Was man in SAS Genetics vergeblich sucht: Allelfrequenzschätzungen bei Dominanz von Allelen mittels EM-Algorithmus

Jäger, Bernd (1); Schüler, Katharina (1); Biebler, Karl-Ernst (1); Rudolph, Paul Eberhard (2)
1: Universtität Greifswald; 2: FBN Leibniz-Institut für Nutztierbiologie, Dummerstorf
bjaeger@biometrie.uni-greifswald.de

Die Schätzung von Allelwahrscheinlichkeiten in Mehrallelensystemen ist nur so lange einfach, wie kein Allel dominiert, alle Genotypen beobachtbar sind. Dann lässt sich die Schätzung analog zum Zweiallelenfall ohne Dominanz durchführen, sie ist eine verallgemeinerte „Genzählmethode“.

Die Schätzung der Allelwahrscheinlichkeit p = P(A) eines beliebigen Allels A wird in einem solchen Fall als relative Häufigkeit der A-Allele bezüglich aller vorhandenen Allele aufgefasst. Eine Stichprobe vom Umfang n enthält insgesamt 2n Allele und beim Durchzählen der A-Allele der Stichprobe muss man berücksichtigen, dass jeder Homozygote AA zwei Allele und jeder Heterozygote A. nur ein A-Allel enthält.

Liegt aber Dominanz vor, wird die Berechnung aufwändig. Der Maximum-Likelihood-Ansatz führt zu einem nichtlinearen Gleichungssystem, dessen numerische Behandlung in der Regel Schwierigkeiten bereitet. Diese Schwierigkeiten zu umgehen ist das Ziel des EM-Algorithmus, eines Iterationsalgorithmus, bei dessen Konstruktion die „Genzählmethode“ Pate stand. Allerdings gehen dabei nicht die beobachteten Genotypen ein, weil sie verborgen hinter den Phänotypen liegen, sondern die erwarteten Häufigkeiten für die Genotypen unter der Bedingung des Phänotyps. Die Anzahlen nicht erkennbarer Genotypen, die einem Phänotypen unterliegen und in die Berechnung der Allelwahrscheinlichkeit eingehen, werden durch die erwarteten Anzahlen bezüglich des Vererbungsmodells ersetzt. So kommt man ausgehend von beliebigen Startwerten für die Allelwahrscheinlichkeiten zu neuen Schätzwerten, die ihrerseits wieder als Startwerte in die folgende Iteration einfließen. Man beendet die Iteration, wenn vorher festgesetzte Genauigkeitsforderungen eingehalten werden.

Der EM-Algorithmus wird als ein ausführlich kommentiertes, die PROC IML mit CALL NLPFDD nutzendes SAS-Programm mitgeteilt, das sich auf beliebige Vererbungssysteme, bisher bekannte aber auch zukünftig hinzukommende, anwenden lässt, sofern man den Erbgang kennt. Ebenso kann mit dem EM-Algorithmus auch die Berechnung von Haplotypenfrequenzen erfolgen. Der Nachweis der Konvergenz der EM-Iteration gegen die MLH-Lösung stammt von EXCOFFIER u. SLATKIN (1995). Ausführlich beschrieben und verallgemeinert findet man den Beweis bei SCHÜLER (2012). Am Beispiel des bekannten AB0-Systems, eines Dreiallelen-Modells (mit Dominanz von A über 0 und B über 0), wird die Vorgehensweise erläutert.