ΠΡΟΓΡΑΜΜΑ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗΣ ΑΕΙ ΓΙΑ ΤΗΝ ΕΠΙΚΑΙΡΟΠΟΙΗΣΗ ΓΝΩΣΕΩΝ ΑΠΟΦΟΙΤΩΝ ΑΕΙ (ΠΕΓΑ)

«Οι σύγχρονες τεχνικές βιο-ανάλυσης στην υγεία, τη γεωργία, το περιβάλλον και τη διατροφή»

Πρόγραμμα Δια Βίου Μάθησης. Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας.

Τίτλος Διάλεξης: Η Βιοπληροφορική και οι ελεύθερα διαθέσιμες Βάσεις Δεδομένων για ανθρώπινες ασθένειες με γενετικό υπόβαθρο.

Γρηγόρης Αμούτζιας, Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική

Προτεινόμενα συγγράμματα για Βιοπληροφορική:

Ελληνικά συγγράμματα:

- Andreas D. Baxevanis & B.F. Francis Quellette. Βιοπληροφορική: Ένας πρακτικός οδηγός για την ανάλυση γονιδίων και πρωτεΐνών.
- Σοφία Κοσσίδα. Βιοπληροφορική - Δυνατότητες & Προοπτικές.

Αγγλικά συγγράμματα:

- Jin Xiong. Essential Bioinformatics. Είναι ένα σύντομο, περιεκτικό και απλά γραμμένο σύγγραμα, ιδανικό για εισαγωγή στο αντικείμενο.
- David W. Mount. Bioinformatics. Sequence and genome analysis. Είναι ένα εκτενές και πολύ αναλυτικό σύγγραμα, ιδανικό για εμβάθυνση.

Τι είναι η Βιοπληροφορική:

Είναι η ανάπτυξη και χρήση τεχνικών και εργαλείων πληροφορικής/μαθηματικών/στατιστικής για την ανάλυση βιολογικών δεδομένων (κυρίως μοριακής βιολογίας). Σήμερα γίνεται διάκριση μεταξύ της Βιοπληροφορικής και της Υπολογιστικής Βιολογίας. Πλέον με τον όρο Βιοπληροφορική εννοούμε κυρίως την ανάπτυξη μεθόδων και προγραμμάτων, ενώ με τον όρο Υπολογιστική Βιολογία εννοούμε κυρίως την χρήση των παραπάνω μεθόδων και προγραμμάτων για την ανάλυση βιολογικών δεδομένων και την εξαγωγή συμπερασμάτων/γνώσης. Συχνά συμβαίνουν και τα δύο ταυτόχρονα και τα σύνορα δεν είναι πάντα ευδιάκριτα. Πλέον, μέσα στον χώρο της Βιοπληροφορικής/Υπολογιστικής Βιολογίας συναντάμε επιστήμονες από πολλές και συμπληρωματικές μεταξύ τους

ειδικότητες όπως π.χ. Βιολογία, Βιοχημεία, Χημεία, Χημική Μηχανική, Μηχανική, Υπολογιστές, Μαθηματικά, Στατιστική κ.α.

Οι βασικοί τομείς της Βιοπληροφορικής

Οι Βάσεις δεδομένων (Databases) αποτελούν τον βασικότερο ίσως τομέα της Βιοπληροφορικής. Είναι υπεύθυνες για την οργάνωση, αποθήκευση και αναζήτηση του τεράστιου όγκου δεδομένων που υπάρχουν πλέον στον χώρο των βιοεπιστημών.

Ένας άλλος σημαντικός τομέας είναι η ανάλυση ακολουθιών DNA, RNA, πρωτεϊνών (Sequence analysis). Ο τομέας αυτός περιλαμβάνει την στοίχιση ακολουθιών, δηλαδή την σύγκριση των αντίστοιχων περιοχών, μεταξύ δύο ή περισσοτέρων ακολουθιών, την αναζήτηση ομόλογων ακολουθιών (μέσω της στοίχισης) και την φυλογενετική ανάλυση, όπου αποκαλύπτονται οι εξελικτικές σχέσεις μεταξύ ομοειδών αντικειμένων όπως π.χ. γονίδια, πρωτεΐνες, οργανισμοί. Ο τομέας της ανάλυσης ακολουθιών είναι από τους πιο παλιούς και στο παρελθόν με τον όρο Βιοπληροφορική εννοούσαμε κυρίως αυτές τις αναλύσεις.

Ένας άλλος τομέας που άνθησε λόγω της αλληλούχισης των γονιδιωμάτων είναι και η Γονιδιακή ρύθμιση/έκφραση (Gene expression), που περιλαμβάνει κυρίως ανάλυση πολύπλοκων δεδομένων από μικροσυστοιχίες και Next-Generation RNA Sequencing. Λόγω του μεγάλου όγκου των δεδομένων αυτών αλλά και λόγω του θορύβου που περιέχουν, ο χώρος αυτός εστίασε στην ανάπτυξη και εφαρμογή στατιστικών εργαλείων, τα οποία όμως μπορούν να χρησιμοποιηθούν και σε πειραματικά δεδομένα μεγάλης κλίμακας (Omics).

Άλλος τομέας είναι η δομική Βιοπληροφορική, που εστιάζεται στην ανάλυση δευτεροταγών και τριτοταγών δομών RNA/πρωτεϊνών (structural biology). Εδώ, σημαντικές δραστηριότητες αποτελούν η πρόβλεψη δευτεροταγούς και τριτοταγούς δομής βιολογικών μακρομορίων όπως επίσης και η ανάλυση πρωτεϊνικών επιφανειών που αλληλεπιδρούν μεταξύ τους.

Την τελευταία 15ετία, λόγω της εκτεταμένης χρήσης επιστημονικών άρθρων και περιοδικών σε ηλεκτρονική μορφή (pdf, Html), δόθηκε η δυνατότητα εξόρυξης δεδομένων από την βιβλιογραφία με αυτοματοποιημένους τρόπους (text mining). Η δραστηριότητα αυτή έχει προσελκύσει μεγάλο ενδιαφέρον και είναι πολύ σημαντική για την αποδοτικότερη και ταχύτερη διάχυση της γνώσης στον επιστημονικό χώρο.

