Gene Prediction

Talk *Gene Prediction* on Jan. 14th, 2011

Mario Stanke
Institut für Mathematik und Informatik
Universität Greifswald

Outline

What you can learn from this talk

- Introduction: What is predicted? Why? Based on what?
- Methods and Results:
  - some intuition behind gene prediction methods
  - approaches:
    evidence integration, comparative, protein-family-based
  - programs and (in)accuracy

Structure of a eukaryotic gene

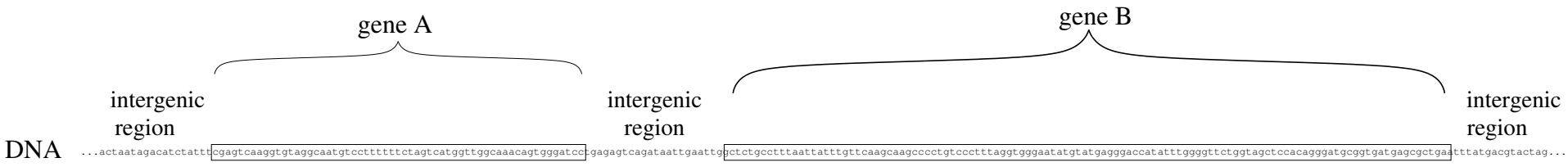
DNA ...actaatagacatctatttcgagtcaaggtgtaggcaatgtccttttttctagtcatggttggcaaacagtgggatcctgagagtcagataattgaattggctctgcctttaattatttgttcaagcaagcccctgtccctttaggtgggaatatgtatgagggaccatatttggggttctggtagctccacagggatgcggtgatgagcgctgaatttatgacgtactag...

Structure of a eukaryotic gene

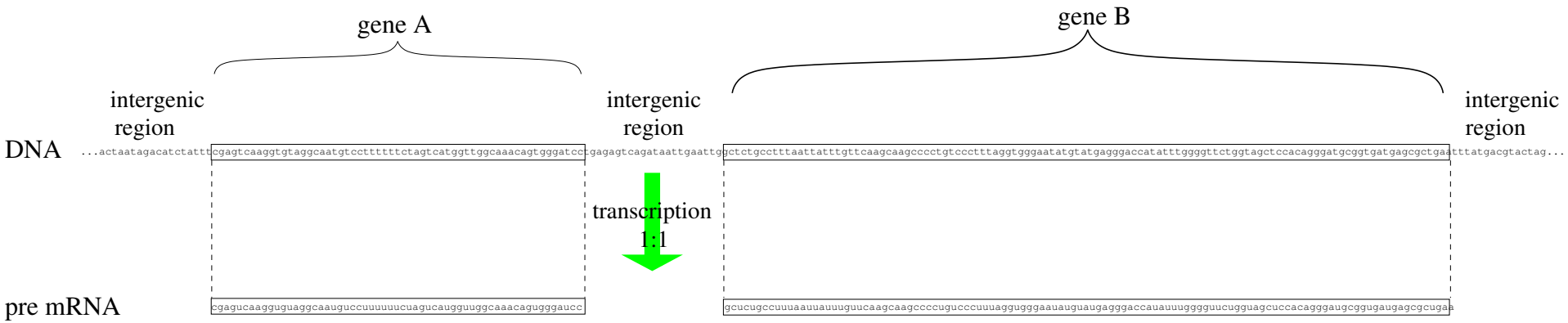
DNA ...actaatagacatctatttcgagtcaaggtgtaggcaatgtccttttttctagtcatggttggcaaacagtg 

aattatttgttcaagcaagcccctgtccctttaggtgggaatatgtatgagggaccatatttggggttctggtagctccacagggatgcggtgatgagcgctgaatttatgacgtactag...

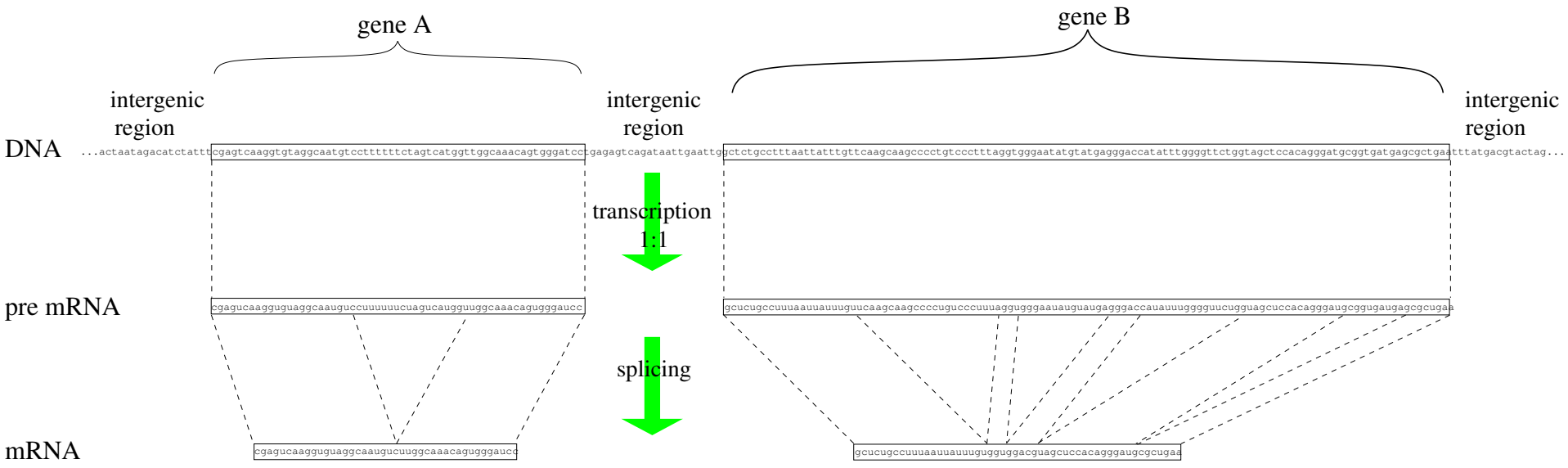
Structure of a eukaryotic gene



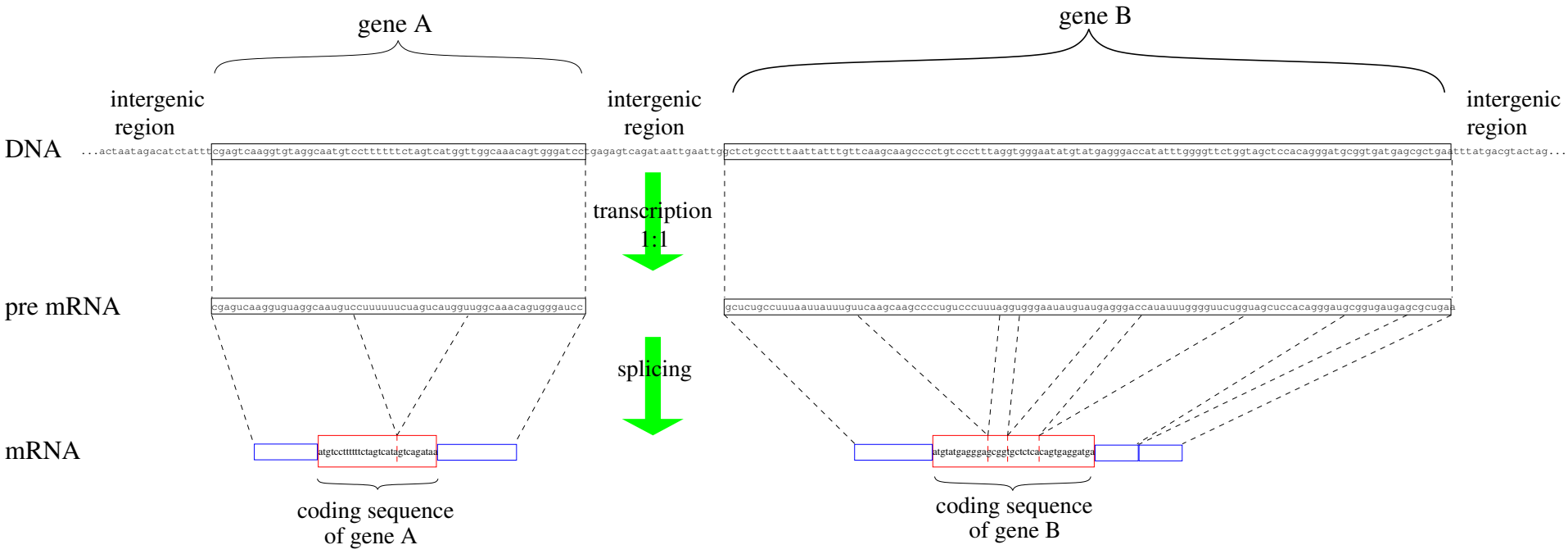
Structure of a eukaryotic gene



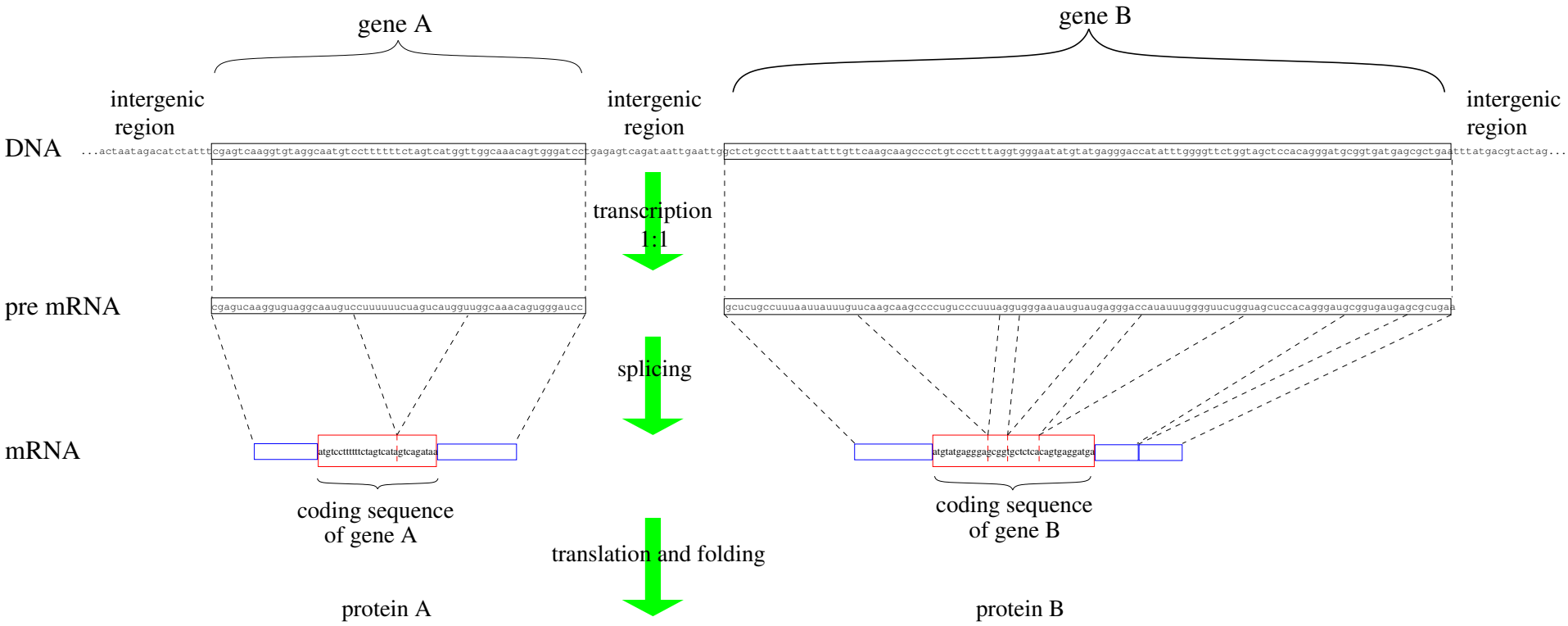
Structure of a eukaryotic gene



Structure of a eukaryotic gene



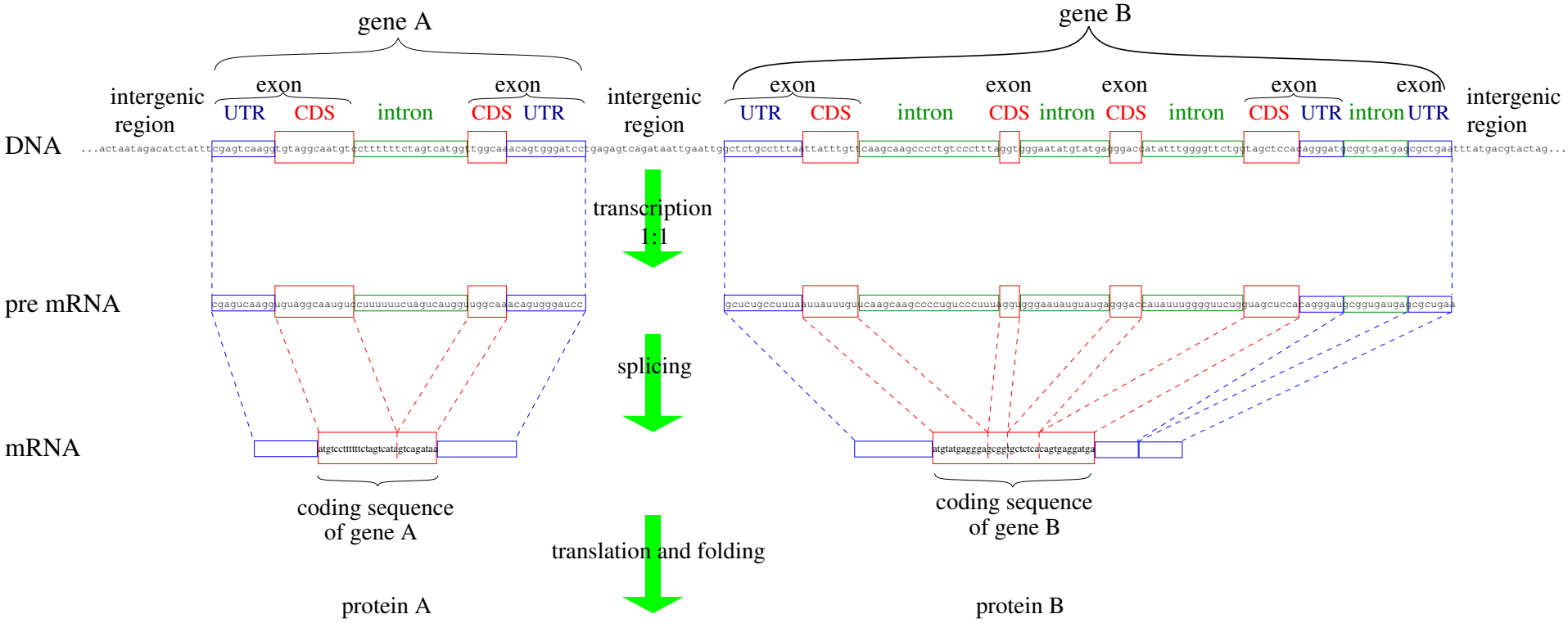
Structure of a eukaryotic gene



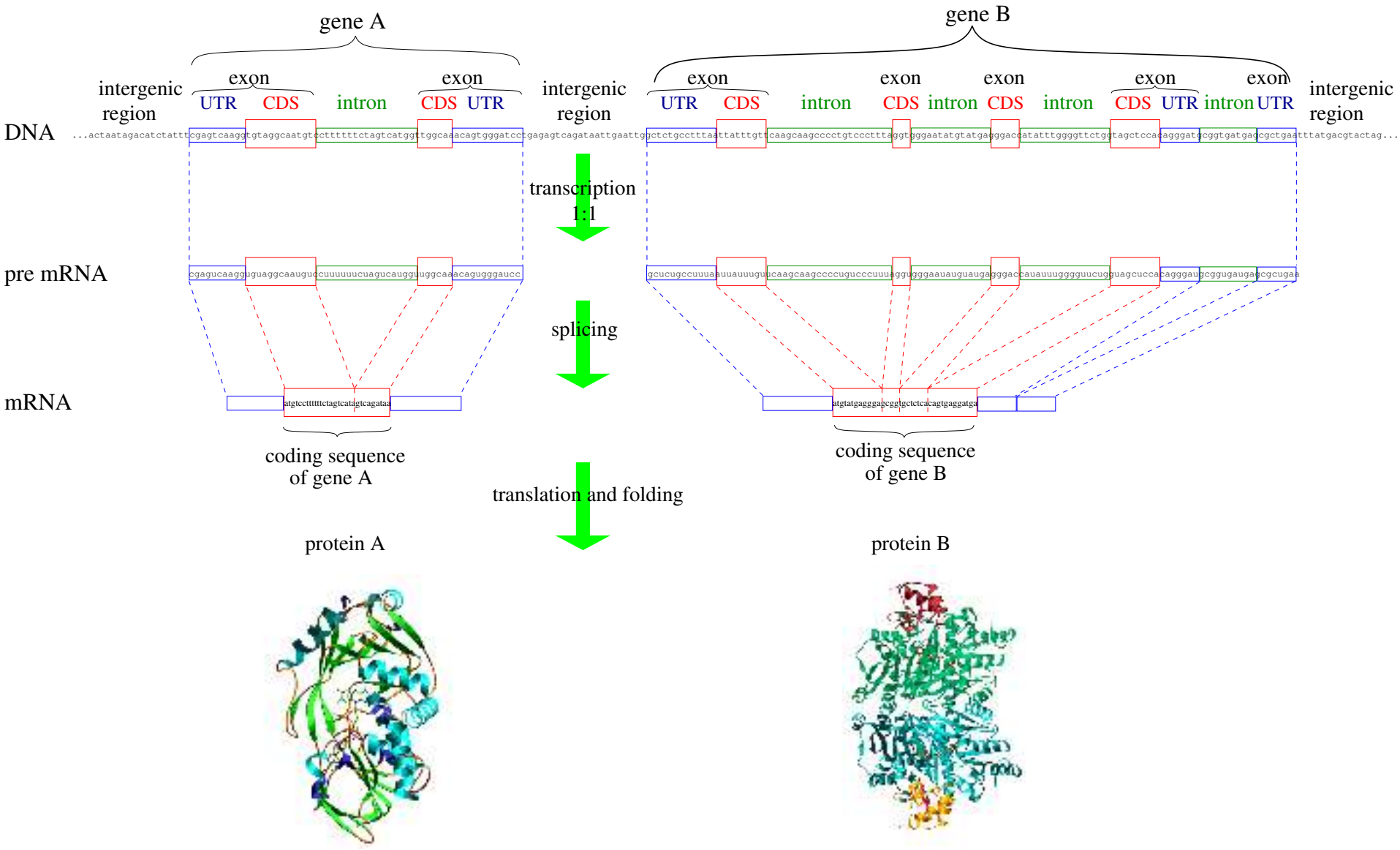
Structure of a eukaryotic gene

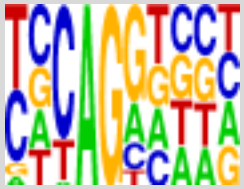
UTR = UnTranslated Region = part of mRNA that is not translated

