Thèse présentée pour obtenir le grade de

Docteur de l'Université Louis Pasteur de Strasbourg

discipline: Aspects moléculaires et cellulaires de la biologie

Par

Stéphanie Boué

Transcripts in Space and Time

Soutenue le 28 avril 2006 devant la commission d'examen:

Dr. James Stévenin	Directeur de Thèse
Dr. Peer Bork	Directeur de Thèse
Dr. Olivier Poch	Rapporteur Interne
Prof Annalisa Pastore	Rapporteur Externe
Dr Toby Gibson	Rapporteur Externe
Prof Jean-Marc Jeltsch	Examinateur

Acknowledgments

I wish to thank Dr Peer Bork for giving me the opportunity to pursue my PhD thesis research in his group as well as present and former members of the Bork group. I wish to thank particularly Mathilde, Ivica, Eoghan, Sean, Francesca, Jeroen and David for sharing with me not only the science but also the life in Heidelberg.

Merci beaucoup à James Stévenin pour son aide précieuse notamment avec les formalités et surtout en fin de thèse. Son enthousiasme est réèllement communicatif.

Thanks a lot to the members of my Thesis Advisory Committee, Peer Bork, Toby Gibson, Iain Mattaj and James Stévenin, thanks to whom I could stay focused and manage my research.

Merci à Jean-Marc Jeltsch et Olivier Poch, grazie mille Annalisa Pastore, thanks to Toby Gibson for accepting to judge my thesis.

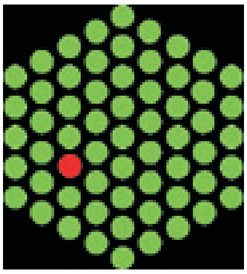
Hvala hrvastka mafijo i prijatelji. Bez vas moj boravak u Heidelbergu bio bi najdosadniji. Sretna sam sto imam kao dobri prijatelji. Dovidjenja i svako dobro ali nije zbogom za sva vremena. Vidimo se u brzo u svoje tajno udruzenje...ili mozda su u sumi! ;-) Filip, Moki, Josipa, Vibor, Ana, Alen, Kreso, thanks for everything, from the support in harder times to the fun at any time. Of course I dont forget my brothers in arms, those "affiliated", who unlike me did not get time to grab some croatian knowledge, so for them: Dilem, Timo, Erwan...we made it into this privileged community, and guess what...they liked it too. And much more than group of strangers in the middle of a croatian mafia, we all are a group of friends, looking forward for a lot more fun.

Drago Maki, hvala ti moja posljednja godina bila je najbolja, and I am sure the best is still to come.

Merci enfin à ma famille, ainsi que Julie et Cécile, qui même si elles n'ont pas toujours compris ce que je faisais et pourquoi, m'ont toujours soutenue.

My PhD thesis was financed through the ASD (Alternative Splicing Database) consortium funding by the EC Grant QLK3-CT-2002-02062, under its Fifth Framework Programme (FP5) - "Quality of Life and living resources" (QoL) section.

The work presented in this thesis has been conducted
in Peer Bork's group
from the Computational and Structural Biology Department
at the EMBL - Meyerhofstrasse 1, D-69117 Heidelberg, Germany



Abbreviations

AS	**A**lternative **s**plicing or **A**lternatively **s**pliced
ATP	**A**denosine **t**ri**p**hosphate
CDS	**C**o**d**ing **s**equence
ChIP	**Ch**romatin **i**mmuno**p**recipitation
CTD	**C**arboxy-**t**erminal **d**omain
DNA	**D**eoxyribo**n**ucleic **a**cid
EJC	**E**xon **j**unction **c**omplex
mRNA	**M**essenger RNA
NAS	**N**onsense-associated **a**ltered **s**plicing
NICD	**N**otch **i**ntra**c**ellular **d**omain
NMD	**N**onsense (mRNA) **m**ediated **d**ecay
NPC	**N**uclear **p**ore **c**omplex
ORF	**O**pen **r**eading **f**rame
PTC	**P**remature (translation) **t**ermination **c**odon
RNA	**R**ibo**n**ucleic **a**cid
RNAi	RNA **i**nterference
RNAP	RNA **p**olymerase
rRNA	**R**ibosomal RNA
snRNA	**S**mall **n**uclear RNA
snRNP	**S**mall **n**uclear **r**ibo**n**ucleo**p**rotein
tRNA	**T**ransfer RNA

Amino acids

One-letter-code	Three-letter-code	Full name
Non polar amino acids (hydrophobic)		
G	Gly	glycine
A	Ala	alanine
V	Val	valine
L	Leu	leucine
I	Ile	isoleucine
M	Met	methionine
F	Phe	phenylalanine
W	Trp	tryptophan
P	Pro	proline
Polar (hydrophilic)		
S	Ser	serine
T	Thr	threonine
C	Cys	cysteine
Y	Tyr	tyrosine
N	Asn	asparagine
Q	Gln	glutamine
Electrically charged (negative and hydrophilic)		
D	Asp	aspartic acid
E	Glu	glutamic acid
Electrically charged (positive and hydrophilic)		
K	Lys	lysine
R	Arg	arginine
H	His	histidine

Abstract

Molecular biologists aim at the understanding of organisms at the molecular level. The ultimate goal is to have the possibility to safely manipulate cells and/or organisms in order to heal genetic diseases, eradicate contagious diseases or for example improve nutrient qualities of food. Currently the most accurate and practical way to capture the functioning of an organism is to look at its transcriptome and its spatial and temporal variations. Following this logic, the focus of my PhD thesis has been two folds: (1) estimate the importance of alternative splicing in the generation of transcript diversity (2) study the transcriptomes of two model organisms: *Mus musculus* and *Drosophila melanogaster*, respectively in a spatial and in a temporal dimension.

Along these years of research I gathered interesting findings on gene expression and its regulation. First, alternative splicing proved to be an important mechanism both in terms of frequency (alternative transcripts are generated for a vast majority of genes and in many species) and evolution (it seems to allow a gene to evolve with manageable consequences for the organism). Moreover we were able to prove that levels of gene expression at the transcript level do not automatically imply function: there is a non negligible amount of neutral expression which has to be taken into account when inferring function according to similarities in expression patterns. Lastly we investigated time series microarray data by applying an innovative technique which allowed grouping of genes into classes according to an original expression profiles criterion (”consistent changes”), and could show that this grouping makes biological sense, and hence that unknown or poorly characterized genes within these groups might be worth investigating further.

An inestimable insight on molecular biology has been and will be gained thanks to studies of the transcriptomes of different organisms in various conditions. However, the full picture seems to only be accessible with proteomics data due to the number of regulatory steps still present after the transcript level, among which alternative splicing.

Résumé

Les biologistes moléculaires cherchent à comprendre comment fonctionnent les organismes au niveau moléculaire. Le but ultime de ces recherches est d'offrir la possibilité de manipuler sans risque des cellules et/ou des organismes afin de combattre des maladies génétiques, d'éradiquer les maladies contagieuses ou par example d'améliorer les qualités nutritives de l'alimentation. Actuellement, la manière la plus précise et pratique de comprendre le fonctionnement d'un organisme est d'étudier son transcriptome et ses variations dans l'espace et le temps. Suivant cette logique, le but de ma thèse de doctorat a été double: (1) estimer l'importance de l'épissage alternatif qui engendre une diversité des transcripts (2) étudier les transcriptomes de deux organismes modèles : *Mus musculus* et *Drosophila melanogaster*, respectivement dans l'espace et le temps.

Durant ces années de recherche, j'ai rassemblé des découvertes intéressantes concernant l'expression des gènes et sa régulation. D'abord, l'épissage alternatif s'est avéré être un méchanisme important non seulement en terme de fréquence (des transcripts alternatifs sont générés pour une vaste majorité des gènes, et ce dans de multiples espèces), mais aussi en terme d'évolution (l'épissage alternatif semble permettre à un gène d'évoluer sans conséquences trop négatives pour l'organisme). Par ailleurs nous avons prouvé que le niveau d'expression de transcripts n'est pas en soi synonyme de fonction: il y a en effet une quantité non négligeable d'expression neutre, qui doit être prise en compte lors de l'assignation d'une fonction à un gène, uniquement basée sur la similarité de son profil d'expression par rapport à celui d'un gène de fonction connue. Enfin, nous avons étudié des séries de puces à ADN appliquées à l'embryogenèse de la mouche dans le temps, en utilisant une technique non conventionnelle pour ce type d'approche. Nous avons réparti les gènes en différentes classes selon leurs profils d'expression. Nous avons pu prouver que ces classes de gènes ont des critères biologiques en commun, ce qui laisse supposer que les gènes inconnus ou mal caractérisés qui tombent dans ces classes sont d'interessants points de départ pour de futures recherches.

Des découvertes inestimables ont été et seront encore faites en biologie moléculaire grâce à l'étude des transcriptomes dans des organismes variés, analysés dans différentes conditions. Cependant, il est devenu clair qu'à cause de la présence de nombreuses étapes de régulation après la transcription, dont l'épissage alternatif, seule l'analyse des protéomes permettra d'obtenir une vision complète de la biologie de la cellule.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Bioinformatics

Ever since the announcement of the first protein sequence (bovine insulin, 51 amino acids) by Frederick Sanger in 1955 [1], and the first nucleotide sequence a decade later, protein and nucleotide sequences have accumulated around the world. With the automation of sequencing techniques, and the development of large scale sequencing projects (see table 1.1), sequences accumulated at an exponential speed, creating the need for a careful storage, organization and indexing of sequence information. Information science has been applied to biology to produce the field called ”bioinformatics”. Dr Margaret Dayhoff was in this sense a pioneer in the usage of computers in biology and is considered as the founder of bioinformatics with her Atlas of Protein Sequence and Structure published in 1965 [2].

The first tasks accomplished in bioinformatics concern the creation and maintenance of databases of biological information. Typically, a database includes an archive of information, a logical organization of that information and tools to gain access to it. Sequence databases (of nucleic acids or proteins) comprise the majority of biological databases. They include not only the raw sequences but also annotation and usually cross-references to other databases. The growth of one of the major protein databases, SwissProt is illustrated in figure 1.1.

However today’s bioinformatics is more than database depository: computer algorithms are developed and used to make inferences from the archived data and make connections among them in order to derive useful and interesting predictions. This field, called ”computational biology”, involves sequence information analysis in order to (list not exhaustive):

- Find the genes in the DNA sequences of various organisms.
- Develop methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences.

Organism	Complete	Draft Assembly	In progress	Total
Prokaryotes	294	190	371	855
Archaea	25	3	23	51
Bacteria	269	187	348	804
Eukaryotes	18	66	140	224
Animals	4	22	62	88
Mammals	2	9	17	28
Birds		1		1
Fishes		2	2	4
Insects	1	6	26	33
Flatworms			2	2
Roundworms	1	2	3	6
Amphibians			1	1
Reptiles				0
Other animals		2	13	15
Plants	2	1	29	32
Land plants	2	1	23	26
Green Algae			6	6
Fungi	9	33	19	61
Ascomycetes	7	29	15	51
Basidiomycetes	1	3	2	6
Other fungi	1	1	2	4
Protists	3	10	27	40
Apicomplexans		5	9	14
Kinetoplasts	1	2	3	6
Other protists	2	3	14	19
total:	312	256	511	1079

Table 1.1: Genome sequencing projects statistics from (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html), as of 01/15/2006

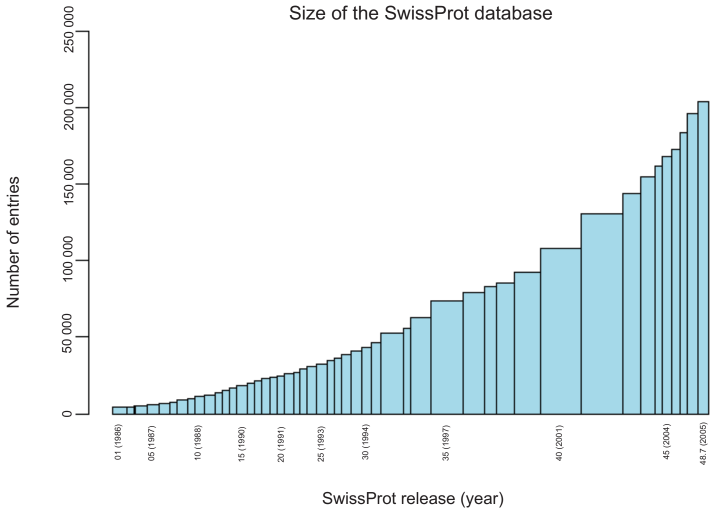


Figure 1.1: **Growth of the SwissProt database**

- Cluster protein sequences into families of related sequences and develop protein models.
- Align similar proteins and generate phylogenetic trees to examine evolutionary relationships.

1.1 Sequence analysis

The focus on sequence analysis was induced not only by the multiplication of genome sequencing projects (see table 1.1), but also by an aggravation of the sequence/structure deficit. One of the main challenges of bioinformatics is to convert sequence information into biochemical and biophysical knowledge in order to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences. However structures analysis, which was earlier an important part of bioinformatics cannot keep the pace of genome sequencing. Structure determination is indeed more tricky and not as easily scalable as sequencing techniques. Ultimately, the challenge is to fully understand nucleic acid language (and derived amino acid language) and to be able to use it to design our own proteins.

In the field of sequence analysis, two directions are taken: first, pattern recognition techniques are applied to detect similarities between sequences and infer related structure and function. Secondly, *ab initio* prediction methods are built, which will infer the function of a protein by deducing its 3D structure from the linear sequence. *Ab initio* prediction methods are still far from being reliable. On the contrary, pattern recognition techniques are already quite advanced. Indeed, since sequences are available from various organisms, sequence comparison algorithms have been developed, starting with the Needleman-Wunsch one in 1970 [3].

Using sequence analysis techniques, it is possible to identify similarities between novel (predicted or newly sequenced) query sequences of unknown function and database sequences whose structures and functions have been elucidated. This remains quite straightforward as long as the similarity level remains high enough, but below 50% similarity, it really becomes difficult to establish relationships in a reliable manner. When speaking about similarity, a distinction should be made between homology and analogy. Sequences are said to be homologous if they are related by divergence from a common ancestor. Analogous relationships are the result from convergence to similar biological solutions from different evolutionary starting points. Moreover there is a distinction among homologous sequences between proteins that perform the same function in different species (orthologs) and those that perform different but related functions within one organism (paralogs). They also give different insights into evolution mechanisms. It is important that the distinction is made when assigning properties based on sequence similarity. Function is indeed conserved only among orthologs.

1.2 Major databases

Databases gathering data on about any subject flourished during the last decade, but the curation, annotation and quality control levels of those databases make them more or less reliable, and some databases imposed themselves as standards. Thus the main databases used during this project are the following:

- **Literature retrieval**
  - **Pubmed**: Available via the NCBI Entrez retrieval system, Pubmed was developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), located at the National Institutes of Health (NIH). It was designed to provide access via a text search to citations from biomedical literature. The database contains over 15 million citations dating back to the 1950's. Coverage is worldwide, but most records are from English-language sources or have English abstracts. It is available at the following url:
    http://www.ncbi.nih.gov/entrez/query.fcgi?db=PubMed
  - **Google scholar**: Google Scholar provides a simple way to broadly search for scholarly literature. From one place, you can search across many disciplines and sources: peer-reviewed papers, theses, books, abstracts

and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations. Google ranking technology considers the full text of each article, the author, the publication in which the article appeared, and how often the piece has been cited in other scholarly literature in order to provide the most useful references first. It is available at the following url:
http://scholar.google.com/

- **Protein databases**
  - **UniProt Knowledgebase**: Can be queried at the following url: http://www.expasy.org/ or
    http://www.ebi.uniprot.org/uniprot-srv/index.do, is composed of:
    - **UniProtKB/Swiss-Prot**: A manually curated protein sequence database established in 1986 and maintained since 2003 by the UniProt Consortium, a collaboration between the Swiss Institute of Bioinformatics (SIB) and the Department of Bioinformatics and Structural Biology of the Geneva University, the European Bioinformatics Institute (EBI) and the Georgetown University Medical Center's Protein Information Resource (PIR). It strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with more than 60 other databases. UniProtKB/Swiss-Prot Release 49.1 of 21-Feb-2006 contains 208,005 entries.
    - **UniProtKB/TrEMBL**: A computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot. UniProtKB/TrEMBL Release 32.1 of 21-Feb-2006 contains 2,618,388 entries.
- **Nucleotide databases**
  - **EMBL nucleotide sequence database**: The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). The EMBL Nucleotide Sequence Database was frozen to make Release 85 on 30-NOV-2005. The release contains 64,739,883 sequence entries comprising 116,106,677,726 nucleotides, of which 12,088,383 entries (59,629,958,692 nucleotides) are WGS (whole genome shotgun) data. It is available at the following url:
    http://www.ebi.ac.uk/embl/Access/index.html

  - **Entrez nucleotides database**: The Entrez Nucleotides database is a collection of sequences from several sources, including GenBank, RefSeq, and PDB. The number of bases in these databases continues to grow at an exponential rate. As of June 2005, there are over 89 billion bases in GenBank and RefSeq alone. It is available at the following url: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
  - **Genome databases**: Two main depository of genomes are available to blast against:
    - Genbank
    - UCSC genome browser: Fully sequenced genomes for human, chimp, rhesus macaque, dog, cow, mouse, rat, opossum, chicken, Xenopus, zebrafish, tetraodon, and fugu are available for browsing or querying against at the following url: http://genome.ucsc.edu/

- **Protein interaction network, pathways**
  - **STRING** (Search Tool for the Retrieval of Interacting Genes/Proteins): A database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:
    - Genomic Context
    - High-throughput Experiments
    - (Conserved) Co-expression
    - Previous Knowledge

    STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently contains 736,429 proteins in 179 species. STRING is available at the following url: http://string.embl-heidelberg.de/
  - **KEGG** (Kyoto Encyclopedia of Genes and Genomes): A suite of databases and associated software, integrating the current knowledge on molecular interaction networks in biological processes (PATHWAY database), the information about the universe of genes and proteins (GENES/SSDB/KO databases), and the information about the universe of chemical compounds and reactions (COMPOUND/DRUG/GLYCAN/REACTION databases). We mainly used the pathway database, which comprises 35,779 pathways

in different species generated from 281 reference pathways. It is available at the following url:
http://www.genome.jp/kegg/pathway.html

- **Domain databases**

  - **SMART** (Simple Modular Architecture Research Tool): Allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signaling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phylogenetic distributions, functional class, tertiary structures and functionally important residues. Each domain found in a non-redundant protein database as well as search parameters and taxonomic information are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa. SMART is available at the following url:
    http://smart.embl-heidelberg.de/
  - **INTERPRO**: A database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. Release 12.0 contains 12,542 entries. It is available at the following url:
    http://www.ebi.ac.uk/interpro/

- **Ontology databases**

  - **Gene ontology**: Provides a controlled vocabulary to describe gene and gene product attributes in any organism. The GO collaborators are developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. The use of GO terms by several collaborating databases facilitates uniform queries across them. The three organizing principles of GO are molecular function, biological process and cellular component. A gene product has one or more molecular functions and is used in one or more biological processes; it might be associated with one or more cellular components.

    * **Molecular function**: Describes activities, such as catalytic or binding activities, at the molecular level. GO molecular function terms

    represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products.

    ∗ **Biological process**: A series of events accomplished by one or more ordered assemblies of molecular functions.
    ∗ **cellular component**: A component of a cell but with the proviso that it is part of some larger object, which may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

  It is available at the following url:
  http://www.geneontology.org/

- **Databases dedicated to *Drosophila melanogaster***
  – **FlyBase**: A database of genetic and molecular data for *Drosophila*. FlyBase includes data on all species from the family Drosophilidae; the primary species represented is *Drosophila melanogaster*. FlyBase includes notably the following information:
    ∗ Information on genes and mutant alleles;
    ∗ Information about the expression and properties of transcripts and proteins;
    ∗ Information on the functions of gene products;
    ∗ Nucleic acid accession numbers linked from gene records;
    ∗ Protein sequence accession numbers linked from protein records;
    ∗ Images that illustrate *Drosophila* anatomy and development terms;
    ∗ A bibliography of *Drosophila* citations;
    ∗ *Drosophila* genetic, cytological, and molecular map information.

    It is available at the following url:
    http://www.flybase.org
  – **The interactive fly**: A cyberspace guide to *Drosophila* development and metazoan evolution. It is available at the following url:
    http://flybase.bio.indiana.edu/allied-data/lk/interactive-fly/aimain/1aahome.htm

  - **Berkeley Drosophila Genome Project**: Stores, among others, information on expression patterns for *Drosophila* during embryogenesis. They use a high throughput 96-well plate RNA in situ protocol to determine patterns of gene expression during embryogenesis for *Drosophila* genes represented in non-redundant sets of *Drosophila* ESTs. At the end of the production pipeline, gene expression patterns are documented by taking a large number of digital images of individual embryos. The quality and identity of the captured image data are verified by independently derived microarray time-course analysis of gene expression using Affymetrix GeneChip technology. Gene expression patterns are annotated with controlled vocabulary for developmental anatomy of *Drosophila* embryogenesis. So far the expression of 5,270 genes has been examined and documented; 3,012 of them with 56,644 digital photographs. It is available at the following url:
    http://www.fruitfly.org/cgi-bin/ex/insitu.pl

- **Alternatively spliced variants** See section 3.5.

1.3 Literature mining

Data mining (also known as Knowledge Discovery in Databases - KDD) has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [4]. A very fast-growing subtype of data mining is literature mining, which aims at automatically retrieving relevant information from the millions of scientific publications available (reviewed in [5]). Because the generation of experimental and *in silico* data mostly happens in a large scale fashion, and trustworthy datasets are needed to evaluate them, a large variety of tools are constructed to extract published data automatically. Moreover, because of the exponential growth of scientific literature, literature-mining tools are really much needed by all researchers to keep an overview on their research field. Tools can be categorized into the following types:

- **Information retrieval tools**: They mainly aim at identifying relevant papers of a topic (e.g. pubmed, google scholar)

- **Information extraction tools**: They go a level deeper, and seek biological information within the scientific publication and extract it. Those tools developed more recently just start to be accurate enough to be used meaningfully.

Chapter 2

Gene expression

One of the big mysteries in life is how the complexity of living organisms is achieved, and able to evolve. It is indeed intriguing that all cells in one organism inherit the same genetic information in the form of DNA molecules, but are able to carry very different functions, and have various morphologies. There are nowadays tools which allow scientists to investigate this ”life wonder” in an extensive way. A lot of understanding has been gained, and although it often brings even more questions than it gives answers, we know now that cell differentiation usually depends on differential transcriptional programs, and not on DNA polymorphisms or gene loss in the respective cell lineages. That is a reason why transcriptomes are extensively studied, among others using microarray technology.

2.1 Genomes

Each species of living organism is united by a common set of inherited traits, observable characters that set it apart from all other species. These traits are passed from one generation to the next thanks to the molecule of heredity, DNA. The gene complement of one cell, i.e. the part of the DNA molecules that gives eventually rise to traits, is called genome.

2.1.1 Gene number and organism complexity

The existence of genes and the rules governing their transmission from generation to generation were discovered by Gregor Mendel in 1866. Numerous advances have lead us to our current knowledge, notably the discovery of the double helix structure of DNA, the genetic code for proteins, the central dogma...

The genetic complement of a cell constitutes its genome. In eukaryotes, this term is commonly used to refer to one complete haploid set of chromosomes, such as found in germ cells. Viral genomes are typically in the range 10-1000kb, bacterial genomes typically in the range 1-10Mb, and eukaryotic genomes in the range 10-1000Mb (see table 2.1).

According to the latest estimates, a human cell genome contains between 20,000 and 25,000 genes [6], which is much less than first estimated, and interestingly not much more than lower eukaryote species such as *Drosophila melanogaster* (See table 2.1). Complexity can hence not be explained only by the gene number of a species.

Organism	Genome size (bp)	Gene number	Notes
Human mitochondrion	16,569	37	
Epstein-Barr virus (EBV)	172,282	80	Causes mononucleosis
Pelagibacter ubique	1,308,759	1,354	Smallest genome yet found in a free-living organism
Escherichia coli	4,639,221	4,377	
Schizosaccharomyces pombe	12,462,637	4,929	Fission yeast
Saccharomyces cerevisiae	12,495,682	5,770	Budding yeast
Caenorhabditis elegans	100,258,171	19,427	First multicellular organism to be sequenced
Arabidopsis thaliana	115,409,949	28,000	Flowering plant
Drosophila melanogaster	122,653,977	13,379	Fruit fly
Anopheles gambiae	278,244,063	13,683	
Homo sapiens	3.3 x 10^9	20,000-25,000	
Rice	3.9 x 10^8	37,544	

Table 2.1: Gene number estimates in haploid fully sequenced genomes, from http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html, as of 01/31/2006

2.1.2 Genome composition

Thus, in nearly all higher animals and plants, the actual number of genes has little relationship to genome size. The C-value is the basal genome size of an organism, defined as the content of DNA (measured by weight or number of base pairs) in a single copy of the entire sequence of DNA found within cells of that organism. The C-value paradox is a term used to describe the discrepancy between nuclear genome size and the number of genes among eukaryotic species.

The discovery of non-coding DNA in the early 1970s resolved the C-value paradox. It is no longer a mystery why genome size does not reflect gene number in eukaryotes: most eukaryotic (but not prokaryotic) DNA is non-coding and therefore does not consist of genes, and as such total DNA content is not determined by gene number in eukaryotes. The human genome, for example, is comprised of only about 1.5% protein-coding genes, with the other 98.5% being various types of non-coding DNA (especially transposable elements [7], see figure 2.1).

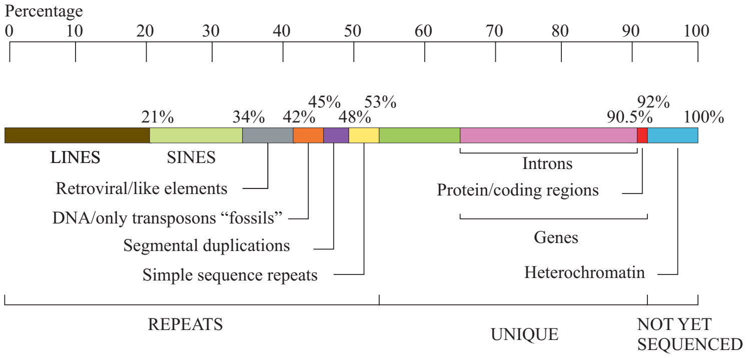


Figure 2.1: **Representation of the nucleotide sequence content of the human genome.** (Adapted from Unveiling the Human Genome, Supplement to the Wellcome Trust Newsletter. London: Wellcome Trust, February 2001.)

2.2 Gene regulation in eukaryotes

Genes can be classified according to their expression patterns:

- Genes expressed in all cells all the time. These so-called housekeeping genes are responsible for the routine metabolic functions (e.g. respiration) common to all cells.
- Genes expressed as a cell enters a particular pathway of differentiation.
- Genes expressed all the time in only those cells that have differentiated in a particular way (for example immune cells).
- Genes expressed only as conditions around and in the cell change. For example, the arrival of a hormone may turn on (or off) certain genes in that cell.

The control of eukaryotic transcription requires the sequential interaction and precise coordination of a variety of large enzymatic complexes that are recruited by sequence-specific promoter/enhancer-binding proteins. Regulation is exerted at all levels of the process:

- **Chromatin structure**: By changing the accessibility of chromatin to transcriptional regulatory proteins and RNA polymerase.

- **Transcriptional initiation**: Mainly depends on the strength of promoter elements, the presence or absence of enhancer sequences and the interaction between multiple activators and silencers.

- **Transcript processing and modification**: Capping, polyadenylation and splicing.

- **RNA transport**

- **RNA editing**: Alteration of the sequence of nucleotides in the RNA after it has been transcribed from DNA but before it is translated into protein. RNA editing occurs by two distinct mechanisms:

  - Substitution editing: Chemical alteration of individual nucleotides (the equivalent of point mutations), catalyzed by enzymes that recognize a specific target sequence of nucleotides.
  - Insertion/deletion editing: Insertion or deletion of nucleotides in the RNA.

- **Transcript stability**: Variates greatly, depending on the presence of signal for rapid degradation, usually in 3' UTR region of the transcript.

