Hidden Markov Models for sequence analysis

Yves Moreau
Master of Bioinformatics
Katholieke Universiteit Leuven
2001-2002

Overview

- Sequence alignment

- Hidden Markov Models
- Profile HMMs
- Gene finding

Hidden Markov Models

Regular expressions

- Alignment

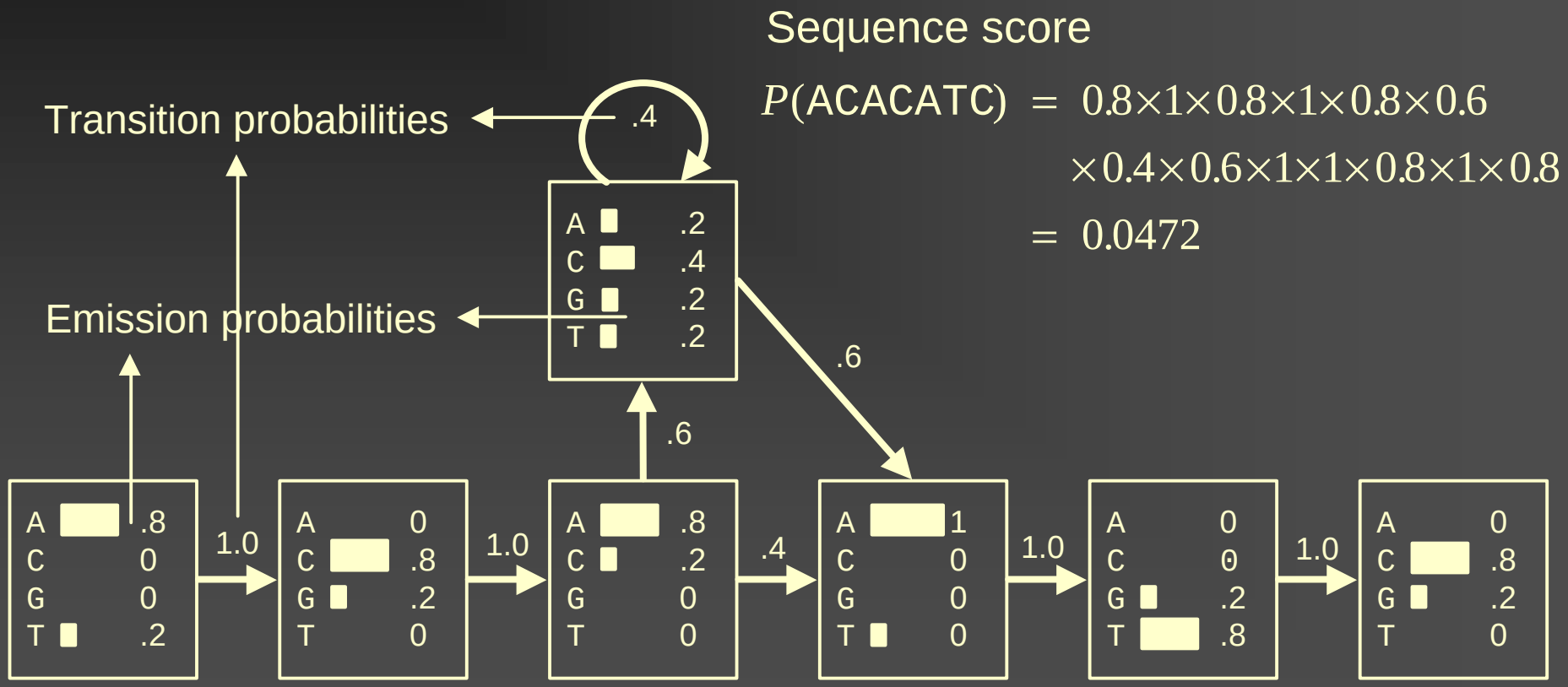
```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

- Regular expression

```
[AT][CG][AC][ACGT]*A[TG][GC]
```

- Problem: regular expression does not distinguish
  - Exceptional `TGCTAGG`
  - Consensus `ACACATC`

Hidden Markov Models

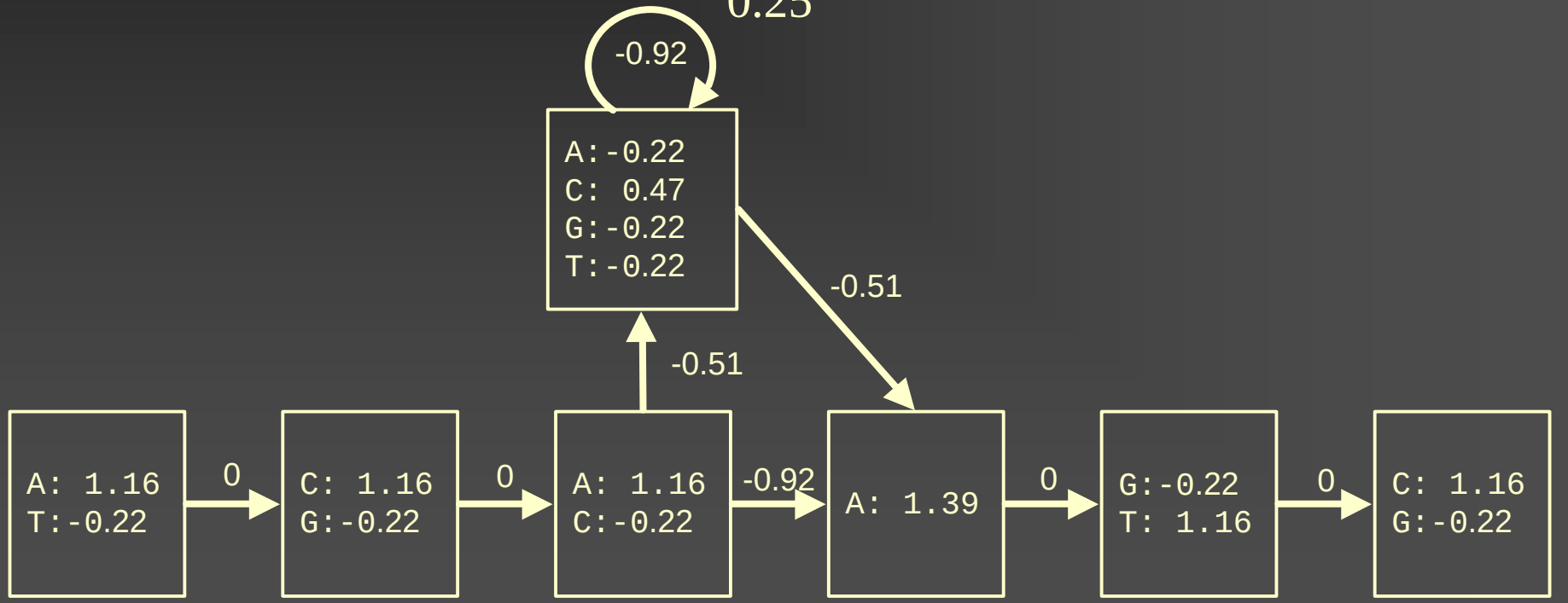


Sequence score

$$P(\text{ACACATC}) = 0.8\times1\times0.8\times1\times0.8\times0.6 \times0.4\times0.6\times1\times1\times0.8\times1\times0.8 = 0.0472$$

Log odds

- Use logarithm for scaling and normalize by random model
- Log odds for sequence S: $\log \frac{P(S)}{0.25^L} = \log P(S) - L \log 0.25$



Log odds

$$\log\text{-odds}(\text{ACACATC}) = 1.16 + 0 + 1.16 + 0 + 1.16 - 0.51 + 0.47 - 0.51 + 1.39 + 1.16 + 0 + 1.16 = 6.65$$

Sequence	Log odds
ACAC--ATC (consensus)	6.7
ACA---ATG	4.9
TCAACTATC	3.0
ACAC--AGC	5.3
AGA---ATC	4.9
ACCG--ATC	4.6
TGCT--AGG (exceptional)	-0.97

Training HMMs

