Упоређивање две ДНК секвенце

Циљеви часа

- Коришћење тачкастих графика (dot plot-а)
- Интерпретација најчешћих облика dot plot-а
- Коришћење Dotlet-а
- Коришћење Lalign-а за локална поравнања две секвенце

План часа

- Разлози за упоређивање две секвенце
- Основни принципи коришћења dot-plot-a
- Коришћење Dotlet-a
- Коришћење Lalign-a

Зашто упоређујемо две секвенце?

- Претраге база података (BLAST) су корисне за налажење хомолога
- Међутим ове претраге не омогућавају детаљно упоређивање две секвенце
- Прецизнији методи су потребни да би се анализирале секвенце у детаљима
  - Dot-plot за графичку анализу
  - Локална или глобална поравнања за анализу резидуал по резидуал
- Поравнање две секвенце се зове поравнање у пару (*pairwise* alignment)

Неке примене поравнања у пару

- Убедите себе да су две секвенце хомологи
- Идентификујте заједничке домене
- Идентификујте дуплиране регионе
- Лоцирајте битне делове секвенце као сто су:
  - Каталитички домени
  - Дисулфидни мостови
- Упоредите ген и његов продукт

Шта је Dot Plot ?

- Dot plot је графичка репрезентација сличности две секвенце
- Базиран на једноставном концепту
- Нацртате једну секвенцу наспрам друге на x и y оси
- Међутим, релативно лако можете да откријете чак и комплексне зависности између две секвенце.
- Базиран је на најсофистицираније машини за статистичку анализу . . . људском мозгу

Како да изаберете две секвенце

- Нерационално је да упоређујете по паровима велики број секвенце
- Зато користите BLAST да ефикасно изаберете ваше секвенце
  - Више од 70% идентитета за ДНК
  - Више од 25% идентитета за протеине
- Водите рачуна да ако су ваше секвенце превише сличне, њихово упоређивање не води ка корисној информацији

Упоређивање секвенце са самом собом

- Почните са упоређивањем секвенце са самом собом
- Можете да откријете
  - Домене који се понављају
  - Мотиве који су поновљени више пута (low complexity)
  - Региони који одговарају слици у огледалу (палиндроми)

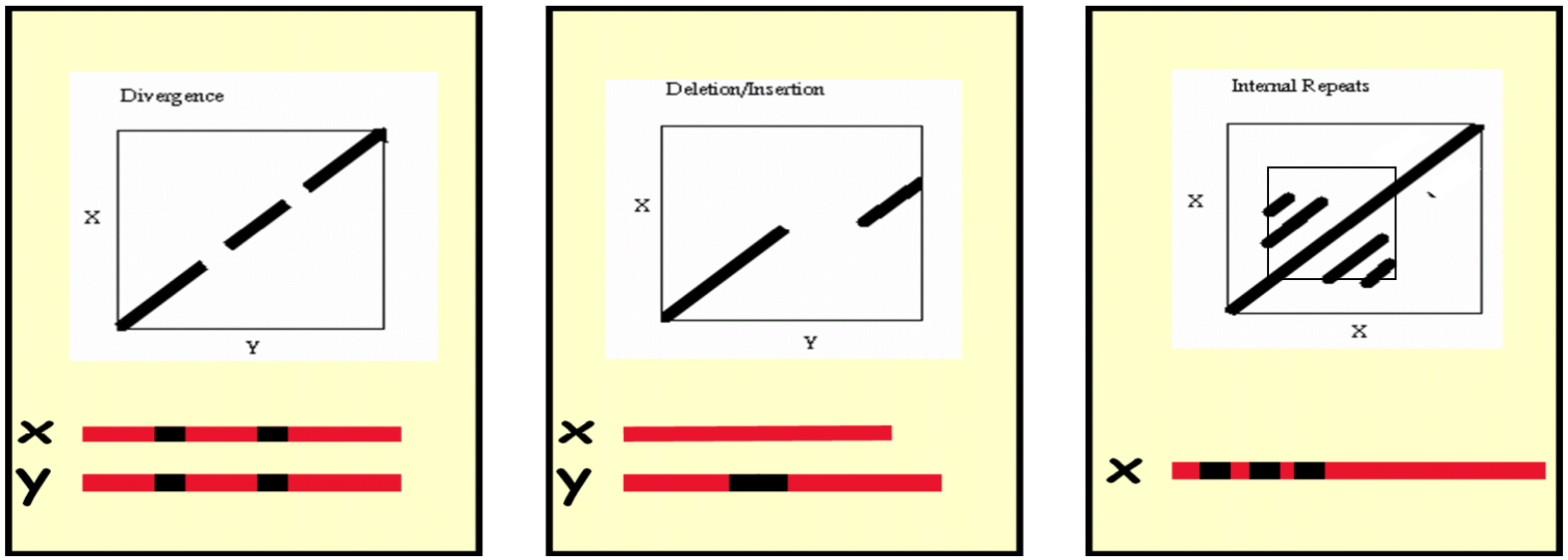
Шта можете да анализирате са Dot Plot-ом ?

