n treatments and n blocks will be used (HINKELMANN, 1966).

For (ii): This is a combination of partial diallel crosses of Type I and II. At the population level the methods for constructing a partial diallel cross Type II can be used. A correspondence is set up between the populations and the treatments of a PBIB design with blocks of size two. If treatments i and j occur together in a block then $P_i \times P_j$ will be present (HINKELMANN and KEMPTHORNE, 1963). At the individual level the same principles will be applied as in (i) for every $P_i \times P_j$ sampled.

Summary

A two-level diallel mating design has been defined. A model for the observations from this design has been given together with an appropriate analysis that yields information about various types of combining abilities. The problem of a genetic interpretation of combining ability variances has been discussed. Finally some modifications of the mating design have been mentioned.

Key *words:* Diallel, **mating design, inter-population crosses, combining abilities, incomplete mating design.**

Zusammenfassung

Ein zweistufiger dialleler Kreuzungsplan wird definiert. Ein Modell für die Beobachtungen nach diesem Plan wird vorgeschlagen mit der dazugehörigen Auswertung. Hieraus erhält man Information über verschiedene Arten von Kombinationseignungen. Das Problem der genetischen Interpretation der Kombinationseignungsvarianzen wird kurz diskutiert. Schließlich werden noch einige Modifikationen des Kreuzungsplanes erwähnt.

References

GRIFFING, B.: Concept of general and specific combining ability in relation to diallel cross systems. Austral. J. Biol. Sci. 9, 463—493 (1956). — HARVEY, W. R.: Least-squares analysis of data with unequal sub-class numbers. ARS-20-8, U. S. Dept. of Agriculture (1960). — HINKELMANN, K.: Unvollständige diallele Kreuzungspläne. Biom. Zeit. 8, 242—265 (1966). — HINKELMANN, K., and KEMPTHORNE, O.: Two classes of group divisible partial diallel crosses. Biometrika 50, 281—291 (1963). — HINKELMANN, K., and STERN, K.: Kreuzungspläne zur Selektionszüchtung bei Waldbäumen. Silvae Genetica 9, 121—133 (1960). — KEMPTHORNE, O.: Personal communication (1972). — LIBBY, W. J., STETTLER, R. F., and SEITZ, F. W.: Forest genetics and forest tree breeding. Ann. Rev. Genet. 3, 469—494 (1969). — MATTHEWS, J. D.: General introduction. Unasylva 18 (2—3), 1—5 (1964). — STERN, K.: Plusbäume und Samenplantagen. Frankfurt, 1960. — STERN, K.: Population genetics as a basis for selection. Unasylva 18 (2—3), 21—29 (1964). — STUBER, C. W., and COCKERHAM, C. C.: Gene effects and variances in hybrid populations. Genetics 54, 1279—1286 (1966). — WRICHT, J. W.: Species hybridization in the white pines. Forest Sci. 5, 210—222 (1959). — WRIGHT, J. W.: Genetics of forest tree improvement. Rome, 1960. — WRIGHT, J. W.: Personal communication (1973).

Genetischer Abstand zwischen Populationen

I. Zur Konzeption der genetischen Abstandsmessung

Von HANS-ROLF GREGORIUS

Lehrstuhl für Forstgenetik und Forstpflanzenzüchtung der Forstlichen Fakultät der Universität Göttingen

(Eingegangen im Januar 1974)

Das Ziel der vorliegenden Ausführungen besteht darin, aus der Diskussion einiger in der Praxis häufiger zur Anwendung gelangter genetischer Abstandsmaße Kriterien herzuleiten, die geeignet sind, eine klare Konzeption des Begriffes 'genetischer Abstand zwischen Populationen' zu formulieren. Hieraus werden sich auf natürliche Weise mehrere konkrete Vorschläge für die genetische Abstandsmessung ergeben.

Diskussion einiger gebräuchlicher genetischer Abstandsmaße

Ein Vergleich zweier Populationen auf genetischer Grundlage wird im allgemeinen an zwei verschiedenen Stufen interessieren, an der Stufe des Genotyps oder des Gens. Die unmittelbarste Art und Weise, einen solchen Vergleich anzustellen besteht wohl darin zu klären, bis zu welchem Grade die Genotyp- bzw. Genhäufigkeiten, d. h. also die genetischen Strukturen bzw. die genetischen Kompositionen beider Populationen miteinander identifiziert werden können. Andererseits leitet sich die genetische Struktur einer Population aufgrund spezieller Paarungsverhältnisse im Zusammenhang mit Selektion, Drift etc. von ihrer genetischen Komposition ab, so daß also der genetischen Komposition die elementarere Bedeutung zukommt. Es ist daher sinnvoll, den als Grad der Abweichung von der Identität aufgefaßten Abstand zwischen zwei Populationen auf die Genhäufigkeiten zu beziehen und jede Population durch ihre genetische Komposition darzustellen. Die Messung eines Abstandes zwischen Populationen (bzw. deren genetischen Kompositionen) geschieht in der Mathematik mit Hilfe einer Metrik, die definiert ist auf der Menge aller miteinander zu vergleichenden Populationen. Ein 'genetisches Abstandsmaß' sollte möglichst alle Eigenschaften einer solchen Metrik besitzen, d. h. es sollte 1) nur nichtnegative reelle Werte annehmen, 2) symmetrisch sein, d. h. Population A sollte zu Population B den gleichen Abstand wie Population B zu Population A haben, 3) den Wert 0 nur genau dann annehmen, wenn die beiden verglichenen Populationen identisch sind, 4) der Dreiecksungleichung genügen, damit die Abstände einer Population zu zwei anderen miteinander verglichen werden können. Diese Forderungen bringen offensichtlich zum Ausdruck, was man sich intuitiv unter einem 'Abstand' vorstellt.

