Sequencing alignment

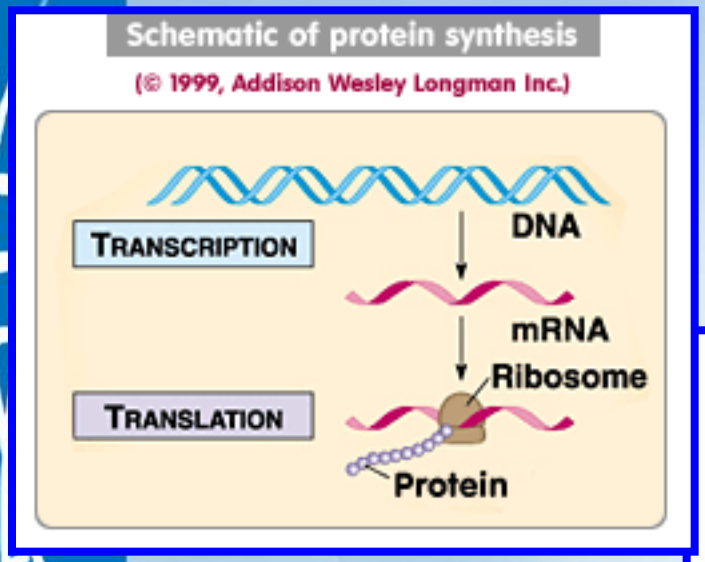
Ameer Effat M. Elfarash

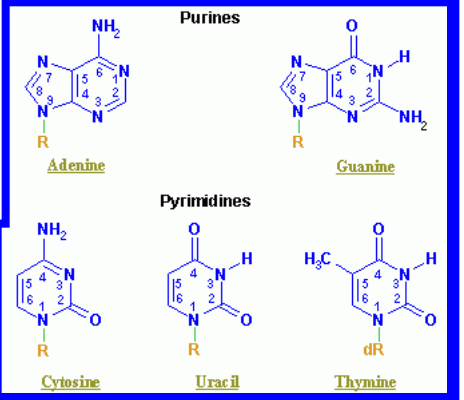
Dept. of Genetics
Fac. of Agriculture, Assiut Univ.
aelfarash@aun.edu.eg

Why perform a multiple sequence alignment?

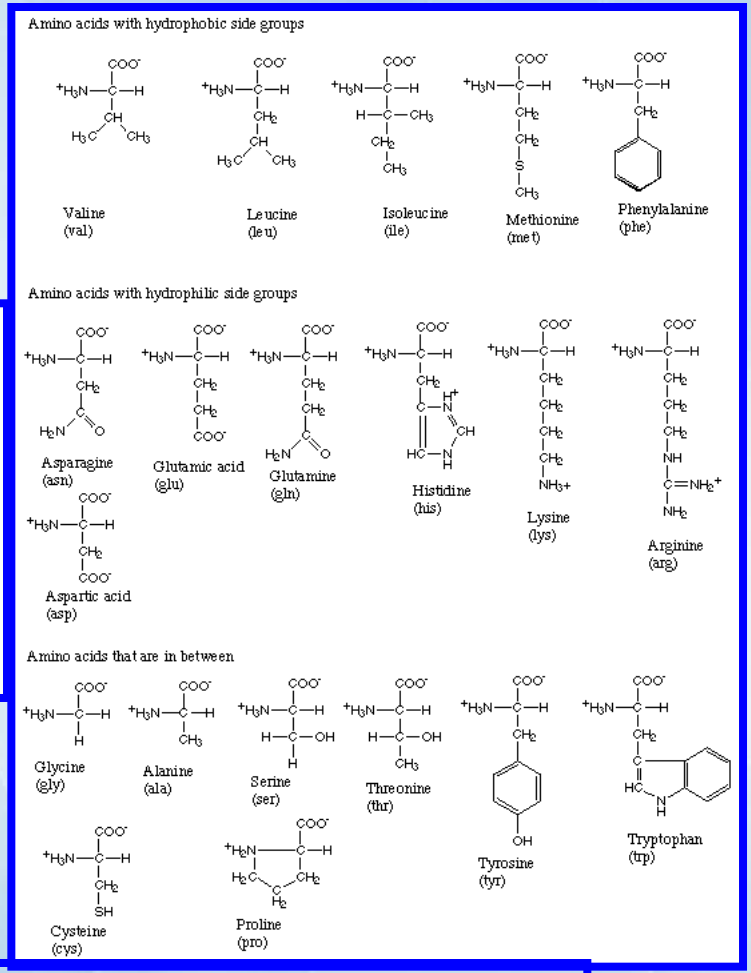
MSAs are at the heart of comparative genomics studies which seek to study evolutionary histories, functional and structural aspects of sequences, and to understand phenotypic differences between species

The Building Blocks…





ATGC



VLMFNQEDHKRCSTPYW

Three types of nucleotide changes:

Substitution – a replacement of one (or more) sequence characters by another: **AAGA → AACA**

Insertion - an insertion of one (or more) sequence characters: **AAG A**

Deletion – a deletion of one (or more) sequence characters: **A A GA**

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

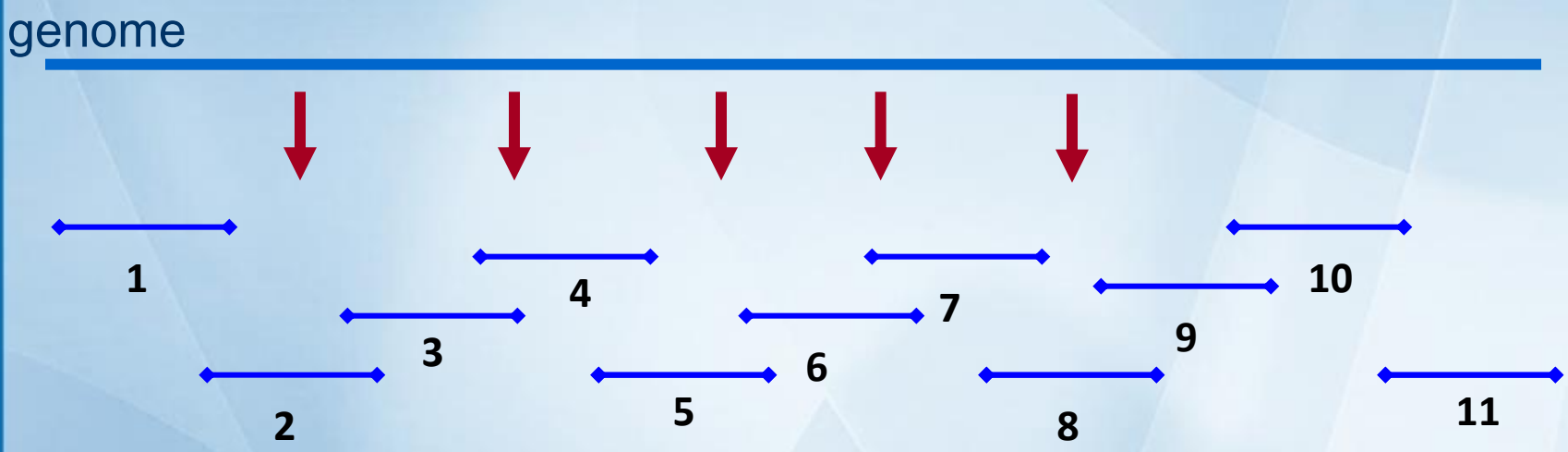
Alignment: Comparing two (pairwise) or more (multiple) sequences. Searching for a series of identical or similar characters in the sequences.

