Письма в ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ
Letters to Vavilov Journal of Genetics and Breeding
2024 • 10 • 4 • 199-203

pismavavilov.ru
DOI 10.18699/letvjgb-2024-10-23

Методы и протоколы

Анализ расщеплений количественных признаков, основанный на фундаментальных свойствах нормального распределения

Д.В. Речкин ✉

Независимый исследователь

Аннотация: Описана математическая модель расщепления, основанная на фундаментальных свойствах нормального распределения. Предложены классификация расщеплений и их соотнесение с методикой исследования, ориентированной на преимущественное использование количественных (измеряемых) признаков. Описан алгоритм последовательного разделения би- и мультимодальных выборок на отдельные группы с применением свойства симметрии нормального распределения. Представлен метод балансировки групп, повышающий точность деления исходной выборки и унифицирующий подсчет количества объектов в группах. Продемонстрирована применимость описываемого метода к сложным распределениям различного вида, обеспечивающая определение формулы расщепления для выявленных групп. Приведены сведения о доступе к исполняемому модулю и исходным текстам специально разработанного инструментального средства.

Ключевые слова: генетика количественных признаков; расщепление; математическая модель; нормальное распределение; статистика; неменделевское наследование.

Для цитирования: Речкин Д.В. Анализ расщеплений количественных признаков, основанный на фундаментальных свойствах нормального распределения. *Письма в Вавиловский журнал генетики и селекции*. 2024;10(4):199-203. DOI 10.18699/letvjgb-2024-10-25

Methods and protocols

Analysis of quantitative traits segregation based on fundamental properties of the normal distribution

D.V. Rechkin ✉

Independent researcher

Abstract: A mathematical segregation model based on the fundamental properties of the normal distribution is described. A classification of segregations and their correlation with a research methodology oriented towards the predominant use of quantitative (measured) traits is proposed. An algorithm for sequential division of bi- and multimodal samples into separate groups is described, which uses the symmetry property of the normal distribution. A method for balancing groups is proposed, which improves the accuracy of dividing the original sample and unifies the calculation of the number of objects in groups. The applicability of the described method to complex distributions of various types is demonstrated, which ensures the determination of the segregation formula for the identified groups. Information is provided on access to the executable module and source codes of a specially developed tool.

Key words: genetics of quantitative traits; splitting; mathematical model; normal distribution; statistics; non-Mendelian inheritance.

For citation: Rechkin D.V. Analysis of quantitative traits segregation based on fundamental properties of the normal distribution. *Pisma v Vavilovskii Zhurnal Genetiki i Selektsii = Letters to Vavilov Journal of Genetics and Breeding.* 2024;10(4):199-203. DOI 10.18699/letvjgb-2024-10-25 (in Russian)

dimer@mail.ru

Введение

Классическая работа Грегора Иоганна Менделя (Mendel, 1865; Мендель, 1935) посвящена анализу проявляющихся в потомстве растений закономерностей, в частности расщеплений по качественным признакам (одному или нескольким). Известно, что Мендель сознательно уклонялся от изучения количественных признаков, поскольку так и не смог установить четкие критерии однозначной классификации групп расщепления. Любое измерение, выражающееся не заключением «признак присутствует – признак отсутствует», связано как с выбором надлежащей точности измерений, так и ошибками (погрешностями) собственно измерений. В свое время анализ подобных ошибок привел К.Ф. Гаусса к исследованию свойств нормального распределения (Gauss, 1821), названного позднее в честь его исключительных заслуг в этом вопросе гауссовым распределением (также распределение Гаусса, распределение Гаусса – Лапласа) (Вентцель, 1999).

В современной генетике часто появляются работы, посвященные исследованию тех или иных особенностей наследования количественных признаков (Авдеев, 2010; Гончарова и др., 2013; Белоногова, 2014; Костылев и др., 2018, 2020). Так, все авторы отмечают сложности с выделением в массиве имеющегося материала отдельных групп, характеризующихся определенными границами измеряемых признаков, позволяющих от подсчета количества объектов перейти непосредственно к анализу расщепления как в классической генетике.

Ранее сделана попытка формализации и разработки специальной программы «Полиген А» (Мережко, 2005), однако выбор платформы реализации и неясность пользовательского интерфейса ограничили сферу ее применения всего до одного-двух экземпляров. Отсутствие документации (модели, инструкций) еще менее способствовало распространению программы.

