Algorithmen der Bioinformatik I
WS 2017/2018

Burkhard Morgenstern
Peter Meinicke

Dept. Bioinformatics
Institute of Microbiology and Genetics (IMG)
University of Göttingen

January 9, 2018

Gene finding, recap.

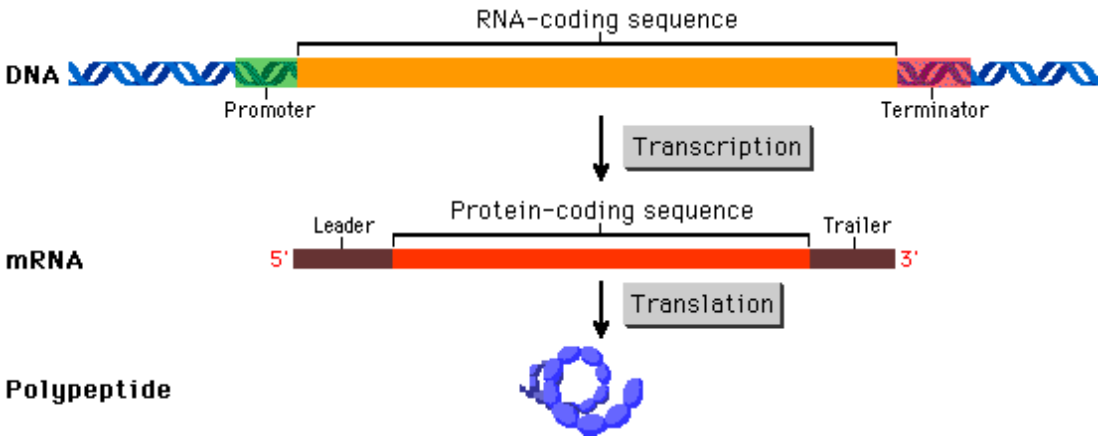


Figure: In *prokaryotes*: coding regions not interrupted by introns.

Simple models for prokaryotes, recap.

Modelling start and stop codons at begin and end of coding region.

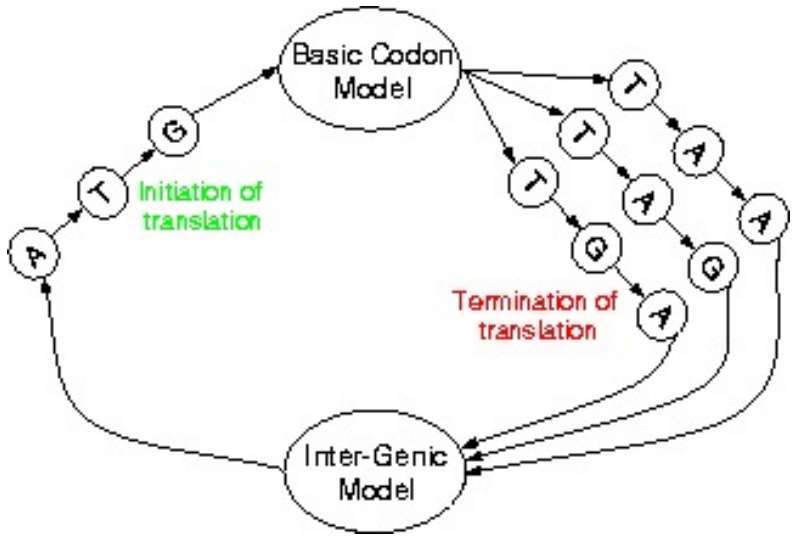


Figure: HMM for gene finding with start and stop codons
(source: `stat.berkeley.edu`)

Note: codon frequencies can be coded through transition probabilities or emission probabilities.

EcoParse, recap.

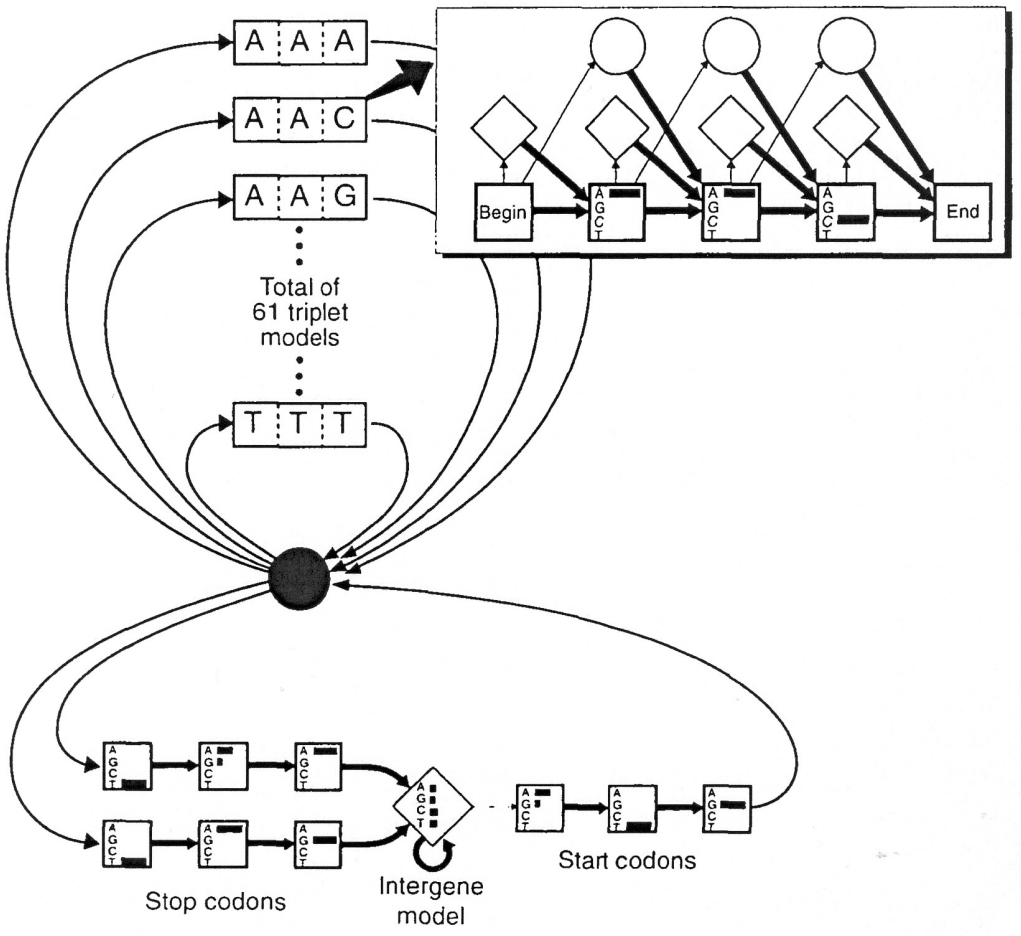


Figure: Structure ('topology') of HMM for gene finding in *EcoParse* (Krogh *et al.*, 1994)

Complex HMMs, recap.

(d) Generalized HMMs (GHMMs)

Explicit modelling of time spent in given state.

So far: time in state depends on transition probabilities. Result: *geometric* probability distribution for length of genes and intergenic regions.

Let p be probability to leave given state a. Probability for model to stay *exactly* n times in state a:

$$p^n \cdot (1 - p)$$

Complex HMMs, recap.

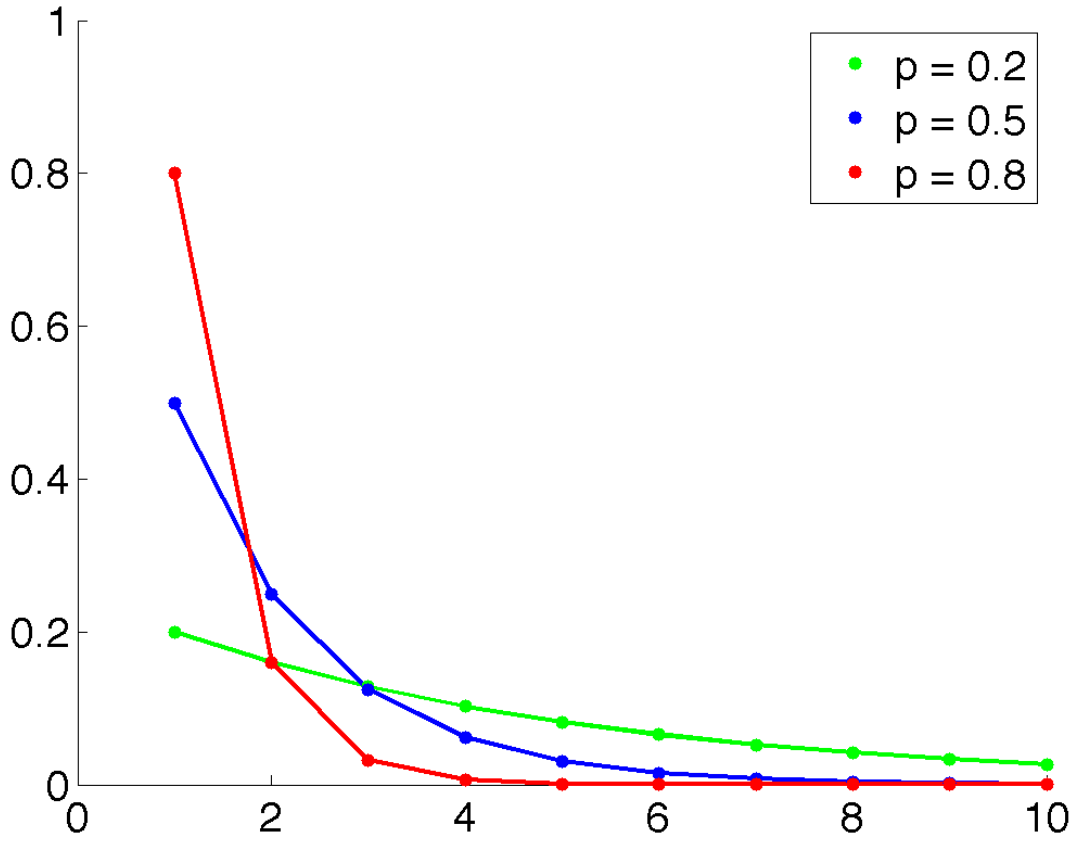


Figure: Geometric distribution for different parameters p (Wikipedia)

Complex HMMs, recap.

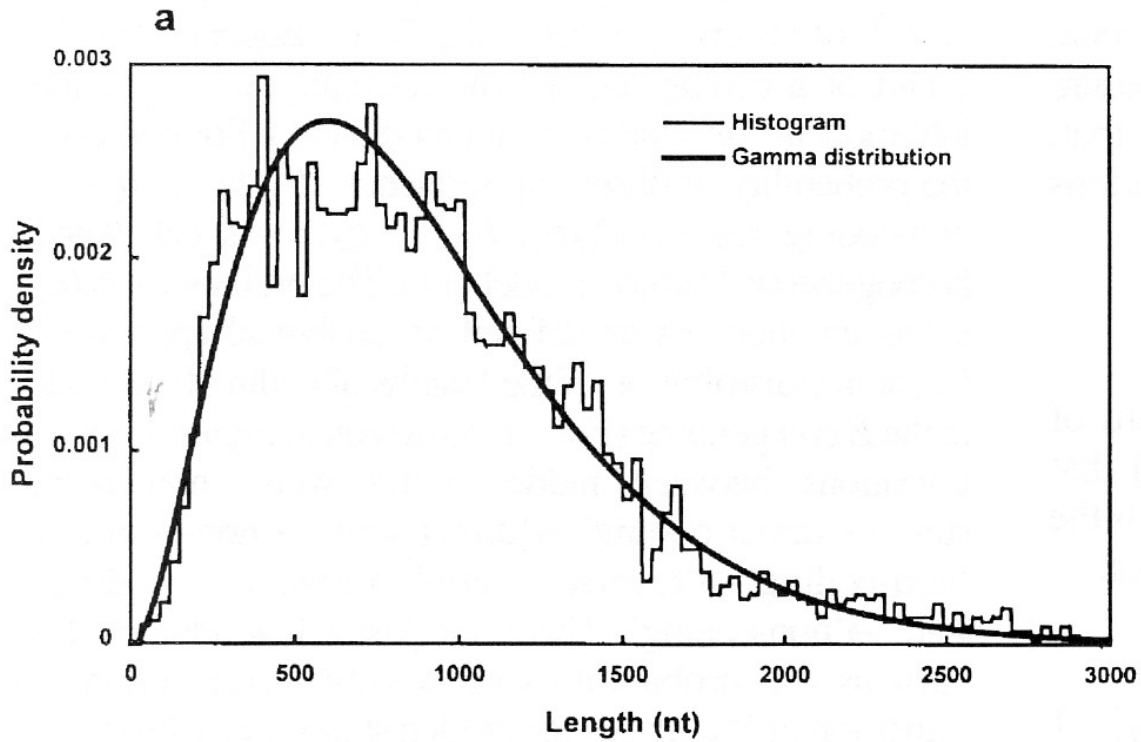


Figure: Real length distribution of protein-coding regions in *E. coli* (Lukashin and Borodovsky, 1998)

Complex HMMs, recap.

