Βιοπληροφορική

΄Υλη

Πηγές στο διαδίκτυο για συγγράμματα με μορφή pdf

- Filecrop: http://www.filecrop.com/
- 4shared: http://www.4shared.com/

Προσοχή σε περιπτώσεις που παραβιάζονται πνευματικά δικαιώματα!

Προτεινόμενα συγγράμματα

- 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. (Σύντομο, περιεκτικό και απλά γραμμένο σύγγραμα).
  - David W. Mount. Bioinformatics. Sequence and genome analysis. (Εκτενές και πολύ αναλυτικό σύγγραμα)

Baxevanis & Quellette

Κεφάλαια:

- 3 Β.Δ. Genbank.
- 5 Β.Δ. Δομών
- 8 Στοίχιση ακολουθιών και αναζήτηση σε Β.Δ.
- 9 Πολλαπλές στοιχίσεις.
- 12 ESTs
- 14 Φυλογένεση
- 15 & 16 Ανάλυση γονιδιωμάτων.

Essential bioinformatics

- Για όσους θέλουν να εμβαθύνουν στην Βιοπληροφορική, προτείνεται να διαβάσουν και από το αγγλικό σύγγραμμα του Jin Xiong, και ειδικότερα τα κεφάλαια 1-7, 10,11 & 18.

Βιοπληροφορική

Εισαγωγή

Βιοπληροφορική: τι 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
  - Βιοπληροφορική: Η ανάπτυξη μεθόδων και προγραμμάτων.
  - Υπολογιστική Βιολογία: Η χρήση των παραπάνω μεθόδων και προγραμμάτων για την ανάλυση βιολογικών δεδομένων.

- Συχνά συμβαίνουν και τα δύο ταυτόχρονα και τα σύνορα δεν είναι πάντα ευδιάκριτα

- Πολλές και συμπληρωματικές μεταξύ τους ειδικότητες (από Βιολογία, Βιοχημεία, Χημεία, Χημική Μηχανική, Μηχανική, Υπολογιστές, Μαθηματικά, Στατιστική κ.α.) συνεργάζονται σήμερα στο χώρο της Βιοπληροφορικής

Βιοπληροφορική: βασικοί τομείς

- Βάσεις δεδομένων (Databases)
  - Οργάνωση, αποθήκευση, αναζήτηση των δεδομένων.

- Ανάλυση ακολουθιών DNA, RNA, πρωτεϊνών. (Sequence analysis)
  - Στοίχιση ακολουθιών: Σύγκριση των αντίστοιχων/ομόλογων περιοχών, μεταξύ δύο ή περισσοτέρων ακολουθιών.
  - Φυλογενετική ανάλυση: Οι εξελικτικές σχέσεις μεταξύ ομοειδών αντικειμένων (γονίδια, πρωτεΐνες, οργανισμοί).

- Γονιδιακή ρύθμιση/έκφραση (Gene expression)

  Ανάλυση δεδομένων από μικροσυστοιχίες, RNA-seq.

- Δομή RNA/πρωτεϊνών (structural biology):

  Πρόβλεψη δευτεροταγούς και τριτοταγούς δομής. Ανάλυση πρωτεϊνικών επιφανειών που αλληλεπιδρούν μεταξύ τους.

- Εξόρυξη δεδομένων από βιβλιογραφία (text mining).

- Βιολογικά δίκτυα/μονοπάτια, Βιολογία Συστημάτων (FBA, MCA).

- Οντολογίες (Ontologies)

  Η χρήση ενός ελεγχόμενου λεξιλογίου (με ιεραρχική δόμηση), για την περιγραφή των ιδιοτήτων και των λειτουργιών ομοειδών αντικειμένων (π.χ πρωτεϊνών).

Ιστορική αναδρομή

- 1965: Η πρώτη έκδοση του Atlas of protein sequence and structure (Margaret Dayhoff), πρόδρομος της βάσης δεδομένων πρωτεϊνικών ακολουθιών PIR (protein information resource).
  - Ακολουθούν και άλλες βάσεις δεδομένων. 1986: Swissprot, Geneva
- 1970: Αλγόριθμος Needleman-Wunsch για την σύγκριση ακολουθιών
- 1990: Blast
- 1990s: Αρχή του Human genome project, που ‘ολοκληρώθηκε’ το 2001. Κινητήριος δύναμη για την αλματώδη ανάπτυξη της Βιοπληροφορικής.

Παρόν/μέλλον

- Μέχρι το 2000, Βιοπληροφορική σήμαινε κυρίως ανάλυση ακολουθιών.
- Η γενωμική αποτέλεσε το ερέθισμα για την ανάπτυξη τεχνολογιών που κάνουν μετρήσεις ευρείας κλίμακας.
- Από το 2000 και μετά, η Βιοπληροφορική καλείται επίσεις να διαχειριστεί και να αναλύσει μεγάλα και πολύπλοκα δεδομένα από το χώρο της γενωμικής, της γονιδιακής έκφρασης, της πρωτεομικής κ.α.
- 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. Βέβαιη η εισδοχή περισσότερων μαθηματικών, στατιστικής και πληροφορικής (προγραμματισμός) μεσοπρόθεσμα στο πρόγραμμα σπουδών.

Bioinformatics Market - Advanced Technologies, Global Forecast and Winning Imperatives (2009 - 2014)

- Απόσπασμα από:
  - http://www.marketsandmarkets.com/Market-Reports/bioinformatics-39.html

- The market for bioinformatics platforms is growing at a significant pace with the increasing demand from U.S. and Europe.
- This trend is supported by the increasing demand for sequencing platforms with increasing life science research using techniques such as gene expression analysis, sequence analysis, and protein expression analysis.
- The global bioinformatics market is expected to reach **$8.3 billion** by 2014 at a high CAGR of 24.8% from 2009-2014. While knowledge management formed the largest submarket is 2009 at $1.3 billion, the bioinformatics platforms market is expected to have greatest market share in 2014 at an estimated $3.9 billion, due to rising demand from the U.S. and Europe.

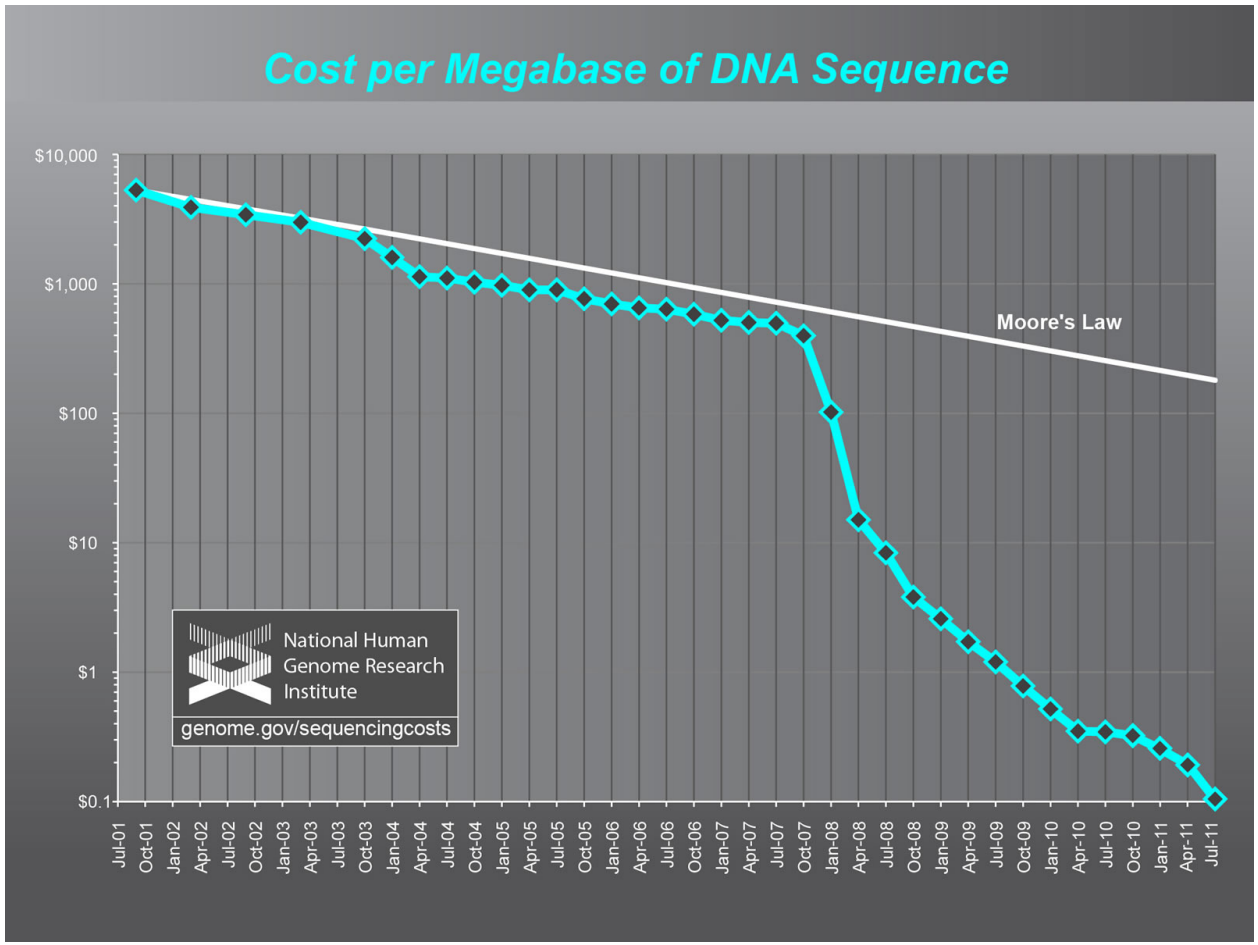
- Συμβουλευτική (δουλειά από το σπίτι)?

Χαμηλό κόστος γενωμικών τεχνολογιών θα οδηγήσει σε καθημερινές εφαρμογές.

- Κόστος αλληλούχισης πέφτει διαρκώς.
  - Illumina -> 1 lane: 19GBp, ~ €3000, 10 βακτηριακά γενώματα.
- Τα δείγματα αποστέλλονται σε κέντρα με μεγάλες εγκαταστάσεις και χαμηλό κόστος λειτουργίας (οικονομία κλίμακας). Η ανάλυση των δεδομένων όμως δεν υπόκειται σε όρους οικονομίας κλίμακας.
- Πλέον, ένα σημαντικό μέρος του ολικού κόστους είναι η βιοπληροφορική ανάλυση.
- Μηχανήματα αλληλούχισης ακριβά (Illumina ~ €600.000) - service φτηνό.
- Μισθός ακριβός (ίσως ένα νέο μοντέλο συμβουλευτικής?)
- Υπολογιστής φτηνός (€3-5.000), εφόσον πρόκειται για μικρά γονιδιώματα (de novo assembly), ή για re-sequencing.

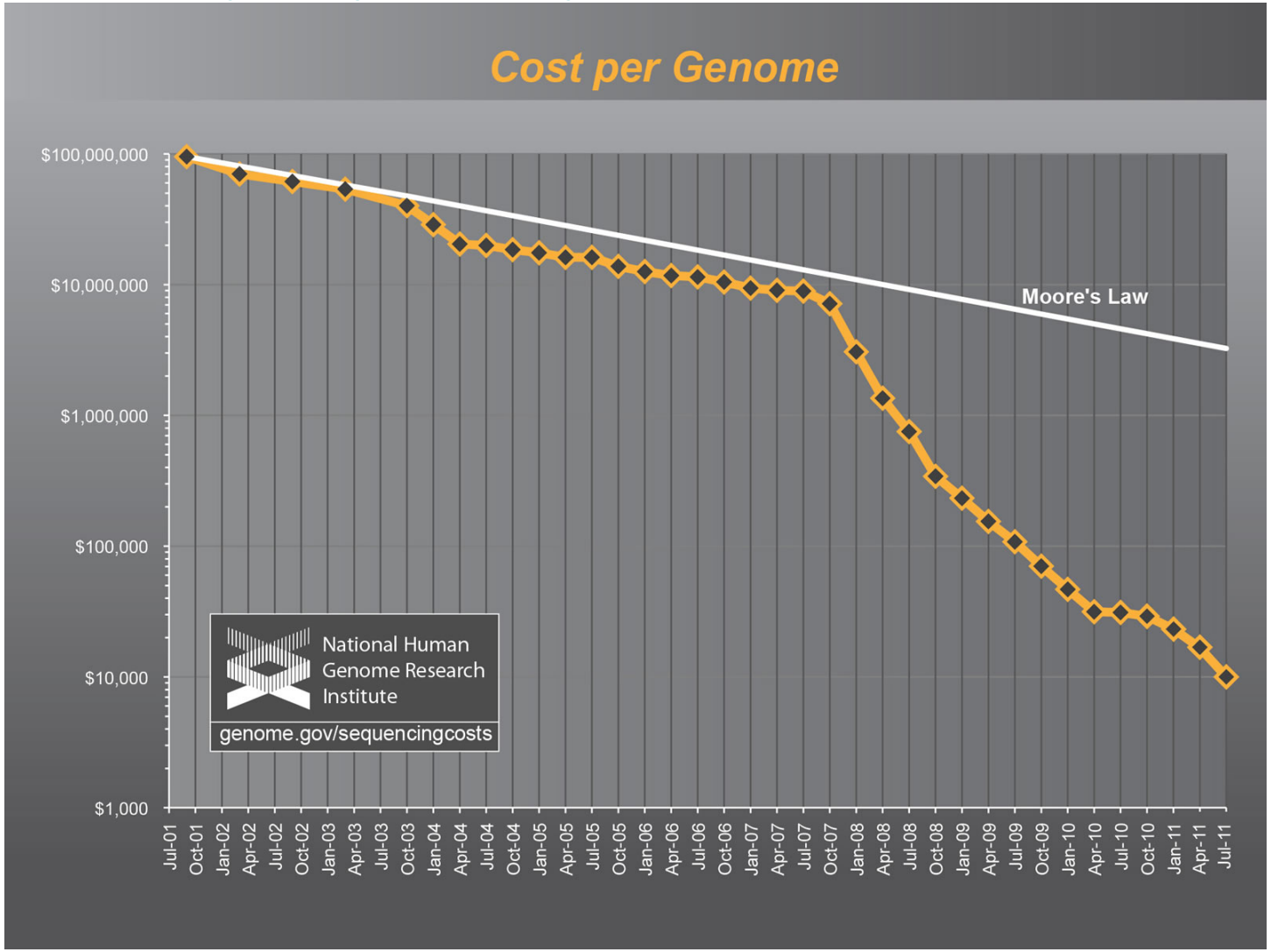
Χαμηλό κόστος γενωμικών τεχνολογιών θα οδηγήσει σε καθημερινές εφαρμογές

- Κόστος αλληλούχισης
  - http://www.genome.gov/sequencingcosts/
- Ο νόμος του Moore προβλέπει διπλασιασμό της υπολογιστικής ισχύς κάθε δύο χρόνια.



Χαμηλό κόστος γενωμικών τεχνολογιών θα οδηγήσει σε καθημερινές εφαρμογές

- Κόστος αλληλούχισης
  - http://www.genome.gov/sequencingcosts/



Εφαρμογές

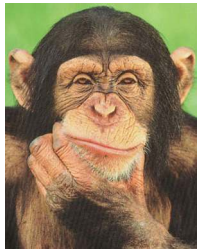
‘Ελεγχος εξελικτικών υποθέσεων -

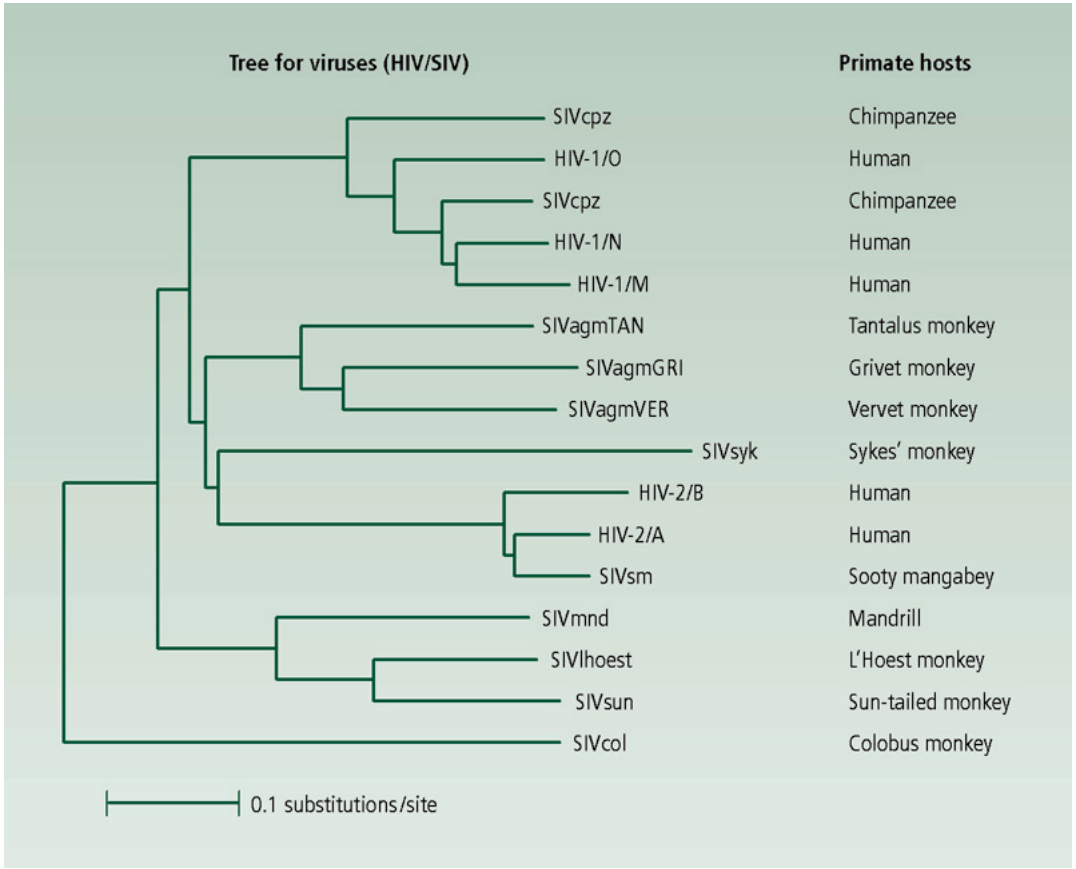
Προέλευση -

Επιδημιολογία

Έλεγχος εξελικτικών υποθέσεων

Από που προήλθε ο ιός HIV;





Πρωτοεμφανίστηκε μυστηριωδώς στις αρχές της δεκαετίας του 1980.

Ο τύπος HIV-1 εισήλθε στους ανθρώπους, ίσως περισσότερες από μια φορές, από τον χιμπατζή.

