MODELS FOR GENETICS

Wolfgang Balzer

and

Chris M. Dawe

We have some copies left. If you write to the first author:
Prof. Dr. Wolfgang Balzer, Seminar PLW, Ludwigstr. 31,
D-80539 München we send you a copy without charge.

PREFACE

This book has grown out of the authors' joint work extending over ten years now. We began by formalising the picture of the structure and comparison of major branches of genetic theories as drawn in Dawe's dissertation (Dawe,1982). First results were published in two articles (Balzer & Dawe,1986a,b). There was extensive demand for preprints following this publication, which we interpreted as indicating significant interest among geneticists and other scientists in the foundations of this rapidly developing discipline. So we decided to go further, and work out a detailed manuscript. Since 1990, copies of our manuscript circulated among colleagues interested in the foundations of genetics. When we recognized that our work began to be taken up in other publications we thought it should officially be published.

The book we present here is a foundational work. We want to provide clear and precise conceptual models for the most important branches of genetics. The value of such work is in unification and simplification. By this alone, no new problems in practical genetics are solved. Neither do we analyse in detail any practical examples from the current frontiers of genetics. However, scientific value cannot be assessed merely by virtue of immediate applicability. History abounds with examples of important theoretical advances which were of little, if any, contemporary value. Foundational work takes some time to penetrate into scientific practise and teaching, but is no less essential for that.

One point of making precise models is to make the field accessible to computers. Our models are of this kind. Genetics already experiences substantial application of computers at the molecular level where the configurations of complex molecules are modelled, and first steps of AI methods are seen, for instance in the MOLGEN programme (Stefik,1981a,b). It is not difficult to transform the axioms which characterise our models into programs in high level computer languages like LISP and PROLOG. In fact, in (Dawe & Dawe, 1994), some first pieces of code are given, based on our models. Such programs may be taken as the kernel for comprehensive expert systems. By adding special features characteristic for special and possibly complicated applications, a somewhat intelligent tool becomes available. Indeed, at one time we had contemplated including substantial parts of programs, but this is a matter beyond our present scope, and one which we shall address on another occasion.

A second point of being precise is to make the field more accessible to beginners. The textbook tradition in genetics, beginning with (Sinnot & Dunn, 1925) focusses on the explanation of various important assumptions, principles or hypotheses, and their application to conrete, 'textbook' cases. This is an efficient way to further the novice's ability to produce certain standard solutions to standard problems. Textbooks are written this way in all disciplines. However, problem solving is just one important scientific activity, finding new hypotheses is another. For the latter it is necessary to have a firm command of the precise form of the primitives. So the book should be profitable for students of genetics

as a companion to 'ordinary' textbooks.

We should like to acknowledge the guidance given by C. Smith, P. Holgate, J. D. Sneed and the late A. Birnbaum at the earlier stages of this research. T. A. F. Kuipers and N. Roll-Hansen gave helpful comments. We thank Janet and Phillio for their patience and tolerance, especially during mutual visits when often we were occupied for several days with our models. We thank Phillio Marcou and Martin Dawe for providing the drawings.

München and Petham
January 1997

CONTENTS

Chapter 1

The Unity of Genetics

Genetics is both the study of the hereditary transmission of traits as well as the study of the underlying mechanisms responsible for those traits. As a matter of historical fact, the study of the hereditary transmission preceded that of the underlying mechanisms. In the 19th century, the first regularities were observed in the transmission of characters distinguished in plants. Mendel's famous experiments may be regarded as marking the origin of genetics as a science. The basic method of observing, counting, and systematizing the occurrences of various characters, and tracing them through various generations has been steadily developed and refined. It is now firmly established and recently gained new impetus from the studies of pedigrees. According to this approach the characters and their observed expressions are taken as given, no attempt is made to explain how a particular expression arises in the development of an organism. On the other hand there is a development from early embryology and 19th century chemistry to cytology which strongly depended on other technological achievements like the microscope, methods of organic chemistry and X-ray diffraction. In this line, it is difficult to fix a special date as the definite beginning of genetic thinking. The doctrine that higher life forms are constituted mainly of multitudes of cells dates back to Schleider and Schwann in 1835. It came from the use of light microscopes with greater magnification power. Lebedeff designed and built the first interference light microscope and Zermicke the phase contrast microscope in 1932. Such developments made it possible to make direct observation of mitosis and meiosis, rather than inference from static situations. Nonetheless, by 1879, Flemming had already indirectly observed the doubling of chromosome number approaching cell division, and the causal role of the chromosome in replication became evident thereafter. The idea that nucleic acids found to be present in the chromosome were the genetic material, was discounted by geneticists until the middle of this century however. This was because of the mistaken belief that DNA contained a simple repeating base sequence. A fuller understanding of the biochemistry of DNA thus provided a missing link to the chain. From there, a steady development can be stated which basically has two dimensions. In one dimension the internal structure of the genetic material was uncovered in ever greater detail. In the other dimension, the ways were studied in which this material gets transmitted during mitosis and meiosis, and in which it governs the processes in the cell. Research along the first dimension led to the famous model of DNA put forward by Watson and Crick. By bringing the structure of the genetic material down to

the level of molecules this model marks a 'final' level of detail at which research can differentiate for quite some time. Studies along the other dimension led to models of transcription and transmission of genetic material in the cell on the basis of the Watson Crick model, as well as to more macroscopic accounts of mitosis and meiosis.

According to this development genetics is usually thought of as having two main branches, and we see no reason to question this. There are transmission genetics and molecular genetics. However, stronger claims about the relation of these two branches are frequently discussed. Some philosophers have claimed that such studies of the underlying mechanisms have superseded those of hereditary transmission. For example, that molecular genetics can replace transmission genetics[1] or that molecular genetics will ultimately 'reduce' transmission genetics.[2] The vocabulary used in such discussions borrows from developments in physics, mainly from the transition of phenomenological macroscopic thermodynamics to statistical mechanics. Also, the notion of scientific revolutions as introduced by Thomas Kuhn[3] is sometimes brought into play, because the introduction of a new theory which is able to reduce the original one constitutes a decisive change, a 'revolution'. Such a revolution, if it had taken place in genetics, would be constituted by the introduction of the Watson Crick model which alledgedly marks the origin of molecular genetics ready to supersede the transmission branch. Though we do not feel informed enough about the historical events in the middle of this century to reject such a claim on the basis of historical material, using Kuhn's characterization of scientific revolutions we have strong reservations concerning the correctness or adequacy of that view. To mention just one point, nothing seems to indicate that the introduction of the Watson Crick model led to specific claims of supersedence among geneticists, or to the claim that transmission genetics is not entirely adequate and needs correction by the molecular account.

Concerning the claim of reduceability of transmission to molecular genetics the situation is less clear. The problem here is that such claims are not based on any rigorous explanation of what is intended by the two theories or approaches under discussion. Nor is the analysis of their relation carried out rigorously. Moreover, there is not one single definition of reduction on which we might agree. Rather, a whole family of such notions can be found in the literature.[4] If reduction is understood in such a strong sense that it makes the reduced theory redundant then we here also have to express reservations about the claim of reduceability. The point will be taken up in Chap.7 in more detail.

As opposed to claims of the kind considered we uphold the thesis of unity of genetics. If we do not want to fall into the same pit we have just made for 'reduction' and 'revolution' we have of course to say precisely what we mean by unity. We understand the unity of genetics as provided by three necessary conditions which may be discussed separately though they are obviously strongly

[1](Hull, 1974).
[2](Schaffner,1969a).
[3](Kuhn,1970).
[4]See, for instance, (Nickles,1973), (Schaffner,1967), (Sklar,1967), or (Sneed,1971).

related to each other. First, the field has to exemplify a smooth historical development. Scientific revolutions are accompanied by bitter, irrational argument and dispute among the rival groups, reduction in the strong sense mentioned implies a rather quick and substantial reorientation of research towards the new approach. In the absence of such occurrences the development may be called smooth. In genetics the historical development was smooth in this sense.

Second, unity is provided by common methods, methods that are applied across various different areas of the discipline. In the case of genetics this condition is satisfied, if understood in the right way. By a method we mean the activity leading from concrete observations to some theoretic result. It is not required (but also not excluded) that the activity be that of one single person nor is it required that only one kind of apparatus in the laboratory is involved. A method may consist in a mixture of various techniques involving different apparatus, and it may be performed piecewise by different teams specialised on different steps. Also, a method may consist of a complex sequence of 'partial' methods. In genetics we observe a certain interplay of techniques from the transmission- and the molecular branch. Many molecular applications presuppose a previous localization of the area 'where to work' on the chromosome by means of transmission methods. In this sense methods from transmission genetics enter into the molecular method. Conversely, molecular methods may be used to detect errors in transmission experiments which escape the methods of transmission genetics. If we regard such applications in a comprehensive way we have to admit that methods of one branch appear in the other and conversely. In the remainder of this chapter we will substantiate this view.

Before turning to the details, we have to state the third necessary condition for unity. This is structural identity on a basic level. By this we mean the following. Two branches of a discipline are structurally identical on a basic level, if they both employ one common model (or to say it differently: two structurally identical models) so that both branches differ only in the way in which they refine the basic model. Both notions involved here, that of a structure (and structural identity) and that of refinement will be substantiated in Chaps.2 and 4 to 7 which in this sense may be regarded as showing the unity of genetics.

We now turn to the interplay, and thus the unity, of transmission and molecular genetics from a methodological point of view. By transmission genetics, we understand that branch of genetics which studies the transmission of a trait through two or more generations. Studies are essentially probabilistic, and concern the proportion of a population exhibiting a specified trait. As such, although a knowledge of the underlying mechanism giving rise to that character may also be studied, this is not an essential part of transmission genetics. Although direct studies of chromosomal structure or molecular structure are not possible in transmission genetics, insights into that structure can be obtained. Thus, linkage maps can be obtained through comparisons of progeny with differing assortments of characters. These indicate a linear ordering of the hypothetical factors deduced to be related to the appearance of specific character differences.

The other main branch of genetics is usually held to be molecular genetics.

This is often narrowly interpreted as the Watson Crick model associated with DNA and RNA. The original theory has been expanded greatly over the years, but with few modifications of the basic account. Many would find it difficult to see molecular genetics as a proper theory, but render it as 'empirical fact'. Putting to one side just what is intended by 'empirical', this view neglects the fact that although almost all of the early work in molecular genetics was carried out with haploid organisms, it can hardly be claimed that molecular genetics is limited to haploids. What happens in the study of non-haploid organisms is that there is considerably reduced certainty about the nature of the progeny. This is due to crossing over, translocation, transduction, and other effects on the genetic material. Many of these effects are observed in the study of entire chromosomes, and in this respect studies of the chromosome are very closely related to the studies of the genetic material of which they are in part composed. We shall include such chromosomal studies under the heading of molecular genetics. Indeed, we were tempted to include molecular genetics as a part of chromosomal genetics, which would be more logical, but avoided this on account of the preponderance of the term 'molecular genetics'.

We will now explain in more detail what we mean by transmission genetics and molecular genetics. Transmission genetics has two areas; Mendelian and linkage genetics. In both cases, hereditary character differences are assigned hypothetical factors, with which they are associated, sometimes in a quite complex manner. This relation would only be one-one in a haploid organism, which case would be trivial for transmission genetics. In the diploid, a typical example is of dominant-recessive antagonism. Given that two factors A and a have been related to a character difference, it is found that while A and a gives one expression of that character, and while A and A gives the same expression, a and a does not.

In Mendelian inheritance, factors related to the same character segregate independently in passing from one generation to next. Thus, given the mating of a parent with factor content AA with a parent with factor content aa, the progenal contents AA, Aa, Aa, aa, are all equally likely to appear. Similarly, factors related to different characters assort independently. Thus, given the parental factor contents $AABB$ and $aabb$, the progenal contents $AABB, AABb, ..., aabb$, are all equally likely to occur. Since the relation between factor content and character is generally not one-one, this does not mean that all of the different phenotypes will be equally likely to appear.

In linkage genetics, although factors related to the same character still segregate independently, factors for different characters do not assort independently. Instead, there is a degree of 'linkage' between them. Thus, given the factor contents $AABBCCDD$ and $aabbccdd$ in parents, it might be found that the combinations in which BC is not replaced by Bc or bC, appear more frequently than would have been expected from independent assortment. The complexities of relating the phenotype to factor content remain, and present the geneticist with challenging problems, which may require careful control of the parental factor content in mating experiments where these are possible. In the study of human transmission, the pedigree may require careful study.

The linkage observed in linkage genetics does not occur in a haphazard manner. Indeed, it is possible to order factors on a linkage map. This is constructed by studying the proportions of progeny in which there has been interchange of factors between those of parents, or 'crossing over' (see Chap.3). The higher the level of crossing over, the more distantly are the factors placed on the map. Some twenty years after this realisation, it became possible to identify chromosomal features and to relate changes in these to the linkage map. Two points are important here. First, that the linkage map could provide an ordering to the genetic material, without the necessity for direct observation. Second, that the geometrical map distances do not correspond to the linkage map distances, even if the ordering does. We feel it is important not to blur the way in which transmission genetics operates with the process of direct observation. This is especially important, since the term 'factor' and the term 'locus' are often used interchangeably with the 'gene'. We wish to keep Mendel's original term 'factor' for the hypothetical and probabilistic entity of transmission genetics.

In some respects, molecular genetics can be seen as a refinement of the study of the genetic material of the chromosome. The tools employed and the conceptual apparatus, are, however, those of the biochemist. Furthermore, molecular genetics provides a mechanism for the replication and transcription of genetic material. Nonetheless, there are aspects of genetics which molecular genetics does not yet appear to be active upon. We refer to the mixing of parental genes, accepted as important and studied for a century by the transmission geneticist. Although molecular genetics provides some degree of certainty in predicting the progeny of a haploid, this is removed for non-haploid cases. This is because the interchange of genetic material from the two parents can not yet be predicted. In the case of closely neighbouring loci, even transmission studies would not help, since the numbers of matings required for such probabilities of crossing over would be too large except for species which breed extremely rapidly. Neither does it appear that the crossing over process is random. Indeed it may be under genetic control. An understanding of the biochemical control of crossing over would be of great value. For example, the transmission of genetic disorders due to non-allelic recessive genes might be controlled.

One of the earliest applications of transmission genetics was by Mendel in his studies of pea colour.[5] In fact, he discovered that grey pea colour was dominant to white. That is to say that if we use G to indicate the presence of the factor for grey, and w to indicate the factor for white, that either GG or Gw would give grey seed colour, while ww would give white. The unknown factor content of parents could be determined from the proportion of grey and white progeny, as in the following example.

Let a grey seeded plant fertilise a grey seeded plant. Suppose that 118 progeny are grey seeded while 39 are white seeded. Consider the possible combinations of parental factor contents (after Strickberger 1985).

[5](Mendel,1901).

GG x GG =	all progeny GG with appearance grey
ww x ww =	all progeny with appearance white
...	...
Gw x Gw =	GG, Gw, wG, ww.

Gw, wG and GG all give grey appearances, ww gives white appearance. Because of independent assortment, in this case 3/4 will be grey, while 1/4 will be white.

Thus Gw is the factor content of both parents, in particular both parents are heterozygous.

It is seen from this example that the observed proportions of 39/158 and 118/158 only approximate the expected probabilities. Indeed it would be extremely improbable that they would exactly equal it. As a matter of interest, Mendel's own results were improbably good![6]

Here is another application of Mendelian genetics (after Strickberger, 1985). This time it is to silky feathers in fowl, the factor for which is recessive to that for normal feathers. 98 birds were raised from a cross between individuals that were heterozygous for this factor. The number which were silky and the number which were normal can be calculated as follows. Taking + to signify normal, and s to signify the silky factor, we obtain:

$$+s \times +s = ++, +s, s+, ss.$$

Now, $+s, s+$ and $++$ all give the normal phenotype, while ss gives silky feathers. Thus 3/4 will be normal and 1/4 will have silky feathers. That is to say that 24 will apear silky and the rest normal.

For both of the above applications, the phenotypes may be described as those of gross characteristics. However, Mendelian genetics does not only involve such gross characters. In particular, there may be some advantage in describing the characters in biochemical terms. Conversely, the use of biochemical terminology does not imply that molecular genetics is now involved. The character differences for Mendelian genetics may be microbiological, electrophoretic, numerical or other, provided the choice enables unique identification of the individuals which carry that trait. The human disease phenylketonuria is biochemically characterised by an alteration of the body's metabolism for phenylalanine. As a result, pyruvic acid is excreted in the urine. Clinically, the disease shows a fairly well defined symptomology, although in cases of doubt, biochemical evidence is taken as confirmatory. Garrod[7] postulated that hereditary distribution of the disease might be explicable by Mendelian genetics. Jervis[8] collected information on over 20 000 patients and their relatives concerning their physiological state and the results of ferric chloride and 2,4 dinitrophenylhydrazine tests for the presence of phenylpyruvic acid in the urine. A number of possible genetic hypotheses existed, and Jervis suggested that the gene was recessive autosomal (non sex-linked). Given the ratio of affected children to normal children,

[6] See (Edwards, 1986).
[7] (Garrod, 1902, 1909).
[8] (Jerwis, 1954).

this suggestion was upheld. In such an application to a human disease, two complications occurred. First, a number of heterozygous parents would escape detection, since their families did not contain any actual cases. Second, some families will have a greater than random proportion of affected children. Note that in this application, there is a choice between the way in which the character difference can be described, whether in terms of the pathology of the disease or through the results of biochemical tests. It may well be that a disease which has a poorly defined symptomology can lead to erroneous genetic analysis until a suitably precise method of diagnosis appears. We shall see later that there is at least one paradigm case of this occurring, namely in the disease called sickle cell anaemia.

Turning to Mendelian independent assortment, a further case in which the character difference may be described as either biochemical or gross will be given. DeVries[9] and Wheldale[10] suggested that the inheritance of corolla colour in antirrhinum majus could be explained in terms of the following factors:

Y: yellow lips/ivory tube	I: ivory lips
L: magenta lips	T: magenta tube

y, i, l, t can be used to denote the absence of these factors. This hypothesis was upheld. Later Onslow[11] and Lawrence[12] found that Y produces a yellow flavone luteolin in the lips and the less oxidised ivory apigenin in the tube of the flower. I suppresses luteolin, apigenin being formed throughout. If Y is present, 6L produces a tinge of red anthocyanin in the lips (delilah) and this is extended to the tube when T is also present. Again, whether the character differences are expressed in terms of colours, or in terms of biochemicals makes little difference to the application of Mendelian genetics.

Much of the early work on linkage genetics concerned Drosophila. Most of the important character differences found for Drosophila had been located on a linkage map by 1920. We will consider the map for the X- chromosome of Drosophila Melanogaster as provided by Morgan[13] and Dobzhansky.[14] In particular, we shall deal with three of the twelve factors they discussed, namely the mutations 'yellow body colour', 'white eye', and 'forked'. The normal characters associated with these mutations are wild type body colour, red eye colour and not forked respectively. Experimental matings of individuals having all three mutations with normal individuals gave a fraction of 875/81299 progeny with yellow body colour, white eye, but not forked and a fraction of 1676/3664 with wild type body colour, but white eye and forked. In this way, the probability of 'crossing over' occurring between white/red eye and forked/not forked was seen to be 0.011. The probability of crossing over occurring between wild type/yellow

[9] (DeVries, 1900).
[10] (Wheldale, 1907).
[11] See (Wheldale, 1907).
[12] (Lawrence, 1950).
[13] (Morgan, 1916).
[14] (Dobzhansky, 1932).

body colour and white/red eye colour was 0.457. These probabilities require some modification before producing linkage map values. This involves multiplication by 100 and correcting for the possibility that double crossing over had occurred. It is thus established that, if we take the map position of forked/not forked as zero, red/white eye is at 1.1 units, while yellow/wild type body is at 56.8 map units.

The cytological map of the X-chromosome of D.Melanogaster agrees with the ordering of the linkage map. However, the loci when relativised are now at 0, 33.0, and 56.8. The reason for this discrepancy has been sought in the structure of the chromosome and its constituents. It has been considered from the position of the variation of stiffness of the chromosome as distance from the centromere increases. More recently, it has been argued that crossing over is under genetic, indeed evolutionary control. The molecular mechanism is still sought. One feature of our analysis is that the models produced in the following can encompass the entire spectrum of such research.

It was mentioned earlier that the relation between a character difference and its factors may be quite complex, although until now only cases of antagonistic dominance and recessiveness have been considered. Thus, in partial dominance, the heterozygote has a character difference which is intermediate between the dominant and the recessive trait. An application is seen in the experiments of Rasmussen on the flowering time of peas.[15] In over-dominance, the heterozygote produces a character difference which is in excess of the dominant one. Over-dominance is usually seen in features which affect biological fitness, such as size, productivity and viability. In co-dominance, each factor can be thought of as contributing to the final character difference. Only in the heterozygote is the full trait realised. An example of co-dominance is found in the blue Andalusian fowl. In this the blue colour is due to a fine mosaic of black and white areas, the blue colouration is only seen if factors for black and white are present. Selecting the correct relation between factors and their related character differences is a major difficulty with transmission genetics. Indeed, in the case of the fowl mentioned the appearance was at first thought to be due to incomplete dominance between black and white colour factors. Closer observation of the trait clarified matters. Unfortunately, factors at different positions of the linkage map frequently affect the same character, and indeed one factor may control whether another can be related to a character difference, or be 'epistatic' to it. A further problem involves whether the effect of a factor can 'penetrate' and be observed. Thus, in Huntingdon's chorea, the effects are not evident until later in life, even though the gene is present from birth. A further complication is that of 'multiple allelism'. In this a number of factors may occupy the same position on the linkage map. Only two of this allelic series are actually involved in any case. Such multiple allelism is important in quantitative inheritance.

It may well be agreed, considering the enormous complexity of transmission genetics, that a simpler theory would fall on fertile ground. Early accounts of molecular genetics may have shown such a promise. However, molecular

[15](Rasmussen, 1935).

genetics is as complex in at least two respects. First, by necessitating that investigations be carried out at the molecular level, astronomical quantities of data must be handled. Second, although much is being learned about the process of transcription and of replication in the haploid, little is known about the control of diploid replication.

As an application of molecular genetics, we shall consider the A polypeptide chain of the tryptophan synthetase enzyme of E.Coli. This chain, approximately 280 amino acids long, shows a large number of mutations whose genetic map has been determined in great detail. Corresponding to this map is a known amino acid sequence of protein A, in which single amino acid changes can be assigned to specific mutations.[16] There is redundancy in the genetic code, in that more than one nucleotide triplet can give rise to a specific amino acid.

For example, adenine, thymine and adenine or adenine, thymine and guanine can both give rise as codon 174 to tyrosine. Applications such as this, however, tend to give the impression that the genotype and phenotype of progeny can be given with certainty. As has been mentioned previously, such examples are to haploid organisms and avoid problematic mixing of parental loci. Also, the vast numbers of loci present in studies of even simple organisms make it necessary for some form of focussing to occur prior to use of molecular genetics. Traditionally, transmission genetics has suggested a certain region of the chromosome and the DNA as being important for the study of a trait, and molecular genetics has then concentrated on this.

Thus, molecular and transmission genetics do not operate as independent theories, nor does it appear that molecular genetics is within sight of replacing or taking over transmission genetics. At the same time, the techniques used by the molecular geneticist are generally quite different from those of the transmission geneticist. Accordingly, the type of resources required will be different, and so will the practical expertise required. These factors, coupled with the fact that transmission and molecular genetics have their roots in different disciplines, have led to an institutional schism. We hope, that one of the benefits of our study of the structure and dynamics of genetics will be some indication of the way in which this schism might be healed. This in turn may lead to more efficient and productive use of the resources available to both the molecular and the transmission geneticist.

An illustration of the interplay of molecular and transmission genetics is found in the study of sickle cell anaemia. This study also raises one or two other interesting matters.

In some individuals, red cells may undergo reversible alteration in shape when the partial pressure of oxygen changes in a cell. This change is from a normal cell shape to a 'sickle' cell shape. In fact, a more and a less severe form of disease is associated with this sickling. Sickle cell anaemia involves a network of deleterious effects including abdominal pain, increased breakdown of erythrocytes and renal effects. These in turn lead to compensatory haemopoiesis, anaemia, jaundice, HbF formation, failure to concentrate urine, and so on.

[16] A full listing of the loci and the associated amino acids is given in (Dawe, 1982).

Sickle cell trait is far milder and is characterised almost entirely by the appearance of cell sickling in the presence of reduced oxygen pressure.

Originally, the two forms of the disease were thought to be extremes of a single character.[17] It was postulated that the ability to sickle was due to a single dominant gene. The difference was thought to be due to a variation in expression between individuals. However, Neel and Beet[18] independently postulated a gene which, in heterozygous condition resulted in sickle cell trait but in homozygous condition resulted in sickle cell anaemia. One way of settling the issue was through the following observation: If the homozygous-heterozygous hypothesis is correct, both parents of any patient with sickle cell anaemia would sickle in reduced pressure of oxygen. A dominant gene with variable expression would only require one parent to show sickling under the same conditions. In fact, every parent of a child with sickle cell anaemia had blood cells which sickled.

Electrophoretic studies by Pauling[19] indicated that the cause of the disease was a molecular change in the structure of the haemoglobin molecule. In fact, the abnormality is in the beta chain. Fragmentation and sequential electrophoretic and chromatographic studies indicated that the difference was simply in the location of a single peptide.[20] This was a substitution of a valine for a glutamatic acid group in the sixth amino acid. As a result of this change, differences of solubility accompanied by markedly increased viscosity led to a less flexible erythrocyte of a sickle shape. The molecular genetics of the disease can now be appreciated, but it is also seen that this was achieved through the offices of transmission genetics. Transmission geneticists effectively told molecular geneticists where to look, but could not say exactly what would be found. Notice, however, that in this example, the evidence provided by transmission genetics does not involve large numbers of progeny. This is because of the difficulties involved in human populations, and the ethical reasons preventing experimental matings. In such cases, the study of pedigree can compensate for these difficulties to some degree. A further feature of the example is the part played by a refined characterisation of the disease in establishing the correct genetic hypothesis.

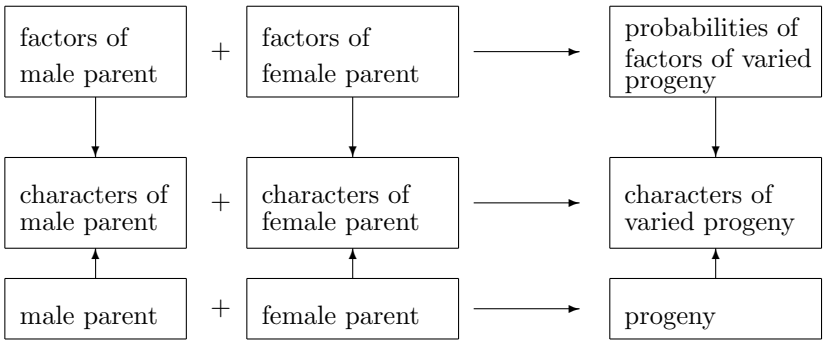
We now can abstract from what has been said so far. Transmission genetics concerns populations of individuals and matings between these. Associated with each population is a type, or set of character differences. In any given application only a few of these characters will be studied, sometimes only one pair. Associated with each character difference is at least one factor. Knowing the character does not, however enable the factors responsible to be known. A hypothesis is made and the results of mating experiments will hopefully validate the hypothesis. Alternatively, the results of mating experiments may enable a correct deduction of the factors involved to be made. The procedure is to mate two individuals and to look at the ratios of progeny with different types. The extent to which these ratios match up with the predictions from calculations on

[17](Taliaferro and Huck, 1923).
[18](Neel, 1949).
[19](Pauling, 1949).
[20](Ingram, 1957).

the basis of independent assortment or segregation then validates or invalidates the hypothesis. However, in the case of linkage, by comparison of crossing over frequencies associated with the factors of a character difference under study with established values, the map position of the new factor can be established. In practice it is extremely improbable that the theoretically expected probabilities would be obtained in an actual experiment. Statistical analysis can then be employed to establish whether the variation from expected values is acceptable as due to chance, or whether the original hypothesis might be wrong. When small numbers are involved, recourse to statistical analysis may be unreliable, as in the case of human traits. The wrong hypothesis may then go undetected.

Figure 1-1 shows a schema to illustrate the structure of transmission genetics.

Fig.1-1

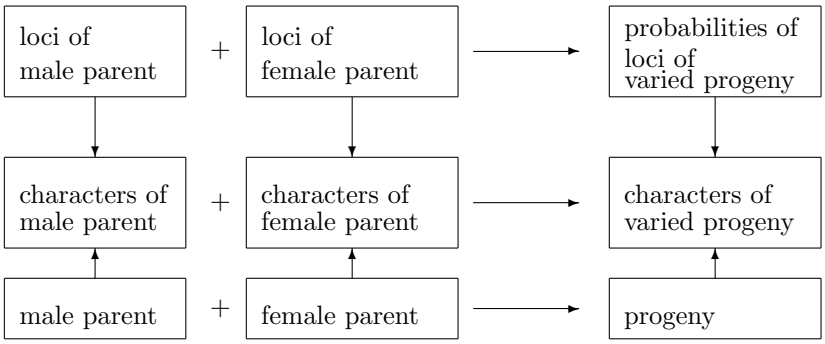


However, the situation has to be generalised to one in which the parents are themselves part of populations. A geneticist given the appearances of the male and female parents would like to know and to predict the appearances of the progeny and the proportions of progeny with different appearances. This cannot be achieved directly. Instead, a factor content is attributed to the male parent and a factor content is attributed to the female parent. Either by calculations based on independent assortment and segregation, or by reference to linkage maps, the probabilities associated with varied types of progeny can be established theoretically. This process constitutes the top line of the figure. To test the hypothesis, the cardinality of the male and female populations is established as are the cardinalities of the varied types of progeny. This is shown in the bottom line of the schema, and might be held to constitute the empirical content of the theory. Both the populations and the factor contents are connected to the traits (vertical arrows). The connection between a character and its factor content may be quite complicated, as has been mentioned previously. The relation between a population and its trait is usually quite straightforward, but may also become subject to error, as in the case of the andalusian fowl mentioned earlier, in which the plumage thought to be blue was actually a mosaic of white and black feathers. The cardinalities of populations of progeny are related back to their types, and these in turn are related to the factor contents. The proportions

of progeny with different types should match up with the expected probabilities to within accepted levels of statistical variation.

A similar schema can be drawn to illustrate the structure of molecular genetics, and this is shown in Figure 1-2.

Fig.1-2



As with transmission genetics, the geneticist given the appearance of two parents would like to predict the appearance of progeny, or the numbers of different types of progeny which would be produced. This is represented in the middle horizontal row. It is not possible to achieve this result directly, however. Instead, loci are associated with the appearances of the parents and progeny. As in transmission genetics, the relation between appearances and loci may become extremely complex. Given the dictionary provided by Watson and Crick, any amino acid character can be related securely back to a small number of nucleotide triplets, and in this respect, matters are far simpler. Relating such amino acids to their polypeptides and hence through to the gross structures frequently studied in transmission genetics removes much of this advantage, however. The process by which the loci of progeny are established from that of parents is also, at first sight, quite straightforward. It has been said previously that this is only the case if we restrict ourselves to the haploid cases and ignore the effects of crossing over, transduction, translocation etc. As soon as such more general situations are considered the loci of progeny can be calculated only with probability. The theoretical content of molecular genetics is established by the top line of the figure which pays attention to the probabilities of progeny's loci. However, unlike transmission genetics, molecular genetics is based on the mating of individuals rather than populations. The bottom line of the schema indicates the mating of two individuals to provide various progeny. In molecular genetics, the details of how the probabilities of different loci are achieved in the non-haploid and taking account of the effects mentioned previously are still to be worked out. Roughly, the observed proportions of individuals are related through their characters to the locus contents of varied progeny, as in the right hand column of the schema.

This schema can already improve our understanding of the development of

genetics. In general, a mapping is first noticed at the level of traits, that is the central horizontal row. The fact that there is similarity between certain features of progeny and their parents and ancestors is the cue to use a genetic theory. The numbers of progeny with such characters may then be observed in transmission genetics, and this is represented by the lowest row of Figure 1-1. A hypothesis concerning the factor content of the individuals is next made. At the same time, a hypothesis is made about the relation between these factors and the characters. Also, a hypothesis has to be made about the nature of the mapping on the top line of Figure 1-1, denoting the parental to progeny factor mapping. If all three hypotheses are correct, the probabilities predicted by theory will agree with the proportions of individuals with different types. Unfortunately, it is possible for two of the hypotheses to be wrong and still get agreement. This is because the effects of one wrong hypothesis may be countered by the effects of another wrong hypothesis. By varying the choice of characters in mating experiments, hopefully the mistake would be found. In a sense, merely repeating similar experiments would not be so helpful, but is the more immediate way of checking the choice of hypotheses.

There may thus be some doubt, even after the most careful application of transmission genetics as to whether the three hypotheses are correct. One of the features of transmission genetics, mentioned before is the linkage map. It has been mentioned that the ordering of factors on the linkage map is the same as that of loci on the chromosome, even though the relative distances between factors is not usually the same as that between loci. The co-ordering may provide added support for a hypothesis, or lead to its abandonment. If the nucleotide triplets responsible for a character can be identified the matter can be considered as finally settled. Molecular genetics thus adds certainty to the solution of a problem in genetics which was brought into near focus by transmission genetics. This passing of a problem from transmission to molecular genetics appears to be the normal way in which a problem in medical genetics is resolved, as was illustrated in the preceding discussion of sickle cell anaemia.

The schemata provided in Figures 1-1 and 1-2 only relate to a transition from one generation to the next. However there is in fact no difficulty in extending them for any number of generations. The precise mechanics of this iteration require some formalism, however, and will be returned to in Chap.6. Work on genetic algebras is especially significant in understanding the way in which these structures may be extended to problems of the n-th generation. Furthermore, 'peeling' and other processes can be applied for pedigree studies. Indeed, the interaction between evolution and molecular processes may be more readily understood once the unifying features of genetics are made explicit. We do not pursue this last issue here, however. In any case, our analysis goes way beyond the traditional questions posed by philosophers concerning the relation between molecular and transmission genetics. Not only can we throw light on that issue, but can discuss the overall structure of genetics with a breadth which would be of value both to scientist and philosopher alike.

Chapter 2

A Model for Genetics

The term 'model' is used in a variety of ways. In arts, we may speak of a model as the thing which is depicted. In this usage the model is the original. Second, a model may mean an image of some original, like a map or a toy-railway. Here, the model is not the original. A third usage, in particular in science, is that of a model as a hypothetical construct to be used in order to understand, or to deal with, or even to visualize, 'to depict' a given real system which is not entirely observable or understandable on its phenomenological 'surface'. We call these *conceptual models*. Clearly, the three kinds of models are related to each other. In all cases there is a relation between the model and some other entity, and the two have to be similar or homomorphic. In science we find many models in the second sense, models which function as an image, like 2-dimensional drawings of chemical formulae, and macro-models of molecules. Abstract, conceptual models with strong hypothetical features are not frequent.

In genetics, the basic attitude up to now is that of empirical research. Bold theorizing is neither necessary nor highly esteemed. However, with the degree of maturity achieved, knowledge becomes very complex, scattered, and specialized. Scientific interest in the metatheory of a discipline is most intense when there is some kind of crisis. It may well be that biology approaches an 'information crisis' in which the amount of empirical data generated becomes unintelligible without an understanding of the metatheory. In this way appropriate theories may be developed more readily. Conceptual models at this stage become useful in several ways.

First, they may provide a reduction of complexity. Various facts, or various different local images are subsumed under one conceptual model and thus integrated into a larger unit. Usually, the conceptual model will be described with only a small sample of all the concepts used in the field, the other remaining concepts being definable in terms of the 'primitives' used to describe the model. This does not mean that the primitives are really more central. They will be central of course but so will other, non-primitive concepts. The point here is that, in order to describe the model the number of concepts is reduced. Nevertheless, since the other concepts are definable in terms of those of the model, and therefore connections not made explicit in the model may be obtained by means of definition and derivation, nothing gets lost. It seems that this reductive potential of hypothetical models has not yet been fully appreciated within genetics.

Second, abstract models provide an easy survey and thus also a good basis

for comparisons. In particular, they make teaching much more efficient.

Third, they provide some guidance for sepcializing, for finding and evaluating various complex applications. The Watson Crick model, for instance, though still more of an image than of a conceptual model, induced a series of new experiments and applications.

Last but not least, today conceptual models serve as a basis for computer applications, for computer assisted search or calculation, but above all for expert systems and discovery programs. We witness today the first running computer programs intended to create new hypotheses obtaining good results[21] and we are already used to expert systems like MOLGEN in genetics.[22] Clear conceptual models greatly facilitate the construction of such programs.

In this book we aim at introducing several models for genetics which are all of the third kind listed above: abstract, conceptual models with hypothetical components. Further advantages for the unity of genetics or questions of comparison will become evident during the presentation.

In order to substantiate what has been said in the previous chapter we have to go into some detail. However, in contrast to the situation at the frontier of research we cannot afford to concentrate on one detail and leave 'the rest' for a while as though it was not important. Since we are talking about genetics as one unit, we have to keep sight of this unity.

The best way not to lose sight of some important feature is to incorporate it into an overall model. In fact, the main advantage of models and theoretical hypotheses is that they systematize and thus make easily comprehensible and storable large groups of facts and features. Models integrate large arrays of isolated data into intuitively comprehensible, 'simple' wholes.

In this chapter we set up a simple, precise model for genetic theories. The model covers transmission as well as molecular genetics and thus provides a strong argument for our thesis of unity. As with every model, concrete applications will require further special assumptions. This holds true for special applications within each branch of genetics, and those assumptions will differ in particular for applications of transmission genetics as contrasted to molecular genetics.

More precisely, this presupposes a particular view of how models are applied. The picture we have in mind here is this.[23] By an *intended system* we mean a concrete real system, in the laboratory or elsewhere clearly delineated from its surrounding, to which a geneticist draws his attention, and to which he intends to apply, or indeed applies, a genetic model or hypothesis. One or several hypotheses may be regarded as characterizing a set of possible models. Each model is an abstract, possible entity or system about which the hypotheses make sense and which, in addition, satisfies the hypotheses. So two kinds of systems are involved: intended systems representing 'reality' and abstract systems (models) representing hypotheses. To a first approximation a theory

[21] See (Langley, Simon et al., 1987), for example. On pp.274 they also discuss briefly the possibility of a programme finding Mendelian hypotheses.

[22] See (Stefik, 1981a,b).

[23] Compare, for instance, (Stegmueller, 1976).

T therefore may be seen as consisting of two components, a class M of models and a set I of intended systems:

$$T = \langle M, I \rangle.$$

It has to be stressed that the notion of a system does not imply a purely static view. There are 'dynamical systems' covering transitions, evolutions or other kinds of processes. It is not necessary here to spell out in which way the set of intended systems is determined. It certainly is not characterized in abstract terms, and also not in the form of a definite list. Rather, there is some amount of agreement in the geneticists' community that this and that system is intended for a particular theory, whereas others are not. By *application* we then mean a process in which it is attempted to 'subsume' some system intended for a theory under the models of that theory. Usually, in the course of this process three stages may be distinguished, at least conceptually (often they are intermingled in time). First, the real system has to be conceptualized in an appropriate way: the geneticist has to come to believe that all his theoretical concepts refer to entities or features of the system. Next, he has to make out those 'parts' of the system he really knows. This amounts to gathering observed data formulated in the theory's observational vocabulary. Finally, he tries to extend the data previously obtained by 'adding' further hypothetical or theoretical 'parts' or 'functions' so that the complete set of data plus the information contained in the theoretical functions satisfy the hypotheses. This process is successful if and only if, at the end, a model corresponding to the initial intended system is obtained.

The point about this picture is that in most cases the process of subsumption requires further assumptions which are not regarded as basic to the theory in question. These assumptions vary in status. Some of them may still express laws but laws of a scope more restricted than that of the theory's basic laws. Some may be idealizing assumptions about the absence of further relevant influences. Some are law-like *ad hoc* assumptions, some just concern the choice or determination of certain parameters. So the above picture of how models are used in the process of application has to be refined as follows. First, each theory has its corresponding class of basic models, that is, possible systems which satisfy all the theory's basic laws. Second, in the process of subsumption of a real case under a model, further assumptions are necessary which may vary from application to application. The special assumptions are not completely independent from the basic model. Rather, the basic model may be said to guide and to direct their choice. In this sense, the process of application -even though being analyzable into distinct subprocesses- forms a unit in which basic models and special assumptions are tightly bound together. To a better approximation therefore a theory has to be seen as a net of model classes plus some assignment of intended systems to each class such that there is one distinguished class of basic models from which all others can be obtained by some kind of specialization or refinement.

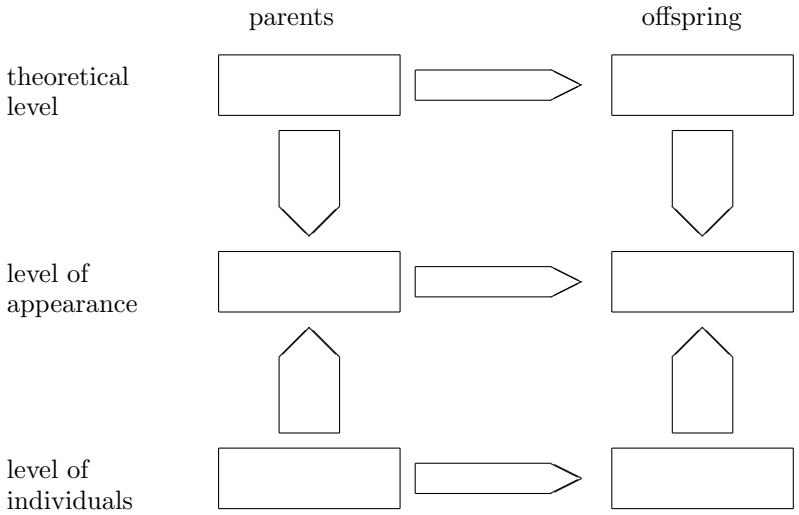
The genetic model to be presented here is the *basic model* or *core model* which may be used in all genetic applications. Its refinement by special laws

and other assumptions required in special, restricted ranges of application will be the subject of the following chapters.

Our model comprises two different levels: the level of *observational concepts*, and a *theoretical* level. The observational level is however further differentiated into a 'most basic' level of individuals, and a level of properties of these individuals. It will sometimes be convenient not to conflate the latter two. Orthogonal to this vertical dimension there is a horizontal dimension which covers the actual happenings in the course of time: mating and the production of offspring. In this dimension our basic model contains the distinction between parents and progeny, and their respective appearances, namely the phenotypes.

The items on the two sides of each distinction are related by operators. There is an operator assigning to each individual its 'properties', its 'appearance', 'structure' or 'function' and another one relating the theoretical level with that of such 'appearances'. Horizontally, at each level we have an operator relating the parental side with that of offspring. We therefore may depict the basic structure of the model by six boxes arranged in three layers with a 'before-after' distinction at each layer, and with arrows (operators) communicating between the boxes.

Fig.2-1



The distinction between boxes and arrows may be regarded as a coarse distinction between individual and relational concepts. One might say that each box represents an individual concept, each arrow a relational one. Next the content of the boxes has to be specified, and the meaning of the arrows to be explained.

We begin at the 'lowest' level, that of individuals. Here, a slight complication arises from the fact that, in transmission genetics, 'individuals' will be popula-

tions. However, nothing to be said in this chapter will become false if we extend the term to apply to proper individuals as well as to populations. For the sake of terminological uniqueness, we will use the term *genetic individual* whenever we want to be so abstract as to cover both meanings. The decisive feature of genetic individuals is that they mate, and thereby produce offspring. We have to distinguish parental genetic individuals (parents, parental populations) and those individuals produced by the parents (progeny). Since two parents may produce many kinds of offspring we have to enumerate. The parents are denoted by

PARENT_1, PARENT_2

and their offspring by

PROGENY_1,...,PROGENY_n

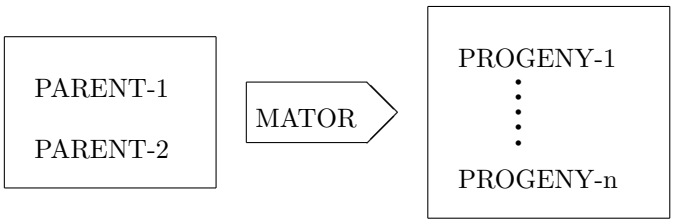
where n is the number of *different* genetic individuals occurring in the offspring. In the population case the difference between genetic individuals basically is decided in terms of phenotypes. It often is natural to take a population as a maximal set of individuals all of which have the same phenotype. This must not be regarded as a strict definition of populations, however. Usually, there are other, additional criteria to distinguish populations: from trivial ones like separation in space and/or time to subtle ones like reference to differences in the immune system. Only in Chap.6 will we use this 'definition' of populations, and only for reasons of simplicity.

