БИОНФОРМАТИКА 5

ПРОФ. ПЛАМЕНКА БОРОВСКА

Алгоритъм на динамичното програмиране

- В математиката,компютърните науки,икономиката и биоинформатиката, **динамичното програмиране** е метод за решаване на изключително сложни проблеми посредством разбиването им на по-прости подпроблеми.
- Алгоритмите на ДП се използват за оптимизация
- Методът е изключително ефективен в случаите, когато броят на повтарящите се подпроблеми расте експоненциално като функция на размера на входните данни.
- ДП е интелигентен brute-force метод, който ни дава всички възможни решения за да изберем най-доброто
- Методът на ДП гарантира намирането на оптималното решение

Алгоритъм на динамичното програмиране

- Прилага се за решаването на проблеми, които съдържат припокриващи се подпроблеми и оптимални субструктури.
- Основната идея на динамичното програмиране – за решаването на даден проблем е необходимо да решим различни части от него (подпроблеми), след което комбинираме решенията на подпроблемите за получаване на цялостното решение на проблема.
- При използването на по-прости методи, много от подпроблемите се генерират и решават много пъти. При динамичното програмиране всеки подпроблем се решава еднократно при което значително се намалява броят на изчисленията му.
- След като веднъж е изчислено, решението на дадения подпроблем се съхранява -it is "memo-ized“. При всеки следващ момент, когато това решение е необходимо, просто резултатът се извиква.

Алгоритъм на динамичното програмиране

- В компютърните науки ***memoization*** е оптимизационна техника, използвана предимно за ускоряване на изпълнението на програмата чрез използване на function calls за да се избегне повторението на вече изпълнени изчисления
- В контекста на някои езици за логическо програмиране, memoization е известна като tabling (lookup table).
- Терминът *"memoization"* за пръв път е използван от Donald Michie (Оксфорд, Единбург, специалист по изкуствен интелект) през 1968 г. и произхожда от латинската дума *memorandum* (*да се запомни*), или *"memo"* - да се запомнят резултатите на функциите.
- Терминът *memoization* да не се бърка с *memorization*

АЛГОРИТЪМ ПО МЕТОДА НА ДИНАМИЧНОТО ПРОГРАМИРАНЕ ЗА ПОДРЕЖДАНЕ НА БИОЛОГИЧНИ СЕКВЕНЦИИ

- Динамично програмиране е изчислителен метод, който се използва за подреждане на две секвенции от протеини или от нуклеинови киселини.
- Този метод е от изключително значение за анализа на биологични секвенции, тъй като дава оптималното подреждане за две секвенции
- При метода на динамичното програмиране се извършва сравнение на всяка двойка символи в двете секвенции и се генерира подреждане на секвенциите.
- Това подреждане включва съвпадащи и несъвпадащи символи и празнини в двете секвенции, които са разположени така, че броят на съвпаденията между еднакви или свързани знаци, е максималния възможен.

АЛГОРИТЪМ ПО МЕТОДА НА ДИНАМИЧНОТО ПРОГРАМИРАНЕ ЗА ПОДРЕЖДАНЕ НА БИОЛОГИЧНИ СЕКВЕНЦИИ

- *Математически е доказано, че този метод генерира оптималното подреждане на две секвенции* при зададено множество на условията за съвпадения.
- Оптималните подреждания предоставят полезна информация за биолозите за релациите между секвенциите като напр., кои символи в секвенциите трябва да се позиционират в една и съща колона при подреждането, и като резултат се виждат позициите на вмъкване в едната секвенция (или заличаванията съответно в другата секвенция).
- Получената информация е важна за прогнозирането на функцията, структурата и еволюцията

АЛГОРИТЪМ ПО МЕТОДА НА ДИНАМИЧНОТО ПРОГРАМИРАНЕ ЗА ПОДРЕЖДАНЕ НА БИОЛОГИЧНИ СЕКВЕНЦИИ

- Софтуерът за глобално подреждане на секвенции се базира на алгоритъма *Needleman-Wunsch*, докато софтуерът за локално подреждане на секвенции се базира на алгоритъма *Smith-Waterman*
- Методът на динамичното програмиране може да се използва и за подреждане на множество секвенции, но само за малък брой секвенции, тъй като сложността на алгоритъма нараства значително за повече от две секвенции
- Програми за подреждане на секвенции се предлагат като част от софтуерните пакети за обработка и анализ на секвенции като напр., широко използвания пакет *Genetics Computer Group GAP (global alignment) and BESTFIT (local alignment) programs*.

Цел на анализа

- Откриване на общи участъци в структурите на два протеина или еднакви особености на структурата?
- Принадлежат ли изследваните протеини към едно и също семейство по отношение на дадена биологична фунция
- Дали протеините споделят общ прародител (общ произход)
- Важни аспекти за изборът на типа на използвания софтуер, дали да се направи глобално или локално подреждане, избор на вида на оценъчната матрица и стойността на санкциите за наличието на празни позиции

THE DYNAMIC PROGRAMMING ALGORITHM

- S_{ij} *е оценката за позиция* i *в секвенция* a *и позиция* j *в секвенция* b
- $s(a_i b_j)$ е оценката за подреждането на символите в позиции i *и* j
- w_x *е санкцията за участък от* x *празнини* в секвенция **a**
- w_y *е санкцията за участък от* y *празнини* в секвенция ***b***
- *Алгоритъмът прохожда всяка позиция в матрицата като оценява стойностите на всяко* S_{ij}
- Когато всички позиции на матрицата (всички S_{ij}) *са запълнени, най-високата оценка на подреждането* се намира в последния ред и колона
- *Използването на метода на ДП изисква система за оценка при сравняването на двойка символи (нуклеотиди при ДНК или амино-киселини при протеиновите секвенции), както и схема за санкции при вмъкване/заличаване (GAP penalties).*

Алгоритъм по метода на ДП

```
1.  SCORE OF NEW      =    SCORE OF PREVIOUS +  SCORE OF NEW
    ALIGNMENT               ALIGNMENT (A)        ALIGNED PAIR

V  D  S  -  C  Y          V  D  S  -  C           Y
V  E  S  L  C  Y          V  E  S  L  C           Y

      15                =        8             +     7


II. SCORE OF          =    SCORE OF PREVIOUS +  SCORE OF NEW
    ALIGNMENT (A)           ALIGNMENT (B)        ALIGNED PAIR

V  D  S  -  C             V  D  S  -              C
V  E  S  L  C             V  E  S  L              C

      8                 =       -1             +     9

III. REPEAT REMOVING ALIGNED PAIRS UNTIL END OF ALIGNMENT IS REACHED.
```

Метод на динамичното програмиране за подреждане на секвенции

- Методът на динамичното програмиране предполага да се следва най-добрата конфигурация на подреждането, получена до момента.
- Правят се опити за подреждането на две секвенции посредством вмъкването на празнини в различни позиции, така че да се максимизира оценката на съвпаденията при това подреждане
- **Оценката се определя на базата на "match award“ – награда за съвпаденията, "mismatch penalty" – санкции за разликите, и "gap penalty“ – санкции за празнините.** Колкото е по-висока оценката, толкова е по-добро подреждането.

Метод на динамичното програмиране за подреждане на секвенции

- Ако санкциите за несъвпаденията и празнините се фиксират на 0, то целта е да се открие подреждане с максимален брой съвпадения.
- **Maximum match** = дефинира се като максималния брой съвпадения, които могат да се получат чрез всички възможни изтривания на различните позициии в другата секвенция
- Използва се при сравнението на протеинови или ДНК секвенции, за предсказване на сходства в тяхната функционалност.
- примери: **Needleman-Wunsch(1970), Sellers(1974), Smith-Waterman(1981)**

Алгоритъм на Needleman-Wunsch

- Осъществява глобално подреждане на две секвенции (A и B) и се прилага за подреждане на протеинови или нуклеотидни секвенции.
- Основава се на метода на динамичното програмиране и гарантира изчисляването на максималното подреждане
- Оценките за подредените символи се специфицират от матрица на преходите $\sigma(i,j)$: сходство на символите i и j.

Глобално подреждане: алгоритъм на Needleman-Wunsch

- Методът на динамичното програмиране е разширен с подобрена система за оценка от Smith и Waterman
- Оптималната оценка за всяка позиция в матрицата се изчислява като се прибавя оценката за текущата позиция към сумарната оценка от предишните позиции и се изваждат санкциите за празнините
- Оценката (сумарният брой точки) за всяка позиция на матриците може да бъде положително число, отрицателно числи, или 0.
- Алгоритъмът на Needleman-Wunsch максимизира броя на позициите със съвпадения между секвенциите по цялата им дължина.

