TAXON 21(4): 381-406. AUGUST 1972

THE DEFINITION OF SYSTEMATIC CATEGORIES IN BIOLOGY*

*Pierre Legendre***

Summary

Biological taxonomy, now combined with cytogenetics and mathematical philosophy, has become a new synthetic theory of evolution. The purpose of this paper is to derive a comprehensive, united series of formal descriptions of the results of evolution, which are the systematic categories as understood by biosystematics. In a first move, differences between individuals are found by comparing their chromosomal arrangements, and algebraic measures of feasibility of pairing are derived. Individuals are also compared with regard to their genes, and an algebraic measure of genic similarity between individuals is defined. A chain is also defined, which unites in clusters the groups of individuals which are equivalent with regard to a relation which is to be defined in each case. With these mathematical tools, a species is defined as a group of individuals which cluster when one applies on them a relation showing their possibility of crossing freely. A genus is defined as a group of species which cluster after a chain is formed on pairs of species between which there is a calculated possibility of occasional hybridization. A local population is defined as a group of organisms located within pairing distance of each other. A subspecies is defined as a major subdivision of the specific gene pool which corresponds to a geographical subdivision of the species' range. The usefulness of the semispecies as a category different from the subspecies is discussed according to biosystematic principles. It is also suggested that an environmental multi-dimensional space could be of major usefulness for determining the major adaptive peaks reached by supra-generic taxa.

Résumé

La taxonomie biologique, qui s'appuie maintenant sur la cytogénétique et la philosophie mathématique, est devenue une nouvelle théorie synthétique de l'évolution. Ce travail présente une série de descriptions formelles des résultats de l'évolution; ces résultats sont les catégories systématiques telles que comprises par la biosystématique. En premier lieu, les différences entre individus sont établies par comparaison de leurs chromosomes et leur capacité d'accouplement est décrite algébriquement. La comparaison des gènes des individus mène à une mesure algébrique de la similarité génique des individus. On définit aussi une chaîne qui groupe les individus qu'une relation (à déterminer dans chaque cas) définit comme équivalents. À l'aide de ces instruments mathématiques, les individus qui se groupent lorsqu'on leur applique une relation montrant leur capacité de croisement, sont définis comme formant une espèce. Les espèces qui se groupent à l'aide d'une chaîne montrant leur capacité de former occasionnellement des hybrides, sont définies comme formant un genre. La population locale est définie comme un groupe d'organismes situés à proximité suffisante les uns des autres pour qu'ils puissent s'accoupler. La sous-espèce est définie comme une subdivision génétique majeure de l'espèce qui correspond de plus à une subdivision géographique. L'utilité de la semi-espèce comme catégorie différente de la sous-espèce est évaluée selon les principes biosystématiques. L'on suggère aussi qu'un espace écologique multi-dimensionnel pourrait être très utile pour déterminer quels sont les sommets évolutifs atteints par les taxa de niveau supra-générique.

---

* Part of a thesis (Legendre, 1971b).

** Department of Biology, University of Colorado, Boulder, Colorado 80302, U.S.A. *Present address:* Centre de Recherches écologiques de Montréal, Université du Québec à Montréal, C. P. 8888, Montréal 101, Québec.

1. Logical background

Ten score years ago, any treatment of natural history would begin with the description of those parts of living nature that were most familiar and, therefore, supposed to be the most understood by humans — the beasts, the birds, the fishes, the flowers, and the trees, passing on perhaps to less well-known parts of life, the reptiles, the insects, the grasses, the fungi, and the algae; then it might go on to include the worms in the soil and even stones. The general guiding idea was that of degrees of perfection, the great ladder of life, beginning with man as nearest to the angels and going on to the less perfect realizations of the divine archetype.

With the triumph of the theory of evolution as presented by Darwin (1859), this approach was abandoned. From this point on, it was thought to be more appropriate to begin with the simple unicellular plants and protozoa and then to go on to the higher and more evolved species following the ever more branching evolutionary tree.

The question, how to proceed with the study of the evolution of life, has been dominated by experimental biologists, though it is no less the task of those recognizing the need of the theoretical approach. A combined experimental and theoretical approach is especially important in the field of classification, which has become considered by many as the poor relative of experimental biology. However, during recent decades biological taxonomy has been revitalized with its new role as an approach to evolutionary principles and results. Hence, its combination with cytogenetics and mathematical philosophy has helped to transform it into a kind of a new synthetic theory of evolution. As a first order, this synthesis uses and formalizes the concepts of classification which are the most important bases for the studies that may solve the problem concerning where to proceed with the explanations of the principles and processes of biological evolution.

We can consider this trend to have started with the later works of Darwin himself, or at least with the geneticists of the early decades of this century, even though they were more concerned with uncovering the principles of heredity and the processes of genetics than with speculations about their significance for the theory of evolution or the classification of living matter. A synthesis of such speculations and experiments was attempted by Turesson (1922), who proposed a new species concept which was essentially genetical but bore little relation to the ideas and experience of taxonomists. An equally abortive attempt was made by Danser (1929), who lost sight of the evolutionary synthesis in his jungle of unwieldy terms. The lack of a genetical background for the attempts at new definitions of the basic biological concepts by Du Rietz (1930) prevented their acceptance as a basis for the waiting synthesis, as did the same faults of numerous other such attempts, so that even when Hultén (1937, 1968) tried to replace cytogenetics with geographical observations, he was unable to apply the definitions without serious misgivings. The real experimental basis for a new theoretical synthesis that led to the discovery of reproductive isolation as the only basis for evolutionary classification was first furnished by Müntzing (1930) when he discovered, in *Galeopsis*, that differentiation within the species is based on gene mutation and genetic recombination, whereas that between species is caused by chromosomal rearrangements, linear or numerical. Unfortunately, his limited interest in theoretical taxonomy prevented him from seeing the significance of

this discovery, and so it was left to Dobzhansky (1941), Mayr (1942), Löve and Löve (1942), and possibly several others to observe the significance of this as a basis for the concept of the evolutionary, or biological, species.

The biological species concept may be said to have been conceived already by pre-Linnaean biologists, like Ray (1686) who proposed that a species name should be given only to those plants which breed true from seeds within their own limits. This was ignored by Linnaeus (1735, 1737, 1751) and his followers up to the present time. Although de Candolle (1813) presented a clearly evolutionary species concept, this was overlooked by Darwin (1859) and almost all later evolutionists, and the same fate came to similar attempts by others until the time was ripe for its rephrasing in the early years of the fifth decade of the present century. Since then, this concept has caused much discussion which has circled about its applicability at the same time as it has led to a re-evaluation of the species and other taxonomic categories, and even to a rational questioning about the validity of the neo-Darwinian approach itself. The large amount of literature published in the past 10 or 15 years on this subject, in *Taxon*, *Systematic Zoology* and other journals, is a proof of the profound interest and of the intensity of the revolution caused by the new concepts. On one side, there are those who try to reconciliate the new and the old theories, and on the other, those who simply investigate the new concepts. In a line of thought somewhat parallel to the latter, are found those interested in developing numerical methods for taxonomic purposes, based or not on the theoretical foundations of systematics.

The need for theoretical, descriptive models of biological phenomena has been acknowledged by some recent authors, and notably by Estabrook (1970). This author points out that true science proceeds by making assumptions (empirical laws) about reality, assumptions that are mathematically reworded in theorems (models), which in turn are used to make predictions that can be tested experimentally against reality. Then, the reality can be used, together with what is now known about the truth value of the models, to formulate an integrated series of assumptions that we call a theory. The theory of evolution is a good example. Most of the work in systematics has been done at the level of observation of reality. Fewer people have been interested in the second step, that is, thinking about the nature of the systematic categories. Ray and Linnaeus were among those. And more recently, people like Dobzhansky, Mayr, and Löve and Löve have made better guesses which led to the development of the biological species concept. From this and other bases, mainly genetical and cytological, has been developed the theory of biosystematics. This theory is, in our opinion, so stabilized that we can proceed to its formal rewording. We will first define the organisms, and by applying the theory of biosystematics to them, we will try to get a comprehensive, united series of descriptions of the results of evolution. (Symbols will be used only in order to get a better step-by-step reasoning. Our aim being primarily to describe the units of evolution, we do not intend to build here a mathematical theory of evolution). If the models correspond well to reality, we can conclude that the assemblage (theory, models) is satisfactory. In the present paper, however, we are mainly concerned with the rewording process.

Our conviction that the theory of biosystematics is ready for this attempt of formalization is based partly upon the trend that can be seen in recent literature. Indeed, many recent papers about biosystematics try to general-

ize evolutionary ideas as precisely as possible, using a language related to that of mathematics. For example, let us consider this sentence: "There are important clusters separated by relatively empty spaces in the pattern of diversity" (Cronquist, 1969: 180). Almost every word corresponds to a concept that would be easily translated into mathematical terms. This is what we will attempt in the following pages, concerning the theory of biosystematics, trying to relate the systematic phenomena to a limited set of genetic factors. This exercise, however, is more than a simple rewording, since it will force us to question the precise meaning of the terms used, and it will also illuminate the gaps (or, in other words, it will expose the terms that cover a lack of inquiry about parts of the process). These gaps, when they are small, are easily bridged. In the case of those large ones that we will see, we will try to present a temporary solution, the adequacy of which will be questioned, we do hope, so that new solutions may be found.

