



دانشگاه صنعتی امیرکبیر

(پلیتکنیک تهران)

دانشکده مهندسی کامپیوتر

## درس بیوانفورماتیک

تمرین دوم ۲

امیرمهدی زرین نژاد

۹۷۳۱۰۸۷

نیمسال دوم ۱۴۰۰



١٦٣

$$S_{i,j} = \max \begin{cases} S(i-1, j-1) + \delta(x_i, y_j) \\ S(i-1, j) - \gamma \\ S(i, j-1) - \gamma \end{cases}$$

،  $\gamma$  = gap penalty

« در جمل اسیارات داریم )  $M, N$  کل تعداد مکالمات (  $\rightarrow$  اسیارات سه = آم و سه = آم ، می باید  $\rightarrow$  این .  
 $\Rightarrow \boxed{\text{gap Penalty} = \pm 1}$

$$\left( \begin{array}{l} s(i,0) = 8x - i, \quad s(0,j) = 8x - j \Rightarrow \underline{\underline{8=+1}} \\ s(0,1) = -1, \quad s(0,2) = -2; \dots \end{array} \right) \quad : \text{مربع}.$$

$$\begin{aligned}
 & \text{(A ≠ T) Case mismatch} \\
 & S_{1,1} = \max \left\{ \begin{array}{l} S(0,0) + 6(x_1, y_1) \\ S(0,1) - 1 \\ S(1,0) - 1 \end{array} \right\} \\
 & \Rightarrow S_{1,1} = \max \left\{ \begin{array}{l} 0 + 6(x_1, y_1) \\ -1 - 1 = -2 \\ -1 - 1 = -2 \end{array} \right\} = \boxed{-1} \Rightarrow 6(x_1, y_1) = \boxed{\text{mismatch value}} = \boxed{-1}
 \end{aligned}$$

$$S_{x,y} = \max \begin{cases} S(1,1) + b(n_x, y_x) \\ S(1,x) - r \\ S(y,1) - \delta \end{cases} = \max \begin{cases} -1 + b(n_x, y_x) \\ -x - 1 = -10 \\ -y - 1 = -14 \end{cases} = 1 \Rightarrow b(n_x, y_x) = +10 + 1 = 11$$

match زیرا  $= 11$

gap penalty = +1  
mismatch = -1, match = 1

(.)

	-	T	G	C	A	T	T	A	C	G	G	A
-	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
A	-1	-1	-2	-3	-1	-1	-2	-3	-1	-1	-2	-3
G	-1	-2	1	1	0	-1	-2	-3	-1	-1	-2	-3
T	-1	1	1	1	0	1	1	0	-1	-1	-2	-3
T	-1	0	0	0	0	1	1	0	-1	-1	-2	-3
G	-1	-1	-1	-1	-1	0	0	0	1	1	1	1
G	-1	-2	1	1	1	1	1	1	1	1	1	1
A	-1	-1	1	1	0	1	1	1	1	1	1	1

$\Rightarrow \text{Score} = 14$

1) AG -- TT-C-GA      2) AG - TT-CG-A  
T G C A T T A C G G A      ' T G C A T T A C G G A

ج) همان قدر در میان میم، ۲ هم تکراری ممکن است (مجد دارد). جواب از درست میتوان به اسلز مانند و میتواند باشد.  
(بین خانه)  $\frac{10,9}{G,G}$  ،  $\frac{10,8}{C,C}$  (مجد دارد)

سوال (۳)

match = 1, mismatch = -1, gap penalty = -2

	A	T	A	T	A	G	C
A	0	0	0	0	0	0	0
G	.	.	0	0	.	0	0
T	.	2	0	2	0	0	1
T	.	1	1	2	1	0	0
A	.	0	2	2	2	0	0
T	0	2	2	2	2	2	1
C	0	0	1	2	0	2	0

