BIOINFORMATICS 1

or why biologists need computers

http://www.bioinformatics.uni-muenster.de/teaching/courses-2011/bioinf1/index.hbi

TOPICS TO BE COVERED IN THIS COURSE

- Introduction to bioinformatics from the evolutionary perspective. [WM]
- Principles of heredity. Mutations, substitutions and polymorphisms. [CA]
- Distances and models. Synonymous and nonsynonymous substitutions. Basics of the neutral theory. [CA]
- Sequence alignment and similarity search. [WM]
- Gene prediction. [WM]
- Phylogenetic inference. [CA]
- Protein analyses. [WM]

HANDS ON COMPUTER LAB

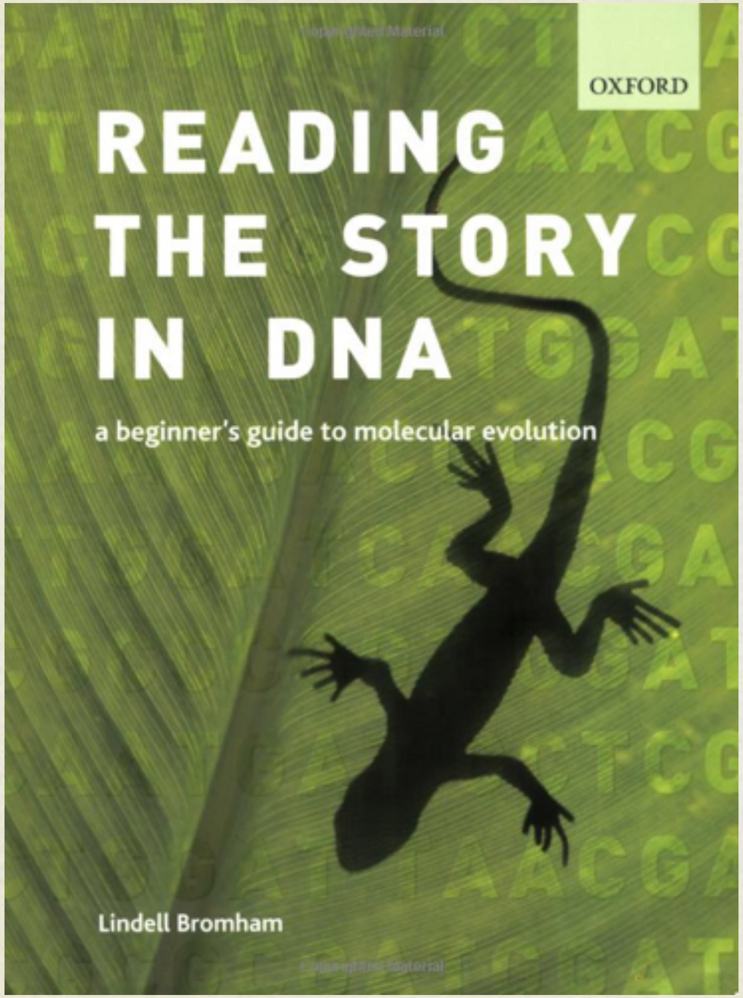
Computer Lab B, Schlossplatz 2b

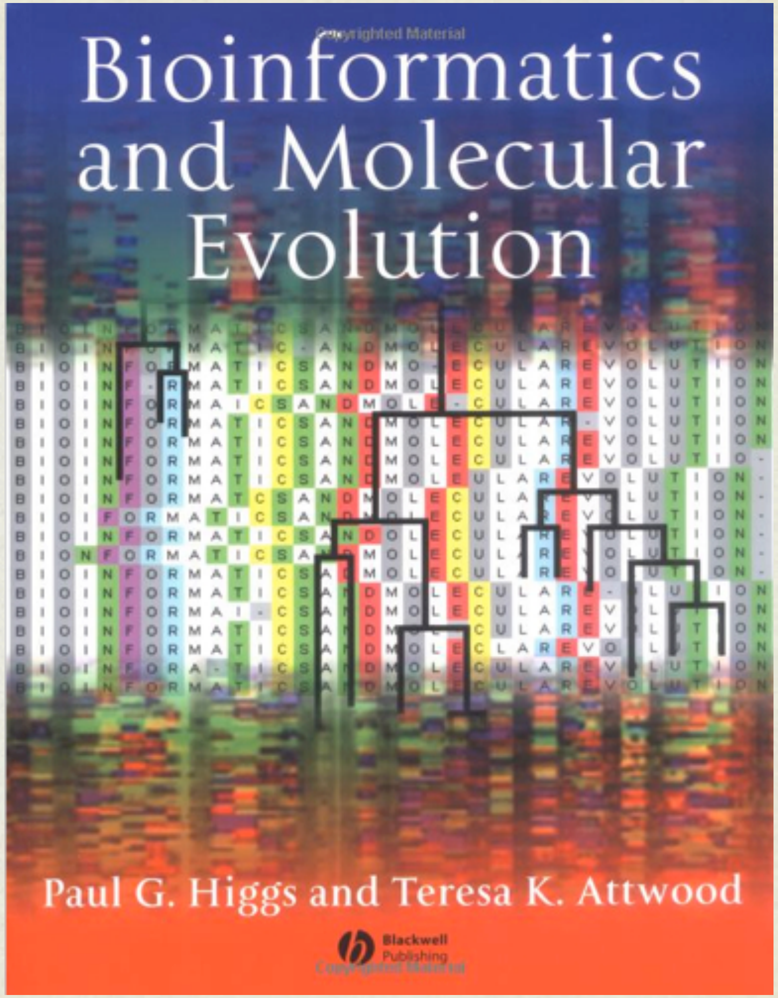
- Alignment and BLAST [November 14]
- Gene prediction [November 21]
- Phylogenetic inference [November 28]
- registration at http://www.bioinformatics.uni-muenster.de/cgi-bin/teaching/coursereg.cgi

CONTACT

- Prof. Claudia Acquisti claudia.acquisti@uni-muenster.de
- Prof. Wojciech Makałowski wojmak@uni-muenster.de
- Robert Fuerst rfuerst@uni-muenster.de (lab coordinator)
- http://www.bioinformatics.uni-muenster.de/teaching/courses-2011/bioinf1/index.hbi
- office hours - see the web site

RECOMMENDED BOOKS



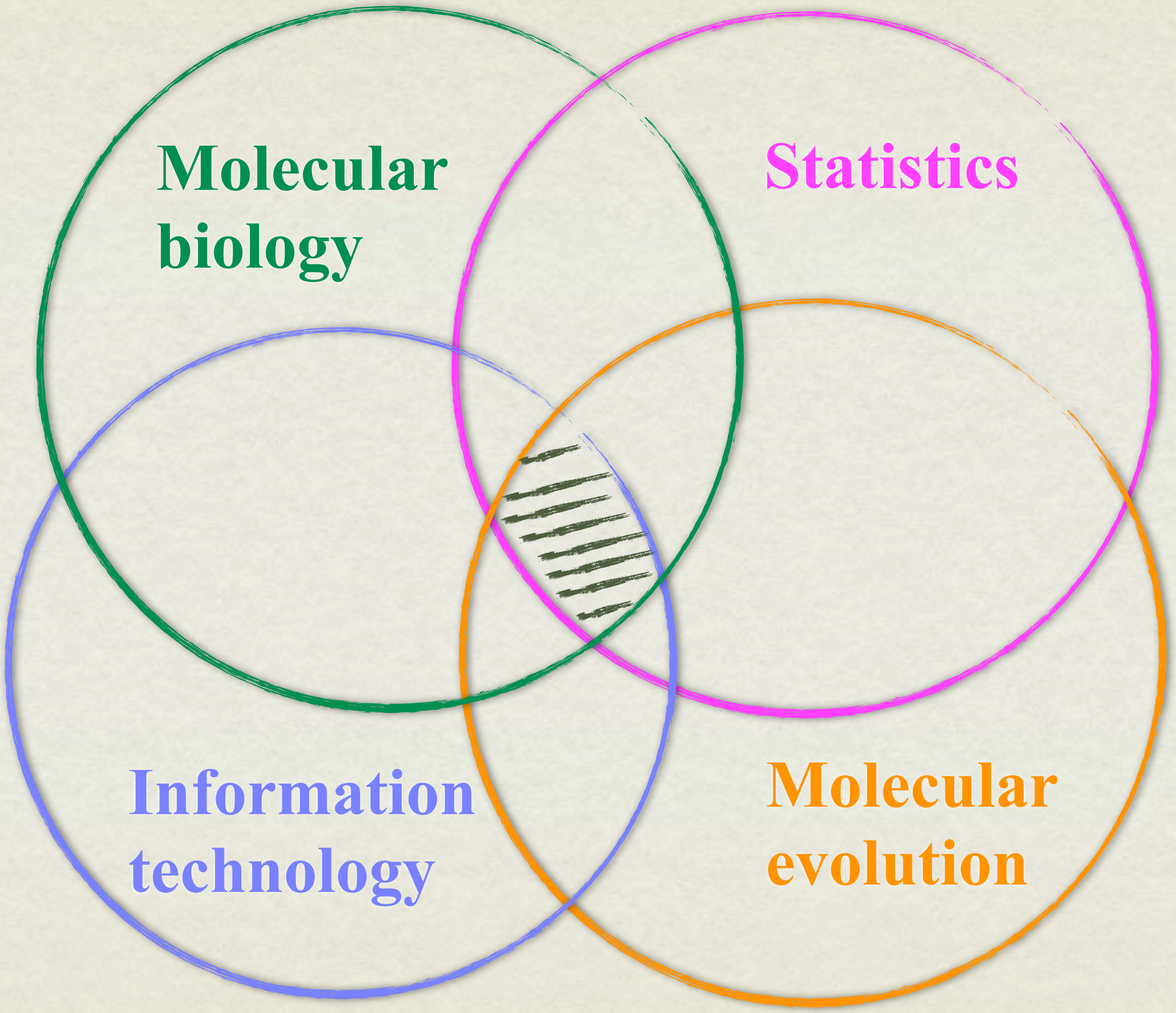


THE ORIGIN OF THE FIELD

Paulien Hogeweg coined the term *bioinformatica* to define “the study of informatic processes in biotic systems”. Hesper B, Hogeweg P (1970) Bioinformatica: een werkconcept. Kameleon 1(6): 28–29. (In Dutch.) Leiden: Leidse Biologen Club.

