Introduction to Phylogenetics
Week 2

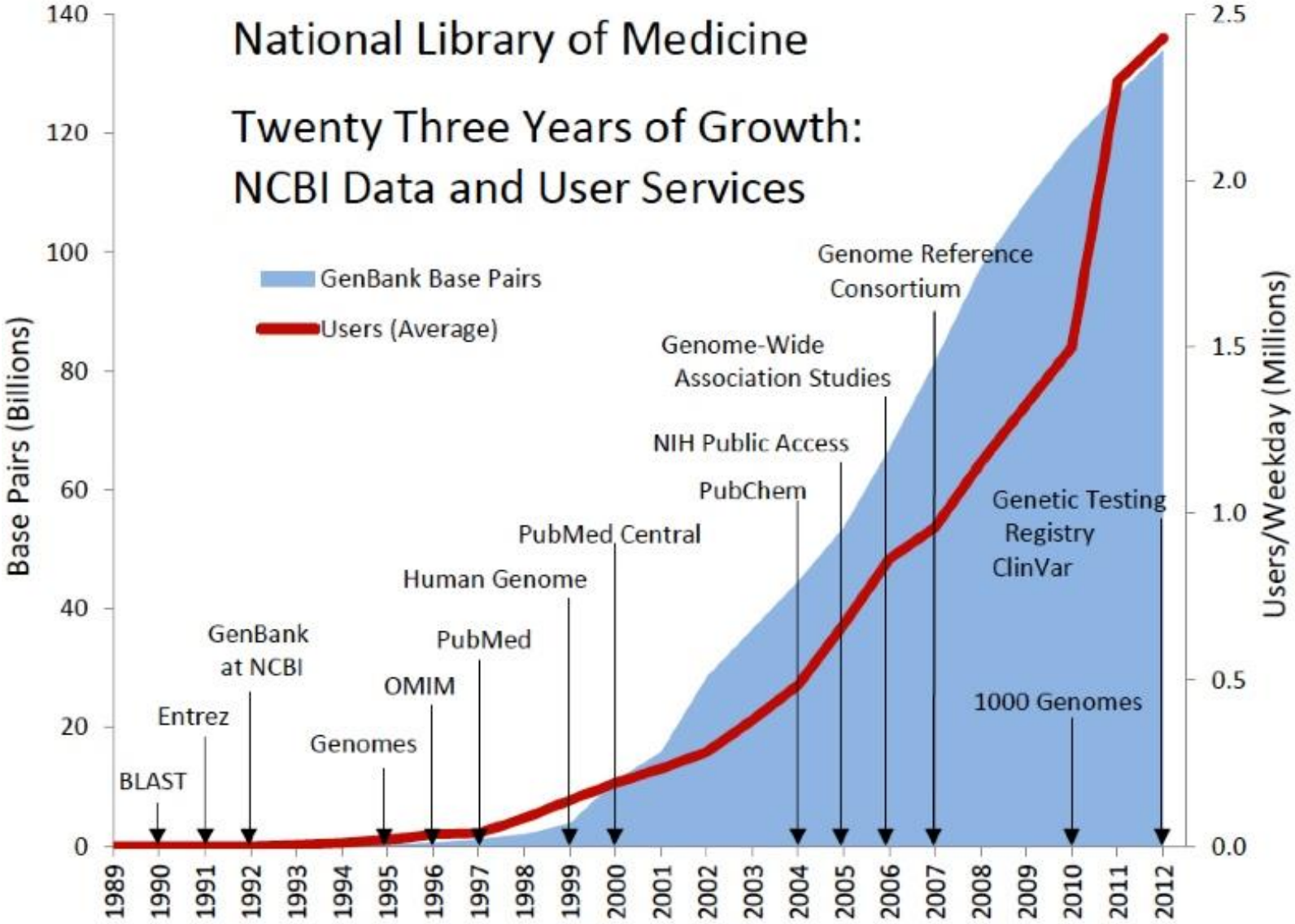
Databases and Sequence Formats

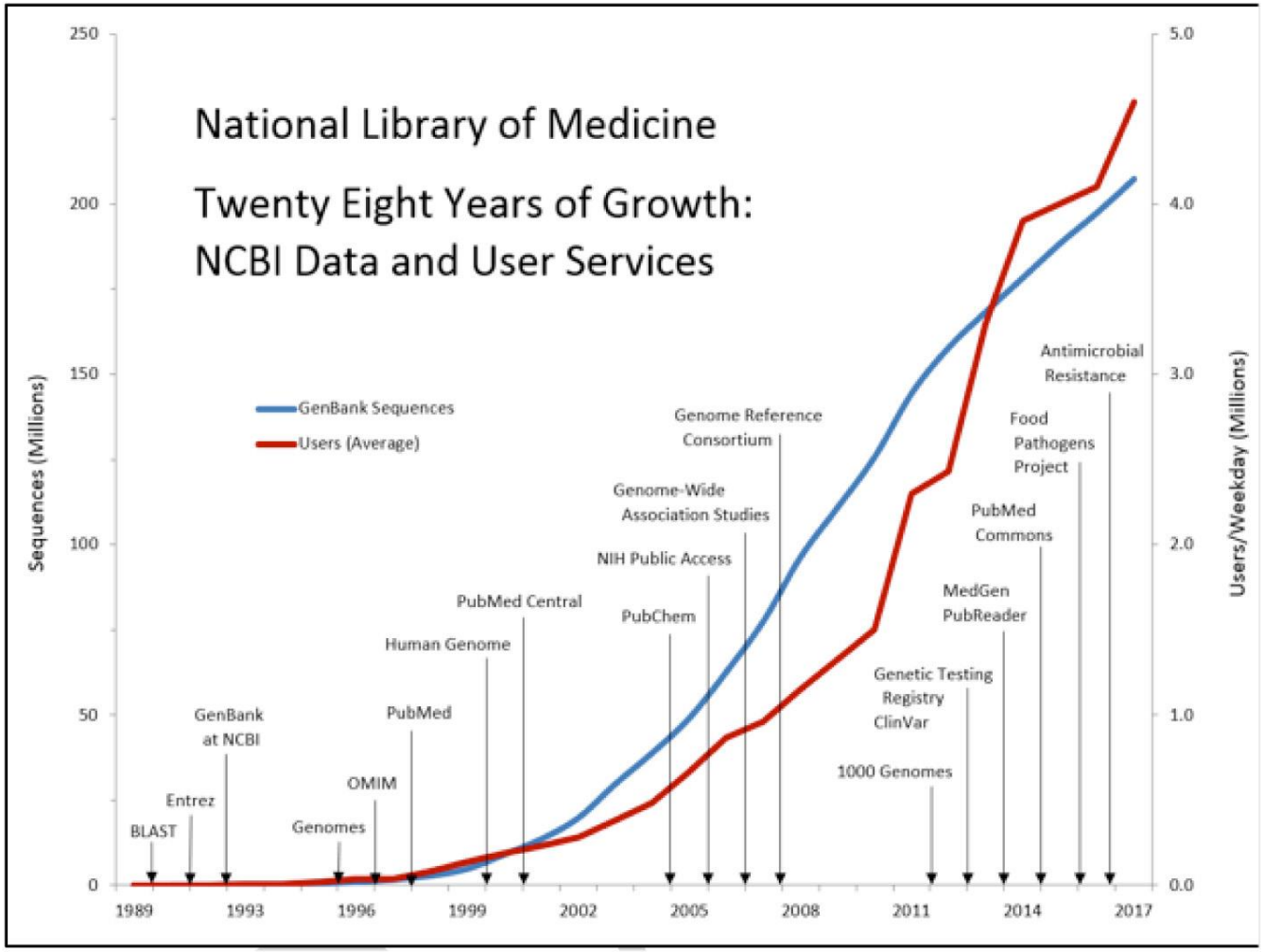
I. Databases

- Crucial to bioinformatics
- The bigger the database, the more comparative research data
- Requires scientists to upload data
- Requires high-quality DNA sequence data
- Requires recognition of what's good/bad sequence data

I. Databases

- Major databases
  - National Center for Biotechnology Information (NCBI - Genbank)
  - European Molecular Biology Laboratory (EMBL)


National Library of Medicine
Twenty Three Years of Growth:
NCBI Data and User Services
GenBank Base Pairs
Users (Average)
Base Pairs (Billions)
Users/Weekday (Millions)
140
120
100
80
60
40
20
0
2.5
2.0
1.5
1.0
0.5
0.0
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
BLAST
Entrez
GenBank at NCBI
Genomes
OMIM
PubMed
Human Genome
PubMed Central
PubChem
NIH Public Access
Genome-Wide Association Studies
Genome Reference Consortium
Genetic Testing Registry ClinVar
1000 Genomes

