Georg-August-Universität Göttingen
Zentrum für Informatik

ISSN 1612-6793
Nummer ZFI-BM-2005-27

Bachelorarbeit

im Studiengang „Angewandte Informatik“

Improving Gene Prediction in Human Using Alignments with Mouse Genome Sequences

Ana Tzvetkova

am Institut für Mikrobiologie und Genetik
der Georg-August-Universität Göttingen,
Abteilung für Bioinformatik

Bachelor- und Masterarbeiten
des Zentrums für Informatik
an der Georg-August-Universität Göttingen

26. September 2005

Georg-August-Universität Göttingen
Zentrum für Informatik

Lotzestraße 16-18
37083 Göttingen
Germany

Tel.	+49 (5 51) 39-1 44 14
Fax	+49 (5 51) 39-1 44 15
Email	office@informatik.uni-goettingen.de
WWW	www.informatik.uni-goettingen.de

Ich erkläre hiermit, daß ich die vorliegende Arbeit selbständig verfaßt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Göttingen, den 26. September 2005

Bachelor Thesis

Improving Gene Prediction in Human Using Alignments with Mouse Genome Sequences

Ana Tzvetkova

26. September 2005

Supervised by Prof. Dr. Burkhard Morgenstern and Dr. Mario Stanke
Institute of Microbiology and Genetics,
Department of Bioinformatics,
Georg-August-Universität Göttingen

Table of Contents

1. Abstract

The recognition of the genes along the human genome is an important step in its annotation. The complete euchromatic sequence of the human and the mouse enable the usage of cross-species comparison techniques in the identification of protein-coding genes in the human genome. Here we present an approach that uses comparative genome sequence analysis to generate extrinsic information as an input of the gene-prediction program AUGUSTUS. This extrinsic information was derived from alignments between human and mouse syntenic sequences. The alignments were made with the alignment program DIALIGN. DIALIGN generates alignments as a chain of gap-free local alignments. We process these local alignments and retrieve evidence about coding sequences from them. This evidence is given as additional input to AUGUSTUS in order to improve the prediction accuracy. The reliability of the extrinsic information generated in this way was automatically evaluated on a training set of human sequence regions and considered in the gene prediction. The above combined approach was tested and compared to AUGUSTUS *ab initio* and to five comparative-genomics-based programs. The tests and the comparisons were made on the annotation of selected human sequence regions which were chosen by the **Enc**yclopedia **o**f **D**NA **E**lements (ENCODE) project initiated by American National Human Genome Research Institute. Our comparative approach improves the accuracy of the gene finding program AUGUSTUS both regarding the specificity (with 1.4 to 6.1% depending on the feature level) and the sensitivity (1.5% to 10.3%). Moreover, it showed the highest sensitivity on base level (88.14%) among the compared programs. Generally, some other dual- or multiple-genome gene finding programs to which our approach was compared show at some aspects better results(performance), but at some other aspects it shows better performance even than already established programs. The simplicity of our comparative approach gives much room for further improvements.

2. Introduction

The genetic information in the living organisms is encoded by a double-stranded polymer molecule called deoxyribose nucleic acid (DNA). DNA is build of monomers called nucleotides. Nucleotides consist of sugar - phosphate residues, that are building the backbone of the polymer molecule and purine or pyrimidine residues (called bases), that are responsible for the coding of the genetic information. There are only four different bases: adenine (A), cytosine (C), guanine (G) and thymine (T) in the DNA molecules of all cellular organisms. The various combinations of them in the distinct DNA sequences, however, are enough to encode the whole abundance of existing organisms. Each nucleotide of the one DNA-strand is chemically bound with a complementary nucleotide on the other DNA-strand. A is the complement of T and reversely. C and G are complementary in the same way. Thus, the nucleotide sequence on the one strand determines the nucleotide sequence on the other strand. Both strands have a chemically determined direction and are respectively called *forward* or *plus strand* and *reverse* or *minus strand*.

The cellular organisms on Earth are classified in two major groups: *eukaryotes* and *prokaryotes* that significantly differ in their cell structure and genome organization. The group of *prokaryotes* includes *euobacteria* and *archea*. They are mono-cellular organisms that do not posses a separate, membrane-bounded cell nucleus or other membrane-defined organelles. The coding regions in their genomes are with extreme high density, i.e. they are very compact located in the genome. The group of *eukaryotes* includes the animal (inclusive human), plant and fungi kingdoms. Their genomic DNA is localized in a double-membrane-bounded cell nucleus. Their genomes consist of more than one DNA molecules called chromosomes. The coding regions in eukaryotic genomes are with low density but their organization is rather complex compared to those of *prokaryotes*.

The proteins are the functional and structural elements building the organisms. They consist of amino acids. The segments of DNA coding for ribonucleic acids (RNA) or proteins are called genes. The genes in *eukaryotes* are constituted of exons and introns. The introns are located between the exons. The exons are parts of the gene that are encoding the so called mature messenger RNA (mRNA) sequence. A protein is synthesized from a DNA sequence region containing gene in the following way. In a process called transcription the DNA region is copied in a so called pre-mature mRNA. The locations of the exon-intron junction are called splice sides. They are part of the introns and are two of a type: acceptor splice site

(ASS) (at the beginning of an intron) and donor splice site (DSS) (at the end of an intron). The introns are cut off in a process known as splicing and in this way is the mature mRNA synthetized. Not the whole mRNA sequence is coding for protein. Two parts called 5'- and 3'-untranslated regions (5'- respectively 3'-UTR) are flanking the coding part. Although not coding for proteins they are part of the genes' exons. The coding part of the mRNA is sequentially translated to amino acid sequence in a process called translation. Each three consecutive nucleotides form a so called codon. The codons are translated according to the genetic code. In the genetic code, three bases are required to specify one amino acid. An mRNA can be read off the DNA in three different reading frames, depending on the starting base. For a DNA region representing an exon reading frame '0' means that its first base corresponds to the first base of a codon. '1' indicates that there is one extra base, i.e. that the second base of the region corresponds to the first base of a codon, and '2' means that the third base of the region is the first base of a codon (Fig. 1). The translation starts with the start codon ATG, goes on in the same frame and stops directly after the first appearance of one of three stop codons: TAA, TAG, TGA in the used frame.

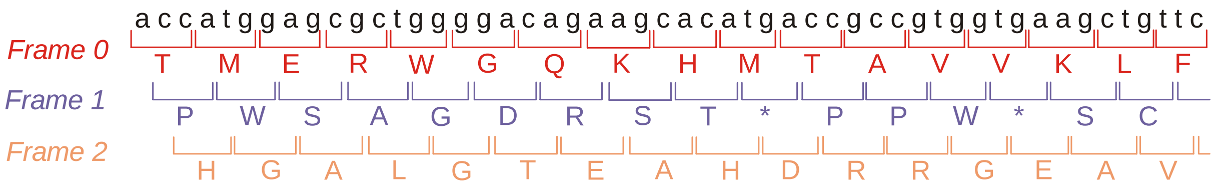

a c c a t g g a g c g c t g g g g a c a g a a g c a c a t g a c c g c c g t g g t g a a g c t g t t c
Frame 0
T M E R W G Q K H M T A V V K L F
Frame 1
P W S A G D R S T * P P W * S C
Frame 2
H G A L G T E A H D R R G E A V

Fig. 1 The tree reading frames for an example DNA sequence. All three possible reading frames (numbered from 0 to 2) are shown. The DNA sequence is shown in small and the possible amino acid sequences in capital letters according to the IUPAC-approved abbreviations. Stop codons are marked with asterisks.

There are various ways of splicing out the introns in a pre-mature mRNA (e.g. some exons or part of them could be also removed by the splicing), implying that several different proteins can be made from the same gene. The different variants of exon-intron composition of one gene are called *transcripts* of this gene.

By July 2005 the genomes of 36 eukaryotes have been sequenced (see http://www.genomesonline.org/). 494 additional eukaryotic genomes are currently sequenced. The human and mouse genomes are also among the completely sequenced genomes. The human genome has ~2.85 billion bases separated in 23 DNA molecules called chromosomes. Thereof only 55 Mega bases (~1.9 %) [International Human Genome Sequencing Consortium 2004] are used to code for proteins. The latest studies estimate the number of all protein coding genes in the human genome at 20000-25000 [International Human Genome Sequencing Consortium 2004]. Now the goal is to identify all protein coding gene structures and to annotate them completely using the available genomic DNA sequence data. A special challenge is the identification of coding sequence regions by eukaryotes, because of the above mentioned more complex structure of the eukaryotic genes and their low density in the genomes. Bioinformatics is an integral part of such a process and its automation. There are two main approaches for finding genes in genomic sequences: intrinsic (also called *ab initio*) and extrinsic methods. The intrinsic methods use stochastic models of biological signals like splice sites, composition and length of exons and introns, codon frequencies, etc. The parameters of these models are trained with known genes from the same or closely related species. Once trained, the programs based on this method need solely the DNA sequence under study to make their predictions. Examples for such programs are GENSCAN [Burge and Karlin 1997], GENEID [Parra, Blanco et al. 2000], GENIE [Kulp, Haussler et al. 1996], HMMGene [Kulp, Haussler et al. 1996]. The extrinsic method tools use the so called extrinsic information which lies outside the DNA sequence. These methods can be divided in two further groups: (i) homology-based gene finders using similarities between the considered species and already known proteins, cDNAs or ESTs such as GENOMESCAN [Yeh, Lim et al. 2001], GENEWISE2 [Birney, Clamp et al. 2004]; (ii) comparative-genomics-based gene finders using similarities between the considered species and evolutionary close species, such as TWINSCAN [Flicek, Keibler et al. 2003], N-SCAN [Gross and Brent 2005], SGP-2 [Parra, Agarwal et al. 2003], SLAM [Alexandersson, Cawley et al. 2003].

