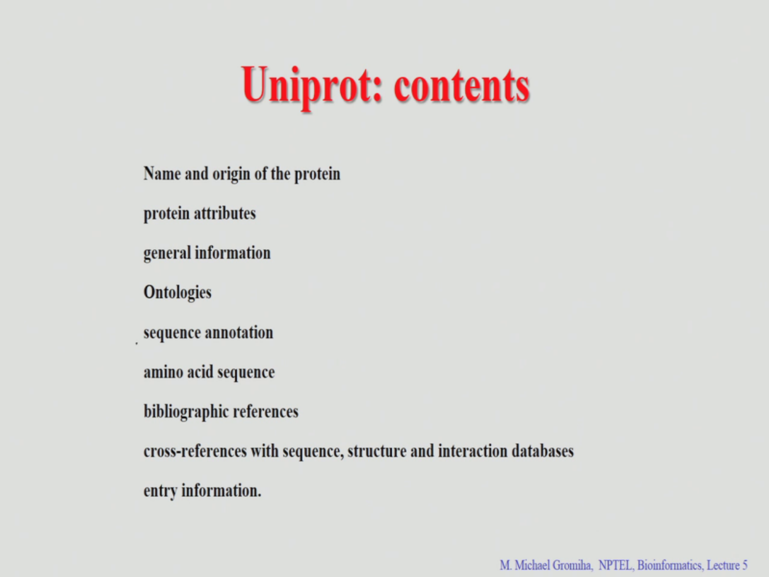
Bioinformatics
Prof. M. Michaele Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 05b
Protein sequence databases II

Now, we will explain more about the contents of UniProt. So, as I discussed earlier it contains a higher notation of the protein sequences. So, it has more information regarding a particular protein. So, what are the major aspects, what are the major contents of Uniprot database? First, they give about the proteins right, the names, origin and attributes and so on and go with the ontologies and then the next the sequence information, right.

(Refer Slide Time: 00:42)



Uniprot: contents

Name and origin of the protein

protein attributes

general information

Ontologies

sequence annotation

amino acid sequence

bibliographic references

cross-references with sequence, structure and interaction databases

entry information.

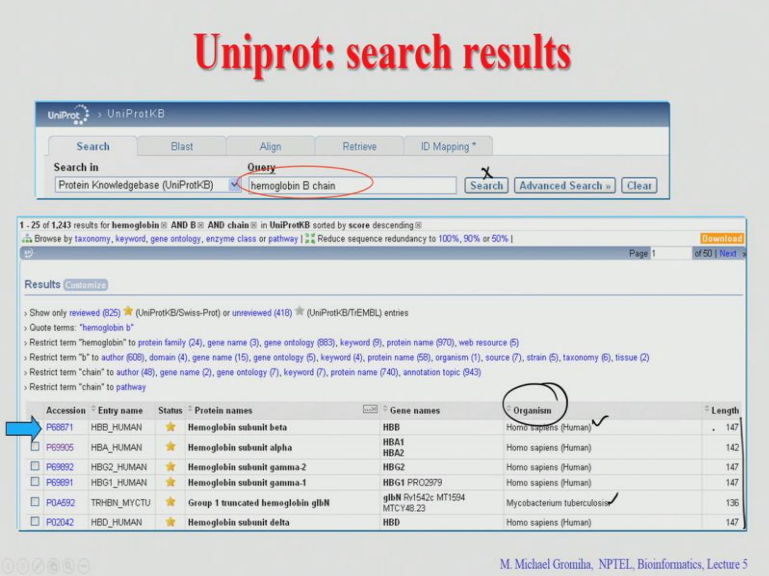
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Then once we give this information then they have to support, this information by other means. So, the data they give in the primary data or supported by the bibliographic references because theories are important to assess the reliability of the data as well as the correctness of the data.

So, all the information they provide in the database right or supported with bibliographic references, where they obtained the original data, original information then they give the cross-reference with a lot of our databases I will show some of the important databases and they give the complete information regarding this particular entries.

So, if you want to search the data for any of these specific proteins right then I will show with one example.

(Refer Slide Time: 01:26)



Uniprot: search results

Search in: Protein Knowledgebase (UniProtKB)

Query: hemoglobin B chain

1 - 25 of 1,243 results for **hemoglobin** ☒ **AND B** ☒ **AND chain** ☒ in **UniProtKB** sorted by **score** descending ☒

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% |

Page 1 of 50 | Next »

Results

› Show only reviewed (825) ★ (UniProtKB/Swiss-Prot) or unreviewed (418) ☆ (UniProtKB/TrEMBL) entries
› Quote terms: "hemoglobin b"
› Restrict term "hemoglobin" to protein family (24), gene name (3), gene ontology (883), keyword (9), protein name (970), web resource (5)
› Restrict term "b" to author (608), domain (4), gene name (15), gene ontology (5), keyword (4), protein name (58), organism (1), source (7), strain (5), taxonomy (6), tissue (2)
› Restrict term "chain" to author (48), gene name (2), gene ontology (7), keyword (7), protein name (740), annotation topic (943)
› Restrict term "chain" to pathway

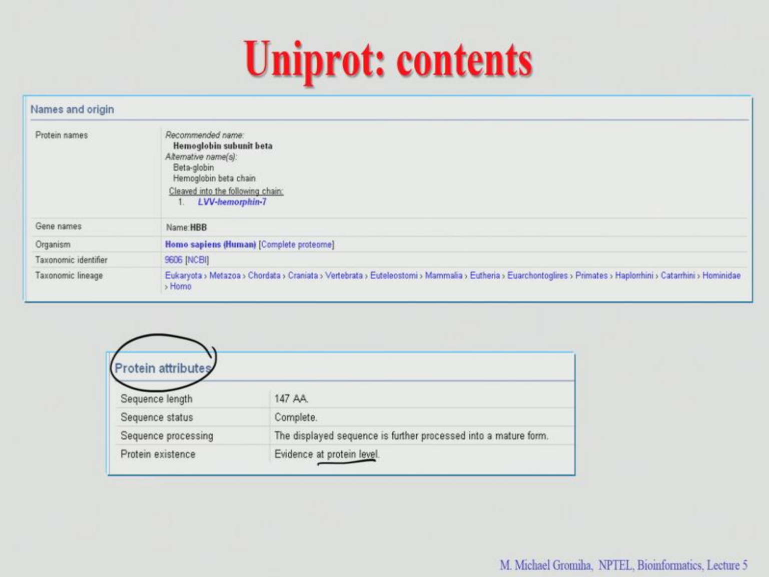
	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
	P68871	HBB_HUMAN	★	**Hemoglobin subunit beta**	**HBB**	Homo sapiens (Human) ✓	147
□	P69905	HBA_HUMAN	★	**Hemoglobin subunit alpha**	**HBA1** **HBA2**	Homo sapiens (Human)	142
□	P69892	HBG2_HUMAN	★	**Hemoglobin subunit gamma-2**	**HBG2**	Homo sapiens (Human)	147
□	P69891	HBG1_HUMAN	★	**Hemoglobin subunit gamma-1**	**HBG1** PRO2979	Homo sapiens (Human)	147
□	P0A592	TRHBN_MYCTU	★	**Group 1 truncated hemoglobin glbN**	**glbN** Rv1542c MT1594 MTCY48.23	Mycobacterium tuberculosis ✓	136
□	P02042	HBD_HUMAN	★	**Hemoglobin subunit delta**	**HBD**	Homo sapiens (Human)	147

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

So, hemoglobin B chain what is the importance of hemoglobin B chain? Yeah, transfer protein, oxygen carrier protein right this is (Refer Time: 01:33) hemoglobin B chain. So, if you click the hemoglobin B chain and click on search typically here. So, this will show you all entries the related with the hemoglobin B chain. So, we saw here. So, these are various accession entries right these are the accession number, this entry name and we have the protein names right we have the subunit beta or subunit alpha, subunit gamma and this is a gene name right and we know the organism sometimes from the human, sometime from this mycobacterium tuberculosis.

So, these are the length of the protein, its 147 amino acid residues in this particular hemoglobin subunit beta.

(Refer Slide Time: 02:06)



Uniprot: contents

Names and origin

Protein names	*Recommended name:* **Hemoglobin subunit beta** *Alternative name(s):* Beta-globin Hemoglobin beta chain Cleaved into the following chain: 1. *LVV-hemorphin-7*
Gene names	Name: **HBB**
Organism	**Homo sapiens (Human)** [Complete proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo

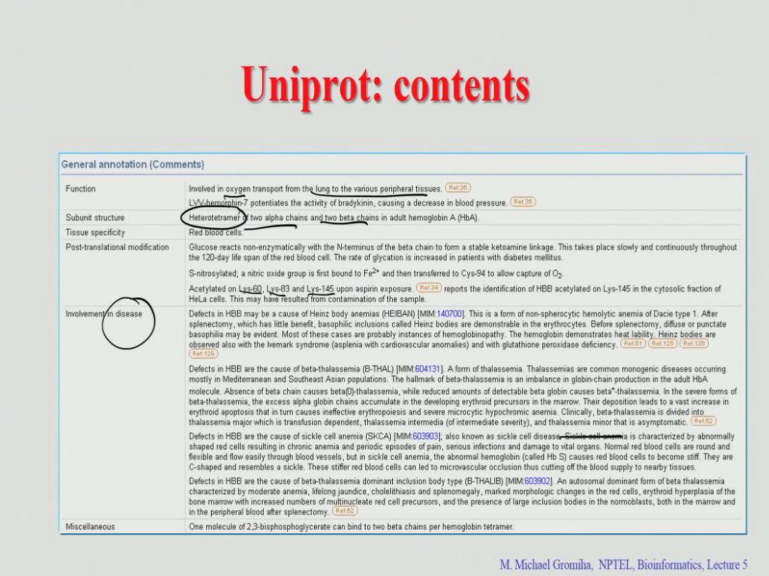
Protein attributes

Sequence length	147 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is further processed into a mature form.
Protein existence	Evidence at protein level.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

So, next they give names right the synonymous names what are the alternate names for these particular protein right. Then give a gene name, organism and so on. Then here you get the protein attributes is 147 amino acid its sequence status is complete because they got all the sequences right and its known at the protein level.