The Needleman-Wunsch algorithm

1. Създава се таблица (m+1)x(n+1) за секвенциите **s** и **t** с дължини m и n, съответно

2. Попълват се в таблицата позиции (m:1) и (1:n) със стойностите:

$$M_{i,1} = \sum_{k=1}^{i} \sigma(\mathbf{s}_k, -), \quad M_{1,j} = \sum_{k=1}^{j} \sigma(-, \mathbf{t}_k)$$

3. Започвайки от горния ляв ъгъл, се изчислява всяка позиция с помощта на рекурсивната релация :

$$M_{i,j} = \max \begin{Bmatrix} M_{i-1,j-1} + \sigma(\mathbf{s}_i, \mathbf{t}_j) \\ M_{i-1,j} + \sigma(\mathbf{s}_i, -) \\ M_{i,j-1} + \sigma(-, \mathbf{t}_j) \end{Bmatrix}$$

4. Прави се обратно прохождане на матрицата от долния десен ъгъл

The Needleman-Wunsch algorithm

Матрица на преходите

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

При санкция за празнини **-5**, оценката ще бъде

$$S(A,C)+S(G,G)+S(A,A)+3\times d+S(G,G)+S(T,A)+S(T,C)+S(A,G)+S(C,T)$$
$$= -3+7+10-3\times 5+7+-4+0+-1+0=1$$

АЛГОРИТЪМ NEEDLEMAN-WUNSCH

- След изчисляването на матрицата F, долният десен ъгъл на матрицата съдържа максималната възможна оценка за произволно подреждане на двете секвенции
- За да се определи подреждането, съответстващо на максималната оценка, стартираме от най-долната лява клетка на матрицата, и сравняваме оценката с трите възможни избора ($Choice_1$, $Choice_2$, и $Choice_3$ отгоре) за да открием посоката.
- В случай на $Choice_1$, то A(i) и B(i) са подредени, при $Choice_2$ - A(i) е подредено с празнина, при $Choice_3$ - B(i) е подредено с празнина.

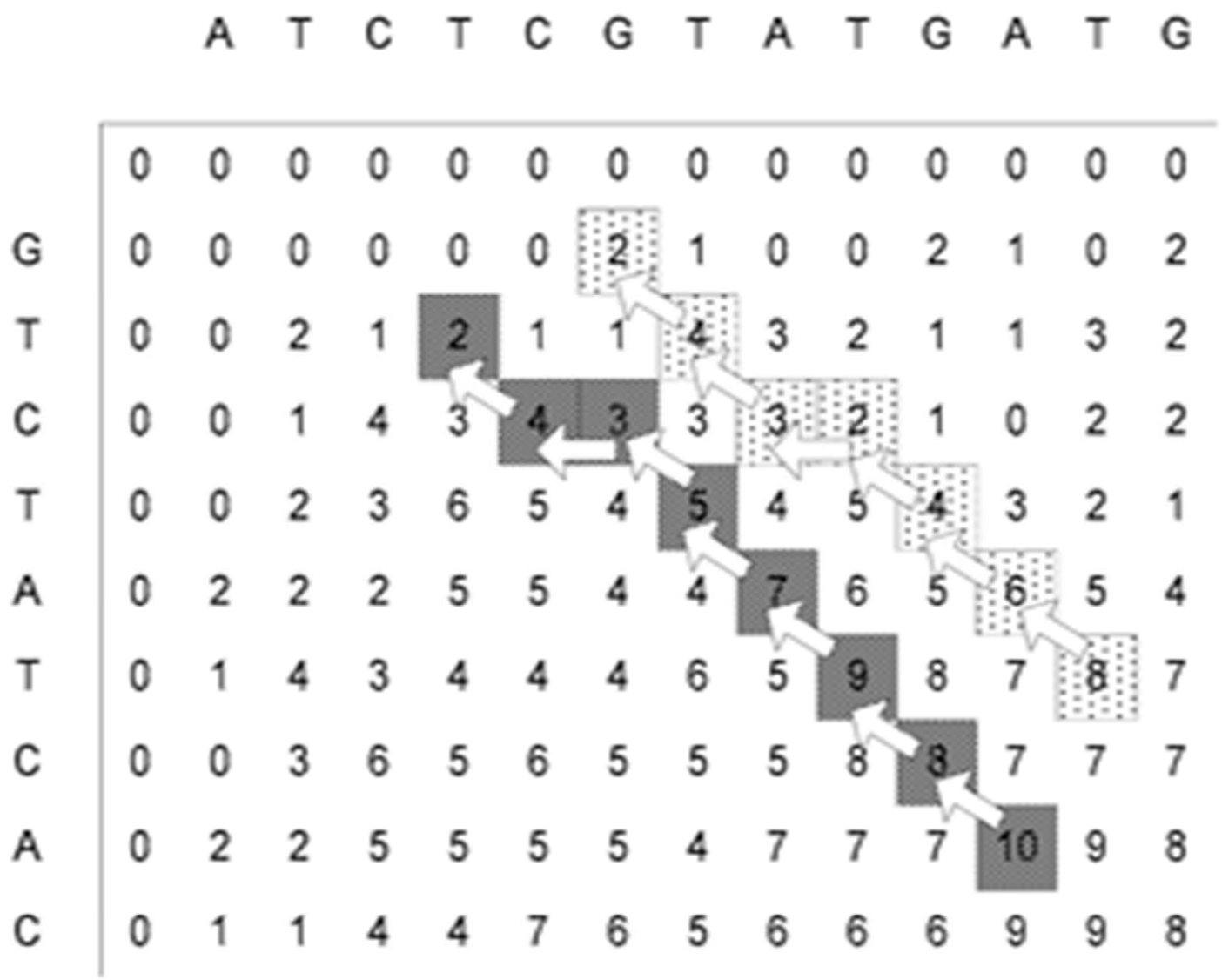
Алгоритъм на ДП

- Съществуват 3 възможни алтернативи в оценъчната матрица за достигане на дадена позиция: (1)ход по диагонала от позиция *i-1,j-1 към позиция i,j без санкции за празнини; (2-3) ход от произволна друга позиция* от колона *j или ред i, със санкция за празнини, която зависи от броя на празните позиции.*
- *За две секвенции* $\mathbf{a=a_1a_2...a_n}$ **and** $\mathbf{b=b_1b_2...b_n}$, ***където*** $S_{ij}=S(a_1a_2 \ldots a_i, b_1b_2..b_j)$ ***then*** (Smith and Waterman 1981)

$$S_{ij} = \max \left\{ \begin{array}{l} S_{i-1,j-1} + s(a_i b_j), \\ \max\limits_{x \geq 1} (S_{i-x,j} - w_x), \\ \max\limits_{y \geq 1} (S_{ij-y} - w_y) \end{array} \right\}$$

The Needleman-Wunsch algorithm

	i	A	B	C	N	J	R	Q	C	L	C	R	P	M
j	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12	-13
A	-1	0 end	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
J	-2	1	0	0	-1	1	0	-1	-2	-3	-4	-5	-6	-7
C	-3	0	0	1	2	1	0	-1	1	0	-1	-2	-3	-4
J	-4	-1	-1	2	2	2	3	2	1	0	-1	-2	-3	-4
N	-5	-2	-2	1	3	3	3	2	1	0	-1	-2	-3	-4
R	-6	-3	-3	0	3	3	4	4	3	2	1	1	0	-1
C	-7	-4	-4	-1	2	2	4	4	5	5	4	3	2	1
K	-8	-5	-5	-2	1	1	3	3	5	6	5	4	3	2
C	-9	-6	-6	-3	0	0	2	2	5	5	7	6	5	4
R	-10	-7	-7	-4	-1	-1	2	1	4	4	6	8	8	7
B	-11	-8	-8	-5	-2	-2	1	0	3	3	5	8	8	7
P	-12	-9	-9	-6	-3	-3	0	-1	2	2	4	7	10	9 ← start



		A	T	C	T	C	G	T	A	T	G	A	T	G
	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	2	1	0	0	2	1	0	2
T	0	0	2	1	2	1	1	4	3	2	1	1	3	2
C	0	0	1	4	3	4	3	3	3	2	1	0	2	2
T	0	0	2	3	6	5	4	5	4	5	4	3	2	1
A	0	2	2	2	5	5	4	4	7	6	5	6	5	4
T	0	1	4	3	4	4	4	6	5	9	8	7	8	7
C	0	0	3	6	5	6	5	5	5	8	8	7	7	7
A	0	2	2	5	5	5	5	4	7	7	7	10	9	8
C	0	1	1	4	4	7	6	5	6	6	6	9	9	8

Similarity matrix of the alignment of sequence ATCTCGTATGATG and GTCTATCA. k=2, substitution cost of +2 if the two characters are identical and -1 otherwise. Both the first and extension gap penalty are -1. Two traceback paths deliver the non-intersecting optimal and near-optimal local alignments.