The transition from the parents to their progeny is represented by a function we call MATOR. MATOR assigns to any two parents their progeny:

MATOR(PARENT_1, PARENT_2) = $\langle$ PROGENY_1,...,PROGENY_n$\rangle$

where the number n may vary with the parents. In other words, MATOR for any two parents specifies their offspring (which may be none, of course). The individual level in the model therefore has the form

Fig.2-2



For example, (Rasmussen, 1935) observed a five day difference in flowering time of peas between parents could give rise to progeny with early, intermediate, and

later flowering. Here, PARENT_1 and PARENT_2 are sets of pea plants, respectively, of different flowering time. To these MATOR assigns their progeny: ⟨ PROGENY_1,PROGENY_2,PROGENY_3⟩ where the PROGENY_i denote sets of pea plants flowering early, intermediate, and late, respectively. As another example, take PARENT_1 and PARENT_2 to be two human individuals whose red cells sickle in reduced pressure, and take, say, PROGENY_1,...,PROGENY_4 as four offspring of these. Each offspring, again, will have sickling red cells, but in this example they are not identified by this condition. Rather, the interpretation here is at the level of individuals proper.

Let us look at the 'middle' level of the model, that of phenotypes. Genetic individuals are distinguished by their appearance. In various applications just one trait and its different expressions, or several traits may be considered. In the case of populations, of course, every individual in the population has to exemplify the 'defining' expressions. A complete characterization of individuals in terms of their appearances will perhaps remain a limit case hardly ever achieved: Applications of genetics usually concentrate on one or few traits. By a *phenotype* we simply mean one or several expressions of these traits. Consequently, the term phenotype is not intended to refer to the complete appearance of a genetic individual. It is used as a technical term to denote one or more expressions of those traits which are relevant in a given application. Just as with individuals we have two parental phenotypes, denoted by

PHENOTYPE_1 and PHENOTYPE_2

and n phenotypes associated with the n different offspring:

PHENOTYPE_OF_PROGENY_1,...,PHENOTYPE_OF_PROGENY_n.

Phenotypes are 'directly observed'. Quantitative consideration of the distribution of phenotypes in offspring formed the starting point of genetics. Accordingly, at the level of phenotypes we have to introduce *distributions* of phenotypes. We use such distributions for the side of progeny only, the model can be easily extended however to include distributions on the parental side (compare Chap.6). Distributions of phenotypes are practically always given by means of relative frequencies. The total number n of offspring from two parents is counted, as well as, for each particular phenotype exemplified in the offspring, the number m_i of individuals of this phenotype in the offspring. $r_i = m_i/n$ then is the relative frequency of occurrence of that particular phenotype, and by collecting all those relative frequencies we obtain a distribution of phenotpyes.

Formally, a distribution is a function which to each element of a given set assigns a real number indicating the 'weight' or 'probability of occurrence' of that element. This notion is more narrow than that of a probability distribution, and also different from the notion of a distribution as used in quantum mechanics. For this reason we will speak of *genetic* distributions, or Γ-distributions for short. In the present case the elements of the set are the different phenotypes in the progeny, and their probability of occurring is approximated by the observed relative frequencies. Since there are only finitely many different phenotypes we may assume a fixed order of those, and write

$$\langle \pi_1, ..., \pi_k \rangle$$

to denote the sequence of phenotypes in that order. With respect to this order we can always write a distribution in explicit form just as a k-tuple of numbers

$$\langle r_1, ..., r_k \rangle \ , \ r_i \leq 0, \ \sum r_i = 1$$

where each number r_i is the weight or probability of phenotype number i occurring in the corresponding sequence of phenotypes. If it is not clear from the context in what order the phenotypes are written down we will include them in the sequence representing a genetic distribution thus writing

$$\langle r_1\pi_1, ..., r_k\pi_k \rangle$$

This notation is just intended to represent a genetic distribution in more explicit form.

The number k of phenotypes need not coincide with the number n of different progeny. In transmission applications these two numbers may be taken identical because each PROGENY_i in this case is a population, and populations are usually distinguished in terms of phenotypes In general the number of phenotypes in offspring may be smaller than the number of offspring, for instance because different individuals in the offspring may have the same phenotype.

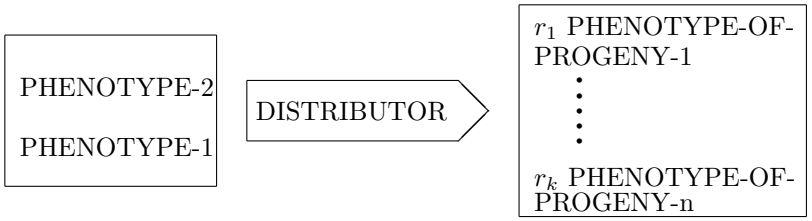
In molecular genetics the distribution of phenotypes may be used in two ways. On the one hand, relative frequencies may be calculated from the offspring of one parental pair of individuals. However, these frequencies may be far off the mark. On the other hand, therefore, relative frequencies may be obtained by iterated experiment or observation of mating identical (or similar) parental pairs, as was the case in the example of sickle cell anemia. The latter use is of course also characteristic for transmission genetics.

The transition from parental phenotypes to distributions of phenotypes in the offspring is described by a function we call DISTRIBUTOR. It takes the two parental phenotypes as arguments, and maps them into a genetic distribution.

DISTRIBUTOR(PHENOTYPE_1,PHENOTYPE_2) =
$\langle r_1\pi_1, ..., r_k\pi_k \rangle$ =
$\langle r_1$PHENOTYPE_OF_PROGENY_1,...,r_kPHENOTYPE_OF_PROGENY_k$\rangle$

where all r_i are positive real numbers such that $\sum_{i=1}^{k} r_i = 1$. As already stated the number k of phenotypes must not be identified with the number n of offspring in general. In our schema the second level therefore may be filled in as follows.

Fig.2-3

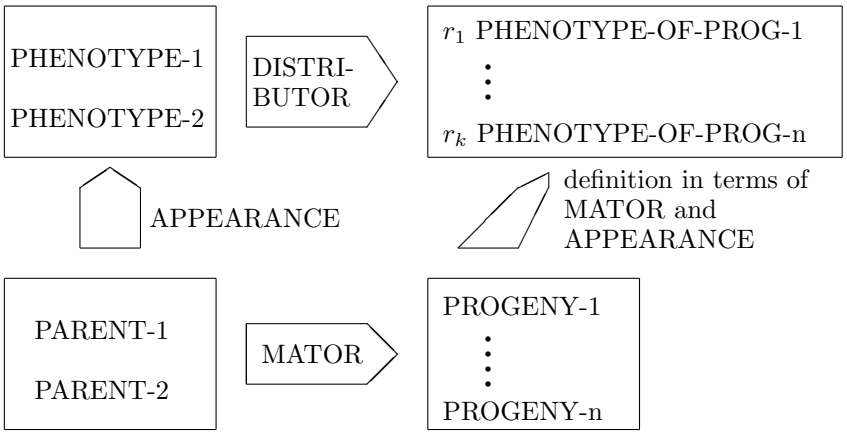


Between these two observational levels there is a natural relation: each genetic individual has its unique phenotype. This correspondence in the model is represented by a function APPEARANCE which assigns to each genetic individual, whether parental or offspring, its phenotype. Thus we have equations

APPEARANCE(PARENT_1) = PHENOTYPE_1
APPEARANCE(PARENT_2) = PHENOTYPE_2
APPEARANCE(PROGENY_i) = PHENOTYPE_OF_PROGENY_j
$(i \leq n, j \leq k)$.

A complete picture of the observational part of the model now has the form shown in figure 2-4. The upward arrow at the right represents the way in which the distribution of phenotypes is determined: we look at the value of MATOR(PARENT_1,PARENT_2), that is at the set of offspring. We look at the value of APPEARANCE(PROGENY_i) for $i \leq n$, that is at the phenotypes occurring in the offspring. We count the total number of offspring as well as the number of offspring showing a given phenotype, and calculate the relative frequency of this phenotype. Obviously, this yields a precise definition of the corresponding distribution of phenotypes which for given forms of MATOR and APPEARANCE can be mechanically evaluated. DISTRIBUTOR may be defined accordingly. For two given phenotypes, we may use APPEARANCE in the reverse direction to obtain the parents, and from there go to the right and upward to obtain the desired function value. On the parental side APPEARANCE may be reversed easily because of the small number of genetic individuals involved.

Fig.2-4



We turn now to the third, theoretical, level. At this level we meet of course the most interesting genetic terms: 'factor', 'gene', 'locus'. These refer to theoretical entities which are held responsible for the occurrence of particular phenotypes, traits and expressions. We use the label GENO-
TYPE to cover all these special entities. The straightforward idea is to think of genotypes as the *causes* of the phenotypes. This idea, however, faces difficulties from two sides. In general, the notion of cause and effect is a difficult one. Only recently there has been some development in probabilistic terms that might be applicable to genetics (but has not been applied as yet).[24] On the other hand, the relation between particular causes and effects -if they can be unravelled at all- in the case of genetics will turn out to be extremely complicated in general. When some difference in phenotype can be traced back to depend on a single peptide in the chromosome, matters are comparatively simple. Indeed most of the observable mutations of Drosophila melanogaster have been determined satisfactorily. When, however, more than one cause is involved, matters are more difficult, although not impossible. Thus, the minute bristle mutation appears on all four chromosomes of Drosophila melanogaster. The difficulty is technical rather than conceptual, and involves demarcing several causes for a single effect. In order not to enter into the philosophy of cause and effect[25] we will avoid causal terminology anyway.

There is a genotype for every phenotype. At the horizontal level we therefore have again two parental genotypes, denoted by

GENOTYPE_1 and GENOTYPE_2

and finitely many genotypes for the offspring, one for each phenotype occurring:

[24]See (Suppes, 1970).

[25]The reader interested in such questions is referred to (Mackie, 1974).

GENOTYPE_OF_PROGENY_1,...,GENOTYPE_OF_PROGENY_s.

The transition from parental genotypes to genotypes of progeny is represented by a function COMBINATOR which to any two parental genotypes assigns a combination or mixture of genotypes of progeny. Analogous to the situation with phenotypes, a quantitative, probabilistic element is needed. However, we can no longer talk about relative frequencies here because we are now on the theoretical level and in general things cannot be directly observed. Instead of relative frequencies we now speak about probabilities proper. The distinction is in certain respects one between experimental and expected probabilities. However, there are many applications where the relative frequencies from earlier experiments are used as a data-base for the estimation of the expected probabilities. A distribution of genotypes may be regarded as a genetic distribution as in the case of phenotypes, i.e. as a function assigning numbers ('weights') to the genotypes. Ordering the finitely many genotypes in a sequence

$$\langle \gamma_1, ..., \gamma_s \rangle$$

we may represent such a function by a similar sequence

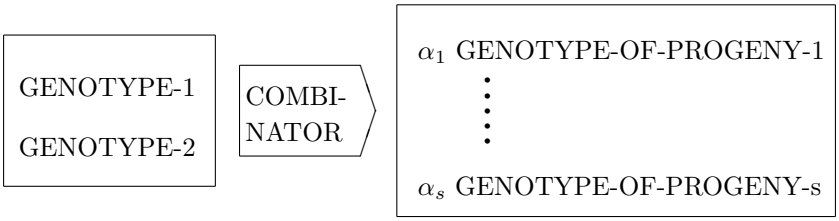
$$\langle \alpha_1, ..., \alpha_s \rangle,\ \alpha_i \geq 0,\ \sum \alpha_i = 1$$

which often is written in the form $\langle \alpha_1\gamma_1, ..., \alpha_s\gamma_s \rangle$ in order to make explicit the underlying ordering of the genotypes. A distribution of genotypes conveys the information that the genotypes $\gamma_1, ..., \gamma_s$ are expected to occur in the progeny with probabilities $\alpha_1, ..., \alpha_s$, respectively. COMBINATOR thus takes the form

$$\text{COMBINATOR(GENOTYPE\_1,GENOTYPE\_2)} = \langle \alpha_1\gamma_1, ..., \alpha_s\gamma_s \rangle = \langle \alpha_1\text{GENOTYPE\_OF\_PROGENY\_1,...,}\alpha_s\text{GENOTYPE\_OF\_PROGENY\_s} \rangle$$

and the boxes on the third level may be filled in accordingly.

Fig.2-5



There is a sharp difference between COMBINATOR and DISTRIBUTOR. The latter is observational, its function values are determined by direct empirical means, and anyway without recourse to the validity of particular genetic hypotheses. COMBINATOR, on the other hand, is a theoretically defined construct. In each non-trivial application COMBINATOR will be given by a theoretical definition which represents the particular hypothesis of how the genotypes

are transmitted in the system considered. On the other hand, there is of course a close connection between the two: COMBINATOR is a kind of theoretical image of DISTRIBUTOR.

An easy example of COMBINATOR in the Mendelian case is this. Take GENOTYPE_1 to be of the form $AABB$, GENOTYPE_2 of the form $aabb$. Then COMBINATOR$(AABB, aabb) = \langle 1/16AABB, 1/16AABb,$
$..., 1/16aabb\rangle$. When we look for examples for COMBINATOR in molecular genetics, however, we are disappointed. The overall impression is that, although molecular geneticists have studied the relationship between genotype and phenotype in ever increasing depth, they have neglected COMBINATOR. Possibly, there is the assumption that COMBINATOR at the molecular level represents a random process. Present thinking is, however, that this process is under genetic control, and it remains a major programme that the biochemistry of COMBINATOR be investigated. It should be stressed that use of the techniques of relative frequency measurement, as per transmission genetics, do not satisfy this programme. Even if the loci and characters are described in biochemical terms this is not sufficient to provide a molecular genetical version of COMBINATOR, although this would be necessary.

It would be too simple to say that genotypes are purely theoretical constructs. Historically, this view might perhaps be defendable for the beginnings of transmission genetics. In the early stages the factors were indeed hypothetical for there was no knowledge of their material basis. After the detection of chromosomes and their part in the transmission of hereditary traits in the first decades of the twentieth century, this hypothetical status became ever more accessible. Today the genotypes in some applications have a status as empirical as anything, and these triumphant applications of course yield credit to their existence in other cases which are still not entirely cleared up. The main reason for this process of getting 'less and less hypothetical' is the development of different independent means of access. Whereas originally the only access to genotypes was via hypotheses about the number and kind of genotypes involved and about the form of COMBINATOR, the situation after that steadily improved and today there are various means of access to genotypes such as electron microscopy, radiography, or restriction endonucleases. In this respect genetics shares company with very few distinguished mature natural sciences such as chemistry and quantum mechanics. The same story does not yet apply to COMBINATOR. Although many stable phases in meiosis are known the present view is that the nature of meiosis as a random process needs further investigation.

From these considerations it follows that the distinction used above between theoretical and observational level, in fact, is very fuzzy. If genotypes were unobservable in the early 19th century they certainly are observable now. So 'theoretical' entities may turn into 'observables'. One might say that 'artefact' becomes 'fact'. We are fully aware of the fuzziness of this distinction, and we do not want to base any important conclusions on it.

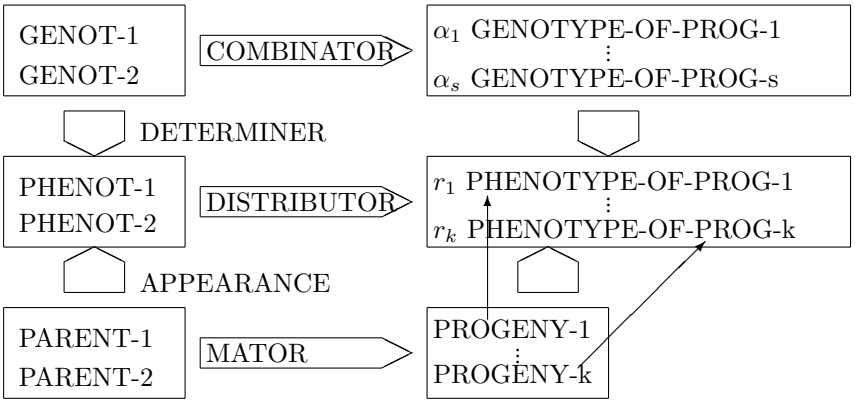
The vertical connection between levels one and two was already established by APPEARANCE. It remains to connect level three (of genotypes) with the other parts of the model. Most naturally this is achieved by relating genotypes

and phenotypes. As already mentioned the relation between a genotype and a corresponding phenotype may be very complex. A genotype may empirically consist of several parts of the chromosome or the DNA which all 'work together' in order to produce some phenotype, while in the absence of any part the phenotype will not occur. Nevertheless, we may assume that one genotype is not involved in the production of several different phenotypes. This is an analytic statement concerning the notion and choice of phenotypes rather than an empirical assumption. We use a function DETERMINER to assign phenotypes to genotypes. The label indicates the 'causal' direction: genotypes determine phenotypes but not the other way round. Note that a given phenotype may happen to be determined by several different genotypes, so DETERMINER in general will not be one-one. This is why we chose different numbers, k as the number of phenotypes, and s as the number of genotypes. In general k will be smaller or equal to s. We write

$$\text{DETERMINER(GENOTYPE\_i)} = \text{PHENOTYPE\_j } ,i \leq s, j \leq k$$

with appropriate indices. DETERMINER can be applied before and after mating so that we now can complete the schematic drawing which represents the overall structure of our model. It has to be noted that the picture does not represent the model in a complete way. What is missing in the picture are the genetic hypotheses required of the various objects and operators. The full model comprises the entities shown below plus the hypotheses postulated for them.

Fig.2-6



On the right hand side we cannot draw the arrows for DETERMINER and APPEARANCE in the same orderly way as on the left hand side because the mappings need no longer be one-one. Different progeny may have the same phenotype, and the genotypes produced by COMBINATOR may be more in number than the phenotypes observed (which would be a strong argument of course against the hypothesis underlying that particular COMBINATOR). We

have inserted dotted arrows in order to show the operation of DETERMINER and COMBINATOR more concretely.

We are now in a position to state an 'abstract' empirical claim associated with the model; 'abstract' because no special hypotheses about the number and kind of genotypes and the forms of COMBINATOR and DETERMINER are presupposed; these are left unspecified. As said before, the two lower levels really belong together since all items occurring are observational in the same way. In particular, DISTRIBUTOR can be defined if we know how the parental phenotypes are related to the two parental individuals (usually this knowledge is trivial). What is happening essentially, in the observational part of the model is that distributions of phenotypes are produced out of parental pairs of phenotypes. Using π_1^*, π_2^* and $\pi_1, ..., \pi_k$ as variables for parental phenotypes and phenotypes of offspring, respectively, we may write

$$\text{DISTRIBUTOR}(\pi_1^*, \pi_2^*) = \langle r_1\pi_1, ..., r_k\pi_k \rangle.$$

On the theoretical level this schema is 'replicated' by COMBINATOR for genotypes instead of phenotypes. By using γ_1^*, γ_2^* and $\gamma_1, ..., \gamma_s$ as variables for parental genotypes and genotypes of progeny, respectively, we may write

$$\text{COMBINATOR}(\ \gamma_1^*, \gamma_2^*) = \langle \alpha_1\gamma_1, ..., \alpha_s\gamma_s \rangle$$

The empirical claim associated with the model now is this:

(1) If the parental genotypes fit with the parental phenotypes then the distribution of genotypes produced by COMBINATOR will fit with the distribution of phenotypes given by DISTRIBUTOR.

This statement may be read as an axiom holding true for genetics in general and thus as *the* basic axiom of genetics. Fit at the parental side is given by DETERMINER. Parental genotypes fit with genotypes of progeny just in case the latter are the function values of the former under DETERMINER:

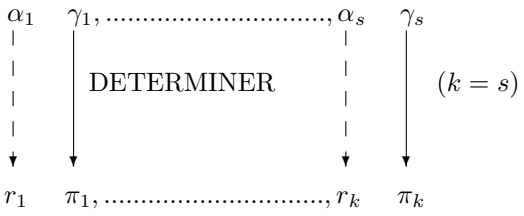
$$\text{DETERMINER}(\gamma_1) = \pi_1, \text{ DETERMINER}(\gamma_2) = \pi_2.$$

Of course, the special form of DETERMINER may be a matter of some theoretical depth involving, in particular, statistical considerations but such statistical considerations are not present *prima facie*. What is interesting is the fit in the model on the side of progeny. Here, statistical means come into play straightforwardly. We say that two genetic distributions $\langle \alpha_1\gamma_1, ..., \alpha_s\gamma_s \rangle$ of genotypes and $\langle r_1\pi_1, ..., r_k\pi_k \rangle$ of phenotypes *fit* with each other iff

i) the numbers k and s are identical
ii) each phenotype π_i arises from some genotype γ_j by means of DETERMINER
iii) the probability coefficients of the items related under ii) fit with each other.

This definition may be visualized if we agree on an appropriate ordering of the γ's such that DETERMINER$(\gamma_i) = \pi_i$. Fit of the two distributions is then given by arrows from the genotypes to the phenotypes, and by corresponding arrows (drawn dotted in Figure 2-7 below) between the coefficients.

Fig.2-7



Statistics is then involved in the fit of the coefficients. Once the order as shown in Figure 2-7 is present we may regard the coefficients $\alpha_1, ..., \alpha_s$ to represent a theoretical distribution, a 'curve', and the relative frequencies $r_1, ..., r_k$ as 'data' to be fitted with the curve according to some statistical procedure. Still more simply we may just compare the 'distances' $\mid \alpha_i - r_i \mid$ and require them not to exceed some given ϵ. The latter method is very crude if the choice of ϵ is not made dependent on the shape of the distribution of the γ's.

The question of goodness of fit has two aspects. First, the choice of a suitable hypothesis about the distribution of COMBINATOR from those allowed by the particular theory. Second, the ability to subsume the empirical data under the given hypothesis to a pre-determined level of accuracy. As an illustration of the latter, consider the following example (Elandt-Johnson, 1971).[26] The segregation ratios at the Ag-B locus in cats differs from the Mendelian ratios 1:2:1 in intercrosses $B'B^4 \times B'B^4$. The main blood group locus in cats is called Ag-B, with which several alleles are associated. The deviations from Mendelian ratios are shown in that

Fig.2-8

GENOTYPE	BB'	$B'B^4$	B^4B^4
observed frequency (r_i)	58	129	13
expected frequency (α_i)	50	100	50

After having chosen a suitable ϵ we may apply, say, a χ^2-test to these data, and see whether the χ^2-value is within the limits of ϵ. In the present case, an ϵ which does not yield rejection of the underlying Mendelian hypothesis would have to be greater than 37, so the hypothesis of simple Mendelian ratios must be rejected. Clearly, the more traits involved, the greater the necessity of some form of statistical analysis such as the above becomes.

A little reflection about the empirical claim just described reveals that this claim is not entirely empirical for the theoretical parts of the model, GENOTYPES, COMBINATOR and DETERMINER, are assumed as given when the

[26](Elandt-Johnson, 1971).

claim is formulated. However, in most applications these components have a hypothetical status, and therefore the claim depends on the corresponding hypotheses. If the claim for a given set of GENOTYPES, COMBINATOR and DETERMINER turns out as untenable because the ϵ needed to produce a fit is too big, another such set may provide a more satisfactory fit and thus a more tenable claim. The question of which set of theoretical entities should be taken is a difficult one. Up to now no readily useable criteria have been put forward for such choice of theoretical entities (if we ignore abstract, purely philosphical accounts). The range of possibilities for the three items under consideration is in principle infinite for a given observational part of the model. This infinity is restricted in practice by the formulation of special hypotheses, special laws, about the kind and number of genotypes and the mathematical forms of COMBINATOR and DETERMINER. Even such special laws are often not sufficient to determine the theoretical components uniquely. Noteably in the case of Mendel's laws, which provide an explicit definition of COMBINATOR there are some degrees of freedom left for the choice of genotypes and of DETERMINER. Leaving Mendel's law untouched we may obtain quite different results by variation of GENOTYPES and DETERMINER.

If there is no unique choice of theoretical components as prescribed by the observational part of the model, which set of theoretical terms should we take in order to state the empirical claim? The standard answer to this question given in philosophy of science[27] is: an arbitrary set from the range of possibilities admitted by the model. This prescription amounts to reducing the empirical claim (1) above to an existential claim.

(2) For any given intended system *there exists* some set of theoretical components which, added to the observational part of the corresponding model, yield a claim of the form (1) with satisfactory measures of fit.

Such a claim may be trivial if the requirements imposed by the model on the theoretical terms are weak. The existence of suitable theoretical parts might then even be provable with pure logic. In fact, this is the case for the model presented here in many cases where there are no data about COMBINATOR and few data about DETERMINER. However, we do not have to abandon the model for that reason as trivial. Recall that this model is intended as a core-model to be applicable in *all* intended systems of genetics. Core-models in other disciplines yield claims of the same trivial kind, for instance in classical mechanics, or in phenomenological thermodynamics.[28] From the core-model we may easily obtain specialized models by means of imposing further hypotheses intended to apply to a small subset of intended systems (like Mendel's laws, or the Watson-Crick model of the double helix). These special models usually yield non-trivial empirical claims of the form (2), and this is why the label 'empirical claim' seems to be justified even in the trivial case of the core-model.

[27] Compare, for instance, (Stegmueller, 1986), Chaps.1 and 2.

[28] See (Balzer, Moulines, Sneed, 1987), Chaps.3 and 4 for details about the examples mentioned.

In order to bring out the non-triviality of special laws let us consider the simplest case of a diploid Mendelian application to one character difference, say, between phenotypes p and P. In this case the following hypotheses have to be added to the core-model.

i) There are exactly two kinds of 'factors' A, a such that each GENOTYPE γ is a pair of two such factors (e.g. $\gamma = \langle A, a\rangle$)
ii) DETERMINER assigns phenotype P to one of the genotypes with identical components (say, to $\langle A, A\rangle$), and p to any other genotype
iii) COMBINATOR for any two parental genotypes produces all possible pairs of factors from the factors present in the genotypes and weighting them equally with 1/4 (e.g. COMBINATOR$(\langle Aa, AA\rangle) = \langle 1/4AA, 1/4AA, 1/4aA, 1/4aA\rangle$).

Clearly, the GENOTYPES as well as COMBINATOR and DETERMINER are uniquely determined by these requirements (up to renaming of the factors). So in a special model for which requirements i)-iii) hold there is not much room for the choice of theoretical components and therefore the existential claim (2) above loses its triviality. In fact, the claim in this special case reduces to the claim that the theoretical terms as determined by i)-iii) fit with the observational part of the model, in particular with the distribution of phenotypes. It is intuitively clear that this claim is true or false according to what the observed frequencies in the distribution of phenotypes look like. This claim, obtained from the abstract claim (2) above by filling concrete hypotheses into our core-model, is a strong, empirically refutable claim indeed.

On the basis of the core-model we may define other concepts important for genetics. The fact that these do not occur as primitives in the model has no bearing on their importance. The choice of primitives for a theory has a large range of conventionality. Of course, one would not take minor concepts as primitives. Once a sufficient basis of primitives is at hand all other terms of the theory should be definable. The point of reducing the number of primitives is that in this way the number of basic hypotheses is also reduced and thus the theory or the model becomes more perspicuous and compact. With respect to definability the situation is similar to that of empiricity. Usually, a new term cannot be defined in the core-model because its definition affords further assumptions which hold only in special cases. So the definition of various genetic terms will go together with the introduction of corresponding assumptions.

The term 'gene' has been much discussed in the history of genetics.[29] Consider the following definition. A gene is a materially identifiable entity that determines the expression of a trait in the phenotype. This definition can be accomodated by our model if we are generous with phenotypes. If the phenotypes are taken to be just different expressions of one trait then genes as just defined are just genotypes. For each gene (genotype) by means of DETERMINER, in fact, determines one expression (phenotype) of the trait under consideration.

[29]See Carlsson (1966) for a good survey.

Though not very subtle, this way of proceeding agrees with real application, for instance, Watson and Crick[30] say 'The various traits are controlled by pairs of factors (which we now call genes), one factor derived from the male parent, the other from the female. For example, pure breeding strains of round peas contain two genes for roundness (RR) whereas pure breeding wrinkled strains have two genes for wrinkledness (rr).' This is a typical use of the term. It was as readily employed to signify chromosomal features and nucleotide sequences.

In general, the definition of 'gene' stated above requires more subtle treatment in terms of our model in order to fit more generally with real applications. Given our frame, the general case that may occur is this. One phenotype in the model may represent a complex of expressions of different traits, and one genotype may analogously represent a collection of materially identifiable entities. In this situation we cannot take the whole genotype as a gene. Also, if the atomic components of the genotype cannot be uniquely related to the different traits and their expressions in the phenotype, we cannot take these atomic components as genes. The situation then depends on the form of DETERMINER. If DETERMINER is such as to relate clusters of material parts of chromosomes, i.e. their representatives in the genotype, into clusters of expressions in a way not further decomposable, it seems inadequate to talk about genes in the sense of the definition stated previously.

These considerations indicate the circumstances in which we may talk about genes, that is, in which the term may be defined by means of our model's primitives. The circumstances are represented by a specialization of our model. We say that, in a given model, DETERMINER is *decomposable* iff DETERMINER can be written as a tuple

$$\langle \text{ DET\_1,...,DET\_r} \rangle$$

such that each DET_i maps a well specified part of the genotype, which is the argument of DETERMINER, on exactly one 'part' of the phenotype such that the phenotype is just the combination (the tuple) of these parts. More precisely, decomposability may be defined as follows.

DETERMINER is *decomposable* iff there exist sets $P_1, ..., P_r$,
$G_1, ..., G_s$, sets of indices $J_i = \{j(i,1), ..., j(i,\sigma(i))\}$ for
$i = 1, ..., r$, and functions DET_1,...,DET_r such that
1) each phenotype π can be represented in the form $\pi = \langle p_1, ..., p_r \rangle$
with $p_1 \epsilon P_1, ..., p_r \epsilon P_r$
2) each genotype γ can be represented in the form $\gamma = \langle \delta_1, ..., \delta_s \rangle$
with $\delta_1 \epsilon G_1, ..., \delta_s \epsilon G_s$
3) the set $\{1, ..., s\}$ of indices is the same as the union of all
the sets J_i, $i \leq r$: $\{1, ..., s\} = \cup\{J_i / i \leq r\}$
4) for all $i \leq r$: DET_i maps genotypes into elements of P_i
5) for all $i \leq r$: DET_i properly depends exactly on all its
arguments with indices $j(i,1), ..., j(i,\sigma(i))$

[30]See (Crick and Watson, 1953).

6) for all genotypes γ:

$$\text{DETERMINER}(\gamma) = \langle\ \text{DET\_1}(\gamma),...,\text{DET\_r}(\gamma)\rangle$$

That is, each phenotype has the form of a tuple $\langle p_1, ..., p_r\rangle$ consisting of 'component phenotypes'. Each set P_i may be regarded as representing a trait, and the elements $p_i \epsilon P_i$ as expressions of this trait. Each genotype consists of a tuple $\langle \delta_1, ..., \delta_s\rangle$ of 'component genotypes'. Each sequence $\langle j(i,1), ..., j(i,\sigma(i))\rangle$ picks out the indices of those components of $\langle \delta_1, ..., \delta_s\rangle$ on which DET_i actually depends, and by 4), DET_i maps the genotype with these components into expressions of the trait P_i.

Now consider the class of models in which DETERMINER is decomposable. This class clearly is a subclass of the class of all possible models. In it, we may define a gene as a combination

$$\langle \delta_{j(i,1)}, ..., \delta_{j(i,\sigma(i))}\rangle$$

of genotype components which are arguments on which one of the 'component determiners' DET_i depends. Note that this definition allows for material genes which are spread out spatially, i.e. which are located at different sites on the chromosome. Still, since they only determine one expression when taken together they are a proper unit to be called a gene according to the definition we started with. On the other hand one is tempted to insist that one gene should determine just one expression of one trait. Apparently, this is the conventional definition. In principle, however, we may choose the structure of phenotypes appropriately in different applications, to use 'expressions' and 'traits' in the model as representing rather complex traits in the real system.

The above definition of gene is not entirely simple, though the idea is easy to grasp. We believe that this fits well with the fact that this notion was disputed for a long time, and still is today.[31]

Another important derived notion is penetrance as defined by the proportion of genotypes that shows one expected phenotype. As in the case of 'gene' this notion cannot be immediately be defined in our model; recurrence to further specification of the model is needed. We have to use a function which assigns genotypes to genetic individuals directly. If ϕ is such a function, i.e. a function of the format given by the expression $\phi(x) = \gamma$ (where x stands for a genetic individual and γ for a genotype), we may define the penetrance of genotype γ with respect to phenotype π as (the number of genetic individuals having genotype γ as stated in terms of ϕ and phenotype π) over (the number of genetic individuals having genotype γ), or formally (with $|z|$ denoting the cardinality of set z):

$$PENETRANCE(\gamma/\pi) = \frac{|\{x/\phi(x)=\gamma \ \ and \ \ APPEARANCE(x)=\pi\}|}{|\{x/\phi(x)=\gamma\}|}$$

[31] See (Kitcher, 1982) for a discussion of the problems with the concept of gene. Though his view of the structure of genetic theory is very different from ours, his account of these problems is largely independent of that view.

Expressivity as the degree to which a particular effect is expressed in genetic individuals can be defined by reference to a refinement of APPEARANCE. If PHENOTYPES are assumed to be tuples of expressions of traits, a scale may be introduced on each trait ordering the expressions of that trait with respect to various degrees. Expressivity of a trait in a genetic individual then is the value of the scale corresponding to that trait for the individual considered.

Pleiotropism as the multiple phenotypic effect of a single gene may be easily defined by refining PHENOTYPES, GENOTYPES and DETERMINER in a suitable way. Roughly, we may proceed as above in connection with genes but take the 'component phenotypes' p_i to represent the multiple effect under study. So p_i will not be one expression of one trait, but rather a set (or sequence) of expressions of a multiplicity of traits.

Epistasis occurs when one gene pair causally screens the effect of another one. We have to employ a counterfactual in connection with DETERMINER in order to define this notion. Consider two genes g, g^* of the form introduced above which are both part of some GENOTYPE γ. Then g *screens* g^* with respect to phenotype π iff:

i) DETERMINER$(\gamma) = \pi$
ii) if g in γ would be replaced so that the resulting genotype was γ' then DETERMINER$(\gamma') = \pi$.

We leave it to the reader to produce similar accounts for the various notions of dominance to be found in the literature.

From a formal point of view the model introduced is still ambiguous as far as the numbers of genetic individuals is concerned. Of course, we do not want to restrict it to a particular number of, say, offspring. What has to be clarified, however, is the number of different matings captured by the model (whether on the level of populations or of individuals). As we left this feature somewhat undecided (deliberately) the impression might have been created that each model just covers exactly one process of mating with subsequent creation of offspring. This is certainly the main case of application. The model does not rule out, however, more complex applications in which more than one mating is considered in one and the same model. The reason for keeping such a level of generality is found in later applications in the context of population genetics proper. There, the formalism of genetic algebras is best suited to describe transitions from a definite number of parental populations to the same number of populations in progeny. The way our models are introduced allows for simple incorporation of that algebraic formalism.

In Chap.4 we will make this explicit. In that context we will have to refer to -and quantify over- the genetic individuals occurring in one model. In order to avoid confusion it is good to have a description of the models paying more attention to the numbers, and sets, of objects involved. We will state such a description now which also satisfies the logicians search for completeness of description. As stated earlier we do not take over the probabilistic notion of a distribution: there is no use of the general features of σ-algebras here. As the

reformulation is essentially a matter of terminology, we may be brief.

If X is a non-empty, finite set then by a Γ-*distribution* over X we mean a function $\mathbf{p} : X \to [0,1]$ such that $\sum_{x \epsilon X} \mathbf{p}(x) = 1$. $[0,1]$ here denotes the closed interval of reals between 0 and 1. If the members of X are ordered in some way, so that a list $\langle x_1, ..., x_n \rangle$ captures exactly all of X's elements we may write down the function values of a γ-distribution $\mathbf{p}$ over X in the same order

$$\langle \mathbf{p}(x_1), ..., \mathbf{p}(x_n) \rangle = \langle \alpha_1, ..., \alpha_n \rangle.$$

Above, the x's were phenotypes or genotypes. For x_i = GENOTYPE_ OF_PROGENY_i, for instance, we wrote α_iPHENOTYPE_OF_PROGENY_i in order to state that α_i 'belongs to' GENOTYPE_OF_PROGENY_i. In the present, abstract notation the GENOTYPE OF PROGENY_i are 'swallowed' by the distribution and reoccur as its arguments:

$$\mathbf{p}(\text{GENOTYPE\_OF\_PROGENY\_i}) = \alpha_i,$$

so that there is no need to write them down additionally. The set of all Γ-distributions over some set X we denote by $D(X)$.

We introduce sets **J**, **P**, and **G** the elements of which are interpreted as the *genetic individuals*, *phenotypes*, and *genotypes* occurring in the model, respectively. Genetic individuals may be individuals proper or populations. The variables i, π, and γ are used to range in these sets, respectively. So we write

$$i \epsilon \mathbf{J},\ \pi \epsilon \mathbf{P} \text{ and } \gamma \epsilon \mathbf{G}$$

to express that i is an arbitrary genetic individual, π is a phenotype, and γ a genotype in the model. By $X \times Y$ we denote the cartesian product of the sets X and Y, i.e. the set of all pairs $\langle x, y \rangle$ with $x \epsilon X$ and $y \epsilon Y$, and by $\mathbf{Po}(X)$ the power set of X.

For a phenotype $\pi \epsilon \mathbf{P}$ and a set of genetic individuals $X \subseteq \mathbf{J}$ we define the *relative frequency of* π *in* X, $RF(\pi/X)$, as follows.

If X is a set of proper individuals then
$RF(\pi/X)$ = (the number of $i \epsilon X$ such that APPEARANCE(i) = π)
over
(the number of elements of X)
and
if X is a set of populations then
$RF(\pi/X)$ = (the number of elements in the sets $i \epsilon X$ for which
APPEARANCE(i) = π)
over
(the number of elements in members of X).

A *model of genetics* is a structure of the form

$\langle \mathbf{J}, \mathbf{P}, \mathbf{G}, \mathbf{APP}, \mathbf{MAT}, \mathbf{DET}, \mathbf{DIST}, \mathbf{COMB} \rangle$ which satisfies the following requirements:

A1 **J**, **P** and **G** are non-empty, finite sets, and pairwise disjoint
A2 **APP**: $\mathbf{J} \rightarrow \mathbf{P}$
A3 **MAT**: $\mathbf{J} \times \mathbf{J} \rightarrow \mathbf{Po(J)}$ is a partial function
A4 **DET**: $\mathbf{G} \rightarrow \mathbf{P}$ is surjective
A5 **DIST**: $\mathbf{P} \times \mathbf{P} \rightarrow D(\mathbf{P})$ is a partial function
A6 **COMB**: $\mathbf{G} \times \mathbf{G} \rightarrow D(\mathbf{G})$
A7 for all $i, i' \epsilon \mathbf{J}$ such that **MAT** is defined for $\langle i, i' \rangle$ and for all $\pi \epsilon \mathbf{P}$:

$$\mathbf{DIST(APP}(i)\mathbf{,APP}(i'))(\pi) = RF(\pi/\mathbf{MAT}(i, i'))$$

A8 for all $i, i' \epsilon \mathbf{J}$ such that **MAT** is defined for $\langle i, i' \rangle$ and for all γ, γ' $\epsilon \mathbf{G}$ such that $\mathbf{DET}(\gamma) = \mathbf{APP}(i)$ and $\mathbf{DET}(\gamma') = \mathbf{APP}(i')$, and for all $\gamma^* \epsilon \mathbf{G}$:

$$\mathbf{COMB}(\gamma, \gamma')(\gamma^*) \approx_\epsilon \mathbf{DIST(DET}(\gamma)\mathbf{,DET}(\gamma'))(\mathbf{DET}(\gamma^*))$$

APP, MAT, DET, DIST, COMB here denote the previous operators APPEARANCE, MATOR, DETERMINER, DISTRIBUTOR and COMBINATOR. MATOR and DISTRIBUTOR are required to be partial functions only, i.e. they need not be defined for all possible arguments. MATOR should be undefined for pairs $\langle i, i' \rangle$ which do not mate. Such pairs necessarily occur in the model, namely among the offspring. DISTRIBUTOR similarly needs not be defined for pairs $\langle \pi, \pi' \rangle$ which do not correspond to mating genetic individuals. One might of course insist that DISTRIBUTOR should be defined for any two phenotypes without consideration of whether these correspond to genetic individuals mating or not. However, this is not in line with our interpretation of distributions of phenotypes as determined by observed relative frequencies in progeny. If DISTRIBUTOR were defined for all pairs of phenotypes DISTRIBUTOR would represent a kind of law or law-like connection in the phenotypes. Such a law should be present only at the level of genotypes, however, and this is why COMBINATOR in A6 is required to be a full function. A7 is the explicit definition of DISTRIBUTOR in terms of relative frequencies, and A8 the basic axiom stating that theoretical frequencies of genotypes as produced by COMBINATOR should coincide -at least approximatively up to a given ϵ - with those observed in progeny as expressed in the corresponding function value of DISTRIBUTOR. Axioms A7 and A8 may be read as stating equalities and approximative equalities of genetic distributions. If we write down the relative frequencies $RF(\pi, \mathbf{MAT}(i, i'))$ in the order of the corresponding phenotypes we obtain a Γ-distribution which may be called the distribution of frequencies. A7 then states that the distribution of frequencies is the same as that given by DISTRIBUTOR. A8 states that the distribution of genotypes, $\mathbf{COMB}(\gamma, \gamma')$ is approximatively the same as the corresponding distribution of phenotypes $\mathbf{DIST(DET}(\gamma)\mathbf{,DET}(\gamma'))$.

The two kinds of applications mentioned above, those covering just one or more than one mating, now can be distinguished by looking at the domain of MATOR, *Domain*(MATOR), i.e. the set of all pairs $\langle i, i' \rangle$ for which MATOR is defined. If the model describes just one mating the domain of MATOR

will contain just one pair of genetic individuals. Whether this is so or not we leave open to decide in each particular application. At the level of model construction we might of course consider the corresponding axiom stating that *Domain*(MATOR) contains only one pair. This axiom added to A1-A8 characterizes a special subclass of models, the difference between this and the full class of models marking the way in which the general models of A1-A8 extend the former.

Chapter 3

Genetic Kinematics

Kinematics is the description of change over time. The general model introduced in the previous chapter is a static model, or better, a quasi-static one. Development in time is covered only in the rudimentary form of one transition from parents to progeny.

The model may be -and has to be- extended to cover kinematical features in two different ways. The first area in which a kinematical model has to be superimposed is the area covered by DETERMINER. In modern genetics the genotypes are no longer mere theoretical terms, and an important part of molecular genetics deals with the transitions from DNA to amino acids and to enzymes. Here, a cyto-chemical model exists describing in detail how DNA gets translated into RNA, and how RNA 'constructs' amino acids.[32] However, if we look at kinematics in a more narrow way to describe sequences in time of the change of definite states of a system, the cyto-chemical model may not easily pass as a kinematical model for it is difficult to identify an homogenous set of 'states' such that the changes may be described as transitions from one state into another one. A proper kinematic model satisfying this condition might be called a model of *transition kinematics*. We will not deal here with transition kinematics for two reasons. First, the cyto-chemical basic processes are well understood and are described now in every textbook. Second, as just noted, the knowledge available is not yet sufficient to construct a homogenous model of transition kinematics.

