



Sequencing alignment

Ameer Effat M. Elfarash

Dept. of Genetics
Fac. of Agriculture, Assiut Univ.
amir_effat@yahoo.com

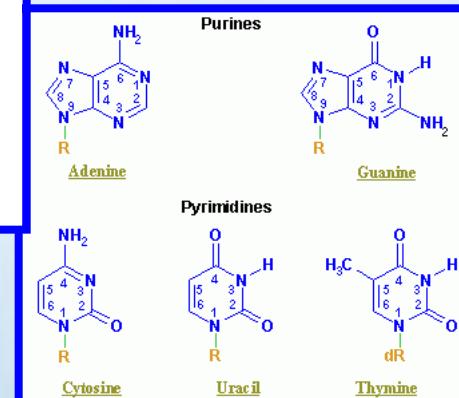
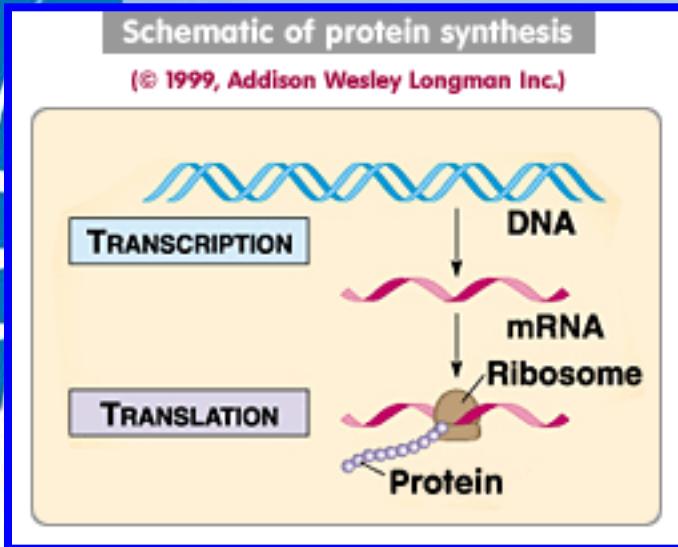


Why perform a multiple sequence alignment?

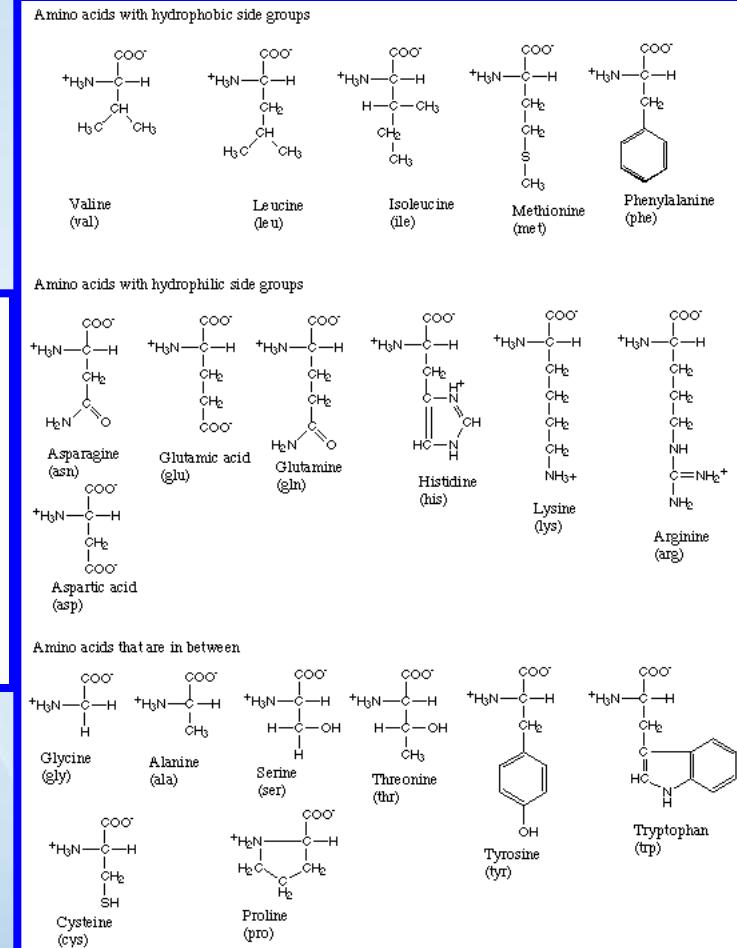
MSAs are at the heart of comparative genomics studies which seek to study evolutionary histories, functional and structural aspects of sequences, and to understand phenotypic differences between species



The Building Blocks...



ATGC



VLMFNQEDHKRCSTPYW



```
MVNLTSDEKTAVLALWNKVDVEDCGGEALGRLLVVYPWTQRFFE...
|| || || || || || || || || || || || || || || || || || |
MVHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFE...
```

Alignment: Comparing two (pairwise) or more (multiple) sequences.
Searching for a series of identical or similar characters in the sequences.

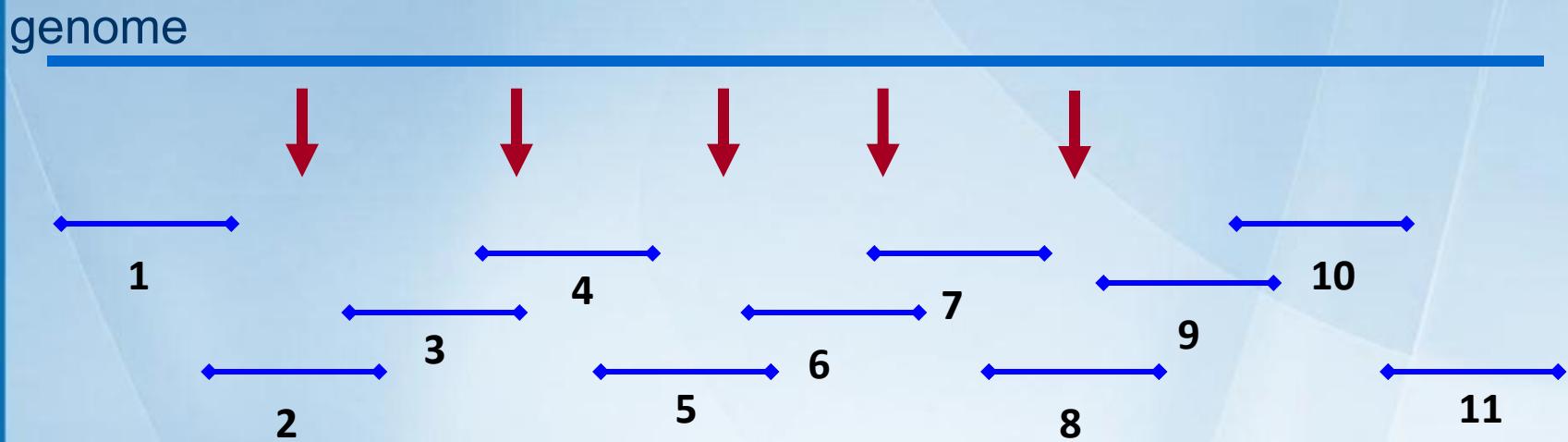


Why Align Sequences?

- Discover functional, structural, and evolutionary information
- Similar Sequences may have similar function
 - Gene Regulation
 - Biochemical Function
 - Similar Structure
- Identify primers and probes to search for homologous sequences in other organisms
- Crucial for genome sequencing



Genome Sequencing Strategy

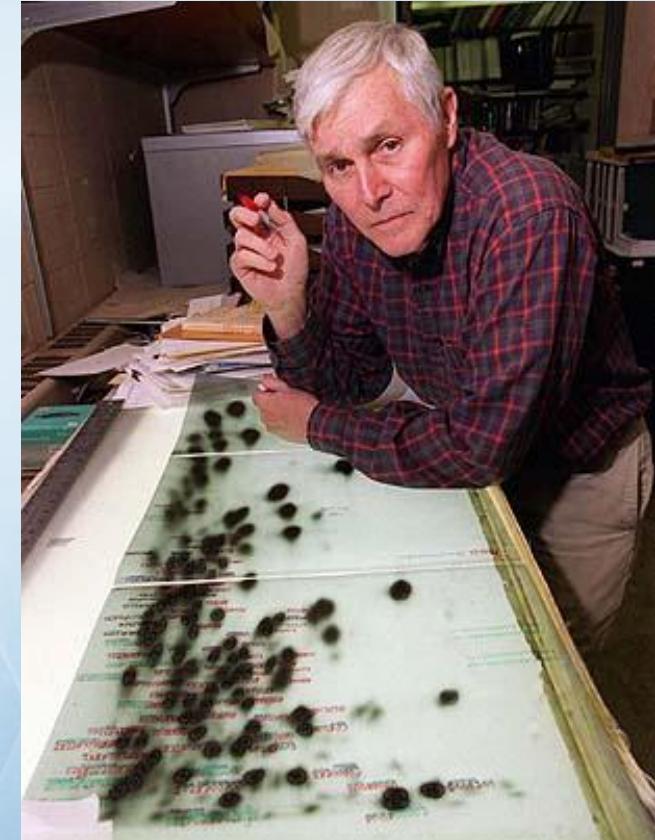


1. Obtain a large collection of BAC clones
2. Map them onto the genome (Physical Mapping)
3. Select a minimum tiling path
4. Sequence each clone in the path with shotgun
5. Assemble
6. Put everything together

