آزمون‌های نهایی
بیست‌ودومین المپیاد
زیست‌شناسی ایران

# بیوانفورماتیک

## آزمون نهایی

**مدت زمان آزمون: ۱۰۵ دقیقه**

**۱. یک توالی پروتئینی ۱۰ اسید آمینه‌ای، از حداکثر چند توالی DNA کدکننده می تواند به دست آمده باشد؟**

آ. $۶^{۱۰}$ ب. $۴^{۱۰}$ ج. $۴^{۳۰}$ د. $۴^{۲۷}$

**۲. کدام گزینه(ها) درست است؟ برای دو توالی دلخواه:**

آ. هر semi-global-alignment یک local-alignment است.

ب. بهترین نمره semi-global-alignment بیش‌تر از بهترین نمره global alignment است.

ج. بهترین نمره local-alignment بیش‌تر از بهترین نمره global alignment است.

د. موارد فوق بستگی به سیستم نمره دهی دارد.

**۳. نمره بهترین global alignment دو توالی زیر را با فرض linear gap pentalty = -2، Mismatch = -1 و match = +2 به دست آورید.**

**TACGAGTACGA**

**ACTGACGACTGAC**

آ. ۲ ب. ۴ ج. ۶ د. ۸

**۴. در ماتریس نمره دهی PAM250 نمره جایگزینی GLY به ARG، ۳- در نظر گرفته شده است. فراوانی نسبی ARG در پایگاه داده پروتئینی ۰.۰۴۱ است. با چه احتمالی بعد از PAM250 زمانی GLY به ARG تبدیل می شود.**

آ. اطلاعات مساله کافی نیست.

ب. ۰.۰۲

ج. $۴۱ \times ۱۰^{-۶}$

د. ۰.۰۰۵

**۵. یک ژن پروکاریوتی، توالی پشت‌سرهمی از کدون‌هاست. یک ژن با کدون ATG آغاز و با یکی از سه کدون TAA ,TAG ,TGA پایان می‌پذیرد. قصد داریم مدلی مبتنی بر زنجیر مارکوف مرتبه‌ی دوم برای احتمال یک توالی بسازیم. این مدل چه تعداد پارامتر دارد؟**

الف. ۲۳۸۵۱۴ ب. ۲۲۶۹۸۱ ج. ۲۳۸۲۰۵ د. ۲۲۶۹۸۹

**۶. یک dot-plot با window size = 8 و stringency = 6 در نظر بگیرید. در این plot، یک جفت قطعات به طول ۸ از جفت توالی x, y را matching pair در نظر می‌گیرند، اگر حداقل ۶ عدد match وجود داشته باشد. در دو توالی به طول ۱۰۰۰، به طور متوسط چند جفت تطابق ( matching pair) مشاهده می‌شود؟ احتمال وقوع نوکلئوتیدهای T,C,A,G برابر است با:**

**P(A)=P(T)= 0.30, P(C)=P(G)=0.20.**

آ. ۲۸۶۰ ب. ۲۶۰۰۰۰ ج. ۲۵۶۳۷۰ د. ۲۹۰۰

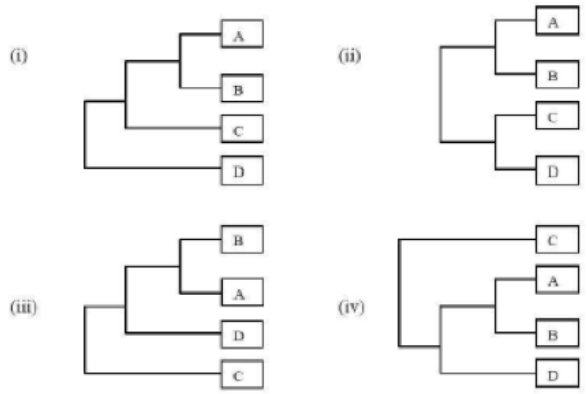
**۷. ماتریس فاصله برای گونه های A, B, C, D, E به صورت زیر داده شده است.**

| | B | C | D | E |
|---|---|---|---|---|
| A | 5 | 4 | 9 | 8 |
| B | | 5 | 10 | 9 |
| C | | | 7 | 6 |
| D | | | | 7 |

**در درخت فیلوژنی به دست آمده با روش Neighbor-Joining فاصله جد مشترک D, E را با W و جد مشترک A, B را با Y نشان می‌دهیم. فاصله W از Y برابر است با:**

آ. ۲ ب. ۳ ج. ۴ د. ۷

**۸. کدام درخت‌های روبه‌رو با هم معادل هستند؟**




آ. i, iii

ب. iii, iv

ج. i,iii,iv

د. هیچکدام

**۹. قاعده once a gap, always a gap**

آ. منجر به multiple alignment با جواب بهینه می‌شود.

