

ПОДРАВНЯВАНЕ НА БИОЛОГИЧНИ СЕКВЕНЦИИ

ПРОФ. ПЛАМЕНКА БОРОВСКА



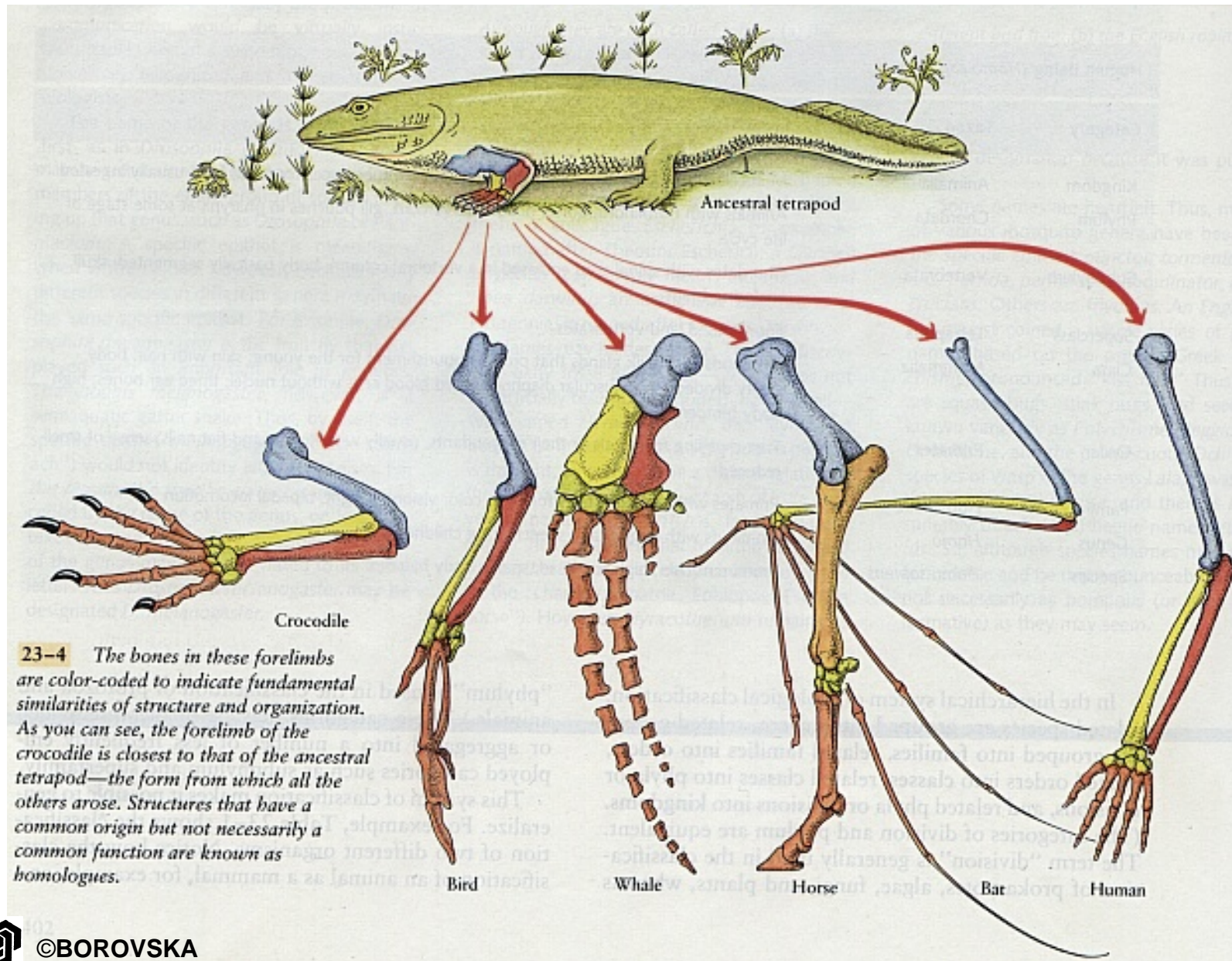
ПОДРАВНЯВАНЕТО НА БИОЛОГИЧНИ СЕКВЕНЦИИ СЕ ИЗПОЛЗВА ЗА:

