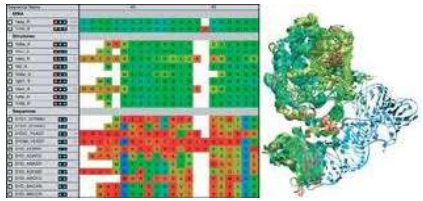
TIN SINH HỌC ĐẠI CƯƠNG
(Introduction to Bioinformatics)

PGS.TS. Trần Văn Lăng
Email: langtv@vast.vn

1



Chương 3:

BẮT CẶP TRÌNH TỰ (SEQUENCE ALIGNMENT)

2

NỘI DUNG

- Giới thiệu
- Bắt cặp hai trình tự
- Bắt cặp nhiều trình tự

3

MỘT SỐ KHÁI NIỆM CHUNG

4

Nhắc lại

- Sinh vật được tạo thành từ tế bào.
- Bên trong mỗi tế bào - ngoại trừ hồng huyết cầu trưởng thành - có nhân (nucleus) chứa tất cả các chỉ thị di truyền (genetic instruction)
- Những chỉ thị này là chức năng của tế bào

- Các tế bào, với các ngăn khác nhau của nó gọi là bào quan, phải đối mặt với một vấn đề là:
  - Tế bào sản xuất các phân tử như kích thích tố, dẫn truyền thần kinh, các cytokine và enzyme
  - Chúng phải được gửi đến nơi khác bên trong tế bào, hoặc xuất ra khỏi tế bào.
  - Việc sản xuất và vận chuyển này phải được thực hiện đúng nơi và đúng lúc.

- Chẳng hạn, mỗi tế bào người có 46 nhiễm sắc thể, được tổ chức thành 23 cặp.
- Mỗi nhiễm sắc thể được cấu thành bởi một trình tự DNA
- DNA phiên mã thành RNA, rồi được dịch mã thành Protein. Từ đó điều chỉnh tất cả các quá trình phát triển của một sinh vật

- Một gene là một đoạn của DNA với trình tự base đặc trưng – cụ thể, gọi là mã di truyền (genetic code), hay chỉ thị di truyền để xác định chức năng của tế bào

Chẳng hạn, về ung thư

- Cơ thể chúng ta liên tục sản xuất ra các tế bào để giúp chúng ta phát triển, để thay thế những tế bào đã chết, hoặc hàn gắn lại những tế bào bị tổn thương sau một chấn thương.
- Có các gene kiểm soát quá trình này; khi những gene này bị gây tổn hại dẫn đến bệnh tật, đặc biệt là ung thư.

- Việc tổn hại này thường xảy ra trong quá trình sinh sống và phát triển của đời người.
- Bên cạnh đó cũng có một số nhỏ thừa hưởng những gene tổn hại này từ cha mẹ.
- Bình thường các tế bào phát triển và nhân lên theo một quy tắc. Tuy nhiên, những gene đã bị tổn hại có thể phát triển không bình thường - phát triển thành những khối gọi là u hay là bướu.

- Những khối u có thể lành tính (không phải ung thư) hoặc ác tính (ung thư).
- Những khối u lành tính không xâm lấn vào các cơ quan và các mô xung quanh của cơ thể. Trong khi đó khối u ác tính sẽ phát triển, có thể xâm lấn vào những mô xung quanh, trở thành ung thư.

Như vậy,

- Ung thư xuất phát từ một tế bào đơn lẻ, sau đó chuyển dạng tế bào bình thường thành tế bào ung thư. Việc này có thể do:
  - Tác nhân vật lý: tia cực tím, bức xạ ion
  - Tác nhân hóa học: như
    Asbestos: các thành phần của khói thuốc lá,
    Aflatoxin: một chất nhiễm bẩn thức ăn
    Arsenic: một chất nhiễm bẩn nước uống
  - Tác nhân sinh học: nhiễm trùng từ virus, vi khuẩn, ký sinh trùng.

- Trong quá trình sinh trưởng và phát triển, các tế bào được thay thế bằng những tế bào mới nhờ quá trình phân bào.

  Ngoại trừ tế bào não và tế bào thần kinh là không được sinh sản thêm ra mà chỉ mất đi.

- Quá trình phân bào phải sao chép và truyền đạt lại toàn bộ và chính xác các thông tin di truyền cho các tế bào con.
- Nên DNA của tế bào mẹ phải được nhân đôi một cách chính xác sao cho mỗi tế bào con đều nhận được DNA giống tế bào mẹ

- Các tác nhân từ ngoài tác động gây ra những tế bào có DNA "bất thường". Nhưng cơ thể con người là một bộ phận để phản ứng với tình trạng này bằng cách:
  - Các tế bào có cơ chế để sửa chữa những DNA bị hư trước khi chúng gây ra một vấn đề nào đó. Hoặc,
  - Hệ thống miễn dịch nhận ra được những tế bào bất thường và tiêu diệt đi.

- Cơ chế làm cho hầu hết các tế bào có DNA bất thường (gọi là các tế bào "tiền ung thư") đều chết trước khi có thể gây ra ung thư.
- Chỉ có một số ít không bị diệt (lọt lưới) để rồi gây ra ung thư

- Các tế bào ung thư được lưu hành trong hệ bạch huyết và máu và tìm đến bộ phận nào yếu nhất trong cơ thể để lưu lại đó.
- Khi đó tế bào ung thư tiếp tục quá trình phân bào để tạo thành khối u ung thư.