Actually, the probabilistic models used in the intrinsic approaches underlie also the above mentioned comparative-genomics-based programs that are using additional evidence to predict genes in genomic DNA. This enhancement of the *ab initio* methods into the comparative approaches aims to improve the gene finding accuracy. The motivation is not only to make the gene finding more reliable but to provide a method identifying genes with low expression rate or novel ones and thus not supported with ESTs or entries in protein databases.

AUGUSTUS is a gene prediction program for eukaryotes implemented by Mario Stanke [Stanke and Waack 2003]. It is based on a stochastic model which uses the genomic DNA sequence of the species, but in addition can consider extrinsic information. The extrinsic information is additional information about potential genes that is obtained outside of the genomic sequence itself. Such extrinsic information could come from cross-species comparison. When genomic regions of the compared species have varying conservation levels this could be regarded as evidence that the regions of high similarity correspond to functional elements of the sequence and therefore are more likely to be coding. We call an individual piece of extrinsic information a *hint* [Stanke 2003].

DIALIGN is an alignment program implemented by Burkhard Morgenstern [Morgenstern 1999]. It constructs the alignments as a chain of gap-free local pairwise alignments called fragments.

There are many research groups developing tools that identify computationally the homologous regions between human and mouse genomic sequences. Our idea was to get these homologous regions and to align them then with DIALIGN. Following we wanted to process the resulting DIALIGN-fragments so that to use them as hints in AUGUSTUS. There are diverse types of extrinsic information (e.g. exon, exonpart) that could be considered in AUGUSTUS [Stanke 2003]. We formulate the extrinsic information obtained through the human-mouse inter-genomic alignments as hints of the type 'exonpart', as we consider a hint we generated as evidence of a part of a coding region.

This thesis is organized as follows:

In chapter 3 the alignment program DIALIGN is presented.

In chapter 4 the gene prediction program AUGUSTUS is presented.

In chapter 5 is shown which regions from human and mouse genomic sequence do we use to make the alignments with DIALIGN. It is also described how we process these alignments to create the hints for AUGUSTUS.

In chapter 6 we describe the annotation of the human genome that we use to make the tests of our combined approach.

In chapter 7 are shown the results of these tests and of the comparisons of our combined approach to other gene prediction programs based on cross-species sequence alignments. It is also discussed which improvements of our approach we could make to boost its performance.

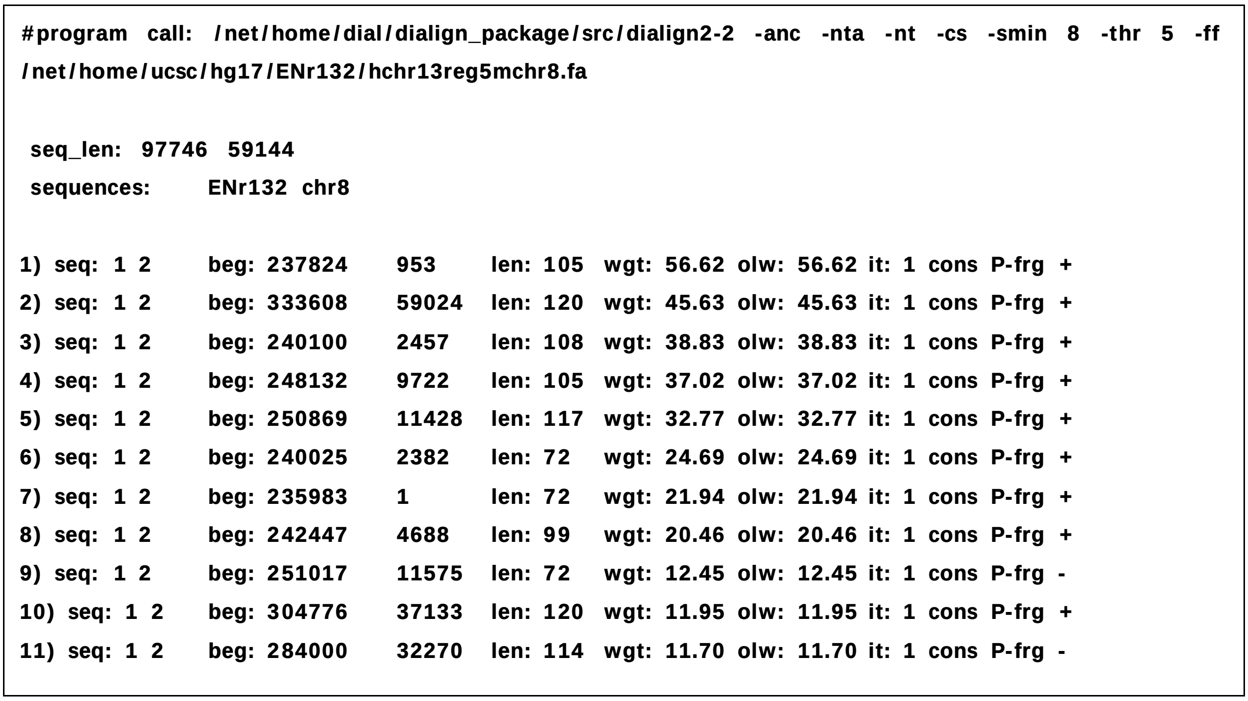
3. DIALIGN

As already mentioned DIALIGN [Morgenstern 1999] is a program for pairwise and multiple alignments of DNA or protein sequences. It constructs the alignments as a collection of so called fragments. The fragments are gap-free local pairwise alignments. To every fragment *f* is assigned a weight score *w*(*f*) reflecting the degree of similarity between two aligned sequence segments. The program selects then a consistent chain of fragments with maximal total weight. For pairwise alignment, it searches a chain of fragments $f_1 \ll f_2 \ll \ldots \ll f_k$ such that the sum $\Sigma_i w(f_i)$ is maximal, where $f_i \ll f_j$ means that, in both sequences, the end positions of f_i are strictly smaller than the respective start positions of f_j [Morgenstern 2000]. DIALIGN offers two possibilities to measure the similarities between two DNA-sequence segments. At the nucleotide level, the segments are compared nucleotide-by-nucleotide and the number of matching nucleotide pairs is considered. At the peptide level, the DNA segments are first translated and then the resulting peptide segments are compared using the BLOSUM 62 substitution matrix [Henikoff and Henikoff 1992]. This option is called the "translation" option. The program calculates the probability of fragments of the same length and at least the same sum of matches or accordingly BLOSUM scores to occur by chance in random sequences of the same length as the input sequences. There is a possibility to calculate the scores of fragments by comparing their segments on both levels. By that mixed alignment the peptide-level similarity is calculated for the plus strand and for the reverse complement. The score of a particular fragment is then the maximum of the three similarity values [Morgenstern et al. 2002].

Our goal is to identify protein-coding sequence regions. For this reason we took the translation option where the DNA-sequence is first translated according to the genetic code and then the segments are compared at the peptide level. DIALIGN searches by default fragments on the plus strand, but there is the so-called Crick-strand option. When this option is chosen the reverse strand will be also considered. Since we want to find the protein-coding genes on both strands we used this option. Another option of DIALIGN determinates a fragment-score threshold, thus generating only fragments with weight score exceeding this threshold value. Using this option could cause generation of fragments other than those generated without it and not simply elimination of those fragments with fewer score.

Anchor points are used by some alignment programs to reduce their search space and running time. For DIALIGN that means that fragments are searched by comparing sequence segments from the first sequence that lie left/right from the anchor point only and only with sequence segments from the second sequence that lie likewise left/right from it. The program CHAOS [Brudno and Morgenstern 2002] searches regions of high sequence similarity between the input sequences. DIALIGN can use regions identified by CHOAS as *anchor points*. The *anchor points* created by CHAOS are speeding up the DIALIGN alignment procedure by one to two orders of magnitude without affecting the quality of the output alignments.

If option –ff is used DIALIGN creates for every aligned DNA-sequence pair not only the alignment of the two sequences but also a so-called fragment file (Fig. 2). For pairwise alignment the fragment file contents the whole information about every fragment contained in the optimal alignment.



```
#program call: /net/home/dial/dialign_package/src/dialign2-2 -anc -nta -nt -cs -smin 8 -thr 5 -ff
/net/home/ucsc/hg17/ENr132/hchr13reg5mchr8.fa


 seq_len:  97746  59144
 sequences:     ENr132  chr8


1) seq: 1 2     beg: 237824    953     len: 105  wgt: 56.62 olw: 56.62 it: 1 cons P-frg +
2) seq: 1 2     beg: 333608    59024   len: 120  wgt: 45.63 olw: 45.63 it: 1 cons P-frg +
3) seq: 1 2     beg: 240100    2457    len: 108  wgt: 38.83 olw: 38.83 it: 1 cons P-frg +
4) seq: 1 2     beg: 248132    9722    len: 105  wgt: 37.02 olw: 37.02 it: 1 cons P-frg +
5) seq: 1 2     beg: 250869    11428   len: 117  wgt: 32.77 olw: 32.77 it: 1 cons P-frg +
6) seq: 1 2     beg: 240025    2382    len: 72   wgt: 24.69 olw: 24.69 it: 1 cons P-frg +
7) seq: 1 2     beg: 235983    1       len: 72   wgt: 21.94 olw: 21.94 it: 1 cons P-frg +
8) seq: 1 2     beg: 242447    4688    len: 99   wgt: 20.46 olw: 20.46 it: 1 cons P-frg +
9) seq: 1 2     beg: 251017    11575   len: 72   wgt: 12.45 olw: 12.45 it: 1 cons P-frg -
10) seq: 1 2    beg: 304776    37133   len: 120  wgt: 11.95 olw: 11.95 it: 1 cons P-frg +
11) seq: 1 2    beg: 284000    32270   len: 114  wgt: 11.70 olw: 11.70 it: 1 cons P-frg -
```

Fig. 2 An example of fragment file generated by DIALIGN. The first line shows the program call: which options are used and which is the file in a FASTA format with the input sequences. The next two lines are showing the length of both aligned sequences and their names. Every numbered line gives information about a particular fragment: begin in both sequences; length and weight score; fragments' type and strand.