- ◉ *Предсказване на функционалност*
- ◉ *Търсене в биологични бази данни*
- ◉ *Откриване на гени*
- ◉ *Откриване на различията
(дивергенция) между секвенциите*
- ◉ *Асемблиране на секвенции*

ПОДОБИЕ (СХОДСТВО) ПРИ СЕКВЕНЦИИТЕ

- ◉ **Хомология:** гени, които произлизат от един общ прародител – наричат се хомоложни гени (*хомолози*)
- ◉ **Ортология:** хомоложни гени в различни организми (*ортолози*)
- ◉ **Паралогия:** хомоложни гени в един организъм, които произхождат от **дублирането на гените** (*паралози*)
- ◉ **Дублиране на гените:** създават се множество копия на един ген, като всяко копие на гена може да еволюира отделно и независимо от другите и да придобива нова функционалност

ХОМОЛОЗИ И ПАРАЛОЗИ



ПРИЧИНИ ЗА СХОДСТВАТА И РАЗЛИЧИЯТА МЕЖДУ СЕКВЕНЦИИТЕ

- **Мутация:** в дадена позиция (определено място) на гена един нуклеотид се замества с **друг** нуклеотид
(напр., АТА → АГА)
- **Вмъкване:** в дадена позиция на гена нов нуклеотид се вмъква между два съществуващи нуклеотида
(напр., АА → АГА)
- **Заличаване:** в дадена позиция на гена се заличава съществуващ нуклеотид
(напр., АСТГ → АС-Г)
- **Вмъкване/заличаване (*indel*):** вмъкване (insertio) или заличаване (deletion)

БИОЛОГИЧНИЯТ ПРОБЛЕМ ЗА ПОДРАВНЯВАНЕ НА СЕКВЕНЦИИ

- Основна цел: откриване на сходствата между две (или повече) секвенции ДНК чрез намиране на сходните им участъци.

ДНК-секвенция 1

tcctctgctctgscatcat---caaccccaagt

||||| ||| ||||| ||||| ||||| ||||| |||||

tcctgtgcatctgcaatcatgggcaaccccaagt

ДНК-секвенция 2

*подравняване (Alignment) –
Привеждане в съответствие*



ДЕФИНИРАНЕ НА ПОДРАВНЯВАНЕТО НА СЕКВЕНЦИИ

- *подравняването на биологични секвенции* представлява подравняването на две или повече секвенции по начин, който дава възможност за *откриване на сходството между тях*.
- В секвенциите се вмъкват необходимият брой празни позиции (gaps), които се отбелязват с тирета, така че, колоните да съдържат **еднаквите символи** в секвенциите

```
tcctctgctctgccatcat---caaccccaagt
||||| ||| ||||| ||||| ||||| |||||
tcctgtgcatctgcaatcatgggcaaccccaagt
```

АЛГОРИТМИ ЗА ПОДРАВНЯВАНЕ НА СЕКВЕНЦИИ

- ◉ Алгоритъм на Needleman-Wunsch – глобално подравняване по двойки (Pairwise global alignment).
- ◉ Алгоритъм на Smith-Waterman – локално или глобално подравняване по двойки (Pairwise, local or global alignment).
- ◉ BLAST - Pairwise – евристично локално подравняване (heuristic local alignment)

