MIKROBIOLOGISEN SELEKTION MATEMAATTISISTA PERUSTEISTA

EEVA KOSKENNIEMI ja HELGE GYLLENBERG

Helsingin yliopiston mikrobiologian laitos

Saapunut 27. 11. 1969

Mikro-organismien tehokkaan hyväksikäytön edellytyksenä maataloudessa ja teollisuudessa on useimmiten, että tavalla tai toisella hyödylliseksi havaitun organismin ihmiselle edullinen aktiivisuus voidaan lisätä moninkertaiseksi alunperin todetusta. Tähän antaa mahdollisuuden mikro-organismien nopea lisääntyminen, suhteellisen korkea spontaanien mutaatioiden frekvenssi ja alttius erilaisille mutageenisille vaikutuksille. Mikro-organismien viljelmät — ns. puhdasviljemätkin — ovat siten geneettisesti varsin heterogeenisiä ja toivotun aktiivisuuden tehostamista voidaankin edistää tarkoituksenmukaisella valinnalla, s e l e k t i o l l a. Toisaalta samat tekijät, jotka myötävaikuttavat aktiivisuuden lisääntymiseen ovat vaikuttamassa myös päinvastaisesti, so. toivottuja ominaisuuksia heikentävästi. Erilaisia selektiotoimenpiteitä joudutaan senvuoksi yleensä soveltamaan myös jo saavutetun tyydyttävänä pidetyn aktiivisuusasteen ylläpitämiseksi.

Huolimatta siitä, että selektiomenetelmät liittyvät jokseenkin itsestäänselvinä mikro-organismien ja nimenomaan yksittäisten mikrobiviljelmien hyväksikäyttöön, mikrobiologian oppi- ja käsikirjat sisältävät varsin vähän osviittaa selektion matemaattisista perusteista ja käytännön sovellutuksista. Selektiomenetelmät ovat senvuoksi enemmän tai vähemmän »omatekoisia». On selvää, että tuntematta menetelmän matemaattisia perusteita ei vallitsevissa olosuhteissa edullista tulosta voida saavuttaa. Tämä puute muodostuu sitäkin arveluttavammaksi, koska selektiotyötä varten ei yleensä ole käytettävissä rajattomasti laboratoriovälineistöä ja aikaa. Erityisen pullonkaulan muodostaa useasti tarjolla oleva tila ravistelupöydässä, joka voidaan ilmaista pöytään sijoitettavissa olevien viljelyastioiden lukumääränä. Näin selektiotehtävä muodostuu optimointiongelmaksi: käytettävissä olevan tilan ja ajan puitteissa on saavutettava paras mahdollinen tulos.

Kun selektion matemaattisia perusteita koskeva informaatio on julkaistu pääasiassa tilastotieteellisessä ja biometrisessä kirjallisuudessa, joka ei yleensä ole mikrobiologien ulottuvilla, eikä aina helposti omaksuttavissa, kirjoittajat ovat pitäneet hyödyllisenä koota tärkeimmät kohdat tästä informaatiosta suomenkielisenä esitettäväksi.

Selektiotapa

Selektion tarkoituksena on löytää tutkijan kannalta edullisia mikro-organismeja. Edullisuus riippuu monista eri tekijöistä, mutta yleensä on rajoituttava tutkimaan kulloinkin kysymyksessä olevan tavoitteen kannalta tärkeintä ominaisuutta, jota seuraavassa sanotaan a k t i i v i s u u d e k s i. Useamman ominaisuuden tarkastelu ei yleensä muuta olennaisesti selektiomenetelmää. Tästä on esimerkkinä COCHRANin (1951) julkaisu.

Selektiossa suoritetaan valintaa saman organismin useamman viljelmän välillä. Seuraavassa sovelletaan termiä k a n t a yksittäisiin viljelmiin, joiden alkuperä ja yleiset ominaisuudet ovat tunnettuja.

Valinta suoritetaan käyttäen kriteerinä aktiivisuutta koeolosuhteissa. Todellisia aktiivisuuden arvoja ei tiedetä, joten päätökset on tehtävä havainnoista saatujen estimaattien avulla. Kun lopullinen valinta suoritetaan, on tämän vuoksi aina valittava useampi kuin yksi kanta, joita voidaan kokeilla käytännössä.

Päätös tutkittavan kannan hyväksymisestä tai hylkäämisestä voidaan tehdä kolmen periaatteessa erilaisen analyysin avulla:

— Kannan aktiivisuus mitataan ja tehdään heti lopullinen päätös.
— Kannan aktiivisuus mitataan. Estimaatin perusteella päätetään hylätäänkö kanta vai suoritetaanko uusi tutkimus. Saatujen tietojen perusteella tehdään päätös.
— Kanta tutkitaan useassa eri vaiheessa. Jokaisen vaiheen jälkeen päätetään siihen mennessä saadun informaation perusteella joko hylätä tai hyväksyä kanta, tai suoritetaan uudet mittaukset. Tätä sanotaan sekvenssimenetelmäksi.

