

آزمایشگاه آموزشی
بیست و دومین المپیاد
زیست‌شناسی ایران

بیوانفورماتیک

Alignment. ادامه ابزارها.

روز هفتم
۹۸/۵/۱۵

اهداف آزمایش:

۱. ادامه آشنایی با ابزارهای بیوانفورماتیک
۱. آشنایی با دانش تئوری مربوط به Alignment

زمان آزمایش: ۹۰ دقیقه



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

تئوری Alignment

PAM1

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Arg	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Asn	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asp	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Cys	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Gln	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Glu	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Gly	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
His	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
Ile	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Leu	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Lys	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Met	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Phe	1	1	0	0	0	0	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Pro	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Ser	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Thr	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Trp	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Tyr	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Val	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

PAM10

A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-20	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8
A																				
R																				
N																				
D																				
C																				
Q																				
E																				
G																				
H																				
I																				
L																				
K																				
M																				
F																				
P																				
S																				
T																				
W																				
Y																				
V																				

PAM10 log odds scoring matrix

PAM250

[illegible]

		BLOSUM62 matrix																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

ادامه معرفی ابزارهای بیوانفورماتیک

EBI

با رفتن به بخش سرویس‌ها می‌توان به تمامی امکانات EBI دسترسی پیدا کرد.

Services

Overview | A to Z | Data submission | Support

Are you using EMBL-EBI's public data resources?
Please help us make them better and share your feedback. Take the survey ►

The European Bioinformatics Institute (EMBL-EBI) maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.

Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our [web services](#) to access our resources programmatically.

— You can read more about our services in the journal *Nucleic Acids Research*

Tools & Data Resources

Search all tools & data resources

Tools

Clustal Omega
Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW

Data resources

Ensembl
Genome browser, API and database, providing access to reference genome annotation

Browse by type

- DNA & RNA
- Gene Expression
- Proteins
- Structures
- Systems
- Chemical biology

Overview | A to Z | Data submission | Support

InterProScan
InterProScan searches sequences against InterPro's predictive protein signatures.
Protein feature detection | Sequence motif recognition

BLAST [protein]
Fast local similarity search tool for protein sequence databases.
Sequence similarity search

BLAST [nucleotide]
Fast local similarity search tool for nucleotide sequence databases.
Sequence similarity search

HMMER
Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.
Sequence similarity search | Protein function analysis

PDBe
The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes.

Europe PMC
A database to search the worldwide life sciences literature

Expression Atlas
An added-value database that shows which genes/proteins are expressed under which conditions, and how expression differs between conditions.

ChEMBL
An open data resource of binding, functional and ADMET bioactivity data.
See all data resources ►

Programmatic access

EMBL-EBI web services allow you to query our large biological data resources programmatically, so that you can develop data analysis pipelines or integrate public data with your own applications. The Web Services technology we use are built on open standards to ensure client and server software from various sources will work well together.

[Browse EMBL-EBI web services](#)

Principles of service provision

Open
Our data and tools are freely available, without restriction. The only exception is potentially identifiable human genetic information, for which access depends on research consent agreements.

Compatible
EMBL-EBI is a world leader in the development of global bioinformatics standards, which are key to data sharing.

Comprehensive
Thanks to our many data-sharing agreements, EMBL-

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.

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

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

Needle (EMBOSS)

EMBOSS Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm.

Stretcher (EMBOSS)

EMBOSS Stretcher uses a modification of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned.

Water (EMBOSS)

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

GeneWise

GeneWise compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

Tools > Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, [Pairwise Sequence Alignment](#) tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

[Launch Clustal Omega](#)

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

[Launch Kalign](#)

MAFFT

MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.


[Launch MAFFT](#)

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.



Expression Atlas

Gene expression across species and biological conditions

Query single cell expression

To Single Cell Expression Atlas

Home

Browse experiments

Download

Release notes

FAQ

Help

Licence

About

Support

Search across 61 species, 3,564 studies, 112,210 assays

Ensembl 96, Ensembl Genomes 43, WormBase ParaSite 12, EFO 3.5

Search

Gene set enrichment

Gene / Gene properties

Enter gene query...

Examples: REG1B, zinc finger, O14777 (UniProt), GO:0010468 (regulation of gene expression)