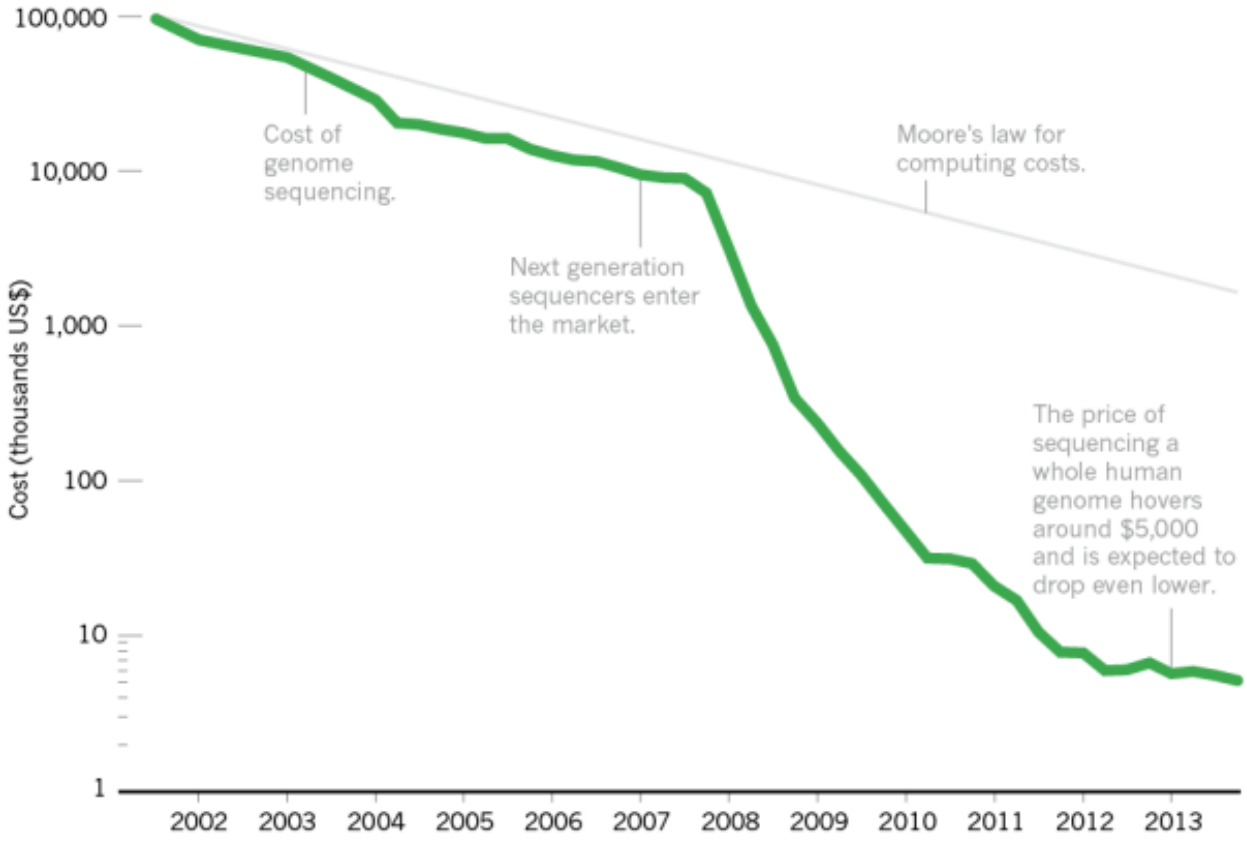

National Library of Medicine
Twenty Eight Years of Growth:
NCBI Data and User Services
Sequences (Millions)
Users/Weekday (Millions)
250
200
150
100
50
0
5.0
4.0
3.0
2.0
1.0
0.0
1989
1993
1996
1999
2002
2005
2008
2011
2014
2017
GenBank Sequences
Users (Average)
BLAST
Entrez
GenBank at NCBI
Genomes
OMIM
PubMed
Human Genome
PubMed Central
PubChem
NIH Public Access
Genome-Wide Association Studies
Genome Reference Consortium
1000 Genomes
Genetic Testing Registry ClinVar
MedGen PubReader
PubMed Commons
Food Pathogens Project
Antimicrobial Resistance