Ο τύπος HIV-2 εισήλθε στους ανθρώπους, από τους sooty mangabees

Έλεγχος εξελικτικών υποθέσεων

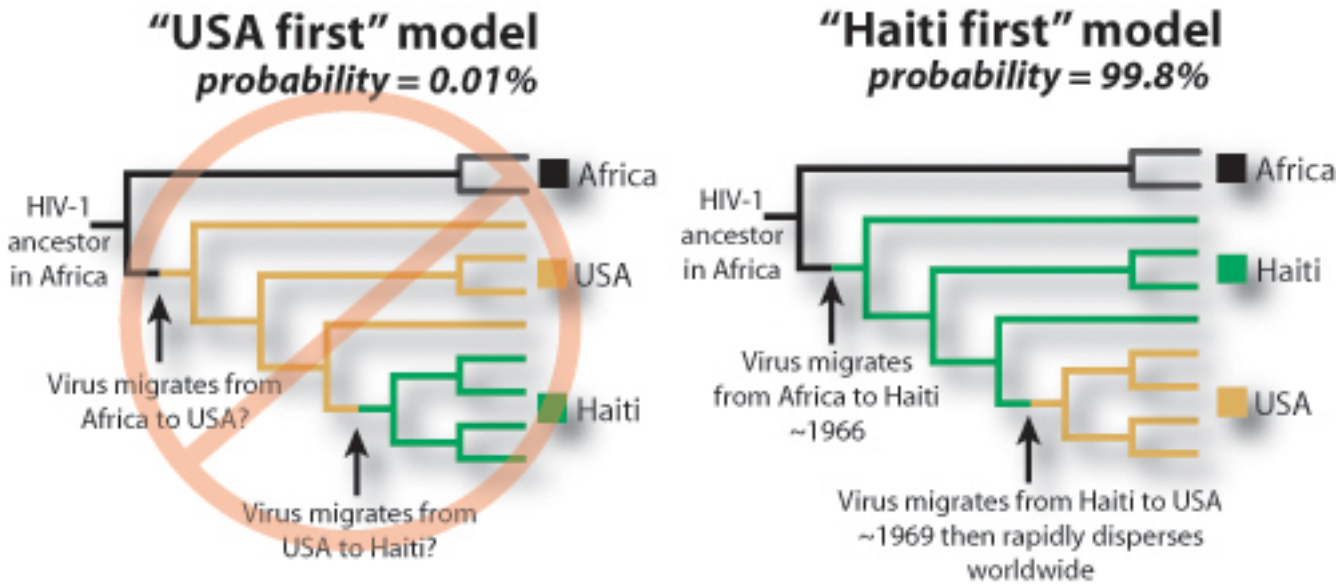
Από που προήλθε ο ιός HIV-1 subtype M; Προέλευση στην Κεντρική Αφρική.

Όταν πρωτοεντοπίστηκε, αρκετοί ασθενείς στην Αμερική ήταν πρόσφατοι Αϊτινοί μετανάστες.

Κάποιοι ισχυρίζονταν ότι πήγε από την Αμερική στην Αϊτή στα μέσα των 70s, λόγω σεξοτουρισμού.

Από την Αϊτή στην Αμερική ή το αντίθετο;

Ο Worobey χρησιμοποίησε ακολουθίες HIV από συντηρημένα δείγματα Αϊτινών ασθενών (1983)



Επιδημία χολέρας στην Αϊτή 2010

•Μετά τον σεισμό στην Αϊτή (Ιανουαριος 2010), ξέσπασε επιδημία χολέρας (Οκτώβριος 2010).
•Το βακτήριο Vibrio cholerae ελευθερώνει μια τοξίνη που προκαλεί έντονες διάρροιες και αφυδάτωση, έως και θάνατο, εντός ολίγων ωρών, αν δεν αντιμετωπιστεί!
•Η μετάδοση γίνεται όταν τα κόπρανα ενός μολυσμένου ατόμου έρθουν σε επαφή με πόσιμο νερό ή τροφή.
•Τα άτομα που δεν παράγουν αρκετό γαστρικό υγρό στο στομάχι τους, ή τα άτομα με ομάδα αίματος Ο είναι πιο ευάλωτα.
•Το Vibrio cholerae υπάρχει σε υδάτινα περιβάλλοντα ανά την υφήλιο και εάν οι συνθήκες είναι ευνοϊκές, μπορεί να ξεσπάσει επιδημία.
•Η χολέρα είναι διαδεδομένη στην Ασία.
•Τα πρώτα κρούσματα παρατηρήθηκαν σε κεντρικές περιοχές του νησιού, στην κοιλάδα Artibonite, μια εβδομάδα μετά την έλευση Νεπαλέζων κυανόκρανων, κοντά στο στρατόπεδό τους.
•Λύμματα από το στρατόπεδο κατέληγαν σε γειτονικό ποταμό.
•Οι κάτοικοι κατηγόρησαν τον ΟΗΕ ότι
    •οι κυανόκρανοι που ήρθαν να βοηθήσουν ευθύνονται για το ξέσπασμα της επιδημίας.
    •ότι ο ΟΗΕ προσπάθησε να αποκρύψει το γεγονός και να μην αναλάβει τις ευθύνες του

Ξέσπασαν ταραχές.

Επιδημία χολέρας στην Αϊτή 2010

•Αλληλούχιση του γονιδιώματος:
- •2 κλινικών στελεχών από την τωρινή επιδημία στην Αϊτή.
- •1 κλινικό στέλεχος από την επιδημία του 1991 στη Νότια Αμερική.
- •2 στέλεχη που απομονώθηκαν στη Νότια Ασία το 2002 και 2008.

•Επίσης χρησιμοποιήθηκαν οι μερικές αλληλουχίες από 23 άλλα στελέχη ανά την υφήλιο (τα τελευταία 98 χρόνια).
•1588 συντηρημένα ορθόλογα γονίδια χρησιμοποιήθηκαν από το κάθε στέλεχος, για να γίνει το φυλογενετικό δένδρο.

Επιδημία χολέρας στην Αϊτή 2010

ORIGIN OF CHOLERA OUTBREAK STRAIN IN HAITI

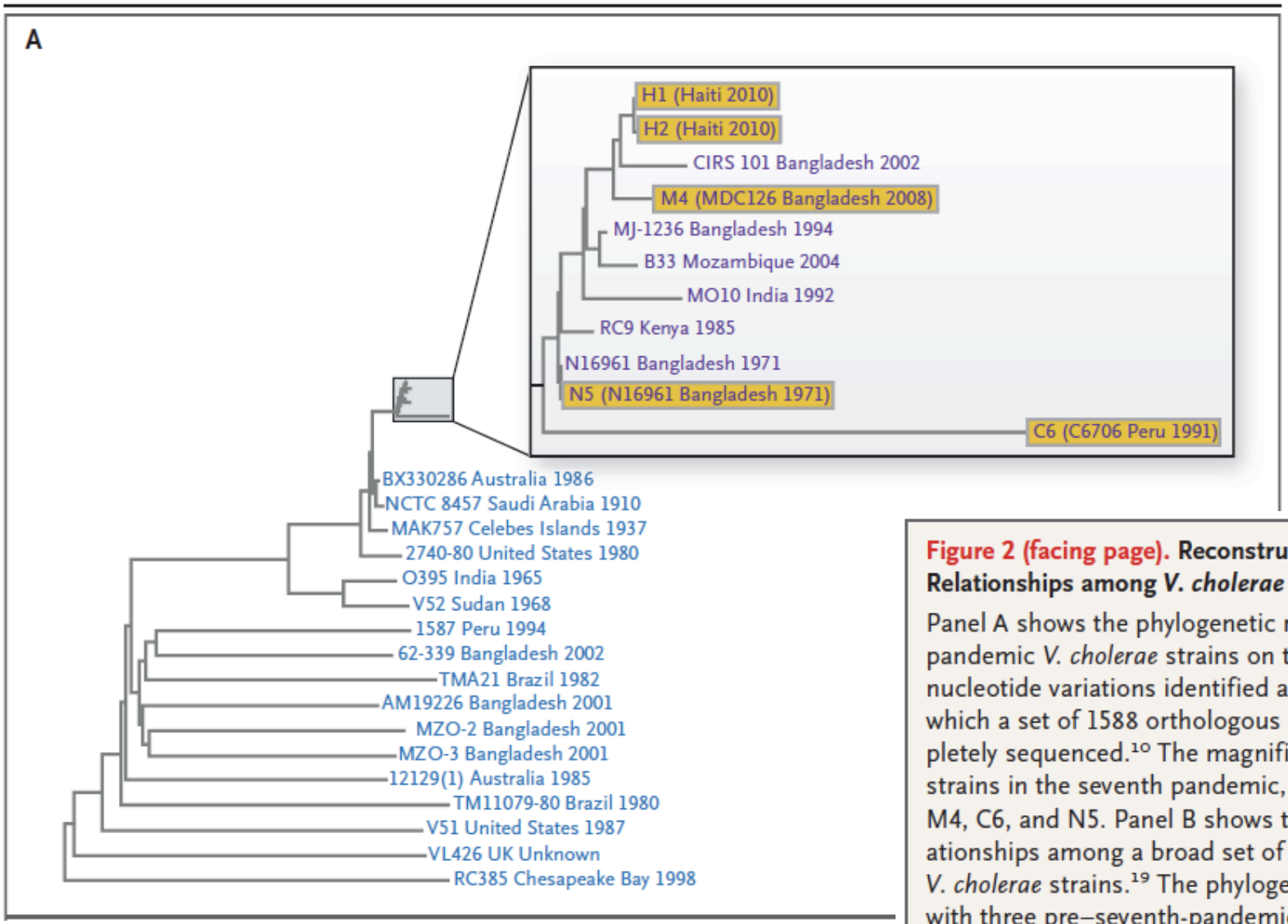


Figure 2 (facing page). Reconstructing Phylogenetic Relationships among *V. cholerae* Strains.

Panel A shows the phylogenetic relationships among pandemic *V. cholerae* strains on the basis of single-nucleotide variations identified among all strains for which a set of 1588 orthologous genes has been completely sequenced.[10] The magnified inset represents strains in the seventh pandemic, including H1, H2, M4, C6, and N5. Panel B shows the phylogenetic relationships among a broad set of seventh-pandemic *V. cholerae* strains.[19] The phylogenetic tree is rooted with three pre–seventh-pandemic strains.

Επιδημία χολέρας στην Αϊτή 2010

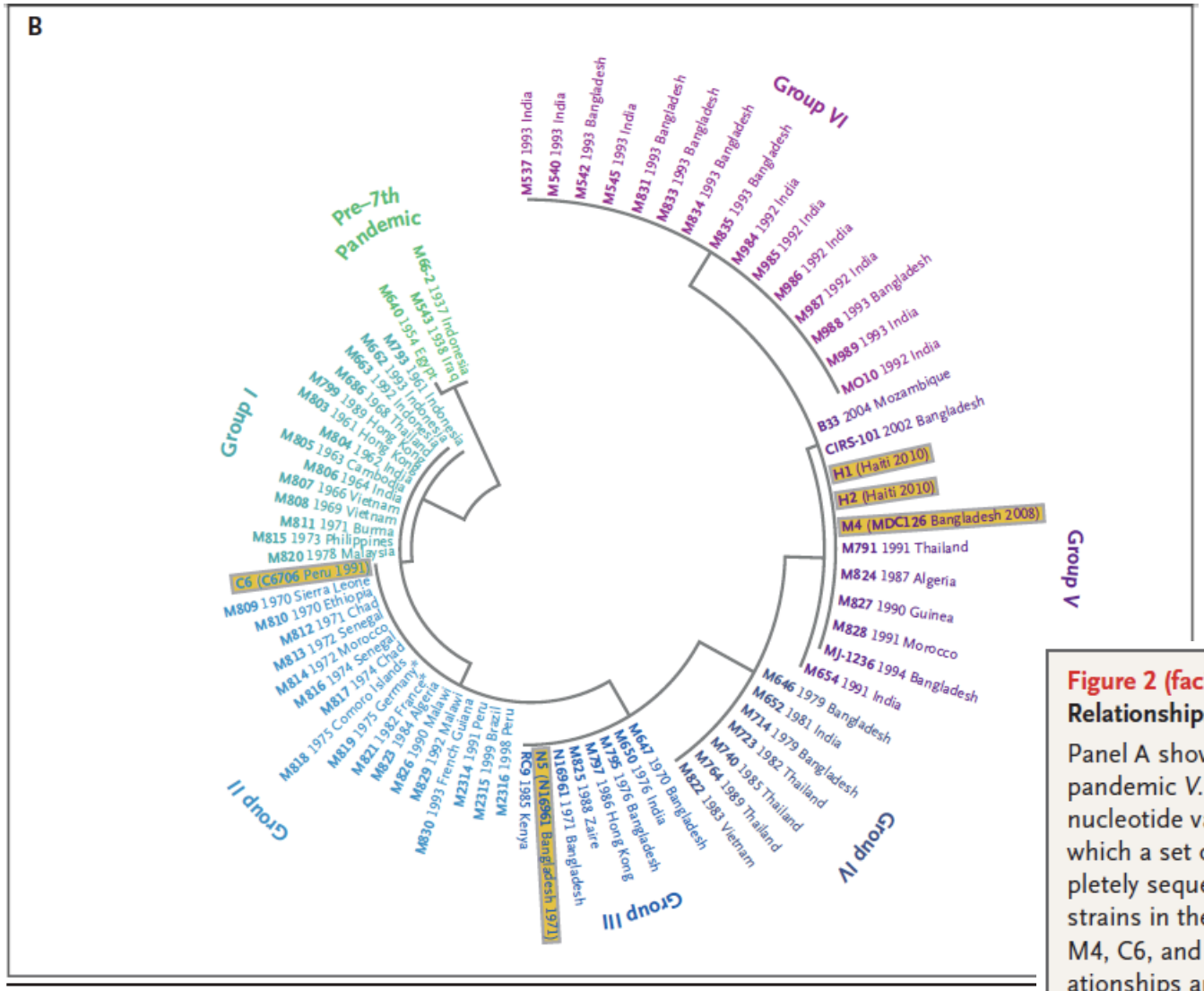


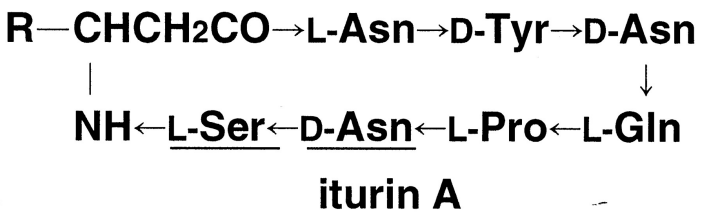
Figure 2 (facing page). Reconstructing Phylogenetic Relationships among *V. cholerae* Strains.

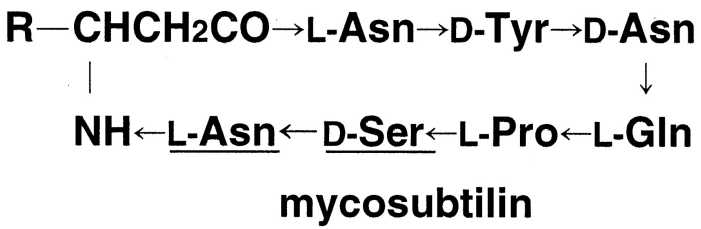
Panel A shows the phylogenetic relationships among pandemic *V. cholerae* strains on the basis of single-nucleotide variations identified among all strains for which a set of 1588 orthologous genes has been completely sequenced.[10] The magnified inset represents strains in the seventh pandemic, including H1, H2, M4, C6, and N5. Panel B shows the phylogenetic relationships among a broad set of seventh-pandemic *V. cholerae* strains.[19] The phylogenetic tree is rooted with three pre–seventh-pandemic strains.

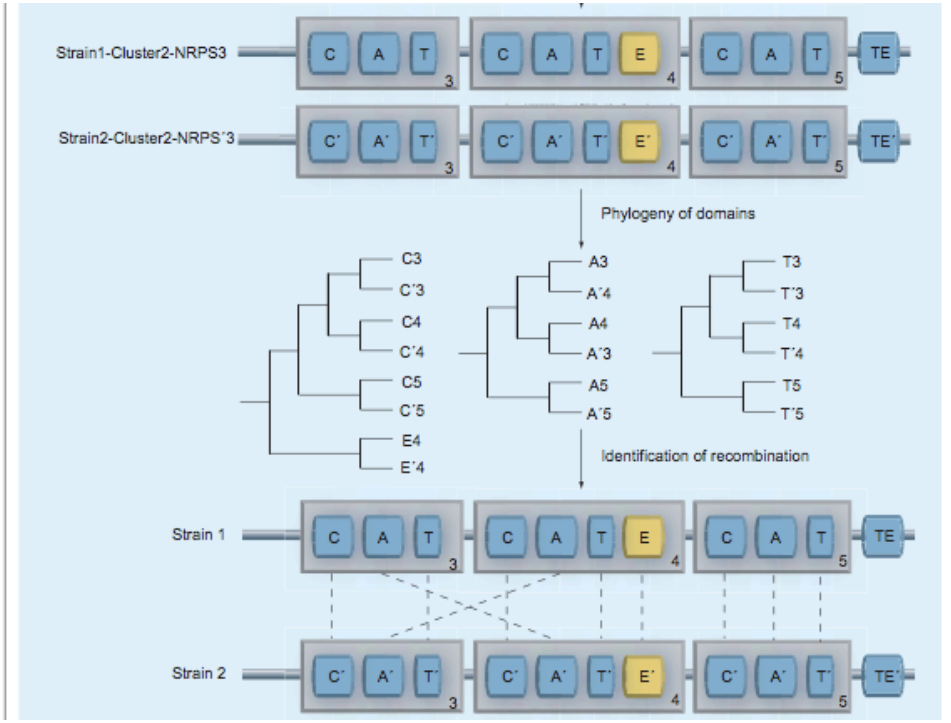
Εξέλιξη αντιβιοτικών και τοξινών

Αλλαγές σε πρωτεϊνες που συνθέτουν αντιβιοτικά (NRPS)
Bacillus subtilis

Strain RB14: Iturin A
Strain ATCC6633: mycosubtillin
μυκητοκτόνα





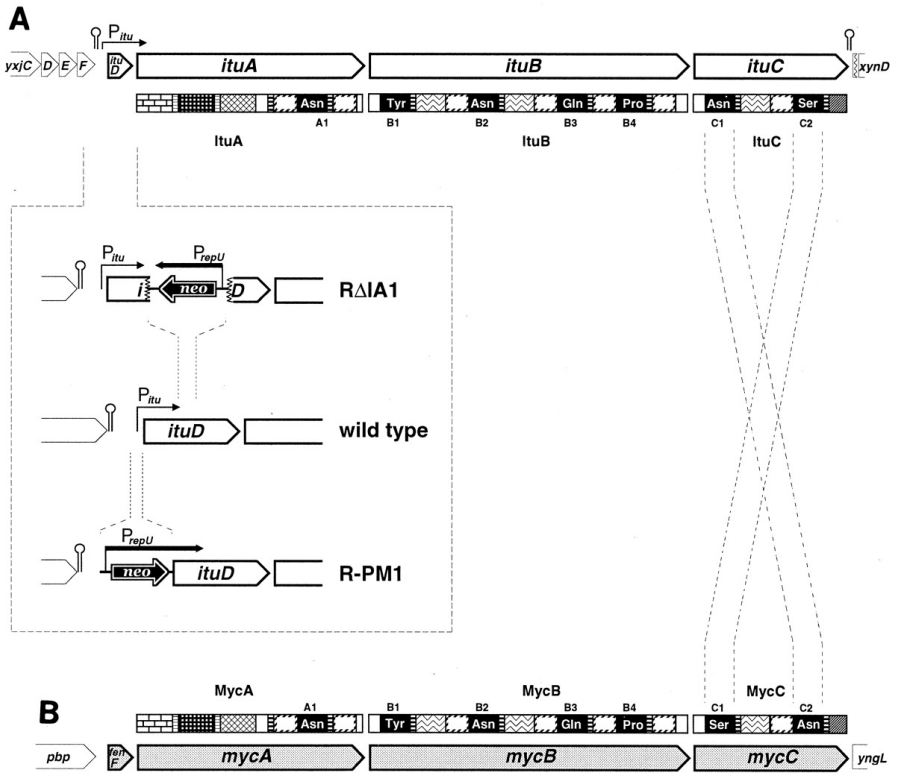


Εξέλιξη αντιβιοτικών και τοξινών

Αλλαγές σε πρωτεϊνες που συνθέτουν αντιβιοτικά (NRPS)
Bacillus subtilis

Strain RB14: Iturin A

Strain ATCC6633: mycosubtillin



Εφαρμογές

Ανίχνευση οργανισμών
-
Μεταγενωμική

Μεταγενωμική

- Παράλληλη ανίχνευση όλων των οργανισμών (μικροβιακών) που απαρτίζουν την υπό μελέτη οικολογική κοινότητα.

- Υπάρχει προοπτική να χρησιμοποιηθεί για περιβαλλοντικές μελέτες/αναλύσεις/ παρακολούθηση (σε βάση ρουτίνας), όταν το κόστος αλληλούχισης (ή μικροσυστοιχιών) μειωθεί περισσότερο.

