Протеинске базе података

Циљеви часа

- Постранслационе модификације протеина
- Разумевање Swiss-Prot уноса
- Специјализоване базе података као KEGG (базе метаболичких путања) или PDB (базе структура)

План часа

1. Прелазак од гена до функционалног протеина
2. Читање UniProt/Swiss-Prot уноса
3. Истраживање метаболичких база као KEGG
4. Информације о поттранслационим модификацијама протеина

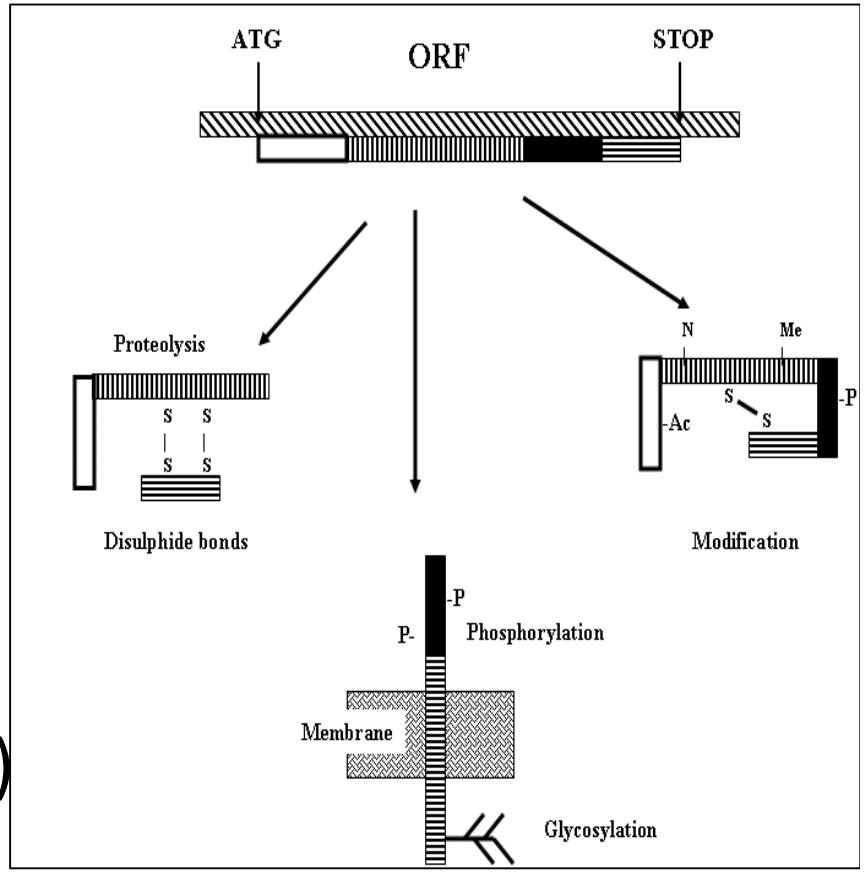
Од гена до функционалног протеина

- Гени који кодирају протеине су транскрибовани у иРНК
- иРНК се транслира у протеине
- Често протеин мора да прође кроз посттранслационе модификације пре него што постане активан.
- Протеини затим морају да се транспортују до своје дестинације
  - Ћелијско једро
  - Митохондрија или нека друга органела
  - Периплазма (бактерија)
  - Изван ћелије
- Протеин је **функционалан** када дође на место где обавља своју функцију!