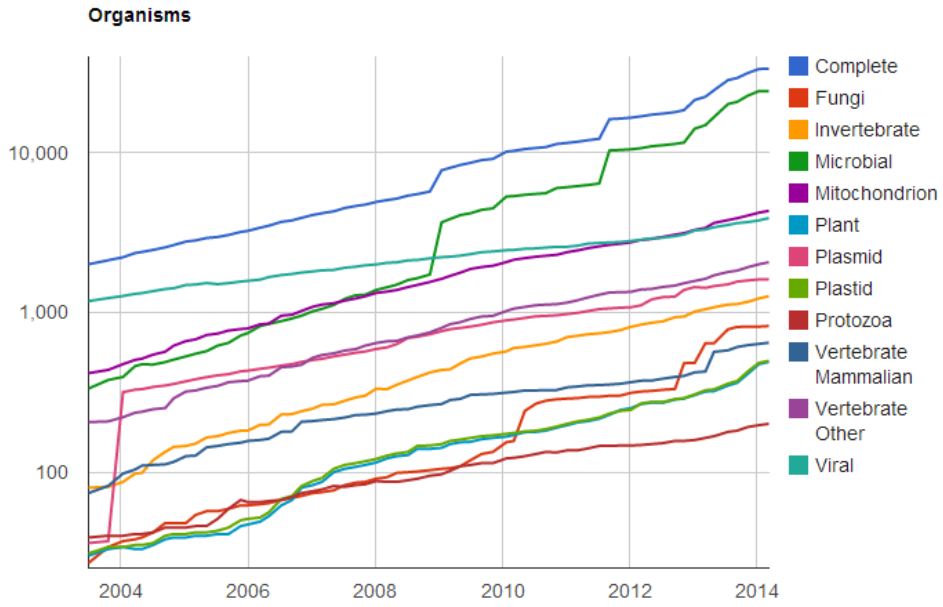
I. Databases

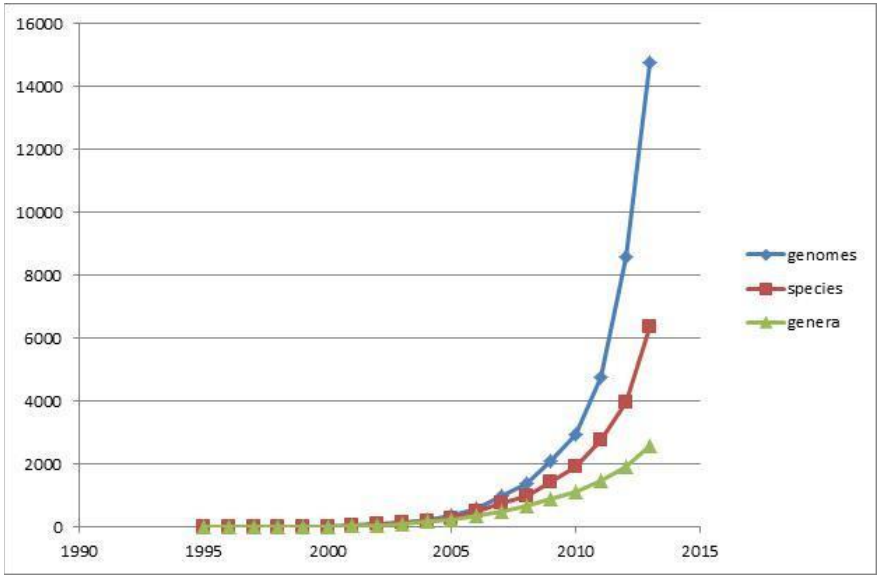
- Major databases
  - National Center for Biotechnology Information (NCBI - Genbank)
  - European Molecular Biology Laboratory (EMBL)
- Curated databases
  - Ribosomal Database Project (RDP – Michigan U)
  - Greengenes (Lawrence Berkeley)
  - SILVA (German Network for Bioinformatics)
  - SwissPlot (University of Geneva)
- Next Generation Sequencing (NGS)
  - Genomic databases (RefSeq, TIGR)

Falling fast

In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore's law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.




Organisms
10,000
1,000
100
2004
2006
2008
2010
2012
2014
Complete
Fungi
Invertebrate
Microbial
Mitochondrion
Plant
Plasmid
Plastid
Protozoa
Vertebrate Mammalian
Vertebrate Other
Viral


16000
14000
12000
10000
8000
6000
4000
2000
0
1990
1995
2000
2005
2010
2015
genomes
species
genera

I. Searching databases

- Knowing what question you wish to ask
- Knowing what gene(s) sequences will help you answer that question
- Knowing where to find sequences
  - Gene identifiers
  - Database searching
  - Other resources (Pubmed)
- Recognizing/trusting data quality

I. Searching databases (Unique Identifier)

- Requires some kind of information in advance (often can be retrieved from journal articles)
  - Read paper by Lee and Kasai
  - Find the accession number for *Escherichia coli* K12
  - Go to NCBI (http://www.ncbi.nlm.nih.gov/)
  - Go to the Nucleotide database
  - Enter the accession number
  - Find the 16S rRNA gene sequence
  - Write the first 10 bp

I. Searching databases (Unique Identifier)

- Accession number: J01695
- 16S rRNA began at base 1268
  - aattgaagag

I. Searching databases (Keyword)

- Can search by keyword
- Issues include:
  - No consistency with organism name (HIV-1 versus HIV1)
  - No consistency with gene names (protease vs. trypsin protease vs. trypsin)
  - Typos
- Requires some amount of trial and error

I. Searching databases (Keyword)

- Keyword search:
  - Go to NCBI (http://www.ncbi.nlm.nih.gov/)
  - Type ‘E. coli 16S’

NCBI Databases

Results found in 29 databases for: **E. coli 16S**

Literature

Bookshelf	160
MeSH	6
NLM Catalog	2
PubMed	3,938
PubMed Central	40,050

Genes

Gene	393
GEO DataSets	164
GEO Profiles	910
HomoloGene	0
PopSet	509
UniGene	0

Genetics

ClinVar	1
dbGaP	0
dbSNP	0
dbVar	0
GTR	3
MedGen	1
OMIM	5

Proteins

Conserved Domains	17
Identical Protein Groups	6,698
Protein	22,186,990
Protein Clusters	17
Sparcle	45
Structure	292

Genomes

Assembly	15,236
BioCollections	2
BioProject	54
BioSample	346
Genome	0
Nucleotide	3,792,549
Probe	19
SRA	1,192
Taxonomy	0

Chemicals

BioSystems	263
PubChem BioAssay	120
PubChem Compound	0
PubChem Substance	45

I. Searching databases (Keyword)

- Keyword search:
  - Go to NCBI (http://www.ncbi.nlm.nih.gov/)
  - Type ‘E. coli 16S’
  - Click on ‘nucleotide’
  - Find the E. coli 16S sequence
  - Type ‘E.coli 16S’
  - Type ‘Escherichia coli 16S’
  - Try to find the *Escherichia coli* K12 16S rRNA gene sequence

I. Searching databases (Keyword)

- Keyword search:
  - Repeat with:
    - Type ‘E.coli 16S’
    - Type ‘Escherichia coli 16S’
  - Try to find the *Escherichia coli* K12 16S rRNA gene sequence

