Inleiding Medisch Technische Wetenschappen

Bioinformatica
Deel 2

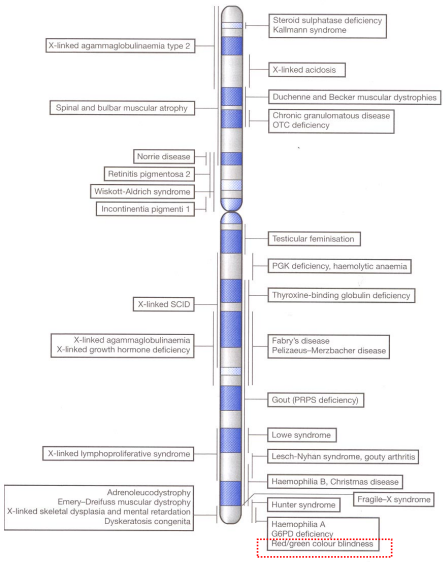
Michael Egmont-Petersen

Universiteit Utrecht

Het menselijk genoom

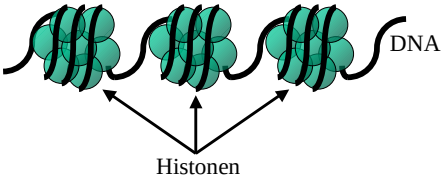
- Het menselijk genoom (DNA) bestaat uit:
  - 3200 Mega Basenparen (MB), A, G, C, T.
- Het menselijk DNA is ingedeeld in:
  - 23 chromosoomparen
    - 22 paren, autosomen, bevinden zich in mannen en vrouwen
    - Een chromosoompaar verschilt, XY (man) versus XX (vrouw)
- De structuur van chromosomen worden in verband gebracht met genetische afwijkingen (ziektes)
  - Voorbeeld is Down's syndroom waarbij meiose resulteert in ongebalanceerde chromosomen

X chromosoom



Gevouwen chromosoom

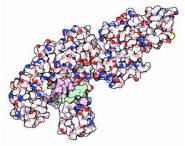
- Elk chromosoom is zeer compact verpakt:
  - Chromosoom #1 is 8 cm lang
  - Hetzelfde chromosoom is 8 µm lang in de cel
  - Elk chromosoom is gevouwen rondom proteïnes (histonen)



- Een chromosoom bevat een hoeveelheid aan genen
- Functionele definitie van een gen:
  - Een DNA subsequentie die bijdraagt aan het fenotype van het organisme

Transcriptie – aflezen van het DNA

- Bij celdeling wordt het gehele DNA gekopieerd (transcriptie) door het template enzym DNA Polymerase



- DNA polymerase is zeer nauwkeurig, het maakt minder dan één fout per miljard basenparen
- Een fout-correctie mechanisme laat een foutief basenpaar weg

- Bij de normale functie in de cel zorgt het enzym RNA Polymerase II voor transcriptie van een deel van het DNA (een gen)

29

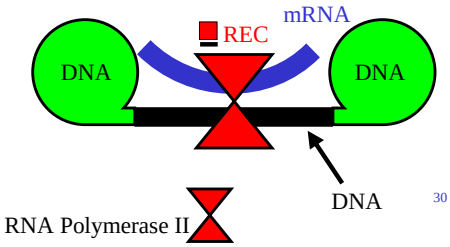
Transcriptie van RNA I

- In de celkern worden stukken van het DNA gekopieerd in RNA, RNA $_{U\subseteq T}$ DNA
  - De operator $_{U\subseteq T}$ geeft aan dat thymine vervangen wordt door uracil

- Transcriptie van RNA begint bij een TATA box

- Transcriptie eindigt door een 'stopsein': AAUAAA

Uitvouwen van DNA
bij transcriptie



30

Transcriptie van RNA II

- mRNA bestaat uit exons en introns
  - exon, codeert voor proteïne
  - intron, geen codering ('spatie')

  - Een paar gegevens:
    - Het histone gen bevat geen intron
    - Het α-globin gen is 0.8 kb lang en bevat 3 introns
    - Het dystrophin gen is 2.4 MB en bevat 79 introns (99.4% van de totale lengte)

- De introns worden verwijderd voordat mRNA de celkern verlaat – splicing
  - De ribonucleoproteïne partikelen, snRNP, zijn verantwoordelijk voor splicing

- Cloning daarentegen maakt exacte kopieën van het DNA – wordt gebruikt bij het ontrafelen van het menselijk genoom in het laboratorium

31

DNA Cloning in het laboratorium

- Polymerase kettingreactie (Polymerase Chain Reaction, PCR) wordt gebruikt om kopieën te maken van het DNA
- Elke kopie is een DNA fragment met een lengte van 600 tot 1200 bp