(Refer Slide Time: 02:28)



Uniprot: contents

General annotation (Comments)

Function	Involved in oxygen transport from the lung to the various peripheral tissues. Ref.35 LVV-hemorphin-7 potentiates the activity of bradykinin, causing a decrease in blood pressure. Ref.35
Subunit structure	Heterotetramer of two alpha chains and two beta chains in adult hemoglobin A (HbA).
Tissue specificity	Red blood cells.
Post-translational modification	Glucose reacts non-enzymatically with the N-terminus of the beta chain to form a stable ketoamine linkage. This takes place slowly and continuously throughout the 120-day life span of the red blood cell. The rate of glycation is increased in patients with diabetes mellitus. S-nitrosylated; a nitric oxide group is first bound to Fe^{2+} and then transferred to Cys-94 to allow capture of O_2. Acetylated on Lys-60, Lys-83 and Lys-145 upon aspirin exposure. Ref.34 reports the identification of HBB acetylated on Lys-145 in the cytosolic fraction of HeLa cells. This may have resulted from contamination of the sample.
Involvement in disease	Defects in HBB may be a cause of Heinz body anemias (HEIBAN) [MIM:140700]. This is a form of non-spherocytic hemolytic anemia of Dacie type 1. After splenectomy, which has little benefit, basophilic inclusions called Heinz bodies are demonstrable in the erythrocytes. Before splenectomy, diffuse or punctate basophilia may be evident. Most of these cases are probably instances of hemoglobinopathy. The hemoglobin demonstrates heat lability. Heinz bodies are observed also with the Ivemark syndrome (asplenia with cardiovascular anomalies) and with glutathione peroxidase deficiency. Ref.51 Ref.125 Ref.126 Ref.129 Defects in HBB are the cause of beta-thalassemia (B-THAL) [MIM:604131]. A form of thalassemia. Thalassemias are common monogenic diseases occurring mostly in Mediterranean and Southeast Asian populations. The hallmark of beta-thalassemia is an imbalance in globin-chain production in the adult HbA molecule. Absence of beta chain causes beta(0)-thalassemia, while reduced amounts of detectable beta globin causes $beta^{+}$-thalassemia. In the severe forms of beta-thalassemia, the excess alpha globin chains accumulate in the developing erythroid precursors in the marrow. Their deposition leads to a vast increase in erythroid apoptosis that in turn causes ineffective erythropoiesis and severe microcytic hypochromic anemia. Clinically, beta-thalassemia is divided into thalassemia major which is transfusion dependent, thalassemia intermedia (of intermediate severity), and thalassemia minor that is asymptomatic. Ref.52 Defects in HBB are the cause of sickle cell anemia (SKCA) [MIM:603903]; also known as sickle cell disease~~. Sickle cell anem~~ia is characterized by abnormally shaped red cells resulting in chronic anemia and periodic episodes of pain, serious infections and damage to vital organs. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resembles a sickle. These stiffer red blood cells can led to microvascular occlusion thus cutting off the blood supply to nearby tissues. Defects in HBB are the cause of beta-thalassemia dominant inclusion body type (B-THALIB) [MIM:603902]. An autosomal dominant form of beta thalassemia characterized by moderate anemia, lifelong jaundice, cholelithiasis and splenomegaly, marked morphologic changes in the red cells, erythroid hyperplasia of the bone marrow with increased numbers of multinucleate red cell precursors, and the presence of large inclusion bodies in the normoblasts, both in the marrow and in the peripheral blood after splenectomy. Ref.52
Miscellaneous	One molecule of 2,3-bisphosphoglycerate can bind to two beta chains per hemoglobin tetramer.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

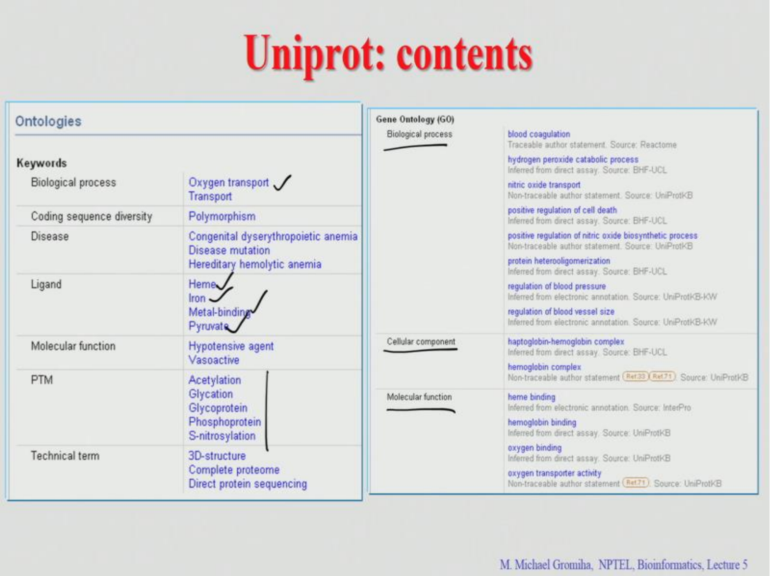
At the protein level, they have the evidence right protein evidence at this protein level. Then you go with general annotation, they try to incorporate try to include most of the data available in the literature.

So, what is the function of this particular protein? They transport oxygen. They transport oxygen from lungs to the tissues. So, this is a major transport. So, then how many subunits or how many small subunits in this particular protein? Totally 4 right because it has 2 alpha chains and 2 beta chains right. We can see the alpha 1, alpha 2, beta 1, beta 2. So, it is the tetramer right. This is heterotetramer right because two different chains this is why hetero, there are 4 chains totally so is a tetramer right.

So, where is it found in tissues? It is in the red blood cells. So, whether any post-translational modifications in this particular protein. So, they mentioned was the glycation and different types of post-translational modifications right at this site.

Mainly the acetylation on Lys 60, Lys 83 and Lys 145 right and they give the proper references also. So, where they have the post-translational modifications; then where this is involved in diseases right they give a lot of information regarding the disease. So, you can see the cause different diseases based on the mutations right. So, they listed up several diseases or we can look all the details right from the UniProt database mainly they give the sickle cell anemia. By the mutation of 6^{th} residue (Refer Time: 03:47) Glutamic acid to valine. So, they mention all the details about the diseases from this particular protein.

(Refer Slide Time: 03:53)



Uniprot: contents

Ontologies

Keywords

Biological process	Oxygen transport ✓ Transport
Coding sequence diversity	Polymorphism
Disease	Congenital dyserythropoietic anemia Disease mutation Hereditary hemolytic anemia
Ligand	Heme ✓ Iron ✓ Metal-binding ✓ Pyruvate ✓
Molecular function	Hypotensive agent Vasoactive
PTM	Acetylation Glycation Glycoprotein Phosphoprotein S-nitrosylation
Technical term	3D-structure Complete proteome Direct protein sequencing

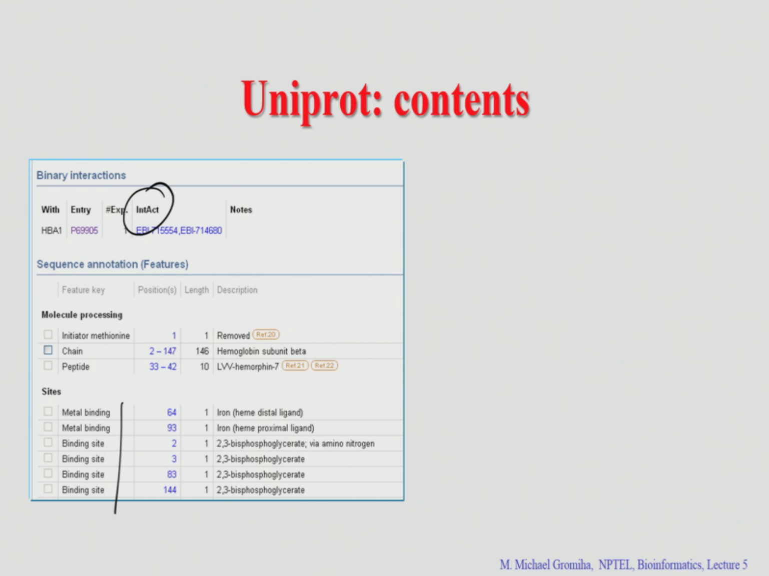
Gene Ontology (GO)

Biological process	blood coagulation Traceable author statement. Source: Reactome hydrogen peroxide catabolic process Inferred from direct assay. Source: BHF-UCL nitric oxide transport Non-traceable author statement. Source: UniProtKB positive regulation of cell death Inferred from direct assay. Source: BHF-UCL positive regulation of nitric oxide biosynthetic process Non-traceable author statement. Source: UniProtKB protein heterooligomerization Inferred from direct assay. Source: BHF-UCL regulation of blood pressure Inferred from electronic annotation. Source: UniProtKB-KW regulation of blood vessel size Inferred from electronic annotation. Source: UniProtKB-KW
Cellular component	haptoglobin-hemoglobin complex Inferred from direct assay. Source: BHF-UCL hemoglobin complex Non-traceable author statement Ref.33 Ref.71. Source: UniProtKB
Molecular function	heme binding Inferred from electronic annotation. Source: InterPro hemoglobin binding Inferred from direct assay. Source: UniProtKB oxygen binding Inferred from direct assay. Source: UniProtKB oxygen transporter activity Non-traceable author statement Ref.71. Source: UniProtKB

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

So, now they give the ontologies what are the various biological process we see oxygen transports. So, any diversity of this coding region, they have polymorphism right they have the changes upon mutations or either any other ligands or any other small molecules they bind to this particular protein. So, what are the different small molecules binds to this protein? Heme, iron, metals, and Pyruvate right these are the different ligands they bind with this protein. So, what are the functions and what are the post-translational modifications occurred in this particular protein. So, various types of PTMs: acetylation, glycation, glycoprotein, phosphoprotein and s nitrosylation the various types of TMs, then we go with the ontology. So, there are different types of gene ontology. They are widely used in the literature, one is a biological process cellular component and molecular function. So, they give the different specific subclasses in this different gene ontological process.