ПОДРАВНЯВАНЕ ПО ДВОЙКИ

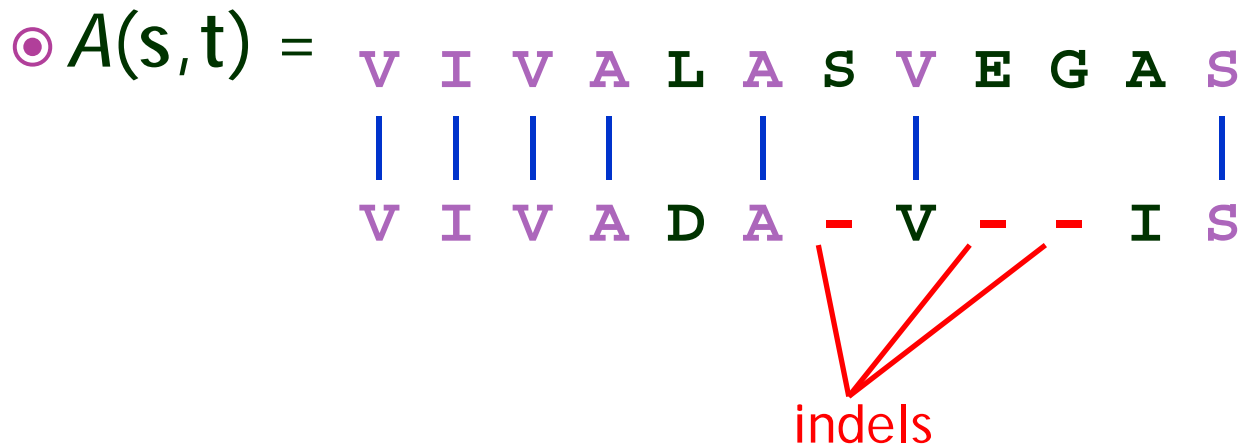
- Основната цел на методите за подравняване на биологични секвенции по двойки е да се направи такова локално или глобално подравняване на нуклеотидните или протеиновите секвенции, което да открива максимален брой еднакви участъци в тях.
- Най-често, целта е откриването на **хомолози** (**роднини**) на специфициран ген или или генни продукти в биологични бази данни с познати гени.
- Получената информация е полезна при даването на отговор на различни биологични въпроси като:
- 1. Идентифицирането на секвенции с **неизвестна** структура или функция.
- 2. Изследването на **молекулярната еволюция**.

ГЛОБАЛНО ПОДРАВНЯВАНЕ

- ◉ Глобалното подравняване на две секвенции обхваща всичките символи (нуклеотиди или протеини) на секвенциите.
- ◉ *подравняване = привеждане в съответствие*
- ◉ Най-често глобалното подравняване се използва за откриването на тясно свързани секвенции (много сходни секвенции).
- ◉ За същата цел се използват успешно и методите за локално подравняване.
- ◉ Освен това, съществуват усложнения при молекулярната еволюция, като напр., разместването на домейни (domain shuffling), които ограничават полезността на методите за глобално подравняване.

ГЛОБАЛНО ПОДРАВНЯВАНЕ

- Намиране на глобално подравняване на две секвенции, което показва всички еднакви символи
- Пример: секвенция $s = \text{VIVALASVEGAS}$ и секвенция $t = \text{VIVADAVIS}$ се подреждат глобално по следния начин:



АЛГОРИТЪМ НА NEEDLEMAN-WUNSCH

- Използва се както при нуклеотидните, така и при протеиновите секвенции
- Оценъчна функция за качеството на подравняването (*Alignment scoring function*)
- *Качеството на подравняването* (alignment cost) на два символа x_i и y_j се оценява с функцията $\sigma(x_i, y_j)$
- *Качеството на подравняването на целите секвенции се оценява посредством сумата от оценките на отделните символи в секвенциите*

$$M = \sum_{i=1}^c \sigma(x_i, y_i)$$

ПРОСТА ОЦЕНЪЧНА ФУНКЦИЯ

- Вмъкване на празна позиция в секвенцията

$$\sigma(-, a) = \sigma(a, -) = -1$$

- Различни символи на секвенциите в една и съща позиция (колона)

$$\sigma(a, b) = -1 \text{ if } a \neq b$$

- Еднакви символи на секвенциите в една и съща позиция (колона)

$$\sigma(a, b) = 1 \text{ if } a = b$$

ОЦЕНЪЧНА ФУНКЦИЯ

- Качеството (цената) на подравняването на двете секвенции $s = \text{VIVALASVEGAS}$ и $t = \text{VIVADAVIS}$:

$$A(s,t) = \begin{array}{cccccccccccc} & \text{V} & \text{I} & \text{V} & \text{A} & \text{L} & \text{A} & \text{S} & \text{V} & \text{E} & \text{G} & \text{A} & \text{S} \\ & | & | & | & | & & | & & | & & & & | \\ \text{V} & \text{I} & \text{V} & \text{A} & \text{D} & \text{A} & - & \text{V} & - & - & \text{I} & \text{S} \end{array}$$

