

Lars Ängquist, Matematikcentrum, Avdelningen för matematisk statistik, Lunds universitet, försvarade 16 mars sin doktorsavhandling Pointwise and Genomewide Significance Calculations in Gene Mapping through Nonparametric Linkage Analysis: Theory, Algorithms and Applications. Fakultetsopponent var docent Staffan Nilsson, Matematiska vetenskaper, Chalmers Tekniska Högskola, Göteborg.

Pointwise and Genomewide Significance Calculations in Gene Mapping through Nonparametric Linkage Analysis: Theory, Algorithms and Applications

av Lars Ängquist, Matematikcentrum, Avdelningen för matematisk statistik, Lunds universitet

I *kopplingsanalys* (linkage analysis), eller i en något mera generell mening vid genletning, så söker man efter sjukdomsgener längs ett *genom*. Här kan man tolka ett genom som en mängd av hela, eller bitar av, olika kromosomer. Med avseende på en mängd sammanhängande flergenerationella släkter så observerar man då, längs genomets kromosombitar, markördata, det vill säga *genotyper* bestående av nedärvda anlag (alleler) från fädernet respektive mödernet. Dessa observationer analyseras sedan tillsammans med iakttagelser gällande individernas *fenotyper*, med vilket här avses aktuell sjukdomsstatus (sjuka/friska/status okänd). Summan av kardemumman är att man vill försöka lokalisera sjukdomsgener genom att finna onormalt starka *kopplingar* mellan nedärvningen av anlag vid vissa kromosompositioner (lokus) och fördelningen av fenotyper över släkternas inkluderade individer. Detta vill man åstadkomma med så god precision som möjligt. En nyckelobservation är då att en, i någon mening, *signifikant avvikelse*, med avseende på kopplingen mellan genotyper och fenotyper, från vad som kan förväntas under hypotesen om *slumpmässig nedärvning* statistiskt sett tyder på en genetisk komponent kopplad till motsvarande observationslokus. (Begreppet slumpmässig nedärvning härstammar från Gregor Mendel.) En intressant avvikelse består vanligtvis av att de olika fenotypgrupperna internt delar fler nedärvda alleler, i någon mening, än vad som kan anses vara rimligt vid slumpmässig nedärvning med avseende på motsvarande lokus.

En försvinnande kort bakgrund

I introduktionen till min avhandling [1] så beskrivs de grundläggande och nödvändiga genetiska begreppen från den statistisk-genetiska disciplinen. Dessutom introduceras basala begrepp som, till exempel, *nedärvningsprocessen* av genetiska anlag, den genetiska *sjukdomsmodellstrukturen* som statistiskt beskriver ramarna för möjliga kopplingar mellan fenotyper och genotyper samt hur utbredd sjukdomen kan tänkas bli, och ibland även var aktuella sjukdomslokus är lokaliserade. Ytterligare begrepp som diskuteras är *datamaterialet* bestående av observerade släkter, *nedärvningsvektorn* som fullständigt beskriver hur nedärvningen av anlag har gått till i en specifik släkt och vidare olika sätt att beskriva mängden av tillgänglig *genetisk information*.

Efter detta så ges en introduktion till så kallad *enlokus icke-parametrisk kopplingsanalys*, där fokus ligger på *signifikansberäkningar* för en viss typ av teststatistika kallad för *NPL score* (Nonparametric linkage score). Be-

greppet icke-parametrisk syftar till att inget explicit antagande om strukturen av den genetiska modellen görs; enlokusanalys är ett uttryck för att man letar efter ett sjukdomslokus i taget längs det aktuella genomet. Vidare så utförs, vagt uttryckt, signifikansberäkningar i syfte att kvantifiera huruvida, vid analysen funna, intressanta resultat avviker, i en statistisk mening, tillräckligt mycket från det normala för att man skall våga tro på att man har hittat något sjukdomsrelaterat genområde. Sådana resultat är naturligtvis kopplade till ett specifikt lokus men metoden kan i princip, med trovärdighet, endast ange att man statistiskt sett har funnit ett sjukdomslokus inom ett område som omringar detta lokus. Vidare, om man letar efter sjukdomar som är kopplade till nedärvningen med avseende på två stycken sjukdomslokus så utför man en *tvålokusanalys*. Även vissa generaliseringar till detta utvidgade fall, samt kopplingar och skillnader till den alternativa analysmetoden *parametrisk kopplingsanalys*, ingår i introduktionen. Vilket kanske kan förstås från relaterad definition ovan så antas vid parametrisk analys kunskap (antagande) om underliggande sjukdomsmodell.

