Πρωτεϊνική Μηχανική

Ιωάννης Παυλίδης
Τμήμα Χημείας | Πανεπιστήμιο Κρήτης

E-mail: ipavlidis@uoc.gr
Τηλ.: +30 2810 54-5130 | Γραφείο Γ211

ΕΠΙΣΗΜΗ ΣΕΛΊΔΑ: HTTP://WWW.CHEMISTRY.UOC.GR/PAVLIDIS/

Εισαγωγή στη βιοπληροφορική

Βιοπληροφορική

Η Βιοπληροφορική είναι η χρήση Η/Υ για την απόκτηση, διαχείριση και ανάλυση βιολογικών δεδομένων

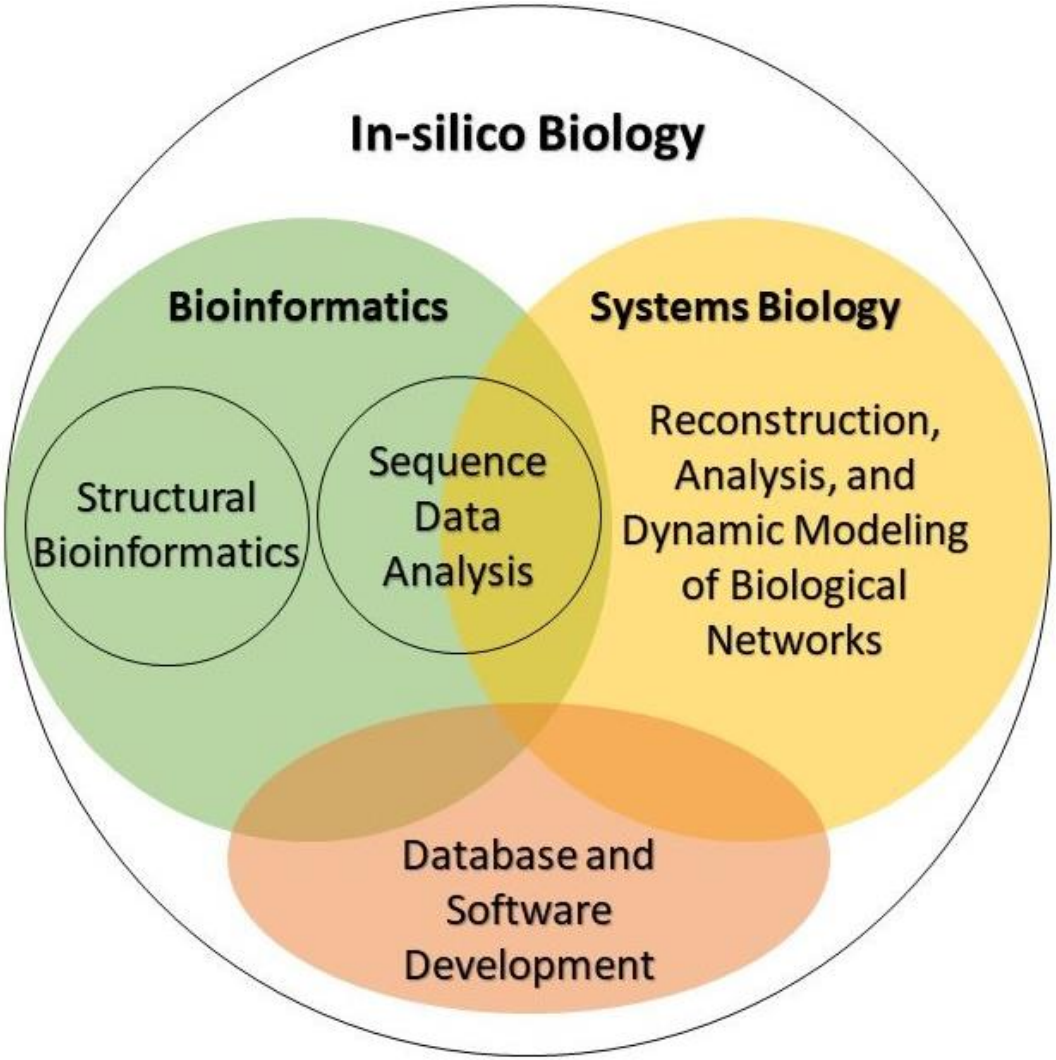
Περιλαμβάνει πολλά πεδία της επιστήμης υπολογιστών:

- Σχεδιασμός βάσεων δεδομένων
- Προγραμματισμός
- Διαχείριση δικτύων και πληροφοριακών συστημάτων

