UNIVERSITY OF TARTU

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Institute of Computer Science

Ilja Livenson

Using Comparative Genomics for Prediction of Regulatory Elements

Bachelor Thesis (4 cp)

Supervisor: Jaak Vilo, Ph. d.

Author: .... "....." May 2006

Supervisor: .... "....." May 2006

Chair: .... "....." 2006

TARTU 2006

Contents

List of Figures

Introduction

> Quidquid latine dictum sit,
> altum sonatur.
>
> Worldly Wisdom

Comparative genomics is a relatively new field that complements a long history of comparison-based disciplines in biology. The recent development of a large dataset of vertebrate genomic sequences (in particular, the recent release of the report of the sequence of human chromosome 1, which is the final chromosome analysis from the Human Genome Project) has aided in global gene predictions as well as in the identification of sequences important in gene regulation. In addition, vertebrate comparative sequence analysis has contributed to the exploration of the genetic bases for differences and similarities among species. In combination with areas of study such as comparative physiology or comparative biochemistry, one might finally manage to understand the genetic explanation for how species have adapted to perform their shared or unique biological functions.

This thesis gives an overview of the area of comparative genomics, describes the employed approaches and provides several references to the implementation of these approaches. As a practical part of the thesis there has been written a tool to make the usage of the scientific tools (in particular, bioinformatics tools) on the BalticGrid system easier by automating the installation process of the software.

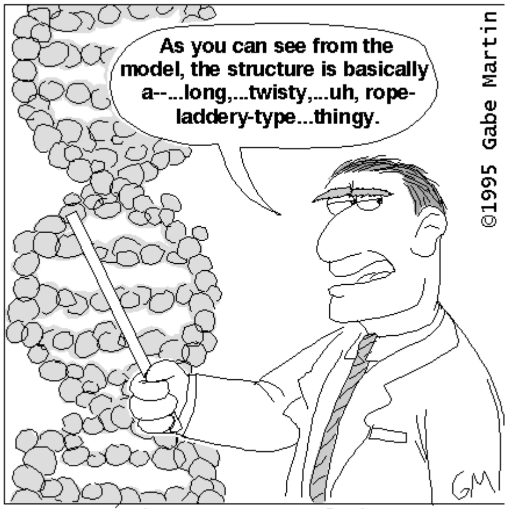
The main text of the thesis consists of four chapters. The first chapter provides a brief biological background required for the understanding of the thesis, the second gives an overview of the computational strategies used for identifying regulatory sequences. The third chapter is dedicated to the review of two specific methods of the comparative genomics (phylogenetic footprinting and shadowing) and tools and databases that provide the functionality needed for the actual prediction of the regulatory elements. The last chapter is the short introduction to the Grid systems with explanation of the software installation process on Grid.

Chapter 1

Biological Background

1.1 What is DNA

Deoxyribonucleic acid (DNA) is a nucleic acid - usually in the form of a double helix - that contains the genetic instructions specifying the biological development of all cellular forms of life, and most viruses. DNA is a long polymer of nucleotides and encodes the sequence of the amino acid residues in proteins using the genetic code, a triplet code of nucleotides [Wik06a].


As you can see from the model, the structure is basically a--...long,...twisty,...uh, rope-laddery-type...thingy.
©1995 Gabe Martin
GM

1953: The structure of the DNA molecule is first described

DNA is responsible for the genetic propagation of most inherited traits. In humans, these traits can range from hair colour to disease susceptibility. During cell division, DNA is replicated and can be transmitted to offspring during reproduction.

DNA is not a single molecule, but rather a pair of molecules joined by hydrogen bonds: it is organized as two complementary strands, head-to-toe, with the hydrogen bonds between them. Each strand of DNA is a chain of chemical "building blocks", called nucleotides, of which there are four types: adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). These allowable base components of nucleic acids can be polymerized in any order giving the molecules a high degree of uniqueness [Wik06a].

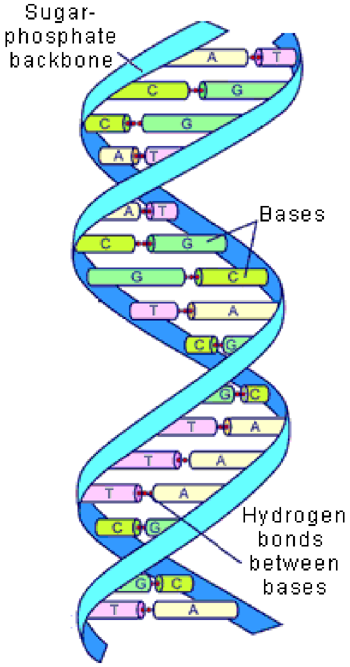


Figure 1.1: DNA Model (Image taken from [Uss98])

1.2 Replication, Transcription, Translation

The central dogma of molecular biology was first formulated by Francis Crick ([ea06a] describes it in more detail) in 1958 and re-stated in a Nature paper published in 1970. It can be stated in a very short and oversimplified manner as “DNA makes RNA makes proteins, which in turn facilitate the previous two steps as well as the replication of DNA”, or simply “DNA > RNA > protein”. This process is therefore broken down into three steps: replication, transcription and translation.

The double-stranded structure of DNA provides a simple mechanism for DNA replication: the DNA double strand is first “unzipped” down the middle, and the “other half” of each new single strand is recreated by exposing each half to a mixture of the four bases. An enzyme makes a new strand by finding the correct base in the mixture and pairing it with the original strand. In this way, the base on the old strand dictates which base will be on the new strand, and the cell ends up with an extra copy of its DNA [Wik06a].

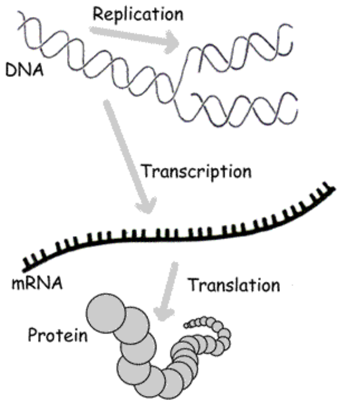


Figure 1.2: “The Central Dogma of Biology” (Image taken from [Uss98])

Transcription is the process through which a DNA sequence is copied by an RNA polymerase to produce a complementary RNA. Or, in other words, the transfer of genetic information from DNA into RNA. In the case of protein-encoding DNA, transcription is the beginning of the process that ultimately leads to the translation of the genetic code (via the mRNA intermediate [Wik06b]) into a functional peptide or protein. Transcription has some proofreading mechanisms, but they are fewer and less effective than the controls for DNA; therefore, transcription has a lower copying fidelity than DNA replication [Wik06d].

Translation is the process of making proteins by joining together amino acids in order encoded in the mRNA. [Wik06e].

1.3 Gene Regulation

DNA sequencing technology has led to a vast data base of unexplored sequence data. The human genome project and other genome projects have greatly accelerated the growth of this data base. Since genes code for proteins, the action molecules of life, the bulk of scientific effort has focused on the genes and their products. Yet only about 3% of the human genome codes for genes. The remainder, sometimes called “junk DNA” has received somewhat less attention. Recent findings suggest that it may be far more important than previously believed. In particular this “junk” has been shown to contain important regulatory signals [ABL+02].

Gene regulation is the fundamental process behind cellular alteration. It is used by single cell organisms to respond to changes in their environment and by multicellular organisms for cell differentiation. The most studied gene regulation involves the binding of regulator proteins to “regulatory elements” which are signal sequences that normally occur in an upstream fragment of the genome called the promoter. While there are experimental methods to identify regulator elements, they are time consuming and difficult, and thus not particularly well suited to examine the vast data base of sequences that is emerging from the genome sequencing projects. Thus there is a need for the development of computational and statistical methods to explore these databases to identify novel regulatory signals that may be hidden within non-coding DNA [ABL+02].

