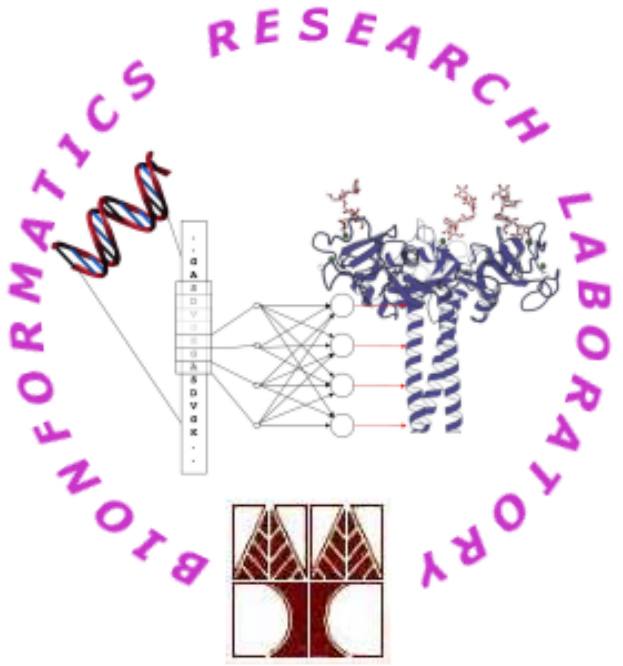
Βάσεις δεδομένων αλληλουχιών



Vasilis Promponas
Bioinformatics Research Laboratory
Department of Biological Sciences
University of Cyprus

ΣΥΝΟΨΗ

- Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών
  - Λίγη ιστορία και μια γεύση από ΒΔ
    - GenBank, EMBL, DDBJ
    - International Nucleotide Sequence Database Collaboration
- Βάσεις δεδομένων αμινοξικών αλληλουχιών
  - Λίγη ακόμη ιστορία ...
    - Atlas of protein sequence and structure,
    - NBRF/PSD, PIR, SwissProt
    - UniProt
- Συζήτηση ..

Πρωτογενείς και Δευτερογενείς ΒΔ

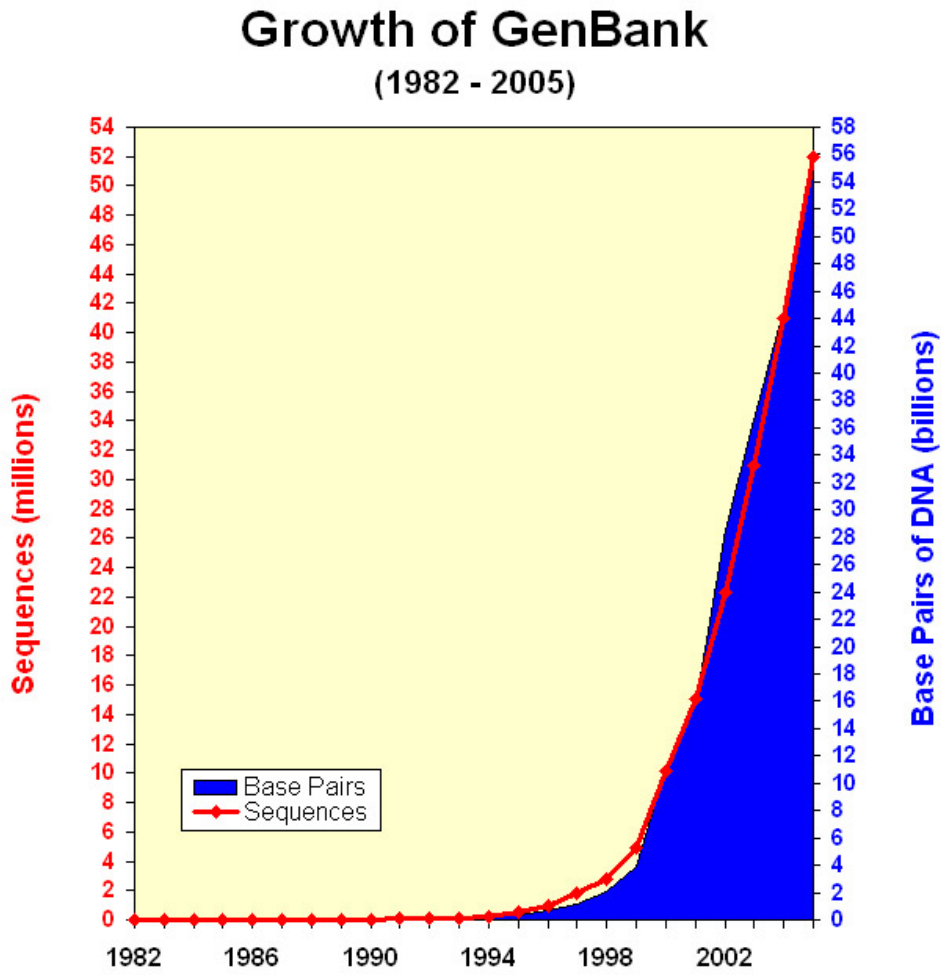
- Πρωτογενείς (primary/archival)
  - Δίνουν το χώρο και το μηχανισμό για την κατάθεση (και την πρόσβαση) σε ΠΕΙΡΑΜΑΤΙΚΑ δεδομένα που αφορούν αλληλουχίες
    - π.χ. προσδιόρισα την αλληλουχία ενός γονιδίου ή τμήματός του
- Δευτερογενείς (secondary/curated)
  - Βασίζονται σε δεδομένα των πρωτογενών βάσεων
  - Περιέχουν επιπλέον σχολιασμό, ο οποίος ΔΕΝ ΕΧΕΙ (απαραίτητα) πειραματική υποστήριξη

Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών

- Λίγη ιστορία ..
  - GenBank
  - EMBL
  - DDBJ
  - Σήμερα: http://www.insdc.org/page.php?page=home

GenBank (US)

http://www.ncbi.nlm.nih.gov/Genbank



http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

- 1979 – Los Alamos Sequence Database (LANL)
- 1982 – GenBank (NIH, NSF, DoE, DoD)
- 1989 – 1992, η GenBank περνά σταδιακά στον έλεγχο του NCBI

EMBL Databank

http://www.ebi.ac.uk/embl/

- European Molecular Biology Laboratory – EMBL (http://www.embl.org)
  - Ευρωπαϊκή προσπάθεια (20 κράτη-μέλη [$$])
  - Βασική έρευνα στη Μοριακή Βιολογία
  - 5 παραρτήματα: Heidelberg (DE), Hinxton (UK), Grenoble (FR), Hamburg (DE), Monterotondo (IT)
- 1980 – Ίδρυση (EMBL-Heidelberg)
- 1982 – Συνεργασία με GenBank
- 1994 – EMBL-EBI (Hinxton)

DNA DataBank of Japan

http://www.ddbj.nig.ac.jp/

- 1984 – 1987 – Ίδρυση (National Institute of Genetics - Mishima)
- 1987 – Συμμετέχει στο INSDC

Τα πέτρινα χρόνια

- Αρχικά τα δεδομένα προέρχονταν (κυρίως) από τη βιβλιογραφία ...

Christian Burks, Michael J. Cinkosky, Paul Gilna, Jamie E. -D. Hayden, Yuki Abe, Edwin J. Atencio, Steve Barnhouse, David Benton, Connie A. Buenafe, Karen E. Cumella, Dan B. Davison, David B. Emmert, Mary Jo Faulkner, James W. Fickett, William M. Fischer, Mark Good, Deborah A. Horne, F. Kay Houghton, Praful M. Kelkar, Tom A. Kelley, et al.

