Tímarit um raunvísindi og stærðfræði
1. árg. 2. hefti 2003 – raust.is/2003/2/08

Venslagreining í fjölskyldum byggð á erfðamarkaröðum

Daníel Guðbjartsson

Íslenskri erfðagreiningu

Vefútgáfa: 19. nóvember 2003

Ágrip Venslagreining er mikið notuð aðferð við leit að meingenum. Aðferðin greinir tenglsaójafnvægi milli samsæta erfðamarka og sjúkdóms. Hefðbundin venslagreining einskorðast við samanburð á sjúklinga- og viðmiðunarhópum sem eru samansettir af óskyldum einstaklingum. Við setjum fram nálgunaraðferð sem tekur tillit til skyldleika sjúklinga og viðmiðunareinstaklinga og ræður við að greina vensl við erfðamarkaröð. Að auki sýnum við fram á að margfeldna sýndarlíkanið er markgildislíkan þegar tengslaójafnvægi við meingen minnkar, óháð sýndarlíkaninu á meingeninu sjálfu.

Inngangur

Flestir litningar æðri lífvera eru geymdir í tvíriti í kjarna hverrar frumu. Þessi tvö afrit eru að stærstum hluta eins, yfirleitt er meira en 99.9% samsvörun milli þeirra. Þegar erfðefni berst frá foreldrum til barns fær barnið eitt afrit frá hvoru foreldra sinna og lögmál Mendels segir að hvort foreldrið um sig láti barni sínu í té annað eintaka sinna og að helmings líkur séu á að sérhvert eintak erfist til barnsins. Kynlitningarnir, X og Y hjá mönnum, eru undantekningin frá því að allir litningar séu geymdir í tvíriti. Hér gerum við ráð fyrir að ekki sé verið að skoða kynlitning.

Þeir staðir litninga sem eru breytilegir meðal manna eru áhugaverðastir, því það eru litningar á slíkum stöðum sem valda því að mönnum er mishætt við að fá hina ýmsu sjúkdóma, þó flestir þeirra hafi engin áhrif á líkamlega starfsemi. Aðferð sem hægt er að nota til að greina arfgerð manneskju á stað sem er breytilegur er kölluð *erfðamark*, hver svo sem hún er. Mögulegar gerðir erfðamarks eru kallaðar *samsætur* þess. Eitt best þekkta erfðamarkið er blóðflokkarnir: A, B og O. Þar sem hver einstaklingur hefur tvær samsætur af hverju erfðamarki eru mögulegar arfgerðir: A/A, A/B, A/O, B/B, B/O og O/O. Mendelskar erfðir valda þá til dæmis því að barn foreldra með arfgerðir A/B og A/O mun hafa eina af arfgerðunum A/A, A/O, A/B eða B/O og að líkurnar á hverri arfgerð séu jafnar eða 0.25. Yfirleitt eru blóðflokkar greindir með því að athuga tilvist mótefna í blóði og því eru mögulegar svipgerðir aðeins færri en arfgerðir, eða A, AB, B og O, en þó eru til aðferðir sem greina blóðflokkana nákvæmlega. Flest nútíma erfðamörk eru þess eðlis að auðvelt er að greina arfgerðir nákvæmlega.

Ef skoðuð eru tvö erfðamörk sem eru mjög nærri hvort öðru á sama litningnum, þá er mögulegt að fylgni sé milli arfgerða einstaklings á þeim. Þessi fylgni er kölluð *tengslaójafnvægi*. Ef fylgni er á milli arfgerðar erfðamarks og líkunum á að einstaklingur fái sjúkdóm er sagt að *vensl* séu milli erfðamarksins og sjúkdómsins. Rannsókn á því hvort vensl séu milli erfðamarks og sjúkdóms er kölluð venslagreining. Hafa ber í huga að fylgni er ekki sama og orsakasamband, því að ekki er víst að erfðamörkin sjálf hafi bein líffræðileg áhrif heldur geta þau einfaldlega verið í tengslaójafnvægi við önnur erfðamörk sem hefur slík áhrif. Líkurnar á sjúkdómi að gefinni arfgerð eru kallaðar *sýnd* og líkan sem segir til um allar sýndir er kallað *sýndarlíkan*.