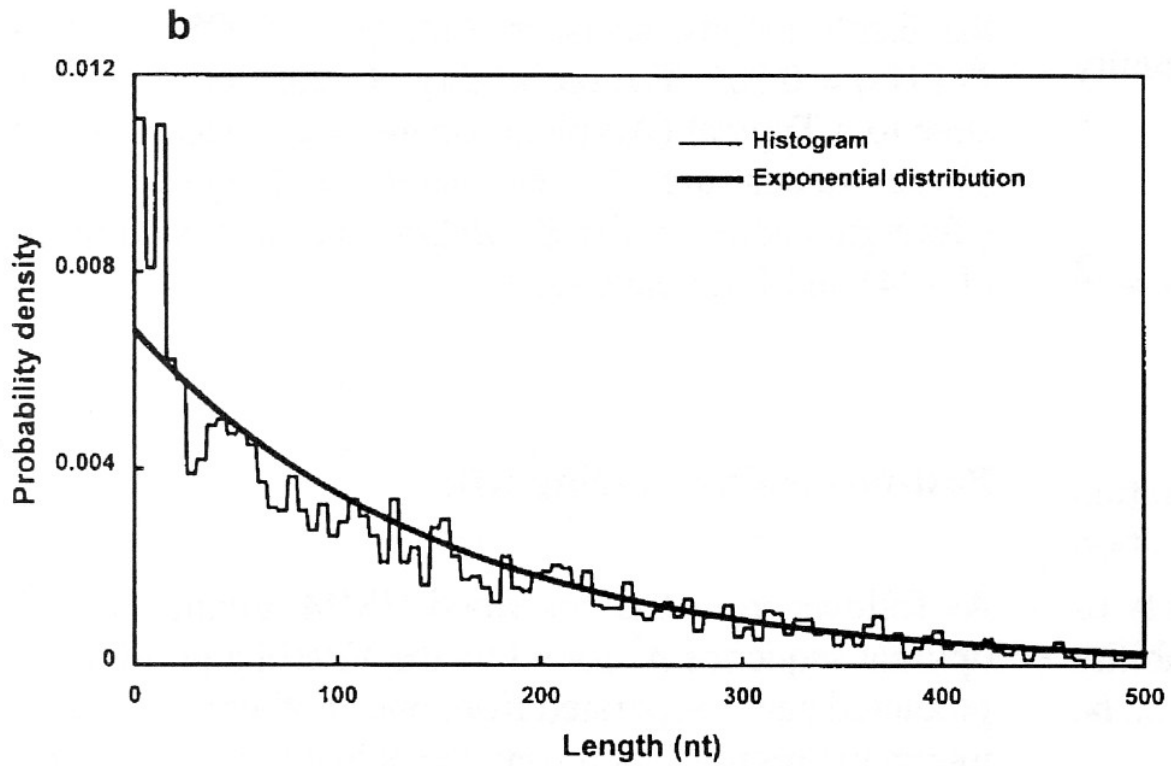


Figure: Real length distribution of non-coding regions in *E. coli* (Lukashin und Borodovsky, 1998)

Complex HMMs, recap.

In GHMM:

- First determined how long model stays in state *a* (according to given distribution)
- Then emissions from *a* generated.

Disadvantage of explicit length distribution:

Longer running time for decoding algorithms (Viterbi, Forward, Backward)

Complex HMMs

Time complexity:

To calculate Viterbi variable $v_k(i)$ (probability of $x_1, \ldots, x_i$ emitted ending in state Z_j):

Consider *all* possible time durations to stay in Z_j.

For sequence of length L and fixed number of states in HMM: $O(L^2)$ running time, instead of $O(L)$.

Complex HMMs

Improved methods for gene prediction in prokaryotes:

- *GeneMark* (or *GeneMark.hmm*), Lukashin und Borodovsky, 1998
- *Glimmer*, Delcher *et al.*, 1999

Predicting genes on both strands of the DNA

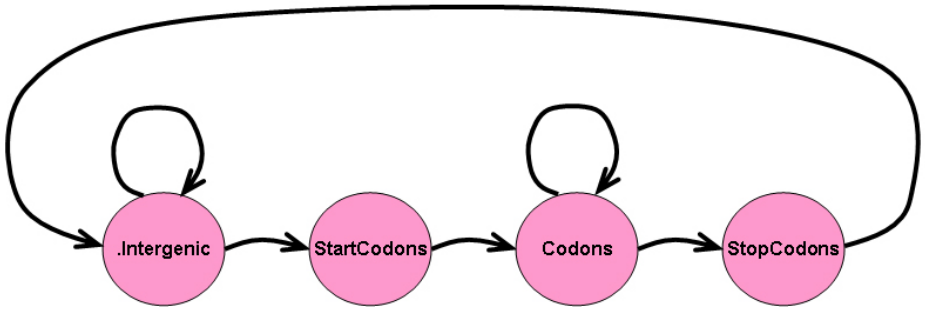


Figure: Basic model for gene finding in prokaryotes on *one* DNA strand (source: `http://cs.wellesley.edu/`)

Predicting genes on both strands of the DNA

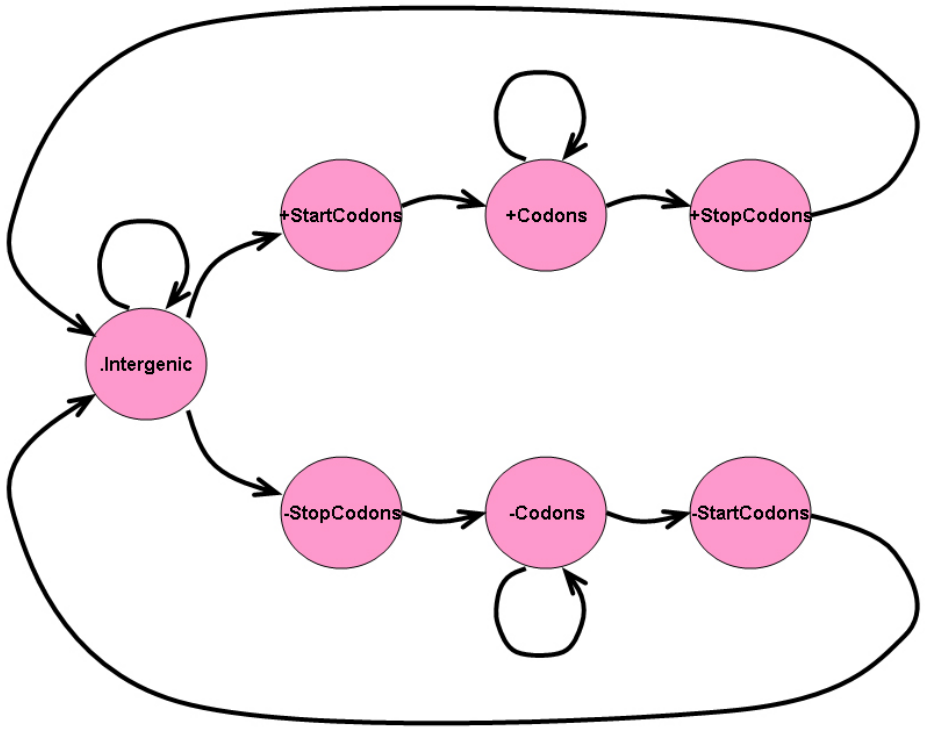


Figure: Predicting genes on *both* DNA strands
(source: `http://cs.wellesley.edu/`)

Gene finding in Eukaryotes

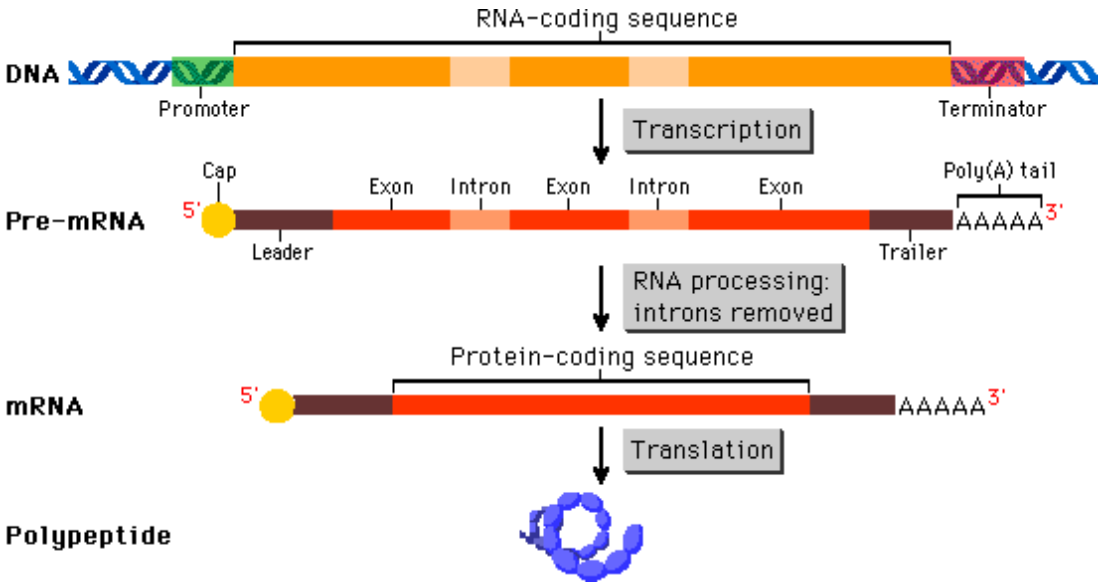


Figure: Protein-coding regions in gene interrupted by *introns*

Gene finding in Eukaryotes

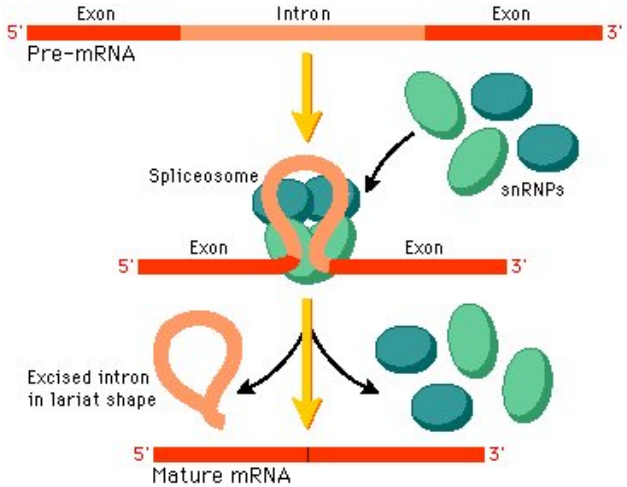


Figure: Intron splicing (source: `www.phschool.com/`)

