نشست دو روزهٔ حل مسأله در بیوانفورماتیک

در روزهای ۲۲ و ۲۳ دی ماه ۸۴ یک دورهٔ دو روزهٔ حل مسأله به پیشنهاد رئیس پژوهشکدهٔ علوم کامپیوتر، دکتر حمید سربازی آزاد و به سرپرستی دکتر روزبه توسرکانی در پژوهشگاه دانش‌های بنیادی برگزار شد که هدف از آن، کسب یک تجربهٔ جدید در استفاده از توان دانشجویان برجستهٔ علوم ریاضی و کامپیوتر برای حل یک مسأله باز و مهم روز بود. با توجه به اهمیت زیاد بیوانفورماتیک به عنوان یک رشتهٔ جدید و در حال رشد و همچنین ماهیت بین‌رشته‌ای آن که به علوم ریاضی و کامپیوتر وابسته است، بررسی یک مسألهٔ بیوانفورماتیک و تلاش برای حل آن به عنوان موضوع فعالیت انتخاب شد. بدین منظور از دکتر مهدی صادقی از پژوهشگاه ملی مهندسی ژنتیک و دکتر چنگیز اصلاحچی از دانشکدهٔ ریاضی دانشگاه شهید بهشتی که در این زمینه فعالیت می‌کنند، دعوت شد تا با طرح یک مسأله، با این دانشجویان در جهت پیدا کردن راه حل آن همکاری کنند. مسأله‌ای با عنوان بخش‌بندی بلوک‌های هاپلوتیپی (Haplotype Block Partitioning) برای دانشجویان مطرح شد که می‌توان آن را به‌طور خلاصه به‌شرح زیر بیان کرد.

مادهٔ ژنتیکی یا DNA را می‌توان به‌صورت یک رشتهٔ پیوسته از چهار مولکول متفاوت (نوکلئوتید) که با نمادهای A، C، G، T مشخص می‌شوند تعریف کرد. طول این رشته در انسان حدود ۳ میلیارد است که در ۲۳ جفت کروموزوم با اندازه‌های مختلف تقسیم شده است. در واقع هر کروموزوم یک زنجیرهٔ بسیار طویل با اندازهٔ متوسط حدود ۱۵۰ میلیون نوکلئوتید است.

نشان داده شده است که در افراد مختلف حدود سه میلیون موقعیت در رشتهٔ DNA وجود دارد که در این موقعیت‌ها، نوع نوکلئوتید قرارگرفته در انسان‌ها متفاوت است و در بقیه موقعیت‌ها (۹۹٫۹٪) نوکلئوتیدها شبیه به هم هستند. همچنین نشان داده شده است که در هر کدام از این موقعیت‌ها فقط دو نوکلئوتید از چهار نوکلئوتید می‌توانند قرار گیرند. این موقعیت‌ها چندریختیِ تک نوکلئوتیدی (Single Nucleotide Polymorphism) یا به اختصار SNP نامیده می‌شوند و دو نوکلئوتید قرارگرفته در این موقعیت‌ها آلل(Allele) نام دارند. بنابراین، چنانچه ترتیب سه میلیارد نوکلئوتید در یک فرد شناخته شده باشد در افراد دیگر فقط با دانستن نوکلئوتید قرارگرفته در موقعیت‌های ترتیب کل توالی نوکلئوتیدها شناخته خواهد شد. از آنجایی که در هر فرد در موقعیت‌های SNP یکی از دو نوکلئوتید می‌تواند قرارگیرد، تعداد کل توالی‌های متفاوت ممکن در انسان 2^n است که n تعداد این موقعیت‌های SNP است. ترتیب نوکلئوتیدها در این موقعیت‌ها در یک فرد، هاپلوتیپ آن فرد نامیده می‌شود. از راه‌های تجربی و نظری نشان داده شده است که در بین انسان‌ها همهٔ این 2^n تنوع نوکلئوتیدی وجود ندارد و تعداد آنها بسیار محدود است. به دلایل تجربی مشخص شده که هاپلوتیپ‌ها را می توان به بلوک‌هایی با تنوع محدودی از هاپلوتیپ‌ها تقسیم‌بندی کرد. در هر بلوک تعداد کمی از SNPها (SNPهای نماینده) کافی خواهند بود که به‌طور منحصر به‌فرد هاپلوتیپ‌های مشترک را تشخیص دهند.

