Korrelasjon mellom genotyper i to lokus

StudForsk-prosjekt NTNU/Olav Thon stiftelsen

Veileder: Øyvind Bakke
Student: Johan Lundin

Innhold

Jeg vil takke Øyvind Bakke for innsikten og støtten han har gitt meg i dette prosjektet. I tillegg vil jeg takke Olav Thon Stiftelsen for den økonomiske støtten som gjør det mulig for oss studenter å gjennomføre interessante prosjekt via StudForsk.

Trondheim desember 2017

1 Introduksjon

Anta at to nabolokus k og $k+1$ har simultan sannsynlighetsfordeling p_{xy} for genotype x på lokus k og genotype y på lokus $k+1$, hvor $x = 0, 1, 2$, $y = 0, 1, 2$ og $\sum_x \sum_y p_{xy} = 1$. Definer marginalsannsynlighetene $p_{x\cdot} = \sum_y p_{xy}$ og $p_{\cdot y} = \sum_x p_{xy}$.

	0	1	2	
0	p_{00}	p_{01}	p_{02}	$p_{0\cdot}$
1	p_{10}	p_{11}	p_{12}	$p_{1\cdot}$
2	p_{20}	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot\cdot}$

(1)

1. La $\rho_{k,k+1} = \mathrm{Corr}(X, Y)$ for genotype X på lokus k og genotype Y på lokus $k+1$. Vi ønsker å se hvorvidt det er mulig å velge verdier p_{xy} for å få en gitt $\rho_{k,k+1} \in [-1, 1]$, og hvordan dette kan gjøres.

2. Anta at vi har flere lokus $k, k+1, k+2, \ldots$ vi ønsker å se hvorvidt det er mulig å velge simultanfordelingene for to og to nabolokus slik at $\rho_{k,k+1}$, $\rho_{k+1,k+2}, \ldots$ får ønskede verdier.

Følgende utledning gir en nyttig formel for $\rho_{k,k+1}$.

$$\begin{aligned}
\mathrm{Corr}(X, Y) &= \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \\
&= \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} \\
&= \frac{p_{11} + 2p_{12} + 2p_{21} + 4p_{22} - (p_{1\cdot}p_{\cdot 1} + 2p_{1\cdot}p_{\cdot 2} + 2p_{2\cdot}p_{\cdot 1} + 4p_{2\cdot}p_{\cdot 2})}{\sqrt{(p_{1\cdot} + 4p_{2\cdot} - (p_{1\cdot} + p_{2\cdot})^2)(p_{\cdot 1} + 4p_{\cdot 2} - (p_{\cdot 1} + p_{\cdot 2})^2)}} \\
\rho_{k,k+1} &= \frac{\frac{1}{4}(p_{11} - p_{1\cdot}p_{\cdot 1}) + \frac{1}{2}(p_{12} - p_{1\cdot}p_{\cdot 2}) + \frac{1}{2}(p_{21} - p_{2\cdot}p_{\cdot 1}) + p_{22} - p_{2\cdot}p_{\cdot 2}}{\sqrt{(\frac{1}{4}p_{1\cdot} + p_{2\cdot} - (\frac{1}{2}p_{1\cdot} + p_{2\cdot})^2)(\frac{1}{4}p_{\cdot 1} + p_{\cdot 2} - (\frac{1}{2}p_{\cdot 1} + p_{\cdot 2})^2)}}
\end{aligned} \tag{2}$$

Vi valgte å løse problemet numerisk ved å finne et intervall (ρ_{min}, ρ_{max}) for gitte marginalsannsynligheter. Algoritmene som er beskrevet i seksjon 2 finner disse intervallene.

2 Algoritmer

2.1 Korrelasjonsintervall for gitte marginalsannsynligheter

Algoritme 1 tar marginalsannsynlighetene for X og Y som parametere og finner største og minste mulige korrelasjon mellom X og Y. For-løkkenes steglengde kan økes (eller senkes) på bekostning av presisjon (eller kjøretid) i korrelasjonsintervallene.

