1 General introduction

Over the past seven years we have witnessed a revolution in (molecular) biology, namely the sequencing of complete genomes of cellular organisms, starting with the simple parasitic bacterium *Haemophilus influenzae* (Fleischmann *et al.* 1995) and culminating in the draft of the human genome sequence (Venter *et al.* 2001, Lander *et al.* 2001). Complete genome sequences, which are mainly obtained through (whole) genome shotgun sequencing, are a unique type of data, because they represent in principle everything that together makes an organism. In a way one could say that we now have a complete list of the pieces that in still largely unknown ways, together and in interaction with the environment, constitute the puzzle of life. It is however not immediately clear what we concretely can do with all these genomes. Obviously they should function as bench for "wet biologist" allowing for example the rapid identification of proteins by their mass spectrometry signature (Gavin *et al.* 2002), but what, if anything, can we learn based 'solely' on this data. For one thing, their availability presents us with an unprecedented wealth of data to study evolution. Since genome data is relatively new and our picture of genome evolution is still very incomplete, such studies entail 'blind' pattern analysis to search for the basic concepts in which we can describe and understand genomes as well as their evolution. Comparative genome analysis thereby provides us with some idea of how genomes came to be. Apart from its intrinsic interest, this understanding is necessary for the efficient usage of complete genomes, for example to evaluate whether the presence of a certain gene is surprising. In general, complete genome sequences allow the study of protein function within the framework of the complete cellular and genomic context. This thesis will deal with a set of bioinformatic analyses that cover different levels of comparative genome analysis (Bork *et al.* 1998). In this introduction I will (1) make the case for studying complete genome sequence data through integrated evolutionary and bioinformatic analysis, (2) introduce comparative sequence analysis, (3) introduce comparative genome analysis, and (4) describe major results from comparative genome analysis that will provide a background for the work described in the main body of this thesis.

**Complete genome sequences, bioinformatic analysis, and evolution**

**Complete genome sequences**

Complete genome sequencing projects provide us with huge amounts of data. These data obviously need computer or informatics analyses to create and maintain them. What is probably more important is the subsequent data analysis to create new biological knowledge from the (complete genome) sequences. Large scale databases of
DNA/protein sequences and protein structures were studied extensively already before complete genomes, because they were already available and because they are very suitable for formal analysis (Bork and Koonin 1998). As such a substantial body of tools and concepts have already been developed to analyze them (Thompson et al. 1994, Felsenstein 1989, Smith and Waterman 1981, Altschul et al. 1990). Presently, many other types of data are being generated by large scale biological experiments such as gene expression (by microarrays, or SAGE (Hughes et al. 2000, Cho et al. 1998)), genomic mutations screens (Winzeler et al. 1999, Tong et al. 2001), proteomics 2D gels (Fey and Larson 2001), peptide/protein chips (Houseman et al. 2002), mass spectrometry (Gavin et al. 2002, Ho et al. 2002), i.e. the whole batch of other 'omic' data. Complete genome sequence data (more so than other sources of sequence data such as EST's) are noise free compared to newer 'omic' large scale biological data. Moreover genome data are inherently of a discrete nature and their formalization is well established. These intrinsic features of genome sequence data and the aforementioned existing body of sequence analysis tools, is probably why complete genome sequence are analyzed more frequently and more successfully than other types of large scale biological data.

**Evolution and bioinformatic analysis**

The intimate relationship between evolution and bioinformatic analysis is nicely illustrated by the fact that one of the first computational analyses on sequences has been phylogenetic analysis, i.e. molecular evolution (Fitch and Margoliash 1967). Based on these bioinformatic studies of sequences many important and intrinsically relevant results for the study of evolution have been obtained. It has revolutionized taxonomy and our understanding of the interplay between phenotype and genotype (Olsen et al. 1994, Clarke et al. 1989). On the other hand, as much as bioinformatic tools are used for the analysis of molecular evolution, they are also based upon it. This is apparent from the lowest levels of analysis such as gene prediction where homology based gene prediction is the most successful computational gene prediction method (Guigo et al. 2000), to higher levels such as the ability to delineate functional modules of interacting proteins through the analysis of evolutionary conserved genomic fingerprints (chapter 6). The relationship between evolutionary and bioinformatic analysis is evidently reciprocal and synergistic. We therefore, to paraphrase Dobzhansky, study genome evolution not only for the sake of evolution itself but also because nothing in genomic biology makes sense except in the light of (genome) evolution. As such evolution and its study, are one of the keys to "unlock nature's warehouses" which complete genomes hold.

