ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ AGRICULTURAL UNIVERSITY OF ATHENS

Βιοπληροφορική

Ενότητα 14:

Μοντέλα Πολλαπλής Στοίχισης (2/2), 1.5ΔΩ

Τμήμα: Βιοτεχνολογίας

Όνομα καθηγητή: **Τ. Θηραίου**

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Μαθησιακοί Στόχοι

- παρουσίαση των μοντέλων πολλαπλής στοίχισης.
- κατανόηση των εφαρμογών και των περιορισμών τους.

Λέξεις Κλειδιά

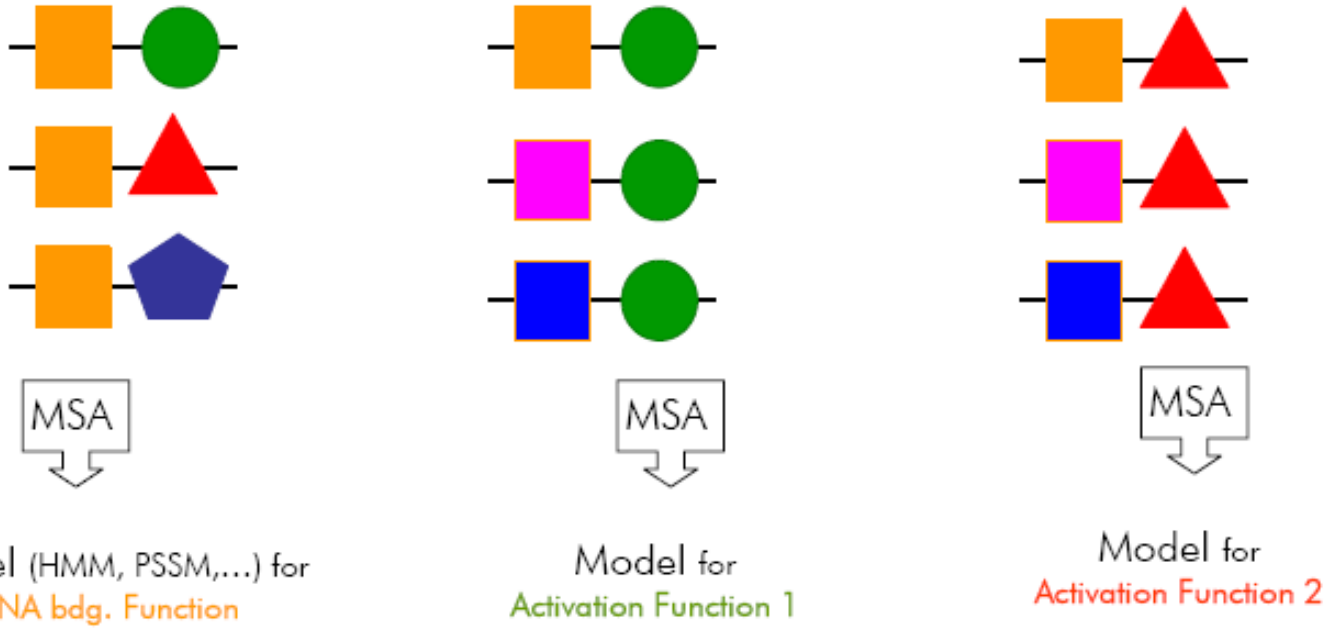
- Markov Chain Model
- Hidden Markov Models (HMMs)
- Profile HMMs