- Bắt cặp trình tự, hay là sắp xếp thẳng hàng trình tự (Sequence Alignment)
- Mục đích đạt đến sự giống nhau đến mức tối đa của các trình tự
- Việc bắt cặp được thực hiện bằng cách thêm các “gap” vào các vị trí có thể sao cho các cột giống nhau hoặc tương tự nhau

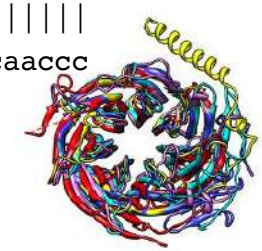
- Việc thêm các gap biểu thị sự đột biến mất nucliotide đã xãy ra tại vị trì này trên trình tự.
- Trong tin học, việc thêm ký tự gap là khoảng trống (“-”) giúp cho việc tạo ra 2 chuỗi ký tự gần giống nhau nhất.

Ví dụ

- Ví dụ bắt cặp 2 trình tự
  - GAATTCAGTTA
  - GGATCGA
- Kết quả
  - GAATTCAGTTA
  - | || | | |
  - GGAT-C-G--A

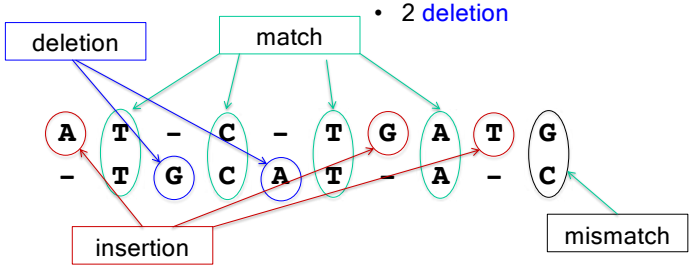
- Hoặc 2 trình tự
  - ACGCTG
  - CATGT
- Kết quả
  - ACGCTG-
  - - | | |
  - --C-ATGT

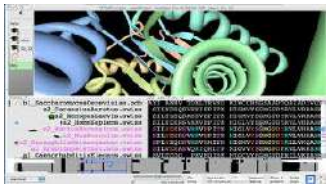
- Tương tự, với 2 trình tự dài hơn
  - tcctctgcctctgccatcat---caaccc
  - |||| ||| ||||| ||||| ||||||
  - tcctgtgcatctgcaatcatgggcaaccc



Cho 2 trình tự:
u = **ATCTGATG**
v = **TGCATAC**

Nếu lấy v làm căn cứ, thì u có:
- 4 match
- 1 mismatch
- 3 insertion
- 2 deletion





- Mục đích
  - Bắt cặp trình tự nhằm nghiên cứu về các trình tự
  - Hoặc để tìm kiếm, so sánh mức độ tương đồng giữa các trình tự

- Sự bắt cặp trình tự không chỉ dừng lại ở trình tự DNA mà cả trình tự protein.
- Trong đó, việc chỉ có 4 ký tự được thay bởi 20 ký tự.
- Tuy nhiên, do protein có đặc điểm bảo tồn cấu trúc và chức năng cao (bởi nếu mất chức năng sẽ gây bất lợi)

- Vì vậy, trong qua trình biến đổi có khuynh hướng chỉ thay thế các amino acid có cấu trúc tương tự, ít làm thay đổi đến cấu trúc và chức năng protein
- Những trình tự protein trong cùng một họ thường có sự thay thế giữa các amino acid có cùng đặc tính hóa lý.

- Thế nào là sự bắt cặp tốt, tiêu chuẩn nào.
- Có thể cho điểm tốt đối với giá trị match, điểm xấu với các trường hợp ngược lại.
- Tuy nhiên, với trình tự protein việc thay thế một amino acid khác vẫn bảo toàn cấu trúc và chức năng cũng không thể là điểm xấu

- Chính vì vậy, với việc bắt cặp trình tự protein có các ma trận điểm thay thế để xem xét khả năng thay thế amino acid mà không ảnh hưởng này.
- Có 2 loại ma trận điểm thay thế:
  - Ma trận PAM (Percentage Accepted Mutation)
  - Ma trận BLOSUM (BLOck SUbstitution Matrix)

- Sự tương tự giữa PAM và BLOSUM:
  - PAM100 ~ BLOSUM90
  - PAM160 ~ BLOSUM62
  - PAM250 ~ BLOSUM45
- PAM được tạo ra từ khoảng cách trong các trình tự liên quan.
  - Chẳng hạn, PAM100 có khoảng cách 100 lần đột biến trên 100 gốc amino acid

- BLOSUM được tính toán thông qua tần suất thay thế của các cặp amino acid trong việc bắt cặp các trình tự có độ tương đồng cao.
  - Chẳng hạn, BLOSUM45 gồm các nhóm trình tự giống nhau 45%

Bảng giá trị của BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9	-1	1	0	2	1	1	2	1	2	0	0	2	4	1	5	1	2	-2	5	C
S	-1	4	0	-2	0	-1	0	0	0	1	0	0	0	1	0	1	-1	1	1	-1	S
T	-1	1	5	-2	-1	-1	0	0	0	0	0	0	-1	0	-1	1	0	1	1	3	T
P	-3	-1	-1	7	2	-1	-2	-2	-1	0	0	1	1	0	0	1	0	1	1	2	P
A	0	1	0	-1	4	0	-1	-1	-2	0	1	1	0	0	-1	0	-1	1	2	4	A
G	-3	0	-2	-2	0	6	0	-1	-1	-1	0	-1	0	0	0	0	-1	0	0	0	G
N	-3	1	0	-2	-2	0	6	2	-1	-1	0	0	2	2	1	-1	0	0	2	1	N
D	-3	0	-1	-1	-2	-1	1	6	1	0	0	-2	0	1	1	-1	0	0	1	3	D
E	-4	0	-1	-1	-1	-2	0	2	5	2	-1	0	1	0	1	0	-1	0	1	2	E
Q	-3	0	-1	-1	-1	-2	0	0	2	5	-1	-1	0	-1	-1	1	0	1	3	-4	Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8	1	-2	-1	-1	1	1	2	3	1	H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	-2	-1	-1	-1	0	1	2	4	R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-1	1	0	0	1	3	K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	-1	-1	0	-1	1	2	M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	0	0	1	2	4	I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	-1	-1	-2	1	L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	2	V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	-1	1	F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

- Đánh giá sự bắt cặp trình tự nucleotide: dùng hàm đánh giá.
- Chẳng hạn, nếu
  - Match (Giống nhau ở cùng vi trí): giá trị là +2
  - Mismatch (Không giống nhau): giá trị là -1
  - Gap (Thêm vào hoặc bị loại bỏ): giá trị là -2
- Hàm đánh giá có giá trị càng cao thì sự giống nhau càng nhiều.

