

# Κρυφά μοντέλα Markov (hidden Markov models – HMM)

# Σύνοψη

- Νησίδα CG
- Το «Καζίνο με τα Δίκαια Στοιχήματα»
- Κρυφό μοντέλο Markov
- Αλγόριθμος αποκωδικοποίησης
- Προς τα εμπρός-προς τα πίσω αλγόριθμος
- HMM για προφίλ (αλληλουχιών)
- Εκτίμηση παραμέτρων HMM
- Αλγόριθμος Viterbi
- Αλγόριθμος Baum-Welch

# Νησίδια CG

- Έχοντας 4 νουκλεοτίδια: η πιθανότητα εμφάνισης είναι  $\sim 1/4$ . Άρα, η πιθανότητα εμφάνισης ενός δινουκλεοτιδίου είναι  $\sim 1/16$ .
- Όμως, οι συχνότητες των δινουκλεοτιδίων στις αλληλουχίες DNA διαφέρουν πολύ.
- Συγκεκριμένα, το CG δεν αντιπροσωπεύεται επαρκώς (η συχνότητα του CG είναι συνήθως  $< 1/16$ )

## ΝησίδιαCG : γιατί;

- Το CG είναι το λιγότερο συχνό δινουκλεοτίδιο επειδή το C στο CG μεθυλιώνεται εύκολα και έχει την τάση να μεταλλάσσεται σε T στη συνέχεια.
- Όμως, η μεθυλίωση καταστέλλεται συχνά γύρω από τα γονίδια σε ένα γονιδίωμα. Έτσι, το CG εμφανίζεται με σχετικά μεγάλη συχνότητα μέσα σε αυτά τα νησίδια CG.
- Άρα, η εύρεση των νησιδίων CG σε ένα γονιδίωμα είναι σημαντικό πρόβλημα

## Τα νησίδια CG και το «Καζίνο με τα Δίκαια Στοιχήματα»

- Το πρόβλημα των νησιδίων CG μπορεί να μοντελοποιηθεί από ένα πρόβλημα που ονομάζεται *«Το Καζίνο με τα Δίκαια Στοιχήματα»*
- Το παιχνίδι συνίσταται στη ρίψη νομισμάτων, η οποία έχει μόνο δύο δυνατά αποτελέσματα: κορώνα (**H**) ή γράμματα (**T**).
- Το ιδανικό (**F**) νόμισμα θα έρθει κορώνα ή γράμματα με την ίδια πιθανότητα  $\frac{1}{2}$ .
- Το «πειραγμένο» (**B**) νόμισμα θα έρθει κορώνα με πιθανότητα  $\frac{3}{4}$ .

# Το «Καζίνο με τα Δίκαια Στοιχήματα»

(συνέχεια)

- Οπότε, ορίζουμε τις πιθανότητες:
  - $P(H|F) = P(T|F) = \frac{1}{2}$
  - $P(H|B) = \frac{3}{4}, P(T|B) = \frac{1}{4}$
  - Ο ανέντιμος κρουπιέρης αλλάζει το ιδανικό και το «πειραγμένο» νόμισμα με πιθανότητα 10%

## Το πρόβλημα του Καζίνου με τα Δίκαια Στοιχήματα

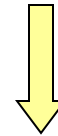
- **Είσοδος:** Η ακολουθία  $x = x_1 x_2 x_3 \dots x_n$  για τις ρίψεις του νομίσματος που δημιουργείται από δύο πιθανά νομίσματα ( $F$  ή  $B$ , ιδανικό ή πειραγμένο).
- **Έξοδος:** Η ακολουθία  $\pi = \pi_1 \pi_2 \pi_3 \dots \pi_n$ , όπου το κάθε  $\pi_i$  έχει την τιμή  $F$  ή  $B$ , γεγονός που υποδεικνύει ότι το  $x_i$  είναι το αποτέλεσμα που προκύπτει από τη ρίψη του ιδανικού ή του πειραγμένου νομίσματος, αντίστοιχα.

# Πρόβλημα...

*Το πρόβλημα του  
Καζίνου με τα Δίκαια  
Στοιχήματα*

Οποιοδήποτε  
παρατηρούμενο  
αποτέλεσμα ρίψεων θα  
μπορούσε να έχει  
παραχθεί από  
οποιαδήποτε ακολουθία  
καταστάσεων!

Πρέπει να βρούμε έναν  
τρόπο να δίνουμε  
διαφορετικές βαθμολογίες  
στις διαφορετικές ακολουθίες.



*Το πρόβλημα της  
Αποκωδικοποίησης*



## $P(x|\text{ιδανικό νόμισμα})$ και $P(x|\text{πειραγμένο νόμισμα})$

- Πρώτα, ας υποθέσουμε ότι ο κρουπιέρης δεν αλλάζει ποτέ νομίσματα. Μερικοί ορισμοί:
  - $P(x|\text{ιδανικό νόμισμα})$ : πιθανότητα ότι ο κρουπιέρης χρησιμοποίησε το νόμισμα  $F$  και παράγαγε το αποτέλεσμα  $x$ .
  - $P(x|\text{πειραγμένο νόμισμα})$ : πιθανότητα ότι ο κρουπιέρης χρησιμοποίησε το νόμισμα  $B$  και παράγαγε το αποτέλεσμα  $x$ .

## $P(x|\text{ιδανικό νόμισμα})$ και $P(x|\text{πειραγμένο νόμισμα})$

- $P(x|\text{ιδανικό νόμισμα}) = P(x_1 \dots x_n | \text{ιδανικό νόμισμα}) = \prod_{i=1,n} p(x_i | \text{ιδανικό νόμισμα}) = (1/2)^n$
- $P(x|\text{πειραγμένο νόμισμα}) = P(x_1 \dots x_n | \text{πειραγμένο νόμισμα}) = \prod_{i=1,n} p(x_i | \text{πειραγμένο νόμισμα}) = (3/4)^k (1/4)^{n-k} = 3^k / 4^n$
- $k$  - ο αριθμός των κορωνών ( $H$ ) στην ακολουθία  $x$ .

## $P(x|\text{ιδανικό νόμισμα})$ και $P(x|\text{πειραγμένο νόμισμα})$

- $P(x|\text{ιδανικό νόμισμα}) = P(x|\text{πειραγμένο νόμισμα})$

$$1/2^n = 3^k/4^n$$

$$2^n = 3^k$$

$$n = k \log_2 3$$

για  $k = n / \log_2 3$  ( $k \sim 0,67n$ )

# Λογαριθμικός λόγος πιθανοτήτων

- Ορίζουμε το **λογαριθμικό λόγο πιθανοτήτων** ως εξής:

$$\log_2(P(x|\text{ιδανικό νόμισμα}) / P(x|\text{πειραγμένο νόμισμα}))$$

$$= \sum_{i=1}^k \log_2(p^+(x_i) / p^-(x_i))$$

$$= n - k \log_2 3$$

## Υπολογισμός του λογαριθμικού λόγου πιθανοτήτων σε κυλιόμενα παράθυρα

$$x_1 x_2 \boxed{x_3 x_4 x_5 x_6 x_7} x_8 \dots x_n$$

Θεωρήστε ένα *κυλιόμενο παράθυρο* της ακολουθίας των αποτελεσμάτων. Βρείτε το λογαριθμικό λόγο πιθανοτήτων γι' αυτό το μικρό παράθυρο.



Μειονεκτήματα:

- Το μήκος του νησιδίου CG δεν είναι γνωστό εκ των προτέρων
- Διαφορετικά παράθυρα μπορεί να ταξινομήσουν την ίδια θέση διαφορετικά

# Κρυφό μοντέλο Markov (HMM)

- Μπορεί να θεωρηθεί ότι είναι μια αφηρημένη μηχανή με  $k$  κρυφές καταστάσεις που εκπέμπει σύμβολα από ένα αλφάβητο  $\Sigma$ .
- Κάθε κατάσταση έχει τη δική της κατανομή πιθανοτήτων, και η μηχανή μεταβαίνει στις καταστάσεις σύμφωνα με αυτή την κατανομή πιθανοτήτων.
- Όταν βρίσκεται σε μια ορισμένη κατάσταση, η μηχανή πρέπει να πάρει 2 αποφάσεις:
  - Σε ποια κατάσταση θα μεταβεί στη συνέχεια;
  - Ποιο σύμβολο – από ένα αλφάβητο  $\Sigma$  – θα εκπέμψει;

## Γιατί «κρυφό»;

- Οι παρατηρητές μπορούν να δουν τα εκπεμπόμενα σύμβολα ενός HMM αλλά δεν έχουν τη δυνατότητα να ξέρουν την κατάσταση στην οποία βρίσκεται το HMM αυτή τη στιγμή.
- Άρα, ο στόχος τους είναι να συμπεράνουν τις πιο πιθανές κρυφές καταστάσεις του HMM με βάση τη δεδομένη ακολουθία των συμβόλων που εκπέμπονται.

