Поравнање више секвенци

Циљеви часа

- Разумевање ситуација у којима је поравнање више секвенци корисно
- Како да се исправно направи поравнање више секвенци
- Разумевање предности и мана прогресивног алгоритма за поравнање више секвенци

План часа

- Зашто се прави поравнање више секвенци?
- Како се бирају одговарајуће секвенце
- Како ради прогресивни алгоритам за поравнање
- Интерпретација поравнања више секвенци

Шта је поравнање више секвенци ?

- Поравнање више секвенци
- MSAs = multiple-sequence alignments

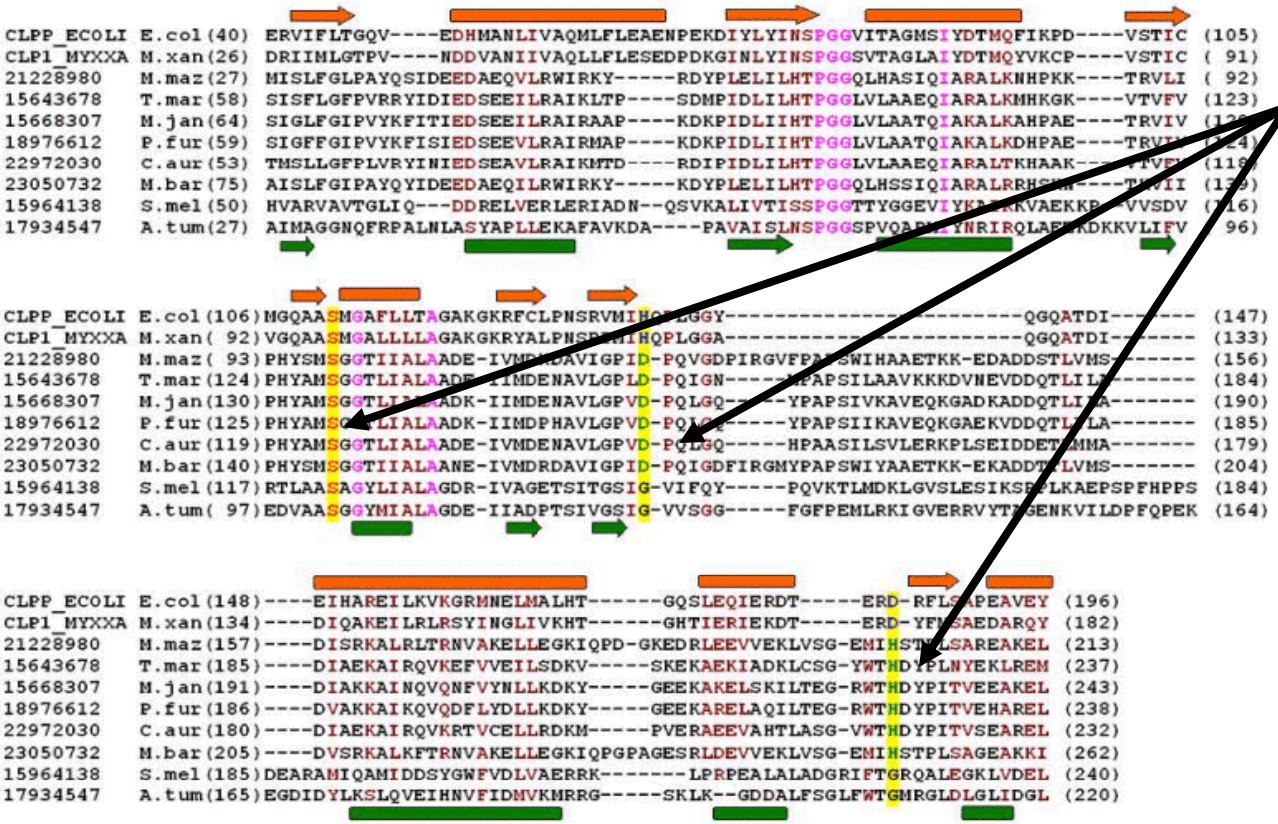
- Постоје два циља MSA:
  - Утврђивање који делови секвенце одговарају једни другима
  - Налажење позиција које су очуване

- Кључни кораци у MSA су:
  - Исправно одабирање секвенци које се поравнавају
  - Коришћење одговарајућег MSA метода алгоритма
  - Интерпретација поравнања

Кратко о очуваним позицијама

- Аминокиселине случајно мутирају
- На мутације онда делује позитивна селекција (бивају прихваћене) или негативна селекција (бивају одбачене)
- Ако је мутација штетна, на њу делује негативна селекција
  - Нестаје из генома
  - Никада је не видите
- Мутације битних позиција (нпр каталитички активна места) су готово увек штетна
- Можете да препознате битне позиције пошто готово никад не мутирају!
- MSAs открива *очуване* позиције

Пример очуване позиције:



Зашто правимо MSA ?