Denkt man sich die Allel-Wahrscheinlichkeiten an den zu betrachtenden Loci in Vektorform angeordnet, so erhält man auf natürliche Weise eine Repräsentation der einzelnen Populationen als Punkte (Ortsvektoren) in einem

euklidischen (kartesischen) Raum, dessen Dimension gleich ist der Anzahl der Allele an einem Locus, summiert über alle betrachteten Loci. Sollen in einer Menge von Populationen genetische Abstände gemessen werden, die sich etwa auf m Loci beziehen, an welchen höchstens jeweils n_i ($i = 1, \ldots, m$) Allele sitzen, und bezeichnet $p_{ik}^{(l)}$ die Wahrscheinlichkeit des k-ten Allels am i-ten Locus in der Population l ($\sum_{k=1}^{n_i} p_{ik}^{(l)} = 1$), so wird die genetische Komposition dieser Population also durch einen Vektor

$$(p_{11}^{(l)}, \ldots, p_{1n}^{(l)}, p_{21}^{(l)}, \ldots, p_{2n_2}^{(l)}, \ldots, p_{m1}^{(l)}, \ldots, p_{mn_m}^{(l)}),$$

im ($\sum_{i=1}^{m} n_i$) — dimensionalen euklidischen Raum dargestellt.

Die auf diesem Raum zu erklärende Metrik wird aufgrund ihrer Bedeutung als genetisches Abstandsmaß die Anzahl der Loci, die für den Vergleich herangezogen werden, berücksichtigen müssen. Andererseits sollte keine Abhängigkeit der Metrik von den genetischen Kompositionen der für eine Abstandsmessung speziell gewählten Populationen bestehen, da sich sonst der genetische Abstand zwischen zwei Populationen ändern könnte, falls eine weitere Population zu den bereits vorhandenen Populationen hinzugenommen würde. Es bereitet keinerlei Schwierigkeiten, eine Metrik zu definieren, welche den oben angeführten Bedingungen genügt; der wohlbekannte 'euklidische Abstand' z. B. wäre bereits anwendbar.

In der Praxis ist es jedoch in nur sehr seltenen Fällen möglich, eine vollständige Kenntnis der genetischen Komposition einer Population zu erlangen, vielmehr ist man vor allem auf Schätzungen der genetischen Komposition aus endlichen zufälligen Stichproben angewiesen. Dieses Problem ist in der Literatur der letzten Jahre von mehreren Autoren unter verschiedenen Gesichtspunkten und Bezugnahmen behandelt worden. Einige der Vorschläge jener Autoren zur genetischen Abstandsmessung wurden inzwischen in größerem Umfange auf Fragestellungen aus der experimentellen Populationsgenetik angewandt. Es handelt sich vornehmlich um die Maße:

G_s von Sanghvi (1952, 1953); E von Edwards u. Cavalli-Sforza (1964, 1972) und Edwards (1971); D_k von Steinberg *et al.* (1967); B von Balakrishnan u. Sanghvi (1968); D von Nei (1972).

Mit Ausnahme von Nei, der das Stichprobenproblem nicht behandelt, gehen die hier angeführten Autoren alle von Maximum-Likelihood-Schätzungen der Gen-Wahrscheinlichkeiten aus und stellen diese auf die oben beschriebene Weise als Punkte (Ortsvektoren) im euklidischen Raum dar. Jeder solche Vektor setzt sich nun also aus Komponenten zusammen, welche die relativen Häufigkeiten der einzelnen Allele an den untersuchten Loci in einer Stichprobe von bestimmtem Umfang wiedergeben und unterliegt damit zufallsbedingten Schwankungen. Insbesondere sind die relativen Häufigkeiten der Allele eines beliebigen, nur für sich betrachteten Locus multinomialverteilt, d. h. die vermittels der einzelnen Loci erklärten Randverteilungen entsprechen einer Multinomialverteilung.

Die oben zitierten Arbeiten befassen sich ohne Ausnahme mit dem Fall ungekoppelter Loci und definieren von dieser Annahme ausgehend im ersten Schritt Abstandsmaße für einen Locus, um dann im zweiten Schritt durch einfache Addition der Abstände bzgl. der einzelnen Loci zu einem Gesamtabstand zu kommen. Diese Vorgehensweise erscheint teilweise gerechtfertigt, wenn man bedenkt, daß im Falle gekoppelter Loci ja lediglich der Übergang von den Allel-Wahrscheinlichkeiten zu den Gameten-Wahrscheinlichkeiten nötig ist, um dasselbe, ursprünglich für einen Locus definierte Abstandsmaß verwenden zu können.