CDS = CcoDing Sequence = part of mRNA (exon) that is translated

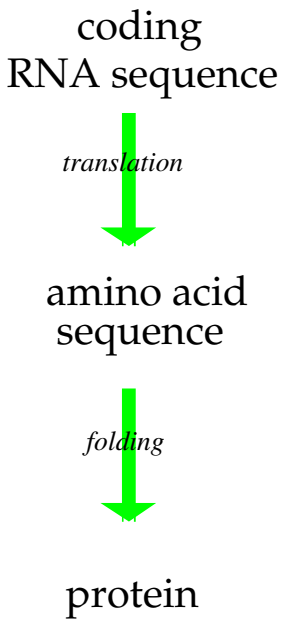


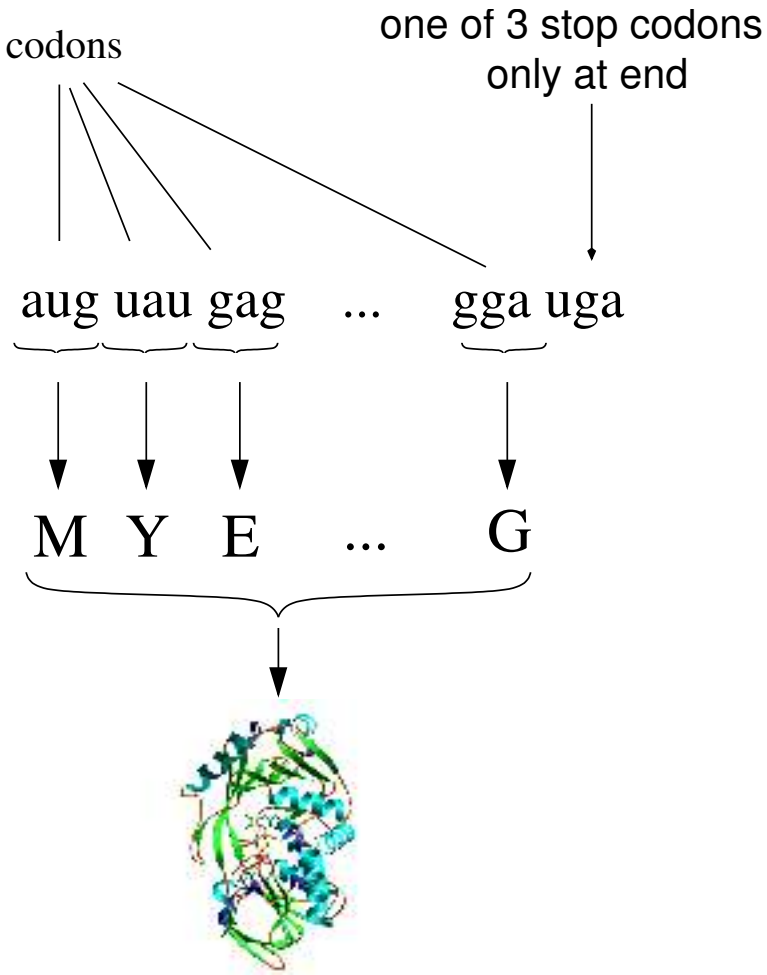
Structure of a eukaryotic gene





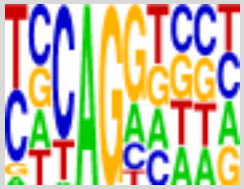
Translation





"universal"
genetic code

codon (DNA)		amino-acid
aaa	↦	K
aac	↦	N
aag	↦	K
aat	↦	N
.		.
.		.
.		.
atg	↦	M
.		.
.		.
.		.
61 codons		20 amino-acids



Translation

coding RNA sequence

translation

amino acid sequence

folding

protein

codons

one of 3 stop codons only at end

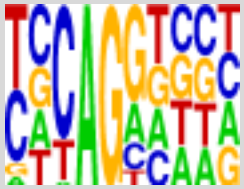
intron

aug uau gag ... gga uga

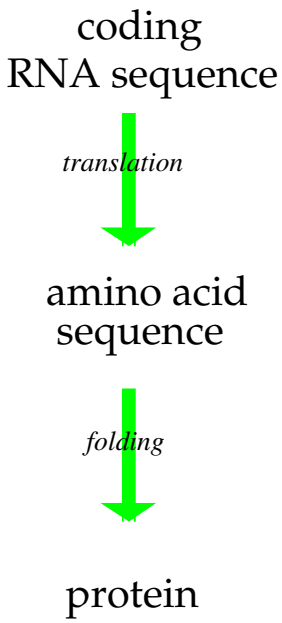
M Y E ... G

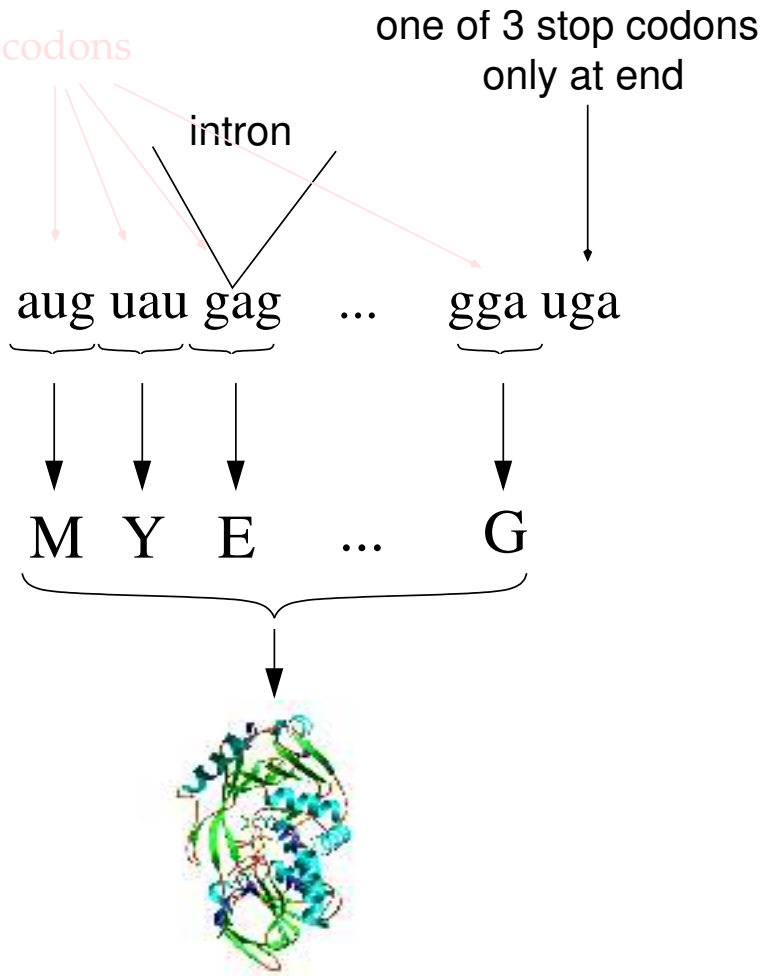
"universal" genetic code

codon (DNA)		amino-acid
aaa	↦	K
aac	↦	N
aag	↦	K
aat	↦	N
.		.
.		.
.		.
atg	↦	M
.		.
.		.
.		.
61 codons		20 amino-acids



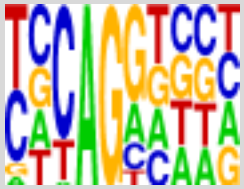
Translation



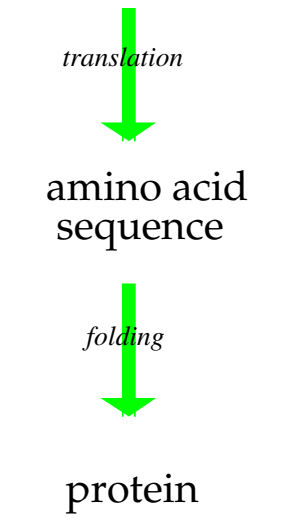


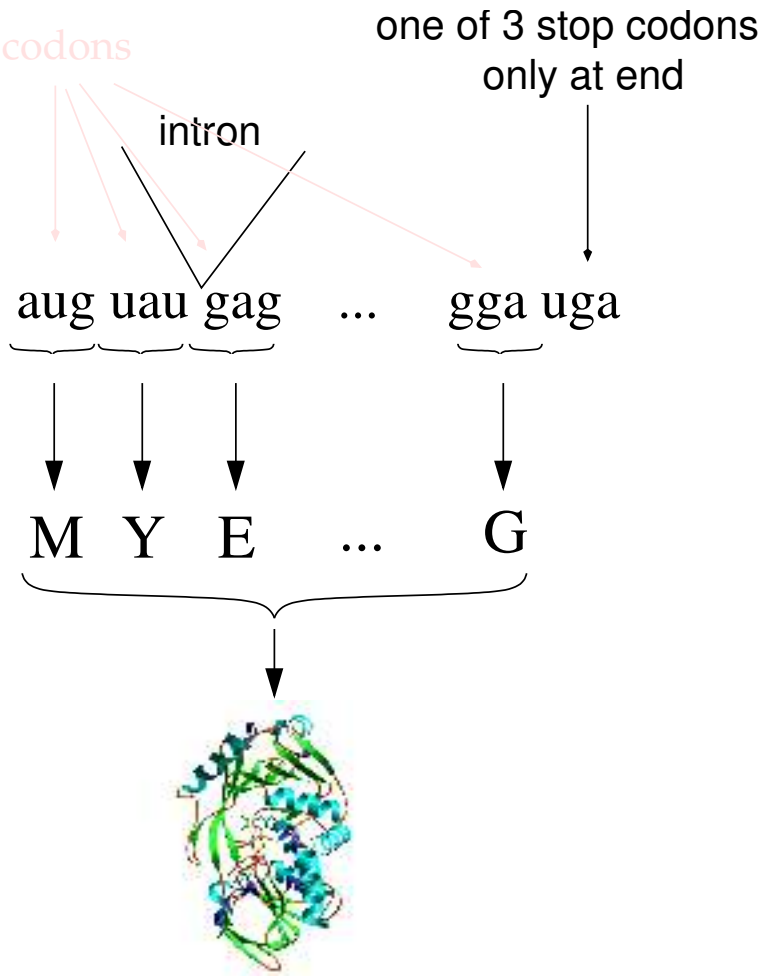
"universal" genetic code

codon (DNA)		amino-acid
aaa	↦	K
aac	↦	N
aag	↦	K
aat	↦	N
.		.
.		.
.		.
atg	↦	M
.		.
.		.
.		.
61 codons		20 amino-acids



Translation





"universal" genetic code

codon (DNA)		amino-acid
aaa	↦	K
aac	↦	N
aag	↦	K
aat	↦	N
.		.
.		.
.		.
atg	↦	M
.		.
.		.
.		.
61 codons		20 amino-acids

Gene Finding Problem

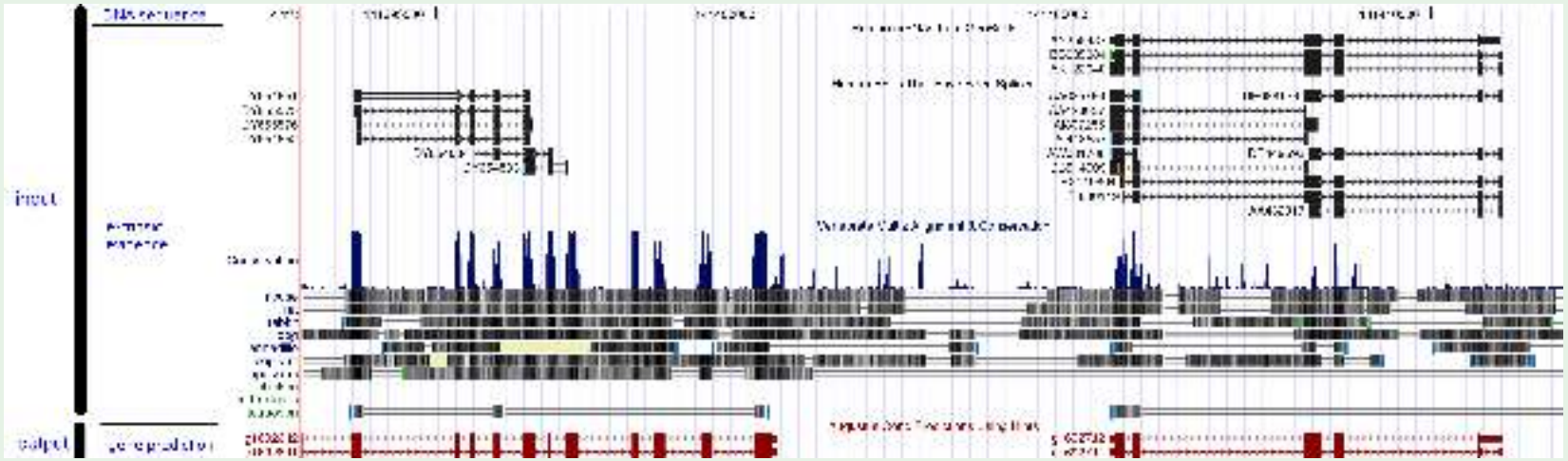
Input

- target DNA sequence, e.g. a chromosome
- optional: extrinsic evidence, e.g. from cDNA sequences or conservation

Output

- start- and end positions of genes, coding regions, exons and introns
- predicted amino acid sequences

Example (19000 bp of human chromosome 7)



Need for Gene Prediction

Sequencing of New Species

- Bejing Genome Institute (BGI):
  500 plants and 500 animals
- Genome 10K:
  10 000 species, one per genus of vertebrae

No Alternative to Gene Prediction Yet

- genomes are annotated *in silico*
- RNA-Seq does not appear to solve the problem

Major Approaches to Gene Prediction

Approaches

approach	extrinsic evidence used
ab initio	-

Major Approaches to Gene Prediction

Approaches

approach	extrinsic evidence used
ab initio	-
transcript based	transcript seqs (both native and alien)

Major Approaches to Gene Prediction

Approaches

approach	extrinsic evidence used
ab initio	-
transcript based	transcript seqs (both native and alien)
protein homology	protein sequences

Major Approaches to Gene Prediction

Approaches

approach	extrinsic evidence used
ab initio	-
transcript based	transcript seqs (both native and alien)
protein homology	protein sequences
comparative (*de novo*)	additional unannotated genomes

Major Approaches to Gene Prediction

Approaches

approach	extrinsic evidence used
ab initio	-
transcript based	transcript seqs (both native and alien)
protein homology	protein sequences
comparative (*de novo*)	additional unannotated genomes
proteogenomics	peptides from mass spectrometry

Major Approaches to Gene Prediction

Approaches

approach	extrinsic evidence used
ab initio	-
transcript based	transcript seqs (both native and alien)
protein homology	protein sequences
comparative (*de novo*)	additional unannotated genomes
proteogenomics	peptides from mass spectrometry
combiners	other gene predictions + transcript seqs + proteins + ?

Major Approaches to Gene Prediction

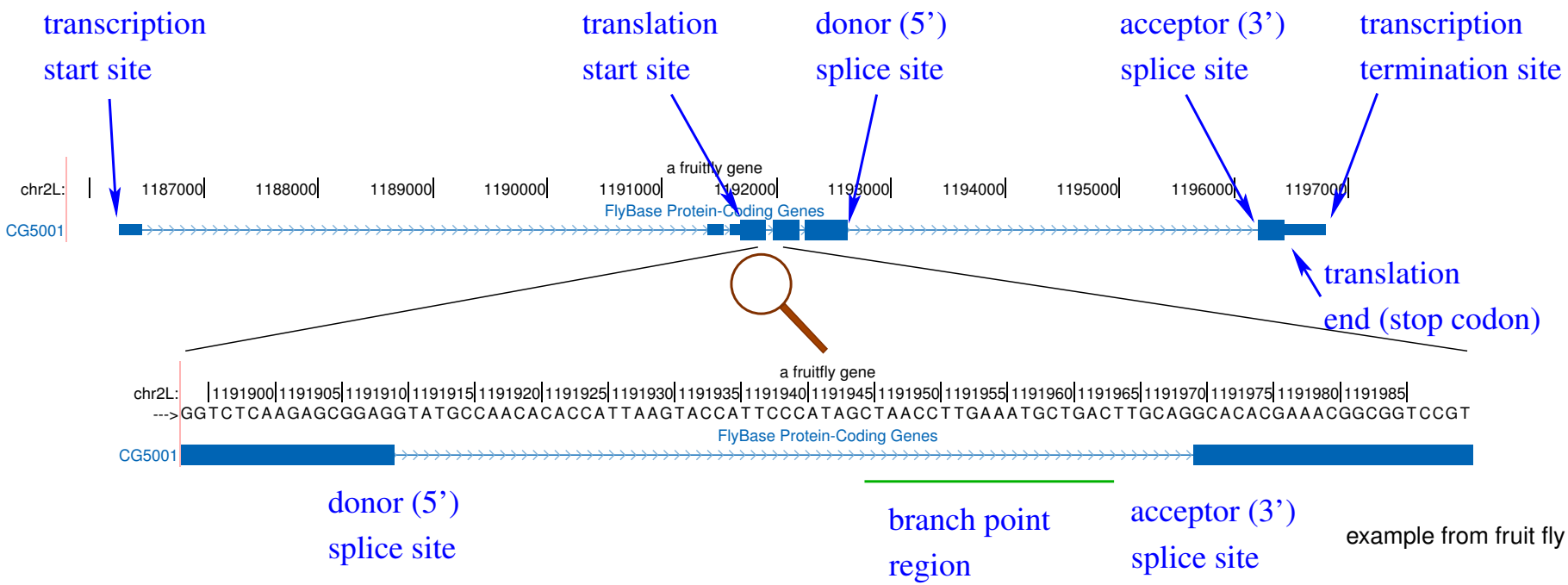
Approaches

approach	extrinsic evidence used
ab initio	-
transcript based	transcript seqs (both native and alien)
protein homology	protein sequences
comparative (*de novo*)	additional unannotated genomes
proteogenomics	peptides from mass spectrometry
combiners	other gene predictions + transcript seqs + proteins + ?

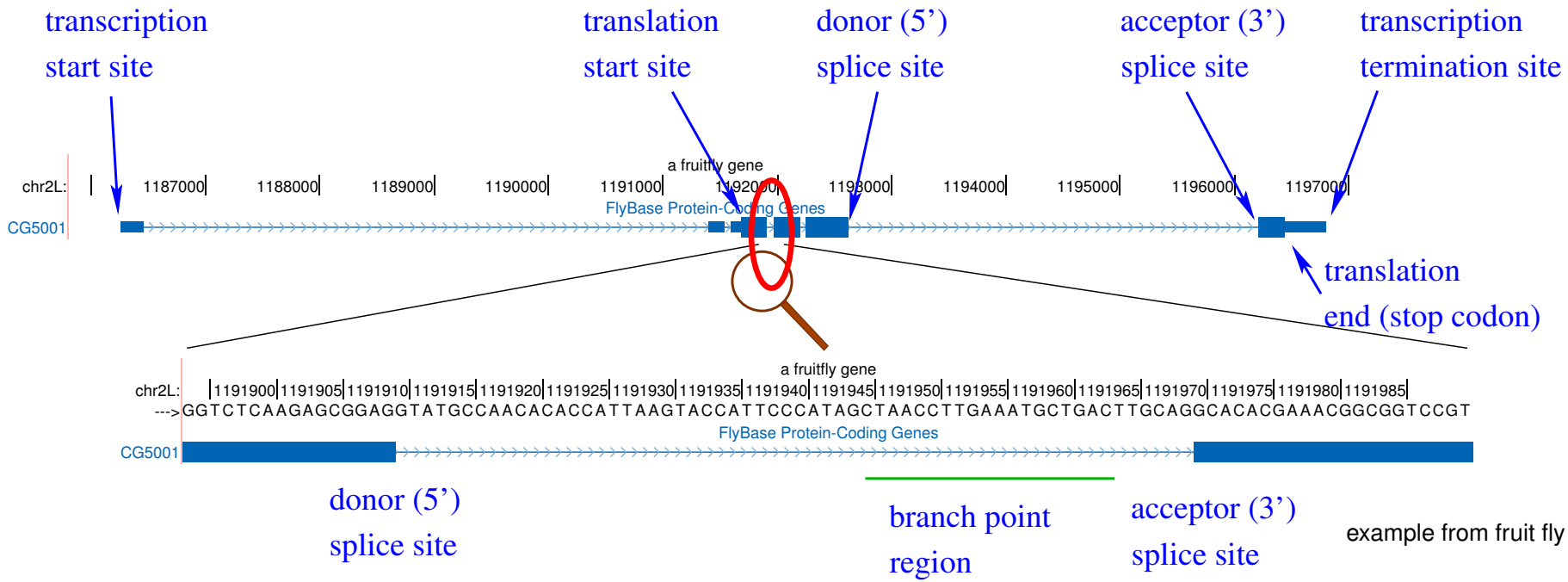
The annotation of most genomes requires a combination of approaches:

Use for every part of a gene all evidence available for that gene or region.

