

Нуклеинске базе података

Циљеви предавања

- Разликовање структуре еукариотских и прокариотских гена
- Разумевање GenBank формата уноса
- Разумевање разлике између GenBank-а и геноцентричних база података

План часа

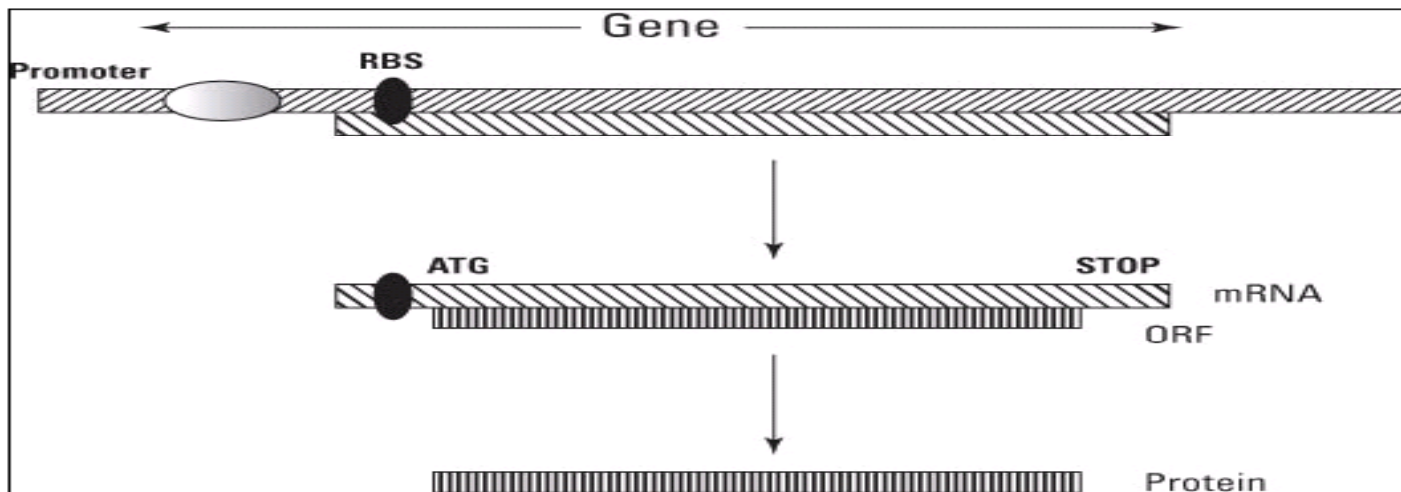
1. Подсетник о генима и геномима
2. Претраживање GenBank-а (ДНК базе података)
3. Коришћење геноцентричних база
4. Анализа генома бактерија

Типични геном прокариота

- ✓ Прокариоти су микроскопски организми
- ✓ Имају кружни геном
- ✓ Дужина је пар милиона bp (0.6 – 10 Mb)
- ✓ Прокариоти имају 1 ген по Kbp
- ✓ 70 % њихових генома кодира протеине

Типични ген прокариота који кодира протеине

- Ген има непрекидну ДНК секвенцу
- Прокариотска mRNA садржи
 - Везивно место за рибозом (RBS)
 - The Open Reading Frame (ORF) – секвенца која кодира протеин
 - У оперонима, транскрипт може да садржи више ORFs