- Hidden Markov Model
  - The observations are the emitted symbols
  - The state transitions are hidden
- Estimation and training of HMMs can be done with several dynamic programming methods
- We can compute the probability of observing a sequence of observations given the sequence of the states (see previous slides)

Training HMMs

- In general, we do not know the sequence of the states (as the states are hidden) but we can find which sequence of states gives the highest probability of producing the sequence of observations = *Viterbi algorithm*
- Also, we can estimate the probability of producing a sequence of observations independently of the state sequence = *forward algorithm*
- Further, we can estimate the probability of the different states at each observation in the observation sequence = *backward algorithm*

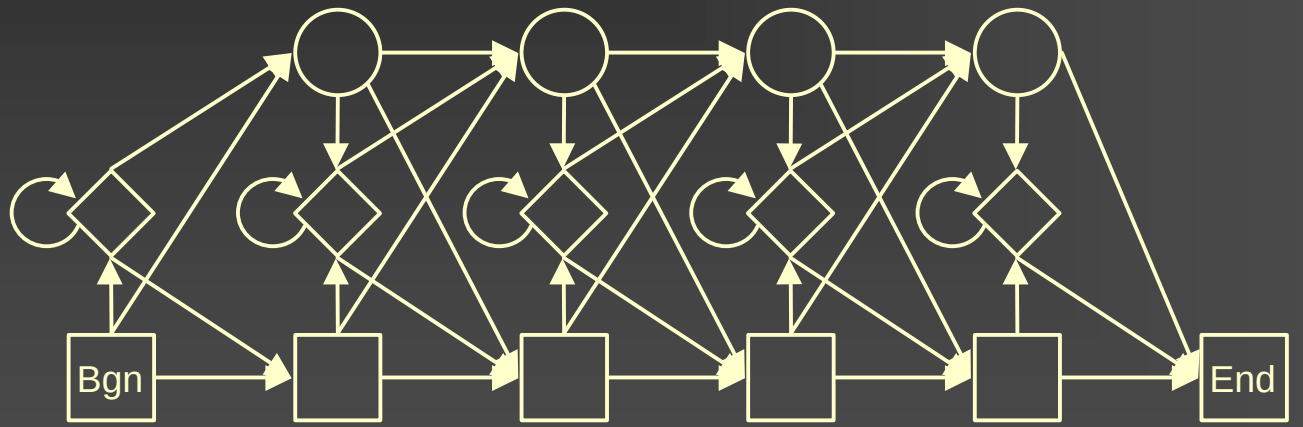
Training HMMs

- Parameter estimation
  - Based on a set of sequences, we want to find an HMM that could have produced this set of sequences
  - If the state sequence is known, estimating the transition and emission probabilities amounts to counting
  - If the state sequence is not known, we optimize the parameters so as to maximize the overall likelihood of the data