В данной работе предпринята попытка популярно изложить математические основания (модель) анализа расщеплений количественных признаков с целью сделать соответствующие идеи и инструментарий доступными для широкой аудитории генетиков и селекционеров.

Математическая модель

Базовые представления. Проявление количественных признаков в природе в большинстве случаев имеет нормальное распределение плотности вероятности, обозначаемое как распределение величины X, зависящей от математического ожидания μ и дисперсии σ^2:

$$X \sim N(\mu, \sigma^2) \quad (1)$$

Функция распределения выглядит как

$$\Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)} dt \quad (2),$$

а функция плотности вероятности – как

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \quad (3)$$

При анализе вариации количественного признака (в простейшем случае – единственного) доля экземпляров с тем или иным значением измерения этого признака x будет пропорциональна плотности вероятности $\varphi(x)$, а гистограмма, простроенная с заданной точностью измерений значений признака и нормированная до максимума функции плотности вероятности, будет повторять вид кривой ожидаемых частот (рис. 1).

Мономодальное распределение. При исследовании вариации единственного признака, определяемого через единственный количественный показатель, закономерно ожидать, что функция плотности вероятности будет представлена ***мономодальной*** кривой (см. рис. 1).

Бимодальное распределение. Если на такой количественный показатель оказывают влияние хотя бы два признака, для которых параметры их нормальных распределений отличаются ($X_1 \sim N(\mu_1, \sigma_1^2)$; $X_2 \sim N(\mu_2, \sigma_2^2)$; $\mu_1 \neq \mu_2$), значения суммарной вероятности определяются для несовместных событий по формуле (4) , а для совместных – по формуле (5):

$$P(A \cup B) = P(A) + P(B) \quad (4)$$

$$P(A + B) = P(A) + P(B) - P(AB) \quad (5)$$

В последнем случае при анализе гистограммы значений количественного показателя очевидно, что для получения более точных результатов следует вводить поправку на совместное проявление обоих признаков. Также не нуждается в доказательстве обстоятельство, согласно которому при различных модах проявления признаков μ_A и μ_B суммарная функция плотности вероятности будет представлена ***бимодальной*** кривой (рис. 2).

Мультимодальное распределение и расщепление количественных признаков. Рассмотрим аспекты распределения значений количественных признаков для гипотетического аллеля *G*, определяющего полную высоту растения в фазе цветения.

Предположим, что для аллеля *G* характерно ***неполное доминирование***, влияющее на прибавку высоты растения; у генотипа *gg* прибавка никак не выражается, для генотипа *Gg* составляет 25 %, а для генотипа *GG* – 50 %. В таком случае при наличии достаточного количества экземпляров первого поколения гибридов от скрещивания чистых линий *GG* и *gg* распределение по показателю высоты растения будет носить ярко выраженный мономодальный характер. Этого и следует ожидать, поскольку выборка образована исключительно растениями с генотипом *Gg* (единообразная 25 %-ная прибавка). Скрещивание F_1 гибридов между собой (как и самоопыление гибридов) в следующем поколении F_2 гибридов даст генотипы *GG*, *Gg* и *gg* в соотношении 1:2:1 (мы по-прежнему полагаем, что законы Менделя работают). Таким образом, четверть растений F_2 должна иметь прибавку высоты 50 %, половина – прибавку высоты 25 %, а оставшаяся четверть – стандартную высоту без прибавки.

Полученная в результате анализа гибридов F_2 гистограмма будет иметь тримодальный вид, причем площадь пика должна вдвое превышать площади пиков μ_{GG} и μ_{gg} (рис. 3). Столкнувшись с таким характером распределения гибридов F_2 по высоте, генетик может зафиксировать в качестве результатов своего исследования как неполное доминирова-

ние для аллеля *G*, так и характерное для него расщепление в соотношении 1:2:1.