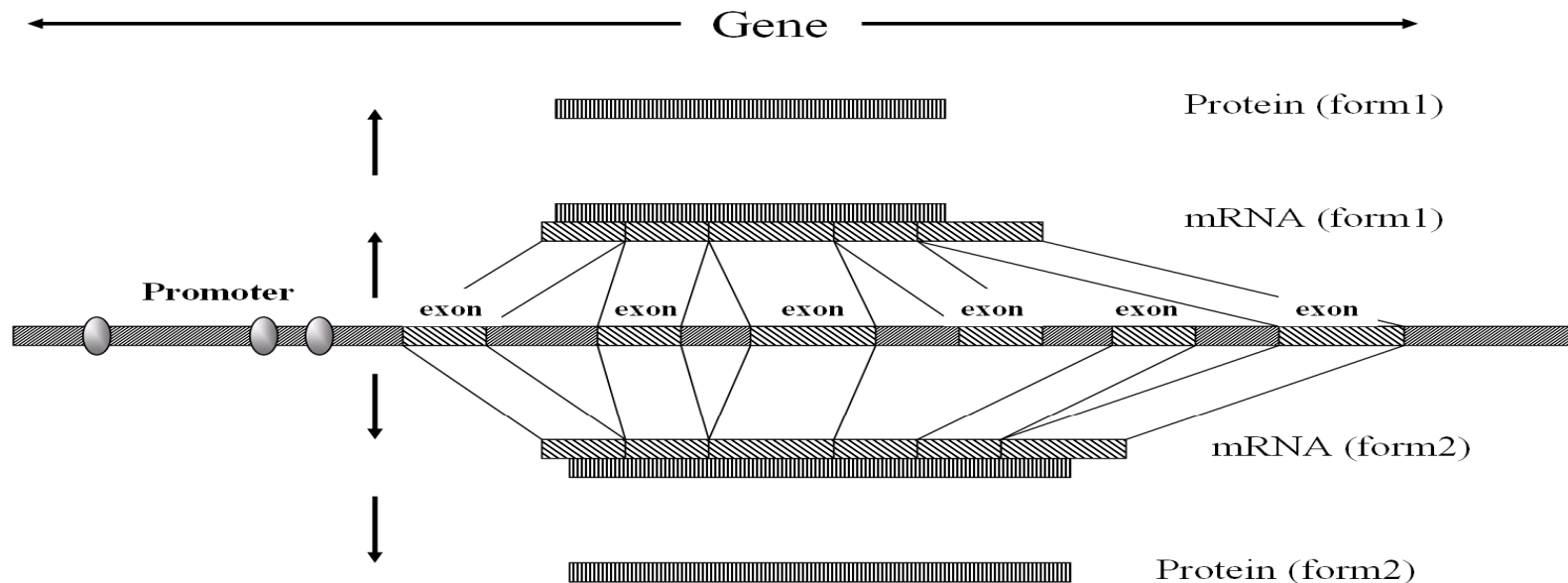


Типични еукариотски геном

- Еукариоти могу бити мали (квасац) или велики (кит)
- Геноми се састоје од више линеарних делова који се зову хромозоми
- Један хромозом: 10 to 700 Mb
- Хумани геном
 - Садржи 22+1 хромозома
 - Дуг је 3 Gb long
- Један ген на сваких 100 Kb (човек)
- 5 % генома кодира за протеин

Типични еукариотски ген који кодирала протеин

- Кодирајућа секвенца се састоји од кодирајућих **егзона** који су прекинути **интронима**
- Интрони се исеку из ДНК секвенце а егзони се слепа да би направили ORF
- Један ген кодирала више алтернативних протеина: **алтернативно спајање (splicing)**



Прокариоти vs. еукариоти

- Прокариоти

- Геноми=један дугачки кружни хромозом + пар малих хромозома (плазмиди)
- 0.5 to 8 Mb / хромозому
- Гени у једном делу
- 70% генома је кодирајуће
- 1 ген по Kb

- Еукариоти

- Геном= много великих линеарних хромозома
- 10 до 700 Mb / хромозому
- Гени су подељени
- 5% генома је кодирајуће
- 1 ген/ 100 Kb (човек)

GenBank

- Налази се у Националном центру за биотехнологију (NCBI)
- GenBank је “меморија” биолошких наука
- Садржи **СВАКУ** ДНК секвенцу која је икад објављена
- GenBank је оригинални извор информација за већину других база података
- GenBank је компликованији за употребу од геноцентричних база података