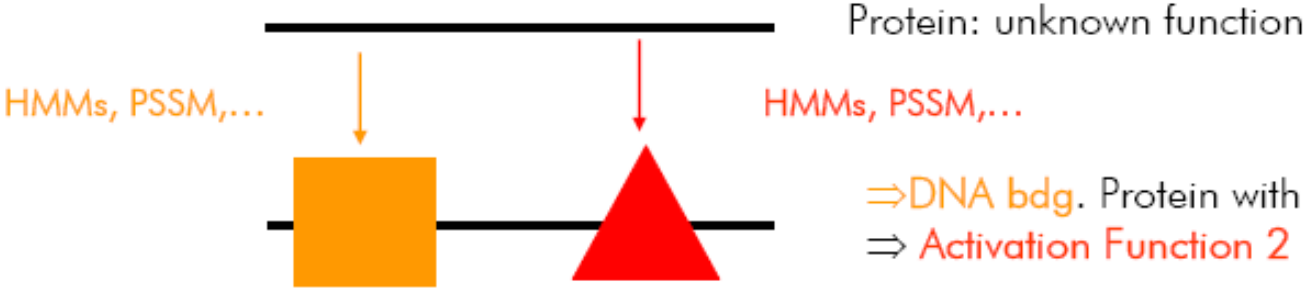
Μοντέλα Πολλαπλής Στοίχισης

- Consensus sequences
- Patterns and regular expressions
- Position Specifc Scoring Matrices (PSSMs)
- Generalized Profiles
- Hidden Markov Models (HMMs)

- Στοίχιση νέων ακολουθιών
- Αναζήτηση σε βάσεις δεδομένων
- Σχολιασμός νέων ακολουθιών

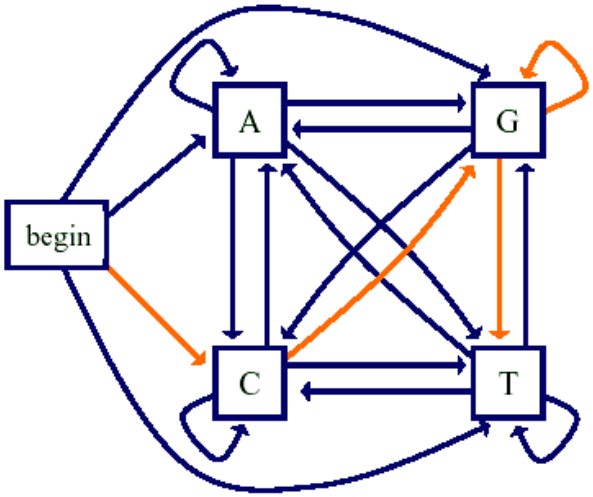
Από την Ακολουθία στη Λειτουργία





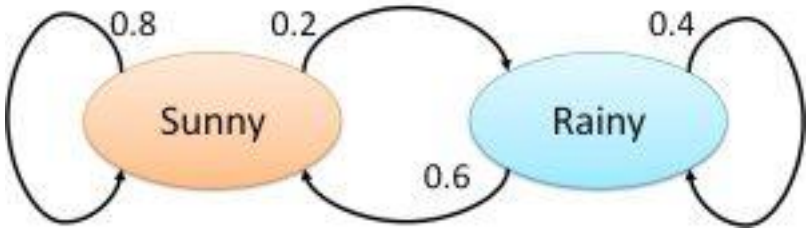
Hidden Markov Models (HMMs) 1/13

- **Markov Chain Model**
  - Σύνολο καταστάσεων
  - **Πιθανότητα μετάβασης** από τη μία κατάσταση στην άλλη
    - π.χ. μία ακολουθία DNA όπου η πιθανότητα εμφάνισης ενός νουκλεοτιδίου εξαρτάται μόνο από το προηγούμενο νουκλεοτίδιο
    - $P(\mathbf{x}) = P(x_1)P(x_2|x_1) \ldots P(x_N|x_{N-1})$
    - $P(CGGT) = P(C)\, P(G|C)\, P(G|G)\, P(T|G)$