The second area in which a kinematical model has to be inserted in the basic model of Chap.2 is the area covered by COMBINATOR. Here, too, modern genetics has achieved detailed knowledge about how strands of DNA separate and recombine during meiosis. In this area it is possible to describe the genetic processes in the form of sequences of definite states so that a kinematical model in the more narrow sense may be constructed. We call it a model of *combination kinematics* for it deals with how parental DNA is combined during meiosis, and subsequently is combined in fertilization. The model is intended to cover both these stages: meiosis as well as possible combination in fertilization. Its main area of application is of course to meiosis proper. A special case of particular interest is that of recombination in the well known technical sense to which we will turn below.

In the general model of the previous chapter, the genotypes have the status of abstract, theoretical entities. The model itself does not imply that the γ's have

[32] Compare (Strickberger, 1985), Sec.4 and 5, for example.

to be interpreted by concrete material objects. Indeed, such a general approach is necessary in order to cover the earliest genetic applications, for instance by Mendel, in which 'factors' were used as mere theoretical, combinatorial means.

Part of the history of success in genetics was that these initially abstract terms gradually became more and more concrete. As early as 1903, a connection was drawn by Sutton between the material in the cell which was thought to be the basis of inheritance, and Mendelian factors. In the course of two decades this connection proved valid; chromosomes were identified, and their crucial role in mitosis and meiosis began to be investigated. From the outset, chromosomes were postulated to function as the 'causes' of phenotypes and thus to be material counterparts to Mendelian factors. The full causal chain leading from some particular chromosomes to the complete phenotype is far from being known, but there is enough evidence to show that variation in the chromosomes implies variation in the corresponding phenotype. For instance, at a very early stage it was clear that there were differences in chromosome content related to sex determination. A comparison of the cytological and genetic maps for the X-chromosome of Drosophila melanogaster was provided by Morgan and Dobzhanski.[33] The geometrical ordering of chromosomal features being identified with that of hypothetical factors as established from studies of relative frequencies. Moreover, a great deal is now known on the structure of chromosomes. The strings of DNA making up the chromosomes are well known down to their chemical composition for many important species, and in many cases (like sickle cell anaemia) particular parts of the DNA have been shown to be partial causes of certain definite traits.

Less is known about how and why the DNA molecules wind up, separate, and join just the way they do in the different stages of meiosis. These phenomena being a major issue in genetics at present, and very likely in the near future, a model is desirable in which the change over time of configurations of chromosomes and DNA can be described. Such a model we call a *combination kinematics*. As the label indicates, the present task is just to *describe how* the chromosomes change their configurations in time, and not to *explain why* they do so. The latter task would afford a truly dynamical model in which the 'forces' responsible for the changes are made explicit.

In the present chapter we want to refine the previous model to include such a combination kinematics. Since we want the refined model to be as general as possible (to cover as much of genetics as possible) we do not bring the full cyto-chemical machinery into play. Most recombination studies to-day do not use cyto-chemical knowledge in a substantial way (the well known techniques of chemical or radioactive labelling notwithstanding). In order to achieve a general approach we use a more abstract terminology covering structure and change at the level of chromosomes as well as at that of DNA.

On both these levels it is undisputed that we are dealing with material objects of a lengthy form, made up of different material 'parts' which can be separated. Essentially, the lengthy objects may be conceptualized as sequences

[33]See (Morgan and Bridges, 1916) and (Dobzhansky, 1932).

of smaller kinds of objects. The objects of the type of sequences we call *strands*, the objects making up those strands we call *quanta*. We do not go into further details about the chemical composition of the quanta, nor about the particular kinds of chemical bonds that may obtain between them. The only feature we want to make explicit is that quanta are ordered in a certain sequence to form a strand. Conceptually, this amounts to their forming a *linear order* with respect to some order relation $<$. A linear order consists of a set Q (in our case a set of quanta) and a binary relation $<$ on this set which satisfies the axioms AO1-AO3 below. Using q, q', q'' as variables for quanta we may write:

$$q < q' \text{ (`}q \text{ is smaller than } q'\text{', or `}q \text{ comes before } q'\text{').}$$

The axioms characteristic for a linear order are the following:

AO1 For all q, q', q'' in Q: if $q < q'$ and $q' < q''$ then $q < q''$
('Transitivity')
AO2 For all q in Q: not($q < q$) ('Anti-reflexivity')
AO3 For all q, q' in Q: $q < q'$ or $q' < q$ ('Connectedness')

As we are interested only in strands composed of finitely many quanta, we may define the notion of a strand as follows.

> s is a *strand* iff s consists of a finite set Q and a binary relation $<$ on Q which is transitive, anti-reflexive, and connected.

If s is a strand we say that q is a quantum *of* (or *in* s) if q is an element of Q. In genetic context, a variety of interpretations of this notion are possible. At the level of chromosomes, the quanta refer to identifiable features and banding of the chromosome, the line of the chromosome, the $<$-relation refers to their order on the chromosome, and the notion of a strand refers to the whole chromosome or to some part of it (which actually may be quite short). At the level of DNA quanta may be regarded as denoting the bases or the triplets of bases, $<$ as denoting their ordering along the DNA molecule, and a whole sequence of bases in a molecule (or a part of such a sequence) would be a referent for the term 'strand'. Note that under this interpretation of the term 'quantum' it is not the whole DNA molecule (or a section of it) that forms a strand. Rather it is 'one half' of the double helix (or a section of it) only, i.e. a sequence of phosphodiester bridged bases.

We note that the notion of a linear order has weak implications only for the topological form of a strand in space. A strand may have an arbitrarily complex shape being wound up along different axes in whatever regular or irregular way. The only constraints imposed by linearity are first, that the strand is connected, not falling apart into separate sub-strands, and second, that it does not contain any loops (one of its ends being connected to its other end). Also, nothing is implied about the distances between the different quanta in a strand.

As the set of quanta in a strand is supposed to be finite, it is easy to define some natural concepts we use in talking about strands. For instance, each strand has two 'ends' which may be defined as the unique *minimal* and *maximal* quantum in the strand. The unique minimal quantum in $s = \langle Q, < \rangle$ is defined as the quantum q in Q such that there is no other quantum q' in Q smaller than q ($q' < q$), the definition of 'maximal' is analogous. Furthermore, we may define the notion of two quanta being *neighboured*. If $s = \langle Q, < \rangle$ and q, q' are quanta in Q then q and q' are *neighboured* (*in* s) if there is no q'' in Q such that q'' is in-between q and q' (i.e. $q < q'' < q'$ or $q' < q'' < q$). Clearly, if we start with the minimal quantum in s, proceed to its neighboured quantum, and iterate this procedure we run through all the quanta of s in the order specified by $<$. The unique number of steps needed in this process to reach some given quantum q in s may be called *the position of* q *in* s. Thus if the position of quantum q in strand s is n then q is the n-th quantum on the string, counted from the minimal end of s. Of course, to start from the minimal end here is a mere convention, we might as well start 'from above'. As usual, we define '$q \leq q'$' as an abbreviation for '$q < q'$ or $q = q'$'. It is easy to show that this relation again is transitive and connected. It is not anti-reflexive, however. Instead, it is anti-symmetric, i.e. from $q \leq q'$ and $q' \leq q$ it follows that $q = q'$.

Strands may be 'cut down' and concatenated. The first procedure yields one or several sub-strands. We define sub-strands such that they consist of 'connected' parts of the original strand from which they are formed. More precisely,

s^* is called a *sub-strand* of a strand $s = \langle Q, < \rangle$ iff
there exist $Q^*, <^*$ and quanta q, q' in Q such that
i) Q^* is the set of all quanta q^* such that $q \leq q^* \leq q'$
ii) $<^*$ is the binary relation defined on Q^* by:
for all $q_1, q_2 : (q_1 <^* q_2$ iff $q_1 < q_2)$
iii) $s^* = \langle Q^*, <^* \rangle$.

Obviously, any sub-strand of a strand again is itself a strand. By the concatenation of two strands we mean the new strand which is obtained by putting together the given strands at two of their ends. We agree to put together the maximal end of the first with the minimal end of the second. Thus the *concatenation* of two strands $s = \langle Q, < \rangle, s' = \langle Q', <' \rangle$ which we denote by $s \circ s'$ is defined as $s^* = \langle Q^*, <^* \rangle$ where Q^* is the union of Q and Q', and $<^*$ is defined by

for all q, q' in Q^*:
$q <^* q'$ iff either q, q' are both in Q and $q < q'$
or q, q' are both in Q' and $q <' q'$
or q is in Q and q' is in Q'.

The concatenation of two strands can be shown to be a strand, too. The genetic content of an individual being composed of several chromosomes we have to use at least a *set of* strands as representing a genotype as occurring in the general model. A mere set of strands, however, is not suited as a basis to describe the

structure and change of strands. A combination kinematics that deserves this label has to include at least some means of describing the spatial configurations of the different strands making up one genotype. Since these spatial configurations are of great complexity, a correspondingly general notion is needed for their representation. In general, the easiest way to describe such configurations is by mapping them into three dimensional space. Although in special applications such mappings may be replaced by more efficient representations, in general they are not. We therefore define a *configuration of strands* to consist of a finite set N of strands, and a mapping ψ which to each quantum in each of the strands assigns a position in three dimensional space $\mathbb{R}^3$. The function ψ we call *position function* for it assigns a position to each quantum. In other words a configuration of strands may be described as a list C:

$$C = \langle N, \mathbb{R}^3, \psi \rangle$$

where N is a finite set of strands, each s in N being of the form $s = \langle Q_s, <_s \rangle$ described above, $\mathbb{R}^3$ denotes the set of 3-dimensional vectors of real numbers, and ψ is a mapping assigning to each quantum in each set Q_s a vector in $\mathbb{R}^3$ as its position. We write

$$\psi(q) = \langle \alpha_1, \alpha_2, \alpha_3 \rangle$$

and we call $\psi(q)$ q's *position* (*in* the configuration of strands C). It has to be excluded, of course, that different quanta in one configuration C get assigned the same position. More realistically, we have to take into account that the quanta have some extension in space, and therefore their positions must have certain minimal distances from each other. Thus we require the following axiom for configurations.

(*) There is some real $\epsilon > 0$ such that for all quanta q, q' in strands of C, if $q = q'$ then $|\psi(q) - \psi(q')| \geq \epsilon$.

It is easy to show that any subset of strands taken from the set N of strands of a configuration of strands again yields a configuration of strands, provided the position function ψ is restricted to the quanta occurring in the strands of that subset. Any such configuration we call a *sub-configuration* of the initial one. The number ϵ in (*) has to be given externally, and will vary from application to application.

The notion of a configuration of strands on the one hand is simple enough for easy application, on the other hand it seems rich enough to express all kinds of configurations actually occurring in genetics. Of course, it does not comprise all kinds of genetic concepts used in connection with combination kinematics, like 'centromere' or 'chiasma'. It provides just a means of describing various (and possibly complicated) structures of chromosomes and DNA, and thus provides a basis for descriptions of the change of such structures over time, i.e. a basis for combination kinematics.

A first step in refinement of the general model from Chap.2 now may be performed by assuming that the genotypes occurring in the model have the form

of configurations of strands. In other words, we require that each genotype γ, in fact, *is* a configuration $C = \langle N, \mathbb{R}^3, \psi \rangle$ of strands. Using the variables C, C', C_i instead of $\gamma, \gamma', \gamma_i$ we may write the connection established by COMBINATOR between genotypes of parents and progeny in the form

$$\text{COMBINATOR}(C, C') = \textstyle\sum_i \alpha_i C_i.$$

We note that this shift in notation by itself still does not commit us to interpret the letters 'C' etc. as referring to material objects. We still may, if we wish, interpret C just as an abstract factor or genotype. In such an abstract interpretation the internal structure of configurations of strands, however, would seem redundant. This indicates that the transition from the general models of Chap.2 to the refined models we begin to describe here roughly corresponds to the transition from transmission genetics proper (in which genotypes are just abstract, theoretical terms) to a form of genetics in which genotypes have a material (chromosomal,DNA) basis. In a way, this transition is analogous to that from early chemistry (in which atoms function as abstract units of combinatorial analysis) to atomic physics (where the atom acquires a proper, 'material' status at the same time when it gets its internal structure).

A second step in refining the general model suggests itself after the first one just described. On the basis of the notion of a configuration we may formulate a principle of conservation of basic genetic material. The genetic material inherent in configurations of strands are the quanta occurring in the different strands. The principle of conservation amounts to a conservation of these quanta. In other words, all the quanta occurring in strands of some configuration C_i of progeny have to be taken from the quanta occurring in the two configurations of the parents C, C', and conversely, all 'parental' quanta have to occur in some configuration of offspring. Somewhat more formally, this principle may be formulated as follows. Let $C, C', C_1, ..., C_n$ be configurations of strands such that COMBINATOR(C,C') $= \sum i = 1^n \alpha_i C_i$. These configurations have the form $\langle N, \mathbb{R}^3, \psi \rangle$, $\langle N', \mathbb{R}^3, \psi' \rangle$, $\langle\ N_1, \mathbb{R}^3, \psi_1 \rangle, ..., \langle N_n, \mathbb{R}^3, \psi_n \rangle$, respectively. Let s be a variable ranging over the strands occurring in $N, N', N_1, ..., N_n$, i.e. $s \epsilon N \cup N' \cup N_1 \cup ... \cup N_n$. For arbitrary such s let $Q(s)$ denote the set of quanta occurring in s. The principle of conservation then states that the union of all sets $Q(s)$ with s in N or in N'

$$\{Q(s)/s \text{ in } N \text{ or } s \text{ in } N'\}$$

and the union of all sets $Q^*(s^*)$ with s^* in N_1 or...or in N_n

$$\{Q^*(s^*)/(s^* \text{ in } N_1) \text{ or...or } (s_* \text{ in } N_n)\}$$

are identical. Often this principle is not strictly satisfied, deletion may occur as well as insertion. Mutation, which might be seen as a combination of deletion and insertion, will also contradict that principle of conservation. Insertion does not create a problem as long as the material inserted is present in the parental genetic material. The set of quanta in progeny in this case is not enlarged

by insertion. On the other hand, in the case of deletion the set of quanta in offspring may get smaller. In order to deal with deletion a weaker form of conservation stating that all quanta of offspring are included among the parental quanta, has to be used. In other words, the principle says that $\{Q^*(s^*)/(s^*$ in $N_1)$ or...or $(s^*$ in $N_n)\}$ is a subset of $\{Q(s)/s$ in N or s in $N'\}$. In the following, we will understand the principle of conservation always in this weaker form. The objection that mutation still provides counterexamples does not seem convincing. Mutation does not provide intended systems for the basic form of molecular genetics considered here. On the contrary: mutation transcends the frame given by the present genetic theories. Its proper treatment is achieved in a different theory, the theory of evolution. We certainly do not want to exclude the theory of evolution from using elements of the genetic models presented here but the phenomenon of mutation shows that the theory of evolution cannot simply be regarded as a refinement of molecular genetics. These are two different though closely related theories the proper relation of which is an important subject but not treated in this book.

A model of genetics in which genotypes have the form of configurations of strands, and which satisfies the (weak) principle of conservation of the basic genetic material we call a *model with material genetic basis*. Such a model will have to be used whenever the genetic material is analysed at the level of chromosomes or at the finer cyto-chemical level. The unrefined model, on the other hand, is restricted to cases in which COMBINATOR has an entirely theoretical status.

Any model with material genetic basis already contains features of combination kinematics. For COMBINATOR transforms parental configurations of strands (representing sets of chromosomes) into those of progeny, that is, COMBINATOR describes a kinematics of chromosomes. It does not seem satisfactory, however, to identify combination kinematics with the refined models with material genetic basis for two reasons. First, in contrast to the kinematical situation in physics we cannot expect a deterministic function expressing how parental configurations transform into those of progeny. Due to the probabilistic features at the level of phenotypes we have to expect similar probabilistic relations at the level of genotypes. Intuitively, we would like to reserve the label of combination kinematics to models dealing with the transition from just one pair of parental chromosomes to one definite chromosome of the offspring so that the transition might be -in principle- clarified by future investigation to an ever increasing degree. Ultimately, the picture we still have in mind is that of some deterministic connection governing meiosis and fertilization. The model thus contemplated, describing just one isolated transition of configurations of strands, is compatible with the models with material genetic basis. Starting from the equation

$$\text{COMBINATOR}(C, C') = \sum \alpha_i C_i$$

we may fix one particular C_i and consider the transition from C and C' to C_i as giving rise to a more local 'model of combination kinematics'. Under

this perspective a model with material genetic basis may be regarded as being 'composed of' n different such local 'models of combination kinematics'.

A second reason to keep combination kinematics distinct from models with material genetic basis is this. The most important application of combination kinematics at the moment is to recombination which occurs during meiosis. As this process does not involve the distinction between parents and progeny, the models of combination kinematics also should be neutral with resepct to this distinction. As it happens, the kinematics of recombination also has the form of a transition from two 'initial' to one 'final' entity. In a model with material genetic basis the initial entities are the parental strands, the final entity is a strand occurring in progeny. Analogously, the two initial entities in recombination are two pairs of chromatids (in the diploid case), and the final entity is the pair of chromatids resulting from these by recombination. In order to cover both kinds of applications we have to avoid the terminology of parental strands and strands in offspring and use a more abstract terminology. We simply will speak of two *initial* configurations of strands, and one such *final* configuration. In most recombination applications, the initial configuration consists of the four strands making up two pairs of chromatids.

In order to obtain a precise definition of a combination kinematics, let us reflect on what items necessarily have to be incorporated. First, as already mentioned, we have two initial configurations of strands forming an 'initial state', and one such configuration as a 'final state'. It is not difficult to take the two initial configurations together to form one bigger common initial configuration serving as the initial state. This 'joining' may be regarded as a purely conceptual operation, and in no way affects the spatial positions of the quanta occurring in the two strands. In recombination, we may conceptually put together the initial pairs of chromatides, and still have the same spatial configurations as before. In this way we have to deal with the transition from one initial to one final state. Between these two states there is a sequence of transitional states taken on during the process of recombination, meiosis, or fertilization. We cannot say that there is a continuous transition from the initial state to the final one, because phenomena like crossing over or deletion or insertion of genetic material represent discontinuities. It seems better to consider a discrete sequence of transitional states. Obviously, the order of this sequence is that of time. The transition is a process in time which to some extent can be made directly visible. We therefore should include a time interval, or at least a finite, discrete set of instants, which are linearly ordered:

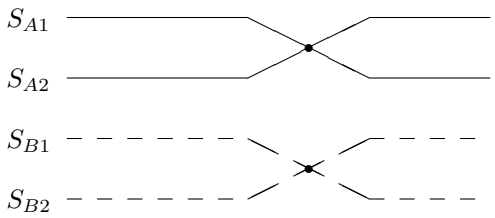
$$\{t_1, ..., t_r\} \text{ where } t_1 < ... < t_r$$

and $<$ is interpreted as 'later than'. In meiosis as well as in recombination proper the kinematic process involves a kind of reduction of the genetic material insofar as the final configuration of strands represents only a part of the material present in the initial configurations. In recombination this is due to the fact that only two of the four chromatids take part in the exchange, in meiosis in general it is due to the fact that the number of chromosomes is reduced by half. In abstract

terms we may deal with this feature just by ignoring part of the initial strands; those which are not involved in combination.

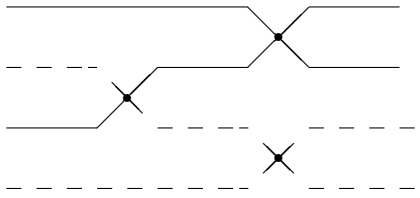
In recombination the process takes the following form. Initially two pairs of strands of homologous DNA or the chromatids of which they are part, synapse.

Fig.3-1



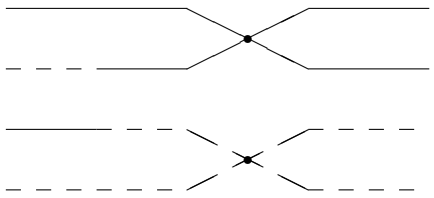
Chiasma formation, (crossing over), then occurs (see Figure 3-2 in which only one chiasma is illustrated, but many may occur).

Fig.3-2



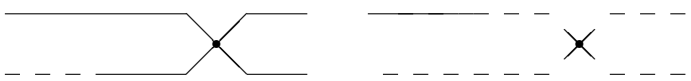
The structures then pull apart, thus forming two new genetic combinations (see Figure 3-3).

Fig.3-3



Anaphase of the first meiotic division then results in diploid cells of the genotypes shown in Figure 3-4.

Fig.3-4



Finally the second (reduction) division occurs, giving four haploid germ cells of the genotypes shown in Figure 3-5.

Fig.3-5

In woman, the primary oocyte gives rise to only one mature gamete, the other three 'polar bodies' degenerating, in man all four develop into mature gametes from the primary spermatocyte.

Primary oocytes remain in prophase and do not finish their first meiotic division before puberty is reached, then providing the germ line for the next generation. After puberty, the mature oocyte divides again meiotically and if fertilised it may produce a new individual. Progeny correspond to whichever one of the four genotypes present after second division actually matured. Fertilization may then occur between the spermatocyte and the oocyte, resulting in the conjugation of chromosome content.

We may formulate a version of the conservation principle stated above so that the quanta making up the final configuration have to be such as making up a distinguished subset of the strands in the initial configurations. The subset of strands distinguished in this way is the set of those involved in combination while all other strands 'are ignored' in the sense of not contributing to the final configuration.

As already noted it is convenient to join the two initial configurations to form one bigger configuration. Conceptually, we may just take the union of the sets of strands occurring in the two given configurations, and make sure that the positions of the quanta do not conflict in the sense that two different quanta from the two configurations do not get the same position. This situation might in fact occur if the frames of reference in the two configurations were badly chosen. We may avoid this situation by imposing a requirement on the two initial configurations. We say that two configurations of strands C, C' are *compatible* if the two 'position functions' ψ and ψ' from C and C' assign different values to any two different quanta. More sharply, we require that, for some given $\epsilon > 0$, for any two quanta q, q' occurring in C and in C' respectively, $|\psi(q) - \psi(q')| > \epsilon$. If two configurations $C = \langle N, \mathbb{R}^3, \psi \rangle, C' = \langle N', \mathbb{R}^3, \psi' \rangle$ of strands are compat-

ible we define *the union* $C \sqcup C'$ *of* C *and* C' by $C \sqcup C' = \langle N_*, \mathbb{R}^3, \psi^* \rangle$ where N^* is the union of N and N', and ψ^* is the union of ψ and ψ'. Using these auxiliary definitions we now can define the notion of combination kinematics.

x is a *combination kinematics* iff there exist $T, <, \mathbf{C}, \theta, C, C', C^*$ and N_0 such that
1) $x = \langle T, <, \mathbf{C}, \delta \rangle$
2) T is a finite set ($T = \{t_1, ..., t_r\}$) and $<$ is a linear order on T such that $t_1 < ... < t_r$
3) $\mathbf{C}$ is a set of r configurations of strands
4) θ is a function from T into $\mathbf{C}$, and one-one
5) C, C', and C^* are configurations of strands in $\mathbf{C}$, and C, C' are compatible
6) $\theta(t_1) = C \sqcup C'$ and $\theta(t_r) = C^*$
7) N_0 is a subset of the set of all strands occurring in $C \sqcup C'$
8) each strand occurring in C^* is a configuration of sub-strands occurring in N_0.

In other words, a combination kinematics essentially is a sequence $C \sqcup C' = C_1, ..., C_{r-1}, C_r = C^*$ of configurations of strands ordered in time by a function θ such that the strands of the 'last' configuration are obtained from a subset N_0 of the 'first' configuration by 'breaking them apart' and concatenating the parts anew. The intermediate transitions from C_1 to C_2 etc. are not specified any further. They may just consist in a change of the positions of the different strands and quanta relative to each other. Recombination proper, i.e. crossing over, breaking, and re-concatenation of strands, will usually occur only at one instant, it need not occur at all. The 'mechanism' according to which the process takes place is not made explicit in this definition. At the moment, it is not fully known. Its full description requires further items, like for instance the centromeres.

We may consider three special cases of combination kinematics. First, the *regular* case in which the final strands in C^* are composed of parts of the strands from N_0 such that all the material present in the latter is used up.

x is a *regular* combination kinematics iff $x = \langle T, <, \mathbf{C}, \theta \rangle$ is a combination kinematics with respect to C, C', C^* and N_0 and each strand occurring in N_0 is a concatenation of sub-strands of strands occurring in C^*.

Among the regular combination kinematics there are those in which just whole strands from N_0 are combined with each other. Another important regular case is that of recombination. The irregular cases may be divided into two further classes. First, there are cases with deletion. In these cases not all the genetic material, i.e. all the strands in N_0 are used up in forming the strands of C^*. During the process of formation of the strands of C^* part of the original strands in N_0 are not used, i.e. they are deleted. We have to require in this case that

no complete strands from N_0 are 'deleted' entirely. That is, each strand in N_0 at least contributes some sub-strand to those of C^*. Without this assumption N_0 might have been chosen 'too big'.

x is a combination kinematics *with deletion* iff $x = \langle T, <, \mathbf{C}, \theta \rangle$ is a
combination kinematics with respect to C, C', C^* and N_0 and
1) x is not regular
2) for each strand s in N_0 there is some sub-strand of s
occurring in a strand from C^*

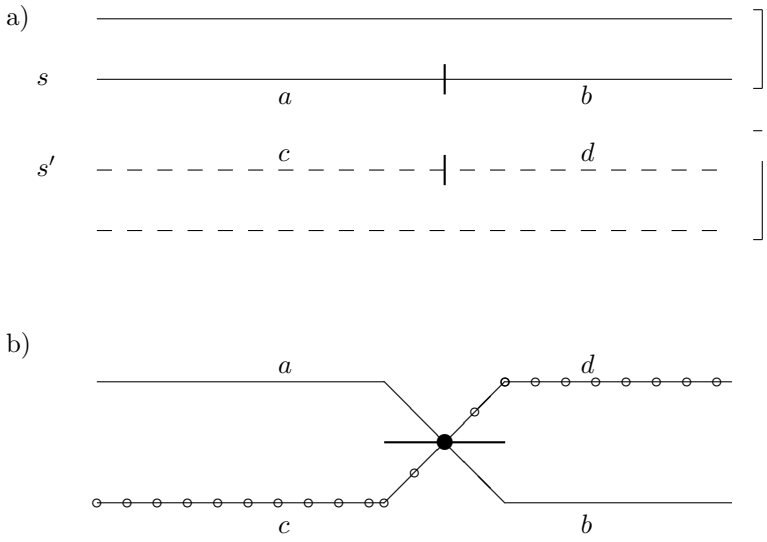
The third case is that of insertion. Here, N_0 has to be imagined as splitting up into two parts, N_0^+ and N_0^* where the material in N_0^+ is completely used up in the strands occurring in C^* while the remaining strands, those in N_0^*, are not entirely used up but provide material which is inserted in the formation of the strands in C^*. The decisive point is that the number of strands occurring in the final configuration is the same as in the set N_0.

x is a combination kinematics *with insertion* iff $x = \langle T, <, \mathbf{C}, \theta \rangle$ is
a combination kinematics with respect to C, C', C^* and N_0 and there
exists a partition of N_0 into two sets N_0^+ and N_0^* such that
1) the number of strands in N_0^+ is the same as in C^*
2) all quanta occurring in strands from N_0^+ also occur in the
strands in C^*
3) for each strand s in N_0^* there is some sub-strand of s occurring
in a strand in C^*

It is known that deletion and insertion go together with particular, favourable topological forms of the strands. Deletion may typically occur when a strand contains a 'loop-like' sub-strand, i.e. a sub-strand the ends of which are very close together in space. The precise causes of such phenomena however are not fully understood yet -eventhough the chemical conditions under which strands may join and break are.

Two textbook examples may serve to illustrate this model. Consider, first, an abstract schema of recombination, say, of the form depicted in Figure 3-6 below.

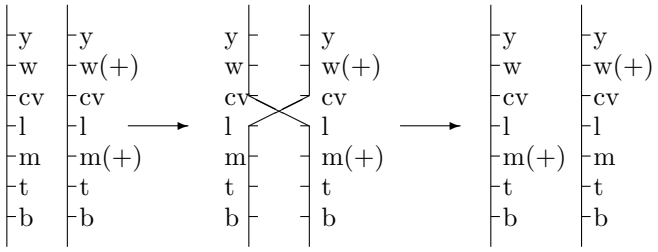
Fig.3-6



Such a process may be modelled by taking $T = \{t_1, t_2\}, C$ and C' as the initial configurations depicted in a), C^* as the final configuration in b), and N_0 as the set of strands involved in the crossing over (s and s'). As a second example consider a less abstract case. Morgan[34] noted that when he mated a white miniature male Drosophila to a homozygous wild-type female, the first filial females were wild type as expected. The males could, however, be wild-type, white miniature, white or miniature. For independent assortment, second filial males should show such phenotypes equally. In fact 37.6% represented the new recombinant types white and miniature, while the rest were either wild type or white with miniature. The siutation is illustrated in Figure 3-7 in which a sample of other loci have been included for clarity. The exact point of crossover could in principle be anywhere between white and miniature, and phenotypes varying in other respects would then be observed.

[34](Morgan, 1911).

Fig.3-7



Key:

y	yellow body
w	white eyes
w(+)	white eyes absent (normal or wild type)
cv	crossveinless wings
l	lozenge wings
m	miniature wings
m(+)	miniature wings absent (normal or wild type)
t	tiny bristles
b	bar eyes

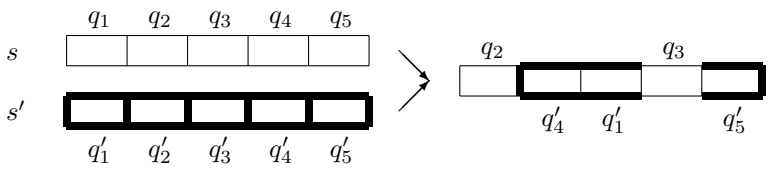
The most important kind of application of combination kinematics is in recombination studies. For this reason we want to further analyse the recombination case. This may best be done in a kind of classification of different types of combination kinematics corresponding to important branches of genetics. We distinguish four such types, three of which are directly connected to particular genetic applications. All types may be described as specialisations of the general model of combination kinematics above.

A first specialisation obtains in cases of *complete linkage*. Complete linkage means that the initial strands occurring in configurations C and C' of a combination kinematics are not altered during the process of combination. So the final strands in C^* are identical with strands in N_0. What may change is just the configuration of those strands. In Figure 3-8 a simple case is depicted.

Fig.3-8

A second specialisation deals with the opposite extreme. This case does not seem of much relevance in genetics. It is of merely conceptual interest. We speak of *unrestricted combination* here. The idea is that the ordering of the quanta in the strands is completely irrelevant during the process of combination. Thus each quantum from some initial strand in C or C' may occur at whatever position in each final strand of C^*. This would be a completely combinatorial account which might be further substantiated by adding statistical hypotheses about the distributions of quanta in the initial and final strands or configurations. A simple schematic example is shown in Figure 3-9.

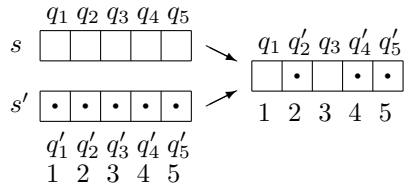
Fig.3-9



A case in-between the two extremes just introduced is that of *Mendelian* combination. This kind of combination uses the positions in which the quanta occur on the strands. All strands involved in the combination kinematics are supposed to have the same 'length', i.e. the same number of quanta, and thus the same number of positions. The combination of quanta during the transition from initial to final strands is restricted to the respective positions at which the quanta occur. The quantum occurring at position number j, say, in a final strand has to be one of those occurring at the same position in one of the initial strands. Besides this constraint quanta may be combined freely, that is, no linkage has

to obtain. Two initial strands thus may be completely 'mixed up' as shown in Figure 3-10. A fuller account of *Mendelian* combination will be given in the next chapter.

Fig.3-10



The fourth and most important specialisation is that of recombination occurring. Recombination consists in one or several occurrences of crossing over. The general case may best be defined recursively. We first describe a situation with just one crossing over, and then iterate the definition.

Let $C = \langle\{x, y\}, \mathbb{R}^3, \psi\rangle$ and $C' = \langle\{u, v\}, \mathbb{R}^3, \pi'\rangle$ be configurations of strands. We say that C' *is obtained from* C *by simple* (or *1-fold*) *crossing over* iff there exist strands a, b, c, d such that

i) $x = a \circ b$ and $y = c \circ d$
ii) $u = a \circ d$ and $v = c \circ b$.

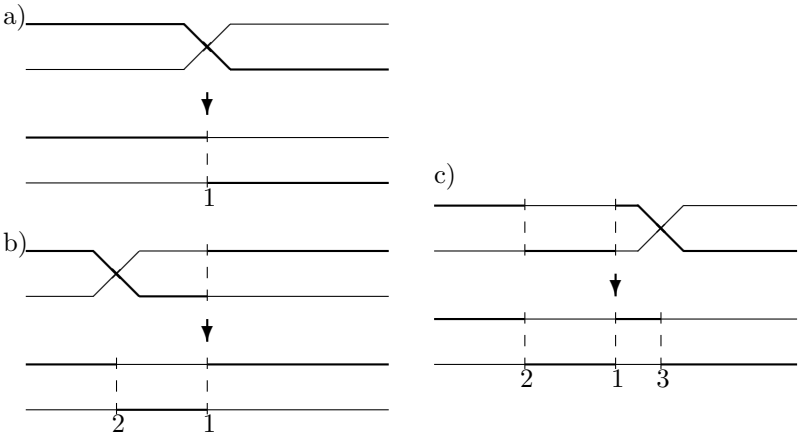
Now let $C = \langle\{x, y\}, \mathbb{R}^3, \psi\rangle$ and $C' = \langle\{u, v\}, \mathbb{R}^3, \psi'\rangle$ be configurations of strands. We say that C' is obtained from C *by (n+1)-fold crossing over* iff there exists a configuration of strands $C^* = \langle\{x_1, x_2\}, \mathbb{R}^3, \psi^*\rangle$ such that

i) C^* is obtained from C by n-fold crossing over
ii) C' is obtained from C^* by 1-fold crossing over.

We say that a combination kinematics *contains recombination* iff at least one sub-configuration of a configuration in C^* consisting of two strands is obtained from corresponding sub-configurations of the initial configurations by n-fold crossing over for some n.

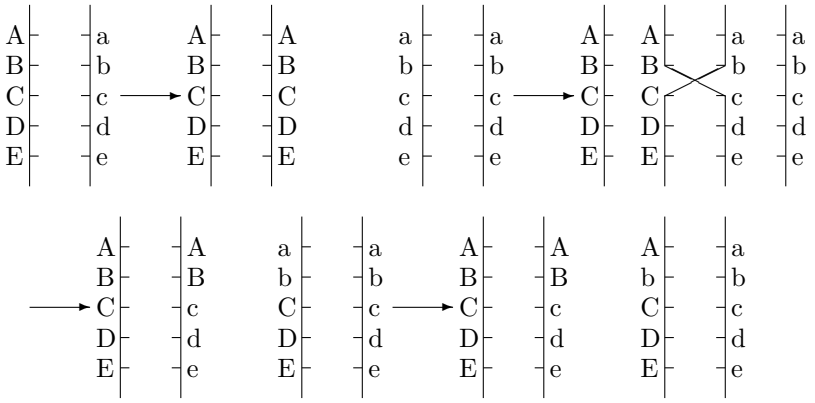
In Figure 3-11 a case of 3-fold crossing over is depicted in its three states relevant to the definition.

Fig.3-11



In general, crossing over may occur not only between two strands but between more than two. Anderson[35] demonstrated that crossing over could occur during the four stranded stage of meiosis. He used the attached X-chromosome mutant of Drosophila and studied the Bar locus. Figure 3-12 schematises such crossing at the four strand stage. A similar schema might be provided for polyploidy.

Fig.3-12



(after Strickberger 1964)

As such phenomena are studied only rarely, a fully general definition need not be given here. Instead, let us briefly look at the case of four strands being involved. The generalisation of 1-fold crossing over is this. Let $C = \langle \{x_1, ..., x_4\}, \mathbb{R}e^3, \psi \rangle$ and $C' = \langle \{y_1, ..., y_4\}, \mathbb{R}^3, \psi' \rangle$ be configurations of strands. Then C' is obtained from C by 1-fold crossing over iff there exist strands $a_1, ..., a_4, b_1, ..., b_4$ such that

[35](Anderson, 1925).

i) $x_i = a_i \circ b_i$ for $i = 1, ..., 4$
ii) $y_i = a_{j_i} \circ b_{j_i}$ for $i = 1, ..., 4$ (where repetition of indices is not allowed).

Similarly, the notion of (n+1)-fold crossing over may be defined.

In practice, an alternative definition is often used which is slightly stronger than the one just given in that it excludes any strand resulting in crossing over from being identical with one of the initial strands. This definition is easier to apply, for on the phenotypic level it is of course impossible to detect whether crossing over of this kind, in which the final strand is identical with one of the initial strands, has occurred. The stronger definition may be stated as follows. Let $C = \langle N, \mathbb{R}^3, \psi \rangle$ and $C' = \langle N', \mathbb{R}^3, \psi' \rangle$ be configurations of strands, and s be a strand. We say that s *is new with respect to* C and C' iff s neither occurs in C nor in C'. We say that C' is obtained from C by *strong* simple (1-fold) crossing over (and that C' is a simple recombination of C) iff there is a strand s in N' which is new with respect to C and C'.

When joined to the above definition of a combination kinematics, the new strands have to consist of quanta occurring in the initial strands. By this feature the notion of 'new' strands is restricted to reasonable combinations.

Chapter 4

Transmission Genetics

Roughly, transmission genetics comprises those applications in which inheritance among populations (as contrasted to individuals proper) is studied without the necessary use of chemical means. The central methodology is to establish data for MATOR, that is, for the probabilities of different PHENOTYPES occurring in the progeny. These data are systematized and explained by means of genetic hypotheses involving reference to COMBINATOR and DETERMINER.

We begin by introducing a general model for diploid cases which may easily be extended to arbitrary ploidy. Our model is obtained by specialising the general model of Chap.2. The manner of specialisation affects all levels of the model.

A first special assumption of transmission genetics concerns the level of MATOR which provides the 'ontology', the entities possessing the phenotypes. Here, transmission genetics is definitely committed to populations, for mere consideration of an individual mating does not yield reliable frequencies of traits in the offspring. In most cases the offspring of individual parents will not even exhaust all phenotypes which could possibly arise, not to speak of providing reliable relative frequencies for the occurrence of these phenotypes. We treat populations as non-empty sets of 'individuals' the nature of which does not really matter, for PHENOTYPES are assigned to populations rather than to individuals in the model. Of course, the real carriers of phenotypes are individuals but the models become slightly simpler the way we proceed. So PARENT and PROGENY in transmission genetics will always refer to populations of parents and offspring. Accordingly, APPEARANCE assigns PHENOTYPES to populations.

Though populations in this way seem to acquire the status of proper objects -things to which attributes are ascribed- their identification proceeds from the opposite direction. Populations are sets of individuals possessing the same phenotype. If considered in progeny they even may be taken as maximal in this respect.

In order to determine relative frequencies of an expression in offspring one has to count the numbers of individuals exemplifying that expression. Using the notion of populations this procedure may be described as follows. We consider all populations in which (i.e. in whose phenotypes) this expression occurs, and count the sizes of these populations. By adding up the numbers thus obtained we get the desired number of occurrences of the trait in progeny. The size of a population is just the cardinality of the set of individuals which make up the

population. Such cardinalities we indicate by $\|\ \|$ so that $\|$PARENT_1$\|$, for instance, denotes the number of individuals occurring in the set PARENT_1.

The next items to be specialised are the PHENOTYPES. In transmission genetics, these may be represented in a simple schematic form. A PHENOTYPE is defined as a combination, more precisely as a tuple, of EXPRESSIONS. We assume a fixed number, k, of EXPRESSIONS making up the phenotypes of one model. Thus each PHENOTYPE has the form

$$\langle\ \text{EXPRESSION\_1,...,EXPRESSION\_k}\ \rangle$$

The expressions possible in a model are collected in a set of expressions. The set of k-tuples of elements of this set may be regarded as a space of possible PHENOTYPES. However, only a few of those possibilities are ordinarily used, often this space is further restricted by reference to traits or characters.

Characters are classifications, i.e. sets, of EXPRESSIONS. If characters are used, a particular position in a PHENOTYPE may be restricted to be filled by EXPRESSIONS of this character only. Consider the paradigm of Mendel's experiments studying the well known expressions: smooth or wrinkled form of seed, yellow or green colour of seed, grey or white colour of seed's coat, full or constricted form of pods, green or yellow colour of pods, place of flower and pods along the stem or on top of it, and long or short stem. Each pea plant will exemplify exactly seven such EXPRESSIONS, so PHENOTYPES are taken to consist of 7-tuples over this 14-element set of expressions. The set of expressions is naturally classified into seven characters: form of seed, colour of seeds etc., and the PHENOTYPES may be restricted to combinations of EXPRESSIONS in which just one EXPRESSION is taken from each character. Thus, the first position of PHENOTYPE might be reserved to expressions of seedform, that is, if PHENOTYPE = $\langle$ EXPRESSION_1,...,EXPRESSION_7 $\rangle$ then EXPRESSION_1 has to be SMOOTH or WRINKLED.

As will be seen below in connection with linkage genetics it is not always useful to be too strict about characters, so we do not introduce them as primitives (though we acknowledge them to be very helpful in many applications, of course). Also, it has to be noted that the combinations of EXPRESSIONS constituting the PHENOTYPES are by no means required to be complete. The richness of a concrete phenotype always enforces some choice for systematic treatment, and in many cases one is simply not interested in many of the EXPRESSIONS that might be differentiated and studied. The set of expressions as well as the number indicating the number of EXPRESSIONS making up one PHENOTYPE are often chosen rather coarse and small -as most convenient to the respective application.

DISTRIBUTOR correspondingly operates on PHENOTYPES of this format. Any 'parental' pair of k-tuples of EXPRESSIONS is mapped into a distribution of phenotypes which according to our conventions we may write

$$\langle \alpha_1\ \text{PHENOTYPE\_OF\_PROGENY\_1,...,}\alpha_r\ \text{PHENOTYPE\_OF\_PROGENY\_r}\ \rangle$$

where $\sum_{i=1}^{r} \alpha_i = 1$ and each PHENOTYPE_OF_PROGENY_i again is a tuple of EXPRESSIONS of the above format. These PHENOTYPES and their transmission as described by DISTRIBUTOR constitute the data which are theoretically systematized at the level of genotypes. These data are not purely observational. In many cases the expressions may be determined by direct observation, but in other cases (for instance in quantitative characters, like size) even the expressions have to be determined by means of acknowledged procedures. More obviously, the coefficients α_i are not observable. They represent relative frequencies of occurrences of a corresponding phenotype in total progeny, and thus have to be determined by determining and counting expressions, and calculating ratios.

It is important to understand the connection in application between phenotypes on the one hand and populations and frequencies on the other hand. When phenotypes are not required to represent all features really present in the individuals but only those of interest in a given application there is an implication for the notion of a population: a given set of concrete individuals may be partitioned into populations in different ways. This depends on how complete and how fine we choose the set of characters, and consequently, how many expressions occur in one phenotype. In one application we may be interested, for instance, in eye-colour and wing-shape as the only characters each of which with two expressions, say black-red and normal-short. A set of individuals as shown in Figure 4-1 then gives rise to four populations A,B,C,D, characterised, respectively, by $\langle blackeyes, normalwings\rangle$,..., $\langle redeyes, shortwings\rangle$. In another application we may be interested *only* in eye-colour and decide to study just one character (eye-colour), say, again with two expressions (black-red). We then have only two populations, namely A ∪ C and B ∪ D in Figure 4-1.

Fig.4-1

colour (eye) / wing shape	black	red
normal	A	B
short	C	D

The model in this sense may be said to be homogenous with respect to the choice of characters. This does not mean that the empirical findings as expressed in hypotheses will be the same irrespective of how the characters are chosen. We only want to emphasize the different possibilities of application (which may result in different hypotheses).

This homogeneity becomes important as soon as the notion of chromosome comes into play. For in connection with chromosomes one is tempted to imagine a PHENOTYPE as comprising all relevant expressions an individual possesses on the basis of its chromosomes. But this picture represents just one out of var-

ious possibilities. More frequently, there are just two or three expressions under study -as in crossing over experiments. We may then decide to use phenotypes to contain only those expressions, and consequently, to differentiate populations according to the difference in these expressions.

