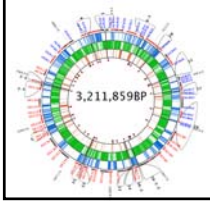
Computational Genomics

10-810/02-710, Spring 2009

Gene Finding and HMM

Eric Xing

Lecture 4, January 26, 2009

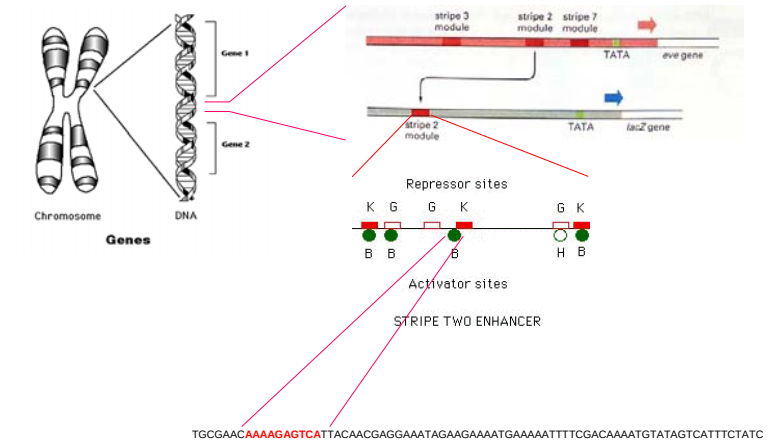


Reading: Durbin chap 3.

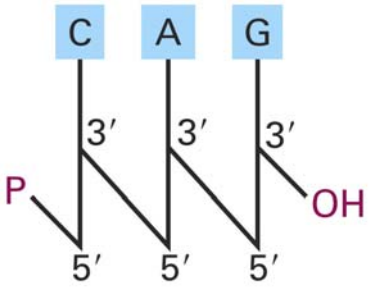
Please correct

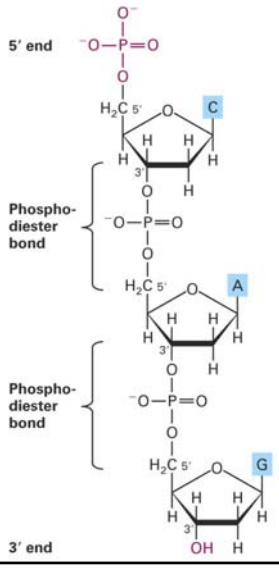
- 10-810-09s-instr@cs should be 10810-09s-instr@cs
- 10-810-09s-announce@cs should be 10810-09s-announce@cs.

Hierarchical structure of the genome



The DNA strand has a chemical polarity





Writing DNA sequence

- One strand is written by listing its bases in 5' to 3' order

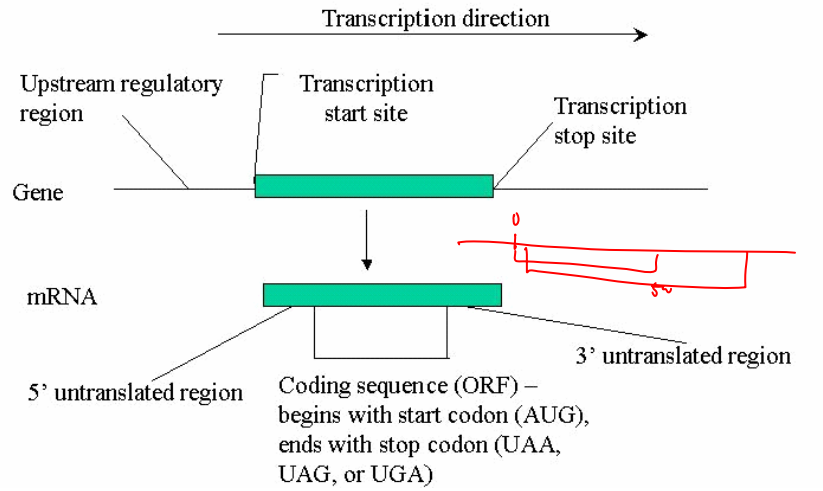
  5' ACCGTTACT 3'

- Each strand uniquely determines the complementary strand, which runs in the opposite direction:

  5' ACCGTTACT 3'
  3' TGGCAATGA 5'