"GenBank: Current status and future directions"

Methods in Enzymology, 1990, 183, 3-22

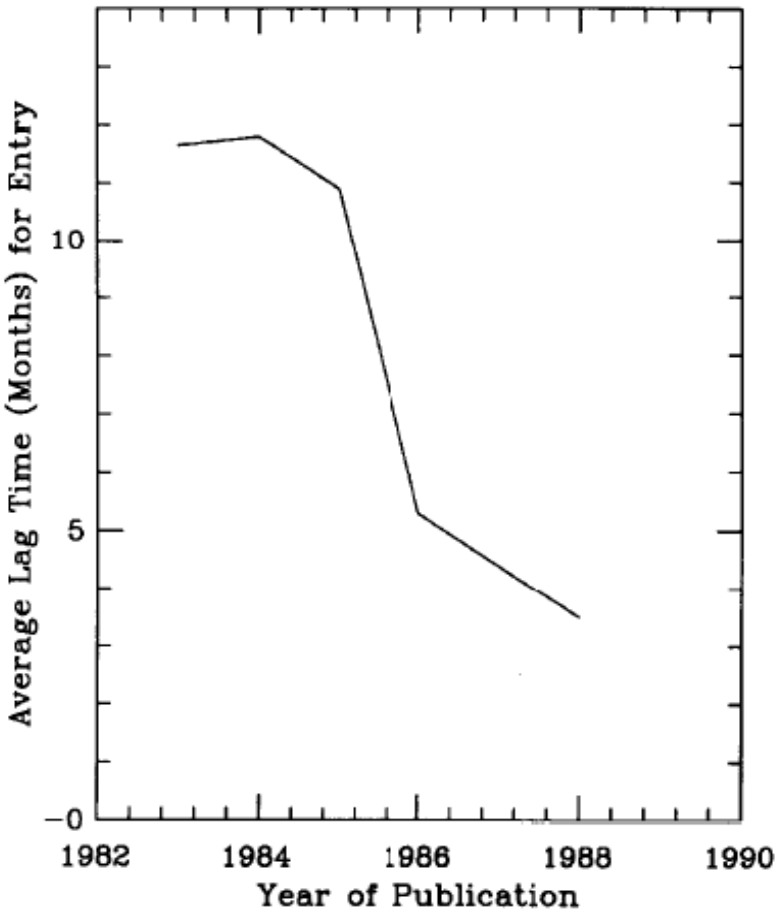


FIG. 2. Lag time for data collected by GenBank to enter the database. The lag time corresponds to the delay between appearance of sequence data in a published journal article and the first appearance of those data in a GenBank release. Note that figures are based on only articles appearing in journals scanned directly by GenBank.

Τα πέτρινα χρόνια (II)

- ... και διανέμονταν με υπερσύγχρονα μέσα ...

Patricia Kahn and Graham Cameron

"EMBL Data Library"

Methods in Enzymology, 1990, 183, 23-31

DISTRIBUTION CHARGES

Subscriber category	Fee per release[b]		Fee per year[b]	
	Tape	CD-ROM	Tape	CD-ROM
Academic users				
EMBL member states[a]	DM 50	DM 100	DM 200	DM 400
Nonmember states	DM 100	DM 200	DM 400	DM 800
Industrial/commercial users	DM 250	DM 500	DM 1000	DM 2000

Magnetic Tape Subscriptions

The main way of distributing copies of the entire database continues to be by mailing magnetic tapes. Much of this distribution is done by the Data Library, and some is done by secondary distributors, such as groups which supply the data along with sequence analysis software. The four releases of nucleotide sequence data in 1988 were supplied to an estimated total user community of approximately 10,000 scientists.

