Βιοπληροφορική Ι

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

Μέθοδοι πρόγνωσης

- Μέθοδοι πρόγνωσης για πρωτεΐνες
  - Δευτεροταγής δομή
  - Διαμεμβρανικά τμήματα
  - Σηματοδοτικές αλληλουχίες
  - Στόχευση
  - Μετα-μεταφραστικές τροποποιήσεις
  - Αλληλεπιδράσεις, δομική κατάταξη κλπ
- Μέθοδοι πρόγνωσης DNA/RNA
  - Έύρεση γονιδίων
  - Εύρεση υποκινητών
  - Σημεία συρραφής
  - TIS
  - Poly-A
  - miRNA

Δευτεροταγής Δομή

Sequence – structure gap

- Today we have much more sequenced proteins than protein's structures.
- **The gap is rapidly increasing**.

Problem:

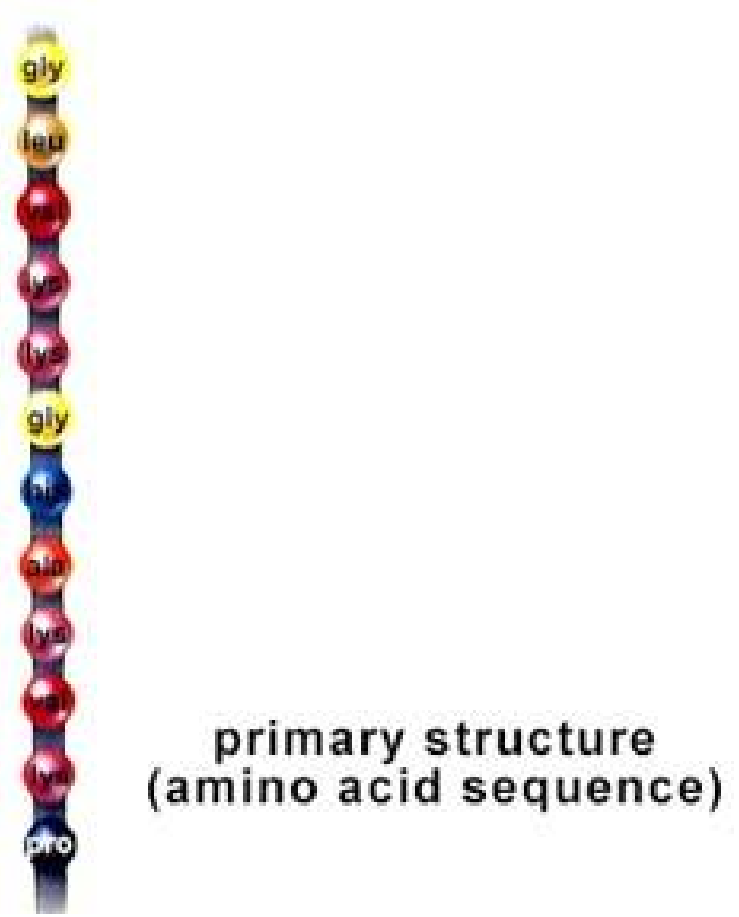
Finding protein structure isn't that simple.

Solution:

A good start : find secondary structure.

Primary Structure

- The sequence of amino acids in the polypeptide chain
- Described as a string from a finite alphabet $\sum_{aa}$
  - $|\sum_{aa}| = 20$

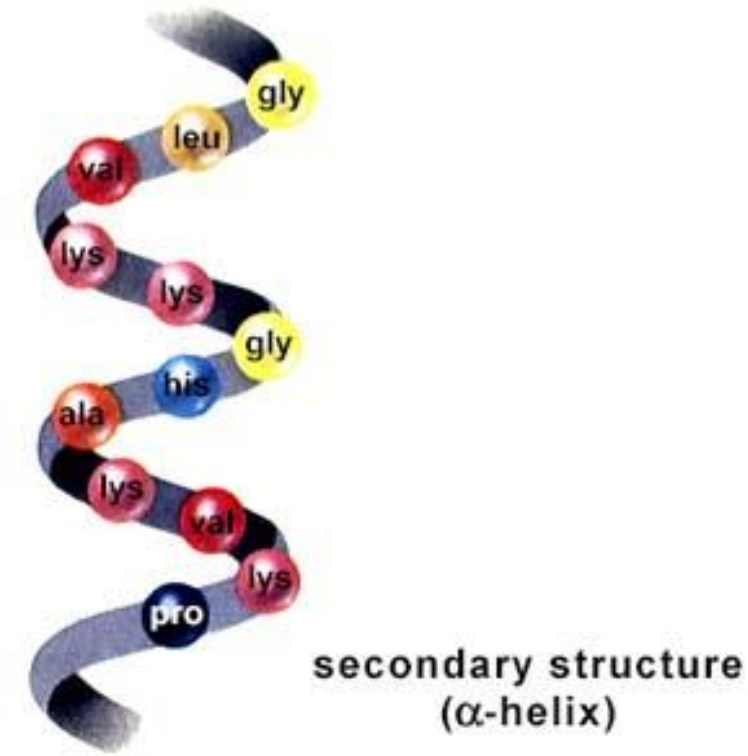


Secondary structure

- Every amino acid in the sequence belongs to one of the three structural motifs
  - α-helix (H)
  - β-sheet (E)
  - Loop or coil (C)
- flatted to a string from an alphabet
  - $\sum_{ss} = \{H, E, C\}$

Secondary structure

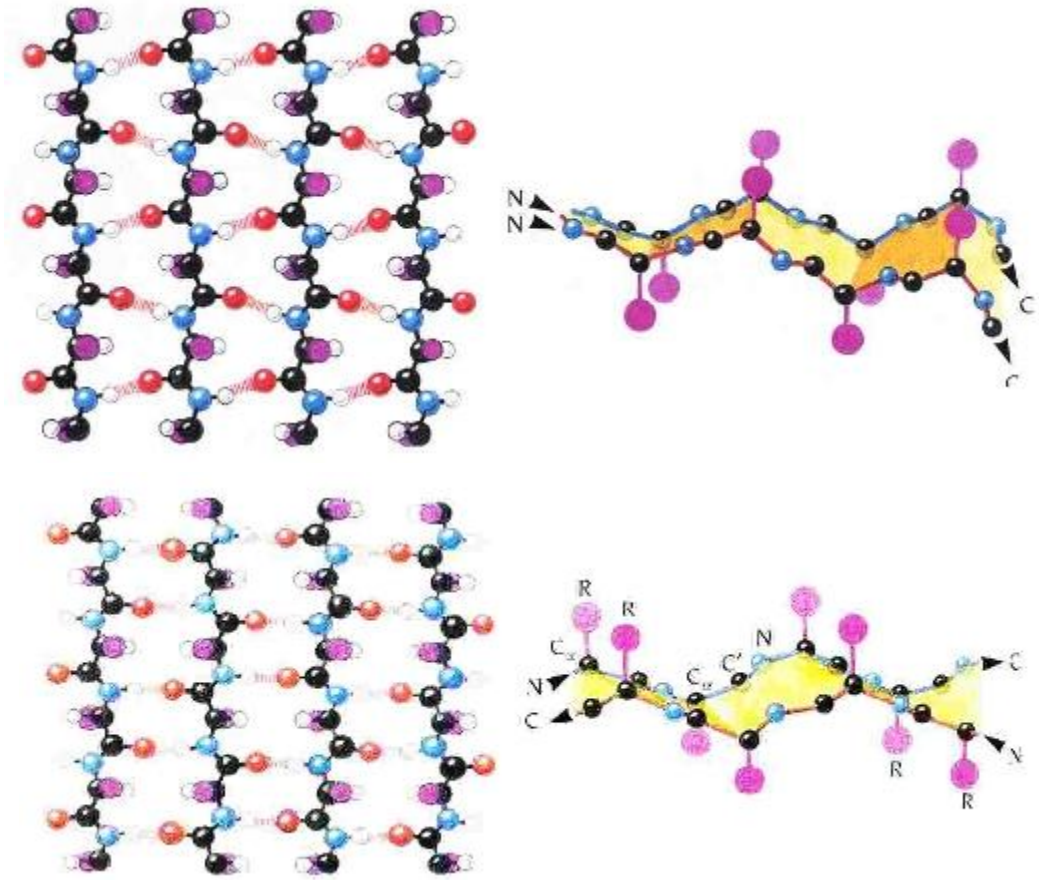
- **α-helix (H)**
  - Built up from one continuous region in the sequence
    - Through the formation of hydrogen bonds between residues in position i and i+4



secondary structure
(α-helix)

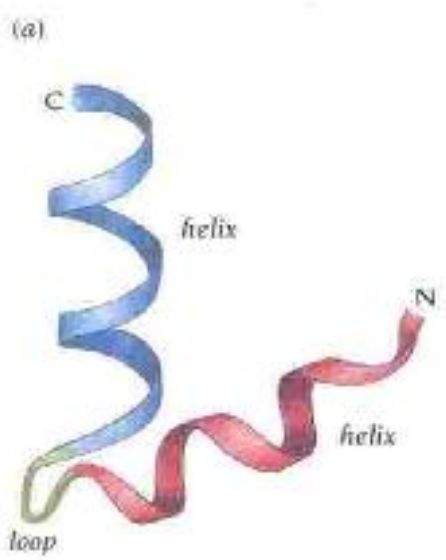
Secondary structure

- **β-sheet (E)**
  - parallel β-sheet
    - Amino acids have the same biochemical direction.
  - Anti parallel β-sheet
    - Amino acids have alternating direction.

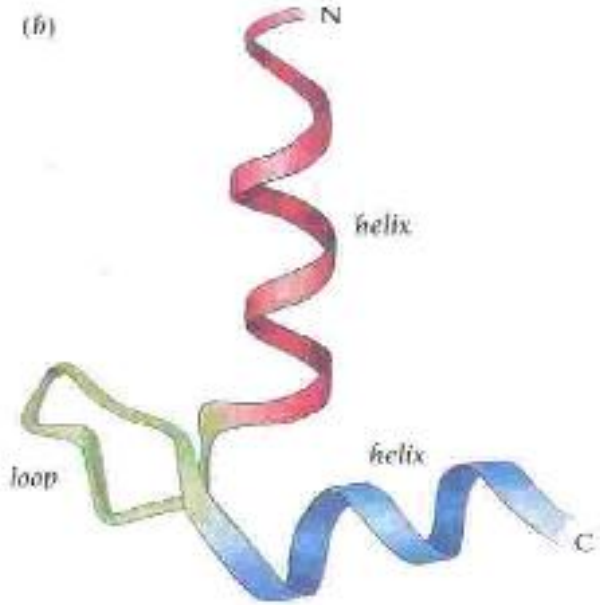


Secondary structure

- **Loop or coil (C)**
  - α-helix and β-sheet are often connected by ***loop regions***.



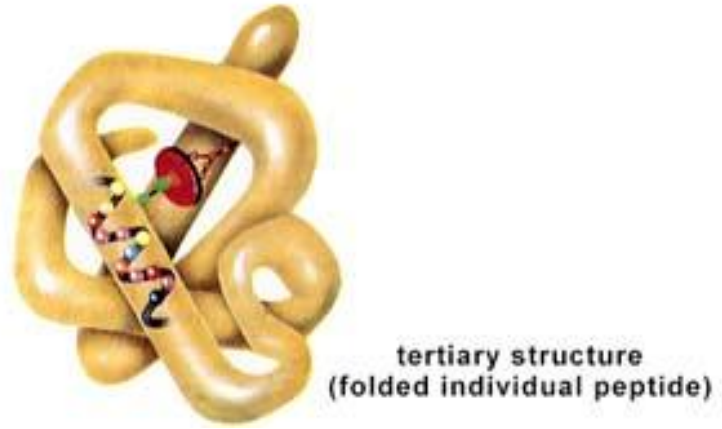
(a) DNA binding motif

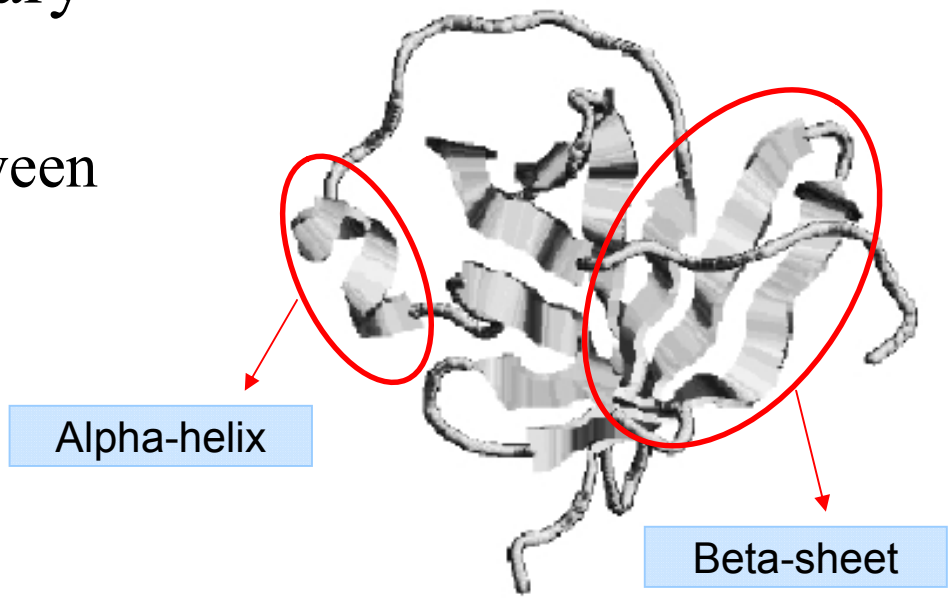


(b) Ca++ binding motif

Tertiary structure

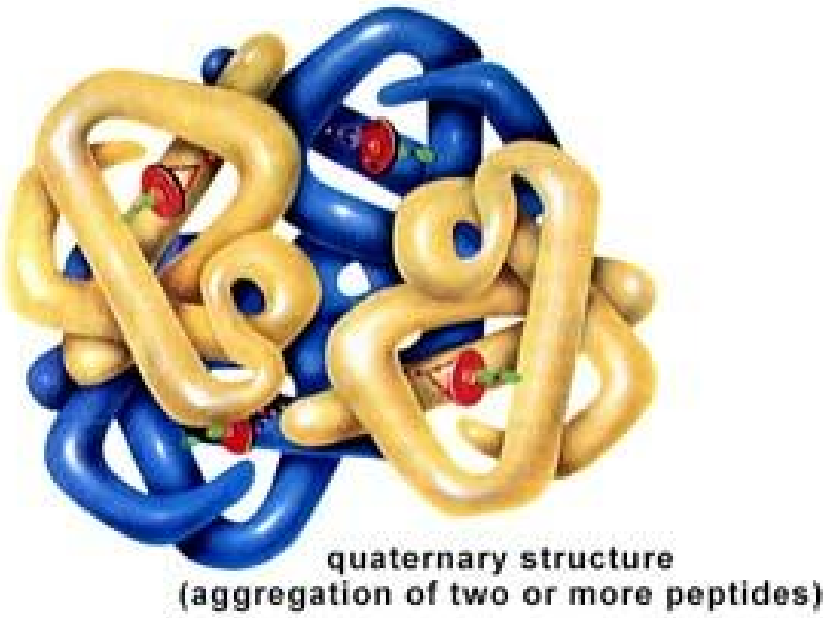
- The 3-dimensional organization of polypeptide chain atoms
- The result of the combinations of secondary structure elements
  - Due to interactions between the amino acids and the solvent



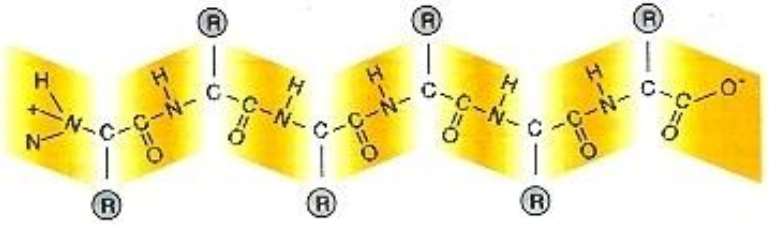


Quaternary structure

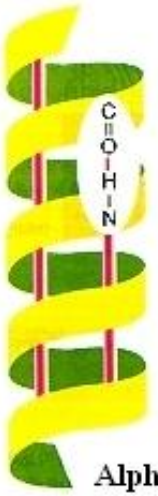
- The complex spatial conformation of a protein composed of many distinct polypeptide chains (*multimeric* protein)



Protein Structure

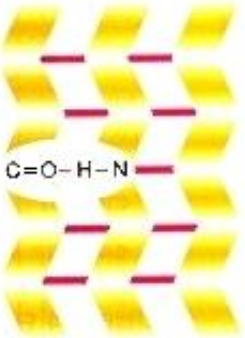


Primary structure
-polypeptide chain-



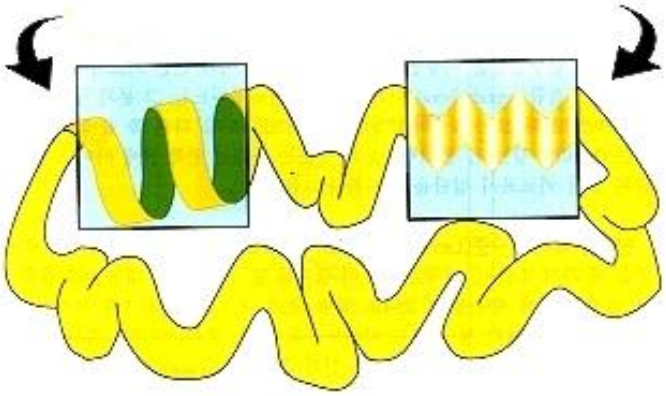
Alpha-helix

Secondary structure



Beta-sheet

Quaternary structure

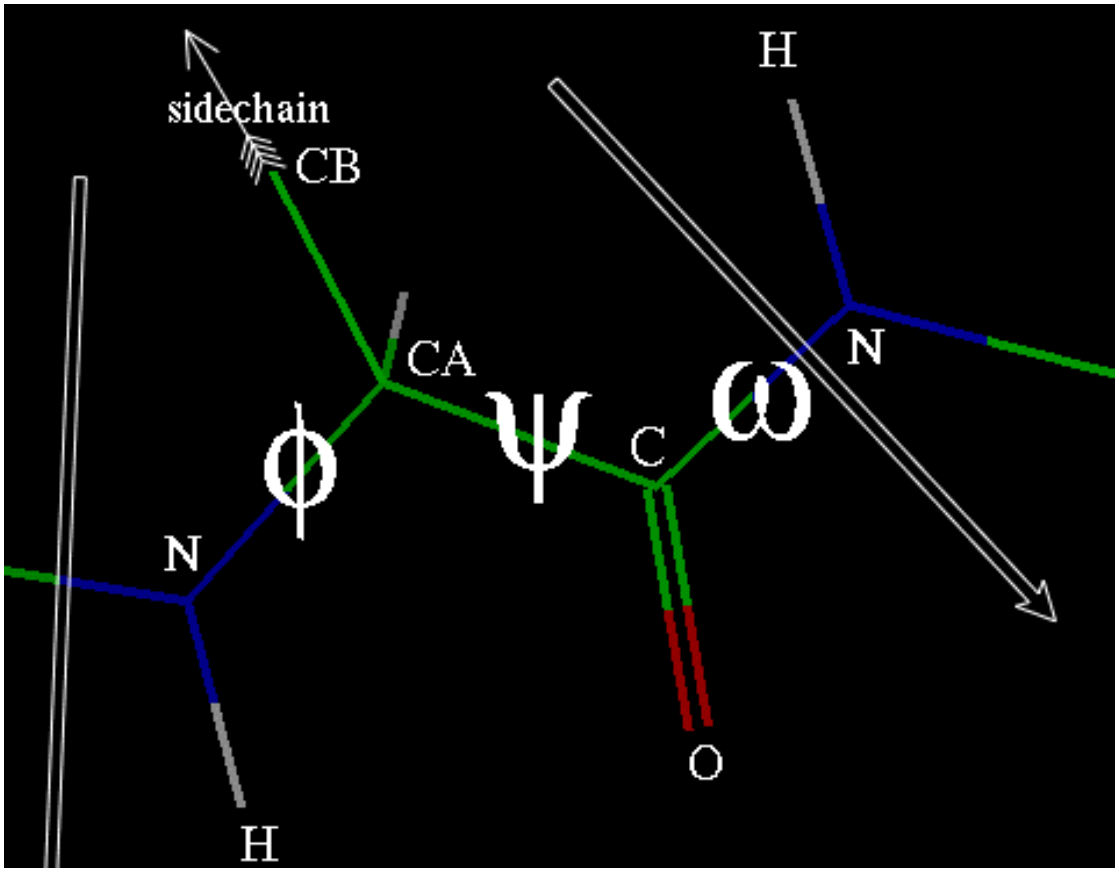


Tertiary structure

Classification of secondary structure

- Defining features
  - Dihedral angles
  - Hydrogen bonds
  - Geometry
- Assigned manually by crystallographers or
- Automatic
  - DSSP (Kabsch & Sander,1983)
  - STRIDE (Frishman & Argos, 1995)
  - Continuum (Andersen et al.)

Dihedral Angles



From http://www.imb-jena.de

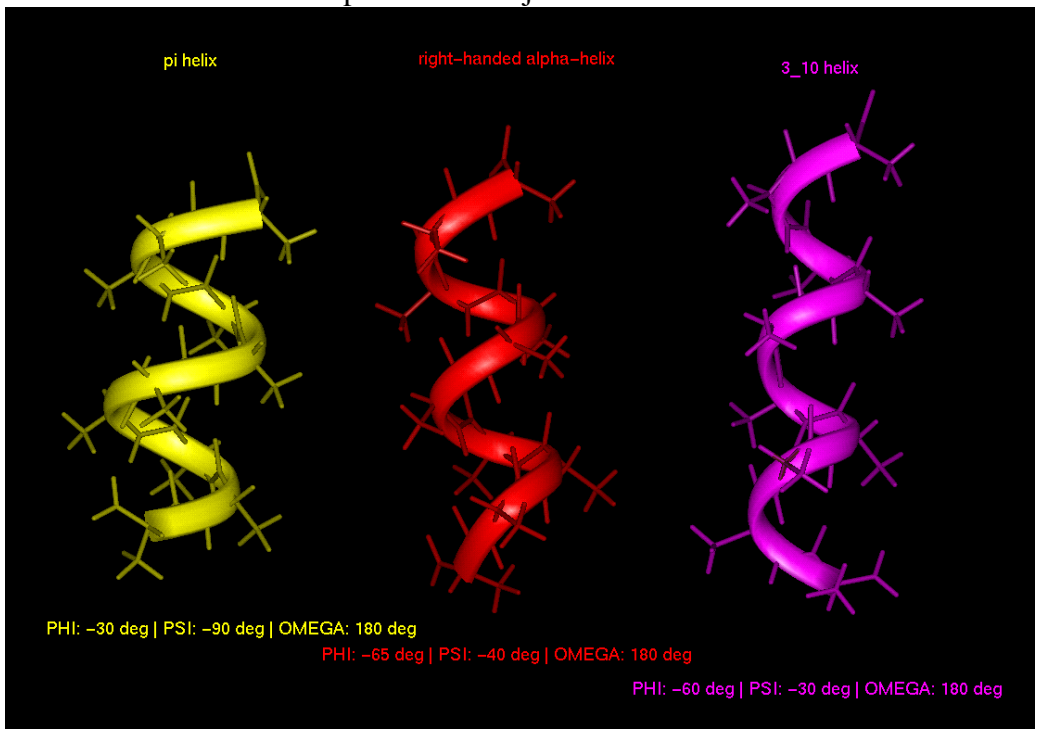
phi - dihedral angle about the N-C*alpha* bond
psi - dihedral angle about the C*alpha*-C bond
omega - dihedral angle about the C-N (peptide) bond

Alpha helices

	phi(deg)	*psi*(deg)	H-bond pattern
right-handed alpha-helix	-57.8	-47.0	i+4
pi-helix	-57.1	-69.7	i+5
3-10 helix	-74.0	-4.0	i+3

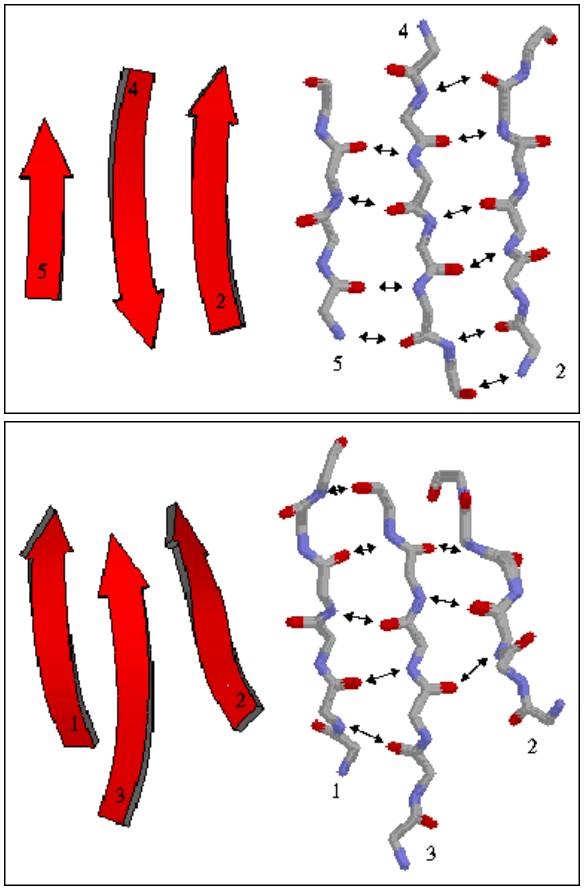
(omega is 180 deg in all cases)

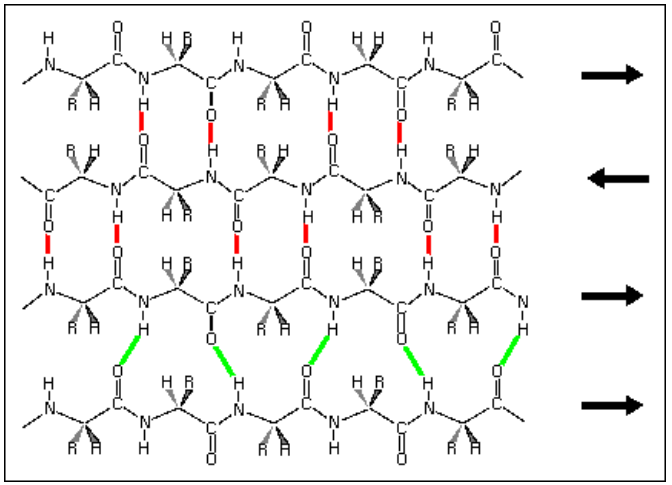
From http://www.imb-jena.de



Beta Strands

	phi(deg)	*psi*(deg)	*omega* (deg)
beta strand	-120	120	180





Hydrogen bond patterns in beta sheets. Here a four-stranded beta sheet is drawn schematically which contains three antiparallel and one parallel strand. Hydrogen bonds are indicated with red lines (antiparallel strands) and green lines (parallel strands) connecting the hydrogen and receptor oxygen.