ПРИМЕР:
EYELESS GENE HOMEOBOX

- Сравнение на гените на Drosophila Melanoganster с човешкия ген aniridia.
- Наблюдават се регулаторни участъци създаващи протеини, които управляват дълги каскади от гени.
- Определени сегменти в Drosophila melanogaster и човешкия aniridia са почти идентични.
- Най-важният сегмент кодира участъка PAX (paired-box), секвенция от 128 аминокиселини, чиято функция е да свързва специфични секвенции от ДНК.
- Друг общ семенг е HOX (homeobox), който се съдържа в повече от 0.2% от общия брой на гените на всички гръбначни.

ПРИМЕР:
EYELESS GENE HOMEOBOX

Compare the HOX domain

Here we compare the HOX domain of human and fly. The peptide sequences can be obtained from the GenBank database using **getgenPept** to download the data from the NCBI site. First read the human protein information into MATLAB.

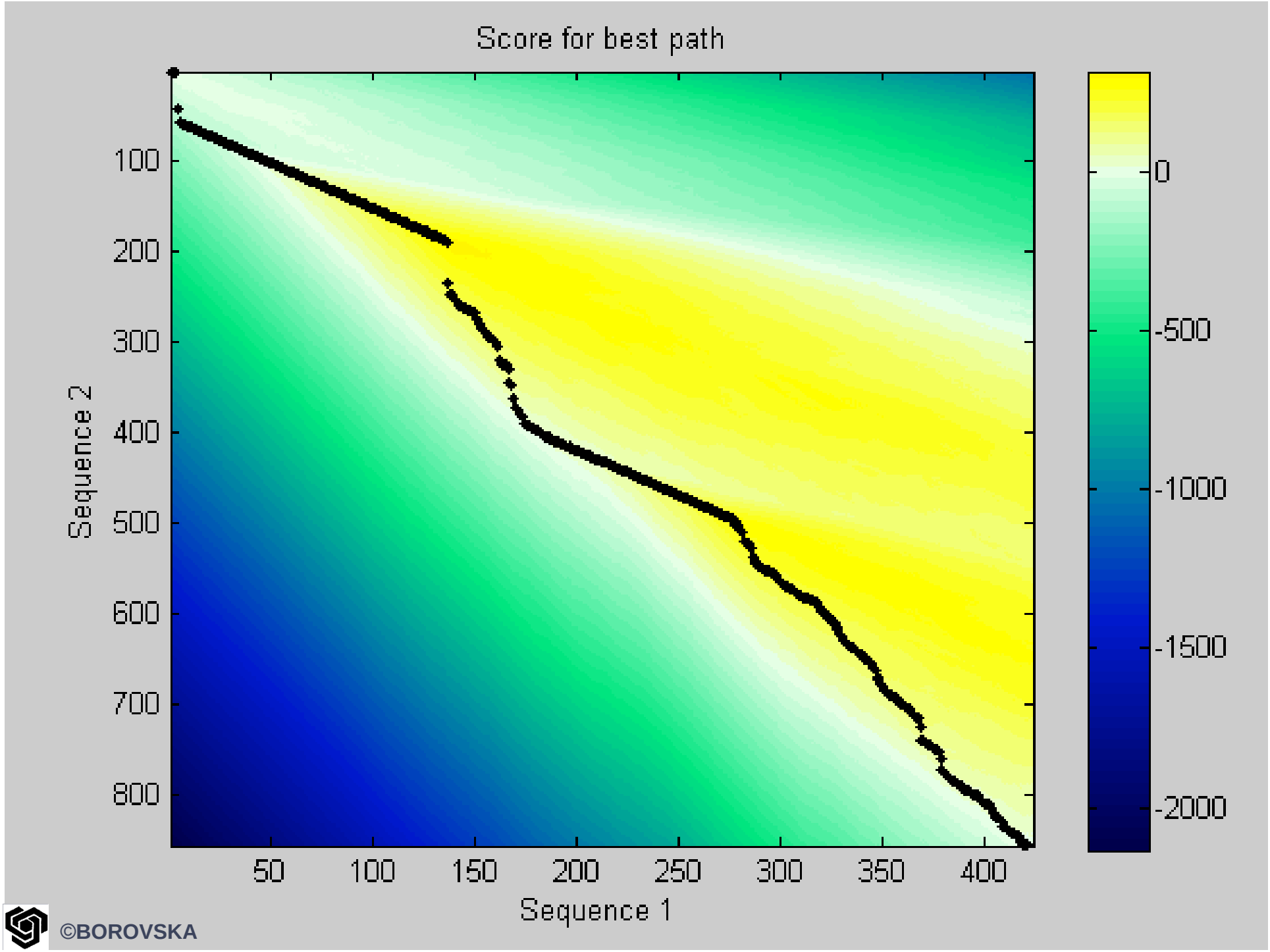
Code Input

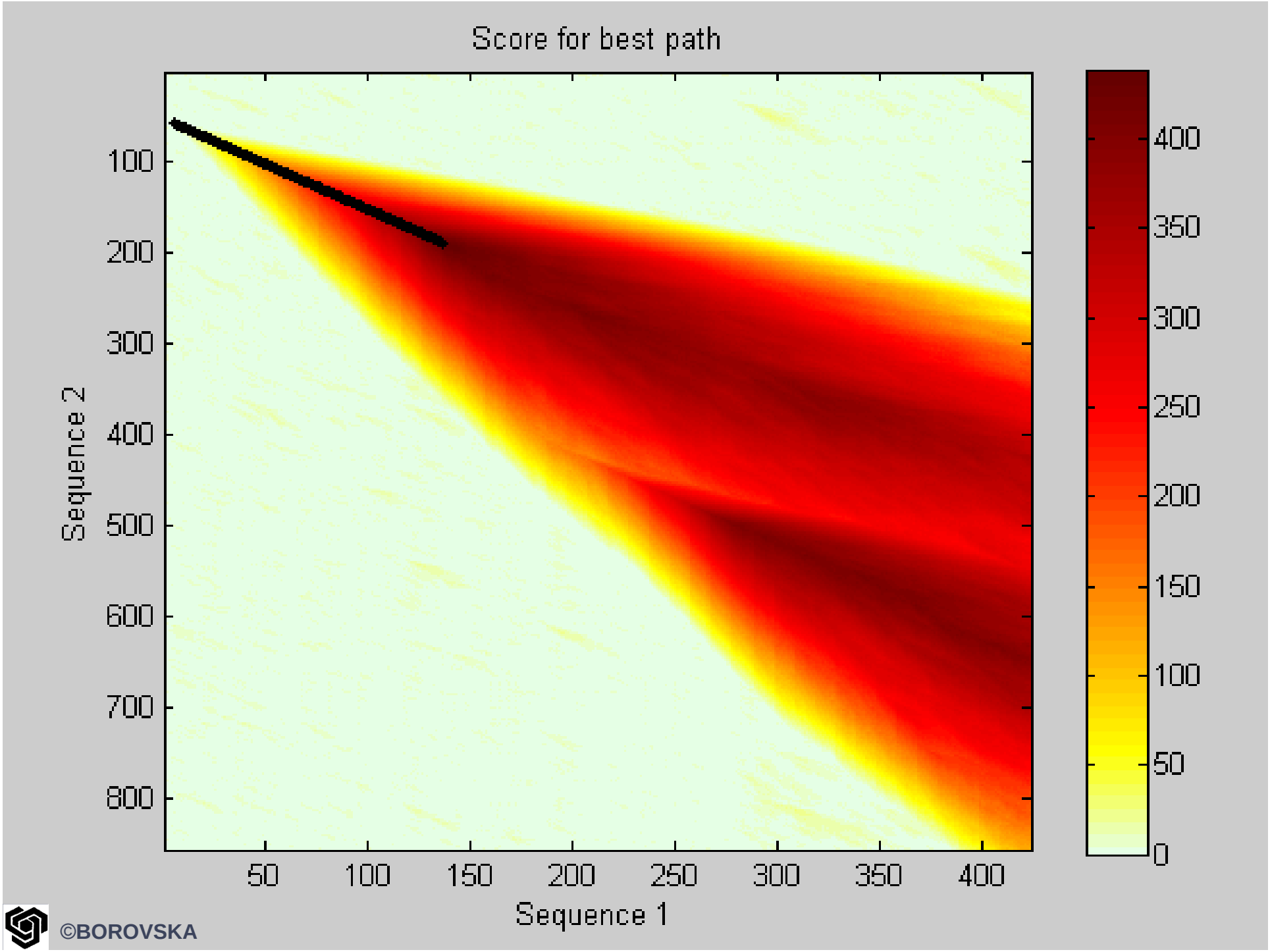
```
human = getgenPept('AAD01939','SequenceOnly',true);
```

Then look at the Drosophila protein (GenBank accession number X79493).