# Παράμετροι ενός HMM

**Σ**: σύνολο των χαρακτήρων που εκπέμπονται.

π.χ.,  $\Sigma = \{H, T\}$  για τη ρίψη νομισμάτων

$\Sigma = \{1, 2, 3, 4, 5, 6\}$  για τη ρίψη ζαριών

**Q**: σύνολο κρυφών καταστάσεων, καθεμία από τις οποίες θα εκπέμψει σύμβολα από το  $\Sigma$ .

$Q = \{F, B\}$  για τη ρίψη νομισμάτων



# Παράμετροι ενός HMM (συνέχεια)

$A = (a_{kl})$ : μια μήτρα διαστάσεων  $|Q| \times |Q|$  που περιγράφει την πιθανότητα να μεταβεί το HMM από την κατάσταση  $k$  στην κατάσταση  $l$ .

$$a_{FF} = 0,9 \quad a_{FB} = 0,1$$

$$a_{BF} = 0,1 \quad a_{BB} = 0,9$$

$E = (e_k(b))$ : μια μήτρα διαστάσεων  $|Q| \times |\Sigma|$  που περιγράφει την πιθανότητα να εκπέμψει το HMM το σύμβολο  $b$  ενώ βρίσκεται στην κατάσταση  $k$ .

$$e_F(0) = 1/2 \quad e_F(1) = 1/2$$

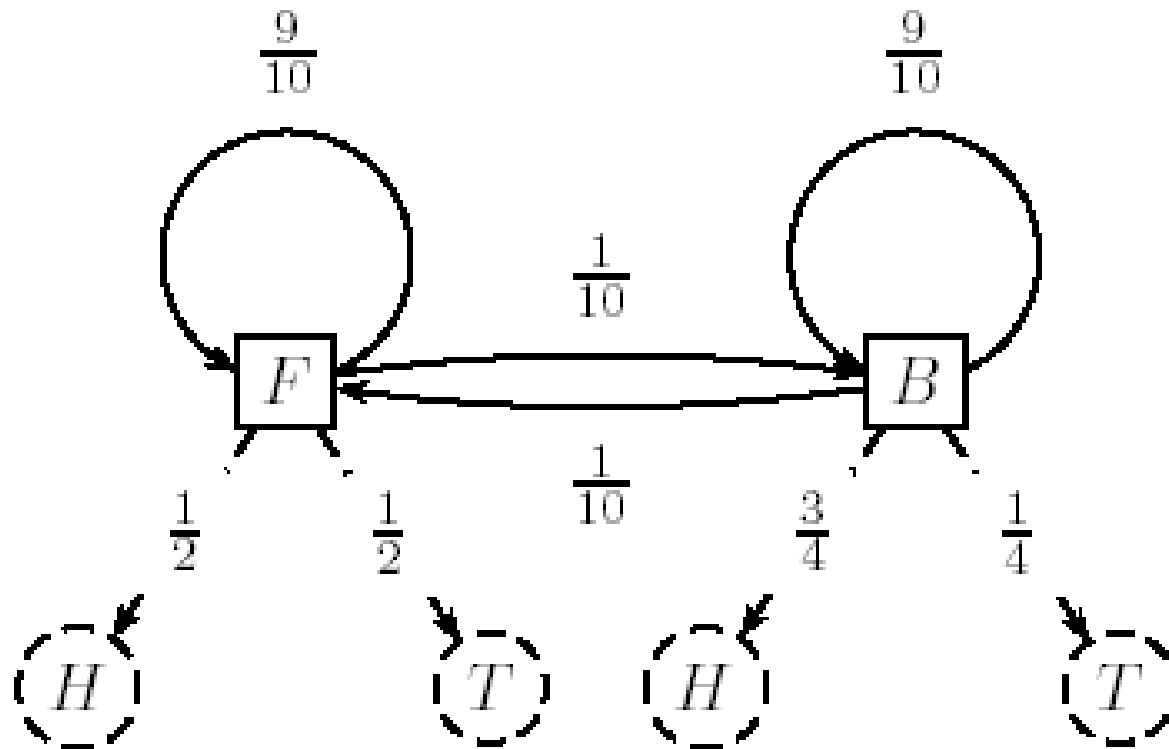
$$e_B(0) = 1/4 \quad e_B(1) = 3/4$$

# HMM για το Καζίνο με τα Δίκαια Στοιχήματα

- Το Καζίνο με τα Δίκαια Στοιχήματα ως HMM:  
 $\Sigma = \{0, 1\}$  (0 για γράμματα (**T**) και 1 για κορώνες (**H**))  
 $Q = \{F, B\}$  (F για ιδανικό και B για πειραγμένο νόμισμα)  
 Πιθανότητες μεταβολής κατάστασης A \*\*\* Πιθανότητες εκπομπής E

|            | Ιδανικό        | Πειραγμένο     |            | Γράμματα<br>(0)        | Κορώνες<br>(1)         |
|------------|----------------|----------------|------------|------------------------|------------------------|
| Ιδανικό    | $a_{FF} = 0,9$ | $a_{FB} = 0,1$ | Ιδανικό    | $e_F(0) = \frac{1}{2}$ | $e_F(1) = \frac{1}{2}$ |
| Πειραγμένο | $a_{BF} = 0,1$ | $a_{BB} = 0,9$ | Πειραγμένο | $e_B(0) = \frac{1}{4}$ | $e_B(1) = \frac{3}{4}$ |

# HMM για το Καζίνο με τα Δίκαια Στοιχήματα (συνέχεια)



Μοντέλο HMM για το πρόβλημα του Καζίνου με τα Δίκαια Στοιχήματα

# Κρυφές διαδρομές

- Μια **διαδρομή**  $\pi = \pi_1 \dots \pi_n$  στο HMM ορίζεται ως μια ακολουθία καταστάσεων.
- Θεωρήστε τη διαδρομή  $\pi = \text{FFFB BBBBFFF}$  και την ακολουθία  $x = 01011101001$

Πιθανότητα εκπομπής του  $x_i$  από την κατάσταση  $\pi_i$

|                                  |               |                |                |                |                |                |                |                |                |                |                |
|----------------------------------|---------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $x$                              | 0             | 1              | 0              | 1              | 1              | 1              | 0              | 1              | 0              | 0              | 1              |
| $\pi$                            | F             | F              | F              | B              | B              | B              | B              | B              | F              | F              | F              |
| $P(x_i \pi_i)$                   | $\frac{1}{2}$ | $\frac{1}{2}$  | $\frac{1}{2}$  | $\frac{3}{4}$  | $\frac{3}{4}$  | $\frac{3}{4}$  | $\frac{1}{4}$  | $\frac{3}{4}$  | $\frac{1}{2}$  | $\frac{1}{2}$  | $\frac{1}{2}$  |
| $P(\pi_{i-1} \rightarrow \pi_i)$ | $\frac{1}{2}$ | $\frac{9}{10}$ | $\frac{9}{10}$ | $\frac{1}{10}$ | $\frac{9}{10}$ | $\frac{9}{10}$ | $\frac{9}{10}$ | $\frac{9}{10}$ | $\frac{1}{10}$ | $\frac{9}{10}$ | $\frac{9}{10}$ |

