KTH Computer Science
and Communication

Taking advantage of phylogenetic trees in comparative genomics

ÖRJAN ÅKERBORG

Doctoral Thesis
Stockholm, Sweden 2008

TRITA CSC-A 2008:09
ISSN 1653-5723
ISRN KTH/CSC/A–08/09–SE
ISBN 978-91-7178-987-7

KTH School of Computer Science
and Communication
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i datalogi onsdagen den 4 juni 2008 klockan 09.30 i FD05, Albanova, Roslagstullsbacken 21, Stockholm.

Tryck: Universitetsservice US AB

Abstract

Phylogenomics can be regarded as evolution and genomics in co-operation. Various kinds of evolutionary studies, gene family analysis among them, demand access to genome-scale datasets. But it is also clear that many genomics studies, such as assignment of gene function, are much improved by evolutionary analysis. The work leading to this thesis is a contribution to the phylogenomics field. We have used phylogenetic relationships between species in genome-scale searches for two intriguing genomic features, namely *functional pseudogenes* and *A-to-I RNA editing*. In the first case we used pairwise species comparisons, specifically human-mouse and human-chimpanzee, to infer existence of functional mammalian pseudogenes. In the second case we profited upon later years' rapid growth of the number of sequenced genomes, and used 17-species multiple sequence alignments. In both these studies we have used non-genomic data, gene expression data and synteny relations among these, to verify predictions. In the A-to-I editing project we used 454 sequencing for experimental verification.

We have further contributed a maximum *a posteriori* (MAP) method for fast and accurate dating analysis of speciations and other evolutionary events. This work follows recent years' trend of leaving the strict molecular clock when performing phylogenetic inference. We discretised the time interval from the leaves to the root in the tree, and used a dynamic programming (DP) algorithm to optimally factorise branch lengths into substitution rates and divergence times. We analysed two biological datasets and compared our results with recent MCMC-based methodologies. The dating point estimates that our method delivers were found to be of high quality while the gain in speed was dramatic.

Finally we applied the DP strategy in a new setting. This time we used a grid laid out on a species tree instead of on an interval. The discretisation gives together with speciation times a common timeframe for a gene tree and the corresponding species tree. This is the key to integration of the sequence evolution process and the gene evolution process. Out of several potential application areas we chose gene tree reconstruction. We performed genome-wide analysis of yeast gene families and found that our methodology performs very well.

Contents

Acknowledgments

I started the work leading to this thesis with the intention of "taking a few years career brake to do some interesting research". The result has been just that but also very very hard work, more than a few stressful periods and quite a lot of hesitation about whether this was at all a good idea. Therefore some of the following are not only mere thanks for help and a good time but in fact acknowledgments of people without whom I would have quit a long time ago.

First and foremost there is my supervisor Jens. Everyone who has had the opportunity of working with Jens can certify how brilliant he is and how inspiring he can be. It is not often that you meet people whose mind is so bright and clear that it affects your entire thinking. Jens is such a person. I took me a while to learn how to communicate efficiently with him, but after I learned how it should be done it has been a delight to speak with him about all sorts of things.

The second most important person has been my dad, Per Svensson. This may seem unusual to some but then again he is an unusual person. He is not only the best father you could hope for, he is also one of my best friends, very good at correcting faulty English, knowledgeable about all sorts of science and always eager to learn more. He is the only person who really knows both me and my work.

Warm thanks also to those with whom I have collaborated, Lars Arvestad, Bengt Sennblad, Mats Ensterö, and Marie Öhman, to those I have shared an office with, Johannes Frey-Skött, Jesper Lundström, Lena Milchert, Samuel Andersson, and Ali Tofigh, to those who has meant inspiration, Arne Elofsson, Gunnar von Heijne, Per Kraulis, Erik Sonnhammer. Also warm thanks to doctorand colleagues who has shown how disputations should be done, how Pärk and Texas hold'em should not be played, and similar important things.

A special thank you to those who have helped a "computer idiot" into becoming a doctor of computer science. These are Ali, Bengt, Lars and Johannes who are all already mentioned, but also system administrator Erik Sjölund and last but not least all the people at PDC - Parallelldatorcentrum.

Finally there are all the others. A substantial number of great friends who have supported me by asking "when will you ever be finished?", "should you not get a job again, sometime?" and similar things. I would not have wanted them any other way. And my two families: Kerstin, Sigrid, and Ragni who mean everything to me, and Susanna and Joel who mean even more. I cannot say what I would have done without you since it is impossible to imagine such a situation. A special acknowledgment to Joel whose nine pre-birth and first two post-birth months coincided with me finalising this work. I hope he did not take any damage from a stressed-up dad while developing.

Publications included in this thesis

- **Paper I: Genome-wide survey for biologically functional pseudogenes**
  Örjan Svensson, Lars Arvestad, and Jens Lagergren
  PLoS Computational Biology 2006

- **Paper II: Birth-death prior on phylogeny and speed dating**
  Örjan Åkerborg, Bengt Sennblad, and Jens Lagergren
  BMC Evolutionary Biology 2008

- **Paper III: A computational screen for site selective A-to-I editing**
  Mats Ensterö, Örjan Åkerborg, Daniel Lundin, Bei Wang, Terrence S. Furey, Marie Öhman, and Jens Lagergren

- **Paper IV: Gene tree analysis reaching maturity**
  Örjan Åkerborg, Bengt Sennblad, Lars Arvestad, and Jens Lagergren

Chapter 1

Foreword and introduction

This thesis summarises my time as PhD student at the School of Computer Science and Communication (CSC) and Stockholm Bioinformatics Center (SBC). Already my joint affiliation suggests that this text will have an interdisciplinary character. It concerns computational biology, which in essence is use of computer science and mathematics in addressing problems in (molecular) biology. When I started my PhD studies, my biological training was limited to the very basic courses that were compulsory in Swedish upper secondary school. And although my level of knowledge in biology is higher now, my work scarcely qualifies me calling myself a biologist. A typical research week at the office has included three days of design and implementation of computer programs, one day of reading and writing, a few hours of mathematics and only very little thinking on biological issues. But even so, none of my many computer programs would have had any meaning without the underlying biology. The addressed biological phenomena will therefore be the main focus in this text and I believe that an understanding of these is the most important factor for appreciating the content. I have for this reason included an appendix giving a brief account of some basic cell biology. The reader who would like a more thorough exposition is referred to [4].

In this first chapter I will summarise the results of the presented work and, very briefly, put these results into context. In chapter 2 is described in some detail the research area in which I and my colleagues work and the biological questions we have addressed. Chapter 3 describes the computational tools that we have used and how we have applied and extended them in order to obtain the biological results. Chapter 3 has the structure of the computational biology discipline itself in that it contains a mixture of mathematics, computer science, probability theory, and data mining. Chapter 4 has four subsections, each one introducing one of the four papers that are the main contributions of this work. In chapter 5 there is room for conclusions and reflections.

Introduction

During evolution the survival of organisms has largely been dependent on their ability to adapt to changes in the environment. Mutations which have signified advantageous innovations for an individual have spread and become fixed in the population. Such innovations have enabled genes to accomplish their tasks more efficiently, or caused existing genes to take on new functions, or even led to the appearance of entirely new genes. A common hypothesis has therefore been that the complexity of an organism should be reflected in its number of genes.

Now, when the genome sequence of human and several other species are known, it is evident that this gene-count – complexity hypothesis has been largely incorrect. Table 1.1 shows a comparison between human and some well-studied model organisms. It is clear that the apparent complexity differences between for example the roundworm *Caenorhabditis elegans* and ourselves is not reflected in the gene count. So the obvious question is then – from where does the complexity difference arise?

Species	Cells	Neurons	Nucleotides	Genes
E. coli	1	0	$4.6 \cdot 10^6$	4,377
Rice	10^7	0	$4 \cdot 10^8$	37,500
C. elegans	1,000	302	10^8	19,400
Human	10^{14}	10^{11}	$3 \cdot 10^9$	22,000

Table 1.1: The table illustrates the gene count paradox, i.e., that the complexity of an organism is not by necessity reflected in its number of genes. The data on rice and human are rough estimates.

The natural way to attack this question is to consider differences between “simpler” and more “complex” groups of species. But it is important to realise that while some issues may be resolved by comparing the very distant, e.g. human and bacteria, other issues may need comparison of the very close, e.g. human and chimpanzee.

For bacteria it seems that the cellular machinery is dominated by the Central dogma processes, i.e., the molecular biology of the prokaryote cell can be relatively well understood by considering only the protein coding genes. Also, the regulation of these genes is rather simple. Groups of genes, *regulons*, are controlled simultaneously from short regulatory regions and a majority of the genome is translated. The limited physical sizes and short generation times of these organisms have clearly favoured small and compact genomes.

By contrast there is an apparent waste of space in the human and other eukaryote genomes. It is now clear that only a single percent of the human genome is in fact translated, and ever since this was discovered the functionality of the remaining, non-coding, DNA has been much debated. Early on there were few plausible explanations and the term *junk DNA* was established. Accordingly, genome research has to a large extent been synonymous with studies of the coding genome and the proteins. When *GenBank*, i.e., the database holding all publicly available nucleotide sequences and their protein translations, was first designed in the 60's there were voices urging that only coding sequences should be sequenced and included (Gary Stormo, personal communication).

The junk DNA term is still in use but it is getting more and more unclear which parts of the genome actually deserve this labeling. The sequencing of the human and mouse genomes has revealed that around 5% of these genomes are evolutionarily conserved [70], i.e., five times more than the coding percentage. In fact, a recent study [12] has revealed the existence of 481 *ultraconserved* regions, i.e., regions $\geq$ 200 base-pairs long, exactly identical between human and mouse. The majority of these are found in regions devoid of protein coding genes. Thus, it is clear that there is a lot going on which is important to the organism but cannot be explained by invoking the Central dogma. As for the 95% non-conserved sequence, it is nowadays agreed that its significance is poorly understood and therefore an important research topic.

Huge recent interest in primate evolution, fueled by the sequencing of the chimpanzee and rhesus macaque genomes, has also given some clues on how complexity can be achieved. Human and chimpanzee are so evolutionarily close that even neutrally evolving orthologous sequences are 98.5% identical [2]. But the process of gene duplication and loss is separating the two species four times faster than does point mutations; 1,418 out of 22,000 genes, i.e., 6.5% are reported to be lineage specific [17]. That is, even though human-chimpanzee orthologous genes are similar, and even though the total gene count in the two species is similar, the difference in gene *composition* is large.

Our contribution

A fast growing number of sequenced genomes have in later years facilitated detection and detailed analysis of a number of genomic phenomena. Projects previously requiring huge laboratory resources can now to a large extent be performed in front of the computer screen. One bio-science discipline whose importance has increased greatly in this new era is *phylogenomics*, i.e., the integration of genome analysis and evolutionary studies. On one hand, genome analysis can help resolve relationships between species. On the other hand, evolutionary relationships can help putting comparative genomics results in perspective and give an understanding of how differences between genes and differences between genomes arose [20].

The work presented here can be seen as a phylogenomic contribution to the effort of further elucidating how organismal complexity can be genetically motivated. We have developed new phylogenetic methods as well as extended and adapted existing ones.

The four articles presented concern a diverse biological area. We have performed genome-scale searches for two intriguing genomic features, namely *functional pseudogenes* and *A-to-I RNA editing*. These are very different phenomena, but have in common that they challenge, or at least extend, the Central dogma. They were also both first observed relatively recently and can be expected to be more important for more complex organisms, such as ourselves. Whether these features in the future will be considered important contributors to complexity is difficult to say. But it should be noted that it was not many years ago that RNA was considered a passive intermediate between DNA and protein, while the belief of today is that the importance of RNA genes are well comparable to that of protein coding ones.