- Главни разлог је да би га користили у даљим применама
- MSA се често користи у прављењу различитих модела
- Следећа 2 слајда дају најчешће примене MSA

Примене MSA

Примена	Процедура
Екстраполација	Одређивање функције протеина
Филогенетска анализа	Прављење филогенетског дрвета
Идентификација образаца (pattern)	Открите битне делове секвенце
Идентификација домена	Претворите поравнање у профил домена
Идентификација регулаторних елемента у ДНК	Анализа промотера
Предвиђање структуре	Предвидите секундарну структуру РНК и протеина
PCR анализа	Предвиђање PCR прајмера

Одабирање одговарајућих секвенци

- Први корак у MSA је да одаберете одговарајуће секвенце
- Два главна фактора у одабирању секвенци:
  - Број секвенци
  - Природа секвенци

- Разуман број секвенци: 15 до 50
  - Идеалан за већину метода
  - Мања поравнања се лако приказују и анализирају

- Типови секвенци
  - Добро селектоване секвенце ⇔ информативно поравнање

Нека упутства за одабирање одговарајућих секвенци

Problem	*Diagnostic*
Proteins or DNA	Use proteins whenever possible.
Many sequences	Start with 10~15 sequences, 50 at max
Very different sequences	Avoid sequences very different from the rest of the set
Identical sequences	Avoid identical sequences; they never help
Partial sequences	Can trigger errors. Avoid them
Repeated domains	Can trigger errors. Extract the domains with Dotlet and align each domain.

Нека упутства за одабирање одговарајућих секвенци

Проблем	Решење
Протеини или ДНК	Користите протеине кад год је могуће
Колико секвенци	Почните са 10-15, максимално 50
Веома различите секвенце	Избегавајте секвенце веома различите од осталих
Идентичне (јако сличне) секвенце	Никад не помажу
Парцијалне секвнеце	Избегавајте их
Домени који се понављају	Извуците домене који се понављају помоћу dotlet-a

Одабирање одговарајућих секвенци

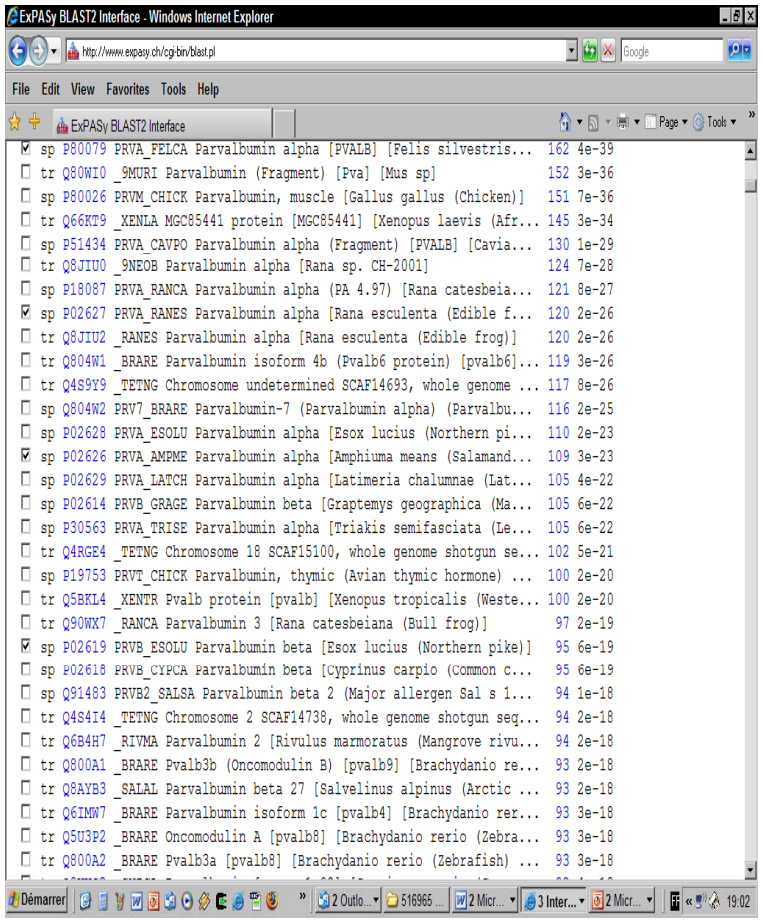
- Поравнање је корисно ако . . .
  - Су секвенце које поравнавате исправно изабране
- Да би добили резултате, секвенце морају да буду
  - Не превише сличне
  - Не превише различите
- Секвенце које су веома сличне . . .
  - Су једноставне да се добро поравнају
  - Нису превише информативне ⇔ лоша предвиђања
- Секвенце које су веома различите . . .
  - Су тешке да се поравнају
  - Су веома информативне ⇔ потенцијално добра предвиђања