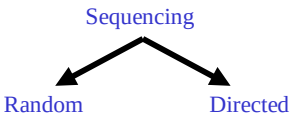
- Polymerase werd ontdekt in de jaren 60 in Yellowstone National Park (VS)
- De bacterie thermus aquaticus leeft daar in warme bronnen
- Zijn enzymen zijn bestaand tegen hoge temperaturen

- Cloning:
1. DNA, polymerase en een 'primer' worden gemengd
2. De twee DNA ladders worden eerste gesepareerd door warmte, 90-95 °C
3. Afkoeling tot 75 °C leidt tot binding met de primers en vervolgens kopiering door polymerase

32

Sequencing van een gen

- Er bestaan twee algemene technieken voor sequencing



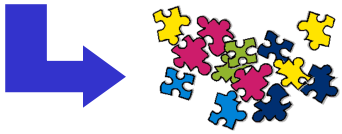
- Random sequencing, maak kopieën van stukken van het DNA
- Het toeval bepaalt welk stuk DNA in welk fragment gekopieerd wordt

DNA: AAGTGACCCGTGAGGATATA
Fragmenten: AAGTG, ACCCG, TGAGG, ATATA %

- Overlap nodig om het DNA te kunnen reconstrueren!

Random sequencing met overlap

DNA: AAGTGACCCGTGAGGATATA
Fragmenten: AAGTG, TGACC, ACCCG, CCGTG, GTGAG, GAGGA, TATA

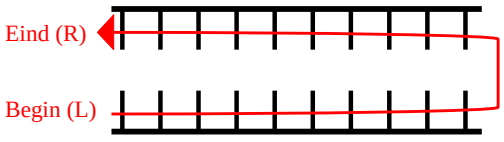


- Probleem: geen overlap tussen TATA en de rest
  - Een deel (AAGTGACCCGTGAGGA) van het DNA kan wel gereconstrueerd worden
- Probleem: Door toeval worden sommige stukken DNA veelvoudig gerepliceerd – andere helemaal niet
- Probleem: Om de kans de minimaliseren dat een stuk DNA niet wordt gerepliceerd, moeten onnodig vele kopieën worden gemaakt
- Probleem: Lange repetitieve sequenties kunnen niet met elkaar gematched worden
  - Bijvoorbeeld 2000 AG-paren achter elkaar (AGAG…..AG)

Gerichte sequencing

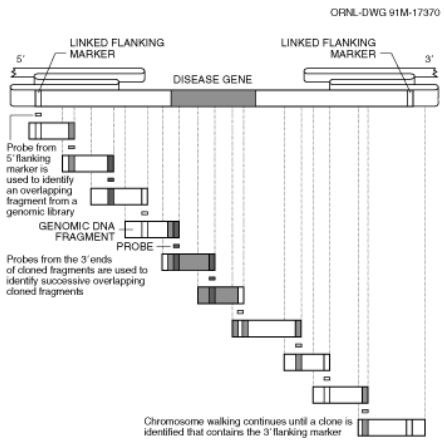
- Gerichte seqencing
  - Cloning beginnend bij een bepaalde (bekende) positie op het DNA
  - Normaliter worden twee DNA ladders (L en R) gelezen van begin (L) tot eind (L) en vanaf het eind (R) tot het begin (R)

Gerichte sequencing van het DNA



Primer walking

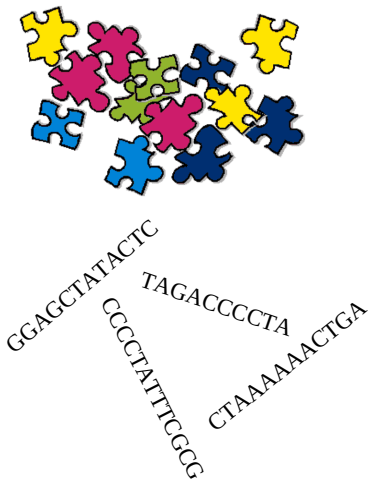
Illustratie van Primer walking



Voor- en nadelen – gerichte sequencing

- Voordeel: overlap is gegarandeerd
- Voordeel: Alle stukken DNA worden gerepliceerd
- Probleem: Lange repetitieve sequenties kunnen steeds niet met elkaar gematched worden
- Probleem: Gerichte sequencing is veel ingewikkelder – je kunt bijvoorbeeld de weg kwijt raken tijdens primer walking
- Probleem: Het constant opnieuw aanmaken van primers gaat langzaam – en kost veel ressources
- Niet elk gen-laboratorium kan gerichte sequencing in de praktijk uitvoeren

De puzzel oplossen – alignment



Hoe?