It has to be kept in mind that the real difficulty in making a model is not to fit a theory, but rather to fit reality. In this process, we will have to propose some new concepts, the truth value of which we cannot fully appreciate until they can be subjected to experience. That will clearly illustrate the fact that modeling is an essential step in the process of testing theories, because the defects in a theory are made obvious when one tries to formulate it with a precise language.

2. The organism

Our immediate purpose is to arrive at a better understanding of the systematic categories; the organisms will therefore here be described in a comparative manner (it is this element of comparison that was lacking in our previous model of species: Legendre and Vaillancourt, 1969).

We will first describe, at the level of the individuals, two comparative elements that will be used later. These are the chromosomal and genic differences. Only individuals whose reproduction is based on allogamy, at least part of the time, will be considered.

2.1. Symbols used

The following list is simply intended as a reference dictionary to be consulted as one reads the text, and not as a list of formal definitions.

B: set of objects under study

C: subset of B

E_e: equivalence relation defined over level of similarity e

F (g_1 (L′), g_1 (M′)): match function that compares the corresponding gene in cell L′ of individual $\bar{L}$ and in cell M′ of individual $\bar{M}$

G ($\bar{L}$): function of gamete production of individual $\bar{L}$

H_e: chain defined over level of similarity e

J. ($\bar{L}$, $\bar{M}$): general measure of similarity of two individuals, $\bar{L}$ and $\bar{M}$

K ($\bar{L}$, $\bar{M}$): general measure of distance of two individuals, $\bar{L}$ and $\bar{M}$

L, M: set of gametes produced by individuals $\bar{L}$ and $\bar{M}$, respectively

L′, M′: cells of individuals $\bar{L}$ and $\bar{M}$, respectively

$\bar{L}$, $\bar{M}$, $\bar{N}$: three individuals

P_d (l, m): difficulty of pairing of gametes l and m

P_f ($\bar{L}$, $\bar{M}$): feasibility of pairing of individuals $\bar{L}$ and $\bar{M}$, so that the result would be a viable and fertile offspring

P'_f ($\bar{L}$, $\bar{M}$): feasibility of pairing of individuals $\bar{L}$ and $\bar{M}$, so that the result would be a viable, but not necessarily fertile offspring

Q_i: individuals
R ($\bar{L}$, $\bar{M}$): genic resemblance between individuals $\bar{L}$ and $\bar{M}$
S ($\bar{L}$, $\bar{M}$): difference of sex of two individuals $\bar{L}$ and $\bar{M}$
S_d (l, m): sex chromosome difference between gametes l and m
T: threshold on the scale of chromosomal difference, between the possibility and the impossibility of producing fertile offspring
T′: threshold on the scale of chromosomal difference, between the possibility and the impossibility of producing a zygote
a, b, c: three indices of difficulty of pairing
g_i: the various genes
l, m: gametes produced by individuals $\bar{L}$ and $\bar{M}$, respectively
l_i, m_i: i-th gamete of individuals $\bar{L}$ and $\bar{M}$, respectively
n: total number of genes
n (L′): number of genes in cell L′
v: value of R ($\bar{L}$, $\bar{M}$)

2.2. *Chromosomal differences between individuals*

Our intention is to define in algebraic terms various measures that could be used to determine the possibility of mating between two individuals, and what the result will be in terms of fertility or sterility of the offspring. We will first define what we mean by chromosomal differences, and then what the effect of these differences can be expected to be.

Since we are looking for a theoretical model, we can feel free to use measurements that are unpractical because it would be too much time consuming to get the data. Hereafter, we will assume that we can see, count and identify the chromosomal mutations and the genes.

2.2.1. The chromosomal differences

The differences between gametes l and m could be identified using a gene-by-gene comparison, and defined as follows:

2.2.1.1. Sex chromosome differences

The sexual differences can be identified by looking at the sex chromosomes, or in cases where there is no sex chromosome difference in the given species, by looking at the genic combination that determines sex. Then, for the pair of gametes l and m, the sexual difference S_d can be defined as:

S_d (l, m) = 0 when there is no difference, or
S_d (l, m) = 1 when the two gametes carry different sexual determinators.

2.2.1.2. Chromosomal mutations

The term chromosomal mutation is used here to designate both the chromosomal modifications and the genome mutations.

The chromosomal mutations are somewhat more difficult to handle, because there exist different types of them. An index of difficulty of pairing will be defined for each type of mutation, according to the type of pairing difficulty they create when in heterozygous condition. Then the index of difficulty of pairing will be multiplied by the fraction giving the relative length of the segment involved in the mutation, by reference to the length of the karyotype, of all the chromosomes of the gametes involv-

ed, in order to get a measure of the pairing difficulty due to a given mutation. The indices of difficulty of pairing are the following positive values: a deficiency and a duplication will each be assigned an index *a* of difficulty of pairing, since they cause the same type of distortion of one of the chromosomes of a heterozygous pair; a shift (*sensu* Solbrig, 1970: 154) will be given an index *2a*, since it causes the formation of two loops of a similar kind as in the case of deficiency or duplication; a translocation is given an index *b*, and an inversion an index *c*; to a duplication of chromosome, or in general the presence of any unmatched chromosome, is attributed an index 1, by comparison to which the relative value of a, b and c could be established after experimental work.

The difficulty of pairing P_d of gametes l and m can then be defined as follows:

$$P_d(l, m) = \sum_{\substack{\text{all chromosomal differences i, but sexual differences}}} (\text{index attributed to the i-th mutation}) \times \frac{\text{length of segment included in i-th mutation}}{\max(\text{length of karyotype of gamete l, length of karyotype of gamete m})}$$

(See also section 2.2.2.2. below).

2.2.2. Effects of chromosomal differences

The effects of chromosomal differences can be studied as follows:

2.2.2.1. Sex chromosome differences

Two gametes can give rise to a zygote only when they come from individuals of different sexes (monoecious plants and hermaphrodite animals are not discussed in detail here, but one can easily make the extrapolation). The ability of two individuals $\bar{L}$ and $\bar{M}$ to mate and produce a zygote can be described by a value that we will call S, which is defined by the formula:

$$S(\bar{L}, \bar{M}) = \left| \max_{l_i, l_j \varepsilon L} S_d(l_i, l_j) - \max_{m_i, m_j \varepsilon M} S_d(m_i, m_j) \right|$$

that will give a 0 value when both the individuals are of the same sex, and a 1 value when they are of opposite sexes. In the formula,

$$L = \{ l \,/\, l \text{ is a gamete produced by individual } \bar{L} \} \text{ and}$$

$$M = \{ m \,/ m \text{ is a gamete produced by individual } \bar{M} \} .$$

The function S_d is as defined above. Let us take an example to see how the formula works: suppose that $\bar{L}$ is a female individual and $\bar{M}$ is a male, and that the sex determining mechanism is the XX/XY system. Then the maximum value that can be obtained for $S_d(l_i, l_j)$, for any pair of gametes formed by $\bar{L}$, is 0, since all the gametes carry a X chromosome. But in the case of individual $\bar{M}$, that is an XY male, even if the comparison of two gametes with X, or with Y chromosomes, would give a 0 value, the maximum value of $S_d(m_i, m_j)$ would nevertheless be 1, since the comparison of a gamete carrying an X and one with a Y gives a S_d value of 1. Then we are left with $|0 - 1| = |-1|$ that is equal to 1 when we take the absolute value of the number inside. By the same procedure, comparison of two females

or of two males would give a 0 value. One should note that this equation works as well with the XX/XY, XX/XO, ZW/ZZ or ZO/ZZ systems, either in their basic form or with neo-sex-chromosomes. It is also applicable in the case of genic sex determinators without visible chromosomal differences.

2.2.2.2. Chromosomal mutations

A certain degree of chromosomal polymorphism is found in many species. Such chromosomal differences do not necessarily cause sterility. On the other hand, biosystematics tells us that real reproductive isolation between species, and consequently also between higher taxa, is caused by a sufficient amount of chromosomal differentiation or polyploidy. Consequently, there must be a threshold value on the scale of chromosomal differentiation between fertility and sterility. Another refinement that we will need when modeling the genus is a difference in the offspring, when it is either fertile or sterile. The concept of gene flow between taxa requires that we take this difference into account, supposing the existence, on the scale of chromosomal differentiation, and in the range of the fertility of two gametes, of another threshold value between fertility and sterility of the offspring. These threshold values can be defined *a posteriori* by saying that two gametes can

— give rise to a fertile offspring (capable of back-crossing indefinitely, that rules out fertile panallopolyploids), when there is no chromosomal difference between them, or when the differences do not cause a difficulty of pairing larger than a threshold value T; or,

— give rise to a zygote that will not develop into a fertile individual (capable of back-crossing indefinitely), or else the individual cannot give rise to fertile descendants, when the difficulty of pairing of the chromosomes is larger than T but smaller than another threshold value T′; or,

— cannot produce a non-alloploid zygote, when the difficulty of pairing of the chromosomes is larger than T′.

