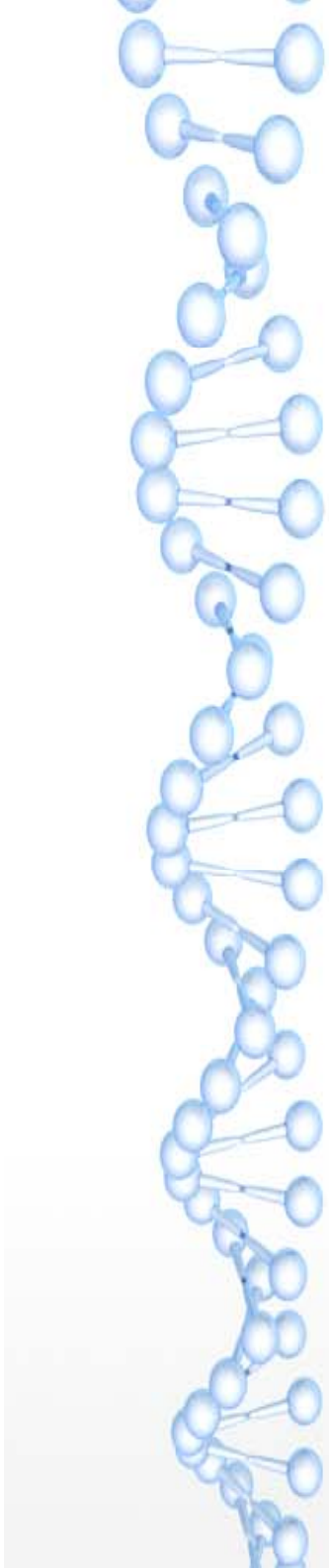


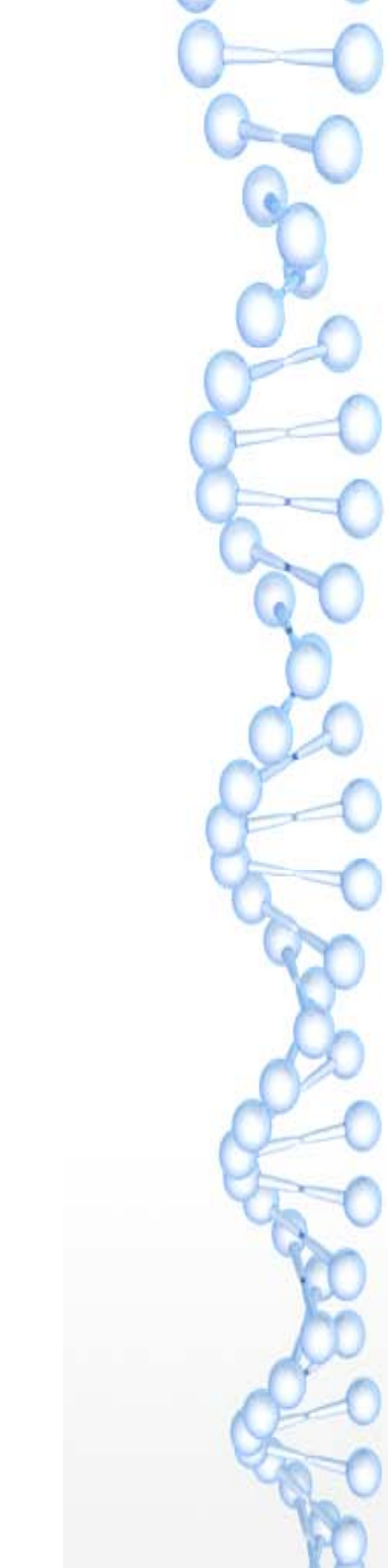
Introduction into Genome Analysis

Kim Weißer

Introduction into Genome Analysis

- Basics

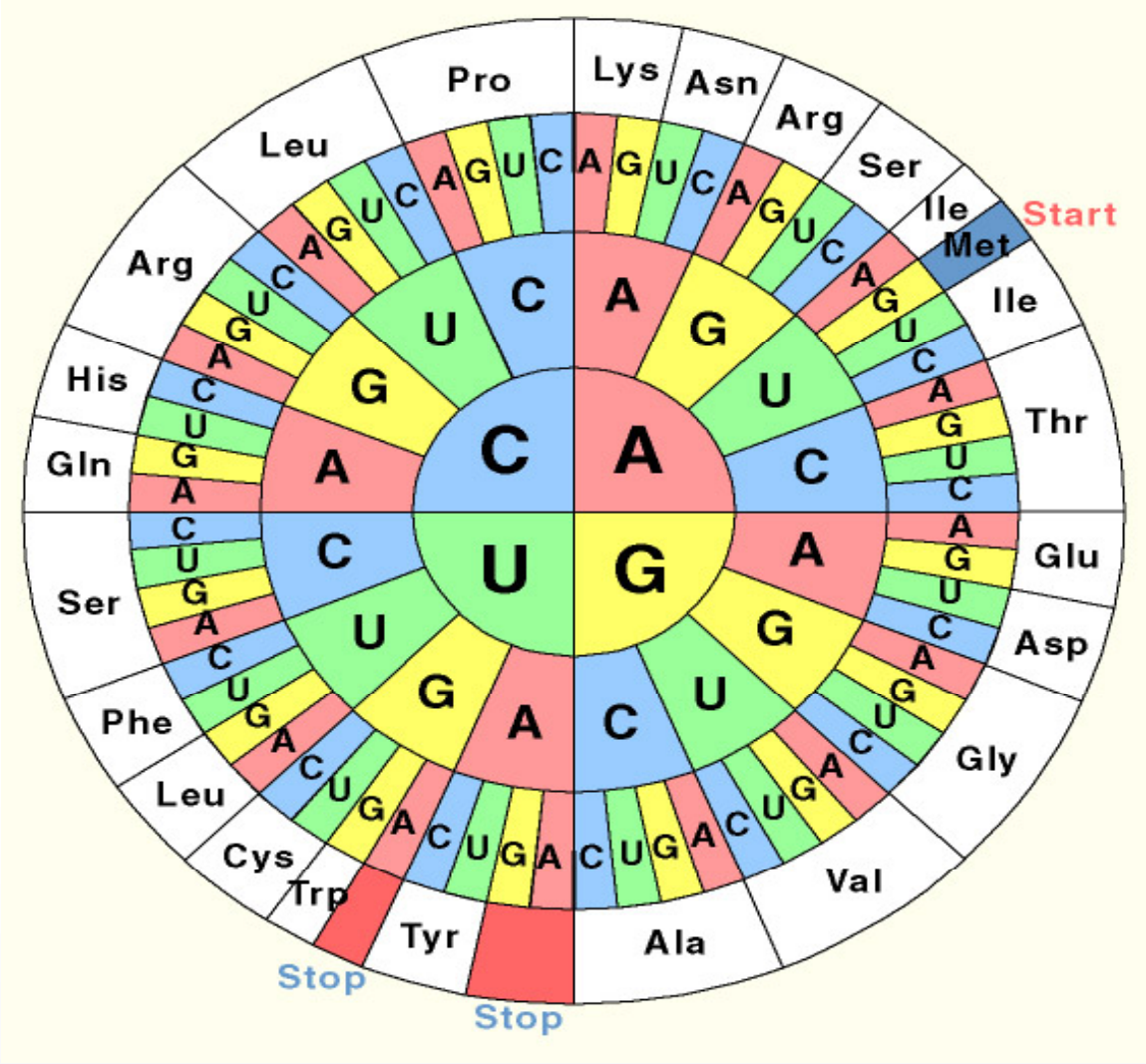
- ORF (Open reading frame)
- Accuracy of predicting programs
- Analyzing tRNA
- Homology searches
- Exon prediction
- Splice site prediction
- Promoter-prediction
- Conclusion



