ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

BIO 230 - Εισαγωγή στην Υπολογιστική Βιολογία

Φυλλάδιο Εργαστηριακών Ασκήσεων και Φροντιστηρίων

Βασίλης Ι. Προμπονάς

Λευκωσία 2015-2019

1η Εργαστηριακή Άσκηση

Διαδικτυακές βιβλιογραφικές πηγές (Μοριακής) Βιολογίας και Βιοπληροφορικής

Στόχοι άσκησης

1. Εξοικείωση των φοιτητών με έννοιες από το πεδίο της Ανάκτησης Πληροφορίας (π.χ. Boolean αναζήτηση, ακρίβεια–precision, ανάκληση–recall, αστοχία–fall-out/false positive rate).
2. Εξοικείωση των φοιτητών με τις βασικές λειτουργίες διαθέσιμες στη βάση δεδομένων PubMed®.
3. Αυτόνομη διερεύνηση από τους φοιτητές άλλων σχετικών διαδικτυακών πηγών (π.χ. arXiv.org, bioRxiv, CiteSeerX, Google Scholar, Scopus), των κύριων μοντέλων δημοσίευσης (π.χ. ανοικτής πρόσβασης) και του ηλεκτρονικού καταλόγου της βιβλιοθήκης του Πανεπιστημίου Κύπρου.

ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ

Ανάκτηση Πληροφορίας – βασικά στοιχεία

Ορισμός:

Με τον όρο Ανάκτηση Πληροφορίας (Information Retrieval – IR) αναφερόμαστε στην "*εύρεση υλικού* (συνήθως εγγράφων) το οποίο είναι εν γένει *μη δομημένο* (συνήθως σε μορφή κειμένου[1]) που ικανοποιεί μια *ανάγκη πληροφόρησης* μέσα από *μεγάλες συλλογές* (συνήθως σε ψηφιακή μορφή)" (Manning *et al*., 2008).

Παραδείγματα:

1. Αναζήτηση ιστοσελίδων του διαδικτύου που αφορούν ένα συγκεκριμένο αντικείμενο (π.χ. τη Βιοπληροφορική).
2. Αναζήτηση στον υπολογιστή μας για αρχεία που περιέχουν πληροφορίες για ένα συγκεκριμένο μάθημα (π.χ. BIO003).

[1]Το πεδίο της Ανάκτησης Πληροφορίας παραδοσιακά αναπτύχθηκε για την εύρεση υλικού σε απλά αρχεία κειμένου, έχει όμως επεκταθεί κατάλληλα για την εύρεση υλικού άλλων τύπων (π.χ. εικόνες, video). Τα αρχεία κειμένου, παρότι μπορούν να θεωρηθούν εγγενώς μη δομημένα, ακολουθούν τις δομές της ανθρώπινης (φυσικής) γλώσσας.

3. Αναζήτηση στα μηνύματα του ηλεκτρονικού μας ταχυδρομείου για να εντοπίσουμε ανακοινώσεις του μαθήματος BIO003.

Θα δώσουμε μερικά παραδείγματα εφαρμογών χρησιμοποιώντας την παρακάτω (απλοϊκή) συλλογή κειμένων (μη δομημένα έγγραφα):

K1:	Η Βιοπληροφορική είναι ένα γνωστικό πεδίο που σχετίζεται με τη Βιολογία και την Πληροφορική.
K2:	Τα πεδία της Γονιδιωματικής, και Επιγονιδιωματικής βασίζονται στη χρήση Βιοπληροφορικών τεχνικών για την ανάλυση μεγάλων όγκων δεδομένων.
K3:	Στο μάθημα BIO001 δεν διδάσκονται αντικείμενα σχετικά με τη Βιοπληροφορική.
K4:	Ο Γιαννής και ο Κωστής είναι αδέλφια.
K5:	Αυτό το εξάμηνο παρακολουθώ το μάθημα BIO003 γιατί με ενδιαφέρει η Βιοπληροφορική, κλάδος που καταπιάνεται με την διαχείριση και ανάλυση βιολογικών δεδομένων.