- **Translation initiation**: Depends on the recognition of the start codon among different methionine codons.

- **Post-translational modifications**: Numerous, among which: phosphorylation, methylation, glycosylation, acetylation or disulfide bond formation (see review [8]). They play a very important biological role by modifying the protein structure and/or function, and have been implicated in aging [9] and auto-immunity [10] for example.

- **Protein transport**

- **Control of protein stability**

2.2.1 Chromosomal DNA and its packaging in the chromatin fiber

In eukaryotes, the DNA in the nucleus is divided between a set of different chromosomes. For example, the human genome is distributed over 24 different chromosomes (22 pairs of autosomes and the sexual chromosomes X and Y). Each human cell contains approximately 2 meters of DNA if stretched end-to-end; yet the nucleus of a human cell, which contains the DNA, is only about 6 micrometer in diameter. The complex task of packaging DNA is accomplished by specialized proteins that bind to and fold the DNA, generating a series of coils and loops that provide increasingly higher levels of organization, preventing the DNA from becoming an unmanageable tangle.

The first level of compaction, also known as nucleosome core particle, is organized by histone proteins. Each individual nucleosome core particle consists of a complex of eight histone proteins - two molecules each of histones H2A, H2B, H3, and H4 - and double-stranded DNA that is 146 nucleotide pairs long. Each nucleosome core particle is separated from the next by a region of linker DNA, which can vary in length from a few nucleotide pairs up to about 80. Although nearly every DNA sequence can, in principle, be folded into a nucleosome, the spacing of nucleosomes in the cell can be irregular, depending on DNA composition, and/or on the presence of other proteins binding the DNA molecule.

Although long strings of nucleosomes form on most chromosomal DNA, chromatin in a living cell probably rarely adopts the extended ”beads on a string” form. Instead, the nucleosomes are packed on top of one another, generating regular arrays in which the DNA is even more highly condensed, and form the 30-nm fiber.

In addition to the proteins involved in packaging the DNA, chromosomes are also associated with many proteins required for the processes of gene expression, DNA replication, and DNA repair.

2.2.2 Chromatin remodeling

The organization of chromatin poses a barrier to transcription because it prevents the transcription machinery from interacting directly with promoter DNA sequences. That is why one of the earliest steps of gene activation involves the mobilization of energy-dependent chromatin remodeling complexes, protein machines that use the energy of ATP hydrolysis to change the structure of nucleosomes temporarily so that DNA becomes less tightly bound to the histone core (see figure 2.2) and is accessible to auxiliary transcription factors (for review see [11]).

There are two classes of chromatin remodeling enzymes:

- Those that covalently modify nucleosomal histone proteins through acetylation, phosphorylation, or methylation;

- Those that alter chromatin structure through hydrolysis of ATP.

Some histone-modifying enzymes and some ATP-dependent remodeling enzymes directly interact with gene-specific activators to ensure that chromatin remodeling is targeted to the correct gene, in the proper cell, and at the right time [12].

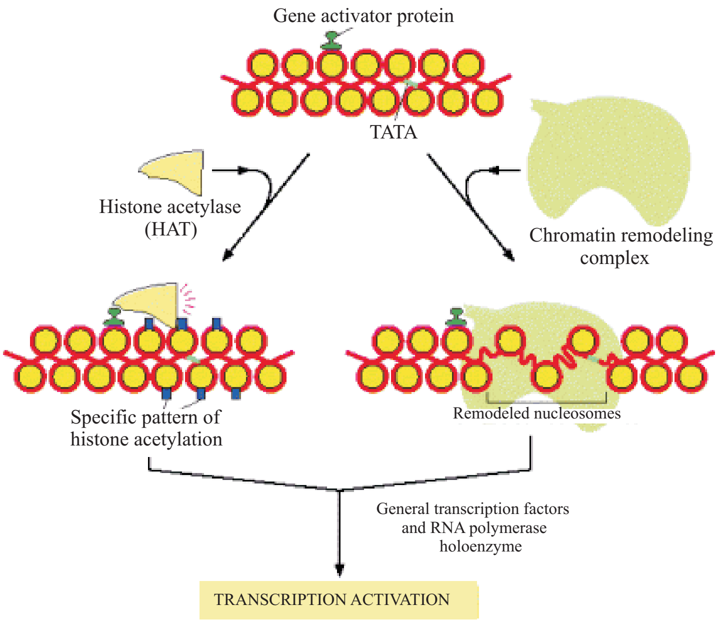

Gene activator protein
TATA
Histone acetylase (HAT)
Chromatin remodeling complex
Specific pattern of histone acetylation
Remodeled nucleosomes
General transcription factors and RNA polymerase holoenzyme
TRANSCRIPTION ACTIVATION

Figure 2.2: **Local alterations in chromatin structure directed by eukaryotic gene activator proteins.**
From [13].

2.2.3 Transcription

Transcription is the transfer of genetic information from the archival copy of DNA to the short-lived messenger RNA. The enzyme RNA polymerase, which is a nucleotidyltransferase, binds to a particular region of the DNA (the promoter) and starts to make a strand of mRNA with a base sequence complementary to the DNA template that is downstream of the RNA polymerase binding site. When the transcription is finished, the portion of the DNA that coded for a protein, i.e. a gene, is represented by a messenger RNA molecule that can be used as a template for translation.

2.2.3.1 RNA polymerase II

RNA polymerases (RNAP) are large, multisubunit complexes.

Eukaryotes have several types of RNAP:

- **RNAP I**: Synthesizes a pre-rRNA 45S, which matures into 28S, 18S and 5,8S rRNAs which will form the major RNA sections of the ribosome.
- **RNAP II**: Synthesizes precursors of mRNAs and most snRNA.
- **RNAP III**: Synthesizes tRNAs, 5S rRNA and other small RNAs found in the nucleus and cytosol.
- Other RNAP types in mitochondria and chloroplasts.

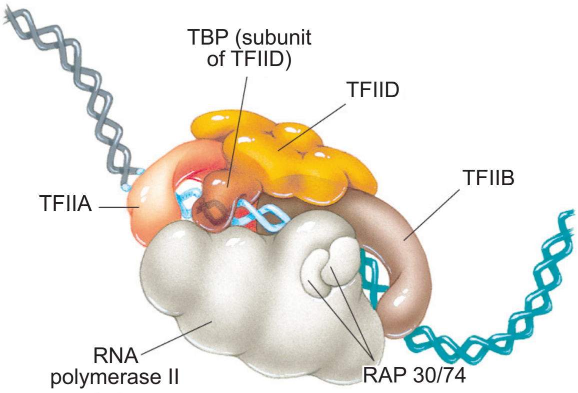


Figure 2.3: **RNAP II holoenzyme complex bound to a promoter**
This model shows various transcription factors bound to RNAP II at the promoter. The transcription factors are often larger and more complex than those shown in this diagram. From [14]

The RNAP II transcription machinery is the most complex, with a total of nearly 60 polypeptides, only a few of which are required for transcription by the other nuclear polymerases (see table 1 in review [15] and [16]).

RNAP II itself is composed of subunits that can be classified into three overlapping categories:

- Subunits of the core domain having homologous counterparts in bacterial polymerase (Rpb1, 2, 3, and 11);
- Subunits shared between all three nuclear polymerases (Rpb5, 6, 8, 10, and 12);
- Subunits specific to RNAP II but not essential for transcription elongation (rpb4, 7, and 9).

The largest subunit of RNAP II has a unique domain, not related to regions in any known protein, at its carboxyl terminus, termed the carboxy-terminal domain (CTD). The CTD contains 25-52 repeats of the tandemly repeated heptad sequence YSPTSPS, with both Ser2 and Ser5 the sites of phosphorylation. Regulatory phosphorylation and dephosphorylation of the CTD is part of the transcription cycle. The CTD acts hence as a platform for assembly of factors that regulate transcription initiation, elongation, termination, and mRNA processing.

2.2.3.2 Promoter

Promoter sequences promote the ability of RNAP II to recognize the nucleotide at which initiation begins. Recognition of the core promoter of the RNAP II is essential for correct positioning and assembly by RNAP II and the general factors. The strength of the binding of RNAP II to different promoters varies greatly, which causes differences in the extent of expression from one gene to another.

Specific DNA elements within the core promoter bind the factors that nucleate the assembly of a functional preinitiation complex and integrate stimulatory and repressive signals from factors bound at distal sites. However, the core promoter recognized by the RNAP II is not constant (see review [17]).

The core promoter includes DNA elements that can extend approximately 35 bp upstream and/or downstream of the transcription initiation site. The basal promoter contains different regions:

- **The TATA box**: A sequence of 7 bases (TATAAAA) usually present 25 to 30 bp upstream of the transcription start site. It is bound by a large complex of some 50 different proteins, including:
  - Transcription Factor IID (TFIID) which is a complex of
    - TATA-binding protein (TBP), which recognizes and binds to the TATA box;
    - 14 other protein factors which bind to TBP, and to each other, but not to the DNA.
  - Transcription Factor IIB (TFIIB) which binds both the DNA and RNAP II.
- Another sequence, the **CCAAT-box** (consensus GGT/CCAATCT) resides 50 to 130 bases upstream of the transcriptional start site. It is bound by the C/EBP (for CCAAT-box/Enhancer Binding Protein).

The basal or core promoter is found in all protein-coding genes. This is in sharp contrast to the upstream promoter whose structure and associated binding factors differ from gene to gene, conferring different expression patterns for different genes.

2.2.3.3 Transcription initiation

The complex for initiation of transcription is formed by four types of proteins (see figure 2.4):

1. **Basal transcription factors**: In response to activators, they position the RNAP II at the start of the protein-coding genes.
2. **Activators**: They bind to genes at enhancer sites and determine which genes are turned on, and at which speed they are transcribed.
3. **Coactivators**: Adaptor molecules integrating the signals from activators and repressors to relay the result to basal transcription factors.
4. **Repressors**: They bind to selected genes at silencer sites and interfere with activators to slow down transcription.

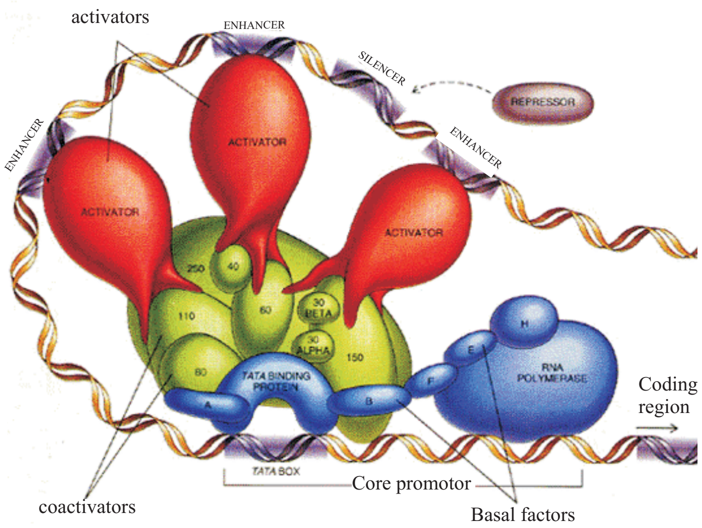


Figure 2.4: **Transcription initiation machinery**
A drawing of the transcription machine, showing where the ”heart” binds to the TATA box in the promoter upstream of the gene (right), turning the gene on. Each colored blob is a complex of several proteins, all of which work together to transcribe the gene into messenger RNA. Scientific american image.

To begin transcription the RNAP II requires a number of general transcription factors (called TFIIA, TFIIB, and so on). The TATA box is recognized and bound by the transcription factor TFIID, which then enables the adjacent binding of TFIIB. The rest of the general transcription factors, as well as the RNAP II itself, assemble at the promoter. TFIIH then uses ATP to pry apart the DNA double helix at the transcription start point, allowing transcription to begin. TFIIH also phosphorylates RNAP II on the CTD domain, changing its conformation so that the polymerase is released from the general factors and can begin the elongation phase of transcription. In some cases, the general factors are thought to first assemble with the polymerase, with the whole assembly subsequently binding to the DNA in a single step.

2.2.4 Ubiquitination

The life span of intracellular proteins varies from as short as a few minutes for mitotic cyclins, which help regulate passage through mitosis, to as long as the age of an organism for proteins in the lens of the eye. Cells have several intracellular proteolytic pathways for degrading misfolded or denatured proteins, normal proteins whose concentration must be decreased, and foreign proteins taken up by the cell. The best-understood pathway, the ubiquitin-mediated pathway, involves two steps: addition of a chain of 76 amino acids (ubiquitin) to an internal lysine side chain of a target protein and proteolysis of the ubiquitinated protein by a proteasome. The numerous proteasomes present in the cell cytosol proteolytically cleave ubiquitin-tagged proteins in an ATP-dependent process that yields peptides and intact ubiquitin molecules. To be targeted for degradation by the ubiquitin mediated pathway, a protein must contain a structure that is recognized by a ubiquitinating enzyme complex. Different conjugating enzymes recognize different degradation signals in target proteins. Internal sequences enriched in proline, glutamic acid, serine, and threonine (PEST sequences) are recognized by some of these enzymes.

Moreover the life span of many cytosolic proteins is correlated with the identity of the N-terminal residue, suggesting that certain residues at the N-terminus favor rapid ubiquitination. For example, short-lived proteins that are degraded within 3 minutes in vivo commonly have Arg, Lys, Phe, Leu, or Trp at their N-terminus. In contrast, a stabilizing amino acid such as Cys, Ala, Ser, Thr, Gly, Val, or Met is present at the N-terminus in long-lived proteins that resist proteolytic attack for more than 30 hours.

Ubiquitin has also been involved in transcriptional activation [18, 19]. Indeed many transcriptional activators are unstable proteins degraded by the ubiquitin-proteasome pathway. This instability is often used to restrain the activity of an activator at times when its target genes should not be expressed, or to quench its activity so that the transcriptional response to a stimulus can be quickly dampened. The proteasome may help reorganize or disassemble the preinitiation complex, freeing RNAP II to progress into elongation. In this context, the proteasome might not be executing its proteolytic function, but might instead use its ATPases as chaperones for remodeling protein conformations or interactions. Indeed, ubiquitination does not always presage destruction. The presence of ubiquitin attached to an activator protein results in enhanced recruitment of P-TEFb and thus enhanced transcriptional elongation.

2.3 Neutral expression

The assumption ”gene activity equals gene function” descends from the one-to-one correspondences assumed by the Mendelian genetics and adaptive evolution theory: ”gene-trait” and ”trait-function”. The operon model of gene regulation as well as many observations from *Drosophila* experimental genetics fitted this assumption: genes are either ”on” or ”off” depending on if they have to carry a function in the cell or not. Artificial ectopic mutants were hence engineered to analyze phenotypic changes due to expression of a gene in tissues where it is normally not expressed.

It became natural after a while to scientifically study the actual correlation between mRNA and protein expression levels in order to prove or disprove this assumption. First, smaller-scale studies (for example on human liver proteins [20]) showed that the correlation was not as high as expected. Then, larger-scale methods were used to analyze this correlation genome-wide. They are reviewed in [21]. Basically, the most commonly used methods to measure genome-wide mRNA expression are Affymetrix chips and microarrays, while the methods of choice to measure protein levels in a large scale are based on 2-dimensional electrophoresis and mass spectrometry methods.

Reasons invoked then by the lack of correlation between mRNA and protein levels of expression are the following [21]:

- There are many complicated and varied post-transcriptional mechanisms involved in turning mRNA into protein;
- Proteins may differ substantially in their in vivo half lives;
- There is a significant amount of error and noise in both protein and mRNA experiments that limit our ability to get a clear picture.

As a matter of fact, and despite a multitude of possible regulation steps, gene expression was proven recently to not be as tightly regulated as previously thought (review in [22]). A certain amount of neutral expression is there, which does not seem to negatively affect the cells, and as a consequence, expression profiles do not always match functional profiles.

Genomic sequencing efforts have made possible a new approach to genetics called functional genomics, which focuses on genome-wide patterns of gene expression and the mechanisms by which gene expression is coordinated. As the cellular environment changes, through aging, stress, disease, the patterns of expression change as well. Functional genomics allows to study and understand those changes.

Chapter 3

Alternative splicing

One of the major challenges of the post-genomic era is the description and functional characterization of the proteome expressed by a given organism at a given time and in a given cell. However this task is not as straightforward as first thought.

During a long period, the central dogma of molecular biology first enunciated by Crick in 1958 [23], and reaffirmed in 1970 [24] was simplified into the statement "DNA makes RNA makes proteins" (see figure 3.1).
But the discovery of interrupted genes and mRNA splicing in 1977 [25, 26] revolutionized the molecular biology world by adding a new step, mRNA splicing, in the central dogma which was until then broken down into three major steps: replication, transcription and translation.

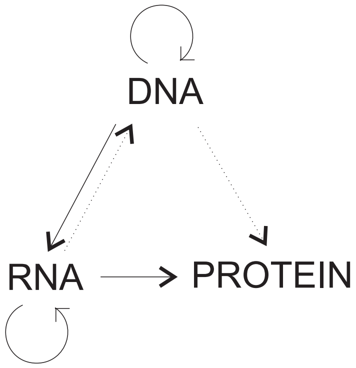


Figure 3.1: **Central dogma** as enunciated by Crick in 1970.
Solid arrows show general transfers; dotted arrows show special transfers.

3.1 RNA splicing

RNA splicing is essential to precisely remove internal non-coding regions of pre-mRNA (introns) and join the remaining segments (exons) in order to yield a mature translatable mRNA .
The process of pre-mRNA splicing can be divided into three stages (see figure 3.2):

1. **Formation of the commitment complex**: The precise recognition of intron-exon junctions (splice sites) and the correct pairing of the 5' splice site with

its cognate 3' splice site is critical for splice site selection. It is during the formation of the commitment complex that splice sites are first recognized by spliceosomal components, with the aid of non-spliceosomal proteins.

2. **Creation of catalytic sites**: A number of dynamic interactions including snRNA-pre-mRNA interactions as well as pre-mRNA-protein and protein-protein interactions bring the reactive sites on the pre-mRNA together and create the catalytic sites for the trans-esterification reactions.

3. **The trans-esterification reactions**: The cleavage and ligation reaction required for intron removal and exon ligation proceeds via two trans-esterification reactions. In the first reaction the 5' exon is cleaved and the 5' end of the intron is joined to the branch point creating the intron lariat structure. The second reaction occurs when the free 3' end of the 5' exon is joined to the downstream exon resulting in exon ligation and release of the intron sequence.

These catalytic reactions are performed and regulated by the spliceosome, a huge assembly of five complexes of RNA and proteins: U1, U2, U4, U5 and U6 snRNPs and more than 100 extrinsic protein factors (review in [27]). The spliceosome assembly has been studied extensively with proteomics approaches yielding models like the one in figure 3.2.

3.2 Coupling transcription, mRNA splicing and mRNA export

Different steps of mRNA processing, namely capping, splicing and 3' end polyadenylation, occur in the nucleus and then the mRNA is exported in the cytoplasm through the nuclear pores to be translated. A coupling of all these processes exists to enable the proofreading and streamlining of the entire process of gene expression.
RNA processing is indeed carried out in close proximity to the site of transcription, allowing for co-transcriptional regulation of alternative pre-mRNA splicing (review in [28]). The coupling of RNAP II transcription and pre-mRNA splicing is required for efficient gene expression. Nascent pre-mRNAs are protected from nuclear degradation because the local concentration of the splicing machinery is sufficiently high to ensure its association over interactions with nucleases. Thus, the coupling of transcription and pre-mRNA splicing guarantees an efficient transfer from the transcription complex to the splicing machinery. The export machinery is also physically and functionally coupled to the splicing one [28].

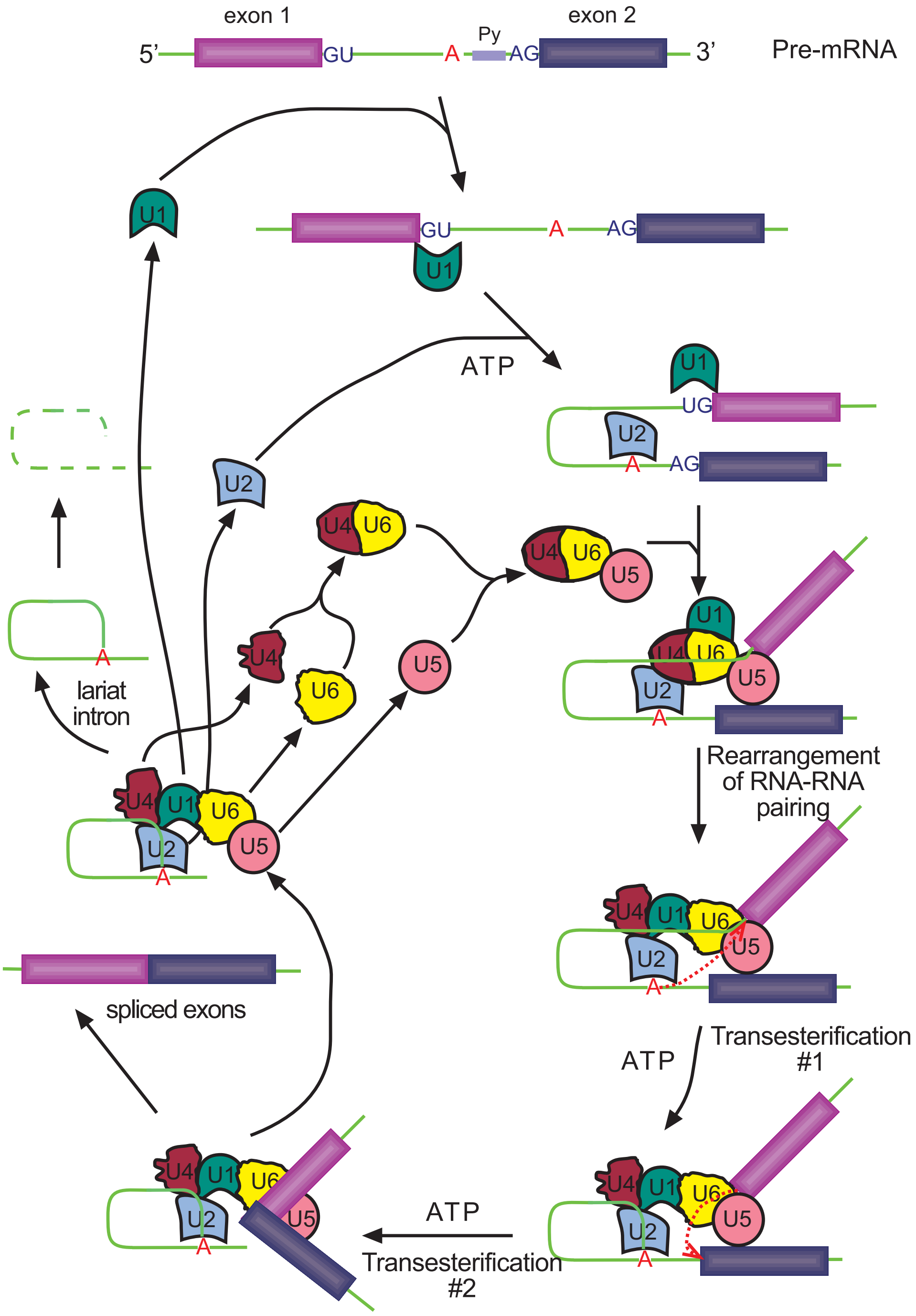


Figure 3.2: **Splicing cycle.**
The splicing snRNPs (U1, U2, U4, U5, and U6) associate with the pre-mRNA and with each other in an ordered sequence to form the spliceosome. This large ribonucleoprotein complex then catalyzes the two transesterification reactions that result in splicing of the exons and excision of the intron as a lariat structure. The branch-point A in pre-mRNA is indicated in red.

3.3 Alternative splicing

The term "alternative pre-mRNA splicing" regroups different kind of mechanisms which, by combining a gene's exons in different manners, give rise to variants of the complete protein. It is a central mode of genetic regulation in higher eukaryotes. Variability in splicing patterns is indeed a major source of protein diversity from the genome, possibly compensating for the low number of genes relatively to the organism's complexity.

Soon after the discovery of interrupted genes, several examples of AS in different tissues were reported [29, 30, 31, 32]. But alternative splicing was thought to be an exceptional event until a thorough analysis of human chromosome 22 and improvement of detection techniques hinted that virtually every human gene is likely to undergo alternative splicing [33].

Not only is alternative splicing a frequent event, it also has a non negligible impact on structure and function [34, 35]. Changes in splice site choice can have different effects on the encoded protein. Small changes in peptide sequence can alter ligand binding, enzymatic activity, allosteric regulation, or protein localization. They can also add or remove complete protein domains, or insert an early stop codon, leading to mRNA degradation (see section on mRNA surveillance).

Moreover AS often leads to the production of tissue or time-specific transcripts at relatively low cost, and has been implicated in numerous human diseases [36] leading to the idea that modifications of splicing pathways are interesting therapy approaches [37]. It is as well of fundamental importance for complete programs of cell differentiation and development, the best documented example being sex determination in *Drosophila melanogaster* (see reviews [38, 39]).

3.4 Detection of AS events

In a typical multiexon RNA, the splicing pattern can be altered in many ways (see figure 3.3). Most exons are constitutive: they are always included in the mature mRNA. A regulated exon that is sometimes entirely included, and sometimes entirely skipped is called a cassette exon. In some cases, and often in case of duplicated exons [40], multiple cassette exons are mutually exclusive: one of them will be included in the final transcript, but never more. Exons can also be lengthened or shortened on either (or both) 3' or 5' side by usage of an alternate 5' or 3' splice site, or an intron can be entirely retained in the mature transcript. Two other situations can happen, which are not considered as alternative splicing events as such, but which also lead to alternative transcripts: usage of alternative promoters, often leading to changes in the 5' part of the transcript or usage of alternative polyadenylation sites, leading to different 3' end of the mature transcript. Both of those events are often associated to tissue or time-specific transcripts.

Alternative splicing discovery became in the last years a very prominent field

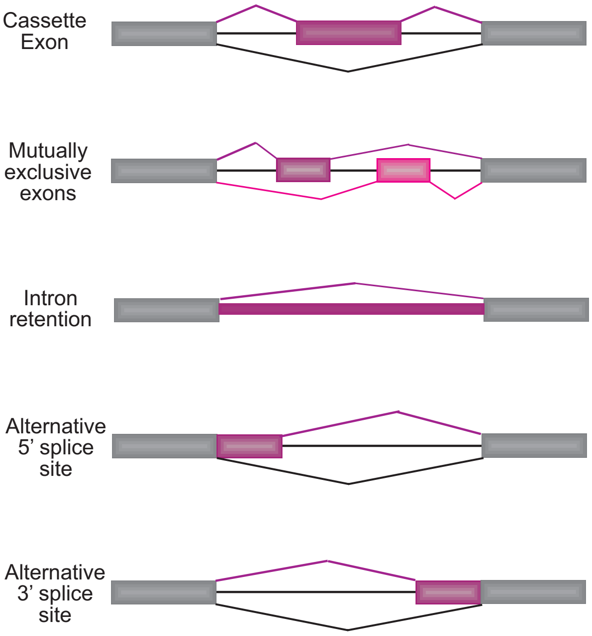


Figure 3.3: **AS types**

and a great challenge for bioinformatics (see review [41]). The first AS events were case-reports found during in depth-study of single genes and their transcripts. Later, large scale studies of AS started with the alignment of ESTs to reference mRNAs [42] or alignment of ESTs to genomic sequence [43]. At that time, those methods estimated the rate of AS in human to be at least 35% and worked quite well for organisms such as human and mouse, that have extensive EST coverage, but rather poorly for other organisms. However, even when EST coverage is quite extensive, many rare alternative splicing events or internal splicing events can still be missed because EST coverage is heavily biased toward the 5' and 3' ends of genes. Another issue is EST sequencing quality, which might produce a number of false positives. Although not perfect, the usage of ESTs allowed to get an overview on splicing, and discover its involvement in tissue and time-specificity [44].

The availability of multiple full genome sequences opened new paths for the discovery of alternative splicing events. EST and genome data of related organisms such as mouse and human can be combined to increase the discovery rate of events [45]. Furthermore more ambitious methods are looking only at genomic alignments to predict alternatively spliced exons [46]. They learn rules from known AS cases, and seek for similar properties in the genome. These rules are the following:

- AS exons have higher sequence identity than constitutive ones;
- the exon length tends to be a multiple of 3 (to not disturb the ORF);

- flanking introns are more conserved at the sequence level.