Search

Clear

Species

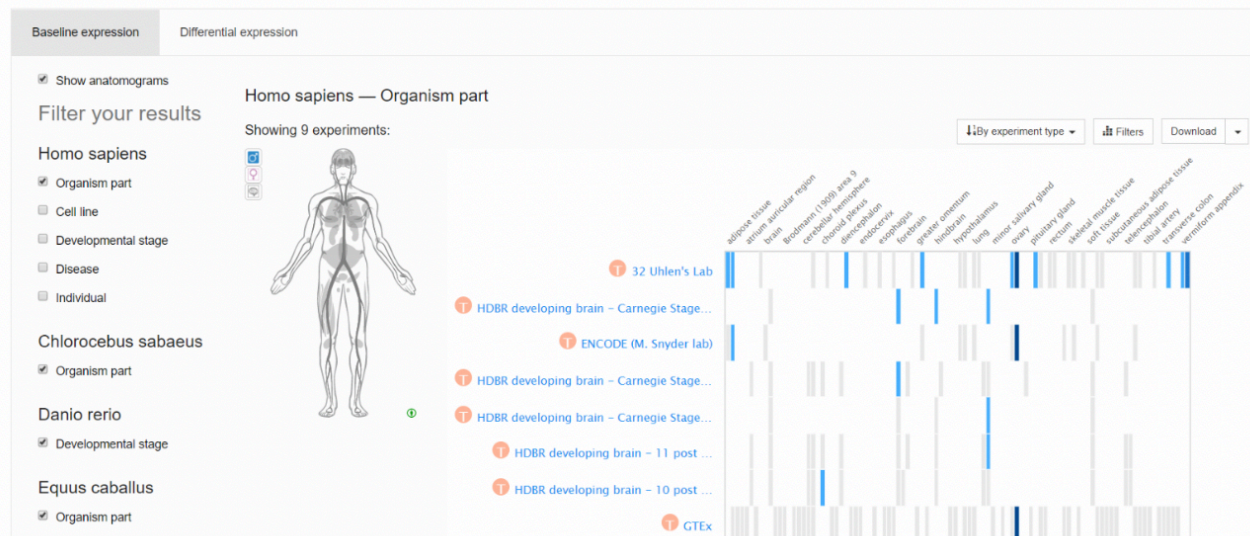
Any

Biological conditions

Examples: lung, leaf, valproic acid, cancer

Show experiments of all species

Results for ins

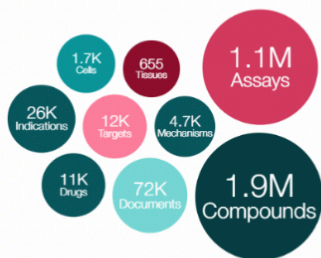


ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs.

Explore ChEMBL

Description: Shows a summary of the ChEMBL entities and quantities of data for each of them.

Instructions: Click on a bubble to explore a specific ChEMBL entity in more detail.



[Browse all ChEMBL](#)

[See all visualisations](#)

ChEMBL Dopamine

All Results 8733 Compounds 79 Targets 53 Assays 7368 Documents 1233 Cells 0 Tissues 0

Compounds

Show Full Query

79 Compounds
0 Selected - Select All
Browse Activities

Table Cards Graph Heatmap

Filters

Type

Small molecule 79

Max Phase

0 39
1 4
2 7
3 1
4 28

#RO5 Violations

0 65

Showing 1-24 out of 79 records

Records per page: 24

Select All

CHEMBL59
Name: DOPAMINE
Max Phase: 4
Full Mwt: 153.18
Aloggp: 0.6

CHEMBL1557
Name: DOPAMINE HYDROCHLORIDE
Max Phase: 4
Full Mwt: 189.64
Aloggp: 0.6

CHEMBL457419
Name: PALMITOYL DOPAMINE
Max Phase: 0
Full Mwt: 391.6
Aloggp: 6.24

CHEMBL138921
Name: ARACHIDONOYL DOPAMINE
Max Phase: 0
Full Mwt: 439.64

CHEMBL138040
Name: LINOLEOYL DOPAMINE
Max Phase: 0
Full Mwt: 413.6

CHEMBL456176
Name: No Data
Max Phase: 0
Full Mwt: 413.6

ChEMBL Search in ChEMBL

EBI > Databases > Chemical Biology > ChEMBL Database > ChEMBL59

Compound Report Card

Name And Classification

Structure Search

ID: ChEMBL59
Name: DOPAMINE
Max Phase: 4 Approved
Molecular Formula: C₈H₁₁NO₂
Molecular Weight: 153.18
ChEMBL Synonyms: Carbilev, DOPAMINE, Dopamine, Intropin, Parcopa, Sinemet
Synonyms From Alternate Forms: ASL-279, DOPAMINE HYDROCHLORIDE
Trade Names From Alternate Forms: DOPAMINE HYDROCHLORIDE, DOPMIN, Dopamin-Natterman, Dynatra, INTROPIN
SELECTAJET, Sabax Dopamin
Molecule Type: Small molecule

Name And Classification

Representations

Sources

Molecule Features

Mechanism Of Action

Indications

Clinical Data

Activity Charts

Literature

Target Predictions

Calculated Properties

Structural Alerts

Cross References

UniChem Cross References

UniChem Connectivity Layer Cross References

Alternative Forms

سوالات

سوال ۱. دو توالی DNA داده شده را به روش برنامه ریزی پویا به صورت سراسری هم ردیف کنید و نتیجه نهایی هم ردیفی و امتیاز آن را بنویسید .

Pairwise Global Alignment By Dynamic Programing !

Match = +5 / Mis Match = - 1 / Gap = 0

1) GTTC

2) ATTGC

سوال ۲. با توجه به جدول پنالتی‌های داده شده برای هم‌ردیفی بالا درستی یا نادرستی گزاره‌های زیر را مشخص کنید.

الف. توالی داده شده احتمالاً مربوط به اینترون یک ژن می‌باشد.

ب. توالی داده شده احتمالاً مربوط به یک ژن بسیار حفاظت شده می‌باشد.

ج. نرخ جهش حذف در جانداران مورد بررسی ما کمتر از میانگین مورد انتظار در جاندارن مشابه می‌باشد.

د. نتیجه نهایی هم‌ردیفی این دو توالی نسبت به تغییر جدول پنالتی‌ها پایدار است (حساس نیست).

سوال ۳. فرمول توزیع گامبل (توزیع فراوانی Score های نهایی هم‌ردیفی در صورت استفاده از توالی‌های رندم) برای هم‌ردیفی سوال ۱ به صورت $P = 1 - e^{-0.014 * e^{Score}}$ می‌باشد. مقدار **p-value** این هم‌ردیفی را بدست آورید و بگویید آیا هومولوژی / آنالوژی دو توالی بالا رد می‌شود یا خیر؟ (آلفا برابر ۵ درصد)

راهنمایی یک: اگر در نظر بگیریم تست آماری ما one-tail می‌باشد در این صورت میزان p-value برابر مساحت زیر نمودار توزیع گامبل از اسکور هم‌ردیفی مورد نظر ما تا اسکور مثبت بی‌نهایت است! البته حواستون به one-tail یا two-tail بودن تست‌مون هم باشه!