Ειδικά Θέματα Βιοπληροφορικής

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

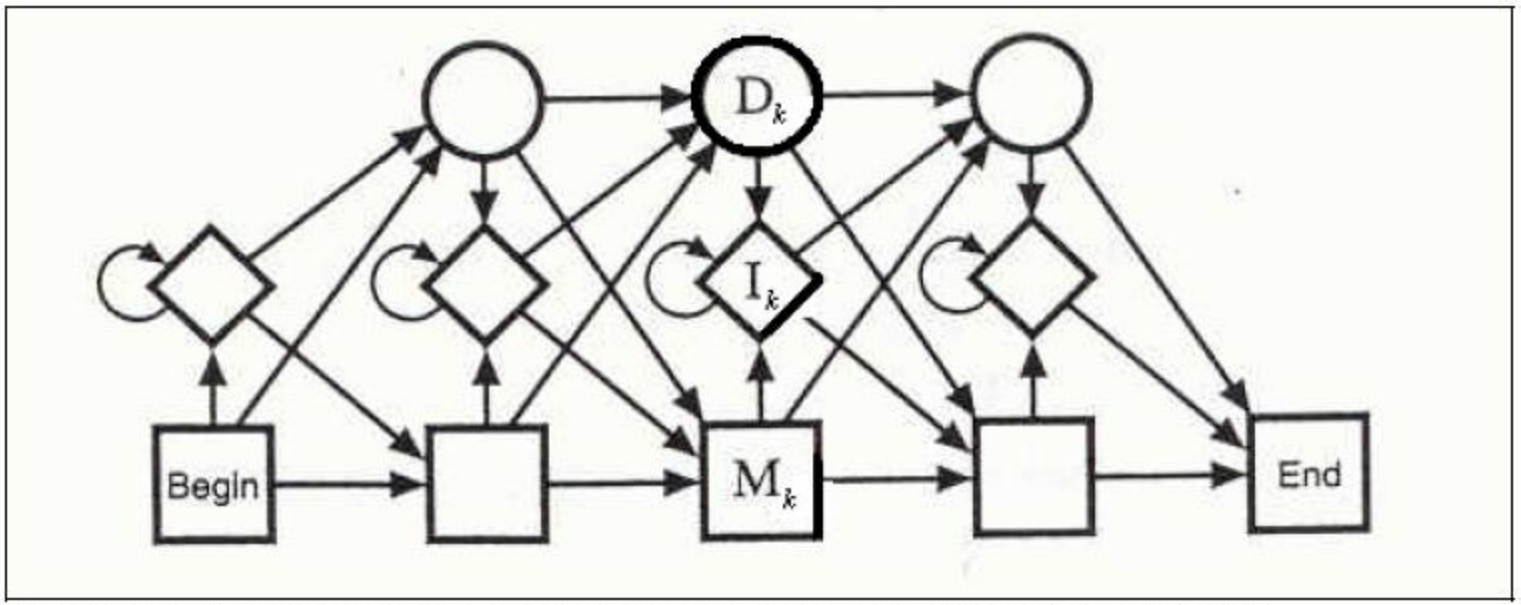
Διάλεξη 5

Profile Hidden Markov Models και
Transformational Grammars

Profile HMM

- Ένα HMM με “left-to-right” αρχιτεκτονική
- Μπορεί να ειδωθεί σαν επέκταση των profiles
- Διαθέτει ειδικά states τα οποία δεν έχουν emissions (non-emitting states)
- Περιγράφουν στατιστικά μια πολλαπλή στοίχιση
- Προτάθηκαν από τους Hughey and Krogh (1996)

Γενική αρχιτεκτονική pHMM



Εικόνα 2.12 Σχηματική αναπαράσταση ενός τυπικού profile Hidden Markov Model

Οι καταστάσεις που παρατηρούνται σε ένα τέτοιο μοντέλο (εκτός αυτών της εκκίνησης και του τερματισμού) χωρίζονται σε 3 κατηγορίες :

Καταστάσεις Σύμπτωσης (Match states)	M_k	*τετράγωνα*
Καταστάσεις Εισαγωγής (Insertion states)	I_k	*ρόμβοι*
Καταστάσεις Απαλοιφής (Deletion states)	D_k	*κύκλοι*