I den tredje delen av introduktionen så beskrivs översiktligt vissa angränsande och/eller alternativa samt kompletterande forskningsfält inom ramen för den statistisk-genetiska kontexten. Slutligen så sammanfattas innehållet i de i avhandlingen fyra olika inkluderade artiklarna. En fullständig avhandlingsintroduktion återfinns via www.maths.lth.se/matstat/staff/larsa/.

Vad vi har gjort (eller försökt att göra)

Allmänt kan sägas att om man letar efter gener över substantiellt stora kromosomområden så ger detta upphov till signifikansmässiga tolkningsproblem på grund av så kallad *multipel testning*. Huvudinriktningen för avhandlingens fyra papper är att på ett rimligt sätt utföra signifikansberäkningar (även i form av statistiska styrkeberäkningar) i olika situationer relaterade till såväl enlokus som tvålokus icke-parametrisk kopplingsanalys i samband med genomvid, i ovanstående mening, multipel testning.

I de två första artiklarna behandlas enlokusanalys vari det första (med Ola Hössjer) förbättrar och utvidgar vissa existerande *analytiska approximationer* för att utföra relaterade signifikansberäkningar [2, 3, 4]. Med analytiska approximationer så menas här att man härleder formler (slutna uttryck) så att man däri kan sätta in aktuella värden på inkluderade, fixa eller skattade, para-

metrar och således direkt få fram numeriska approximationsvärden. Själva uttrycken bygger här på *extremvärdesteori* för sannolikheten att överskrida höga trösklar med avseende på Ornstein-Uhlenbeck-likastokastiska processer. Vår metod bygger, till exempel, på en kvantilkoppling mellan standard normalfördelningen och en kontinuerlig version av den aktuella marginella NPL scorefördelningen. I slutändan leder detta till en korrigeringsmetod med avseende på normalavvikelse för den senare fördelningen.

Artikel nummer två (med Ola Hössjer) behandlar samma problematik, men här är approximationerna baserade på så kallade *Monte Carlo simuleringar* istället för beräkningar med hjälp av fasta analytiska approximationsformler. Denna typ av simuleringar innebär, löst uttryckt, att man slumpmässigt (exakt eller approximativt) genererar (simulerar) fram förlopp eller processer av den typ man är intresserad av och sedan analyserar utfallet av dessa förlopp. I det traditionella fallet med Monte Carlo simuleringar med avseende på genomvid icke-parametrisk kopplingsanalys så uppkommer i vissa fall en beräkningsmässig problematik då det tar, i någon mening, för lång tid att generera tillräckligt många förlopp som är analysmässigt intressanta. Detta beror på att den teststatistika, det vill säga den slumpvariabel som stoppas in i de analytiska eller simuleringsbaserade approximationsformlerna, då generellt sett alltför ofta antar för låga värden under vårt analysscenario, vår nollhypotes. För att lösa detta så inför vi, för att uttrycka det enkelt, ett slumpmässigt placerat artificiellt sjukdomslokus som då i allmänhet, vid simuleringar, leder till högre värden på teststatistikan i närheten av detta lokus. För att få en korrekt probabilistisk tolkning så korregerar vi för införandet av denna procedur genom att på ett visst sätt väga samman de olika förloppens resultat med avseende på approximationsformeln. Detta utförs med hjälp av så kallad *vägd simulering* (importance sampling). Här jämförs resultaten med avseende på beräkningstid för fix varians (cost-adjusted relative efficiency), med gott resultat, med den relaterade metoden beskriven i [5].