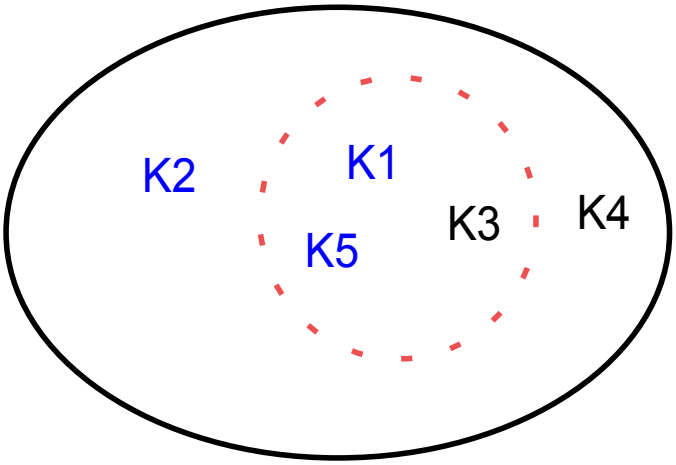
Έστω ότι έχουμε την ανάγκη να συλλέξουμε πληροφορίες σχετικά με το αντικείμενο της Βιοπληροφορικής, χρησιμοποιώντας την παραπάνω συλλογή κειμένων. Η λογική στρατηγική που θα ακολουθούσαμε είναι να επιλέξουμε από τη συλλογή τα κείμενα που αφορούν το αντικείμενο της Βιοπληροφορικής, και στη συνέχεια να συλλέξουμε τις πληροφορίες που χρειαζόμαστε από αυτά.

Ο απλούστερος τρόπος να επιλέξουμε τα κείμενα που είναι σχετικά με το αντικείμενο που μας ενδιαφέρει είναι με το να πραγματοποιήσουμε μια αναζήτηση με μία ή περισσότερες *λέξεις-κλειδιά* ή αλλοιώς *όρους* (keywords, terms) [2].

Στο παραπάνω παράδειγμα, ένα κατάλληλο keyword αναζήτησης θα ήταν ο όρος "Βιοπληροφορική". Προφανώς, ένα ερώτημα (query) με τον όρο "Βιοπληροφορική" μας επιστρέφει τα κείμενα εκείνα στα οποία εμφανίζεται ο όρος, στο παράδειγμά μας τα K1, K3, K5.

[2]Στην πράξη χρησιμοποιούνται και άλλα "πλαίσια" αναπαράστασης των κειμένων που μας επιτρέπουν περισσότερο ευέλικτες αναζητήσεις. Στα πλαίσια του παρόντος εργαστηρίου μας αρκούν τυπικές αναζητήσεις με keywords σε μη δομημένα (ASCII) κείμενα.

Παρατηρώντας προσεκτικά τα κείμενα της παραπάνω συλλογής, εύκολα αντιλαμβανόμαστε ότι το βέλτιστο υποσύνολο κειμένων για την πληροφοριακή μας ανάγκη είναι το K1, K2, K5. Χρησιμοποιώντας για ευκολία την αναπαράσταση του παρακάτω διαγράμματος



Εικόνα 1: Αναπαράσταση αναζητησης στην απλή συλλογή κειμένων με τον όρο "Βιοπληροφορική". Η κόκκινη διακεκομμένη γραμμή υποδεικνύει το σύνολο των κειμένων που ανακτώνται από την αναζήτηση, ενώ με μπλε χρώμα επισημαίνονται οι "επιθυμητές" απαντήσεις.

αντιλαμβανώμαστε ότι ένα ερώτημα Q μπορεί για διάφορους λόγους να αποτύχει να μας δώσει την επιθυμητή απάντηση. Με βάση την Εικόνα 1 μπορούμε να κατατάξουμε τα αποτελέσματα της αναζήτησής μας σε 4 κατηγορίες:

Αληθώς θετικά (True Positives): είναι τα κείμενα τα οποία **καλώς** ανακτήθηκαν από το ερώτημα Q που πραγματοποιήσαμε (K1 και K5).