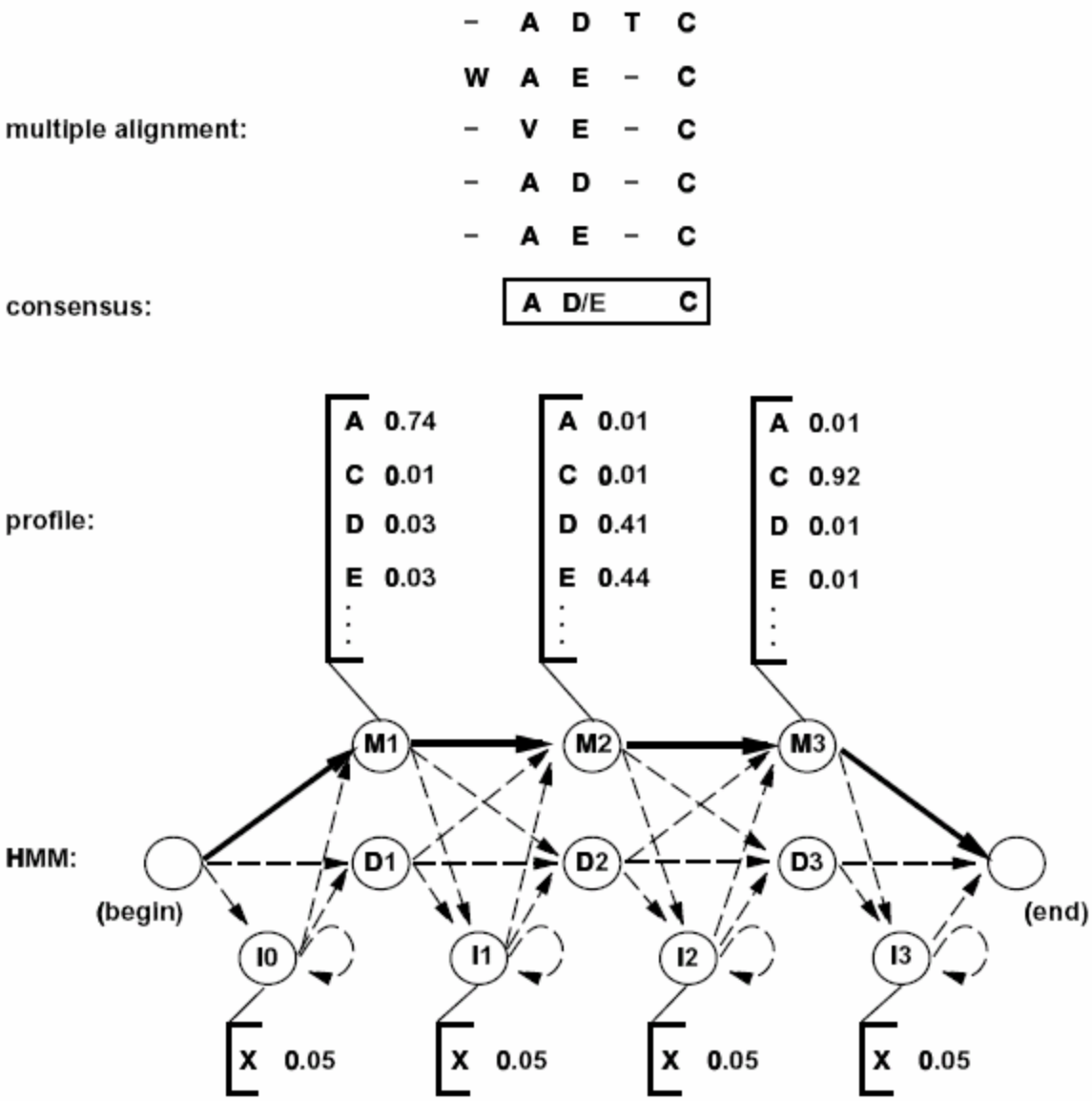
και συνδέονται με τις αντίστοιχες πιθανότητες μεταβάσεως, που συμβολίζονται με βέλη.

Ιδιαιτερότητες

- Αντίστοιχα ορίζονται οι πιθανότητες γέννησης οι οποίες γεννούν τα σύμβολα σε κάθε κατάσταση.
- Έτσι υπάρχει και εδώ μια αλληλουχία καταστάσεων η οποία είναι κρυφή και μια αλληλουχία συμβόλων που είναι φανερή, και θεωρούμε ότι παράγεται από την αλληλουχία των καταστάσεων.
- Οι καταστάσεις σύμπτωσης και εισαγωγής, είναι κανονικές καταστάσεις οι οποίες συνδέονται μέσω των πιθανοτήτων γέννησης με την εμφάνιση συμβόλων.
- Οι μεν καταστάσεις σύμπτωσης αντιστοιχούν σε στήλες της πολλαπλής στοίχισης οι οποίες στοιχίζονται καλά και άρα αντιστοιχούν σε περιοχή με ομοιότητα, ενώ οι καταστάσεις εισαγωγής, αντιστοιχούν σε περιοχές στις οποίες έχουμε εισαγωγή χαρακτήρων που δεν στοιχίζονται καλά.
- Οι περιοχές αυτές, οι οποίες δεν υπάρχουν στις υπόλοιπες ακολουθίες, εμφανίζονται ως κενά τα οποία μοντελοποιούνται μέσω των σιωπηρών καταστάσεων απαλοιφής.

Ιδιαιτερότητες

- Εκπαίδευση με full Baum-Welch, Simulated annealing ή Viterbi training
- Για την κατασκευή μπορεί να απαιτείται μια πολλαπλή στοίχιση ή και όχι
- Στοίχιση της ακολουθίας με το μοντέλο με κλασσικό Viterbi
- Πολλαπλή στοίχιση μέσω στοίχισης ακολουθιών με το μοντέλο


multiple alignment:
– A D T C
W A E – C
– V E – C
– A D – C
– A E – C
consensus:
A D/E C
profile:
A 0.74
C 0.01
D 0.03
E 0.03
A 0.01
C 0.01
D 0.41
E 0.44
A 0.01
C 0.92
D 0.01
E 0.01
HMM:
(begin)
M1
M2
M3
D1
D2
D3
(end)
I0
I1
I2
I3
X 0.05
X 0.05
X 0.05
X 0.05

Πλεονεκτήματα

- Με την εισαγωγή των διαφορετικών καταστάσεων σύμπτωσης και εισαγωγής, γίνεται μια σημαντική τομή σε σχέση με τις κλασσικές μεθόδους στοίχισης, οι οποίες καθώς δεν προϋποθέτουν ένα μοντέλο δεν διαχωρίζουν τις πληροφοριακές θέσεις στην στοίχιση από τις απλές τυχαίες εισαγωγές.
- Επιπλέον, για πρώτη φορά οι ποινές για την εισαγωγή κενών (gap penalties), δεν τίθενται εκ των προτέρων αλλά εκτιμώνται από τα δεδομένα και αναπαρίστανται με καθαρά πιθανοθεωρητικό τρόπο, αποκλείοντας την υποκειμενική παρέμβαση.
- Έτσι, με τέτοια μοντέλα, είμαστε σε θέση να πραγματοποιήσουμε ιδιαίτερα ευαίσθητες αναζητήσεις και να εντοπίσουμε απομακρυσμένες ομολογίες (remote homologies), τις οποίες οι παραδοσιακοί αλγόριθμοι στοίχισης δεν θα μπορούσαν να εντοπίσουν.

