Циљеви часа

- Разумевање значаја претраге по сличности
- Разумевање разлика између хомологије, сличности и идентитета
- Коришћење BLAST-а и интерпретација излазних резултата
- Разумевање концепта E-вредности
- Како да одговарате на различита питања помоћу BLAST-а

План часа

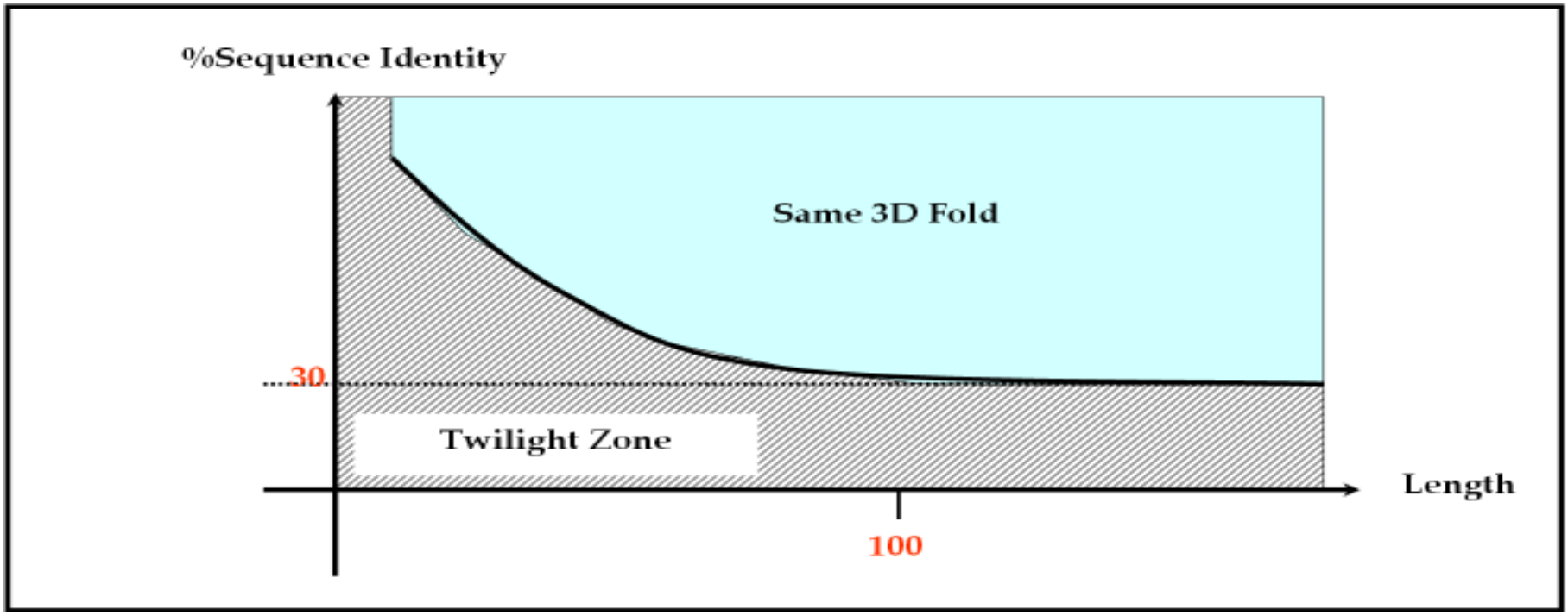
- Биолошки значај сличности између секвенци
- Хомологија, идентитет и сличност
- Коришћење BLAST-а
- Интерпретирање BLAST излаза
- Биолошка анализа помоћу BLAST-а
- Коришћење PSI-BLAST-а (најновије верзије BLAST-а)

Хомологија

- Хомологне секвенце су оне које имају истог претка
  - Истог претка (и протеини и ДНК)
- Хомологне секвенце (протеини) имају
  - Сличну 3D структуру
  - Често сличну функцију
- Две протеинске секвенце са више од 25 % идентитета (ако су дуже од 100 аминокиселина) су вероватно хомологне
- Две ДНК секвенце са више од 70 % идентитета (ако су дуже од 100 нуклеотида) су вероватно хомологне

Хомологија (наставак)

- Када два протеина имају мање од 25% идентитета, немогуће је рећи да ли су протеини хомолгни или нису.
  - Овај регион се зове “Зона сумрака”