The theoretical machinery used in order to explain the frequencies in distributions of phenotypes actually observed in principle was introduced by Mendel: every expression is represented by a sequence of abstract *factors* and some definite hypothesis is put forward about how these factors have to be processed so that the resulting factors and their coefficients yield some acceptable representation of the observed distributions. In our model this means specifying the GENOTYPES to the form of sequences of factors and then stating an explicit definition of COMBINATOR corresponding to a hypothesis of how transmission takes place.

Given the format of PHENOTYPES and the assumption that each expression is theoretically represented as a sequence of factors, there is a straightforward way to specialise GENOTYPE. Each GENOTYPE has the form of a sequence

$$\langle \theta_1, ..., \theta_k \rangle$$

where k matches with the index used for PHENOTYPES, and where each θ_i is a sequence of *allelic factors*. Since we want to restrict the details to the diploid case, each sequence θ_i will be just a pair of *factors*, or *alleles* as it is said:

$$\theta_i = \langle a^i, b^i \rangle.$$

We note that the order implicit in writing down a GENOTYPE as a sequence $\langle \theta_1, ..., \theta_k \rangle$ rather than a mere set is not really necessary in all variants of transmission genetics. In *Mendelian* genetics, for instance, we could get along with a mere set, but at the cost of formal complication. In those cases, treating GENOTYPES as sequences may be regarded as a mere matter of convenience. In contrast to this, the allelic pairs δ_i necessarily have to be pairs, and not sets. For there is no difference between the sets $\{a\}$ and $\{a, a\}$, but there is a difference between two identical allelic factors being present, or only one of them. Thus in ordinary dominance $\langle a, a \rangle$ will give the recessive character, while $\langle A, a \rangle$ will give a dominant character. Furthermore, the triploid $\langle a, a, a \rangle$ will in general give rise to a different phenotype than either the diploid $\langle a, a \rangle$ or the haploid $\langle a \rangle$, if indeed all are viable.

In analogy with PHENOTYPES we write

$$\text{GENOTYPE} = \langle\langle\ \text{FACTOR}(1,1),\text{FACTOR}(1,2)\ \rangle, ..., \langle\ \text{FACTOR}(k,1), \text{FACTOR}(k,2)\rangle\rangle$$

with FACTOR(i, j) for the j-th factor ($j = 1, 2$) occurring in the i-th section δ_i. By collecting all FACTOR(i, j) occurring in a model for a given i we obtain a set, which we call SET_OF_FACTORS_i. The SETS_OF_FACTORS_i define a space of tuples of factors from which the GENOTYPES are chosen. A standard

configuration of GENOTYPES is given for $k = 2$, that is when there are just two SETS_OF_FACTORS, denoted by $\{A, a\}$ and $\{B, b\}$. GENOTYPES over these SETS_OF_FACTORS have the form $\langle\langle A, A\rangle, \langle B, B\rangle\rangle$, $\langle\langle A, A\rangle, \langle B, b\rangle\rangle$, $\langle\langle A, A\rangle, \langle b, B\rangle\rangle$ etc.

With these specific forms of PHENOTYPE and GENOTYPE it becomes possible to set up precise requirements for COMBINATOR and DETERMINER. The general assumption for DETERMINER (in the diploid case) in transmission genetics is this. DETERMINER maps GENOTYPES into PHENOTYPES such that each allelic pair yields a unique EXPRESSION_i. In other words, there are functions DET_1,...,DET_k such that each DET_i maps allelic pairs into expressions

$$\text{DET\_i}(\theta_i) = \text{EXPRESSION\_i}$$

and DETERMINER is defined as the tuple of all DET_i as follows:

(3) $\text{DETERMINER}(\theta_1, ..., \theta_k) = \langle\ \text{DET\_1}(\delta_1),...,\text{DET\_k}(\delta_k)\rangle$

where the latter expression upon evaluation yields some PHENOTYPE: $\langle$EXPRESSION_1,...,EXPRESSION_k$\rangle$.

In Chap.2 the property of being decomposable was generally discussed for DETERMINER. It is easy to see that (3) is an instance of decomposability. Accordingly, the allelic sequences θ_i which make up the GENOTYPES may be regarded as genes as soon as there is some indication for their being materially distinguishable. It is worth repeating that originally, at the time of Mendel's experiments no such indication was at hand. His pairs of factors therefore may be called genes only with hindsight.

Finally, the most interesting specialisation occurs for COMBINATOR. We have to state how COMBINATOR operates on two given GENOTYPES of the form

(4) $\langle\langle\ \text{FACTOR}(1,1),\text{FACTOR}(1,2)\ \rangle, ...,\langle\ \text{FACTOR}(k,1),\ \text{FACTOR}(k,2)\ \rangle\rangle$

In order to keep things legible let's represent these sequences in the form

$$g_1^* = \langle a^1, b^1, ..., a^k, b^k\rangle,\ g_2^* = \langle c^1, d^1, ..., c^k, d^k\rangle,$$

respectively. Out of these two sequences, COMBINATOR has to produce a distribution $\langle\alpha_1 g_1, ..., \alpha_s g_s\rangle$ where each g_i again is a sequence of the form (4). If we denote an arbitrary g_i by $\langle e^1, f^1, ..., e^k, f^k\rangle$ then there are two general requirements g_i has to satisfy in order to occur in a distribution of the desired form. The first requirement refers to the SET_OF_FACTORS_i introduced before, and states that, for the factors e^i, f^i, both should be elements of the SET_OF_FACTORS_i. This requirement expresses independent assortment of factors. The factors occurring in the i-th position of a GENOTYPE for offspring have to come from a small SET_OF_FACTORS_i, and no other set. So each position has its corresponding SET_OF_FACTORS, and the factors occurring there

are not affected by factors of 'other kinds', i.e. from SETS_OF_FACTORS_j with $j \neq i$. Note that this requirement is implicit, 'built in' in the syntax. A second requirement is that δ_i should consist only of factors actually occurring in the parental GENOTYPES g_1^*, g_2^*. That is, each e^i and f^i is one of the a^i, b^i, c^i, d^i. This is an instance of the conservation principle discussed in Chap.3. Roughly, it may be rephrased as stating that the genetic material is a stable 'genidentical' entity: no new factors enter the scene in the course of transmission.

If we denote the SET_OF_FACTORS_i by F_i then, clearly, each GENOTYPE is an element of the cartesian product

$$\mathbf{F} = (F_1 \times F_1) \times ... \times (F_k \times F_k).$$

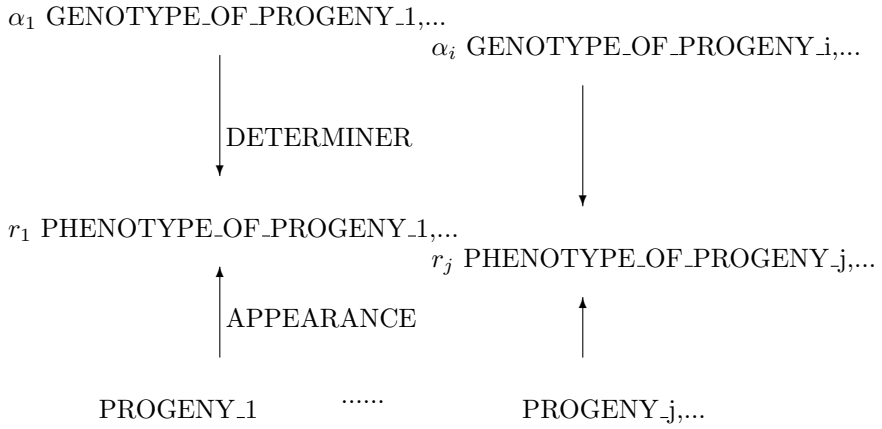
Using the notion of a γ-distribution over $\mathbf{F}$ we may say that COMBINATOR is a function in the set $\mathbf{D(F)}$ of all Γ-distributions over $\mathbf{F}$:

$$\text{COMBINATOR: } \mathbf{F} \times \mathbf{F} \rightarrow \mathbf{D}(\mathbf{F}).$$

With some redundancy in notation each value of COMBINATOR may also be written in the form $\langle \alpha_1\gamma_1, ..., \alpha_s\gamma_s \rangle$, where $\alpha_i \epsilon \mathbb{R}, \alpha_i \geq 0, \Sigma\alpha_i = 1$, s is some natural number, and $\gamma_i \epsilon \mathbf{F}$. Strictly speaking the numbers α_i are the values of a Γ-distribution COMBINATOR(γ, γ'). The first requirement above is then met automatically, by the definition of $\mathbf{F}$.

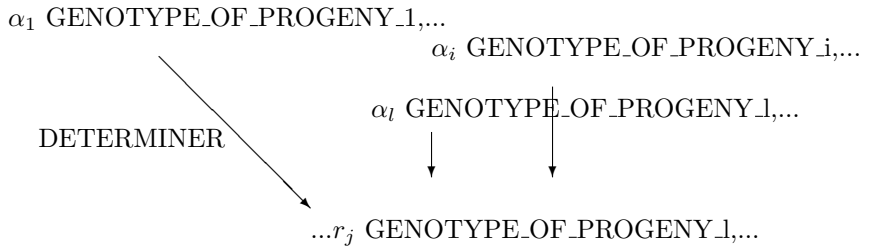
The model being specified thus far, we may consider in more detail the claim (2) stated in Chap.2 associated with it in applications, or in other words, the central axiom distinguishing proper models from structures in which the primitives are realized 'by chance'. In Chap.2 this claim was formulated in broad terms as a claim of fit between the coefficients $\alpha_1, ..., \alpha_s$ and $r_1, ..., r_k$ occurring in the distribution of genotypes and the distribution of phenotypes. We now can specify the origin of the relative frequencies $r_1, ..., r_k$ *within* the model, whereas in Chap.2 this was entirely a matter of interpretation external to the model. The specification of the r_i is in terms of sizes of populations in the offspring. We may use DETERMINER to match genotypes of progeny, and APPEARANCE to match the latter with populations. In this way we might proceed as suggested by Figure 4-2.

Fig.4-2



We count β_j =‖PROGENY_j ‖, the number of individuals in the j-th population, and by taking its ratio to the overall number of progeny,
μ =‖PROGENY_1 ‖ + ...+ ‖PROGENY_k ‖, obtain the relative frequency $r_j = \beta_j/\mu$ of occurrences of phenotypes of kind j, that is, phenotypes assigned to PROGENY_j by APPEARANCE. It is tempting now to use DETERMINER in a similar way to relate the frequency r_j to its theoretical counterpart α_i, namely to look for that GENOTYPE_OF_PROGENY_i which by DETERMINER is mapped on PHENOTYPE_OF_PROGENY_j, and to relate its coefficient α_i to r_j. This procedure overlooks, however, that the theoretical combinations, the new genotypes obtained by COMBINATOR, need not uniquely characterize one phenotype each. DETERMINER is not, and cannot be, required to be one-one. In other words, different GENOTYPES_OF_PROGENY produced by COMBINATOR might yield the same value under DETERMINER. Thus the situation as depicted in Figure 4-3 will occur.

Fig.4-3



We therefore must not match the α_i one-one with the r_j. There are too many α_i, in general, to allow for this. Rather, we have to group together all those α_i which give rise to the same PHENOTYPE_OF_PROGENY, and match the sum

of their coefficients with that of the resulting PHENOTYPE_OF_PROGENY. If, in the situation of Figure 4-3, besides the three GENOTPYES depicted there are no others in the model which are mapped on PHENOTYPE_OF_PROGENY_j then the correct match of the coefficients will be

$$\alpha_1 + \alpha_i + \alpha_l = r_j.$$

r_j is the relative frequency of individuals in the offspring having PHENOTYPE_OF_PRPOGENY_j. α_1, α_i and α_l are the weights of those genotypes which by DETERMINER are mapped onto PHENOTYPE_OF_
PROGENY_j, that is, those and only those genotypes theoretically produced by COMBINATOR from the parental ones which give rise to PHENOTYPE_j. So $\alpha_1 + \alpha_i + \alpha_l$ is the expected frequency of PHENOTYPE_j to occur in the offspring, expected on the basis of the particular forms of DETERMINER and COMBINATOR which produce the considered situation.

In order to obtain a general formulation let us introduce, for given PHENOTYPE_OF_PROGENY_j ($j \leq k$), and given (parental) GENOTYPE_1 γ and GENOTYPE_2 γ^*, the set $C(\gamma, \gamma^*, j)$ of all coefficients α_i in the distribution of genotypes COMBINATOR$(\gamma, \gamma^*) = \langle \alpha_1$GENOTYPE_OF_
PROGENY_1, ...,α_sGENOTYPE_OF_PROGENY_s $\rangle$ for which

DETERMINER(GENOTYPE_OF_PROGENY_i) = PHENOTYPE_OF_
PROGENY_j.

In Figure 4-3, $C(\gamma, \gamma_*, j)$ would be $\{\alpha_1, \alpha_i, \alpha_l\}$. The central axiom of fit to hold in models of transmission genetics may then be formulated in two ways. The first, simpler version is restricted to the two upper levels of the model:

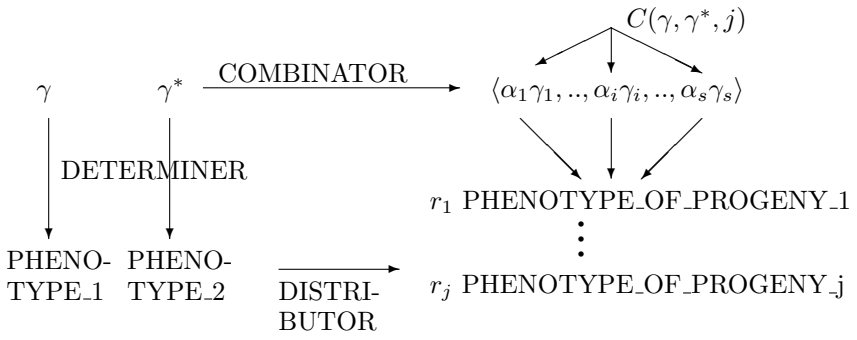
(5) If
- DISTRIBUTOR(PHENOTYPE_1,PHENOTYPE_2) =
$\langle r_1$ PHENOTYPE_OF_PROGENY_1,..., r_kPHENOTYPE_OF_
PROGENY_k$\rangle$
- DETERMINER(γ) = PHENOTYPE_1
- DETERMINER(γ^*) = PHENOTYPE_2
- COMBINATOR$(\gamma, \gamma^*) = \langle \alpha_1\gamma_1, ..., \alpha_s\gamma_s \rangle$
- $j \leq k$

then

$$\sum_{\alpha \epsilon C(\gamma, \gamma^*, j)} \alpha = r_j$$

The situation is depicted in Figure 4-4. The two levels fit, as required, if the sum of all α_i in $C(\gamma, \gamma^*, j)$ equals r_j.

Fig.4-4



In this formulation PHENOTYPE_1, PHENOTYPE_2, $r_1, ..., r_k$, PHENOTYPE _OF_PROGENY_1,...,PHENOTYPE_OF_PROGENY_k, $\gamma, \gamma^*, \alpha_1, ..., \alpha_s, \gamma_1, ..., \gamma_s$ and j have the status of variables. The axiom requires that for all instances of the premisses which are possible in the model, the conclusion holds true. In particular, the equation among coefficients has to hold for *all* $j \leq k$. Note that j occurs also at the left hand side of the final equation, namely in the definition of $C(\gamma, \gamma^*, j)$.

A second, more elaborated version is obtained when we include the first level, and define relative frequencies r_j in terms of sizes of populations.

(6) If

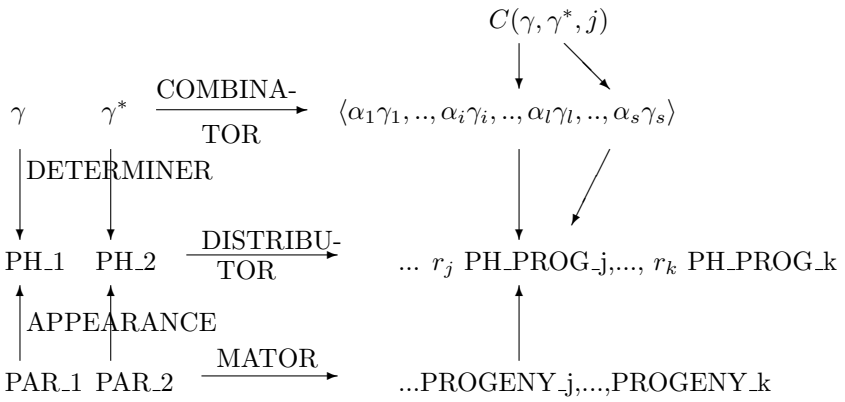
- PARENT_1 and PARENT_2 are populations
- MATOR(PARENT_1,PARENT_2) = $\langle$ PROGENY_1,...,PROGENY_k $\rangle$
- $j \leq k$
- APPEARANCE(PROGENY_j) = PHENOTYPE_OF_PROGENY_j
- γ, γ^* are GENOTYPES
- DETERMINER(γ) = APPEARANCE(PARENT_1)
- DETERMINER(γ^*) = APPEARANCE(PARENT_2)
- COMBINATOR(γ, γ^*) = $\langle \alpha_1\gamma_1, ..., \alpha_s\gamma_s \rangle$

then

$$\sum_{\alpha \epsilon C(\gamma,\gamma^*,j)} \frac{\|PROGENY\_j\|}{\|PROGENY\_1\|+...+\|PROGENY\_k\|}.$$

The right hand side of the equation is just the *definiens* for the relative frequency r_j. Thus the corresponding picture is obtained by extending Figure 4-4 on the bottom.

Fig.4-5



In summary, a general model of transmission genetics is a general model as described in Chap.2 which satisfies the following additional axioms:

There exist a number k and sets SET_OF_FACTORS_i for all $i \leq k$
such that
AT1 each PHENOTYPE has the form $\langle$ EXPRESSION_1,...,
EXPRESSION_k $\rangle$
AT2 each GENOTYPE has the form
$\langle\langle$ FACTOR$(1,1)$,FACTOR$(1,2)$ $\rangle$, ..., $\langle$ FACTOR$(k,1)$,FACTOR$(k,2)\rangle\rangle$
where for all $i \leq k$ and $j = 1, 2$, FACTOR(i, j) is an element of
SET_OF_FACTORS_i
AT3 there exist functions DET_1,...,DET_k such that DETERMINER can
be written in the form
DETERMINER($\langle$ FACTOR$(1,1)$,...,FACTOR$(k,2)$ $\rangle$ =
$\langle$ DET_1(FACTOR$(1,1)$,FACTOR$(1,2)$),..., DET_k(FACTOR$(k,1)$,
FACTOR$(k,2)$ $\rangle$ =
$\langle$ EXPRESSION_1,...,EXPRESSION_k $\rangle$
AT4 COMBINATOR is such that for all $\gamma, \gamma^*, \alpha_1\gamma_1, ..., \alpha_s\gamma_s$:
if COMBINATOR$(\gamma, \gamma^*) = \langle \alpha_1\gamma_1, ..., \alpha_s\gamma_s \rangle$ then
for all $i \leq s$ all components of γ_i are among the components of
γ and γ^*
AT5 the basic axiom of fit holds in one of the two forms (5) or (6)
above, for all $r_1, ..., r_k$,
all PHENOTYPE_OF_PROGENY_1,...,PHENOTYPE_OF_
PROGENY_k,
all γ, γ_*, all $\alpha_1, ..., \alpha_s, \gamma_1, ..., \gamma_s$,
all PARENT_1, PARENT_2, PROGENY_1,...,PROGENY_k , and
all PHENOTYPE_1, PHENTOYTPE_2 for which DISTRIBUTOR
is defined, and which occur in the system.

We note that our account for ploidy $p = 2$ can be extended without essential change to ploidy $p > 2$. The only change is in the form of GENOTPYES which for arbitrary p are tuples of allelic sequences of factors of length p (instead of 2). That is, each GENOTYPE takes the form

$$\langle\langle \text{ FACTOR}(1,1),...,\text{FACTOR}(1,p) \ \rangle, ..., \langle FACTOR(k,1), ..., FACTOR(k,p)\rangle\rangle.$$

Note further that axiom AT5 is not implied by the general form of the axiom of fit stated in (2) in Chap.2. Whereas in (2) the connection between the coefficients occurring in the distributions of phenotypes and genotypes is left unspecified, in AT5 it is spelled out in a more special way.

Two further remarks may be added. First, one might contemplate the possibility of further relaxing the requirement of independent assortment as expressed in the form of GENOTYPES and of COMBINATOR by means of the different SETS_OF_FACTORS_i. This would yield a model with one single overall set of factors which might be relevant in complex, simultaneous combinations of different expressions. In fact, such situations occur, in fine structure genetics, for example, the semi-dominant 'bar' locus in Drosophila.[36] We believe, however, that the principle of independent assortment provides a core principle for versions of *classical* transmission genetics of Mendelian and non-Mendelian types. Relaxation of this principle opens a new field and should be conceptually distinguished by giving rise to another model. Essentially, the classical model as just described assumes a relation

one gene - one expression

which goes together with independent assortment. If the latter assumption is given up, the former will be difficult to maintain.

Second, we did not require independent segregation. This formally would amount to requiring that COMBINATOR, in fact, produces all possible combinations of a certain type with equal probability coefficients. This special case is typical for Mendelian models proper.

A first submodel of transmission genetics is thus provided by *Mendelian* genetics. This model is obtained by imposing further requirements on COMBINATOR which, in fact, amount to an explicit definition, and on DETERMINER. In addition to the stipulations for the general model there are four new features. First, all possible combinations of parental factors -combinations within one SET_OF_FACTORS, to be sure- have to be considered, and all these combinations are equally probable. This amounts to Mendel's law of independent segregation. Second, we now have an explicit restriction to the diploid case (but still the definition of COMBINATOR to be given below can be extended to general ploidy). Third, we have to introduce characters in more detail. There are exactly k CHARACTERS corresponding to the length k of

[36](Sturtevant and Morgan, 1923), (Sturtevant, 1925).

the tuple representing PHENOTYPES. Each CHARACTER_i, $i \leq k$, is a set containing exactly two EXPRESSIONS, denoted by EXPRESSION$(i,1)$ and EXPRESSION$(i,2)$. Fourth, we need the distinction between dominant and recessive factors. This is introduced by a stipulation for DETERMINER. It is required that for any two EXPRESSIONS of one CHARACTER there are two corresponding unique FACTORS, both from the same SET_OF_FACTORS such that exactly one pair of these two FACTORS yields one EXPRESSION (the recessive one) while the three other pairs that can be formed from the two FACTORS yield the other EXPRESSION. The four pairs of two factors a, b are, of course, $\langle a,a\rangle, \langle a,b\rangle, \langle b,a\rangle, \langle b,b\rangle$. For a precise definition of this requirement it is convenient -though not strictly necessary- that DETERMINER is decomposable, as assured in the general transmission model.

Formally, these requirements are expressed by defining COMBINATOR in an explicit way. To this end we have to introduce a formal operation of multiplication for distributions of genotypes. We agree that the *concatenation* of two tuples

$$\gamma = \langle x_1, ..., x_n\rangle, \gamma^* = \langle y_1, ..., y_m\rangle, \text{ denoted by } \gamma\gamma^*,$$

is simply the tuple

$$\langle x_1, ..., x_n, y_1, ..., y_m\rangle.$$

We further agree on abbreviating distributions of genotypes of the form

$$\langle \alpha_1\gamma_1, ..., \alpha_s\gamma_s\rangle$$

by

$$\textstyle\sum_{i=1}^{s} \alpha_i\gamma_i \text{ , or by } \alpha_i\gamma_i + ... + \alpha_s\gamma_s.$$

Two distributions of genotypes $\sum_{i=1}^{t} \alpha_i\gamma_i$ and $\sum_{i=1}^{t} \beta_i\gamma_i^*$ are formally multiplied as follows

(7) $(\sum_{i=1}^{s} \alpha_i\gamma_i)(\sum_{i=1}^{t} \beta_i\gamma_i^*) = \alpha_1\beta_1\gamma_1\gamma_1^* + ... + \alpha_1\beta_t\gamma_1\gamma_t^* + ... + \alpha_s\beta_1\gamma_s\gamma 1^* + ... + \alpha_s\beta_t\gamma_s\gamma_t^*.$

This definition may of course be iterated by 'multiplying' (7) from the right with another distribution and so on. The result of such iterated multiplication of n distributions

$$\textstyle\sum_{i=1}^{s} \alpha_i^j\gamma_i^j,\ j = 1, ..., n$$

(all of equal 'length' s) is written as

$$(...((\textstyle\sum_{i=1}^{s} \alpha_i^1\gamma_i^1)(\sum_{i=1}^{s} \alpha_i^2\gamma_i^2))...(\sum_{i=1}^{s} \alpha_i^n\gamma_i^n))$$

or, more briefly:

$$\prod_{j=1}^{n}(\sum_{i=1}^{s}\alpha_i^j\gamma_i^j).$$

Note that in the course of this formal procedure the tuples of 'genetic factors' get longer, for instance the 'length' (i.e. the number of components of the tuples) of $\gamma\gamma^*$ is the sum of the lengths of g and g^*.

With the help of these formal definitions COMBINATOR can be defined as follows. Let $g = \langle a^1, b^1, ..., a^k, b^k\rangle$ and $g^* = \langle c^1, d^1, ..., c^k, d^k\rangle$ be two GENOTYPES. Then

$$\text{COMBINATOR}(\gamma, \gamma^*) = \prod_{j=1}^{k}(1/4a^jc^j + 1/4a^jd^j + 1/4b^jc^j + 1/4b^jd^j).$$

For $k = 1$, this yields

$$\text{COMBINATOR}(\langle a, b\rangle, \langle c, d\rangle) = 1/4ac + 1/4ad + 1/4bc + 1/4bd$$

and for $k = 2$

$$\text{COMBINATOR}(\langle a^1, b^1, a^2, b^2\rangle, \langle c^1, d^1, c^2, d^2\rangle) =$$
$$(1/4a^1c^1+1/4a^1d^1+1/4b^1c^1+1/4b^1d^1)(1/4a^2c^2+1/4a^2d^2+1/4b^2c^2+1/4b^2d^2) =$$
$$1/16a^1c^1a^2c^2 + 1/16a^1c^1a^2d^2 + ... + 1/16b^1d^1b^2d^2.$$

By eliminating the '+' and inserting commas instead, the last expression is just a distribution of genoytpes $\langle\alpha_1\gamma_1, ..., \alpha_{16}\gamma_{16}\rangle$ where each $\alpha_i = 1/16$. Note that the increase in length of GENOTYPES in the course of formal mulitplication is counterbalanced by reducing the length of the tuples occurring on the right hand side in the definition of COMBINATOR after the symbol for the product. These tuples have all length 2, like $\langle a^j, b^j\rangle$. Intuitively, this corresponds to taking only combinations of FACTORS that contribute to the same CHARACTER.

By way of summary we obtain the following description of a model of Mendelian genetics.

A model of *Mendelian genetics* is a model of transmission genetics
with corresponding number k, SETS_OF_FACTORS_i ($i \leq k$),
and component-functions DET_1,...,DET_k of DETERMINER subject
to the following additional requirements:

MEND1: For all GENOTYPES $\gamma = \langle a^1, b^1, ..., a^k, b^k\rangle, \gamma^* = \langle c^1, d^1, ..., c^k, d^k\rangle$
COMBINATOR$(\gamma, \gamma^*) = \prod_{j=1}^{k}(1/4a^jc^j + 1/4a^jd^j + 1/4b^jc^j + 1/4b^jd^j)$

MEND2: For all $i \leq k$ there is a set CHARACTER_i consisting of two
elements EXPRESSION$(i, 1)$ and EXPRESSION$(i, 2)$

MEND3: For any number $i \leq k$, and any two EXPRESSION(i, j_1),
EXPRESSION(i, j_2) there exist exactly two FACTOR$(i, 1)$,
FACTOR$(i, 2)$ such that
a) DET_i(FACTOR$(i, 1)$,FACTOR$(i, 1)$) = EXPRESSION(i, j_1)

b)

$$\begin{array}{ll} \text{DET\_i(FACTOR}(i,1)\text{,FACTOR}(i,2)) & \\ \text{DET\_i(FACTOR}(i,2)\text{,FACTOR}(i,1)) & = \text{EXPRESSION}(i,j_2) \\ \text{DET\_i(FACTOR}(i,2)\text{,FACTOR}(i,2)) & \end{array}$$

The structure of GENOTYPES referring to SETS_OF_FACTORS, together with the definition of COMBINATOR in MEND1 covers Mendel's laws of independent segregation and assortment. EXPRESSION(i, j_1) in MEND3-a is called *recessive*, and EXPRESSION(i, j_2) in MEND3-b *dominant*.

An application of Mendel's law is now considered in which characters may be described in biochemical or gross terms. It is to flower colour in Antirrhinum majus. DeVries and Wheldale[37] suggested that inheritance of corolla colour could be explained in terms of the characters yellow lips, ivory tube, ivory lips, magenta lip and magenty tube. Later Onslow, nee Wheldale, and Bassett[38] gave biochemical characterisations of the situation. It is worth noting that whether a character is expressed in biochemical terms or gross terms this does not affect which of transmission or molecular genetics is employed. However, as will be seen elsewhere, a biochemical characterisation is a pre-requisite for the application of molecular genetics.

In the present example we will use I to signify the character ivory lips, and i to signify its absense. T will signify magenta tube and t the absense thereof. The corresponding factors will be A and a, and B and b. Then MEND3 reads as follows:

$$\begin{array}{l} \text{DET}(A,A) = \text{DET}(A,a) = \text{DET}(a,A) = I \\ \text{DET}(a,a) = i \\ \text{DET}(B,B) = \text{DET}(B,b) = \text{DET}(b,B) = T \\ \text{DET}(b,b) = t. \end{array}$$

Let the factor content of the parents be given by MEND1 as $\gamma = \langle A, a, T, t\rangle$ and $\gamma_* = \langle A, A, T, T\rangle$. Then

$$\text{COMBINATOR}(\gamma, \gamma_*) = 1/8AABB + 1/8AAbB + 1/8AaBB + 1/8AabB + 1/8aABB + 1/8aAbB + 1/8aaBB + 1/8aabB$$

Applying MEND3 we get

$$3/4IT + 1/4iT$$

as expected probabilities of progeny.

A second specialisation of the general transmission model yields a model for linkage genetics. The essential addition in this submodel is some weak reference to chromosomes. The idea that genes are sections of the chromosomes is constitutive for linkage genetics. We do not need, however, to introduce a full concept

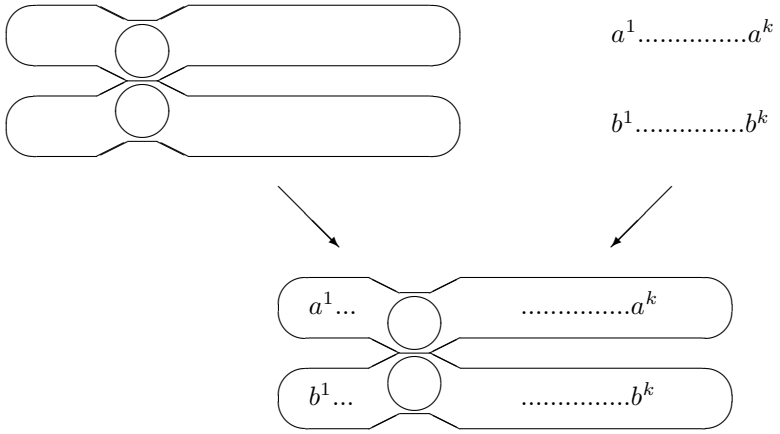
[37](DeVries, 1900) and (Wheldale, 1907).
[38](Onslow and Bassett, 1913) and (Bassett, 1931).

of a chromosome in order to describe the theoretical linkage model. All we need is the idea that genes are linearly ordered along the chromosomes, or along pairs of homologous chromosomes. Restricting ourselves to the diploid case (as before) the following interpretation of genotypes is most natural. Any GENOTYPE of the form $\langle\langle a^1, b^1\rangle, ..., \langle a^k, b^k\rangle\rangle$ has a natural order built in, as given by the indices $1, ..., k$. Furthermore, it consists of two strands, defined as follows.

> If $\gamma = \langle\langle a^1, b^1\rangle, ..., \langle a^k, b^k\rangle\rangle$ is a GENOTYPE the two *strands of* γ are given by the tuples $\langle a^1, ..., a^k\rangle$ and $\langle b^1, ..., b^k\rangle$.

These strands can be easily interpreted as representing the chromosomes of a pair of homologous chromosomes. Thus strand $\langle a^1, ..., a^k\rangle$ represents one chromosome, and strand $\langle b^1, ..., b^k\rangle$ the other as shown in Figure 4-6).

Fig.4-6



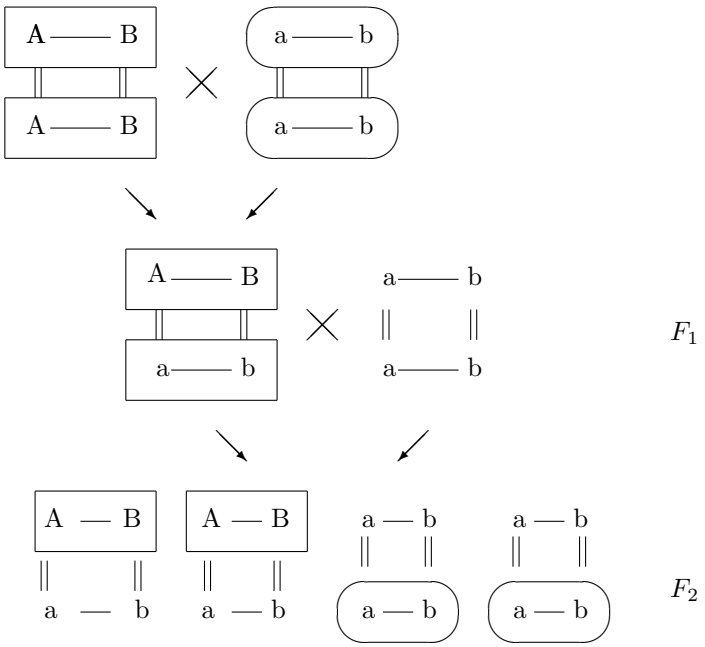
As is well known, the chromosomes duplicate during the first meiotic division, each of the connected copies being called a *chromatid*. The two connected chromatids are copies of each other, and therefore structurally identical. Under the present interpretation each chromatid also is represented by a strand, and connected chromatids have to be represented by the same kind of strand.

It was observed as early as Bateson at al.[39] who worked on linkage between pollen shape and flower colour in sweet peas that the hypothesis of whole strands combining with each other cannot be upheld together with the other parts of the transmission model in the light of the empirical data. Consider the abstract example of two parental GENOTYPES of the form $\langle A, A, B, B\rangle$ and $\langle a, a, b, b\rangle$ where capital and small letters as usual denote dominant and recessive factors. Let us assume that there is sufficient evidence to ascribe these GENOTYPES to real parents. The strands formed in meiosis according to the above definition are these: $\langle A, B\rangle, \langle A, B\rangle$, and $\langle a, b\rangle, \langle a, b\rangle$ (compare Figure 4-7 below). Now if

[39](Bateson et al., 1905).

the strands combine only as wholes we would obtain offspring of GENOTYPE $\langle A, a, B, b\rangle$ in the first generation, and when mating these with individuals of the kind $\langle a, a, b, b\rangle$, offspring of the kind $\langle A, a, B, b\rangle$ or $\langle a, a, b, b\rangle$ in the next generation.

Fig.4-7



Offspring of GENOTYPES $\langle A, a, b, b\rangle$ and $\langle a, a, B, b\rangle$ should not occur. Actually, such offspring are encountered. Think of the mutations yellow body colour, white eye and forked in Drosophila, studied by Morgan and Sturtevant. Thus the assumption of strands combining as wholes (complete linkage) is untenable in general, it occurs only in a limited range of applications. This assumption is replaced by the weaker one that parts of strands also can combine under appropriate conditions. In cytological terms such combinations of parts are represented as occurrences of crossing over. Two chromosomes may exchange only part of their material as depicted in the well-known schema in Figure 4-8.

Fig.4-8

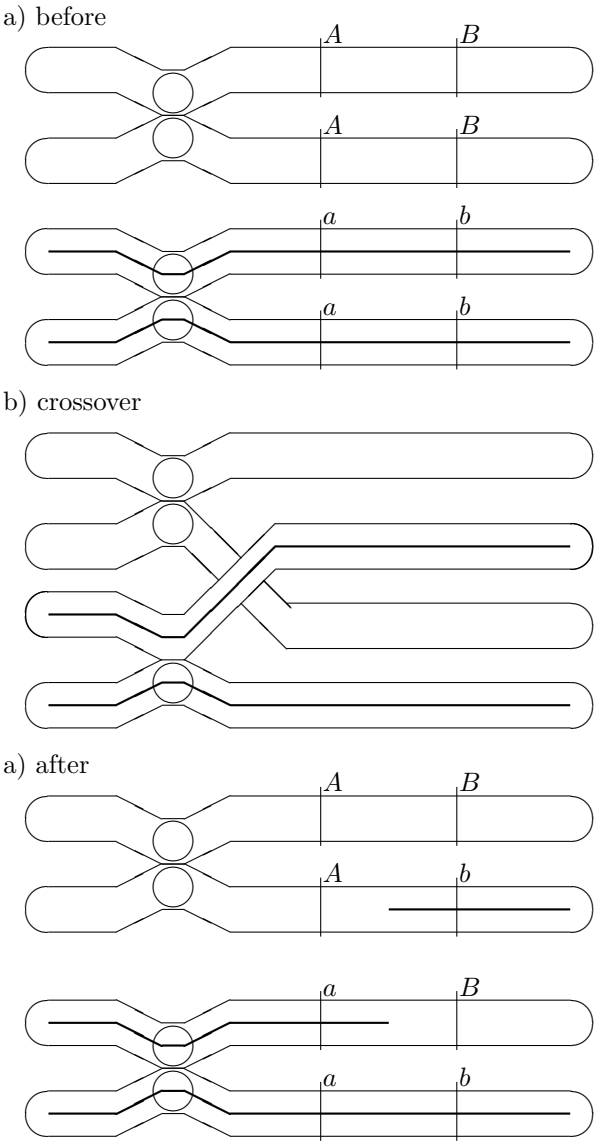


Figure 4-8 is also typical insofar as crossing over has been shown to occur between chromatids and not between whole homologous chromosomes.

The central hypothesis of linkage genetics in first approximation then says that the further two parts of a parental chromosome are away from each other the greater is the frequency of crossing over occurring between them, if the chromosome is involved in reproduction.

By means of crossing over GENOTYPES may be obtained which are not found among the parental ones, as $\langle a, a, B, b\rangle$ in Figure 4-8-c. These GENOTYPES are new with respect to the given parental ones in the sense explained in Chap.3. The definition was this. If $\gamma = \langle a^1, b^1, ..., a^k, b^k\rangle$ and γ' $= \langle c^1, d^1, ..., c^k, d^k\rangle$ are two (parental) $GENOTYPES$ then $\gamma^* = \langle e^1, f^1, ..., e^k, f^k\rangle$ is new with respect to γ and γ' if at least one of the strands $\langle e^1, ..., e^k\rangle$, $\langle f^1, ..., f^k\rangle$ of γ^* is different from all strands occurring in γ and γ': $\langle a^1, ..., a^k\rangle$,$\langle b^1, ..., b^k\rangle$, $\langle c^1, ..., c^k\rangle$, $\langle d^1, ..., d^k\rangle$. In Figure 4-8-c, the inner two strands are different from all the original ones. Informally, a strand is *new* with respect to (at least two) given strands if it is different from all those.

In order to get access at relative frequencies at the phenotypic level we must not talk about strands but about homologous chromosomes, i.e. about pairs of strands or about genotypes for it is these who give rise to phenotypic differences. Also, we must not stick to the configuration as shown in Figure 4-8 but consider two parental genotypes, and one in progeny. Moreover, the notion of a 'part' of a chromosome has to be clarified. Note first, that we are not interested in the concrete material parts, for the positions at which these occur in the chromosome may be occupied by different material objects (at least in the range admitted inside one SET_OF_FACTORS). What counts is not the material but the position on the chromosome. Second, in order to have phenotypic relevance, we need pairs of such positions -as corresponding to 'one position' on a homologous pair of chromosomes. Given the structure of genotypes it is easy to identify such *positions*. Since the factors in $\gamma = \langle\langle a^1, b^1\rangle, ..., \langle a^k, b^k\rangle\rangle$ are linearly ordered by their indices, we may simply take those indices as their positions. Each index $i \leq k$ thus represents one *position* which may be 'occupied' by different pairs $\langle a^i, b^i\rangle, \langle c^i, d^i\rangle$ for which a^i, b^i, c^i, d^i are from the SET_OF_FACTORS_i. Under our interpretation each such position marks a position on two paired homologous chromosomes, as shown in Figure 4-9.

We define the set of LOCI of a model x as the set of these positions.

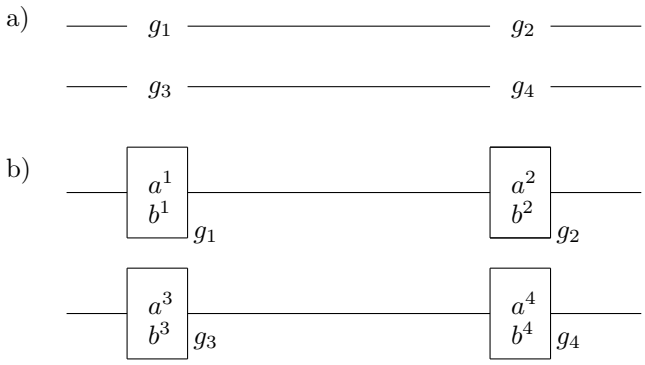
$$\text{LOCI}(x) = \{1, ..., k\}.$$

In order to avoid confusion this definitions should be applied only when the model x is intended to cover just one kind of chromosomes. This assumption does not restrict linkage genetics for linkage as well as genetic maps are of course notions relative to one kind of chromosome. There is no linkage between chromosomes of different kinds, and each genetic map represents one kind of chromosome. In second approximation the central hypothesis of linkage genetics now takes the following form. The farther away two loci are on a genotype the

greater is the frequency of crossing over occurring between these loci, if the genotype gets involved in reproduction.

The models in which recombination frequencies are determined are usually chosen such that populations are coarse, and just reflect the differences in two or three loci. Thus, the mutations yellow body colour, white eye and

Fig.4-9



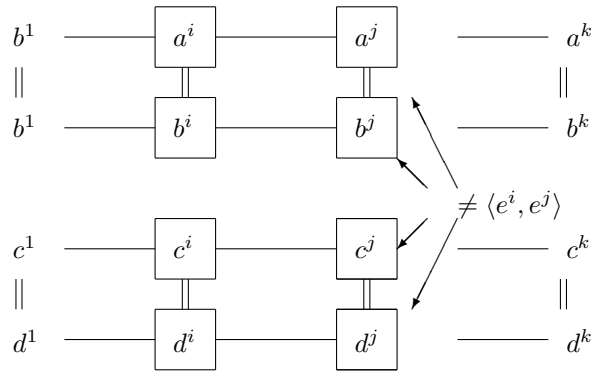
forked are assumed to relate to specific genes. Restricting consideration to particular loci on the genotypes the previous definition of a strand being new may be restated in a more restricted form which takes into account only those factors occupying those loci. If γ, γ' are GENOTYPES and i, j are loci we define

A strand s is *new wrt* γ, γ',i and j iff
1) s has the form $\langle e^1, ..., e^k \rangle$
2) γ and γ' have the forms $\langle \langle a^1, b^1 \rangle, ..., \langle a^k, b^k \rangle \rangle$ and
$\langle \langle c^1, d^1 \rangle, ..., \langle c^k, d^k \rangle \rangle$, respectively
3) $\langle e^i, e^j \rangle$ is different from each of the four pairs:
$\langle a^i, a^j \rangle, \langle b^i, b^j \rangle, \langle c^i, c^j \rangle, \langle d^i, d^j \rangle$

This definition is schematically depicted in Figure 4-10. Accordingly, we say that a GENOTYPE γ^* is *new with respect* to given genotypes γ, γ' and loci i, j if at least one of the two strands of γ^* is new with respect to γ, γ',i and j.

