МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
МИКОЛАЇВСЬКИЙ НАЦІОНАЛЬНИЙ АГРАРНИЙ УНІВЕРСИТЕТ

Факультет технології виробництва і переробки продукції тваринництва, стандартизації та біотехнології

Використання інформаційно-статистичних Та багатовимірних методів У селекції великої рогатої худоби та свиней

(виробничо-практичні рекомендації)

Миколаїв
2022

УДК 636.082
В-43

Рекомендовано до друку рішенням вченої ради Миколаївського національного аграрного університету від «20» грудня 2022 р., протокол № 4.

Укладачі:

С. І. Луговий - д-р с.-г. наук, доцент, завідувач кафедри біотехнології та біоінженерії, Миколаївський національний аграрний університет;

С.С. Крамаренко - д-р біол. наук, професор, професор кафедри біотехнології та біоінженерії, Миколаївський національний аграрний університет;

О. С. Крамаренко - канд. с.-г. наук, доцент кафедри технології переробки, стандартизації та сертифікації продукції тваринництва, Миколаївський національний аграрний університет.

Рецензенти:

П. А. Ващенко – д-р с.-г. наук, старший науковий співробітник, професор кафедри технології виробництва продукції тваринництва, Подільський державний аграрний університет;

Є. В. Баркарь – канд. с.-г. наук, доцент, доцент кафедри біотехнології та біоінженерії, Миколаївський національний аграрний університет.

ЗМІСТ

ТЕМА 1. МАТЕМАТИКО-СТАТИСТИЧНІ МЕТОДИ АНАЛІЗУ МОЛЕКУЛЯРНО-ГЕНЕТИЧНОЇ МІНЛИВОСТІ ДЛЯ МАРКЕРІВ З ДОМІНАНТНИМ ТИПОМ УСПАДКУВАННЯ (RAPD, ISSR, AFLP)

ВСТУП

При аналізі молекулярно-генетичної мінливості для маркерів з домінантним типом успадкування (RAPD, ISSR, AFLP) існує два різних підходи. Один в більшою мірою є фенетичним і базується на розгляді ДНК профілів досліджуваних об’єктів. При цьому мінливість визначається за принципом наявності або відсутності смуги для відповідного продукту ампліфікації та такий підхід ґрунтується на обліку наявності/відсутності відповідних продуктів ампліфікації (band-based approach). Передбачається, що кожна смуга відповідає одному локусу. На основі мультилокусного аналізу групи об’єктів, надалі будується бінарна матриця, що містить нулі та одиниці (0/1-матриця) для відповідних об’єктів щодо відповідних локусів. Цей підхід є орієнтованим на окрему особину, вірніше, її мультилокусний генотип.

Принципово інший підхід, навпаки, орієнтується на популяції загалом і виходить з частот алелів по кожному локусу (frequencies alleles approach). Однак, оскільки для домінантних маркерів є специфічні особливості (насамперед, у співвідношенні фенотипів та генотипів), дана група методів має свої специфічні особливості

Оскільки дуже часто розрахунки, необхідні для отримання відповідних оцінок, дуже громіздкі, ми також розглядаємо можливість застосування спеціалізованих комп’ютерних програм, розрахованих на аналіз молекулярно-генетичних маркерів.

МЕТОДИКИ, ЩО ҐРУНТУЮТЬСЯ НА ОБЛІКУ НАЯВНОСТІ/ВІДСУТНОСТІ ВІДПОВІДНИХ ПРОДУКТІВ АМПЛІФІКАЦІЇ (BAND-BASED APPROACH)

1. МЕТОДИ ОЦІНКИ РІЗНОМАНІТНОСТІ

Основні методи математико-статистичного аналізу мінливості маркерів з домінантним типом успадкування були розроблені В. Повелом зі співавторами (Powell W., Morgante M., Andre C., Hanafey M., Vogel J., Tingey, S., Rafalski A. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis // Molecular Breeding. – 1996. – V. 2. – P. 225-238.). Під час аналізу результатів електрофореграм вони запропонували оцінювати такі показники:

Число поліморфних смуг (n_p) – і, відповідно, число мономорфних стрічок (n_{np}) для даного праймера у вибірці організмів; сума цих показників дає загальну кількість продуктів, що ампліфікуються (смуг);

Частка поліморфних смуг (β) – відношення числа поліморфних стрічок до загального числа стрічок, що ампліфікуються:

$$\beta = \frac{n_p}{n_p + n_{np}};$$

Очікувана гетерозиготність поліморфного локусу ($He_{(p)}$):

$$He_{(p)} = 1 - p^2 - q^2,$$

де p, q - частота домінантного і рецесивного алелів за відповідним локусом.

Цей показник відповідає показнику *PIC* (Polymorphism Information Content), який забезпечує оцінку дискримінаційної сили локусу, враховуючи не тільки число алелів, але також і їх відносні частоти (Ghislain M., Zhang D., Fajardo D., Huamann Z., Hijmans R.J. Marker-assisted sampling of the cultivated Andean potato *Solanum phureja* collection using RAPD markers // Genet.Res.Crop Evol. – 1999. – V. 46. – P. 547-555.)

Крім того, розраховується середня очікувана гетерозиготність по всім поліморфним локусам ($\overline{He}_{(p)}$) :

$$\overline{He}_{(p)} = \frac{\sum He_{(p)}}{n_p}$$

і, відповідно, середня очікувана гетерозиготність по всім локусам ($\overline{He}$):

$$\overline{He} = \beta \cdot \overline{He}_{(p)}.$$

Сумма ефективного числа алелей (SENA) розраховується по формулі:

$$SENA = \sum_{j=1}^{L} \left[\frac{1}{\sum_{i=1}^{2} p_i^2} - 1 \right],$$

де *j* – число аналізованих локусів; *i* - число алелей по кожному локусу. Цей показник враховує тільки поліморфні локуси, оскільки величина в дужках для мономорфних локусів дорівнює нулю.

Важливим показником, що використовується при порівнянні різних молекулярних маркерів (ISSR, RAPD, AFLP), є ефективне мультиплексне відношення (*EMR*), яке являє собою відношення числа всіх зареєстрованих смуг по всіх праймерах, що використовуються, до числа використовуваних праймерів. Цей показник, крім того, використовується для розрахунку середньої кількості поліморфних смуг на одну гель-доріжку (lane) – показника, який називається Маркерний індекс (*MI*). Нині запропоновано два підходи до оцінки Маркерного Індексу.