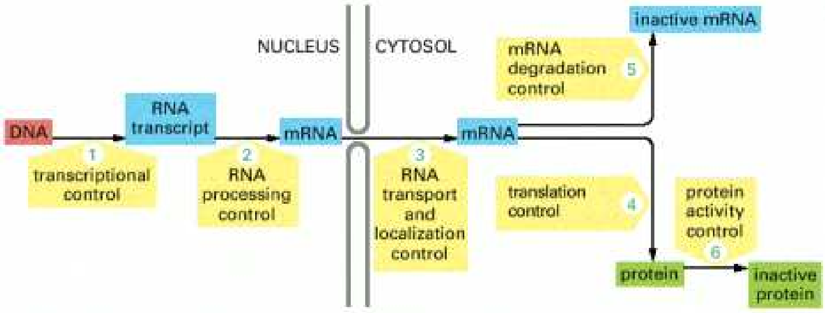


Figure 1.3: Six steps at which eucaryotic gene expression can be controlled. Image taken from [ABL+02]

Chapter 2

Strategies for Identifying Regulatory Sequences

Classical searches for *cis*-regulatory sequences (i.e. sequences co-located on the same chromosome of a homologous pair) have typically involved various trial-and-error strategies. The focus on the identification of regulatory elements for individual genes has included several experimental approaches: the generation of deletion constructs to determine the minimal sequences necessary for transcription in cell-culture-based systems; *DNaseI hypersensitivity studies*[1] to identify sequences potentially available for transcription factor binding; and *in vitro* approaches, such as *DNase footprinting*[2] and *Gel shifts*[3], to determine sequences that bind various regulatory proteins. Screens to identify *cis*-regulatory elements have also been carried out in transgenic mice, albeit in an extremely laborious and low-throughput manner. In addition, a limited number of large-scale promoter and enhancer trapping studies have been done. Most of these gene regulatory studies have consisted of

[1] Identifies regions of the genome that lack nucleosome structure and are therefore readily degraded by the enzyme DNaseI. Such regions tend to be associated with transcriptional activity.

[2] An assay that identifies a region of DNA that is protected from digestion by DNaseI (usually due to the binding of a protein, such as a transcription factor).

[3] A gel-based assay in which proteins that bind to a DNA fragment are detected by virtue of the reduced migration of the DNA. The assay is often used to detect transcription factor binding.

largely unguided searches of genomic sequence for those with gene regulatory properties [Har03].

Computational sequence analysis provides three broadly different approaches for scanning genomic sequence to identify those regions predicted to participate in gene regulation. The first approach for predicting regulatory elements is sequence analysis of co-regulated genes within a species. This strategy is based on the fact that few transcription factors influence only single genes; rather, most bind to conserved sites in several genes to coordinate their expression. Accordingly, genes are thought to be co-regulated because they respond to similar regulatory pathways owing to shared noncoding sequence motifs that direct the binding of specific sets of shared transcription factors. The second approach for the identification of gene regulatory sequences involves generating and analysing databases of known transcription factor binding sites and characterizing promoter regions. Finally, inter-species sequence comparisons have been used to identify non-coding sequences that have a reasonable likelihood of having gene regulatory properties. This is possible because sequences that mediate gene expression tend to be conserved between species.

2.1 Short Overview of Computational Methods

2.1.1 Sequence Analysis of Co-regulated Genes within a Species

This approach can be called called “multiple genes, single species” approach. It proposes that a degenerate motif is embedded in some or all of the otherwise unrelated input sequences and tries to describe a common motif and identify its occurrences ([BJVU98, vHACV98, CLLL03]). It is often used for co-regulated genes identified through experimental approaches.

2.1.2 Generating and Analysing the TFBS

Within this approach different data mining algorithms are applied to the sets of transcription factor binding sites (TFBS) in hope to detect TFBS specificities and use them in *de novo* site discovery ([BVUV97, Lon04]).

2.1.3 Inter-species Sequence Comparisons

In this case genomic data of different organisms is analyzed. The general idea is that TFBS and regulatory elements should be well conserved among different species. The potential functions of conserved non-coding sequences are numerous, and include roles in chromosomal assembly and replication as well as gene regulation. Compelling support for the conservation of sequence-based regulatory information across species comes from a diverse set of experimental approaches. Most importantly, this support includes the DNA sequence conservation of experimentally defined regulatory elements among mammals.

The latter approach is reviewed further in this work with examples of databases and tools for analysis.

Chapter 3

Comparative genomics

A complete genome sequence of an organism can be considered to be the ultimate genetic map, in the sense that the heritable characteristics are encoded within the DNA and that the order of all the nucleotides along each chromosome is known. However, knowledge of the DNA sequence does not tell us directly how this genetic information leads to the observable traits and behaviors (phenotypes) that we want to understand [PR01].

The major principles of comparative genomics are straightforward. It is the analysis and comparison of genomes from different species. The purpose is to gain a better understanding of how species have evolved and to determine the function of genes and noncoding regions of the genome. Common features of two organisms will often be encoded within the DNA that is conserved between the species. More precisely, the DNA sequences encoding the proteins and RNAs responsible for functions that were conserved from the last common ancestor should be preserved in contemporary genome sequences. Likewise, the DNA sequences controlling the expression of genes that are regulated similarly in two related species should also be conserved. Conversely, sequences that encode (or control the expression of) proteins and RNAs responsible for differences between species will themselves be divergent.

Researchers have learned a great deal about the function of human genes by examining their counterparts in simpler model organisms such as the mouse. Genome researchers look at many different features when comparing genomes: sequence similarity, gene location, the length and number of coding regions (called exons) within genes, the amount of noncoding DNA in each genome, and highly conserved regions maintained in organisms as simple as bacteria and as complex as humans [Pro06a].

Comparative genomics involves the use of computer programs that can line up multiple genomes and look for regions of similarity among them.

Mice and humans (and most or all mammals including dogs, cats, rabbits, monkeys, and apes) have roughly the same number of nucleotides in their genomes - about 3 billion base pairs. This comparable DNA content implies that all mammals contain more or less the same number of genes.

Different questions can be addressed by comparing genomes at different phylogenetic distances (figure 3). Broad insights about types of genes can be gleaned by genomic comparisons at very long phylogenetic distances, e.g., greater than 1 billion years since their separation. For example, comparing the genomes of yeast, worms, and flies reveals that these eukaryotes encode many of the same proteins, and the non-redundant protein sets of flies and worms are about the same size, being only twice that of yeast.

The more complex developmental biology of flies and worms is reflected in the greater number of signaling pathways in these two species than in yeast. Over such very large distances, the order of genes and the sequences regulating their expression are generally not conserved. At moderate phylogenetic distances (roughly 70-100 million years of divergence), both functional and nonfunctional DNA is found within the conserved DNA. In these cases, the functional sequences will show a signature of purifying or negative selection, which is that the functional sequences will have changed less than the nonfunctional or neutral DNA. Not only does comparative genomics aim to discriminate conserved from divergent and functional from nonfunctional DNA, this approach is also contributing to identifying the general functional class of certain DNA segments, such as coding exons, noncoding RNAs, and some gene regulatory regions. Examples of analyses at this distance include comparisons among enteric bacteria, among several species of yeast, and between mouse and human. In contrast, very similar genomes, such as those of humans and chimpanzees (separated by about 5 million years of evolution), are particularly well for finding the key sequence differences that may account for the differences in the organisms. These are sequence changes under positive selection. Comparative genomics is thus a powerful and prospering discipline that becomes more and more informative as genomic sequence data accumulate [Har03].

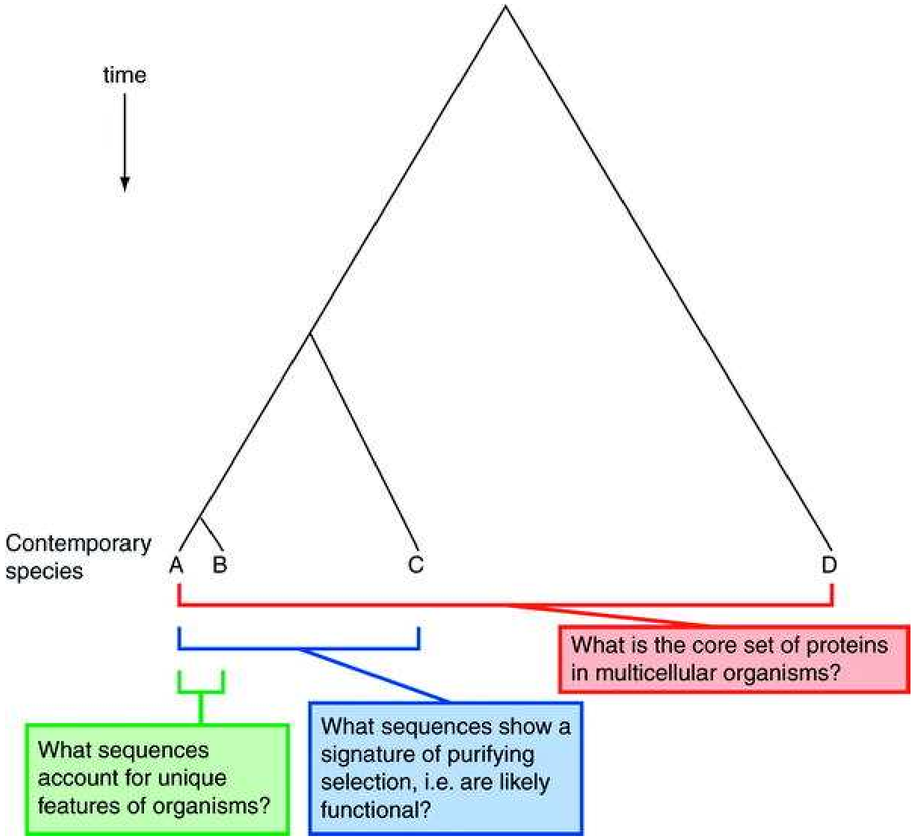

time
Contemporary species
A
B
C
D
What is the core set of proteins in multicellular organisms?
What sequences show a signature of purifying selection, i.e. are likely functional?
What sequences account for unique features of organisms?

