Szekvencia igazításról és a BLAST algoritmusról

Szörényi Balázs

A feladat egy analizálandó input q szekvencia (pl. DNS szekvenciák vagy fehérjeláncok) összehasonlítása az adatbázisunkban található minden szekvenciával. Ennek során arra keresünk választ, hogy q-nak és egy adatbázisbeli d szekvenciának milyen "hasonló" részei vannak. (A tapasztalat azt mutatja ugyanis, hogy a "nagy hasonlóságot" mutató részek funkciója is nagyon hasonló, ezért ennek a problémának hatalmas a jelentősége a bioinformatikában pl. génkutatásoknál.) A hasonlóságot valamilyen M illeszkedési táblázat (avagy hasonlósági mátrix, pl. a PAM vagy a BLOSUM mátrixcsaládokat) segítségével szokás meghatározni, ami minden lehetséges szimbólumpárra (pl. a bennünket érdeklő fehérje párokra) megad egy "igazítási jóságot", ami lényegében egy tapasztalati/intuitív érték arra vonatkozóan, hogy mennyire felel meg a két szimbólum egymásnak. Ezen felül, amennyiben szimbólumok törlését (vagy másként: rések beillesztését) is megengedjük, szükséges még egy g érték, amivel ezen törlésekhez (avagy résekhez) rendelünk egy pontszámot.

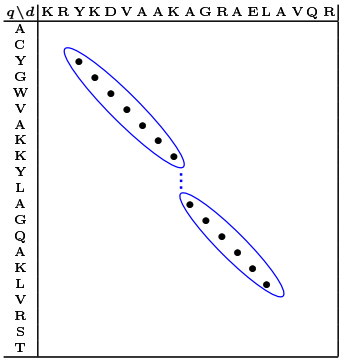
Tekintsük például a q=`ACYGWVAKKYLAGQAKLVRST...` harmadik szimbólumától a tizenhetedik szimbólumáig tartó $\hat{q}$ és a d=`KRYKDVAAKAGRAELAVQR...` harmadik szimbólumától a tizenhatodik szimbólumáig tartó $\hat{d}$ részének egy lehetséges igazítását:

$\hat{q}'$	=	Y	G	W	V	A	K	K	Y	L	A	G	Q	A	K	L
$\hat{d}'$	=	Y	K	D	V	A	A	K	-	-	A	G	R	A	E	L
		10	-2	-7	4	2	-1	5	-3	-3	2	5	1	5	0	6

Az alsó sorban az egyes szimbólumilleszkedéseknek a PAM250 mátrix alapján számított értéke szerepel (lásd pl. Dayhoff és szerzőtársai eredeti cikkét [1]), g értékeként pedig 3-at használtunk. Ennek az igazításnak az összértéke

$$\text{ertek}(\hat{q}', \hat{d}') = \sum_{k=1}^{|\hat{d}'|} \text{ertek}(\hat{q}'_k, \hat{d}'_k) = \sum_{k=1}^{|\hat{q}'|} \text{ertek}(\hat{q}'_k, \hat{d}'_k) = 24,$$

ahol két szimbólum esetén ertek az M mátrix által meghatározott értéket adja vissza, egy rés szimbólum (azaz "-") esetén pedig a $-g$ értéket. Az ilyen igazításokat grafikusan is reprezentálhatjuk az alábbi módon:



Nyilván annál jobbnak tartunk egy igazítást, minél magasabb az értéke.

1 A megoldandó feladat formálisan

Egy $\mathcal{C}$ véges ábécé feletti $q \in \mathcal{C}^m$ és $d \in \mathcal{C}^n$ szekvenciák (szimbólumsorozatok) *igazítása* alatt egy olyan q', d', $\mathcal{C} \cup \{\texttt{-}\}$ feletti, azonos hosszúságú szekvenciapárost értünk, melyekből kitörölve a `"-"` szimbólumot visszakapjuk a q illetve d szekvenciákat.

Először tekintsünk egy egyszerűbb feladatot, amikor nem részszavakat kell igazítani, hanem a teljes szekvenciákat.

