



Sequenzvergleich – wo kommt unser Genom her?

Seit dem Human Genome Project hat das Problem der Genomsequenzierung, also der Entschlüsselung des (menschlichen) Erbguts, die Nachrichten und Zeitungen erreicht. Dieses grundlegende biologische Problem ist ohne Informatikmethoden nicht zu lösen, so dass sich ein neues Gebiet der Informatik gebildet hat. In Deutschland hat sich hierfür die Bezeichnung Bioinformatik durchgesetzt, wobei aber der Begriff „computational molecular biology“ treffender ist. Schließlich kommen die Probleme aus der Biologie und nicht aus der Informatik und es wird auch nur ein Teil der Biologie, nämlich die Molekularbiologie, betrachtet. Die heutigen Endrundenaufgaben sind so konzipiert, dass Biologiekenntnisse nicht nötig sind (wir sind ja auch ein Informatikwettbewerb).

In der Bioinformatik geht es häufig um Sequenzen (Folgen) von Einheiten, wobei in der Molekularbiologie sowohl die vier Basen Adenin, Guanin, Cytosin und Thymin (oder Uracil) als auch die 20 Aminosäuren eine besondere Rolle spielen. Das menschliche Genom ist eine Sequenz aus ungefähr $3 \cdot 10^9$ Basen. Wir haben es also mit sehr langen Sequenzen zu tun. Daher lassen sich nur zeit- und platzeffiziente Algorithmen erfolgversprechend einsetzen. Zu den zentralen Problemen der Molekularbiologie gehört die Entschlüsselung der biologischen Bedeutung einzelner Sequenzabschnitte. Auf dem Weg zur Lösung des Problems wird angenommen, dass alle Genome durch Veränderungen aus einem Urogenom entstanden sind. Es ist daher wichtig herauszufinden, welche Genome wie verwandt sind und wie lange in der Entwicklungsgeschichte sie einen gemeinsamen Weg genommen haben. Damit kommen wir zu dem Basisproblem, die Ähnlichkeit (oder Verschiedenartigkeit) zweier Sequenzen zu bewerten.

Sequenzen sind für uns im Weiteren einfach Folgen von Buchstaben wie z.B. A, G, C und T, die für die vier genannten Basen stehen. Wir gehen von zwei Sequenzen $x = (x_1, \dots, x_n)$ und $y = (y_1, \dots, y_m)$ aus, wobei alle x_i und y_j aus einem nicht sehr großen Alphabet stammen. Für ein Sequenzenpaar (x, y) werden häufig Alignments (Anpassungen) betrachtet. Das sind Paare (x^*, y^*) , wobei x^* und y^* die gleiche Länge haben und durch Auffüllen von x bzw. y um so genannte Lücken „–“ entstehen (wenn in x^* alle Lücken entfernt werden, erhalten wir wieder x , analog für y^*). Z.B. ist für $x = \text{ACCCGAT}$ und $y = \text{CCTATA}$ ein mögliches Alignment

ACCCGAT–
–CC–TATA

Zur Ähnlichkeitsbewertung von Sequenzen werden nun Alignments bewertet. Die wichtigste Bewertungsfunktion geht von einer paarweisen Bewertung $s(a, b)$ aus, wobei a und b Buchstaben oder Lücken sein können. Insbesondere ist $s(-, -) < 0$. Die Bewertung eines Alignments (x^*, y^*) ist die Summe aller $s(x_i^*, y_i^*)$, es werden also alle untereinander stehenden Paare bewertet. So kann man alle möglichen Alignments von x und y bewerten; die Ähnlichkeit der Sequenzen x und y ist dann die maximale dieser Alignment-Bewertungen. Um mit dem Modell vertraut zu werden, sollte am Anfang kurz diskutiert werden, warum dieses Modell sinnvoll ist.

- 1.) Welches biologische Modell könnte hinter diesen Definitionen stehen? Warum ist es sinnvoll, die Summe (und nicht das Produkt oder eine andere Funktion) der paarweisen Bewertungen zu verwenden? Wie können die Werte $s(a, b)$ ermittelt oder geschätzt werden?

Bei den folgenden algorithmischen Fragestellungen werden „gute“ Algorithmen gesucht. Natürlich ist ein Algorithmus gut, wenn er in möglichst kurzer Zeit mit geringem Speicherplatzbedarf eine optimale Lösung liefert. Aber es sind auch Algorithmen „erlaubt“, die nicht immer die Berechnung einer optimalen Lösung garantieren, dafür aber besonders effizient sind.

- 2.) Entwerfe einen effizienten Algorithmus zur Berechnung global optimaler (oder guter) Alignments. Ein Alignment heißt global optimal, wenn es unter allen Alignments die beste Bewertung hat.