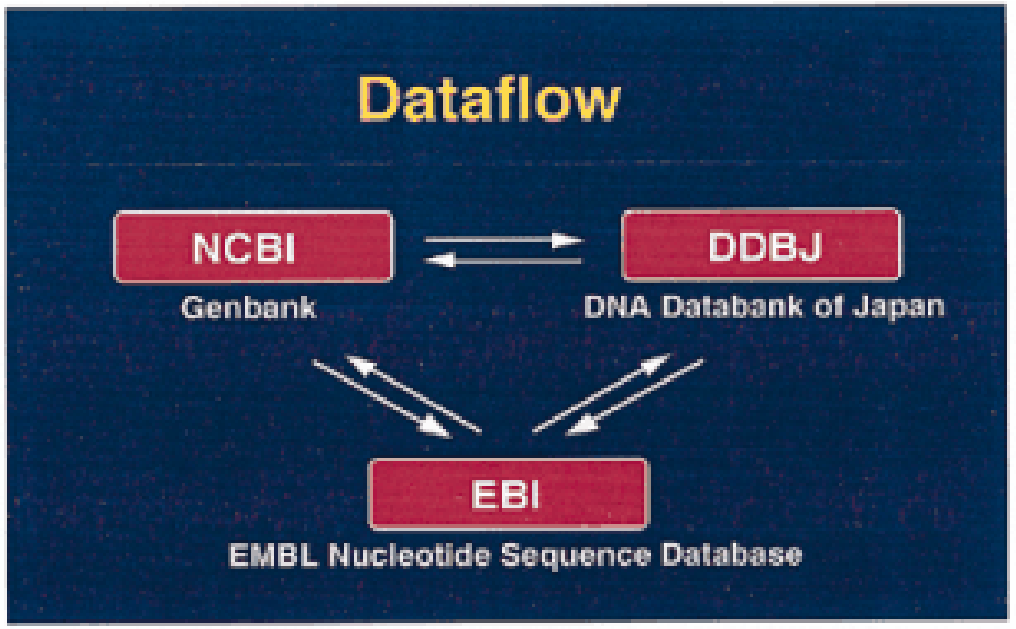
CD-ROM

CD-ROM is attractive as a medium on which to distribute the sequence databases because it represents an inexpensive way to store large quantities data and because the devices required to read it are within financial reach of the typical personal computer user. In early 1989 the Data Library produced a prototype disk which includes prototype CD-ROM retrieval

Network Access

The rapid pace of research in molecular biology has generated a requirement for better and more rapid access to the databases than that provided by quarterly releases. As part of an attempt to meet this need, EMBL set up in early 1988 a file server which enables researchers worldwide to retrieve entries from the major databases available at EMBL via

International Nucleotide Sequence Database Collaboration



http://www.ebi.ac.uk/embl/Contact/dataflow.gif

- Η κάθε ΒΔ έχει δικό της μηχανισμό κατάθεσης (submission) και αναθεώρησης (update) δεδομένων
- Ανάκτηση
  - GenBank – NCBI Entrez
  - EMBL – SRS
  - DDBJ – getentry
- Κοινός μηχανισμός "αμοιβαίας" ενημέρωσης

Μορφές αναπαράστασης δεδομένων (database formats)

- Αλληλουχίες σε ψηφιακή μορφή
  - Αναγνώσιμες από Η/Υ (machine readable)
  - Περιεκτικές
  - Περιγραφικές
- Διαφορετικά επίπεδα λεπτομέρειας – ανάγκες
- Flatfile (text) VS binary formats

Η απλή λύση – FASTA format

```
>gi|394765|emb|X70508.1| Homo sapiens mRNA for insulinoma pre-proinsulin
GCTGCATCAGAAGAGGCCATCAAGCACATCACTGTCCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCC
CCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGC
TCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGCC
GGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCC
CTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCATCTGCTCCCTC
TACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCCCCCACCCGCCGCCTCCTGCACCG
AGAGAGATGGAATAAAGCCCTTGAACCAGC
```

Τίτλος/Περιγραφή

```
>gi|394765|emb|X70508.1| Homo sapiens mRNA for insulinoma pre-proinsulin
```

Identifiers

Description

Αλληλουχία

```
GCTGCATCAGAAGAGGCCATCAAGCACATCACTGTCCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCC
CCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAACCAACACCTGTGCGGC
TCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGCC
GGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCC
CTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCATCTGCTCCCTC
TACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCCCCCACCCGCCGCCTCCTGCACCG
AGAGAGATGGAATAAAGCCCTTGAACCAGC
```