I. Searching databases (Keyword)

- Keyword search:
  - Go to Taxonomy Browser
  - Type ‘Escherichia coli’
  - Find *Escherichia coli* K12
  - Click on Nucleotide link
  - Try to find the 16S rRNA gene sequence

I. Searching databases (Keyword)

- Keyword search:
  - Go to myRDP (https://rdp.cme.msu.edu/index.jsp)
  - Click ‘Hierarchy Browser’
  - We only want ‘isolates’
  - E. coli is in the Proteobacteria/Gammaproteobacteria/Escherichia/
  - Size should be >1,200
  - Quality should be good

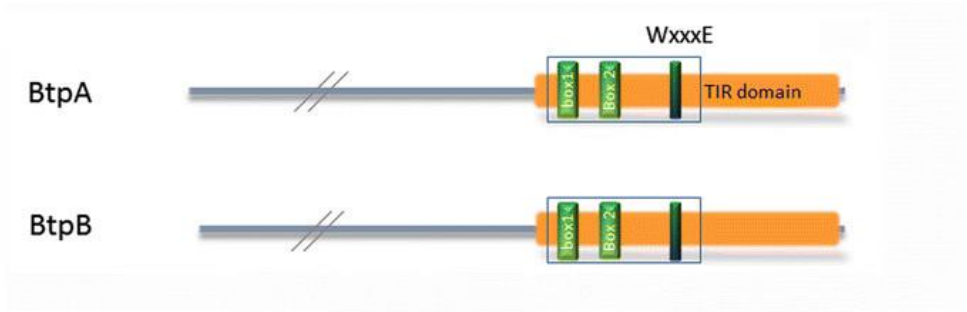
I. Searching databases (Keyword)

- Keyword search:
  - In myRDP, download selected sequence as an unaligned file in FASTA format (as a ‘testdrive’)
  - Remove all gaps
  - Open in text editor

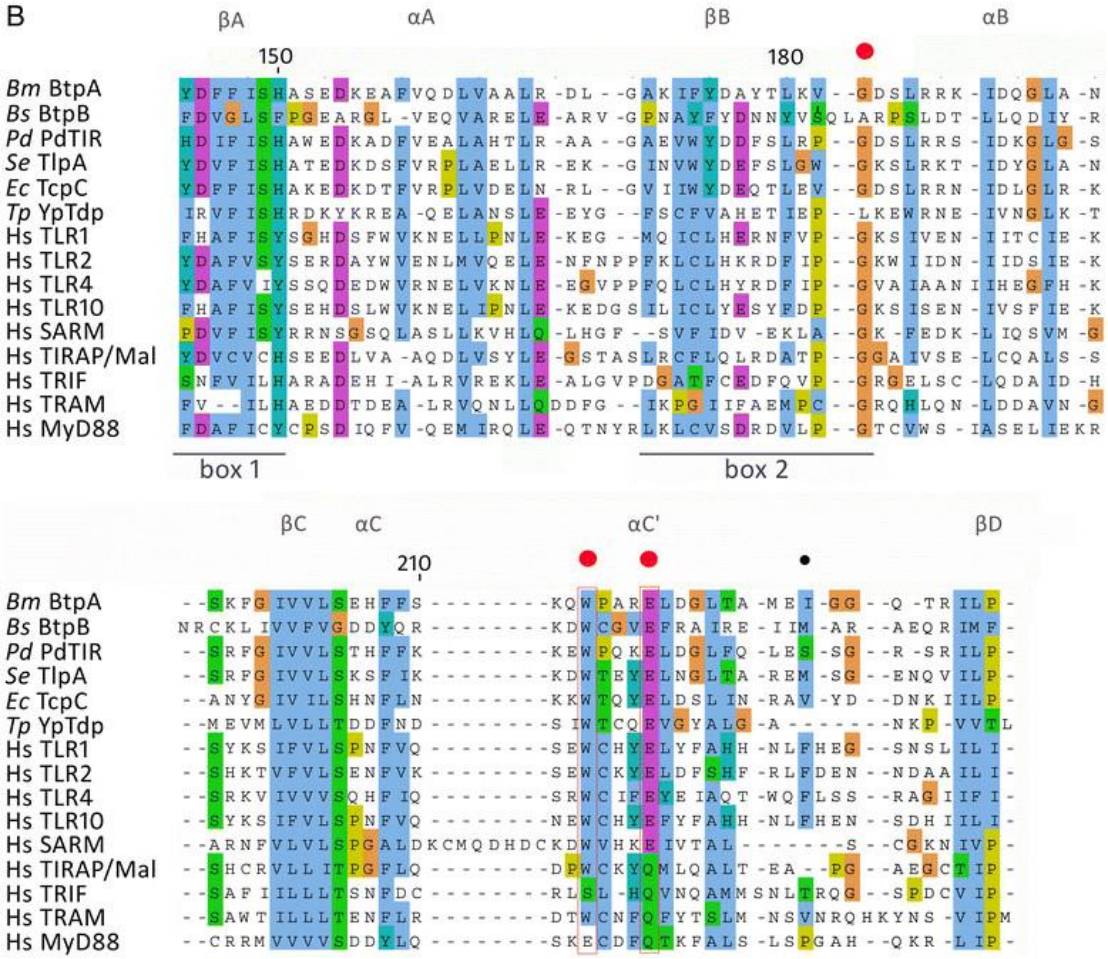
I. Searching databases (homology)