Basics

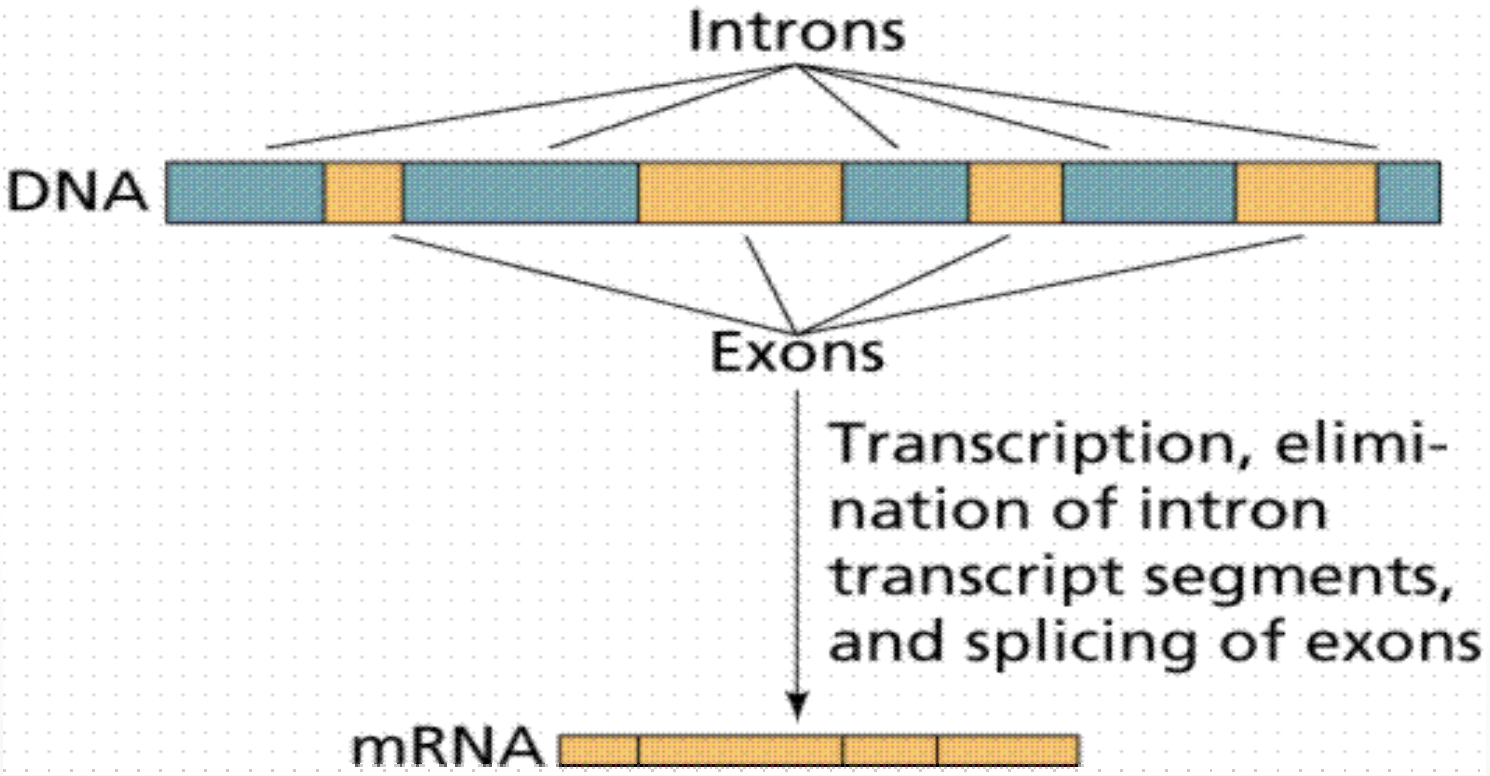
- Proteins, Peptides, functional RNA
- MRNA, tRNA, cRNA, ncRNA
- read in codons

Codon Wheel


Pro
Lys
Asn
Arg
Ser
Ile
Start
Met
Leu
Arg
His
Gln
Ser
Phe
Leu
Cys
Trp
Stop
Tyr
Stop
Ala
Val
Gly
Asp
Glu
Thr
Ile
U
C
A
G

Exons/Introns

- only in Eukaryotes
- Splicen is called the process which cuts out of heterogeneous nuclear RNA (hnRNA) and combinds the exons to mRNA



Open Reading Frame

- The region which codes a gene
- Starts with Startcodon (AUG)
- Ends with Stopcodon (UAA, UAG , UGA)

Possible consequences for the translated protein of mistakesin the prediction of an exon

Start of exon	Length of exon	Effect on translation of this exon	Effect on translation of correctly starting next exon
Correct	Correct	Correct	Correct
	Incorrect, correct frame	Correct, but extra or missing residues	Correct exept possibly the first residue
	Incorrect,wrong frame	Correct, but extra or missing residues	Incorrect
Incorrect, correct frame	Correct	Correct, but extra or missing residues	Correct exept possibly the first residue
	Incorrect, correct frame	Correct, but extra or missing residues	Correct exept possibly the first residue
	Incorrect,wrong frame	Correct, but extra or missing residues	Incorrect
Incorrect, wrong frame	Correct	Incorrect	Incorrect
	Incorrect, correct frame	Incorrect	Incorrect
	Incorrect,wrong frame	Incorrect	Possibly correct if the two first exon frameshifts cancel