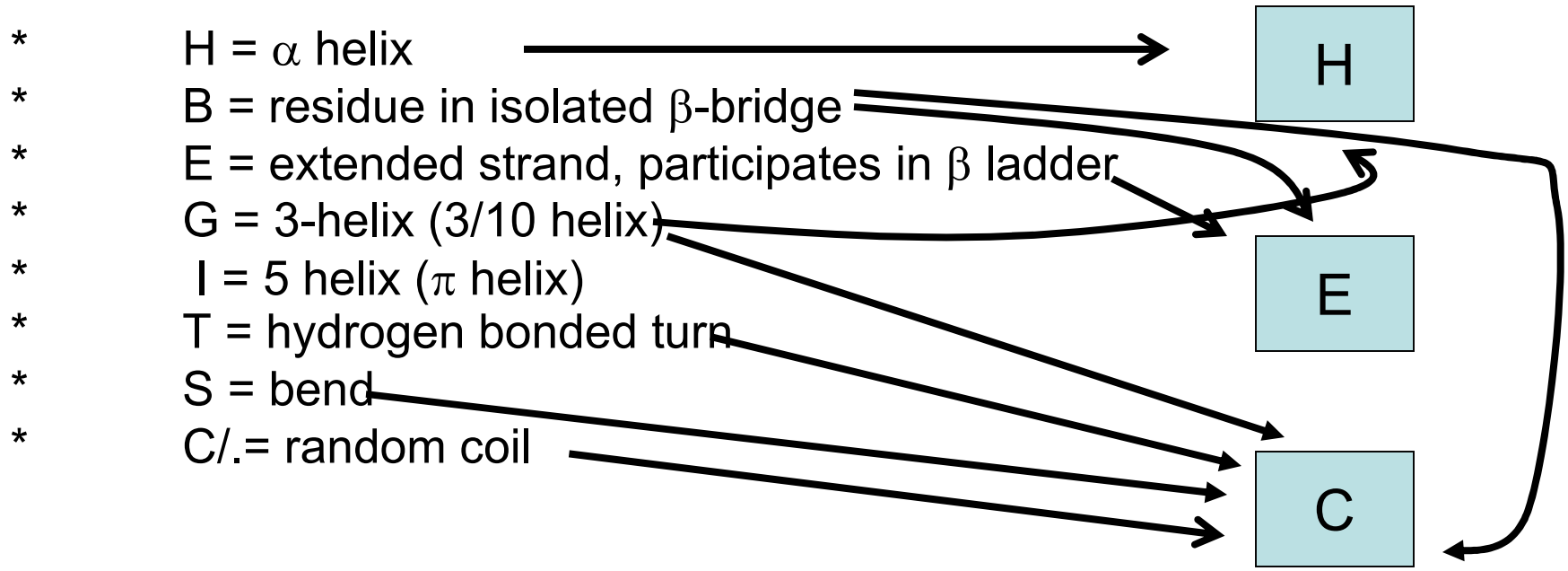
From http://broccoli.mfn.ki.se/pps_course_96/

Secondary Structure Types

* H = alpha helix
* B = residue in isolated beta-bridge
* E = extended strand, participates in beta ladder
* G = 3-helix (3/10 helix)
* I = 5 helix (pi helix)
* T = hydrogen bonded turn
* S = bend

Secondary Structure Prediction

- What to predict?
  - All 8 types or pool types into groups



Simplifications

- Identification of secondary structures focused on
- α-helices
- β -strands
- others (turns, coils, other helices) are collectively called "coils"

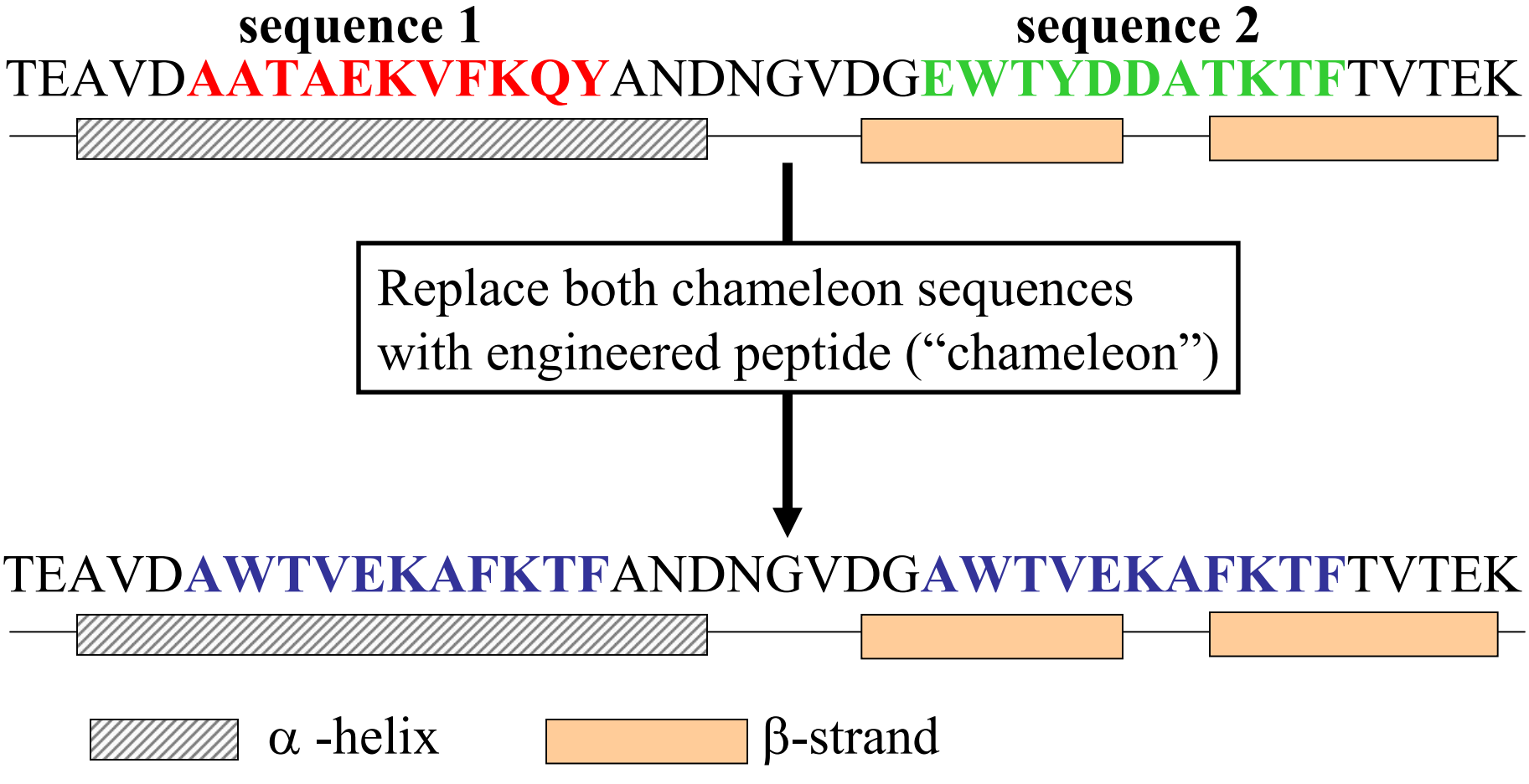
A surprising result !

- Can secondary structure prediction algorithms predict structures of engineered proteins ?

- Test case “the chameleon” sequence

- Algorithm: PHDsec with alignment (PHD 30) and without alignment (PHD no)

The “Chameleon” sequence



Source: Minor and Kim 1996, Nature, 380, 730-734

Secondary Structure Prediction

- Simple alignments.
- Heuristic Methods (e.g., Chou-Fasman, 1974)
- Neural Networks (different inputs)
  - Raw Sequence (late 80's)
  - Blosum matrix (e.g., PhD, early 90's)
  - Position specific alignment profiles (e.g., PsiPred, late 90's)

Improvement of accuracy

1974 Chou - Fasman	~50-53%	
1978 Garnier	63%	
1987 Zvelebil	66%	
1988 Quian - Sejnowski		64.3%
1993 Rost - Sander		70.8-72.0%
1997 Frishman - Argos	<75%	
1999 Cuff - Barton		72.9%
1999 Jones		76.5%

Chou-Fasman

- General applicable
- Works for sequences with no solved homologs
- Low Accuracy

Παράδειγμα

- Ξεκινάμε ορίζοντας ως *fj*(*i*) = τη συχνότητα εμφάνισης του αμινοξέος *i* στην κατάσταση *j* (helix, sheet, turn).
- Στη συνέχεια υπολογίζουμε τη μέση συχνότητα <*fj*> ως τη μέση τιμή όλων των *f* για όλα τα αμινοξέα της κατηγορίας *j*.
- Τέλος, υπολογίζουμε τη στερεοδιαταξική παράμετρο *Pj*(*i*) για κάθε αμινοξύ *i* και κατάσταση *j* ως *Pj*(*i*) =*fj*(*i*)/<*fj*>.
- Για παράδειγμα, στο σύνολο εκπαίδευσης υπήρχαν 228 Αλανίνες (119 σε α-έλικα, 38 σε β-πτυχωτή επιφάνεια και 71 σε τυχαία δομή).
- Άρα, οι παράμετροι θα είναι *fH*(*A*) = 0.522, *fE*(*A*) = 0.167 και *fC*(*A*) *fc* = 0.311. Για την α-έλικα οι μέσες τιμές είναι <*fH*>=890/2473 = 0.359, για τη β-πτυχωτή επιφάνεια <*fE*>=424/2473 = 0.171 και για την τυχαία δομή, <*fC*>=1159/2473 = 0.469.
- Κατά συνέπεια, οι στερεοδιαταξικές παράμετροι για την Αλανίνη θα είναι *PH*(A)= 0.522/0.359 = 1.45, *PE*(A)= 0.167/0.171 = 0.97 και *PC*(A)=0.311/0.469 = 0.63.

Πίνακας

aminoacid	P(helix)	P(sheet)	P(coil)
A (Ala)	1.420	0.830	0.660
R (Arg)	0.980	0.930	0.950
N (Asn)	0.670	0.890	1.560
D (Asp)	1.010	0.540	1.460
C (Cys)	0.700	1.190	1.190
Q (Gln)	1.110	1.100	0.980
E (Glu)	1.510	0.370	0.740
G (Gly)	0.570	0.750	1.560
H (His)	1.000	0.870	0.950
I (Ile)	1.080	1.600	0.470
L (Leu)	1.210	1.300	0.590
K (Lys)	1.160	0.740	1.010
M (Met)	1.450	1.050	0.600
F (Phe)	1.130	1.380	0.600
P (Pro)	0.570	0.550	1.520
S (Ser)	0.770	0.750	1.430
T (Thr)	0.830	1.190	0.960
W (Trp)	1.080	1.370	0.960
Y (Tyr)	0.690	1.470	1.140
V (Val)	1.060	1.700	0.500

Άλλες μέθοδοι

- Chou-Fasman http://cho-fas.sourceforge.net/
- **GOR IV** (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html), η οποία κάνει χρήση μόνο της αμινοξικής αλληλουχίας, να φτάνει ένα ποσοστό σωστών προγνώσεων της τάξης του 64%,
- με την **GOR V** (http://gor.bb.iastate.edu/), η οποία χρησιμοποιεί πολλαπλές στοιχίσεις με τη μορφή προφίλ του PSI-BLAST, φτάνει πλέον σε ένα ποσοστό ακρίβειας της τάξης του 74%.

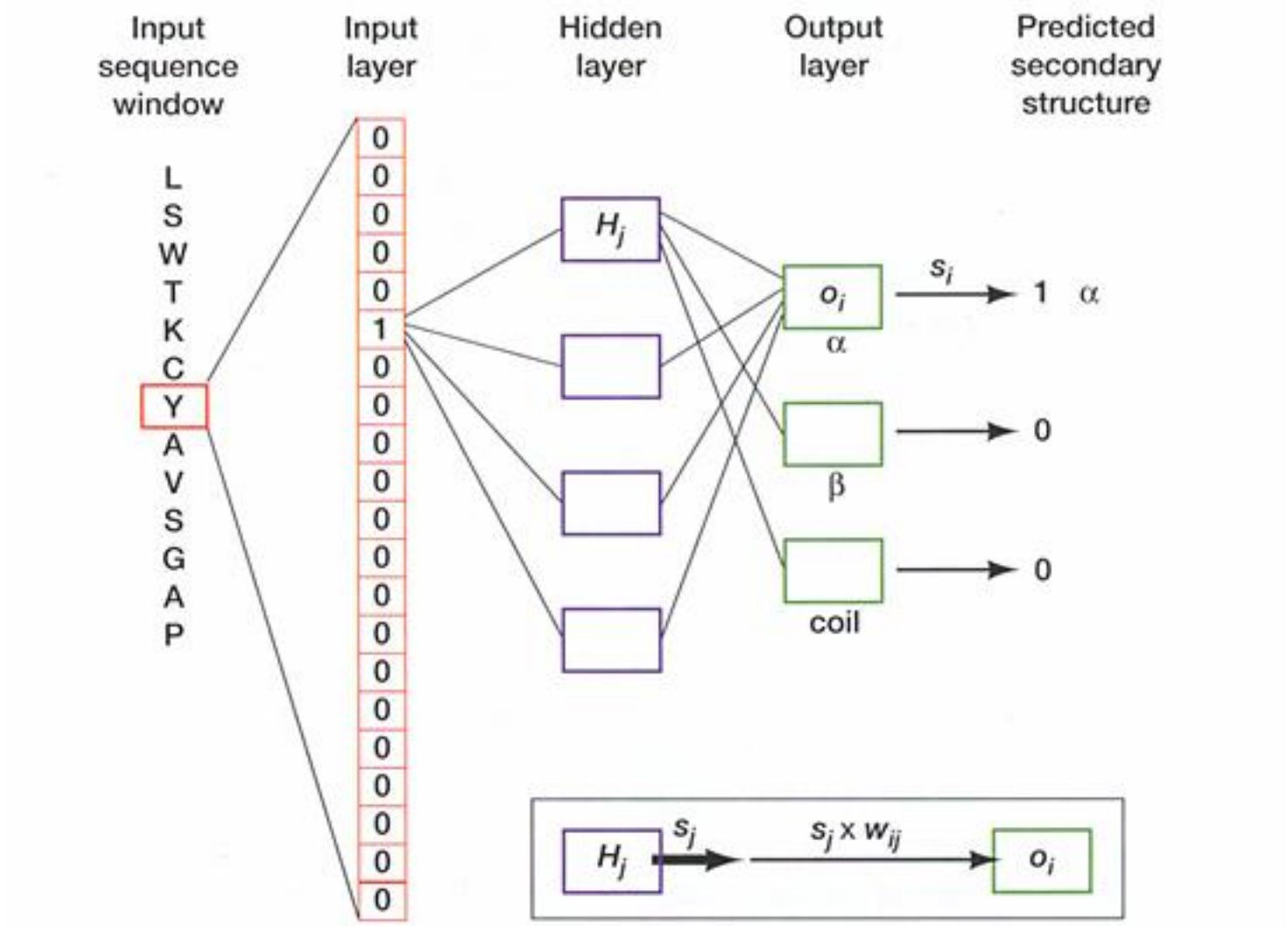
Neural Networks

- Benefits
  - General applicable
  - Can capture higher order correlations
  - Inputs other than sequence information
- Drawbacks
  - Needs many data (different solved structures)
  - Risk of overtraining

Measuring prediction accuracy

- not a clear cut process
- traditional Qindex and Q3
- SOV (Segment Overlap) measure

  *- Rost et al. - JMB. **1994**, 235, 13-26*
- Correlation coefficient
  - *Mathews 1975*


Input sequence window
Input layer
Hidden layer
Output layer
Predicted secondary structure
L
S
W
T
K
C
Y
A
V
S
G
A
P
0
0
0
0
0
1
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
H_j
o_i
α
β
coil
s_i
1 α
0
0
H_j
s_j
$s_j \times w_{ij}$
o_i

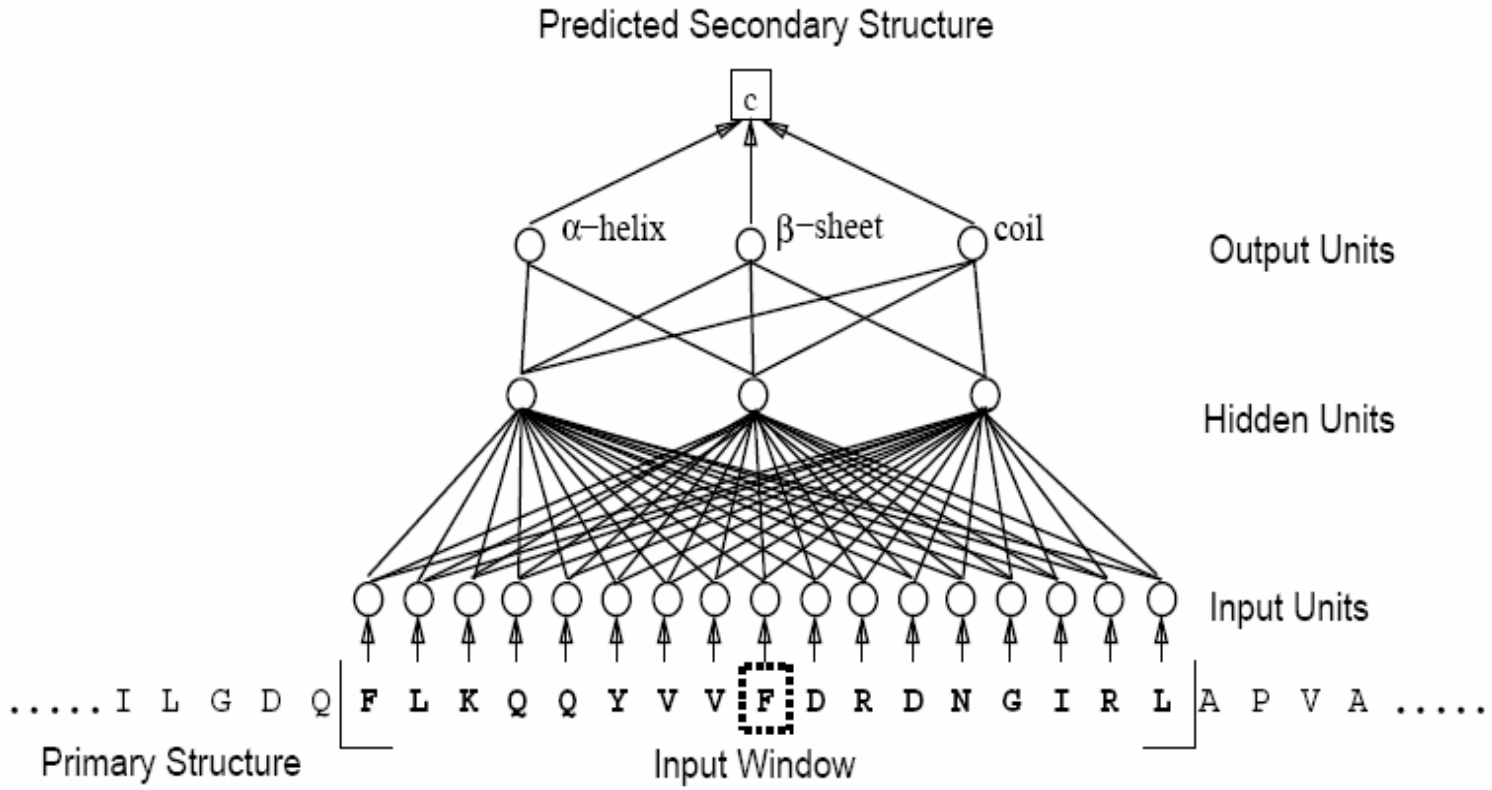
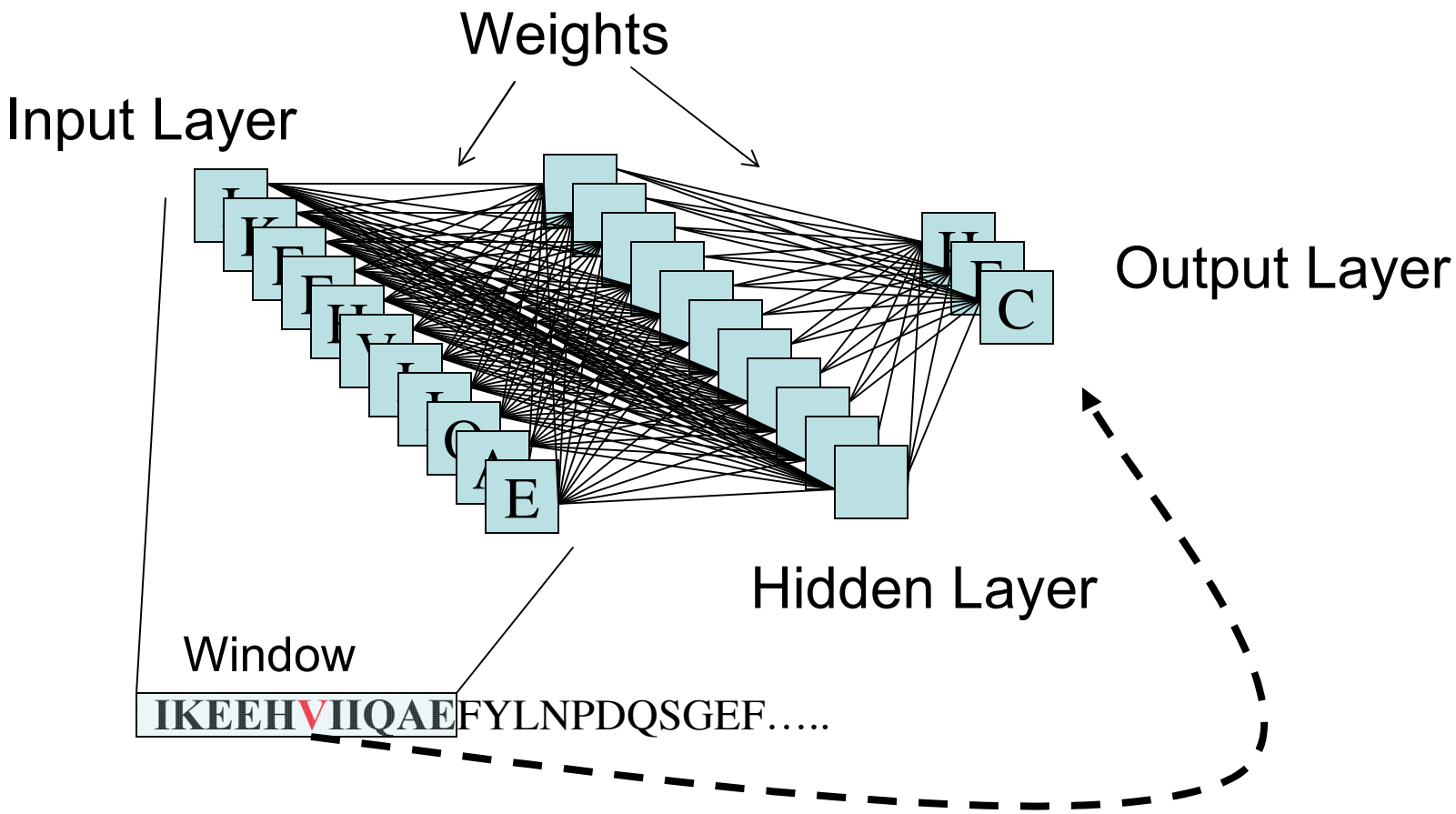


Figure 5. General neural network architecture used by Holley & Karplus and Qian & Sejnowski.

