Unieke factoren in het menselijk DNA

Jeroen F. J. Laros

Begeleiders: Peter Taschner
Hendrik Jan Hoogeboom
Walter Kosters

DNA is opgebouwd uit 4 letters: A, T, C en G

Er zijn ongeveer 3×10^9 van deze letters in het menselijk genoom.

We zoeken naar unieke stukjes van hoogstens lengte 18. Er zijn $4^{18} \approx 6.9 \times 10^{10}$ mogelijke stukjes van lengte 18.

Statistisch gezien zou ongeveer 5% van alles wat we tegenkomen uniek moeten zijn.

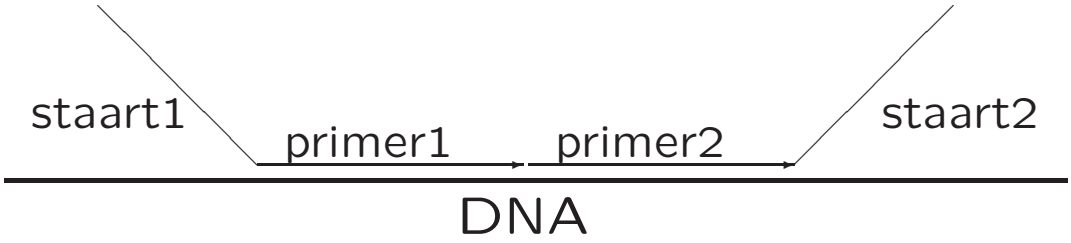
Het vinden van unieke stukjes DNA is het beginpunt voor bepaalde technieken binnen de biologie en de geneeskunde.

Enkele voorbeelden zijn

- MPLA Multiplex Ligation-dependent Probe Amplification

- Micro Arrays

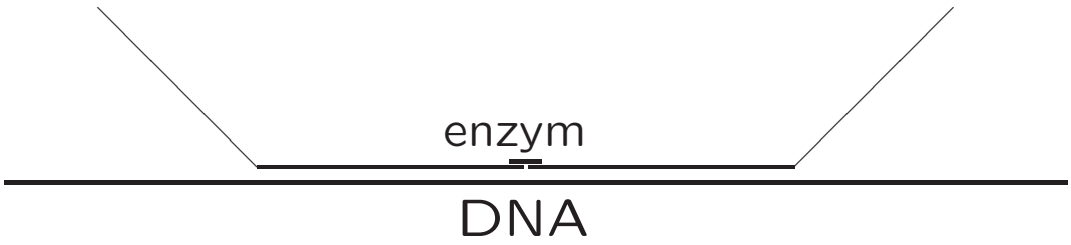
MPLA



Onder een bepaalde temperatuur plakken de primers op het DNA.

```
        primer
      ATGGTAATCCGA
      ||||||||||||
GCCGTATACCATTAGGCTTTGAA
        DNA
```