Analysis Programs

MZEF	Michael Zhang's Exon Finder Webbased,
GeneMark	A family of gene prediction programs developed at Georgia Institute of Technology, Atlanta, Georgia, USA.
FirstEF	FirstExonFinder Webbased, especially used for predicting the first exon. decision tree
Orpheus	Vergleiche, Codonstatistiken, Ribiosom-Bindestellen Bakterielles Genom
Glimmer	Interpolated Markov Modeler, prokaryote-gene finding tool, free (including source code) with registration for non-commercial use
FGENESH	HMM-based gene structure prediction
GRAIL	Gene Relationships Among Implicated Loci

HMMGene	Web based The program is based on a hidden Markov model.
AAT (Analysis and Annotation Tools)	Includes two sets of programs, one for comparing the query sequence with a protein database and the other for comparing the query with a cDNA database.
GeneBuilder	based on prediction of functional signals and coding regions by different approaches in combination with similarity searches in proteins and databases.
Twinscan	uses similarity between species

Introduction into Genome Analysis

- Basics
- ORF (Open reading frame)
- Accuracy of predicting programs
- Analyzing tRNA
- Homology searches
- Exon prediction
- Splice site prediction
- Promoter-prediction
- Conclusion

Confirming Predictions

Program	Sensitivity	Specificity	Missed Exons %	Wrong Exons %
FGENSH	77.1	65.7	9.6	23.2
GenScan	66.5	44.9	12.0	40.9
HMMGene	69.6	36.6	15.5	55.5

Confirming Predictions

- TP: True positive
- TN: True negative
- FP: False positive

	Reality		
Predicted		c	nc
	c	TP	FP
	nc	FN	TN

Sensitivity

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

Specificity

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

Confirming Predictions

AC: approximate correlation coefficient

ACP: average conditional probability

$$AC = 2 \square ACP - 1$$

$$ACP = \frac{1}{4}\left[\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right]$$

