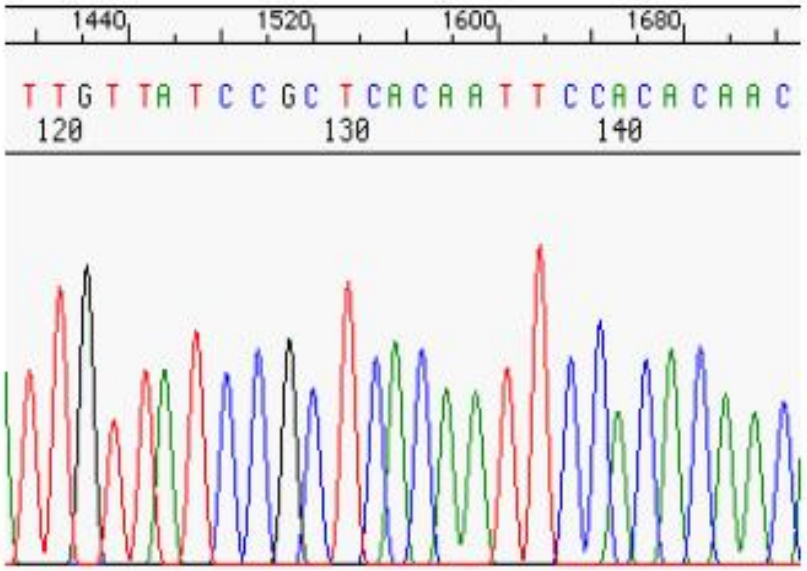
Interpretation of DNA sequencing results

Ameer Effat M. Elfarash

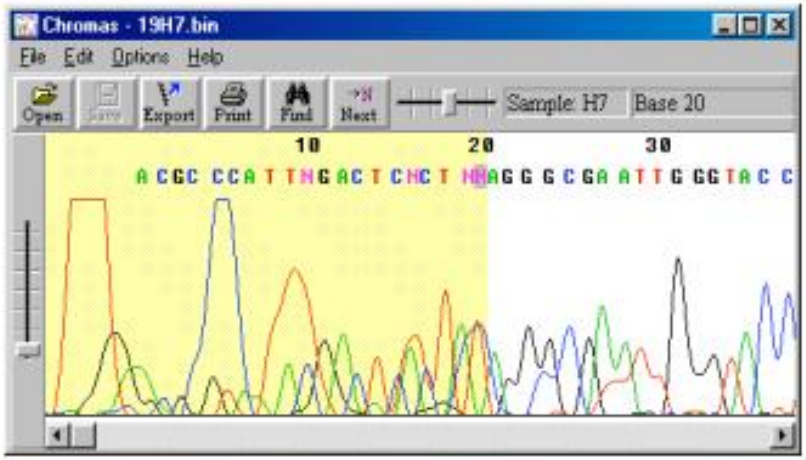
Dept. of Genetics
Fac. of Agriculture, Assiut Univ.
aelfarash@aun.edu.eg



>GXP_210035 loc=GXL_175098|sym=FAM149A|taxid=9606|spec=Homo sapiens|chr=4|ctg=NC_000004|str=(+)|start=187065495|end=187066181|len=687|comm=Promoter Region

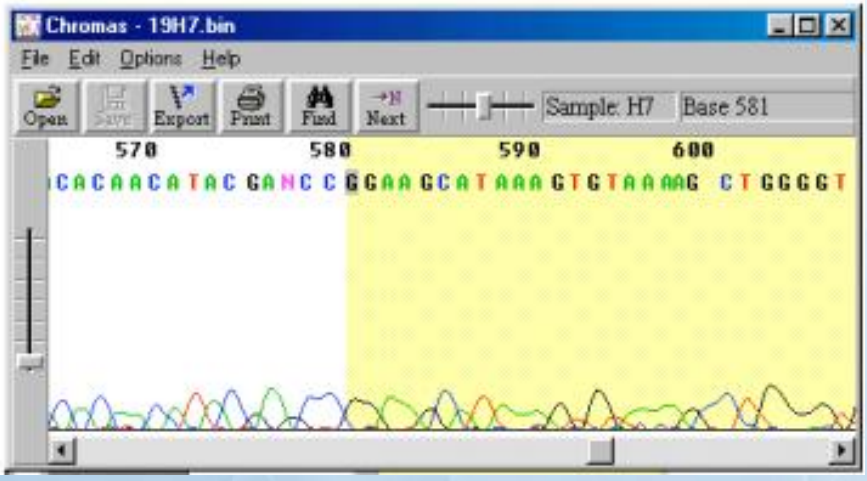
GGACGGGCGTGGAAGGGTCCACGTCTTTAGTRTGCATGCTTAGATCTAGCGTTCCTGTTGATGGAGTAATGGTTCTGCA
TTGACCAGATCCGGGGCTTCATTTTTTAAACCTCATTCGTCCACTCCCCACCCCAGCCTGGTGTGCGCACCCTTTGATGG
GGCGGGGATAGGCGAGATGGTCCTGTGGTTCTCTGCCTTCTTCTGGTGAATTAAAATCCGATTTGGAAGAGAGAAGGGCA
GCCAGCACCAGTATGCACAGCCCCCGGCCCCAGAGACCCGGGAAGGAGTAGGGAGCCCGGGCCGTGCGCGGAGGAGTGGC
CGCTGGGTTGGAAACCCGGCCCGGCAGGGAGCGGGGAAGGCGCGCTTTCCCGGAGGTCGGCGCGGGGCGGGGCCGGGGC
CGGGGCCCGGAGCGGGGATGGGCGGGCGCAGCCGGGATTAGCTGGCGGGCGAGGGCGCAGCGCAGGGAGGAGGAGGGGAG
GCGGCGCCGGCGCGGGCGGGGCGGAGGATCTGGAGAGGGAAGGGGCGTGCGAGCCCCGCGGACCCCGGGCGCGCCCGGGC
CGCCTGAGCTGGGCCAGCCGCGCGGCGGGCGCGGGCGCGGGCGCGGGCGCGGGCGCGGGCGCGGGCGGGTGGGGAGCCCCAGCCCC
GGGGCCGCGGGGGCGCGTGACCGGCTGTCTGCGTGGGGCCCGCGCGC

Good sequence generally begins roughly around base 20.