У роботі В.Повелла зі співавторами (Powell W., Morgante M., Andre C., Hanafey M., Vogel J., Tingey, S., Rafalski A. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis // Molecular Breeding. – 1996. – V. 2. – P. 225-238.) Маркерний Індекс розраховується як добуток ефективного мультиплексного відношення та середньої очікуваної гетерозиготності для поліморфного локусу:

$$MI = EMR \cdot \overline{He}_{(p)}.$$

Тоді як у роботі О.Превоста і М.Вілкінсона (Prevost A., Wilkinson M.J. A new system of comparing PCR primers applied to ISSR fingerprinting of potato cultivars // Theor.Appl.Genet. – 1999. – V. 98. – P. 107–112.) даний показник пропонується оцінювати за такою формулою:

$$MI = EMR \cdot \overline{Ib},$$

де $\overline{Ib}$ - Середня інформативність смуг, яка розраховується за формулою (Prevost A., Wilkinson M.J. A new system of comparing PCR primers applied to ISSR fingerprinting of potato cultivars // Theor.Appl.Genet. – 1999. – V. 98. – P. 107–112.):

$$\overline{Ib} = \frac{1}{L} \cdot \sum_{j=1}^{L} \left[1 - \left(2 \cdot \left| 0{,}5 - P_L \right| \right) \right],$$

де L - число локусів; P_L – частка особин у вибірці, для яких відзначена ампліфікація цієї стрічки.

Крім того, використовуючи той же показник інформативності смуг (*Ib*) А.Превост та М.Вілкінсон запровадили таке поняття, як роздільна сила (Resolving Power - *Rp*) праймера. Цей показник розраховується як сума оцінок інформативності смуг для всіх ампліфікованих даним праймером смуг:

$$Rp = \sum_{j=1}^{L} Ib_j = \sum_{j=1}^{L} \left[1 - \left(2 \cdot \left|0{,}5 - P_L\right|\right)\right].$$

З іншого боку, мірою ефективності праймера є показник, який називається індекс праймера (*PI*), введений М. Гіслайн зі співавторами (Ghislain M., Zhang D., Fajardo D., Huamann Z., Hijmans R.J. Marker-assisted sampling of the cultivated Andean potato *Solanum phureja* collection using RAPD markers // Genet.Res.Crop Evol. – 1999. – V. 46. – P. 547-555.). Він може бути віднесений як до RAPD-маркера і тоді він позначається як *RPI* (RAPD Primer Index), або до ISSR-маркера – *IPI* (ISSR Primer Index). Розраховується даний показник, як сума всіх оцінок *PIC* для всіх локусів, що ампліфікуються одним і тим же праймером.

2. МЕТОДИ ОЦІНКИ КОЕФІЦІЄНТІВ ПОДІБНОСТІ (COEFFICIENTS OF SIMILARITY)

При аналізі ступеня подібності двох, випадково обраних з вибірки, особин щодо наявності/відсутності в них тих чи інших смуг для продуктів ампліфікації виникає чотири різні ситуації:

		Особина 2	
		Смуга є (1)	**Смуга відсутня (0)**
Особина 1	**Смуга є (1)**	***a***	***b***
	Смуга відсутня (0)	***c***	***d***

Для аналізу ступеня близькості пари особин щодо мультилокусного генотипу домінантного маркера (RAPD, ISSR, AFLP) використовується цілий набір показників, що враховують інформацію про наявність/відсутність у них тих чи інших смуг.

Найбільш популярним є коефіцієнт Жаккара (Jaccard P. Nouvelles recherches sur la distribution florale // Bulletin de la Societe Vaudoise Des Sciences Naturelles. – 1908. – V. 44. – P. 223-270.):

$$I_J = \frac{a}{a+b+c}.$$

Даний коефіцієнт бере в розрахунок тільки смуги, присутні, по щонайменше, в однієї з аналізованих особин, і, отже, нечутливий до тих випадків, коли відсутність смуг викликано результатами гомоплазії (тобто, коли відсутність однієї і тієї ж смуги у різних особин викликано різними мутаціями).

Ще одним із поширених коефіцієнтів є коефіцієнт Л.Дайса (Dice L.R. Measures of the amount of ecologic association between species // Ecology. – 1945. – V. 26. – P. 297-302.):

$$I_D = \frac{2 \cdot a}{2 \cdot a + b + c}.$$

Даний коефіцієнт еквівалентний коефіцієнту, запропонованому Т.Сьоренсеном (Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons // Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter. – 1948. – V. 5. P. 1-34.). Ще пізніше подібний за змістом коефіцієнт був запропонований М. Неєм та В. Лі (Nei M., Li W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases // PNAS – 1979. – V. 76. – P. 5269–5273.) для аналізу нуклеотидної різноманітності та нуклетидної диференціації при використанні RFLP маркерів. І в цьому випадку, даний коефіцієнт подібності має такий вигляд:

$$I_{NL} = \frac{2 \cdot Nxy}{Nx \cdot Ny},$$

де *Nx*, *Ny* - число смуг, зазначених у особин *x* і *y*, відповідно; *Nxy* - число загальних смуг, зазначених у особин *x* і *y*. У такому вигляді коефіцієнт М.Нея та В.Лі отримав велике поширення при аналізі ступеня подібності особин як метрика для побудови дендрограм методами UPGMA та NJ.

Порівняно з коефіцієнтом Жаккара, коефіцієнт Л.Дайса (і його аналоги) надає великої ваги смугам, зазначеним одночасно в обох індивідуумів.

Коефіцієнт простого узгодження (*SMC* - simple-matching coefficient) вперше запропонований Р.Сокалом і К.Міхенером в 1958 році (Sokal R.R., Michener C.D. A statistical method for evaluating systematic relationships // University of Kansas Science Bulletin. – 1958. – V. 38. – P. 1409–1438.):

$$SMC = \frac{a + d}{a + b + c + d}.$$

Таким чином, коефіцієнт простого узгодження максимізує всю отриману від ДНК профілів домінантних маркерів інформацію за допомогою розгляду всіх локусів, котрим є продукти ампліфікації. Одночасна присутність обох смуг і одночасна відсутність обох смуг у двох аналізованих

особин має однаковий біологічний (молекулярно-генетичний) сенс, за винятком випадків гомоплазії.

Крім того, цей коефіцієнт ще цікавий і тим, що має характерні властивості Евклідової метрики, і це дозволяє використовувати його при аналізі молекулярної мінливості (процедура AMOVA; див. нижче).

В даний час розроблено цілу групу програм або пакетів прикладних програм, здатних за матрицею бінарних даних розраховувати всі зазначені вище коефіцієнти подібності і ще низку інших. З пакетів прикладних програм загального призначення можна виділити пакет SPSS (Бююль А., Цєфель П. SPSS: мистецтво обробки інформації. Аналіз статистичних даних і відновлення прихованих закономірностей: Пер. з нім . : ТОВ «ДіаСофтЮП», 2001. - 608 с.), який, крім описаних вище, може розраховувати два десятки різноманітних коефіцієнтів подібності.