Иная картина будет наблюдаться в случае ***полного доминирования*** аллеля *G* над аллелем *g*. Получив во втором поколении F_2 гибридов то же соотношение генотипов *GG*, *Gg* и *gg* (1:2:1), исследователь не сможет по фенотипу отличить *GG* от *Gg*, так как все они будут иметь одинаковую прибавку в росте по сравнению с *gg*. Соотношение количества фенотипов будет выражаться как 3:1, и соответствующая гистограмма будет иметь бимодальный вид (рис. 4), а площадь пика μ_G будет втрое больше площади пика μ_g.

Таким образом, с помощью анализа количества и характера групп, выделяемых при оценке распределения признака, можно так же надежно, как в классической модели (качественных признаков), установить характер доминирования изучаемого аллеля и формулу расщепления.

Разделение выборки на группы. Гистограмма мономодального распределения повторяет вид функции плотности вероятности, а гистограмма распределения мультимодального – вид суммы таких функций. Остается определить для набора функций $\varphi(x)$ параметры математического ожидания (моды) μ_i и дисперсии σ_i^2.

В любом случае анализ сложного распределения следует начинать с «открытого» края (слева или справа), где отсутствует «наложение» двух соседних функций плотности вероятности. Для определенности выберем вариант «слева» (рис. 5).

Используем такой алгоритм:

1. Двигаясь по оси абсцисс (измеряемый показатель) от минимальных значений к максимальным, находим первый по порядку максимум и по соответствующему объекту исходной выборки точно определяем значение m_μ (может отличаться от значения для столбца гистограммы).
2. Выделяем все объекты со значением измеряемого показателя $m_i < m_\mu$ в отдельную выборку M.
3. Для каждого объекта этой выделенной выборки (кроме объекта с максимальным значением показателя) создаем «пару», то есть объект со значением показателя $m_i' = m_\mu + (m_\mu - m_i)$. Объединяем выборки M и M', получая новую выборку M.
4. Вычисляем выборочную дисперсию σ_M^2 для выборки M, определяем величину стандартного отклонения σ_M как квадратный корень из величины σ_M^2.
5. Из исходной выборки выделяем объекты, у которых $m_\mu - \sigma_M \le m_i \le m_\mu + \sigma_M$, в выборку M^*. Вычисляем математическое ожидание (моду) μ.
6. Полученные значения μ_{M^*} и $\sigma_{M^*}^2$ являются параметрами первого выделенного нормального распределения $X^* \sim N(\mu_{M^*}, \sigma_{M^*}^2)$.
7. Из исходной выборки удаляем объекты, вошедшие в выборку M^*, продолжаем до тех пор, пока исходная выборка не станет пустой.

Каждое последующее нормальное распределение можно анализировать, выбирая между вариантами «слева» и «справа». Результатом служит набор функций нормального распределения, сумма которых и определяет общую вариацию исходной выборки.

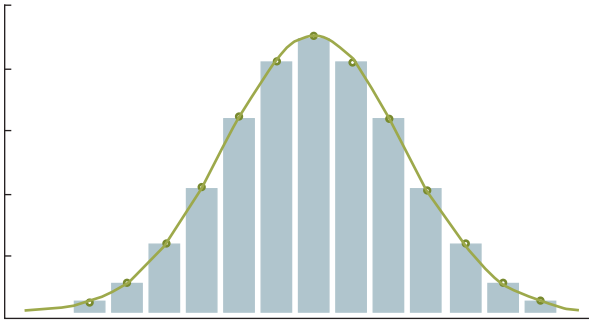

Рис. 1. Соответствие гистограммы и функции плотности нормального распределения

Fig. 1. Correspondence of the histogram and the density function of the normal distribution