Ψευδώς θετικά (False Positives): είναι τα κείμενα τα οποία **κακώς** ανακτήθηκαν από το Q (K3).

Αληθώς Αρνητικά (True Negatives): είναι τα κείμενα τα οποία **καλώς** δεν ανακτήθηκαν από το Q (K4).

Ψευδώς Αρνητικά (False Negatives): είναι τα κείμενα τα οποία **κακώς** δεν ανακτήθηκαν από το Q (K2).[3]

[3]Συχνά, στη σχετική βιβλιογραφία, μπορείτε να δείτε ότι αυτές οι παράμετροι αναπαριστώνται σε μορφή 2x2 πίνακα, ο οποίος ονομάζεται "πίνακας ασάφειας" (confusion matrix).

Αν συμβολίσουμε με TP, FP, TN, FN τα πλήθη των κειμένων στις παραπάνω κατηγορίες, τότε μπορούμε να υπολογίσουμε αριθμητικά μέτρα της αποτελεσματικότητας της παραπάνω διαδικασίας:

ακρίβεια–precision: $$precision = \frac{TP}{(TP + FP)} \quad (1)$$

ανάκληση–recall: $$recall = \frac{TP}{(TP + FN)} \quad (2)$$

αστοχία–fall-out/false positive rate: $$FPR = \frac{FP}{(FP + TN)} \quad (3)$$

Προφανώς, μια ιδανική διαδικασία αναζήτησης έχει *precision=1*, *recall=1*, *FPR=0*.

Ερώτηση για την τάξη

Να υπολογίσετε την ακρίβεια, ανάκληση και αστοχία για την παραπάνω αναζήτηση.

Η παραπάνω αναζήτηση πραγματοποιήθηκε με τη χρήση ενός μόνο keyword. Πιο πολύπλοκα ερωτήματα μπορούν να τεθούν με τη χρήση περισσότερων του ενός keywords και τη χρήση λογικών τελεστών (μοντέλο Boolean ανάκτησης)[4]. Το μοντέλο αυτό βασίζεται στο συνδυασμό των keywords μέσω λογικών τελεστών για την κατασκευή λογικών εκφράσεων της άλγεβρας Boole. Οι λογικοί τελεστές που χρησιμοποιούνται είναι οι τελεστές: AND (σύζευξη), OR (διάζευξη) και NOT (άρνηση). Οι τελεστές AND και OR είναι δυαδικοί (δηλ. έχουν δύο ορίσματα) ενώ ο NOT μοναδιαίος (δηλ. έχει ένα μόνο όρισμα).

Παράδειγμα:

Έστω οι λέξεις–κλειδιά: Βιοπληροφορική, Βιολογία, Πληροφορική. Οι παρακάτω ερωτήσεις

Q1 = Βιοπληροφορική AND Πληροφορική

Q2 = Βιοπληροφορική OR Βιοπληροφορικών

Q3 = NOT Βιοπληροφορική

εάν πραγματοποιηθούν έναντι της παραπάνω απλής συλλογής κειμένων επιστρέφουν αντίστοιχα:

Q1→{K1}

Q2→{K1,K2,K3,K5}

[4]Περισσότερες πληροφορίες για όσους ενδιαφέρονται μπορούν να βρεθούν στο βιβλίο των Manning et al., 2008 ή στις σημειώσεις διαλέξεων Μανωλόπουλου Ι. και Παπαδόπουλου Α.Ν. για το μάθημα "Ανάκτηση πληροφορίας" του Τμήματος Πληροφορικής του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης (http://delab.csd.auth.gr/courses/c_ir/irbook.pdf).

Q3→{K2,K4}

Σύνθετες λογικές εκφράσεις είναι δυνατόν να δημιουργηθούν με συνδυασμό περισσότερων τελεστών, ενώ προσοχή απαιτείται για την ορθή κατασκευή τους με τη χρήση παρενθέσεων που ρυθμίζουν την προτεραιότητα εκτέλεσής τους.

