Gensøk-algoritmer

Forelesning INF3350/4350
5. sept 2007

Ole Christian Lingjærde
Gruppen for bioinformatikk
Institutt for Informatikk, UiO

Hvor er genene?

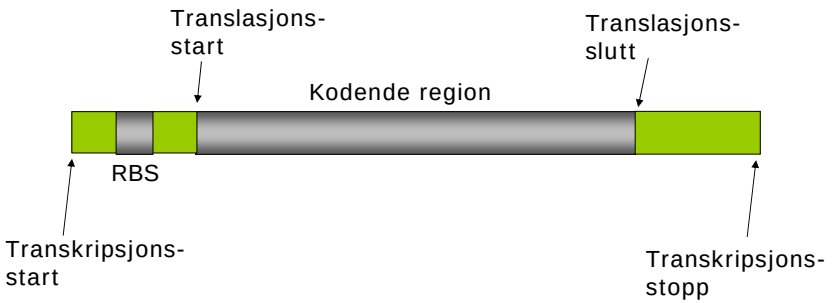
En viktig del av kartleggingen av et genom er å finne ut hvor genene og deres regulatoriske elementer ligger. Mønstergjenkjenningsproblem!

GATCAGTTTCTTTAAGCCGATGGGTCCAGACTTTTCAGCCCTGCCAGAGAATTCCTAAT
TCCATCTCTCAGGTTTTCCAGTGGTAATGAAAAGCTAGCCAAGTTTGGCTATGCTAACC
AAAGCGGGTTCAGTGTGTGTTGTCAGTAAATATTAGTCTATGTGATGTTAATAATCAAAC
TTATCTTGTGTGGGACCACT**ATGCTGAATGAACTTTGTACTGTTATCTCATTTAATTCTG**
AGGATAGCTCTTAAGGTAAGTATTATGATAGCCCTTGATTTACACTTGAGGAAACCAA
GGCATAGAGAGATTAAGTAGTGTGTCTAAAGTCACACTACTAGAAAGTGCAAGAGCCT
GAACTCAACCCAGGCAGTCTGACTCTGGAGCCCAGCTTGTGAGCTCCATGCTAGTCTG
TCACCTTACCTTACCAGTCCTTGGACTACAAAGCTGCTAGTTCTGGTACTGTATCCTTGA
GTGTCACGCGCGTCCGTGTGAAGAGACCACCAAACAGGCTTTGTGTGAGCAATAAA.....

.

Vanskelig problem, fordi gener varierer mye i struktur, basekomposisjon og lengde. Å skille kodende DNA fra ikke-kodende DNA krever ofte en kombinasjon av flere teknikker.

Genenes anatomi (prokaryoter)

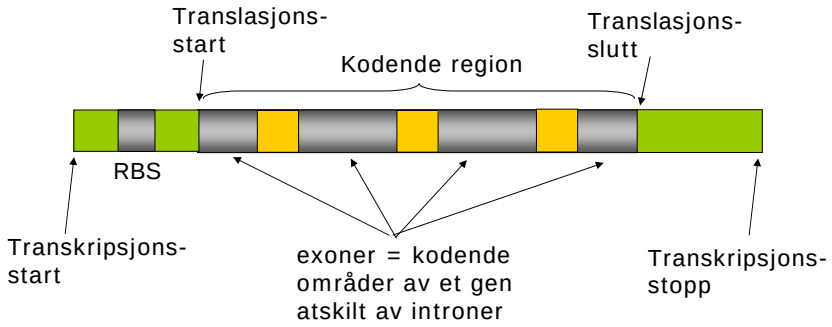


RBS = Ribosomalt bindingssete
Transkripsjon = kopiering av DNA til mRNA (hvert gen for seg)
Translasjon = oversettelse fra mRNA til protein (skjer på ribosomene)

Gener i prokaryoter

- Høy gentetthet (> 90% kodende)
- Få repetitive sekvenser
- Et gen har en enkelt sammenhengende kodende sekvens
- De fleste gener har ATG (=methionin) som startkodon
- Tre mulige stoppkodoner (TAG, TAA, TGA)
- Ofte felles transkripsjon av flere gener – operoner – med karakteristisk termineringssignal (rho-uavhengig terminator)
- Mellom transkripsjonsstart og translasjonsstart er et ribosomalt bindingssete (RBS), Shine-Dalgarno sekvensen, med konsensussekvens AGGAGG
- Tredje posisjon i kodoner er GC-rik
- Tredje nukleotid i kodoner har tendens til å repetere seg