The pseudogene scan is introduced in section 4.1 and the corresponding article is included in this thesis as paper I. The A-to-I RNA editing paper is introduced in section 4.3 and included as paper III.

We have further considered the various processes governing gene and genome evolution. More specifically, we have modeled the *sequence evolution process* and the *gene evolution process*, and analysed sequence families with respect to these processes and the phylogenetic trees that relate them. Section 4.2 and paper II presents a novel method for inference of phylogenetic trees and divergence times in the maximum likelihood framework. Special consideration is taken of the effects of sequence evolution *rates* varying over lineages. Section 4.4 and paper IV concerns a method integrating sequence evolution on one hand and the duplication and loss process on the other.

Chapter 2

On genomic motivations for organismal complexity

2.1 Duplications

In the years around 1970 Susumu Ohno, by means of a series of bold and controversial statements, considerably popularised research on gene and genome evolution [64]. In the book *Evolution by Gene Duplication* [54] he postulated gene duplication as the single most important factor in evolution. Taking the argument even further he proposed that without duplicated genes, the evolution from unicellular organisms would have been impossible. The argument was that new genes is a necessary requirement for major evolutionary steps, such as the invention of metazoans, vertebrates, and mammals. But even 30 years prior to Ohno, similar opinions were raised, claiming that the difference between, for example, human and amoeba could not be explained by mutations of the same set of genes [28, 52].

Massive investigations during later years have largely given Ohno and his predecessors right, the primary evidence being the widespread existence of *gene families* [35]. A gene family can be defined as all genes descended from a single ancestral gene in their last common ancestral species [69]. Within a gene family we distinguish between *orthologs* and *paralogs*, that is, sets of genes descending from a speciation and a duplication respectively (see Figure 2.1). The size and composition of a family are results of speciations of the ancestral species and the rate with which genes are duplicated and lost. There are several obstacles related to the classification of genes into families. Even if conservative numbers are used it is clear that gene duplication is of great importance in species from bacteria to human. Of the 4,014 genes encoded in the genome of the bacterium *Bacillus subtilis*, 1,888 (47%) have one or more obvious paralogs (the biggest family consists of 77 genes) [4]. In human an estimated 70% of genes have paralogs and the mean gene family size has been estimated to be about 7 [71].

It is common that gene family composition changes over short evolutionary

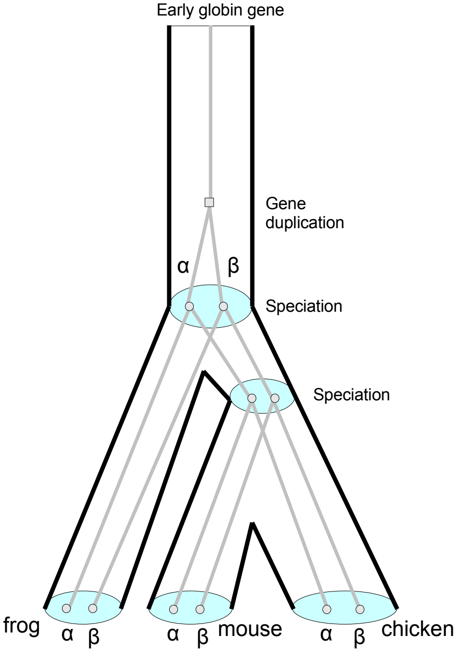


Figure 2.1: The figure shows the evolution of a globin gene family, containing the globin α and globin β genes from species frog, mouse, and chicken, respectively. The gene tree, drawn with grey thin edges, evolves within the species tree. The mouse α and mouse β globin genes are paralogs since they originate from a duplication (the one at the top of the tree). The mouse α and chicken α are, correspondingly, orthologs since they originate from the mouse-chicken species split.

distances. Families have expanded in one lineage but are contracted or completely erased in another. An example is the evolution of the olfactory and Opsin gene families since the ancestor of human and mouse. The former family is expanded in mouse, enabling its delicate sense of smell, while an Opsin duplication has enabled the human eye to distinguish three different wavelengths, vital for colour vision, while mice distinguish only two [35].

There are various mechanisms through which duplications can arise. Three of these, unequal cross-over, non-homologous duplication, and *retrotransposition* are illustrated in Figure 2.2. The first two are DNA-mediated processes resulting in sequences positioned in *tandem*, i.e., next to each other along the genome. Retrotransposition occurs when mRNA is re-inserted into the genome by way of the enzyme *reverse transcriptase*. Genes originating from retrotransposition are said to

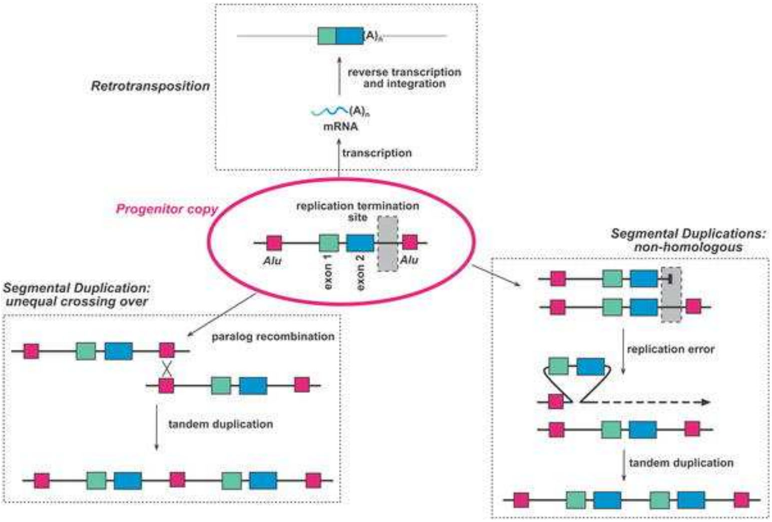


Figure 2.2: Three mechanisms of gene duplication, figure from [35]. (Top) An mRNA sequence is, after intron splicing, re-inserted in some genome position by reverse transcriptase. (Left) Recombination resulting in unequal cross-over is an important source of duplicated sequences. In [9] is shown that *Alu-elements* are enriched at duplication breakpoints, implicating that these repeat sequences can be important generators of duplications in primates. (Right) Duplication breakpoints are also enriched at replication termination sites [43], suggesting that replication dependent chromosome breakage may be an important source of duplicated sequences.

be *processed*, i.e., because of the splicing of the original gene these duplicates lack introns and UTRs. These features make them relatively easy to detect and diagnose. By contrast, DNA mediated duplications are not so. It is highly non-trivial to separate the result of a large block duplication where most genes have been deleted or moved elsewhere, from the result of a few duplication events acting on individual genes. This diagnosing difficulty is a problem – and therefore an active research area – since the ability to diagnose duplication events is important for our interest in reconstructing evolutionary history but also since functional prediction of the resulting genes are often aided by knowledge of which mechanisms have created them [18].

In a much-cited study Lynch and Conery [45] discuss the possible outcomes of a gene duplication event. Three variants are studied, namely *nonfunctionalisation* (one copy is silenced by deleterious mutations), *neofunctionalisation* (one copy acquires a novel beneficial function), and *subfunctionalisation* (the two copies each take up a fraction of the functionality of the original gene). The conclusions in the paper are the following: (1) duplicate genes arise in eukaryotes at a rate of 0.01 per gene per million years, a surprisingly high rate, well comparable to that of substitutions of individual nucleotides, (2) most duplicates experience a short period of relaxed selective pressure which ends with a silencing mutation, and (3) the few duplicates surviving long enough to develop new functions are subsequently put under strong selective pressure. Lynch and Conery also reiterate a point stated many years before [16], namely that duplications could lead to the origin of new species.

Whole genome duplications

Most gene duplication events occur at the level of individual genes or even parts of genes. Duplication of entire genomes is, however, perfectly possible. This phenomenon is rare in animals, contrasting to the situation in plants where plenty of examples are known. In this matter, Ohno laid forward the so-called 2R-hypothesis which suggested that two (or more) whole genome duplications (WGDs) have been vital in early vertebrate evolution [54, 47]. At the time, the evidence in data was scarce and the hypothesis depended on the influential personality behind it. Later on, it seemed that a strong evidence was the difference in gene count comparing species descendant from the WGD, for example human, with species unaffected by the WGD. If *C.elegans* has around 20,000 genes and human around 80,000 genes, which was the belief not more than 10 years ago [5], this would agree well with the four-fold increase of gene count that would result from a double WGD. It is now known that the human genome only hosts around 22,000 genes, so this argument cannot be used in favor of the 2R-hypothesis. In addition to that, various tests have been carried out in later years giving Ohno's hypothesis little support.

However, there are other WGD hypotheses which have proved to be correct. One of them has occurred in the lineage leading to the teleost fishes, possibly resulting in the remarkable diversification among these fishes. Also the yeast species tree [40, 60, 69] harbours a WGD (see Figure 2.3). In this case the WGD has been followed by so many gene losses that the gene count difference between the pre-WGD and post-WGD yeasts is small. The WGDs in the 2R-hypothesis must have occurred much earlier in evolution than the one in yeast, and are much harder to detect because of the many gene losses and nucleotide substitutions that have taken place since then.

2.2 Mobile elements

In 1948 Barbara McClintock, by studying various mutation patterns in maize, discovered that segments of DNA had the ability to move around in the genome [50]. These segments were mobile elements or *transposons* (sometimes also called "jumping genes") and they have later on been found to be not "of rare occurrence" which McClintock herself thought, but rather abundant in most species. A fascinating

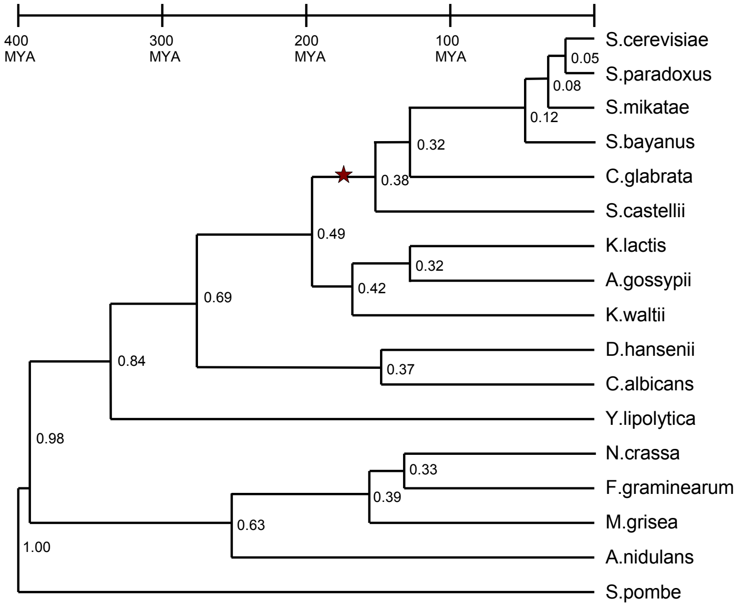


Figure 2.3: Phylogenetic tree relating a number of yeast species, including the model organism *Saccharomyces cerevisiae* – baker's yeast. The whole genome duplication occurring in the tree is marked with a star. The tree topology was obtained with the neighbour-joining algorithm [59] while the branching times were estimated using the dating algorithm presented in paper II and calibrated according to [65]. The species tree – with times – was used in order to obtain the results presented in paper IV.

example is the *P elements* in the fruit fly *Drosophila melanogaster*. The P elements were first observed around 1950 and they are believed to have originated a few years prior to this. Nowadays, only some sixty years later, these elements have been subject to a massive selective sweep resulting in them being spread to every population of the species. The P elements are believed to have been introduced through a rare horizontal transmission event in which one or more autonomous P element copies were acquired by *Drosophila melanogaster* from another *Drosophila* species [42].