Посттранслационе модификације

Посттранслационе модификације укључују:

- Уклањање делова протеина
- Сечење крајева протеина
- Хемијске модификације
- Фосфорилација
- Везивање липида или угљених хидрата (гликација)



Информација о протеину

- Да би разумели како ваш протеин ради, треба да разумете
  - Његове посттранслационе модификације
  - Како се транспортује
  - Механизам функционисања
- Сва ова информација мора да се експериментално одреди
- Ако је урађено, онда је најчешће (нажалост не увек) у Swiss-Prot-у

Swiss-Prot база података

- Уноси описују протеине са познатим функцијама
- Релативно мала база, без преклапања: укупно 500,000 протеина
  - trEMBL садржи 70 милиона предвиђених протеина из GenBank
  - Swiss-Prot садржи подскуп од trEMBL са познатом функцијом
- У Swiss-Prot-у сви уноси су ручно анотирани
- Најтачнија база са протеинским функцијама
- Приступ Swiss-Prot је са www.expasy.ch

Задатак

- Нађите Swiss Prot унос за “human epidermal growth factor receptor”.
- Идентификујте “primary accession number” за овај протеин.
- Пробајте да укуцате старији “primary accession number” за овај протиен нпр. P06268.
- Испитајте значење различитих поља у Swiss prot уносу.