De två avslutande artiklarna riktar in sig på tvålokusanalys. Den första av dessa (med Dragi Anevski och Holger Luthman) behandlar så kallad *obetingad* tvålokusanalys, vilket innebär att man simultant (samtidigt) letar efter två olika sjukdomsgener. I vårt fall består här mängden av släkter enbart av så kallade sjuka syskonpar (affected sib-pairs); vi har med andra ord oss tillhanda ett homogent familjematerial vari varje familj består av ett par föräldrar och ett par affekterade barn till dessa föräldrar. En generell grundkontext målas upp med begreppsapparat samt diskussion av olika angreppssätt och olika typer av relaterade signifikansberäkningar (signifikansnivåer och styrka) med avseende på diverse möjliga situationer. Bland annat så tittar vi på sammansatta nollhypoteser i den meningen att vi tillåter inkluderande av (högst) ett sjukdomslokus vid motsvarande tvålokusdefinitioner. Vi härleder sedan, till exempel, en konservativ analytisk approximationsformel med avseende på *minst gynnsamma fördelning* (least favourable distribution) samt metoder

för att uppskatta statistisk signifikans genom att parameterskatta aktuell nollhypotes. Detta innebär, mer explicit, att bestämma instans med avseende på sammansättningen, det vill säga att skatta sjukdomsmodellen under nollhypotesen.

Slutligen, i den sista artikeln (med Ola Hössjer och Leif Groop), utvecklas ett generellt angreppssätt för signifikansberäkningar etcetera gällande så kallad *betingad* tvålokusanalys. Den betingade analysen kan ses som en hybrid mellan enlokus- och tvålokusanalys där man betingar med avseende på någon typ av *information* från ett första *betingningslokus* innan man sedan letar efter ett andra lokus. Här kan betingningslokusen vara givna a priori eller skattade utifrån en initial enlokusanalys. Vidare så kan informationen man betingar på vara enlokusresultat, från teststatistikan, i vårt fall i form av NPL scorer, eller motsvarande underliggande nedärvningsvektorer. Här har det första alternativet beskrivits i [6] medan det andra är en utvidgning utförd i denna artikel. Detta ger alltså upphov till sekventiella snarare än simultana tvålokusmetoder. Man kan också notera att av central betydelse i detta sammanhang är begreppet *icke-centralitetsparameter* (noncentrality parameter), vilket här enkelt uttryckt är ekvivalent med väntevärdet av aktuell teststatistika under en väldefinierad sjukdomsmodell (alternativhypotes). I artikeln härleds optimala versioner av NPL scoren med avseende på icke-centralitetsparametern. Man kan också notera att i detta sammanhang så är denna parameter nära besläktad med styrkefunktionen. ■

Referenser

- [1] Ängquist, L. (2007). *Pointwise and Genomewide Significance Calculations in Gene Mapping through Nonparametric Linkage Analysis: Theory, Algorithms and Applications*. Doctoral thesis 2006:15, Department of Mathematical Statistics, Lund University.
- [2] Feingold, E., Brown, P. O. and Siegmund, D. (1993). Gaussian Models for Genetic Linkage Analysis Using Complete High-Resolution Maps of Identity by Descent. *American Journal of Human Genetics* **53**, 234-251.
- [3] Lander, E. S. and Kruglyak, L. (1995). Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Results. *Nature Genetics* **11**, 241-247.
- [4] Tang, H. K. and Siegmund, D. (2001). Mapping Quantitative Trait Loci in Oligogenic Models. *Biostatistics* **2**, 147-162.
- [5] Malley, J. D., Naiman, D. Q. and Bailey-Wilson, J. E. (2002). A Comprehensive Method for Genome Scans. *Human Heredity* **54**, 174-185.
- [6] Cox, N. J., Frigge, M., Nicolae, D. L., Concannon, P., Hanis, C. L., Bell, G. I. and Kong, A. (1999). Loci on Chromosomes 2 (NIDDM1) and 15 Interact to Increase Susceptibility to Diabetes in Mexican Americans. *Nature Genetics* **21**, 213-215.