Ερώτηση για το σπίτι

Να παραστήσετε τα αποτελέσματα των αναζητήσεων με τα ερωτήματα Q1, Q2, Q3 όπως το ερώτημα Q στην Εικόνα 1 και να υπολογίσετε την ακρίβεια, ανάκληση και αστοχία τους εάν χρησιμοποιηθούν για τον ίδιο σκοπό.

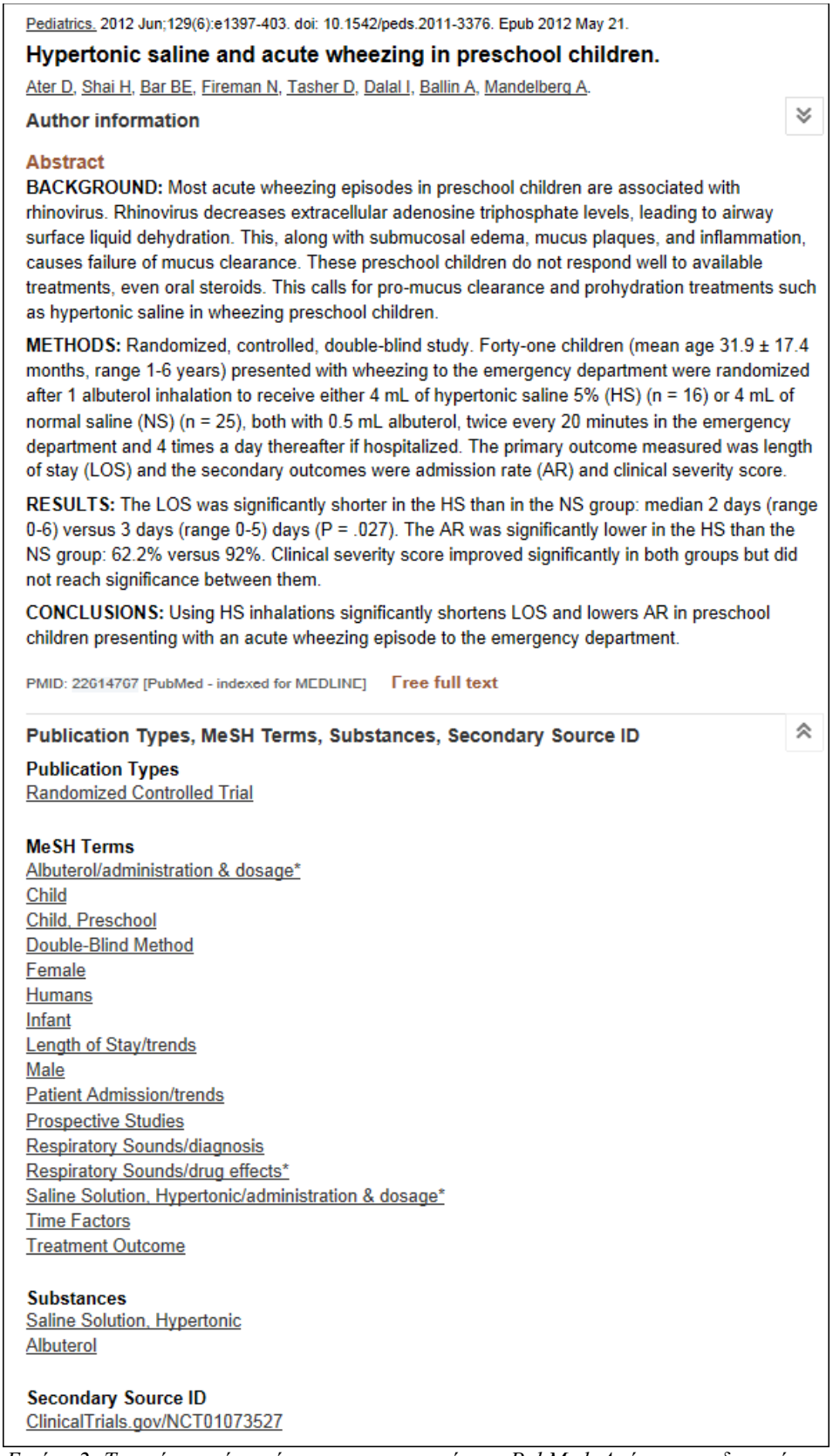
Η βάση δεδομένων PubMed – βασικά στοιχεία

