Algorithmen der Bioinformatik I
WS 2017/2018

Burkhard Morgenstern
Peter Meinicke

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

January 8, 2018

Gene finding

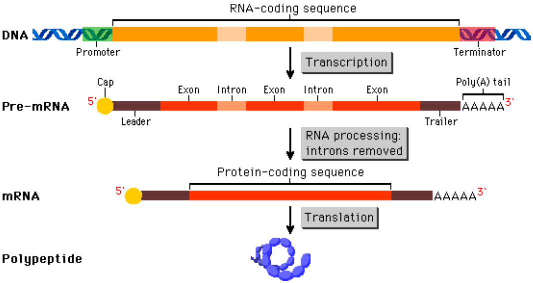


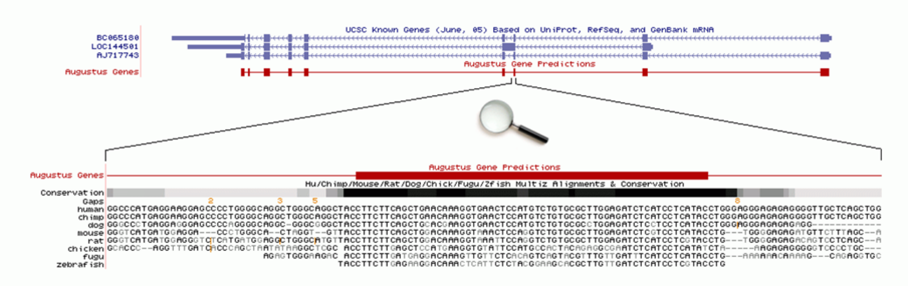
Figure: First step in genome analysis: computational prediction of gene structures. In *eukaryotes*: coding regions (*exons*) separated by *introns*.

Gene finding

Sources of information for gene finding:

- Intrinsic:
  - Short signals: Start/stop codons, splice sites
  - Statistical properties of genome components: Hidden-Markov-Models
- Extrinsic:
  - Comparison to known genes/proteins
  - Transcriptomics sequences
  - Comparative genome analysis: Alignment of genomic sequences

Gene finding


UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
BC065180
LOC144501
AJ717743
Augustus Gene Predictions
Augustus Genes
Augustus Gene Predictions
Augustus Genes
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
Gaps
2
3
5
8

```
    human GGCCCATGAGGAAGGAGCCCCTGGGGCAGGCTGGGCAGGCTACCTTCTTCAGCTGAACAAAGGTGAACTCCATGTCTGTGCGCTTGGAGATCTCATCCTCATACCTGGGAGGGAGAGAGGGGTTGCTCAGCTGG
    chimp GGCCCATGAGGAAGGAGCCCCTGGGGCAGGCTGGGCAGGCTACCTTCTTCAGCTGAACAAAGGTGAACTCCATGTCTGTGCGCTTGGAGATCTCATCCTCATACCTGGGAGGGAGAGAGGGGTTGCTCAGCTGG
      dog GGGCCCTGAGGAGGGAGCCCCAGGGGCAGGC-GGGCGGGCTACCTTCTTCAGCTGCACGAAGGTGAACTCCATGTCTGTGCGCCTGGAGATCTCCTCCTCATACCTGGGAGGGAGAGAGAGG------------
    mouse GGGTCATGATGGAGGA---CCCTGGGCA--CTAGGT--GTTACCTTCTTCAGCTGGACAAAGGTAAACTCCAGGTCTGTGCGCTTGGAGATCTCGTCCTCATACCTG---TGGGGAGAGATGTTCTTTAGC--A
      rat GGGTCATGATGGAGGGTCTCATGATGGAGGCTGGGCATGTTACCTTCTTCAGCTGGACAAAGGTAAATTCCAGGTCTGTGCGCTTGGAGATCTCATCCTCGTACCTG---TGGGGAGAGACAGTCCTCAGC--A
  chicken GCACCC--AGGTTTGATCACCCAGCTAATATAAGGCTCGC ACCTTCTTGAGCTCCATGAAGGTGTATTCCATGCCACTACAGAGGCGAATCTCATCCTCATATCTA-----AAGAGAGGGGGTCCCCTGC---
     fugu                            AGAGTGGGAAGAC ACCTTCTTGATGAGGACAAAGTTGTTCTCACAGTCAGTACGTTTGTTGATTTCATCCTCATACCTG---AAAAAACAAAAG---CAGAGGTGC
zebrafish                                          TACCTTCTTGAGAAGGACAAACTCATTCTCTACGGAAGCACGCTTGTTGATCTCATCCTCGTACCTG
```

