دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

مقدمه‌ای بر بیوانفورماتیک

پاییز ۱۳۹۹

علی شریفی‌زارچی، سمیه کوهی

آزمون پایانترم

مهلت ارسال: ساعت ۱۲:۱۵

سوالات (۱۰۰ + ۵ نمره)

۱. (۱۰ نمره) رشته‌ی x را یک سوپررشته برای رشته‌ی v می‌نامیم هرگاه رشته‌ی v یک زیر رشته از رشته‌ی x باشد. (یادآوری: رشته‌ی s زیررشته‌ی رشته‌ی t نامیده می‌شود هرگاه s با حذف صفر یا تعداد بیش‌تری کاراکتر از رشته‌ی t بدون تغییر ترتیب کاراکترها به دست آید.)
با داشتن دو رشته‌ی v و w یک الگوریتم dynamic programming ارائه دهید که در زمان $O(mn)$ کوتاه‌ترین سوپررشته‌ی این دو رشته را پیدا کند.

۲. (۱۵ نمره) همانطور که می‌دانید یکی از روش‌های بازسازی ژنوم پیدا کردن دور همیلتنی در گراف k-merهای آن ژنوم است. علت استفاده نشدن از این روش را توضیح دهید، سپس با استفاده از گراف de Bruijn اویلری کوتاه‌ترین رشته‌ای که شامل تمامی زیررشته‌های زیر باشد را بیابید.

{AATAGA, ACGTAG, ACGTGGTA, GACGT, GGTAG, GTAGA, GTAGT, TACGT, TAGAATA, TAGTACG}

۳. (۱۵ نمره) ماتریس فاصله‌ی زیر را در نظر بگیرید.

	$v_۱$	$v_۲$	$v_۳$	$v_۴$
$v_۱$	۰	۵	۵	۴
$v_۲$	۵	۰	۹	۷
$v_۳$	۵	۹	۰	۴
$v_۴$	۴	۷	۴	۰

آ. از میان دو الگوریتم Additive Phylogeny و UPGMA کدام یک برای پیدا کردن درخت فیلوژنی این ماتریس مناسب است؟ چرا؟

ب. الگوریتم انتخابی را مرحله به مرحله روی ماتریس فاصله‌ی بالا اجرا کنید.

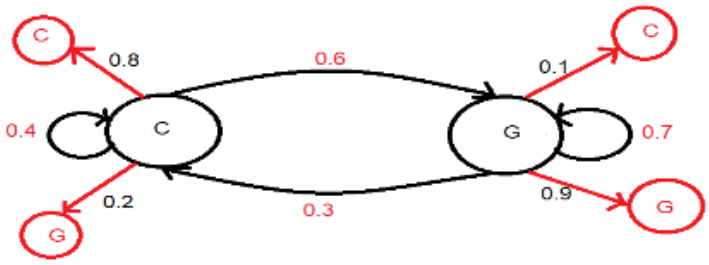
۴. (۱۰ نمره) با کمک MedianString یک 3-mer پیدا کنید که d(Pattern, DNA) را از میان تمامی 3-merها کمینه کند. (توجه داشته باشید که باید راه حل کامل باشد و اگر چند جواب دارد یک جواب کافی‌است. منظور از راه حل کامل نوشتن همه ۶۴ حالت ممکن نیست. باید تشخیص دهید کمینه جواب چیست و یک حالت آنرا بیان کنید.)

ATAA
ACAC
AGGC
GATT
AACC

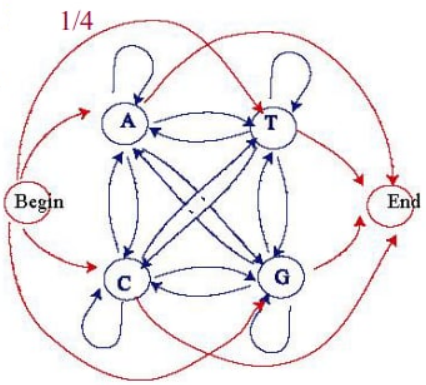
۵. (۱۰ نمره) مسئله Motif Finding و Gold Bug Problem را از دو منظر شباهت و تفاوت مقایسه کنید.

۶. (۱۵ نمره) دستگاه‌های توالی‌یاب، معمولا دارای درصدی خطا هستند. به این معنی که توالی‌هایی که توسط این دستگاه‌ها نمایش داده می‌شود، لزوما توالی اصلی نخواهد بود. این خطا را می‌توان با استفاده‌از یک HMM مدل کرد. در صورتی که در یک ناحیه‌ی خاص از ژن (CpG island) تنها دو باز C و G مشاهده شود و احتمال خطا و transition probability بین این دو باز نیز با روابط زیر مشخص شده‌باشد و این ناحیه را با یک HMM دارای ۲ عدد hidden state مدل کرده‌باشیم، و توالی «CGGCG» را ببینیم، محتمل‌ترین توالی اصلی چیست؟

P(transition from G to C) = 0.3 P(observing C | C) = 0.8
P(transition from G to G) = 0.7 P(observing G |C) = 0.2
P(transition from C to C) = 0.4 P(observing C | G) = 0.1
P(transition from C to G) = 0.6 P(observing G|G) = 0.9
P(start from C) = 0.5 P(start from G) = 0.5



۷. (۱۵ نمره) یکی از کاربردهای زنجیره‌ی مارکوف در زیست‌شناسی، بررسی احتمال وجود نواحی CpG island در توالی‌هاست. در صورتی که دو ماتریس زیر، نشان‌دهنده‌ی transition probability بین بازهای مختلف در نواحی معمولی و نواحی CpG island باشد، احتمال این که توالی «AATCGT» مربوط به یک CpG island باشد را به دست آورید. (احتمال آغاز توالی از هر یک از بازهای A،C،G و T را در هر دو حالت 1/4 در نظر بگیرید.)



CpG	A	C	G	T
A	0.2	0.3	0.4	0.1
C	0.1	0.4	0.3	0.2
G	0.1	0.4	0.4	0.1
T	0.1	0.3	0.4	0.2

general	A	C	G	T
A	0.3	0.2	0.3	0.2
C	0.3	0.3	0.1	0.3
G	0.2	0.3	0.3	0.2
T	0.2	0.2	0.3	0.3

۸. (۱۵ نمره) وجود پیوندهای هیدروژنی بین دو رشته DNA به پایداری آن کمک شایانی کرده است. از طرفی انرژی پیوند هیدروژنی به طور قابل ملاحظه‌ای از انرژی پیوندهای بین اتمی (کووالانسی) کمتر است. مزیت این کمتر بودن نسبی انرژی پیوندهای هیدروژنی در ساختار DNA کجاست؟