Όντας στην εποχή της γονιδιωματικής και των πειραματικών δεδομένων μεγάλης κλίμακας έχουμε πλέον κατανοήσει ότι τα βιομορία δεν λειτουργούν ως μεμονωμένα συστήματα μέσα στο κύτταρο, αλλά αλληλεπιδρούν μεταξύ τους και δημιουργούν περίπλοκες βιολογικές μηχανές, ή μονοπάτια μεταφοράς της πληροφορίας. Για τον λόγο αυτό, ένα σημαντικό μέρος της σύγχρονης Βιοπληροφορικής εστιάζει στα Βιολογικά δίκτυα/μονοπάτια όπως επίσης και στην Βιολογία Συστημάτων, όπου πειράματα μεγάλης κλίμακας από πολλά επίπεδα της βιολογίας, των δομικών συστατικών αλλά και της βιολογικής ρύθμισης ενσωματώνονται, για να δημιουργηθεί μια όσο το δυνατόν πλήρης εικόνα του τι συμβαίνει μέσα στο κύτταρο. Απώτερος σκοπός είναι η αποτύπωση όλης αυτής της πληροφορίας σε μαθηματικά μοντέλα τα οποία θα μας επιτρέπουν να προβλέψουμε την συμπεριφορά και τον φαινότυπου του κυττάρου ή και του οργανισμού με βάση τον γονότυπο και τα εξωτερικά ερεθίσματα.

Τέλος, ένας τομέας εξίσου σημαντικός με τους άλλους είναι και οι Οντολογίες (Ontologies), που στοχεύουν στην δημιουργία και χρήση ενός ελεγχόμενου λεξιλογίου (με ιεραρχική δόμηση), για την περιγραφή των ιδιοτήτων και των λειτουργιών ομοειδών αντικειμένων (π.χ πρωτεϊνών).

Πλέον ο όρος ‘βιοπληροφορική’ είναι τόσο εξειδικευμένος/γενικός, όσο και ο όρος ‘μοριακή βιολογία’! Βρισκόμαστε σε μια μεταβατική περίοδο για τις βιολογικές επιστήμες, όπως η φυσική πριν πολλά χρόνια. Θεωρείται πλέον βέβαιη η εισδοχή περισσότερων μαθηματικών, στατιστικής και πληροφορικής (προγραμματισμός) μεσοπρόθεσμα στα προγράμματα σπουδών με αντικείμενο τις βιοεπιστήμες.

Βάσεις δεδομένων

Η πρώτη βάση μοριακών δεδομένων δημιουργήθηκε το 1965 από την Margaret Dayhoff με το Atlas of protein sequence and structure, πρόδρομος της βάσης δεδομένων πρωτεϊνικών ακολουθιών PIR (protein information resource). Στη συνέχεια, η ανάπτυξη της τεχνολογίας και η δημιουργία νέων μοριακών δεδομένων με όλο και ταχύτερους ρυθμούς οδήγησε σε μια έκρηξη του αριθμού των βάσεων δεδομένων, οι οποίες στην συντριπτική πλειοψηφία τους είναι ελεύθερα προσβάσιμες.

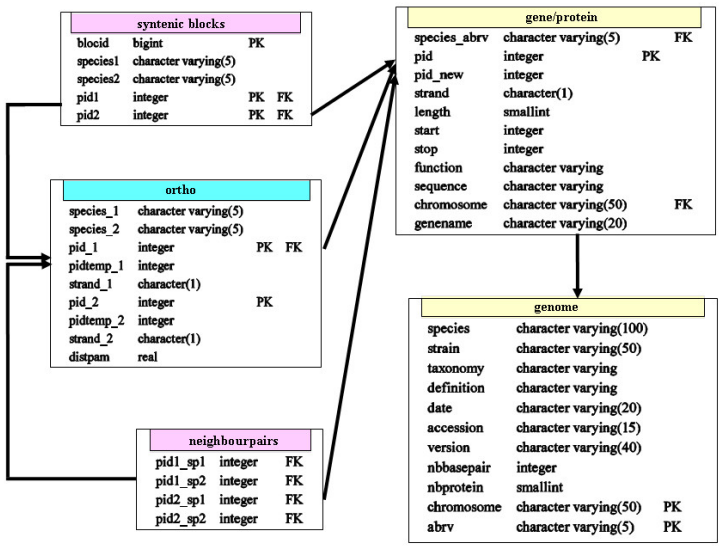
Οι βασικοί τρόποι αποθήκευσης δεδομένων είναι με καταλόγους επίπεδης οργάνωσης (Flat-files), με σχεσιακές βάσεις (relational databases), όπως π.χ. η SQL, με αντικειμενοστρεφείς βάσεις (object orientated), οι οποίες όμως δεν είναι ιδιαίτερα διαδεδομένες στον χώρο των βιοεπιστημών και τελευταία με βάσεις δεδομένων που βασίζονται σε γράφους (Graph-databases). Από τους παραπάνω τύπους, οι πιο απλές είναι οι κατάλογοι επίπεδης οργάνωσης, όπου κάθε μία εγγραφή έχει συγκεκριμένα πεδία τα οποία συμπληρώνονται. Όλες οι εγγραφές αποθηκεύονται η μία κάτω από την άλλη σε ένα μεγάλο αρχείο (Εικόνα 1α).

Αντίθετα, οι σχεσιακές βάσεις είναι πιο περίπλοκες αλλά και πολύ πιο διαδεδομένες πλέον σε σχέση με τους καταλόγους επίπεδης οργάνωσης. Εδώ, η πληροφορία οργανώνεται σε πίνακες (Εικόνα 1β) που σχετίζονται μεταξύ τους. Έτσι αποφεύγεται η επανάληψη και συσσώρευση δεδομένων.

```
LOCUS       name of locus, length and type of sequence,
            classification of organism, data of entry
DEFINITION  description of entry
ACCESSION   accession numbers of original source
KEYWORDS    key words for cross referencing this entry
SOURCE      source organism of DNA
ORGANISM    description of organism
REFERENCE
COMMENT     biological function or database information
FEATURES    information about sequence by base position or range of positions
        source          range of sequence, source organism
        misc_signal     range of sequence, type of function or signal
        mRNA            range of sequence, mRNA
        CDS             range of sequence, protein coding region
        intron          range of sequence, position of intron
        mutation        sequence position, change in sequence for mutation
BASE COUNT  count of A, C, G, T and other symbols
ORIGIN      text indicating start of sequence
      1 gaattcgata aatctctggt ttattgtgca gtttatggtt ccaaaatcgc
     51 atatactcac agcataactg tatatacacc caggggcgg aatgaaagcg
//                database symbol for end of sequence
```

Figure 2.5. GenBank DNA sequence entry.

Εικόνα 1α. Κατάλογος επίπεδης οργάνωσης για μία εγγραφή στην Genbank. Στα αριστερά των γραμμών αναγράφεται το είδος της πληροφορίας (πεδίο) που καταγράφεται στην συνέχεια. Το σύμβολο // απεικονίζει το τέλος αυτής της εγγραφής. Σε ένα κατάλογο επίπεδης οργάνωσης, από κάτω ακολουθεί μια νέα εγγραφή.