Хомологија, сличност и идентитет

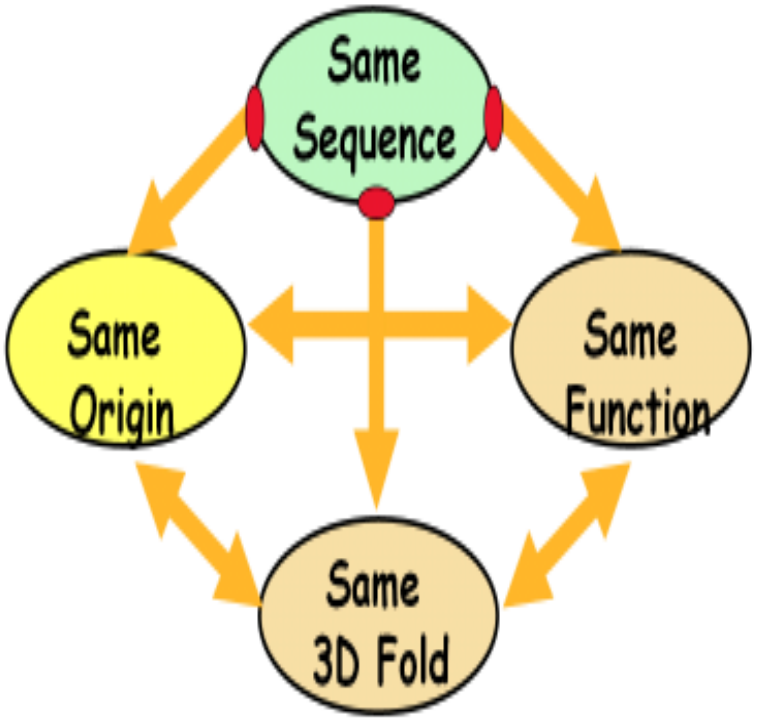
- Идентитет је *мера* која је директно везана са поравнањем
  - Нпр. секвенца А може бити “32 % идентична са” секвенцом В
- Идентитет може да се изражава у процентима (нпр 30%)
- Хомологија је бинарна особина, т.ј. она или постоји или не постоји
  - Секвенца А или **ЈЕСТЕ** или **НИЈЕ** хомологна са секвенцом В
  - Секвенца А **не може бити** “40% хомологна” са секвенцом В
- Хомологија се успоставља на основу измереног идентитета између две секвенце

Како се установљава хомологија

- Упоредите протеин А са сваким другим протеином у бази података као што је Swiss-Prot
- Рецимо да сте идентификовали протеин В који је 40% идентичан са вашим протеином
  - Још боље од % идентитета је да користите Е-вредности (више о томе доле)
- Пошто су А и В веома слични, можете да закључите да су највероватније хомологи
  - Логика је слична као код исказа, “Милица и Ненад су највероватније брат и сестра пошто су веома слични.”
- Ако знате структуру и функцију од В, онда и А највероватније има сличну структуру и функцију

Виртуелни експеримент

- Ако успоставите да су два протеина хомологни (A and B), можете да екстраполирате оно што знате о једном протеину на други протеин.
- То је као да правите виртуелни експеримент.