Why Align Sequences?

- Crucial for genome sequencing
- Discover functional, structural, and evolutionary information because similar Sequences may have similar function
- Identify primers and probes to search for homologous sequences in other organisms

Why Align Sequences?

- Genome Sequencing Strategy



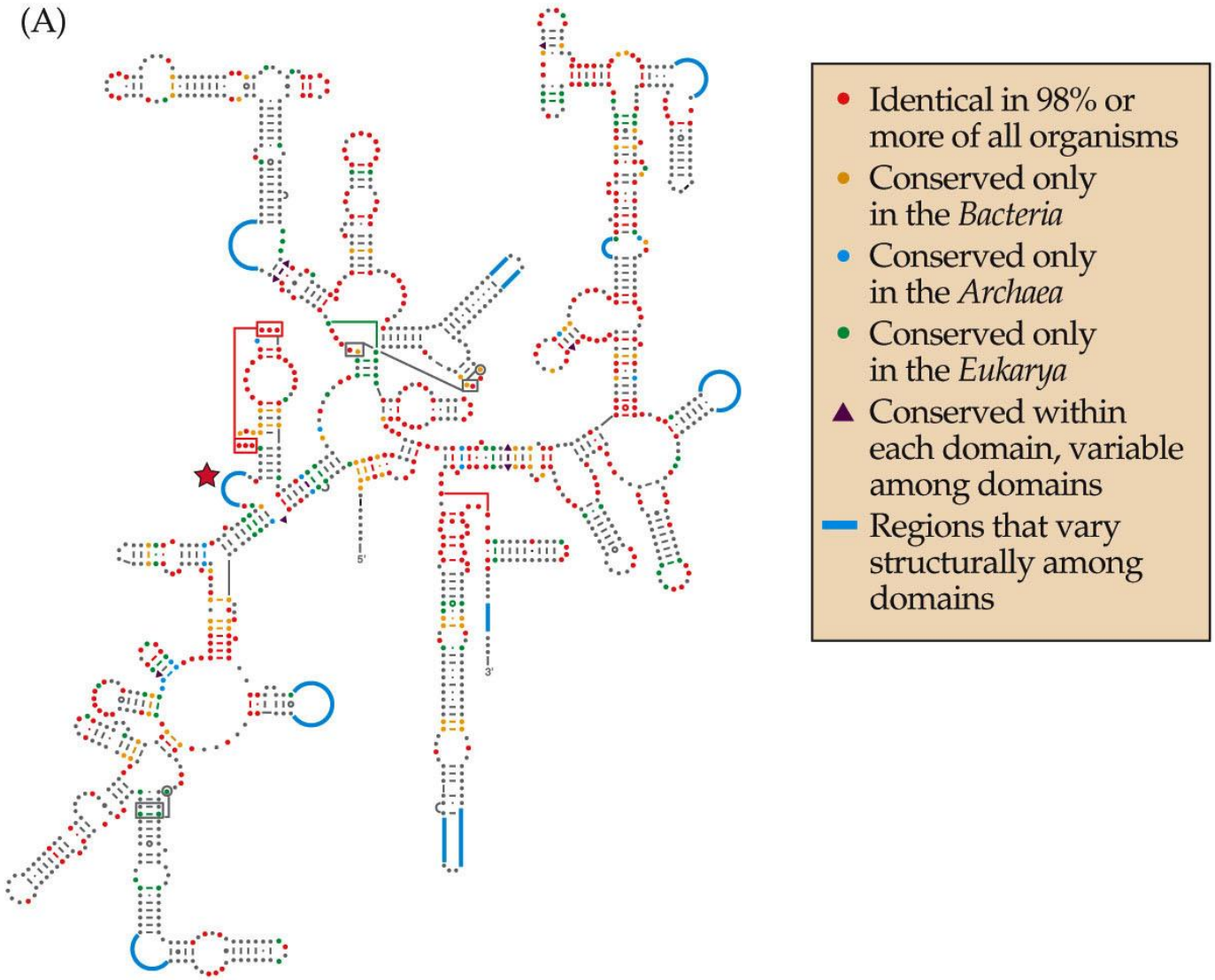
1. Obtain a large collection of BAC clones
2. Map them onto the genome (Physical Mapping)
3. Select a minimum tiling path
4. Sequence each clone in the path with shotgun
5. Assemble
6. Put everything together

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



***Microbial Life 2e*, Figure 17.3 (Part 1)**

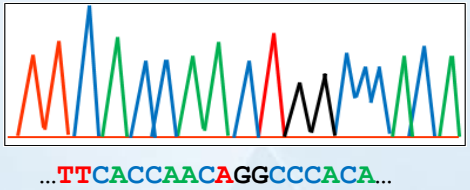
How DNA Sequence Data is Compared

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

Genetic Data

Extract DNA from Cells

Sequence DNA



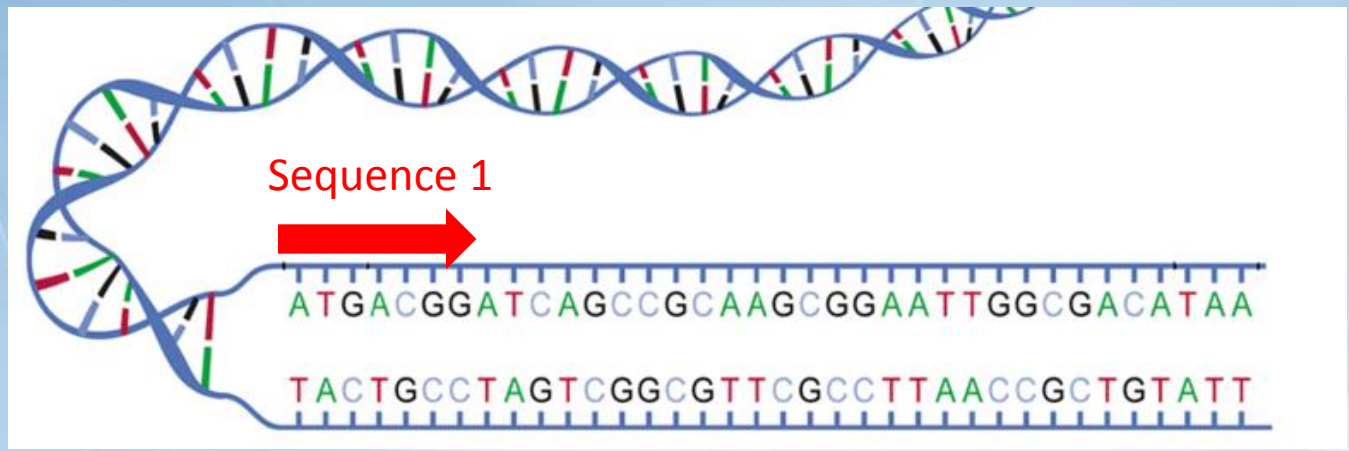
Compare DNA Sequences to One Another

```
TTCAACAACAGGCCCAC
TTCACCAACAGGCCCAC
TTCATCAACAGGCCCAC
```