Gene finding in Eukaryotes

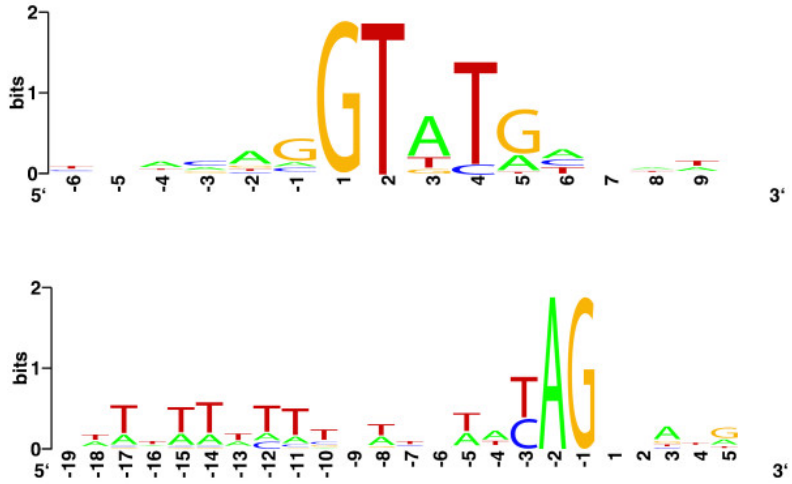


Figure: Donor and acceptor splice signals in eukaryotic genes (Slamovits and Keeling, 2006, *BMC Evolutionary Biology*) 6:34

Gene finding in Eukaryotes

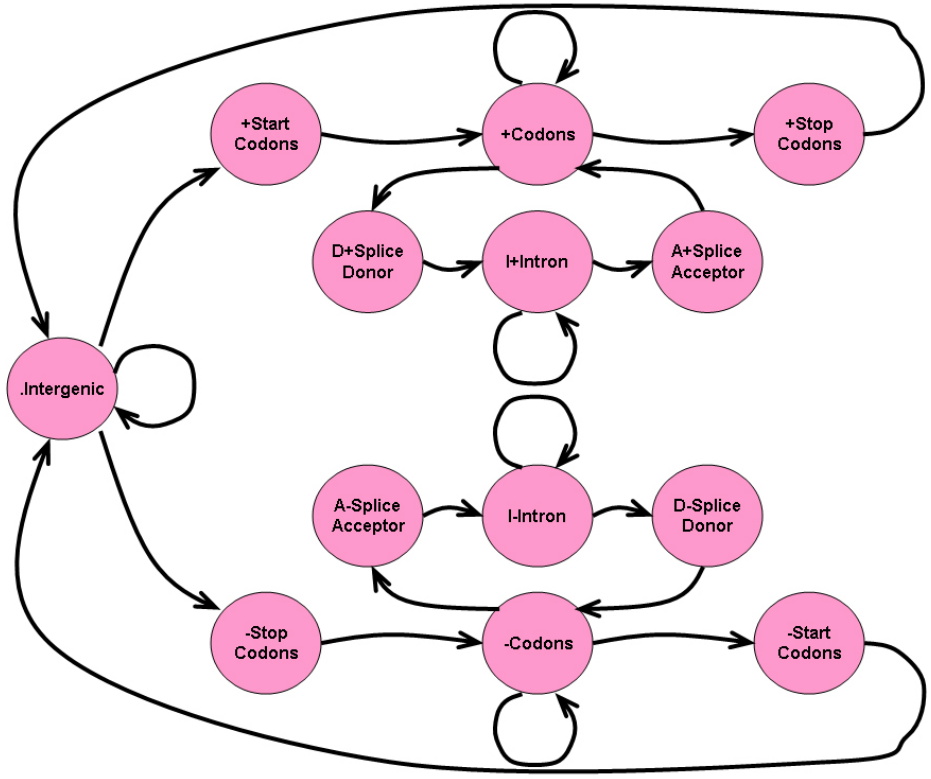


Figure: Predicting genes in eukaryotes
(source: `http://cs.wellesley.edu/`)

Gene finding in Eukaryotes

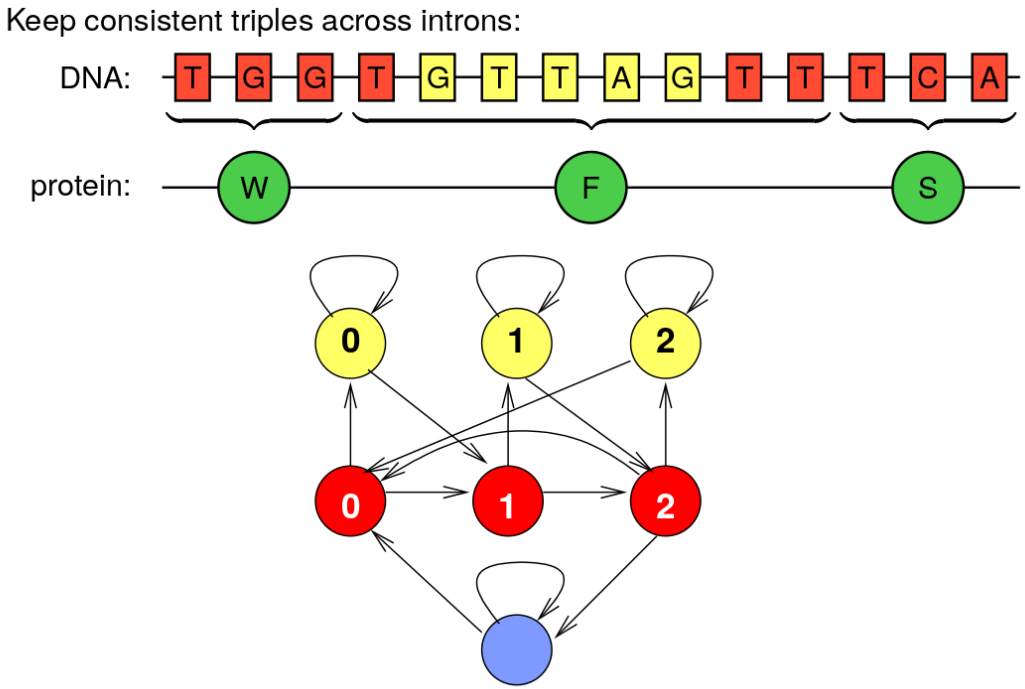


Figure: Codons can be interrupted by introns! (Broňa Brejová)

Gene finding in Eukaryotes

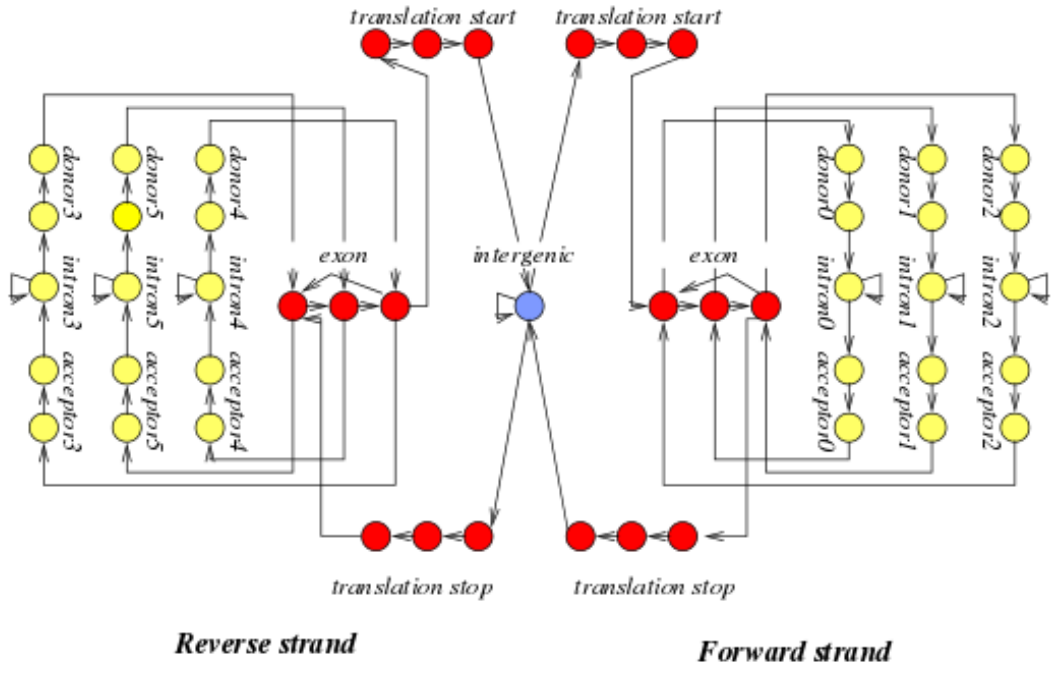


Figure: Three sub-models for introns to encode *phase* in which intron is inserted. (Broňa Brejová)

Gene finding in Eukaryotes

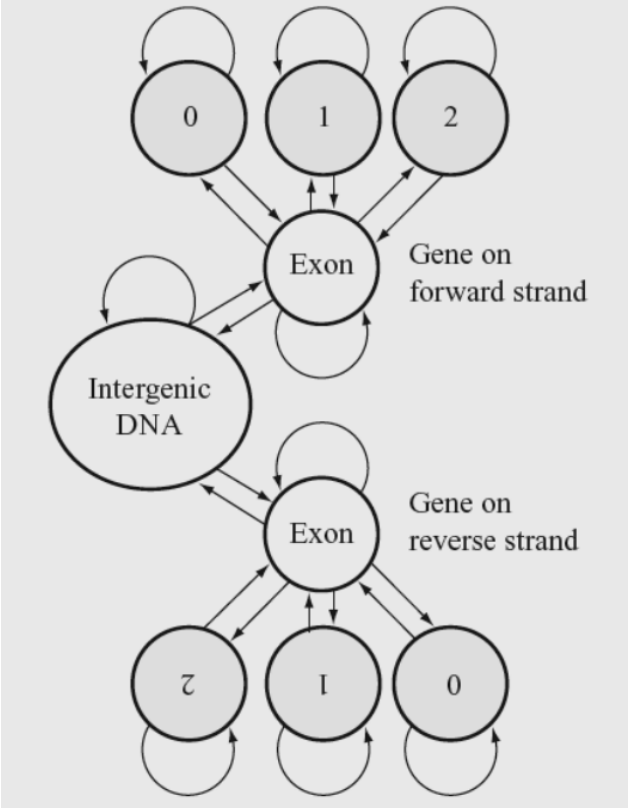
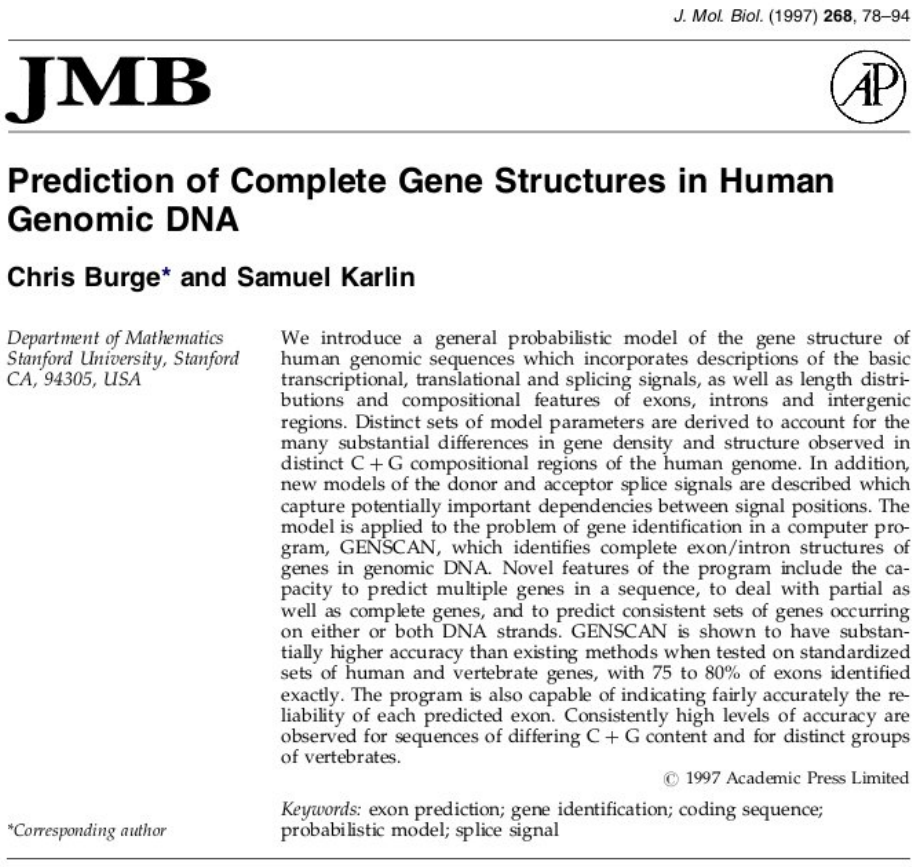