Figure: Gene finding by comparative sequence analysis: exons more conserved in genome than non-coding regions

Gene finding

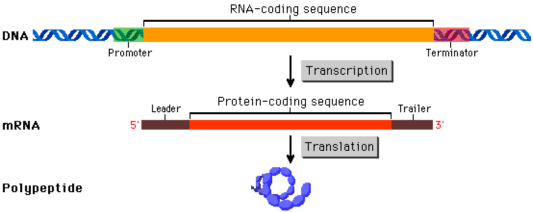


Figure: In *prokaryotes*: coding regions not interrupted by introns.

Hidden Markov Models (recap.)

HMM consists of

- States
- Possible transitions between states
- Emission of characters from states.

HMM generates:

- Path π through states
- Sequence S 'emitted' by states

Hidden Markov Models (recap.)

Parameters: probabilities

- for ‘emitting’ characters from states
- for transitions to next state

Wanted: path π maximizing *conditional* probability

$$P(\pi|S)$$

given observed sequence S of observations.

Hidden Markov Models (recap.)

By definition:

$$P(\pi|S) = \frac{P(\pi, S)}{P(S)}$$

For constant (observed) sequence S: maximize $P(\pi, S)$

$$P(\pi, S) = P(S|\pi) \cdot P(\pi)$$

Probabilities $P(S|\pi)$ and $P(\pi)$ easy to calculate as product of emission and transition probabilities!

Hidden Markov Models (recap.)

Trade-off between simpler and more complex HMMs:

- Complex models
  - use more information
  - more accurate, if enough training data available
- Simple models
  - easy to understand / develop
  - usually faster decoding
  - need fewer training data

First question if HMM developed: *what kind of information* is used?

Simple models for prokaryotes

Most basic model for gene finding uses *frequencies* of nucleotides in coding regions (‘exons’) and non-coding regions (‘introns’) and length of coding/non-coding regions.

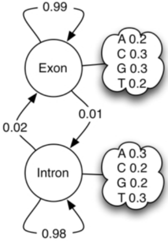


Figure: Two-state HMM for gene prediction, analogous to *Casino* model (source: Ian Korf)

Simple models for prokaryotes

More complex model considers probability of nucleotides depending on position in codon.

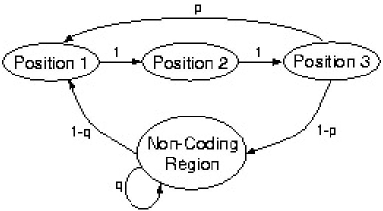


Figure: Four-state HMM for gene prediction in prokaryotes, distinguishes frequencies of nucleotides at different positions in codon (source: stat.berkeley.edu)

Simple models for prokaryotes

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

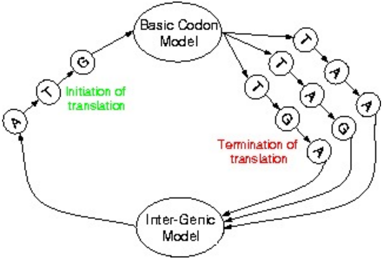
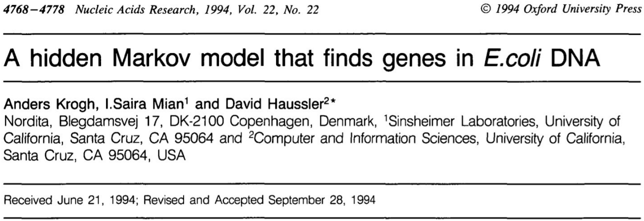


Figure: HMM for gene finding with start and stop codons (source: `stat.berkeley.edu`)

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

EcoParse



4768–4778 Nucleic Acids Research, 1994, Vol. 22, No. 22 © *1994 Oxford University Press*

A hidden Markov model that finds genes in *E.coli* DNA

Anders Krogh, I.Saira Mian[1] and David Haussler[2]*
Nordita, Blegdamsvej 17, DK-2100 Copenhagen, Denmark, [1]Sinsheimer Laboratories, University of California, Santa Cruz, CA 95064 and [2]Computer and Information Sciences, University of California, Santa Cruz, CA 95064, USA

Received June 21, 1994; Revised and Accepted September 28, 1994

EcoParse first HMM-based approach to gene finding in *E. coli*

EcoParse

Most important source of information: codon frequencies.