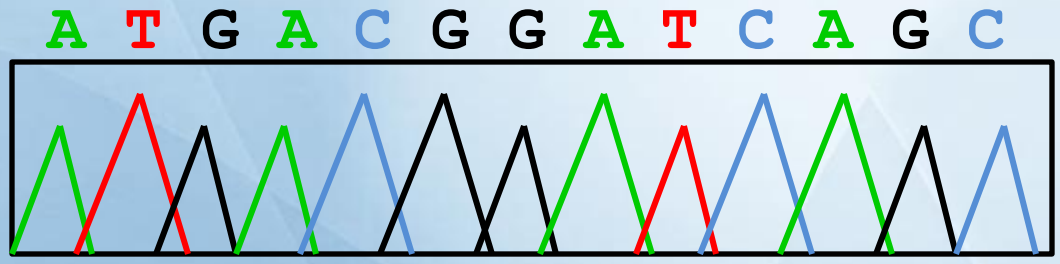
GOALS:

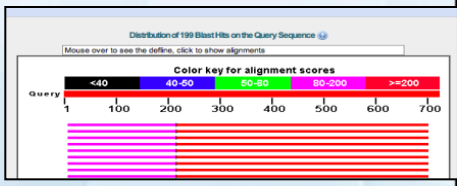
- Identify the organism from which the DNA was obtained.
- Compare DNA sequences to each other.

confirm insert sequence



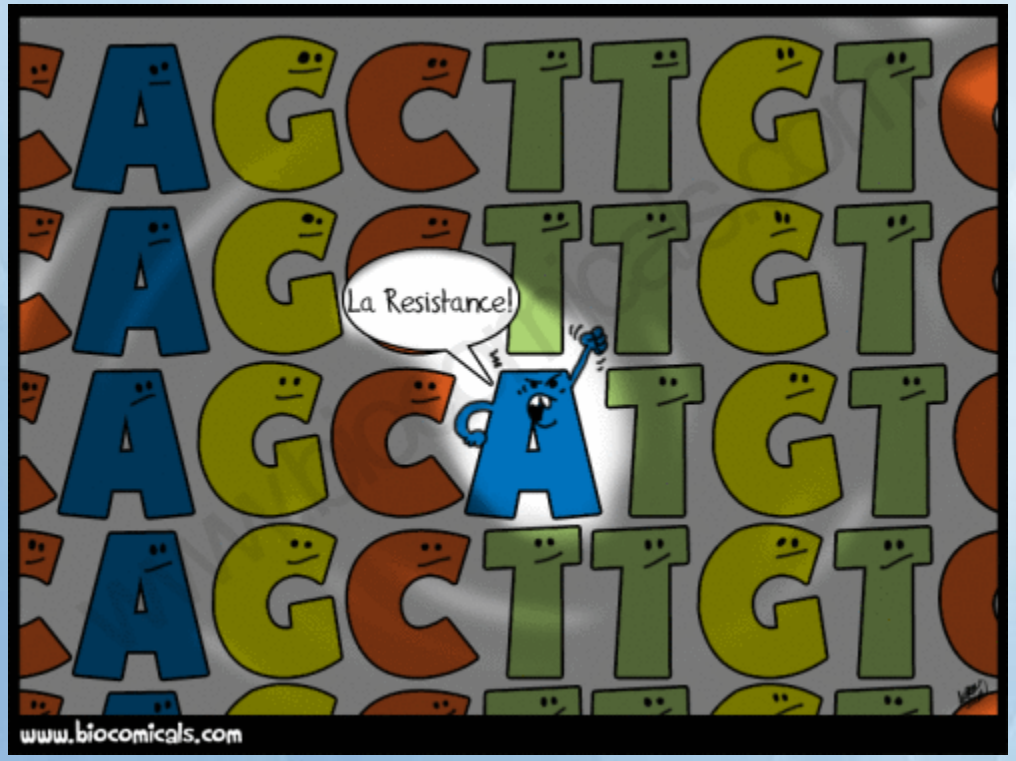
Sequence :





Comparing two sequences

```
MVNLTSDEKTAVLALWNKVDVEDCGGE
||  ||   |||||  |||  ||  ||    ||
MVHLTPEEKTAVNALWGKVNVDAVGGE
```



Multiple sequence alignment

```
Seq1 VTISCTGSSSNIGAG-NHVKWYQQLPG
Seq2 VTISCTGTSSNIGS--ITVNWYQQLPG
Seq3 LRLSCSSSGFIFSS--YAMYWVRQAPG
Seq4 LSLTCTVSGTSFDD--YYSTWVRQPPG
Seq5 PEVTCVVVDVSHEDPQVKFNWYVDG--
Seq6 ATLVCLISDFYPGA--VTVAWKADS--
Seq7 AALGCLVKDYFPEP--VTVSWNSG---
Seq8 VSLTCLVKGFYPSD--IAVEWWSNG--
```

Each row represents an individual sequence

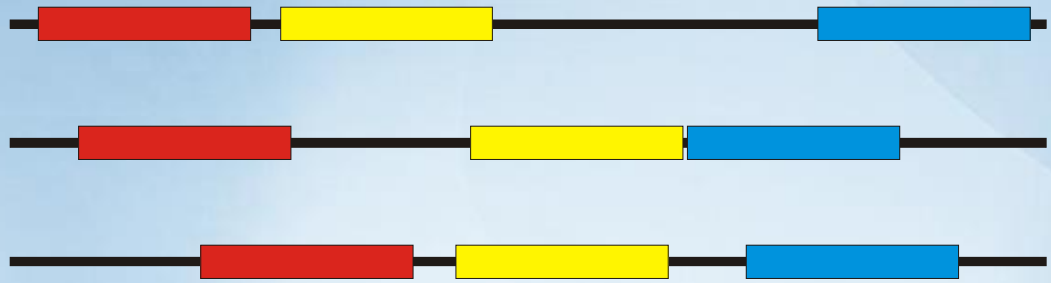
Each column represents the ‘same’ position