- Πλεονέκτημα: Δεν χρειάζεται να καλλιεργηθούν
  - Κλινικά δείγματα
  - Περιβαλλοντικά δείγματα

Genome assembly

Key steps in de novo assembly

1. Find reads that overlap by a specified number of bases (the k-mer size)

2. Merge overlapping, "good" reads into longer contigs

3. Link contigs to form scaffolds using paired-end information

Diagrams from S. Batzoglou, Stanford

Metagenomics

- Environmental Protection Agency (EPA)
- The Clean Water Act: **Fecal Source Identification.**
- **Απόσπασμα από Microbial Source tracking guide Document (Ιούνιος 2005).**
- "The Clean Water Act establishes that the states must adopt water quality standards that are compatible with pollution control programs to reduce pollutant discharges into waterways. In many cases the standards have been met by the significant reduction of loads from point sources under the National Pollutant Discharge Elimination System (NPDES). Point sources are defined as "any discernable, confined and discrete convey- ance, including but not limited to any pipe, ditch or concentrated animal feeding operation from which pollutants are or may be discharged". However, more than 30 years after the Clean Water Act was implemented, a significant fraction of the U.S. rivers, lakes, and estuaries continue to be classified as failing to meet their designated uses due to the high levels of fecal bacteria (USEPA, 2000b). As a consequence, protection from fecal microbial contamination is one of the most important and difficult challenges facing environmental scientists trying to safeguard waters used for:
  - recreation (primary and secondary contact),
  - public water supplies,
  - propagation of fish and shellfish.

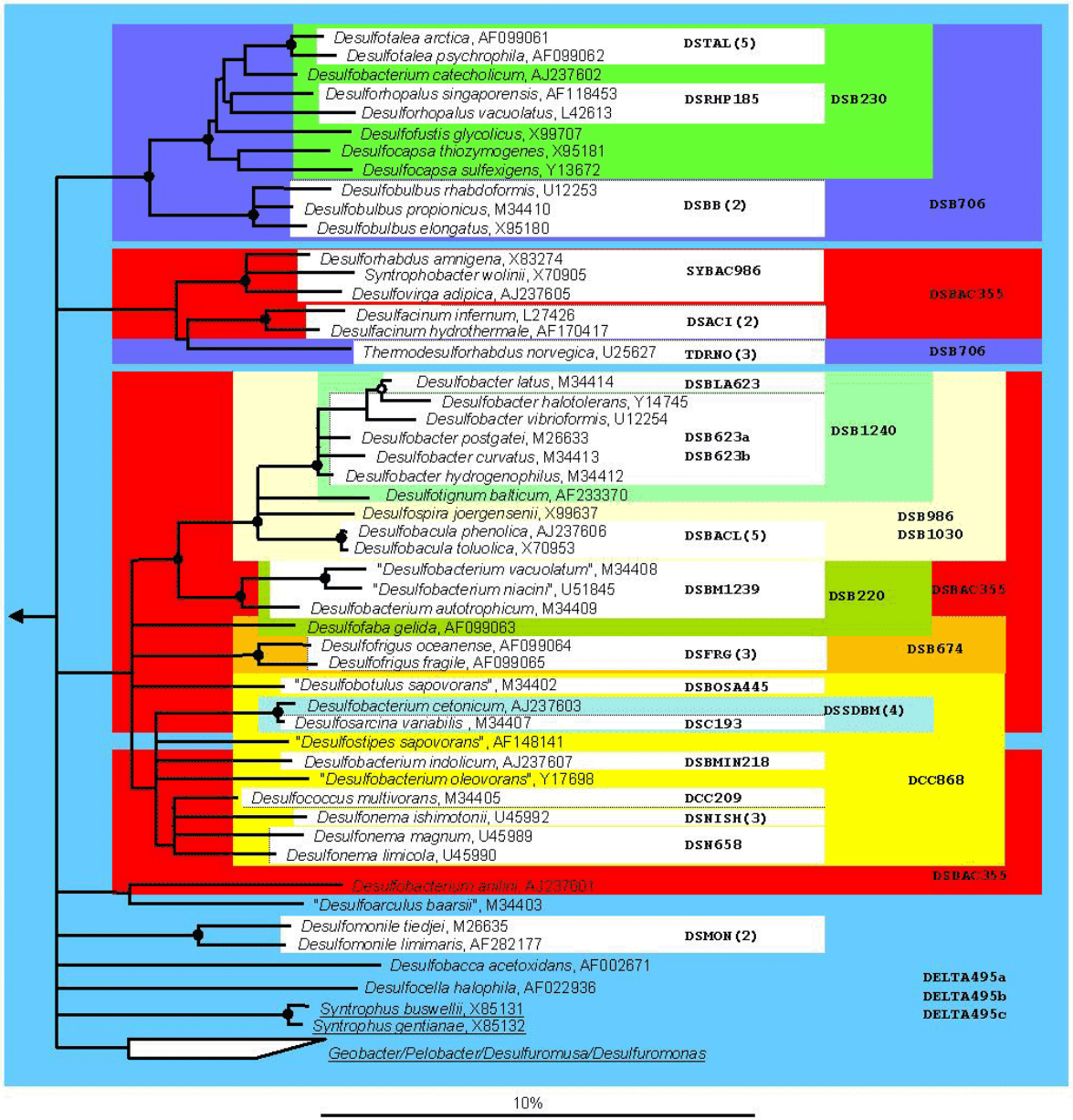
- Fecally contaminated waters not only harbor pathogens and pose potential high risks to human health, but they also result in **significant economic loss** due to closure of shellfish harvesting areas and recreational beaches."

Phylochip

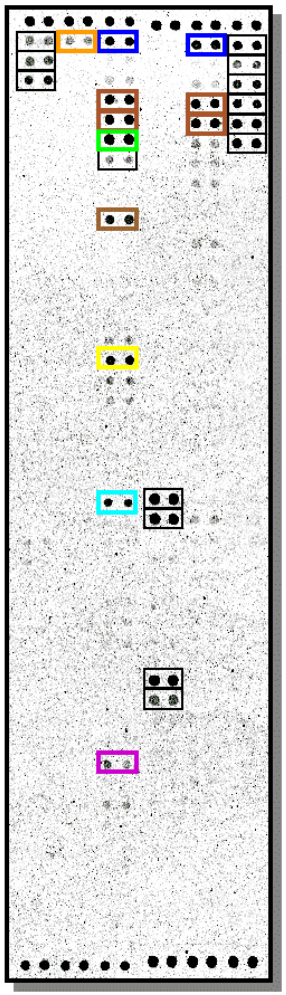
- Affymetrix
- Μικροσυστοιχία που βασίζεται στον υβριδισμό κομματιών DNA (από το περιβαλλοντικό δείγμα-μίγμα) πάνω σε καθηλωμένα (στο chip) probes.
  - Probes βασίζονται σε RNA γονίδια.
    - RNA γονίδια αποτελούνται από βαθιά συντηρημένες και από λίγο συντηρημένες περιοχές. Στον σχεδιασμό του chip, επιλέγουμε την περιοχή ανάλογα με το βαθμό διαχωρισμού που επιθυμούμε
      - Βαθιά συντηρημένες περιοχές για διαχωρισμό μεταξύ εξελικτικά απομακρυσμένων οργανισμών.
      - Υψηλά μεταβλητές περιοχές για διαχωρισμό μεταξύ εξελικτικά κοντινών συγγενικών οργανισμών (π.χ. Στελέχη ενός μικροβίου)

Phylochip

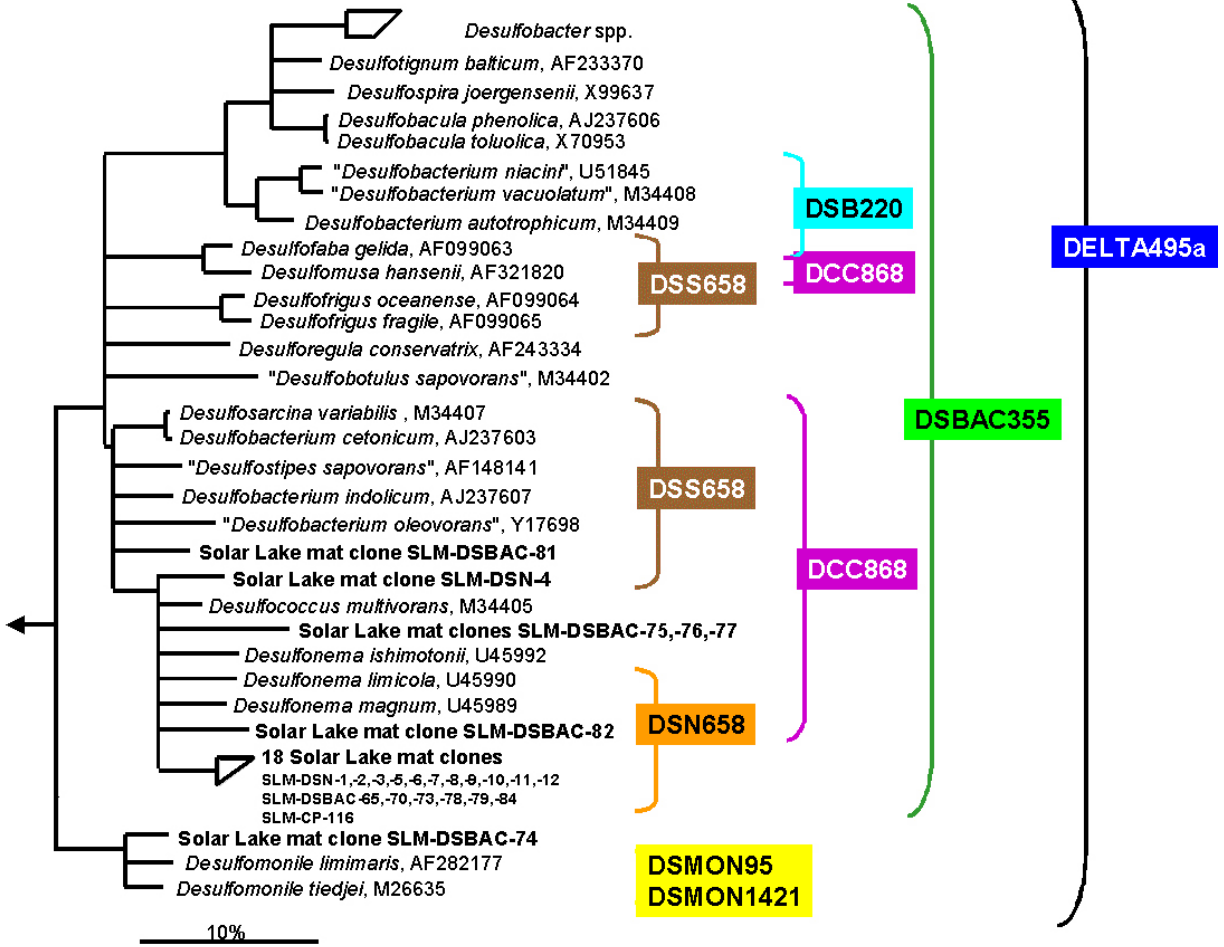
Fig. 1. Applied multiple probe concept. 16S rRNA-based phylogenetic consensus tree of all recognized sulfate-reducing bacteria of the orders "Desulfobacterales" and "Syntrophobacterales" showing examplarily the hierarchical and parallel specificity of oligonucleotide probes.



Phylochip




Cyanobacterial Mat


Desulfobacter spp.
Desulfotignum balticum, AF233370
Desulfospira joergensenii, X99637
Desulfobacula phenolica, AJ237606
Desulfobacula toluolica, X70953
"Desulfobacterium niacini", U51845
"Desulfobacterium vacuolatum", M34408
Desulfobacterium autotrophicum, M34409
Desulfofaba gelida, AF099063
Desulfomusa hansenii, AF321820
Desulfofrigus oceanense, AF099064
Desulfofrigus fragile, AF099065
Desulforegula conservatrix, AF243334
"Desulfobotulus sapovorans", M34402
Desulfosarcina variabilis , M34407
Desulfobacterium cetonicum, AJ237603
"Desulfostipes sapovorans", AF148141
Desulfobacterium indolicum, AJ237607
"Desulfobacterium oleovorans", Y17698
Solar Lake mat clone SLM-DSBAC-81
Solar Lake mat clone SLM-DSN-4
Desulfococcus multivorans, M34405
Solar Lake mat clones SLM-DSBAC-75,-76,-77
Desulfonema ishimotonii, U45992
Desulfonema limicola, U45990
Desulfonema magnum, U45989
Solar Lake mat clone SLM-DSBAC-82
18 Solar Lake mat clones
SLM-DSN-1,-2,-3,-5,-6,-7,-8,-9,-10,-11,-12
SLM-DSBAC-65,-70,-73,-78,-79,-84
SLM-CP-116
Solar Lake mat clone SLM-DSBAC-74
Desulfomonile limimaris, AF282177
Desulfomonile tiedjei, M26635
10%
DSB220
DCC868
DSS658
DSS658
DCC868
DSN658
DSMON95
DSMON1421
DSBAC355
DELTA495a

In vitro διαγνωστικά τεστ που βασίζονται σε μικροσυστοιχίες

FDA: In Vitro Diagnostic Multivariate Index Assays (IVDMIAs)

- FDA's **In Vitro Diagnostic Product Database**
- http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfivd/index.cfm

- http://www.ivdtechnology.com/article/exploring-fda-approved-ivdmias

- Some IVDMIAs are laboratory-developed tests (LDTs). LDTs are tests that are developed by a single clinical laboratory for use only in that laboratory.

- http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm079148.htm

- IVDMIAs raise significant issues of safety and effectiveness. These types of tests are developed based on observed correlations between multivariate data and clinical outcome, such that the clinical validity of the claims is not transparent to patients, laboratorians, and clinicians who order these tests. Additionally, IVDMIAs frequently have a high risk intended use. FDA is concerned that patients are relying upon IVDMIAs with high risk intended uses to make critical healthcare decisions when FDA has not ensured that the IVDMIA has been clinically validated and the healthcare practitioners are unable to clinically validate the test themselves. Therefore, there is a need for FDA to regulate these devices to ensure that the IVDMIA is safe and effective for its intended use.

Mammaprint - Tissue of origin

- http://www.ivdtechnology.com/article/exploring-fda-approved-ivdmias
- **MammaPrint.**
  The first IVDMIA, the MammaPrint system, made by Agendia Inc., is a qualitative IVD test service performed in a single lab outside the United States using a 70-gene expression profile of fresh frozen breast cancer tissue samples to assess a breast cancer patient¬s risk for distant metastasis. FDA approved MammaPrint in February 2007 under de novo classification procedures.

- **Tissue of Origin Test**
  In July 2008, the Tissue of Origin Test, made by Pathwork Diagnostics, was cleared. This microarray RNA profiling test is to be used on clinical, formalin-fixed, paraffin-embedded (FFPE) biopsy tissue to aid in the classification of the origin of the tumor tissue. In June 2010 a second clearance introduced a different specimen and specimen-preparation method, and the algorithm for analysis of the expression data to create a diagnostics report and interpretation. The test uses microarray technology by Affymetrix Inc. and advanced analytics to measure the gene-expression patterns of challenging tumors, including metastatic, poorly differentiated, and undifferentiated cancer. It is intended to measure the degree of similarity between the RNA expression patterns in a patient’s tumor tissue with the RNA expression patterns in a database of fifteen known tumor types.

Εφαρμογές στην Τοξικολογία

Εφαρμογές στην τοξικολογία/ τοξικογενωμική

- Μέτρηση της γονιδιακής έκφρασης μετά από έκθεση σε τοξικό παράγοντα μπορεί να δείξει τον μοριακό μηχανισμό δράσης του παράγοντα.
- Μπορεί να αποτελέσει μοναδική μοριακή υπογραφή του συγκεκριμένου τοξικού παράγοντα, για μελλοντική ανίχνευσή του.
  - Ομαδοποίηση τοξικών παραγόντων με κοινή δράση, με βάση την ομοιότητα των μοριακών προφιλ τους

Μοριακό προφίλ τοξικότητας

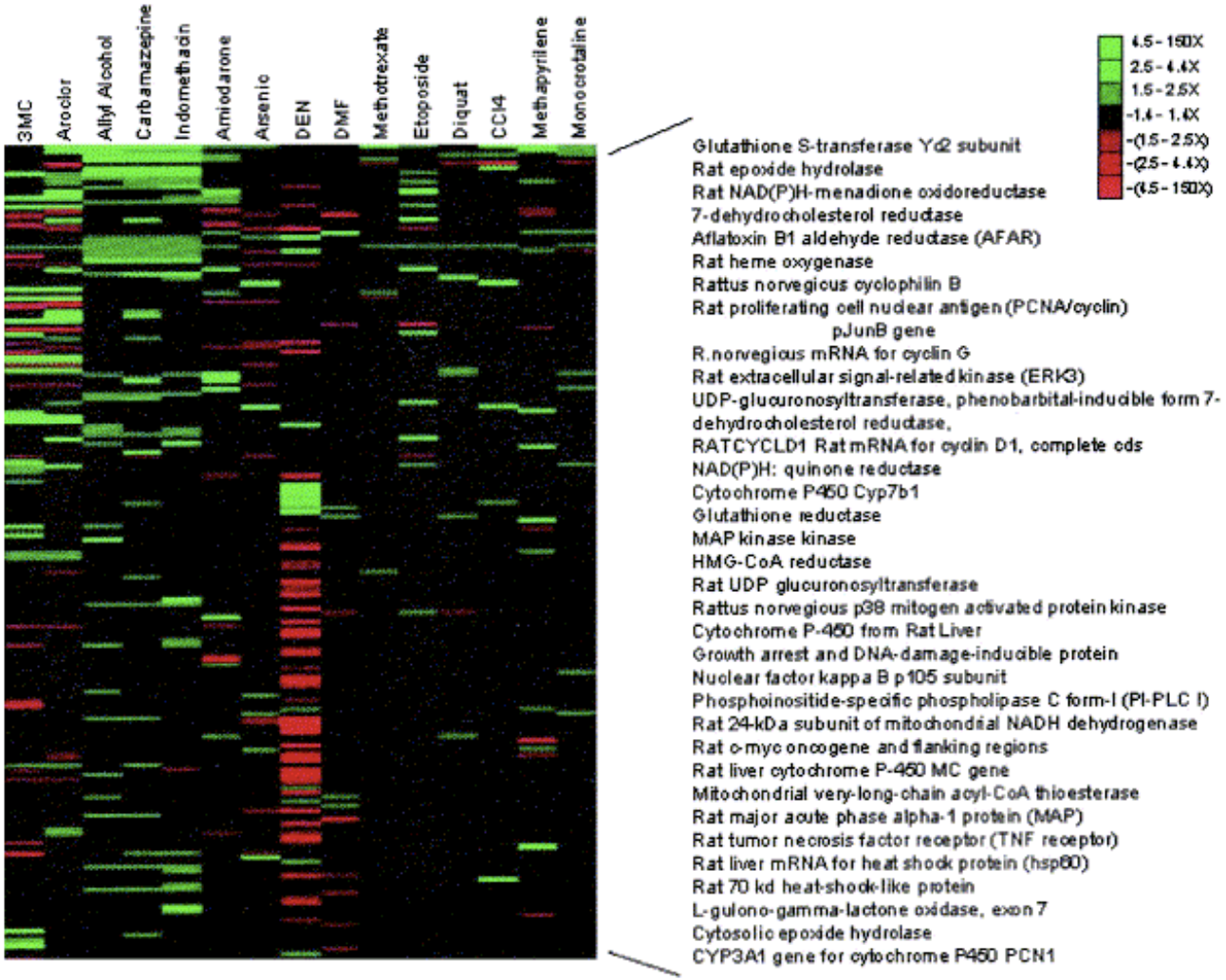


Fig. 2. Graph showing the gene changes occurring in livers from rats treated with the 15 known hepatotoxins. A total of 179 genes were shown to be regulated at least two-fold by at least one compound. Some of these genes are shown to the right of the figure.

