

آزمایشگاه آموزشی
بیست و دومین المپیاد
زیست‌شناسی ایران

بیوانفورماتیک

دیتابیس‌ها.

روز چهارم
۹۸/۵/۳

اهداف آزمایش:

۱. آشنایی با مقدمات بیوانفورماتیک

زمان آزمایش: ۹۰ دقیقه



این فایل به منظور آموزش عملی دانش‌پژوهان المپیاد زیست‌شناسی ایران گردآوری شده است.



معرفی ابزارها و دیتابیس‌های بیوانفورماتیک

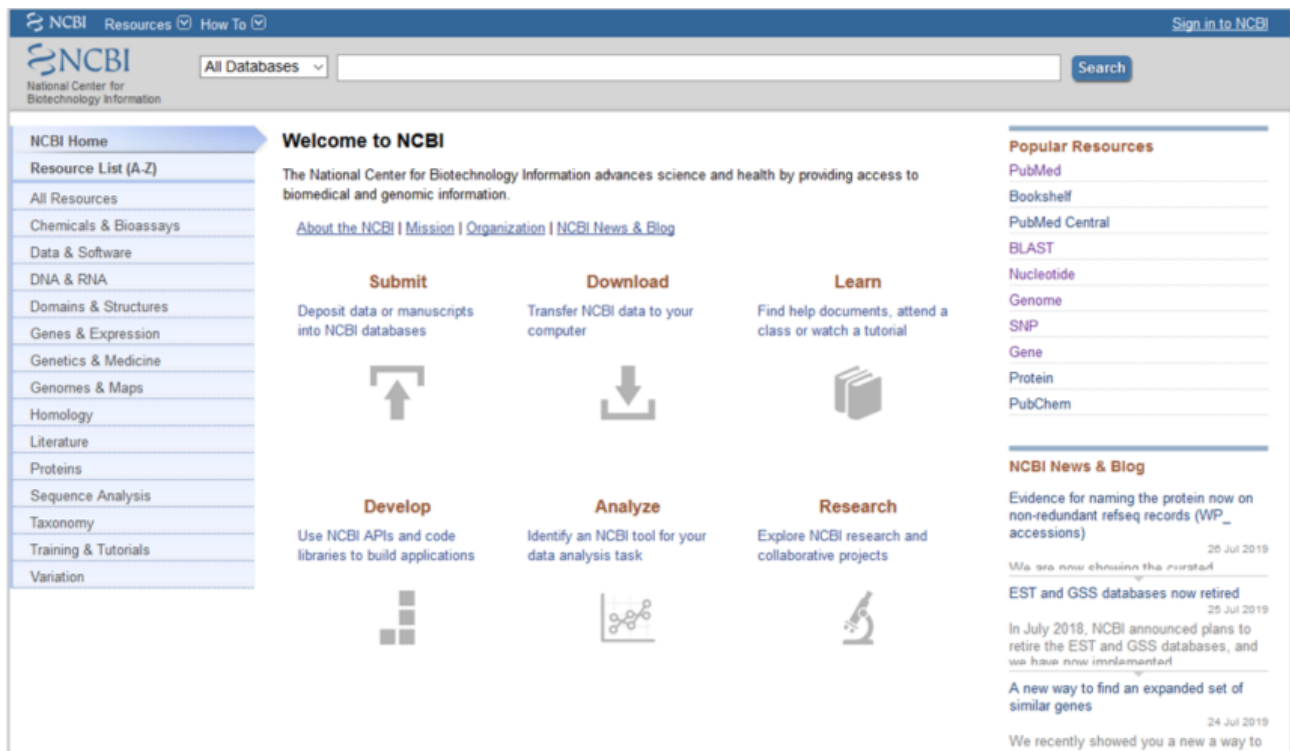
NCBI

NCBI (National Center for Biotechnology Information) حاوی پایگاه‌های داده عمومی می‌باشد، تحقیقات در زمینه زیست‌شناسی محاسباتی را هدایت می‌کند، نرم‌افزارهایی برای آنالیز اطلاعات توسعه می‌دهد و اطلاعات زیست‌پزشکی را منتشر می‌کند. در زیر مهم‌ترین منابع (Resources) ارائه شده توسط NCBI آمده است:

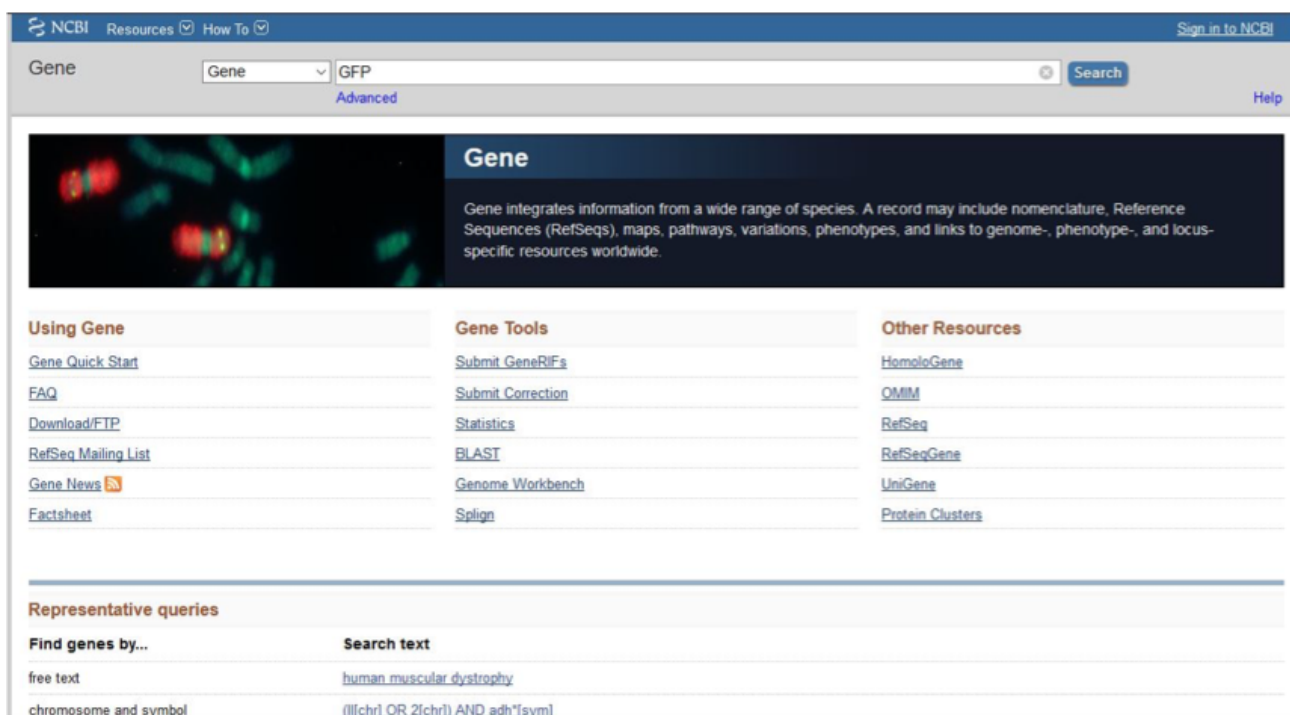
- PubMed: موتور جستجوی رایگان برای دسترسی به مقالات زیست‌پزشکی می‌باشد.
- BLAST: ابزار جستجوگر بر پایه شباهت توالی‌های نوکلئوتیدی یا پروتئینی که پس از جستجو در تمام پایگاه داده خود تعدادی از شبیه‌ترین توالی‌ها به توالی مدنظر شما را نمایش می‌دهد.
- BLAST-Primer: این ابزار برای طراحی پرایمر بصورت اتوماتیک و البته با تنظیماتی که شما انجام می‌دهید استفاده می‌شود و توالی جفت پرایمرهای FORWARD و REVERSE که مناسب‌تر از بقیه هستند را در اختیار شما قرار می‌دهد.

اکنون به جستجو در پایگاه داده NCBI پرداخته و به دنبال ژن خاصی می‌گردیم.

پس از وارد شدن به سایت NCBI برای دسترسی به تمام منابع سایت می‌توانید از زبانه Resources در بالای صفحه استفاده کنید اما امکانات محبوب سایت را می‌توانید در قسمت Popular Resources در سمت راست صفحه مشاهده کنید.



روی Gene کلیک کنید تا صفحه جدیدی باز شود.



نام ژن یا پروتئین مورد نظر را وارد کنید، می توانید برای اختصاصی تر شدن جستجو نام علمی گونه را هم در ادامه وارد کنید و روی دکمه search کلیک کنید. بدلیل انتخاب گزینه Gene در هر صورت توالی نوکلئوتیدی برایتان به نمایش در می آید.

The screenshot shows the NCBI Gene database search results for the query 'GFP'. The search results are displayed in a table with columns: Name/Gene ID, Description, Location, and Aliases. The results are sorted by Relevance and show 1 to 20 of 1551 items. The first few results are:

Name/Gene ID	Description	Location	Aliases
gfp ID: 7011691	green fluorescent protein [<i>Neisseria gonorrhoeae</i>]	NC_011521.1 (4419..5135)	pCmGFP_001
NAL1 ID: 4336986	Protein NARROW LEAF 1 [<i>Oryza sativa Japonica Group</i> (Japanese rice)]	Chromosome 4, NC_029259.1 (31203502..31214734)	OSNPB_040615000, GFP, LSCHL4, OsJ_16147, SPIKE, qFLW4
1700016L21Rik ID: 72208	RIKEN cDNA 1700016L21 gene [<i>Mus musculus</i> (house mouse)]	Chromosome 1, NC_000067.6 (80445932..80475660)	CD11c-DTR, DCKO, DTR-GFP, DTR- GRP, Jr3, Tg(tgax-DTR/EGFP)57Lan
LOC107331116 ID: 107331116	GFP-like fluorescent chromoprotein amFP486 [<i>Acropora digitifera</i>]		
LOC107331118 ID: 107331118	GFP-like fluorescent chromoprotein amFP486 [<i>Acropora digitifera</i>]		

The right sidebar shows filters for 'Results by taxon' and 'Find related data'. The 'Results by taxon' section lists top organisms: *Saccharomyces cerevisiae* (416), *Lachanea lanzarotensis* (270), *Homo sapiens* (144), *Mus musculus* (114), *Arabidopsis thaliana* (89), and All other taxa (516). The 'Find related data' section shows a dropdown menu for 'Database' and a 'Find items' button. The 'Search details' section shows the search criteria: 'GFP[All Fields] AND alive[prop]'.

از بین ژن های نشان داده شده روی نام (Gene ID/Name) ژن مورد نظر خود کلیک کنید.

