

Practical Session 1: Introduction to databases and homology-based functional inference

A- UniProt

Exercise 1

You were flipping through a past issue of Science and came across the following article:

Using ancient protein kinases to unravel a modern cancer drug's mechanism

C. Wilson,^{1*} R. V. Agafonov,^{1*} M. Hoemberger,¹ S. Kutter,¹ A. Zorba,¹ J. Halpin,¹
V. Buosi,¹ R. Otten,¹ D. Waterman,¹ D. L. Theobald,² D. Kern^{1†}

Macromolecular function is rooted in energy landscapes, where sequence determines not a single structure but an ensemble of conformations. Hence, evolution modifies a protein's function by altering its energy landscape. Here, we recreate the evolutionary pathway between two modern human oncogenes, Src and Abl, by reconstructing their common ancestors. Our evolutionary reconstruction combined with x-ray structures of the common ancestor and pre-steady-state kinetics reveals a detailed atomistic mechanism for selectivity of the successful cancer drug Gleevec. Gleevec affinity is gained during the evolutionary trajectory toward Abl and lost toward Src, primarily by shifting an induced-fit equilibrium that is also disrupted in the clinical T315I resistance mutation. This work reveals the mechanism of Gleevec specificity while offering insights into how energy landscapes evolve.

You want to find all human protein kinases in UniProt that have a 3D structure associated with them.

A- How would you do this?

In order to find all human protein kinases in UniProt, we need to use the advanced search. We have to select the taxonomy Homo Sapiens, protein name Kinase and cross-references – 3d structure databases. The final query is:

```
database:(type:pdb) name:kinase AND organism:"Homo sapiens (Human) [9606]"
```

B- How many proteins have you found?

We find 498 proteins, 475 reviewed and 3 unreviewed.

Exercise 2. You are helping colleagues analyse the protein products of the gene *tp53* in different organisms. They have sent you a list of gene accessions that they are interested in. They are interested in analysing the corresponding protein products for these genes.

How would you download the corresponding UniProt protein sequences for these gene accessions?

We should go to Retrieve/ID mapping and submit our identifiers, selecting the different options.

16 out of 17 EMBL/GenBank/DDBJ identifiers were successfully mapped to 8988 UniProtKB IDs in the table below.

[Click here to download the 1 unmapped identifier.](#)

Filter byⁱ

Reviewed (3,524)
Swiss-Prot

Unreviewed (5,464)

BLAST Align **Download** Add to basket Columns

	Your list:...0A3A5Q	Entry	Entry name	Protein names	Gene names
<input type="checkbox"/>	AP002032	Q9FG25	Q9FG25_ARATH	Uncharacterized protein	At5g06810

Quiz 1

1- UniProtKB consists of two types of entries, Reviewed (Swiss-Prot) and _____.

D- Unreviewed (TrEMBL)

2- UniProt allows you to convert other database identifiers to UniProt identifiers but not vice versa.

A- True

3- Which of the following is not a UniProt dataset?

A- Pathogens

4- To find functional information about a protein, which UniProt section should you consult?

C- UniProtKB

5- Which of the following are tools provided on the UniProt website for protein sequence analysis? Choose all that apply

B- BLAST

B- InterProScan

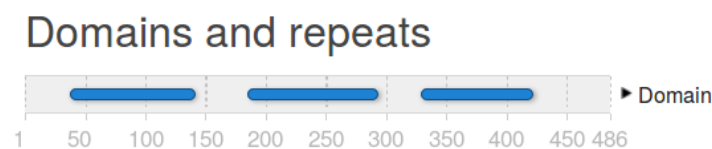
Exercise 3

Find information about the protein in my_protein.txt using InterPro.

What family does this protein belong to?

The protein belongs to Membrane-bound alcohol dehydrogenase, cytochrome c subunit (IPR014353)

What domains does it have?



The domain it has is Cytochrome c-like domain.

What processes is it involved in?

