

Kopplingsanalys - ett led i forskningen om folksjukdomars uppkomst¹

Lars Ängquist²

2003-03-05 / 2003-08-05

¹Denna populärvetenskapliga artikel är ursprungligen skriven, i ett något annorlunda format och utformande, som ett delmoment till doktorandkursen i kommunikationsteknik för naturvetenskapliga fakulteten VT-2003.

²Avdelningen för matematisk statistik, Matematikcentrum, Lunds universitet.

1 Kopplingsanalys

Ingen människa vill ha cancer. Helst slipper man nog också åldersdiabetes. Det är därför inte konstigt att det finns ett stort intresse för forskning som på något sätt kan kopplas till försök att lösa de komplicerade gåtorna bakom uppkomsten av olika folksjukdomar. Kopplingsanalys kan användas till att försöka finna gener som kan vara en del, större eller mindre, av ett potentiellt svar.

1.1 Introduktion

Inom kopplingsanalysen så söker man, med hjälp av olika typer av matematiska metoder, efter så kallade genetiska kopplingar. Detta innebär i praktiken att man försöker lokalisera olika positioner på genomet (gener) som kan vara *kopplade* till t.ex. olika typer av sjukdomar. En möjlig väg, för att uppnå detta, kan då beskrivas som ett sökande efter en sannolikhet.

1.2 En liten dos genetik

Arvsbetingade karakteristika hos en individ bestäms utifrån hennes nedärvda genetiska material (DNA). En enskild *gen* definieras som en DNA-sekvens inom en kromosom som kodar aminosyrasekvensen för en speciell del av ett protein. Vidare så definieras ett *lokus* som en position på en viss kromosom och en *allele* kan ses som en speciell sorts DNA-sekvens tillhörande ett specificerat lokus. Olika sekvenser tolkas därför som olika alleles.

En individ har en uppsättning av två olika *anlag*, av allelisk sort, per lokus. Ett anlag är nedärvt från modern och ett anlag härstammar från fadern. Dessa två anlag utgör individens *genotyp*. Med avseende på speciella karakteristika så kan man bestämma individuella, kvalitativa eller kvantitativa, *fenotyper*. En typisk tvådelad, sjukdomsrelaterad, fenotyp är kategoriseringen av individer som friska/sjuka.

Ett enkelt totalexempel (Tabell 1), relaterat till ovanstående definitioner, är den klassiska uppdelningen av mänskligheten i fyra olika blodgruppsklasser (fenotyper), med tillhörande sex olika typer av alleliska tvåkombinationer (genotyper).

Table 1: Genetik - blodgruppsexempel.

Alleler:	A, B, O.
Genotyper:	AA, AB, AO, BB, BO, OO.
Blodgrupper (fenotyper):	A (AA, AO) B (BB, BO) AB (AB) O (OO)

1.3 Nedärvning av anlag

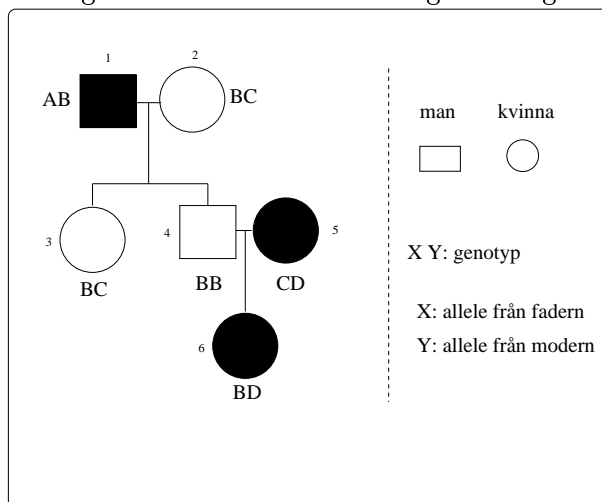
Enligt de Mendelska lagarna så ärver avkomman ned anlag från föräldrarna enligt ett helt slumpmässigt urval, dvs. det är 50 % chans att anlag 1 ärvs ner från modern (fadern) och 50 % chans att anlag 2 ärvs ned. Nedärvningen på olika kromosomer är helt oberoende, men om man simultant granskar en enskild kromosom så ser man att arvet sker blockvis. Detta innebär att man växelvis, över bitar av kromosomen, ärver ned anlag från mormor (farmor) respektive morfar (farfar). De nämnda växlingarna av anlagsgivare sker på slumpmässiga positioner, där så kallade *överkorsningar* har ägt rum.

Nedärvningen, gällande ett lokus x , hos en släkt med n stycken avkommor, kan beskrivas på ett enkelt sätt med hjälp av en binär så kallad *nedärvningsvektor*, $v(x)$, definierad som

$$v(x) = (p_1, m_1, \dots, p_n, m_n), \quad (1)$$

där p_1 (m_1) relateras till nedärvningen av anlag från faderns (moderns) sida för avkomma nummer 1 och kodas som en 0:a om anlaget härstammar från farfar (morfar) och som en 1:a om det härrör från farmor (mormor).

Figure 1: Genetik: nedärvning av anlag.



Om två identiska anlag, sådana som har ärvts ner från samma ursprungsperson, återfinns i genotypen hos två olika personer så säger man att dessa delar anlaget *IBD* (identical by descent). Som exempel kan man ta två helsyskon som kan dela 0-2 anlag IBD och två kusiner som enbart kan dela 0-1 anlag. Nedärvningen, beskriven av $v(x)$, används för att räkna ut IBD-delningen för datamaterialet. Detta ligger sedan till grund för den fortsatta analysen när man söker efter genetiska kopplingar. Ovan beskrivs ett enkelt nedärvningsexempel (Figur 1) där t.ex individ 5 och 6 delar en allele IBD, medan individ 3 och 6 med 50% sannolikhet delar en allele IBD och med 50% sannolikhet inte delar en enda enskild allele IBD.