مسئلہ ۴

$\Rightarrow \boxed{TAT} \cdot G \boxed{TATAC}$

اصلی نہیں

$$\text{از نیز } H_{i,j} = \max_{x[i-i], y[1-j]} \left\{ \begin{array}{l} H_{i-1, j-1} + S(a_i, b_j) \\ \max_{k \geq 1} \{ H_{i-k, j} - w_k \} \\ \max_{l \geq 1} \{ H_{i, j-l} - w_l \} \\ 0 \end{array} \right.$$

برای back trace

#### سوال (۴)

الف) برای این کار می‌توانیم به طول رشته‌ها و شباهتشان توجه کنیم. رشته‌های A و B از نظر طولی بسیار نزدیک به هم هستند و خود رشته‌ها شباهت زیادی به یکدیگر دارند(به عبارتی فاصله تکاملی کمی دارند). پس insertion و deletion کمتر بین آن‌ها وجود دارد و برای مقایسه‌شان بهتر است هم‌ترازی سراسری یا همان Global alignment را به کار ببریم.

اما در مقایسه‌ی دو رشته‌ی A و C با توجه به اینکه از نظر طولی فاصله دارند و شباهت قابل توجهی هم بین کاراکترهایشان وجود ندارد؛ فاصله تکاملی‌شان زیاد است و نمی‌توان هم‌ترازی سراسری مناسبی به دست آورد بلکه برخی قسمت‌ها از دو رشته با یکدیگر شباهت دارند و می‌توان بین آن‌ها هم‌ترازی‌ها در نظر گرفت. پس به کارگیری Local alignment برای مقایسه‌ی این دو رشته مناسب‌تر است.

#### ب)

نتایج هم‌ترازی سراسری و محلی برای زوج‌های ذکر شده در ۴ صفحه‌ی بعدی آمده‌اند. با توجه به امتیازات و میزان شباهت و ... به دست آمده می‌بینیم که امتیازات هم‌ترازی سراسری و محلی دو رشته‌ی A و C بیشتر از زوج A و B است که شباهت بیشتر بینشان را نشان می‌دهد. چراکه از نظر طولی و کاراکتری نزدیک به هم و شبیه هستند. در نتیجه GAP‌های کمتری هم دارند که به معنای deletion و mismatch های کمتر و امتیاز بالاتر است. در مقابل دو رشته A و C امتیاز کمتر و تقریباً نصف زوج قبلی را دارند که به دلیل شباهت کمتر و GAP و mismatch های بیشتر است که در بخش الف نیز ذکر شد. هم‌چنین با مقایسه‌ی امتیاز Global alignment های هر زوج می‌بینیم که امتیازشان نزدیک به هم است اما در Local alignment کمی بیشتر شده است چراکه در این هم‌ترازی تنها بخش (زیررشته)‌های قابل توجه را align می‌کنیم که gap‌هایی کمتر و شباهت و امتیاز بیشتر را نتیجه می‌دهد. در خصوص زوج A و C هم این تفاوت بیشتر است زیرا همان‌طور که در بخش الف گفته شد، از نظر طولی فاصله دارند و شباهت قابل توجهی هم بین کاراکترهایشان وجود ندارد و هم‌تراز کردن برخی قسمت‌ها از دو رشته که با یکدیگر شباهت دارند مناسب‌تر از هم‌ترازی سراسری است و امتیاز بیشتری می‌دهد.

- Global(A, B) : امتیاز ۸۷۲ برای هم ترازی سراسری رشته های A و B: (دیگر موارد نیز هایلایت شده اند)