There are no biological processes predicted for this protein.

Quiz 2

1- How can you know the type (family, domain, repeat, site) of an InterPro entry?

A- The entry type is indicated by a specific icon before the name and identifier on every InterPro entry page

2- When an InterPro entry consists of several signatures, this means:

C- Those signatures are predicting the same biological entity: a protein family, domain, repeat or site

3- If you have a novel uncharacterised protein sequence you can use InterPro:

A- To predict the function of the protein and the presence of important domains or sites

C- Ensembl

Exercise 4

ESR1 encodes an estrogen receptor, a ligand-activated transcription factor composed of several domains important for hormone binding, DNA binding, and activation of transcription.

The protein localizes to the nucleus where it may form a homodimer or a heterodimer with estrogen receptor 2. Estrogen and its receptors are essential for sexual development and reproductive function, but also play a role in other tissues such as bone.

You must search for ESR1 Human gene in ensembl and obtain the follow information:

- Which is the ensembl gene identifier?

The ensembl gene identifier is: ENSG00000091831¹.

- How many transcripts has this gene?

The gene has 15 transcripts.

- How many homologs are in chimpanzees?

It has 193 homologs, 185 orthologues and 8 paralogues. In chimpanzees, there are 1 ortholog.

- Download the Genomic sequences. Make a multiple sequence alignment between Human and Bonobo, Chimpanzee, Crab-eating macaque and Gorilla.

In order to make a multiple sequence alignment, we can go to the genomic alignment option found in the comparative genomics category. There, we can select the primate species we want and download the alignment.

Location: 6:151,656,691-152,129,619 Gene: ESR1

Gene: ESR1 ENSG00000091831

Description estrogen receptor 1 [Source:HGNC Symbol;Acc:HGNC:3467]

Gene Synonyms ER-alpha, ESR, Era, NR3A1

Location [Chromosome 6: 151,656,691-152,129,619](#) forward strand.
GRCh38:CM000668.2

About this gene This gene has 15 transcripts ([splice variants](#)), [185 orthologues](#), [8 paralogues](#), is a member of [1 Ensembl protein family](#) and is associated with [62 phenotypes](#).

Transcripts [Show transcript table](#)

1 http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000091831;r=6:151656691-152129619