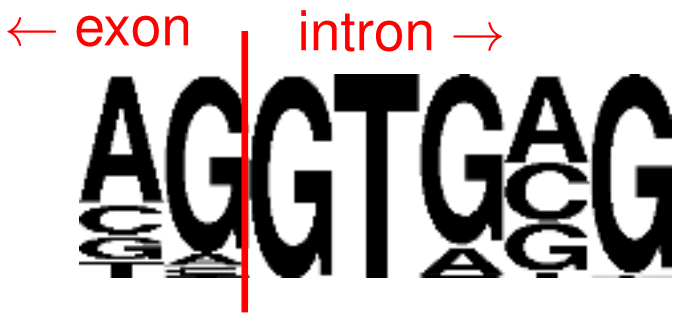
Signals


transcription start site
translation start site
donor (5') splice site
acceptor (3') splice site
transcription termination site
a fruitfly gene
chr2L: 1187000 1188000 1189000 1190000 1191000 1192000 1193000 1194000 1195000 1196000 1197000
FlyBase Protein-Coding Genes
CG5001
translation end (stop codon)
a fruitfly gene
chr2L: 1191900 1191905 1191910 1191915 1191920 1191925 1191930 1191935 1191940 1191945 1191950 1191955 1191960 1191965 1191970 1191975 1191980 1191985
---> GGTCTCAAGAGCGGAGGTATGCCAACACACCATTAAGTACCATTCCCATAGCTAACCTTGAAATGCTGACTTGCAGGCACACGAAACGGCGGTCCGT
FlyBase Protein-Coding Genes
CG5001
donor (5') splice site
branch point region
acceptor (3') splice site
example from fruit fly

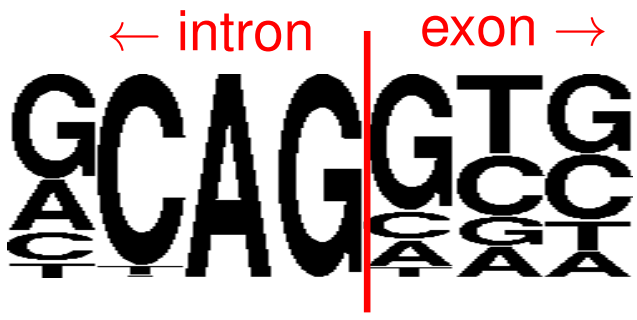
Signals



example from fruit fly



donor splice site (DSS) signal

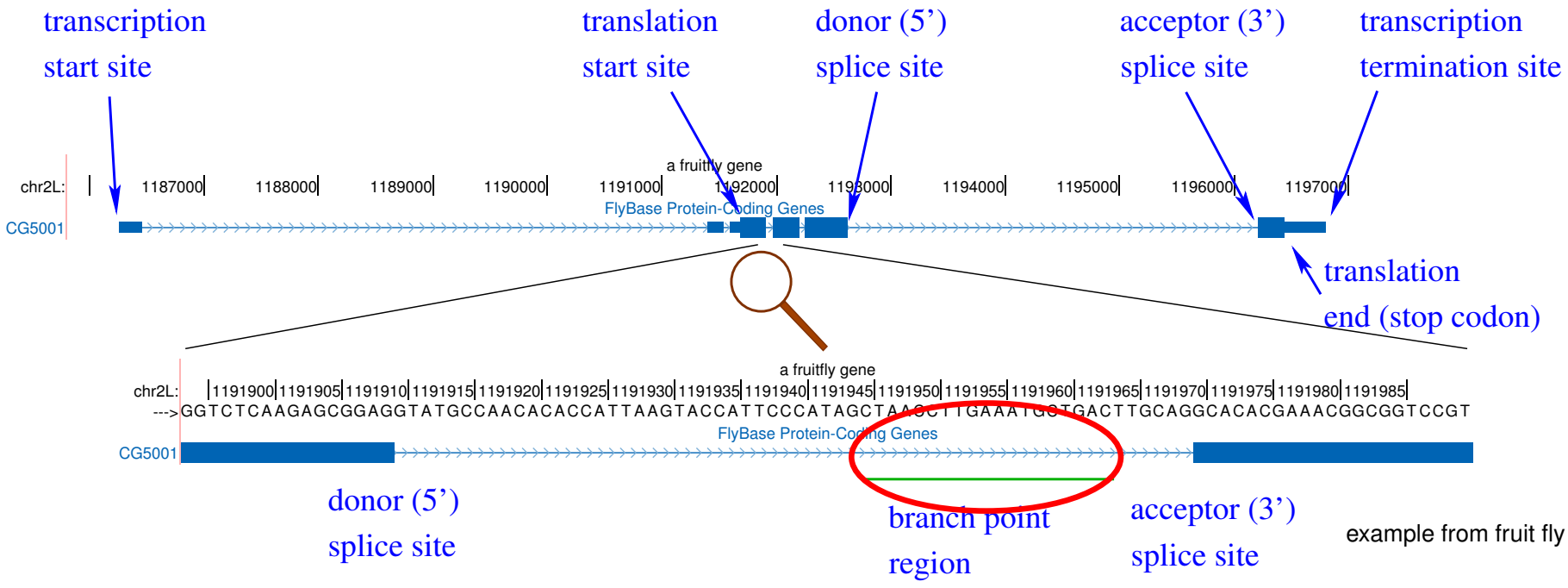


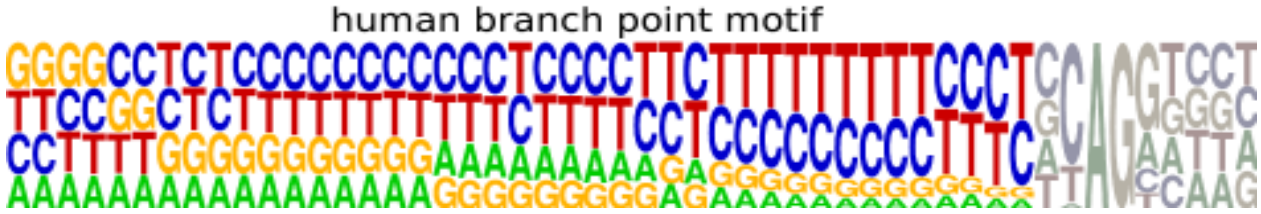
acceptor splice site (ASS) signal

Frequency of the nucleotides at positions relative to splice site.

from green algae *Chlamydomonas*

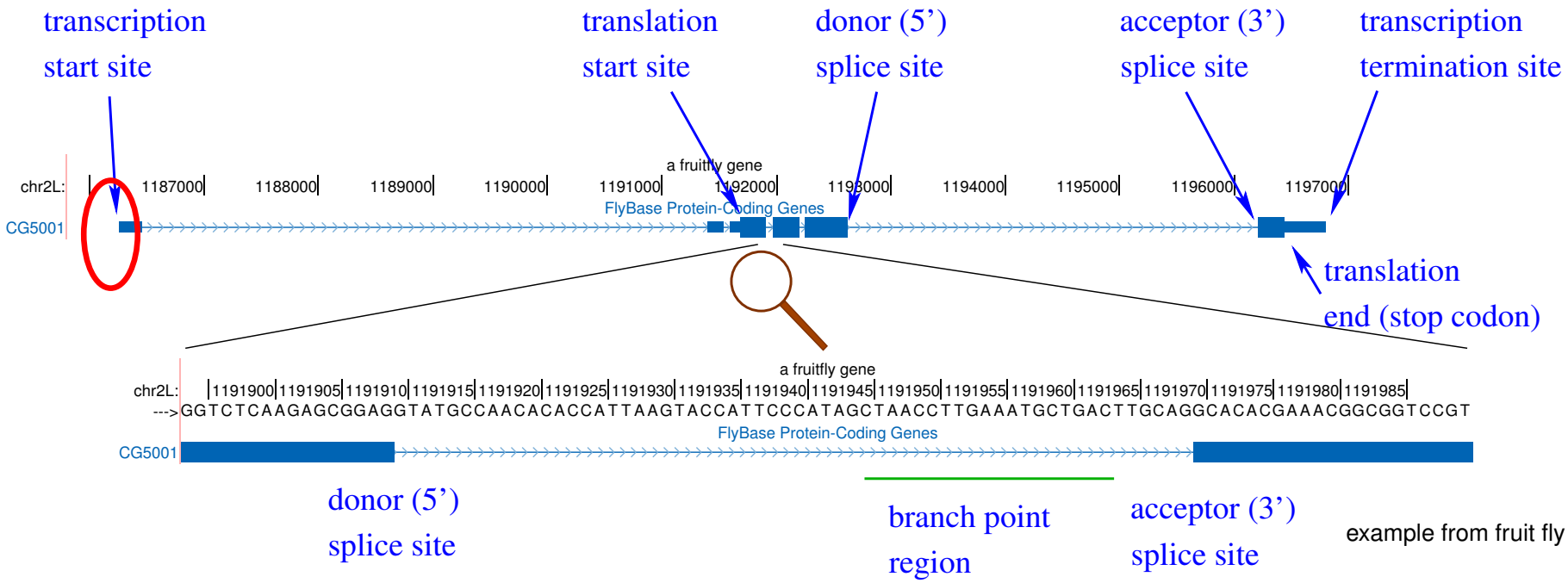
Signals





Branch point: upstream of 3' splice site, a single conserved adenine at variable distance to 3' splice site ($\approx$ -30), a splicing complex binds to it, pyrimidine (C,T) rich in human

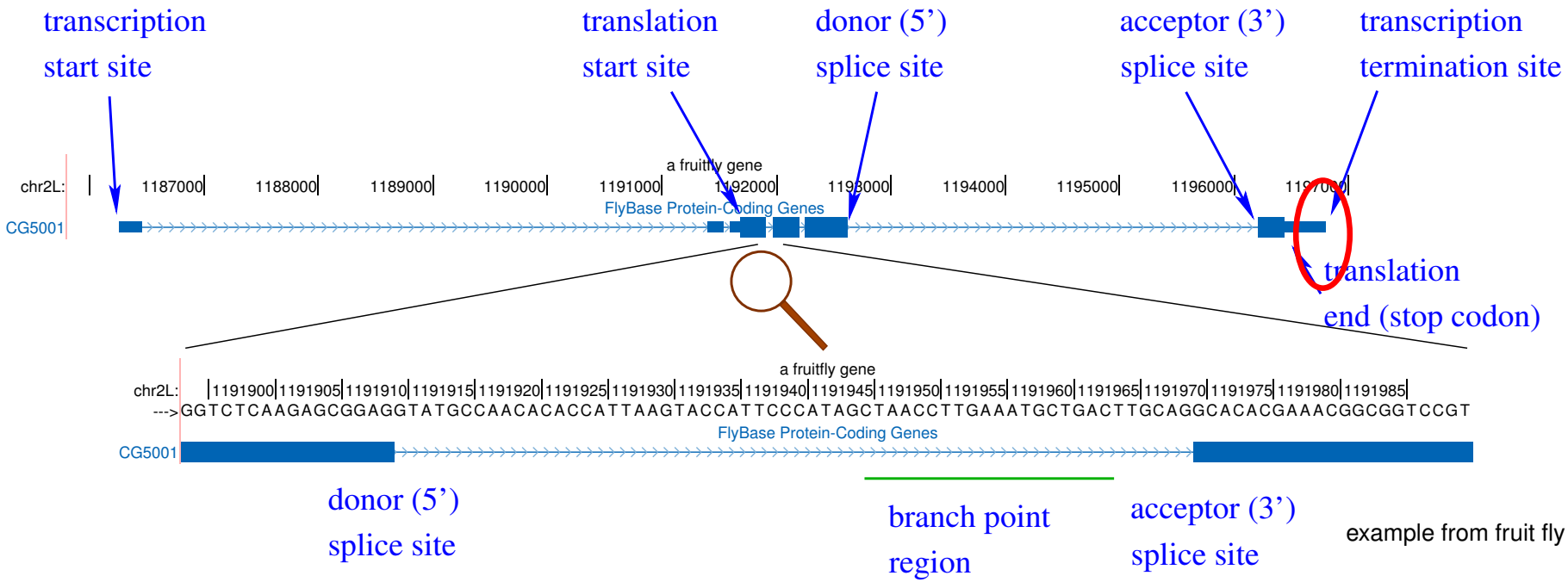
Signals



Transcription start site: Transcription from DNA to RNA by RNA polymerase starts here facilitated by promoter elements.
Promoter elements are diverse and their profiles tend to contain little info:

- diverse transcription factor binding sites at very variable positions
- sometimes TATA-box
- "CpG islands"

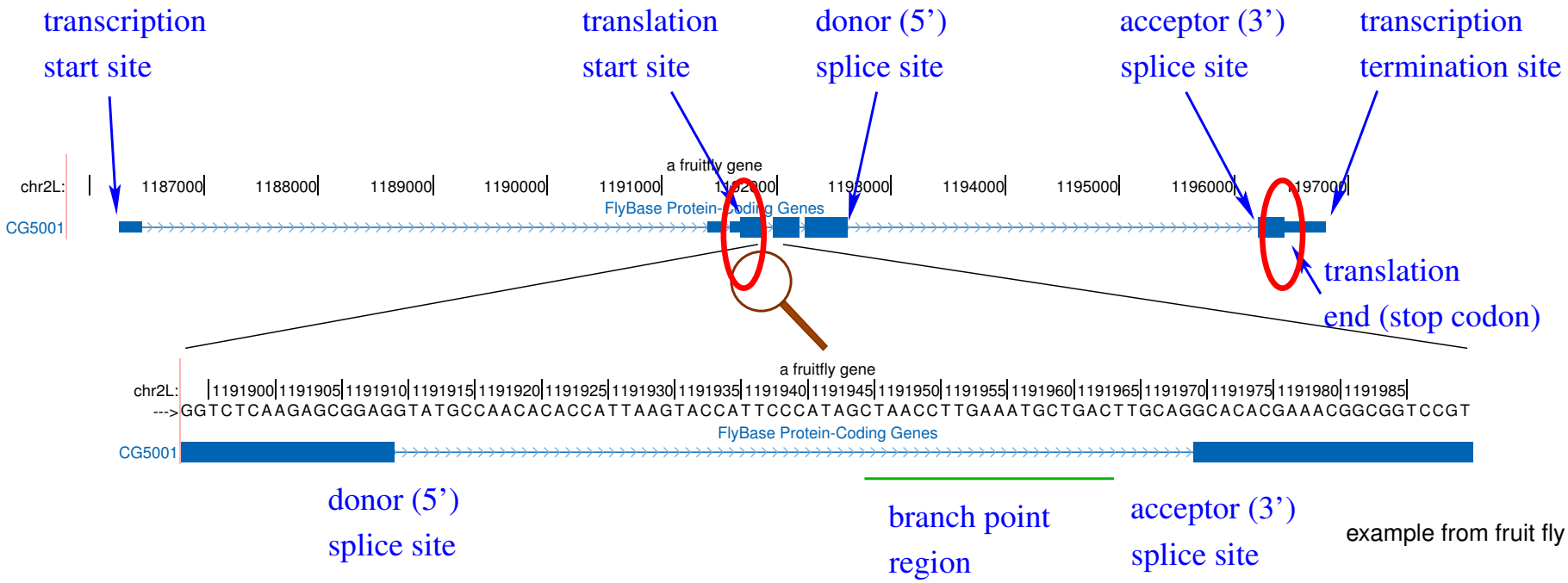
Signals



Transcription termination site (TTS):

- cleavage of the transcript.
- some non-templated A's are appended (polyadenylation).
- polyadenylation is triggered in many species in many genes by the hexamer aataaa roughly 15 bp upstream of the TTS.

Signals



Start and stop codon:

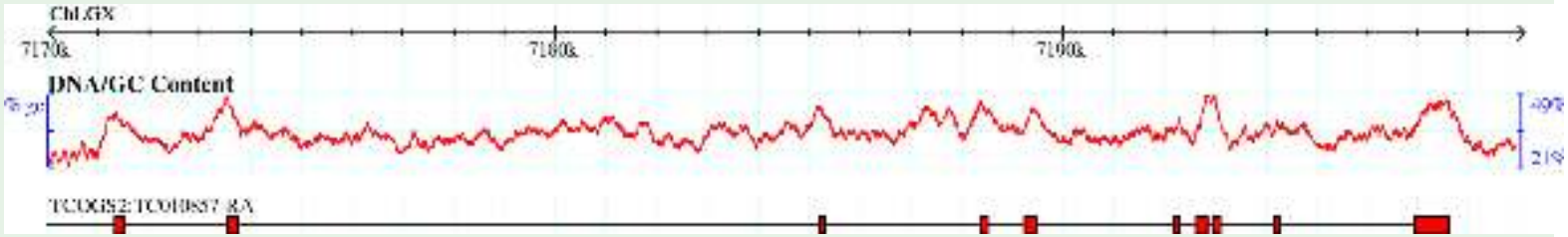
- start codon: ATG
- stop codons: TAA, TAG, TGA
  In some species the genetic code is altered and a "stop codon" is actually coding for an amino acid.

Nucleotide Composition of Coding and Noncoding Regions

Sequence Content

Besides the signals, position-unspecific frequencies of nucleotide patterns can be used to guess biological classification (e.g. CDS, non-coding, CpG-island) of longer sequence intervals.

Example (GC content in red flour beetle)



Typically, higher order patterns are examined:
E.g. reading-frame dependent k-mer frequencies ($k = 5, 6$) for protein-coding regions.

Remark

Sequence content is usually only indirect evidence.

Signal and Content Models

Scores

- usually, we cannot decide with certainty whether something is an exon, splice site, etc
- $\Rightarrow$ therefore assign a number (score s or probability) to exon candidates, splice site candidates
- aim:
  - the larger the score the more likely there is a true signal
  - the score is small for false signals

Signal and Content Models

Scores

- usually, we cannot decide with certainty whether something is an exon, splice site, etc
- $\Rightarrow$ therefore assign a number (score s or probability) to exon candidates, splice site candidates
- aim:
  - the larger the score the more likely there is a true signal
  - the score is small for false signals

Signal and Content Models

- Signal Model: to score whether a biological signal (e.g. splice site) is present at a certain position
- Content Model: to score whether a sequence region belongs to a given class (e.g. coding)

Example Signal Score

Example (DSS position weight matrix)

Have position specific frequency matrix for DSS

$$m(i, b) \quad (i = 1, 2, \ldots, 7, b \in \text{A,C,G,T}),$$

$$m(i, \text{A}) + m(i, \text{C}) + m(i, \text{G}) + m(i, \text{T}) = 1$$

Have "background" distribution of nucleotides $q(b)$

$$q(\text{A}) + q(\text{C}) + q(\text{G}) + q(\text{T}) = 1.$$

Define log-odds score: $s = \log \prod_{i=1}^{7} m(i, w_i)/q(w_i)$.

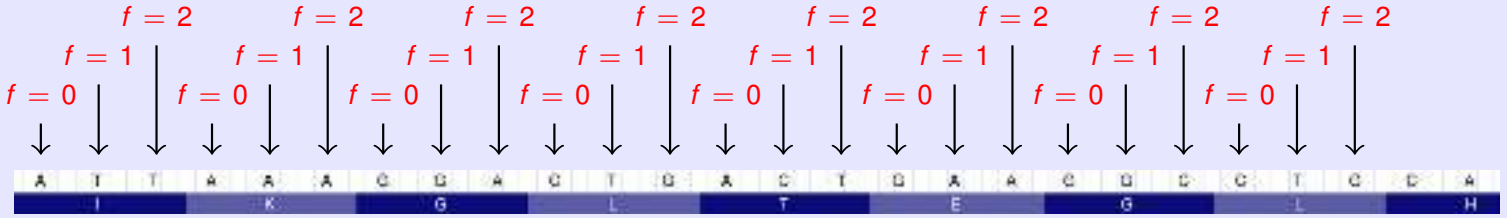
Decision rule

$$\text{predict splice site} :\Leftrightarrow s \geq \text{threshold}$$

guaranteed to have most true predictions given a limit on false positive rate under certain assumptions (Neyman-Pearson).

Example Content Score

Base composition is frame-dependent



nucleotide frequencies in human:

	coding sequence				noncoding
	$f=0$	$f=1$	$f=2$	all f	sequence
A	0.248	0.291	0.146	0.229	0.26
C	0.264	0.243	0.351	0.286	0.24
G	0.321	0.201	0.312	0.278	0.24
T	0.166	0.265	0.190	0.207	0.26

Example Content Score

Example (frame-dependent Markov chain of order k)

Let w be the DNA word of length n to be scored as CDS.
Let $f \in \{0,1,2\}$ be the frame of the first position of w.

$$P(w) := p_f(w_1,\ldots,w_k) \cdot \prod_{i=k+1}^{n} p_{f(i)}(w_i \mid w_{i-k},\ldots,w_{i-1})$$

Here:
- p_f is a start probability for the first k bases
- $f(i) \in \{0,1,2\}$ such that $f(i) \equiv f - 1 + i \mod 3$ is the frame of the i-th position of w

Define $s(w) = \log(P(w)/Q(w))$,
where $Q(w)$ is the probability of w in a "background" model (e.g. non-coding).