The screenshot shows the NCBI Genomic regions, transcripts, and products page for the sequence NC_011521.1. The page displays a genomic map with various features, including genes, STS markers, and misc features. The 'Genomic regions, transcripts, and products' section shows the sequence NC_011521.1 and its associated features. The 'Bibliography' section lists related articles in PubMed, including 'Phasevarions mediate random switching of gene expression in pathogenic *Neisseria*' by Srikantha YN, et al. PLoS Pathog, 2009 Apr. PMID 19390608, Free PMC Article. The 'General protein information' section shows the preferred names for the green fluorescent protein. The right sidebar shows a list of related sites, including BLAST, Genome, BioProject, Genome Data Viewer, and GEO.

در این صفحه اطلاعات مربوط به ژن از قبیل توالی آن به همراه annotation ها و همچنین در قسمت RefSeq لینک مربوط به پروتئین مربوطه در UniProt آمده است که با استفاده از آن می توانید اطلاعات بیشتری درباره خود پروتئین کسب کنید.

برای دانلود توالی ژن در ابتدا در بخش Genomic Region روی FASTA کلیک کنید.

The screenshot shows the NCBI GenBank interface. At the top, there's a search bar and a 'Send to' dropdown menu. Below the search bar, the 'Change region shown' section is visible, with 'Selected region' selected. The 'Update View' button is present. The main content area displays the FASTA sequence for 'Neisseria gonorrhoeae plasmid pCmGFP, complete sequence'. The sequence is shown in a monospaced font, with line numbers on the left. The sequence starts with >NC_011521.1:4419-5135 and ends with ATGACTATACAAATAA.

در این قسمت می توانید توالی مربوطه را با فرمت FASTA ببینید اما برای دانلود فایل متنی آن روی Send to کلیک کرده و در قسمت choose Destination گزینه File را انتخاب کرده و create file کنید.

در این قسمت در ضمن می توانید coding sequence را انتخاب کرده و با استفاده از FASTA Protein توالی آمینواسیدی را دانلود کنید.

Pairwise alignment: برای مقایسه‌ی دوتایی توالی‌ها استفاده میشود.

برای استفاده از این ابزار مرورگر خود را بر روی آدرس ebi.ac.uk/tools/msa قرار دهید. در این صفحه الگوریتم‌های مختلف برای همراهی توالی‌ها وجود دارد که هر کدام را میتوان بر روی توالی‌های نوکلئوتیدی یا آمینواسیدی اجرا کرد.

Tools > Pairwise Sequence Alignment

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

By contrast, **Multiple Sequence Alignment (MSA)** is the alignment of three or more biological sequences of similar length. From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.

Global Alignment

Global alignment tools create an end-to-end alignment of the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

Needle (EMBOSS)
EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.
[Protein](#) [Nucleotide](#)

Stretcher (EMBOSS)
EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.
[Protein](#) [Nucleotide](#)

Local Alignment

Local alignment tools find one, or more, alignments describing the most similar region(s) within the sequences to be aligned. There are separate forms for protein or nucleotide sequences.

Water (EMBOSS)
EMBOSS Water uses the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.
[Protein](#) [Nucleotide](#)

Matcher (EMBOSS)
EMBOSS Matcher identifies local similarities between two sequences using a rigorous algorithm based on the LALIGN application.
[Protein](#) [Nucleotide](#)

LALIGN
LALIGN finds internal duplications by calculating non-intersecting local alignments of protein or DNA sequences.
[Protein](#) [Nucleotide](#)

Genomic Alignment

Genomic alignment tools concentrate on DNA (or to DNA) alignments while accounting for characteristics present in genomic data.

با توجه به پروژه‌ی خود الگوریتم مربوطه را انتخاب کنید. در این مثال Global Alignment برای دو توالی پروتئینی نشان داده شده است.

توالی‌های FASTA خود را در پنجره های مربوطه کپی کنید. با کلیک کردن روی دکمه‌ی more options می‌توانید

The screenshot shows the EMBOSS Stretcher web interface. At the top, there's a navigation bar with links to EMBL-EBI, Services, Research, Training, Industry, About us, and a search icon. Below this, the page title is "EMBOSS Stretcher". A breadcrumb trail shows "Tools > Pairwise Sequence Alignment > EMBOSS Stretcher". The main heading is "Pairwise Sequence Alignment (PROTEIN)". A brief description states: "EMBOSS Stretcher calculates an optimal global alignment of two sequences using a modification of the classic dynamic programming algorithm which uses linear space. This is the form for protein sequences. Please go to the [nucleotide](#) form if you wish to align DNA or RNA sequences."

STEP 1 - Enter your protein sequences

Enter or paste your first **protein** sequence in any supported format:

Or, [upload a file:](#) Choose file no file selected [Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

AND

Enter or paste your second **protein** sequence in any supported format:

Or, [upload a file:](#) Choose file no file selected

STEP 2 - Set your pairwise alignment options

MATRIX	GAP OPEN	GAP EXTEND	OUTPUT FORMAT
BLOSUM62	12	2	pair

STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

[Submit](#)

If you plan to use these services during a course please [contact us](#).

تنظیمات همراستایی را از حالت پیش فرض تغییر دهید. ماتریس جهش مورد استفاده و گپ پناستی از جمله این موارد هستند.

در آخر دکمه‌ی submit را زده و منتظر شوید تا همراهی انجام شود. در پایان با چنین صفحه‌ای مواجه میشوید که خلاصه تنظیمات و نتایج همراهی را در بر دارد.