ب. فرض می‌کند که گپ‌ها میان توالی‌هایی که نزدیک‌ترین به هم هستند باید در multiple alignment حفظ شود.

ج. فرض می‌کند که توالی‌هایی که اخیراً در تکامل از هم جدا شده‌اند دارای اولویت از نظر ترتیب در ساخت multiple alignment هستند.

د. ممکن است گپ‌های ایجاد شده با اضافه کردن توالی جدید در مراحل ساخت multiple alignment به صورت نامناسبی قرار بگیرد.

**۱۰. بازدهی برنامه‌های multiple alignment را چگونه می‌توان ارتقا داد؟**

آ. با اضافه کردن اطلاعات مربوط به ساختار سوم پروتئین

ب. با انجام دادن psi-blast

ج. با داشتن اطلاعات pair-wise alignment

د. هیچکدام

**۱۱. درستی و نادرستی گزاره‌های زیر را مشخص کنید.**

آ. درخت ساخته شده توسط UPGMA یک Cladogram است.

ب. دو روش out group و midpoint عمده‌ترین روش‌های ریشه‌دار کردن درخت هستند.

ج. Bootstrapping اغلب برای تایید طول شاخه‌های درخت به کار گرفته می‌شود.

د. درخت ایجاد شده توسط neighbor-joining یک درخت ultrametric است.

**۱۲. توالی DNA از ۱۰۰۰ ویروس که هر کدام ۵۰۰ نوکلئوتید دارند را در نظر بگیرید. اگر بخواهید بدانید کدام جفت توالی تقریبا یکسان هستند کدام روش زیر ماراتر است؟**

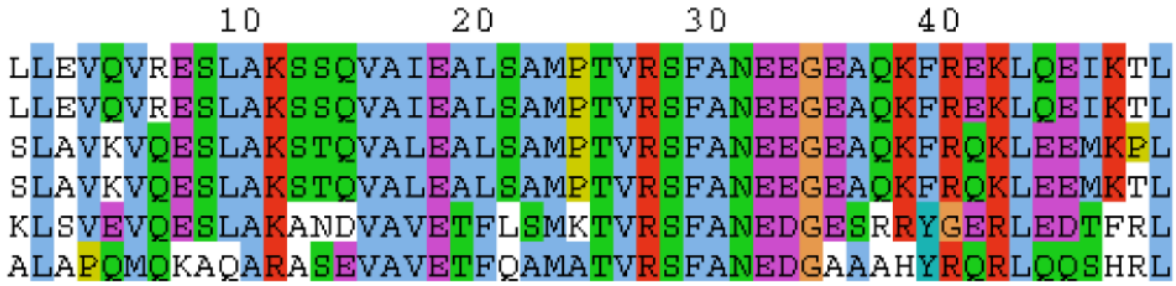
آ. BLAST

ب. Neighbor joining phylogenetic

ج. STAR

د. PSI-Blast

**۱۳. Multiple alignment زیر مربوط به توالی پروتئینی ۶ گونه‌ی انسان، شامپانزه، موش، رت، مرغ و کوسه است.**




آ. اسید آمینه‌ها در conserved positions با اسید آمینه‌های conserved در PAM250 همخوانی دارند.

ب. اگر توالی اول مربوط به انسان و توالی سوم مربوط به rat باشد، توالی دوم شامپانزه است.

ج. اگر توالی اول مربوط به انسان و توالی سوم مربوط به rat باشد، توالی چهارم موش است.

د. اگر توالی اول مربوط به انسان و توالی سوم مربوط به rat باشد، توالی پنجم مرغ است.

**۱۴. برای پیش‌بینی عملکرد پروتئینی از دو سایت (site) برای جستجو به روش BLAST استفاده کردیم. نتایج این دو سایت در زیر آمده است.**

```
 Site 1:
 >gb|AAC60279.1| ubiquitin/ribosomal protein [Gallus gallus]

Length=156
 Score = 47.8 bits (112),  E-value = 1e-04
 Identities = 47/95 (49%), Positives = 50/95 (52%), Gaps = 36/95 (37%)
Query  1    IRKETTLHKVLRLWGGAYKDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXKKKSY  60
            I+KE+TLH VLRL GGA K                                   KKKSY
Sbjct  61   IQKESTLHLVLRLRGGAKK----------------------------------RKKKSY  85
Query  61   TMPXXXXXXXXXXX-AVLPYYKIDEYGKISRFRRE  94
            T P            AVL YYK+DE GKISR RRE
Sbjct  86   TTPKKNKHKRKKVKLAVLKYYKVDENGKISRLRRE  120

Site 2:
>sp|P42568|AF9_HUMAN Protein AF-9 (Myeloid/lymphoid or mixed-lineage leukemia associated)
Length=568
 Score = 68.9 bits (167),  E-value = 5e-11
 Identities = 40/52 (76%), Positives = 44/52 (84%), Gaps = 0/52 (0%)
Query  21   SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSKKKSYTMPKKNKHKHKK  72
            SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS   S++ P K   +HK+
Sbjct  154  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSTSFSKPHKLMKEHKE  205
```