Beginning of Sequence

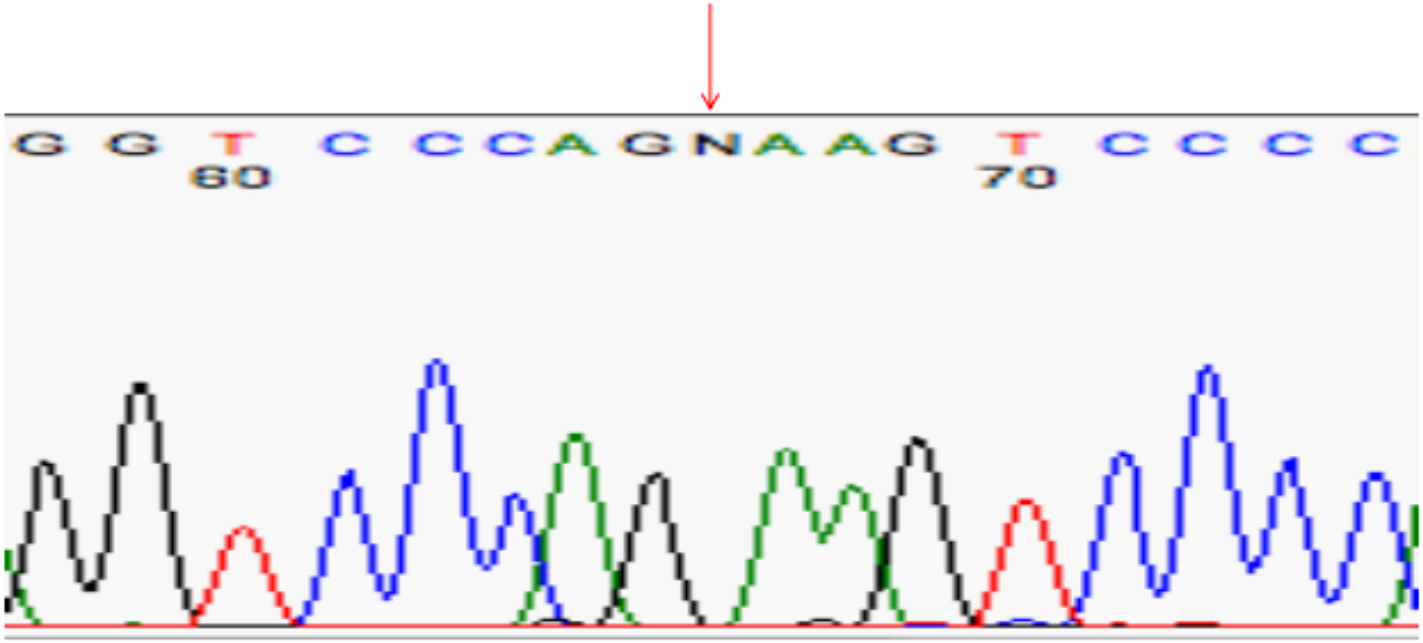
End of sequence



4- Mis-Called

(b) Mis-call a nucleotide:

Sometimes the computer will mis-call a nucleotide when a human could do better. Most often, this occurs when the basecaller calls a specific nucleotide, when the peak really was ambiguous and should have been called as 'N'.

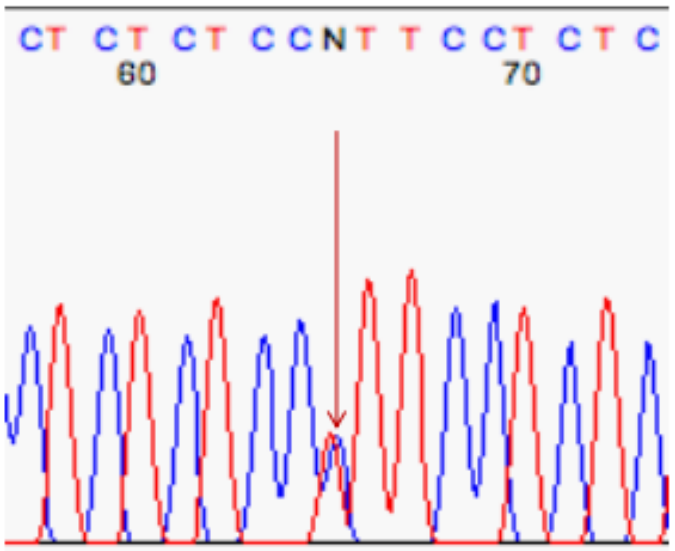


5- Heterozygous (double) peaks:

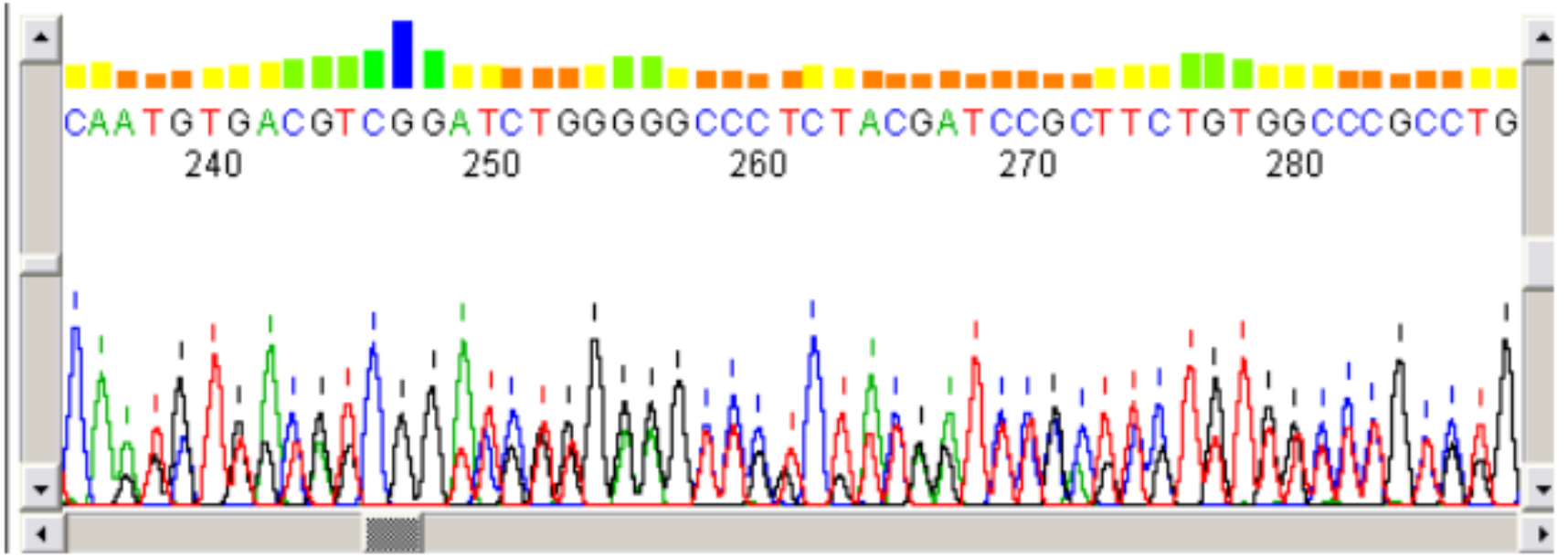
A single peak position within a trace may have but two peaks of different colors instead of just one.