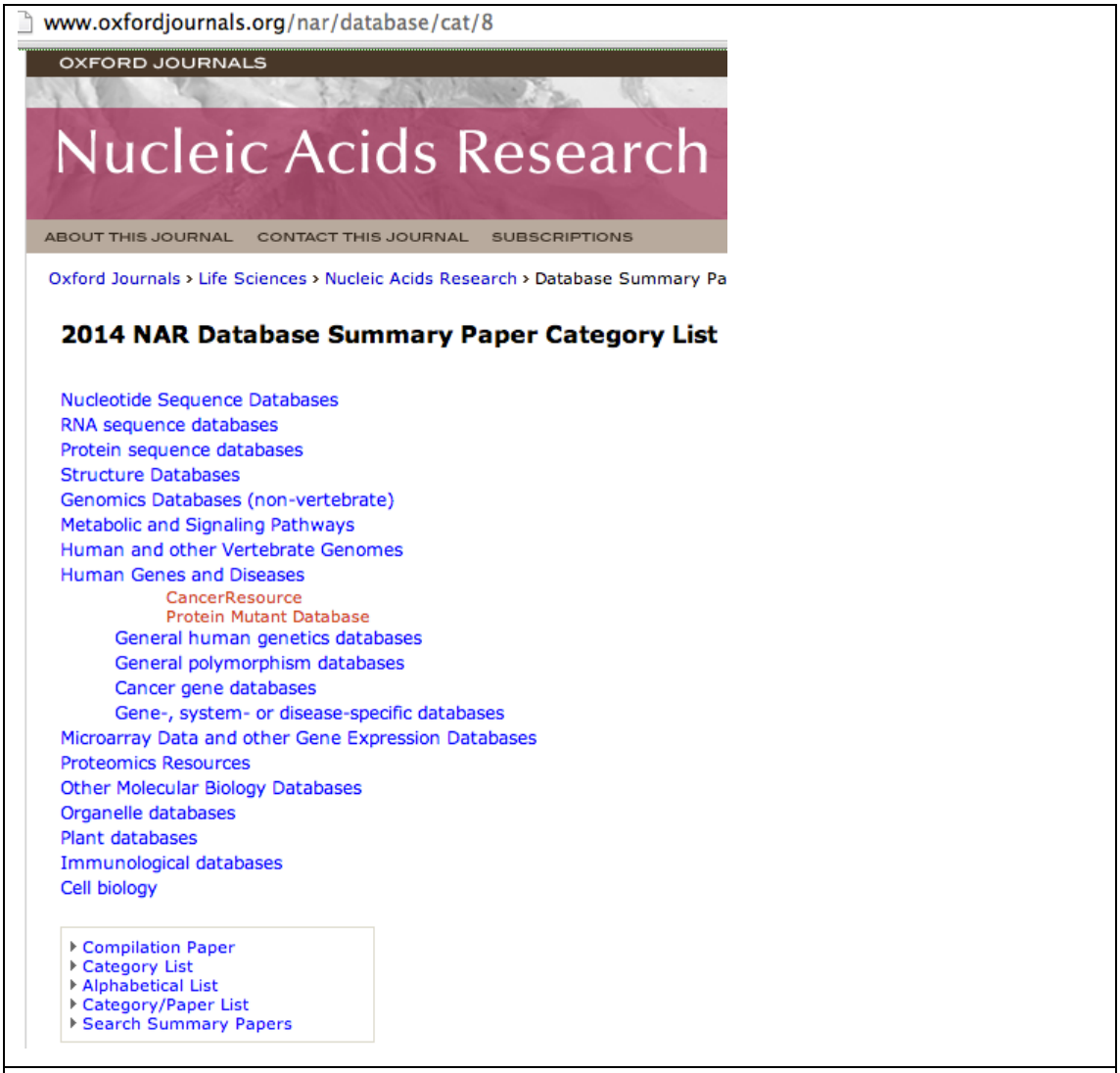
Εικόνα 1β. Σχεσιακή βάση δεδομένων όπου τα δεδομένα είναι οργανωμένα σε πίνακες. Οι πίνακες σχετίζονται μεταξύ τους και για μία εγγραφή ανακτώνται όλα τα σχετικά δεδομένα από τους πίνακες.

Οι βάσεις δεδομένων, ανάλογα με το τι είδους δεδομένα αποθηκεύουν διακρίνονται σε αρχειακές/πρωτεύοντες και δευτερεύοντες. Στις αρχειακές αποθηκεύονται πειραματικά δεδομένα που έχουν δημιουργηθεί για πρώτη

φορά από τους επιστήμονες. Π.χ., η Protein Data Bank (PDB) είναι αρχειακή, γιατί εκεί θα αποθηκευθούν οι τριτοταγείς δομές βιομορίων (συντεταγμένες των ατόμων στον χώρο) που έχουν ανακαλυφθεί με την βοήθεια ακτίνων Χ. Στις δευτερεύοντες βάσεις δεδομένων δεν κατατίθενται νέα πειραματικά δεδομένα. Σε αυτές γίνονται μετα-αναλύσεις δεδομένων που υπάρχουν στις αρχειακές. Παράδειγμα ενός τέτοιου τύπου βάσης δεδομένων αποτελούν οι CATH & SCOP, όπου βιοπληροφορικοί αλγόριθμοι αλλά και ειδικοί μελετούν τις τριτοταγείς δομές που κατατίθενται στην PDB και τις ταξινομούν ιεραρχικά με βάση την ομοιότητά τους (ως τρισδιάστατα αντικείμενα).

Ο Ετήσιος κατάλογος βάσεων δεδομένων του περιοδικού Nucleic Acids Research

Ο τεράστιος όγκος των βιολογικών δεδομένων αλλά και η πολύπλοκη φύση τους έχουν οδηγήσει τα τελευταία χρόνια σε μια έκρηξη του αριθμού των βάσεων δεδομένων που είναι διαθέσιμες στους βιοεπιστήμονες. Ένας σχετικά μικρός αριθμός από αυτές είναι γενικής φύσεως και χρησιμοποιούνται από σχεδόν όλους τους βιοεπιστήμονες σε καθημερινή βάση. Δεν θα μπορούσε να υπάρξει πλέον μοριακή βιολογία δίχως αυτές τις Β.Δ. Όμως, υπάρχει και ένας πολύ μεγάλος αριθμός Β.Δ. που είναι αρκετά εξειδικευμένες και χρησιμοποιούνται από μικρές και πολύ εξειδικευμένες κοινότητες επιστημόνων. Είναι πλέον αδύνατον να παρακολουθήσει κάποιος από μόνος του το τι συμβαίνει στον χώρο των Β.Δ. και το τι νέο έχει εμφανιστεί. Για τον λόγο αυτό, το διεθνές περιοδικό Nucleic Acids Research (NAR) έχει αναλάβει μια πρωτοβουλία να καταγράφει και να οργανώνει (σε θεματικές ενότητες – εικόνα 2) σε ένα ειδικό τεύχος (Database Issue) κάθε Ιανουάριο ένα ετήσιο κατάλογο με τις διάφορες Β.Δ. που υπάρχουν, ανανεώνονται ή δημιουργούνται εκ νέου (http://www.oxfordjournals.org/nar/database/c/). Χαρακτηριστικά, για το 2014 αυτός ο κατάλογος έχει 1552 καταχωρημένες Β.Δ. Ο κατάλογος είναι επίσης οργανωμένος και αλφαβητικά.