Серед програм, розрахованих на аналіз молекулярно-генетичної мінливості, на першому місці слід зазначити програму FreeTree А.Павлічека та співавторів (Pavlicek A., Hrda S., Flegr J. FreeTree - Freeware program for construction of phylogenetic trees on the basis of distance data and bootstrap/jackknife analysis of the tree robustness. Application in the RAPD analysis of the genus *Frenkelia* // Folia Biologica (Praha). – 1999. – V. 45. – P. 97-99.). Крім того, що вона має зручний інтерфейс, програма FreeTree здатна за матрицею бінарних даних розраховувати шість коефіцієнтів подібності (а, крім того, і коефіцієнти генетичних дистанцій; див. нижче), а також здатна будувати дендрограми методами UPGMA та NJ, і, нарешті, перевіряти стійкість їхньої топології методами чисельного ресамплінгу (bootstrapping- та jackknifing- процедури).

Іншою, не менш популярною, є програма WINBOOT (Yap I., Nelson R. J. (1996). WINBOOT : A program for performing bootstrap analysis of binary data to determine the confidence limits of UPGMA-based dendrograms.). Вона здатна розраховувати 16 коефіцієнтів подібності і на основі отриманої матриці будувати дендрограми методом UPGMA, а також перевіряти стійкість їхньої топології bootstrap-методом.

Нарешті, програма Famd (Schlüter P.M., Harris S.A. Analysis of multilocus fingerprinting data sets containing missing data // Mol.Ecol.Notes. – 2006. – V. 6. – P. 569-572.) має можливість розраховувати вказані вище коефіцієнти, враховуючи можливі пропуски в даних.

Необхідно також відзначити, що дендрограми (а також інші процедури, що оцінюють ступінь близькості між об’єктами) використовують не стільки міри подібності, скільки дистанції між об’єктами (dissimilarity coefficients, distance), оцінки яких можна отримати, використовуючи формули:

$$D = 1 - S,$$

$$D = \sqrt{1 - S}\ .$$

3. МЕТОДИ АНАЛІЗУ СТУПЕНЯ ГЕНЕТИЧНОЇ ДИФЕРЕНЦІАЦІЇ МІЖ ПОПУЛЯЦІЯМИ

Для оцінки ступеня генетичної диференціації між популяціями на основі молекулярних маркерів з домінантним типом успадкування можуть бути використані методи, засновані на обліку наявності/відсутності тієї чи іншої смуги у різних особин як у межах популяції, так і у особин із різних популяцій. Найбільш популярними є запропонований М.Раймондом і Ф.Россе (Raymond M.L., Rousset F. An exact test for population differentiation // Evolution. – 1995. – V. 49. – P. 1280-1283.) точний тест. У цьому випадку розраховується оцінка відмінності між парою порівнюваних популяцій на підставі частот народження у особин цих популяцій «0» і «1» по всіх аналізованих локусах окремо. А потім знаходиться інтегральний показник диференціації, враховуючи адитивну властивість критерію Хі-квадрат Пірсона та його аналогів (наприклад, G-критерія або критерію Хі-квадрат, розрахованого по формулі максимальної правдоподібності). Крім того, у разі наявності малих частот може бути використаний точний тест, який використовує метод Марківських ланцюгів Монте-Карло (MCMC).

Подібна процедура реалізована в програмах TFPGA (Miller M.P. 1997. Tools for population genetic analyses (TFPGA) 1.3: A Windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by author.), PopGene (Yeh F.C., Yang R.Y., Boyle T. 1999. PopGene v.1.32. Microsoft Windows-based freeware for population genetic analysis.) і GENEPOP (Raymond M., Rousset F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism // J. Heredity. – 1995. – V. 86. – P. 248-249.).

Іншим не менш розповсюдженим методом аналізу ступеня генетичної диференціації є аналіз молекулярної мінливості (Analysis of Molecular Variation – AMOVA), запропонований у 1992 р. Л. Екскоффієром та співавторами (Excoffier L., Smouse P., Quattro J. Analysis of molecular variance inferred for metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data // Genetics. – 1992. – V. 131. – P. 479-491.). Дана процедура базується на алгоритмі дисперсійного аналізу Р.Фішера і таким чином розкладає всю наявну мінливість на дві компоненти: мінливість між популяціями та мінливість у середині популяцій. Як міра мінливості використовується квадрат Евклідової дистанції між кожною парою об’єктів (δ_{ij}^2):

$$\delta_{ij}^2 = \begin{cases} 0, & if\ i = j; \\ 1, & if\ i \neq j; \end{cases}.$$

Особливістю даного підходу є можливість аналізу ієрархічно згрупованого комплексу, коли вибірки згруповані в популяції, популяції, відповідно, в групи популяцій, групи популяцій – у регіони тощо. І тут є можливість оцінити міру впливу диференціації на різних рівнях.

Досить добре ця процедура реалізована у програмі GenAnIEx (Peakall R., Smouse P. GENAIEX 6: genetic analysis in Excel. Population genetic software for teaching and research // Molecular Ecology Notes. – 2006. – №6. – P. 288-295.), хоча і має обмежену кількість рівнів ієрархії.

МЕТОДИКИ, ЗАСНОВАНІ НА РОЗРАХУНКУ ЧАСТОТ АЛЕЛЕЙ МОЛЕКУЛЯРНО-ГЕНЕТИЧНИХ МАРКЕРІВ (FREQUENCIES ALLELES APPROACH)

1. МЕТОДИ ОЦІНКИ ЧАСТОТИ РЕЦЕСИВНОГО АЛЕЛЯ

При розробці методів оцінки частоти алелів молекулярних маркерів з домінантним типом успадкування базувалися на таких припущеннях:

- дані маркери мають Менделевський тип успадкування;
- популяція, що аналізується (чи вибірка з неї) перебуває у стані генетичної рівноваги, тобто, частоти генотипів можна оцінити із рівнянь Гарді-Вайнберга;
- рецесивна (нуль) алель за будь-яким локусом є ідентичною за станом (identical in state) у всіх особин будь-якої популяції;
- домінантна аллель за будь-яким локусом є ідентичною за станом (identical in state) у всіх особин будь-якої популяції.

Останні дві умови передбачають відсутність гомоплазії (homoplasy). Гомоплазія серед алелів має місце в тому випадку, коли негомологічні фрагменти займають одне й те ж саме положення на електрофоретичному профілі, або ж, коли різні мутації призводять до втрати одного і того ж фрагмента (Simmons M.P., Zhang L.B., Webb C.T., Müller K. A penalty of using anonymous dominant markers (AFLPs, ISSRs, and RAPDs) for phylogenetic inference // Molecular Phylogenetics and Evolution. – 2007. – V. 42. – P. 528–542.).

У такому разі серед усіх досліджуваних зразків можна виділити особин, які мають смугу (band) за відповідним локусом (що відзначається символом «1») та особин, які не мають такої смуги (що відзначається символом «0»). Особи першої групи, мають домінантний фенотип, є сумішшю гомозигот по домінантному алелю (*AA*) і гетерозигот (*Aa*) та їх число у вибірці позначимо як D. Особи другої групи, мають рецесивний фенотип, є рецесивними гомозиготами (*aa*) та їх кількість вибірці позначимо як R. Природно, якщо обсяг вибірки дорівнює N особинам, то:

$$D + R = N,$$

або, що тотожно:

$$d + r = 1,$$

де d і r – частка особин з домінантним і рецесивним фенотипом в вибірці, відповідно.