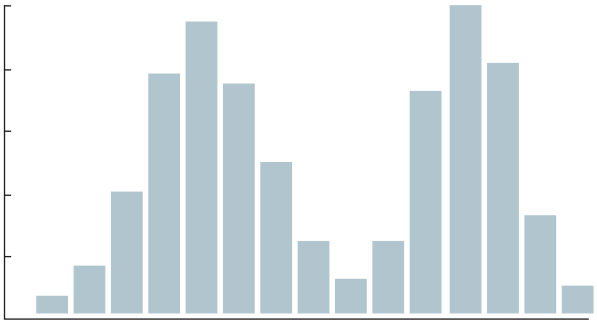

Рис. 2. Гистограмма бимодального распределения

Fig. 2. Bimodal distribution histogram sample

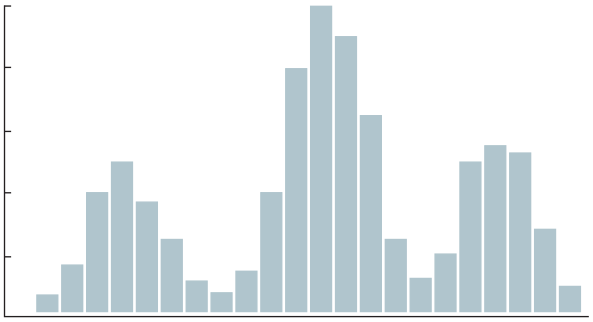

Рис. 3. Гистограмма тримодального распределения при неполном доминировании

Fig. 3. Trimodal distribution histogram sample on case of incomplete dominance

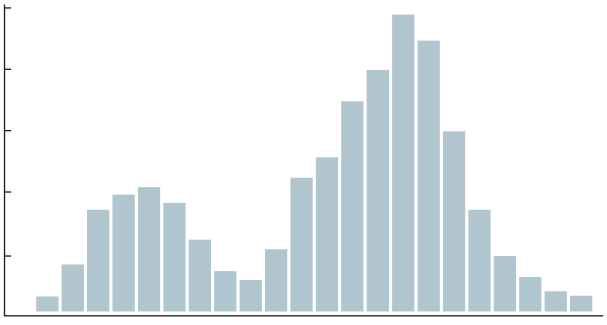

Рис. 4. Гистограмма бимодального распределения при полном доминировании

Fig. 4. Bimodal distribution histogram sample on case of complete dominance

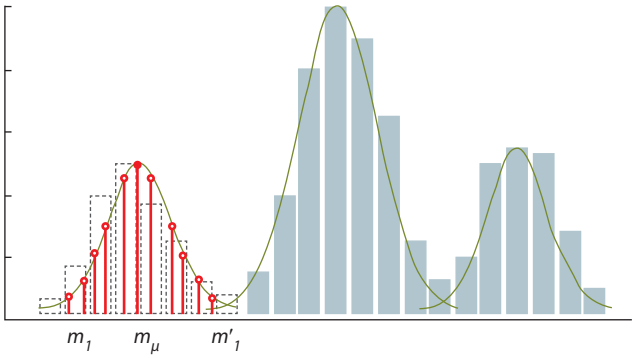

Рис. 5. Использование свойства симметрии нормального распределения
Fig. 5. Using the symmetry property of the normal distribution

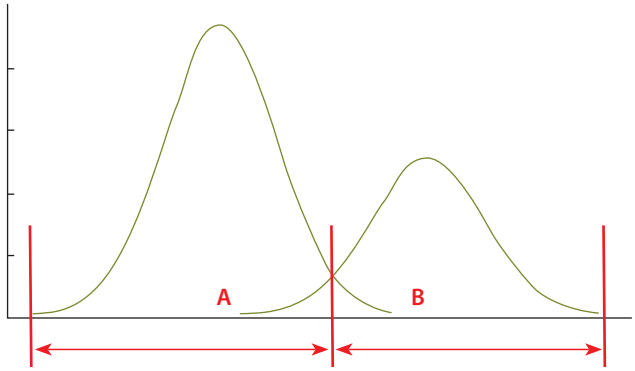

Рис. 6. Сбалансированное разделение групп
Fig. 6. Balanced division of groups

Определение границ между группами. Для всех вычисленных нормальных распределений в качестве границ (минимального и максимального значений признака для объектов, включаемых в группу) изначально применяют значения $x_{min} = \mu - 3\sigma$, $x_{max} = \mu + 3\sigma$ (правило трех сигм). Однако при таком подходе общее количество объектов, определяемое как сумма количества объектов во всех выборках, зачастую оказывается заведомо больше, чем количество объектов исходной выборки (ведь соседние распределения перекрываются, и в «зоне перекрытия» объекты считаются дважды). Поэтому следующий шаг должен приводить к сбалансированному перераспределению объектов соседних групп, чтобы избежать их вхождения более чем в одну группу.