(Refer Slide Time: 04:53)



Uniprot: contents

Binary interactions

With	Entry	#Exp.	IntAct	Notes
HBA1	P69905	1	EBI-715554,EBI-714680	

Sequence annotation (Features)

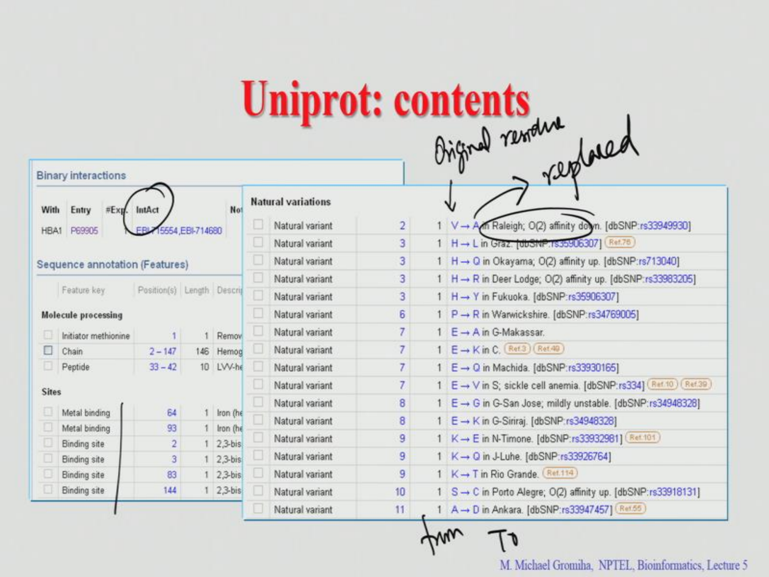
	Feature key	Position(s)	Length	Description
Molecule processing				
□	Initiator methionine	1	1	Removed Ref.20
□	Chain	2 – 147	146	Hemoglobin subunit beta
□	Peptide	33 – 42	10	LVV-hemorphin-7 Ref.21 Ref.22
Sites				
□	Metal binding	64	1	Iron (heme distal ligand)
□	Metal binding	93	1	Iron (heme proximal ligand)
□	Binding site	2	1	2,3-bisphosphoglycerate; via amino nitrogen
□	Binding site	3	1	2,3-bisphosphoglycerate
□	Binding site	83	1	2,3-bisphosphoglycerate
□	Binding site	144	1	2,3-bisphosphoglycerate

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Then you go with this interactions, they have the interaction in other proteins. So, this the database for the protein interactions that intact.

So, they give the interaction with the other proteins, how they are interacting with the different other proteins you had different sites. What are the metal binding sites, where different binding sites they give the information regarding the binding site of a particular protein?

(Refer Slide Time: 05:14)



Uniprot: contents

Original residue

replaced

Binary interactions

With	Entry	#Exp.	IntAct	Not
HBA1	P69905	1	EBI-715554,EBI-714680	

Sequence annotation (Features)

	Feature key	Position(s)	Length	Descri
	Molecule processing			
□	Initiator methionine	1	1	Remov
□	Chain	2 – 147	146	Hemog
□	Peptide	33 – 42	10	LVV-he
	Sites			
□	Metal binding	64	1	Iron (he
□	Metal binding	93	1	Iron (he
□	Binding site	2	1	2,3-bis
□	Binding site	3	1	2,3-bis
□	Binding site	83	1	2,3-bis
□	Binding site	144	1	2,3-bis

Natural variations

□	Natural variant	2	1	V → A in Raleigh; O(2) affinity down. [dbSNP:rs33949930]
□	Natural variant	3	1	H → L in Graz. [dbSNP:rs35906307] Ref.76
□	Natural variant	3	1	H → Q in Okayama; O(2) affinity up. [dbSNP:rs713040]
□	Natural variant	3	1	H → R in Deer Lodge; O(2) affinity up. [dbSNP:rs33983205]
□	Natural variant	3	1	H → Y in Fukuoka. [dbSNP:rs35906307]
□	Natural variant	6	1	P → R in Warwickshire. [dbSNP:rs34769005]
□	Natural variant	7	1	E → A in G-Makassar.
□	Natural variant	7	1	E → K in C. Ref.3 Ref.49
□	Natural variant	7	1	E → Q in Machida. [dbSNP:rs33930165]
□	Natural variant	7	1	E → V in S; sickle cell anemia. [dbSNP:rs334] Ref.10 Ref.39
□	Natural variant	8	1	E → G in G-San Jose; mildly unstable. [dbSNP:rs34948328]
□	Natural variant	8	1	E → K in G-Siriraj. [dbSNP:rs34948328]
□	Natural variant	9	1	K → E in N-Timone. [dbSNP:rs33932981] Ref.101
□	Natural variant	9	1	K → Q in J-Luhe. [dbSNP:rs33926764]
□	Natural variant	9	1	K → T in Rio Grande. Ref.114
□	Natural variant	10	1	S → C in Porto Alegre; O(2) affinity up. [dbSNP:rs33918131]
□	Natural variant	11	1	A → D in Ankara. [dbSNP:rs33947457] Ref.55

from To

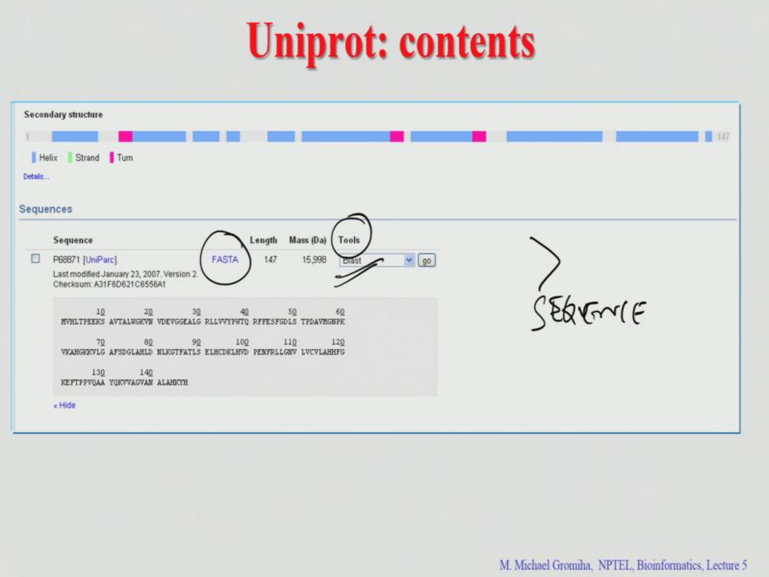
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Now the variants, there are various variants in this particular protein right. So, they give the wild-type residue, this is the residue which are real in the original residue in the protein.

This is the residue which is replaced. So, this is replaced, for example, valine is replaced to alanine, then they will happen to what is the major disease. So, here give you all the mutations ‘from’ this is ‘to’ and which type of diseases, and they give the references. Then they can give all the information essentially if you see UniProt is the unique resource which contains all the information regarding any particular protein right.

Now, you go with the next step, now still now they mention about a general information right what are the functions and what are the gene ontology, what are the binding sites what are the PTM sites all the information they give. Now they go with the sequence level right the major aspect of this UniProt. So, UniProt contains not only the sequence, it contains the data for other different functions and the structure of a particular protein.

(Refer Slide Time: 06:22)



Uniprot: contents

Secondary structure

Sequences

	Sequence		Length	Mass (Da)	Tools
□	P68871 [UniParc]. Last modified January 23, 2007. Version 2. Checksum: A31F6D621C6556A1	FASTA	147	15,998	Blast go

```
        10         20         30         40         50         60
MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPK

        70         80         90        100        110        120
VKAHGKKVLG AFSDGLAHLD NLKGTFATLS ELHCDKLHVD PENFRLLGNV LVCVLAHHFG

       130        140
KEFTPPVQAA YQKVVAGVAN ALAHKYH
```

« Hide

>

SEQUENCE

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

So, here this is the sequence. How many residues is in this protein? 41, 42, 43, 44, 45, 46, 47 they have 147 residues right. So, it gives the complete sequence right and here they give the secondary structure because as you know this protein right this is the alpha-helical protein this is predominantly with the alpha helices, I will discuss the secondary structure in the later classes.

So, if you see the blue once they are mainly helices right and the green one strands, but here it does not have any strand and we have some turns here and there we can see some turns right. This is the secondary structure information regarding their particular protein. So, you can see this sequence in two different formats, either you can see this is the UniProt format and also you can see the FASTA format what is FASTA format? FASTA format is a format which starts with a greater than symbol, and here then we can see this is the command line and here the sequence will start.

So, this is a kind of format which you adopt in bioinformatics generally for the treating the sequences as well as for our large-scale analysis. So, to separate two sequences they use this specific format. Then you can also see if you see a sequence any other sequences in any other organisms they have similar to this sequence right. So, there are various tools if you see here there are various tools available, one is BLAST I will discuss the details in later classes right this will help you to see the proteins, which are related with your original protein sequence.

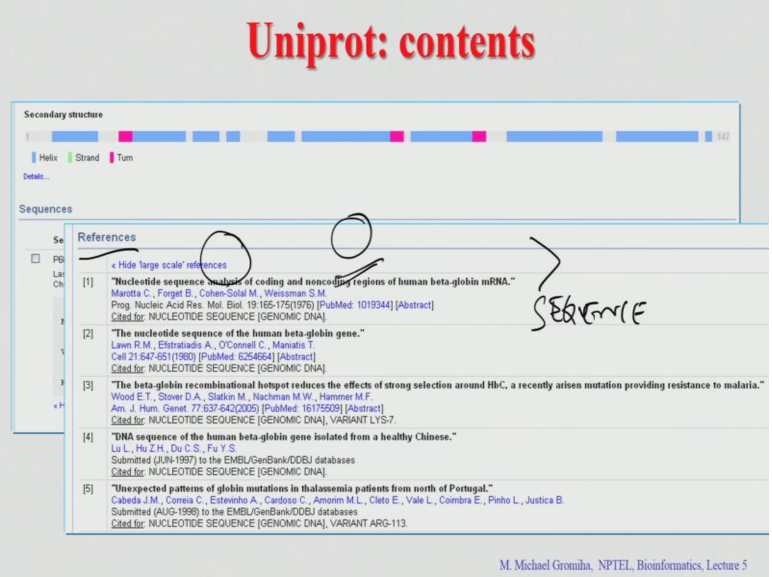
So, you have your own sequence, if you click on this blast this will give you what are the other sequences related with your own sequence. So, now, we have a reference because this is an important part, because how they obtain the information. Because they cannot get directly from again is from outside right that is why we have to use any reliable resources. So, only one major resource reliable resource, so that is the published articles, where shall we get this information.

Student: PUBMED.

PUBMED database right now we discussed in the previous class previous classes right. So, PUBMED provides the resource for all the published articles mainly in biology right and medicine. So, we have this different papers we can get the information from PUBMED and they integrate all the data in the database and wherever they collect the information from the literature. So, they give the references.

(Refer Slide Time: 18:42)

Uniprot: contents



References

« Hide 'large scale' references

[1] "Nucleotide sequence analysis of coding and noncoding regions of human beta-globin mRNA."
Marotta C., Forget B., Cohen-Solal M., Weissman S.M.
Prog. Nucleic Acid Res. Mol. Biol. 19:165-175(1976) [PubMed: 1019344] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA].

[2] "The nucleotide sequence of the human beta-globin gene."
Lawn R.M., Efstratiadis A., O'Connell C., Maniatis T.
Cell 21:647-651(1980) [PubMed: 6254664] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA].

[3] "The beta-globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria."
Wood E.T., Stover D.A., Slatkin M., Nachman M.W., Hammer M.F.
Am. J. Hum. Genet. 77:637-642(2005) [PubMed: 16175509] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA], VARIANT LYS-7.