Not only do bioinformatic and evolutionary analysis support each other, their combined effort feeds back results into the larger molecular biological community. Among these results are many concrete findings for individual proteins. For example a gene reported as causing breast cancer but without a known molecular function, was subsequently by independent bioinformatic research convincingly predicted to be nuclear signaling receptor (Koonin et al. 1996). On a larger scale, important general biological findings are obtained by integrated bioinformatic and evolutionary analysis. For example, the large scale analysis of the number of known alternative splice variants corrected for the sizes of existing EST databases argue that the perceived increase in human complexity relative to fly or worm cannot be explained by an increase in alternative splicing (Brett et al. 2002) as was hypothesized earlier. Moreover the ability and knowledge of how to digest large scale biological databases, can be formalized in the form of (web) tools and distilled into
higher level information in the form of databases, making them available to the community (Tatusov et al. 1997, Kanehisa and Goto 2000).

**Comparative sequence analysis**

**Comparative sequence analysis and comparative genome analysis**

Comparative genome analysis is relatively new. As such it tries to define the basic concepts in which to describe and understand genome evolution. However these attempts do not start from scratch: its most important tools by and large stem from the realm of classical sequence analysis. In fact, many insightful comparative genome analyses are in effect uncomplicated comparisons that apply sequence analysis tools on a genome wide scale. The analysis of complete genomes has not only benefited from existing sequence analysis tools. Rather, the wealth of data generated by genome sequencing projects stimulates the improvement of existing tools and creation of new sequence analysis tools. The development of new and improved conventional sequence analysis tools is (i) needed to deal with the sheer amount of data (e.g. DbClustal, Thompson et al. 2000), and (ii) to exploit the new possibilities that this data offers (e.g. PSI-BLAST Altschul et al. 1997).

**Pairwise homology searches**

Arguably the most important task in sequence analysis is establishing whether two sequences are homologous, i.e. if they stem from the same ancestral sequence. One of the most widespread applications of establishing homology is function prediction, because homologous sequences tend to have similar functions (Bork and Koonin 1998). The extent of homology roughly corresponds to different levels of conservation of molecular properties. Very similar sequences are likely to have the same substrate specificity, while proteins with intermediate levels of sequence similarity tend to catalyze the same reaction, albeit on related but different substrates. When two sequences are distant homologs, only the most general characteristics of the protein such as its 3D structure, i.e. 'fold', can be assigned because this is the most conserved property of a protein. Note that the type of function that one predicts this way is the *molecular/enzymatic* function of a protein. Other dimensions of protein function are cellular and biological process in which a protein plays a role, or its localization in the cell. Currently there are systematic formalized vocabularies, i.e. ontologies, being constructed that attempt to deal with this challenge such as the Gene Ontology project (The Gene Ontology Consortium 2001). Moreover, detecting homology is also the first step in the complicated task of determining what can be considered to be the corresponding gene between two genomes (Fitch 1970; see below).

In practice establishing whether two sequences are homologous is performed in the context of a search of a query sequences against a database of many other sequences. Different tools (using different heuristics) align the query sequence consecutively to each query sequence from the whole database. Based on our condensed knowledge of sequence evolution in the form of amino acid substitution matrices and gap opening/extension penalties, a score is computed for each alignment. Taking into account
database size and composition, the score is used to compute a expected chance of similarity. This allows a statistically sound assessment of whether two sequences are homologous or whether the observed similarity could be due to chance alone (Altschul 1990). However we cannot assess the absence of homology. Moreover note that two genes are not necessarily homologous across their full length. Different modules can be attached to the N-terminus, C-terminus, or even in the middle of a protein. These modules that form evolutionary, functionally and structurally independent units, are referred to as protein domains (Schultz et al. 1998).