Multiple sequence alignment

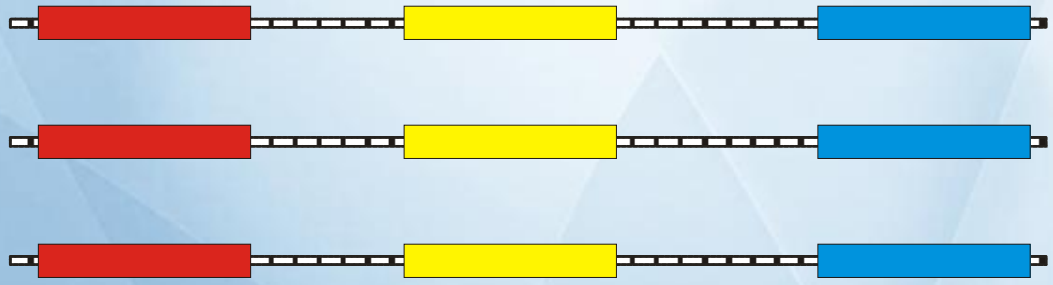
```
Seq1  VTISCTGSSSNIGAG-NHVKWYQQLPG
Seq2  VTISCTGTSSNIGS--ITVNWYQQLPG
Seq3  LRLSCSSSGFIFSS--YAMYWVRQAPG
Seq4  LSLTCTVSGTSFDD--YYSTWVRQPPG
Seq5  PEVTCVVVDVSHEDPQVKFNWYVDG--
Seq6  ATLVCLISDFYPGA--VTVAWKADS--
Seq7  AALGCLVKDYFPEP--VTVSWNSG---
Seq8  VSLTCLVKGFYPSD--IAVEWWSNG--
```

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!

Choosing an alignment:

- Many different alignments between two sequences are possible:

AAGCTGAATTCGAA

AGGCTCATTTCTGA

```
AAGCGAAATTCGAAC
A-G-GAA-CTCGAAC
```

```
AAGCGAAATTCGAAC
AGG---AACTCGAAC
```

How do we determine which is the best alignment?

Assessing the significance of an alignment score

True

AAGCTGAATTCGAA AGGCTCATTTCTGA	AAGCTGAATTC-GAA AGGCTCATTTCTGA- 28.0

Random

AGATCAGTAGACTA GAGTAGCTATCTCT	AGATCAGTAGACTA----- ----GAGTAG-CTATCTCT 26.0
.	
.	
CGATAGATAGCATA GCATGTCATGATTC	CGATAGATAGCATA--------- ---------GCATGTCATGATTC 16.0

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

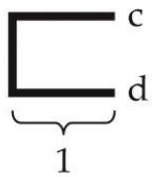
1 Construct matrix showing relatedness between strains.

First matrix

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

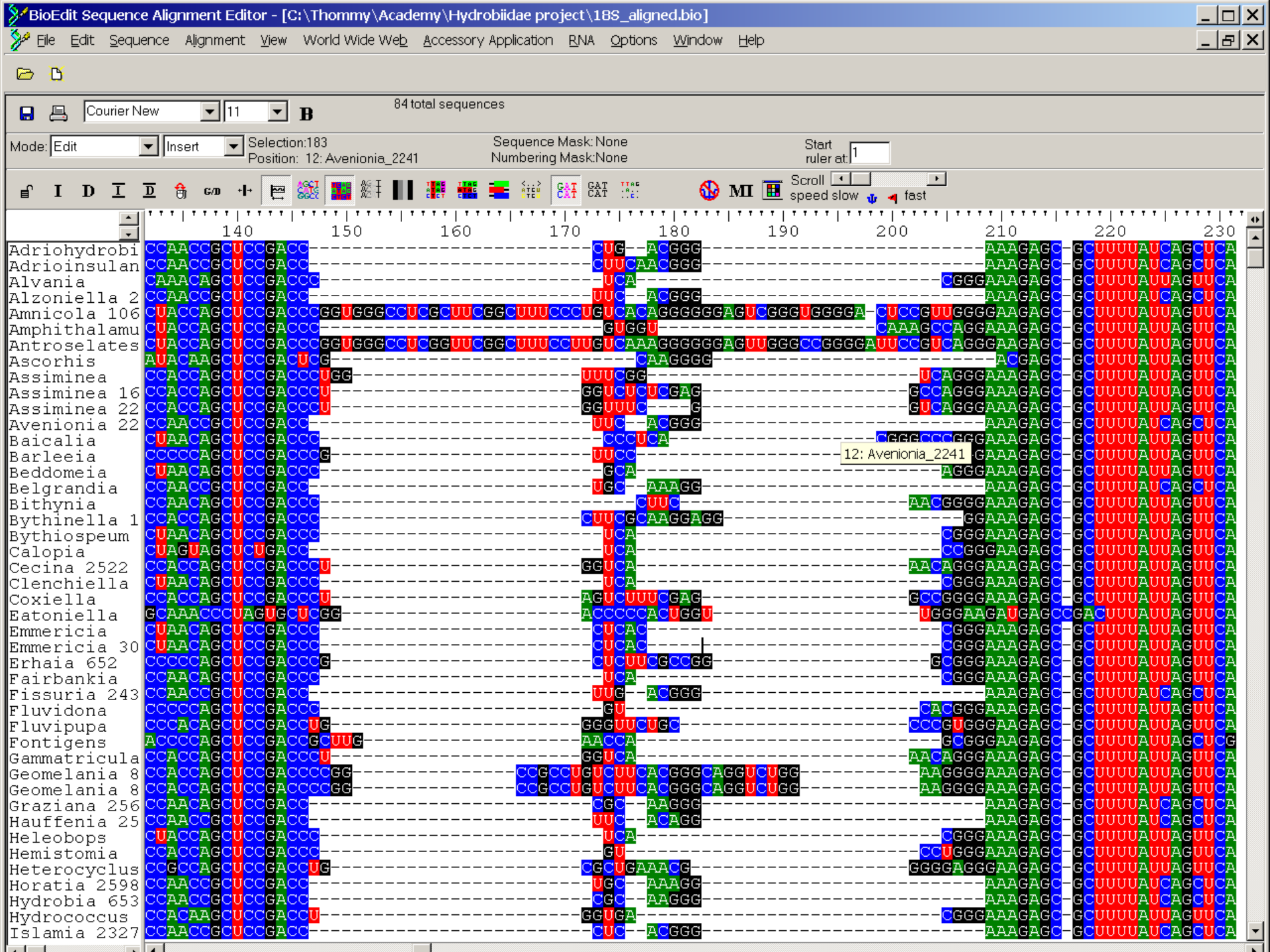
2 Diagram relatedness between strains.

Beginning tree



where $\frac{d_{cd}}{2} = 1$

MICROBIAL LIFE, **Figure 17.6 (Part 1)** © 2002 Sinauer Associates, Inc.