Code Input

```
fly = getgenPept('AAQ67266','SequenceOnly',true);
```


Score for best path
Sequence 2
100
200
300
400
500
600
700
800
50
100
150
200
250
300
350
400
Sequence 1
0
-500
-1000
-1500
-2000

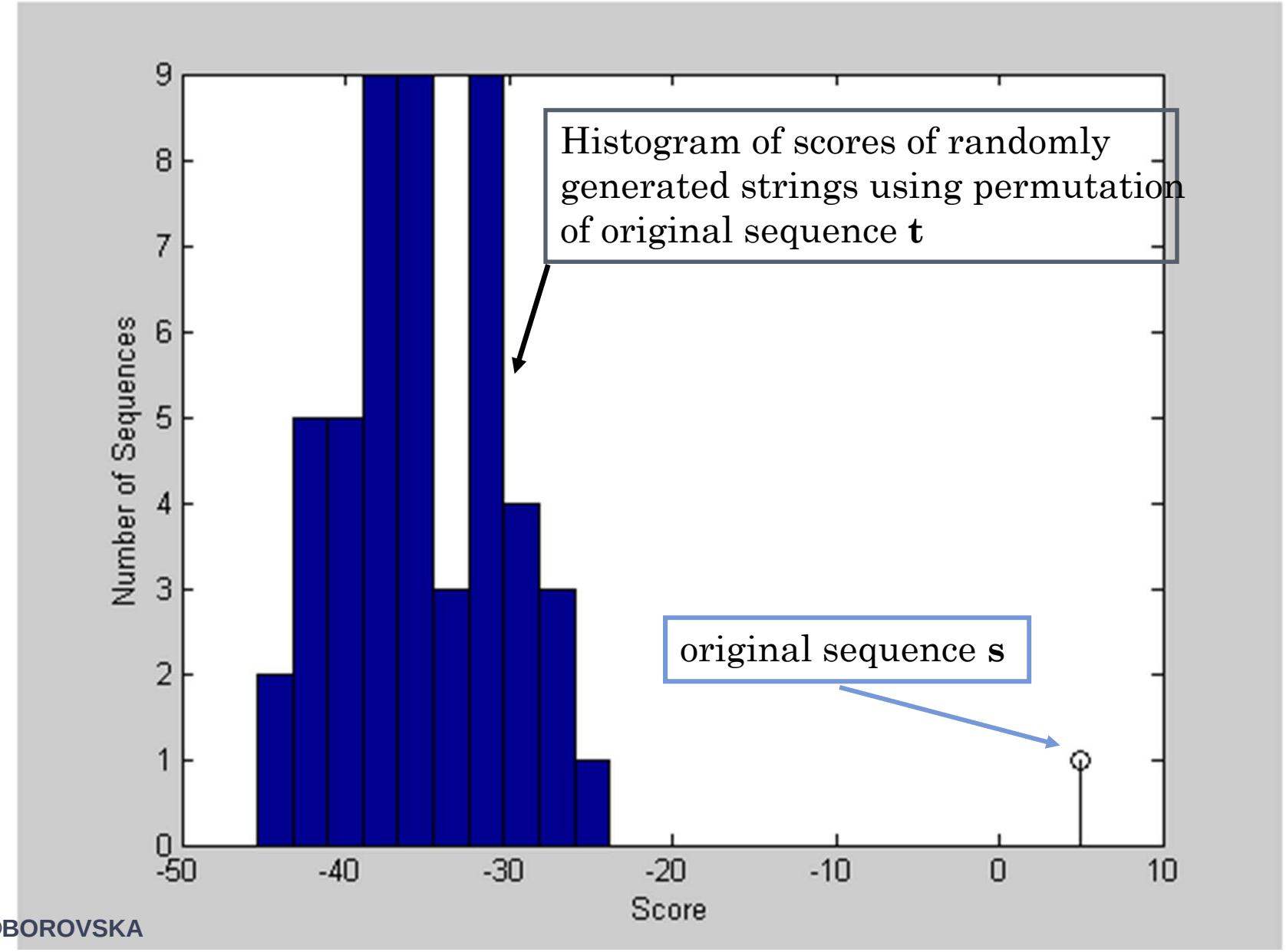

Score for best path
Sequence 2
Sequence 1
100
200
300
400
500
600
700
800
50
100
150
200
250
300
350
400
0
50
100
150
200
250
300
350
400
©BOROVSKA

Статистически анализ на подреждането

Аналогичен на откриването на гени:

- Генериране на рандомизирани секвенции на база втория стринг
- Определяне на оптималните подреждания на първата секвенция с рандомизираните секвенции
- Изчисляване на хистограма и ранкиране на оценките в хистограмата
- Относителната позиция дефинира т.нар. *p*-стойност.

Статистически анализ на подреждането



Използване на оценка за дистанция (разлики) при подреждането на секвенции (Distance Scores)

- Първоначално предложеният алгоритъм на динамичното програмиране от Needleman и Wunsch и Smith и Waterman се базира на сходството или идентичността на символите в секвенциите
- Алтернативен метод за оценка на подреждането на секвенциите се базира на разликите между секвенциите и символите в секвенциите – *колко промени са необходими за да трансформираме едната секвенция в другата*
- Този метод дава възможност да се прецени, че колкото е по-голяма дистанцията между секвенциите, толкова по-дълго еволюционно време е отнело развитието на секвенциите от общия предшественик.
- Следователно, метод с оценка на дистанцията осигурява адекватен биологичен критерий за сравнение на секвенциите, който е по-ефективен от биологична гледна точка, в сравнение с метода, базиран на сходствата.