Πιθανότητα μεταβολής από την κατάσταση  $\pi_{i-1}$  στην κατάσταση  $\pi_i$

# Υπολογισμός της $P(x|\pi)$

- $P(x|\pi)$ : η πιθανότητα παραγωγής της ακολουθίας  $x$  μέσα από τη διαδρομή  $\pi$ :

$$P(x|\pi) = P(\pi_0 \rightarrow \pi_1) \cdot \prod_{i=1}^n P(x_i | \pi_i) \cdot P(\pi_i \rightarrow \pi_{i+1})$$

$$= a_{\pi_0, \pi_1} \cdot \prod e_{\pi_i}(x_i) \cdot a_{\pi_i, \pi_{i+1}}$$

# Υπολογισμός της $P(x|\pi)$

- $P(x|\pi)$ : η πιθανότητα παραγωγής της ακολουθίας  $x$  μέσα από τη διαδρομή  $\pi$ :

$$P(x|\pi) = P(\pi_0 \rightarrow \pi_1) \cdot \prod_{i=1}^n P(x_i | \pi_i) \cdot P(\pi_i \rightarrow \pi_{i+1})$$

$$= a_{\pi_0, \pi_1} \cdot \prod e_{\pi_i}(x_i) \cdot a_{\pi_i, \pi_{i+1}}$$

$$= \prod e_{\pi_{i+1}}(x_{i+1}) \cdot a_{\pi_i, \pi_{i+1}}$$

αν μετράμε από  $i=0$  αντί για  $i=1$

# Το πρόβλημα της Αποκωδικοποίησης

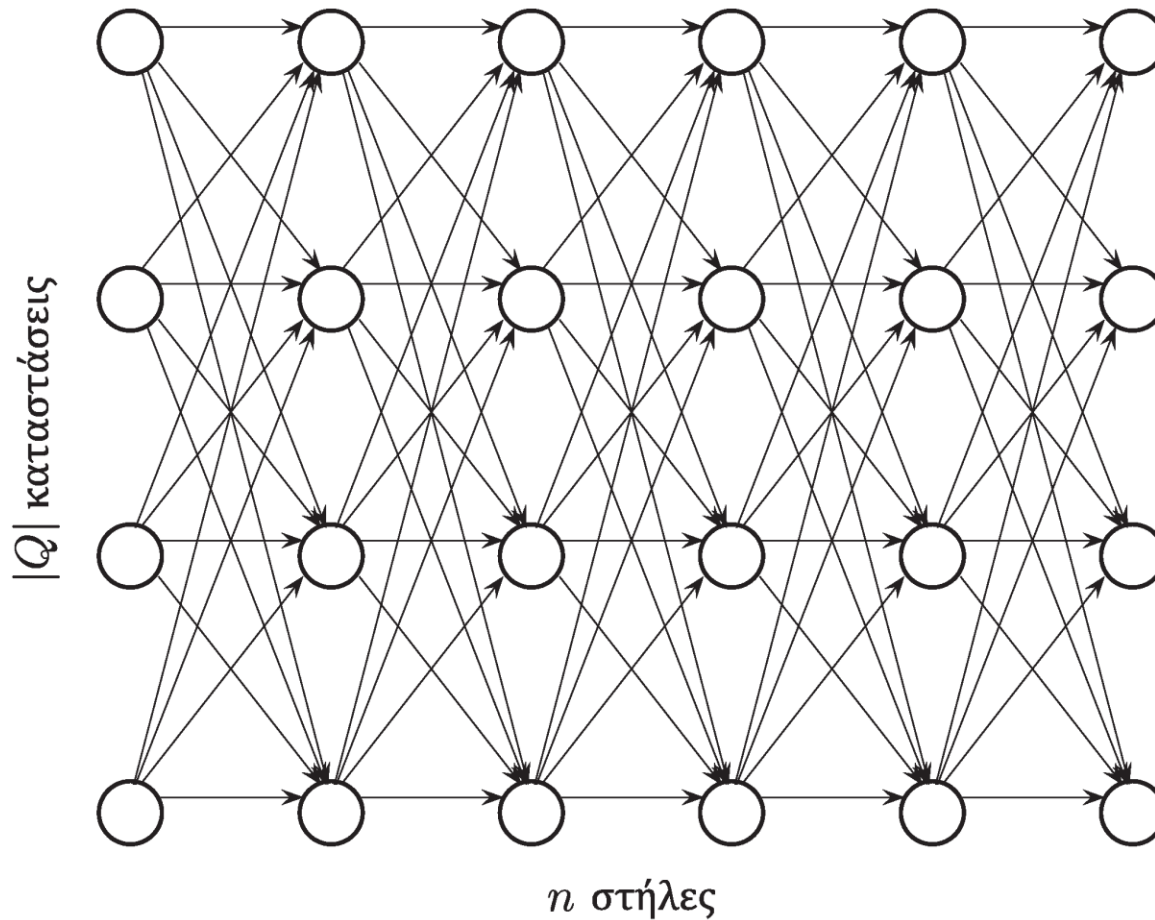
- **Στόχος:** Βρείτε τη βέλτιστη κρυφή διαδρομή καταστάσεων με δεδομένες παρατηρήσεις.
- **Είσοδος:** Η ακολουθία των παρατηρήσεων  $x = x_1 \dots x_n$  που παράγεται από το HMM  $M(\Sigma, Q, A, E)$ .
- **Έξοδος:** Η διαδρομή που μεγιστοποιεί την πιθανότητα  $P(x|\pi)$  για όλες τις δυνατές διαδρομές  $\pi$ .

## Κατασκευή πλέγματος Μανχάταν για το πρόβλημα της Αποκωδικοποίησης

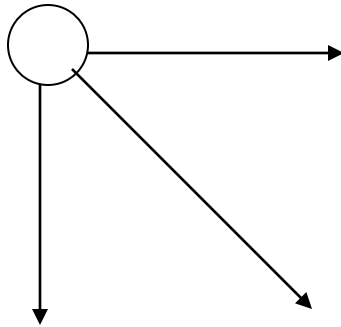
- Ο Andrew Viterbi χρησιμοποίησε το μοντέλο του πλέγματος Μανχάταν για να λύσει το *πρόβλημα της Αποκωδικοποίησης*.
- Κάθε επιλογή της  $\pi = \pi_1 \dots \pi_n$  αντιστοιχεί σε μια διαδρομή στο γράφημα.
- Η μοναδική έγκυρη κατεύθυνση στο γράφημα είναι προς τα *ανατολικά*.
- Το γράφημα αυτό έχει  $|Q|^2(n-1)$  ακμές.



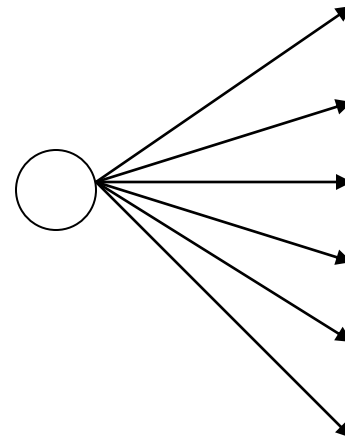
# Γράφημα μετασχηματισμού για το πρόβλημα της Αποκωδικοποίησης



## Το πρόβλημα της Αποκωδικοποίησης και το πρόβλημα της Στοίχισης



Έγκυρες κατευθύνσεις  
στο πρόβλημα της  
Στοίχισης.



Έγκυρες κατευθύνσεις  
στο πρόβλημα της  
Αποκωδικοποίησης.

## Το πρόβλημα της Αποκωδικοποίησης ως εύρεση της μεγαλύτερης διαδρομής σε ένα DAG

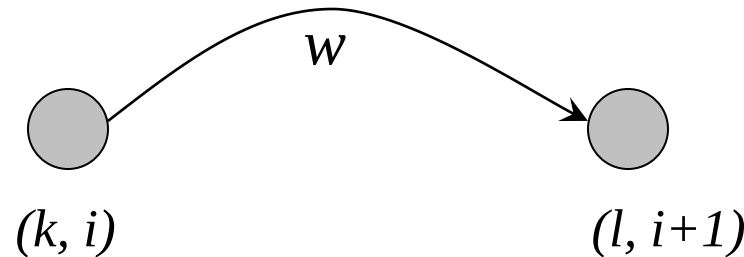
- Το πρόβλημα της Αποκωδικοποίησης ανάγεται στην εύρεση της μεγαλύτερης διαδρομής στο *κατευθυνόμενο ακυκλικό γράφημα (DAG)* παραπάνω.
- **Σημείωση:** το μήκος της διαδρομής ορίζεται ως το *γινόμενο* και όχι ως το *άθροισμα* των συντελεστών στάθμισης των ακμών της.

# Το πρόβλημα της Αποκωδικοποίησης

(συνέχεια)

- Κάθε διαδρομή στο γράφημα έχει την πιθανότητα  $P(x|\pi)$ .
- Ο αλγόριθμος του Viterbi βρίσκει τη διαδρομή που μεγιστοποιεί την  $P(x|\pi)$  με αναζήτηση στο χώρο όλων των πιθανών διαδρομών.
- Ο αλγόριθμος του Viterbi εκτελείται σε χρόνο  $O(n|Q|^2)$ .