Transposons are commonly classified into groups based on their mechanism of transposition. Most common are *DNA transposons*, which move directly from one genome position to another using a transposase enzyme, and *retrotransposons*, which are being transcribed to RNA and then back to DNA using reverse transcriptase. These processes can be thought of as "cut and paste" and "copy and paste" respectively. Both these processes, the latter in particular, have profound impact on most eukaryotic genomes. The human genome harbours enormous amounts of retrotransposons, including the *Long INterspersed Elements* (LINEs) and *Short IN-*

terspersed Elements (SINEs). There are some 850,000 LINE:s constituting 21% of the genome and more than one million copies of SINEs (including the famous *Alu* family). SINEs represent 11% of the genome.

LINEs and SINEs, because there are so many of them and since they tend to cluster together, are examples of what is known as *repetitive* sequences, the existence of which has some interesting side effects. Repetitive sequences increase the probability of unequal cross-over, which, as we have seen, is one basis for the origin of duplications. From a more technical viewpoint repetitions make the already challenging task of genome sequencing even harder. The most effective sequencing method, and really the one enabling successful completion of the human genome sequencing project, is called *shotgun sequencing* [67]. The genome sequence is split up into short segments which are sequenced individually and then puzzled together. If huge amounts of segments are identical or very similar, as is the case when repetitive sequences are sequenced, the puzzling task becomes nearly hopeless.

The functional significance of most of these repetitive sequences is, however, unclear. Are these the real junk DNA or is their abundance, especially among primates, important for the complexity of these species? The answer lies probably somewhere between these extremes. There are intriguing examples of unexpected conservation patterns in repetitive sequences. In [37] is reported a large family of repeat sequences, the so-called MER121, which is highly conserved in mammals. The authors speculate about a possible regulating or structural role for these sequences and conclude: "Clearly, there exist some hidden treasures among the supposed junk in the human genome.". In [11] is told the story about a previously unknown SINE family, recently active in the "living fossil" Indonesian coelacanth, from which has been derived one ultraconserved exon and an enhancing regulatory element. In [62], by contrast, is reported numerous *transposon-free* regions (TFR), that is, stretches of DNA $\geq 10,000$ base-pairs long, protected from the shower of transposons affecting the genome in general. The expected number of TFRs is, assuming that transposons are randomly distributed over the genome, very close to zero, 0.002 to be more precise [62]. This should be compared with the reported number of TFR:s which is 860! The obvious hypothesis is that these TFRs harbour sequences of extreme importance for the organism, which therefore need protection from transposons disrupting them. But since many TFRs show moderate or low conservation, alternative or complementary explanations are called for. A simple such explanation, favoured by the authors of [62], is that the estimated fraction of 5% conserved sequences in the human genome has been obtained with too blunt methods and that many regions thought to evolve neutrally are in fact selected for.

Retrogenes

As mentioned it can happen that an mRNA sequence is re-inserted back into the genome by way of the enzyme reverse transcriptase. Since the resulting retrotransposed sequences lack introns and UTRs and are removed from the regulatory environment of the original gene, they have been believed to constitute only a small fraction of functional paralogs in most genomes [18].

Indeed non-functionality is probably the case for most lower organism retrosequences, but whether this is a general fact is strongly questioned by the authors of [36]. They compare the relative impact of segmental duplication and retrotransposition in five mammalian genomes and report that the non-functionalisation of retrotransposed sequences are compensated by the high rate by which they are occurring, resulting in a near equal contribution of new segmentally duplicated genes and new *retrogenes*. Kaessman and colleagues found that the mammalian X chromosome was particularly prone to generate and recruit functional retrogenes [21, 68]. They concluded [68] that there is “compelling evidence that much of the extensive transcriptional activity of retrocopies does not represent transcriptional noise but has been profoundly shaped by natural selection”.

2.3 Not all genes are protein coding

In later years, the importance of *non-coding RNA genes* has become more and more evident. Transfer RNA (tRNA), which transfer the correct amino acid to a growing polypeptide chain during translation, and ribosomal RNA (rRNA), the primary constituent of ribosomes, have been known for a long time. But other than this, RNA has been thought of as a passive intermediate between DNA and protein. We now acknowledge the existence of tens of thousands of genes whose final product are not protein, but RNA molecules. These are often short, hence the names small nucleolar RNA (snoRNA), micro-RNA, small interfering RNA (siRNA), and small hairpin RNA (shRNA), but exceptions are abundant, in which cases the common name long non-coding RNA (long ncRNA) is often used.

Prasanth and Spector summarise the current status of our knowledge of eukaryotic noncoding *regulatory* RNAs in a recent review [57]. A point of departure of theirs is the fact that 98% of the transcriptional output of the human genome represents RNA that does not encode protein [48]. And that this fraction in fact follows organism complexity. It is indeed clear that non-coding RNA are abundant in eukaryotes – in [57] is reported that although we acknowledge only some 22,000 human protein coding genes, the total number of genes predicted from transcription analysis is around 69,000. The function of the majority of non-coding RNA:s is unknown and it is possible that many, if not most, are non-functional. But many are known to be important or even essential.

The 2006 Nobel Prize in Physiology or Medicine was awarded for the discovery of RNA interference (RNAi) [24], a process by which mRNA is targeted for destruction by specific double-stranded forms of RNA. The experiments leading to the discovery was performed on mRNA from a muscle gene in *C. elegans*, but it was soon to be demonstrated that RNAi is an active pathway in most organisms and applicable to all genes. The latter fact has since then heavily been taken advantage of in laboratories in order to turn off a specific gene.

The recognition of the importance of RNA genes is an example of the many attacks that the gene concept has had to endure. In [26] researchers in the ENCODE (short for ENCyclopedia Of DNA Elements) project reviews the history of the gene concept, from the early view of a gene as a discrete unit of heredity, to the 1940s view of a gene as a blueprint for a protein, to what they propose as an appropriate gene definition of today: "a union of genomic sequences encoding a coherent set of potentially overlapping functional products". That is, genes are not necessarily consecutive, they are often overlapping, not always protein coding, etc., but they should all have a function that affect the *phenotype*, i.e., the observed quality, of the organism in question.

Pseudogenes

Pseudogenes are copies of genes for which the functionality has been disrupted by mutations. If the pseudogene used to be protein coding, typical deleterious mutations, *disablements*, are such that change the reading frame or such that introduce premature stop codons. The normal faith of a pseudogene, because of its lost function, is that it will be released from selective pressure and thus by accumulating more and more mutations it will no longer be recognisable as formerly protein coding.

Studies of pseudogenes have been motivated by the dilemma they constitute for gene finders and hybridisation experiments because of their similarity to ordinary genes. Given their non-functionality, pseudogene sequences can also be viewed as a molecular fossil and have been used to measure background genomic substitution rates [29, 73]. There has been a number of surveys of the pseudogene population in human and related genomes [74, 66, 72]. The methodology has typically been to locate the pseudogenes that have functional paralogs in the genome in question. If the pseudogene is not too old it will still be recognisable by its similarity to the functional counterpart.

The origin of a pseudogene is generally either a segmental duplication or a retrotransposition. The resulting pseudogene is in the latter case called *processed* and in the former duplicated or *non-processed.* The authors of [74] and [66] both report around 20,000 human pseudogenes out of which some 8,000 show evidence of processing. The authors of [55] used more restrictive criteria, and identified around 3,600 human processed pseudogenes.

Given these large numbers and the inventiveness of eukaryotic cellular machinerys it is not unreasonable that some previously protein-coding pseudogenes have acquired functionality different from the original ones after the disabling event. Indeed, in [32] was reported that the mouse gene *Makorin1* has a paralog *Makorin1-p1* which cannot be protein coding but is nevertheless essential for the organism. A follow-up study [56] established that Makorin1-p1 is in fact evolutionary conserved across several mouse species.

However, it seems clear that this phenomenon of genes first losing protein coding capabilities, then evolving an alternative function is rare. Not long after the publication of the Makorin papers, a study [30] was performed that strongly contradicted the findings in these. The title of the latter paper is: "The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans" and the main conclusion in the paper is that mammalian pseudogenes should also henceforth be considered evolutionary relics.

On the other hand, support *for* the functional-pseudogene idea has also recently been published. The famous *Xist* gene, which is a key player in X-chromosome deactivation, was found to have evolved from pseudogenisation of a protein coding gene [19].

2.4 Gene regulation on different levels is likely to increase complexity

A consistently recurring theme when genome and organism complexity is discussed is the importance of gene regulation. It is commonly believed that the complexity of the gene regulatory machinery follows the organismal complexity better than the gene count does; that nature to some extent has evolved more elaborate control systems based on the already existing genes, rather than changing the genes present and their number. We are only beginning to understand the richness of eukaryote gene regulation, but it is clear that regulation is going on at many levels. Most genes are controlled by a combination of specific transcription factors, activators, repressors, and enhancers, reacting on various stimuli. The same set of genes can thus be expressed differently under different developmental stages, under different environmental conditions and in different body tissues. On a higher level stretches of DNA can be made inaccessible by structural modifications acting on the *histone* proteins onto which DNA is wound. Genes can be turned off by chemical modifications of DNA, *DNA methylation* is the most well known such modification. Such gene silencing can be effective for only a short time, but it can also be inherited *epigenetically* from parent to child.

Great, and increasing, interest is also directed towards the various processes modifying the gene product. We distinguish here between *post-transcriptional* and *post-translational* modifications. The latter is a common name for chemical transformations of a protein after translation. These include processes that remove certain amino acids from the protein, others that cut the peptide and others still that, like phosphorylation, methylation, and glycosylation, control the behaviour of the protein.

Post-transcriptional modifications include the processes which transform precursor RNA into mature RNA. The most important of these processes is RNA splicing, in which the introns are removed from the pre-mRNA and the remaining exons reconnect to a continuous molecule. In eukaryotes it is very common for pre-mRNAs to have alternative splice sites so that the same gene can result in a large set of mRNA molecules. Each one of these constitutes a unique subset of the original gene's sequence and set of exons. It is understandable that this phenomenon, *alternative splicing*, has been believed to be a substantial source of complexity. Whether this is the case is still largely an open question.

Operating on a more modest scale is the – also post-transcriptional – phenomenon RNA editing. The enzyme family ADAR (*A*denosine *D*eamination that *A*ct on *R*NA) performs substitutions of certain Adenosines (A) into Inosines (I) by binding to double stranded stem loop structures formed by preRNA. The nucleotide Inosine is read by the translational machinery as Guanosine (G), which means that when A-to-I editing is performed on exon sequences it has the effect of changing an A into a G; with possible amino-acid change as result. Current knowledge on RNA A-to-I editing is reviewed in [49] which has the title: “Breaking the central dogma by RNA editing”.

There are two types of A-to-I edited sites: hyper-edited and selectively edited. The former is very common in the 5’ and 3’ UTRs and in introns. It is *non-specific* in nature meaning that largely any A in the vicinity can be edited. Hyper-editing has no known function. Site-selective editing is in turn rare and vital for the organism. In human some 15 genes are known to be affected by this phenomenon. Interestingly, all these genes are implicated in neurotransmisson and in many cases primarily in early embryo development. Some of the known selectively edited sites are found on stems subject to extreme conservation across species.

Chapter 3

Techniques, tools, and databases

The focus of bioinformatics has traditionally very much been developing *useful* methodologies, not seldom at the expense of the soundness of the underlying mathematics. We do not argue against this, but since our primary contribution to the field – as computer scientists – are precisely mathematics and computer science skills it is important for us that we sacrifice neither the mathematical soundness nor the biological relevance. A difference between our group and maybe many others is, however, that we seek biological applications for our methods rather than the other way around.