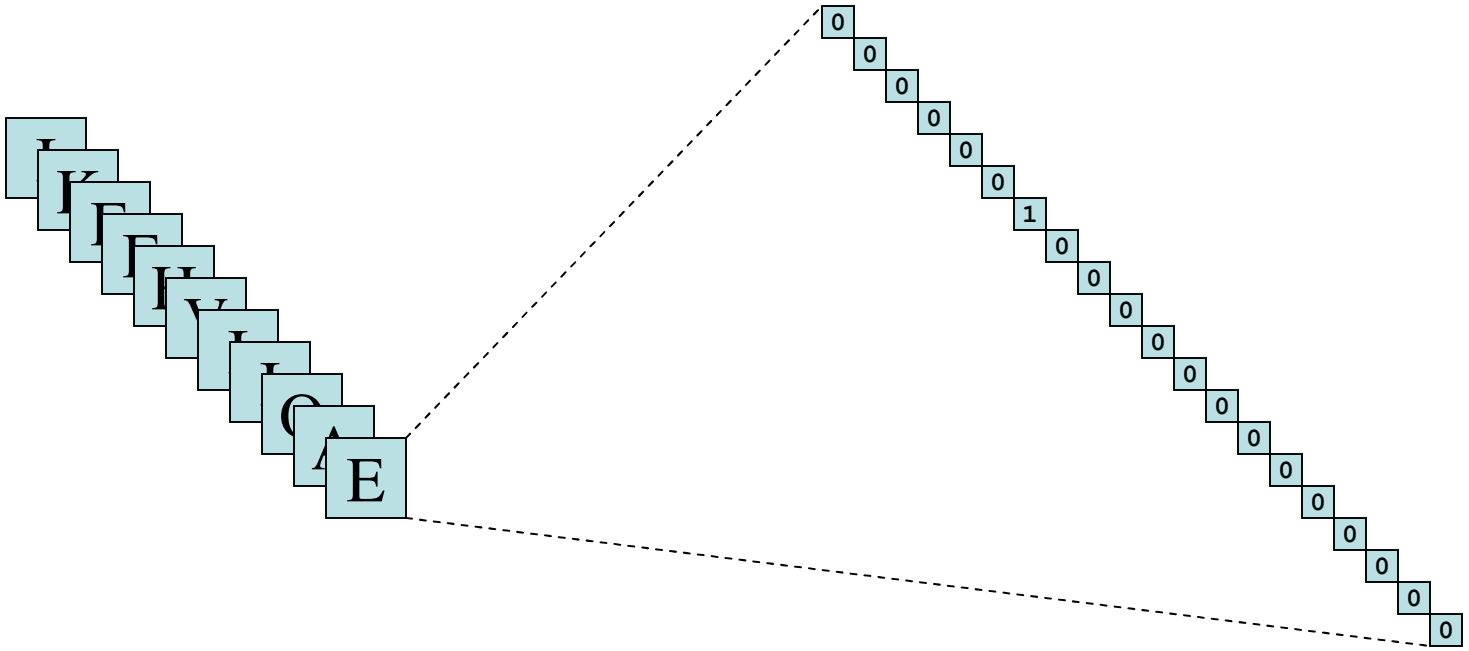
Architecture



Sparse encoding

Inp Neuron	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
AAcid																				
A	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

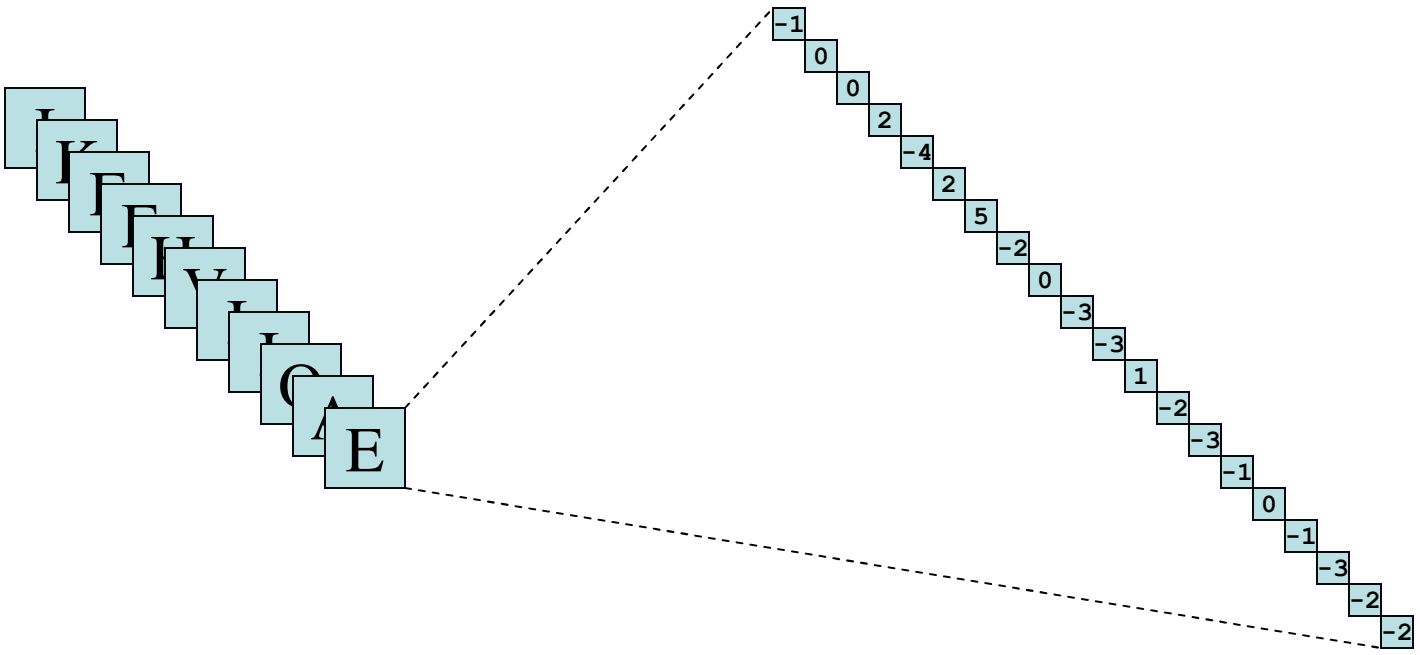
Input Layer



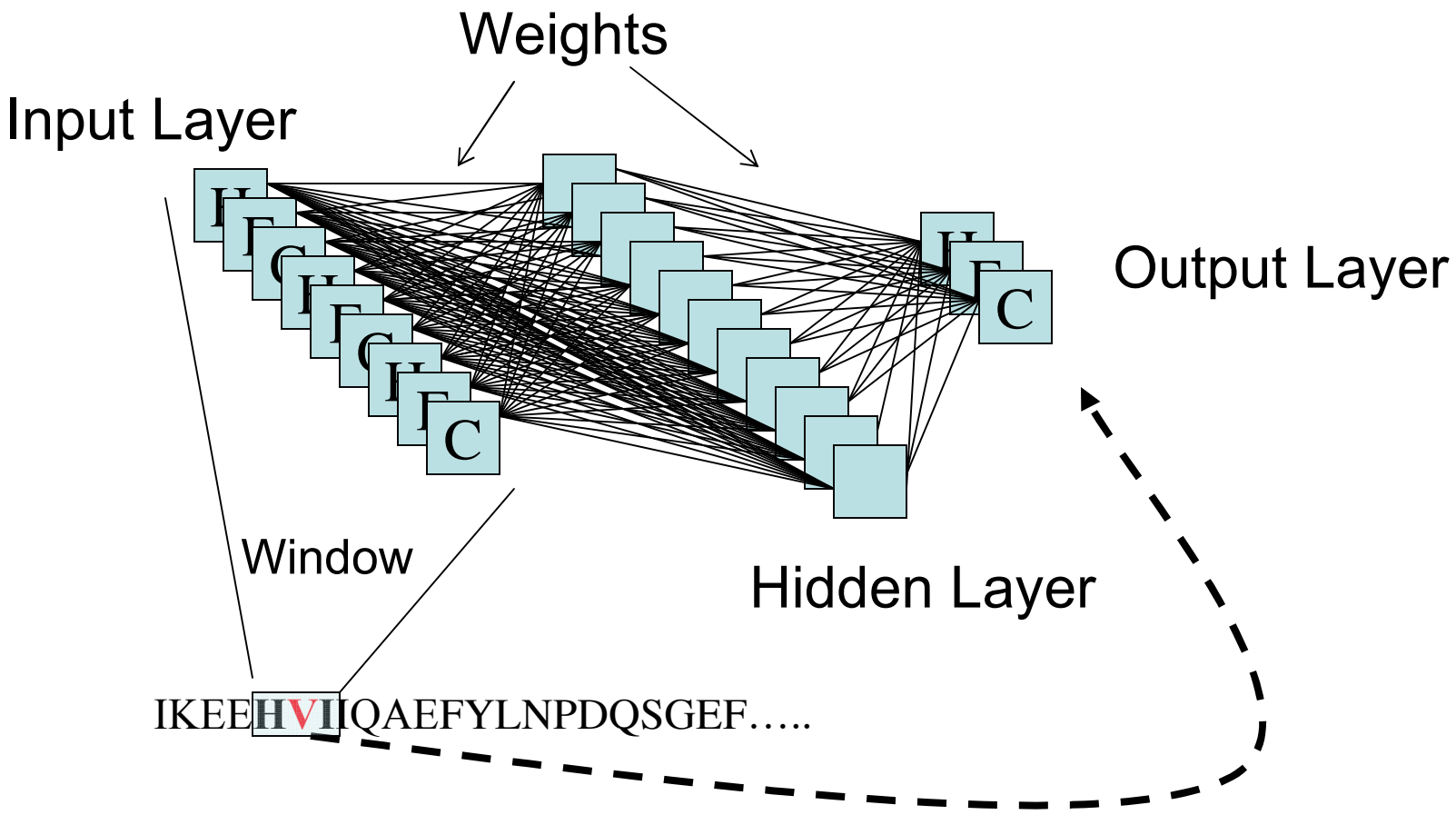
BLOSUM 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4

Input Layer



Structure to Structure



PHD method (Rost and Sander)

- Combine neural networks with sequence profiles
  - 6-8 Percentage points increase in prediction accuracy over standard neural networks
- Use second layer “Structure to structure” network to filter predictions
- Jury of predictors
- Set up as mail server

Position specific scoring matrices
(BLAST profiles)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	I	-2	-4	-5	-5	-2	-4	-4	-5	-5	6	0	-4	0	-2	-4	-4	-2	-4	-3	4
2	K	-1	-1	-2	-2	-3	-1	3	-3	-2	-2	-3	4	-2	-4	-3	1	1	-4	-3	2
3	E	5	-3	-3	-3	-3	3	1	-2	-3	-3	-3	-2	-2	-4	-3	-1	-2	-4	-3	1
4	E	-4	-3	2	5	-6	1	5	-4	-3	-6	-6	-2	-5	-6	-4	-2	-3	-6	-5	-5
5	H	-4	2	1	1	-5	1	-2	-4	9	-5	-2	-3	-4	-4	-5	-3	-4	-5	1	-5
6	V	-3	0	-4	-5	-4	-4	-2	-3	-5	1	-2	1	0	1	-4	-3	3	-5	-3	5
7	I	0	-2	-4	1	-4	-2	-4	-4	-5	1	0	-2	0	2	-5	1	-1	-5	-3	4
8	I	-3	0	-5	-5	-4	-2	-5	-6	1	2	4	-4	-1	0	-5	-2	0	-3	5	-1
9	Q	-2	-3	-2	-3	-5	4	-1	3	5	-5	-3	-3	-4	-2	-4	2	-1	-4	2	-2
10	A	2	-4	-4	-3	2	-3	-1	-4	-2	1	-1	-4	-3	-4	1	2	3	-5	-1	1
11	E	-1	3	1	1	-1	0	1	-4	-3	-1	-3	0	3	-5	4	-1	-3	-6	-3	-1
12	F	-3	-5	-5	-5	-4	-4	-4	-1	-1	1	1	-5	2	5	-1	-4	-4	-3	5	2
13	Y	3	-5	-5	-6	3	-4	-5	-2	-1	0	-4	-5	-3	3	-5	-2	-2	-2	7	1
14	L	-1	-3	-4	-2	1	5	1	-1	-1	-1	1	-3	-3	1	-5	-1	-1	-2	3	-2
15	N	-1	-4	4	1	5	-3	-4	2	-4	-4	-4	-3	-2	-4	-5	2	0	-5	0	0
16	P	-2	4	-4	-4	-5	0	-3	3	2	-5	-4	0	-4	-3	0	1	-2	-1	5	-3
17	D	-3	-2	1	5	-6	-2	2	2	-1	-2	-2	-3	-5	-4	-5	-1	2	-6	-3	-4

PSI-Pred (Jones, DT)

- Use alignments from iterative sequence searches (PSI-Blast) as input to a neural network
- Better predictions due to better sequence profiles
- Available as stand alone program and via the web

Several different architectures

- Sequence-to-structure
  - Window sizes 15,17,19 and 21
  - Hidden units 50 and 75
  - 10-fold cross validation => 80 predictions
- Structure-to-structure
  - Window size 17
  - Hidden units 40
  - 10-fold cross validation => 800 predictions

Αλλες μέθοδοι

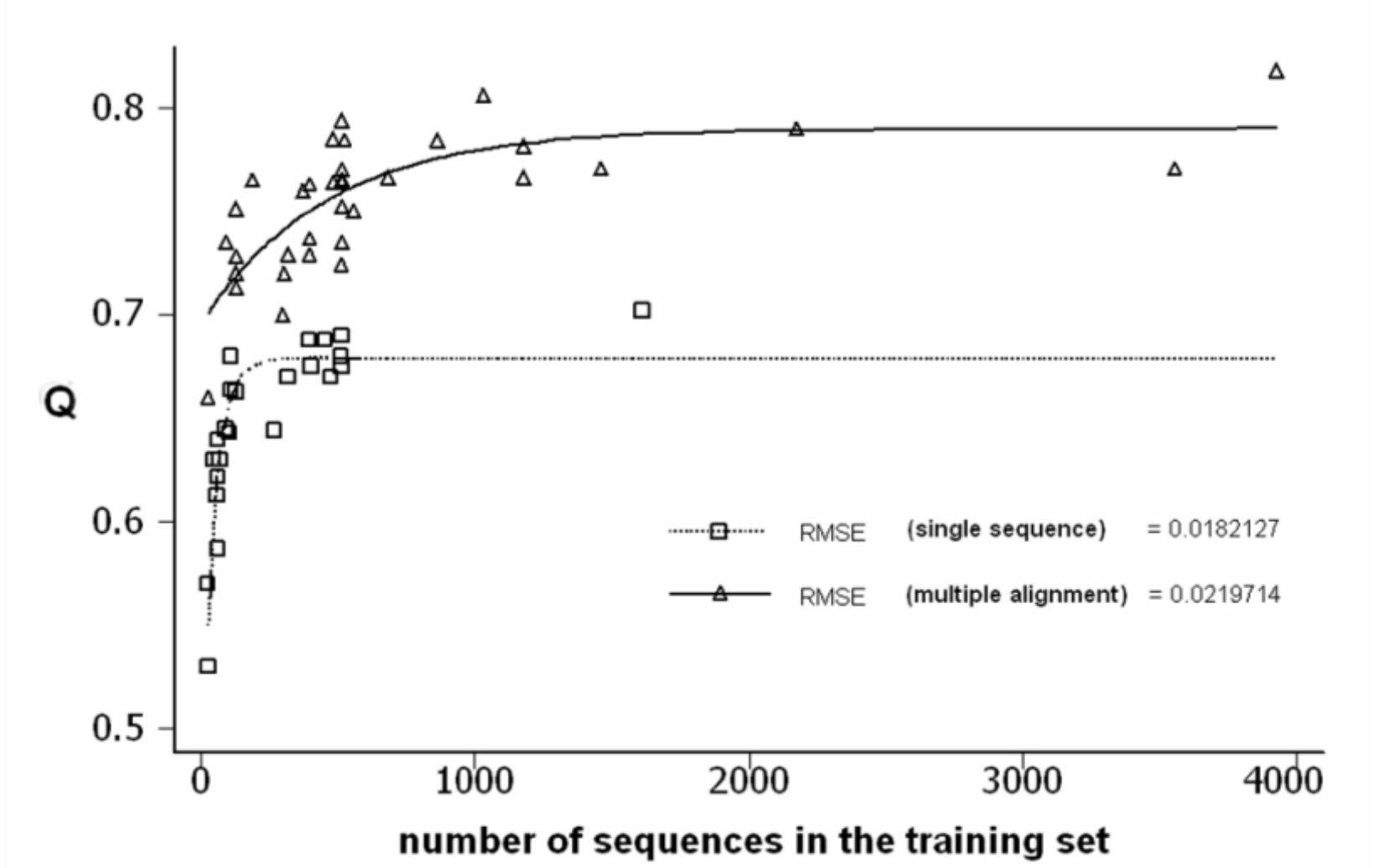
- Μια από τις πρώτες προσπάθειες είχε γίνει το 1988 όταν ο Hamodrakas (Hamodrakas, 1988) δημοσίευσε ένα συνδυαστικό αλγόριθμο που έκανε χρήση των τότε διαθέσιμων μεθόδων (Chou-Fasman, GOR, Lim, Dufton-Hider, Burgess, Nagano). Η μέθοδος αυτή έδειξε μια βελτίωση της τάξης του 2-3% και μετέπειτα έγινε και διαθέσιμη σαν διαδικτυακή εφαρμογή με το όνομα **SecStr** (http://athina.biol.uoa.gr/SecStr/).
- Το **JPRED** (http://www.compbio.dundee.ac.uk/jpred/) ήταν ίσως η πρώτη μέθοδος που χρησιμοποίησε συνδυασμό μεθόδων και ταυτόχρονα έκανε χρήση εξελικτικής πληροφορίας το 1998 (Cuff, Clamp, Siddiqui, Finlay, & Barton, 1998). Στην πρώτη έκδοση έκανε χρήση του JNET και μιας σειράς άλλων αλγορίθμων της εποχής (NNSSP, DSC, PREDATOR, MULPRED, PHD, ZPRED) και ανέφερε σημαντικά βελτιωμένη απόδοση. Σήμερα, η μέθοδος έχει φτάσει στην έκδοση 4 (JPRED4) και συγκαταλέγεται ανάμεσα στις καλύτερες μεθόδους, έχοντας αυτοματοποιημένη πρόσβαση μέσω διαδικτυακής εφαρμογής και πολλές επιλογές, όπως γραφικές παραστάσεις των αποτελεσμάτων ή τη δυνατότητα ο χρήστης να δώσει τη δική του πολλαπλή στοίχιση.
- Μια άλλη γνωστή από παλιά συνδυαστική μέθοδος είναι η **NPS@** (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html) η οποία κάνει συνδυαστική πρόγνωση με χρήση των μεθόδων SOPM, SOPMA, HNN, MLRC, DPM, DSC, GOR I, GOR III, GOR IV, PHD, PREDATOR, SIMPA96 ενώ δίνει στο χρήστη τη δυνατότητα να επιλέξει ποιες από αυτές θα χρησιμοποιηθούν.
- Άλλες πιο πρόσφατες συνδυαστικές μέθοδοι είναι το **CONCORD** (http://helios.princeton.edu/CONCORD/) το οποίο χρησιμοποιεί τα PSIPRED, DSC, GOR IV, Predator, Prof, PROFphd, και SSpro, και το **SYMPRED** (http://www.ibi.vu.nl/programs/sympredwww/) το οποίο κάνει χρήση των PHDpsi, PROFsec, SSPro, Predator, YASPIN, JNet και PSIPRED.
- Πρέπει να τονίσουμε σε αυτό το σημείο, ότι η σύγχρονη τάση των μεγάλων εργαστηρίων είναι να διαθέτουν σε μια διαδικτυακή εφαρμογή όλες τις σχετικές μεθόδους πρόγνωσης (δευτεροταγούς δομής, προσβασιμότητας του διαλύτη, διαμεμβρανικών τμημάτων κ.ο.κ.). Έτσι, οι μέθοδοι του B. Rost βρίσκονται όλες μαζί στην ιστοσελίδα **PREDICTPROTEIN** (www.predictprotein.org/), στην ιστοσελίδα του **PSI-PRED** (http://bioinf.cs.ucl.ac.uk/psipred/) διατίθενται εκτός από την ομώνυμη εφαρμογή και άλλες μέθοδοι πρόγνωσης πρωτεϊνών του εργαστηρίου, ενώ αντίστοιχες μέθοδοι διατίθενται στο **SCRATCH** (http://scratch.proteomics.ics.uci.edu/index.html).

Benchmarking secondary structure predictions

- Από τη δεκαετία του 1990 οι επιστήμονες δημιούργησαν το συνέδριο **CASP** (Critical Assessment of Structure Predictions http://predictioncenter.org/). Σε αυτή την προσπάθεια, εντοπίζονται μετά από επικοινωνία με τους κρυσταλλογράφους οι αλληλουχίες των πρωτεϊνών που είναι «έτοιμες» να προσδιοριστούν πειραματικά. Αφού ελεγχθεί ότι οι αλληλουχίες αυτές δεν εμφανίζουν ομοιότητα με καμία άλλη πρωτεΐνη γνωστής δομής, οι αλληλουχίες ανακοινώνονται και οι διάφοροι αλγόριθμοι δοκιμάζονται. Όταν φτάσει ο καιρός του συνεδρίου τα αποτελέσματα των αλγορίθμων ανακοινώνονται και συγκρίνονται με τις πραγματικές δομές που στο μεταξύ έχουν προσδιοριστεί αλλά παραμένουν μυστικές.
- Μια άλλη προσπάθεια για συνεχή παραγωγή τέτοιων ανεξάρτητων συνόλων, είχε δημιουργήσει ο Rost. Το πρόγραμμα ονομάζεται **EVA** (Koh et al., 2003) και πραγματοποιούσε κάθε μήνα αναζήτηση στην PDB για νέες δομές και πραγμάτωνε τη σύγκριση με τα γνωστά σύνολα εκπαίδευσης όλων ή των περισσότερων, γνωστών μεθόδων. Έτσι, υπάρχει ένα συνεχώς ανανεωνόμενο σύνολο ανεξάρτητου ελέγχου για κάθε μέθοδο, οπότε με σύγκριση των συνόλων αυτών θα μπορεί ανά πάσα στιγμή να κατασκευαστεί ένα σύνολο που να είναι κατάλληλο για τη σύγκριση δύο ή περισσότερων αλγορίθμων.