There are different formats for the alignment, such as FASTA, clustal or stockholm. I decided to select clustalw format.

```

CLUSTAL W (1.81) multiple sequence alignment

homo_sapiens/1-985162      GAG--CCTCAAATATCTCCAAATCTGATACCAATCCTTTTGATTGTGAATTATATTCTG
pan_paniscus/1-985162     GAG--CCTCAAATATCTCCAAATCTGATACCAATCCTTTTGATTGTGAATTATATTCTG
pan_troglodytes/1-985162  GAG--CCTCAAATATCTCCAAATCTGATACCAATCCTTTTGATTGTGAATTATATTCTG
gorilla_gorilla/1-985162  GAG--CCTCAAATATCTCCAAATCTGATACCAATCCTTTTGATTGTGAATTATATTCTG
macaca_fascicularis/1-985162 GAG--CCTCAAATATCTCCAAATCTGATACCAATCCCTTTGATTGTGAATTATATTCTG
***      *****

homo_sapiens/1-985162      TAGCTACCAAAGAAGGTAAGTTTTTATTTT-TTCTACTCTATTAACCTTCCCTTGGACAA
pan_paniscus/1-985162     TAGCTACCAAAGAAGGTAAGTTTTTATTTT-TTCTACTCTATTAACCTTCCCTTGGACAA
pan_troglodytes/1-985162  TAGCTACCAAAGAAGGTAAGTTTTTATTTT-TTCTACTCTATTAACCTTCCCTTGGACAA
gorilla_gorilla/1-985162  TAGCTACCAAAGAAGGTAAGTTTTTATTTT-TTCTACTCTATTAACCTTCCCTTGGACAA
macaca_fascicularis/1-985162 TAGCTACCAAAGAAGGTAAGTTTTTATTTT-TTCTACTCTATTAACCTTCCCTTGGGCAA
*****

homo_sapiens/1-985162      CTGAATATTAAGATGACTATGTAAGGAGGTTATCAGACCAAGGCCTCACACATCAGGATA
pan_paniscus/1-985162     CTGAATATTAAGATGACTATGTAAGGAGGTTATCAGACCAAGGCCTCACACATCAGGATA
pan_troglodytes/1-985162  CTGAATATTAAGATGACTATGTAAGGAGGTTATCAGACCAAGGCCTCACACATCAGGATA
gorilla_gorilla/1-985162  CTGAATATTAAGATGACTATGTAAGGAGGTTATCAGACCAAGGCCTCACACATCAGGATA
macaca_fascicularis/1-985162 CTGAATATTAAGATGACTATGTAAGGAGGTTATCAGACCAAGGCCTCACACATCAGGATA
*****

homo_sapiens/1-985162      AAAGCACATGCCATAGAAAGAACATTTGTGTCTCAAAGGTGATACCAAGACAAGGCTGT
pan_paniscus/1-985162     AAAGCACATGCCATAGAAAGAACATTTGTGTCTCAAAGGTGATACCAAGACAAGGCTGT
pan_troglodytes/1-985162  AAAGCACATGCCATAGAAAGAACATTTGTGTCTCAAAGGTGATACCAAGACAAGGCTGT
gorilla_gorilla/1-985162  AAAGCACATGCCATAGAAAGAACATTTGTGTCTCAAAGGTGATACCAAGACAAGGCTGT
macaca_fascicularis/1-985162 AAAGCACATGCCATAGAAAGAACATTTGTGTCTCAAAGGTGATACCAAGGCAAGGCTAT
*****

homo_sapiens/1-985162      GGGATATATATGGGCACAATGGTTGATACCTTCAAAGACTTCATACATGGTGTGGAGGTT
pan_paniscus/1-985162     GGGATATATATGGGCACAATGGTTGATACCTTCAAAGACTTCATACATGGTGTGGAGGTT
pan_troglodytes/1-985162  GGGATATATATGGGCACAATGGTTGATACCTTCAAAGACTTCATACATGGTGTGGAGGTT
gorilla_gorilla/1-985162  GGGATATATATGGGCACAATGGTTGATACCTTCAAAGACTTCATACATGGTGTGGAGGTT
macaca_fascicularis/1-985162 GGGATATATATGGGCACAATGT-TGATATCTTCAAAGACTTGATACATAGTCTGGAGGTT
*****

homo_sapiens/1-985162      TTTGGAGATTT-TAATTTATAATGACAATCTTCCAGTTAGGAGAATTT-TTGGACTGTA

```

:

:

D- GO enrichment

Exercise 5

Search for GO enrichment terms in Biological Process using the list of genes from Arabidopsis thaliana (Athaliana_identifiers.txt). To do this, you must use the Gene Ontology Consortium website <http://geneontology.org/>. Download the results.



















Analysis Type:	PANTHER Overrepresentation Test (Released 20190308)
Annotation Version and Release Date:	GO Ontology database Released 2019-02-02
Analyzed List:	upload_1 (Arabidopsis thaliana)