So far this text has almost exclusively dealt with the application area – molecular biology – and not the methods we used to analyse problems in it. This chapter is intended to remedy this imbalance. It will however not be a thorough exposé of computational biology methods, there are plenty of books presenting such, but only provide a summary of the methods that I and my colleagues have used and sometimes extended, and of the tools and databases that we have used. It is also possible that I in this chapter will lose readers not familiar with concepts such as dynamic programming, likelihood, or mathematical notation in general.

3.1 Modeling molecular evolution

A common theme in the four papers presented in this thesis, is that they all model molecular evolution using aligned sequences. The evolutionary process considered is most often what we refer to as sequence evolution, i.e., DNA or protein sequences evolving by means of substitutions and insertions/deletions of individual nucleotides. The basic idea is that each nucleotide has a certain probability of changing into another during a certain period of time. Sometimes all nucleotides are considered independent from each other, but it can also be that we would like to model a biological phenomenon of which we have some prior knowledge, and that this knowledge can constrain some nucleotides. An obvious example of the latter is the case of protein-coding sequences. Due to the structure of the genetic code, we expect the rate of evolution to be lower for the first and second nucleotide in a codon – these are expected to be conserved, and higher for the third nucleotide in a codon – which is expected to evolve neutrally.

In the following is shown an example on how a substitution model can be formalised. The presented model – which we used in paper I – is known as the Goldman-Yang model [27, 13]. A Markov process is used to describe substitutions between codons. The substitution rate from codon i to codon j is specified by the instantaneous rate matrix $Q = \{q_{ij}\}$. The equilibrium frequency of codon j is π_j. Selective constraints acting on substitutions affect the substitution rate. The rate is multiplied by κ if the change involves a *transition*, e.g. a mutation changing a purine to another purine nucleotide (A ↔ G) or a pyrimidine to another pyrimidine nucleotide (C ↔ T). Normally $\kappa > 1.0$, since transitions are more frequent that transversions. The rate is correspondingly multiplied by ω, which generally is < 1.0, if the change is nonsynonymous. μ is a normalising rate factor. $Q = \{q_{ij}\}$ can now be calculated according to:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one} \\ & \text{position in a codon triplet} \\ \mu\pi_j, & \text{differ by a synonymous transversion} \\ \mu\kappa\pi_j, & \text{differ by a synonymous transition} \\ \mu\omega\pi_j, & \text{differ by a nonsynonymous transversion} \\ \mu\omega\kappa\pi_j, & \text{differ by a nonsynonymous transition} \end{cases}$$

The probability that codon i is substituted by codon j after time t can be calculated as $P(t) = \{p_{ij}\} = e^{Qt}$. The scaling factor μ allows us to measure time by the *genetic distance* between the sequences, i.e., the expected number of substitutions per codon.

The gene sequence evolution model

In paper IV, we modeled the process of gene evolution, in which the items modeled are not individual nucleotides but whole genes. Here it is fruitful to consider the corresponding sequences as units which can be duplicated and lost, rather than as collections of nucleotides. Arvestad et al. presented the *gene evolution model* [6, 8], the first probabilistic model of how a gene family evolves with respect to duplications and losses. The process generates *reconciliations* of gene trees and species trees, i.e., specifications of when speciations and duplications occurred since the time of the ancestral gene. A toy example of a reconciliation is shown in Figure 2.1. The key component is the modeling of gene duplications and losses by means of a birth-death process [41], where duplications correspond to births and losses to deaths. The gene tree is modeled as evolving within the species tree which constrains where speciations have occurred in the gene tree. This work was further extended [7] to consider the gene evolution model and models of sequence evolution simultaneously.

3.2 Inference of and with phylogenetic trees

Another common theme among the papers is – as is also indicated by the title of this thesis – that in the modeling we aim to benefit from using the sequences' phylogenetic relationships. In paper I and paper II, we modeled gene sequences evolving over a gene tree. In paper III, we worked with aligned sequences picked from a number of species and we made use of prior knowledge regarding the corresponding species tree. In paper IV, we performed reconciliations of gene trees and species trees.

There are three main methodologies for phylogenetic inference; parsimony methods, distance methods and likelihood-based methods. A detailed description of these methods is out of scope for this text (the reader is referred to Joseph Felsensteins book "Inferring Phylogenies" [23]), but the intuition is the following: with a parsimony approach we aim to find the tree explaining our sequence data using a minimal number of events, e.g. substitutions, with a distance approach we aim to find the tree best explaining (somehow calculated) distances between sequences, and with a likelihood-based approach we aim to calculate the best way of fitting a parameterised mathematical model to our data.

These methods all have strengths and weaknesses. Distance methods are "fast but inaccurate", likelihood methods are "descriptive but computationally intensive" and so on. We have most often been interested not only in the tree shape but also in values of various parameters related to the tree. Because of that, a parameterised, likelihood-based, framework has been suitable.

A phylogenetic tree harbours two main types of information, the topological relationship between the sequences, represented by the tree itself, and the (genetic) distance between the sequences, represented by the lengths of the edges in the tree. So, what one ideally would like to do, is to find the tree with edge lengths that best describes the set of sequences at hand. Unfortunately there are too many trees to examine, i.e., exponentially many, and too many possible length combinations for any given tree. Instead, in order to find the lengths, a dynamic programming algorithm is most often used. This method was originally presented by Felsenstein [22]. To find the best tree one applies "branch-swapping", a strategy where small perturbations, e.g., moving some subtree from one place in the tree to another, is repeatedly performed on the tree. Whether this strategy will be successful, i.e., often provide an almost optimal tree, is normally not known beforehand but there are methods to pinpoint unacceptably bad results.

Likelihood analysis of phylogenetic trees is today a standard procedure. However, even with the many approximations made, it is not a very fast way of finding a tree explaining the sequence data. But it has the advantage of being highly adaptable. We have been interested in ways of expressing lengths of edges in the tree not by their genetic distances, but rather by the time measured in years. The length of an edge in the tree can be seen as the product of the time passing on the edge and the (mean) substitution rate on that edge. In fact, standard maximum likelihood can be seen as a simplification of this more general case, in that all rates

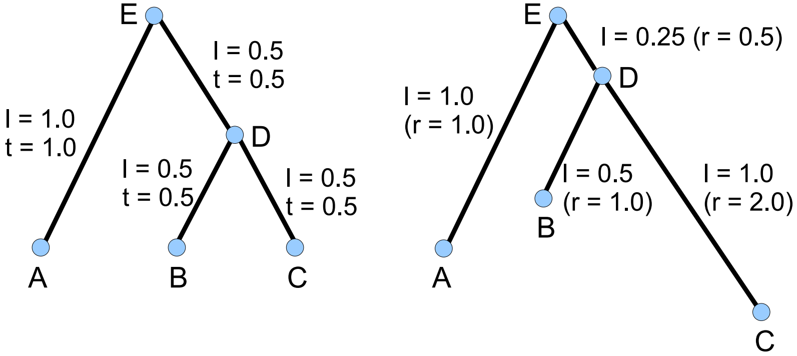


Figure 3.1: Two different phylogenetic trees. Both figures show a phylogenetic tree with three leaves {A,B,C}. In the left figure the substitution rates are all assumed to be equal to unity, i.e., the molecular clock situation. Therefore the lengths and the times are the same for each edge. In the right figure the influence of the rates is included. The edges, drawn in scale according to their lengths, are now very different from what they were in the left figure.

are assumed to be equal and therefore their influence vanishing. Whether this assumption, known as the "strict molecular clock", is acceptable or not is very much a question of the problem analysed. Sometimes it can be expected to influence results only marginally, but sometimes it does has an influence, see Figure 3.1 for a toy example of the latter; and if the assumption is made there is generally no way to infer from the results which was actually the case.

Bayesian MAP and MCMC for phylogenies

Bayesian inference involves collecting evidence supporting or contradicting a given hypothesis. As evidence accumulates the degree of belief in the hypothesis changes. We talk about the *posterior* belief which is the degree of belief for the hypothesis after including evidence and the *prior* belief which is the degree of belief before including evidence. In the present context a hypothesis is a phylogenetic tree with lengths, rates, and times, and the evidence is the sequence data. Our aim is to calculate the probability of the rates $\mathbf{r}$, the times $\mathbf{t}$, and the tree T, given the data D. This is done according to:

$$P[\mathbf{r},\mathbf{t},T|D] = \frac{\int_{\Omega_T}\int_{r\times t} P[D|\mathbf{r},\mathbf{t},T]p[\mathbf{r}]p[\mathbf{t}|T]p[T]d\mathbf{r}d\mathbf{t}}{P[D]}. \tag{3.1}$$

where the integrals are over Ω_T, the space of all trees having s leaves, the same number as the number of sequences, and over $r \times t$, the rates and times space, respectively.

This setting requires us to specify prior belief on the rates, the times, and the tree. In the rates case it is not at all obvious how this should be done – we most often model the rates as being *iid*-Γ, independent and identically Γ-distributed across edges in the tree. In the case of times and tree there is a natural and appealing choice, namely to model the tree and the edge times as obtained by a birth-death process where births correspond to speciations and deaths to extinctions of lineages.

The integrals in 3.1 cannot be evaluated algebraically nor numerically with any efficiency, compelling us to use other methods. In paper II we used a maximum *a posteriori* (MAP) methodology, i.e., the integral was approximated by its maximum, and in paper IV we performed Markov chain Monte Carlo (MCMC) integration. In both cases the space was explored by suggesting small local changes to the current *state*, i.e., the tree with lengths or rates and times. When using MAP we perform hill-climbing while the search algorithm in MCMC is the well-known Metropolis-Hastings [51, 31] proposal-acceptance scheme.

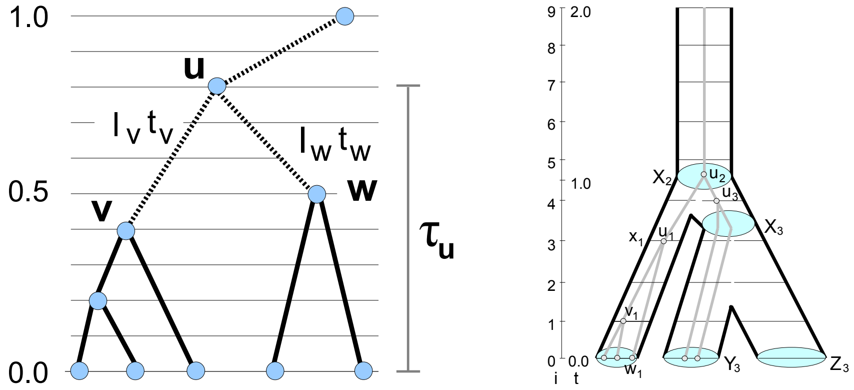
To speed up the MAP search, and avoid local optima that can be a problem for hill-climbing methodologies, we developed a dynamic programming (DP) algorithm. It was used to factorise the branch-lengths of the current state optimally into rates and times. To do this we discretised the time interval from the leaves to the root in the tree (see Figure 3.2(a)). The root and the leaves were given times zero and one respectively and all inner nodes were given times corresponding to the equidistant grid that was the result of the discretisation. The DP algorithm works from the leaves up, and when the root is reached the optimal factorisation and its probability can be retrieved.

In paper IV we applied the DP strategy in a new setting. We used the same type of positioning of gene tree nodes on a discretised grid, but this time the grid was laid out on a species tree (see Figure 3.2(b)) instead of on an interval. That is, paper II concerns a special case of the problem addressed in paper IV.

3.3 Tools

When performing a genome-wide analysis of some phenomenon, as we have done on two occasions, there is always a multitude of tools involved. These are invoked to retrieve and format data, to test ideas, to perform searches, and to analyse results. Most important are the various programs using the BLAST algorithms. BLAST (or Basic Local Alignment Search Tool as is the full name) uses clever heuristics to search huge databases in order to find sequences similar to the query sequence at hand, and very clever statistics to evaluate their significance. We used TBLASTN, i.e., BLAST with protein queries and DNA databases to find candidate pseudogenes in paper I. In paper III we used BLASTZ to find sequence pairs being approximately reverse complementary, and thus putative targets for RNA editing.