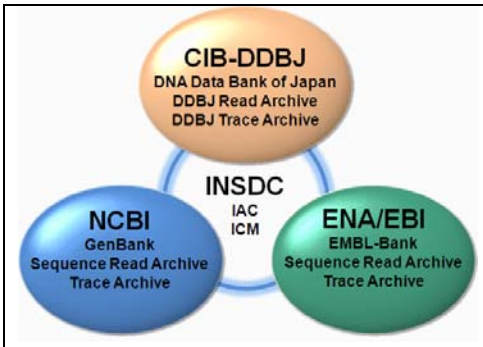
Εικόνα 2. Ο Ετήσιος κατάλογος των Β.Δ. του διεθνούς περιοδικού Nucleic Acids Research οργανωμένος ανά θεματική ενότητα.

Σημαντικές Β.Δ. ευρύτερου ενδιαφέροντος

Αρχειακές βάσεις νουκλεοτιδικών δεδομένων

Η EMBL-Bank στην Ευρώπη, η Genbank στην Αμερική και η DNA Databank of Japan στην Ιαπωνία αποτελούν 3 αρχειακές βάσεις δεδομένων όπου κατατίθενται νουκλεοτιδικές ακολουθίες από όλους τους οργανισμούς. Και οι 3 ΒΔ ανήκουν στο International nucleotide sequence database collection (INSDC) (εικόνα 3). Κάθε μέρα ανταλλάσουν δεδομένα μεταξύ τους, οπότε η ίδια ακολουθία εμφανίζεται και στις 3 ταυτόχρονα. Επομένως, αρκεί μια αναζήτηση σε οποιαδήποτε από τις 3 αυτές Β.Δ. Μια νέα ακολουθία

κατατίθεται όμως μόνο σε μία από τις 3 Β.Δ. η οποία είναι υπεύθυνη για την εγγραφή αυτή. Επίσης έχει την δυνατότητα (μόνο αυτή η Β.Δ.) να αναθεωρήσει την εγγραφή.



Εικόνα 3. Οι 3 ΒΔ που ανήκουν στο International nucleotide sequence database collection.

Από το 2009, το INSDC ξεκίνησε να καταχωρεί και αμορφοποίητα δεδομένα από μεγάλης κλίμακας αλληλουχίσεις (Sequencing projects), είτε αυτά προέρχονται από κλασσικές μεθόδους αλληλούχισης (Trace archive) (capillary sequencing), είτε από μεθόδους αλληλούχισης νέας γενιάς (Read Archive).