Note that the base caller may list that base position as an 'N', or it may simply call the larger of the two peaks.

Here's a great example of a PCR **single-nucleotide polymorphism (SNP).**



11- DNA contamination:

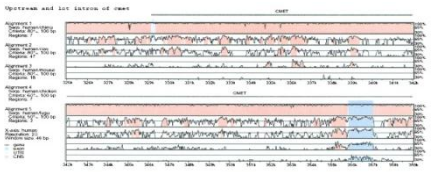


```
MVNLTSDEKTAVLALWNKVDVEDCGGEALGRLLVVYPWTQRFFE...
||  ||   ||||| ||| ||  ||    ||||||||||||||||||||
MVHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFE...
```

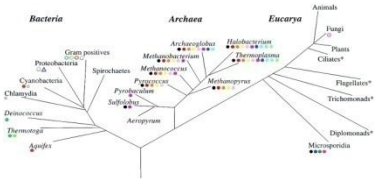
Alignment: Comparing two (pairwise) or more (multiple) sequences. Searching for a series of identical or similar characters in the sequences.

Central role of multiple alignments

Comparative genomics

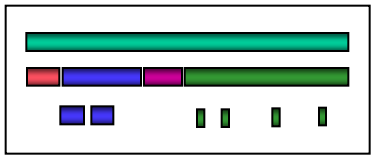


Phylogenetic studies

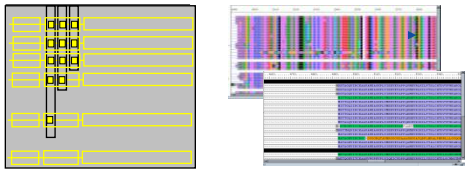


Hierarchical function annotation:

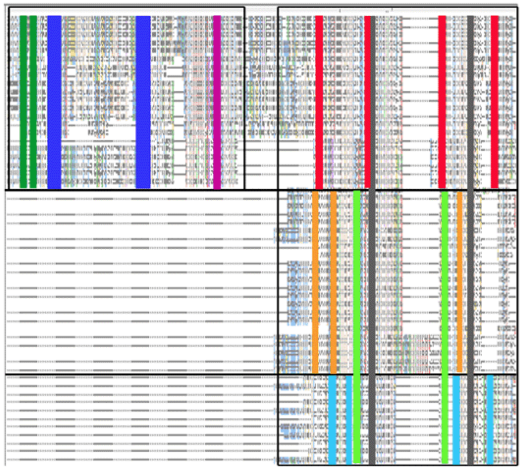
homologs, domains, motifs



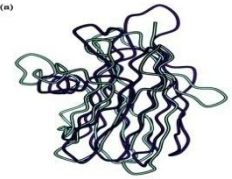
Gene identification, validation



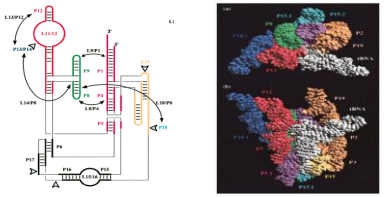
Multiple alignment



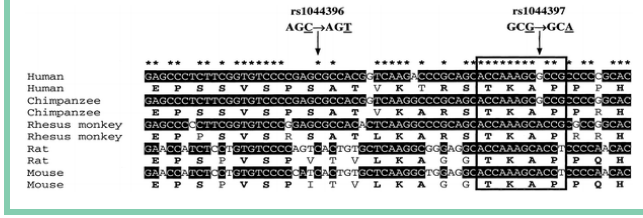
Structure comparison, modelling



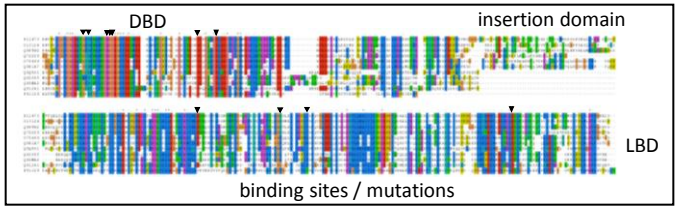
RNA sequence, structure, function



Human genetics, SNPs



Therapeutics, drug design



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?

Three types of nucleotide changes:

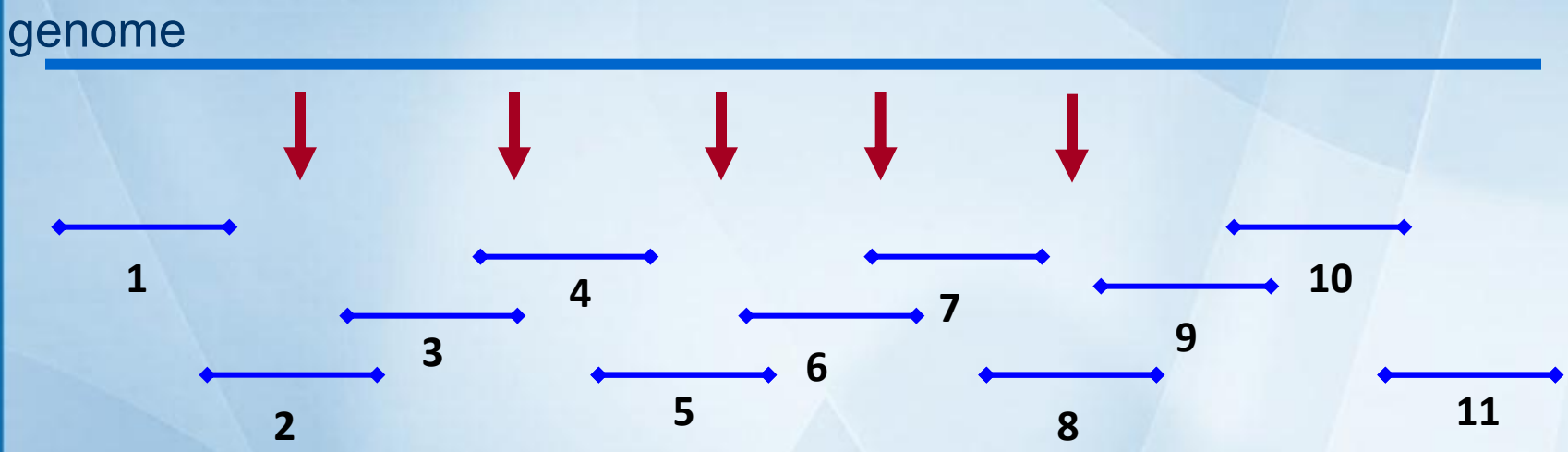
Substitution – a replacement of one (or more) sequence characters by another: AAGA → AACA