Folglich verbleibt nur noch die Frage, inwiefern die Multinomialverteilung der Allele eines Locus in die Konstruktion eines genetischen Abstandsmaßes eingehen sollte. Balakrishnan-Sanghvi sowie Edwards stimmen in der Forderung überein, eine Tranformation des ursprünglichen euklidischen Raumes in sich selbst vorzunehmen derart, daß 'ein Abstand dieselbe Signifikanz hat, in welcher Richtung und welchem Teil innerhalb des neuen Raumes er auch immer gemessen wird', Edwards präzisiert diese Forderung, indem er zum Ausdruck bringt, daß eine solche Transformation die Multinomialverteilung in eine sphärisch symmetrische Verteilung überführen müsse, deren Varianz unabhängig von ihrem Erwartungswert ist. Die Autoren wählen verschiedene Transformationen, wobei allerdings Balaktrishnan-Sanghvi versäumen, für ihre Transformation die Gültigkeit der von ihnen selbst aufgestellten Forderung nachzuweisen. Weiterhin treten sie keinerlei Beweise für die durchaus nicht triviale Behauptung an, daß ihr Maß B der Dreiecksgleichung genüge. Gegen eine Verwendung von B spricht ebenfalls die Tatsache, daß dieses Maß von den für eine Abstandsmessung speziell gewählten Populationen abhängt. Damit sind wenigstens zwei der eingangs begründeten Eigenschaften, welche ein genetisches Abstandsmaß besitzen sollte, nicht gegeben. Ähnliches gilt teilweise auch für die Maße G_s und D_k, die außerdem, wie Kurcynski (1970) zeigte, identisch sind.

Andererseits erfüllt Edwards' E all die hier aufgeführten Bedingungen und sollte daher allen bis hierher erwähnten, auf zufälligen Stichproben gründenden genetischen Abstandsmessungen vorgezogen werden:

Bezeichnet $p_{ik}^{(l)}$ ($l = 1, 2$; $i = 1, \ldots, m$; $k = 1, \ldots, n_i$) die relative Häufigkeit des k-ten Allels am i-ten Locus in einer Stichprobe aus der Population l (wobei die Stichprobenumfänge für $l = 1$ und $l = 2$ gleich sein müssen), so erhält E die Darstellung

$$E^2 = \sum_{i=1}^{m} \frac{8\left(1 - \sum_{k=1}^{n_i} \sqrt{p_{ik}^{(1)} \cdot p_{ik}^{(2)}}\right)}{\left(1 + \sum_{k=1}^{n_i} \sqrt{p_{ik}^{(1)}/n_i}\right)\left(1 + \sum_{k=1}^{n_i} \sqrt{p_{ik}^{(2)}/n_i}\right)}$$

Zu diesem Ergebnis gelangen Edwards u. Cavalli-Sforza, indem sie zwei Transformationen der relativen Häufigkeiten nacheinander ausführen: 1) die wohl ursprünglich auf Fisher (1958) bzw. Bhattacharyya (1946) zurückzuführende sog. Winkel-Transformation der Allelhäufigkeiten, die gleichbedeutend mit der Zuordnung

$$p_{ik}^{(l)} \to \sqrt{p_{ik}^{(l)}}$$

ist. Hierdurch wird die Überführung der Multinomialverteilung in eine annähernd sphärisch-symmetrische Verteilung mit ebenfalls annähernd vom Erwartungswert unabhängigen Varianzen erreicht. Allerdings besteht noch eine Abhängigkeit vom Stichprobenumfang, woraus sich auch die Forderung erkärt, für alle miteinander zu vergleichenden Populationen gleiche Stichprobenumfänge zu wählen. 2) Die stereographische Projektion, welche es ermöglicht, E als den euklidischen Abstand in dem auf diese Weise zweifach transformierten euklidischen Raum zu definieren.

Wie eine einfache Überlegung zeigt, ist der maximale Wert, den E^2 annehmen kann gleich

$$\sum_{i=1}^{m} \frac{8}{(1+\sqrt{1/n_i})^2},$$

so daß die folgende Normierung von E^2 vorgenommen werden kann:

$$E'^2 := \frac{E^2}{\sum_{i=1}^{m} \frac{8}{(1+\sqrt{1/n_i})^2}}, \text{ also } 0 \leqslant E'^2 \leqslant 1.$$

Es war eingangs bereits erwähnt worden, daß Nei's Maß D nicht auf den relativen Allelhäufigkeiten in zufälligen, endlichen Stichproben aus bestimmten Populationen aufbaut und daher auch all die Schwierigkeiten ausklammern kann, welche aus gewissen Verteilungsannahmen von Stichprobenwerten erwachsen würden. Voraussetzung für die Anwendung von D ist eine vollständige Kenntnis der genetischen Kompositionen miteinander zu vergleichender Populationen und nicht nur deren Schätzungen aus Stichproben. D mißt also den 'wahren' Abstand zwischen den genetischen Kompositionen zweier Populationen. Das Maß ordnet Populationen, die keine Gene gemeinsam enthalten immer den gleichen maximalen Abstand voneinander zu, wie auch immer diese Gene ansonsten verteilt sein mögen, eine Feststellung, welche z. B. für E nicht zutrifft. Dieser positiven Eigenschaft steht allerdings als Nachteil entgegen, daß D nicht die Dreiecksungleichung erfüllt und Werte annehmen kann, die nach oben nicht beschränkt sind.