Figure: Three sub-models for introns to encode *phase* in which intron is inserted. (source: `http://www.ece.drexel.edu/`)

Genscan



J. Mol. Biol. (1997) 268, 78–94

JMB

Prediction of Complete Gene Structures in Human Genomic DNA

Chris Burge* and Samuel Karlin

Department of Mathematics Stanford University, Stanford CA, 94305, USA

We introduce a general probabilistic model of the gene structure of human genomic sequences which incorporates descriptions of the basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic regions. Distinct sets of model parameters are derived to account for the many substantial differences in gene density and structure observed in distinct C + G compositional regions of the human genome. In addition, new models of the donor and acceptor splice signals are described which capture potentially important dependencies between signal positions. The model is applied to the problem of gene identification in a computer program, GENSCAN, which identifies complete exon/intron structures of genes in genomic DNA. Novel features of the program include the capacity to predict multiple genes in a sequence, to deal with partial as well as complete genes, and to predict consistent sets of genes occurring on either or both DNA strands. GENSCAN is shown to have substantially higher accuracy than existing methods when tested on standardized sets of human and vertebrate genes, with 75 to 80% of exons identified exactly. The program is also capable of indicating fairly accurately the reliability of each predicted exon. Consistently high levels of accuracy are observed for sequences of differing C + G content and for distinct groups of vertebrates.

Keywords: exon prediction; gene identification; coding sequence; probabilistic model; splice signal

*Corresponding author

Figure: Gene-finding program *GenScan* (Karlin and Burge, 1997)

Genscan

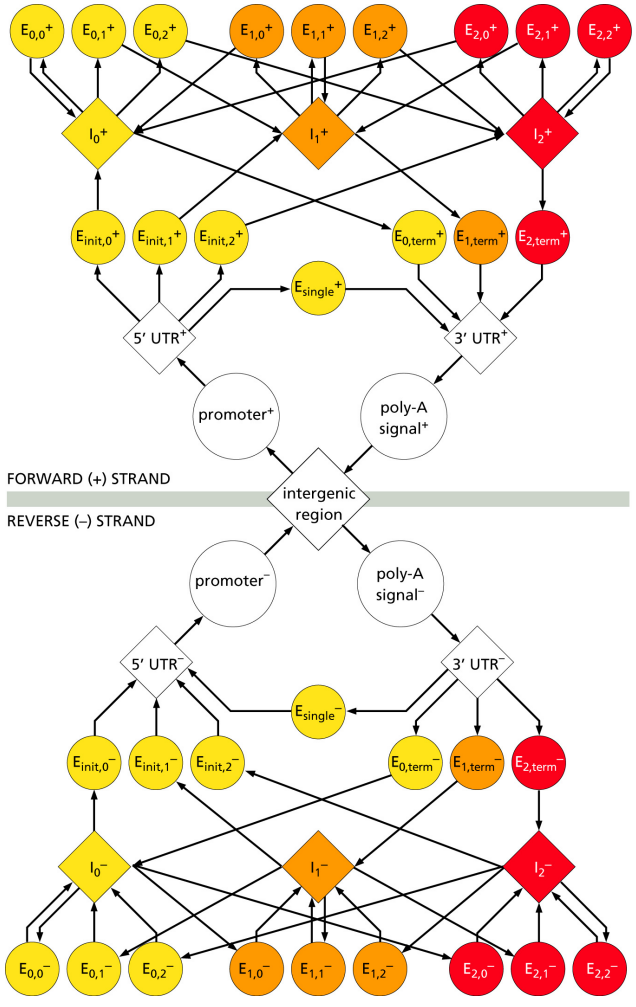
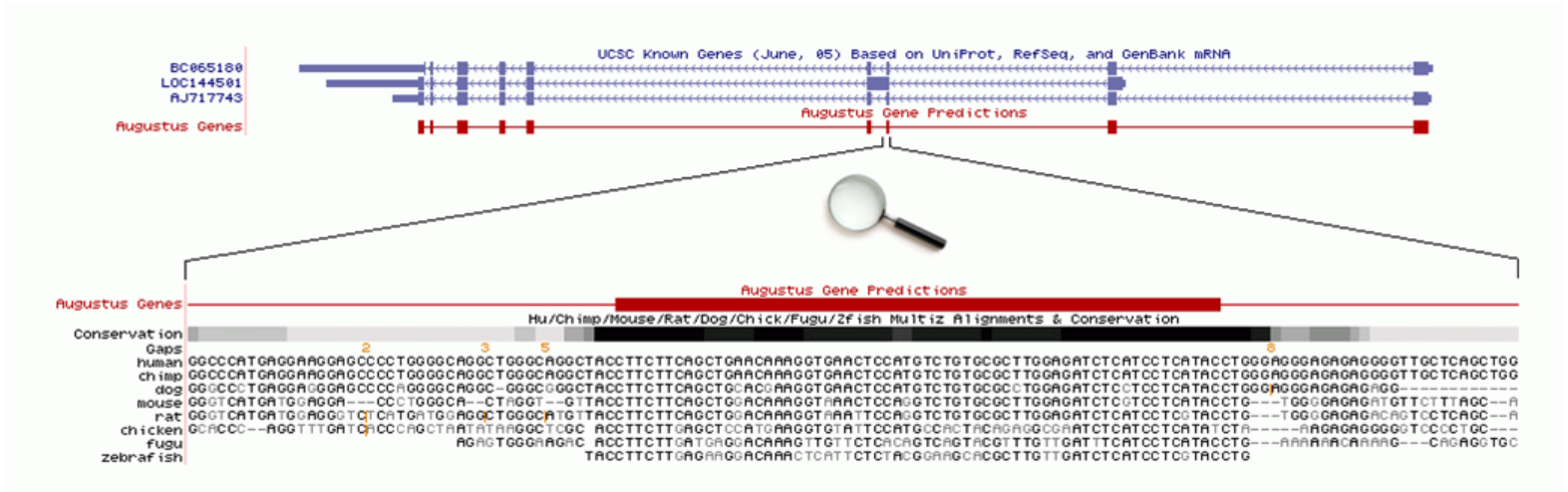


Figure: Topology of *GenScan* (Karlin and Burge, 1997)

Gene Finding by Comparative Sequence Analysis



Protein-kodierende Bereiche im Genom stärker konserviert als nicht-kodierende Bereiche. Mögliche kodierende Bereiche durch Sequenzalignment gefunden (M. Stanke).

Gene Finding by Comparative Sequence Analysis

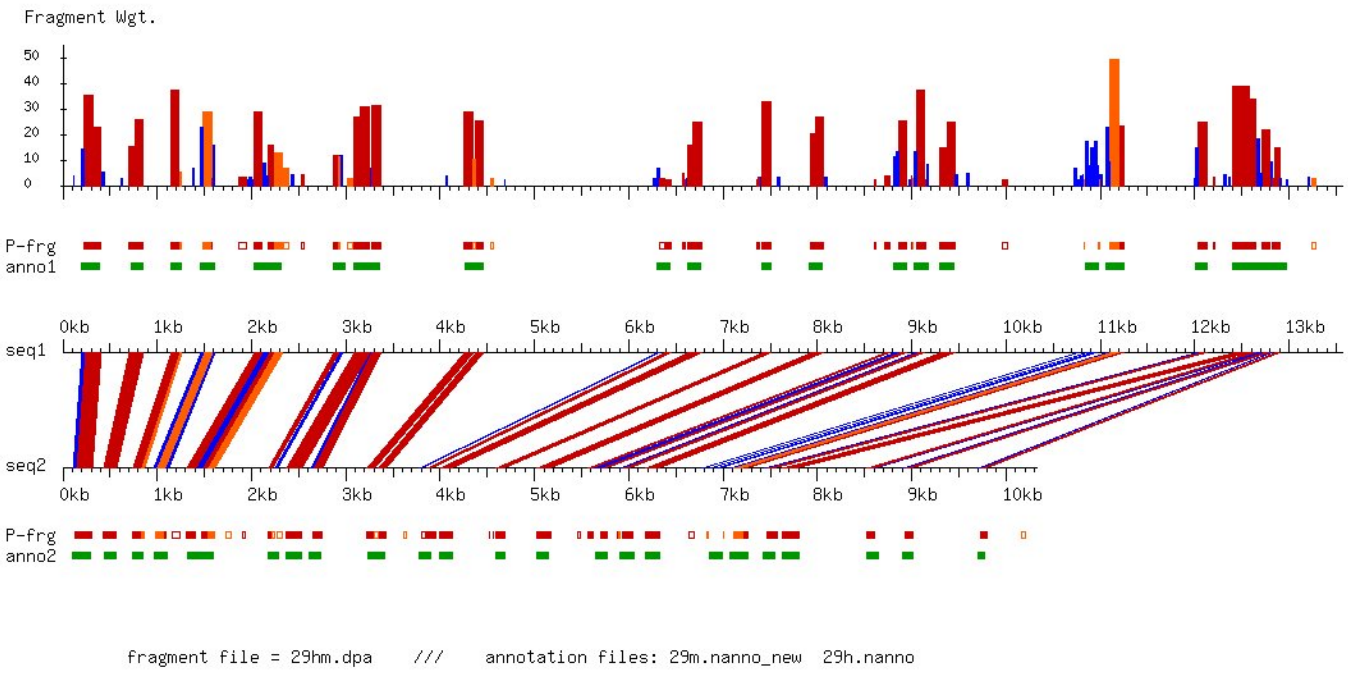


Figure: *DIALIGN* Alignment von genomischen Sequenzen von Mensch und Maus: Bekannte Exons (grün) und gefundene lokale Sequenzähnlichkeiten (rot, orange, blau)

Gene Finding by Comparative Sequence Analysis

AGenDA (Alignment-based Gene Detection Algorithm)