Remark:
Division by background $\Rightarrow$ good exon candidates get positive score

Example Content Score - Continued

Example

$w = \text{ATTCTGC}$
frame $f = 2$, i.e. with these codon breaks: A||TTC||TGC
$k = 2$

$$\begin{aligned} P(\text{ATTCTGC}) &= p_2(\text{AT})p_1(\text{T}|\text{AT})p_2(\text{C}|\text{TT}) \\ &\quad p_0(\text{T}|\text{TC})p_1(\text{G}|\text{CT})p_2(\text{C}|\text{TG}) \end{aligned}$$

Choice of order k

- probabilities $p_r(x \mid y_1, \ldots, y_k)$ can be estimated on known coding sequences
- if $k \geq 2$ above content model can reflect codon usage
- usually: the higher k the bigger the difference between coding and noncoding

Example Content Score - Continued

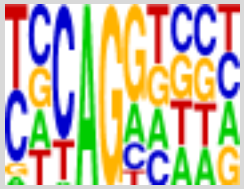
Example

$w = \text{ATTCTGC}$
frame $f = 2$, i.e. with these codon breaks: A||TTC||TGC
$k = 2$

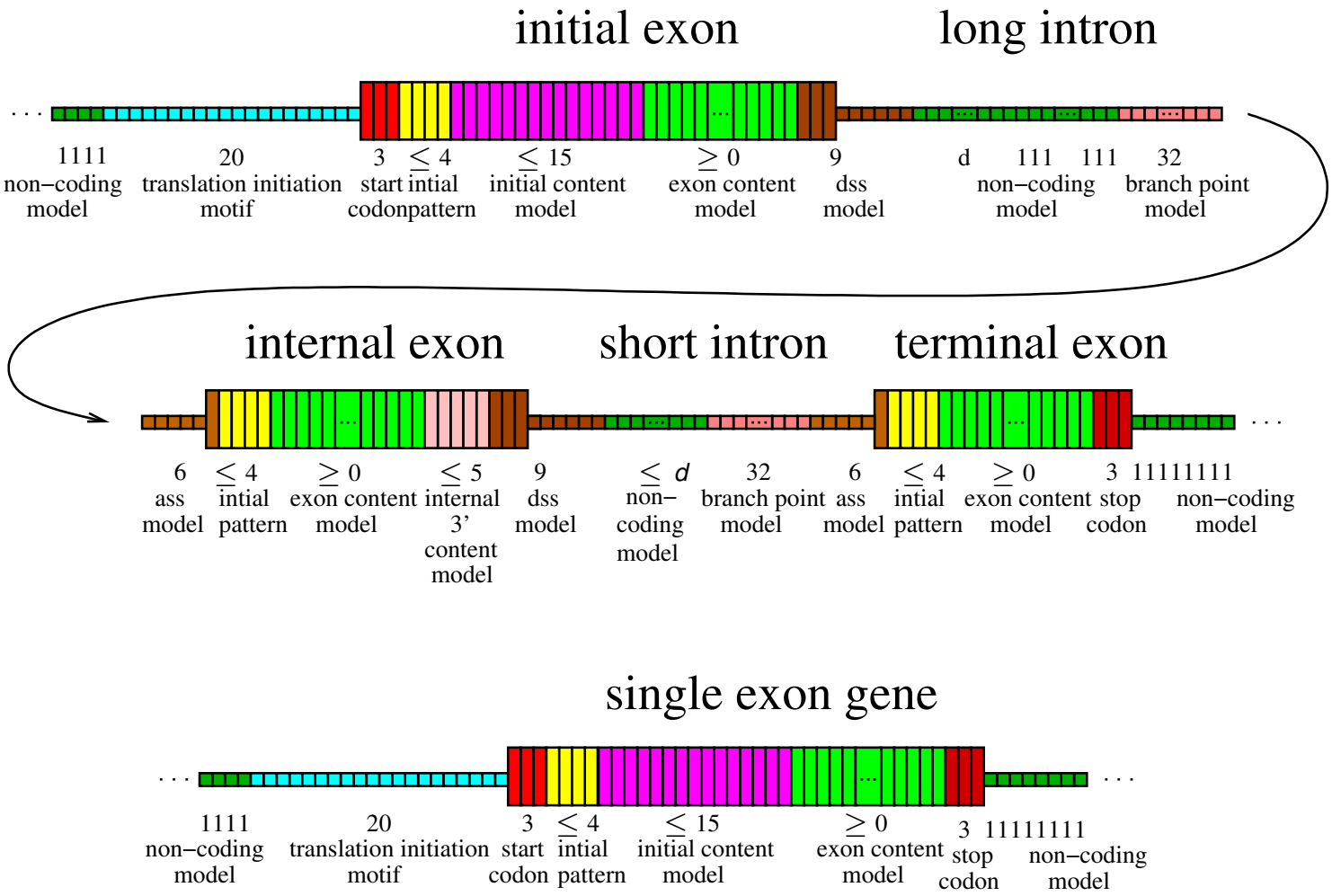
$$\begin{aligned} P(\text{ATTCTGC}) &= p_2(\text{AT})p_1(\text{T}|\text{AT})p_2(\text{C}|\text{TT}) \\ &\quad p_0(\text{T}|\text{TC})p_1(\text{G}|\text{CT})p_2(\text{C}|\text{TG}) \end{aligned}$$

Choice of order k

- probabilities $p_r(x \mid y_1, \ldots, y_k)$ can be estimated on known coding sequences
- if $k \geq 2$ above content model can reflect codon usage
- usually: the higher k the bigger the difference between coding and noncoding
- but: content model $\neq$ reality (dead genes)
- typical: $k = 4$ or $k = 5$



Signal and Content Models of AUGUSTUS (without UTR)



What (Most) Eukaryotic Species Have in Common

In Common:

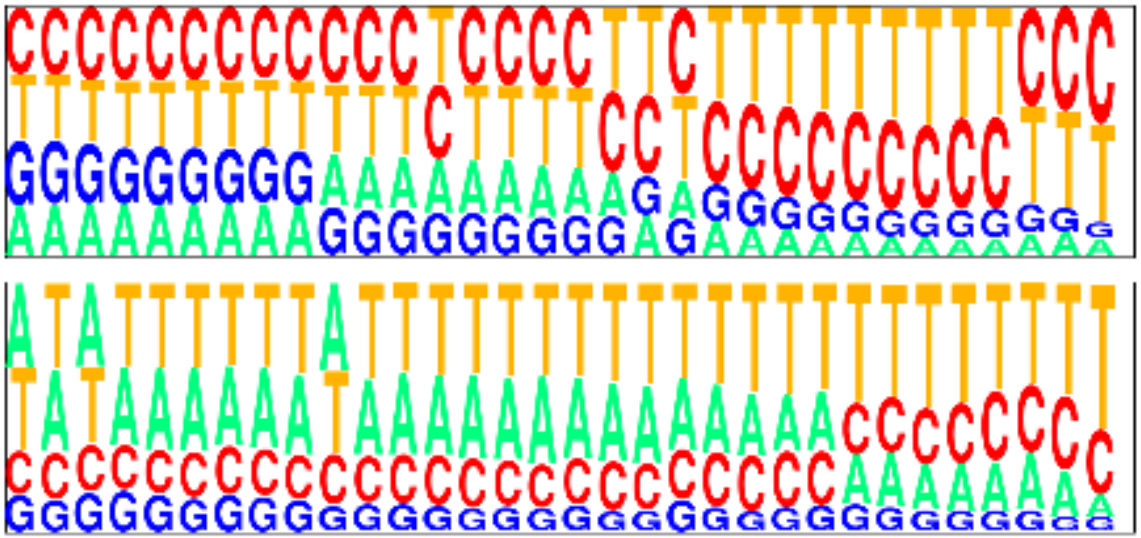
- same genetic code, including start and stop codons
- genes can have introns, may have many
- genes rarely overlap in sequence
- introns start almost always with GT, end with AG (some introns GC/AG)
- more non-coding sequence than coding sequence

How Species-Specific Must Gene Finding Models Be?

Differences:

- distribution at signals, e.g. branch point region

top: human / bottom: fly



How Species-Specific Must Gene Finding Models Be?

Differences:

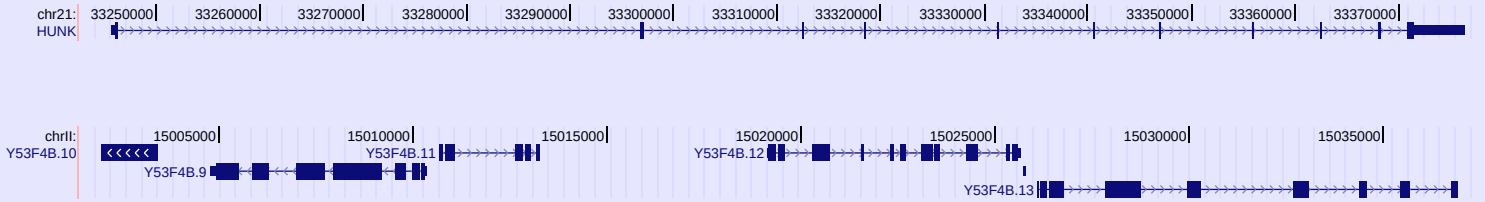
- distribution at signals, e.g. branch point region
- GC content highly variable

How Species-Specific Must Gene Finding Models Be?

Differences:

- distribution at signals, e.g. branch point region
- GC content highly variable
- number and length distribution of introns

top: human / bottom: *C. elegans*



How Species-Specific Must Gene Finding Models Be?

Differences:

- distribution at signals, e.g. branch point region
- GC content highly variable
- number and length distribution of introns
- length distribution of UTRs

How Species-Specific Must Gene Finding Models Be?

Differences:

- distribution at signals, e.g. branch point region
- GC content highly variable
- number and length distribution of introns
- length distribution of UTRs
- gene density

Training: Estimate Species-Specific Parameters

"Training Set"

- input: set of annotated sequences

$$(x^{(k)}, y^{(k)})_{k=1,\ldots,N},$$

  such that the parse $x^{(k)}$ represents the gene structure of DNA sequence $y^{(k)}$.
- frequently a few hundred genes constructed from transcript sequences
- used to estimate species-specific parameters
- self-training programs: use their own preliminary prediction as training set and iterate prediction and training (GeneMark-ES, Borodovsky et al.)
  (does not work in all species)

Training: Estimate Species-Specific Parameters

"Training Set"

- input: set of annotated sequences

$$(x^{(k)}, y^{(k)})_{k=1,\ldots,N},$$

  such that the parse $x^{(k)}$ represents the gene structure of DNA sequence $y^{(k)}$.
- frequently a few hundred genes constructed from transcript sequences
- used to estimate species-specific parameters
- self-training programs: use their own preliminary prediction as training set and iterate prediction and training (GeneMark-ES, Borodovsky et al.)
  (does not work in all species)

Training: Estimate Species-Specific Parameters

"Training Set"

- input: set of annotated sequences

$$(x^{(k)}, y^{(k)})_{k=1,\ldots,N},$$

  such that the parse $x^{(k)}$ represents the gene structure of DNA sequence $y^{(k)}$.
- frequently a few hundred genes constructed from transcript sequences
- used to estimate species-specific parameters
- self-training programs: use their own preliminary prediction as training set and iterate prediction and training (GeneMark-ES, Borodovsky et al.)
  (does not work in all species)

Training: Estimate Species-Specific Parameters

"Training Set"

- input: set of annotated sequences

$$(x^{(k)}, y^{(k)})_{k=1,\ldots,N},$$

  such that the parse $x^{(k)}$ represents the gene structure of DNA sequence $y^{(k)}$.
- frequently a few hundred genes constructed from transcript sequences
- used to estimate species-specific parameters
- self-training programs: use their own preliminary prediction as training set and iterate prediction and training (GeneMark-ES, Borodovsky et al.)
  (does not work in all species)

Problems and General Ansatz

Problems

- known signal models do not carry much information
- false positive signals because of low number of true positives
- sequence content can be misleading (pseudogenes, repeats)

Ansatz

- combine all individual weak info to boost discriminatory power
- enforce standard gene structure:
  - reading frame consistency between exons
  - minimal splice site consensus (GT/AG, maybe GC/AG)
  - no in-frame stop codons
  - minimal intron length ($\approx$ 40 bp)

Problems and General Ansatz

Problems

- known signal models do not carry much information
- false positive signals because of low number of true positives
- sequence content can be misleading (pseudogenes, repeats)

Ansatz

- combine all individual weak info to boost discriminatory power
- enforce standard gene structure:
  - reading frame consistency between exons
  - minimal splice site consensus (GT/AG, maybe GC/AG)
  - no in-frame stop codons
  - minimal intron length ($\approx$ 40 bp)

Problems and General Ansatz

Problems

- known signal models do not carry much information
- false positive signals because of low number of true positives
- sequence content can be misleading (pseudogenes, repeats)

Ansatz

- combine all individual weak info to boost discriminatory power
- enforce standard gene structure:
  - reading frame consistency between exons
  - minimal splice site consensus (GT/AG, maybe GC/AG)
  - no in-frame stop codons
  - minimal intron length ($\approx$ 40 bp)

Problems and General Ansatz

Problems

- known signal models do not carry much information
- false positive signals because of low number of true positives
- sequence content can be misleading (pseudogenes, repeats)

Ansatz

- combine all individual weak info to boost discriminatory power
- enforce standard gene structure:
  - reading frame consistency between exons
  - minimal splice site consensus (GT/AG, maybe GC/AG)
  - no in-frame stop codons
  - minimal intron length ($\approx$ 40 bp)

Problems and General Ansatz

Problems

- known signal models do not carry much information
- false positive signals because of low number of true positives
- sequence content can be misleading (pseudogenes, repeats)

Ansatz

- combine all individual weak info to boost discriminatory power
- enforce standard gene structure:
  - reading frame consistency between exons
  - minimal splice site consensus (GT/AG, maybe GC/AG)
  - no in-frame stop codons
  - minimal intron length ($\approx$ 40 bp)

Gene Structure Formally

observed variables:

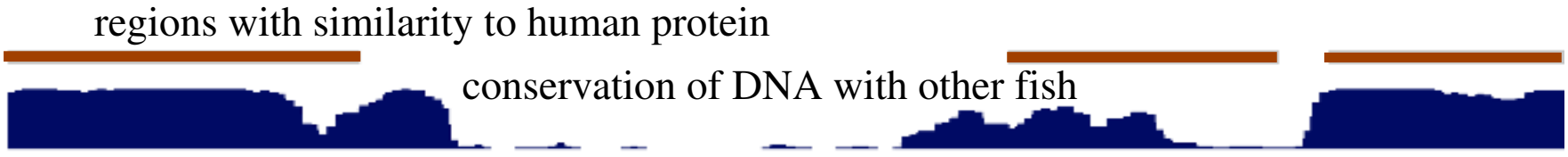
DNA
S

region of Fugu DNA

hints
H

regions with similarity to human protein

conservation of DNA with other fish

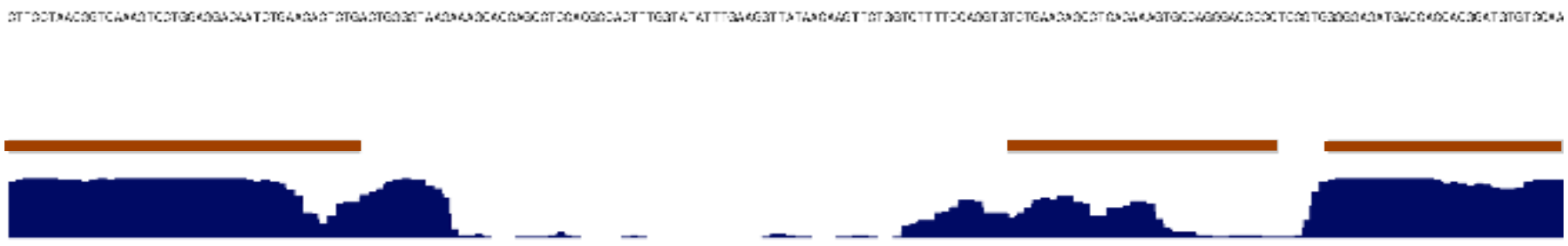


Gene Structure Formally

unobserved variables:

DNA
S

hints
H



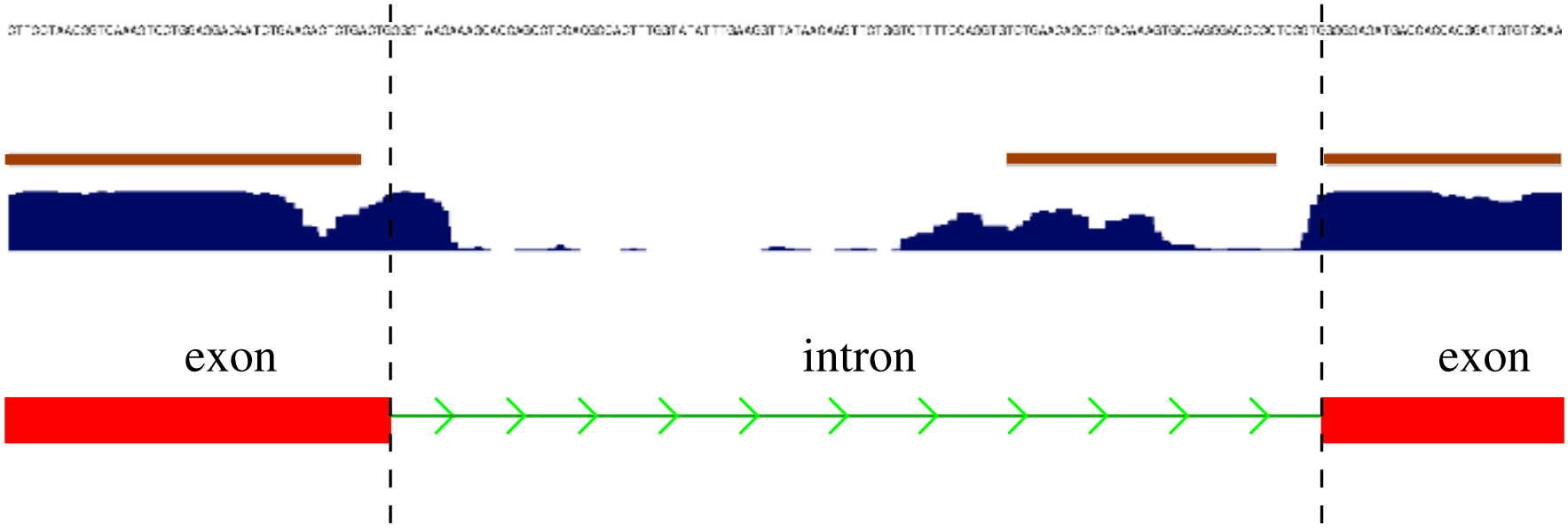
exon

Gene Structure Formally

unobserved variables:

DNA
S

hints
H

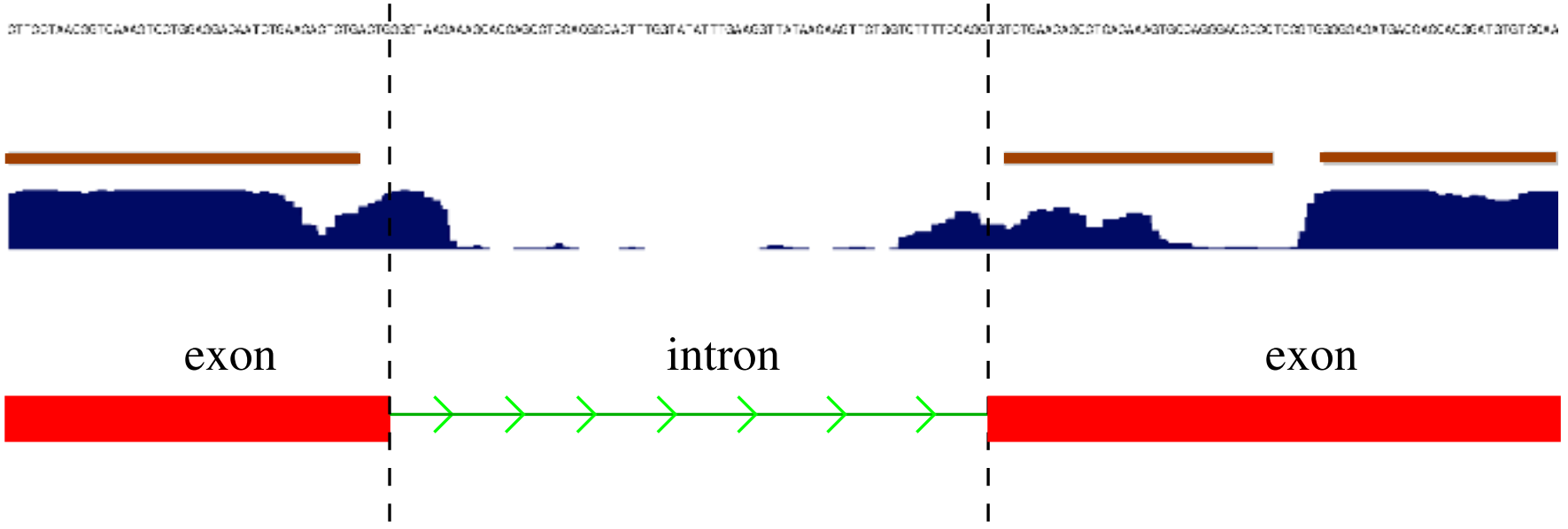


Gene Structure Formally

unobserved variables:

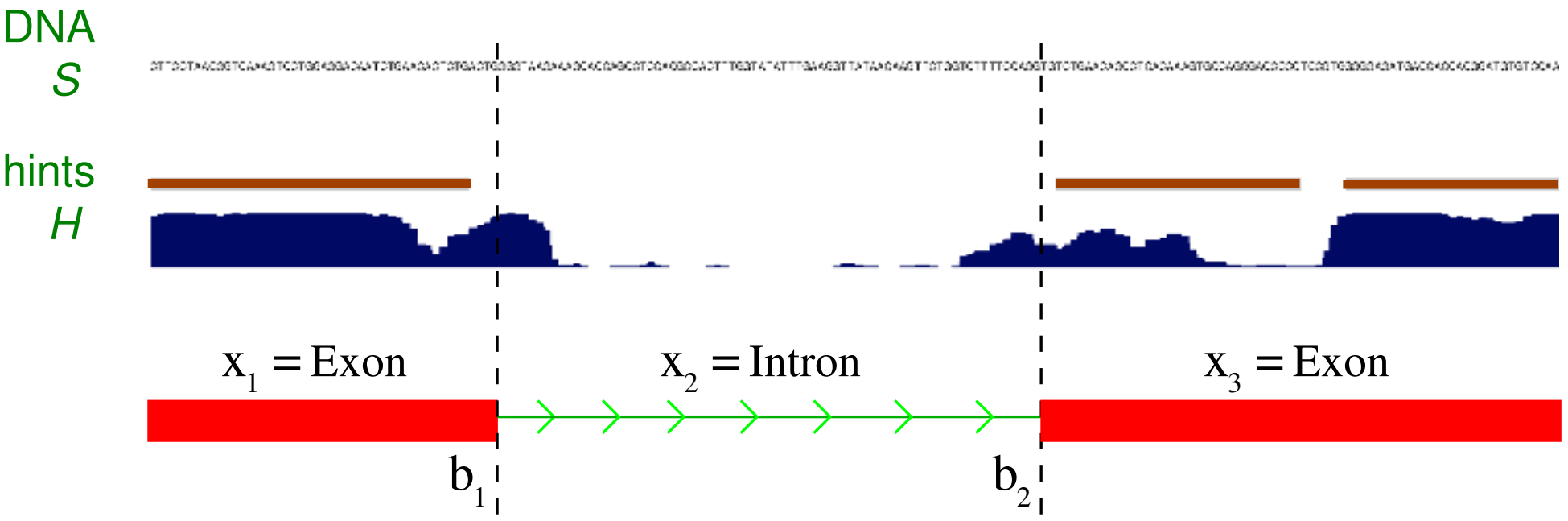
DNA
S

hints
H



Gene Structure Formally

unobserved variables:



gene structure $x = ((x_1, b_1), (x_2, b_2), \ldots, (x_n, b_n))$ is a labelled segmentation of DNA

$x_1, x_2, \ldots$ biological labels, e.g. "exon", "intron", "intergenic region"
$b_1, b_2, \ldots$ segment boundaries

Generic Approach

Input

sequence: DNA sequence to find genes in
evidence: optinal additional evidence, e.g. conservation

Output

gene structure x: start end end positions of exons, genes

Gene finding as optimization problem

maximize

$$s(\text{gene structure } x, \text{sequence } S, \text{evidence } H)$$

subject to constaint

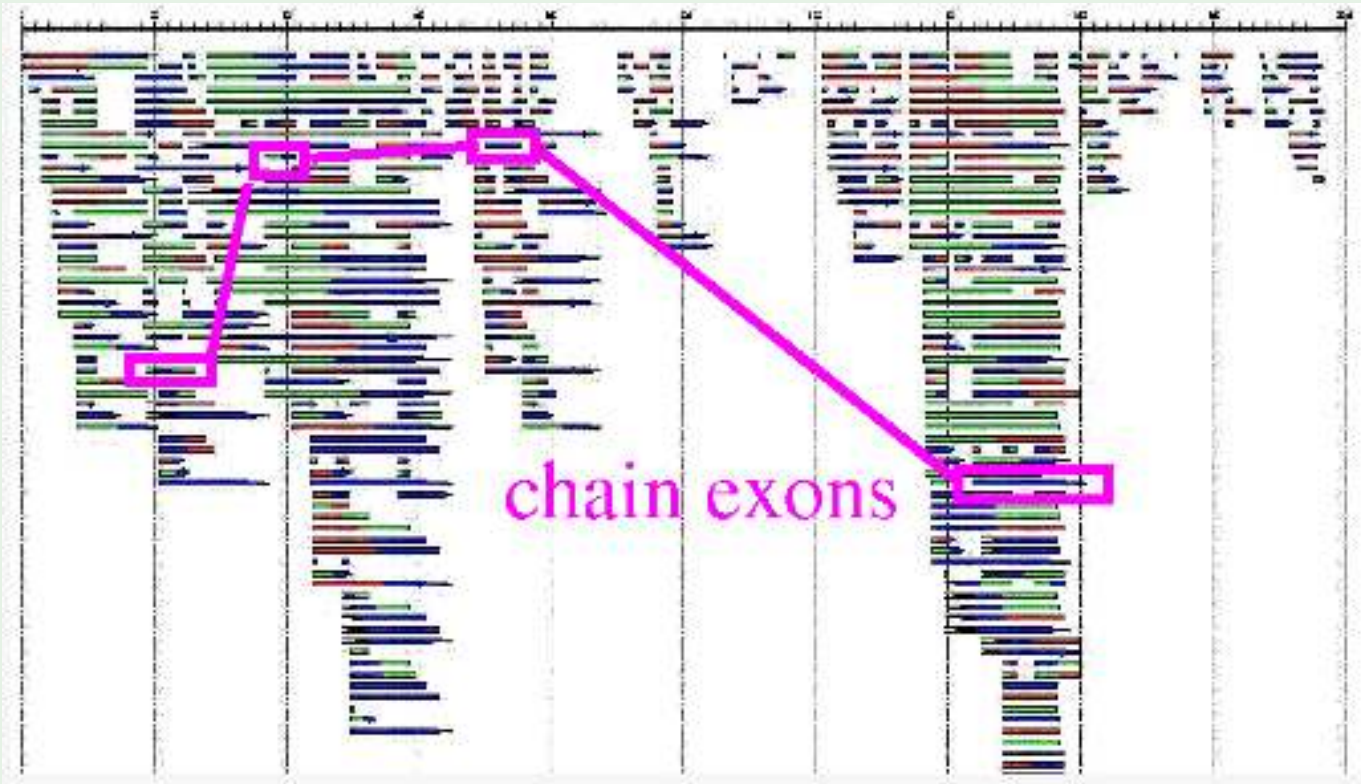
gene structure x makes biologically sense

One Step Closer to Concreteness

Exon Chaining

- find nonoverlapping compatible exon candidates with highest sum of scores
- e.g. done by GENEID (Parra et al., 2000)

Example (exon candidates in a DNA of length 2000)



http://monstrel.imim.es/courses

exon candidates of the program GENEID

One Step Closer to Concreteness

Exon Chaining

- find nonoverlapping compatible exon candidates with highest sum of scores
- e.g. done by GENEID (Parra et al., 2000)

Hidden Markov Model (HMM)

GENSCAN (Burge and Karlin, 1996),
GENIE (Kulp et al., 1996),
GENEMARK (Lukashin and Borodovsky, 1998),
FGENESH (Salamov and Solovyev, 2000),
AUGUSTUS (Stanke and Waack, 2003)

Conditional Random Field (CRF)

CRAIG (Bernal et al., 2007),
CONTRAST (Gross et al., 2007),
AUGUSTUS (unpublished)

One Step Closer to Concreteness

Exon Chaining

- find nonoverlapping compatible exon candidates with highest sum of scores
- e.g. done by GENEID (Parra et al., 2000)

Hidden Markov Model (HMM)

GENSCAN (Burge and Karlin, 1996),
GENIE (Kulp et al., 1996),
GENEMARK (Lukashin and Borodovsky, 1998),
FGENESH (Salamov and Solovyev, 2000),
AUGUSTUS (Stanke and Waack, 2003)

Conditional Random Field (CRF)

CRAIG (Bernal et al., 2007),
CONTRAST (Gross et al., 2007),
AUGUSTUS (unpublished)

Conditional Random Field for Gene Prediction

Definition (CRF)

A (Semi-)Conditional-Random-Field is a conditional distribution of gene structures x given the observation y of the following form

$$P(x \mid y) = \frac{1}{Z(y)} \exp \sum_{j=1}^{n} f(x_{i-1}, x_i, b_{i-1}, b_i, y).$$

CRF

- gene structure is scored locally:
  score may only locally depend on one labelled segment
- not scorable, e.g.: gene should have a total exon length of 1000bp

Conditional Random Field versus Hidden Markov Model

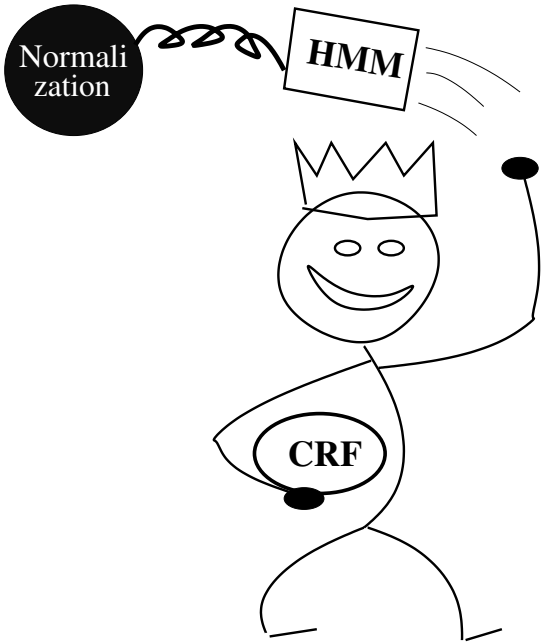
HMM

$$P(x, y) = \prod_{i=1}^{n} P(x_i|x_{i-1}) \cdot P(y(b_{i-1}, b_i] \mid x_i)$$

is a generative model: everything that is used in prediction (y), must have a distribution.
E.g. in comparative gene finding: Must define probability of any observable alignment column.

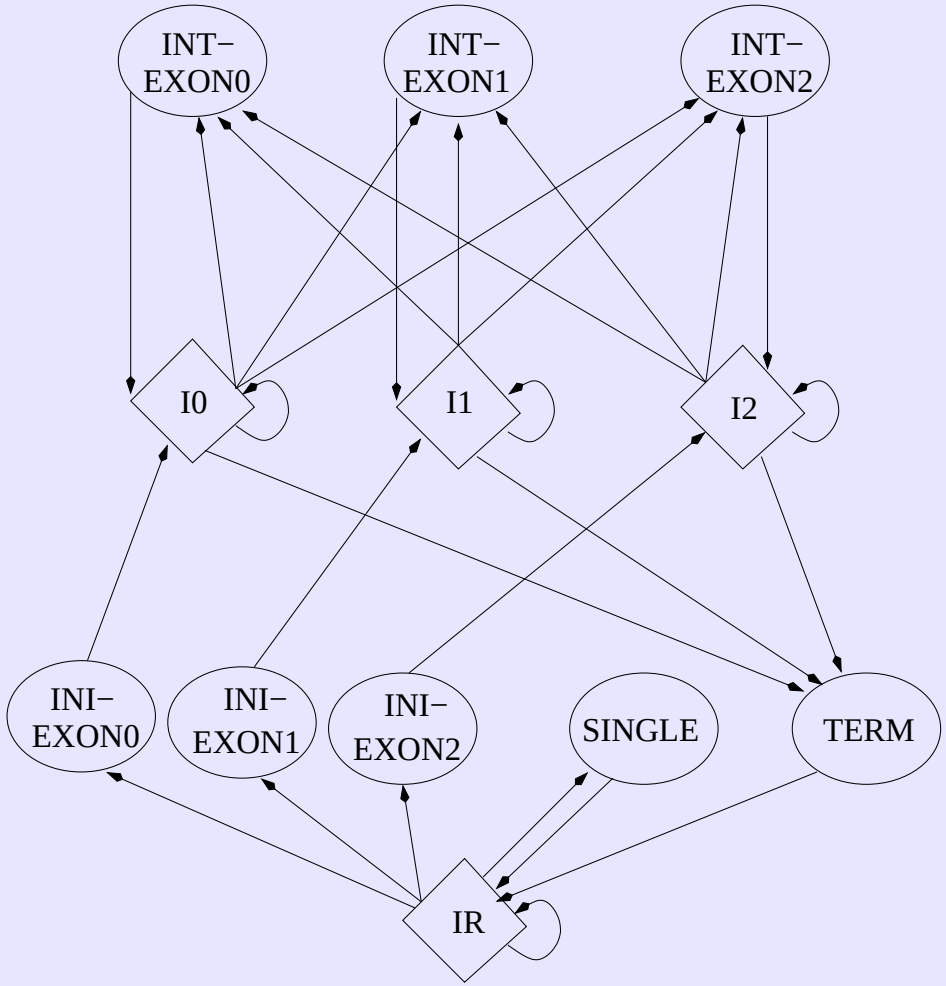
CRF

- for every HMM there is a CRF with same $P(x|y)$
- only *globally* normalizaion necessary
- can easier be dicriminatively trained

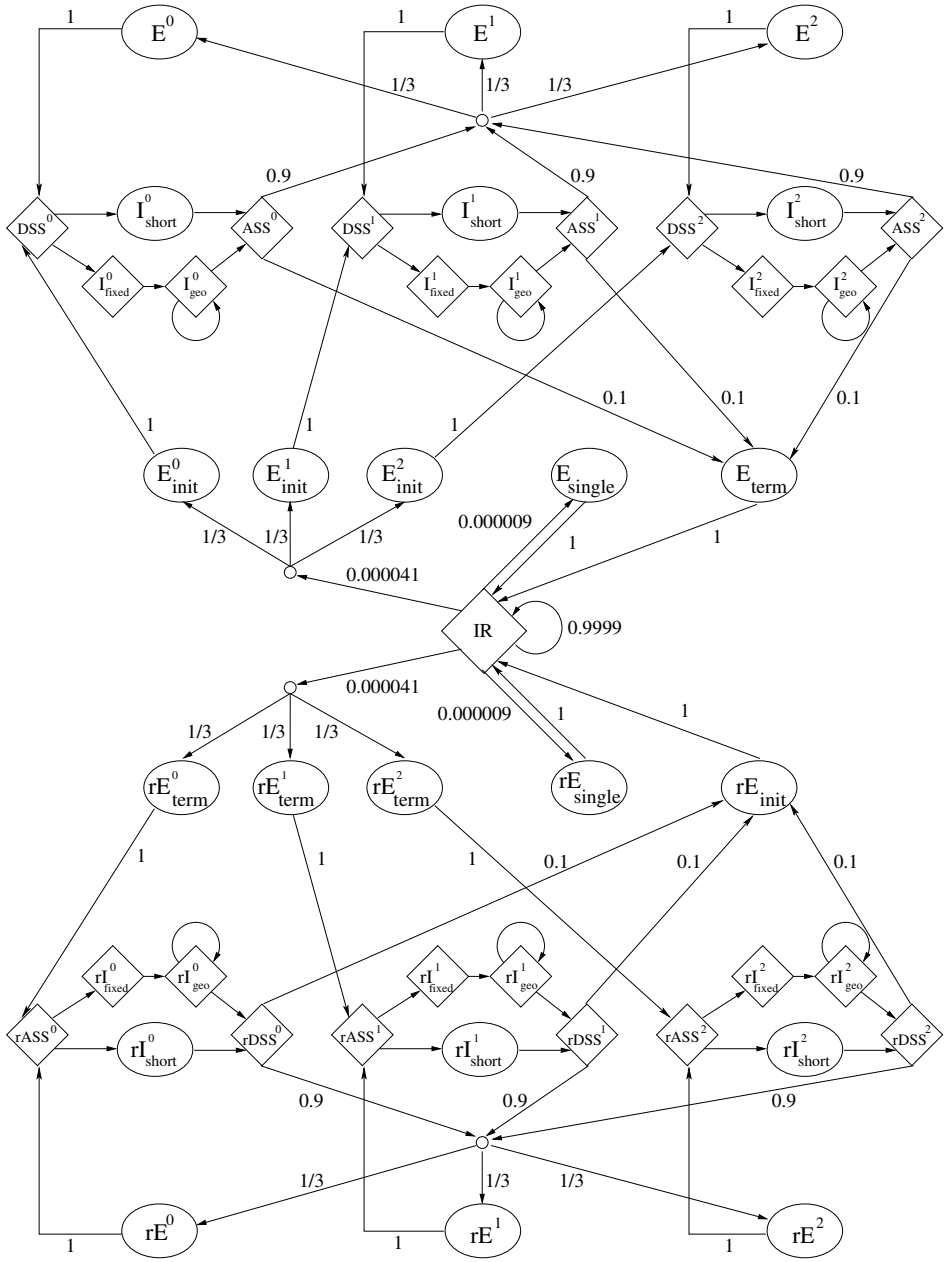


A Simple (Semi-Markov) HMM for Gene Finding: Model Topology

Model for (multiple) eukaryotic genes on forward strand



Transition Matrix of AUGUSTUS (without UTRs)



Conservation of Gene Structure and Sequence

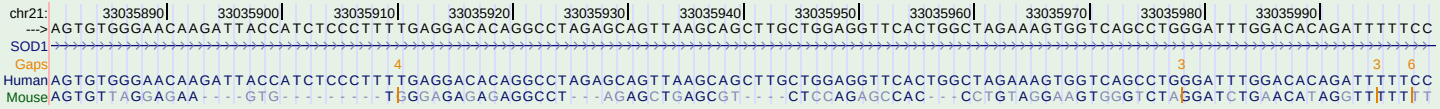
Observation

Protein sequences and rough structure of genes are often conserved between species that are tens of millions of years separated.

Example (Human-Mouse: 75 million years)

- 95% of orthologous gene pairs have same number of exons in human and mouse
- coding sequence to ≈ 85% identical

- noncoding sequence to ≈ 35% identical



Conservation of Gene Structure and Sequence

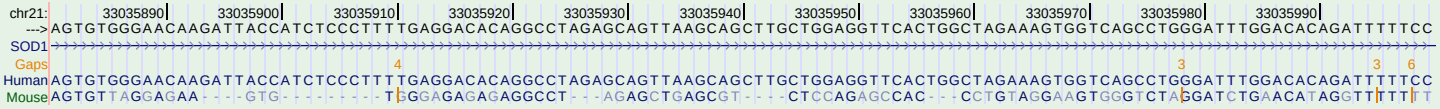
Observation

Protein sequences and rough structure of genes are often conserved between species that are tens of millions of years separated.