се оценява по следния начин:

$$\begin{aligned} M(A) &= 7 \text{ съвпадения} + 2 \text{ разлики} + 3 \text{ празни позиции} \\ &= 7 - 2 - 3 = 2 \end{aligned}$$

МАТРИЦА НА ЗАМЕСТВАНИЯТА (THE SUBSTITUTION MATRIX)

- По-реалистична оценъчна функция,
която е инспирирана от биологията:

-	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

ОПТИМАЛНО ГЛОБАЛНО ПОДРАВНЯВАНЕ

- Оптималното глобално подравняване A^* на две секвенции s и t е такова подравняване $A(s,t)$, при което се получава максимална стойност на общата оценъчна функция $M(A)$ в сравнение с всички възможни подравнявания.

$$A^* = \max M(A_i)$$

- Намирането на оптималното глобално подравняване A^* е комбинаторен оптимизационен проблем и обхваща следните стъпки:*
 - Генериране на всички възможни подреждания;
 - Изчисляване на качеството на всички подреждания M ;
 - Селекция на оптималното подравняване A^* с максимално качество M^* ;

ЛОКАЛНО ПОДРАВНЯВАНЕ

- ◉ *Методите за локално подравняване откриват сходни участъци в рамките на секвенциите* – подравняването обхваща подмножества от символите в изследваните секвенции.
- ◉ Напр., **позиции 30-59** в секвенция A могат да бъдат подредени с **позиции 30-59** в секвенция B.
- ◉ Тази техника се счита за по-гъвкава от глобалното подравняване и има предимството, че сходни участъци, които са подредени по различен начин в изследваните секвенции (*domain shuffling*) могат да бъдат идентифицирани като сходни.
- ◉ Това не е възможно при методите за глобално подравняване.

АЛГОРИТЪМ НА SMITH WATERMAN

- Алгоритъмът Smith-Waterman (1981) се използва за откриване на сходните участъци в две нуклеотидни или протеинови секвенции.
- Базира се на техниките на динамичното програмиране
- Представява подобрение на алгоритъма на Needleman-Wunsch
- *Осигурява гарантирано откриване на оптималното локално подравняване* по отношение на използваната оценъчна функция, която включва матрицата на заместванията и схема за оценка (“наказания”) на вмъкнатите празни позиции.

АЛГОРИТЪМ НА SMITH WATERMAN

- Поради факта, че алгоритъмът на Smith-Waterman изчерпва всички възможни подравнявания за да гарантира намирането на оптималното подравняване, той изисква мощни изчислителни ресурси и консумира изключително голямо изчислително време – за подравняването на две секвенции с дължини m и n , сложността на алгоритъма се оценява на $O(mn)$
- Практически, най-популярен е алгоритъмът BLAST, който се базира на евристични техники, и осигурява откриването на добро (суб-оптимално) подравняване за приемливо време.

БИОЛОГИЧНАТА ИНТЕРПРЕТАЦИЯ НА ПОДРАВНЯВАНЕТО НА СЕКВЕНЦИИ

- подравняването на биологичните секвенции е ефективен метод *за изследването на еволюцията и определянето на общ произход.*
- *Разликите между биологичните секвенции при подравняването съответстват на мутациите, а празните позиции (gaps) съответстват на вмъкване или заличаване на ген (протеин).*
- подравняването на биологични секвенции се използва, също така, за подреждания на потенциално несвързани секвенции в биологичните бази данни.

ПОДРАВНЯВАНЕ НА ДВОЙКА СЕКВЕНЦИИ

1. *Анализ с точкова матрица*
2. *Алгоритми, базирани на динамичното програмиране (DP)*
3. *Методи с обработка по думи (k-tuple methods), използвани от софтуера FASTA и BLAST*

АНАЛИЗ С ТОЧКОВА МАТРИЦА

- Освен в случаите, когато е известно, че секвенциите са много сходни, препоръчително е първо да се използва метода с точкова матрица, тъй като с негова помощ могат да се наблюдават всички възможни съответствия като диагонали на матрицата.
- Анализът на основата на точкова матрица може лесно да открие наличието на вмъквания и заличавания, както и повторенията (вкл. и инвертираните повторения), които се откриват по-трудно с други методи
- Основното ограничение на този метод, е че повечето програми, основани на анализ с точкова матрица, не показват реално подравняване