ATPase activity (GO:0016887)	639	18	6.83	+	2.63	2.49E-04	2.95E-02			
nucleoside-triphosphatase activity (GO:0017111)	875	24	9.36	+	2.56	3.55E-05	6.07E-03			
pyrophosphatase activity (GO:0016462)	920	24	9.84	+	2.44	1.12E-04	1.73E-02			
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides (GO:0016818)	922	24	9.86	+	2.43	1.14E-04				
1.67E-02										
hydrolase activity, acting on acid anhydrides (GO:0016817)	928	24	9.93	+	2.42	1.20E-04	1.60E-02			
ATP binding (GO:0005524)	1547	40	16.55	+	2.42	3.61E-07	1.59E-04			
adenyl ribonucleotide binding (GO:0032559)	1569	40	16.78	+	2.38	4.96E-07	1.91E-04			
adenyl nucleotide binding (GO:0030554)	1576	40	16.86	+	2.37	5.49E-07	1.88E-04			
purine ribonucleoside triphosphate binding (GO:0035639)	1777	43	19.01	+	2.26	7.00E-07	1.96E-04			
ribonucleotide binding (GO:0032553)	1822	44	19.49	+	2.26	8.01E-07	2.06E-04			
purine ribonucleotide binding (GO:0032555)	1798	43	19.23	+	2.24	1.27E-06	2.80E-04			
purine nucleotide binding (GO:0017076)	1807	43	19.33	+	2.22	1.36E-06	2.79E-04			
carbohydrate derivative binding (GO:0097367)	1860	44	19.89	+	2.21	1.07E-06	2.53E-04			
nucleotide binding (GO:0000166)	2231	52	23.86	+	2.18	1.50E-07	9.24E-05			
nucleoside phosphate binding (GO:1901265)	2231	52	23.86	+	2.18	1.50E-07	7.70E-05			
drug binding (GO:0008144)	1908	43	20.41	+	2.11	4.92E-06	8.91E-04			
anion binding (GO:0043168)	2254	49	24.11	+	2.03	3.12E-06	6.00E-04			
small molecule binding (GO:0036094)	2605	56	27.86	+	2.01	6.52E-07	2.01E-04			
protein binding (GO:0005515)	4527	90	48.42	+	1.86	3.13E-09	4.82E-06			
hydrolase activity (GO:0016787)	3162	58	33.82	+	1.71	4.77E-05	7.73E-03			
organic cyclic compound binding (GO:0097159)	5689	104	60.85	+	1.71	7.86E-09	8.08E-06			
heterocyclic compound binding (GO:1901363)	5667	103	60.61	+	1.70	1.22E-08	9.39E-06			
binding (GO:0005488)	10024	161	107.21	+	1.50	2.53E-10	7.80E-07			
molecular_function (GO:0003674)	21868	259	233.90	+	1.11	2.08E-04	2.67E-02			
Unclassified (UNCLASSIFIED)	5713	36	61.10	-	.59	2.08E-04	2.56E-02			































B- Using Quick GO (<https://www.ebi.ac.uk/QuickGO/>), lists all terms that are direct descendants of the GO terms obtained from A.

Here, are some lists of the direct descendants of the GO terms from A.

























This table lists all terms that are direct descendants (child terms) of GO:0015662

Child Term	Relationship to GO:0015662
GO:0008556   potassium-transporting ATPase activity	is_a
GO:0016463   zinc-exporting ATPase activity	is_a
GO:0008554   sodium-exporting ATPase activity, phosphorylative mechanism	is_a
GO:0008553   proton-exporting ATPase activity, phosphorylative mechanism	is_a
GO:0008551   cadmium-exporting ATPase activity	is_a
GO:0015444   magnesium-importing ATPase activity	is_a
GO:0015445   silver-exporting ATPase activity	is_a
GO:0005388   calcium-transporting ATPase activity	is_a
GO:0043682   copper-transporting ATPase activity	is_a

This table lists all terms that are direct descendants (child terms) of GO:0042623

Child Term	Relationship to GO:0042623
GO:0008186    RNA-dependent ATPase activity	is_a
GO:0140083    protein-DNA unloading ATPase activity	is_a
GO:0036402    proteasome-activating ATPase activity	is_a
GO:0030898    actin-dependent ATPase activity	is_a
GO:0070463    tubulin-dependent ATPase activity	is_a
GO:0030899    calcium-dependent ATPase activity	is_a
GO:0061860    DNA clamp unloader activity	is_a
GO:1990939    ATP-dependent microtubule motor activity	is_a
GO:1990621    ESCRT IV complex	capable_of
GO:0070615    nucleosome-dependent ATPase activity	is_a