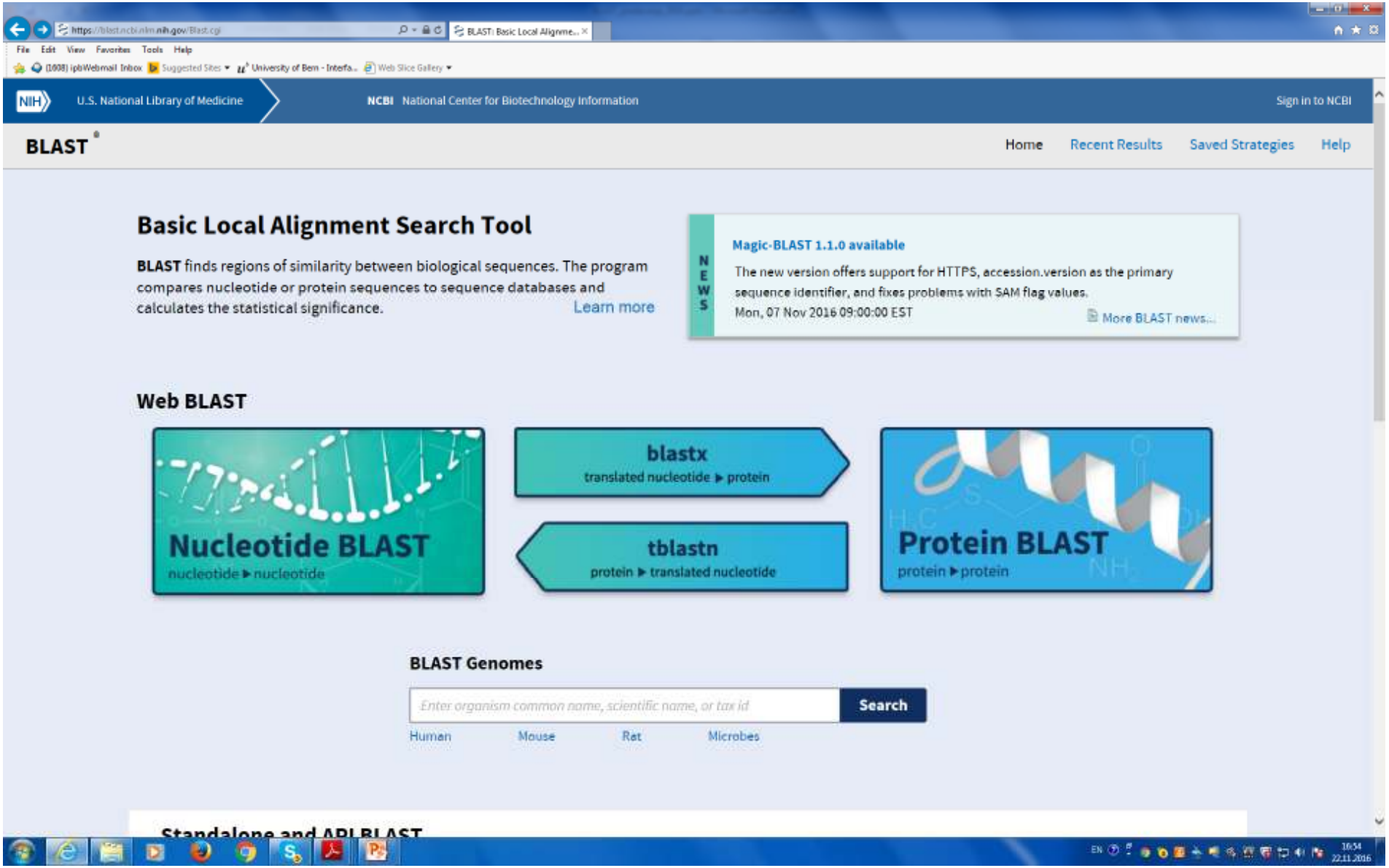
BLAST

- BLAST: Basic Local Alignment Search Tool
- BLAST је програм који вам омогућава да упоредите вашу секвенцу са свим осталим секвенцама у бази података
- BLAST може да упоређује
  - ДНК секвенце
  - Протеинске секвенце
- BLAST је тачнији у упоређивању протеинских секвенци него ДНК секвенци

BLAST (наставак)

- BLAST прави локално поравнање (local alignment)
  - Поравнава само оно што може бити поравнато
  - Игнорише све остало
- BLAST је веома брз
  - Треба вам око минута да би претражили Swiss-Prot на стандардном PC-ју
- Више врста BLAST који решавају различите задатке су на располагању

Различите врсте BLAST-а . . .

