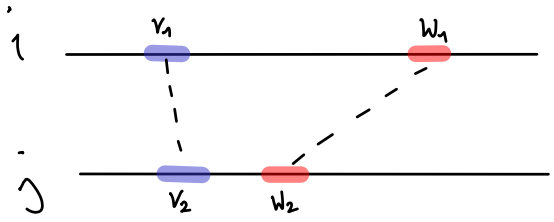
Multiples Alignment von k-meren

Hintergrund

Nachdem in der Vergangenheit eher die Genome von neuen Spezies sequenziert wurden, die die vorher sequenzierten Genome ergänzt haben und von ihnen eher weit entfernt verwandt waren, werden nun zunehmend näher verwandte Genome sequenziert, etwa 10 000 Vertebratenspezies, oder menschliche Individuen oder verschiedene Mausstämme.
Eine für verschiedene Aufgaben der vergleichenden Genomik wichtige Aufgabe ist dann, syntenische Bereiche zu identifizieren und zu alignieren. Ein syntenischer Bereich ist ein Tupel von Sequenzregionen, die homolog zueinander sind, also aus einer gemeinsamen Vorfahrsequenz entstanden sind. Diese sind oft ortholog und in verschiedenen Genomen, können aber auch im selben Genom liegen und durch größere Duplikationen verursacht sein.
In nah verwandten Eukaryoten enthalten syntenische Regionen typischerweise mehrere Gene und umfassen mehrere hunderttausend Nukleotide. Diese Regionen sind jedoch auf einer kleineren Skala "durchlöchert" durch evolutionäre Ereignisse wie Inversionen, Duplikationen, Insertionen (z.B. von repetitiven Sequenzen) und Deletionen, die lokal die Kollinearität oder Ähnlichkeit zerstören. Beim Vergleich von Bakteriengenomen enthalten syntenische Regionen oft eine einzelne transkribierte Einheit (Operon) und der Baum, der die Verwandtschaft der Sequenzen beschreibt ist ein anderer als der, der die Verwandschaft der Genome beschreibt (z.B. wegen horizontalem Gentransfer).
Syntenie wird typischerweise mit paarweisen oder multiplen Sequenzalignments (MSAs) von Genomen identifiziert. Genom-MSAs sind eine Menge von kleineren MSAs im klassischen Sinn. Programme, die Genom-MSAs von vielen Genomen konstruieren, verwenden bisher eine Abfolge von *paarweisen* Alignments (z.B. Cactus, MULTIZ, progressiveMauve), z.B. in einer Reihenfolge, die durch einen Guide-Tree gegeben ist. Solch ein voranschreitender Zugang ("progressive alignment") kann allerdings einen Anteil von klar syntenischen Regionen verfehlen.
Es wird zur Zeit viel an effizienten gemeinsamen Repräsentationen von vielen verwandten Genomen in einem Graph geforscht. Bestimmte solche Graphen sind De-Bruijn-Graphen, die alle k-mere von allen Genomen enthalten für ein festes, geeignet gewähltes k.

Aufgabe



Ein Multiples Sequenzalignment (MSA) kann als eine Äquivalenzrelation auf den Positionen aufgefasst werden, bei der miteinander alignierte Positionen in einer Äquivalenzklasse sind. Es dürfen hier nur Positionen in eine Äquivalenzklasse, an denen das gleiche k-mer vorkommt. Ein MSA muss auch folgende Konsistenzeigenschaft erfüllen: Wenn Positionen