- So the reverse complement of ACCGTTACT is written TGGCAATGA
- In general people write one strand and in 5' to 3' order
  - This is the ordering that a polymerase or a ribosome scan the sequence
  - Establishes a common standard for genome nomenclatures

Gene structure in prokaryotes

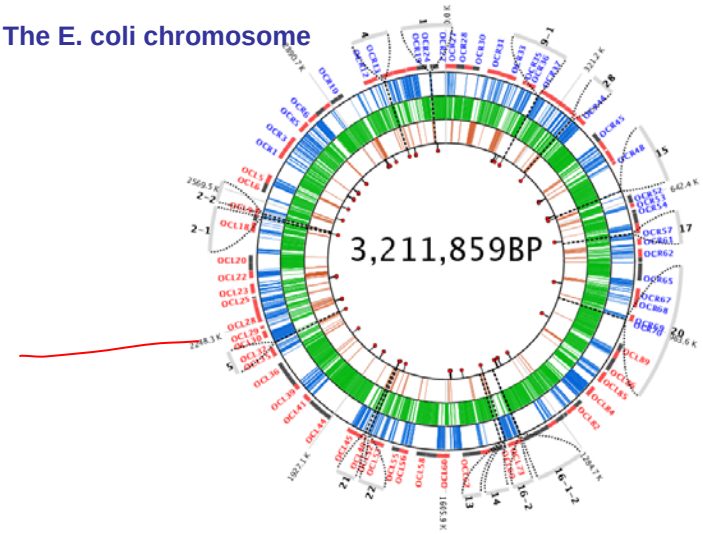


Gene structure in prokaryotes

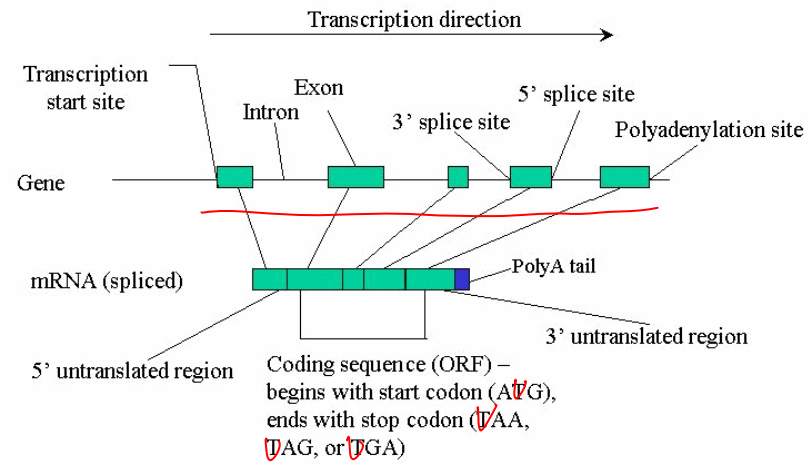
- **A protein-coding gene consists of the following, in 5' to 3' order**
  - An ***upstream regulatory region***, generally < 50 bp, which turns transcription on and off.
  - A ***transcription start site*** where RNA polymerase incorporates 1st nucleotide into nascent mRNA.
  - A 5' *untranslated region*, generally < 30bp, that is transcribed into mRNA but not translated.
  - The **translation start site** marking the start of the coding region. Consists of a **start codon**, which causes the start of translation
  - The ***coding region*** of the gene (typically=1000bp), consisting of a sequence of codons.
  - The ***translation stop site*** marking the end of coding region. Consists of a **stop codon**, which causes the release of the polypeptide at conclusion of translation.
  - A 3' *untranslated region*, transcribed into RNA but not translated.
  - The ***transcription stop site*** marking where the RNA polymerase concludes transcription.