Example (Human-Mouse: 75 million years)

- 95% of orthologous gene pairs have same number of exons in human and mouse
- coding sequence to ≈ 85% identical

- noncoding sequence to ≈ 35% identical



Conservation of Gene Structure and Sequence

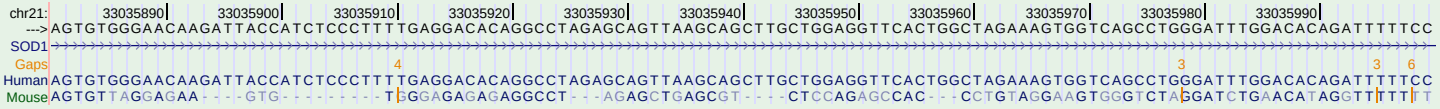
Observation

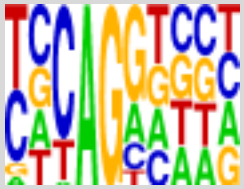
Protein sequences and rough structure of genes are often conserved between species that are tens of millions of years separated.

Example (Human-Mouse: 75 million years)

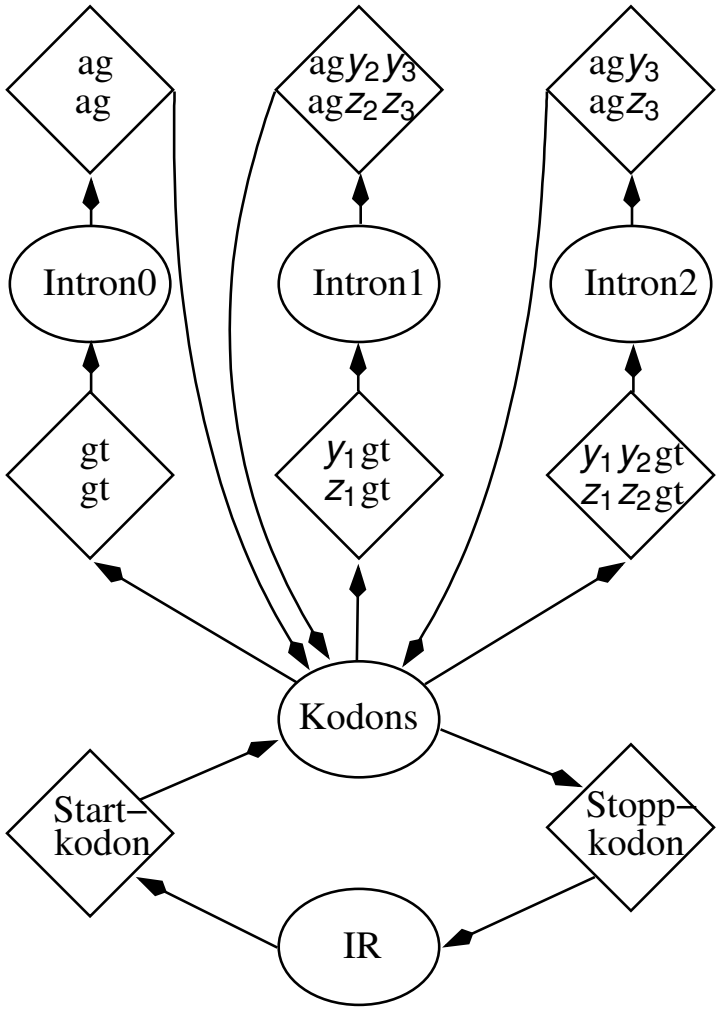
- 95% of orthologous gene pairs have same number of exons in human and mouse
- coding sequence to $\approx$ 85% identical

- noncoding sequence to $\approx$ 35% identical





Pair HMMs for Eukaryotic Gene Finding



- DOUBLESCAN (Meyer and Durbin, 2002)
- does alignment of two genomics sequences and gene prediction in both at same time
- running time quadratic in sequence length
- does not generalize efficiently to >2 species
- not popular

CONTRAST (Gross et al., 2007)

CONTRAST

- still best *de novo* predictor for human (?)
- discriminatory model (SVM+CRF)
- uses 11 other vertebrate genomes as informants
- roughly for 50% of human genes at least one splice form is correctly predicted

```
human       ACAGGTGAGGAGGCG
macaque     ...............
mouse       ...............
rat         ACAGGTGAGAAAG..
rabbit      ...............
dog         ACAGGTGAGGAGTCG
cow         ACAGGTGAGCAGTCG
armadillo   ACAGGTGAGGAG_CA
elephant    ...............
tenrec      ...............
opossum     CCAGGGAAG......
chicken     CCAGGTGA.......
EST         SSSSIIIIIIIIIII
```

Using Extrinsic Evidence to Improve the Accuracy of AUGUSTUS

Extrinsic evidence

additional information to aid gene prediction, not contained in the sequence

AUGUSTUS currently can use

- mRNA/EST/454 alignments, RNA-Seq
- genomic conservation
- annotation of related species using a syntenic alignment
- protein alignments, peptide mass spectrometry

Hint

piece of additional local information about gene structure, in general uncertain

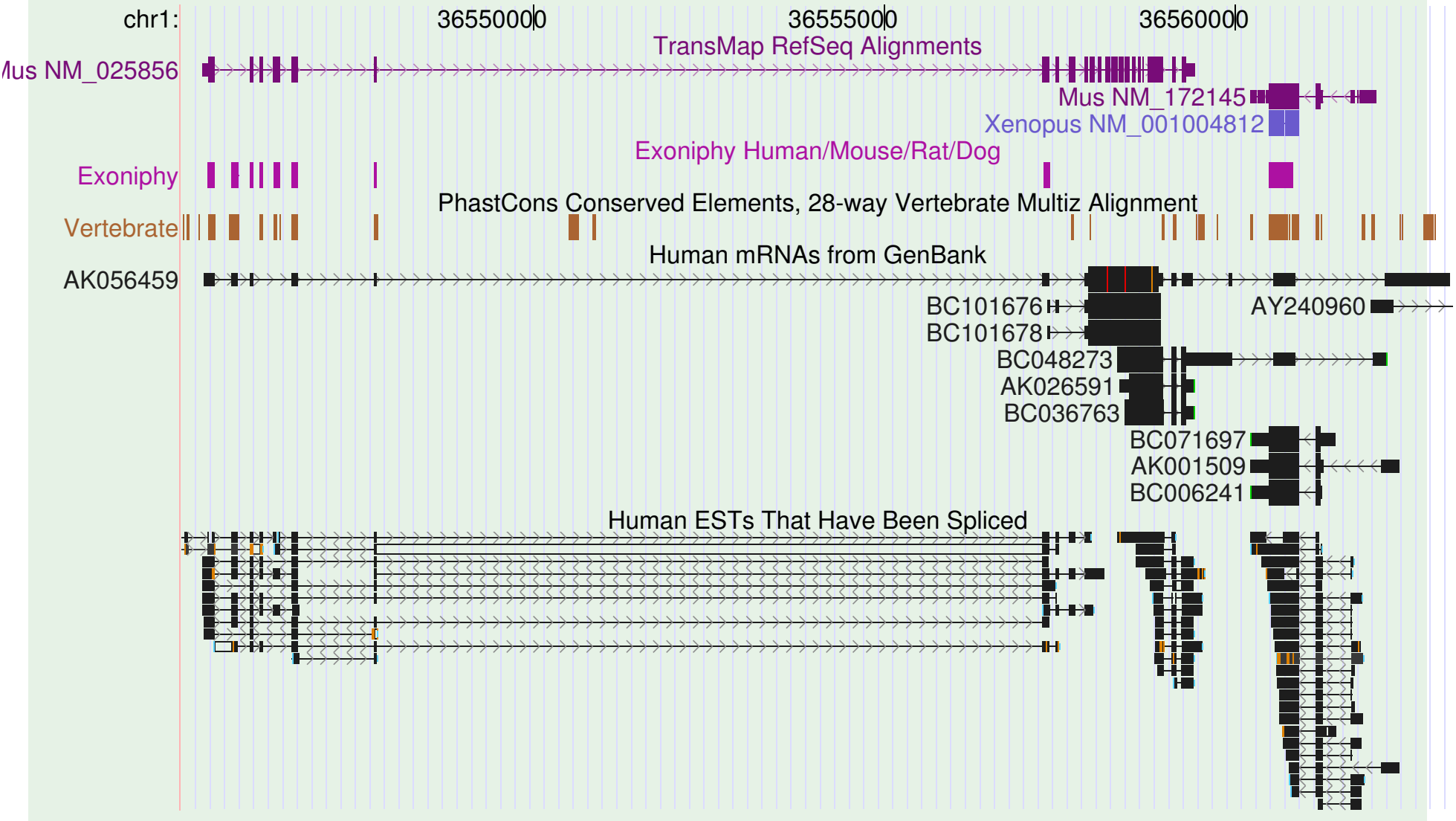
Types of Hints

AUGUSTUS understands these hints...

type	meaning	source of information
start	start codon	protein, TransMap
stop	stop codon	protein, TransMap
tss	transcription start site	TransMap, CpG islands, promoter predictor
tts	transcription termination site	TransMap, 3' ESTs
ass	acceptor splice site	cDNA, proteins, TransMap, conservation
dss	donor splice site	cDNA, proteins, TransMap, conservation
exonpart	part of biological exon	cDNA
exon	biological exon	cDNA
intronpart	part of intron	TransMap
intron	intron	cDNA, proteins, TransMap
CDSpart	part of coding part of exon	protein, TransMap, conservation
CDS	coding part of exon	protein, conservation
UTRpart	part of non-coding part of exon	TransMap
UTR	non-coding part of exon	
irpart	part of intergenic region	mRNA, TransMap, cDNA
nonexonpart	intron or intergenic region	retro genes, repeats

Sources of Hints - Example

Example



Integration of Hints into the Model

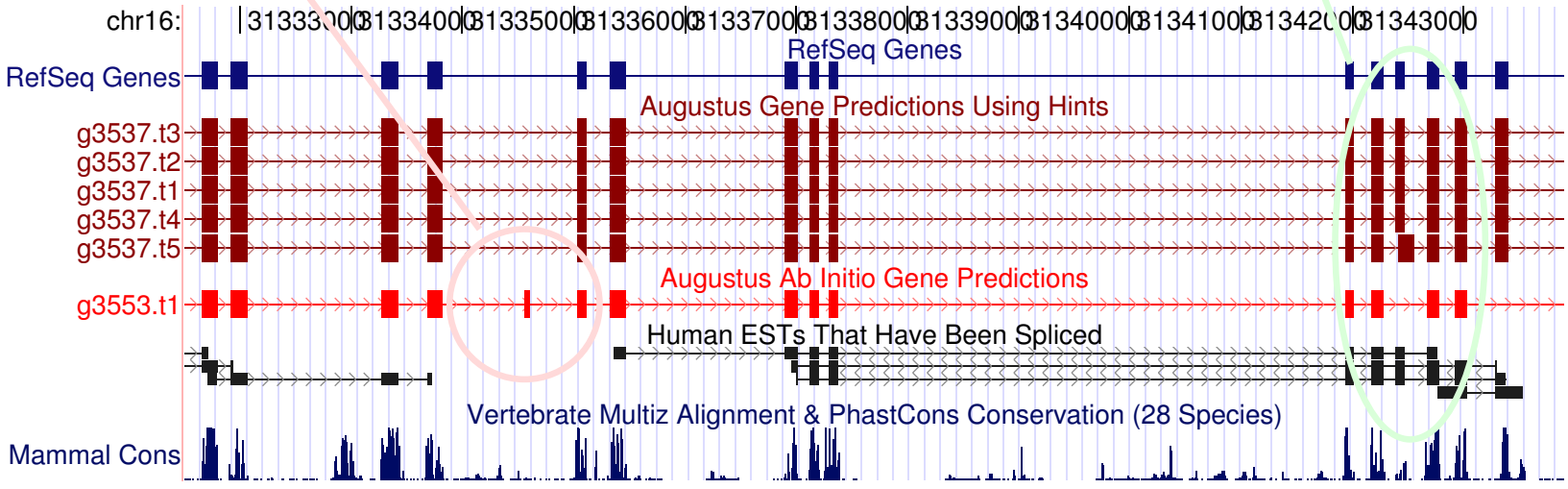
Hints modify the joint probability

$$P(\text{gene structure}, \text{DNA}, \text{hints})$$

by a factor that depends on the compatibility of the gene structure with the hints.

Hints have two effects

1. gene structure candidates that obey a hint get a relative bonus.
2. gene structure candidates with unsupported exons, introns or signals get a relative malus (penalty)



Integration of Hints into the Model

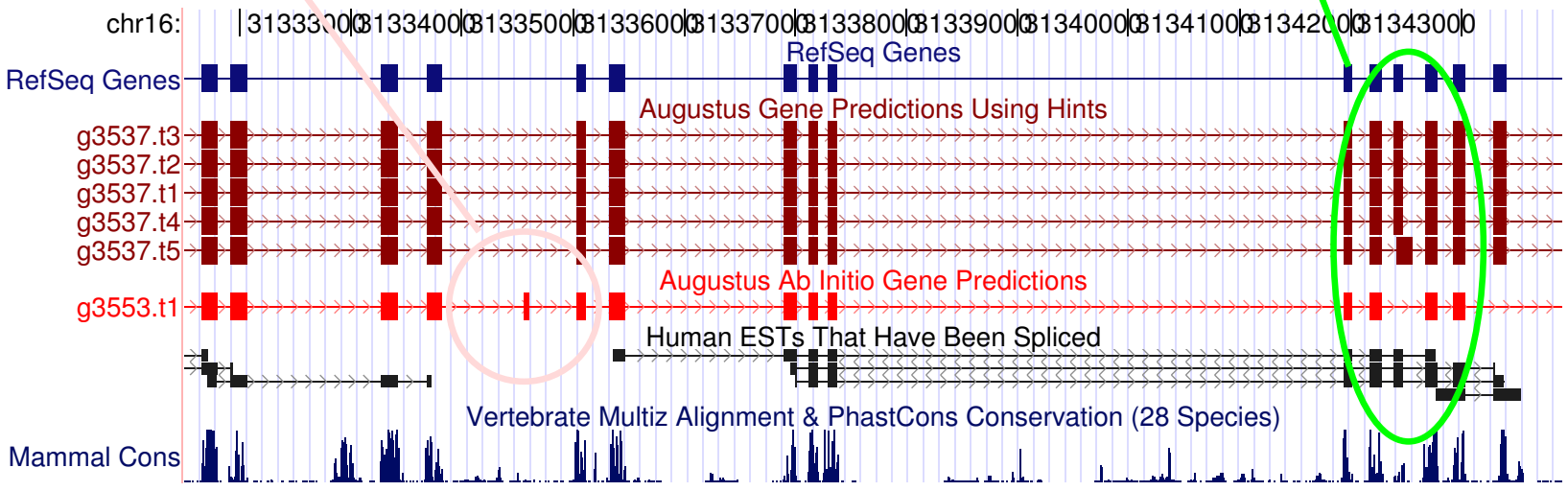
Hints modify the joint probability

$$P(\text{gene structure}, \text{DNA}, \text{hints})$$

by a factor that depends on the compatibility of the gene structure with the hints.

Hints have two effects

1. gene structure candidates that obey a hint get a relative bonus.
2. gene structure candidates with unsupported exons, introns or signals get a relative malus (penalty)



Integration of Hints into the Model

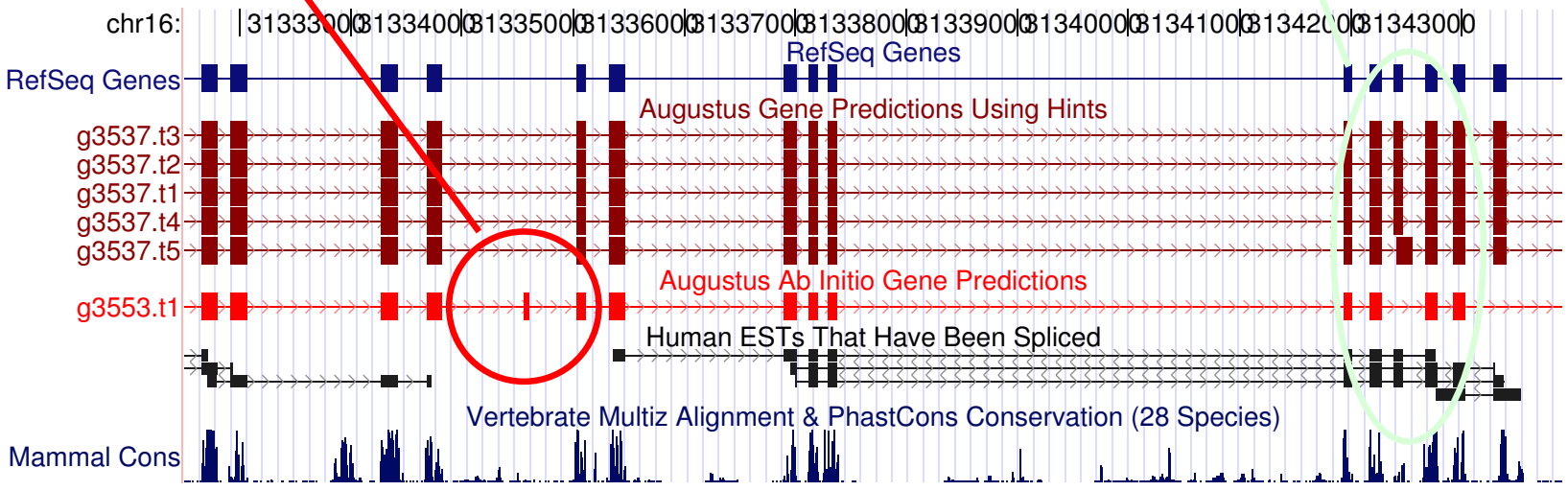
Hints modify the joint probability

$$P(\text{gene structure}, \text{DNA}, \text{hints})$$

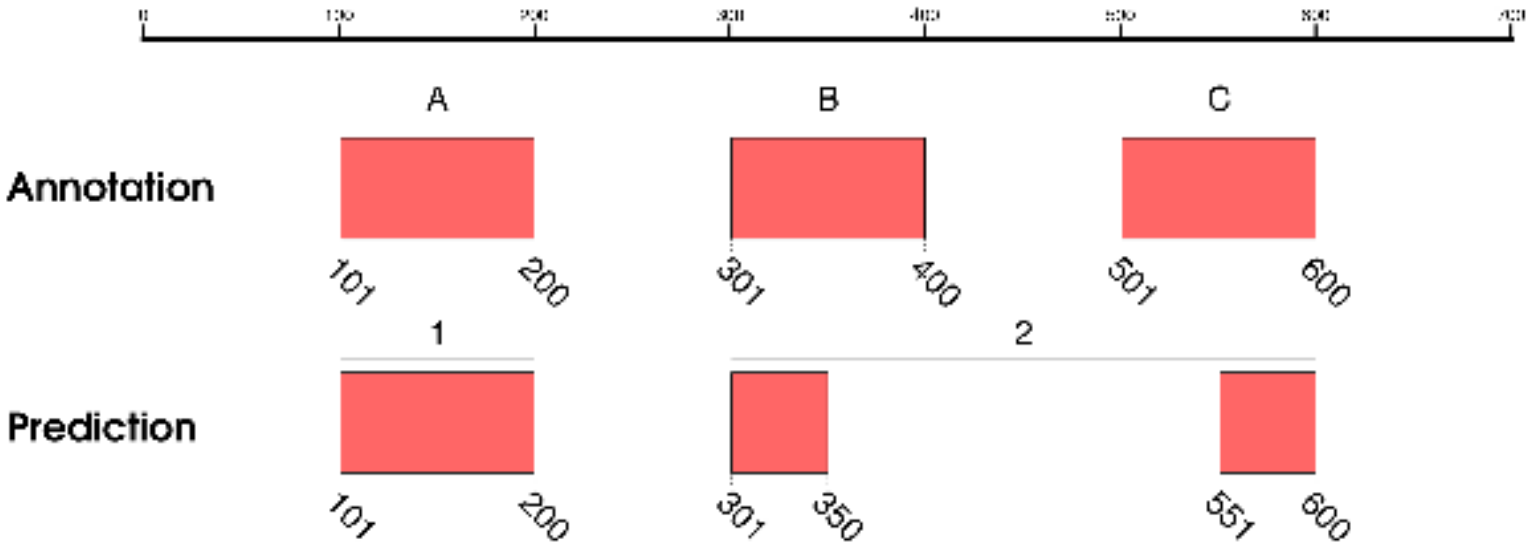
by a factor that depends on the compatibility of the gene structure with the hints.