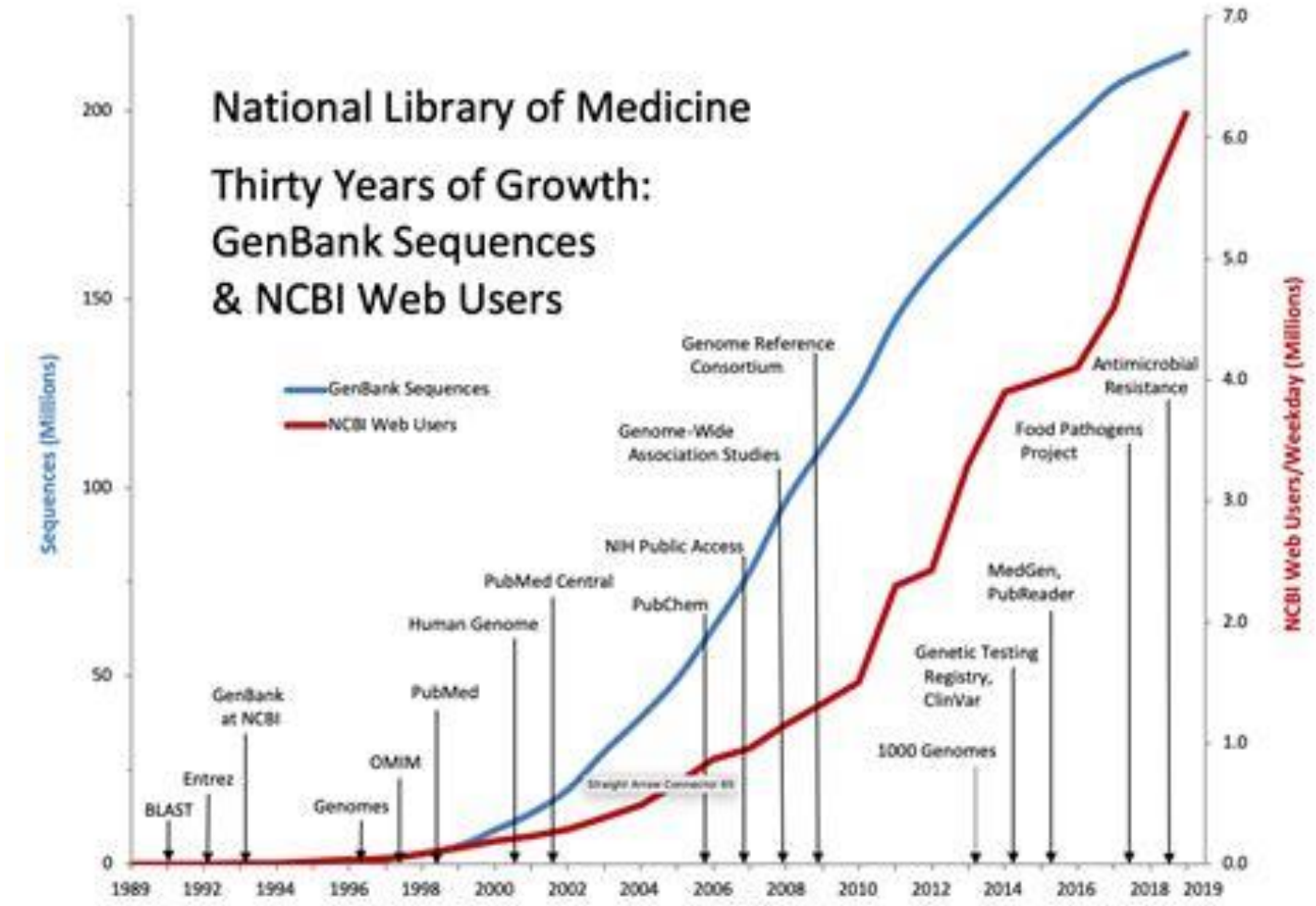
Μπορεί να εφαρμοστεί σε πολλά πεδία:

- Μοριακή βιολογία
- Γενετική
- Πρωτεϊνική βιοτεχνολογία

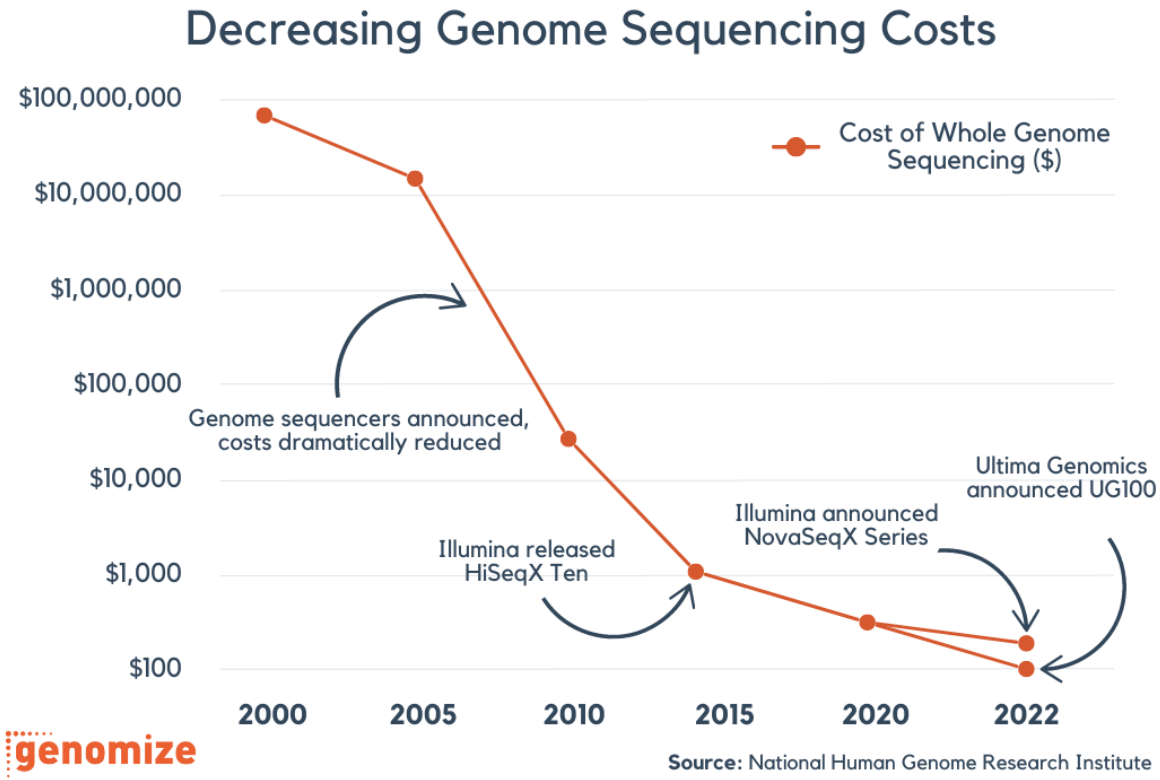
Βιοπληροφορική



Ραγδαία παραγωγή δεδομένων



Ραγδαία παραγωγή δεδομένων



Sequencing by Expansion (SBX) platform

«Αντιστροφή» της επιστημονικής μεθόδου

Η επιστημονική μέθοδος:

Σχηματισμός Υπόθεσης → **Συλλογή δεδομένων για επιβεβαίωση**

Η «νέα» επιστημονική μέθοδος με την βιοπληροφορική:

Ανάλυση (πολλών) δεδομένων → **Δημιουργία υπόθεσης**

Μία θεωρία γεννάται από τα δεδομένα

Data-mining: Νέα υπολογιστικά εργαλεία παρέχουν πρόσβαση και νέους τρόπους ανάλυσης της αχανούς πληροφορίας στις βάσεις δεδομένων

Ιστορία της βιοπληροφορικής

1960s: Πρώτα «βήματα» της βιοπληροφορικής

Η **Dayhoff** δημιούργησε την **πρώτη πρωτεϊνική βάση δεδομένων** και το πρώτο πρόγραμμα **για την συναρμολόγηση αλληλουχιών**

