

A. ΑΝΑΖΗΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΕ ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Η αναζήτηση και εύρεση κατάλληλης επιστημονικής βιβλιογραφίας αποτελεί το πρώτο και κυριότερο στάδιο για την προσέγγιση και επίλυση ερευνητικών προβλημάτων. Η βιβλιογραφική αναζήτηση πραγματοποιείται με πρόσβαση σε βάσεις δεδομένων χρησιμοποιώντας κατάλληλες λέξεις-κλειδιά. Το αποτέλεσμα και η συνάφεια των αποτελεσμάτων εξαρτάται από την επιλογή των λέξεων καθώς και τη θέση τους στο κείμενο (τίτλος, περίληψη, κυρίως σώμα). Στην πλειοψηφία των βάσεων δίνεται η δυνατότητα εφαρμογής φίλτρων ώστε να περιοριστούν τα αποτελέσματα της αναζήτησης (ημερομηνία, ονόματα συγγραφέων κ.ά.).

- Η **PubMed** περιλαμβάνει περισσότερες από 21×10⁶ βιβλιογραφικές αναφορές από τη **MEDLINE** (βιβλιογραφική βάση δεδομένων της U.S. National Library of Medicine, **NLM**), από περιοδικά επιστημών υγείας, χημείας και βιοτεχνολογίας, καθώς και βιβλία. Δημιουργήθηκε και παρέχεται δωρεάν από το National Center for Biotechnology Information (**NCBI**). Οι εγγραφές περιλαμβάνουν τις περιλήψεις των άρθρων, αλλά και συνδέσμους στο πλήρες κείμενο των άρθρων, στους ιστότοπους των εκδοτών και σε άλλες βάσεις μοριακών δεδομένων του **NCBI**. Το περιεχόμενο των άρθρων περιγράφεται από τα **MeSH Terms**, ένα ελεγχόμενο λεξιλόγιο βιοϊατρικών όρων της **NLM**. Ο χρήστης της **PubMed** έχει τη δυνατότητα να επιλέξει τον τρόπο εμφάνισης των αποτελεσμάτων, να εφαρμόσει φίλτρα προκειμένου να περιορίσει τα αποτελέσματα της αναζήτησης και να ανακτήσει σχετικές πληροφορίες από άλλες βάσεις δεδομένων. Παράλληλα, χρησιμοποιώντας την επιλογή **Advanced search**, μπορεί να συνδυάσει ποικίλα κριτήρια ώστε να προβεί σε σύνθετες αναζητήσεις.
- Το **ScienceDirect** είναι μια βάση δεδομένων επιστημονικών άρθρων από 2500 περιοδικά και 11000 βιβλία. Αποτελεί τμήμα του εκδοτικού οίκου **Elsevier**, που δραστηριοποιείται στους τομείς της επιστήμης, της τεχνολογίας και της ιατρικής. Ο οίκος **Elsevier** έχει ψηφιοποιήσει πληθώρα περιοδικών που είχαν εκδοθεί πριν το 1995, επιτρέποντας πρόσβαση μέσω του Η/Υ σε άρθρα από το 1823 (**The Lancet**). Το **ScienceDirect** παρέχει τη δυνατότητα απλής και σύνθετης αναζήτησης με χρήση λογικών τελεστών, την εμφάνιση των άρθρων με τις περισσότερες αναφορές, τη μαζική ανάκτηση πολλών άρθρων, την ανάκτηση συμπληρωματικού υλικού όπως αρχείων ήχου και video, καθώς

και την πρόσβαση στο πλήρες κείμενο από μη πιστοποιημένες διευθύνσεις με τη χρήση κωδικών πρόσβασης.

ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΑΣΚΗΣΗΣ

Αναζήτηση στην PubMed.