Á síðustu tíu til tuttugu árum hafa sprottið upp nýjar tegundir erfðamarka og nýjar leiðir til að greina þessi erfðamörk á ódýran og hagkvæman hátt þannig að raunhæft er orðið að gera sér vonir um að finna erfðamörk sem hafa vensl við sjúkdóma.

Oft er fjöldi erfðamarka arfgerðagreindur á litlu svæði og þá er hægt að fá auknar upplýsingar með því að skoða samsætur á nokkrum erfðamörkum saman. Listi yfir samsætur nokkurra ólíkra erfðamarka er kallaður *erfðamarkaröð*.

Í mörgum tilfellum má gera venslagreiningu öflugri með því að skoða vensl milli erfðamarkaraða og sjúkdóms í stað þess að skoða vensl milli einstakra erfðamarka og sjúkdóms. Hafa ber í huga að með nútíma arfgerðagreiningu eru arfgerðir einstakra erfðamarka skoðaðar einar sér sem þýðir að erfðamarkaraðir sjást ekki með beinum hætti. Sem dæmi má hugsa sér að tvö aðliggjandi erfðamörk séu til rannsóknar, bæði með tvær samsætur. Það fyrra með A og B og það seinna með 1 og 2. Þá eru mögulegar erfðamarkaraðir A1, A2, B1 og B2. Ef við fáum einstakling með arfgerð A/B á fyrra erfðamarkinu og 1/2 á því seinna þá eru tveir möguleikar: Hann getur haft erfðamarkaraðirnar A1 og B2 eða hann getur haft erfðamarkaraðirnar A2 og B1. Þessi aukna óvissa er helsta ástæða þess að flækjustig reiknirita sem byggja á erfðamarkaröðum er hærra en hinna sem byggja á stökum erfðamörkum.

Í 1. kafla er rifjað upp í smáatriðum hvernig hefðbundin venslagreining er framkvæmd. Í 2. kafla er venslagreiningin útvíkkuð til erfðamarkaraða og almennra fjölskyldna. Í 3. kafla er margfeldna sýndarlíkanið skilgreint og sýnt að það er markgildislíkan þegar vensl milli sjúkdóms og erfðamarka minnka. Í 4. kafla er loks sýnt hvernig nálga má útreikninga með Monte Carlo hermun.

1. Hefðbundin venslagreining

Erfðaefni úr N sjúklingum og M viðmiðunareinstaklingum er greint fyrir eitt erfðamark sem hefur tvær samsætur, 1 og 2. Mögulegar arfgerðir í þessu tilfelli eru þá 1/1, 1/2 og 2/2. Látum X vera fjölda eintaka samsætu 1 sem greinast í sjúklingunum og Y vera fjölda eintaka sömu samsætu sem greinast í viðmiðunareinstaklingunum. Ef sjúklingarnir og viðmiðunareinstaklingarnir eru óskyldir þá hafa X og Y tvívalsdreifingu með stika p og $2N$ annars vegar og q og $2M$ hins vegar, þar sem p er þýðistíðni samsætu 1 í sjúklingum og q er þýðistíðni samsætu 1 í viðmiðunareinstaklingum. Markmiðið er að sýna fram á vensl milli sjúkdómsins og erfðamarksins. Ef engin vensl eru til staðar þá eru þýðistíðnirnar þær sömu, $p = q$, svo að við setjum fram eftirfarandi núll- og gagntilgátu til að prófa tilvist vensla:

$$H_0 : p = q = p_0 \quad \text{og}$$
$$H_A : p \neq q,$$

þar sem p_0 er sameiginlegur tvívalsstiki X og Y undir núlltilgátunni H_0. Þetta próf er nýtanlegt til að sýna fram á tilvist vensla ef sjúklingar og viðmiðunareinstaklingar eru valdir úr sama þýðinu. Ef þeir eru ekki úr sama þýðinu er túlkunin um tilvist vensla ef núlltilgátunni er hafnað ekki lengur gild þar sem hverskonar munur á þýðununm sem sjúklingarnir og viðmiðunareinstaklingarnir voru dregnir úr getur orðið til þess að breyta þýðistíðni samsætu 1.