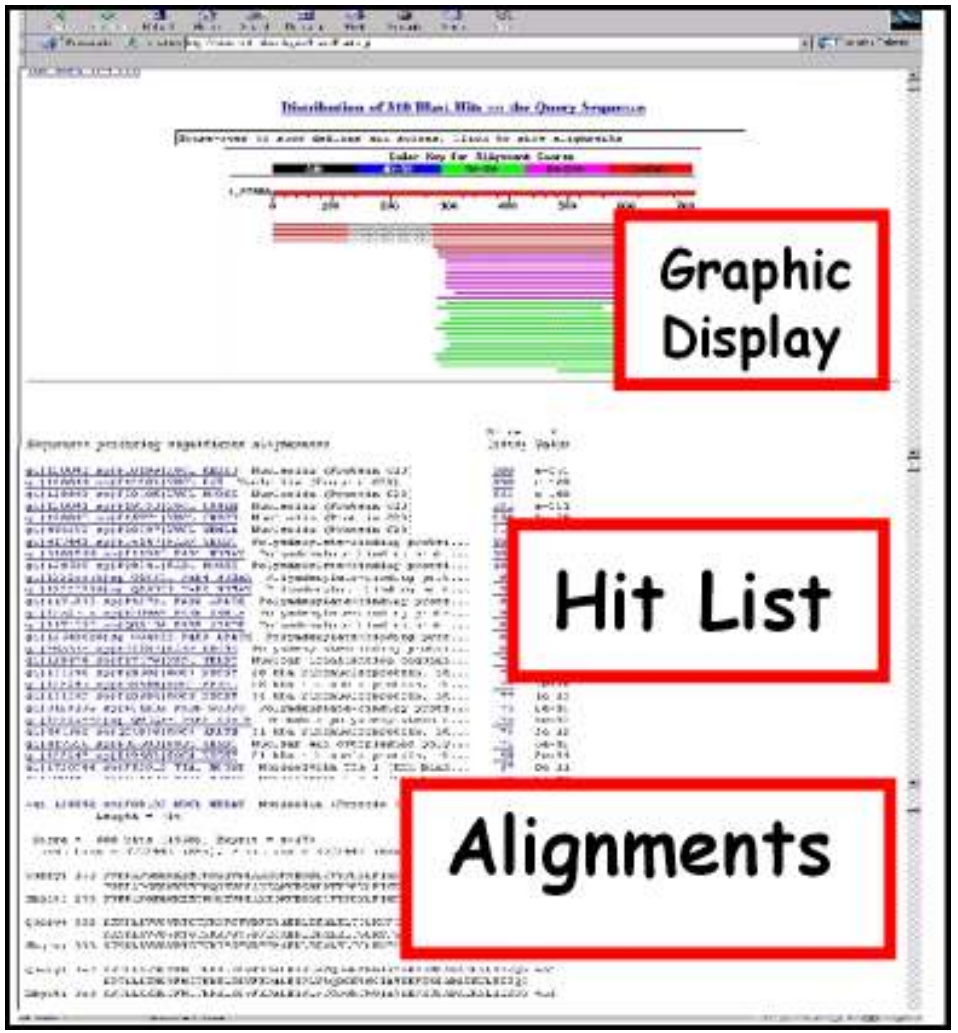


Коришћење blastp (Protein BLAST)

- Изаберите неки од сервера са BLAST-ом
  - NCBI www.ncbi.nlm.nih.gov/blast
  - EBI www.ebi.ac.uk/blast
  - EMBNet www.expasy.ch/blast
- Изаберите базу података коју желите да претражујете:
  - NR да би нашли било коју протеинску секвенцу
  - Swiss-Prot да би нашли протеине са познатим функцијама
  - PDB да би нашли протеине са познатим функцијама
- Исеците и залепите вашу секвенцу
- Стисните **BLAST** дугме

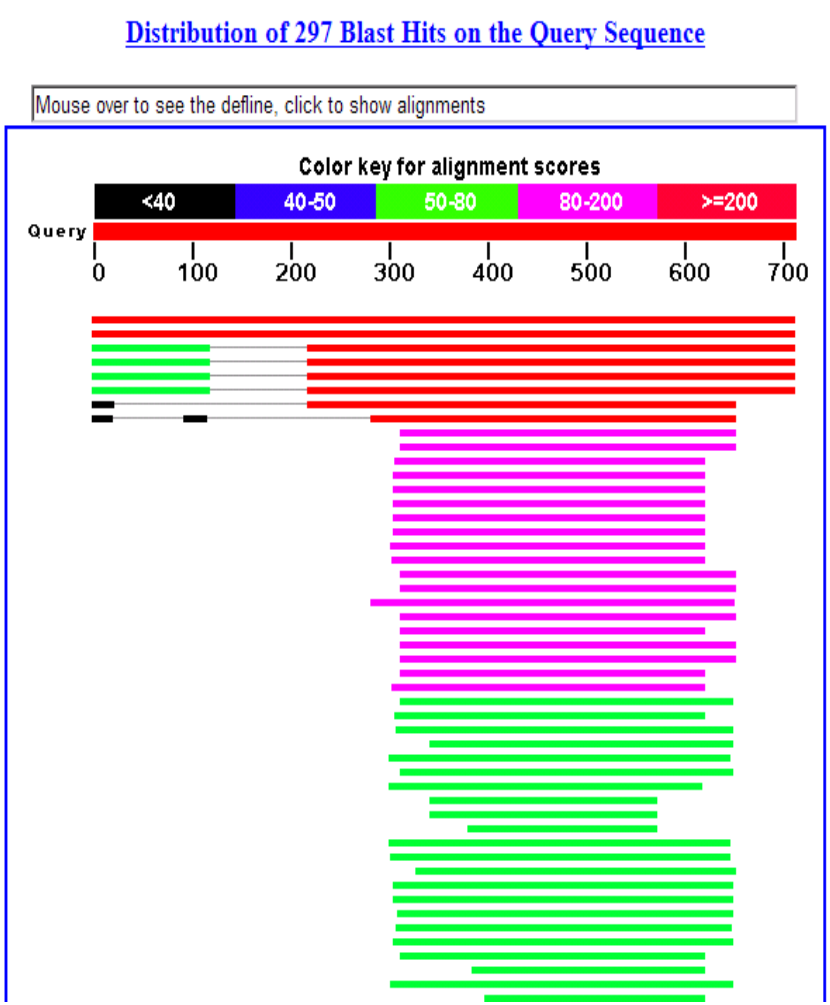
Читање BLAST излаза

- Графички Излаз
  - Приказ поравнања
- Листа са поготцима
  - Укључујући статистички значај сваког поготка
- Поравнања (Alignments)
  - Детаљи о сваком поравнању