Μοριακό προφίλ τοξικότητας

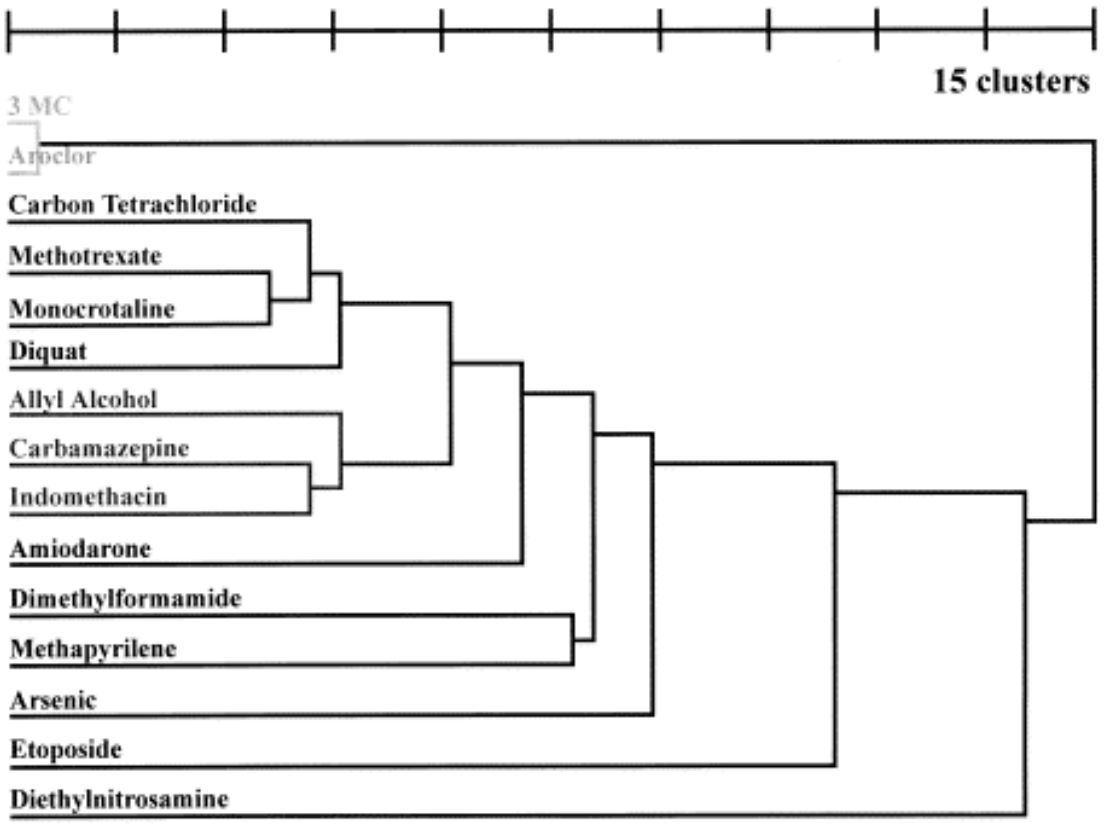


Fig. 3. Dendrogram showing the clustering of the hepatotoxins based on gene regulation. The clustering was hierarchical using correlation as the distance (see Section 2).

Hierarchical cluster analysis showed a close association in gene expressional responses between aroclor 1254 and 3-methylcholanthrene.

Environmental Protection Agency (EPA)

- http://www.epa.gov/osa/spc/pdfs/genomics.pdf

- **Genomics** methodologies are expected to provide valuable insights for **evaluating how environmental stressors affect cellular/tissue functions** and bow changes in gene expression may relate to adverse effects.

- However, the relationships between changes in gene expression and adverse effects are **unclear at this time** and may likely be difficult to elucidate.

- Nonetheless, EPA believes that some of these changes **may prove to be predictive of subsequent adverse effects**. Changes in gene expression can be informative when a weight-of-evidence approach for human and ecological health assessments is performed, particularly when used to explore the possible link between exposure. mechanism(s) of action, and adverse effects. In addition, genomics information may be useful to EPA in setting priorities, in ranking of chemicals for further testing, and in supporting possible regulatory actions. While genomics data may be considered in decision-making at this time, these data alone are insufficient as a basis for decisions. For assessment purposes, **EPA will consider genomics information on a case-by-case basis**. Before such information can be accepted and used, agency review will be needed to determine adequacy regarding the quality, representativeness, and reproducibility of the data.

Βάσεις Δεδομένων

Βάσεις Δεδομένων: Εισαγωγή

Χρησιμοποιούνται για:

- Οργάνωση
- Αποθήκευση
- Επεξεργασία
- Αναζήτηση/επαναπόκτηση

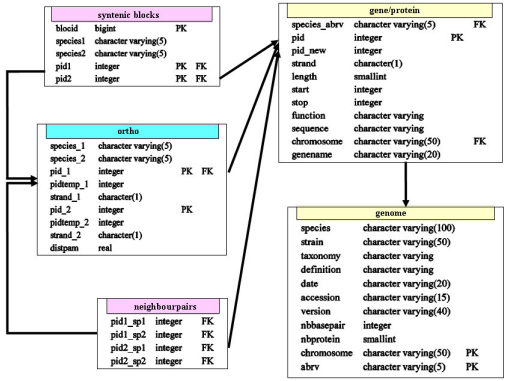
της βιολογικής πληροφορίας

```
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 cagggggcgg aatgaaagcg
//              database symbol for end of sequence
```

Figure 2.5. GenBank DNA sequence entry.

Κύρια είδη:

Επίπεδης οργάνωσης (Flat-files:) Το ποιό απλό είδος. Ουσιαστικά είναι κατάλογοι



Σχεσιακές βάσεις. Πιο περίπλοκες και πλέον πολύ διαδεδομένες . Π.χ., SQL. Η πληροφορία οργανώνεται σε πίνακες που σχετίζονται μεταξύ τους. Έτσι αποφεύγεται η επανάληψη και συσσώρευση δεδομένων

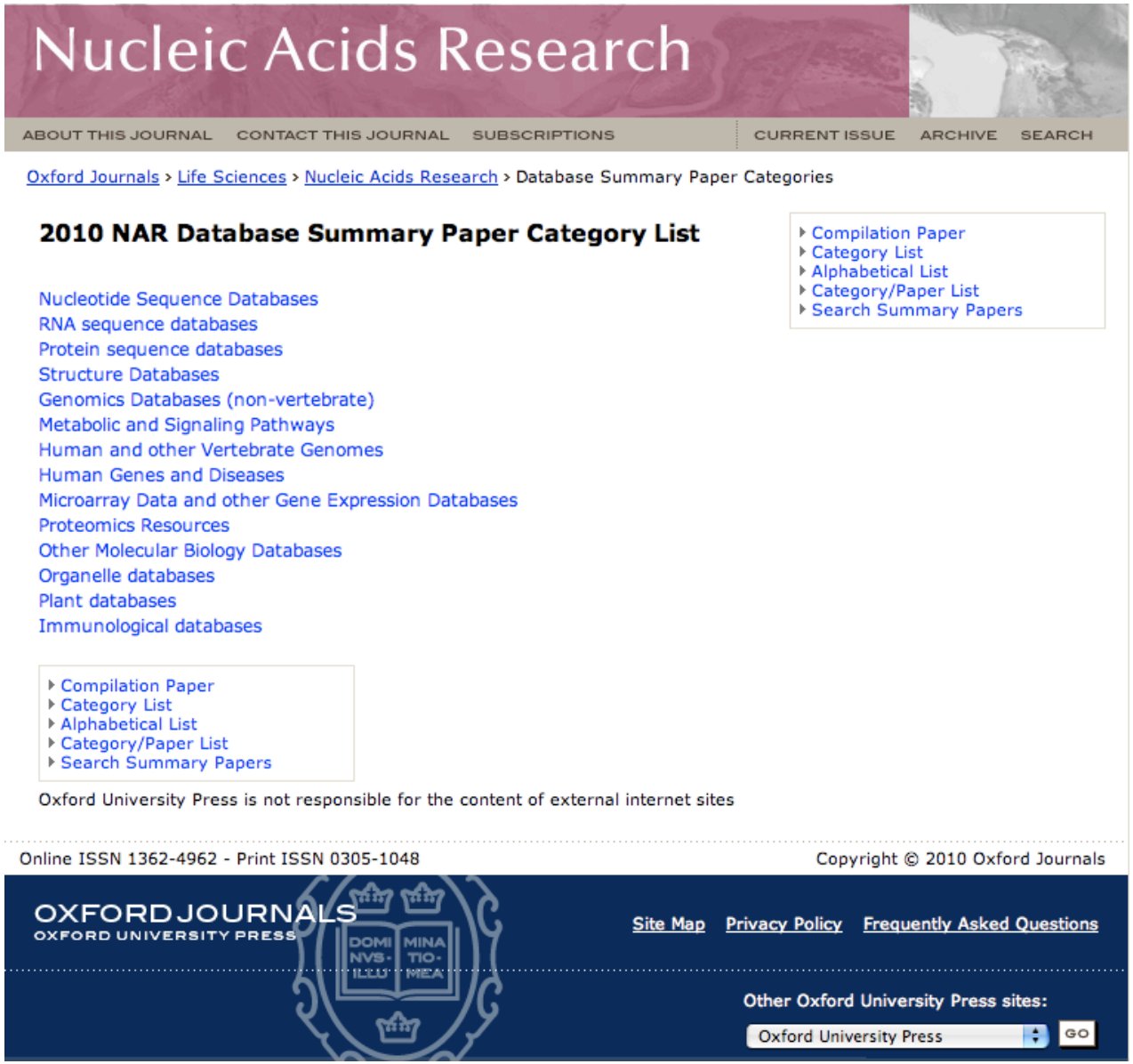
Αντικειμενοστρεφείς βάσεις κ.α.

Διακρίνονται κυρίως σε αρχειακές/πρωτεύοντες και δευτερεύοντες

Στις αρχειακές γίνεται κατάθεση δεδομένων ενώ στις δευτερεύοντες τα δεδομένα είναι περαιτέρω επεξεργασμένα/σχολιασμένα/ αλληλοσυνδεδεμένα

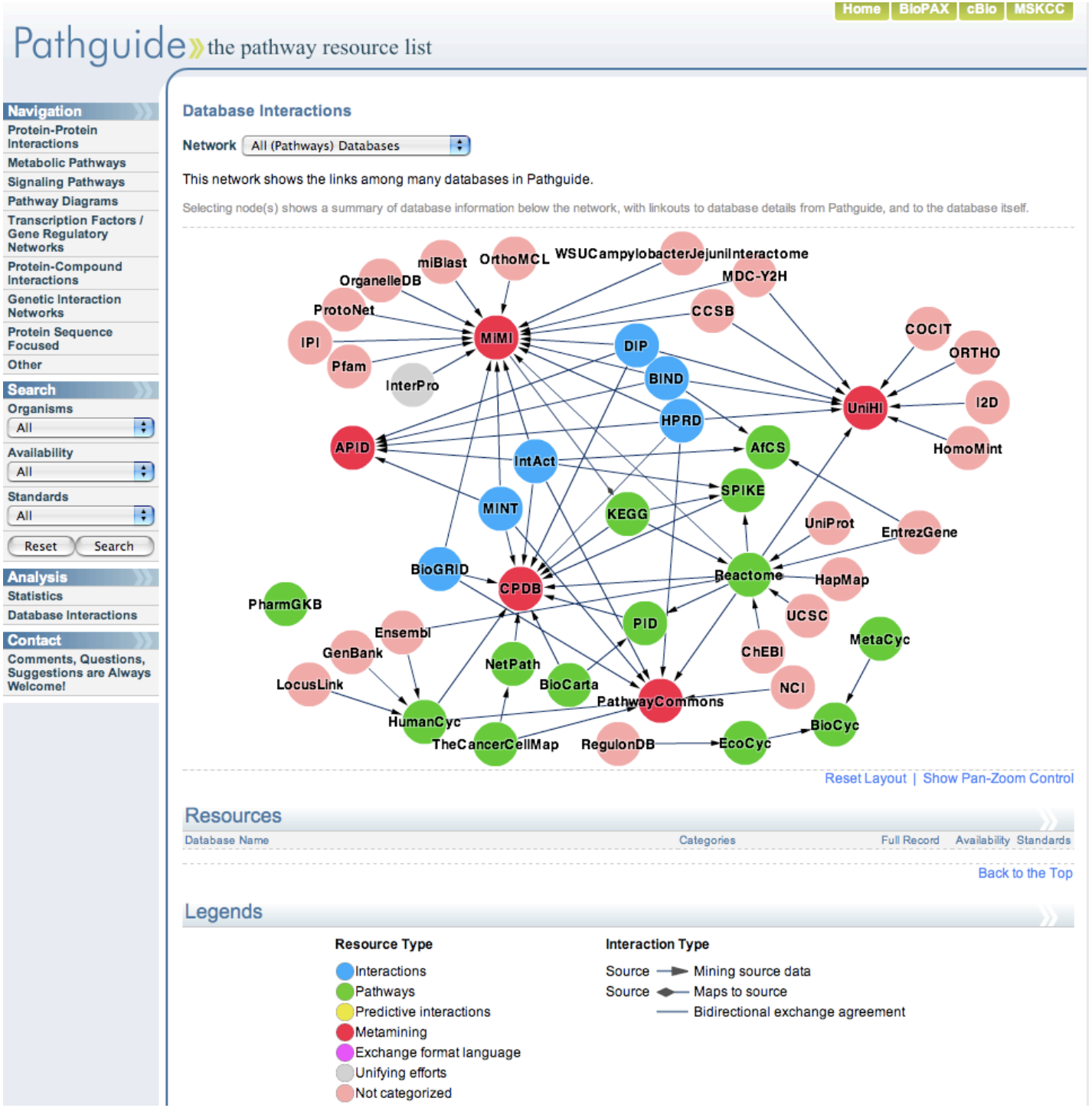
Ετήσιος κατάλογος ΒΔ

- Κάθε Ιανουάριο στο Nucleic Acids Research (Special database issue)
- 2010: 58 νέες και 73 ανανεωμένες
- Σύνολο: 1230
- 5% ετήσια ανάπτυξη
- Επίσης υπάρχει το περιοδικό Database: the journal of biological databases and curation

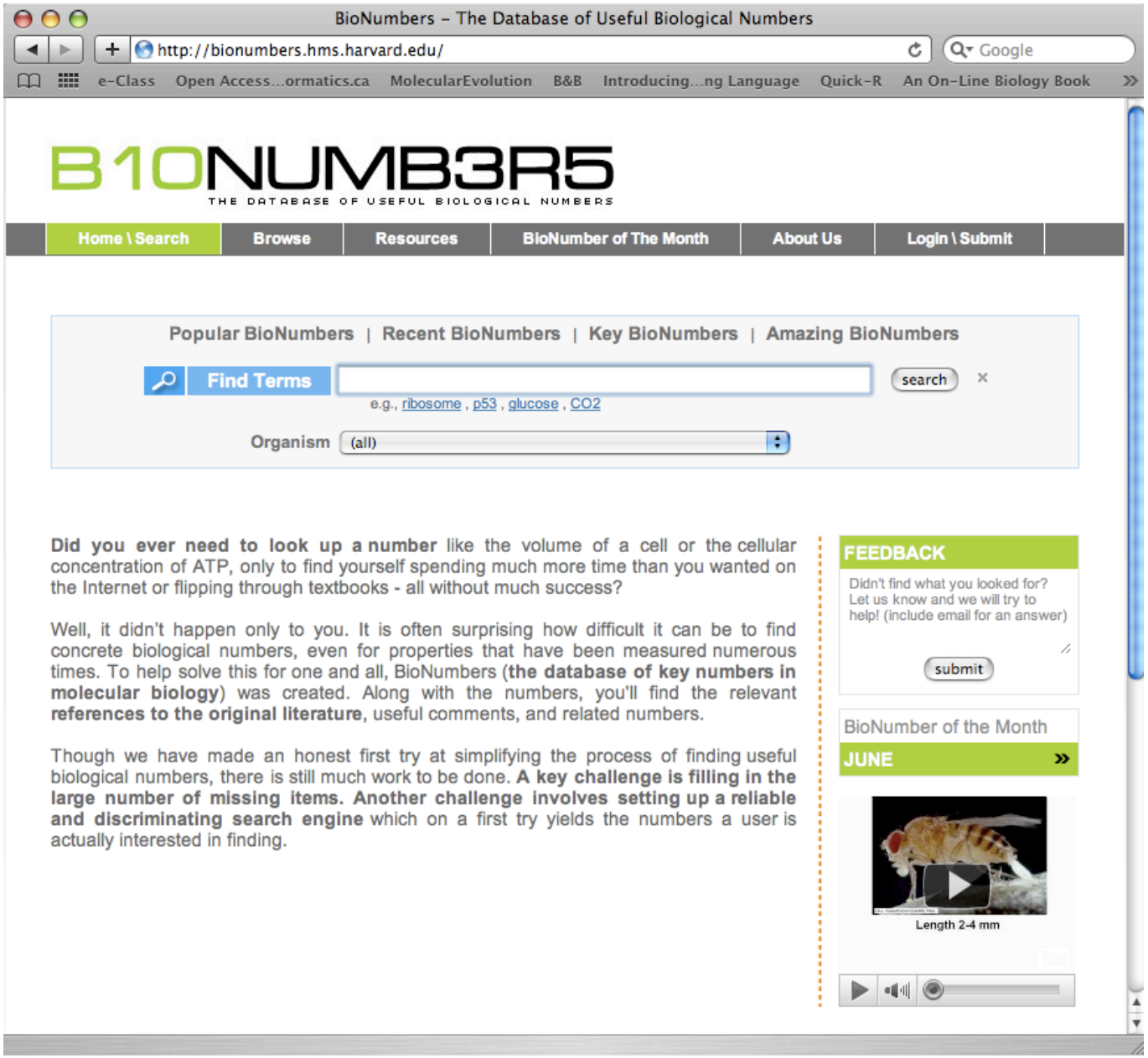


Κατάλογος με ΒΔ: Pathguide

- http://www.pathguide.org/



Bionumbers


BioNumbers – The Database of Useful Biological Numbers
http://bionumbers.hms.harvard.edu/
Google
e-Class
Open Access...ormatics.ca
MolecularEvolution
B&B
Introducing...ng Language
Quick-R
An On-Line Biology Book
B10NUMB3R5
THE DATABASE OF USEFUL BIOLOGICAL NUMBERS
Home \ Search
Browse
Resources
BioNumber of The Month
About Us
Login \ Submit
Popular BioNumbers | Recent BioNumbers | Key BioNumbers | Amazing BioNumbers
Find Terms
search
e.g., ribosome , p53 , glucose , CO2
Organism
(all)
Did you ever need to look up a number like the volume of a cell or the cellular concentration of ATP, only to find yourself spending much more time than you wanted on the Internet or flipping through textbooks - all without much success?
Well, it didn't happen only to you. It is often surprising how difficult it can be to find concrete biological numbers, even for properties that have been measured numerous times. To help solve this for one and all, BioNumbers (the database of key numbers in molecular biology) was created. Along with the numbers, you'll find the relevant references to the original literature, useful comments, and related numbers.
Though we have made an honest first try at simplifying the process of finding useful biological numbers, there is still much work to be done. A key challenge is filling in the large number of missing items. Another challenge involves setting up a reliable and discriminating search engine which on a first try yields the numbers a user is actually interested in finding.
FEEDBACK
Didn't find what you looked for? Let us know and we will try to help! (include email for an answer)
submit
BioNumber of the Month
JUNE
Length 2-4 mm

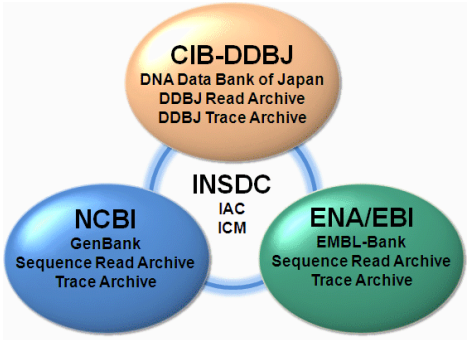
Βάσεις νουκλεοτιδικών δεδομένων (Ι)