Πολλαπλή Στοίχιση

EGF domain structural alignment:

```
1ixa       VDGDQCE    SNPCLNGGSCKDD  INSYECWCPFGFEGKNCEL
1apo       KDGDQCE    GHPCLNQGHCKDG  IGDYTCTCAEGFEGKNCEFSTR
1epi       NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR
4tgf     VVSHFNDCPDSHTQFCFH GTCRFLVQEDKPACVCHSGYVGARCEHADLLA
```

ClustalW alignment:

```
1ixa       VDGDQCESN   P CLNGGSCKDD    INSYECWCPFGFEGKNCE L
1apo       KDGDQCEGH   P CLNQGHCKDG    IGDYTCTCAEGFEGKNCE FSTR
1epi       NSYPGCPSSYDGYCLNGGVCMHIES LDSYTCNCVIGYSGDRCQTRDLRWWELR
4tgf  VVSHFND CPDSHTQF CFHG TCRFLVQEDKP ACVCHSGYVGARCE HADLLA
```

HMM simulated annealing alignment:

```
1ixa       V  DGDQCE    SNPCLNGGSCKDD  INSYECWCPFGFEGKNCE  L
1apo       K  DGDQCE    GHPCLNQGHCKDG  IGDYTCTCAEGFEGKNCEF STR
1epi       N  SYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR
4tgf       VVSHFNDCPDSHTQFCFH GTCRFLVQEDKPACVCHSGYVGARCE  HADLL  A
```

Χρήσεις

- Πολλαπλή στοίχιση
- Κατασκευή χαρακτηριστικών profiles
- Εντοπισμός remote homologues
- Αναζητήσεις σε βάσεις δεδομένων

Πολλαπλή στοίχιση

a=CAEFDDH **b**=CDAEFPDDH

Viterbi

a = CAEFDDH

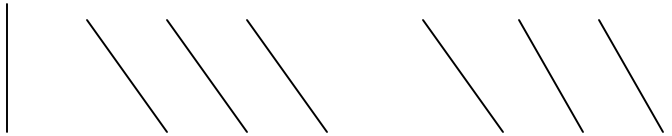
Pa=$m_0m_1m_2m_3m_4d_5d_6m_7m_8m_9m_{10}$

b = CDAEFPDDH

Pb=$m_0m_1i_1m_2m_3m_4d_5m_6m_7m_8m_9m_{10}$

C A E F D D H

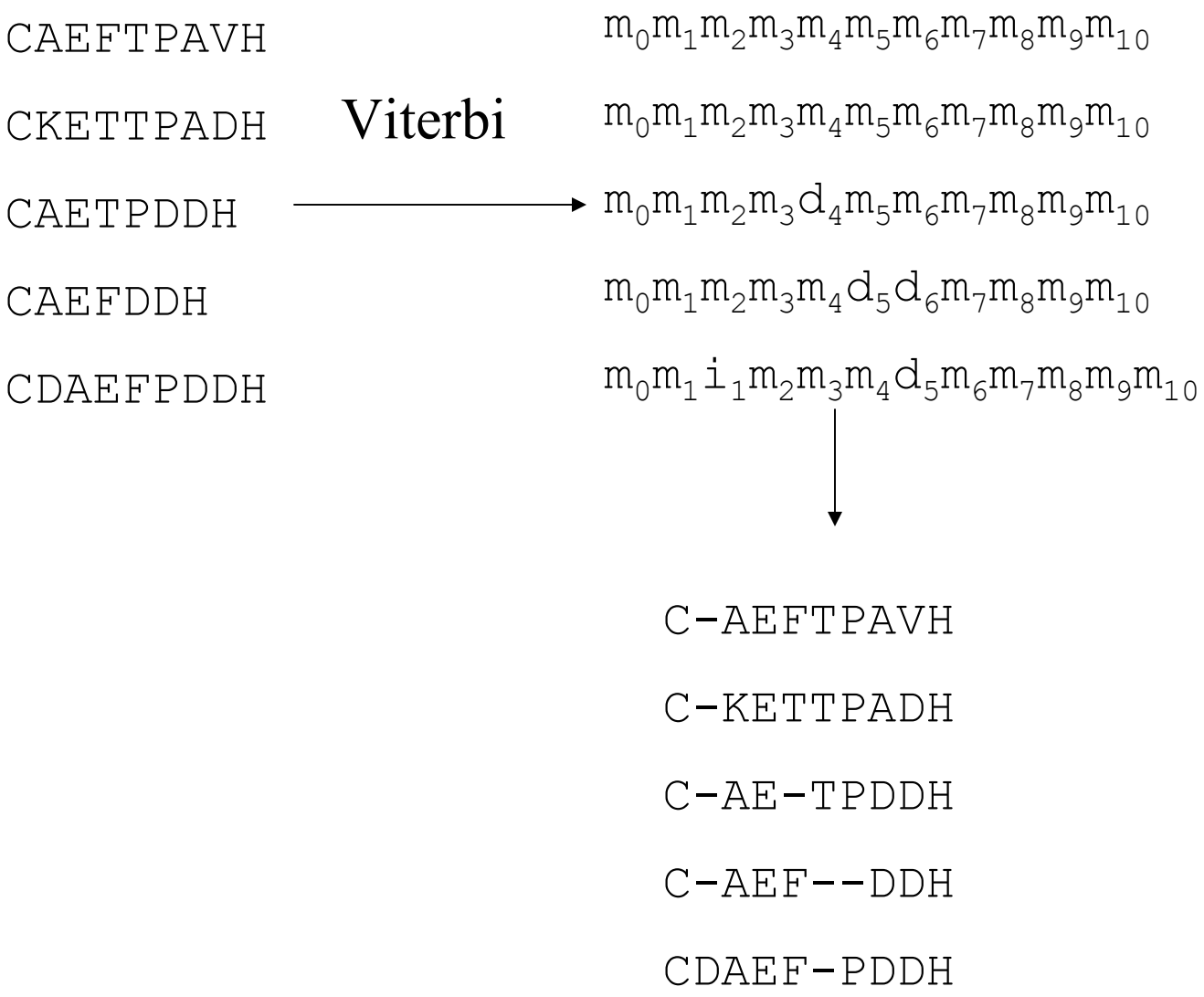
m_0m_1 $m_2m_3m_4d_5d_6m_7m_8m_9m_{10}$



C D A E F P D D H

m_0m_1 $i_1m_2m_3m_4d_5m_6m_7m_8m_9m_{10}$

Γενικά


CAEFTPAVH
CKETTPADH
CAETPDDH
CAEFDDH
CDAEFPDDH
Viterbi
$m_0m_1m_2m_3m_4m_5m_6m_7m_8m_9m_{10}$
$m_0m_1m_2m_3m_4m_5m_6m_7m_8m_9m_{10}$
$m_0m_1m_2m_3d_4m_5m_6m_7m_8m_9m_{10}$
$m_0m_1m_2m_3m_4d_5d_6m_7m_8m_9m_{10}$
$m_0m_1i_1m_2m_3m_4d_5m_6m_7m_8m_9m_{10}$
C-AEFTPAVH
C-KETTPADH
C-AE-TPDDH
C-AEF--DDH
CDAEF-PDDH