Для разделения соседних групп находим точку пересечения соседних функций распределения, как показано на рисунке 6 (в большинстве случаев существование такой точки, притом единственной, обусловлено характером функций плотности распределения для выборок $X_1 \sim N(\mu_1, \sigma_1^2)$ и $X_2 \sim N(\mu_2, \sigma_2^2)$; $\mu_1 \neq \mu_2$ поскольку они определены на интервале $x \in (-\infty, +\infty)$. При попадании объекта точно на границу групп для определенности считаем, что он относится к «левой» группе. В исключительных случаях точка пересечения отсутствует, и тогда очевидно, что одна из выборок полностью «покрыта» другой.

Подсчет количества объектов в сбалансированных группах дает их естественное соотношение, которое легко преобразовать в искомую формулу расщепления.

Многомерные решения. В ходе исследования может выясниться, что анализируемые признаки не являются независимыми. Более того, на один и тот же количественный показатель могут влиять сразу несколько признаков. В таком случае следует собрать максимально подробные данные измерений (или иных количественных оценок проявлений изучаемых признаков) и использовать их как входные данные (исходная матрица данных) для применения метода главных компонент. В результате вычислений будут получены: (а) вектор собственных чисел, отражающий статистический вес (информативность) сформированных собственных векторов, и (б) собственные векторы, отражающие вклад каждого из исходных признаков в формирование того или иного орта (оси координат) преобразованного пространства; в этом пространстве исходные объекты могут группироваться в виде облаков рассеяния.

Для определения природы сочетаний признаков необходимо анализировать значимость главных компонент и вклады признаков в собственные векторы. Для оценки характера распределений (выделение моно-, би- и мультимодальных распределений и их использование при генетическом анализе материала) рекомендуется применять вышеописанный метод.

Заключение

Предлагаемая методика реализована автором в программе Quantic Cat[1], имеющей удобный и наглядный (графический) интерфейс в среде Microsoft Windows, а также все необходимые средства проведения расчетов.

Исследование многомерных взаимозависимостей количественных признаков удобно проводить с помощью авторского калькулятора многомерной статистики STATIC (Речкин, 1985) в современной реализации (Jacobi-Static[2]). Калькулятор использует программы на языке ЯКОБИ (Ефимов, Речкин, 1985); перенос исходных данных и результатов обработки методом главных компонент существенно облегчается тем, что обе программы имеют один и тот же формат записи данных, CSV (comma-separated values).

Подсчитываемые программой Quantic Cat границы значений признака для различных групп можно соотнести с селекционной ценностью отдельных растений, входящих в ту или иную группу. При массовом отборе для последующей селекционной работы целесообразно использовать для разведения только экземпляры, обладающие значением признака, укладывающимся в обозначенные границы. Такой подход позволит значительно сократить объем затрачиваемых ресурсов (материальных, людских), необходимых для селекции перспективных в хозяйственном отношении сортов и линий.

Таким образом, разработана математическая модель и реализованы программы, удовлетворяющие потребностям исследований анализа расщеплений по количественным признакам. Программы общедоступны, распространяются с открытым исходным кодом, что полностью снимает эконо-

[1] Доступно: http://quantic-cat.sourceforge.net (дата обращения 27 июля 2023 г.)

[2] Доступно: http://sourceforge.net/projects/jacobi-static (дата обращения 14 июня 2023 г.)

мические и юридические вопросы и создает предпосылки для их применения в исследовательской работе.

Список литературы / References

Авдеев Ю.И. Генетический анализ количественных признаков растений (монография). *Успехи современного естествознания*. 2010;(2):87-88

[Avdeev Yu.I. Genetic analysis of quantitative traits of plants (monograph). *Advances in Current Natural Sciences*. 2010;(2):87-88 (in Russian)]

Белоногова Н.М. «Прямая» и «обратная» генетика. Генетика количественных признаков. *Вавиловский журнал генетики и селекции*. 2014;18(1):147-157

[Belonogova N.M. "Direct" and "reverse" genetics. Genetics of quantitative traits. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2014;18(1):147-157 (in Russian)]

Вентцель Е.С. Теория вероятностей: Учебник для вузов, изд. 6-е, стер. М.: Высш. школа, 1999

[Wentzel E.S. Theory of Probability: Textbook for High Schools (6th ed., steryotyped). Moscow: High School Publ., 1999 (in Russian)]