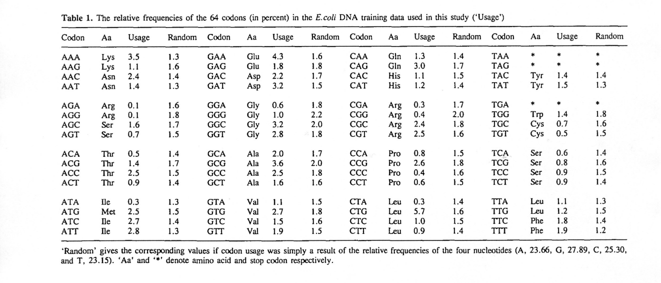


Table 1. The relative frequencies of the 64 codons (in percent) in the *E.coli* DNA training data used in this study ('Usage')

Codon	Aa	Usage	Random	Codon	Aa	Usage	Random	Codon	Aa	Usage	Random	Codon	Aa	Usage	Random
AAA	Lys	3.5	1.3	GAA	Glu	4.3	1.6	CAA	Gln	1.3	1.4	TAA	*	*	*
AAG	Lys	1.1	1.6	GAG	Glu	1.8	1.8	CAG	Gln	3.0	1.7	TAG	*	*	*
AAC	Asn	2.4	1.4	GAC	Asp	2.2	1.7	CAC	His	1.1	1.5	TAC	Tyr	1.4	1.4
AAT	Asn	1.4	1.3	GAT	Asp	3.2	1.5	CAT	His	1.2	1.4	TAT	Tyr	1.5	1.3
AGA	Arg	0.1	1.6	GGA	Gly	0.6	1.8	CGA	Arg	0.3	1.7	TGA	*	*	*
AGG	Arg	0.1	1.8	GGG	Gly	1.0	2.2	CGG	Arg	0.4	2.0	TGG	Trp	1.4	1.8
AGC	Ser	1.6	1.7	GGC	Gly	3.2	2.0	CGC	Arg	2.4	1.8	TGC	Cys	0.7	1.6
AGT	Ser	0.7	1.5	GGT	Gly	2.8	1.8	CGT	Arg	2.5	1.6	TGT	Cys	0.5	1.5
ACA	Thr	0.5	1.4	GCA	Ala	2.0	1.7	CCA	Pro	0.8	1.5	TCA	Ser	0.6	1.4
ACG	Thr	1.4	1.7	GCG	Ala	3.6	2.0	CCG	Pro	2.6	1.8	TCG	Ser	0.8	1.6
ACC	Thr	2.5	1.5	GCC	Ala	2.5	1.8	CCC	Pro	0.4	1.6	TCC	Ser	0.9	1.5
ACT	Thr	0.9	1.4	GCT	Ala	1.6	1.6	CCT	Pro	0.6	1.5	TCT	Ser	0.9	1.4
ATA	Ile	0.3	1.3	GTA	Val	1.1	1.5	CTA	Leu	0.3	1.4	TTA	Leu	1.1	1.3
ATG	Met	2.5	1.5	GTG	Val	2.7	1.8	CTG	Leu	5.7	1.6	TTG	Leu	1.2	1.5
ATC	Ile	2.7	1.4	GTC	Val	1.5	1.6	CTC	Leu	1.0	1.5	TTC	Phe	1.8	1.4
ATT	Ile	2.8	1.3	GTT	Val	1.9	1.5	CTT	Leu	0.9	1.4	TTT	Phe	1.9	1.2

'Random' gives the corresponding values if codon usage was simply a result of the relative frequencies of the four nucleotides (A, 23.66, G, 27.89, C, 25.30, and T, 23.15). 'Aa' and '*' denote amino acid and stop codon respectively.

Figure: Table with frequencies of codons in *E. coli* compared to probabilities of random occurrence (Krogh *et al.*, 1994)

EcoParse

HMM to detect genes in *E.coli* contains:

- Single state generating non-coding sequence
- States (sub-models) generating codons
- States (sub-models) generating start/stop codong
- Allow for insertions and deletions within codons

EcoParse

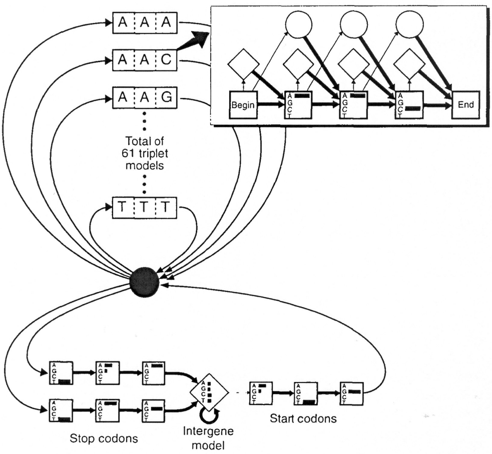


Figure: Structure ('topology') of HMM for gene finding in *EcoParse* (Krogh *et al.*, 1994)

EcoParse

Note:
EcoParse predicts genes only on one strand of the genome.