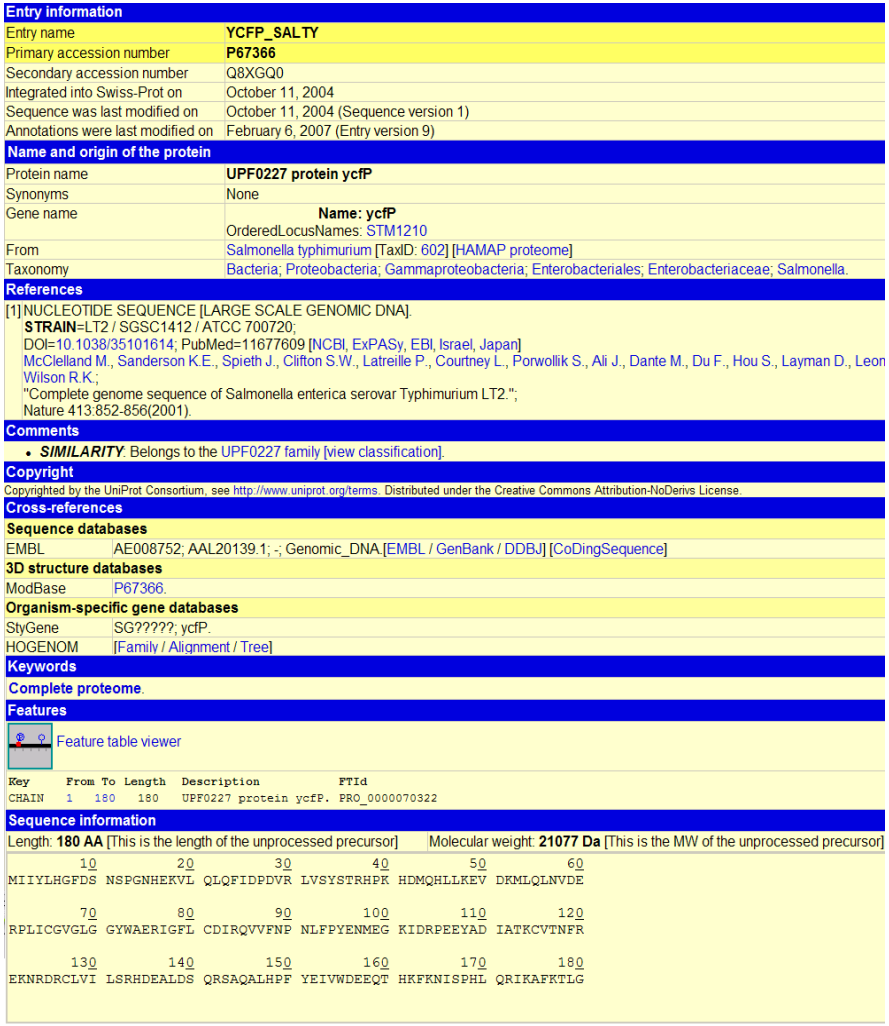
Swiss-Prot унос

Овај унос је на www.expasy.org/uniprot/P00533

Entry information	
Entry name	**EGFR_HUMAN**
Primary accession number	**P00533**
Secondary accession numbers	O00688 O00732 P06268 Q14225 Q92795 Q9BZS2 Q9GZX1 Q9H2C9 Q9H3C9 Q9UMD7 Q9UMD8 Q9UMG5
Integrated into Swiss-Prot on	July 21, 1986
Sequence was last modified on	November 1, 1997 (Sequence version 2)
Annotations were last modified on	March 20, 2007 (Entry version 111)
Name and origin of the protein	
Protein name	**Epidermal growth factor receptor [Precursor]**
Synonyms	**EC 2.7.10.1** **Receptor tyrosine-protein kinase ErbB-1**
Gene name	**Name: EGFR** Synonyms: ERBB1
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
References	
[1]	NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1). DOI=10.1038/309418a0; PubMed=6328312 [NCBI, ExPASy, EBI, Israel, Japan] Ullrich A., Coussens L., Hayflick J.S., Dull T.J., Gray A., Tam A.W., Lee J., Yarden Y., Libermann T.A., Schlessinger J., Downward J., Mayes E.L.V., Whittle N., Waterfield M.D., Seeburg P.H.; "Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells."; Nature 309:418-425(1984).

Основне секције Swiss-Prot уноса

- Општа информација
  - Accession number (идентификациони број)
- References
  - Радови који се односе на протеинску секвенцу
- Comment section
  - Информација о функцији
- Cross-references
  - Линкови на друге базе података
- Feature table
  - Мапирање сваке познате функције
- Sequence
  - Секвенца протеина



Entry information

Entry name	**YCFP_SALTY**
Primary accession number	**P67366**
Secondary accession number	Q8XGQ0
Integrated into Swiss-Prot on	October 11, 2004
Sequence was last modified on	October 11, 2004 (Sequence version 1)
Annotations were last modified on	February 6, 2007 (Entry version 9)

Name and origin of the protein

Protein name	**UPF0227 protein ycfP**
Synonyms	None
Gene name	**Name: ycfP** OrderedLocusNames: STM1210
From	Salmonella typhimurium [TaxID: 602] [HAMAP proteome]
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Salmonella.

References

[1] NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
STRAIN=LT2 / SGSC1412 / ATCC 700720;
DOI=10.1038/35101614; PubMed=11677609 [NCBI, ExPASy, EBI, Israel, Japan]
McClelland M., Sanderson K.E., Spieth J., Clifton S.W., Latreille P., Courtney L., Porwollik S., Ali J., Dante M., Du F., Hou S., Layman D., Leon
Wilson R.K.;
"Complete genome sequence of Salmonella enterica serovar Typhimurium LT2.";
Nature 413:852-856(2001).

Comments

- ***SIMILARITY***: Belongs to the UPF0227 family [view classification].

Copyright

Cross-references

Sequence databases	
EMBL	AE008752; AAL20139.1; -; Genomic_DNA.[EMBL / GenBank / DDBJ] [CoDingSequence]
3D structure databases	
ModBase	P67366.
Organism-specific gene databases	
StyGene	SG?????; ycfP.
HOGENOM	[Family / Alignment / Tree]

Keywords

Complete proteome.

Features

Feature table viewer

Key	From	To	Length	Description	FTId
CHAIN	1	180	180	UPF0227 protein ycfP.	PRO_0000070322

Sequence information

Length: **180 AA** [This is the length of the unprocessed precursor] | Molecular weight: **21077 Da** [This is the MW of the unprocessed precursor]

```
        10         20         30         40         50         60
MIIYLHGFDS NSPGNHEKVL QLQFIDPDVR LVSYSTRHPK HDMQHLLKEV DKMLQLNVDE

        70         80         90        100        110        120
RPLICGVGLG GYWAERIGFL CDIRQVVFNP NLFPYENMEG KIDRPEEYAD IATKCVTNFR

       130        140        150        160        170        180
EKNRDRCLVI LSRHDEALDS QRSAQALHPF YEIVWDEEQT HKFKNISPHL QRIKAFKTLG
```