Читање прокариотских GeneBank података

- **ACCESSION** је идентификациони број
 - Јединствен за сваки унос
 - Трајан
- **LOCUS** садржи информацију о дужини гена
- **ORGANISM** дефинише информацију о организму из које долази ген
- **REFERENCE** даје информацију ко је одредио секвебцу
- **FEATURES** наводи неке функционалне податке о гену
- GenBank унос може да садржи информацију о више од једном гену

NCBI Sequence Viewer - Netscape 6

File Edit View Search Go Bookmarks Tasks Help

NCBI

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display default Save Text Add to Clipboard

1: X01714. E. coli dut gene ... [gi:41296] Related Sequences, Protein, PubMed, Taxonomy

LOCUS ECDUT 1609 bp DNA linear BCT 12-SEP-1993

DEFINITION E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine 5'-triphosphate nucleotidohydrolase).

ACCESSION X01714

VERSION X01714.1 GI:41296

KEYWORDS dUTPase; unidentified reading frame.

SOURCE Escherichia coli.

ORGANISM [Escherichia coli](#)
Bacteria; Proteobacteria; gamma subdivision; Enterobacteriaceae; Escherichia.

REFERENCE 1 (bases 1 to 1609)
AUTHORS Lundberg,L.G., Thoresson,H.O., Karlstrom,O.H. and Nyman,P.O.
TITLE Nucleotide sequence of the structural gene for dUTPase of Escherichia coli K-12
JOURNAL EMBO J. 2 (6), 967-971 (1983)
MEDLINE [84057777](#)
PUBMED [6139280](#)

COMMENT Data kindly reviewed (25-NOV-1985) by L. Lundberg.

FEATURES Location/Qualifiers

source	1..1609 /organism="Escherichia coli" /db_xref="taxon:562"
promoter	286..291 /note="-35 region"
promoter	310..316 /note="-10 region"
misc feature	322..324 /note="put. transcription start region"
RBS	330..333 /note="put. rRNA binding site"
CDS	343..798

Démarrer NCBI ... Adobe... C:WI... Chapt... 21:58

FEATURE секција GenBank уноса

- Промотор
 - Даје прецизне координате промотера
 - Може да постоји више од једног промотера
- RBS даје координате места за везивање рибозома (Ribosome Binding Site)
- CDS даје особине **CoDing Sequence** (кодирајуће секвенце) који кодира протеине



```
NCBI Sequence Viewer - Netscape 6
File Edit View Search Go Bookmarks Tasks Help
FEATURES
Location/Qualifiers
source
1..1177
/organism="Homo sapiens"
/db_xref="taxon:9606"
/map="15q15-q21.1"
gene
order(AFO18429.1:<1..1735,1..1177,AFO18431.1:1..45,
AFO18432.1:658..732,AFO18432.1:884..954,
AFO18432.1:1391..>1447)
/gene="DUT"
mRNA
join(AFO18429.1:<282..561,AFO18429.1:1034..1172,
560..651,AFO18431.1:1..45,AFO18432.1:658..732,
AFO18432.1:884..954,AFO18432.1:1391..>1447)
/gene="DUT"
/product="dUTPase"
/note="alternatively spliced; encodes mitochondrial form
of the protein"
CDS
join(AFO18429.1:282..561,AFO18429.1:1034..1172,560..651,
AFO18431.1:1..45,AFO18432.1:658..732,AFO18432.1:884..954,
AFO18432.1:1391..1447)
/gene="DUT"
/note="DUT-M; alternatively spliced; mitochondrial form of
the protein; similar to H. sapiens dUTPase encoded by
GenBank Accession Number U90224"
/codon_start=1
/product="dUTPase"
/protein_id="AAB71393.1"
/db_xref="GI:2443580"
/translation="HTPLCPRPALCYHFLTSLLRSAQNARGTAEGRSRGLRARPAP
RPPAAQHGIPLPLSSAGRLSQCGASTVGAAGWKGELPKAGGSPAPGPETPAISPSK
RARPAEVEGGHQLRFARLSEHATAPTRGSARAAGYDLVSYDYTIIPPEKAVVKTDIQI
ALPSGCYGRVAPRSLAAKHFIDVGAQVIDEDYRGNVGVVLFNFGKEKFEVFKGDRIA
QLICERIFYPEIEEVQALDDTERGSGGFGSTGKN"
mRNA
join(AFO18429.1:<1018..1172,560..651,AFO18431.1:1..45,
AFO18432.1:658..732,AFO18432.1:884..954,
AFO18432.1:1391..>1447)
/gene="DUT"
/product="dUTPase"
/note="alternatively spliced; encodes nuclear form of the
protein"
CDS
join(AFO18429.1:1018..1172,560..651,AFO18431.1:1..45,
AFO18432.1:658..732,AFO18432.1:884..954,
AFO18432.1:1391..1447)
/gene="DUT"
```