... but its origin can be tracked back many decades earlier.

BIOINFORMATICS EMERGED AS AN INTERACTION BETWEEN DIFFERENT DISCIPLINES



BIOINFORMATICS - DEFINITION

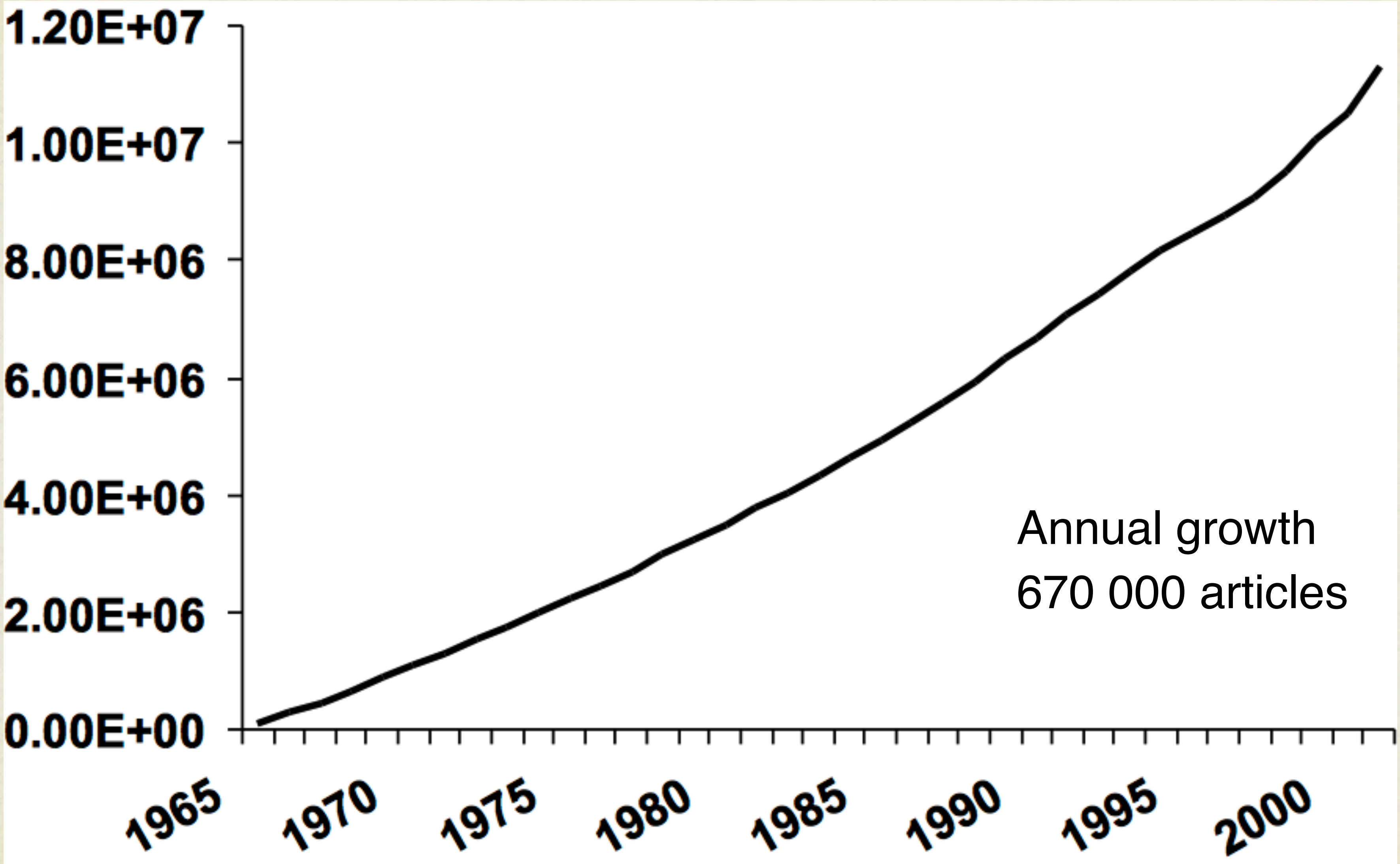
- research, development, or application of computational tools and approaches for expanding the use of biological data, including those to acquire, store, organize, archive, analyze, or visualize such data.
- its goal is to enable biological discovery based on existing information or in other words transform biological information into knowledge

ROLE OF BIOINFORMATICS IN MODERN BIOLOGY

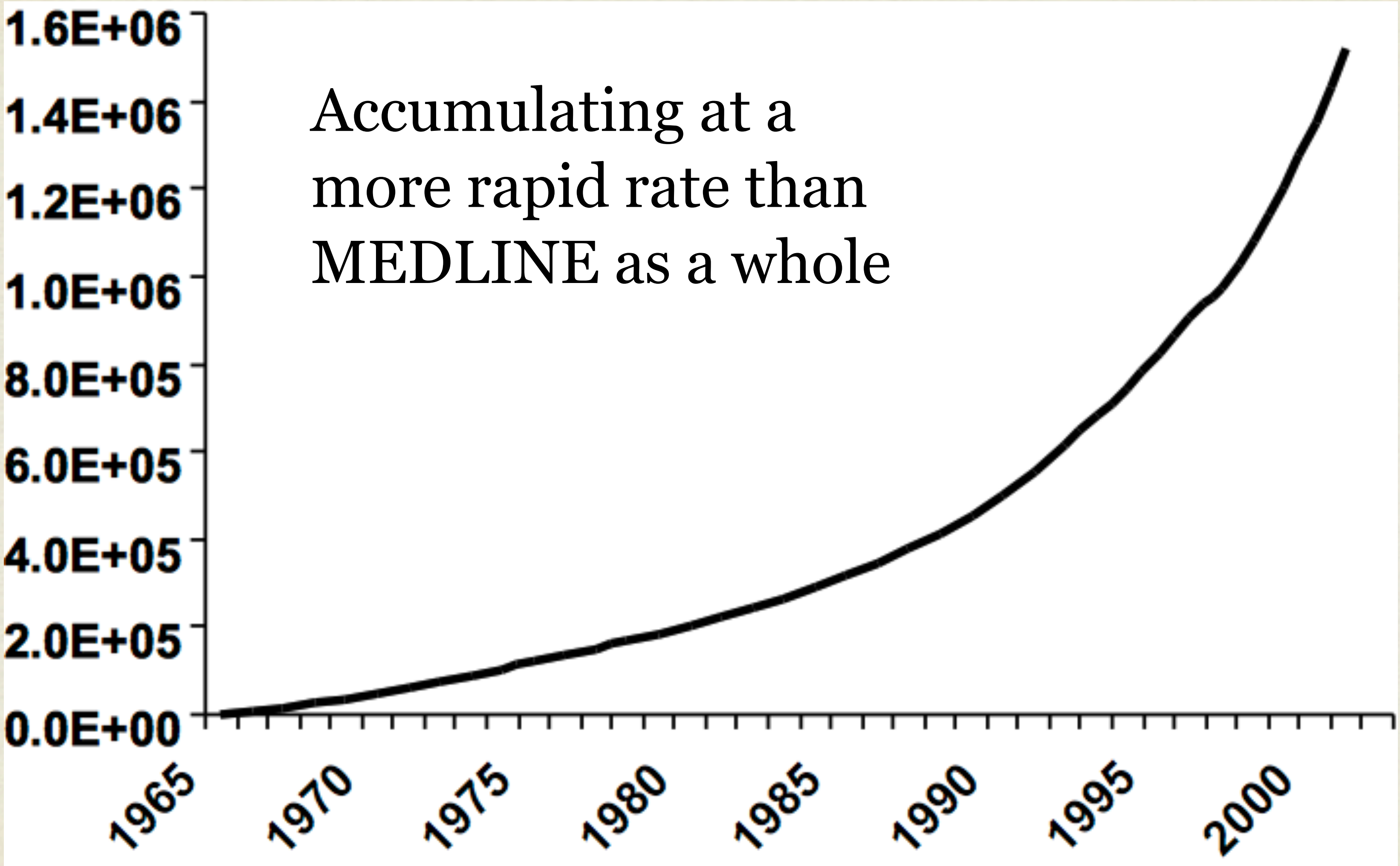
- molecular biology
- molecular evolution
- genomics
- system biology
- protein engineering
- drug design
- personalized medicine