4. AUGUSTUS

AUGUSTUS is a program for finding genes coding for proteins in eukaryotes. It finds particularly the coding part of the genes. The program is based on a generalized Hidden Markov Model (GHMM) where especially the intron lengths are more precisely modeled [Stanke and Waack 2003].

In the probabilistic model of AUGUSTUS extrinsic information can be also incorporated. The following types of extrinsic information can be considered by AUGUSTUS:

- start – presumable translation start site of a gene (a nucleotide triple where the translation of the coding sequence starts)
- stop – presumable translation termination site of a gene (a nucleotide triple after that the translation of the coding sequence stops)
- ASS – presumable acceptor splice site of a gene
- DSS – presumable donor splice site of a gene
- exonpart – a segment of the sequence presumably coding: part of an exon. The actual exon may properly contain this segment or may be equal to the segment.
- exon – a complete presumable coding exon.

We call each individual piece of extrinsic information a hint. The hints that we are using in this thesis are of type 'exonpart'.

Bonuses and maluses for each gene structure are introduced in AUGUSTUS to support the search of an optimal gene structure [Stanke 2003]. To every gene structure is assigned a value giving the probability of that structure to get sampled. The value of a gene structure increases if it gets a bonus. For every hint that is compatible with a gene structure, this gene structure gets a bonus. However, it does not automatically mean that only structures compatible with hints are chosen. A gene structure that is not supported with hints could be preferred rather then one that is, only because the first one has much higher value. The value of a gene structure decreases if it gets malus. A gene structure gets malus for every predicted coding base that does not lie in a hint's interval.

The extent of the bonus that a gene structure gets may depend on the score of its hints (if they have any). When this possibility is chosen the hints are classified according to their score values in several classes. To each class is assigned a scale coefficient by which the bonus is multiplied. If the score of a hint is high the bonus for the gene structure will be also high.

Thus the gene prediction is influenced. Since the hints that we generate have scores we use this possibility to improve the performance of our combined approach.

Both DIALIGN and AUGUSTUS program are available at: http://gobics.de/department/

5. Generation of Hints for AUGUSTUS

To generate the hints for AUGUSTUS we use cross-species sequence alignments. Comparative genomics is based on the phylogenetic footprinting principal which states that coding sequence regions are usually more conserved between the species than the non-coding ones. The mouse genome is considered as one of the genomes that are appropriate for comparing to the human genome. One reason for this is that the mouse is an established species suitable for labor experiments. Another reason is that the mouse is evolutionary close enough to human. That facilitates the alignment of sequences between both species which are orthologous (Two sequences are called orthologous if they evolved from a common ancestor.). On the other hand the mouse is sufficiently evolutionary distant from human. Thus the regions which remained conserved between the two species are often regions that are functionally important. As the Mouse Genome Sequencing Consortium [Waterston, Lindblad-Toh et al. 2002] reports: over 90% of the mouse and human genomes can be partitioned into corresponding regions of synteny (conserved gene order), reflecting segments in which the gene order in the most recent common ancestor has been conserved in both species. Also, at the nucleotide level, approximately 40% of the human genome can be aligned to the mouse genome representing thus most of the ortologous sequences that remain in both lineages from the common ancestor.

As a first step, we need all orthologous genomic regions between human and mouse to analyze them in detail using DIALIGN.

5.1. Intergenomic alignment maps

It is quite challenging to find all the orthologous regions between two species especially if they both have large genomes and are of high complexity. There are various research centers that construct maps of alignments between human genomic sequence and mouse or other eukaryotic genomes. Following are some browsers through which such maps are accessible:

NCBI (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606),

Ensembl (http://www.ensembl.org/),

UCSC (http://genome.ucsc.edu/),

Softberry (http://sun1.softberry.com/berry.phtml?topic=index&group=synteny), etc.

5.1.1. Ensembl

Ensembl is a joint project between EMBL-EBI (European Bioinformatics Institute) and the Sanger Institute. The main aim of the Ensembl project is to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Through the Ensembl Genome Browser one can get maps of similarities between these genomes (Fig. 3). The problem by these maps is that they are too general (some regions of one of the genomes are 30 MBase long and more) and it is quite complicated to get the files with the data over the aligned sequence segments for the whole genome.

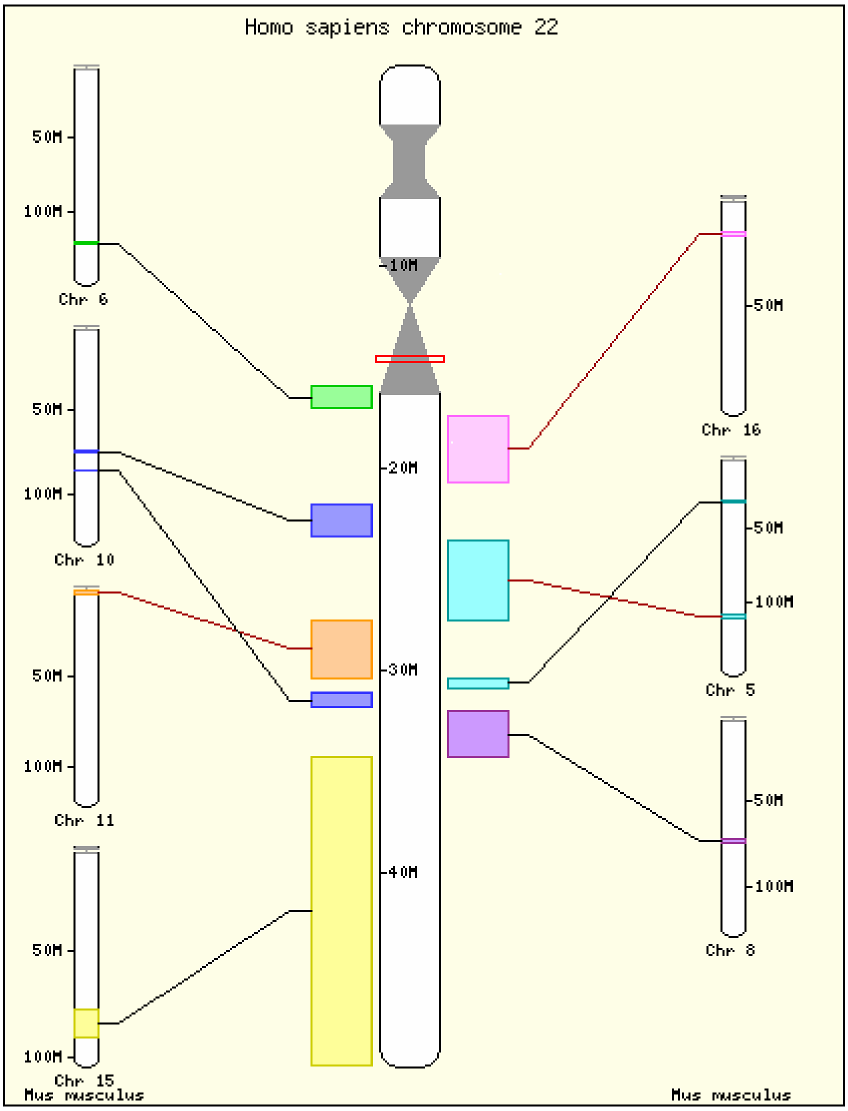


Fig. 3. Ensembl's syntenyview of human chromosome 22 vs. mouse genome. Regions of the chromosome that are aligned with regions from the same chromosome in mouse (Mus musculus) have the same colour.

5.1.2. Source of used data

5.1.2.1. "Genome Bioinformatics" site of UCSC (University of California at Santa Cruz)

The "Genome Bioinformatics" site of UCSC (University of California at Santa Cruz) contains the reference sequence and working draft assemblies for a large collection of genomes including the human and murine genomes. It also provides many tools to explore these sequences. Sequence and annotation data of these species is available for downloading. Varieties of pairwise alignments are also available for the diverse human and mouse assemblies.

We used the May 2004 human assembly (also known as build35 or hg17) vs. the May 2004 mouse assembly (also known as build33 or mm5). We downloaded the data from the axtTight directory (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsMm5/) which contains a highly conserved subset of the best alignments for any part of the human genome. The alignments are produced by the enhanced BLASTZ alignment program [Schwartz, Kent et al. 2003], appropriate for aligning whole-genomic DNA sequences. BLASTZ is available at http://www.bx.psu.edu/miller_lab/.