Општа информација у Swiss-Prot уносу

- **Entry Name**
  - Идентификује унос
  - Може да се промени ако се два уноса споје у један
- **Primary Accession Number**
  - Има форму PXXXX
  - Сталан је број и никад се не мења
- **Last Modified** – информација када је унос задњи пут модификован
- **Protein Name** (име протеина) и **Synonyms (**синоними) дају нека уобичајена имена протеина
- **From** и **Taxonomy** указују одакле протеин долази
- **References** садржи листу свих радова који се користе да би се саставио тај унос

Секција са коментарима

- **Comments** секција наводи све познате функције протеина.
- Ова секција је драгоцен документ који “ручно”
- Коментари се односе на скуп “стандардних” тема (види табелу)

Topic	Description
ALTERNATIVE PRODUCTS	Description of the existence of related protein sequence(s) produced by alternative splicing of the same gene or by the use of alternative initiation codons
BIOTECHNOLOGY	Description of the use of a specific protein in a biotechnological process
CATALYTIC ACTIVITY	Description of the reaction(s) catalyzed by an enzyme [1]
CAUTION	This topic warns you about possible errors and/or grounds for confusion
COFACTOR	Description of an enzyme cofactor
DATABASE	Description of a cross-reference to a network database/resource for a specific protein [2]
DEVELOPMENTAL STAGE	Description of the developmental specific expression of a protein
DISEASE	Description of the disease(s) associated with a deficiency of a protein
DOMAIN	Description of the domain structure of a protein
ENZYME REGULATION	Description of an enzyme regulatory mechanism
FUNCTION	General description of the function(s) of a protein
INDUCTION	Description of the compound(s) which stimulate the synthesis of a protein
MASS SPECTROMETRY	Reports the exact molecular weight of a protein or part of a protein as determined by mass spectrometric methods [3]
MISCELLANEOUS	Any comment which does not belong to any of the other defined topics
PATHWAY	Description of the metabolic pathway(s) with which a protein is associated
PHARMACEUTICAL	Description of the use of a specific protein as a pharmaceutical drug
POLYMORPHISM	Description of polymorphism(s)
PTM	Description of a posttranslational modification
SIMILARITY	Description of the similarity(s) (sequence or structural) of a protein with other proteins
SUBCELLULAR LOCATION	Description of the subcellular location of the mature protein
SUBUNIT	Description of the quaternary structure of a protein
TISSUE SPECIFICITY	Description of the tissue specificity of a protein

Секција са коментарима за P00533

Comments

- ***FUNCTION***: Receptor for EGF, but also for other members of the EGF family, as TGF-alpha, amphiregulin, betacellulin, he control of cell growth and differentiation.
- ***FUNCTION***: Isoform 2/truncated isoform may act as an antagonist.
- ***CATALYTIC ACTIVITY***: ATP + a [protein]-L-tyrosine = ADP + a [protein]-L-tyrosine phosphate.
- ***SUBUNIT***: Binds RIPK1. CBL interacts with the autophosphorylated C-terminal tail of the EGF receptor. Part of a complex PIK3C2B, maybe indirectly. Interacts with PELP1.
- ***INTERACTION***:
  Self; NbExp=1; IntAct=EBI-297353, EBI-297353;
  Q53FG0:-; NbExp=2; IntAct=EBI-297353, EBI-913954;
  P22681:CBL; NbExp=1; IntAct=EBI-297353, EBI-518228;
  P22682:Cbl (xeno); NbExp=1; IntAct=EBI-297353, EBI-640919;
  P13987:CD59; NbExp=1; IntAct=EBI-297353, EBI-297972;
  P01133:EGF; NbExp=2; IntAct=EBI-297353, EBI-640857;
  P04626:ERBB2; NbExp=2; IntAct=EBI-297353, EBI-641062;
  P21860:ERBB3; NbExp=2; IntAct=EBI-297353, EBI-720706;
  Q15303:ERBB4; NbExp=2; IntAct=EBI-297353, EBI-80371;
  P62993:GRB2; NbExp=2; IntAct=EBI-297353, EBI-401755;
  O00750:PIK3C2B; NbExp=4; IntAct=EBI-297353, EBI-641107;
  P98083:Shc1 (xeno); NbExp=1; IntAct=EBI-297353, EBI-300201;
  P63104:YWHAZ; NbExp=1; IntAct=EBI-297353, EBI-347088;
- ***SUBCELLULAR LOCATION***: Cell membrane; single-pass type I membrane protein. Isoform 2: Secreted protein.
- ***ALTERNATIVE PRODUCTS***: 4 named isoforms [FASTA] produced by alternative splicing.