EVA results (Rost et al., 2001)

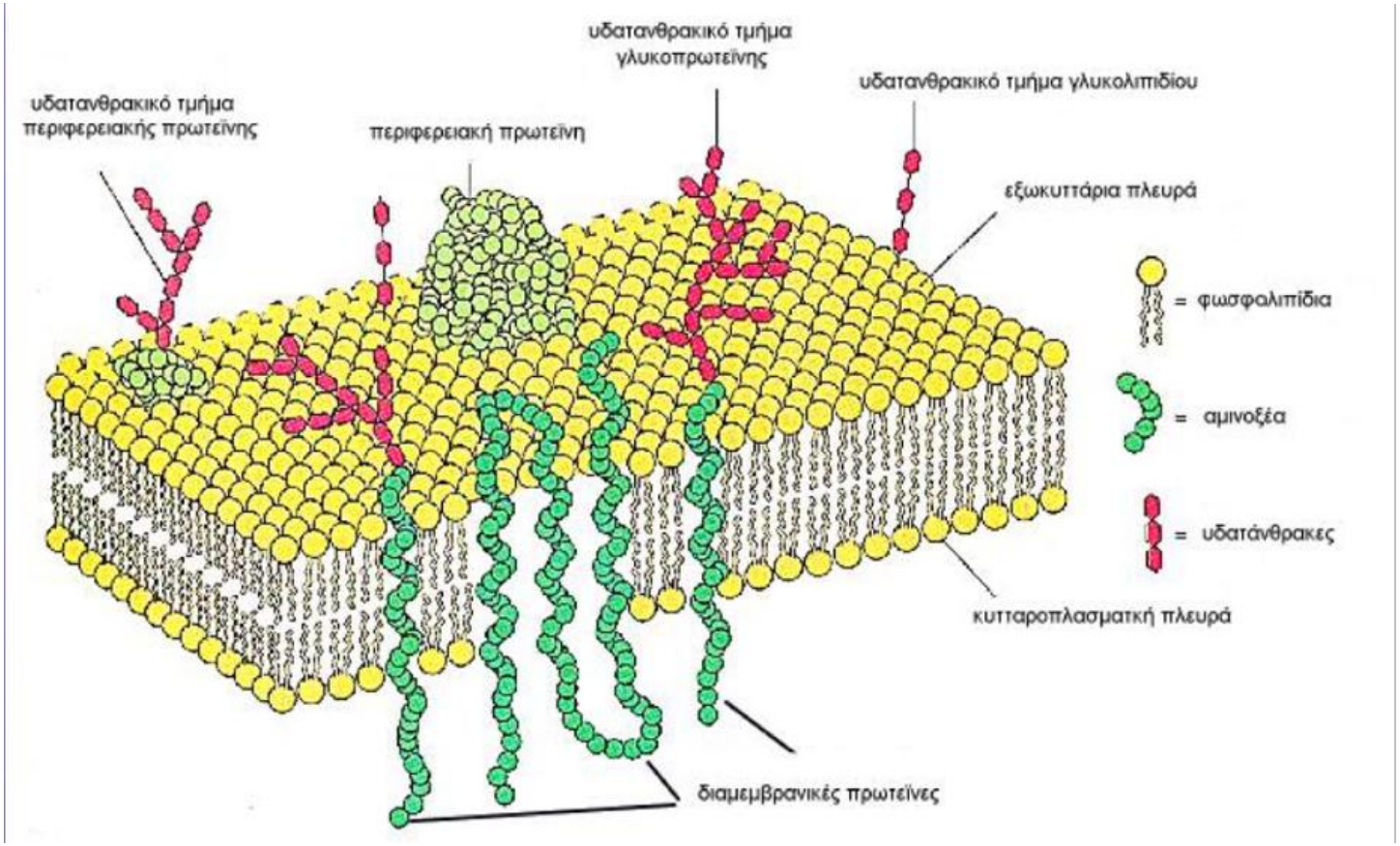
- PROFphd 77.0%
- PSIPRED 76.8%
- SAM-T99sec 76.1%
- SSpro 76.0%
- Jpred2 75.5%
- PHD 71.7%
  - Cubic.columbia.edu/eva


0.8
0.7
0.6
0.5
Q
0
1000
2000
3000
4000
number of sequences in the training set
RMSE (single sequence) = 0.0182127
RMSE (multiple alignment) = 0.0219714

Practical Conclusion

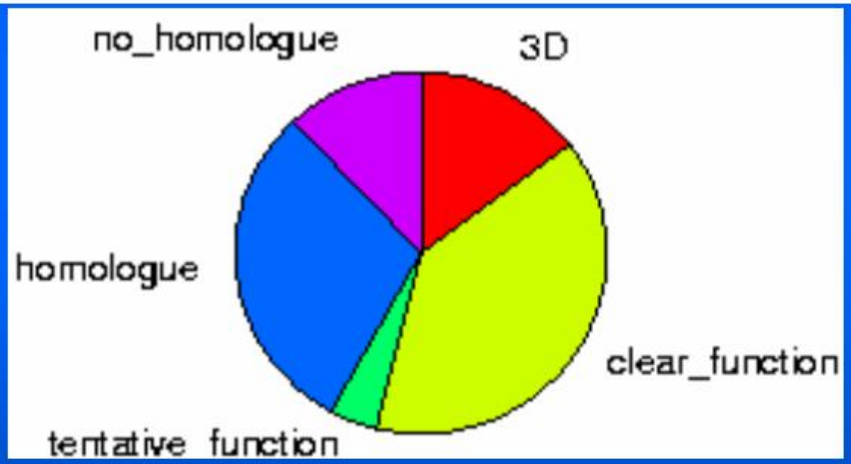
- If you need a secondary structure prediction use one of the newer ones such as
  - ProfPHD,
  - PSIPRED, and
  - JPred
- And *not* one of the older ones such as
  - Chou-Fasman, and
  - Garnier

Διαμεμβρανικά τμήματα



Why do we need prediction methods ?

Automatic similarity-based annotation for the complete genome of the extremophile archaeon *Methanococcus jannaschii* by GeneQuiz. http://jura.ebi.ac.uk:8765/ext-genequiz//genomes/mj0005/index.html



- **Sequence Similarity Methods, frequently, fail to aid in gene and/or protein characterization**
- **Predictive Methods are recruited to 'fill the gap'**
- **Prediction of 2D-3D features is helpful**

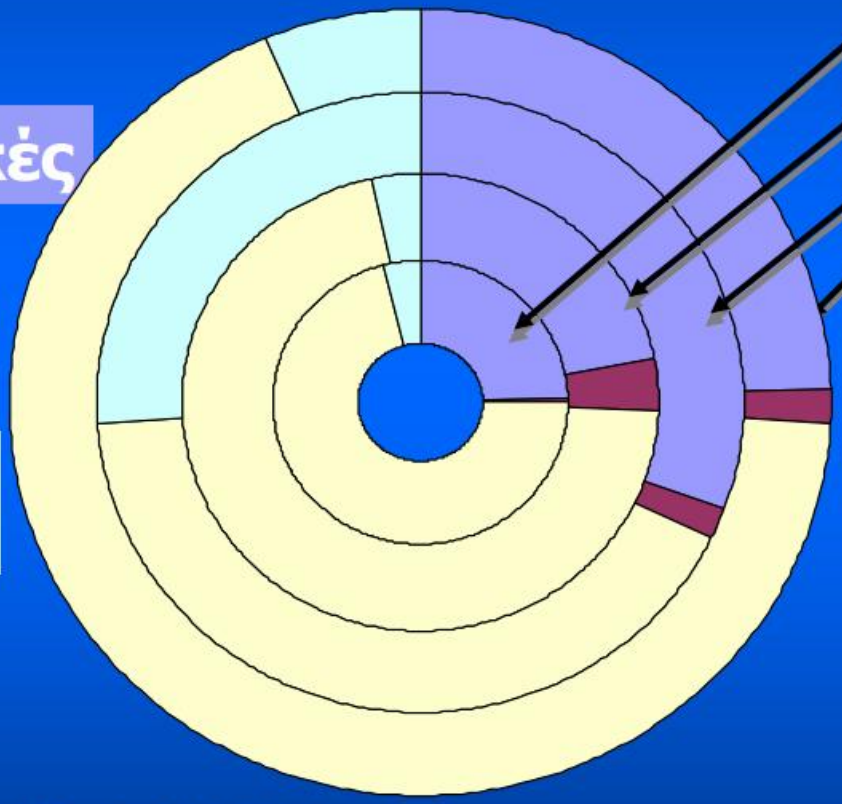
Διαμεμβρανικές πρωτεΐνες σε πλήρη γονιδιώματα

Διαμεμβρανικές

Ινώδεις

Σφαιρικές

Μικτές
(Σφαιρικές + Ινώδεις)

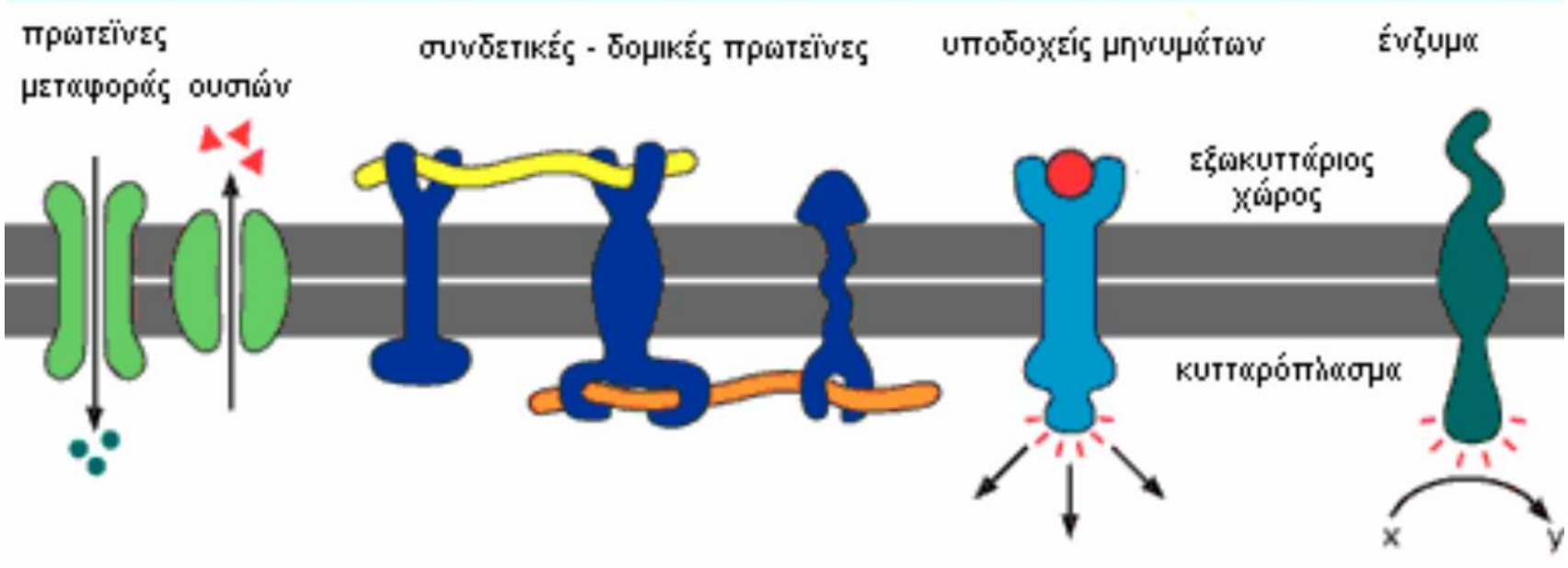


Βακτήρια (21)
Αρχαία (6)
Ευκάρυα (3)
Μέσος όρος

Πρόγνωση με το PRED-CLASS
Pasquier C, Promponas VJ, Hamodrakas SJ (2001),
Proteins; **44(3): 361-369.**

Λειτουργίες «ενσωματωμένων» μεμβρανικών πρωτεϊνών

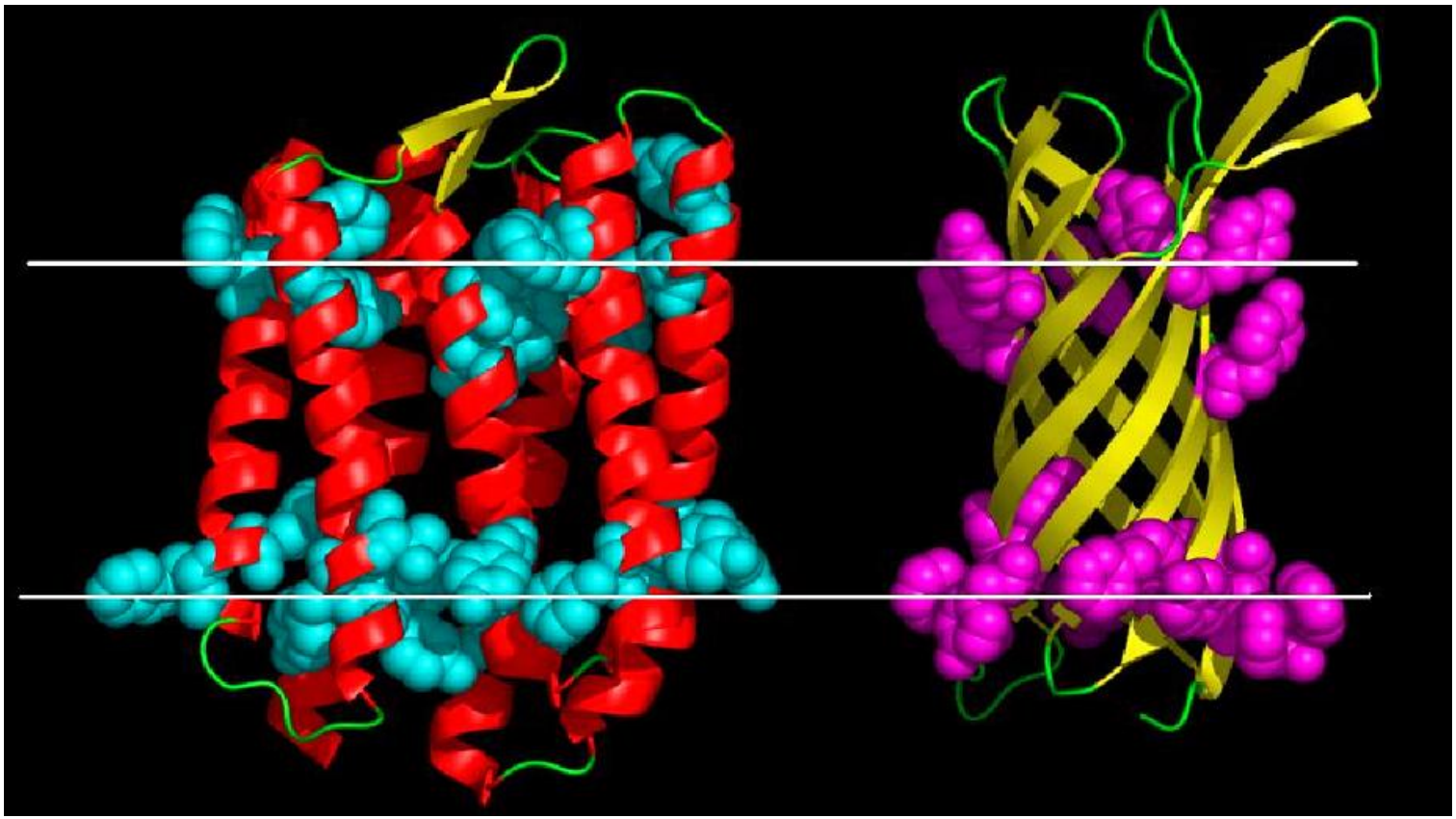
1. Υποδοχείς
2. Ένζυμα
3. Μεταφορείς ουσιών
4. Επικοινωνία
5. Συγκόλληση (κυτταρική)
6. Μετατροπείς ενέργειας

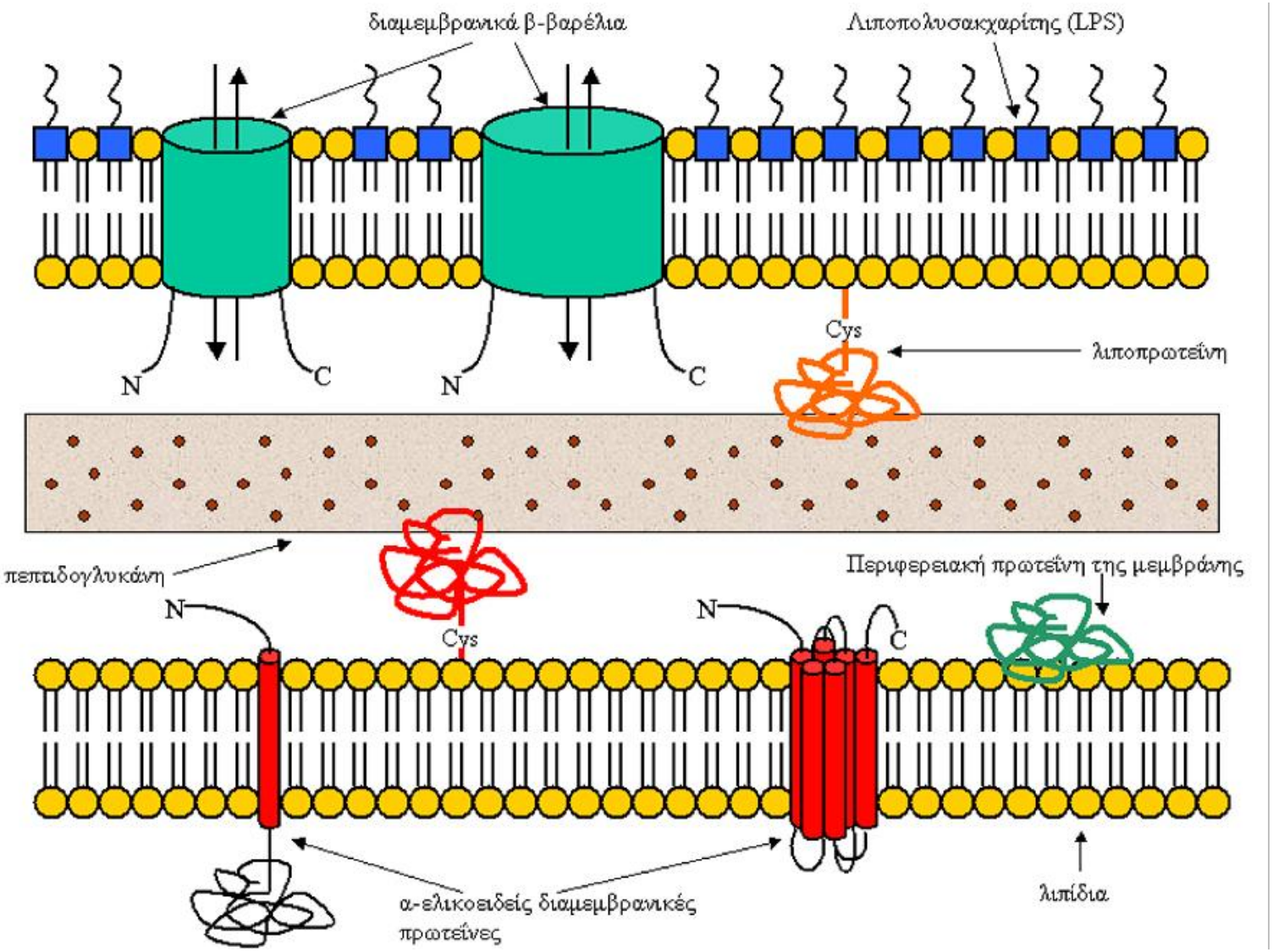


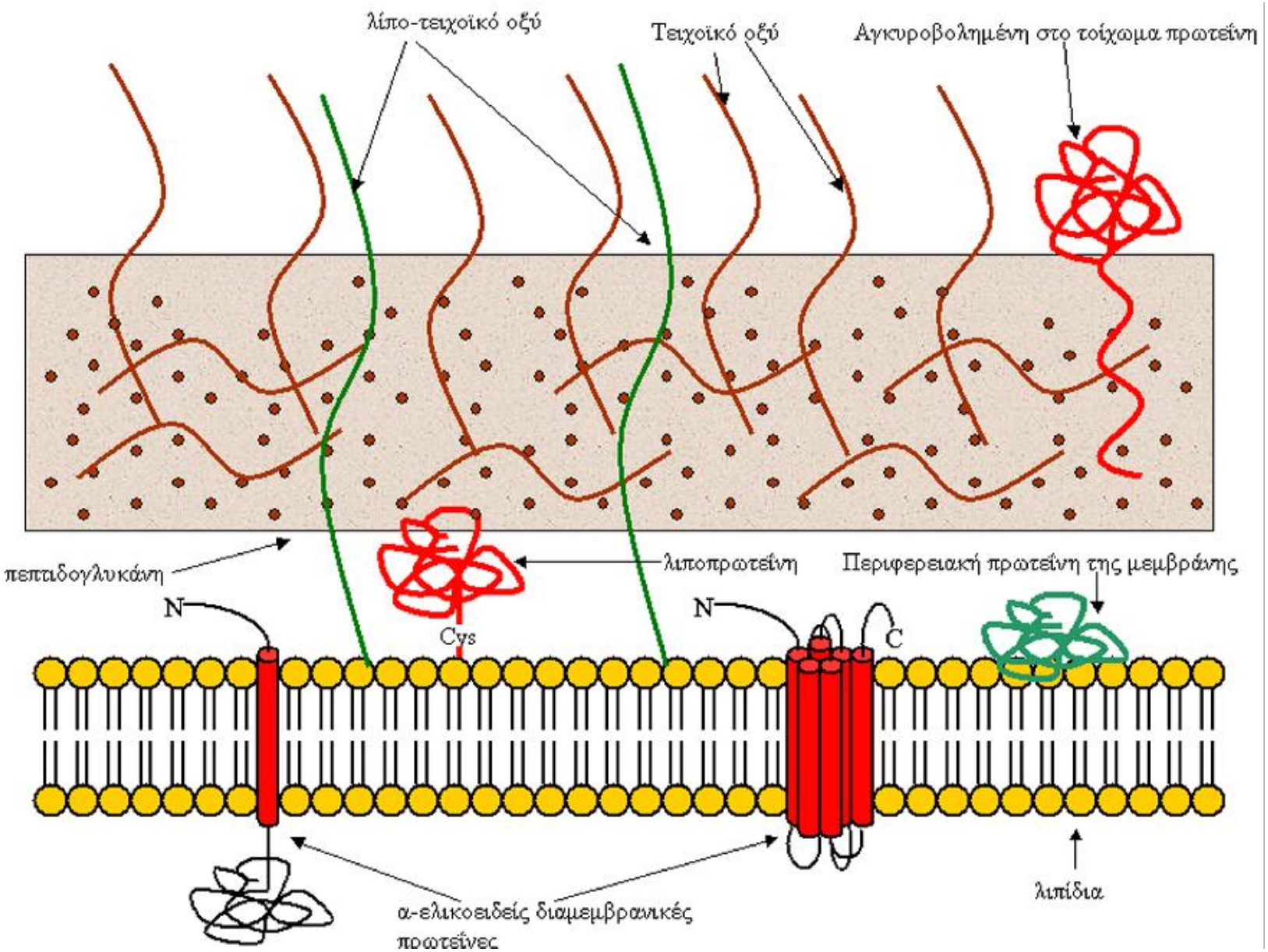
Γιατί είναι σημαντική η πρόγνωση;

- Οι μεμβρανικές πρωτεϊνες είναι πολλές (30%), με σημαντικές λειτουργίες
- Η δομή τους είναι δύσκολο να προσδιοριστεί
- Η πρόγνωση όμως είναι ευκολότερη

Διαμεμβρανικές πρωτεΐνες




διαμεμβρανικά β-βαρέλια
Λιποπολυσακχαρίτης (LPS)
N
C
N
C
Cys
λιποπρωτεΐνη
πεπτιδογλυκάνη
Περιφερειακή πρωτεΐνη της μεμβράνης
N
Cys
N
C
α-ελικοειδείς διαμεμβρανικές πρωτεΐνες
λιπίδια


λίπο-τειχοϊκό οξύ
Τειχοϊκό οξύ
Αγκυροβοληµένη στο τοίχωµα πρωτεΐνη
πεπτιδογλυκάνη
λιποπρωτεΐνη
Περιφερειακή πρωτεΐνη της µεµβράνης
N
Cys
N
C
α-ελικοειδείς διαµεµβρανικές πρωτεΐνες
λιπίδια