The screenshot shows the EMBOSS Stretcher web interface. The top navigation bar includes links for EMBL-EBI, Services, Research, Training, Industry, About us, and a search icon. The main header is "EMBOSS Stretcher". Below it, there are tabs for "Protein alignment", "Nucleotide alignment", "Web services", "Help & Documentation", and "Also in this section". The "Protein alignment" tab is selected. The page displays the results for a job named "emboss_stretcher-120180824-162531-0081-81745245-p2m". The results are shown in a text format, including the command line used, the sequences being aligned, and the alignment itself. The alignment shows a high degree of similarity between the two sequences, with a score of 648.

Multiple alignment: با استفاده از این ابزار میتوانید چندین توالی را با هم مقایسه کنید. امکان استفاده از خروجی این نوع همراهی برای کشیدن درخت فیلوژنتیک نیز وجود دارد. برای شروع مرورگر خود را بر روی آدرس ebi.ac.uk/tools/msa قرار دهید. در این صفحه الگوریتم‌های مختلف برای همراهی چندگانه را مشاهده میکنید. در این مثال از الگوریتم Clustal Omega برای مقایسه‌ی چند توالی پروتئین استفاده می‌کنیم.

The screenshot shows the "Multiple Sequence Alignment" page on the EMBL-EBI website. The page title is "Multiple Sequence Alignment". It provides an overview of MSA and lists several tools available for performing MSA. The tools listed are Clustal Omega, Kalign, MAFFT, MUSCLE, MView, T-Coffee, and WebPRANK. Each tool is accompanied by a brief description of its capabilities and a link to launch the tool. The page also includes a footer with information about the tools and a link to the EMBL-EBI support page.

در اولین پنجره Enter or paste a set of FASTA مورد نظر خود را در پنجره مربوطه کپی کنید. دقت داشته باشید که بین پایان یک توالی و آغاز توالی بعد باید یک خط فاصله باشد.

The screenshot shows the Clustal Omega web interface. The header includes navigation links: EMBL-EBI, Services, Research, Training, Industry, About us, and a search bar. The main title is 'Clustal Omega'. Below the title, there are links for 'Input form', 'Web services', 'Help & Documentation', 'Bioinformatics Tools FAQ', 'Feedback', and 'Share'. The main content area is titled 'Multiple Sequence Alignment' and includes a brief description of the tool. It also features a 'STEP 1 - Enter your input sequences' section with a text area for pasting sequences and a file upload option. Below this is a 'STEP 2 - Set your parameters' section with various dropdown menus for output format, alignment input, and clustering options. The final section is 'STEP 3 - Submit your job' with a 'Submit' button and a checkbox for email notifications.

EMBL-EBI Services Research Training Industry About us Q EMBL-EBI Hinxton

Clustal Omega

[Input form](#) [Web services](#) [Help & Documentation](#) [Bioinformatics Tools FAQ](#) [Feedback](#) [Share](#)

[Tools](#) > [Multiple Sequence Alignment](#) > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our [pairwise sequence alignment tools](#).

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

Or, upload a file: no file selected [Use a example sequence](#) [Clear sequence](#) [See more example inputs](#)

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

[ALIGN INPUT SEQUENCES](#) [MSB-LIKE CLUSTERING GUIDE](#) [MSB-LIKE CLUSTERING ITERATION](#) [NUMBER of COMBINED ITERATIONS](#)

no	yes	yes	default(9)
MAX GUIDE TREE ITERATIONS	MAX HMM ITERATIONS	ORDER	
default	default	aligned	

STEP 3 - Submit your job

☐ Be notified by email (Tick this box if you want to be notified by email when the results are available)

[Submit](#)

If you plan to use these services during a course please [contact us](#).

Please read the [FAQ](#) before seeking help from our support staff.

1.

EMBL-EBI

[Services](#)
[Research](#)
[Training](#)
[Industry](#)
[About us](#)

EMBL-EBI Hinxton

Clustal Omega

[Input form](#)
[Web services](#)
[Help & Documentation](#)
[Bioinformatics Tools FAQ](#)

[Feedback](#)
[Share](#)

Tools > Multiple Sequence Alignment > Clustal Omega

Results for job [clustalo-I20180824-163858-0479-29210239-p2m](#)

[Alignments](#)
[Result Summary](#)
[Phylogenetic Tree](#)
[Submission Details](#)

[Download Alignment File](#)
[Show Colors](#)
[View result with Jalview](#)
[Send to Simple Phylogeny](#)
[Send to MView](#)