Algoritme 1 Correlation interval

function CorrInterval($p_{0\cdot}, p_{1\cdot}, p_{2\cdot}, p_{\cdot 0}, p_{\cdot 1}, p_{\cdot 2}$)
  $corrMin \leftarrow 0$
  $corrMax \leftarrow 0$
  for $p_{00} \leftarrow max(0, p_{0\cdot} + p_{\cdot 0} - 1)$ **to** $min(p_{0\cdot}, p_{\cdot 0})$ **do** ▷ For-loops stepsize = 0.01
    for $p_{01} \leftarrow max(0, p_{0\cdot} - p_{\cdot 2} - p_{00})$ **to** $min(p_{0\cdot} - p_{00}, p_{\cdot 1})$ **do**
      for $p_{10} \leftarrow max(0, p_{\cdot 0} - p_{2\cdot} - p_{00})$ **to** $min(p_{1\cdot}, p_{\cdot 0} - p_{00})$ **do**
        for $p_{11} \leftarrow max(0, 1 - p_{2\cdot} - p_{\cdot 2} - p_{00} - p_{01} - p_{10})$ **to** $min(p_{\cdot 1} - p_{01}, p_{1\cdot} - p_{10})$ **do**
          $p_{02} \leftarrow p_{0\cdot} - p_{00} - p_{01}$
          $p_{12} \leftarrow p_{1\cdot} - p_{10} - p_{11}$
          $p_{20} \leftarrow p_{\cdot 0} - p_{00} - p_{10}$
          $p_{21} \leftarrow p_{\cdot 1} - p_{01} - p_{11}$
          $p_{22} \leftarrow p_{\cdot 2} - p_{02} - p_{12}$
          if $\rho_{k,k+1} > correMax$ **then**
            $corrMax \leftarrow \rho_{k,k+1}$
          else if $\rho_{k,k+1} < corrMin$ **then**
            $corrMin \leftarrow \rho_{k,k+1}$
          end if
        end for
      end for
    end for
  end for
  return $corrMin, corrMax$
end function

2.2 Korrelasjonsintervall for alle marginalsannsynligheter

I algoritme 2 er X og Y vektorene som henholdsvis inneholder marginalsannsynligheten $(p_{0\cdot}, p_{1\cdot}, p_{2\cdot})$ og $(p_{\cdot 0}, p_{\cdot 1}, p_{\cdot 2})$. Algoritmen løper gjennom alle mulige konfigurasjoner av marginalsannsynligheter og finner største og minste mulige korrelasjon ved bruk av algoritme 1. For-løkkenes steglengde kan økes (eller senkes), igjen på bekostning av presisjon (eller kjøretid). Vi brukte språket C++. Programmet ble kjørt på en maskin med CPU klokkehastighet 3.40 GHz (én kjerne brukt) og 8 GB RAM, og kjøretiden var omtrent 15 timer.

Algoritme 2 Correlation intervals for fixed marginal probabilities

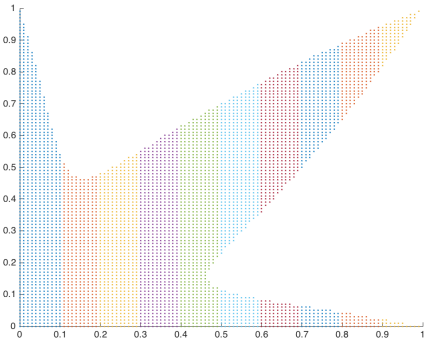
function CorrelationRunThroughMarginals
  $corrMin \leftarrow 0$
  $corrMax \leftarrow 0$
  $X \leftarrow (0, 0, 0)$
  $Y \leftarrow (0, 0, 0)$
  for $X(0) \leftarrow 0$ **to** 1 **do** ▷ For-loops stepsize = 0.01
    for $X(1) \leftarrow 0$ **to** $1 - X(0)$ **do**
      for $Y(0) \leftarrow 0$ **to** 1 **do**
        for $Y(1) \leftarrow 0$ **to** $1 - Y(0)$ **do**
          $X(2) \leftarrow (1 - X(0) - X(1))$
          $Y(2) \leftarrow (1 - Y(0) - Y(1))$
          $(corrMin, corrMax) =$CorrInterval$(X(0), X(1), X(2), Y(0), Y(1), Y(2))$
          (optional)$(textfile \leftarrow (X(0), X(1), X(2), Y(0), Y(1), Y(2), corrMin, corrMax))$
        end for
      end for
    end for
  end for
end function