For each human chromosome in the axtTight directory there is a file in "axt" format with the corresponding alignments to the mouse genome. An "axt" file has the following structure:

```
0 chr19 3001012 3001075 chr11 70568380 70568443 - 3500
TCAGCTCATAAATCACCTCCTGCCACAAGCCTGGCCTGGTCCCAGGAGAGTGTCCAGGCTCAGA
TCTGTTCATAAACCACCTGCCATGACAAGCCTGGCCTGTTCCCAAGACAATGTCCAGGCTCAGA

1 chr19 3008279 3008357 chr11 70573976 70574054 - 3900
CACAATCTTCACATTGAGATCCTGAGTTGCTGATCAGAATGGAAGGCTGAGCTAAGATGAGCGACGAGGCAATGTCACA
CACAGTCTTCACATTGAGGTACCAAGTTGTGGATCAGAATGGAAAGCTAGGCTATGATGAGGGACAGTGCGCTGTCACA

2 chr19 3008482 3008533 chrX 7236511 7236590 + 5300
CACATCTGGAGCACAGATGGCCCTCTCAAGGTAATTTATTGTATGCATTGACTGTTTACCAAACAAATGTCTTACTATGT
CACATCTGGAGCACATATGGCCTTCTCAAGGTGATTTATTTTATGCATTTACTGTTTACTGAATAACTGTCTGACTGTGT
```

Each alignment is with an individual block with 3 lines represented. The first line contains chromosome, start and end position in both primary and aligning organism, strand of the aligning organism and BLASTZ score of the alignment. The blocks are by blank lines separated. The alignments are ordered by position on the plus DNA strand of the respective human chromosome. For more detailed description of the "axt" format see http://genome.ucsc.edu/goldenPath/help/axt.html

5.1.2.2. Postprocessing of the human-mouse sequence alignments

For each human chromosome we use the alignments to the mouse assembly to generate a file with assumed orthologous regions in both sequences. To create this file, we join the alignments from the corresponding "axt" file in order to obtain larger sequence segments that will be later aligned with DIALIGN. We call the union of the "axt"-alignments a *large alignment* and that are the human and mouse sequence segments beginning at the corresponding start positions of the first and ending at the end positions of the last involved "axt"-alignment. There are several conditions for the alignments to be included in a *large alignment*.

- The alignments must follow each other directly in the "axt" file. As the exons of a gene are consecutively located in a genomic sequence, we use the fact that in the "axt" file the conserved regions between the two species are ordered by position on the plus DNA strand of the respective human chromosome.
- The involved mouse chromosome and mouse strand must be identical (Fig. 4a). This condition is necessary as we search orthologous regions between human and mouse and a gene lies only in one chromosome and only in one DNA strand.
- The physical distance between the corresponding human chromosome segments <u>and</u> between the corresponding mouse chromosome segments must be less then a certain threshold value T1 (Fig. 4b). The gaps between the aligned segments could represent introns or intergenic regions. Our idea was to assign T1 a value representing maximal intron length such that the most of introns remain under this threshold respective their lengths [Sakharkar, Chow et al. 2004]. Thus we try to keep the exons belonging to same gene in one *large alignment*.
- The entire length of the larger human region <u>and</u> the corresponding larger mouse region must not exceed a certain threshold value T2 (Fig. 4c). If the joining of a new alignment cause such an exceedence it would not be joined, but will initiate a new *large alignment*. Thus a *large alignment* is separated in more, but smaller alignments. The running time of DIALIGN increases significantly if very large sequences are aligned with it. Reducing the length of the DNA segments from both species that have to be aligned with DIALIGN decreases radically the running time and still the all primary "small" alignments are processed.

A *large alignment* is included in the file with assumed orthologous regions if its entire length exceeds a threshold value (T3). We introduce this threshold value considering the minimal length of the human exons [Sakharkar, Chow et al. 2004].

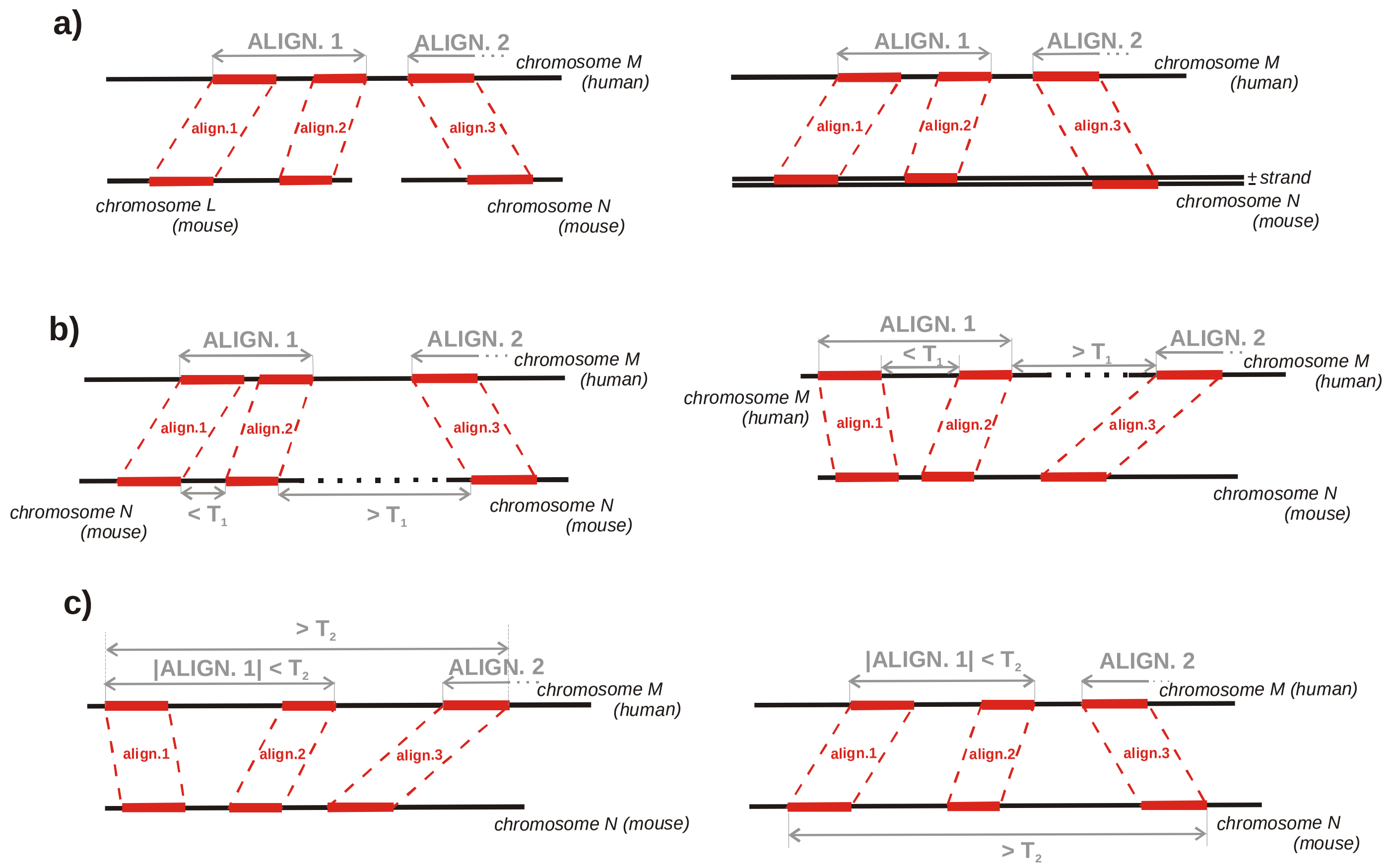


Fig. 4: Processing the UCSC alignments between human and mouse genomes. 'align.X' and 'ALIGN.X' denote the UCSC alignments and the resulting *large alignments* respectively. To create a *large alignment* align.1-align.3 are required to appear in the same order in the "axt" file. In all cases alignment align.3 will not be joined to the *large alignment* ALIGN.1 but initiates a new large alignment ALIGN.2. **a)** On the left, align.3 lies on a different mouse chromosome than align.1 and align.2; **a)** On the right, align.3 lies on the same mouse chromosome, but on a different DNA strand than align.1 and align.2; **b)** the distance between align.3 and align.2 in human (right) or in mouse (left) is bigger then threshold value T1; **c)** the length by human (left) or mouse (right) of the ALIGN.1 will exceed the threshold value T2 if align.3 joint.

The file with orthologous regions is then used to obtain the FASTA file with the DNA sequence segments of the two species that we want to align with DIALIGN (Fig. 5). We first write the DNA sequence region of human and then the DNA sequence region of mouse which is presumably orthologous with the considered human region. For each region in human which is presented in the file with orthologous regions is created a corresponding FASTA file. The start and the end positions are included in the region. If the strand of mouse is 'minus' we take the reverse complement of the plus-strand region with the corresponding coordinates.

```
Homo_sapiens chr20: 3210279 3223291   Mus_musculus chr2:  130479804 130488508 +
Homo_sapiens chr20: 3225396 3288253   Mus_musculus chr2:  130489310 130555984 +
Homo_sapiens chr20: 3289146 3290992   Mus_musculus chr3:  99534153  99535988  -
Homo_sapiens chr20: 3302682 3366078   Mus_musculus chr2:  130559832 130620079 +
Homo_sapiens chr20: 3393153 3394405   Mus_musculus chrX:  34380657  34382675  -
Homo_sapiens chr20: 3399719 3405135   Mus_musculus chr2:  130655548 130661092 +
Homo_sapiens chr20: 3417950 3419668   Mus_musculus chr4:  31038592  31056150  -
Homo_sapiens chr20: 3463880 3563041   Mus_musculus chr2:  130682389 130767719 +
Homo_sapiens chr20: 3567456 3667369   Mus_musculus chr2:  130769945 130883150 +
Homo_sapiens chr20: 3668101 3752949   Mus_musculus chr2:  130883582 130961901 +
Homo_sapiens chr20: 3783249 3876962   Mus_musculus chr2:  130987794 131057576 +
Homo_sapiens chr20: 3892282 3951227   Mus_musculus chr2:  131061956 131113683 +
Homo_sapiens chr20: 3952369 3953260   Mus_musculus chr4:  123874406 123875340 +
Homo_sapiens chr20: 3970812 3975320   Mus_musculus chr2:  131135710 131142144 +
```

Fig. 5 An example of a file with assumed orthologous regions in two sequences. Each line contains the species names, the chromosomes, begin and end position of the sequences' segments that will be aligned with DIALIGN. The coordinate of the first species are always concerning the plus DNA strand. The DNA strand of the sequence segment of the second species is given then at the end of the line.