- Αρχειακές ΒΔ για νουκλεοτιδικές αλληλουχίες:
  - EMBL-BANK. European Nucleotide Archive (ENA), EBI. Hinxton, UK.
  - GENBANK. NCBI, NIH. Bethesda, USA
  - DNA databank of Japan (DDBJ). National institute of Genetics,.Mishima, JP

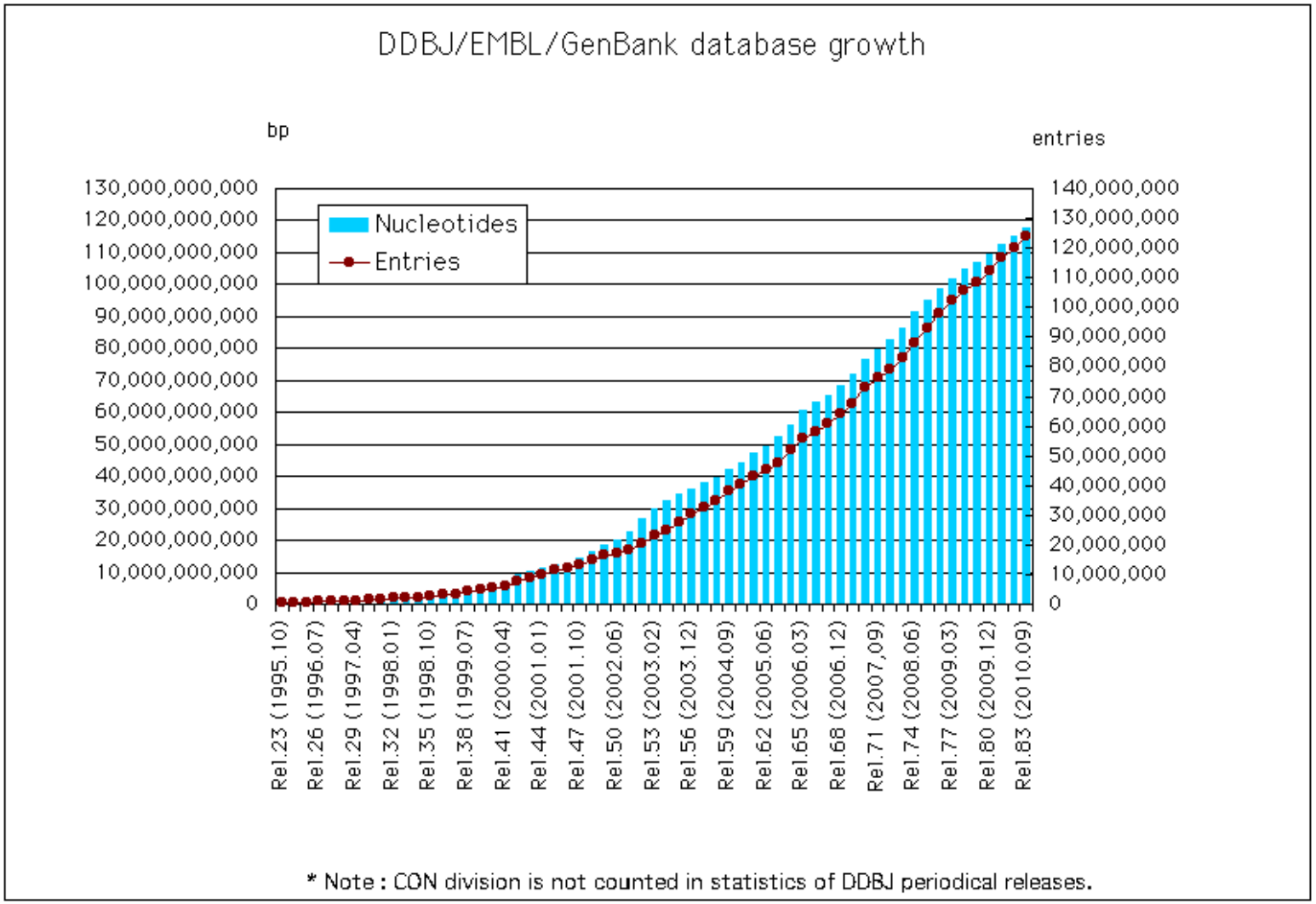
- Η ακολουθία κατατίθεται σε μία από τις ΒΔ, η οποία έχει και τη δυνατότητα να την αναθεωρήσει (μόνο αυτή, για αποτροπή 'συγκρούσεων')

- Και οι 3 ΒΔ ανήκουν στο International nucleotide sequence database collection (INSDC). Κάθε μέρα ανταλλάσουν δεδομένα. Η ίδια ακολουθία Χ3. Νέα έκδοση ανά δίμινο.

- 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 2ης γενιάς (Read Archive) (454, Solexa, Solid, Helicos)



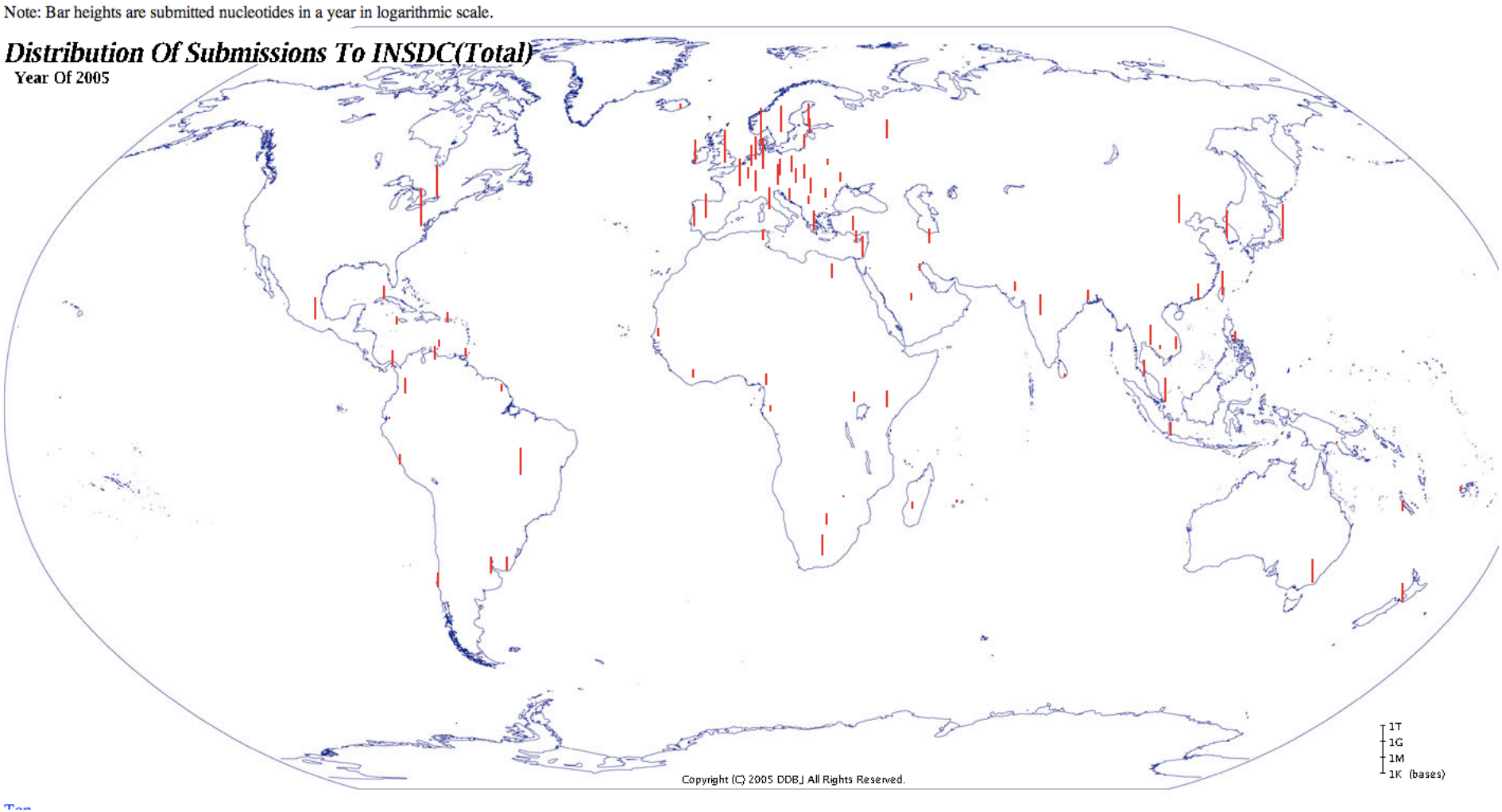
Βάσεις νουκλεοτιδικών δεδομένων (ii)



Πάνω από 100 Δις βάσεις στο INSDC. Σύντομα αναμένεται πληθώρα προσωπικών γενωμάτων.

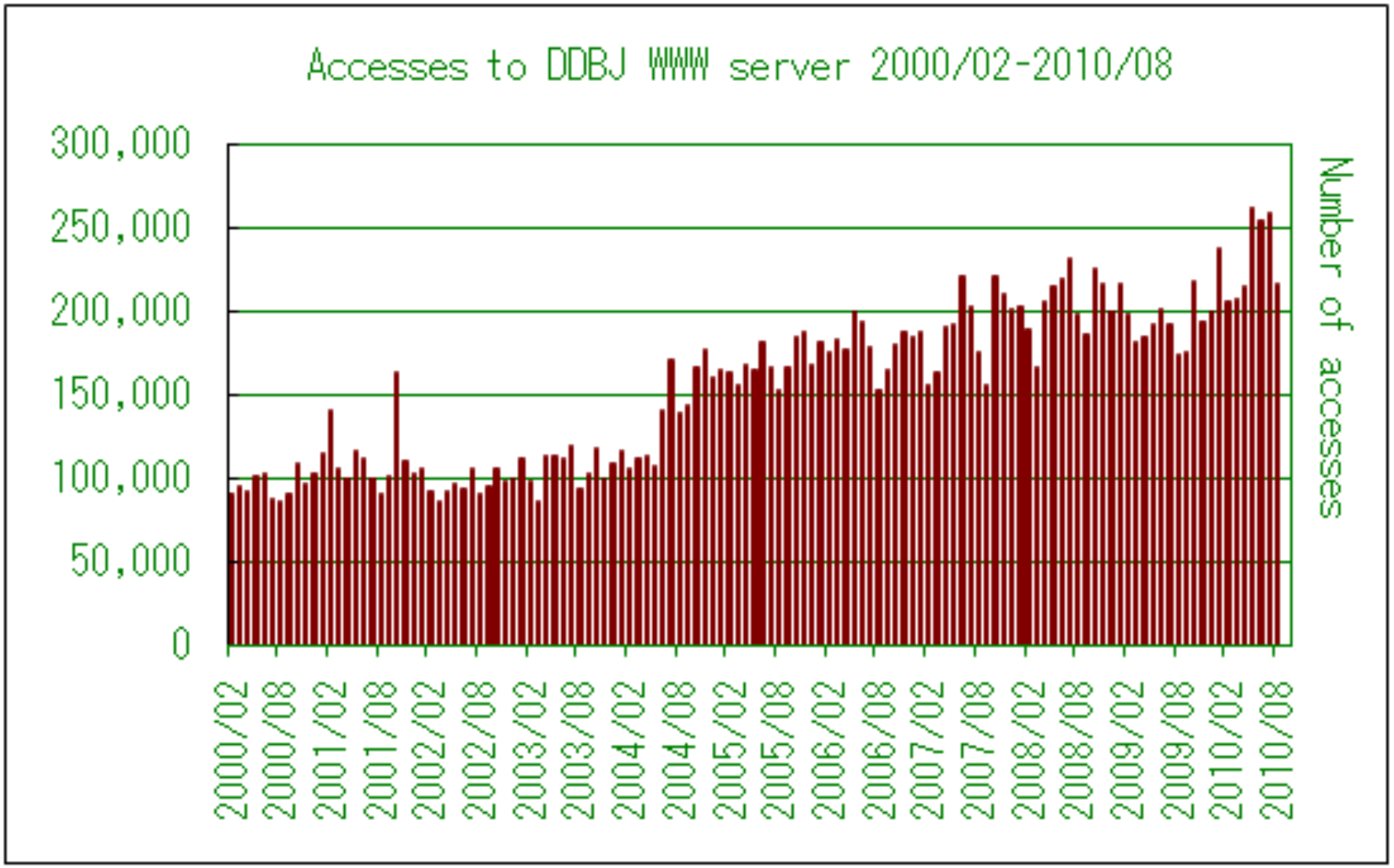
Εγείρονται προβληματισμοί για την αποθήκευση όλων αυτών των δεδομένων!

Βάσεις νουκλεοτιδικών δεδομένων (III)



2005: Ελλάδα: 2,7 MB USA: 7GB. Από DDBJ

Βάσεις νουκλεοτιδικών δεδομένων (iv)



Βάσεις νουκλεοτιδικών δεδομένων. EMBL format

```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC   fabids; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN   [5]
RP   1-1859
RX   DOI; 10.1007/BF00039495
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL   Hughes M.A., University of Newcastle Upon Tyne, Medical School, Newcastle
RL   Upon Tyne, NE2 4HH, UK
XX
FH   Key             Location/Qualifiers
FH
```

Βάσεις νουκλεοτιδικών δεδομένων. EMBL format

```
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
     cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
     tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
     aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
     tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta       300
     caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc       360
     ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaatcaa       420
     atattacaac aaccttatca acgaactatt ggctaacggt atacaaccat ttgtaactct       480
     ttttcattgg gatcttcccc aagtcttaga agatgagtat ggtggtttct taaactccgg       540
     tgtaataaat gattttcgag actatacgga tctttgcttc aaggaatttg gagatagagt       600
     gaggtattgg agtactctaa atgagccatg ggtgtttagc aattctggat atgcactagg       660
     aacaaatgca ccaggtcgat gttcggcctc caacgtggcc aagcctggtg attctggaac       720
     aggaccttat atagttacac acaatcaaat tcttgctcat gcagaagctg tacatgtgta       780
     taagactaaa taccaggcat atcaaaaggg aaagataggc ataacgttgg tatctaactg       840
     gttaatgcca cttgatgata atagcatacc agatataaag gctgccgaga gatcacttga       900
     cttccaattt ggattgttta tggaacaatt aacaacagga gattattcta agagcatgcg       960
     gcgtatagtt aaaaaccgat tacctaagtt ctcaaaattc gaatcaagcc tagtgaatgg      1020
     ttcatttgat tttattggta taaactatta ctcttctagt tatattagca atgccccttc      1080
     acatggcaat gccaaaccca gttactcaac aaatcctatg accaatattt catttgaaaa      1140
     acatgggata cccttaggtc caagggctgc ttcaatttgg atatatgttt atccatatat      1200
     gtttatccaa gaggacttcg agatcttttg ttacatatta aaaataaata taacaatcct      1260
     gcaatttttca atcactgaaa atggtatgaa tgaattcaac gatgcaacac ttccagtaga      1320
     agaagctctt ttgaatactt acagaattga ttactattac cgtcacttat actacattcg      1380
     ttctgcaatc agggctggct caaatgtgaa gggtttttac gcatggtcat ttttggactg      1440
     taatgaatgg tttgcaggct ttactgttcg ttttggatta aactttgtag attagaaaga      1500
     tggattaaaa aggtacccta agctttctgc ccaatggtac aagaactttc tcaaaagaaa      1560
     ctagctagta ttattaaaag aactttgtag tagattacag tacatcgttt gaagttgagt      1620
     tggtgcacct aattaaataa aagaggttac tcttaacata ttttttaggcc attcgttgtg      1680
     aagttgttag gctgttattt ctattatact atgttgtagt aataagtgca ttgttgtacc      1740
     agaagctatg atcataacta taggttgatc cttcatgtat cagtttgatg ttgagaatac      1800
     tttgaattaa aagtcttttt ttatttttt aaaaaaaaaa aaaaaaaaaa aaaaaaaaa        1859
//
```

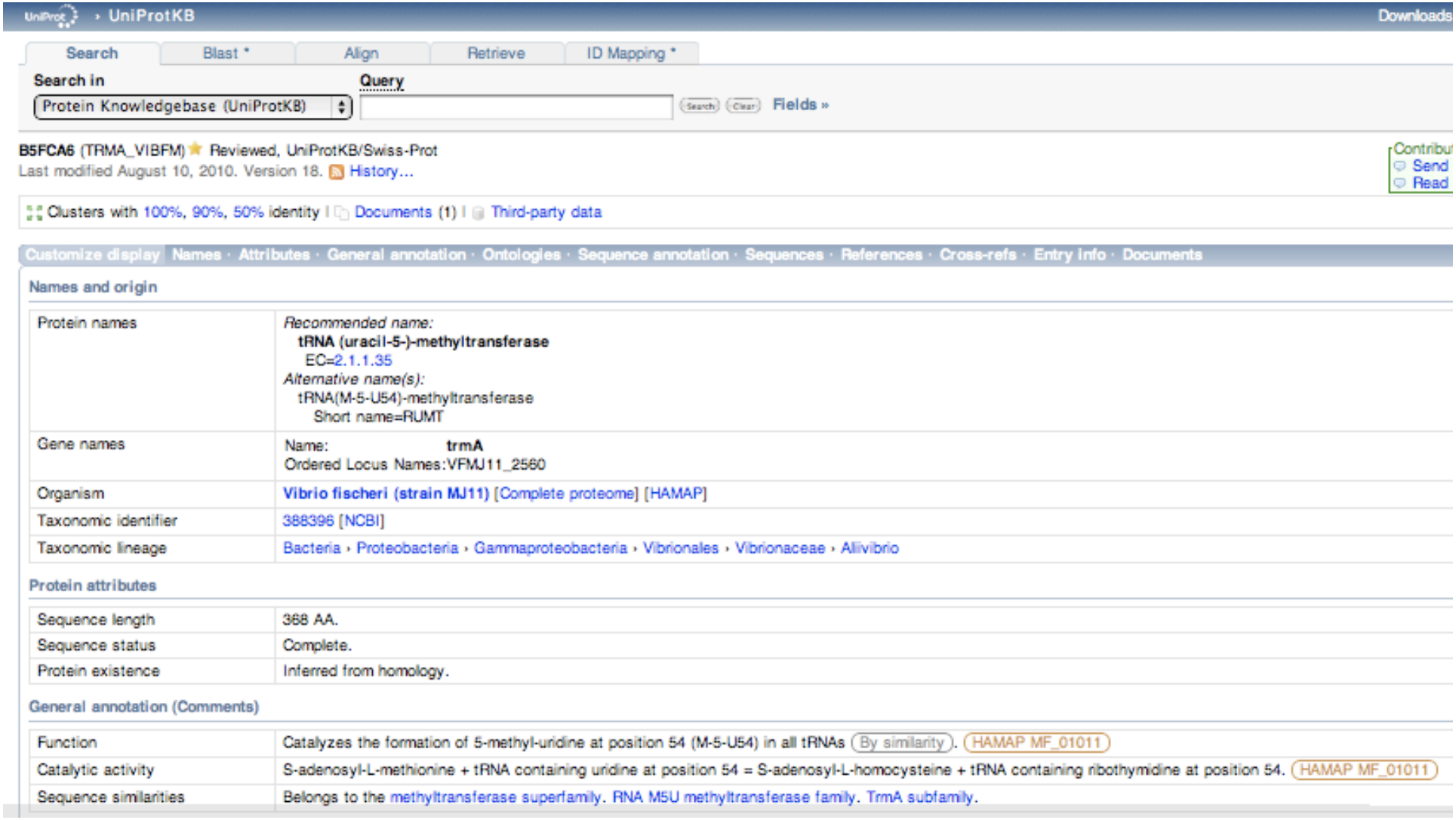
Βάσεις νουκλεοτιδικών δεδομένων. FASTA format

```
>gi|44890066|ref|NM_002228.3| Homo sapiens jun proto-oncogene (JUN), mRNA
GACATCATGGGCTATTTTTAGGGGTTGACTGGTAGCAGATAAGTGTTGAGCTCGGGCTGGATAAGGGCTC
AGAGTTGCACTGAGTGTGGCTGAAGCAGCGAGGCGGGAGTGGAGGTGCGCGGAGTCAGGCAGACAGACAG
ACACAGCCAGCCAGCCAGGTCGGCAGTATAGTCCGAACTGCAAATCTTATTTTCTTTTCACCTTCTCTCT
AACTGCCCAGAGCTAGCGCCTGTGGCTCCCGGGCTGGTGTTTCGGGAGTGTCCAGAGAGCCTGGTCTCCA
GCCGCCCCCGGGAGGAGAGCCCTGCTGCCCAGGCGCTGTTGACAGCGGCGGAAAGCAGCGGTACCCACGC
GCCCGCCGGGGGAAGTCGGCGAGCGGCTGCAGCAGCAAAGAACTTTCCCGGCTGGGAGGACCGGAGACAA
GTGGCAGAGTCCCGGAGCGAACTTTTGCAAGCCTTTCCTGCGTCTTAGGCTTCTCCACGGCGGTAAAGAC
```

Βάσεις πρωτεϊνικών δεδομένων

- Swissprot. 1987, Uni Geneva + SIB. Σχολιασμός των δεδομένων από επιστήμονες
- TrEMBL. 1996. SIB + EBI. Αυτόματη μετάφραση των ακολουθιών που βρίσκονται στην EMBL. Δεδομένα στην ίδια μορφή με την Swissprot. Μπορεί να είναι υποθετικές ή ο σχολιασμός να μην είναι εκτενής, όπως στην Swissprot.
- PIR. 1984, USA
- UniProt. 2002. Ενώθηκαν οι παραπάνω βάσεις.
- UniMes: για μεταγενωμικά δεδομένα, όπου δεν γνωρίζουμε από ποιά είδη προέρχονται οι ακολουθίες.