CLUSTAL O(1.2.4) multiple sequence alignment

```

tr [013169] [013169] CYP3A
sp [P69905] HBA_HUMAN
sp [P01942] HBA_MOUSE
sp [P13786] HBAE_CAPI1
      NLSGSDRAAVYLAHAKISFPAADDIGASALGMLVYPTGYTFYAINSGLSGSGPYKRR      60
      NYLSFADKTVYKANGKYGVAAGEYGAELERHLSFPTTKTTFYFD-LSHGSAQVKGH      59
      NYLSGEDRINKIKAHNGK1GGCAETGAEALERHLSFPTTKTTFYFD-YSHGSAQVKGH      59
      NLSLTWERTTILSLNKK1STGADYIGTTELEKPLFCYQDAKTYTFYFD-LSHGSAQVAH      59
      * * 111 1 1 *..1.. 1.. *1*1* *11 1* 1**** *1* 1 **..11 *

tr [013169] [013169] CYP3A
sp [P69905] HBA_HUMAN
sp [P01942] HBA_MOUSE
sp [P13786] HBAE_CAPI1
      GKVMGAVADAVEK1DGLVGLASLSHASKLVDPANFKILAHNVIVVIGLPGDFG      120
      GKRYADALINAVAYDMPHLSALSOLJAHKLVDPNFKLSHCLVTLAAHLPAAFT      119
      GKRYADALASAGHSDGLSALSOLJAHKLVDPNFKLSHCLVTLASRHHPAFT      119
      GKRYAVAGGAVYS1DPTLSALSGEJAVYLVDPNFKLSHCLVTLASRHHPAFT      119
      *.. 1 *1.. 1*11 1..*1 **** *.....*1*1*11*1.. 1..1*

tr [013169] [013169] CYP3A
sp [P69905] HBA_HUMAN
sp [P01942] HBA_MOUSE
sp [P13786] HBAE_CAPI1
      PEVYNSVQKFFQSLALALEKTYR      143
      PAVHSLQKFLASVITVLTSEKTYR      142
      PAVHSLQKFLASVITVLTSEKTYR      142
      ADARAAMKFLSIVSQVLTSEKTYR      142
      .. 1 ***: 11 ..*1,***
  
```

PLEASE NOTE: Showing colors on large alignments is slow.

EMBL-EBI

Services

[By topic](#)
[By name \(A-Z\)](#)
[Help & Support](#)

Research

[Publications](#)
[Research groups](#)
[Postdocs & PhDs](#)

Training

[Train at EBI](#)
[Train outside EBI](#)
[Train online](#)
[Contact organisers](#)

Industry

[Members Area](#)
[Workshops](#)
[SME Forum](#)
[Contact Industry programme](#)

About EMBL-EBI

[Contact us](#)
[Events](#)
[Jobs](#)
[News](#)
[People & groups](#)

EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. +44 (0)1223 49 44 44

Covovicht © EMBL-EBI 2018 | EMBL-EBI is part of the European Molecular Biology Laboratory | Terms of use

Intranet

با کلیک بر روی تب Result summary با صفحه‌ی زیر مواجه خواهید شد. هر کدام از لینک های موجود در صفحه اطلاعاتی در مورد همراستایی انجام شده به شما میدهد.

The screenshot shows the Clustal Omega web interface. At the top, there's a navigation bar with links like 'EMBL-EBI', 'Services', 'Research', 'Training', 'Industry', and 'About us'. Below this, the 'Clustal Omega' header is visible. The main content area shows the 'Result Summary' tab for a job with ID 'clustalo-120180824-163858-0479-29210239-p2m'. The summary includes links for 'Input Sequences', 'Tool Output', 'Alignment in CLUSTAL format with base/residue numbering', 'Phylogenetic Tree', and 'Percent Identity Matrix'. On the right, there are options to 'View result with Jalview', 'Simple Phylogeny', 'MView', and 'Send to MView'. The footer contains EMBL-EBI contact information and a copyright notice for 2018.

Uniprot

Uniprot یکی از مهم ترین Data Base های بیوانفورماتیکی به ویژه در زمینه پروتئین هاست. از این سایت می‌توانید برای اهداف زیر استفاده کنید :

- (1) پیدا کردن توالی پروتئین با داشتن اسم آن
- (2) پیدا کردن توالی پروتئین با داشتن قطعه ای از توالی آن
- (3) به دست آوردن اطلاعات پیشرفته در رابطه با پروتئین ها (برهمکنش ها ، بیماری ها و ...)

در ادامه به توضیح بخش های مختلف این سایت می پردازیم . عکس های مربوط به این بخش در قالب ضمیمه به شما داده شده است .

بخش اول: صفحه اصلی (عکس شماره 1)

بخش اصلی این سایت به نام UniProtKB (knowledgebase) متشکل از داده های پروتئین ها به صورت مجزاست (هر پروتئین کد و صفحه مخصوص به خود را دارد) .

فلش ۱: از این قسمت با جستجو نام پروتئین یا ژن کد کننده آن می توانید به اطلاعات آن پروتئین دسترسی پیدا کنید.

فلش ۲: UniProtKB از دو دسته داده پروتئین تشکیل شده است :

- دسته اول یا Prot-Swiss شامل پروتئین هایی است که علاوه بر پیش بینی وجود و توالی شان از روی داده های DNA ، به صورت مستقیم (خود پروتئین) نیز بررسی و تایید شده است .
- دسته دوم یا TrEMBL همان طور که از اسم اش نیز مشخص می شود (Translate of EMBL) تنها از پیش بینی وجود و توالی پروتئین از روی توالی های DNA موجود در DataBase سایت EMBL بوجود آمده است و از اعتبار کمتری بر خوردار می باشد

فلش ۳: UniRef دربردارنده خانواده های پروتئینی است (آستانه ای که دو توالی را در یک خانواده در نظر بگیرد در سه سطح 50 و 90 و 100 درصد قابل تنظیم است) .