Figure 3.1: A generalized phylogenetic tree is shown, leading to four different organisms, with A and D the most distantly related pairs. Examples of the types of questions that can be addressed by comparisons between genomes at the different distances are given in the boxes (Image taken from [Har03]).

Information on sequence similarity among genomes is a major resource for finding functional regions and for predicting what those functions are.

3.1 Phylogenetic footprinting

3.1.1 Idea

Since co-regulated genes are recognized by the same transcription factor, they must have the same motif in their upstream sequences. Thus, in principle, we can expect to understand the co-regulation relationship of genes by finding their common motif. However, it is generally difficult to find unknown motifs because they are usually short and are not strictly conserved. One way to reduce unavoidable noises is to use the evolutionary information. Namely, one can postulate that most binding sites of transcription factors are phylogenetically conserved (the method is often called the *phylogenetic footprinting*).

The simple idea underlying phylogenetic footprinting is that selective pressure causes functional elements to evolve at a slower rate than that of nonfunctional sequences. This means that unusually well conserved sites among a set of orthologous regulatory regions are excellent candidates for functional regulatory elements. This approach has proved successful for the discovery of regulatory elements for many genes, including ϵ-*globin* ([TKG$^+$88]), γ-*globin* ([TKG$^+$88]), *rbcL* ([MN95]) and many others.

The major advantage of phylogenetic footprinting over the single genome, multi-gene approach mentioned earlier is that the latter requires a reliable method for assembling the requisite collection of co-regulated genes. In contrast, phylogenetic footprinting is capable of identifying regulatory elements specific even to a single gene, as long as they are sufficiently conserved across many of the species considered. Genome projects are quickly producing sequences from a wide variety of organisms, so the data necessary for phylogenetic footprinting are becoming increasingly available.

The standard method that has been used for phylogenetic footprinting is to construct a global multiple alignment of the orthologous regulatory sequences and then to identify conserved regions in the alignment ([Bla01] has a nice overview of these methods).

To see why this approach to phylogenetic footprinting does not always work, consider typical lengths of the sequences involved. Regulatory elements tend to be quite short (5 to 20 nucleotides long) relative to the entire

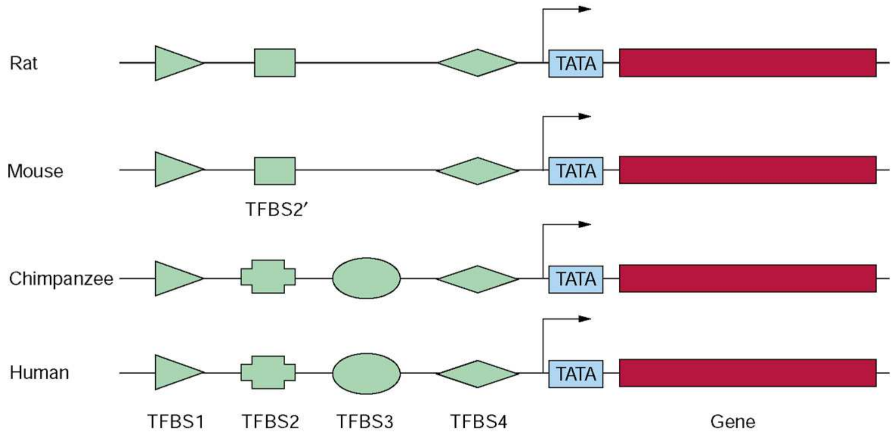
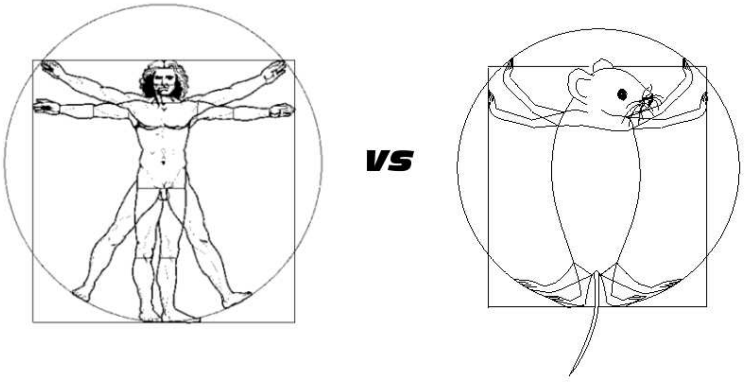


Figure 3.2: Using phylogenetic footprinting to detect conserved TFBSs. This schematic diagram shows a hypothetical human gene aligned with its orthologs from three other mammals. Cross-species sequence comparison reveals conserved TFBSs in each sequence. Sequence motifs of the same shape (colored in green) represent binding-sites of the same class of transcription factors. TFBS1 and TFBS4 are conserved in all four mammals; TFBS3 represents a newly acquired, primate-specific binding site. TFBS2 and TFBS2' represent orthologous regulatory sites that have diverged significantly between the primate and rodent lineages. Blue rectangles represent TATA boxes. (Image taken from [ZG03]).

regulatory region in which we search for them (a 1000-bp promoter region would be typical). Given these relative lengths, if the species are somewhat diverged, it is likely that the noise of the diverged nonfunctional background will overcome the short conserved signal. The result is that the alignment may not align the short regulatory elements together. In that case, the regulatory elements would not appear to belong to conserved regions and would go undetected. Thus, when the entire regulatory regions considered are moderately to highly diverged, global multiple alignment is likely to miss significant signals.

Cliften *et al.* (2001) made similar observations in conjunction with their comparative analysis of several *Saccharomyces* species. They discovered that if the species are too closely related, the sequence alignment is obvious but uninformative, because the functional elements are not sufficiently better conserved than the surrounding nonfunctional sequence. On the other hand, if the species are too distantly related, it is difficult or impossible to find an accurate alignment ([MB02]). That is why most of the algorithms and tools described below use some modifications of the straight-forward approach incorporating some additional biological knowledge.

3.1.2 Of Mice and Men



The inherent problem of comparative genomics is the question which species should be compared with each other to most reliably identify functional regions in the genomic sequence.

Human-mouse comparisons have been used extensively to identify potential regulatory regions which in many cases proved to be functional. Human and mouse diverged 75-90 million years ago. The divergence rate between their genomes has been low enough that one can still align orthologous sequences, but high enough to allow the discrimination of functional elements by their greater conservation. To assess the usefulness of human-mouse comparisons, several previous studies ([der, LSM+03, LH02]) addressed the point to what extent experimentally known TFBSs can be identified by phylogenetic footprinting. The data collections of these studies comprehended between 99 and 481 TFBSs of which about 60-68% could be detected by human-mouse comparisons. A more recent study on the human-rodent comparison is done in Sauer *et al* ([TS06]) with approximately the same results.

3.2 Phylogenetic shadowing

In contrast to footprinting, “phylogenetic shadowing” examines sequences of closely related species and takes into account the phylogenetic relationship of the set of species analyzed. This approach enabled the localization of regions of collective variation and complementary regions of conservation, facilitating the identification of coding as well as noncoding functional regions ([NP03, BGvdB$^+$05]).

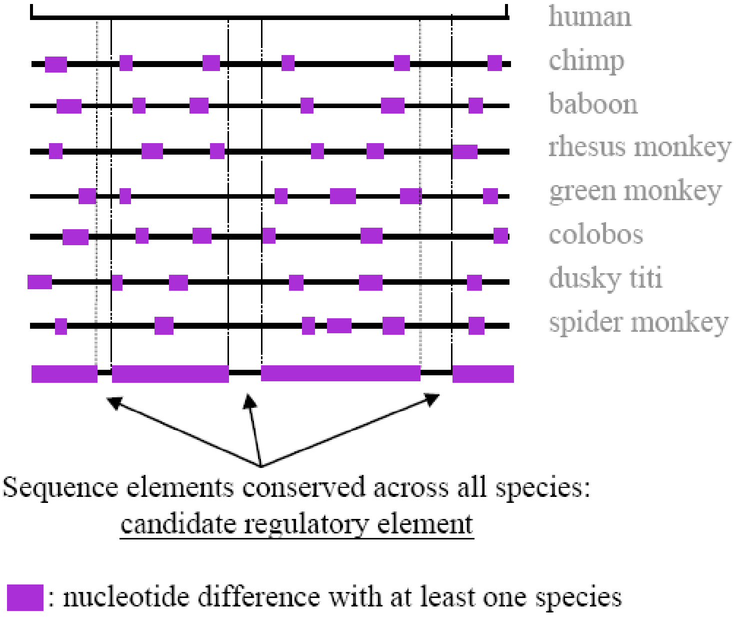

human
chimp
baboon
rhesus monkey
green monkey
colobos
dusky titi
spider monkey
Sequence elements conserved across all species:
candidate regulatory element
: nucleotide difference with at least one species

Figure 3.3: By sequencing specific segments within each of the genomes of the different primates being analyzed, the researchers found enough small differences between genomes in the non-human primates that could be combined to create a phylogenetic ”shadow”, which could then be compared to the human genome. (Image taken from [NP03]).