Тоді, використовуючи закон Гарді-Вайнберга, частоту рецесивної (нуль) алелі можна, можливо розрахувати за формулою:

$$q = \sqrt{r} = \sqrt{\frac{R}{N}} .$$

Однак, Дж. Холдейн (Haldane J.B.S. Almost unbiased estimates of functions of frequencies // Sankhya: Indian J. Stat. – 1956. – V. 17. – P. 201-208.) вказав на те, що ця оцінка є зміщеною, і запропонував поправку, яка має зменшити це зміщення:

$$q = \sqrt{\frac{4 \cdot R + 1}{4 \cdot N + 1}} .$$

Однак навіть ця формула не давала повністю незміщену вибіркову оцінку частоти рецесивної алелі і тому подальші дослідники намагалися по можливості зменшити негативний вплив цього зсуву, особливо для малочисельних вибірок.

Так, у 1994 р. М.Лінч та Б.Мілліган запропонували свою формулу для оцінки частоти рецесивного алеля при аналізі RAPD-маркерів (Lynch M., Milligan B.G. Analysis of population genetic structure with RAPD markers // Molecular Ecology. – 1994. – V. 3.– P. 91–99.):

$$q = \frac{\sqrt{r}}{1 - \frac{r \cdot (1 - r)}{8 \cdot N \cdot r^2}} ,$$

що відрізняється від «класичної» формулу наявністю поправки, що залежить як від особливостей розподілу частоти рецесивного алеля у вибірці, так і від обсягу вибірц. Природньо, що при зростанні обсягу вибірки (тобто при $N \to \infty$) вплив поправки знижується.

Свій варіант поправки запропонували у 1999 році П.Джорде зі співавторами (Jorde P.E., Palm S., Ryman N. Estimating genetic drift and effective population size from temporal shifts in dominant gene marker frequencies // Molecular Ecology. – 1999. – V. 8.– P. 1171-1178.):

$$q = \sqrt{r + \frac{1 - r}{4 \cdot N}} .$$

Принципово інший підхід до оцінки частоти рецесивного алеля для ознак з домінантним типом успадкування ще в 1980 р. запропонували

К.Хьюзер та Е.Мерфі (Huether C.A., Murphy E.A. Reduction of bias in estimating of frequency of recessive genes // Am.J.Hum.Genet. – 1980. – V. 32. – P. 212-222.). Цей метод ґрунтується на одній із методик чисельного ресамплінгу, а саме, на процедурі «складного ножа» (jackknifing). У цьому випадку знизити усунення оцінки рецесивного алеля можна, використовуючи формулу:

$$q = N \cdot \sqrt{\frac{R}{N}} - (N-1) \cdot \left[\frac{R}{N} \cdot \sqrt{\frac{R-1}{N-1}} + \frac{(N-R)}{N} \cdot \sqrt{\frac{R}{N-1}} \right].$$

Ще один принципово інакший метод для оцінки шуканого показника був запропонований в 1999 р. Л.А. Животовським (Zhivotovsky L.A. Estimating population structure in diploids with multilocus dominant DNA markers // Molecular Ecology. – 1999. – V. 8. P. 907–913.). Він базується на використанні методу Байєса.

У тому випадку, якщо заздалегідь відомо, що вихідна популяція не знаходиться в стані генетичної рівноваги і, відповідно, відома для неї оцінка коефіцієнта інбридингу (*Fis*), наприклад, в результаті аналізу з використанням маркерів, що мають ко-домінантний тип успадкування (мікросателітів або алозимів), оцінка частоти рецесивної алелі може бути отримана, враховуючи, що частка особин з рецесивним фенотипом тоді буде дорівнювати (Lynch M., Milligan B.G. Analysis of population genetic structure with RAPD markers // Molecular Ecology. – 1994. – V. 3.– P. 91–99.):

$$r = q^2 \cdot (1 - Fis) + q \cdot Fis$$

Звідки частота рецесивної алелі може бути виражена наступним чином (Kremer A., Caron H., Cavers S., Colpaert N., Gheysen G., Gribel R., Lemes M., Lowe A.J., Margis R., Navarro C., Salgueiro F. Monitoring genetic diversity in tropical trees with multilocus dominant markers // Heredity. – 2005. – V. 95. – P. 274- 280.):

$$q = \frac{\sqrt{Fis^2 + \left[4 \cdot (1 - Fis) \cdot r\right]} - Fis}{2 \cdot (1 - Fis)}.$$

Характерно, що для повністю самозаплідних видів, дана формула (при відомих спрощеннях) дає потрібну величину: q = 0,5. В інших випадках залежність між частотою рецесивної алелі та часткою особин з рецесивним фенотипом у популяції (і, відповідно, вибірці) стає більш лінійною при підвищенні коефіцієнта інбридингу в ній.

В даний час розроблено кілька програм для аналізу генетичних даних, які, в числі всього іншого, дають можливість оцінити частоту рецесивної алелі для молекулярно-генетичних маркерів із домінантним типом

успадкування. Всі вони, звичайно, дають оцінку частоти, виходячи з припущення про генетичну рівновагу популяції. Однак, при цьому, у програмах TFPGA (Miller M.P. 1997. Tools for population genetic analyses (TFPGA) 1.3: A Windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by author.) і AFLPsurv (Vekemans X., Beauwens T., Lemaire M., Roldan-Ruiz I. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size // Molecular Ecology – 2002. – V. 11. – P. 139-151.) є можливість отримати оцінку частоти рецесивної алелі на підставі формули М.Лінча та Б.Міллігана (Lynch M., Milligan B.G. Analysis of population genetic structure with RAPD markers // Molecular Ecology. – 1994. – V. 3.– P. 91–99.).

Крім того, програма AFLPsurv (Vekemans X., Beauwens T., Lemaire M., Roldan-Ruiz I. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size // Molecular Ecology – 2002. – V. 11. – P. 139-151.) наводить оцінку частоти рецесивного алеля з використанням методу Байєса (Zhivotovsky L.A. Estimating population structure in diploids with multilocus dominant DNA markers // Molecular Ecology. – 1999. – V. 8. P. 907–913.).

Серед інших програм, що можуть бути використані для аналізу молекулярно-генетичних маркерів з домінантним типом успадкування, слід також вказати на програми PopGene (Yeh F.C., Yang R.Y., Boyle T. 1999. PopGene v.1.32. Microsoft Windows-based freeware for population genetic analysis.) і GenAnIEx (Peakall R., Smouse P. GENAIEX 6: genetic analysis in Excel. Population genetic software for teaching and research // Molecular Ecology Notes. – 2006. – №6. – P. 288-295.).