Confirming Predictions

AE: actual exons

CE: correct exons

PE: predicted exons

ME: missing exons

WE: wrong exons

Sensitivity

$$Sn_1 = \frac{CE}{AE}$$

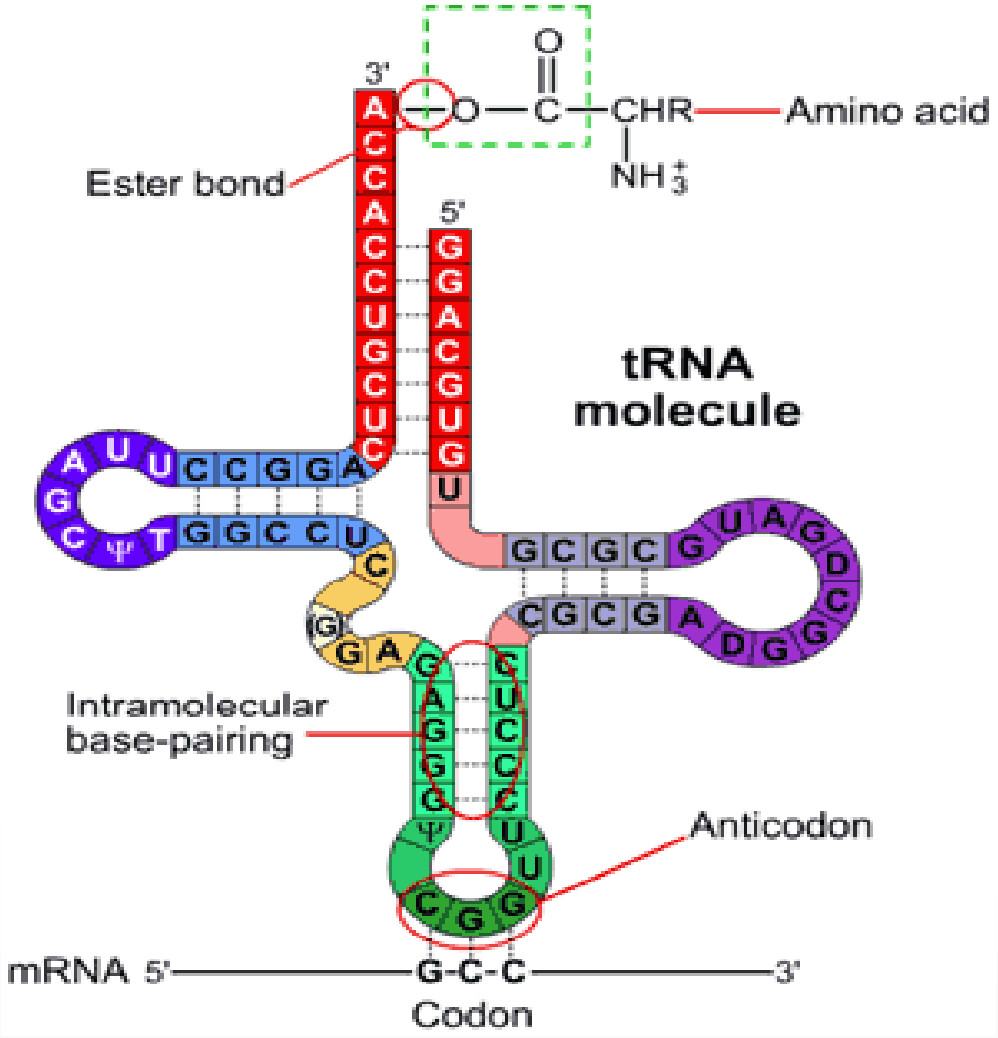
$$Sn_2 = \frac{ME}{AE}$$

Specificity

$$Sp_1 = \frac{CE}{PE}$$

$$Sp_2 = \frac{WE}{PE}$$

tRNA


3'
5'
Ester bond
Amino acid
tRNA molecule
Intramolecular base-pairing
Anticodon
mRNA 5'
G-C-C
3'
Codon

GeneMark

- One of the few prokaryontic gene prediction programs which are as well use full for Eukaryots

$$P(a|b_1 b_2 b_3 b_4 b_5)$$

- developed in 1993

GeneMark

$$P(a|b_1 b_2 b_3 b_4 b_5) = \frac{n_{b_1 b_2 b_3 b_4 b_5 a}}{\sum_{x=A,C,G,T} n_{b_1 b_2 b_3 b_4 b_5 x}}$$

Each frame has its own probabilities

$$P_1(a|b_1 b_2 b_3 b_4 b_5), P_2(a|b_1 b_2 b_3 b_4 b_5), \ldots$$

GeneMark

$$x = x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9$$

Possibility of this sequence to be in the 2. reading frame

$$P(x|2) = P_1(x_1 x_2 x_3 x_4 x_5) * P_1(x_6|x_1 x_2 x_3 x_4 x_5) * P_2(x_7|x_2 x_3 x_4 x_5 x_6) * P_3(x_8|x_3 x_4 x_5 x_6 x_7) * P_1(x_9|x_4 x_5 x_6 x_7 x_8)$$

Likelihood that a segment of x is in coding frame 2 (P(2|x))

