



H3ABioNet

Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics Online Course: IBT

Multiple Sequence Alignment

Building Multiple Sequence Alignment

Lec1 Building a Multiple Sequence Alignment

Learning Outcomes

- 1- Understanding Why multiple sequence alignment is useful for scientists
- 2-Identifying situations where multiple alignments do not help
- 3-Main Criteria for Building a Multiple Sequence Alignment
- 4- Main Applications of Multiple Sequence Alignments
- 5-What are the kinds of sequences you're looking for?
- 6- Tips for Naming sequences
- 7- Tips for difficult MSA to interpret
- 8- Comparing sequences you cannot align



In the coming lectures we will learn

1- Gathering the sequences you need to make a multiple sequence alignment

2- Differences between some famous multiple sequence alignment programs

COBALT (Constraint-based Multiple Alignment Tool) New

ClustalW (everybody uses it),

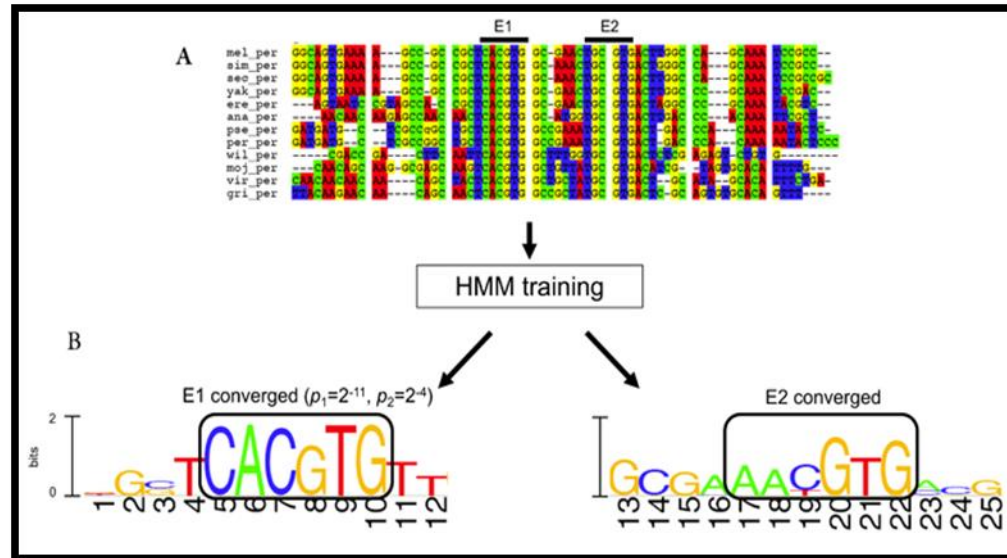
MUSCLE (very fast)

TCOFFEE (accurate and combine sequences and structures)

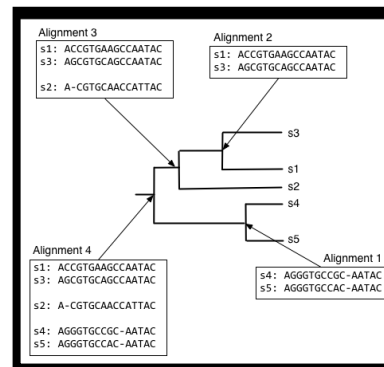
3- Creating and comparing multiple sequence alignments with -
Comparing sequences you cannot align

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Building a Multiple Sequence Alignment (1)

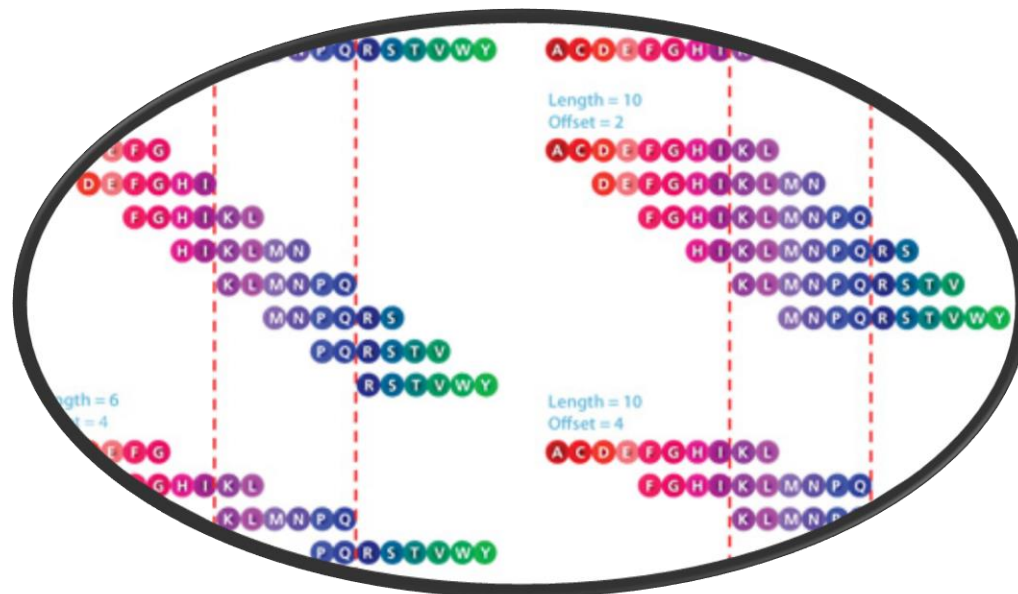


- “In many ways, multiple sequence alignments are to bioinformatics what *Swiss knives* are to *MacGyver*”
- Building multiple sequence alignments is far from an exact science
- In fact, it’s more art than science, requiring that you use **everything you know in bioinformatics and in biology.**”

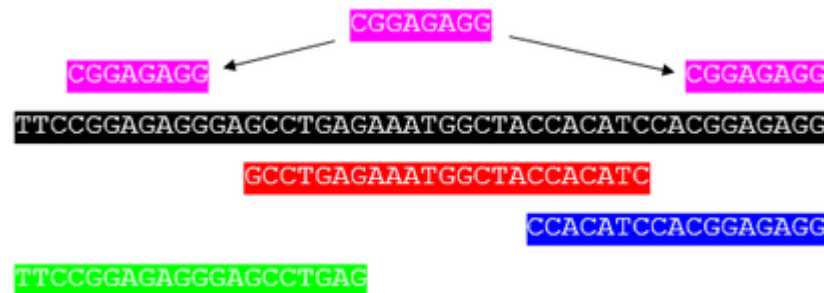


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Identifying situations where multiple alignments do not help