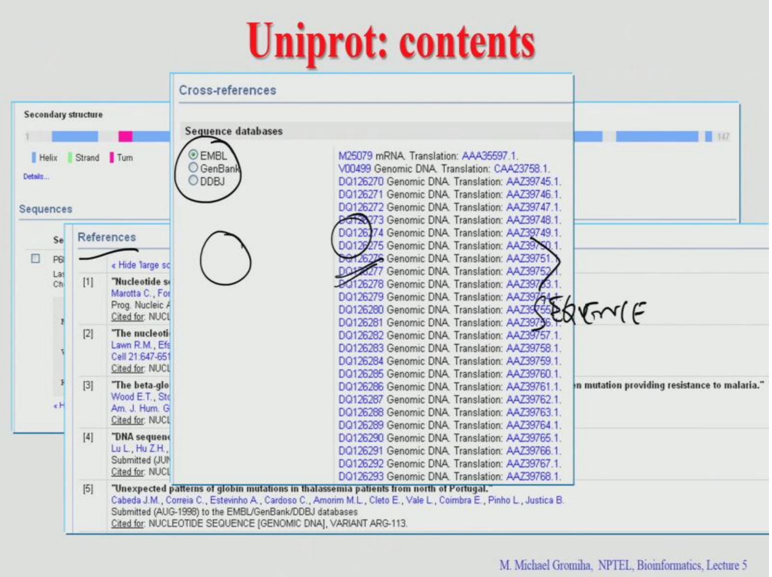
[4] "DNA sequence of the human beta-globin gene isolated from a healthy Chinese."
Lu L., Hu Z.H., Du C.S., Fu Y.S.
Submitted (JUN-1997) to the EMBL/GenBank/DDBJ databases
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA].

[5] "Unexpected patterns of globin mutations in thalassemia patients from north of Portugal."
Cabeda J.M., Correia C., Estevinho A., Cardoso C., Amorim M.L., Cleto E., Vale L., Coimbra E., Pinho L., Justica B.
Submitted (AUG-1998) to the EMBL/GenBank/DDBJ databases
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA], VARIANT ARG-113.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Right these are the references. To provide all the information its very time-consuming. So, manual curation is a very hard this is a reason why the manually curated sequences are less compared with the computer translated sequences. So, where they can collect the information by keyword searching and they put the collect the data from the resources and put it as it is right there we do not we have to work on that, but what the reliability. So, we need to do the manual curation.

(Refer Slide Time: 09:07)



Uniprot: contents

Cross-references

Sequence databases

● EMBL
○ GenBank
○ DDBJ

M25079 mRNA. Translation: AAA35597.1.
V00499 Genomic DNA. Translation: CAA23758.1.
DQ126270 Genomic DNA. Translation: AAZ39745.1.
DQ126271 Genomic DNA. Translation: AAZ39746.1.
DQ126272 Genomic DNA. Translation: AAZ39747.1.
DQ126273 Genomic DNA. Translation: AAZ39748.1.
DQ126274 Genomic DNA. Translation: AAZ39749.1.
DQ126275 Genomic DNA. Translation: AAZ39750.1.
DQ126276 Genomic DNA. Translation: AAZ39751.1.
DQ126277 Genomic DNA. Translation: AAZ39752.1.
DQ126278 Genomic DNA. Translation: AAZ39753.1.
DQ126279 Genomic DNA. Translation: AAZ39754.1.
DQ126280 Genomic DNA. Translation: AAZ39755.1.
DQ126281 Genomic DNA. Translation: AAZ39756.1.
DQ126282 Genomic DNA. Translation: AAZ39757.1.
DQ126283 Genomic DNA. Translation: AAZ39758.1.
DQ126284 Genomic DNA. Translation: AAZ39759.1.
DQ126285 Genomic DNA. Translation: AAZ39760.1.
DQ126286 Genomic DNA. Translation: AAZ39761.1.
DQ126287 Genomic DNA. Translation: AAZ39762.1.
DQ126288 Genomic DNA. Translation: AAZ39763.1.
DQ126289 Genomic DNA. Translation: AAZ39764.1.
DQ126290 Genomic DNA. Translation: AAZ39765.1.
DQ126291 Genomic DNA. Translation: AAZ39766.1.
DQ126292 Genomic DNA. Translation: AAZ39767.1.
DQ126293 Genomic DNA. Translation: AAZ39768.1.

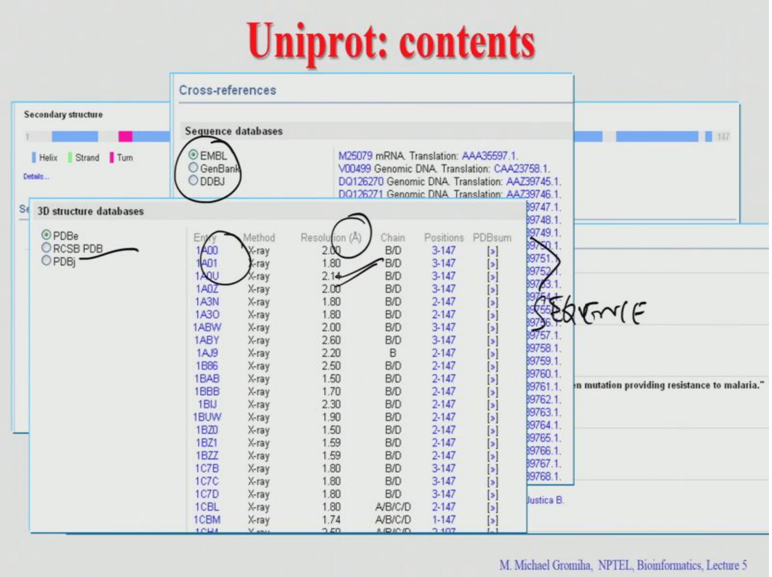
[5] "Unexpected patterns of globin mutations in thalassemia patients from north of Portugal."
Cabeda J.M., Correia C., Estevinho A., Cardoso C., Amorim M.L., Cleto E., Vale L., Coimbra E., Pinho L., Justica B.
Submitted (AUG-1998) to the EMBL/GenBank/DDBJ databases
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA], VARIANT ARG-113.

SEQUENCE

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

So, now give the sequence databases right we can see the; what EMBL Genbank and DDBJ which sequence database? Nucleic acid sequence database, where they give the complete sequence databases and the translated sequences they provide the information.

(Refer Slide Time: 09:21)



Uniprot: contents

Cross-references

Sequence databases

● EMBL
○ GenBank
○ DDBJ

3D structure databases

● PDBe
○ RCSB PDB
○ PDBj

Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
1A00	X-ray	2.00	B/D	3-147	[»]
1A01	X-ray	1.80	B/D	3-147	[»]
1A0U	X-ray	2.14	B/D	3-147	[»]
1A0Z	X-ray	2.00	B/D	3-147	[»]
1A3N	X-ray	1.80	B/D	2-147	[»]
1A3O	X-ray	1.80	B/D	2-147	[»]
1ABW	X-ray	2.00	B/D	3-147	[»]
1ABY	X-ray	2.60	B/D	3-147	[»]
1AJ9	X-ray	2.20	B	2-147	[»]
1B86	X-ray	2.50	B/D	2-147	[»]
1BAB	X-ray	1.50	B/D	2-147	[»]
1BBB	X-ray	1.70	B/D	2-147	[»]
1BIJ	X-ray	2.30	B/D	2-147	[»]
1BUW	X-ray	1.90	B/D	2-147	[»]
1BZ0	X-ray	1.50	B/D	2-147	[»]
1BZ1	X-ray	1.59	B/D	2-147	[»]
1BZZ	X-ray	1.59	B/D	2-147	[»]
1C7B	X-ray	1.80	B/D	3-147	[»]
1C7C	X-ray	1.80	B/D	3-147	[»]
1C7D	X-ray	1.80	B/D	3-147	[»]
1CBL	X-ray	1.80	A/B/C/D	2-147	[»]
1CBM	X-ray	1.74	A/B/C/D	1-147	[»]

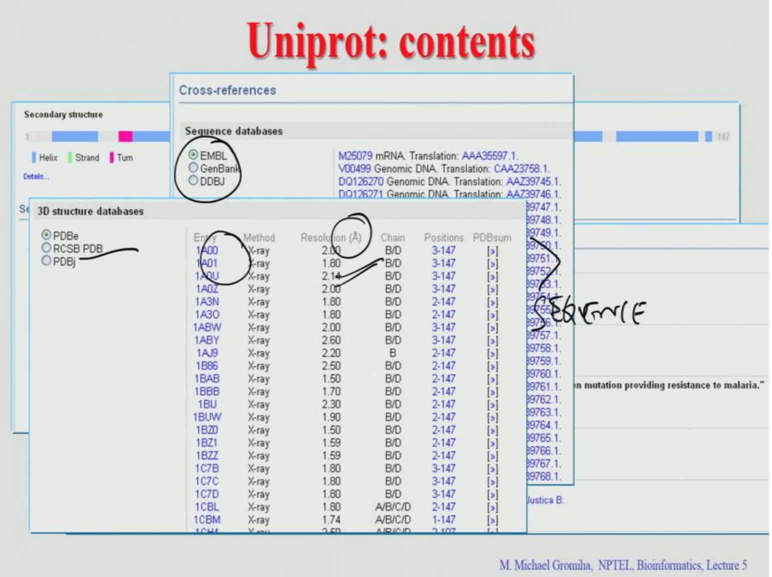
SEQUENCE

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Then this is the 3D structures. So, this is a method how they obtained a data right and the resolution and the positions and so on.

More details I will discuss in later classes.

(Refer Slide Time: 09:32)



Uniprot: contents

Cross-references

Sequence databases

● EMBL
○ GenBank
○ DDBJ

M25079 mRNA. Translation: AAA35597.1.
V00499 Genomic DNA. Translation: CAA23758.1.
DQ126270 Genomic DNA. Translation: AAZ39745.1.
DQ126271 Genomic DNA. Translation: AAZ39746.1.