Now the frequency of crossing over between two loci may be defined in two steps. First, consider some given genotype γ^* of the progeny produced by γ and γ'. The frequency of 'occurrence' of γ^* is given in the transmission model as the coefficient which in the genetic distribution COMB(γ, γ') is associated with genotype γ^*. If COMB(γ, γ') has the form $\sum \alpha_i \gamma_i$, and γ^* is γ_i, then the frequency of γ^* is just α_i. If crossing over has occurred a

Fig.4-10

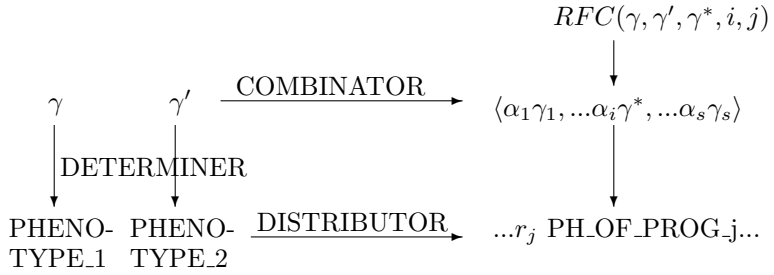


new genotype will be 'observed', i.e. will be inferred from observations of phenotypes by means of extra assumptions about DETERMINER pertaining to the case. Taking this new genotype as γ^* we get it's frequency from the transmission model as just described. This frequency tells how many times some crossing over resulting in γ^* has occurred. In other words, the frequency of crossing over between loci i and j for given parental genotypes γ, γ' which results in genotype γ^* is given by the coefficient of γ^* in the general model. We define

> If γ^* is new with respect to γ, γ',i and j then the
> *recombination frequency in loci* i, j *of* γ and γ' *resulting in* γ^*,
> abbreviated by $RCF(\gamma, \gamma', \gamma^*, i, j)$, is defined by
> $RCF(\gamma, \gamma', \gamma^*, i, j) = \text{COMB}(\gamma, \gamma')(\gamma^*)$.

This definition refers to nearly all components of the transmission model, and is depicted in Figure 4-11 for the case of DETERMINER being one-one.

Fig.4-11



In this definition, the three genotypes γ, γ' and γ^* are assumed to be given. We arrive at the recombination frequency for loci i, j on given parental genotypes

γ, γ' by considering all possibilities of recombination, calculating the corresponding frequencies, and adding them up. In this way we obtain the *recombination frequency* for γ, γ' *in loci* i, j:

$$RCF(\gamma, \gamma', i, j) = \sum RCF(\gamma, \gamma', \gamma^*, i, j)$$

where summation is over all genotypes γ^* which are new with respect to γ, γ', i and j.

Note that these definitions do not contain direct reference to the material objects studied in cytology. It has to be stressed, however, that this truthfully reflects the status of linkage genetics. Though a connection between chromosomes and the ordering of genes and factors along them was hypothesised from the first days of linkage genetics, the linkage theory -as described in the following model- does not systematically refer to, or use, this connection. Linkage genetics entirely relies on recombination frequencies.

Recombination frequencies are used in order to construct *linkage-* or *genetic maps*. Such maps are representations of loci on the line of real numbers such that order and distances as appearing in the genetic material are homomorphically represented by the order and distances of the representing numbers. Note that even the order is not given *apriori* which is not surprising in view of the complicated topological structure of the strands if considered as chromosomes or DNA. For genetic maps, the order is established by comparing the measured distances and fitting them so that additivity makes sense. If, for instance, $d(\alpha, \beta)$ and $d(\beta, \gamma)$ are both smaller than $d(\alpha, \gamma)$ then β must be between α and γ. Clearly, the definition has to be relativised to one given kind of chromosomes or, formally, to one given kind of GENOTYPES. The loci occurring in GENOTYPES of that kind are ordered by the genetic map, while comparison of loci belonging to different kinds of GENOTYPES obviously makes no sense. This relativisation is best achieved by restricting application of the model. We only work with a model whose set of GENOTYPES is interpreted as containing only GENOTYPES 'of the same kind', i.e. corresponding to one kind of chromosomes. Under this interpretation, the relativisation of the genetic map needs not to be made explicit.

We may introduce genetic maps as follows. Let x be a model of transmission genetics. The GENOTYPES in x have the form $\langle a^1, b^1, ..., a^k, b^k \rangle$ where a^i and b^i vary in the SETS_OF_FACTORS_i for $i = 1, ..., k$. A *genetic map for* x is defined as a function

$$f\colon \text{LOCI}(x) \to \mathbb{R}$$

subject to the following requirements:

AL1 for all $i \in$ LOCI(x): $f(i) \geq 0$
AL2 for all GENOTYPES γ, γ' in x and all $i, j \leq k$
$100 \cdot RCF(\gamma, \gamma', i, j) = \mid f(i) - f(j) \mid$

The factor 100 is inserted in AL2 to get percentages rather than relative frequencies. By adding to the models of transmission genetics a genetic map we obtain models of linkage genetics. In other words, a *model of linkage genetics* is defined to be a structure

$$\langle x, f \rangle$$

where x is a model of transmission genetics and f:LOCI(x) $\rightarrow$ $\mathbb{R}$ is a genetic map for x.

Some further remarks are necessary in order to clarify this definition. We require the genetic map to be defined for *all* the loci occurring in the model. Often, not all the f-values of such a map will be known. But this feature f shares with many other concepts, like COMBINATOR and DETERMINER. A notion may be introduced even if we do not know its complete extension in all cases. In this connection it has to be noted (again) that the model is homogenous with respect to application. We may consider a rather restricted set of loci in one model, a set much smaller than the set of 'all real' loci which belong to the population under study simply by using more restricted GENOTYPES. Even if the real system contains many loci, the fact that we don't know them (yet) does not prevent us from application of our model.

Furthermore, we observe that axiom AL2 becomes false if different genotypes (of the same kind) yield different recombination frequencies. This might happen when the loci i, j are occupied by different factors in two parental sets of genotypes. The axiom excludes such situations from the models of linkage genetics, and thus goes beyond a mere 'definition' of the genetic map f. It requires the data to be consistent (at least approximately) in this respect. Only if they are, linkage genetics is able to produce genetic maps.

The actual determination of a genetic map proceeds by a kind of trial and error. First, some assumption about the order is put forward. Then some f-values are determined through recombination experiments. If the f-values fit with the assumption about their order, fine. If they don't fit, we have to modify the assumption about their order and start again. The usual way to proceed in determining f-values is to determine as many recombination frequencies as possible. From the equations in AL2 the f-values may be determined, and, if correct, their order falls out naturally.

The theme of characters, played down in connection with transmission models, may now be reconsidered. The original idea linking the notions of character, gene, and chromosome was that genes are materialised along the chromosome, each gene giving rise to, or causing, one expression. When expressions are classified into characters then, so the original account runs, different expressions of the same character are brought about by different genes, but genes at the same locus of the chromosome. In other words, replacing materially a gene by another one which belongs to the same character will yield a different expression of that character, and by this process of replacement all expressions of a character may be obtained. This picture has turned out as wrong in the development of linkage genetics, however. It was found that different expressions of the same character

may occupy different positions on the genetic map, and therefore also on the chromosome. Think of Drosophila, and its different positions for shape of wings or eye-colour. So in linkage genetics the notion of a character can not be used in its ordinary meaning. If it is used at all, we have to provide for the possibilitiy of different expressions which belong to the same ordinary 'character' falling under different CHARACTERS of the model. This led us to drop the notion of character as a primitive in the general transmission model altogether.

In the light of our previous remarks on application the following objection may be raised against the models of linkage genetics just introduced. As we allow for very small models it might occur that we have two models describing mating experiments within the same species. Our model in this case does not contain any hint at an identity of the linkage maps in both models, which one would certainly want. We are aware of this difficulty. It could be avoided by enlarging the models such that each model already contains all populations of a species. We think that our approach has decisive advantages to this alternative. First, on the alternative account, there is a big problem with application. Application of a model would mean applying it to a whole species. Such a process may be imagined, but it represents a heuristic limit rather than an observable process as actually taking place in scientific practice. Since we want to have the process of application of a model be part of ordinary scientific practice rather than a Platonic ideal we have to stick to small, 'local' models. Second, our approach does not preclude such 'big' models, on the other hand. A model covering a whole species may be subsumed under the previous definition without difficulty. Third, we learn from other sciences that identities of the kind just met for genetic maps within a species are a major feature of empirical theories and therefore should be treated as explicitly as possible (and not hidden in assumptions about the size of a model). Examples to the point are identities of masses in mechanics, of chemical formulas in chemistry, or of utility functions in economics.

The best way to make such identities explicit is to treat them as constraints on models.[40] Different models of a certain kind are constrained to contain identical parts. Different models of linkage genetics representing cases from the same species are required to have identical linkage distances.

It is helpful for the understanding of our models to see what can be made of the notion of a species in terms of such models. We must not expect too much, however. Delineating species is a deep problem which cannot be adequately performed as long as we stay within the boundaries of genetics.[41] Nevertheless, something can be achieved in this direction even when restricting ourselves to the models of transmission genetics. On the basis of these models we define a SPECIES to be a set X of models of transmission genetics subject to the following requirements:

[40] This important idea was introduced by (Sneed, 1971).
[41] Compare (Mayr, 1967) for a survey.

AS1 for any two models x, y in X:
1.1 the two numbers k_x and k_y characteristic for the lengths of PHENOTYPES in x and y, are identical: $k_x = k_y$
1.2 for all $i \leq k$ ($=k_x = k_y$) the i-th expressions in x are the same as in y
1.3 for all $i \leq k$: the i-th set of factors is the same both in x and in y: (SET_OF_FACTORS_i)(x) = (SET_OF_FACTORS_i)(y)
1.4 the determiners are the same in x and y: DETERMINER(x) = DETERMINER(y)
AS2 X is maximal with respect to AS1.

Thus in any two models of a species the structure of phenotypes and genotypes is the same, as well as DETERMINER. AS2 guarantees that no model is 'overlooked'. We need not require that genotypes be identical for this follows from AS1.1 and AS1.3. Also, AS1.1 and AS1.2 imply that the set of all possible phenotypes in any model of a species is the same. This definition provides a first approximation. It does not take into account the well known problems which occur on closer inspection, as for instance changes in chromosome number in Down's syndrome. Here the DETERMINER operates differently owing to changes in GENOTYPE, but the PHENOTYPE is still human.

On the basis of this concept of a species the above constraint on models of linkage genetics requiring identity of genetic maps may be formulated in precise terms.

A set X of models of linkage genetics *satisfies the constraint for the genetic map* iff

CL1 X is a species[42] (and the GENOTYPES in models of X have length $2k$)
CL2 for any two models x, y in X and for all γ, γ' such that γ and γ' are GENOTYPES both in x and in y, and for all $i, j \leq k$ $RCF_x(\gamma, \gamma', i, j) = RCF_y(\gamma, \gamma', i, j)$

The assumption expressed in this constraint by and large may be taken as an empirical hypothesis which is well corroborated. Recombination frequencies for a given set of loci remain identical in different experiments of the same kind, or in experiments within the same species.

We call a set X of models of linkage genetics *recombination complete* if, for every model x in X and for every pair i, j of loci in x there are GENOTYPES γ, γ' in x such that the recombination frequency $RCF(\gamma, \gamma', i, j)$ in x is non-zero. This condition expresses a strong idealization, but is necessary to prove the following

[42] Very strictly, one has to omit the genetic maps from the structures in X in requiring CL1, for SPECIES were defined at the general level of transmission genetics.

THEOREM 1 If X is a set of models of linkage genetics satisfying the constraint for the genetic map, and if X is recombination complete then for all models x, y in X there exists a real number α such that:
the genetic maps f_x and f_y of x and y respectively, are identical up to the addition of α.

Proof: Let X be a set of models of linkage genetics satisfying the constraint for the genetic map, and let X be recombination complete. Let x, y ϵ X. From AS1.3 it follows that the sets of loci in x and y are identical: LOCI(x) = LOCI(y). Let i be some locus in x. If LOCI(x) does not contain any other element, there is nothing to prove. So we may assume that LOCI(x) contains some j different from i. By the assumption of recombination completeness there exist GENOTYPES γ, γ' in x such that $RCF(\gamma, \gamma', i, j)$ is non-zero in x. By AL2 we obtain $100 \cdot RCF_x(\gamma, \gamma', i, j) = \mid f_x(i) - f_x(j) \mid$. By AS1.3 γ, γ' also are GENOTYPES in y, so from CL2 we obtain: $RCF_x(\gamma, \gamma', i, j) = RCF_y(\gamma, \gamma', i, j)$ and from AL2: $100 \cdot RCF_y(\gamma, \gamma', i, j) = \mid f_y(i) - f_y(j) \mid$. So $(*)$ $\mid f_x(i) - f_x(j) \mid = \mid f_y(i) - f_y(j) \mid$. As i, j were arbitrary $(*)$ holds for all pairs of loci in LOCI(x) = LOCI(y). So the distances between any two arguments of f_x and f_y are the same in their common domain. A theorem of real analysis then states that f_x and f_y can differ at most by some constant $\alpha : f_x = f_y + \alpha$.

This theorem may be rephrased by saying that in a set X satisfying all the assumptions stated the genetic maps in all models of X are equal up to the choice of some unit. Thus Theorem 1 may be regarded as an instance of a representation theorem governing the introduction of a numerical function on the basis of 'observational data'.[43] For a SPECIES X in which the two additional conditions stated in Theorem 1 are satisfied we may say that in X the genetic map is uniquely determined up to addition of some α.

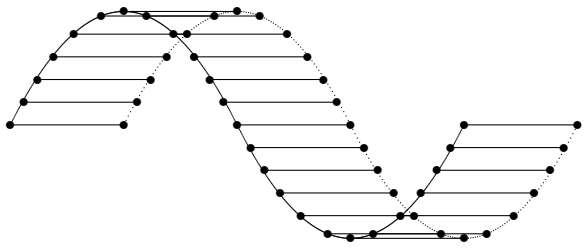
[43] Such representation theorems are studied in the literature about measurement, see (Krantz et al., 1971), for example.

Chapter 5

Molecular Genetics

It is difficult to give an adequate definition of molecular genetics which clearly relates it to the rest of genetics, and to transmission genetics in particular. We do not want to state an attempted definition now but first try to get a more concise idea about molecular genetics. Our general model may then serve to draw distinctions in this area. Let us begin by recalling some basics of molecular genetics.

Much of molecular genetics is concerned with the structure of DNA, in particular with the Watson Crick model of DNA.[44] According to this model, DNA contains two purine bases, adenine (A) and guanine (G) and two pyrimidine bases, thymine (T) and cytosine (C). It is polarised, with a base sequence such as AGTCG having its first member at the 3'-OH group and the last one at the 5'-OH end of the molecule. A base sequence is thus said to be written in the 5' to 3' direction. The DNA molecule has a double helical structure as shown in Figure 5-1.
Fig.5-1



[44](Crick and Watson, 1953).

The two polynucleotide chains have opposite polarity. However, adenine always pairs with thymine while guanine always pairs with cytosine. There is no restriction to the sequence of bases, and it is this sequence which carries genetic information. Typically, a DNA molecule may contain millions of base pairs. Although DNA as encountered in the chromosomes of eukaryotes is generally linear, some prokaryotes have a circular DNA. Similarly, packets of DNA used in transduction experiments, in which the cell genome is deliberately changed, are frequently circular. In the eukaryote, a single DNA molecule is the main constituent of each chromosome. The molecule forms a linear and unbranched string folded in a very complex way. Approximately half of the mass of chromosomes is due to DNA, the rest to small proteins called histones. The chromosome is physically composed of chromatin fibers. Each chromatin fiber is a flexibly jointed chain of nucleosomes each containing around 200 base pairs of DNA wound around the outside of a core of histones. By virtue of this arrangement, the DNA occupies very little volume in comparison to its length. Not all DNA is in the nucleus, however. The mitochondrion contains its own DNA. A distinguishing feature of eukaryotes is that they carry repetitive sequences of DNA. These are termed satellite DNA sequences and are located at the centromeres. Codon sequences for histones are repeated in tandem many times. By contrast, many important proteins only have a single codon sequence. However, when this is the case, these are separated by some hundreds of bases. Furthermore, the codon sequences for almost all proteins in the higher eukaryotes are split into distinct exons separated by an intron many hundreds of base pairs in length. Transcriptional activation of part of a chromosome has been related to 'puffing' or loosening the chromosome in that region.

It is not now necessary to repeat the details of the chemical composition of the different bases here which can be found in many textbooks.[45] This is not to say that the chemical details are not important to molecular genetics. Quite the contrary is true. We omit these details because in further filling out our model we will not reach a level on which the chemical 'fine structure' is used in a substantial way. Such a level, namely the level of genetic control of phenomena like crossing over, exists, to be sure, and forms an area of excited research at the moment. However, since no model of these phenomena has been put forward we can only make a few sketchy remarks about our model's bearing on that issue.

The most frequent form of occurrence of DNA is in linear form. Neglecting circular occurrences -which may be treated in a model of their own- we may apply the general concept of a strand as stated in Chap.3. Each linear DNA molecule has the structure of a strand. It forms an ordered sequence of quanta, where the order is given by the order in which the different base pairs are chemically bound together via the well known phosphodiester bonding. There is an alternative to taking the whole DNA molecule as forming a strand: instead of taking the whole molecule we might consider only one string of the double helix as forming a strand in the sense of Chap.3. There are two reasons which make the latter choice preferable. First, it is more economic. Second, and

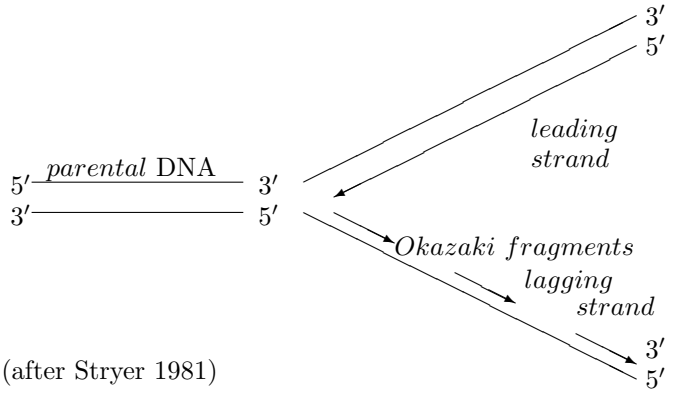
[45] For instance (Strickberger,1985) and (Goodenough and Levine, 1974).

more importantly, the notion of a strand was introduced in order to clarify the structure of genotypes, genotypes were analyzed as configurations of strands in Chap.3. Genotypes carry the genetic information. If we look at how this information is processed in transcription we see that only one of the two DNA strings serves as a template. The genetic information is actually contained in one (either) of the two strings of the DNA molecule. We therefore will use the notion of a strand to apply to just one such string. Comparing the quanta there also is an alternative. We may take quanta to be single bases or to be base triplets. As will become clear below, the latter alternative is more adequate. To summarize, we apply the notion of a strand to DNA molecules in such a way that a strand is given by the sequence of nucleotide bases which makes up one string of the double helix. We do not insist that a strand always covers the full length of a DNA molecule, connected substrings will also provide useful applications of the concept.

A second main part of molecular genetics focusses on the way in which the information encoded in DNA is used to build up other, bigger constituents. Originally, the double helical string of nucleotides in DNA was claimed to have the ability to replicate and also to produce the structural materials for all known life forms. It was, however, early on realized that the DNA molecule in isolation had neither of these properties. In fact, replication could only occur in the presence of many other cell components. Furthermore, the DNA molecule does not directly produce proteins, but acts by intermediate RNA molecules which again, need the assistance of other macromolecules. In appropriate conditions of pH, temperature, and availability of water, proteins of the kind required for the viability of an organism are produced. DNA is not able to replicate itself, but uses DNA polymerase which is itself instructed indirectly by the DNA. Other materials must also be present for replication to occur. Growth of the chain proceeds in the 5' to 3' direction, as the bases are written. Remarkably, DNA polymerase has the property of being able to edit the growing strand. It edits in the reverse direction to replication. DNA ligase is also required for rejoining fragmented DNA.

During replication, DNA maintains its helical form. However, some unwinding is necessary during the synthesis of new DNA. The site of this process is called a replication fork. DNA replication starts at particular places and then proceeds in opposite directions. However, one strand, the leading strand, is replicated continuously, while the other, the lagging strand is replicated discontinuously.

Fig.5-2



(after Stryer 1981)

The process of replication is assisted by DNA gyrase, which changes the sense of supertwisting of the molecule. Error rate in copying DNA is estimated as around one base in ten million. However, damage may occur by physical or chemical means. Even this is reduced by appropriate enzymes. A DNA endonuclease may, for instance, locate and nick a damaged section. A polymerase may then produce the correct sequence. Finally, a ligase may seal the DNA up again.

The DNA molecule is transcribed to produce an RNA molecule which is then translated into a protein. We shall not enter into the nature of RNA, but three forms are present in the cell, distinguished by their location or function. mRNA, or messenger RNA is the template for RNA synthesis. tRNA, transfer RNA, carries amino acids in the sequence dictated by the mRNA. rRNA is a major component of ribosomes, but its precise function is still unclear. The bases of RNA are those of DNA but with uracil (U) replacing thymine. In transcription, the bases of RNA are complementary to those of DNA. The initiation and termination of transcription are closely controlled by the DNA molecule. However, there is no RNA polymerase editing action, so that the transcription process is of much lower fidelity than that of replication. Termination may be signified by a sequence of bases with some common characteristic. The resulting transcript RNA is generally prepared in a piece larger than that subsequently translated, and this post-transcriptional modification is central to the transport of RNA from the nucleus.

The polynucleotide chain of the DNA strand and its transcript RNA strand carries genetic information in the form of a code. Every triplet of three consecutive bases, called a codon, dictates a specific amino acid for the final translation process. There are 64 possible codons, but three of these are in fact signals for chain termination. The code is degenerate, that is to say there is more than one codon for most amino acids. In the eukaryote, most genes are discontinuous. The coding sections are called exons, and the intervening sequences which are not expressed are called introns. Mutations to the genetic material may occur if there is a change of base sequence. There may be substitution of one base

for another. There may be insertion or deletion of bases. The effects here may go far beyond the immediate locality, since the entire reading frame for transcription may be shifted, resulting in almost totally erroneous decoding. If a purine is replaced by a purine, or a pyrimidine by a pyrimidine, this is a transition. A transversion occurs when purine is replaced by a pyrimidine or vice versa. Spontaneous tautomerisation or other changes to individual bases are also possible. In the eukaryote replication occurs at many thousands of places simultaneously. At each of these forks new histones are formed and assemble on the DNA of the lagging strand.

Clearly, the main mechanism here is the way in which amino acids and sequences of amino acids are produced in the cell which is driven and initiated by DNA molecules. This process we labelled transition kinematics in Chap.3. As stated already in Chap.3 the investigation and knowledge of transition kinematics marked a decisive step in the origin of molecular genetics. So transition kinematics forms an essential part of molecular genetics. This suggests to work out a model of it. We could try to do so, but abstain for two reasons. First, the 'mechanism' of transcription and translation, and of how amino acids are built up is very clearly understood, the pictures indicating this model are found in every textbook.[46] Second, and more important from our standpoint, this model is quite different from the basic model presented in Chap.2, and therefore not in the scope of this book. In the following it will suffice for our purposes to introduce a function EX which maps codons on amino acids. We may regard this function as just capturing the input-output scheme but leaving the precise way of constructing the output unanalyzed. The precise way in which amino acids are constructed would be captured by a model of transition kinematics, if we had introduced such a model.

Let us finish our survey by considering a third main area of research in molecular genetics which deals with combination and recombination of DNA, or, more precisely, of strands as introduced before. Although there are two strings of polynucleotide, in replication only one is copied. This is known as semiconservative replication. Single such strings, or strands in our sense, combine during meiosis which after fertilisation leads to the formation of a new genotype. From this picture the immediate impression might be obtained that the GENOTYPE_OF_PROGENY could be stated with certainty if those of the parents are known, and if we know exactly how the 'parental' strands combine during meiosis. This view may be admissible for the haploid case, it is not, however, for a ploidy of two or more. The process of crossing over, of recombination and other rearrangements of the genetic material must then be considered. Though these processes at the moment are still under investigation, it is recognized that they do not occur in a random manner. Some parts of a strand are more prone to break than others, and the process may itself be under genetic control. In any case, however the 'mechanisms' may turn out, we may describe all the ways in which parental strands combine to form GENOTYPES_OF_PROGENY in terms of our model of combination kinematics from Chap.3. The meaning given

[46] For instance (Strickberger,1985) and (Goodenough and Levine, 1974).

previously to the notion of a strand in molecular genetics immediately provides us with molecular genetic configurations of strands. According to the definition of Chap.3 these are sets of strands the quanta of which are consistently mapped into corresponding spatial positions. A strand being taken as one string of a DNA molecule we may pass over from a strand to its chromosome, namely the chromosome consisting of the DNA molecule obtained by completing the strand through its 'missing half' (plus the additional material of the chromosome). Thus a configuration of strands yields a set of chromosomes with base-triplets on the chromosomes (the quanta) endowed with respective spatial positions. By observing such a configuration through its development over time we obtain a model of combination kinematics. Under the present interpretation of the technical concepts such a model describes the spatial rearrangements of chains of nucleotides, and we may easily restrict attention to those phenomena which occur during the 'recombinatorial' phase in meiosis.

We may now try to integrate the three 'parts' of molecular genetics described: DNA, transition kinematics, and recombination, into one comprehensive picture. Not unexpectedly, our model provides a nice frame for such integration. Clearly, the structure of the DNA molecule as uncovered in molecular genetics gives further detail, and material substance to the model component of GENOTYPES. In accordance with the previous discussion we may specialise the notion of a GENOTYPE in molecular genetics to refer to configurations of (molecular genetic) strands. A GENOTYPE thus consists of a set N of chains of nucleotides, each chain being just 'one-half' of a DNA molecule, and each member of the set corresponding to one chromosome, plus a position function ψ assigning spatial positions to the nucleotides. With respect to the quanta the finding that it is triplets of bases, codons, which encode one amino acid each is decisive. There would be no gain by choosing quanta smaller than such triplets.

More precisely, let us understand by a *triplet* a sequence of three of the four bases: A,G,T,C with phosphodiester bonds among any two succeeding bases. Once quanta are chosen as triplets in this particular way the ordering of quanta required in the definition of strands may be defined by reference to chemical notions. We say that two triplets B, B' are *ordered neighbours*: $B \prec B'$ iff they are bound by a phosphodiester bridge such that the 3'-OH group of B meets the 5'-OH group of B'. The complete order of a set Q of quanta now can be defined as follows. Two quanta B, B' in Q stand in the order relation $<$ (i.e. $B < B'$) iff there are quanta $B_1, ..., B_n$ in Q such that $B = B_1 \prec ... \prec B_n = B'$. The ordering so defined we call the *natural chemical ordering*. Altogether, then, the genotypes in molecular genetics may be characterised as configurations of strands the quanta of which are triplets, such that the ordering relation in each strand is given by the previous definition. According to our interpretation of strands the number of strands occurring in one GENOTYPE corresponds to the number of chromosomes of the individuals considered in the model. This number will be the same for all GENOTYPES that occur in one application, so we may require that in one model all GENOTYPES have the same number, k, of strands. By joining these requirements, we obtain a first axiom for molecular models which describes the form of molecular GENOTYPES.

AM1 There is a number k such that all GENOTYPES γ are configurations of strands of the form $\langle N, \mathbb{R}^3, \psi\rangle$ such that
1) N has exactly k elements ($N = \{s_1, ..., s_k\}$)
2) for all $i \leq k$: if strand s_i in N has the form $\langle Q_i, <_i\rangle$ then
2.1) each element of Q_i is a triplet
2.2) $<_i$ is a natural chemical ordering on Q_i

Condition 1 of AM1 entails that the orderings $<_i$ are linear, for linearity was required in Chap.3 to hold for strands in general. Note that linearity does not follow from the definition of a natural chemical ordering. AM1 therefore excludes well known applications in haploids in which genotypes are circular. We choose this restrictive version for matters of simplicity. It would be possible to start with a weaker notion of strands in Chap.3 allowing for the circular case, and to adjust the corresponding notions in this chapter and in Chap.7 below. The adjustment, however, involves some technicalities, and will not be undertaken in this book.

The second area of research considered above was transition kinematics: the transition from DNA to sequences of amino acids. As DNA was stipulated as the essential ingredient of the GENOTYPES in molecular genetics, and as GENOTYPES are the arguments of DETERMINER it follows that transition kinematics covers the model component of DETERMINER. We only have to choose the PHENOTYPES appropriately. Here some complexity arises from the fact that GENOTYPES consist of whole sets of strands (sets of 'chromosomes'). Each strand giving rise to a sequence of amino acids, we encounter a set of sequences of amino acids as being produced from one GENOTYPE. Though this complexity is ultimately unavoidable we may reduce it for analytic reasons and consider just the propagation of one strand into one sequence of amino acids. Such reduction is admissible as long as no phenomena of interference of one chromosome with the production of amino acids by another one is known. Concentrating on just one strand the situation is such that the strand has to be considered as a sequence of base triplets, each triplet possibly determining the production of exactly one amino acid. Thus if the strand consists of k triplets (and so of $3k$ nucleotides) it will give rise to a sequence of k, or less than k, amino acids, and the ordering of the amino acids is determined by that of the triplets on the strand. This suggests we assume that PHENOTYPES consist of sets of ordered sequences of amino acids. Each sequence in a PHENOTYPE corresponds to a strand of a GENOTYPE, and the sets of sequences and strands are thus mapped one-one onto each other.

Now, in fact, an ordered sequence is nothing but a strand, so a PHENOTYPE is just a set of strands. In order to avoid confusion we distinguish $strands_G$ and $strands_P$ in GENOTYPES and PHENOTYPES, when necessary. The elements out of which the $strands_P$ are formed are amino acids; avoiding the term 'quanta' for these and using the symbol $\preceq$ to refer to the order of amino acids on $strands_P$ settles the notation. As with GENOTYPES we may assume that all PHENOTYPES occurring in one molecular genetic model have the same number of strands. Moreover, applications up to now are such

that the number of strands in a PHENOTYPE is the same as that of strands in the corresponding GENOTYPE. Very roughly, each chromosome gives rise to 'its' strand of amino acids.

AM2 There is a number r such that each PHENOTYPE is a set of r $strands_P$ of amino acids, and r is identical with the number k of $strands_G$ in the GENOTYPES.

A $strand_P$ thus is a structure $\langle A, \prec \rangle$ where A denotes the set of amino acids, and $\preceq$ its ordering 'along the strand'.

We might distinguish the different strands occurring in pheno- and genotypes according to different types (different 'kinds of chromosomes'), but such distinctions vary from application to application and thus must not be included in the general model. Moreover, the occurrence of repetitions of genetic material as for instance in Drosophila forbids the assumption that strands of the 'same kind' in different genotypes (or phenotypes) have the same lengths.

GENOTYPES and PHENOTYPES being specified we can turn to DETERMINER. As stated above, not all codons give rise to an amino acid, and among the others there is some redundancy: some codons yield the same amino acids. However, there is complete knowledge about which codon yields which amino acid. So there is a function, call it EX (for expression), mapping codons into amino acids, which moreover we know explicitly. It is given by the following well known list which constitutes another axiom of molecular genetics.

AM3 EX(TTT) = EX(TTC) = Lysine
EX(TTA) = EX(TTG) = EX(CTA) = EX(CTG) = Asparagine
EX(TCT) = EX(TCC) = EX(GCT) = EX(GCC) = EX(GCG) = EX(GAT) = EX(GCA) = Arginine
EX(TCA) = EX(TCG) = EX(AGT) = EX(AGC) = Serine
EX(TGT) = EX(TGC) = EX(TGA) = EX(TGG) = Threonine
EX(CTT) = EX(CTC) = Glutamine
EX(CCT) = EX(CCA) = EX(CCC) = EX(CCG) = Glycine
EX(CAT) = EX(CAC) = EX(CAA) = EX(CAG) = Valine
EX(CGC) = EX(CGA) = EX(CGG) = Alanine
EX(ATA) = EX(ATG) = Tyrosine
EX(ACC) = TRP
EX(ACA) = EX(ACG) = EX(GAC) = EX(GAA) = EX(GAG) = Cysteine
EX(AAT) = EX(AAC) = Leucine
EX(TAT) = EX(TAA) = EX(TAG) = Isoleucine
EX(TAC) = Methionine
EX(AAA) = EX(AAG) = Phenylalanine
EX(GTT) = EX(GTC) = Glutamina N.
EX(GGT) = EX(GGC) = EX(GGA) = EX(GGG) = Proline
EX(CTA) = EX(CTG) = Aspartic Acid
EX(CTT) = EX(CTC) = Glutamic Acid

As already stated the letters A,G,C,T stand for: adenine, guanine, cytosine and thymine, respectively.

In order to characterise DETERMINER in more detail we need a function, called COR (for *correlation*) which to each $strand_G$ of a GENOTYPE assigns its corresponding $strand_P$ in the PHENOTYPE. This function is necessary because on both sides there are sets of strands so that it is not obvious which $strand_G$ corresponds, or causes, which $strand_P$. We may assume that this function does not depend on the environment of a $strand_G$ as given by the other $strands_G$ in the GENOTYPE considered, or by the environment of the whole GENOTYPE. So COR may be introduced as a function defined on the set of all $strands_G$ occurring in the different genotypes of a model. Let us write STRAND($\mathbf{G}$) to denote this set (where $\mathbf{G}$ denotes the model's set of GENOTYPES), and similarly STRAND($\mathbf{P}$) to denote the set of all $strands_P$ occurring in the model which has $\mathbf{P}$ as its set of PHENOTYPES. For a given model of molecular genetics with sets $\mathbf{G}$ and $\mathbf{P}$ of GENOTYPES and PHENOTYPES, respectively, we then can write briefly

$$\text{COR: STRAND}(\mathbf{G}) \rightarrow \text{STRAND}(\mathbf{P}).$$

Having identified the $strand_P$ which 'belongs to' a $strand_G$ we still do not know the precise internal structure of that $strand_P$, that is, precisely which amino acids in which order occur in it. In order to produce that additional information we have to refer to the function EX defined above which assigns amino acids to triplets. What was said about COR also applies here. EX is independent of the environment of a codon as given by the strand or the whole genotype in which the codon occurs. So EX may be treated as a function defined on the set of all triplets, TRIPLET, into the set of amino acids, A-ACID:

$$\text{EX: TRIPLET} \rightarrow \text{A-ACID}.$$

In order to cope with the fact that some triplets just serve as markers, and do not themselves give rise to an amino acid, we agree that A-ACID contains one dummy element which is not an amino acid, onto which all and only the marker triplets are mapped.

Using these two functions DETERMINER may be specified as follows. For a given GENOTYPE containing a set of $strands_G$, each $strand_G$ $s = \langle Q, < \rangle$ of this set is mapped by means of function COR into a $strand_P$: COR(s), and COR(s) $= \langle Q^*, \preceq \rangle$ is further specified by means of function EX as follows. We start with the minimal element q_1 of Q ('minimal' with respect to $<$) and map it to EX(q_1). EX(q_1) has to be the element in Q^* minimal with respect to $\preceq$. Then the 'next' element of Q with respect to $<$, q_2 is mapped into EX(q_2), and EX(q_2) has to be 'greater than' EX(q_1), i.e. EX(q_1) $\preceq$ EX(q_2), and minimal in this respect, i.e. there is no quantum in Q^* which lies properly 'between' EX(q_1) and EX(q_2). This procedure is iterated until we reach the end of strand s, i.e. the maximal quantum of Q with respect to $<$. The set of all $strands_P$ of the form COR(s) obtained in this way forms a PHENOTYPE which is uniquely determined by the GENOTYPE from which we start. So

this PHENOTYPE may be taken as the function value of DETERMINER for the initial GENOTYPE. We may formalize this definiton of DETERMINER as follows.

AM4 If **P** and **G** are the sets of PHENOTYPES and GENOTYPES, respectively, then there exist functions
COR: STRAND(**G**) $\rightarrow$ STRAND(**P**) and
EX: TRIPLET $\rightarrow$ A-ACID
such that for all PHENOTYPES π in **P** and all GENOTYPES γ in **G** the following holds:
If π has the form $\{s_1^*, ..., s_k^*\}$ and γ has the form $\langle N, \mathbb{R}^3, \psi\rangle$ with $N = \{s_1, ..., s_k\}$ then:
DETERMINER$(\gamma) = \pi$ iff
1) COR, restricted to N, is onto $\{s_1^*, ..., s_k^*\}$
2) for all $i, j \leq k$, if COR$(s_i) = s_j^*$ and s_i, s_j^* have the forms
$s_i = \langle Q_i, <_i\rangle, s_j^* = \langle Q_j^*, \prec_j\rangle$, respectively, then
2.1) EX, restricted to Q_i, is onto Q_j^*
2.2) EX, restricted to Q_i, is order preserving, i.e.
for all $q, q' \epsilon Q_i : q <_i q'$ iff EX$(q) \preceq_j$ EX(q')

1) may be rephrased as saying that function COR maps the $strands_G$ occuring in the GENOTYPE one-one onto the $strands_P$ occurring in the PHENOTYPE. Similarly, requirement 2) means that the function EX maps the quanta of each given strand one-one onto the quanta of the corresponding (via COR) strand such that the order among the quanta is preserved, i.e. such that triplets being bound by a phosphodiester bond are mapped on amino acids which are ordered neighbours. Note that the lengths of $strands_G$ and $strands_P$ need not be identical. The number of codons in a $strand_G$ may be greater than the number of amino acids in its COR-image.

The third 'part' of molecular genetics addressed above deals with combination and recombination of DNA strings during meiosis and fertilization. This is the area captured by COMBINATOR in our model. On the basis of the previous specifications of GENOTYPES as configurations of strands we may directly apply the notion of a combination kinematics from Chap.3. Recall that a combination kinematics essentially consists of a sequence of configurations of strands such that the final configuration consists of concatenations of sub-strands taken from the two initial configurations (which are concatenated for the sake of simplicity). Two GENOTYPES which form the arguments of COMBINATOR, and which we may assume to be compatible, are taken as the initial configurations C, C' of the definition in Chap.3. Their strands are considered through a sequence of changes involving combination, recombination, deletion, insertion and others, until a final configuration C^* is obtained. The set N_0 of strands from C and C' has to be chosen appropriately so that all strands in C^* have been obtained as concatenations of sub-strands of strands in N_0. The final configuration by definition is a GENOTYPE of molecular genetics. So combination kinematics in its broadest possible application yields a transition from two parental GENOTYPES to one GENOTYPE_OF_PROGENY.

This raises the question about the role and status of the probability coefficients occurring in the general model in the function values of COMBINATOR. Are they redundant in molecular genetics? Would this mean that molecular genetics is deterministic in some sense? At the present stage of development of molecular genetics, these are important questions. In meiosis the GENOTYPES physically conjugate and exchange varying amounts of material. This is not a random process, however. Certain parts of the GENOTYPE are more prone to change than others, and there may even be genetic control over the process. On the other hand, as a matter of fact, differences in progeny occur within certain, relatively stable probabilities, if large numbers of matings are considered. In principle, there are two ways of accomodating for this fact. First, we might schedule our models so that they just describe one individual occurrence of mating. In this case, no probability coefficients are necessary, all progeny occur 'with certainty', and so do GENOTYPES and PHENOTYPES. On the basis of such a model we might treat the probabilistic facts on large numbers of matings as constraints on the models. In this way the coefficients would naturally be interpreted as relative frequencies (of models in which particular EXPRESSIONS and GENOTYPES did occur as a consequence of mating). The second possibility is to incorporate the coefficients 'directly' in the models in the form described in Chap.2 as probability coefficients of distributions of genotypes as produced by COMBINATOR. In this case their interpretation is still that of relative frequencies, the difference from the first approach being that these frequencies cannot be calculated from the entity under consideration (one model in the second, a set of models in the first case). We think that this difficulty in calculation does not yield a serious objection, and is outweighed by the gain in homogeneity. So we adopt the second way.

There is another argument in favour of this. Molecular genetics by going to the finest possible level of detail ultimately should be able to uncover the 'mechanism' of control which leads to the occurrence of *stable* probabilities in (re)combination. If this were achieved, the coefficients could be explained in a non-probabilistic, non frequentist way. We have to reserve judgement about whether this goal will be achieved, witnessing just the first steps in this direction.

Having agreed that COMBINATOR in molecular genetics should also produce distributions rather than single genotypes we have to admit that COMBINATOR cannot be completely determined by means of combination kinematics. We may state as an axiom

AM5 For all GENOTYPES γ, γ', $\gamma_1, ..., \gamma_n$ and all $\alpha_1, ..., \alpha_n$:
if COMBINATOR$(\gamma, \gamma') = \sum \alpha_i \gamma_i$ then for each $i \leq n$:
γ_i can be obtained from γ and γ' by means of a combination kinematics

This axiom allows for indeterminacy in its 'can be' mode. Given two GENOTYPES γ, γ', in general we simply do not know in advance which γ_i will occur after mating.

Having located the three main areas of molecular genetics in three distinguished parts of our model we may now start from the model and ask about its

remaining components: DISTRIBUTOR, MATOR, and APPEARANCE. As we took the PHENOTYPES to consist of sets of strands of amino acids, there is a far way to go from such PHENOTYPES to the full individuals, as well as to gross expressions of individuals. The detailed study of the transition involved here does not form an essential part of molecular genetics, it is better regarded as belonging to biochemistry, cytology and embryology. However, the transition itself has to be included in our model. It is represented by APPEARANCE which to each individual assigns its molecular genetic PHENOTYPE. MATOR thus keeps its original meaning: it maps parental pairs into offspring produced by them.

The reason why this underlying level is still needed in molecular genetics becomes clear when we turn to DISTRIBUTOR. Here again, the possibility we met in connection with COMBINATOR of formally treating the probability coefficients comes up. Again we choose that version in which the function values of DISTRIBUTOR are distributions rather than single PHENOTYPES. This leaves us inside the boundaries set by our model. The probability coefficients occurring in those distributions have to be interpreted as relative frequencies (of the occurrence of the respective PHENOTYPE in a large number of matings with parents of the same parental PHENOTYPES). Now such frequencies can be determined, and are meaningful, only if we refer to the individual level. If we do not know how to associate amino acids with full individuals we loose any rule for counting them, and the coefficients become meaningless. It must be noted that this account of the coefficients is not entirely uncontroversial.

For the moment, we may summarize the above considerations as stating that molecular genetic models can be obtained, in fact, as specialisations of our basic genetic model. The GENOTYPES are specialised to (sets of) strands of nucleotides as occurring in DNA, the PHENOTYPES are specialised to (sets of) strands of amino acids, DETERMINER is specialised to function as described by transition kinematics, and COMBINATOR as described by combination kinematics. A gap remains with respect to COMBINATOR since the combination kinematics leading from parental $strands_G$ to those of progeny do not tell us precisely what $strands_G$ in progeny we should expect. This gap will perhaps be closed in the future by the development of a 'combination dynamics', but only the first steps towards this end can be preceived to-day.

The models for molecular genetics obtained in this way are still pretty general, they may serve as a basis for further specialisation. This should be expected from the general remarks made in Chap.2 on the structure of comprehensive theories. No doubt, molecular genetics is a comprehensive theory, so we expect it to exemplify the strucutre of a theory-net: with basic models on top and a tree-like structure of specialisations originating from there downwards. Let us indicate very briefly some of the specialisations that can be made. First, we may of course add further details about the chemical structure of DNA, RNA etc.: the atoms, the structure of the molecules, the kinds of chemical bonds, distances and spatial configurations. Such models are particularly useful for computer applications on the one hand, and may be needed in order to investigate the genetic control in recombination on the other hand.