Multiple sequence alignments, trees, and profiles

When comparing sequences one has to find out which positions in the sequences at hand are equivalent. This is called a sequence alignment. Aligning two sequences (pairwise sequence alignment) is necessary to determine whether they are homologous. Hence this is crucial for the homology searches described above. Comparing more than two sequences, i.e. making multiple sequence alignment, gives more information, thereby opening new possibilities. Multiple sequence alignment poses a big algorithmic and computational challenge, but adequate programs do exist, such as CLUSTALW (Thompson et al. 1994) and T_COFFEE (Notredame et al. 2000). In general, a multiple alignment is useful so see which positions, or combinations of positions (i.e. motifs), are conserved, and thus important for the function of that protein. Multiple sequences alignments form the necessary prerequisite for reconstructing reliable phylogenetic trees of genes, because they allow the detection of their evolutionary differences at the equivalent positions. Moreover phylogenetic trees themselves albeit it based only on pairwise alignments are used as so called ‘guide trees’ by programs such as CLUSTALW and T_COFFEE to make reliable multiple sequence alignments. In any case, phylogenetic trees from these alignments can be used to determine the relationships between species (molecular systematics). Most notably, the systematic collection and subsequent phylogenetic analysis of ribosomal RNA sequences (well conserved and essential genes present in all living organisms), have established the current view of life on earth as being divided in three kingdoms (Olsen et al. 1994). Tree building of single genes also allows the study of protein evolution and its relation to function (Copley and Bork 2000). Lastly, because multiple alignments allow us to evaluate the amino acid conservation at certain positions, they open up the possibility to use this information for searching divergent homologs. This is done by constructing profiles (or patterns). Profiles can be either based (i) on an ad hoc alignment of the sequences that are found during the search of the database anchored to the single initial query sequence (PSI-BLAST, Altschhul et al. 1997), or (ii) on a (manually curated) explicitly reconstructed multiple alignment (HMMER Eddy 2000, SearchWise, Birney 1996).

Comparative genome analysis

When one considers genomes as bags of marbles, what are the marbles?

Comparative genome analysis obviously is much younger than sequence analysis. As with sequence analysis, we need to establish equivalency among the components we
compare (i.e. homology among sequences). Unlike sequence analysis however,
comparative genome analysis is on a more fundamental level first faced with the question
of which components in the genome we want to compare. Similar problems are
encountered in classical comparative studies like comparative zoology or morphology.
This question actually is central in comparative genome analysis, namely: what are the
components or characters for which we would like to establish equivalency? If we see
genomes as bags of marbles, what are the marbles? Nucleotides, gene functions,
intergenic sequences, proteins, regulatory elements, protein-protein interactions,
metabolic pathways, and of course genes, are all possible characters. Note that we can
already here encounter big technical problems in identifying 'the marbles' within the
genome due to the multi-level nature of the data. For example, when we want to compare
genomes at the level of genes, as we do here in chapter 4, identifying the coding regions
in the genome, i.e. the genes, is a non-trivial task. These severe problems in gene
prediction thus seriously influence any analysis that wants to compare genomes as bags of
genomes. Similarly, when comparing the metabolic pathways in two species, one is
primarily interested in the presence and absence of certain enzymatic functions.
Compiling such a list of which enzymes are present in a genome is difficult, because even
for the best studied organisms this involves reliable function prediction for all genes in
the genome, which is difficult to attain (Huynen et al. 1999). Technical difficulties in
obtaining these characters aside, the question of which character to study, probably
depends on the research inquiry and tools at hand. In the aforementioned study of
metabolic pathways the primary interest in the presence and absence of enzymatic
functions makes the question whether the enzymes that code for these activities are
homologous of less relevance. Thus the multi level nature of genomes is reflected in
different levels of functional analysis (Bork et al. 1998). This thesis mainly deals with
genomes as bags of genes and their relations. However to offer a general perspective,
some lower levels of genome comparison are discussed first.

Genomes as bags of nucleotides and amino acids

On a most basic level, one can see the genomes as bags of nucleotides or encoded amino
acids. When one leaves out the strict evolutionary requirement of common ancestry, i.e.
homology, and instead opts for simple equivalency, the classification is trivial. For
example one can make an analysis of genomes by taking its complete DNA or all its ORF
sequences, and considering them as bags of nucleotides and amino acids to obtain average
statistics. The most obvious example is Guanine-Cytosine (GC) content, which is a
classic taxonomic indicator of microbial genomes. Complete genome sequences have
confirmed the previously found biases for certain species that were determined with
isopycnic centrifugation in CsCl (Enea and Zinder 1975). Using complete genome
sequences more complicated analyses that search for genomes as bags of short nucleotide
words, have shown that there is unique fingerprints for all genomes even to the extent that
one can differentiate strains of the same species (Sandberg et al. 2001). What furthermore
has become possible is to find regions within the genome that significantly differ in GC
content (Lawrence and Ochman 1998). Detecting such regions has been a fruitful
approach to find regions in the genome that might have arrived there through horizontal
gene transfer.