## Το πρόβλημα της Αποκωδικοποίησης: συντελεστές στάθμισης των ακμών

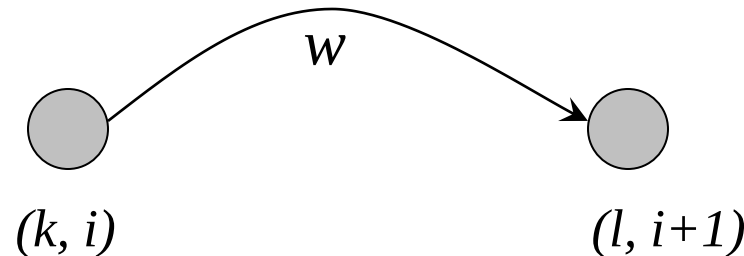


Ο συντελεστής στάθμισης  $w$  δίνεται από τη σχέση:

;

## Το πρόβλημα της Αποκωδικοποίησης: συντελεστές στάθμισης των ακμών

$$P(x|\pi) = \prod_{i=0}^{n-1} e^{\pi_{i+1}(x_{i+1})} \cdot a_{\pi_i, \pi_{i+1}}$$

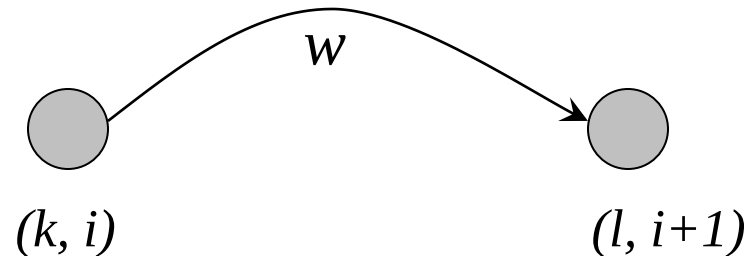


Ο συντελεστής στάθμισης  $w$  δίνεται από τη σχέση:

;;

## Το πρόβλημα της Αποκωδικοποίησης: συντελεστές στάθμισης των ακμών

$$i\text{-οστός όρος} = e_{\pi_{i+1}}(x_{i+1}) \cdot a_{\pi_i, \pi_{i+1}}$$

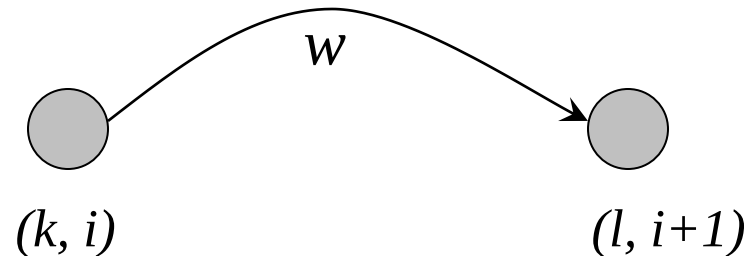


Ο συντελεστής στάθμισης  $w$  δίνεται από τη σχέση:

;

## Το πρόβλημα της Αποκωδικοποίησης: συντελεστές στάθμισης των ακμών

$i$ -οστός όρος =  $e_{\pi_i}(x_i) \cdot a_{\pi_i, \pi_{i+1}} = e_l(x_{i+1}) \cdot a_{kl}$  για  $\pi_i = k, \pi_{i+1} = l$



Ο συντελεστής στάθμισης  $w = e_l(x_{i+1}) \cdot a_{kl}$



# Το πρόβλημα της Αποκωδικοποίησης: δυναμικός προγραμματισμός

$$S_{l,i+1} = \max_{k \in Q} \{s_{k,i} \cdot \text{συντελεστής στάθμισης της ακμής μεταξύ } (k,i) \text{ και } (l,i+1)\} =$$

$$\max_{k \in Q} \{s_{k,i} \cdot a_{kl} \cdot e_l(x_{i+1})\} =$$

$$e_l(x_{i+1}) \cdot \max_{k \in Q} \{s_{k,i} \cdot a_{kl}\}$$

# Το πρόβλημα της Αποκωδικοποίησης

(συνέχεια)

- Αρχικές τιμές:
  - $s_{begin,0} = 1$
  - $s_{k,0} = 0$  για  $k \neq begin$
- Έστω ότι η  $\pi^*$  είναι η βέλτιστη διαδρομή.  
Τότε,

$$P(x|\pi^*) = \max_{k \in Q} \{s_{k,n} \cdot a_{k,end}\}$$

# Ο αλγόριθμος του Viterbi

- Η τιμή του γινομένου μπορεί να γίνει υπερβολικά μικρή, γεγονός που οδηγεί σε υπερχείλιση (overflow).
- Για να αποφευχθεί η υπερχείλιση, χρησιμοποιούνται λογαριθμικές τιμές.

$$s_{k,i+1} = \log_e(x_{i+1}) + \max_{k \in Q} \{s_{k,i} + \log(a_{kl})\}$$

## Το πρόβλημα «προς τα εμπρός-προς τα πίσω»

**Δίνεται:** μια ακολουθία ρίψεων ενός νομίσματος που παράγεται από ένα HMM.

**Στόχος:** βρείτε την πιθανότητα ότι ο κρουπιέρης χρησιμοποίησε «πειραγμένο» νόμισμα σε μια συγκεκριμένη χρονική στιγμή.

# Προς τα εμπρός αλγόριθμος

- Ορίζουμε την  $f_{k,i}$  (προς τα εμπρός πιθανότητα) ως την πιθανότητα εκπομπής του προθέματος  $x_1 \dots x_i$  και μετάβασης στην κατάσταση  $\pi_i = k$ .
- Η σχέση επανάληψης για τον προς τα εμπρός αλγόριθμο:

$$f_{k,i} = e_k(x_i) \cdot \sum_{l \in Q} f_{l,i-1} \cdot a_{lk}$$

# Προς τα πίσω αλγόριθμος

- Ωστόσο, η *προς τα εμπρός πιθανότητα* δεν είναι ο μοναδικός παράγοντας που επηρεάζει την πιθανότητα  $P(\pi_i = k|x)$ .
- Η ακολουθία των μεταβάσεων και των εκπομπών στις οποίες υποβάλλεται το HMM ανάμεσα στις καταστάσεις  $\pi_{i+1}$  και  $\pi_n$  επηρεάζει επίσης την πιθανότητα  $P(\pi_i = k|x)$ .



# Προς τα πίσω αλγόριθμος (συνέχεια)

- Ορίζουμε την *προς τα πίσω πιθανότητα*  $b_{k,i}$  ως την πιθανότητα να βρίσκεται το HMM στην κατάσταση  $\pi_i = k$  και να εκπέμπει το *επίθεμα*  $x_{i+1} \dots x_n$ .
- Η σχέση επανάληψης για τον *προς τα πίσω αλγόριθμο*:

$$b_{k,i} = \sum_{l \in Q} e_l(x_{i+1}) \cdot b_{l,i+1} \cdot a_{kl}$$

## Προς τα πίσω-προς τα εμπρός αλγόριθμος

- Η πιθανότητα να έχει ο κρουπιέρης πειραγμένο νόμισμα τη στιγμή  $i$ :

$$P(\pi_i = k|x) = \frac{P(x, \pi_i = k)}{P(x)} = \frac{f_k(i) \cdot b_k(i)}{P(x)}$$

Η  $P(x)$  είναι το άθροισμα των  $P(x, \pi_i = k)$  για όλα τα  $k$



## Εύρεση μακρινών μελών μιας οικογένειας πρωτεϊνών

- Με δεδομένη μια οικογένεια πρωτεϊνών με λειτουργικά σχετιζόμενες βιολογικές αλληλουχίες, τα «μακρινά ξαδέρφια» μπορεί να έχουν μικρές ομοιότητες (ανά ζεύγος) με κάθε μέλος της οικογένειας και έτσι να αποτύχουν στον έλεγχο της (στατιστικής) σημαντικότητας.
- Όμως, ενδέχεται να έχουν μικρές ομοιότητες με **πολλά** μέλη της οικογένειας.
- Ο στόχος είναι να στοιχίσουμε μια αλληλουχία με **όλα** τα μέλη της οικογένειας ταυτόχρονα.
- Μια οικογένεια σχετιζόμενων πρωτεϊνών μπορεί να αναπαρασταθεί από την πολλαπλή στοίχισή τους και το αντίστοιχο προφίλ.

