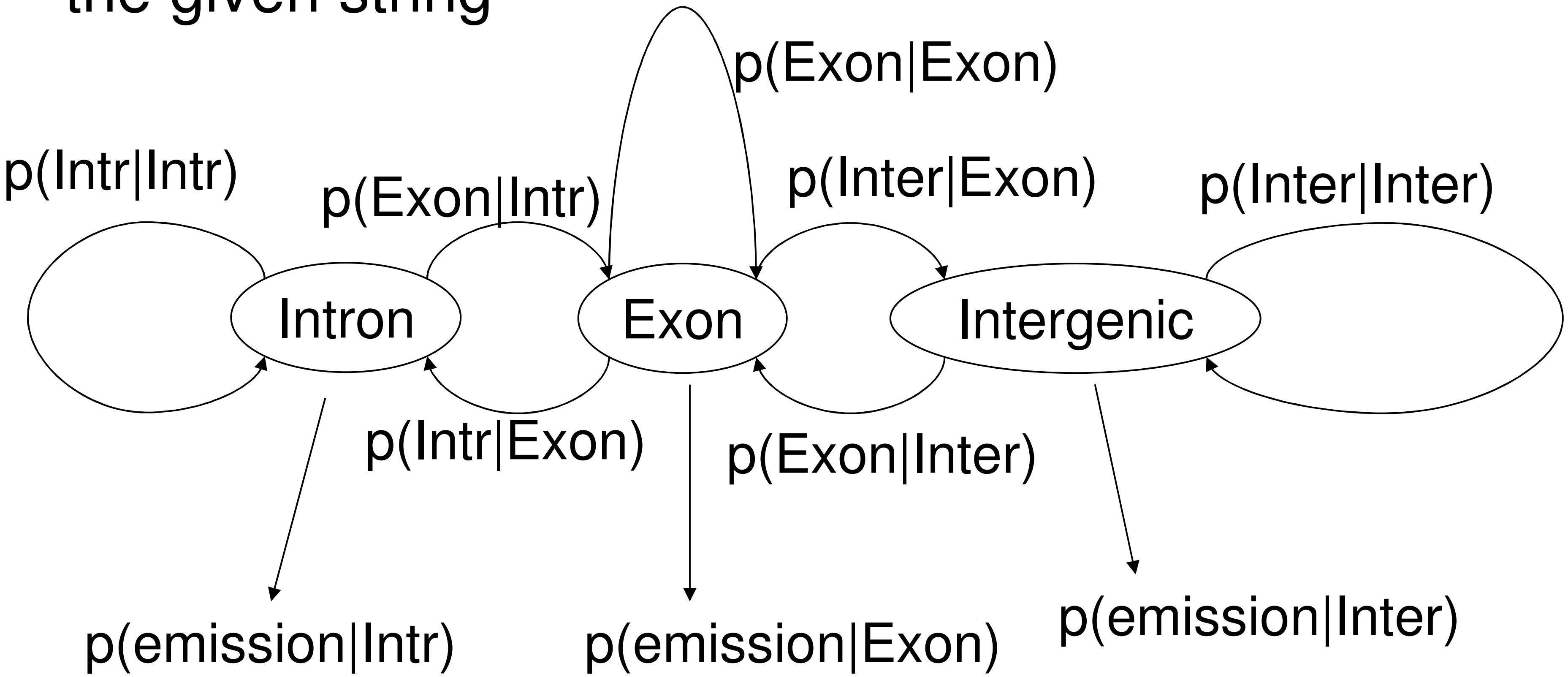
CMSC423: Bioinformatic Algorithms, Databases and Tools Lecture 19