Let us decide that the maximum value that can be taken by P_d (l, m) is 1, in the case where the chromosomes of the two gametes are completely different. The minimum value of P_d (l, m) is 0, as can be seen on the formula of section 2.2.1., since we have defined a, b and c as positive values. T and T′ will then have to be between 0 and 1, but smaller than 1, and so will a, b and c, since none of the mutations to which they correspond can create a difficulty of pairing larger than that created by a completely different chromosome, to which case an index 1 was attributed. This leads obviously to the conclusion that panalloploid offspring cannot be used in the establishment of T and T′, since it is formed from gametes between which there is a difficulty of pairing of 1, or a value very close to 1. This is why the gametes that produced panalloploid offspring, like the well-known *Raphanobrassica*, or *Hylandra* (Löve, 1961), do not exhibit a pairing difficulty smaller than T′, so that they are not indicators of generic relationships between their parent species (see section 3.3. below). On the other hand, in hemialloploids formed from hybridizing species of various ploidy levels of a genus which contains a polyploid series, some of the chromosomes of the parental gametes pair, so that their pairing difficulty P_d (l, m) is sufficiently below 1 to be smaller than the value of T′. So is also the case with allopolyploids as found, for instance, in the genus *Avena*: in this example, the diploids, tetraploids and hexaploids are members of

the same genus, since they all share the A genome, and the A, B and CD genomes also show some homology (Rajhathy and Morrison, 1959).

We cannot tell at this point if universal values of T and T′ do exist, or if constant values could be found only for restricted groups. Perhaps these values would vary significantly from one pair of species to another. However, at this point we can propose an experimental scheme that would lead to a good approximation of these values, at least for a restricted group of species. It presupposes again that the arrangement of the segments on the chromosomes is known, which is not an unrealistic assumption as far as modern cytological techniques of investigation reach: even on material without polythene chromosomes, the Q (Caspersson *et al.*, 1968) and G (Sumner *et al.*, 1971; Seabright, 1972; Kato and Yosida, 1972) banding techniques could provide the necessary information. In order to find T, the largest value of intra-specific chromosomal pairing difficulty found is taken as the lower bound of T. The lowest value of inter-specific pairing difficulty found is taken as the higher bound of T: this test should be done by comparing the chromosomal arrangements of the species under study with those of the most closely related species. T will then be somewhere between the higher and lower bounds. The accuracy of determination of T can be increased, to a certain extent, by studying more cases of upper and lower bounds. In this process, a good approximation of the values of indices a, b and c, for at least a group of related species, could probably be found by studying each pair of species in the group. T′ could be estimated by successive approximations of the same type as in the case of T, although a cytological investigation would be made more difficult by the increasing occasionality of formation of zygotes, when approaching the T′ value after which the chromosomal difference between the species is so large as to lead to complete sterility.

On the other hand, experimental data could also show if the formula of P_d (l, m) above can bring stable enough values of a, b, c, T and T′ for a group of related species, or if it should be replaced by a more complex formula in which the length of each mutation would be compared only to the length of the chromosome in which it takes place. This formula would be of the type

$$\frac{1}{\text{n. of chromosomes}} \times \min \sum_{\substack{\text{all chromosomal differences i, but sexual differences}}} (\text{index attributed to i-th mutation}) \times \frac{\text{length of segment included in i-th mutation}}{\max (\text{length of chromosome of gamete l involved, length of chromosome of gamete m involved})}$$

The minimum of the sum is used here to indicate that we are interested to compare only the corresponding chromosomes of the two gametes.

The use that we make of the parameters T, T′, a, b, and c is not a function of the ease with which these parameters can be determined. It depends only on their existence, since what we seek is a theoretical model. The values of a, b, c, T and T′, and consequently also the values of P_f ($\bar{L}$, $\bar{M}$) of sections 2.2.4. below, could now be estimated only in the few cases, like *Drosophila*, where the cytological studies have been so extensive that the detail of the chromosomal morphology is known.

2.2.3. Sterility

Before proceeding with more sophisticated measurements, we should first take into account the fact that a given individual may be sterile, independently of its age. For instance, most F_1 inter-specific hybrids are. We will describe this phenomenon by saying that G(L) is 1 when individual $\bar{L}$ can produce gametes, and 0 when it cannot; and similarly for individual $\bar{M}$.

2.2.4. Feasibility of pairing

We can now define the feasibility of pairing P_f of individuals $\bar{L}$ and $\bar{M}$:

$$P_f(\bar{L}, \bar{M}) = \max_{\substack{l \varepsilon L \\ m \varepsilon M}} (T - P_d(l, m)) \times S(\bar{L}, \bar{M}) \times G(\bar{L}) \times G(\bar{M})$$

in which $P_d(l, m)$, $S(\bar{L}, \bar{M})$, $G(\bar{L})$ and $G(\bar{M})$ are as defined above. The properties of this definition are as follows:

– when at least one of the individuals cannot produce gametes, then $G(\bar{L})$ or $G(\bar{M})$ is 0, and consequently $P_f(\bar{L}, \bar{M})$ is equal to 0;

– when two individuals are of the same sex, $S(\bar{L}, \bar{M})$ is 0 and then $P_f(\bar{L}, \bar{M})$ is equal to 0;

– when two individuals are of opposite sexes and can produce gametes, then $S(\bar{L}, \bar{M}) \times G(\bar{L}) \times G(\bar{M})$ is equal to 1, and consequently $P_f(\bar{L}, \bar{M})$ is equal to the maximum value that can be taken by $(T - P_d(l, m))$. The maximum of $(T - P_d(l, m))$, for a fixed T, corresponds to substracting from T the minimum value of pairing difficulty that can be found for all the gametes produced by $\bar{L}$ and $\bar{M}$. The use of this procedure can be justified by the following example: suppose that we are trying to cross a homozygous with a heterozygous individual. It can happen in border cases that, even if these two individuals are of the same species, the chromosomes of some recombination gametes of the second individual could not pair with those of the first one, but some other recombination gametes could. It would be unrealistic to say that the pairing difficulty of these two individuals is larger than T. Instead, by using the minimum of $P_d(l, m)$, we get a pairing difficulty smaller than T;

– when the pairing difficulty $P_d(l, m)$ is larger than the threshold value T, $P_f(\bar{L}, \bar{M})$ is negative. When it is equal to T, then $P_f(\bar{L}, \bar{M})$ is equal is 0.

The biological significance of the feasibility of pairing P_f can be seen by looking at the concept of fecundity. The fecundity of a cross between two individuals would be a function, probably non-linear, of P_f. However, P_f is obviously not the only factor which determines fecundity, which is also a function of the fitness of the overall allele combination in the given environment.

2.3. Genic similarity

In the study of subspecies, we will need to compare individuals with regard to their genic similarity. It is this measure of similarity $R(\bar{L}, \bar{M})$ between two individuals which we intend to define here.

2.3.1. The simplest case

The study of the genic similarity between two individuals is complicated by phenomena like sex chromosome differences, differences in chromosomal arrangements and the presence of segments of chromosomes in one individual and not in the other, and these make it impossible in real cases to

make a straight forward comparison of the genes on the chromosomes. It is also complicated by the multiplicity of similarity coefficients available, which have been discussed by Sokal and Sneath (1963: 128-139). These authors point out (p. 128), however, that the fundamental formula of all these coefficients "consits of the number of matches divided by a term implying the possible number of comparisons but varying in its detailed composition".

Consequently, we will first consider the simplest case possible, in order to help us choose the measure of similarity that is best adapted to our needs. This case is that in which there are no sex chromosome differences, no differences in chromosomal morphology, and where there are exactly the same number of genes on the chromosomes, disposed in corresponding sequence. In this case, the obvious choice is the coefficient called Simple Matching by Sokal and Sneath (1963: 129), which is equal to the number of matching pairs of genes divided by the total number of pairs of genes (n).

But we must first ask if it is possible to apply this coefficient on the 2n somatic number of chromosomes. To answer this question, we may reason as follows: suppose that we are using the somatic cells for the comparison of individuals $\bar{L}$ and $\bar{M}$, and suppose that $\bar{L}$ and $\bar{M}$ are monozygotic twins heterozygous for many loci. If we compare the similar chromosomes, then all pairs of genes will match and the similarity ratio R ($\bar{L}$, $\bar{M}$) will be equal to 1. But if the chromosomes are compared in heterozygous pairs, then the number of matching pairs of genes will be lower than the total number of pairs of genes, and R ($\bar{L}$, $\bar{M}$) will be smaller than 1. But the obvious answer in this case of monozygotic twins is 1, and the problem consists in finding a method to get it every time.

One method would be to compare the genes on the chromosomal complements of sufficiently many cells of the two individuals, and to take as R value the maximum of all the similarity ratios found. In practice, this comparison could be done by a method like DNA hybridization. The formula expressing this treatment is

$$R(\bar{L}, \bar{M}) = \max_{\substack{L' \text{ is a cell of } \bar{L} \\ M' \text{ is a cell of } \bar{M}}} \left| \frac{\sum_{i=1}^{n} F(g_i(L'), g_i(M'))}{n} \right|$$

where n is the total number of pairs of genes in the comparison, that is in this simple case the total number of genes in each cell. The various $g_i(L')$ are the various genes in the cell L′ of individual $\bar{L}$, where i varies from the 1-st gene to the n-th gene. Similarly for $g_i(M')$. $F(g_i(L'), g_i(M'))$ is the match function, that takes the value 1 when $g_i(L')$ and $g_i(M')$ are identical, and the value 0 when they are different.