  - The strategy is to find the best state sequences given the current parameters and to reestimate the parameters given the best state sequences; this procedure is then iterated until convergence
    - Viterbi training uses the Viterbi algorithm in the iteration
    - Baum-Welch training uses the forward-backward algorithm in the iteration

Profile HMMs

Profile HMMs

- Hidden Markov Models for multiple alignments
- Match, insert, and delete states

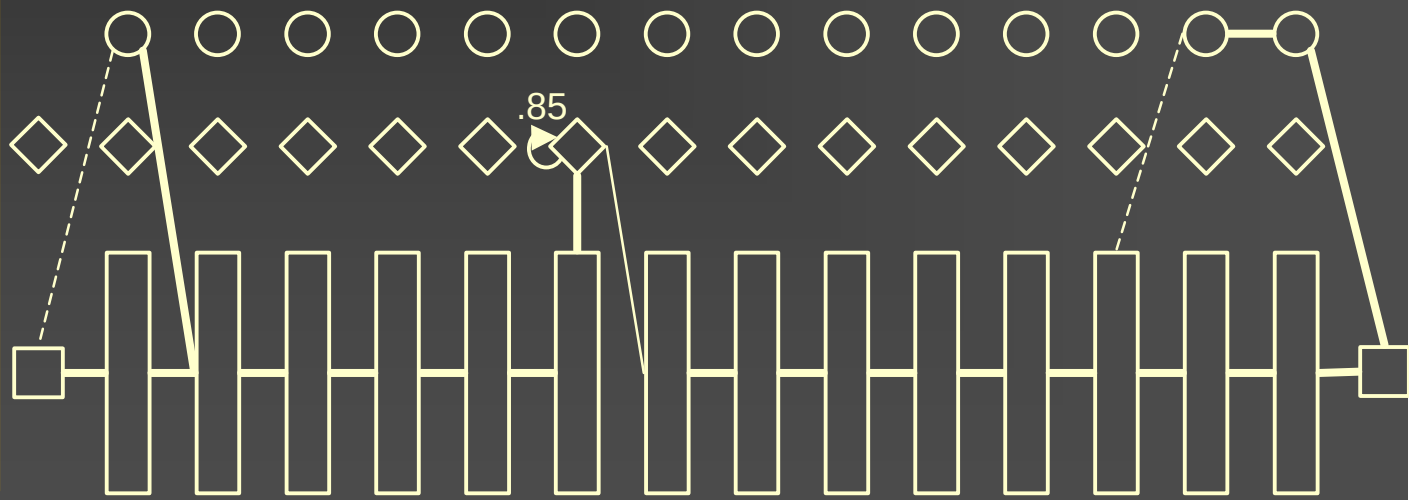


HMM from multiple alignment

```
GGWWRGdy.ggkkqLWFPSNYV
IGWLNGynettgerGDFPGTYV
PNWWEGql..nnrrGIFPSNYV
DEWWQArr..deqiGIVPSK--
GEWWKAqs..tgqeGFIPFNFV
GDWWLArs..sgqtGYIPSNYV
GDWWDAel..kgrrGKVPSNYL
-DWWEArslssghrGYVPSNYV
GDWWYArslitnseGYIPSTYV
GEWWKArslatrkeGYIPSNYV
GDWWLArslvtgreGYVPSNFV
GEWWKAkslsskreGFIPSNYV
GEWCEAgt.kngq.GWVPSNYI
SDWWRVvnlttrqeGLIPLNFV
LPWWRArd.kngqeGYIPSNYI
RDWWEFrsktvytpGYYESGYV
EHWWKVkd.algnvGYIPSNYV
IHWWRVqd.rngheGYVPSSYL
KDWWKVev..ndrqGFVPAAYV
```

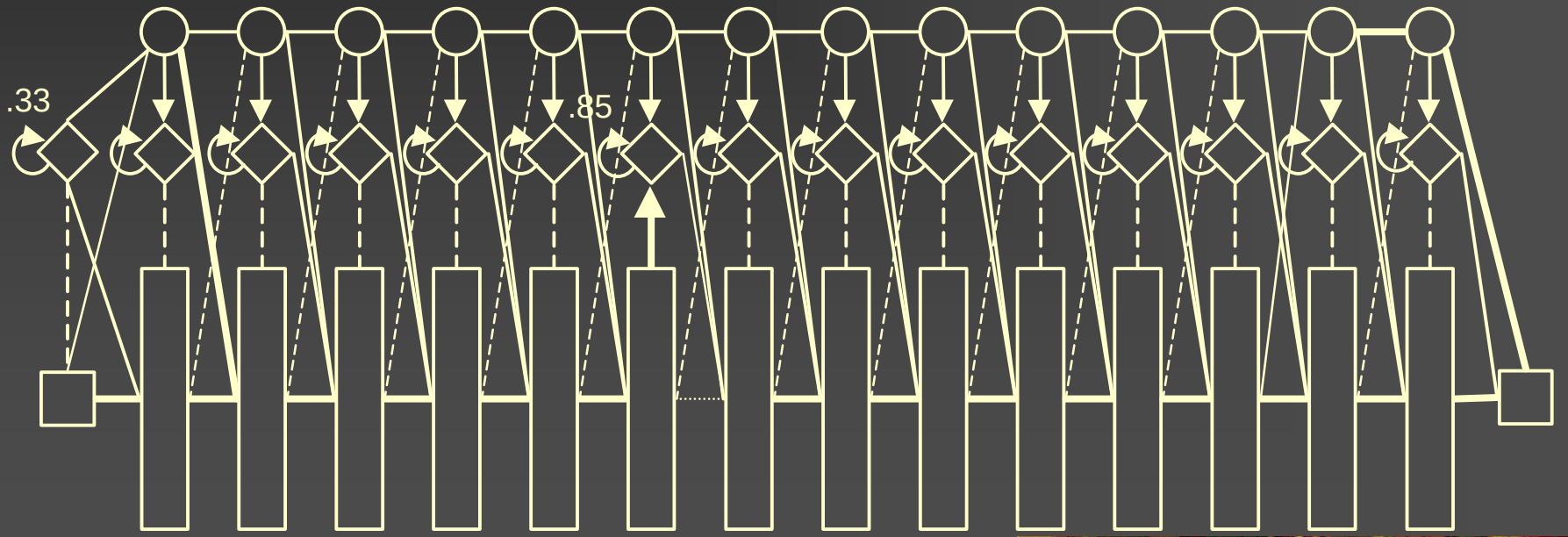
Multiple alignment (+ conserved columns)

Corresponding profile HMM



Pseudocounts