Fjölmargar leiðir eru til að prófa þessar tilgátur. Mikið notuð leið byggir á líknafallinu,

$$L(p, q) = p^X (1 - p)^{2N-X} q^Y (1 - q)^{2M-Y}. \tag{1}$$

Ef N og M eru stór þá hefur líknahlutfallið, $2 \log \dfrac{L(\hat{p}, \hat{q})}{L(\tilde{p}_0, \tilde{p}_0)}$, u.þ.b. χ^2-dreifingu með eina frígráðu, þar sem $\hat{p} = \frac{X}{2N}$, $\hat{q} = \frac{Y}{2M}$ og $\tilde{p}_0 = \frac{X+Y}{2N+2M}$ eru hálíknamöt p, q og p_0.

2. Útvíkkun líknafallsins fyrir fjölskyldur og erfðamarkaraðir

Margir hafa orðið til að útvíkka líknafallið í (1) til að taka tillit til erfðamarkaraða og fjölskyldna (t.d. [5] og [6]), en það hefur þó ætíð verið gert í samhengi við einfaldar kjarnafjölskyldur, þ.e. systkinahópa. Allar útvíkkanirnar eru jafngildar, enda um náttúrulega útvíkkun að ræða. Hins vegar er misjafnt hversu langt þær ganga eftir því hversu flóknar fjölskyldur eru til rannsóknar.

Látum A tákna sjúkdómsgreiningar allra einstaklinga sem eru til rannsóknar. Í flestum erfðarannsóknum er rétt að skilyrða á A í stað þess að líta á A sem hluta gagnanna. Ekki þurfti að taka það fram fyrir (1), en fyrir almennar fjölskyldur er það nauðsynlegt.

Sérhverri erfðamarkaröð h tengjum við þýðistíðni θ_h. Látum β vera sýndarlíkan sem gefur sérhverri arfgerð h, k sýnd, þ.e. skilgreinir $P(A|\beta, h, k)$. Köllum niðurstöður allra arfgerðagreininga g. Almennt veitir g ekki fullkomnar upplýsingar um erfðamarkaröð sérhvers einstaklings og því þarf meira til að geta skrifað niður líknafall fyrir stikana θ og β. Til þess skilgreinum við stærðir sem innihalda fullkomnar upplýsingar um hinar óséðu erfðir. Látum z innihalda upplýsingar um erfðamarkaraðir allra forfeðra, þ.e. allra sem ekki eiga foreldra í fjölskyldunum sem eru til rannsóknar, og v vera upplýsingar um hvernig erfðaefni berst milli kynslóða í þessum sömu fjölskyldum. Gefið z og v, eru g og A óháðar stærðir, þ.e.

$$P(g, A|\beta, \theta, z, v) = P(g|\beta, \theta, z, v)\, P(A|\beta, \theta, z, v)\,.$$

Að auki gildir að $P(g|\beta, \theta, z, v) = P(g|z, v)$ og að $P(A|\beta, \theta, z, v) = P(A|\beta, z, v)$. Fyrri jafnan leiðir af því að gefið z skipta þýðistíðnirnar θ ekki máli, auk þess er g aðeins háð β í gegnum A svo að í fjarveru þess er g óháð β. Seinni jafnan gild vegna þess að gefið z skiptir θ ekki máli.

Við getum nú skrifað líknafallið fyrir β og θ sem

$$\begin{aligned} L(\beta, \theta; g|A) = P(g|\beta, \theta, A) = \frac{P(g, A|\beta, \theta)}{P(A|\beta, \theta)} &= \frac{1}{P(A|\beta, \theta)} \sum_{z,v} P(z, v|\beta, \theta)\, P(g, A|\beta, \theta, z, v) \\ &= \frac{1}{P(A|\beta, \theta)} \sum_{z,v} P(z|\theta)\, P(v)\, P(g|z, v)\, P(A|\beta, z, v)\,. \quad (2) \end{aligned}$$

3. Margfeldið sýndarlíkan

Líknafallið í (1) er jafngilt margfeldna sýndarlíkaninu, þ.e.

$$P(A|\beta, h, k) \propto \beta_h \beta_k.$$

Þetta líkan hefur ýmsa kosti. Í fyrsta lagi fækkar það frígráðum úr $\frac{n(n-1)}{2}$ í n, þar sem n er fjöldi mögulegra erfðamarkaraða. Reyndar fækkar frígráðunum í $n-1$ í flestum tilfellum, þar sem yfirleitt er aðeins hægt að greina fólk sem sjúkt en ekki sem heilbrigt og því er hægt að velja eina erfðamarkaröð sem grunnerfðamarkaröð og miða breytingu á sýnd við þá erfðamarkaröð. Í öðru lagi einfaldar það alla útreikninga. Í þriðja lagi hefur verið sýnt fram á að margfeldna sýndarlíkanið er ekki viðkvæmt fyrir því að vera ekki rétt líkan [5]. Í fjórða lagi sýnum við fram á að það er markgildislíkan eftir því sem tengslajafnvægi erfðamarkanna sem eru til rannsóknar við meingen, sem hefur bein áhrif á sjúkdóminn, minnkar.