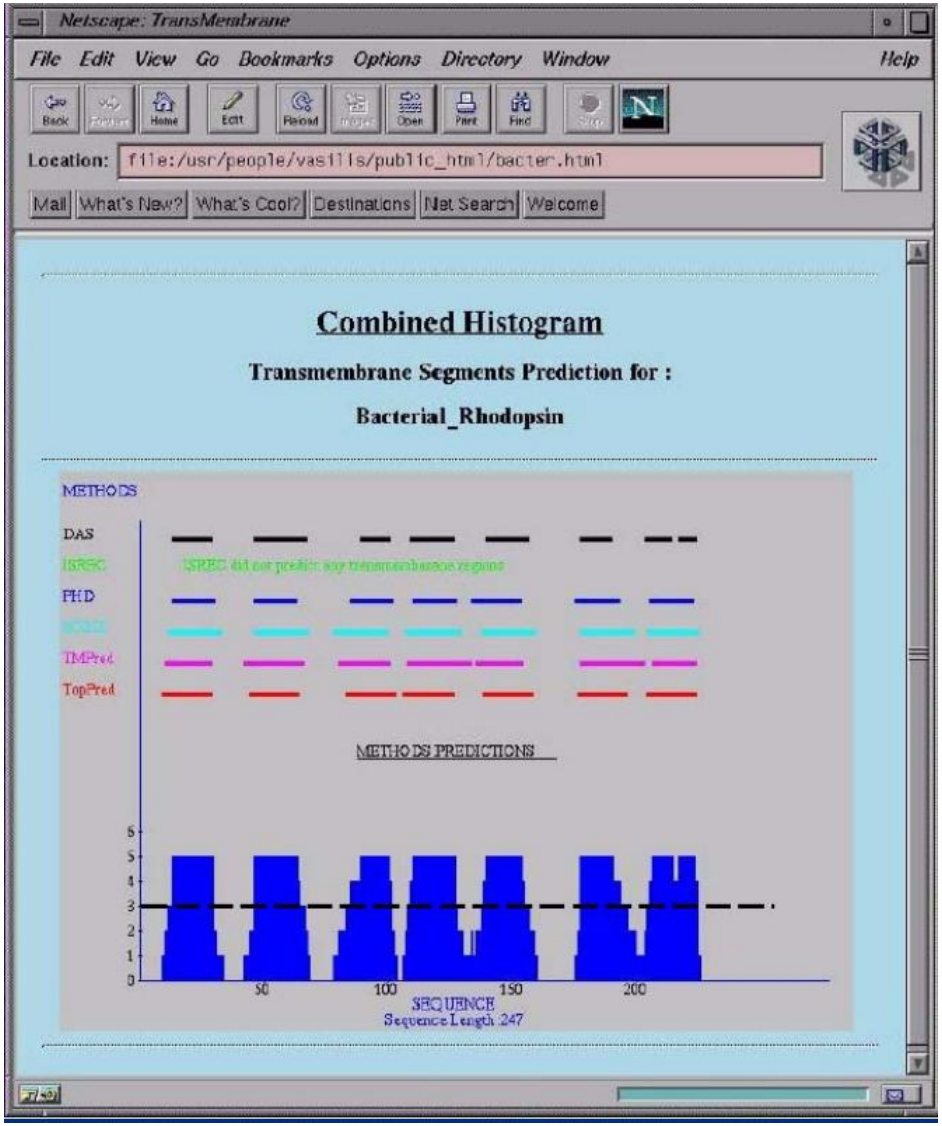
Οι πρώτοι αλγόριθμοι

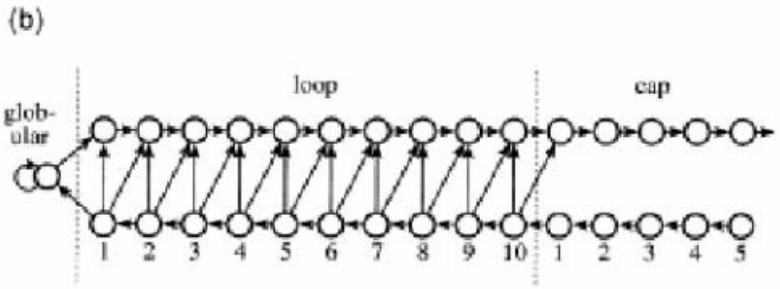
- Οι πρώτοι αλγόριθμοι πρόγνωσης της τοπολογίας των α-ελικοειδών μεμβρανικών πρωτεϊνών βασίστηκαν σε κινούμενα παράθυρα κατά μήκος της αμινοξικής αλληλουχίας. Αρχικά γινόταν χρήση παραθύρων σε συνδυασμό με κάποια κλίμακα υδροφοβικότητας αλλά και με τον κανόνα positive-inside.
- Έτσι, ένας από τους πρώτους αλγόριθμους πρόγνωσης ήταν το **TopPred** (Claros & von Heijne, 1994)(διαθέσιμο στη διεύθυνση http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::toppred).
- Το **TMpred** (http://www.ch.embnet.org/software/TMPRED_form.html) ήταν επίσης ένας από τους αρχικούς αλγόριθμους που βασιζόταν σε στατιστικές προτιμήσεις για την εμφάνιση των αμινοξέων.
- Την ίδια εποχή, εμφανίστηκε και το **MEMSAT** (το οποίο βέβαια έχει εξελιχθεί από τότε), που στηριζόταν σε ένα log-odds score βασισμένο σε στατιστικές προτιμήσεις αμινοξέων και βελτιστοποιούσε τα αποτελέσματα με χρήση δυναμικού προγραμματισμού (http://bioinf.cs.ucl.ac.uk/?id=756).
- Το **PRED-TMR** ήταν επίσης μια παρόμοια μέθοδος που αναπτύχθηκε λίγο αργότερα, από Έλληνες επιστήμονες (Pasquier et al., 1999)(διαθέσιμο στη διεύθυνση http://athina.biol.uoa.gr/PRED-TMR/) ενώ, καθώς προέβλεπε μόνο την παρουσία των διαμεμβρανικών ελίκων, έπρεπε να συνδυαστεί με έναν άλλον αλγόριθμο, το **orienTM** (http://athina.biol.uoa.gr/orienTM/), το οποίο βασιζόταν επίσης σε στατιστικές προτιμήσεις των αμινοξέων για να προβλέψει τη διευθέτηση των ήδη προβλεφθέντων διαμεμβρανικών περιοχών (Liakopoulos, Pasquier, & Hamodrakas, 2001).

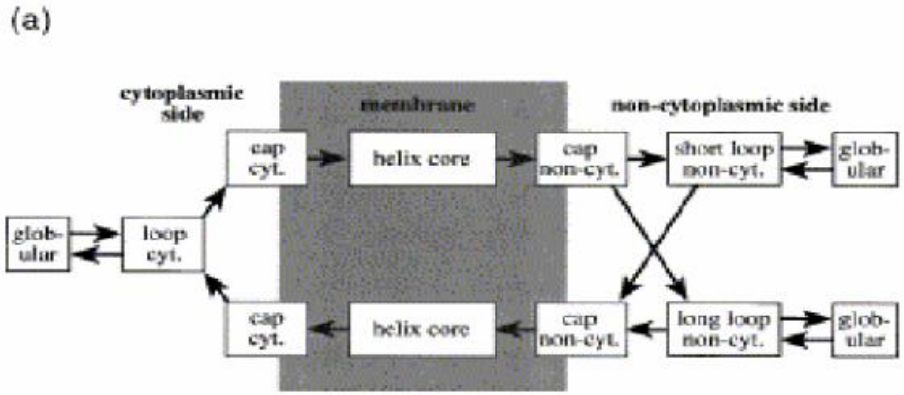
Νευρωνικά δίκτυα-HMM

- Η πρώτη προσπάθεια εφαρμογής Νευρωνικών Δικτύων, συνδυασμένη με πληροφορία από πολλαπλές στοιχίσεις, έγινε το 1996 με το **PHDtm** (www.predictprotein.org), ενώ από τότε έχουν εμφανιστεί πολλοί παρόμοιοι αλγόριθμοι.
- Ο πρώτος αλγόριθμος βασισμένος σε HMM εμφανίστηκε το 1998 (Sonnhammer, von Heijne, & Krogh, 1998), είναι το **TMHMM** (http://www.cbs.dtu.dk/services/TMHMM/) και θεωρείται ακόμα και σήμερα, ένας από τους καλύτερους αλγορίθμους της κατηγορίας (τουλάχιστον όσον αφορά τους αλγορίθμους που βασίζονται μόνο στην αμινοξική αλληλουχία). Παρόμοιος αλγόριθμος, αν και κάπως διαφορετικός στην υλοποίηση του μοντέλου είναι το **HMMTOP** (http://www.enzim.hu/hmmtop/) (Tusnady & Simon, 2001).
- Ένας από τους πρώτους αλγόοριθμους, που χρησιμοποίησαν συνδυαστική πρόγνωση, ήταν το **CoPreThi** (http://athina.biol.uoa.gr/CoPreTHi/) που αναπτύχθηκε στην Ελλάδα και βασιζόταν στους διαθέσιμους εκείνη την εποχή αλγόριθμους SOSUI, Tmpred, ISREC, DAS, TopPred, PHDtm και PRED-TMR (Promponas, Palaios, Pasquier, Hamodrakas, & Hamodrakas, 1999).

CoPreThi


Netscape: TransMembrane
File Edit View Go Bookmarks Options Directory Window Help
Location: file:/usr/people/vasilis/public_html/bacter.html
Mail What's New? What's Cool? Destinations Net Search Welcome
Combined Histogram
Transmembrane Segments Prediction for :
Bacterial_Rhodopsin
METHODS
DAS
PHD
TMPred
TopPred
METHODS PREDICTIONS
0
1
2
3
4
5
6
50
100
150
200
SEQUENCE
Sequence Length :247

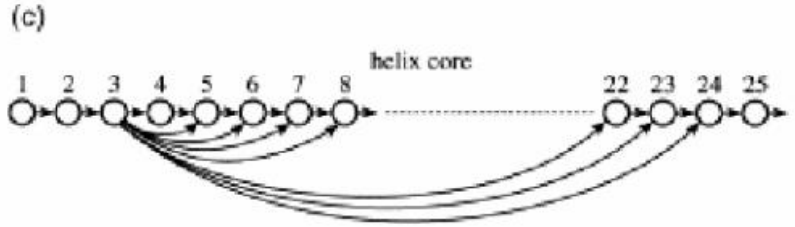




Structures of loop and transmembrane core

TM core 5-25 residues

TM helix from 15-35 residues



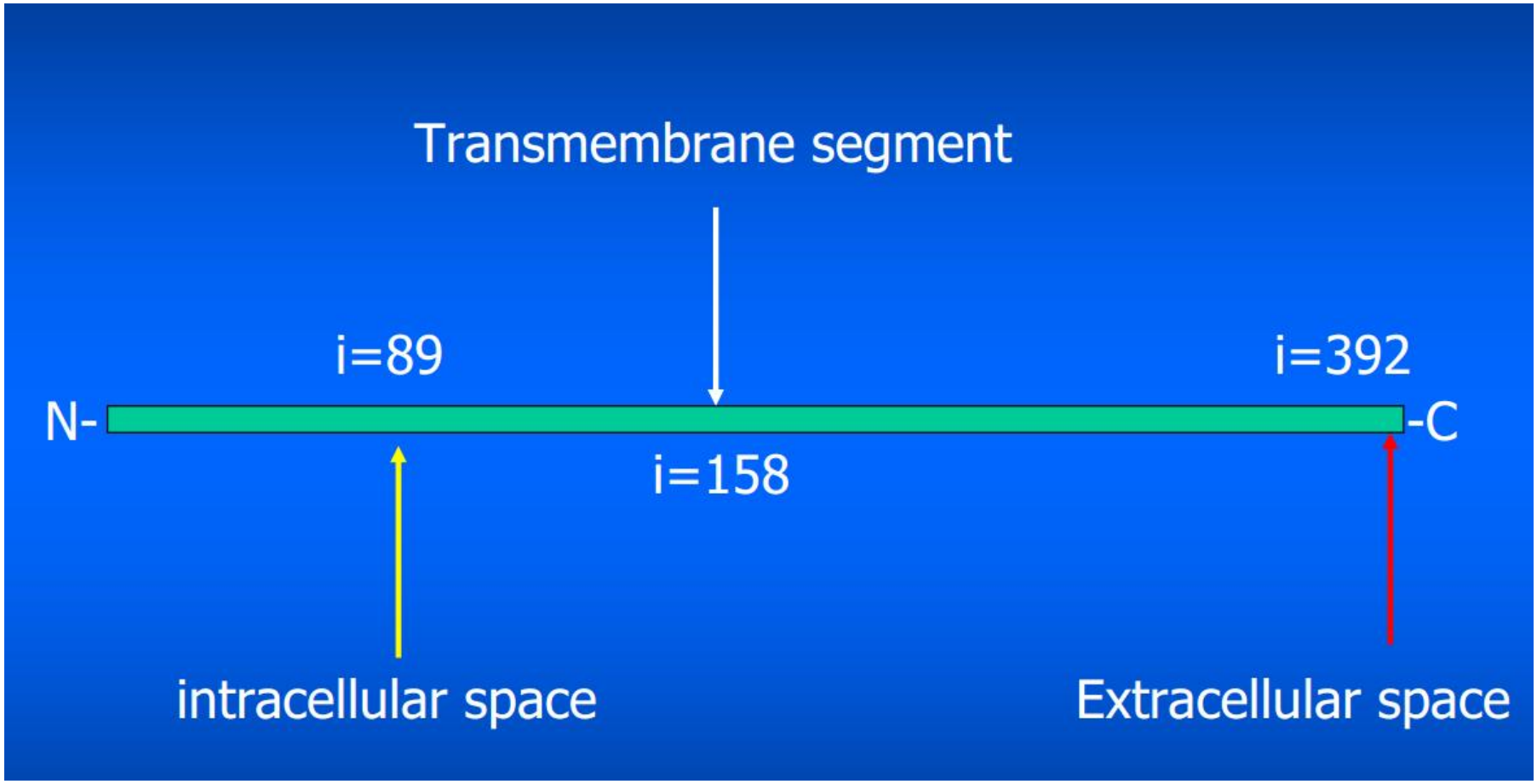
(Krogh et al, 2001, JMB)

Επεκτάσεις

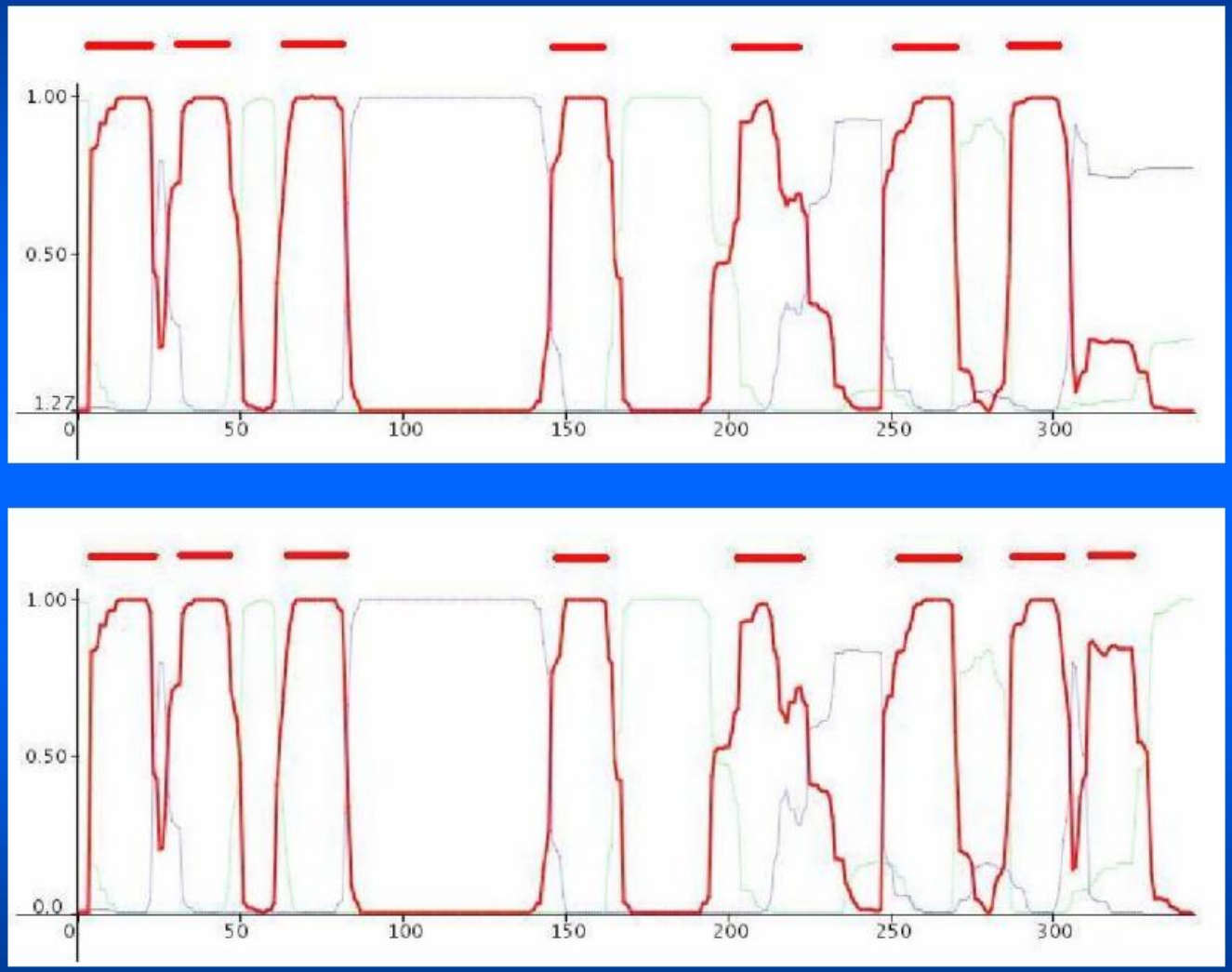
- Μια άλλη μεγάλη κατηγορία μεθόδων που έκαναν την εμφάνισή τους, ειδικά βασισμένοι σε χρήση των HMM, ήταν οι μέθοδοι που έκαναν ταυτόχρονη πρόγνωση των διαμεμβρανικών τμημάτων και των πεπτιδίων οδηγητών. Η βάση αυτής της μεθοδολογίας βρισκόταν στην παρατήρηση ότι τα αμινοτελικά πεπτίδια οδηγητές (βλ. επόμενη ενότητα) έχουν μια μεγάλη υδρόφοβη περιοχή που μοιάζει με διαμεμβρανική α-έλικα, και κατά συνέπεια πολλοί αλγόριθμοι πρόγνωσης των διαμεμβρανικών τμημάτων τα μπερδεύουν με διαμεμβρανικές περιοχές.
- Η πρώτη μέθοδος που έκανε αυτή την επέκταση ήταν το **Phobius** (Kall, Krogh, & Sonnhammer, 2004) (διαθέσιμο στη διεύθυνση http://phobius.sbc.su.se/), ενώ αργότερα εμφανίστηκε και το **SPOCTOPUS** (http://octopus.cbr.su.se/index.php?about=SPOCTOPUS).
- Μια άλλη παρόμοιας φύσεως επέκταση, έχει να κάνει με την ταυτόχρονη πρόγνωση τόσο των διαμεμβρανικών περιοχών όσο και των θέσεων μετα-μεταφραστικών τροποποιήσεων. Οι τροποποιήσεις αυτές, έχουν ειδική στόχευση στην αλληλουχία, αλλά συμβαίνουν και σε διακριτά τμήματα του κυττάρου. Έτσι, μια πρόγνωση για γλυκοζυλίωση μπορεί να βοηθήσει και την πρόγνωση των διαμεμβρανικών τμημάτων καθώς η γλυκοζυλίωση γίνεται σε περιοχές της πρωτεΐνης που βρίσκονται εκτεθειμένες στον εξωκυττάριο χώρο. Αντίθετα, οι θέσεις φωσφορυλίωσης βρίσκονται πάντα στην πλευρά που βρίσκεται στο κυτταρόπλασμα. Η μόνη μέθοδος που προσφέρει μέχρι στιγμής αυτή τη δυνατότητα, είναι το **HMMpTM** (http://bioinformatics.biol.uoa.gr/HMMpTM), το οποίο με αυτόν τον τρόπο πετυχαίνει βελτιωμένη πρόγνωση τόσο στην περίπτωση της διαμεμβρανικής τοπολογίας, όσο και στην περίπτωση των θέσεων γλυκοζυλίωσης και φωσφορυλίωσης (Tsaousis, Bagos, & Hamodrakas, 2014).

Ενσωμάτωση πειραματικής πληροφορίας

- Από τις ήδη διαθέσιμες προγνωστικές μεθόδους, το TMHMM και το HMMTOP (Tusnady & Simon, 2001), προσφέρουν στο χρήστη την επιλογή να ενσωματώσει στην πρόγνωσή του, πειραματικά προσδιορισμένη πληροφορία για την τοπολογία. Παρόμοια επιλογή, προσφέρεται και από την συνδυασμένη πρόγνωση διαμεμβρανικών α-ελίκων και πεπτιδίων οδηγητών, με τη μέθοδο **Phobius** (Kall et al., 2004).
- Το **HMM-TM** το οποίο αναπτύχθηκε από την ομάδα μας (http://bioinformatics.biol.uoa.gr/HMM-TM/), ήταν η πρώτη μέθοδος που ενσωμάτωνε τέτοιου είδους πληροφορία σε κάθε αλγόριθμο αποκωδικοποίησης των HMM, ενώ παράλληλα έδινε και τη θεωρητική τεκμηρίωση για αυτήν την τροποποίηση.

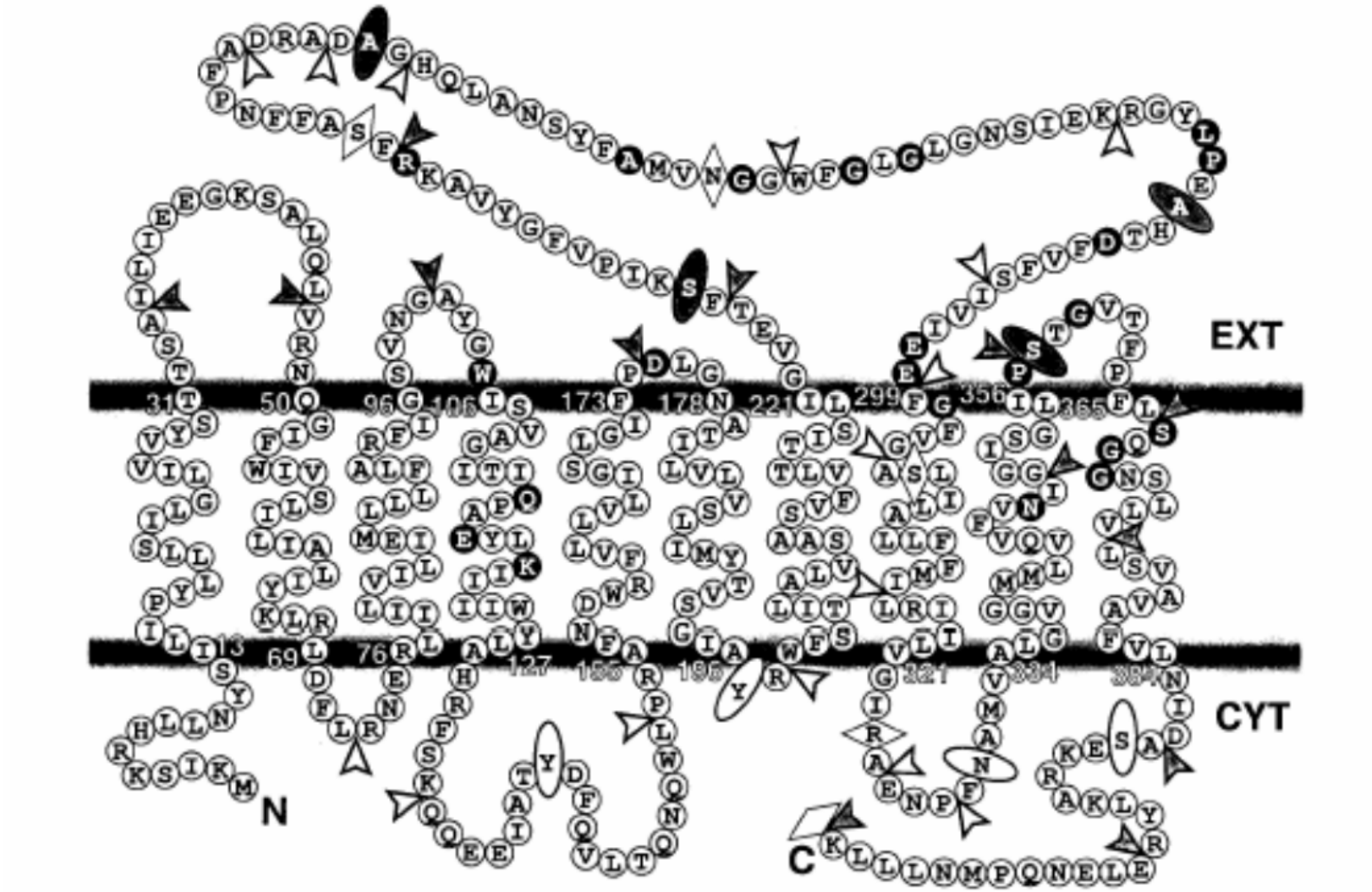

Transmembrane segment
i=89
i=392
N-
-C
i=158
intracellular space
Extracellular space