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

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

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

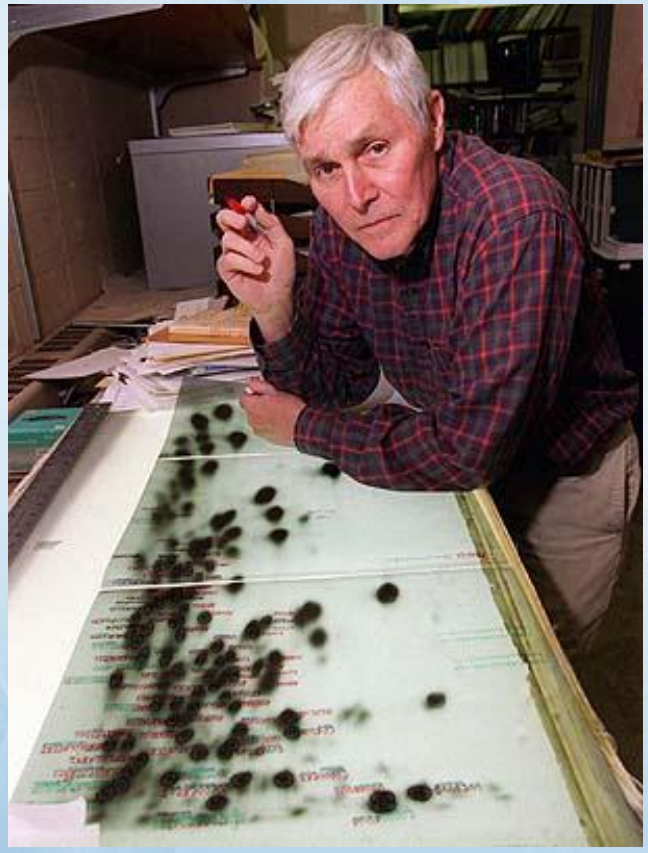
Why Align Sequences?

- Homology
  - Similar sequences may have a common ancestor

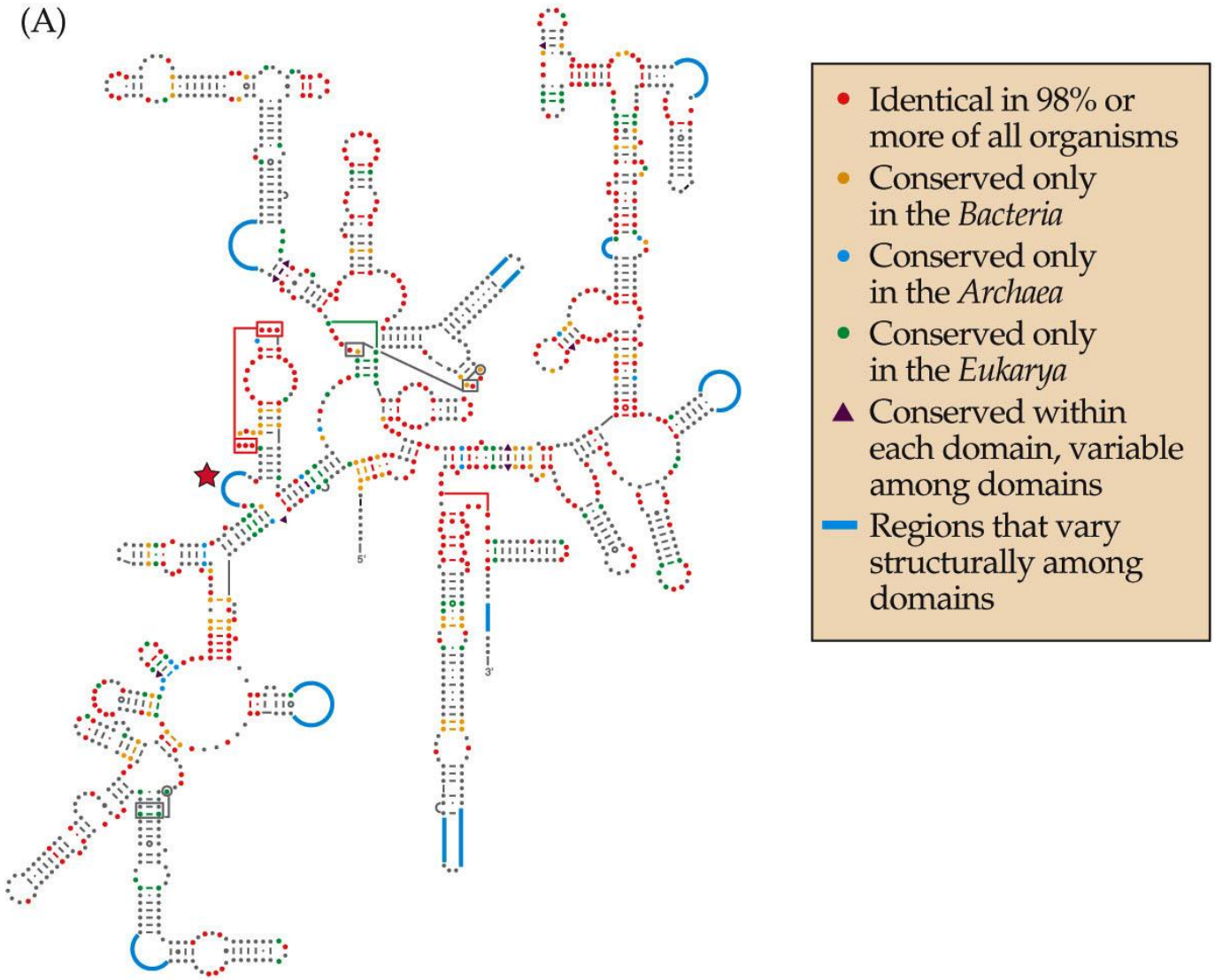
- In 1990 Carl Woese and colleagues proposed the Tree of Life, using a molecular phylogenetic approach.

- It is based on sequencing rRNA (16S and 18S), which all organisms share.