Hints have two effects

1. gene structure candidates that obey a hint get a relative bonus.
2. gene structure candidates with unsupported exons, introns or signals get a relative malus (penalty)



Prediction Evaluation Measures



$$\text{Sensitivity sn} = \frac{\text{true positives}}{\text{annotated positives}}$$

$$\text{Specificity sp} = \frac{\text{true positives}}{\text{predicted positives}}$$

Gene sn	33.3%
Gene sp	50%
Transcript sn	33.3%
Transcript sp	50%
Exon sn	33.3%
Exon sp	33.3%
Base sn	66.7%
Base sp	100%

Accuracy of Genome Annotation

Two independent published gene prediction assessments since 2000:

- EGASP on human ENCODE regions, 2005:

  genes correctly predicted:

ab initio	24% (AUGUSTUS)
using any available info	73% (JIGSAW)

  (Guigo et al., *Genome Biology*, 2006)

- nGASP on *C. elegans*, 2007:

  genes correctly predicted:

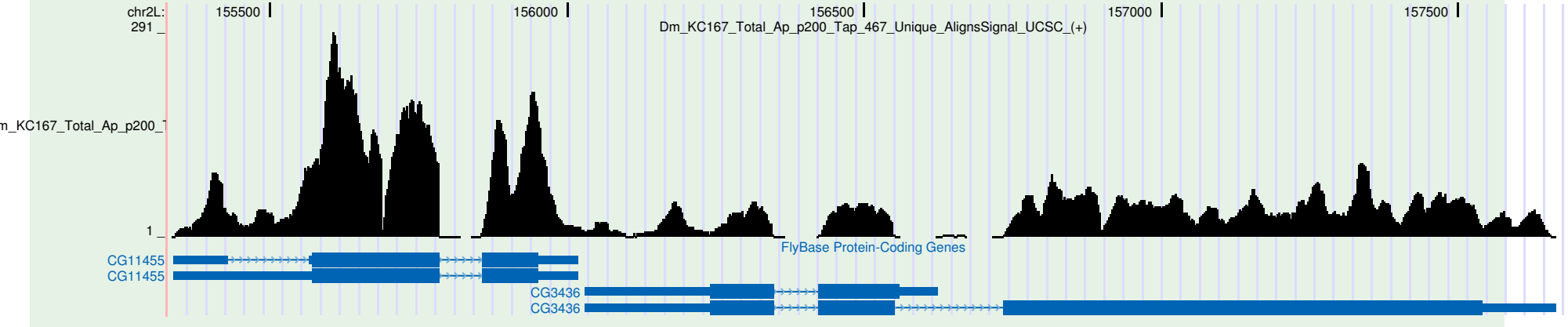
ab initio	61% (AUGUSTUS, mGENE, FGENESH)
using ESTs & proteins	80% (AUGUSTUS, FGENESH, mGENE)
best-performing combiner	80% (JIGSAW)

  (Coghlan et al., *BMC Bioinformatics*, 2008)

Note: Most species have less available evidence than these two!

Next Generation Sequencing of Transcriptomes (RNA-Seq)

Example (coverage per position in genome)



Chances and Challenges

- genes expressed in sample have usually good (unprecedented) coverage
- intuitively, should give boost to gene prediction accuracy
- but how?

RGASP: RNA-Seq Genome Annotation Assessement Project

RGASP

- 17 participating groups submitting 1 year ago, all on same data
- on human, fly, worm
- Illumina, SOLiD, Helicos
- paired, unpaired reads
- unstranded, stranded reads

Two major approaches

- evidence integration into gene finder
  1. align reads to genome
  2. integrate evidence from coverage and spliced alignments into gene finder
- de novo assembly
  1. assemble transcriptome reads into transcript contigs
  2. use contigs for gene finding or just align them

RGASP: RNA-Seq Genome Annotation Assessement Project

RGASP

- 17 participating groups submitting 1 year ago, all on same data
- on human, fly, worm
- Illumina, SOLiD, Helicos
- paired, unpaired reads
- unstranded, stranded reads

Two major approaches

- evidence integration into gene finder
  1. align reads to genome
  2. integrate evidence from coverage and spliced alignments into gene finder
- de novo assembly
  1. assemble transcriptome reads into transcript contigs
  2. use contigs for gene finding or just align them

RGASP: preliminary results

My Conclusions

- RNA-Seq did not help as much as expected:
  - substantial intronic transcription
  - mapping ambiguity due to self-similarity of genomes
- still a lot of potential for improvement

Proteomics and Genomics: Proteogenomics

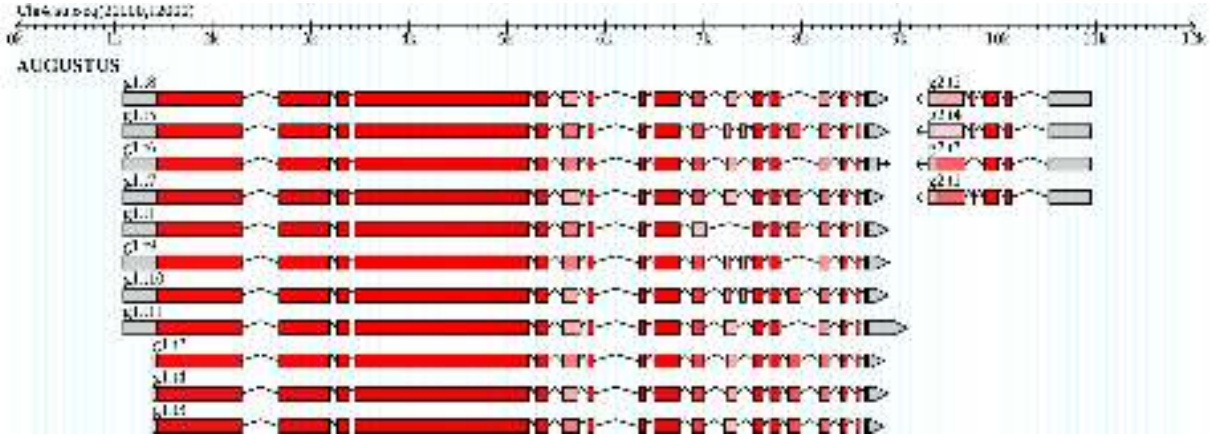
Peptide from Tandem Mass Spectrometry (MS/MS) for Gene Prediction

- unknown proteins of interest cleaved into smaller peptides
- masses of peptides measured with a mass spectrometer
- masses compared *in silico* with peptides from a candidate protein database
- result: set of peptides that is actually translated

Candidate Protein Database

should be very inclusive

1. six-frame translation of genome
2. known genes
3. predicted genes:
   - sample many gene structures with AUGUSTUS

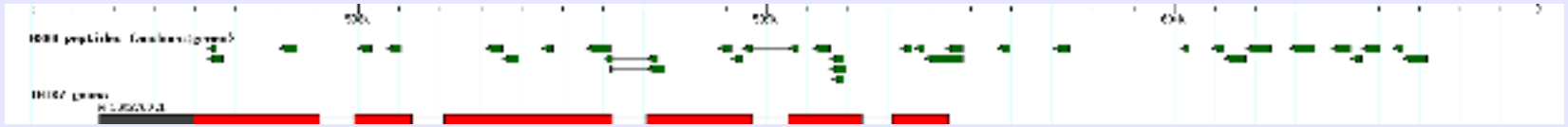


Map Peptides to Genome

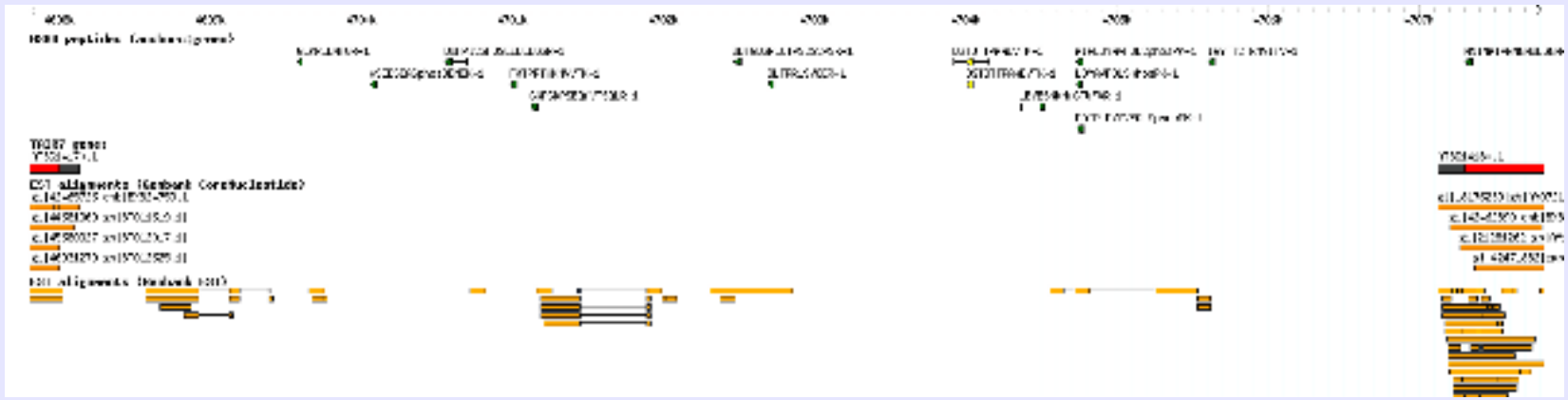
Peptides can

- match protein from set of “known genes”: Partially confirm gene structure, existence of gene
- not match any known gene: What then?
  Extend known gene? Join two known genes? New splice form?
  Replace splice form of “known” gene?

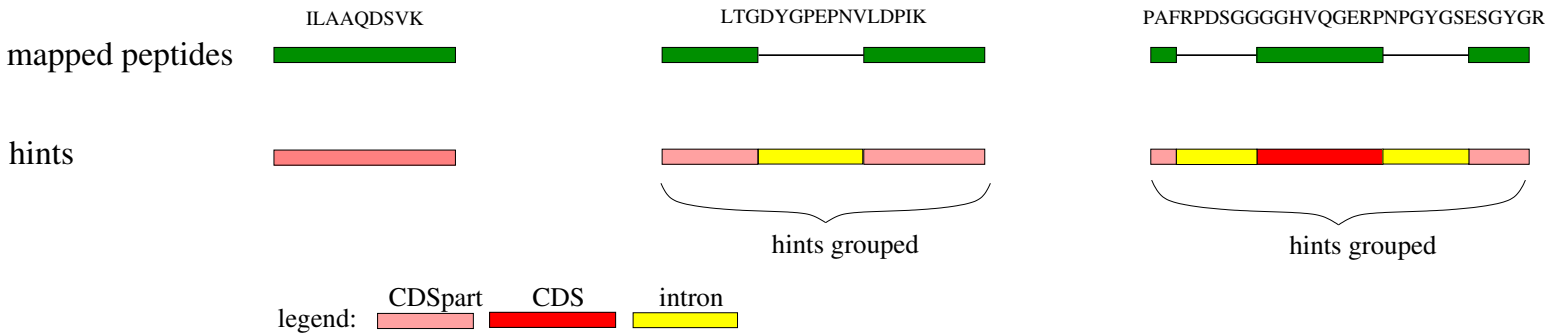
example 1:



example 2:



Peptides to “Hints”



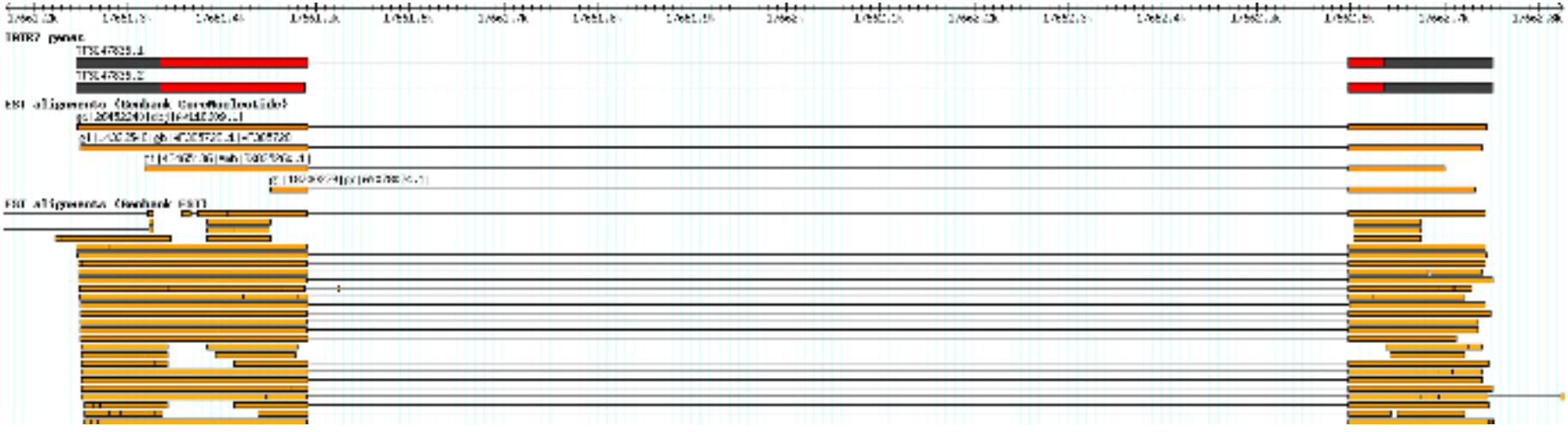
(average peptide length: 15.2)

CDSpart hint:	region is likely protein-coding
CDS hint:	region is exact protein-coding exon
intron hint:	region is likely an intron with exact boundaries

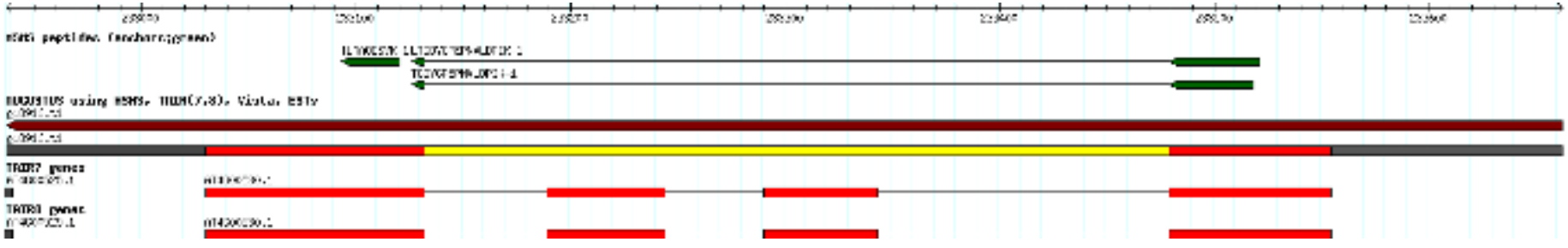
Hints can include strand and frame information.

Examples for Gene Corrections

Who would have doubted this gene?



precious *spliced* peptides



Conclusions on Proteogenomic Evidence

MS/MS

- peptides not only good for “verification” of given gene structures but also valuable, additional orthogonal source of evidence for predicting gene structures
- new translation-level info (reading frame)
- not enough coverage for identification of alternative splicing based on MS/MS alone

Incorporating Protein Homology: "The Pipeline Way"

Use known proteins from other species

1. align proteins to genome
   (e.g. with BLASTX, TBLASTN, Exonerate)
2. use matches from step 1 as evidence for coding regions in gene finding
   (many gene finders)

Disadvantages

- step 1 misses info: no weak homology or many false positives
- gene finders as subject to split gene errors
- protein family MSAs hold more info than individual proteins

Incorporating Protein Homology: "The Pipeline Way"

Use known proteins from other species

1. align proteins to genome
   (e.g. with BLASTX, TBLASTN, Exonerate)
2. use matches from step 1 as evidence for coding regions in gene finding
   (many gene finders)

Disadvantages

- step 1 misses info: no weak homology or many false positives
- gene finders as subject to split gene errors
- protein family MSAs hold more info than individual proteins

Protein Homology in One Step

Genewise (Birney 2004)

- marry (multiply) two HMMs:
  - simple (species-unspecific) HMM for gene prediction with
  - profile HMM for protein family (like HMMer)
- one big (product) HMM that does both:
  - gene finding and
  - protein to protein-family alignment + membership decision

Disadvantages

- slow
- poor performance for weak homology
  cannot cope with missing homology
- no integration with other evidence (RNA-Seq, ESTs)

Protein Homology in One Step

Genewise (Birney 2004)

- marry (multiply) two HMMs:
  - simple (species-unspecific) HMM for gene prediction with
  - profile HMM for protein family (like HMMer)
- one big (product) HMM that does both:
  - gene finding and
  - protein to protein-family alignment + membership decision

Disadvantages

- slow
- poor performance for weak homology
  cannot cope with missing homology
- no integration with other evidence (RNA-Seq, ESTs)

Just Published...

Bioinformatics Advance Access published January 6, 2011

A novel hybrid gene prediction method employing protein multiple sequence alignments

Oliver Keller [1,*], Martin Kollmar [2], Mario Stanke [3,*] and Stephan Waack [1]

AUGUSTUS-PPX: Protein Profile eXtension

- find members of given protein family (MSA) in DNA
- seamless extension of AUGUSTUS:
  - same program, same species parameters, etc.
  - UTR prediction possible
  - additional hints integration possible, e.g. RNA-Seq
  - defaults to normal AUGUSTUS for non-member genes and gene parts without homology
- **Try out this afternoon!**

AUGUSTUS-PPX

Block

Ungapped and highly conserved section of a MSA

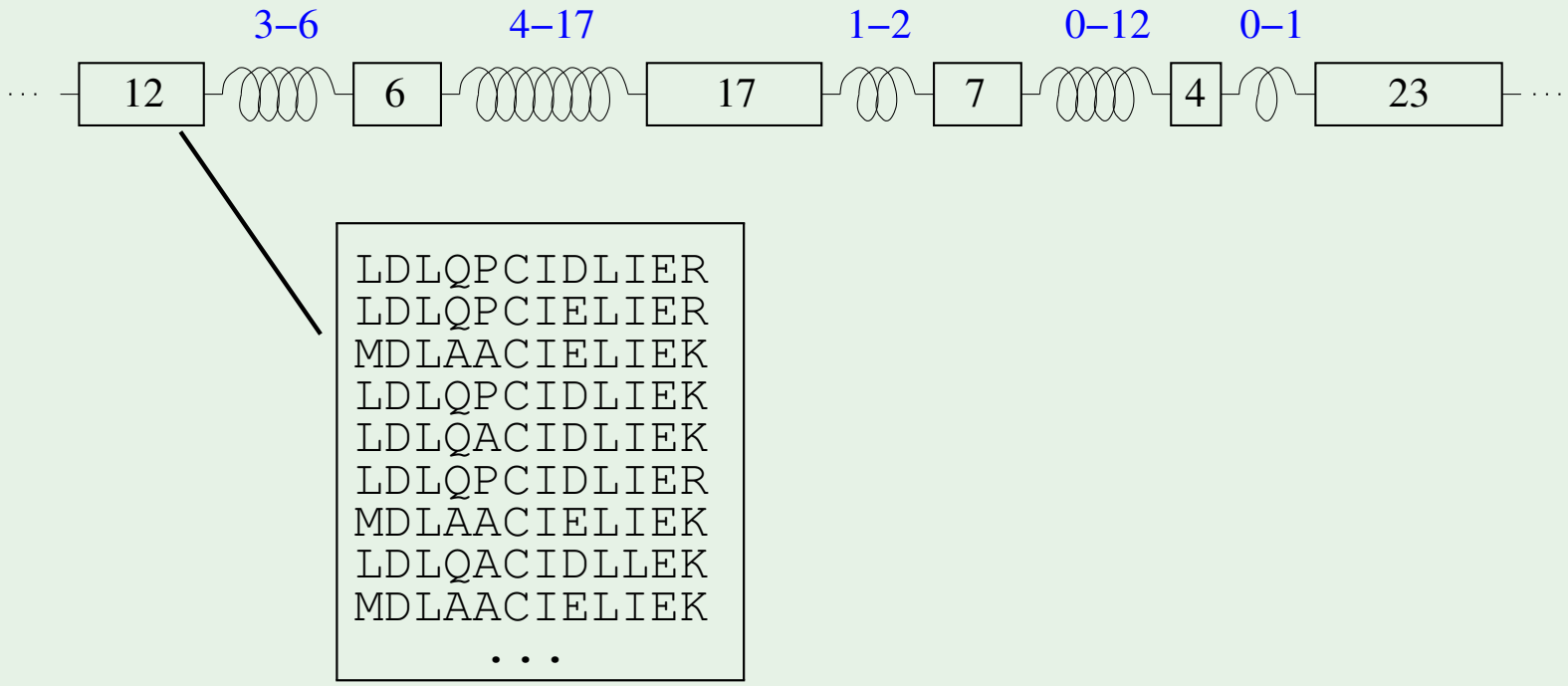
Example

AUGUSTUS-PPX

Block Profile

- frequency table for each block column
- distances between blocks constrained by intervals $[d_{\text{min}}, d_{\text{max}}]$
- no insertions or deletions
- order of blocks preserved