Die wesentlichen Ergebnisse dieser Diskussion lassen sich somit wie folgt zusammenfassen: Soweit die hier angeführten Maße sich auf das Stichprobenproblem beziehen, stellen sie offenbar den Versuch dar, einen Abstand zwischen verschiedenen Stichprobenverteilungen zu beschreiben, und eine Schätzung dieses Abstandes anzugeben. Damit wird also eine Fragestellung aufgegriffen, die für normalverteilte Zufallsvariablen bereits von Mahalanobis (1936) behandelt wurde. Leider jedoch ist unserer Meinung nach aus allen Arbeiten keine klare Definition des Abstandsbegriffes ersichtlich (was auch auf die Arbeit von Nei zutrifft), und es wird weiterhin darauf verzichtet nachzuweisen, in welchem Zusammenhang Unterschiede zwischen Stichprobenverteilungen mit solchen zwischen genetischen Kompositionen bzw. Strukturen stehen. Letzteres aber sollte ja gerade die Grundlage sein, auf welcher eine genetische Abstandsmessung aufbauen könnte. Aus diesem Grunde erscheint es sinnvoll, an dieser Stelle einen Vorschlag für die Konzeption des genetischen Abstandsbegriffes zu unterbreiten.

Konzeption des genetischen Abstandes zwischen Populationen

Jede Population wird durch ihre genetische Komposition bzw. Struktur repräsentiert, wobei diese beiden im Falle hypothetisch unendlich großer Populationen mit Hilfe von Wahrscheinlichkeiten und im Falle endlich großer Populationen mit Hilfe von relativen Häufigkeiten beschrieben werden. Der „genetische Abstand" zwischen jeweils zwei Populationen wird als Abstand zwischen deren genetischen Kompositionen bzw. Strukturen aufgefaßt.

Die anschließenden Ausführungen für genetische Kompositionen lassen sich sinngemäß auch auf genetische Strukturen übertragen. Wir wollen uns vorerst auf die Betrachtung nur eines Locus mit beliebig, aber endlich vielen Allelen beschränken.

Wie bereits eingangs erläutert, können die genetischen Kompositionen der einzelnen Populationen als Vektoren im euklidischen Raum, dessen Dimension gleich der Anzahl der verschiedenen Allele am betrachteten Locus ist, dargestellt werden. Die Komponenten eines solchen Vektors sind die Allel-Wahrscheinlichkeiten bzw. relativen Allelhäufigkeiten in der derart repräsentierten Population. Da sich in einem geschlossenen System die Allel-Wahrscheinlichkeiten bzw. die relativen Allelhäufigkeiten immer zu 1 summieren und nicht-negativ sind, liegen alle Vektoren in einem bestimmten Teilraum Θ_n eines euklidischen Raumes etwa der Dimension n:

$$\Theta_n := \{x \mid x = (x_1, \ldots, x_n);\ x_i \geqslant 0 \text{ für } i = 1, \ldots, n;\ \sum_{i=1}^{n} x_i = 1\},$$

Θ_n wird im allgemeinen als n-dimensionales Simplex bezeichnet. Berücksichtigt man nun all die zuvor aufgezählten Eigenschaften, welche ein genetisches Abstandsmaß besitzen sollte, so läuft dies hinaus auf die Konstruktion von auf Θ_n definierten, beschränkten Metriken d(. , .), welche je zwei Populationen, die keine Allele gemeinsam haben, den größtmöglichen Abstand zuordnen. Sind z. B. $x, y \varepsilon \Theta_n$ solche Populationen, die also im Sinne der Geometrie orthogonal aufeinander stehen — oder gleichbedeutend hiermit — deren skalares Produkt

$$(x \mid y) := \sum_{i=1}^{n} x_i y_i$$

gleich 0 ist, dann soll die Beziehung

$d(x, y) = \sup_{u, v \varepsilon \Theta_n} d(u, v)$ (= kleinste obere Schranke von d(. , .) auf Θ_n) Gültigkeit haben.

Daß solche Metriken existieren, mögen die folgenden, leicht interpretierbaren Beispiele zeigen.

Sei $x = (x_1, \ldots, x_n)$ ein Vektor des n-dimensionalen euklidischen Raumes, dann ist die euklidische Norm (Länge) von x wie folgt erklärt

$$||x||^2 := \sum_{i=1}^{n} x_i^2.$$

Mit Hilfe dieser Norm und des bereits eingeführten skalaren Produktes von Vektoren definieren wir für $x, y \varepsilon \Theta_n$:

$$d_0(x, y) := \sum_{i=1}^{n} |x_i - y_i|$$

$$d_1(x, y) := \left|\left| \frac{x}{||x||} - \frac{y}{||y||} \right|\right| = \sqrt{2\left(1 - \frac{(x|y)}{||x|| \cdot ||y||}\right)}$$

$$d_2(x, y) := \arccos \frac{(x|y)}{||x|| \cdot ||y||}$$

$d_3(x, y) := ||T(x) - T(y)|| = \sqrt{2(1 - (T(x) \mid T(y)))}$; hierbei ist $T(x_1, \ldots, x_n) = (\sqrt{x_1}, \ldots, \sqrt{x_n})$ die Winkeltransformation.