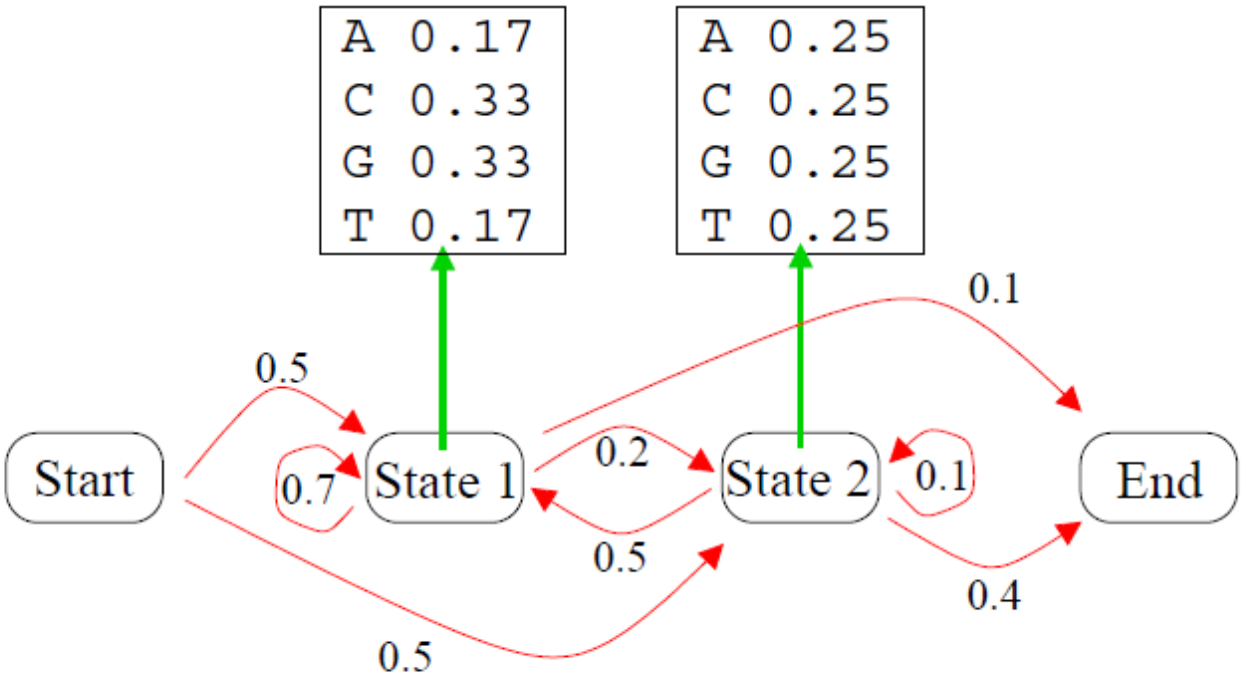
Hidden Markov Models (HMMs) 2/13

- Markov Chain Model
  - transition matrix (πίνακας μεταβάσεων)



		Weather today	
		Sunny	**Rainy**
Weather	**Sunny**	0.8	0.2
yesterday	**Rainy**	0.4	0.6

Hidden Markov Models (HMMs) 3/13



START 1 1 1 1 2 2 1 1 1 2 END

G C A G C T G G C T

Hidden Markov Models (HMMs) 4/13

$M = (\Sigma, Q, \theta)$

- Σ: Αλφάβητο συμβόλων $\Sigma = \{ b_1, b_2, \ldots, b_M \}$
- Q: Σύνολο δυνατών καταστάσεων $Q = \{ 1, \ldots, K \}$
- θ: Σύνολο πιθανοτήτων
  - Πιθανότητες Μετάβασης (Transition) από κατάσταση σε κατάσταση

    $a_{ij}: i \rightarrow j$

    $a_{i1} + \ldots + a_{iK} = 1, i = 1 \ldots K$
  - Πιθανότητες Γεννήσεως (Emission) συμβόλων σε κάθε κατάσταση

    $e_i(b) = P(x_i = b \mid \pi_i = k)$

    $e_i(b_1) + \ldots + e_i(b_M) = 1, i = 1 \ldots K$

Hidden Markov Models (HMMs) 5/13

- Όταν το HMM βρίσκεται σε μία δεδομένη κατάσταση, πρέπει να πάρει δύο αποφάσεις:
  - Σε ποια κατάσταση θα μεταβεί;
    - Κάθε χρονική στιγμή t, η επόμενη κατάσταση εξαρτάται μόνο από την τρέχουσα κατάσταση
  - Ποιο σύμβολο του αλφάβητου θα "γεννήσει";
- Ένα σύμβολο μπορεί να "γεννηθεί" από διαφορετικές καταστάσεις.
  - Όταν παρατηρείται ένα σύμβολο, δεν είναι γνωστή η κατάσταση στην οποία βρίσκεται το HMM.
- **Γνωστή σειρά συμβόλων / Άγνωστη σειρά καταστάσεων**

Hidden Markov Models (HMMs) 6/13

1. Εκτίμηση (Evaluation)

- Δεδομένου ενός μοντέλου M και μιας ακολουθίας x, ποια είναι η πιθανότητα της ακολουθίας (forward algorithm)