MPLA

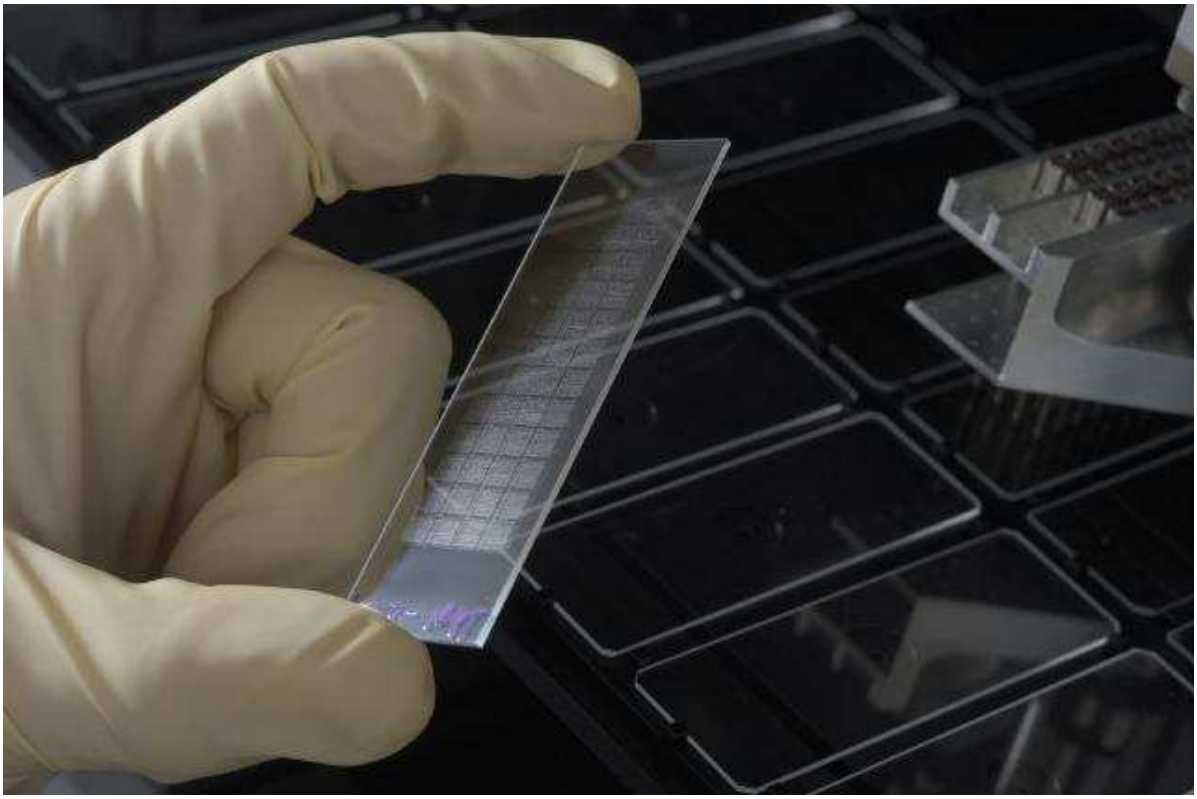


De primers worden aan elkaar “gelijmd” met een enzym.