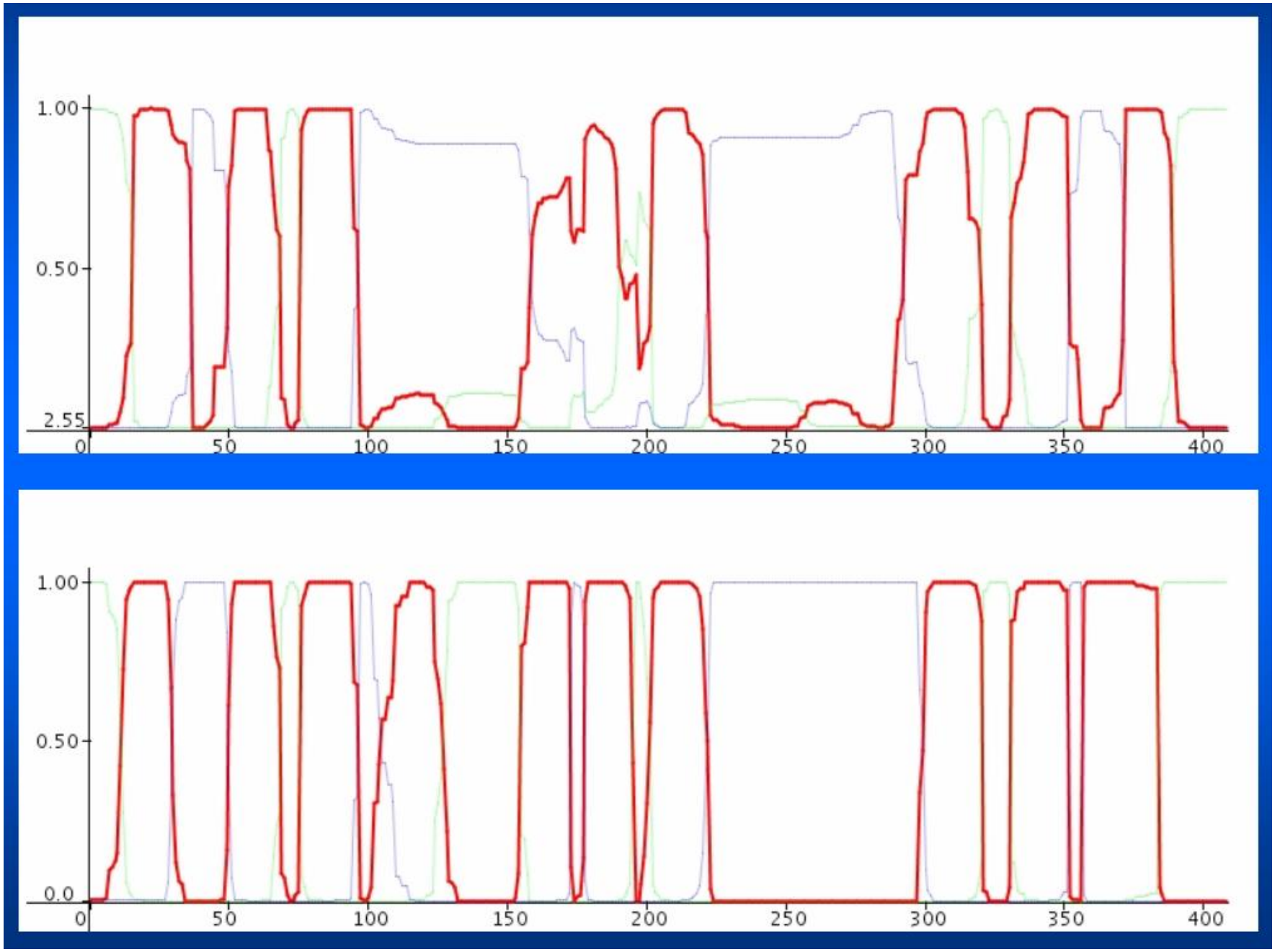
A sample prediction (YDGG_ECOLI)



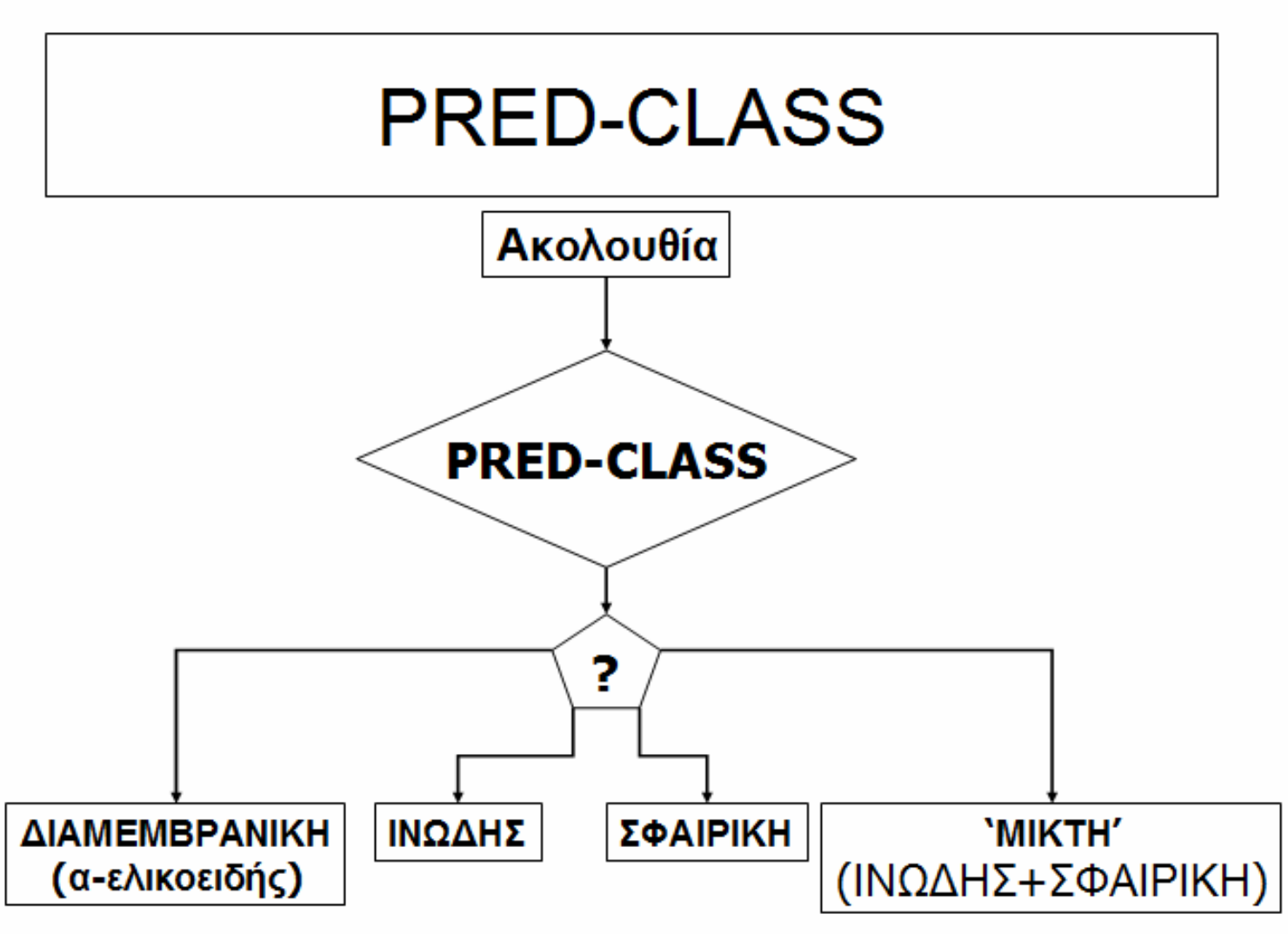
http://bioinformatics.biol.uoa.gr/HMM-TM/

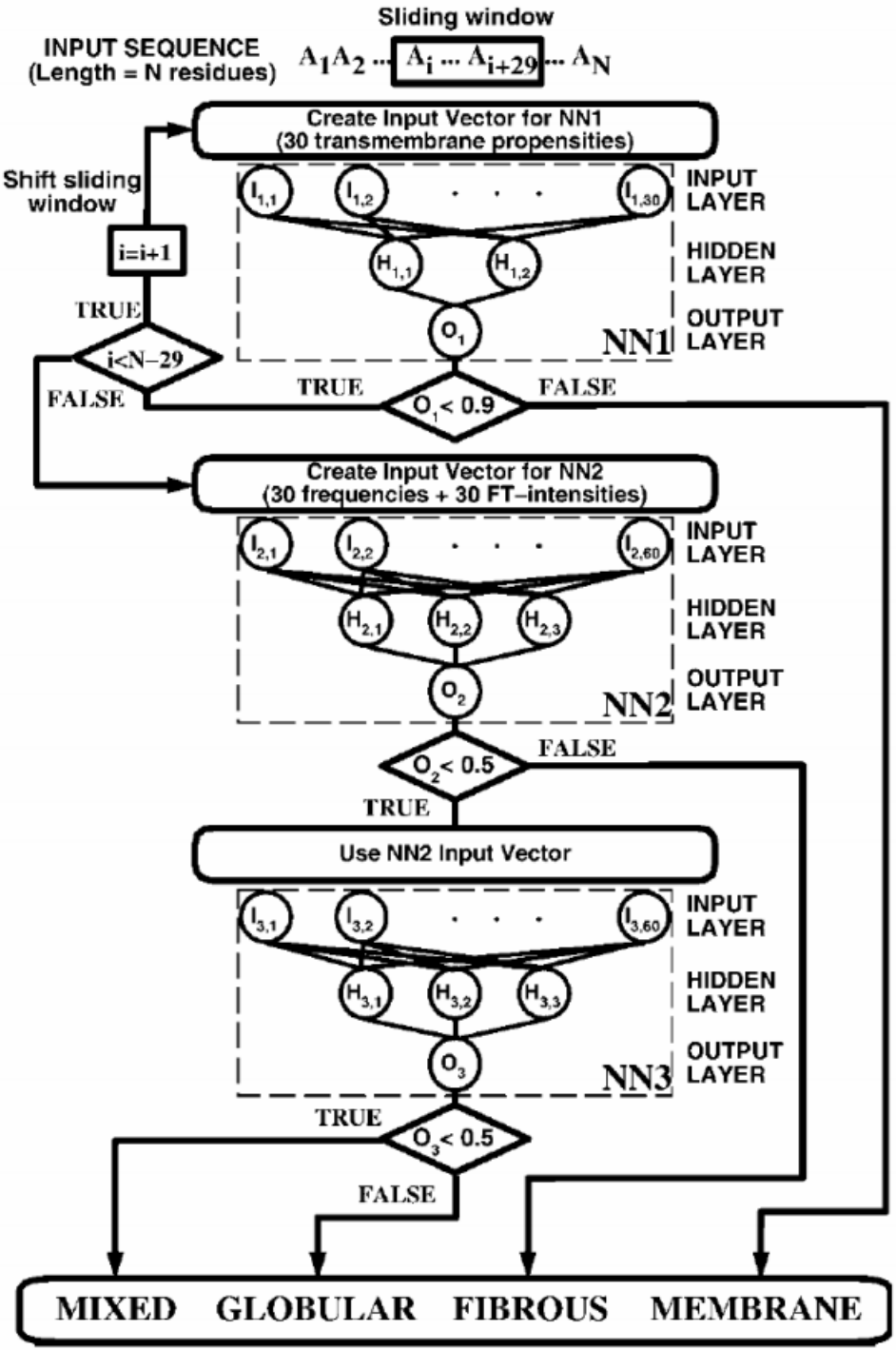
FtsW *S. pneumoniae* (Q8DPW6_STRR6)

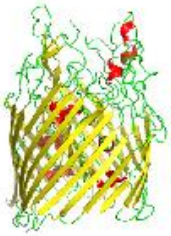



1.00
0.50
2.55
0
50
100
150
200
250
300
350
400
1.00
0.50
0.0
0
50
100
150
200
250
300
350
400

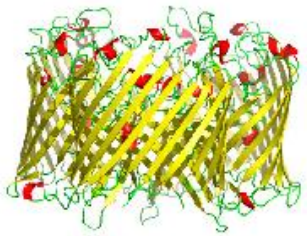
PRED-CLASS


PRED-CLASS
Ακολουθία
PRED-CLASS
?
ΔΙΑΜΕΜΒΡΑΝΙΚΗ
(α-ελικοειδής)
ΙΝΩΔΗΣ
ΣΦΑΙΡΙΚΗ
'ΜΙΚΤΗ'
(ΙΝΩΔΗΣ+ΣΦΑΙΡΙΚΗ)


Sliding window
INPUT SEQUENCE
(Length = N residues)
A1A2 ... Ai ... Ai+29 ... AN
Create Input Vector for NN1
(30 transmembrane propensities)
Shift sliding window
i=i+1
TRUE
i<N−29
FALSE
I1,1
I1,2
I1,30
H1,1
H1,2
O1
INPUT LAYER
HIDDEN LAYER
OUTPUT LAYER
NN1
TRUE
O1< 0.9
FALSE
Create Input Vector for NN2
(30 frequencies + 30 FT−intensities)
I2,1
I2,2
I2,60
H2,1
H2,2
H2,3
O2
NN2
O2< 0.5
FALSE
TRUE
Use NN2 Input Vector
I3,1
I3,2
I3,60
H3,1
H3,2
H3,3
O3
NN3
TRUE
O3< 0.5
FALSE
MIXED
GLOBULAR
FIBROUS
MEMBRANE



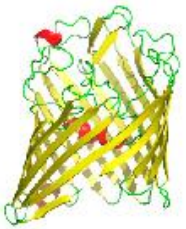
FepA (1FEP)



Sucroporin (1A0S)

Omp32 (1E54)

Porin (2POR)

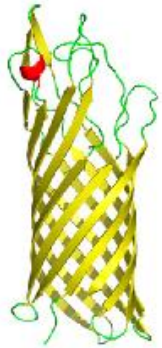


OmpF (2OMF)

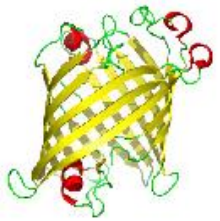
OmpX (1QJ8)

OmpA (1QJP)

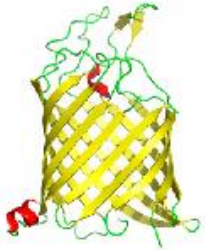
OmpT (1I78)



OpcA (1K24)



OmplA (1QD5)

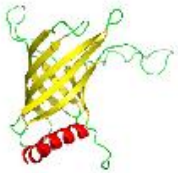


Tsx (1TLY)

FADL (1T1L)

NspA (1P4T)

NalP (1UYN)



PagP (1MM4)

χαρακτηριστικά

- Οι διαμεμβρανικοί β-κλώνοι είναι κατά βάση αμφιπαθικοί, καθώς εμφανίζουν εναλλαγή υδρόφοβων-πολικών καταλοίπων. Τα υδρόφοβα κατάλοιπα αλληλεπιδρούν με τις υδρόφοβες ουρές των λιπιδίων της μεμβράνης, ενώ τα πολικά στρέφονται προς το εσωτερικό του βαρελιού και άρα αλληλεπιδρούν με το υδάτινο περιβάλλον του πόρου.
- Τα αρωματικά κατάλοιπα έχουν την τάση να εμφανίζονται με μεγαλύτερη συχνότητα στις επιφάνειες επαφής με τις πολικές κεφαλές των λιπιδίων, σχηματίζοντας έτσι τις λεγόμενες «αρωματικές ζώνες» στην περιφέρεια του βαρελιού.
- Και το αμινοτελικό και το καρβοξυτελικό άκρο των πρωτεϊνών αυτών, είναι τοποθετημένα στον περιπλασμικό χώρο (εσωτερικά σε σχέση με την εξωτερική μεμβράνη). Σε κάποιες περιπτώσεις, μεγάλες αμινοτελικές και καρβοξυτελικές δομικές περιοχές, με μήκος μεγαλύτερο των 100 καταλοίπων, είναι δυνατόν να σχηματίζονται.
- Τα τμήματα της ακολουθίας, τα οποία συνδέουν τους διαμεμβρανικούς κλώνους, και τα οποία βρίσκονται στον περιπλασμικό χώρο (εσωτερικές στροφές) είναι γενικά μικρότερου μήκους από τα τμήματα τα οποία βρίσκονται στον εξωκυττάριο χώρο (εξωτερικές θηλιές). Οι στροφές του περιπλασμικού χώρου, σε όλες σχεδόν τις γνωστές δομές, έχουν μήκος 12 ή και λιγότερα κατάλοιπα ενώ αυτές του εξωκυττάριου χώρου μπορεί να έχουν μήκος και πάνω από 30 κατάλοιπα. Αυτό είναι επιτρεπτό λόγω της διαμόρφωσης του μαιάνδρου που υιοθετείται από το β-βαρέλι.
- Το μήκος των διαμεμβρανικών β-κλώνων ποικίλει ανάλογα με την κλίση του κλώνου σε σχέση με τον άξονα του βαρελιού και παίρνει τιμές από 6 έως και 22 κατάλοιπα. Παρ' όλα αυτά, σε αρκετές περιπτώσεις, μόνο ένα μικρό τμήμα του κλώνου είναι βυθισμένο στην λιπιδική διπλοστιβάδα, και το υπόλοιπο προεξέχει μακριά από το επίπεδο της μεμβράνης προς τον εξωκυττάριο χώρο, σχηματίζοντας εύκαμπτες φουρκέτες.
- Οι διαμεμβρανικές πρωτεΐνες με μορφή β-βαρελιού εμφανίζουν μικρότερη συντηρητικότητα στις ακολουθίες τους, σε σχέση με τις σφαιρικές-υδατοδιαλυτές πρωτεΐνες. Ακόμα μικρότερη είναι η συντηρητικότητα στις εξωκυττάριες στροφές, οι οποίες δρουν συχνά σαν αντιγονικοί καθοριστές. Το γεγονός αυτό συνεπάγεται, ότι πρωτεΐνες με πολύ μικρή ομοιότητα σε επίπεδο ακολουθίας είναι δυνατόν να διπλώνονται με απολύτως όμοιο τρόπο, αλλά παρ' όλα αυτά οι μέθοδοι αναζήτησης με βάση την ομοιότητα στην ακολουθία να μην μπορούν να τις ανιχνεύσουν.
- Οι γειτονικοί β-κλώνοι συνδέονται με ένα δίκτυο δεσμών υδρογόνου, το οποίο σταθεροποιεί τη δομή του βαρελιού.

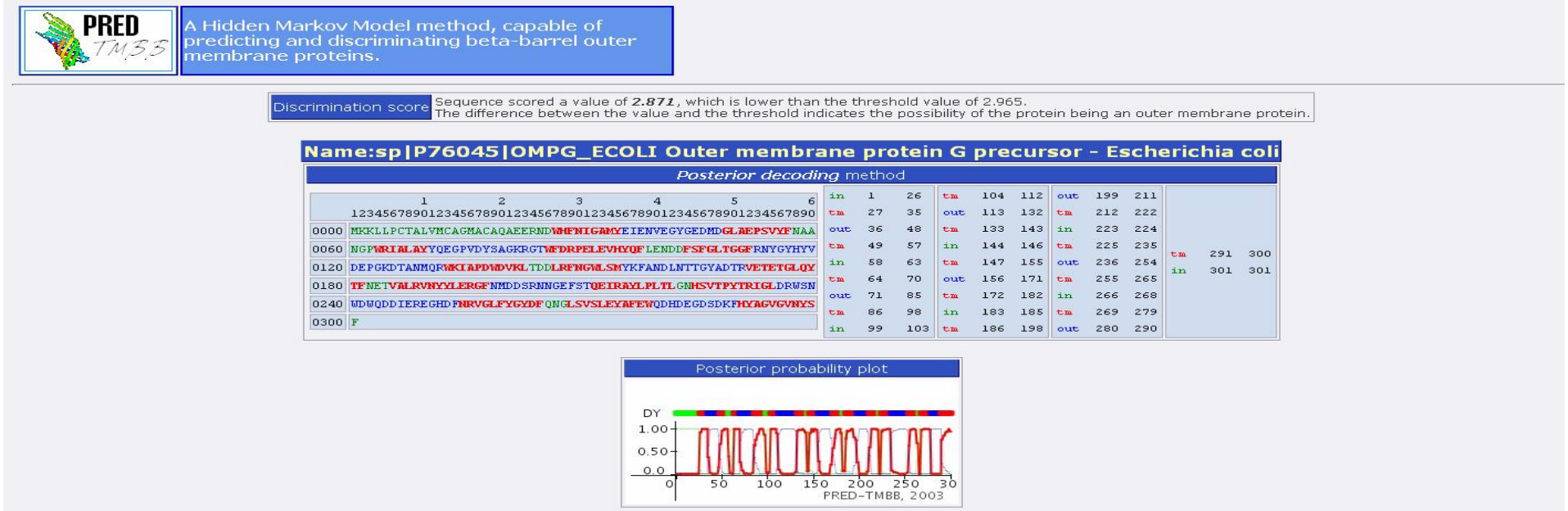
Νευρωνικά δίκτυα

- Οι μέθοδοι πρόγνωσης των διαμεμβρανικών β-βαρελιών, επίσης, διακρίνονται σε μεθόδους που βασίζονται στην υδροφοβικότητα, σε στατιστικές τεχνικές και σε μεθόδους μηχανικής μάθησης.
- Αξίζει να σημειωθεί, ότι είναι άλλο το πρόβλημα της πρόγνωσης της διαμεμβρανικής τοπολογίας των β-βαρειλιών και άλλο το πρόβλημα του εντοπισμού τους. Κατά συνέπεια, έχουν αναπτυχθεί και διαφορετικές μεθοδολογίες για τις παραπάνω περιπτώσεις, αν και κάποιοι από τους αλγόριθμους αυτούς επιτυγχάνουν και τις δύο λειτουργίες.
- Η πρώτη προσπάθεια εφαρμογής μεθόδων μηχανικής μάθησης για την πρόγνωση της τοπολογίας των διαμεμβρανικών β-βαρελιών, πραγματοποιήθηκε από τον Diederichs και του συνεργάτες του (Diederichs, Freigang, Umhau, Zeth, & Breed, 1998) αλλά πλέον η μέθοδος αυτή δεν είναι διαθέσιμη.
- Το **B2TMPRED** που αναπτύχθηκε λίγο αργότερα χρησιμοποίησε Νευρωνικά Δίκτυα με ταυτόχρονη χρήση εξελικτικής πληροφορίας αλλά και επιπλέον φιλτράρισμα των αποτελεσμάτων με αλγόριθμο δυναμικού προγραμματισμού (Jacoboni, Martelli, Fariselli, De Pinto, & Casadio, 2001) και είναι διαθέσιμο στη διεύθυνση http://gpcr.biocomp.unibo.it/cgi/predictors/outer/pred_outercgi.cgi.
- Σε Νευρωνικά Δίκτυα βασίζονται επίσης και το **TBBpred** (http://www.imtech.res.in/raghava/tbbpred/) και το
- **TMBETA-NET** (http://psfs.cbrc.jp/tmbeta-net/) τα οποία χρησιμοποιούν μόνο την αμινοξική αλληλουχία, αλλά και το **TMBETAPRED-RBF** (http://rbf.bioinfo.tw/~sachen/BARRELpredict/TMBETAPRED-RBF.php) και το
- **TMBpro** (http://tmbpro.ics.uci.edu/) τα οποία χρησιμοποιούν εξελικτική πληροφορία με τη μορφή πολλαπλών στοιχίσεων.

HMM

- Οι πρώτες μέθοδοι βασισμένες σε Hidden Markov Model (HMM) εμφανίστηκαν επίσης στις αρχές της δεκαετίας του 2000, και από τότε η μεθοδολογία αυτή έχει κυριαρχήσει (Bagos, Liakopoulos, Spyropoulos, & Hamodrakas, 2004a, 2004b; Bigelow, Petrey, Liu, Przybylski, & Rost, 2004; Hayat & Elofsson, 2012; Liu, Zhu, Wang, & Li, 2003; Martelli, Fariselli, Krogh, & Casadio, 2002; Savojardo, Fariselli, & Casadio, 2013; Singh, Goodman, Walter, Helms, & Hayat, 2011).
- Η πρώτη μέθοδος ήταν το **HMM-B2TMR**, το οποίο χρησιμοποιούσε πολλαπλές στοιχίσεις αλλά έγινε δημόσια διαθέσιμο αργότερα (http://gpcr.biocomp.unibo.it/predictors/),
- ενώ πλέον έχει εμφανιστεί και μια συνδυαστική μέθοδος από την ίδια ομάδα, το **BetAware** (http://www.biocomp.unibo.it/~savojard/betawarecl).
- Το **PRED-TMBB** (http://bioinformatics.biol.uoa.gr/PRED-TMBB/) παρουσιάστηκε λίγο αργότερα από εμάς, και ήταν ιδιαίτερα πετυχημένο, καθώς παρ' όλο που χρησιμοποιούσε μόνο πληροφορία από την αμινοξική αλληλουχία, χρησιμοποίησε ένα διαφορετικό κριτήριο για την εκτίμηση των παραμέτρων του μοντέλου, αλλά και διαφορετικούς αλγόριθμους για την εκπαίδευση και την αποκωδικοποίησή του.
- Ταυτόχρονα είχε εμφανιστεί το **PROFtmb** (https://www.predictprotein.org/) το οποίο έκανε χρήση εξελικτικής πληροφορίας ενώ αργότερα εμφανίστηκαν και άλλες μέθοδοι, όπως το **TMBHMM** και το **TMBhunt**.
- Η τελευταία και πιο αξιόπιστη μέθοδος, είναι το **BOCTOPUS** (http://boctopus.cbr.su.se/), το οποίο χρησιμοποιεί ένα συνδυασμό Support Vector Machines και HMMs ενώ κάνει και χρήση εξελικτικής πληροφορίας.