The bacterial genome

The E. coli chromosome



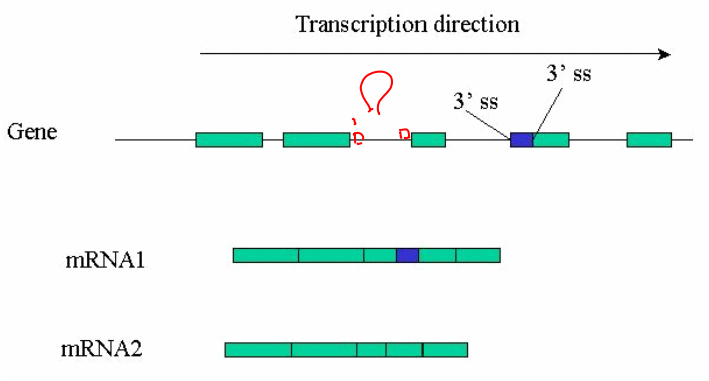
Gene structure in eukaryotes



Gene structure in eukaryotes

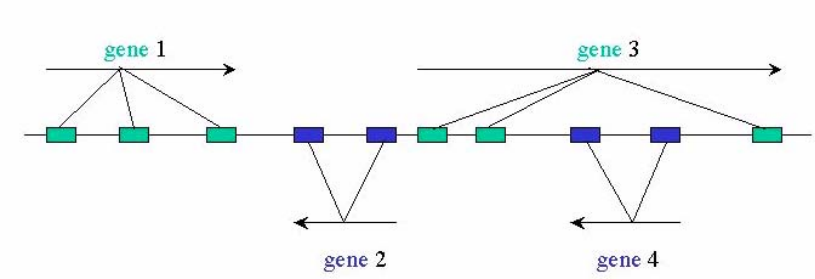
- **A typical gene consist of the following, in 5' to 3' order**
  - An ***upstream regulatory region***, often larger and more complex than in prokaryotes, parts of which may be several thousand bases or more upstream of transcription start site.
  - A ***transcription start site***.
  - A **5' *untranslated region***, often larger than in prokaryotes, and which may include sequences playing a role in translation regulation.
  - The ***coding sequence***, which unlike the case with prokaryotes, may be interrupted by non—coding regions called introns. These are spliced out of the transcript to form the mature mRNA (and sometimes the splicing can occur in more than one way).
  - The ***translation stop site***.
  - A **3' *untranslated region***, which may contain sequences involved in translational regulation.
  - **A *polyadenylation (playA) signal***, which indicates to the cell's RNA processing machinery that the RNA transcript is to be cleaved and a poly-adenine sequence (AAAAAA…) tail appended to it
  - The ***transcription stop site***.

Alternative splicing

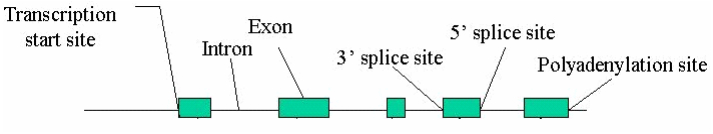


Eukaryotic genome structure

- Genes may be transcribed in either direction, and can overlap



Probabilistic segmentation



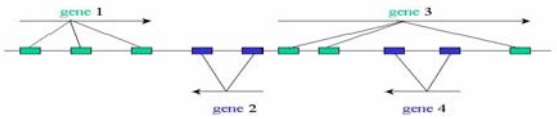
Probabilities on Sequences

- Let $\mathcal{S}$ be the space of DNA or protein sequences of a given length n. Here are some simple assumptions for assigning probabilities to sequences:
  - **Equal frequency assumption**: All residues are equally probable at any position; i.e., $P(X_{i,r})=P(X_{i,q})$ for any two residues r and q, for all i.
    - this implies that $P(X_{i,r})=\theta_r=1/|\mathcal{A}|$, where $\mathcal{A}$ is the residue alphabet (1/20 for proteins, 1/4 for DNA)
  - **Independence assumption**: whether or not a residue occurs at a position is independent of what residues are present at other positions.
    - probability of a sequence

$$P(X_1, X_2, ..., X_N) = \theta_r \cdot \theta_r \cdot ..., \cdot \theta_r = \theta_r^N$$

Failure of Equal Frequency Assumption for (real) DNA

- For most organisms, the nucleotides composition is significantly different from 0.25 for each nucleotide, e.g.,
  - *H, influenza* .31 A, .19 C, .19 G, .31 T
  - *P. aeruginosa* .17 A, .33 C, .33 G, .17 T
  - *M. janaschii* .34 A, .16 C, .16 G, .34 T
  - *S. cerevisiae* .31 A, .19 C, .19 G, .31 T
  - *C. elegans* .32 A, .18 C, .18 G, .32 T
  - *H. sapiens* .30 A, .20 C, .20 G, .30 T
- Note symmetry: A≅T, C≅G, even thought we are counting nucleotides on just one strand. Explanation:
  - although individual biological features may have non-symmetric composition, usually features are distributed ~ randomly w.r.t. strand, so get symmetry.