Het alignment probleem

- Random sequencing resulteert in miljoenen DNA-fragmenten, zeg m
- De fragmenten hebben een gemiddelde lengte van $m^{-1} \sum_i n_i$ basenparen
- Paarsgewijs koppelen van m sequenties vergt $\frac{1}{2}m(m-1)$ matchingen
- Twee sequenties met de lengtes n_1 en n_2 hebben $(n_1 - n_2+1)$ volledig overlappende combinaties plus $2(n_2 - 1)$ gedeeltelijk overlappende combinaties

GGAGCTATACTC

CTAAAAAACTGA ⋯▶

- In het ergste geval $((n_1 - n_2+1)+ 2\,(n_2 - 1))\,(\frac{1}{2}\,m\,(m-1))$ vergelijkingen
- Voorbeeld: m=900, n_1= 1200, n_2=600
  - Resulteert in $((1200-600+1)+ 2\,(600 - 1))\,(\frac{1}{2}\,900\,(900-1)) \approx 728$ miljoen vergelijkingen

Het alignment probleem in de praktijk

- Gewoonweg alle mogelijke matches uitproberen kost te veel tijd – en is ook niet efficiënt
- Matching kan herhalingen die langer zijn dan 1200 bp niet hanteren
- Partiele matching ook interessant, want twee bijna gelijke sequenties kunnen hetzelfde fenotype hebben

  GGATTGAAGC

  GG –TTGAAGC

  Hier zijn subseqenties gematched

  Het '–' teken is ingezet op de niet-gematchte plaats
- Hoe kan het matchproces geoptimaliseerd worden?

Matching door dynamisch programmeren

- Dynamisch programmeren is een techniek die complete en partiele matches kan vinden
- Hoe werkt dynamisch programmeren?
- Men begint met de lege match tekst:

s1="GGATTGAAGC"
s2="GGTTGAAGC"

// Algoritme 1 – tellen aantal matches (pseudocode)

```
#Match=0; Substring="";
∀i, ∀j:
  if Matching(s1(i),s2(j))
   #Match=#Match+1;
  else
   #Match=max(score(Matchletter(s1)),
          score(Matchletter(s2)));
  end;
```

41

Dynamisch programmeren I

- Begin met twee sequenties
  GAGA
  TGA
- Maak tabel met randen

		G	A	G	A
	0	0	0	0	0
T	0	0	0	0	0
G	0	1	1	1	1
A	0	1	2	2	2

Berekening van tabel

// Algoritme 2 – construeer tabel (pseudocode)

```
for all i>1
   for all j>1
      if s1[i]==s2[j] // Match, basenpaar
         T[i,j]=T[i-1,j-1]+1;
      else
         T[i,j]=max(T[i-1,j],T[i,j-1]);
```

42

Dynamisch programmeren II

- Reconstrueer substring uit tabel

		G	A	G	A
	0	0	0	0	0
T	0	0	0	0	0
G	0	1	1	1	1
A	0	1	2	2	2

Match substring

// Algoritme 3 – reconstrueer match (pseudocode)

```
i=len(s1)+1; j=len(s2)+1; // len(s)=lengte(s)
s="";
while i>1
    while j>1
        if s1[i]==s2[j]
            s=cat(s1(i),s); // Concatenatie
            i=i-1; j=j-1;
        else
            if T[i,j-1]>=T[i-1,j]
                j=j-1;
            else
                i=i-1;
```

- Resulteert in GAGA of GAGA
  TGA GAT

43

Dynamisch programmeren III

- Waarom werkt dynamisch programmeren?

Laten we een paar voorbeelden bekijken:
TAAAGC
TCAAA

		T	A	A	A	G	C
	0	0	0	0	0	0	0
T	0	1	1	1	1	1	1
C	0	1	1	1	1	1	2
A	0	1	2	2	2	2	2
A	0	1	2	3	3	3	3
A	0	1	2	3	4	4	4

- De hoogste score, tot nu toe, wordt altijd doorgegeven, → ↓ ↘
- De volgorde van de basen ligt vast:
  TAAAGC, ~~AAAGT~~

44

Dynamisch programmeren IV

- Voorbeeld – geen match

  TTTGC

  AAA

		T	T	T	G	C
	0	0	0	0	0	0
A	0	0	0	0	0	0
A	0	0	0	0	0	0
A	0	0	0	0	0	0

Een score van 0

- Voorbeeld – totale match

  TTT

  TTT

		T	T	T
	0	0	0	0
T	0	1	1	1
T	0	1	2	2
T	0	1	2	3

Een score van 3

45

Samenvatting

- Het menselijk genoom bestaat uit lange sequenties van basen (nucleotiden): A, C, G, T
- Het genoom is ingedeeld in 23 chromosomen – elk chromosoom bevat een aantal genen
- Elk gen draagt bij aan de functie van de cel
- Stukken DNA worden “gekopieerd” door transcriptie ⇒ RNA
- Sequencing bootst transcriptie na:
  - Random sequencing
  - Gerichte sequencing
- Sequencing resulteert in veel DNA-fragmenten, 600–1200 bp lang
- Reconstructie van het DNA vergt matching
  - Dynamisch programmeren is een techniek voor matching van substrings
  - Niet-matches mogelijk bij dynamisch programmeren
- Het herhalingsprobleem (>1200 bp) is niet opgelost

46