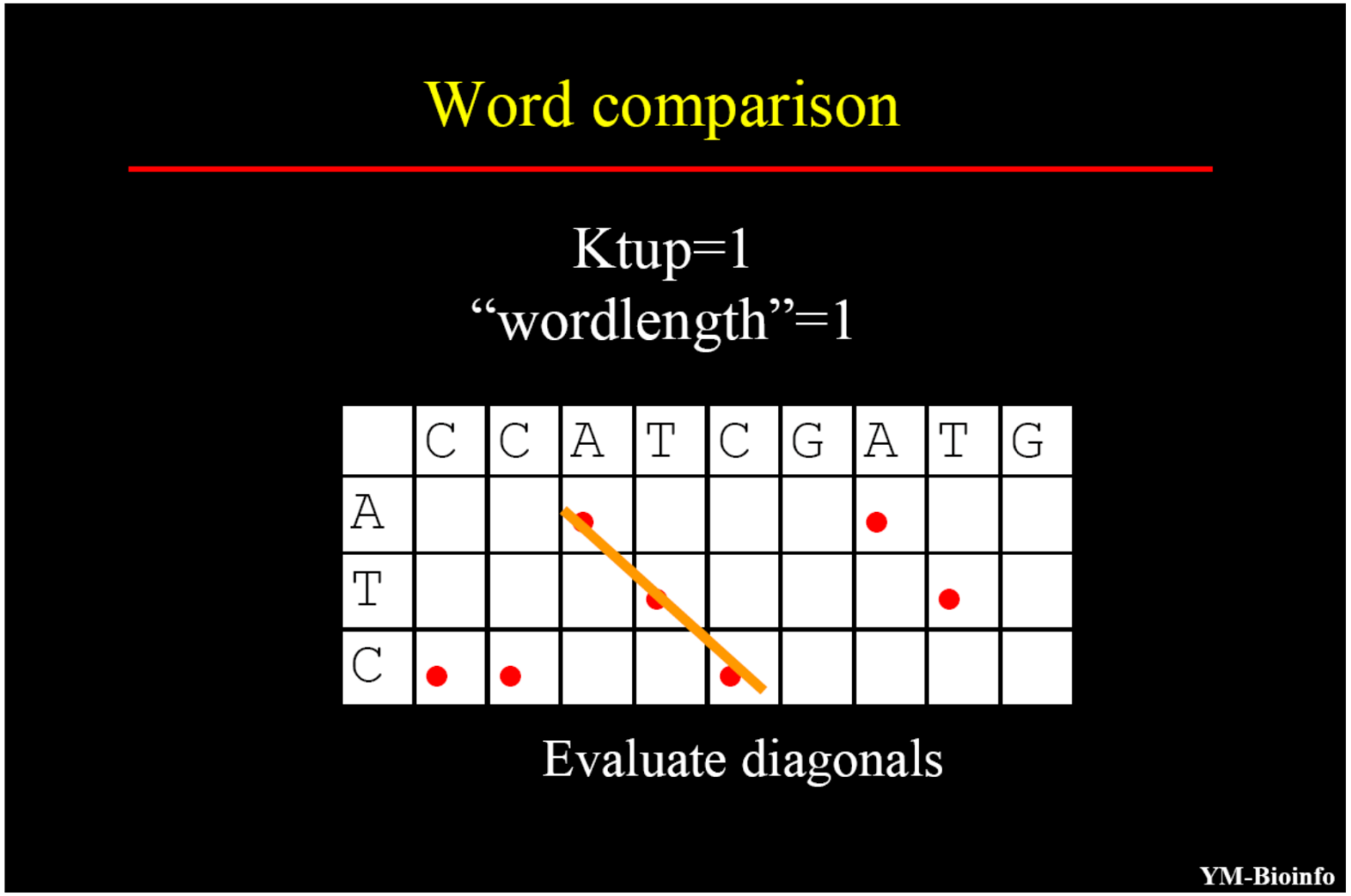
K-TUPLE METHODS
МЕТОДИ С ДУМИ

- Използват се от алгоритмите *FASTA и BLAST*
- Тези методи осигуряват бързо подреждане на секвенциите, като търсенето се осъществява по множество подредени символи, наричани думи или к-торки (words or *k-tuples) и впоследствие съединявайки тези думи посредством подреждане по метода на динамичното програмиране.*
- Тези методи са достатъчно бързи за да осигурят ефективно търсене в цели бази данни на секвенции, които най-добре съвпадат с входната тестова секвенция
- Методите FASTA и BLAST са евристични – емпирични методи при компютърното програмиране, при които "правилата на палеца" (практически правила) се прилагат за намиране на решения, като се използват обратни връзки за повишаване на производителността
- Тези методи са надеждни в аспекта на статистиката, и в общия случай осигуряват надеждно подреждане.

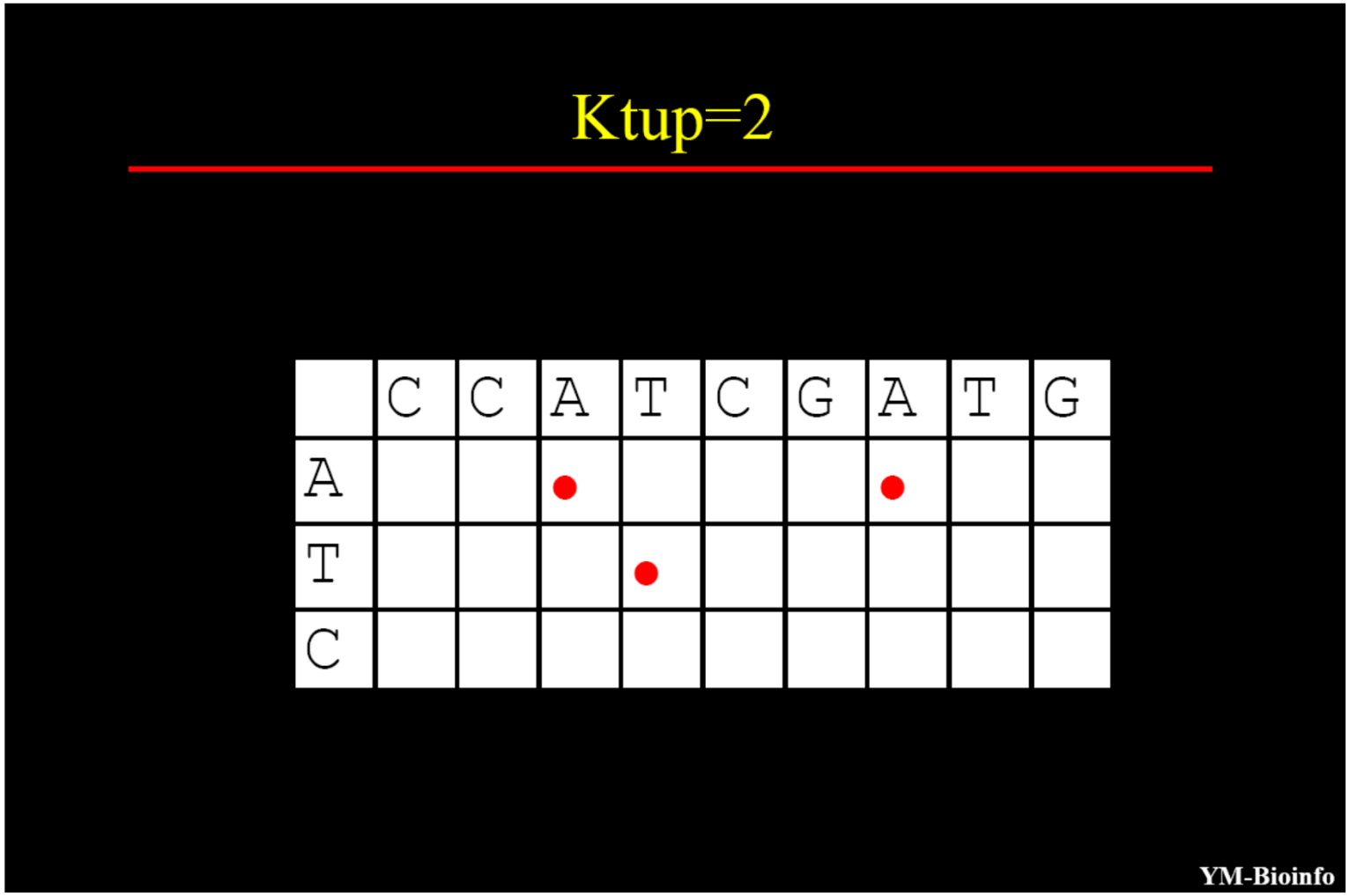
K-TUPLE METHODS
МЕТОДИ С ДУМИ

- Определя се размера на прозореца (window size, w)
- Едновременно се сравняват няколко символа (остатъка)
- Оценява се броя на сравненията между w двойки от остатъците