Ensimmäinen menettely on aivan kiinteä. Jokainen tutkittava käsitellään samalla tavalla. Toisessa menetelmässä kiinnostavat kannat tutkitaan tarkemmin kuin heti huonoiksi todetut. Tutkimisvaiheita voi olla useampi kuin kaksi, mutta niiden maksimimäärä on vakio. Kolmannessa menetelmässä muuttuu vaiheitten määrä tapauksen mukaan. Tätä pidetään useassa tapauksessa taloudellisimpana menetelmänä.

GURNOW (1962) on selostanut useita selektiomenetelmiä, jotka on kehitetty lähinnä kasvinjalostusta varten. Sekvenssimenetelmiä on selostettu WALDIN (1947) kirjassa. Sekvenssimenetelmän sovellutuksesta biologiassa on esimerkki DAVIESin (1958) julkaisussa.

Selektion päämäärä

Kun selektiomenetelmä on tarkoin selvillä voidaan periaatteessa laskea todennäköisyys sille, että aktiivisuuden x omaava kanta tulee hyväksytyksi. Jos edellä mainittua todennäköisyyttä merkitään $\eta(x)$:llä, ja tutkittavien kantojen aktiivisuuksien jakautuma on f(x)dx, hyväksytään kannoista keskimäärin P:s osa kun

$$P = \int_{-\infty}^{+\infty} \eta(x) f(x) dx.$$

Todennäköisyyttä P sanotaan selektiointensiteetiksi.

Hyväksyttyjen kantojen odotusarvo tulee olemaan

$$M = \frac{1}{P} \int_{-\infty}^{+\infty} x \, \eta(x) f(x) dx.$$

Kantoja, joiden aktiivisuus on vähintään α on tutkittavien joukossa keskimäärin P_g:s osa, kun

$$P_g = \int_{\alpha}^{+\infty} f(x)dx.$$

Sellaisia kantoja, joiden aktiivisuus on vähintään α hyväksytään tutkittavista keskimäärin P_α:s osa kun

$$P_\alpha = \int_{\alpha}^{+\infty} \eta(x)f(x)dx.$$

Todennäköisyys sille, että kanta jonka aktiivisuus on vähintään α hylätään, on

$$P_1 = \int_{\alpha}^{+\infty} f(x)\ (1-\eta(x))dx = P_g - P_\alpha.$$

Todennäköisyys sille, että kanta jonka aktiivisuus on pienempi kuin α hyväksytään, on

$$P_2 = \int_{-\infty}^{\alpha} f(x)\eta(x)dx = P - P_\alpha.$$

Jos tutkija on kiinnostunut kannoista, joiden aktiivisuus on suurempi tai yhtä suuri kuin α, muodostavat P_1 ja P_2 riskin väärän päätöksen teolle.

Daviesin (1958) mukaan ovat seuraavat päämäärät tarkoituksenmukaisimpia biologisessa selektiossa:

— Määrätään selektiointensiteetti halutun suuruiseksi ja suoritetaan valinta siten että tutkijan kannalta hyviä kantoja tulee hyväksytyksi mahdollisimman paljon kustannuksiin nähden.
— Selektiointensiteetin ollessa halutun suuruinen pyritään hyväksyttyjen kantojen keskiarvo saamaan mahdollisimman suureksi kustannuksiin nähden.
— Määrätään hyväksyttyjen hyvien kantojen ja hyväksyttävien kantojen keskimääräinen suhde halutun suuruiseksi ja pyritään maksimoimaan P_α kustannuksiin nähden.

Kun pyrkimyksenä on löytää uusia tietyn ominaisuuden omaavia kantoja, on ensimmäinen ja kolmas päämäärä useimmiten tarkoituksenmukaisin. Kun kyseessä on kannan kehittäminen, on toinen päämäärä yleisin.

Käytännössä on numeerisia arvoja hyvin vaikea saada, ellei tehdä yksinkertaistavia oletuksia. Kun numeeriset ratkaisut on saatu, voidaan verrata toisiinsa erilaisia selektiomenetelmiä parhaan löytämiseksi.

Esimerkkejä

Seuraavien esimerkkien tarkoituksena on havainnollistaa edellä esitettyjä käsitteitä. Ensin tarkastellaan selektiota, joka suoritetaan yhdessä vaiheessa. Esimerkki perustuu KEULSIN ja SIEBENIN (1955) julkaisuun ja toisaalta FINNEYN (1958) tutkimuksiin. Seuraavaksi tarkastellaan FINNEYN (1958) esimerkkiä selektiosta kahdessa vaiheessa, joka johtaa DAVIESIN (1964) sovellutukseen antibiootteja muodostavien mikro-organismien kehittämisessä.

S e l e k t i o y h d e s s ä v a i h e e s s a. Seuraavassa tarkastellaan hyvin suurta määrää kantoja, joiden aktiivisuuksien jakautuma on normaalin keskiarvona ξ ja hajontana σ. Voidaan ajatella, että kannat ovat muodostuneet aktiivisuuden ξ omaavasta kannasta eristetyistä pesäkkeistä. Aktiivisuutta mitattaessa tehdään virhe, jonka jakautuma on myös normaalinen odotusarvona 0 ja hajontana $\sqrt{\gamma}\,\sigma$. Virheen jakautuman oletetaan pysyvän jatkuvasti samanlaisena. Kannan aktiivisuus mitataan n kertaa ja aktiivisuuden estimaatiksi y valitaan tulosten keskiarvo. Estimaatin virhe on silloin jakautunut normaalisesti odotusarvona 0 ja hajontana $\sqrt{\gamma/n}\,\sigma$. Aktiivisuutta x vastaava estimaatti y on siis myös jakautunut normaalisti odotusarvona x ja hajonta $\sqrt{\gamma/n}\,\sigma$. Jos päätetään hyväksyä sellaiset kannat, joiden aktiivisuuden estimaatit ovat suurempia kuin y*, saa todennäköisyys $\eta(x)$ arvon

$$\eta(x) = \int_{y^*}^{\infty} \frac{e^{-\frac{1}{2}\frac{(y-x)^2}{\gamma/n\,\sigma^2}}}{\sqrt{2\pi}\,\sqrt{\gamma/n}\,\sigma}\,dy.$$