$d_4(x, y) := \arccos (T(x) \mid T(y))$.

d_1 ist der euklidische Abstand zwischen den auf die Länge 1 normierten Vektoren und d_2 der im Bodenmaß gemessene Winkel zwischen diesen. Für d_3 und d_4 gelten die gleichen Interpretationen, jedoch bezogen auf die transformierten Vektoren. Außer der Gültigkeit der Dreiecksungleichung bei d_2 und d_4 (die z. B. bei Rinow, 1961, S. 4—5 nachgewiesen ist) sind in allen fünf Fällen die oben geforderten Bedingungen trivialerweise erfüllt. Es sei noch angemerkt, daß die Metriken d_3 und d_4 bereits in einer Arbeit von Edwards-Cavalli-Sforza (1964) zur genetischen Abstandsmessung benutzt wurden, dort jedoch aus einer anderen Zielsetzung heraus resultieren. Bhattacharyya (1946) benutzt d^2_4 (Δ^2) zur Messung der Divergenz zwischen Multinomialverteilungen und leitet eine approximative Verteilung einer Schätzung D'^2 von Δ^2 her. Die größten Werte, welche d_0, d_1 und d_3 bzw. d_2 und d_4 annehmen können sind 2, $\sqrt{2}$ bzw. $\Pi/2$, so daß also zusätzlich eine Normierung der

Metriken auf Werte zwischen 0 und 1 vorgenommen werden kann. (Weitere Erläuterungen der $d_i(.,.)$ entnehme man dem Anhang).

Eine Ausdehnung der genetischen Abstandsmessung an einem Locus auf beliebig, aber endlich viele Loci kann unter vornehmlich zwei verschiedenen Gesichtspunkten vorgenommen werden:

1) Jeder Gamet wird als genetische Informationseinheit betrachtet, und der Abstand zwischen zwei Populationen soll bzgl. dieser Einheiten gemessen werden;
2) Es sollen die Beiträge der einzelnen Loci zum Gesamtabstand getrennt voneinander, jedoch mit unter Umständen verschiedenen Gewichtungen berücksichtigt werden.

Zu 1): Anstelle des Alleles an einem Locus tritt hier nun der Gamet, so daß alle vorangegangenen Bemerkungen über den genetischen Abstand an einem Locus Gültigkeit behalten. Es sollte noch darauf hingewiesen werden, daß in diesem Falle zwei Populationen bereits dann als genetisch völlig verschieden voneinander betrachtet werden müssen, wenn sie an wenigstens einem Locus keine Allele gemeinsam besitzen.

Zu 2): Die Forderung, daß die Beiträge der einzelnen Loci zum Gesamtabstand getrennt voneinander zu berücksichtigen sind, kann erfüllt werden, indem man z. B. für jeden Locus einen Abstand im obigen Sinne definiert und den Gesamtabstand als Funktion dieser einzelnen Abstände darstellt.

Sollen etwa k Loci in die Betrachtungen eingehen, wobei der i-te Locus mit n_i verschiedenen Allelen besetzt sein kann ($i = 1, \ldots, k$), dann müssen Abstände $d_i(.,.)$ auf Θ_{n_i} definiert sein (die selbstverständlich alle vom geichen Typ sein können). Wir nehmen o. B. d. A. an, daß alle $d_i(.,.)$ auf Werte zwischen 0 und 1 normiert sind. Weiterhin wird, wie eingangs bereits erläutert, jede Population als Vektor, dessen Komponenten die Allel-Wahrscheinlichkeiten an den einzelnen Loci enthalten, repräsentiert. Folglich ist ein solcher Vektor als Element des kartesischen Produktes der Θ_{n_i} ($i = 1, \ldots, k$) aufzufassen, und der Gesamtabstand $d(.,.)$ muß auf dem Teilraum

$$\Theta := \bigtimes_{i=1}^{k} \Theta_{n_i} \quad \text{(kartesisches Produkt)}$$

des $(\sum_{i=1}^{k} n_i)$-dimensionalen euklidischen Raumes erklärt werden.

Ist daher eine Population als $x \varepsilon \Theta$ gegeben, so läßt sie sich immer in der Form $x = (x_1, \ldots, x_k)$ mit $x_i \varepsilon \Theta_{n_i}$ ($i = 1, \ldots, k$) darstellen.

Eine sehr einfache und leicht zu interpretierende Abhängigkeit des Gesamtabstandes von den einzelnen Abständen ist z. B. durch die folgende lineare Funktion gegeben:

Seien $a_1, \ldots, a_k$ irgendwelche positive reelle Zahlen, deren Summe gleich 1 ist, dann wird ein Gesamtabstand $d(.,.)$ auf Θ wie folgt erklärt

$$d(x, y) := \sum_{i=1}^{k} a_i \cdot d_i(x_i, y_i).$$

Wie man leicht nachweist, hat $d(.,.)$ alle Eigenschaften einer Metrik und nimmt nur Werte zwischen 0 und 1 an. Weiterhin ordnet der Abstand $d(.,.)$ unabhängig von den Gewichtungen a_i zwei Populationen genau dann den maximalen Abstand 1 zu, wenn dies auch in bezug auf jeden einzelnen Locus zutrifft. Somit gilt auch für mehrere Loci, was zuvor für nur einen Locus als Bedingung formuliert wurde, nämlich daß zwei Vektoren genau dann maximalen Abstand voneinander haben sollen, wenn sie orthogonal aufeinander stehen. Verschiedene Gewichtungen a_i der einzelnen Loci sind z. B. dann von Bedeutung, wenn der genetische Abstand zwischen Populationen bzgl. einer gewissen Anzahl von Loci gemessen werden soll, die polygen ein Merkmal kontrollieren, und von denen einige in mehr oder weniger starkem Umfange die Ausprägung dieses Merkmals beeinflussen. Falls andererseits kein Anlaß gegeben ist, irgendwelche Bewertungen der einzelnen Loci vorzunehmen, sind die $a_i = \frac{1}{k}$ ($i = 1, \ldots, k$) zu setzen.