Name	**1**
Synonyms	p170
Isoform ID	P00533-1
This is the isoform sequence displayed in this entry.	

Name	**2**
Synonyms	p60, Truncated, TEGFR
Isoform ID	P00533-2
Features which should be applied to build the isoform sequence: VSP_002887, VSP_002888.	

Name	**3**
Synonyms	p110
Isoform ID	P00533-3
Features which should be applied to build the isoform sequence: VSP_002889, VSP_002890.	

Name	**4**
Isoform ID	P00533-4
Features which should be applied to build the isoform sequence: VSP_002891, VSP_002892.	

- ***TISSUE SPECIFICITY***: Expressed in placenta. Isoform 2 is also expressed in ovarian cancers.
- ***PTM***: Phosphorylation of Ser-695 is partial and occurs only if Thr-693 is phosphorylated.
- ***PTM***: Monoubiquitinated and polyubiquitinated upon EGF stimulation; which does not affect tyrosine kinase activity or signa 'Lys-63', but linkage through 'Lys-48', 'Lys-11' and 'Lys-29' also occur.
- ***DISEASE***: Defects in EGFR are associated with lung cancer [MIM:211980].

Cross-reference секција

- Садржи линкове ка уносима из других база
- Информације се аутоматски ажурирају

Protein-protein interaction databases	
DIP	DIP:405N; -.
IntAct	P00533; -.
PTM databases	
GlycoSuiteDB	P00533; -.
Polymorphism databases	
NIEHS-SNPs	EGFR.
2D gel databases	
SWISS-2DPAGE	P00533; HUMAN.
Organism-specific gene databases	
HGNC	HGNC:3236; EGFR.
GeneCards	EGFR.
GeneLynx	EGFR; Homo sapiens.
GenAtlas	EGFR.
HPA	CAB000035; -.
MIM	131550; gene. [NCBI / EBI] 211980; phenotype. [NCBI / EBI]
HOVERGEN	[Family / Alignment / Tree]
Gene expression databases	
CleanEx	HGNC:3236; EGFR.
ArrayExpress	P00533; -.
GermOnline	ENSG00000146648; Homo sapiens.

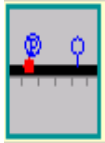
Неке битне Cross-References

- **EMBL**: Ориганална ДНК секвенца из GenBank-а
- **PDB**: Експериментална структура протеина
- **DIP**: Протеини који интерагују са вашим протеином
- **GlycoSuiteDB**: Додавање шећера (гликолизација)
- **MIM:** Листа генетских болести са којима је повезан протеин
- **Ontologies:** Функција протеина
- **Profiles:** Познати домени унутар протеина
- **ENSEMBL:** Полозај протеина у геному

Features секција

- На протеинској секвенци прецизно локализује сваку познату функцију протеина.
- TRANSMEM: Трансмембрански домен
- ACT_SITE: Хемијски активна места
- BINDING: Места за везивање
- DISULPHID: Дисулфидне везе