General Hypothesis Regarding Unequal Frequency

- Neutralist hypothesis: mutation bias (e.g., due to nucleotide pool composition)
- Selectionist hypothesis: natural selection bias

Models for Homogeneous Sequence Entities

- Probabilities models for long "homogeneous" sequence entities, such as:
  - exons (ORFs)
  - introns
  - inter-genetic background
  - protein coiled-coil (other other structural) regions
- Assumptions:
  - no consensus, no recurring string patterns
  - have distinct but uniform residue-composition (i.e., same for all sites)
  - every site in the entity are iid samples from the same model
- The model:
  - a single multinomial: $X \sim \text{Mul}(1, \theta)$

The Multinomial Model for Sequence

- For a site i, define its residue identity to be a multinomial random vector:

$$X_i = \begin{bmatrix} X_{i,A} \\ X_{i,C} \\ X_{i,G} \\ X_{i,T} \end{bmatrix}, \quad \text{where} \quad X_{i,j} = [0,1], \text{ and } \sum_{j \in \{A,C,G,T\}} X_{i,j} = 1$$
$$X_{i,j} = 1 \text{ w.p. } \theta_j, \quad \sum_{j \in \{A,C,G,T\}} \theta_j = 1 \ .$$

- The probability of an observation s_i=A (i.e, $x_{i,A}$=1) at site i:

$$p(x_i = (\text{say}, A)) = P(\{X_{i,j} = 1, \text{where } j = A \text{ index the observed nucleotide}\})$$
$$= \theta_A = \theta_A^{x_A} \times \theta_C^{x_C} \times \theta_G^{x_G} \times \theta_T^{x_T} = \prod_k \theta_k^{x_k} = \theta^x$$

- The probability of a sequence $(x_1, x_2, \ldots, x_N)$:

$$P(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} P(x_i) = \prod_{i=1}^{N} \prod_k \theta_k^{x_{i,k}}$$
$$= \prod_k \theta_k^{\sum_{i=1}^{N} x_{i,k}} = \prod_k \theta_k^{n_k}$$

Parameter Estimation

- Maximum likelihood estimation: $\theta = \arg\max_{\theta} P(D \mid \theta)$
  - multinomial parameters:

$$\{\theta_1, \theta_2 \ldots\} = \arg\max_{\theta} \prod_k \theta_k^{n_k}, \quad \text{s.t.} \sum_k \theta_k = 1$$

$$\text{It can be shown that}: \theta_k^{\text{ML}} = n_k / N$$

- Bayesian estimation:
  - Dirichlet distribution:

$$P(\theta) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} = C(\alpha) \prod_k \theta_k^{\alpha_k - 1}$$

  - Posterior distribution of θ under the Dirichlet prior:

$$P(\theta \mid x_1, \ldots, x_N) \propto \prod_k \theta_k^{\alpha_k} \prod_k \theta_k^{n_k} = \prod_k \theta_k^{\alpha_k - 1 + n_k}$$

  - Posterior mean estimation:

$$\theta_k = \int \theta_k P(\theta \mid D) d\theta = \int \theta_k \prod_k \theta_k^{\alpha_k - 1 + n_k} d\theta = \frac{n_k + \alpha_k}{N + |\alpha|}$$

MAP

Models for Homogeneous Sequence Entities, ctd

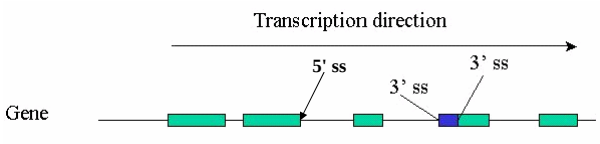
- Limitations
  - non-uniform residue composition (e.g., CG rich regions)
  - non-coding structural regions (MAR, centromere, telomere)
  - di- or tri- nucleotide couplings
  - estimation bias
  - evolutionary constrains