Koska $f(x)dx = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}\frac{(x-\xi)^2}{\sigma^2}}\,dx$, saadaan normaalisen jakautuman ominaisuuksien perusteella seuraavat kaavat:

$$P = \frac{1}{\sqrt{2\pi}} \int_{T(P)}^{+\infty} e^{-\frac{1}{2}t^2}\,dt$$

$$P_{\xi+d} = \frac{1}{2\pi}\,\frac{1}{\sqrt{1-\rho^2}} \int_{d/\sigma}^{\infty} dx \int_{T(P)}^{\infty} e^{-\frac{1}{2}\frac{x^2+y^2-2\rho xy}{1-\rho^2}}\,dy$$

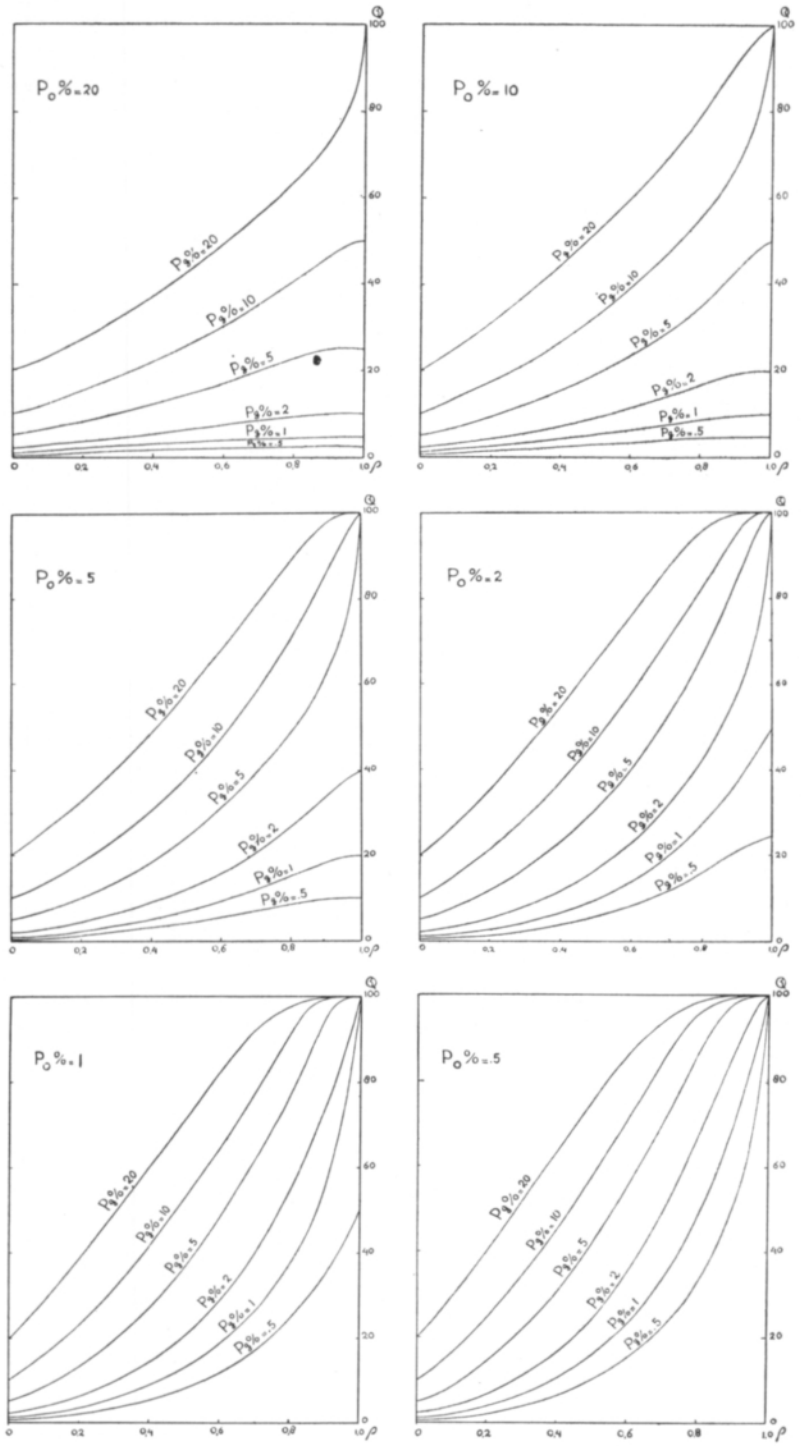
(1) $\quad M = \xi + \rho\sigma\nu(P) = \xi + G.$

Kaavoissa $T(P) = \frac{y-\xi}{\sigma}\,\frac{1}{\sqrt{1+\frac{\gamma}{n}}} = \frac{y-\xi}{\sigma}\rho$. Luku ρ on x:n ja y:n korrelaatiokerroin.