Swissprot (I)


UniProt › UniProtKB
Downloads
Search
Blast *
Align
Retrieve
ID Mapping *
Search in
Protein Knowledgebase (UniProtKB)
Query
Search
Clear
Fields »
B5FCA6 (TRMA_VIBFM) Reviewed, UniProtKB/Swiss-Prot
Last modified August 10, 2010. Version 18. History...
Contribu
Send
Read
Clusters with 100%, 90%, 50% identity | Documents (1) | Third-party data
Customize display Names · Attributes · General annotation · Ontologies · Sequence annotation · Sequences · References · Cross-refs · Entry info · Documents
Names and origin
Protein names
Recommended name:
tRNA (uracil-5-)-methyltransferase
EC=2.1.1.35
Alternative name(s):
tRNA(M-5-U54)-methyltransferase
Short name=RUMT
Gene names
Name: trmA
Ordered Locus Names:VFMJ11_2560
Organism
Vibrio fischeri (strain MJ11) [Complete proteome] [HAMAP]
Taxonomic identifier
388396 [NCBI]
Taxonomic lineage
Bacteria › Proteobacteria › Gammaproteobacteria › Vibrionales › Vibrionaceae › Allivibrio
Protein attributes
Sequence length
368 AA.
Sequence status
Complete.
Protein existence
Inferred from homology.
General annotation (Comments)
Function
Catalyzes the formation of 5-methyl-uridine at position 54 (M-5-U54) in all tRNAs By similarity. HAMAP MF_01011
Catalytic activity
S-adenosyl-L-methionine + tRNA containing uridine at position 54 = S-adenosyl-L-homocysteine + tRNA containing ribothymidine at position 54. HAMAP MF_01011
Sequence similarities
Belongs to the methyltransferase superfamily. RNA M5U methyltransferase family. TrmA subfamily.

Swissprot (ι)

Ontologies

Keywords

Biological process	tRNA processing
Ligand	S-adenosyl-L-methionine
Molecular function	Methyltransferase Transferase
Technical term	Complete proteome

Gene Ontology (GO)

Biological process	tRNA processing Inferred from electronic annotation. Source: UniProtKB-KW
Molecular function	tRNA (uracil-5-)-methyltransferase activity Inferred from electronic annotation. Source: EC

Complete GO annotation...

Sequence annotation (Features)

	Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing						
	Chain	1 – 368	368	tRNA (uracil-5-)-methyltransferase HAMAP MF_01011		PRO_1000198560
Sites						
	Active site	326	1	Nucleophile By similarity		
	Binding site	190	1	S-adenosyl-L-methionine By similarity		
	Binding site	218	1	S-adenosyl-L-methionine; via carbonyl oxygen By similarity		
	Binding site	223	1	S-adenosyl-L-methionine By similarity		
	Binding site	239	1	S-adenosyl-L-methionine By similarity		
	Binding site	301	1	S-adenosyl-L-methionine By similarity		

Swissprot (II)

Sequences

Sequence		Length	Mass (Da)	Tools
B5FCA6-1 [UniParc].	FASTA	368	42,807	Blast go

Last modified October 14, 2008. Version 1.
Checksum: 01F6E61F385BA072

```
        10         20         30         40         50         60
MTQSVMNPEN YQVQLDEKAE ALSAMESEFN VPELEVFSSP AENYRMRAEF RVWHEGDEMY

        70         80         90        100        110        120
YVMFNQETKE KYRVDYFLPA SRLINDLMPL LTEAVKESKT LRYKMFQVDF LSTLSGEILV

       130        140        150        160        170        180
SMLYHRQLDD AWKEEAKALK QRLNDEGFNL NIIGRARKMK IVLDQEFVIE KLKVNDDILT

       190        200        210        220        230        240
YKQVENSFTQ PNGIVAQKML EWAVDCTQNS QGDLLELYCG NGNFSLALAK NFDRVLATEL

       250        260        270        280        290        300
AKPSVDSAQY NIAANNIDNV QIIRMSAEDF TDAMEGKREF RRLKDQNIDL KSYNCNTIFV

       310        320        330        340        350        360
DPPRSGMDEG TCKMVQGYER IMYISCNPET LKENLEILSQ THNITRFALF DQFPYTHHME


AGVFLERK
```

« Hide

References

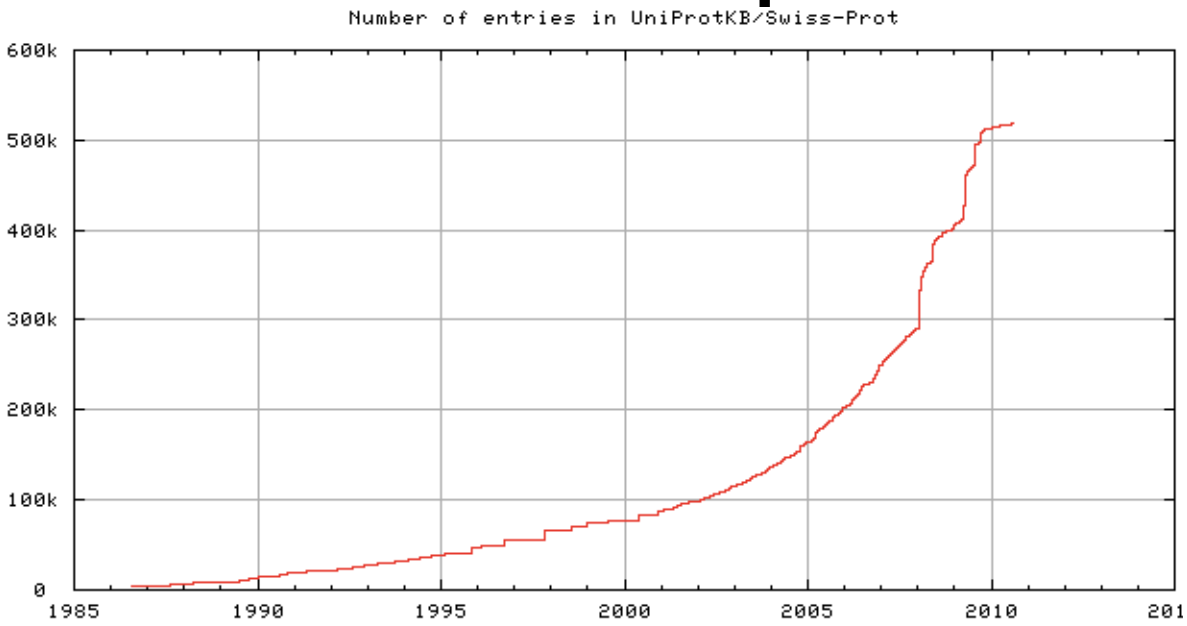
[1] **"Complete sequence of Vibrio fischeri strain MJ11."**
Mandel M.J., Stabb E.V., Ruby E.G., Ferriera S., Johnson J., Kravitz S., Beeson K., Sutton G., Rogers Y.-H., Friedman R., Frazier M., Venter J.C.
Submitted (AUG-2008) to the EMBL/GenBank/DDBJ databases
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].

Swissprot (III)

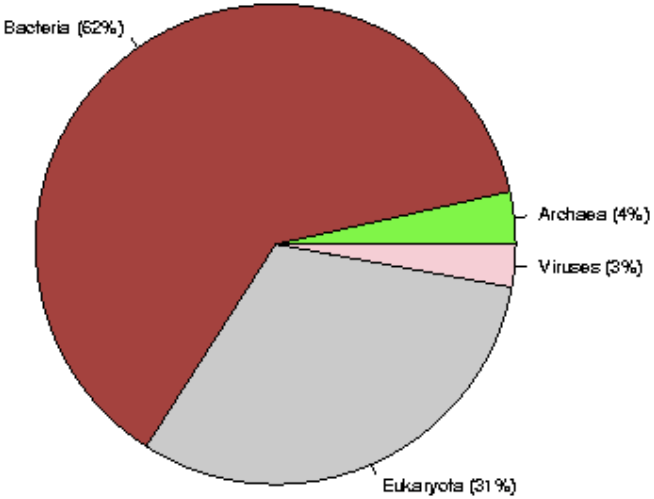
Cross-references

Sequence databases	
EMBL GenBank DDBJ	CP001139 Genomic DNA. Translation: ACH67360.1.
RefSeq	YP_002157222.1.
3D structure databases	
ProteinModelPortal	B5FCA6.
SMR	B5FCA6. Positions 1-368.
ModBase	Search...
Genome annotation databases	
GeneID	6807703.
GenomeReviews	Gene locus VFMJ11_2560 in contig CP001139_GR.
KEGG	vfm:VFMJ11_2560.
Organism-specific databases	
CMR	Search...
Phylogenomic databases	

Swissprot-statistics



2.3 Taxonomic distribution of the sequences



Kingdom	sequences	(% of the database)
Archaea	18324	(4%)
Bacteria	324101	(62%)
Eukaryota	162009	(31%)
Viruses	14914	(3%)

ΒΔ γονιδιακής έκφρασης

Microarray Data and other Gene Expression Databases

- 4DXpress
- 5'SAGE
- *Drosophila* microarray centre
- ABA - Ascidian Body Atlas
- ArrayExpress
- Axeldb
- BarleyBase
- BGED - Brain Gene Expression Database
- BloodExpress
- BodyMap
- BodyMap-Xs
- CAGE
- CATMA - Complete Arabidopsis Transcriptome MicroArray
- CEBS
- CGED - Cancer Gene Expression Database
- CleanEx
- CycleBase
- dbERGEII
- Edinburgh Mouse (EMAP) Atlas
- EMAGE
- EPConDB
- EpoDB - Erythropoiesis Database
- FLIGHT
- GEISHA
- Gene Aging Nexus
- Gene Expression in Tooth Database
- GeneNote
- GenePaint
- GeneTide
- GeneTrap
- GENSAT
- GEO - Gene Expression Omnibus
- GermOnline
- GermSAGE
- GPX-Macrophage
- GXA
- GXD - Mouse Gene Expression Database
- HemBase
- HemoPDB - Hematopoietic Promoter Database
- HPMR - Human Plasma Membrane Receptome
- HugeIndex - Human Gene Expression Index
- Interferon Stimulated Gene Database
- International Gene Trap Consortium Database
- ITTACA
- Kidney Development Database
- LOLA
- M3D
- MAGEST
- MAMEP - Molecular Anatomy of the Mouse Embryo Project
- MEPD: A Medaka gene expression pattern database
- Mouse SAGE
- NASCarrays
- NetAffx
- OncoMine

- ArrayExpress. EBI, UK. Δέχεται δεδομένα από το 2002
- Gene expression omnibus (GEO). NCBI, USA.
- Κάθε εβδομάδα το ArrayExpress ενσωματώνει δεδομένα από το GEO.
- Unigene (Expressed sequence tags)

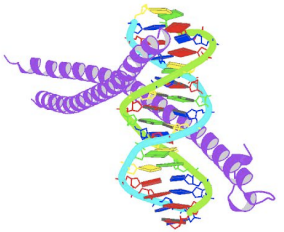
- Αν τα δεδομένα προέρχονται από μικροσυστοιχίες, τότε κατατίθενται με τη μορφή MIAME (minimum information about a microarray experiment).
- Αν τα δεδομένα προέρχονται από τεχνολογία αλληλούχισης, τότε κατατίθενται με τη μορφή MINSEQE (minimum information about a high-throughput sequencing experiment).

ΒΔ πρωτεομικής

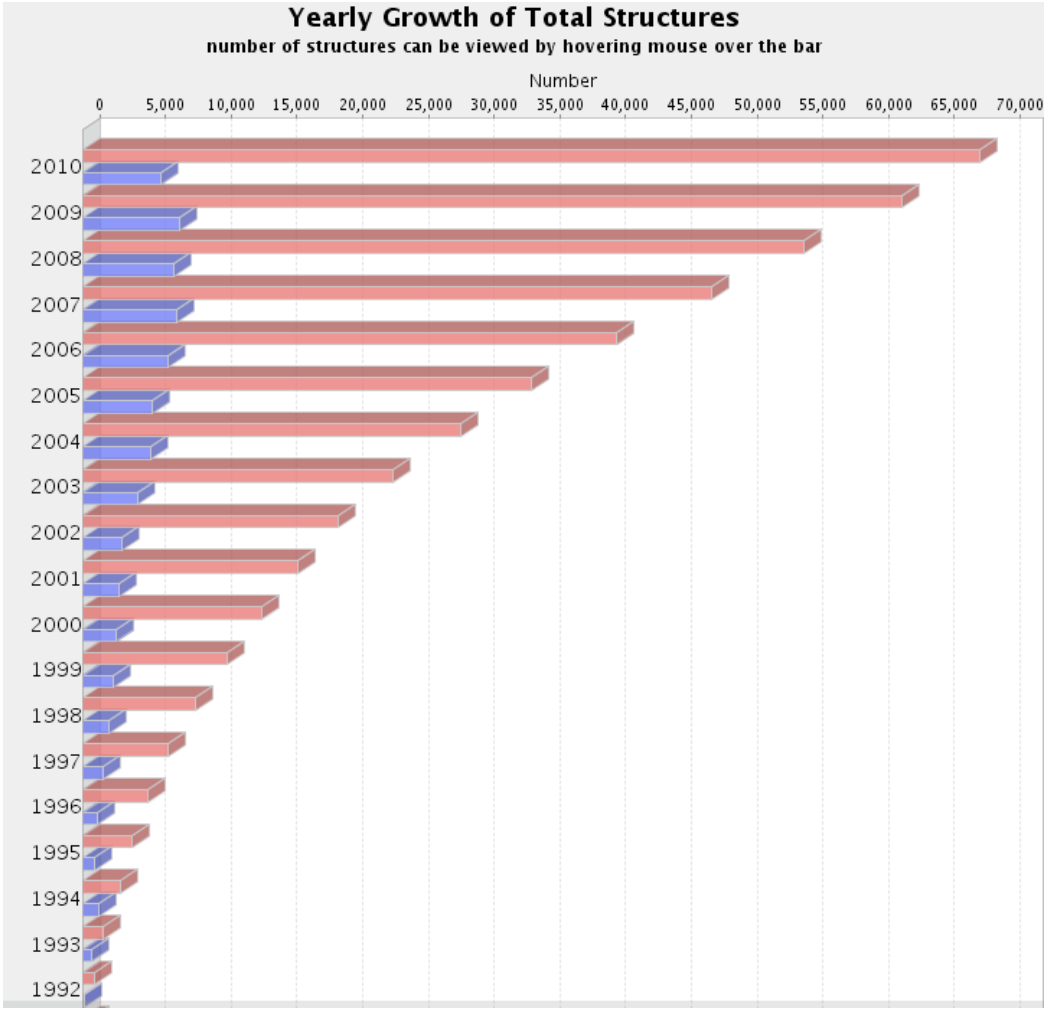
Proteomics Resources

- 2D-PAGE
- AAindex
- Biodefense Proteomics Resource Center
- BIOZON
- CutDB
- dbLEP
- dbPTM
- DynaProt 2D
- GELBANK
- MAPU
- Open Proteomics Database
- PEP: Predictions for Entire Proteomes
- PepSeeker
- PeptideAtlas
- PlantMarkers
- PRIDE
- RESID
- SWISS-2DPAGE
- Sys-BodyFluid
- SysPIMP

ΒΔ τρισδιάστατων δομών

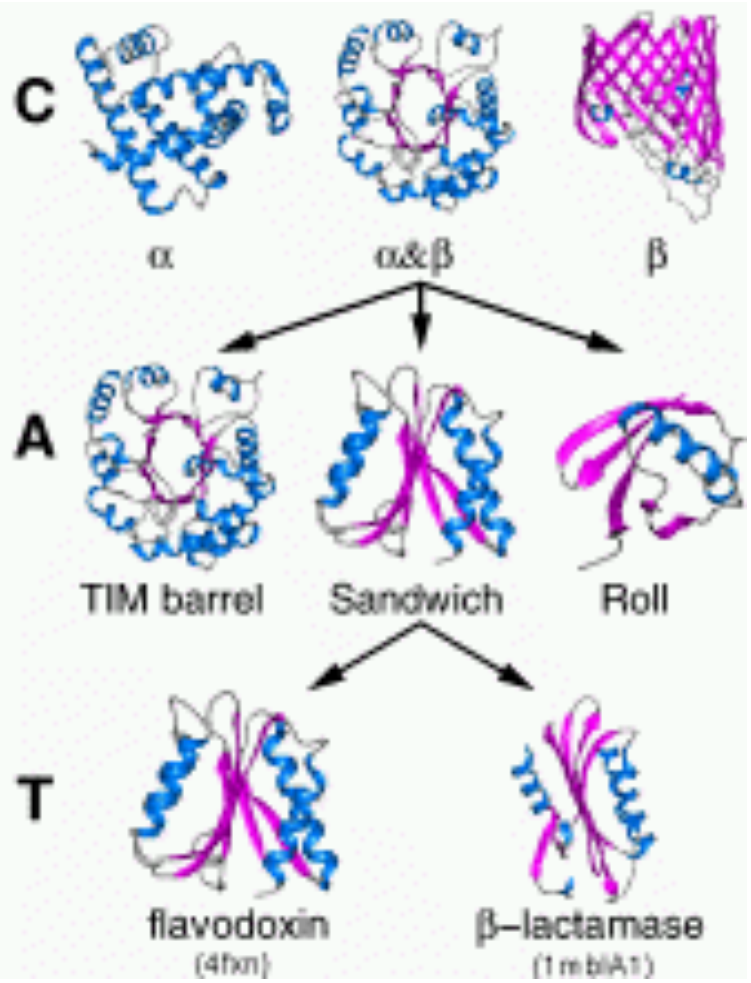


- Protein Data Bank (PDB)
  - Πρωτεΐνες
  - Νουκλεϊκά οξέα
  - Σύμπλοκα των παραπάνω
- Μέθοδοι
  - X-ray (~59000)
  - NMR (~8500)
  - Κρύο-ηλεκτρονική μικροσκοπία (~300)
- Οι παραπάνω μέθοδοι βρίσκουν τις συντεταγμένες (3D) των ατόμων του βιολογικού μορίου.
- Τα αρχεία με τις συντεταγμένες διαβάζονται από ειδικά προγράμματα (π.χ Rasmol) που απεικονίζουν την δομή στο χώρο



Βάσεις τρισδιάστατων δομών

- CATH: κατηγοριοποιεί τις τρισδιάστατες δομές των πρωτεϊνικών επικρατειών ιεραρχικά, σε 4 βασικά επίπεδα.
- Η κατηγοριοποίηση γίνεται με ένα συνδυασμό αυτόματων μεθόδων και ανθρώπινης κρίσης.