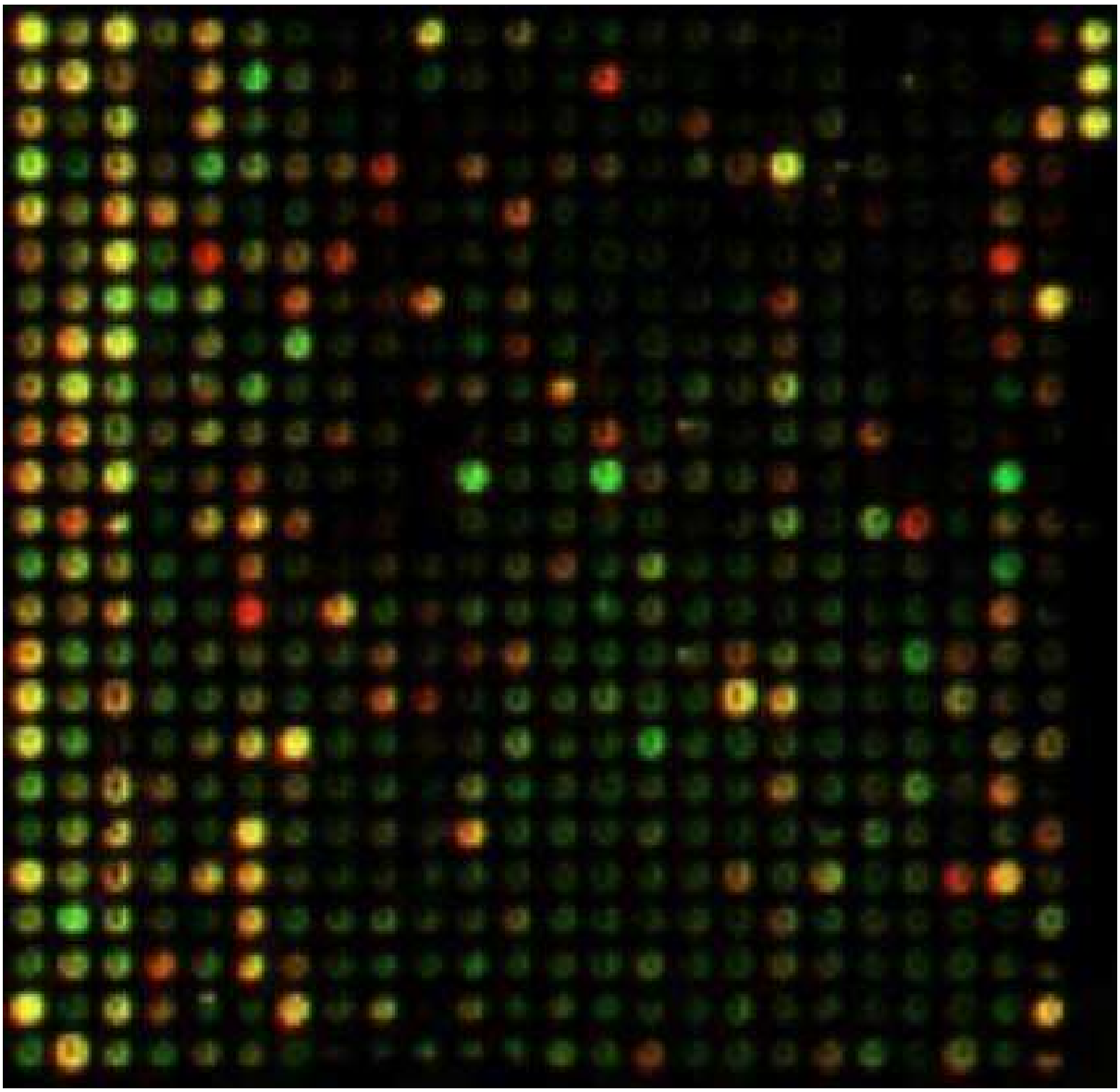
MPLA

Boven een bepaalde temperatuur laat het product los.

Micro arrays

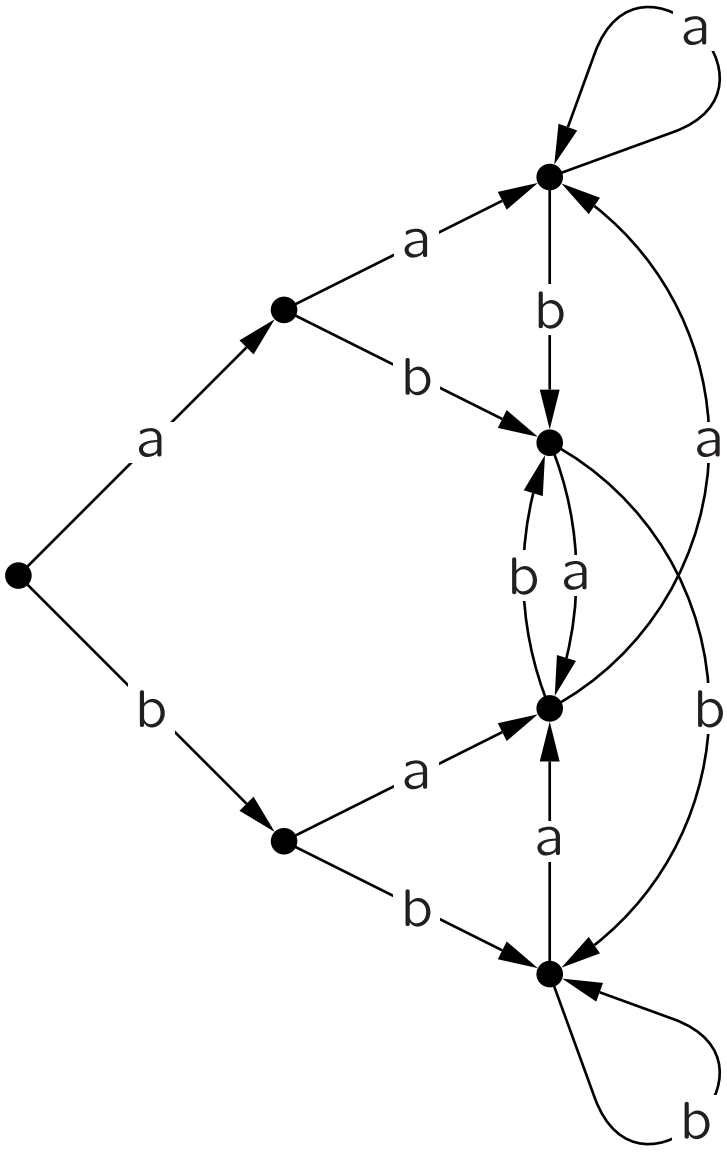


Micro arrays



De uitlezing van een micro array.

Unieke factoren zoeken



Een trie (de Aho-Corasick methode).