## Αναπαράσταση οικογενειών πρωτεϊνών με προφίλ

- Οι στοιχισμένες αλληλουχίες DNA μπορούν να αναπαρασταθούν με μια μήτρα προφίλ διαστάσεων  $4 \cdot n$  που περιέχει τις συχνότητες των νουκλεοτιδίων σε κάθε στοιχισμένη θέση.

|          |      |      |      |      |      |      |      |      |
|----------|------|------|------|------|------|------|------|------|
| <b>A</b> | 0,72 | 0,14 | 0    | 0    | 0,72 | 0,72 | 0    | 0    |
| <b>T</b> | 0,14 | 0,72 | 0    | 0    | 0    | 0,14 | 0,14 | 0,86 |
| <b>G</b> | 0,14 | 0,14 | 0,86 | 0,44 | 0    | 0,14 | 0    | 0    |
| <b>C</b> | 0    | 0    | 0,14 | 0,56 | 0,28 | 0    | 0,86 | 0,14 |

- Η οικογένεια πρωτεϊνών μπορεί να αναπαρασταθεί από ένα προφίλ διαστάσεων  $20 \cdot n$  που δείχνει τις συχνότητες των αμινοξέων.

# Προφίλ και HMM

- Τα HMM μπορούν επίσης να χρησιμοποιηθούν για τη στοίχιση μιας αλληλουχίας με ένα προφίλ που αναπαριστά μια οικογένεια πρωτεϊνών.
- Ένα προφίλ  $P$  διαστάσεων  $20 \cdot n$  αντιστοιχεί σε  $n$  σειριακά συνδεδεμένες καταστάσεις ταιριάσματος  $M_1, \dots, M_n$  στο HMM του προφίλ  $P$ .

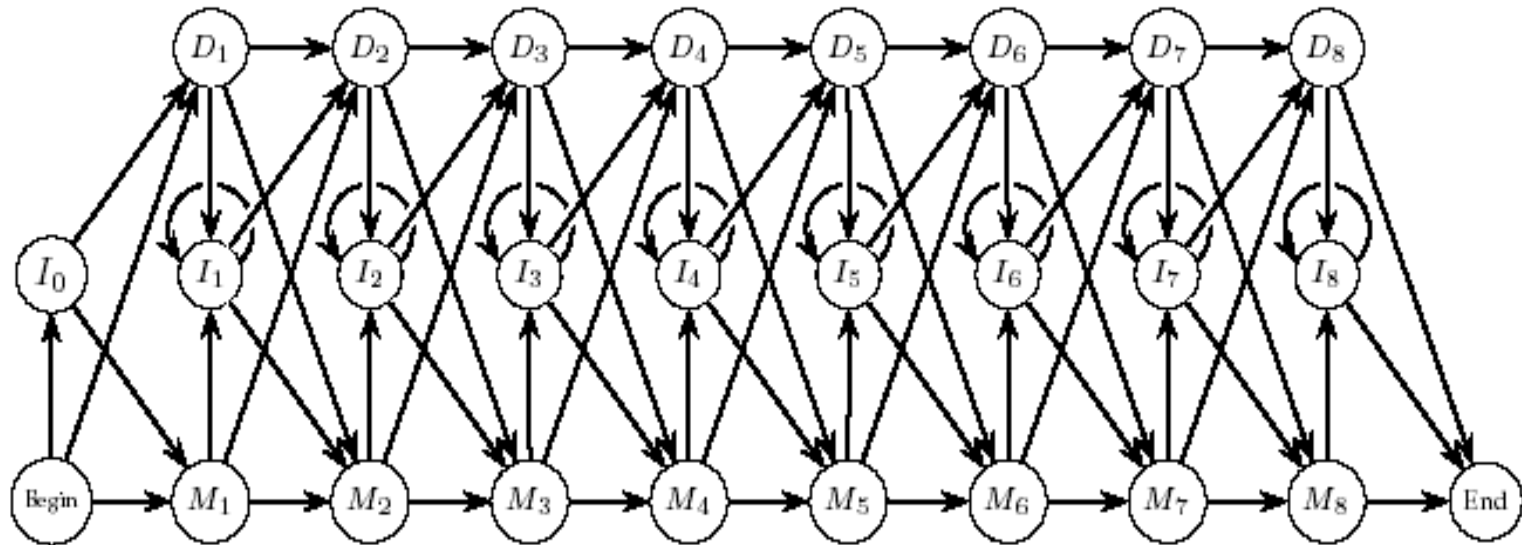
## Πολλαπλές στοιχίσεις και ταξινόμηση οικογενειών πρωτεϊνών

- Η πολλαπλή στοίχιση μιας οικογένειας πρωτεϊνών δείχνει τις διαφορές στη συντήρηση κατά μήκος μιας πρωτεΐνης
- Παράδειγμα: μετά από τη στοίχιση πολλών πρωτεϊνών σφαιρινών, οι βιολόγοι αναγνώρισαν ότι οι περιοχές των ελίκων στις σφαιρίνες είναι περισσότερο συντηρημένες από άλλες περιοχές.

# Τι είναι το HMM ενός προφίλ;

- Το HMM ενός προφίλ είναι η αναπαράσταση μιας πολλαπλής στοίχισης με πιθανότητες.
- Μια δεδομένη πολλαπλή στοίχιση (μιας οικογένειας πρωτεϊνών) χρησιμοποιείται για τη δημιουργία ενός HMM για το προφίλ.
- Το μοντέλο αυτό μπορεί να χρησιμοποιηθεί στη συνέχεια για την εύρεση και τη βαθμολόγηση λιγότερο προφανών πιθανών ταιριασμάτων για νέες πρωτεϊνικές αλληλουχίες.

# Το HMM ενός προφίλ



Το HMM ενός προφίλ

# Δημιουργία HMM για προφίλ

- Χρησιμοποιούμε την πολλαπλή στοίχιση για την κατασκευή του μοντέλου HMM.
- Αντιστοιχίζουμε μια κατάσταση *Ταιριάσματος* του HMM σε κάθε στήλη. Προσθέτουμε τις καταστάσεις *Προσθήκη* και *Αφαίρεση*.
- Εκτιμάμε τις πιθανότητες εκπομπής με βάση το πλήθος των αμινοξέων στη στήλη. Διαφορετικές θέσεις στην πρωτεΐνη θα έχουν διαφορετικές πιθανότητες εκπομπής.
- Εκτιμάμε τις πιθανότητες μεταβολής κατάστασης για τις καταστάσεις *Ταιριάσματος*, *Αφαίρεσης* και *Προσθήκης*.
- Το μοντέλο HMM εκπαιδεύεται για να βρίσκει τις βέλτιστες παραμέτρους.

# Καταστάσεις ενός HMM για προφίλ

- Καταστάσεις ταιριάσματος  $M_1 \dots M_n$  (συν καταστάσεις *begin/end*)
- Καταστάσεις προσθήκης  $I_0 I_1 \dots I_n$
- Καταστάσεις αφαίρεσης  $D_1 \dots D_n$



## Πιθανότητες μεταβολής κατάστασης σε ένα HMM για προφίλ

- $\log(a_{MI}) + \log(a_{IM}) =$  ποινή ενεργοποίησης (εισαγωγής) κενού
- $\log(a_{II}) =$  ποινή επέκτασης κενού

## Πιθανότητες μεταβολής κατάστασης σε ένα HMM για προφίλ

Πιθανότητα εκπομπής ενός συμβόλου  $a$  σε μια κατάσταση προσθήκης  $I_j$ :

$$e_{I_j}(a) = p(a)$$

όπου  $p(a)$  είναι η συχνότητα εμφάνισης του συμβόλου  $a$  σε όλες τις αλληλουχίες.

# Στοίχιση προφίλ με HMM

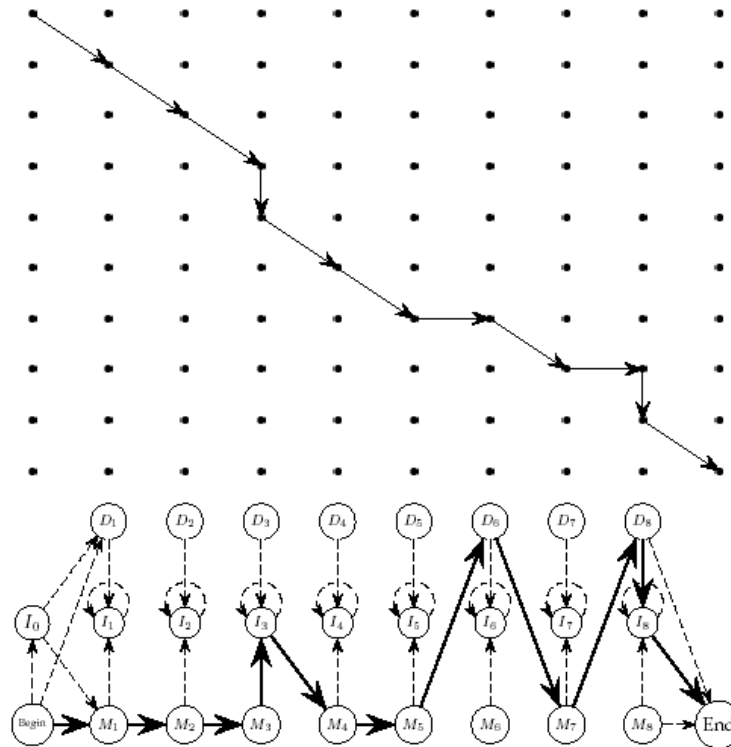
- Ορίζουμε το  $v_j^M(i)$  ως τη λογαριθμική βαθμολογία πιθανοφάνειας της καλύτερης διαδρομής για το ταίριασμα της συμβολοσειράς  $x_1..x_i$  με το HMM ενός προφίλ που τελειώνει με την εκπομπή του συμβόλου  $x_i$  από την κατάσταση  $M_j$ .
- Οι βαθμολογίες  $v_j^I(i)$  και  $v_j^D(i)$  ορίζονται με παρόμοιο τρόπο.