According to this formula, the R value chosen will be the one that corresponds to the best fit of corresponding chromosomes, and it can take any value between 0 (complete unmatch) and 1 (complete match).

2.3.2. Extra chromosomal segments in one of the individuals

The first complication that we will bring into the model is the case where the only non-genic difference between two individuals consists in

the presence, in one of the individuals, of segments of chromosomes that are not found in the other. These segments are likely to be the result of duplications, in the real case, but let us assume that they can hold also genes that are not found elsewhere on the chromosome complement of the holder. This is for the purpose of the present discussion, since we are on our way to more general cases. We can also assume that these segments will be situated at the ends of the chromosomes, so that the pairing of the chromosomes will be facilitated. We will study the effects of chromosomal rearrangements in section 2.3.3.

We have obviously to take into account the presence of these extra genes on the complement of one of the individuals. This can be done by replacing the denominator n of the equation of section 2.3.1. by the maximum of the number of genes found in the cells L′ and M′. This can be written:

$$\max_{\substack{L' \text{ is a cell of } \bar{L} \\ M' \text{ is a cell of } \bar{M}}} (n(L'), n(M'))$$

where n(L′) and n(M′) are the number of genes found respectively in the cells L′ and M′.

2.3.3. Differences in chromosomal arrangements

Since we are after a measure of genic similarity between individuals, we want to leave out the chromosomal mutations. We can assume that these mutations have very little effect on the phenotype when they cause breakages of the chromosomes at the level of the punctuations separating the genes. If, on the contrary, they break genes, our method of measuring will then consider the segments as separate genes and will make the comparison accordingly. On the other hand, we do not want to leave genes out of the comparison just because they are locked in a chromosomal rearrangement, since these genes still have an effect on the phenotype, and we want to measure here all the genic causes that determine the phenotype, so that we may obtain a measure of intra-specific diversity.

But let us consider for a moment the method that we said we were using, practically, to determine genic differences, which is a method related to the DNA hybridization technique. These techniques already imply breaking the chromosomes into small pieces before making the comparisons. Consequently, geographical rearrangements on the chromosomes would not influence the formula. The chromosomal mutations that imply duplication or disappearance of segments have already been taken care of in section 2.3.2.

2.3.4. Sex chromosomes

The genic differences due to sex chromosomes have to be excluded from the comparison, since they are not in any way indicators of differences between gene pools. One way to achieve this would be to supplement the formula of genic similarity ratio with a function based on $S(\bar{L}, \bar{M})$, so that the comparison would be made only between individuals of the same sex. However this solution would be finally quite unpractical, since we would end up with dividing allogamous taxa into two groups of compared objects, males and females, between which no comparison would be possible.

Consequently, we will instead have to use an experimentally unrealistic

system, of the kind we have used to define P_d (l, m), in which we will simply state that the sex chromosomes are excluded from the comparison. This can be done by saying simply that the various $g_i(L')$ are the various genes found on all chromosomes, except the sex chromosomes, of cell L′ of individual $\bar{L}$. Similarly for $g_i(M')$. n becomes the number of genes found on all chromosomes, except the sex chromosomes. With these restrictions, the formula is now

$$R(\bar{L}, \bar{M}) = \max_{\substack{L' \text{ is a cell of } \bar{L} \\ M' \text{ is a cell of } \bar{M}}} \left[\frac{\sum_{i=1}^{n} F(g_i(L'), g_i(M'))}{\max_{\substack{L' \text{ is a cell of } \bar{L} \\ M' \text{ is a cell of } \bar{M}}} (n(L'), n(M'))} \right]$$

3. The systematic categories

The systematic categories are the result of an evolutionary process which begins at the level of the local population, or deme, and proceeds towards the higher categories, through the variety, subspecies, species, genus, family, order, etc. Its basic processes are those of increasing variation, which is based on mutations and hybridization between completely interfertile individuals or demes of the same gene pool, followed by natural selection. At the level of species a new process is added when reproductive isolation sets in to conserve favorable combinations; after that level is reached, natural selection resulting in extinction of intermediates becomes the main process that decides the formation of all higher categories, be they related genera that are naturally grouped into families, or families grouped into orders. The main distinction at these higher levels, where miscibility of genes and forming of new gene combinations is effectively prevented, is connected with evolutionary distance, or the distance in time from the source of evolution which is the deme of the interfertile gene pool. The older the taxon is in such terms, the greater are its differences as compared to other taxa at the same level and the clearer the agreement on its classification by whatever methods are employed.

This distance in history explains the difference in distinction between the taxonomic categories, and also the ease with which they can be defined. Since the processes with which the gene pool is being differentiated are essentially the same, at the level of deme up to the level of subspecies, these categories are difficult to define in such a way that a full agreement can be reached as to their limitations; even the geographical distinction between variety and subspecies attempted by numerous authors and sharpened by Hultén (1968) is not generally accepted, so that several authors still regard as a subspecies races that others find to be typical varieties, and vice versa, and even the same author may not be consistent in his choice within the same limited group of biota. A clear evolutionary definition is possible at the level of species because of its reproductive barrier (Mayr, 1942), whereas the level of genus again is not free from difficulties. When defined by Linnaeus (1751) in Philosophia Botanica, the genus was regarded as the second most important category, since it united related species which, according to the later opinion of this author, might have

developed from a generic prototype. Later definitions have failed to improve this first one to any conceivable degree, except that the unwritten biological definition of recent decades seems to tend to require of a genus that all its species could have evolved from its prototype by gradual or abrupt speciation without a dysploid change of basic number (cf. Kirpicznikov, 1968, and Löve, 1963). The limits between related but distinct genera are sharpened by natural selection through extinctions of intermediates in a considerably higher degree than when selection broadens the gaps between gradual species.

Definitions of families are all based on their morphological distinctions, which are usually great and indisputable, although some few plant families are still perceived a little differently by different authors. At the level of orders and higher categories, however, definitions seem unnecessary because most of the taxa are distinct that no one would have any difficulty in seperating them at a glance.

Using the building blocks that were defined in the previous chapter, it is possible to derive models of various systematic categories in order to strengthen their limitations and sharpen their definition. However, we will limit our discussion to the categories below the family level, because objective criteria have been defined only for them by aid of the biosystematic approach. Indeed, the categories at and above the family level lack objective criteria and are thus somewhat subjective, as exemplified by the paper of Dillon (1963) in which all living beings are classified in the plant kingdom, although these taxa usually are surprisingly stable and commonly agreed upon; the cause for their stability seems to be their distinction which is caused by their considerable distance from their evolutionary origin and by increasingly severe selection through a long period of time. A good example of such a stability is the orders of insects, which seem satisfactory to most entomologists. This can be due to one, or both, of two reasons: either there are unrecognized objective criteria for the establishment of taxa in these categories, a process that is already facilitated by the fact that species and genera can be defined according to objective criteria, and that kingdoms and even phyla are so few as to be easily defined, establishing fixed markers at both ends of the categoric scale; or, most systematists are more interested in alpha- or beta-taxonomy than in taxonomy of higher order of abstraction. This second hypothesis is most probably correct for many groups of organisms: it will suffice to mention as an example the work of Jarvik and especially his 1960 publication, in which comparative anatomical evidence led to a complete reworking of the classification of the vertebrates at the level of the higher categories. In this case, some zoologists accept the conclusions of Jarvik, many do not, but very few venture to discuss them.

3.1. The H-chain

Our purpose in the next sections will be to group the biological objects into taxa pertaining to various systematic categories, each category having its own characteristics which are defined by the theory of biosystematics. This grouping of the objects, up to the generic level, will be done by clustering them with a technique derived from the one described by Estabrook (1966), that is based on a graph theory model.

A cluster is defined by the relation of hybridization (*sensu lato*) called H, the precise meaning of which will be defined for each category. The general form of this relation is defined as follows: for any pair of objects $\bar{L}$ and $\bar{M}$,

$\bar{L}$ H_e $\bar{M}$ if and only if J ($\bar{L}$, $\bar{M}$) $>$ e.

This is to be read: '$\bar{L}$ is connected with $\bar{M}$ at level of similarity e if, and only if, the given measure of similarity, determined by function J, between $\bar{L}$ and $\bar{M}$ is larger than e'. The relation H for a pair of objects can as well be defined in a similar way by a distance function called K:

$\bar{L}$ H_e $\bar{M}$ if and only if K ($\bar{L}$, $\bar{M}$) $\leqslant$ e.

The properties of symmetry and reflexivity of H_e are clear, since $\bar{L}$ H_e $\bar{M}$ if and only if $\bar{M}$ H_e $\bar{L}$, and $\bar{L}$ H_e $\bar{L}$ always. This fits what we want a partition of objects into taxa to be: every two objects to which a similarity value larger than e, or a distance at least as small as e, has been attributed, will be included in the same cluster.