Suureella $\nu(P)$ on merkitty suhdetta

$$\nu(P) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}T(P)^2} \Big/ \frac{1}{\sqrt{2\pi}} \int_{T(P)}^{+\infty} e^{-\frac{1}{2}t^2}\,dt.$$

Normeerattu arvo T(P) määrää P:n ja siis sen, kuinka suuri osa tutkittavista keskimäärin halutaan hyväksyä. $P_{\xi+d}$ ilmoittaa todennäköisyyden sille, että kanta jonka aktiivisuus on vähintään $\xi + d$ tulee hyväksytyksi. Hyväksyttyjen joukossa on siis tällaisten



Kuva 1. Hyvien kantojen suhteellinen määrä hyväksyttyjen joukossa (Q) selektiointensiteetin (P_o) eri arvoilla, kun tutkittavassa joukossa on P_g prosenttia hyviä kantoja. ρ on havaitun ja todellisen aktiivisuuden välinen korrelaatiokerroin.

kantojen suhteellinen frekvenssi $P_{\xi+d}/P$. Suure G ilmoittaa kuinka paljon hyväksyttyjen kantojen keskiarvo keskimäärin poikkeaa kantojen jakautuman odotusarvosta.

Kuva 1 on KEULSin ja SIEBENin (1955) julkaisusta. P_g osoittaa hyvinä pidettyjen kantojen suhteellista määrää koko tutkittavassa aineistossa, kun hyvät kannat ovat sellaisia, joiden aktiivisuus ylittää tietyn arvon $\xi + d$. P_o edustaa selektiointensiteettiä ja Q suhdetta $P_{\xi+d}/P_o$. Kuvista näkyy siis hyväksyttyjen hyvien kantojen suhteellinen frekvenssi eri P_g:n ja P_o:n arvoilla ρ:n muuttuessa.

Seuraavassa valitaan selektiointensiteetiksi 5 %. Olkoon $\gamma = 3$. Hyviä kantoja on tutkittavista 10 %. Kustannukset syntyvät siten, että kannan käsittelyn hinta on 30 kertainen yhden mittauksen suorittamisesta koituviin menoihin nähden. Hinta h on siis verrannollinen lausekkeen $30 + n$ arvoon. Seuraavassa tarkastellaan suhteita $P_{\xi+d}/h$ ja G/h. Niiden maksimikohdat saadaan yleensä eri luvun n arvoilla, sillä G riippuu ρ:n ja σ:n lisäksi vain T(P):stä. Kun P_o on vakio, saavuttaa $P_{\xi+d}/h$ maksimin samalla kun Q/h. G/h ja ρ/h saavuttavat maksimin samalla n:n arvolla. Korrelaatiokertoimen ja toistojen lukumäärän välinen yhteys käy ilmi yhtälöstä $n = \frac{\rho^2}{1-\rho^2} \cdot \gamma$, Koska $\gamma = 3$, on hinta verrannollinen lausekkeeseen

$$10 + \frac{\rho^2}{1-\rho^3} = h'.$$

Voidaan siis tarkastella suhteita Q/h' ja ρ/h'. (Taulukko 1). Kaksi ensimmäistä saraketta on KEULSin ja SIEBENin julkaisusta. Muut taulukon arvot on laskettu näiden lukujen perusteella.

TAULUKKO 1.

$P_o = 0.05$
$P_g = 0.10$

ρ	Q	$\rho^2/1-\rho^2$	h'	Q/h' 10^2	ρ/h' 10^2
0.447	0.345	0.25	10.25	3.4	4.4
0.707	0.591	1.0	11.0	5.4	6.4
0.895	0.854	4.0	14.0	6.1	6.4
0.95	0.944	9.2	19.2	4.9	4.9
0.98	0.985	25.0	35.0	2.8	2.8

Tämän mukaan $P_{\xi+d}/h$:n suurin arvo saadaan välillä $0.707 < \rho < 0.95$, joten yhden estimaatin määräämiseksi tulisi tehdä 3—28 mittausta. Vastaavat arvot suhteen G/h kohdalla olisivat $0.707 < \rho < 0.895$, mikä vastaa 3—12 mittausta. Tämän perusteella saadaan hyvin hatara käsitys sopivasta toistojen määrästä. Tarkemman tutkimisen suorittamiseksi joudutaan käyttämään taulukoita, joista saadaan todennäköisyys $P_{\xi+d}$. Arvot on julkaissut esimerkiksi LEE (1927). G/h:ta maksimoitaessa on helppo saada ratkaisu differentiaalilaskulla. Tässä esimerkissä on optimi $\rho^2 = {}^2/_3$, jota vastaava $n = 6$.

Seuraavassa tapauksessa selektiointensiteetti ei ole määrätty. P_g on edelleen 0.10 ja ρ on 0.707. Väärän valinnan todennäköisyys on

$P_1 + P_2 = P_g + P_o - 2P_{\xi+d}$.