- Global(A, C) : امتیاز ۴۱۰.۵ برای هم ترازی سراسری رشته های A و C: (دیگر موارد نیز هایلایت شده اند)

```

#####
# Program: needle
# Rundate: Wed  6 Apr 2022 14:45:19
# Commandline: needle
#
#   -auto
#   -stdout
#   -asequence emboss_needle-I20220406-144516-0907-78748088-p2m.asequence
#   -bsequence emboss_needle-I20220406-144516-0907-78748088-p2m.bsequence
#   -datafile EBLOSUM62
#   -gapopen 10.0
#   -gapextend 0.5
#   -endopen 10.0
#   -endextend 0.5
#   -aformat3 pair
#   -sprotein1
#   -sprotein2
# Align_format: pair
# Report_file: stdout
#####
=====

#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 390
# Identity: 101/390 (25.9%)
# Similarity: 140/390 (35.9%)
# Gaps: 177/390 (45.4%)
# Score: 410.5
#
#
=====

EMBOSS_001      1 -----MEAMNVEKASADGNLPEVI      19
EMBOSS_001      1 MTXPALSLHTPLSTSFTPAVWYNMGWSILSKIGAINIENAVGGGKLLEVD 50
EMBOSS_001      20 SNIKETLKIVSRTPVNITMAGDSGNGMSTFISALRNTGHEKGASPPTELV 69
EMBOSS_001      51 SIVRGTLTEESSTPV-XAVTGDSSNVMSFIKALRVTHXEEATAPGMV 99
EMBOSS_001      70 KATQRQCASYFSHSFNVLWDLPGTGSATTLENYLMEMQFNRYD-FIMV 118
EMBOSS_001      100 RTTXIPTCYSSSYIPSVVLWDMPGTGTATQSPE NYLEEMHFSXYELFIIS 149
EMBOSS_001      119 ASAQFSMNHVMLAKTAEDMGKKFYIVWTKLDMDSLSTGALPEVQLLQ-IRE 167
EMBOSS_001      150 ISEQFSMNLIKLAQIIQSLGKRFYIWTKLDRDLSTS AFWEWLQNIQE 199
EMBOSS_001      168 NVLENLQKERLACHEKYLKSTPE---NSTRPR-----NIPS 201
EMBOSS_001      200 NIQKNLLKEGVCEPIIFLVSIDPLLHNFPVPRDTLHIRYHGPLENLPYT 249
EMBOSS_001      202 RKLYVN----LLRIFNS----- 214
EMBOSS_001      250 HEKVINYEVISLXVKIAS KFFQDTLGFQNADDLGECLKAYHLLFXVD 299
EMBOSS_001      215 ----- 214
EMBOSS_001      300 LQQVAQHMGKPMEEYKTIMKSQDLHTVHPRETLALYWMNCNTASYISQI 349
EMBOSS_001      215 ----- 214
EMBOSS_001      350 PLLDDTIINYTRQXKYRQFLGIVTKDTKTILKKILQDFII 389
=====

#
# -----
# -----

```

- امتیاز ۸۷۶ برای هم ترازی محلی رشته های A و B: (دیگر موارد نیز هایلایت شده اند) Local(A, B)

**امتیاز ۴۱۹ برای هم ترازی محلی رشته های A و C: (دیگر موارد نیز هایلایت شده اند)**

```
#####
# Program: water
# Rundate: Wed  6 Apr 2022 14:50:01
# Commandline: water
#   -auto
#   -stdout
#   -asequence emboss_water-I20220406-144958-0528-33013500-p2m.asequence
#   -bsequence emboss_water-I20220406-144958-0528-33013500-p2m.bsequence
#   -datafile EBLOSUM62
#   -gapopen 10.0
#   -gapextend 0.5
#   -aformat3 pair
#   -sprotein1
#   -sprotein2
# Align_format: pair
# Report_file: stdout
#####

=====
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 176
# Identity:      92/176 (52.3%)
# Similarity:    120/176 (68.2%)
# Gaps:          3/176 ( 1.7%)
# Score: 419.0
#
#
=====

EMBOSS_001      3 AMNVEKASADGNLPEVISNIKETLKIVSRTPVNITMAGDGSNGMSTFISA      52
                 |:|:|.|...|.|.|.|.:|||:..|.| || ..:|||.|||:|||.|.
EMBOSS_001      34 AINIENAVGGGKILLEVDSIVRGTLTEESSTPV-XAVTGDSSNVMSFFKA     82
                 ||.|||..:||:|||.:|||...|..|||...:|||:||||:|||:|||..|
EMBOSS_001      53 LRNTGHEGKASPPTELVKATQRCASYFSSHFSNVLWDLPGTGSATTLE     102
                 ||.|||..:||:|||.:|||...|..|||...:|||:||||:|||:|||..|
EMBOSS_001      83 LRVIGHXEATAPTMVRRTXIPTCYSSSYIPSVVLWDMPGTGTATQSPE    132
                 ||.|||..:||:|||.:|||...|..|||...:|||:||||:|||:|||..|
EMBOSS_001      103 NYLMEMQFNRYD-FIMVASAQFSMNHVMLAKTAEDMGKKFYIWWTKLMDM    151
                 |||..||.||.:|||...|..|||...:|||...:|||:|||:||||.|
EMBOSS_001      133 NYLEEMHFSXYELFIISISEQFSMNLIKLAQIIQLSLGKRFYIWTKLDRD    182
                 |||..||.||.:|||...|..|||...:|||...:|||:|||:||||.|
EMBOSS_001      152 LSTGALPEVQLLQ-IENVLENLQKE      176
                 |||..||.||.:|||...|..|||...:|||...:|||:|||:||||.|
EMBOSS_001      183 LSTSASFWEWLQNQENIQKNLLKE      208

-----
```