Even if those rules seem to generalize quite well and could recover half of known AS exons, they also detect false positives at a quite high rate (25%). The same kind of analysis has been conducted in two *Drosophila* species [47] with comparable efficiency. Although not perfect this type of method can allow AS discovery in species for which no ESTs or too few data are available.

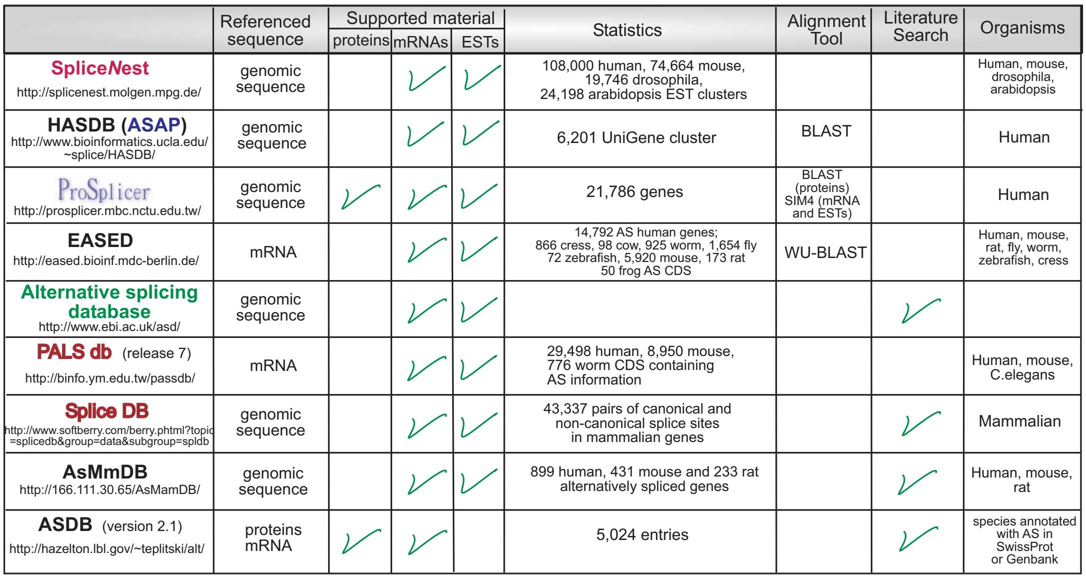
The most recent developments in AS detection involve microarrays. The first microarray designed especially for detecting AS events [48] reported exon skipping only and was conducted on an Agilent array platform covering 10,000 human genes and surveyed 50 human tissues. More than 800 new events have been reported, which were not detectable by ESTs. A tiling microarray study performed on an Affymetrix platform on chromosomes 21 and 22 and profiling RNA samples from 11 tissues allowed to detect AS events also in the middle of the gene, and not mainly on the 5' and 3' ends like the previous array (due to a bias in EST coverage). Another splicing microarray is available, which was designed to identify different types of AS [49]. After a first period of prototype-demonstration stage, AS microarray are becoming a real tool that researchers can use to make discoveries, and will surely offer a new dimension to AS studies.

3.5 Alternative splicing variant databases

Virtually every development of a detection method of AS gave rise to a database of AS variants. AS information is as well present in the annotation of the SwissProt database (under the data field VARSPLIC). The most famous AS databases are summarized in the figure 3.4.

3.6 mRNA surveillance

Proof-reading mechanisms following mRNA splicing assess the position of the protein stop codon within the mature mRNA as a measure of the accuracy of RNA splicing. Usually the protein stop codon is located in the last exon of the mRNA. However, if the mRNA surveillance machinery detects an in frame stop codon upstream of the last exon then mechanisms are initiated to either degrade the transcript or remove the premature stop codon to allow expression of almost full length protein. If the stop codon is absent then the mRNA is also targeted for destruction. Three separate mechanisms have been identified:



	Referenced sequence	Supported material			Statistics	Alignment Tool	Literature Search	Organisms
		proteins	mRNAs	ESTs				
Splice*Nest* http://splicenest.molgen.mpg.de/	genomic sequence		✓	✓	108,000 human, 74,664 mouse, 19,746 drosophila, 24,198 arabidopsis EST clusters			Human, mouse, drosophila, arabidopsis
HASDB (ASAP) http://www.bioinformatics.ucla.edu/~splice/HASDB/	genomic sequence		✓	✓	6,201 UniGene cluster	BLAST		Human
ProSplicer http://prosplicer.mbc.nctu.edu.tw/	genomic sequence	✓	✓	✓	21,786 genes	BLAST (proteins) SIM4 (mRNA and ESTs)		Human
EASED http://eased.bioinf.mdc-berlin.de/	mRNA		✓	✓	14,792 AS human genes; 866 cress, 98 cow, 925 worm, 1,654 fly 72 zebrafish, 5,920 mouse, 173 rat 50 frog AS CDS	WU-BLAST		Human, mouse, rat, fly, worm, zebrafish, cress
Alternative splicing database http://www.ebi.ac.uk/asd/	genomic sequence		✓	✓			✓	
PALS db (release 7) http://binfo.ym.edu.tw/passdb/	mRNA		✓	✓	29,498 human, 8,950 mouse, 776 worm CDS containing AS information			Human, mouse, C.elegans
Splice DB http://www.softberry.com/berry.phtml?topic=splicedb&group=data&subgroup=spldb	genomic sequence		✓	✓	43,337 pairs of canonical and non-canonical splice sites in mammalian genes		✓	Mammalian
AsMmDB http://166.111.30.65/AsMamDB/	genomic sequence		✓	✓	899 human, 431 mouse and 233 rat alternatively spliced genes		✓	Human, mouse, rat
ASDB (version 2.1) http://hazelton.lbl.gov/~teplitski/alt/	proteins mRNA	✓	✓		5,024 entries		✓	species annotated with AS in SwissProt or Genbank

Figure 3.4: **AS databases**

- **Nonsense-mediated decay (NMD)**: Detects the presence of a premature stop codon in the mRNA and initiates mRNA degradation (review in [50]).
- **Nonsense-associated altered splicing (NAS)**: Induces alternative splicing in order to remove the premature stop codon, thereby correcting the error to generate a near full length transcript [51].
- **Non-stop decay**: Detects the absence of an in frame stop codon in a mature mRNA and initiates transcript degradation [52].

3.7 Alternative splicing and evolution

As additional genome sequences become available, comparative genomics provides new insights into alternative splicing: its conservation and role in evolution. Thus it is possible to not only look for overall frequency conservation of AS, but also to go deeper and look at single events, and their conservation. Moreover, hypotheses can now be made concerning the benefits that AS can bring concerning the evolvability of organisms. This has been one of the achievements of this thesis (see Results).

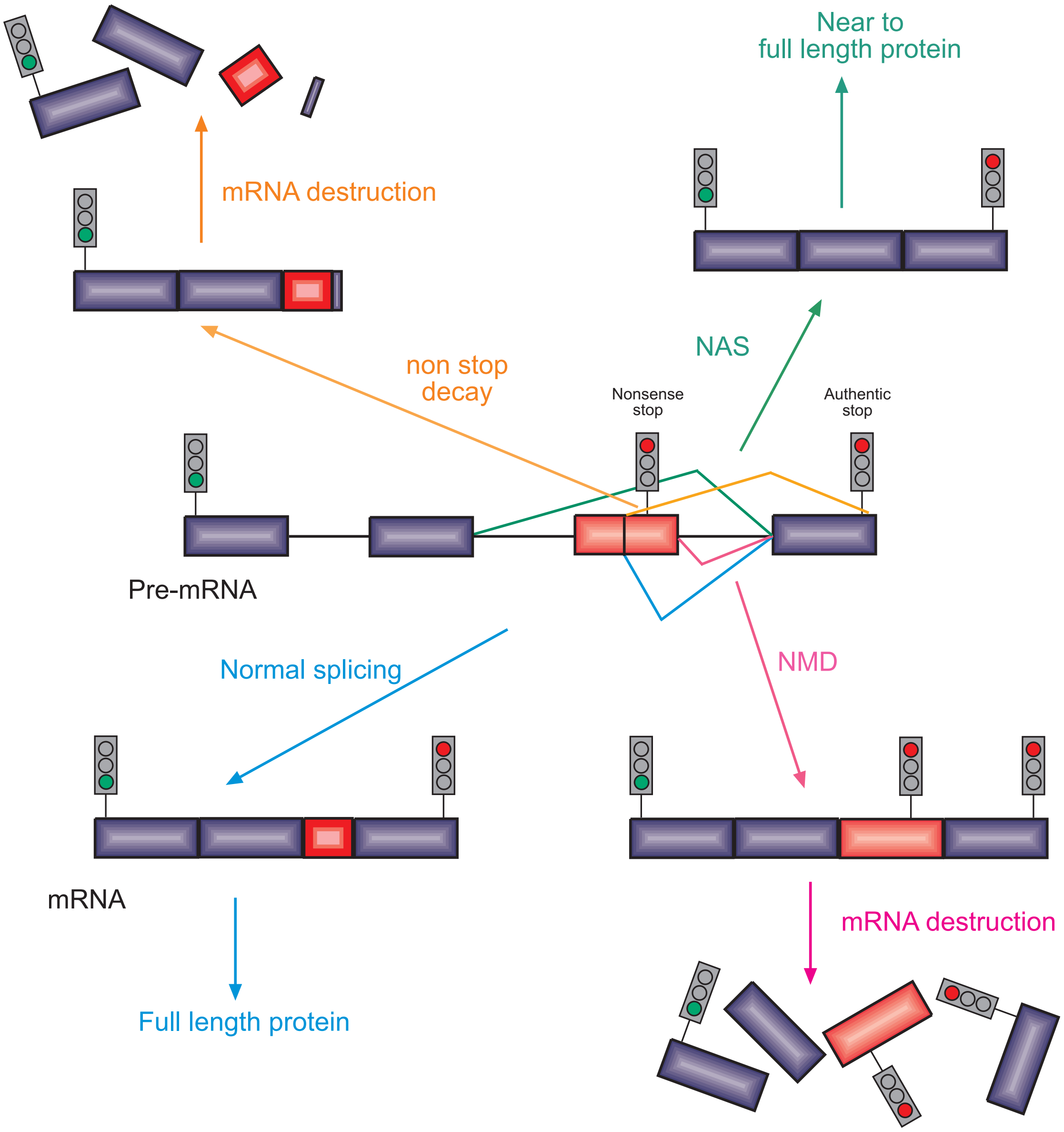


Figure 3.5: **mRNA surveillance.**
The presence of an in frame stop codon upstream of the last exon results in the activation of either nonsense mediated decay (NMD) or nonsense-associated altered splicing (NAS). The deletion of all in frame stop codons results in degradation of the message by non-stop decay mechanisms. Adapted from [51].

Chapter 4

Microarray

4.1 Principle and applications

Large scale analyses are usually undertaken to reveal common expression patterns. To achieve this, many methods have been developed, which usually cluster the "genes" according to their expression profiles and create trees from those related clusters. Methods can usually be classified into two groups: the supervised and unsupervised methods. Broadly used unsupervised clustering methods are K-means or Principal Component Analysis [53].

The assessment of transcript expression levels changes between two different states or tissues has been of interest for many years because they are thought to directly reflect functional differences. They were first studied using northern blot experiments which are labor intensive and not usable as such for larger scale studies. Thus, microarray have been developed, which are able to deal with large amount of reactions. The first report of usage of this technique [54] is now more than 10 years old. Progresses have been fast and numerous since them, allowing the application of this technique to essentially every situation. Microarray were seen as the promise of understanding in details how cells work or where diseases are coming from. However, already in 2001 doubts were emitted concerning the achievability of this goal with microarray [55].

Microarrays can be generalized into two categories (see figure 4.1):

- **cDNA (oligonucleotide) spotted arrays**: cDNA arrays are mostly used in 2-dye (also known as 2-channel) experiments. This means that 2 samples of RNA are taken from 2 separate sources and cDNA is created from each by using a technique called RT-PCR (Reverse Transcriptase Polymerase Chain Reaction). Each cDNA sample is labeled with its own specific dye, and then both are hybridized to a single array slide, after which dye intensities on each spot is calculated using a fluorescence camera. cDNA arrays are typically less dense than Affymetrix-like arrays, containing approximately 20,000 sequences, each of which is about 60 to 80 nucleotides long.

- **High-density synthetic oligonucleotide (Affymetrix-like) arrays**: Affymetrix-like arrays can only be treated with a single sample (i.e. RNA from a single source). No RT-PCR is necessary. These arrays can potentially have a density of up to 100,000 sequences, each of which being no longer than 40 nucleotides.

4.2 Experimental procedure

A typical microarray experiment involves 6 steps:

1. Experimental design;
2. Prepare the DNA chip with the chosen target DNA;
3. Generate a hybridization solution containing a mixture of fluorescently labeled cDNAs;
4. Incubate the hybridization mixture containing fluorescently labeled cDNAs with the DNA chip;
5. Detect bound cDNA using laser technology and store data in a computer;
6. Analyze data using computational methods.

4.2.1 Experimental design

The type of experimental design depends on the biological question that needs to be answered thanks to the microarray experiment. A careful design is particularly important if the resulting experiment is to be maximally informative, given the effort and the resources (for review, see [56]).
A key choice in microarray design is whether to use direct or indirect comparisons; that is, whether to make comparisons within or between slides. Generally, cDNA arrays comparisons are best made directly, i.e. the two conditions being compared should be put on the same array, one with the green dye and the other with the red dye. However, for some special cases, like a big number of experimental samples, comparison to a reference sample (constant sample) can be used. Moreover, to increase the statistical power of the analysis, many repeats (biological as well as

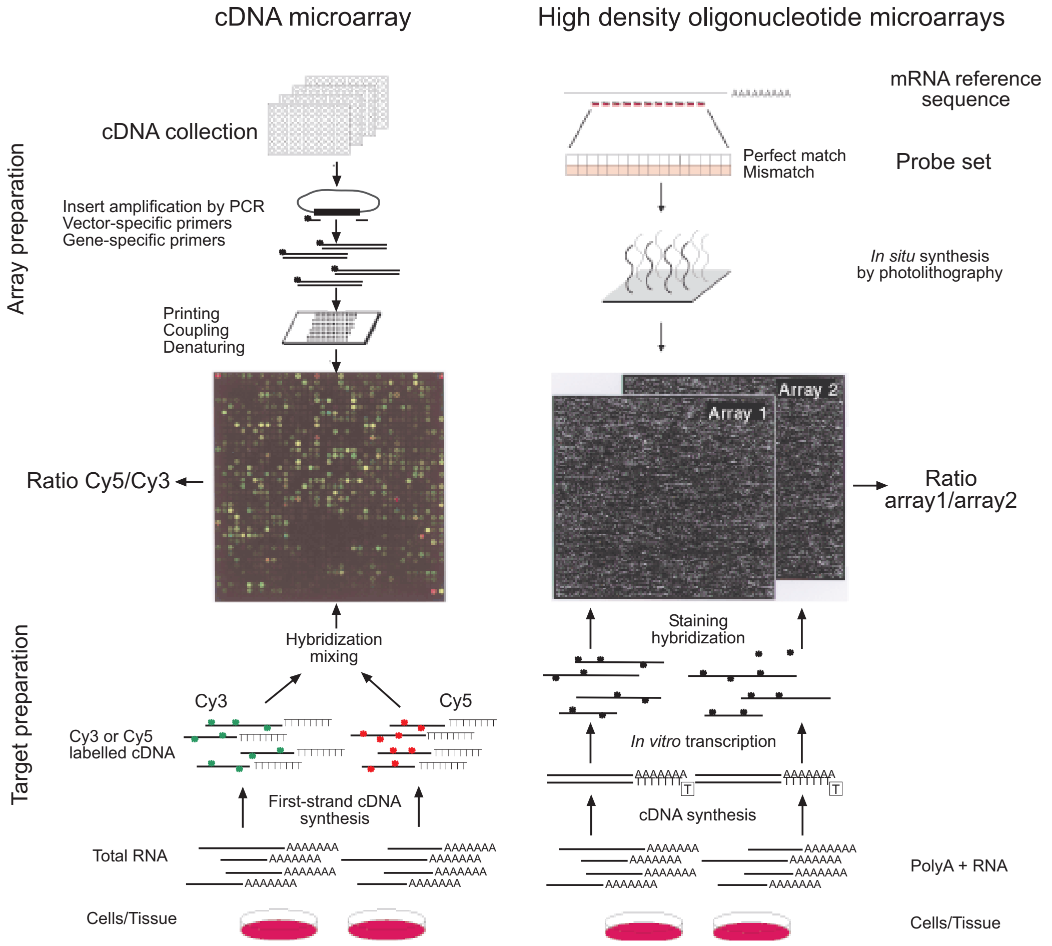


Figure 4.1: **Schematic overview of probe array and target preparation for spotted cDNA microarrays and high-density oligonucleotide microarrays.**
Adapated from [55].

technical) should be done, for example the use of different dyes (dye-swap experiment) in order to rule out specific dye-bias.

A spotted array would ideally have these basic sequences:

- **Genes of interest**: around 25,000 gene-complementary sequences on cDNA array or more than 100,000 on high-density oligonucleotide arrays.
- **Controls**
  - Negative (including non-organism-of-interest sequences);
  - Positive (including housekeeping genes).

4.2.2 MIAME: Minimum Information About a Microarray Experiment

As the number of microarray experiment data produced and to be produced is enormous, a need for standardization occurred, and a collaboration between the leaders in the microarray field issued a framework for describing an array experiment: MIAME [57] (Minimum Information About a Microarray Experiment) outlines the content, not the format, of annotation for an array experiment. Using this, data are stored in the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/).

MIAME includes 6 basic components to describe a microarray experiment:

1. **Experimental design**: This section contains general information about the experiments as a whole, including the biological question being explored, replicates, information about the authors, etc.

2. **Array design**: This section contains all information concerning physical attributes of the arrays themselves.

3. **Samples**: This section contains all information concerning the origin of the RNA that is hybridized to the array surface, including any chemical parameters, e.g. RNA preparation, etc.

4. **Hybridizations**: Contains any information pertaining to the physical parameters of hybridization.

5. **Measurements**: Contains experimental results (raw data). Includes gene expression matrix.

6. **Normalization controls**: Contains all information pertaining to normalization procedures (including which sequences on the array are used for such procedures) so that arrays can be made comparable.

4.3 Data interpretation

The data of a microarray experiment typically constitutes a long list of spot intensities and intensity ratios, generated either by pairwise comparison or by comparing different samples to a common control. The main challenge consists then in extracting the biologically meaningful information from this load of data, thereby recognizing and leaving out the technological noise. Replication of the experimental setup and statistical analysis of the data have proven to be essential for extracting more significant information from the data.

4.3.1 Data normalization

Every repeat experiment will give rise to a certain amount of variation. These changes are termed systematic and random variations and together make up the experimental error inherent in the procedure. A statistical method must be applied to minimize these variations, which in turn allows one to compare the expression levels between multiple microarray experiments. Normalization procedures rely on the fact that gene expression data can follow a normal distribution and therefore the entire distribution can be transformed about the population mean and median without affecting the standard deviation (i.e. the variation of the data). For cDNA arrays, the array itself is already hybridized with two types of mRNA or cDNA, each of which has been labeled with a different fluorescent dye. Most often these are Cy3 and Cy5. A within-slide normalization procedure is therefore usually applied to account for the systematic variations in measured dye intensity due to hybridization preferences. In contrast, Affymetrix chips need cross-slide (or between-slide) normalization procedures applied to them since one chip will only ever be treated with one RNA sample. Cross-slide normalization will be routinely applied to cDNA arrays because a single experiment with one array and 2 RNA samples will not yield enough information for a statistically valid analysis.

There are several simple types of normalization that can be applied in this case (see figure 4.2):

- Mean or median normalization: corrects the data such that all arrays have the same mean or median.
- Scaling normalization: the distributions about the median for each array are made to be equivalent.
- Rank distribution normalization: makes the entire distribution of every array identical.

4.3.2 Clustering and classification methods

Clustering involves grouping data points together according to some measure of similarity. One goal of clustering is to extract trends and information from raw data sets. There are two general types of clustering: supervised and unsupervised clustering. Supervised clustering uses a set of example data to classify the rest of the data set. It is hence really much dependent on a clever choice of example data. Unsupervised clustering, on the other hand, tries to discover the natural groupings inside a data set without any input from a trainer. The main input a

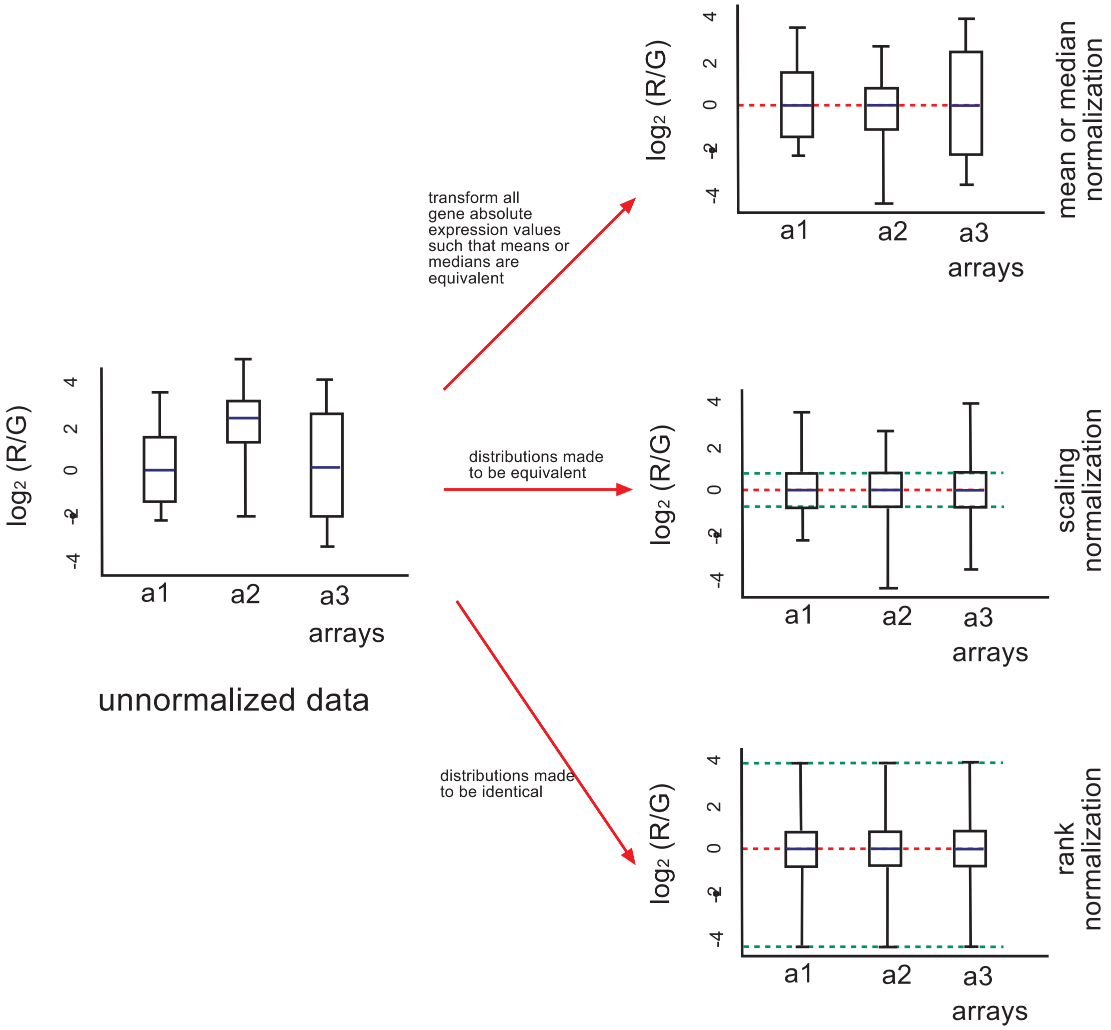

log2 (R/G)
4
2
0
-2
-4
a1
a2
a3
arrays
unnormalized data
transform all gene absolute expression values such that means or medians are equivalent
distributions made to be equivalent
distributions made to be identical
mean or median normalization
scaling normalization
rank normalization

Figure 4.2: **Schematic overview of standard microarray data normalization procedures**

typical unsupervised clustering algorithm takes is the number of classes it should find. One of the most important characteristics of any supervised or unsupervised clustering process is how to measure the similarity of two data points. Generally, unsupervised methods perform the job of clustering while supervised methods are more suited to classification of datasets.

The purpose of any clustering method is to group entities on the basis of similarity of features. In the case of gene expression data derived from multiple microarray experiments, genes can be clustered on the basis of similar expression profiles. What this means is that if one were to perform many microarray experiments and record each gene expression result, one could group together genes that are expressed together most frequently.

4.3.2.1 Hierarchical clustering

Developed in 1998 [58], it is nowadays the most widely used clustering method to arrange genes according to similarity in pattern of gene expression. This similarity can be expressed in many ways: Euclidean distance, angle or dot-products of the two n-dimensional vectors representing a series of n measurements. The result of the clustering is then represented by a tree, very much like the ones used to show sequence similarities between genes or proteins.

There are two types of hierarchical clustering - agglomerative and divisive. Agglomerative clustering takes each entity (i.e. gene) as a single cluster to start off with and then builds bigger and bigger clusters by grouping similar entities together until the entire dataset is encapsulated into one final cluster. Divisive hierarchical clustering works the opposite way around - the entire dataset is first considered to be one cluster and is then broken down into smaller and smaller subsets until each subset consists of only a single entity.

Linkage is the criterion by which the clustering algorithm determines the actual distance between two clusters by defining single points that are associated with the clusters in question. Hierarchical clustering can be implemented in three main ways:

- **Single linkage**: Defines the distance between any two clusters as the minimum distance between them. It tends to force clusters together due to single entities being close to each other regardless of the positions of other entities in that cluster (chaining phenomenon).
- **Complete linkage**: Defines the distance between any two clusters as the maximum distance between them. Outliers are given more weight in the cluster decision.
- **Average linkage or UPGMA (Unweighted Pair-Group Method using Arithmetic averages)**: Takes the mean distance between all possible pairs of entities of the two clusters in question. More computationally intensive, but also avoids the problem of chaining, and does not give too much weight to outliers, so it is the most popular method.

The agglomerative hierarchical clustering algorithm proceeds in 3 steps:

1. Derive vector representations for each entity (i.e. gene expression values for each experiment make up the vector elements for a specific gene)
2. Compare every entity with all other entities by calculating a distance. Input that distance into a matrix. Calculation of the distance depends on (i) the

linkage method being implemented; (ii) the method of calculation of actual distances (euclidean, pearson correlation coefficient...)

3. Group the closest two entities (or clusters) together (which makes a new cluster) and go back to step 2, counting the new cluster as a single entity, until all entities are contained within one big cluster.

4.3.2.2 K-means method

The k-Means method is known as a partitional method since the user must first predefine the number of clusters (and actually identify their centers) after which the algorithm partitions the data iteratively until a solution is found. These are the basic steps to follow given a raw dataset (i.e. microarray data):

1. Initialization
   - Define the number of clusters (k).
   - Designate a cluster center (a vector quantity that is of the same dimensionality of the data) for each cluster.
2. Assign each data point to the closest cluster center. That data point is now a member of that cluster.
3. Calculate the new cluster center (the geometric average of all the members of a certain cluster).
4. Calculate the sum of within-cluster sum-of-squares. If this value has not significantly changed over a certain number of iterations, exit the algorithm. If it has, or the change is insignificant but has not been seen to persist over a certain number of iterations, go back to Step 2.

4.3.2.3 Principal component analysis or PCA

PCA is commonly used in microarray research as a cluster analysis tool [59]. It is designed to capture the variance in a dataset in terms of principle components (PC). Actually, it tries to reduce the dimensionality of the data to summarize the most important (i.e. defining) parts while simultaneously filtering out noise. PCA can be imposed on datasets to capture the cluster structure (just using the first few PC's) prior to cluster analysis (e.g. before performing k-Means clustering to determine a good value for K).