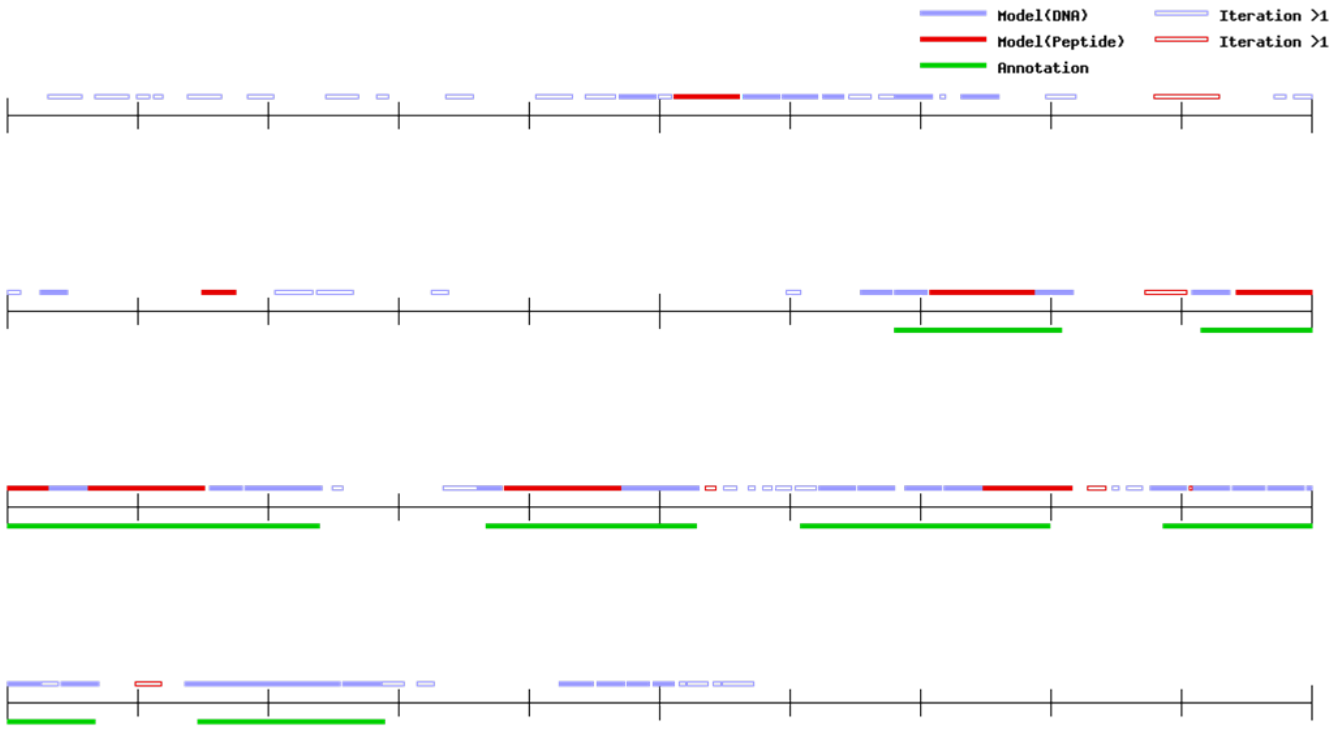
- Finde lokale Homologien zwischen Genomen durch Alignment (*Dialign*)
- Finde konservierte Splice-Stellen bzw. Start/Stop-Codons am Rand der konservierten Sequenzen

Gene Finding by Comparative Sequence Analysis

- Definiere ‘candidate exons’ als Segmente mit starker Ähnlichkeit, begrenzt durch konservierte Splicestellen bzw. Start/Stop-Codons D.h. ‘candidate exons’ können überlappen
- Finde optimale Kette von ‘candidate exons’ mit *Dynamischem Programmieren*

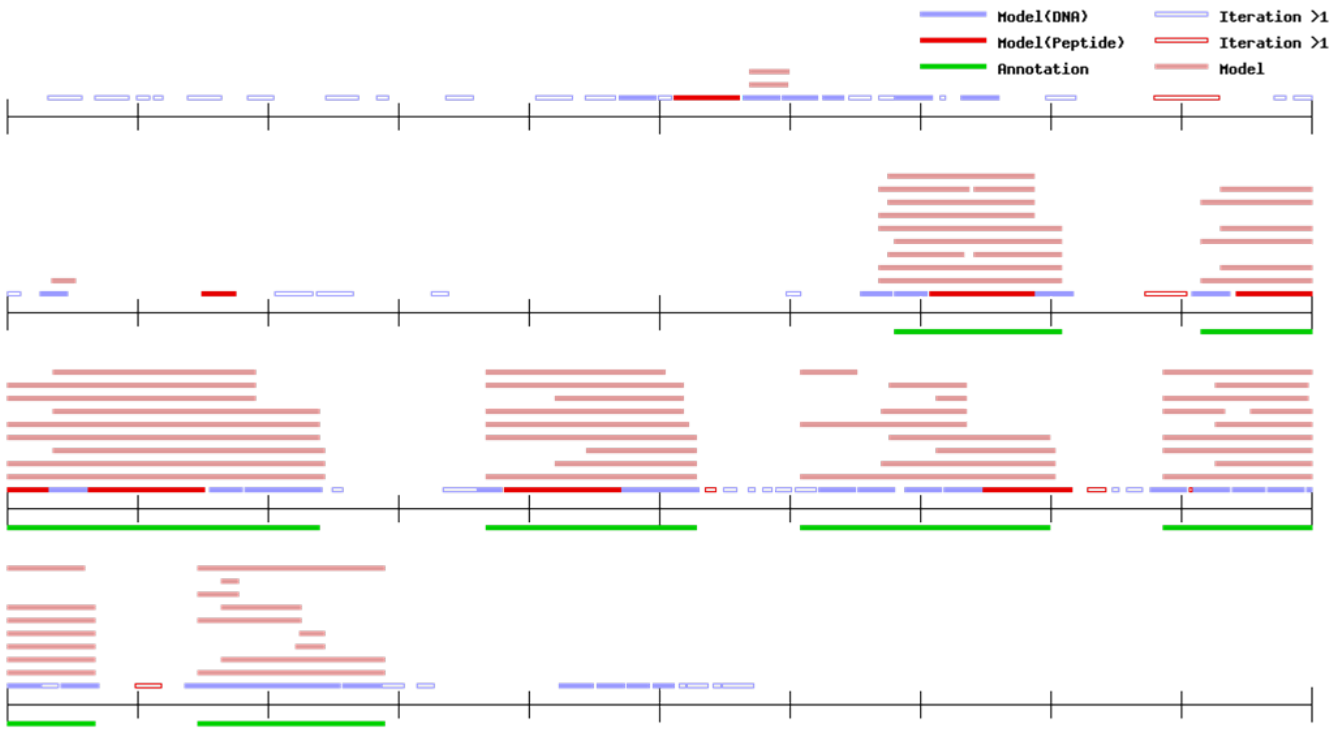
Achtung: Kette von ‘candidate exons’ muss zusätzliche Bedingungen erfüllen bzgl. Reihenfolge von *Splice Sites* und *Start/Stop Codons*. Daher Variante des *DP* Algorithmus für Intervall-Verkettung verwendet.

Gene Finding by Comparative Sequence Analysis



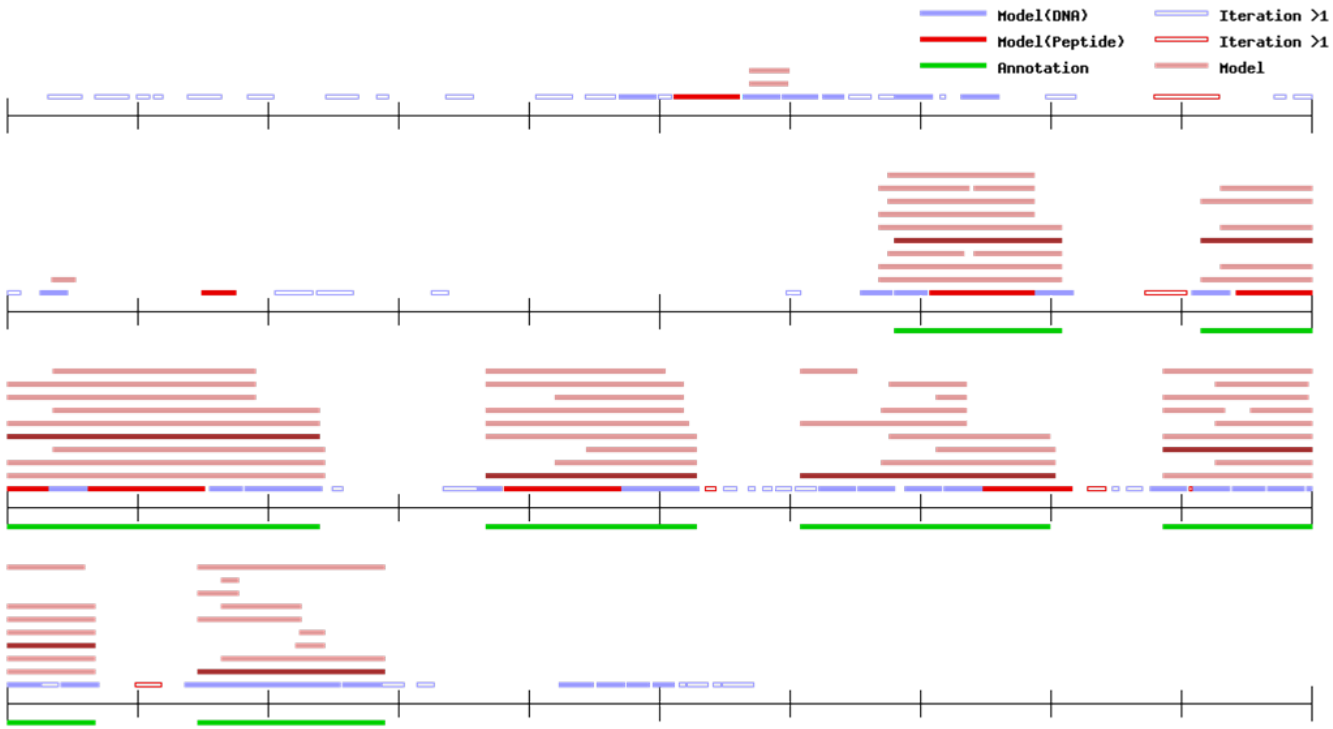
Konservierte Regionen zwischen Genomsequenzen von Mensch und Maus (blau und rot) und bekannte Exons (grün)

Gene Finding by Comparative Sequence Analysis



Potenzielle Exons, vorhergesagt auf Grundlage von Sequenzähnlichkeit und konservierten Splice-Signalen bzw. Start- und Stop-Codons

Gene Finding by Comparative Sequence Analysis



Optimale Kette von potenziellen Exons

Gene Finding by Comparative Sequence Analysis

Evaluierung und Vergleich von Methoden zur Genvorhersage:

- Verwende Genomsequenzen mit *zuverlässig* annotierter Genstruktur
- Wende Methode auf Sequenzen an, vergleiche von Methode vorhergesagte Gene mit annotierten Genen