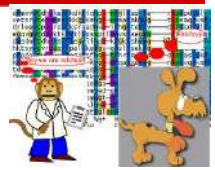
- **<u>Định nghĩa</u>**: Mức độ tương đồng (điểm đánh giá) của 2 trình tự bắt cặp S_1' và S_2' là đại lượng:

$$n_a \text{ x match} + n_i \text{ x mismatch} + n_g \text{ x gap}$$

- Trong đó,
  - n_a, n_i, n_g: tương ứng là số phần tử giống nhau (match), không giống nhau (mitmatch) và số gap.
  - match, mismatch, gap: tương ứng là giá trị tính toán để đánh giá.
  - Thông thường, điểm dương cho match, điểm âm cho sự đột biến (mismatch và gap)

Ví dụ

- Với
  - match = 2
  - mismatch = -1
  - gap = -2



```
GAATTCAGTTA
|  || |  |   |
GGAT-C-G--A
```

- Điểm đánh giá: 6 x (+2) + 4 x (-2) + 1 x (-1) = 3


```
AC--GCTG
 |   | |
-CATG-T-
```

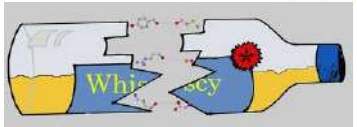
- Điểm đánh giá: 3 x (+2) + 5 x (-2) + 0 x (-1) = -4

```
ACGCTG-
 |   ||
-C-ATGT
```

- Điểm: 3 x (+2) + 3 x (-2) + 1 x (-1) = -1

```
tcctctgcctctgccatcat---caaccc
|||| ||| ||||| |||||   ||||||
tcctgtgcatctgcaatcatgggcaaccc
```

- Điểm: 23 x (+2) + 3 x (-2) + 3 x (-1) = 37



- Có 2 loại:
  - **<u>Bắt cặp toàn cục</u>** (Global alignment): được áp dụng trên toàn bộ trình tự để tìm sự tương đồng giữa các trình tự.
  - Thường được sử dụng khi 2 trình tự có độ tương đồng cao, chiều dài xấp xỉ nhau

    Thuật toán sử dụng: Needleman - Wunsch

  - **<u>Bắt cặp cục bộ</u>** (Local alignment): chỉ thực hiện trên một vùng trình tự con tương đồng nằm ở các vị trí khác nhau trên hai trình tự.
  - Mục đích tìm ra vùng trình tự tương đồng nhất.
  - Sử dụng khi so sánh 2 trình tự có chiều dài khác nhau, mức độ tương đồng trên toàn bộ là thấp.

    Thuật toán: Smith - Waterman

Bắt cặp hai trình tự

- **<u>Bài toán</u> (Pairwise Sequence Alignment - PSA)**: cho 2 trình tự sinh học S_1, S_2. Hãy tìm 2 trình tự S_1', S_2' bằng cách thêm các ký tự '-' sao cho:
  - Điểm đánh giá Score(S_1', S_2') là lớn nhất với giá trị match, mismatch và gap cho trước
  - Chiều dài S_1', S_2' là bằng nhau ($|S_1'| = |S_2'|$)
  - Nếu loại bỏ ký tự gap từ S_1', S_2' sẽ nhận được S_1, S_2 ban đầu

Ví dụ PSA

- S_1 = "ACGCTG"
- S_2 = "CATGT"

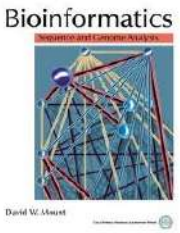
- S_1' = "-ACGCTG"
- S_2' = "CATG-T-"

Bắt cặp đa trình tự

- **<u>Bài toán</u> (Multiple Sequence Alignment - MSA)**: Cho k trình tự sinh học $S_1, S_2, \ldots, S_k$. Hãy tìm k trình tự $S_1', S_2', \ldots, S_k'$ bằng cách thêm các ký tự '-' sao cho:
  - Mức độ tương đồng của các trình tự này là cao nhất
  - $|S_1'| = |S_2'| = \ldots = |S_k'|$
  - Nếu loại bỏ ký tự gap từ $S_1', S_2', \ldots, S_k'$ sẽ nhận được $S_1, S_2, \ldots, S_2$ tương ứng

Ví dụ MSA

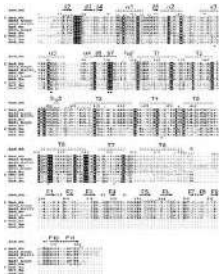
- Thông thường, bắt cặp đa trình tự được sử dụng khi cần tìm kiếm một trình tự đại diện trong tập hợp nhiều trình tự sinh học.