Site Models

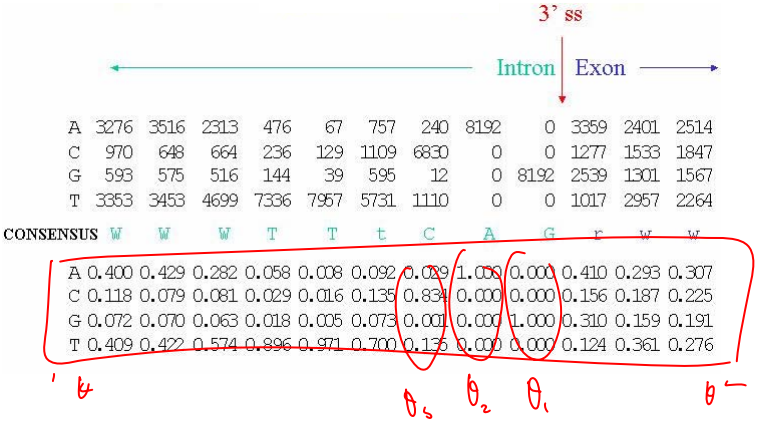
- Probabilities models for short sequences, such as:
  - splice sites
  - translation start sites
  - promoter elements
  - protein "motifs"
- Assumptions:
  - different examples of sites can be aligned without indels (insertions/deletions) such that tend to have similar residues in same positions
  - drop equal frequency assumption; instead have position-specific frequencies
  - retain independence assumption (for now)

Site Models ctd.

- Applies to short segments (<30 residues) where precise residue spacing is structurally or functionally important, and certain positions are highly conserved
  - DNA/RNA sequence binding sites for a single protein or RNA molecule
  - Protein internal regions structurally constrained due to folding requirements; or surface regions functionally constrained because bind certain ligands
- Example: C. elegans splice sites



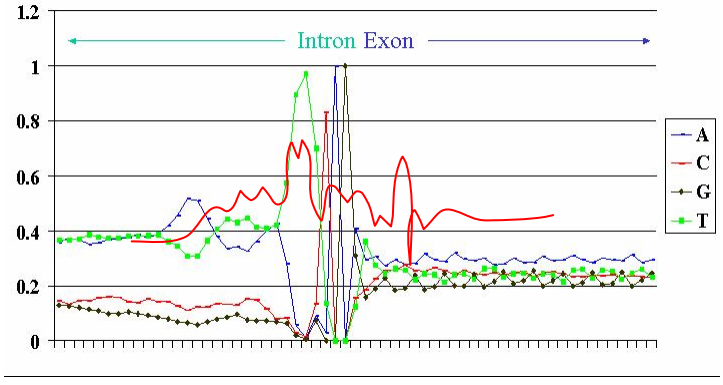
Nucleotide Counts for 8192 C. elegans 3' Splice Sites



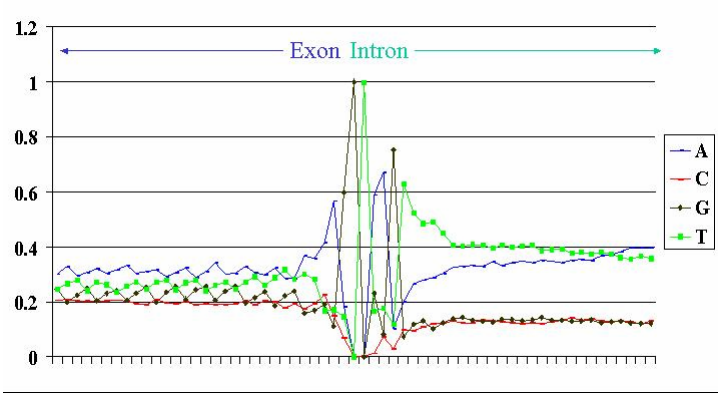
A	3276	3516	2313	476	67	757	240	8192	0	3359	2401	2514
C	970	648	664	236	129	1109	6830	0	0	1277	1533	1847
G	593	575	516	144	39	595	12	0	8192	2539	1301	1567
T	3353	3453	4699	7336	7957	5731	1110	0	0	1017	2957	2264
CONSENSUS	W	W	W	T	T	t	C	A	G	r	W	W