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
84 total sequences
Mode: Edit
Insert
Selection:183
Position: 12: Avenionia_2241
Sequence Mask: None
Numbering Mask:None
Start ruler at: 1
Scroll speed slow fast
140 150 160 170 180 190 200 210 220 230
12: Avenionia_2241
Adriohydrobi
Adrioinsulan
Alvania
Alzoniella 2
Amnicola 106
Amphithalamu
Antroselates
Ascorhis
Assiminea
Assiminea 16
Assiminea 22
Avenionia 22
Baicalia
Barleeia
Beddomeia
Belgrandia
Bithynia
Bythinella 1
Bythiospeum
Calopia
Cecina 2522
Clenchiella
Coxiella
Eatoniella
Emmericia
Emmericia 30
Erhaia 652
Fairbankia
Fissuria 243
Fluvidona
Fluvipupa
Fontigens
Gammatricula
Geomelania 8
Geomelania 8
Graziana 256
Hauffenia 25
Heleobops
Hemistomia
Heterocyclus
Horatia 2598
Hydrobia 653
Hydrococcus
Islamia 2327

"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

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

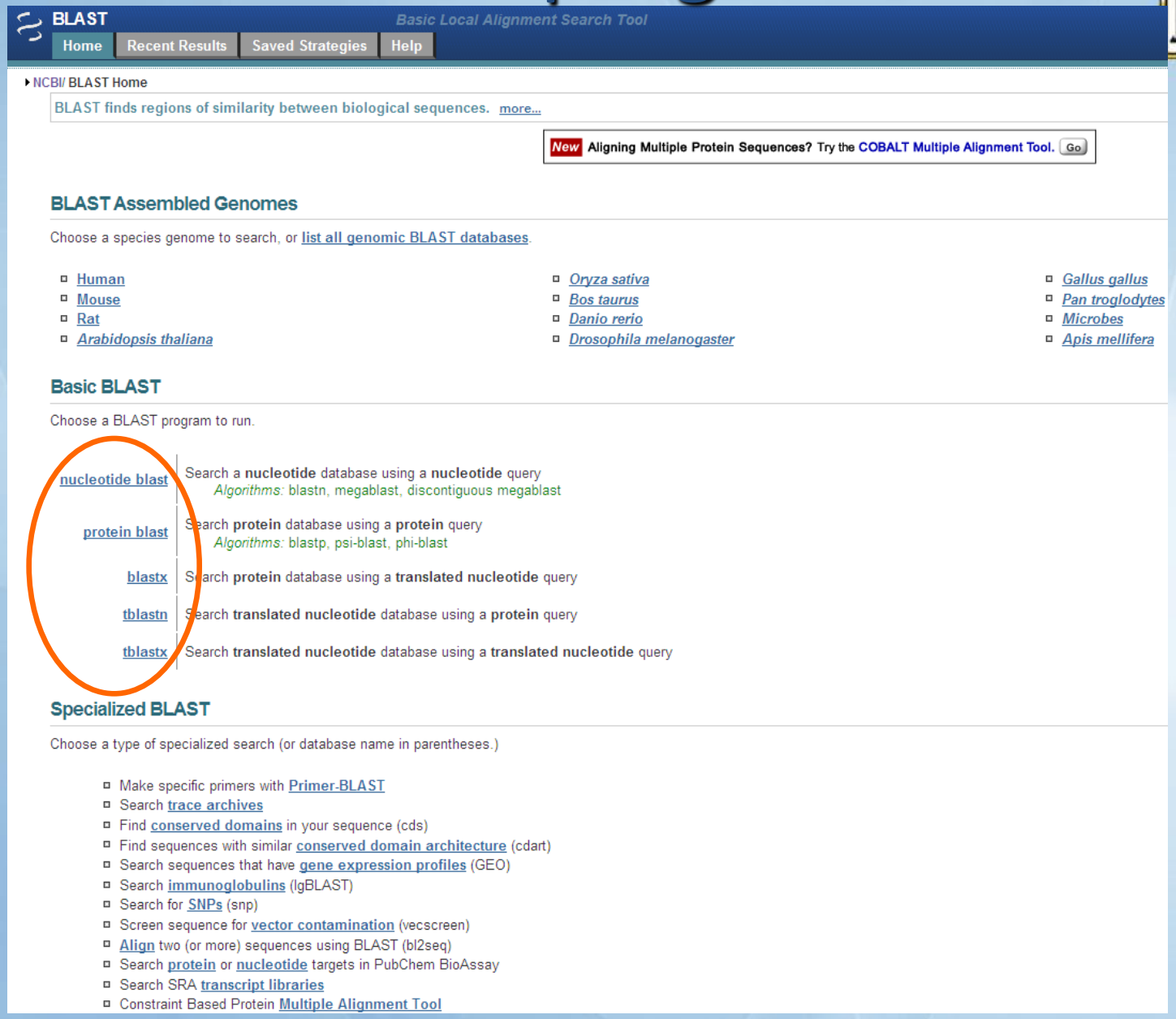
Success of the alignment will depend on the similarity of the sequences. If sequence variation is great it will be very difficult to find an optimal alignment

Web servers for pairwise alignment

NCBI


www.ncbi.nlm.nih.gov
NCBI Resources How To
Sign in to NCBI
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.
GO
You Tube
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
New NCBI YouTube video: Create custom databases for BLAST
Mar 28, 2014
In the newest NCBI video on YouTube,

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.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* 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

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 (vecscreen)
- 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