Kun P_o on vakio, pienenee $P_1 + P_2$ korrelaatiokertoimen kasvaessa. Kun ρ on vakio ja P_o sensijaan muuttuu, saadaan KEULSin ja SIEBENin mukaan väärän valinnan riskille

pienin todennäköisyys silloin kun $y^* = \xi + \frac{d}{\rho^2}$. Tällöin T(P) on $\frac{d}{\sigma\rho}$. Kun P_g on 10 %, on $\frac{d}{\sigma} = 1.2816$ ja optimi $T(P) = 1.2816/0.707 = 1.813$, joka vastaa suunnilleen kolmea prosenttia. Seuraavasta taulukosta käy ilmi tarkasteltavien suureiden muuttuminen selektiointensiteetin muuttuessa. Kaksi ensimmäistä saraketta on KEULSin ja SIEBENin julkaisusta. Taulukossa 2 esitetyt arvot on laskettu näiden lukujen perusteella.

TAULUKKO 2.

$P_g = 0.10$
$\rho = 0.707$

P_o	Q	$P_{\xi+d}$	P_1+P_2	G/σ
0.005	0.86	0.004	0.096	2.0
0.01	0.78	0.008	0.094	1.9
0.02	0.715	0.014	0.091	1.7
0.05	0.591	0.030	0.091	1.5
0.10	0.472	0.047	0.106	1.2
0.20	0.350	0.070	0.160	1.0

Virheitä voidaan pienentää siis joko lisäämällä mittauskertoja tai valitsemalla selektiointensiteetti sopivaksi. Usein on tarkoituksenmukaista määrätä hyvien kantojen suhteellinen frekvenssi valittujen kantojen joukossa halutun suuruiseksi. Sama arvo Q saadaan suurentamalla tai pienentämällä P_o:ta ja ρ:ta samanaikaisesti. Kun $\rho = 1$, on $P_{\xi+d}$ pienempi todennäköisyyksistä P_g ja P_o. Yhtälö $QP_o = P_g$ määrää siis suurimman mahdollisen teoreettisen arvon todennäköisyydelle P_o, kun Q ja P_g on määrätty. Seuraavat arvoparit P_o,ρ on laskettu käyttäen hyväksi LEEN (1927) taulukoita (taulukko 3). Edellisen esimerkin mukaisesti on laskettu kustannuksiin verrannollinen luku h' ja suhde $P_{\xi+d}/h'$.

TAULUKKO 3.

$P_g = 0.10$
$Q = 0.80$

P_o	$P_{\xi+d}$	q	$n = \rho^2/1-\rho^2 \cdot \gamma$	h'	$P_{\xi+d}/h' \cdot 10^4$
0.005	0.004	0.665	0.8 · 3	10.8	3.7
0.01	0.008	0.711	1.0 · 3	11.0	7.3
0.02	0.016	0.767	1.4 · 3	11.4	14.0
0.05	0.040	0.863	2.9 · 3	12.9	31
0.10	0.080	0.956	10.6 · 3	20.6	39
0.11	0.088	0.967	14.4 · 3	24.4	36

Paras tulos kustannuksiin nähden saavutetaan siis silloin, kun selektiointensiteetti on 5 ja 11 prosentin välillä. Tämän mukaan kannattaa hyväksyä suhteellisen monta kantaa tarkkojen mittausten perusteella.

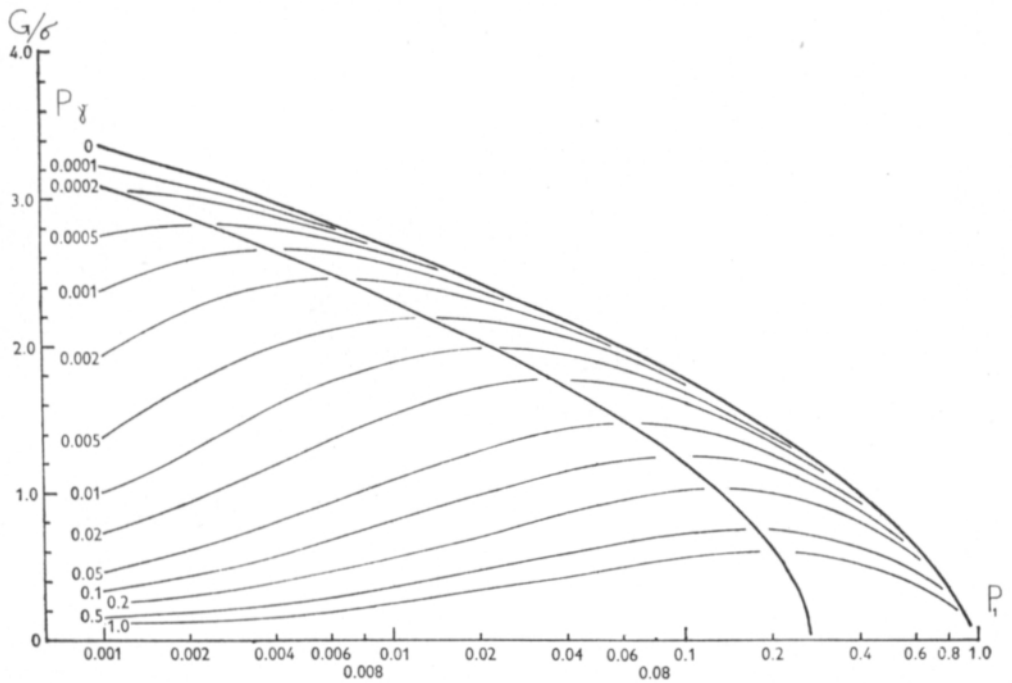
Seuraavassa tapauksessa käsitellään selektiota hieman toiselta kannalta. Usein on kaikkien tutkittavien aktiivisuus määrättävä samaan aikaan. Esimerkiksi kasvinjalostuk-