3D structure databases

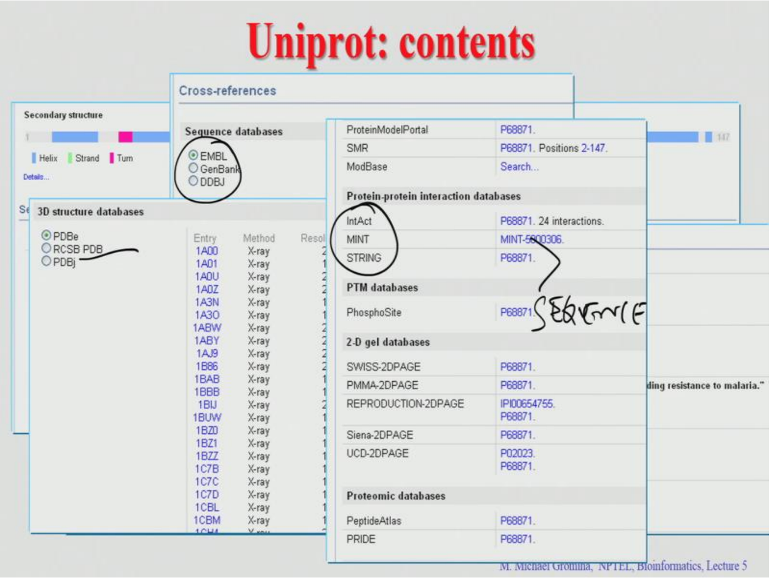
● PDBe
○ RCSB PDB
○ PDBj

Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
1A00	X-ray	2.00	B/D	3-147	[»]
1A01	X-ray	1.80	B/D	3-147	[»]
1A0U	X-ray	2.14	B/D	3-147	[»]
1A0Z	X-ray	2.00	B/D	3-147	[»]
1A3N	X-ray	1.80	B/D	2-147	[»]
1A3O	X-ray	1.80	B/D	2-147	[»]
1ABW	X-ray	2.00	B/D	3-147	[»]
1ABY	X-ray	2.60	B/D	3-147	[»]
1AJ9	X-ray	2.20	B	2-147	[»]
1B86	X-ray	2.50	B/D	2-147	[»]
1BAB	X-ray	1.50	B/D	2-147	[»]
1BBB	X-ray	1.70	B/D	2-147	[»]
1BIJ	X-ray	2.30	B/D	2-147	[»]
1BUW	X-ray	1.90	B/D	2-147	[»]
1BZ0	X-ray	1.50	B/D	2-147	[»]
1BZ1	X-ray	1.59	B/D	2-147	[»]
1BZZ	X-ray	1.59	B/D	2-147	[»]
1C7B	X-ray	1.80	B/D	3-147	[»]
1C7C	X-ray	1.80	B/D	3-147	[»]
1C7D	X-ray	1.80	B/D	3-147	[»]
1CBL	X-ray	1.80	A/B/C/D	2-147	[»]
1CBM	X-ray	1.74	A/B/C/D	1-147	[»]

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Right.

(Refer Slide Time: 09:34)



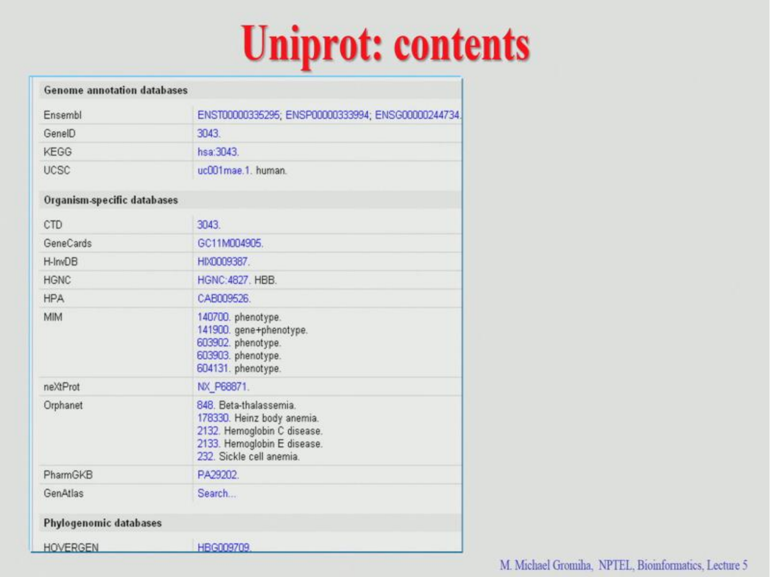
Uniprot: contents

ProteinModelPortal	P68871.
SMR	P68871. Positions 2-147.
ModBase	Search...
Protein-protein interaction databases	
IntAct	P68871. 24 interactions.
MINT	MINT-5000306.
STRING	P68871.
PTM databases	
PhosphoSite	P68871.
2-D gel databases	
SWISS-2DPAGE	P68871.
PMMA-2DPAGE	P68871.
REPRODUCTION-2DPAGE	IPI00654755. P68871.
Siena-2DPAGE	P68871.
UCD-2DPAGE	P02023. P68871.
Proteomic databases	
PeptideAtlas	P68871.
PRIDE	P68871.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

So, these are the protein-protein interaction databases, MINT, IntAct, and STRING. So, it will give you the particular proteins and how this interact with other proteins. Then they give other protein databases and the post translational modification database and so on right.

(Refer Slide Time: 09:49)



Uniprot: contents

Genome annotation databases	
Ensembl	ENST00000335295; ENSP00000333994; ENSG00000244734.
GeneID	3043.
KEGG	hsa:3043.
UCSC	uc001mae.1. human.
Organism-specific databases	
CTD	3043.
GeneCards	GC11M004905.
H-InvDB	HIX0009387.
HGNC	HGNC:4827. HBB.
HPA	CAB009526.
MIM	140700. phenotype. 141900. gene+phenotype. 603902. phenotype. 603903. phenotype. 604131. phenotype.
neXtProt	NX_P68871.
Orphanet	848. Beta-thalassemia. 178330. Heinz body anemia. 2132. Hemoglobin C disease. 2133. Hemoglobin E disease. 232. Sickle cell anemia.
PharmGKB	PA29202.
GenAtlas	Search...
Phylogenomic databases	
HOVERGEN	HBG009709.

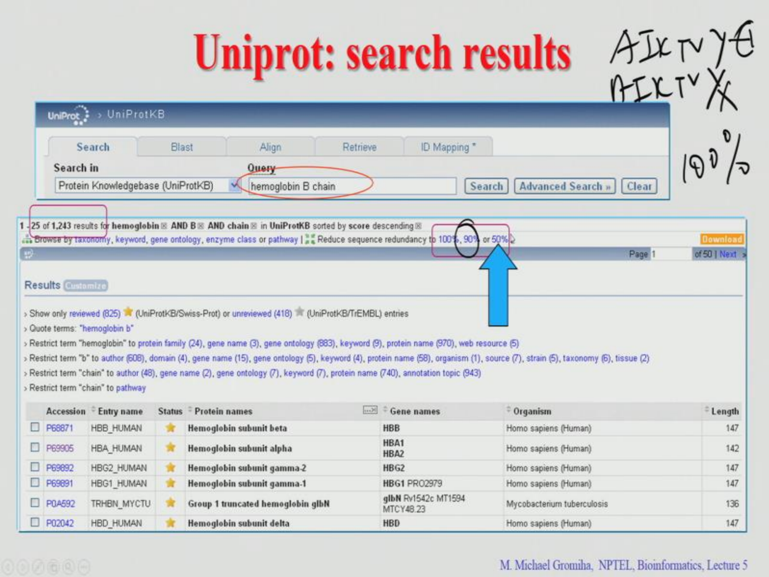
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Then also they provide genome annotations, organism-specific database as well as the phylogenomic databases. This is the one part; the first part is a general information, the second part is a sequence and the structure and the functional information plus the links and now they give the different one type annotations.

They give the enzyme in pathways, where they have this reaction and gene expression databases right and family domain database and as well as the other resources, for example, the drug bank and other resources. Fine, then they have the other relevant documents like chromosome and other polymorphism disease mutation and so on.

Now, you are familiar with the all the contents of UniProt database right when you have time you will look into the UniProt and take your own protein and if you search and if you read, then you will be comfortable with understanding all the aspects of a particular protein. So, here the hemoglobin B chain, in the first part we give all the functional information regarding the general information and second part with the protein sequence along with the secondary structure or the tertiary structure and the links with other databases, and the last part they give the enzymes and pathways and the interactions and so on. If you look search with the hemoglobin B chain, for example, it will give you lot of data. So, for example, if you see this, many data and from that some of them are redundant. If you see the 2 sequences there are sometimes 90 percent identity, sometimes 80 percent identity right.

(Refer Slide Time: 11:14)



Uniprot: search results

AIKTVYE
AIKTVYK

100%

Search in: Protein Knowledgebase (UniProtKB)

Query: hemoglobin B chain

1 - 25 of 1,243 results for **hemoglobin** ☒ **AND B** ☒ **AND chain** ☒ in **UniProtKB** sorted by **score** descending ☒

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50%

› Show only reviewed (825) ★ (UniProtKB/Swiss-Prot) or unreviewed (418) ★ (UniProtKB/TrEMBL) entries
› Quote terms: "hemoglobin b"
› Restrict term "hemoglobin" to protein family (24), gene name (3), gene ontology (883), keyword (9), protein name (970), web resource (5)
› Restrict term "b" to author (608), domain (4), gene name (15), gene ontology (5), keyword (4), protein name (58), organism (1), source (7), strain (5), taxonomy (6), tissue (2)
› Restrict term "chain" to author (48), gene name (2), gene ontology (7), keyword (7), protein name (740), annotation topic (943)
› Restrict term "chain" to pathway

	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
□	P68871	HBB_HUMAN	★	**Hemoglobin subunit beta**	**HBB**	Homo sapiens (Human)	147
□	P69905	HBA_HUMAN	★	**Hemoglobin subunit alpha**	**HBA1** **HBA2**	Homo sapiens (Human)	142
□	P69892	HBG2_HUMAN	★	**Hemoglobin subunit gamma-2**	**HBG2**	Homo sapiens (Human)	147
□	P69891	HBG1_HUMAN	★	**Hemoglobin subunit gamma-1**	**HBG1** PRO2979	Homo sapiens (Human)	147
□	P0A592	TRHBN_MYCTU	★	**Group 1 truncated hemoglobin glbN**	**glbN** Rv1542c MT1594 MTCY48.23	Mycobacterium tuberculosis	136
□	P02042	HBD_HUMAN	★	**Hemoglobin subunit delta**	**HBD**	Homo sapiens (Human)	147

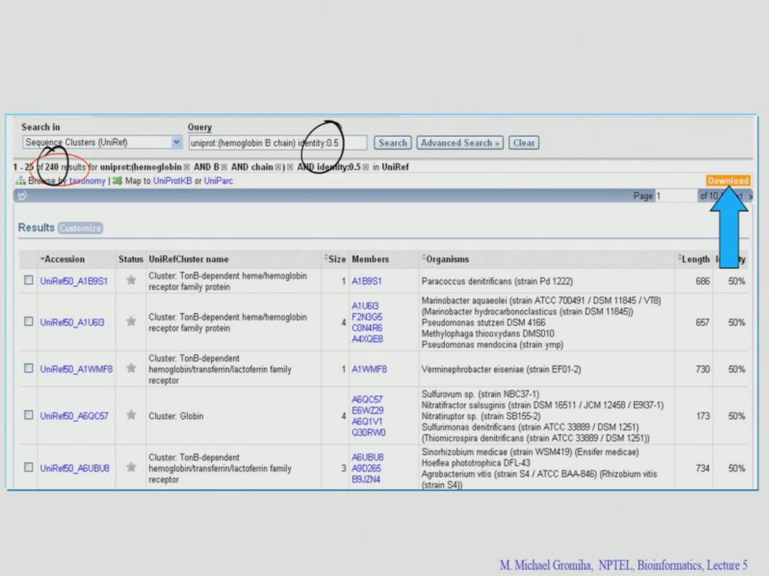
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

For example, if you have these two sequences, what is the identity? 100 percent right. So, we can see the 100 percent identity. So, in this case, this UniProt provides an option to select these sequences with some level of redundancy for example, if you want to reduce to 90 percent; that means, if two sequences if the similarities identity is less than 90 percent you can get, but more than 90 percent you will take one and discard this one. So, also you can get the redundant reduce redundancy up to 50 percent. They will check the two sequences and see the identity if it is more than 50 percent they will keep one and discard the other.