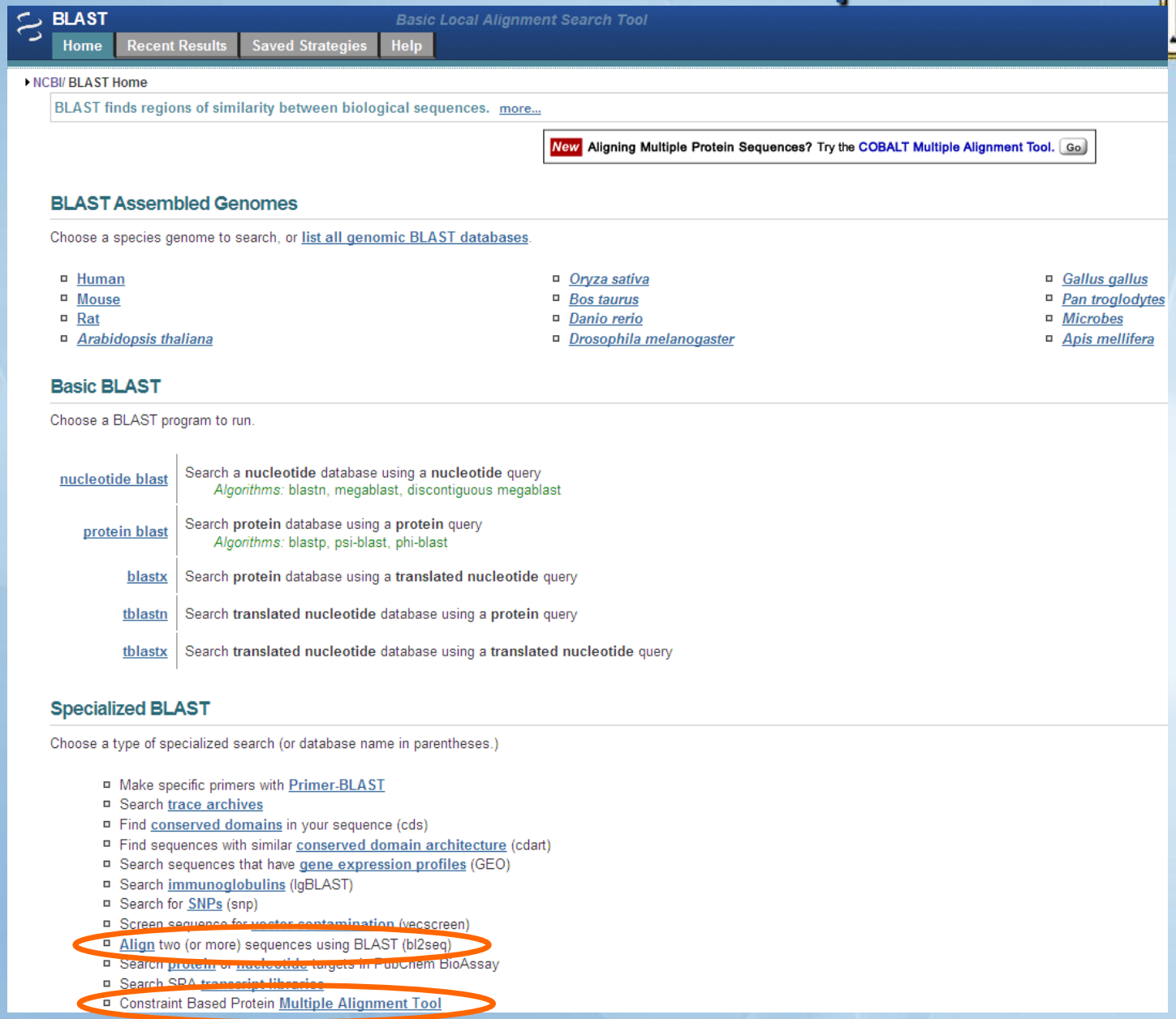
Query type: AA or DNA?

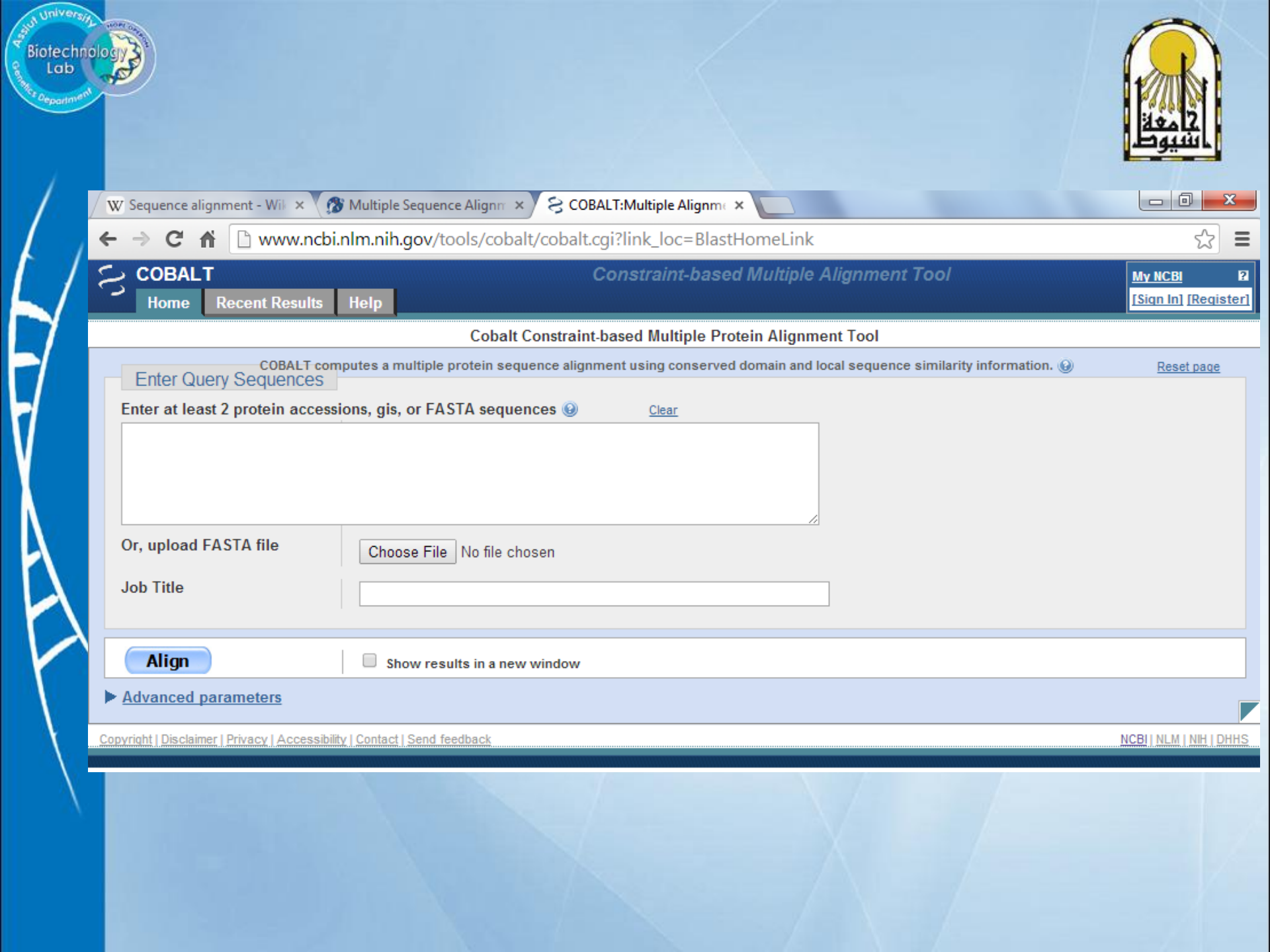
- For coding sequences, AA (protein) data are better
  - Selection operates most strongly at the protein level → the homology is more evident
  - AA – 20 char' alphabet DNA - 4 char' alphabet