## سوال ۵

همانطور که می‌دانیم در دنیای واقعی mismatch ها و match ها بین کاراکترهای مختلف امتیاز برابر مشابهی ندارند چراکه میوتیشن و تبدیلات بینشان با احتمالات یکسانی رخ نمی‌دهد و برخی محتمل‌تر و برخی نادر تر هستند و ماتریس‌هایی برای امتیازدهی درنظر گرفته می‌شود که این امتیازات بین کاراکترهای مختلف را از آن‌ها به دست آورد (معمولًا آن‌هایی از نظر فیریکی-شیمیایی به یکدیگر نزدیک‌تر هستند؛ پنالتی کمتری دارند).

ماتریس امتیازدهی PAM یا Point Accepted Mutation معرفی شد، بر اساس مدل تکاملی پیاده‌سازی می‌شود یعنی دنباله‌های پروتئینی نزدیک به هم را با هم تراز و امتیازدهی می‌کند. در این ماتریس در سطرهای و ستون‌ها ۲۰ آمینواسید وجود دارد که هر درایه‌اش احتمال جایگزینی آن دو آمینواسید را نشان می‌دهد. پس این ماتریس احتمال جهش بین توالی‌های پروتئینی را نگهداری می‌کند و مربوط به ۷۱ خانواده پروتئین می‌شود که با یکدیگر بیش از ۸۵ درصد شباهت داشته‌اند و ۱۵۷۲ جهش از آن‌ها به دست می‌آید. ماتریس PAM نسخه‌های ۱ تا ۲۵۰ دارد که PAM1 برای دنباله‌هایی استفاده می‌شود که شباهت زیادی دارند (فاصله تکاملی کمتر) و PAM250 برای دنباله‌هایی استفاده می‌شود که کمترین شباهت و بیشترین فاصله تکاملی را دارند. و اعداد ۱ تا ۲۵۰ به معنای درصد جهش‌ها هستند و PAM1 یعنی ۱ جهش در ۱۰۰ توالی و PAM250 یعنی ۲۵۰ جهش در ۱۰۰ توالی.

محاسبه‌ی این ماتریس نیز بدین صورت است که ابتدا ماتریسی با کمتر از ۱ درصد جهش (PAM1) محاسبه می‌شود و سپس ماتریس‌های بعدی (...، PAM2، PAM3) از برونيابی این ماتریس به صورت بازگشتی محاسبه می‌شوند. این ماتریس کاربردهایی از جمله در بازسازی درخت‌های فیلوجنی (phylogenetic trees) و یا شناخت برخی بیماری‌های ژنتیکی و یا در ارتباط با DNA دارد.