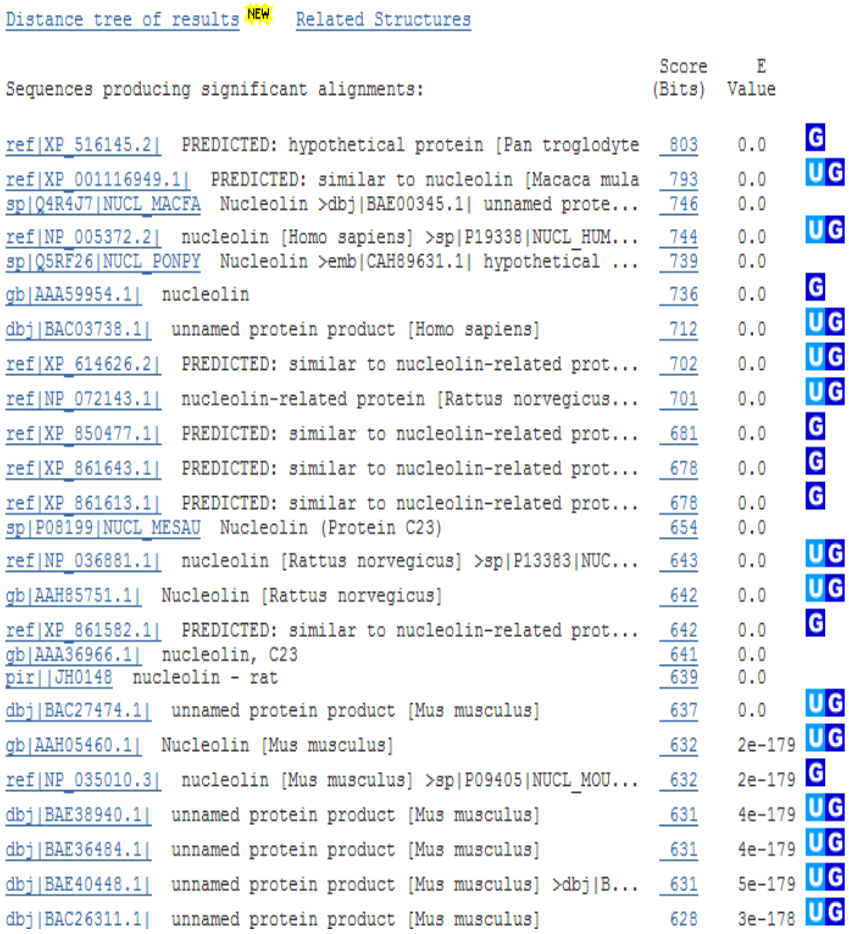
Графички приказ

- Хоризонтална оса (0-700) одговара вашем протеину (query)
- Боја указује на квалитет преклапања
  - Црвено: врло добар
  - Зелено: прихватљив
  - Црно: лош
- Танке линије повезују независне поготке на истој секвенци



Листа погодака

- Sequence accession number
  - Идентификује секвенцу
- Description (опис)
  - Taken from the database
- Bit score (поени)
  - **Високи** bit score = **добар** погодак
- E-вредност
  - **Ниска** E-вредност = **добар** погодак
- Links
  - Геном (G)
  - Uniref (U), база транскрипта



Distance tree of results NEW Related Structures

Sequences producing significant alignments:	Score (Bits)	E Value	
ref|XP_516145.2| PREDICTED: hypothetical protein [Pan troglodyte	803	0.0	G
ref|XP_001116949.1| PREDICTED: similar to nucleolin [Macaca mula	793	0.0	UG
sp|Q4R4J7|NUCL_MACFA Nucleolin >dbj|BAE00345.1| unnamed prote...	746	0.0	
ref|NP_005372.2| nucleolin [Homo sapiens] >sp|P19338|NUCL_HUM...	744	0.0	UG
sp|Q5RF26|NUCL_PONPY Nucleolin >emb|CAH89631.1| hypothetical ...	739	0.0	
gb|AAA59954.1| nucleolin	736	0.0	G
dbj|BAC03738.1| unnamed protein product [Homo sapiens]	712	0.0	UG
ref|XP_614626.2| PREDICTED: similar to nucleolin-related prot...	702	0.0	UG
ref|NP_072143.1| nucleolin-related protein [Rattus norvegicus...	701	0.0	UG
ref|XP_850477.1| PREDICTED: similar to nucleolin-related prot...	681	0.0	G
ref|XP_861643.1| PREDICTED: similar to nucleolin-related prot...	678	0.0	G
ref|XP_861613.1| PREDICTED: similar to nucleolin-related prot...	678	0.0	G
sp|P08199|NUCL_MESAU Nucleolin (Protein C23)	654	0.0	
ref|NP_036881.1| nucleolin [Rattus norvegicus] >sp|P13383|NUC...	643	0.0	UG
gb|AAH85751.1| Nucleolin [Rattus norvegicus]	642	0.0	UG
ref|XP_861582.1| PREDICTED: similar to nucleolin-related prot...	642	0.0	G
gb|AAA36966.1| nucleolin, C23	641	0.0	
pir||JH0148 nucleolin - rat	639	0.0	
dbj|BAC27474.1| unnamed protein product [Mus musculus]	637	0.0	UG
gb|AAH05460.1| Nucleolin [Mus musculus]	632	2e-179	UG
ref|NP_035010.3| nucleolin [Mus musculus] >sp|P09405|NUCL_MOU...	632	2e-179	G
dbj|BAE38940.1| unnamed protein product [Mus musculus]	631	4e-179	UG
dbj|BAE36484.1| unnamed protein product [Mus musculus]	631	4e-179	UG
dbj|BAE40448.1| unnamed protein product [Mus musculus] >dbj|B...	631	5e-179	UG
dbj|BAC26311.1| unnamed protein product [Mus musculus]	628	3e-178	UG