The notion of a H_e-chain can now be introduced. A H_e-chain is said to exist from $\bar{L}$ to $\bar{M}$ if there is a series of objects $\bar{Q}_1$, $\bar{Q}_2$, $\bar{Q}_3$, ..., $\bar{Q}_i$, in the set B of objects under study, such that $\bar{L}$ H_e $\bar{Q}_1$ and $\bar{Q}_1$ H_e $\bar{Q}_2$ and $\bar{Q}_2$ H_e $\bar{Q}_3$ and ... and $\bar{Q}_i$ H_e $\bar{M}$.

The equivalence relation E_e can now be defined: $\bar{L}$ E_e $\bar{M}$ if and only if there exists a H_e-chain from $\bar{L}$ to $\bar{M}$. This means that two objects $\bar{L}$ and $\bar{M}$ will be in relation with each other (will be in the same cluster) if there exists a connection between them, connection that can be established through other intermediate objects. E_e is an equivalence relation since it has the following properties:

– it is clearly reflexive: $\bar{L}$ E_e $\bar{L}$ since $\bar{L}$ H_e $\bar{L}$ always;

– since H_e is symmetric, the H_e-chains can be turned around, and then the existence of a chain from $\bar{L}$ to $\bar{M}$ implies the existence of a chain from $\bar{M}$ to $\bar{L}$;

– if a H_e-chain exists from $\bar{L}$ to $\bar{M}$ and another exists from $\bar{M}$ to $\bar{Q}$, then these two chains can be combined; consequently, $\bar{L}$ E_e $\bar{M}$ and $\bar{M}$ E_e $\bar{Q}$ imply that $\bar{L}$ E_e $\bar{Q}$, or in other words the relation is transitive.

The result of this process is to group the objects in various clusters. Consequently the relation H_e defines a partition of the objects. This is just what we expect that a classification into taxa of a given category will do.

Technically, a cluster is a connected subgraph: the relation H_e together with the set B of objects under study is called a graph, and a subgraph of (B, H_e) is a subset of B, called C, together with the relation H_e restricted to the objects included in C. A subgraph (C, H_e) is connected if there exists a H_e-chain from $\bar{L}$ to $\bar{M}$ for all pairs of objects $\bar{L}$ and $\bar{M}$ which are elements of C.

3.2. Definition of the species category

We have chosen to define the species category first because the categories just above and below it are defined relatively to it: genera are groups of species with special attributes, and subspecies are divisions of species also with special attributes.

Although many practicing taxonomists still use a typological or nominalistic species concept, the so-called biological species concept seems to be universally accepted among those who write about the theory of systematics, as it is defined by Mayr (1942, 1970) and Löve (1964a, 1964b). The biological species concept has at least two advantages over the morphological one. First, the biological species corresponds to an evolutionary unit, since it marks the point of differentiation where phyletically related lines become isolated from one another. Secondly, and as pointed out by Lehman (1971), the biological species concept is the only one of all the species concepts proposed, to include in its definition an objective criterion for the delimitation of species, contrarily to the typological or nominalistic concepts, which leave the delimitation of species to the arbitrariness of the taxonomist.

In the following definition of the species category, we will consider as

true the basic dogma of biosystematics, discovered by Müntzing (1930) and re-emphasized by many authors, like Löve and Löve (1967), who say that the process which isolates gene pools from each other, that is, the process leading to speciation, consists primarily of the chromosomal differentiation of populations.

In the process of modeling the species category, we will find a great help in the tools that we have defined above: the H-chain, and the feasibility of pairing $P_f(\bar{L}, \bar{M})$ defined by the threshold T (fig. 1).

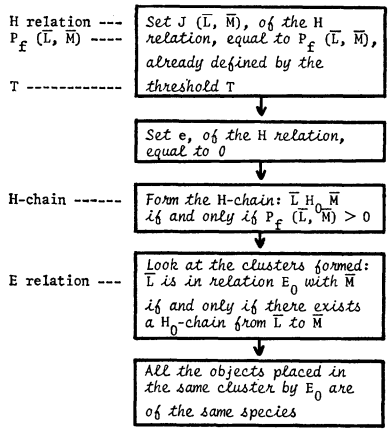

FIG. 1. Diagram showing the theoretical pathway to follow for defining species.

The operation consists of applying the pairing feasibility $P_f(\bar{L}, \bar{M})$ to all pairs of objects under consideration. As discussed in section 2.2.4., this measure has essentially the property of changing its sign when the pairing difficulty becomes larger than the threshold value T, which has been defined as a marker on the scale of chromosomal differentiation after which the hybrids, if produced, are sterile.

Graphically, if we had lines connecting all pairs of objects under study, we could represent the values of $P_f(\bar{L}, \bar{M})$ as modifiers of these lines, by making the thickness of the lines proportional to $P_f(\bar{L}, \bar{M})$. In these conditions, if $P_f(\bar{L}, \bar{M})$ is zero or negative, then the line disappears, and the groups of objects that still form connected subgraphs are our species. But this can be represented formally using the H-chain definition. We simply have to define that

- the general similarity relation $J(\bar{L}, \bar{M})$ takes in this case the values of $P_f(\bar{L}, \bar{M})$, or $J(\bar{L}, \bar{M}) = P_f(\bar{L}, \bar{M})$
- since the value of the threshold T has already been taken into account in the formula of $P_f(\bar{L}, \bar{M})$, then the value that we need for e is 0.

Consequently, the H-chain will be defined as follows:

$\bar{L}\ H_0\ \bar{M}$ if and only if $P_f(\bar{L}, \bar{M}) > 0$.

We can now define what a species is: all the objects which are placed in the same cluster by the equivalence relation E_0 are of the same species.

3.3. Definition of the genus category

There seem to be two main tendencies with regard to the concept of the genus. The definition given by Mayr (1969: 403) represents well the most common opinion: "Genus. A category for a taxon including one species or a group of species, presumably of common phylogenetic origin, which is separated from related similar units (genera) by a decided gap, the size of the gap being in inverse ratio to the size of the unit (genus)".

The first problem with this model lies in the fact that the knowledge about the common phylogenetic origin is based upon cladistic, unverifiable assumptions. The second one is that the determination of the size of the genus is left completely to the worker. It is true that in many instances in the past, the intuition of the experienced taxonomist has successfully accounted for the non-existence of recognized objective criteria. But in many cases also, the results have been modified by biosystematists working with the theory that supplied an objective criterion capable of solving these two problems. This criterion, which has been known for a very long time and has been re-emphasized by the biosystematists, is that "whereas hybridization is possible between species of a genus, hybridization between genera should be excluded" (Löve, 1963: 45). This criterion is not cladistic. It establishes the closeness of the gene pools by looking at their amount of differentiation, mainly chromosomal, that is a much better measure than the cladistic approach (Cronquist, 1969: 179; Legendre, 1971a). Secondly, it supplies an objective criterion for the establishment of the boundaries of the genus, as we will see.

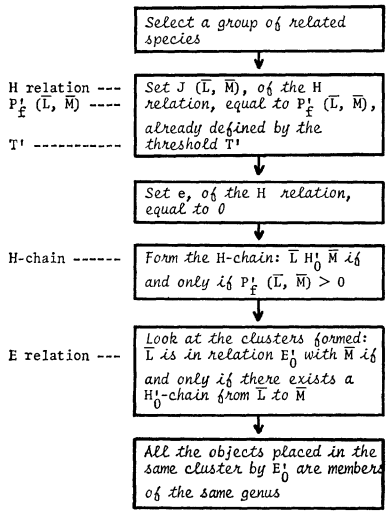

FIG. 2. Diagram showing the theoretical pathway to follow for defining genera.

For the definition of the genus category (fig. 2), we will use a measure of pairing feasibility which we have not defined yet, because there are two possibilities: one can build the genus either from individuals, or from species. If one uses individuals, the only modification that has to be done on the P_f ($\bar{L}$, $\bar{M}$) equation is to replace the threshold value T by T′, which delimits the amount of chromosomal differentiation after which the for-

mation of hybrids, even sterile, becomes impossible. The formula then becomes:

$$P'_f(\bar{L}, \bar{M}) = \max_{\substack{l \varepsilon L \\ m \varepsilon M}} (T' - P_d(l, m) \times S(\bar{L}, \bar{M}) \times G(\bar{L}) \times G(\bar{M})$$

When one prefers to use species, or species-representatives, instead of individuals, the terms $S(\bar{L}, \bar{M})$, $G(\bar{L})$ and $G(\bar{M})$ loose their meaning, and the formula to use is

$$P'_f(\bar{L}, \bar{M}) = \max_{\substack{l \varepsilon L \\ m \varepsilon M}} (T' - P_d(l, m))$$

The $P'_f(\bar{L}, \bar{M})$ values can be used in the same way as the $P_f(\bar{L}, \bar{M})$ values in the case of the species: either graphically, as modifiers of the thickness of the connecting lines between individuals or species, or as the basis for the definition of a chain, that we will call H', by defining the equality

$$J(\bar{L}, \bar{M}) = P'_f(\bar{L}, \bar{M})$$

Here again, since the value of T' has already been taken into account in the equation establishing $P'_f(\bar{L}, \bar{M})$, the value of e is 0, and the H'-chain is defined as follows:

$$\bar{L}\ H'_0\ \bar{M} \text{ if and only if } P'_f(\bar{L}, \bar{M}) > 0.$$

The genus is then defined as follows: all the objects which are placed in the same cluster by the equivalence relation E'_0 are members of the same genus.