Features



Feature table viewer

Key	From	To	Length	Description
SIGNAL	1	24	24	
CHAIN	25	1210	1186	Epidermal growth factor receptor.
TOPO_DOM	25	645	621	Extracellular *(Potential)*.
TRANSMEM	646	668	23	*Potential*.
TOPO_DOM	669	1210	542	Cytoplasmic *(Potential)*.
REPEAT	75	300	226	Approximate.
REPEAT	390	600	211	Approximate.
DOMAIN	712	979	268	Protein kinase.
NP_BIND	718	726	9	ATP *(By similarity)*.
COMPBIAS	1025	1071	47	Ser-rich.
ACT_SITE	837	837		Proton acceptor *(By similarity)*.
BINDING	745	745		ATP *(By similarity)*.

Више података о посттранслационим модификацијама протеина

- Протеини се често посттранслационо модификују да би постали активни
- Модификацвија може да значи додавање масти или шећера
- Користите те ресурсе да би нашли податке о посттранслационим модификацијама

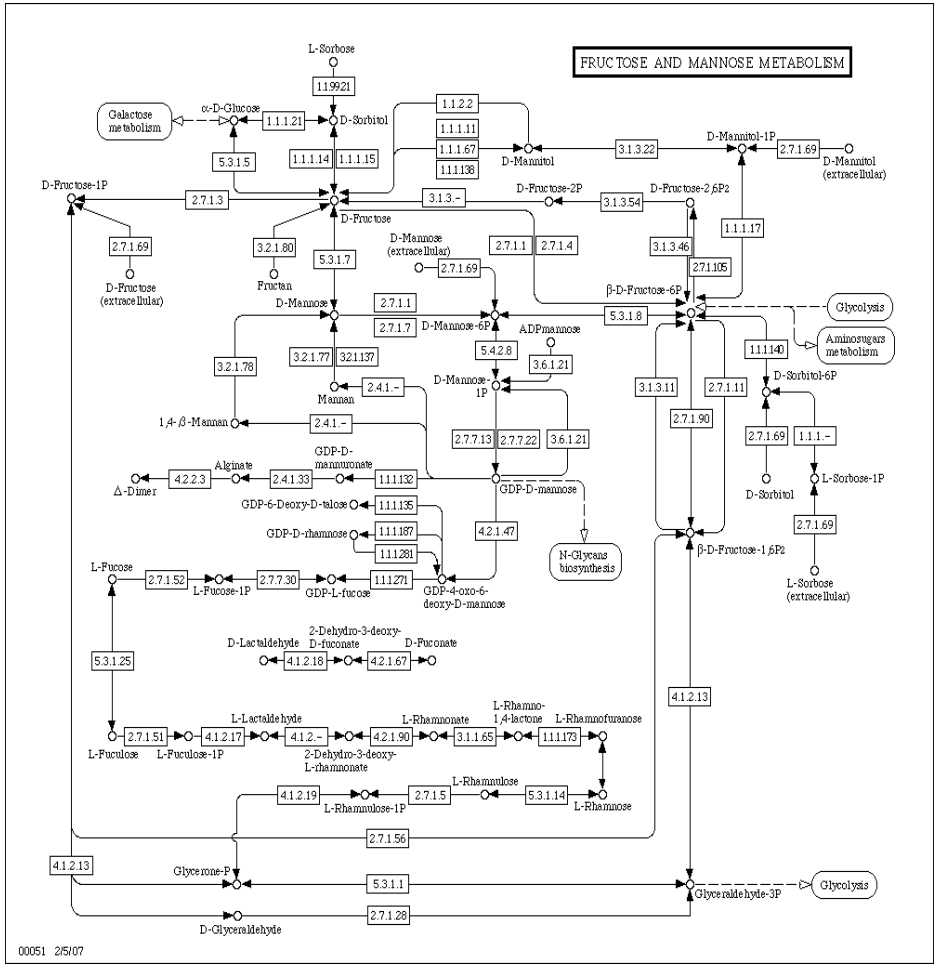
- www.ebi.ac.uk/RESID
  - Садржи сваку познату посттранслациону модификацију
- www.glycosuite.com
  - Комплетна база свих познатих шећера нађених у протеинима
- www.lipidbank.jp
  - База за масти