A	0.400	0.429	0.282	0.058	0.008	0.092	0.839	1.000	0.000	0.410	0.293	0.307
C	0.118	0.079	0.081	0.029	0.016	0.135	0.834	0.000	0.000	0.156	0.187	0.225
G	0.072	0.070	0.063	0.018	0.005	0.073	0.001	0.000	1.000	0.310	0.159	0.191
T	0.409	0.422	0.574	0.896	0.971	0.700	0.136	0.000	0.000	0.124	0.361	0.276

3' Splice Site - C. elegans



5' Splice Sites - C. elegans



Limitation of Homogeneous Site Models

- Failure to allow indels means variably spaced subelements are "smeared", e.g.:
  - branch site, for 3' splice sites;
  - coding sequences, for both 3' and 5' sites
- Independence assumption
  - usually OK for protein sequences (after correcting for evolutionary relatedness)
  - often fails for nucleotide sequences; examples:
    - 5' sites (Burge-Karlin observation);
    - background (dinucleotide correlation, e.g., GC repeats).

Why Correlation?

- Splicing involves pairing of a small RNA with the transcription at the 5' splice site.
- The RNA is complementary to the 5' srRNA consensus sequence.
- A mismatch at position -1 tends to destabilize the pairing, and makes it more important for other positions to be correctly paired.
- Analogy can be easily drew for other DNA and protein motifs.

Comparing Alternative Probability Models

- We will want to consider more than one model at a time, in the following situations:
  - To differentiate between two or more hypothesis about a sequence
  - To generate increasingly refined probability models that are progressively more accurate
- First situation arises in testing biological assertion, e.g., "is this a coding sequence?" Would compare two models:
  1. one associated with a hypothesis H_{coding} which attaches to a sequence the probability of observing it under experiment of drawing a random coding sequence from the genome
  2. one associate with a hypothesis $H_{noncoding}$ which attaches to a sequence the probability of observing it under experiment of drawing a random non-coding sequence from the genome.

Likelihood Ratio Test

- The posterior probability of a model given data is:

$$P(M|D) = P(D|M)P(M)/P(D)$$

- Given that all models are equally probable *a priori*, the posterior probability ratio of two models given the same data reduce to a *likelihood ratio*:

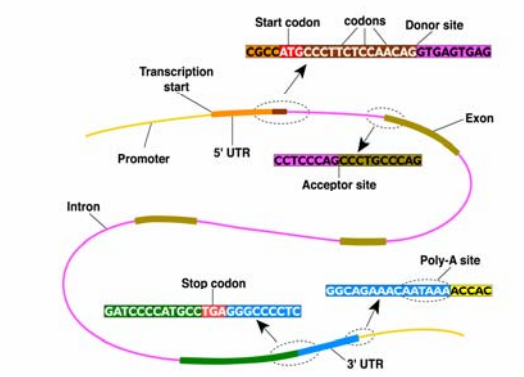
$$LR(M_a, M_0 \mid D) = \frac{P(D \mid M_a)}{P(D \mid M_0)}$$

  - the numerator and the denominator may both be very small!

- The log likelihood ratio (LLR) is the logarithm of the likelihood ratio:

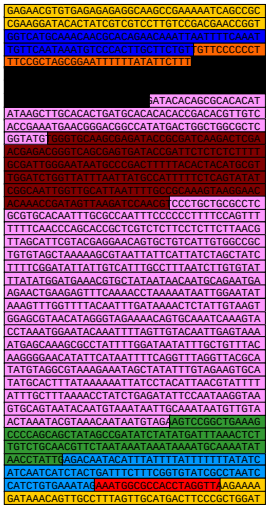
$$LLR(M_a, M_0 \mid D) = \log P(D \mid M_a) - \log P(D \mid M_0)$$

The Hidden Markov Models for sequence parsing



Gene Finding

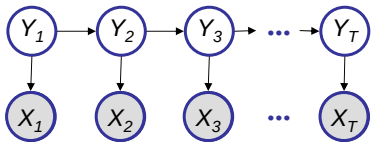
- Given un-annotated sequences,
- delineate:
  - transcription initiation site,
  - exon-intron boundaries,
  - transcription termination site,
  - a variety of other motifs: promoters, polyA sites, branching sites, etc.
- The hidden Markov model (HMM)