Since a GC bond is stronger than an Adenine-Thymine bond, thermophilic organisms
might be expected to have a bias in their GC content to stabilize their genomic DNA. This
is however not the case. Instead they seem to use reverse gyrase to stabilize their genomic DNA by supercoiling (Forterre 2002). They do however need more GC bonds in their RNA genes to maintain their functionality. This observation has allowed the finding of new RNA genes in the genomes by searching of regions with significantly higher GC content (Omer et al. 2000), similar to the finding of putative horizontally transferred genes based on different signatures (see above).

Not only do genomes have distinct GC contents, there is also a difference in usage of amino acids for the proteins. Comparison of global genome statistics that treat all ORFs as one big pool of amino acids, have found two significant trends: the first is that the GC content correlates strongly with the Arginine content, while having a strong anti correlation with the lysine content (Kreil and Ouzounis 2001; Cambillau and Claverie 2000). This is an almost purely mechanistic result of the underlying GC bias. The second finding from complete genomes with regard to amino acid content has been that hyperthermophily is characterized by a sharp increase of charged residues, notably Lysine and Glutamate, at the expense of polar non charged residues, mainly Glutamine. We thus find effects of a feature of the highest level of organismal phenotype, i.e. the temperature at which it lives, onto the lowest levels of molecular observation, the amino acid content of its proteins and the nucleic acid contents of its RNA genes. On the other hand, there are pure statistical (seemingly random) biases on the DNA level that also affect the amino acid content. These deviating amino acid or nucleic acid compositions provide us with examples of the relation between the habitat and its composing parts, which stand at the core of genome function and evolution. These biases in the sequence composition probably affect homology detection and phylogenetic inference, however in practice they are not (yet?) taken into account.

Comparing genomes as bags of genes means establishing equivalency among the genes: homology and orthology

Since genomes are basically very long sequences, one might be tempted to align them just as normal sequences. Thereby one would obtain at the lowest possible level a strict evolutionary equivalency for each nucleotide to each other nucleotide. However, this is only possible with very closely related genomes because of the fast rate of genome shuffling (Suyama and Bork 2001). Hence the need for a higher level, more modular, analysis: at the level of genes. In general comparative genomics mostly operates at this bag of genes level (Huynen and Bork 1998). Having established what the characters are, in order to perform comparative genome analysis, we now must establish which is gene is equivalent to which other gene. The starting point for this is finding homologous genes. Applying the sequence analysis tools described above on completely sequenced genomes thus yields the basic data for performing comparative genome analysis. However the evolutionary dynamics of genes relative to the evolutionary dynamics of the species wherein they reside, has given rise to the insight that homology as a definition for 'the same gene' in different species is conceptually insufficient due to gene loss, and ancient as well as recent gene duplications.

The concept that seems to offer the best solution for these complications is orthology (Fitch 1970). Two genes in two organisms are defined as being orthologs when they are homologous and they diverged from each other at the same time as the two species diverged from each other, i.e. they are related by speciation rather than by gene
duplication. The simplest operational definition for orthology when comparing two species that has been put forward, is the bidirectional best hit (Tatusov et al. 1996). This approach has proven to be very useful for such comparisons (Huynen and Bork 1998, Overbeek et al. 1999, Tamames 2001) and we also employ it in chapters 2, 3, and 5. However, operational orthology definition becomes more complicated when we compare more than two genomes. As orthology is defined with respect to speciation, when we compare multiple species, then it is the last common ancestor of all these genomes, and we obtain an orthologous group of genes which does not necessarily includes a single gene per genome. In the case that the comparison spans all completely sequences genomes, the relevant ancestor is the last common ancestor of all extant life. An orthologous groups in that case includes all genes that stem from one single gene in the last common ancestor of all extant life. Obviously many gene evolution events (most notably gene duplication, gene loss, and horizontal gene transfer) can have occurred to an orthologous group of genes since this ancestor. This principle of group orthology is what underlies the methods we use in chapter 4 and 6, and also the COG (clusters of orthologous groups) database (Tatusov et al. 1997).