$$P(2|x) = \frac{P(x|2)P(2)}{P(x|nc)P(nc) + \sum_{m=1}^{6} P(x|m)P(m)}$$

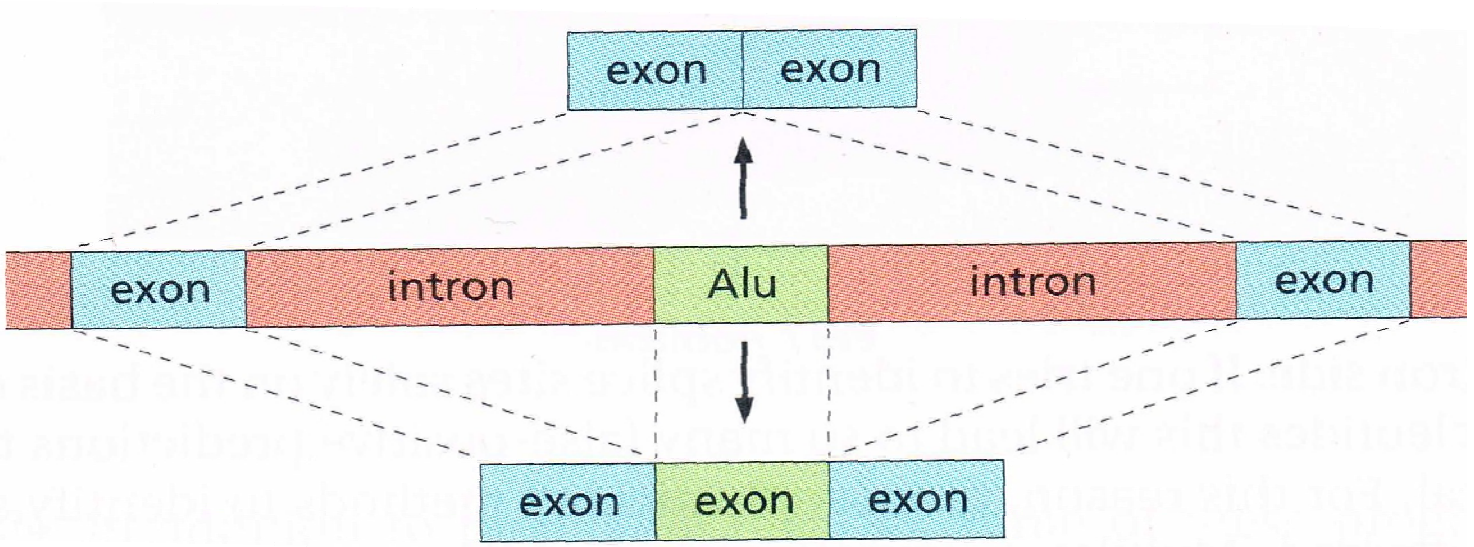
Introduction into Genome Analysis

- Basics
- ORF (Open reading frame)
- Accuracy of predicting programs
- Analyzing tRNA
- Homology searches
- Exon prediction
- Splice site prediction
- Promoter-prediction
- Conclusion

Homology searches

- Is possible for both, Pro- and Eukaryonte
- Comparison with allready detected sequences
- Diffrent spezies