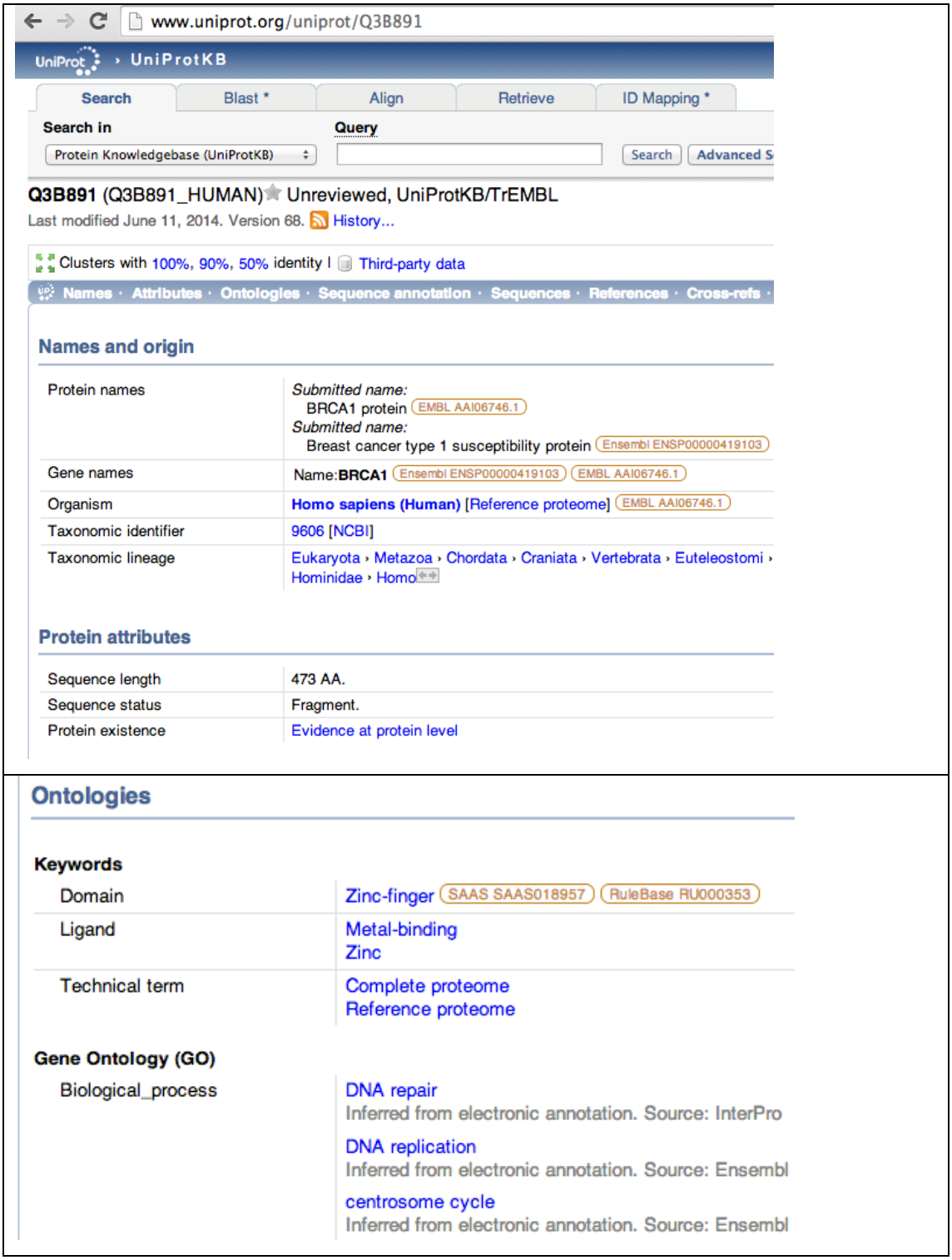
Β.Δ. Πρωτεϊνών

Ίσως η πιο σημαντική Β.Δ. πρωτεϊνών είναι η Swissprot και δημιουργήθηκε το 1987 στο Πανεπιστήμιο της Γενεύης. Κάθε εγγραφή είναι μια μοναδική πρωτεΐνη, όπου κάποιος βιοεπιστήμονας την μελέτησε μέσα από την διεθνή βιβλιογραφία και συμπλήρωσε τα αντίστοιχα πεδία στην Β.Δ. Λόγω του ελέγχου που υπόκειται η κάθε εγγραφή από ειδικούς βιοεπιστήμονες (database curators), η βάση αυτή αποτελεί σημείο αναφοράς είτε για βιοεπιστήμονες είτε για άλλες Β.Δ. Η Swissprot “μιλάει” με πολλές άλλες Β.Δ. και επομένως, μέσω συνδέσμων μπορεί κάποιος να πλοηγηθεί από την Swissprot σε μια άλλη πιο εξειδικευμένη Β.Δ. και να δει τι πληροφορίες υπάρχουν εκεί για την ίδια πρωτεΐνη.

Τα βασικά πεδία που έχει μια εγγραφή (π.χ. BRCA1: http://www.uniprot.org/uniprot/Q3B891) στην Swissprot είναι (εικόνα 4):

- το όνομα, συνώνυμα, ταξινόμηση του είδους
- Ιδιότητες της πρωτεΐνης, όπως το μήκος της
- Οντολογίες για την πρωτεΐνη, όπου ένα ελεγχόμενο λεξιλόγιο ιεραρχικής δόμησης χρησιμοποιείται για να περιγράψει τις βασικές ιδιότητες και λειτουργίες της πρωτεΐνης
- Η ακολουθία της πρωτεΐνης
- Βιβλιογραφικές αναφορές που χρησιμοποιήθηκαν για να γίνει αυτή η εγγραφή
- Σύνδεσμοι σε άλλες Β.Δ. που εμπεριέχουν εξειδικευμένες πληροφορίες για αυτή την πρωτεΐνη.

Προέκταση της Swissprot αποτελεί η TrEMBL (Translated EMBL), η οποία δημιουργήθηκε το 1996 και περιέχει πρωτεΐνες οι οποίες προέρχονται από την αυτόματη μετάφραση των νουκλεοτιδικών ακολουθιών της EMBL-Bank. Η διαφορά από την Swissprot είναι ότι οι εγγραφές δημιουργούνται αυτόματα και δεν υπόκεινται τον έλεγχο που περνάει μια εγγραφή της Swissprot από ειδικούς βιοεπιστήμονες. Επομένως, στην TrEMBL περιέχονται πολλές περισσότερες εγγραφές, η ποιότητα των δεδομένων τους όμως είναι σε κάποιες περιπτώσεις αμφίβολη. Μια Τρίτη Β.Δ. πρωτεϊνών είναι η PIR (Protein Information Resource) που εδράζεται στην Αμερική. Το 2002, οι παραπάνω 3 Β.Δ. ενώθηκαν κάτω από την ομπρέλα του Uniprot Knowledge Base (UniprotKB).



Ontologies

Keywords

Domain	Zinc-finger SAAS SAAS018957 RuleBase RU000353
Ligand	Metal-binding Zinc
Technical term	Complete proteome Reference proteome

Gene Ontology (GO)

Biological_process	DNA repair Inferred from electronic annotation. Source: InterPro DNA replication Inferred from electronic annotation. Source: Ensembl centrosome cycle Inferred from electronic annotation. Source: Ensembl

Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view
Experimental info					
☐	Non-terminal residue	473	1	Ensembl ENSP00000419103 EMBL AAI06746.1	

Sequences

	Sequence		Length	Mass (Da)	Tools
☐	Q3B891 [UniParc].	FASTA	473	53,351	Blast go

Last modified November 22, 2005. Version 1.
Checksum: 9A2D4ED5DA4B2AFA

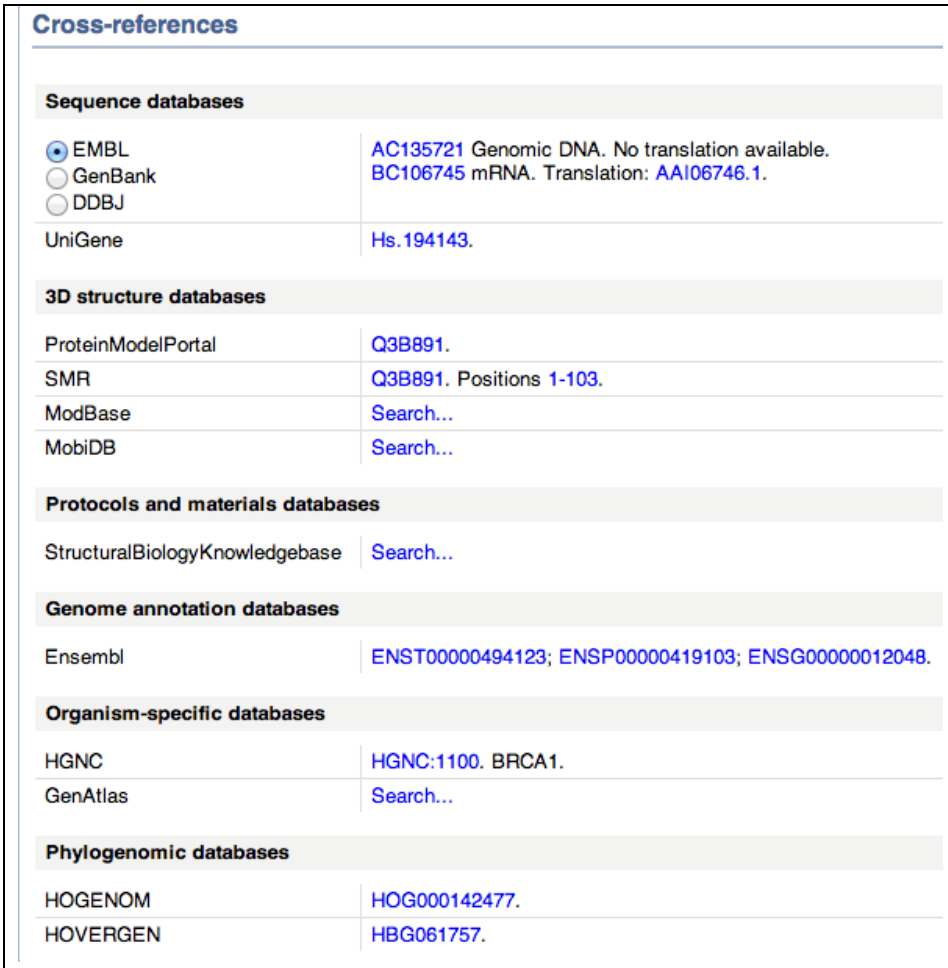
```
        10         20         30         40         50         60
MDLSALRVEE VQNVINAMQK ILECPICLEL IKEPVSTKCD HIFCKFCMLK LLNQKKGPSQ

        70         80         90        100        110        120
CPLCKNDITK RSLQESTRFS QLVEELLKII CAFQLDTGLE YANSYNFAKK ENNSPEHLKD

       130        140        150        160        170        180
EVSIIQSMGY RNRAKRLLQS EPENPSLQET SLSVQLSNLG TVRTLRTKQR IQPQKTSVYI

       190        200        210        220        230        240
ELGSDSSEDT VNKATYCSVG DQELLQITPQ GTRDEISLDS AKKAACEFSE TDVTNTEHHQ

       250        260        270        280        290        300
PSNNDLNTTE KRAAERHPEK YQGSSVSNLH VEPCGTNTHA SSLQHENSSL LLTKDRMNVE

       310        320        330        340        350        360
KAEFCNKSKQ PGLARSQHNR WAGSKETCND RRTPSTEKKV DLNADPLCER KEWNKQKLPC

       370        380        390        400        410        420
SENPRDTEDV PWITLNSSIQ KVNEWFSRSD ELLGSDDSHD GESESNAKVA DVLDVLNEVD

       430        440        450        460        470
EYSGSSEKID LLASDPHEAL ICKSERVHSK SVESNIEDKI FGKTYRKKAS LPN
```