Βάσεις τρισδιάστατων δομών

CATH Domain: 1cukA01 XML

PDB 1cuk, Chain A, Domain 1

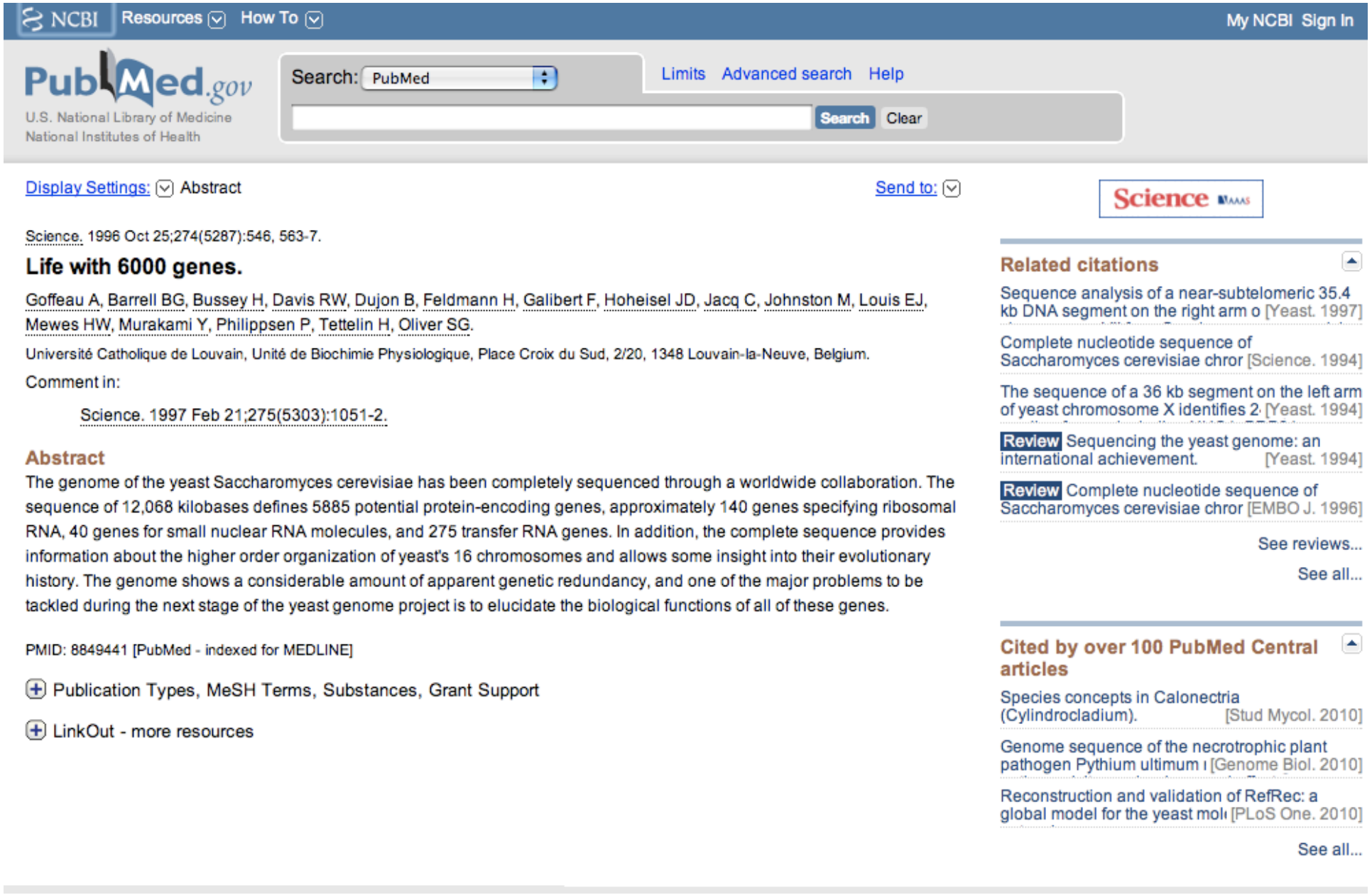
CATH Code	Level Description	Links
C 2	Mainly Beta	
A 2.40	Beta Barrel	
T 2.40.50	OB fold (Dihydrolipoamide Acetyltransferase, E2P)	
H 2.40.50.140	Nucleic acid-binding proteins	Gene3D
S 2.40.50.140.47		
O 2.40.50.140.47.1		
L 2.40.50.140.47.1.1		
I 2.40.50.140.47.1.1.1		
D 2.40.50.140.47.1.1.1.1		

1cukA01

Pubmed

- ΒΔ του NCBI. Ξεκίνησε τον Ιανουάριο του 1996.
- Καταχωρεί όλες τις δημοσιευμένες εργασίες που προέρχονται από τον ευρύτερο χώρο της βιοϊατρικής
- ~20 εκατομύρια εργασίες καταχωρημένες (Ιούλιος 2010)
- Όταν μια εργασία γίνεται δεκτή από το περιοδικό, κατατίθεται και στην Pubmed
- Η Pubmed δίνει ένα μοναδικό κωδικό εγγραφής (PMID) και λέξεις κλειδιά που χαρακτηρίζουν το περιεχόμενο της εργασίας (MeSH terms).
- Από το 2007, το NIH απαιτεί όποιες ερευνητικές εργασίες έχουν χρηματοδοτηθεί από αυτό, τα αποτελέσματά τους να γίνονται προσβάσιμα σε όλους, μέσω του Pubmed Central (εντός 12 μηνών από την ημερομηνία δημοσίευσης). (~ 1 εκατομύριο εργασίες)

Pubmed


NCBI
Resources
How To
My NCBI
Sign In
PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
Search: PubMed
Limits
Advanced search
Help
Search
Clear
Display Settings: Abstract
Send to:
Science
Science. 1996 Oct 25;274(5287):546, 563-7.
Life with 6000 genes.
Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG.
Université Catholique de Louvain, Unité de Biochimie Physiologique, Place Croix du Sud, 2/20, 1348 Louvain-la-Neuve, Belgium.
Comment in:
Science. 1997 Feb 21;275(5303):1051-2.
Abstract
The genome of the yeast Saccharomyces cerevisiae has been completely sequenced through a worldwide collaboration. The sequence of 12,068 kilobases defines 5885 potential protein-encoding genes, approximately 140 genes specifying ribosomal RNA, 40 genes for small nuclear RNA molecules, and 275 transfer RNA genes. In addition, the complete sequence provides information about the higher order organization of yeast's 16 chromosomes and allows some insight into their evolutionary history. The genome shows a considerable amount of apparent genetic redundancy, and one of the major problems to be tackled during the next stage of the yeast genome project is to elucidate the biological functions of all of these genes.
PMID: 8849441 [PubMed - indexed for MEDLINE]
Publication Types, MeSH Terms, Substances, Grant Support
LinkOut - more resources
Related citations
Sequence analysis of a near-subtelomeric 35.4 kb DNA segment on the right arm o [Yeast. 1997]
Complete nucleotide sequence of Saccharomyces cerevisiae chror [Science. 1994]
The sequence of a 36 kb segment on the left arm of yeast chromosome X identifies 2 [Yeast. 1994]
Review Sequencing the yeast genome: an international achievement. [Yeast. 1994]
Review Complete nucleotide sequence of Saccharomyces cerevisiae chror [EMBO J. 1996]
See reviews...
See all...
Cited by over 100 PubMed Central articles
Species concepts in Calonectria (Cylindrocladium). [Stud Mycol. 2010]
Genome sequence of the necrotrophic plant pathogen Pythium ultimum i [Genome Biol. 2010]
Reconstruction and validation of RefRec: a global model for the yeast mol [PLoS One. 2010]
See all...

Pubmed

```
PMID- 8849441
OWN - NLM
STAT- MEDLINE
DA  - 19961122
DCOM- 19961122
LR  - 20090929
IS  - 0036-8075 (Print)
IS  - 0036-8075 (Linking)
VI  - 274
IP  - 5287
DP  - 1996 Oct 25
TI  - Life with 6000 genes.
PG  - 546, 563-7
AB  - The genome of the yeast Saccharomyces cerevisiae has been completely sequenced
      through a worldwide collaboration. The sequence of 12,068 kilobases defines 5885
      potential protein-encoding genes, approximately 140 genes specifying ribosomal
      RNA, 40 genes for small nuclear RNA molecules, and 275 transfer RNA genes. In
      addition, the complete sequence provides information about the higher order
      organization of yeast's 16 chromosomes and allows some insight into their
      evolutionary history. The genome shows a considerable amount of apparent genetic
      redundancy, and one of the major problems to be tackled during the next stage of
      the yeast genome project is to elucidate the biological functions of all of these
      genes.
AD  - Universite Catholique de Louvain, Unite de Biochimie Physiologique, Place Croix
      du Sud, 2/20, 1348 Louvain-la-Neuve, Belgium.
FAU - Goffeau, A
AU  - Goffeau A
```

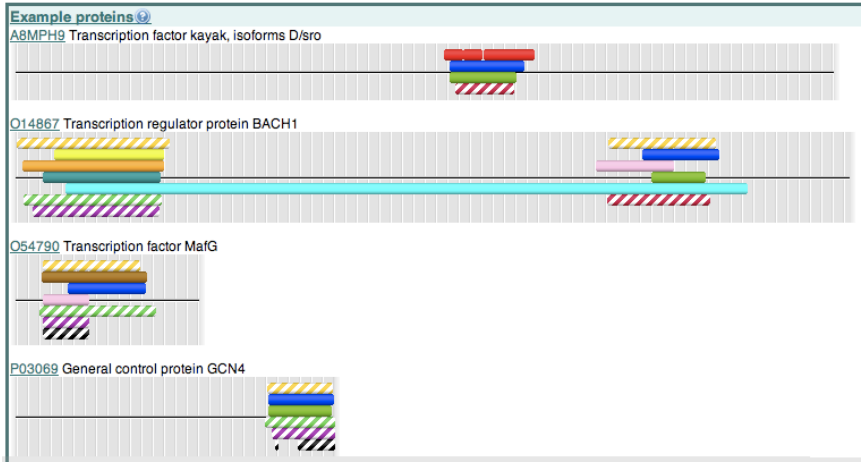
Pubmed

```
FAU - Oliver, S G
AU  - Oliver SG
LA  - eng
GR  - Wellcome Trust/United Kingdom
PT  - Journal Article
PT  - Review
PL  - UNITED STATES
TA  - Science
JT  - Science (New York, N.Y.)
JID - 0404511
RN  - 0 (DNA, Fungal)
RN  - 0 (Fungal Proteins)
RN  - 0 (RNA, Fungal)
SB  - IM
CIN - Science. 1997 Feb 21;275(5303):1051-2. PMID: 9054002
MH  - Amino Acid Sequence
MH  - Base Sequence
MH  - *Chromosome Mapping
MH  - Chromosomes, Fungal/genetics
MH  - Computer Communication Networks
MH  - DNA, Fungal/genetics
MH  - Evolution, Molecular
MH  - Fungal Proteins/chemistry/genetics/physiology
MH  - Gene Library
MH  - *Genes, Fungal
MH  - *Genome, Fungal
MH  - International Cooperation
MH  - Multigene Family
MH  - Open Reading Frames
MH  - RNA, Fungal/genetics
MH  - Saccharomyces cerevisiae/*genetics
MH  - Sequence Analysis, DNA
RF  - 86
EDAT- 1996/10/25
MHDA- 1996/10/25 00:01
CRDT- 1996/10/25 00:00
PST - ppublish
SO  - Science. 1996 Oct 25;274(5287):546, 563-7.
```

ΒΔ πρωτεϊνικών επικρατειών

- Πρωτεϊνική επικράτεια: Μια περιοχή της πρωτεΐνης με συγκεκριμένη λειτουργία/ δομή και καλά συντηρημένη.
- Διάφορες βάσεις δεδομένων, όπως:
  - PROSITE
  - Pfam
  - PRINTS
  - ProDom
  - SMART
  - TIGRFAMs
  - PIR superfamily
  - Superfamily
- Έχουν ενσωματωθεί στο INTERPRO
- Το INTERPRO περιέχει πρωτεϊνικές επικράτειες. Το πρόγραμμα INTERPROscan ανιχνεύει αυτές τις επικράτειες στις πρωτεΐνες.

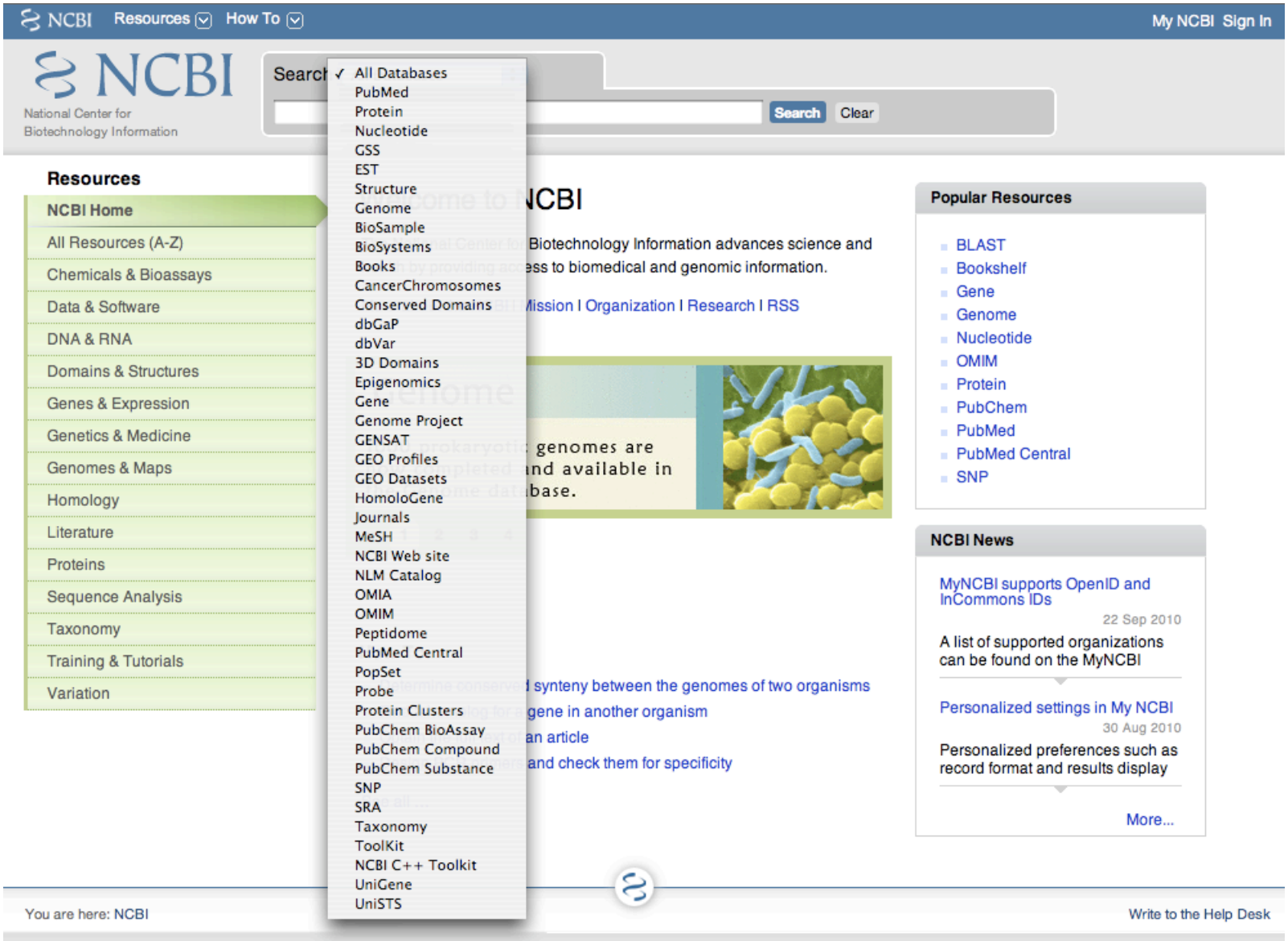
INTERPRO



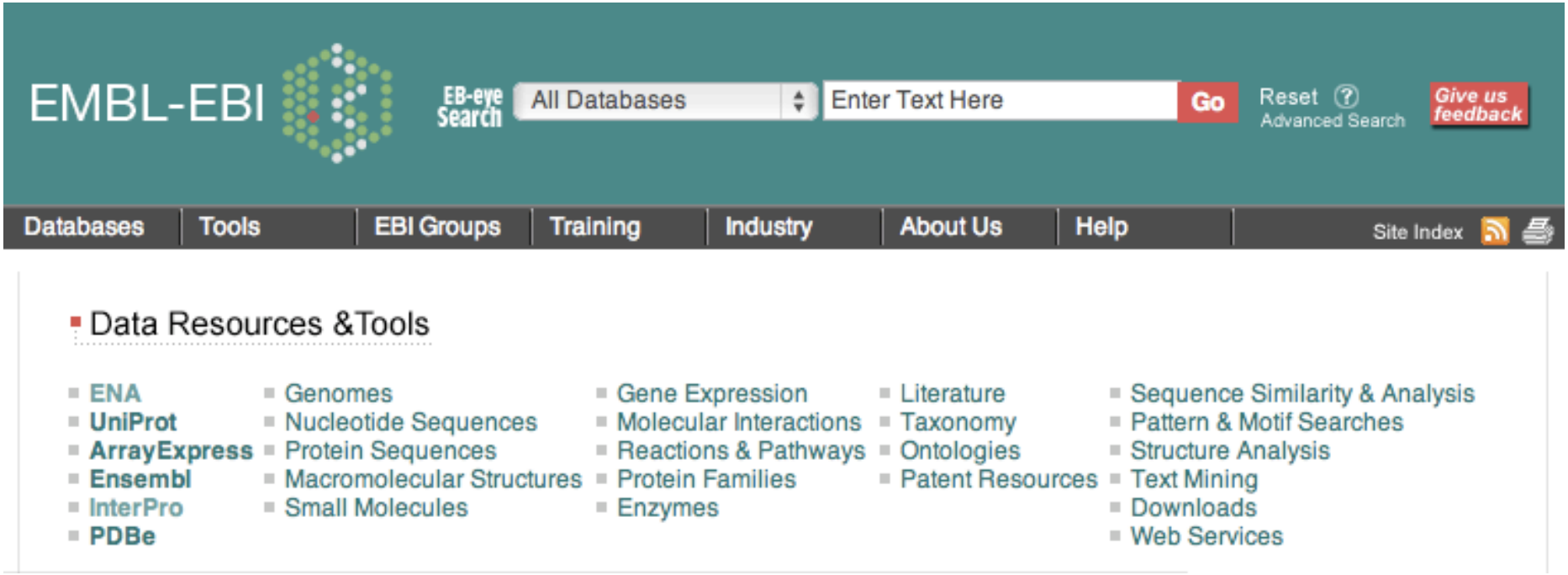
Example Proteins Key

InterPro entry accession number/name and structure databases		Colour code
IPR013089	Kelch related	
IPR013069	BTB/POZ	
IPR011333	BTB/POZ fold	
IPR000837	Fos transforming protein	
IPR000210	BTB/POZ-like	
IPR004827	Basic-leucine zipper (bZIP) transcription factor	
IPR008917	Eukaryotic transcription factor, Skn-1-like, DNA-binding	
IPR004826	Maf transcription factor	
IPR011616	bZIP transcription factor, bZIP-1	
	PDB Chain	
	ModBase	
	CATH Domain	
	SWISS-MODEL	
	SCOP Domain	