4.3.2.4 ANalysis Of VAriance between groups or ANOVA

Basically, we seek a statistic that quantifies the signal-to-noise ratio. The variance between groups is thought of as a signal of group differences. The variance within groups is thought of as background noise. When the variance between groups (signal) is much larger than the variance within groups (background noise), we conclude that the evidence points to real group differences (see figure 4.3).

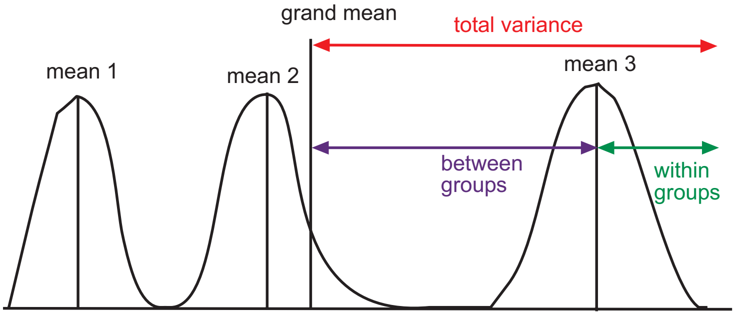


Figure 4.3: **ANOVA principle**

The analysis of variance, known as ANOVA, is used to test hypotheses about differences between two or more means. The question being asked to an ANOVA test is whether the expression level for one gene (taken one at a time) changes significantly as a function of time (in our case). To do this, it compares the variability within replicates for a given time point and a given probe to the variability caused by different time points. Thus, the null and alternative hypotheses are:

$$H0 : \mu_1 = \mu_2 = ... = \mu_k$$

H1: H0 is false (at least one population mean differs) where μ_i represents the population mean of group i.

The ANOVA computes a p-value, which is basically the probability to tell that there is a significant difference between groups, or probes although there is not.

Chapter 5

Drosophila development

5.1 Drosophila as a model organism

The fruit fly *Drosophila melanogaster* has the longest history of any model organism and has been widely used to study genetics and developmental biology. The fruit fly is a small insect that feeds and breeds on spoiled fruit. It has been used as a model organism for over 100 years and thousands of scientists around the world work on it. Part of the reason for this is historical. Scientists today choose to study the fruit fly because so many others have done so before them, which implies that there are established methods for handling flies in the laboratory and an immense volume of data has accumulated about fly biology.

As with most of the long-established model organisms, the initial choice was for practical reasons. The fruit fly is small and has a simple diet. Therefore, large numbers of flies can be maintained inexpensively in the laboratory. The life cycle is also very short, taking about two weeks, so large-scale crosses can be set up and followed through several generations in a matter of months. Due to these advantages, fruit flies were extensively used in the early 20th century to work out the principles of genetics. Mutants are available for a large number of genes and new mutations can be induced very easily by exposing flies to radiation or adding mutagenic chemicals to their food. This ability to recover mutants means that flies can be used to investigate the genetic basis of any conceivable biological process. The fly genome, which was sequenced in the year 2000 [60], contains approximately 14 000 genes.

Figure 5.1: **Fruit fly: *Drosophila melanogaster***

5.2 Embryonic development

The *Drosophila* life cycle consists of a number of stages: embryogenesis, three larval stages, a pupal stage, and the adult stage. We will concentrate here on the embryonic development, for which the 17 stages as defined by Volker Hartenstein and José Campos-Ortega are depicted in figures 5.2 and 5.3.

Following fertilization, mitosis (nuclear division) begins. However, cytokinesis (division of the cytoplasm) does not occur in the early *Drosophila* embryo, resulting in a multinucleate cell called a syncytium, or syncytial blastoderm. The shared cytoplasm allows morphogen gradients to play a key role in pattern formation. At the tenth nuclear division, the nuclei migrate to the periphery of the embryo, and the germ line is formed from about 10 pole cells set off at the posterior end; the pole cells undergo two additional divisions and are reincorporated within the embryo by invagination. At the thirteenth division, the 6,000 or so nuclei are partitioned into separate cells during an event called cellularization, which gives rise to the blastoderm. Although not yet evident, the major body axes and segment boundaries are determined. Subsequent development results in an embryo with morphologically distinct segments.

The genes that control *Drosophila* embryonic development can be divided into three classes:

1. **Maternal-effect genes** that specify egg polarity and the spatial coordinates of the egg and future embryo;

2. **Segmentation genes**, including the gap, pair-rule, and segment polarity classes of genes, that determine the number and polarity of the body segments;

3. **Homeotic genes** that determine the identification and sequence of the segments.

5.3 The Notch pathway

The building of an organism from a single cell to a complex set of organs requires coordination of genes to direct the fate of individual cells. Two principal mechanisms progressively restrict the cell fate, or developmental outcome, of cells within a lineage:

- Developmental restriction may be autonomous: it is determined by genetically programmed changes in the cells themselves;

- Cells may respond to positional information: developmental restrictions are imposed by the position of cells within the embryo, and more precisely they are mediated by signaling interactions between neighboring cells or by gradients in concentration of particular molecules.

Moreover, development seems to make reiterative use of a small set of essential molecular signals: Wingless, Hedgehog, transforming growth factor beta, receptor tyrosine kinase/phosphatase and Notch pathways. As a matter of fact, the Notch pathway which is responsible for this essential coordination during embryogenesis appears to be ubiquitous in virtually all cell-cell contacts in all metazoan animals, and is best known and most throughout studied in *Drosophila melanogaster*. Signals exchanged between neighboring cells through the Notch receptor can amplify and consolidate molecular differences, which eventually dictate cell fates. In *Drosophila* the Notch signaling pathway regulates, with both positive and negative signals, the differentiation of at least central and peripheral nervous system and eye, wing disc, oogenesis, segmental appendages such as antennae and legs, and muscles, through lateral inhibition or induction.

The gene encoding the Notch receptor was discovered in mutant flies with partial loss of function of Notch resulting in notches at the wing margin [61]. The notch gene encodes a 300kD single-pass transmembrane receptor. The large extracellular domain contains EGF repeats. A PEST sequence is found within the intracellular domain. Genetic and molecular interaction studies provided a number of proteins involved in the transmission or the regulation of the notch signal (see review [62]).

In general, the pathway, depicted in figure 5.4, works as follows: Notch is at the cell surface, where it is activated by its ligands Serrate or Delta from the neighboring cell. Then, the receptor is cleaved twice by a proteolytic mechanism in which Presenilin plays an important role, and the intracellular domain (NICD) is transferred to the nucleus, where it, together with the Suppressor of Hairless protein, constitutes a transcription factor which activates the Notch target genes, mainly located in the Enhancer of Split complex, and which encode repressors. In the absence of Notch, CSL(CBF1, Su(H), Lag1) acts as a transcriptional repressor. CSL binds to at least two corepressor complexes: the SMRT/NcoR/histone deacetylase 1 (HDAC1) and CIR/HDAC2/SAP30 complexes.

Although the core mechanism of the Notch pathway starts to be well understood, it also becomes clear that a lot of cross-talk is going on between the Notch and other developmental pathways, and interaction with specific proteins modulate the pathway depending on the cell type or the time.

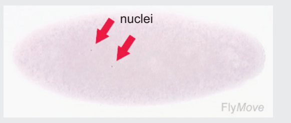
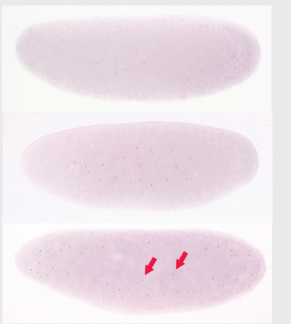
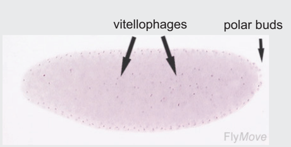
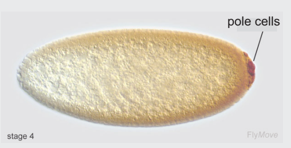
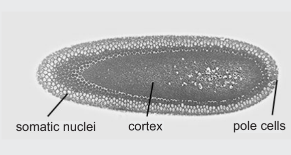
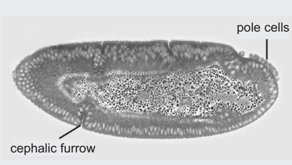
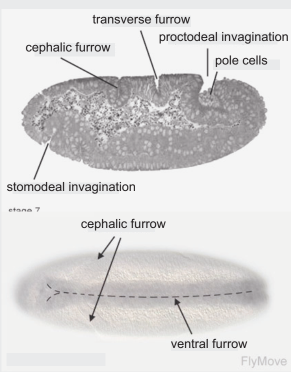
stage	time	pictures	Key event
1	0 - 0:25 h		First two nuclear divisions
2	0:25 - 1:05 h		Syncytial divisions 3-8 Cytoplasmic clearing
3	1:05 - 1:20 h		Syncytial division 9 Polar bud formation
4	1:20 - 2:10 h		Syncytial divisions 10-13 Pole cell formation
5	2:10 - 2:50 h		Cellularization
6	2:50 - 3:00 h		Onset of gastrulation, formation of ventral and cephalic furrow, dorsal shift of pole cells
7	3:00 - 3:10 h	 transverse furrow cephalic furrow proctodeal invagination pole cells stomodeal invagination cephalic furrow ventral furrow FlyMove	Completion of gastrulation Pole cells in a pocket
8	3:10 - 3:40 h		Rapid phase of germ band extension, to 60% egg length

Figure 5.2: ***Drosophila melanogaster* embryonic stages 1-8**.
Time is defined for a development at 25°C; stages are as defined by Volker Hartenstein and José Campos-Ortega

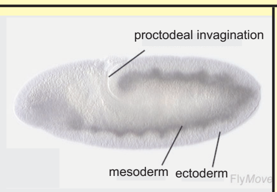
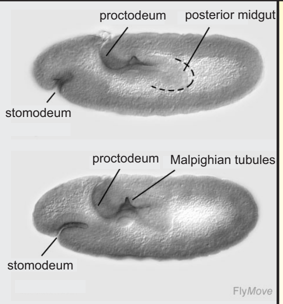
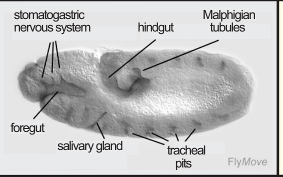
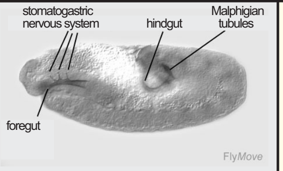
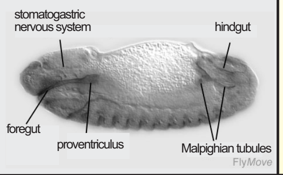
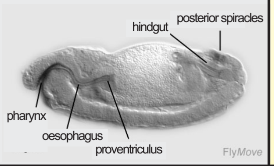
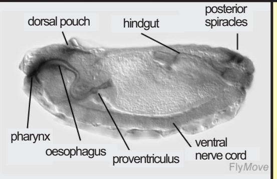
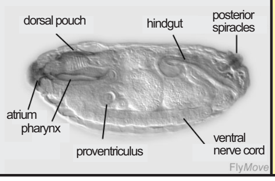
stage	time	pictures	Key event
9	3:40 - 4:20 h		Germ band elongation to 70% egg length; early neuroblast delamination
10	4:20 - 5:20 h	 proctodeum posterior midgut stomodeum proctodeum Malpighian tubules stomodeum FlyMove	Germ band elongates to 75% egg length; stomodeum invaginates
11	5:20 - 7:20 h	 stomatogastric nervous system hindgut Malphigian tubules foregut salivary gland tracheal pits FlyMove	Segmentation, tracheal pits arise, posterior midgut invagination reaches the posterior pole
12	7:20 - 9:20 h	 stomatogastric nervous system hindgut Malphigian tubules foregut FlyMove	Onset of germ band retraction, fusion of anterior and posterior midgut
13	9:20 - 10:20 h	 stomatogastric nervous system hindgut foregut proventriculus Malpighian tubules FlyMove	
14	10:20 - 11:20 h	 hindgut posterior spiracles pharynx oesophagus proventriculus FlyMove	Dorsal closure to 80% along the dorsoventral axis, head involution
15	11:20 - 13:00 h	 dorsal pouch hindgut posterior spiracles pharynx oesophagus proventriculus ventral nerve cord FlyMove	Dorsal closure of epidermis and midgut
16	13:00 - 16:00 h		Gastric caecae are formed, somatic musculature becomes visible
17	16:00 - hatching	 dorsal pouch hindgut posterior spiracles atrium pharynx proventriculus ventral nerve cord FlyMove	Completion of organogenesis, movements of the embryo within the vitelline envelope

Figure 5.3: ***Drosophila melanogaster* embryonic stages 9-17**
Time is defined for a development at 25°C as defined by Volker Hartenstein and José Campos-Ortega

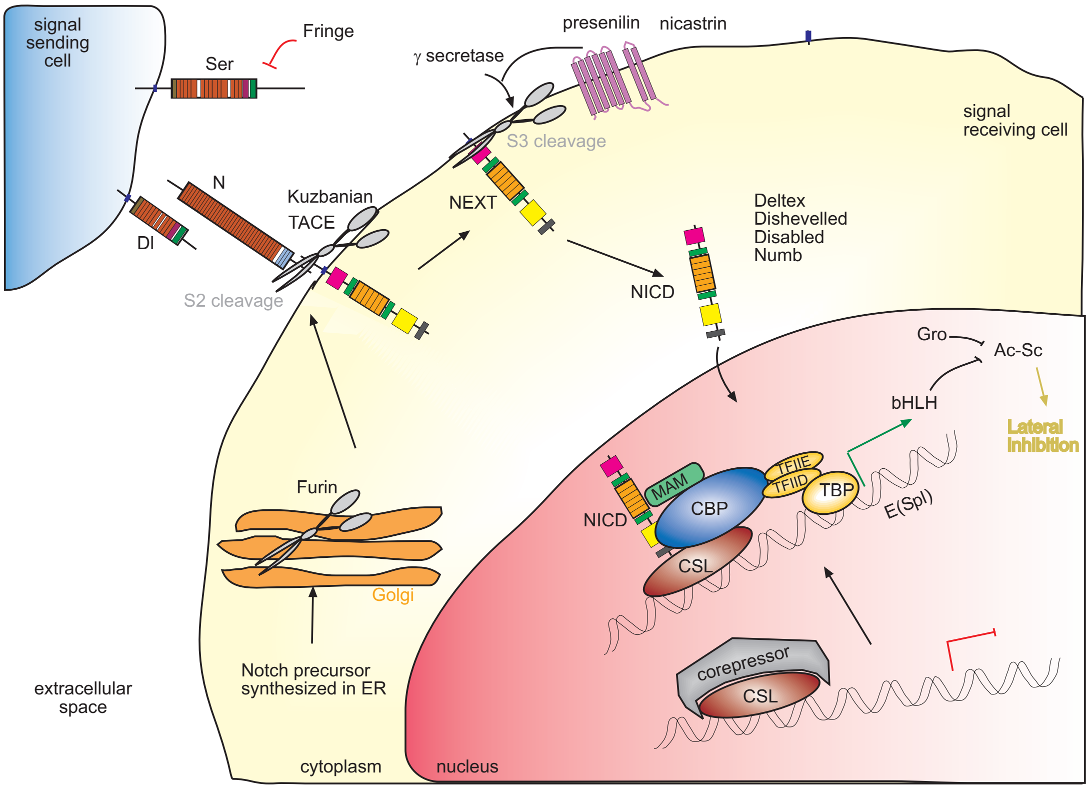


Figure 5.4: **Simplified representation of the Notch pathway**
Protein domains defined by SMART are color-coded as follow: red= EGF repeat, green=MNNL(N terminus of Notch ligand), violet=DSL(Delta Serrate ligand), blue=transmembrane, brown=VWC(von Willebrand factor type C domain), light blue=LIN, orange=ankyrin repeats, dark grey=PEST sequence, light green=nuclear localization signal, pink=RAM domain, yellow=transactivation domain.
The gene symbols used are the following: Ser=Serrate Dl=delta N=Notch

Chapter 6

Aim and achievements of the thesis - Publication list

The aim of this thesis done at the European Molecular Biology Laboratory (EMBL) under the co-direction of Dr Peer Bork and Dr James Stévenin was to study the transcriptome of different organisms. This study has taken two directions: first, I studied the repercussion of alternative splicing on the transcriptome. Secondly I studied transcriptomes in multiple dimensions, namely in time, during *Drosophila melanogaster* development and in space, in many tissues of *Mus musculus* and *Homo sapiens*.

The most straightforward computational approach for studying the transcriptome of an organism is to look at the coding potential of its genes. However, the dogma ”one gene gives rise to one protein” is no longer valid, and the diversity brought at different levels between transcription and post-translation, which has to be taken into account, is still poorly predictable from the sole gene sequence. That is the reason why a lot of attention has been focused on using large scale experiments, namely EST sequencing and microarray mainly, to get a snapshot of the transcriptome at a given time, and in a given tissue or organism. The dimensions available depend on the experiment setting: it is possible to assess the transcript diversity generated from a single gene if the focus is alternative splicing, or the expression of a gene (generally not at the single transcript level though) in a particular condition.

Specialized microarrays for the study of alternatively spliced variants started to appear at the beginning of my thesis, but started to become reliable only recently, and although the possibilities brought by this technique are without doubt considerable, this was too early days to be the focus of my thesis. I instead concentrated on assessing further alternative splicing relevance both in terms of frequency and evolution. This work summarized in chapter 7 gave rise to two publications:

- One on alternative splicing and evolution:
  Boue, S., Letunic, I. and Bork, P. (2003) ”Alternative splicing and evolution” *BioEssays* **25**(11)1031-1034

- One on the estimation of AS frequency in different species:
  Harrington, E., **Boue, S.**, Valcarcel, J., Reich, J., Bork, P. (2004) ”Estimating rates of alternative splicing in mammals and invertebrates” *Nature genetics* **36**(9)916-917

As most of AS databases are computationally generated, and usually lack any experimental validation, it is difficult to estimate how trustworthy they can be. That's why the creation of a benchmark database regrouping only experimentally validated AS events is of great interest, and will allow to compare the performances of different prediction methods. Such a database has been created thanks to a literature mining agent based on support vector machine and trained to extract AS events from existing literature. The tool developed to create this database is presented in:

- Shah, P., Jensen, L.J., **Boué,S.**, Bork, P. (2005) ”Extraction of transcript diversity from scientific literature” *PLoS computational biology* **1**e10.

Large scale production of experimental measurements of gene expression of thousands of genes as well as the rapid generation of full genome sequences, created in the scientific community the hope that very soon cell physiology and pathology would be understood in the littlest details. However, the gap between the production of raw data and a good understanding of metabolic or genetic pathways has certainly been underestimated. There are indeed numerous checkpoints and regulatory mechanisms between the DNA strand as it is sequenced, and the function of the protein it encodes at specific time and space points. In this respect, mRNA transcript, although not perfect because not definitive, has already been subjected to numerous checkpoints and modifications, and can give already a precise overview on what is going on in different cells. This is moreover facilitated by a solid experimental procedure history. We realized two kinds of studies to get a snapshot at gene expression properties in space on the one hand and in time on the other hand.

The study in space was based on microarray gene expression data for human and mouse in more than 50 different tissues or cell types. Expression profiles of 16,400 mouse gene in 47 tissues permitted to construct a ”phylogenetic tree” of tissues, which reveals functional relationships, but also shows that numerous genes are neutrally expressed, i.e. they are expressed innocuously also in tissues where they are not believed to have a function. Tests showed that this neutral expression most probably is the result of expression leakage, which could be explained by the increased accessibility to a gene caused by the chromatin remodeling exerted for the expression of a neighbor gene, which conducts a function in this tissue.

- Yanai, I.*, Korbel, J.O.*, **Boue, S.***, McWeeney, S.K., Bork, P. and Lercher, M. (2006) "Similar gene expression profiles do not imply similar tissue functions" *Trends in genetics* **22**(3)132-138

  * These authors contributed equally to this work

The survey of transcripts in time consists in the analysis of series of genome-wide microarray measurements during *Drosophila melanogaster* development from fertilization to hatching. The experiments have been conducted by collaborators in the Yale medical school. The originality of our study is the application of a statistical method particularly adapted to time-series microarrays, which takes into account the succession in time of the measurements of expression, and does not consider the data points as a number of independent measures of the same gene expression, as would the classical ANOVA do. This method allowed us to classify "significant profiles" (profiles where a significant variation is observed) into three categories, depending on the observation of consistent increase or decline of expression (or both) during the studied embryogenesis. If only a decline of expression is observed, the genes are called "maternal"; if an activation followed by a decline are observed, the genes are named "transient" and if only a increase is observed "activated". This classification highlighted the over representation of known developmental pathways in the "transient" group, mainly members of the Notch pathway, which hints at the presence of new members of this pathway within this category. This work has been submitted, but referees are asking for experimental proofs of co-expression of the "incriminated" genes by *in situ* hybridization, which will delay the publication of this work.

- Hooper, S.*, **Boue, S.***, Krause, R.*, Jensen, L.J., Mason, C., Ghanim, M., White, K. P., Furlong, E. E. and Bork, P. "Identification of tightly regulated groups of transiently expressed genes during *Drosophila melanogaster* embryogenesis." submitted.

  * These authors contributed equally to this work

It is hence already possible to obtain biologically interesting results from large scale transcriptomics studies. However because of the constant discovery of new regulatory mechanisms and despite technical difficulties, it becomes clear that large scale experiments will have to investigate proteomes in order to get a real functional overview of the molecular biology of the cell.

Additional publication:

A collaboration with an experimentalist group within EMBL led to the following publication, which will not be discussed in this thesis.

- Ulbert, S., Platani, M., **Boue, S.** and Mattaj, I. (2006) "Direct transmembrane protein-DNA interactions required early in nuclear envelope assembly" *Journal of cell biology* In press.

As most of the results of this work have been published, in the results part, I will present the publications, highlight the motivation, and sum up the results.

Part II

Methods and results

Chapter 7

Alternative splicing

7.1 Evaluate the rate of AS in different species

As mentioned in the introduction, after realizing that the complexity of organisms is not accounted for by the gene content of a genome, other mechanisms of complexity generation have been sought. In this quest, alternative splicing revealed itself as the holy grail. It turned out to not be an exceptional event at all, on the contrary, and can generate a multitude of transcripts from a single gene, the most amazing example being the Dscam gene in *Drosophila melanogaster* able to generate more than 44,000 transcripts (see review [63]). However importance of AS has to be evaluated in different organisms, also in "lower" eukaryotes in order to estimate its relative functional importance in "higher" organisms compared to lower ones.

Withal in order to make a meaningful comparison, one has to use the same detection method in every species, and should not infer events in one specie based on another one under the assumption that AS events are conserved, as it would imply a circular reasoning. The method of choice to compare AS frequency among species is (was) the alignment of ESTs to a reference sequence, mRNA or genome. Applying this methodology, Brett *et al.* [64] showed in 2002 that, contrary to common expectations, AS rates (both in terms of percentage of genes subjected to AS, and to the number of AS forms per gene) are similar in 7 different species (human, mouse, rat, cow, fly, worm and plant, see figure 7.1) and that observed differences are due to different coverage of EST databases, in terms of raw number of ESTs, but also of the variety of tissues and conditions sampled.

These results have been challenged in 2004 by Kim *et al.* [65], who used a different method to prevent this EST-bias in the estimation of AS rate. They based their approach on a modified version of the one that Ewing and Green used to estimate the total human gene count [66]. Thereby they estimated the extent of alternative splicing in *Caenorhabditis elegans, Drosophila melanogaster, Mus musculus* and *Homo sapiens.* These data, which they showed are not influenced by EST coverage, indicated that mice and humans have a higher rate of alternative splicing than do fruit flies and nematodes.

As the original paper [64] has been produced by researchers from our group, we were given the opportunity to analyze the Kim *et al.* analysis, and to answer their comments [67]. The full text version of both their comment and our answer are reproduced in the next pages. What is important to retain from our own analysis is summarized in the following points:

- Kim *et al.* data agree with our previous results before the normalization for EST redundancy: the more ESTs, the higher the AS rate estimation.
- Their study is not flawless: part of it is impossible to reproduce, and their result get biased both by the length of the contigs used, and the EST coverage, as demonstrated by the calculation of AS rate with their method (as far as we could reproduce it) in mouse and rat. Rat has a much lower EST coverage and has also a much lower AS rate estimation. However it is expected that rat and mouse have similar behaviors concerning AS patterns, which is far from being the case using this method: rat seems closer to invertebrates than to mouse, which seems doubtful (see figure 1 of the answer).
- The usage of EST data for estimating AS rate has limitations due to EST coverage and sampling differences among organisms, and should be seen as an indication, but by no means as a definite result. It is only the best bioinformatics could do up to now, and one will probably have to wait for technological advances (maybe the AS microarrays) to touch the end of the suspense, and how big is the role of AS for producing the functional complexity observed in animals.

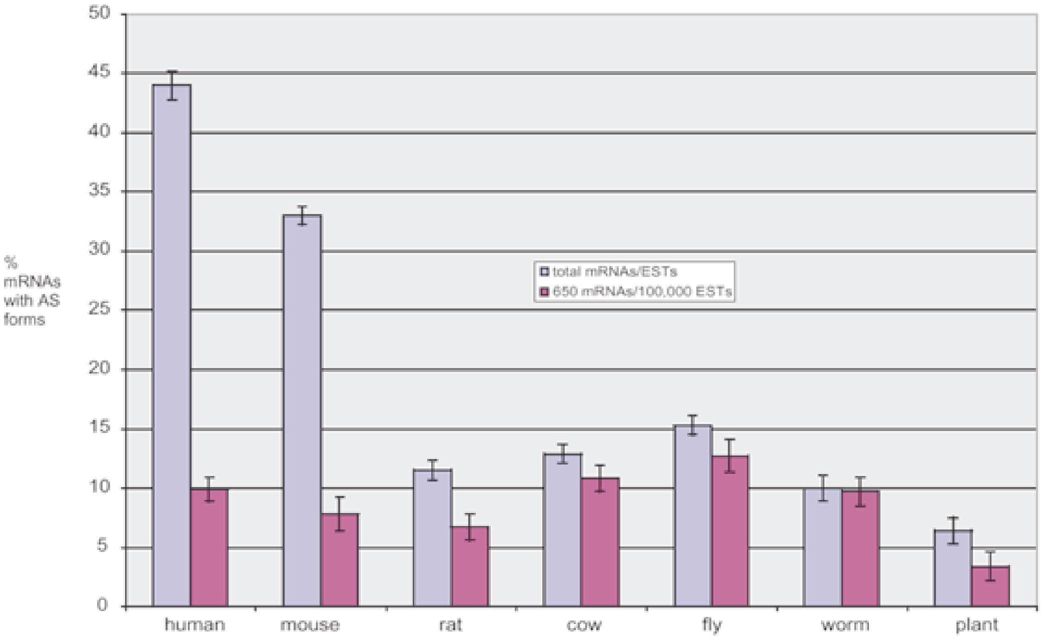


Figure 7.1: Original figure 1 from [64] **EST estimations of alternative splicing from different eukaryotes.**
To identify alternative splicing (AS), we identified high-scoring ESTs with more than 98% identity over 100 bps to mRNA sequences (mRNAs), using the BLASTN search tool with slightly modified parameters. We recorded deleted or inserted sequences in each matching EST after filtering internal repeats. The bar to the left (light purple) for each organism shows the estimate of alternative splicing based on all ESTs and the total number of published mRNA sequences and ESTs for the species: Homo sapiens, 23,161 mRNAs and 3.1 million ESTs; Mus musculus, 9,682 mRNAs and 1.9 million ESTs; Rattus norvegicus, 5,803 mRNAs and 263,362 ESTs; Drosophila melanogaster, 2,973 mRNAs and 115,191 ESTs; Bos taurus, 1,370 mRNAs and 159,130 ESTs; Caenorhabditis elegans, 18,821 mRNAs and 108,115 ESTs; Arabidopsis thaliana, 3,084 mRNAs and 112,112 ESTs. As the rate of detectable alternative splicing depends on EST coverage, we created comparable subsets for each organism, comprising a random set of 650 mRNA sequences with a coverage of 100,000 ESTs (to allow inclusion of cow with the smallest number of public mRNA sequences and ESTs of the seven organisms). Only mRNAs with 3-20 EST matches were included in the set, to account for biases in EST coverage. The darker red bar to the right for each organism is the outcome of this subset. Error bars were created with normal binomial s.d. We carried out random re-sampling with smaller subsets of ESTs (100,000) tested against the total mRNA sets. We observed a similar distribution and scale of error bars.