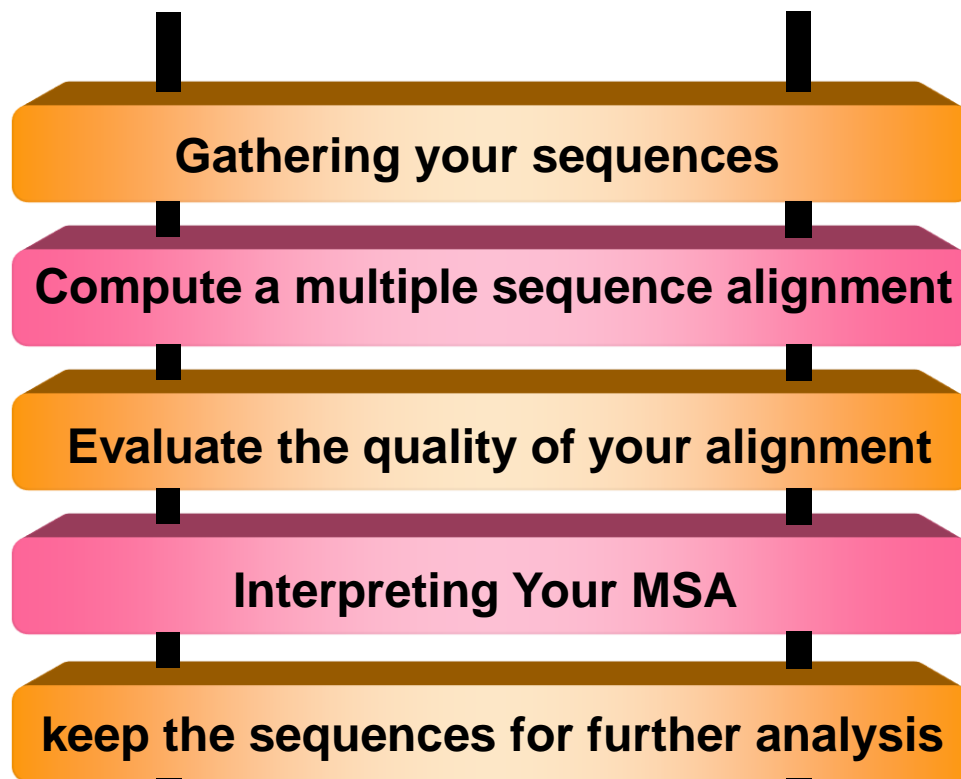


- Don't work well for **assembling the sequence** pieces in a sequencing project.
- if you want to turn an **EST cluster** into a gene sequence
- When the sequence you're interested in has **no homologue** in any of the sequence databases (in this case you can use functional criteria and conducting a pattern search).



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Building informative alignments



[Mansour A, Jaime A. Teixeira da Silva, Gábor Gyulai \(2009\) Assessment of molecular \(dis\)similarity: The role of multiple sequence alignments \(MSA\) programs in biological research. *Genes, genomes and genomics* 30-23 :\(1 eussl laicepS\)3 .Print ISSN \) \(0383-1749Bioinformatics SI.\(](#)

What we are looking for with MSA?

“The idea behind a multiple alignment is to put amino acids or nucleotides **in the same column** because they’re similar according to some criterion. You can use **four major criteria** to build a multiple alignment of sequences that all have different properties.”

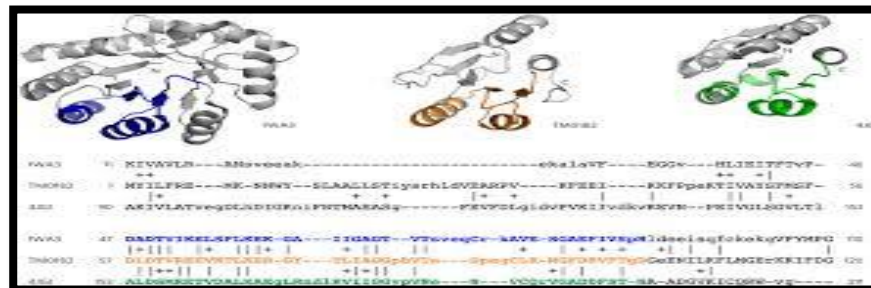
Scarites	C	T	T	A	G	A	T	C	G	T	A	C	C	A	A	-	-	-	A	A	T	A	T	T	A	C
Carenum	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	A	-	T	A	C	-	T	T	T	A	C
Pasimachus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	T	A	T	A	A	G	T	T	T	A	C
Pheropsophus	C	T	T	A	G	A	T	C	G	T	T	C	C	A	C	-	-	-	A	C	A	T	A	T	A	C
Brachinus armiger	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	T	C
Brachinus hirsutus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	A	C
Aptinus	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	C	A	A	T	T	A	C
Pseudomorpha	C	T	T	A	G	A	T	C	G	T	A	C	C	-	-	-	-	-	A	C	A	A	A	T	A	C

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Main Criteria for Building a Multiple Sequence Alignment

1- Structural similarity





Amino acids that play the same role in each structure are in the same column. **Structure-superposition** programs are the only ones that use this criterion.



Main Criteria for Building a Multiple Sequence Alignment

2- Evolutionary similarity

Amino acids or nucleotides related to the same amino acid (or nucleotide) in **the common ancestor** of all the sequences are put in the same column. No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.

