УДК: 578.087.1:575.113:582.47

А.Е. Демкович

ПРОГРАММА «GENRES» ДЛЯ АНАЛИЗА ДАННЫХ ПОПУЛЯЦИОННО-ГЕНЕТИЧЕСКИХ ИССЛЕДОВАНИЙ ХВОЙНЫХ

компьютерная обработка данных, прикладные программы, популяционно-генетические исследования, хвойные

В большинстве случаев для проведения статистической обработки результатов генетического анализа используются свободно доступные пакеты программ и программы. Например, отделом промышленной ботаники и популяционой генетики донецкого ботанического сада НАН Украины (ДБС НАНУ) PopGen, GenAlEX, BYOSYS-1 [8, 10, 12] – для расчета основных популяционно-генетических параметров; MLTR [11] – для исследования системы скрещивания популяций; PHYLIP [7] – для задач филогении (например, кластеризации выборок с использованием бутстрэп-метода). Все вышеупомянутые программы рассчитаны на работу с различными популяционно-генетическими маркерами: основаниями ДНК, данными по кодирующим или некодирующим участкам ДНК, аминокислотными последовательностями, изоферментами. При этом, каждой морфе, отличной от других, присваивается уникальный идентификатор, что позволяет описывать любые типы маркеров, не меняя формата данных, которыми оперирует та или иная программа. Однако, в зависимости от задач исследования, входные данные могут представлять собой как матрицы межвыборочных дистанций, так и профили аллельных частот сравниваемых выборок, и даже перечисление аллелей каждого члена отдельной выборки. В пределах одного пакета программ эти данные, как правило, совместимы, однако форматы данных для разных пакетов отличаются. Эта особенность присуща всем программам, используемым в статистической обработке результатов популяционно-генетических исследований. Она проистекает из того, что их написанием занимаются сами ученые, по мере необходимости восполняя пробелы в программном обеспечении и реализуя передовые методы статистической обработки. Однако, написанные специалистами в своей области, такие программы достаточно сложны в работе и несовместимы. Объемы файлов данных для пакетов анализа могут достигать десятков тысяч дат, что создает сложности при переходе от одной программы статобработки к другой. В нашем случае проблему унификации данных можно обойти, пользуясь пакетом анализа генетических данных, объемлющим все методы статобработки, либо создав инструмент, автоматизированно преобразующий форматы данных из одного, выбранного основным, к форматам остальных пакетов анализа. Разработка подобного инструмента стала первой нашей задачей.

При создании программного обеспечения, мы воспользовались языком Visual Basic 6.0 для приложений на платформе Microsoft 2000. Созданный нами пакет программ – GENRES был оформлен в виде надстройки Excel. Для хранения данных мы использовали стандартные файлы таблиц Excel. Благодаря этому стали возможны различные преобразования, форматирование, сортировка данных предусмотренными в Excel средствами. Естественной стала унификация разрабатываемого пакета с программой GenAlEX, также предназначенной для использования с Excel.

Чем ближе формат данных, используемых для анализа, к системе первоначальной записи, тем меньше труда необходимо для их подготовки к анализу. Оптимальным, в таком случае, является непосредственное использование формата лабораторного журнала наблюдений в качестве входного для программ дальнейшей статобработки. При разработке соответствующего программного обеспечения мы остановили свой выбор на удобном и неизбыточном формате данных, аналогичном формату в программе MLTR, и использовали его как базовый формат для ведения лабораторного журнала. Разработанной нами программный код автоматизирует задачу преобразований, позволяя вести общий журнал наблюдений, совместимый с форматом данных для программы MLTR, из которого легко сгенерировать входной файл для MLTR. Это обстоятельство позволяет отказаться от достаточно кропотливой и сложной работы непосредственно с входными файлами для MLTR.

Изоферменты являются наиболее исследованным и пока наиболее доступными генетическими маркерами. [1]. Для хвойных растений большим подспорьем в проведении изоферментного анализа является возможность при исследованиях семенного материала изучать гаплоидные ткани мегагаметофита (эндосперм семени), генотип которых идентичен половинному генотипу спорофита (материнского растения). Интерпретация получаемых при этом фореграмм гаплоидного материала упрощается, благодаря отсутствию внутрилокусных гетеромеров [6]. Приходится, однако, анализировать не менее 7 мегагаметофитов [4], и в дальнейшем определять по этой совокупности генотип спорофита, тогда как при анализе диплоидных тканей достаточно одного образца. Процедура сравнения гаплотипов макрогаметофитов и определения по их совокупности генотипа спорофита была нами реализована программно. Появилась возможность из файла лабораторного журнала генерировать файл с генотипами материнских растений. Одновременно мы реализовали обработчик ошибок ведения лабораторного журнала. Например, при обнаружении более двух аллелей для локуса одного растения, в соответствующем месте выходного файла записывается сообщение об ошибке. Необходимо отметить, что весь вышеописанный модуль рассчитан на работу с данными о диплоидных организмах, и в случае иной плоидности организмов корректно работать не сможет, поскольку при этом изменится количество аллелей, одновременно могущих быть представленными в одном локусе конкретного организма.