It's sink or swim as a tidal wave of data approaches
Nature 399:517 10 June 1999

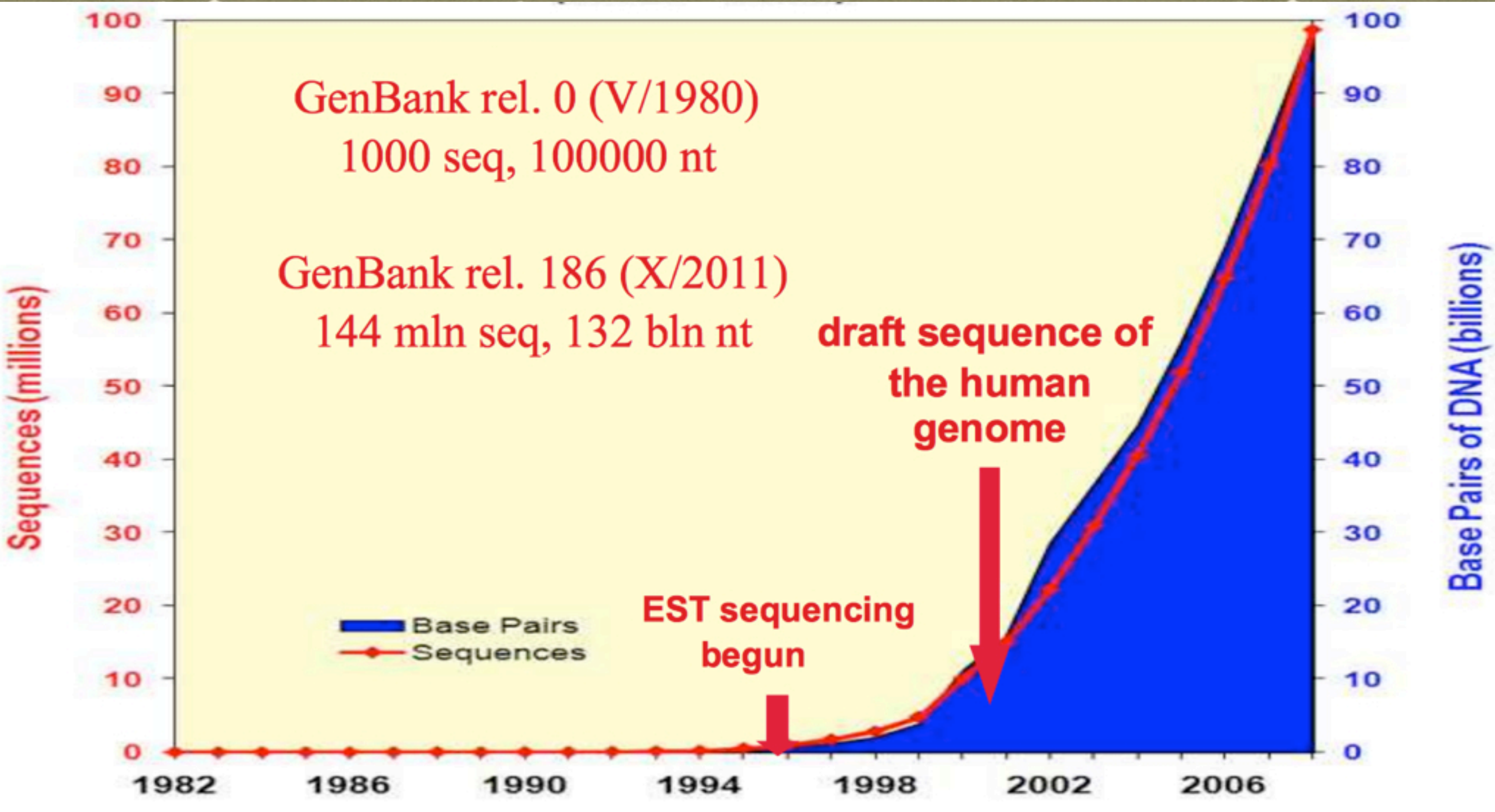
GROWTH OF BIOMEDICAL INFORMATION - MEDLINE



GROWTH OF BIOMEDICAL INFORMATION - SECTION G5 OF MEDLINE - MOL BIOL & GENETICS



GROWTH OF BIOMEDICAL INFORMATION - GENBANK



BIOLOGICAL DATABASES

BIOLOGICAL DATABASES

- organized sets of large amount of data, usually coupled with a software that enables data search, information extraction, and data update
- databases should be characterized by
  - easy data access
  - the possibility to extract only the information that is desirable

INFORMATION IN DATABASES

- Databases and resources may contain many different kinds of information. Each item of entry is typically called an entry. Regardless of the type of resource, each entry comprises two main parts, each broken into one or more fields
- Descriptive information - Annotation
  - Description
  - Literature references
- The raw data – sequence or observations
- The most valuable information is frequently the annotation with the raw data providing a scaffold to organize this curated information.

HISTORICAL (?) LOOK AT DATABASES

- Early systems were file based
  - One entry - one file
  - Lookup based on computer system functions such as grep
- Drawbacks to file-based systems
  - Concurrency
  - No way to check consistency
    - Are values appropriate for fields?
    - Have you updated all necessary information?
  - Unable to limit queries to specific fields
  - Queries and especially updates may be slow and require special programming skills

GENBANK RECORD

```
LOCUS       AF062069     3808 bp    mRNA            INV       02-MAR-2000
DEFINITION  Limulus polyphemus myosin III mRNA, complete cds.
ACCESSION   AF062069
VERSION     AF062069.2  GI:7144484
KEYWORDS    .
SOURCE      Atlantic horseshoe crab.
  ORGANISM  Limulus polyphemus
            Eukaryota; Metazoa; Arthropoda; Chelicerata; Merostomata;
            Xiphosura; Limulidae; Limulus.
REFERENCE   1  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     A myosin III from Limulus eyes is a clock-regulated phosphoprotein
  JOURNAL   J. Neurosci. (1998) In press
REFERENCE   2  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (29-APR-1998) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
REFERENCE   3  (bases 1 to 3808)
  AUTHORS   Battelle,B.-A., Andrews,A.W., Calman,B.G., Sellers,J.R.,
            Greenberg,R.M. and Smith,W.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-MAR-2000) Whitney Laboratory, University of Florida,
            9505 Ocean Shore Blvd., St. Augustine, FL 32086, USA
  REMARK    Sequence update by submitter
COMMENT     On Mar 2, 2000 this sequence version replaced gi:3132700.
```

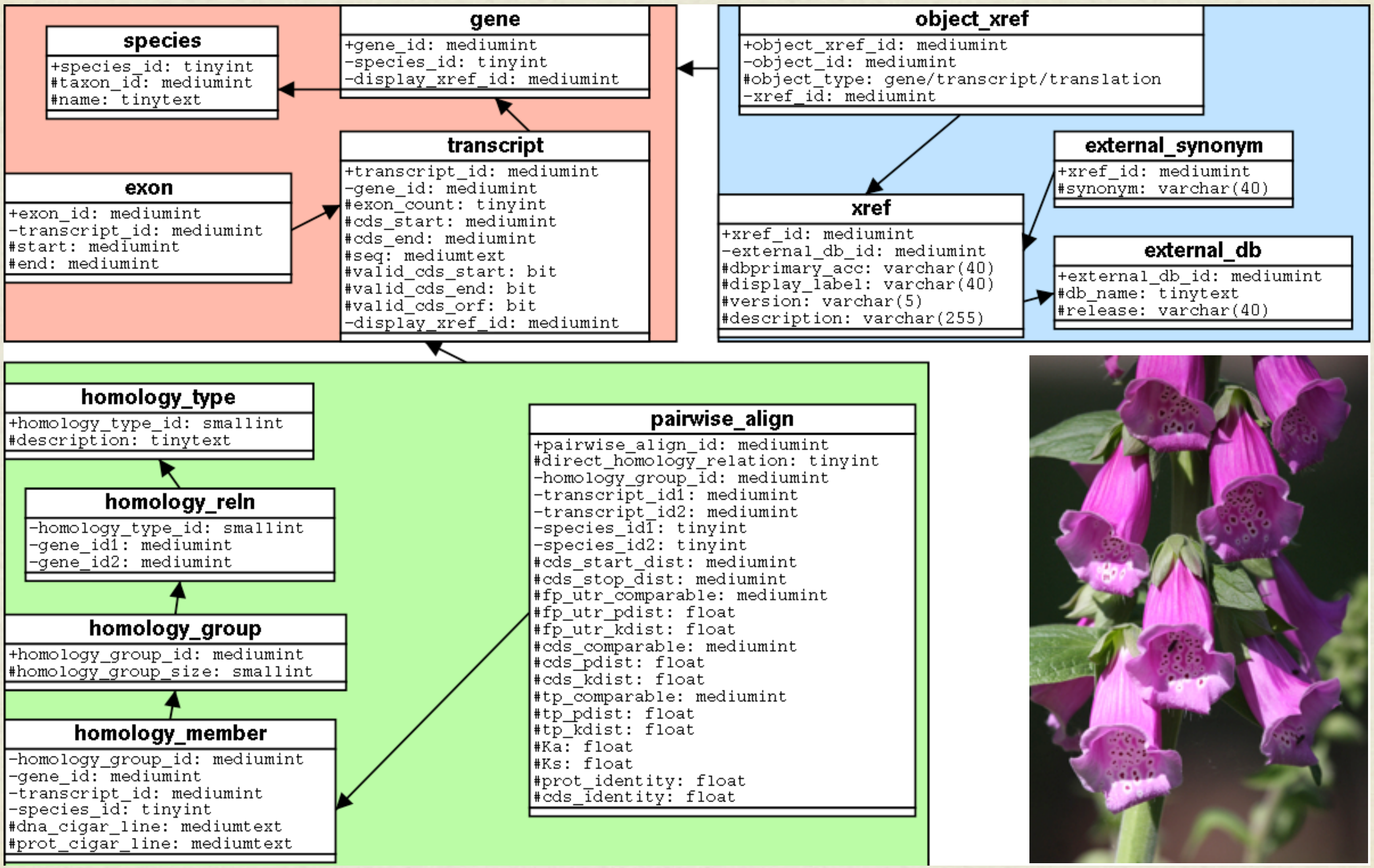
GENBANK RECORD

```
FEATURES             Location/Qualifiers
     source          1..3808
                     /organism="Limulus polyphemus"
                     /db_xref="taxon:6850"
                     /tissue_type="lateral eye"
     CDS             258..3302
                     /note="N-terminal protein kinase domain; C-terminal myosin
                     heavy chain head; substrate for PKA"
                     /codon_start=1
                     /product="myosin III"
                     /protein_id="AAC16332.2"
                     /db_xref="GI:7144485"
                     /translation="MEYKCISEHLPFETLPDPGDRFEVQELVGTGTYATVYSAIDKQA
                     NKKVALKIIGHIAENLLDIETEYRIYKAVNGIQFFPEFRGAFFKRGERESDNEVWLGI
                     EFLEEGTAADLLATHRRFGIHLKEDLIALIIKEVVRAVQYLHENSIIHRDIRAANIMF
                     SKEGYVKLIDFGLSASVKNTNGKAQSSVGSPYWMAPEVISCDCLQEPYNYTCDVWSIG
                     ITAIELADTVPSLSDIHALRAMFRINRNPPPSVKRETRWSETLKDFISECLVKNPEYR
                     PCIQEIPQHPFLAQVEGKEDQLRSELVDILKKNPGEKLRNKPYNVTFKNGHLKTISGQ
BASE COUNT     1201 a    689 c    782 g   1136 t
ORIGIN
        1 tcgacatctg tggtcgcttt ttttagtaat aaaaaattgt attatgacgt cctatctgtt
     3781 aagatacagt aactagggaa aaaaaaaa
//
```