Δημιουργός και του κώδικα των αμινοξέων με ένα γράμμα

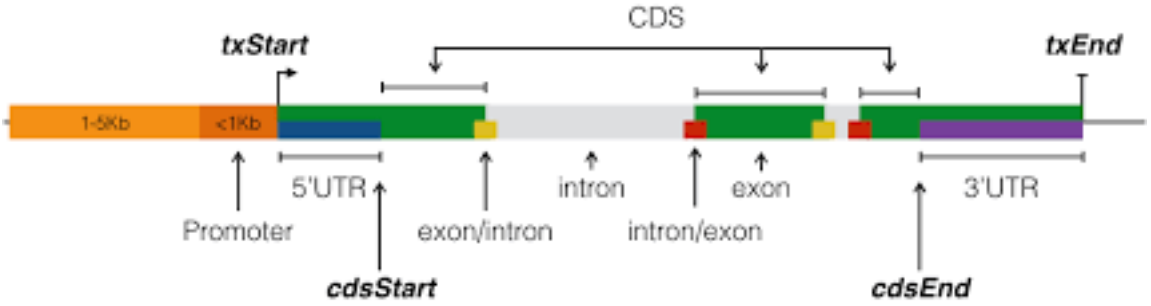
Margaret Oakley Dayhoff

IBM 7090 computer

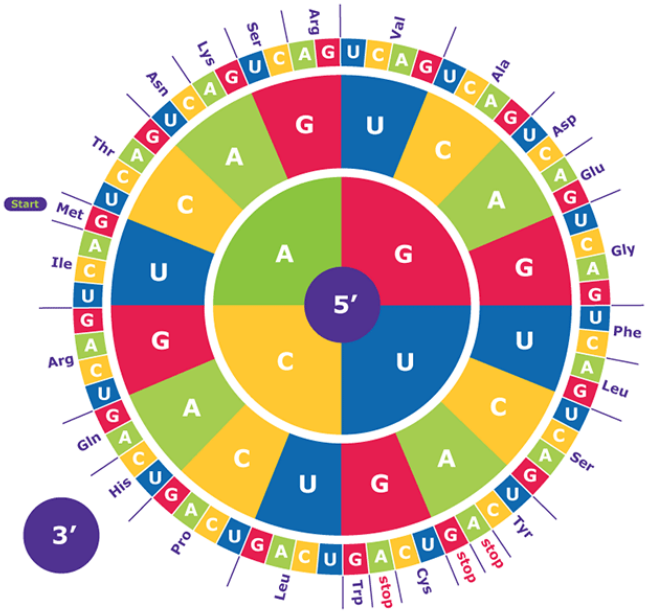
Αλγόριθμοι για ανάλυση δεδομένων

Ανάλυση αλληλουχιών DNA

Δομή του DNA και περιοχές στο γονιδίωμα



Μετάφραση της αλληλουχίας του DNA σε αμινοξική αλληλουχία



Ανάλυση αλληλουχιών DNA

```
CGCCTTTGACCGCCGCGGCTTTGGCCGCTCGGACCAACCCTGGACCGGCAACGACTACGACACCT
TCGCCGACGACATCGCCCAGTTGATCGAACACCTGGACCTCAAGGAGGTGACCCTGGTGGGCTTC
TCCATGGGCGGCGGCGATGTGGCCCGCTACATCGCCCGCCACGGCAGCGCACGGGTGGCCGGCCT
GGTGCTGCTGGGCGCCGTCACCCCGCTGTTCGGCCAGAAGCCCGACTATCCGCAGGGTGTCCCGC
TCGATGTGTTCGCAAGGTTCAAGACTGAGCTGCTGAAGGATCGCGCGCAGTTCATCAGCGATTTC
AACGCACCGTTCTATGGCATCAACAAGGGCCAGGTCGTCTCCCAAGGCGTGCAGACCCAGACCCT
GCAAATCGCCCTGCTGGCCTCGCTCAAGGCCACGGTGGATTG
```

- Από που ξεκινάω να διαβάζω;
- Πόσους τρόπους έχω να διαβάσω την συγκεκριμένη αλληλουχία;
- Είναι προκαρυωτικό ή ευκαρυωτικό γονίδιο (αν είναι όντως κωδικοποιούσα περιοχή);

Hints:

Πάντα χρησιμοποιείτε την γραμματοσειρά **`Courier`** για αλληλουχίες!

Κρατά ίσο πλάτος για κάθε γράμμα

.fasta: Γενικός τύπος αρχείων για αλληλουχίες DNA/πρωτεϊνών.

Συνήθως χρησιμοποιούνται κεφαλαία για αμινοξέα, μικρά για DNA

.gb: Γενικός τύπος αρχείων για αλληλουχίες DNA με σημειώσεις

Ανάλυση αλληλουχιών DNA

Πως αναγνωρίζω ένα ανοιχτό πλαίσιο ανάγνωσης;

1. **Kozak consensus sequence**: Αλληλουχία κοντά στο κωδικόνιο έναρξης στους ευκαρυώτες CC(A/G)CCATGG
2. **RBS**: Αλληλουχία πρόσδεσης του ριβοσώματος στους προκαρυώτες
3. **Χρήση κωδικονίων** διαφορετική μεταξύ κωδικοποιούσας και μη κωδικοποιούσας περιοχής στο DNA
4. **Θέσης Wobble (3η)** προτίμηση σε G/C ή A/T ανά οργανισμό
5. **Στοίχιση** με γνωστές αλληλουχίες

Στοίχιση αλληλουχιών

Πρακτικές αναζήτησης σε βάσεις δεδομένων:

- Αναζήτηση κειμένου:

  "*show all adrenaline receptors*"

- Αναζήτηση με ομοιότητα αλληλουχιών:

  Βρες όλες τις παρόμοιες αλληλουχίες με την αλληλουχία του ανθρώπινου υποδοχέα αδρεναλίνης

- Ο δεύτερος τρόπος είναι πιο ορθολογικός στην επιστήμη υπολογιστών
- Οι σημειώσεις μπορεί να είναι λάθος

Σύγκριση αλληλουχιών

Hamming distance: Σε δεδομένη αλληλουχία ίσου μήκους, είναι ο αριθμός των διαφορών (χαμηλότερο σκορ είναι καλύτερο)