- Било који пар секвенци
  - ДНК
  - Протеине
  - РНК
- Секвенце дуже од 1000 симбола се тешко анализирају преко сервера

  - За то треба локално (на вашем компјутеру) да инсталирате програм

Нека типична Dot-plot упоређивања

- Дивергентне секвенце где је само један сегмент хомолог
- Дугачка убацивања (insertions) или брисања (deletions)
- Тандемска понављања
  - За њих је карактеристичан квадратни облик као на слици

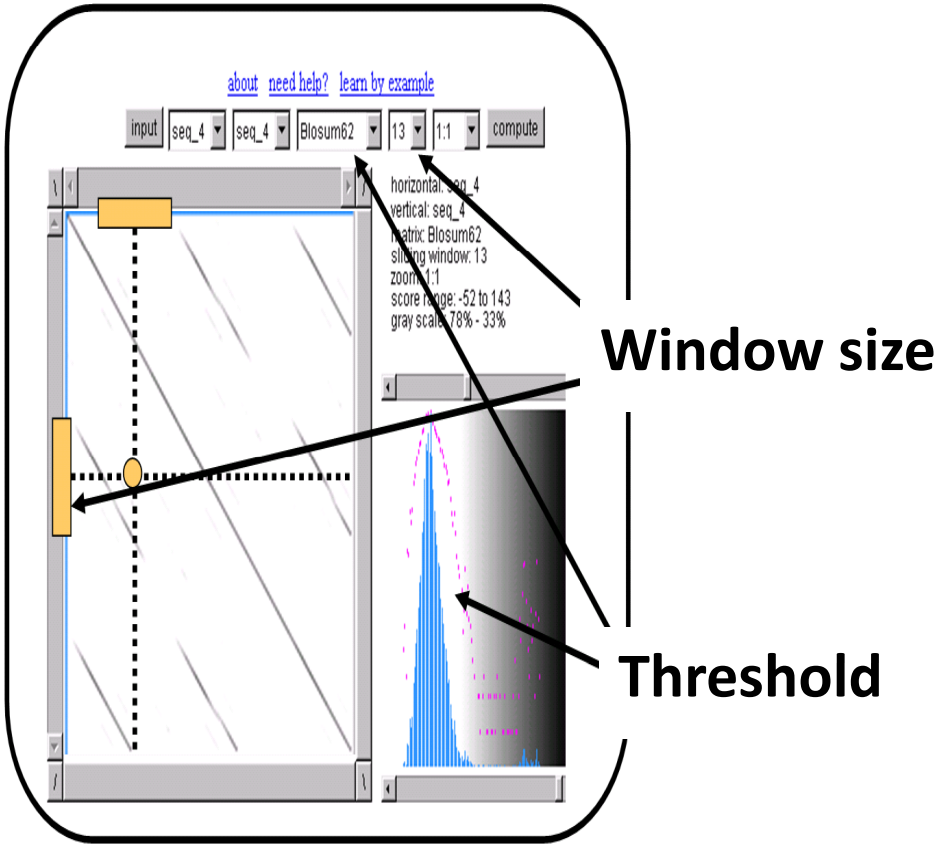


Коришћење Dotlet-a

- Dotlet је један од најпогоднијих метода за коришћење dot-plot-a
- Dotlet је Java aplikacija
- Отворите и скините апликацију са следећег сајта:
  - www.isrec.isb-sib.ch/java/dotlet
- Користите Firefox или IE (ако један не ради користите други)