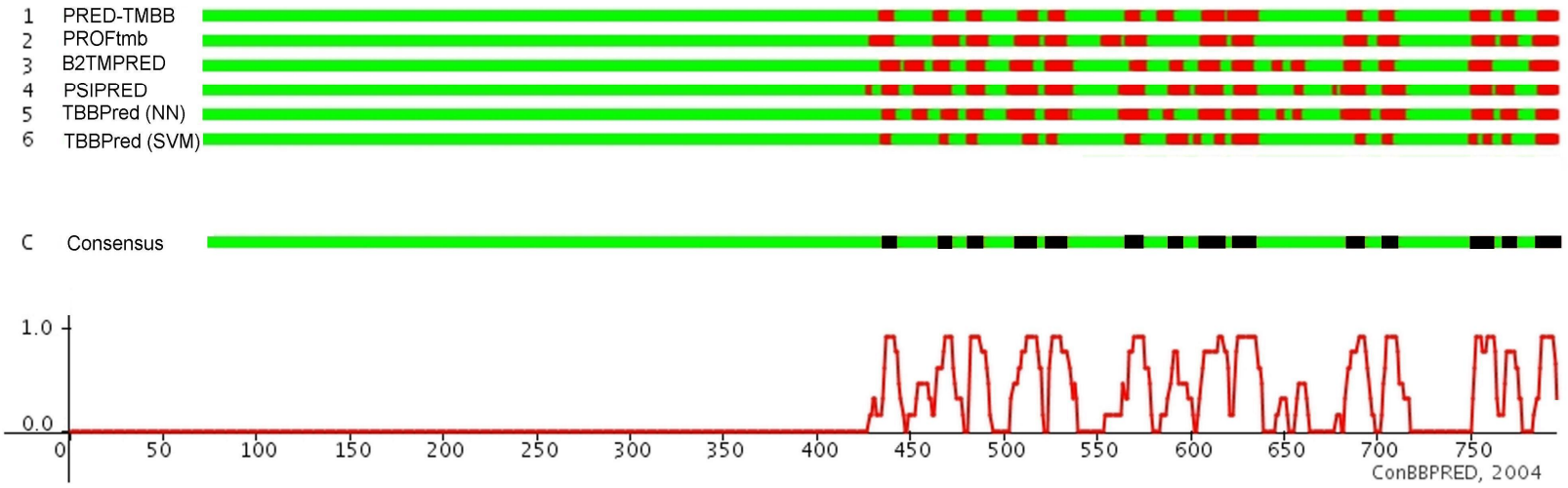
PRED-TMBB

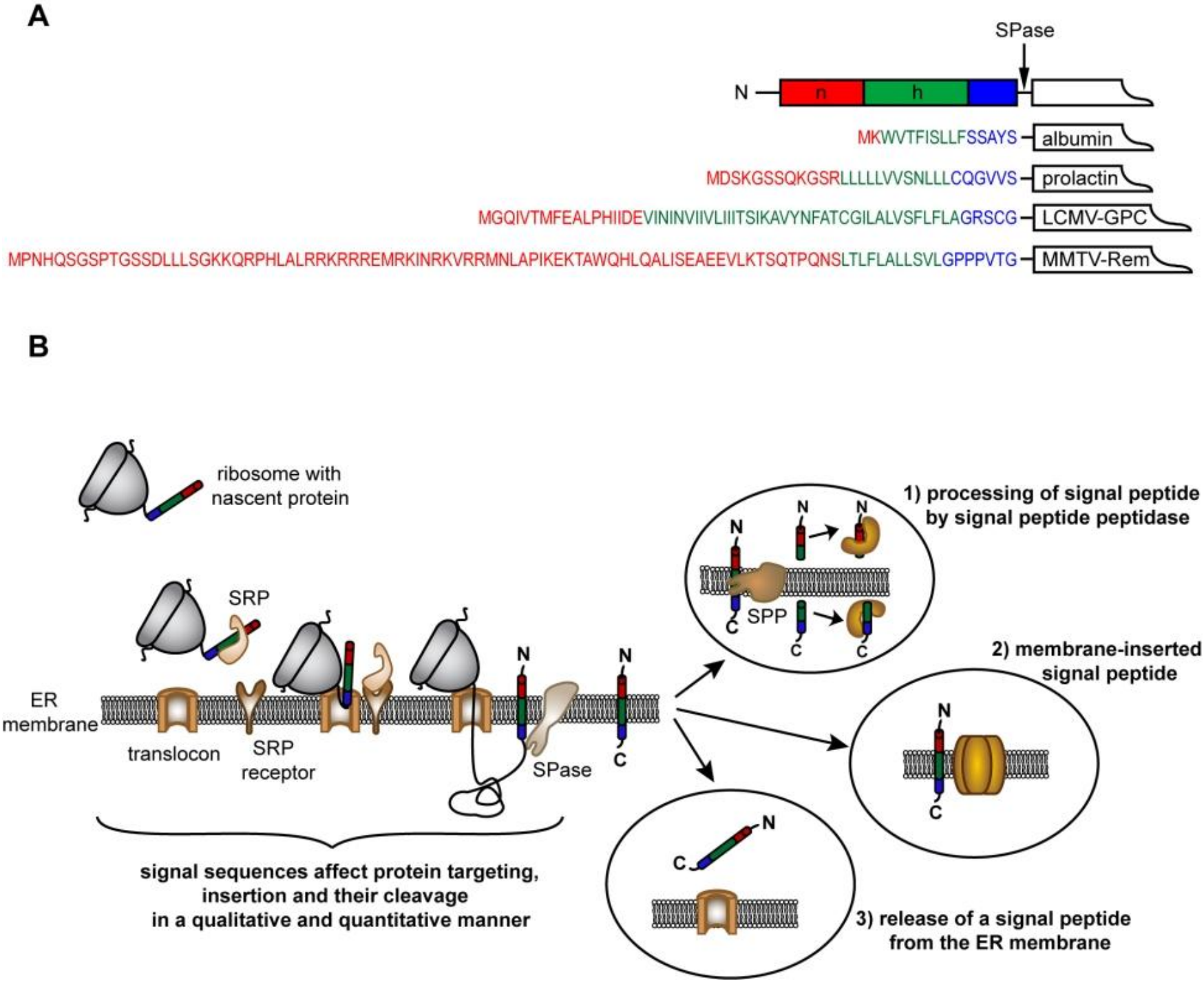

PRED TM55
A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins.
Discrimination score
Sequence scored a value of 2.871, which is lower than the threshold value of 2.965.
The difference between the value and the threshold indicates the possibility of the protein being an outer membrane protein.
Name:sp|P76045|OMPG_ECOLI Outer membrane protein G precursor - Escherichia coli
Posterior decoding method
1 2 3 4 5 6
1234567890123456789012345678901234567890123456789012345678901234567890
0000 MKKLLPCTALVMCAGMACAQAEERNDWHFNIGAMYEIENVEGYGEDMDGLAEPSVYFNAA
0060 NGPWRIALAYYQEGPVDYSAGKRGTWFDRPELEVHYQFLENDDFSFGLTGGFRNYGYHYV
0120 DEPGKDTANMQRWKIAPDWDVKLTDDLRFNGWLSMYKFANDLNTTGYADTRVETETGLQY
0180 TFNETVALRVNYYLERGFNMDDSRNNGEFSTQEIRAYLPLTLGNHSVTPYTRIGLDRWSN
0240 WDWQDDIEREGHDFNRVGLFYGYDFQNGLSVSLEYAFEWQDHDEGDSDKFHYAGVGVNYS
0300 F
in 1 26
tm 27 35
out 36 48
tm 49 57
in 58 63
tm 64 70
out 71 85
tm 86 98
in 99 103
tm 104 112
out 113 132
tm 133 143
in 144 146
tm 147 155
out 156 171
tm 172 182
in 183 185
tm 186 198
out 199 211
tm 212 222
in 223 224
tm 225 235
out 236 254
tm 255 265
in 266 268
tm 269 279
out 280 290
tm 291 300
in 301 301
Posterior probability plot
DY
1.00
0.50
0.0
0 50 100 150 200 250 30
PRED-TMBB, 2003

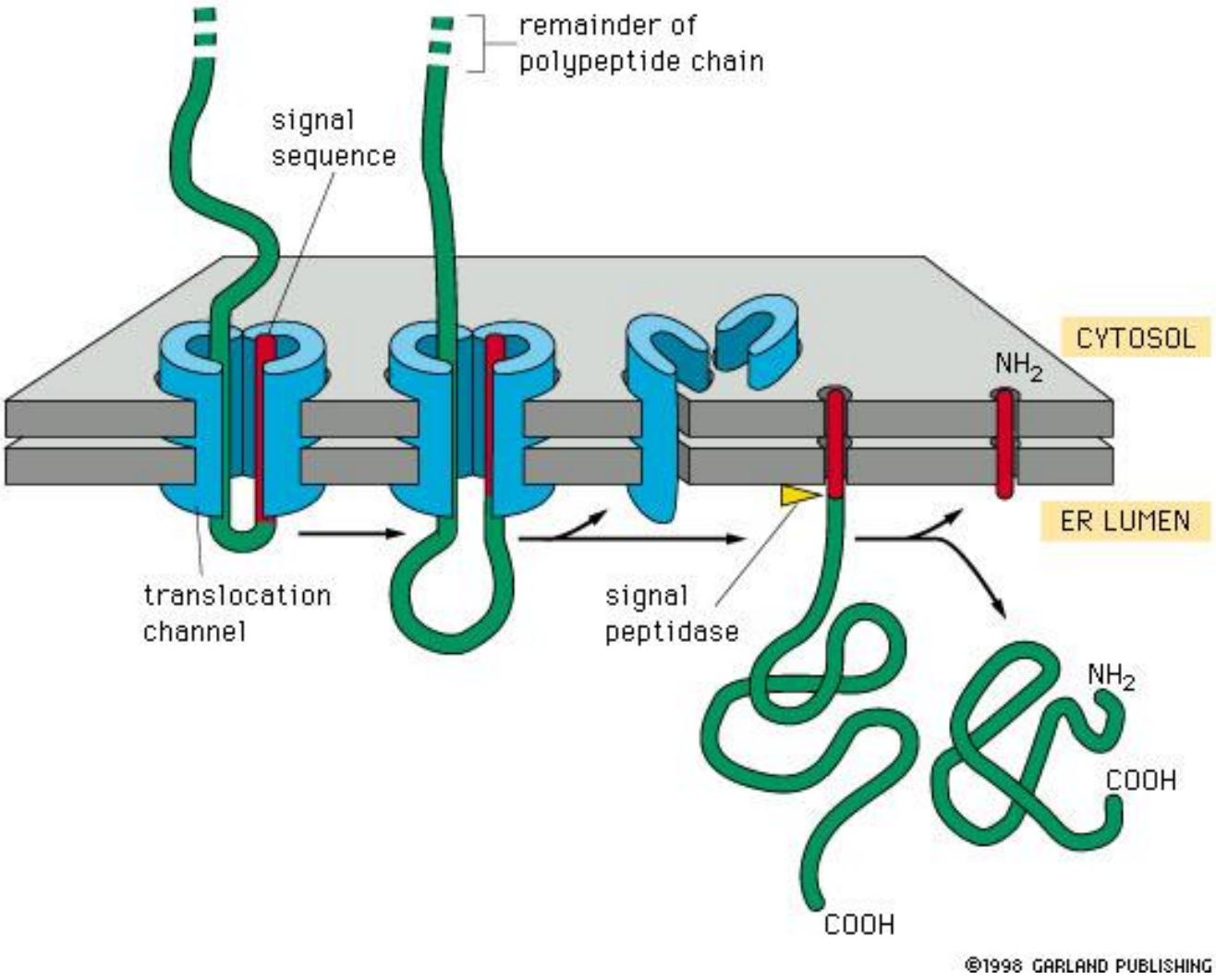
Συνδυαστικές μέθοδοι

- Όμοια με τις α-ελικοειδείς μεμβρανικές πρωτεΐνες, εκτεταμένες εμπειρικές αναλύσεις έχουν δείξει ότι οι μέθοδοι που βασίζονται σε κάποια γραμματική δομή όπως τα HMM, είναι κατά κανόνα καλύτερες για την πρόγνωση των διαμεμβρανικών β-βαρελιών σε σχέση με τις πιο απλές στατιστικές μεθόδους, αλλά και σε σχέση με τα Νευρωνικά Δίκτυα.
- Επίσης, τόσο ο συνδυασμός πολλών μεθόδων όσο και η χρήση εξελικτικής πληροφορίας με τη μορφή πολλαπλών στοιχίσεων είναι παράγοντες που αυξάνουν σημαντικά την απόδοση των μεθόδων αυτών.
- Με βάση τα παραπάνω, το 2005, παρουσιάσαμε τον μοναδικό μέχρι στιγμής συνδυαστικό αλγόριθμο πρόγνωσης των β-βαρελιών, το **ConBBPRED** (http://bioinformatics.biol.uoa.gr/ConBBPRED/). Το ConBBPRED δίνει τη δυνατότητα στο χρήστη να επιλέξει ποιες μεθόδους θα συμπεριλάβει στη συνδυαστική πρόγνωση ενώ επιπλέον βελτιστοποιεί την τελική πρόγνωση με έναν αλγόριθμο δυναμικού προγραμματισμού. Με τον τρόπο αυτό, η μέθοδος ξεπερνάει σε επιτυχία όλες τις επιμέρους μεθόδους που χρησιμοποιούνται στην πρόγνωση. Το μειονέκτημα της μεθόδου είναι το γεγονός ότι ο χρήστης πρέπει να έχει λάβει μόνος του τα αποτελέσματα από τις επιμέρους μεθόδους και να τα επικολλήσει στην αντίστοιχη φόρμα της διαδικτυακής εφαρμογής (Bagos, Liakopoulos, & Hamodrakas, 2005).
- Παρ' όλο που είδαμε ότι ακόμα και για τα β-βαρέλια η αύξηση του μεγέθους του συνόλου εκπαίδευσης δεν οδηγεί σε γραμμική αύξηση της απόδοσης, το μέγεθος παίζει κάποιο ρόλο, ειδικά αν αναλογιστούμε ότι οι πρώτες μέθοδοι ήταν εκπαιδευμένες σε μόλις 10-20 τέτοιες πρωτεΐνες.
- Έτσι, είναι κατανοητό ότι οι πιο σύγχρονες μέθοδοι όπως το BOCTOPUS, που έχουν εκπαιδευθεί σε μερικές δεκάδες αλληλουχίες, θα είναι πιο αποδοτικές. Παρ' όλα αυτά, οι αλγοριθμικές επιλογές αλλά και ο σωστός σχεδιασμός του μοντέλου καθιστούν ακόμα και σήμερα το PRED-TMBB μια ιδιαίτερα ανταγωνιστική μέθοδο. Εκτός από το PRED-TMBB και το BOCTOPUS, οι πιο αξιόπιστες μέθοδοι σύμφωνα με τα τελευταία δεδομένα είναι το PROFtmb, το BetAware και το HMM-B2TMR. Μια προσπάθεια να επανεκπαιδευθεί το PRED-TMBB σε νέα δεδομένα αλλά και να χρησιμοποιήσει εξελικτική πληροφορία, έχει δώσει εξαιρετικά μέχρι στιγμής αποτελέσματα και αναμένουμε να δημοσιευτεί σύντομα. Η μέθοδος αυτή, το **PRED-TMBB2** (www.compgen.org/tools/PRED-TMBB2), φαίνεται ότι είναι πλέον η πιο αξιόπιστη μέθοδος, ενώ μια νέα εφαρμογή για συνδυαστική πρόγνωση βρίσκεται υπό κατασκευή.

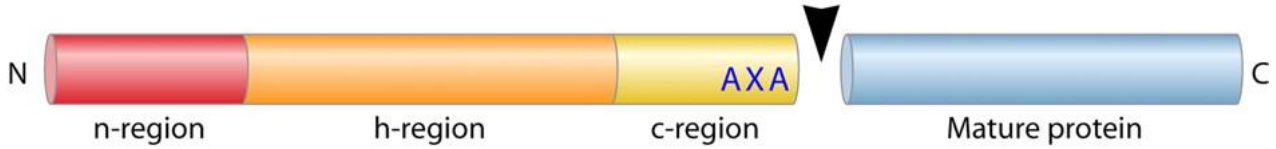
ConBBPRED


1 PRED-TMBB
2 PROFtmb
3 B2TMPRED
4 PSIPRED
5 TBBPred (NN)
6 TBBPred (SVM)
C Consensus
1.0
0.0
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
ConBBPRED, 2004


A
SPase
N
n
h
MKWVTFISLLFSSAYS
albumin
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVS
prolactin
MGQIVTMFEALPHIIDEVININVIIVLIIITSIKAVYNFATCGILALVSFLFLAGRSCG
LCMV-GPC
MPNHQSGSPTGSSDLLLSGKKQRPHLALRRKRRREMRKINRKVRRMNLAPIKEKTAWQHLQALISEAEEVLKTSQTPQNSLTLFLALLSVLGPPPVTG
MMTV-Rem
B
ribosome with
nascent protein
SRP
ER
membrane
translocon
SRP
receptor
SPase
N
C
signal sequences affect protein targeting,
insertion and their cleavage
in a qualitative and quantitative manner
1) processing of signal peptide
by signal peptide peptidase
SPP
2) membrane-inserted
signal peptide
3) release of a signal peptide
from the ER membrane

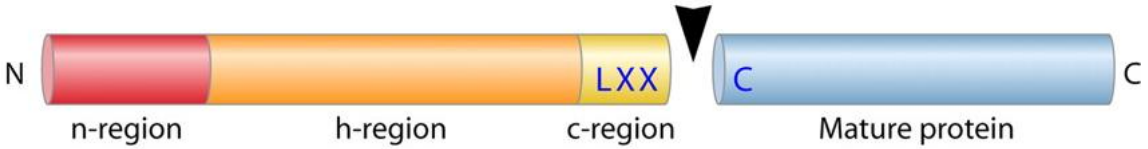


Bacterial signal peptide (SPI)

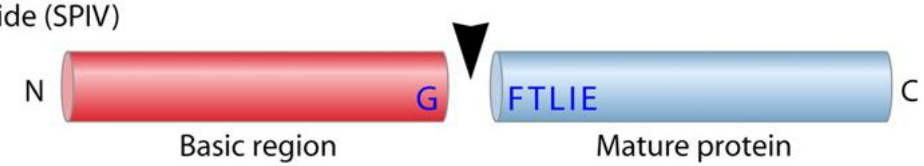


Tat signal peptide (SPI)

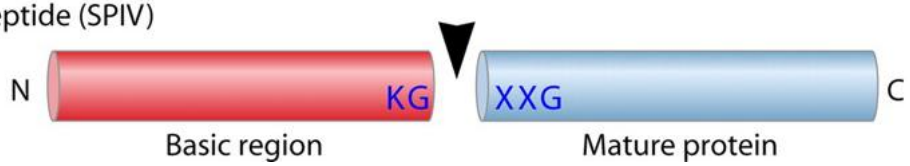
Lipoprotein signal peptide (SPII)



Bacterial prepilin signal peptide (SPIV)



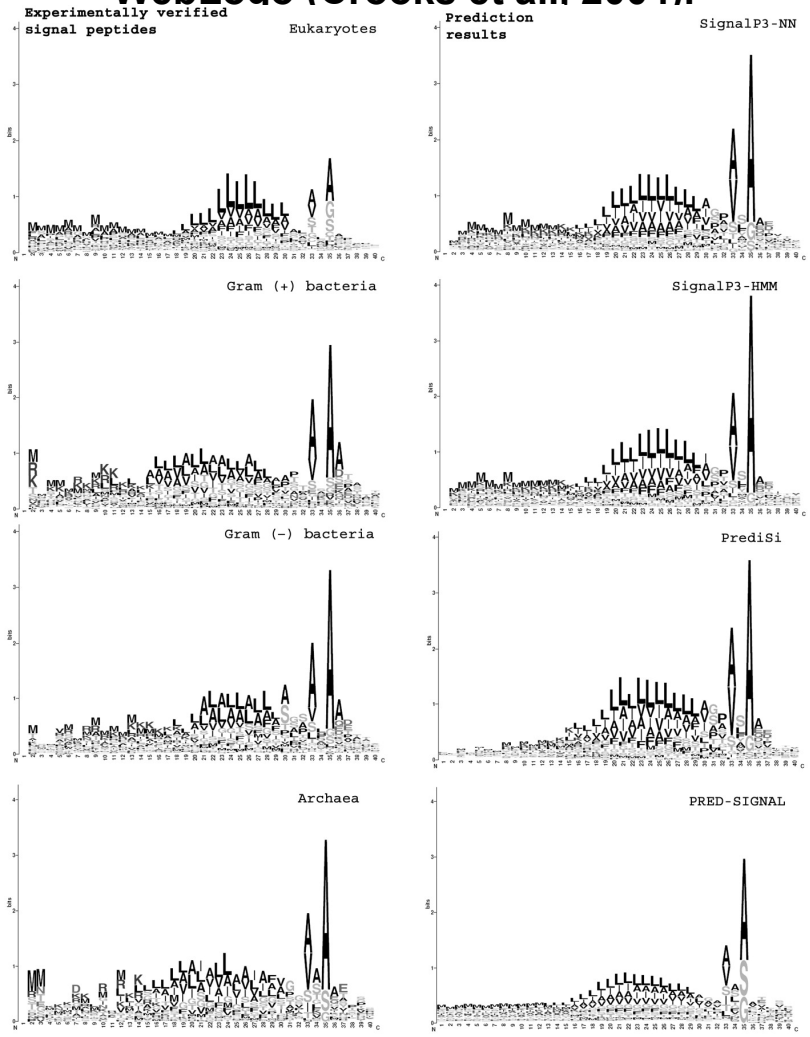
Archaeal preflagellin signal peptide (SPIV)



Μέθοδοι

- Η υπολογιστική πρόγνωση των πεπτιδίων οδηγητών, αλλά και των άλλων σηματοδοτικών αλληλουχιών, ήταν ένα σημαντικό πρόβλημα, ήδη από τη δεκαετία του 1980. Αρχικά χρησιμοποιήθηκαν weight matrices βασισμένοι στην ανάλυση του Gunnar von Heijne (von Heijne, 1986), και ο πιο γνωστός αλγόριθμος που βασίζεται σε αυτή τη μέθοδο, είναι το **SigCleave,** το οποίο υπάρχει διαθέσιμο σε πολλές εκδόσεις (http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/sigcleave.html).
- Μια πιο σύγχρονη μέθοδος βασισμένη σε weight matrices, η οποία έχει εκπαιδευθεί σε περισσότερα και καλύτερης ποιότητας δεδομένα, είναι το **PrediSi** (http://www.predisi.de/). Η μέθοδος αυτή, όπως και οι περισσότερες σύγχρονες μέθοδοι, έχει διαφορετικές εκδόσεις για τις τρεις μεγάλες κατηγορίες οργανισμών (Ευκαρυωτικοί, αρνητικά κατά Gram Βακτήρια, θετικά κατά Gram Βακτήρια).
- Οι πιο αποδοτικές όμως σύγχρονες μεθοδολογίες, βασίζονται σε μεθόδους μηχανικής μάθησης όπως τα νευρωνικά δίκτυα και τα HMM. Η πιο καλή και η πιο γνωστή από τις σύγχρονες μεθόδους, είναι το **SignalP** (http://www.cbs.dtu.dk/services/SignalP), το οποίο έχει φτάσει ήδη την έκδοση 4.1, και εκτός του ότι διαθέτει ξεχωριστά εργαλεία για την κάθε ομάδα οργανισμών και δυο διαφορετικές μεθόδους (νευρωνικά δίκτυα και HMM), ενώ βασίζεται στην εξαιρετική βιβλιογραφική αναζήτηση για την κατάρτιση του συνόλου εκπαίδευσης, περιλαμβάνοντας έτσι πολλές πρωτεΐνες, αλλά και απομακρύνοντας λάθος καταχωρίσεις (Bendtsen, Nielsen, von Heijne, & Brunak, 2004).
- Όπως ήδη αναφέραμε, κάποιες μέθοδοι πρόγνωσης διαμεμβρανικών πρωτεϊνών διαθέτουν επιπλέον την ικανότητα να προβλέπουν τα πεπτίδια οδηγητές. Οι μέθοδοι αυτές είναι το **Phobius**, διαθέσιμο στη διεύθυνση http://phobius.sbc.su.se/ (Kall et al., 2004; Kall, Krogh, & Sonnhammer, 2007) και το **Philius** (Reynolds, Kall, Riffle, Bilmes, & Noble, 2008), το οποίο είναι διαθέσιμο στη διεύθυνση http://noble.gs.washington.edu/proj/philius/, οι οποίες χρησιμοποιούν γραφικά μοντέλα (HMM και Bayesian network, αντίστοιχα), ενώ αργότερα εμφανίστηκε και το **SPOCTOPUS** (http://octopus.cbr.su.se/index.php?about=SPOCTOPUS).

Left panel (from top to bottom): the sequence logos of experimentally verified eukaryal, gram-positive, gram-negative and archaeal signal peptides (SPs), respectively, produced by WebLogo (Crooks et al., 2004).