Χωρίς στοίχιση, HD:3

```
AGGVLIIQV
AGGVLIQVG
```

Με στοίχιση, HD:1

```
AGGVLIIQV
AGGVLI-QVG
```

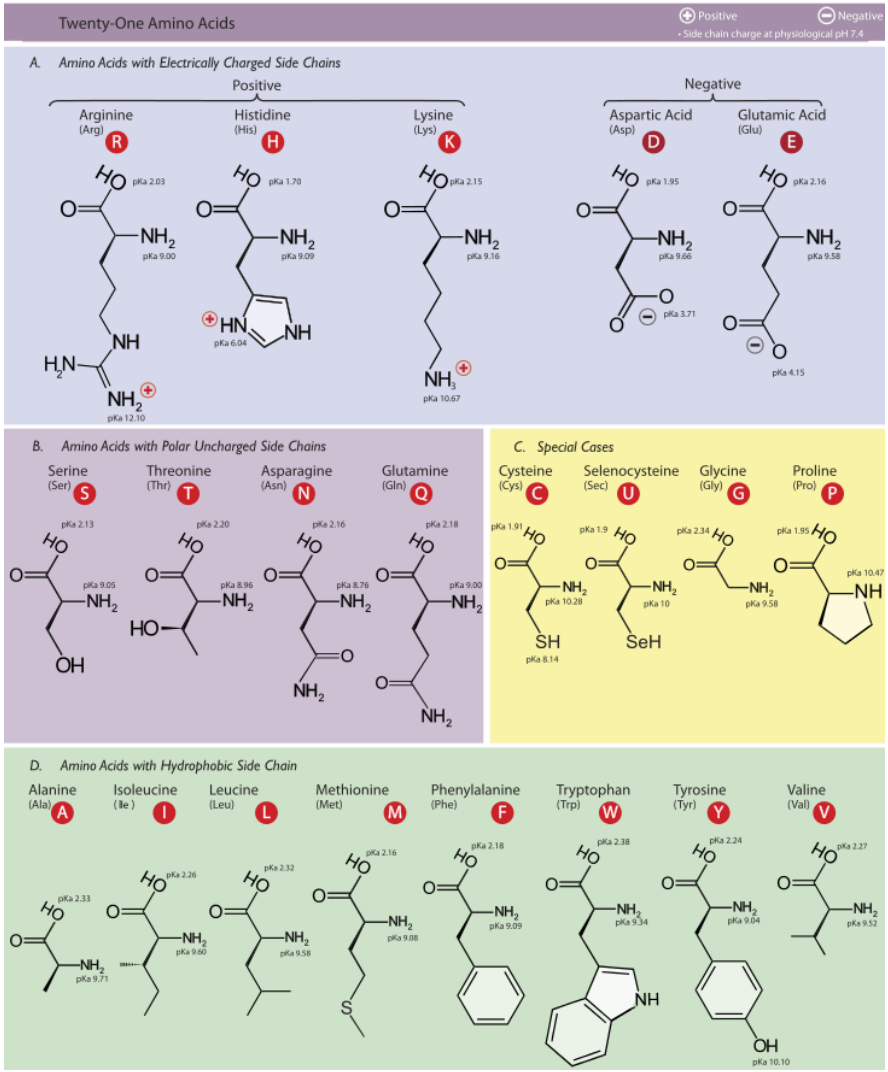
Σε πραγματικές αλληλουχίες, το ερώτημα είναι πιο περίπλοκο, καθώς οι αλληλουχίες είναι μεγαλύτερες και όχι ίδιου μεγέθους

Ταυτότητα και ομοιότητα

Η εύρεση ταυτότητας σε αλληλουχίες είναι εύκολη υπολογιστικά

Όταν μελετάμε και την ομοιότητα, το ερευνητικό ερώτημα γίνεται πιο περίπλοκο

- Πως στοιχίζουμε;
- Που επιτρέπουμε μία ασυμφωνία (mismatch) για να βελτιστοποιήσουμε το σκορ;



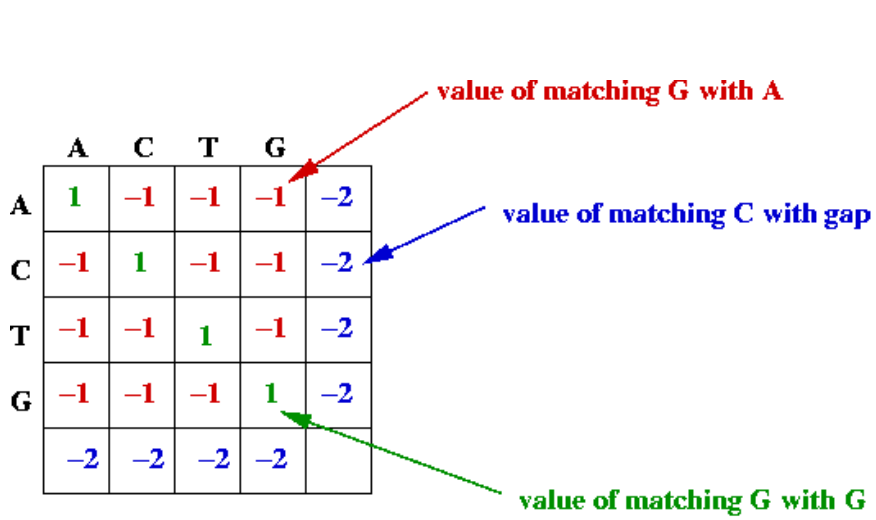
Σκορ στοίχισης

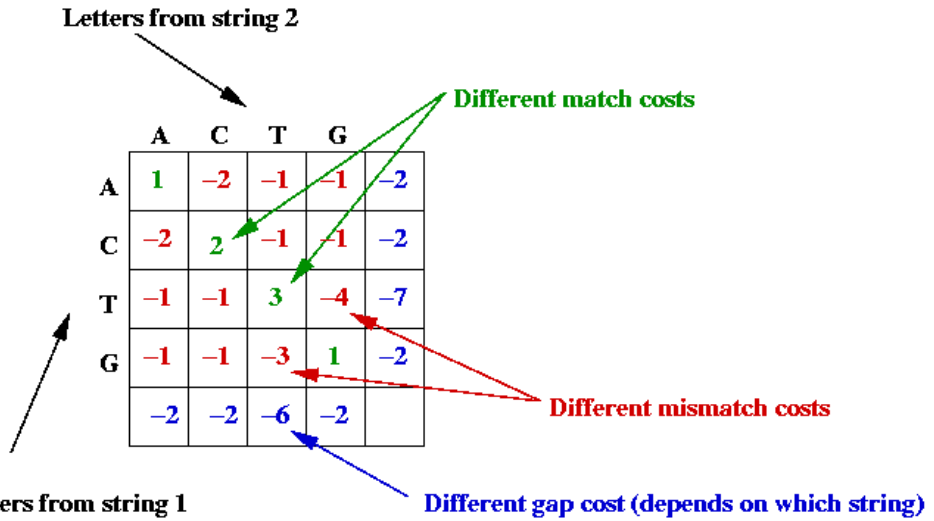
Το σκορ της στοίχισης έρχεται από συγκεκριμένους πίνακες:

- όταν ταιριάζουν οι θέσεις, έχει θετικό σκορ
- όταν δεν ταιριάζουν, έχει ένα αρνητικό σκορ

Ίση επίδραση

Πιθανή ασυμμετρία





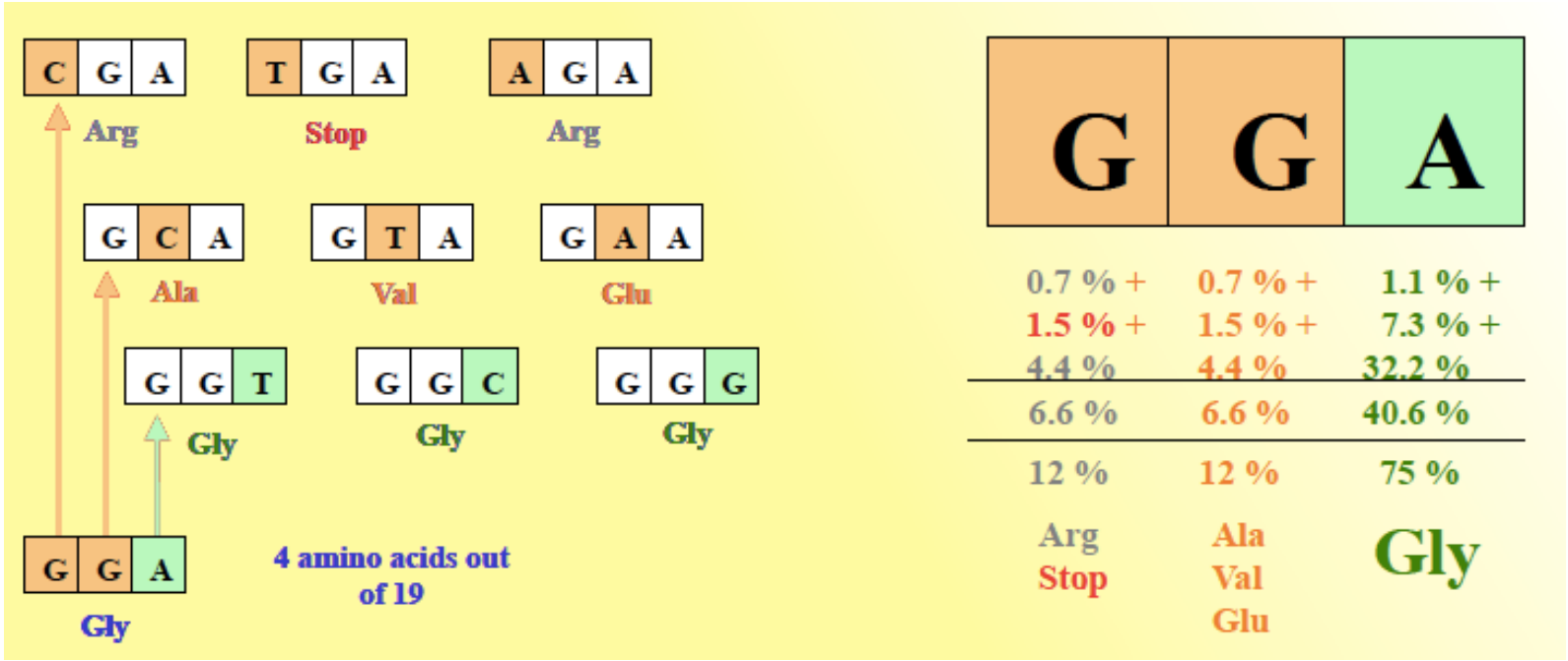
Αλγόριθμοι στοίχισης

Αλγόριθμος Dayhoff : βασισμένος στο PAM matrix

PAM: Point accepted mutation

Πιθανότητα της αλλαγής ενός αμινοξέος σε ένα άλλο σε 100 θέσεις

Πίνακες για μεγαλύτερες αλληλουχίες έχουν προετοιμαστεί με πολλαπλασιασμό πινάκων



Αλγόριθμοι στοίχισης

PAM 250 matrix

Log (πιθανότητα αλλαγής / πιθανότητα φυσικής παρουσίας)

Προσθέτουμε για να πάρουμε το τελικό σκορ

τιμές >0 είναι πιθανές μεταλλάξεις

τιμές = 0 ουδέτερες (τυχαίες μεταλλάξεις)

τιμές <0 μη πιθανές μεταλλάξεις

	C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp
C Cys	12																			
S Ser	0	2																		
T Thr	−2	1	3																	
P Pro	−3	1	0	6																
A Ala	−2	1	1	1	2															
G Gly	−3	1	0	−1	1	5														
N Asn	−4	1	0	−1	0	0	2													
D Asp	−5	0	0	−1	0	1	2	4												
E Glu	−5	0	0	−1	0	0	1	3	4											
Q Gln	−5	−1	−1	0	0	−1	1	2	2	4										
H His	−3	−1	−1	0	−1	−2	2	1	1	3	6									
R Arg	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
K Lys	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
M Met	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
I Ile	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
L Leu	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
V Val	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
F Phe	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
Y Tyr	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10	
W Trp	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17
	C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp

PAM-250 Log Odds Substitution Matrix

Αλγόριθμοι στοίχισης

Ο αλγόριθμος της Dayhoff δουλεύει ικανοποιητικά για ταυτότητα >85%

BLOSUM: BLock SUbstitution Matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Ο αλγόριθμος Dayhoff πολλαπλασιάζει τους πίνακες για μεγαλύτερες αλληλουχίες, δημιουργώντας έτσι τεχνικά σφάλματα

Ο BLOSUM κάνει τοπικές στοιχίσεις (blocks) και υπολογίζει τη μέση συνεισφορά κάθε στοιχείου (π.χ. αμινοξέος) σε κάθε θέση

Διάφοροι πίνακες, βασισμένοι στο ποσοστό ταυτότητας για κάθε block

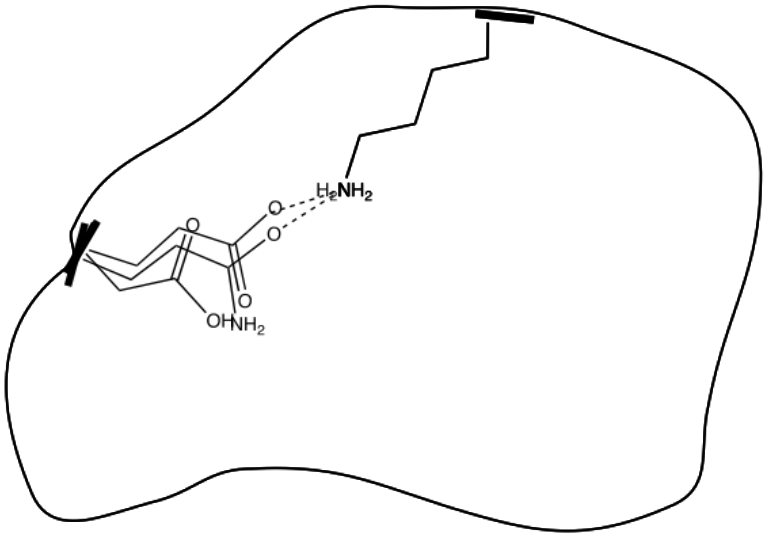
BLOSUM62: Identity≥62%

Στοίχιση βασισμένη στη δομή

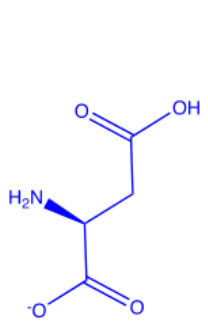
Στοιχίστε τις παρακάτω αλληλουχίες:

```
PIFENHGT      PIFENHGT      PIFENHGT
PI-DQHGT      PID-QHGT      PIDQ-HGT
```

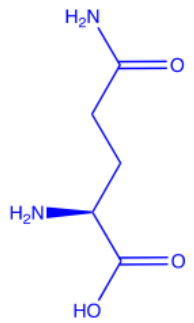
Αν το γλουταμικό οξύ παρέχει ένα δεσμό υδρογόνου σημαντικό για τη σωστή αναδίπλωση της πρωτεΐνης;