Til að sýna fram á þetta látum við sem svo að eitt erfðamark með tvær samsætur, H og h, sé til rannsóknar og að meingenið hafi aðeins tvær samsætur, D og d. Þýðistíðni samsætu H táknum við með g og samsætu D með f. Látum p_0, p_1 og p_2 vera sýndina gefið að einstaklingur sé með 0, 1 eða 2 eintök af D samsætunni, og á sama hátt q_0, q_1 og q_2 vera sýndina gefið að einstaklingur sé með 0, 1 eða 2 eintök af samsætu H af erfðamörkunum. Tengslaójafnvægisstikinn δ er skilgreindur á hefðbundinn hátt með jöfnunum:

$$\begin{aligned} P(HD) &= gf + \delta, \\ P(Hd) &= g(1-f) - \delta, \\ P(hD) &= (1-g)f - \delta, \\ P(hd) &= (1-g)(1-f) + \delta. \end{aligned}$$

Þegar δ er 0, þá er erfðamarkið og meingenið í tengslajafnvægi og þar með sjúkdómurinn.

Nú gildir að

$$
\begin{aligned}
q_0 &= \frac{1}{(1-g)^2}\left(P\left(hD\right)^2 p_2 + 2P\left(hD\right)P\left(hd\right)p_1 + P\left(hd\right)^2 p_0\right)\\
&= R - \frac{R'}{(1-g)}\delta + K_0\delta^2,\\
q_1 &= \frac{1}{g(1-g)}\Big(P\left(HD\right)P\left(hD\right)p_2 +\\
&\quad \left(P\left(HD\right)P\left(hd\right) + P\left(Hd\right)P\left(hD\right)\right)p_1 + P\left(Hd\right)P\left(hd\right)p_0\Big)\\
&= R + \frac{1}{2}\left(\frac{R'}{g} - \frac{R'}{1-g}\right)\delta + K_2\delta^2,\\
q_2 &= \frac{1}{g^2}\left(P\left(HD\right)^2 p_2 + 2P\left(HD\right)P\left(Hd\right)p_1 + P\left(Hd\right)^2 p_0\right)\\
&= R + \frac{R'}{g}\delta + K_2\delta^2,
\end{aligned}
$$

þar sem K_0, K_1 og K_2 eru fastar óháðir δ, og

$$
\begin{aligned}
R &= f^2 p_2 + 2f(1-f)p_1 + (1-f)^2 p_0,\\
R' &= \frac{\partial R}{\partial f} = 2\left(fp_2 + (1-2f)p_1 - (1-f)p_0\right).
\end{aligned}
$$

Þessar útvíkkanir gefa fyrsta stigs nálganir á $\log(q_2/q_1)$ og $\log(q_1/q_0)$ sem

$$
\begin{aligned}
\log(q_2/q_1) &= \frac{1}{g(1-g)}\frac{R'}{2R}\delta + O(\delta^2) \quad \text{og}\\
\log(q_1/q_0) &= \frac{1}{g(1-g)}\frac{R'}{2R}\delta + O(\delta^2).
\end{aligned}
$$

Þetta sýnir að sýndirnar stefna á margfeldna sýndarlíkanið hraðar en þær stefna á 0, auk þess að gefa sjálft markgildislíkanið.

4. Nálgun líknafallsins með Monte Carlo hermun

Hefðbundnar aðferðir sem nýta sér upplýsingar frá mörgum erfðamörkum samtímis ([4, 3, 1]) gera ráð fyrir að öll erfðamörk séu í tengslajafnvægi. Þær er því ekki hægt að nota með beinum hætti við mat á líknafallinu í (2). Við stingum upp á að nálga líknafallið með Monte Carlo hermun undir forsendum tengslajafnvægis og með því að endurvigta hermuðu gildin á viðeigandi hátt. Þessi aðferð er skyld mikilvægishermun [2].