Genenes anatomi (eukaryoter)



Gener i eukaryoter

- Lav gentetthet (~3% i det humane genomet)
- Mange repetitive sekvenser og transposoner (= sekvenser av DNA som kan flytte rundt til forskjellige posisjoner i genomet i en enkelt celle med hjelp fra naturlig forekommende enzymer)
- Et gen består normalt av flere kodende delsekvenser (exoner) atskilt av ikke-kodende sekvenser (introner)
- Felles transkripsjon av flere gener (operoner) er uvanlig, med noen få unntak (nematoder = rundormer)
- RBS rett oppstrøms for (= før) translasjonsstart : Kozak-sekvensen med konsensussekvens (GCC)[AG]CCATGG
- Start- og stoppkodoner som for prokaryoter

(GCC) = mindre konservert sekvens
[AG] = A eller G
CCA = konservert sekvens

To typer metoder

- Homologibaserte metoder
- Ab initio baserte metoder

Homologibaserte metoder

Sammenlikner sekvensen direkte med kjente gensekvenser.

Kan brukes til å finne gener med sekvens som likner det vi finner i et kjent gen.

Homologibaserte metoder

Ideen som utnyttes her er at to organismer (f.eks. menneske og mus) har et felles opphav – en "forfar-organisme" som begge stammer fra.

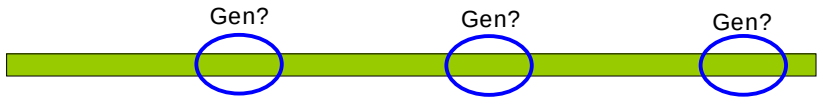
Dermed vil vi forvente at de to organismene fortsatt bærer på endel felles gener som stammer fra forfar-organismen.

Men etter at organismene A og B skilte lag kan:

- A og/eller B ha mistet noen av de opprinnelige genene
- A og/eller B ha ervervet seg nye gener
- De felles genene ha utviklet seg forskjellig i A og B (divergens)

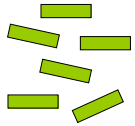
Homologibaserte metoder håndterer ikke de to første, men det siste punktet kan langt på vei håndteres med gode sammenstillingsmetoder.

Homologibaserte metoder



Sentralt spørsmål for denne typen metoder: når er sekvens-likheten stor nok til at vi kan fastslå slektsskap?

(Tema for senere forelesning)



Kjente gensekvenser (kan være fra andre organismer)

Ab initio baserte metoder

Sammenlikner ulike *egenskaper* ved sekvensen med tilsvarende egenskaper hos kjente gener.

Med slike metoder kan vi også finne gener som ikke har noen sekvenslikhet med kjente gener, såfremt de har andre fellestrekk med gener.

Forskjellige metoder

Vi har sett at prokaryoter og eukaryoter har tildels svært forskjellige egenskaper.

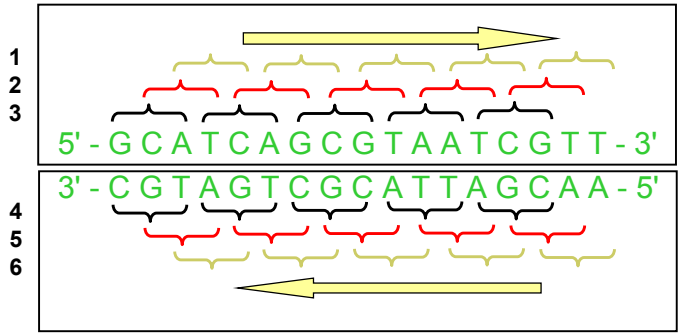
Ab initio metoder for genprediksjon i prokaryoter er derfor forskjellige fra tilsvarende metoder for eukaryoter.

Ab initio metoder : generelt

- **Leter etter spesielle kjennetegn for gener**, bl.a.
  - konserverte sekvenser nær start/slutt av genet
  - konserverte sekvenser i overgangene mellom introner og exoner
  - statistiske egenskaper hos exonene (basekomposisjon m.m.)
- **To hovedtyper** av ab initio metoder:
  - Signalsøk: Søk etter sekvenselementer som matcher gitte mønstre, f.eks. Shine-Dalgarno eller Kozak konsensussekvensen.
  - Innholdssøk: Leting etter sekvensregioner med statistiske egenskaper som matcher de vi finner i kjente gensekvenser, f.eks. med hensyn til hyppighet av GC i tredje base i kodoner.