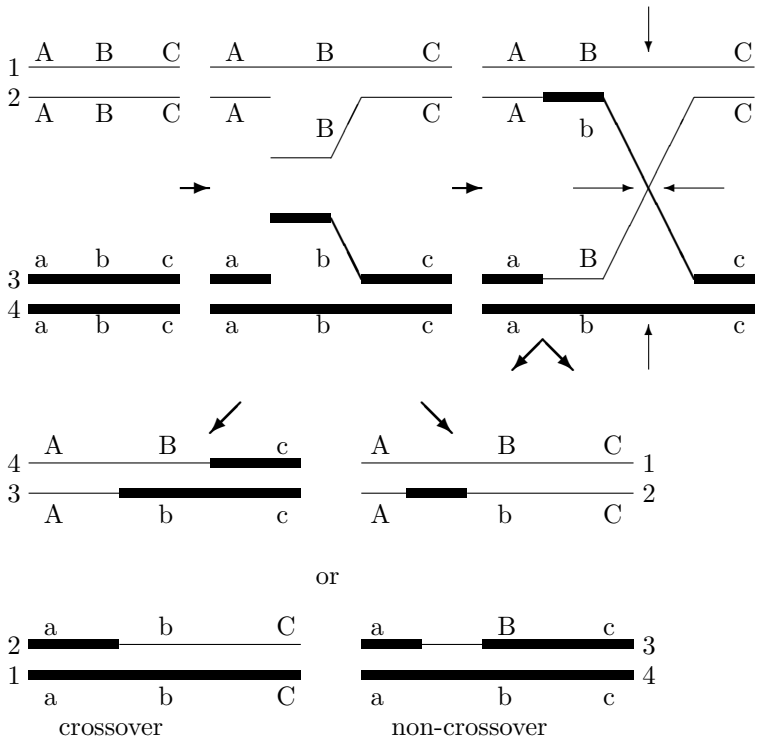
Second, we may specialise the expression function EX introduced in connection with DETERMINER to be the result, in fact, of a model of transition kinematics. That is, we add a transition kinematics and require that EX-values are produced along the lines of such kinematics. A third specialisation is obtained when we insist that the $strands_G$ should have the full length of DNA molecules. In the present formulation there is no requirement to that effect. The strands may be rather short in comparison to those of DNA. This generality is intended and necessary, for in most applications only a small portion of a full DNA molecule is investigated. On the other hand it seems difficult to formulate a requirement that $strands_G$ should cover complete DNA molecules in general, i.e. in a way independent of the kind of individual, and even of the particular chromsome under investigation. For different chromosomes have different lengths, and still greater variety obtains from species to species. Thus it seems that specialisations concerning the length and possibly the way in which nucleotides follow one another have to be formulated in conncetion with a simultaneous restriction to particular species and even chromosomes of those. This is not at all disturbing. On the contrary, in this way we obtain a very large number of different specialisations, a very big theory-net, and thus the picture of a very comprehensive theory.

Further ways of specialisation are found in connection with recombination. Formally, particular models of recombination have to be added to the combination kinematics which is already present governing COMBINATOR. There are different possibilities here. In genetic recombination, a new strand is formed by the breaking and rejoining of existing strands. In transposition, material is moved from one site to another at the same or another chromosome or DNA molecule. Single stranded regions of DNA are intermediates in genetic recombination and enzymes are important in the process. Rejoining of DNA occurs at regions of high homology where the sequence of codons for the two strands is very similar. Rejoining can occur at many points of the DNA, subject to this previous requirement. In E.Coli such general recombination is dependent on the rec family of genes. One property of these is to produce single stranded DNA which can then enter a DNA molecule.

Such general recombination can change the pairing of nucleotides, but does not introduce or delete material. In practice translocatable elements such as the plasmids may be used for the purpose of transduction of bacteria. The location of insertion for such elements is not defined by their homology, but by DNA-protein-interaction. This is dependent on the element having insertion sequences at both ends. Such elements have been invaluable in the development of genetic engineering.

There are three main models of recombination. In the Holiday model (see Figure 5-3), DNA strands for the same polarity are first nicked at homologous sites and then exchanged to produce symmetrical heteroduplex DNA.

Fig.5-3



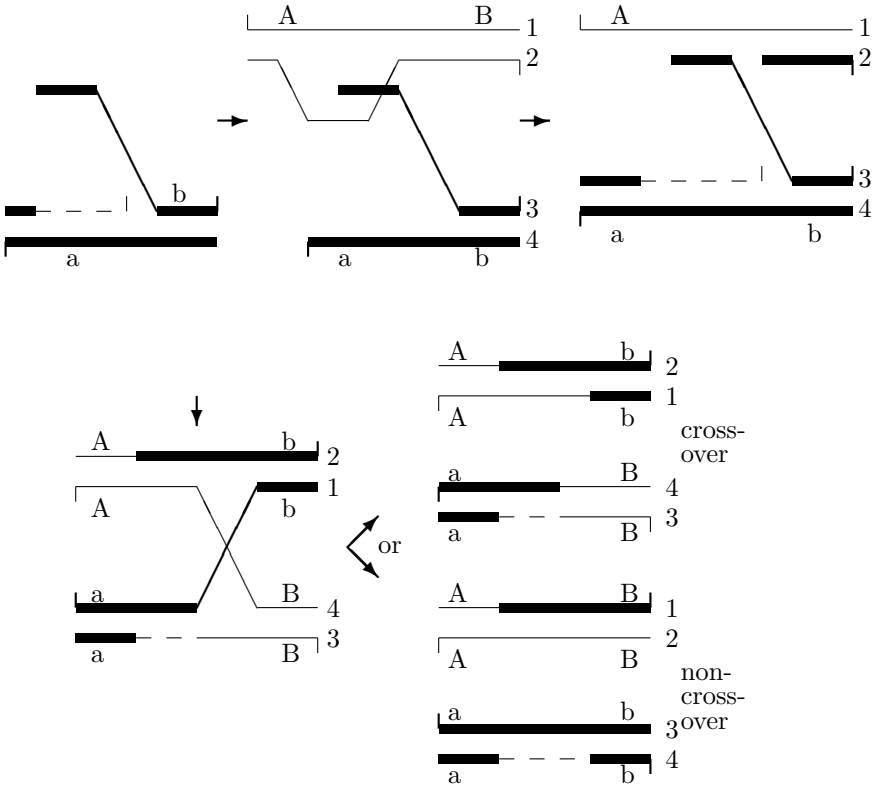
(after Devoret 1988)

The crossed strands known as *Holiday* junction may then be resolved with or without exchange of flanking markers, leading to crossover or non crossover relative to these.

Meselson and Radding varied this model[47] (see Figure 5-4). Here recombination is initiated by a single nick which primes the DNA repair process. As a result, a single strand is displaced which can pair with a homologous region of the other chromatid. However, the joint molecule is degraded and the asymmetrical heteroduplex DNA enlarged by DNA synthesis on the donor chromatid coupled with degradation on the recipient duplex. Branch migration and ligation of the nicks produces a Holiday junction which is isomerised. Symmetrical heteroduplex DNA can be formed by branch migration of the Holiday junction. Again, resolution can lead either to a crossover or non crossover configuration.

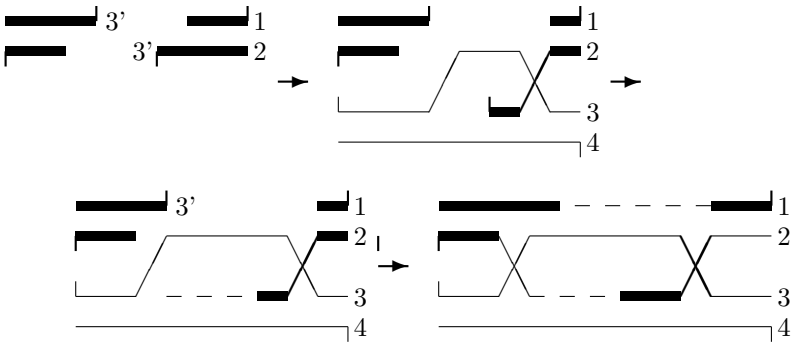
[47](Meselson and Radding, 1975) : Proc.Nat.Acad.Sci.USA 72: 358-61.

Fig.5-4

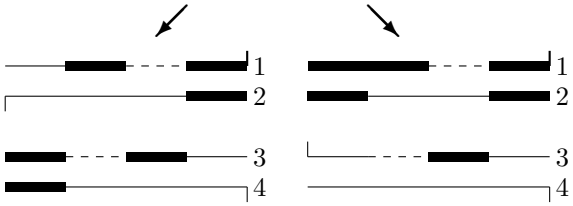


Third, a double strand break repair model was proposed by Szostak et al.[48] (see Figure 5-5).

Fig.5-5 (after Szostak 1983, left hand side: crossover)

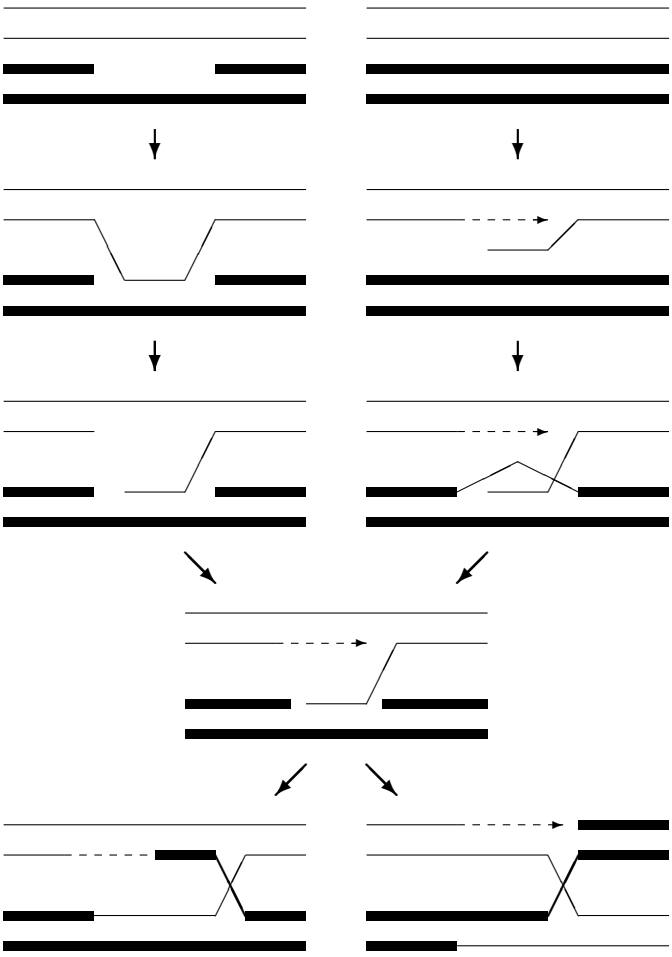


[48](Szostak et al., 1983).



Here, recombination is initiated by a double strand break in one chromatid, which is enlarged by exonucleases so as to form a gap with 3' single stranded ends. One of these then invades a homologous region on the other intact chromatid, forming a small joint molecule. This is enlarged by repair

Fig.5-6



(after Devoret 1988)

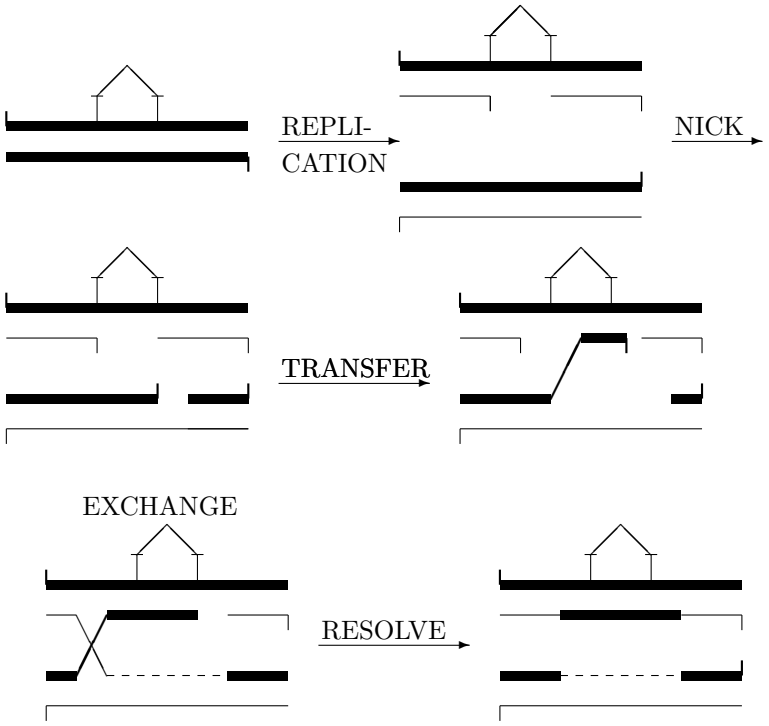
synthesis which is primed by the invading 3' end. In this way the original chromatid is regenerated as a single strand. Finally, repair synthesis fills the gap, using the 3' end as primer. In other words, a double stranded gap is repaired by two rounds of repair synthesis.

There are three stages in all of these models. First, the formation of single stranded DNA regions. Second, the formation of heteroduplex DNA with the generation of Holiday junctions. Finally, the resolution of the cross stranded structure by cutting and ligation of exchanged ends.

Under physiological conditions, one replication fork is the major source of single stranded DNA. The synthesis of single stranded replicative fragments on

the lagging strand leaves single stranded DNA between those fragments already replicated. Figure 5-7 shows a model for daughter strand gap repair.[49]

Fig.5-7



(after Rupp 1971)

Let us now look at the application process corresponding to the molecular genetic models. Application of the models of transmission genetics had a clear direction: one has to work 'bottom up', i.e. first to gather data about the levels of MATOR and DISTRIBUTOR, and then fill in the theoretical *Ueberbau* as given by GENOTYPES and COMBINATOR, together with special hypotheses about the form of the latter, if necessary. The situation in molecular genetics is less clear. On the one hand, GENOTYPES have now become identifiable material objects, and the processes of combination and recombination of strands of the GENOTYPES come ever closer to being 'directly observable'. From this one might expect that the respective parts of the model now can be filled in in a more direct way, by gathering data in the same way as this is done for MATOR and DISTRIBUTOR in transmission genetics. It seems that GENOTYPES and COMBINATOR can be determined in a non-hypothetical way.

For GENOTYPES this is indeed the case. There are various different meth-

[49](Rupp et al., 1971).

ods to get direct information about the structure of DNA and its spatial ordering in the chromosomes: from electron microscopy to chemical or radioactive markers. For COMBINATOR the situation is more involved. It is still difficult to obtain direct data for combination kinematics. More is known about recombination in vitro than in vivo, and comparatively little is known about recombination in higher eukaryotes. In particular, the probabilities of differences in large numbers of offspring are still far from being explained by reference to the physico-chemical structures of the strands during meiosis and fertilization. Up to now there is no access to the coefficients in the values of COMBINATOR which avoids relative frequencies as calculated from observed progeny. In other words, there is no specific molecular access to these coefficients. Usually, they are determined as in transmission genetics: by observing large numbers of processes of matings with parents of the same kind, and determining relative frequencies of an expression in the progeny produced. Such procedures at best yield the coefficients at the level of PHENOTYPES. In order to identify the coefficients from PHENOTYPES with those occurring at the level of GENOTYPES a further hypothesis is necessary. Relative frequencies at the level of strands of DNA are not available yet.

We conclude that although the change from transmission genetics to molecular genetics provided new access to parts of the genetic models that were purely hypothetical before, it did not provide full access to all parts of the model. Some parts still retain a kind of hypothetical character and have to be assumed hypothetically. Roughly, then, the process of application of molecular genetic models fits into the general scheme presented in Chap.2.

In the course of molecular genetic applications a tight interplay with transmission genetics is usually observed. One well established pattern is this. A gross phenotype or some of its expressions are related to a genotype or some genes in transmission genetics. By studying recombination frequencies, the linkage map of the genes is determined. Now for some of the expressions a chromosomal locus may be found by cytological studies. On the hypothesis that the linkage map corresponds to the way in which the genetic material is ordered on the chromosome, this leads to a distinction of various parts of the chromosome as corresponding to the expressions under investigation. Once this -hypothetical- relation is accepted one may start to investigate the distinguished chromosomal loci by molecular means proper. A classical study in which an application of transmission genetics prepared the way for an application of molecular genetics is that of Yanofsky and his colleagues.[50] He appreciated the need for recombination studies in locating specific DNA sequences. In order to reach the small genetic map distances which would correspond to the scale of the DNA molecule, however, it was necessary to use a bacterium or a bacterial virus, since these could replicate rapidly, and provide the vast numbers of crossings needed. Distances were inferred from the frequency with which parent organisms, each with at least one mutation in the same gene, gave rise to offspring in which neither mutation is present. Yanofsky worked with the tryptophan synthetase

[50](Yanofsky, 1964).

enzyme of E.Coli. A set of bacterial mutants with mutations at many sites on the A gene was used in transduction experiments. In this procedure, virus progeny act rather as sperm, and the status of MATOR and COMBINATOR is little affected. However, for such small map distances, and so many mutations, merely ordering the sites is sufficient.

The molecular application began with analysis of the amino acid sequence of the enzymes produced. Ultimately, the identity and location of all 267 amino acids in wild type and mutant strands was established. From this, the possible DNA triplets responsible could be given. In general, such exhaustive analysis is extremely difficult, although the difficulty is technological rather than fundamental. Once a trait has been related to a specific protein, it must further be related to individual polypeptide chains. Only then could relations to specific sequences be hoped for. Nonetheless, such exhaustive analysis of the human genome is currently being contemplated and undertaken.[51]

A more immediate approach is to locate the polypeptide and hence the amino change responsible for a specific character difference, and then relate this to any DNA changes possible. In sickle cell anaemia what is of more direct significance than the entire structure of the haemoglobin chains is that there is a change from glutamic acid to valine at the sixth position. This is probably due to the substitution of an adenine group for a thymine group in the DNA sequence responsible.

The question may be raised whether this type of interplay is not just frequent but even necessary for molecular genetics. In other words, the question is whether molecular genetics can be applied independently of transition genetics at all. This question is important in order to obtain a clear identification of molecular genetics. If every molecular genetic application depends on previous transmission genetic knowledge then molecular genetics would strongly depend on transmission genetics at least in a practical sense. We think that the question can be answered in favour of independence. There seem to be cases -however rare- in which molecular genetics applies without reference to non-molecular chromosomal structure or to a linkage map. These cases are probably restricted to in vitro situations.

[51] (McCusick, 1989).

Chapter 6

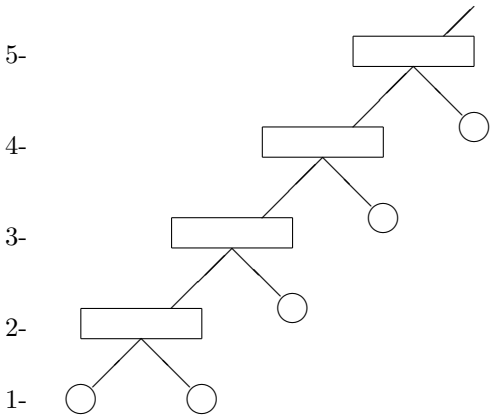
Stochastic Models

The models considered up to now were static, or at least quasi-static. They covered only the transition from parents to progeny, or from one generation to the next. Often, the data available for one such transition are insufficient in order to cut down the genetic content to a sufficient degree of determinateness. In population studies often it is difficult or impossible to prepare pure parental populations, i.e. populations all of whose individuals have one genotype. In the study of human inheritance experimentation is hardly feasible, statistics of whole populations usually incomplete, and transitions to the next generation slow so that one has to be satisfied and to work with sparse data scattered over several generations or over a population in a non-random way. In population studies generalisation to 'impure' cases in which parental populations are characterised by distributions of genotypes rather than by single genotypes, straightforwardly leads to the formalism of genetic algebra which is convenient to trace and calculate the development of distributions of genotypes through many generations. In the human area pedigrees became an important area of research. In this chapter we extend our model so that developments over many generations -not just two- can be covered. It is difficult to describe such extended models for the two levels of applications: individual level and level of populations, in exactly the same terms. We therefore introduce two different models, the first covering only populations proper the second dealing with inheritance at the individual level. Both extended models we call *stochastic* models because they comprise the properties of stochastic processes.

In Chap.2 we used the term 'genetic individual' to cover individuals proper as well as populations. Populations were considered as sets of individuals, and were distinguished from each other by means of differences in phenotypes. A population might be regarded as *pure* if all its individuals have the same phenotype. However, a population may be pure in this sense even if its members have different genotypes. So the notion of purity is a difficult one, and should not be positioned in the centre of interest. In particular, we cannot generally adopt the characterisation of populations as sets of individuals with equal phenotype as a strict definition. Many authors use the term quite freely, presupposing that the individuals making up a population under study can be distinguished from others by whatever criteria, possibly peculiar only to the particular case. We adopt this liberal use here and in the following we will understand by a population any set of individuals, leaving the criteria of identity to *ad hoc* treatment in each case.

The natural approach to extend our model as applying to populations is to 'glue' one such model to the next so that a population of offspring in the preceding model is taken as one of the two parental populations which occur in the succeeding model. In this way we obtain a sequence as depicted in Figure 6-1 in which circles denote parental populations and rectangles denote populations functioning as parental as well as as offspring populations, the bars at the left indicating the generations.

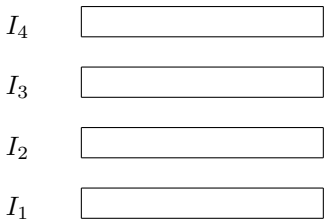
Fig.6-1



Such a model looks inhomogenous, or 'open', because the 'new' parental population needed in each step is added externally. In order to close the model we have to take each of the 'filial' parental populations as being offspring populations, too, i.e. to turn all circles (except the two initial ones) into rectangles. But further adjustment is necessary in order to obtain a satisfactory picture. In particular, we have to take into account that several different populations of offspring may result from two parental populations which brings us back to the problem just mentioned, of how these should be distinguished from each other.

The most elegant extension is obtained by using a general, unspecific notion of population, by lumping together all offspring into one big such population in which different phenotypes may be present, and by generalising the two parental populations to form one big 'parental' population, also containing different phenotypic individuals. In this way a homogenous model is created which consists of just one comprehensive population in each generation, depicted by the rectangles in Figure 6-2.

Fig.6-2



The sets of individuals in each generation are denoted by I_t, and we will refer to these sets as 'generations' in the following.

The distinctions between the different parental populations and the different populations of offspring are now no longer expressed as distinctions between populations, but as distinctions of phenotypes within the 'same' population. Thus by splitting a comprehensive population **pop** into two: $\mathbf{pop} = \mathbf{pop}_1 \cup \mathbf{pop}_2$ where individuals in $\mathbf{pop}_1$ have phenotype π_1 and those in $\mathbf{pop}_2$ have phenotype π_2 we obtain our initial distinction of two parental populations. Similarly, by splitting **pop** into several subpopulations: $\mathbf{pop} = \mathbf{pop}_1 \cup ... \cup \mathbf{pop}_n$ we may represent the previous distinction between different populations in offspring. Formally, it is not necessary to introduce the corresponding subsets of APPEARANCE and PHENOTYPE. If **pop** is a set of individuals, and π a PHENOTYPE then the set of those members of **pop** which have π as their phenotype, $\mathbf{pop}_\pi$, is defined as

$$\mathbf{pop}_\pi = \{i \,\epsilon\, \mathbf{pop} / \text{ APPEARANCE}(i) = \pi\}.$$

By letting π vary through all phenotypes present, the subsets $\mathbf{pop}_\pi$ exhaust the original population **pop**. Instead of referring to the sets $\mathbf{pop}_\pi$ it is usually sufficient to talk about their numbers of elements. The cardinality of $\mathbf{pop}_\pi$ is just the relative frequency of phenotype π's occurring in population **pop**. So the way in which a given population **pop** splits up into different subgroups is uniquely represented by the corresponding genetic distribution of relative frequencies. This distribution, denoted by **p**, assigns a number to each phenotype, and therefore is a function from the set of PHENOTYPES present in a model to the set of real numbers. In the present context we denote the set of PHENOTYPES in a model by PHENO, so that

$$\mathbf{p}\text{: PHENO} \rightarrow \mathbb{R}$$

is such that all function values are non-negative and sum up to one. $\mathbf{p}(\pi) = \alpha$ means that α is the relative frequency of phenotype π in the population considered.

In this way, the earlier distinctions between populations are replaced by genetic distributions on the populations present in each generation. Using t as an index for different generations we may write $\mathbf{p}_t$ for the distribution of phenotypes in the population I_t of the t-th generation. Though it is not at

all easy to distinguish the succeeding generations (more to this below), in the present model this distinction is adopted as primitive. That is, we regard it as being determinable by external means, varying with the model under study.

Accepting this basic structure let us see what happens to the different primitives of our original model. Clearly, the vertical operators, APPEARANCE and DETERMINER should be taken over as they are, and in addition should be kept constant over time. The latter requirement may be expressed as a constraint over different generations. If an individual occurs in different generations it keeps its phenotype, and if a genotype occurs in different generations, its corresponding phenotype as given by DETERMINER also remains identical. With respect to individuals one might try to sharpen the notion of a generation so that one individual cannot belong to different generations. This approach works well at the individual level of pedigrees to be considered below. At the level of populations, however, it creates an unrealistically strict notion of a generation. Often, we are not able to trace all the individuals, and to determine whether all individuals in generation number $t+1$ are offspring from generation number t -whatever the way in which generations are empirically determined.

By constrast, all the horizontal operators of the original model have to be changed. In the generalized picture we may have more than two phenotypically distinct sets of individuals which in one generation function as 'parental populations'. So there is no natural way to establish just one MATOR function between two succeeding generations. We would need several of them, according to the number of parental sets involved. But even if we use several MATORS to provide a connection between succeeding phenotypic sets the problem is how to tell which set(s) of offspring belong to which parental sets of individuals. However, since our interpretation here is in terms of populations conceived as sets of individuals, we may circumvent the problem by retreating to a MATOR function operating at the individual level which we denote by INDMATOR. INDMATOR assigns sets of individuals (offspring) to certain pairs of individuals (parents). Now the link from one generation to the next is established simply by taking the latter as consisting of the union of all sets of progeny produced by INDMATOR from pairs in the former population for which INDMATOR is defined (i.e. which produce offspring at all). We choose INDMATOR not to depend on time because, in a certain sense, the genetically relevant time scale is determined by means of INDMATOR.

For DISTRIBUTOR, a similar situation arises. Since populations in one generation may contain individuals of more than two phenotypes we cannot apply just one DISTRIBUTOR of the format introduced in Chap.2. Rather, we had to apply several of them but again, there would be a problem of keeping them apart. Adopting the notation just introduced, we may see the effect of DISTRIBUTOR as providing a transition from one distribution of phenotypes in one generation to another such distribution in the succeeding generation. So DISTRIBUTOR is replaced by transitions of the form

$$\mathbf{p}_t \Longrightarrow \mathbf{p}_{t+t}$$

where $t+1$ denotes the point or short period of time immediately succeed-

ing t. It is not necessary to introduce a new function symbol to express such transitions. Adopting the well known approach in probability theory we may regard the sequence $\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_t, ...$ of genetic distributions as describing one underlying process of change which expresses itself in the change of probability distributions,[52] that is, as a *stochastic process*. Formally, a stochastic process is a sequence of probability measures, all over one common σ-algebra.[53]

Concerning COMBINATOR we proceed in exactly the same way by referring to genotypes instead of phenotypes. The function values of COMBINATOR were genetic distributions of the form

$$\sum_{i=1}^{n} \alpha_i \gamma_i$$

where $\{\gamma_1, ..., \gamma_n\}$ is the set of all genotypes under consideration (ordered in some conventional way). Generalising at the 'parental side' we obtain a function mapping distributions of genotypes as occurring in one generation to such distributions in the succeeding generation. If we denote the distributions by $\mathbf{p}^*_t$, t indicating the generation, we have transitions of the form

$$\mathbf{p}^*_t \Longrightarrow \mathbf{p}^*_{t+1}$$

just as in the case of phenotypes. The main difference, indicated by the asterisk is that we now deal with distributions of genotypes rather than of phenotypes. Again, the sequence $\mathbf{p}^*_1, \mathbf{p}^*_2, ..., \mathbf{p}^*_t, ...$ of these distributions may be regarded as a stochastic process. The extended model therefore (among other things) contains two stochastic processes $\mathbf{p}$ and $\mathbf{p}^*$, $\mathbf{p}$ describing the change of the distributions of phenotypes 'over time', i.e. in the course of succeeding generations, and $\mathbf{p}^*$ describing the change of distributions of genotypes. This central feature justifies the label 'stochastic model'.

Summarizing these considerations we obtain an extended model of the following form:

$$\langle T, <, J, \text{INDMATOR}, \text{PHENO}, \text{GENO}, \text{APP}, \text{DET}, \mathbf{p}, \mathbf{p}^* \rangle$$

the different components of which have the following meaning. T is a finite set of indices for generations which also may be regarded as a set of points or periods of time, and $<$ is a linear ordering of the elements of T. Since T is required to be finite we can speak of that element of T which immediately succeeds a given one, say t (provided t is not the 'last' one), and denote it by $t+1$. In the 'generations' interpretation $t+1$ is the index of the generation I_{t+1} immediately succeeding the generation with index t.

J is the overall set of (proper) individuals occurring in the model, i.e. the set of all individuals occurring in the different generations, and INDMATOR a function

$$\text{INDMATOR: } J \times J \rightarrow \mathbf{Po}(J)$$

[52] As already noted we do not use the notion of a probability distribution in its full strength but use the weaker notion of a genetic distribution.

[53] See (Bauer, 1974), for details.

assigning to all pairs of individuals the set of offspring produced by those. $\mathbf{Po}(J)$ denotes the power set of J. The function value INDMATOR(i, j) can be empty, in particular the empty set will be produced for pairs which did not coexist in one generation. We do not try to formally disentagle the problem arising from the possibility of one individual mating with another one as well as with further offspring resulting from that mating for the distinction between different populations. The model assumes that this problem is somehow solved in each particular application. A trivial solution would consist of adding an index t to any individual and to INDMATOR, denoting the generation in which the individual or a process of mating is observed. However, this would only shift the problem to the determination of T and of its ordering, $<$.

PHENO and GENO are finite sets, of *phenotypes* and *genotypes*, as before, the elements of which are denoted by π, π_i and γ, γ_i, respectively:

$$\text{PHENO} = \{\pi_1, ..., \pi_r\},\ \text{GENO} = \{\gamma_1, ..., \gamma_s\}.$$

In general we do not assume any particular order of these pheno- and genotypes, but when we switch to the 'sum' notation $\sum \alpha_i \pi_i$ or $\sum \alpha_i \gamma_i$, some conventional order (which really is irrelevant to the formalism) is introduced. Though the occurrence of new pheno- or genotypes, as well as the deletion of some of them, occurs in reality, we choose to exclude these possibilities from the present, basic model for reasons of simplicity. One effect of this choice is that the overall collections of pheno- and genotypes may be regarded as relevant to each generation.

APP and DET are as in the original model. APP assigns a phenotype to each individual

$$\text{APP: } J \rightarrow \text{PHENO}$$

and DET assigns a phenotype to each genotype:

$$\text{DET: GENO} \rightarrow \text{PHENO}.$$

No dependency on time or generations is provided for. In case of APP this amounts to individuals keeping their phenotypes when they occur in more than one generation. DET's independence of generations follows from that of the two sets of pheno- and genotypes just stipulated.

The sets of pheno- and genotypes give rise to corresponding sets of genetic distributions D(PHENO) and D(GENO). Members $\mathbf{p}$ of D(PHENO) formally are functions

$$\mathbf{p}\text{:PHENO} \rightarrow \mathbb{R}$$

such that $\mathbf{p}(\pi) \geq 0$ for all $\pi \in$ PHENO and $\sum_{\pi \epsilon PHENO} \mathbf{p}(\pi) = 1$. Analogously, members of D(GENO) are functions $\mathbf{p}^*$

$$\mathbf{p}^* : \text{GENO} \rightarrow \mathbb{R}$$

such that for all $\gamma \epsilon$ GENO, $\mathbf{p}^*(\gamma) \geq 0$ and $\sum_{\gamma \epsilon GENO} \mathbf{p}^*(\gamma) = 1$. As before, we will often write such distributions in the more convenient form $\sum \alpha_i \pi_i$ and $\sum \beta_i \gamma_i$ where the π_i and γ_i vary in the sets GENO and PHENO, respectively, and α_i, β_i are the corresponding function values of the distributions, i.e. $\alpha_i = \mathbf{p}(\pi_i)$ and $\beta_i = \mathbf{p}^*(\gamma_i)$. This notation assumes that PHENO and GENO are ordered, but the order is not made explicit.

Finally, we require that $\mathbf{p}$ and $\mathbf{p}^*$ are *stochastic genetic processes* over PHENO and GENO wrt. T, respectively. By this we mean that

$$\mathbf{p} : T \times \text{ PHENO} \rightarrow \mathbb{R}, \mathbf{p}^* : T \times \text{ GENO} \rightarrow \mathbb{R}$$

are functions assigning non-negative real numbers to points of time and phenotypes (respectively genotypes), such that for fixed time index t, $\mathbf{p}_t$ and $\mathbf{p}_t^*$ are genetic distributions. Here $\mathbf{p}_t$ and $\mathbf{p}_t^*$ are formally defined by setting $\mathbf{p}_t(\pi) = \mathbf{p}(t, \pi)$ and $\mathbf{p}_t^*(\gamma) = \mathbf{p}^*(t, \gamma)$, so that $\mathbf{p}_t$: PHENO $\rightarrow \mathbb{R}$ and $\mathbf{p}_t^*$: GENO $\rightarrow \mathbb{R}$ have the right type to be Γ-distributions. This completes the description of the extended model for the case of populations.

In retrospect let us make precise the connection between the stochastic Γ-processes $\mathbf{p}$ and $\mathbf{p}^*$ and the corresponding original operators DISTRIBUTOR and COMBINATOR. Consider some fixed generation $I_t \subseteq J$ different from the last one in the extended model. Suppose that I_t contains individuals of just two phenotypes $\pi_1, \pi 2$. Then I_t can be split into two subsets $I_{t,1}$ and $I_{t,2}$ such that members of $I_{t,i}$ have phenotype π_i, respectively. These two populations wrt. to DISTRIBUTOR may be regarded as parental so that DISTRIBUTOR$(\pi_1, \pi_2) = \sum \alpha_i \pi_i$ represents the distribution of phenotypes of their offspring in the following generation I_{t+1}. In the extended model this distribution is given by $\mathbf{p}_{t+1}$. So, in the case considered, $\mathbf{p}_{t+1}$ equals DISTRIBUTOR(π_1, π_2). In the general case I_t will contain members of more than two phenotypes, so I_t has to be characterised by a full Γ-distribution $\mathbf{p}_t = \sum \delta_i \pi_i$. Splitting I_t into two subsets $I_{t,1}$ and $I_{t,2}$ in whatsoever way, the subsets will have their own Γ-distributions, say $\sum \alpha_i \pi_i$ and $\sum \beta_i \pi_i$. Now, if for each pair $\langle \pi_i, \pi_j \rangle$ of phenotypes we know the resulting Γ-distribution DIST$(\pi_i, \pi_j) = \sum_k \delta_k^{ij} \pi_k$ which DISTRIBUTOR assigns to that pair we may calculate the overall distribution for the next generation as

(6.1) $$\sum_k (\sum_{i,j} \alpha_i \delta_k^{ij} \beta_j) \pi_k,$$

on the assumption that matings between any two individuals are equally probable. To see that this formula is correct, note that for each pair $\langle \pi_i, \pi_j \rangle$ the probability of mating is $\alpha_i \beta_j$ so the probability that this pair produces phenotypes π_k in the next generation is

$$\alpha_i \beta_j \delta_k^{ij,},$$

and the overall probability that phenotype π_k is produced is obtained by summing over all pairs $\langle i, j \rangle$:

$$\sum_{i,j} \alpha_i \beta_j \delta_k^{ij}.$$

This is just the coefficient of π_k in the Γ-distribution (6.1). Rearranging summation in (6.1) we obtain:

$$\sum_{i,j} \alpha_i \beta_j \sum_k \delta_k^{ij} \pi_k$$

i.e.

(6.2) $$\sum \alpha_i \beta_j \text{ DISTRIBUTOR}(\pi_i, \pi_j).$$

In this way the transition from $\mathbf{p}_t$ to $\mathbf{p}_{t+1}$ can be expressed in terms of DISTRIBUTOR by (6.2) under the assumption of random mating.

In the same way a connection is established between $\mathbf{p}^*$ and COMBINATOR. If generation I_t splits up into two subpopulations with genetic content γ_1, γ_2, respectively, we may calculate $\text{COMBINATOR}(\gamma_1, \gamma_2) = \sum \alpha_i \gamma_i$ which, in this case is just $\mathbf{p}^*_{t+1}$. In the general case, by splitting the Γ-distribution $\sum \delta_i \gamma_i$ of I_t into two distributions corresponding to some partition of $I_t, \sum \alpha_i \gamma_i, \sum \beta_i \gamma_i$, we may calculate the coefficients of $\mathbf{p}^*_{t+1}$ as before

(6.3) $$\mathbf{p}^*_{t+1} = \sum_k \sum_{i,j} \alpha_i \delta_k^{ij} \beta_j \gamma_k$$

where δ_k^{ij} are the coefficients occurring in $\text{COMBINATOR}(\gamma_i, \gamma_j) = \sum_k \delta_k^{ij} \gamma_k$.

Written differently, (6.3) becomes

(6.4) $$\sum \alpha_i \beta_j \text{ COMBINATOR}(\gamma_i, \gamma_j).$$

In these calculations only one step from generation I_t to the next, I_{t+1}, is considered due to the corresponding 'local' nature of our operators DISTRIBUTOR and COMBINATOR. Of course, the extended model contains several succeeding transitions of that kind. The question therefore arises of whether in a sequence of transitions of the above form we may use the same DISTRIBUTOR and COMBINATOR in each step. Intuitively, assuming that DISTRIBUTOR and COMBINATOR do not change over time means that the rates of segregation, and therefore the law governing the whole process of reproduction, remain stable. A priori such an assumption might definitely be wrong. However, genetic experience shows that there is a wide range of phenomena in which some stability of this kind can be found.

A distinction has to be made here between the levels of pheno- and genotypes. Strictly speaking, the genetic laws in their idealized, theoretical form are formulated for genotypes, so the question of time-independent laws is primarily a matter concerning COMBINATOR. The distributions of phenotypes in most applications are empirically determined and therefore subject to the inaccuracy to be met in all areas of quantitative data. Nevertheless, if the genetic laws remain stable over time, and if there is sufficient fit between genotypic and phenotypic distributions, the latter also have to change in essentially the same way as the former throughout the process, and therefore the 'laws' governing the change of phenotypic distributions will also be stable over time.

If an assumption of stability is made we arrive in the field of genetic algebras. The essential feature of genetic algebras is a multiplicative operation on entities

of the form of genetic distributions. Consider a finite set $G = \{\gamma_1, ..., \gamma_n\}$ and the set $\mathbf{D}(G)$ of all formal expressions of the form

$$\sum \alpha_i \gamma_i$$

over G which is a superset of $D(G)$. On this set a multiplication is defined by reference to given numbers δ_k^{ij}, $i, j, k = 1, ..., n$:

(6.5) $$(\sum \alpha_i \gamma_i)(\sum \beta_i \gamma_i) = \sum_k (\sum_{i,j} \alpha_i \delta_k^{ij} \beta_j) \gamma_k$$

where the δ_k^{ij} are required to satisfy:

(6.6) $$0 \leq \delta_k^{ij} \leq 1 \text{ and } \sum_k \delta_k^{ij} = 1 \text{ for all } i, j \leq n$$

The set $\mathbf{D}(G)$ endowed with its natural structure of a vector space and a multiplication defined by (6.5) is called a *gametic algebra* and the numbers δ_k^{ij} which serve to define the multiplication are called *segregation coefficients*. This label is justified for the right hand side of (6.5) may be rewritten as

$$\sum_{i,j} \alpha_i \beta_j (\sum \delta_k^{ij} \gamma_k)$$

which, in case of proper Γ-distributions is nothing but

$$\sum_{i,j} \alpha_i \beta_j \text{COMBINATOR}(\gamma_i, \gamma_j).$$

So the δ_k^{ij} are the coefficients according to which genotypes γ_i and γ_j combine in producing genotype γ_k.

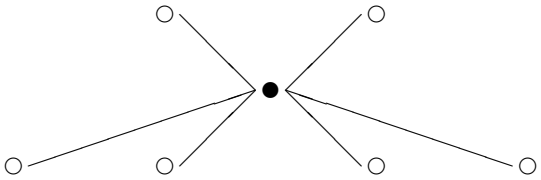
Gametic algebras are obviously non-commutative. They provide one example of genetic algebras. The general notion of a genetic algebra as introduced by Schafer[54] requires some technical algebraic preliminaries, in particular the notion of baric algebras, which are not central to our subject. So we do not attempt to go further and introduce a precise definition of genetic algebras. Precise definitions are found, for instance, in (Woerz-Busekros, 1980).[55]

Turning now to the level of inheritance among individuals without regard of populations we take a new, direct way to pedigrees. We start from a general notion of a pedigree which may be specified in various ways so as to yield models covering sequences of generations. Our model leans heavily on the work of Cannings et al..[56]

Roughly, a pedigree can be imagined as a tree-like graph the nodes of which denote individuals and matings (which also are called marriages in the literature). Two individuals mating and producing offspring is depicted as follows.

[54](Schafer, 1949).
[55](Woerz-Busekros, 1980), p.40.
[56](Cannings et al., 1978).

Fig.6-3



The white nodes denote individuals in genetic order top-down, while the black node denotes a marriage. The same scheme may be conceptualised in a more functional way by saying that the two parental individuals on top are related by a binary relation which we call $marr$, and the individuals at the bottom are the offspring of the latter. This may be expressed by means of a function $offs$ which assigns sets of individuals (i.e. of offspring) to pairs of individuals (parents). In genetic context each individual has at least a phenotype, and we will provide for the possibility of assigning even genotypes to some or all individuals. Thus a *pedigree* is a structure of the following form

$$\langle J, P, G, marr, offs, pheno, geno \rangle$$

which satisfies the follwing requirements:

AP1 J, P, G are non-empty, finite sets, and pairwise disjoint

AP2 $marr$ is a binary relation on J

AP3 $offs : J \times J \rightarrow \mathbf{Po}(J)$ is a partial function

AP4 $pheno : J \rightarrow P$ is a partial function

AP5 $geno : J \rightarrow G$ is a partial function

AP6 for all $i, j \in J : offs$ is defined for $\langle i, j \rangle$ iff $marr(i, j)$

AP7 for all $i, j \in J : offs(i, j) = offs(j, i)$ and
$(marr(i, j)$ iff $marr(j, i))$

AP8 for all $i, j, k, l \in J$: if $\{i, j\} \neq \{k, l\}$ then $offs(i, j)$ and
$offs(k, l)$ are disjoined, and $offs(i, i) = \emptyset$

AP9 any two individuals i, j in J are connected by a chain of marriage or offspring

The sets J, P, G are similar to **I,P,G** introduced at the end of Chap.2. J is a set of individuals proper, P a set of phenotypes, and G a set of genotypes. $marr(i, j)$ symbolises that individuals i, j mate. By AP6 this holds if and only if the offspring function $offs$ is defined for the pair $\langle i, j \rangle$. We agree that by writing

down $offs(i,j)$ in the following we always presuppose that $offs$ is defined for $\langle i,j\rangle$. Thus AP7, for example, presupposes that $\langle i,j\rangle$ and $\langle j,i\rangle$ are both in the domain of $offs$. The function value $offs(i,j)$ may be the empty set. Besides the trivial case this provides room for the theoretically more interesting case of lethal genes. $offs$ being defined only for *marr*ied pairs can only be a partial function. The pheno- and genotypes are assigned to individuals by functions *pheno* and *geno* which also may be partial. This allows for situations where the presence of some individual in a pedigree is known but not its phenotype, not to speak of its genotype. AP7 requires marriage and offspring to be symmetric in the parents. Their order does not matter. AP8 bears some empirical content. It states that different parental pairs produce different offspring. In other words, an individual cannot be produced by two different pairs of parents. Also, it cannot be produced by one parent mating with itself (which may be the case for genetic individuals which are populations). AP9 requires the whole pedigree to be connected. There are no 'isolated' individuals which are neither married to, nor offspring or parents of, other individuals in J. More precisely, AP9 requires that for any two individuals i,j in J there exists a chain $i_1,...,i_n$ in J such that $i=i_1,...,i_n=j$, and any two succeeding i_k,i_{k+1} in the chain are related either by $marr$, or such that i_k is a parent of i_{k+1} or i_{k+1} is a parent of i_k.