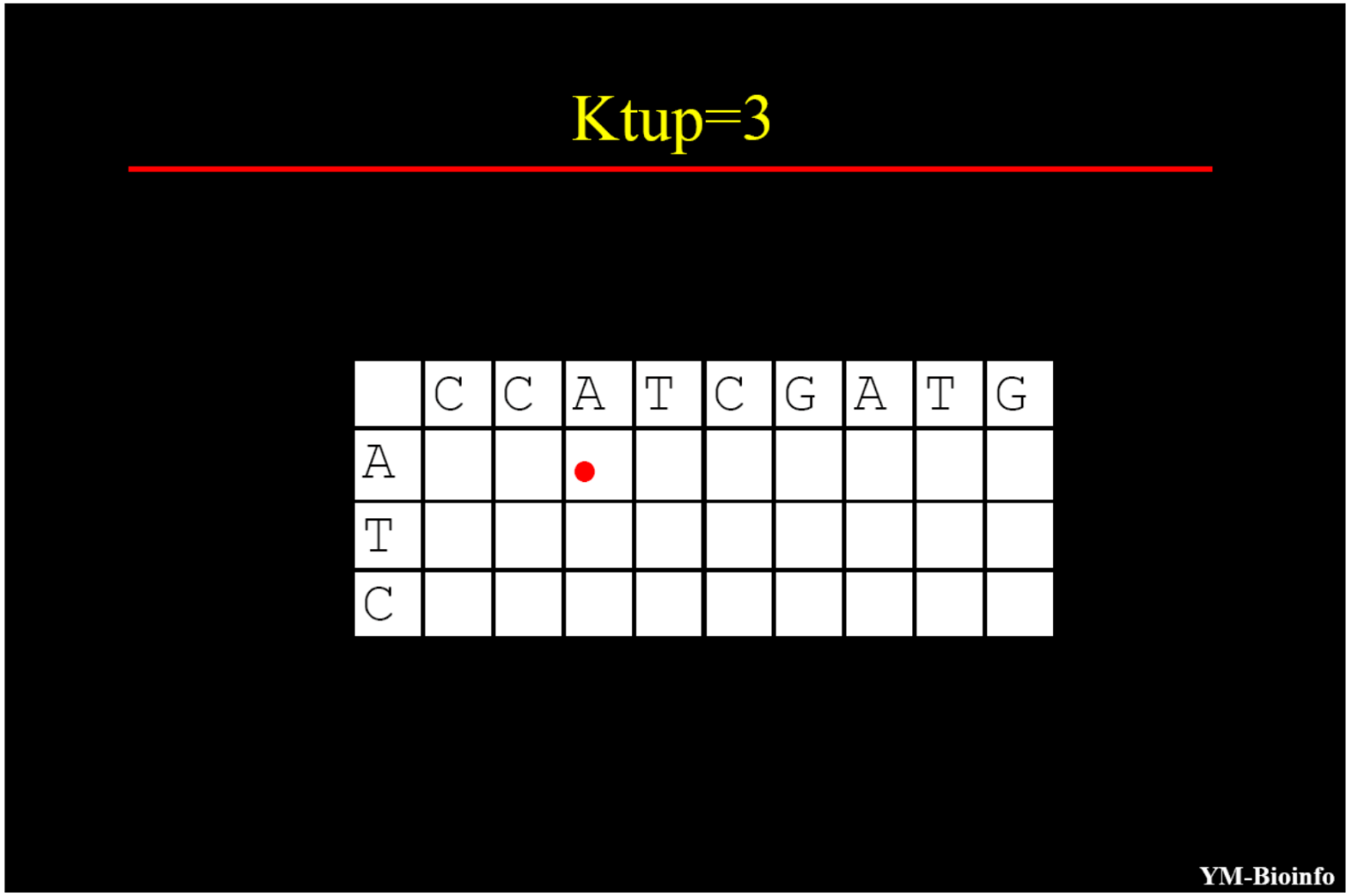
Сравнение на думи



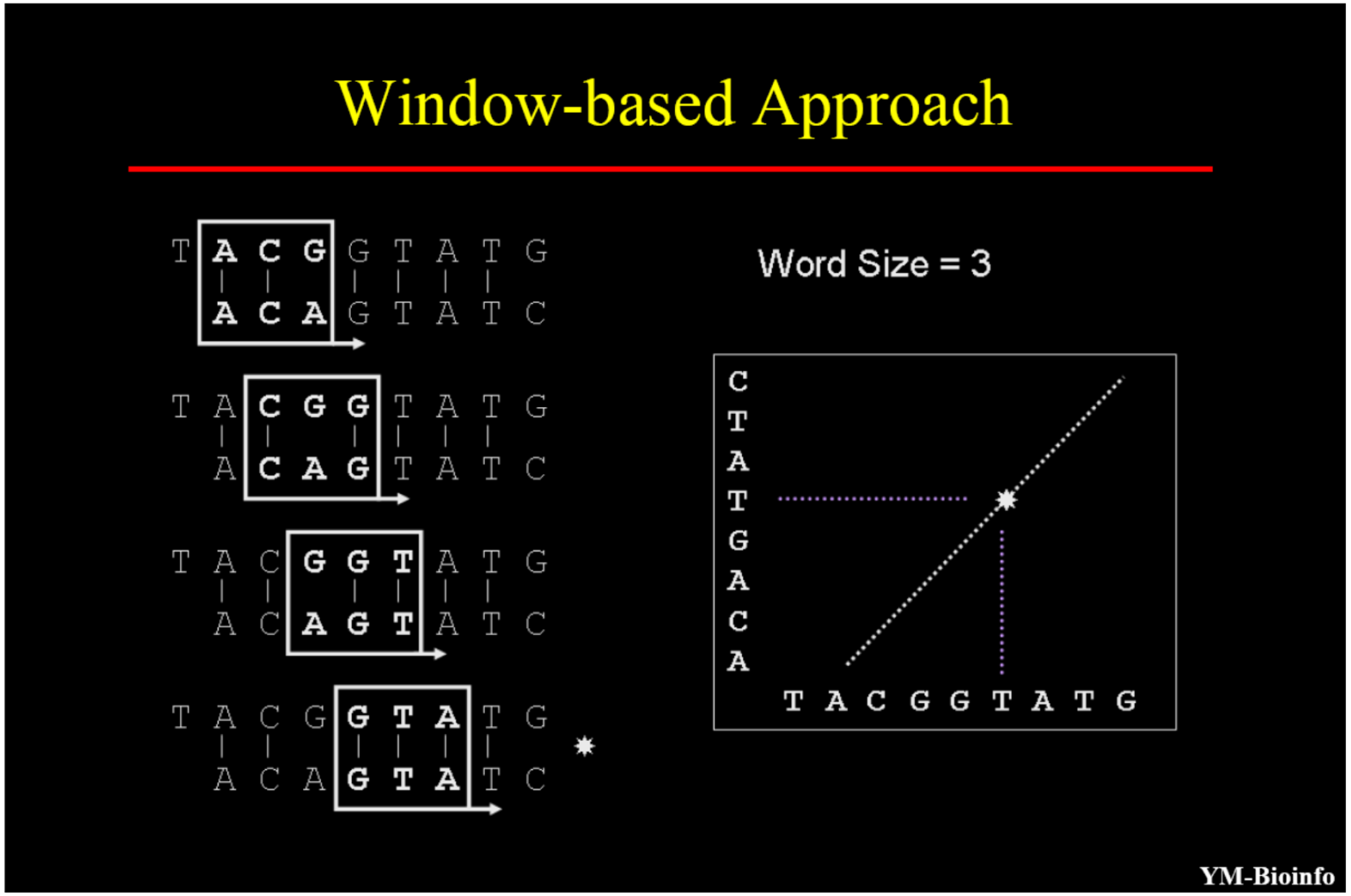
Сравнение на думи



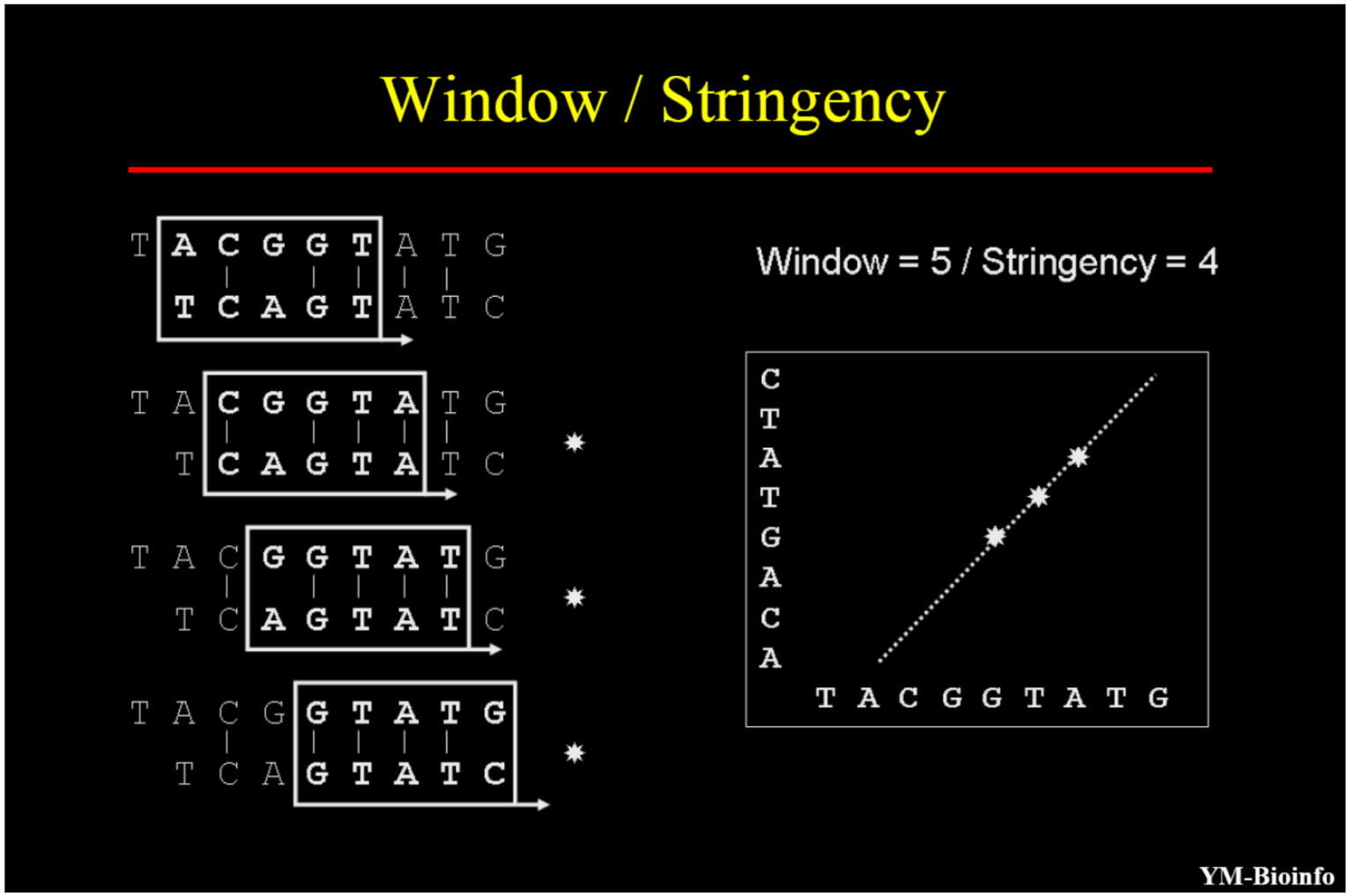
Сравнение на думи



Търсене с прозорци



WINDOW & STRINGENCY
ПРОЗОРЦИ И КОЕФИЦИЕНТ НА СЪВПАДЕНИЕ В ПРОЗОРЕЦА



Алгоритъмът FASTA

- *Алгоритъмът FASTA най-често се използва за търсене в бази данни на биологични секвенции*, като осигурява метод, алтернативен на този на динамичното програмиране (ДП) за подреждане на биологични секвенции.
- Локализират се късите еднакви участъците в двете секвенции.
- Като стартова точка за подреждане по алгоритъма на ДП се използва поредица от множество съвпадащи участъци, които се наблюдават в един и същи ред и в двете секвенции
- По-старите версии на FASTA осъществяваха глобално подреждане, но по принцип съвременните версии правят локално подреждане със статистически оценки на резултатите.

Алгоритъмът FASTA

- Програмата PLFASTA в състава на пакета FASTA създава диаграма на участъците с най-висок процент на съвпадение, аналогично на метода на точковата матрица, и по този начин осигурява възможност за генериране на алтернативни подреждания.
- *FASTA suite* се поддържа от Genestream на http://vega.igh.cnrs.fr/. Програмите включват *ALIGN (global, Needleman-Wunsch alignment), LALIGN (local, Smith-Waterman alignment), LALIGNO (Smith-Waterman alignment, no end gap penalty), FASTA (local alignment, FASTA method), and PRSS* (локално подреждане с разбъркани копия (scrambled copies) на втората секвенция за статистически анализ).

Оценка на качеството на подреждането със санкции за празни позиции (gap penalty)

```
sequence 1   V   D   S   -   C   Y
sequence 2   V   E   S   L   C   Y
SCORE        4   2   4 -11   9   7   SCORE = SUM OF AMINO ACID PAIR SCORES
(26)                                 MINUS SINGLE GAP PENALTY (11) = 15
```

- Празни позиции се вмъкват при подреждането по такъв начин, че еднаквите или подобни киселини да се позиционират една под друга, съответно, в двете секвенции.
- В най-добрия случай, при подреждането на сходни секвенции, се получават дълги участъци от идентични или сходни двойки аминокиселини и много малко празнини.
- При много различни секвенции (голяма дистанция), подреждането обхваща големи участъци от различни символи и голям брой празнини.

Оценъчни матрици при подреждането на секвенции от аминокиселини

- Матриците Dayhoff PAM се базират на еволюционен модел на измененията на протеина
- Матриците BLOSUM са предназначени за откриване на протеини, които принадлежат на едно и също семейство протеини
- Повечето програми за подреждане се предлагат с препоръчителни оценъчни матрици и санкции за празнини, които са полезни в повечето случаи
- При някои по-нови методи (като напр., Bayesian methods) се използват едновременно множество оценъчни матрици и санкции за празнини за генерирането на най-качествените подреждания на секвенциите

Оценъчни матрици Dayhoff PAM250 и BLOSUM62 за заместванията на аминокиселините при еволюцията на протеините (substitution matrices)