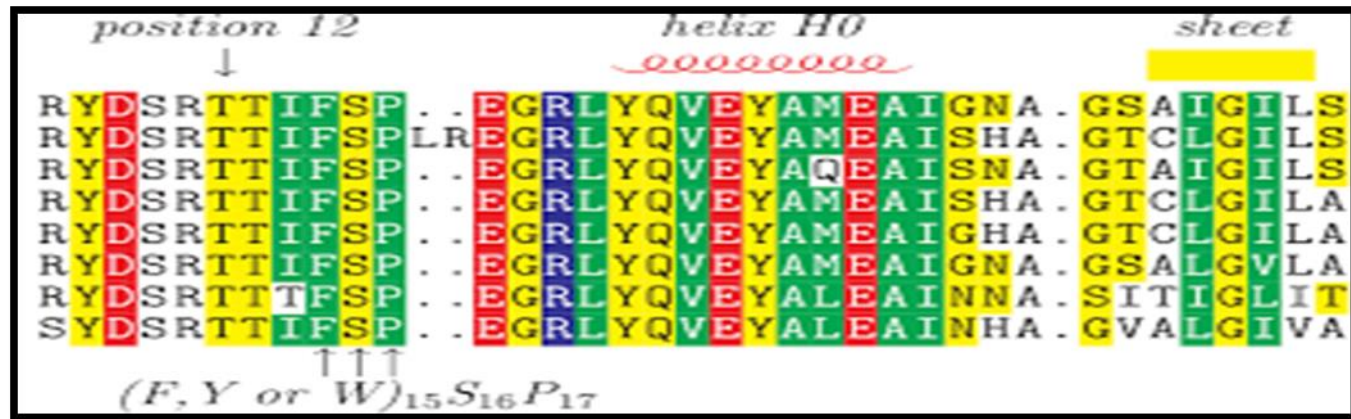
Organism	 CHIMP	 MOUSE	 CHICKEN	 FRUIT FLY
Gene Conservation with Humans (%)	99.5	88	75	60

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Main Criteria for Building a Multiple Sequence Alignment

3- Functional similarity

Amino acids or nucleotides with the **same function** are in the same column. No automatic program explicitly uses this criterion, but if the information is available, you can **force some programs to respect it** — or you can edit your alignment manually.

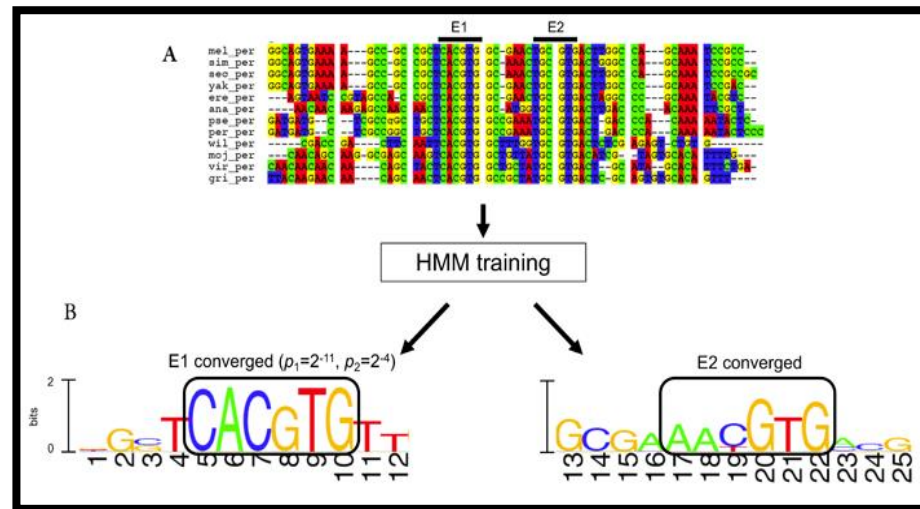


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Main Criteria for Building a Multiple Sequence Alignment

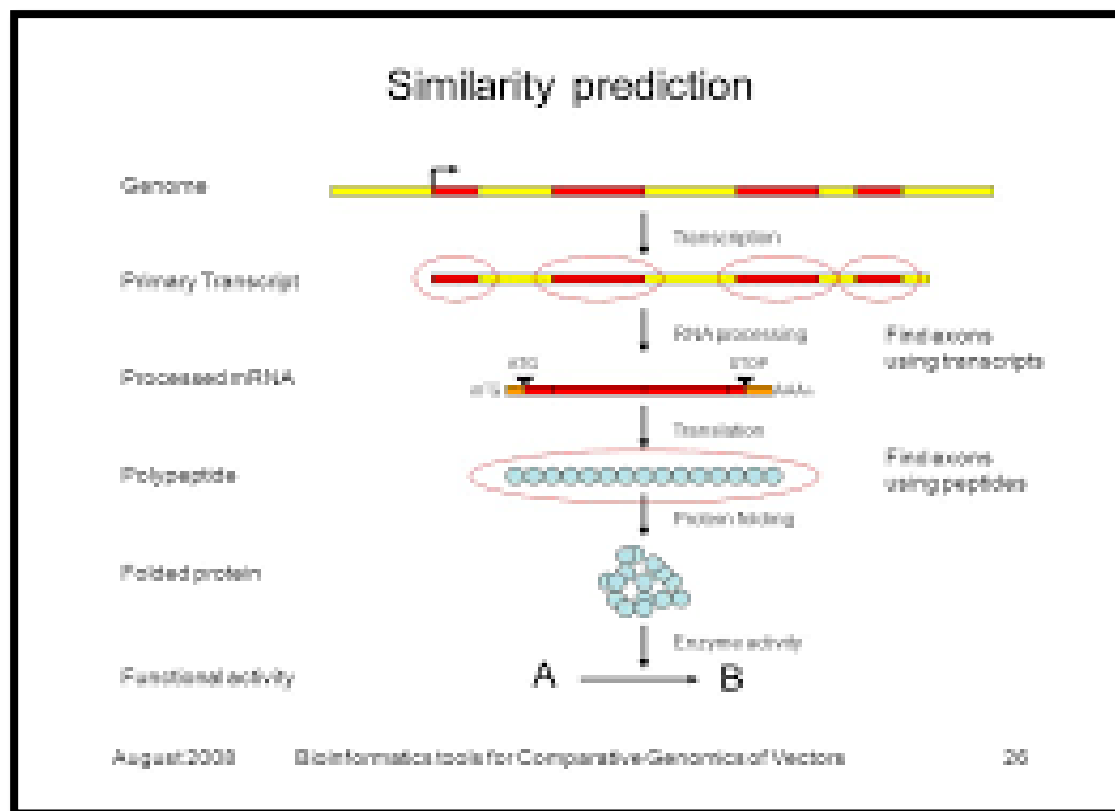
4- Sequence similarity

“Amino acids in the same column are those that yield an alignment with maximum similarity. Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, their **structural, evolutionary, and functional** similarities are **equivalent to sequence similarity**”.