مسألهٔ اصلی این است: چگونه می‌توان این توالی SNPها را به تعدادی بلوک تقسیم‌بندی کرد به‌طوری که تعداد SNPهای نماینده مینیمم باشد.

برای حل مسأله چند شرط را باید در نظر گرفت. هاپلوتیپ‌های مشترک، هاپلوتیپ‌هایی هستند که بیشتر از یک بار در یک بلوک دیده شوند. بنابراین ما نیاز داریم که در نتیجهٔ نهایی تقسیم‌بندی، بخش قابل توجهی از هاپلوتیپ‌ها در هر بلوک، هاپلوتیپ‌های مشترک باشند (مثلاً ۸۰٪). فرض کنید که k هاپلوتیپ (از k فرد مختلف) با n SNP توالی داریم. $r_i, i = ۱, ۲, \ldots, n$ یک بردار k بعدی است که $r_i(k) = ۰, ۱, ۲$ آلل kامین هاپلوتیپ در موقعیت SNP، i است که ۱ و ۲ دو آلل مختلف است و صفر به معنای آن است که وضعیت آلل در آن موقعیت مشخص نشده است (به دلایل تجربی ممکن است نتوان نوکلئوتید قرارگرفته در یک موقعیت را تشخیص داد).

یک بلوک با ترتیب $r_i, \ldots, r_j$ تعریف می‌شود. دو هاپلوتیپ kوk' باهم سازگار گفته می‌شوند اگر آلل‌ها در دو هاپلوتیپ در موقعیت‌هایی که صفر نباشند، یکسان باشند. به‌عبارت دیگر: به ازای هر l، $i \le l \le j$، به‌طوری‌که $r_j(k)r_l(k) \neq ۰$، $r_i(k) = r_i(k')$. یک هاپلوتیپ در یک بلوک «مبهم» است اگر با دو هاپلوتیپی که باهم ناسازگار هستند سازگار باشد. با در نظر گرفتن شرط سازگار بودن، ما می‌توانیم هاپلوتیپ «نامبهم» را به گروه‌های مجزا یا بلوک‌ها تقسیم کنیم. دو هاپلوتیپ در یک گروه خواهند بود اگر باهم سازگار باشند. با تعریف ذکر شده دربارهٔ هاپلوتیپ‌های نامبهم می‌توان تابع $\text{block}(r_i, \ldots, r_j) = ۱$ را تعریف کرد، اگر حداقل α درصد (مثلا ۸۰ درصد) از هاپلوتیپ‌های نامبهم در بلوک بیشتر از یک بار ظاهر شده باشند. بنابراین بلوک‌های به‌دست آمده در تقسیم‌بندی نهایی باید این شرایط ذکر شده را داشته باشند. همان‌گونه که قبلاً ذکر شد، می‌خواهیم تقسیم‌بندی اپتیمم بلوک‌ها را پیدا کنیم (مثلا $B_r, \ldots, B_l$) به‌طوری‌که تعداد کل SNPهای نماینده $(f(B_۱) + \ldots + f(B_l))$ مینیمم باشد.

در طول دو روز برگزاری این نشست، الگوریتم‌های متعددی که توسط دانشجویان مطرح، پیاده سازی و بررسی شد که متأسفانه هیچ کدام از آنها به حل درستی از مسأله منجر نشد. پس از آن تعدادی از دانشجویان کار

بر روی حل مسأله را ادامه دادند که نهایتاً یک فرمول‌بندی بر اساس نظریهٔ اطلاعات به نتایج قابل توجهی منجر شد که کار بر روی آن و استخراج اطلاعات از یک پایگاه داده واقعی مربوط به ژنوم انسان و نهایتاً تهیهٔ یک نرم افزار مناسب توسط امین صادقی، دانشجوی کارشناسی مهندسی کامپیوتر، دانشگاه صنعتی شریف در حال انجام است که انتظار می‌رود به‌زودی تکمیل شود.