Διαδικασία

- Μια ακολουθία για την οποία υπάρχουν πειραματικές ενδείξεις για την λειτουργία ή τη δομή της
- Αναζήτηση σε βάσεις δεδομένων (BLAST, PSI-BLAST)
- Συλλογή ομολόγων, επιλογή ξεσκαρτάρισμα κλπ
- Πολλαπλή στοίχιση (μπορεί και τροποποίηση αυτής με το χέρι)
- Προγνώσεις (Secondary structure, TM κλπ)
- Κατασκευή HMM, αξιολόγηση του
- Αναζήτηση εκ νέου σε βάσεις δεδομένων

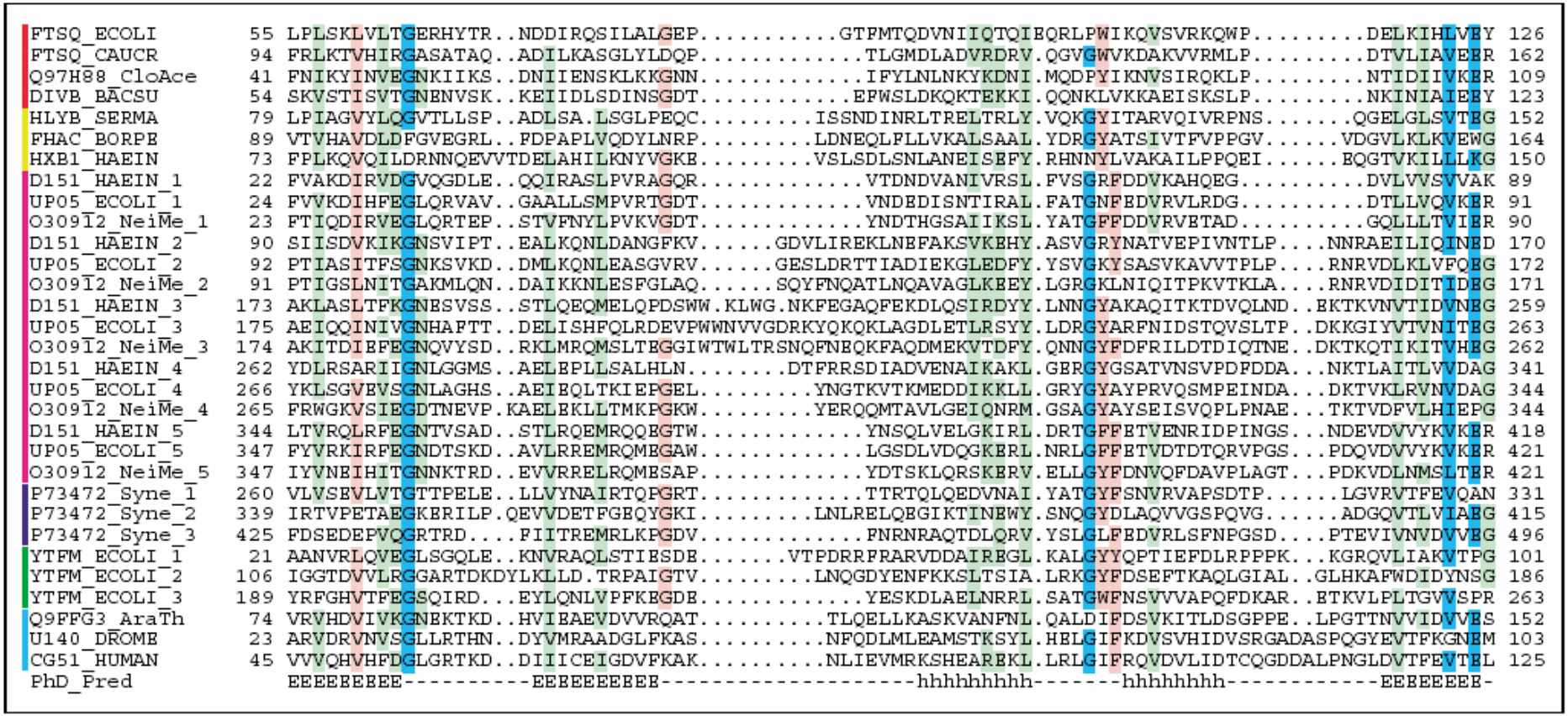


Figure 1. Representative multiple alignment of the POTRA (for polypeptide-transport-associated domain) domain. The alignment was produced with HMMer [10] and T-Coffee [23] using default parameters and was slightly refined manually. It is viewed with the Belvu program (http://www.sanger.ac.uk/Software/Pfam/help/belvu_setup.shtml). The colour scheme indicates the average BLOSUM62 score (correlated to amino-acid conservation) in each alignment column: cyan, >1.6; light red, 1–1.6; and light green, 0.3–1. The boundaries of the domains are indicated by the residue positions on each side. Consensus PHD secondary-structure prediction [11] is shown below the alignment, with E indicating a β strand and H an α helix, in upper and lower case for high and low accuracy, respectively. The sequences are named with their SWISSPROT or SPTREMBL identifications. Multiple alignments and trees for each family, profiles and other information are accessible at: http://www.pdg.cnb.uam.es/POTRA. A larger version of this multiple sequence alignment (alignment number ALIGN_000590) has been deposited at the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000590.dat). The species abbreviations are: AraTh, *Arabidopsis thaliana*; BACSU, *Bacillus subtilis*; BORPE, *Bordetella pertussis*; CAUCR, *Caulobacter crescentus*; CloAce, *Clostridium acetobutylicum*; DROME, *Drosophila melanogaster*; ECOLI, *Escherichia coli*; HAEIN, *Haemophilus influenzae*; NeiMe, *Neisseria meningitidis*; SERMA, *Serratia marcescens*; Syne, *Synechocystis sp.* The numbering after the protein name indicates the domain-repeat number when more than one is detected in the sequence. Different groups identified by sequence similarity are shown by coloured lines to the left of the alignment: red, FtsQ; yellow, ShlB; violet, D15; blue, Toc75; green, YTFM; and cyan, CGI51.