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

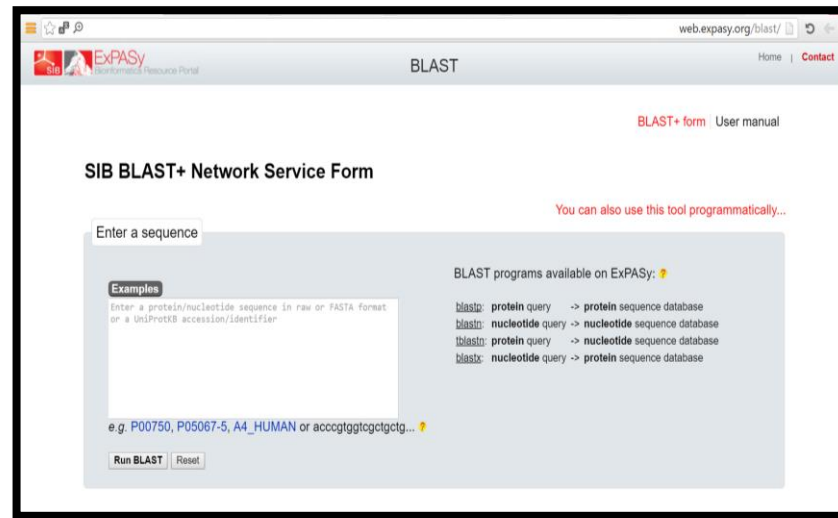
Main Applications of Multiple Sequence Alignments•



Main Applications of Multiple Sequence Alignments•

Extrapolation

“A good multiple alignment can help convince you that an **uncharacterized sequence is really a member of a protein family**. Alignments that include Swiss-Prot sequences are the most informative. Use the **ExPASyBLAST** server (at www.expasy.ch/tools/blast/) to gather and align them”.

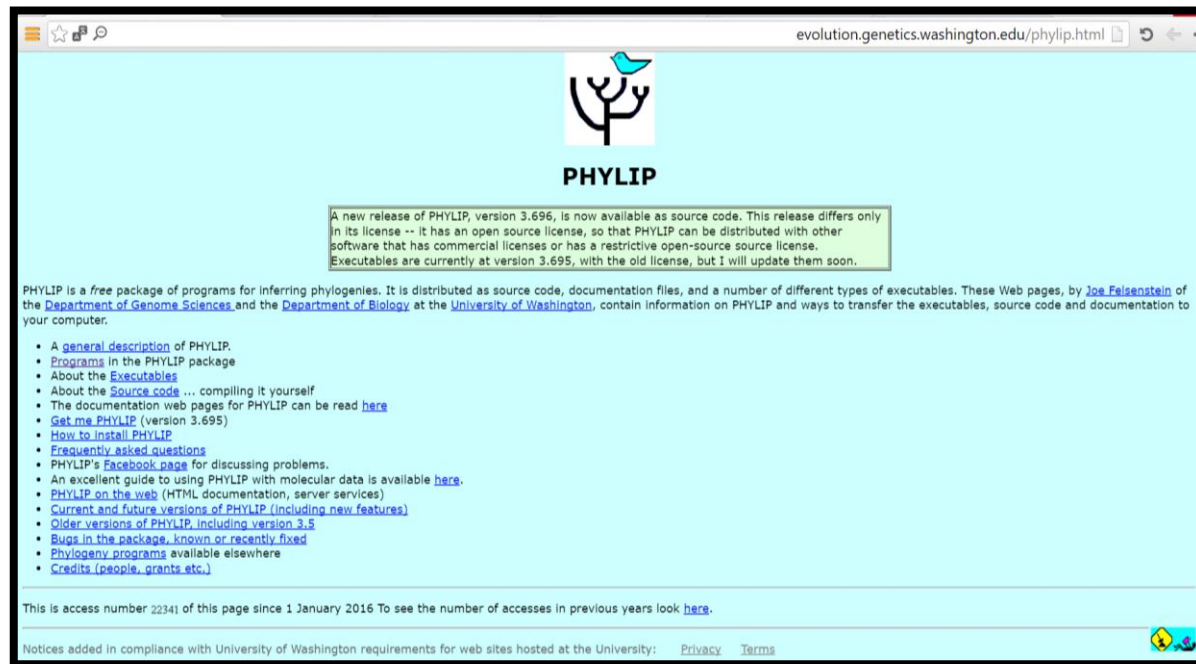


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Main Applications of Multiple Sequence Alignments

Phylogenetic Analysis

“If you carefully choose the sequences you include in your analysis multiple alignment, you can *reconstruct the history of these proteins*. Use the Pasteur Phylip server at bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html.”

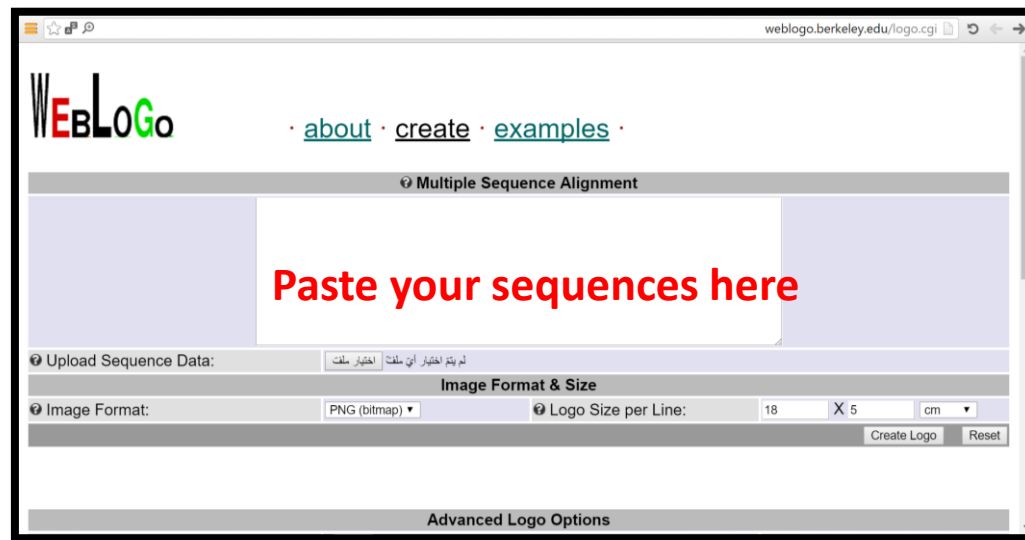


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Main Applications of Multiple Sequence Alignments