MODERN RESOURCES

- Relational Database Management Systems (RDBMS)
  - Introduced in the 1970s
  - Commercial, off-the-shelf software
    - Oracle, DB2, MySQL
  - High level declarative language - SQL
  - Concurrency
  - Transaction control
  - Consistency

RELATIONAL DATABASES - AN EXAMPLE



CRITICAL ISSUES FOR BIOLOGICAL DATABASES

- Annotation
  - Correctness
  - Consistency
  - Quality
- Archival Quality
- Updates
  - Raw data
  - Annotation

CRITICAL ISSUES ANNOTATION

- Correctness – many genes are annotated primarily based on sequence comparisons. Annotation is copied from a similar sequence to a novel sequence. This may cause some problems
  - Comparison may have been done when the data was less complete
  - If sequence is incorrectly annotated, this error propagates through the database

CRITICAL ISSUES ANNOTATION QUALITY

- Who supplies the annotation? An expert, or a non-expert at the database

- Many databases have defined groups of "experts" to help annotated genes or gene families, but there is no peer-review of information in databases

- What is the vocabulary?

CRITICAL ISSUES ARCHIVAL QUALITY

- Databases have been torn between trying to be archival – to simply report information as experts publish it (*primary databeses*), or curated – to provide the best editorially reviewed data on a topic (*secondary DB*).
- Can the same entry be recovered later?
  - Accession numbers are more stable than entry or locus names
  - Many databases do not note that there have been changes to the data! What you retrieve today may be different than yesterday

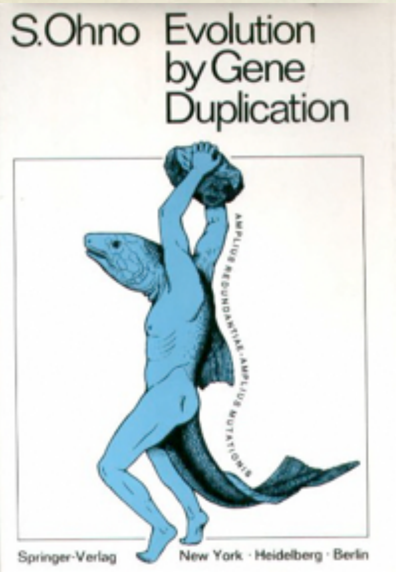
CRITICAL ISSUES UPDATES

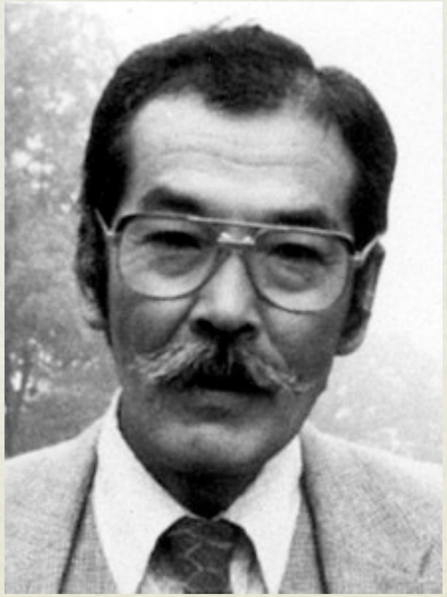
- How often are updates done? Major databases take direct submissions.
- Generally, only the original submitter can change an entry, even if you can prove it is wrong. This is tied to the question of archival versus curated.
- How is annotation updated as more knowledge is available? Who decides?

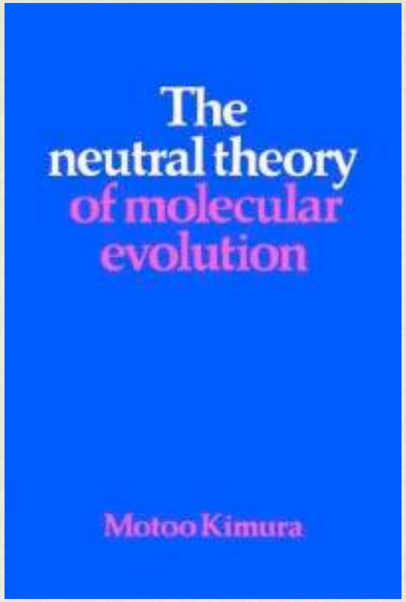
SECONDARY (SPECIALIZED) DATABASES

- Boom of biological databases
- Every year first issue of *Nucleic Acids Research* dedicated to biological databases
    - http://nar.oxfordjournals.org/content/vol39/suppl_1/index.dtl
    - this year's database issue includes 1330 databases - 100 more than last year's list
    - the first collection published in 1993 contained description of 24 databases