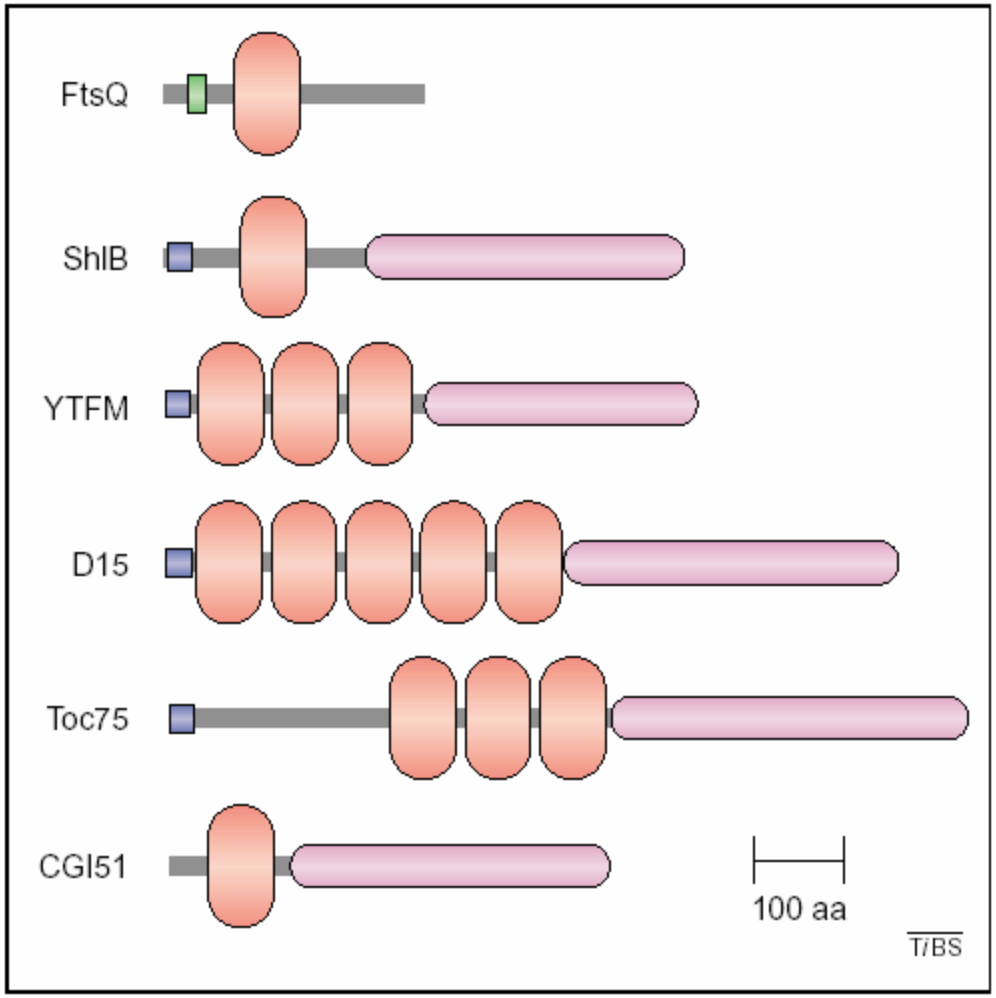


Figure 2. Schematic representation of the domain architectures of a representative set of POTRA (for polypeptide-transport-associated domain) domain-containing proteins. Corresponding to SWISSPROT identifiers: CGI51, SW:Q9Y512: CG51_HUMAN; D15, SW:P46024:D151_HAEIN; FtsQ, SW:P06136:FTSQ_ECOLI; ShlB, SW:P15321:HLYB_SERMA; Toc75, SPTREMBL:P73472; YTFM, SW:P39320: YTFM_ECOLI. The proteins are drawn approximately to scale and colour coded as follows: transmembrane region, green; signal peptide, blue; POTRA domain, pale red; β barrel, pink. Hypothetical signal peptides were predicted with Signal P [24]. Transmembrane β-sheets were predicted using B2TMPRED [25].