**Comparing genomes on the level of genes: gene content evolution**

**Gene family evolution within genomes**

Whether two genes, or the proteins domains they are composed of, belong to the same gene family is an operationally relatively well defined question, thanks to tools from sequence analysis. The study of gene family dynamics within the genome, is therefore a fertile and successful example of applying conventional sequence analysis tools to genes on a genome wide scale. There are various levels of relatedness in defining gene families: three levels on which gene family dynamics within the genome has been studied are recent gene duplications (since the speciation from intermediately close relatives)(Jordan et al. 2001), conventional homology by sequences similarity based gene families (Huynen and Nimwegen 1998), and the fold level (Qian et al. 2001). Note that only for a few genes within a genome the 3D structure is known. Therefore an important spin-off from approaches studying the number of different genes in a genome that are of a certain fold, is fold prediction through sensitive distant homology searches (Huynen et al. 1998, Teichman et al. 1998). There are also different approaches to detect these families: either bottom up by all against all sequence comparisons, or top down by scanning a genome with profiles. Top down searches seem to be more powerful and easier, but are only made possible in the first place by manually curated bottom up searches that are used to create their profiles.

Irrespective of the conceptual or heuristic approach, the results all point to the same thing: the frequency distribution of gene families in all genomes follows a power law. This distribution can be explained by a deletion/duplication model in which related genes have a similar chance of being deleted or duplicated. This is probably due to related genes having similar function, and are thus under a similar selection regime as first shown and proposed by Huynen and Nimwegen (1998) and more recently by Qian and co-workers (2001). An analysis of recent gene duplications (Jordan et al. 2001), shows similar
patterns despite the fact that the duplications that gave rise to fold or gene families by and
large have occurred much longer ago. The result thus holds for different time scales.

**Gene evolution versus genome evolution**

Obviously an organism obtains most of its gene from its direct ancestors. One would
therefore expect that phylogeny is the major determinant in gene content similarity.
Initially it was shown that when comparing shared gene content of complete genomes
with some measure of evolutionary time (like protein sequence evolution), it correlates
with the evolutionary proximity (Huynen and Bork 1998). However other types of
analysis, which do not focus on the presence and absence of genes, but rather compare
trees of genes with those of the presumed organismal tree, suggest that many gene trees
are inconsistent with organismal tree (Doolittle and Logsdon 1998). This has prompted
the notion that horizontal gene transfer (HGT) is a substantial or, maybe even, dominating
force in determining gene content. Similar estimates for the dominance of HGT come
from studies that use deviating GC content or codon usage to determine which genes have
recently been transferred (Lawrence and Ochman 1998). The apparent ubiquity of HGT
has resulted in a number of publications that cast doubt on the very notion of an
organismal phylogeny (Doolittle 1999). Still, as will be discussed in this thesis and shown
by Tekaia and coworkers (1999) and Fitz-Gibbon and House (1999), the gene content
contains a quantitatively dominant phylogenetic signal.

From all this emerges a picture where for one the most fundamental properties of
genomes, its gene content, we struggle to reach an understanding of how it comes to be.
This in contrast to sequences and their multiple alignment, for which heuristics do exist in
the form of substitution matrices. Although these substitution matrices are not a perfect
model for sequence evolution, they have provided us with useful tools for studying
sequence evolution. The lack of insight in the gene content evolution of complete
genomes as a fundamental evolutionary process, presents us with no basic or neutral
expectation for behavior of genes. Among other effects this also limits the assessment of
how surprising the absence or presence of a gene is. It thereby illustrates the need for
strategies such as the one outlined in chapter 4 that explicitly reconstruct which
transformations have occurred over the course of genome evolution.