References

« Hide 'large scale' references

[1] **"The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)."**
The MGC Project Team
Gerhard D.S., Wagner L., Feingold E.A., Shenmen C.M., Grouse L.H., Schuler G., Klein S.L., Old S., Rasooly R., Good P. W., Sherry S., Feolo M. Malek J.
Genome Res. 14:2121-2127(2004) [PubMed] [Europe PMC] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].

[2] **"DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage."**
Zody M.C., Garber M., Adams D.J., Sharpe T., Harrow J., Lupski J.R., Nicholson C., Searle S.M., Wilming L., Young S.K. Bugalter B.E., Butler J., Chang J.L. Nusbaum C.
Nature 440:1045-1049(2006) [PubMed] [Europe PMC] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].



Cross-references

Sequence databases	
EMBL GenBank DDBJ	AC135721 Genomic DNA. No translation available. BC106745 mRNA. Translation: AAI06746.1.
UniGene	Hs.194143.
3D structure databases	
ProteinModelPortal	Q3B891.
SMR	Q3B891. Positions 1-103.
ModBase	Search...
MobiDB	Search...
Protocols and materials databases	
StructuralBiologyKnowledgebase	Search...
Genome annotation databases	
Ensembl	ENST00000494123; ENSP00000419103; ENSG00000012048.
Organism-specific databases	
HGNC	HGNC:1100. BRCA1.
GenAtlas	Search...
Phylogenomic databases	
HOGENOM	HOG000142477.
HOVERGEN	HBG061757.

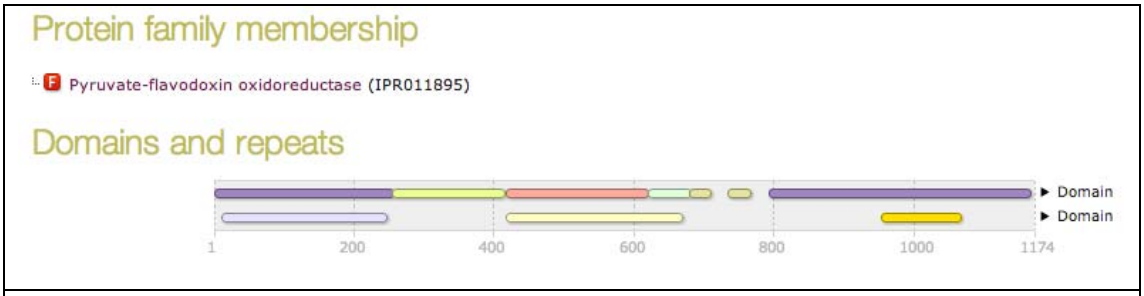
Εικόνα 4. Τα βασικά πεδία της εγγραφής για την ανθρώπινη πρωτεΐνη BRCA1 στην Swissprot.

Β.Δ. πρωτεϊνικών επικρατειών

Οι πρωτεϊνικές επικράτειες (domains) είναι περιοχές (σε μια πρωτεΐνη) με συγκεκριμένη λειτουργία, τριτοταγή δομή και συνήθως είναι καλά συντηρημένες. Ο εντοπισμός μια επικράτειας σε μια άγνωστη πρωτεΐνη μας βοηθάει να προβλέψουμε σε ένα αρκετά ικανοποιητικό βαθμό τις βιοχημικές/μοριακές λειτουργίες αυτής της πρωτεΐνης. Π.χ., αν σε μια άγνωστη πρωτεΐνη εντοπιστεί μια επικράτεια υπεύθυνη για σύνδεση στο DNA, τότε πιθανόν η άγνωστη πρωτεΐνη να λειτουργεί ως μεταγραφικός παράγοντας. Επιπλέον, μια μετάλλαξη που συναντάται σε μία βαθιά συντηρημένη επικράτεια μπορεί να μας βοηθήσει να προβλέψουμε τι

επιπτώσεις θα έχει αυτή η μετάλλαξη στο μοριακό μονοπάτι και στον φορέα της.

Στο παρελθόν υπήρξαν πολλές και ανεξάρτητες προσπάθειες να καταγραφούν οι πρωτεϊνικές επικράτειες σε Βάσεις Δεδομένων. Η κάθε μία ερευνητική ομάδα μπορεί να περιέγραφε μια συγκεκριμένη επικράτεια με ελαφρά διαφορετικό τρόπο από μια άλλη ομάδα. Επιπλέον, μια Β.Δ. μπορεί να περιείχε πληροφορίες για μια επικράτεια που δεν συναντούνταν σε άλλη Β.Δ. Διάφορες βάσεις δεδομένων, όπως η PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIR superfamily, Superfamily έχουν πλέον ενοποιηθεί κάτω από την ομπρέλα του INTERPRO (εικόνα 5) (http://www.ebi.ac.uk/interpro/). Επιπλέον, το πρόγραμμα INTERPROscan ανιχνεύει αυτές τις επικράτειες στις πρωτεΐνες.



Εικόνα 5. Παράδειγμα οργάνωσης πρωτεϊνικών επικρατειών σε μια πρωτεΐνη, με βάση το Interpro.