- Zero probabilities in HMM causes the rejection of sequences containing previously unseen residues
- To avoid this problem, add pseudocounts (add extra counts as if prior data was available)



New profile HMM

Extensions

- More sophisticated pseudocounts are possible
- Different types of local alignments can be done with HMMs (e.g., Smith-Waterman)
- Profile HMMs can be estimated from unaligned sequences
  - This produces also the multiple alignment
  - Need large number of sequences

Software for profile HMMs

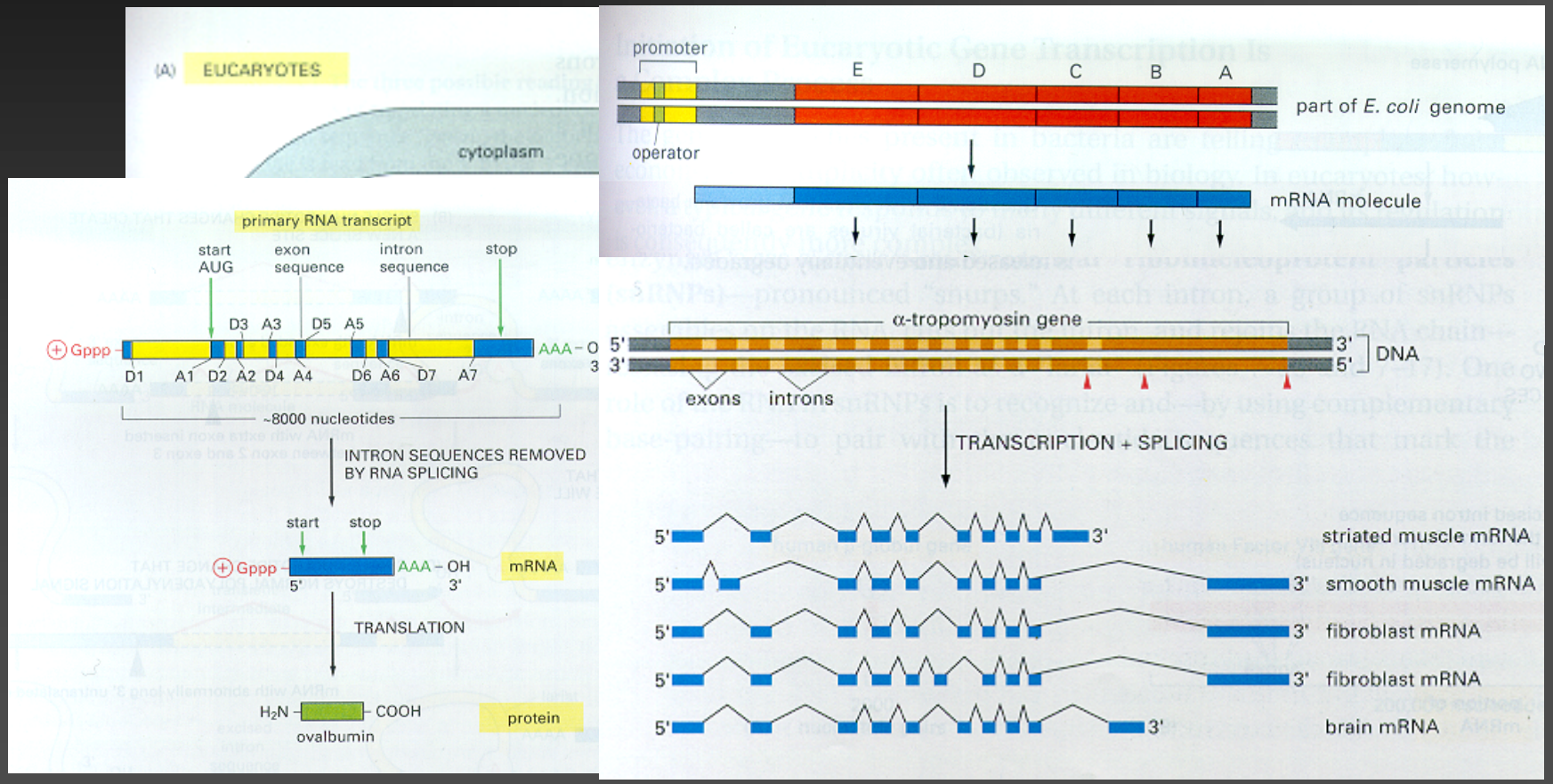
- SAM: University of California Santa Cruz
  - http://www.cse.ucsc.edu/research/compbio/sam.html
  - Web service: http://www.cse.ucsc.edu/research/compbio/HMM-apps/HMM-applications.html (takes time)
- Hmmer ('hammer'): Washington University, St. Louis
  - http://genome.wustl.edu/eddy/hmmer.html

Protein families

- PFAM
  - http://www.sanger.ac.uk/Software/Pfam/search.shtml
  - Collection of protein families and protein domains
  - Provides multiple alignment of the protein families for the domains
  - Provides the domain organization of proteins
  - Provides profile HMMs of the domains

Gene finding

DNA makes RNA makes proteins



Elements for gene prediction