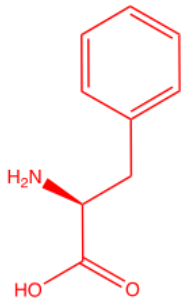
D



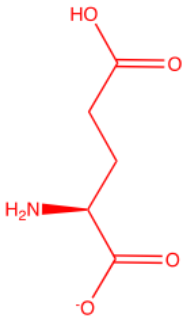
Q



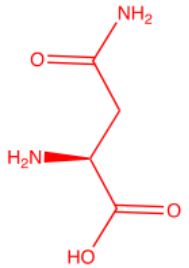
F



E



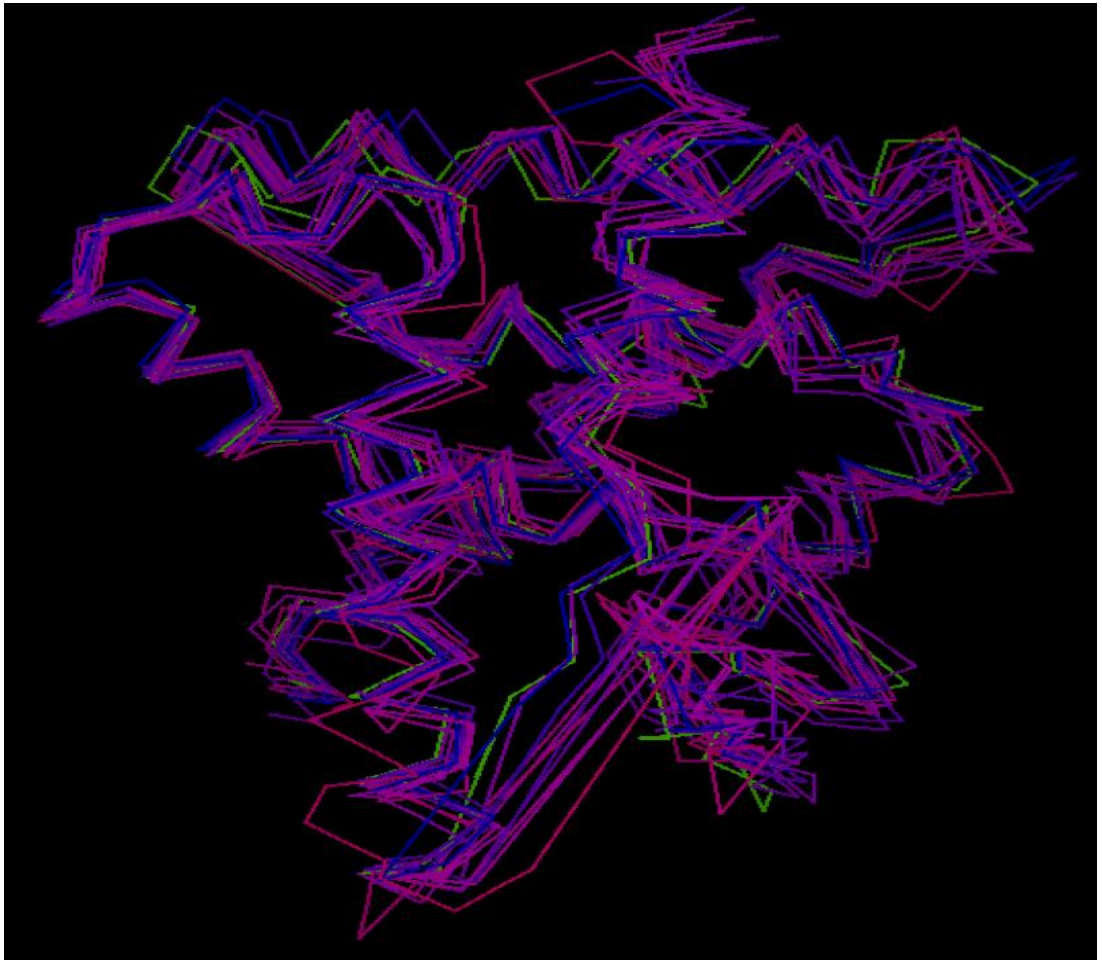
N



Στοίχιση βασισμένη στη δομή

Στις στοιχίσεις πάντα θα πρέπει να λαμβάνουμε υπόψη την δομή

Η στοίχιση 2 πρωτεϊνών συγκρίνει ταυτόχρονα και την 3D δομή τους



Στοίχιση βασισμένη στη δομή

Συντηρημένα αμινοξέα – λειτουργικότητα πρωτεϊνών

Υπάρχουν συντηρημένες και μη συντηρημένες περιοχές

Χρησιμοποιώντας αυτή τη γνώση, μπορούμε να κάνουμε στοίχιση αλληλουχιών με πολύ χαμηλή ταυτότητα

```
QLIPPLINLLMSIEPDVIYA LLTSLNQLGERQLLSVVKWSKSLPGFRNLHIDDQITLIQYSWMSLMVFGLGW
DEEWELIKTVTEAHVATNAQ AFSHFTKIITPAITRVVDFAKKLPMFCELPCEDQIILLKGCCMEIMSLRAAV
PEVGELIEKVRKAHQETFPA LWDKFSELSTKCIIKTVEFAKQLPGFTTLTIADQITLLKAACLDILILRICT
PQLEELITKVSKAHQETFPS LWDKFSELATKCIIKIVEFAKRLPGFTGLSIADQITLLKAACLDILMLRICT
ANEDMPVERILEAELAVEPK PVTNICQAADKQLFTLVEWAKRIPHFSELPLDDQVILIRAGWNELLIASFSH
EEQRMMIRELMDARMKTFDT LLPHMADMSTYMFKGIISYAKVISYFRDLPIEDQISLRKGAAFELCQLRFNT
ADLKSLAKRIYEAYLKNFNM IFHCCQCTSVETVTELTELAKAIPGFANLDLNDQVTLLKYGVYEAIFAMLSS
KPYNKIVSHLLVAEPEKIYA ALTTLCDLADRELVVIIGWAKHIPGFSTLSLADQMSLIQSAWMEILILGVVY
QLTPTLVSLLEVIEPEVLYA IMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDDQMTLIQYSWMSLMAFALGW
LSPEQLVLTLLEAEPPHVLI MMMSLTKLADKELVHMISWAKKIPGFVELSLFDQVRLIESCWMEVLMMGLMW
SEIDRIAQNIIKSHLETCQY LWQQCAIQITHAIQYVVEFAKRITGFMELCQNDQILLLKSGCLEVVLVRMCR
PEEWDLIHIATEAHRSTNAK AFSEFTKIITPAITRVVDFAKKLPMFSELPCEDQIILLKGCCMEIMSLRAAV
```

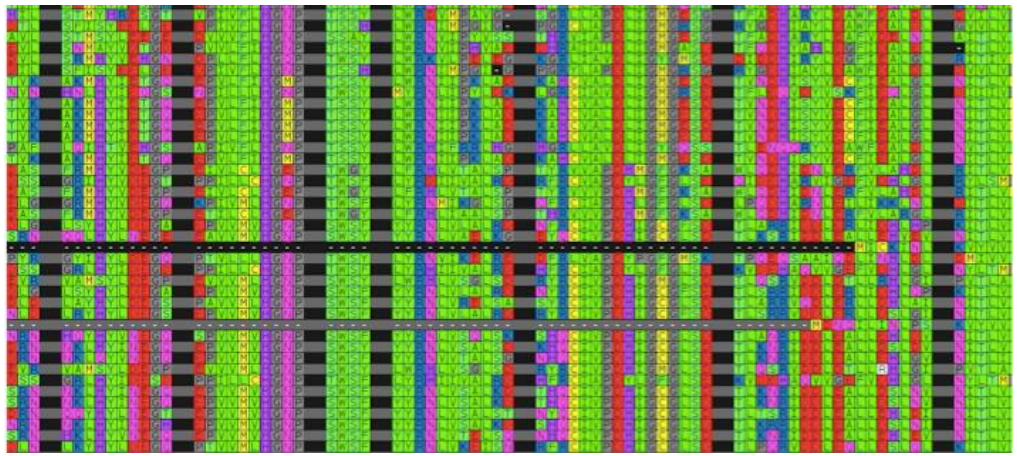
Στοίχιση βασισμένη στη δομή

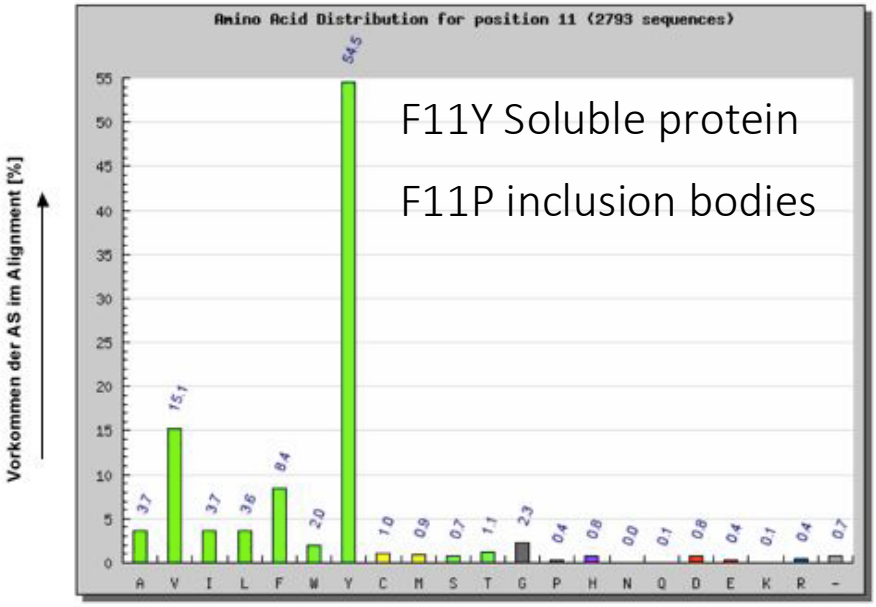
Μας βοηθάει στον ήμι-ορθολογικό σχεδιασμό πρωτεϊνών