Β.Δ. Μοριακών μονοπατιών και ασθενειών.

Από τις ποιό σημαντικές Β.Δ. αυτής της κατηγορίας είναι η KEGG (Kyoto encyclopedia of genes and genomes). Περιέχει μεταβολικά/μοριακά μονοπάτια και πληροφορίες για τα συμμετέχοντα γονίδια. Οι πληροφορίες για τα μοριακά μονοπάτια έχουν συγκεντρωθεί από ειδικούς Βιοεπιστήμονες (curators) που μελετήσαν τη σχετική Βιβλιογραφία και επομένως θεωρείται Β.Δ. υψηλής ποιότητας. Εκτός από τα μοριακά μονοπάτια έχει και πληροφορίες για μοριακές ασθένειες που σχετίζονται με αυτά τα μονοπάτια, αλλά και για σχετικά φάρμακα (εικόνα 6).

www.genome.jp/kegg/

KEGG Search Help

» Japanese

KEGG Home
Release notes
Current statistics
Plea from KEGG

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies

KEGG Software
KegTools
KEGG API
KGML

KEGG FTP
Subscription

GenomeNet

DBGET/LinkDB

Feedback

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (See Release notes for new and updated features).
Please see: Renewed plea to support KEGG

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents	Update notes

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps	[Pathway list]
KEGG BRITE	BRITE functional hierarchies	[Brite list]
KEGG MODULE	KEGG modules	[Module list]
KEGG ORTHOLOGY	Ortholog groups	[KO system]
KEGG GENOME	Genomes	[KEGG organisms]
KEGG GENES	Genes and proteins	Release history
KEGG COMPOUND	Small molecules	[Compound classification]
KEGG REACTION	Biochemical reactions	[Reaction modules]
KEGG DISEASE	Human diseases	[Cancer | Infectious disease]
KEGG DRUG	Drugs	[ATC drug classification]
KEGG MEDICUS	Health information resource	[Drug labels search]

Organism-specific entry points

KEGG Organisms Enter org code(s) Go hsa hsa eco

Analysis tools

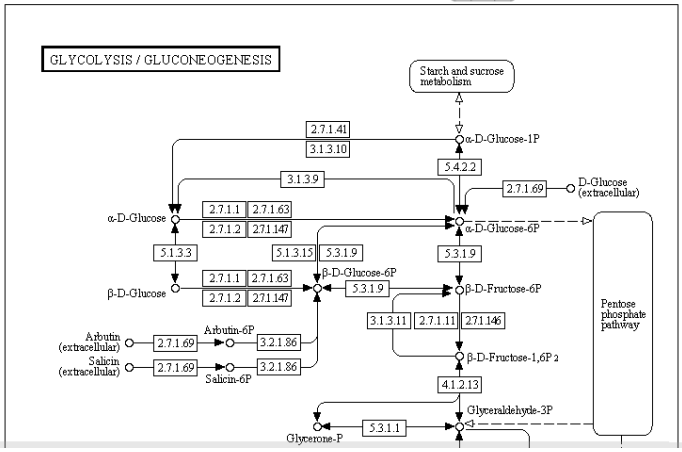
KEGG Mapper KEGG PATHWAY/BRITE/MODULE mapping tools

Glycolysis / Gluconeogenesis - Reference pathway

[Pathway menu | Pathway entry | Hide description]

Glycolysis is the process of converting glucose into pyruvate and generating small amounts of ATP (energy) and NADH (reducing power). It is a central pathway that produces important precursor metabolites: six-carbon compounds of glucose-6P and fructose-6P and three-carbon compounds of glycerone-P, glyceraldehyde-3P, glycerate-3P, phosphoenolpyruvate, and pyruvate [MD:M00001]. Acetyl-CoA, another important precursor metabolite, is produced by oxidative decarboxylation of pyruvate [MD:M00679]. When the enzyme genes of this pathway are examined in completely sequenced genomes, the reaction steps of three-carbon compounds from glycerone-P to pyruvate form a conserved core module [MD:M00002], which is found in almost all organisms and which often corresponds to operon structures in bacterial genomes. Gluconeogenesis is a synthesis pathway of glucose from noncarbohydrate precursors. It is essentially a reversal of glycolysis with minor variations of alternative paths [MD:M00003].

Reference pathway Go



Εικόνα 6. Η Β.Δ. KEGG και τα είδη της πληροφορίας που ενσωματώνει.

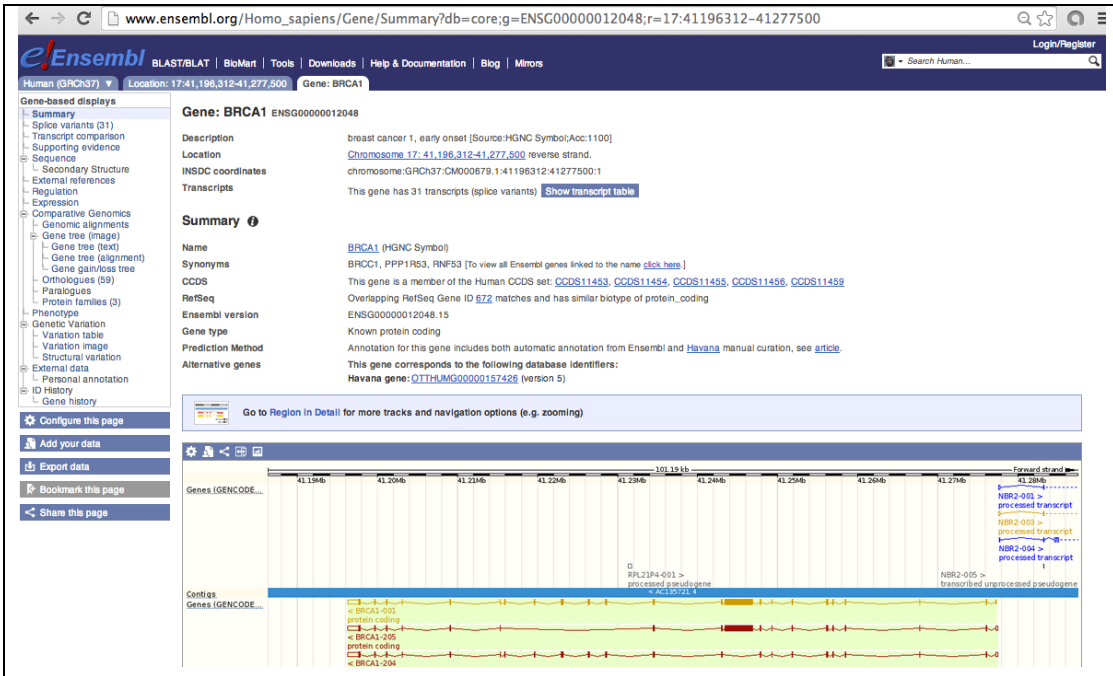
Β.Δ. που εξειδικεύονται στα ανθρώπινα γονίδια & ασθένειες.