**صحت گزاره‌ی زیر را تعیین کنید. (نمره منفی دو برابر نمره سوال)**

- نتایج Site 1 معتبرتر از نتایج Site 2 است.

**۱۵. گزاره(های) صحیح کدام است؟**

آ. E-value تعداد مورد انتظار false positive در جستجوی BLAST را گزارش می‌کند.

ب. تغییر ماتریس نمره‌دهی از PAM30 به APM120 باعث می‌شود BLAST توالی‌های کم‌تری بدهد.

ج. در الگوریتم BLAST لیستی از "word"ها تهیه می‌شود. "word"هایی که نمره آن‌ها بزرگ‌تر از آستانه‌ی T باشد، hit نامیده می‌شوند. Hitها در database جستجو می‌شوند تا exact match یافته شوند و سپس آن‌ها از هر دو طرف گسترش داده می‌شوند.

د. normalized BLAST Score که Bit Score هم نامیده می‌شود بدون واحد (unitless) است.

**۱۶. رایج ترین روش در multiple alignment:**

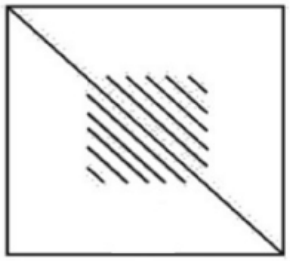
آ. تمام توالی‌ها یک‌باره الاین شوند.

ب. توالی‌ها دوبه‌دو و با روش پیش‌رونده progressive الاین می‌شوند.

ج. توالی‌ها یکی پس از دیگری الاین می‌شوند.

د. هیچکدام

**۱۷. Dot-plot یک توالی در مقابل خودش در شکل زیر داده شده است. نواحی حول قطر اصلی چه چیزی را نشان می‌دهد؟**



الف) palindrome region
ب) low complexity region
ج) local alignment
د) repeated region

**۱۸. گزاره‌های صحیح را انتخاب کنید.**

آ. ساخت ماتریس‌های PAM مبتنی بر مدل مارکوف است.

ب. در ساخت ماتریس نمره‌دهی PAM فرض شده که نرخ جهش هر اسیدآمینه در هر پوزیشن روی توالی با هم برابر است.

ج. ماتریس‌های BLOSUM مبتنی بر الاینمنت پروتئین‌های خویشاوند نزدیک است.

د. Henikoff and Henikoff در سال ۱۹۹۲ نشان دادند که ماتریس BLOSUM62 عملکرد بهتری از ماتریس‌های PAM در جستجوی BLAST دارد.

**۱۹. گزاره‌های صحیح را انتخاب کنید.**

آ. ساعت مولکولی فرضیه‌ای است که بیان می‌کند توالی‌های مولکولی با نرخ‌های متفاوتی تکامل یافته‌اند.

ب. طول شاخه‌ها (branch length) در یک phylogram متناسب با زمان تکامل است.

ج. درخت تولید شده توسط روش UPGMA یک phylogram است.

د. درخت تولید شده توسط روش nighbor-joining یک درخت ultrametric است.

**۲۰. دو توالی x = ACGGTAC و y = GAGGT با سیستم نمره‌دهی match = +1 و mismatch=-1 و d = 3 و e = 2 که نمره گپ با طول k به صورت $w(k) = -d + (k - 1)e$ محاسبه می‌شود. گزاره(های) صحیح کدام است؟**

آ. $I_x(7,5) = -4$

ب. $M(7,5) = -4$

ج. $I_y(7,5) = -4$

د. $M(7,5) = -6$

**۲۱. گزاره(های) صحیح کدام است؟**

آ. برای مقایسه دو توالی نسبتاً دور از ماتریس نمره‌دهی BLOSUM45 استفاده می‌کنیم.

ب. BLOSUMها برای الاینمنت پروتئین‌های خیلی دور مناسب هستند.

ج. برای local alignment ماتریس‌های BLOSUM مناسب‌تر از PAM هستند.

د. اگر دو توالی بعد از گلوبال الاینمنت، ۵۰ درصد یکسانی (identity) داشته باشند در این صورت به طور متوسط ۸۰ تغییر رخ داده است.