How to reduce this redundancy I will explain in the subsequent classes right. So, you can get that if you do this here now for example, currently you get 1243 results, when you look into this specific redundancy right now we can get 240 results. So, here we use the identity of 0.5. So, we reduce the data from 1243 to 240. Because to do any analysis it is very important to have the data from different ways, not to use the same several times because you will introduce bias because of the same sequence right. So, in this case, it is important to have a non-redundant data. So, here you can get with the two different cutoffs right.

For example, this is 100 percent, 90 percent, and 50 percent right, if you want to have other redundancies we have to use other programs available in the literature.

(Refer Slide Time: 12:50)



Search in: Sequence Clusters (UniRef)

Query: uniprot:(hemoglobin B chain) identity:0.5

Search | Advanced Search » | Clear

1 - 25 of 240 results for **uniprot:(hemoglobin ☒ AND B ☒ AND chain ☒) ☒ AND identity:0.5 ☒** in **UniRef**

Browse by taxonomy | Map to UniProtKB or UniParc

Download

Page 1 of 10

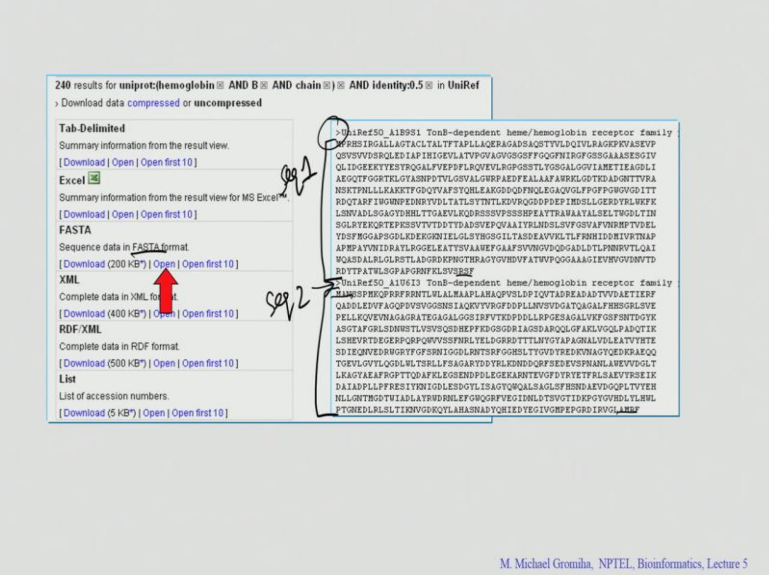
Results Customize

	Accession	Status	UniRefCluster name	Size	Members	Organisms	Length	Identity
□	UniRef50_A1B9S1	★	Cluster: TonB-dependent heme/hemoglobin receptor family protein	1	A1B9S1	Paracoccus denitrificans (strain Pd 1222)	686	50%
□	UniRef50_A1U6I3	★	Cluster: TonB-dependent heme/hemoglobin receptor family protein	4	A1U6I3 F2N3G5 C0N4R6 A4XQE8	Marinobacter aquaeolei (strain ATCC 700491 / DSM 11845 / VT8) (Marinobacter hydrocarbonoclasticus (strain DSM 11845)) Pseudomonas stutzeri DSM 4166 Methylophaga thiooxydans DMS010 Pseudomonas mendocina (strain ymp)	657	50%
□	UniRef50_A1WMF8	★	Cluster: TonB-dependent hemoglobin/transferrin/lactoferrin family receptor	1	A1WMF8	Verminephrobacter eiseniae (strain EF01-2)	730	50%
□	UniRef50_A6QC57	★	Cluster: Globin	4	A6QC57 E6WZ29 A6Q1V1 Q30RW0	Sulfurovum sp. (strain NBC37-1) Nitratifractor salsuginis (strain DSM 16511 / JCM 12458 / E9I37-1) Nitratiruptor sp. (strain SB155-2) Sulfurimonas denitrificans (strain ATCC 33889 / DSM 1251) (Thiomicrospira denitrificans (strain ATCC 33889 / DSM 1251))	173	50%
□	UniRef50_A6UBU8	★	Cluster: TonB-dependent hemoglobin/transferrin/lactoferrin family receptor	3	A6UBU8 A9D265 B9JZN4	Sinorhizobium medicae (strain WSM419) (Ensifer medicae) Hoeflea phototrophica DFL-43 Agrobacterium vitis (strain S4 / ATCC BAA-846) (Rhizobium vitis (strain S4))	734	50%

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

So, these are the entries now you have 240 results. So, these are the different entries right obtained from this particular search.

(Refer Slide Time: 12:57)



240 results for **uniprot:(hemoglobin ☒ AND B ☒ AND chain ☒) ☒ AND identity:0.5 ☒** in **UniRef**

› Download data compressed or **uncompressed**

Tab-Delimited

Summary information from the result view.

[Download | Open | Open first 10]

Excel

Summary information from the result view for MS Excel™.

[Download | Open | Open first 10]

FASTA

Sequence data in FASTA format.

[Download (200 KB*) | Open | Open first 10]

XML

Complete data in XML format.

[Download (400 KB*) | Open | Open first 10]

RDF/XML

Complete data in RDF format.

[Download (500 KB*) | Open | Open first 10]

List

List of accession numbers.

[Download (5 KB*) | Open | Open first 10]

```
>UniRef50_A1B9S1 TonB-dependent heme/hemoglobin receptor family
MPRHSIRGALLAGTACLTALTFTAPLLAQERAGADSAQSTYVLDQIVLRAGKPKVASEVP
QSVSVVDSRQLEDIAPIHIGEVLATVPGVAGVGSGSFFGQGFNIRGFGSSGAAASESGIV
QLIDGEEKYYESYRQGALFVEPDFLRQVEVLRGPGSSTLYGSGALGGVIAMETIEAGDLI
AEGQTFGGRTKLGYASNPDTVLGSVALGWRPAEDFEALAAFAWRKLGDTKDADGNTTVRA
NSKTPNLLLKAKKTFGDQYVAFSYQHLEAKGDDQDFNQLEGAQVGLFPGFPGWGVGDITT
RDQTARFIWGWNPEDNRYVDLTATLSYTNTLKDVRQGDDPDEPIMDSLLGERDYRLWKFK
LSNVADLSGAGYDHHLTTGAEVLKQDRSSSVPSSSHPEAYTRAWAAYALSELTWGDLTIN
SGLRYEKQRTEPKSSVTVTDDTYDADSVEPQVAAIYRLNDSLSVFGSVAFVNRMPTVDEL
YDSFMGGAPSGDLKDEKGKNIELGLSYHGSGILTASDEAVVKLTLFRNHIDDMIVRTNAP
APMPAYVNIDRAYLRGGELEATYSVAAWEFGAAFSVVNGVDQDGADLDTLPNNRVTLQAI
WQASDALRLGLRSTLADGRDKPNGTHRAGYGVHDVFATWVPQGGAAAGIEVHVGVDNVTD
RDYTPATWLSGPAPGRNFKLSVSRSF
>UniRef50_A1U6I3 TonB-dependent heme/hemoglobin receptor family
MANSSPMKQPRRFRRNTLWLALMAAPLAHAQPVSLDPIQVTADREADADTVVDAETIERF
QADDLEDVFAGQPDVSVGGSNSIAQKVYVRGFDDPLLNVSVDGATQAGALFHHSGRLSVE
PELLKQVEVNAGAGRATEGAGALGGSIRFVTKDPDDLLRPGESAGALVKFGSFSNTDGYK
ASGTAFGRLSDNWSTLVSVSQSDHEPFKDGSGDRIAGSDARQQLGFAKLVGQLPADQTIK
LSHEVRTDEGERPQRPQWVVSSFNRLYELDGRRDTTTLNYGYAPAGNALVDLEATVYHTE
SDIEQNVEDRWGRYFGFSRNIGGDLRNTSRFGGHSLTYGVDYREDKVNAGYQEDKRAEQQ
TGEVLGVYLQGDLWLTSRLLFSAGARYDDYRLKDNDDQRFSEDEVSPNANLAWEVVDGLT
LKAGYAEAFRGPTTQDAFKLEGSENDPDLEGEKARNTEVGFDYRYETFRLSAEVYRSEIK
DAIADPLLPFRESIYKNIGDLESDGYLISAGYQWQALSAGLSFHSNDAEVDGQPLTVYEH
NLLGNTMGDTWIADLAYRWDRNLEFGWQGRFVEGIDNLDTSVGTIDKPGYGVHDLYLHWL
PTGNEDLRLSLTIKNVGDKQYLAHASNADYQHIEDYEGIVGMPEPGRDIRVGLAMRF
```

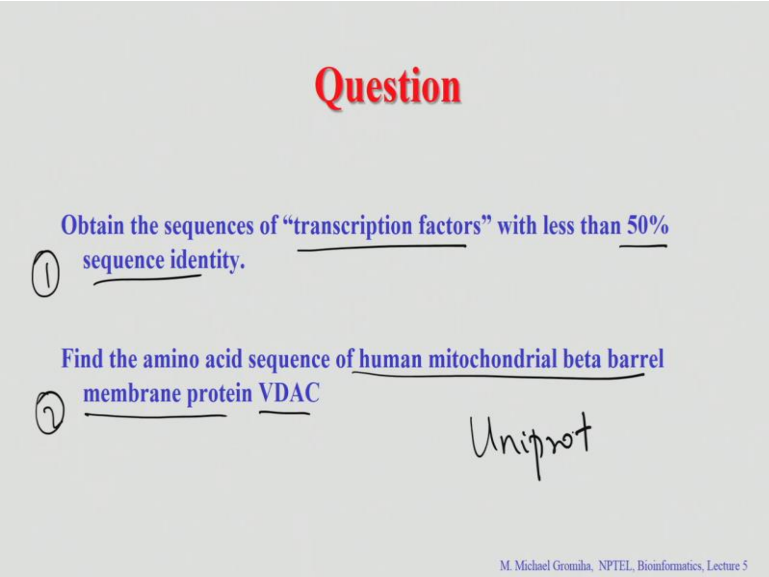
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Now we can also download the data the various options to download the data right they have the various options. For example, you can see a tab delimited and you can use excel, you can download in FASTA and you can download XML and you can list the accession numbers right as well as you can get the XML or RDF format.