- Prokaryotes vs. eukaryotes
  - Higher gene density in prokaryotes
  - No introns in prokaryotes
  - In prokaryotes, large ORFs (>300-500bp) likely to contain a gene but smaller genes are harder to detect
- Eukaryotes
  - Promoter
  - CpG islands
  - 5’UTR
  - Exons
  - Introns
  - 3’UTR
  - Splice sites
  - Repeats

Evidence for gene prediction

- Sources of evidence (positive and negative)
  - Similarity to features not likely to overlap coding sequence (e.g., Alu repeats)
  - Sequence similarity to known genes (e.g., found by BLASTX)
  - Statistical measure of codon bias
  - Template matches to functional sites (e.g., splice site)
- The structure must respect the biological grammar (promoter, exon, intro, ...)

Search by signal vs. search by content

- Search by signal
  - Detect short signals in the genome
  - E.g., splice site, signal peptide, glycosylation site
  - Neural networks can be useful here
- Search by content
  - Detect extended regions in the genome
  - E.g., coding regions, CpG islands
  - Hidden Markov Models are useful here
- Gene finding algorithms combine both

Probabilistic prediction vs. homology

- Hidden Markov Models can be used to predict genes
- Homology to a known gene is also a strong methods for detecting genes
- More and more gene prediction packages combine both approaches

CpG islands

- In mammalians, CpG islands have higher G+C and CG dinucleotide content than the rest of the DNA
- CpG islands arise in active regions where no deactivation by methylation takes place (CG dinucleotides in methylated regions disappear by deamination)
- CpG islands may be used as gene markers in mammalians

Repeats

- Repeats make up a large part of the human genome
  - Alu repeats
  - Long Interspersed Elements (LINEs)
  - Short Interspersed Elements (SINEs)
- Important to mask repeats when searching for genes

Promoter region

- Promoter region contains the element that control the expression of the gene
  - Prediction of the promoter region (e.g., prediction of the TATA-box) is difficult
  - Current methods in human do not significantly better than the prediction of CpG islands

Translation initiation site

- Translation is initiated at an AUG codon (= methionine)

Splice sites

- Introns usually
  - Begin with GU dinucleotide (donor splice site)
  - End with AG dinucleotide (acceptor splice site)

