1.3

Recognizing signals

Recognizing signals in the sequence

(1) Signals are functional sites
- E.g., start/stop, intron to extron (splice site), etc.

(2) Sometimes signals follow characteristic patterns
- E.g. splice site A G G T (A | G) A G T

(3) Describing patterns
- Regular expressions
- Statistical models

(4) Recognizing patterns

(5) Learning patterns from examples
- Determining statistical coefficients, e.g. transition probabilities

Regular expressions

(1) Describe a set of words (language)

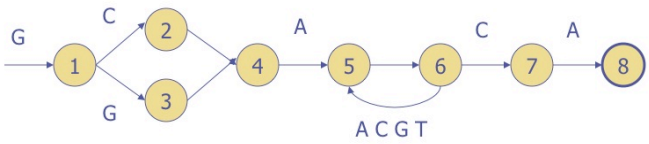
- The letters A C G T denote a word consisting of that letter
- Concatenation of two expressions x and y denotes all concatenations that can be built from words denoted by x and y
- (x | y) denotes all words denoted either by x or y
- { x } * denotes 0 or more repetitions of the words denoted by x

(2) Example: G (C | G) A { G | A | C | T }* C A

- Denotes all DNA sequences that start with either G C A or G G A and end with C A

Finite automata

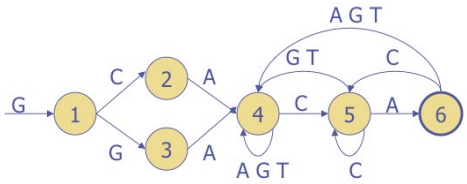
(1) Efficient match for words



Non-deterministic automaton

Word is accepted if there exists a corresponding path through the graph from start to end state

Deterministic finite automata



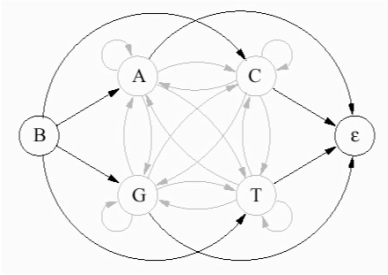
Can always be constructed, efficient table representation

	0	1	2	3	4	5	6
A	x	x	4	4	4	6	4
C	x	2	x	x	5	5	5
G	1	3	x	x	4	4	4
T	x	x	x	x	4	4	4

Markov models

(1) Going from a fixed language to transition probabilities

- A Markov chain is a process where the probability of appearance of a state (character) depends only on the previous state (character), not on the complete history



Define transition probabilities for each state transition (outgoing probabilities must sum to 1)

Example: CpG islands

(1) The probability of C G sequences in the genome is lower than random

- Reason: C in this combination is typically methylated and has a tendency to mutate into T
- Methylation is suppressed in biologically interesting regions, such as around promotors and start regions of genes. The probability of C G sequences is higher there (CpG islands)

(2) A Markov chain can distinguish between CpG islands and regular sequences

- Take a number of example sequences of each and calculate the transition probabilities (relative frequency vs. other pairs)

Example transition probabilities

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

CpG islands

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

Regular sequences

Log-likelihood ratio for a transition: logarithm of quotient between +model and –model (e.g., log (0.274/0.078) for an observed C G transition)

To score a region, sum up the log-likelihood ratios of the occurring transitions and divide by the length, positive results are indicators for the +model

Sequence families

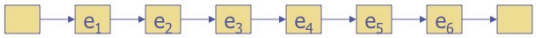
(1) Multiple alignment of related sequences
- E.g., multiple proteins with a known similar structure
- Manually align along structural information (loops and helices)
- Manually align key positions with known functionality

(2) Hidden Markov models to describe the „pattern"
- To check whether all of the structural elements are conserved
- Thus including „biological semantics" and not only substitution probabilities

(3) Profile HMMs
- Given manual alignment of example sequences
- Build model of structural features
- Estimate model parameters from example sequences
- Calculate most probable path and probability for new sequences

Position-specific score matrices

(1) Simple model for the position-specific probabilities of short ungapped segments
- e_i (a): probability that the amino acid a is observed in position i
- Equivalent to a HMM with n states



(2) Can be used to find pattern by scoring a segment from a larger sequence
- Iterate to find high-scoring segments
- Known segments can be stored in a database (BLOCKS)

Summary: Recognizing signals

(1) Regular grammars
- For short patterns
- Deterministic

(2) Position-specific scoring matrices
- Also called blocks or matrices
- For ungapped longer blocks

(3) Profile HMMs
- For carefully annotated patterns
- Most powerful, but require careful parameter estimation