Διαθέσιμα πακέτα

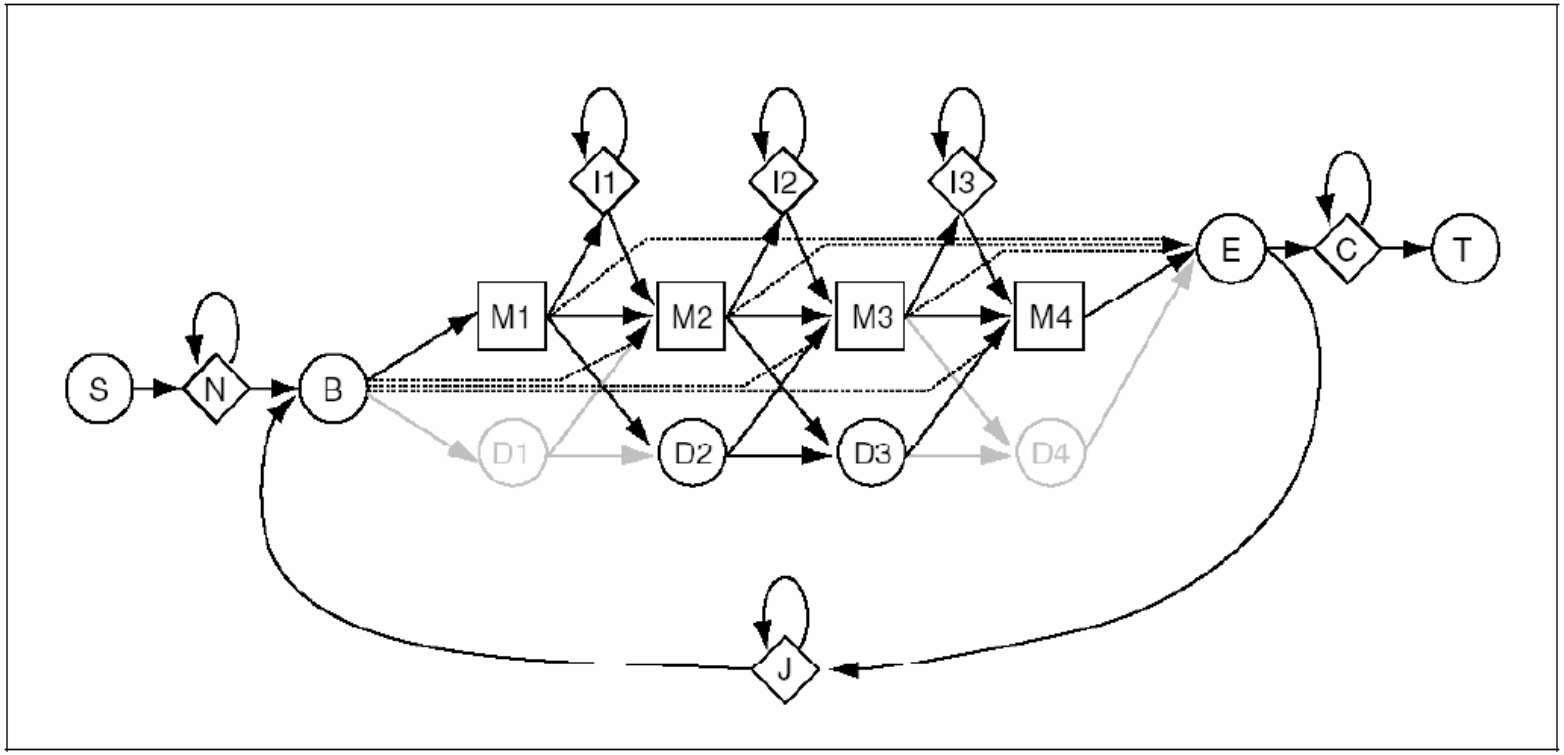
- HMMER
  - (http://hmmer.wustl.edu/)
- HMMpro
  - (http://www.netid.com/)
- SAM
  - (http://www.cse.ucsc.edu/research/compbio/sam.html)
- PFTOOLS
  - (http://www.isrec.isb-sib.ch/ftp-server/pftools/)

  [PROSITE PATTERNS]

HMMER

- GNU license
- Ανανεώνεται συνεχώς
- Τρέχει σε όλες τις πλατφόρμες
- Ποικιλία εργαλείων
- Ευέλικτη αρχιτεκτονική