**Genome evolution beyond a bag of genes**

**Evolution of gene order**

In all analyses described above the only information from the genome that is used, is that
it consists of a certain bag set of genes. And even that information is only used to increase
for example the number of observations of genes that show a characteristic x (such as
being shared with another genome, or having a TIM-barrel fold). Naturally there are
approaches that do exploit the unique additional information from complete genomes.
When doing that, the same tools and concepts as described above are used, while at the
same time operating at a higher level of genome description. One of the most immediate
analyses beyond a bag of genes that uses tools from conventional sequence analysis, is
the most simple link between genes, namely their order on the chromosome. Gene order
as a step beyond gene content has been studied in mitochondrial genome analysis with the aim of recovering phylogenies (Boore and Brown 1998). Actually in many ways mitochondrial genomes have provided pilot studies for analyzing larger nuclear genomes. Hence gene order is studied quite extensively insofar as genomes are available. Based on the first available prokaryotic genomes it was concluded that gene order is not, or only very poorly, conserved (Mushegian and Koonin 1996). More quantitative approaches similar to shared gene content over evolutionary time, show that the amount gene order conservation decreases more rapidly than other measures of evolutionary time like protein sequence identity, but that even over large evolutionary distances some conservation can be observed (Huynen and Bork 1998). Interestingly, those gene pairs that (Galperin and Koonin 1996) are conserved seem to be functionally interacting genes. Studies on gene order in the complete genomes of eukaryotes show that here it evolves faster than in prokaryotes, with hardly any shared gene order left, at distances where prokaryotes still share a substantial number of gene pairs (Huynen et al. 2001).

**Predicting interactions between proteins using complete genomes**

As has been done for gene order, we can study the evolution of a diverse set of genomic relations between genes. Many of these relations tend to evolve relatively quickly as is observed for gene order (Huynen and Bork 1998). Therefore when these genomic links are conserved, selection is probably operating to keep them intact. As mentioned above, this for example has already been suggested to be the case for conserved gene order because the gene pairs tended have some functional link (Galperin and Koonin 1996). Subsequent in depth analysis various types of relations between genes have found some genomic associations that were shown to reflect functional associations (reviewed in Huynen et al. 2000). These genomic associations are the result of evolutionary pressure and thus reflect the traces left in genomes by the selection on functionally interacting proteins.

Until now three different types of genomic associations have been introduced. Firstly, the most general type of genomic association is the tendency for genes to be absent and present together from the genome (Huynen and Bork 1998, Pellegrini et al. 1999, Tatusov et al. 2001). This co-occurrence of genes in genomes (phylogenetic profiles) indicates that they have been lost and gained together, which in turn has been shown to be indicative of a functional interaction. Secondly, as mentioned above, one can observe that gene pairs whose order is conserved seem to be functionally interacting genes (Galperin and Koonin 1996). This in turn has stimulated more systematic large scale complete genome comparisons that have systematized and established conserved gene order as a very powerful tool for the prediction of functional interactions based on this 'conserved local genomic context' (Dandekar et al. 1998, Overbeek et al. 1998, Huynen et al. 2000). Note that the conservation of the gene order is more important than the presence of two genes in the same operon, because (i) there are cases known where the gene order is conserved but the gene cluster consists of different transcriptional units in different organisms (Suh et al. 1996), and (ii) genes in the same operon but only in one species do not necessarily necessarily have a functional association (Salgado et al. 2000). Finally the most intimate form of genomic association is the fusion of two genes into one polypeptide. This type of associations has been shown to be a very strong predictor that the two genes have a functional interaction, albeit with relatively low coverage (Enright et al. 1999, Marcotte et al. 1999, Yanai et al. 2001)
These genomic context, or genomic association, approaches go beyond comparative genome analysis as a bag of genes, because they actually look at the relations between the genes. Since they predict functional interactions between genes rather than molecular functions of genes themselves, they are orthogonal to conventional function prediction by means of homology searches (see above).

**This thesis**

This thesis deals with a set of bioinformatic analyses that cover different types of comparative genome analysis on the level of genes and their relations (Bork *et al.* 1998). The chapters follow the build up from defining the equivalency among genes across genomes (orthology), to the basic evolutionary pattern in gene content evolution, to gene order evolution, and large scale analysis of the genomic associations between genes.