Estimating rates of alternative splicing in mammals and invertebrates

Heebal Kim, Robert Klein, Jacek Majewski & Jurg Ott

Nature Genetics, *2004, volume 36, N° 9, Pages 915-917*

Pages 915 à 917 :

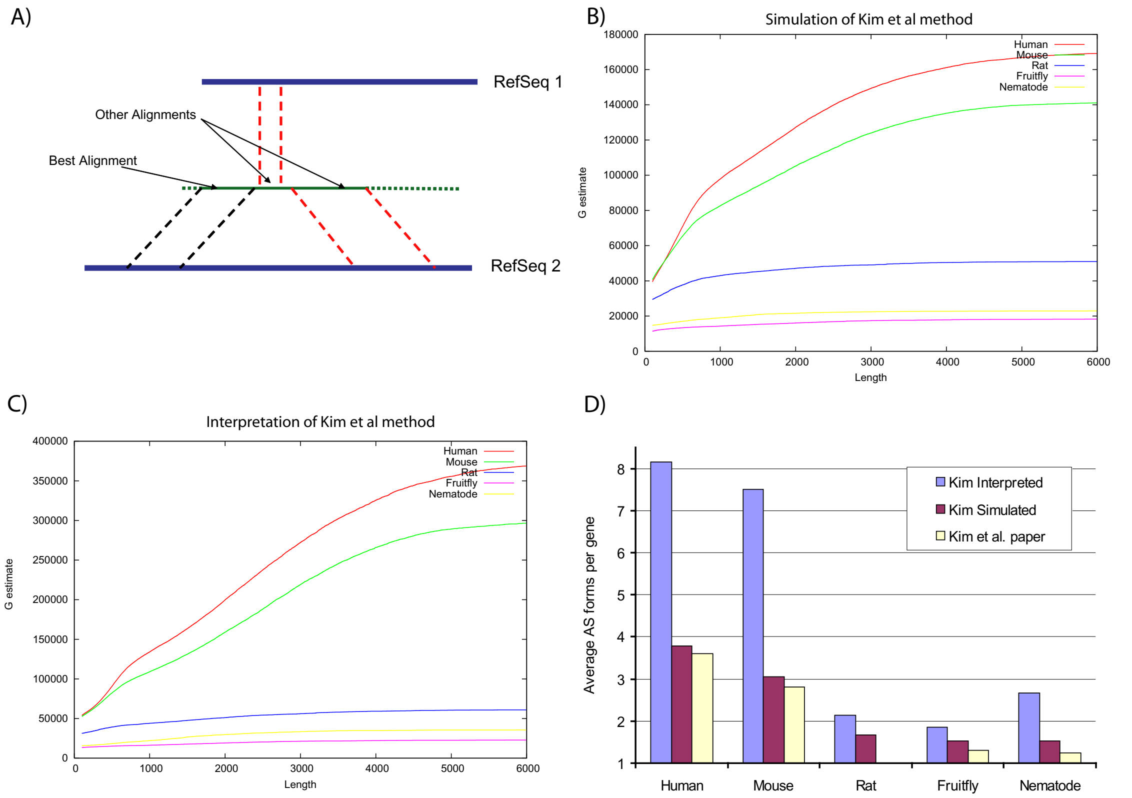


Figure S 1. Simulation and interpretation of the method of *Kim et al.*. A) Possible counting error introduced by the use of the cross_match -masklevel option. This option allows an EST contig to match more than once as long as 98% (for -masklevel 2) of the nucleotides in an alignment are not part of a better alignment. Therefore with a masklevel set to 0 the EST contig in S2A would be represented as a single alignment, whereas with a masklevel value of 2 it would be represented as 3. B) Simulation result that fits best the graphs by Kim et al. (cf with Fig.1A in *Kim et al.*). It assumes the cross_match -masklevel 2 option and a parsing error. C) Our interpretation of the Kim et al. method using cross_match -masklevel 0 and an 'overhang' threshold of 30bp (see S.III for discussion). D) Estimated rate of alternative splicing per gene for the 'interpreted' and 'simulated' methods as well as the values reported by *Kim et al.*.

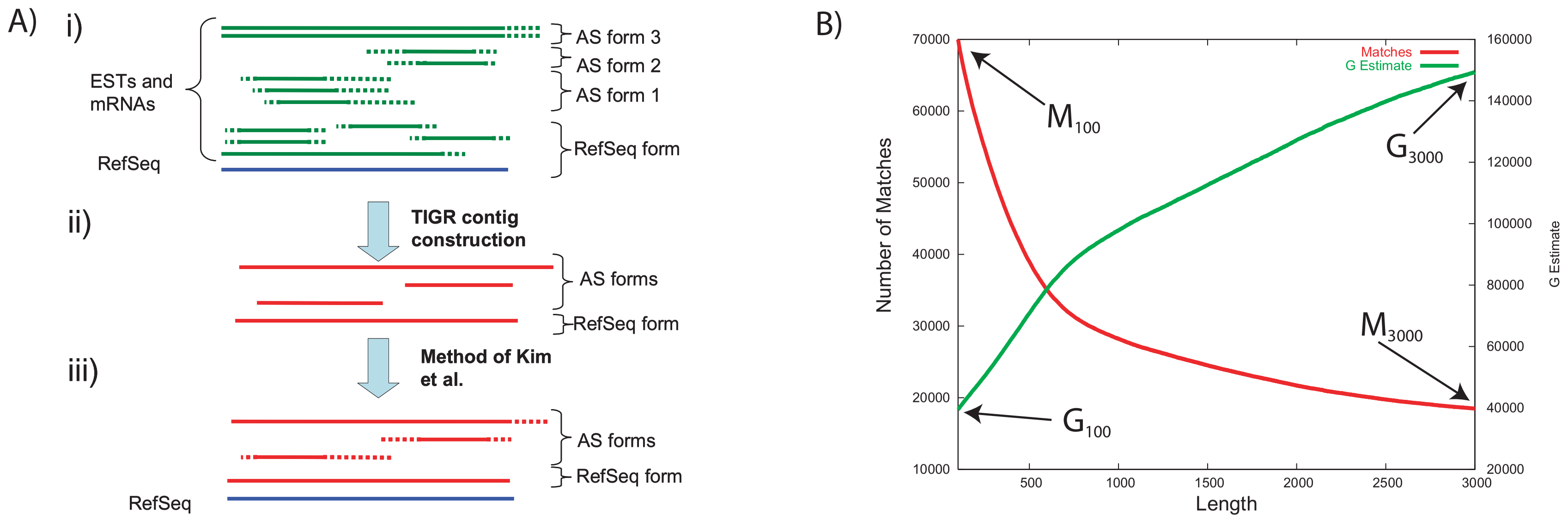


Figure S 2. A) Flow chart of the method of Kim et al. including the TIGR contig construction process. i) TIGR contigs are produced by comparing original EST, mRNA and RefSeq sequences to each other (RefSeq shown in blue); two sequences will belong to the same cluster if they align over 40 base pairs or more, have greater than 96% identity in the aligned region and have less than 30 bases unaligned overhang at the ends. ii) During the TIGR contig construction process each RefSeq sequence and the ESTs that match it with the above criteria are placed in the same cluster and therefore represented as a single TIGR contig (shown in red). Conversely alternative splice forms of this RefSeq sequence will not pass the match criteria and therefore are represented as distinct contigs. (Note that TIGR exploits this fact already for their own AS prediction; http://www.tigr.org/tigr-scripts/tgi/splnotes.pl?species=human) iii) Kim et al.. then align these contigs to the same RefSeq sequences used in step i) (blue). B) For every length x between 100 and 3000bp, Kim et al. calculate Mx, the number of EST contigs that match with >98% identity over that length or that align with less than 50bp of sequence unaligned at either end of the contig. Their estimate for the number of splice forms, G_x, at a given length is a constant (the number of RefSeq sequences used times the number of EST contigs used) divided by the value for M_x. Therefore the value G_{3000}/G_{100} is equal to M_{100}/M_{3000}.

	Rate of AS depending on the number of unaligned bp at end of EST contigs that are tolerated in alignment procedure							
	Interpreted Method				Simulated Method			
	30bp	**50bp**	**70bp**	**100bp**	**30bp**	**50bp**	**70bp**	**100bp**
Human	8.16	6.83	5.54	4.45	4.60	3.78	3.38	3.07
Mouse	7.50	5.63	4.57	3.72	3.72	3.04	2.72	2.48
Rat	2.14	1.95	1.85	1.72	1.75	1.67	1.62	1.55
Fruitfly	1.87	1.72	1.66	1.59	1.56	1.52	1.57	1.47
Nematode	3.22	2.64	2.03	1.78	1.84	1.52	1.43	1.36

Table S1. The rates of alternative splicing detected using ‘unaligned EST’ values of 30, 50, 70 and 100 bp.

Supplementary Information

SI.I Reproducing the method and implementation of *Kim et al.*

We had difficulty in reproducing the results of *Kim et al.* due to i) changes in data sources, ii) unreported non-default parameter choices and iii) possible erroneous protocols. We followed their method as closely as possible, using the same TIGR contig versions but more recent releases of UniGene and RefSeq. As non-default parameter choices are not indicated in the method section, *Kim et al.* kindly explained some of their parameter choice on request. They produced their alignments using the program cross_match with the –masklevel option set to 2. (Figure S1A). Their use of this option was apparently meant to detect the longest EST contig in case of multiple matches either to the same RefSeq sequence or even different RefSeq sequences. It implies though that a single EST contig could be counted as several different splicing forms. However, even when we used this flag our estimates for the rate of splicing were much higher than those of *Kim et al.*. We carried out simulations of possible parameter choices to determine the cause of this discrepancy and found that the parameter set most consistent with the results of *Kim et al.* included a parsing error in which the numbers for unaligned EST contig sequence and unaligned RefSeq sequence were interchanged (Figure S1B). We refer to this method as the 'simulated' method.

We also implemented our interpretation of the method of *Kim et al*., allowing an EST contig to match only once (-masklevel 0), extending the analysis to 6000bp (the human and mouse curves are not yet asymptotic at 3000bp in this analysis), and using a value of 30bp overhang tolerance to fully comply with the TIGR contig assembly procedure (see SI.III). This analysis is shown in Fig. S1C and is referred to as the 'interpreted' method. The estimates for the average number of AS forms per gene in each organism are summarized in Fig. S1D. Although the results vary a lot between the interpreted and simulated implementations, the overall trend of mouse and human having much higher rates of AS than rat, worm and fruitfly remains.

SI.II Testing the dependence of the methods on EST coverage

In order to understand the test of *Kim et al.* that is supposed to show the independence of their method from EST coverage it is necessary to consider the TIGR contig construction process (figure S2A). The construction of TIGR contigs removes the redundancy in EST and mRNA data, therefore for each RefSeq sequence with alternative splice forms there is a single contig which is the length of the RefSeq (the RefSeq form) and then one or more contigs for each alternative splice form (AS forms) whose lengths are proportional to the length of the cDNA/ESTs that support this splice form. *Kim et al.* then align these contigs to the original RefSeq sequences and based on these alignments estimate the rate of alternative splicing per gene as G_{3000}, the estimate of G at a length cut-off of 3000, divided by G_{100} (Figure S2B). As the value G_x at a given length cut-off x is merely a constant divided by M_x (the number of EST contigs that match a RefSeq at that length cut-off), the rate of alternative splicing is equal to M_{100}/M_{3000} (Figure S2B). The theory behind this is that all AS forms of a sequence will match at a low length stringency, whereas only the RefSeq form contigs match at the highest length stringency. In other words M_{100} is a count of both the RefSeq and AS forms and M_{3000} is a count of the RefSeq forms only (see Figure S2B). Therefore we can reformulate the estimate for the rate of alternative splicing as (number of RefSeq forms + number of AS forms)/(number of RefSeq forms), which simplifies to 1+(AS forms/RefSeq forms). Thus when *Kim et al.* resample using lower numbers of EST contigs they are reducing both AS and RefSeq forms equally which means that the ratio and therefore the rate of alternative splicing is unaffected. *Kim et al.* interpret this as proof that their method is independent of EST contig coverage, which is true, but does not mean that it is independent of EST coverage. Methods of AS detection based on EST-mRNA[1] and EST-Genome[2] alignments are known to be sensitive to the levels of EST coverage and as the TIGR contig construction process is based on EST-EST and EST-mRNA alignments it too will be subject to the same limitation. Therefore to test effectively for independence of EST coverage, one would have to resample the EST data before it enters the TIGR pipeline.

In addition the above explains the correlation between the cumulative distribution of contig lengths (fig 1A) and *et al.*'s estimation of G (*Kim et al.* fig.1). When *Kim et al.* count the number of matches at a given length cut-off they are indirectly measuring the length of AS forms.

On a different note, the authors rely on the preprocessing of EST data by the TIGR procedure which reduces the redundancy in EST data, but might also introduce artifacts and additional biases (coincidentally, Ewing and Green, from whom the method is borrowed, used a different EST assembly[3] for their human gene number estimate and their results differed several fold from an estimate based on TIGR contigs[4]).

SI.III Illustrating the sensitivity of the method to parameter choice

One of the parameters used in the construction of TIGR contigs is the so called 'overhang' tolerance, the maximum number of bases of unaligned sequence allowed at the end of ESTs in the same cluster. This is used to take account of the low quality of EST sequences towards the ends. For the construction of TIGR contigs a value of 30bp is currently used (20bp in earlier releases). The method of *Kim et al.* uses a value of 50 base pairs for the maximum number of unaligned overhang at the ends. Therefore we tested the sensitivity of the method to variation of this parameter. Table S1 shows that the value of this parameter has a significant impact on the estimation of the AS rate.

SI References

1. Brett, D *et al. Nature Genetics* **30,** 29-30 (2002)
2. Kan, Z *et al. Genome Research* **12**, 1837-1845 (2002)
3. Ewing, B and Green, P. *Nature Genetics* **25**, 232-234 (2000)
4. Liang, F *et al. Nature Genetics* **25**, 239-240 (2000)

7.2 Alternative splicing and evolution

Alternative splicing seems to be literally fascinating biologists in the last decade. Its significance in terms of function, time and spatial specificity as well as its involvement in numerous diseases are now established facts and turn AS from an oddity into a very powerful mechanism developed during evolution to regulate gene expression. Furthermore, as additional genome sequences become available, comparative genomics brings a new dimension to AS: its conservation and role in evolution. Thus it becomes possible to speculate about the benefits that AS can bring concerning the evolvability of organisms.

The pieces of research building the groundwork for elaborating evolution theories basically tried to trace AS events from one species to another one and hence evaluated the degree of conservation of AS not only at the mechanistic point of view, but also at the single event point of view. I wrote a paper reproduced in the following pages reviewing those studies. The key points of this are the following:

- Most of the genes have constant exon number in human and mouse.

- Exon duplication accounts for most of the remaining genes.

- Exon length is much more conserved than intron length.

- Constitutive exons (present in all transcripts investigated) are equally well conserved than the major form of AS exons (present in more than 50% of the known transcripts), and more conserved than the minor form of alternatively spliced exons (present in less than 50% of the investigated transcripts): the rate of conservation of an exon between species seems to be an accurate predictor of its inclusion in transcripts.

- AS, by allowing modification of the gene structure to remain silent in most of the transcripts and hence for them to normally accomplish their function, also allows minor forms to evolve relatively harmlessly for the cell, until an advantage can be gained and selected for. By decreasing the selective pressure on genes, AS grants the organism an option of trial/error approach for gene evolution.

What the papers say
Alternative splicing and evolution

Stephanie Boué, Ivica Letunic, and Peer Bork

BioEssays, 2003, Vol. 25, Pages 1031–1034

7.3 Alternative splicing detection and databases

Reliability is a major issue in database usage. Depending on how it was built: from a manual collection and annotation of data or automatically, the quality of a database can indeed greatly vary. It is why manually curated databases are developed, which can be used as benchmark sets with whom automatically generated databases are tested for their ability to retrieve a maximum of the right data, and not too much wrong one.

The usual way a manually curated database is developed is that human curators read the scientific literature to extract relevant information, and manually fill up the database with them. This is of course the most reliable method, but is very much time and money consuming. An other challenge is the exponential growth of published scientific articles, which make it uneasy to keep the pace. For those reasons, a new field emerged few years ago to overcome this barrier, namely literature mining. The main idea is to train a computer to get "smart" enough to accomplish a curator task.

The method of choice, which wont be developed here, but which is briefly explained in the following publication, is support vector machine (SVM). It basically processes in two steps:

1. **Learning phase** The SVM receives a list, as exhaustive as possible, of sentences and/or words describing a particular event (here AS) which is manually compiled (what I did here for AS events) and a list of sentences extracted from the same environment but not describing this event, and learns to recognize them. Its efficiency (recall and precision) is calculated by seeing how many true events are retrieved from a set of true and false ones, how many are missed, how many are retrieved which should not be, and how many are correctly not retrieved (manually determined). This phase is iterative: the SVM is trained with sentences until it performs well enough.

2. **Information retrieval and extraction phase**: a subset of documents describing events or scenarios of interest is identified (information retrieval [IR]), and then facts are extracted from these documents and deposited into structured fields (information extraction [IE]) to form a database.

The tool developed to extract information about AS in scientific literature was applied only to abstracts, as full text is not always available, and AS information usually resides in the abstract if the AS event was experimentally proven in the analyzed publication, which is what we are looking for. It is described in the following article.

The main achievements of this work are the following:

- The trained SVM identified 31,123 putative transcript diversity-containing sentences from the MEDLINE database (12,948,515 abstracts);
- False positives were removed manually, leaving 20,549 TD-containing sentences in 13,892 abstracts, which corresponds to a precision of 66%.
- Thanks to semantic parsing, words are assigned a type, which helps classify them into eight different semantic categories: event mechanism (AS type), gene name, tissue name, species, number of described isoforms, difference in structure/function, experimental method used to prove the AS event, specificity of the event (in time, space, physiological state);
- Gene names retrieved were mapped to protein and nucleotide databases in order to get sequence, functional or structural information;
- In total we extracted 9,503 instances of event mechanisms from as many abstracts and 5,028 instances of tissues with associated gene names. Overall, the database contains 3,063, 874, and 207 non-redundant instances of AS, differential promoter usage, and alternative polyadenylation associated with genes and tissues.
- Among the events retrieved, many were not annotated in current databases having AS annotation.
- We identified 959 events describing tissue specificity in AS. These represented 675 AS events for pairs of tissues and 284 events where only one tissue was reported. This information allowed to design a scheme of tissue utilization of AS, which underlines the high frequency of AS in nervous system for example.
- The retrieved information is deposited in a database called LSAT, standing for Literature Support for Alternative Transcripts, which is freely available at the following url:
  http://www.bork.embl.de/LSAT/.

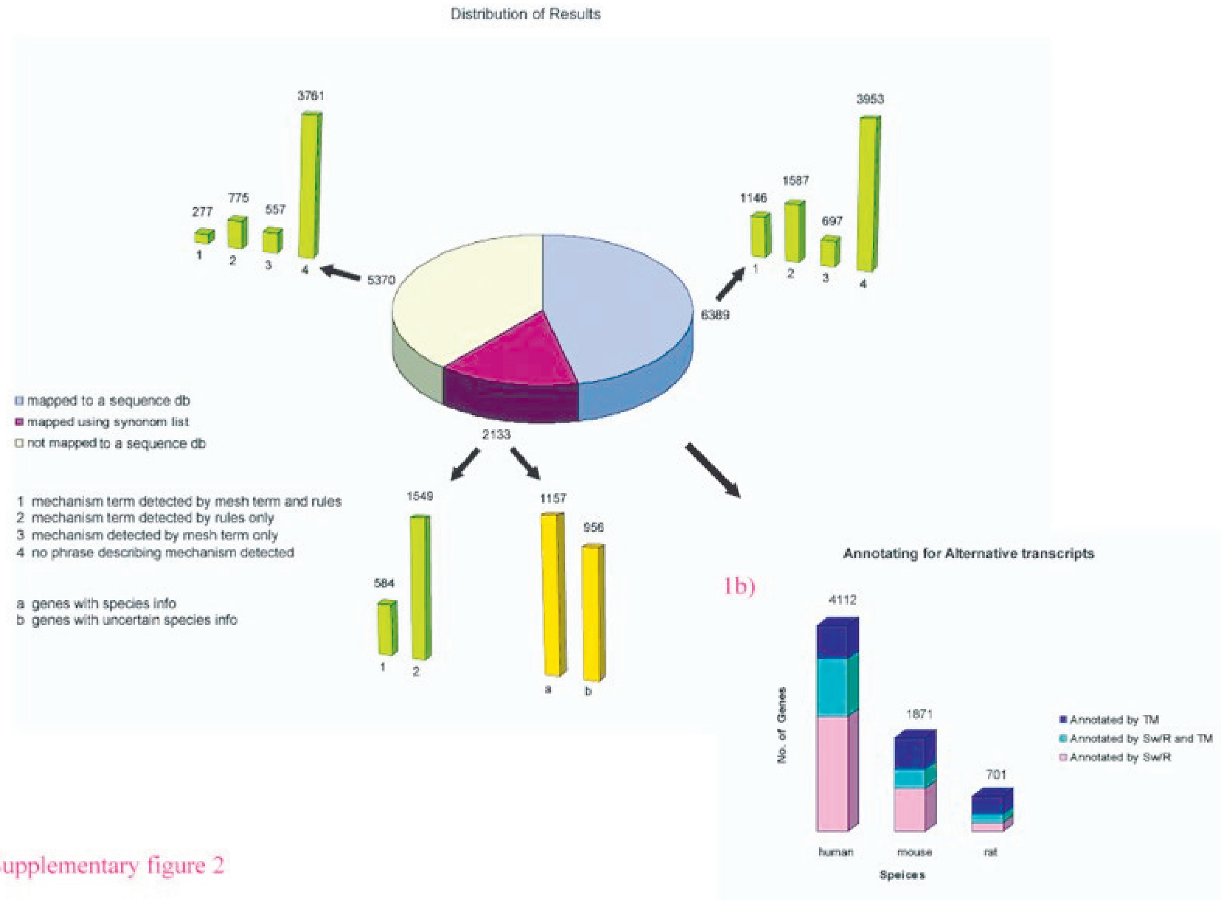
Extraction of Transcript Diversity from Scientific Literature

Parantu K. Shah, Lars J. Jensen, **Stéphanie Boué**, Peer Bork

PLoS Computational Biology, 2005, Vol.1, N°1, e10, Pages 67-73

Entry: 60	Pmid 10102990
Title: A novel form of human neuropsin, a brain related serine protease, is generated by alternative splicing and is expressed preferentially in human and adult brain.	
Genbank: AB008390, AB008927	Ensembl: ENSG00000188879.1
Refseq Report:	
Identfiers	NM_144506, NM_144507, NM_144505, NM_007196
species	Human
Gene Definition	Kallikrein 8 (neuropsin/ovasin) (KLK8)
Comment	REVIEWED
Transcript variants	Four transcript variants in refseq
Swissprot Annotations	
Identifier	KLK8_HUMAN
Description	Neuropsin precursor (EC 3.4.21.-) (Kallikrein 8) (Ovasin) (serine protease)
Annotations	Alternative Splicing
Text Extraction Data	
Gene name	neuropsin
Event	Alternative splicing; was a species-specific splice variant
Experimental method	Sequence analysis of the 946 bp genomic DNA spanning the region encoding the insertion sequence
Tissue	brain
Species	Human, mouse
Isoform-number	two
Specificity	Species-specific

Supplementary Figure 1



Supplementary figure 2

Supplementary Information

Description of the training set

The mechanisms for generating transcript diversity have been studied experimentally, using various biochemical methods including variants of PCR, S1 nuclease assays and blot hybridizations. The conclusion about the mechanism(s) involved, can be reached after nucleotide sequencing and computational analysis. Hence, sentences describing events that generate TD (Supp figure 3) may contain event mechanisms, results of the experimental methods or statements describing observations or presumptions.

The information generally available from these sentences includes gene names, experimental methods, tissue and species specificity, alternative exon function and other biologically interesting properties. The amount of information retrievable from different sentences varies much: from most of this information to a really partial one (categories 1-3; Supp. figure 3). For lack of space, absence of conclusive experimental evidence, or stylistic reasons, the event mechanism may not be mentioned in the abstract text (e.g., [1,2,3]). In these cases, the event may be missing but the presence of other word chunks may give enough bases to consider them positive sentences (category 3). Such events, which we aim to catch automatically, can be verified by manual/automated curation of article full-text, or computational analysis. For example, we detected sentences for DCLK gene, from the articles published in 2004 [4] and 1999 [5]. The article published by Engels *et al*., describes alternative splicing as the event that was perhaps not fully described by Sossey-Alaoui and co-workers. By using the event description our classifier detects the event.

Part of speech tagging

The task of POS-tagging is to assign part of speech tags (e.g., verb or noun) to words reflecting their syntactic category.

Inductive Learning

In the process of inductive learning, positive and negative learning examples are provided to a learning method. The learning performance is then assessed on the set of examples the learner haven't seen before. The process is repeated till the classifier achieves satisfactory performance.

Predicate argument structures

A verb which indicates a particular type of event conveyed by a sentence can exist in its verbal form, its participial modifier format or its nominal form. For example, the normal form of a verb used to describe the event "finding presence of something" would be *detect*, its participial modifier format would be *detecting* or *detected*, and its nominal format would be *detection*. Sentence constituents holding meaningful roles to complete the meaning of an event indicated by the verb are called arguments. (also see below)

Merging multiple syntactic patterns to semantic patterns

For example, in the sentence, 'Northern blot analysis detected the presence of a 2.4kb transcript and a 3.2 kb transcript in brain, liver and pancreas', the phrases 'Northern blot analysis' and 'brain, liver and pancreas' would serve the role of arguments to the verb *detect* with semantic labels of *experimental methods* and *tissues*, respectively. It is clear that variation of the sentence as 'Detection of 2.4 kb and 3.2 kb transcripts present in brain, liver and pancreas by northern blot analysis' would not change the semantic role assigned to constituent 'northern analysis' and 'brain, liver and pancreas'. At the same time in sentence, 'Using RT-PCR and nucleotide sequencing, alternative splicing was confirmed in liver, brain and testis', phrases 'RT-PCR and nucleotide sequencing' and 'liver, brain and testis' would serve roles of *experimental methods* and *tissues*, respectively.

Rules for extracting semantic patterns

For example, a rule to find out the role of the variable region in alternatively spliced transcripts in terms of structure or function could be summarized as follows: "*Take Noun phrase chunks right to different forms of verbs 'lack' (Figure 2; sentence 4) and 'differ'. Terminate when any of the end condition is encountered".* The end condition includes encounter of end of line, break in the sentence, different forms of 'be', words like 'through', 'due to' and 'because'. The rule for extracting experimental methods can be described as follows: "*Take chunks left to the different verbs 'show' and 'detect' (Figure 2; sentence 4, 6, 8, and 9) containing certain keys words (e.g., PCR or blot). Take the chunks to the right if passive form of verbs is used*".

Apart from the phrases extracted using predicate argument structure analysis, event mechanisms were extracted based on bi-gram and tri-gram lists. Tissue specificity was identified by tagging the word 'specific*' that may follow the tagged tissue name or part of the word describing the tissue (e.g. brain-specific). Similarly, 'number of isoforms' was extracted by the fact that such numbers always preceded the tagged event mechanisms. Tissues were tagged using a dictionary compiled from Swissprot and Refseq. Gene names were tagged using an entity tagger [6].

Example entry from the database

Information extracted from the Medline abstracts and different sequence databases is incomplete. Such incompletion resulted in variability in contents for our database entries. For example, information about AS in the human *neuropsin* has been well annotated in Swissprot and RefSeq (Supp. figure 1). Text extraction data in this case added information about tissue, experimental methods, and species-specificity observed in these alternative splicing events.

Supplementary figure legends

Supplementary figure 1: **An example database entry**

Entries in our database have three distinct parts. First part includes the pubmed identifier and title of the abstract. Second part contains mappings from sequence databases like Swissprot, Refseq, GenBank and Ensembl. The third part includes knowledge derived from text with extraction rules.