Особенности гаметогенеза хвойных упрощают изучение у них системы скрещивания и исследования пулов материнских и отцовских гамет, принявших участие в образовании семян. В таких исследованиях проводят параллельный электрофоретический анализ эндоспермов и зародышей семян. При этом эндосперм, являясь остатком мегагаметофита, имеет генотип равный половинному генотипу материнского спорофита и идентичный половинному генотипу зародыша семени. Генотип зародыша семени является суммой генотипов макрогаметофита-макрогаметы и генотипа микрогаметофита-микрогаметы (половинного генотипа отцовского спорофита). Для записи данных по таким исследованиям достаточно одной матрицы наблюдений, а для дальнейшей обработки генетических данных по пулам материнских, отцовских гамет требуются их повторные выборки из матрицы. Разработанный нами пакет программ, используя журнал наблюдений, совместимый с MLTR, позволяет формировать и выводить на листы Excel вышеупомянутые выборки в формате совместимом с пакетом программ GenAlEX, который имеет встроенный модуль преобразования собственного формата данных в форматы многих других программ, используемых в популяционно-генетических исследованиях, что еще более расширяет возможности анализа. Все вышеупомянутые подпрограммы собраны в первый модуль нашей программы, предназначенный для работы с файлами данных.

Оценка внутри- и межпопуляционного разнообразия и гетерогенности является отдельной задачей популяционно-генетических исследований. Маркеры, используемые в межпопуляционных сравнениях, могут быть как фенотипическими, так и чистогенетической природы. Оценки межпопуляционной гетерогенности, пригодные как в первом, так и во втором случае позволяют проводить корректные сравнения выборок с использованием обоих типов маркеров. Одним из критериев, применимых в этом случае, является r-критерий идентичности популяций Животовского [3], который вместе с сопутствующими ему статистиками, такими как показатель внутрипопуляционного разнообразия, доля редких морф в популяции, и ошибок этих показателей соответственно был реализован с изменениями и с улучшенной оболочкой [2] во втором модуле представленного пакета. В том же модуле мы реализовали расчеты дистанций и идентичностей Нея – D_N, I_N [9] – наиболее используемых мер дистанции и идентичности по генетическим данным. В качестве входных данных для этих расчетов возможно использование как частот морф (фенотипов), так их численностей, или того и другого одновременно. Важным отличием r-критерия Животовского от I_N является возможность вычисления ошибки и тестирования достоверности отличий между сравниваемыми выборками с помощью специального критерия идентичности I [3], распределенного как χ^2. Нами была реализована процедура сравнения расчетного критерия I с табличными значениями χ^2. В выводимых данных значения критерия идентичности и r-критерия сходства выборок, отличных на 95, 99, 99,9 уровнях значимости маркируются голубым, желтым и красным цветами, соответственно.

Параллельное применение различных методов статистической обработки даных позволяет контролировать ошибки, обусловленные как неправильным использованием, так и самой природой применяемых методов. В нашем случае, для проверки и уточнения результатов, получаемых с использованием I-критерия, мы воспользовались тестом гетерогенности выборок по частотам морф с помощью χ^2 критерия, который, также тестирует межвыборочную гетерогенность [4]. Программный код χ^2 теста гетерогенности для случая можества выборок был написан в виде функций, представленных в третьем модуле пакета программ. Эти три функции ориентированы на непосредственную работу с данными рабочих листов и аналогичны стандартным функциям Excel. Первая из них вычисляет значение χ^2, вторая рассчитывает степени свободы, третья, по выбору пользователя, выводит уровень значимости (в виде *, **, ***, n.s.) либо ближайшее табличное значение χ^2, используя результаты, сообщаемые первыми двумя функциями. Третья функция может быть использована для определения уровней значимости в случае любых сравнений с использованием χ^2 критерия. При попарном сравнении выборок значение критерия и количество степеней свободы идентичны обычному χ^2 критерию [5], что позволяет использовать вышеописанные функции для чрезвычайно широкого круга задач парных сравнений с использованием немодифицированного χ^2 критерия.