**۲۲. یک tandem repeat از الگویی مانند $P = p_1, p_2, ...p_m$، توالی به طول mxk است که از کنار هم قرار گرفتن k نسخه از P به دست می آید و آن را با $P^k$ نمایش می دهیم. به عنوان مثال برای الگوی P = GGT و توالی T = ACCGGTGCT بزرگترین tandem repeat با k = 3 به صورت زیر به دست می آید.**

**ACCGGTGCTG-TAA**

**GGTGGTGGT**

**مقدار tandem repeat در گزاره‌های آ، ب، ج و د را از کوچک به بزرگ مرتب کنید.**

آ.

$$F(0,j) = 0,\ 0 \le j \le n$$

$$F(i,0) = -i,\ 0 \le i \le nm$$

$$F(i,j) = max\begin{cases} F(i-1,j-1) + S(P^n(i),T(j)) \\ F(i-1,j) + S(P^n(i),-) \\ F(i,j-1) + S(-,T(j)) \end{cases}$$

نمره ماکسیمم در ماتریس F روی سطر km که $0 \le k \le n$ جواب مساله است.

ب.

$$F(0,j) = 0,\ 0 \le j \le n$$

$$F(i,0) = 0,\ 0 \le i \le nm$$

$$F(i,j) = max\begin{cases} F(i-1,j-1) + S(P^n(i),T(j)) \\ F(i-1,j) \\ F(i,j-1) \end{cases}$$

نمره ماکسیمم در ماتریس F روی سطر km که $0 \le k \le n$ جواب مساله است.

ج.

$$F(0,j) = 0,\ 0 \le j \le n$$

$$F(i,0) = 0,\ 0 \le i \le nm$$

$$F(i,j) = max\begin{cases} F(i-1,j-1) + S(P^n(i),T(j)) \\ F(i-1,j) + S(P^n(i),-) \\ F(i,j-1) + S(-,T(j)) \end{cases}$$

نمره ماکسیمم در ماتریس F روی سطر km که $0 \le k \le n$ جواب مساله است.

د.

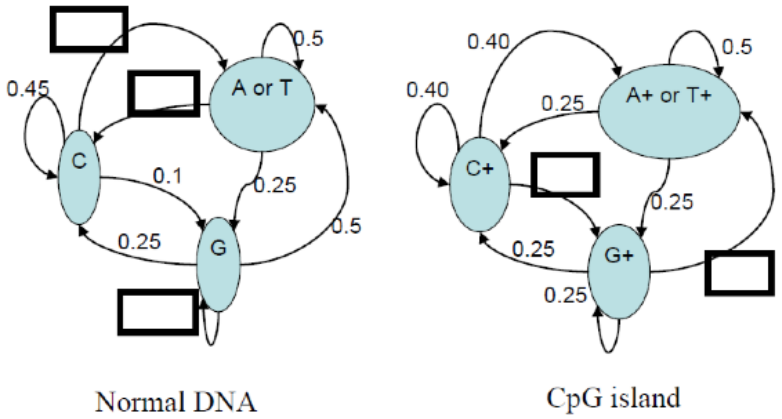
$$F(0,j) = 0,\ 0 \le j \le n$$

$$F(i,0) = -i,\ 0 \le i \le nm$$

$$F(i,j) = max\begin{cases} F(i-1,j-1) + S(P^n(i),T(j)) \\ F(i-1,j) \\ F(i,j-1) \end{cases}$$

نمره ماکسیمم در ماتریس F روی سطر km که $0 \le k \le n$ جواب مساله است.

**۲۳. زنجیره مارکوف مرتبه‌ی اول زیر را در نظر بگیرید. محققی احتمالات انتقال که روی یال‌ها درج شده را گزارش کرده و مستطیل‌ها خالی مانده است.**




**فرض کنید احتمال آغاز توالی با C برابر یک است. با توجه به مدل، توالی CCGCACG**

آ. محتمل‌تر است که در CpG island باشد.

ب. محتمل‌تر است که در Normal DNA باشد.

ج. احتمال آن که در CpG island باشد برابر است با احتمال آن که در Normal DNA باشد.

د. اطلاعات داده شده برای پاسخ دادن به این سوال که کدام مدل محتمل‌تر است کافی نیست.

**۲۴. دو الاینمنت زیر را در نظر بگیرید. با نمره‌دهی match = +1، mismatch = -1 برای چه مقداری از g در linear gap الاینمنت i نمره بهتری نسبت به الاینمنت ii دارد؟**

**i.** **–CGCATG**
**ACG–A–G**

**ii.** **CGCATG**
**ACGA–G**

آ. ۱-  ب. ۲-  ج. ۳-  د. ۴-