Пример

Идите на Genbank и искористите accession number X01714 да би лоцирали податке наведене на претходним слајдовима.

Задатак

CRP (cyclic AMP receptor protein) је транскрипциони фактор (протеин који регулише транскрипцију) код бактерија. Лоцирајте унос у Genbank-у који одговара ДНК секвенци *crp* гена код *E coli* (K12).

Решење

- Идите на Swiss Prot базу података и лоцирајте протеин који одговара CRP *E coli*
- Ка дну Swiss Prot стране идите на Sequence Databases под Cross References.
- Копирајте први accession number - J01598 (у пракси некад морате да испробате више опција)
- Идите са овим accession number – ом у GenBank.
- Прођите кроз овај унос и коментаришите.

Читање еукариотског GeneBank уноса

- Секције су исте као у прокариотском уносу
- SOURCE садржи **map** секцију која даје податке о хромозому који садржи ген
- GENE садржи информацију како да се реконструише CDS из гена
- Запамти: Еукариотски гени су испрекидани интронима

Реконструкција CDSs из GenBank уноса

- **Ген, mRNA, и CDS** секције вам дају информацију који делови морају да се споје да би се реконструисао ген, mRNA, или CDS

```
gene          order (AF018429.1:<1..1735,1..1177,AF018431.1:1..45,  
          AF018432.1:658..732,AF018432.1:884..954,  
          AF018432.1:1391..>1447)  
          /gene="DUT"  
mRNA         join(AF018429.1:<282..561,AF018429.1:1034..1172,  
          560..651,AF018431.1:1..45,AF018432.1:658..732,  
          AF018432.1:884..954,AF018432.1:1391..>1447)  
          /gene="DUT"  
          /product="dUTPase"  
          /note="alternatively spliced; encodes mitochondrial form  
          of the protein"  
CDS         join(AF018429.1:282..561,AF018429.1:1034..1172,560..651,  
          AF018431.1:1..45,AF018432.1:658..732,AF018432.1:884..954,  
          AF018432.1:1391..1447)
```

Спајање CDS-а из GenBank уноса

- Ген може да кодира више mRNA молекула
- Пример: dUTPase ген кодира
 - Митохондријску dUTPase
 - dUTPазу у језгру

Table 3-1 Mitochondrial or Nuclear dUTPase mRNAs in AF018430

<i>mRNA</i>	<i>AF018429</i>	<i>AF018430</i>	<i>AF018431</i>	<i>AF018432</i>
Type 1 (Mitochondria)	282-561 1034-1172	560-651	1-45	658-732 884-954 1391-1447 ->
Type 2 (Nuclear)	<1018-1172	560-651	1-45	658-732 884-954 1391-1447 ->