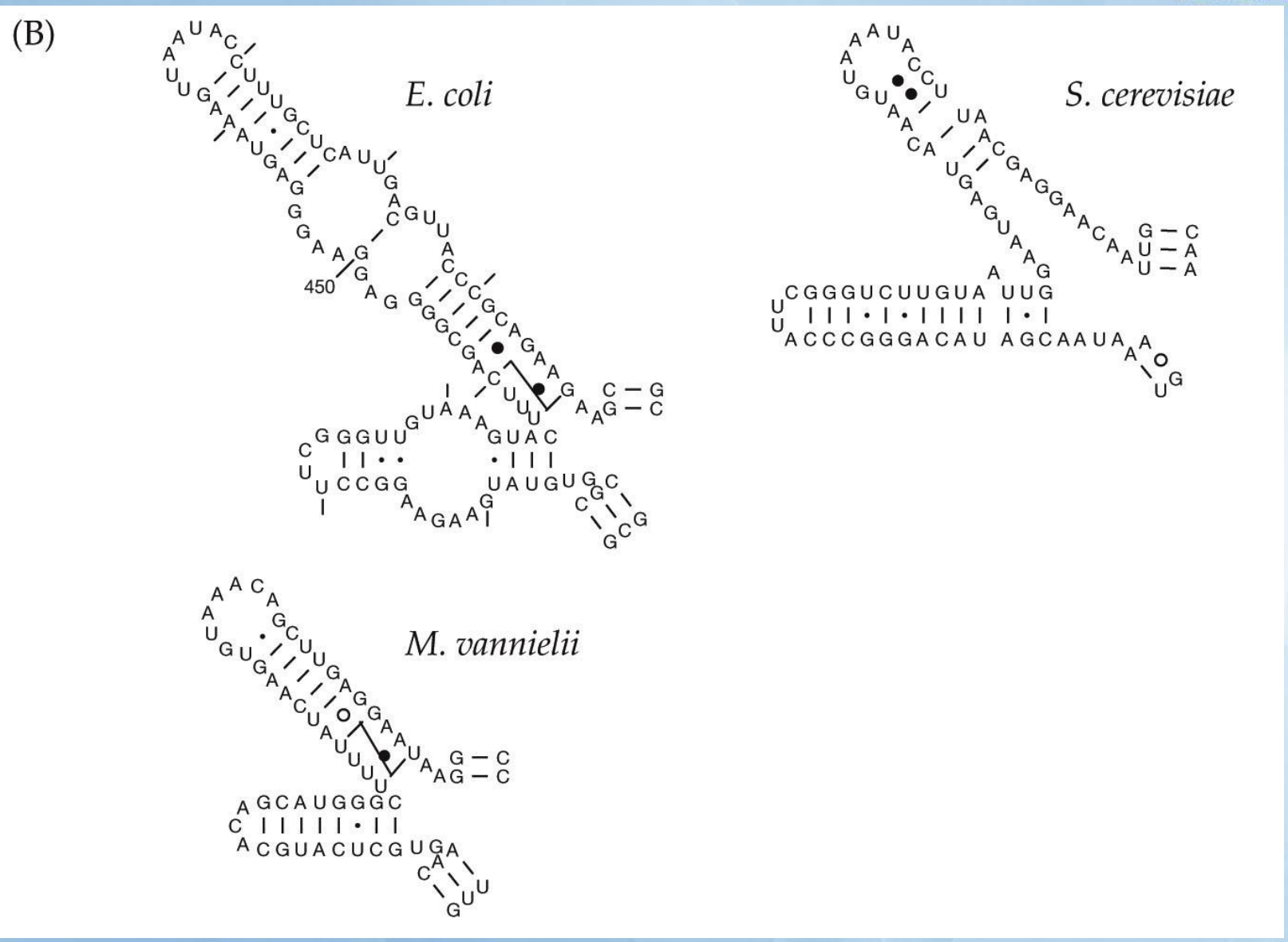
- All organisms were separated into three domains: *Bacteria*, *Archaea*, *Eukarya*.



16S RNA

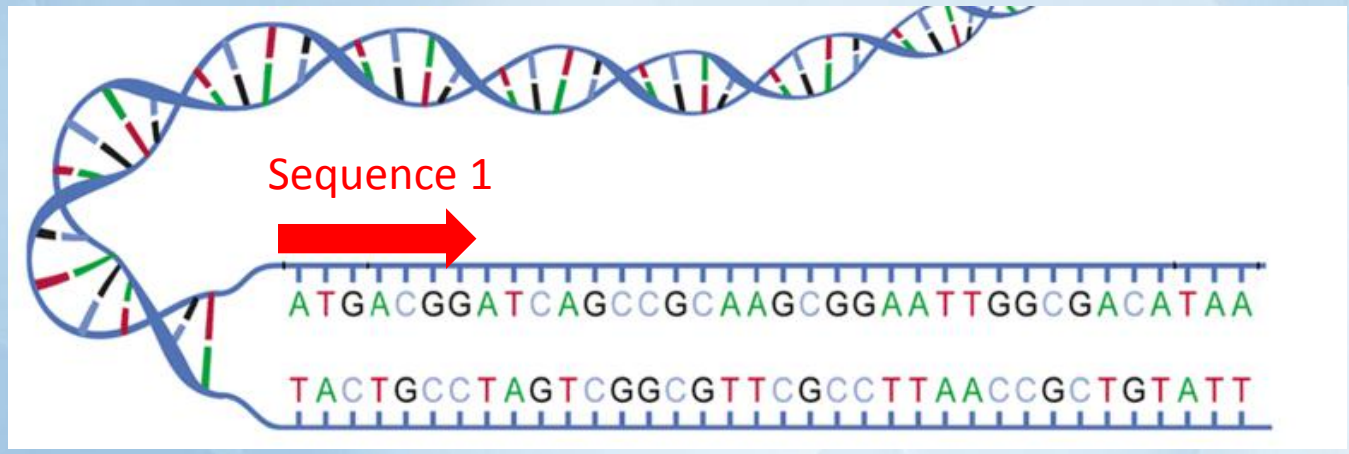


***Microbial Life 2e*, Figure 17.3 (Part 1)**

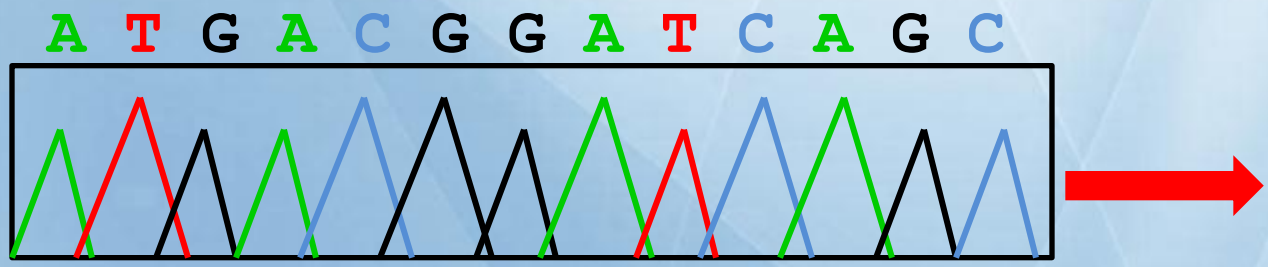

(B)
E. coli
450
S. cerevisiae
M. vannielii

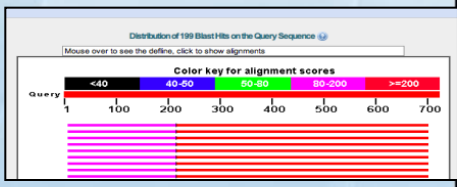
Why Align Sequences?

confirm insert sequence



Sequence :





Why Align Sequences?