sessa on pitkän kasvukauden takia tarkoituksenmukaisinta kasvattaa kaikki kasvit samana kasvukautena. Tällöin ei voida käyttää rajattomasti maa-alaa tutkimuksiin.

Nyt oletetaan, että mittaus voidaan suorittaa korkeintaan N kertaa. Tällöin voidaan tutkia enintään N kantaa. Selektiointensiteetti P on etukäteen määrätty ja sen mukaan päätetään hyväksyä NP aktiivisuudeltaan parasta kantaa. Tässä ei aseteta rajaa y* estimaatille, vaan valitaan määrätty määrä kantoja. Jos mittauksen virhe on suuri, saattaa olla edullista satunnaisesti hylätä osa kannoista ja tutkia jäljelle jääneet P'N kantaa suorittamalla kullekin kannalle 1/P' mittausta. Testattavien hyvien kantojen absoluuttinen määrä vähenee, mutta estimaatin tarkkuus suurenee. Testattavista P'N:stä kannasta valitaan sitten $P'P_1N = PN$ kantaa. Tarkoituksena on määrätä luku P' siten, että valittujen kantojen aktiivisuuksien keskiarvon odotusarvo saadaan mahdollisimman suureksi. Tässä tapauksessa $\rho = \frac{1}{\sqrt{1 + \frac{P_\gamma}{P_1}}}$. Keskiarvon odotusarvo on kaavan (1) mukaan $M = \xi + \rho\sigma\nu(P_1)$. Finney on johtanut vastaavan kaavan myös yleisen jakautuman ollessa kyseessä. Normaalisen jakautuman tapauksessa saadaan maksimi, kun

$$\left(1 + \frac{P_1}{P_1 + P_\gamma}\right) = 2T(P_1)/\nu(P_1).$$

Seuraava kuva 2 on FINNEYn (1958) julkaisusta. Siinä näkyy usealla tulon P_γ arvolla suureen G/σ muuttuminen P_1:n muuttuessa. P_1:n on oltava luonnollisesti aina vähintään



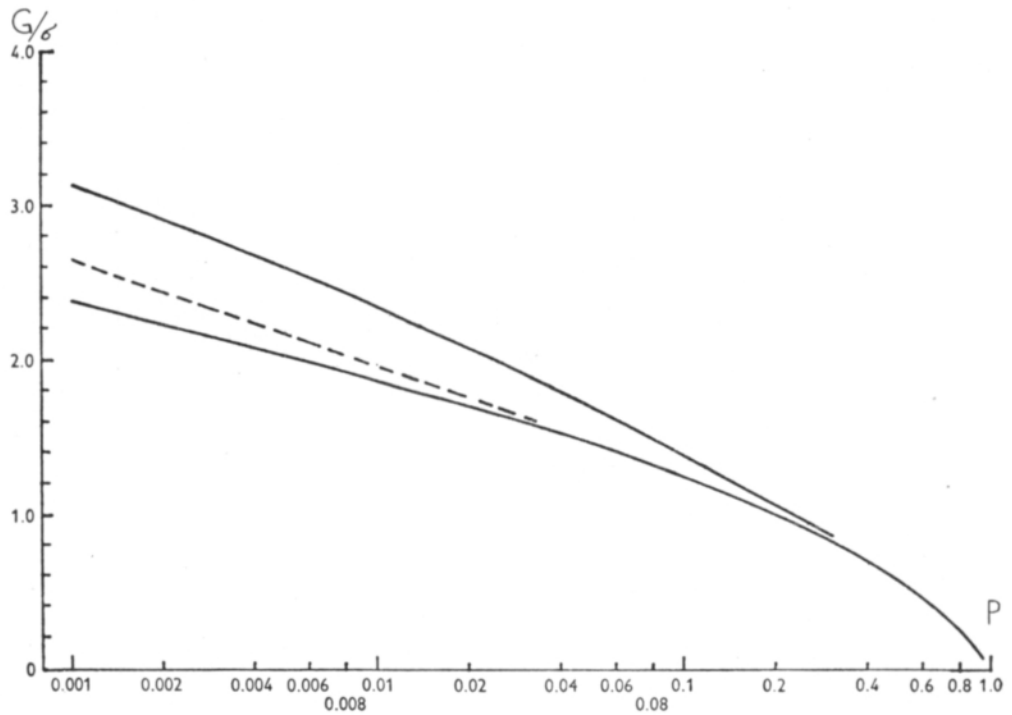
Kuva 2. Selektion antaman hyödyn (G/σ) riippuvuus satunnaisesta karsinnasta, kun selektiointensiteetti on P ja yhden määrityksen standardivirheen ja tutkittavan joukon hajonnan suhde on $\sqrt{\gamma}$. Kun ennen määrityksiä suoritetaan satunnaiskarsintaa, jonka selektiointensiteetti on P', on näin saadussa joukossa valittava selektiointensiteetiksi $P_1 = P/P'$.