Qualität der Vorhersage gemessen als *Sensitivität* und *Spezifität*

Gene Finding by Comparative Sequence Analysis

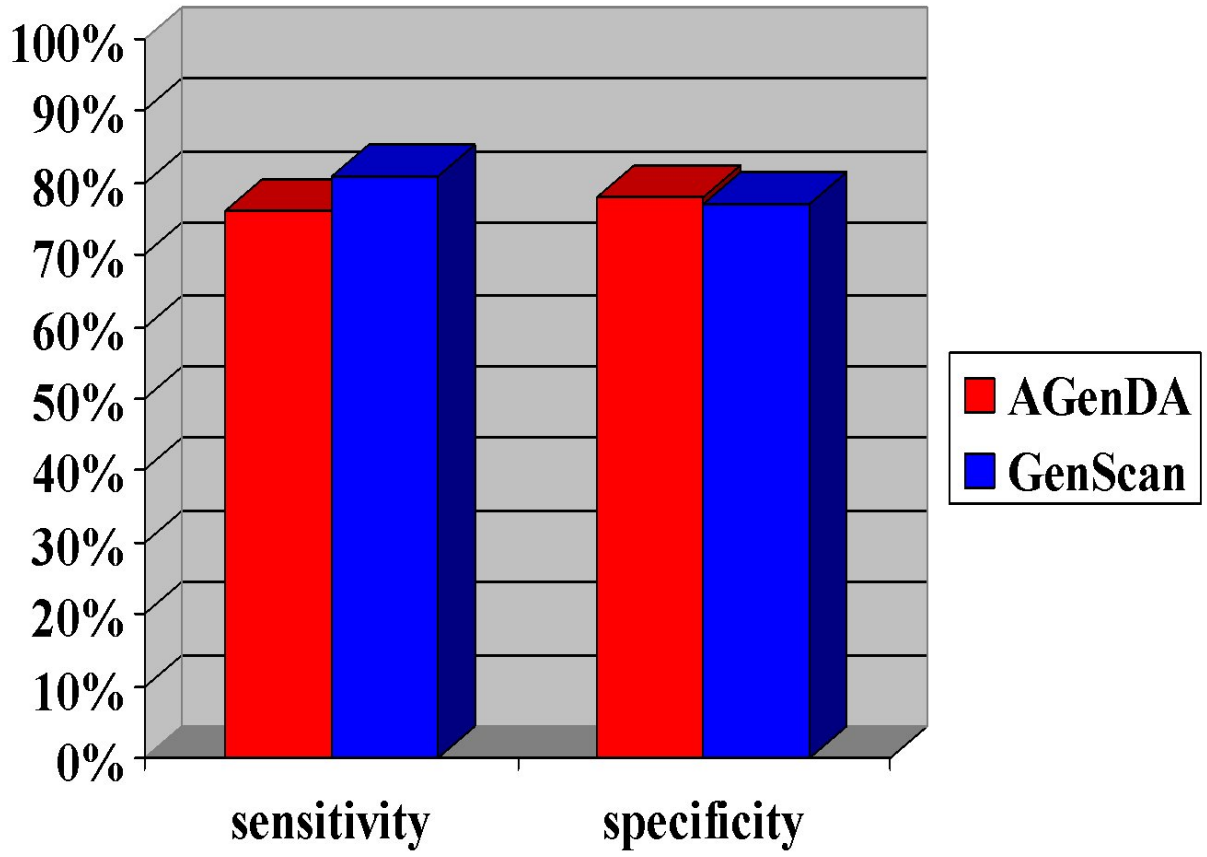


Figure: Benchmark-Resultate von *AGenDA* und *GenScan*

Gene Finding by Comparative Sequence Analysis

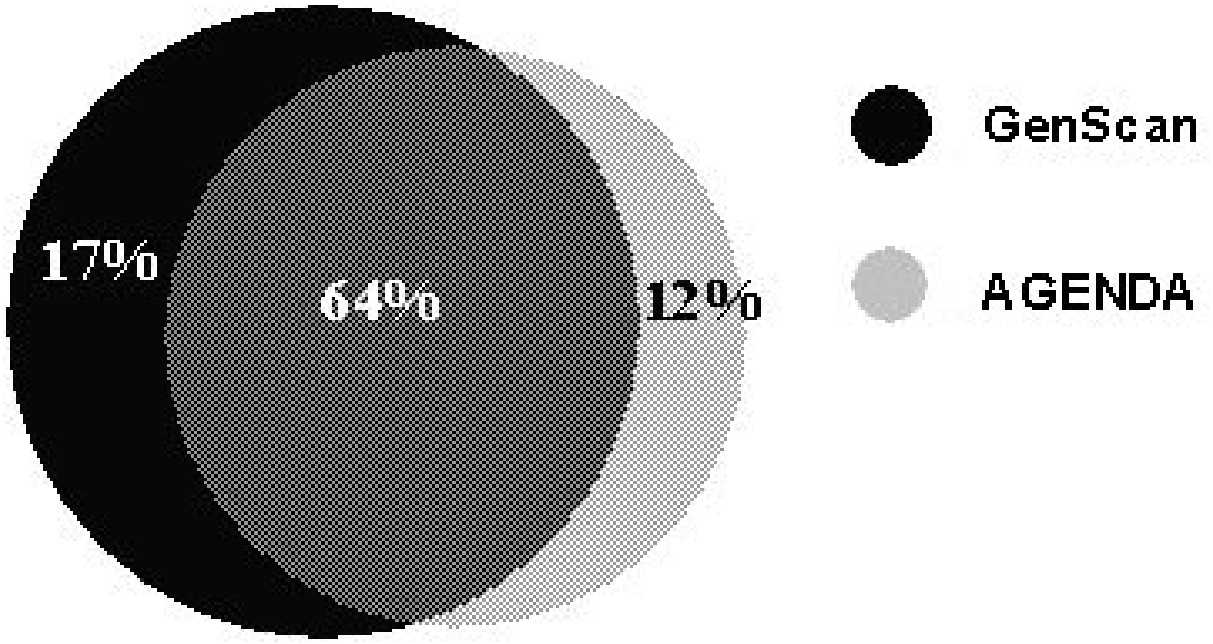


Figure: Benchmark-Resultate von *AGenDA* und *GenScan*

Gene Finding by Comparative Sequence Analysis

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 19 no. 12 2003, pages 1575–1577*
DOI: 10.1093/bioinformatics/btg181

AGenDA: homology-based gene prediction

Leila Taher[1,], Oliver Rinner[2], Saurabh Garg[1], Alexander Sczyrba[3], Michael Brudno[4], Serafim Batzoglou[4] and Burkhard Morgenstern[1,5]*

[1]International Graduate School for Bioinformatics and Genome Research, University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany, [2]GSF Research Center, MIPS / Institute of Bioinformatics, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany, [3]Faculty of Technology, Research Group in Practical Computer Science, University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany, [4]Computer Science Department, Stanford University, Stanford, CA 94305, USA and [5]University of Göttingen, Institute of Microbiology and Genetics, Goldschmidtstr. 1, 37077 Göttingen, Germany

Received on August 19, 2002; revised on November 30, 2002; accepted on February 14, 2003

ABSTRACT

Summary: We present a www server for homology-based gene prediction. The user enters a pair of evolutionary related genomic sequences, for example from human and mouse. Our software system uses CHAOS and DIALIGN to calculate an alignment of the input sequences and then searches for conserved splicing signals and start/stop codons around regions of local sequence similarity. This way, candidate exons are identified that are used, in turn, to calculate optimal gene models. The server returns the constructed gene model by email, together with a graphical representation of the underlying genomic alignment.

Availability: http://bibiserv.TechFak.Uni-Bielefeld.DE/agenda/

Contact: ltaher@TechFak.Uni-Bielefeld.DE

such as genes and regulatory sites tend to be more conserved than non-functional sequences; therefore, local sequence simliarity usually indicates biological function. One problem with traditional gene-prediction approaches is that they rely heavily on information derived from already known genes of the same or a closely related species. Thus, they succeed only where such information is available, and they are unable to detect genes with different properties. By contrast, the new comparative approaches rely more on sequence conservation and less on features of previously-known genes, and therefore are more likely to identify genes with new features and different statistical composition.

Rinner and Morgenstern (2002) recently proposed a homology-based gene-finding program called AGenDA (Alignment-based Gene-Detection Algorithm). The pro-

Figure: Veröffentlichung über AGenDA

Gene Finding by Comparative Sequence Analysis

BIOGNOSYS
PROTEOMICS – NEXT GENERATION

About us | Technology | Products | Contract Research | News Room | Online Sh
Our Team | Career | Contact | Legal info

The People Behind Biognosys

Management

Dr. Oliver Rinner, Founder, CEO, board director

Oliver graduated at the University of Tübingen in biochemistry and psychology , and then received his PhD in 2005 for his work in the field of psychophysics and molecular genetics from the University of Zurich. He joined the group of Ruedi Aebersold at the ETH Zurich as a PostDoc, where he published key papers and patents in the field of targeted proteomics and founded Biognosys in 2008.

AUGUSTUS

BIOINFORMATICS

Vol. 19 Suppl. 2 2003, pages ii215–ii225
DOI: 10.1093/bioinformatics/btg1080

Gene prediction with a hidden Markov model and a new intron submodel

Mario Stanke [1,] and Stephan Waack [2]*

[1]Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Universität Göttingen, Goldschmidtstraße 1, Göttingen, 37077, Germany and [2]Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Lotzestraße 16-18, Göttingen, 37083, Germany

Received on March 17, 2003; accepted on June 9, 2003

AUGUSTUS

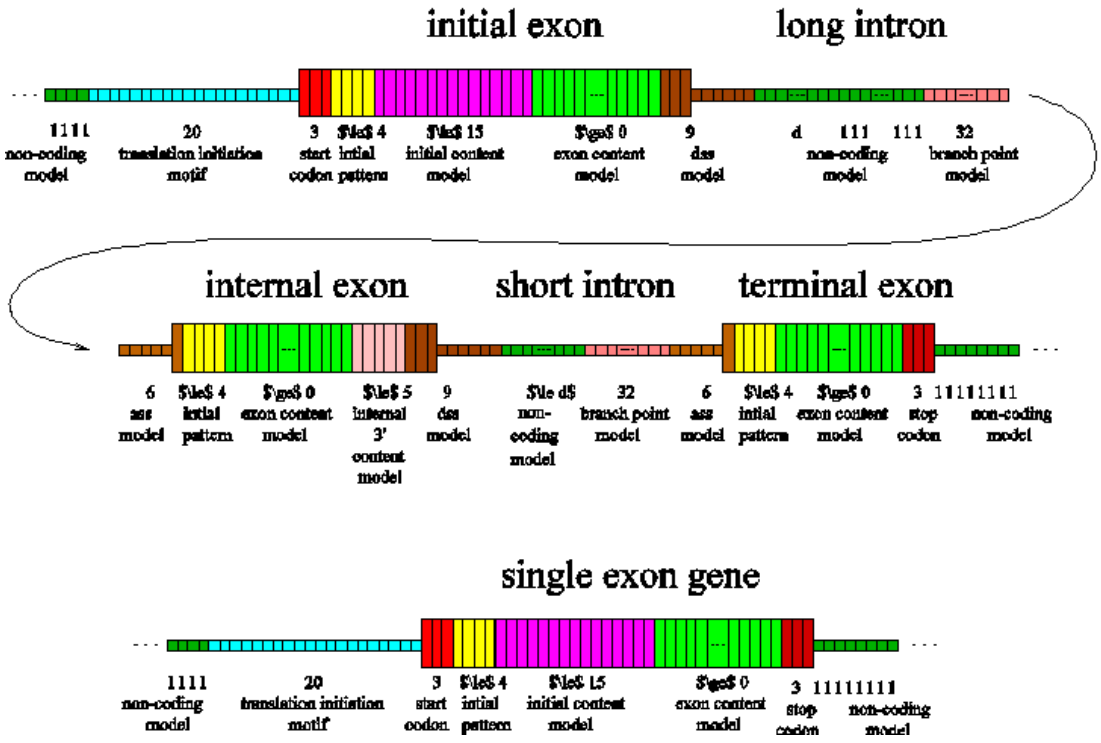


Figure: Hidden-Markov-Model for statistical composition of sub-structures of genes and intergenic regions