This table lists all terms that are direct descendants (child terms) of GO:0016887

Child Term	Relationship to GO:0016887
GO:0032780    negative regulation of ATPase activity	negatively_regulates
GO:0042030    ATPase inhibitor activity	negatively_regulates
GO:0042623    ATPase activity, coupled	is_a
GO:0032781    positive regulation of ATPase activity	positively_regulates
GO:1904949    ATPase complex	capable_of
GO:0060590    ATPase regulator activity	regulates
GO:0043462    regulation of ATPase activity	regulates
GO:0001671    ATPase activator activity	positively_regulates

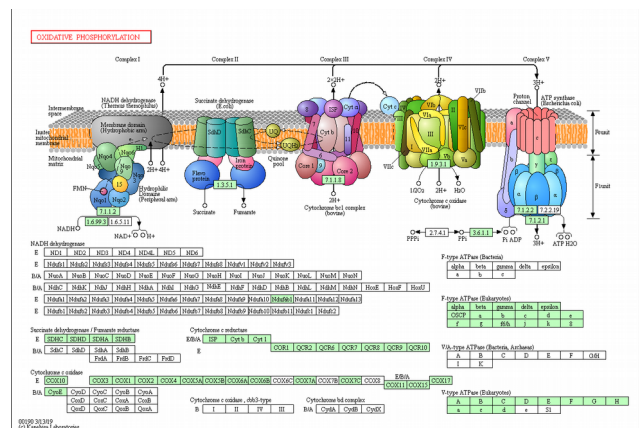
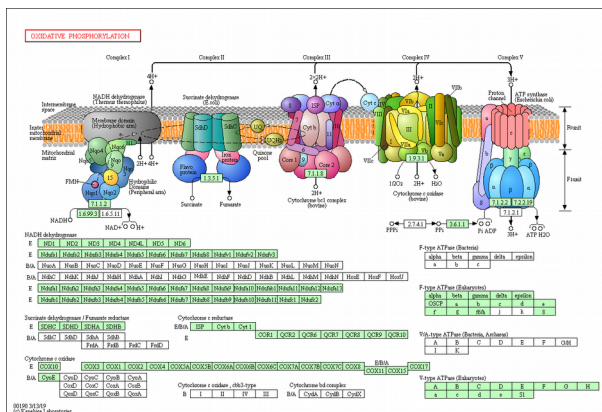
E- KEGG.

Exercise 7

Oxidative phosphorylation is the process in which ATP is formed as a result of the transfer of electrons from NADH or FADH₂ to O₂ by a series of electron carriers. This process, which takes place in mitochondria, is the major source of ATP in aerobic organisms.

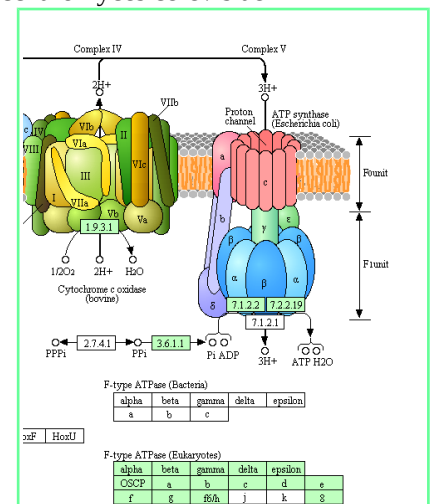
In eukaryotes, these redox reactions are carried out by a series of protein complexes within the inner membrane of the cell's mitochondria, whereas, in prokaryotes, these proteins are located in the cells' intermembrane space.

A- Which is the main difference in the Oxidative phosphorylation pathway between Human and the yeast *Saccharomyces cerevisiae*?