Λεπτομερή formats

- Ας δούμε την προηγούμενη εγγραφή στη GenBank ...
  - http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nuccore&id=X70508
- EMBL ...
  - http://www.ebi.ac.uk/cgi-bin/expasyfetch?X70508
- DDBJ ...
  - http://getentry.ddbj.nig.ac.jp/search/get_entry?accnumber=X70508

Υποδιαιρέσεις / Προέλευση δεδομένων

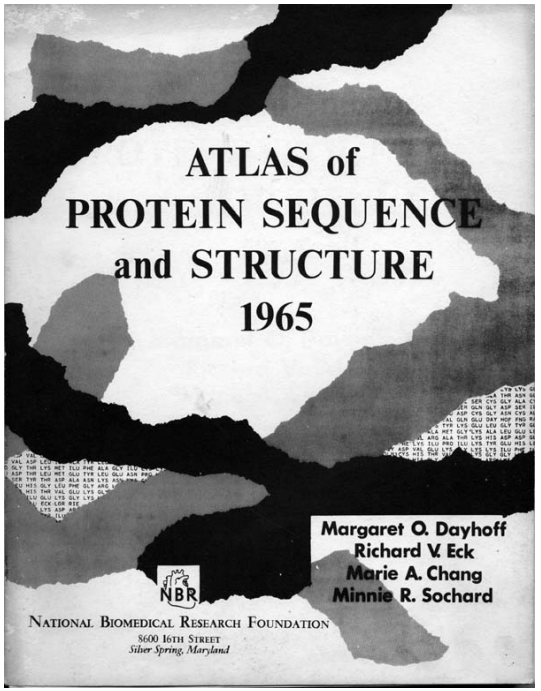
- EST (Expressed Sequence Tags)
- STS (Sequence-Tagged Sites)
- GSS (Genome Survey Sequences)
- HTG (High-Throughput Genome Sequences)
- HTC (unfinished sequences from HTGs)
- WGS (Whole Genome Shotgun sequences)
- PAT (Patent sequences)
- CON (Constructed chromosomes, genomes, etc)

Πλεονασμός (??)