AUGUSTUS

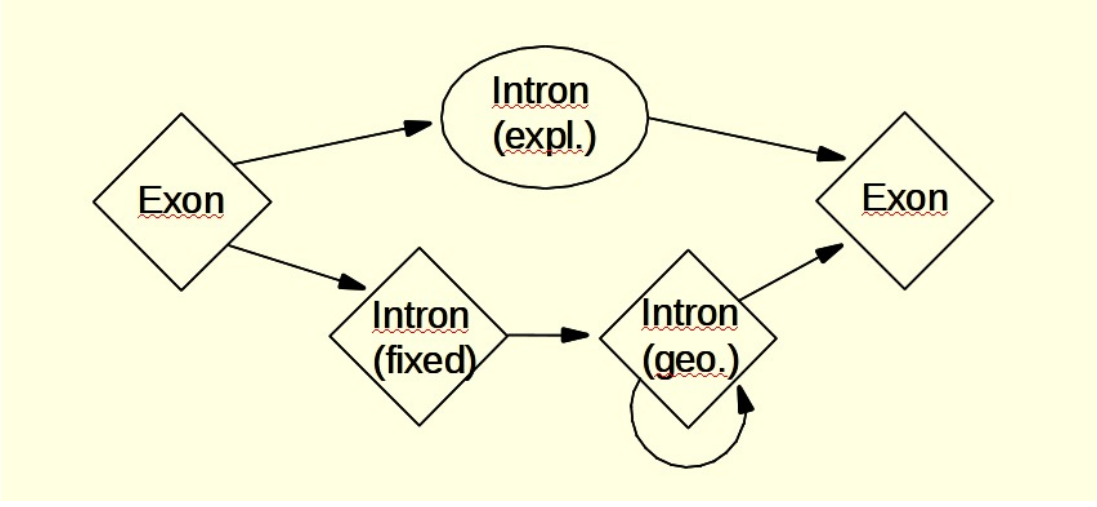


Figure: Sub-model for introns with explicit length modelling for short introns (generalized HMM) and implicit length modelling (geometric distribution) for long introns

AUGUSTUS

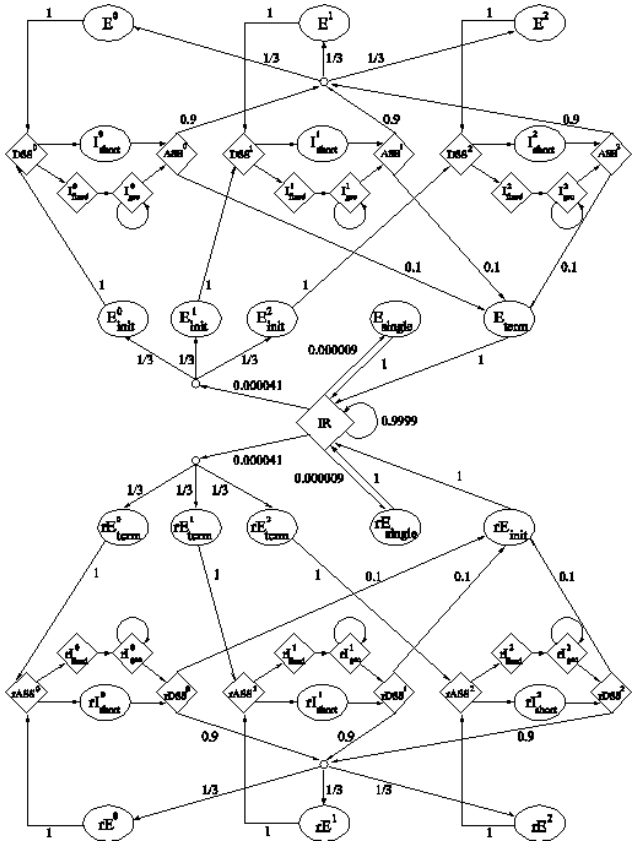


Figure: Hidden-Markov-Model for gene finding in eukaryotes. *States* of model (circles, diamonds) *emit* symbols (nucleotides)

AUGUSTUS

Features of AUGUSTUS:

- Intron length model
- Initial pattern for exons
- Similarity-based weighting for splice sites
- Interpolated HMM

AUGUSTUS

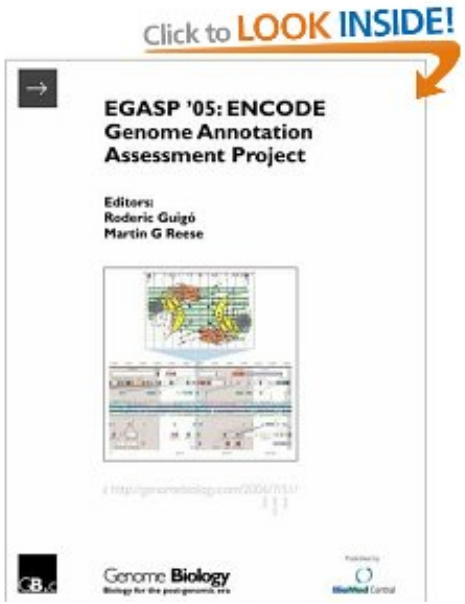


Figure: First systematic evaluation of gene-finding programs at *EGASP* workshop at *Sanger Center* (Genome Biology, special issue 2006).

Stanke, Zvetkova, Morgenstern (2006) *Genome Biology* 7, S11

AUGUSTUS

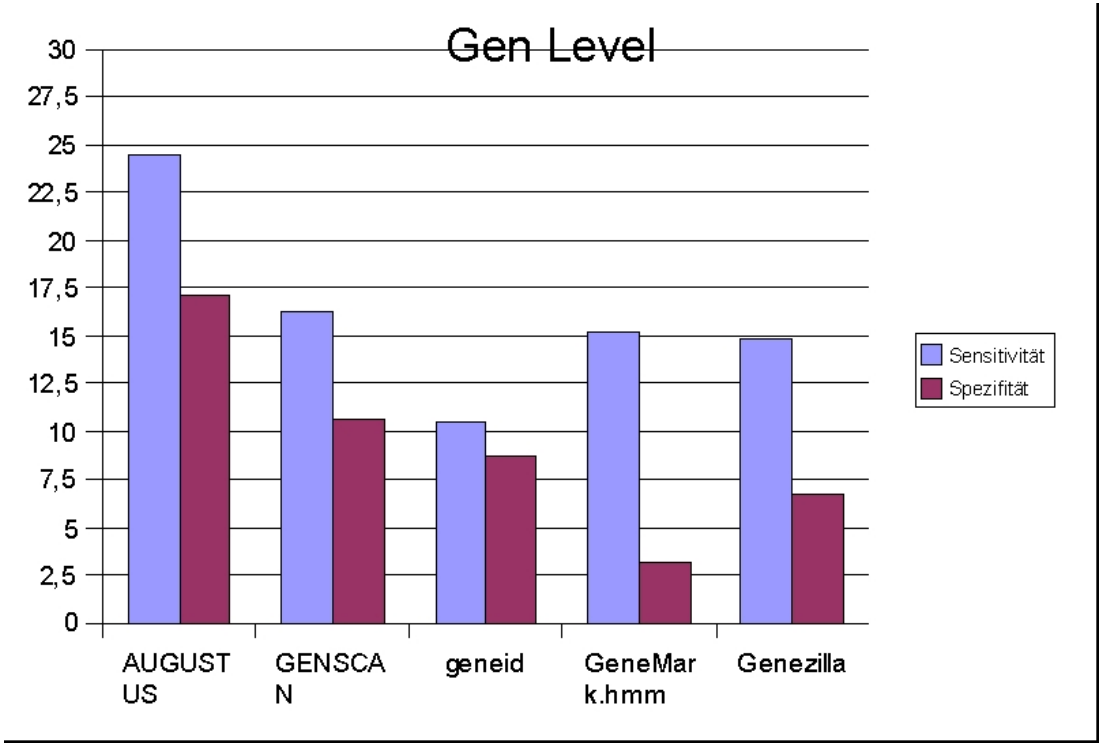


Figure: Result at EGASP: AUGUSTUS best method in the category of *intrinsic* gene-finding methods.

AUGUSTUS plus

Combination of *intrinsic* and *extrinsic* gene finding in one HMM.

Use external *hints* to support gene finding:

- Transcriptome sequences
- Matches to protein sequences
- Cross-species sequence alignments
- User-defined hints

AUGUSTUS plus

Model considers:

- Different *types* of hints pointing to full exons, partial exons, splice sites *etc.*
- Different *grades* for hints, depending on source of information

$h_{i,t}$ Information about hint of type t at position i in sequence:

$$h_{i,t} = \text{(grade,strand[,length,reading frame])}$$

At most one hit of each type t allowed at a position i

AUGUSTUS plus

HMM in AUGUSTUS+ generates gene structure (path) ϕ, sequence of nucleotides S *and* set of hints h with *joint probability*

$$P(\phi, S, h)$$

Goal: maximize

$$P(\phi|S, h),$$

equivalent to maximizing joint probability for given (observed) S and h.

AUGUSTUS plus

Calculate joint probability as

$$\begin{aligned} P(\phi, S, h) &= P(\phi, S) \cdot P(h|\phi, S) \\ &= P(\phi, S) \cdot \prod_{i,t} P(h_{i,t}|\phi, S) \end{aligned}$$

Assumption: hints $h_{i,t}$ independent of each other.

AUGUSTUS plus

Also, assumption that $h_{i,t}$ only depend on t and g (not on i!).

Then for given ϕ, S:

$$P(h_{i,t}|\,\phi, S) \quad = \quad \begin{cases} q^+(t,g) & \text{if } h_{i,t} \text{ is compatible with } \phi; \\ q^-(t,g) & \text{if } h_{i,t} \text{ is compatible with } s \text{ but not with } \phi; \\ 0 & \text{if } h_{i,t} \text{ is not compatible with } S. \end{cases}$$

Values $q^+(t,g), q^-(t,g)$ learned from training data.

AUGUSTUS plus

Results:

- Gene structures supported by hints get higher probabilities.
- Similarly: gene structures *not* supported by hints get *lower* probabilities