As a proof of principle, “phylogenetic shadowing” proved successful for the identification of both exons as well as putative gene regulatory elements ([B$^+$03]). In this study, 13-17 primate sequences of several orthologous genomic segments were generated and compared. For a single exon from four independent genes, highly conserved “shadows” matched strongly with these functionally important protein-encoding regions. In addition, analysis of the human apolipoprotein gene (*apo(a)*) revealed highly conserved intervals embedded within the upstream promoter region, and functional studies of these “phylogenetic shadows” compared to more variable flanking DNA supported their role in regulating *apo(a)* expression ([B$^+$03]). The success of this approach suggests that a genome-wide comparison of a handful of primate species will aid in the identification of both human exons and gene regulatory elements.

3.3 Existing Tools and Databases

To actually use the approaches described above one would need the tools and test inputs taken from the databases. It’s up to the researcher to select which tool to use (or perhaps improve) and which data type will be used for research/validation of the algorithm. Below is short overview of some major databases and tools that one could use for comparative analysis.

3.3.1 Databases

GenBank

GenBank is a comprehensive sequence database that contains publicly available DNA sequences for more than 119 000 different organisms, obtained primarily through the submission of sequence data from individual laboratories and batch submissions from large-scale sequencing projects. Most submissions are made using the BankIt (web) or Sequin programs and accession numbers are assigned by GenBank staff upon receipt. Daily data exchange with the EMBL Data Library in the UK and the DNA Data Bank of Japan helps ensure worldwide coverage. GenBank is accessible through NCBI’s retrieval system, Entrez, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. BLAST provides sequence similarity searches of GenBank and other sequence databases. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. To access GenBank and its related retrieval and analysis services, go to the NCBI home page at: http://www.ncbi.nlm.nih.gov

There are approximately 59,750,386,305 bases in 54,584,635 sequence records in the traditional GenBank divisions and 63,183,065,091 bases in 12,465,546 sequence records in the WGS division as of February 2006.

The number of eukaryote genomes for which both coverage and assembly are good is increasing rapidly and now includes *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Plasmodium falciparum*, *Caenorhabditis elegans*, *Drosophila melanogaster*, Anopheles gambiae, *Arabidopsis thaliana*, *Mus musculus* and *Homo sapiens*.

Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references, and a table of features listing areas of biological significance, such as coding regions and their protein translations, transcription units, repeat regions, and sites of mutations or modifications.

Each GenBank record, consisting of both a sequence and its annotations, is assigned a stable and unique identifier, the accession number. The accession number remains constant over the lifetime of the record even when there is a change to the sequence or annotation. Each DNA sequence in GenBank is assigned another unique identifier, called a "gi". The gi numbers appear on the VERSION line of GenBank records following the accession number. When a change is made to a sequence given in a GenBank record, a new gi number is assigned to the new sequence version associated with the record while the accession number for the record remains unchanged. The older sequence version retains the old gi. [B$^+$04]

ENCODE

ENCODE, or the Encyclopedia Of DNA Elements, is a project started in September 2003 by the National Human Genome Research Institute (NHGRI), to identify all functional elements in the human genome sequence.

For use in the ENCODE Pilot Project, defined regions of the human genome - corresponding to 30Mb, roughly 1 percent of the total human genome - have been selected. These regions serve as the foundation on which to test and evaluate the effectiveness and efficiency of a diverse set of methods and technologies for finding various functional elements in human DNA.

A component of ENCODE data production involves the generation of sequencing information from a number of different genomes in order to extract the maximum amount of information about the human genome through comparative analyses. Efforts are already underway at the NHGRI, University of British Columbia and the NIH Intramural Sequencing Center to identify, map and sequence, respectively; BAC clones for regions syntenic to the human ENCODE targets will be made in additional mammalian species. In addition to these ENCODE-directed efforts, sequence data generated through whole genome sequencing projects will be used in comparative analyses to help scientists better understand the human sequence.

Galaxy

Galaxy is a genome alignment and annotation database ([GRH$^+$05]). It features connections to UCSC Table Browser, EnsMart, and contains hundreds of tools. Some examples:

- extracting multiple alignments corresponding to a genomic region;
- finding exons overlapping SNPs, computing phastCons scores for a set of genomic ranges;
- building histograms, computing correlations, drawing scatterplots.

dbERGE II

dbERGE II stores experiment and result details for various types of experiments ([EGS$^+$05]). The current types of experiments available are:

- DNA transfer experiments (Transfections and Transgenic mice)

- Binding assays (Gel shift, *in vivo* footprint, *in vitro* footprint and Methylation interference)
- Hypersensitive sites
- ChIP-on-chip experiments

OMGProm

OMGProm is a database of orthologous mammalian gene promoters, which can be use as a platform for comparative genomics of transcriptional regulation, in order to facilitate the identification of gene regulatory elements such as core promoters and transcription factor binding sites that are conserved in the upstream regions of orthologous genes.

Extensive molecular research in the field of transcription regulation has produced invaluable promoter sequence data that are being deposited into GenBank. In parallel, recent advances in sequencing technologies have generated full-length cDNAs of mammalian genes. The OMGProm team has systematically integrated the cDNA and genome sequence data and curated a set of 8550 promoters of orthologous mammalian genes. This data repository might be a valuable control set for designing novel promoter prediction tools and for testing the sensitivity of existing programs, such as FirstEF. The database serves to complement similar databases such as DBTSS ([SYNS02]) and PromoSer ([HLW03]).

3.3.2 Tools

The main tools used for comparative genomics are the sequence alignment tools. Basically, it is an arrangement of two or more sequences, highlighting their similarity. The sequences are padded with gaps (usually denoted by dashes) so that wherever possible, columns contain identical or similar characters from the sequences involved. Mismatches in the alignment correspond to mutations, and gaps correspond to insertions or deletions ([Wik06c]).

One can speak of two main types of alignments: global and local. A global alignment between two sequences is an alignment in which all the characters in both sequences participate in the alignment. Global alignments are useful mostly for finding closely-related sequences. As these sequences are also easily identified by local alignment methods global alignment is now somewhat deprecated as a technique. Local alignment methods find related regions within sequences - in other words they can consist of a subset of the characters within each sequence. For example, positions 20-40 of sequence A might be aligned with positions 50-70 of sequence B.

This is a more flexible technique than global alignment and has the advantage that related regions which appear in a different order in the two proteins can be identified as being related. This is not possible with global alignment methods.

BLAST

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families ([AGM$^+$90]).

This method uses a pre-computed hash table to serve as an index for short sequences. Given a query sequence, the sub-sequences are looked up in the index to reduce the amount of time and searching involved. Several parameters need to be provided to make this method faster or more accurate. Once patterns that match the search sequence are found, more accurate and intensive algorithms may be applied.

BLAST uses a pairwise local search and uses a number of methods to increase the speed of the original Smith-Waterman algorithm ([SW81]).

Mulan

Mulan performs local multiple DNA sequence alignments of finished and draft-quality sequences. It identifies transcription factor binding sites evolutionarily conserved across multiple species.

Mulan brings together several novel algorithms: the TBA multi-aligner program for rapid identification of local sequence conservation, and the multiTF program for detecting evolutionarily conserved transcription factor binding sites in multiple alignments.

The Mulan alignment engine consists of several data analysis and visualization schemes for high-throughput identification of functional coding and noncoding elements conserved across large evolutionary distances. Mulan determines phylogenetic relationships among the input sequences and generates phylogenetic trees, constructs graphical and textual alignments, dynamically detects evolutionary conserved regions (ECRs) in alignments, and presents users with several visual display options for the generated conservation profiles. This tool is also able to implement the phylogenetic shadowing strategy for identifying slow-mutating elements in comparisons of multiple closely related species.

Mulan employs two alignment strategies that allow for comparative analysis of multiple sequences that are present either as draft or finished configuration. The first approach (threaded blockset aligner, TBA) allows for the construction of an alignment for multiple draft-quality sequences and subsequently for effective order-and-orientation (O&O) of unfinished sequences based on the reference genome.

TBA builds a threaded blockset under the assumption that all matching segments occur in the same order and orientation in the given sequences; inversions and duplications are not addressed.