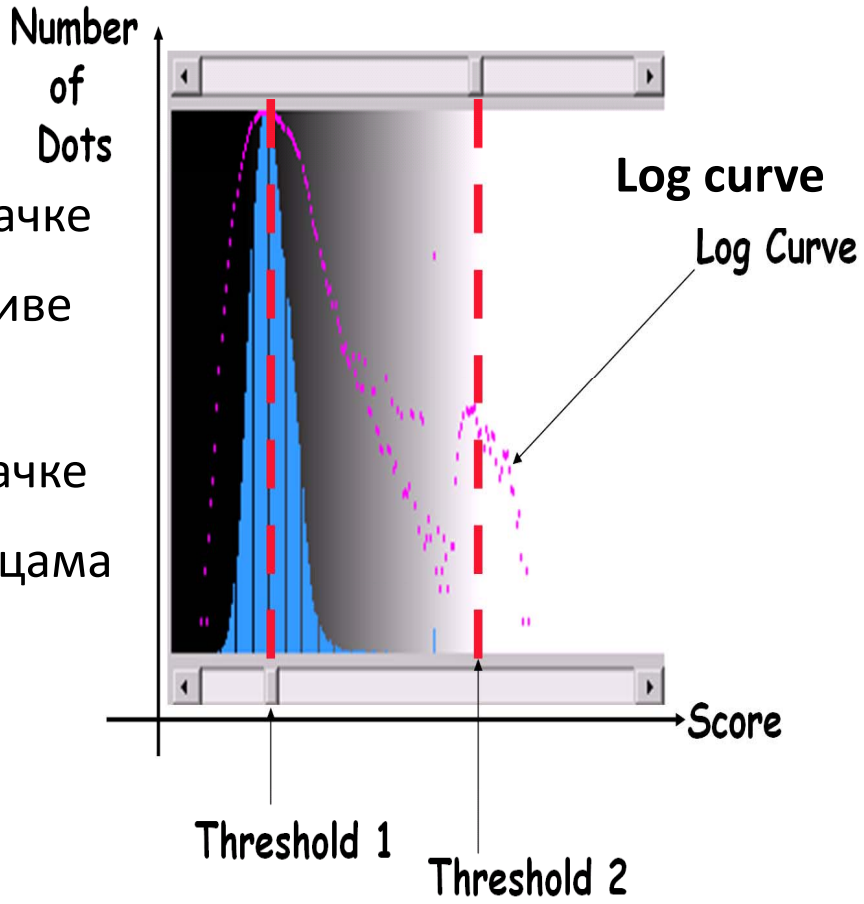
Подешавање Dotlet параметара

- Dotlet помера прозор дуж сваке секвенце
- Ако су прозори сличнији од одређене граничне вредности (threshold), Dotlet штампа тачку на њиховом пресеку
- Програм омогућава да подесите вредност ове граничне вредности