2. МЕТОДИ ОЦІНКИ КОЕФІЦІЄНТА ІНБРИДИНГУ

При аналізі молекулярно-генетичних маркерів з домінантним типом успадкування, особини з гетерозиготним генотипом не відрізняються від особин з домінантним гомозиготним генотипом і, отже, прямий аналіз рівня інбридингу не можливий, як у разі використання маркерів з ко-домінантним типом успадкування. Однак навіть у цьому випадку можна отримати непрямі оцінки коефіцієнта інбридингу (*Fis*), використовуючи розумні припущення.

Наприклад, М.Лінч та Б.Мілліган (Lynch M., Milligan B.G. Analysis of population genetic structure with RAPD markers // Molecular Ecology. – 1994. – V. 3.– P. 91–99.) використовували для цих цілей співвідношення між частотою рецесивного алеля у двох послідовних поколіннях («материнському» та «дочірньому»). У цьому випадку оцінка коефіцієнта інбридингу може бути отримана з використанням формули:

$$Fis = \frac{r_p - r_o}{\sqrt{r_o} - r_o} \cdot \left[1 - \frac{[(r_o + 3 \cdot r_p) \cdot (1 - 3 \cdot \sqrt{r_o}) + 8 \cdot r_o \cdot r_p] \cdot Var(q_o)}{2 \cdot r_o \cdot (r_p - r_o) \cdot (1 - \sqrt{r_o})^2} \right],$$

де r_p, r_0 – спостерігається частота рецесивного фенотипу серед особин «материнського» та «дочірнього» покоління, відповідно, а варіансу частоти рецесивного алеля (у нашому випадку «дочірнього» покоління) розраховується за формулою:

$$Var(q) = \frac{1 - r}{4 \cdot N}.$$

У тій ж самій роботі (Lynch M., Milligan B.G. Analysis of population genetic structure with RAPD markers // Molecular Ecology. – 1994. – V. 3.– P. 91–99.) наведено наближену формулу для оцінки вибіркової варіанси коефіцієнта інбридингу та розглянуто методи перевірки нульової гіпотези (тобто *Fis* = 0) для нього.

Принципово інший метод було запропоновано у 2002 році у роботі К.Хользінгера та співавторів (Holsinger K.E., Lewis P.O., Dey D.K. A Bayesian approach to inferring population structure from dominant markers // Molecular Ecology. – 2002. – V. 11. – P. 1157–1164.). Даний метод базується на припущенні, що для різних локусів, що використовуються в аналізі, оцінка коефіцієнта інбридингу має бути однаковою. А крім того, що розподіл частоти певного алеля в різних популяціях може бути апроксимований бета-розподілом із середньою і варіансою, що залежать від оцінки коефіцієнта генетичної диференціації (*Fst*). Він використовує MCMC-процедуру (тобто процедуру Марківського ланцюга Монте Карло) і реалізований авторами у програмі Hickory (Holsinger K.E., Lewis P.O. (2005) Hickory: A Package for Analysis of Population Genetic Data v.1.0.4.).

Зовсім нещодавно М. Фолл і співавтори (Foll M., Beaumont M.A., Gaggiotti O. An approximate Bayesian computation approach to overcome biases that arise when using AFLP markers to study population structure // Genetics. – 2008. – V. 179. – P. 927-939.). запропонували новий метод для оцінки коефіцієнта інбридингу, заснований на підході Байєса. Вони назвали свій метод ABC-процедурою (тобто Approximate Bayesian Computation) і реалізували її у програмі ABC4F (Foll M. (2008). ABC4F: The program estimates *Fst* and *Fis* for each local population from AFLP markers.).

Необхідно зазначити, що процедура оцінювання та відповідні оцінки, отримані з використанням методу К.Хользінгера та співавторів (Holsinger K.E., Lewis P.O., Dey D.K. A Bayesian approach to inferring population structure from dominant markers // Molecular Ecology. – 2002. – V. 11. – P. 1157– 1164.) може давати іноді серйозне зміщення, на що вказують і самі автори. А, крім того, програма ABC4F, в якій реалізований метод М. Фолла і співавторів (Foll M., Beaumont M.A., Gaggiotti O. An approximate Bayesian computation

approach to overcome biases that arise when using AFLP markers to study population structure // Genetics. – 2008. – V. 179. – P. 927-939.) має набагато простіший інтерфейс і, насамперед, формат даних, що вводяться. Хоча, оскільки процедура оцінки ітераційна, при великих кількостях аналізованих одночасно об'єктів та локусів, ця програма вимагає досить значних запасів машинного часу та оперативної пам'яті.

3. МЕТОДИ ОЦІНКИ ГЕННОГО РІЗНОМАНІТТЯ В ПОПУЛЯЦІЯХ

Найбільш звичайною мірою генного різноманіття (хоча цей термін може бути використаний до молекулярних маркерів з певною обережністю, оскільки частіше всього вони містять некодуючі ділянки ДНК) є оцінка:

$$H = 2 \cdot q \cdot (1 - q) ,$$

яка дорівнює ймовірності того, що два гена, випадковим чином обрані з даної популяції, відрізняються за цим локусом. Більш того, цей захід еквівалентний очікуваній гетерозиготності для ко-домінантних маркерів, що перебувають у стані рівноваги Гарді-Вайнберга і також може розглядатися, як ймовірність того, що випадкова пара алелів міститиме один «0» і одну «1».

М.Лінч і Б.Мілліган запропонували наступну незміщену оцінку генного різноманіття для випадків використання молекулярних маркерів з домінантним типом успадкування для одного локусу (Lynch M., Milligan B.G. Analysis of population genetic structure with RAPD markers // Molecular Ecology. – 1994. – V. 3.– P. 91–99.):

$$H_i = 2 \cdot q_i \cdot (1 - q_i) + 2 \cdot Var(q_i) ,$$

де варіансу частоти рецесивного алеля розраховується за наведеною вище формулою.

Варіанса ж самої оцінки генного різноманіття по даному локусу дорівнюватиме:

$$Var(H_i) = 4 \cdot \left[1 - 2 \cdot q_i\right]^2 \cdot Var(q_i) .$$

Для всіх використаних локусів може бути розрахована середнє генне різноманіття:

$$\overline{H} = \frac{1}{L} \cdot \sum_{i=1}^{L} H_i ,$$

з відповідною варіансою:

$$Var(\overline{H}) = \frac{1}{L \cdot (L-1)} \cdot \sum_{i=1}^{L} \left(\overline{H} - H_i\right)^2 ,$$

де L - число використаних локусів.

Більшість програм, що використовують молекулярні маркери з домінантним типом успадкування, з метою оцінки генного розмаїття використовують формули М.Нея (зміщену чи незміщену оцінку). Нам відома лише одна програма, в якій реалізовано метод М.Лінча та Б.Міллігана для оцінки відповідного показника; це - програма AFLPsurv (Vekemans X., Beauwens T., Lemaire M., Roldan-Ruiz I. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size // Molecular Ecology – 2002. – V. 11. – P. 139-151.). Крім стандартної оцінки $\overline{H}$, ця програма також знаходить її компоненти, викликані відмінностями між локусами, між особинами і між популяціями.