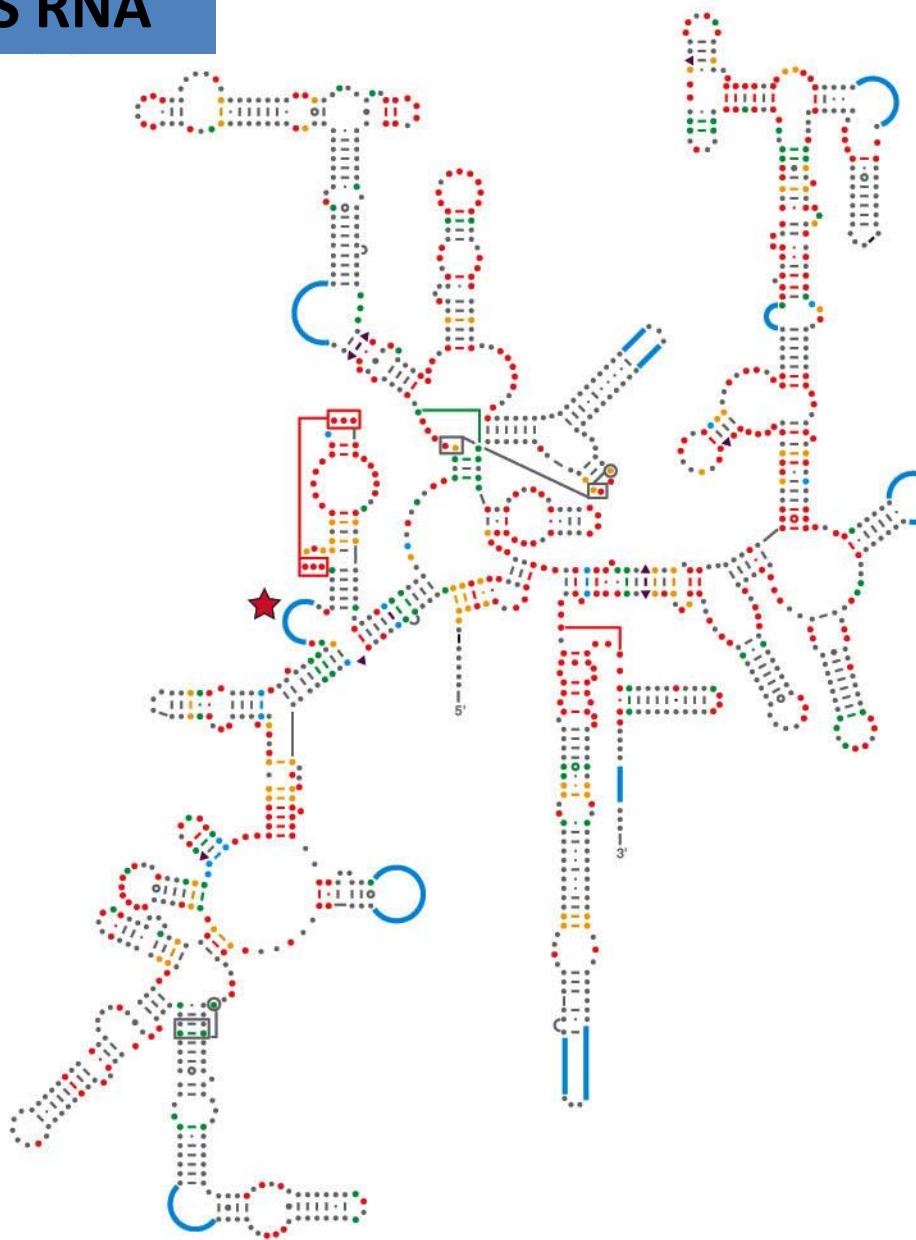
Why Align Sequences?



- Homology
 - Similar sequences may have a common ancestor
- In 1990 Carl Woese and colleagues proposed the Tree of Life, using a molecular phylogenetic approach.
- It is based on sequencing rRNA (16S and 18S), which all organisms share.
- All organisms were separated into three domains: *Bacteria*, *Archaea*, *Eukarya*.



16S RNA

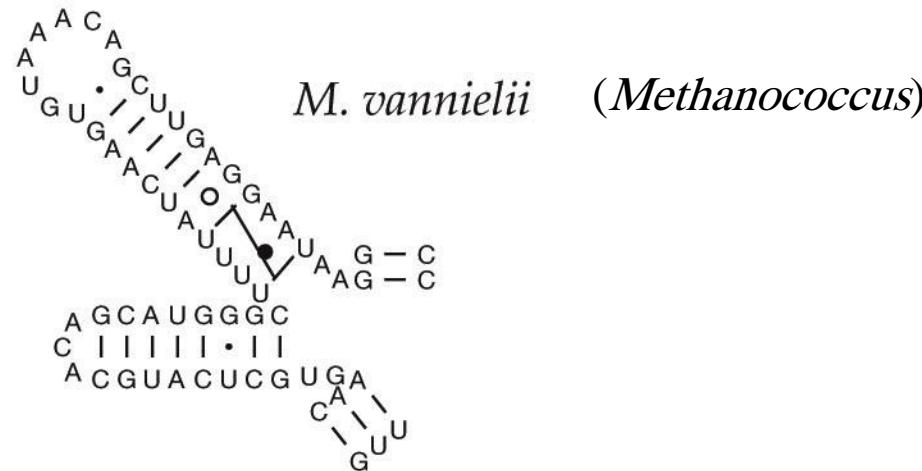
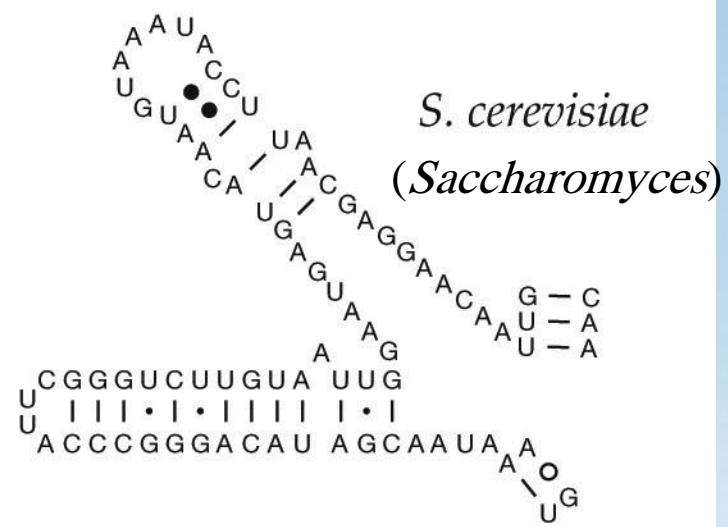
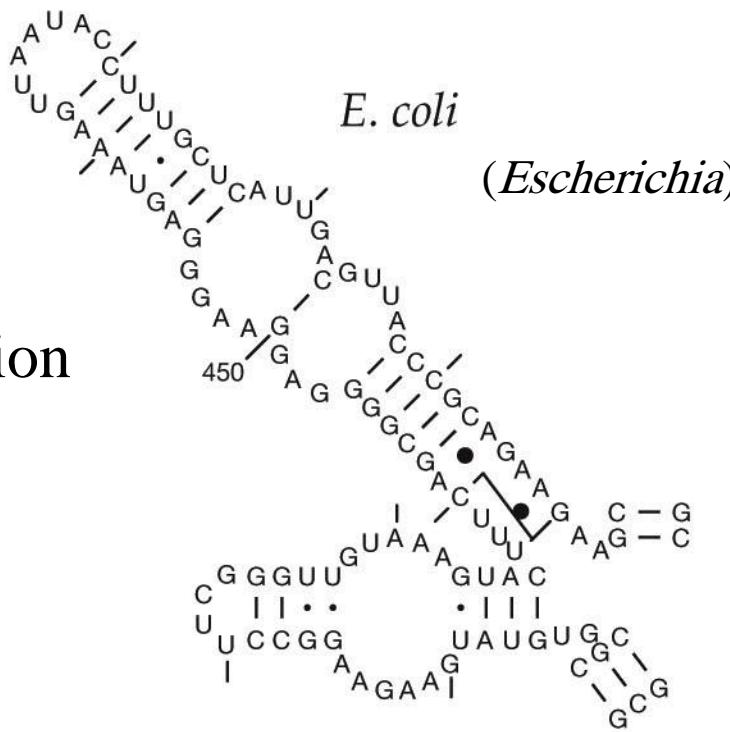