- Alignment uses the homology between sequences to infer relationship
- Deletions/insertions/non-homology can be used to generate similarity score
  - Nucleic acids – mismatch score
  - Proteins – uses a ‘log odds matrix’
- Heuristic search (e.g. BLAST) – allows rapid searching, no guarantee of finding highest alignment scores

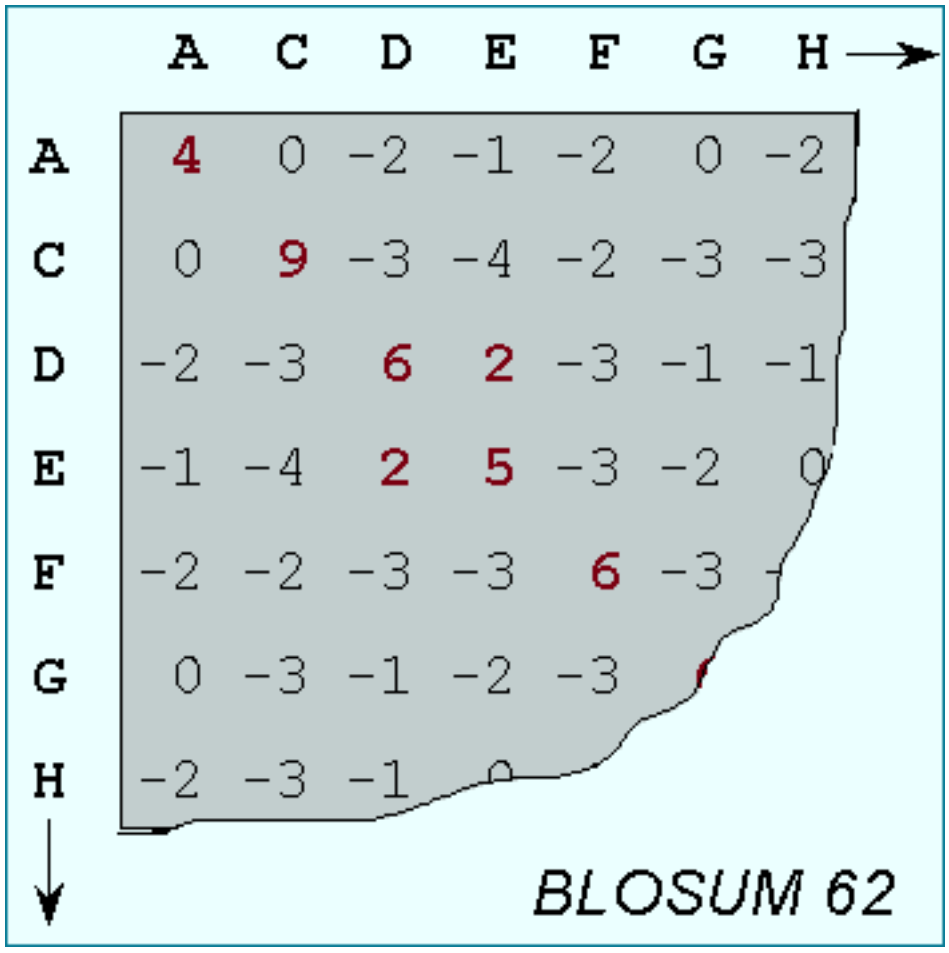
A


WxxxE
BtpA
box1
Box 2
TIR domain
BtpB
box1
Box 2

B


βA
αA
βB
αB
150
180
Bm BtpA
Bs BtpB
Pd PdTIR
Se TlpA
Ec TcpC
Tp YpTdp
Hs TLR1
Hs TLR2
Hs TLR4
Hs TLR10
Hs SARM
Hs TIRAP/Mal
Hs TRIF
Hs TRAM
Hs MyD88
box 1
box 2
βC
αC
αC'
βD
210
Bm BtpA
Bs BtpB
Pd PdTIR
Se TlpA
Ec TcpC
Tp YpTdp
Hs TLR1
Hs TLR2
Hs TLR4
Hs TLR10
Hs SARM
Hs TIRAP/Mal
Hs TRIF
Hs TRAM
Hs MyD88