Η βάση δεδομένων PubMed αποτελεί μια online υπηρεσία που προσφέρεται ελεύθερα από την National Library of Medicine (NLM) των ΗΠΑ μέσω του National Center for Biotechnology Information (NCBI) από το διαδικτυακό τόπο http://www.ncbi.nlm.nih.gov/pubmed. Η PubMed αποτελεί το σημείο πρόσβασης στις περιλήψεις (abstracts) βιβλιογραφικών αναφορών (με κύρια προέλευση το ευρετήριο βιβλιογραφικών παραπομπών MEDLINE®) σχετικών με τις βιοϊατρικές επιστήμες (>3 *10^7, Σεπτέμβριος 2019, πηγή http://www.ncbi.nlm.nih.gov/pubmed). Στην πραγματικότητα, η PubMed αποτελεί ένα από τα υποσυστήματα του ENTREZ, ενός μεγάλου ενοποιημένου συστήματος ανάκτησης πληροφοριών από πλήθος ετερογενών βιοϊατρικών βάσεων δεδομένων (NCBI Resource Coordinators, 2015).

Βασικά χαρακτηριστικά μιας εγγραφής PubMed®/MEDLINE®

Κάθε βιβλιογραφική εγγραφή μπορεί να αποτελείται από διάφορα πεδία που αναφέρονται σε διαφορετικούς τύπους πληροφοριών[5] (π.χ. τίτλος του άρθρου, ονόματα συγγραφέων, πηγή – όνομα επιστημονικού περιοδικού). Μια τέτοια εγγραφή φαίνεται στην Εικόνα 2, ενώ μπορείτε να δείτε τη συγκεκριμένη εγγραφή στην PubMed μέσω του συνδέσμου http://www.ncbi.nlm.nih.gov/pubmed/22614767 .

[5]Όπως γίνεται φανερό, οι εγγραφές της PubMed έχουν δομή. Εάν περιορίσουμε την αναζήτησή μας μόνο στις περιλήψεις των άρθρων (φυσική γλώσσα) τότε μπορούμε να θεωρήσουμε ότι ισχύουν όσα αναφέραμε για την ανάκτηση πληροφορίας.



Pediatrics. 2012 Jun;129(6):e1397-403. doi: 10.1542/peds.2011-3376. Epub 2012 May 21.

Hypertonic saline and acute wheezing in preschool children.

Ater D, Shai H, Bar BE, Fireman N, Tasher D, Dalal I, Ballin A, Mandelberg A.

Author information

Abstract

BACKGROUND: Most acute wheezing episodes in preschool children are associated with rhinovirus. Rhinovirus decreases extracellular adenosine triphosphate levels, leading to airway surface liquid dehydration. This, along with submucosal edema, mucus plaques, and inflammation, causes failure of mucus clearance. These preschool children do not respond well to available treatments, even oral steroids. This calls for pro-mucus clearance and prohydration treatments such as hypertonic saline in wheezing preschool children.

METHODS: Randomized, controlled, double-blind study. Forty-one children (mean age 31.9 ± 17.4 months, range 1-6 years) presented with wheezing to the emergency department were randomized after 1 albuterol inhalation to receive either 4 mL of hypertonic saline 5% (HS) (n = 16) or 4 mL of normal saline (NS) (n = 25), both with 0.5 mL albuterol, twice every 20 minutes in the emergency department and 4 times a day thereafter if hospitalized. The primary outcome measured was length of stay (LOS) and the secondary outcomes were admission rate (AR) and clinical severity score.