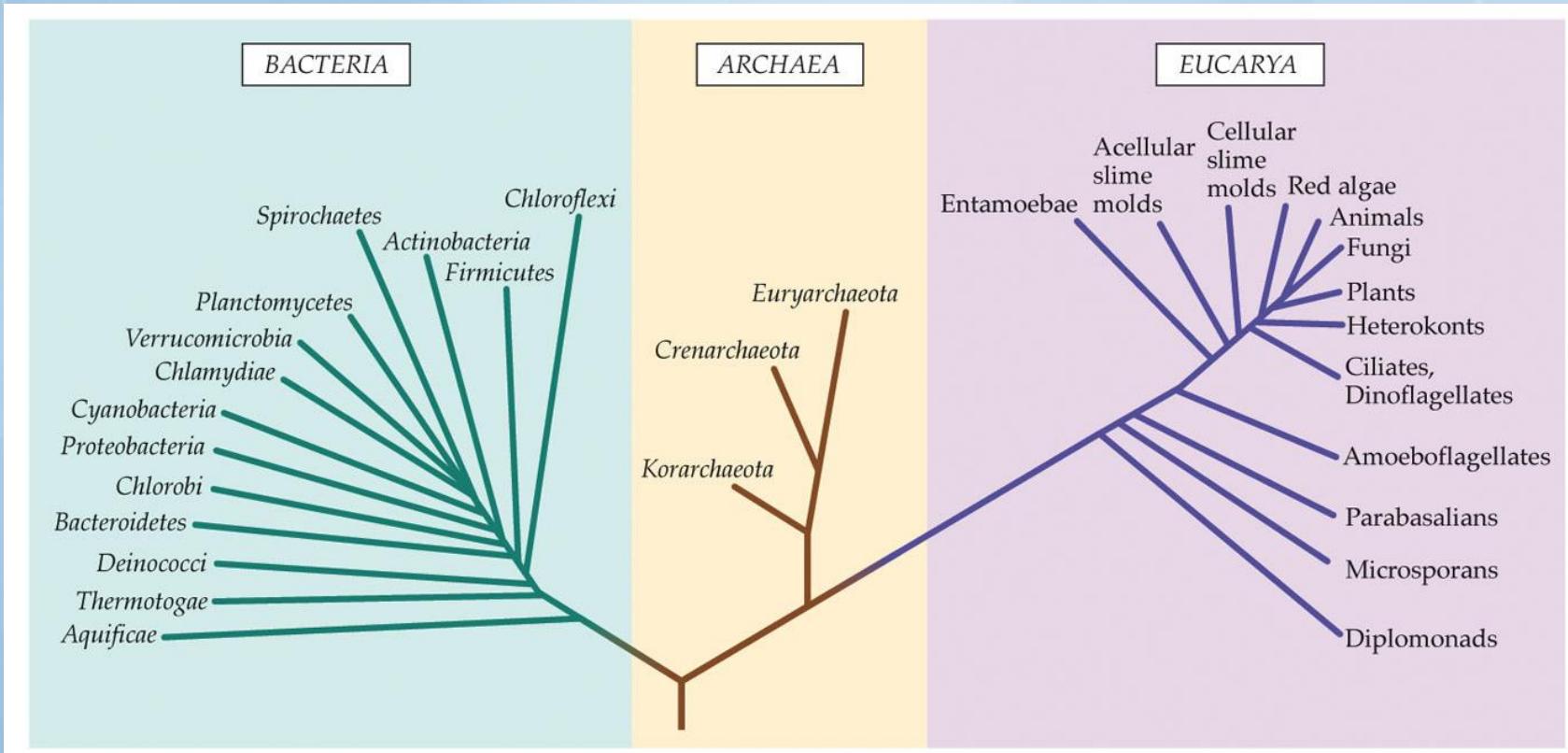
- Identical in 98% or more of all organisms
- Conserved only in the *Bacteria*
- Conserved only in the *Archaea*
- Conserved only in the *Eukarya*
- ▲ Conserved within each domain, variable among domains
- Regions that vary structurally among domains

(B)



Region







Evolutionary changes in sequences

Three types of nucleotide changes:

1. **Substitution** – a replacement of one (or more) sequence characters by another: **AAGA** → **AACA**
2. **Insertion** - an insertion of one (or more) sequence characters: **AAG A**
3. **Deletion** – a deletion of one (or more) sequence characters: **AA G A**

Insertion + Deletion → Indel



How DNA Sequence Data is Compared

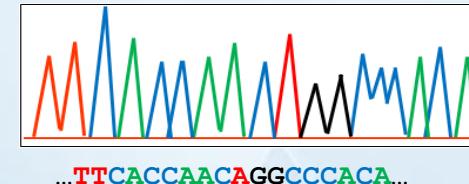
Obtain Samples:
Blood , Saliva, Hair
Follicles, Feathers, Scales



Genetic Data

Extract DNA from Cells

Sequence DNA



**Compare
DNA
Sequences to
One Another**

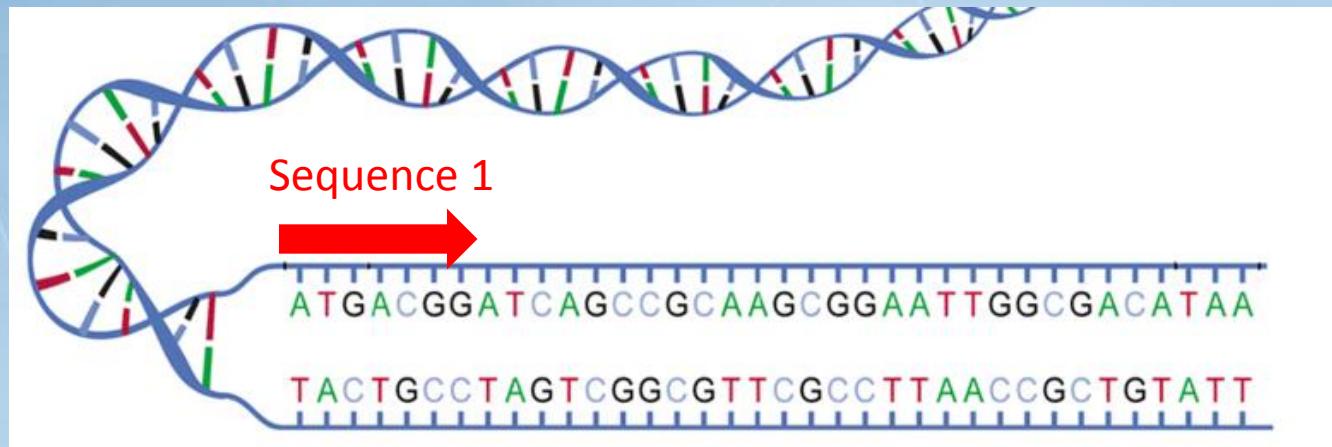
TTCAACAAACAGGCCAC
TTCACCAACAGGCCAC
TTCATCAACAGGCCAC

GOALS:

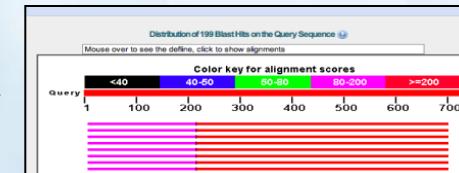
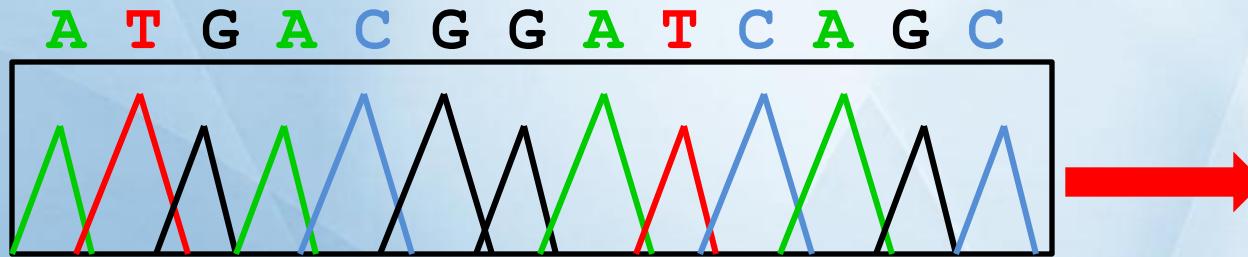
- Identify the organism from which the DNA was obtained.
- Compare DNA sequences to each other.