So, various formats you can download the data. So, if you give the FASTA format right. So, you give the FASTA format this is the result. So, as we discussed earlier in the FASTA format its start with the greater than symbol and this is the name and followed by the sequence you can see from here to here this amino acid sequence, sequence number 1 and here again the greater than symbol started this is sequence 1 right and here this symbol started again. So, it is started with sequence 2. So, and this is a sequence.

So, you can get the all the 240 results based on the greater than symbol, you can understand the first sequence starts from' and end with the 'RSF'. The second sequence started from 'MAN' and ended with this 'AMRF'. Then you can separate data and you can use these data for the analysis. Now I have two questions. The first question is obtain the sequences of transcription factors with less than 50 percent sequence identity.

(Refer Slide Time: 14:10)



Question

1 Obtain the sequences of "transcription factors" with less than 50% sequence identity.

2 Find the amino acid sequence of human mitochondrial beta barrel membrane protein VDAC

Uniprot

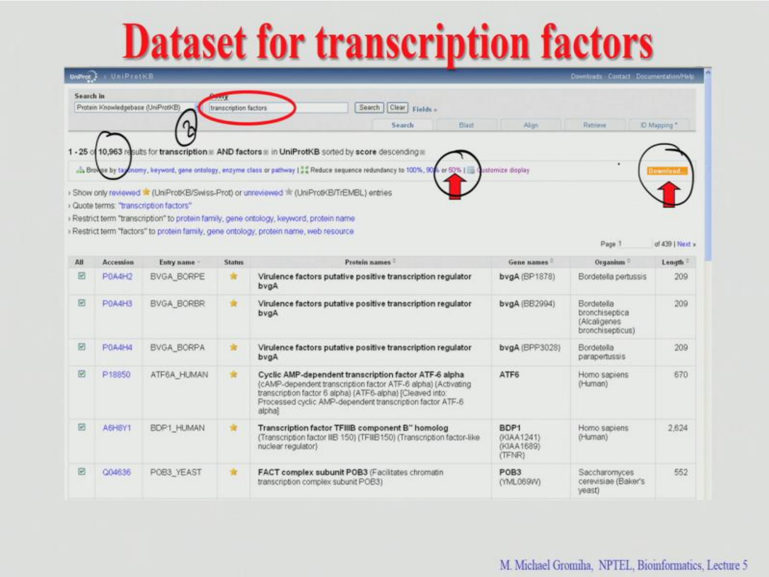
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Transcription factors are DNA binding proteins right they have specific functions. So, to understand the features of this transcription factors, sequences, I would like to collect the sequence from the transcription factors and reduce the data with 50 percent sequence identity. This is question number one.

Question number 2 I am interested to understand the function of a particular protein, for example, this human mitochondrial beta-barrel membrane protein VDAC, two questions. So, to understand the first question, how to obtain this information how to obtain the transcription factors with less than 50 percent sequence identity, first you have to go to

UniProt, right. The first step is to go to UniProt database and search for transcription factors.

(Refer Slide Time: 14:59)



Dataset for transcription factors

UniProt › UniProtKB
Downloads · Contact · Documentation/Help
Search in
Protein Knowledgebase (UniProtKB)
Query
transcription factors
Search
Clear
Fields »
Search
Blast
Align
Retrieve
ID Mapping *
2
Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Customize display
Download...
Page 1 of 439 | Next »

1 - 25 of 10,963 results for **transcription** AND **factors** in **UniProtKB** sorted by **score** descending

› Show only reviewed ★ (UniProtKB/Swiss-Prot) or unreviewed ★ (UniProtKB/TrEMBL) entries
› Quote terms: "transcription factors"
› Restrict term "transcription" to protein family, gene ontology, keyword, protein name
› Restrict term "factors" to protein family, gene ontology, protein name, web resource

All	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
☑	P0A4H2	BVGA_BORPE	★	**Virulence factors putative positive transcription regulator bvgA**	**bvgA** (BP1878)	Bordetella pertussis	209
☑	P0A4H3	BVGA_BORBR	★	**Virulence factors putative positive transcription regulator bvgA**	**bvgA** (BB2994)	Bordetella bronchiseptica (Alcaligenes bronchisepticus)	209
☑	P0A4H4	BVGA_BORPA	★	**Virulence factors putative positive transcription regulator bvgA**	**bvgA** (BPP3028)	Bordetella parapertussis	209
☑	P18850	ATF6A_HUMAN	★	**Cyclic AMP-dependent transcription factor ATF-6 alpha** (cAMP-dependent transcription factor ATF-6 alpha) (Activating transcription factor 6 alpha) (ATF6-alpha) [Cleaved into: Processed cyclic AMP-dependent transcription factor ATF-6 alpha]	**ATF6**	Homo sapiens (Human)	670
☑	A6H8Y1	BDP1_HUMAN	★	**Transcription factor TFIIIB component B" homolog** (Transcription factor IIB 150) (TFIIB150) (Transcription factor-like nuclear regulator)	**BDP1** (KIAA1241) (KIAA1689) (TFNR)	Homo sapiens (Human)	2,624
☑	Q04636	POB3_YEAST	★	**FACT complex subunit POB3** (Facilitates chromatin transcription complex subunit POB3)	**POB3** (YML069W)	Saccharomyces cerevisiae (Baker's yeast)	552

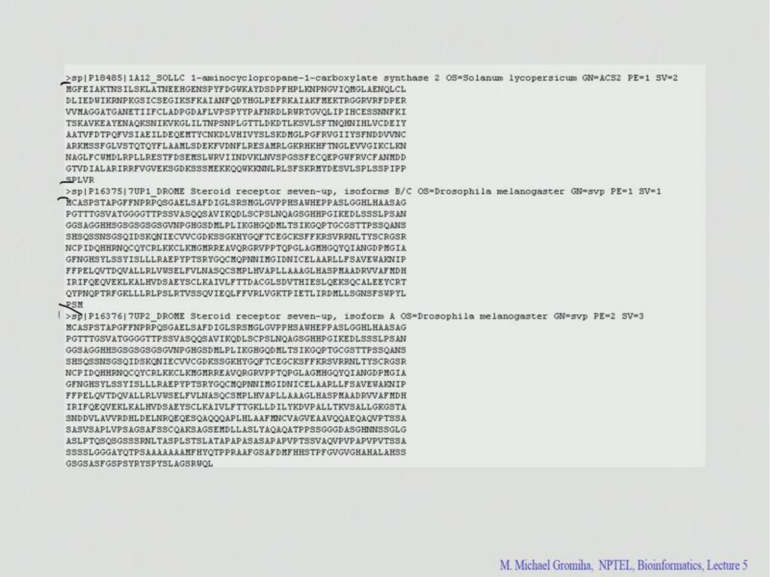
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

So, go to there and we can see the UniProt. So, we have to search with transcription factors. So, we search this is the second step.

So, when you search transcription factors what will happen? We will get a list of sequences which contain ‘transcription factors’ in any field because if you use the symbol if you symbol search right you can get the ‘transcription factors’, which are matched with anywhere or you can they search in this you can have the different options, if you want to get the only the title you can give title or any other MESH term (Refer Time: 15:30) you want to use it we can use any MESH(Refer Time: 15:32) terms.

So, in a transcription factors how many data we get? So, more than 10,000 data right now if you use now you will get more than that. So, now, what is the question? So, we need the transcription factors or with less than 50 percent sequence identity. So, what to do with this? You have to choose the 50 percent and if you click we will get the data. So, to download the data. So, there is an option called download here, click here download what will happen? We will ask for the option which format do you want, if you ask for the FASTA format, we will get the data in FASTA format right.

(Refer Slide Time: 16:00)