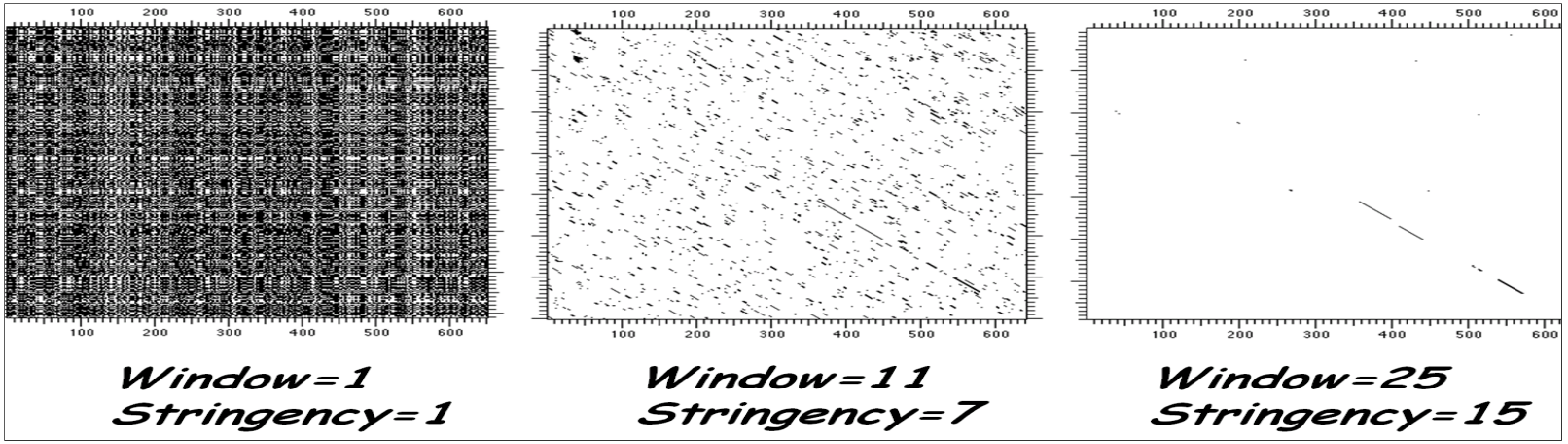
Гранична вредност у Dotlet-у

- Свакој тачки одговара одређени број поена добијен упоређивањем два прозора
- Када је број поена
  - Испод граничне вредности 1 ⇔ црне тачке
  - Између граничних вредности1 и 2 ⇔ сиве тачке
  - Изнад граничне вредности 2 ⇔ беле тачке
- Плава крива је расподела поена по секвенцама
- Максимум ⇔ најчешћи број поена,
  - најчешћи ⇔ најмање информативан



Правилно подешавање Dot Plot-a

- Величина прозора и гранична вредност одређују аспекте вашег dot plot-a
  - “Строги” параметри = Чист dot plot, мало сигнала
  - “Либерални” параметри = Велик шум, превише сигнала
- Играјте се са граничном вредношћу, све док не добијете одговарајућу граничну вредност