## Στοίχιση προφίλ με HMM: δυναμικός προγραμματισμός

$$v_j^M(i) = \log(e_{M_j}(x_i)/p(x_i)) + \max \begin{cases} v_{j-1}^M(i-1) + \log(a_{M_{j-1}, M_j}) \\ v_{j-1}^I(i-1) + \log(a_{I_{j-1}, M_j}) \\ v_{j-1}^D(i-1) + \log(a_{D_{j-1}, M_j}) \end{cases}$$

$$v_j^I(i) = \log(e_{I_j}(x_i)/p(x_i)) + \max \begin{cases} v_j^M(i-1) + \log(a_{M_j, I_j}) \\ v_j^I(i-1) + \log(a_{I_j, I_j}) \\ v_j^D(i-1) + \log(a_{D_j, I_j}) \end{cases}$$

## Διαδρομές στο γράφημα μετασχηματισμού και στο HMM ενός προφίλ



Μια διαδρομή μέσα από γράφημα μετασχηματισμού, και η αντίστοιχη διαδρομή μέσα από το HMM ενός προφίλ

## Δημιουργία συλλογής HMM για οικογένειες πρωτεϊνών

- Χρησιμοποιούμε το Blast για να χωρίσουμε μια βάση δεδομένων πρωτεϊνών σε οικογένειες σχετιζόμενων πρωτεϊνών
- Κατασκευάζουμε μια πολλαπλή στοίχιση για κάθε οικογένεια πρωτεϊνών.
- Κατασκευάζουμε ένα μοντέλο HMM για προφίλ και βελτιστοποιούμε τις παραμέτρους του μοντέλου (πιθανότητες μεταβολής κατάστασης και εκπομπής).
- Στοιχίζουμε την αλληλουχία-στόχο με κάθε HMM για να βρούμε την καλύτερη προσαρμογή μεταξύ μιας αλληλουχίας-στόχου και ενός HMM

## Εφαρμογή των HMM για προφίλ στη μοντελοποίηση πρωτεϊνών σφαιρινών

- Οι σφαιρίνες αντιπροσωπεύουν μια μεγάλη συλλογή πρωτεϊνικών αλληλουχιών.
- Από το σύνολο των σφαιρινών επιλέχθηκαν τυχαία 400, οι οποίες χρησιμοποιήθηκαν για την κατασκευή μιας πολλαπλής στοίχισης.
- Η πολλαπλή στοίχιση χρησιμοποιήθηκε για την αντιστοίχιση ενός αρχικού HMM.
- Στη συνέχεια, το μοντέλο εκπαιδεύτηκε επανειλημμένα με μήκη τυχαία επιλεγμένα μεταξύ των τιμών 145 και 170, για να προκύψει ένα μοντέλο HMM με βελτιστοποιημένες πιθανότητες.

## Πόσο καλό είναι το HMM για τη σφαιρίνη;

- Οι υπόλοιπες 625 αλληλουχίες σφαιρινών της βάσης δεδομένων στοιχίστηκαν με το HMM που κατασκευάστηκε, με αποτέλεσμα να προκύψει μια πολλαπλή στοίχιση. Αυτή η πολλαπλή στοίχιση εμφανίζει υπερβολικά μεγάλη συμφωνία με τη στοίχιση που προέκυψε δομικά.
- 25.044 πρωτεΐνες επιλέχθηκαν τυχαία από τη βάση δεδομένων και συγκρίθηκαν με το HMM για τη σφαιρίνη.
- Το πείραμα είχε σαν αποτέλεσμα τον τέλειο διαχωρισμό της οικογένειας των σφαιρινών από αυτή των μη σφαιρινών.



# Η βάση δεδομένων PFAM

- Η βάση Pfam περιγράφει **οικογένειες πρωτεϊνών**
- Κάθε οικογένεια πρωτεϊνών στην Pfam έχει:
  - Μια στοίχιση με φύτρο: πολλαπλή στοίχιση ενός αντιπροσωπευτικού συνόλου αλληλουχιών η οποία έχει επαληθευτεί με μη αυτόματο τρόπο.
  - Ένα *HMM* κατασκευασμένο από τη στοίχιση με φύτρο για επιπλέον αναζητήσεις στη βάση δεδομένων.
  - Μια *πλήρη στοίχιση* που έχει *παραχθεί αυτόματα* από το HMM
- Η διάκριση μεταξύ των στοιχίσεων με φύτρο και των πλήρων στοιχίσεων διευκολύνει την ενημέρωση της βάσης Pfam.
  - Οι στοιχίσεις με φύτρο είναι σταθεροί πόροι.
  - Τα HMM των προφίλ και οι πλήρεις στοιχίσεις μπορούν να ενημερώνονται καθώς προσδιορίζονται νέες αμινοξικές αλληλουχίες.

# Χρήσεις της PFAM

- Τα HMM της Pfam καλύπτουν ολόκληρες περιοχές (πρωτεϊνών) που περιλαμβάνουν περισσότερο συντηρημένα μοτίβα αλλά και λιγότερο συντηρημένες περιοχές με προσθήκες και αφαιρέσεις.
- Έχουν ως αποτέλεσμα τη μοντελοποίηση ολόκληρων περιοχών, η οποία διευκολύνει τον καλύτερο σχολιασμό των αλληλουχιών και οδηγεί σε μια πιο ευαίσθητη ανίχνευση.

# Εκτίμηση παραμέτρων HMM

- Μέχρι τώρα, υποθέσαμε ότι οι πιθανότητες μεταβολής κατάστασης και εκπομπής είναι γνωστές.
- Ωστόσο, στις περισσότερες εφαρμογές των HMM, οι πιθανότητες δεν είναι γνωστές. Η εκτίμησή τους είναι πολύ δύσκολη.

# Το πρόβλημα της Εκτίμησης Παραμέτρων HMM

- Δίνονται
  - Το HMM με καταστάσεις και αλφάβητο (χαρακτήρες εκπομπής)
  - Ανεξάρτητες ακολουθίες εκπαίδευσης  $x^1, \dots, x^m$
- Βρείτε τις παραμέτρους  $\Theta$  του HMM (δηλαδή τα  $a_{kl}, e_k(b)$ ) που μεγιστοποιούν την

$$P(x^1, \dots, x^m \mid \Theta)$$

δηλαδή την από κοινού πιθανότητα των ακολουθιών εκπαίδευσης.