P:n suuruinen. Jos P = 0.1 ja $\gamma = 1$ saataisiin suurin arvo M:lle, kun $P_1 = 0.1$ ja P' = 1. Tällöin $G/\sigma \approx 1.24$. Jos selektio intensiteetti on vain 0.01, on sopivin P_1:n arvo 0.02, mikä merkitsee testattavien karsimista puoleen. Tällöin $G/\sigma \approx 1.98$.

Selektio kahdessa vaiheessa. Yllä esitetty selektio voidaan myös suorittaa kahdessa vaiheessa. Satunnaisen karsinnan jälkeen jää P'N kantaa, joiden testaamiseen käytetään osa tutkimuspaikoista. Tämän vaiheen jälkeen hyväksytään osa lopulliseen testaukseen. Näistä $P'P_1N$:stä kannasta valitaan, käyttämällä jäljelle jääneitä testauspaikkoja estimaatin y määräämiseksi, $PN = P'P_1P_2N$ kantaa.

Tehtävänä on siis löytää sellainen testausalan jako ja luvut P',P_1 että hyväksyttyjen kantojen keskiarvon odotusarvo tulee mahdollisimman suureksi. Finney (1957) on johtanut kaavan tälle suureelle jakautuman ollessa normaalinen. Edullisimpien arvojen löytäminen on työlästä. Voidaan kuitenkin todeta, että satunnaisesta karsinnasta ei ole yleensä sanottavaa hyötyä. Jos siis valitaan P' = 1, jää jäljelle testausalan sopiva jako ja P_1:n määrääminen. Tutkijat ovat havainneet, että päästään lähelle maksimia, jos suoritetaan ns. symmetrinen selektio. Tällaisen valinnassa käytetään kummassakin vaiheessa yhtä suuret alat mittauksiin ja samaa selektiointensiteettiä.

Symmetrinen selektio sopii luontevasti tapauksiin, missä suoritetaan kahden vaiheen selektio siten, että molemmissa kokeissa käytetään kaikkia testauspaikkoja hyväksi. Jos tutkittavana on N kantaa ja yhdellä kerralla voidaan suorittaa korkeintaan N mittausta, on kahden vaiheen selektion suorittaminen selvää. Ensimmäisessä vaiheessa jokaisen kan-



Kuva 3. Selektion tuottama hyöty (G/σ), kun $\gamma = 2$ eri selektiotapojen tuloksena selektiointensiteetin (P) muuttuessa. Ylinnä symmetrinen kahden vaiheen selektio, keskellä karsinnalla optimoitu yhden vaiheen selektio ja alinna selektio yhdessä vaiheessa ilman satunnaista karsintaa.

nan aktiivisuus mitataan. Jos selektiointensiteetti on P, hyväksytään jatkoon $\sqrt{P}$ N kantaa, jotka tutkitaan suorittamalla kullekin kannalle $1/\sqrt{P}$ mittausta. Näin saatujen aktiivisuuden estimaattien perusteella valitaan $\sqrt{P} \cdot \sqrt{P}\,N = PN$ parasta kantaa. Yhteensä suoritetaan aktiivisuuden mittauksia siis 2N kertaa. Kuvasta 3, joka on FINNEYN julkaisusta (1958) käy ilmi eri menetelmillä saavutettu hyöty, kun $\gamma = 2$.

Jos kahden vaiheen selektiossa hyväksyttyjen kantojen keskiarvon odotusarvoa merkitään M_2:lla ja G:llä suuretta $M_2 - \xi$, osoittaa G selektion avulla saavutettua keskimääräisen aktiivisuuden nousua. Kuvan ylin viiva antaa G/σ:n eri selektiointensiteetin P arvoilla, kun on käytetty symmetristä selektiota. Katkoviiva osoittaa tulosta yhden vaiheen selektiosta, kun P':lle on annettu sopivin arvo. Alin viiva kuvaa tapahtumaa, kun optimointia ei ole suoritettu yhden vaiheen selektiossa. Kaikissa tapauksissa tutkittavia kantoja on N ja testauspaikkoja 2N kappaletta.

Kun yhden mittauksen virhe kasvaa, vähenee kahden vaiheen selektion etu yhdenvaiheen optimoituun selektioon nähden. Optimoinnin merkitys sensijaan kasvaa.

CURNOW (1961) on todennut, että symmetristä selektiota voidaan käyttää hyvin monen jakautuman ollessa kyseessä menettämättä kovinkaan paljoa suurimmasta mahdollisesta edusta.

Edellinen koski tapausta, missä valittujen kantojen keskiarvon odotusarvo pyrittiin maksimoimaan. DAVIES (1958) on esittänyt toisen tyyppisen esimerkin selektiosta kahdessa vaiheessa, kun hyvien hyväksyttyjen kantojen lukumäärä pyritään maksimoimaan kustannusten suhteen.