E-вредност

- E-вредност значи очекивана вредност (*expectation value)*
- E-вредност је најчешће коришћена мера за утврђивање сличности између секвенци
- E-вредност даје колико пута се очекује да ће дати или бољи погодак да се јави случајно
- Ако је погодак високо неочекиван, вероватно потиче од нечега што није случајност
  - Заједничко порекло је највероватније објашњење
  - Због тога се из ниске E-вредности изводи закључак о хомологији

Колика E-вредност је довољно ниска ?

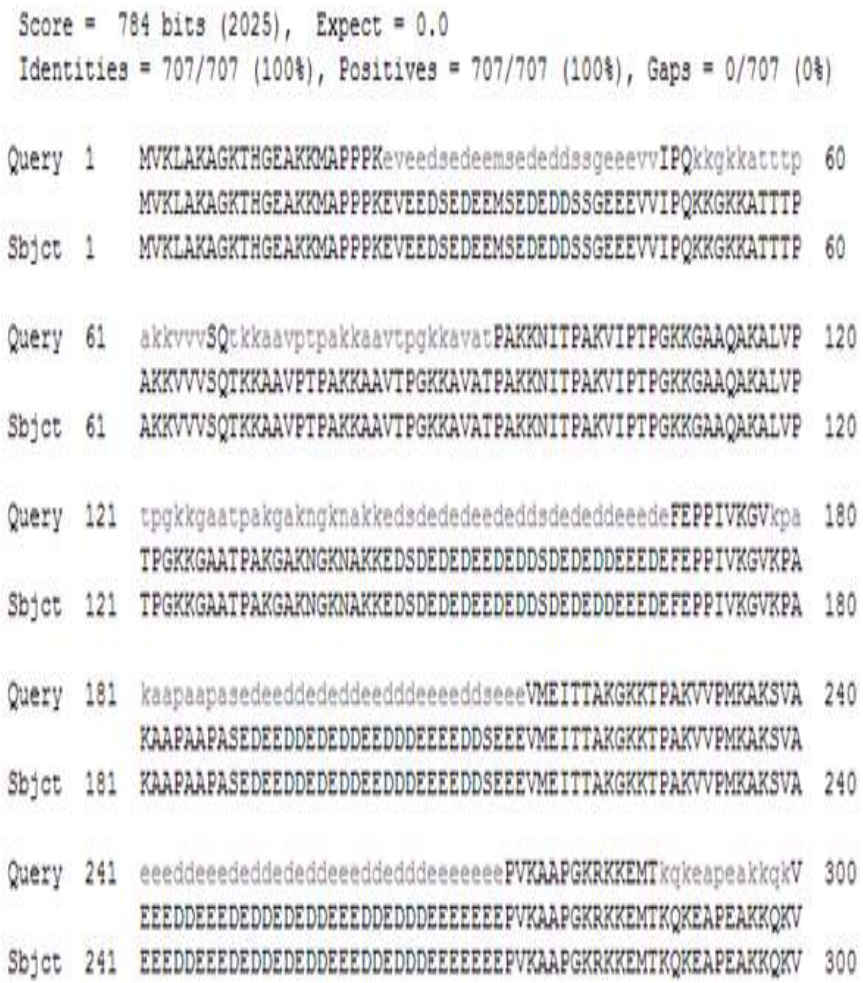
- Ниска E-вредност ⇔ добар погодак
  - 1 = лоша E-вредност
  - 10^{e-3} = гранична E-вредност
  - 10^{e-4} = добра E-вредност
  - 10^{e-10} = врло добра E-вредност
- E-вредности ниже од 10^{e-4} указују на могућу хомологију
- E-вредности више од 10^{e-4} захтевају додатне доказе да би се успоставила хомологија

Зашто користите E-вредности?