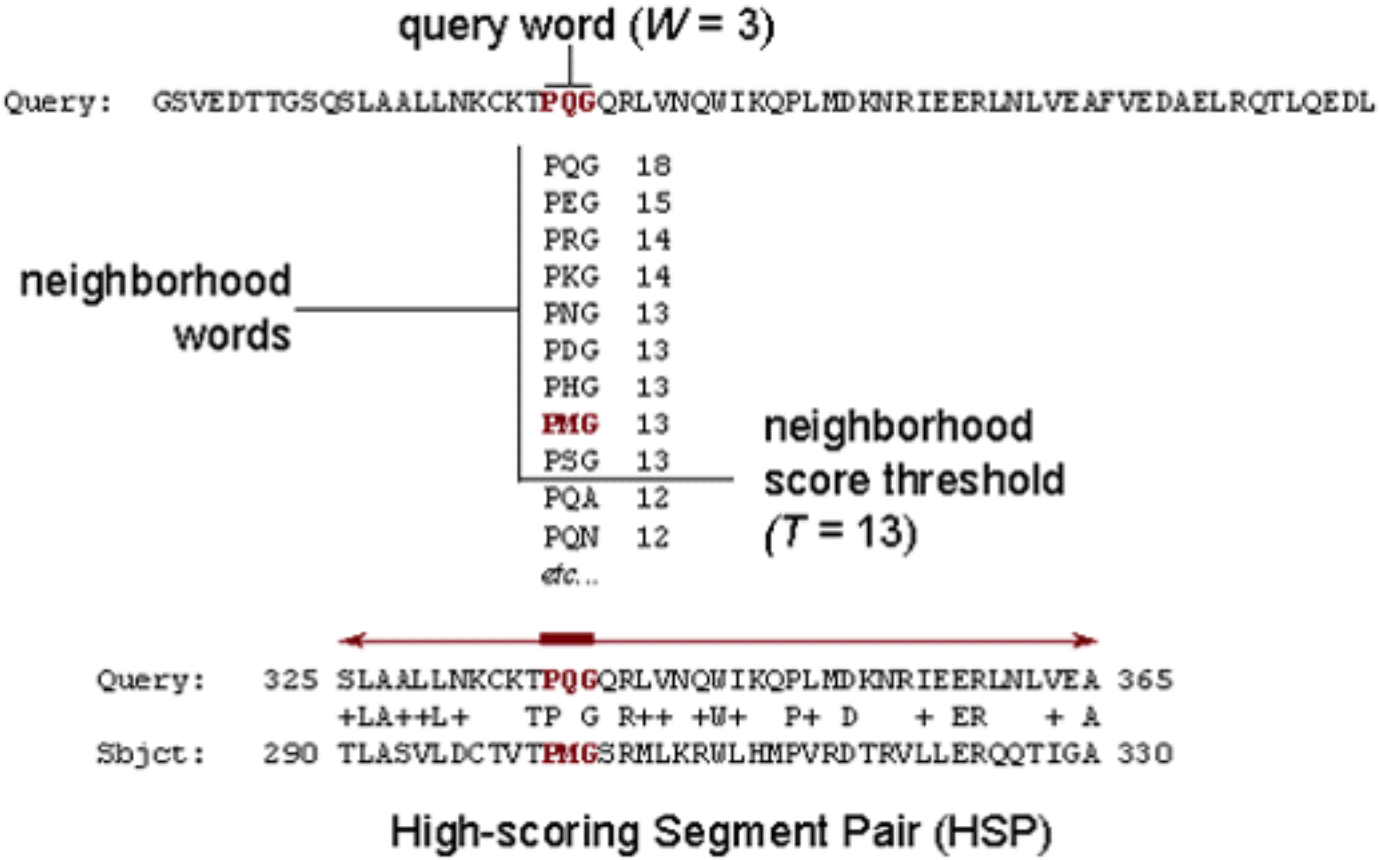
$$M_{i,j} = \log_2 \frac{q_{i,j}}{p_i \cdot p_j} = \log_2 \frac{observed\ frequency}{expected\ frequency}$$



I. Searching databases (BLAST)

- Local alignment
- Speed over sensitivity
- Searches for similar sequences
  - Nucleotides – identical neighbor bp
  - Proteins – identical/similar neighbor aa
  - Growing alignment scored with match/mismatch scoring matrix

The BLAST Search Algorithm



High-scoring Segment Pair (HSP)

Program	Description
BLASTP	Compares an amino acid query sequence against a protein sequence or a database.
BLASTN	Compares a nucleotide query sequence against a nucleotide sequence or a database.
BLASTX	Compares a nucleotide query sequence translated in all reading frames against a protein sequence or a database. This option is useful to find potential translation products of an uncharacterized nucleotide sequence.
TBLASTN	Compares a protein query sequence against a nucleotide sequence or a database dynamically translated in all open reading frames.
TBLASTX	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence or a database. The TBLASTX program is the most computationally intensive. It is useful when trying to find distant homologies between coding DNA sequences.

I. Searching databases (Similarity)

- **P-value:** How likely is that BLAST is producing an alignment between two sequences that are not functionally related.
- A threshold of 0.05 means you are 95% sure that the result is significant.
- **E-value.** The E-value is computed by multiplying the p-value times the size of the database, giving the expected number of times that the given score would appear in a random database of a given size.
- A p-value of 0.001 and a database of 1,000,000 sequences, the corresponding E-value is 0.001 x 1,000,000 = 1,000.
- The expected **number** of distinct alignments that we would obtain with a score greater or equal to a given value, by chance, in a database search. The higher the E-value, the less significant the match.
- An e-value of 2 means that we expect 2 alignments to occur just by chance for the given score. Similarly, an e-value of 0.01 means that we expect one random match in every hundred for the given score.