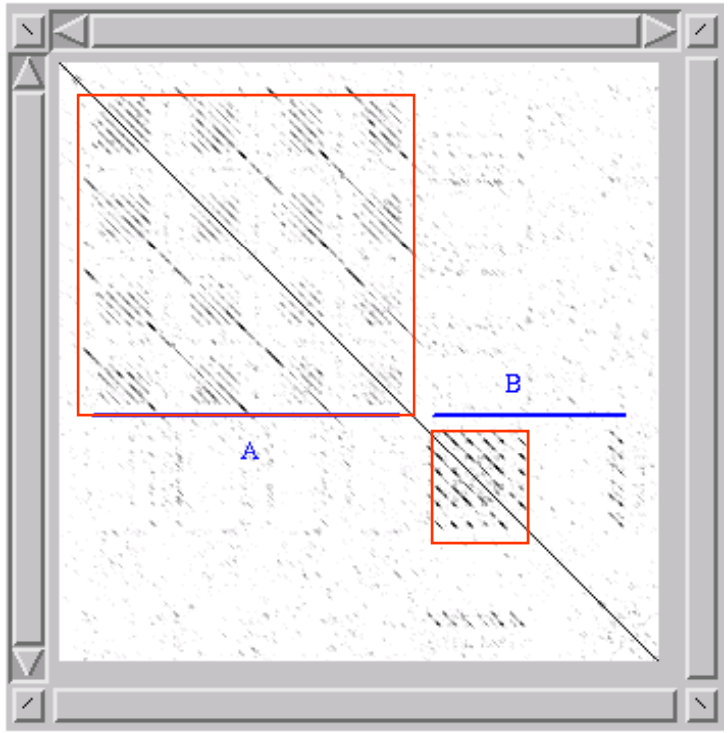


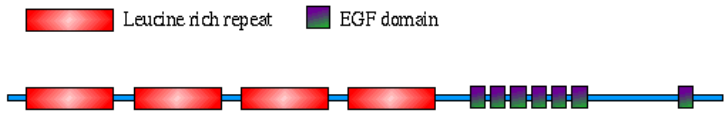
Величина прозора?

- Дугачак прозор
  - Чист dot plot
  - Мала осетљивост
- Кратак прозор
  - Доста шума у dot plot-у
  - Веома осетљиви
- Дужина прозора треба отприлике да одговара дужини елемента који тражите
  - Конзервирани домени: 50 аминокиселина
  - Трансмембрански сегменти: 20 аминокиселина
- Скратите прозор да би упоредили удаљене секвенце

Упоређивање поновљених домена са Dotlet-om

- Квадратни облик је типичан за тандемска понављања
- Понављања нису савршена пошто секвенце дивергирају после дупликације





Пример-очувани домени

- Дата су два протеина са Swiss Prot accession бројем P05049 и P08246. Поравнање та два протеина помоћу BLAST-а води ка маргинално значајној E вредности (10^-4). Искористите Dotplot да би установили да ли постоји конзервирани домен између ове две секвенце.

Пример - идентификовање тандемских понављања

- Протеин са Swiss Prot приступним бројем Q9P255 одговара могућем транскрипционом фактору код човека. Испитајте да ли овај протеин има тандемска понављања, и ако да, нађите њихов број.

Поравнање секвенци

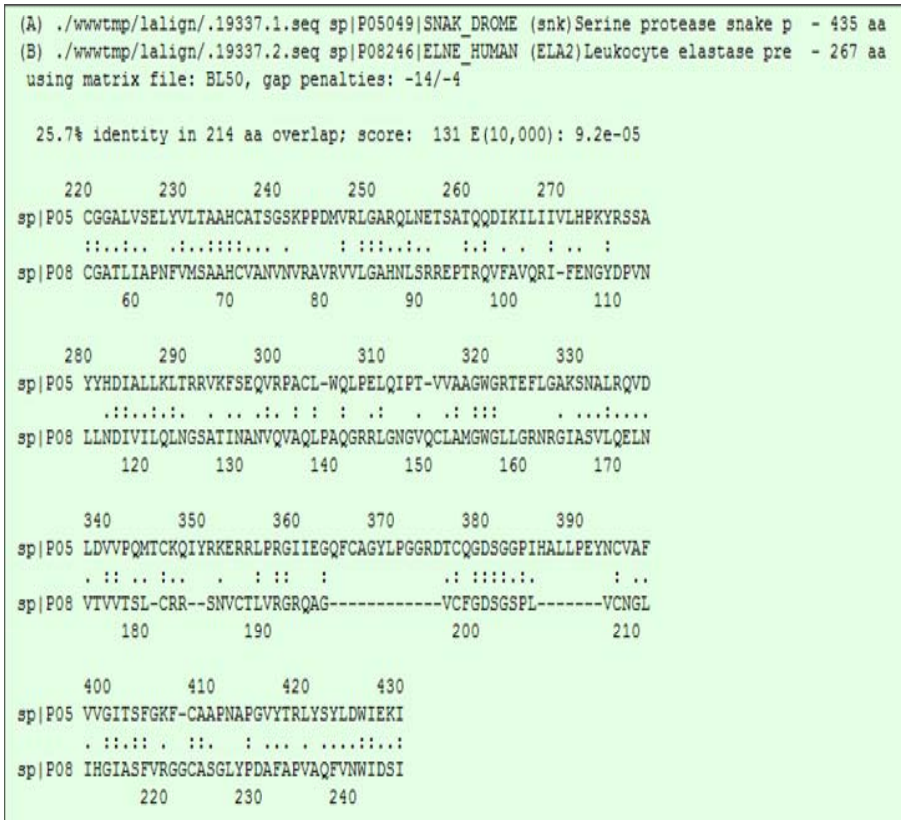
- Dotlet dot plot-ови су веома добар начин да се добије глобална слика
- Dot plots не омогућавају анализу резидуал по резидуал
- За ово вам треба експлицитно поравнање
- Најпогоднији алат за прављење прецизног локалног поравнања је Lalign