In this general form the concept of a pedigree bears little genetic content. It gets content if further special assumptions on *pheno*, *geno*, or the tree-structure are imposed. These are left for specialisations of the general notion. The notion is not trivial however. For instance, it uniquely determines a *depth function* which to each individual assigns the number of matings between it and its most distant ancestor(s) in the pedigree. For a given pedigree $y=\langle J,P,G,marr,offs,pheno,geno\rangle$ we define $depth_y$ to be a function into the natural numbers, $\mathbb{N}$ (zero included):

$$depth_y : J \rightarrow \mathbb{N}$$

satisfying the following two requirements:

AD1 for all $i \epsilon J$, if i has no ancestors j,k in J such that $i \epsilon offs(j,k)$ then $depth_y = 0$

AD2 for all $i \epsilon J$, if i has ancestors j,k in J such that $i \epsilon offs(j,k)$ then $depth_y(i) = max\{depth_y(j), depth_y(k)\}+1$

It is easy to see that $depth_y$ is well defined, and uniquely determined in y. For if $i \epsilon J$ has ancestors j,k in J then by AP8 $\langle i,j\rangle$ is uniquely determined. We may therefore proceed by induction and transfer uniqueness of $\langle j,k\rangle$ to uniqueness of $depth_y(j)$ and $depth_y(k)$, and thus to $depth_y(i)$. As J is required to be finite there is a maximal number k_0 such that all numbers $depth_y(i)$ are smaller than or equal to k_0. We will refer to k_0 by $max(depth_y)$.

Pedigrees are definitely to be interpreted at the individual level. If members of J were taken to be populations rather than individuals AP8 could not be maintained, for at the level of populations difference in parental populations

might come up in a trivial way, for example by one population being a proper subset of another one. A population genetic version of AP8 would have to assume that the respective parental populations are disjoined, a strong assumption which applies only under very good experimental conditions. In the following, interpretation of members of J as populations is ruled out anyway.

The notion of a pedigree gets more substance when reference to our original models is added. We may consider models of genetics as being *contained* in a pedigree. This amounts to the model's individuals, phenotypes and genotypes being included in those of the pedigree, and to the model's functions MATOR, APPEARANCE, DISTRIBUTOR, DETERMINER and COMBINATOR coinciding with those of the pedigree as far as possible. A precise definition is this. Let

$$x = \langle \mathbf{I}, \mathbf{P}, \mathbf{G}, \mathbf{MAT}, \mathbf{APP}, \mathbf{DET}, \mathbf{DIST}, \mathbf{COMB} \rangle$$

be a model of genetics in the format described at the end of Chap.2 and let

$$y = \langle J, P, G, marr, offs, pheno, geno \rangle$$

be a pedigree. We say that x *is contained in* y (or that y *contains* x) iff the following requirements are satisfied.

AC1 $\mathbf{I} \subseteq J, \mathbf{P} \subseteq P, \mathbf{G} \subseteq G$

AC2 $marr$, restricted to $\mathbf{I}$, is identical with the domain of $\mathbf{MAT}$

AC3 $offs$, restricted to $\mathbf{I}$, is identical with $\mathbf{MAT}$

AC4 $pheno$, restricted to $\mathbf{I}$, is identical with $\mathbf{APP}$

AC5 for all $i \in J$, if there is $\gamma \in \mathbf{G}$ such that $\mathbf{DET}(\gamma) = \mathbf{APP}(i)$
then $geno$ is defined for $i, geno(i) \in \mathbf{G}$, and $\mathbf{DET}(geno(i)) = \mathbf{APP}(i)$

AC6 for all $i, j \in \mathbf{I}$, if $pheno(i) = pheno(j)$ then $geno(i) = geno(j)$

By AC3 and 4 the functions $offs$ and $pheno$, when restricted to the individuals in the model of genetics yield MATOR and APPEARANCE of that model. In case of $geno$ a similar requirement cannot be formulated for $geno$ is of a different type from DETERMINER. While DETERMINER operates in the direction from genotypes to phenotypes, $geno$ is defined to operate from individuals to genotypes. Roughly, AC5 says that $geno$, when restricted to the model of genetics, produces genotypes compatible with the connections drawn in that model. $geno(i)$ is a genotype which belongs to phenotype $\mathbf{APP}(i)$ in the sense of $geno(i)$ and $\mathbf{APP}(i)$ being related by DETERMINER. AC6 further narrows down $geno$'s range of variability in the model of genetics. If $\mathbf{DET}$ is known to be one-one, this requirement is implied by AC5.

The model x contained in a pedigree y gives genetic content to the latter. Roughly, the model covers one or more processes of mating plus the production of offspring, and the information captured by the model is transferred to the

pedigree. If all the matings of the pedigree are covered in this way we say that the pedigree has been put on a proper genetic basis. We state this as a formal definition. If $y = \langle J, P, G, marr, offs, pheno, geno\rangle$ is a pedigree then y *has a genetic basis* iff for all $i, j, i_1, ..., i_n \epsilon J$ such that $offs(i, j) = \{i_1, ..., i_n\}$ there exists a model of genetics $x = \langle \mathbf{I}, \mathbf{P}, \mathbf{G}, \mathbf{MAT}, \mathbf{APP}, \mathbf{DET}, \mathbf{DIST}, \mathbf{COMB}\rangle$ such that x is contained in y.

In a pedigree with genetic basis all matings are captured in detail by genetic models. In particular, the transition of genotypes in each case is governed by some COMBINATOR in the 'contained' model of genetics. We may sit back here for a moment and think about the roles of *pheno* and *geno* in pedigrees with genetic basis. At a first glance it seems that *pheno* and *geno* just duplicate what is already present in the form of **APP** and **DET**. In fact, this is so in each single model of genetics contained in the pedigree. The positive role of *pheno* and *geno* comes to bear only if we look at different genetic models contained in the same pedigree. Consider, for example, two such models with parents $\langle i, j\rangle$ and $\langle j, k\rangle$, respectively. It may happen that individual j which thus occurs in both models gets assigned different genotypes in the two models. If $x_1 = \langle \mathbf{I}_1, \mathbf{P}_1, \mathbf{G}_1, \mathbf{MAT}_1, \mathbf{APP}_1, \mathbf{DET}_1, \mathbf{DIST}_1, \mathbf{COMB}_1\rangle$ and $x_2 = \langle \mathbf{I}_2, \mathbf{P}_2, \mathbf{G}_2, \mathbf{MAT}_2, \mathbf{APP}_2, \mathbf{DET}_2, \mathbf{DIST}_2, \mathbf{COMB}_2\rangle$ are the models in question this means that $j \epsilon \mathbf{I}_1 \cap \mathbf{I}_2$, but there are genotypes $\gamma_1 \epsilon \mathbf{G}_1$ and $\gamma_2 \epsilon \mathbf{G}_2$ such that $\gamma_1 = \gamma_2, \mathbf{DET}_1(\gamma_1) = \mathbf{APP}_1(j)$ and $\mathbf{DET}_2(\gamma_2) = \mathbf{APP}_2(j)$. A similar situation may occur even at the level of phenotypes. It might be the case that $\mathbf{APP}_1(j) = \mathbf{APP}_2(j)$ for instance because in x_1 other expressions are studied than in x_2. Such situations are ruled out by the connections to *pheno* and *geno* stated in AC4 and AC5. For, by AC4, $\mathbf{APP}_1(j) = pheno(j) = \mathbf{APP}_2(j)$. This, by AC5, implies $\mathbf{DET}_1(geno(j)) = \mathbf{APP}_1(j) = \mathbf{APP}_2(j) = \mathbf{DET}_2(geno(j))$. So **APP** and **DET** in both models produce the same values for j. Stated differently, *pheno* and *geno* serve to make the different models of genetics contained in a pedigree consistent. If we had chosen to put together several models of genetics in a consistent way in order to obtain a concept similar to that of a pedigree, *pheno*'s and *geno*'s role would be played by constraints on those models.

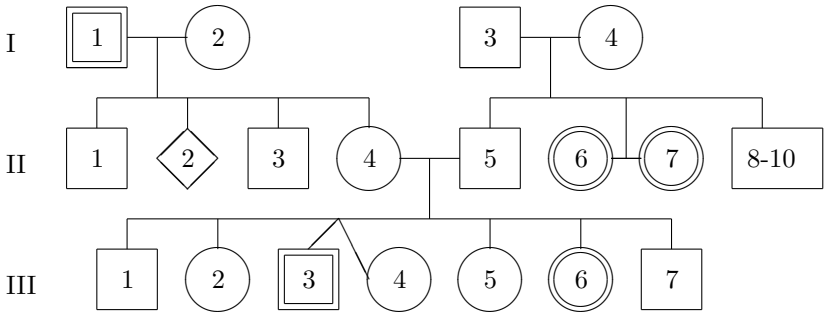
Consider the following example of the technique of pedigrees, which involves segregation for a single character difference.[57] It concerns albinism which, in humans is due to a recessive gene. In the homozygote this causes a very light skin, white hair and pink/red eyes. Using C to stand for the gene for normal pigmentation and c to stand for albinism, we have:

$\mathrm{DET}(C, C) = \mathrm{DET}(C, c) = \mathrm{DET}(c, C) = \text{normal}$
$\mathrm{DET}(c, c) = \text{albino}.$

An example of a pedigree for this trait is shown in Figure 6-4.

[57] See (Strickberger, 1985), p.103.

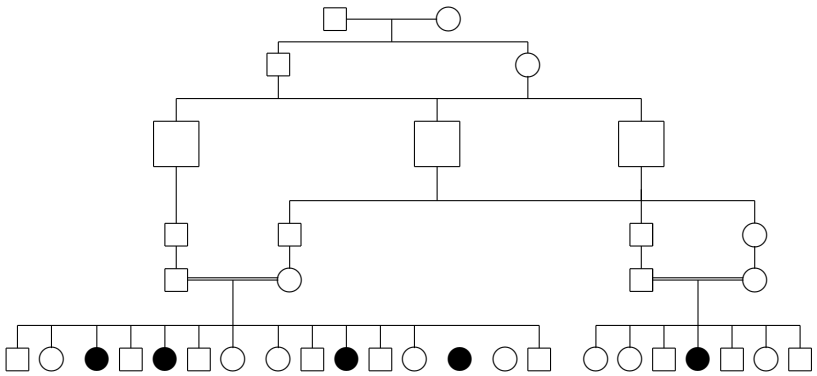
Fig.6-4



In this, shading represents the presence of albinism. Circles are females, squares are males and diamonds are of undetermined sex. Parents are connected by marriage links, consanguinous marriages by double lines. Offspring are connected by sibship lines. Offspring are listed in order of birth, through successive generations, I, II, III and so on. The individuals within a generation are numbered 1,2,3 etc. If sibs are not individually listed, as at II8 to II10, the number of individuals is placed within the symbol. If two lines are conjoined as at II6 and II7, the individuals are identical twins, that is they arise from the splitting of a single fertilised zygote. They may not, however, necessarily have the same GENOTYPE and may have the same GENOTYPE but differing PHENOTYPE, owing to penetrance. Progeny connected separately to the same sibship line are fraternal.

Many traits can be traced in this way, and such diagrams may lead to an understanding of the genetic basis, even when no experiments may be performed, say, for ethical reasons in the case of humans. For example, the human disease microcephaly in which the affected homozygote has a small head and is mentally retarded (see Figure 6-5, after McKusick[58]).

Fig.6-5



[58](McKusik, 1969).

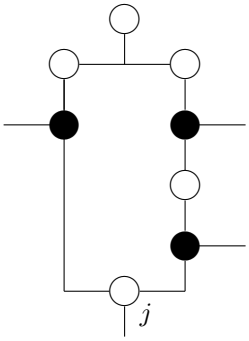
It is found that the only individuals exhibiting the trait are from consanguinous marriages. This would support the presence of a recessive gene for the disease. In practice, arriving at such a diagram requires sampling of a larger population. If only 'interesting' pedigrees are analysed, there may be a bias for or against a particular mode of inheritance. On the other hand, including all individuals is uneconomic, and some reference to previous sampling would seem better. Cannings and Thompson have discussed this central problem in pedigree formation.[59]

In a pedigree we may define generations by reference to the depth function. If $y = \langle J, P, G, marr, offs, pheno, geno \rangle$ is a pedigree then U is a *generation in* y iff there is some number k such that

$$U = \{i \in J / depth_y(i) = k\} \text{ , provided this set is not empty.}$$

In other words, a generation in y consists of all individuals of the same depth. We write I_k for the generation of depth k. By itself, this definition is not very satisfactory. If the pedigree has loops of the form

Fig.6-6



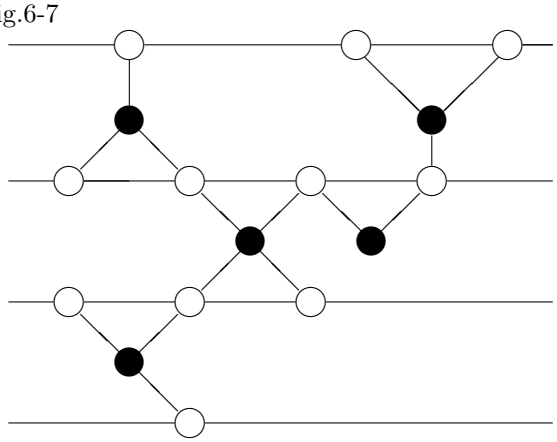
(with black circles for marriage and white circles for individuals, as above) then the two individuals i, i' on top have the same depth, say k, and the individual j at the bottom has depth $k+2$. However, in terms of generations the problem is that j is offspring of i's on the one hand, and so belongs to the generation following i. On the other hand, the definition of depth will follow the right-hand branch which leads to $depth_y(j) = k+2$, i.e. j is in I_{k+2}. In general kinds of pedigrees, containing loops of the kind indicated, for instance, there is no satisfactory definition of generations. If we want this notion to satisfy our intuitions there is no other resolution than to rule out configurations of the kind just considered. If this is done, the pedigree becomes neatly stratified in terms of the depth function, and all offspring of parents at one level (depth) belongs to the following level. Formally, let us define a pedigree to be *regular* iff for all numbers k, I_{k+1} is the set of all offspring obtained from I_k. More precisely, for all $i \in J$:

[59](Cannings & Thompson, 1977).

$i \epsilon I_{k+1}$ iff there exist $j, l \epsilon I_k$ such that $i \epsilon offs(j, l)$.

The graph of a regular pedigree looks like this:

Fig.6-7



Each individual belongs to exactly one generation, indicated by the horizontal line. Between any two generations, there are the instances of matings.

Some implications of these definitions may be noted. Firstly, the generations I_k form a partition of the set J of individuals in the pedigree. By definition, each I_k is non-empty, and each $i \epsilon J$ belongs to some I_k. Moreover, I_k is disjoined from I_r if $k \neq r$, for *depth* is a function, as shown above. Secondly, we note that for all $k \leq max(depth_y), I_k \neq \emptyset$. Thirdly, in a regular pedigree for each $i \epsilon I_k$ with $k \neq 0$ there exist two unique parents $r, s \epsilon I_{k-1}$ such that $i \epsilon offs(r, s)$.

In a regular pedigree y we may consider the sequence $I_0, I_1, ...$ of generations from the point of view of population genetics, that is, as a sequence of matings of different populations. To this end we have to partition each generation into populations according to the phenotypes realised. In each generation I_k we consider the subsets

$$\Omega_{\pi,i} = \{i \epsilon I_k / pheno(i) = \pi\}$$

For each phenotype $\pi \epsilon P$ such a subset is called a *homogenous* population (in generation k).

The theoretically interesting point about pedigrees is that they capture longer sequences of genetic transitions. This allows for more homogenous and subtle technical treatment at the level of phenotypes as well as of genotypes. The transitions from one to the next generation in a pedigree with genetic basis are captured by single COMBINATORS or DISTRIBUTORS which need not have much theoretical connection; there is no theoretical 'thread' running through the sequence of generations. A more adequate conceptualisation requires concepts covering all of the generations at once. Such concepts are available in the form of stochastic processes. As stated above, a stochastic process is given by a sequence of probability measures $\mathbf{p}_k$, $k = 0, 1, 2,...$ (all defined over some common

σ-algebra). In applying this notion to the genetic frame developed we may use a simplified version in which the structure of the σ-algebra does not play any role. Considering a mere set instead we may simply work with Γ-distributions over some set X as defined in Chap.2. By adding such a stochastic process at the level of phenotypes, the notion of a regular pedigree is substantially enriched.

We define a *stochastic pedigree* z to consist of a regular pedigree y together with a stochastic process $\mathbf{p}$, i.e.

$$z = \langle y, \mathbf{p} \rangle, \text{ where}$$

1) y is a regular pedigree, $y = \langle J, P, G, marr, offs, pheno, geno \rangle$

2) $\mathbf{p} : \{0, 1, 2, ..., max(depth_y)\} \times P \rightarrow [0, 1]$ is such that for all $k \epsilon \{0, 1, 2, ..., max(depth_y)\}, \mathbf{p}_k : P \rightarrow [0, 1]$ is a Γ-distribution over P

Here $\mathbf{p}_k$ is of course defined by $\mathbf{p}_k(\pi) = \mathbf{p}(\pi, k)$. The stochastic process $\mathbf{p}$ may be regarded from two points of view. On the one hand, it may be regarded to comprise the information present in *pheno*. In order to make this more clear let us assume that *pheno* were defined for all $i \epsilon J$. Then in each generation I_k we simply could calculate the relative frequencies of all phenotypes occurring, i.e. $RF(\pi, I_k)$. Defining a Γ-distribution θ over P by $\theta(\pi) = RF(\pi / I_k)$ it seems natural that θ should be identical with $\mathbf{p}_k$. So if *pheno* were fully defined we might define $\mathbf{p}$ in terms of the relative frequencies given by *pheno*. However, *pheno* is not usually defined for all individuals in a pedigree. Usually, some information is simply missing. Then *pheno* cannot serve as a basis in order to define the stochastic process $\mathbf{p}$, and $\mathbf{p}$ acquires a more hypothetical status. Seen from this more theoretical point of view, $\mathbf{p}$ is a new, independent theoretical concept which is used in order to systematise incomplete knowledge about *pheno* and the development of distributions of phenotypes over time. There is no need to choose one or the other point of view here. Both are adequate in certain contexts. If we want to apply methods working already on the basis of incomplete knowledge, like processes of peeling used in pedigree analysis,[60] we may well adopt the first point of view and regard $\mathbf{p}_k$ as 'observed distribution'. If, on the other hand, we want to study general laws governing the transition of *pheno*-distributions from one generation to the next, it may be better to look at the distributions as somewhat idealized, theoretical objects which need not completely fit with observed frequencies.

Similar observations hold at the level of genotypes. We may define a pedigree *with genetic basis* z as a structure

$$z = \langle y, \mathbf{p}^* \rangle$$

where $y = \langle J, P, G, marr, offs, pheno, geno \rangle$ is a regular pedigree and

$$\mathbf{p}^* : \{0, 1, 2, ..., max(depth_y)\} \times G \rightarrow [0, 1]$$

[60] (Cannings et al., 1978).

a stochastic process as in the previous definition. Since much less usually is known about *geno* in observational terms than about *pheno*, it seems appropriate to treat $\mathbf{p}^*$ as a theoretical construct from the outset. Though $\mathbf{p}^*$ could be defined in terms of *geno* if *geno* were defined for all $i \epsilon J$, this possibility is very unrealistic. Even in a pedigree with genetic basis where each mating is governed by a model of genetics yielding hypothetical genotypes and a COMBINATOR, the assignment of genotypes to individuals is not uniquely determined, and the function values of *geno* contain some features of arbitrariness. So it seems more adequate to regard $\mathbf{p}^*$ as a means to be used for easier theoretical investigation of the stochastic features at the level of genotypes.

In the two previous definitions no requirements were made connecting the stochastic processes $\mathbf{p}$ and $\mathbf{p}^*$ with their respective counterparts *pheno* and *geno*. This was on purpose because such connections cannot be drawn straightforwardly. If *pheno* and *geno* are undefined for some $i \epsilon I_k$ the relative frequencies $RF(\pi/I_k)$ and $RF(\gamma/I_k)$ will usually deviate from the values of the 'true' Γ-distributions $\mathbf{p}_k$ and $\mathbf{p}_k^*$, if we think of the latter as defined in the hypothetical case where *pheno* and *geno* are defined for all individuals. In order to achieve some conceptually clear description of these connections let us introduce the *observational distributions* $\mathbf{op}_k$ in each generation I_k of a regular pedigree. $\mathbf{op}_k$ is defined by the observed relative frequencies at the level of phenotypes, i.e.

$$\mathbf{op}_k(\pi) = RF(\pi/I_k) \text{ for all } \pi \epsilon P.$$

In the same way, at the genotypic level we may define

$$\mathbf{op}_k^*(\gamma) = RF(\gamma/I_k) \text{ for all } k \leq max(depth_y) \text{ and } \gamma \epsilon G.$$

The connection between these observational distributions and the stochastic processes introduced before has to be expressed by some approximation. The most natural approximation is given by reference to some $\epsilon > 0$, setting

$$|\ \mathbf{op}_k(\pi) - \mathbf{p}_k(\pi)| < \epsilon, \text{ for all } \pi \epsilon P$$

and similarly for $\mathbf{op}_k^*$. If such an approximation holds for suitably small ϵ the stochastic processes $\mathbf{p}_k$ and $\mathbf{p}_k^*$ may be called *useful*.

Summarising this discussion we may introduce the notion of an *applied pedigree z with ϵ fit* to consist of a regular pedigree y together with stochastic processes $\mathbf{p}$, $\mathbf{p}^*$

$$z = \langle y, \mathbf{p}, \mathbf{p}^* \rangle$$

such that

1) $\langle y, \mathbf{p} \rangle$ is a stochastic pedigree

2) $\langle y, \mathbf{p}^* \rangle$ is a pedigree with genetic basis

3) for all $k \leq max(depth_y)$ and all $\pi \epsilon P$, $\gamma \epsilon G$:

$$| \mathbf{op}_k(\pi) - \mathbf{p}_k(\pi)| < \epsilon \text{ and } | \mathbf{op}^*_k(\gamma) - \mathbf{p}^*_k(\gamma)| < \epsilon.$$

The choice of ϵ has to depend on the amount of information lacking about *pheno* and *geno* relative to the total set of individuals considered, but also on the observed relative frequencies. If the relative frequencies in a generation are all of the same order, ϵ may be chosen with respect to that number. If relative frequencies greatly vary, one has to concentrate on relative frequencies for phenotypes or genotypes occupied by few individuals for these will change considerably if lacking values of *pheno* or *geno* turn out to contribute to the latter. In such cases an ϵ leading to fit will typically be greater than in the first case.

Comparison of the frequencies of pheno- and genotypes in the transition from parents to progeny as given by the stochastic processes $\mathbf{p}$, $\mathbf{p}^*$ on the one hand, and by the COMBINATORS of the genetic models describing these transitions in a pedigree with genetic basis on the other hand is more involved. Let us look at the level of genotypes first.

In a regular pedigree with genetic hypothesis and genetic basis the transition of genotypes from generation I_k to I_{k+1} may be described as follows. Let γ, γ' be the genotypes of two parents in generation I_k and γ^* the genotype of one of their offspring in the following generation I_{k+1}. Since the model has a genetic basis, the mating process is described by a model of genetics with COMBINATOR, i.e.

$$\text{COMBINATOR}(\gamma, \gamma') = \sum \alpha_i \gamma_i$$

and γ^* is, say, γ_r. Now in I_k the probabilities of γ and γ' occurring are given by $\mathbf{p}^*_k(\gamma)$ and $\mathbf{p}^*_k(\gamma')$. On the assumption that mating in I_k occurs with equal probability among any pair of individuals the probability of matings with parents of genotypes γ and γ' in I_k is $\mathbf{p}^*_k(\gamma) \cdot \mathbf{p}^*_k(\gamma')$. On the other hand, the probability of creating offspring of genotype γ_r in each such mating is α_r, so the total probability of γ_r occurring in $I_{k+1}, \mathbf{p}^*_{k+1}(\gamma_r)$, is $\alpha_r \cdot \mathbf{p}^*_k(\gamma) \cdot \mathbf{p}^*_k(\gamma')$. By regarding COMBINATOR(γ, γ') as a Γ-distribution over G, α_r may be written as COMBINATOR$(\gamma, \gamma')(\gamma_r)$. So we arrive at the following equation (with γ^* instead of γ_r):

(6.7) for all $k < max(depth_y)$ and all $\gamma, \gamma', \gamma^* \in G$:
$\mathbf{p}^*_{k+1}(\gamma^*) = \mathbf{p}^*_k(\gamma) \cdot \mathbf{p}^*_k(\gamma') \cdot$COMBINATOR$(\gamma, \gamma')(\gamma^*)$.

This is a general statement governing the stochastic process $\mathbf{p}^*$ in all cases where the probability of mating is equally distributed within each generation. For different kinds of distributions of mating probabilities, different versions of (6.7) may be derived in an analogous way.

We note that the connection expressed in (6.7) is entirely concerned with idealised, theoretical functions. If we switch from $\mathbf{p}^*$ to relative frequencies in terms of *geno*, an approximative version of (6.7) may be obtained connecting relative frequencies in the pedigree as expressed by $\mathbf{op}^*$ with the theoretical values of COMBINATOR in the underlying model of genetics.

At the level of phenotypes we may consider an analogous connection between the value of DISTRIBUTOR, DISTRIBUTOR(π, π'), for two parental phenotypes in generation I_k, and the value of $\mathbf{p}_{k+1}$ as given in the pedigree. Repeating the derivation which led to (6.7) we obtain a similar formula for phenotypes in the case of equal probabilities of mating in each generation.

(6.8) for all $k < max(depth_y)$ and all $\pi, \pi', \pi^* \epsilon P$:
$\mathbf{p}_{k+1}(\pi^*) = \mathbf{p}_k(\pi) \cdot \mathbf{p}_k(\pi') \cdot \text{DISTRIBUTOR}(\pi, \pi')(\pi^*)$.

Replacing the $\mathbf{p}_k$'s by observed numbers, $\mathbf{op}_k$, (6.8) may be transformed into a statement connecting observed relative frequencies as occurring in the pedigree and in the underlying models of genetics.

Finally, let us consider the additional features arising in the relation of models of genetics being contained in a pedigree when the genetic individuals in the 'contained' model are populations. Since the interpretation of a pedigree is in terms of proper individuals we have to introduce some connection between populations in the underlying model and individuals in the pedigree. The natural connection is of course to take populations $\Omega_{\pi,k}$ as defined above to be identical with homogenous populations occurring in the 'contained' genetic model. As before, the genetic model may cover less generations or matings than present in the pedigree. For reasons of simplicity let us consider just one transition between two succeeding generations in the pedigree. A natural modification of the requirements stated above in the definition of 'contained in' then is the following.

Let $x = \langle \mathbf{I}, \mathbf{P}, \mathbf{G}, \mathbf{MAT}, \mathbf{APP}, \mathbf{DET}, \mathbf{DIST}, \mathbf{COMB} \rangle$ be a model of population genetics and $y = \langle J, P, G, marr, offs, pheno, geno \rangle$ be a regular pedigree.

x is contained in y iff

1) there is some $k < max(depth_y)$ such that each genetic individual $i \epsilon \mathbf{I}$ is a homogenous population $\Omega_{\pi,k}$

2) $\mathbf{P} \subseteq P$ and $\mathbf{G} \subseteq G$

3) for all $i, j, i_1, ..., i_n \epsilon \mathbf{I}$: if $\mathbf{MAT}(i, j) = \{i_1, ..., i_n\}$ then $offs(i, j) = \cup_{j \leq n} i_j$

4) for all $i \epsilon \mathbf{I}$ and all $j \epsilon i$: if *pheno* is defined for j then $\mathbf{APP}(i) = pheno(j)$

5) for all $i \epsilon \mathbf{I}$, if there is some $\gamma \epsilon \mathbf{G}$ such that $\mathbf{DET}(\gamma) = \mathbf{APP}(i)$ then for all $j \epsilon i$:
geno is defined for j and $geno(j) \epsilon G$ and $\mathbf{DET}(geno(j)) = \mathbf{APP}(i)$

6) for all $i \epsilon \mathbf{I}$ and all $j, k \epsilon i$: if *geno* is defined for j and k then $geno(j) = gen(k)$.

Requirements 2-6 are essentially taken over from AC1 and AC3 to AC6, with adjustments to the population case. The definition of a pedigree with genetic basis can be taken over, we only have to change the label into 'pedigree with

population genetic basis'.

The relation of population models to stochastic models is much the same as in the individual case. Consider a stochastic pedigree $y = \langle J, P, G, marr, offs, pheno, geno\rangle$, a model of population genetics $x = \langle \mathbf{I}, \mathbf{P}, \mathbf{G}, \mathbf{MAT}, \mathbf{APP}, \mathbf{DET}, \mathbf{DIST}, \mathbf{COMB}\rangle$ contained in y, as well as parental phenotypes π, π' in generation I_k, and a phenotype π^* in generation I_{k+1}. Let $\Omega_{\pi,k}, \Omega_{\pi',k}$ and $\Omega_{\pi^*,k}$ be the corresponding homogenous populations and suppose the parental populations occur in x. Then DISTRIBUTOR in x assigns a Γ-distribution to the two parental phenotypes π, π' :DISTRIBUTOR$(\pi, \pi') = \sum \alpha_i \pi_i$. By definition of the models of population genetics each α_i here is the relative frequency of offspring with phenotype π_i (in population $\Omega_{\pi_i,k+1}$) in the total offspring $\Omega_{\pi,k}$ and $\Omega_{\pi',k}$. As before, we may consider the probability for offspring of phenotype π_i from parents of phenotypes π and $\pi' : \mathbf{p}_k(\pi) \cdot \mathbf{p}_k(\pi') \cdot \alpha_i$. In terms of populations these numbers also make sense. $\mathbf{p}_k(\pi)$ is the relative size of population $\Omega_{\pi,k}$ in I_k, and so is $\mathbf{p}_k(\pi')$ with respect to $\Omega_{\pi',k}$. The relative frequency of individuals with phenotype π_i in the offspring, $\mathbf{p}_{k+1}(\pi_i)$, then, is given as before by $\mathbf{p}_k(\pi) \cdot \mathbf{p}_k(\pi') \cdot \alpha_i$, that is

$$\mathbf{p}_{k+1}(\pi_i) = \mathbf{p}_k(\pi) \cdot \mathbf{p}_k(\pi') \cdot \text{ DISTRIBUTOR}(\pi, \pi')(\pi_i).$$

As before, this equation holds only on the assumption that probabilities are equally distributed. A similar derivation at the level of genotypes yields the same formula with γ's instead of π's, and COMBINATOR instead of DISTRIBUTOR.

Chapter 7

Diversity

Our approach stresses the unity of different genetic models. In fact, it not only stresses it, but proves it, in a sense. For all the models of the different branches: transmission-, *Mendelian*-, linkage-, and molecular genetics are obtained by refinement or specialisation of the basic model introduced in Chap.2. In a strict sense, therefore, all the different branches' models are structurally identical at the basic coarse level where phenotypes and genotypes are not yet looked at in detail. This was one of the necessary conditions put forward for unity in a scientific field in Chap.1

This observation seems rather surprising, given the many discussions about whether molecular genetics can fully replace transmission genetics, and whether the latter can be reduced to the former.[61] In fact, if the relation were so obvious one would wonder how such a discussion is possible. It has to be admitted that even at the coarse level at which phenotypes and genotypes are not further specified, a claim of structural identity of different models has not been strongly defended, neither among geneticists nor among methodologists. Why? There is a simple explanation: because such a claim presupposes sufficient conceptual 'carving' to produce models which can be structurally compared, and because there were no attempts to construct precise overall models for genetic theories up to now.[62] On the informal level of textbook presentations a claim of structural identity hardly arises because at this level no systematic distinction is made between central assumptions and concepts characteristic for all kinds of applications, and other specific assumptions and concepts used only in particular applications. There are no easily identifiable classes of assumptions one might compare. In order to arrive at a stage where comparison is easy the task of clustering together the various concepts and assumptions in the right way has to be performed first. Anybody who tries to figure out which assumptions are central for one branch of genetics, and which are not, will agree that this is by no means a trivial procedure.

When a discipline reaches a stage of sufficient maturity, as is now the case for genetics, there is some inherent drive to deal with foundational issues. If things get so specialised that scientists in different subdisciplines have difficul-

[61] This discussion is mainly entertained among philosophers of biology, to be sure. We mention the exchange of papers between Hull and Schaffner concerning this point: (Hull, 1969,1972), (Schaffner, 1967,1969a,1969b).

[62] We do not think that (Woodger, 1959) and (Kyburg, 1968) prove the contrary. In addition to using the somewhat clumsy syntax of first-order predicate calculus their accounts deal with one special model only, namely the Mendelian model.

ties understanding each other's articles there is some need for unification, that is, clarification, simplification, and comparison. Concepts and assumptions that were introduced in the first phase of struggle with the yet unknown phenomena get conceptually clarified by subsequent careful study of their relations to existing models. This automatically leads to various forms of comparison, and sometimes simplification follows. We think that such issues will become increasingly important in genetics though it is admitted that at the moment the development is still so fast that most activity is attracted by immediate, practically relevant research.

The kind of structural *identity* considered holds only at the coarse level of the basic model, as mentioned. It is lost once we go into the detailed specification of phenotypes and genotypes. There is enough room for diversity under the frame given by the general models. In considering this diversity in more detail, questions about comparison naturally come up. Let us begin by restating in detail the points in which any two of our models differ, and by attempting to achieve clarity about how any two models can be compared, and are interrelated.

For reasons of comparison some general terminology concerning the components of the models is useful. We distinguish between the *objects* occurring in a model on the one hand, and the *functions* and *relations* occurring in it on the other hand. By objects we mean genetic individuals (the entities denoted by PARENT and PROGENY), the PHENOTYPES, the GENOTYPES (of parents and progeny), the EXPRESSIONS and the FACTORS. All other items occurring in the models are functions: MATOR, APPEARANCE, DISTRIBUTOR, DETERMINER and COMBINATOR, the genetic map f in linkage models, the component functions DET_1,..., DET_k. A third kind of entities not covered by this distinction are numbers. These we regard as objects, but of an auxiliary kind.

We begin with the easy cases, considering first the relation between the general models of Chap.2 and the models of transmission genetics. Each model of transmission genetics by definition is also a general model. So the kinds of objects and the functions of the general models all reappear in the transmission models. But they are not simply taken over, they get endowed with additional inner structure. Let us go through the list of components of the general model: PARENTS, PROGENY, PHENOTYPES, GENOTYPES, MATOR, DISTRIBUTOR, APPEARANCE, DETERMINER, COMBINATOR. Whereas PARENTS and PROGENY in the general models are treated as unspecific basic objects, they are required to be non-empty sets of individuals in the transmission models. The PHENOTYPES in the general models are regarded as primitive objects which are not further analysed. In the transmission models the PHENOTYPES are also present, but now they are no longer primitive entities, they acquire some internal structure, being defined as k-tuples of EXPRESSIONS. The same happens with GENOTYPES. They change their status of unanalysed, basic objects into that of sequences of k pairs of FACTORS. Thus all three kinds of objects in the general models get *refined* in the same way: the objects are not eliminated, they are taken over, but they change their status from unanalysed,

'last' elements to more complex, defined structures. In order to achieve the required definition it is of course necessary to introduce other, 'final' building blocks in terms of which the previous ones, PARENT, PROGENY, PHENOTYPES, and GENOTYPES, may be defined. In fact, new 'atomic' objects are introduced in the transmission models: EXPRESSION, FACTORS, and INDIVIDUALS. We did not explicitly introduce a set of individuals such that each population is a subset of it, but only for reasons of economy. A set of individuals which form the populations may be easily introduced, however.[63]

So the transmission models contain more kinds of objects: EXPRESSIONS, FACTORS, and INDIVIDUALS are new. If we are keen on conceptual sparsity we could eliminate some of the 'old' kinds of objects, and treat them as explicitly defined in terms of the new ones. This might be done for GENOTYPES and PHENOTYPES, it does not work for POPULATIONS. We do not make use of this possibility.

The addition and refinement of objects induces corresponding refinements of the functions. MATOR, which originally mapped pairs of objects (PARENTS) into sets of objects (set of PROGENY) now maps pairs of sets of objects (sets of INDIVIDUALS, parental populations) into sets of such sets (sets of populations of progeny). Similar observations hold for DISTRIBUTOR and COMBINATOR.

Such addition of new objects and subsequent refinement of 'old' objects and functions constitutes one step in the transition from general- to transmission models which may be subsumed under the label *conceptual extension*. The term 'extension' is used because all syntactic stipulations about the 'old' concepts also hold in the 'new' models. A second step now consists in the addition of further assumptions which, in a certain sense, operate in narrowing down the additional possibilities obtained by the conceptual extension. Three additional assumptions were introduced in Chap.4: that DETERMINER be decomposable, that the coefficients of the distributions of phenotypes are defined as relative frequencies, and that the factors assigned to offspring are all among the factors assigned to their parents. All three assumptions place considerable restrictions on the terms involved, the first on DETERMINER, the second on DISTRIBUTOR, and the third on COMBINATOR. Though we hesitate to call the first two assumptions genetic laws, it seems admissible to lump together the effect of narrowing down the extensions of the terms exerted by all three assumptions under the label *specialisation of laws*. For this is their overall effect when seen from the point of view of the general models. The laws of the general models (the axiom of fit), get specialized by means of the three additional assumptions on DETERMINER, DISTRIBUTOR and COMBINATOR.

The result of this formal comparison may then be stated as follows. The models of transmission genetics are obtained from the general models by conceptual extension plus specialisation of laws. We suggest the term *refinement* for intertheoretical relations of this form. In abstract terms, a refinement consists in the addition of further kinds of objects and perhaps also of further functions

[63]See our treatment in (Balzer & Dawe, 1986a,b).

such that the 'old' objects can be defined as complex structures of 'new' objects, *and* in the introduction of further law-like assumptions concerning the 'old' and 'new' items in addition to the 'old' laws' holding also in the 'new' models. It is not difficult to see, and could be proved in a more precise setting, that refinement is transitive: if a class of models M_2 is a refinement of the models in M_1, and M_3 a refinement of M_2, then M_3 is a refinement of M_1.

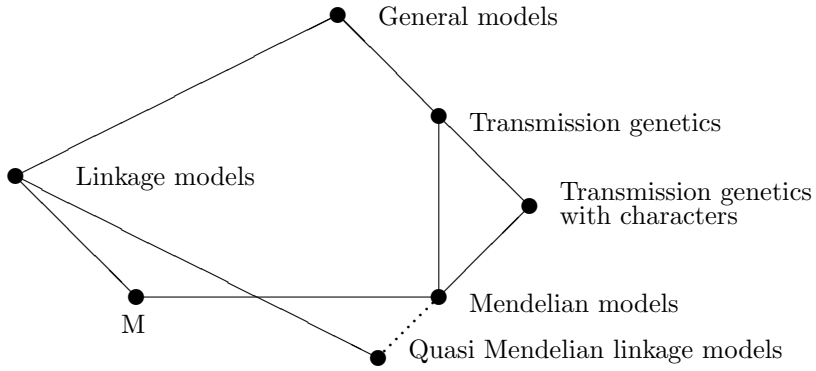
Let us next consider the relation between Mendelian- and transmission genetics. As defined in Chap.4, each Mendelian model is a transmission model. In addition, Mendelian models are conceptual extensions, for they contain the new sets of CHARACTERS and they contain specialisations of laws. The first special law is about DETERMINER and formulates the distinction between dominant and recessive factors,the second special law is the definition of COMBINATOR as yielding all possible combinations of factors with equal coefficients. So here we also have a relation of refinement. Mendelian models are a refinement of transmission models. From the transitivity of the refinement relation it follows, that Mendelian models also are a refinement of the general models.

The same relation obtains between linkage models and transmission models. By definition, each linkage model is a model of transmission genetics to which a genetic map f is added. So we have a case of conceptual extension, a case in which the objects on both sides are strictly identical, and only a new function is added. There is also a specialisation of laws given by the axioms for the genetic map. So linkage models are a refinement of transmission models.

The comparison of linkage and Mendelian models in terms of the above formal notion of refinement, on the other hand, yields a negative result. Mendelian models contain objects new with respect to linkage models, namely CHARACTERS, and vice versa, linkage models contain the genetic map which is new with respect to Mendelian models. So no side is a conceptual extension of the other. The same holds for laws. Linkage models are not required to satisfy the Mendelian form of COMBINATOR, and a Mendelian COMBINATOR will in general not satisfy the recombination axioms which impose features of linkage on COMBINATOR. So the two classes of models are independent of each other in the sense that no one is a refinement of the other. This is not to say that formal relations could not be produced, outside of our present reconstruction. For example, Mendelian genetics might be seen as a limiting case of linkage genetics when the genetic map distances are infinite (scientific practise does not appear to operate in this way, however). What we say here is strictly in respect of the models of genetics which have been produced in this book.

We may depict the result of these comparisons by drawing a little graph, the knots of which represent classes of models and the threads of which denote the relation of refinement (from above below):

Fig.7-1



This graph may be considerably enlarged by adding further special assumptions to those characterising the models introduced. For example, by adding the assumption about dominant and recessive factors to linkage models, we obtain a refinement of linkage models labelled 'quasi-Mendelian linkage models' in Fig.7-1. By adding characters to the models of transmission genetics we obtain a refinement of the transmission models, called 'Transmission genetics with characters' in Fig.7-1. By assuming that all assumptions of linkage and of Mendelian models are satisfied we would obtain another class of models, M^* in Fig.7-1 which, however, is of little scientific interest. Note that Mendelian models are also a refinement of those of transmission genetics with characters. That is, the lines in the graph may also form 'upward forks'. The occurrence of upward forks is compatible with our intuitive understanding of 'refinement': a class of models may be obtained from two different classes of models by refinement, i.e. refining two different classes in different ways may yield the same result.

A second dimension of comparison is given by the constraints. The constraint for the general models required that populations with the same phenotypes get assigned the same genotypes in all models in which they occur. This constraint also holds for the models in transmission genetics, as well as for Mendelian, and linkage models. In linkage genetics a further constraint was introduced, requiring that the map-values of a gene be the same in a species, provided the gene was assigned to populations with identical phenotypes (compare Chap.4). The pattern here is much the same as at the level of models. New 'objects' (species) are introduced, and a more special contraint is added to the 'old' ones. Thus the label 'refinement' may be extended to cover also the level of constraints, and Figure 7-1 above also represents the situation with respect to this extension.[64]

These two dimensions of comparison are concerned with the formal parts of the theories involved. However, empirical theories must not be seen as mere

[64]Strictly speaking, the lines now represent the *conjunction* of both kinds of requirements: for models *and* constraints.

formal entities. As pointed out in Chap.1 and Chap.2 there are at least two further, non-formal features of major importance. First, there are the real systems at whose explanation the theory aims, which we called intended systems. Their determination ultimately has to involve pragmatics in the form of *ostensions* or *ad hoc* decisions. Second, there is what we called the process of application of the theory to its intended systems. For two theories to be compared with each other these two features also have to be related in an appropriate way in order to obtain a 'real', satisfactory intertheoretic relation.

With respect to the intended systems the situation for most of the pairs of theories just considered is clear. In all cases where there is a relation of refinement it is pretty clear that the intended systems of the refined theory are also intended systems of the 'coarser' theory, but not necessarily vice versa. That is, the intended systems of the refined theory form a subset of the set of intended systems of the 'coarser' theory. Any intended system of transmission or Mendelian or linkage genetics is also an intended system of our basic genetic model, and every system intended for linkage and Mendelian genetics is also intended for the general transmission model. More precisely, if the community of linkage geneticists intends to apply their model of the genetic map to some particular real system (say a population of Drosophila) then the same community also intends to apply the general transmission model to that system. This is a consequence of the refinement relation obtaining between the two models. As the linkage model 'contains' the general transmission model any application of a linkage model to a Drosophila population automatically amounts to an application of the general transmission submodel to the same population. The only pair of models not standing in the refinement relation was that of linkage and Mendelian genetics. The lack of a nice inclusion between the models in this case is accompanied by a corresponding lack of inclusion of the intended systems. Though the sets of intended systems for both models have a large array of overlap there are other intended systems of either model which are not intended systems for the other model.