Вышеописанный пакет программ был написан и тестировался в системах MS Windows (версии 98, ME, XP). Для обозначения аллелей использовали цифры. Принципиальным ограничением является размер листа MS Excel (65536*256 дат), на котором размещаются как входные, так и выходные данные и связанный с этим ряд ограничений.

Для подпрограмм первого модуля:

1. Максимальное количество локусов в файле лабораторного журнала – 255. Максимальное количество образцов – 65535.

2. Максимальное количество локусов для генерирования генотипов материнских растений – 127.

Для подпрограмм второго модуля:

1. Максимальное количество выборок – 255.

2. Максимальное количество образцов – около 65400.

Для формул третьего модуля суммарный объем выборок менее 65533.

Для больших массивов данных могут существовать ограничения, связанные с объемом памяти компьютера и временем выполнения.

Разработанный нами пакет программ ориентирован на совместную работу с программами MLTR и GenAlEX и не предоставляет достаточных возможностей для всестороннего генетического анализа только своими средствами. Он, однако, способствует интеграции разрозненных программных инструментов путем унификации форматов данных. Дополнительно нами были реализованы расчеты некоторых интересующих нас статистик, среди которых формулы для критерия χ^2.

1. *Алтухов Ю.П.* Генетические процессы в популяциях. – 3-е изд. – М.: ИКЦ "Академкнига", 2003. – 431 с.
2. *Демкович А.Е.* Использование Microsoft Visual Basic для статистической обработки в популяционно-генетических исследованиях. // Актуальні проблеми фізіології, генетики та біотехнології рослин і ґрунтових мікроорганізмів: Тези доп. IX конф. мол. дослідників . – Київ 2005. – С. 55.
3. *Животовский Л.А.* Показатель сходства популяций по полиморфным признакам // Журн. общ. биологии. – 1979. – 40, № 4. – С. 587–602.
4. *Животовский Л.А.* Популяционная биометрия. М.: Наука, 1991. – 271 с.
5. *Лакин Г. Ф.* Биометрия. – М.: Высш. шк. – 1973. – 343 с.
6. *Раутиан А.С.* О природе генотипа и наследственности // Журн. общ. биологии. – 1993. – 54, № 2. – С. 131 – 148.
7. *Felsenstein J.* PHYLIP (Phylogeny Inference Package) version 3.5 c. http://evolution.gs.washington.edu/phylip.html.
8. *Francis C. Yeh* POPGENE VERSION 1.31 Microsoft Window-based Freeware for Population Genetic Analysis. Avialable via Http://www.ualberta. ca/~fyeh
9. *Nei M.* Genetic distance between populations // Amer. Naturalist. – 1972. – Vol. 106. – P. 283 – 292.
10. *Peakall, R., and P. E. Smouse.* 2005. GenAlEx V6: Genetic analysis in Excel. Population genetic software for teaching and research. Australian National University, Canberra. Available via http://www.anu.edu.au/BoZo/GenAlEx.
11. *Ritland K.* Extensions of models for the estimation of mating systems using n independent loci // Heredity. – 2002. – 88. – P. 221 – 228.
12. S*wofford D. L., Selander R. B.* BIOSYS-1: a FORTRAN program for the comprehensive analysis of electrophoretic data in population genetics and systematics // J. Hered. – 1981. – 72, № 4. – P. 281 – 283.

Донецкий ботанический сад НАН Украины Получено 23.07.2007

УДК 578.087.1:575.113:582.47

ПРОГРАММА «GENRES» ДЛЯ АНАЛИЗА ДАННЫХ ПОПУЛЯЦИОННО-ГЕНЕТИЧЕСКИХ ИССЛЕДОВАНИЙ ХВОЙНЫХ

А.Е. Демкович

Донецкий ботанический сад НАН Украины

На базе Microsoft Excel разработан пакет прикладных программ для оптимизации ведения лабораторного журнала и облегчения работы с большими массивами данных, используемых в популяционно-генетических исследованиях хвойных. Также реализованы r-критерий идентичности Животовского с сопутствующими ему статистиками и формулы для вычисления χ^2 критерия.

UDC 578.087.1:575.113:582.47

«GENRES PROGRAM» FOR CONIFEROUS POPULATION GENETIC ANALYSIS

A.Ye. Demkovich.

Donetsk Botanical Gardens, Nat. Acad. of Sci. of Ukraine

On Microsoft Excel base an application program package is worked out for laboratory record-keeping optimization and easy work with mass datum used in population-and-genetic researches of conifers. Gyvotovsky r-kriterium of identity with companion statistics and chi-square test formulas are reliased as well.