Sequence Both Strands of DNA



Sequence :

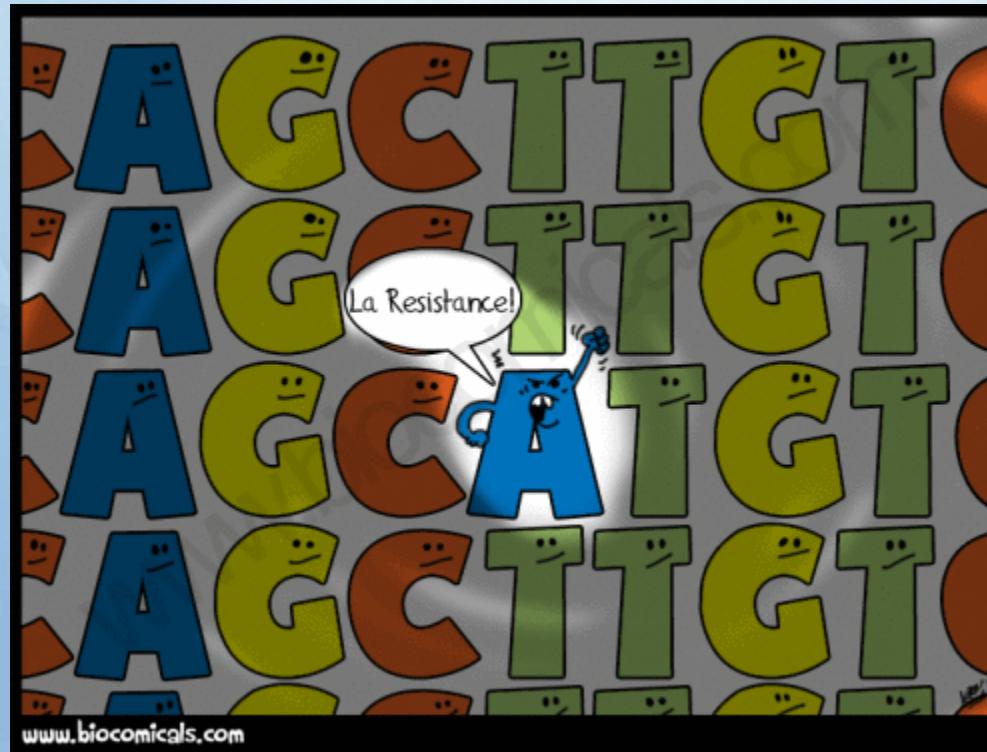


Bioinformatics tools like **BLAST** can be used to compare the sequences from Different individuals.



Comparing two sequences

MVNLT**SDEKTAVLALWNKVDVEDCGGE**
||| ||| | | | | | | | | | | | | | | |
MVHLT**PEEKTAVNALWGKVNVDAVGGE**





Multiple sequence alignment

Seq1 VTISCTGSSSNIGAG-NHVWKWYQQLPG

Seq2 VTISCTGTSSNIGS--ITVNWYQQLPG

Seq3 LRLSCSSSGFIFSS--YAMYWVRQAPG

Seq4 LSLTCTVSGTSFDD--YYSTWVRQPPG

Seq5 PEVTCVVVDVSHEDPQVKFNWYVDG--

Seq6 ATLVC LISDFYPGA--VTVAWKADS--

Seq7 AALGCLVKDYFPEP--VTWSWNSG---

Seq8 VSLTCLVKGFYPSD--IAVEWWNSNG--

Similar to pairwise alignment BUT **n** sequences are aligned instead of just **2**

Each row represents an individual sequence

Each column represents the ‘same’ position



Multiple sequence alignment

Seq1 VTIS**C**TGSSSN**I****GAG**-NHV**KW**YQ**QL****PG**

Seq2 VTIS**C**TGTSSNI**GS**--ITV**NW**YQ**QL****PG**

Seq3 LRLS**C**SSSGFIF**SS**--YAM**YW**VR**QA****PG**

Seq4 LSLT**C**TVSGTSF**DD**--YYST**TW**VR**QP****PG**

Seq5 PEVT**C**VVVDVSH**ED**PQVKF**NW**YV**DG**--

Seq6 ATL**C**LISDFYP**GA**--VTV**AW**KADS--

Seq7 AALG**C**LVKDYFP**EP**--VTV**SW**NSG---

Seq8 VSLT**C**LVKGFY**PSD**--IAV**EW**WS**NG**--





Multiple Sequence Alignment: Approaches

- Local alignment – finds regions of high similarity in parts of the sequences

| | |
|-----------------|--------|
| ADLGAVFALCDRYFQ | |
| | |
| ADLGRTQN | CDRYYQ |

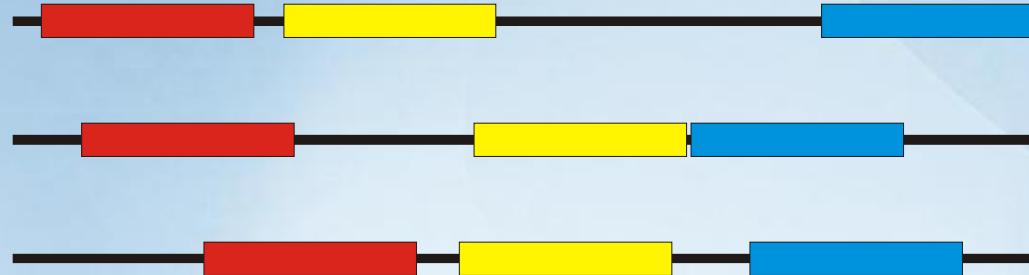
- Global alignment – finds the best alignment across the entire two sequences

| | |
|-----------------|--------|
| ADLGAVFALCDRYFQ | |
| | |
| ADLGRTQN | CDRYYQ |

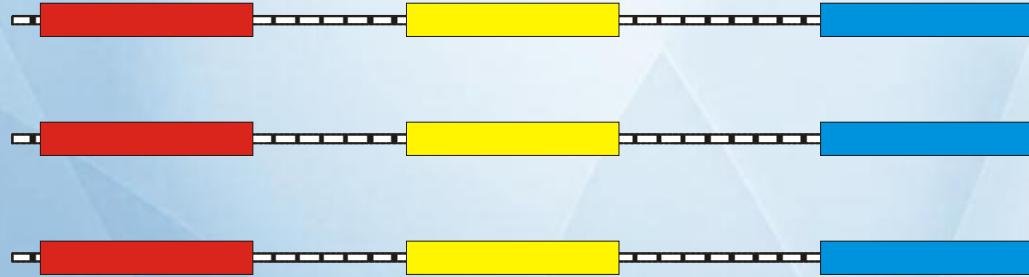


How to optimize alignment algorithms?

- ◆ Sequences often contain highly conserved regions



These regions can be used for an initial alignment



By analyzing a number of small, independent fragments,
the algorithmic complexity can be drastically reduced!

BioEdit Sequence Alignment Editor - [C:\Thommy\Academy\Hydrobiidae project\18S_aligned.bio]

File Edit Sequence Alignment View World Wide Web Accessory Application RNA Options Window Help

Courier New 11 B

84 total sequences

Mode: Edit Insert Selection: 183 Position: 12: Avenionia_2241 Sequence Mask: None Numbering Mask: None Start ruler at: 1