Pattern identification

“By discovering very conserved positions, you can *identify a identification region that is characteristic of a function* (in proteins or in nucleic-acid sequences). Use the **Weblogo server** <http://weblogo.berkeley.edu/logo.cgi>”



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Main Applications of Multiple Sequence Alignments•

Domain identification

“It is possible to turn a multiple sequence alignment into a profile that describes a **protein family or a protein domain** (PSSM). You can use this profile to scan databases for new members of the family. Use **PROSITE** (<http://prosite.expasy.org/>)”

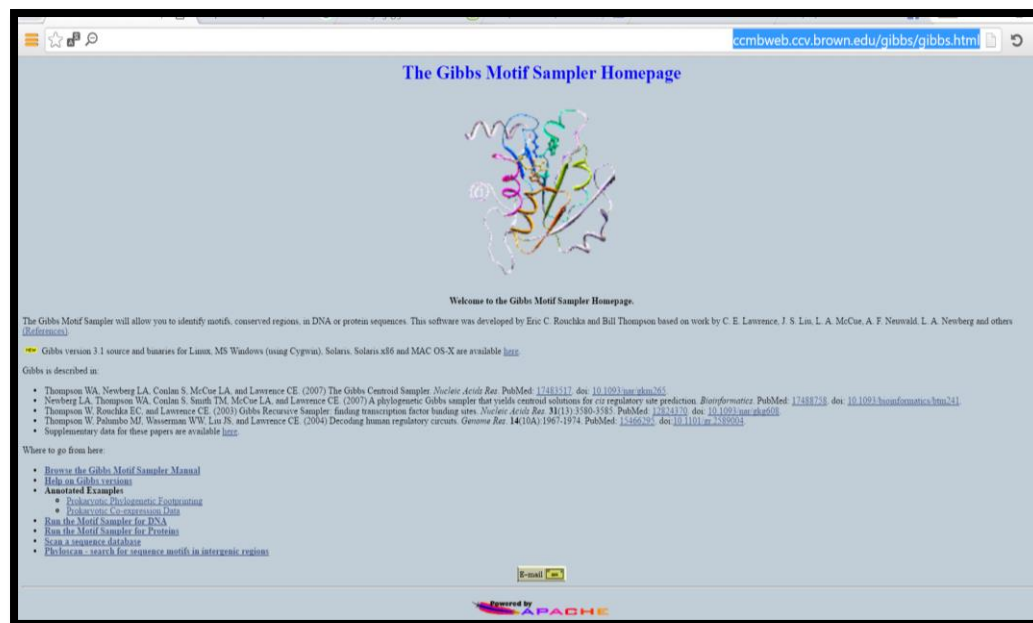


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Main Applications of Multiple Sequence Alignments

DNA regulatory elements

“You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potentially elements similar binding sites. Use the **Gibbs sampler** to identify these sites:
<http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html>”

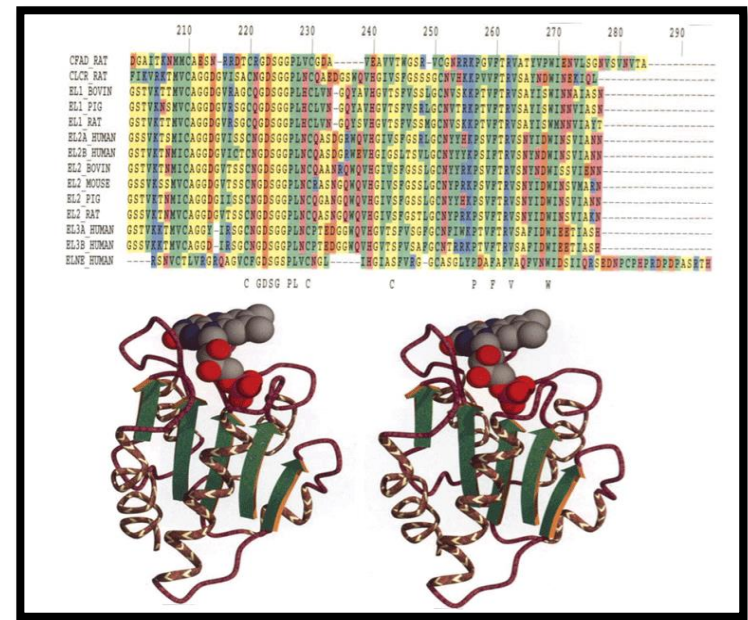
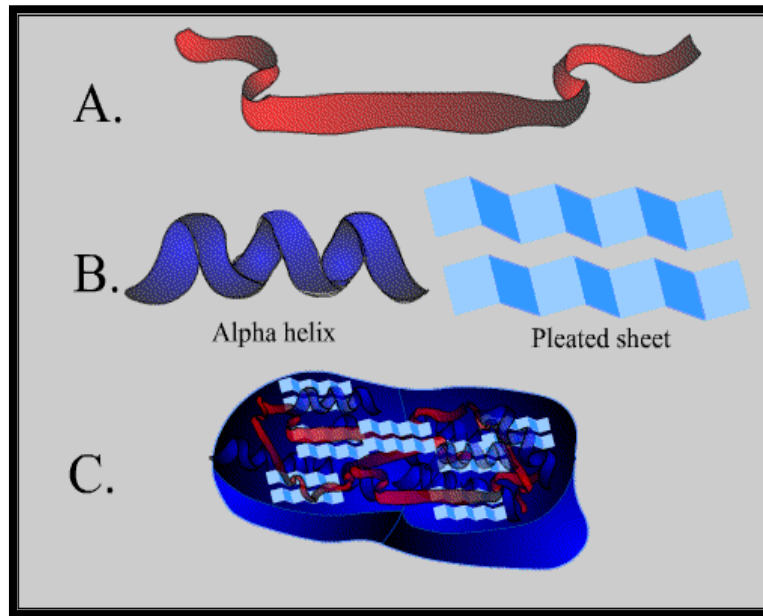


Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Main Applications of Multiple Sequence Alignments•