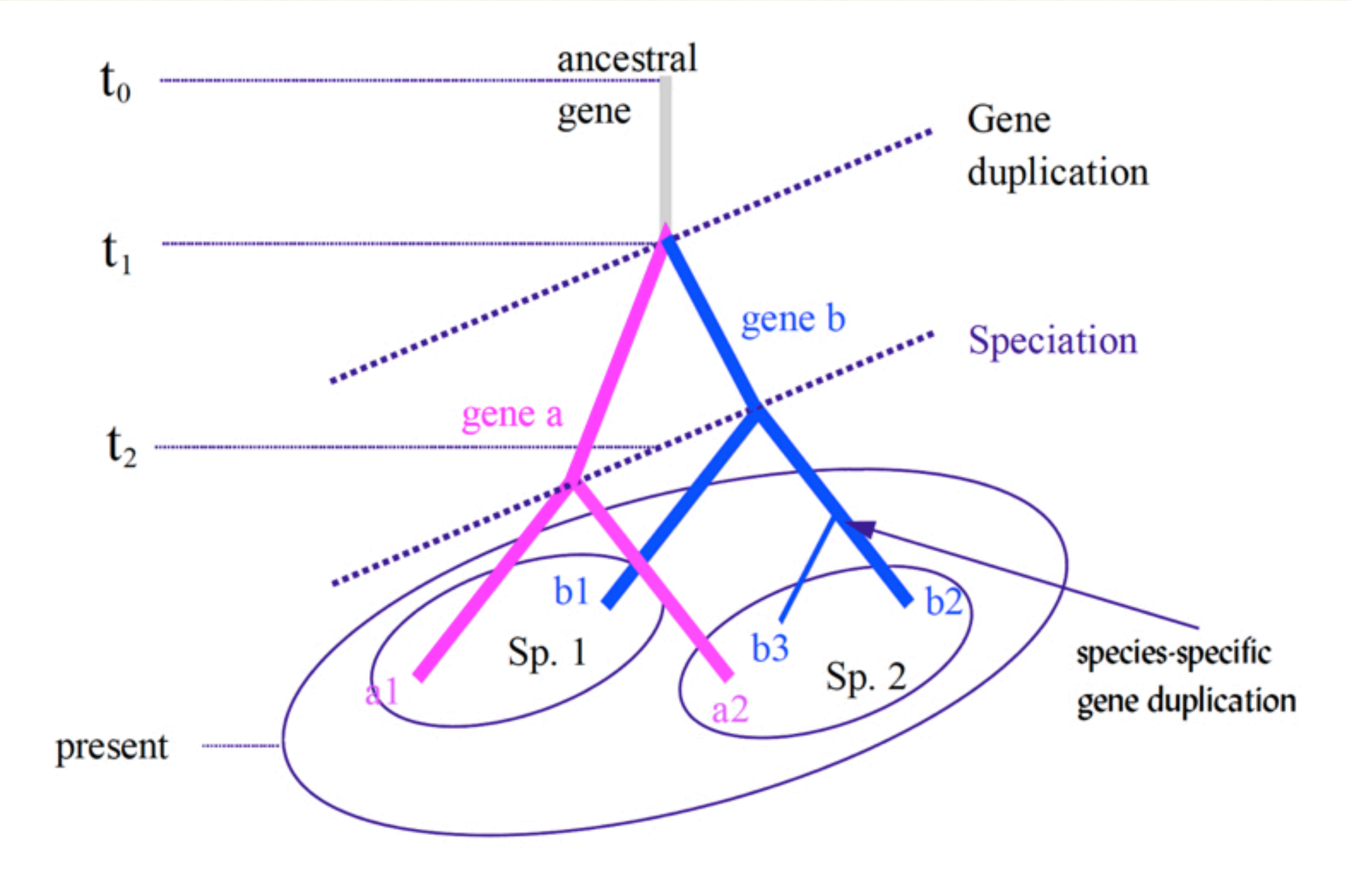
EVOLUTIONARY BASIS OF BIOINFORMATICS



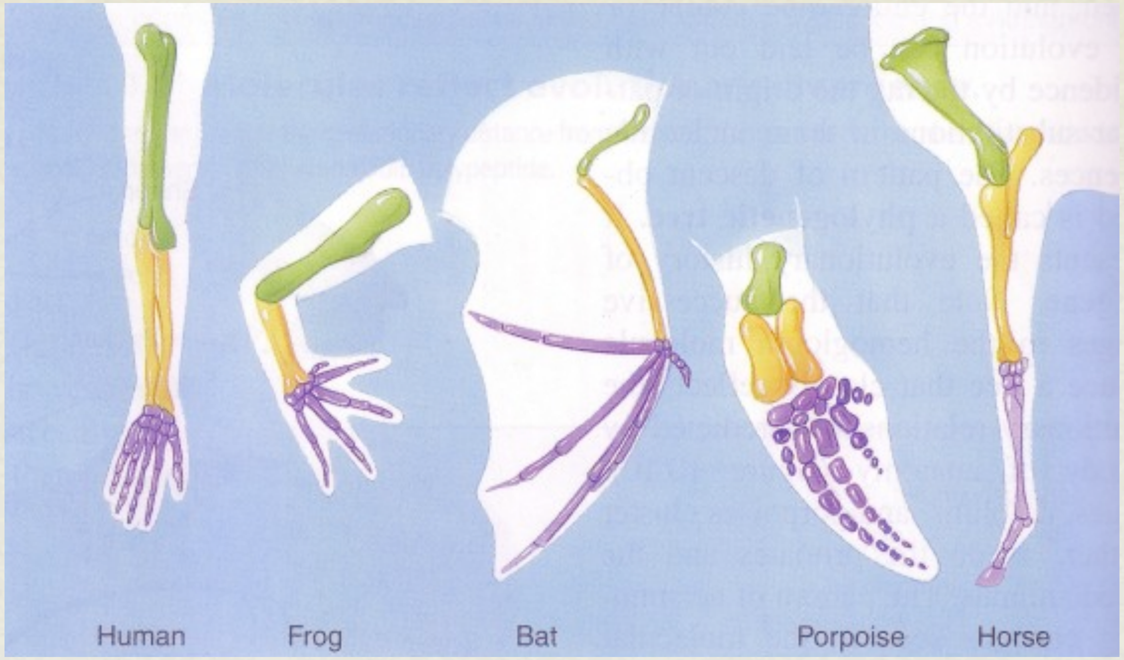




EVOLUTIONARY BASIS OF BIOINFORMATICS



HOMOLOGS



Two anatomical structures or behavioral traits within different organisms which originated from a structure or trait of their common ancestral organism. The structures or traits in their current forms may not necessarily perform the same functions in each organism, nor perform the functions it did in the common ancestor. An example: the wing of a bat, the fin of a whale and the arm of a man are homologous structures.

http://www.everythingbio.com/glos/definition.php?ID=3385

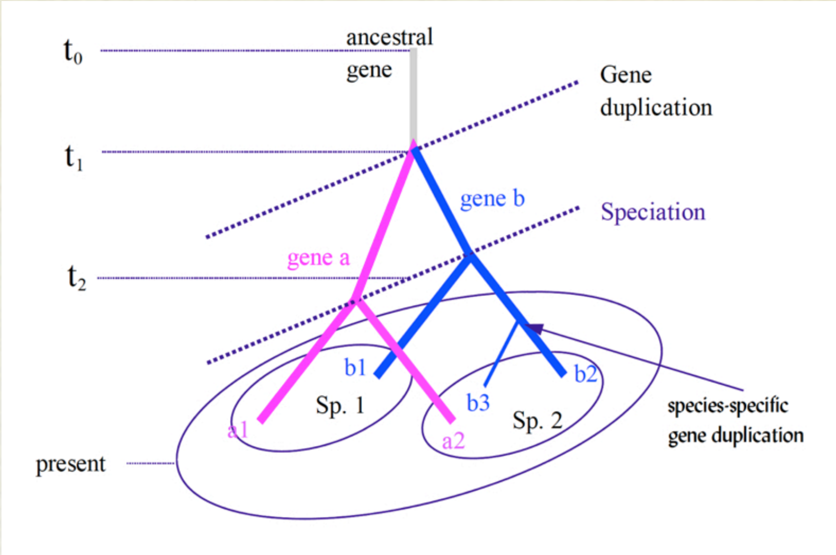
HOMOLOGS AT THE MOLECULAR LEVEL

```
cow      ATG---ACTAACATTCGAAAGTCCCACCCACTAATAAAAATTGTAAAC
sheep    ATG---ATCAACATCCGAAAAACCCACCCACTAATAAAAATTGTAAAC
goat     ATG---ACCAACATCCGAAAGACCCACCCATTAATAAAAATTGTAAAC
horse    ATG---ACAAACATCCGGAAATCTCACCCACTAATTAAAATCATCAAT
donkey   ATG---ACAAACATCCGAAAATCCCACCCGCTAATTAAAATCATCAAT
ostrich  ATGGCCCCCAACATTCGAAAATCGCACCCCCTGCTCAAAATTATCAAC
emu      ATGGCCCCTAACATCCGAAAATCCCACCCTCTACTCAAAATCATCAAC
turkey   ATGGCACCCAATATCCGAAAATCACACCCCCTATTAAAAACAATCAAC
```

Two sequences that share common ancestry. Significant sequence similarity usually suggests homology, however sequence similarity may occur also by chance and some homologous sequences may diverge beyond detectable similarity.

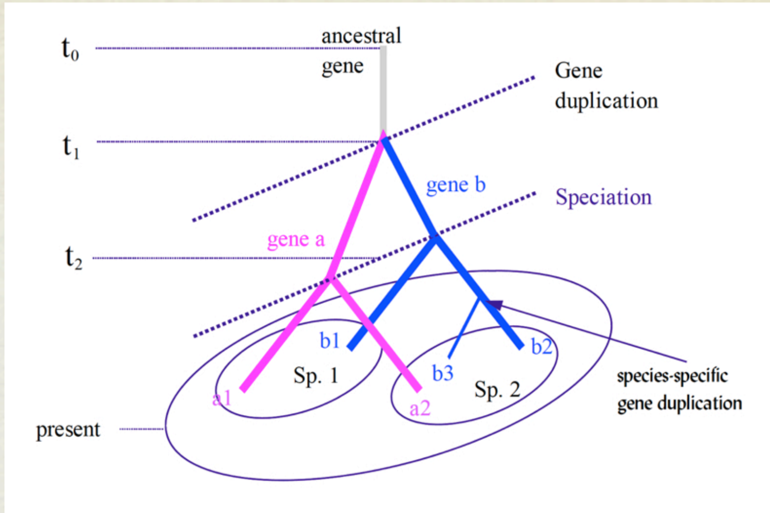
HOMOLOGS: ORTHOLOGS AND PARALOGS

ORTHOLOGS. Genes or sequences that result from a speciation event followed by a sequence divergence. Such genes may not exist side by side in the same genome. The last common ancestor of two orthologous sequences existed just before speciation event.

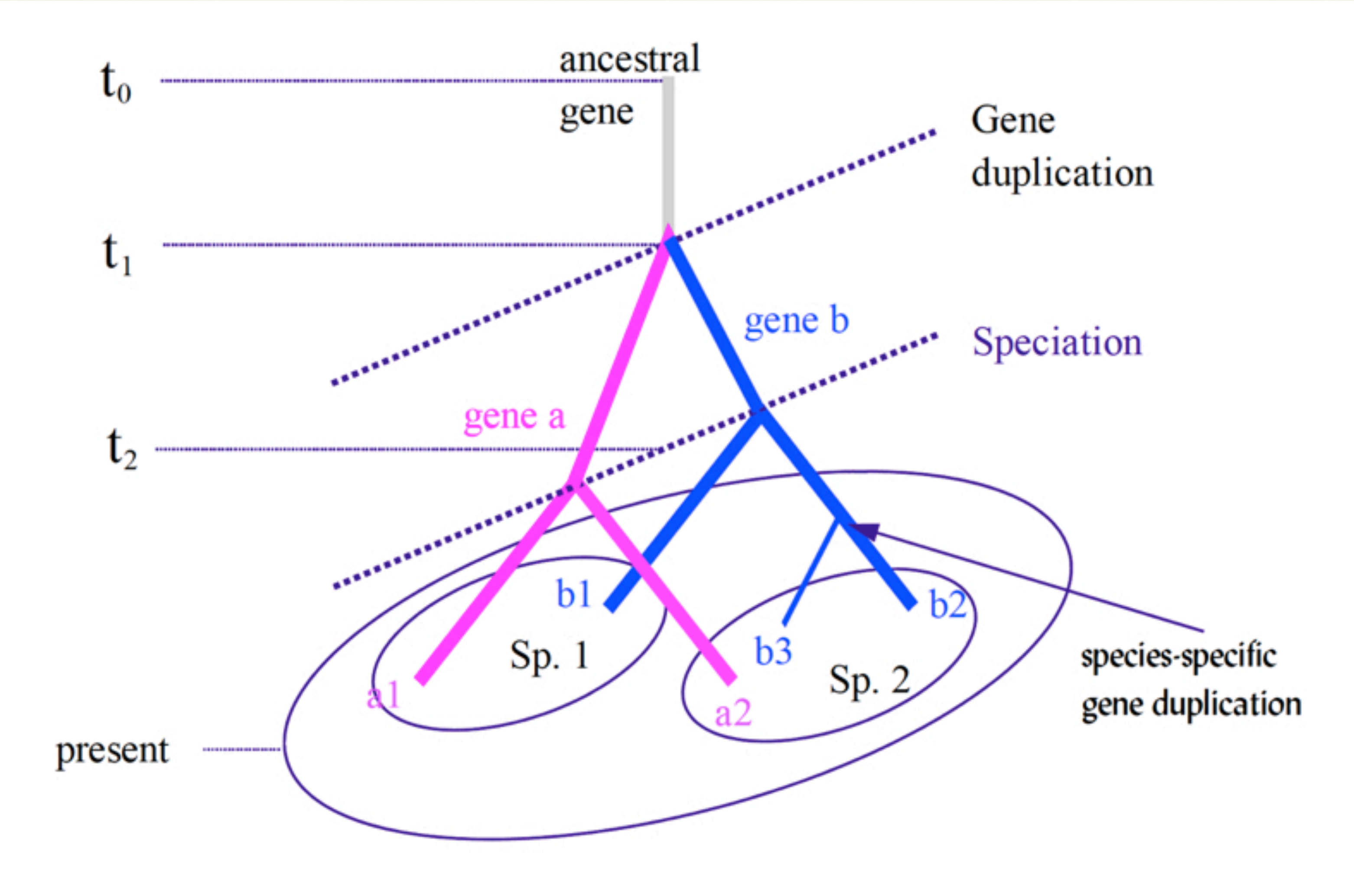


HOMOLOGS: ORTHOLOGS AND PARALOGS

PARALOGS. Genes or sequences that resulted from duplication of genetic material followed by a sequence divergence. Such genes may descend and diverge while existing side by side in the same genome. If speciation occurs after gene duplication, then two paralogous genes may exist in two different genomes. The last common ancestor of two paralogous sequences existed just before duplication event.