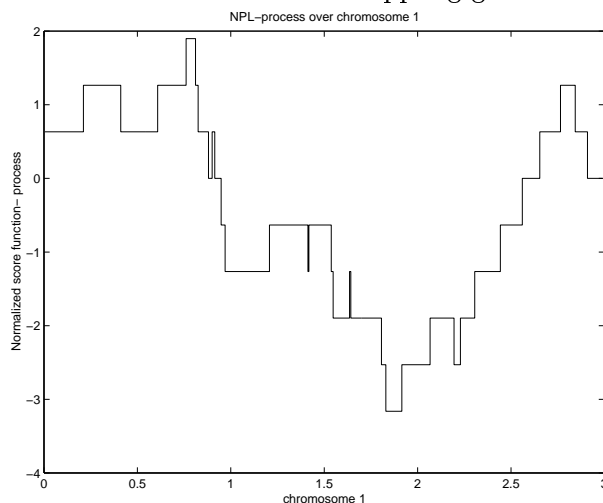
Vidare kan nämnas att det datamaterial som man har tillgängligt vid analys oftast består av nedärvningsdata för ett större antal olika släkter.

1.4 Vad innebär genetisk koppling?

Om individers fenotyper, i någon mån, bestäms av vilka anlag som ärvs ner på ett visst lokus så säger man att det finns en genetisk koppling mellan fenotyp och lokus. En sådan korrelation kan upptäckas genom att individer tillhörande en viss fenotyp onormalt ofta delar anlag IBD.

Hur avgör man då om det verkar finnas någon koppling närvarande? Fenotyp och intressanta kromosomområden definieras. Man introducerar också en så kallad *score-funktion*, $S(x)$, som numeriskt beräknar ett värde på hur pass gemensam nedärvningen är - givet fenotyp och lokus. Det finns en mängd olika föreslagna funktioner som alla har olika typer av fördelar och nackdelar. Värdet som man då finner kallas för den aktuella *NPL-scoren* (NPL=nonparametric linkage), och kontentan blir att höga värden blir ett tecken på en eventuell koppling och att låga värden tyder på att detta inte verkar föreligga. Ett exempel på en möjlig *NPL-process* ges i Figur 2, där lokus längs hela den första kromosomen undersöks.

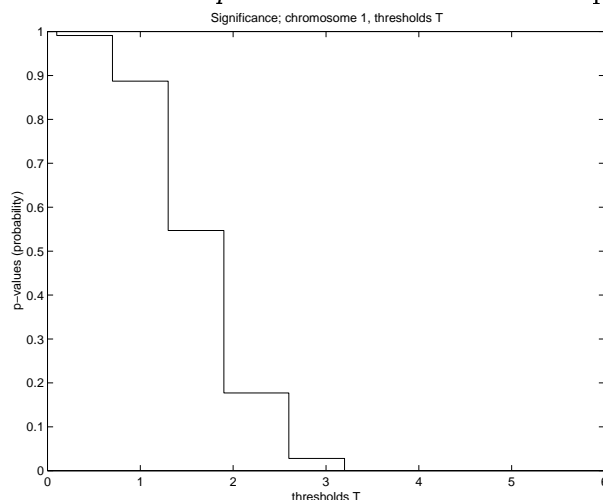
Figure 2: Matematik - sökandet efter koppling gällande kromosom 1.



Ett av de enklaste exemplen på en score-funktion är den så kallade parfunktionen, som här för enkelhets skull ej definieras helt formellt. Om vi tänker oss att den intressanta fenotypen är sjuka individer och att vi granskar lokus x , så beräknas $S_{pairs}(x)$ genom att välja ut alla möjliga par av sjuka individer och att sedan summera hur många anlag, totalt sett, som dessa delar IBD.

1.5 Sökandet efter sannolikheten

När man har beräknat NPL-värden för alla inkluderade lokus så intresserar man sig för det största värdet z_{\max} . Det lokus där detta resultat beräknades är då den position som tydligast visar tecken på genetisk koppling till den granskade fenotypen. Man vill nu beräkna ett så kallat *p-värde*. Detta sannolikhetsvärde är ett svar på frågan: Hur sannolikt skulle det vara att finna ett värde lika stort (eller större) än z_{\max} , någonstans längs de undersökta områdena, om någon verklig koppling ej existerar? Ett lågt värde på denna sannolikhet ger en indikation på att det är troligt att det kan finnas en koppling mellan denna position och den aktuella fenotypen. För ett exempel se figur 3 (samma material som förut).

Figure 3: Matematik - p -värden för ett antal värden på z_{\max} .

Det som nu har beskrivits är en grundläggande introduktion till forskningsområdet *icke-parametrisk kopplingsanalys* (man definierar ingen matematisk modell för hur sjukdomen är kopplad till nedärvningen). Eftersom man här främst jobbar statistiskt, med sannolikheter, så kan man med denna metod ej ge några definitiva svar. Dock så kan man, vilket man också har gjort, bidra till förståelsen kring vilka gener som har sina genetiska fingrar med i det mycket komplicerade miljö- och arvmässiga spel som ligger bakom några av våra vanligaste, och kanske forskningsmässigt mest intressanta, sjukdomar.