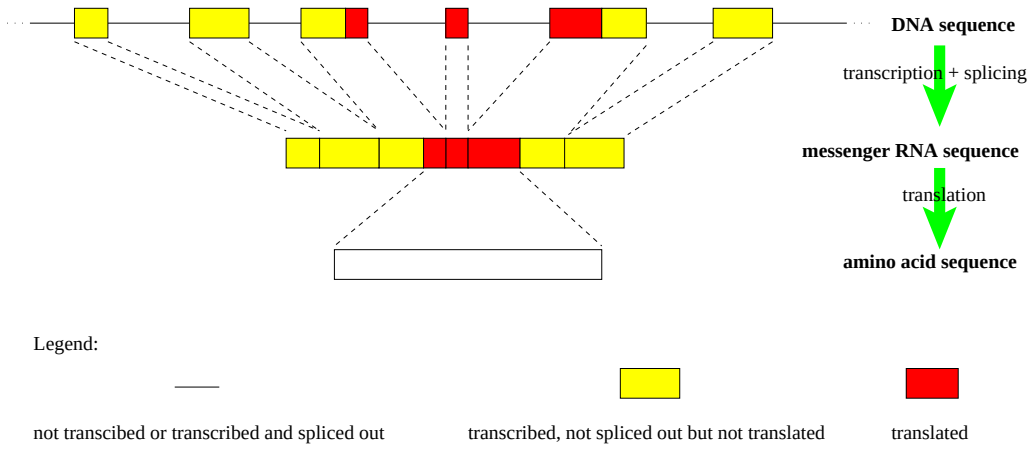
Gene Prediction

October 2005
Mario Stanke, mstanke@gwdg.de

So far, 39 eukaryotic genomes have been sequenced and their sequence published (Genomes OnLine Database GOLD). Examples are the bakers yeast (1997), the worm *Caenorhabditis elegans* (1998), the fruit fly *Drosophila melanogaster* (2000), the plant *Arabidopsis thaliana*, the human (2001), the malaria parasite *Anopheles gambiae* (2002) and the mouse (2002). 531 additional eukaryotic genome sequencing projects are currently under way. For prokaryotes these figures are even higher.
These sequencing efforts generate a large amount of raw data as the DNA sequence of an eukaryote is often longer than a hundred million base pairs. The genome of humans has approximately a size of $3.2 \cdot 10^9$ base pairs. The annotation of these sequences by biological means can by far not keep pace with the speed with which the data is accumulated and therefore computational tools to find genes are needed. Locating the genes is helpfull and often even a prerequisite for further analysis such as the characterisation of the function of the gene product, determining the phylogeny of different species or understanding gene regulation.

The picture below shows a eukaryotic genomic sequence containing one gene. The red (dark) parts are coding exons. Gene prediction programs usually try to predict these coding exons and group predicted exons to genes.



Methods of Gene Prediction:

- signal sensors
- content sensors

} ab initio methods

- alignment with cDNA
- alignment with ESTs
- protein homology
- cross-species DNA comparison

} extrinsic methods

The first two of these methods are so-called *ab initio* methods. 'Ab initio' here means that only the sequence under investigation itself is used for prediction. These methods base solely on statistical models of genes, mainly signal and content sensors. A signal sensor tries to identify signals by locally examining the sequence around a possible signal site. Signals are short sequence segments of the DNA, that control translation or transcription (such as the promoter region, the splice sites, the translation start region). Content sensors try to distinguish coding from non-coding sequences. The most commonly used content sensors of current gene prediction programs base on reading frame dependent hexamer frequencies. Ab initio methods have the advantage that no other evidence about the gene structure such as ESTs or homology to known proteins is needed. However, they are on average less accurate species-specific.

Extrinsic methods use other external information about the input sequence such as the results of a database search (protein (e.g. GenomeScan), EST or cDNA data) or a second syntenic DNA sequence (that codes for the same or a similar protein). Programs that make use of extrinsic methods often also include statistical models as the ab initio methods.

The reliability of gene prediction depends on the organism and – if applicable – strongly on the quality and quantity of the extrinsic information. For humans the currently best *ab initio* methods have about the following accuracy on large contigs.

- 60% of the exons (of all splice variants) are predicted correctly (exon correct if both splice sites correct)
- 25% of all genes are predicted completely correct (one splice variant is predicted)

	this course	http://gobics.de/mario/methodscourse/
	GOLD	http://www.genomesonline.org
	AUGUSTUS	http://augustus.gobics.de
	AGenDA	http://bibiserv.techfak.uni-bielefeld.de/agenda/
Web resources:	GENSCAN	http://genes.mit.edu/GENSCAN.html
	GenomeScan	http://genes.mit.edu/genomescan.html
	DoubleScan	http://www.sanger.ac.uk/Software/analysis/doublescan/
	BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
	gff2ps	http://bioweb.pasteur.fr/seqanal/interfaces/gff2ps.html