UniRef - Cluster: MoeD5 (90%)

Unreviewed members
Published: June 5, 2019
Cluster ID: UniRef90_A0A001 (built on seed sequence¹ A0A001)

Expand cluster to 30% identity.
List component clusters with 100% identity.
1 to 9 of 9 Show (25)

Cluster members	Entry name	Protein name	Organism	Related clusters	Length	Role
A0A001	A0A001_STRVD	MoeD5	Streptomyces viridosporus	UniRef100_A0A001	591	Representative & Seed
A0A2U3HSU7	A0A2U3HSU7_SACTN	ABC transporter ATP-binding protein	Streptomyces sp. NWU49	UniRef100_A0A2U3HSU7	591	
D6A7F5	D6A7F5_STRV1	Uncharacterized protein	Streptomyces viridosporus (strain ATCC 14672 / DSM 40746 / JCM 4963 / KCTC 9882 / NRRL B-12104 / FH 1290) (Streptomyces ghanaensis)	UniRef100_D6A7F5	591	
A0A1ASPG65	A0A1ASPG65_SACTN	ABC transporter	Streptomyces sp. H-KF8	UniRef100_A0A1ASPG65	590	
A0A0C4Y377	A0A0C4Y377_SACTN	Putative ABC transporter (Fragment)	Streptomyces prasinus	UniRef100_A0A0C4Y377	296	
UP100098FFE6C		ABC transporter ATP-binding protein	Streptomyces viridosporus	UniRef100_UP100098FFE6C	591	
UP100000AA7E5		ABC transporter ATP-binding protein	Streptomyces viridosporus	UniRef100_UP100000AA7E5	591	
UP100068B640E		ABC transporter ATP-binding protein	Streptomyces prasinus	UniRef100_UP100068B640E	590	
UP100068240A1		ABC transporter ATP-binding protein	Streptomyces prasinus	UniRef100_A0A0C4Y377	590	

Sequence

Representative sequence ¹	Length	Mass (Da)	Tools
A0A001 Checksum: 4F6121D422B63694	591	61,726	BLAST

فلش ۴: UniParc در واقع DataBase ای است که از یکی کردن توالی های که در UniProtKB بیش از یک بار با نام های متفاوت تکرار شده بودند به وجود آمده است .

UniParc results

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. Proteins may exist in different source databases and in multiple copies in the same database. UniParc avoids such redundancy by storing each unique sequence only once and giving it a stable and unique identifier (UPI).

A UPI is never removed, changed or reassigned. UniParc contains only protein sequences. All other information about the protein must be retrieved from the source databases using the database cross-references.

Entry	Organisms	UniProtKB	First seen	Last seen	Length
UP0000000001	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	P07612	1988-11-01	2019-07-03	250
	Marburg virus (MARV)	P07612.1 (obsolete)			
	Rabbitpox virus	ADA212H01			
	Vaccinia virus	Q00N26			
	Cowpox virus (CPV)	Q482L4			
	synthetic construct	Q711T2			
		Q76QK2			
		Q76ZT7.1 (obsolete)			

فلش ۵: Proteomes دربردارنده تمام توالی های پروتئینی مربوط به یک ارگانیسم می باشد . اکثر این توالی ها از پیش بینی وجود و توالی پروتئین از روی ژنوم این ارگانیسم ها به دست آمده است .

Proteomes results

What are proteomes?
A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

What are reference proteomes?
Some proteomes have been (manually and algorithmically) selected as **reference proteomes**. They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny.

Filter by:

- 16,274 Reference proteomes
- 200,586 Other proteomes
- Superkingdom
- 117,763 Bacteria
- 96,280 Viruses
- 1,185 Archaea
- 1,732 Eukaryota
- Map to
- UniProtKB
- UniParc (for redundant proteomes)
- Demo

Proteome ID	Organism	Organism ID	proteomecount
UP000000863	Emiliana huxleyi virus 86 (isolate United Kingdom/English Channel/1999) (EHV-86)	654925	472
UP000000872	Amesoba moorei entomopoxvirus (AmEPV)	28321	278
UP000006714	Penicillium chrysogenum virus (isolate Caston/2003) (PCV)	654932	4
UP000001099	Lettuce infectious yellows virus (isolate United States/92) (LIYV)	651355	9
UP000000680	White bream virus (isolate Blicca bjoerkna L./Germany/DF24/00) (WBV)	766180	5
UP000002418	Bdellovibrio phage phiRH2K (Bacteriophage phiRH2K)	145579	11
UP000001670	Poinsettia latent virus (isolate Euphorbia pulcherrima/Germany/Siepen/2005) (PLV) (Poinsettia cryptic virus)	686943	4
UP000007021	Ostreid herpesvirus 1 (isolate France) (OshV-1) (Pacific oyster herpesvirus)	654903	116
UP000006715	Potato mop-top virus (isolate Potato/Sweden/Sw) (PMTV)	652839	7

فلش 6: برای جست و جوی بیماری ها و از آن طریق فهمیدن پروتئین های دخیل در آن بیماری ها به کار می رود .

فلش 7: برای جست و جوی ساختار های سلولی و پروتئین های موجود در آن ساختار ها به کار می رود .

بررسی یک پروتئین

برای این کار ابتدا نام پروتئین مد نظر (در این مورد prion) را در نوار جست و جو UniProtKB وارد می کنیم .
صفحه زیر ظاهر می شود:

The screenshot shows the UniProtKB search results for the query 'prion'. The page is titled 'UniProtKB results' and includes a search bar at the top. Below the search bar, there is a summary of the results, indicating that the UniProt Knowledgebase (UniProtKB) is the central hub for protein information. The results are categorized into 'Reviewed (Swiss-Prot) - Manually annotated' and 'Unreviewed (TrEMBL) - Computationally analyzed'. A table of results is displayed, showing columns for Entry, Entry name, Protein names, Gene names, Organism, and Length. The table lists several prion proteins from various organisms, including Human, Mouse, Bovine, and Sheep.