- ✓ Επισκεφθείτε την ιστοσελίδα του NCBI (<http://www.ncbi.nlm.nih.gov/>).
- ✓ Στο πεδίο *All Databases* επιλέξτε **MeSH**, αναζητήστε και καταγράψτε τον ορισμό για τις πρωτεΐνες **odorant receptors**.
- ✓ Στο *Related information*, επιλέξτε **PubMed** ώστε να μεταβείτε στην σελίδα της PubMed που περιλαμβάνει βιβλιογραφία σχετική με τις πρωτεΐνες **odorant receptors**. Πόσα άρθρα βρήκατε;
- ✓ Χρησιμοποιώντας τις επιλογές **Items per page** και **Sort by** από το *Display Settings* (Σχήμα 1.1) εμφανίστε στην οθόνη **50** αποτελέσματα ταξινομημένα κατά **ημερομηνία δημοσίευσης** (*Pub Date*).
- ✓ Περιορίστε τα αποτελέσματα της αναζήτησης σε όσα άρθρα έχουν ελεύθερα προσβάσιμο (*text availability*) το πλήρες κείμενό τους (**Free full text available**) κάνοντας χρήση του αντίστοιχου φίλτρου. Πόσα άρθρα βρήκατε; Καταγράψτε τις **5 πρώτες βιβλιογραφικές αναφορές** (*συγγραφείς, τίτλο, περιοδικό, έτος, σελίδες*).
- ✓ Στο πάνω τμήμα της σελίδας επιλέξτε *Advanced*, ώστε να προβείτε σε μία σύνθετη αναζήτηση. Διαγράψτε τα φίλτρα που έχετε εφαρμόσει (**clear all**).
- ✓ Αναζητήστε άρθρα που έχουν στον τίτλο ή στην περίληψή τους (**Title/Abstract**) τις λέξεις - κλειδιά **odorant receptors** και **anopheles** και έχουν δημοσιευθεί από το **2000** και μετά (*Date - Publication*). Πόσα άρθρα βρήκατε;
- ✓ Στο πεδίο *Find related data* επιλέξτε **Nucleotide**, ώστε να αναζητήσετε εγγραφές της βάσης δεδομένων GenBank που αναφέρονται στα άρθρα που βρήκατε. Καταγράψτε τα **Accession Numbers** των **5** πρώτων εγγραφών και αποθηκεύστε τις ακολουθίες τους σε μορφή *FASTA*.
- ✓ Επιστρέψτε στην προηγούμενη σελίδα και περιορίστε τα αποτελέσματα της αναζήτησης σε όσα άρθρα έχουν **ελεύθερα προσβάσιμο το πλήρες κείμενό τους** (*Free full text available*) κάνοντας χρήση του αντίστοιχου φίλτρου. Επιλέξτε ένα από τα άρθρα και ακολουθώντας τους συνδέσμους μεταβείτε στο πλήρες κείμενο. Καταγράψτε τη βιβλιογραφική αναφορά και την πρώτη πρόταση του κυρίως κειμένου.

B. ΑΝΑΖΗΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΕ ΒΙΟΛΟΓΙΚΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Μια βιολογική βάση δεδομένων χρησιμοποιείται για την **οργάνωση, αποθήκευση, επεξεργασία, αναζήτηση** και **ανάκτηση** της βιολογικής πληροφορίας. Τα δεδομένα τα οποία συναντώνται στις βιολογικές βάσεις είναι αυτά που παράγονται από τη βιολογική έρευνα (κυρίως της μοριακής βιολογίας). Ενδεικτικοί τύποι δεδομένων:

- νουκλεοτιδικές ακολουθίες (DNA και mRNA) και ολόκληρα γονιδιώματα
- πρωτεϊνικές ακολουθίες
- 3-D δομές πρωτεϊνών και νουκλεϊνικών οξέων
- δεδομένα γονιδιακής έκφρασης
- δεδομένα γενετικής ποικιλότητας (πολυμορφισμοί)

Πληροφορίες σχετικά με τις διαθέσιμες **BB** καθώς και εργαλεία λογισμικού μπορούν να αναζητηθούν σε εξειδικευμένα επιστημονικά περιοδικά ή ιστότοπους, όπως:

- Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection (<http://nar.oxfordjournals.org/>)
- Database: The Journal of Biological Databases and Curation (<http://database.oxfordjournals.org/>)
- NCBI Resource Guide (<http://www.ncbi.nlm.nih.gov/guide/all/>)

Υπάρχουν πολλές εξειδικευμένες βάσεις δεδομένων, ωστόσο στο πλαίσιο της συγκεκριμένης άσκησης θα επικεντρωθούμε στις βασικές κατηγορίες και στις πιο αντιπροσωπευτικές κάθε κατηγορίας, που είναι ελεύθερα προσβάσιμες στην επιστημονική κοινότητα.