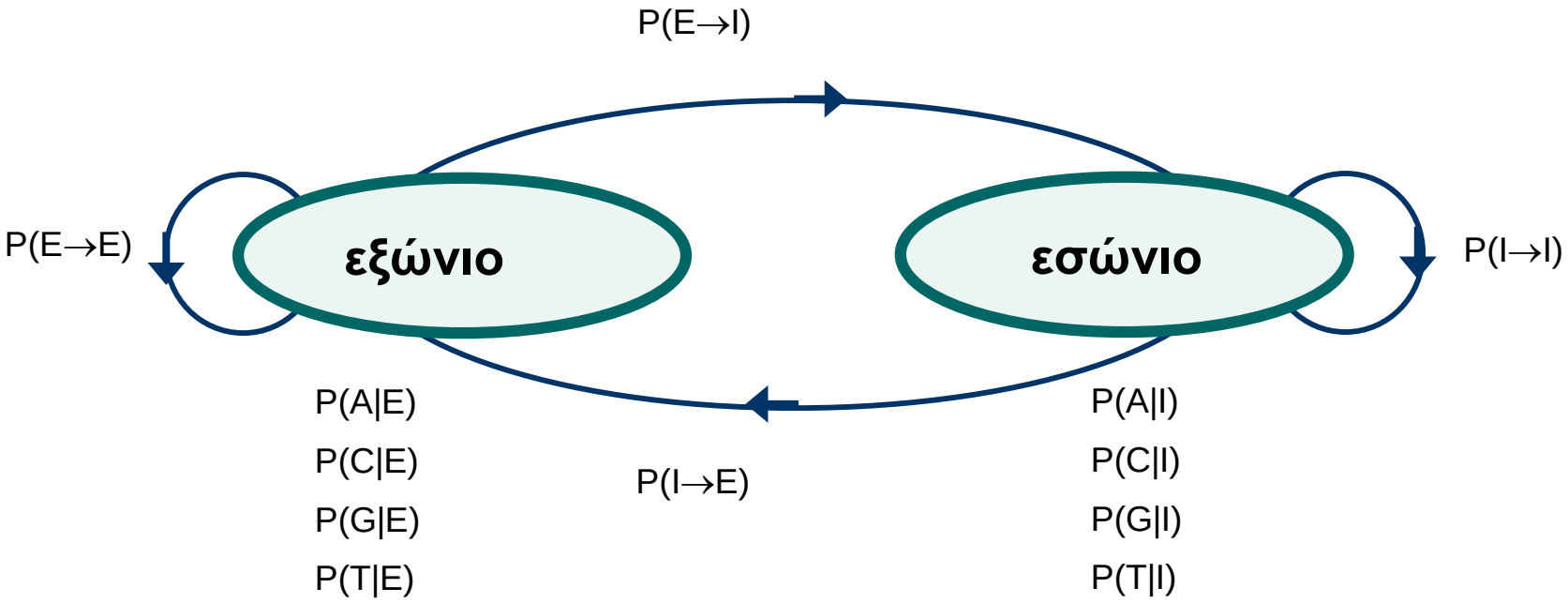
2. Αποκωδικοποίηση (Decoding)

- Δεδομένου ενός μοντέλου M και μιας ακολουθίας x, ποια είναι η πιθανότερη ακολουθία καταστάσεων για τη μοντελοποίηση της ακολουθίας (viterbi algorithm)

3. Εκπαίδευση (Learning)

- Πως μπορούν να προσδιορισθούν οι παράμετροι ενός μοντέλου M (πιθανότητες μετάβασης / γεννήσεως) από μία ομάδα ακολουθιών (forward-backward algorithm, Baum-Welch expectation maximization)