Structure prediction

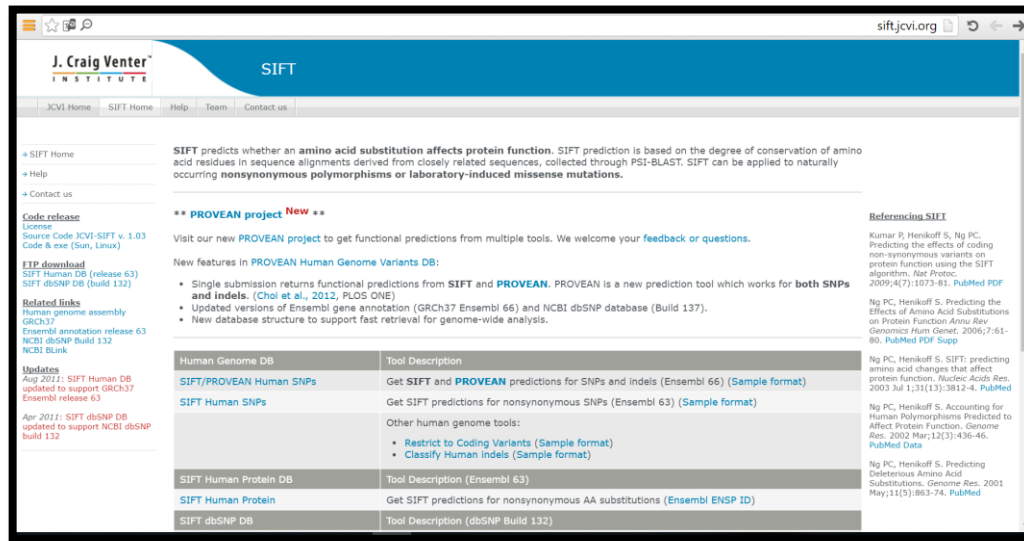
“A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for both proteins and RNA. Sometimes it can also **help in the building of a 3-D model**”.



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Main Applications of Multiple Sequence Alignments• nsSNP analysis

“Various gene alleles often have different amino-acid sequences. Multiple alignments can help you **predict whether a Non-Synonymous Single-Nucleotide Polymorphism** is likely to be **harmful**. See the **SIFT** site for more details: <http://sift.jcvi.org/>”



The screenshot shows the SIFT (Sorting Intolerant From Tolerant) website. The header includes the J. Craig Venter Institute logo and the SIFT logo. The main content area describes the SIFT prediction tool, which predicts whether an amino acid substitution affects protein function based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI-BLAST. SIFT can be applied to naturally occurring nonsynonymous polymorphisms or laboratory-induced missense mutations.

Key features and links include:


- Code release:** Source Code JCvi-SIFT v. 1.03, Code & exe (Sun, Linux)
- FTP download:** SIFT Human DB (release 63), SIFT dbSNP DB (build 132)
- Related links:** Human genome assembly GRCh37, Ensembl annotation release 63, NCBI dbSNP Build 132, NCBI BLink
- Updates:** Aug 2011: SIFT Human DB updated to support GRCh37, Ensembl release 63; Apr 2011: SIFT dbSNP DB updated to support NCBI dbSNP build 132
- Referencing SIFT:** Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 2009;4(7):1073-81. [PubMed PDF](#)
- Other human genome tools:**
 - Restrict to Coding Variants (Sample format)
 - Classify Human Indels (Sample format)

Human Genome DB	Tool Description
SIFT/PROVEAN Human SNPs	Get SIFT and PROVEAN predictions for SNPs and indels (Ensembl 66) (Sample format)
SIFT Human SNPs	Get SIFT predictions for nonsynonymous SNPs (Ensembl 63) (Sample format)
SIFT Human Protein DB	Tool Description (Ensembl 63)
SIFT Human Protein	Get SIFT predictions for nonsynonymous AA substitutions (Ensembl ENSP ID)
SIFT dbSNP DB	Tool Description (dbSNP Build 132)

Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

PCR analysis

CODEHOP:
Consensus-**D**Egenerate **H**ybrid **O**ligonucleotide **P**rimer*s*



NOTICE: This version of CODEHOP is no longer maintained.
Please try this site instead: [CODEHOP](#) at the Viral Bioinformatics Resource Center in Victoria, British Columbia, Canada

PCR primers designed from protein multiple sequence alignments

- [Getting started](#)
- [Full Help file](#)
- [The CODEHOP algorithm](#)
- [The CODEHOP nomenclature](#)
- [Stats identified using CODEHOP](#)

The input should be a set of local multiple alignments (blocks) of a group of related protein sequences. The alignments must be in [Blocks Database format](#), such as in [Block Maker](#) output.

Unaligned parts of Clustal- or FASTA-formatted global multiple alignments can be automatically turned into blocks by the [Blocks multiple alignment processor](#). You can also manually reformat multiple sequence alignments with the [Blocks formatter](#).

The output of all these programs contains links that send the resulting blocks to this page.

If your sequences align globally you will get better multiple alignment results from [Clustal](#) than from the motif finders used by Block Maker.

Blocks are processed using sequence weights (the numbers following each sequence segment). To [emphasize particular sequences](#) in the block(s) manually adjust the sequence weights. Increase the number to give a sequence more weight.

Paste your block(s) below:

[Close](#) (degenerate 3' region)

[Cleanup](#) (non-degenerate 5' region)

- [degeneracy](#) [default=128]:
 - [strings](#) [default=0.0]:
 - [temperature](#) [default=60.0]:
 - [positions](#) [default=5]:

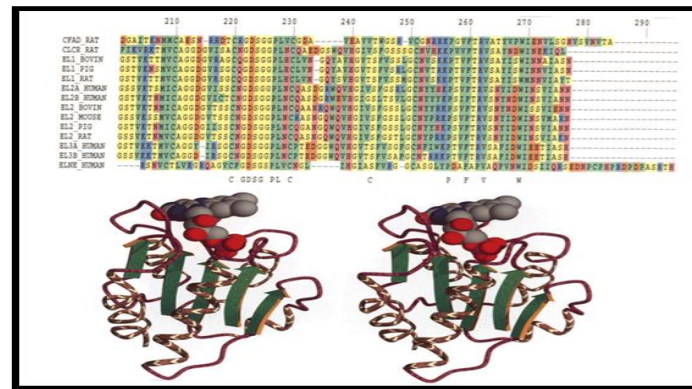


What are the kinds of sequences you're looking for?

Always bear in mind that in evolution:

1- **Important** amino acids (or nucleotides) are **NOT** allowed to **mutate**. For instance, active sites of enzymes are much conserved.

2- **Less-important** residues **change** more **easily** — sometimes randomly —and sometimes in order to **adapt a function**.