II. File Formats

```
LOCUS       HUMBETGLA               468 bp    DNA     linear   UNC 22-JAN-2015 ORIGIN
  1  atggtncayy tnacnccngt ggagaagtcy gcygtnacng cnctntgggg yaaggtnaay
 61  gtggatgaag yyggyggyga ggccctgggc agnctgctng tggtctaccc ttggacccag
121  aggttcttng antcnttygg ggatctgnnn acnccngang cagttatggg caaccctaag
181  gtgaaggctc atggcaagaa agtgctcggt gcctttagtg atggcctggc tcacctggac
241  aacctcaagg gcacctttgc cacactgagt gagctgcact gtgacaagct ncaygtggat
301  cctgagaact tcaggctnct nggcaacgtg ytngtctgyg tgctggccca tcactttggc
361  aaagaattca ccccaccagt gcangcngcc tatcagaaag tggtngctgg tgtngctaat
421  gccctggccc acaagtatca ctaagctngc yttyttgytg tccaattt
//
```

II. File Formats: FASTA

- Most common format for bioinformatics

  >name

  sequence

- Common mistakes:
  - Non-unique identifier in name
  - > in the wrong place
  - Extra spaces, weird returns

II. File Formats: FASTA

>J01859.1 Escherichia coli 16S ribosomal RNA, complete sequence
AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAG
TCGAACGGT
AACAGGAAGAAGCTTGCTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGGGAAA
CTGCCTGATG
GAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAAGAGG
GGGACCTTCG
GGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAACGGCTCAC
CTAGGCGACG
ATCCCTAGCTGGTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTC
CTACGGGAGG

II. File Formats: NEXUS

- Used in PAUP* and PHYLIP
  - Really difficult to generate by hand (use generator)
- Common mistakes
  - Tabs in the wrong place
  - Matrix doesn’t match entered data
  - Missing/non-standard symbols not defined
  - Names too long
- See Chapter 8, page 289

II. File Formats: NEXUS

```
#nexus

begin DATA;
dimensions ntax=5 nchar=51;
format datatype=dna missing=? gap=-;
begin data;
matrix
Ephedra     TTAAGCCATGCATGTCTAAGTATGAACTAATTCCAAACGGTGAAACTGCG
Gnetum      TTAAGCCATGCATGTCTATGTACGAACTAATCAGAACGGTGAAACTGCGG
Welwit      TTAAGCCATGCACGTGTAAGTATGAACTAGTC-GAAACGGTGAAACTGCG
Ginkgo      TTAAGCCATGCATGTGTAAGTATGAACTCTTTACAGACTGTGAAACTGCG
PinusTTAAGCCATGCATGTCTAAGTATGAACTAATTGCAGACTGTGAAACTGCG
;
end;
```

II. File Formats: PHYLIP

- Because why not create a format just for your own program
- Common mistakes
  - Matrix doesn’t match entered data
  - Missing/non-standard symbols not defined
  - Names too long

II. File Formats: PHYLIP

```
5 51
Ephedra   TTAAGCCATGCATGTCTAAGTATGAACTAATTCCAAACGGTGAAACTGCG
Gnetum
      TTAAGCCATGCATGTCTATGTACGAACTAATCAGAACGGTGAAACTGCGG
Welwit    TTAAGCCATGCACGTGTAAGTATGAACTAGTC-
GAAACGGTGAAACTGCG
Ginkgo    TTAAGCCATGCATGTGTAAGTATGAACTCTTTACAGACTGTGAAACTGCG
Pinus     TTAAGCCATGCATGTCTAAGTATGAACTAATTGCAGACTGTGAAACTGCG
```

II. File Formats: CLUSTAL

- Because why not create a format just for your own program
- Common mistakes
  - Includes spaces
  - Blocks of text exceed 60 characters
  - Names too long
- Provides information on alignment

*	residues or nucleotides in column are identical
:	conserved substitutions observed
.	semi-conserved substitutions observed
	no match

II. File Formats: CLUSTAL

```
CLUSTAL W (1.82) multiple sequence alignment

FOSB_MOUSE          MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
FOSB_HUMAN          MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
                    ************************************************************

FOSB_MOUSE          ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 120
FOSB_HUMAN          ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 120
                    ********************************.***************:*.**:****** 

FOSB_MOUSE          GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT 180
FOSB_HUMAN          GGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT 180
                    ****** ***** .********************************************** 

FOSB_MOUSE          DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD 240
FOSB_HUMAN          DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD 240
                    ************************************************************

FOSB_MOUSE          LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY 300
FOSB_HUMAN          LPGSAPAKEDGFSWLLPPPPPPPLPFQTSQDAPPNLTASLFTHSEVQVLGDPFPVVNPSY 300
                    ****:.******.**************:*:**************************.*** 

FOSB_MOUSE          TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL 338
FOSB_HUMAN          TSSFVLTCPEVSAFAGAQRTSGSDQPSDPLNSPSLLAL 338
                    **********************:*************** 
```