Сакупљање секвенци помоћу BLAST-у

- Најпогоднији начин да селектујете секвенце је да користите BLAST сервер
- Неки од BLAST сервера су интегрисани са методама за поравнање више секвенци:
  - www.expasy.ch
  - srs.ebi.ac.uk
  - npsa-pbil.ibcp.fr

Сакупљање секвенци помоћу BLAST-а

- Равномерно селектујте секвенце од врха до дна
- Идеја је да имате равномеран прелаз између секвенци

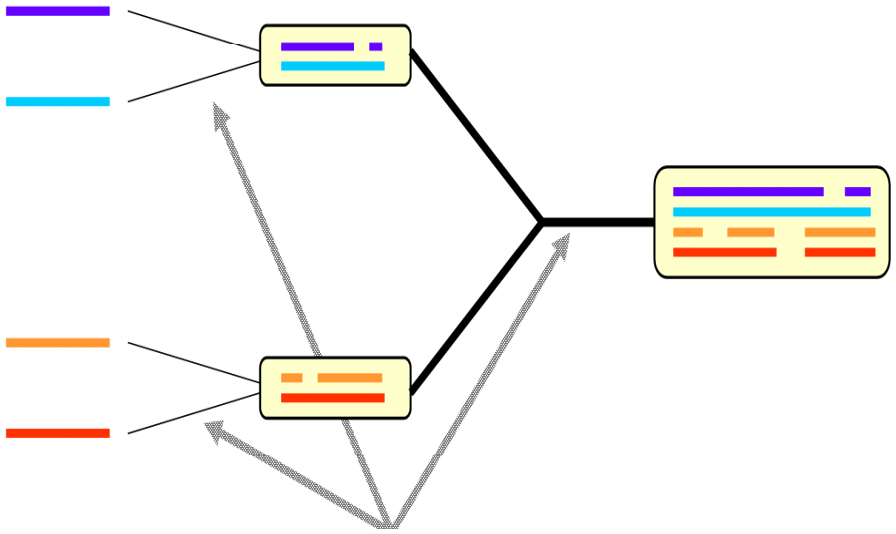


Поравнање секвенци

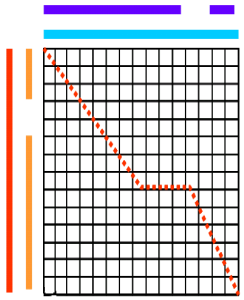
- Генерално веома тежак проблем
- Сви методи су апрокцимативни
- Методи за поравнање користе прогресивни алгоритам
  - Секвенце се упоређују две по две
  - Прави се дрво које дефинише процес упоређивања
  - Секвенце се поравнавају по реду које утврђује то дрво

Прогресивни алгоритам

- Секвенце се групишу по сличности (прави де дрво)
- Секвенце се поравнавају 2 по 2
- Добијена поравнања се онда даље поравнавају 2 по 2



Dynamic programming using a substitution matrix



Прогресивни алгоритам (наставак)

- Главна предност му је брзина
- Главна мана му је похлепа
  - Не враћа се на секвенце које је поравнао на почетку
  - Ране грешке се не отклањају
- Сакупите податке са доста прелазних секвенци
- Замислите да је свака секвенца део каменог моста преко реке:
  - Није битно колико је широка река, ако је каме ње довољно близу
  - Није битно колико су секвенце међусобно различите, докле год свака има блиског рођака

Селекција метода

- Постоје многе MSA методе
- Готово сви користе **прогресивни алгоритам**
- Сви методи су апроксимативни
- Ниједан није гарантован да води ка најбољем поравнању
- Сви постојећи методи имају добре и лоше стране
  - **ClustalΩ (наследник ClustalW)** је најпопуларнији (21,000 цитата)
  - T-Coffee и ProbCons су тачнији али спорији
  - MUSCLE је веома брз, идеалан за веома велике базе

Неки коментари о MSA . . .