Seks leserammer

Vi vet ikke på forhånd hvilken leseramme en proteinkodende sekvens benytter:



To gener i samme DNA-sekvens kan benytte forskjellige leserammer.

Ab initio prediksjon i prokaryoter

Strategi 1:

Finn alle åpne leserammer (= Open Reading Frames = ORF'er).

En ORF er en delsekvens som:

- Har baselengde som er delelig på 3
- Starter med startkodonet ATG
- Slutter med en av stoppkodonene TAA, TAG, TGA
- Ikke har noen stoppkodoner i midten

Ab initio prediksjon *forts.*

Problem: slike sekvenser finner en også i ikkekodende DNA!

Løsning: Anta for enkelhets skyld at ikkekodende DNA er helt tilfeldig sammensatt av A, T, G og C. I en gitt leseramme er da sannsynligheten for å unngå stoppkodon i k påhverandre følgende basetripler gitt ved $(61/64)^k$.

Betrakt en sekvens i ikkekodende DNA av formen:

62 eller flere basetripler uten stoppkodon

ATG XXX XXX XXX ..... XXX XXX TAG

Sannsynligheten for å finne dette er $(61/64)^{62} = 0.051$, altså svært lav! M.a.o. hvis vi begrenser ORF-søk til de som har lengde minst 64 basetripler, får vi bare noen få falske positive.

Ab initio prediksjon *forts.*

Strategi 2:

Søk etter Shine-Dalgarno sekvenser.

I praksis holder det å søke rett oppstrøms for de ORF'er som vi har identifisert ved strategi 1.

Ab initio prediksjon *forts.*

Strategi 3:

Velg en leseramme og lag et plott som viser hyppigheten av GC i tredje base. Gjør tilsvarende for alle andre leserammer.

Strategi 4:

Velg en leseramme og lag et plott som viser hyppigheten av repetisjon i tredje base. Gjør tilsvarende for alle andre leserammer.

Ab initio metoder i eukaryoter

Må da identifisere:

- Start og slutt for genet
- Start og slutt for hvert exon i genet

For å identifisere exoner benyttes typisk en *kombinasjon* av signalsøk (signaler ved genets start og slutt, spleisesignaler mellom introns og exons) og innholdssøk (f.eks. CpG-øyer).

Signalsøk i eukaryoter

Start- og stoppkodoner

Promotere

forteller RNA polymerase hvor transkripsjonen starter.
Ex: GC boks, TATA boks, CAAT boks.

Spleisesteder

exon-intron grenser (donorer) og intron-exon grenser (akseptorer). F.eks. kan akseptorer ha formen

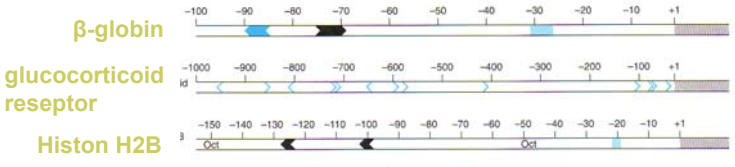
5' - PyPyPyPyPyPyNCAG - 3'

hvor Py $\in$ {C,T} og N $\in$ {A,C,G,T}.

Terminatorer, ribosomale bindingsseter, ...

Promotere

Eukaryote promotere består av en samling korte sekvenser plassert med relativt konstant avstand fra stedet hvor transkripsjonen starter. Fungerer som bindingssete for RNA polymerase for å initiere transkripsjon. Eksempler:



GC boks	:	GGGCGG
TATA boks	:	TATAAA
CAAT boks	:	CCAAT

Søk etter promotere og spleisesteder

- Promotere og spleisesteder er mer kompliserte å finne enn f.eks. start- og stoppkodoner, fordi det er betydelig sekvensvariasjon.
- Eksempel: konsensussekvensen for en av promoterne til *E.coli* er TATAAT, men i et studium av 263 promoter-regioner hadde ingen akkurat denne sekvensen.
- Søkemetoder for slike signaler inkluderer motif-baserte søkemetoder. Vi skal se nærmere på dette.