Lykilatriði í notagildi Monte Carlo hermunarinnar er sú staðreynd að líkurnar á g gefið z og v eru óháðar því hvort gert sé ráð fyrir tengslajafnvægi eða ekki. Með öðrum orðum er $P\left(g|z,v\right) = P_E\left(g|z,v\right)$, þar sem P er byggt á tengslaójafnvægi og P_E er byggt á tengslajafnvægi. Af þessu, og því að hvernig erfðaefni berst milli kynslóða er óháð því hvort gert sé ráð fyrir tengslaójafnvægi eða tengslajafnvægi, þ.e. $P\left(v\right) = P_E\left(v\right)$, leiðir

að líknafallið í (2) má setja fram sem

$$\begin{aligned} L(\beta,\theta;g|A) &= \frac{1}{P\left(A|\beta,\theta\right)}\sum_{z,v} P\left(z|\theta\right)P\left(v\right)P\left(g|z,v\right)P\left(A|\beta,z,v\right) \\ &= \frac{1}{P\left(A|\beta,\theta\right)}\sum_{z,v} P\left(z|\theta\right)P\left(v\right)P_E\left(g|z,v\right)P\left(A|\beta,v,z\right) \\ &= \frac{1}{P\left(A|\beta,\theta\right)}\sum_{z,v} P\left(z|\theta\right)P\left(v\right)\frac{P_E\left(z,v|g\right)P_E\left(g\right)}{P_E\left(z,v\right)}P\left(A|\beta,z,v\right) \\ &\propto \frac{1}{P\left(A|\beta,\theta\right)}\sum_{z,v} \frac{P\left(z|\theta\right)}{P_E\left(z\right)}P_E\left(z,v|g\right)P\left(A|\beta,z,v\right). \end{aligned}$$

Nú er ljóst hvernig má nálga líknafallið með því að herma z og v úr $P_E\left(z,v|g\right)$. Áður hefur verið sýnt fram á hvernig má herma slík gildi á reikningslega fýsilegan og hagkvæman hátt [1].

Summary: Association studies are being used extensively in the search of genes that modify the risk of contracting diseases. These studies detect linkage disequilibrium between alleles of genetic markers and diseases. Classical case-control studies are expanded to include general families and haplotypes. We also show that the penetrance model based on a biallelic marker, in linkage disequilibrium with a risk modifying mutation, tends to the multiplicative model as the disequilibrium between the marker and the mutation decreases, regardless of the penetrance model at the site of the mutation itself.

Heimildir

[1] Daníel F. Guðbjartsson, Kristján Jónasson, M. L. Frigge og A. Kong, *Allegro, a new computer program for multipoint linkage analysis*, Nature Genetics, **25:1**, 12-13 (2000).
[2] J. M. Hammersley og D. C. Handscomb, *Monte Carlo Methods*, Wiley, (1964)
[3] L. Kruglyak og M. J. Daly, M. P. Reeve-Daly og E. S. Lander, *Parametric and nonparametric linkage analysis: a unified multipoint approach*, American Journal of Human Genetics, **58:6**, 1347-1363, (1996).
[4] E. S. Lander og P. Green, *Construction of multilocus genetic linkage maps in humans.*, Proc Natl Acad Sci U S A, **84:8**, 2363-2367, (1987).
[5] D. J. Schaid, *General score tests for associations of genetic markers with disease using cases and their parents*, Genet Epidemiol., **13:5**, 423-449 (1996).
[6] A. S. Whittemore og I. P. Tu, *Detection of disease genes by use of family data. I. Likelihood-based theory*, American Journal of Human Genetics, **66:4**, 1328-1340, (2000).

Um höfundinn: Daníel Fannar Guðbjartsson er fæddur 1973 í Hafnarfirði. Hann lauk stúdentsprófi frá Fjölbrautaskóla Suðurnesja 1993, BS prófi í stærðfræði frá Háskóla Íslands 1996 og doktorsprófi í tölfræði frá Duke-háskóla 2001. Daníel hefur starfað hjá Íslenskri erfðagreiningu frá 1999.

Íslensk erfðagreining
Sturlugötu 8
IS-101 Reykjavík
dfg@decode.is

Móttekin: 15. mars 2002