5.2. DIALIGN-fragments processing

For training we took thirteen annotated regions of the human genomic sequence (see chapter 6.3.) and their homologous regions in the mouse genome. The DNA sequences we used were all masked where repetitive elements or low complexity DNA sequences identified with the RepeatMasker program (http://www.repeatmasker.org/). We then created the corresponding FASTA files and gave them as an input to DIALIGN. Then we ran DIALIGN on these regions. With the resulting fragment files we made PostScript graphics to check whether there is some correlation between the fragments and the exons as expected. To create the graphics we used the program gff2ps [Abril and Guigo 2000]. The program and its manual are available at: http://www1.imim.es/software/gfftools/GFF2PS.html

On the pictures some regularities are noticeable. The fragments that match a particular exon lie often relative close to each other (Fig. 6a). Consequently we decided to merge the fragments with a distance smaller than a certain threshold value in a larger fragment. Our idea was to use the DIALIGN-fragments to retrieve evidence for coding sequence segments as an input to AUGUSUTS. The hints we wanted to create are of type exonpart. It occurs frequently that a fragment matching exon is longer than this exon at one of the exon's ends (Fig. 6b). One reason can be that the splice sites which are part of the introns are highly conserved and thus included in the fragments. For that reason we shortened the fragments at both ends. Another observation brought us to a next idea. The fragments with a lower weight score often lie outside the protein-coding regions (Fig. 6c). Therefore, we left out all fragments with score under a certain threshold value. The motivation is that the functional elements that are non-coding (e.g. regulatory elements) are less conserved than the exons, but still are more conserved than the non-functional elements. That implies that fragments are also found where non-coding elements are located though these fragments generally have lower scores. We eliminate low-scoring fragments because we are interested only in the coding elements.

Since with the protein option of DIALIGN the sequences are first translated and then aligned, the length of the resulting fragments is divisible by three. Consequently we firstly consider that the reading frame of the human sequence segment participating in a fragment is '0' (see Chapter 2). Secondly fragments are only then merged if the distance between them is smaller then the distance-threshold value and if this distance is divisible by three, i.e. the merged fragments have the same reading frame. Thirdly the fragments are cut at the both ends in such a way that their resulting length is also divisible by three.

We discussed also which function we should use to calculate the score of a large fragment when the fragments are merged. We tested the sum of the fragment scores and their maximum. The choice of the function exerted no influence on the test results and finally we took the maximum.

We wanted to prove if there are some regularities by the above mentioned events. Therefore we introduced 3 parameters influencing the hints for AUGUSTUS:

- the minimal score (the fragment-weight threshold);
- the maximal distance between the fragments (the criterion for their merging);
- cutoff value (how many bases are removed at both ends of a fragment).

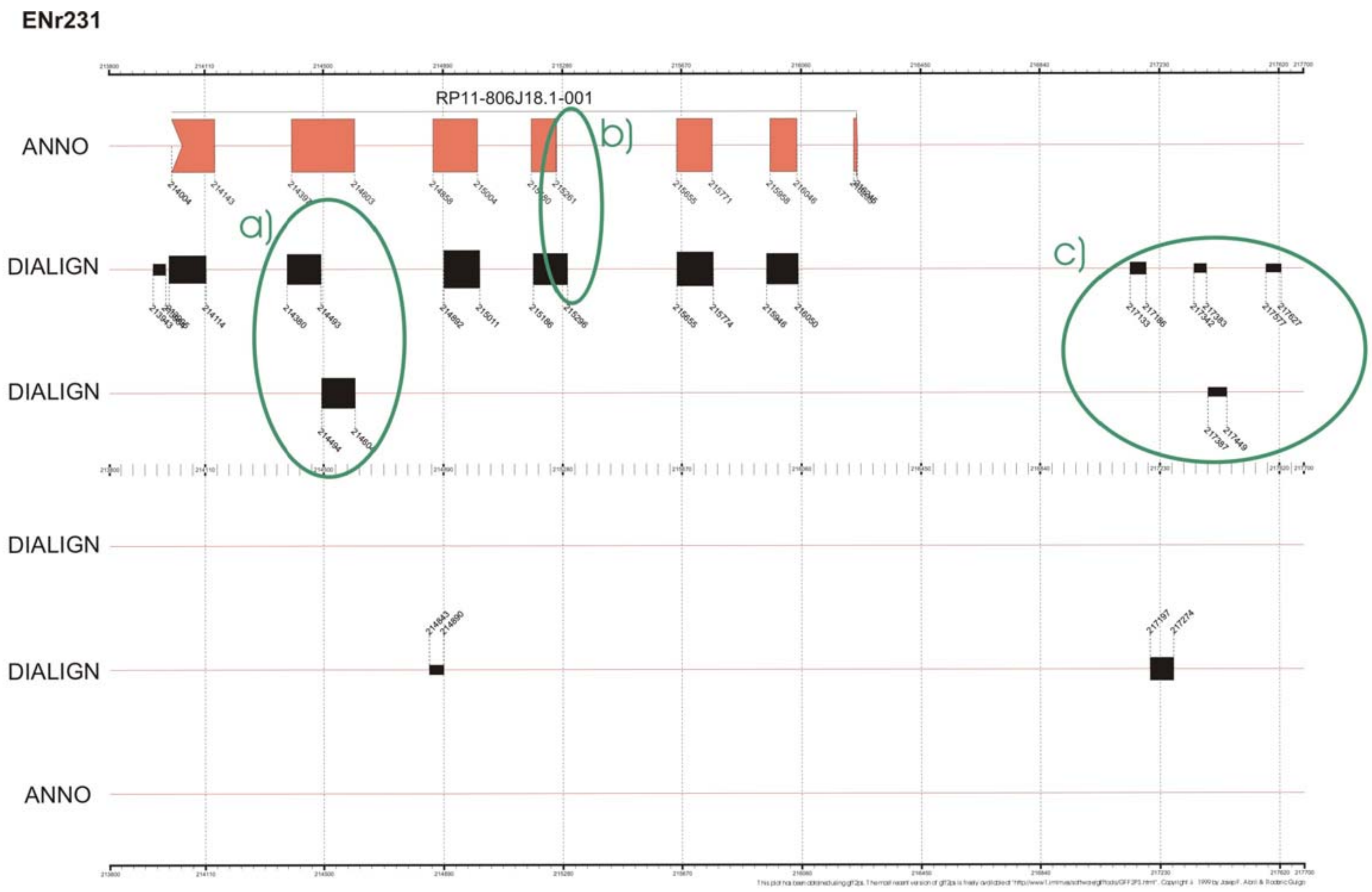


Fig. 6: Congruence between the DIALIGN pairwise local alignments (fragments) and the gene structure according to ENCODE annotation. The line with the ENCODE annotation is denoted with "ANNO" and the annotated Exons are shown as red blocks. The lines showing the DIALIGN alignments are denoted with "dialign". The DIALIGN-fragments are marked as black blocks where the height of the rectangles corresponds to the fragments' weight scores. For clarity two "dialign"-lines were used to represent fragments that lie close to each other. Both plus (above) and minus (bellow) DNA strands are shown. a) fragments appropriate for merging; b) fragment exceeding exon; c) fragments with low score not matching an exon.

The values of the parameters could not be easily determined, since there is no rule defining them. For example, on the one hand there are fragments with a very small score (e.g. short fragments) that are still matching exons and on the other hand fragments with high scores may not match any exon. Our task was to find the optimal values for these parameters.

We placed the hints generated with the above described method in files in General Feature Format (GFF) proposed by Richard Durbin and David Haussler. The full description of the format can be found at: http://www.sanger.ac.uk/Software/formats/GFF/.

If an option 'checkExAcc' of AUGUSTUS is switched on and a GFF file with hints and a file with annotated genomic DNA sequence in a genbank format are available, AUGUSTUS calculates the bonus- and malus-values for each class of score values (if classified) for the given hints (see Chapter 4). AUGUSTUS evaluates thus the reliability of the given hints. In our case that means how many of the hints are part of a coding sequence (or are matching exact an exon) and how many of them are not. We use these evaluations to tune the parameters of our approach. In the next chapter we explain which annotation we use.

6. Annotation of the human genome

To estimate the precision of our gene predictions we needed a comprehensive reliable annotation of the human genome or of a part of it. There are many databases, browsers and internet sites of scientific consortiums, universities, institutes containing versions of such annotation.

6.1 NCBI (National Center for Biotechnology Information)

For the NCBI databases the gene prediction is done in 3 different ways:

- Gene models are located using alignments between the RefSeq RNAs and the genomic sequence.
- Additional gene models are located using alignments of ESTs to the genomic sequence.
- Genes are predicted using the gene prediction program GENOMESCAN additionally aided by hints provided by protein homologies.

Following, depending on some conditions, the obtained gene models are combined to get a consensus set of gene models. A more detailed description of the annotation process by NCBI can be found under:

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.section.1486

6.2 Ensembl genome annotation

The Ensembl gene building pipeline [Curwen, Eyras et al. 2004] incorporates different approaches including *ab initio* gene predictions and homologies to the same or other species. The genes are predicted in three steps:

- All known human genes from SPTREMBL-database [Bairoch and Apweiler 2000] are aligned to the genome with PMATCH (R. Durbin, unpublished) and for each gene the best match is taken. These matches are then refined using GENEWISE [Birney and Durbin 2000] to provide an accurate gene structure.