A block is a rectangular array of symbols such that removing dashes from any row produces a run of one or more consecutive positions in one of the original sequences or their reverse complements. A block can consist of only one row, but no column of a block may consist entirely of dashes (indicating insertions or deletions). A set of such blocks is called a blockset. A “ref-blockset” consists of a blockset in which every block has a designated row, all of which come from the same original sequence, called the reference for that ref-blockset.

A given sequence, S, is said to ”thread” a blockset if every position in S appears precisely once in some block of the blockset. Thus, a ref-blockset is threaded by the reference sequence. If a blockset is threaded by each of the original sequences, it is called a threaded blockset.

Given a threaded blockset, one can generate an S-ref blockset for any original sequence S, an operation called ”projecting onto S“. One merely picks the blocks having a row from S and orders them according to S. In practice one moves the reference row to the top of each block. See Figure 1 for an example. A critical property of projections is that any two ref-blocksets generated by projection from the same threaded blockset are consistent. More precisely, if position x of sequence X aligns to position y of sequence Y in one projection and to position z of Y in another projection, then $y = z$.

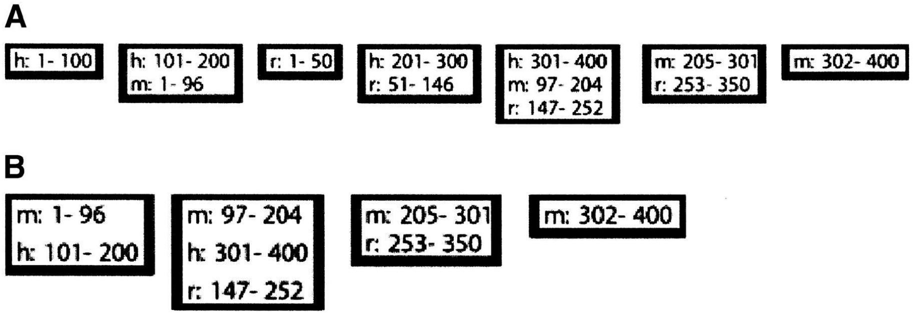


Figure 3.4: (A) Blocks (alignments) of a hypothetical threaded blockset for sequences human (400 bp), mouse (400 bp) and rat (350 bp). Only the range of positions in each alignment is given. (B) Projection of the threaded blockset onto mouse. (Image taken from [BKR$^+$04]).

The second approach operates with multiple high-quality finished sequences.

The ability to accurately predict functional transcription factor binding sites (TFBS) is a powerful approach for sequence-based discovery of gene regulatory sequences and for elucidating gene regulation networks and mechanisms. To combat the overabundance of false-positive computational predictions stemming predominantly from the small size of TFBS footprints and from poorly defined position weight matrices (PWM), evolutionary sequence analysis has been proposed as a robust strategy for filtering out false-positive sites. Mulan incorporates a TFBS analysis tool, multiTF that is similar to pairwise alignment-based rVista 2.0, but implements a different method of detecting TFBS present in all the sequences included in the multiple alignment ([OLG$^+$05]).

CompareProspector

CompareProspector is a sequence motif-finding algorithm, which extends Gibbs sampling by biasing the search in promoter regions conserved across species. Using human-mouse comparison, CompareProspector correctly identified the known motifs for transcription factors *Mef2*, *Myf*, *Srf*, and *Sp1* from a set of human muscle-specific genes. CompareProspector outperformed many other computational motif-finding programs tested, demonstrating the power of comparative genomics-based biased sampling in eukaryotic regulatory element identification ([LLW$^+$04]).

It is built on BioProspector ([LBL01]), which is an extension of the original Gibbs Sampler ([LNL95]).

FootPrinter

Footprinter's approach is based on the simple formalization of the phylogenetic footprinting idea, the Substring Parsimony Problem (SPP).

Given a set of homologous input sequences and the phylogenetic tree T relating them, the algorithm identifies every set of $k$mers, one from each input sequence, that have parsimony score at most d with respect to T, where k and d are parameters specified by the user.

Or, to be a bit more formal:

Given: a set of orthologous sequences $S_1, \ldots, S_n$ from n different species, the phylogenetic tree T relating these species, the size k of the motifs to look for, and an integer d.

Problem: find all sets of substrings $s_1, \ldots, s_n$ of $S_1, \ldots, S_n$ respectively, each of size k, such that the parsimony score of $s_1, \ldots, s_n$ on T is at most d.

The parsimony score is the minimum number of nucleotide substitutions along the branches of T that explain the set of identified $k$mers. It is computed as the minimum, over all possible labelings of the internal nodes with sequences of size k, of the sum of the Hamming distances between the labels of nodes connected by an edge in T. Looking for sets of substrings that achieve a low parsimony score corresponds to searching for highly conserved regions.

Solution: Solution uses dynamic programming approach. The algorithm assumes a rooted tree, so one should first select a root (arbitrary internal node r). The algorithm then proceeds from the leaves up to the root. At each node u of the tree, a table W_u is computed containing 4^k entries, one for each sequence of size k. For a string s of size k, one defines $W_u[s]$ as the best parsimony score that can be achieved for the subtree rooted at u, if u was to be labeled with s (i.e. if the ancestral sequence at u is forced to be s). Let $C(u)$ be the set of children of u, let $d(s,t)$ be the Hamming distance between sequences s and t, and let $\sum = \{A, C, G, T\}$. The tables W can be computed recursively:

$$W_u[S] = \begin{cases} 0 & \text{if } u \text{ is a leaf, } s \subset S_u; \\ +\infty & \text{if } u \text{ is a leaf, } s \not\subset S_u; \\ \sum_{v \in C(u)} \min_{t \in \sum^k} (W_v[t] + d(s,t)) & \text{if } u \text{ is not a leaf} \end{cases}$$

Then, the score of the optimal solution to the SPP is given by $min_{s \in \sum^k}(W_r[s])$. From that point, the ancestral sequences $s_{n+1}, \ldots, s_{|V|}$ and substrings $s_1, \ldots, s_n$ can be recovered by tracing back the recurrence, from the root down to the leaves, for each entry of W_r with score at most d [Bla01, MB03].

Footer

Footer is a novel phylogenetic footprinting algorithm, which combines two statistics in order to score a pair of putative regulatory sites. Given two homologous promoter sequences and a number of putative motifs identified in each of them (by default Footer retains one top scoring motif per TF per 300 bp of promoter sequence), the method performs all pairwise comparisons of the motifs.

A scoring scheme based on two statistics has been employed. The first statistic scores a pair of motifs according to their position conservation in the sequence. The second statistic scores the pair of motifs according to their agreement with the corresponding position weight matrices (PWM) model(s). A PWM model is the most commonly used way to represent the binding preferences of a TF. These models are generated from known mammalian binding sites from the TRANSFAC database.

Typically, a set of aligned sequences is used to calculate a $4xL$ weight matrix (L is the length of the pattern). In each column, the weights correspond to the log-likelihood of the preferences of the TF to each of the four bases (sometimes normalized for the background).

The two statistics Footer employs consist of the P-values of the observed data, under the assumption that the two sites are unrelated. The position-related score is calculated using the following formula:

$$PF_D = P(D_{XY} \leqslant d) = \frac{1}{N} + \sum_{k=1}^{d} \frac{2(N-k)}{N^2}$$

where D_{XY} is the random variable denoting the distance between two putative sites, d is the observed distance of the particular putative sites (measured from the 3' closest conserved region boundary), N is the effective promoter length (i.e. the promoter length minus $L-1$, where L is the length of the pattern). The equation calculates the tail probability that two high-scoring "signals" will be found by chance at a distance d or less in the promoter with effective length N.

The PWM-related score is calculated using the following formula:

$$PF_S = P[(S+T) < (s+t)|M_1, M_2]$$

where M_1 and M_2 are the PWM models for the two species; S and T are random variables following the models' score distributions; and s and t are the observed PWM scores. The PF_S score is calculated using Gaussian approximation of mean and standard deviation estimated through random samplings from the PWM model distributions. The results of the samplings are stored in each model. Similarly to PF_D the latter equation calculates the corresponding tail probability under the assumption that the two high-scoring "hits" are due to chance alone. The combined score, PF, consists of a weighted log-likelihood transformation:

$$PF = -w_D * log(PF_D) - w_S * log(PF_S)$$

.

The weights w_D and w_S are positive numbers that sum to one (default values of the Footer implementation: $w_D = 0.85$; $w_S = 0.15$).

The sample output of Footer is presented in appendix A. [DLC05].

rVista

rVista tool combines pattern recognition with comparative sequence analysis to dramatically reduce the number of false positive TFBS matches and enrich for functional sites. It is a tool that interconnects TFBS motif searches and cross-species sequence analysis with several comparative sequence analysis tools to significantly simplify and expedite its use. Currently, rVISTA accepts blastz alignments submitted at the rVISTA homepage or alignment and gene annotations automatically forwarded from the ECR Browser, zPicture and GALA database. [LOP$^+$02, LO04]. The sample data workflow can be seen in appendix B.