Other than BLAST, we have used a number of alignment programs based on the Smith-Waterman [63] algorithm. The typical use has been to realign BLAST output sequences in order to enhance precision, but alternatives are several. To



(a) (b)

Figure 3.2: (a) A dynamic programming algorithm for branch length factorization. Values $f_u(\tau_u)$ are calculated for all possible choices of τ_u. τ_u is u's divergence time counted from the leaves and it takes values corresponding to the grid. $f_u(\tau_u)$ is a product of rate priors for the edges leading to u, u's contribution to the time prior, and the corresponding probabilities for u's children v and w: $f_v(\tau_v)$ and $f_w(\tau_w)$ respectively. (b) A gene tree, thin grey, evolves inside a discretised species tree. Discretisation points have indexes from 0 to 9, the species tree root has time 2.0 and the first speciation has time 1.0. Three cases are shown with associated subscripts 1, 2, and 3 representing duplication, speciation, and loss, respectively. (1) A duplication u_1 occurs at x_1 resulting in two lineages containing nodes v_1 and w_1. (2) A speciation u_2 occurs at X_2, u_2 has descendants in both planted subtrees of X_2. (3) Losses have occurred since u_3 has descendants in one, S^{Y_3}, but not the other, S^{Z_3}, planted subtree of X.

retrieve alignments to analyse with the method presented in paper IV we used Mafft [39]. To perform the 4-species alignments on which the pseudogene analysis in paper I was based, we used DIALIGN [53].

In paper III the program *StemPrediction* is central. It is developed by Terrence Furey and colleagues at UCSC and uses algorithms from RNAfold [33] in order to predict double-stranded RNA stem loop structures. The input is sequence pairs resulting from BLASTZ and the result is predicted structures with associated free energy.

3.4 Databases

During my PhD student period the increase of publicly available sequence data has been impressive. When we performed genome-wide scans for conserved pseudogenes, the low quality of the contemporary chimpanzee genome was a severe problem. We nevertheless performed human-chimpanzee comparisons in addition to the human-mouse ones that was the main study. We would have liked to perform human-macaque as well but that was – at the time – not possible. Today all these genomes are sequenced to high quality and there are tens of other eukaryote genomes available as well. We benefited from this in paper III. The conservation scoring that we evaluated for our RNA editing predictions was based on sequences from 17 fully sequenced genomes, among them chimpanzee and rhesus macaque.

In the pseudogene scan we thus used human, mouse and chimpanzee genome data downloaded from Ensembl [34]. We further used the Uniprot protein database [3] to assemble protein sequence sets. Important for our methodology were also human-mouse synteny relations downloaded from the GRIMM database [15]. We evaluated our predictions using databases containing EST and mRNA data. These were downloaded from NCBI [46].

In the survey for RNA editing targets we used 17-way sequence data downloaded from the UCSC Golden Path website [38]. This dataset contains alignments between the mouse genome and 16 other eukaryote genomes. Also in this case we relied on EST databases to some extent – potential RNA editing sites should most often be an *A-G mismatch*, i.e., have an A in the genomic template but a G in the expressed sequence. A mouse EST database [14] and a mouse SNP database [61] were used, the latter in order to remove A-G mismatches likely to have polymorphic origin.

Chapter 4

Present Investigations

4.1 Surveying functional pseudogenes - paper I

In 2003, a Nature paper [32] was presented with results indicating the existence of a functional mouse pseudogene, Makorin1-p1. The authors had inserted a foreign sequence in the genome of mice embryo. The insertion was made in a genomic position chosen because it was believed to be devoid of functional material. The transgenic mice, however, were subject to severe body deformities and died shortly after insertion. It turned out that the inserted sequence had interrupted the sequence of Makorin1-p1 and that this was the likely cause for the mice dying. And this in turn indicated that Makorin1-p1 was, after all, functional and important.

Inspired by these findings and the large number of, believed non-functional, pseudogenes in the mouse and human genomes we constructed a pipeline for identification of additional examples of functional mammalian pseudogenes. The reasoning was that the inventiveness of the eukaryote cellular machinery may have made previously protein-coding pseudogenes pick up new functionality after the disabling event.

At the core of our procedure was assembly and evaluation of *sequence quartets*, containing one human gene and its mouse ortholog together with a presumed pseudogene paralog from each species. Taking advantage of the pseudogene-finding methodology published in [74] we were able to assemble 11,146 such quartets. We aimed for cases where the pseudogene had picked up new functionality prior to the human-mouse species split. We compared the likelihood of four alternative scenarios according to the following: (1) the pseudogene originated before the species split and has acquired as well as maintained function, (2) the pseudogene originated independently in the two lineages after the species split, (3) the transition from gene to pseudogene occurred independently in the two lineages after the species split, and (4) the human gene has the mouse pseudogene as sibling in the gene tree. Figure 4.1 illustrates scenario (1),(2), and (3). We never found scenario (4) and it is therefore left out of the figure.

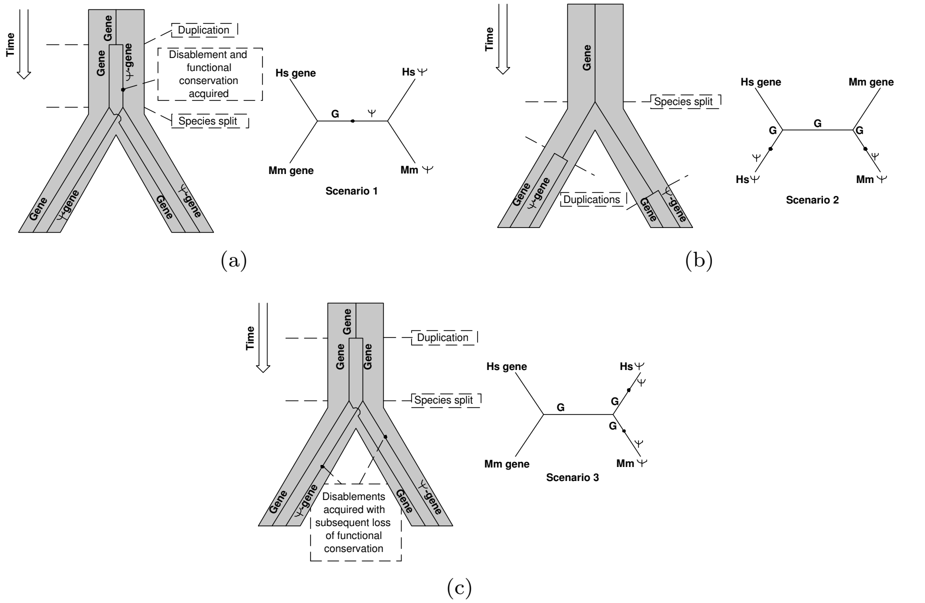


Figure 4.1: The three evolutionary scenarios into which sequence quartets were classified in paper I. In each case we show the gene tree as evolving within the species tree and separate. The gene-to-pseudogene transitions are indicated with black dots.

As many as 30 quartets were found to be best described by scenario (1), which should be compared to 0.15 which was the expected number given that pseudogenes evolve neutrally. We further tested these 30 for synteny, expression, and conservation and arrived at 20 quartets containing conserved, previously uncharacterised pseudogenes.

We repeated the analysis using human-chimpanzee comparisons. In this case, however, the task of filtering out non-functional pseudogenes was much harder than when comparing human and mouse. The assumption that nonfunctional pseudogenes having originated before the species split should have diverged beyond recognition is not true. We found 742 quartets containing pseudogenes which have evolved according to scenario (1). We were, however, unable to state with certainty whether these are functional or not.

4.2 Phylogenetic inference with special consideration taken to rate variation over lineages - paper II

The trend in the phylogenetic inference community has recently been towards leaving the assumption of a molecular clock. Modeling substitution rates and divergence times separately enables formulation of informative priors for branch lengths, as well as separate inference of rates and times. Previous work in this direction has used Markov chain Monte Carlo (MCMC) frameworks with long computation times as an immediate consequence. We wanted to investigate whether it was feasible to perform fast simultaneous inference of rates and times using a hill-climbing methodology. We were further interested in whether we could benefit from discretising the time interval from leaves to root and apply a dynamic programming (DP) algorithm to the problem. The interest stemmed from the fact that a similar procedure was promising for the integration of sequence modeling with rate modeling and gene evolution modeling, which in turn was studied in paper IV.

Two problems were addressed, the *parameter inference problem* and the *phylogeny inference problem*. The former concerns inference of rate and time parameters, i.e., values of the rates and the times for each edge in the tree, using a fixed tree topology. The latter concerns inference of parameters and tree topology simultaneously. It turned out that the parameter inference problem was effectively analysed using our methodology. For easy cases, i.e., on small trees and when the rate variation in the tree was not dramatic, we recovered parameters very well. However, in easy cases the benefit from using the DP was not large when it came to speed. In difficult cases, we gained considerably in speed as well as accuracy.

Topology inference using hill-climbing was, on the other hand, found to be hampered by the local optima problem. We found that parameters were often optimised to an extent that made escape from the current tree impossible, with obvious implication if the tree in question was not the optimal one. We therefore developed the SAL-method with which we favour tree swaps early in the process, while biasing our focus towards rate and time parameter changes later on.

We analysed two biological datasets and compared our results with MCMC-based methodologies. The rates and times point estimates that our method delivers were found to be of high quality. The speed gain was drastic in all cases.

4.3 Surveying ADAR RNA A-to-I editing targets - paper III

Together with colleagues from Marie Öhman's group at Stockholm University, we have compiled a computational screen for ADAR mediated, site selective, RNA A-to-I editing. There were several motivations for the collaboration. A number of the already known sites subject to selective editing are extremely conserved across species, see Figure 4.2. Given that there are additional sites yet to be detected, it seemed reasonable that a methodology relying on conservation should be rewarding. From our viewpoint, our collaborators deep knowledge of the biological phenomenon studied was obviously motivating, as was the opportunity to test our predictions in a laboratory environment.

Our computational screen was based on sequence conservation as well as RNA structure; known ADAR targets are all double-stranded RNA molecules with stem loop structure. For the structure we used StemPrediction. For the conservation we constructed a scoring scheme containing two parts. The first part, intended to capture absolute conservation, was based on the parsimony principle. We counted the number of substitutions needed to explain a certain column in the alignment. We computed a p-value for the event that we observe that number – or less – given a model of neutral evolution. The p-value was converted to a score and each column was given the summed score of all columns in a window surrounding that column. For the second part we reasoned that if editing was a reality for some species but not for others, the conservation could be expected to be limited to a part of the tree containing species subject to editing. We reused the window and counted the total number of substitutions within it, again assuming the parsimony model to be valid. We then computed a p-value for the event that the found number of substitutions should all occur in a subtree of the size in question. This p-value was also converted to a score, the *tree score* (see Figure 4.2 for an example), and added to the absolute conservation score.

We scored the predicted stems using the described scheme. We further evaluated the scheme by means of *A-G mismatch*, i.e., the discrepancy between the genomic DNA template and the RNA sequence. If there is indeed editing present, we expect this to be revealed by an A genomically and a G in the corresponding position of the RNA sequence.