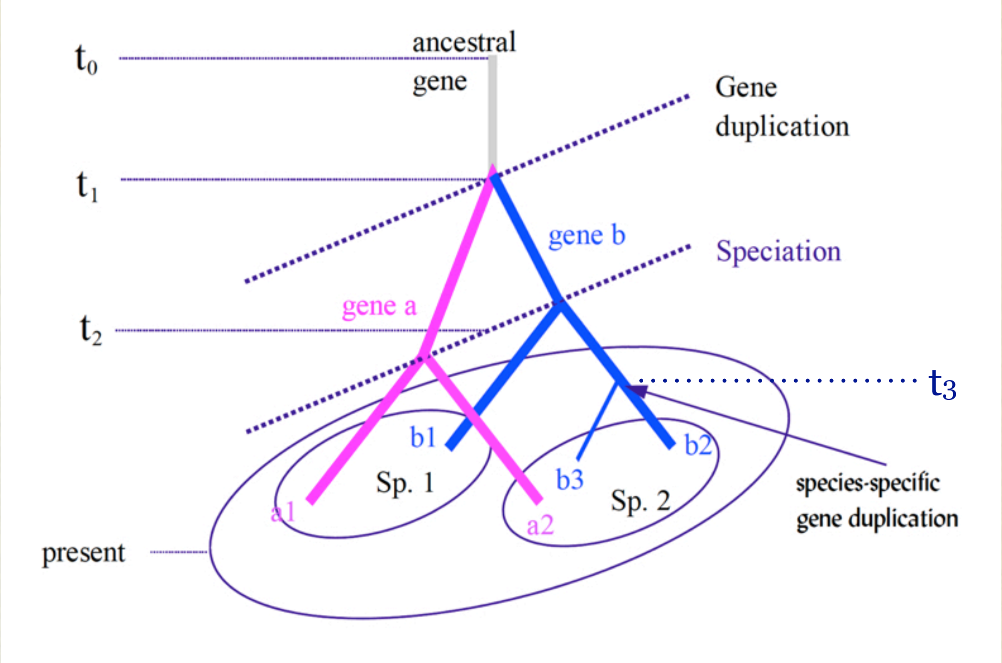


EVOLUTIONARY BASIS OF BIOINFORMATICS



HOMOLOGS: ORTHOLOGS AND PARALOGS

Compared Genes	Relation	Time of last comm. ancestor	Evolutionary event at the time of last common ancestor	Presence in the same species
A - B	paralogy	t_1	gene duplication	yes
A1 - A2	orthology	t_2	speciation	no
A1 - B1	paralogy	t_1	gene duplication	yes
A1 - B2	paralogy	t_1	gene duplication	no
A1 - B3	paralogy	t_1	gene duplication	no
A2 - A1	orthology	t_2	speciation	no
A2 - B1	paralogy	t_1	gene duplication	no
A2 - B2	paralogy	t_1	gene duplication	yes
A2 - B3	paralogy	t_1	gene duplication	yes
B1 - A1	paralogy	t_1	gene duplication	yes
B1 - A2	paralogy	t_1	gene duplication	no
B1 - B2	orthology	t_2	speciation	no
B1 - B3	orthology	t_2	speciation	no
B2 - A1	paralogy	t_1	gene duplication	no
B2 - A2	paralogy	t_1	gene duplication	yes
B2 - B1	orthology	t_2	speciation	no
B2 - B3	paralogy	t_3	gene duplication	yes
B3 - A1	paralogy	t_1	gene duplication	yes
B3 - A2	paralogy	t_1	gene duplication	no
B3 - B1	orthology	t_2	speciation	no
B3 - B2	paralogy	t_3	gene duplication	yes

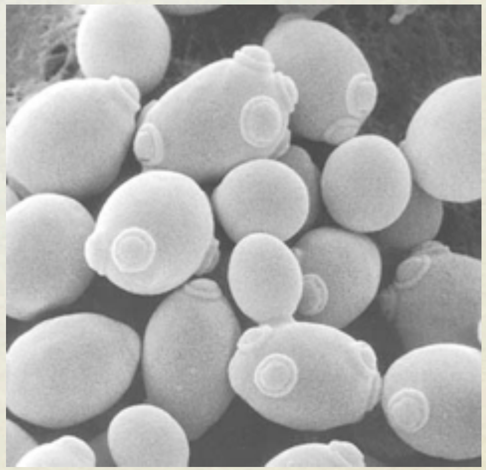


COMPARATIVE GENOMICS

What is true for *E. coli* is also true for elephant.