ENSEMBL

Είναι από τις πιο σημαντικές Β.Δ. για γονιδιώματα (http://www.ensembl.org/). Ενσωματώνει πληροφορίες:

- Γενικές (περιγραφή του γονιδίου)
- Δομή του γονιδίου (Εξόνια/Ιντρόνια)
- Αντίστοιχα μετάγραφα και πρωτεΐνες
- Εξελικτικά δεδομένα από την σύγκριση με άλλα είδη (ορθόλογα/παράλογα)
- Σχετιζόμενους φαινότυπους και πολυμορφισμούς/μεταλλάξεις
- Συνδέσεις σε άλλες πιο εξειδικευμένες Β.Δ

Επιπλέον, η ENSEMBL έχει ένα υπολογιστικό εργαλείο, το BioMart, με το οποίο κάποιος χρήστης μπορεί να υποβάλλει στη Β.Δ. μια επερώτηση, π.χ. μια λίστα με γονίδια και να ζητήσει μια σειρά από διαφορετικές και πολύ συγκεκριμένες πληροφορίες για αυτή την λίστα, π.χ. ακολουθίες, πολυμορφισμούς, και όχι όλα τα υπόλοιπα, αντί να κάνει αναζήτηση για κάθε ένα γονίδιο ξεχωριστά.



Εικόνα 7. Η Β.Δ. ENSEMBL και τα είδη της πληροφορίας που ενσωματώνει.

NCBI Gene

Είναι Β.Δ. του NCBI που περιέχει πληροφορίες για γονίδια, (http://www.ncbi.nlm.nih.gov/gene) όπως:

- Γενικά χαρακτηριστικά (επίσημη ονομασία γονιδίου, συνώνυμα, περιληπτική περιγραφή).
- Γονιδιωματικές πληροφορίες.
- Σχετική Βιβλιογραφία.
- Φαινότυποι που σχετίζονται με το γονίδιο αυτό.
- Πολυμορφισμοί του γονιδίου.
- Μοριακά μονοπάτια στα οποία συμμετέχει το γονίδιο.
- Αλληλεπιδράσεις της πρωτεΐνης του.

κ.α.

MedGen

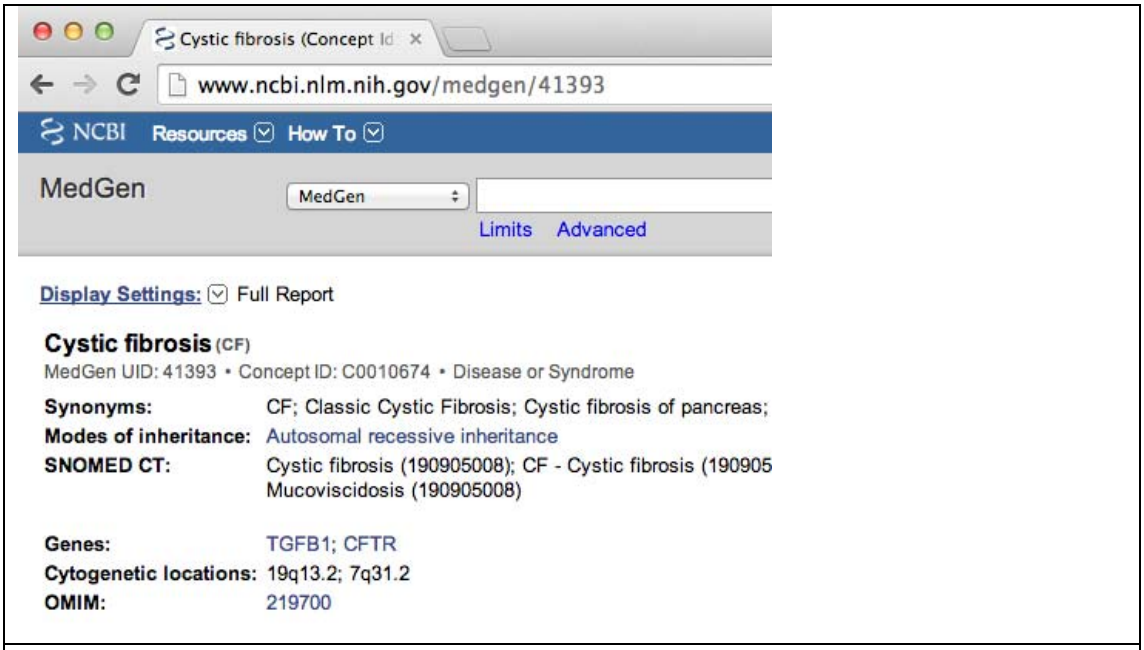
Είναι Β.Δ. του NCBI που περιέχει πληροφορίες για ανθρώπινες ασθένειες και φαινότυπους που έχουν γενετικό υπόβαθρο (http://www.ncbi.nlm.nih.gov/medgen) (εικόνα 8).

Παρόμοια δεδομένα έχει επίσης και μία άλλη πολύ δημοφιλής Β.Δ. η OMIM (Online Mendelian Inheritance in Man) (http://www.ncbi.nlm.nih.gov/omim).

Η MedGen περιέχει πληροφορίες όπως:

- Γενικά χαρακτηριστικά (συνώνυμα, τρόπος κληρονόμισης)
- τα γονίδια που εμπλέκονται στην ασθένεια
- τα χαρακτηριστικά της ασθένειας
- Πληροφορίες από άλλες Β.Δ. όπως η OMIM
- Κλινικά χαρακτηριστικά
- Κλινικές μελέτες
- Διάγνωση
- Θεραπεία
- Πρόγνωση

κ.α.



Εικόνα 8. Η Β.Δ. MedGen και τα είδη της πηροφορίας που ενσωματώνει.

Genetic Testing Registry

Είναι Β.Δ. του NCBI που περιέχει πληροφορίες για τα γενετικά τεστ που είναι διαθέσιμα σε διάφορα εργαστήρια για μια συγκεκριμένη πάθηση/ασθένεια με γενετικό υπόβαθρο (http://www.ncbi.nlm.nih.gov/gtr/) (εικόνα 9).

Περιέχει πληροφορίες όπως:

- Γενικά χαρακτηριστικά της ασθένειας

- Διαθέσιμα γενετικά τεστ
- Ποιά γονίδια εμπλέκονται
- Άλλες σχετιζόμενες παθήσεις
- Κλινικά χαρακτηριστικά

Εικόνα 9. Η Β.Δ. Genetic Testing Registry, του NCBI.