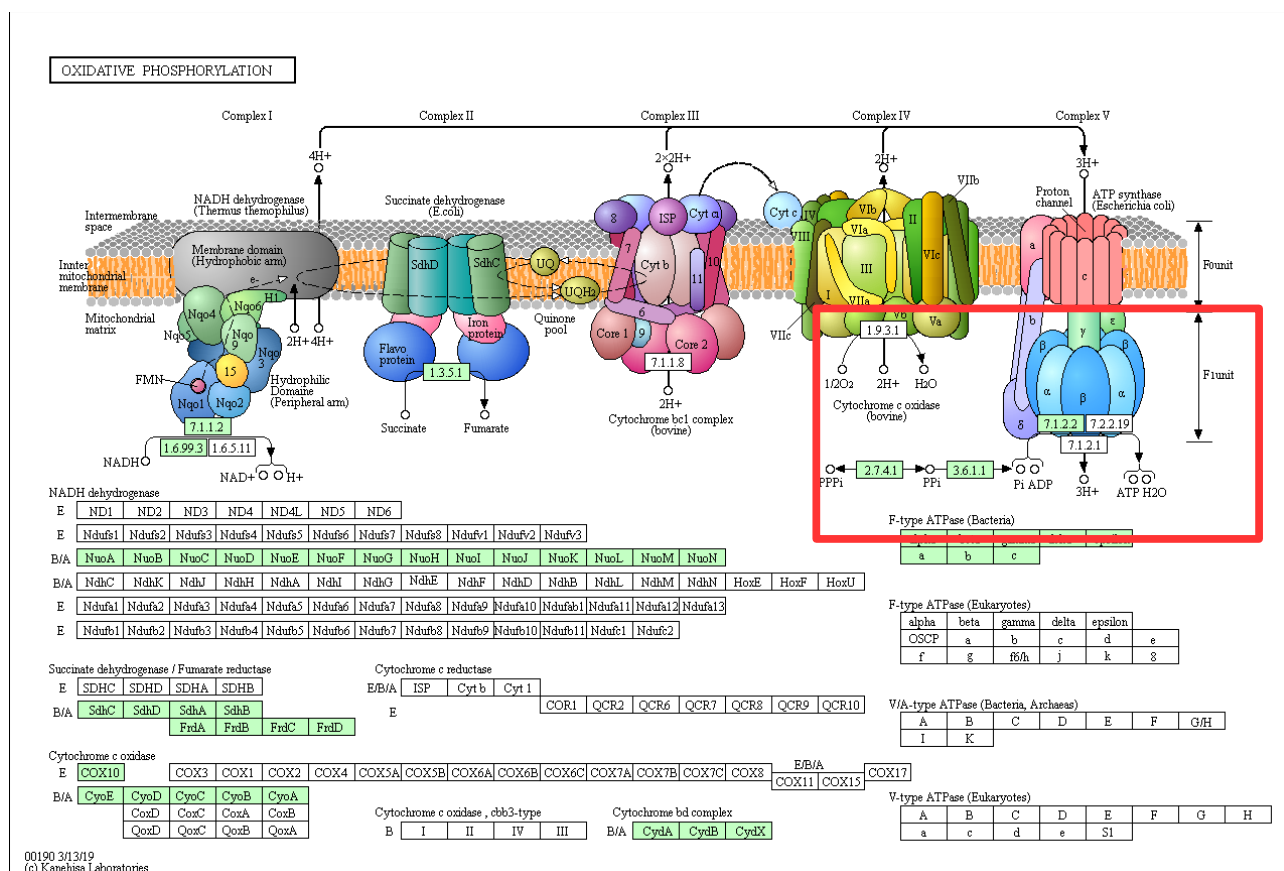
The main difference between Human and the yeast *Saccharomyces cerevisiae* is found in the **green boxes** of the images. The green boxes indicate the presence of the gene in the genomic of the organism and the completeness of the pathway. Therefore, the genes expressed in the organisms are different.