↓

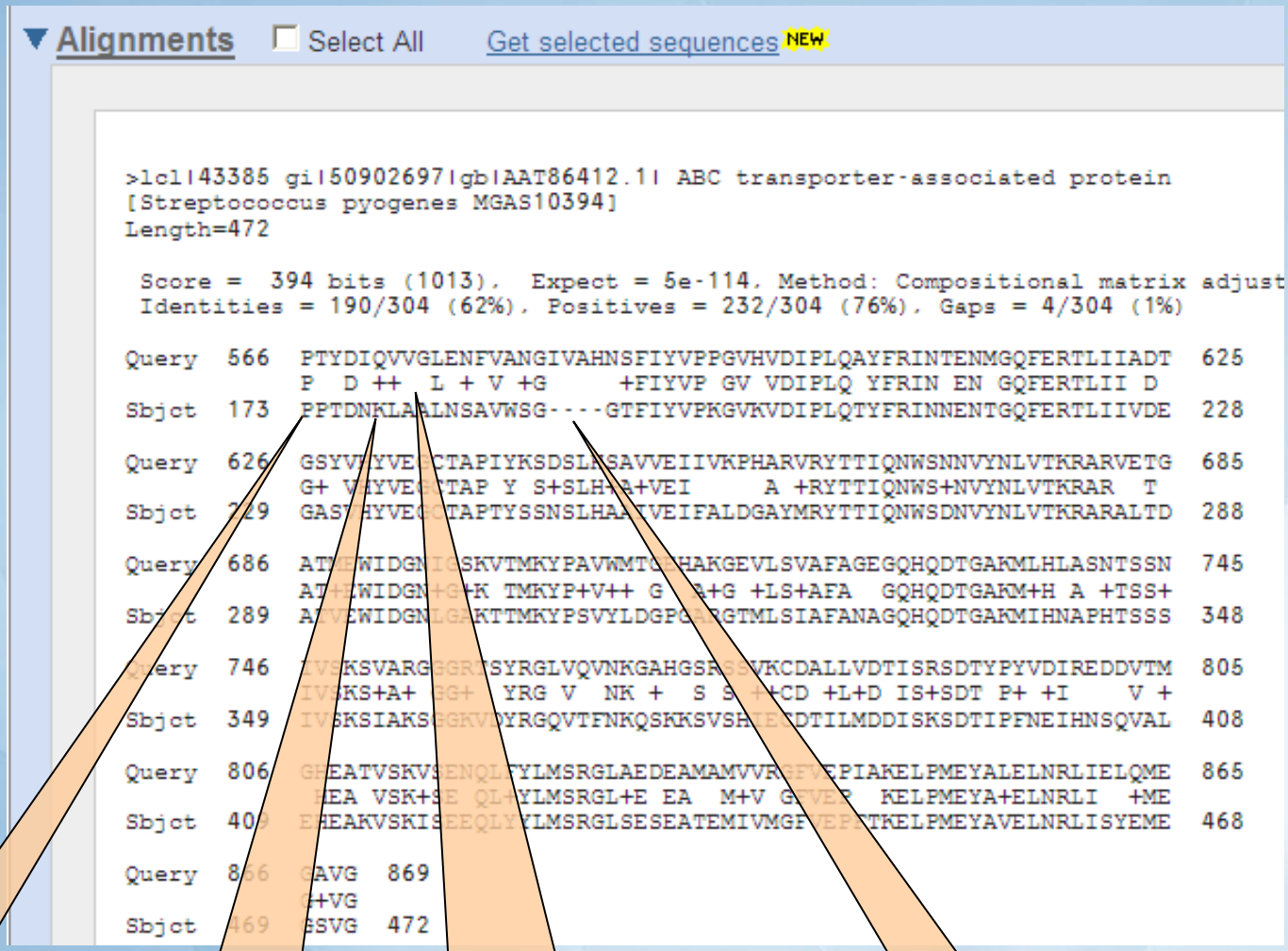
lower chance of random homology for AA

BLAST – bl2seq




Sequence alignment - Wi
Multiple Sequence Alignm
COBALT:Multiple Alignm
www.ncbi.nlm.nih.gov/tools/cobalt/cobalt.cgi?link_loc=BlastHomeLink
COBALT
Constraint-based Multiple Alignment Tool
My NCBI
[Sign In] [Register]
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

results



Alignments Select All Get selected sequences NEW

```
>lcl|43385 gi|50902697|gb|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 ++  L + V +G      +FIYVP GV VDIPLQ YFRIN EN GQFERTLII D
Sbjct  173  PPTDNKLALNSAVWSG----GTFIYVPKGVKVDIPLQTYFRINNENTGQFERTLIIVDE  228

Query  626  GSYVHYVEGCTAPIYKSDSLHSAVVEIIVKPHARVRYTTIQNWSNNVYNLVTKRARVETG  685
            G+ V HYVEGCTAP Y S+SLH+A+VEI     A +RYTTIQNWS+NVYNLVTKRAR  T
Sbjct  229  GASVHYVEGCTAPTYSSNSLHAAIVEIFALDGAYMRYTTIQNWSDNVYNLVTKRARALTD  288

Query  686  ATLEWIDGNIGSKVTMKYPAVWMTGHAKGEVLSVAFAGEGQHQDTGAKMLHLASNTSSN  745
            AT+EWIDGN+G+K TMKYP+V++ G  A+G +LS+AFA  GQHQDTGAKM+H A +TSS+
Sbjct  289  ATVEWIDGNLGAKTTMKYPSVYLDGPGAGTMLSIAFANAGQHQDTGAKMIHNAPHTSSS  348

Query  746  IVSKSVARGGGRISYRGLVQVNKGAHGSRSSVKCDALLVDTISRSDTYPYVDIREDDVTM  805
            IVSKS+A+ GG+  YRG V  NK +  S S  +CD +L+D IS+SDT P+ +I    V +
Sbjct  349  IVSKSIAKSGGKVDYRGQVTFNKQSKKSVSHIECDTILMDDISKSDTIPFNEIHNSQVAL  408

Query  806  GHEATVSKVSENQLTYLMSRGLAEDEAMAMVVRGFVEPIAKELPMEYALELNRLIELQME  865
             HEA VSK+SE QL+YLMSRGL+E EA  M+V G VEP  KELPMEYA+ELNRLI  +ME
Sbjct  409  EHEAKVSKISEEQLYYLMSRGLSESEATEMIVMGFVEPFTKELPMEYAVELNRLISYEME  468

Query  866  GAVG  869
            G+VG
Sbjct  469  GSVG  472
```

Match

Similarity

Dissimilarity

Gaps

that is, the substitution of amino acids whose side chains have similar biochemical properties

MSA input: multiple sequence Fasta file

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

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

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

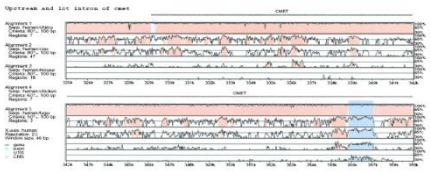
>gi|6715607|ref|NP_000175.1| G-gamma globin [Homo sapiens]
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLT
SLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTGVAS
ALSSRYH

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

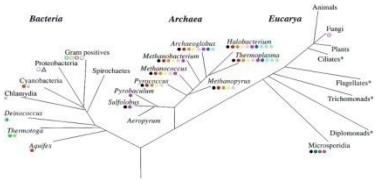
>gi|4885397|ref|NP_005323.1| hemoglobin, zeta [Homo sapiens]
MSLTKTERTIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYFPHFDLHPGSAQLRAHGSKVVAAVGDA
VKSIDDIGGALSKLSELHAYILRVDPVNFKLLSHCLLVTLAARFPADFTAEAHAAWDKFLSVVSSVLTEK
YR