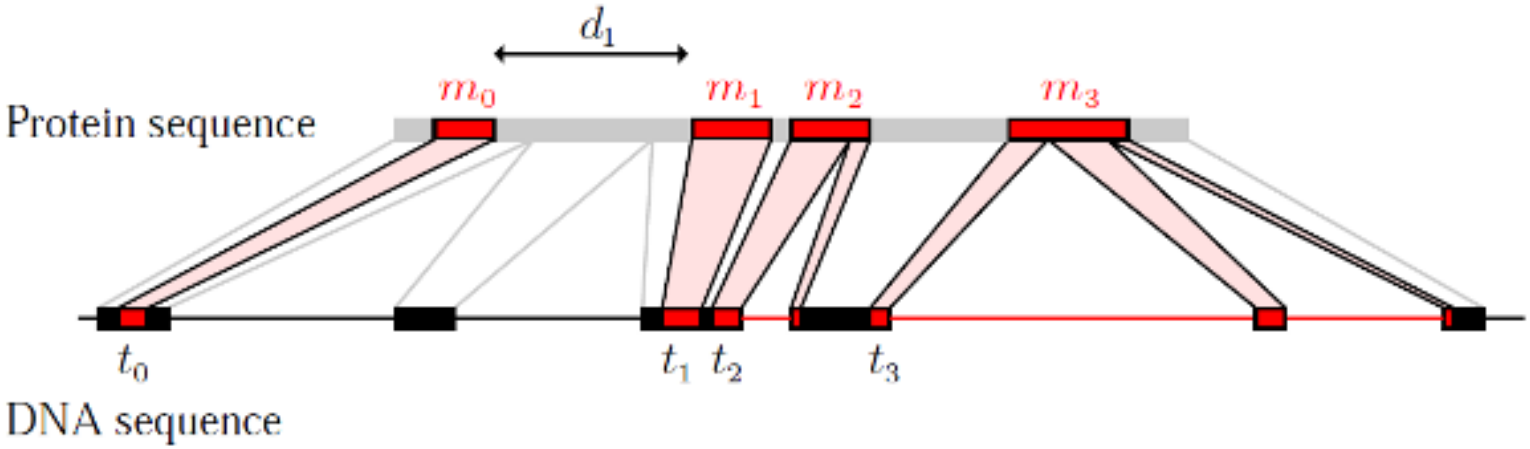
Example (Myosins)



AUGUSTUS-PPX

Problem:

Find gene structure in genome under hypothesis, that predicted protein sequence fits into family.

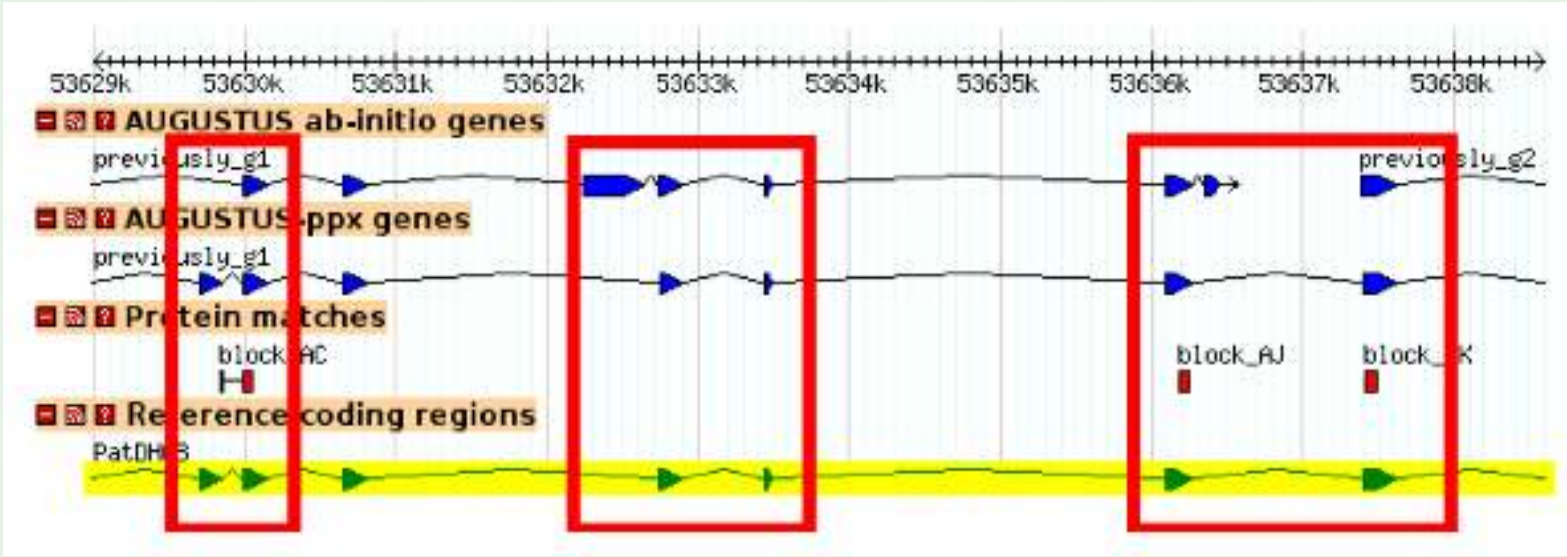


Scoring

$$\begin{aligned} & \text{score}(\text{gene structure}, \text{DNA}, \text{block profile}) \\ = \; & \text{score}(\text{gene structure}, \text{DNA}) \\ & \cdot \text{score}(\text{block profile} \mid \underbrace{\text{gene structure}, \text{DNA}}_{\rightarrow \text{protein sequence cand.}}) \end{aligned}$$

AUGUSTUS-PPX: How it helps

Example



Corrections w.r.t. ab initio

- add exons containing blocks
- remove false positive exons (constraint on inter-block lengths)
- join split genes

AUGUSTUS-PPX: Results

Test on: Dynein Heavy Chain (DHC) family

- > 4000 amino acids, up to 100 exons, several 100 000 bp genomic range
- 42 Blocks with 1214 sites

Accuracy Results

species	AUGUSTUS-PPX		AUG.	Genewise	
	full	ex-ortho	ab-initio	full	ex-ortho
highly accurate genes (%)					
human	62.5	62.5	31.3	93.8	6.3
mouse	72.2	61.1	22.2	55.6	0.0
chicken	58.3	50.0	8.3	16.7	0.0
frog	44.4	38.9	0.0	11.1	0.0
zebrafish	57.1	35.7	57.1	14.3	0.0
snail	44.4	44.4	5.6	11.1	0.0
total	56.3	49.0	11.5	34.4	1.0

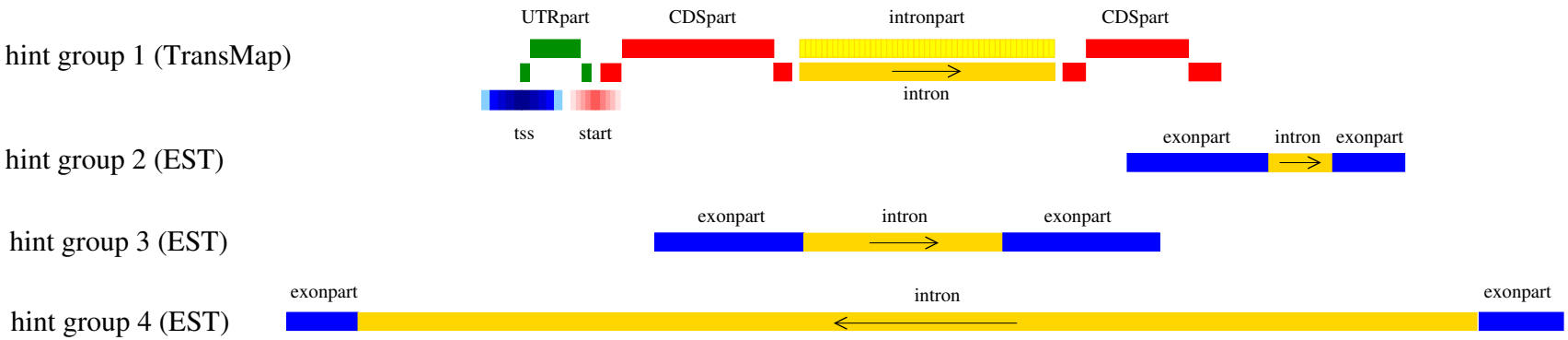
Evidence Suggests Alternative Splicing

AUGUSTUS: Alternative Transcripts Based on Evidence

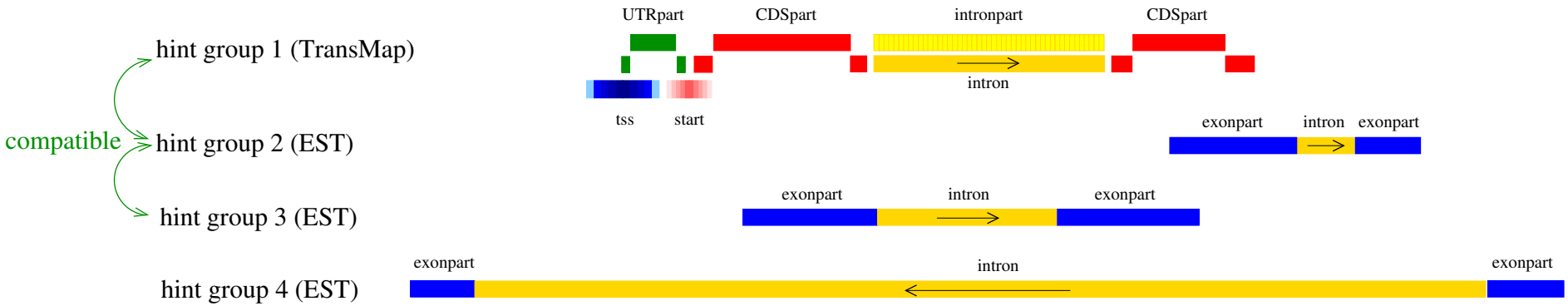
Hints can be grouped to tell AUGUSTUS they are coming from the same transcript.
For example: All hints from one cDNA alignment belong to one group. Based on evidence predict alternative splice forms.

1. Split input sequence between genes based on hint groups and preliminary gene prediction. For each piece, do:
2. Determine compatibility between hint groups.
3. Filter out outlier hint groups.
4. For each hint group g, do:
    1. Deactivate all hint groups h that are incompatible with g.
    2. Run Viterbi algorithm to predict single-transcript genes.
5. Join transcripts to genes

Alternative Transcripts Based on Evidence

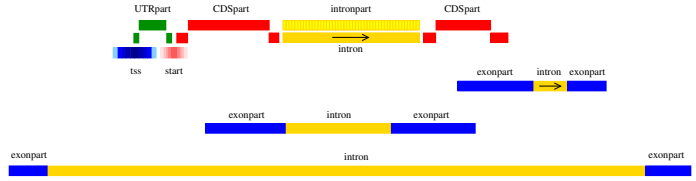


Alternative Transcripts Based on Evidence



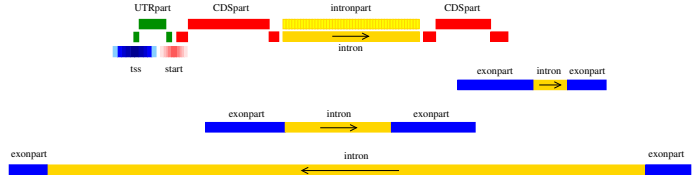
Alternative Transcripts Based on Evidence

prediction run 1



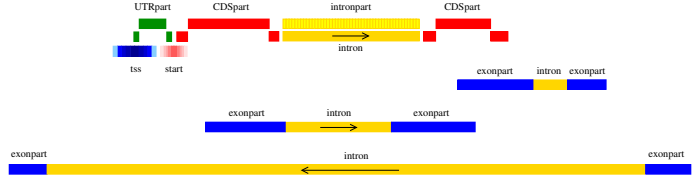
---

prediction run 2



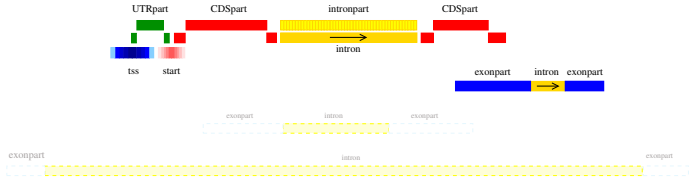
---

prediction run 3

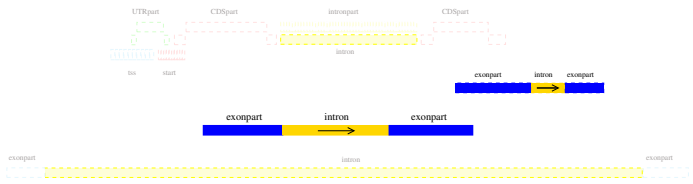


Alternative Transcripts Based on Evidence

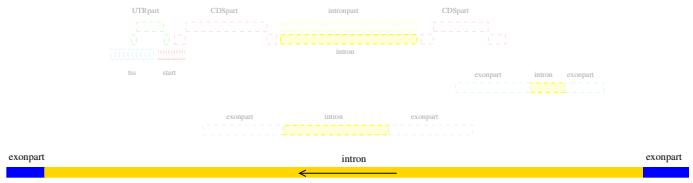
prediction run 1



prediction run 2

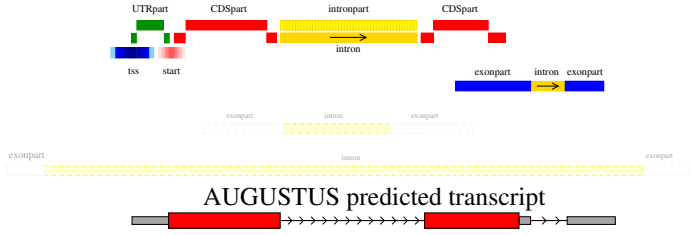


prediction run 3

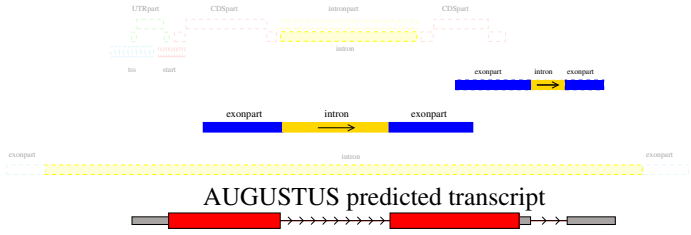


Alternative Transcripts Based on Evidence

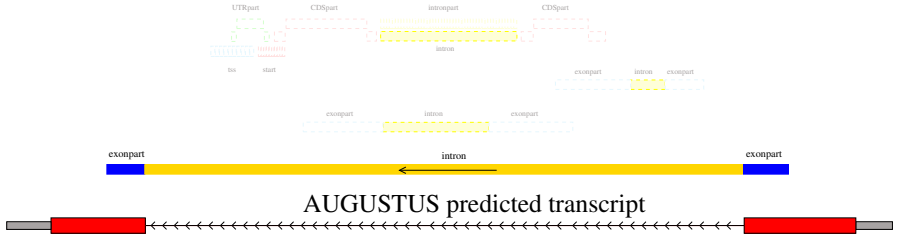
prediction run 1



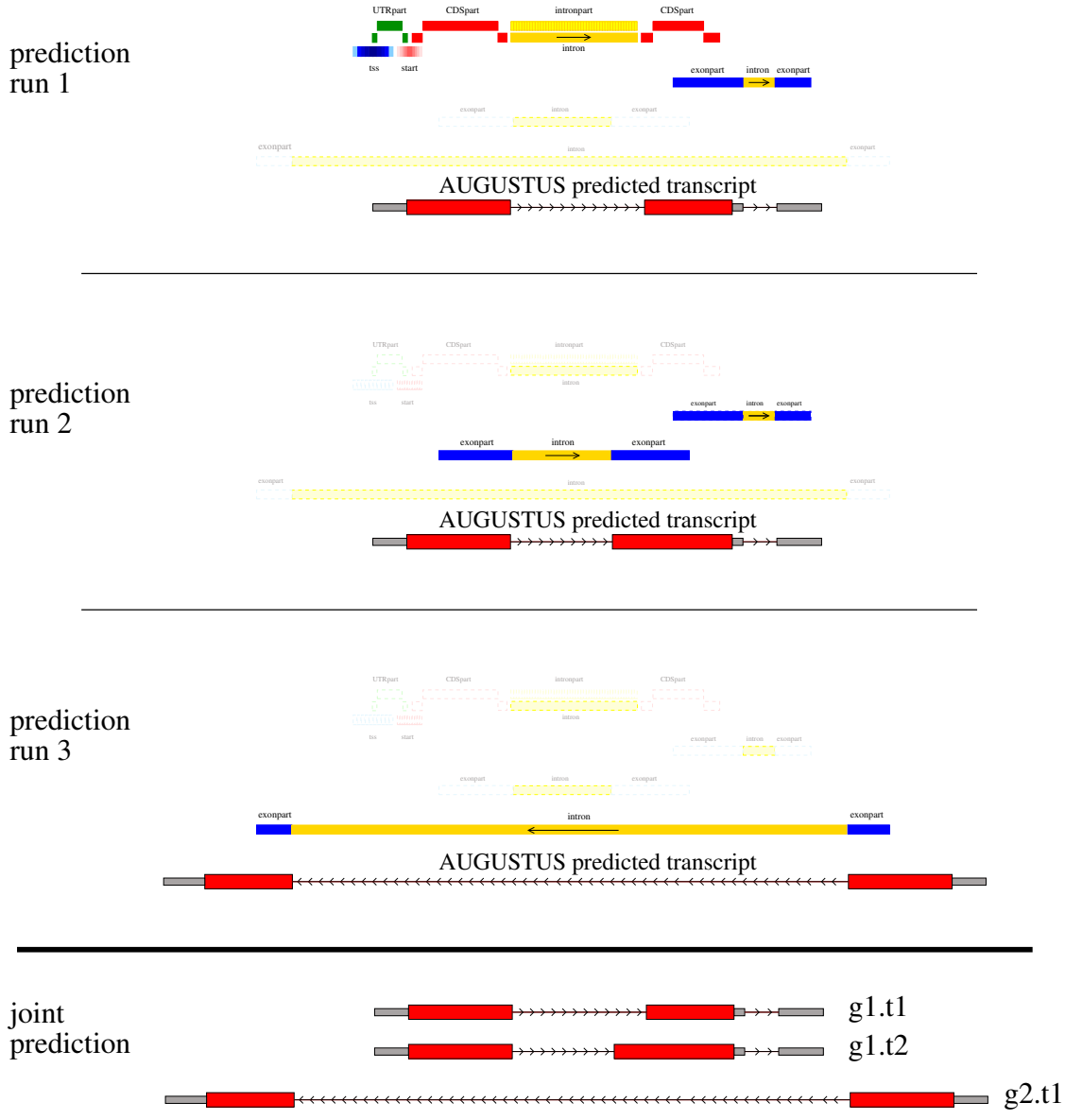
prediction run 2



prediction run 3



Alternative Transcripts Based on Evidence



Acknowledgements

Thanks to

Stephan Waack **Oliver Keller** Burkhard Morgenstern	Göttingen
David Haussler Mark Diekhans	UC Santa Cruz
Michael Hippler Michael Specht	Münster
Vineet Bafna Natalie Castellana Sam Payne	UC San Diego