3 Resultater

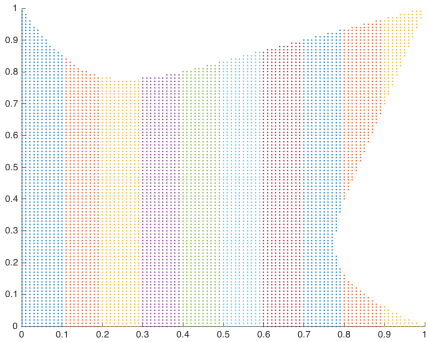
Resultatene foreligger elektronisk, med minimal og maksimal oppnåelig korrelasjon for alle kombinasjoner av marginalsannsynligheter i steg på 0.01.

I plottene under løper $p_{0\cdot}$ langs x-aksen og $p_{\cdot 0}$ langs y-aksen. Koordinat (a, b) er markert hvis den ønskede korrelasjonen kan oppnås med $p_{0\cdot} = a, p_{\cdot 0} = b$, altså det finnes verdier for $p_{1\cdot}, p_{2\cdot}, p_{\cdot 1}, p_{\cdot 2}$ slik at for $p_{0\cdot} = a, p_{\cdot 0} = b$ kan den ønskede korrelasjonen oppnås ved en viss konfigurasjon av p_{xy}.

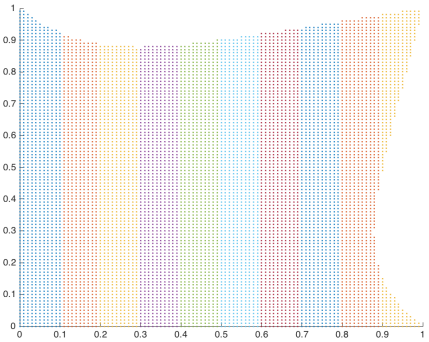
Plottene er laget i MATLAB.