4. МЕТОДИ ОЦІНКИ ГЕНЕТИЧНИХ ДИСТАНЦІЙ МІЖ ПОПУЛЯЦІЯМИ

Крім оцінки ступеня подібності щодо індивідуальних профілів ДНК (див. вище), можуть бути отримані оцінки подібності між окремими популяціями (або вибірками з них), розраховані на підставі частот окремих алелів.

Якщо, наприклад, є дві популяції (*A* і *B*) з відповідними частотами *i*-го алеля – x_i і y_i, то відстань між цими двома популяціями в *q*-мірному просторі (де q - число алелів по даному локусу) можна визначити як:

$$d = \sqrt{\sum_{i=1}^{q} (x_i - y_i)^2} \ .$$

Однак оцінка відстані в цьому випадку буде перебувати в інтервалі від 0 до $\sqrt{2}$.

При цьому остання ситуація буде відповідати тому випадку, коли обидві популяції фіксовані за різними алелями. Оскільки це не дуже зручно, Дж. Роджерс (Rogers J.S. Measures of genetic similarity and genetic distance // In: Studies in Genetics, v. VII. – Austin: Univ. of Texas, 1972. – P. 145-153. (Publication #7213)) модифікував дану відстань таким чином, щоб значення перебували в інтервалі від 0 до 1:

$$d_R = \sqrt{\frac{1}{2}\left[\sum_{i=1}^{q} (x_i - y_i)^2\right]} \ .$$

Л. Каваллі-Сфорца та А. Едвардс (Cavalli-Sforza L., Edwards A. Phylogenetic analysis: Models and estimation procedures // Evolution. – 1967. – V. 21. – P. 550-570.) припустили, що генетична відстань між двома популяціями може бути виміряна довжиною хорди між точками *X* та *Y* на *q*-

мірній гіперсфері. І в цьому випадку оцінка генетичної дистанції між парою популяцій може бути розрахована як:

$$d_{CE} = \frac{2}{\pi} \cdot \sqrt{2 \cdot \left[1 - \sum_{i=1}^{q} \left(x_i \cdot y_i \right) \right]} .$$

Зрештою, М.Ней (Nei M. Genetic distance between populations // Amer. Natur. – 1972. – V. 106. – P. 283-292 .; Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals // Genetics. – 1978. – V. 89. – P. 583-590.) запропонував відстань, названу ним стандартною генетичною відстанню, математичне очікування якої пропорційно еволюційному часу та враховує можливі ефекти мутацій та генетичного дрейфу:

$$d_N = -\ln I ,$$

де

$$I = \frac{J_{XY}}{\sqrt{J_X \cdot J_Y}} .$$

У свою чергу J_X, J_Y, J_{XY} – це незміщені оцінки середніх сум $\sum x_i^2$, $\sum y_i^2$, $\sum x_i . y_i$ по всім використовуваним локусів, відповідно:

$$J_X = \frac{2n_x}{2n_x - 1} \cdot \sum_{i=1}^{q} x_i^2 - 1 ,$$

$$J_Y = \frac{2n_y}{2n_y - 1} \cdot \sum_{i=1}^{q} y_i^2 - 1 ,$$

$$J_{XY} = \sum_{i=1}^{q} x_i \cdot y_i ,$$

де n_x і n_y – кількість особин в популяціях X і Y, відповідно.

Однак, всі ці підходи більшою мірою стосуються молекулярних маркерів з ко-домінантним типом успадкування і безліччю алелів по кожному локусу (наприклад, алозимів або мікросателітів), тому в 1994 М. Лінч і Б. Мілліган (Lynch M., Milligan B.G. Analysis of population genetic structure with RAPD markers // Molecular Ecology. – 1994. – V. 3.– P. 91–99.) запропонували модифіковану формулу для оцінки генетичної дистанції між двома популяціями, на підставі частот алелів молекулярних маркерів з домінантним типом успадкування. Вона відрізняється від стандартної

генетичної відстані М. Нея наявністю поправного коефіцієнта, який, загалом, знижує оцінку відстані, порівняно з випадком множинного алелізму.

Нарешті, нещодавно У. Хілл і Б. Вейр (Hill W.G., Weir B.S. Moment estimation of population diversity and genetic distance from data on recessive markers // Mol. Ecol. – 2004. – V. 13. – P. 895-908.) запропонували нову методику розрахунку величини генетичної відстані між парою популяцій на підставі використання моментів оцінок. Ця методика має ітераційний характер і розрахована, як і підхід М.Лінча і Б.Міллігана, на генетичні ознаки з домінантним типом успадкування.

Переважна більшість програм (у тому числі і зазначені вище), розрахованих на аналіз генетичних даних, дають оцінки генетичних дистанцій між парами популяцій. Найчастіше це оцінки генетичних дистанцій по М. Нею (стандартна генетична відстань), Л. Каваллі-Сфорца і А. Едвардса, або Дж. Роджерса. Поправка М. Лінча і Б. Міллігана для стандартної генетичної відстані М. Нея врахована тільки в програмі AFLPsurv (Vekemans X., Beauwens T., Lemaire M., Roldan-Ruiz I. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size // Molecular Ecology – 2002. – V. 11. – P. 139-151.).

Крім того, є спеціалізована програма Population О.Лангелла, що вільно розповсюджується через Інтернет, і яка дозволяє оцінити генетичні дистанції всім типам молекулярно-генетичних маркерів. Загалом, ця програма розраховує 16 різних оцінок генетичних дистанцій між парою популяцій.

У будь-якому випадку результатом аналізу генетичної відстані між набором популяцій є квадратна матриця, що містить попарні оцінки відстаней. Надалі, для більш детального аналізу (у тому числі і філогенетичного) дана матриця використовується для виявлення найбільш достовірної топології, з використанням різних методик (UPGMA, NJ, parsimony та ін; див. нижче).

Достовірність отриманої топології можна, можливо оцінити з використанням bootstrap-процедури. Така можливість є в програмі TFPGA (Miller M.P. 1997. Tools for population genetic analyses (TFPGA) 1.3: A Windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by author.). В результаті такої перевірки, для кожної гілки (nodes) програма наводить можливість її формування. Чим ближче ця оцінка до 100%, тим вірогідніше поява цієї групи об’єктів і, відповідно, даного кластера.

5. МЕТОДИ ОЦІНКИ ГЕНЕТИЧНОЇ ДИФЕРЕНЦІАЦІЇ ТА ПІДРОЗДІЛЬНОСТІ ПОПУЛЯЦІЙ

Популяції рідко існують у континуальній формі, найчастіше вони представлені набором напівізольованих субпопуляцій. Таким чином, навіть

при повній панміксії в межах таких субпопуляцій, популяція загалом виявляється інбредованою і міру цього інбридингу можна оцінити (у разі діалельної системи) за формулою С. Райта (Wright S. The genetical structure of populations // Ann.Eugen. – 1951. – V. 15. – P. 323-354.):

$$Fst = \frac{Var(q)}{\bar{q} \cdot (1 - \bar{q})},$$

де $\bar{q}$ - середня частота даного алеля в групі субпопуляцій, а Var(q) – варіанса частоти.