# Μεγιστοποίηση της πιθανότητας

- Η  $P(x^1, \dots, x^m | \Theta)$  ως συνάρτηση του  $\Theta$  ονομάζεται **πιθανοφάνεια** του μοντέλου
- Οι ακολουθίες εκπαίδευσης θεωρούνται ανεξάρτητες, συνεπώς

$$P(x^1, \dots, x^m | \Theta) = \prod_i P(x^i | \Theta)$$

- Το πρόβλημα της εκτίμησης των παραμέτρων ψάχνει τις παραμέτρους  $\Theta$  για τις οποίες ισχύει

$$\max_{\Theta} \prod_i P(x^i | \Theta)$$

- Στην πράξη, υπολογίζεται η **λογαριθμική πιθανότητα** για να αποφευχθούν τα σφάλματα υποχείλισης

# Δύο περιπτώσεις

Γνωστές διαδρομές για τις ακολουθίες εκπαίδευσης

- Τα νησίδα CpG σημαδεύονται στις ακολουθίες εκπαίδευσης
- Ένα βράδυ, ο κρουπιέρης στο καζίνο μας επιτρέπει να δούμε πότε αλλάζει ζάρια

Άγνωστες διαδρομές

- Τα νησίδα CpG δεν είναι σημαδεμένα
- Δεν μπορούμε να δούμε πότε αλλάζει ζάρια ο κρουπιέρης στο καζίνο

# Γνωστές διαδρομές

$A_{kl}$  = # φορές που συμβαίνει κάθε μετάβαση  $k \rightarrow l$   
στις ακολουθίες εκπαίδευσης

$E_k(b)$  = # φορές που το σύμβολο  $b$  εκπέμπεται από  
την κατάσταση  $k$  στις ακολουθίες εκπαίδευσης

Υπολογίζουμε τα  $a_{kl}$  και  $e_k(b)$  ως εκτιμητές μέγιστης  
πιθανοφάνειας:

$$a_{kl} = A_{kl} / \sum_{l'} A_{kl'}$$

$$e_k(b) = E_k(b) / \sum_{b'} E_k(b')$$

# Ψευδοπλήθη

- Κάποια κατάσταση  $k$  μπορεί να μην εμφανιστεί στις ακολουθίες εκπαίδευσης. Αυτό σημαίνει ότι ισχύει  $A_{kl} = 0$  για κάθε κατάσταση  $l$  και το  $a_{kl}$  δεν μπορεί να υπολογιστεί με τη δεδομένη εξίσωση.
- Για να αποφύγουμε την **υπερπροσαρμογή** (overfitting), χρησιμοποιούμε προκαθορισμένα **ψευδοπλήθη**  $r_{kl}$  και  $r_k(b)$ .

$$A_{kl} = \# \text{ μεταβάσεων } k \rightarrow l + r_{kl}$$

$$E_k(b) = \# \text{ εκπομπών του } b \text{ από την } k + r_k(b)$$

- Τα ψευδοπλήθη αντικατοπτρίζουν τις εκ των προτέρων μεροληψίες μας σχετικά με τις τιμές των πιθανοτήτων.



## Άγνωστες διαδρομές: εκπαίδευση Viterbi

- Ιδέα: θα χρησιμοποιήσουμε την αποκωδικοποίηση Viterbi για να υπολογίσουμε **την πιο πιθανή διαδρομή** για την ακολουθία εκπαίδευσης  $x$
- Ξεκινάμε με κάποια εικασία για τις αρχικές παραμέτρους και υπολογίζουμε την πιο πιθανή διαδρομή  $\pi^*$  για τη  $x$  με χρήση των αρχικών παραμέτρων
- Επαναλαμβάνουμε μέχρι να μην υπάρχει αλλαγή στην  $\pi^*$  :
  1. Προσδιορίζουμε τις τιμές των  $A_{kl}$  και  $E_k(b)$  όπως προηγουμένως
  2. Υπολογίζουμε τις νέες παραμέτρους  $a_{kl}$  και  $e_k(b)$  με τον ίδιο τύπο που χρησιμοποιήσαμε προηγουμένως
  3. Υπολογίζουμε τη νέα διαδρομή  $\pi^*$  για τη  $x$  και τις τρέχουσες παραμέτρους

# Ανάλυση της εκπαίδευσης Viterbi

- Ο αλγόριθμος **συγκλίνει ακριβώς**  
Υπάρχει πεπερασμένο πλήθος διαδρομών.  
Οι νέες παράμετροι προσδιορίζονται μοναδικά από την τρέχουσα διαδρομή  $\pi^*$ .  
Ενδέχεται να υπάρχουν πολλές διαδρομές για την ακολουθία  $x$  που έχουν την ίδια πιθανότητα, άρα πρέπει να συγκρίνουμε τη νέα  $\pi^*$  με όλες τις προηγούμενες διαδρομές που έχουν τη μεγαλύτερη πιθανότητα.
- **Δεν μεγιστοποιεί την πιθανοφάνεια**  $\prod_x P(x | \Theta)$  αλλά τη συνεισφορά της πιο πιθανής διαδρομής  $\prod_x P(x | \Theta, \pi^*)$  στην πιθανοφάνεια
- Γενικά, έχει χειρότερη απόδοση από τον αλγόριθμο Baum-Welch

# Άγνωστες διαδρομές: Baum-Welch

Ιδέα:

1. Μαντεύουμε τις αρχικές τιμές των παραμέτρων.

τέχνη και εμπειρία, όχι επιστήμη

2. Εκτιμάμε τις νέες (καλύτερες) τιμές των παραμέτρων.

Πώς;

3. Επαναλαμβάνουμε μέχρι να ικανοποιηθούν τα κριτήρια τερματισμού.

Ποια είναι τα κριτήρια;

## Καλύτερες τιμές των παραμέτρων

- Χρειαζόμαστε τις τιμές  $A_{kl}$  και  $E_k(b)$  αλλά δεν μπορούμε να μετρήσουμε (η διαδρομή είναι άγνωστη) και δεν θέλουμε να χρησιμοποιήσουμε την πιο πιθανή διαδρομή.
- Για όλες τις καταστάσεις  $k, l$ , το σύμβολο  $b$  και την ακολουθία εκπαίδευσης  $x$

Υπολογίζουμε τις  $A_{kl}$  και  $E_k(b)$  ως **αναμενόμενες τιμές**, με δεδομένες τις τρέχουσες παραμέτρους

# Σημειογραφία

- Για οποιαδήποτε ακολουθία χαρακτήρων  $x$  που εκπέμπονται κατά μήκος κάποιας άγνωστης διαδρομής  $\pi$ , συμβολίζουμε με  $\pi_i = k$  την υπόθεση ότι η κατάσταση στη θέση  $i$  (όπου εκπέμπεται ο χαρακτήρας  $x_i$ ) είναι η  $k$ .

## Πιθανοκρατικό περιβάλλον για την τιμή $A_{k,l}$

- Με δεδομένες τις ακολουθίες  $x^1, \dots, x^m$ , θεωρήστε ένα **χώρο διακριτών πιθανοτήτων** με τα **στοιχειώδη ενδεχόμενα**

$$\varepsilon_{k,l} = "k \rightarrow l \text{ συμβαίνει στις } x^1, \dots, x^m"$$

- Για κάθε  $x$  στο  $\{x^1, \dots, x^m\}$  και κάθε θέση  $i$  στη  $x$ , έστω ότι η  $Y_{x,i}$  είναι μια **τυχαία μεταβλητή** που ορίζεται ως

$$Y_{x,i}(\varepsilon_{k,l}) = \begin{cases} 1, & \text{αν } \pi_i = k \text{ και } \pi_{i+1} = l \\ 0, & \text{διαφορετικά} \end{cases}$$

- Ορίζουμε την  $Y = \sum_x \sum_i Y_{x,i}$  ως την τυχαία μεταβλητή που μετράει τον # των φορών που συμβαίνει το ενδεχόμενο  $\varepsilon_{k,l}$  στις  $x^1, \dots, x^m$ .

## Η σημασία της τιμής $A_{kl}$

- Έστω ότι η  $A_{kl}$  είναι η αναμενόμενη τιμή (ή μαθηματική ελπίδα) της  $Y$

$$\begin{aligned} E(Y) &= \sum_x \sum_i E(Y_{x,i}) = \sum_x \sum_i P(Y_{x,i} = 1) = \\ &= \sum_x \sum_i P(\{\varepsilon_{k,l} \mid \pi_i = k \text{ και } \pi_{i+1} = l\}) = \\ &= \sum_x \sum_i P(\pi_i = k, \pi_{i+1} = l \mid x) \end{aligned}$$

- Πρέπει να υπολογίσουμε την πιθανότητα  $P(\pi_i = k, \pi_{i+1} = l \mid x)$

## Πιθανοκρατικό περιβάλλον για την τιμή $E_k(b)$

- Με δεδομένες τις ακολουθίες  $x^1, \dots, x^m$ , θεωρήστε ένα **χώρο διακριτών πιθανοτήτων** με τα **στοιχειώδη ενδεχόμενα**

$\varepsilon_{k,b}$  = “το  $b$  εκπέμπεται στην κατάσταση  $k$  στις  $x^1, \dots, x^m$ ”

- Για κάθε  $x$  στο  $\{x^1, \dots, x^m\}$  και κάθε θέση  $i$  στη  $x$ , έστω ότι η  $Y_{x,i}$  είναι μια **τυχαία μεταβλητή** που ορίζεται ως

$$Y_{x,i}(\varepsilon_{k,b}) = \begin{cases} 1, & \text{αν } x_i = b \text{ και } \pi_i = k \\ 0, & \text{διαφορετικά} \end{cases}$$

- Ορίζουμε την  $Y = \sum_x \sum_i Y_{x,i}$  ως την τυχαία μεταβλητή που μετράει τον # των φορών που συμβαίνει το ενδεχόμενο  $\varepsilon_{k,b}$  στις  $x^1, \dots, x^m$ .