ماتریس BLOSUM یا BLOck Substitution Matrix یک ماتریس دیگر برای امتیازدهی است که توسط Henikoff ارائه شد و در آن هم‌ترازی با توجه به جانشینی‌های ممکن آمینواسیدها بین چندین دنباله انجام می‌شود (multiple sequence alignments). این ماتریس براساس بیش از ۲۰۰۰ بلاک که نشان‌دهنده ۵۰۰ گروه دنباله‌پروتئین است به دست می‌آید. بلاک‌ها (الگوی سیکوئنس) درواقع الیمنت‌های بدون گپ برای کمتر از ۶۰ آمینواسید در طول رشته هستند.

در این روش، این بلاک‌ها از سیکوئنس‌ها هستند مورد بررسی قرار می‌گیرند و نه کل دنباله چراکه بخش‌هایی هستند که در آن‌ها جهش کمتر است، تغییرات کمتری نسبت به بقیه دنباله وجود دارد و شباهت و اطلاعات مفید بیشتری می‌توان یافت که به آن‌ها نواحی محافظت شده نیز گفته می‌شود.

پس مقادیر این ماتریس با توجه به نرخ جانشینی آمینواسیدها در این بلاک‌ها محاسبه می‌شود. به عبارتی امتیاز BLOSUM برای یک جفت باقی‌مانده برابر است با لگاریتم نرخ مشاهده جانشینی باقی‌مانده تقسیم بر احتمال موردنظر برای همان باقی‌مانده (یا همان random chance). BLOSUM هم مانند PAM نسخه‌های مختلف دارد و به صورت BLOSUM<sub>n</sub> نشان‌داده می‌شود. مانند BLOSUM45 که (n) ۴۵

نشان دهنده درصد شباهت محتمل و موردنظر است. پس BLOSUM45 برای دنباله‌های با فاصله بیشتر و BLOSUM62 برای دنباله‌های با فاصله کمتر مورد استفاده قرار می‌گیرد. (نسخه‌های ۴۵، ۵۲، ۶۲، و ۸۰ نسخه‌های معمول و پر کاربردش هستند)

ماتریس BLOSUM بیشتر برای جستجو در پایگاه داده‌ها و پیدا کردن نواحی محافظت شده در پروتئین‌ها به کار گرفته می‌شود.

مقایسه‌ها و کاربردها در تعاریف ذکر شد. در ادامه چند مورد مقایسه را با توجه مطالب فوق نام می‌بریم: ماتریس PAM بر اساس مقایسه بین پروتئین‌هایی که نزدیک به هم هستند، برونویابی می‌شوند و این کار بین جفت پروتئین‌ها صورت می‌گیرد. در اما ماتریس BLOSUM بر اساس هم ترازی دنباله‌های پروتئینی که از یکدیگر فاصله گرفته‌اند و به صورت multiple sequence alignment محاسبه می‌شود و بلوک‌هایی در آن‌ها برای این کار درنظر گرفته می‌شوند.

کارکرد عددی مقابله BLOSUM و PAM برعکس هم است و برای PAM هرچه بیشتر باشد شباهت کمتر است اما برای BLOSUM هرچه بیشتر باشد شباهت بیشتر است. چراکه عدد PAM تعداد جهش و عدد BLOSUM درصد شباهت را نشان میدهد.

روش PAM الاینمنت‌ها را بین یکسری رشته پروتئینی که رابطه نزدیکی باهم دارند برقرار می‌کند و بر این اساس، احتمالات را بدست می‌آورد.

روش BLOSUM آن‌هایی که تفاوت و فاصله دارند را درنظر می‌گیرد. برای همین نمی‌تواند کل الاینمنت را دخیل کند. پس یکسری بلاک درنظر می‌گیرد که داخلشان تعداد mutation‌ها کمتر است و شباهت بیشتری دارند. (برخلاف ماتریس PAM که کل توالی پروتئین را مورد بررسی قرار می‌دهد)

پس ماتریس PAM بیشتر در هم ترازی سراسری و ماتریس BLOSUM بیشتر در هم ترازی محلی مناسب‌اند.