- Paralogous human proteins and proteins from other organisms are aligned to the genome thus obtaining a set of novel human genes.
- Predictions are made with the *ab initio* program GENSCAN. Exons from these predictions are taken if they are supported by BLAST matches to proteins, vertebrate mRNA and UniGene clusters.

With these three steps create a set of transcripts which are grouped into genes wherever an exon is shared [Hubbard, Barker et al. 2002].

6.3 The ENCODE (ENCyclopedia Of DNA Elements) project

The ENCODE, the **Enc**yclopedia **o**f **D**NA **E**lements, [The ENCODE Project Consortium 2004] is a public research consortium initiated by the National Human Genome Research Institute (NHGRI) to carry out a project to identify all functional elements in the human genome sequence. The project consists of three phases: a pilot project phase, a technology development phase and a planned production phase. In the pilot phase a number of methods for annotation of the human genome are tested and compared. The aim is to construct a suite of tools for comprehensive finding of functional elements in the human genome. For the evaluation of the approaches are selected 44 diverse regions of the human genome representing ~1% or 30Mb of it (Tab.1). A part of the target regions are manually and a part of them are randomly selected. The intention by the selection of the regions is they to be representative. The aim is all annotated genes in those regions to be biologically confirmed through labor experiments and thus completely proved. The use of a uniform set enables the direct comparison of the different methods that are tested. With the purpose of evaluation of the methods for finding functional elements in genomic DNA sequence the establisher of the project published the annotation (biologically proved where current possible) of 13 of the 44 target regions thus making them a training set for the tested approaches. The rest of the 44 regions are provided as a test set.

The organizers of the ENCODE project have arranged a workshop to discuss and compare the different programs searching genes in the human genome. All the information about the participants in this workshop and the annotation and sequences of the target ENCODE regions can be found at: http://genome.imim.es/gencode/workshop2005.html

We found that the annotation of the human genome regions of the ENCODE project was most appropriate for training and testing of our approach.

Tab.1 The 44 ENCODE project target regions with their names, way or reasons of selection, chromosome to which they belong and approximate size.

Region	Description	Chr	Size (~Mb)
ENm001	CFTR	7	1.9
ENm002	Interleukin	5	1
ENm003	Apo Cluster	11	0.5
ENm004	Chr22 Pick	22	1.7
ENm005	Chr21 Pick	21	1.7
ENm006	ChrX Pick	X	1.2
ENm007	Chr19 Pick	19	1
ENm008	Alpha Globin	16	0.5
ENm009	Beta Globin	11	1
ENm010	HOXA Cluster	7	0.5
ENm011	1GF2/H19	11	0.6
ENm012	FOXP2	7	1
ENm013	Manual	7	1.1
ENm014	Manual	7	1.2
ENr111	Random	13	0.5
ENr112	Random	2	0.5
ENr113	Random	4	0.5
ENr114	Random	10	0.5
ENr121	Random	2	0.5
ENr122	Random	18	0.5
ENr123	Random	12	0.5
ENr131	Random	2	0.5
ENr132	Random	13	0.5
ENr133	Random	21	0.5
ENr211	Random	16	0.5
ENr212	Random	5	0.5
ENr213	Random	18	0.5
ENr221	Random	5	0.5
ENr222	Random	6	0.5
ENr223	Random	6	0.5
ENr231	Random	1	0.5
ENr232	Random	9	0.5
ENr233	Random	15	0.5
ENr311	Random	14	0.5
ENr312	Random	11	0.5
ENr313	Random	16	0.5
ENr321	Random	8	0.5
ENr322	Random	14	0.5
ENr323	Random	6	0.5
ENr324	Random	X	0.5
ENr331	Random	2	0.5
ENr332	Random	11	0.5
ENr333	Random	20	0.5
ENr334	Random	6	0.5

7. Results and discussion

To estimate the accuracy of our method we used the 44 regions of the ENCODE project (see Chapter 6.3). For optimization of the parameters that were introduced we took the 13 training regions with the preliminary given annotation. The tests were then made over the rest 31 target regions.

To assess the gene prediction accuracy we used the standard performance measures sensitivity and specificity. For the predicted features (gene, transcript, exon, and coding nucleotide) TP (True Positives) gives the number of the correctly predicted features, FP (False Positives) gives the number of features that are not correctly predicted (considered as feature, but not a part of the annotation) and FN (False Negatives) gives the number of the annotated features that are not predicted. The sensitivity is then:

$$Sn = \frac{TP}{TP + FN} \text{ or}$$

$$Sn = \frac{\#\{correctly\_predicted\}}{\#\{annotated\}}$$

Paraphrased, the sensitivity gives the probability that an annotated feature is correctly predicted.

The specificity is:

$$Sp = \frac{TP}{TP + FP} \text{ or}$$

$$Sn = \frac{\#\{correctly\_predicted\}}{\#\{predicted\}}$$

The specificity is the probability that a predicted feature is also annotated, i.e. the prediction of that feature was correct.

The accuracy of our approach is estimated at four levels.

Base level: A nucleotide is correctly predicted if it is a part of an annotated coding sequence region.

Exon level: A predicted exon is considered correct if its both splice sites are at the same positions as of an annotated exon.

Gene level: A gene is correctly predicted if all its exons are correctly predicted and no additional exons are predicted.

Transcript level: The same as gene level.

As already mentioned we used the DIALIGN option 'threshold value' for the fragments' weight score.

We introduced 3 threshold values for the human-mouse alignments of the UCSC: the threshold for the maximal distance between the alignments T1, for the maximal length of a large alignment T2, and for the minimum length of a large alignment T3 (see Chapter 5.1.2.2). We introduced also 3 parameters influencing the hints we created: the score threshold, the distance threshold and the reduction of fragments' size (see Chapter 5.2).

We tested a large number of values for all these parameters and on a principal of approximation achieved the following values that among the tested ones provided the best results:

- DIALIGN score threshold – 5
- T1 – 50 Kbase
- T2 – 100 Kbase
- T3 – 800 base
- Fragments' score threshold – 10
- Fragments' distance threshold – 0 (it was after all better not to merge the fragments)
- Reduction of the fragments – at 33 bases at each fragment's end.

To evaluate the accuracy of our approach we compared its' predictions to those of AUGUSTUS *ab initio*, DOGFISH-C, SGP2, TWINSCAN, TWINSCAN-MARS and N-SCAN. N-SCAN is an enhancement of TWINSCAN 2.0 [Gross and Brent 2005]. We took AUGUSTUS *ab initio* to investigate if our approach improves the gene localization. We used the other five programs as they are also based on comparative genomics (dual- or multiple-genome predictors). Another reason for choosing DOGFISH-C, TWINSCAN-MARS and N-SCAN was that they were also presented at the ENCODE gene prediction workshop. Thus their predictions on the same test human DNA sequence regions have been published. SGP2 and TWINSCAN did not participate at ENCODE workshop, but as they are established dual-genome predictors their predictions on the same regions were also published.

The predictions made over the 31 ENCODE test regions by TWINSCAN-MARS, N-SCAN and DOGFISH-C were downloaded from the folder "EGASP_on_all_regions" of the ENCODE gene prediction workshop's site. Those made by SGP2 and TWINSCAN were downloaded from the folder "ucsc_annotations_20050502" of the ftp directory of the ENCODE workshop's site.

The prediction accuracy of all approaches was computed using the annotation data in the folder "genes_known_validated" version dating from 07.06.2005 of the ENCODE gene prediction workshop's site. From the available annotation data only information concerning CDS (coding sequence) were used. The data concerning pseudogenes was excluded. The sources we used were: Known protein coding genes (referenced in Entrez Gene, NCBI) and Novel protein coding genes annotated by Havana (not referenced in Entrez Gene, NCBI).

All results were analyzed and compared using the EVAL package by E. Keibler and M. Brent (http://genes.cse.wustl.edu/) and are summarized in figure 7, with the exact values in table 2.

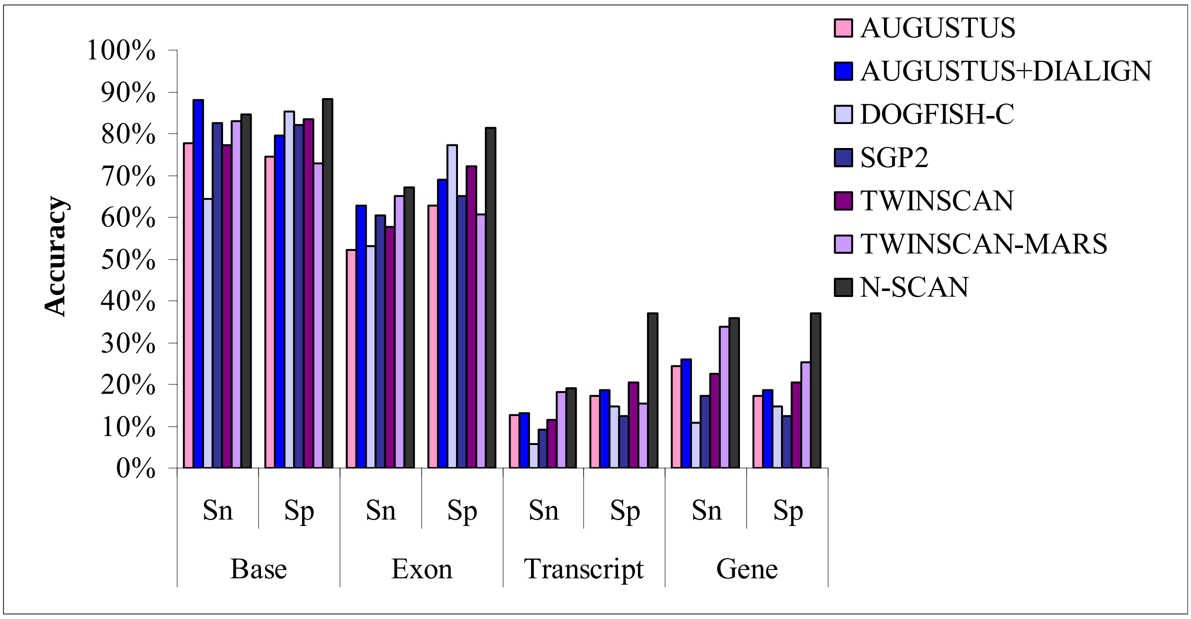


Fig. 7 Accuracy of the compared gene prediction programs. Illustrated are the sensitivity (Sn) and the specificity (Sp) at base, exon, transcript, and gene level for all compared programs.