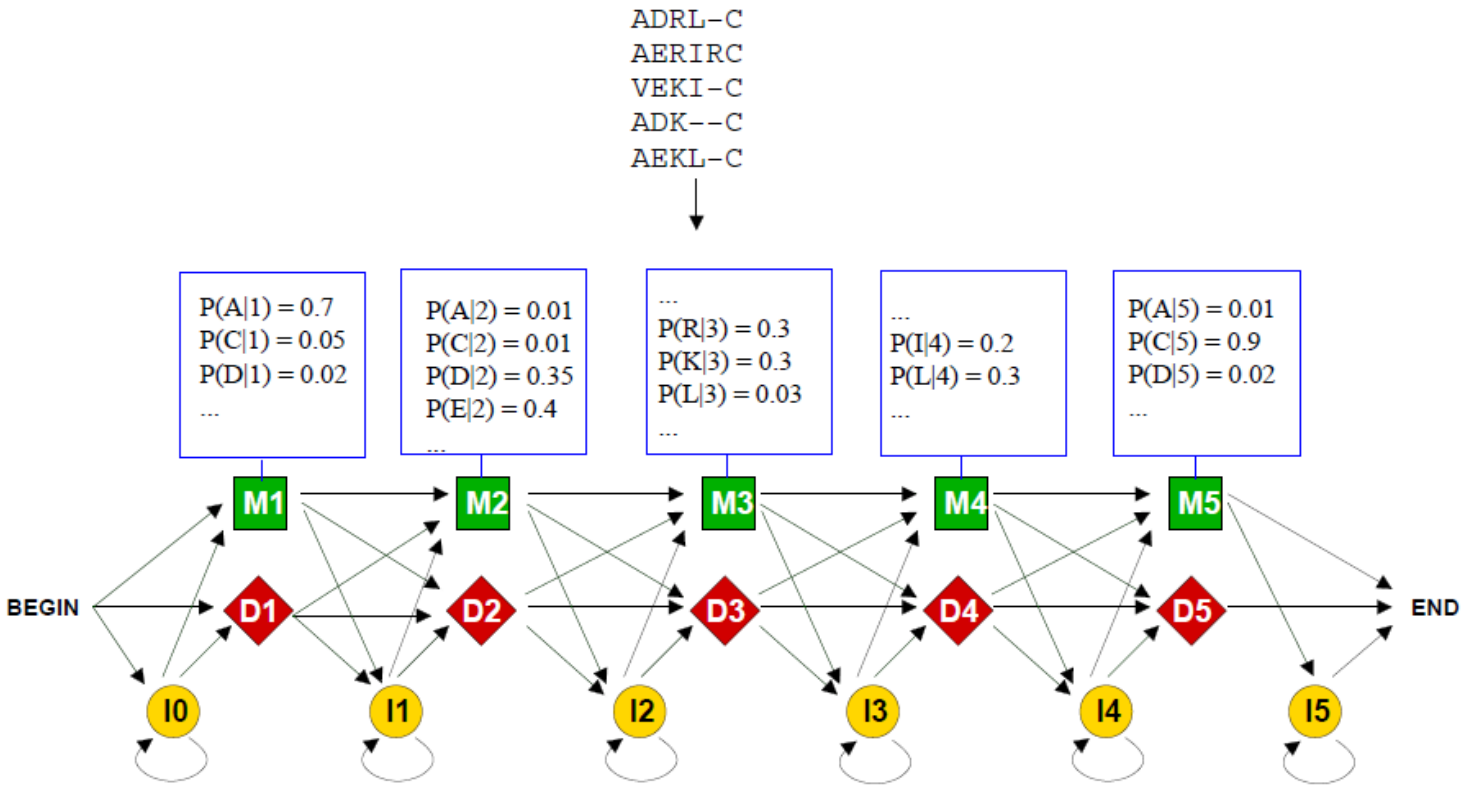
Hidden Markov Models (HMMs) 7/13



- x = A A G T A G T A T C
- π = E E E I I I I I E E

Hidden Markov Models (HMMs) 8/13

- **Profile HMMs**
  - δημιουργία μοντέλου βάσει μιας πολλαπλής στοίχισης

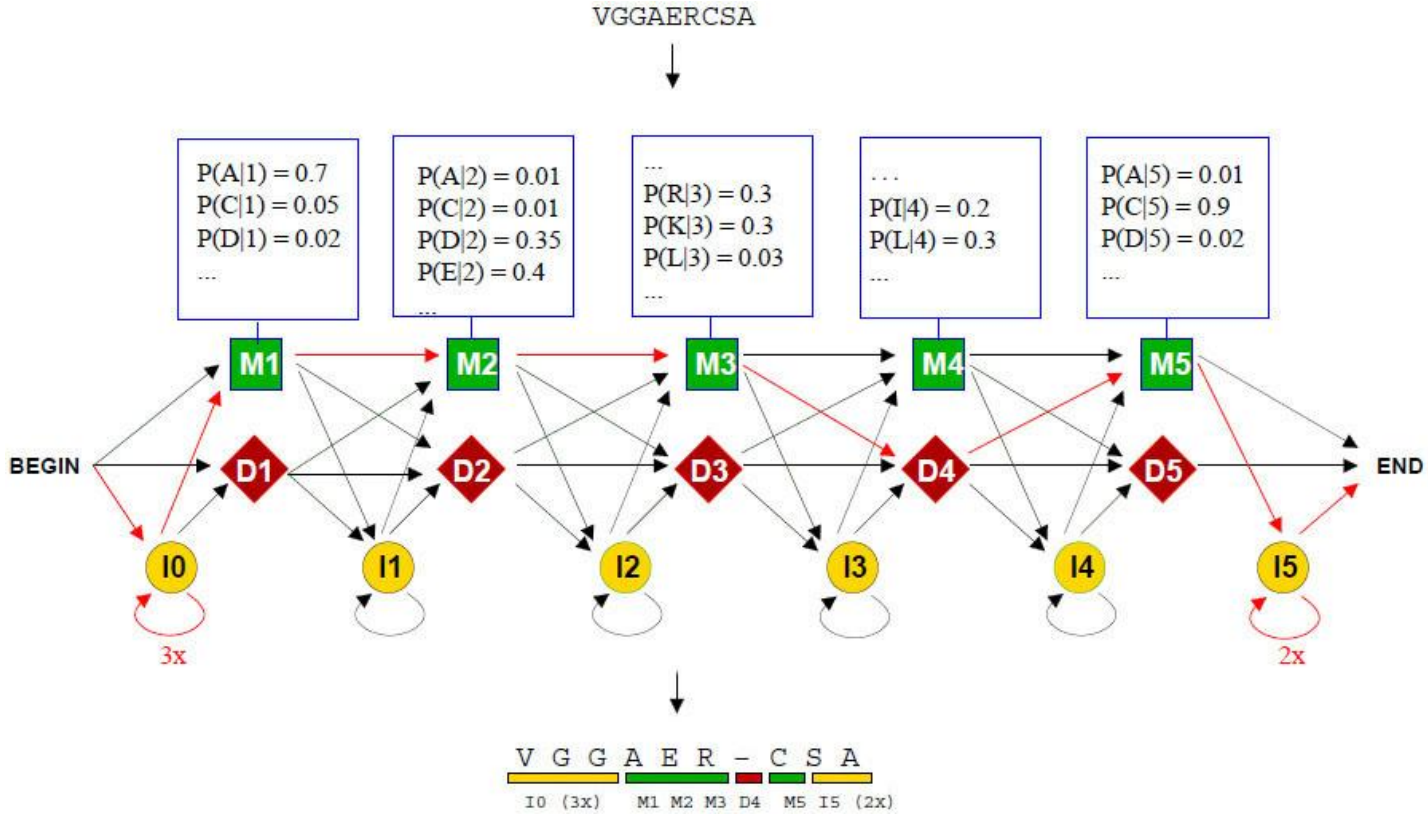


Hidden Markov Models (HMMs) 9/13

- Καταστάσεις:
  - Match (γέννηση καταλοίπου βάσει της αντίστοιχης κατανομής)
  - Insert (γέννηση καταλοίπου βάσει της κατανομής υποβάθρου)
  - Delete
- Overfitting
  - Στρεβλώσεις όταν οι συχνότητες υπολογίζονται από ένα μικρό αριθμό ακολουθιών
- Pseudocounts
  - εξομαλύνουν την παρατηρούμενη συχνότητα χαρακτήρων π.χ. Dirichlet mixture

Hidden Markov Models (HMMs) 10/13

- **Profile HMMs**
  - στοίχιση ακολουθίας με το μοντέλο (viterbi algorithm)



Hidden Markov Models (HMMs) 11/13

- Προγράμματα για τη δημιουργία και χρήση HMMs
  - HMMER http://hmmer.janelia.org/
  - SAM http://compbio.soe.ucsc.edu/sam.html
- Θεωρητικό υπόβαθρο
- Καλύτερη αντιμετώπιση των κενών
- Κατάλληλα για τη μοντελοποίηση domains
- Πιο ευαίσθητα για μακρινές ομολογίες