Supplementary figure 2: Distribution of results

Figure 2a: The pie chart in the middle shows the number of abstracts that could be mapped to sequence databases using literature entries and synonymous list and those that couldn’t (clockwise). The bar graph with categories 1-4 shows number of abstracts in which mechanism could be assigned to genes extracted from those abstracts. We have used MeSH terms and species information to identify gene studied in the abstract (bar graph with categories a, and b).

Figure 2b: We mapped all Swissprot, RefSeq and GenBank sequences to Ensembl genes for human, mouse and rat genomes. Using literature entries present in these databases we mapped our results to Ensembl genes. We could add 674, 637, and 359 annotations for AS for human, mouse and rat genomes, respectively.

Supplementary figure 3: Description of training set

Example sentences from our training set, describing generation of transcript diversity (figure3a) and negative sentences (figure3b) from MEDLINE. Alternative transcripts are generated by many mechanisms or combinations of them. Hence, the SVM classifier has to learn multiple patterns apart from their syntactic variants. The sentences are classified in to various categories and semantic patterns are marked from 1-8. Please see table1 for the pattern labels.

References

1. Rajavashisth TB, Eng R, Shadduck RK, Waheed A, Ben-Avram CM, et al. (1987) Cloning and tissue-specific expression of mouse macrophage colony-stimulating factor mRNA. Proc Natl Acad Sci U S A 84: 1157-1161.
2. Russell DL, Ochsner SA, Hsieh M, Mulders S, Richards JS (2003) Hormone-regulated expression and localization of versican in the rodent ovary. Endocrinology 144: 1020-1031.
3. Lonnerberg P, Ibanez CF (1999) Novel, testis-specific mRNA transcripts encoding N-terminally truncated choline acetyltransferase. Mol Reprod Dev 53: 274-281.
4. Engels BM, Schouten TG, van Dullemen J, Gosens I, Vreugdenhil E (2004) Functional differences between two DCLK splice variants. Brain Res Mol Brain Res 120: 103-114.
5. Sossey-Alaoui K, Srivastava AK (1999) DCAMKL1, a brain-specific transmembrane protein on 13q12.3 that is similar to doublecortin (DCX). Genomics 56: 121-126.
6. Mika S, Rost B (2004) Protein names precisely peeled off free text. Bioinformatics 20 Suppl 1: I241-I247.

Supplementary Information

Description of the training set

The mechanisms for generating transcript diversity have been studied experimentally, using various biochemical methods including variants of PCR, S1 nuclease assays and blot hybridizations. The conclusion about the mechanism(s) involved, can be reached after nucleotide sequencing and computational analysis. Hence, sentences describing events that generate TD (Supp figure 3) may contain event mechanisms, results of the experimental methods or statements describing observations or presumptions.

The information generally available from these sentences includes gene names, experimental methods, tissue and species specificity, alternative exon function and other biologically interesting properties. The amount of information retrievable from different sentences varies much: from most of this information to a really partial one (categories 1-3; Supp. figure 3). For lack of space, absence of conclusive experimental evidence, or stylistic reasons, the event mechanism may not be mentioned in the abstract text (e.g., [1,2,3]). In these cases, the event may be missing but the presence of other word chunks may give enough bases to consider them positive sentences (category 3). Such events, which we aim to catch automatically, can be verified by manual/automated curation of article full-text, or computational analysis. For example, we detected sentences for DCLK gene, from the articles published in 2004 [4] and 1999 [5]. The article published by Engels *et al*., describes alternative splicing as the event that was perhaps not fully described by Sossey-Alaoui and co-workers. By using the event description our classifier detects the event.

Part of speech tagging

The task of POS-tagging is to assign part of speech tags (e.g., verb or noun) to words reflecting their syntactic category.

Chapter 8

Gene expression in space

Nowadays a snapshot of an organism's transcriptome can be analyzed routinely using microarray technology. This technique proved to be very powerful to assess relationships between samples, for example samples taken from different cancer types and/or cancer stages. Hence it became natural to represent microarray data with dendrograms, implicitly linking samples on the sole base of their expression profiles. This was extrapolated and led to the belief that single-gene expression profile is so meaningful that the function of a gene could be assigned based on identity of its gene expression profile to a gene having a known function. This is based on the paradigm that expression is synonym to function. However knowledge is lately accumulating, which disproves this paradigm, and what was called spurious or ectopic expression is in fact quite frequent: a gene can be expressed in a tissue, where it does not seem to have any function, and this without being deleterious; it is called neutral expression.

Using recently published microarray data for mouse and human [68], we produced a tree representing relationships between 47 mouse tissues, and analyzed what causes those relationships. The methods used to produce this analysis are reproduced, as given in supplementary material for the paper after the copy of the article itself.

The key points of this paper can be summarized by the following points:

- We constructed a tissue tree with the neighbor joining method from the Euclidean distances between the gene expression vectors of 16,004 Ensembl genes across 47 adult mouse tissues.

- Many groups of tissues in this tree seem consistent with a clustering by tissue functions. Some other clusters are consistent with ontological relationships (i.e. the developmental history of the tissues).

- Tissue clusters are recovered in the absence of gene function: with two sets of genes with tissue specific function: genes involved in germ cell formation (functioning in testis and oocyte), and genes with brain specific function.

- Even genes that do not experience distinct selective pressures across tissues recover the clusters; it is the case for housekeeping genes which are supposed to exert the same function in every cell, and spliceosomal complex.

- Tissue clusters are recovered from expression of both old and new proteins.

- Similarity in tissue expression profiles is unlikely to represent remnants of ancestral expression patterns.

- Ectopic expression seems to be associated with the tissue-specific expression of neighboring genes, but it is more significant for brain genes than for meiotic genes.

- This paper mainly represents a warning: one should only infer function from gene expression if conservation of expression patterns across species exceeds neutral levels; estimation of such neutral levels must include the effects of transcriptional leakage.

Similar gene expression profiles do not imply similar tissue functions

Itai Yanai, Jan O. Korbel, **Stéphanie Boué**, Shannon K. McWeeney, Peer Bork and Martin J. Lercher

TRENDS in Genetics, 2006, Vol.22, No.3, Pages 132-138

Chapter 9

Gene expression in time: analysis during development

All organisms share common developmental processes. For example, in most metazoan organisms, the basic stages of early embryonic development share similar morphological characteristics and include similar genes. This allows to conduct research on model organisms and transfer this knowledge to the species we are most interested in, usually human. One of these model organisms is *Drosophila melanogaster*. A large body of knowledge about developmental processes has been accumulated thanks to studies in this species, often conducted thoroughly on single-gene mutants.

For the purpose of tracking relative transcript levels during development, microarray analysis has proven to be invaluable. The partial transcriptomes of two major model organisms have already been analyzed; the nematode *Caenorhabditis elegans* [69] during embryogenesis, and an extensive developmental time series in *Drosophila melanogaster* of approximately 30% of all genes covering the entire life span, from the first minutes of development to aging adults [70]. The latter study gave the first insights into global changes of regulation, such as the prominent biphasic expression of many genes in two major stages, either in the embryo and pupa, or in the larva and adult, revealing the molecular similarities in these stages of the life cycle.

Recently, genome-wide expression in fruit fly was measured at the exon level, providing enough resolution to identify alternative splicing in 40% of predicted genes and to identify at least 15% as developmentally regulated [71]. However, as only two time points were considered during early and late embryo development, many transient and tightly regulated processes during embryogenesis would not have been detected.

Here, we apply an extensive analysis of the fly transcriptome, comprising of 12,868 genes (FlyBase 4.0 release [72]), during 30 time points, covering the entire 24h period in which the fertilized egg develops into a larva.

9.1 Material and methods

9.1.1 Generation of the developmental time series data

The samples used in this study are the same samples that were used in Arbeitman et al. [70]. In brief, thirty one-hour time points were collected throughout embryogenesis. To capture the rapid developmental changes that occur during the first half of embryogenesis, overlapping one hour time points were collected for the first six and a half hours of development. The stages of all samples were verified and only tightly staged embryo collections were used for RNA isolation and microarray analysis. The distribution of stages within each time point is shown in figure 9.1. Note that several stages were measured at the same time point due to short length of some of the stages.

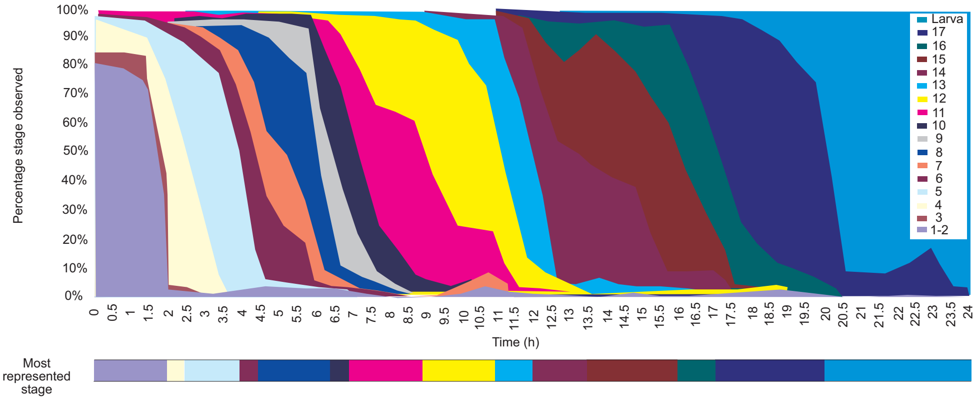


Figure 9.1: **Observed distribution of *Drosophila melanogaster* embryo stages.**
The upper panel shows the actual distribution of stages used for the experiment, whereas the lower panel shows the major stage observed at each measurement period.

Three independent embryo collections were used for each stage of development. Details of the microarray hybridizations are described in [70] and at: http://genome.med.yale.edu/lab/index.htm. All samples were hybridized to a common reference sample. The reference sample, described in Arbeitman *et al.* [70], was made from pooled samples from all stages of the *Drosophila* life cycle. This reference serves as a constant denominator to which the relative levels of expression of each gene in the experimental samples can be compared. The microarrays used for this study consist of PCR fragments of one exon of every predicted *Drosophila* gene [73].

9.1.2 Normalization of microarrays

The spot intensities of the two channels (Cy3 and Cy5) on each microarray were individually normalized using the Qspline method [74] with a log-normal distribution as target (M=ln1000, S=ln1000). The channels were further normalized to correct for spatial biases using a Gaussian smoother with sigma=0.8. After adding a regularization background intensity of 100 to the normalized intensities, a log-ratio was calculated for each gene on each spotted array. This value was semi-empirically chosen to make the spread of log-ratios independent of the spot intensities.

9.1.3 Identification of significantly regulated genes

We performed an ANOVA on each gene in order to determine significant changes in expression as has been done in earlier studies of time series [70, 69].

9.1.4 Local convolution

In order to specify the exact times of activation and suppression, we convolved array data with vectors of 8 integers, for instance x = [0 0 0 0 1 1 1 1]. We selected those matches with correlation coefficients exceeding 0.9, corresponding to p<0.001 ($t = \frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$, t-test).

Throughout testing, trends remained when both the length of the x-vector and the correlation lower limit were either strengthened or weakened.

9.1.5 Global convolution - supervised clustering

When searching for plateaus of expression, we convolved the expression profiles according to $S_{i,j} = c(e, x_{i,j})$, where e is the expression profile vector and

$$x_{i,j} = \begin{cases} i \; for \; i \rightarrow j \\ 0 \; otherwise \end{cases}, j > i$$

.

Here c is the correlation coefficient function. The expression profile is considered to be active from i to j if the maximum of S exceeds 0.8. This approach is conceptually similar to the method employed by Sasik *et al.* [75], although here we actively look for steady plateaus followed by sharp declines in expression.

Furthermore, an advantage of using correlation coefficients is that p-values are given, removing the need for random sampling. This method of clustering was chosen since it does not assume Euclidean distance between genes. A Euclidean distance implies that there is no time dependency, which is not consistent with our expectations. Furthermore, the resulting clusters are more readily explained biologically as contiguous phases of up- and down regulation. For class II, in order to distinguish transcript expression profiles from transcripts which are maternally deposited, we set the first six data points to low expression and varied the remaining 24. Hence, the number of combinations is 276 and not 435.

9.1.6 Orthology

Orthologs of the *Drosophila* genes in *Anopheles gambiae* were retrieved from the STRING database [76].

9.1.7 Lethality

We consider all genes as lethal that were annotated as ”Phenotypic class: lethal” in Flybase [72] unless the time of manifestation was given and stated manifestation of the phenotype only in the larvae or later stages. In total, 711 genes of the 12,868 genes (5.5%) tested were considered as lethal.

9.1.8 GO annotation

Overrepresentation of GO categories was analyzed using the GOSSIP program [77], correcting for multiple testing using the false discovery rate or FDR and the family wise error probability FWER.

9.1.9 Transcription factors

336 genes annotated as ’transcription factor’ were extracted from FlyBase [72] and manually curated (Tobias Doerks, personal communication).

9.1.10 Ubiquitination - PEST degradation signals

The PEST-find program [78] was used to perform proteome-wide computational search for PEST regions. The number of PEST containing proteins was counted among all significantly regulated genes as well as within each ”expression class”. The statistical significance of PEST overrepresentation was calculated for each

”expression class” compared to all significantly regulated genes using the exact hypergeometric test.

9.1.11 *In situ* data

The *in situ* data were retrieved from the Berkeley database [79]. This database contains *in situ* data annotation for 2,152 genes with 211 anatomical terms that are based on pictures taken at 5 different developmental times. Data is available for 253 out of our 842 significantly correlated genes.

9.1.12 Pathways

1,309 genes grouped in 31 pathways or groups of functionally related genes were retrieved from the Interactive Fly database (http://www.sdbonline.org, [80]). 86% of them are spotted on the microarray.

9.2 Results and discussion

9.2.1 Estimating expression changes of genes during embryogenesis

Two different statistical methods were used to identify genes that change in expression during embryonic development:

- The widely used **ANOVA** was applied to identify significant changes in the general expression level. We found significant changes in transcript levels over time for 86% ($p<0.05$, 70% at $p<0.001$) of all genes. This compares to *C. elegans* [69] and an earlier analysis of *D. melanogaster* [70] where 68% and 95% respectively ($p<0.05$) of the genes were found to change expression during embryogenesis. However, the actual implementation of ANOVA may differ slightly.
  The high number of genes with a significant change in expression level could suggest that an ANOVA is not sufficiently specific. It indeed would retrieve short changes up and/or down, which might not be biologically significant.

- To explicitly analyze the temporal dependency of expression levels in individual genes, a **runs test** was used, which, unlike ANOVA, takes the temporal ordering of the data points into account and hence fits better with our subsequent analyses; it suggests temporal changes in transcript levels for 65% of genes during embryogenesis.

These estimates give a global overview of the amount of changes occurring during the entire embryonic development, however they do not pinpoint when transitions in gene expression programs occur. To get a more exact time measure, we searched for changes in expression levels using local convolution methods (see figure 9.2).

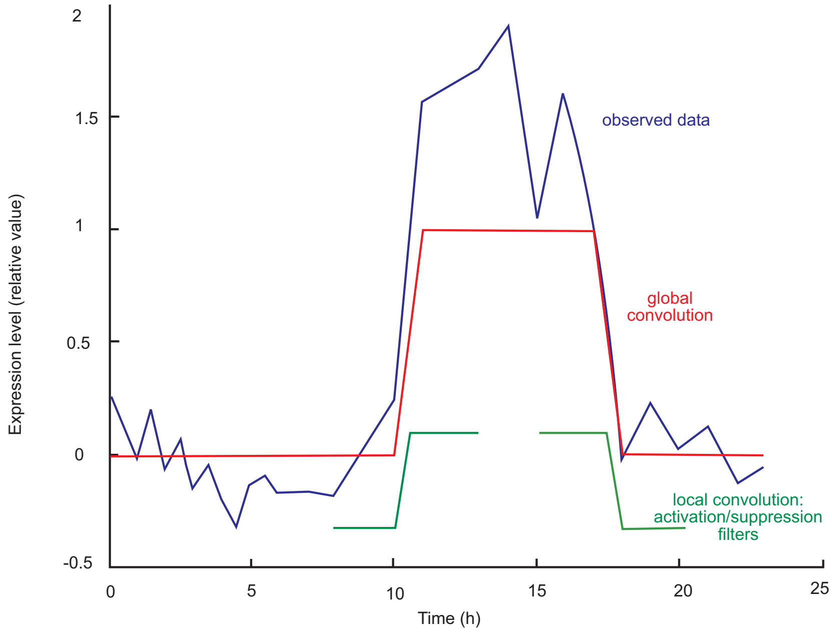


Figure 9.2: **Schematic representation of the features captured by the different methods applied.**
The blue line represents the actual data, while the green and red lines are local and global convolution respectively. Here, the data matches the global convolution vector nicely, and is therefore clustered with other genes that match the convolution profile. The local convolution finds local areas that match the profile. Therefore, it is better suited to pick up general trends, since it is much less stringent than the global convolution.

More specifically, we required four points of low expression and four subsequent points of high expression (or vice versa) even if the amplitude change was relatively low (see methods). This type of convolution not only requires a sharp increase or decrease of expression, but also that the change in transcript level is consistent over a period of time, thereby reducing the rate of false positives due to individual points of high noise.

Using this approach, we mapped 6,233 sharp and consistent changes in transcript levels (2,808 increases and 3,425 decreases) to time points and developmental stages (see figure 9.3).

As already indicated in the study of Arbeitman *et al.* [70], several developmental stages show an increased frequency of transcript level changes during embryogenesis. This can now be confirmed genome-wide. The local convolution analysis

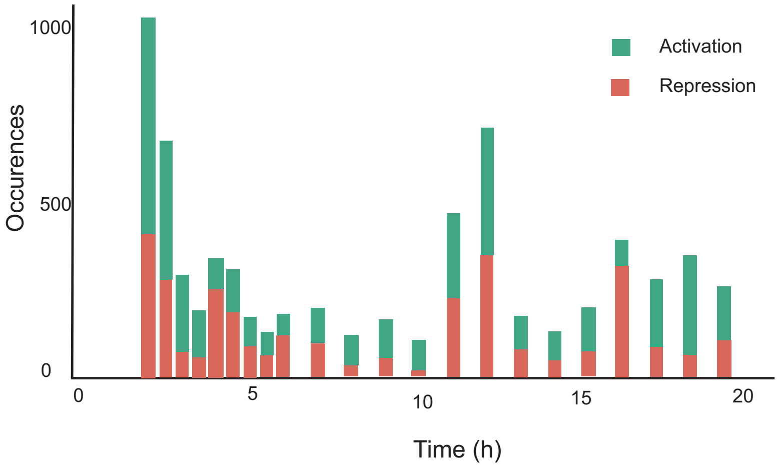


Figure 9.3: **Activation and repression of fly genes during embryogenesis.**
Green bars indicate points of sharp expression changes from low to high (activation) and red bars signify changes from high to low expression (repression).

also revealed that increase in transcript level of one group of genes often coincides with the decrease of transcript expression of another group of genes and vice versa indicating coordinated waves of expression (see figure 9.3).

A high number of expression changes is observed at 2-3 h, representing the initiation of zygotic transcription and the parallel decay of some maternal transcripts. This first stage of dramatic change corresponds to the events just after cellularization. It is consistent with the major morphogenetic changes leading to germ layer formation that happen in that time period in the cellularized embryo. It might also reflect the embryo patterning that begins along both the anterior-posterior and dorsal-ventral axis.

A second major period of transcript expression change (both increase and decrease) was observed at roughly 12 hours, corresponding to the end of the dorsal closure, the terminal differentiation of many tissues, and to the invagination of the epithelial cells that will become the imaginal discs.

A third period of gene expression change is observed at 16h when a discrete set of transcripts decrease their expression levels, followed by an intense increase of transcript levels of another set of genes (17-19h). This could possibly be in preparation of the transition to the larval stage. Generally, sharp decreases of mRNAs seem to be more confined to particular time points.

The correlation between times of increase and decrease in transcript levels suggests the existence of co-regulated groups of genes which drive major developmental events during embryogenesis. The global convolution method allows us to classify genes according to this criterion.

9.2.2 Classification of gene expression behavior

To be able to group genes whose transcripts follow similar patterns over the full 30 time points, and to correlate this with the periods of rapid expression changes, we used global convolution (see methods and figure 9.2). We assigned genes to three general expression classes which are characterized by distinct plateaus of low and high expression during embryogenesis:

- **class I (maternal)**: genes encoding transcripts that start with a high relative transcript level, which subsequently decreases;
- **class II (transient)**: genes whose transcripts levels first increase and later decrease;
- **class III (activated)**: genes encoding transcripts for which we only observe an increase in expression: the level stays high until the end of embryogenesis. Class III gene transcripts, though, are most likely not present at high levels during the entire *Drosophila melanogaster* life cycle. Indeed most of the corresponding genes in the previous study return to low levels at various time points beyond embryogenesis [70].

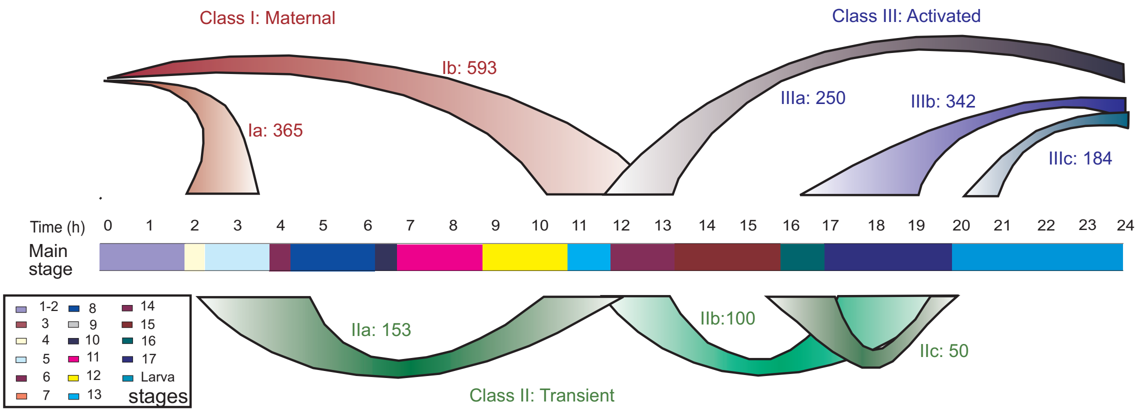


Figure 9.4: **Major classes of transcript levels, as determined by global convolution.**
Arcs represent the dominant subgroups within class I, II and III transcripts. For instance, class I is dominated by two main subgroups I:a and I:b. Time is in hours, and the main stages present at every measurement time point are represented by colored boxes. The time of increase and decrease of the transcript groups coincide with those derived by local convolution.

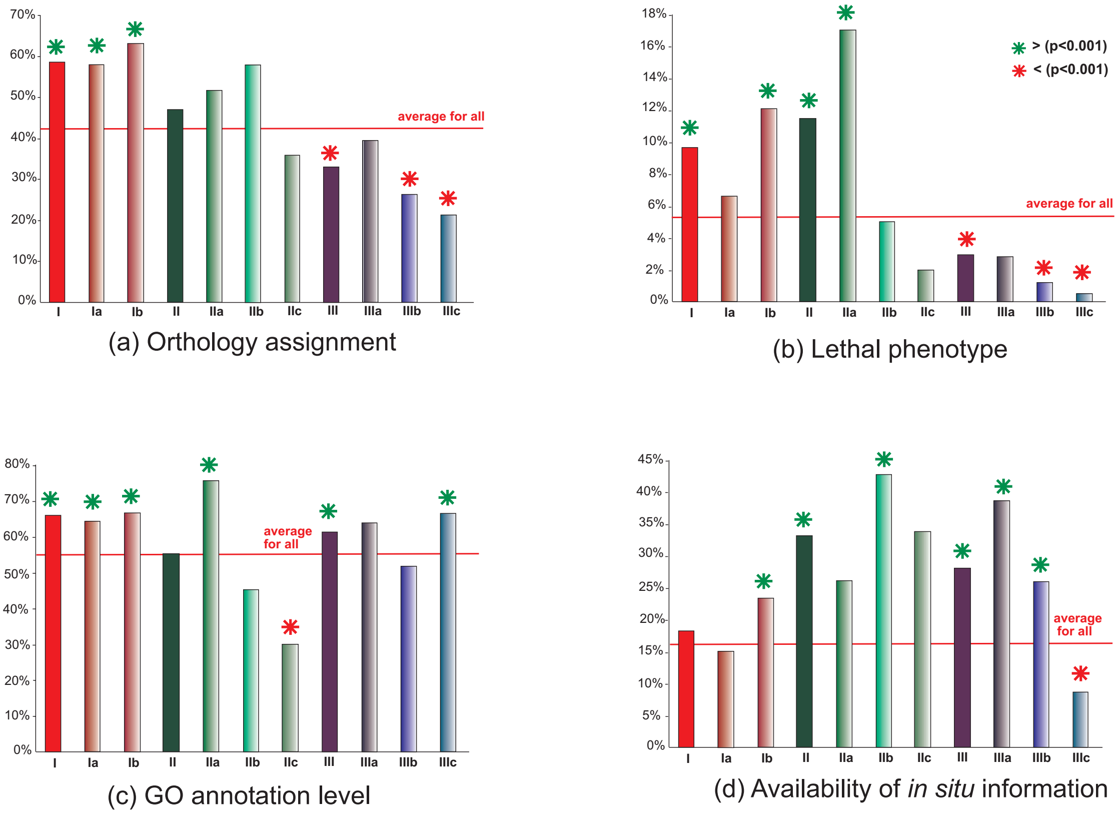
Using global convolution we found significant expression correlation coefficients (r >0.8, p< 10^{-4}, t-test) between transcript levels and global convolution profiles for 26% (3,379) of the transcripts present on the array. Of these, we classified 1,534 as class I, 792 as class II and 1,053 as class III. Many genes do not fit to any of these three classes: genes that are constitutively transcribed, not transcribed at all during embryogenesis, or whose levels vary very quickly. Furthermore, the requirements for assignment were rather stringent; the entire expression profile must fit the categorization to a high degree (30 time points), whereas in the local convolution, only eight time points are considered (corresponding to $p< 10^{-3}$, t-test). This leads to a low rate of false positives at the cost of sensitivity. For instance, we are likely to miss many genes with very short albeit biologically important periods of transcriptional activity, as for example genes that vary cyclically with the cell cycle [81].

Within each of the three global classes, groups of genes can be identified with common times of increase and/or decrease of relative transcript levels (see figure 9.4). More than 62% of the class I (maternal) genes can be classified into two major groups, class Ia and Ib, whose transcript levels decrease at 3-5h and 12-14h, respectively. For the class II genes, even though there are 276 possible combinations of time points of increase and decrease of transcript levels, as many as 38% of the 792 class II genes fall into only three groups: IIa (2.5-12h), IIb (11-20h) and IIc (15-20h) containing 153, 100 and 50 genes, respectively. Of the genes whose transcript levels increase but are not observed to decrease during embryogenesis (class III), more than 73% can be classified into three main groups, class IIIa , IIIb, and IIIc (times of increase at 12-13h, 16-19h, 20h of development).

In order to identify some biological principles underlying these different co-expression groups, and also as a general quality control, we studied these groups using various sources of biological information:

- Orthology assignment to *Anopheles gambiae*
- Lethality in case of mutation of the gene
- Annotation of function thanks to GO terms from the flybase database
- Presence of transcription factors, sensitivity to ubiquitination and hence to degradation by the proteasome (marked by the presence of a PEST sequence)
- RNA *in situ* hybridization data
- Pathways or groups of genes conducting one function, from the flygrid database

The results of these analyses are summarized in figures 9.5 to 9.8.