- E-вредности омогућавају да се упоређују поравнања различитих дужина
- E-вредности се користе од стране већине програма за упоређивање секвенци нпр.
  - PSI-BLAST
  - Претрага за доменима (CD Search)
- E-вредности увек имају исто значење
  - Можете да упоредите излаз различитих програма

Поравнања

- Поравнање показује који делови секвенце су идентификовани као слични.
- Секвенца коју смо унели за претрагу (Query) се поравнава са секвенцама које су BLAST поготци (Subject).
- Обратите пажњу на кластере који указују на идентитете између секвенци.



Пример

a) Нађите протеинске секвенце које су сличне са Xp10 RNA полимеразом користећи BLAST. Користите Swiss Prot да би нашли протеинску секвенцу за Xp10 RNA полимеразу.

Решење

- Идите на http://blast.ncbi.nlm.nih.gov/Blast.cgi
- Кликните на линк за “protein BLAST”
- Отворите текстуални фајл са протеинском секвенцом Xp10 RNA полимеразом
- У поље “Enter Query Sequence” унесите секвенцу из запамћеног текстуалног фајла
- Кликните на “Run” дугме
- Појавиће се прозор који означава да се податци обрађују. Чекајте и не дирајте ништа
- Када добијете резултат коментаришите, који протеини су слични са вашим молекулом

BLAST-овање ДНК секвенци

- Врста BLAST програма који је коришћен зависи од ДНК секвенце:
  - Кодирајућа ДНК
  - Некодирајућа ДНК
- BLAST-овање ДНК секвенци је мање прецизно него BLAST-овање протеинских секвенци.
- Ако је ваша секвенца кодирајућа, blastx и tblastx ће их транслирати у 6 могућих оквира читања.

Пример blastn

- Искористите blastn да би лоцирали ДНК секвенце у бази које су сличне са интергенским регионом бактериофага Xp10 са координатама 42097-43182.

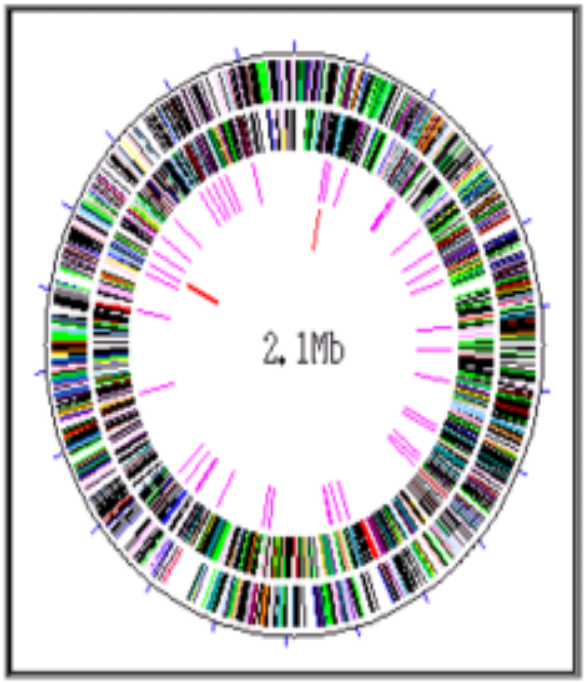
tBLASTn

- Искористите tBLASTn да идентификујете гене који кодирају протеине који су хомологи са sigma фактором који кодира бактерофаг 7-11.

Још неки подаци о BLAST-у

- Оригинални BLAST рад је један од најцитиранијих научних радова уопште
  - ~60,000 цитата (Google Schoolar)
- BLAST је променио многе аспекте модерне биологије
- Слајдови који следе дају неке примене BLAST процедуре

Тражење гена са BLAST-ом

What you need	*The BLAST way*
Предвиђање гена у секвенци	**Исеците ваш геном у мале делове (2~5Kb) секвенци које се међусобно преклапају. Користите blastx да би BLAST-овали вашу секвенцу у односу на NR (the Non Redundant protein database).**
	**Алтернатива је да користите програме за предикцију гена.**

Пример – тражење гена са BLAST-ом

- Користећи blastx одредите гене који се налазе унутар сегмента 7000-12000 геномске секвенце бактериофага “Phage 7-11”.