Chapter 4

Using the Grid for Large-Scale Analysis

The common feature of all above mentioned tools is the need to use a lot of computing power when the analysis is not basic. Thus running them on one CPU or perhaps a small cluster might still require too much time. One possible solution - described below - would be to use the Grid infrastructure, in particular BalticGrid infrastructure.

4.1 What is the Grid

The term “Grid” became popular after the publication of the book by Foster and Kesselman [FK98]. Due to this hype the word was (and still is) often used when speaking about distributed application. The general idea is however that the Grid is an infrastructure that provides access to computational and storage services (see appendix C for an overview of virtualization layers that lead to the Grid).

A more formal definition was given in Foster’s article ([Fos02]) according to which a Grid is a system where the following properties hold:

- computing resources are not administered centrally;
- open standards are used;

- non trivial quality of service is achieved.

Although in perfect world one would imagine to have one global grid, there are currently a lot of rivaling Grid projects producing a lot of middlewares. Perhaps most known are EGEE (with its lcg2/gLite middleware, [Pro06b]), Nordugrid ARC ([ea06b]), Condor ([TWML01]) and Unicore ([E^+02]).

BalticGrid (www.balticgrid.org) is a project intended to develop and integrate the research and education computing and communication infrastructure in the Baltic States into the emerging European Grid infrastructure. BalticGrid uses EGEE middleware, i.e. lcg2/gLite.

4.2 Using the Grid

Although the description below might be applied to most of the Grid systems, it is written with BalticGrid/EGEE in mind.

In order to get access to the Grid resources one must first acquire the certificate signed by some Certification Authority, for instance Estonian ID card comes with a certificate signed by Sertifitseerimiskeskus Ltd, company established by the 3 largest companies in Estonia.

Having acquired the certificate one would have to join a virtual organization (VO) and thus gain access to the resources that this VO is authorized to use.

After these initial steps one is ready to use the Grid resources. The usual use case is that the user describes the job in a script file and submits it to the resource broker and afterwards retrieves the results. However, the assumption that the required program is installed on all computing elements is too strong: it might work for GNU C compiler, but most certainly will fail for some analysis specific soft. Therefore one must think of a way to make sure that the needed software is available prior to running the job. Of course, copying the needed program to the target computing element or even compile it from sources as part of the job will work, but this is neither user-friendly, nor optimal in a sense of reliability and network usage.

4.3 Software Installation

Within BalticGrid infrastructure the software installation is handled by the users having the appropriate role (called *lcgadmin*). These users have the write rights on all the clusters within a EGEE/BalticGrid project (in case BalticGrid VO is authorized to use the computing element) in a special readable by all BalticGrid VO members area (*$VO_BALTICGRID_SW_DIR*).

The straight-forward approach for installing the software is therefore to submit a job to each computing element installing the software in the *$VO_BALTICGRID_SW_DIR* folder. After the successful installation and, perhaps, validation with a small job, one should publish a tag in the information system, giving information on the installed program (name, version, some other information) and making it possible to filter the computing resources according to the installed software.

This approach has been implemented and is available as appendix D (also accessible at `http://cerncms.hep.kbfi.ee/bg/auto_install.html`). The contents is the usage instruction, scripts for generation of program specific jobs and bulk submission, templates and samples.

Võrdleva genoomika kasutamine regulatoorsete järjestuste ennustamiseks

Bakalaureusetöö (4 ap)

Ilja Livenson

Resümee

Võrdlev genoomika on bioloogia valdkond, mis kasutab analüüsiks lähteandmetena erinevate organismide genoomsed andmeid. Üks võimalik kasutusala on regulatoorsete järjestuste ennustamine kasutades erinevate organismide DNA järjestusi ja fülogeneetilisi puid. Käesolevas töös antakse ülevaade fülogeneetilise jalajälje ja fülogeneetilise varjutamise meetoditest ja algoritmidest. Mõlemal juhul on tegemist fülogeneetilise võrdlusega, mis paljastab evolutsiooniliselt konserveerunud funktsionaalsed elemendid homoloogsetes geenides, erinevus seisneb fülogeneetelises kauguses võrdlevate organismide vahel.

Võrdleva genoomika puhul on standardseks meetodiks järjestuste joondamine. Antud töös kirjeldatakse järgmisi algoritme:

- BLAST - Smith-Waterman algoritmil baseeruv meetod.
- Mulan - DNA andmete joondamine threaded blockset aligner algoritmi abil. Töös on antud algoritmi kirjeldus.
- CompareProspector - Gibbs juhusliku valiku meetod.

- FootPrinter - järjestuste joondamine alamstringi säästlikuse probleemi lähenemisel. Töös on toodud selle probleemi formaalne kirjeldus ning lahenduse algoritm.
- Footer - meetod, mis kasutab kahte statistilist meetodit (asukoha konserveerumine ning vastavus PWM mudelile).
- rVista - erinevate programmide kompositsioon regulatoorsete järjestuste ennustamiseks.

Viimases osas on kirjeldatud viis, kuidas saab paigaldada ülalnimetatud programme ja meetodeid Gridil (hajutatud arvutusvõrgul) ja võimaldada suuremahulisi analüüse.

Abstract

Bachelor thesis (4 cp)

Ilja Livenson

Abstract

The thesis gives an overview of the comparative genomics - area of the biology dealing with the analysis and comparison of genomes from different species. Using this data one can enhance the algorithms for the discovery of the regulatory elements and thus to determine the function noncoding regions of the genome.

A number of databases and tools for comparative analysis are reviewed giving the reader general knowledge about data available for the analyses and the applied algorithms. The need of these tools for computational power is the motivation for the last part of the thesis - automated installing of scientific tools on the BalticGrid. The results of the latter are available on the accompanying compact disk.

Bibliography

[ABL$^+$02] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson. *Molecular Biology of the Cell*. 2002.

[AGM$^+$90] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol*, 215(3):403–410, 1990.

[B$^+$03] D. Boffelli et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, 2003.

[B$^+$04] D.A. Benson et al. Genbank. *Nucleic Acids Research*, 32(90001):23–26, 2004.

[BGvdB$^+$05] E. Berezikov, V. Guryev, J. van de Belt, E. Wienholds, RH Plasterk, and E. Cuppen. Phylogenetic shadowing and computational identification of human microrna genes. *Cell*, 120(1):21–24, 2005.

[BJVU98] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting Gene Regulatory Elements in Silico on a Genomic Scale. *Genome Research*, 8(11):1202, 1998.

[BKR$^+$04] M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F.A. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, et al. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research*, 14(4):708, 2004.

[Bla01] M. Blanchette. *Algorithms for phylogenetic footprinting.* ACM Press New York, NY, USA, 2001.

[BVUV97] A. Brazma, J. Vilo, E. Ukkonen, and K. Valtonen. Data mining for regulatory elements in yeast genome. *International Conference Intelligent Systems for Molecular Biology, 5th. Halkidiki, Greece, June*, pages 65–74, 1997.

[CLLL03] E.M. Conlon, X.S. Liu, J.D. Lieb, and J.S. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences*, 100(6):3339–3344, 2003.

[der] Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover.

[DLC05] Panayiotis V. Benos David L. Corcoran, Eleanor Feingold. Footer: a web tool for finding mammalian dna regulatory regions using phylogenetic footprinting. *Nucleic Acids Research*, 33:442–446, 2005.

[E^{+}02] D.W. Erwin et al. UNICOREŮa Grid computing environment. *Concurrency and Computation: Practice and Experience*, 14(13-15):1395–1410, 2002.

[ea06a] Barbara J. Culliton et al. Genetics and genomics timeline: 1957. `http://www.genomenewsnetwork.org/resources/timeline/1957_Crick.php`, 2006. [Online; accessed 21-May-2006].

[ea06b] M.Ellert et al. Advanced resource connector middleware for lightweight computational grids. *To be published in Future Generation Computing Systems*, 2006.

[EGS^{+}05] L. Elnitski, B. Giardine, P. Shah, Y. Zhang, C. Riemer, M. Weirauch, R. Burhans, W. Miller, and R.C. Hardison. Improvements to GALA and dbERGE II: databases featuring genomic sequence alignment, annotation and experimental results. *Nucleic Acids Research*, 33(Supplement 1):D466–D470, 2005.

[FK98] I. Foster and C. Kesselman. *The grid: blueprint for a new computing infrastructure*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1998.

[Fos02] I. Foster. What is the Grid? A Three Point Checklist. *Grid Today*, 1(6):22, 2002.

[GRH$^+$05] B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, et al. Galaxy: A platform for interactive large-scale genome analysis. *Genome Research*, 15(10):1451, 2005.