For example, in the last step of the oxidative phosphorylation pathway, in Human it is formed (from a phosphate and ADP) ATP and water. Whereas, in the *saccharomyces*, it is formed 3H⁺.



B- And between Human and the bacteria *Escherichia coli*?

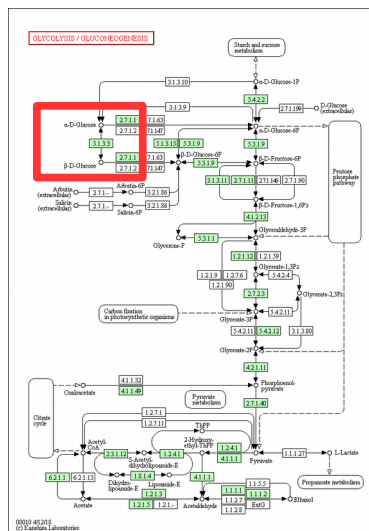
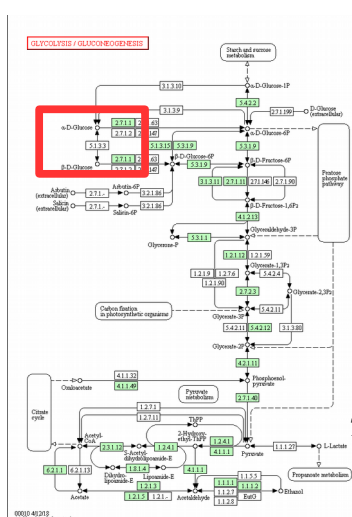
As we have just said in Human, ATP and water is formed in the oxydative phosphorylation pathway. However, in the following image, we can see that in Complex V of *Escherichia coli* neither $3H^+$ nor ATP and water is formed.



Exercise 8

Glycolysis is the process of converting glucose into pyruvate and generating small amounts of ATP (energy) and NADH (reducing power). It is a central pathway that produces important precursor metabolites.

Using KEGG you have to identify differences in the Glycolysis pathway between this fungi species: *Penicillium rubens* and *Tremella mesenterica*.

*Penicillium rubens*

Tremella mesenterica

The principal difference is found in the 5.1.3.3 enzyme.

Enzymes [BR:pcs01000]

5. Isomerases

5.1 Racemases and epimerases

5.1.3 Acting on carbohydrates and derivatives

5.1.3.3 aldose 1-epimerase

Pc13g14400

This enzyme is only found in *penicillium rubens*, whereas it is not found in *tremella mesenterica*. In consequence, beta-D-glucose is not formed in *Tremella mesenterica*.

Exercise 9

Choose a protein from the file ASPCL.fasta and use the web resources that we have seen, in order to annotate that protein (you should get all the information you can).

The protein I have chosen is: ASPCL_0078_02820.

The sequence of the protein is:

>ASPCL_0078_02820

```
MFKKAVKDHSAGSFKPLQSSLFSSNGPAQSKLPPQQSIGVKRKIEMANTGGALGSLHSAVYFDENDFDDIDLDIEEPEPFIPPTKIVRPSIGGEALETSSNAN
SGIMRTNISKKEPITIDLVSPPTDIKYPDLPTIPDEHVPPSSSIQYPWSSSPSHLQKPSTGRTIPWLQKSEIPEEEYNKPQTPARPKSTAPWNKSASAIKEE
QKELRRQHKNQKNDASSQLQRPKIASFLSDEQRHVLDTVVQQGKSIFFTGSAGTGKSVLMREIIKQLRSKYRKEPDRVAVTASTGLAACNIEGVTLSHFAG
IGLGKEPVPELVKKIKRNQKARNRWLRTKVLIIDEVSMVDGDLFDKLEEIARRIRNNGRPFGGIQLVVTGDFFQLPPVPEGSNREAKFAFAAGTWNTSIQHTILL
THVFRQKDPDFADMLNEMRLGKLTPTIETFKLSRPLNFHDSLEATELFPTRHEVEQANSARMVKLSGEMMTFQAVDSGSIQDAQYREKLLANCMAPPVIHLKK
GAQVMLIKNMEDTLVNGSIGRVAFAFMDAATFEYYRDNENEFSGRQENGSDENLSHARKKLKGLGNKDGIVVSRKWPLVCFVQPDGTERHLLCQPEAWKIELPN
GEVQAQRQOVPLILAWALSIHKAQGQTLQRVKVDLGRVFEKGQAYVALSRATSKGLQVTRFDARKVMVHPKVTEFYAKLVSITDVLAPKSSKARQLADKDSKSH
LDEEELLQQLYG
```