MI Scroll speed slow fast

| | 140 | 150 | 160 | 170 | 180 | 190 | 200 | 210 | 220 | 230 | | | |
|--------------|---------|--|-----|-------|------------|-----------------------|----------------|--------------------|------------|---------------|------------|------------|------|
| Adriohydrobi | CCAACC | GUCCGACCC | | CUG | ACGGG | | | AAAGAGC | GC | CUUUUAUCAG | CUCA | | |
| Adrioinsulan | CCAACC | GUCCGACCC | | | CUUCAACGGG | | | AAAGAGC | GC | CUUUUAUCAG | CUCA | | |
| Alvania | CAAA | CAGCUCCGACCC | | UCA | | | | GGGAAAGAGC | GC | CUUUUAUUAG | UUCA | | |
| Alzonella 2 | CCAACC | GUCCGACCC | | UUC | ACGGG | | | AAAGAGC | GC | CUUUUAUCAG | CUCA | | |
| Amnicola 106 | CUACCAG | GUCCGACCCGGUGGGCCUCGCUUCGGCUUUCUCCUGUCACAGGGGGAGUCGGGU | | | GU | | | CCCGUUGGGGAAGAGC | GC | CUUUUAUUAG | UUCA | | |
| Amphithalamu | CUACCAG | GUCCGACCC | | | GUGGU | | | CAAAGCCAGGAAAGAGC | GC | CUUUUAUUAG | UUCA | | |
| Antroselates | CUACCAG | GUCCGACCCGGUGGGCCUCGGCUUUCUUGUCAAAAGGGGGAGUUGGGCCGGGA | | | UUCGUCA | | | UUCCGUCAAGGAAGAGC | GC | CUUUUAUUAG | UUCA | | |
| Ascorhis | AUAC | AAGCUCCGAC | U | | CAAGGGG | | | ACGAGC | GC | CUUUUAUUAG | UUCA | | |
| Assiminea | CCAC | CAGCUCCGACCC | UGG | | UUUCGG | | | UCAGGGAAAGAGC | GC | CUUUUAUUAG | UUCA | | |
| Assiminea 16 | CCAC | CAGCUCCGACCC | | | GGUCUC | UCGAG | | GCCAGGGAAAGAGC | GC | CUUUUAUUAG | UUCA | | |
| Assiminea 22 | CCAC | CAGCUCCGACCC | | | GGUUUC | - | G | GUCAGGGAAAGAGC | GC | CUUUUAUUAG | UUCA | | |
| Avenionia 22 | CCAACC | GUCCGACCC | | UUC | ACGGG | | | AAAGAGC | GC | CUUUUAUCAG | CUCA | | |
| Baicalia | UAA | CAGCUCCGACCC | | | CCCUC | A | | GGGCC | CC | GGGAAAGAGC | GC | CUUUUAUUAG | UUCA |
| Barleeia | CCCCC | AAGCUCCGACCCG | | U | UCC | | | 12: Avenionia_2241 | GAAAGAGC | GC | UUUUUAUUAG | UUCA | |
| Beddomeia | UAA | CAGCUCCGACCC | | | GCA | | | AGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Belgrandia | CCAACC | GUCCGACCC | | U | UGC | AAAGG | | AAAGAGC | GC | UUUUUAUCAG | CUCA | | |
| Bithynia | CCAA | CAGCUCCGACCC | | | CUUC | | | AACGGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Bythinella 1 | CCAC | CAGCUCCGACCC | | | CUUCGCA | AGGGAGG | | GGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Bythiospeum | CUAA | CAGCUCCGACCC | | | UCA | | | CGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Calopia | CUAG | UAGUC | U | | UCA | | | CGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Cecina 2522 | CCAC | CAGCUCCGACCC | | GGUCA | | | AACAGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | | |
| Clenchiella | CUAA | CAGCUCCGACCC | | UCA | | | CGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | | |
| Coxiella | CCAC | CAGCUCCGACCC | U | | AGUCUUUC | GAG | | GCCGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Eatonella | GCAAA | CCUAGUGU | U | | ACUGG | | | GGGAAGA | U | GGAGCCGACUUUA | UUCA | | |
| Emmericia | CUAA | CAGCUCCGACCC | | U | UCAC | | | GGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Emmericia 30 | CUAA | CAGCUCCGACCC | | | CUCAC | - | | GGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Erhaia 652 | CCCCC | AAGCUCCGACCCG | | | CUCUUCGCC | GG | | GGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Fairbankia | CCAA | CAGCUCCGACCC | | U | UCA | | | GGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Fissuria 243 | CCAA | CAGCUCCGACCC | | U | UUG | ACGGG | | AAAGAGC | GC | UUUUUAUCAG | CUCA | | |
| Fluvidona | CCCCC | AAGCUCCGACCC | | U | GU | | | GGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Fluvipupa | CCCAC | AAGCUCCGACCC | U | | GGGUUC | UGC | | CACGGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Fontigenes | ACCC | AAGCUCCGACCC | U | | AACCA | | | CCCGGU | GGGAAAGAGC | GC | UUUUUAUUAG | CUCG | |
| Gammatricula | CCAC | CAGCUCCGACCC | U | | GGUCA | | | GGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Geomelania 8 | CCAC | CAGCUCCGACCCGG | | | CCGCCU | GUUUUACAGGGCAGGGUCUGG | | AAGGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Geomelania 8 | CCAC | CAGCUCCGACCCGG | | | CCGCCU | GUUUUACAGGGCAGGGUCUGG | | AAGGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Graziana 256 | CCAA | CAGCUCCGACCC | | | CGC | AAGGG | | AAAGAGC | GC | UUUUUAUCAG | CUCA | | |
| Hauffenia 25 | CCAA | CAGCUCCGACCC | | UUC | ACAGG | | | AAAGAGC | GC | UUUUUAUCAG | CUCA | | |
| Heleobops | UAA | CAGCUCCGACCC | | | UCA | | | GGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Hemistomia | CCAC | CAGCUCCGACCC | | U | GU | | | CCUGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Heterocyclus | CCGCC | AAGCUCCGACCC | U | | CGCUGAA | CG | | GGGGAGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Horatia 2598 | CCAA | CAGCUCCGACCC | | U | UGC | AAAGG | | AAAGAGC | GC | UUUUUAUCAG | CUCA | | |
| Hydrobia 653 | CCAA | CAGCUCCGACCC | | U | CGC | AAAGG | | AAAGAGC | GC | UUUUUAUCAG | CUCA | | |
| Hydrococcus | CCCA | AAGCUCCGACCC | U | | GGUGA | | | GGGGAAAGAGC | GC | UUUUUAUUAG | UUCA | | |
| Islamia 2327 | CCAA | CAGCUCCGACCC | | | CUC | ACGGG | | AAAGAGC | GC | UUUUUAUCAG | CUCA | | |



Choosing an alignment:

- Many **different** alignments between two sequences are possible:

AAGCTGAATTCTGAA
AGGCTCATTCTGA

AAGCGAAA**T**TCGAAC
A-G-GAA-**C**TCGAAC

AAGCGAAA**T**TCGAAC
AGG---**AA****C**TCGAAC

How do we determine which is the best alignment?



Assessing the significance of an alignment score

True

AAGCTGAATCGAA
AGGCTCATTCTGA

AAGCTGAATTC-GAA
AGGCTCATTCTGA-

28.0

Random

AGATCAGTAGACTA
GAGTAGCTATCTCT

AGATCAGTAGACTA-----
-----GAGTAG-CTATCTCT

26.0

CGATAGATAGCATA
GCATGTCATGATTG

CGATAGATAGCATA-----
-----GCATGTCATGATTG

16.0



“Optimal” vs. “correct” alignment

For a given group of sequences, there is no single “correct” alignment, only an alignment that is “optimal” according to some set of calculations

This is partly due to:

- the complexity of the problem,
- limitations of the scoring systems used,
- our limited understanding of life and evolution

Determining what alignment is best for a given set of sequences is really up to the judgment of the investigator

Success of the alignment will depend on the similarity of the sequences. If sequence variation is great it will be very difficult to find an optimal alignment



Web servers for pairwise alignment



NCBI



C www.ncbi.nlm.nih.gov

NCBI Resources How To

[Sign in to NCBI](#)



National Center for
Biotechnology Information

All Databases ▾

Search

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [NCBI News](#)

Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

NCBI YouTube channel

Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel.



|| 1 2 3 4 5 6 7 8

Popular Resources

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

[PubMed Health](#)

[BLAST](#)

[Nucleotide](#)

[Genome](#)

[SNP](#)

[Gene](#)

[Protein](#)

[PubChem](#)

NCBI Announcements

Coffee Break tutorial: Brown fat and obesity

Apr 1, 2014

The latest Coffee Break tutorial discusses EUMT1 or SPTAN1

New NCBI YouTube video: Create custom databases for BLAST

Mar 28, 2014

In the newest NCBI video on YouTube, we show you how to create custom



Basic Local Alignment Search Tool



BLAST

Basic Local Alignment Search Tool

NCBI/ BLAST/ blastn suite

blastn **blastp** **blastx** **tblastn** **tblastx**

Enter Query Sequence

Enter accession number, gi, or FASTA sequence [?](#) [Clear](#) **Query subrange** [?](#)
From
To

Or, upload file [Browse...](#) [?](#)

Job Title
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence [?](#) [Clear](#) **Subject subrange** [?](#)
From
To

Or, upload file [Browse...](#) [?](#)

Program Selection

Optimize for Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
Choose a BLAST algorithm [?](#)

BLAST [Search nucleotide sequence using Megablast \(Optimize for highly similar sequences\)](#) Show results in a new window

▶ [Algorithm parameters](#)

blastn –
nucleotide
blastp –
protein



BLAST – programs

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

► NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New Aligning Multiple Protein Sequences? Try the [COBALT Multiple Alignment Tool](#). [Go](#)

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

| | | |
|---|--|--|
| <input type="checkbox"/> Human | <input type="checkbox"/> Oryza sativa | <input type="checkbox"/> Gallus gallus |
| <input type="checkbox"/> Mouse | <input type="checkbox"/> Bos taurus | <input type="checkbox"/> Pan troglodytes |
| <input type="checkbox"/> Rat | <input type="checkbox"/> Danio rerio | <input type="checkbox"/> Microbes |
| <input type="checkbox"/> Arabidopsis thaliana | <input type="checkbox"/> Drosophila melanogaster | <input type="checkbox"/> Apis mellifera |

Basic BLAST

Choose a BLAST program to run.

| | |
|----------------------------------|--|
| nucleotide blast | Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontiguous megablast |
| protein blast | Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phi-blast |
| blastx | Search protein database using a translated nucleotide query |
| tblastn | Search translated nucleotide database using a protein query |
| tblastx | Search translated nucleotide database using a translated nucleotide query |

nucleotide blast is circled in orange.