Fysieke beperkingen

Een trie voor lengte 18 is te groot voor het geheugen, het zou honderden Gigabytes aan geheugen kosten, meer dan er op een moderne harde schijf past, laat staan het werkgeheugen.

Een harde schijf van minimaal 64Gb volstaat nog, als we een impliciete structuur gebruiken.

```
AAAA - 0
AAAC - 1
...
TTTG - 254
TTTT - 255
```

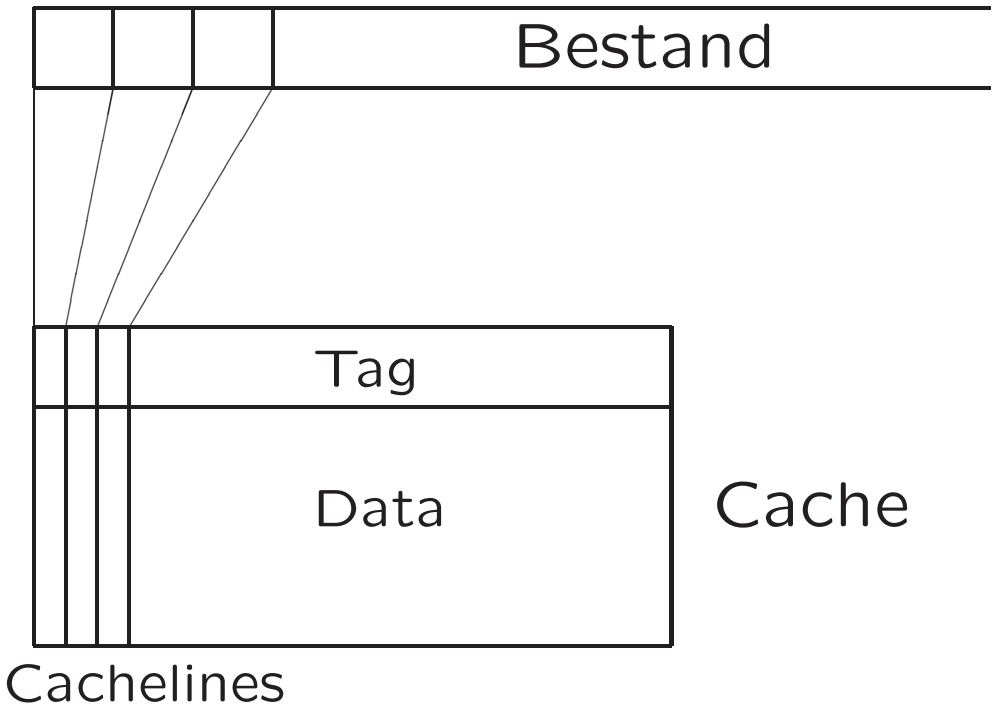
De nadelen hiervan zijn

- Een harde schijf is veel trager dan werkgeheugen.

- Een harde schijf is niet gemaakt om random data weg te schrijven.

Fysieke beperkingen

Bufferen van data is een manier om gebruik te maken van geheugen dat al eens aangesproken is.



Als we 4x de string ATGG tegenkomen, dan kunnen we beter eerst tot 4 tellen en dan de waarde 4 wegschrijven in plaats van 4x naar de harde schijf te schrijven.

Fysieke beperkingen

Als het geheugen vol begint te raken, dan moet er iets naar de harde schijf worden weggeschreven. Als we daar dan toch mee bezig zijn, dan kunnen we maar beter gebruik maken van het feit dat schijven in blokken schrijven.

We kijken hoeveel cachelines er in een blok vallen en schrijven dan al die cachelines weg.

Lineaire schijf toegang

Door gedeeltes van de strings te fixeren kunnen we zonder random harde schijf toegang af.

```
AAAA....
AAAC....
...
TTTT....
```

Wel moeten we na elke keer de getelde data naar schijf schrijven, maar omdat het nu lineair gaat, is dat niet erg.

Dubbelstrengs DNA

DNA kan maar in 1 richting gelezen worden.

We houden bij wat het "reverse complement" is.

AATA $\Leftrightarrow$ TATT.
Als we nu een nieuwe nucleotide lezen (C), dan krijgen we:
AATAC $\Leftrightarrow$ GTATT.

Elk van deze strings is te converteren naar een getal, en een van deze getallen is het kleinst. Dit getal gebruiken we als identificatie.

Als we AAAA of TTTT tegenkomen, dan tellen we 1 op bij offset 0.