Häufig gibt es eine einheitliche Lückenbewertung $d < 0$, es gilt dann für alle Buchstaben $s(a, -) = s(-, a) = d$. Eine Lücke der Länge l erhält auf diese Weise eine Bewertung von ld . Viele kurze Lücken sind dann genauso teuer wie eine große Lücke derselben Gesamtlänge. Dies ist biologisch nicht sinnvoll (warum?). Daher werden so genannte affine Lückenbewertungen eingeführt, bei denen eine Lücke der Länge l eine Bewertung von $ld + D$ ($D < 0$) erhält.

- 3.) Betrachte Aufgabe 2 für affine Lückenbewertungen.

Beim lokalen Alignment von x und y werden zusammenhängende Teilfolgen x' von x und y' von y gesucht, so dass die Bewertung eines besten Alignments von x' und y' maximal ist. (Hierbei ist x_3, x_4, x_5, x_6 eine zusammenhängende Teilfolge von x , nicht aber x_4, x_6, x_{10} .)

- 4.) Betrachte Aufgabe 2 für lokale Alignments.

Typischerweise werden bei drei Sequenzen x, y und z das optimale Alignment von x und y , das optimale Alignment von x und z und das optimale Alignment von y und z nicht „kompatibel“ sein, also nicht zu einem Alignment der drei Sequenzen x, y und z führen (warum?). Ein multiples Alignment von k Sequenzen ist ein gemeinsames Alignment, wobei jede Position als Bewertung die Summe der $\binom{k}{2}$ paarweisen Bewertungen erhält, für eine Spalte mit x_1, y_1 und z_1 also $s(x_1, y_1) + s(x_1, z_1) + s(y_1, z_1)$.

5.) Erweitere die Lösungen zu Aufgabe 2 zu Algorithmen für das multiple Alignment von k Sequenzen.

Hierbei sei darauf hingewiesen, dass keine Algorithmen bekannt sind, die optimale multiple Alignments für nicht sehr kleine k so effizient berechnen, dass sie für nicht ganz kurze Sequenzen praktisch durchführbar sind. Dennoch sind möglichst effiziente Algorithmen für die Berechnung optimaler Lösungen auch von Interesse. Hier sollten aber in jedem Fall auch heuristische Algorithmen diskutiert werden, also Algorithmen, die hoffentlich schnell hoffentlich gute Lösungen berechnen.

Bei der Ausgangsfrage, wo unser Genom herkommt, sind wir an so genannten phylogenetischen Bäumen interessiert. Die Wurzel ist mit dem Urogenom bezeichnet. Die Kinder eines Knotens werden mit den Genomen bezeichnet, die während der Evolution aus dem Genom, mit dem der Elter (also der Knoten selbst) bezeichnet ist, entstanden sind. An den Blättern stehen dann die Genome der zur Zeit existierenden Arten. Wir kennen weder den Baum noch die Markierung der inneren Knoten. Wir kennen nur die Sequenzen, mit denen die Blätter des Baumes markiert sind.

Phylogenetische Bäume sollen die Evolution gut widerspiegeln. Wie können phylogenetische Bäume bewertet werden? Wir betrachten nun ein Distanzmaß für Sequenzen (was muss im Vergleich zu den bisher behandelten Ähnlichkeitsmaßen geändert werden?). Distanzmaße erfüllen die folgenden Bedingungen:

$$D(x, x) = 0,$$

$$D(x, y) = D(y, x),$$

$$D(x, y) \leq D(x, z) + D(z, y) \text{ (Dreiecksungleichung)}.$$

Die Länge einer Kante eines phylogenetischen Baumes ist gleich der Distanz der Sequenzen, mit denen die beiden Endpunkte der Kanten markiert sind. Die Kosten eines phylogenetischen Baumes sind gleich der Summe der Längen aller Kanten des Baumes.

- 6.) Gegeben ist ein binärer Baum, an dessen Blättern Sequenzen stehen. Es soll ein möglichst billiger Baum berechnet werden, wobei jeder innere Knoten mit einer Sequenz markiert werden muss, die an einem der Kinder steht.
- 7.) Ähnlich wie Aufgabe 6, wobei jetzt beliebige Markierungen an den inneren Knoten erlaubt sind.

Die letzte Problemstellung betrifft ein noch offenes Problem. Es sollten Lösungsansätze diskutiert werden, wenn noch genügend Zeit ist.

- 8.) Wie kann zu einer Menge von Sequenzen ein möglichst guter phylogenetischer Baum berechnet werden? Wie sollte in einem Baum die Wurzel gewählt werden, deren Markierung ja im „richtigen“ phylogenetischen Baum das Ur-genom ist?

Viel Spaß und viel Erfolg!