J. Monod, c. 1961

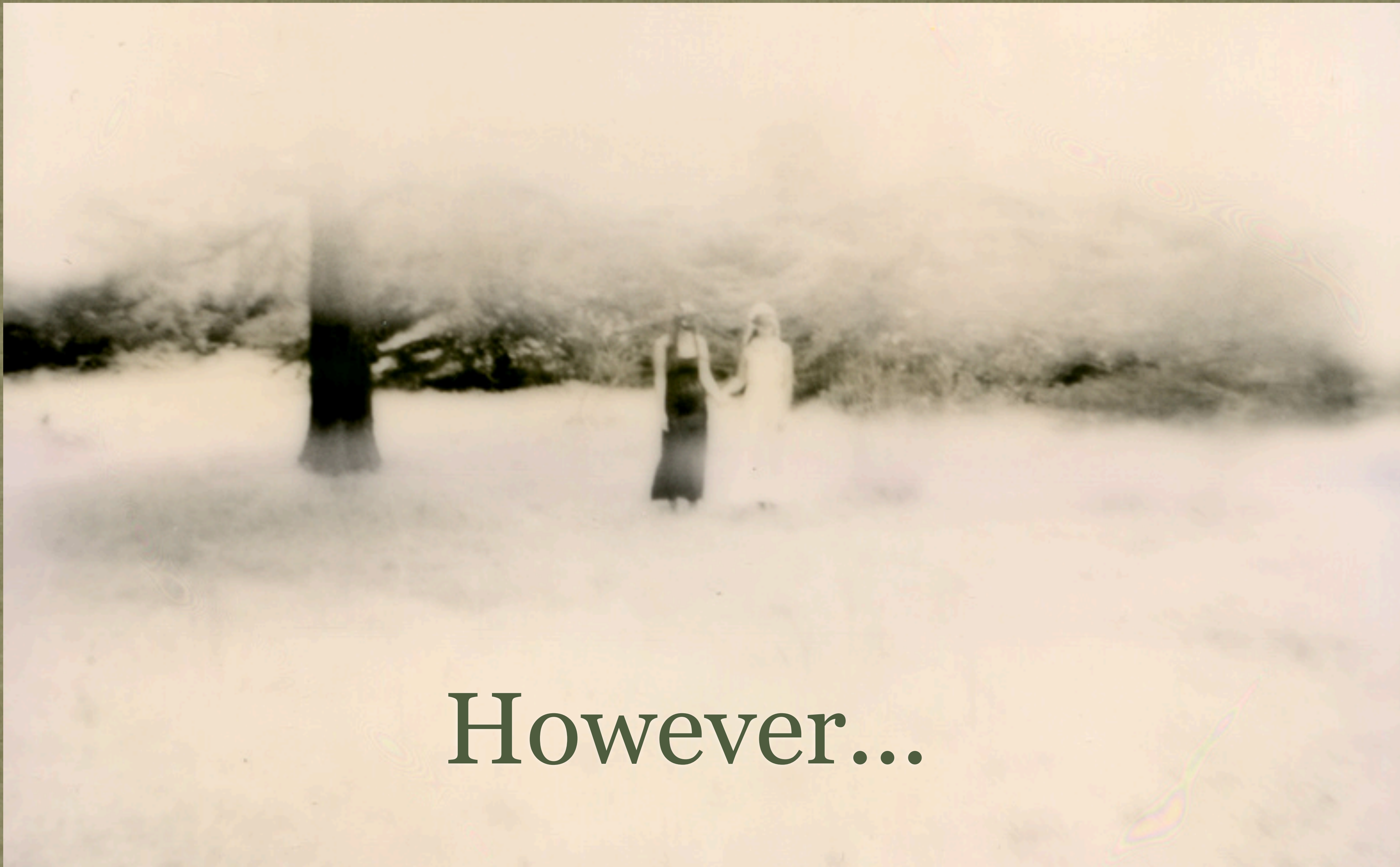
COMPARATIVE GENOMICS



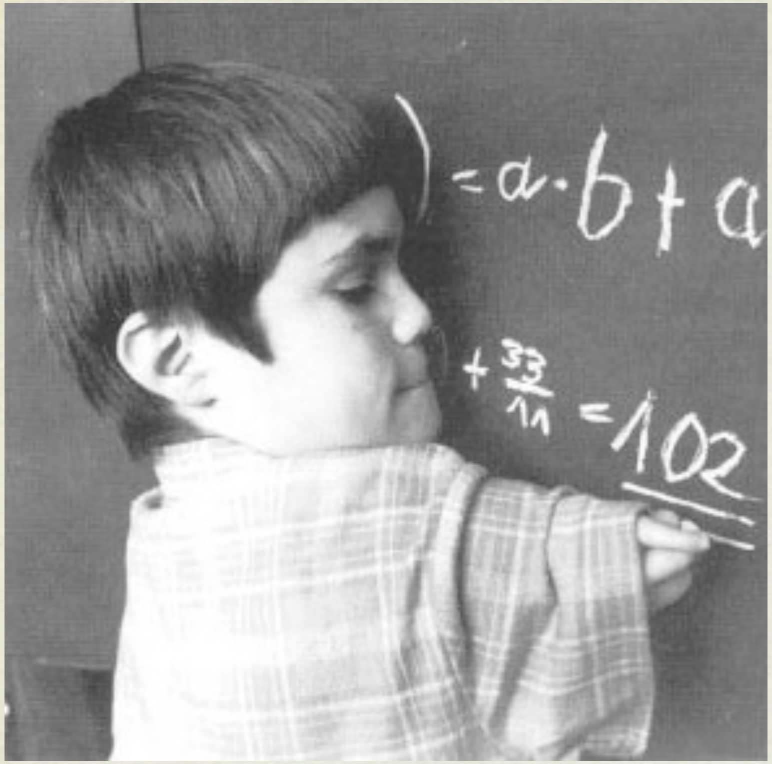
What is true for yeast is also true for human.

D. Botstein, 1988

COMPARATIVE GENOMICS

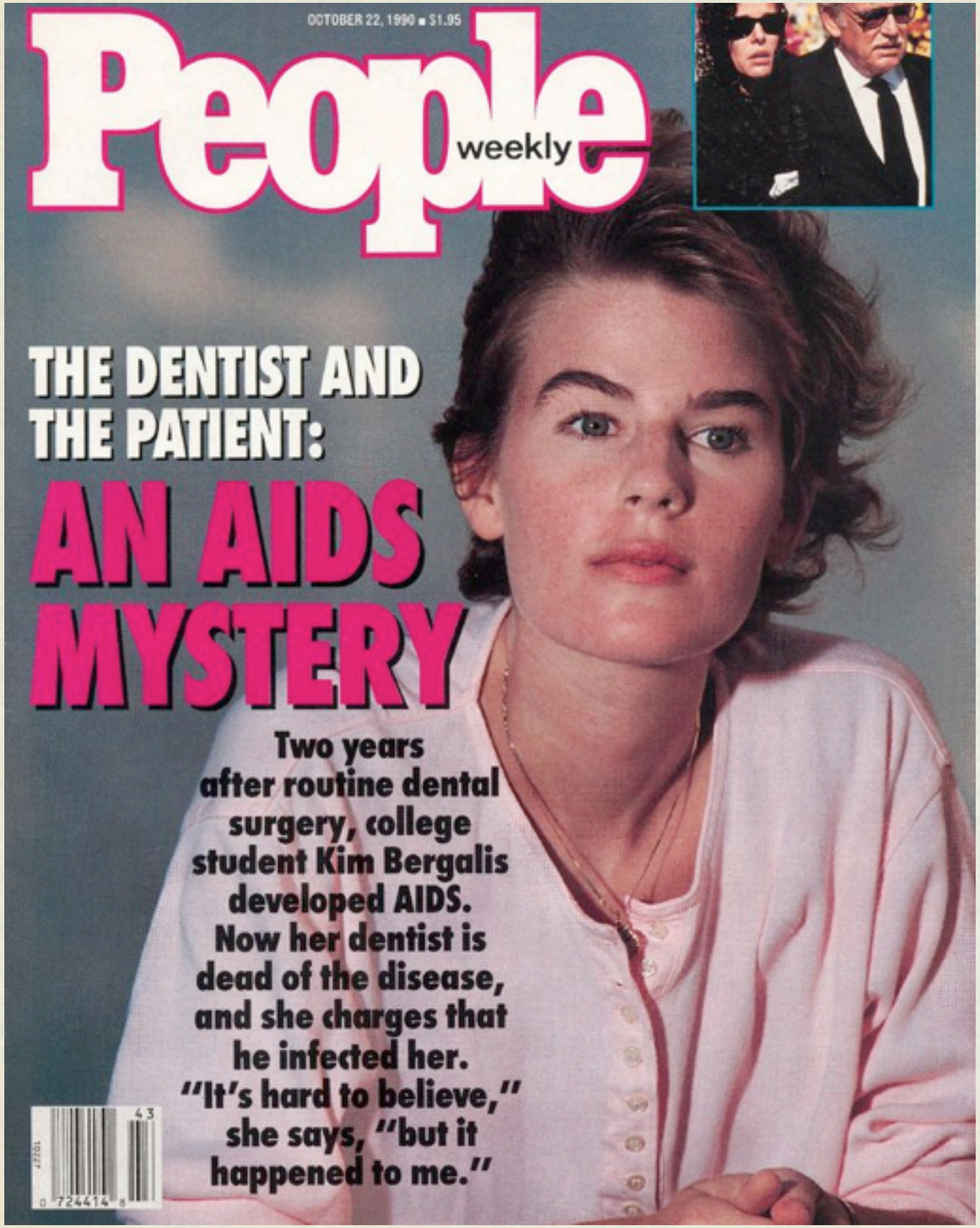


COMPARATIVE GENOMICS



15 000 victims of thalidomide

What is true for mouse is not necessarily true for human...



Did the Florida Dentist infect his patients with HIV?

Kimberly Bergalis

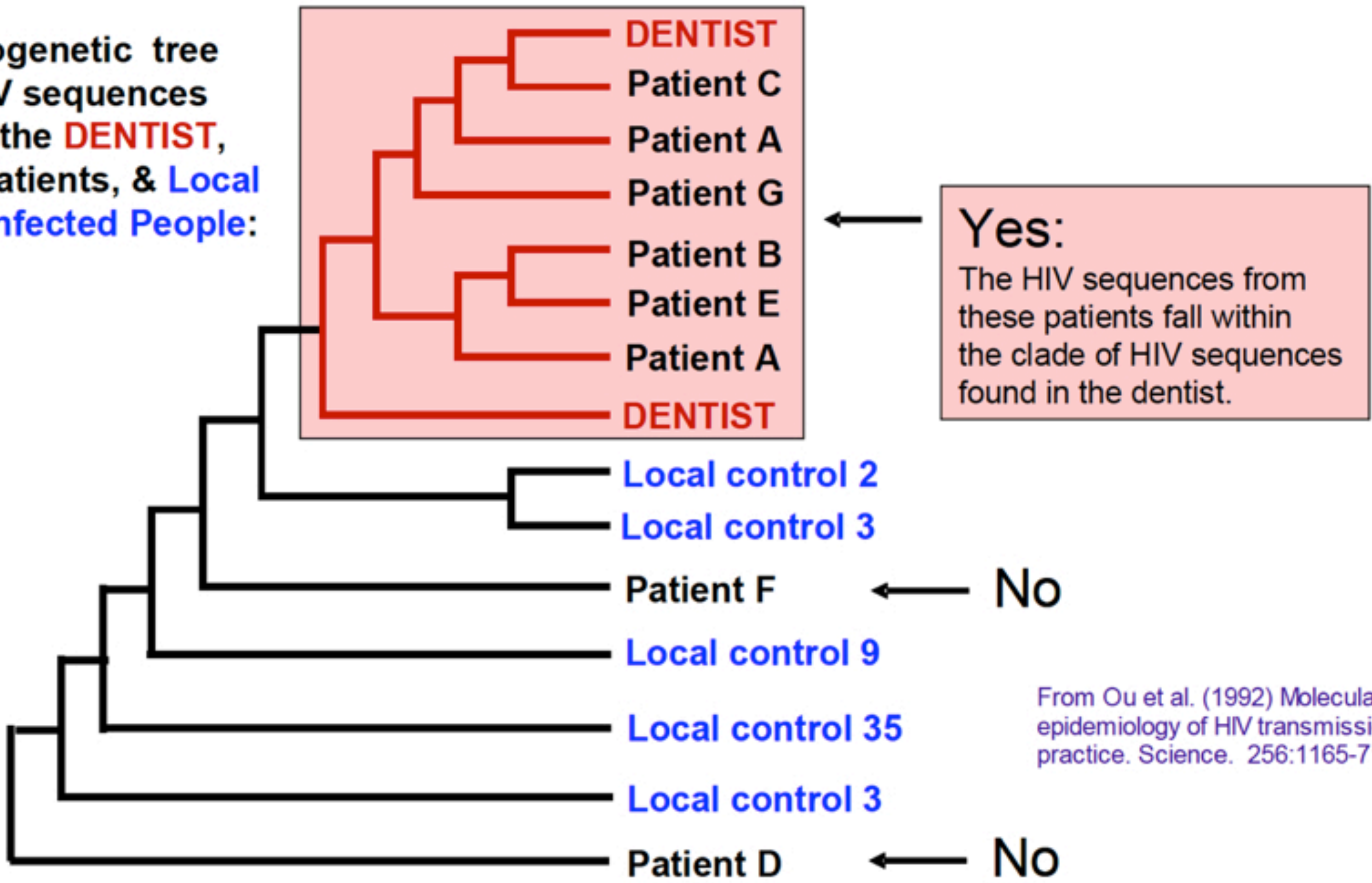
(1968-1991)

David J. Acer

(1940-1990)

DID THE FLORIDA DENTIST INFECT HIS PATIENTS WITH HIV?