Пізніше в 1973 році М.Ней (Nei M. Analysis of gene diversity in subdivided populations // PNAS. – 1973. – V. 70. – P. 3321-3323.) дав інший зміст даному коефіцієнту і, крім того, поширив ідею С. Райта на випадок численних алелів. Згідно М.Нея мірою диференціації субпопулцій може виступати величина:

$$Gst = \frac{Dst}{Ht},$$

де *Dst* - різниця між рівнем генетичної мінливості в межах всієї популяції в цілому (*Ht*) і середньої генетичної мінливості в ряду досліджуваних субпопуляцій (*Hs*):

$$Dst = Ht - Hs .$$

Пізніше, 1984 р. Б.Вейр і Ч.Кокерхам (Weir B.S., Cockerham C.C. Estimating F-statistics for the analysis of population structure // Evolution. – 1984. – V. 38. – P. 1358-1370.) перевизначили F-статистики, запроваджені С.Райтом, у термінах кореляції між алелями. Ними було розроблено принципово новий алгоритм оцінки коефіцієнта генетичної диференціації, що ґрунтується на принципах дисперсійного аналізу. Як міру диференціації вони запровадили оцінку θ, яка близька за змістом до *Gst* М.Нея.

Формули для розрахунку θ досить громіздкі і їх можна знайти в класичній роботі (Weir B.S., Cockerham C.C. Estimating F-statistics for the analysis of population structure // Evolution. – 1984. – V. 38. – P. 1358-1370.) або ж у нещодавно опублікованому огляді з F-статистики (Weir B.S., Hill W.G. Estimating F-statistics // Annu. Rev. Genet. – 2002. – V. 36. – P. 721-750.) .

З програм, що аналізують генетичні дані з повним домінуванням, можна відзначити програму PopGene (Yeh F.C., Yang R.Y., Boyle T. 1999. PopGene v.1.32. Microsoft Windows-based freeware for population genetic analysis.), що дає можливість оцінити величини *Gst* для кожного аналізованого локусу, а крім того – розраховує середню оцінку генетичної диференціації та її 95% довірчий інтервал.

З іншого боку, програма TFPGA (Miller M.P. 1997. Tools for population genetic analyses (TFPGA) 1.3: A Windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by author.) розраховує величину θ Б.Вейра та Ч.Кокерхама (Weir B.S., Cockerham C.C. Estimating F-statistics for the analysis of population structure // Evolution. – 1984.

– V. 38. – P. 1358-1370.), і, використовуючи bootstrap-процедуру, розраховує її 95% довірчий інтервал.

У роботі М.Лінча та Б.Міллігана (Lynch M., Milligan B.G. Analysis of population genetic structure with RAPD markers // Molecular Ecology. – 1994. – V. 3.– P. 91–99.) розглядається проблема аналізу генетичної підрозділності популяцій з використанням молекулярно-генетичних маркерів з повним домінуванням та наводиться аналог *Fst* для ознак такого типу. Формули та розрахунки досить громіздкі, але шукану оцінку можна отримати за допомогою програми AFLPsurv (Vekemans X., Beauwens T., Lemaire M., Roldan-Ruiz I. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size // Molecular Ecology – 2002. – V. 11. – P. 139-151.).

Зовсім нещодавно, у 2004 році У.Хілл і Б.Вейр (Hill W.G., Weir B.S. Moment estimation of population diversity and genetic distance from data on recessive markers // Mol. Ecol. – 2004. – V. 13. – P. 895-908.) розглянули можливість оцінки генетичної диференціації (підроздільності) популяції, використовуючи молекулярно-генетичні маркери з домінантним типом успадкування. І для цього вони використовували оцінки моментів розподілу частот фенотипів, а не частот алелів. Описана процедура має ітераційний характер, але, зазвичай, сходиться вже після 5-6 ітерацій.

Розглядаючи проблему підроздільності, не можна не згадати про близько пов’язану проблему оцінки потоку генів (*Nm*; gene flow) між окремими субпопуляціями (попарно) та загалом для всієї популяції. Оскільки, на відміну генетичних маркерів з ко-домінантним типом успадкування (аллозіми, микросателіти тощо.), для молекулярно-генетичних маркерів типу RAPD, AFLP, ISSR тощо. формула, що пов’язує міру генетичної підроздільності популяції і потоку генів має специфічний вигляд (McDermott J.M., McDonald B.A. Gene flow in plant pathosystems // Annual Review of Phytopathology. – 1993. – V. 31. – P. 353-373.) :

$$Nm = \frac{1}{2} \cdot \left[\frac{1}{Fst} - 1 \right],$$

де в якості величини *Fst* можливо використовувати будь-яку з оцінок міри підроздільності (*Gst, θ, Φst*).

6. МЕТОДИ ПОБУДОВИ ФІЛОГЕНЕТИЧНИХ ДЕРЕВ

Як ми вказали вище, після того, як є матриця генетичних відстаней між парами об’єктів (популяцій, видів і т.д.), що відображають еволюційні відстані, за ними може бути збудовано дерево філогенії. Зараз існує безліч методів побудови дерев на основі матриці генетичних відстаней, найбільш поширені з яких - це методи UPGMA і NJ.

(Слід зазначити, що якщо є матриця подібності (similarity; див. вище) від неї обов'язково необхідно перейти до матриці відстаней, як зазначено вище.)

Найпростіший метод у цій групі – попарного внутрішньогрупового невзваженого середнього (UPGMA). Вважається, що його детальний алгоритм з'явився у роботі Ф.Сніта і Р.Сокела (Sneath P.H.A., Sokal R.R. Numerical taxonomy. – San Francisco: Freeman, 1973.). Дерево, побудоване за допомогою даного методу, ще називають фенограмою, оскільки його спочатку використовували для вирішення питань чисельної таксономії на основі ступеня фенотипної подібності між різними видами.

Але, як виявилося, якщо швидкість замін для будь-якого гена більш менш постійна, цей метод може бути використаний і для молекулярних даних. Прийом, на відміну від інших методів, заснованих на генетичних відстанях, метод UPGMA дозволяє отримувати хороші оцінки філогенії і в тих випадках, коли таке дерево будується за генними частотами (Ней М., Кумар С. Молекулярна еволюція та філогенетика. – К.: КВІЦ, 2004. – 418 с.).

