Hur kan jag göra datamaskiner snabbare?

Baserat på verkliga händelser (högst 90% fabrikation)

Alernativt: Snabba data strukturer för medicin och bioinformatik[*]

Saska Dönges

[*]Med material stulet från Travis Gagies DCC 2023 inledningsanförande

AGENDA

1. Vem är jag?
2. Berättelsen
3. Motivation
4. Resultat

JAG

Doktorand i “Compressed data structures” gruppen, som är en del av supergruppen “algorithmic bioinformatics”.

Borde disputera inom $\sim$ 1.5 år.

Efter gymnasiet och milin studerade jag utomlands i några år tills jag burnouttade och flyttade tillbaka till Finland.

Jobbade några år “inom industrin” tills jag fick nog.

Startade på uni hösten 2015 $\rightarrow$ magister på våren 2021.

BERÄTTELSEN

Jag återkommer till medicin, bioinformatik och mer detaljerad information om forskningsresultat om jag har tid.

Men först, en berättelse om min forsking, som kanske skulle kunna vara sann.

DATAMASKINER ÄR SNABBA

En modern dator kan göra miljontals meningsfulla beräkningar per sekund.

Men datorer är enkla. De gör exakt vad som bes av dem.

Även om det skulle kunna finnas bättre sätt att göra saker.

Gör en svår sak enklare för datorn $\Leftrightarrow$ Datorn kan beräkna saken snabbare.

MULTIPLIKATION

Det är (inte speciellt) svårt att beräkna $70 \cdot 300$.

Det är enklare att räkna $7 \cdot 3 \cdot 10 \cdot 100$. “21 och tre nollor” $= 21000$.

Skulle dethär vara något som vi skulle kunna hjälpa datorer med?

MULTIPLIKATION

Det är (inte speciellt) svårt att beräkna $70 \cdot 300$.

Det är enklare att räkna $7 \cdot 3 \cdot 10 \cdot 100$. “21 och tre nollor” $= 21000$.

Skulle dethär vara något som vi skulle kunna hjälpa datorer med?

Nej. Tyvärr är datorer bra på att multiplicera, och bryr sig inte egentligen om vilka tal som används.

MULTIPLIKATION

Det är (inte speciellt) svårt att beräkna $70 \cdot 300$.

Det är enklare att räkna $7 \cdot 3 \cdot 10 \cdot 100$. “21 och tre nollor” $= 21000$.

Skulle dethär vara något som vi skulle kunna hjälpa datorer med?

Nej. Tyvärr är datorer bra på att multiplicera, och bryr sig inte egentligen om vilka tal som används.

Men kanske division...

DIVISION

Det är (inte speciellt) svårt att beräkna $1000/300$.

Det är enklare att ta $\frac{1000/100}{3}$. “Ta bort 2 nollor och dividera med 3” $= 10/3 = 3\frac{1}{3}$.

Skulle dethär vara något som vi skulle kunna hjälpa datorer med?

DIVISION

Det är (inte speciellt) svårt att beräkna $1000/300$.

Det är enklare att ta $\frac{1000/100}{3}$. “Ta bort 2 nollor och dividera med 3” $= 10/3 = 3\frac{1}{3}$.

Skulle dethär vara något som vi skulle kunna hjälpa datorer med?

Jo! Fast inte så som jag gjorde åvan.

DIVISION

Det visar sig att datorer inte är speciellt bra på att dividera. Det skulle vara bättre att multiplicera i stället.

Så i stället för att göra $1000/300$ kan vi göra $1000 \cdot 0.0033333\ldots$.

Eller om vi arbetar med 32-bitars heltal kan vi göra $1000 \cdot 458129845 \gg 37$ I stället.

Altså vi sparar resultetet av $1000 \cdot 458129845$ som ett 64-bitars heltal och skiftar sedan resultatet till höger med 37 bitar.

DIVISION

$$
\begin{array}{rl}
 & 00000000000000000000000000000000000000000000000000010011100010000 \\
/ & 00000000000000000000000000000000000000000000000000000100101100 \\
\hline
= & 00000000000000000000000000000000000000000000000000000000100001
\end{array}
$$

$$
\begin{array}{rl}
 & 00000000000000000000000000000000000000000000000000010011100010000 \\
\cdot & 00000000000000000000000000000000000011011010011101000000110110101 \\
\hline
= & 00000000000000000000100001010101010101010101010101011001010000
\end{array}
$$

Tyvärr gör moderna kompilatorer det här automatiskt. (Det var inte jag som räknade ut lösningen på förra sidan.)

NÅ MEN STRÄNGÖKNING DÅ?

Strängsökning går ut på att hitta förekomster för en söksträng P med längden m ur en (ofta lång) text T med längden n.

DNA SEKVENSERING

Då en människas genom sekvenseras, görs det i allmänhet med hjälp av ett referensgenom.

En sekvenseringsmaskin läser in små snuttar av en ny människas dna.

Rätt plats för snuttarna söks genom att gämföra till referensen.

referens	G	A	T	A	C	A	T
snutt 1	G	A	T	A			
snutt 2		A	T	A	C		
snutt 3				A	C	A	T

DNA SEKVENSERING

Fungerar ofta ganska bra.

Men om den nya individen har små mutationer som inte kommer överens med referensen kan det uppstå problem.

reference	G	A	T	T	A	C	A	T
read 1	G	A	T	-	A			
read 2		A	-	T	A	G		
read 3					A	G	A	T
output	G	A	T	-	A	C	A	T

I problemfall fyller man bara i med referensgenomen.