SzóIgazítás	
Adott:	$\mathcal{C}$ véges ábécé, $q \in \mathcal{C}^m$ input (analizálandó) szekvencia, $d \in \mathcal{C}^n$ (adatbázisbeli szekvencia), $M : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ hasonlósági értéktábla, $g \in \mathbb{R}$ résbüntetés.
Kell:	q és d azon igazítása, melynek M és g alapján számított értéke a lehető legnagyobb

A $T(i,j) = \max\{T(i-1,j) - g, T(i,j-1) - g, T(i-1,j-1) + M(q_i, q_j)\}$ rekurzió alapján dinamikus programozással megoldható a fenti probléma.

LokálisSzóIgazítás	
Adott:	$\mathcal{C}$ véges ábécé, $q \in \mathcal{C}^m$ input (analizálandó) szekvencia, $d \in \mathcal{C}^n$ (adatbázisbeli szekvencia), $M : \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ hasonlósági értéktábla, $g \in \mathbb{R}$ résbüntetés.
Kell:	q illetve d minden lehetséges résszavának minden lehetséges igazítása közül az, melynek M és g alapján számított értéke maximális

A fentihez hasonlóan ez a probléma is megoldható dinamikus programozással a $T(i,j) = \max\{T(i-1,j) - g, T(i,j-1) - g, T(i-1,j-1) + M(q_i, q_j), 0\}$ rekurzió alapján. (Ez a *Smith-Waterman algoritmus*.)

2 Egy heurisztikus algoritmus: a `gapped-BLAST`

Habár az előző szekció mindkét problémájára $\Theta(mn)$ futásidejű pontos algoritmust adtunk, a gyakorlatban sokszor még ez is lassúnak bizonyul. Ezért szükségessé vált olyan *heurisztikus* algoritmusok kifejlesztése, melyek ugyan nem adják vissza mindig az összes megfelelő illeszkedést, viszont futásidejük már a gyakorlati alkalmazások számára is elfogadható. Ilyenek például a FAST és a BLAST algoritmus családok, melyek megfelelő egyensúlyt találtak a gyorsaság és a pontosság között.

Az alábbiakban a BLAST egyik tagját, a `gapped-BLAST` algoritmust [2] tekintjük át vázlatosan. Az algoritmus olyan igazítás(oka)t keres (ha léteznek), aminek a M és g alapján számított értéke meghaladja egy (paraméterként megadott) V küszöbértéket.

1. lépés. Hasonló, rövid (előre rögzített w hosszúságú) *egybefüggő* részszavak keresése (proteinek esetében pl. $w = 3$ használható), azaz valamely előre rögzített T paraméter melletti hits $= \{(i,j) : \sum_{k=0}^{w-1} M(q_{i+k}, d_{j+k}) \geq T\}$ halmaz meghatározása. (Ehhez mintaillesztési módszerek szükségesek, melyhez tipikusan auomataelméleti eszközöket használnak.)

2. lépés. hits-beli elempárok összefűzhetőségének vizsgálata: két, egymás utáni, legfeljebb egy előre meghatározott A (pl. $A = 40$) távolságra levő ilyen rövid igazítás kibővíthető-e egy (még mindig résmentes, azaz egybefüggő) viszonylag magas pontszámú (mostmár hosszabb) igazítássá. Az így kapott kibővített igazításokat hívjuk *HSP*-knek (High-scoring Segment Pair). Ez a grafikus reprezentációban két, ugyanazon átlón közvetlenül egymást követően elhelyezkedő hits-beli elem összefűzhetőségeként valósul meg.

3. lépés. Az előző pontban talált, elég nagy értékű (egy S_g értéket meghaladó) HSP-k megfelelő kiigazítása (lokális maximum keresése). Ennek során az aktuális HSP egy megfelelő pontját[1] lerögzítjük, és ebből indítunk előrefele és hátrafele is egy-egy keresést a legjobb lokális igazítás megtalálására dinamikus programozási módszereket használva. Az S_g paraméter úgy kerül beállításra, hogy az annál nagyobb értékű HSP-kből nagyjából 50 adatbázisbeli elemben forduljon elő egy.

References

[1] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3*, chapter 22, pages 345–352. National Biomedical Research Foundation, 1978. `http://ibi.zju.edu.cn/bioinplant/courses/dayhoffetal1978.pdf`

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997. `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC146917/`

[1] Ez az úgynevezett *seed*, ami konkrétan az igazítás 11 hosszú egybefüggő részszavai közül a legmagasabb pontszámúnak a középpontja.