RESULTS: The LOS was significantly shorter in the HS than in the NS group: median 2 days (range 0-6) versus 3 days (range 0-5) days (P = .027). The AR was significantly lower in the HS than the NS group: 62.2% versus 92%. Clinical severity score improved significantly in both groups but did not reach significance between them.

CONCLUSIONS: Using HS inhalations significantly shortens LOS and lowers AR in preschool children presenting with an acute wheezing episode to the emergency department.

PMID: 22614767 [PubMed - indexed for MEDLINE] Free full text

Publication Types, MeSH Terms, Substances, Secondary Source ID

Publication Types
Randomized Controlled Trial

MeSH Terms
Albuterol/administration & dosage*
Child
Child, Preschool
Double-Blind Method
Female
Humans
Infant
Length of Stay/trends
Male
Patient Admission/trends
Prospective Studies
Respiratory Sounds/diagnosis
Respiratory Sounds/drug effects*
Saline Solution, Hypertonic/administration & dosage*
Time Factors
Treatment Outcome

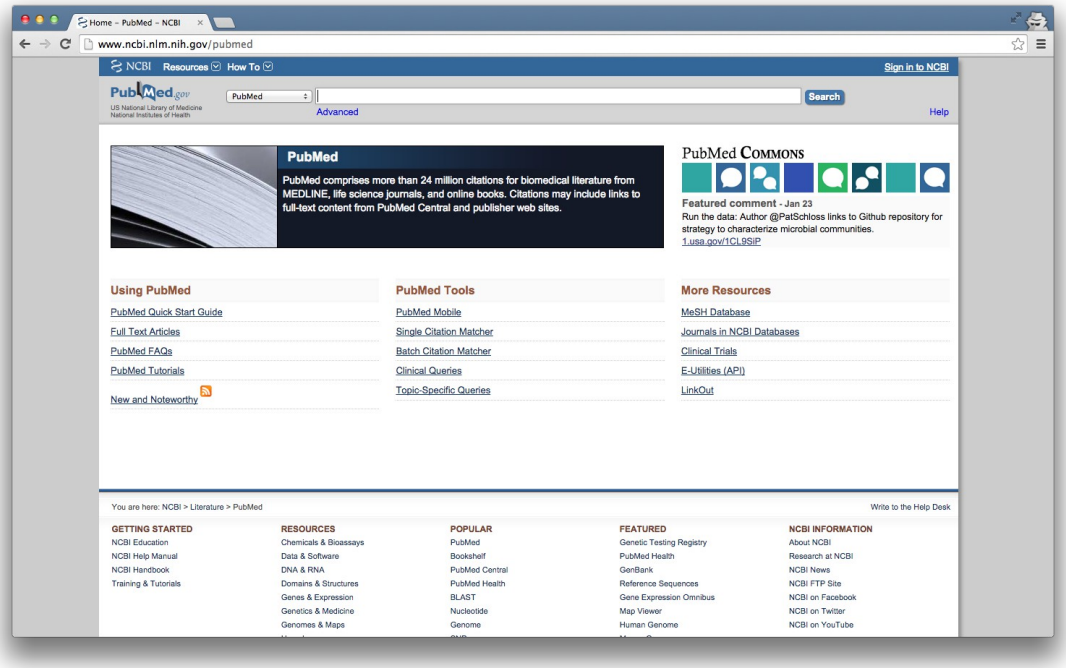
Substances
Saline Solution, Hypertonic
Albuterol

Secondary Source ID
ClinicalTrials.gov/NCT01073527

*Εικόνα 2: Τυπική μορφή εμφάνισης μιας εγγραφής της PubMed. Από το εκπαιδευτικό υλικό της PubMed: http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/010_100.html. Παρατηρήστε, την ύπαρξη ενός **μοναδικού** αριθμού καταχώρησης (PMID: 22614767).*