The enrichment of A-G mismatch in conserved structures was striking. The p-value for the event that the enrichment is due to random phenomena was found

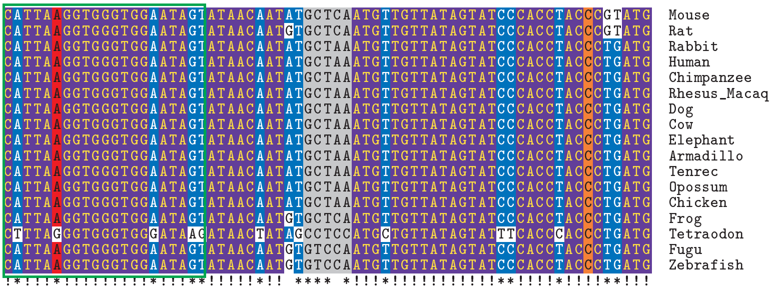

CATTAAGGTGGGTGGAATAGTATAACAATATGCTCAATGTTGTTATAGTATCCCACCTACCCGTATG Mouse
CATTAAGGTGGGTGGAATAGTATAACAATGTGCTCAATGTTGTTATAGTATCCCACCTACCCGTATG Rat
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Rabbit
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Human
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Chimpanzee
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Rhesus_Macaq
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Dog
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Cow
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Elephant
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Armadillo
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Tenrec
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Opossum
CATTAAGGTGGGTGGAATAGTATAACAATATGCTAAATGTTGTTATAGTATCCCACCTACCCTGATG Chicken
CATTAAGGTGGGTGGAATAGTATAACAATGTGCTCAATGTTGTTATAGTATCCCACCTACCCTGATG Frog
CTTTAGGGTGGGTGGGATAAGATAACTATAGCCTCCATGCTGTTATAGTATTTCACCCACCCTGATG Tetraodon
CATTAAGGTGGGTGGAATAGTATAACAATGTGTCCAATGTTGTTATAGTATCCCACCTACCCTGATG Fugu
CATTAAGGTGGGTGGAATAGTATAACAATGTGTCCAATGTTGTTATAGTATCCCACCTACCCTGATG Zebrafish
!*!!!*!!!!!!!!!*!!!**!!!!!*!! **** *!!!*!!!!!!!!!!!**!!!!*!!!!**!!!

Figure 4.2: A 17-species alignment, visualised with TeXshade [10], of the genomic region overlapping known edited site GluR-B R/G is shown. The column corresponding to the edited site is shown in red, while the complementary site is shown in orange. The loop is shown in grey. We note the extreme conservation but also the fact that most substitutions have occurred in Tetraodon. A window of the size used in the analysis, 21 nucleotides, is shown surrounded by a green rectangle. Within this window all five substitutions are found in Tetraodon resulting in a high tree score for that window.

to be $< 10^{-85}$. The experimental testing, however, turned out negative. We do see editing to an extent that is unlikely (p-value $< 10^{-18}$) to be explained by sequencing errors. However, it seems unreasonable that the enrichment of AG-mismatch that we observe is well reflected in our experimental results. Variability due to tissue and expression levels are possible explanations.

4.4 A fast implementation of the Gene Evolution Model - paper IV

Our group has developed a computational framework, PrIME – Probabilistic Integrated Models of Evolution [6, 7, 8], for integration of the single-nucleotide mutation process with the duplication and loss process. Numerous biological questions can be addressed within this framework, one such being reconstruction of reconciliations of gene trees and species trees. In [7] was described how to sample reconciliations γ from the distribution $P[\gamma|G,\theta]$ and also how to sample times t from the distribution $P[t|G,\gamma,\theta]$. By using this we can approximate $P[l|G,\theta]$ according to:

$$P[l|\theta] \approx \frac{1}{n} \sum_{\gamma \sim P[\gamma|G,\theta]} \sum_{t \sim P[t|G,\gamma,\theta]} P[l/t|G,\theta] \tag{4.1}$$

A problem with (4.1) is that on realistically sized gene trees one would need a very large number of samples in order to achieve the desired accuracy.

To alleviate this problem we designed the discretisation procedure previously described (see Figure 3.2(b)). The discretisation and predefined values for the speciation points give a common timeframe for gene tree and species tree. This is the key to the desired integration of the sequence evolution process and the gene evolution process. While the former depends only on edge lengths, the latter depends on the factorisation of lengths into rates and times. We work with a birth-death process giving prior probabilities for times and a relaxed molecular clock from which we can calculate prior probabilities for rates.

We implemented these ideas in the PrIME framework and applied the resulting program, JARDEN, to the problem of reconstructing gene trees. Given sequences for a gene family and the corresponding, dated, species tree we perform MCMC integration over lengths and tree topologies resulting in a posterior distribution over gene trees. The presented method represents a considerable improvement compared to earlier methods for gene tree reconstruction. Few of these take the species tree into account at all, and those that do are unable to integrate the sequence information and the species tree information in a mathematically sound way.

We used JARDEN to analyse the yeast family presented in Figure 2.3 and performed genome-wide analysis of gene families. We compared our method's performance with SYNERGY, a recently presented method [69], and concluded that our method meets competition well. This is a good result for several reasons. (1) We work with sequence families grouped by their methodology, (2) the analysed dataset is affected by the yeast WGD mentioned in chapter 2, a fact that makes this dataset an ideal target for SYNERGY which uses *gene-order information* together with the sequence data, (3) it is further an "easy" dataset in that it does not contain a large number of duplications besides the WGD.

We believe that we should have good chances of outperforming SYNERGY, and any other current method, on datasets where duplications are abundant, such as many animal gene families.

Chapter 5

Conclusions

The central issue when writing this thesis has been finding the common theme for my four articles. While I suspect that paper IV has been in Jens's head for several years, the choice of topic for paper I and paper III was caused by events along the road. Paper II can be seen as a test bench for the ideas implemented in paper IV.

Since the application area covered is so wide, I decided to take a wide grip when writing my main story - chapter 2. Obviously there is content enough for a few thousand more pages under the headline "On genetic motivations for organismal complexity" but spatial and temporal restrictions limited the text to what has been presented.

These past five years have meant dramatic improvements for what can be achieved with *in-silico* computational biology methods. It so happened that the very first seminar I attended as new in this field, was held by Craig Venter. He was discussing various issues related to the release of the draft human genome the previous year, and what the focus should be in this new *post-genomic* era. I do not remember whether he was specifying "integration of data from different sources" as primary, but this has certainly been stated many times since. It is clear that this is not an easy task. Even a seemingly straight-forward process integration project as the one presented in paper IV has required several years; and there is still need for a lot more research to be done. This is for two processes with appealing mathematical formulations.

Another trend that has had implications on my work has been a growing suspiciousness towards biology carried out solely in front of a computer. Today, it is very difficult to get high-impact publications without a visit to the wet lab. I believe that this suspiciousness is well motivated. More than once I have had very promising results impaired by "just one more test". And even when you are certain that fantastic results will come from the lab, some restraint may be called for.

In my opinion, the need for collaboration between biologists and computer scientists has only grown with more available data. It is important to realise that even though a certain problem can harbour little novelty to a computer scientist it can be of great importance to the biologist. But it is also clear that many future discoveries will depend on clever use of computers to analyse data.

Finally I can only confirm that the main result of expanding one's area of knowledge is that the frontier to what one does not know gets longer.

Appendix A

Background biology

The ensemble of organisms that inhabit our planet are commonly divided into three groups; *bacteria*, *archaea*, and *eukarya*. In the case of bacteria and archaea, one individual normally consists of only one *cell* – they are unicellular – while the eukaryotes are most often multicellular. Humans are eukaryotes and consist of some 100 trillion (100,000,000,000,000) cells.

Given their great variation in cell number, physical size, and complexity, all organisms have surprisingly much in common. From bacteria to human, the instructions used in development and function are encoded in the nucleic acid known as *DNA*. Eukaryote DNA is organised into structures called *chromosomes*, humans have 23 pairs of these, and a complete set of chromosomes are stored in each cell. The DNA molecule has a twisted-ladder structure, the so-called double helix, primarily supporting its function as transmitter of information from molecule to molecule and from generation to generation. This is carried out in a process called *DNA replication* in which the two sides of the DNA, the *strands*, are separating and a new double helix is constructed of each. Chemically, DNA is a stretch of units called *nucleotides*, which consist of a backbone made of sugars and phosphate groups joined by ester bonds. To each sugar is attached molecules, bases, which come in four types: Adenine, Cytosine, Guanine, and Thymine, or A, C, G, and T (illustration in Figure A.2). It is the sequence of these bases, the *genome*, that is the information defining the organism in question.

The number of nucleotides in the genome varies between organisms but it is typically enormous. The human genome contains for example approximately 3 billion nucleotides. This is roughly equal to the number of letters in this thesis – times 50,000! It is not until very recently that this information has been humanly readable, *sequenced*, in an efficient manner. To date, some 500 organisms have had their DNA fully sequenced. The majority of these are bacteria, which because of their smaller and less complex genomes are relatively easy and cheap to sequence; but the number of sequenced multicellular eukaryotic organisms are also growing fast. Among the sequenced organisms are those with particular economical interest,

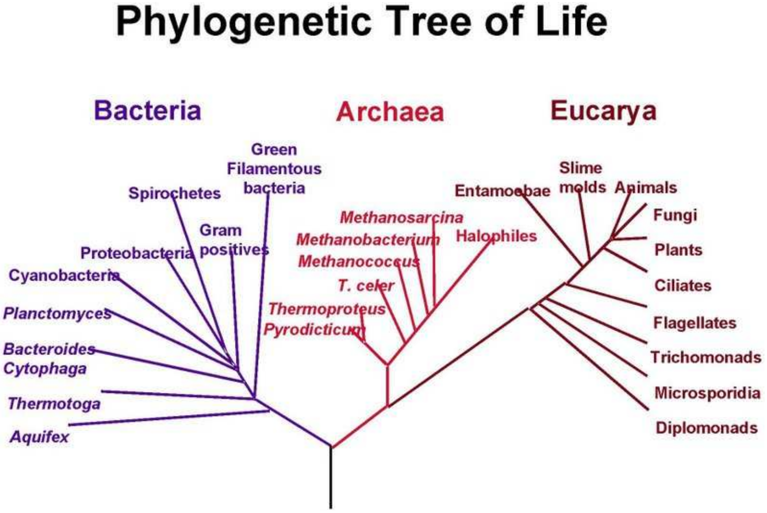


Figure A.1: Of central importance in this thesis is the concept of a tree. Most often our trees will have a root and leaves just like any tree, but it is not necessary that the tree grows upwards. A *phylogenetic tree* is a tree showing the evolutionary relationships among various biological species or other entities of life. An important example is the Tree of Life above (from [1]) showing relations between various systematic groups. Note for example the *animals* leaf harboring, among many other species, human.

for example rice and bakers yeast, or medical importance, for example *E. coli*, malaria, influenza and many others. But there are also a number of organisms, for example mouse, for which the main reason for their being sequenced is their usefulness as comparison objects to human. And, of course, since 2001 there is also a human genome sequence to compare with.

Encoded in the genomes are sequences of nucleotides called *genes*. A gene can be seen as an entity holding instructions on how to carry out a certain function in the cell. These instructions, the gene's sequence, is used by the cellular machinery to construct *proteins* which in turn perform the function.

The protein construction process consists of two major phases, *transcription* and *translation*. In the former the gene's DNA sequence is transcribed into *messenger RNA*. Like the DNA, mRNA is a sequence of nucleotides, but mRNA is single-stranded and the Thymine (T) nucleotides are replaced by Uracil (U) nucleotides. In the translation phase, the mRNA is converted to *amino acids* by way of the genetic code (see Table A.1). It specifies how triplets of nucleotides, *codons*, code for an amino acid. The amino acids are in the process connected into a long stretch – the protein.

Replication, transcription and translation are the processes constituting the

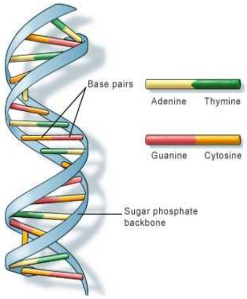


Figure A.2: The DNA double helix structure. Onto the backbone are on each strand bases connected. These bases are pairwise connected to each other in a *complementary* fashion. That is, an A from one strand is always connected to a T on the other, and likewise with C and G. This means that the sequence of bases on one single strand is enough to reconstruct a double-stranded molecule, a fact that is central in the DNA replication process.