Пример - одређивање функције протеина

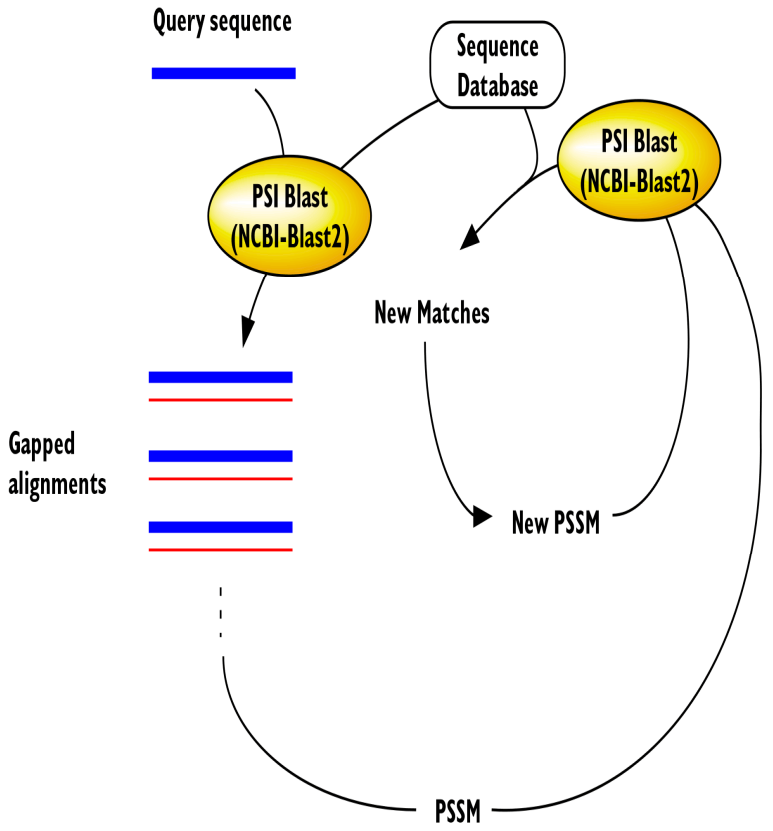
- Овде користите обичан протеин BLAST (pBLAST) као у првом примеру који смо урадили.

PSI-BLAST

- PSI-BLAST је *P*osition-*S*pecific *I*terated BLAST
  - Осетљивији него BLAST: налази поготке које BLAST не може да нађе
  - Специфичнији него BLAST: пријављује мање лажних погодака
  - Нешто спорији него BLAST
- PSI-BLAST је у стању да нађе удаљене хомологе
  - Ово омогућава да се идентификују врло удаљени чланови дате фамилије протеина
- PSI-BLAST користи резултате сваке итерације да би повећао специфицитет претраге

PSI-BLAST Итерације

- PSI-BLAST користи најбоље резултате прве итерације да би направио матрицу која даје вероватноћу појаве дате базе на датој позицији у секвенци (PSSM – Position Specific Scoring Matrix).
- PSI-BLAST користи ту матрицу да би поново претражио базу података и нашао нове поготке.
- PSI-BLAST наставља да понавља ову процедуру све до конвергенције, односно до тренутка када престају да се идентификују нови поготци.



Додатни коментари о PSI-BLAST-у

- Ако се ваш протеин састоји из више домена, претражите домене један по један
- PSI-BLAST је спорији него стандардни BLAST пошто ради итерације
- Можете PSI-BLAST-у вашу сопствену PSSM
  - За то можете да користите NCBI сервер

Пример – PSI-BLAST

PSI-BLAST – Приступате јој са исте стране као и обичном BLAST-у, само што у оквиру “algorithm” кликнете на опцију “PSI-BLAST”. Пробајте сами у оквиру задатака за вежбу, водите рачуна да до конвергенције погодака морате да прођете кроз више BLAST итерација.

Задаци за вежбу

1) Користећи BLAST, нађите протеинске секвенце које су сличне са протеином p7 код бактериофага Xp10. Секвенцу протеина нађите преко Swiss Prot-a. Наведите који од добијених погодака су статистички значајани.

2) Пођите од геномске секвенце бактериофага Xp10 и искористите одговарајућу верзију BLAST-a да нађите које ДНК секвенце су сличне интергенском региону са координатама 42253-42740.

2) Пођите од секвенце протеина у protein_seq_vezba.txt и користећи одговарајућу верзију BLAST-a идентификујте гене у геномским секвенцама које кодирају хомологне протеине.

3) Користећи одговарајућу верзију BLAST-a одредите гене који се налазе унутар сегмента 3000-8000 геномске секвенце бактериофага “Phage 7-11”.

4) Искористите PSI-BLAST да би утврдили вероватну функцију протеина који потиче из новосеквенцованог бактериофага “Phage 7-11”. Секвенца протеина је дата у фајлу означеном са protein_seq_vezba.txt У претрази користите следеће параметре:

- Извршите претрагу у најобимнијој расположивој бази података “Non-redundant protein sequences”
- У оквиру опције “Organism” ограничите претрагу на “phages with short tails”.