P.G. Bagos et al. Protein Engineering, Design and Selection 2009;22:27-35

peds
protein engineering design & selection

Λιποπρωτεΐνες

- Οι βακτηριακές λιποπρωτεΐνες για πολλά χρόνια αναγνωρίζονταν με χρήση κανονικών εκφράσεων της PROSITE, όπως αυτές που αναφέραμε στο κεφάλαιο 5 (π.χ. το PS00013).
- Παρ' όλα αυτά, τα τελευταία χρόνια αναπτύχθηκαν και για αυτές τις πρωτεΐνες πιο σύγχρονες μέθοδοι. Αρχικά αναπτύχθηκε το **LipoP** (http://www.cbs.dtu.dk/services/LipoP), το οποίο βασίστηκε σε HMM και είχε εκπαιδευθεί να αναγνωρίζει λιποπρωτεΐνες από αρνητικά κατά Gram βακτήρια (Juncker et al., 2003). Το LipoP έχει επιπλέον την ειδική ικανότητα να προβλέπει εξίσου καλά και πεπτίδια οδηγητές εκκρινόμενων πρωτεϊνών, αλλά και διαμεμβρανικές έλικες στο αμινοτελικό άκρο και έχει μια επιτυχία της τάξης του 97% στη σωστή ταξινόμηση στις λιποπρωτεΐνες από αρνητικά κατά Gram βακτήρια, ενώ δίνει λάθος προβλέψεις (δηλαδή, σε μη εκκρινόμενες πρωτεΐνες), της τάξης του 0.3%. Παρ' όλα αυτά, όταν χρησιμοποιηθεί σε λιποπρωτεΐνες από θετικά κατά Gram βακτήρια, η ακρίβειά του πέφτει περίπου στο 90-92%.
- Έτσι, σε μια παλιότερη εργασία μας, αφού πραγματοποιήσαμε εκτεταμένη αναζήτηση στη βιβλιογραφία για την εύρεση πειραματικά προσδιορισμένων λιποπρωτεϊνών από θετικά κατά Gram βακτήρια, κατασκευάσαμε το **PRED-LIPO** (http://www.compgen.org/tools/PRED-LIPO), το οποίο αποδίδει καλύτερα σε αυτή την κατηγορία βακτηρίων, ενώ παράλληλα προβλέπει με αρκετά μεγάλη ακρίβεια και τα πεπτίδια των εκκρινόμενων πρωτεϊνών, αλλά και τις διαμεμβρανικές έλικες (Bagos, Tsirigos, Liakopoulos, & Hamodrakas, 2008).

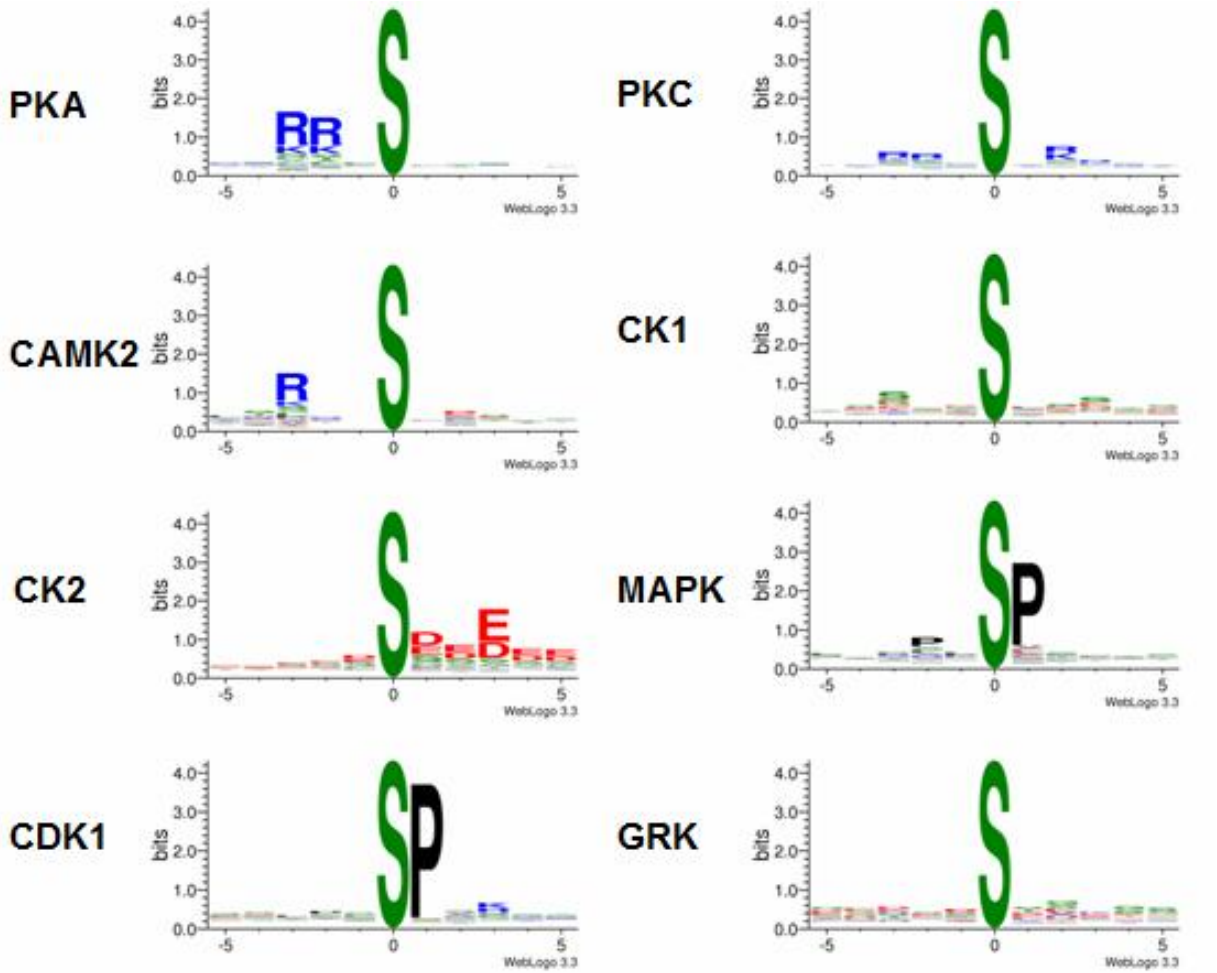
TAT

- Παρ' όλο που πολλές από τις μεθόδους που αναφέραμε, μπορούν να προβλέψουν (μέχρι κάποιο βαθμό) και τα πεπτίδια που οδηγούνται μέσω του συστήματος Tat (χωρίς όμως να μπορούν να τα διαχωρίσουν), έχουν αναπτυχθεί τα τελευταία χρόνια και ειδικές μεθοδολογίες που δουλεύουν καλύτερα στις πρωτεΐνες αυτής της κατηγορίας.
- Η πρώτη τέτοια μέθοδος ήταν το **TATFIND** (http://signalfind.org/tatfind.html), το οποίο βασιζόταν σε ανάλυση υδροφοβικότητας και σε κανονικές εκφράσεις (Rose, Bruser, Kissinger, & Pohlschroder, 2002).
- Λίγα χρόνια αργότερα εμφανίστηκε το **TatP** (http://www.cbs.dtu.dk/services/TatP/), το οποίο χρησιμοποιεί νευρωνικά δίκτυα αλλά και κανονικές εκφράσεις για να διακρίνει την περιοχή RR (Bendtsen, Nielsen, Widdick, Palmer, & Brunak, 2005). Το TatP είναι γενικά αξιόπιστο, αλλά όχι στα επίπεδα του SignalP, ενώ το TATFIND αναγνωρίζει μόνο την ύπαρξη του σήματος RR, αλλά όχι και το σημείο αποκοπής.
- Σε μια προσπάθεια να επιλύσουμε όλα αυτά τα προβλήματα, παρουσιάσαμε πρόσφατα το **PRED-TAT** (http://www.compgen.org/tools/PRED-TAT/), μια μέθοδο βασισμένη στα HMMs, η οποία μπορεί αφενός μεν να διαχωρίσει τα πεπτίδια οδηγητές (Sec και Tat), αφετέρου δε, να προβλέψει και τις θέσεις αποκοπής στις δύο κατηγορίες. Η μέθοδος αυτή, είναι αυτή τη στιγμή, η κορυφαία για τα Tat πεπτίδια οδηγητές, αλλά ταυτόχρονα προβλέπει και τα κλασικά πεπτίδια (Sec) σε ικανοποιητικό βαθμό, ενώ υστερεί ελάχιστα σε αυτή την κατηγορία σε σχέση με το SignalP (Bagos et al., 2010).

Αλληλουχίες στόχευσης

- Σχετικά με τις σηματοδοτικές αλληλουχίες που κατευθύνουν τις πρωτεΐνες στα μιτοχόνδρια και τους χλωροπλάστες, έχουν επίσης αναπτυχθεί εξειδικευμένοι αλγόριθμοι.
- Για τους χλωροπλάστες, ο πιο γνωστός είναι το **ChloroP** (http://www.cbs.dtu.dk/services/ChloroP),
- ενώ το **TargetP** (http://www.cbs.dtu.dk/services/TargetP), είναι ένα ολοκληρωμένο σύστημα που προβλέπει τόσο τις εκκριτικές πρωτεΐνες, όσο και αυτές των μιτοχονδρίων και των χλωροπλαστών.
- Παρόμοιας αρχιτεκτονικής και φιλοσοφίας είναι το κάπως παλιότερο **iPSORT** (http://ipsort.hgc.jp/how.html).
- Άλλα εργαλεία που προβλέπουν τις μιτοχονδριακές σηματοδοτικές αλληλουχίες, είναι το **MitoProt** (https://ihg.gsf.de/ihg/mitoprot.html),
- το **Predotar** (http://urgi.versailles.inra.fr/predotar/predotar.html),
- και το **Tppred2** (http://tppred2.biocomp.unibo.it).
- Για τις πρωτεΐνες των υπεροξεισωμάτων, υπάρχει το **PTS1 predictor** (http://mendel.imp.ac.at/mendeljsp/sat/pts1/PTS1predictor.jsp),
- ενώ για τις πρωτεϊνες που κατευθύνονται στον πυρήνα έχει αναπτυχθεί το **cNLS Mapper** (http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi),
- το **NLStradamus** (http://www.moseslab.csb.utoronto.ca/NLStradamus/),
- το **NucPred** (http://www.sbc.su.se/~maccallr/nucpred/)
- και το **PredictNLS** (https://rostlab.org/owiki/index.php/PredictNLS).

Μετα-μεταφραστικές τροποποιήσεις

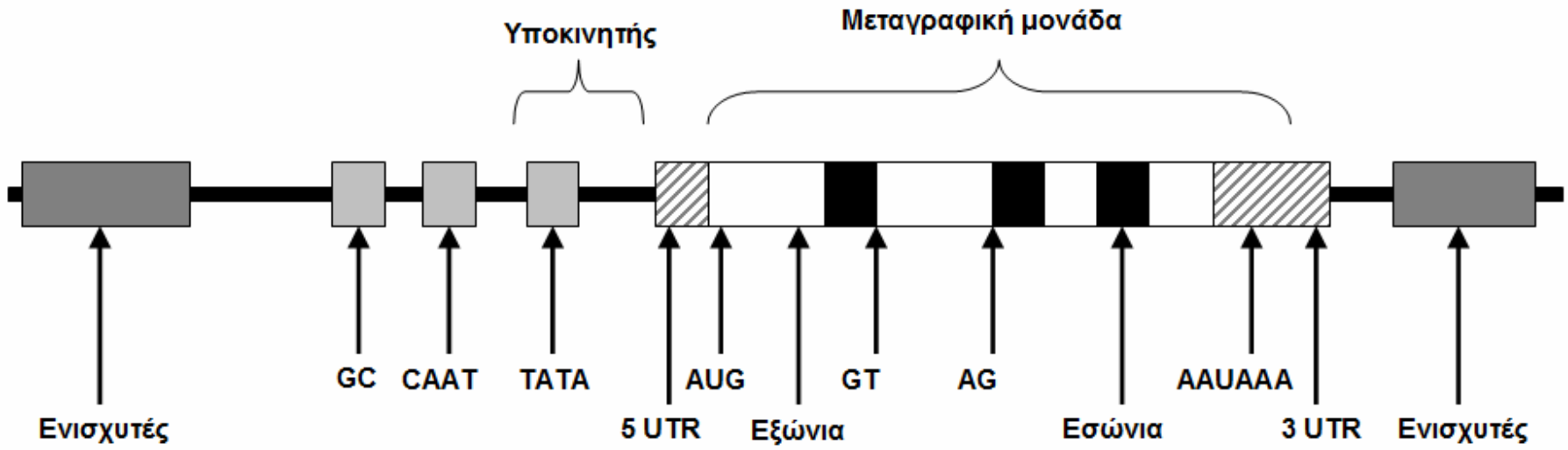


Φωσφορυλίωση

- Η φωσφορυλίωση, είναι επίσης μια πολύ σημαντική κατηγορία τροποποιήσεων που συνίσταται στην προσθήκη φωσφορικής ομάδας, συνήθως στην πλευρική ομάδα της Σερίνης, της Θρεονίνης ή της Τυροσίνης.
- Τα ένζυμα που πραγματοποιούν αυτές τις αντιδράσεις ονομάζονται κινάσες και η διαδικασία αυτή χρησιμεύει σαν αντιστρεπτός μηχανισμός σηματοδότησης και ενεργοποίησης διαφόρων μηχανισμών.
- Η πιο γνωστή μέθοδος πρόγνωσης είναι το **NetPhos** (http://www.cbs.dtu.dk/services/NetPhos/) που βασίζεται σε νευρωνικά δίκτυα, ενώ η πιο εξελιγμένη έκδοση **NetPhosK** (http://www.cbs.dtu.dk/services/NetPhosK/) προβλέπει και το είδος της κινάσης που πραγματοποιεί την κάθε αντίδραση.
- Το **GPS** (http://gps.biocuckoo.org/) είναι ένα άλλο εργαλείο για πρόγνωση της φωσφορυλίωσης (περιέχει και μεθόδους πρόγνωσης και για άλλες μετα-μεταφραστικές τροποποιήσεις). Το **KinasePhos2** (http://kinasephos2.mbc.nctu.edu.tw/) είναι μια ακόμα γνωστή εφαρμογή για πρόγνωση των θέσεων φωσφορυλίωσης που προβλέπει και το είδος της κινάσης και βασίζεται σε HMM.
- Άλλες μέθοδοι είναι το **PhosphoSVM** (http://sysbio.unl.edu/PhosphoSVM/), το **DISPHOS** (http://www.dabi.temple.edu/disphos/), το **pkaPS** (http://mendel.imp.ac.at/sat/pkaPS/) και το **Predikin** (http://predikin.biosci.uq.edu.au/).
- Εμπειρικές μελέτες έχουν δείξει ότι οι υπάρχουσες μέθοδοι πρόγνωσης έχουν σχετικά μικρή ακρίβεια και πολλές φορές δρουν συμπληρωματικά (άλλες έχουν μεγάλη ευαισθησία, άλλες μεγάλη ειδικότητα), κατά συνέπεια, μια συνδυαστική μέθοδος μπορεί να αποδώσει καλύτερα.
- Η μόνη προς το παρόν τέτοια μέθοδος είναι το **MetaPredPS** (http://c1.accurascience.com/MetaPred/MetaPredPS_091201/).
- Πολλές φορές επίσης, σε ειδικές κατηγορίες οργανισμών, οι γενικές μέθοδοι δεν αποδίδουν καλά, οπότε υπάρχει και η ανάγκη για εξειδικευμένες μεθόδους όπως το **NetPhosYeast** (http://www.cbs.dtu.dk/services/NetPhosYeast/) και το **NetPhosBac** (http://www.cbs.dtu.dk/services/NetPhosBac-1.0/).

DNA/RNA

- Έύρεση γονιδίων
- Εύρεση υποκινητών
- Σημεία συρραφής
- TIS
- Poly-A
- miRNA


Υποκινητής
Μεταγραφική μονάδα
GC
CAAT
TATA
AUG
GT
AG
AAUAAA
Ενισχυτές
5 UTR
Εξώνια
Εσώνια
3 UTR
Ενισχυτές

Gene-finders

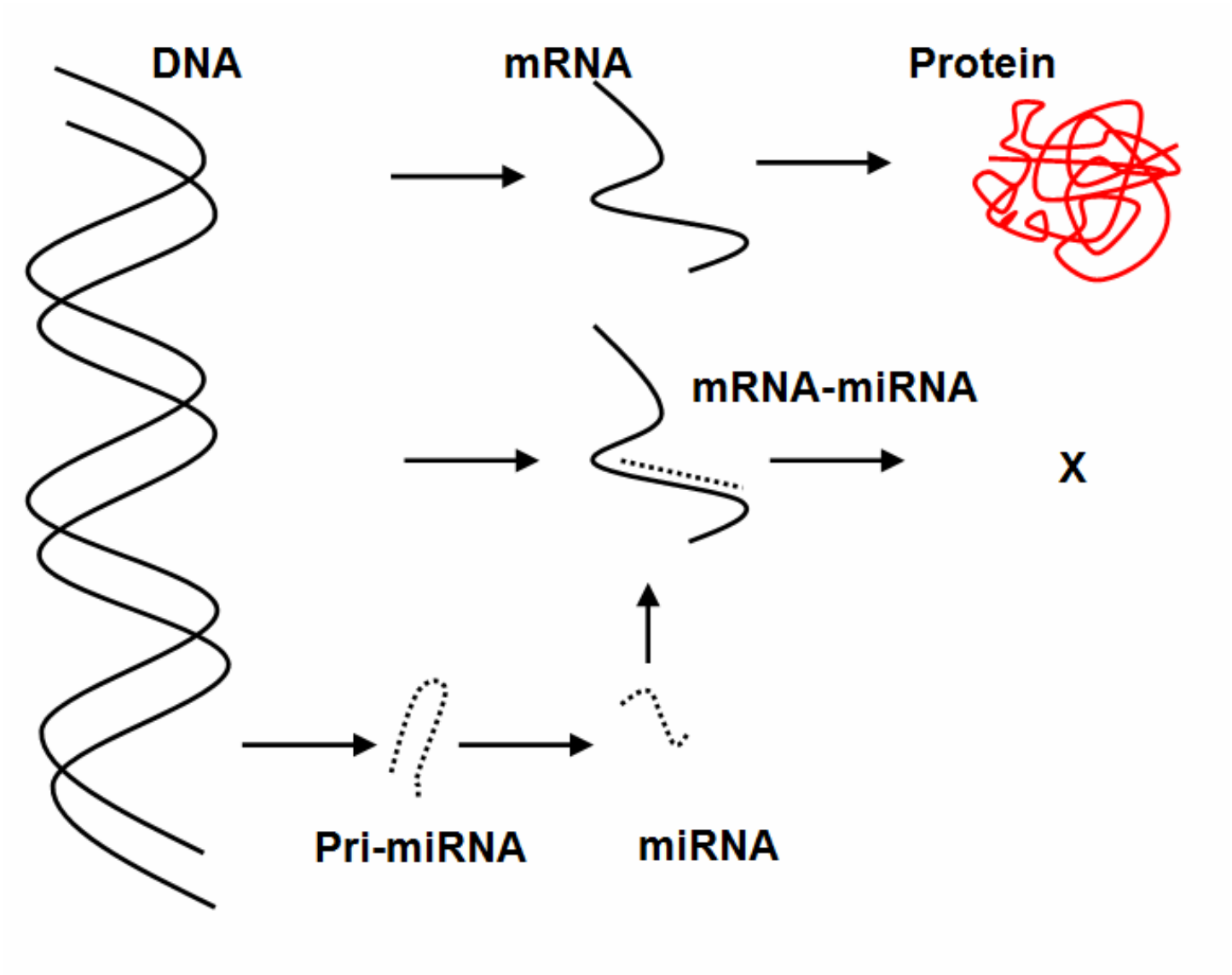
- Για τους προκαρυωτικούς οργανισμούς, τα πιο γνωστά και πετυχημένα εργαλεία περιλαμβάνουν τα:
- **FrameD** (http://tata.toulouse.inra.fr/apps/FrameD/FD)
- **GeneMark** (http://exon.gatech.edu/GeneMark/gmchoice.html)
- **Glimmer** (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi)
- **EasyGene** (http://www.cbs.dtu.dk/services/EasyGene/)
- **FGENESB (**http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=help&subgroup=gfindb)
- **Prodigal** (http://prodigal.ornl.gov/)

Gene-finders

- Αντίστοιχα, για τους ευκαρυωτικούς οργανισμούς, τα πιο πετυχημένα αντίστοιχα εργαλεία είναι:
- **FGENESH** (http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind)
- **GlimmerHMM** (https://ccb.jhu.edu/software/glimmerhmm/)
- **HMMgene** (http://www.cbs.dtu.dk/services/HMMgene/)
- **GeneMark.hmm** (http://exon.gatech.edu/GeneMark/hmmchoice.html)
- **GeneID** (http://genome.crg.es/software/geneid/geneid.html)
- **GeneScan** (http://genes.mit.edu/GENSCAN.html)
- **mGene** (http://raetschlab.org/suppl/mgene)
- **Grail** (http://compbio.ornl.gov/grailexp/)

Άλλα εργαλεία

- Ειδικά εργαλεία για την έναρξη της μεταγραφής (translation initiation) είναι:
  - **ATGpr** (http://atgpr.dbcls.jp/)
  - **NetStart** (http://www.cbs.dtu.dk/services/NetStart/)
  - **TIS Miner** (http://dnafsminer.bic.nus.edu.sg/Tis.html)
  - **StartScan** (http://bioinformatics.psb.ugent.be/webtools/startscan/)
- Για την πολυαδενυλίωση του mRNA τα διαθέσιμα εργαλεία αυτή τη στιγμή είναι:
  - **Poly(A) Signal Miner** (http://dnafsminer.bic.nus.edu.sg/)
  - **PolyAPred** (http://www.imtech.res.in/raghava/polyapred/help.html)
  - **POLYAH** (http://www.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter)
  - **PolyApredict** (http://cub.comsats.edu.pk/polyapredict.htm)
- Τέλος, μέθοδοι που εστιάζονται στην εύρεση των σημείων αποκοπής και συρραφής εσωνίων/εξωνίων σε ευκαρυωτικά γονιδιώματα, είναι:
  - **Human Splice Finder** (http://www.umd.be/HSF3/)
  - **NetGene** (http://www.cbs.dtu.dk/services/NetGene2/)
  - **NetPlant** (http://www.cbs.dtu.dk/services/NetPGene/)
  - **GeneSplicer** (https://ccb.jhu.edu/software/genesplicer/)
  - **SpliceView** (http://bioinfo4.itb.cnr.it/~webgene/wwwspliceview_ex.html)
  - **SplicePredictor** (http://bioservices.usd.edu/splicepredictor/)


DNA
mRNA
Protein
mRNA-miRNA
X
Pri-miRNA
miRNA

miRNA

- **CID miRNA** (http://melb.agrf.org.au:8888/cidmirna/)
- **MiRPara** (https://code.google.com/p/mirpara/)
- **HeteroMirPred** (http://ncrna-pred.com/premiRNA.html)
- **HHMMiR** (http://biodev.hgen.pitt.edu/kadriAPBC2009.html)
- **HuntMi** (http://adaa.polsl.pl/agudys/huntmi/huntmi.htm)
- **MaturePred** (http://nclab.hit.edu.cn/maturepred/)
- **microPred** (http://www.cs.ox.ac.uk/people/manohara.rukshan.batuwita/microPred.htm)
- **MiPred** (http://www.bioinf.seu.edu.cn/miRNA/)
- **miRAbela** (http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi)
- **MiRAlign** (http://bioinfo.au.tsinghua.edu.cn/miralign/)
- **miRBoost** (http://evryrna.ibisc.univ-evry.fr/miRBoost/index.html)
- **mirnaDetect** (http://datamining.xmu.edu.cn/main/~leyiwei/mirnaDetect.html)
- **miRNAFold** (http://evryrna.ibisc.univ-evry.fr/miRNAFold/)
- **MiRscan** (http://genes.mit.edu/mirscan/)
- **novoMIR** (http://www.biophys.uni-duesseldorf.de/novomir/)
- **ProMiR** (http://bi.snu.ac.kr/Research/ProMiR/ProMiR.html)
- **RNAmicro** (http://www.tbi.univie.ac.at/~jana/software/RNAmicro.html)
- **tripletSVM** (http://bioinfo.au.tsinghua.edu.cn/mirnasvm/)
- **SplamiR** (http://www.uni-jena.de/SplamiR.html)
- **SSCprofiler** (http://mirna.imbb.forth.gr/SSCprofiler.html)
- **EumiR** (http://miracle.igib.res.in/eumir/)

miRNA target

- **Diana Micro-T** (http://diana.cslab.ece.ntua.gr/microT/)
- **PicTar** (http://pictar.mdc-berlin.de/)
- **TargetScan** (http://www.targetscan.org/)
- **miRTar** (http://mirtar.mbc.nctu.edu.tw/human/)
- **miRanda** (http://www.microrna.org/microrna/home.do)
- **MaMi** (http://mami.med.harvard.edu/)
- **ComiR** (http://www.benoslab.pitt.edu/comir/) (συνδυαστική μέθοδος)
- **PITA** (http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html)
- **MirMap** (http://mirmap.ezlab.org/)
- **STarMir** (http://sfold.wadsworth.org/starmir.html)