Βασικές λειτουργίες της PubMed

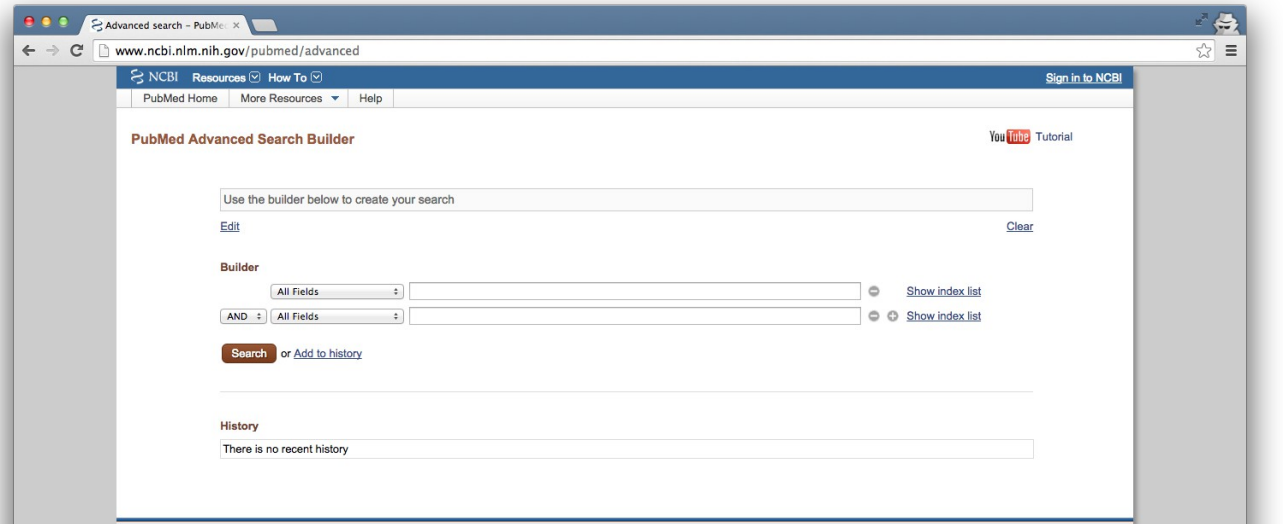
Η κεντρική σελίδα της PubMed (http://www.ncbi.nlm.nih.gov/pubmed) προσφέρει τη δυνατότητα αναζήτησης (μέσω του συστήματος ENTREZ), την πλοήγηση σε σχετικές πηγές δεομένων (π.χ. MeSH database), την πρόσβαση σε εξειδικευμένα εργαλεία αναζήτησης και την παροχή σχετικής βοήθειας προς τους χρήστες.



Εικόνα 3: Η κεντρική σελίδα της PubMed (http://www.ncbi.nlm.nih.gov/pubmed, Ιανουάριος 2015).

Εκτός από την απλή φόρμα αναζήτησης υπάρχει και η δυνατότητα χρήσης λειτουργιών "προχωρημένης" αναζήτησης (Advanced search), όπως φαίνεται στην Εικόνα 4.

Για περισσότερες πληροφορίες μπορείτε να δείτε το σχετικό με την PubMed εκπαιδευτικό υλικό στο URL http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/cover.html.



Εικόνα 4: Η φόρμα "προχωρημένης" αναζήτησης στην PubMed (http://www.ncbi.nlm.nih.gov/pubmed/advanced).

ΠΡΑΚΤΙΚΟ ΜΕΡΟΣ

Ένα πιθανό σενάριο:

Ενδιαφερόμαστε να βρούμε πληροφορίες για τις πρωτεΐνες (protein) που εμπλέκονται στη διαδικασία της μετάφρασης (translation) και έχουν μελετηθεί στον άνθρωπο (human ή "*homo sapiens*") ή στο ποντίκι (mouse ή "*mus musculus*")[6].