[Har03] Ross C. Hardison. Primer: Comparative genomics. *PLoS Biology*, 2(2):156–160, 2003.

[HLW03] A.S. Halees, D. Leyfer, and Z. Weng. PromoSer: a large-scale mammalian promoter and transcription start site identification service. *Nucleic Acids Research*, 31(13):3554–3559, 2003.

[LBL01] X. Liu, D.L. Brutlag, and J.S. Liu. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput*, 6:127–138, 2001.

[LH02] S.V. Levy and S.V. Hannenhalli. Identification of transcription factor binding sites in the human genome sequence. *Mammalian Genome*, 13(9):510–514, 2002.

[LLW$^+$04] Y. Liu, X.S. Liu, L. Wei, R.B. Altman, and S. Batzoglou. Eukaryotic Regulatory Element Conservation Analysis and Identification Using Comparative Genomics. *Genome Research*, 14(3):451, 2004.

[LNL95] J.S. Liu, A.F. Neuwald, and C.E. Lawrence. Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling

Strategies. *Journal of the American Statistical Association*, 90(432), 1995.

[LO04] G.G. Loots and I. Ovcharenko. rvista 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32:217–221, 2004.

[Lon04] F. Long. Genome-wide prediction and analysis of function-specific transcription factor binding sites. *In Silico Biology*, 4(4):395–410, 2004.

[LOP^{+}02] G.G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak, and E.M. Rubin. rvista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Research*, 12(5):832, 2002.

[LSM^{+}03] B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W.W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.

[MB02] Martin Tompa Mathieu Blanchette. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, pages 739–748, 2002.

[MB03] Martin Tompa Mathieu Blanchette. Genbank. *Nucleic Acids Research*, 31(1):23–27, 2003.

[MN95] J.F.R. Manen and A.R. Natali. Comparison of the evolution of ribulose-1, 5-biphosphate carboxylase (rbcl) and atpb-rbcl noncoding spacer sequences in a recent plant group, the tribe rubieae (rubiaceae). *Journal of Molecular Evolution*, 41(6):920–927, 1995.

[NP03] M.A. Nobrega and L.A. Pennacchio. Comparative genomic analysis as a tool for biological discovery. *The Journal of Physiology*, 554:31–39, 2003.

[OLG$^+$05] I. Ovcharenko, G.G. Loots, B.M. Giardine, M. Hou, J. Ma, R.C. Hardison, L. Stubbs, and W. Miller. Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. *Genome Research*, 15(1):184, 2005.

[PR01] Len A. Pennacchio and Edward M. Rubin. Genomic strategies to identify mammalian regulatory sequences. *Nature*, 2(2):100–109, February 2001.

[Pro06a] Human Genome Program. Functional and comparative genomics fact sheet. `http://www.ornl.gov/sci/techresources/Human_Genome/faq/compgen.shtml`, 2006. [Online; accessed 17-March-2006].

[Pro06b] LCG Project. Lhc computing grid project. `http://lcg.web.cern.ch/LCG/`, 2006. [Online; accessed 27-May-2006].

[SW81] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol*, 147(1):195–197, 1981.

[SYNS02] Y. Suzuki, R. Yamashita, K. Nakai, and S. Sugano. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Research*, 30(1):328–331, 2002.

[TKG$^+$88] DA Tagle, BF Koop, M. Goodman, JL Slightom, DL Hess, and RT Jones. Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol*, 203(2):439–55, 1988.

[TS06] Edgar Wingender Tilman Sauer, Ekaterina Shelest. Evaluating phylogenetic footprinting for humanŰrodent comparisons. *Bioinformatics*, 22(4):430–437, 2006.

[TWML01] Todd Tannenbaum, Derek Wright, Karen Miller, and Miron Livny. Condor – a distributed job scheduler. In Thomas Sterling, editor, *Beowulf Cluster Computing with Linux*. MIT Press, October 2001.

[Uss98] David Ussery. Gene expression and regulation. 215, 1998.

[vHACV98] J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol*, 281(5):827–42, 1998.

[Wik06a] Wikipedia. Dna. `http://en.wikipedia.org/wiki/DNA`, 2006. [Online; accessed 23-March-2006].

[Wik06b] Wikipedia. Rna. `http://en.wikipedia.org/wiki/RNA`, 2006. [Online; accessed 23-March-2006].

[Wik06c] Wikipedia. Sequence alignment. `http://en.wikipedia.org/wiki/Sequence_alignment`, 2006. [Online; accessed 27-May-2006].

[Wik06d] Wikipedia. Transcription (genetics). `http://en.wikipedia.org/wiki/Transcription_\%28genetics\%29`, 2006. [Online; accessed 23-March-2006].

[Wik06e] Wikipedia. Translation (genetics). `http://en.wikipedia.org/wiki/Translation_\%28genetics\%29`, 2006. [Online; accessed 25-March-2006].

[ZG03] Z. Zhang and M. Gerstein. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol*, 2(11), 2003.

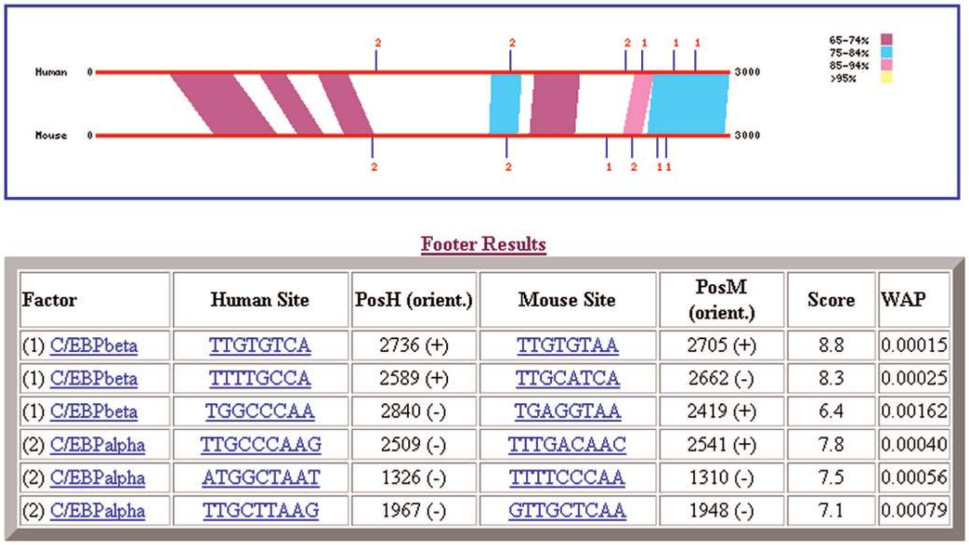
Glossary

***cis*-regulatory sequences**	Sequences co-located on the same chromosome of a homologous pair, 12
in vitro	In an artificial environment, 12
conserved sequence	An invariant sequence found in different DNA or RNA or protein sequences, 13
DNA	Deoxyribonucleic acid, 8
enhancer	short region of DNA that can be bound with proteins to enhance transcription levels of genes, 12
exon	Region of DNA within a gene that is not spliced out from the transcribed RNA, 16
Grid job	Instructions for running the user process, e.g. executable file, arguments, CPU requirements etc, 34
Hamming distance	Number of substitutions required to change one string into the other, 29
orthologous	Two genes are called orthologous if they have a common ancestor, 18

phylogenetic	Concerning the relationships of evolutionary development, 16
phylogenetic distances	number of evolutionary events (e.g. mutations) that occurred in the transition from one genome to the other, 16
promoter	DNA sequence that enables a gene to be transcribed, 12
PWM	matrix of score values that gives a weighted match to any given substring of fixed length, 28
PWM	position weight matrices, 30
regulatory elements	promoter, enhancer or other segment of DNA where regulatory proteins such as transcription factors bind preferentially, 11
RNA	Ribonucleic acid, 9
RNA polymerase	Enzyme responsible for transcribing DNA, 10
SNP	Single-nucleotide polymorphisms or one-letter variations in the DNA sequence, 24
TFBS	a region on a protein, DNA, or RNA to which specific transcription factors bind, 14
transcription factor	protein that binds DNA at a specific promoter or enhancer region or site, where it regulates transcription, 12
upstream	Region towards the 5' end of the DNA or RNA strand, 11
VO	Virtual Organization, 34