It is this generic criterion that was used by Legendre (1970: 1176) to suggest the transfer of the cyprinid fish *Semotilus margarita* to the genus *Phoxinus*: first, this species hybridizes with *Phoxinus* species, and secondly it has the same chromosome number as the *Phoxinus* species investigated, but not the same as the *Semotilus* species studied. The criterion supplied by the basic chromosome number has also been used by D. Löve and R. E. G. Pichi-Sermolli (personal communication) to divide the old fern genus *Dryopteris* into at least 9 genera, pertaining to 3 different families, each genus being characterized by its basic chromosome number, which shows that the old genus *Dryopteris* consisted of units (the new genera) between which hybridization was not possible. Many other such examples are given by Löve (1963: 47).

3.4. Definition of a local population

A local population of a given species of interbreeding organisms, also called deme (Mayr, 1970), is the potentially interbreeding group of organisms of this species that live in a given locality. Although it is not a systematic category, we would like to discuss it here because of its importance as the primary evolutionary unit.

Between local populations of a species, there may or may not be chromosomal and genic differences between the local gene pools (*sensu* Mayr, 1969: 403), although it is likely that at least genic differences would be found between them, if the local populations have been established for a long enough period of time. Consequently, we cannot use the measures of chromosomal and genic differences in the process of qualifying local populations.

The main characteristic of the local population is that it is a geographic unit. The membership is established strictly by the spatial proximity of the individuals, and the degree of proximity necessary is determined, in turn, by the distance of dispersal of most of the gametes (fig. 3). If we say "most of the gametes", it is because inter-populational exchanges of gametes often occur, but in amounts much smaller than the intra-populational exchanges.

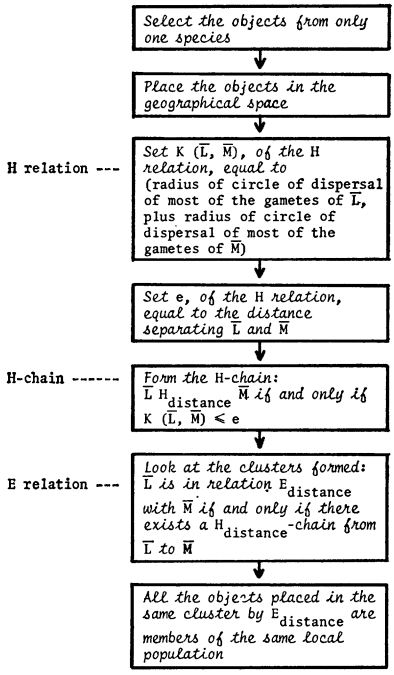

FIG. 3. Diagram showing the theoretical pathway to follow for defining local population.

Even if the location and the borders of the local population are mainly determined by the ecology, we still need to use only the geographical space in this model. This space is formed (depending on the characteristics of the species) of the two or three geographical axes that give a location of each individual in space. This determination should also be done at a given moment in time, and preferably at the beginning of the reproduction period in the case of mobile individuals.

We need now to define the H-chain on the individuals, so as to obtain clusters that will be equated to the local populations. The H-chain will be defined with the help of the second notation of section 3.1., that is, with the distance function K (L̄, M̄), as follows:

L̄ $H_{distance}$ M̄ if and only if the radius of the circle of dispersal of 95% of the gametes of L̄, plus the radius of the circle of dispersal of 95% of the gametes of M̄ ⩽ distance separating L̄ and M̄.

The 95% figure, that we have chosen here because it corresponds approximately to the number of objects within two standard deviations on each

side of the mean of a normal distribution, can be modified in each specific case, although its magnitude should be respected.

The local population, or deme, is then defined as follows: all the objects which are members of each connected subgraph (C, $H_{distance}$), that is, all the objects which are placed in the same cluster by the equivalence relation $E_{distance}$, are members of the same deme, or local population.

3.5. Definition of the subspecific category

The task of defining the subspecies is not as simple, theoretically at least, as what we accomplished in the preceding sections. A subspecies is defined as a group of local populations with a distribution which is a sub-division of that of the species, and with a gene pool which differs from that of the other subspecies. We will not discuss here the so-called polytopic subspecies, which, as far as we are concerned, is too artificial a unit to deserve systematic recognition.

We can notice first that a subspecies is not essentially an incipient species, since the subspecific recognition does not imply any value judgement about reproductive isolation: indeed, two subspecies of a given species are usually completely identical, from the chromosomal point of view. However, genic differences, which affect taxonomic differences, will always be present. The consequence is that we will have to use the index of genic similarity R ($\bar{L}$, $\bar{M}$), and not the pairing feasibility P_f (fig. 4).

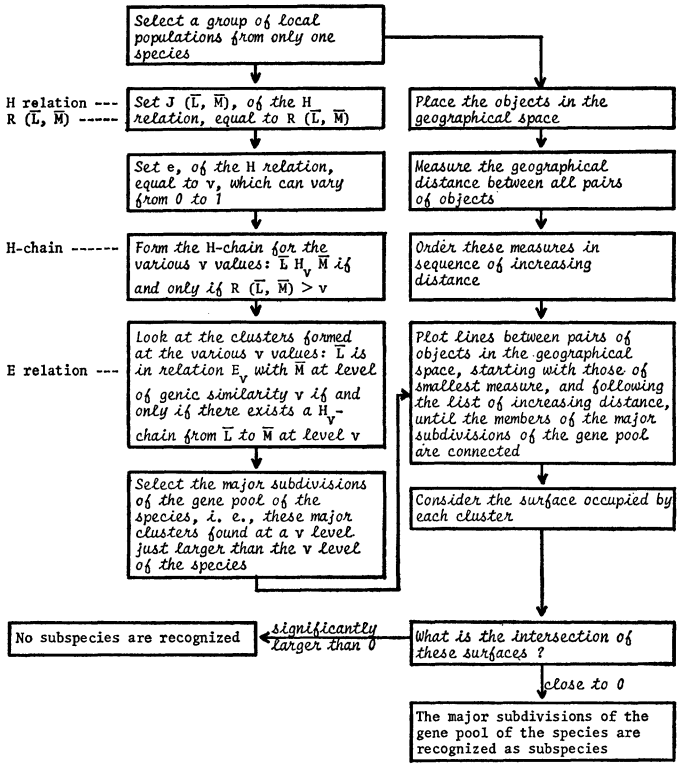

FIG. 4. Diagram showing the theoretical pathway to follow for defining subspecies.

We should also notice that, although the subspecies is defined as a group of populations on one side, and a subdivision of the species on the other side, which forces it as a category between the local population and the species, that we have already defined, there is still no unique, objective criterion for its definition. Indeed, the subspecies is strictly a comparative unit, since one will recognize a subspecies only if one can also recognize one or more other subspecies within the species under study. The consequence is that the subspecies cannot be defined by defining a threshold value on the scale of genic similarity R ($\bar{L}$, $\bar{M}$). Instead, subspecies will be recognized only in the case where the units showing the closeness of their members by their genic similarity correspond also to geographical units.

Thirdly, there may be several levels of subdivisions of the species that qualify, and in this case different authors may disagree on the level to choose. In these cases, the discussion is more often about the size of the geographic unit to choose, and one is usually told to define as subspecies the major geographic races, whereas the minor geographic races are called varieties, at least in botany. Here, we propose instead to consider the size of the gaps of genic differences that can be found, and to define as subspecies the major divisions of the gene pool of the species, if they qualify geographically. If not, we believe that no subspecies should be recognized. But of course, the following model can be adapted to satisfy the needs of those who would disagree with this premise.

We will first find the subdivisions of the gene pool of the species. In this respect, we will use mainly the graph theory model defined in Wirth *et al.* (1966) and in Estabrook (1966), with which a classification of the objects can be obtained, and from which we have already borrowed the idea of the H-chain.

The H-chain in this case will be defined by using the measure of genic similarity R ($\bar{L}$, $\bar{M}$), after stating that

$$J(\bar{L}, \bar{M}) = R(\bar{L}, \bar{M})$$

and by establishing no given threshold value. The partitioning capacity of the chain will be studied for various values v that can be taken by R ($\bar{L}$, $\bar{M}$), where v can vary from 0 to 1, which are the limit values that can be taken by R ($\bar{L}$, $\bar{M}$). The chain is defined as follows:

$\bar{L}\ H_v\ \bar{M}$ if and only if $R(\bar{L}, \bar{M}) > v$.