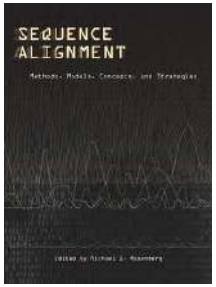
Ví dụ

- Với đoạn trình tự "ACTCGATT"
  - Mất T,C ở vị trí 3, 4: "ACGATT"
  - Đột biết vị trí 2 (thay C bằng G), vị trí 3 (thay G bằng C), vị trí 6 (thay T bằng C): "AGCATC"
  - Thêm TA vào vị trí 4: "AGCTAATC"
- Như vậy, từ "ACTCGATT" sẽ biến thành "AC--GATT", "A**G**--**C**AT**C**", "AG--C**TA**ATC"

- Như vậy, với 2 trình tự:
  - ACTCGATT
  - AGCTAATC
- có thể được bắt cặp là:
  - ACTCG--ATT
  - \| \|
  - AG—CTAATC



- Khi đó, ký tự "gap" vừa:
  - deletion gap: mất đi
  - insertion gap: thêm vào



Sử dụng

PHẦN MỀM CLUSTALX

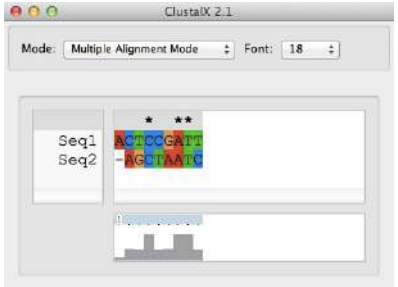
- Có thể tìm tại http://www.clustal.org/clustal2/

- Có hai dạng Clustal trên 3 hệ điều hành khác nhau: Linux, Mac OS X, Windows:
  - ClustalW: thực thi ở chế độ dòng lệnh
  - ClustalX: dùng ở chế độ khung của sổ (window)

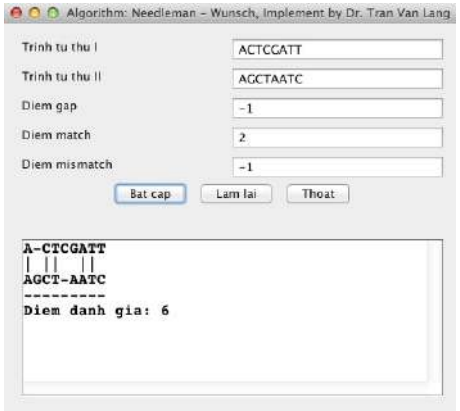
Ví dụ

- Để bắt cặp 2 trình tự,
  - tạo file dạng FASTA
  - chọn File/Load Sequences là file mới tạo
  - chọn Alignment/Do Complete Sequences

```
>Seq1
ACTCCGATT
>Seq2
AGCTAATC
```



Hoặc có thể viết một application



Thuật toán

NEEDLEMAN - WUNSCH

- Do Saul Needleman và Christian Wunsch đưa ra vào năm 1970
- Áp dụng trên toàn bộ trình tự để tìm sự tương đồng giữa toàn bộ 2 trình tự (bắt cặp toàn cục – Gobal Alignment)

- Cho 2 trình tự lần lượt có chiều dài là n và m
- Thuật toán gồm các bước sau:
  - **Bước 1**: Khởi tạo giá trị ban đầu cho ma trận tính toán M như sau:

$$M_{i0} = id, \quad \forall i = \overline{0,n}, \qquad M_{0j} = jd, \quad \forall j = \overline{0,m}$$

$$M_{ij} = \max\left\{M_{i-1,j-1} + \sigma_{ij}, M_{i-1,j} + d, M_{i,j-1} + d\right\},$$

$$\forall i = \overline{1,n}; j = \overline{1,m}$$

$$\sigma_{ij} = \begin{cases} \text{match} \\ \text{mismatch} \end{cases}, \qquad d = gap$$

- Ví dụ cho 2 trình tự
  - CATGT
  - ACGCTG
- Các giá trị tính toán
  - match = 2
  - mismatch = -1
  - d = -2
- Ma trận M như bảng

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8	−7	−3	1	−1

Giải thích kết quả

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1				
C	−4					
G	−6					
C	−8					
T	−10					
G	−12					

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0			
C	−4					
G	−6					
C	−8					
T	−10					
G	−12					

57

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2		
C	−4					
G	−6					
C	−8					
T	−10					
G	−12					

58

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	
C	−4					
G	−6					
C	−8					
T	−10					
G	−12					

59

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4					
G	−6					
C	−8					
T	−10					
G	−12					

60

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0				
G	−6					
C	−8					
T	−10					
G	−12					

61

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2			
G	−6					
C	−8					
T	−10					
G	−12					

62

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−2	
G	−6					
C	−8					
T	−10					
G	−12					

63

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	
G	−6					
C	−8					
T	−10					
G	−12					

64

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6					
C	−8					
T	−10					
G	−12					

65

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2				
C	−8					
T	−10					
G	−12					

66

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1			
C	−8					
T	−10					
G	−12					

67

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3		
C	−8					
T	−10					
G	−12					

68

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	
C	−8					
T	−10					
G	−12					

69

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8					
T	−10					
G	−12					

70

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4				
T	−10					
G	−12					

71

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3			
T	−10					
G	−12					

72

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2		
T	−10					
G	−12					

73

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	
T	−10					
G	−12					

74

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10					
G	−12					

75

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6				
G	−12					

76

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5			
G	−12					

77

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1		
G	−12					

78

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	
G	−12					

79

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12					

80

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8				

81

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8	−7			

82

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8	−7	−3	−3	

83

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8	−7	−3	1	

84

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8	−7	−3	1	−1

- **<u>Bước 2</u>**: Tìm vết dựa trên kết quả tính các giá trị của ma trận trước đó. Xuất phát từ M_{nm} nếu:

  – $M_{i-1,j-1} = M_{ij} - \sigma_{ij}$, theo duong chéo, (i,j) → (i-1,j-1)
  – $M_{i-1,j} = M_{ij} - d$, dich chuyen lên trên, (i,j) → (i-1,j)
  – $M_{i,j-1} = M_{ij} - d$, dich chuyen lui, (i,j) → (i,j-1)

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8	−7	−3	1	−1

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8	−7	−3	1	−1

- **<u>Bước 3</u>**: Bắt cặp trình tự
  - Xuất phát từ phần tử M_{nm}
  - Nếu phần tử kế nằm trên đường chéo: hai ký tự được bắt cặp với nhau
  - Nếu phần tử kế nằm bên trái: thêm gap cho trình tự thứ hai (ở dưới)
  - Ngược lại, nếu phần tử kế nằm ở trên: thêm gap cho trình tự thứ nhất

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8	−7	−3	1	−1

-CA-TGT
ACGCTG-

- Điểm đánh giá = 3x2 + 1x(-1) + 3(-2) = -1

CATG-T-
-ACGCTG

- Điểm: 3x2 + 1(-1) + 3(-2) = -1

		C	A	T	G	T
	0	−2	−4	−6	−8	−10
A	−2	−1	0	−2	−4	−6
C	−4	0	−2	−1	−3	−5
G	−6	−2	−1	−3	1	−1
C	−8	−4	−3	−2	−1	0
T	−10	−6	−5	−1	−3	1
G	−12	−8	−7	−3	1	−1

Xem xét ví dụ khác

- Xét 2 trình tự peptide như sau:
  - U = AlaCysGlyCysAspGly
  - V = CysAlaAspGlyAsp
- Gồm các amino acid: Alanin (A), Cystein (C), Glycine (G), Aspartic acid (D)

- Có thể biểu diễn
  - U = “ACGCDG”
  - V = “CADGD”

- Tạo ma trận đánh giá theo quy tắc:
  - $M_{00} = 0$
  - $M_{i0} = M_{i-1,0} + d$
  - $M_{0j} = M_{0,j-1} + d$
  - $M_{ij} = \text{Max}\ \{M_{i-1,j-1} + \sigma_{ij}, M_{i,j-1} + d, M_{i-1,j} + d\}$
  - $d = -1$
- Trong đó (giả sử không dung ma trận PAM, BLOSUM)
  - $\sigma_{ij} = +2$ nếu U_i và V_j giống nhau
  - $\sigma_{ij} = -1$ nếu U_i và V_j khác nhau

		C	A	D	G	D
	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
G	-3	0	0	-1	2	1
C	-4	-1	-1	-1	1	1
D	-5	-2	-2	1	0	3
G	-6	-3	-3	0	3	2

- Tìm vết bằng cách dùng d = -1 và ma trận σ để so sánh trên 2 trình tự:
- Xuất phát từ M_{65}, nếu:
  - $M_{ij} = M_{i-1,j-1} + \sigma_{ij}$ thì vết (i,j) → (i-1,j-1) đi theo đường chéo
  - $M_{ij} = M_{i,j-1} + d$ thì vết (i,j) → (i,j-1) đi lui
  - $M_{ij} = M_{i-1,j} + d$ thì vết (i,j) → (i-1,j) đi lên

- Trong trường hợp này, có nhiều vết được tạo ra (màu red, blue, green)

		C	A	D	G	D
	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
G	-3	0	0	-1	2	1
C	-4	-1	-1	-1	1	1
D	-5	-2	-2	1	0	3
G	-6	-3	-3	0	3	2

		C	A	D	G	D
	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
G	-3	0	0	-1	2	1
C	-4	-1	-1	-1	1	1
D	-5	-2	-2	1	0	3
G	-6	-3	-3	0	3	2

		C	A	D	G	D
	0	-1	-2	-3	-4	-5
A	-1	-1	1	0	-1	-2
C	-2	1	0	0	-1	-2
G	-3	0	0	-1	2	1
C	-4	-1	-1	-1	1	1
D	-5	-2	-2	1	0	3
G	-6	-3	-3	0	3	2

- Sử dụng kỹ thuật lưu vết theo quy tắc:
  - (i,j) →(i-1,j-1): U_i và V_j được ghi vào
  - (i,j) →(i-1,j): "-" và V_j được ghi và
  - (i,j) →(i,j-1): U_i và "-" được ghi vào

- Vết Red: 3(2) + 1(-1) + 3(-1) = 2

  CADG–D–

  –ACGCDG
- Vết Blue: 3(2) + 1(-1) + 3(-1) = 2

  –CA–DGD

  ACGCDG–
- Vết Green: 3(2) + 1(-1) + 3(-1) = 2

  –C–ADGC

  ACGCDG–

Một ví dụ khác

- Cho 2 trình tự nucletotide:

  GGATCGA

  GAATTCAGTTA

		G	A	A	T	T	C	A	G	T	T	A
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
G	-1	2	1	0	-1	-2	-3	-4	-5	-6	-7	-8
G	-2	1	1	0	-1	-2	-3	-4	-2	-3	-4	-5
A	-3	0	3	3	2	1	0	-1	-2	-3	-4	-2
T	-4	-1	2	2	5	4	3	2	1	0	-1	-2
C	-5	-2	1	1	4	4	6	5	4	3	2	1
G	-6	-3	0	0	3	3	5	5	7	6	5	4
A	-7	-4	-1	2	2	2	4	7	6	6	5	7

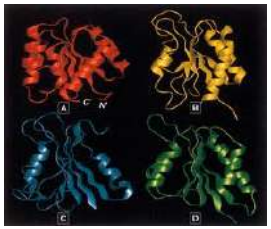
- Bắt cặp 2 trình tự này là:

  GGA–TC–G––A

  | | | || | |

  GAATTCAGTTA

- Kết quả: 6(2) + 4(-1) + 1(-1) = 7



Bài tập

- (P1) Tính toán các giá trị của ma trận với trường hợp tương tự, nhưng:
  - $M_{00} = 0$
  - $M_{i0} = M_{i-1,0} + d$
  - $M_{0j} = M_{0,j-1} + d$
  - $d = -2$
- Rút ra nhận xét

TG Needleman – Wunsch nguyên thủy

```
for i=0 to length(U)
   M(i,0) ← d*i
for j=0 to length(V)
   M(0,j) ← d*j
for i=1 to length(U)
   for j=1 to length(V){
         First  ← M(i-1,j-1) + σ(i,j)
         Second ← M(i-1,j) + d
         Third ← M(i,j-1) + d
         M(i,j) ← max(First, Second, Third)
   }
```



```
AlignmentU ← ""
AlignmentV ← ""
i ← length(U)
j ← length(V)
while (i > 0 and j > 0){
  Value ← M(i,j)
  ValueDiag ← M(i - 1, j - 1)
  ValueUp ← M(i, j - 1)
  ValueLeft ← M(i - 1, j)
  if (Value == ValueDiag + σ(i,j)){
    AlignmentU ← Ui + AlignmentU
    AlignmentV ← Vj + AlignmentV
    i ← i - 1
    j ← j - 1
  }
```

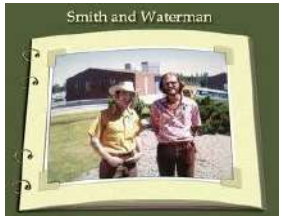


```
  else if (Value == ValueLeft + d){
    AlignmentU ← Ui + AlignmentU
    AlignmentV ← "-" + AlignmentV
    i ← i - 1
  }
  otherwise (Value == ValueUp + d){
    AlignmentU ← "-" + AlignmentU
    AlignmentV ← Vj + AlignmentV
    j ← j - 1
  }
}
```



```
while (i > 0){
  AlignmentU ← Ui + AlignmentU
  AlignmentV ← "-" + AlignmentV
  i ← i - 1
}
while (j > 0){
  AlignmentU ← "-" + AlignmentU
  AlignmentV ← Vj + AlignmentV
  j ← j - 1
}
```

Coi thêm NeedWun.java



Thuật toán

SMITH - WATERMAN

- Do Temple F. Smith và Michael S. Waterman đưa ra vào 1981
- Khác biệt so với thuật toán Needleman - Wunsch là chỉ sử dụng để bắt cặp 2 trình tự trong một đoạn của trình tự (bắt cặp cục bộ - Local Alignment)

- Các bước tính toán hoàn toàn tương tự, chỉ khác một số bước như sau:
  - Cách thức tính ma trận:

$$H_{i0} = 0, \quad \forall i = \overline{0,n}, \qquad H_{0j} = 0, \quad \forall j = \overline{0,m}$$

$$H_{ij} = \max\left\{0, H_{i-1,j-1} + \sigma_{ij}, H_{i-1,j} + d, H_{i,j-1} + d\right\},$$

$$\forall i = \overline{1,n}; j = \overline{1,m}$$

$$\sigma_{ij} = \begin{cases} \text{match} \\ \text{mismatch} \end{cases}, \qquad d = \text{gap}$$

- Do chỉ bắt cặp cục bộ, nên vết được xác định không phải bắt đầu từ giá trị cuối (H_{nm}), mà từ giá trị tốt nhất (điểm cao nhất của ma trận), cho đến khi H có giá trị 0.

Ví dụ

- Với U = "ACA", V = "AGCA", với d = -1 ta có các phần tử của ma trận H như sau:

		A	C	A
	0	0	0	0
A	0	H_{11}	H_{12}	H_{13}
G	0	H_{21}	H_{22}	H_{23}
C	0	H_{31}	H_{32}	H_{33}
A	0	H_{41}	H_{42}	H_{43}

$$H_{11} = \max\{0, H_{00} + \sigma_{11}, H_{01} - 1, H_{10} - 1\} = \max\{0, 0+2, 0-1, 0-1\} = 2$$
$$H_{12} = \max\{0, H_{01} + \sigma_{12}, H_{02} - 1, H_{11} - 1\} = \max\{0, 0-1, 0-1, 2-1\} = 1$$
$$H_{13} = \max\{0, H_{02} + \sigma_{13}, H_{03} - 1, H_{12} - 1\} = \max\{0, 0+2, 0-1, 1-1\} = 2$$

$$H_{21} = \max\{0, H_{10} + \sigma_{21}, H_{11} - 1, H_{20} - 1\}$$
$$= \max\{0, 0-1, 2-1, 0-1\} = 1$$
$$H_{22} = \max\{0, H_{11} + \sigma_{22}, H_{12} - 1, H_{21} - 1\}$$
$$= \max\{0, 2-1, 1-1, 2-1\} = 1$$
$$H_{23} = \max\{0, H_{12} + \sigma_{23}, H_{13} - 1, H_{22} - 1\}$$
$$= \max\{0, 1-1, 2-1, 1-1\} = 1$$

$$H_{31} = \max\{0, H_{20} + \sigma_{31}, H_{21} - 1, H_{30} - 1\}$$
$$= \max\{0, 0-1, 1-1, 0-1\} = 0$$
$$H_{32} = \max\{0, H_{21} + \sigma_{32}, H_{22} - 1, H_{31} - 1\}$$
$$= \max\{0, 1+2, 1-1, 0-1\} = 3$$
$$H_{33} = \max\{0, H_{22} + \sigma_{33}, H_{23} - 1, H_{32} - 1\}$$
$$= \max\{0, 1-1, 1-1, 3-1\} = 2$$

$$H_{41} = \max\{0, H_{30} + \sigma_{41}, H_{31} - 1, H_{40} - 1\}$$
$$= \max\{0, 0+2, 0-1, 0-1\} = 2$$
$$H_{42} = \max\{0, H_{31} + \sigma_{42}, H_{32} - 1, H_{41} - 1\}$$
$$= \max\{0, 0-1, 3-1, 2-1\} = 2$$
$$H_{43} = \max\{0, H_{32} + \sigma_{43}, H_{33} - 1, H_{42} - 1\}$$
$$= \max\{0, 3+2, 2-1, 2-1\} = 5$$