In *chapter 2* we study the occurrence of gene fusion and gene fission on a genome wide scale. Fusion and fission (e.g. the fragmentation or splitting of genes) are two principal processes in molecular evolution. However they are also complicating factors in defining orthology (Huynen and Bork 1998). These processes so far had mainly been recognized and described in individual cases (although they have been studied for large scale function prediction Enright *et al.* 1999, Marcotte *et al.* 1999). The estimates of the frequency of occurrence of gene fission and gene fusion that we obtain are compared to each other and across the various genomes. The quantitative analysis shows a prevalence of fusion, which can be expected because there is a benefit to fusion in that it allows for the physical coupling of functions that are biologically coupled. We separate fission into cases that look more like frameshift sequencing errors or very recent frameshift mutations on the one hand, and cases of established 'genuine' fissions on the other. Interestingly a correlation of the genuine fissions with a thermophilic lifestyle is found. We here argue that this correlation is observed because a split organization actually offers an adaptation to thermophilic lifestyle.

In *chapter 3* we introduce and discuss genome phylogenies. The apparent ubiquity of HGT suggests that the correspondence between the evolution of gene content and of the species might be low or non existent (Doolittle 1999). On the other hand, quantitative studies suggest that the number of shared genes correlates with evolutionary closeness (Huynen and Bork 1998). We here explicitly probe shared gene content for a phylogenetic signal, by constructing a genome tree based on shared genes. We thereby find a good correspondence between the obtained tree and known phylogenies from other sources. Subsequently we discuss the relevance of this work for defining the tree of life, and even for answering whether such a thing as a species phylogeny is feasible. Finally we introduce a web server, SHOT, that makes the construction of genome trees with a diverse set of parameters and species, available to the general community for which such computationally intensive research otherwise would not be possible. The usefulness of the web server is demonstrated by discussing genome trees obtained from a recent comprehensive set of species.

In *chapter 4* we present an integrated approach to reconstruct which genes were present in the Archaeal and Proteobacterial ancestral genomes and how ancestral and present day genomes have been shaped by the processes of gene loss, gene duplication, horizontal
gene transfer (HGT), gene fusion/fission, and gene genesis. In chapter 3 we present a classification of complete genomes. Here we use the thereby obtained tree to actually interpret the presence and absence patterns in terms of genome evolutionary events. The reconstruction suggests that the ancestor of the Proteobacteria contained around 2500 genes, and the ancestor of the Archaea around 2050 genes. Although it is necessary to invoke horizontal gene transfer to explain the content of present day genomes, gene loss, gene genesis, and simple vertical inheritance are quantitatively the most dominant processes in shaping the genome. Together they result in a turnover of gene content such that even the lineage leading from the ancestor of the Proteobacteria to the relatively large genome of *Escherichia coli* has lost at least 950 genes. Gene loss, unlike the other processes, correlates fairly well with time. This clock-like behavior suggests that gene loss is under negative selection, while the processes that add genes are under positive selection.

The repeated occurrence of genes in each other's neighborhood on genomes has been shown to indicate a functional association between the proteins they encode. Since we have been heavily participating in finding the basic patterns of genomic associations, and benchmarking these for function prediction (Huynen and Snel 2000), as well as co-pioneering the use of conserved gene order for function prediction, we introduce in chapter 5 STRING, a Search Tool for Recurring Instances of Neighbouring Genes. STRING is a web server that allows the retrieval and display of the genes a query gene repeatedly occurs with in clusters on the genome. It performs iterative searches and visualizes the results in their genomic context. By finding the genomically associated genes for a query, it delineates a set of potentially functionally associated genes. The usefulness of STRING is illustrated with an example that suggests a functional context for an RNA methylase with unknown specificity.

In chapter 6, we present an analysis of the complete network of genomic associations derived from conserved gene order with the aim of delineating functional modules: sets of proteins that functionally interact. Associations obtained from conserved co-occurrence of two genes within operons indicate a functional interaction between their products. However, many genes end up being indirectly linked to each other. This trend is likely to only get worse with more genomes. We therefore study the properties of the network. Analysis of the giant component reveals that it is a scale-free, small world network with a high degree of local clustering. It consists of locally highly connected subclusters that are connected to each other by linker proteins. By splitting up the giant component at these linker proteins we identify subclusters that tend to have a homogeneous functional composition. It is thereby shown that comparative genome analysis allows the identification of a natural classification of proteins that is complementary to those based on molecular function.

Finally in chapter 7, we provide a summarizing and synthesizing discussion of the chapters presented in this thesis. We moreover describe and summarize a few new and parallel developments that provide the arising context for our results. Partly, these developments are also described because they solve some of the issues that are raised here, or present promising approaches in comparative genome analysis in general.