СРАВНЯВАНЕ НА СЕКВЕНЦИИ С ТОЧКОВА МАТРИЦА

- *Анализът с точкова матрица се използва основно като метод за сравняване на две секвенции с цел търсене на възможни съответствия на символи между секвенциите*
- Методът се използва също така за откриване на преки или инвертирани повторения в протеинови или нуклеотидни секвенции, за предсказване на самодопълващи се участъци в РНК, и следователно, имат потенциал за формиране на вторична структура.
- Препоръчително е, всяка лаборатория, в която се прави анализ на секвенции, да разполага поне с една програма за анализ с точкови матрици.

СРАВНЯВАНЕ НА СЕКВЕНЦИИ С ТОЧКОВА МАТРИЦА

- Програмата DOTTER използва метода на точковата матрица за анализ на нуклеотидни и протеинови секвенции
<http://sonnhammer.sbc.su.se/Dotter.html>
- Програмите COMPARE и DOTPLOT на Genetics Computer Group също се използват за анализ с точкова матрица.
- Въпреки, че не използва метода с точковата матрица, програмата PLALIGN в рамките пакета FASTA се основава на метода на динамичното програмиране и може да се използва за откриване на съответствията между две секвенции върху граф (Pearson 1990).

http://fasta.bioch.virginia.edu/fasta/fasta_list.html

СРАВНЯВАНЕ НА СЕКВЕНЦИИ ПО ДВОЙКИ ПО МЕТОДА С ТОЧКОВА МАТРИЦА

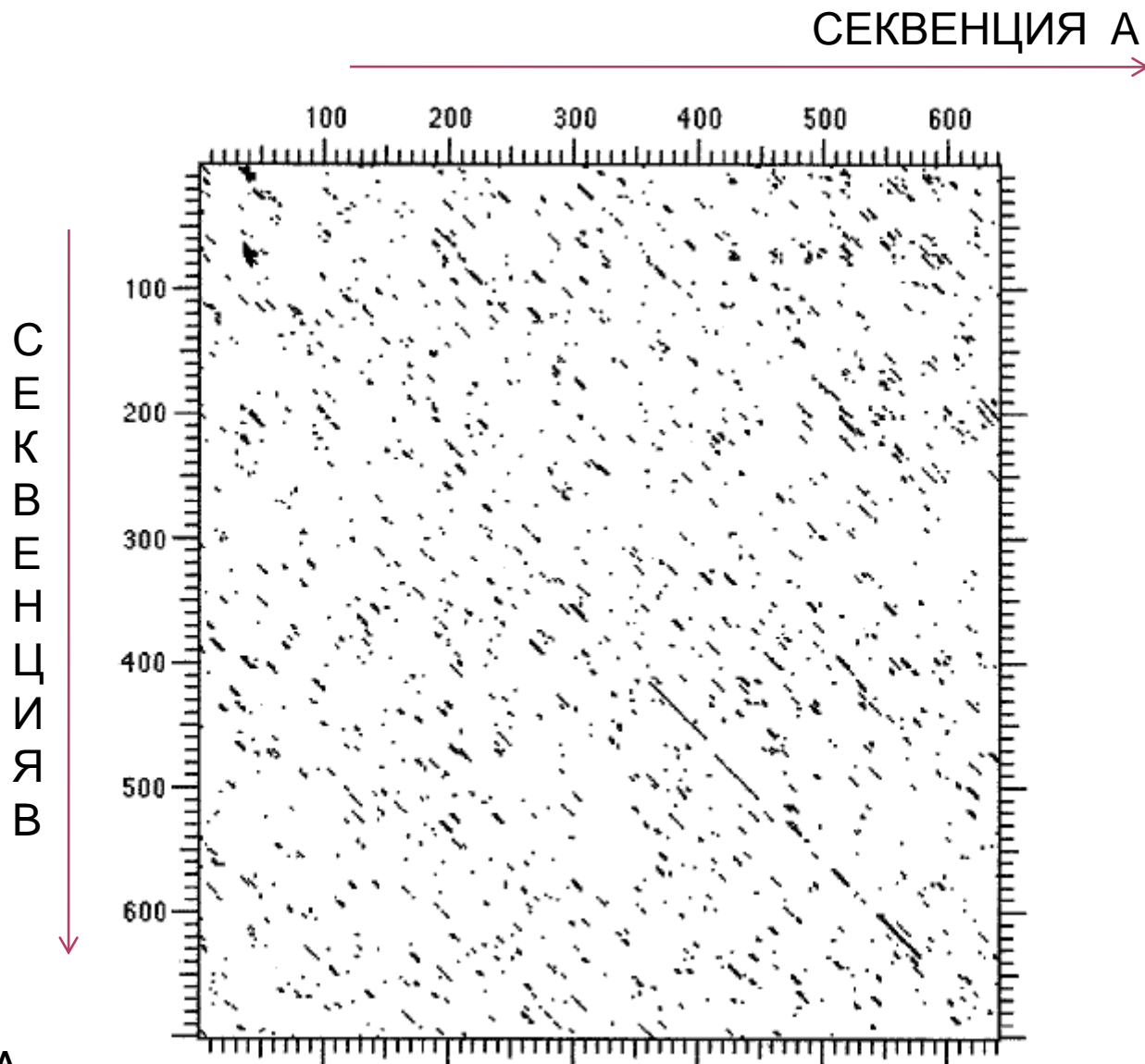
- При метода за сравнение на секвенции с точкова матрица, едната секвенция (А) се разполага хоризонтално в горната част на страницата, а другата секвенция (В) се разполага вертикално в левия край от горе надолу
- Запълването на точковата матрица стартира от първия ред, като първият символ в секвенция В се сравнява последователно със символите на секвенция А. При откриване на съвпадение на символите в двете секвенции във всяка колона на първия ред се поставя точка.
- Следва запълването на втория ред на матрицата, като вторият символ в секвенция В се сравнява последователно със символите на секвенция А. При откриване на съвпадение на символите в двете секвенции във всяка колона на втория ред се поставя точка.