Thus:
Run program on genome sequence *and* on reverse complement

Complex HMMs

(a) Higher-order HMMs

Until now: probability for transition to current state x_i depends only on previous state x_{i-1}.

$$P(x_i|x_{i-1}, \ldots, x_1) = P(x_i|x_{i-1})$$

Thereby modelled: frequency of *pairs* (x_{i-1}, x_i)

Complex HMMs

Generalization: transition depends on previous *n* states:

$$P(x_i | x_{i-1}, \ldots, x_1) = P(x_i | x_{i-1}, \ldots, x_{i-n})$$

For gene finding usually: 5th-order HMM, *i.e.* frequency if 6-tuples (di-codons) modelled.

Remark: *n*-th order HMM for state set $\mathcal{A}$ equivalent to first-order HMM for state set $\mathcal{A}^n$.

Complex HMMs

Example: $\mathcal{A} = \{A, B\}$

Equivalent:

$$P(x_i = B | x_{i-1} = A, x_{i-2} = A)$$

and

$$P(y_i = AB | y_{i-1} = AA)$$

Problem: For higher-order HMMs more training data necessary

Complex HMMs

(b) Interpolated HMMs (IHMMs)

Order of HMM varying, depending on amount of available training data.

(c) Inhomogeneous HMMs

Emissions probabilities depend on position

E.g. 3-periodic inhomogeneous HMMs of 5-th order used by modern gene finders to model coding regions (modelling frequency of *dicodons*).

Complex HMMs

(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

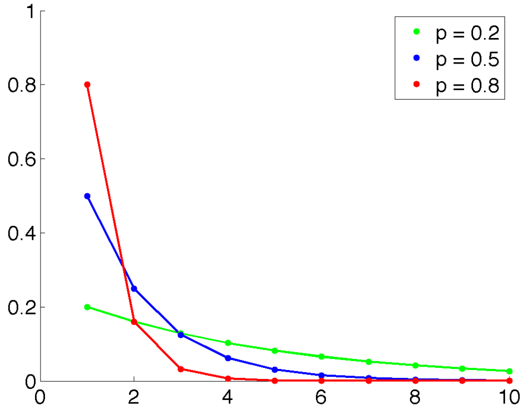


Figure: Geometric distribution for different parameters *p* (Wikipedia)

Complex HMMs

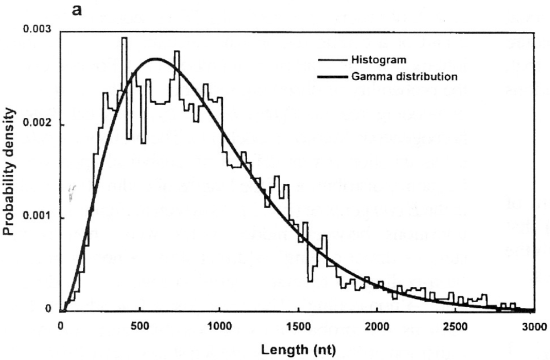


Figure: Real length distribution of protein-coding regions in *E. coli* (Lukashin and Borodovsky, 1998)

Complex HMMs

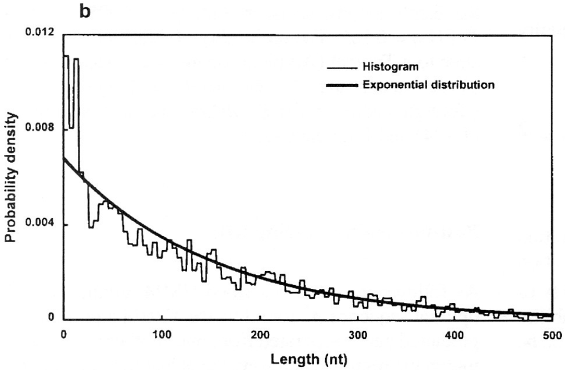


Figure: Real length distribution of non-coding regions in *E. coli* (Lukashin und Borodovsky, 1998)

Complex HMMs

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)