## Η σημασία της τιμής $E_k(b)$

- Έστω ότι η  $E_k(b)$  είναι η **αναμενόμενη τιμή της  $Y$**

$$E(Y) = \sum_x \sum_i E(Y_{x,i}) = \sum_x \sum_i P(Y_{x,i} = 1) = \sum_x \sum_i P(\{\varepsilon_{k,b} \mid x_i = b \text{ και } \pi_i = k\})$$

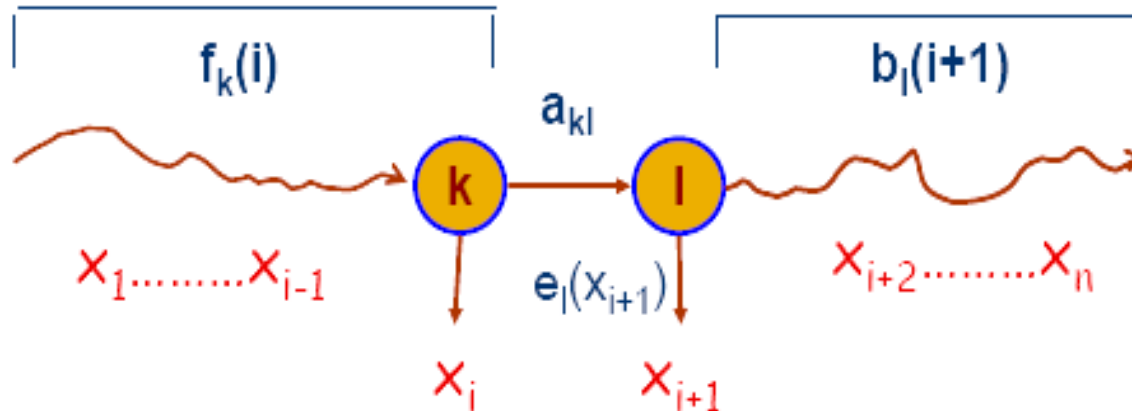
$$\sum_x \sum_{\{i \mid x_i = b\}} P(\{\varepsilon_{k,b} \mid x_i = b, \pi_i = k\}) = \sum_x \sum_{\{i \mid x_i = b\}} P(\pi_i = k \mid x)$$

- Πρέπει να υπολογίσουμε την πιθανότητα  **$P(\pi_i = k \mid x)$**

# Υπολογισμός νέων παραμέτρων

Θεωρήστε την ακολουθία εκπαίδευσης  $x = x_1 \dots x_n$

Επικεντρώνουμε την προσοχή μας στις θέσεις  $i$  και  $i+1$



Χρησιμοποιούμε τις προς τα εμπρός-προς τα πίσω τιμές:

$$f_{ki} = P(x_1 \dots x_i, \pi_i = k)$$

$$b_{ki} = P(x_{i+1} \dots x_n \mid \pi_i = k)$$

# Υπολογισμός της $A_{kl}$ (1)

- Η πιθανότητα να συμβαίνει η μετάβαση  $k \rightarrow l$  στη θέση  $i$  της  $x$   
$$P(\pi_i = k, \pi_{i+1} = l \mid x_1 \dots x_n) = P(x, \pi_i = k, \pi_{i+1} = l) / P(x)$$
- Υπολογίζουμε την  $P(x)$  είτε με προς τα εμπρός είτε με προς τα πίσω τιμές
- Θα δείξουμε ότι  $P(x, \pi_i = k, \pi_{i+1} = l) = b_{li+1} \cdot e_l(x_{i+1}) \cdot a_{kl} \cdot f_{ki}$
- Αναμενόμενος # φορών που συμβαίνει η μετάβαση  $k \rightarrow l$  στις ακολουθίες εκπαίδευσης

$$A_{kl} = \sum_x \sum_i (b_{li+1} \cdot e_l(x_{i+1}) \cdot a_{kl} \cdot f_{ki}) / P(x)$$

## Υπολογισμός της $A_{kl}$ (2)

$$\begin{aligned}
 &P(x, \pi_i = k, \pi_{i+1} = l) = \\
 &P(x_1 \dots x_i, \pi_i = k, \pi_{i+1} = l, x_{i+1} \dots x_n) = \\
 &P(\pi_{i+1} = l, x_{i+1} \dots x_n \mid x_1 \dots x_i, \pi_i = k) \cdot P(x_1 \dots x_i, \pi_i = k) = \\
 &P(\pi_{i+1} = l, x_{i+1} \dots x_n \mid \pi_i = k) \cdot f_{ki} = \\
 &P(x_{i+1} \dots x_n \mid \pi_i = k, \pi_{i+1} = l) \cdot P(\pi_{i+1} = l \mid \pi_i = k) \cdot f_{ki} = \\
 &P(x_{i+1} \dots x_n \mid \pi_{i+1} = l) \cdot a_{kl} \cdot f_{ki} = \\
 &P(x_{i+2} \dots x_n \mid x_{i+1}, \pi_{i+1} = l) \cdot P(x_{i+1} \mid \pi_{i+1} = l) \cdot a_{kl} \cdot f_{ki} = \\
 &P(x_{i+2} \dots x_n \mid \pi_{i+1} = l) \cdot e_l(x_{i+1}) \cdot a_{kl} \cdot f_{ki} = \\
 &b_{li+1} \cdot e_l(x_{i+1}) \cdot a_{kl} \cdot f_{ki}
 \end{aligned}$$

# Υπολογισμός της $E_k(b)$

- Η πιθανότητα εκπομπής του  $x_i$  από τη  $x$  στην κατάσταση  $k$

$$P(\pi_i = k \mid x_1 \dots x_n) = P(\pi_i = k, x_1 \dots x_n) / P(x)$$

$$\begin{aligned} P(\pi_i = k, x_1 \dots x_n) &= P(x_1 \dots x_i, \pi_i = k, x_{i+1} \dots x_n) = \\ &P(x_{i+1} \dots x_n \mid x_1 \dots x_i, \pi_i = k) \cdot P(x_1 \dots x_i, \pi_i = k) = \\ &P(x_{i+1} \dots x_n \mid \pi_i = k) \cdot f_{ki} = b_{ki} \cdot f_{ki} \end{aligned}$$

- Αναμενόμενος # φορών που εκπέμπεται το  $b$  στην κατάσταση  $k$

$$E_k(b) = \sum_x \sum_{i: x_i=b} (f_{ki} \cdot b_{ki}) / P(x)$$

## Τελικά, νέοι παράμετροι

$$a_{kl} = A_{kl} / \sum_{l'} A_{kl'}$$

$$e_k(b) = E_k(b) / \sum_{b'} E_k(b')$$

Μπορούμε να προσθέσουμε ψευδοπλήθη όπως προηγουμένως.

## Κριτήρια τερματισμού

- Στην πραγματικότητα, δεν μπορούμε να φτάσουμε στο μέγιστο (βελτιστοποίηση συνεχών συναρτήσεων)
- Άρα, χρειαζόμαστε κριτήρια τερματισμού
  - Υπολογισμός της λογαριθμικής πιθανότητας του μοντέλου για τις τρέχουσες παραμέτρους  $\Theta$

$$\sum_x \log P(x | \Theta)$$

- Σύγκριση με την προηγούμενη λογαριθμική πιθανότητα
- Τερματισμός αν η διαφορά είναι μικρή
- Τερματισμός μετά από συγκεκριμένο αριθμό επαναλήψεων

# Ο αλγόριθμος Baum-Welch

## Αρχικές τιμές:

Επιλογή της καλύτερης εικασίας για τις παραμέτρους του μοντέλου (ή τυχαία)

## Επανάληψη:

1. Προς τα εμπρός για κάθε  $x$
2. Προς τα πίσω για κάθε  $x$
3. Υπολογισμός των  $A_{kl}, E_k(b)$
4. Υπολογισμός νέων  $a_{kl}, e_k(b)$
5. Υπολογισμός νέας λογαριθμικής πιθανότητας

Μέχρι η λογαριθμική πιθανότητα να μην αλλάζει πολύ



# Βιβλιογραφία

- Διαφάνειες για το μάθημα CS 262 στο Στάνφορντ, το οποίο διδάσκει ο Serafim Batzoglou