Een masker maken

Eerst vertalen we een string DNA naar binaire vorm en we reserveren ruimte voor de score.

ATGCT $\Rightarrow$
00 11 10 01 11 $\Rightarrow$
00 00 11 00 10 00 01 00 11 00

En na de analyse voor lengte 3 zou het er zo uit kunnen zien

00 00 11 00 10 10 01 01 11 11

Terugvertaald is dit

A 0 T 0 G 2 C 0 T 3

Dit betekent dat ATG 3$\times$ voorkomt, TGC 1$\times$ en GCT $\geq$ 4$\times$.

Verdere analyse

Nadat we deze “mal” hebben gemaakt, kan er meer specifieke analyse op worden toegepast. Bijvoorbeeld

- Het bepalen van het GC-gehalte.
- De bindingsengergie bepalen.
- Het eruit filteren van ongewenste stukjes.

GC-gehalte

`ATTAGCAAGAATA` heeft een GC-gehalte van $\frac{3}{13}$.

Het bepalen van een groot aantal van deze gehaltes kan een stuk sneller met een verschuivend venster.

- Bepaal eerst 1 maal het aantal C's en G's.

- Tel 1 bij de score op als we een C of een G toevoegen, haal er 1 vanaf als we een C of een G weggooien.

GC-gehalte

We kijken naar de GC-gehaltes van strings van lengte 4 in `ATTAGCAAGA`. De strings die we tegenkomen zijn:

string	verandering	gehalte
ATTA		$\frac{0}{4}$
ATTAG	$+1$	$\frac{1}{4}$
TTAGC	$+1$	$\frac{2}{4}$
TAGCA		$\frac{2}{4}$
AGCAA		$\frac{2}{4}$
GCAAG	$-1+1$	$\frac{2}{4}$
CAAGA	-1	$\frac{1}{4}$

Bindingsenergie

Ook het bepalen van de bindingsenergie is op deze manier te doen, we moeten alleen de eerste en laatste paren bekijken.

De bindingsenergie van een streng DNA is te benaderen door te kijken naar paren nucleotides. Hiervoor zijn tabellen met expirimenteel waargenomen waarden nodig.

Net als met het bepalen van het GC-gehalte kunnen we veel rekenwerk van te voren doen.

Hiervoor hebben we een tabel gemaakt met alle mogelijke paren.

Als we de bindingsenergie van strings van lengte 6 gaan bepalen in ATTAGCAAGA, dan beginnen we met het berekenen van

$$(AT) + (TT) + (TA) + (AG) + (GC)$$

Hierna kunnen we het berekende resultaat hergebruiken

string	opzoeken
ATTAGCA	$-(AT) + (CA)$
TTAGCAA	$-(TT) + (AA)$
TAGCAAG	$-(TA) + (AG)$
AGCAAGA	$-(AG) + (GA)$

Deze manier van "on the fly" berekenen scheelt erg veel tijd.

Het eruit filteren van ongewenste stukken DNA doen we ook door middel van een trie. Zo hoeven we slechts 1x door de data heen te lopen om stukjes af te keuren.

Zowel bij het afkeuren op GC-gehalte, bindingsenergie alsmede andere criterea, hebben we er voor gekozen het getal 4 aan de string toe te kennen.

Uiteindelijk betekent de 4 dus alleen afgekeurd.

lengte	uitvoer	origineel	gecached	gecached & geblocked	meerdere iteraties
8	64K	$\pm$ 5m 19s	$\pm$ 4.4s		2.4s
9	256K	$\pm$ 5m 28s	$\pm$ 5.2s		2.6s
10	1M	$\pm$ 5m 42s	$\pm$ 6.5s		3.6s
11	4M	$\pm$ 8m 25s	$\pm$ 8.9s		4.3s
12	16M	$\pm$ 30m	$\pm$ 17.6s		4.4s
13	64M	$\pm$ 4h 10m	$\pm$ 55.2s		8.2s
14	256M	$\pm \infty$	$\pm$ 58m	$\pm$ 18m	26.8s
15	1G	$\pm \infty$	$\pm$ 1h 44m	$\pm$ 46m	1m 40s
16	4G	$\pm \infty$	$\pm$ 1h 19m		7m 25s
17	16G	$\pm \infty$	$\pm$ 1h 35m		29m 30s
18	64G	$\pm \infty$	$\pm$ 2h 24m		