Motif-baserte søkemetoder

- Vi lager først en beskrivelse av domenet. Til det trenger vi et sett av kjente instanser (et "treningssett").
- Vi løpet deretter gjennom sekvensen med et vindu av fast lengde og gir hvert segment en score etter hvor godt det passer med beskrivelsen av domenet.
- Dette kan implementeres på flere måter. To sentrale klasser av metoder er bruk av
  - Posisjonsvektmatriser
  - Skjulte Markov modeller

Posisjonsvektmatriser

En posisjonsvektmatrise (profil) beregnes ut fra en multippel sammenstilling av sekvensene i treningssettet og estimerer sannsynligheten for at en bestemt nukleotid skal forekomme i en bestemt posisjon. Eksempel:

```
GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT
```

posisjoner →

	1	2	3	4	5	6	7	8	9
A	0.3	0.6	0.1	0.0	0.0	0.6	0.7	0.2	0.1
C	0.2	0.2	0.1	0.0	0.0	0.2	0.1	0.1	0.2
G	0.1	0.1	0.7	1.0	0.0	0.1	0.1	0.5	0.1
T	0.4	0.1	0.1	0.0	1.0	0.1	0.1	0.2	0.6

tabell over relative frekvenser

Posisjonsvektmatriser *forts.*

Anta at vi har gitt en sekvens:

ACTGTGCCC

Sannsynligheten for at en instans av domenet skal se slik ut:

0.3 * 0.2 * 0.1 * 1.0 * 1.0 * 0.1 * 0.1 * 0.1 * 0.2 = 0.0000012

Høy score betyr god overensstemmelse med domene-beskrivelsen (treningssettet).

	1	2	3	4	5	6	7	8	9
A	0.3	0.6	0.1	0.0	0.0	0.6	0.7	0.2	0.1
C	0.2	0.2	0.1	0.0	0.0	0.2	0.1	0.1	0.2
G	0.1	0.1	0.7	1.0	0.0	0.1	0.1	0.5	0.1
T	0.4	0.1	0.1	0.0	1.0	0.1	0.1	0.2	0.6

Posisjonsvektmatriser *forts.*

Tilfeldig valgte sekvenser kan også få positiv score. For hver sekvens S regner vi derfor ut

- sannsynligheten for å finne S i domenet:

  P(S | domene) = 0.3 * 0.2 * 0.1 * 1.0 * 1.0 * 0.1 * 0.1 * 0.1 * 0.2

- sannsynligheten for å finne S et tilfeldig sted:

  P(S | tilfeldig) = $p_A * p_C * p_T * p_G * p_T * p_G * p_C * p_C * p_C$

Forholdet mellom disse kalles en *likelihood ratio*. Ofte ser vi på logaritmen:

$$\lambda = \ln \frac{P(S \mid domene)}{P(S \mid tilfeldig)}$$

$\lambda < 0$: mest trolig tilfeldig
$\lambda = 0$: like trolig domene og tilfeldig
$\lambda > 0$: mest trolig domene

Posisjonsvektmatriser *forts.*

La posisjonsvektmatrisen være

	1	2	3	4	5	6	7	8	9
A	f_{A1}	f_{A2}	f_{A3}	f_{A4}	f_{A5}	f_{A6}	f_{A7}	f_{A8}	f_{A9}
C	f_{C1}	f_{C2}	f_{C3}	f_{C4}	f_{C5}	f_{C6}	f_{C7}	f_{C8}	f_{C9}
G	f_{G1}	f_{G2}	f_{G3}	f_{G4}	f_{G5}	f_{G6}	f_{G7}	f_{G8}	f_{G9}
T	f_{T1}	f_{T2}	f_{T3}	f_{T4}	f_{T5}	f_{T6}	f_{T7}	f_{T8}	f_{T9}

Da er log-likelihood ratioen til sekvensen S = $n_1 n_2 .... n_9$ gitt ved

$$\lambda = \ln \frac{P(S \mid domene)}{P(S \mid tilfeldig)} = \ln P(S \mid domene) - \ln P(S \mid tilfeldig)$$

$$= \sum_{i=1}^{9} \ln f_{n_i,i} - \sum_{i=1}^{9} \ln p_{n_i} = \sum_{i=1}^{9} \ln \left(\frac{f_{n_i,i}}{p_{n_i}}\right)$$

Posisjonsvektmatriser *forts.*

Definer verdier $h_{n_i,i} = \ln (f_{n_i,i}/p_{n_i})$ og erstatt verdiene i posisjons-vektmatrisen med disse:

	1	2	3	4	5	6	7	8	9
A	h_{A1}	h_{A2}	h_{A3}	h_{A4}	h_{A5}	h_{A6}	h_{A7}	h_{A8}	h_{A9}
C	h_{C1}	h_{C2}	h_{C3}	h_{C4}	h_{C5}	h_{C6}	h_{C7}	h_{C8}	h_{C9}
G	h_{G1}	h_{G2}	h_{G3}	h_{G4}	h_{G5}	h_{G6}	h_{G7}	h_{G8}	h_{G9}
T	h_{T1}	h_{T2}	h_{T3}	h_{T4}	h_{T5}	h_{T6}	h_{T7}	h_{T8}	h_{T9}

Dermed har f.eks. sekvensen ACTGTGCCC score

$$\lambda = \sum_{i=1}^{9} h_{n_i,i} = h_{A1} + h_{C2} + h_{T3} + h_{G4} + h_{T5} + h_{G6} + h_{C7} + h_{C8} + h_{C9}$$

Skjulte Markov modeller

- Posisjonsvektmatriser definerer en sannsynlighetsfordeling over alle mulige sekvenser av en gitt lengde: for hver sekvens kan vi finne dens sannsynlighet.
- En svakhet med metoden er at den antar uavhengighet mellom nukleotidene i en sekvens.
- Skjulte Markov modeller (Hidden Markov Models = HMM) definerer også en sannsynlighetsfordeling over alle mulige sekvenser av en gitt lengde, men antar ikke uavhengighet.

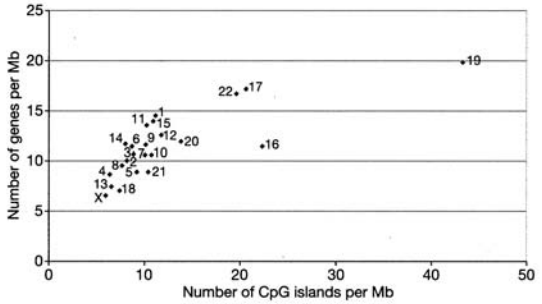
Innholdssøk i eukaryoter

- Søk etter CpG-øyer
- Sammenlikne hyppigheten av ulike basetripler med det vi forventer å finne i kodende og ikke-kodende regioner

CpG-øyer

Områder på 5'-siden av genet som har høyere forekomst av sekvensen 5' - CG - 3' enn ellers i genomet.

Det er en klar sammenheng mellom tetthet av CpG-øyer og tetthet av gener på kromosomet:



CpG-øyer

På genomisk skala er sekvensen

5' - CG - 3' (= CpG)

mer uvanlig hos pattedyr (ca 1%) i forhold til det en statistisk skulle forvente (ca 6%) hvis en ser på frekvensen av C'er og G'er.

Årsak: basen til C metyleres ofte og muterer til T

5' - C G - 3'
3' - G C - 5'

metylering
+
mutasjon

5' - T G - 3'
3' - A C - 5'

Viktig observasjon: Det finnes likevel mange "øyer" av ikkemetylert CpG i genomet, ca 2kb lange, der andelen av CG er nærmere det en skulle forvente statistisk. Og dessuten:

✓ CpG-øyer er ofte rett oppstrøms for gener

✓ Anslagsvis 50-60% av genene hos pattedyr har tilhørende CpG-øy.

Søk etter CpG-øyer

- Eksempel: metoden cpgplot i EMBOSS (European Molecular Biology Open Software Suite)
- Se på flere vinduer (=segmenter) av en gitt lengde L:

5' - A C G T C A C G C G T C G A C T G - 3' w=1

5' - A C G T C A C G C G T C G A C T G - 3' w=2

5' - A C G T C A C G C G T C G A C T G - 3' w=3

Typisk verdi: L = 200

- For hvert vindu beregn:

  GC-andel: (# GC) / (L-1)

  observert/forventet ratio = (GC-andel) / (forventet GC-andel)

  hvor

  (forventet GC-andel) = (#G)/L * (#C)/L

Søk etter CpG-øyer

- Et område defineres som en CpG-øy dersom
  - området har lengde ≥ 200 baser
  - GC andel er over 50%
  - observert/forventet ratio > 0.6
- De ulike parametrene ovenfor kan endres ved behov.

Søk etter CpG-øyer: cpgplot