✓ **Βάσεις νουκλεοτιδικών δεδομένων**

Οι βάσεις δεδομένων νουκλεοτιδικών ακολουθιών καθιστούν ελεύθερα διαθέσιμα, στην επιστημονική κοινότητα, ετερογενή στοιχεία που ποικίλλουν όσον αφορά στην προέλευση του υλικού (π.χ. γονιδίωμα έναντι cDNA), στην ποιότητά του, στην έκταση του σχολιασμού και στην πληρότητα της ακολουθίας σχετικά με το βιολογικό στόχο (π.χ. πλήρους έναντι μερικής κάλυψης ενός γονιδίου ή ενός γονιδιώματος).

Οι τρεις μεγαλύτερες βάσεις νουκλεοτιδικών δεδομένων που είναι ελεύθερα διαθέσιμες είναι οι:

- DNA Data Bank of Japan (**DDBJ**) στο Center for Information Biology (**CIB**) (<http://www.ddbj.nig.ac.jp/>)
- **GenBank** στο National Center for Biotechnology Information (**NCBI**)

(<http://www.ncbi.nlm.nih.gov/genbank/>)

- **EMBL_Bank** στο European Bioinformatics Institute (**EBI**)

(<http://www.ebi.ac.uk/embl/index.html>).

Οι βάσεις αυτές σε συνεργασία έχουν δημιουργήσει την **International Nucleotide Sequence Database Collaboration** (<http://www.insdc.org/>). Η συνεργασία μεταξύ τους περιλαμβάνει τη δημιουργία κοινών κανόνων για την ταξινόμηση και το χαρακτηρισμό των δεδομένων. Μια τυπική εγγραφή στη μορφοποίηση της **GenBank** αποτελείται από γραμμές. Η πρώτη λέξη κάθε γραμμής υποδηλώνει το είδος της πληροφορίας που ακολουθεί (παράρτημα Α). Κάθε εγγραφή τελειώνει με τους χαρακτήρες “//”. Η νουκλεοτιδική ακολουθία μπορεί να ανακτηθεί στη μορφοποίηση **FASTA** (Σχήμα 1). Η πρώτη γραμμή ξεκινά με το χαρακτήρα “>” και δίνει μια σύντομη περιγραφή της ακολουθίας, η οποία εμφανίζεται στις επόμενες γραμμές.

>gi|209751317|gb|EU891385.1| Escherichia coli regulator for SOS regulon (ECs5026) gene, complete cds

```
ATGAAAGCGTTAACGGCCAGGCAACAAGAGGTGTTTGATCTCATCCGTGATCACATC
AGCCAGACAGGTA
TGCCGCCGACGCGTGC GGAAATCGCGCAGCGTTTGGGGTTCCGTTCCCCAAACGCG
GCTGAAGAACATCT
GAAGGCGCTGGCACGCAAAGGCGTAATTGAAATTGTTTCCGGCGCATCACGCGGGA
TTCGTCTGTTGCAG
GAAGAGGAAGAAGGGTTGCCACTGGTAGGTCGTGTGGCTGCCGGTGAACCGCTTCT
GGCGCAACAGCATA
TTGAAGGTCATTATCAGGTCGATCCTTCCTTGTTCAAGCCGAATGCTGATTCCTGCT
GCGCGTCAGCGG
GATGTTCGATGAAAGATATCGGCATTATGGATGGCGACTTGCTGGCAGTGCATAAAA
CTCAGGATGTACGT
AACGGTCAGGTCGTTGTCGCACGTATTGATGACGAAGTTACCGTTAAGCGCCTGAAA
AAACAGGGCAATA
AAGTCGAACCTGTTGCCAGAAAATAGCGAGTTTAAACCAATTGTCGTTGACCTTCGTC
AGCAGAGCTTCAC
CATTGAAGGGCTGGCGGTTGGCGTTATTCGCAACGGCGACTGGCTGTAA
```

Σχήμα 1. Νουκλεοτιδική ακολουθία σε μορφοποίηση FASTA.