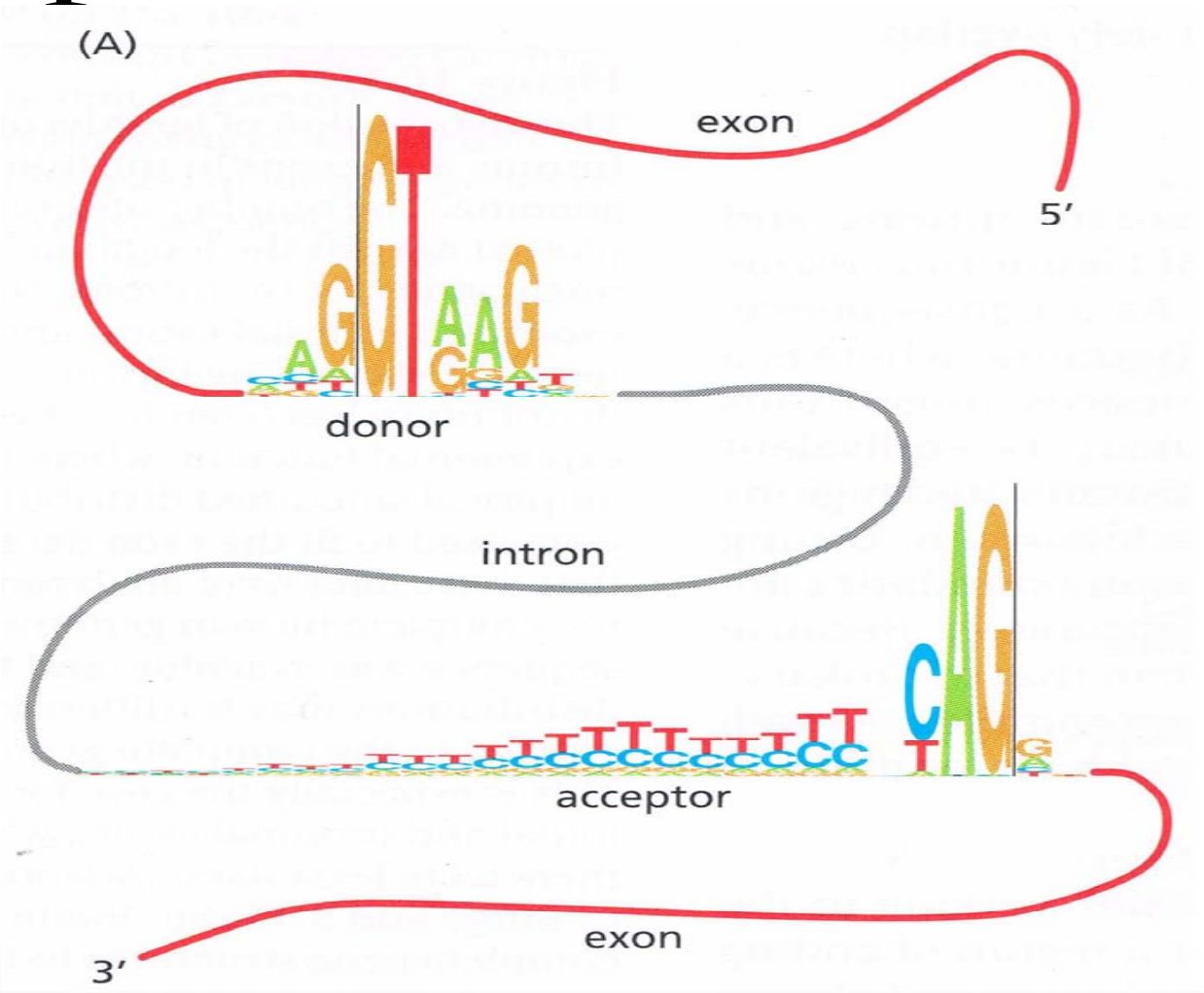
Alternative Splicing

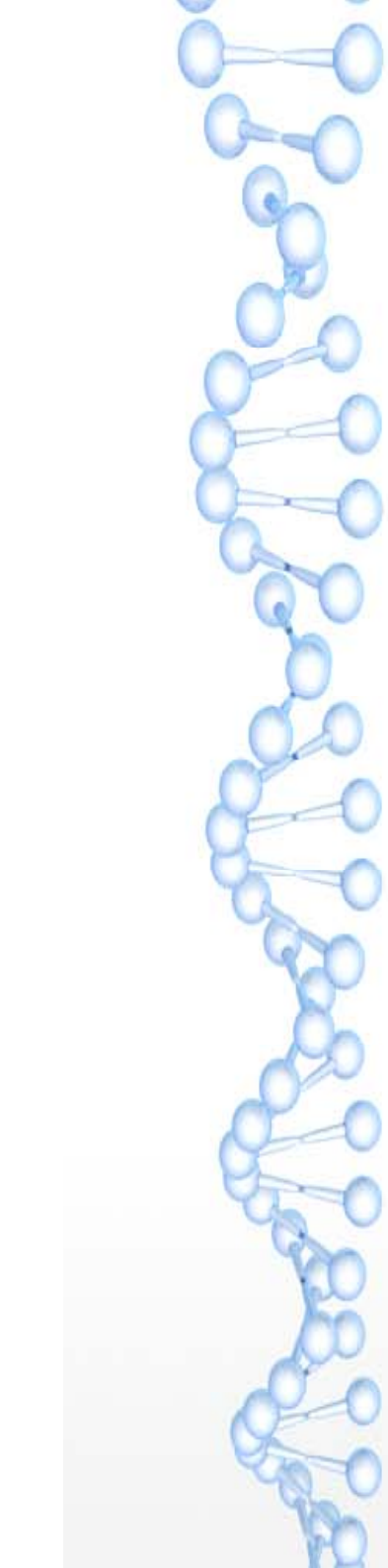


Introduction into Genome Analysis

- Basics
- ORF (Open reading frame)
- Accuracy of predicting programs
- Analyzing tRNA
- Homology searches
- Exon prediction
- **Splice site prediction**
- Promoter-prediction
- Conclusion

Splice site detection





Internal exon detection

- MZFE
- Specially designed to predict internal exons
- Uses search with pattern

Exon detection

Human ALDH10 gene		
Program	**Exon 1**	**Exon 2**
Experimental	1352-1762	2169-2400
MZEF	1601-1762	_
GeneMark	1610-1762	2169-2400
HMMGene	1610-1762	2169-2400
FGENSH	1542-1694	2226-2400
GeneScan	1610-1762	2169-2400
GrailEXP	1610-1762	2169-2459