Appendices

Appendix A
Footer sample output



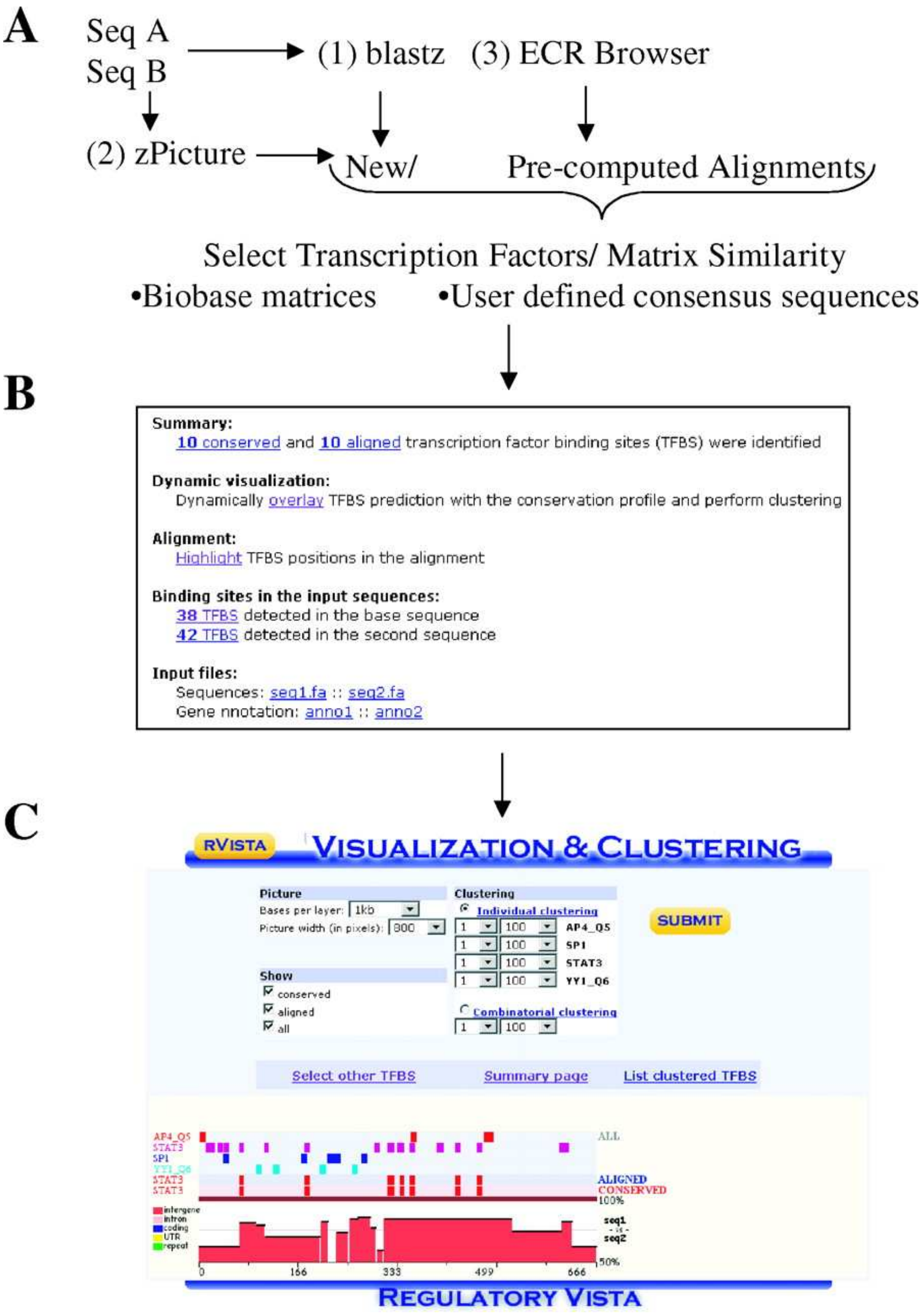
Footer Results

Factor	Human Site	PosH (orient.)	Mouse Site	PosM (orient.)	Score	WAP
(1) C/EBPbeta	TTGTGTCA	2736 (+)	TTGTGTAA	2705 (+)	8.8	0.00015
(1) C/EBPbeta	TTTTGCCA	2589 (+)	TTGCATCA	2662 (-)	8.3	0.00025
(1) C/EBPbeta	TGGCCCAA	2840 (-)	TGAGGTAA	2419 (+)	6.4	0.00162
(2) C/EBPalpha	TTGCCCAAG	2509 (-)	TTTGACAAC	2541 (+)	7.8	0.00040
(2) C/EBPalpha	ATGGCTAAT	1326 (-)	TTTTCCCAA	1310 (-)	7.5	0.00056
(2) C/EBPalpha	TTGCTTAAG	1967 (-)	GTTGCTCAA	1948 (-)	7.1	0.00079

Figure 4.1: Example of FOOTER output. The predicted sites are presented in table format and in the PNG formatted figure. The figure displays the alignment of the two promoter sequences, colour-coded by conservation percentage.

Appendix B
rVista 2.0 Analysis Data Flow

(A) The rVISTA tool can process blastz alignments submitted at the rVISTA homepage, or alignments automatically forwarded from the zPicture alignment program, the ECR Browser or the GALA database. (B) Users select the search criteria, and the results are returned in the same page as the downloadable static data files and dynamic links to visual analysis of TFBS distribution. (C) TFBSs for pre-selected TFs can be visualized above the conservation profile as tick marks, and the clustering module can detect user-specified groups of TFBSs.



Summary:
10 conserved and 10 aligned transcription factor binding sites (TFBS) were identified

Dynamic visualization:
Dynamically overlay TFBS prediction with the conservation profile and perform clustering

Alignment:
Highlight TFBS positions in the alignment

Binding sites in the input sequences:
38 TFBS detected in the base sequence
42 TFBS detected in the second sequence

Input files:
Sequences: seq1.fa :: seq2.fa
Gene nnotation: anno1 :: anno2

Picture
Bases per layer: 1kb
Picture width (in pixels): 800

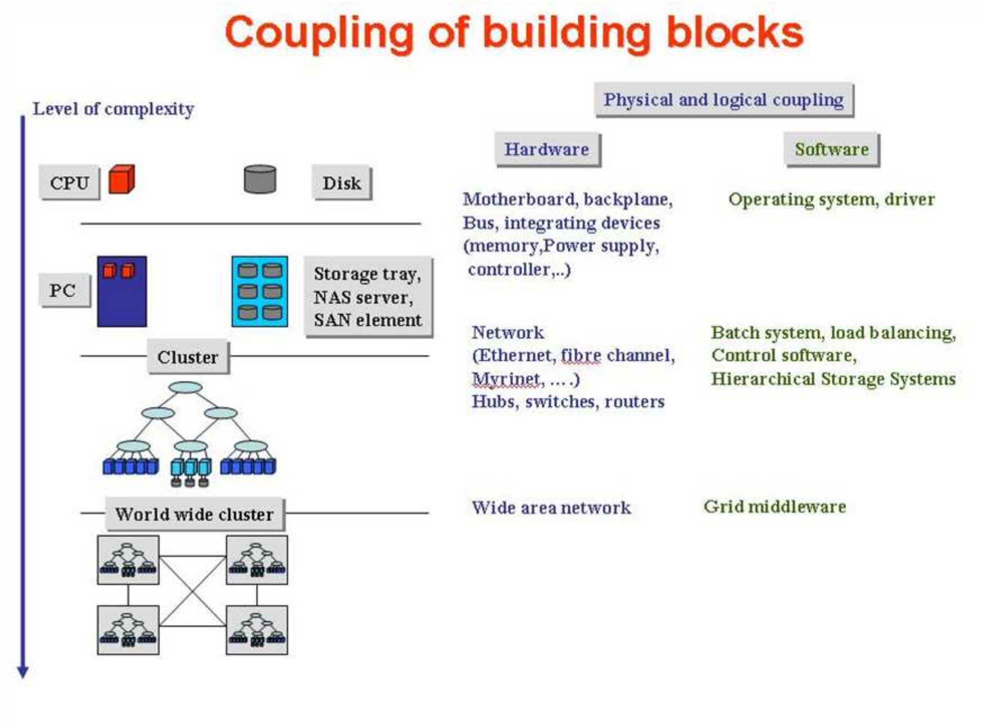
Show
☑ conserved
☑ aligned
☑ all

Clustering
● Individual clustering
1 100 AP4_Q5
1 100 SP1
1 100 STAT3
1 100 YY1_Q6

○ Combinatorial clustering
1 100

Figure 4.2: rVista 2.0 Analysis Data Flow

Appendix C
From Single CPU to Grid



Coupling of building blocks

Level of complexity

Physical and logical coupling

Hardware

Software

CPU

Disk

Motherboard, backplane, Bus, integrating devices (memory,Power supply, controller,..)

Operating system, driver

PC

Storage tray, NAS server, SAN element

Cluster

Network (Ethernet, fibre channel, Myrinet, ...) Hubs, switches, routers

Batch system, load balancing, Control software, Hierarchical Storage Systems

World wide cluster

Wide area network

Grid middleware

Figure 4.3: Grid is a next step in the virtualization of computer resources [Pro06b]

Appendix D
Scripts and templates for automated software installation

Please, see the accompanying compact disk.