Hidden Markov Models

The underlying source:
genomic entities,
dice,

The sequence:
Ploy NT,
sequence of rolls,

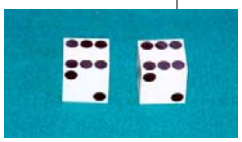


Example: The Dishonest Casino

A casino has two dice:

- Fair die
  P(1) = P(2) = P(3) = P(5) = P(6) = 1/6
- Loaded die
  P(1) = P(2) = P(3) = P(5) = 1/10
  P(6) = 1/2

Casino player switches back-&-forth between fair and loaded die once every 20 turns



Game:

1. You bet $1
2. You roll (always with a fair die)
3. Casino player rolls (maybe with fair die, maybe with loaded die)
4. Highest number wins $2

Puzzles Regarding the Dishonest Casino

GIVEN: A sequence of rolls by the casino player

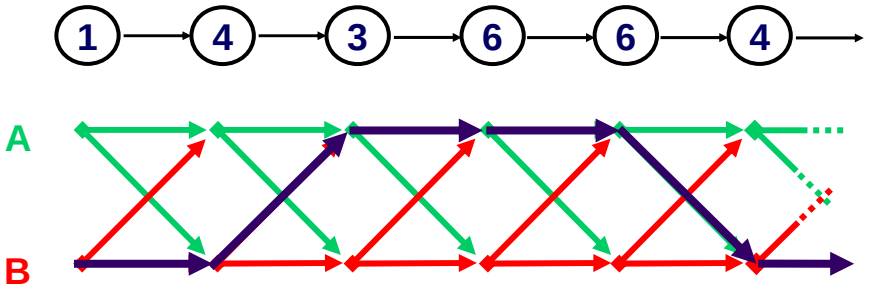
12455264621461461361366616646616366163661636616515615115146123562344

QUESTION

- How likely is this sequence, given our model of how the casino works?
  - This is the **EVALUATION** problem in HMMs
- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
  - This is the **DECODING** question in HMMs
- How "loaded" is the loaded die? How "fair" is the fair die? How often does the casino player change from fair to loaded, and back?
  - This is the **LEARNING** question in HMMs

A Stochastic Generative Model

- Observed sequence:



- Hidden sequence (a parse or segmentation):

Definition (of HMM)

- Observation space
  - **Alphabetic set:** $\mathbb{C} = \{c_1, c_2, \cdots, c_K\}$
  - **Euclidean space:** $\mathbb{R}^d$
- Index set of hidden states

  $\mathrm{I} = \{1,2,\cdots,M\}$
- Transition probabilities between any two states

  $p(y_t^j = 1 \mid y_{t-1}^i = 1) = a_{i,j}, \quad \forall i, j \in \mathrm{I}.$

  or

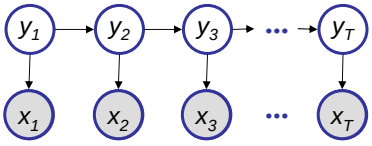
  $p(y_t \mid y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \ldots, a_{i,M}), \quad \forall i \in \mathrm{I}.$
- Start probabilities

  $p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M).$
- Emission probabilities associated with each state

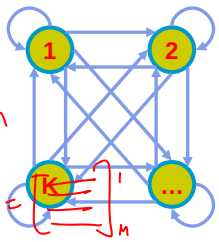
  $p(x_t \mid y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \ldots, b_{i,K}), \quad \forall i \in \mathrm{I}.$

  or in general:

  $p(x_t \mid y_t^i = 1) \sim \mathrm{f}(\cdot \mid \theta_i), \quad \forall i \in \mathrm{I}.$



Graphical model



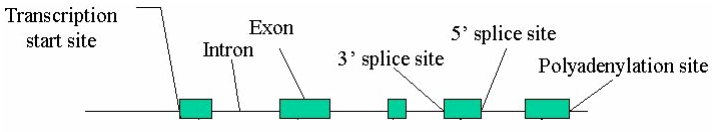
State automata

Probability of a Parse

- What is a parse?

$x_{1:N}$ 1245526462146146136136661664661636616366163616515615115146123562344

$y_{1:N}$ - - - - - - - -



- How to score a parse?

$P(y_t \mid x_{1:N}) = \frac{P(y_t, x_{1:N})}{P(x_{1:N})}$

$P(y_{1:N} \mid x_{1:N}) = P(y_{1:N}, x_{1:N})$