Tab. 2 Sensitivity (Sn) and specificity (Sp) values for the compared gene prediction programs.

		AUGUSTUS	AUGUSTUS+ DIALIGN	DOGFISH-C	SGP2	TWINSCAN	TWINSCAN- MARS	N-SCAN
Base	Sn	77.79	88.14	64.38	82.58	77.15	82.89	84.52
	Sp	74.55	79.58	85.20	82.02	83.37	72.83	88.22
Exon	Sn	52.22	62.81	53.08	60.40	57.69	65.08	67.17
	Sp	62.76	68.90	77.34	64.97	72.22	60.76	81.44
Transcript	Sn	12.55	13.11	5.72	9.21	11.58	18.13	19.11
	Sp	17.22	18.64	14.61	12.35	20.55	15.29	37.06
Gene	Sn	24.32	26.01	10.81	17.23	22.64	33.78	35.81
	Sp	17.22	18.64	14.61	12.35	20.55	25.26	37.06

The graphic shows that our combined AUGUSTUS+DIALIGN approach is more sensitive and more specific then AUGUSTUS *ab initio* at all 4 levels. Usually the intrinsic-methods based programs tend to predict too many genes [Guigo, Agarwal et al. 2000]. The use of syntenic-sequence alignments reduces their false-positive rates and brings additional evidences of possible coding regions. Thus with AUGUSTUS+DIALIGN we achieved higher specificity and sensitivity than AUGUSTUS *ab initio*. A set of exact overlapping of exons of two or more predictions is the intersection of the sets of the predicted by them exons. An exon belongs to such intersection if its both splice sites are predicted by all respective predictions. Figure 8 shows the exact overlapping of the annotated exons and the exons predicted with AUGUSTUS *ab initio* and with AUGUSTUS+DIALIGN at 11 random selected ENCODE regions. It illustrates once more the improvement obtained through usage of extrinsic information. The number of exons overlooked by AUGUSTUS *ab initio* (208) is much larger than the number of exons overlooked by AUGUSTUS+DIALIGN (42), i.e. our combined approach is more sensitive. The number of exons that AUGUSTUS+DIALIGN predicts incorrect (166) is smaller than that predicted by AUGUSTUS *ab initio* (235) i.e. our combined approach is more specific. However the number of incorrectly predicted and overlooked exons by both AUGUSTUS+DIALIGN and AUGUSTUS *ab initio* (304 and 619 respectively) is considerably large.

AUGUSTUS+DIALIGN shows highest sensitivity at the base level among the compared programs. On the other hand its sensitivity at the other three levels is worse than that observed by TWINSCAN-MARS and N-SCAN which could occur due to wrong predicted (slightly shifted) splice sites by our approach, implying a big potential to improve our performance.

DOGFISH-C uses multiple alignments with 8 species (Human, Frog, Mouse, Rat, Dog, Chicken, Fugu, and Zebrafish) instead of pairwise human-mouse alignments. That could be one reason that the program is quite specific at base- and exon-level.

The lower sensitivity rates at transcript- compared to those at gene-level by all programs (excluding TWINSCAN-MARS) could be explained with the fact that they do not predict cases of alternative splicing leading to more then one transcript per gene. That explains also the identical specificity rates at both levels. For the 31 test regions in the annotation the average number of transcripts per gene is 2.42. Among the compared programs only TWINSCAN-MARS finds multiple transcripts from a single gene and thus having different specificity rates on transcript- and gene-level.

Except at base-level sensitivity, N-SCAN has the best accuracy among all compared programs. That could be explained with the fact that in the program are integrated several enhancements of the probabilistic model used by TWINSCAN 2.0. Its GHMM allows usage of multiple alignments, considers evolutionary relationships between the aligned organisms, and has additional states modeling 5' UTR structure [Brown, Gross et al. 2005] and other conserved non-coding sequence [Gross and Brent 2005].

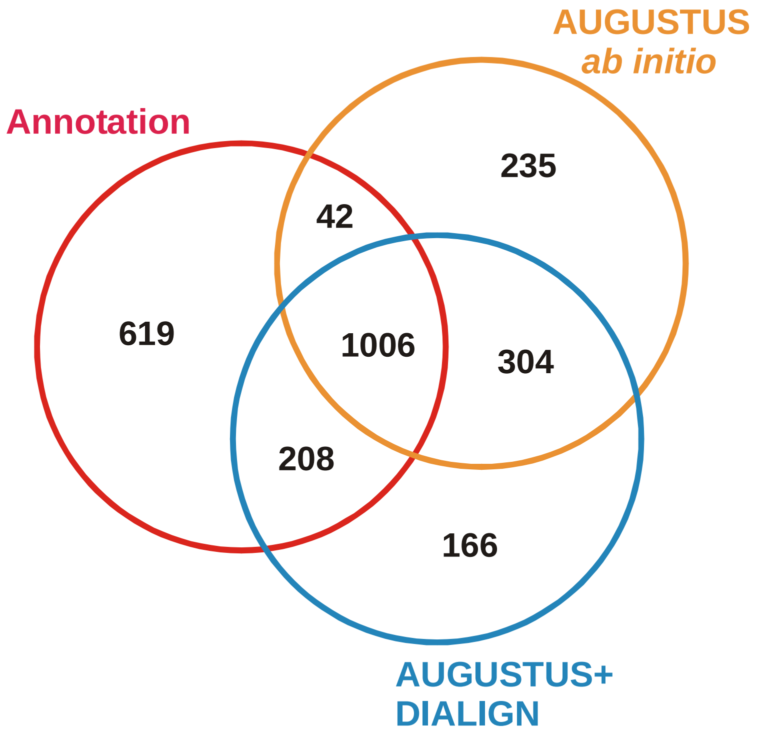


Fig. 8 Exact overlapping of the exons according ENCODE annotation and the exons predicted with AUGUSTUS *ab initio* and with AUGUSTUS+DIALIGN at 11 random selected ENCODE regions.

Figure 9 illustrates the exact overlapping of the annotated exons and the exons predicted with AUGUSTUS+DIALINGN and with N-SCAN at 11 random selected ENCODE regions. The considerably smaller number of incorrect predicted exons by N-SCAN (192) than that by AUGUSTUS+DIALINGN could be explained with the usage by N-SCAN of more than one informant organism thus improving the specificity. The exons that are predicted from both AUGUSTUS+DIALINGN and N-SCAN and are not annotated (52) could be used as base for biological experiments verifying new genes, which is the main motivation for development of gene prediction programs based on cross-species comparisons

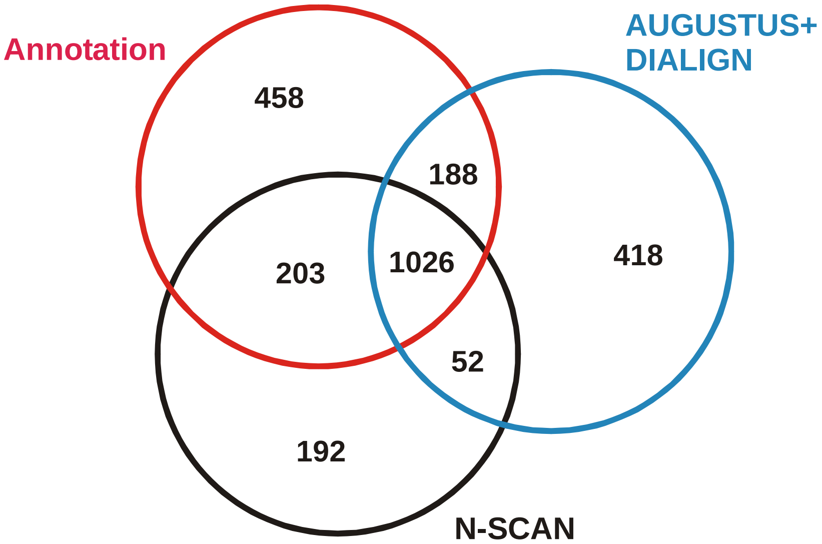


Fig. 9 Exact overlapping of the exons according ENCODE annotation and the exons predicted with AUGUSTUS+DIALIGN and with N-SCAN at 11 random selected ENCODE regions.

The usage of more informant organisms additional to the mouse could improve the accuracy of our approach. There are some more enhancement possibilities of out approach which could boost ifs performance. As shown in figure 10, AUGUSTUS+DIALIGN predicts correctly more exons than AUGUSTUS *ab initio* do. But still, 9 annotated exons are overlooked by both programs despite of the fact that there are hints matching them. One explanation, why these hints are "ignored", is that the hints generated with our approach are not sufficiently reliable. There are exons (as the first annotated one) that have not matching hints as well as there are hints not matching any exon. That implies lower influence on the prediction and is a motivation to modify the processing of the alignments obtained by UCSC. One of the generated hints lies on the opposite DNA strand, which is not rarely observed. That could be regarded through integration of 2 possibilities ("correct" and "wrong" DIALIGN-fragment's strand) in the bonus factors of AUGUSTUS (see Chapter 4). Another accuracy improvement of our approach could be achieved through a more complex usage of all DIALIGN's options and their further tuning.