Пример

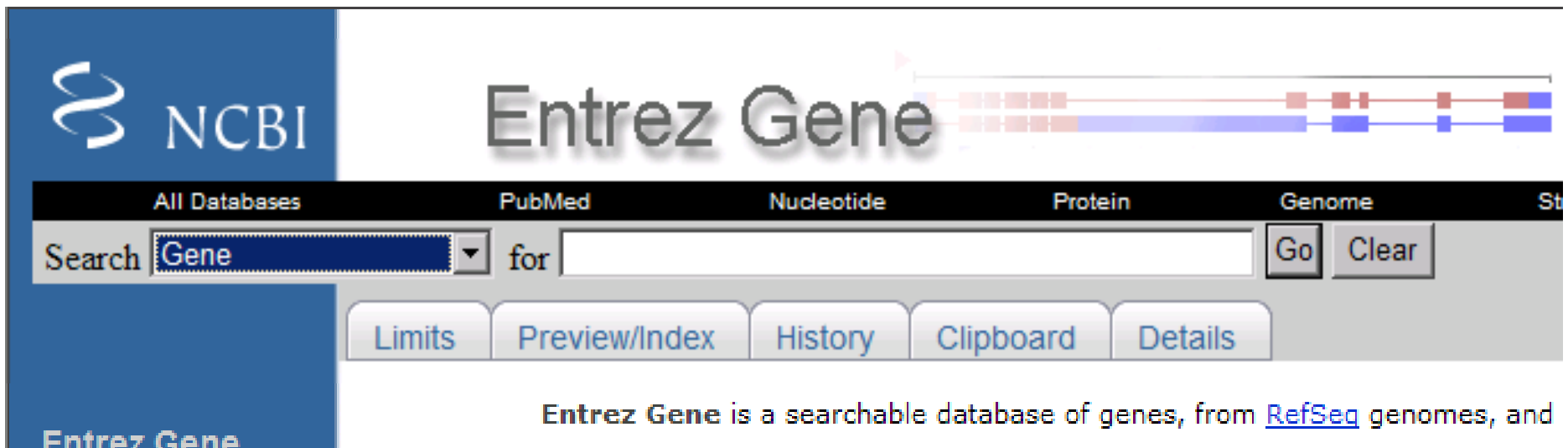
- Преко GenBank-а прикажите унос који одговара приступном броју AF018430. Објасните реконструкцију гена и mRNA.

Ограничења GenBank-а

- GenBank може да садржи
 - Читаве гене
 - Делове гена
 - Много гена
- GenBank унос може бити неједнаког квалитета
 - Могу бити дупликати или нетачни
 - Не постоји механизам за селекцију података
 - Сви подтаци се подједнако третирају
- GenBank унос није последња реч о датом гену
 - Немају ауторитативно биолошко значење
 - Једноставно садрже информацију о ономе што је већ урађено
- Геноцентричне базе податаке су потребне да би се прикупило све што се зна о датом гену и да би се исправиле могуће грешке

Коришћење геноцентричних база: Entrez Gene

- NCBI вам омогућава приступ Entrez Gene-у
- У GenBank-у, сваки унос је једна **секвенца по публикацији**
- In Entrez Gene-у, сваки унос је један **ген**
- Entrez Gene је направљен од GenBank података



The screenshot shows the NCBI Entrez Gene search interface. On the left is the NCBI logo. The main header reads "Entrez Gene" with a diagram of a gene structure. Below the header is a navigation bar with tabs for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", and "St". The search bar contains the text "Gene" and "for" followed by an empty input field. To the right of the search bar are "Go" and "Clear" buttons. Below the search bar are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". At the bottom, a text box states: "Entrez Gene is a searchable database of genes, from [RefSeq](#) genomes, and

Задаци за вежбу

- У Genbank-у лоцирајте ген nirR у E coli. Сачувајте секвенцу овог гена у FASTA формату као текст фајл у Word-у.
- Преко GenBank-а лоцирајте унос који одговара “accession number” AF018432 Запамтите овај унос као html фајл.
- Преко геноцентричне базе (Entrez Gene) лоцирајте CRP ген код E coli.
- Поновите исто за GL3, дакле преко Entrez Gene лоцирајте ген за GL3 Arabidopsis Thaliana.