Stanke *et al.* (2006) *BMC Bioinformatics* 7, 62

AUGUSTUS plus

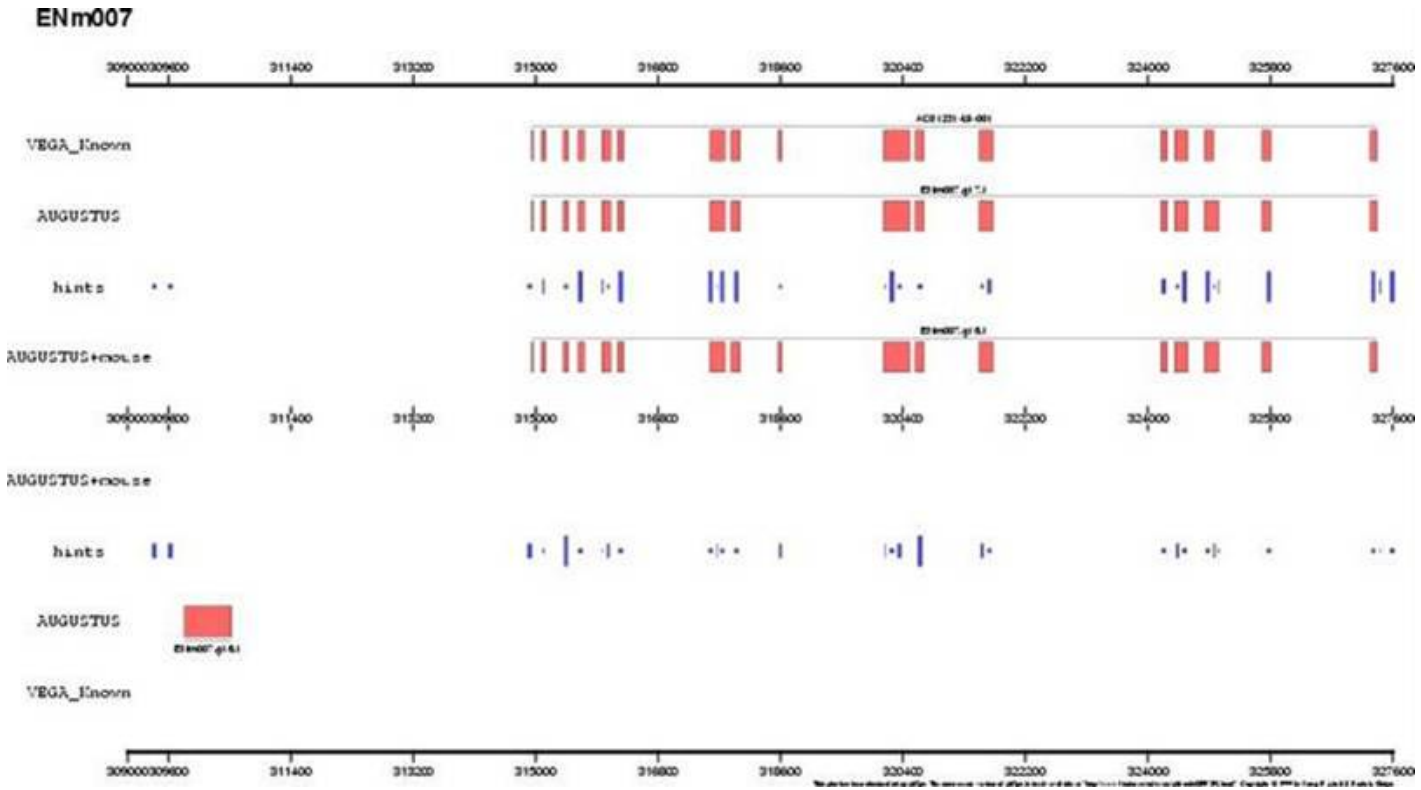


Figure: *AUGUSTUS*: intrinsic gene prediction and gene prediction with hints

AUGUSTUS plus

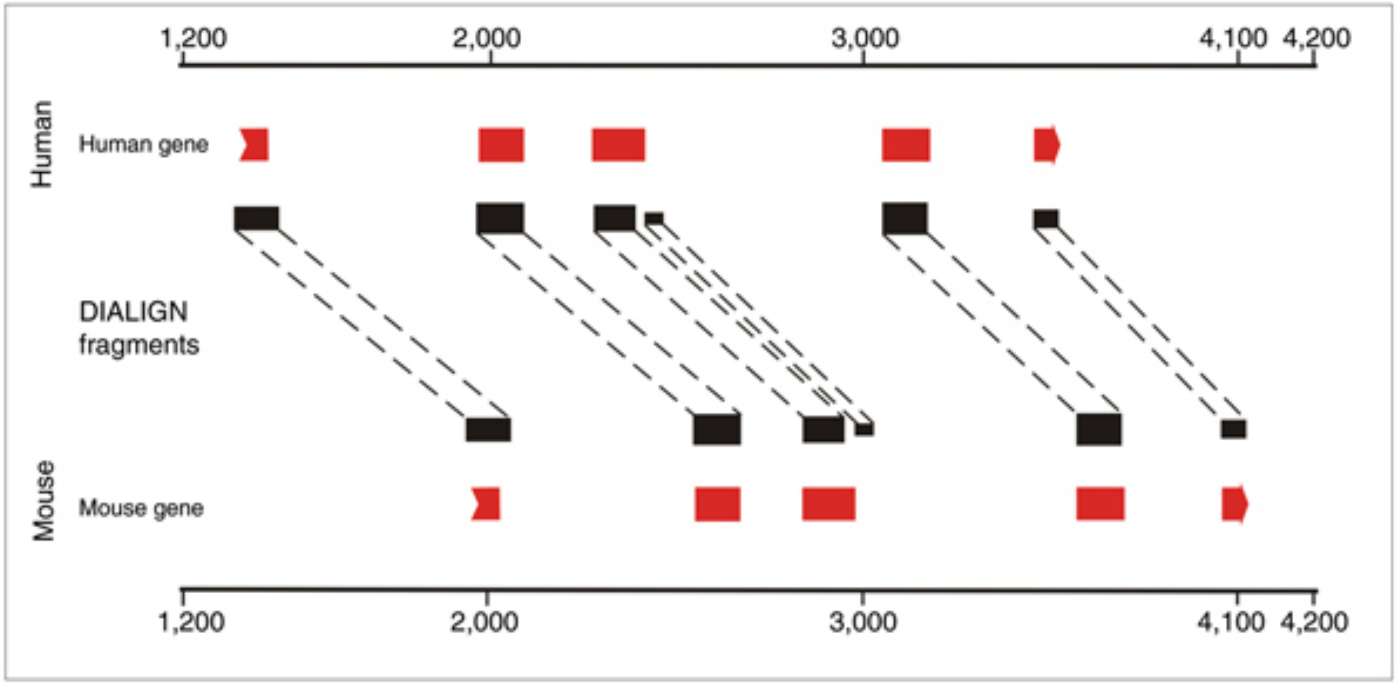


Figure: For EGASP evaluation: hints created using inter-species alignments with *DIALIGN*

Stanke, Tzvetkova, Morgenstern (2006) *Genome Biology* 7, S11

AUGUSTUS plus

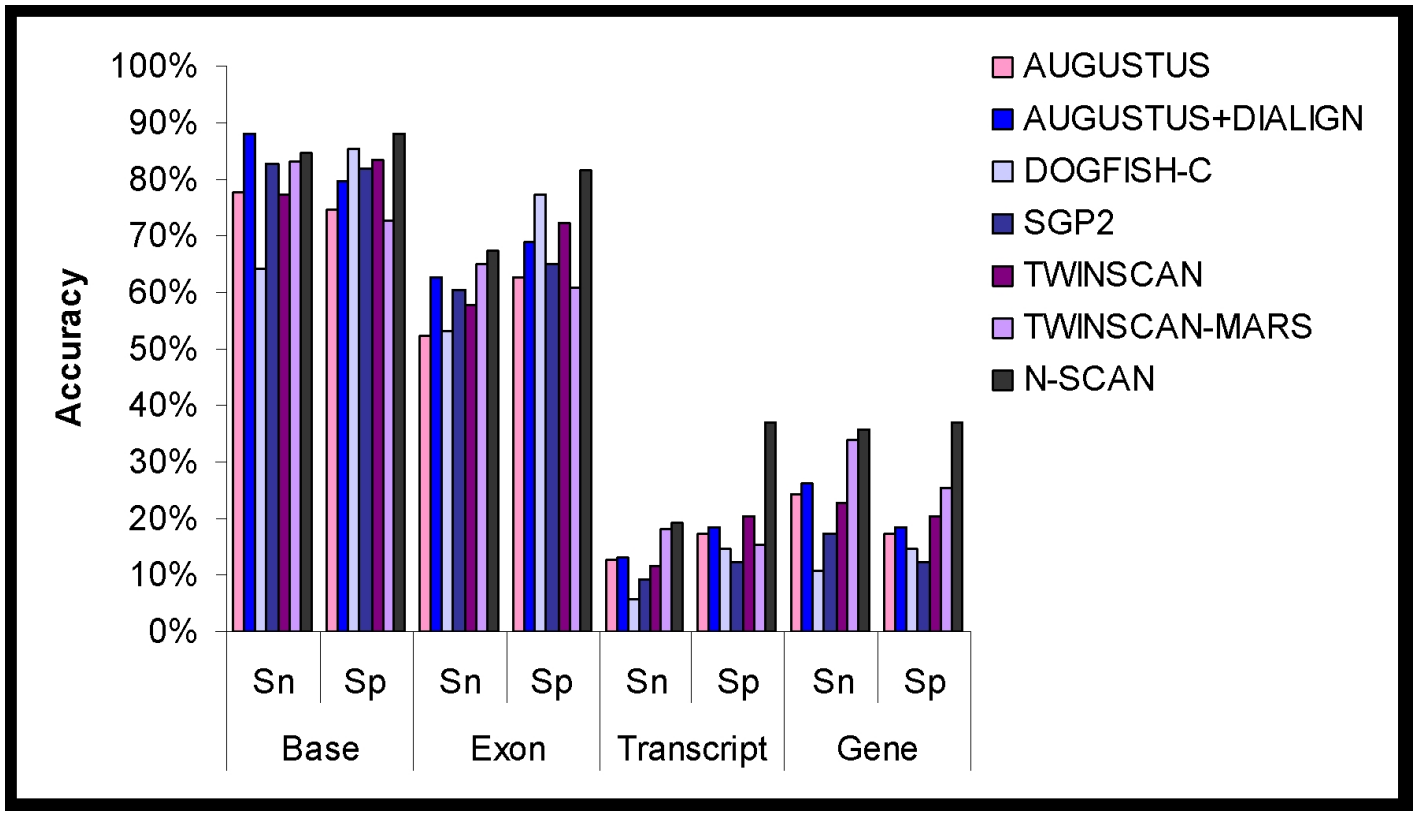


Figure: EGASP evaluation results, AUGUSTUS with *hints* created by *DIALIGN*.

AUGUSTUS plus

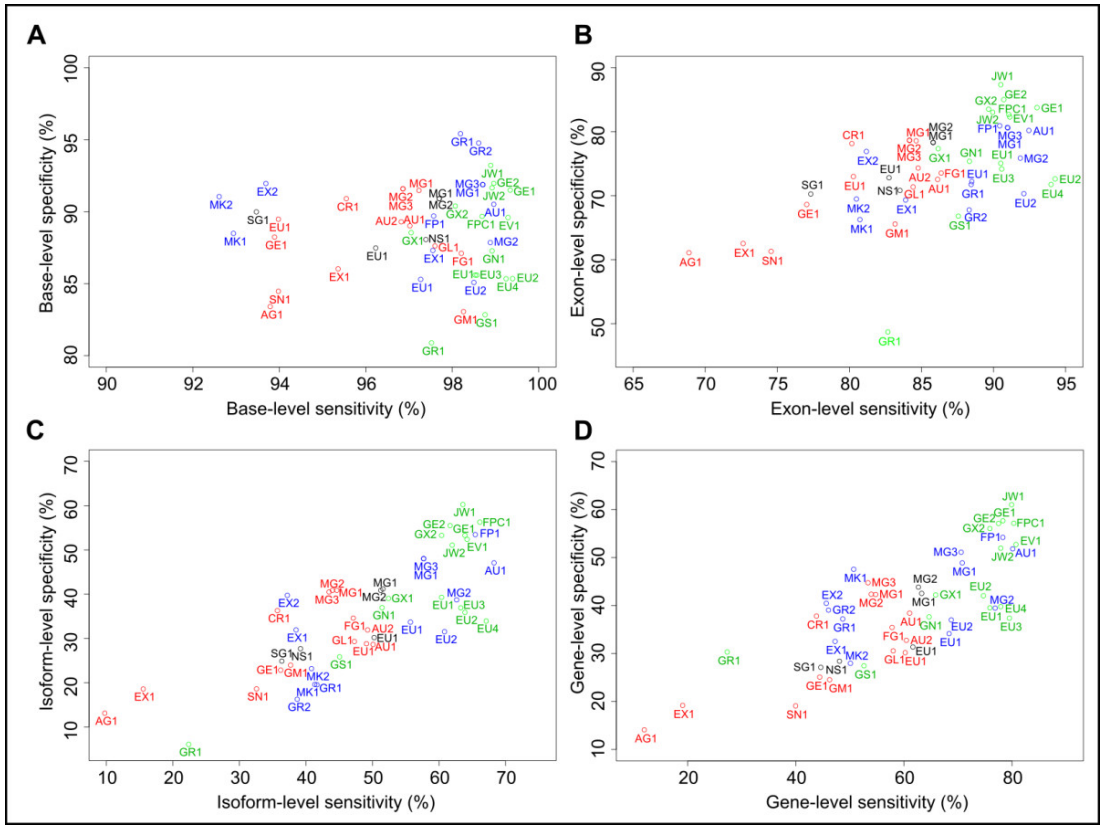


Figure: nGASP evaluation results: ab-initio prediction (red), with genome alignments (black), with transcripts/proteins (blue) and 'combiners' (green)

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