СРАВНЯВАНЕ НА СЕКВЕНЦИИ ПО ДВОЙКИ ПО МЕТОДА С ТОЧКОВА МАТРИЦА

- Процедурата се повтаря като за всеки символ от секвенция В се запълва съответния ред на точковата матрица до изчерпването на всички символи в секвенция В
- Матрицата е изпълнена с точки, представлящи всички възможни съвпадения между символите на секвенция А и символите на секвенция В
- Участъците с еднакви последователности от символи се индицират с диагонал от точки
- Изолираните точки извън диагоналите представят случайни съвпадения, които най-вероятно не са свързани със значимо сходство

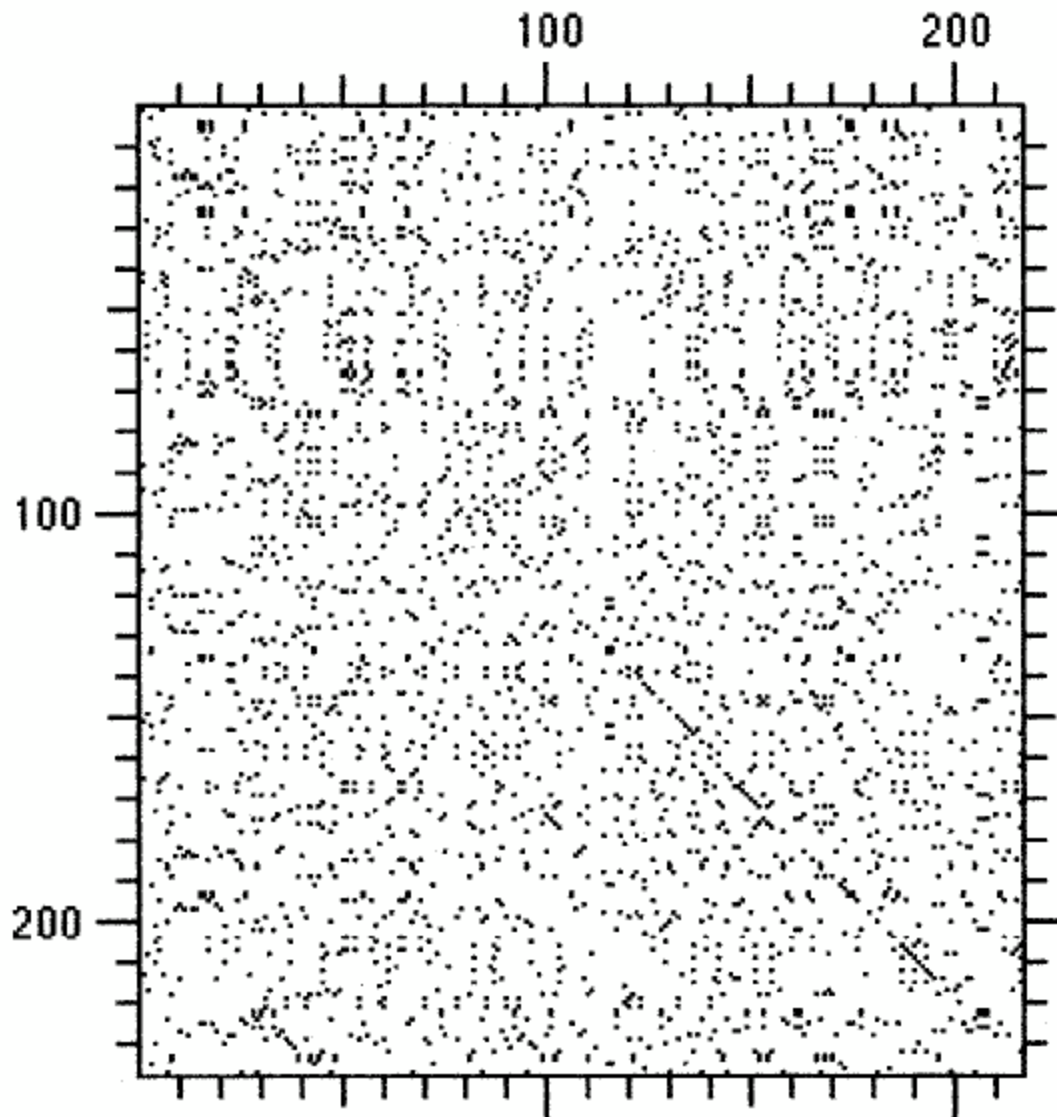
ПРОГРАМА DNA STRIDER ЗА АНАЛИЗ НА НУКЛЕОТИДНИ СЕКВЕНЦИИ



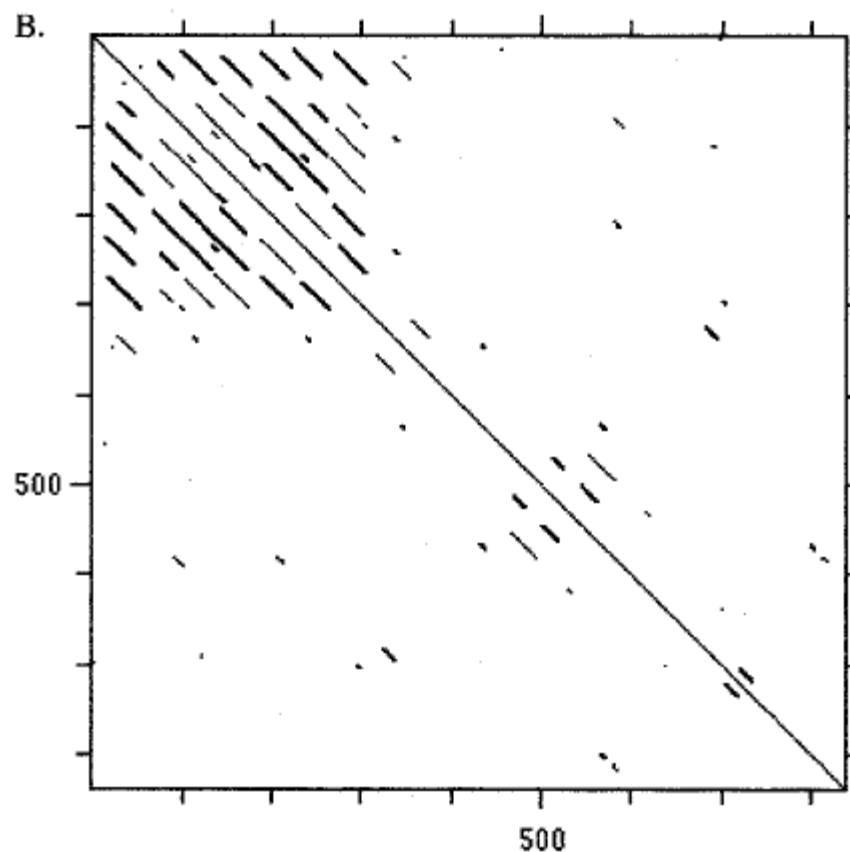
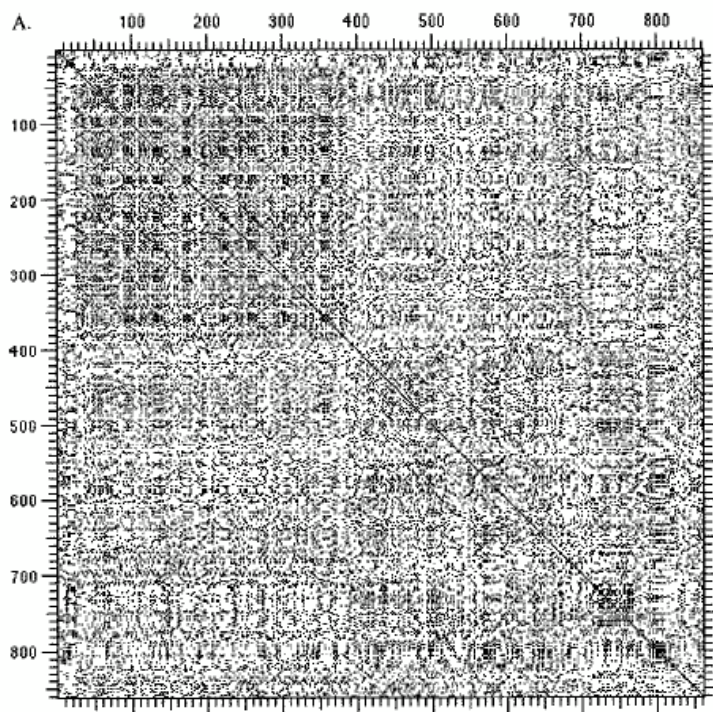