Hidden Markov Models (HMMs) 12/13

- Pfam http://pfam.sanger.ac.uk/
  - Pfam-A
    - στοιχίσεις / HMMs από εξειδικευμένους ερευνητές
  - Pfam-B
    - εγγραφές που προέκυψαν αυτοματοποιημένα
  - Clan
    - ομαδοποίηση των εγγραφών της Pfam-A βάσει ομοιοτήτων σε επίπεδο ακολουθίας, δομής, profile-HMM
  - Εκτενής σχολιασμός

Hidden Markov Models (HMMs) 13/13

- SMART http://smart.embl-heidelberg.de/
  - normal vs genomic
- TIGRFAMs http://www.jcvi.org/cgi-bin/tigrfams/index.cgi
  - J. Craig Venter Institute
- SUPERFAMILY http://supfam.cs.bris.ac.uk/SUPERFAMILY/
  - collection of hidden Markov models, which represent structural protein domains at the SCOP superfamily level

InterPro 1/2

- http://www.ebi.ac.uk/interpro/about.html
- ολοκλήρωση ΒΔ οικογενειών και domains

Proteins matched: SH2 domain (IPR000980)

This Domain is found in the following proteins:

Showing 1 - 20 of 9904 | Next Last

	Protein	Protein name	Species	Family	Domain organisation	Length	Structure
	A0JNB0	Tyrosine-protein kinase Fyn	*Bos taurus (Bovine)*			537	No
	A1A5H8	Tyrosine-protein kinase yes	*Danio rerio (Zebrafish)*			546	No
	A1Y2K1	Tyrosine-protein kinase Fyn	*Sus scrofa (Pig)*			537	No
	A6NKC9	SH2 domain-containing protein 7	*Homo sapiens (Human)*			451	No
	A6QLK6	GRB2-related adapter protein	*Bos taurus (Bovine)*	Neutrophil cytosol factor 2 p67phox		217	No
	A6X942	SH2 domain-containing protein 4B	*Mus musculus (Mouse)*			431	No
	A8XI74	Cell death abnormality protein 2	*Caenorhabditis briggsae*			277	No
	B2RZ59	SH2 domain-containing protein 1A	*Rattus norvegicus (Rat)*	SH2 protein 1A		126	No
	B5KFD7	Growth factor receptor-bound protein 10	*Sus scrofa (Pig)*			589	No
	F1LM93	Tyrosine-protein kinase Yes	*Rattus norvegicus (Rat)*			541	No
	F1N9Y5	Tyrosine-protein kinase SYK	*Gallus gallus (Chicken)*	Tyrosine-protein kinase, non-receptor SYK/ZAP-70		613	No

InterPro 2/2

- CATH/Gene3D at University College, London, UK
- PANTHER at University of Southern California, CA, USA
- PIRSF at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, USA
- Pfam at the Wellcome Trust Sanger Institute, Hinxton, UK
- PRINTS at the University of Manchester, UK
- ProDom at PRABI Villeurbanne, France
- PROSITE and HAMAP at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland
- SMART at EMBL, Heidelberg, Germany
- SUPERFAMILY at the University of Bristol, UK
- TIGRFAMs at the J. Craig Venter Institute, Rockville, MD, US

Βιβλιογραφία

- David Mount, "Bioinformatics: Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press; 2nd edition (March 12, 2013).

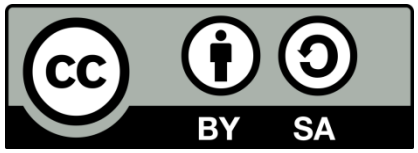
- Jonathan Pevsner, "Bioinformatics and Functional Genomics", Wiley-Blackwell; 2nd edition (May 4, 2009).

- Andreas D. Baxevanis, B. F. Francis Ouellette, "Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins", Wiley-Interscience; 3rd edition (October 29, 2004).

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.

- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδεια χρήσης, η άδεια χρήσης αναφέρεται ρητώς.



Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα Γεωπονικού Πανεπιστημίου Αθηνών**» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.