Функција вашег протеина

- Features и Comments секције вам дају драгоцену информацију о функцији
- Да би разумели функцију вашег протеина, морате да одредите
  - Где ваш протеин ради
  - Метаболичку путању у коју је укључен
  - 3D структуру протеина
  - Којој фамилији протеин припада
- Можете да пратите ове податке тако што пратите линкове у Cross reference секцији

Где ваш протеин ради?

- Протеин је обично део неке метаболичке путање
- Метаболичка путања је као ланац који повезује више различитих протеина
- У метаболичкој путањи се метаболити мењају тако што прелазе од једног ка другом ензиму
- У KEGG путањи, ензими су графички приказани у облику кућица



Неки битни ресурси за метаболичке путање

- www.genome.ad.jp/kegg
  - KEGG је најпотпунија база метаболичких путања
  - Можете да је користите да би упоредили врсте
- www.chem.qmul.ac.uk/iubmb
  - IUBMD одређује EC број који се користи да опише активност датог ензима
- www.ecocy.org
  - Екстензивна листа познатих метаболичких путања у E. coli и другим бактеријама.

Која је структура вашег протеина?

- Протеин мора да има одговарајућу структуру да би обављао своју функцију
- Структура протеина је кључна да би се разумела његова функција
- Предвиђање протеинске структуре је веома тешко
- Прецизна предвиђања захтевају експерименте
  - X-ray кристалографију
  - NMR (нуклеарна магнетна резонанца)
- Предвиђање само из секвенце је могуће али непоуздано

Неке базе података са протеинским структурама

- www.rcsb.org/pdb
  - База протеинских структура
  - “PDB” протеина је често синоним са његовом структуром
- www.ncbi.nlm.nih.gov/Structure
  - Још једна база са протеинским структурама
- swissmodel.expasy.org
  - Предвиђања структуре из секвенци

Неке битне фамилије протеина

- Протеини се могу класификовати у фамилије
- Класификација се заснива и на функцији и на секвенци
- Специјализоване базе података су на располагању за већину важних фамилија

- www.kinasenet.org
  - Kinases контролишу велики број процеса у организму; поремећај у њиховој регулацији је често узрок рака
- imgt.cines.fr
  - Имуноглобулини су кључни елементи у имунитету
- rebase.neb.com
  - Кључан ресурс за рестрикционе ензиме

Закључак

- Предвиђање функције протеина је један од најважнијих циљева у биологији
- Протеинске базе података могу да помогну да се организује знање
- Оне дају материјал да се
  - Развију нови биолошки експерименти
  - Развију нови биоинформатички алгоритми
  - Екстраполирају експериментални податци на нове врсте

Задатак за вежбу

- Испитајте унос за ензим “human dUTPase”
- Који је Accession number за овај протеин?
- Да ли наведена секвенца захтева даље процесирање?
- У којој форми је протеин (мономер, дајмер, трајмер).
- Шта катализује овај ензим?
- Наведи Swiss Prot primary accession протеина са којим dUTP интерагује
- Са колико експеримената је подржана та интеракција

- Колико изоформи има овај протеин?
- У чему се те изоформе разликују од канонске секвенце?
- Који део протеинске секвенце се уклони током посттранслационе модификације?
- Сачувајте FASTA секвенцу за изоформу 2
- Колико радова је коришћено да би се направио унос за овај протеин (примедба: идете на “Display”, па на “Publications”, левом делу)?
- У KEGG бази прикажите метаболичку путању везану за овај протеин.
- Искористите линк на PDB базу, да прикажете експериментално одређену 3D структуру протеина која одговара резидуалима 112-252.