МЕТОД С ТОЧКОВА МАТРИЦА

- Откриването на съвпадащи участъци може да бъде подобро чрез филтриране на случайните съвпадения в точковата матрица.
- Филтрацията се осъществява посредством “плъзгащ се прозорец” за сравнение на двете секвенции.
- Вместо да се сравняват единични позиции в секвенциите, се използва *“прозорец”, обхващащ съседни позиции в двете секвенции, които се сравняват едновременно*
- В матрицата се отпечатва точка, само ако има съвпадение при предефиниран брой съседни позиции на двете секвенции (думи с фиксирана дължина)
- Прозорецът започва от сравняваните позиции в секвенции A и B и обхваща символи в диагонала, като се движи надолу и надясно, сравнявайки всяка двойка, аналогично на подравняването.

АНАЛИЗ ПО МЕТОДА С ТОЧКОВА МАТРИЦА НА СЕКВЕНЦИИ ОТ АМИНО КИСЕЛИНИ



АНАЛИЗ ПО МЕТОДА С ТОЧКОВА МАТРИЦА НА HUMAN LDL РЕСЕРТОР СЪС САМИЯ СЕБЕ СИ КАТО СЕ ИЗПОЛЗВА DNA STRIDER



МЕТОД С ТОЧКОВА МАТРИЦА

	М	А	Т	С	Н	М	А	К	Е	Р
М	*					*				
А		*					*			
К								*		
Е									*	
А		*					*			
М	*					*				
А		*					*			
Т			*							
С				*						
Н					*					
М	*					*				
А		*					*			
К								*		
Е									*	
Р										*

МЕТОД С ТОЧКОВА МАТРИЦА

	М	А	Т	С	Н	М	А	К	Е	Р
М	*					*				
А		*					*			
К							*			
Е								*		
А		*				*				
М	*					*				
А		*					*			
Т			*							
С				*						
Н					*					
М	*					*				
А		*					*			
К							*			
Е		*						*		
Р									*	

Пряко
съвпадение

Обратно
съвпадение

Пълно
съвпадение