		A	C	A
	0	0	0	0
A	0	2	1	2
G	0	1	1	1
C	0	0	3	2
A	0	2	2	5

Tạo vết

- Xuất phát từ $H_{nmax,mmax}$, nếu:
  - $H_{ij} = H_{i-1,j-1} + \sigma_{ij}$ thì vết (i,j) → (i-1,j-1) theo đường chéo
  - $H_{ij} = H_{i,j-1} + d$ thì vết (i,j) → (i,j-1) đi lui
  - $H_{ij} = H_{i-1,j} + d$ thì vết (i,j) → (i-1,j) đi lên

		A	C	A
	0	0	0	0
A	0	2	1	2
G	0	1	1	1
C	0	0	3	2
A	0	2	2	5

Tìm kết quả

- Nếu
  - (i,j) →(i-1,j-1): theo đường chéo
    U_i và V_j được ghi vào
  - (i,j) →(i-1,j): đi lên
    "-" và V_j được ghi vào
  - (i,j) →(i,j-1): đi lui
    U_i và "-" được ghi vào

- Kết quả bắt cặp
  - U' = "A–CA"
  - V' = "AGCA"
- Độ đánh giá: 3(2) + 1(-1) + 0(-1) = 5

Ví dụ

- Với trình tự dài hơn, chẳng hạn:
  U = "ATATGCTAAG"
  V = "ACTACTTAG"
- Chọn d = -1, Match = 2 và Mismatch = -1 cho sự tương đồng và không tương đồng của 2 phân tử trong 2 trình tự.

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

Với kỹ thuật lưu vết như trên

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

- Cũng bằng cách ghi kết quả theo vết, 2 trình tự được bắt cặp:
  - U' = "A–TATGCTAAG"
  - V' = "ACTA––CTTAG"
- Kết quả: 7(2) + 3(-1) + 1(-1) = 10

Ví dụ

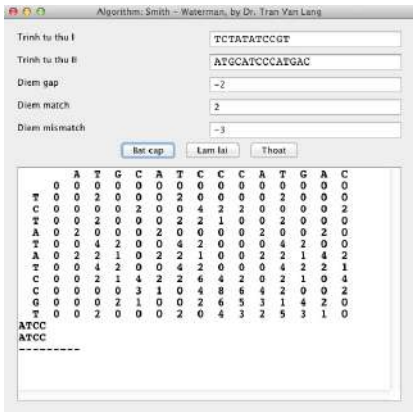
- Với 2 trình tự như hình, có thể tính được

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

- Kết quả trên ứng với match = 2, mismatch = -3 và gap = -2
- Khi đó, 8 là giá trị lớn nhất, nên bắt đầu từ vị trị này để xác định vết.
- Từ đó, kết quả bắt cặp cục bộ của 2 đoạn trình tự:

ATCC

ATCC



- Giá trị σ_{ij} trong các thuật toán Smith - Waterman, hay Needleman –Wunsch khi bắt cặp trình tự DNA chỉ căn cứ vào sự giống nhau (match) hay khác nhau (mismatch) của 2 nucleotide trên 2 trình tự.

- Nếu giống nhau, giá trị này là match, khác nhau giá trị là mismatch
- Khi bắt cặp 2 trình tự protein, giá trị σ_{ij} này được lấy từ ma trận BLOSUM* hoặc PAM (https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt)

* Henikoff S, J.G. Henikoff (1992) - Amino Acid Substitution Matrices from Protein Blocks", *Proc. Natl. Acad. Sci. USA*, **89** (22): 10915–10919 (http://www.pnas.org/content/pnas/89/22/10915.full.pdf).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

- Nhìn lại, với bắt cặp trình từ nucleotide, ma trận tương tự BLOSUM có thể viết cho 4 nucleotide như sau với d = -2, match = 2, mismatch = -1

	A	C	G	T
A	2	-1	-1	-1
C	-1	2	-1	-1
G	-1	-1	2	-1
T	-1	-1	-1	2

Thuật toán

CLUSTAL

Thuật toán Clustal

- Dùng cho việc bắt cặp nhiều trình tự (giải bài toán MSA)
- Lấy ý tưởng từ thuật toán lũy tiến (Progessive Algorithm)

- Thuật toán lũy tiến như sau:
  - Bước 1: giải bài toán PSA trên 2 trình tự bất kỳ được chọn.
  - Bước 2: chọn một trình tự khác rồi sắp hàng với nhóm đã thực hiện.
  - Bước 3: lặp lại Bước 2 cho trình tự khác

Thuật toán Clustal W

- **Bước 1:**
  - Dùng PSA cho tất cả các trình tự
  - Xác định mức độ tương đồng mỗi cặp
  - Xây dựng ma trận khoảng cách tương đồng giữa các trình tự.