Diese Feststellungen bewahren ihre Gültigkeit auch in dem Falle, daß man statt der Einzelabstände $d_i(.,.)$ deren Quadrate in die vorangehende Definition einsetzt und anschließend die Wurzel aus der Summe zieht, i. e. man bildet einen Gesamtabstand

$$d'(x, y) := \sqrt{\sum_{i=1}^{k} a_i \cdot d_i(x_i, y_i)^2}.$$

Im Vergleich zu $d(.,.)$ werden hier jene Loci, die einen mittleren Abstand (etwa 0.5) voneinander besitzen, eine durch die Quadrierung bedingte, geringere Berücksichtigung im Gesamtabstand erfahren.

Schätzung des genetischen Abstandes

Eine Bestimmung des 'wahren' genetischen Abstandes zwischen zwei hypothetisch unendlich großen Populationen ist in praxi nicht möglich. Man ist hier im allgemeinen darauf angewiesen, annähernd zuverlässige Informationen über die genetischen Kompositionen der Populationen und damit über deren Abstand aus zufälligen Stichproben zu entnehmen. D. h. also, auf jeder Population ist eine Stichprobenvariable erklärt, deren Verteilung durch die genetische Komposition der entsprechenden Population gegeben ist; diese Variablen werden als voneinander unabhängig angenommen. Damit stellt sich der (wahre) genetische Abstand zwischen zwei Populationen als Funktionalparameter der gemeinsamen Verteilung der beiden zugehörigen Stichprobenvariablen dar, und eine Schätzung des genetischen Abstandes ist gleichbedeutend mit einer Schätzung dieses Funktionalparameters. Leider existieren jedoch keine Verfahren, welche die Berechnung von 'möglichst guten' Schätzfunktionen für beliebige Funktionalparameter gestatten (wie z. B. das Maximum-Likelihood-Verfahren für Verteilungsparameter). Wir wollen uns deshalb vorerst mit dem Hinweis begnügen, daß der Abstand zwischen den durch relative Häufigkeiten beschriebenen genetischen Kompositionen in zwei Stichproben eine zumindest konsistente Schätzfunktion für den genetischen Abstand der beiden Populationen, aus welchen die Stichproben stammen, ist. Die Konsistenz ergibt sich aus der der relativen Häufigkeiten in Verbindung mit der Stetigkeit der Metrik, soweit sie der euklidischen Metrik topologisch äquivalent ist.

An dieser Stelle sollte noch einmal betont werden, daß zwischen Schätzungen des Abstandes von Verteilungen (wie sie von den eingangs diskutierten Autoren behandelt wurden) und Schätzungen des 'wahren' genetischen Abstandes kein notwendiger innerer Zusammenhang besteht, da beiden verschiedene Abstandsbegriffe zu Grunde liegen.

Bei der Schätzung des Abstandes zwischen genetischen Kompositionen über mehrere Loci stößt man gleich zu Beginn auf eine definitorische Lücke, welche insbesondere die präzise Beschreibung der Stichprobenvariablen unmöglich erscheinen läßt. Der Begriff der genetischen Komposition einer Population ist nur für die Allele *eines* Locus erklärt und muß daher eine möglichst sinnvoll anwendbare Erweiterung auf beliebig viele Loci erfahren. Nun kann an-

dererseits jedes Allel, welches ein Individuum an einem bestimmten Locus trägt als Bestandteil desjenigen haploiden Chromosomensatzes (Genoms) aufgefaßt werden, der von einem Elter als Gamet an dieses Individuum weitergegeben wurde. Folglich liegt es nahe, die genetische Komposition einer Population in allgemeiner Form durch die Verteilung dieser Genome, die also auf elterliche Gameten rückführbar sind, zu definieren. Ein wesentlicher Vorzug dieser Definition liegt darin, daß in vielen Fällen Untersuchungen der genetischen Komposition einer Generation mit solchen der Gametenproduktion der vorangehenden Generation identisch sind. Unsere Stichprobenvariable kann somit auf der Gametenproduktion einer Population erklärt werden. Welche Werte diese Variable annimmt, hängt nun davon ab, ob die einzelnen Loci nicht getrennt oder getrennt voneinander in die genetische Abstandsmessung eingehen sollen, wie dies in den Punkten 1) und 2) des vorigen Abschnittes erläutert wurde. Dort konnte der Punkt 1) auf die Betrachtung nur eines Locus zurückgeführt werden, so daß er sich als Spezialfall des Punktes 2) darstellen ließ.