HMMER

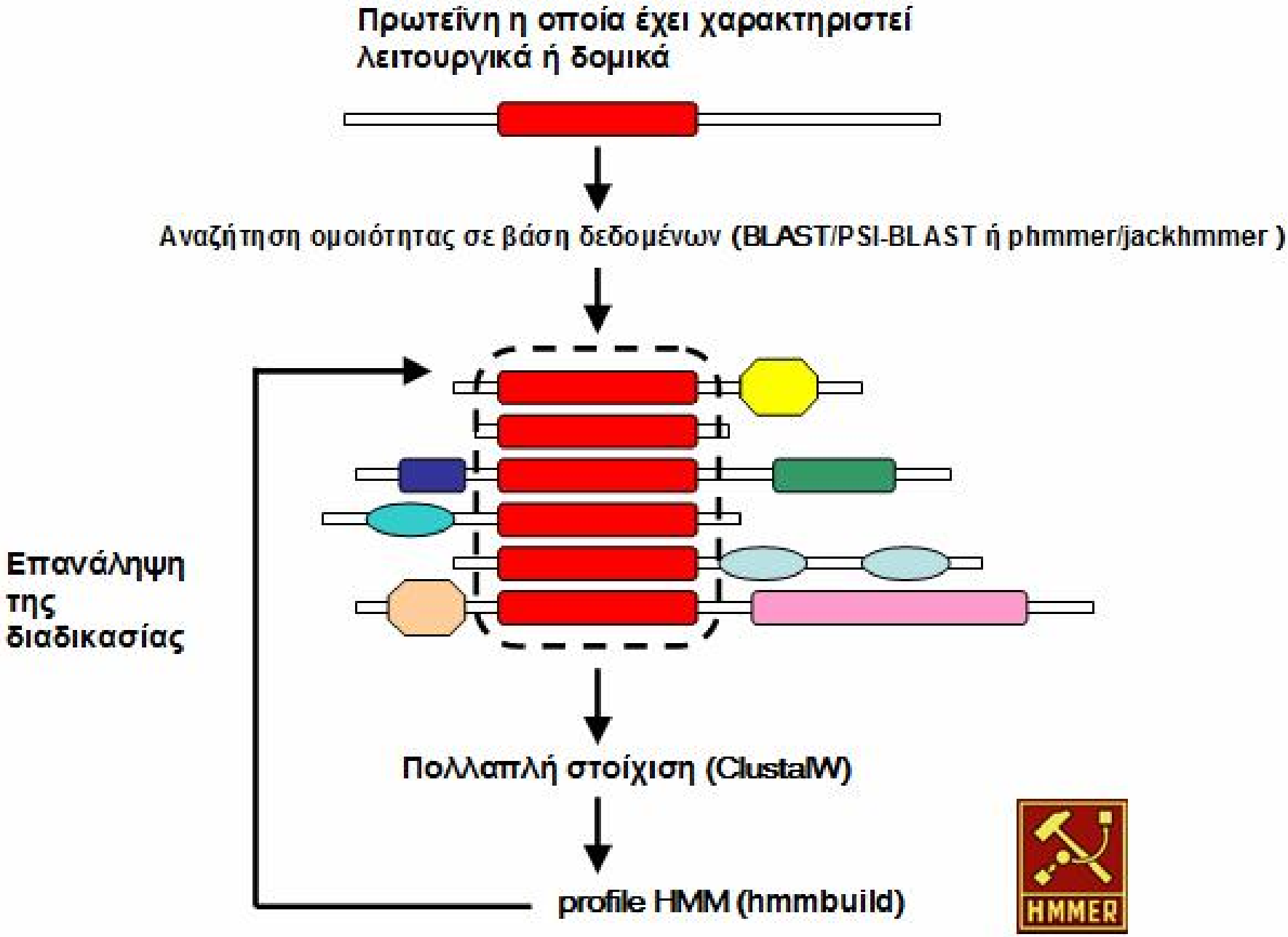


Ρουτίνες του HMMER

- **hmmbuild:** Πρόγραμμα με χρήση του οποίου, ξεκινώντας από μια αρχική πολλαπλή στοίχιση, κατασκευάζεται ένα μοντέλο HMM το οποίο να την περιγράφει.
- **hmmalign:** Πρόγραμμα με το οποίο μια σειρά ακολουθιών οι οποίες προέρχονται από ένα HMM, στοιχίζονται σε μια πολλαπλή στοίχιση. Η πολλαπλή στοίχιση, επιτυγχάνεται μέσω διαδοχικών στοιχίσεων των ακολουθιών με το μοντέλο.
- **hmmsearch:** Πρόγραμμα το οποίο, πραγματοποιεί αναζητήσεις ενός μοντέλου HMM έναντι μιας βάσης ακολουθιών πρωτεϊνών.
- **phmmer:** Πρόγραμμα το οποίο πραγματοποιεί αναζήτηση μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το BLASTP)
- **jackhmmer:** Πρόγραμμα το οποίο πραγματοποιεί επαναληπτικές αναζητήσεις μια πρωτεϊνικής αλληλουχίας έναντι μιας βάσης δεδομένων πρωτεϊνών (ανάλογο με το PSI-BLAST)

Ρουτίνες του HMMER

- **hmmscan:** Πρόγραμμα με το οποίο πραγματοποιούνται αναζητήσεις μιας η περισσότερων ακολουθιών έναντι μιας βάσης δεδομένων από μοντέλα HMM. Πρέπει να τονιστεί εδώ, ότι αν έχουμε μια ακολουθία και ένα HMM, τα δυο παραπάνω προγράμματα επιστρέφουν ακριβώς το ίδιο αποτέλεσμα. Αν διαφέρουν, είτε οι ακολουθίες είτε τα μοντέλα, τότε δίνουν άλλο αποτέλεσμα, λόγω του διαφορετικού τρόπου υπολογισμού της στατιστικής σημαντικότητας.
- **nhmmer:** Πρόγραμμα που πραγματοποιεί αναζήτηση μιας ακολουθίας DNA, μιας στοίχισης ή ενός pHMM, έναντι μιας βάσης ακολουθιών DNA. (ανάλογο με το BLASTN)
- **nhmmscan:** Πρόγραμμα που πραγματοποιεί αναζήτηση μιας ακολουθίας DNA έναντι μιας βάσης δεδομένων από DNA profile HMM.
- **hmmconvert:** Πρόγραμμα που μετατρέπει μοντέλα HMM από και προς τη μορφή του HMMER3.
- **hmmemit:** Πρόγραμμα, με το οποίο 'εκπέμπεται' η καλύτερη (ανάλογα με τον ορισμό) ακολουθία η οποία θα μπορούσε να παραχθεί από το μοντέλο.
- **hmmpress:** Μετατρέπει μια βάση δεδομένων HMM σε δυαδικό κώδικα για το hmmscan.
- **hmmstat:** δείχνει συνοπτικά στατιστικά για μια βάση δεδομένων HMM.

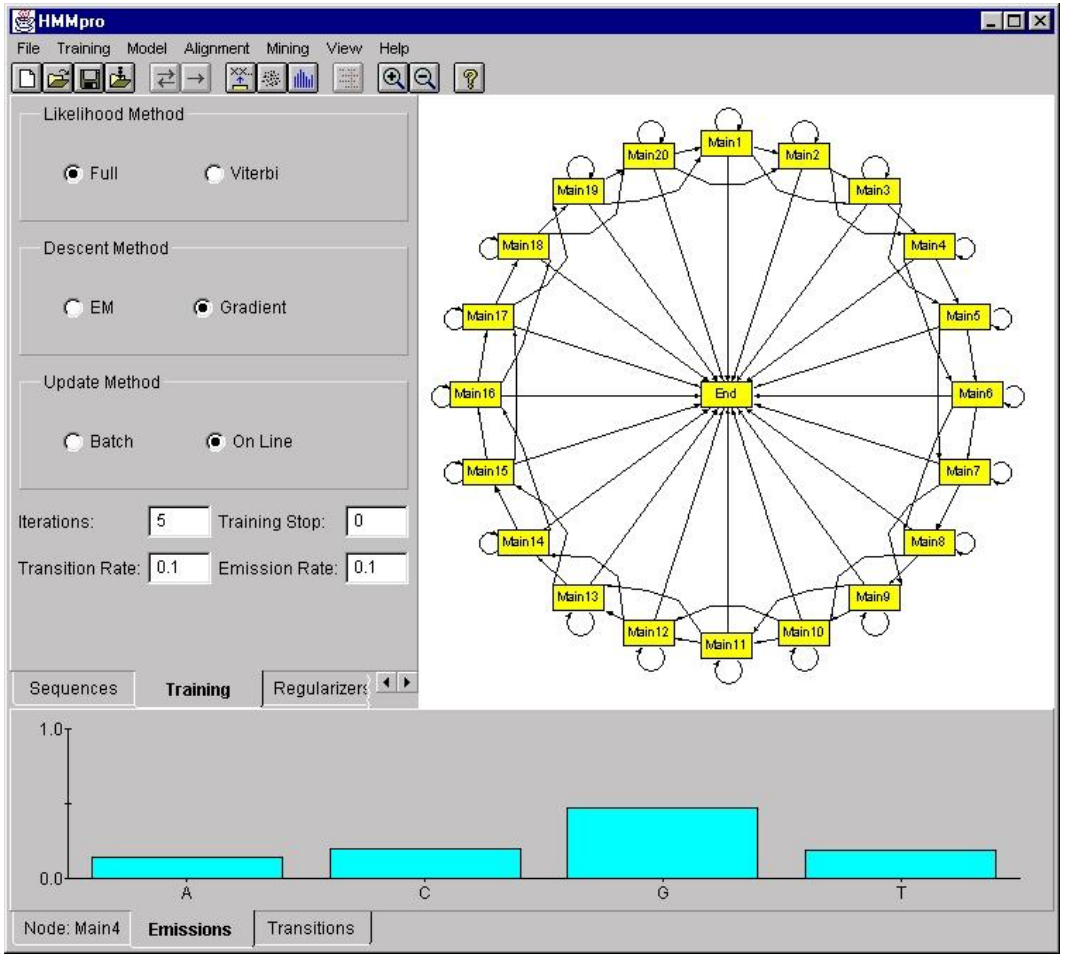

Πρωτεΐνη η οποία έχει χαρακτηριστεί
Λειτουργικά ή δομικά
Αναζήτηση ομοιότητας σε βάση δεδομένων (BLAST/PSI-BLAST ή phmmer/jackhmmer)
Επανάληψη
της
διαδικασίας
Πολλαπλή στοίχιση (ClustalW)
profile HMM (hmmbuild)
HMMER