- У пракси, прављење MSA није егзакна наука
- Одабир секвенци је најбитнији корак
  - Ако поравнање изгледа лоше, промените секвенце
- У пракси, прављење доброг MSA је итеративан (try-again) процес
  - Селектујете пар секвенци
  - Поравнате их
  - Визуелно оцените квалитет MSA
  - Додате или одузмете секвенце
  - Поново их поравнате

Интерпретација MSA

- Користите обрасце конзервације ClustalW
  - ‘*’ Комплетно конзервирана позиција
  - ‘:’ Високо конзервирана позиција
  - ‘.’ Конзервирана позиција
- При интерпретацији вашег поравнања, увек користите биолошку информацију (ако је на располагању)

1. Направите поравнање протеинских секвенци сигма фактора из ECF групе (сигма70 фамилија), датих у фајлу ECF_sigma_1.txt

Напомена: ECF сигма фактори учествују у регулацији експресије бактеријских гена под стресним условима, при чему за интеракцију са ДНК промотора користе домене σ^2 и σ^4.

Упутство: Користите програм ClustalΩ (default параметри) и на поравнању уочите границе домена σ^2 и σ^4.

Додатак: Који од поравнатих региона одговара σ^2, а који σ^4 домену? Да би ово са сигурношћу утврдили, изаберите неку од секвенци (нпр. прву) и пустите на њој CD search. Искористите резултат CD search-а да кратко прочитате о функцији ова два домена.

Локално поравнање више секвенци

- Већина MSA програма подразумева да су ваше секвенце повезане целом дужином
- Када ова претпоставка није тачна, прогресивни приступ не ради
- Једина алтернатива је да урадите локално поравнање више секвенци

Методи за локално MSA

- Gibbs Sampler
  - Игнорише неповезане делове ваших секвенци
  - Идеалан за налажење мотива у ДНК

- Методи за локално MSA
  - Тражи кратке мотиве који су очувани у секвенцама
  - Секвенце не треба да се глобално поравнају

- The most popular motif-discovery methods:
  - MEME, Bioprospector

3. Направите вишеструко локално поравнање ДНК секвенци, датих у фајлу MLSA.txt, при чему је дужина траженог мотива 5bp

Упутство: Користите програм MEME, http://meme-suite.org/tools/meme , MLSA.txt уносите под **Input the primary sequences.** Користите default параметре, осим:

1) Под **Select the site distribution**, користите **oops** опцију.
2) Под **Input job details** унесите вашу мејл дресу, ово је опционо, али типично ћете испробавати различите дужине мотива и њихове дистрибуције у секвенцама (види тачке 4 и 1 изнад), па су резултати овако прегледнији.
3) Под **Select number of motifs** одаберите **1**
4) Под **Advanced options** ставите и **minimum width** и **maximum width** на 5; ова опција одређује дужину мотива коју тражите у секвенцама – у овом случају тражите да мотив буде 5bp дугачак.

Задаци вежба

2. Исто као задатак означен са 1 (глобално поравнање више протеинских секвенци помоћу ClustalΩ), само сада поравнајте протеинске секвенце сигма фактора датих у фајлу ECF_sigma_2.txt. Секвенце такође припадају ECF групи сигма фактора, само су другачије селектоване. Упоредите поравнање са резултатима добијеним у задатку 1 (информативност поравнања, тј. могућност уочавања домена σ^2 и σ^4). Коментаришите у којој мери селекција секвенци може да утиче на поравнање, односно повежите са дискусијом у предавању.

4) Исто као задатак означен са 3 (локално поравнање више секвенци помоћу MEME-а), само сада урадите поравнање и са дужином мотива 4 и 6. Колико је мотив који налазите робустан у односу на овакву промену дужине? Такође, тестирајте и робусност у односу на промену у **Select the site distribution,** нпр. уместо **oops** одаберите **zoops** опцију. Да ли је добијени мотив и даље робустан? Приметите да logo (интуитивну визуелну репрезентацију) мотива можете да видите под **Discovered motifs** (кликом на плаву стрелицу усмерену надоле, лого можете да увећате) – помоћу логоа вам је најлакше да процените да ли добијате исти мотив (скраћен, померен за 1бп налево или надесно итд.) Да ли можете да препознате на шта се односи нађени мотив? У вези са тим, ако су секвенце које претажујете интергенски региони узводно од бактеријских гена (прецизније узводно од оперона), зашто очекујете да се овај мотив појави у свакој од претражибаних секвенци (односно у готово свакој)?