Seien nun wieder k Loci mit jeweils n_i ($i = 1, \ldots, k$) Allelen in die Betrachtungen einbezogen und bezeichne $e_j^{(i)}$ einen Vektor der Länge n_i, welcher in der j-ten Komponente ($j = 1, \ldots, n_i$) die Zahl 1 und ansonsten überall die Zahl 0 enthält ($i = 1, \ldots, k$). Einem Gameten, welcher am i-ten Locus unter allen dort möglichen n_i Allelen das j-te Allel enthält, wird der Vektor $e_j^{(i)}$ zugeordnet; dies wird für alle Loci $i = 1, \ldots, k$ des betrachteten Gameten wiederholt, so daß letztlich ein aus den $e_j^{(i)}$ zusammengesetzter Vektor der Länge $\sum_{i=1}^{k} n_i$ entsteht, welcher genau k Einsen und ansonsten nur Nullen enthält. Dieser Vektor kennzeichnet den Wert, welchen die Stichprobenvariable dem Gameten zuordnet. Die Verteilung der Zufallsvariablen wird durch die Verteilung der Gameten bestimmt. Einer Stichprobe vom Umfang N aus der Gametenproduktion entsprechen dann also N solcher Vektoren der Länge $\sum_{i=1}^{k} n_i$. Ihre Summe dividiert durch N spiegelt somit in den ersten n Komponenten die relativen Häufigkeiten der Allele am ersten Locus in der Stichprobe wider; gleiches gilt für die nächsten n_2 Komponenten bzgl. des zweiten Locus, etc. Diese relativen Häufigkeiten gehen dann in die o. g. Schätzung des genetischen Abstandes ein. Hieraus mag man nun ersehen, daß die Konsistenz der Schätzfunktion erhalten bleibt, ungeachtet ob die einzelnen Loci abhängig voneinander bzw. gekoppelt sind oder nicht.

Anhang

Im vorletzten Abschnitt hatten wir die dort definierten Metriken $d_0(.,.)$ bis $d_4(.,.)$ (in ihrer normierten Form) zur genetischen Abstandsmessung empfohlen und damit zugleich darauf hingewiesen, daß möglicherweise noch weitere Metriken mit den gewünschten Eigenschaften existieren. Die Entscheidung für die Wahl einer dieser Metriken (Abstände) wird nun davon abhängen, welche zusätzlichen Forderungen man an sie stellt.

Vorab sollte jedoch insbesondere die folgende Invarianzbedingung für alle in dem hier angesprochenen Rahmen zulässigen Metriken weitgehend erfüllt sein: Trifft für die genetischen Kompositionen u, v, x, y vierer Populationen bei Verwendung einer gewissen zulässigen Metrik $d(.,.)$ die Beziehung $d(u, v) \leqslant d(x, y)$ zu, dann soll dies auch für *alle* zulässigen Metriken gelten. Andernfalls nämlich könnten je nach Wahl einer bestimmten Metrik die Abstandsverhältnisse beliebig verändert werden, was dazu führen würde, daß die Ergebnisse kaum mehr eine Aussagekraft besäßen.

Leider läßt sich an einigen Gegenbeispielen zeigen, daß unsere zulässigen Metriken diese Invarianzbedingung nicht in allen Fällen erfüllen. Es erweist sich daher als notwendig eine bestimmte Metrik auszuzeichnen, und im Anschluß daran nur jene zulässigen Metriken zu akzeptieren, welche relativ zu der ausgezeichneten der Invarianzbedingung weitgehend genügen. Das Problem der besten Wahl eines Abstandsmaßes kann dann unter verschiedenen Gesichtspunkten (wie z. B. einfache rechentechnische Behandlung etc.) angegangen werden.

In seiner elementarsten und damit auch anschaulichsten Form bezieht sich der intuitive Abstandsbegriff auf die Längenmessung, d. h. also auf den Absolutbetrag der Differenz zwischen reellen Zahlen. Da nun diese Absolutbeträge direkt und ohne weitere Transformationen additiv in die Definition von $d_0(.,.)$ eingehen, sollte diese Metrik ausgezeichnet werden. Weiterhin ist $d_0(.,.)$ sogar zulässig, so daß die Verwendung einer anderen im oben beschriebenen engeren Sinne zulässigen Metrik nur dann zu vertreten ist, wenn zwingende Gründe vorliegen.

Die anschließenden graphischen Darstellungen sollen exemplarisch die Beziehungen der Metriken $d_1(.,.)$ und $d_3(.,.)$ zur Metrik $d_0(.,.)$ und zugleich mögliche Verletzungen der Invarianzbedingung illustrieren. Zu diesem Zwecke wählen wir die genetische Komposition $\underline{y}$ einer festen Bezugspopulation und betrachten $d(\underline{y},x)$ als Funktion der genetischen Komposition x einer variablen Population. Es wird das Beispiel eines Locus mit zwei Allelen für drei Bezugspopulationen demonstriert.

Man entnimmt den Darstellungen, daß im wesentlichen $d_1(.,.)$ zu größeren Werten als $d_0(.,.)$ und $d_3(.,.)$ zu kleineren als $d_0(.,.)$ führt. Verletzungen der Invarianzbedingungen treten nur dann auf, wenn x und z 'rechts' und 'links' von y liegen; weiterhin sind die Verletzungen bei $d_3(.,.)$ gravierender als bei $d_1(.,.)$, denn es gilt z. B. für

$\underline{y} = \binom{0.3}{0.7}$, $z = \binom{0}{1}$, $x = \binom{0.7}{0.3}$:

$d_0(\underline{y}, z) = 0.3 < d_0(\underline{y}\, x) = 0.4$ aber

$d_3(\underline{y}, z) = 0.404 > \bar{d}_3(\underline{y}\, x) = 0.289$

und für $x = \binom{0.55}{0.45}$:

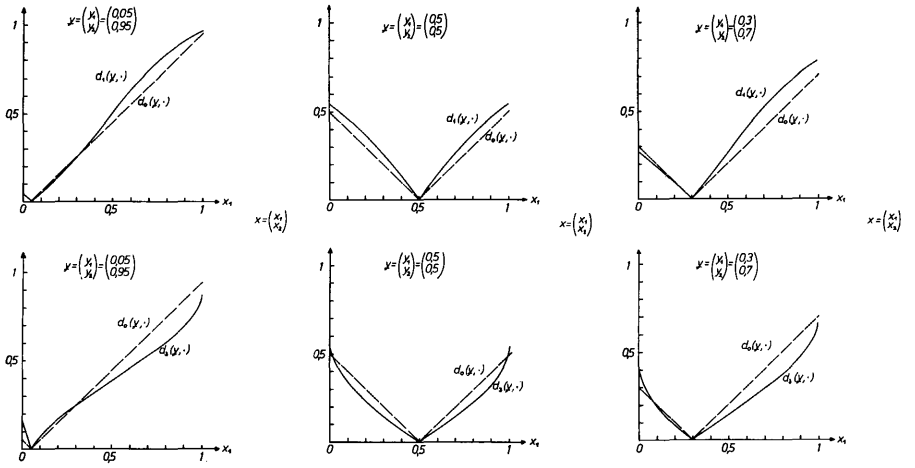
$d_0(\underline{y}, z) = 0.3 > d_0(\underline{y}, x) = 0.25$ aber

$d_1(\underline{y}, z) = 0.284 < \bar{d}_1(\underline{y}, x) = 0.336$.

Diese Feststellungen sind selbstverständlich nicht in direkter Weise auf den allgemeinen Fall übertragbar und bedeuten auch nicht etwa, daß es keine weiteren Metriken gibt, welche die Invarianzbedingung erfüllen, wie das folgende Beispiel zeigt: $\frac{2 \cdot d_0(x, y)}{1 + d_0(x, y)}$ ist eine Metrik (siehe z. B. Rinow, 1961, S. 70) und auch im engeren Sinne zulässig, wie man leicht nachweist.

Zusammenfassung

Die kurzgefaßte Diskussion einiger gebräuchlicher genetischer Abstandsmaße führte zu der Feststellung, daß in den meisten Fällen weitgehende definitorische Unklarheiten und Unzulänglichkeiten bestehen. Insbesondere wurde darauf hingewiesen, daß die Erklärung eines Abstandes zwischen Verteilungen (Stichprobenverteilungen) keine befriedigende Antwort auf die eigentliche Frage nach dem



'wahren' genetischen Abstand zwischen Populationen zu geben vermag. Aus diesem Grunde wurde der Versuch unternommen, eine Konzeption des genetischen Abstandsbegriffes zu geben, welche auf mehreren von der Problemstellung bedingten grundlegenden Forderungen aufbaut. Innerhalb des derart abgesteckten Rahmens konnten einige konkrete Abstandsmaße angeführt werden, von denen sich wiederum eines dadurch auszeichnete, daß es intuitive Vorstellungen am besten widerspiegelte. Das Problem der Schätzung des wahren Abstandes aus Stichproben wurde nur vorläufig umrissen.

Summary

A brief discussion of some measures of genetic distance in use, led to the conclusion that in most cases there exist open points as well as insufficiencies in definition. We particularly pointed out that the interpretation of genetic distance between distributions (sample distributions) provides no satisfactory answer to the proper question about the 'true' genetic distance between populations. For this reason we tried to establish a perspective on the concept of genetic distance based on several fundamental requirements which are implied by the statement of the problem. Within this scope we specified a few actual measures of distance from which one has been distinguished because of its good accordance with intuitive ideas. The problem of estimation of the true distance from random samples has been only preliminarily outlined.

Schlagwort: Genetischer Abstand zwischen Populationen (Genetic distance between populations).

Zitierte Literatur

Balakrishnan, V., and Sanghvi, L. D.: Distance between populations on the basis of attribute data. Biometrics **24**, 859—65 (1968). — Bhattacharyya, A.: On a measure of divergence between two multinomial populations. Sankhya **7**, 401—406 (1946). — Edwards, A. W. F.: Distances between populations on the basis of gene frequencies. Biometrics **27**, 873—81 (1971). — Edwards, A. W. F., and Cavalli-Sforza, L. L.: Reconstruction of evolutionary trees. In Phenetic and Phylogenetic Classification, Publ. No. 6, 67—76, Systematics Association, London, 1964. — Edwards, A. W. F., and Cavalli-Sforza, L. L.: Affinity as revealed by differences in gene frequencies. In the Assessment of Population Affinities in Man. Clarendom Press, Oxford, 1972. — Fisher, R. A.: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, 1958. — Kurczynski, T. W.: Generalized distance and discrete variables. Biometrics **26**, 525—34 (1970). — Nei, M.: Genetic distance between populations. The American Naturalist **106**, No. 949, 283—92 (1972). — Mahalanobis, P. C.: On the generalized distance in statistics. Proc. Nat. Inst. Sci. India 2, 49—55, 1936. — Rinow, W.: Die innere Geometrie der metrischen Räume. Springer Verlag, Berlin—Göttingen—Heidelberg, 1961. — Steinberg, A. G., Bleibtreu, H. K., Kurczynski, T. W., Martin, A. O., and Kurczynski, E. M.: Genetic studies on an inbred human isolate. Proc. Third Int. Congress Hum. Genetics. Eds. J. F. Crow and J. V. Neel. Johns Hopkins Press, Baltimore, 1967.