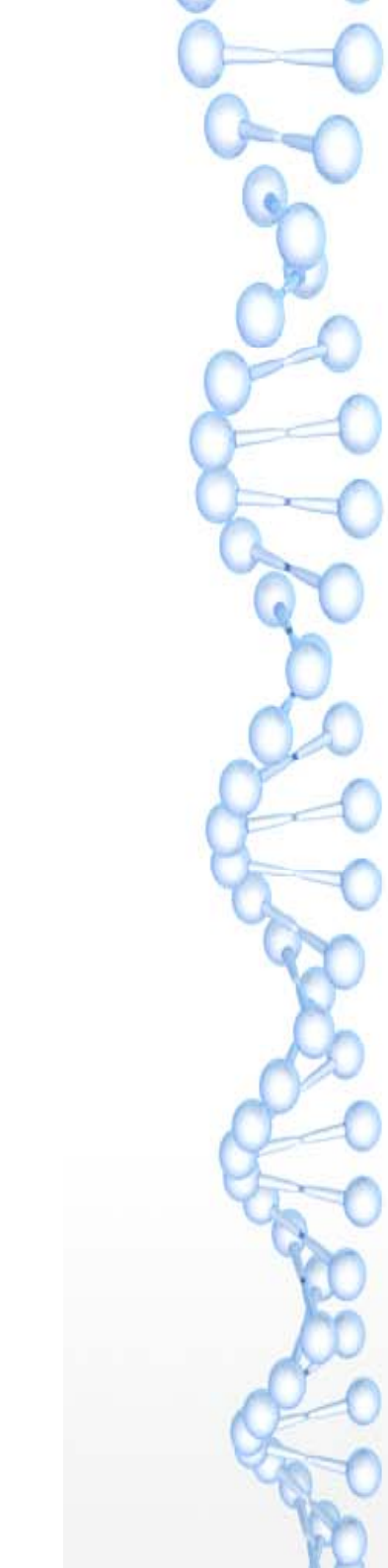
Promoter

- Transcription start site
- TATA-Box
- E.coli: Pribnow-Box (TATAAT), AT-rich region, TTGACA-Box
- CG-Islands

Analysis Programs

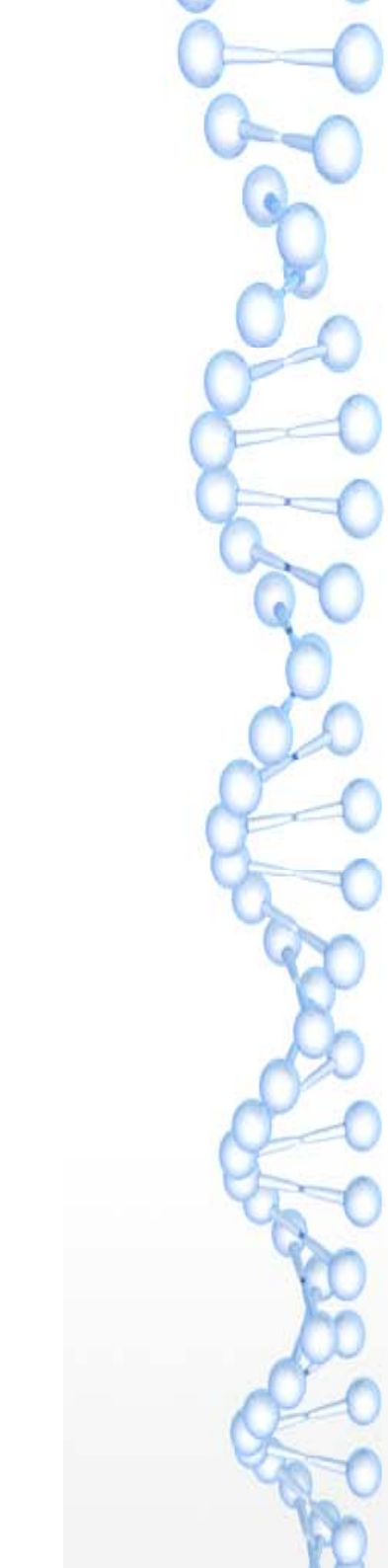
MZEF	Michael Zhang's Exon Finder Webbased,
GeneMark	A family of gene prediction programs developed at Georgia Institute of Technology, Atlanta, Georgia, USA.
FirstEF	FirstExonFinder Webbased, especially used for predicting the first exon. decision tree
Orpheus	Vergleiche, Codonstatistiken, Ribiosom-Bindestellen Bakterielles Genom
Glimmer	Interpolated Markov Modeler, prokaryote-gene finding tool, free (including source code) with registration for non-commercial use
FGENESH	HMM-based gene structure prediction
GRAIL	Gene Relationships Among Implicated Loci

HMMGene	Web based The program is based on a hidden Markov model.
AAT (Analysis and Annotation Tools)	Includes two sets of programs, one for comparing the query sequence with a protein database and the other for comparing the query with a cDNA database.
GeneBuilder	based on prediction of functional signals and coding regions by different approaches in combination with similarity searches in proteins and databases.
Twinscan	uses similarity between species



Steps in eukaryontic gene prediction

- Submit DNA sequence to exon prediction programs
- Take average or consensus exon prediction
- Translate predicted



?