Μαθαίνουμε από την φυσική εξέλιξη των πρωτεϊνών

Μπορούμε να μεταφέρουμε μεταλλάξεις μεταξύ ομόλογων πρωτεϊνών

Μπορούμε να παρατηρήσουμε δίκτυα συσχετισμένων θέσεων





Βιοπληροφορικά εργαλεία για ανάλυση πρωτεϊνών

Βιοπληροφορικά εργαλεία για πρωτεΐνες

Μελέτες σχέσης δομής και λειτουργίας

Ανάπτυξη μοντέλων νέων πρωτεϊνών ή μεταλλαγμάτων

- πρόβλεψη / κατανόηση καταλυτικής συμπεριφοράς

- ορθολογικός σχεδιασμός πρωτεϊνών

Μοντέλα ομολογίας

Ανάλυση docking

- docking μικρών μορίων στις πρωτεΐνες

Ειδικά εργαλεία για την καθοδήγηση της εξέλιξης πρωτεϊνών

- βάσεις δεδομένων με δεδομένα στοίχισης με βάση τη δομή

NCBI

National Center for Biotechnology Information

https://www.ncbi.nlm.nih.gov

- Κύρια βάση δεδομένων για αλληλουχίες (γονίδια, γονιδιώματα, πρωτεΐνες)
- Κατάθεση δημοσιευμένων αλληλουχιών
- Διασύνδεση με τη βιβλιογραφία (**PubChem**) και εργαλεία (**BLAST**)
- Μοναδικός κωδικός για κάθε αλληλουχία
- Πηγή εύρεσης νέων γονιδίων/πρωτεϊνών από γνωστά μοτίβα

Protein Data Bank

http://www.rcsb.org/pdb

- Κύρια βάση δεδομένων για 3D δομές πρωτεϊνών
- Περίπου 170.000 δομές (κατάθεση ~10.000 / έτος)
- Πηγή pdb αρχείων - βασική μορφή αρχείου για τέτοια δεδομένα
- Σύνδεση με βιβλιογραφία
- Δεδομένα ποιότητας της δομής
- 3D viewer
- Δομικές πληροφορίες και πληροφορίες για την αλληλουχία
- Εύρεση παρόμοιων πρωτεϊνών
- Κωδικός πάντα με 4 ψηφία!

ExPASy

Expert **P**rotein **A**nalysis **Sy**stem

https://www.expasy.org

Μία τεράστια βιβλιοθήκη βιοπληροφορικών εργαλείων για πρωτεΐνες και DNA

Στα πρωτεωμικά εργαλεία:

- **UniProt**: Λειτουργικές πληροφορίες ενζύμων
- **ENZYME**: Κατηγοριοποίηση ενζύμων
- **Biochemical pathways**
- **Peptide cutter**: προβλέπει θέσεις κοπής
- **Phyre2**: Μοντέλα ομολογίας

Other databases

UniProt: https://www.uniprot.org

Πληροφορίες σχετικά με την λειτουργία των πρωτεϊνών

Διαφορετικοί κωδικοί από την NCBI!

Δίνει σημαντικά σημεία στην αλληλουχία

KEGG: https://www.genome.jp/kegg/

Κυρίως δίνει δεδομένα σε σχέση με συγκεκριμένα μεταβολικά μονοπάτια

BRENDA: https://www.brenda-enzymes.org

Δομικά και λειτουργικά δεδομένα

Συσχέτιση και με την SCOP (Structural Classification of Proteins)

Εύρεση πρωτεϊνών με παρόμοια δομή

MolProbity

http://molprobity.biochem.duke.edu

- Έλεγχος ποιότητας της δομής
- Μπορεί να σχεδιάσει διάγραμμα Ramachandran
- Δίνει λίστα με δομικά προβλήματα όπως στερικές παρεμποδίσεις, περίεργες γωνίες δεσμών κτλ
- Μπορεί να στρίψει πλευρικές ομάδες με ασυμμετρία (His, Asn, Gln)
- Πολύ σημαντικό εργαλείο, ειδικά για την μελέτη ποιότητας μοντέλων ομολογίας

Protein viewer

Pymol 0.99 (newer versions are not freeware)

Το λογισμικό που χρησιμοποιούν οι περισσότεροι

Πλήρης ρύθμιση – πολλά plug-in για να δουλέψεις

Chimera

Freeware – χρησιμοποιείται αρκετά τα τελευταία χρόνια

YASARA

Freeware ως viewer

Πέρα της επισκόπησης της δομής, μπορεί να χρησιμοποιηθεί και για molecular docking, δημιουργία μοντέλων ομολογίας κτλ

Bio-prodict και 3DM

https://www.bio-prodict.com

Η 3DM είναι μία πλατφόρμα που συλλέγει, αποθηκεύει, συνδέει και οπτικοποιεί δεδομένα από μία υπεροικογένεια και χρησιμοποιεί μοναδικό αριθμό θέσης για τις αλληλουχίες.

1. *Αποδελτίωση δεδομένων από βιβλιογραφία*
2. *Κατηγοριοποίηση σε μία υπο-οικογένεια, γύρω από μία ομόλογη PDB δομή*
3. *Δημιουργία στοίχισης βασισμένης στη δομή για όλες τις αλληλουχίες*
4. *Δημιουργία δεδομένων από την μετα-ανάλυσης, όπως συσχετισμός θέσεων, συντήριση δεδομένων, μοντέλα ομολογίας κτλ*
5. *Οπτικοποίηση στο YASARA*

Βιβλιογραφία

- Attwood and Parry-Smith (1999) Introduction to bioinformatics. Adison Wesley Longman
- Misener and Krawetz (2000) Methods and protocols. Humana press
- Higgins and Taylor (2000) Bioinformatics: Sequence, structure and databanks. Oxford University Press