Claverie J, Notredame C (2007). Bioinformatics for Dummies (2nd Edn). Wiley publishing, Inc. 436 pp.

Tips for Naming sequences



Never use white spaces



Do not use special symbols



Never use names longer than 15 characters



Never give the same name to two different sequences

[Mansour A, Jaime A. Teixeira da Silva, Gábor Gyulai \(2009\) Assessment of molecular \(dis\)similarity: The role of multiple sequence alignments \(MSA\) programs in biological research. *Genes, genomes and genomics* 30-23 :\(1 eussl laicepS\)3 .Print ISSN \) \(0383-1749Bioinformatics SI.\(](#)

Tips for difficult MSA to interpret

- 1 Remove insertions/deletions
- 2 Redo MSA with the smaller set
- 3 Keep trimming to interpret

[Mansour A, Jaime A. Teixeira da Silva, Gábor Gyulai \(2009\) Assessment of molecular \(dis\)similarity: The role of multiple sequence alignments \(MSA\) programs in biological research. *Genes, genomes and genomics* 30-23 :\(1 eussl laicepS\)3 .Print ISSN \) \(0383-1749Bioinformatics SI.\(](#)

Enhancing Alignments

Remove gaps

Remove extremities

Keep informative blocks

*Enhancing
your
Alignment*

[Mansour A, Jaime A. Teixeira da Silva, Gábor Gyulai \(2009\) Assessment of molecular \(dis\)similarity: The role of multiple sequence alignments \(MSA\) programs in biological research. *Genes, genomes and genomics* 30-23 :\(1 eussl laicepS\)3 .Print ISSN \) \(0383-1749Bioinformatics SI.\(](#)

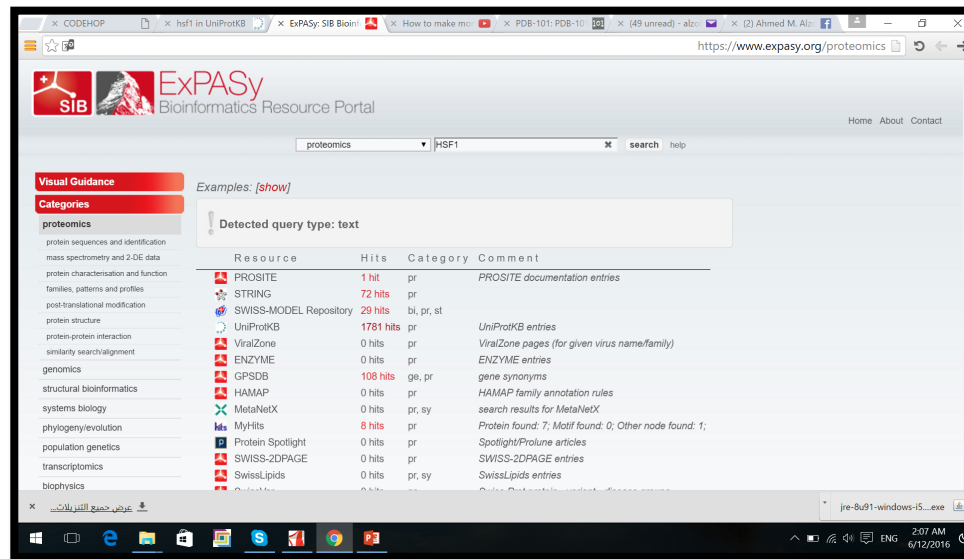
(to try in your own time)

Searching sequences on the ExPASy server

Only to retrieve *protein* sequences in FASTA format

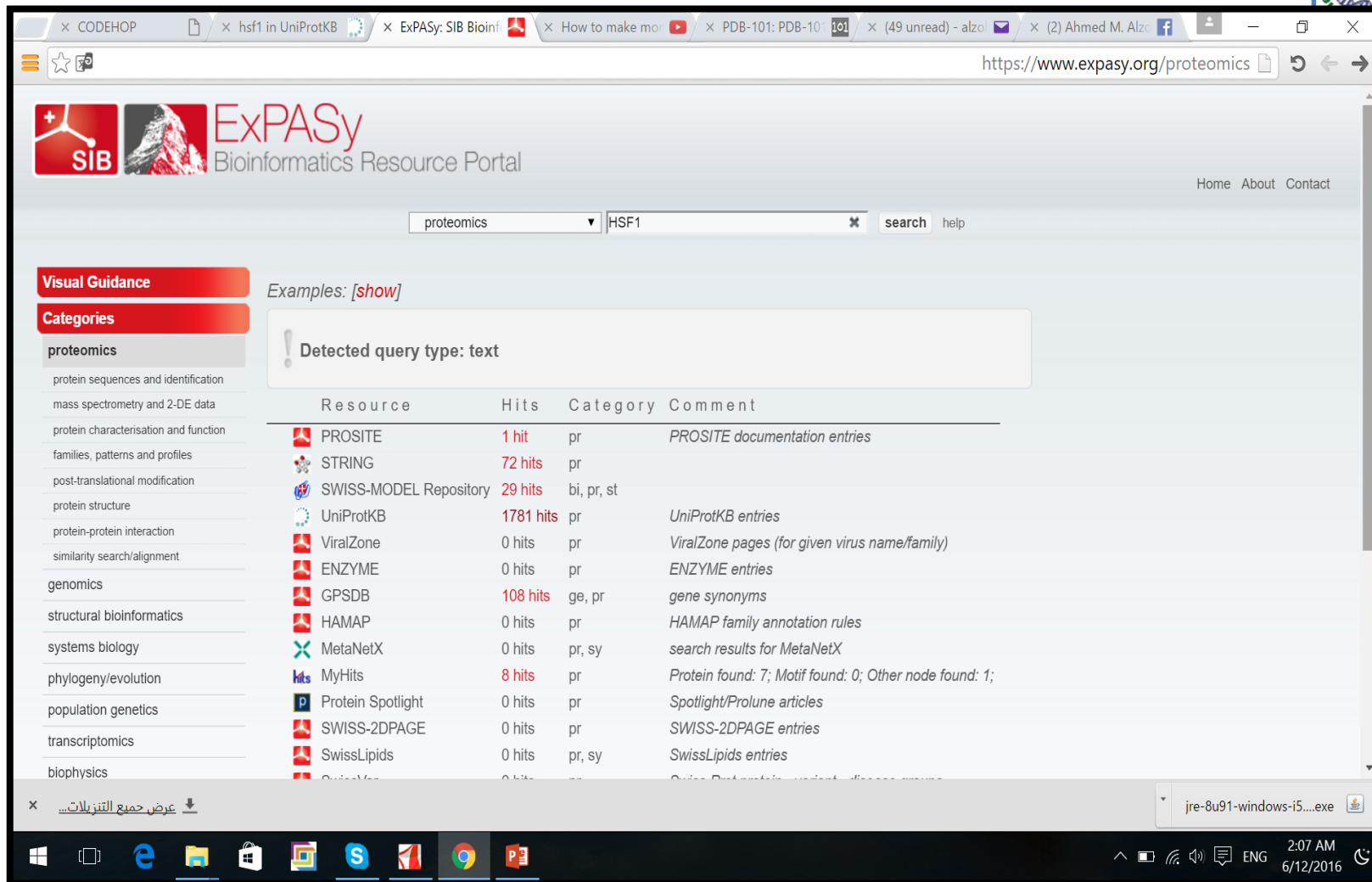
Example: Heat shock factor 1 (HSF1)