Translation:

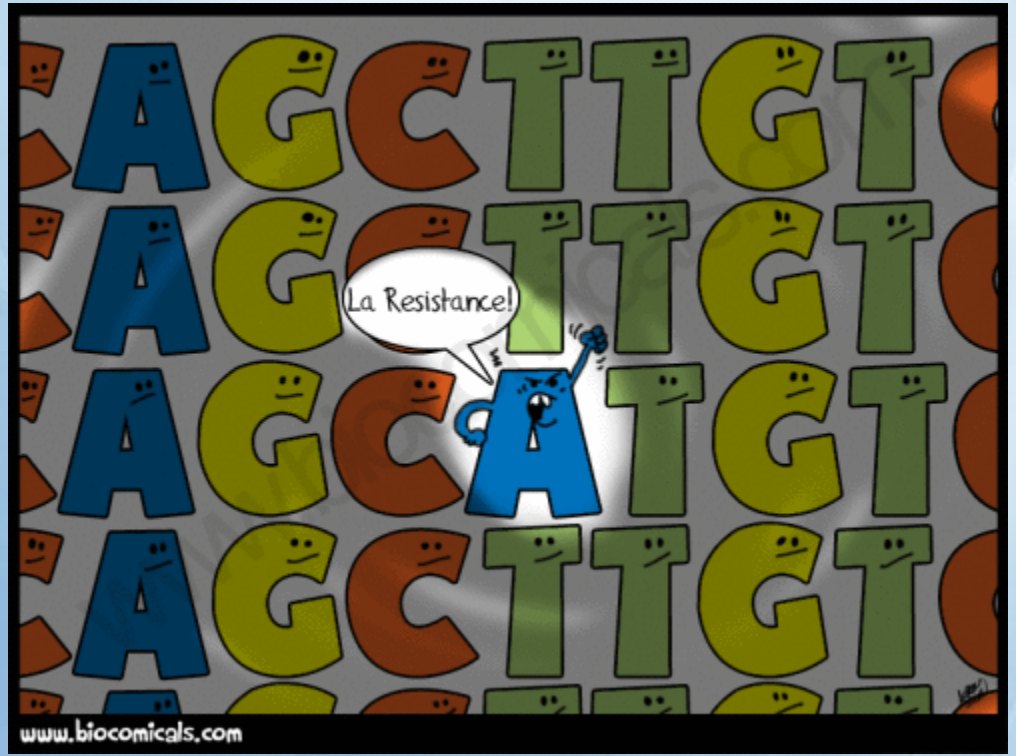
Each sequence must be translate to its amino acids (aa) by using

Expasy.translatesoftware

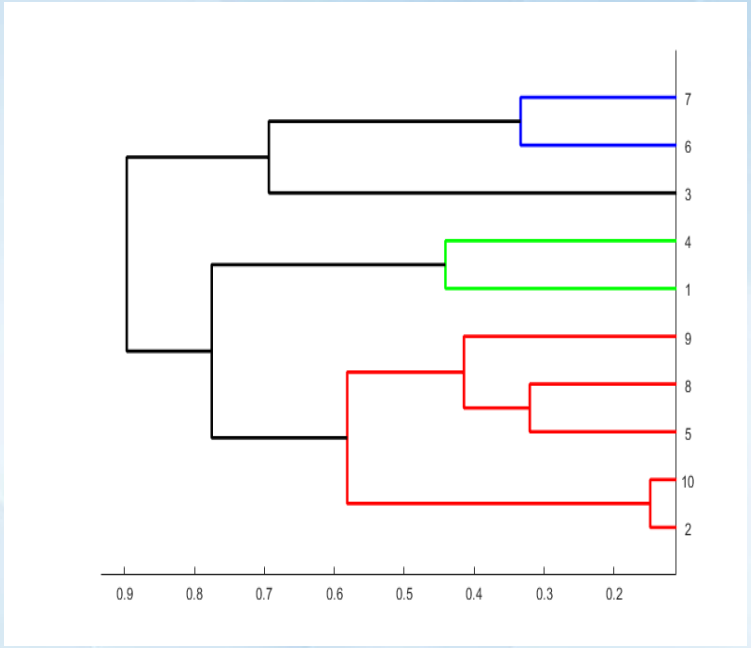
Why Align Sequences?

Comparing two sequences

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



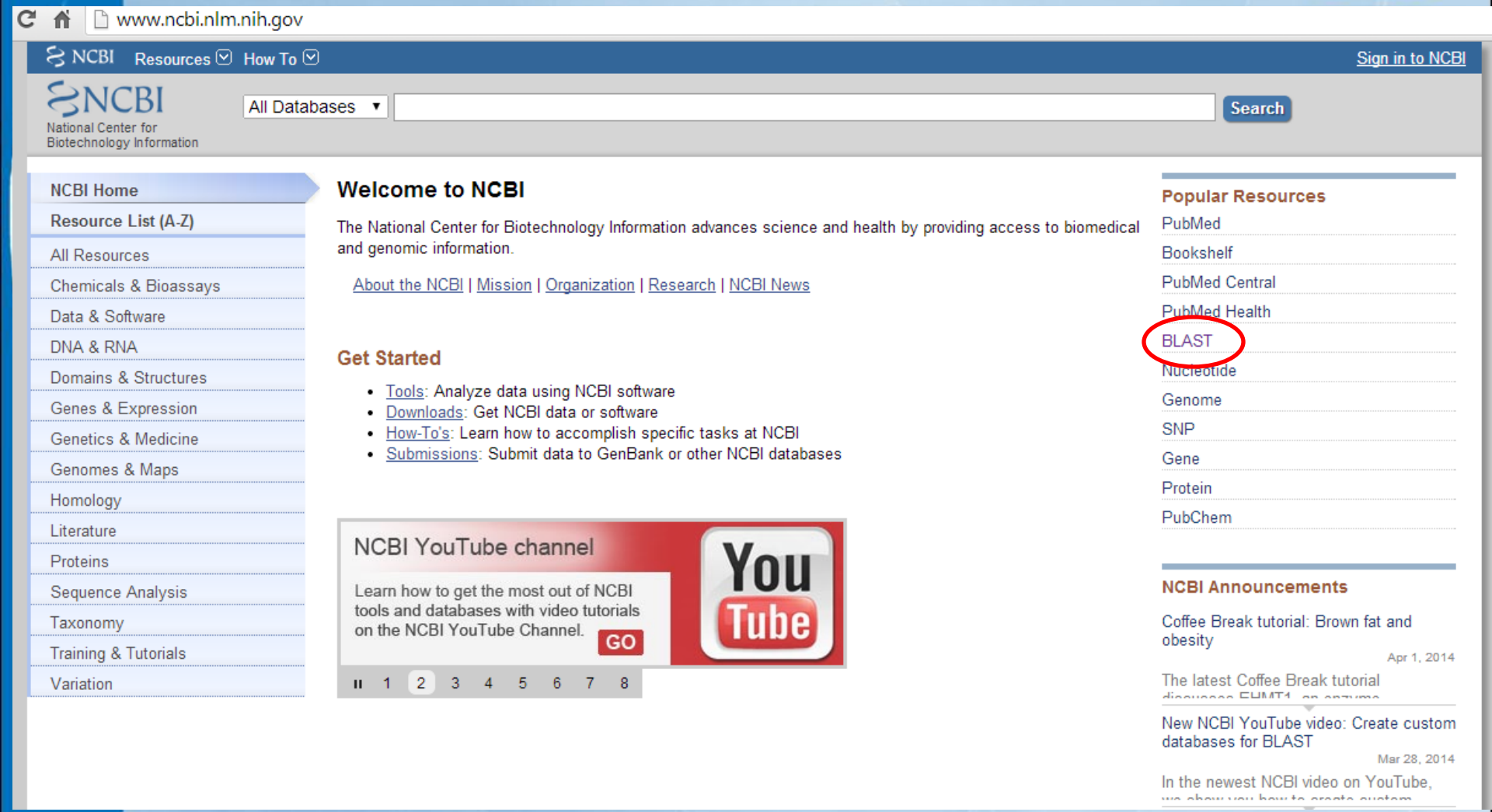
Why Align Sequences?