(a) Orthology assignment

(b) Lethal phenotype

(c) GO annotation level

(d) Availability of *in situ* information

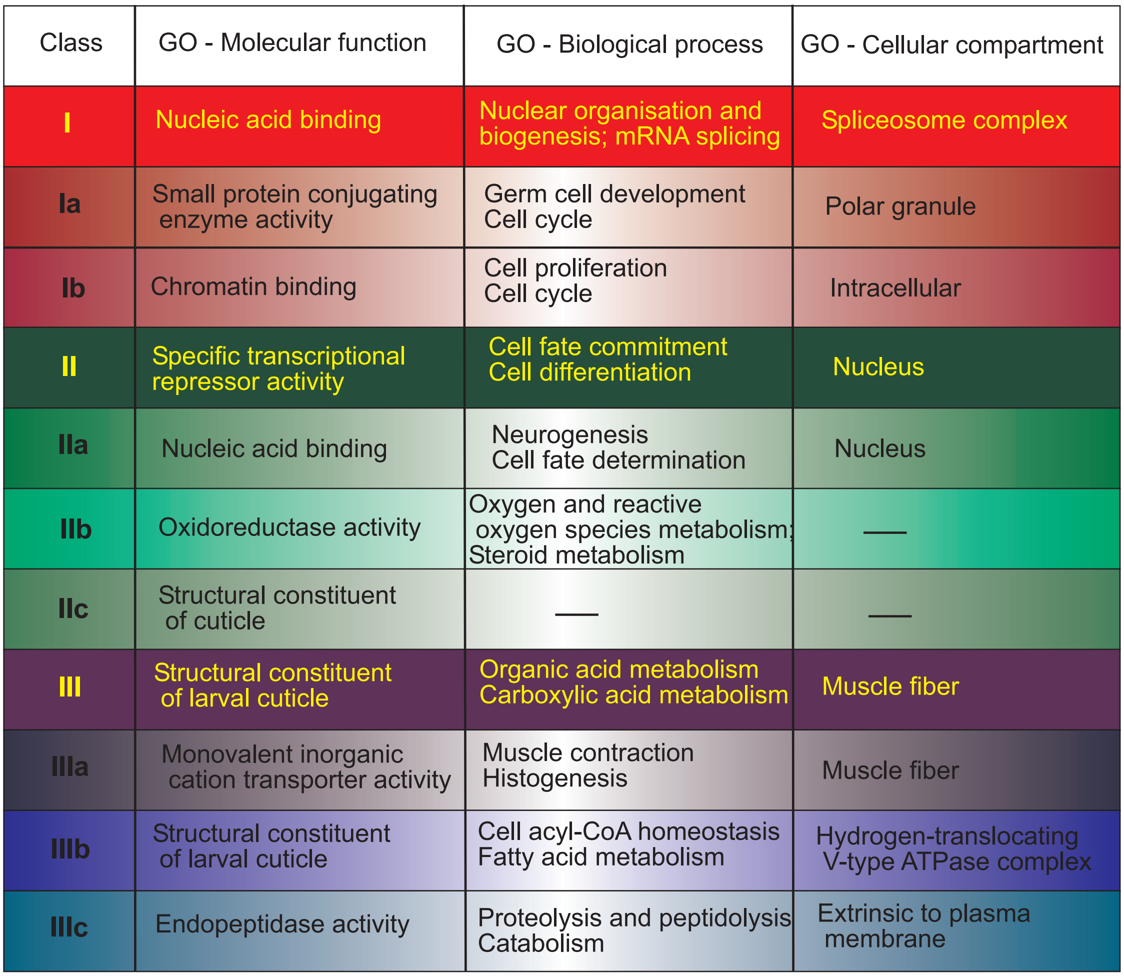
Figure 9.5: **Diverse statistics on the gene classes.**
The red line represents the average for all available data. The green star means over representation, and the red star underrepresentation for one class, compared to the average, with the p-value $p < 0.001$. (a) **Orthology assignment**: proportion of genes for each class for which *Anopheles gambiae* ortholog could be assigned. (b) **Lethal phenotype**: proportion of genes which are marked as having a lethal phenotype. (c) **GO annotation level**: proportion of genes for which GO annotation is available. (d) ***In situ* information**: proportion of genes for which *in situ* data is available.

9.2.2.1 Class I (maternal) genes

With 1,534 of the 12,868 genes present on the microarray, class I genes represent the most populated of the three major expression classes. It is long known that a large number of transcripts are deposited in the oocyte during gametogenesis. Among other vital functions, they have been shown to be responsible for establishing the major body axes and for the initiation of zygotic transcription[82].

The analysis of available data highlighted the following properties of the class I genes:

- The importance of these genes is reflected by:
  - a higher than average fraction of orthologs in this group shared with *Anopheles gambiae*: these genes are more conserved than average (t-test, $p < 0.001$, see panel (a) from figure 9.5)



Class	GO - Molecular function	GO - Biological process	GO - Cellular compartment
I	Nucleic acid binding	Nuclear organisation and biogenesis; mRNA splicing	Spliceosome complex
Ia	Small protein conjugating enzyme activity	Germ cell development Cell cycle	Polar granule
Ib	Chromatin binding	Cell proliferation Cell cycle	Intracellular
II	Specific transcriptional repressor activity	Cell fate commitment Cell differentiation	Nucleus
IIa	Nucleic acid binding	Neurogenesis Cell fate determination	Nucleus
IIb	Oxidoreductase activity	Oxygen and reactive oxygen species metabolism; Steroid metabolism	—
IIc	Structural constituent of cuticle	—	—
III	Structural constituent of larval cuticle	Organic acid metabolism Carboxylic acid metabolism	Muscle fiber
IIIa	Monovalent inorganic cation transporter activity	Muscle contraction Histogenesis	Muscle fiber
IIIb	Structural constituent of larval cuticle	Cell acyl-CoA homeostasis Fatty acid metabolism	Hydrogen-translocating V-type ATPase complex
IIIc	Endopeptidase activity	Proteolysis and peptidolysis Catabolism	Extrinsic to plasma membrane

Figure 9.6: **Statistics on function of the classes**
Result of the Gossip analysis of annotation of molecular function (best scoring term), biological process (two best scoring terms), and cellular compartment (best scoring term).

  - a high proportion of lethal phenotype genes (hereafter to referred to as lethals) (t-test, $p < 10^{-3}$, see panel (b) from figure 9.5)

- The analysis of Flybase GO annotation [72] revealed that the class I genes are better characterized than the average (t-test, $p < 0.001$, see panel (c) from figure 9.5). Moreover, they show a significant over representation of genes involved in "nuclear organization and biogenesis" and "nuclear mRNA splicing" (see figure 9.6). These genes facilitate the organization of the chromosomes and nuclei during the very rapid cell divisions in the precellular blastoderm embryo.

- They are enriched in transcription factors ($p < 0.05$).

- The cell cycle, chromatin organization, and DNA replication enzymes and protein cofactors genes are enriched in maternal genes (t-test, $p < 0.05$, see

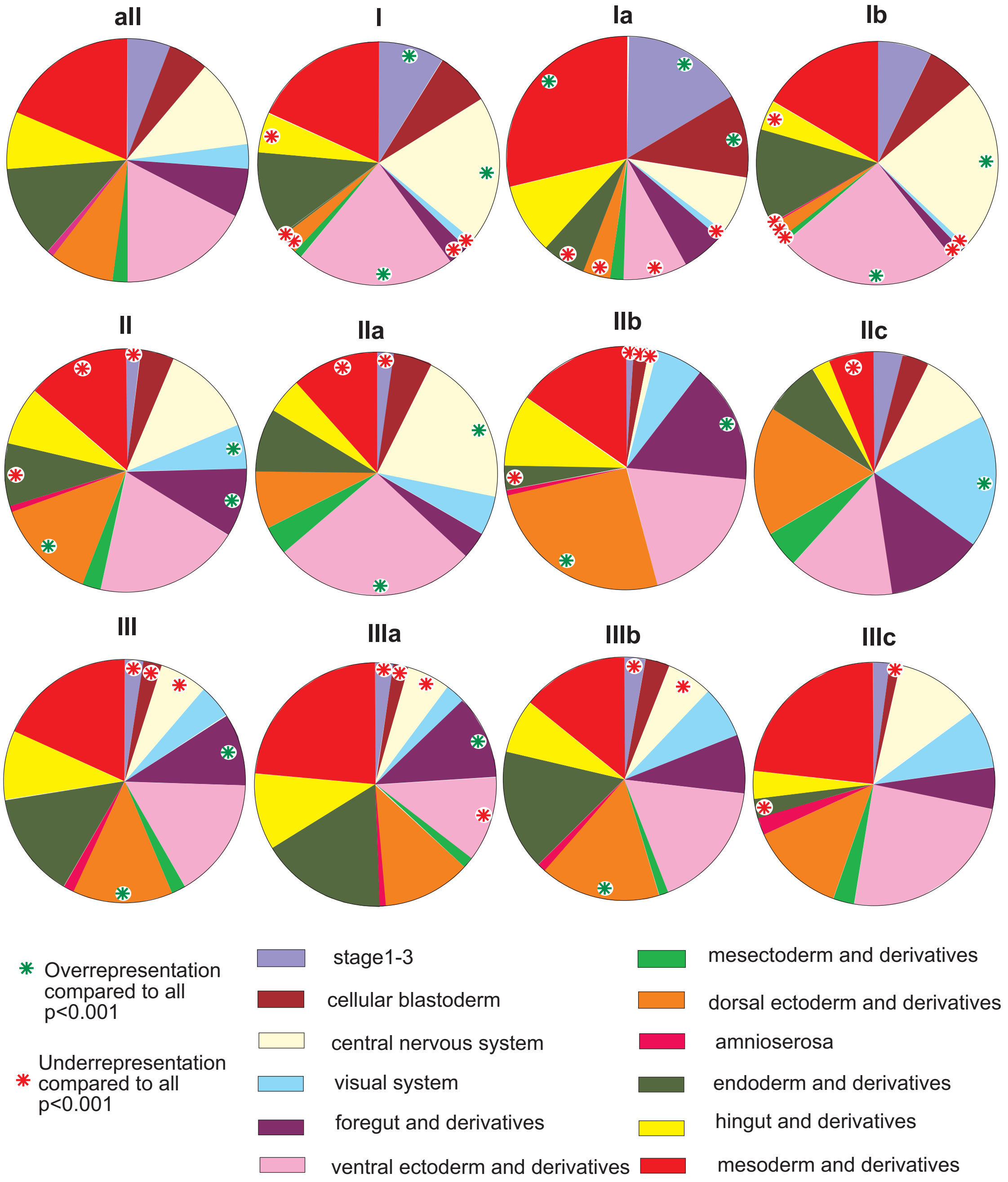
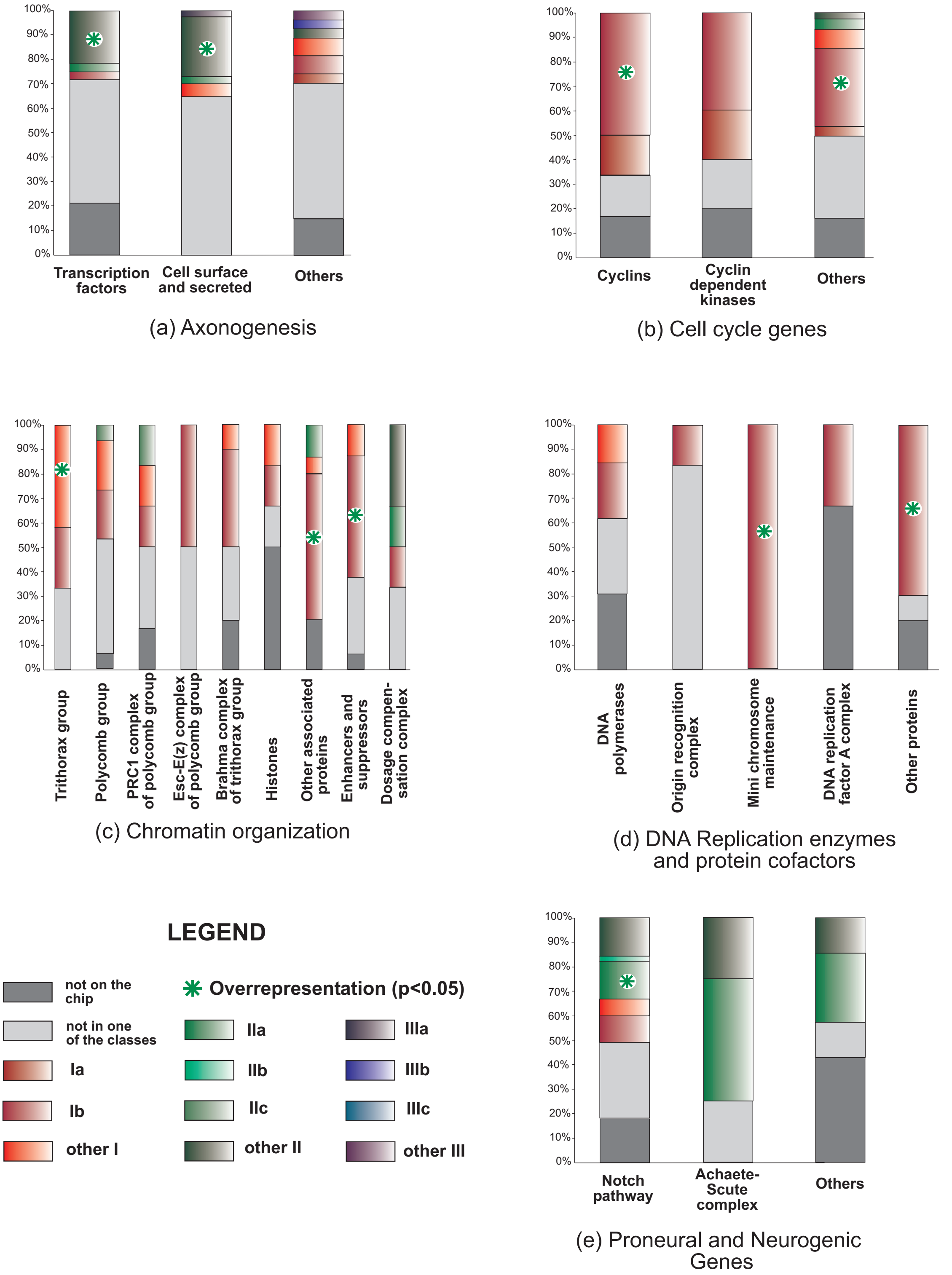


Figure 9.7: **Statistics on the *in situ* data available for the gene classes** Distribution of tissues groups for each class of genes, and for all *in situ* data available; a red star indicates a significant underrepresentation of one tissue group compared to average ($p < 0.001$), a green star indicates over representation.

panels (b), (c) and (d) from figure 9.8).

365 of the 1,534 maternal genes change transcript levels at 1.5-3h (group class Ia, see figure 9.4). As expected, the functionally characterized genes of the class Ia group are mostly involved in early development and in the cell cycle according to the interactive fly database [80].



(a) Axonogenesis

(b) Cell cycle genes

(c) Chromatin organization

(d) DNA Replication enzymes and protein cofactors

(e) Proneural and Neurogenic Genes

Figure 9.8: **Gene classes distribution of chosen Flygrid pathways or groups of genes**

For each of the panel, genes were retrieved from the Flygrid database, which have a common function or are involved in common pathways, and mapped to the classes. A green star means a significant over representation of one gene class within one pathway compared to the average distribution of gene classes over all pathways or gene groups.

The class Ib group consists of 593 genes, which encode transcripts with decreased levels of expression by 10-11h. These genes are on average well characterized, and have a higher annotation in the interactive fly database of genes involved in the cell cycle and cell proliferation processes. They also have an over representation of chromatin binding function. Furthermore, class Ib is enriched in lethal genes ($p < 0.001$). Some group of genes (defined in Flygrid) involved in chromatin organization, as well as the mini-chromosome maintenance family and some other proteins involved in DNA replication are found significantly more within this group than in other classes or unclassified genes.

9.2.2.2 Class II (transient) genes

Despite the stringent requirements on class II genes to have both, a sharp increase and a sharp decrease in transcript levels, 792 fly genes were classified in this class. This is consistent with earlier reports in *D. melanogaster* [70] and also *C. elegans* [69]. The class II genes are enriched in transcription factors and lethal phenotype genes and not surprisingly, this class shows an over representation of genes with functions involved in development (cell fate commitment and cell differentiation for example).The class II group is also enriched in genes that encode known transcription factors ($p < 0.05$).

Of the class II genes, 303 (38%) are found in only three groups: IIa, IIb and IIc, each defined by specific times of increase and decrease of transcript levels (see figure 9.4). In addition to the common features above, the groups also differ from each other. For example, more genes from class IIa (with a plateau of increased transcript levels starting at 3-6 h and decreasing at 12-13 h) have been functionally characterized than average for the genome ($p < 0.01$). Conversely, genes in class IIb (13-14 to 19-21h) and IIc (17-18 to 21h) are poorly characterized (see panel (c) from figure 9.5). The class IIa genes seem also to be more generally involved in neurogenesis, and cell fate determination (see figure 9.6), as is also shown with their over representation in the proneural and neurogenic genes (see panel (e) from figure 9.8).

The analysis of the *in situ* data available (see figure 9.7) shows that genes from the class IIa group are expressed in the procephalic ectoderm, ventral ectoderm, sensory complex and central brain neurons. Again, this strongly suggests a role for class IIa genes in nervous tissue and brain development.

Despite the low proportion of functionally characterized genes in the class IIb group (51 of 100 genes), there is an enrichment in oxidoreductase and peroxidase functions (GO classification).

The class IIc group encompasses 50 genes. Only 19 of these are annotated, and most are involved in cuticle formation. The Drosophila embryo secrets a hard proteinaous material, which forms a thick protective cuticle surrounding the larvae. Furthermore, class IIb is linked via in situ data to the dorsal ectoderm, suggesting that also many genes in class IIb may contribute to the cuticle formation. Therefore, groups IIb and IIc may be of interest when designing insecticides or planning experiments which target the cuticle.

Moreover, another neurogenesis-related mechanism, namely axonogenesis is over representated among the class II genes which are not clustered in one of the groups.

9.2.2.3 Class III (activated) genes

Of the 1,053 transcripts with sharp increase in expression levels but without a subsequent decrease during embryogenesis (class III), 250 are activated at 11-12h (IIIa), 342 at 16-18h (IIIb), and 184 at 20h (IIIc). These genes are the most species-specific of the three categories: we found indeed significantly fewer orthologs to predicted genes from *Anopheles gambiae* than for the genes represented on the microarray as a whole (see panel (a) from figure 9.5). Furthermore, known transcription factors are significantly underrepresented in this group ($p < 0.001$), which implies that the mRNA expression of transcriptional regulators is likely to be under tight regulation. Genes of the class III are less associated with lethal phenotype than the average for the whole chip. According to the function annotation in the FlyBase database, genes of the class III are usually associated with the metabolism or catabolism functions of the cell, most probably preparing for the transition from the embryo to the larva stage.

9.2.3 Coordinated regulation of transcripts and protein products

The precise regulation of a function can not be only regulated at the transcript level. Indeed, from the decrease in transcript levels during embryogenesis discussed above to the degradation of the respective protein, and hence the extinction of the function, it might take hours if regulation only takes place at the mRNA level, through transcriptional repression and/or mRNA degradation. Such a slow decay in protein activity could be harmful if the protein is supposed to act only during a certain stage of embryogenesis, mostly if it regulates other genes or proteins. We thus hypothesize that the protein products of most transcriptionally repressed genes are inactivated, for example through targeted degradation controlled by ubiquitination. This mechanism has previously been suggested to be responsible for the degradation of maternal proteins in *C. elegans* [69] as well as for proteins that are periodically expressed during the mitotic cell cycle in *S. cerevisiae* [81].

To test this hypothesis, we used a computational method to systematically predict PEST regions in the *D. melanogaster* proteome and compared the percentage of PEST-containing proteins encoded by the genome to that of the genes in each expression class. The highest percentages of PEST-containing proteins are encoded by class I (39%) and class II genes (37%). Both groups are significantly enriched in PEST regions compared to the proteome-wide content (31%) ($p < 0.001$). The difference between class I and II is not statistically significant. In contrast, PEST regions are found in only 21% of the proteins encoded by class III genes, which is significantly less than expected ($p < 0.001$). These observations are consistent with the hypothesis of a coordinated down-regulation of genes and their protein products during embryogenesis.

9.2.4 The Notch pathway: insights gained by this study

The most tightly regulated pathway according to our method (= the one containing the highest over-representation of the transient class genes) is formed by the proneural and neurogenic genes (http://www.sdbonline.org/fly/aignfam/neuropro.htm). According to the Interactive fly database [80], there are 45 such genes. In depth literature research, as well as investigation of genes having a similar sharp decrease of their expression level at 12h (like Notch itself), added 34 members (due to high number of papers, not cited here), most probably shown to be part of the Notch pathway after the last update of the Interactive fly database. Of those 79 genes, 68 were assayed on the array and 43 fit significantly to the category of class II genes. Thus, >54% of the Notch pathway genes are transient, compared to an average 6.5% on the whole chip.

It is interesting to see that although the Notch pathway is broadly used in development, there is a general shutdown of many of its members (among which Notch itself) around the same time point (12h). Hence cell fate commitment has to be done by this time, at least as far as embryogenesis is concerned. However, the Notch pathway will be used again later to define the fate of some cell lines, or maintain some cell populations like the neural stem cell one [83], with somewhat different role though (cell cycle).

Thanks to this specific sharp decrease at 12h of development, we propose a list of genes distributed in three categories (see tables 9.1 and 9.2):

1. ”best” candidate genes: have not yet been shown to be involved directly in the Notch pathway, but are involved more generally in development;

2. "less certain" candidates: have the same sharp decrease, but might or might not have anything to do with Notch itself: no further evidence was found;

3. members of the Notch pathway, included in the interactive fly definition of this pathway, or which are proven in the literature to be involved with Notch.

We hope that this list in table 9.2 will be used by experimentalists and Notch experts to investigate further the Notch pathway.

As a "positive control", genes from the class IIa showing the same sharp decrease, and proven to be involved with the Notch pathway are described in the table 9.1.

cat	F Id	corr	start	stop	symbol	full name	Loc
3	FBgn0000216	0.85	6	18	Brd	bearded	71A4
3	FBgn0029123	0.87	6	17	SoxN	sox neuro	29F2
3	FBgn0003448	0.88	6	18	sna	snail	35D2
3	FBgn0000022	0.89	8	17	ac	achaete	1A6
3	FBgn0002735	0.89	10	18	HLHm g	HLHm gamma	96F9
3	FBgn0002631	0.89	7	17	HLHm5	HLHm5	96F10
3	FBgn0002561	0.90	7	17	l(1)sc	lethal of scute	1B1
3	FBgn0002932	0.90	6	17	neur	neuralized	85C2-3
3	FBgn0040808	0.91	6	17	BobA	brother of bearded A	71A2
3	FBgn0001325	0.91	6	18	Kr	kruppel	60F5
3	FBgn0003326	0.92	6	17	sca	scabrous	49C3-D3
3	FBgn0002734	0.92	9	18	HLHm delta	HLHm delta	96F9
3	FBgn0040296	0.93	7	18	Ocho	ocho	71A4
3	FBgn0000180	0.93	6	17	bib	big brain	30F5

Table 9.1: Known genes belonging the Notch pathway from the class II a. The F Id is the Flybase identifier. Loc is the physical map location of the gene.

cat	F Id	corr	start	stop	symbol	full name	Loc	comments
1	FBgn0025776	0.87	8	17	ind	intermediate neurone defective	71B2	role in brain development interacts with Delta and Sox Neuro
1	FBgn0001983	0.90	10	18	wor	worniu	35D2	role in brain development linked to snail and escargot
1	FBgn0000411	0.91	6	18	D	dichaete	70D3	transcription factor involved in neuroectoderm development
1	FBgn0004394	0.91	8	17	pdm2	POU domain protein 2	33F2-3	transcription factor involved in neuroectoderm development
1	FBgn0001148	0.92	9	18	gsb	gooseberry	60F2	transcription factor involved in neuroectoderm development
2	FBgn0002970	0.84	9	18	nub	nubbin	33F1-2	transcription factor involved in cell fate determination
2	FBgn0000117	0.84	7	18	arm	armadillo	2B14	regulated by Notch to modulate wnt signaling (pmid:15772135)
2	FBgn0036559	0.85	9	18	Notum	notum	72C3-D1	involved in wnt signaling
2	FBgn0003117	0.85	9	17	pnr	panier	89A13-B1	transcription factor involved in cell fate determination
2	FBgn0013725	0.85	6	17	phyl	phyllopod	51A2	role in cell fate determination acts epistatically to Notch (pmid: 11526076)
2	FBgn0024195	0.87	9	18	lea	leak	22A1	functions in axon guidance interacts with kuzbanian, member of the Notch pathway (pmid: 11472832)
2	FBgn0024234	0.90	9	18	gbb	glass bottom boat	60A3-4	cell-cell signaling
2	FBgn0026319	0.90	7	18	Traf1	TNF receptor associated factor 1	24E1-4	Jnk cascade and Toll signaling pathway
2	FBgn0025739	0.95	9	18	pon	partner of numb	4C10	acts where Notch pathway does not regulate cell fate (pmid: 9857191)
2	FBgn0021874	0.85	9	17	Nle	notchless	2100	

Table 9.2: Candidate genes from the class II a belonging to the Notch pathway Genes are classified in 2 categories (cat): (1) best candidate genes (2) candidate genes. The F Id is the Flybase identifier. Loc is the physical map location of the gene.

9.3 Summary of this study

Our approach focused on consistent changes in transcript levels in order to identify genes which may be subject to tight regulation. The biological signals investigated which are shared among expression groups, and also the signals which distinguish groups from each other both lend support to our analysis strategy. Although these signals come at the cost of low sensitivity (groups are likely to have more members than this study can identify), the expression categories and groups seem to be a starting point for exploring different functional features. For instance, we note that the class II genes are likely to play vital roles in development, based on the behavior of their transcripts. Overall, we find a strong consistency between the global clustering method which is conceptually based on time-dependent data, and various sources of purely biological information. This further underlines the advantages of time series arrays as opposed to non-temporal ones.

Taken together, our initial analysis of embryonic gene expression in *Drosophila melanogaster* confirmed a number of known expression patterns but also revealed a surprisingly low number of well-defined expression groups that are sharply activated and suppressed together.

Part III

Conclusion and perspectives

Molecular biology is a very active field, where dogma tend not to live long due to the high complexity of life. As any experiment can only be interpreted with the current knowledge and hypotheses are made based on the current picture, history often takes time to prove them right or wrong. It is hence a challenge to collect as many evidences as possible to strengthen any theory. Recent years have seen the development of genome-wide studies, such as genome sequencing or microarray analyses, which brought along the expectation of a full understanding of molecular biology in a very near future. However, the closer we think we come to a precise understanding of biology, the more complicated it gets, and exceptions not rarely become the rule.

In the topics addressed during my PhD thesis, there are a handful of examples of this. Alternative splicing is one of them. First thought of as a really peculiar event, it is now clear that it generates a great deal of diversity in the transcripts, which is needed to account for the complexity of eukaroytes, as it turned out that the gene content of the genome can not explain it alone. But here again, not only has alternative splicing a role in generating diversity, but it can also be responsible for creating a temporal, a spatial, or a functional specificity. Moreover, acting on AS pathways is a very promising therapy approach. Antisense compounds are being developed for manipulation of alternative splicing patterns: oligonucleotides can be used to silence mutations that cause aberrant splicing, thus restoring correct splicing and function of the defective gene. In this way, the ratio of different splice variants can be altered, and the function of a gene changed. Considering the wide usage of alternative splicing, and its implication in numerous diseases, as for example cancer, cystic fibrosis or thalassemia, the possible range of applications of such technique is almost unlimited. Among the most promising applications of this technique is the correction of mis-splicing of the beta-globin gene caused by a point mutation creating a new splice-site and inserting a non-sense mutation in the transcript leading to the lack of beta-globin, and hence of oxygen-carrying capacity of red blood cells [84]. Another example is the play on the bcl-xL/bcl-xS transcript ratios (produced by alternative splicing). In summary, the long form, bcl-xL, has antiapoptotic properties, while the short form, bcl-xS, has proapoptotic properties. Both forms are required for the normal cellular function. Changing the ratio of those isoforms can induce cell death, which could target cancer cells or adipose tissue and reduce obesity. Such a fascinating topic is without doubt promised a great deal of attention in the future years. A new therapeutic area is born.