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscren)
- Align two (or more) sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript libraries](#)
- Constraint Based Protein [Multiple Alignment Tool](#)



BLAST – bl2seq



BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

► NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. [Go](#)

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

| | | |
|---|--|--|
| <input type="checkbox"/> Human | <input type="checkbox"/> Oryza sativa | <input type="checkbox"/> Gallus gallus |
| <input type="checkbox"/> Mouse | <input type="checkbox"/> Bos taurus | <input type="checkbox"/> Pan troglodytes |
| <input type="checkbox"/> Rat | <input type="checkbox"/> Danio rerio | <input type="checkbox"/> Microbes |
| <input type="checkbox"/> Arabidopsis thaliana | <input type="checkbox"/> Drosophila melanogaster | <input type="checkbox"/> Apis mellifera |

Basic BLAST

Choose a BLAST program to run.

| | |
|----------------------------------|--|
| nucleotide blast | Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontiguous megablast |
| protein blast | Search protein database using a protein query <i>Algorithms:</i> blastp, psi-blast, phr-blast |
| blastx | Search protein database using a translated nucleotide query |
| tblastn | Search translated nucleotide database using a protein query |
| tblastx | Search translated nucleotide database using a translated nucleotide query |

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscren)
- Align two (or more) sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript libraries](#)
- Constraint Based Protein [Multiple Alignment Tool](#)



Sequence alignment - Wiki x Multiple Sequence Alignm x COBALT:Multiple Alignm x

www.ncbi.nlm.nih.gov/tools/cobalt/cobalt.cgi?link_loc=BlastHomeLink

COBALT Constraint-based Multiple Alignment Tool

Home Recent Results Help Cobalt Constraint-based Multiple Protein Alignment Tool

COBALT computes a multiple protein sequence alignment using conserved domain and local sequence similarity information. [?](#) [Reset page](#)

Enter Query Sequences

Enter at least 2 protein accessions, gis, or FASTA sequences [?](#) [Clear](#)

Or, upload FASTA file [Choose File](#) No file chosen

Job Title

Align Show results in a new window

► [Advanced parameters](#)

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback

NCBI | NLM | NIH | DHHS



BL2seq results

| Alignments | | <input type="checkbox"/> Select All | Get selected sequences NEW |
|---|--|-------------------------------------|--|
| <pre>>loc 43385 gi 509026971gb AAT86412.1 ABC transporter-associated protein [Streptococcus pyogenes MGAS10394] Length=472 Score = 394 bits (1013). Expect = 5e-114. Method: Compositional matrix adjust Identities = 190/304 (62%). Positives = 232/304 (76%). Gaps = 4/304 (1%) Query 566 PTYDIQVVGLENFVANGIVAHNSFIYVPPGVHVDIPLQAYFRINTENMGQFERTLIIADT 625 P D ++ I + V +G +FIYVP GV VDIPLQ YFRIN EN GQFERTLII D Sbjct 173 PPTDNKLAALNSAVWSG----GTIFIYVPKGVKVDIPLQTYFRINNENTGQFERTLIIIVDE 228 Query 676 GSYVHIVETCTAPIYKSDSLRSAVVEIIIVKPHARVRYTTIQNWSNNVYNLVTKRARVETG 685 G+ V +IVECTAP Y S+S+H+A+VEI A +RYTTIQNWS+NVYNLVTKRAR T Sbjct 229 GASVYVECCTAPTYSSNSLHAATVEIFALDGAYMRYTTIQNWSDNVYNLVTKRARALTD 288 Query 686 ATMEWIDGNIGSKVTMKYPAVWMTCGAKGEVLISVAFAGEGQHQDTGAKMLHLASNTSSN 745 AT+EWIDGN+G+K TMKYP+V++ G +G +LS+AFA GQHQDTGAKM+H A +TSS+ Sbjct 289 ATVEWIDGNLGAKTTMKYPSVYLDGPAGTMLSIAFANAGQHQDTGAKMIHNAPHTSSS 348 Query 746 IVSKSVARGGGRTSYRGLVQVNKGAGHGSRSVKCDALLVDTISRSDTYPYVDIREDDVTM 805 IVSKS+A+GG+ YRG V NK + S S +LCD +L+D IS+SDT P+ +I V + Sbjct 349 IVSKSIAKSGGKVIVYRGQVTFNKQSKKSVSHKEETDILMDDISKSDTIPFNEIHNSQVAL 408 Query 806 GHEATVSKVSENQLFYLMMSRLGLADEEAMAMVVRGFPIAKELPMYEAYALELNRLIELQME 865 HEA VSK+SE QL-YLMSRGL+E EA M+V GVER KELPMYEYA+ELNRLI +ME Sbjct 406 EHEAKVSKISEEQLYLYLMSRGLSESEATEMIVMGFEPIKELPMYEAYAELNRLISYEME 468 Query 866 GAVG 869 G+VG Sbjct 469 GSVG 472 GSVG</pre> | | | |

Match

Similarity

Dissimilarity

Gaps



MSA input: multiple sequence Fasta file

>gi|4504351|ref|NP_000510.1| delta globin [Homo sapiens]

MVHLTPEEKAVNALWGKVNVDAVGGEALGRLLVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFSQLSELHCDKLHVDPENFRLGNVLVCVLARNFGKEFTPQMQAAYQKVVAGVAN
ALAHKYH

>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]

MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH

>gi|4885393|ref|NP_005321.1| epsilon globin [Homo sapiens]

MVHFTAAEKAATSLWSKMNVEEAGGEALGRLLVYPWTQRFFDSFGNLSSPSAILGNPKVKAHGKKVLT
SFGDAIKNMDNLKPAFAKLSELHCDKLHVDPENFKLLGNVMVIILATHFGKEFTPEVQAAWQKLVSAAV
ALAHKYH

>gi|6715607|ref|NP_000175.1| G-gamma globin [Homo sapiens]

MGHFTeedKATITSLWGKVNVEDAGGETLGRLLVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLT
SLGDAIKHDDLKGTFQLSELHCDKLHVDPENFKLLGNVLTVLAIHFGKEFTPEVQASWQKMVTGVAS
ALSSRYH

>gi|28302131|ref|NP_000550.2| A-gamma globin [Homo sapiens]

MGHFTeedKATITSLWGKVNVEDAGGETLGRLLVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLT
SLGDATKHDDLKGTFQLSELHCDKLHVDPENFKLLGNVLTVLAIHFGKEFTPEVQASWQKMVTAVAS
ALSSRYH

>gi|4885397|ref|NP_005323.1| hemoglobin, zeta [Homo sapiens]

MSLTKTERTIIVSMWAKISTQADTIGTETLERLFLSHPQTCKTYFPHFDLHPGSAQLRAHGSKVVAAGDA
VKSIDDIGGALSKLSELHAYILRVDPVNFKLLSHCLLVTLAARFPADFTAEEHAAWDKFLSVVSSVLTEK
YR



Query type: AA or DNA?