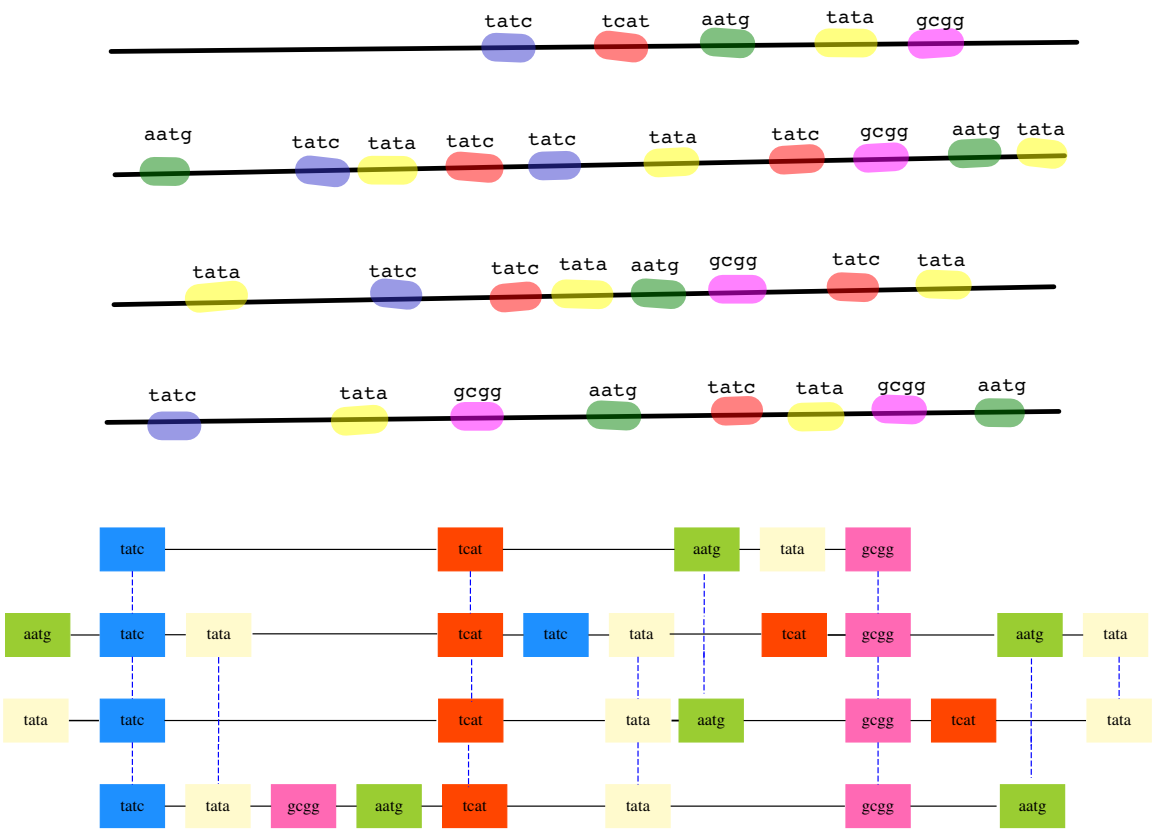


Abbildung 1: Genomalignment-Beispiel. *Oben:* Vier Genomsegmente mit solchen k-meren hervorgehoben, die in mehreren Sequenzen vorkommen. Gleiche Farben markieren gleiche k-mere, verschiedene Farben verschiedene k-mere. *Unten:* MSA der k-mere von oben. Durch blaue gestrichelte Linien verbundene k-mere sind miteinander aligniert. k-mere dürfen auch unaligniert bleiben.

v_1 und w_1 von Sequenz i jeweils mit Positionen v_2 bzw. w_2 von Sequenz j aligniert sind, dann gilt

$$(w_1 \geq v_1 \wedge w_2 \geq v_2) \text{ oder } (w_1 \leq v_1 \wedge w_2 \leq v_2).$$

Dies bedeutet, dass nach horizontalem Platzieren der Knoten (Einfügen von Lücken) sich in Abbildung 1 die blauen Kanten nicht überschneiden dürfen.
Entwickle und schreibe ein multiples Alignmentprogramm, das k-mere miteinander aligniert, die in mehreren Genomen vorkommen. Dieses neue Programm soll lediglich das Vorhandensein von identischen k-meren verwenden. Es soll keine paarweisen Sequenzvergleiche machen, und gut skalieren, wenn die Genomanzahl wächst.
Das Optimierungskriterium ist ein noch genauer zu definierender Alignmentscore. Naheliegend ist eine Summe

$$\sum_A s(A)$$

über die Äquivalenzklassen (Alignmentspalten), wobei z.B. $s(A) := \binom{|A|}{2}$ eine Möglichkeit wäre (Sum-of-Pairs-Score).

Mögliche Herangehensweise

Speichere zunächst alle k-mere von allen Genomen in einer geeigneten Datenstruktur, z.B. eine Hashtabelle mit den k-meren als Schlüssel. Lösche oder ignoriere k-mere, die in sehr wenigen Genomen (z.B. einem einzelnen) vorkommen. Für die verbleibenden k-mere werden jeweils alle gespeichert. Aus Effizienzgründen kann es sinnvoll sein, die Positionen auf "ungefähre" Positionen zu vergröbern, z.B. Position v auf "Kachel" $\lfloor v/1000 \rfloor$ abzubilden. Dann wäre ein Vorkommen ein Tupel (Speziesname, Sequenzname, Kachel). Für ein k-mer $p \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}^k$ sei $V(p)$ die Menge aller Vorkommen. Z.B. könnte man dann ein Abstandsmaß auf k-meren einführen, dem zufolge k-mere p und q nahe sind, wenn sie in vielen Genomen an ähnlichen Positionen vorkommen, etwa

$$d(p, q) := |V(p) \cap V(q)|.$$

k-mere könnten basierend auf so einem Abstandsmaß geclustert werden und besonders große und klar abgegrenzte Cluster könnten in einem iterativen Algorithmus als Äquivalenzklasse realisiert werden. Nachem einige Äquivalenzklassen (Alignmentspalten) gewählt wurden, ist wegen der Konsistenzbedingung die weitere Auswahl eingeschränkt und die Datenstruktur muss entsprechend geupdatet werden.
Das Programm sollte in `C++` programmiert werden, da bei einer großen Anzahl von großen Genomen hohe Effizienz sehr wichtig ist.

Eingabedaten und Anwendungsbeispiel

Ich werde zum Testen eine Menge vermutlich syntenischer Regionen bereitstellen, etwa von 20 Mausgenomen oder von ca. 10 Primatengenomen. Auch Sequenzen aus simulierter Evolution kommen für erste Tests in Frage.