Метод UPGMA відноситься до групі ієрархічних методів кластеризації є одним із найстаріших алгоритмів кластеризації даних. Особливістю методів ієрархічної кластеризації є те, що вони розбивають документи на кластери шляхом розбиття їх у ієрархічні групи, тобто одержувана множина кластерів має ієрархічну структуру. Принцип роботи полягає у послідовному об'єднанні груп елементів, спочатку найближчих, та дедалі більш віддалених друг від друга. Основна суть цих методів полягає у виконанні наступних кроків:

1. Обчислення значень близькості між елементами та одержання матриці близькості.
2. Визначення кожного елемента в свій окремий кластер.
3. Злиття в один кластер найбільш близьких пар елементів.
4. Оновлення матриці близькості шляхом видалення колонок і рядків для кластерів, які були злиті з іншими та подальшого перерахунку матриці.
5. Перехід на крок 3 до тих пір, поки не спрацює зупинний критерій.

У цілому алгоритм UPGMA має наступний вигляд. На першому етапі будується матриця попарних еволюційних (генетичних) відстаней між об'єктами (популяціями, видами, т.п.). Кластеризація об'єктів починається з пари, що має найменшу відстань. Така пара об'єднується в один кластер, а точка розгалуження (вузол) для неї знаходиться, як середина величини їх генетичної відстані, тобто, передбачається, що об'єкти, що розглядаються, розташовані на однаковій відстані від точки розгалуження. Таким чином, ці два об'єкти об'єднуються в один кластер, і заново обчислюється відстань між цим кластером і рештою всіх об'єктів ($d_{(12)\text{-}j}$), як середня арифметична

відстань від кожного з включених в перший кластер об’єктів до кожного з тих, що залишилися:

$$d_{(12)-j} = \frac{d_{1j} + d_{2j}}{2}.$$

Далі, для отриманої матриці відстаней знову знаходиться найменше і даний об’єкт включається до нового кластеру, для якого розраховується відповідна точка розгалуження.

І так відбувається доти, доки всі об’єкти не будуть включені в дендрограму. Більше детально алгоритм побудови філогенетичних дерев з використанням методу UPGMA наведено у книзі (Ней М., Кумар С. Молекулярна еволюція та філогенетика. – К.: КВІЦ , 2004. – 418 с.).

Як ми зазначали вище, оцінкою достовірності отриманої топології може бути bootstrap-процедура. Для молекулярно-генетичних маркерів з домінантним типом успадкування дендрограма може бути побудована на основі методу UPGMA програмою TFPGA (Miller M.P. 1997. Tools for population genetic analyses (TFPGA) 1.3: A Windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by author.). Ця програма дає можливість оцінити отриману топологію за допомогою bootstrap-процедури.

Іншим поширеним методом побудови філогенетичних дерев є методом об’єднання сусідів (NJ). Він був запропонований Н.Саітоу і М.Неєм (Saitou N., Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees // Mol. Biol. Evol. – 1987. – V. 4. – P. 406-425.) і ґрунтується на принципі мінімуму еволюції. У цьому методі не розглядаються всі можливі топології, але на кожному етапі об’єднання об’єктів використовується принцип мінімуму еволюції.

Ядро методу складає концепція сусідів – двох об’єктів на дереві без кореня, з’єднаних через один внутрішній вузол. Більш детальний опис алгоритму методу NJ наведено у книзі (Ней М., Кумар С. Молекулярна еволюція та філогенетика. – К.: КВІ Ц , 2004. – 418 с.).

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Вдовиченко Ю., Жарук П. Генетичні ресурси овець в Україні. *Вісник аграрної науки*. 2019. Т. 97, № 5. С. 38-44.
2. Войтенко С., Сидоренко О. Збереження генофонду та підвищення продуктивності худоби білоголової української породи. *Вісник аграрної науки*. 2021. Т. 99, № 2. С. 41–51.
3. Войтенко С. Л., Порхун М. Г., Сидоренко О. В., Ільницька Т. Є. Генетичні ресурси сільськогосподарських тварин України на початку третього тисячоліття. *Розведення і генетика тварин*. 2019. Вип. 58. С. 110-119.
4. Наукові та організаційні аспекти розведення, генетики, біотехнології та збереження генофонду у тваринництві / М. В. Гладій та ін. *Розведення і генетика тварин*. 2018. Вип. 56. С. 5-14.
5. Дзіцюк В. В., Типило Х. Т., Гузеватий О. Є. Цитогенетика сільськогосподарських і домашніх тварин : монографія. Київ : Аграрна наука, 2021. 127 с.
6. Кругляк О. В. Генетичні ресурси молочного скотарства України. *Економіка АПК*. 2018. № 1. С. 33-39.
7. Методологія оцінки генотипу тварин за молекулярно-генетичними маркерами у тваринництві України : монографія / К. В. Копилов, О. М. Жукорський, К. В. Копилова та ін. ; наук. ред. М. В. Гладій. Київ : Аграрна наука, 2015. 208 с.
8. Почукалін А. Є., Прийма С. В., Різун В. Забезпеченість генетичними ресурсами скотарства України. *Вісник Сумського національного аграрного університету. Серія «Тваринництво».* 2022. № 1. С. 59-64.
9. Селекційно-генетичний моніторинг у конярстві / за ред. І. В. Ткачової. Київ : Аграрна наука, 2018. 204 с.
10. Сідашова С. О., Ковтун С. І. Генетичні ресурси племінних молочних стад: селекційний потенціал кращих корів та ефективність їх відтворення. *Розведення і генетика тварин*. 2018. Вип. 55. С. 209-219.
11. Супрун І. Генетичні ресурси рисистого конярства в Україні. *Вісник Сумського національного аграрного університету. Серія «Тваринництво»*. 2020. № 3. С. 67-76.
12. Хмельничий Л. М., Павленко Ю. М. Генетичні маркери в селекції та збереженні генофонду бурої худоби Сумського регіону. *Вісник Сумського національного аграрного університету. Серія «Тваринництво»*. 2021. № 3. С. 3-6.
13. The Genetics of the Pig / Edited by M. Rothschild, A. Ruvinsky. CABI Publishing, 2011. 520 p.
14. The Genetics of Cattle / Edited by D. Garrick, A. Ruvinsky. CABI Publishing, 2014. 634 p.

Навчально-наукове видання

ВИКОРИСТАННЯ ІНФОРМАЦІЙНО-СТАТИСТИЧНИХ ТА БАГАТОВИМІРНИХ МЕТОДІВ У СЕЛЕКЦІЇ ВЕЛИКОЇ РОГАТОЇ ХУДОБИ ТА СВИНЕЙ

Виробничо-практичні рекомендації

Укладачі:

Луговий Сергій Іванович

Крамаренко Сергій Сергійович

Крамаренко Олександр Сергійович

Формат 60 × 84/16. Ум. друк. арк. 2,0.
Тираж 25 прим. Зам. №523.

Надруковано у видавничому відділі
Миколаївського національного аграрного університету
54008, м. Миколаїв, вул. Георгія Гонгадзе, 9

Свідоцтво суб'єкта видавничої справи ДК № 4490 від 20.02.2013 р.