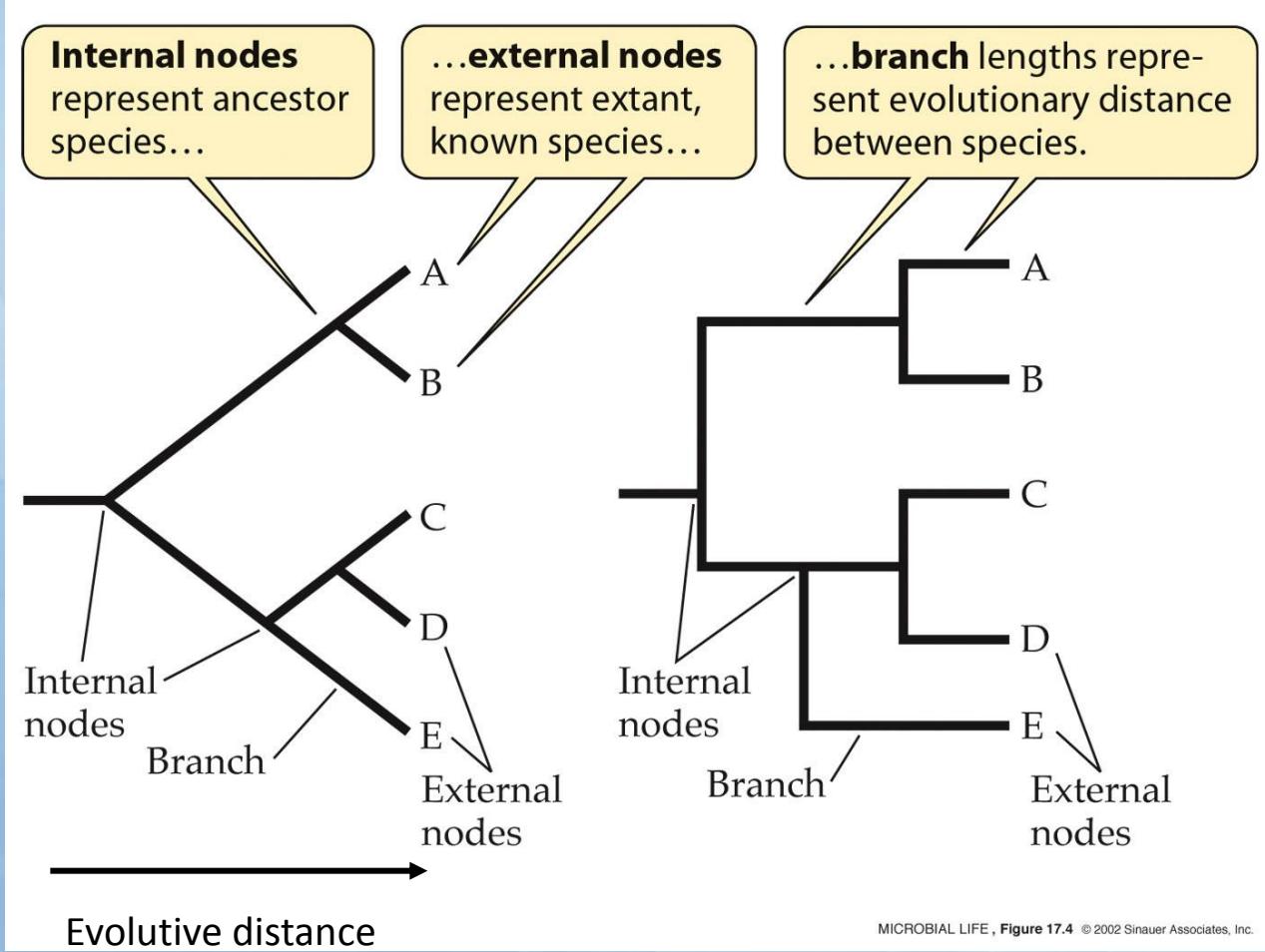
- For coding sequences, AA (protein) data are better
 - Selection operates most strongly at the protein level → the homology is more evident
 - AA – 20 char' alphabet DNA - 4 char' alphabet



lower chance of random homology for AA



Phylogenetic trees

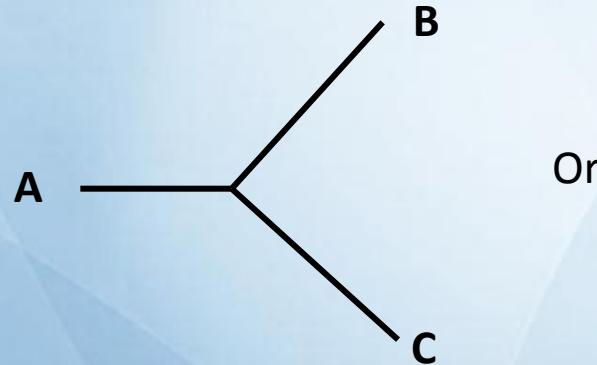




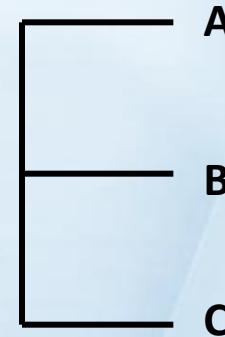
Phylogenetic trees

Rooted and unrooted trees

- **Unrooted trees:** compare one feature of a group of related organisms



Or



Only one shape for a tree with 3 species

Distance matrix: UPGMA



- UPGMA = unweighted pair group method with arithmetic mean

(A) UPGMA method

Table shows sequence of nine-base region of rRNAs of four strains.

| Organism (strain) | Site number | | | | | | | | |
|----------------------|-------------|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| a | G | C | G | G | A | C | A | A | A |
| b | G | A | C | G | C | C | A | A | G |
| c | G | A | A | A | U | C | U | A | A |
| d | G | A | A | A | G | C | U | A | G |

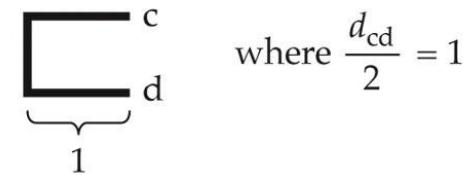
First matrix

1 Construct matrix showing relatedness between strains.

| | a | b | c |
|---|--------------|--------------|--------------|
| b | $d_{ab} = 4$ | — | — |
| c | $d_{ac} = 5$ | $d_{bc} = 5$ | — |
| d | $d_{ad} = 6$ | $d_{bd} = 4$ | $d_{cd} = 2$ |

Beginning tree

2 Diagram relatedness between strains.





Distance matrix: UPGMA

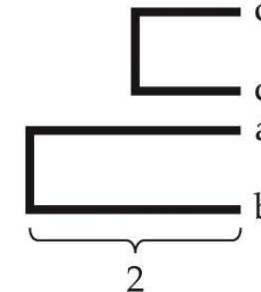
c & d are considered as one unit

Second matrix

| | (cd) | a |
|---|--------------------|--------------|
| a | $d_{(cd)a} = 11/2$ | — |
| b | $d_{(cd)b} = 9/2$ | $d_{ab} = 4$ |

- 3 Construct matrix to assess distance between (cd) and (a and b).

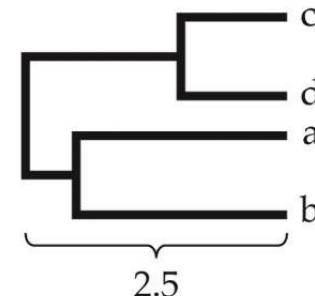
Second tree



- 4 Determine and diagram their relatedness.

$$\text{where } \frac{d_{ab}}{2} = 2$$

Final tree



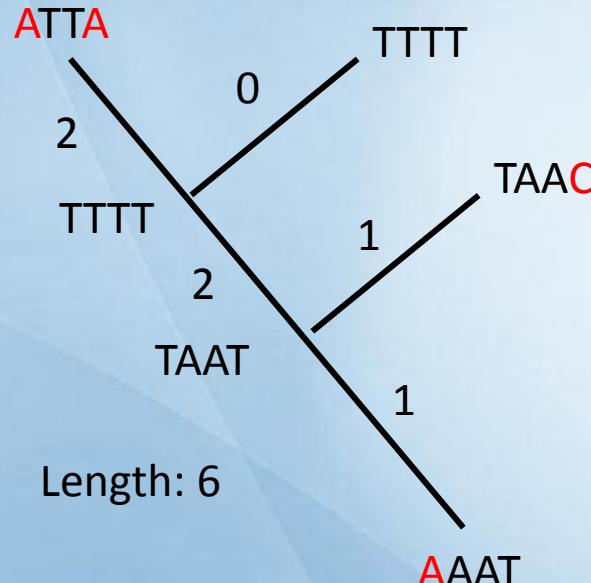
- 5 Determine relatedness between cd and a and b.

$$\text{where } d_{(cd)(ab)} = \frac{(11/2 + 9/2)/2}{2} = 2.5$$

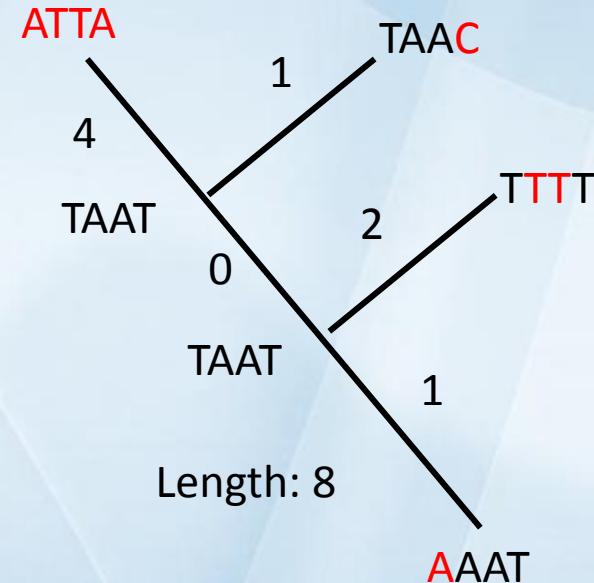


Other method: maximum parsimony

Sequence 1: ATTA
Sequence 2: TAAC
Sequence 3: AAAT
Sequence 4: TTTT



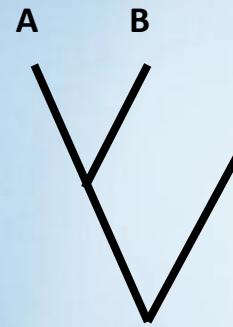
Most likely tree



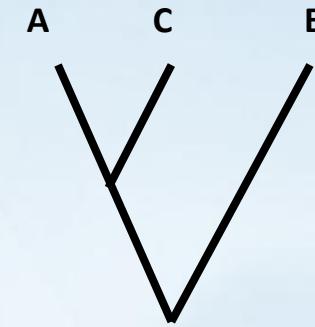


The alignment problem

- ◆ What happens when a sequence alignment is wrong?