- Съдържат положителни и отрицателни стойности, отразявайки степента на сходството на всяка аминокиселина в протеините.
- Високата оценка на дадено подреждане на секвенциите показва висока степен на сходство на сравняваните протеини.
- Стойностите на оценките за наличието на празнини в сравняваните протеини са отрицателни, защото по същество те отразяват различия в двата протеина

PAM Matrix: *Point Accepted Mutation (точкови мутации)*

Използва "еволюционен модел", базиращ се на установени различия в тясно свързани протеини

- Моделът включва дефинирана скорост за всеки тип изменение на секвенцията
- Суфиксът (*n*) отразява изминалото "време":

  Скорост на прогнозираната мутация ако (if) n% от аминокиселините са се променили
- Така напр., матрицата *PAM1 оценява очакваната* скорост на заместванията ако *1% от аминокиселините в състава на протеина са се променили*
- PAM1 се използва като база за създаването на други матрици като се добавя възможността за изследване на възникването на множество замествания в едни и същи позиции на секвенциите
- PAM1 - използва се за сходни секвенции (по-кратко еволюционно време)
- PAM250 – за различаващи се секвенции (по-дълго еволюционно време)

40

BLOSUM: *BLOCK SUBSTITUTION MATRIX*

Базира се на % заместванията, наблюдавани при блокове от консервативни участъци в секвенциите на протеини с различно еволюционно развитие (в BLOCKS бази данни)

- Не се основава на специфичен еволюционен модел
- Суфиксът (*n*) отразява очакваното сходство:

  Среден процент (avg %) на сходство при множественото подреждане на секвенциите, от което е генерирана матрицата
- *Матрицата BLOSUM62* се генерира от подреждания на секвенции на протеините при установено сходство на по-малко от *62%*
- Базите данни Blocks съдържат *подредени сегменти без празнини*, съответстващи на най-консервативните участъци на протеините *(с най-малко изменения)*

- BLOSUM45 – за различаващи се секвенции
- BLOSUM62 – за сходни секвенции

41

Оценъчни матрици или матрици на заместванията

PAM & BLOSUM

- ***PAM = Point Accepted Mutation***

  използва “еволщционен модел" базиран на открити различия при подреждането на тясно свързани протеини
- ***BLOSUM = BLOck SUbstitution Matrix***

  базира се на % замествания на аминокиселини, в блокове от консервативни участъци в секвенциите на протеини с различно еволюционно развитие

Web sites за подреждане на двойки секвенции

- Bayes block aligner
  http://www.wadsworth.org/res&res/bioinfo
- BCM Search Launcher:Pairwise sequence alignment

http://dot.imgen.bcm.tmc.edu:9331/seq-search/alignment.html

- SIM—Local similarity program for finding alternative alignments

http://www.expasy.ch/tools/sim.html

- Global alignment programs (GAP, NAP)

http://genome.cs.mtu.edu/align/align.html

- FASTA program suite

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

- BLAST 2 sequence alignment

http://www.ncbi.nlm.nih.gov/gorf/bl2.html

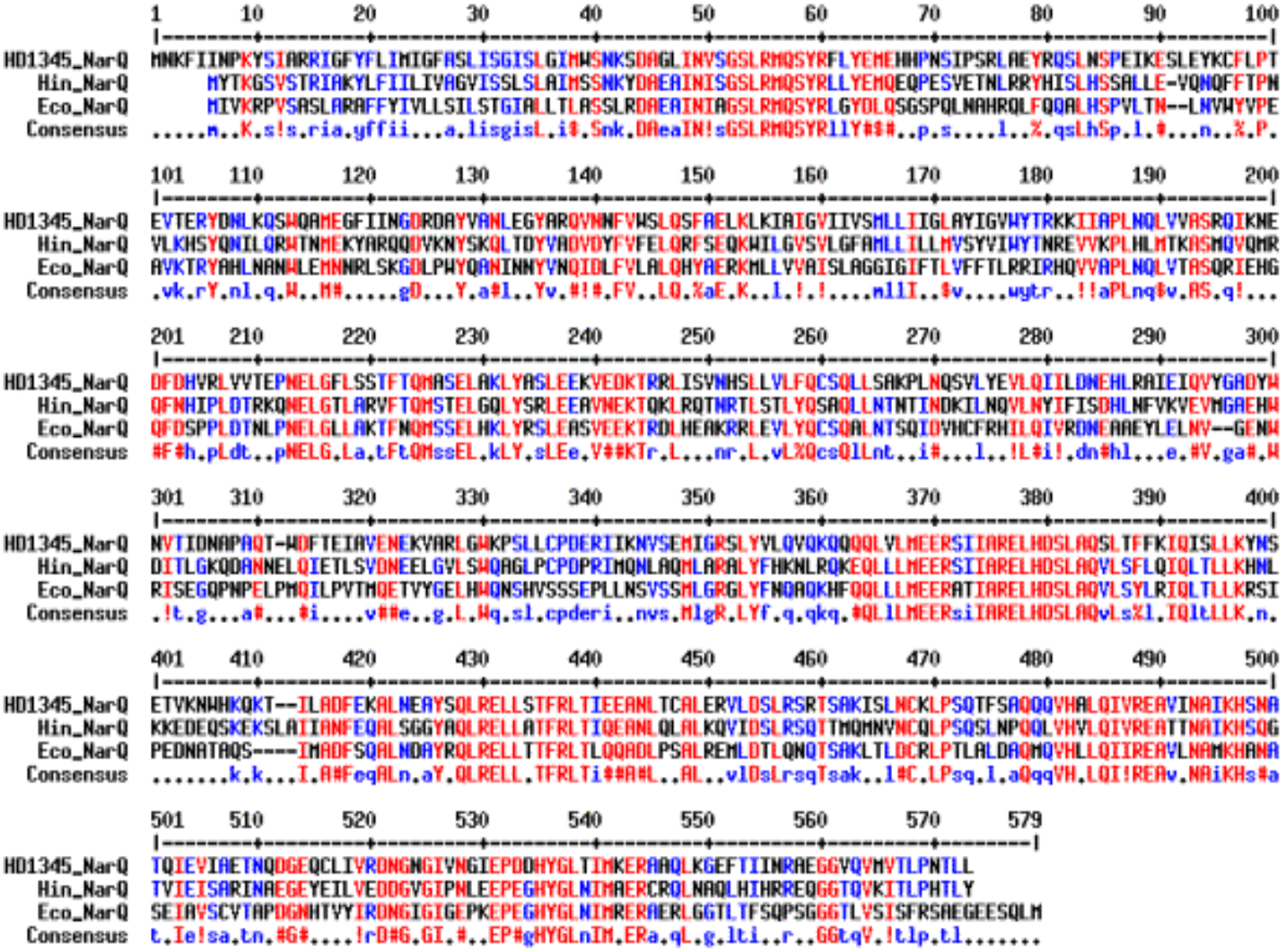
- Likelihood-weighted sequence alignment lwa