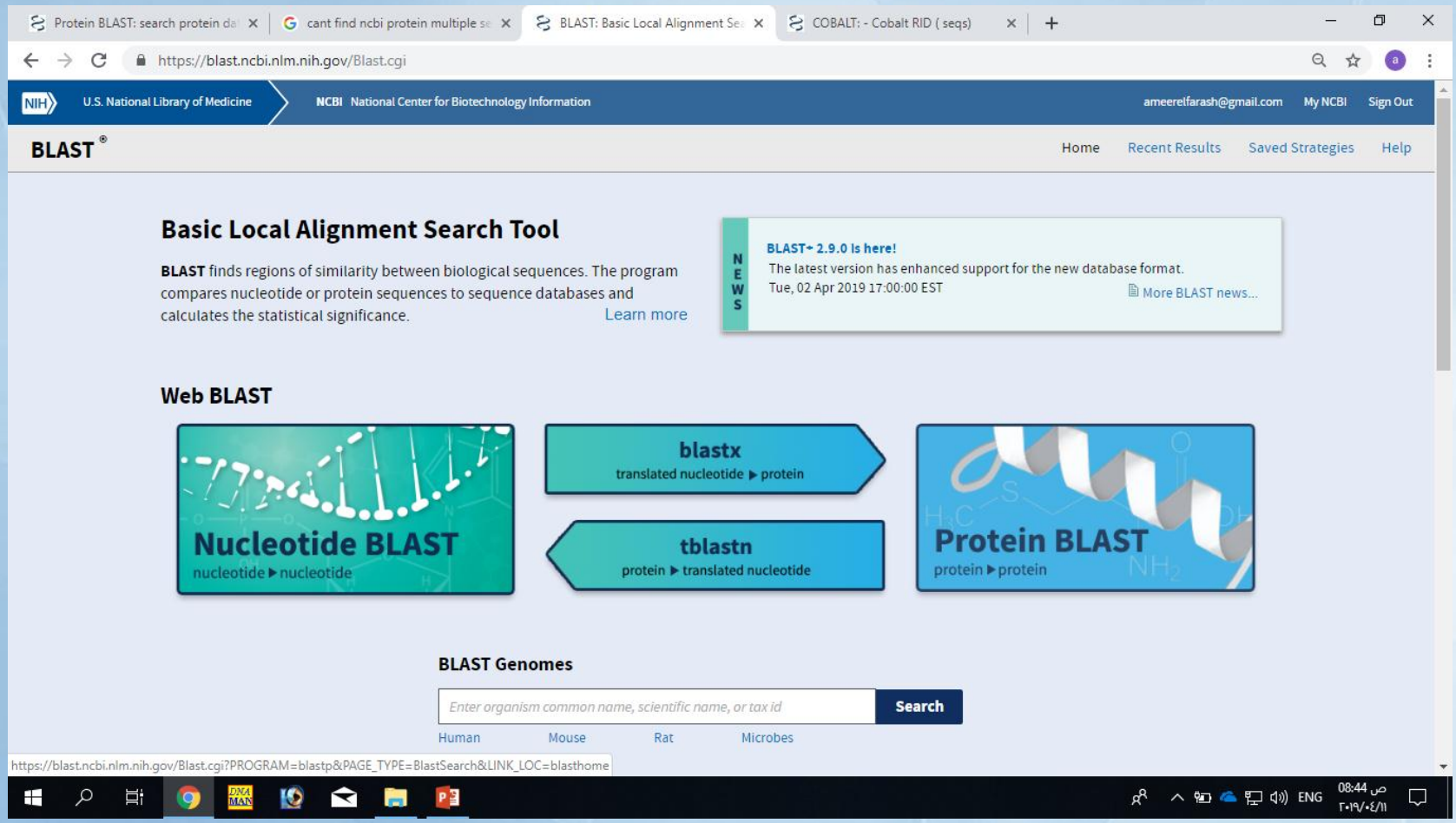
Dendrogram or phylogenetic tree

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



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

MSA input: multiple sequence Fasta file

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

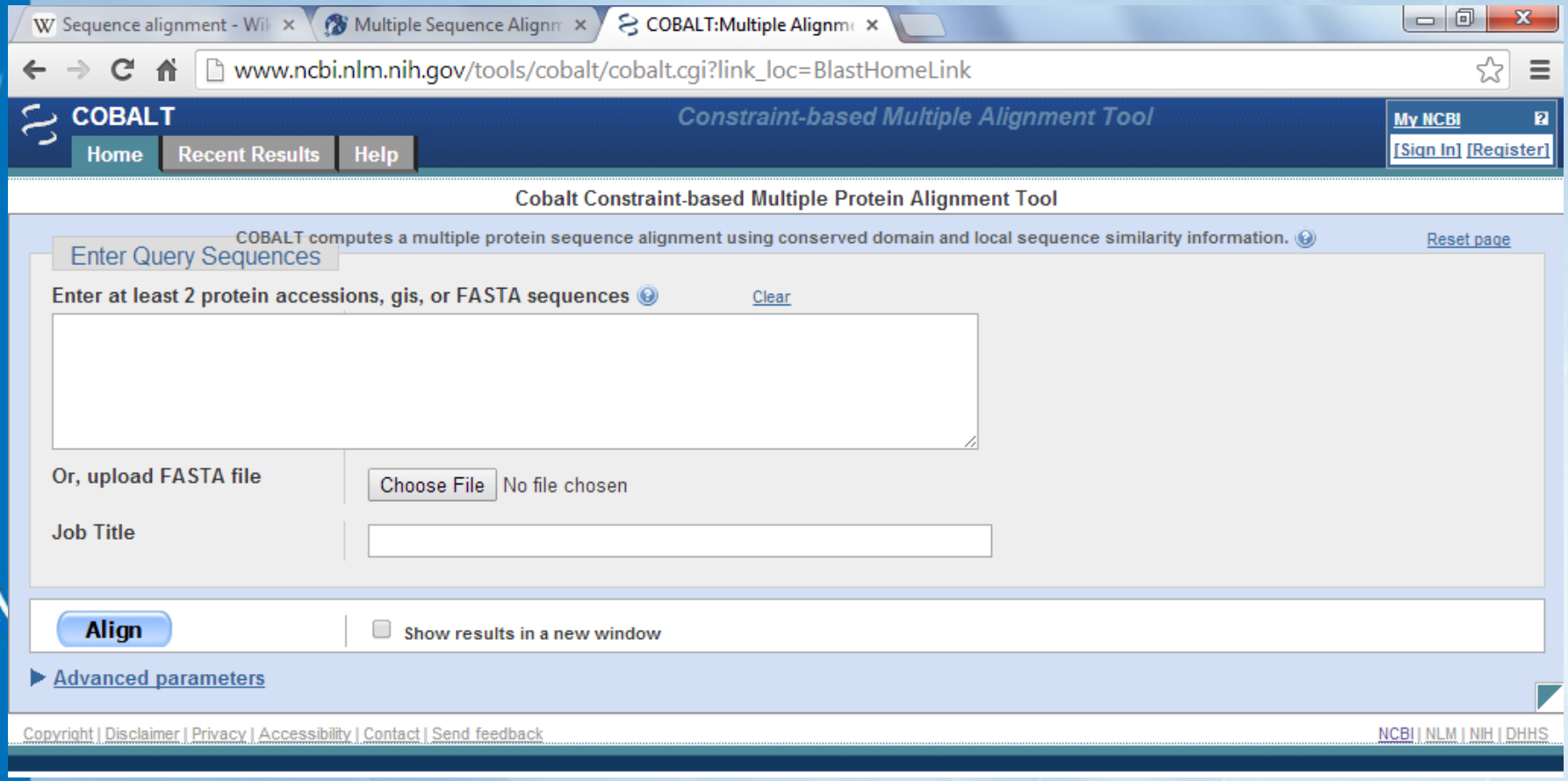
>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH

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

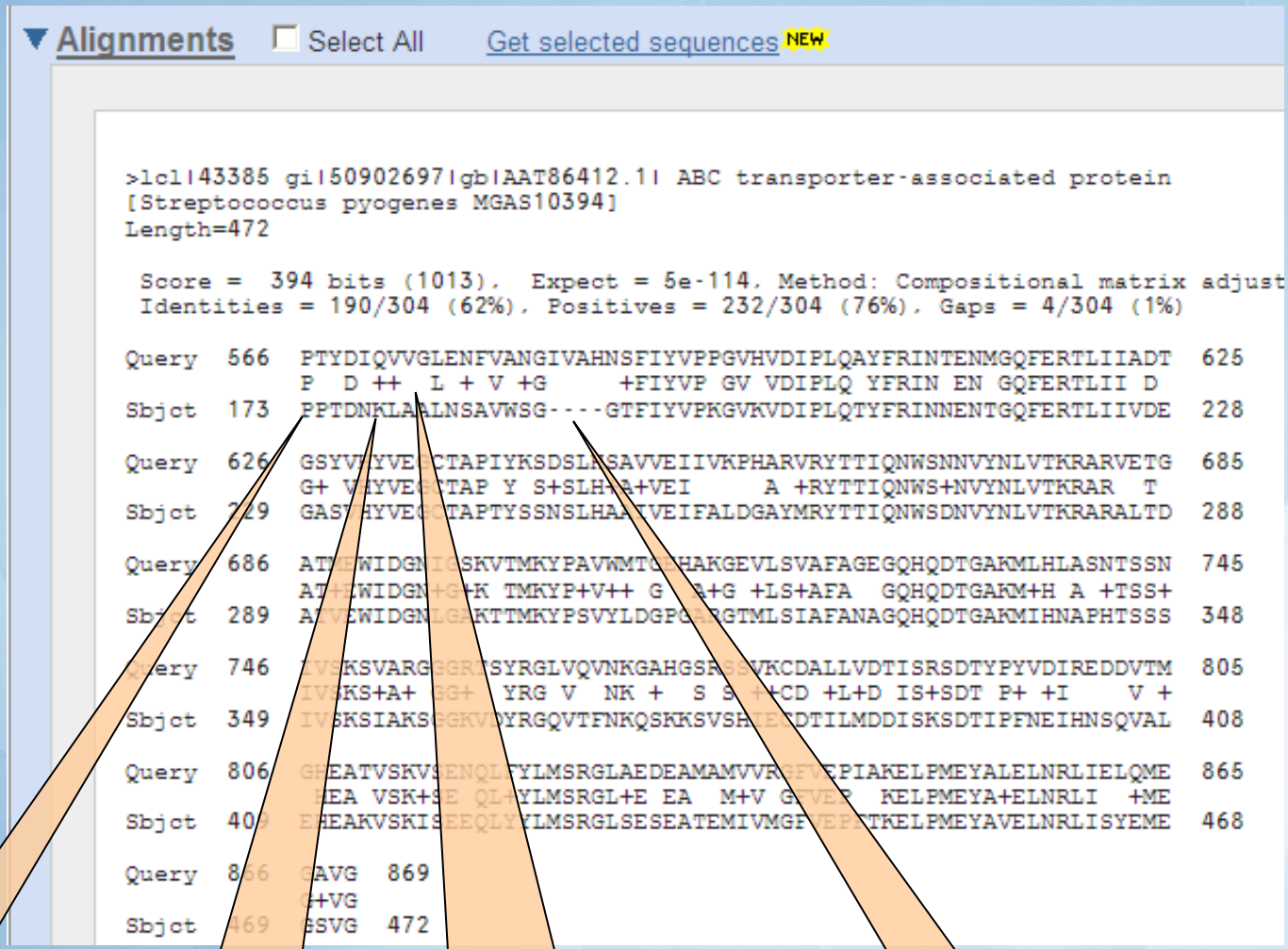
>gi|6715607|ref|NP_000175.1| G-gamma globin [Homo sapiens]
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLT
SLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTGVAS
ALSSRYH

>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


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  GSYVHYVEQCTAPIYKSDSLHSAVVEIIVKPHARVRYTTIQNWSNNVYNLVTKRARVETG  685
            G+ V HYVEQ TAP Y S+SLH+A+VEI     A +RYTTIQNWS+NVYNLVTKRAR  T
Sbjct  229  GASVHYVEQCTAPTYSSNSLHAAIVEIFALDGAYMRYTTIQNWSDNVYNLVTKRARALTD  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