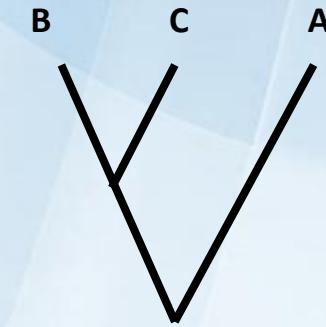


A: AGT
B: AT
C: ATC



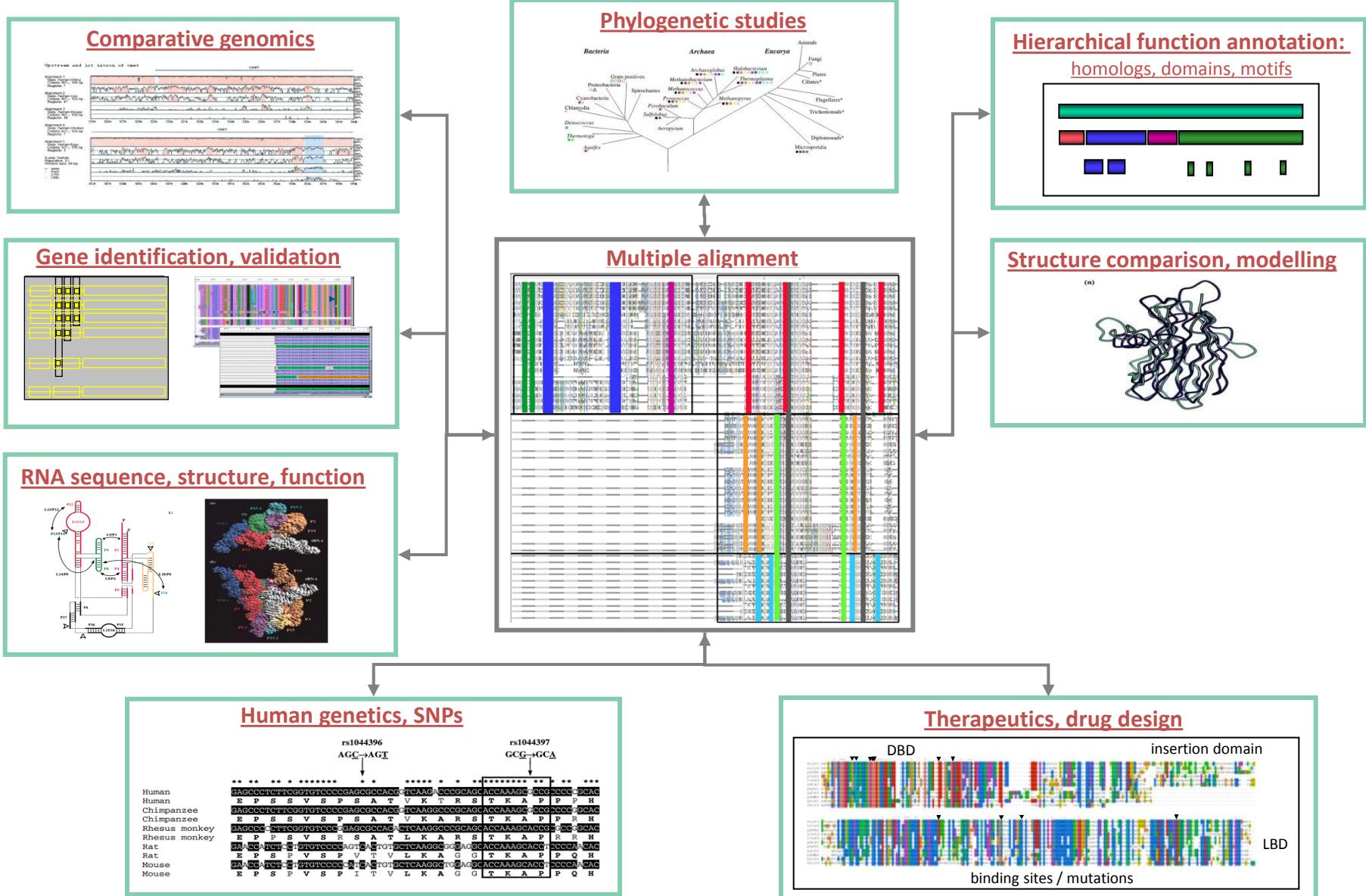
A: AGT
B: A -T
C: ATC

A: AGT -
B: A -T -
C: A -TC



A: AGT
B: AT -
C: ATC

Central role of multiple alignments

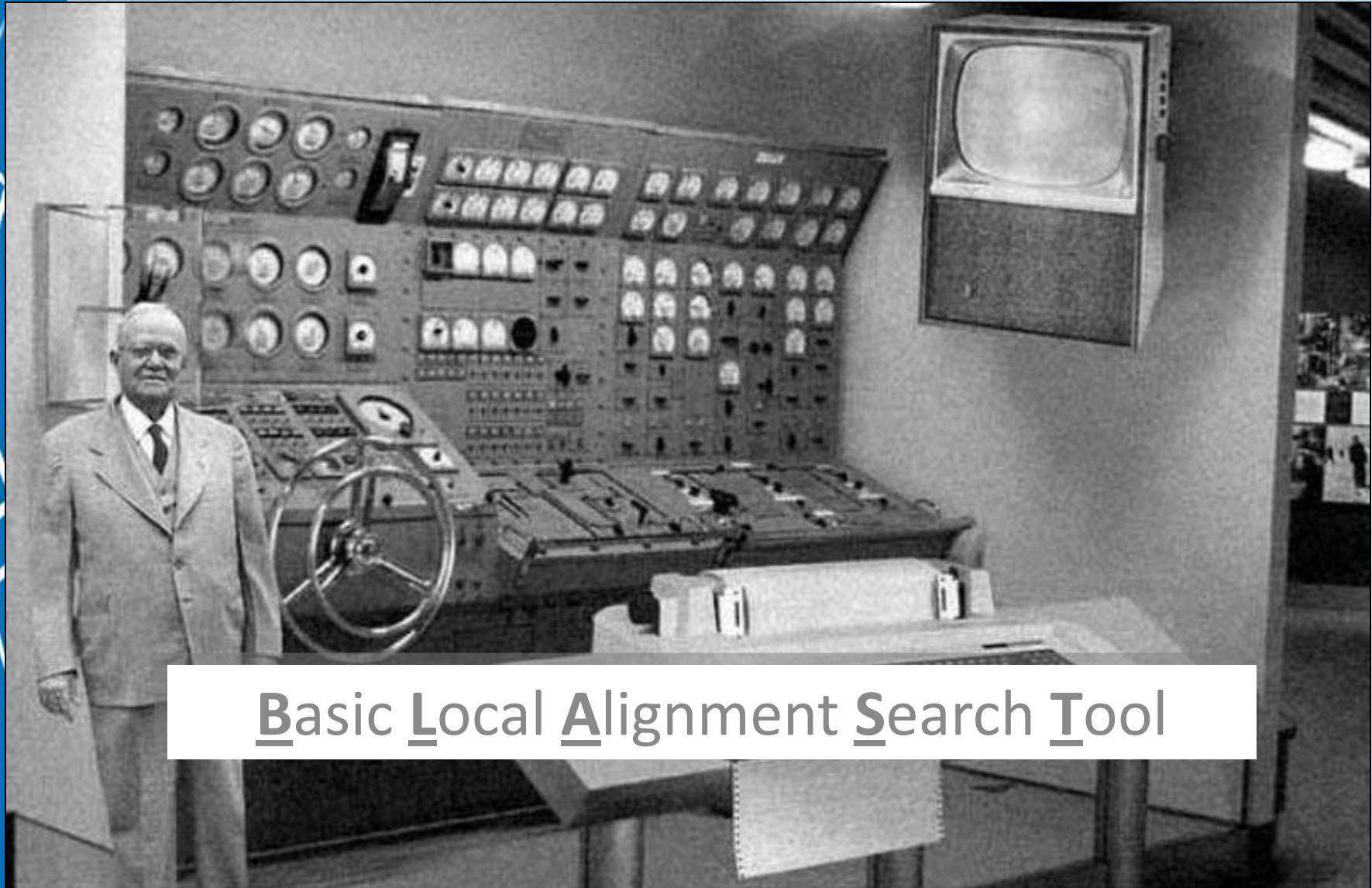


Most important sequence databases

- **Genbank** – maintained by USA National Center for Biology Information (NCBI)
 - All biological sequences
 - www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html
 - Genomes
 - www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome
- **Swiss-Prot** - maintained by EMBL- European Bioinformatics Institute (EBI)
 - Protein sequences
 - www.ebi.ac.uk/swissprot/



Sequence Similarity Searching



Basic Local Alignment Search Tool