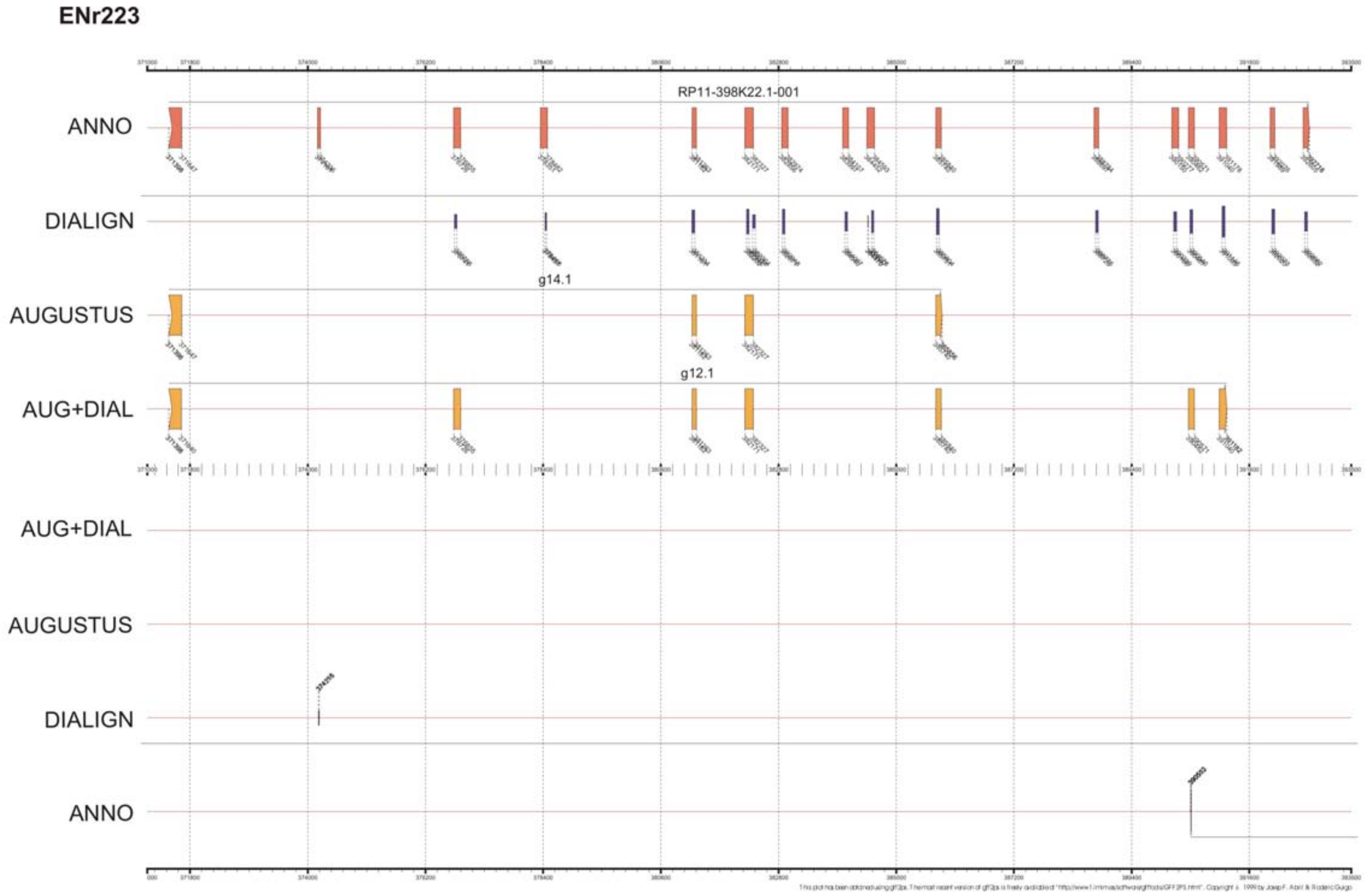


Fig. 10 Comparison between the generated hints, AUGUSTUS *ab initio*, AUGUSUTS+DIALIGN predictions and the gene structure according to ENCODE annotation. The lines with the ENCODE annotation (red), the hints (blue), the predicted exons from AUGUSTUS *ab initio* and AUGUSTUS+DIALIGN (orange) are according denoted with "ANNO", "DIALIGN", "AUGUSTUS" and "AUG+DIAL". The initial, internal and terminal exons are shown respectively as arrow-start, block and arrow-end. Both plus (above) and minus (bellow) DNA strands are shown. The figure was created using gff2ps by J.F. Abril and R. Guig□.

8. Conclusions

We have shown that the integration of appropriately processed human-mouse alignments generated from DIALIGN program improves both the sensitivity and the specificity of the *ab initio* gene prediction program AUGUSTUS. Further, the combination between DIALIGN and AUGUSTUS showed the best sensitivity at base level in comparison with other dual- or multiple alignments based programs, i.e. DOGFISH-C, SGP2, TWINSCAN, TWINSCAN-MARS and N-SCAN. However, at all other levels (exon, transcript, and gene) is the program N-SCAN more sensitive. N-SCAN is also more specific at all 4 levels. That is a motivation for us to include more organisms in our alignments to make the hints we generate more specific. Further improvements of the processing of these hints could also make them more reliable for the prediction with AUGUSTUS.

References

Abril, J. F. and R. Guigo (2000). "gff2ps: visualizing genomic annotations." Bioinformatics **16**(8): 743-4.

Alexandersson, M., S. Cawley, et al. (2003). "SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model." Genome Res **13**(3): 496-502.

Bairoch, A. and R. Apweiler (2000). "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000." Nucleic Acids Res **28**(1): 45-8.

Birney, E., M. Clamp, et al. (2004). "GeneWise and Genomewise." Genome Res **14**(5): 988-95.

Birney, E. and R. Durbin (2000). "Using GeneWise in the Drosophila annotation experiment." Genome Res **10**(4): 547-8.

Brown, R. H., S. S. Gross, et al. (2005). "Begin at the beginning: predicting genes with 5' UTRs." Genome Res **15**(5): 742-7.

Brudno, M. and B. Morgenstern (2002). "Fast and sensitive alignment of large genomic sequences." Proc IEEE Comput Soc Bioinform Conf **1**: 138-47.

Burge, C. and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA." J Mol Biol **268**(1): 78-94.

Curwen, V., E. Eyras, et al. (2004). "The Ensembl automatic gene annotation system." Genome Res **14**(5): 942-50.

Flicek, P., E. Keibler, et al. (2003). "Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map." Genome Res **13**(1): 46-54.

Gross, S. S. and M. R. Brent (2005). Using Multiple Alignments to Improve Gene Prediction. RECOMB 2005. S. Miyano. Berlin Heidelberg, Springer-Verlag.

Guigo, R., P. Agarwal, et al. (2000). "An assessment of gene prediction accuracy in large DNA sequences." Genome Res **10**(10): 1631-42.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A **89**(22): 10915-9.

Hubbard, T., D. Barker, et al. (2002). "The Ensembl genome database project." Nucleic Acids Res **30**(1): 38-41.

International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome." Nature **431**(7011): 931-45.

Kulp, D., D. Haussler, et al. (1996). "A generalized hidden Markov model for the recognition of human genes in DNA." Proc Int Conf Intell Syst Mol Biol **4**: 134-42.

Morgenstern, B. (1999). "DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment." Bioinformatics **15**(3): 211-8.

Morgenstern, B. (2000). "A space-efficient algorithm for aligning large genomic sequences." Bioinformatics **16**(10): 948-9.

Morgenstern, B. et al. (2002). "Exon discovery by genomic sequence alignment." Bioinformatics **18**(6): 777-87.

Parra, G., P. Agarwal, et al. (2003). "Comparative gene prediction in human and mouse." Genome Res **13**(1): 108-17.

Parra, G., E. Blanco, et al. (2000). "GeneID in Drosophila." Genome Res **10**(4): 511-5.

Sakharkar, M. K., V. T. Chow, et al. (2004). "Distributions of exons and introns in the human genome." In Silico Biol **4**(4): 387-93.

Schwartz, S., W. J. Kent, et al. (2003). "Human-mouse alignments with BLASTZ." Genome Res **13**(1): 103-7.

Stanke, M. (2003). Gene Prediction with a Hidden Markov Model. Goettingen, Georg-August-Universitaet, Germany.

Stanke, M. and S. Waack (2003). "Gene prediction with a hidden Markov model and a new intron submodel." Bioinformatics **19 Suppl 2**: II215-II225.

The ENCODE Project Consortium (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." Science **306**(5696): 636-40.

Waterston, R. H.,K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Yeh, R. F., L. P. Lim, et al. (2001). "Computational inference of homologous gene structures in the human genome." Genome Res **11**(5): 803-16.

Acknowledgments

On the first place, I want to thank Prof. Dr. Burkhard Morgenstern for giving me the possibility to work in his department and for his support during my work. Special thanks are given to Dr. Mario Stanke for the comprehensive supervising of my work and for all his experience and knowledge that he readily shared with me. My biggest acknowledgments are for my husband for his patience and steadily support. I want to thank all my colleges from the department of bioinformatics for their help and for the friendly ambient there.