Gene finding

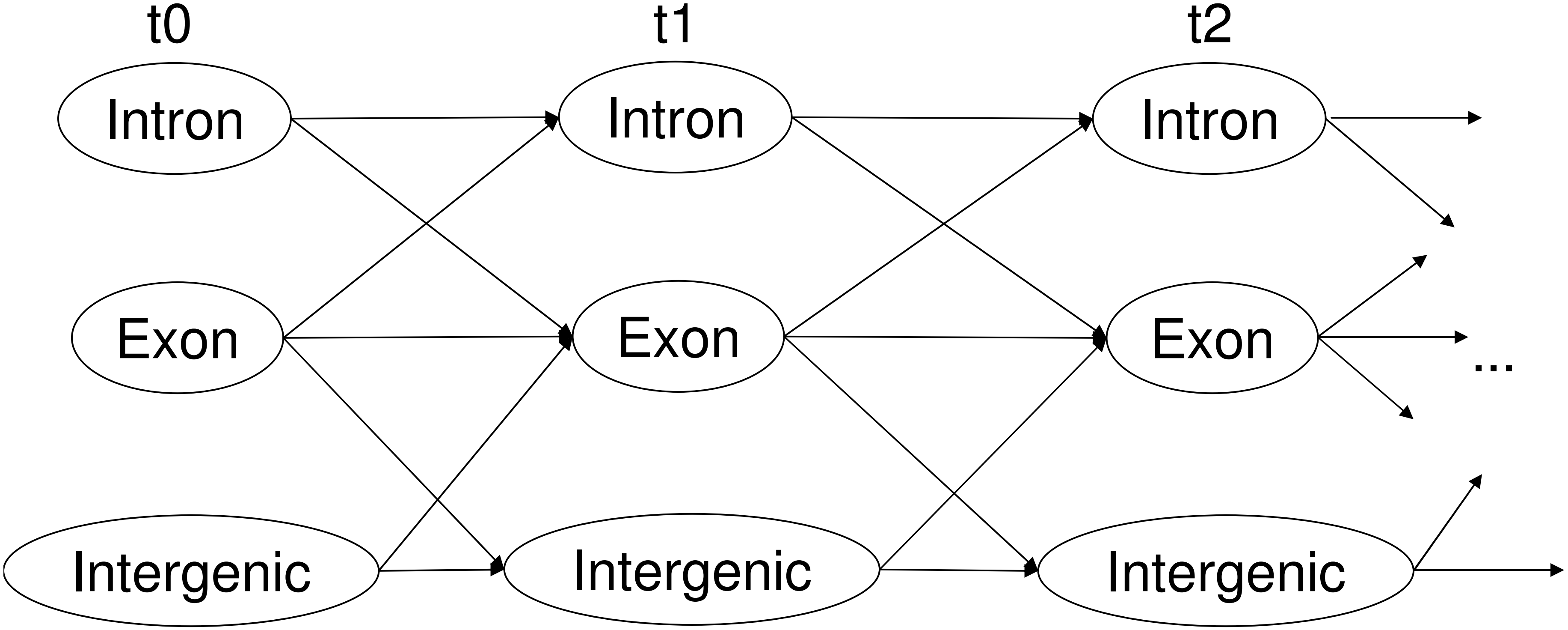
Motif finding

Viterbi algorithm

- Given an HMM and an output string, compute the most likely path through the HMM that would result in the given string



Viterbi algorithm



Observations:

x0 x1 x2

maximize $\prod_{0}^{n} e_{statej}(x_j)\, p(state_j \mid state_{j-1})$ over all possible state paths

dynamic programming algorithm

Viterbi algorithm

- S(k,i) – most likely path for x0..xi ends in state k
- $S(l, i + 1) = \max_k \{ S(k, i) * p(l|k) * p(\text{emission of } x_{i+1}|l))$
  $= p(\text{emission of } x_{i+1}|l) * \max_k \{S(k,i) * p(l|k)\}$

- The optimal path is found by back-tracking
- Note: Viterbi is equivalent to finding longest path in a graph
- Implementation problem: underflow – many products of very small values
- Solution: work in log-space
  - instead of probabilities use logarithm of probabilities
  - instead of products use sums

Forward-backward algorithm

- Given an HMM and an output string of length n, what is the probability that the HMM was in state k at time i < n?

- Similar dynamic programming as Viterbi however done twice:
  - from t0 to ti (forwards)
  - from tn to ti (backwards)
- In Viterbi recurrence replace max with ∑
  - likelihood is a sum of probabilities - all possible paths that go through state k at time i

Notes on training an HMM

- Gene finder output
  - a set of predictions (exon, intron, intergenic, etc.)
  - a probability (likelihood) for each prediction
- In addition to setting parameters for the model you also need to pick a threshold – how high should the probability be before you "believe" it.

Picking the "right" threshold

- Cross-validation (hold-out cross validation)
  - divide training set into Training and Hold sets
  - train in "Training"
  - test result on "Hold" – adjust threshold until results look best

- k-fold cross-validation
  - divide training set into K sub-sets
  - train on K-1 sets and test on one of them
  - repeat for different choices of "test" set

Assessing accuracy

- Confusion matrix: compare predictions to truth

truth

		Gene	Not-gene
prediction	Gene	True positive	False positive Type I error
	Not-gene	False negative Type II error	True negative

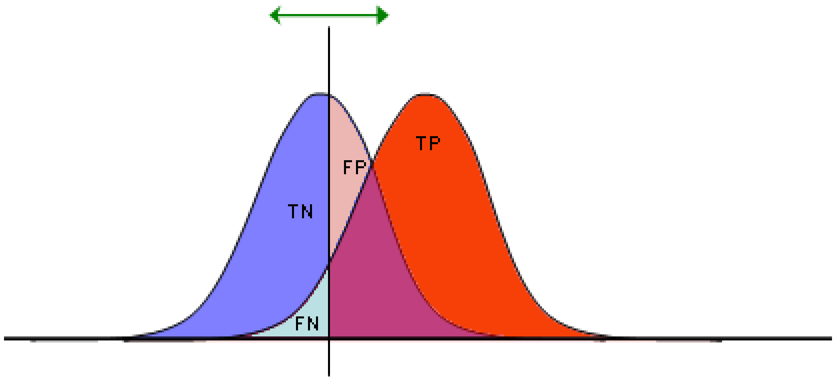
Measures of accuracy

- Sensitivity (Sn, recall) – TP/TP+FN
- Specificity (Sp) – TN/TN+FP
- Precision – TP/TP+FP

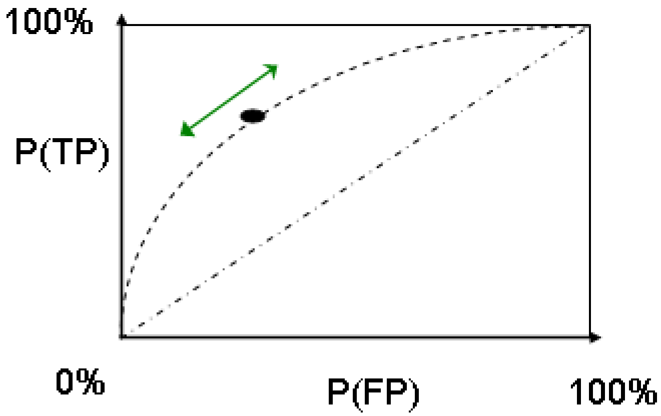
- Usually reported as (Sp, Sn), or (precision, recall).
- Also:
  F-score = 2*Precision*Recall/(Precision + Recall)

TP	FP
FN	TN

Receiver operating characteristic



TP	FP
FN	TN
1	1


P(TP)
P(FP)