Jatkuva selektio

Edellisessä tapauksessa oli oletettava, että tarkasteltiin hyvin suurta joukkoa kantoja. Kuten aikaisemmin mainittiin, on usein vaikea saada numeerisia tuloksia ilman yksinkertaistavia oletuksia. DAVIES (1964) on tutkinut selektiota, jossa pyritään jatkuvasti kehittämään mikro-organismien aktiivisuutta. Tällöin pidettiin onnistumisen mittana valittujen kantojen aktiivisuuksien keskiarvon nousua tietyssä ajassa. Tutkimus suoritettiin simuloimalla tietokoneella.

Daviesin systeemi oli seuraava: Daviesilla oli 22 paikkainen ravistelupöytä, ja jokaisella selektiokierroksella valittiin k kantaa. Kullekin aiheutettiin mutaatioita ja muodostuneista pesäkkeistä eristettiin yhtä monta jokaisesta alkuperäisestä kannasta. Uudet kannat kasvatettiin ravistelijassa ja aktiivisuus mitattiin. Yhdessä tai useammassa vaiheessa valittiin k parasta kantaa. Tällöin oli valintaperusteena eri vaiheissa saatujen aktiivisuuden arvojen keskiarvo.

Davies tutki tapahtumaa kolmen eri jakautuman avulla. Hän nimitti jakautumiaan ylärajan-, normaalin- ja alarajan jakautumiksi sen mukaan, kuinka usein edullisia mutaatioita tapahtui.

Ensin Davies tutki pätevätkö Finneyn ja Curnowin tulokset tällaisessa prosessissa. Daviesin simulointi tuki symmetrisen selektion pätevyyttä. Jos virhe oli pieni ja P suuri, voitiin ylärajan tapauksessa pitää parempana yhden vaiheen selektiota. Päinvastaisessa tapauksessa oli kolmen vaiheen symmetrinen selektio edullisin. Tällöin $P_1 = P_2 = P_3 = \sqrt[3]{P}$. Eri selektiotapoja verrattaessa laskettiin saavutettu odotusarvon kasvu ajassa, joka kului kymmeneen kierrokseen kahden vaiheen selektiota. Tuloksiin vaikutti siis aika, jonka katsottiin kuluvan eri kasvatusvaiheisiin.

Koska kahden vaiheen symmetrinen selektio todettiin yleensä pätevimmäksi, tutki Davies ainoastaan tässä tapauksessa edullisinta arvoa k. Hän totesi, ettei normaalijakautuman tapauksessa voitu havaita suuria eroja arvojen 1—5 välillä. Kun kyseessä oli hyvin edullinen jakautuma, voitiin tyytyä vain yhden kannan valitsemiseen. Päinvastaisessa tapauksessa oli syytä valita joka kierrosta varten kymmenkunta kantaa.

Lopuksi on syytä korostaa sitä, että matemaattiset selektiomallit tarjoavat joukon muodollisia sääntöjä. Niiden avulla lienee mahdollista välttää sitä ajanhukkaa ja epäonnistumista, joka yleensä seuraa suunnittelemattomista kokeiluista. Toisaalta muodolliset säännöt eivät voi korvata tutkijan intuitiota ja kokemusta, mutta intuitio ja kokemus johtavat nopeimmin edullisimpaan tulokseen, jos ne voidaan tukea perusteltuun teoriaan.

KIRJALLISUUTTA

Cochran, W. G. 1951. Improvements by means of selection. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, p. 449—470.

Curnow, R. E. 1961. Optimal programmes for varietal selection. J. R. statist. Soc. B. 23: 282—318.

Davies, O. L. 1958. The desing of screening tests in the pharmaceutical industry. Bull. int. statist. Inst. 36: 226—241.

—»— 1964. Screening for improved mutants in antibiotic research. Biometrics. 20: 576—591.

Finney, D. J. 1957. The consequences on selection for a variate subject to errors of measurements. Rev. Inst. int. de Statist. 24: 22—29.

—»— 1958. Statistical problems of plant selection. Bull. int. statist. Inst. 36: 242—268.

Keuls, M.& Sieben, J. W. 1955. Two statistical problems in plant selection. Euphytica 4, 1: 34—44.

Lee, a. 1927. Supplementary tables for determining correlations from tetrachoric groupings. Biometrica 19: 354—405.

Wald, A. 1947. Sequential Analysis. New York & London.

SUMMARY

THE MATHEMATICAL BASIS OF MICROBIOLOGICAL SELECTION

Eeva Koskenniemi and H. G. Gyllenberg

Department of Microbiology, University of Helsinki, Finland

For the effective utilization of microorganisms, either in biotechnology or agriculture, it is necessary to increase their original activity considerably. This can be performed by selection because even pure cultures of microorganisms are genetically heterogenous. The determination of the activity in a given strain is carried out in cultivation experiments. The error in the figures obtained decreases with repetition of the process. However, repetition of the determination raises the expenses. This can be compensated by reducing the number of strains included in the repeated experiments. This may involve a loss of profitable strains. The problem thus lies in choosing a selection procedure which minimizes labour and costs. In looking for the most suitable selection procedure mathematical methods may be used. However, in most cases it is difficult to get numerical results, even when assumptions are introduced for the sake of simplification. Computer simulation provides an alternative to solving some of the problems. For other problems it may be possible to find intuitive solutions which may not be the best ones, but are close to them.