Entry	Entry name	Protein names	Gene names	Organism	Length
Q9UKY0	PRND_HUMAN	Prion-like protein doppel	PRND DPL, UNQ1830/PRO3443	Homo sapiens (Human)	176
P04156	PRIO_HUMAN	Major prion protein	PRNP ALTRP, PRIP, PRP	Homo sapiens (Human)	253
P04925	PRIO_MOUSE	Major prion protein	Prnp Prn-p, Prp	Mus musculus (Mouse)	254
P10279	PRIO_BOVIN	Major prion protein	PRNP PRP	Bos taurus (Bovine)	264
P04273	PRIO_MESAU	Major prion protein	PRNP PRP	Mesocricetus auratus (Golden hamster)	254
P23907	PRIO_SHEEP	Major prion protein	PRNP PRP, SIP	Ovis aries (Sheep)	256
P13852	PRIO_RAT	Major prion protein	Prnp Prn, Prp	Rattus norvegicus (Rat)	254
Q95211	PRIO_RABIT	Major prion protein	PRNP PRP	Oryctolagus cuniculus (Rabbit)	252
P27177	PRIO_CHICK	Major prion protein homolog	PRNP PRN-P	Gallus gallus (Chicken)	273
O18754	PRIO_FELCA	Major prion protein	PRNP PRP	Felis catus (Cat)	256

در این صفحه لیستی از پروتئین ها با نام های مشابه یا یکسان با نام وارد شده مشاهده می کنید با انتخاب یکی از پروتئین ها می توانید وارد صفحه جزئیات آن شوید .
*از ستون سمت چپ می توان برای انتخاب ارگاناسمی که پروتئین متعلق به آن است استفاده کرد .
*با انتخاب چند پروتئین می توان Alignment انجام داد .

اطلاعات پروتئین (تصویر شماره ۲):

این صفحه شامل اطلاعات متنوعی در مورد پروتئین مد نظر ما است که در چند دسته بندی قرار گرفته اند .
برای مشاهده هر کدام از این قسمت ها باید تیک کنار آن قسمت در جدول سمت چپ صفحه زده شده باشد.
* پیشنهاد می شود که ابتدا همه تیک ها را برداشته (بوسیله دکمه None) و سپس تک تک قسمت ها را بررسی کنید .

None

<input checked="" type="checkbox"/>	Function
<input checked="" type="checkbox"/>	Names & Taxonomy
<input checked="" type="checkbox"/>	Subcellular location
<input checked="" type="checkbox"/>	Pathology & Biotech
<input checked="" type="checkbox"/>	PTM / Processing
<input checked="" type="checkbox"/>	Expression
<input checked="" type="checkbox"/>	Interaction
<input checked="" type="checkbox"/>	Structure
<input checked="" type="checkbox"/>	Family & Domains
<input checked="" type="checkbox"/>	Sequences (2+)
<input checked="" type="checkbox"/>	Similar proteins
<input checked="" type="checkbox"/>	Cross-references
<input checked="" type="checkbox"/>	Entry information
<input checked="" type="checkbox"/>	Miscellaneous

▲ Top

1. Function : این قسمت نمایی کلی از ساختار و عملکرد پروتئین به ما می‌دهد .
2. Taxonomy & Name : نام رسمی و کد مخصوص این پروتئین را در این سایت و پایگاه داده های دیگر می‌دهد .
3. Subcellular location : جایگاه پروتئین مد نظر را در سلول نشان می‌دهد
4. Biotech & Pathology : نشان دهنده بیماری های مرتبط با این پروتئین و جهش هایی در پروتئین است که باعث آن بیماری ها میشوند .
5. Processing/PTM : نشان دهنده تغییرات بعد از ترجمه پروتئین (سیگنال پپتید ، گلیکوزیلاسیون و ...) است .
6. Expression : حاوی لینک پایگاه داده های بیان این پروتئین در بافت های مختلف است .
7. Interaction : نشان دهنده برهمکنش دو به دو پروتئین مد نظر ما با دیگر پروتئین ها است .
8. Structure : نشان دهنده ساختار سه بعدی پروتئین می باشد .
9. Domains & Family : نشان دهنده دمین های ساختاری یا عملکردی موجود در پروتئین است .
10. Sequences : این قسمت شامل توالی پروتئین است که میتوان آن را به صورت فایل FASTA ذخیره کرد و همچنین جهش ها و پلی مورفیسم موجود برای این پروتئین است .
11. Similar proteins : این قسمت پروتئین های شبیه به این پروتئین از نظر ساختار اول را نشان می‌دهد .
12. Entry information و references-Cross : حاوی اطلاعاتی مربوط به کد مخصوص این پروتئین در این پایگاه داده یا کد معادل در پایگاه داده های دیگر است .

سوالات عملی

تنها ابزار های مجاز برای استفاده در این بخش سایت های EBI , NCBI , Uniport می باشند و استفاده از اینترنت (غیر از موارد ذکر شده) و یا ابزار های آفلاین (مانند نوت پد یا اکسل) ممنوع می باشد و با فرد خاطی برخورد خواهد شد .