Choose <http://www.uniprot.org/>



The screenshot shows the ExPASy Bioinformatics Resource Portal search results for the query "HSF1". The search results are displayed in a table with columns: Resource, Hits, Category, and Comment. The results are as follows:

Resource	Hits	Category	Comment
PROSITE	1 hit	pr	PROSITE documentation entries
STRING	72 hits	pr	
SWISS-MODEL Repository	29 hits	bl, pr, st	
UniProtKB	1781 hits	pr	UniProtKB entries
ViralZone	0 hits	pr	ViralZone pages (for given virus name/family)
ENZYME	0 hits	pr	ENZYME entries
GPSDB	108 hits	ge, pr	gene synonyms
HAMAP	0 hits	pr	HAMAP family annotation rules
MetaNetX	0 hits	pr, sy	search results for MetaNetX
MyHits	8 hits	pr	Protein found: 7; Motif found: 0; Other node found: 1;
Protein Spotlight	0 hits	pr	Spotlight/Prolume articles
SWISS-2DPAGE	0 hits	pr	SWISS-2DPAGE entries
SwissLipids	0 hits	pr, sy	SwissLipids entries



CODEHOP | hsf1 in UniProtKB | ExPASy: SIB Bioinformatics Resource Portal | How to make more | PDB-101: PDB-101 | (49 unread) - alzo | (2) Ahmed M. Alzohairy

https://www.expasy.org/proteomics

ExPASy
Bioinformatics Resource Portal

Home About Contact

proteomics | HSF1 | search | help

Visual Guidance

Categories

proteomics

- protein sequences and identification
- mass spectrometry and 2-DE data
- protein characterisation and function
- families, patterns and profiles
- post-translational modification
- protein structure
- protein-protein interaction
- similarity search/alignment
- genomics
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics

Examples: [show]

! Detected query type: text

Resource	Hits	Category	Comment
PROSITE	1 hit	pr	PROSITE documentation entries
STRING	72 hits	pr	
SWISS-MODEL Repository	29 hits	bi, pr, st	
UniProtKB	1781 hits	pr	UniProtKB entries
ViralZone	0 hits	pr	ViralZone pages (for given virus name/family)
ENZYME	0 hits	pr	ENZYME entries
GPSDB	108 hits	ge, pr	gene synonyms
HAMAP	0 hits	pr	HAMAP family annotation rules
MetaNetX	0 hits	pr, sy	search results for MetaNetX
MyHits	8 hits	pr	Protein found: 7; Motif found: 0; Other node found: 1;
Protein Spotlight	0 hits	pr	Spotlight/Prolune articles
SWISS-2DPAGE	0 hits	pr	SWISS-2DPAGE entries
SwissLipids	0 hits	pr, sy	SwissLipids entries
SwissVar	0 hits	pr	SwissVar protein variants database

عرض جميع النتائج

jre-8u91-windows-i5...exe

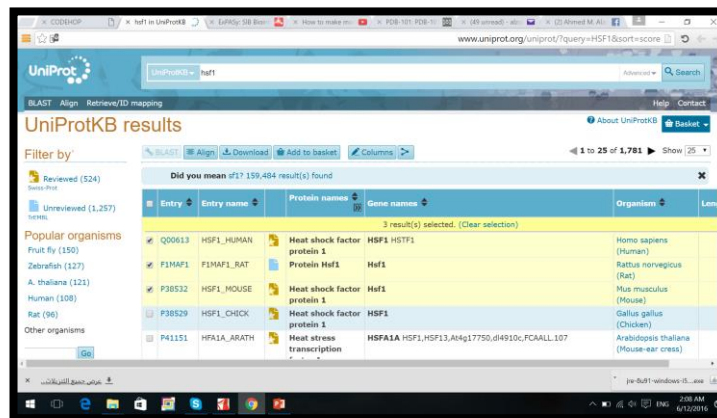
2:07 AM 6/12/2016

(to try in your own time)

Select the sequences you want

**This is the most delicate part of the process
you can use the following guidelines**

- Select the top sequence.
- For a first analysis, you want to select ten sequences or fewer.
- check it's similar to the query sequence - along its entire length.



(to try in your own time)

Methods to export your sequences

- **FASTA:** Generates a file that contains your sequences in FASTA format.
- **ClustalW, Tcoffee, and MAFFT:** These are MSA packages running on the EMBnet server.
- **Reduce Redundancy:** This option will extract the most meaningful sequences from your dataset.
- **Pratt:** Will search for conserved motifs in your sequences without aligning them

Practical

(to try in your own time)

- Go to <https://www.expasy.org/proteomics>
- Search for HSF1
- Click on ([UniProtKB](#))
- Retrieve your protein sequences (eg. Heat shock Factor1 “HSF1”) from different organisms
- This will take you to <http://www.uniprot.org/uniprot/?query=HSF1&sort=score>
- Select your organism (Human, Rat, Mouse, Arabidopsis, Chicken, Pig)
- Click Download (Download Selected) then (Go)
- Save it in FASTA format in one text file.
- Align the sequences using Clustal Omega
- Checking the gene-based phylogenetics tree
- Add one more sequence NOT related sequence (Out Group)
- Checking the change on the gene-based phylogenetics tree

