دانشکده‌ی مهندسی کامپیوتر

مقدمه‌ای بر بیوانفورماتیک

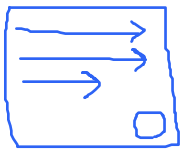
نیم‌سال اول ۱۴۰۰ـ۱۳۹۹

مدرس: دکتر شریفی زارچی، دکتر کوهی

میانترم

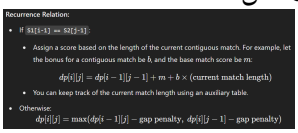
مسئله‌ی ۱.

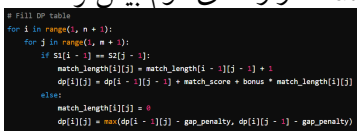
دو رشته‌ی ACGTCATCA و TAGTGTCA را در اختیار داریم. جدول alignment این دو رشته را تشکیل دهید و تغییرات لازم در رشته‌ی اول برای این‌که به رشته‌ی دوم تبدیل شود را ذکر کنید.



مسئله‌ی ۲.

در مسئله‌ی alignment فرض کنید که هرچه‌قدر نواحی مشابه طولانی‌تر باشند برای ما مطلوب‌تر است. برای مثال در align کردن دو رشته‌ی CAGCACTTGGATTCTCGG و CAGCGTGG ، رشته‌ی تطبیق‌یافته‌ی CAGC - - - - - G - T - - - - GG - - - CAGCGTGG - - - - - - - ترجیح می‌دهیم. با این‌که تعداد match در رشته‌ی دوم بیش‌تر است. راه حلی ارائه دهید که بتواند این مسئله را حل کند.





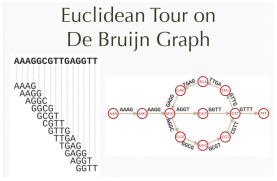
مسئله‌ی ۳.

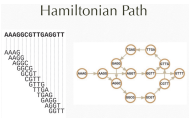
با استفاده از گراف *de Bruijn* اویلری کوتاه‌ترین رشته‌ای که شامل تمامی زیررشته‌های زیر باشد را بیابید.

{ACGT, AGAC, CGTG, GACG, GGTA, GTAC, GTAG, GTGT, TACG, TAGA, TGTA}

سپس توضیح دهید در صورتی که در هنگام بازسازی یک ژنوم از k-merهای آن، بیش از یک رشته با کوتاه‌ترین طول که شامل همه‌ی زیررشته‌ها باشد وجود داشته باشد، چه گونه می‌توان رشته‌ی مربوط به ژنوم را از سایر رشته‌ها تشخیص داد؟

https://chatgpt.com/c/674e0d73-b488-8005-b71d-9ec4b6034ff4





مسئله‌ی ۴.

یک رشته پس از اعمال تبدیل BWT [۱] به فرم ATTAT$G درآمده‌است، ابتدا رشته‌ی اولیه را بدست آورید سپس توضیح دهید که چگونه می‌توان با انجام پیش‌پردازش بر روی تبدیل BWT این رشته، تعداد تکرارهای زیررشته‌ی TA در رشته‌ی اصلی را محاسبه کرد.

https://chatgpt.com/c/674e0d73-b488-8005-b71d-9ec4b6034ff4

last to first
count array

s = TAGTTA$

مسئله‌ی ۵.

آ. الگوریتم additive phylogeny را مرحله به مرحله روی ماتریس فاصله‌ی زیر اجرا کنید.

	v_1	v_2	v_3	v_4
v_1	۰	۹	۱۰	۹
v_2	۹	۰	۹	۸
v_3	۱۰	۹	۰	۷
v_4	۹	۸	۷	۰

Introducing *Count* Array

i	*FirstColumn*	*LastColumn*	*LastToFirst(i)*	Count					
				$	a	b	m	n	p	s
0	$1	s1	13	0	0	0	0	0	0	0
1	a1	m1	8	0	0	0	0	0	0	1
2	a2	n1	9	0	0	0	1	0	0	1
3	a3	p1	12	0	0	0	1	1	0	1
4	a4	b1	7	0	0	0	1	1	1	1
5	a5	n2	10	0	0	1	1	1	1	1
6	a6	n3	11	0	0	1	1	2	1	1
7	b1	a1	1	0	0	1	1	3	1	1
8	m1	a2	2	0	1	1	1	3	1	1
9	n1	a3	3	0	2	1	1	3	1	1
10	n2	a4	4	0	3	1	1	3	1	1
11	n3	a5	5	0	4	1	1	3	1	1
12	p1	$1	0	0	5	1	1	3	1	1
13	s1	a6	6	1	5	1	1	3	1	1
				1	6	1	1	3	1	1

$Count_{symbol}(i, LastColumn)$:
#occurrences of *symbol* in the first *i* positions of *LastColumn*

[۱] Burrows–Wheeler Transform

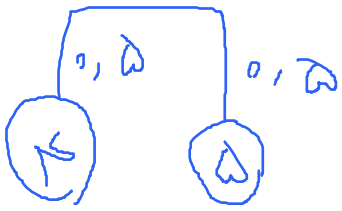
ب. وزن یال‌ها در درخت فیلوژنی می‌تواند نشان‌دهنده‌ی چه چیزی باشد؟

مسئله‌ی ۶.

آ. الگوریتم UPGMA را مرحله به مرحله روی ماتریس فاصله‌ی زیر اجرا کنید.سوال 6 تمرین 2

	v_1	v_2	v_3	v_4	v_5
v_1	۰	۲	۵	۴	۳
v_2	۲	۰	۴	۳	۵
v_3	۵	۴	۰	۲	۳
v_4	۴	۳	۲	۰	۱
v_5	۳	۵	۳	۱	۰

not ultrametric
سرعت تکامل گونه ها یکی نباشد



ب. الگوریتم UPGMA در چه صورتی ممکن است درختی اشتباه را به عنوان خروجی تولید کند؟

The UPGMA algorithm (Unweighted Pair Group Method with Arithmetic Mean) can produce an incorrect phylogenetic tree due to its underlying assumptions and limitations. These issues arise primarily because UPGMA assumes a molecular clock hypothesis, which means it expects all lineages to evolve at the same constant rate. If this assumption is violated or there are inconsistencies in the input data, UPGMA may construct a tree that does not accurately represent the evolutionary relationships.

UPGMA assumes that evolutionary distances are ultrametric, meaning the distance from the root to any leaf is the same for all leaves.If some lineages(دودمان) evolve faster than others (rate heterogeneity), the ultrametric property is violated, and UPGMA fails to correctly infer relationships.
For example:Consider three taxa: A, B, and C.If A and B have evolved slowly and C has evolved rapidly, UPGMA might group A and B correctly but place C incorrectly because it interprets the greater distance to C as a later divergence, rather than faster evolution.

UPGMA assumes that pairwise distances are additive (i.e., the distance between two taxa equals the sum of the branch lengths connecting them).If distances are not strictly additive (e.g., due to homoplasy or back mutations), UPGMA may produce an inaccurate tree.

UPGMA clusters taxa step-by-step, and once two taxa or groups are merged, the decision cannot be undone.If an incorrect merge occurs early, subsequent steps propagate the error, producing a globally incorrect tree.