Ερώτηση για την τάξη

Ποιό/ά από τα παρακάτω πιστεύετε ότι είναι το πιο κατάλληλο/α ερώτημα/τα προς τη βάση δεδομένων PubMed ώστε να ανακτήσουμε σχετικά έγγραφα;

Q1: (protein OR translation) AND (human OR mouse OR "homo sapiens" OR "mus musculus")
Q2: (protein AND translation) OR (human OR mouse OR "homo sapiens" OR "mus musculus")
Q3: protein AND translation AND (human OR mouse OR "homo sapiens" OR "mus musculus")
Q4: (protein AND translation) OR (human AND mouse AND "homo sapiens" AND "mus musculus")

Βήμα 1: Αφού απαντηθεί η παραπάνω ερώτηση, να υποβάλλετε σε βήματα το ερώτημα που επιλέξατε στην PubMed επιλέγοντας ολοένα και περισσότερους όρους αναζήτησης.

Ερώτηση για την τάξη

Σχολιάστε τα αποτελέσματα και αποθηκεύστε τα τελικά αποτελέσματα σε ένα αρχείο κειμένου ASCII με όνομα `results1.txt`.

Θεωρείτε ότι η αναζήτηση που πραγματοποιήσατε μπορεί να σας βοηθήσει να απαντήσετε το ερώτημα με το οποίο ξεκινήσατε;

Βήμα 2: Χρησιμοποιήστε τα φίλτρα που είναι διαθέσιμα στο αριστερά μέρος της σελίδας των αποτελεσμάτων ώστε να επιλέξετε μόνο τα άρθρα ανασκόπησης (Review) που δημοσιεύθηκαν από την 01/01/2014 και για τα οποία είναι ελεύθερα διαθέσιμο το πλήρες κείμενο (Free full text).

Ερώτηση για την τάξη

Συγκρίνετε τα αποτελέσματα που πήρατε με τα αποτελέσματα από την προηγούμενη ερώτηση. Αποθηκεύστε τα αποτελέσματα σε ένα αρχείο κειμένου ASCII με όνομα `results2.txt`.

[6]Με την εισαγωγή περισσότερων από ενός όρων σε διπλά εισαγωγικά υποδηλώνουμε ότι οι όροι αυτοί θέλουμε να εμφανίζονται διαδοχικά (με τη συγκεκριμένη σειρά) στο κείμενο που θα ανακτηθεί.

Βήμα 3: Χρησιμοποιήστε την προχωρημένη αναζήτηση και περιορίστε την αναζήτηση του όρου "translation" στον τίτλο (Title) των εγγραφών.

Ερώτηση για την τάξη

Συγκρίνετε τα αποτελέσματα που πήρατε με τα αποτελέσματα από την προηγούμενη ερώτηση. Αποθηκεύστε τα αποτελέσματα σε ένα αρχείο κειμένου ASCII με όνομα `results3.txt`. Συγκρίνετε τα μεγέθη των αρχείων που προέκυψαν από τις 3 αναζητήσεις και συσχετίστε τις με τον αριθμό των εγγραφών που κάθε φορά ανακτήσατε.

Κατ'οίκον εργασία

1. Απαντήστε στην ερώτηση της σελίδας 5.
2. Σας ενδιαφέρει να βρείτε πληροφορίες που αφορούν πρωτεΐνες του ανθρώπου οι οποίες εμπλέκονται στη διαδικασία της μετάφρασης και έχουν συσχετισθεί με κάποια ασθένεια (disease) πώς θα διαμορφώνατε το ερώτημα για αναζήτηση στην PubMed;
3. Να ετοιμάσετε συνοπτική αναφορά με τις απαντήσεις στις ερωτήσεις που απαντήθηκαν στην τάξη χρησιμοποιώντας όπου είναι απαραίτητο εικόνες και διαγράμματα.

Παράδοση εργασίας: Θα διευθετηθεί την ώρα του μαθήματος.

Σχετική Βιβλιογραφία

Manning CD, Raghavan P, Schütze H, Introduction to Information Retrieval, Cambridge University Press, 2008.

NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2015 Jan 28;43(Database issue):D6-D17.

Διαδικτυακές Πηγές

PubMed: http://www.ncbi.nlm.nih.gov/pubmed

PubMed tutorial: http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/cover.html