Phylogenetic tree of HIV sequences from the DENTIST, his Patients, & Local HIV-infected People:

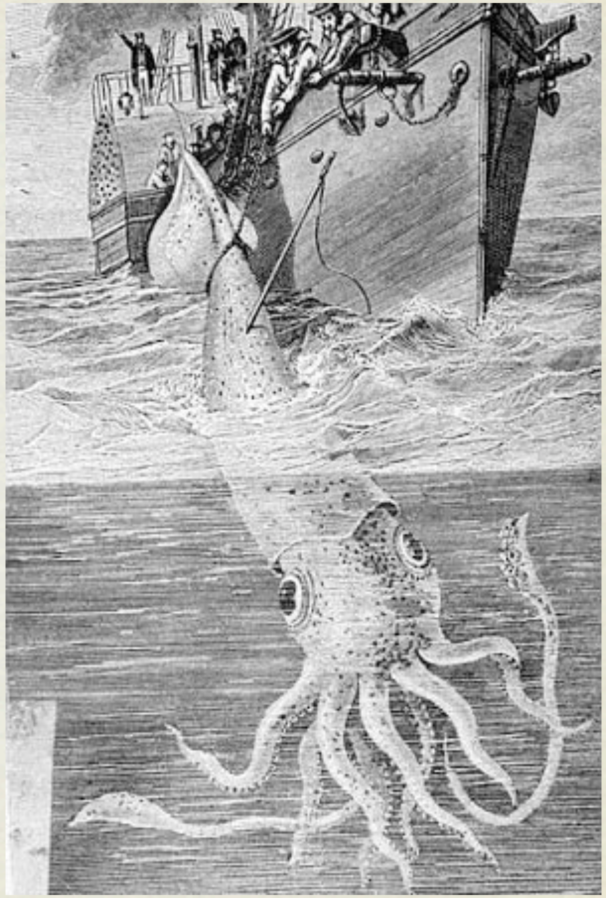


From Ou et al. (1992) Molecular epidemiology of HIV transmission in a dental practice. Science. 256:1165-71.

THE MYSTERY OF THE CHILEAN BLOB

THE MYSTERY OF THE CHILEAN BLOB

```
>Chilean_Blob
TAATACTAACTATATCCCTACTCTCCATTCTCATCGGGG
GTTGAGGAGGACTAAACCAGACTCAACTCCGAAAAATTA
TAGCTTACTCATCAATCGCCCACATAGGATGAATAACCA
CAATCCTACCCTACAATACAACCATAACCCTACTAAACC
TACTAATCTATGTCACAATAACCTTCACCATATTCATAC
TATTTATCCAAAACTCAACCACAACCACACTATCTCTGT
CCCAGACATGAAACAAAACACCCATTACCACAACCCTTA
CCATACTTACCCTACTTTCCATAGGGGGCCTCCCACCAC
TCTCGGGCTTTATCCCCAAATGAATAATTATTCAAGAAC
TAACAAAAAACGAAACCCTCATCATACCAACCTTCATAG
CCACCACAGCATTACTCAACCTCTACTTCTATATACGCC
TCACCTACTCAACAGCACTAACCCTATTCCCCTCCACAA
ATAACATAAAAATAAAATGACAATTCTACCCCACAAAAC
GAATAACCCTCCTGCCAACAGCAATTGTAATATCAACAA
TACTCCTACCCCTTACACCAATACTCTCCACCCTATTAT
AG
```



THE MYSTERY OF THE CHILEAN BLOB

Lineage Report

```
Cetacea       [whales & dolphins]
. Odontoceti     [whales & dolphins]
. . Physeteridae [whales & dolphins]
. . . Physeter catodon ------------------------- 1085  3 hits [whales & dolphins]  Physeter catodon NADH dehydrogenase subunit 2 (nad2) gene,
. . . Kogia breviceps ..........................  638  1 hit  [whales & dolphins]  Kogia breviceps complete mitochondrial genome
. . Orcaella brevirostris ----------------------  593  1 hit  [whales & dolphins]  Orcaella brevirostris isolate 97 mitochondrion, complete ge
. . Grampus griseus ............................  593  1 hit  [whales & dolphins]  Grampus griseus mitochondrion, complete genome
. . Feresa attenuata ...........................  592  2 hits [whales & dolphins]  Feresa attenuata isolate 36 mitochondrion, complete genome
. . Tursiops truncatus (bottle-nosed dolphin) ...  592  1 hit  [whales & dolphins]  Tursiops truncatus mitochondrion, complete genome
. . Globicephala melas .........................  586  3 hits [whales & dolphins]  Globicephala melas isolate GlomelG42 mitochondrion, partial
. . Peponocephala electra ......................  580  2 hits [whales & dolphins]  Peponocephala electra isolate M6 mitochondrion, complete ge
. . Globicephala macrorhynchus .................  580  4 hits [whales & dolphins]  Globicephala macrorhynchus isolate Glomac65 mitochondrion,
. . Pseudorca crassidens .......................  577  3 hits [whales & dolphins]  Pseudorca crassidens mitochondrion, complete genome
. . Orcinus orca (Orca) ........................  569 54 hits [whales & dolphins]  Orcinus orca isolate ENPTGA2 mitochondrion, complete genome
. . Sotalia fluviatilis ........................  569  2 hits [whales & dolphins]  Sotalia fluviatilis haplotype 10 NADH dehydrogenase subunit
. . Platanista minor ...........................  569  1 hit  [whales & dolphins]  Platanista minor complete mitochondrial genome
. . Steno bredanensis ..........................  566  2 hits [whales & dolphins]  Steno bredanensis isolate StebreS9 mitochondrion, partial g
. Megaptera novaeangliae -----------------------  636  5 hits [whales & dolphins]  Megaptera novaeangliae voucher GOM9049 NADH dehydrogenase s
. Balaenoptera bonaerensis .....................  630  1 hit  [whales & dolphins]  Balaenoptera bonaerensis mitochondrial DNA, complete genome
. Eubalaena japonica ...........................  619  1 hit  [whales & dolphins]  Eubalaena japonica mitochondrial DNA, complete genome
. Balaenoptera brydei ..........................  614  2 hits [whales & dolphins]  Balaenoptera brydei mitochondrial DNA, complete genome, iso
. Balaena mysticetus (Greenland right whale) ....  614  2 hits [whales & dolphins]  Balaena mysticetus mitochondrial DNA, complete genome
. Balaenoptera musculus ........................
. Balaenoptera edeni ...........................
. Balaenoptera omurai ..........................
. Eschrichtius robustus (California gray whale) .
. Balaenoptera borealis ........................
. Caperea marginata ............................
. Balaenoptera physalus (finback whale) .........
```

THE MYSTERY OF THE CHILEAN BLOB

```
>emb|AJ277029.2| D Physeter macrocephalus mitochondrial genome
Length=16428

 Score = 1074 bits (581),  Expect = 0.0
 Identities = 585/587 (99%), Gaps = 0/587 (0%)
 Strand=Plus/Plus

Query  1     TAATACTAACTATATCCCTACTCTCCATTCTCATCGGGGGTTGAGGAGGACTAAACCAGA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4400  TAATACTAACTATATCCCTACTCTCCATTCTCATCGGGGGTTGAGGAGGACTAAACCAGA  4459

Query  61    CTCAACTCCGAAAAATTATAGCTTACTCATCAATCGCCCACATAGGATGAATAACCACAA  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4460  CTCAACTCCGAAAAATTATAGCTTACTCATCAATCGCCCACATAGGATGAATAACCACAA  4519

Query  121   TCCTACCCTACAATACAACCATAACCCTACTAAACCTACTAATCTATGTCACAATAACCT  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4520  TCCTACCCTACAATACAACCATAACCCTACTAAACCTACTAATCTATGTCACAATAACCT  4579

Query  181   TCACCATATTCATACTATTTATCCAAAACTCAACCACAACCACACTATCTCTGTCCCAGA  240
             ||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4580  TCACCATATTCACACTATTTATCCAAAACTCAACCACAACCACACTATCTCTGTCCCAGA  4639

Query  241   CATGAAACAAAACACCCATTACCACAACCCTTACCATACTTACCCTACTTTCCATAGGGG  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4640  CATGAAACAAAACACCCATTACCACAACCCTTACCATACTTACCCTACTTTCCATAGGGG  4699

Query  301   GCCTCCCACCACTCTCGGGCTTTATCCCCAAATGAATAATTATTCAAGAACTAACAAAAA  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4700  GCCTCCCACCACTCTCGGGCTTTATCCCCAAATGAATAATTATTCAAGAACTAACAAAAA  4759

Query  361   ACGAAACCCTCATCATACCAACCTTCATAGCCACCACAGCATTACTCAACCTCTACTTCT  420
             ||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4760  ACGAAGCCCTCATCATACCAACCTTCATAGCCACCACAGCATTACTCAACCTCTACTTCT  4819

Query  421   ATATACGCCTCACCTACTCAACAGCACTAACCCTATTCCCCTCCACAAATAACATAAAAA  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4820  ATATACGCCTCACCTACTCAACAGCACTAACCCTATTCCCCTCCACAAATAACATAAAAA  4879

Query  481   TAAAATGACAATTCTACCCCACAAAACGAATAACCCTCCTGCCAACAGCAATTGTAATAT  540
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4880  TAAAATGACAATTCTACCCCACAAAACGAATAACCCTCCTGCCAACAGCAATTGTAATAT  4939

Query  541   CAACAATACTCCTACCCCTTACACCAATACTCTCCACCCTATTATAG  587
             |||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4940  CAACAATACTCCTACCCCTTACACCAATACTCTCCACCCTATTATAG  4986
```

BIOINFORMATICS CREDO

- Remember about biology
- Do not trust the data
- Use comparative approach
- Use statistics
- Know the limits
- Remember about biology!!!