The clusters at any level of genic similarity v are formed by the usual equivalence relation E_v. The result of this process is to group the objects in various clusters at any given level of genic similarity. Consequently, the relation E_v defines, at any level v, a partition of the objects such that each object is in one and only one cluster. Furthermore, the partitioning process is hierarchical when considered along the axis of decreasing genic similarity, since the clusters will connect to each other as the similarity value drops. One could theoretically consider all the values taken by v between 1 and 0. However, in practice, it would be simpler to divide the range of v into, for instance, 100 values that can be considered one after the other or to consider only the values of v which correspond to a change in the membership of at least one of the clusters, as the authors mentioned above do.

At high similarity values, there are many small clusters. We may expect the first clusters formed to correspond rapidly to the local populations, or to small assemblages thereof. The clusters become fewer and larger as

the similarity value drops, finally attaining the point where there is only one cluster that includes all the objects of the species. Furthermore, pairs of objects that are in the same cluster at high similarity value remain inseparable for all the lower genic similarity values.

What we propose to consider as subspecies are the primary major subdivisions of the gene pool of the species, that is, these major clusters which exist at the v levels just before everything gets lumped into a single specific cluster — if, of course, these units correspond also to geographic subdivisions of the range of the species. In practice, the taxonomist obviously has to work with a limited set of characters, instead of the genic composition of the individuals. However, what we want to emphasize here is that a subspecies is a major subdivision of the gene pool of the species, so that instead of relying on the 75% rule applied to one character at a time, the taxonomist should rather use as many characters together as he can and try to get a 100% differentiation of the divisions of the specific gene pool. This could be done by using, for instance, the similarity measure of Estabrook and Rogers (1966), or a discriminant function analysis. One would then obtain a much better evaluation of the divisions of the gene pool of the species.

In order to find the geographic subdivisions in the distribution of the species, one may use a geographical sub-space of CD of the type that we used in the case of local populations. This time, however, one needs only the two dimensions known as longitude and latitude. By looking at the distribution of points in this space, one will see that they form groups, and these groups form larger groups, and so on.

We will not need to form an H-chain that would analyse the hierarchical geographical grouping of the individuals. All we need the distribution pattern for is to test if the major clusters found above correspond to geographic subdivisions of the distribution of the species.

It is not simple to map the distribution of the alleged subspecies. For instance, we cannot simply draw a circle or an ellipse around the points: the contour has to follow very closely the edges of the range, since it is possible to have two subspecies that will be very close to each other in certain parts of their range, even to the point where a local population of a subspecies would be situated between two marginal local populations of another subspecies. So, we tentatively propose the following operational procedure, that should be efficient enough according to the literature on the distribution of the continuous subspecies (by opposition to the polytopic subspecies, that we have decided not to consider). First, the geographical distance between all pairs of objects is calculated. Secondly, these measures are ordered in a sequence of increasing distance. Thirdly, lines are plotted between pairs of individuals, starting with those that are closest to each other, and following the list of increasing distance, until we get a connected subgraph for each alleged subspecies, at which point we stop drawing the connecting lines. In this process, the local populations should get connected first, then the larger sub-units, and so on up to the point where all the objects are connected in a single cluster. If, as we can expect, the demes are less densely packed to each other near the margin of the range of the alleged subspecies, then the periphery of this range should be very irregular in shape.

The network of connecting lines determines a surface occupied by each alleged subspecies. If the intersection of the surfaces so determined for

the alleged subspecies of a given species is nil — or very close to nil, to account for the rigidity of the operational drawing procedure —, then these alleged subspecies can be considered as subspecies, and given a trinomen.

This operational definition does not apply, however, to certain types of organisms, like in certain migratory species and in certain parasites with subspecies that occur in different, but sympatric hosts, as noted by Mayr (1969: 41). In both of these cases, it is the geographical overlapping test which has to be modified. In the first case, it should be enough to make this test at the breeding time. In the second case, the operational procedure should be based upon the knowledge of the life cycle of the organism.

This procedure for the determination of subspecies has been followed by Rogers and Appan (197), who first discovered the major divisions of the gene pools of species in the genus *Manihot*, using the original version of the graph theory model designed for phenotypic systematic studies (Estabrook, 1966), and then correlated these divisions with geography so as to establish what subspecies should be recognized.

3.6. Note on the semispecies

The term semispecies refers hereafter only to the incipient species, and not to the members of a superspecies, as in Mayr (1969: 53). It is not surprising that no categoric rank is attributed to the semispecies by the codes of botanical and zoological nomenclature, since many biologists still do not see the difference between the processes leading to speciation and those by which subspecies are formed, so that the subspecific category has been allowed to include also the semispecies, a fact that has resulted in much confusion as to the evolutionary importance of the former. When the difference between these two mechanisms will be more widely understood, the semispecies will most probably be seen as a useful systematic category. In systematic studies of freshwater fishes, for instance, the need for this category is obvious. A nomenclatural procedure could then be easily established for it.

A semispecies is here defined as a group of actually or potentially interbreeding populations, which are chromosomally somewhat distinct, but not effectively reproductively isolated from other such groups. This means that under experimental conditions at least, individuals from related semispecies can breed and form almost regular hybrids; they are gradual incipient species characterized by the same number of chromosomes. However, such groups of populations are isolated from other semispecies of the species complex by secondary isolation mechanisms, not controlled by genes in the case of geographical isolation, or controlled genically, for instance in the case of sympatric ecological, mechanical, ethological or seasonal isolation.

The essential difference between a semispecies and a subspecies is that the semispecies is a real incipient species, that is, it can eventually lead to the formation of a new species if the secondary isolation mechanism is maintained long enough for the chromosomes to differentiate sufficiently, whereas the subspecies, genically differentiated from other subspecies, is not and cannot be on its way to become a different species. As a consequence, all forms of secondary isolation, and not only geographical isolation as in the case of the subspecies, are effective as means of preserving and accumulating the acquired chromosomal distinctions.

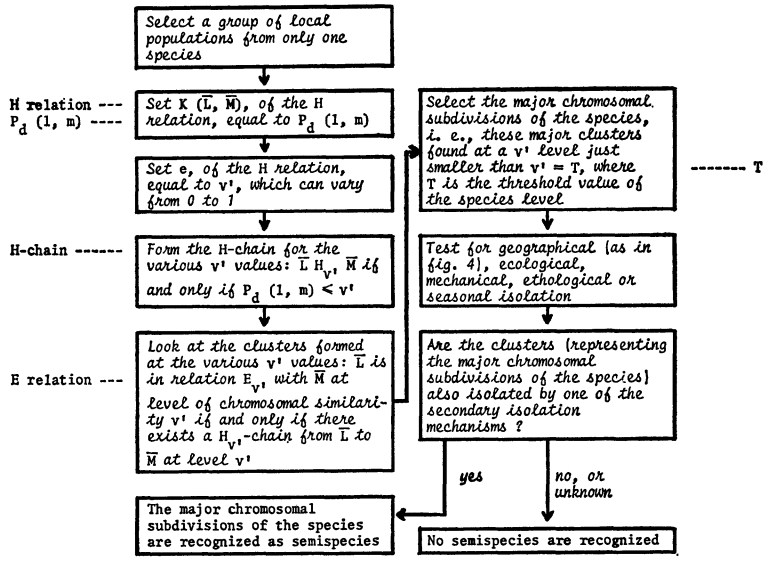

FIG. 5. Diagram showing the theoretical pathway to follow for defining semispecies.

However, the semispecies have in common with the subspecies the arbitrariness of their definitions, in border cases, so that a formal definition of the semispecies (fig. 5) would be even more hazardous than in the case of the subspecies. Indeed, one could first classify the populations according to their chromosomal similarity, with the help of the graph theory model defined in Estabrook (1966) and explained in section 3.5., so that divisions of the species into groups of chromosomally similar local populations would be found. Then, one would have to look for the clusters to which can be attributed any one of the secondary isolating mechanisms that can suffice to isolate this group from the other semispecies. And to achieve this formally, one would have to develop other mathematical tools than those of chapters 2 and 3 above.

Even without recognition of the semispecies as a systematic category, one should nevertheless be acquainted with the concept, since it clearly illustrates how gradual speciation occurs.

3.7. The categories above the generic level

We do not intend to discuss here in great detail the procedure to follow in the case of the categories above the level of genus, since no objective criterion has so far been proposed. But we may mention a few general ideas, after the discussion of this chapter.

— The categories above the genus are somewhat squeezed between the genus, formally defined, and the top of the classification. Even at the kingdom level, there are more than one opinion: many authors recognize two kingdoms, plants and animals; Dillon (1963) recognizes only one, that is called *Plantae* because of the rule of priority; at the other end of the scale, Grant (1963: 85) recognizes 5 smaller, more homogenous kingdoms, in which Margulis (1970) and Whittaker (1969) agree. But this is due mainly to our lack of knowledge about the early stages of evolution. The

farther down we get into the classification, the more advanced is our knowledge of the type of inter-group relationships. If one considers first only the main systematic categories, family, order, class, phylum and kingdom, then it should be possible, because of the squeezing effect that we just mentioned, on one hand, and because of the possibility to find the major adaptive groups, that we will discuss below, on the other hand, to obtain a classification into groups of corresponding importance.