✓ Βάσεις πρωτεϊνικών δεδομένων

Η **UniProt** (<http://www.uniprot.org/>) αποτελεί την περιεκτικότερη παγκόσμια συλλογή πληροφοριών για **πρωτεΐνες** (ακολουθία και λειτουργία), η οποία προέκυψε από τη συνεργασία των

- **Swiss-Prot**, που παρέχει ένα υψηλό επίπεδο σχολιασμού (όπως περιγραφή της λειτουργίας μιας πρωτεΐνης, των μετα-μεταφραστικών τροποποιήσεων, κ.λ.π.), ένα ελάχιστο επίπεδο πλεονασμού και υψηλό επίπεδο ολοκλήρωσης – διασύνδεσης με άλλες **BB**.
- **TrEMBL**, που αποτελεί το συμπλήρωμα της Swiss-Prot. Περιλαμβάνει εγγραφές που δεν έχουν ακόμα ενσωματωθεί στη Swiss-Prot και έχουν προκύψει από τη μετάφραση των καταχωρημένων νουκλεοτιδικών εγγραφών της EMBL. Ο σχολιασμός της γίνεται αυτοματοποιημένα.
- **PIR**, που προέκυψε από τον Atlas of Protein Sequence and Structure (1965-1978) της Margaret Dayhoff και εξελίχθηκε σε μια ολοκληρωμένη πηγή δεδομένων και αναλυτικών εργαλείων.

✓ Βάσεις πρωτεϊνικών οικογενειών και domains

Οι βάσεις **πρωτεϊνικών οικογενειών** και **domains** προέκυψαν από την ανάλυση των πρωτογενών βάσεων πρωτεϊνικών ακολουθιών. Παρέχουν πληροφορίες σχετικά με περιοχές των πρωτεϊνών που έχουν καλά συντηρημένη ακολουθία και συγκεκριμένη λειτουργία/δομή, και έχουν εξελιχθεί σε σημαντικά εργαλεία για την εύρεση απομακρυσμένων σχέσεων σε νέες ακολουθίες, καθώς και την εξαγωγή συμπερασμάτων για την πρωτεϊνική λειτουργία. Διαφέρουν μεταξύ τους ως προς τις μεθόδους ανάλυσης των πρωτεϊνικών ακολουθιών, όπως είναι οι κανονικές εκφράσεις (**regular expressions**), τα προφίλ (**profiles**) και τα μοντέλα HMMs (**Hidden Markov Models**). Κάποιες από τις βάσεις αυτές είναι:

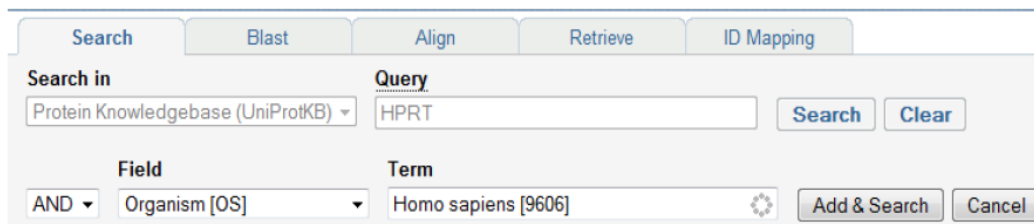
- **PROSITE** (<http://prosite.expasy.org/>)
- **PFAM** (<http://pfam.sanger.ac.uk/>)
- **PRINTS** (<http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/>)
- **PRODOM** (<http://prodom.prabi.fr/>)
- **SMART** (<http://smart.embl-heidelberg.de/>)

Η **InterPro** (<http://www.ebi.ac.uk/interpro/index.html>) είναι το αποτέλεσμα της ολοκλήρωσης των προαναφερθέντων **BB**, με στόχο την περιεκτικότερη εποπτεία των διαθέσιμων πόρων. Το πρόγραμμα **InterProScan** επιτρέπει την αναζήτηση των πρωτεϊνικών μοτίβων σε μια ακολουθία.

ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΑΣΚΗΣΗΣ

A. Αναζήτηση δεδομένων στην UniProt

- ✓ Μεταβείτε στην ιστοσελίδα της **UniProt** (<http://www.uniprot.org/>) και αναζητήστε πληροφορίες σχετικά με την πρωτεΐνη “μεταφορά της φωσφοριβοσυλιωμένης γουανίνης” (Hydroxanthineguanine phosphoribosyltransferase ή **hprt** ή **hgprt**). Εισάγετε τον όρο **hprt** στο πεδίο αναζήτησης, πατήστε **Advanced Search** και στο πεδίο **Organism** εισάγετε τον όρο **homo sapiens** (Σχήμα 2) . Από τα αποτελέσματα επιλέξτε την εγγραφή με κωδικό P00492.