۱. برای هر کدام از بیماری های ژنتیکی زیر : در کل 50 نمره
 - توضیح کوتاهی از پاتولوژی بیماری به فارسی بنویسید . 1 نمره
 - ژن یا ژن ها و پروتئین یا پروتئین های درگیر در بیماری را بیابید . 2 نمره
 - عملکردهای احتمالی پروتئین یا پروتئین های درگیر در بیماری را بنویسید . 1 نمره
 - توالی بخش CDC ژن یا ژن های درگیر در بیماری را در فرمت FASTA در فایل Q1.TEXT در پوشه Q1 ذخیره کنید (اسم هر بیماری را قبل توالی های مورد نظر ذکر کنید) . 3 نمره
 - جهش یا جهش های عامل بیماری را مشخص کنید (در صورت تعداد بالا 3 مورد ذکر شود) . 2 نمره
 - مکان تجمع و قرارگیری پروتئین ها را معین کنید . 1 نمره

فاویسم (Favism)

هانتینگتون (Huntington) :

فنیل کتونوریا (Phenylketonuria) :

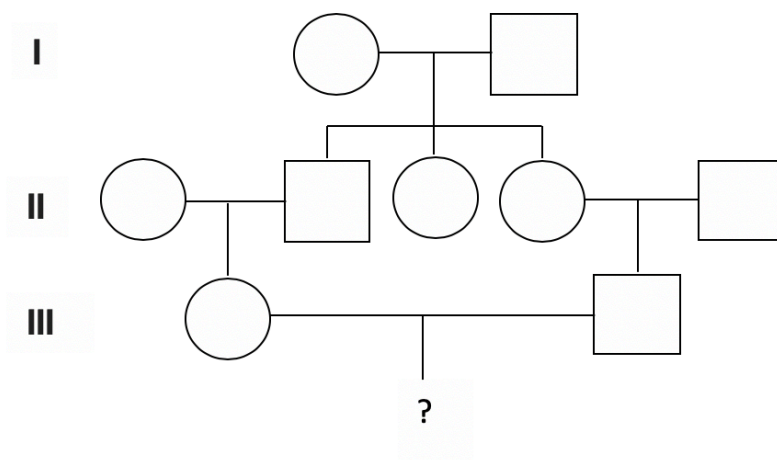
آلکاپتونوریا (Alkaptonuria) :

دیستروفی عضلانی دوشن (Duchenne muscular dystrophy) :

برای سوالات بعدی به توضیحات زیر توجه کنید :

کم‌خونی داسی شکل یک بیماری ژنتیکی و یک اختلال خونی است که با هموگلوبین معیوب (HbSS) شناخته می‌شود. سلول‌های هموگلوبین سلول داسی سفت و به شکل داس می‌باشند. این مولکول‌های هموگلوبین تمایل دارند به شکل خوشه‌ای و در کنار یکدیگر قرار گیرند، بنابراین به راحتی از مویرگ‌های خونی عبور نمی‌کنند. سلول‌های طبیعی بیش از ۱۲۰ روز زنده هستند، سلول‌های داسی بعد از ۲۰-۱۰ روز از بین می‌روند. این روند طی یک دوره مزمن منجر به بروز کم‌خونی می‌گردد.

هر دو نسخه ژن بتا هموگلوبین افراد یک شجره نامه توالی یابی شد و توالی آن‌ها در پوشه Q2 در اختیار شما گرفته است، برای هر فرد دو توالی (دو کروموزوم) وجود دارد که با L1 و L2 مشخص شده اند، با توجه به توالی‌ها و اطلاعات بالا به سوالات زیر پاسخ دهید. فرض کنید فراوانی آلل این بیماری در جامعه برابر 10 درصد می‌باشد.



۱. ژنوتیپ تمام افراد شجره را مشخص کرده و بنویسید. در مواردی که پاسخ به صورت قطعی مشخص نیست حالت‌های ممکن و احتمال هر یک را بنویسید. 9 نمره

۲. چه قدر احتمال دارد فرزند مشخص شده با علامت سوال , دختری بیمار باشد ؟ 1 نمره

۳. صحت صحت گزاره های زیر را بررسی کنید . 20 نمره (هر پاسخ درست +25 هر پاسخ نادرست -2)

الف) برای تشخیص اولیه هانتینگتون در جمعیت باید از تستی با حساسیت بالا استفاده کنیم .

ب) انتظار می رود در خون بیمار مبتلا به فنیل کتونوریای کنترل نشده شاهد مقادیر افزایش یافته هورمون های تیروئیدی آزاد باشیم .

پ) بیماری گلبول قرمز داسی شکل را می توان یک بیماری اتوزومی غالب با تجلی پذیری پایین در حالت هتروزیگوت در نظر گرفت .

ت) هموگلوبین معیوب (Hbss) در کانفورماسیون Tight نسبت به کانفورماسیون Relax تمایل بیشتری به تجمع و پلیمر شدن دارد .

ج) به دلیل فشار سهمی اکسیژن کمتر در خون جنین , تجلی پذیری بیماری گلبول قرمز داسی شکل در دوران جنینی بیشتر است.

د) در بیماری فاویسم میزان بیان ژن گلوکاتیون در گلبول های قرمز افزایش می یابد.

ه) با توجه به اطلاعات داده شده , خانواده مورد بررسی ما احتمالا از مناطقی با شیوع بالای مالاریا انتخاب شده است .

ی) در بیماری بتا تالاسمی , جهش ایجاد شده در زنجیره بتا از نوع Missense می باشد.