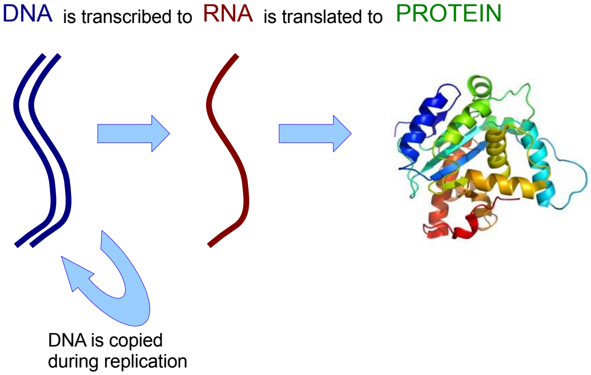

DNA is transcribed to RNA is translated to PROTEIN
DNA is copied during replication

Figure A.3: The Central dogma of molecular biology.

so-called Central dogma in molecular biology (see Figure A.3). It describes the information flow for most genes and has been a cornerstone in molecular biology knowledge during the last half-century.

On occasion, *mutations* occur in the genome. This can be due to errors made by the DNA replication machinery, but radiation, viruses and even processes under cellular control can also cause mutations. Most often, a mutation is meaningless in that it has no effect on the individual. But sometimes a mutation does have an effect and is selected for (if the effect is positive for the individual) or against (if the effect is negative). If a mutation occurs in a *germ cell*, i.e., a sperm or an egg, it can be passed on from generation to generation and spread in the population. This process, the *evolution*, creates differences between individuals in a population, which in turn creates new species when the differences between sub-populations makes cross-mating impossible.

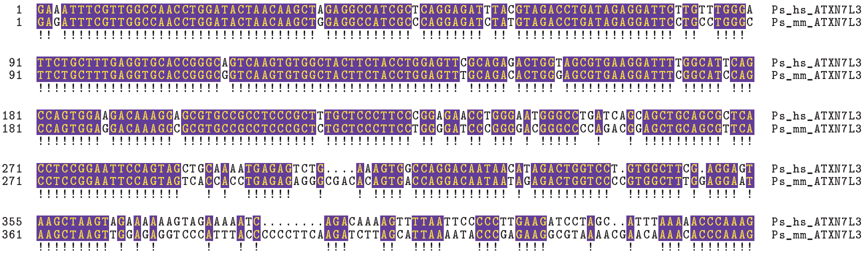
The best known type of mutation is maybe *point mutations*, i.e., where a single nucleotide is substituted. In eukaryotes, point mutations occur in the germ cell, with a rate of about 10^{-5} per nucleotide per generation [58]. That may seem very little, but it is enough to produce several thousand differences between parent and child. More rare are global mutations which is a common name for events that affect a whole sequence of nucleotides. Among these are sequence copying, *duplications*, sequence deletions, *losses*, and various other processes affecting genome segments, entire chromosomes, or, in the extreme case, entire genomes.

Amino acid	DNA codon	Amino acid	DNA codon
Isoleucine	ATT, ATC, ATA	Leucine	CTT, CTC, CTA, CTG, TTA, TTG
Valine	GTT, GTC, GTA, GTG	Phenylalanine	TTT, TTC
Methionine	ATG	Cysteine	TGT, TGC
Alanine	GCT, GCC, GCA, GCG	Glycine	GGT, GGC, GGA, GGG
Proline	CCT, CCC, CCA, CCG	Threonine	ACT, ACC, ACA, ACG
Serine	TCT, TCC, TCA, TCG, AGT, AGC	Tyrosine	TAT, TAC
Tryptophan	TGG	Glutamine	CAA, CAG
Asparagine	AAT, AAC	Histidine	CAT, CAC
Glutamic	GAA, GAG	Aspartic	GAT, GAC
Lysine	AAA, AAG	Arginine	CGT, CGC, CGA, CGG, AGA, AGG

Table A.1: The genetic code. The table shows the 20 amino acids and the DNA triplets coding for them. It should be noted that the genetic code is *redundant*, i.e., several codons are translated into the same amino acid. An example is the quartet of codons, GGT, GGC, GGA, GGG all coding for Glycine. In this case the third nucleotide in the codon is unimportant, it is said to be *synonymous*, while the first two are *non-synonymous.*

A.1 Genes and genomes

While many basic features of the cellular machinery are the same for all of today's lifeforms, e.g., the Central dogma and the genetic code, the structure of the genomes, as well as the genes therein differs substantially. In Figure A.5 is shown schematics of the genomes of the extensively studied bacterium *Escherichia coli* (a), and human (b), here representing a prokaryote and a eukaryote genome, respectively. One of the more obvious differences between them is their genomes sizes. *E. coli* has a genome size of about $4.6 \cdot 10^6$ nucleotides, three orders of magnitude less than human. Another difference is genome content. The *E. coli* genome is compact in the sense that it stores its genes in a near optimal way. It thus consists of 60% *coding regions*, i.e., more than half the genome is converted from DNA to protein.

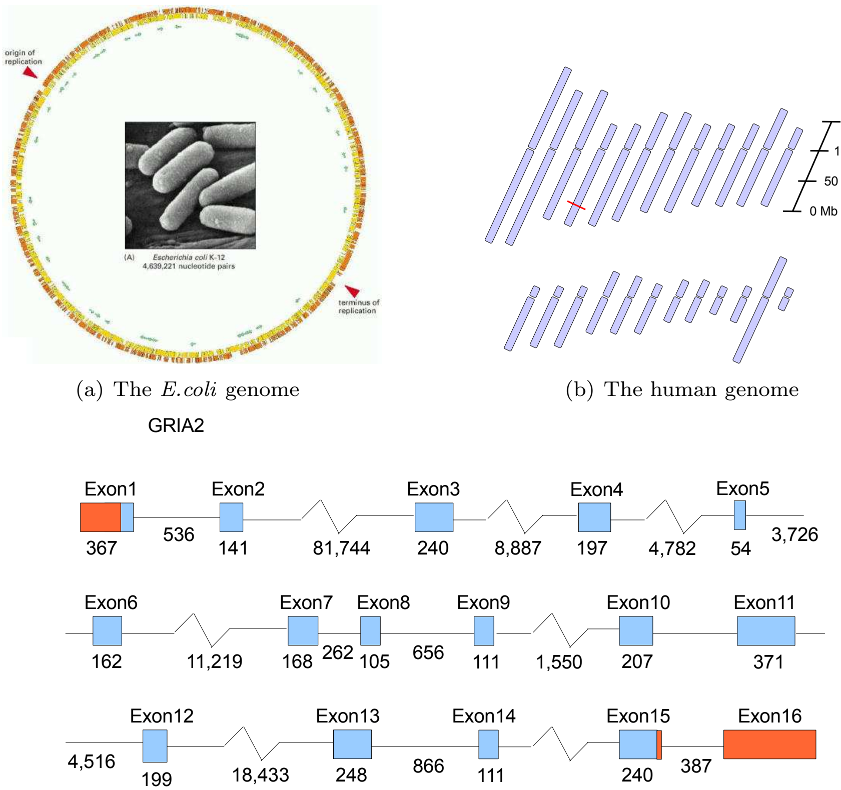


```
  1 GAAATTTCGTTGGCCAACCTGGATACTAACAAGCTAGAGGCCATCGCTCAGGAGATTTACGTAGACCTGATAGAGGATTCTTGTTTGGGA  Ps_hs_ATXN7L3
  1 GAGATTTCGTTGGCCAACCTGGATACTAACAAGCTGGAGGCCATCGCCCAGGAGATCTATGTAGACCTGATAGAGGATTCCTGCCTGGGC  Ps_mm_ATXN7L3
    !! !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! !!!!!!!!!!! !!!!!!!! !! !!!!!!!!!!!!!!!!!!!! !!  !!!!

 91 TTCTGCTTTGAGGTGCACCGGGCAGTCAAGTGTGGCTACTTCTACCTGGAGTTCGCAGAGACTGGTAGCGTGAAGGATTTTGGCATTCAG  Ps_hs_ATXN7L3
 91 TTCTGCTTTGAGGTGCACCGGGCGGTCAAGTGTGGCTACTTCTACCTGGAGTTTGCAGACACTGGGAGCGTGAAGGATTTCGGCATCCAG  Ps_mm_ATXN7L3
    !!!!!!!!!!!!!!!!!!!!!!! !!!!!!!!!!!!!!!!!!!!!!!!!!!!! !!!!! !!!!! !!!!!!!!!!!!!! !!!!! !!!

181 CCAGTGGAAGACAAAGGAGCGTGCCGCCTCCCGCTTTGCTCCCTTCCCGGAGAACCTGGGAATGGGCCTGATCAGCAGCTGCAGCGCTCA  Ps_hs_ATXN7L3
181 CCAGTGGAGGACAAAGGCGCGTGCCGCCTCCCGCTCTGCTCCCTTCCTGGGGATCCCGGGGACGGGCCCCAGACGGAGCTGCAGCGTTCA  Ps_mm_ATXN7L3
    !!!!!!!! !!!!!!!! !!!!!!!!!!!!!!!!! !!!!!!!!!!! !! !! !! !!! ! !!!!!  !   ! !!!!!!!!!! !!!

271 CCTCCGGAATTCCAGTAGCTGCAAAATGAGAGTCTG....AAAGTGGCCAGGACAATAACATAGACTGGTCCT.GTGGCTTCG.AGGAGT  Ps_hs_ATXN7L3
271 CCTCCGGAATTCCAGTAGTCACCACCTGAGAGAGGGCGACACAGTGACCAGGACAATAATAGAGACTGGTCCCCGTGGCTTTGGAGGAAT  Ps_mm_ATXN7L3
    !!!!!!!!!!!!!!!!!!   ! !  !!!!!!   !    ! !!!! !!!!!!!!!!!! ! !!!!!!!!!!  !!!!!!! ! !!!! !

355 AAGCTAAGTAGAAAAAAGTAGAAAAATC........AGACAAAAGTTTTAATTCCCCCTTGAAGATCCTAGC..ATTTAAAAACCCAAAG  Ps_hs_ATXN7L3
361 AAGCTAAGTTGGAGAGGTCCCATTTACCCCCCTTCAAGATCTTAGCATTAAAATACCCGAGAAGGCGTAAAACGAACAAAACACCCAAAG  Ps_mm_ATXN7L3
    !!!!!!!!! ! ! !      !   ! !        !!!    !!  !!!!    !!!  !!!!     !    !   !!! !!!!!!!!
```

Figure A.4: A sequence feature that is important for the organism, for example a gene, is generally *conserved*. The cellular machinery will not let mutations alter the gene's sequence since that might be harmful for its function. A comparison between such sequence segments from two different species will therefore show a higher degree of similarity than surrounding, less important, segments. Here is shown a segment of the pseudogene ATXN7L3 (reduced version of Figure 10 in paper I). In the upper part the degree of conservation is high – almost all nucleotides are identical for mouse and human. In the lower part the conservation is much lower, in fact the number of non-identical nucleotides are approaching what would be expected for non-conserved sequence segments.

Knowing this, the discovery of the eukaryotic gene and genome structure was a major surprise for the molecular biology community. Eukaryotic genomes are sparse, the human genome has only 1% coding regions, and there are large areas where there are hardly any genes to be found. The genes are very long, but only a small fraction of the genome sequence defined by the start- and endpoints of the gene is actually translated into amino acids. Instead, a typical eukaryotic gene – exemplified in Figure A.5 (c) by the gene Gria2 – consists of a long sequence segment before and after the first translated nucleotide, the so-called *5' and 3' untranslated regions* (5' and 3' UTR:s). Also, the regions that are in fact translated, the *exons*, are interrupted by (often very long) segments of untranslated sequence, *introns*. On average a human gene has 9 exons of 145 nucleotides each; these are interrupted by 8 introns of 374 nucleotides each. The transcribed region begins with 300 nucleotides 5' UTR and ends with 770 nucleotides 3' UTR (numbers from [44]). That is, the average gene has 5,367 transcribed nucleotides, but only 1,340 nucleotides are translated.