Гончарова Э.А., Шумлянская (Почепня) Н.В., Щедрина З.А. Методология анализа количественных признаков в разработке технологий создания исходного материала для селекции растений. *Овощи России*. 2013;(3):30-31

[Goncharova E.A., Shumlyanskaya (Pochepnya) N.V., Shchedrina Z.A. Methodology for the analysis of quantitative traits in the development of technologies for creating initial material for plant breeding. *Vegetable Crops of Russia*. 2013;(3):30-31 (in Russian)]

Ефимов В.М., Речкин Д.В. ЯКОБИ – входной язык пакетов прикладных программ статистической обработки биологических данных. *Науч.-техн. бюл*. Новосибирск: ВАСХНИЛ, Сиб. отд-ние. 1985;(48):12-17

[Efimov V.M., Rechkin D.V. JACOBI - the input language of packages of applied programs for statistical processing of biological data. *Scientific and Technical Bulletin*. Novosibirsk: VASKhNIL, Siberian Department. 1985;(48):12-17 (in Russian)]

Костылев П.И., Краснова Е.В., Аксенов А.В. Наследование ряда количественных признаков у гибрида риса Карлик 1 × LK. *Зерновое хозяйство России*. 2018;(3):43-47. DOI 10.31367/2079-8725-2018-57-3-43-47

[Kostylev P.I., Krasnova E.V., Aksenov A.V. Inheritance of a number of quantitative traits of the rice hybrid Karlik 1 × LK. *Grain Economy of Russia*. 2018;(3):43-47. DOI 10.31367/2079-8725-2018-57-3-43-47 (in Russian)]

Костылев П.И., Краснова Е.В., Аксенов А.В., Балюкова Э.С. Анализ наследования количественных признаков у гибрида риса Кубояр × Гагат. *Аграрный вестник Урала*. 2020;(11(202)):64-75. DOI 10.32417/1997-4868-2020-202-11-64-75

[Kostylev P.I., Krasnova E.V., Aksenov A.V., Balyukova E.S. Analysis of the inheritance of quantitative traits in the rice hybrid Kuboyar × Gagat. *Agrarian Bulletin of the Urals*. 2020;(11 (202)):64-75. DOI 10.32417/1997-4868-2020-202-11-64-75 (in Russian)]

Мендель Г. Опыты над растительными гибридами. М.; Л.: ОГИЗ-Сельхозгиз, 1935;112

[Mendel G. Experiments on plant hybrids. Moscow – Leningrad: OGIZ-Selkhozgiz Publ., 1935;112 (in Russian)]

Мережко А.Ф. Использование менделевских принципов в компьютерном анализе наследования варьирующих признаков. В: Экологическая генетика культурных растений: Материалы школы молодых ученых РАСХН. Краснодар: ВНИИ риса, 2005; 107-117

[Merezhko A.F. The use of Mendelian principles in computer analysis of the inheritance of varying traits. In: Ecological genetics of cultivated plants: Materials of the school of young scientists of the Russian Academy of Agricultural Sciences. Krasnodar: All-Russian Research Institute of Rice Publ., 2005;107-117 (in Russian)]

Речкин Д.В. Реализация входного языка ЯКОБИ для мини-ЭВМ «Электроника-60». Пакет STATIC. *Науч.-техн. бюл*. Новосибирск: ВАСХНИЛ, Сиб. отд-ние. 1985;(48):18-24

[Rechkin D.V. Implementation of the JACOBI input language for the Elektronika-60 minicomputer. Package STATIC. *Scientific and Technical Bulletin*. Novosibirsk: VASKhNIL, Siberian Department. 1985;(48):18-24 (in Russian)]

Gauss C.F. Theoria combinationis observationum: erroribus minimis obnoxiae. Göttingen: Societas Regia Scientiarum Gottingensis, 1821

[Gauss C.F. Theory of the combination of observations least subject to errors. Göttingen: Societas Regia Scientiarum Gottingensis, 1821 (in Latin)]

Mendel G. Versuche über Pflanzen-Hybriden. In: Verhandlungen des naturforschenden Vereins in Brünn. IV. Band. Abhandlungen 1865. Brünn: Im Verlage des Verein, 1866;3-47

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию 28.07.2023. После доработки 11.09.2023. Принята к публикации 12.09.2024.