Lalign

- Lalign је веома прецизан BLAST
- Упоређује само дате две секвенце
- Спорији је од BLAST-а
- Користи се за краће секвенце
- За дате две секвенце вам даје десет најбољих поравнања – за разлику од BLAST-а који даје само једно најбоље поравнање
- Ради боље са протеинима него са ДНК

Излаз Lalign-a

- Lalign даје излаз који је сличан секцији са поравнањем BLAST-а
- E-value указује на статистички значај
- Ниска E-вредност ⇔ добро поравнање



```
(A) ./wwwtmp/lalign/.19337.1.seq sp|P05049|SNAK_DROME (snk)Serine protease snake p  - 435 aa
(B) ./wwwtmp/lalign/.19337.2.seq sp|P08246|ELNE_HUMAN (ELA2)Leukocyte elastase pre  - 267 aa
 using matrix file: BL50, gap penalties: -14/-4

  25.7% identity in 214 aa overlap; score:  131 E(10,000): 9.2e-05

     220       230       240       250       260       270
sp|P05 CGGALVSELYVLTAAHCATSGSKPPDMVRLGARQLNETSATQQDIKILIIVLHPKYRSSA
       ::..:..  .:..::::... .    : :::..:..  :.: . .  : ..  :
sp|P08 CGATLIAPNFVMSAAHCVANVNVRAVRVVLGAHNLSRREPTRQVFAVQRI-FENGYDPVN
           60        70        80        90       100       110

     280       290       300       310       320       330
sp|P05 YYHDIALLKLTRRVKFSEQVRPACL-WQLPELQIPT-VVAAGWGRTEFLGAKSNALRQVD
       .::..:.:.  . .. .:. : :  :  .:   .  .: :::      . ...:....
sp|P08 LLNDIVILQLNGSATINANVQVAQLPAQGRRLGNGVQCLAMGWGLLGRNRGIASVLQELN
           120       130       140       150       160       170

       340       350       360       370       380       390
sp|P05 LDVVPQMTCKQIYRKERRLPRGIIEGQFCAGYLPGGRDTCQGDSGGPIHALLPEYNCVAF
       . :: .. :..   .   : ::   :           .: ::::.:.       : ..
sp|P08 VTVVTSL-CRR--SNVCTLVRGRQAG-----------VCFGDSGSPL-------VCNGL
           180         190                 200             210

       400       410       420       430
sp|P05 VVGITSFGKF-CAAPNAPGVYTRLYSYLDWIEKI
       . ::.:: .  ::.   : ... . ....::..:
sp|P08 IHGIASFVRGGCASGLYPDAFAPVAQFVNWIDSI
            220       230       240
```

Пример - Lalign

- Искористите Lalign да би поравнали секвенце протеина са приступним бројевима P05049 и P08246.

Вежбе

1) Искористите Dotplot да би нашли конзервиране домене између протеинских секвенци које одговарају sigma70 (“sigma_70_Ecoli_sekvenca.txt”) и sigmaS (“stress_induced_sigma_factor.txt”)
2) Помоћу Dotplot-а установите да ли постоји тандемско понављање код протеина са Swiss Prot приступним бројем P03001 (фајл “TF_tandem_vezba.txt”). Напомена: водите рачуна да дуплицирани делови могу у знатној мери да дуплирају.
3) Искористите Lalign да нађете поравнање између две секвенце у првом задатку.