διαδικασία χαρακτηρισμού μιας πρωτεϊνικής οικογένειας

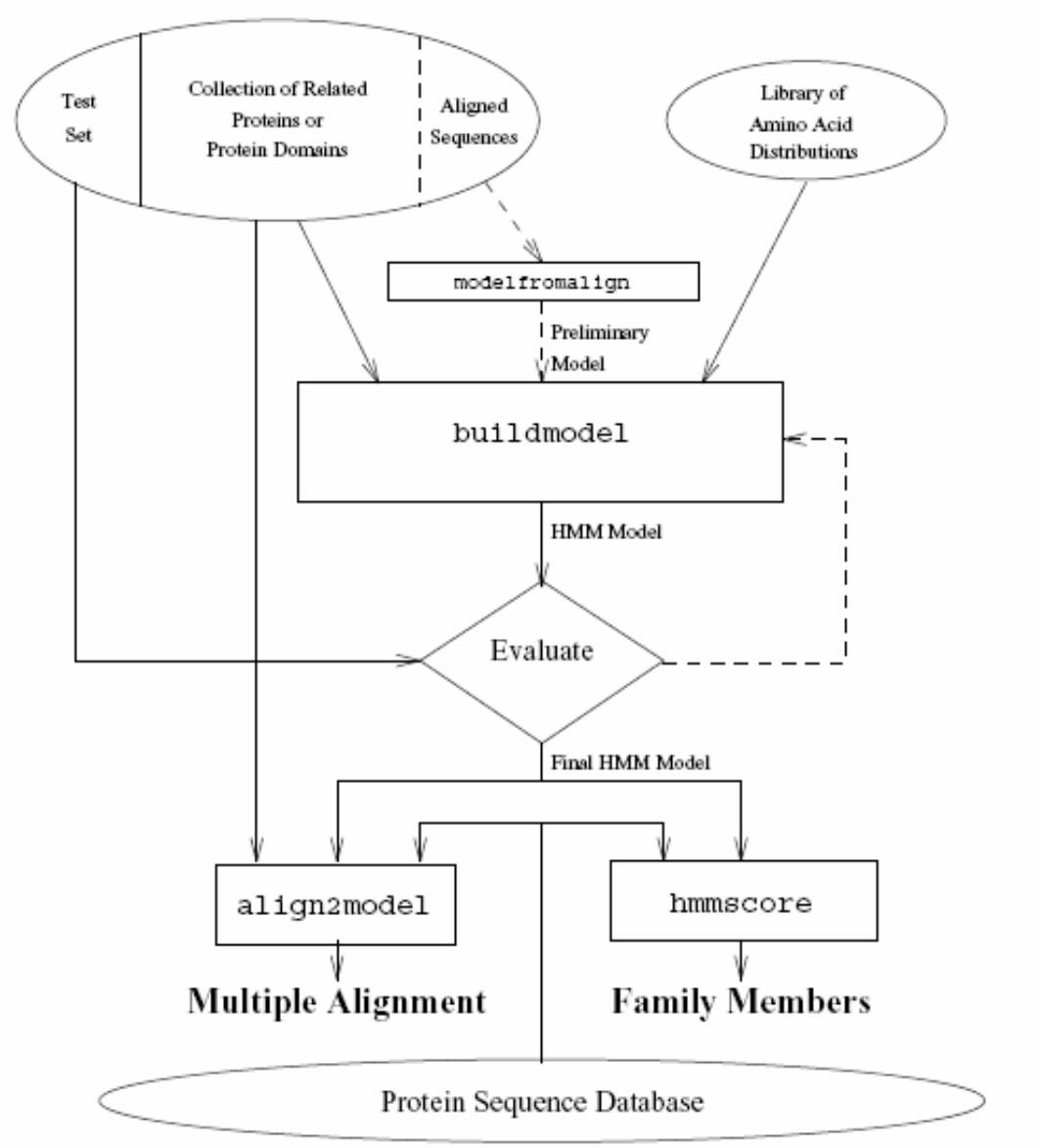
- Στην αρχή, ξεκινάμε με μια ακολουθία για την οποία υπάρχουν πειραματικές ενδείξεις για τη λειτουργία ή τη δομή της
- Γίνεται αναζήτηση σε βάσεις δεδομένων (BLAST, PSI-BLAST ή πλέον, με το HMMER)
- Συλλογή ομολόγων, επιλογή και ξεσκαρτάρισμα
- Γίνεται μια πολλαπλή στοίχιση (μπορεί και τροποποίηση αυτής με το χέρι)
- Ανάλογα με την περίπτωση, πραγματοποιούνται προγνώσεις (δευτεροταγούς δομής, διαμεμβρανικών τμημάτων ή οποιουδήποτε άλλου χρήσιμου χαρακτηριστικού)
- Γίνεται κατασκευή HMM και αξιολόγηση του (HMMER)
- Αναζήτηση εκ νέου σε βάσεις δεδομένων, μέχρι να μην προκύπτουν νέα μέλη της οικογένειας.

HMMpro

- Commercial Software
- Παραθυρικό interface



SAM


Test
Set
Collection of Related
Proteins or
Protein Domains
Aligned
Sequences
Library of
Amino Acid
Distributions
modelfromalign
Preliminary
Model
buildmodel
HMM Model
Evaluate
Final HMM Model
align2model
hmmscore
Multiple Alignment
Family Members
Protein Sequence Database