چه کشوری سریعترین رشد را در تولیدات علمی دارد؟

ممکن است بگویید چین؛ ولی اشتباه می‌کنید. طبق تحقیقی که اخیراً مؤسسهٔ بازرگانی و صنعت بریتانیا (DTI) دربارهٔ رتبه‌بندی کشورها از لحاظ علمی انجام داده، کشور دیگری حائز این رتبه است. در این بررسی، انواع شاخص‌های ورودی و خروجی برای به‌دست آوردن تصویر نسبتاً کاملی از توزیع فعالیت و تأثیر علمی در سراسر جهان مورد توجه قرارگرفته است. علی‌رغم پاسخ تعجب‌آوری که عنوان این نوشته دارد (و در پایان مطلب آن را افشا خواهم کرد)، این گزارش رشد سریع چین را به مثابهٔ یک قدرت علمی که برتریِ سابقاً بی‌چون و چرایِ آمریکا را به چالش طلبیده، تأیید می‌کند و نیز (در چشم‌انداز محدود ملی) مؤید عملکرد نسبتاً قوی بریتانیاست که نسبت به رقبایش پول کمتری صرف تحقیق می‌کند و پژوهشگران کمتری دارد ولی در مقایسه، دانش بیشتری با تأثیر زیاد تولید می‌کند.

در زمینهٔ بودجهٔ تحقیقات علمی، صعود چین بیش از هر کشوری چشمگیر است. بودجهٔ پژوهشی چین (با در نظر گرفتن ملاحظات مربوط به قدرت خرید) طی دههٔ گذشته چهار برابر شده و اکنون بیشتر از هر کشوری در جهان بجز آمریکا و ژاپن است و به نصف کل بودجهٔ تحقیقات در اتحادیهٔ اروپا رسیده است. از لحاظ میزان انتشارات علمی، سهم چین ۵٪ کل جهان است و در دههٔ گذشته سه برابر شده و امروز بیشتر از فرانسه است. اگر هر کشوری جداگانه در نظر گرفته شود، ایالات متحدهٔ آمریکا هنوز هم درصدر قرار دارد و تقریباً یک سوم کل محصول علمی جهان را تولید می‌کند، ولی اتحادیهٔ اروپا در مجموع از آمریکا پیشی گرفته و سهم آن ۳۷/۹٪ از کل انتشارات است. بریتانیا با حدود ۹٪ مقام دوم را در بین تک تک کشورها دارد و اخیراً ژاپن را پشت سر نهاده است. گروه کشورهای آسیاپاسیفیک (چین، کره، تایوان، و سنگاپور) روی هم رفته ۱۰٪ از تولید علمی جهان را در دست دارند.

و اما دربارهٔ کیفیت و تأثیر. در این مورد آمریکا هنوز به وضوح پیشگام است؛ اگر سهم هر کشور از پر ارجاع‌ترین مقاله‌ها (با در نظر گرفتن ۱% بالا در هر رشته) به‌عنوان شاخصی از تأثیر جهانی قلمداد شود، آمریکا با ۶۱% در رأس قرار دارد. بریتانیا سهم خود را از مقاله‌های پر ارجاع افزایش داده و به ۱۳% رسانده است. چین از این لحاظ هنوز نسبت به صدرنشینان عقب‌تر است ولی فاصله‌اش از آنها در حال کم شدن است به خصوص که شمار ارجاعات، یک شاخص تأخیری است یعنی مدتی طول می‌کشد تا ثمرهٔ صرف بودجه در تحقیقات به صورت انتشار مقاله‌ها ظاهر شود و پس از آن، پژوهشگران دیگر به آن مقاله‌ها استناد کنند.

کشوری که رشد انتشارات علمی آن طی دههٔ گذشته بیش از همه بوده، ایران است که میزان تولید آن ده برابر شده است (هرچند نسبت به میزانی اندک در آغاز دهه). باید دید که آیا با توجه به تحولات سیاسی اخیر، ایران می‌تواند همچنان به عملکرد خوب خود در این زمینه ادامه دهد یا خیر.

منبع:

http://www.softmachines.org/wordpress/