As we can see, the first step will be search the protein into **Uniprot**. However, we cannot search the name of the protein as we only have the sequence. So, we are going to use the **BLAST²** tool from uniprot. There, we obtain a protein called **ATP-dependent DNA helicase PIF1** from *Aspergillus clavatus* organism. The protein can be found on the nucleus or the mitochondria.

The function of the protein is: *DNA-dependent ATPase and 5'-3' DNA helicase required for the maintenance of both mitochondrial and nuclear genome stability*.

Then, we are going to submit our sequence in **Interproscan**, in order to study the family and the domains.

Our protein belongs to DNA helicase Pif1-like family.

It is formed by AAA+ ATPase domain.

2 BLAST from uniprot: <https://www.uniprot.org/uniprot/A1CBS2>

Domains and repeats



The protein is involved in two different **biological processes**: telomere maintenance (GO:0000723) and DNA repair (GO:0006281). In addition, it has one **molecular function**: DNA helicase activity (GO:0003678).

Moreover, it is important to get information of the gene in Ensembl. However, we need to search in <https://fungi.ensembl.org/>, as our organism is a fungi³.

e!EnsemblFungi | HMMER | BLAST | BioMart | Tools | Downloads |

Search:

e.g. **NAT2** or **alcohol***

The gene has 1 transcript and 302 orthologues and no paralogues.

In Ensembl, we can also look for the gene ontologies. In relation to Biological process and molecular function we find:

GO: Biological process					
Accession	Term	Evidence	Annotation source	Mapped using	Transcript IDs
GO:0000002	mitochondrial genome maintenance	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0000723	telomere maintenance	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0006281	DNA repair	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0006310	DNA recombination	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0006974	cellular response to DNA damage stimulus	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0005208	DNA duplex unwinding	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0043049	mitochondrial DNA repair	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0044806	G-quadruplex DNA unwinding	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:1902983	DNA strand elongation involved in mitotic DNA replication	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:1903469	removal of RNA primer involved in mitotic DNA replication	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:1905467	positive regulation of G-quadruplex DNA unwinding	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190

GO: Molecular function					
Accession	Term	Evidence	Annotation source	Mapped using	Transcript IDs
GO:0000166	nucleotide binding	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0003677	DNA binding	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0003678	DNA helicase activity	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0003723	RNA binding	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0004386	helicase activity	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0005524	ATP binding	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0016767	hydrolase activity	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0016887	ATPase activity	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0017116	single-stranded DNA-dependent ATP-dependent DNA helicase activity	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0039892	ATP-dependent 5'-3' DNA/RNA helicase activity	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0043141	ATP-dependent 5'-3' DNA helicase activity	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0051889	G-quadruplex DNA binding	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0061995	ATP-dependent protein-DNA complex displacement activity	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:0070338	flap-structured DNA binding	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190
GO:1905816	DNA-DNA annealing activity	IEA	UniProtKB/TrEMBL:A1C8S2		EAU13190

It is clear that the molecular function and biological processes match with the information from Interpro.

3 EnsemblFungi: https://fungi.ensembl.org/Aspergillus_clavatus/Gene/Summary?g=ACLA_016360;r=DS027049;2234265-2236615;t=EAU13190;db=core