- Neural networks hav been applied for analysis of splice sites, they are able to detect complex correlation between positions in a functional site

Polyadenylation signal

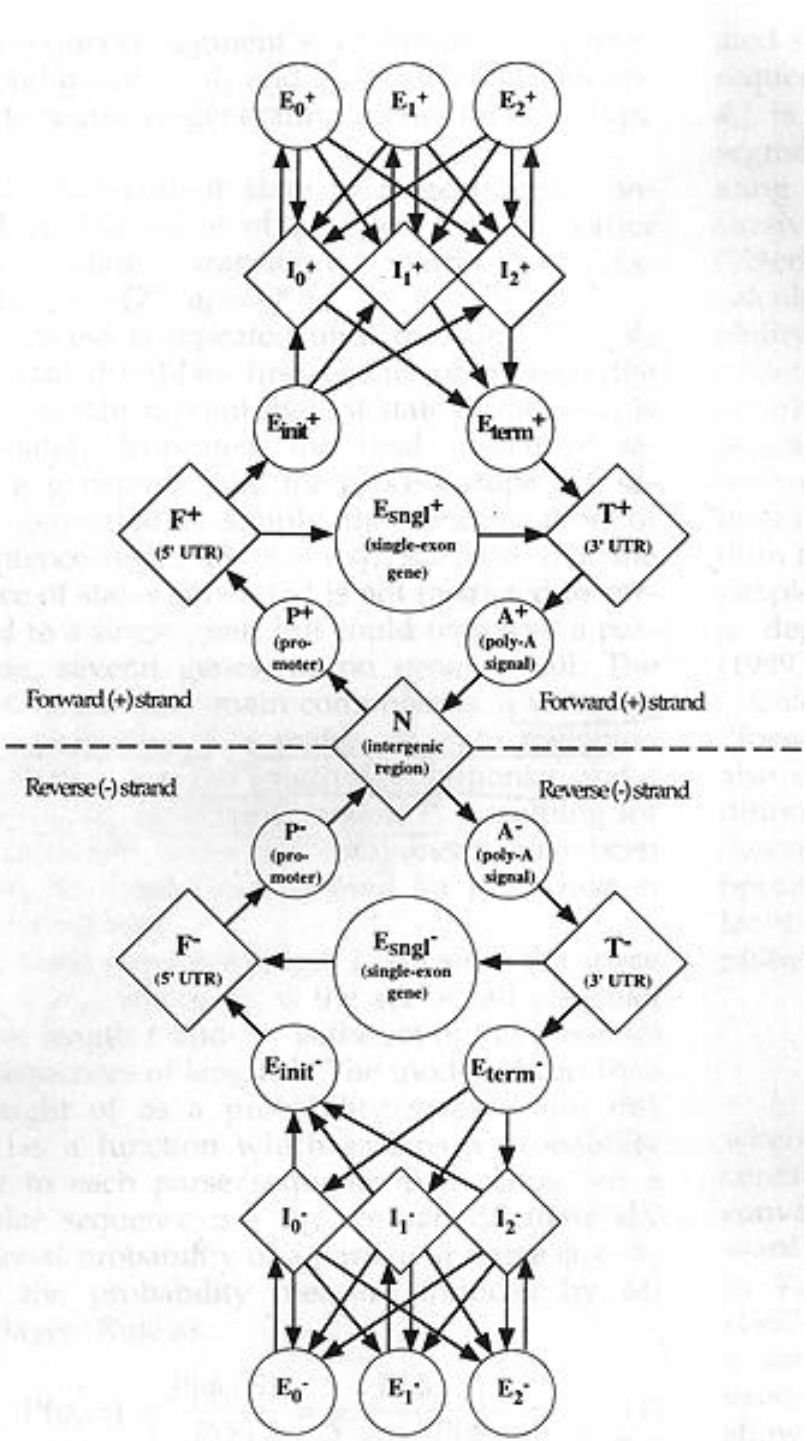
- Polyadenylation (cleavage of pre-mRNA 3' end and synthesis of poly-(A) tract) is a very important early step of pre-mRNA processing
- The most well-known signal involved in this process is AATAAA, located 15-20 nucleotides upstream from the poly-(A) site (site of cleavage)
- Real AATAAA signals can differ from AATAAA consensus sequence. The most frequent natural variant, ATTAAA, is nearly as active as the canonical sequence.

Coding potential

- Most coding potentials are based on analysis of codon usage
- The HMMs keeps track of some kind of average coding potential around each position
- The increase and decrease of the coding potential will “push” the HMM in and out of the exons

Gene prediction architecture

- Architecture of GENSCAN



Gene structure prediction

- Current performance on exon prediction is acceptable
- However, grouping the correct exons into the genes is still problematic
- In many cases, the majority of the predicted genes will not be correct