Central role of multiple alignments

Comparative genomics

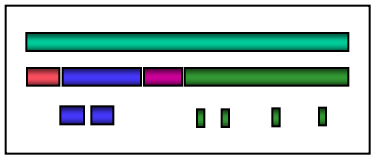


Phylogenetic studies

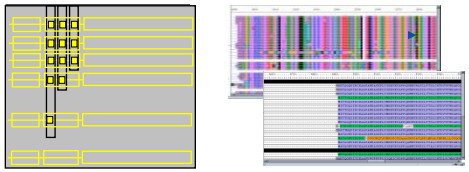


Hierarchical function annotation:

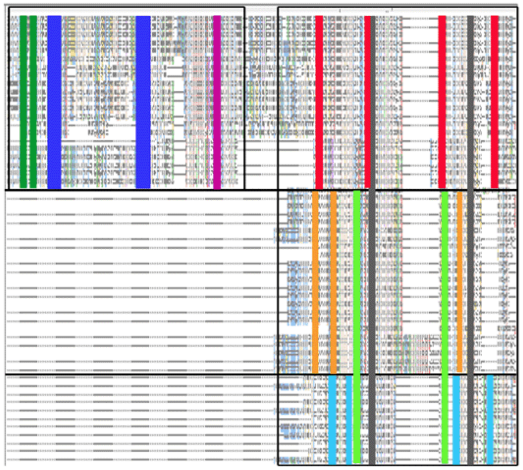
homologs, domains, motifs



Gene identification, validation

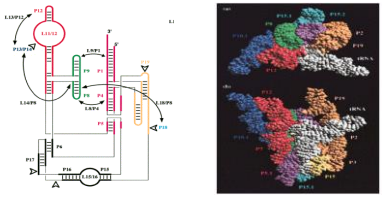


Multiple alignment

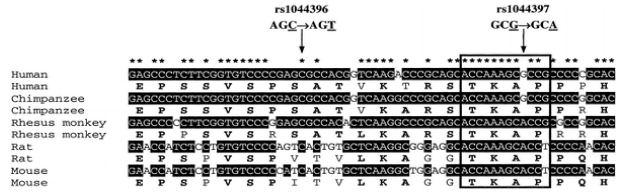


Structure comparison, modelling

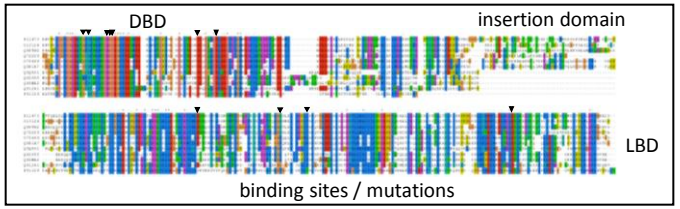
RNA sequence, structure, function



Human genetics, SNPs



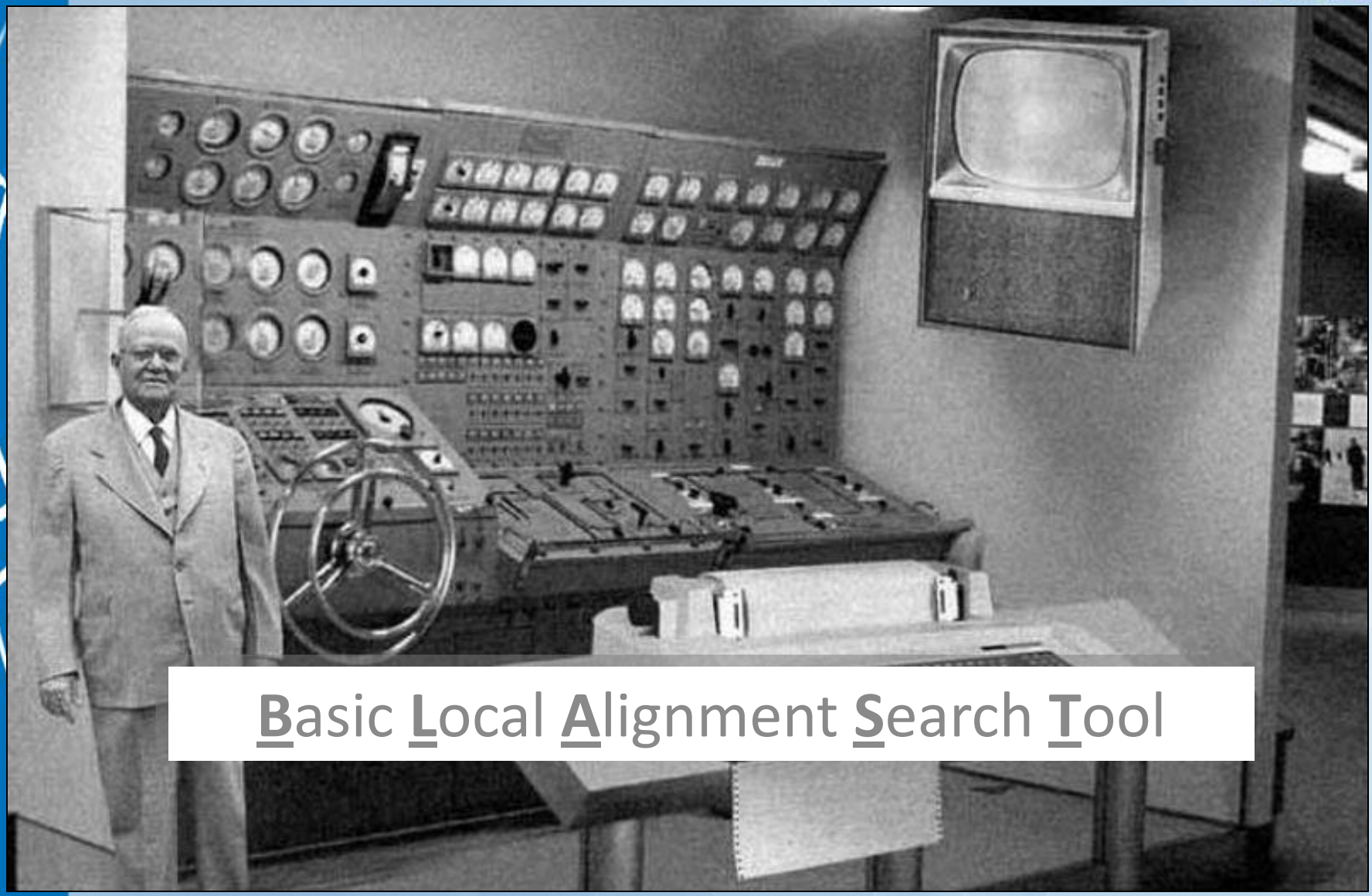
Therapeutics, drug design



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 **L**ocal **A**lignment **S**earch **T**ool