الگوریتم‌های احتمالی عضویت – تمرین ۱

۱- الگوریتم ادغام و الگوریتم افزایش اندازهٔ فیلتر خارج‌قسمت پیاده و تحلیل شود. مستندات و استفاده از مجموعه داده الزامی است.

۲- یکی از دو مورد زیر را انجام دهید:

I- روش بررسی املا متن با استفاده از فیلتر بلوم پیاده شود. جهت پیشبرد کار یافتن مجموعه داده مناسب از لغات فارسی و ایجاد محیطی که متن را خوانده یا بارگذاری کرده اهمیت دارد. ایجاد مجموعه داده و فیلتر مذکور، و سپس بررسی متن از اهم موارد است.

II- استفاده از فیلترهای عضویت برای جستجوی k-mer در توالی‌های ژنومی. ژنوم‌چینی فرآیند بازسازی توالی کامل و پیوسته ژنوم اصلی موجود زنده از قطعاتی پراکنده است، که می‌توان آن را به حل پازلی بسیار بزرگ تشبیه کرد.

با الهام از روش‌هایی که در ابزارهای ژنوم‌چینی[۱] مانند مینیا و بایفراست (مقالات به پیوست هستند) استفاده می‌شوند. در ابزارهای مذکور میلیاردها قطعه کوتاه دی‌ان‌ای به نام $k-mer$ ذخیره و جستجو می‌شوند، بنابراین استفاده از ساختارهای حافظه‌کارا مانند فیلتر بلوم ضروری است.

مفاهیم اولیه

۱. توالی دی‌ان‌ای: ژنوم موجود زنده رشته‌ای طولانی متشکل از چهار نوکلئوتید A, C, G, T است. مثلا، توالی کوتاه ACGTTGCAACGT نمونه‌ای از رشته حاصل از چهار نویسه مذکور است. k-mer به تمام زیررشته‌های طول k از توالی دی‌ان‌ای گفته می‌شود. مثلا، اگر k = ٤ باشد و توالی ACGTTGCA را در نظر بگیریم k-mer-های آن عبارتند از: ACGT، CGTT، GTTG، TTGC، TGCA.

در بسیاری از الگوریتم‌های ژنومی، داده‌ها به شکل مجموعه‌ای از k-merها ذخیره و پردازش می‌شوند. در اینجا پای فیلتر بلوم به میان می‌آید. در پیوست قطعه کوچکی از دی‌ان‌ای در قالب فستا[۲] داده قرار گرفته است. حال

الف- بخش اول: استخراج k-merها. ۱. مقدار k را برابر ۱۵ در نظر بگیرید. ۲. از روی توالی داده شده، تمام k-merها را استخراج کنید.

ب- بخش دوم: پیاده‌سازی فیلتر بلوم. یک فیلتر بلوم بسازید.

ج- بخش سوم: پرس‌وجوی[۳] عضویت-یک مجموعه از k-mer های آزمایشی بسازید که شامل دو نوع باشد:

۱. واقعاً در توالی وجود دارند. ۲. به صورت تصادفی ساخته شده‌اند. برای هر پرس‌وجو بررسی کنید فیلتر بلوم چه پاسخی می‌دهد.

د- بخش چهارم: اندازه‌گیری خطا: برای مجموعه پرس‌وجوها موارد تعدادو نسبت مثبت صادق و تعداد مثبت کاذب را محاسبه کنید.

هـ بخش پنجم: آزمایش با پارامترهای مختلف- مقادیر مختلف اندازه k-mer و فیلتر بلوم را امتحان کنید و بررسی کنید نسبت خطا و مصرف حافظه چگونه تغییر می‌کند.

ز-فرض کنید دو نمونه ژنومی داریم. ۱. فیلتر بلوم را با k-mer های نمونه A بسازید. ۲. k-merهای نمونه B را پرس‌وجو کنید. ۳. درصد k-merهایی که در فیلتر بلوم پیدا می‌شوند را محاسبه کنید. این مقدار تخمینی از شباهت دو توالی است.

خروجی‌های تمرین عبارتند از

- کد پیاده‌سازی فیلتر بلوم
- k-mer-های استخراج شده
- نسبت مثبت کاذب
- نتایج آزمایش با پارامترهای مختلف

---

[۱] Genome assembly
[۲] FASTA
[۳] query

- توضیح درباره نتایج مشاهده‌شده

استفاده از کتابخانه‌های آماده به جز توابع درهم مجاز نیست. جهت پیشبرد کار توالی ژنومی ویروس سیرکو و مقالات مینیا و بایفراست پیوست شده‌اند.

۳- بیت‌های متاداده فیلتر خارج‌قسمت دارای هشت حال هستند. هر یک از هشت حالت را بررسی کنید.

۴- با استفاده از منابع عملکرد فیلترهای بلوم، بلوم شمارنده، خارج‌قسمت، و فاخته را در درج، آزمون عضویت، حذف، تغییر اندازه، و حافظه بررسی کنید. ب- موارد را برای مجموعه داده تمرین دو آزمایش کنید.

۵- بر اساس میزان پر بودن حافظه (ضریب بارگذاری) عملکرد فیلترها را برای درج، آزمون عضویت تصادفی و جستجوهای موفق تحلیل کنید. رسم نمودار و تحلیل توصیه می‌شود.

۶- الف- در الگوریتم خارج‌قمست میزان احتمال مثبت کاذب و همچنین نحوه محاسبه طول باقیمانده گزارش شد. به صورت تحلیل معادلات مذکور را بدست آورید. ب- در فیلتر فاخته احتمال مثبت کاذب، طول اثر انگشت و کران پائین طول فیلتر گزارش شد. مقادیر آنها را به صورت تحلیل به دست آورید.

۷- تحلیل زمانی اجرای الگوریتم‌های درج و آزمون عضویت را برای فیلترهای بلوم و خارج‌قسمت و فاخته گزارش کنید. نحوه سنجش مرتبه زمانی باید در ابتدا به روشنی مشخص و سپس مراحل الگوریتم‌ها تحلیل شوند.

۸- با وجود جابجایی‌های فراوان در حین درج مقداری جدید که هر دو جای آن اشغال شده است، آیا آزمون عضویت الگوریتم فاخته به درستی کار می‌کند؟ چرا؟

تمرین‌ها در گروه‌های یک یا دو نفره ممکن است. مراجع در صورت استفاده باید دقیق باشد و صرفا به مدل‌های زبانی تکیه نشود.