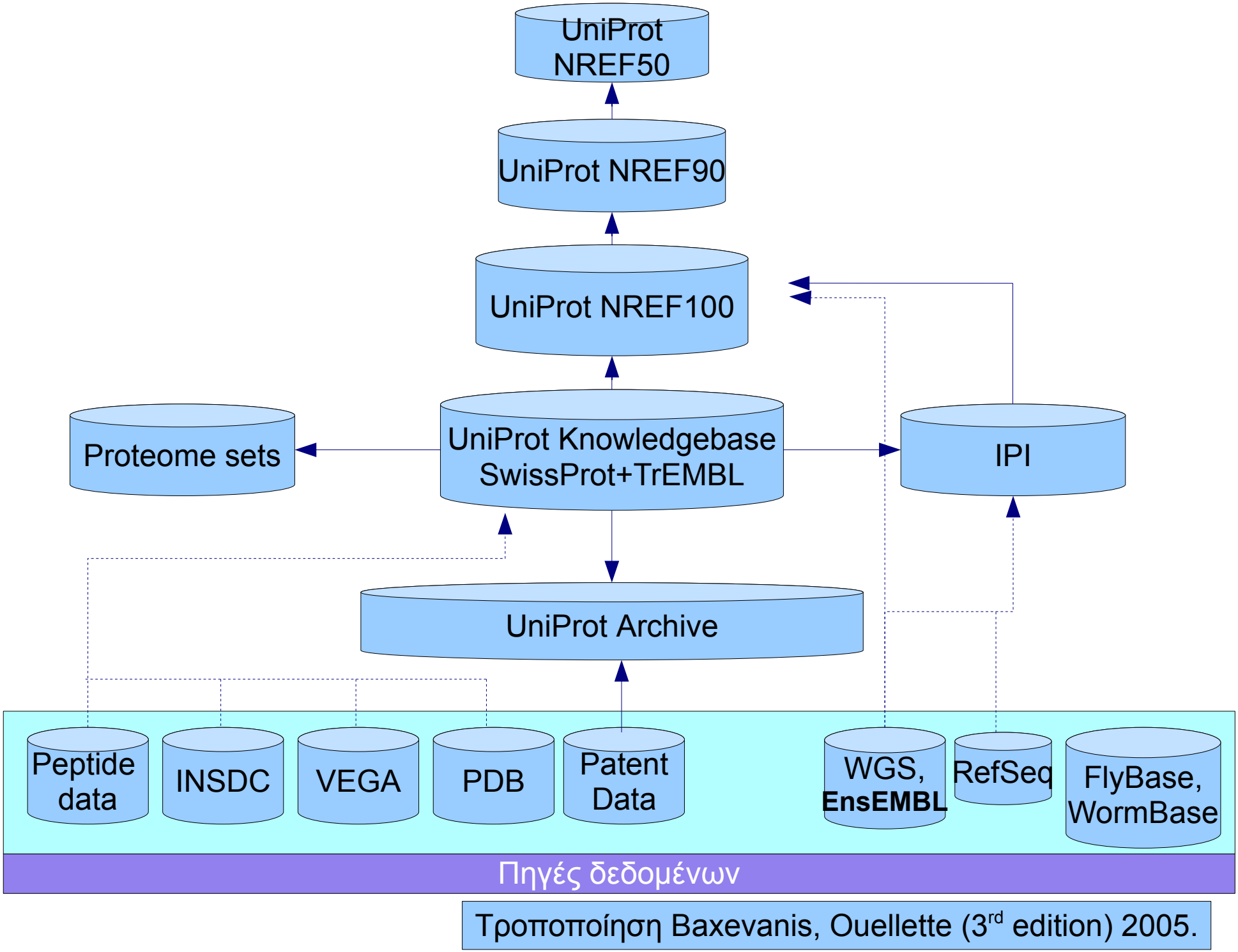
- Π.χ. αλληλουχίες του ίδιου γονιδίου, από διαφορετικό άτομο του ίδιου οργανισμό
- Ορισμένες φορές χρήσιμος (π.χ. μελέτη πολυμορφισμών – πληθυσμιακών χαρακτηριστικών)
- Άλλες φορές πρόβλημα (π.χ. ταυτοποίηση γονιδίων)
- Λύσεις: Σύνολα αναφοράς – Ομαδοποίηση
  - π.χ. NCBI-RefSeq (δευτερογενής ΒΔ)

Βάσεις δεδομένων πρωτεϊνικών αλληλουχιών

- Λίγη ιστορία ..
  - Atlas of protein sequence and structure (1965-1978)



- PIR-PSD (NBRF/MIPS/JIPID), SwissProt
- UniProt (EBI/SIB/PIR)
  - UniParc (archive)
  - UniProtKB
  - UniRef
- GenPept
- RefSeq
- TREMBL



Τροποποίηση Baxevanis, Ouellette (3rd edition) 2005.

Μορφές αναπαράστασης δεδομένων (database formats)

- Αλληλουχίες σε ψηφιακή μορφή
  - Αναγνώσιμες από Η/Υ (machine readable)
  - Περιεκτικές
  - Περιγραφικές
- Διαφορετικά επίπεδα λεπτομέρειας – ανάγκες
- Flatfile (text) VS binary formats

Η απλή λύση – FASTA format

```
>P01308|INS_HUMAN Insulin [Contains: Insulin B chain; Insulin A chain] - Homo sapiens (Human).
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

Τίτλος/Περιγραφή

```
>P01308|INS_HUMAN Insulin [Contains: Insulin B chain; Insulin A chain] - Homo sapiens (Human).
```

Identifiers

Description

Αλληλουχία

```
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

Λεπτομερή formats

http://www.expasy.ch/uniprot/P01308

Συζήτηση ...