The screenshot shows the UniProt search interface. At the top, there are tabs for 'Search', 'Blast', 'Align', 'Retrieve', and 'ID Mapping'. The 'Search' tab is active. Below the tabs, there are two main search sections. The first section is labeled 'Search in' and 'Query'. The 'Search in' dropdown is set to 'Protein Knowledgebase (UniProtKB)'. The 'Query' input field contains 'HPRT'. There are 'Search' and 'Clear' buttons to the right. The second section is labeled 'Field' and 'Term'. The 'Field' dropdown is set to 'AND' and the 'Term' dropdown is set to 'Organism [OS]'. The 'Term' input field contains 'Homo sapiens [9606]'. There are 'Add & Search' and 'Cancel' buttons to the right.

1. Καταγράψτε το **μήκος της αμινοξικής ακολουθίας (Sequence length)**, τη **λειτουργία (Function)** και τον **ενδοκοιτταριο εντοπισμό (Subcellular location)** της πρωτεΐνης.
2. Με ποια **ασθένεια** σχετίζεται η συγκεκριμένη πρωτεΐνη (Involvement in disease); Καταγράψτε μία **μετάλλαξη** που συνδέεται με την εν λόγω ασθένεια (**Natural variations**).
3. Ακολουθήστε τον υπερσύνδεσμο στην **PROSITE** και καταγράψτε το **μοτίβο (Consensus pattern)** που χαρακτηρίζει την οικογένεια ενζύμων στην οποία ανήκει η **hprt**.
4. Επιστρέψτε στην σελίδα της **UniProt** και ακολουθήστε τον υπερσύνδεσμο στην **KEGG**.

Με ποια μεταβολικά μονοπάτια (**Pathway**) σχετίζεται το συγκεκριμένο ένζυμο;

B. Αναζήτηση δεδομένων στο Entrez

- ✓ Επισκεφθείτε την ιστοσελίδα του NCBI (<http://www.ncbi.nlm.nih.gov/ncbisearch/>) και εντοπίστε τη γενωμική ακολουθία του γονιδίου **cystic fibrosis transmembrane conductance regulator (CFTR)**.
- ✓ Επιλέξτε τη **BB Nucleotide** και στο πεδίο **Search** χρησιμοποιήστε την αναζήτηση: **CFTR[Title]**

- ✓ Με τη βοήθεια των επιλογών στο δεξιό μέρος της οθόνης, περιορίστε τα αποτελέσματά σας στις εγγραφές της **RefSeq (Reference Sequence collection)**, που παρέχει ένα περιεκτικό και καλά σχολιασμένο σύνολο ακολουθιών, και στον οργανισμό **Homo Sapiens**. Επιλέξτε την εγγραφή με κωδικό *Accession: NG_016465.3*
 - Ποιο είναι το μήκος της ακολουθίας (**LOCUS**);
 - Σε ποιο χρωμόσωμα και σε ποια θέση βρίσκεται το γονίδιο (**FEATURES _ source _ /map**);
 - Ποιος είναι ο αριθμός των εξονίων και ποιες οι 10 πρώτες βάσεις της γενωμικής ακολουθίας;

- ✓ Επιλέξτε τον υπερσύνδεσμο **Reference sequence information _ RefSeq mRNA** και μεταβείτε στην αντίστοιχη εγγραφή.
 - Ποιος είναι ο κωδικός αριθμός της εγγραφής και ποιο το μήκος της ακολουθίας (**LOCUS**);
 - Ποιες είναι οι 10 πρώτες βάσεις της mRNA ακολουθίας;
Χρησιμοποιώντας τον υπερσύνδεσμο **Reference sequence information _ RefSeq protein product**, καταγράψτε το μήκος της πρωτεΐνης και τα 10 πρώτα αμινοξέα.
 - Στη συνέχεια, μεταβείτε στην **OMIM**, μελετήστε την εγγραφή για το γονίδιο **CFTR** με κωδικό ***602421** και καταγράψτε τις ασθένειες με τις οποίες συνδέεται το συγκεκριμένο γονίδιο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- 1.** Bioinformatics A Practical Guide to the Analysis of Genes and Proteins. A. D. Baxevanis, B.F.F. Ouellette, Wiley Interscience, **2001**.
- 2.** Advanced Data Mining Technologies in Bioinformatics. H.-H. Hsu, Idea Group Publishing, **2006**.
- 3.** Bioinformatics Algorithms Techniques and Applications. I. I. Măndoiu, A. Zelikovsky, John Wiley & Sons, Inc., **2001**.