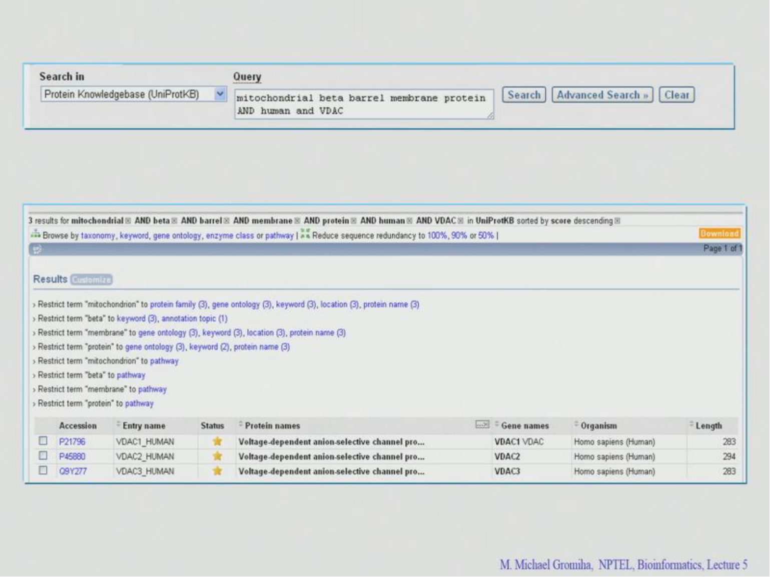


```
>sp|P18485|1A12_SOLLC 1-aminocyclopropane-1-carboxylate synthase 2 OS=Solanum lycopersicum GN=ACS2 PE=1 SV=2
MGFEIAKTNSILSKLATNEEHGENSPYFDGWKAYDSDPFHPLKNPNGVIQMGLAENQLCL
DLIEDWIKRNPKGSICSEGIKSFKAIANFQDYHGLPEFRKAIAKFMEKTRGGRVRFDPER
VVMAGGATGANETIIFCLADPGDAFLVPSPYYPAFNRDLRWRTGVQLIPIHCESSNNFKI
TSKAVKEAYENAQKSNIKVKGLILTNPSNPLGTTLDKDTLKSVLSFTNQHNIHLVCDEIY
AATVFDTPQFVSIAEILDEQEMTYCNKDLVHIVYSLSKDMGLPGFRVGIIYSFNDDVVNC
ARKMSSFGLVSTQTQYFLAAMLSDEKFVDNFLRESAMRLGKRHKHFTNGLEVVGIKCLKN
NAGLFCWMDLRPLLRESTFDSEMSLWRVIINDVKLNVSPGSSFECQEPGWFRVCFANMDD
GTVDIALARIRRFVGVEKSGDKSSSMEKKQQWKKNNLRLSFSKRMYDESVLSPLSSPIPP
SPLVR
>sp|P16375|7UP1_DROME Steroid receptor seven-up, isoforms B/C OS=Drosophila melanogaster GN=svp PE=1 SV=1
MCASPSTAPGFFNPRPQSGAELSAFDIGLSRSMGLGVPPHSAWHEPPASLGGHLHAASAG
PGTTTGSVATGGGGTTPSSVASQQSAVIKQDLSCPSLNQAGSGHHPGIKEDLSSSLPSAN
GGSAGGHHSGSGSGSGSGVNPGHGSDMLPLIKGHGQDMLTSIKGQPTGCGSTTPSSQANS
SHSQSSNSGSQIDSKQNIECVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNLTYSCRGSR
NCPIDQHHRNQCQYCRLKKCLKMGMRREAVQRGRVPPTQPGLAGMHGQYQIANGDPMGIA
GFNGHSYLSSYISLLLRAEPYPTSRYGQCMQPNNIMGIDNICELAARLLFSAVEWAKNIP
FFPELQVTDQVALLRLVWSELFVLNASQCSMPLHVAPLLAAAGLHASPMAADRVVAFMDH
IRIFQEQVEKLKALHVDSAEYSCLKAIVLFTTDACGLSDVTHIESLQEKSQCALEEYCRT
QYPNQPTRFGKLLLRLPSLRTVSSQVIEQLFFVRLVGKTPIETLIRDMLLSGNSFSWPYL
PSM
>sp|P16376|7UP2_DROME Steroid receptor seven-up, isoform A OS=Drosophila melanogaster GN=svp PE=2 SV=3
MCASPSTAPGFFNPRPQSGAELSAFDIGLSRSMGLGVPPHSAWHEPPASLGGHLHAASAG
PGTTTGSVATGGGGTTPSSVASQQSAVIKQDLSCPSLNQAGSGHHPGIKEDLSSSLPSAN
GGSAGGHHSGSGSGSGSGVNPGHGSDMLPLIKGHGQDMLTSIKGQPTGCGSTTPSSQANS
SHSQSSNSGSQIDSKQNIECVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNLTYSCRGSR
NCPIDQHHRNQCQYCRLKKCLKMGMRREAVQRGRVPPTQPGLAGMHGQYQIANGDPMGIA
GFNGHSYLSSYISLLLRAEPYPTSRYGQCMQPNNIMGIDNICELAARLLFSAVEWAKNIP
FFPELQVTDQVALLRLVWSELFVLNASQCSMPLHVAPLLAAAGLHASPMAADRVVAFMDH
IRIFQEQVEKLKALHVDSAEYSCLKAIVLFTTGKLLDILYKDVPALLTKVSALLGKGSTA
SNDDVLAVVRDHLDELNRQEQESQAQQQAPLHLAAFMNCVAGVEAAVQQAEQAQVPTSSA
SASVSAPLVPSAGSAFSSCQAKSAGSEMDLLASLYAQAQATPPSSGGGDASGHNNSSGLG
ASLPTQSQSGSSSRNLTASPLSTSLATAPAPASASAPAPVPTSSVAQVPVPAPVPVTSSA
SSSSLGGGAYQTPSAAAAAAAMFHYQTPPRAAFGSAFDMFHHSTPFGVGVGHAHALAHSS
GSGSASFGSPSYRYSPYSLAGSRWQL
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

This is the one and here this is the second one and the third one and so on. So, you get all the data in FASTA format right the first question now it is done, now we can use this for the analysis right. What is the second question? Yeah, I want to see amino acid sequence as well as some of the functions of this specific protein, because this protein is a recently published one. So, it has several functions right, eukaryotic protein, it is a first eukaryotic protein right in beta barrel membrane protein.

(Refer Slide Time: 16:32)



Search in: Protein Knowledgebase (UniProtKB)

Query: mitochondrial beta barrel membrane protein AND human and VDAC

3 results for **mitochondrial** ☒ **AND beta** ☒ **AND barrel** ☒ **AND membrane** ☒ **AND protein** ☒ **AND human** ☒ **AND VDAC** ☒ in **UniProtKB** sorted by **score** descending ☒

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% |

› Restrict term "mitochondrion" to protein family (3), gene ontology (3), keyword (3), location (3), protein name (3)
› Restrict term "beta" to keyword (3), annotation topic (1)
› Restrict term "membrane" to gene ontology (3), keyword (3), location (3), protein name (3)
› Restrict term "protein" to gene ontology (3), keyword (2), protein name (3)
› Restrict term "mitochondrion" to pathway
› Restrict term "beta" to pathway
› Restrict term "membrane" to pathway
› Restrict term "protein" to pathway

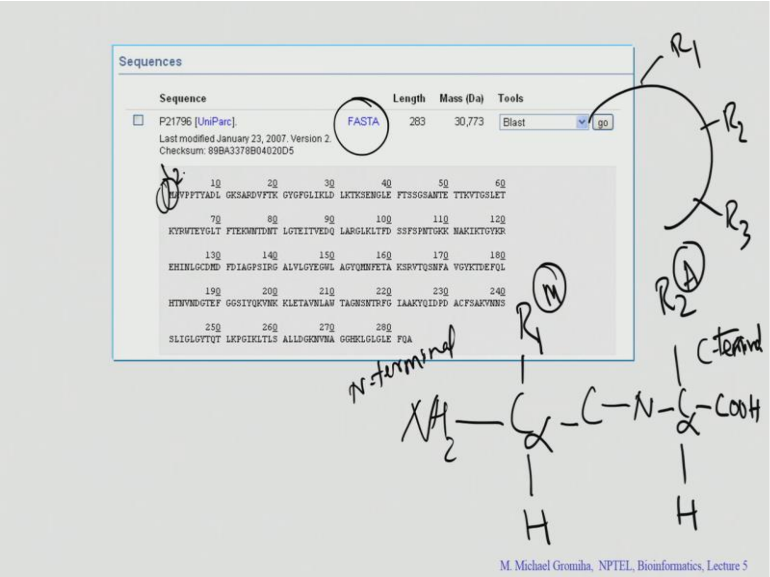
	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
□	P21796	VDAC1_HUMAN	★	**Voltage-dependent anion-selective channel pro...**	**VDAC1** VDAC	Homo sapiens (Human)	283
□	P45880	VDAC2_HUMAN	★	**Voltage-dependent anion-selective channel pro...**	**VDAC2**	Homo sapiens (Human)	294
□	Q9Y277	VDAC3_HUMAN	★	**Voltage-dependent anion-selective channel pro...**	**VDAC3**	Homo sapiens (Human)	283

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

So, if you do this right just for go to the UniProt and if you search with this keyword, mitochondrial beta barrel membrane protein and human and VDAC right then you will get these specific entries and if you click on any of these things right, you will get the data this is a sequence. So, these are your sequences if you want to get the FASTA format you can click into FASTA format and then you will get the sequence in FASTA format. Now again in the UniProt, you can obtain the information on different perspectives; for example, if you are interested on any of the particular protein, you want to do that. So, we can do it and get the information. If you want to get the information regarding post translational modification sites right if you give the codes, if you can get all the post translational modification sites or if you want to collect the sequences of a set of data.

For example DNA binding proteins or the RNA of binding proteins or any information right if you give the see the in the search, you search the correctly you get the data and it is also possible to get the data with any specific sequence redundancy. So, it is a unique resource, it contain lot of information regarding protein sequences right. So, this is the reason why several researchers they are using the UniProt database in their research. So, now, we will recollect again, what are the various aspects we discussed in the class today. Primary structures, what is a primary structure? Specific arrangement of amino acid residues in a protein right like it has a main chain and it has a site chain. So, main chain it is the same right what is the main chain?

(Refer Slide Time: 17:57)



Sequences

	Sequence		Length	Mass (Da)	Tools
□	P21796 [UniParc]. Last modified January 23, 2007. Version 2. Checksum: 89BA3378B04020D5	FASTA	283	30,773	Blast go

```
        10         20         30         40         50         60
MAVPPTYADL GKSARDVFTK GYGFGLIKLD LKTKSENGLE FTSSGSANTE TTKVTGSLET

        70         80         90        100        110        120
KYRWTEYGLT FTEKWNTDNT LGTEITVEDQ LARGLKLTFD SSFSPNTGKK NAKIKTGYKR

       130        140        150        160        170        180
EHINLGCDMD FDIAGPSIRG ALVLGYEGWL AGYQMNFETA KSRVTQSNFA VGYKTDEFQL

       190        200        210        220        230        240
HTNVNDGTEF GGSIYQKVNK KLETAVNLAW TAGNSNTRFG IAAKYQIDPD ACFSAKVNNS

       250        260        270        280
SLIGLGYTQT LKPGIKLTLS ALLDGKNVNA GGHKLGLGLE FQA
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 5

Student: (Refer Time: 17:59).

The amino acid side there is a NH2 right.

Student: C alpha.

C alpha.

Student: C

C N C alpha C it goes and right then you can put the COOH, this is the amino terminal N terminal this called the N terminal or amino terminal. This is C terminal. So, we can see this is the information we will get, but C alpha is connected one side with the.

Student: R group.

R group. Another with the?

Student: H.

H right. So, this is H here this is R1 this is R2. So, now, this is the main chain this is chain is the same only the side chains are different. So, you get R 1, we will get R 2, R 3 and so on this can be same amino acid or the different amino acid because this determine the sequence. So, here if you see this one what is R1 here in this sequence this is M R 2.

Student: Alanine.

Alanine right because M is here right and then this is 1 2. So, this how we get the sequence there is a primary sequence we know the link, we know the sequence arrangement, but we do not need anything else if we do not know the locations right.

So, now what is the primary resource for the protein sequences?

Student: PIR.

PIR and the (Refer Time: 19:20) SWISS-PROT right this is the earlier developed databases right then they merge together to form.

Student: UniProt.

UniProt database UniProt is the universal protein database right now it is widely used in literature. Now what are the different contents of UniProt sequences? Functions and the structural information and the various interactions and the pathways and the assemblies right and they supplemented with the original literature, then what are the major applications of the UniProt? Yeah different functions and the different sequences different distribution of amino acids say different functions right we can use for higher order specific sequence or any correctly sequences right and you can use it for the further applications right.

This UniProt is the unique resource for protein sequence databases. Next class I will discuss about what are the various aspects you can derive from this amino acid sequences that we will discuss in the later classes.

Thanks for the kind attention.