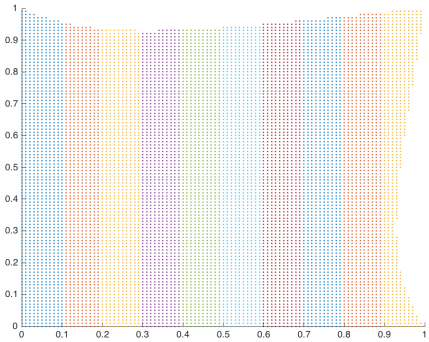
Figur 1: $\rho = 0.9$



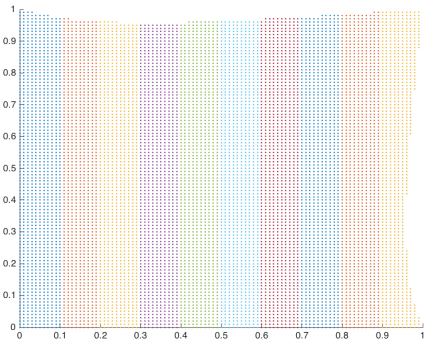
Figur 2: $\rho = 0.8$



Figur 3: $\rho = 0.7$

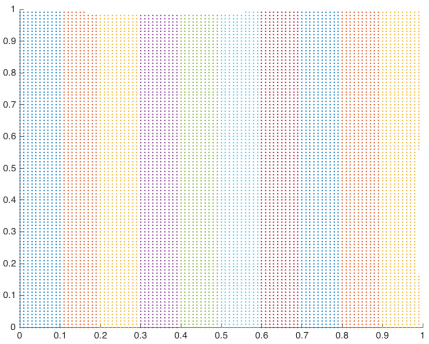


Figur 4: $\rho = 0.6$

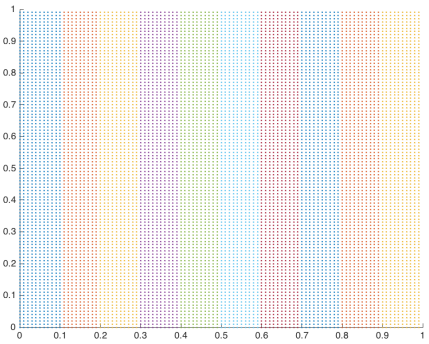


Figur 5: $\rho = 0.5$

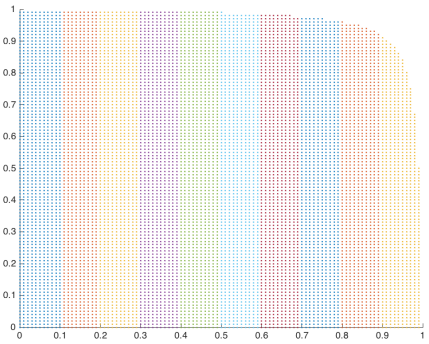
Figur 6: $\rho = 0.4$



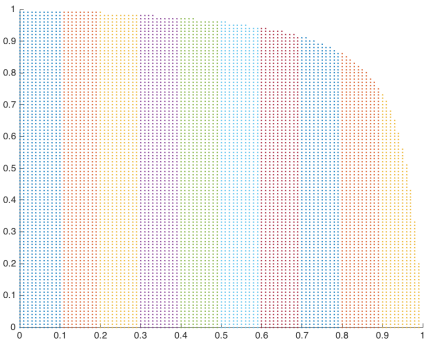
Figur 7: $\rho = 0.3$



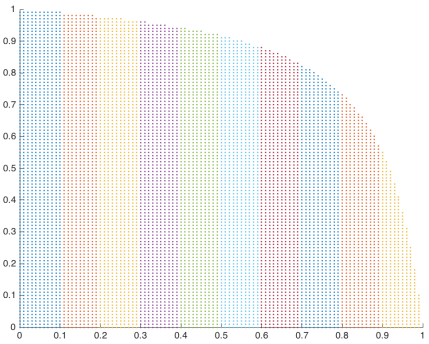
Figur 8: $\rho = 0.2, \rho = 0.1, \rho = 0.0$