— The resulting classification should not be cladistic. However, the internal logic of cladism should be applied to the dendrogram of informational similarity between the taxa to be classified (Legendre, 1971a) so as to point out the commonness of their adaptation. The units of common adaptation could be defined by logical operations, that could be called "last common adaptation" and the reciprocal, "set of objects with the given adaptation", referring here not to a given taxonomic character, but rather to a set of properties that make the individuals adapted in some general way.

— An ecological-adaptive space (Legendre and Vaillancourt, 1969 Hutchinson, 1957; Whitaker, 1972) becomes very important for the delimitation of the higher categories, since it provides an appropriate space for the clustering activity. It is awell-recognized fact that the higher categories correspond to broader and more differentiated adaptive zones (Simpson, 1953 *in* Mayr, 1970: 353).

— Consequently, the method for finding the higher categories would involve, first, the correlation of the adaptive zones with the units of phylogenetic adaptations, and secondly, a repartition of the natural units found on the range of the systematic categories available, trying first with the few, main categories, and then refining the definitions by addition of intermediate categories. Morphological characters have always brought and will always bring the first indications to consider when looking for higher taxa, since in most groups of organisms they bear indications as to the phylogenetic relations and the adaptive zone occupied. However, one should look for correlations of morphological patterns with other evidences, be they related to karyology, study of the adaptive zones, biochemistry, serology, embryology or anatomy, in order to get meaningfully uniform higher taxa.

Acknowledgments

The author wishes to thank Dr. Áskell Löve for his friendly interest and learned advice during this study, and Dr. David J. Rogers who also spent much time discussing various aspects of the problem. Dr. Doris Löve, Mrs. Elizabeth A. Kaersvang, Dr. Stanislaw Ulam, Dr. Jane Bock and Dr. Hobart M. Smith reviewed the manuscript and suggested many improvements. Personal financial support during this study was provided by the National Research Council of Canada.

References

CANDOLLE, A. P. DE 1813 — Théorie élémentaire de la botanique. Déterville, Paris. viii + 500 + 27 pages.

CASPERSSON, T., S. FARBER, G. E. FOLEY, J. KUDYNOWSKI, E. J. MODEST, E. SIMONSSON, U. WAGH and L. ZECH 1968 — Chemical differentiation along metaphase chromosomes. Exp. Cell Res. 49: 219-222.

CRONQUIST, A. 1969 — On the relationship between taxonomy and evolution. Taxon 18 (2): 177-187.

DANSER, B. H. 1929 – Ueber die Begriffe Komparium, Kommiskuum und Konvivium und ueber die Entstehungweise der Konvivien. Genetica 11: 399-450.
DARWIN, C. 1859 – On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London. 507 pages.
DILLON, L. S. 1963 – A reclassification of the major groups of organisms based upon comparative cytology. Systematic Zoology 12 (2): 71-82.
DOBZHANSKY, T. 1941 – Genetics and the origin of species. 2 nd ed. Columbia University Press, New York. 446 pages.
DU RIETZ, G. E. 1930 – The fundamental units of biological taxonomy. Svensk Bot. Tidskr. 24: 333-428.
ESTABROOK, G. F. 1966 – A mathematical model in graph theory for biological classification. J. Theoret. Biol. 12: 297-310.
ESTABROOK, G. F. 1970 – Appendix to: Theoretical and practical considerations on data structuring for a computerized information retrieval system, by D. J. Rogers. Pp. 154-157 *in*: Archéologie et calculateurs. Marseille, 7-12 avril 1969. Colloques internationaux du Centre National de la Recherche Scientifique. Editions du Centre National de la Recherche Scientifique, Paris.
ESTABROOK, G. F. and D. J. ROGERS 1966 – A general method of taxonomic description for a computed similarity measure. BioScience 16 (11): 789-793.
GRANT, V. 1963 – The origin of adaptations. Columbia University Press, New York. x + 606 pages.
HULTÉN, E. 1937 – Outline of the history of arctic and boreal biota during the Quaternary period. Bokförlags Aktiebolaget Thule, Stockholm. 168 pages.
HULTÉN, E. 1968 – Flora of Alaska and neighboring territories. Stanford University Press, Stanford. XXIII + 1008 pages.
HUTCHINSON, G. E. 1957 – Concluding remarks. Cold Spring Harbor Symp. Quant. Biol. 22: 415-427
JARVIK, E. 1960 – Théorie de l'évolution des vertébrés. Masson & Cie, Paris. 104 pages.
KATO, H. and T. H. YOSIDA 1972 – Banding patterns of chinese hamster chromosomes revealed by new techniques. Chromosoma 36: 272-280.
KIRPICZNIKOV, M. E. 1968 – On the concept of genus in flowering plants. Bot. Zhurnal 53: 190-202.
LEGENDRE, P. 1970 – The bearing of *Phoxinus* (Cyprinidae) hybridity on the classification of its North American species. Canadian Journal of Zoology 46 (6): 1167-1177.
LEGENDRE, P. 1971a – Circumscribing the concept of the genus. Taxon 20 (1): 137-139.
LEGENDRE, P. 1971b – Some formal aspects of the theory of biological evolution. Ph. D. dissertation, Department of Biology, University of Colorado, Boulder. viii + 98 pages.
LEGENDRE, P. and P. Vaillancourt 1969 – A mathematical model for the entities species and genus. Taxon 18 (3): 245-252.
LEHMAN, H. 1971 – Classification and explanation in biology. Taxon 20(2/3): 257-268.
LINNAEUS, C. 1735 – Systema naturae, sive regna tria naturae systematice proposita per classes, ordines, genera et species. Theodorus Hauk, Lugduni Batavorum. 13 pages.
LINNAEUS, C. 1737 – Critica botanica in qua nomina plantarum generica, specifica, et variantia examini subjiciuntur, selectiona confirmantur, indigna rejiciuntur, simulque doctrina circa denominationem plantarum traditur. Conradus Wishoff, Lugduni Batavorum. xiv + 270 pages.
LINNAEUS, C. 1751 – Philosophia botanica in qua explicantur fundamenta botanica cum definitionibus partium, exemplis terminorum, observationibus rariorum, adjectis figuris aeneis. Godofr. Kiesewetter, Stockholmiae. 362 pages.
LÖVE, Á. 1961 – Hylandra – a new genus of Cruciferae. Svensk Bot. Tidsk. 55 (1): 211-217.
LÖVE, Á. 1963 – Cytotaxonomy and generic delimitation. Regnum Vegetabile 27: 45-51.
LÖVE, Á. 1964a – The evolutionary framework of the biological species concept. Genetics Today, Proceedings of the XI International Congress of Genetics, pp. 409-415.
LÖVE, Á. 1964b – The biological species concept and its evolutionary structure. Taxon 13 (2): 33-45.
LÖVE, Á and D. LÖVE 1942 – Chromosome numbers of Scandinavian plant species. Bot. Notiser 1942: 19-59.

Löve, Á and D. Löve 1967 – Evolution and the Linnaean species. Univ. Babes Bolyai din Cluj Grăd. Bot. Contrib. Bot., pp. 203-210.
Margulis, L. 1970 – Origin of eukaryotic cells. Yale Univ. Press. New Haven. 349 pages.
Mayr, E. 1942 – Systematics and the origin of species. Columbia University Press, New York. 334 pages.
Mayr, E. 1969 – Principles of systematics zoology. McGraw-Hill, New York. xi + 428 pages.
Mayr, E. 1970 – Populations, species, and evolution. Harvard University Press, Cambridge. xv + 453 pages.
Müntzing, A. 1930 – Outlines to a genetic monograph of the genus *Galeopsis*. Hereditas, 13: 185-341.
Rajhathy, T. and J. W. Morrison 1959 – Chromosome morphology in the genus *Avena*. Canadian Journal of Botany 37: 331-337.
Ray, J. 1686 – Historia plantarum. Vol. 1. Mariae Clark, Londini. 983 pages.
Rogers, D. J. and S. G. Appan 197 – The North American species of *Manihot* delimited by computer aided taximetric methods. Colorado Associate Univ. Press (in press).
Seabright, M. 1972 – The use of proteolytic enzymes for the mapping of structural rearrangements in the chromosomes of man. Chromosoma 36: 204-210.
Simpson, G. G. 1953 – The major features of evolution. Columbia University Press, New York. 434 pages.
Sokal, R. R. and P. H. A. Sneath 1963 – Principles of numerical taxonomy. Freeman and Co., San Francisco. xvi + 359 pages.
Solbrig, O. T. 1970 – Principles and methods of plant biosystematics. Macmillan, New York. xiii + 226 pages.
Sumner, A. T., H. J. Evans and R. A. Buckland 1971 – New technique for distinguishing between human chromosomes. Nature (Lond.) New Biol. 232: 31-32.
Turesson, G. 1922 – The genotypic response of the plant species to the habitat. Hereditas 3: 211-350.
Whittaker, R. H. 1969 – New concepts of kingdoms of organisms. Science 163: 150-160.
Whittaker, R. H. 1972 – Evolution and measurement of species diversity. Taxon 21 (2/3): 213-251.
Wirth, M., G. F. Estabrook and D. J. Rogers 1966 – A graph theory model for systematic biology, with an example for the Oncidiinae (Orchidaceae). Systematic Zoology 15 (1): 59-69.