In addition to these practical issues involving AS, studies recently highlighted its involvement in evolution: by diminishing the selective pressure on genes, it allows them to evolve more freely. Indeed because new gene functions may arise from insertion or deletion of an exon, it was suggested that alternatively spliced exons (ASEs) can accelerate gene evolution. This topic, though, remains quite controversial with different studies leading to different results (see [85] for a recent study on evolutionary forces exerted on AS and constitutive exons). According to this study, different evolutionary forces act on AS and constitutive exons, such that ASEs show a increased rate of non-synonymous substitutions, suggesting that they contributed more to protein diversity than constitutive exons. This discussion will certainly keep going in the next years, as evolution is a topic for which no irrefutable proof is easily accessible.

Another example of challenged dogma is the relation gene-transcript-protein-function. It was indeed thought that a gene would be transcribed into one mRNA transcript, which then would be translated into one protein in order to conduct a function in the cell (or in the extracellular compartment). It seems however that the situation is not that trivial. In addition to the numerous roles already played at the level of RNA molecules, a number of regulation steps along the road intervene: genes can be transcribed in a cell without any obvious functional reason, but also without disturbing cellular functions (=neutral expression), or the transcript can be produced and quickly be silenced by NMD. As more regulation steps of gene expression are discovered, each of them seems not to need to be as systematic as previously thought. This neutral expression could be seen as a waste, but it is important to keep in mind that repressing is also costly for the cell. Hence a balance has to be found between the regulation of every single gene thanks to multiple combinations of activators and repressors (not very likely as it would require a huge number of regulatory genes), and the allowance of leaky expression (which is also partly due to structural constraints of the chromatin).

The molecular biology technique that revolutionized the past few year is without doubt the microarray technology. Its field of application is constantly growing, and almost seems to be unlimited. The first generation of it was able to investigate some hundreds of transcripts in a particular condition compared to a control. They then got able to do it genome wide, and as the technology became more easily available, virtually any condition could be tested. Recent developments allowed to follow particular events (cell cycle [81], development [69, 70]) during time, and reveal general patterns of expression. As alternative splicing's importance was revealed, special microarrays were designed to specifically investigate it. But even this seems not to be enough when one wants to investigate genuine functionality of the gene-coding products. Protein arrays are nowadays still under development to fill up this gap and the scientific community is impatiently waiting for them to reveal yet more of this complexity that fascinates us.

Bibliography

[1] A.P. Ryle, F. Sanger, L.F. Smith, and R. Kitai. The disulphide bonds of insulin. *Biochemical journal*, 60:541–556, 1955.

[2] Margareth Dayhoff, Eck, Chang, and Sochard. *Atlas of protein sequence and structure*. Black cover, 1965.

[3] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *JMB*, 48(3):443–453, 1970.

[4] W. Frawley, G. Piatetsky-Shapiro, and C. Matheus. Knowledge discovery in databases: An overview. *AI Magazine*, pages 213–228, 1992.

[5] L.J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature reviews genetics*, 7:119–129, 2006.

[6] L.D. Stein. Human genome: end of the beginning. *Nature*, 431:915–916, 2004.

[7] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[8] F. Wold. In vivo chemical modification of proteins (post-translational modification). *Annual Review of Biochemistry*, 50:783–814, 1981.

[9] S.I. Rattan, A. Derventzi, and B.F. Clark. Protein synthesis, posttranslational modifications, and aging. *Ann N Y Acad Sci.*, 663:48–62, 1992.

[10] H.A. Doyle and M.J. Mamula. Post-translational protein modifications in antigen recognition and autoimmunity. *Trends in Immunology*, 22, 2001.

[11] J.T. Kadonaga. Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell*, 92:307–313, 1998.

[12] C.F. Fry and C.L. Peterson. Unlocking the gates to gene expression. *Science*, 295:1847–1848, 2002.

[13] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell.* Garland Publishing, 4th edition, 2002.

[14] Horton, Moran, Ochs, Rawn, and Scrimgeour. *Principles of biochemistry.* Prentice Hall, 2002.

[15] S. Hahn. Structure and mechanism of the RNA polymerase II transcription machinery. *Nature structural and molecular biology*, 11(5):394–403, 2004.

[16] Y. Hirose and J.L. Manley. RNA polymerase II and the integration of nuclear events. *Genes and development*, 14:1415–1429, 2000.

[17] S.T. Smale and J.T. Kadonaga. The RNA polymerase II core promoter. *Annual review of biochemistry*, 72:449–479, 2003.

[18] F.J. Herrera and S.J. Triezenberg. What ubiquitin can do for transcription. *Current biology*, 14:R622–R624, 2004.

[19] T. Kurosu and B.M. Peterlin. VP16 and ubiquitin: binding of P-TEFb via its activation domain and ubiquitin facilitates elongation of transcription of target genes. *Current biology*, 14:1112–1116, 2004.

[20] L. Anderson and J. Seilhamer. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, 18:533–537, 1997.

[21] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology*, 4:117, 2003.

[22] F. Rodriguez-Trelles, R. Tarrio, and F.J. Ayala. Is ectopic expression caused by deregulatory mutations or due to gene-regulation leaks with evolutionary potential? *BioEssays*, 27(6):592–601, 2005.

[23] F. Crick. The biological replication of macromolecules. In *Symp. Soc. Exp. Biol.*, 1958.

[24] Francis Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[25] S.M. Berget, C. Moore, and P.A. Shart. Spliced segments at the 5'terminus of adenovirus 2 late mRNA. *PNAS*, 74:3171–3175, 1977.

[26] L.T. Chow, R.E. Gelinas, T.R. Broker, and R.J. Roberts. An amazing sequence arrangement at the 5'ends of adenovirus 2 messenger RNA. *Cell*, 12:1–8, 1977.

[27] M.S. Jurica and M.J. Moore. Pre-mRNA splicing: awash in a sea of proteins. *Molecular cell*, 12:5–14, 2003.

[28] R. Reed. Coupling transcription, splicing and mRNA export. *Current opition in cell biology*, 15:326–331, 2003.

[29] R. Breathnach, N. Mantei, and P. Chambon. Correct splicing of a chicken ovalbumin gene transcript in mouse L cells. *Proc.Natl.Acad.Sci.USA*, 77(2):740–744, 1980.

[30] F.M. DeNoto, D.D. Moore, and H.M. Goodman. Human growth hormone DNA sequence and mRNA structure: possible alternative splicing. *Nucleic acids research*, 9(15):3719–3730, 1981.

[31] E.C. Mariman, R.J. van Beek-Reinders, and W.J. van Venrooij. Alternative splicing pathways exist in the formation of adenoviral late messenger RNAs. *Journal of molecular biology*, 163(2):239–256, 1983.

[32] C.R. King and J. Piatigorsky. Alternative RNA splicing of the murine alpha A-crystallin gene: protein-coding information within an intron. *Cell*, 32(3):707–712, 1983.

[33] D. Kampa, J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekkiranov, G. Helt, H. Tammana, and T.R. Gingeras. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome research*, 14:331–342, 2004.

[34] S. Boue, M. Vingron, E. Kriventseva, and I. Koch. Theoretical analysis of alternative splicing forms using computational methods. *Bioinformatics*, Suppl.2:S65–73, 2002.

[35] S. Stamm. Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Human molecular genetics*, 11(20):2409–2416, 2002.

[36] N.A. Faustino and T.A. Cooper. Pre-mRNA splicing and human disease. *Genes and development*, 17(4):419–437, 2003.

[37] M.A. Garcia-Blanco, A.P. Baraniak, and E.L. Lasda. Alternative splicing in disease and therapy. *Nature biotechnology*, 22(331-342), 2004.

[38] P. Forch and J. Valcarcel. Splicing regulation in *Drosophila* sex determination. *Prog Mol Subcell Biol*, 2003.

[39] D.L. Black. Mechanisms of alternative pre-messenger RNA splicing. *Annu.Rev.Biochem.*, 72(291-336), 2003.

[40] I. Letunic, R.R. Copley, and P. Bork. Common exon duplication in animals and its role in alternative splicing. *Human molecular genetics*, 11(13):1561–1567, 2002.

[41] C. Lee and Q. Wang. Bioinformatics analysis of alternative splicing. *Briefings in bioinformatics*, 6(1):23–33, 2005.

[42] J. Hanke, D. Brett, I. Zastrow, A. Aydin, S. Delbruck, G. Lehmann, F. Luft, J. Reich, and P. Bork. Alternative splicing of human genes: more the rule than the exception? *Trends in genetics*, 15(10):389–390, 1999.

[43] A.A. Mironov, J.W. Fickett, and M.S. Gelfand. Frequent alternative splicing of human genes. *Genome research*, 9(12):1288–1293, 1999.

[44] S. Gupta, D. Zink, B. Korn, M. Vingron, and S.A. Haas. Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC genomics*, 5(1):72, 2004.

[45] Z. Kan, J. Castle, J.M. Johnson, and N.F. Tsinoremas. Detection of novel splice forms in human and mouse using cross-species approach. *Proceedings of the 9th Pacific symposium on biocomputing*, pages 42–53, 2004.

[46] R. Sorek, R. Shemesh, Y. Cohen, O. Basechess, G. Ast, and R. Shamir. A non-EST-based method for exon-skipping prediction. *Genome research*, 14(8):1617–1623, 2004.

[47] D.L. Philipps, J.W. Park, and B.R. Graveley. A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA*, 10(12):1838–1844, 2004.

[48] J.M. Johnson, J. Castle, P. Garrett-Engele, and *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarray. *Science*, 302:2141–2144, 2003.

[49] P. Fehlbaum, C. Guihal, L. Bracco, and O. Cochet. A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic acids research*, 33(5):e47, 2005.

[50] M.W. Hentze and A.E. Kulozik. A perfect message: RNA surveillance and nonsense-mediated decay. *Cell*, 96(3):307–310, 1999.

[51] M.J. Moore. RNA events. No end to nonsense. *Science*, 298(5592):370–371, 2002.

[52] P.A. Frischmeyer, A. van Hoof, K. O'Donnell, A.L. Guerrerio, R. Parker, and H.C. Dietz. An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science*, 295:2258–2261, 2002.

[53] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.

[54] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[55] Almut Schulze and Julian Downward. Navigating gene expression using microarrays - a technology review. *NCB*, 3:E190–E195, 2001.

[56] Y.H. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nature reviews genetics*, 3:579–588, 2002.

[57] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, f.C.P. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nature Genetics*, 29:365–371, 2001.

[58] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci*, 95:14863–14868, 1998.

[59] K.Y. Yeung and W.L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[60] M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, S.E. Lewis, S. Richards, M. Ashburner, S.N. Henderson, G.G. Sutton, J.R. Wortman, M.D. Yandell, Q. Zhang, L.X. Chen, R.C. Brandon, Y.C.

Rogers, R.G. Blazej, M. Champe, B.D. Pfeiffer, K.H. Wan, C. Doyle, E.G. Baxter, G. Helt, C.R. Nelson, G.L. Gabor Miklos, J.F. Abril, A. Agbayani, H-J. An, C. Andrews-Pfannkoch, D. Baldwin, R.M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E.M. Beasley, K.Y. Beeson, P.V. Benos, B.P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M.R. Botchan, P. Bouck, J.and Brokstein, P. Brottier, K.C. Burtis, D.A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J.M. Cherry, S. Cawley, C. Dahlke, L.B. Davenport, P. Davies, B. Pablos, A. Delcher, Z. Deng, A.D. Mays, I. Dew, S.M. Dietz, K. Dodson, L.E. Doup, M. Downes, S. Dugan-Rocha, B.C. Dunkov, P. Dunn, K.J. Durbin, C.C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A.E. Gabrielian, N.S. Garg, W.M. Gelbart, K. Glasser, A. Glodek, F. Gong, J.H. Gorrell, Z. Gu, P. Guan, M. Harris, N.L. Harris, D. Harvey, T.J. Heiman, J.R. Hernandez, J. Houck, D. Hostin, K.A. Houston, T.J. Howland, M-H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G.H. Karpen, Z. Ke, J.A. Kennison, K.A. Ketchum, B.E. Kimmel, C.D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A.A. Levitsky, J. Li, Z. Li, X. Liang, Y.and Lin, X. Liu, B. Mattei, T.C. McIntosh, M.P. McLeod, D. McPherson, G. Merkulov, N.V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S.M. Mount, M. Moy, B. Murphy, L. Murphy, D.M. Muzny, D.L. Nelson, D.R. Nelson, K.A. Nelson, K. Nixon, D.R. Nusskern, J.M. Pacleb, M. Palazzolo, G.S. Pittman, S. Pan, J. Pollard, V. Puri, M.G. Reese, K. Reinert, K. Remington, R.D.C. Saunders, F. Scheeler, H. Shen, B.C. Shue, I. Sidén-Kiamos, M. Simpson, M.P. Skupski, T. Smith, E. Spier, A.C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A.H. Wang, X. Wang, Z-Y. Wang, D.A. Wassarman, G.M. Weinstock, J. Weissenbach, S.M. Williams, T. Woodage, K.C. Worley, D. Wu, S. Yang, Q.A Yao, J. Ye, R-F. Yeh, J.S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X.H. Zheng, F.N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H.O. Smith, R.A. Gibbs, E.W. Myers, G.M. Rubin, and J.C. Venter. The Genome Sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–2195, 2000.

[61] O.L. Mohr. Character changes caused by mutation of an entire region of a chromosome in *Drosophila*. *Genetics*, 4:275–282, 1919.

[62] S. Artavanis-Tsakonas, M.D. Rand, and R.J. Lake. Notch signaling: cell fate control and signal integration in development. *Science*, 284:770–776, 1999.

[63] C.W. Smith. Alternative splicing - when two is a crowd. *Cell*, 123(1):1–3, 2005.

[64] D. Brett, H. Pospisil, J. Valcarcel, J. Reich, and P. Bork. Alternative splicing and genome complexity. *Nature genetics*, 30(1):29–30, 2002.

[65] H. Kim, R. Klein, J. Majewski, and J. Ott. Estimating rates of alternative splicing in mammals and invertebrates. *Nature genetics*, 36:915–916, 2004.

[66] B. Ewing and P. Green. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature genetics*, 25(2):232–234, 2000.

[67] E. Harrington, S. Boue, J. Valcarcel, J.G. Reich, and P. Bork. Estimating rates of alternative splicing in mammals and invertebrates. *Nature genetics*, 36(916-917), 2004.

[68] A.I. Su, T. Witshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, and J.B. Hogenesh. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc.Natl.Acad.Sci.USA*, 101(16):6062–6067, 2004.

[69] L.R. Baugh, A.A. Hill, D.K. Slonim, E.L. Brown, and C.P. Hunter. Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development*, 130(5):889–900, 2003.

[70] M.N. Arbeitman, E.E.M. Furlong, F. Imam, E. Johnson, B.H. Null, B.S. Baker, M.A. Krasnow, M.P. Scott, R.W. Davis, and K.P. White. Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, 297:2270–2275, 2002.

[71] V. Stolc, Z. Gauhar, C. Mason, G. Halasz, M.F. van Batenburg, S.A. Rifkin, S. Hua, T. Herreman, W. Tongprasit, P.E. Barbano, H.J. Bussemaker, and K.P. White. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science*, 306:655–660, 2004.

[72] R.A. Drysdale, M.A. Crosby, and Flybase Consortium. FlyBase: genes and gene models. *Nucleic acids research*, 33(Database issue):D390–395, 2005.

[73] T.R. Li and K.P. White. Tissue-specific gene expression and ecdysone-regulated genomic networks in *Drosophila*. *Developmental cell*, 5(1):79–92, 2003.

[74] C. Workman, L.J. Jensen, H. Jarmer, R. Berka, L. Gautier, H.B. Bielser, H.H. Saxild, C. Nielsen, S. Brunak, and S. Knudsen. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome biology*, 3(9):research0048, 2002.

[75] R. Sasik, E. Calvo, and J. Corbeil. Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics*, 18(12):1633–1640, 2002.

[76] C. von Mering, L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen, and P. Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research*, 33(Database issue):D433–437, 2005.

[77] N. Bluthgen, S.M. Kielbasa, and H. Herzel. Inferring combinatorial regulation of transcription *in silico*. *Nucleic acids research*, 33(1):272–279, 2005.

[78] M. Rechsteiner and S.W. Rogers. PEST sequences and regulation by proteolysis. *Trends in biochemical sciences*, 21(7):267–271, 1996.

[79] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S.E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S.E. Celniker, and G.M. Rubin. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome biology*, 3(12):RESEARCH0088, 2002.

[80] T. Brody. The interactive fly: gene networks, development and the internet. *Trends in genetics*.

[81] U. De Lichtenberg, L.J. Jensen, S. Brunak, and P. Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727, 2005.

[82] S. Luschnig, B. Moussian, J. Krauss, I. Desjeux, J. Perkovic, and C. Nusslein-Volhard. An F1 genetic screen for maternal-effect mutations affecting embryonic pattern formation in *Drosophila melanogaster*. *Genetics*, 167(1):325–342, 2004.

[83] T.O. Alexson, S. Hitoshi, B.L. Coles, A. Bernstein, and D. van der Kooy. Notch signaling is required to maintain all neural stem cell populations–irrespective of spatial or temporal niche. *Developmental neuroscience*, 28(1-2):34–48, 2006.

[84] P. Sazani and R. Kole. Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing. *J. Clin. Invest.*, 112:481–486, 2003.

[85] F.C. Chen, S.S. Wang, C.J. Chen, W.H. Li, and T.J. Chuang. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Molecular biology and evolution*, 23(3):675–682, 2006.

Glossary

Accession number A unique number or code given to mark the entry of a sequence or pattern to a database.

Algorithm The logical sequence of steps by which a task can be performed.

Allele One of two or more alternative forms of a gene at a particular locus.

Alternative splicing Different ways of combining a gene's exons to make variants of the complete protein.

Amino acid The fundamental building block of proteins. There are 20 naturally occurring amino acids in animals.

Analogs Non-homologous proteins that have similar folding architecture, or similar functional sites, which are believed to have arisen through convergent evolution.

Apoptosis Programmed cell death, is a series of programmed steps that cause a cell to die via "self digestion" without rupturing and releasing intracellular contents into the surrounding environment.

Assembly The process of aligning overlapping sequence fragments into a contig or a series of contigs.

ATP Adenosine triphosphate is the primary molecule for storing chemical energy in the cell.

Basepair (bp) Any possible pairing between bases in opposing strands of DNA or RNA. Adenine pairs with thymine in DNA, or with uracil in RNA; and guanine pairs with cytosine.

BLAST Basic Local Alignment Search Tool is a computer program widely used to search large databases of DNA or amino acid sequences, providing sequences that have regions of similarity to the sequence of interest provided by the user.

Blastoderm the layer of cells in an insect embryo that surrounds an internal yolk mass. The cellular blastoderm develops from a syncytium by surrounding the cleavage nuclei with membranes derived from the enfolding of the surrounding membrane.

Capping Attachment of a modified guanine (G) to the 5' end of the pre-mRNA as it emerges from the RNA polymerase II. The cap protects the RNA from being degraded by enzymes that degrade RNA from the 5' end.

cDNA The DNA copy of a eukaryotic messenger RNA molecule, produced *in vitro* by enzymatic synthesis and used for producing cDNA libraries or probes for isolating genes in genomic libraries.

cDNA library A gene library composed of cDNA inserts synthesized from mRNA using reverse transcriptase.

Cell The fundamental unit of life; cells may be organized in organs that are relatively autonomous but cooperate in the functioning of the organism.

Cell cycle The sequence of events between one cell mitotic division and another in a eukaryotic cell. Mitosis (M phase) is followed by a growth (G1) phase, then by DNA synthesis (S phase), then by another growth (G2) phase, and then by another mitosis.

Central dogma A fundamental principal of molecular biology essentially stating that the transfer of information from nucleic acid to nucleic acid, and from nucleic acid to protein is possible, while transfer from protein to nucleic acid or from protein to protein is impossible.

Chromosomes The paired, self replicating genetic structures of cells that contain the cellular DNA.

Clone A population of identical cells often containing identical recombinant DNA molecules.

Cluster analysis A method of hierarchically grouping taxa or sequences on the basis of the similarity or minimum distance.

Coding sequence (CDS) A region of DNA or RNA whose sequence determines the sequence of amino acids in a protein.

Complementary DNA (cDNA) DNA that is synthesized from a messenger RNA template using the enzyme reverse transcriptase.

Conserved sequence A sequence of bases in a DNA molecule (or an amino acid sequence in a protein) that has remained essentially unchanged during evolution.

Convergent evolution The independent development of similar (analogous) structures in different groups; convergent evolution is thought to be the result of similar environmental selection pressures on different groups.

Dendrogram A branched diagram that represents the evolutionary history of a group of organisms.

Discontinuous gene A gene in which the genetic information is separated into two or more different exons by an intervening sequence (intron) which typically is noncoding. Most eukaryotic genes are discontinuous.

Divergent evolution Divergent evolution is the process by which initially similar gene copies diverge to perform different functions as they lose the selective pressure initially present on them. New selective pressures then take over to evolve the genes towards totally different functions. Note that these gene products may have common structural aspects and may involve the same step mechanistically. However, the reaction involved will be widely different.

DNA (deoxyribonucleic acid) The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bond between basepairs of nucleotides. The four nucleotides in DNA contain the bases: adenine (A), guanine (G), cytosine (C) and thymine (T). In nature, basepairs form only between A and T and between C and G; this the base sequence of each single strand can be deduced from that of its partner.

DNA sequence The linear sequence of base pairs, whether in a fragment of DNA, a gene, a chromosome or an entire genome.

Ectopic expression The occurrence of gene expression in a tissue in which it is normally not expressed.

Enhancer Sequences of DNA that can increase transcription of neighboring genes over long distances up or downstream of the gene and in either possible orientation.. A gene epistatic to another masks the expression of the second gene.

Epistatic Epistatis id the nonreciprocal interaction of nonallelic genes.

Euchromatin Regions of eukaryotic chromosome that appear less condensed and stain less well with DNA-specific dyes than other segments of the chromosome.

Euclidean distance Distance between objects or values that is computed as a straight line.

eukaryote An organism with cells containing a membrane-bound nucleus that reproduces by meiosis. Cells divide by mitosis. Oxidative enzymes are packaged within mitochondria.

Exon One of the coding regions of a discontinuous gene.

Gene A segment of DNA that codes for a RNA and/or a polypeptide molecule. It includes regions preceding and following the coding region, as well as introns.

Genome The total complement of DNA in an organism.

Histones Basic proteins that make up nucleosomes and have a fundamental role in chromosome structure. of different sequence that code for the same gene product.

Homologous genes Two genes from different organisms and therefore

Homology Being related by the evolutionary process of divergence from a common ancestor. Homology is not a synonym for similarity.

Housekeeping genes Genes whose products are required by the cell for normal maintenance.

Hybridisation The process of joining two complementary strands of DNA or one each of DNA and RNA to form a double-stranded molecule.

***in situ* hybridization** The pairing of complementary DNA and RNA strands, or the pairing of complementary DNA single strands to produce a hybrid molecule in intact chromosomes or cells. Pairing is detected by some form of label.

Introns The sequence of DNA bases that interrupts the protein-coding sequence of a gene; these sequences are transcribed into RNA but are edited out of the message before it is translated into protein.

Lethal mutation Mutation of a gene to yield no product, or a defective gene product, resulting in the death of the organism because the gene product is essential to life.

Library An unordered collection of clones (i.e., cloned DNA from a particular organism), generated from genomic DNA or cDNA.

Messenger RNA or mRNA RNA molecules which code for proteins and which are translated on the ribosomes.

Microarray A tool for analyzing gene expression that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern.

Model system A biological system used to represent other, often more complex, systems in which similar phenomena either do, or are thought to, occur (e.g.,*D.melanogaster, M.musculus, S.cerevisiae, C.elegans, E.coli*).

Normalized library cDNA library generated such that all the genes in the library are represented at the same frequency.

Nucleosome A basic structure by which eukaryotic chromosomes are organized and compacted. Nucleosomes comprise an octamer of histone proteins with DNA coiled around them and are connected to other nucleosomes by linker DNA.

Nucleotide A molecule consisting of a nitrogenous base (A, G, T or C in DNA; A, G, U or C in RNA), a phosphate moiety and a sugar group (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form a DNA or RNA molecule.

Oligonucleotide or oligo A short fragment of a single-stranded DNA that is typically 5 to 50 nucleotides long.

Open reading frame (ORF) A series of DNA codons, including a 5' initiation codon and a termination codon, that encodes a putative or known gene.

orthologs Homologous proteins that perform the same function in different species.

Paralogues Homologous proteins that perform different but related functions within one organism.

Phenotype The observable characteristics of an organism that are determined by both genotype and environment.

Phylogenetic tree A graphical representation of the putative evolutionary relationships between groups of organisms, e.g. as calculated from multiple protein or nucleic acid sequence alignments.

Polyadelylation Addition of a stretch of adenine (A) nucleotides. When a special poly(A) attachment site is found in the pre-mRNA that emerges from the RNA polymerase II, the transcript is cut there, and a poly(A) tail is attached to the exposed 3' end. This completes the mRNA molecule, which is now ready for export to the cytosol.

Polymerase chain reaction (PCR) A method for amplifying a DNA base sequence using a heat-stable polymerase and two primers, one complementary to the (+)-strand at one end of the sequence to be amplified and the other complementary to the (-)-strand at the other end. The faithfulness of reproduction of the sequence is related to the fidelity of the polymerase. Errors may be introduced into the sequence using this method of amplification.

Pre-mRNA The unprocessed transcript of a protein-coding gene.

Primary structure The linear sequence of amino acids in a protein molecule.

Principle Components A set of variables that define a projection that encapsulates the maximum amount of variation in a dataset and is orthogonal (and therefore uncorrelated) to the previous principle component of the same dataset.

Promoter A site on DNA to which RNA polymerase will bind and initiate transcription.

Protein A molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene coding for the protein. Proteins are required for the structure, function and regulation of cells, tissues and organs, each protein having a specific role.for destruction by the addition of ubiquitin.

Proteome The protein complement of a cell.

Proteosome A large protein complex in the cytoplasm of eukaryotic cells that contains proteolytic enzymes. Proteosomes break down proteins that have been tagged

Ribosomal RNA or rRNA The RNA that acts as a structural component of ribosomes.

Ribosome A self-assembling cellular organelle made up of proteins and RNA in which translation of mRNA occurs. Ribosomes consist of two subunits, each composed of RNA and proteins.

RNA (Ribonucleic acid) A molecule chemically similar to DNA that plays a central role in protein synthesis. The structure of RNA is similar to that of DNA but is inherently less stable. There are several classes of RNA molecule, including messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and other small RNAs, each serving a different purpose.

RNA editing RNA editing involved altering the mRNA after transcription. This results in different proteins being produced from a single gene. The molecular mechanisms include single or multiple base insertions or deletions, as well as base substitutions.

RNA polymerase An enzyme capable of synthesizing a RNA copy of a DNA template.

RNA surveillance A system in eukaryotic cells to degrade aberrant mRNAs.

Sequencing Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule, or the order of amino acids.

Shotgun method Cloning of DNA fragments randomly generated from a genome.

Spliceosome The RNA and protein particules in the nucleus that remove introns from pre-messenger RNA molecules.

Splicing The process by which introns, non-coding regions, are excised out of the primary messenger RNA transcript and exons (i.e., coding regions) are joined together to generate mature messenger RNA.

Stop codon One of the three mRNA codons (UAG, UAA, and UGA) that prevent further polypeptide synthesis.

Syncytium A mass of protoplasm containing many nuclei not separated by cell membranes.

Synteny Synteny refers to the fact that many genes remain groped together in the same relative positions in the genome across taxa.

Tertiary structure The overall fold of a protein sequence, formed by the packing of its secondary structure elements

Transcription The synthesis of a RNA copy from a sequence of DNA (a gene); the first step in gene expression.

Transcriptome The transcriptome is the profile of the genes that are expressed or transcribed from genomic DNA within a cell or tissue, with the goal of understanding cell phenotype and function. The transcriptome is dynamic and changes rapidly in response to stress or during normal cell processes.

Transfer RNA or tRNA A family of small RNA molecules that serve as adapters for bringing amino acids to the site of protein synthesis on the ribosome.

Translation The process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids.

True-negative A false match that correctly fails to be recognized by a discriminator.

True-positive A true match correctly recognized by a discriminator.

Ubiquitination The attachment of the protein ubiquitin to lysine residues of other molecules, often as a tag for their rapid cellular degradation.

Upstream Further back in the sequence of a DNA molecule, with respect to the direction in which the sequence is being read.