In addition to this, the genes onset-offset machinery differs somewhat between prokaryotes and eukaryotes. The basic principle is, in both cases, that the enzyme *RNA polymerase* is recruited and binds to a specific region upstream, "before", the gene's start, the so-called *promotor*. The recruitment is in turn performed by proteins called (specific) transcription factors which themselves bind to a nearby DNA sequence as well as to the RNA polymerase. Since there is a need for regulation of when a certain gene should be *expressed*, that is, called into function, one may imagine this as a network of transcription factors enhancing or repressing the



(a) The *E.coli* genome

(b) The human genome

(c) The human gene Gria2

Figure A.5: Comparison between a prokaryote and a eukaryote genome. (a) The circular genome of bacterium *E. coli*, figure from [4]. Protein coding genes are shown as yellow or orange bars depending on the DNA strand from which they are transcribed. The number of nucleotides are 4,639,221. (b) The 24 chromosomes constituting the human genome are shown. Chromosome lengths vary from 57,700,000 (chromosome Y) to 247,200,000 (chromosome 1). (c) Exon-intron structure of the human gene Gria2, position on chromosome 4 is marked by a red line in (b). Gria2 is a glutamate receptor gene and it is important for brain development. It is also the most important example of a gene affected by A-to-I RNA editing, which is the topic in section 3.3 and paper III.

function of each other. In the prokaryote genome this activity is usually going on in an area of several hundred nucleotides upstream the promotor. This regulatory region is often common for a set of genes, an *operon*, expressed together. And it is typically only one or a few proteins that regulate the expression of any particular gene. By contrast, it is common for a eukaryote gene to have tens of regulators

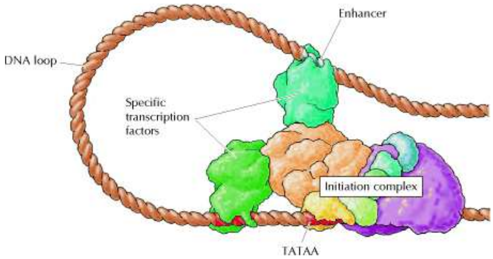


Figure A.6: The eukaryotic transcriptional machinery is very complex. The initiation complex has many components other than the key player *RNA polymerase*, here shown in purple, figure from [25]. The number of specific transcription factors, here exemplified by the green initiator and the blue-green enhancer, can also be large. Note that although the enhancer, by means of its binding to the initiation complex, is close in three dimensional space, its DNA binding site is separated from the rest by the possibly very long DNA loop.

working from distances ranging from very close to millions of nucleotides away, or from an intron of the gene itself (see Figure A.6).

Bibliography

[1] http://commons.wikimedia.org/wiki/Image:PhylogeneticTree.png.

[2] Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, Sep 2005.

[3] The universal protein resource (UniProt). *Nucleic Acids Res*, 36(Database issue):D190–5, Jan 2008.

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

[5] F. Antequera and A. Bird. Predicting the total number of human genes. *Nat Genet*, 8(2):114, Oct 1994.

[6] L. Arvestad, A-C. Berglund, J. Lagergren, and B. Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, 19:Suppl 1:7–15, 2003.

[7] L. Arvestad, A.-C. Berglund, J. Lagergren, and B. Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *RECOMB 2004: Proceedings of the eighth annual international conference on computational molecular biology*, pages 326–335, New York, 2004. ACM Press.

[8] L. Arvestad, J. Lagergren, and B. Sennblad. The gene evolution model and computing its likelihood. *Submitted to JACM*, 2007.

[9] J. A. Bailey, G. Liu, and E. E. Eichler. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet*, 73(4):823–34, Oct 2003.

[10] E. Beitz. TeXshade: shading and labeling of multiple sequence alignments using LaTeX2e. *Bioinformatics*, 16(2):135–139, Feb 2000.

[11] G. Bejerano, C. B. Lowe, N. Ahituv, B. King, A. Siepel, S. R. Salama, E. M. Rubin, W. J. Kent, and D. Haussler. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441(7089):87–90, May 2006.

[12] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325, May 2004.

[13] J.P. Bielawski and Z. Yang. Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J Struct Funct Genomics*, 3(1-4):201–212, 2003.

[14] M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev. dbEST–database for "expressed sequence tags". *Nature Genetics*, 4(4):332–3, Aug 1993.

[15] G. Bourque, P. A. Pevzner, and G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res*, 14(4):507–16, Apr 2004.

[16] C. B. Bridges. The bar "gene" a duplication. *Science*, 83(2148):210–211, Feb 1936.

[17] J. P. Demuth, T. De Bie, J. E. Stajich, N. Cristianini, and M. W. Hahn. The evolution of mammalian gene families. *PLoS ONE*, 1:e85, 2006.

[18] D. Durand and R. Hoberman. Diagnosing duplications–can it be done? *Trends Genet*, 22(3):156–64, Mar 2006.

[19] L. Duret, C. Chureau, S. Samain, J. Weissenbach, and P. Avner. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*, 312(5780):1653–1655, Jun 2006.

[20] J. A. Eisen and C. M. Fraser. Phylogenomics: intersection of evolution and genomics. *Science*, 300(5626):1706–1707, Jun 2003.

[21] J. J. Emerson, H. Kaessmann, E. Betran, and M. Long. Extensive gene traffic on the mammalian X chromosome. *Science*, 303(5657):537–40, Jan 2004.

[22] J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17:368–376, 1981.

[23] J Felsenstein. *Inferring phylogenies.* Sinauer Associates, Sunderland, MA, 2004.

[24] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–11, Feb 1998.

[25] M. Geoffrey. *The cell: A molecular approach.* Sinauer Associates, Inc., 2000.

[26] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-ENCODE? History and updated definition. *Genome Res*, 17(6):669–81, Jun 2007.

[27] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol Biol Evol*, 11(5):725–36, Sep 1994.

[28] R. Goldschmit. *The material basis of evolution.* Yale University Press, 1940.

[29] D. Graur, Y. Shuali, and W.H. Li. Deletions in processed pseudogenes accumulate faster in rodents than in humans. *J Mol Evol*, 28(4):279–285, 1989.

[30] T. A. Gray, A. Wilson, P. J. Fortin, and R. D. Nicholls. The putatively functional Mkrn1-p1 pseudogene is neither expressed nor imprinted, nor does it regulate its source gene in trans. *Proc Natl Acad Sci U S A*, 103(32):12039–44, Aug 2006.

[31] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[32] S. Hirotsune, N. Yoshida, A. Chen, L. Garrett, and F. Sugiyama et al. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature*, 423(6935):91–96, 2003.

[33] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie*, 125:167–188, 1994.

[34] T. J. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res*, 35(Database issue):D610–7, Jan 2007.

[35] M. Hurles. Gene duplication: the genomic trade in spare parts. *PLoS Biol*, 2(7):E206, Jul 2004.

[36] J. Jun, P. Ryvkin, E. Hemphill, I. Mandoiu, and C. Nelson. The birth of new genes by retrotransposition and segmental duplication during mammalian evolution. *In preparation.*

[37] M. Kamal, X. Xie, and E. S. Lander. A large family of ancient repeat elements in the human genome is under strong selection. *Proc Natl Acad Sci U S A*, 103(8):2740–2745, Feb 2006.

[38] D. Karolchik, R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, and W.J. Haussler, D.and Kent. The UCSC genome browser database. *Nucleic Acids Res*, 31(1):51–54, Jan 2003.

[39] K. Katoh, K. Kuma, H. Toh, and T. Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–518, 2005.

[40] M. Kellis, B. W. Birren, and E. S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature*, 428(6983):617–24, Apr 2004.

[41] D. G. Kendall. On the generalized "birth-and-death" process. *Ann. Math. Stat.*, 19:1–15, 1948.

[42] M. G. Kidwell. Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. *Proc Natl Acad Sci U S A*, 80(6):1655–1659, Mar 1983.

[43] R. Koszul, S. Caburet, B. Dujon, and G. Fischer. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J*, 23(1):234–43, Jan 2004.

[44] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield,

K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, and Y. J. Chen. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[45] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, Nov 2000.

[46] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 35(database issue):D26–31, Jan 2007.

[47] W. Makalowski. Are we polyploids? A brief history of one hypothesis. *Genome Res*, 11(5):667–70, May 2001.

[48] J. S. Mattick. The functional genomics of noncoding RNA. *Science*, 309(5740):1527–1528, Sep 2005.

[49] O. Maydanovych and P. A. Beal. Breaking the central dogma by RNA editing. *Chem Rev*, 106(8):3397–411, Aug 2006.

[50] B. McClintock. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*, 36(6):344–55, Jun 1950.

[51] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machine. *J Chem Phys*, 21:1087–1091, 1953.

[52] C.W. Metz. Duplication of chromosome parts as a factor in evolution. *Am. Nature*, 81:81–103, 1947.

[53] B. Morgenstern. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3):211–218, 1999.

[54] S. Ohno. *Evolution by gene duplication.* Springer-Verlag, New York, 1970.

[55] K. Ohshima, H. Masahira, T. Yada, T. Gojobori, and Y. Sakaki et al. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol*, 4(11):R74, 2003.

[56] O. Podlaha and J. Zhang. Nonneutral evolution of the transcribed pseudogene Makorin1-p1 in mice. *Mol Biol Evol*, 21(12):2202–2209, 2004.

[57] K. V. Prasanth and D. L. Spector. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev*, 21(1):11–42, Jan 2007.

[58] L. B. Russell and W. L. Russell. Frequency and nature of specific-locus mutations induced in female mice by radiations and chemicals: a review. *Mutat Res*, 296(1-2):107–27, Dec 1992.

[59] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–25, Jul 1987.

[60] D. R. Scannell, K. P. Byrne, J. L. Gordon, S. Wong, and K. H. Wolfe. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082):341–345, Mar 2006.

[61] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–11, Jan 2001.

[62] C. Simons, M. Pheasant, I. V. Makunin, and J. S. Mattick. Transposon-free regions in mammalian genomes. *Genome Res*, 16(2):164–72, Feb 2006.

[63] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, Mar 1981.

[64] J. S. Taylor and J. Raes. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*, 38:615–43, 2004.

[65] T.N. Taylor, H. Hass, and H. Kerp. The oldest fossil ascomycetes. *Nature*, 399(6737):648, Jun 1999.

[66] D. Torrents, M. Suyama, E. Zdobnov, and P. Bork. A genome-wide survey of human pseudogenes. *Genome Res*, 13(12):2559–2567, 2003.

[67] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–51, Feb 2001.

[68] N. Vinckenbosch, I. Dupanloup, and H. Kaessmann. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A*, 103(9):3220–3225, Feb 2006.

[69] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449(7158):54–61, Sep 2007.

[70] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S. P. Yang, E. M. Zdobnov, M. C. Zody, and E. S. Lander. Initial sequencing

and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, Dec 2002.

[71] J. Yang, R. Lusk, and W. H. Li. Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci U S A*, 100(26):15661–15665, Dec 2003.

[72] Z. Zhang, N. Carriero, and M. Gerstein. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet*, 20(2):62–67, 2004.

[73] Z. Zhang and M. Gerstein. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.*, 31(18):5338–5348, 2003.

[74] Z. Zhang, P.M. Harrison, Y. Liu, and M. Gerstein. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res*, 13(12):2541–2558, 2003.