NCBI/Entrez


NCBI
Resources
How To
My NCBI
Sign In
National Center for Biotechnology Information
Search
All Databases
PubMed
Protein
Nucleotide
GSS
EST
Structure
Genome
BioSample
BioSystems
Books
CancerChromosomes
Conserved Domains
dbGaP
dbVar
3D Domains
Epigenomics
Gene
Genome Project
GENSAT
GEO Profiles
GEO Datasets
HomoloGene
Journals
MeSH
NCBI Web site
NLM Catalog
OMIA
OMIM
Peptidome
PubMed Central
PopSet
Probe
Protein Clusters
PubChem BioAssay
PubChem Compound
PubChem Substance
SNP
SRA
Taxonomy
ToolKit
NCBI C++ Toolkit
UniGene
UniSTS
Search
Clear
Resources
NCBI Home
All Resources (A-Z)
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Popular Resources
BLAST
Bookshelf
Gene
Genome
Nucleotide
OMIM
Protein
PubChem
PubMed
PubMed Central
SNP
NCBI News
MyNCBI supports OpenID and InCommons IDs
22 Sep 2010
A list of supported organizations can be found on the MyNCBI
Personalized settings in My NCBI
30 Aug 2010
Personalized preferences such as record format and results display
More...
You are here: NCBI
Write to the Help Desk

EBI


EMBL-EBI
EB-eye Search
All Databases
Enter Text Here
Go
Reset
Advanced Search
Give us feedback
Databases
Tools
EBI Groups
Training
Industry
About Us
Help
Site Index
Data Resources &Tools
ENA
UniProt
ArrayExpress
Ensembl
InterPro
PDBe
Genomes
Nucleotide Sequences
Protein Sequences
Macromolecular Structures
Small Molecules
Gene Expression
Molecular Interactions
Reactions & Pathways
Protein Families
Enzymes
Literature
Taxonomy
Ontologies
Patent Resources
Sequence Similarity & Analysis
Pattern & Motif Searches
Structure Analysis
Text Mining
Downloads
Web Services

EBI: Μηχανή αναζήτησης EB-eye

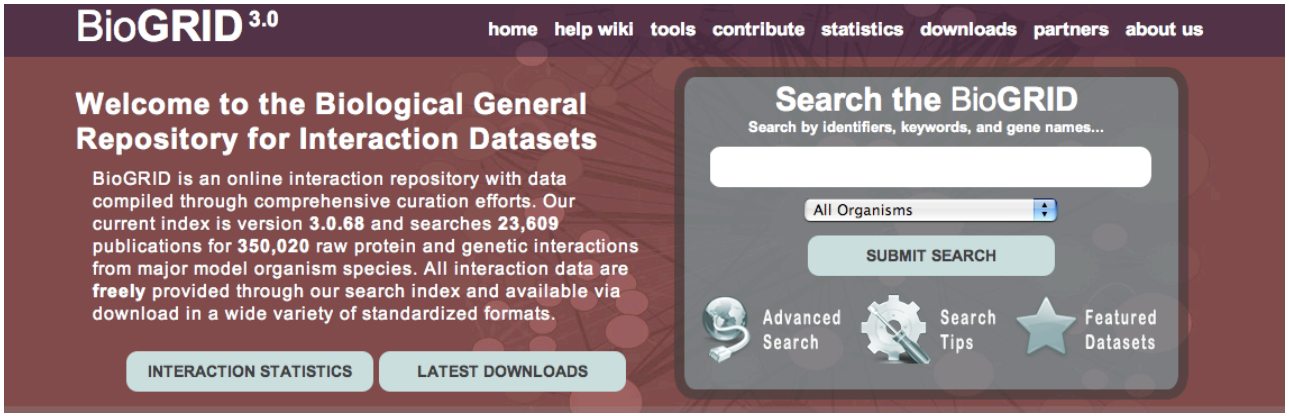
What is included in the EB-eye Search?

There are several data resources available organised into knowledge 'domains' that should make it easier for a user to find results relevant to a text query. Whenever you perform a search you can see the summary page of results grouped into different domains.

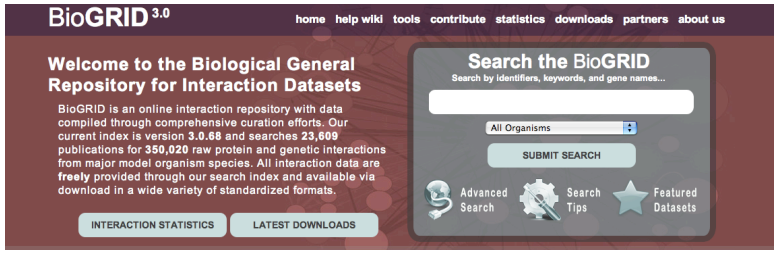
Genomes	• Integr8 • Ensembl
Nucleotide Sequences	• ASTD •EMBL-Bank •EMBL-Bank (Coding Sequence)
Protein Sequences	• Pride • UniProtKB • UniRef • UniParc
Macromolecular Structures	• PDBe
Small Molecules	• ChEBI • Ligands • RESID
Gene Expression	• ArrayExpress (Repository of Microarray data) • ArrayExpress (Warehouse of Microarray experiments) • ArrayExpress (Warehouse of gene expression profiles)
Molecular Interactions	• IntAct Experiments • IntAct Interactions • IntAct Interactors
Reactions & Pathways	• BioModels • Reactome
Protein Families	• GPCRDB • InterPro • MEROPS Peptidases • MEROPS Peptidase Clans • MEROPS Peptidase Families
Enzymes	• Intenz
Literature	• Medline • Patents
Ontologies	• GO • SBO • Taxonomy
EBI Website	• Main sections • EBI Staff • EBI Members & Groups • 2can Support Portal

Πρωτεϊνικές αλληλεπιδράσεις

- Metabolic and Signaling Pathways
  - Enzymes and enzyme nomenclature
  - Metabolic pathways
  - Protein-protein interactions
    - 3D-Interologs
    - AffinDB
    - aMAZE
    - Argonaute
    - AtPID
    - Bacteriome.org
    - Binding MOAD
    - BioCarta
    - BioGRID
    - CellCircuits
    - ConsensusPathDB
    - CORUM
    - CutDB
    - Cyclonet
    - DDIB
    - DIMA
    - DIP - Database of Interacting Proteins
    - Dockground
    - DOMINE
    - DOMINO
    - DRC - Database of Ribosomal Crosslinks
    - DroID - Drosophila Interactions Database
    - Dynamic Proteomics
    - EciD
    - EndoNet
    - GeneNet
    - GLIDA
    - Hedgehog signalling
    - Het-PDB Navi
    - HotSprint
    - hp-DPI
    - HPID - Human Protein Interaction database
    - I2D - Interologous Interaction Database
    - IMOTdb
    - IntAct
    - Inter-Chain Beta-Sheets
    - InterDom
    - KDBI
    - LigAsite
    - MHC-Peptide Interaction Database
    - MiMI
    - MiMI - Michigan Molecular Interactions
    - MINT
    - MiST - Microbial Signal Transduction database
    - Mpact
    - NetworKIN
    - NPInter
    - PANTHER
    - Pathguide
    - PathoPlant



Πρωτεϊνικές αλληλεπιδράσεις



Workbook1

	A	B	C	D	E	F	G	H	I	J	K	L
1	Brief Description of the Columns:											
2												
3	A.) INTERACTOR_A											
4	B.) INTERACTOR_B											
5	C.) OFFICIAL_SYMBOL_FOR_A											
6	D.) OFFICIAL_SYMBOL_FOR_B											
7	E.) ALIASES_FOR_A											
8	F.) ALIASES_FOR_B											
9	G.) EXPERIMENTAL_SYSTEM											
10	H.) SOURCE											
11	I.) PUBMED_ID											
12	J.) ORGANISM_A_ID											
13	K.) ORGANISM_B_ID											
14												
15												
16	INTERACTOR_A	INTERACTOR_B	OFFICIAL_SYMBOL_A	OFFICIAL_SYMBOL_B	ALIASES_FOR_A	ALIASES_FOR_B	EXPERIMENTAL_SYSTEM	SOURCE	PUBMED_ID	ORGANISM_A	ORGANISM_B_ID	
17	ETG6416	ETG2318	MAP2K4	FLNC	JNKK1|SERK1|PRK	ABPL|ABPA|ABP-28	Two-hybrid	Marti A (1997)	9006895	9606	9606	
18	ETG84665	ETG88	MYPN	ACTN2	MYOP	CMD1AA	Two-hybrid	Bang ML (2001)	11309420	9606	9606	
19	ETG90	ETG2339	ACVR1	FNTA	ACVRLK2|ACTRI|S	PTAR2|FPTA|PGGT	Two-hybrid	Wang T (1996)	8599089	9606	9606	
20	ETG2624	ETG5371	GATA2	PML	MGC2306|NFE1B|F	RNF71|TRIM19|PP	Two-hybrid	Tsuzuki S (2000)	10938104	9606	9606	
21	ETG6118	ETG6774	RPA2	STAT3	REPA2|RPA32	APRF|MGC16063|H	Two-hybrid	Kim J (2000)	10875894	9606	9606	
22	ETG375	ETG23163	ARF1	GGA3	N/A	KIAA0154	Two-hybrid	Dell'Angelica EC (20	10747089	9606	9606	
23	ETG377	ETG23647	ARF3	ARFIP2	N/A	POR1	Two-hybrid	Kanoh H (1997)	9038142	9606	9606	
24	ETG377	ETG27236	ARF3	ARFIP1	N/A	MGC117369|HSU5	Two-hybrid	Kanoh H (1997)	9038142	9606	9606	
25	ETG10327	ETG54512	AKR1A1	EXOSC4	MGC1380|DD3|ALI	Ski6p|FLJ20591|hI	Two-hybrid	Lehner B (2004)	15231747	9606	9606	
26	ETG54464	ETG226	XRN1	ALDOA	DKFZp686B22225|	MGC17767|GSD12	Two-hybrid	Lehner B (2004)	15231747	9606	9606	
27	ETG351	ETG10513	APP	APPBP2	ABPP|PN2|AD1|CV	PAT1|HS.84084|KI	Two-hybrid	Zheng P (1998)	9843960	9606	9606	
28	ETG333	RP6-239D12.2	APLP1	DAB1	APLP	N/A	Two-hybrid	Homayouni R (1999	10460257	9606	9606	
29	ETG10370	ETG7020	CITED2	TFAP2A	P35SRJ|MRG1	BOFS|AP2TF|AP-2|	Two-hybrid	Braganca J (2003)	12586840	9606	9606	
30	RP1-85F18.1	ETG7020	EP300	TFAP2A	p300|KAT3B	BOFS|AP2TF|AP-2|	Two-hybrid	Braganca J (2003)	12586840	9606	9606	

Tab delimited format

Μεταβολικά μονοπάτια

- Metabolic and Signaling Pathways
  - Enzymes and enzyme nomenclature
  - Metabolic pathways
    - BioCarta
    - BioCyc
    - Bionemo
    - BioSilico
    - BRITE - Biomolecular Relations in Information Transmission and Expression
    - BSD - Biodegradative Strain Database
    - HMDB
    - HMDB - The Human Metabolome Database
    - KEGG - Kyoto Encyclopedia of Genes and Genomes
    - Klotho
    - LIGAND
    - MedicCyc
    - MetaCrop
    - MetaCyc
    - Metagrowth
    - MMCD
    - MODOMICS
    - NMPDR - National Microbial Pathogen Data Resource
    - Pathguide
    - PMAP
    - PUMA2
    - SYSTOMONAS
    - UM-BBD
  - Protein-protein interactions
  - Signalling pathways

KEGG pathways

- Kyoto encyclopedia of genes and genomes
- 2010: 374 μεταβολικά μονοπάτια

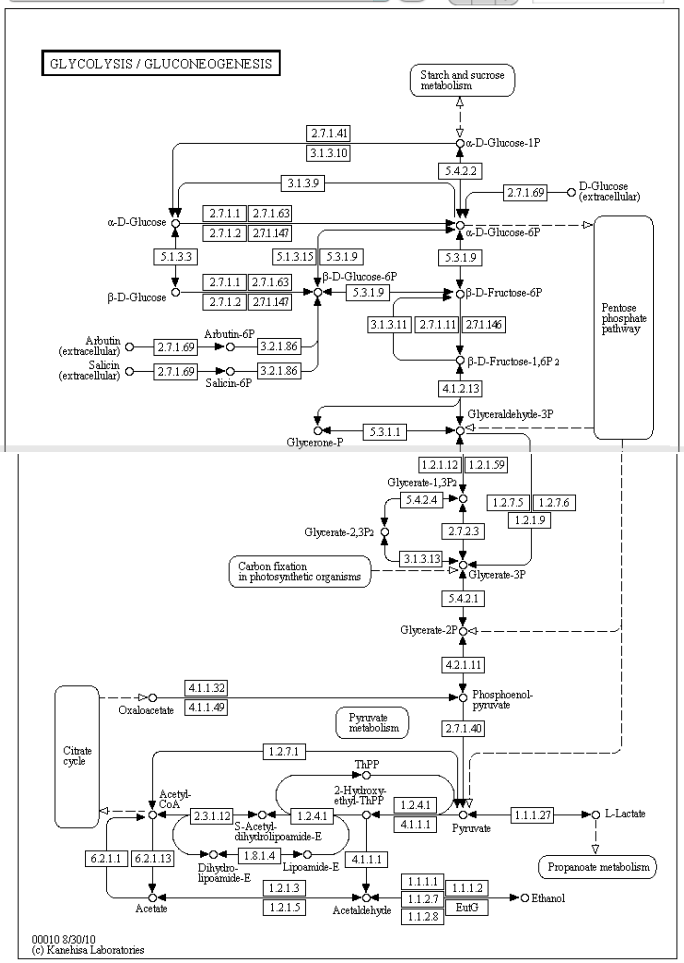
KEGG pathways

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



ENZYME: 3.1.3.9

Help

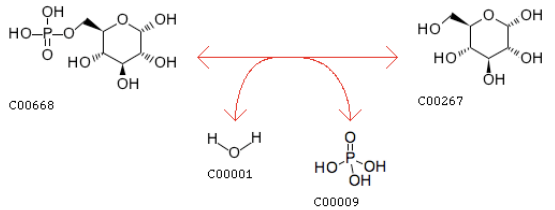
Entry	EC 3.1.3.9 Enzyme
Name	glucose-6-phosphatase; glucose 6-phosphate phosphatase
Class	Hydrolases; Acting on ester bonds; Phosphoric-monoester hydrolases BRITE hierarchy
Sysname	D-glucose-6-phosphate phosphohydrolase
Reaction(IUBMB)	D-glucose 6-phosphate + H2O = D-glucose + phosphate [RN:R00303]
Reaction(KEGG)	R00303 > R01788 Show all
Substrate	D-glucose 6-phosphate [CPD:C00092]; H2O [CPD:C00001]
Product	D-glucose [CPD:C00031]; phosphate [CPD:C00009]
Comment	Wide distribution in animal tissues. Also catalyses potent transphosphorylations from carbamoyl phosphate, hexose phosphates, diphosphate, phosphoenolpyruvate and nucleoside di- and triphosphates, to D-glucose, D-mannose, 3-methyl-D-glucose or 2-deoxy-D-glucose [cf. EC 2.7.1.62 (phosphoramidate---hexose phosphotransferase), EC 2.7.1.79 (diphosphate---glycerol phosphotransferase) and EC 3.9.1.1 (phosphoamidase)].
Pathway	ec00010 Glycolysis / Gluconeogenesis ec00052 Galactose metabolism ec00500 Starch and sucrose metabolism ec01100 Metabolic pathways
Orthology	K01084 glucose-6-phosphatase
Genes	HSA: 2538(G6PC) 57818(G6PC2) PTR: 741431(G6PC2) MCC: 712053 MMU: 14377(G6pc) RNO: 25634(G6pc) CFA: 403492(G6PC) BTA: 538710(G6PC) SSC: 100134959(G6PC)

All links

Ontology (5)
KEGG BRITE (5)
Pathway (89)
KEGG PATHWAY (89)
Disease (1)
OMIM (1)
Chemical substance (6)
KEGG COMPOUND (6)
Chemical reaction (12)
KEGG ENZYME (4)
KEGG REACTION (2)
KEGG RPAIR (5)
KEGG RCLASS (1)
Genome (2)
KEGG GENOME (2)
Gene (58)
KEGG ORTHOLOGY (1)
KEGG GENES (20)
KEGG DGENES (8)
KEGG EGENES (29)
Protein sequence (64)
UniProt (25)
PRF (3)
RefSeq(pep) (25)
PMD (11)
DNA sequence (50)
RefSeq(nuc) (30)
GenBank (10)
EMBL (10)
Protein domain (2)
InterPro (1)
Pfam (1)
Literature (3)
PubMed (3)
Enzyme (4)
BRENDA (1)
EXPASY-ENZYME (1)
EXPLORENZ (1)
IUBMB (1)
All databases (296)

REACTION: R01788

Help

Entry	R01788 Reaction
Name	alpha-D-Glucose 6-phosphate phosphohydrolase
Definition	alpha-D-Glucose 6-phosphate + H2O <=> alpha-D-Glucose + Orthophosphate
Equation	C00668 + C00001 <=> C00267 + C00009
	
RPair	RP00216 C00267_C00668 main RP05676 C00001_C00009 leave RP06709 C00009_C00668 leave
Enzyme	3.1.3.9
Pathway	rn00010 Glycolysis / Gluconeogenesis rn00052 Galactose metabolism rn00500 Starch and sucrose metabolism rn01100 Metabolic pathways
Orthology	K01084 glucose-6-phosphatase [EC:3.1.3.9]

All links

Ontology (2)
KEGG BRITE (2)
Pathway (8)
KEGG PATHWAY (8)
Chemical substance (4)
KEGG COMPOUND (4)
Chemical reaction (5)
KEGG ENZYME (1)
KEGG RPAIR (3)
KEGG RCLASS (1)
Gene (1)
KEGG ORTHOLOGY (1)
All databases (20)

ΒΔ για μαθηματικά μοντέλα μοριακών μονοπατιών

- Biomodels. EBI. 2010: 249 ελεγμένα/σχολιασμένα μοντέλα
- Αποθηκευμένα σε μορφή SBML (Systems Biology Markup Language)

⊟ The following fields are used to describe a model:

- *BioModels ID* → A unique string of characters associated with the model, which will never be re-used even if the model is deleted from the BioModels Database.
- *Name* → The name of the model, as written in the model itself by its creator(s).
- *Publication ID* → The unique identifier of the reference publication describing the model, specified either as a PubMed identifier (linked to the EBI Medline database), or as a DOI (linked to the original publication through a DOI resolver), or as an URL. Being all published, all models must have one publication identifier, and the same identifier can be shared amongst several models if they have been described in the same publication.
- *Last Modified* → The date when the model was last modified.

⇦ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 ⇨ 10 | 50 | 100 | All

BioModels ID	Name	Publication ID	Last Modified
BIOMD0000000001	Edelstein1996_EPSP_AChEvent	8983160	2009-06-05T11:40:04+00:00
BIOMD0000000002	Edelstein1996_EPSP_AChSpecies	8983160	2009-08-13T12:24:26+00:00
BIOMD0000000003	Goldbeter1991_MinMitOscil	1833774	2010-03-17T00:25:38+00:00
BIOMD0000000004	Goldbeter1991_MinMitOscil_ExplInact	1833774	2009-08-10T15:46:55+00:00
BIOMD0000000005	Tyson1991_CellCycle_6var	1831270	2010-02-12T14:00:50+00:00
BIOMD0000000006	Tyson1991_CellCycle_2var	1831270	2009-08-10T14:30:30+00:00
BIOMD0000000007	Novak1997_CellCycle	9256450	2009-10-15T16:07:54+00:00
BIOMD0000000008	Gardner1998_CellCycle_Goldbeter	9826676	2009-08-10T15:44:34+00:00
BIOMD0000000009	Huang1996_MAPK_ultrasens	8816754	2010-02-07T20:45:50+00:00
BIOMD0000000010	Kholodenko2000_MAPK_feedback	10712587	2009-07-30T11:35:19+00:00