- Bước 2:
  - Xây dựng cây cây tương đồng (similarity tree) hay cây hướng dẫn (guide tree) bằng cách dùng thuật toán gom nhóm Neighbor – Joining.
  - Cây hướng dẫn hể hiện mối quan hệ tương đồng giữa các trình tự với nhau

- Bước 3: Thực hiện quá trình lũy tiến
  - Căn cứ vào cây hướng dẫn xác định những nhánh có cặp trình tự tương đồng lớn nhất
  - Thực hiện PSA trên từng cặp
  - Kết hợp những cặp đó lại thu được kết quả đa trình tự.

Minh họa

- Xét 5 trình tự:
  - S_1 = "ARDFGI"
  - S_2 = "AKHGL"
  - S_3 = "ADFIKF"
  - S_4 = "ARFGLI"
  - S_5 = "AKDILM"

- Lần lượt bắt cặp:
  - S_1' = "ARDFGI"
  - S_2' = "A-KHGL"

  - S_1' = "ARDFGI--"
  - S_3' = "A-DF-IKF"

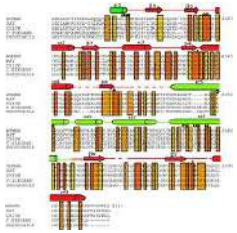
  - S_1' = "ARDFG-I"
  - S_4' = "AR-FGLI"

  - S_1' = "ARDFGI--"
  - S_5' = "AKD--ILM"

– S_2' = "A---KHGL"
– S_3' = "ADFIK--F"

– S_2' = "AKHGL-"
– S_4' = "ARFGLI"

– S_2' = "AKHGL-"
– S_5' = "AKDILM"



– S_3' = "ADF--IKF"
– S_4' = "ARFGLI--"

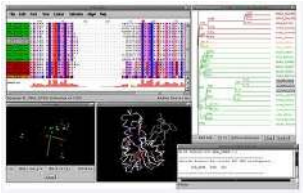
– S_3' = "A-DFIKF"
– S_5' = "AKD-ILM"

– S_4' = "ARFGLI"
– S_5' = "AKDILM"

- Khoảng cách D(S_1',S_2') giữa 2 trình tự bằng tỷ số giữa m và n. Trong đó
  - m = số mismatch giữa 2 trình tự (không tính gap)
  - n = số cặp không phải là gap giữa 2 trình tự
- Ví dụ:
  - S_1' = "ARDFGI"
  - S_2' = "A-KHGL"

Có m = 3, n = 5. Suy ra D(S_1',S_2') = 3/5 = 0,6

- Ví dụ:
  - S_1' = "ARDFGI--"
  - S_3' = "A-DF-IKF"



Có m = 0, n = 4. Suy ra D(S_1',S_3') = 0/4 = 0

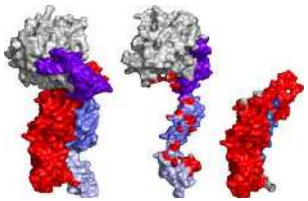
- Ví dụ:
  - S_1' = "ARDFGI--"
  - S_5' = "AKD--ILM"

  Có m = 1, n = 4. Suy ra D(S_1',S_5') = 1/4 = 0.25

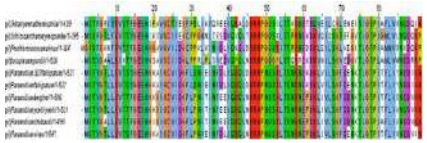
Ma trận khoảng cách

	S_1	S_2	S_3	S_4	S_5
S_1	-				
S_2	0,60	-			
S_3	0	0,33	-		
S_4	0	0,40	0,25	-	
S_5	0,25	0,60	0,40	0,66	-

- Theo ma trận khoảng cách, có thể S_1 và S_3 là nhỏ nhất, nên mức độ gần nhau là nhiều nhất.
- Hoặc S_1 và S_4



	S_{13}	S_2	S_4	S_5
S_{13}				
S_2	(0,6+0,33)/2 = 0,465			
S_4	(0+0,25)/2 = 0,125	0,4		
S_5	(0,25+0,4)/2 = 0,325	0,6	0,66	

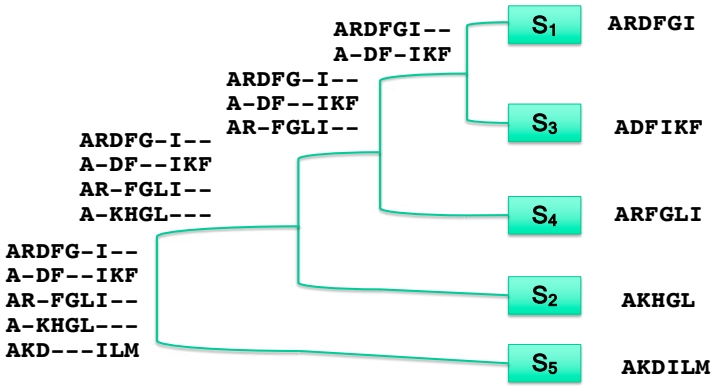
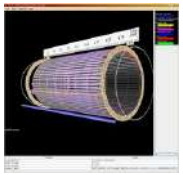


- Tiếp tục, khoảng cách giữa S_{13} và S_4 là nhỏ nhất.
- Nên S_{13} và S_4 là gần nhau nhất

	$S_{13,4}$	S_2	S_5
$S_{13,4}$			
S_2	(0,465+0,4)/2 = 0,4325		
S_5	(0,325+0,66)/2 = 0,4925	0,6	

- Tiếp tục, còn S_{134} và S_2 là nhỏ nhất

	$S_{134,2}$	S_5
$S_{134,2}$		
S_5	(0,4925+0,6)/2 = 0,54625	



- Ở đây kết quả có được bằng cách gióng từng cặp:
  - S_1, S_3
  - Lấy kết quả S_1 có được để bắt cặp với S_4
  - Tương tự, với S_2
  - Rồi với S_5