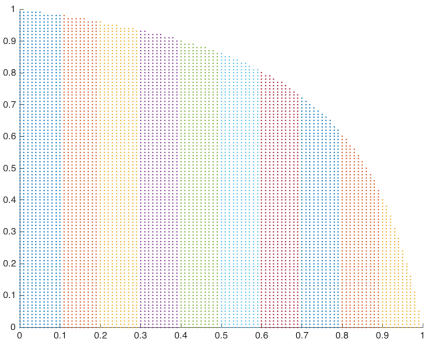
Figur 9: $\rho = -0.1$



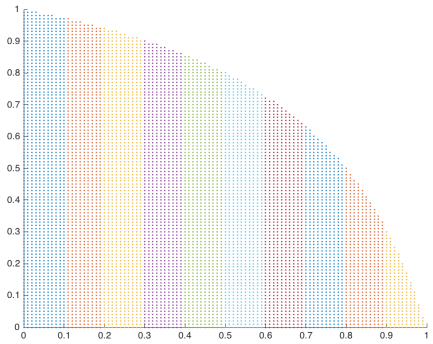
Figur 10: $\rho = -0.2$



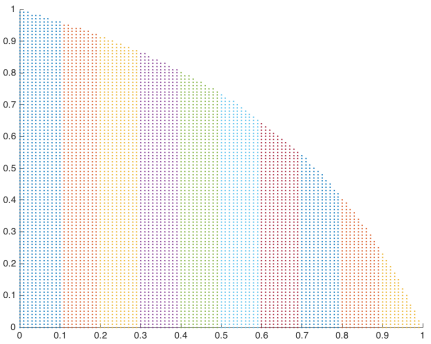
Figur 11: $\rho = -0.3$



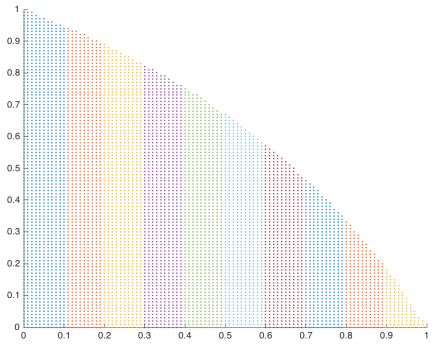
Figur 12: $\rho = -0.4$



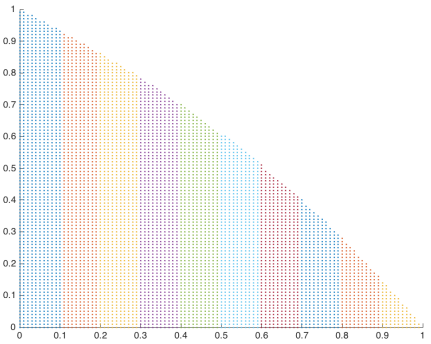
Figur 13: $\rho = -0.5$



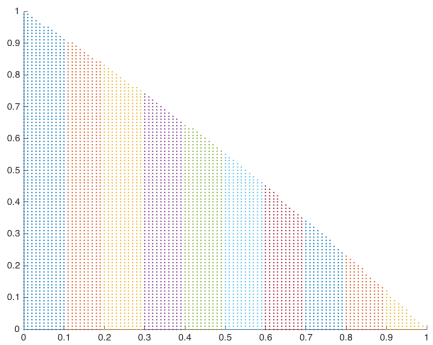
Figur 14: $\rho = -0.6$



Figur 15: $\rho = -0.7$



Figur 16: $\rho = -0.8$



Figur 17: $\rho = -0.9$

4 Diskusjon

Punkt 1 i seksjon 1 er i stor grad besvart av plottene. For en gitt korrelasjon $\rho_{k,k+1}$ gir plottene for hvilke verdier av $p_{0\cdot}$ og $p_{\cdot 0}$ den korrelasjonen er oppnåelig, og problemet reduseres til å finne verdier for $p_{1\cdot}, p_{2\cdot}, p_{\cdot 1}, p_{\cdot 2}$. Algoritme 1 kan enkelt modifiseres til å søke etter ønsket korrelasjon $\rho \pm \epsilon$ for $\epsilon > 0$ gitt $p_{0\cdot}$ og $p_{\cdot 0}$.

Det samme gjelder punkt 2. Gitt korrelasjonene $\rho_{k,k+1}, \rho_{k+1,k+2}$ for genotypene X, Y, Z i lokus $k, k+1, k+2$ indikerer plottene hvilke verdier $p_{0\cdot}, p_{\cdot 0}$ for X, Y som er mulige for $\rho_{k,k+1}$. Videre gir plottene hvilke verdier $p_{0\cdot}, p_{\cdot 0}$ for Y, Z som er mulige for $\rho_{k+1,k+2}$. Igjen kan algoritme 1 enkelt modifiseres til å lete etter simultanfordelingene som gir disse korrelasjonene.

Korrelasjonen $\rho_{k,k+1}$ er en kontinuerlig funksjon fra en sammenhengende delmengde av $\mathbb{R}^4$ til $\mathbb{R}$, så verdimengden er også kontinuerlig. Altså, for minimal og maksimal korrelasjon ρ_{min}, ρ_{max} kan alle korrelasjoner $\rho \in [\rho_{min}, \rho_{max}]$ oppnås.