http://www.ibc.wustl.edu/servive/lwa.html

Множествено подреждане

- Множественото подреждане е разширение на подреждането по двойки с цел да се осигури сравнението на повече от две секвенции
- Методите за множествено подреждане обхващат всички секвенции в дадено множество
- Най-популярният софтуерен инструмент за множествено подреждане е **CLUSTAL**.
- Множественото подреждане на секвенции се класифицира като изчислителен проблем с NP-сложност.

MULTIPLE ALIGNMENT



Редактори и форматиращи програми за множествено подреждане на секвенции

- След получаването на крайната конфигурация на множественото подреждане (msa program), може да се наложи подредените секвенции да се редактират ръчно за подобряване на подреждането.
- Фактори, които се вземат предвид при избора на редактор на секвенциите (sequence editor), който е препоръчително да съдържа колкото се може повече от долуизброените свойства:
- (1) ясно и цветно представяне на символите на секвенцията
- (2) разпознаване на формата на представянето на секвенцията като изход от msa програмата и поддържане на адекватен формат до завършването на редактирането
- (3) подходящ и удобен потребителски интерфейс, осигуряващ възможност за прибавяне, изтриване и местене на символи и/или части от секвенциите и актуализиране на представянето на секвенциите след редактиране на подреждането.
- Осигуряване на допълнителни възможности като напр., като маркиране на консервирани участъци в подредените секвенции

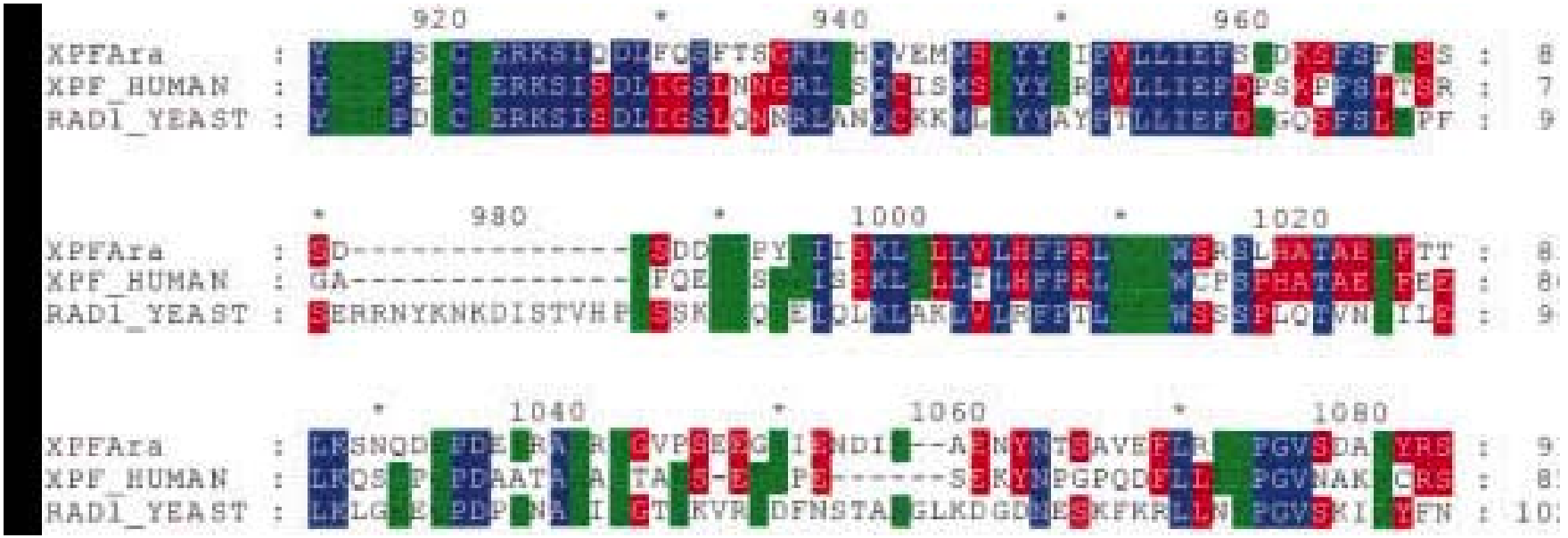
Формати за множествено подреждане на секвенции

- Два от най-популярните формати са Genetics Computer Group's MSF format и CLUSTALW ALN format. Because these formats follow a precise outline, one may be readily converted to another by computer programs.
- Софтуерът READSEQ (D.G.Gilbert от Indiana University, Bloomington може да се изпълни на почти всяка компютърна платформа (FTP - ftp.bio.indiana.edu/molbio/readseq)
- Web-базиран интерфейс за READSEQ от Baylor College of Medicine http://dot.imgen.bcm.tmc.edu:9331/seq-util/sequtil.html/
- Софтуерният пакет SEQIO осигурява програмни модули на C за конвертиране на файлове от секвенции от един формат във друг - FTP - ftp.pasteur.fr/pub/GenSoft/unix/programming/seqio-1.2.tar.gz; документация - http://bioweb.pasteur.fr/docs/doc-gensoft/seqio/.

Редактори на секвенции

- CINEMA (Color Interactive Editor for Multiple Alignments) – програма с богата функционалност за редактиране и анализ на секвенции, вкл. Анализ с точкова матрица – аплет, който се изпълнява под Web browser и следователно, на всяка компютърна платформа
- GDE (Genetic Data Environment) – осигурява общ интерфейс към UNIX-базирани компютри за анализ на секвенции, редактиране на подреждане на секвенции, и дисплей (Smith et al. 1994) – достъпна от няколко FTP сайта
- GeneDoc - редактиране на подреждане на секвенции, и дисплей (K. Nicholas и H. Nicholas от Pittsburgh Supercomputing Center)
- MACAW – софтуер за множествено подреждане на секвенции и софтуерен инструмент за редактиране на секвенции (Schuler et al. 1991). В зададено множество от секвенции, програмата намира блокове без празнини в секвенциите и изчислява статистически тяхното значение

GeneDoc



Форматиращи програми за секвенции
Sequence formatters

- Boxshade е форматираща програма (автор - K. Hofmann) за маркиране на идентични или сходни участъци на секвенциите при множествено подреждане (msas) с щриховане, достъпен от FTP -

http://www.isrec.isb-sib.ch/sib-isrec/boxshade

- Web сървър

http://www.ch.embnet.org/software/BOX_form.html
приема файл с множествено подреждане във формат Genetics Computer Group MSF или формат CLUSTAL ALN, при различни опции за формата на изходния файл

Форматиращи програми за секвенции
Sequence formatters

- CLUSTALX е софтуерен инструмент за форматиране на секвенции, осигуряващ Windows интерфейс за CLUSTALW msa за широка гама компютърни платформи
- ftp-igbmc.u-strasbg.fr/pub/ClustalX/