Referens bias

To the scientists' puzzlement, however, the boy's sequence showed no sign of the mutation in the gene known to cause Baratela Scott, called XYLT1. Nor did the DNA of the next boy with the disorder, or the next. As they tried to compare the boys' DNA sequences to the reference genome, it was like trying to check a spelling in a Webster's from which a prankster had torn handfuls of pages. Many pieces of the boys' genomes, called short reads, "weren't in the reference genome at all," . . . There was no way to check them for disease-causing misspellings.

— Sharon Begley, *Stat News*, March 11th, 2019

REFERENS BIAS

This bias limits the kind of genetic variation that can be detected, leaving some patients without diagnoses and potentially without proper treatment. What is more, people who share less ancestry with the man from Buffalo will probably benefit less from the incoming era of precision medicine, which promises to tailor healthcare to individuals.

[O]ur understanding of diversity within populations of European descent is now so good that we can start to use it for precision medicine. But for other populations, "We do not have the same kind of data . . . [This] is going to increase healthcare disparities above and beyond what they are today." . . . [A] huge new project is offering a different solution with the aim to represent global diversity: a human pangenome.

— Ida Emilie Steinmark, *Guardian*, January 29th, 2023

Referens bias

[T]he project is not just about sequencing more diverse data. "We need to come up with a better data structure to encode that information," . . . That data structure is called a genome graph. In contrast to the current reference, which is just a long string of letters, the genome graph shows variation between genomes as detours on an otherwise shared path. That will enable researchers and doctors to map short reads to the version of the path that best fits their sample.

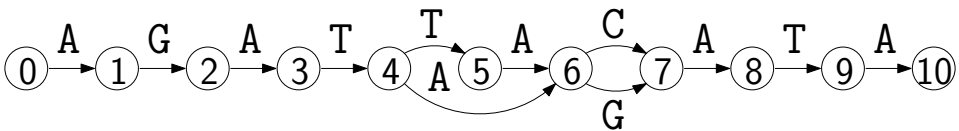
— Ida Emilie Steinmark, *Guardian*, January 29th, 2023

PANGENOM GRAF

Flere referensgenom:

- `GATTACAT`
- `AGATACAT`
- `GATACAT`
- `GATTAGAT`
- `GATTAGATA`

Pangenom graf utgående från referenserna:

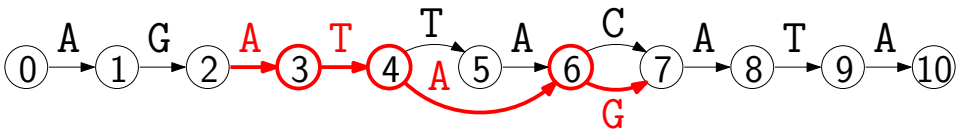


PANGENOM GRAF

Referensgenom:

- GA3TTACAT
- AGATACAT
- GATACAT
- GATTAGAT
- GATTAGATA

Ser ut som om “ATAG” skulle vara en känd delsekvens

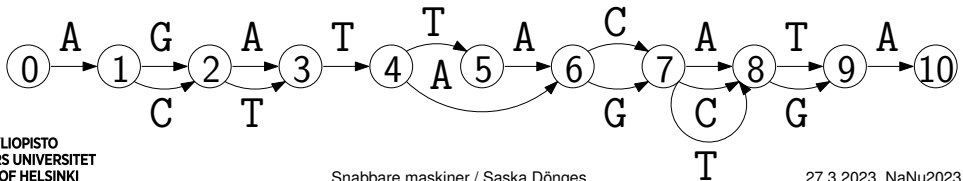


PANGENOM GRAF

Ännе flere referensgenom:

- GATTACAT
- AGATACAT
- GATACAT
- GATTAGAT
- GATTAGATA
- CATTACAT
- GTTAGAT
- GATTCCATA
- GATTACAGA

Ännu svårare att veta vad som är riktigt och vad som inte är det



PANGENOM LÖSNING

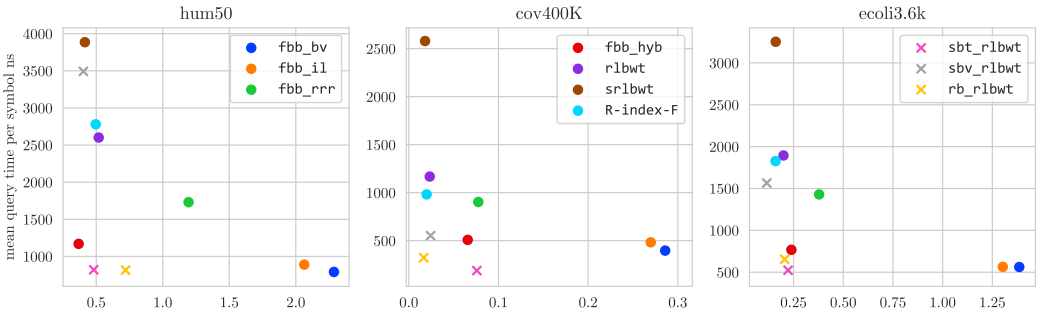
Lösningen är att inte använda grafen för att hitta var en snutt hör till, utan att helt enkelt leta rätt på positionen genom att söka i alla referensgenom.

Problemet är att ett genom är hyfsat stort $\sim$ 700 Megabit (lätt komprimerat).

Om vi vill söka i 100000 referensgenom, tar lätt komprimerade referenserna 70 terabyte utrymme.

Ett effektivt komprimerat index kanske bara tar $\sim$ 1 terabyte, vilket fortfarande är ganska mycket, men rymms i minnet av superdatorer.

VÅR IMPLEMENTATION ÄR BRA



Våra index indikeras med “x”. De kräver mindre utrymme och är lika snabba som de snabbaste konkurrenterna.