Similar observations hold for the whole process of application. If one theory is a refinement of another one then any successful process of application of the former comprises a successful process of application for the latter. Think of some intended system of Mendelian genetics, for instance, and the process of application of Mendelian theory to that system. According to Chap.2 this amounts to finding out which of the Mendelian concepts are realised in the system, determining as much data as possible which can be expressed in these concepts, and finally to checking whether the data can be embedded in a full Mendelian model. Consider as a first case a model of transmission genetics. The concepts used in a model of transmission genetics are just the same as used in the Mendelian model. So all the Mendelian concepts realised in the system will also be concepts of transmission genetics, and the first step in the process of application is the same for both models. This extends to the second step. Every determination of data for the Mendelian model may be regarded as a determination of the same data for the general transmission model, for the data are expressed in the form of concepts which are the same for both

models. The third step of checking whether the data obtained fit into a full model is also unproblematic. If the data fit into a Mendelian model then *a forteriori* they fit into a transmission model. This is due to the relation of refinement obtaining between the two models. By its definition, the Mendelian model satisfies all requirements stated for the transmission model. Now 'fit' amounts to an existential claim: 'there is a model into which the data can be incorporated'. So if there is a Mendelian model into which the data can be incorporated there also will be a transmission model into which they fit. For every Mendelian model by definition also *is* a transmission model. So each successful application of the Mendelian model yields a corresponding successful application of the transmission model.

The same reflections could be made for any other pair of models among which a relation of refinement was stated. As a second case let us consider the relation between Mendelian models and the basic genetic models described in Chap.2. In this case the refined (Mendelian) model uses concepts additional to those used in the basic model: EXPRESSION, FACTOR, SET_OF_FACTORS, and the component determiners DET_i. In principle therefore we might encounter the following situation: the only concepts realised in the intended system under study are those new ones typical for the *Mendelian* case, while none of the concepts from the basic model is realised. Though *a priori* possible this case will not occur because all the new Mendelian concepts more or less presuppose the 'old' concepts of the basic model as meaningful. The notion of an EXPRESSION is used to refine (and redefine) that of a PHENOTYPE, so whenever in an intended system we can realise EXPRESSIONS we also will realise some kind or other of PHENOTYPE. The same holds for FACTOR and GENOTYPE, and DET_i and DETERMINER. The only 'new' concept which does not really rely on 'old' concepts is that of a SET_OF_FACTORS according to which the factors are clustered to form allelic sequences. Now this concept in the Mendelian theory is the most theoretical one, and unlikely to be directly determined in the form of data. So it will play a role in the process of application only in the third step when collected data are fitted into a full model. In the present case this means that each Mendelian model (in which SETS_OF_FACTORS are used) also satisfies the requirements set forth for the basic genetic model (in which the SETS_OF_FACTORS do not occur). This implication simply holds by the definition of Mendelian models. To summarise, a successful process of application of the Mendelian model will yield a similar successful process for the basic genetic model. Among the Mendelian concepts realised in an intended system there will be most of the observational concepts of the basic model, and the collection of data will be the same for these when regarded from the point of view of the two models. Finally, if there is a Mendelian model into which the 'Mendelian data' can be fitted then there will be a basic model into which the subsets of basic observational data can be fitted.

Again, in the only non-comparable case that has occurred so far, that of linkage and Mendelian models, also the processes of application do not stand in a relation of inclusion with each other. There are applications of linkage genetics which cannot be regarded as applications of Mendelian genetics, and vice versa.

Not unexpectedly, the relations considered so far were unproblematic because we did not reach molecular genetics. We expect a different kind of relation to hold between molecular genetics and the other, 'classical' versions. The folklore is that, since chromosomes as well as processes in the cell ultimately consist in chemical reactions between macromolecules, molecular genetics yields a model or picture adequate to deal with all genetic applications. Therefore, it should be only a matter of time until all successful applications of the other models are reproduced by means of molecular models. This is a rather speculative view, and nearly of the same quality as other, philosophical reduction claims, like that of biology reducing to chemistry, chemistry to physics, or mathematics to set-theory. The point about such claims is not that they are right or wrong: they are of little practical import because they have no empirical content. They represent promises rather than empirical statements. If they are interpreted as empirical claims, they lack support.

If we try to compare molecular models with other models, which models should we consider at the other side? On our account the models of molecular genetics *by definition* are (general) models of genetics as described in Chap.2. So taking the latter as a counterpart for comparison will not yield thrillingly new insights. The models of molecular genetics are refinements of the general models. They are obtained from the latter in two steps. First, new objects are introduced: QUANTA at the level of DNA and of amino sequences, and the component functions DET_i of DETERMINER. PHENOTYPES and GENOTYPES are defined in terms of these new objects: they come out as configurations of strands of QUANTA. So molecular models are a conceptual extension of the general models. In a second step, the general laws are specialised by assuming that DETERMINER and COMBINATOR take special forms as required in AM4 and AM5. This means that we also have a specialisation of laws, and therefore the molecular models are indeed a refinement of the general ones.

The interesting questions of comparison turn up only when we consider transmission genetics, or one of its refinements, on the 'classical' side. Let us first concentrate on the general transmission models, and see how they compare with the molecular models. As the claim of most interest is that of a reduction of the former to the latter, let us first concentrate on the question whether transmission genetics in some sense is reduceable to molecular genetics.

There are no commonly accepted criteria for reduceability,[65] so we have to choose one particular approach to guide our investigation. We prefer an approach similar to the 'structuralist' account[66] the basic ideas of which will be sketched briefly. Before doing so, some terminological clarification seems in place. We will speak of the *reduced* theory and the *reducing* theory in order to refer to the two candidates to be compared. By this we do not want to imply that the reduced theory in fact reduces to the reduced theory. Rather, the labels are used even at a stage where just the possibility of reduction is investgated. In

[65] See (Nickles, 1973), (Schaffner, 1967), (Sklar, 1967), (Sneed, 1971), (Balzer, Moulines and Sneed, 1987), Chap.6, and (Pearce, 1987), Chap.4 for a sample.

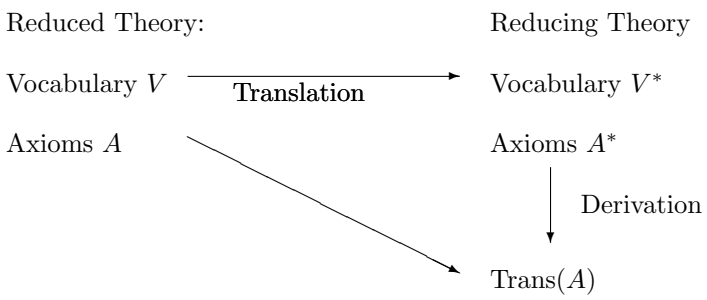
[66] (Sneed, 1971), (Balzer, Moulines and Sneed, 1987), Chap.6.

the present case the reduced theory would be transmission genetics, the reducing theory molecular genetics, but it is still open whether a relation of reduction,in fact, exists between the two.

There are at least three dimensions in which the two theories, reduced and reducing, have to be studied in order to see whether a relation of reduction really exists between them. These dimensions we met already in the comparisons made previously. First, there is the dimension of formal comparison of the models, second, there is the dimension of comparing the intended systems, and third, there is that of comparing the processes of applications on both sides.

Turning to the first dimension of formal comparison of the models, we have to investigate whether such comparison is possible. If it is possible we will say that there exists a relation of *formal correspondence*. The notion of a formal correspondence comprises two parts. First, it consists of a translation of the reduced theory's concepts into concepts of the reducing theory, and second, of a derivation of the translated axioms of the reduced theory from those of the reducing theory. In Figure 7-2 this is depicted schematically.

Fig.7-2

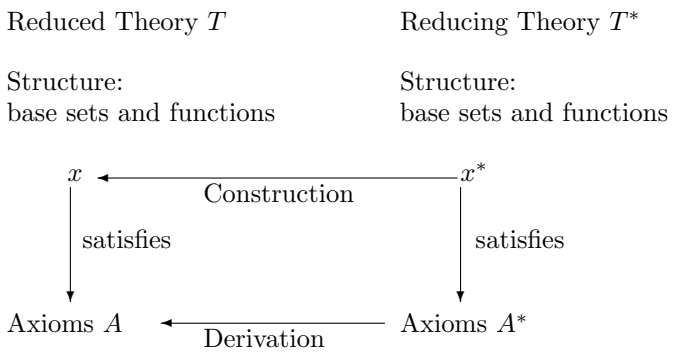


In our case this amounts to translating the transmission genetic terms into molecular terms, to restating the assumptions specific for the transmission models (AT1 to AT5) in molecular terms, and then deriving these translated assumptions from the axioms set forth for molecular models (AM1 to AM5).

In model theoretic terms this procedure roughly[67] amounts to the following. We start from a model of molecular genetics and try to define, or construct from the 'material' present in this model, the sets of objects, and the functions, a new model which is a model of transmission genetics. In abstract terms, we start from a model of the reducing theory and try to construct a model of the reduced theory out of it. This is schematically depicted in Figure 7-3.

[67]The connection between translatability and the model theoretic constructions to be used is not entirely straightforward, and requires some technical assumptions which, however, are of no interest here. A recent discussion is found in (Pearce, 1987).

Fig.7-3



The dotted arrows indicate that from the way of constructing x and from the assumption that x^* satisfies the axioms of the reducing theory, we may derive that x, in fact satisfies the axioms of the reduced theory.

The process of construction is not committed to start from an *arbitrary* model on the reducing side. Rather, we may choose the models we start with in a way which is most appropriate. The only requirement to be satisfied in this construction is that we be able to construct *all* models of the reduced theory (transmission genetics) in this way. In other words, we must be able to find, for every transmission model, some appropriate model of the molecular theory out of which we can construct the former. This requirement captures the idea that all models of the reduced theory must be reproduceable in some sense in the reducing theory. Instead of construction we also might speak of definition in this context, we prefer however the 'construction' terminology.

More precisely, such construction on the basis of a given molecular model involves the following steps. First, we have to construct or define the sets of objects for a transmission model, that is, the sets of genetic individuals, of PHENOTYPES,EXPRESSIONS,GENOTYPES and SETS_OF_FACTORS. Second, the same has to be done for the functions occurring in a transmission model: MATOR, APPEARANCE, DETERMINER, DISTRIBUTOR, COMBINATOR, as well as the component functions DET_i of DETERMINER. If these steps succeed, then in a final step we have to show that the base sets and functions thus constructed satisfy the axioms of transmission genetics provided the initial model is in fact a model of molecular genetics (i.e. its sets of objects and functions satisfy the axioms for molecular models).

The idea of actually constructing a model of the reduced theory in concrete applications turns out as rather difficult to realise. For this reason we will be satisfied with a weaker condition in which the model to be constructed is already given to some extent. Instead of defining this model from the beginning, we start with a structure of the type of such a model, and relate it to the given model of molecular genetics in a way which hopefully on further elaboration would give rise to a proper construction. On this account, we start with two structures: one model of molecular genetics (which is denoted by x^* in the following), and

one structure of the type of a model of transmission genetics (denoted by x). The model of molecular genetics need not be completely general, we may choose it to satisfy further special requirements as long as these in general still allow for a construction of every transmission model out of an appropriate molecular model satisfying the special requirements. The task then is to establish or define a relation of formal correspondence, denoted by μ in the following between the two structures which has two properties. First, it allows us to show that if x^* is a model of molecular genetics, and x is μ-related to x^* it follows that x is a model of transmission genetics. Second, the relation μ has to be such that it can be interpreted as a recipe for actually constructing x out of x^*. Intuitively, it will be helpful in the following to imagine both x^* and x to be descriptions of the same real genetic system, x^* describes the system in molecular terms whereas x describes it in transmission terms.

In order to establish such a formal correspondence μ the heuristic idea is of course to relate the pheno- and genotypes of both models appropriately in the first place. We will try to match the molecular GENOTYPES which are configurations of strands with the transmission GENOTYPES which are tuples of the form $\langle\langle a^1, b^1\rangle, ..., \langle a^k, b^k\rangle\rangle$. Such a match will, however, remain unsatisfactory as long as the ploidy explicitly present in the form of the transmission GENOTYPE is not represented in the molecular GENOTYPE.

The immediate reaction to this observation might be to say that our account of molecular genetics is inadequate, for we did not incorporate ploidy in the basic model. To this there are three replies. First, we may state that up to now ploidy does hardly play any role in applications of molecular genetics. Though molecular genetic applications in their majority stem from the two areas of haploid and diploid systems, this distinction itself is not relevant to the process of application itself. Second, we may of course introduce the notion of ploidy in molecular genetic models (see below). Third, and most importantly, we know from other cases of reduction[68] that reduction *always* involves one further step not yet mentioned. Reduction practically never obtains between the reduced theory and the *full* reducing theory. Rather, the normal case is that the reducing theory has to be refined (in the technical sense discussed previously) in order to achieve reduction. Thus the relation of formal correspondence does not obtain between the two theories in their basic form, it obtains between the reduced theory in its basic form on the one hand, and between a refinement of the reducing theory on the other hand. In order to reduce collision mechanics to Newtonian mechanics, for instance, we must not use the full range of Newtonian models as the reducing theory. The class of models to be used is that of Newtonian models which, in addition, satisfy the law of conservation of momentum. Similarly, in order to reduce rigid body mechanics to mechanics in general, we have to use models of mechanics at the reducing side which are more special in that their particles do not move relative to each other ('rigid' systems). Along these lines it seems most natural to choose a refinement of the molecular theory in which the notion of ploidy is introduced, as a reducing counterpart for the

[68] See (Balzer, Moulines, Sneed, 1987), Chap.6.

intended relation of formal correspondence.

To put it differently, the attempted formal correspondence will be established in two steps. In step one the molecular theory is refined by introducing the notion of ploidy. Only in step two then, can we try to establish a formal correspondance between transmission models and this refined class of models. Similar considerations are in place for the feature of proper populations to which transmission genetics is restricted. As the molecular theory was held neutral in this respect it is necessary to further specialise it to population genetics proper in order to obtain a theory which might reduce transmission genetics.

We therefore introduce a refinement of molecular models in which ploidy is present, and in which the models are restricted to populations proper. With respect to ploidy we will restrict ourselves to the diploid case, as we did in transmission genetics. Extension of our treatment to the haploid case, as well as to ploidy greater than two, is easily achieved.

In material terms, the basis of diploidy is the presence of two chromosomes of the same kind in each cell. So the introduction of ploidy in molecular models has to identify pairs of strands as 'belonging together'. In order to pair the 'right' chromosomes, it is necessary to characterise that two chromosomes, or strands in a configuration of strands, are of the 'same kind'. Instead of attempting a definition of 'same kind' in terms of the chromosomes' spatial form, we will take a more general, and simpler route. We introduce new basic concepts $R_1, ..., R_s$ each R_i denoting a set of strands of the same kind. In pairing, we may then just require that any two paired strands belong to the same set R_i. In order to simplify matters in the following, we will also introduce some order among the (pairs of) strands occurring in a molecular genotype, that is, we will pass over from a set N of strands to a tuple of pairs of strands $\langle\langle s_1, s_2\rangle, ..., \langle s_{r-1}, s_r\rangle\rangle$ such that $s_1, ..., s_r$ are exactly the elements of N. Each pair in such a tuple represents a pair consisting of two chromosomes of the same kind. This may be expressed by requiring that s_i, s_{i+1} are both in one of the sets R_j (for appropriate indices i). A similar ordering is performed with the sets of strands occurring in the PHENOTYPES.

All of these additional requirements are still not sufficient for the establishment of a satisfactory formal correspondance. There is a further problem arising in connection with the SETS_OF_FACTORS. In order to construct these sets out of the molecular strands, some corresponding distinction has to be introduced for the strands because in their present form we would not know which 'part' of a strand in fact 'is' a factor of a particular kind. Having put together all the strands occurring in one molecular GENOTYPE we may introduce the distinction of different appropriate 'parts' by means of a sequence of indices denoting the different places where to 'cut' the whole strand in order to obtain the different relevant parts. This will only work, however, if we make the further assumption that the overall 'length' i.e. number of QUANTA, in the concatenation of all the strands in a molecular GENOTYPE is the same for all GENOTYPES in one model. This assumption we did not want to make for molecular models in general (see Chap.5) but without it we see no way to establish a 'reasonable' formal correspondence. Under this assumption the two

overall strands obtained from the concatenations $s_1 \circ s_3 \circ s_5 \circ ... \circ s_{r-1}$ and $s_2 \circ s_4 \circ ... \circ s_r$ are 'cut' into an equal number of 'parts' by the indices $\tau_1, ..., \tau_{k^*}$, and any two such parts occurring in the same position in the two strands we call *opposed*. Adding the assumption that the genetic individuals be proper populations, we may summarise these requirements, and obtain the definition of a refinement of the models of molecular genetics, which we call models of diploid molecular genetics.

A *model of diploid molecular genetics* is defined as a structure

$$\langle \mathbf{I}^*, \mathbf{P}^*, \mathbf{G}^*, \mathbf{MAT}^*, \mathbf{APP}^*, \mathbf{DET}^*, \mathbf{DIST}^*, \mathbf{COMB}^* \rangle$$

for which there exist sets $\mathbf{P}$, $\mathbf{G}$, functions EX and COR, and numbers t^0, k^* and $\tau_1, ..., \tau_{k^*}$ such that

1) $\langle \mathbf{I}^*, \mathbf{P}, \mathbf{G}, \mathbf{MAT}^*, \mathbf{APP}^*, \mathbf{DET}^*, \mathbf{DIST}^*, \mathbf{COMB}^* \rangle$ together with EX and COR is a model of molecular genetics (as described in Chap.5)

2) the number r of strands in the GENOTYPES of $\mathbf{G}$ is even, and for $s := r/2$ there exist sets $R_1, ..., R_s$ such that each R_i $(i \leq s)$ is a set of strands occurring in GENOTYPES of $\mathbf{G}$, the R_i are pairwise disjoint, and together exhaust the set of all strands occurring in GENOTYPES of $\mathbf{G}$.

3) $\mathbf{G}^*$ is obtained from $\mathbf{G}$ by ordering the strands $s_1, ..., s_r$ of each set N in each GENOTYPE $\gamma = \langle N, \mathbb{R}^3, \psi \rangle$ of $\mathbf{G}$ in the form $\langle \langle s_1, s_2 \rangle, ..., \langle s_{r-1}, s_r \rangle \rangle$ such that for each odd index $j \leq r$, s_j and s_{j+1} are both in $R_{(j+1/2)}$

4) the number of QUANTA occurring in the concatenations $s_1 \circ s_3 \circ s_5 \circ ... \circ s_{r-1}$ and $s_2 \circ s_4 \circ s_6 \circ ... \circ s_r$ is equal to t^0 for all GENOYPTES of $\mathbf{G}$, and $\tau_1, ..., \tau_{k^*}$ are such that
   4.1) $1 < \tau_i < t^0$ for all $i \leq k^*$
   4.2) $\tau_1 < \tau_2 < ... < \tau_{k^*}$

5) each PHENOTYPE in $\mathbf{P}$ contains exactly s strands of amino acids, and $\mathbf{P}^*$ is obtained from $\mathbf{P}$ by ordering the strands of each PHENOTYPE $\pi = \{s_1, ..., s_s\}$ into a sequence $\langle s_1, ..., s_s \rangle$

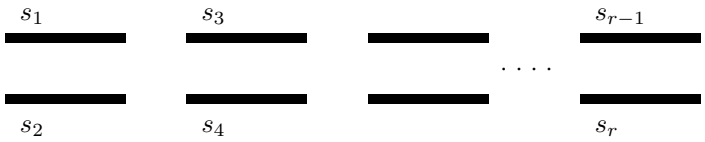
6) the number of QUANTA occurring in the concatenation $s_1 \circ ... \circ s_s$ is equal to t^0 for all PHENOTYPES in $\mathbf{P}$

7) DETERMINER* is such that the correlation function COR assigns a transmission strand to each pair $\langle s_i, s_{i+1} \rangle$ occurring in the concatenated strand according to 3), and there is a function EX*, compatible with EX such that EX* maps pairs of opposed parts into parts of the sequence of amino acids which forms the value of DETERMINER* such that parts of the GENOTYPE are mapped into parts of the PHENOTYPE with equal position function values

8) each element of $\mathbf{I}^*$ is a non-empty set.

The indices in 3) can best be understood from Figure 7-4.

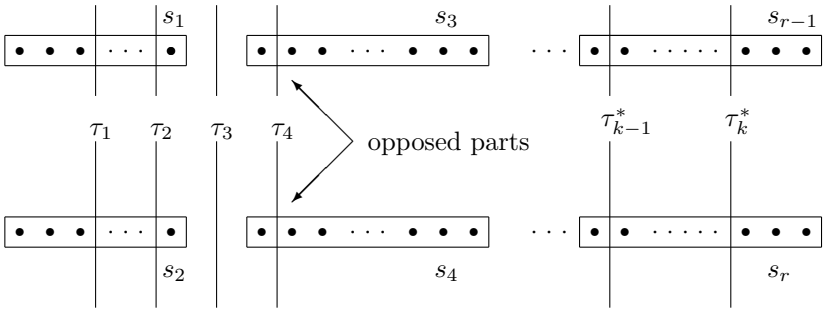
Fig.7-4



$s_1, s_2 \in R_1$, s_3,$s_4 \in R_2$, $s_{r-1}, s_r \in R_{r/2}$

Note that the spatial ordering of the strands $s_1, s_3, s_5, ..., s_{r-1}$ one after the other in this figure is artificial and does not depict any real feature. This ordering as well as the corresponding concatenation to be used in the following is performed purely conceptually. Note further that a similar definition for arbitrary ploidy can be obtained by replacing the number s in 2) by r/p, and the pairs in 3) and 7) by p-tuples. The numbers t^0, $\tau_1, ..., \tau_{k^*}$ may be called a *frame*. In Figure 7-5 a concatenated sequence of strands as occurring in $\mathbf{G}^*$ is depicted in the upper row with an indication of the QUANTA of each strand. t^0 is the overall number of QUANTA occurring, and each number τ_i marks a position of cut indicated by the vertical lines. These cuts divide the whole string into $k^* + 1$ 'parts'. At the left, part of the second strand is depicted with two opposed parts being indicated.

Fig.7-5

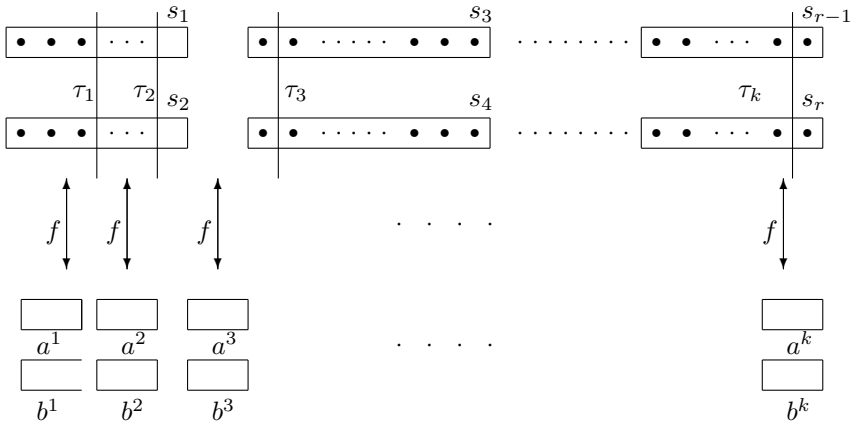


Condition 7) adjusts DETERMINER to the diploid case, and also decomposes it with respect to the parts into which the concatenated strands are cut by the frame $\tau_1, ..., \tau_{k^*}$.

We now may establish a formal correspondence between models of transmission genetics and appropriate models of diploid molecular genetics. At the level of GENOTYPES, for two given GENOTYPES on either side what we have to do is to put the strands of the molecular GENOTYPE together so that they

correspond in a natural way to the two 'strands' $\langle a^1, ..., a^k \rangle$, $\langle b^1, ..., b^k \rangle$ present in the transmission genotype. To this end we concatenate the pairs of strands occurring in the sequence $\langle \langle s_1, s_2 \rangle, ..., \langle s_{r-1}, s_r \rangle \rangle$ so that we obtain one big comprehensive pair of 'superstrands', each 'superstrand' of which is obtained from $r/2$ of the original strands of the GENOTYPE by concatenation in the technical sense introduced in Chap.3. As stressed earlier, this concatenation has to be imagined as a purely conceptual operation. The pair of superstrands thus obtained now may be matched with the strands of the transmission model way as depicted in Figure 7-6.

Fig.7-6



The third vertical double arrow from the left indicates a case where QUANTA from two STRANDS together form one transmission FACTOR. We have not ruled this out, though it might easily be done. For the two GENOTYPES to be in correspondence there has to exist a function g mapping the (pairs of) parts defined by the frame $t^0, \tau_1, ..., \tau_{k^*}$ in the molecular GENOTYPE one-one onto the (pairs of) factors occurring in the transmission GENOTYPE. As the different parts occuring in the overall strand may be chemically identical we have to add some further index in order to differentiate between such parts, identical otherwise. Such an index is provided by the spatial positions as given by ψ in the configuration of strands. So g, in fact, has to map pairs of parts and their respective positions onto pairs of factors. Moreover, this mapping should respect the order as depicted in Figure 7-5: parts on the same 'superstrand' should be mapped into factors on the same 'transmission strand', and the ordering induced among the parts by the ordering $<$ of quanta should go over into the ordering by indices of their images, for instance, for two parts p, p' on the 'upper' strand:

$$\text{if } p < p', g(p) = a^i \text{ and } g(p') = a^j \text{ then } i < j.$$

The whole sets of GENOTYPES of two models x and x^* correspond to each other if there is a function h mapping each GENOTYPE from the molecular set

one-one onto some GENOTYPE from the transmission set such that between the two GENOTYPES there is a function g as just described. We note that this kind of correspondence can be sharpened with a little extra effort to yield an effective recipe for constructing the transmission GENOTYPE from the molecular one.

PHENOTYPES being defined as k-tuples of EXPRESSIONS on the one side and as sequences of strands of amino acids on the other (in diploid models) can be formally matched as follows. Given one phenotype at either side we simply take all the strands of the molecular sequence and join them conceptually to form one big 'superstrand' which may be cut into appropriate pieces by means of the frame $t^0, \tau_1, ..., \tau_{k^*}$ just as was done for the GENOYTPES. The parts of this overall strand may then be mapped one-one onto the EXPRESSIONS of the PHENOTYPE from the transmission model in a way that preserves order. This correspondence cannot be turned into a construction proper. In order to go from the amino acids to the EXPRESSIONS in the transmission PHENOTYPE we would have to go beyond the possibilities of definition and construction. The real EXPRESSIONS are caused by the amino acids in a very complex way which cannot be described in the vocabulary of the genetic models presented here. This is why we have to resort to an abstract kind of match as described above in which the transmission PHENOTYPE is not constructed but taken as given, and its components are related to the quanta occurring in the molecular phenotype in some suitable way. Note that the number k^* has to be related to the number k of EXPRESSIONS by $k^* + 1 = k$.

Formally, the match between PHENOTYPES on both sides will be provided by a function i mapping the molecular PHENOTYPES in models x^* one-one onto the transmission PHENOTYPES of the structure x such that, for each π^* of the form $\langle s_1, ..., s_k \rangle$ there is a function j mapping the parts of $s_1 \circ ... \circ s_k$ as given by the frame $t^0, \tau_1, ..., \tau_{k^*}$ one-one onto the EXPRESSIONS of $i(\pi^*)$.

The populations in the transmission model can be obtained from a molecular model in which just these populations are taken as genetic individuals. As stated previously, there is some freedom in the choice of the molecular models, which is exploited now.

It is easy to see that these correspondences between the sets of objects on either side yield natural correspondences of the functions, too. Let $x = \langle \mathbf{I}, \mathbf{P}, \mathbf{G}, \mathbf{MAT}, \mathbf{APP}, \mathbf{DET}, \mathbf{DIST}, \mathbf{COMB} \rangle$ and $x^* = \langle \mathbf{I}^*, \mathbf{P}^*, \mathbf{G}^*, \mathbf{MAT}^*, \mathbf{APP}^*, \mathbf{DET}^*, \mathbf{DIST}^*, \mathbf{COMB}^* \rangle$ be two structures such that x^* is a model of diploid molecular genetics, x is a structure of the type of the models of transmission genetics, and let the base sets **I,P,G** and $\mathbf{I}^*, \mathbf{P}^*, \mathbf{G}^*$ be related as just described. In this case MATOR may be taken as identical with MATOR*, and APPEARANCE can be defined by

(**9**) $$\text{APPEARANCE}(u) = i(\text{APPEARANCE}^*(u))$$

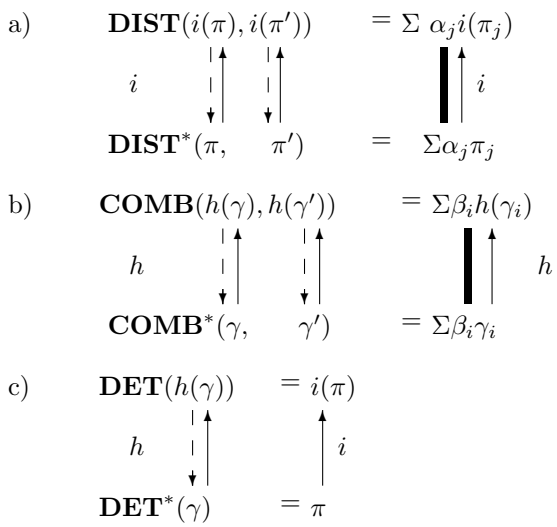
for each population u in $\mathbf{I} = \mathbf{I}^*$. Similarly, DISTRIBUTOR, COMBINATOR and DETERMINER on the transmission side may be defined by

(**10**) $$\text{DISTRIBUTOR}(i(\pi), i(\pi')) = \sum \alpha_i i(\pi_i)$$
$$\text{COMBINATOR}(h(\gamma), h(\gamma')) = \sum \beta_i h(\gamma_i), \text{ and}$$

$$\text{DETERMINER}(h(\gamma)) = i(\text{DETERMINER}^*(\gamma))$$

where $\text{DISTRIBUTOR}^*(\pi, \pi') = \sum \alpha_i \pi_i$, $\text{COMBINATOR}^*(\gamma, \gamma') = \sum \beta_i \gamma_i$, and h, i are the functions described previously. These definitions are depicted in Figure 7-7.

Fig.7-7 a)



The dotted arrows indicate that the respective functions are calculated in reverse direction. As both h and i are one-one and onto, this is possible. The thick arrows indicate identity.

Let us summarise this account of a formal correspondence between the transmission and the molecular levels by the definition of a formal correspondence μ between structures of the type of transmission models, and models of diploid molecular genetics. We write

$$\mu(x, x^*)$$

to abbreviate that the transmission structure $x = \langle \mathbf{I}, \mathbf{P}, \mathbf{G}, \mathbf{MAT}, \mathbf{APP}, \mathbf{DET}, \mathbf{DIST}, \mathbf{COMB} \rangle$ and the diploid molecular model $x^* = \langle I^*, P^*, G^*, MAT^*, APP^*, DET^*, DIST^*, COMB^* \rangle$ formally correspond to each other by μ. We stipulate that this relation $\mu(x, x^*)$ holds if and only if there exist functions g, h, i, j such that

1) $\mathbf{I} = \mathbf{I}^*$

2) the number k^* given by the frame for x^* and the number k of EXPRESSIONS in x as well as of pairs of FACTORS in GENOTYPES of x, are related as $k^* + 1 = k$, and
$i : \mathbf{P}^* \rightarrow \mathbf{P}$ is one-one, and onto

$h : \mathbf{G}^* \rightarrow \mathbf{G}$ is one-one, and onto

3) for all γ^* in $\mathbf{G}^*$, g maps the parts defined by the frame t^0, $\tau_1, ..., \tau_{k^*}$ occurring in strands of γ^* and their positions one-one onto the FACTORS occurring in $h(\gamma^*)$ such that order is preserved

4) for all π^* in $\mathbf{P}^*$, j maps the parts defined by the frame t^0, $\tau_1, ..., \tau_{k^*}$ occurring in the strand of π^* one-one onto the EXPRESSIONS in $i(\pi^*)$ such that order is preserved

5) $\mathbf{MAT} = \mathbf{MAT}^*$ and **APP, DET,DIST**, and **COMB** are defined by (9) and (10) above.

Note that we start from models of *diploid* molecular genetics, so π^* and γ^* are not sets but sequences of strands.

According to the general explanations on reduction given previously the following two conditions for μ have to be investigated. First, we have to see whether from $\mu(x, x^*)$ we may infer that x is a proper model of transmission genetics, provided x^* is a model of diploid molecular genetics, and second we have to see whether for all models x of transmission genetics there is some appropriate model x^* of diploid molecular genetics such that $\mu(x, x^*)$ holds. As far as the second condition is concerned, we may state the following

Theorem 2 For every model x of transmission genetics there is a model x^* of diploid molecular genetics such that $\mu(x, x^*)$ holds.

The proof is straightforward, but tedious, and suppressed here. The first condition cannot be met, it fails for AT5 which does not follow from the molecular axioms and the stipulation for μ. We believe that nevertheless there is some interest in looking more closely at those transmission axioms which can be proved, and we will return to the overall failure after this.

Theorem 3 For all x, x^*, if x^* is a model of diploid molecular genetics and x is a structure of the type of models of transmission genetics, and if $\mu(x, x^*)$ holds then A7, A8 from Chap.2 and AT2 to AT4 from Chap.4 are satisfied in x.

We will give the proof in an informal way. Of structure x it is required that x be of the right form, that is, x should have the form $x = \langle \mathbf{I}, \mathbf{P}, \mathbf{G}, \mathbf{MAT}, \mathbf{APP}, \mathbf{DET}, \mathbf{DIST}, \mathbf{COMB} \rangle$ precisely described at the end of Chap.2. It is *not* assumed that x satisfies the axioms of transmission genetics, nor those general axioms which go beyond fixing x's type. In detail, the axioms not assumed to hold are A7 and A8 from Chap.2, and AT2 - AT5 of Chap.4. It has to be proved, then, that these axioms follow from those holding for x^*, and from the stipulations put forward for μ. A7: By 1) and 5) of the definition of μ, the right hand side of A7 is the same both for x and x^*. So it suffices to show that the same is true for the left hand side. This follows from the first equation in (10). A8: A8 is assumed to hold in x^*, i.e.

$\mathbf{COMB}^*(\gamma,\gamma')(\gamma^*) \approx_\epsilon \mathbf{DIST}^*(\mathbf{DET}^*(\gamma),\mathbf{DET}^*(\gamma'))(\mathbf{DET}^*(\gamma^*))$. Mapping γ,γ' and γ^* into $h(\gamma),h(\gamma')$ and $h(\gamma_*)$, respectively, we obtain from (10) that the left hand side is transformed identically into $\mathbf{COMB}(h(\gamma),h(\gamma'))(h(\gamma^*))$. The right hand side is shown to be equal to

$$\mathbf{DIST}(\mathbf{DET}(h(\gamma)),\mathbf{DET}(h(\gamma')))(\mathbf{DET}(h(\gamma^*))).$$

So the approximate equality has to hold in x, too. AT2: We have to note that AT2 comprises two requirements, one of typification, and one goying beyond that. The requirement of typification is that each GENOTYPE is a tuple of length $2k$. This requirement cannot be derived but has to be presupposed. In order to prove the 'rest' of AT2 we first define the SETS_OF_FACTORS_i for $i = 1,..,k$. Let $i \leq k$ be given. By stipulation 2) for μ the number k^* of 'cuts' in the GENOTYPES of $\mathbf{G}^*$ is equal to $k-1$, so that they cut each overall strand in just k 'parts'. Each such part has a position as determined by the ordering of QUANTA on the strands, and by the order of concatenating the original strands. We then set SET_OF_FACTORS_i to be the set of all function values of the form $g(p)$ where p is a part in position number i of some GENOTYPE in $\mathbf{G}^*$. From this definition we obtain immediately that there are exactly k SETS_OF_FACTORS, and that any two FACTORS e^i, e^j such that i is odd and $j = i+1$ are from the same SET_OF_FACTORS. AT3: The component DETERMINERS DET_i are defined as follows. Let $s \leq k$ be given. Then DET_s is defined on pairs of factors from the SET_OF_FACTORS_s defined previously, and its value is obtained as follows. Let $\gamma = \langle\langle$ FACTOR$(1,1)$,FACTOR$(1,2)\rangle,...,$ $\langle$ FACTOR$(k,1)$,FACTOR$(k,2)\rangle\rangle$ be a GENOTYPE in $\mathbf{G}$. As function h was assumed to be onto, there is some GENOTYPE γ^* in $\mathbf{G}^*$ such that $h(\gamma^*) = \gamma$. Take the opposed parts of γ^* present at position number s in γ^*, and consider the corresponding value of the function EX* required in 7) of the correlation of the base sets. This is the part of DETERMINER$^*(\gamma^*)$ in position number s. Now consider i(DETERMINER$^*(\gamma^*)$) and take the s-th part of it. By 4) of the definition of μ this part, call it z, is just the j-value of the original part of DETERMINER$^*(\gamma^*)$. So z is uniquely determined by the whole procedure. We then set DET_s(FACTOR$(s,1)$,FACTOR$(s,2)$) equal to z. By 4), z is an EXPRESSION in x, so the second equation in AT3 is satisfied. We still have to show the first equation in AT3, namely that DETERMINER($\langle\langle$ FACTOR$(1,1)$, ...,FACTOR$(k,2)\rangle\rangle$ = $\langle\langle$ DET_1(FACTOR$(1,1)$,FACTOR$(1,2)\rangle$,...,$\langle$ DET_k(FACTOR$(k,1)$,FACTOR$(k,2)$)$\rangle\rangle$. This follows from equation 3 of (10) and the construction of the DET_s, $s \leq k$. AT4: This follows from AM5 and the weak conservation principle built into the definition of combinations kinematics, together with (10).

As already stated, AT5 cannot be proved in this way. There is no assumption in the molecular models by which the coefficients of the DISTRIBUTIONS_OF_PHENOTYPES are defined as relative frequencies. Such a requirement cannot be put forward simply because the molecular models are intended to deal with individual (non-population) applications as well. Why not add a corresponding assumption in the definition of *diploid* molecular models? Well, this might be done, and we could then extend the proof of Theorem 3 to a full

proof that x is a proper model of transmission genetics.

Purely formally, these results show that between the models of molecular genetics and those of transmission genetics a formal correspondence of the 'reductive' type can be established if sufficiently strong refinements are first made at the molecular side. With respect to the full class of molecular models this means that a formal correspondence exists between transmission genetics and an appropriate refinement of molecular genetics. The two conditions expressed by Theorems 2 and 3 may be rephrased as follows. Theorem 2 says that each transmission model has a counterpart in the molecular theory, or can be reproduced in the molecular theory. Theorem 3 says roughly that the axioms for transmission models (except AT5) follow from those for diploid molecular models, if these are translated along the lines of μ.

Can we conclude from these formal results that transmission genetics is reduceable to molecular genetics? Besides the purely formal dimension considered there are the other two dimensions discussed previously which are as important as the formal one.

The second dimension of comparison was that of intended systems. For a reduction relation to obtain, all intended systems of the reduced theory should 'be', or correspond to, intended systems of the reducing theory: every system to which the community of geneticists intends to apply transmission genetics also should be such that the community intends to apply molecular genetics to it. This condition is of course very hard to evaluate. We certainly have to acknowledge that there is an area of overlap in the intended systems of both theories. The example of sickle cell anaemia which provides intended systems for both was discussed in detail. This is an example in which geneticists not only intend to apply the theories, but in which both theories actually have been applied. However, this kind of situation is not so frequent as to be dominant. There are many other cases in which systems intended for transmission genetics simply are of no interest to molecular genetics. Think of the spread of an 'ordinary' trait like wrinkled seed in a population, or of the phenomenon of mixing of genes mentioned in Chap.1. It can be stated that systems of this kind have not given rise to molecular application proper up to now.

We must be clear about the difference between this observation and the claim that such systems are not intended systems of molecular genetics. The latter claim is much stronger, and amounts to claiming that molecular geneticists do not intend at all to apply the molecular theory to such systems. As such intentions may be entirely about future actions and developments the fact that no molecular application has taken place yet, cannot taken as confirming or disconfirming evidence. We therefore cannot claim that the typical transmission systems just referred to definitely are not taken as intended systems in molecular theory. The best that can be said in this direction is that the present situation allows for doubts about the inclusion of intended transmission systems in those of molecular theory.

This brings us to the third dimension of comparison, namely that of the processes of application. For a proper relation of reduction it is necessary that all application processes of the reduced theory be reproduceable in the reducing

theory. In the case before us, all processes of applying transmission genetics should be reproduceable in molecular genetics. It is at this point that the extravagancy of a reduction claim becomes visible. Consider some 'ordinary' application of transmission genetics to the spread of some gross trait in a population. As described in Chap.4 the process of application in this case amounts to establishing data about the relative frequencies of the trait, to make up an hypothesis about COMBINATOR and DETERMINER, and to see whether the theoretical distribution obtained by evaluating COMBINATOR fits to that of relative frequencies with sufficient degree. Even in cases of systems to which the molecular model is applied this process of application can hardly be said to be reproduced on the molecular side. The process of applying molecular genetics differs markedly. Here, the collection of data consists of identifying strands and the chemical quanta on them. Hypotheses about COMBINATOR are usually taken over from the transmission side, and only DETERMINER is fixed by molecular theory.

It is hard to see how in such cases it may be claimed that the process of applying the molecular theory can be reproduced on the side of transmission theory. Given a molecular process of application there are no hints at how the DETERMINER should be chosen in a corresponding application of transmission genetics, and the same is true for COMBINATOR. Moreover, the molecular application does not provide any data about DISTRIBUTOR in a corresponding transmission model. Even without attempting further clarification of what is meant by 'reproduction' of a process of application it is obvious that the molecular processes of application cannot be reproduced in transmission genetics.

Rather, the picture emerging from consideration of such processes is a picture of completion. When seen from our basic model of genetics the processes of applying molecular theory seem to complete those of applying transmission genetics, as was already noted in Chap.1 and Chap.5. Perhaps not always, but in many cases anyway, it is only after a first round of observation and fitting of appropriate hypotheses to observed distributions of gross characters that molecular methods come into play. Without these foundations in the transmission branch, molecular applications proper would concern only a small part of our model, centering on DETERMINER in either of its (molecular or transmission) versions.

These considerations may be summarised by stating that even though there is the possibility of formal correspondence between the models of transmission and diploid molecular genetics, we cannot claim that this amounts to a relation of reduction between the two branches. The situation with respect to intended systems is undecided, but with respect to the processes of application there is no overall, homogenous way to correlate such processes on both sides. Even on the purely formal side the situation is not very satisfactory. We have gone into the details of spelling out a formal correspondence and the details of the proof of Theorem 3 in order ot make clear that this correspondence is more of a formal exercise than of doing gentics. Roughly, the strategy that becomes visible in this formal part is this. In order to reproduce transmission models, try to do so, and, whenever there is some point of failure, add a corresponding assumption

to those of molecular theory so that you can go on. It cannot be said that the assumptions added about ploidy, decomposition into parts, populations proper, are nonsense from the genetic point of view. What can be said, however, is that these assumptions are added *ad hoc*. They were not founded upon independent consideration of molecular theory, but were 'detected' only in the context of comparison. To put it differently, these assumptions did not have a standing of their own in molecular genetics.

There are further negative points, however. Up to now we only compared the general transmission models with those of (diploid) molecular genetics. In order to have a full reduction of transmission to molecular genetics we should also be able to find molecular counterparts for Mendelian and linkage genetics. We might set out and look for formal correspondences for these two refinements, too. In fact, we can find appropriate refinements of molecular models which make such formal correspondence possible. The strategy for their production is that described previously. Further assumptions have to be made, using new basic concepts in order to obtain refinements with which Mendelian and linkage models could be correlated.

Refinements obtained in this way again have an air of being contrived, however. They have no standing in molecular genetics. They would be artificially invented just for the sake of reduction. Although the possibility of their construction seems to demonstrate the richness of molecular models, this by itself cannot be taken as an argument for overall reduceability. At the level of comparison of whole theory nets it is not sufficient for reduction that for each refinement on the reduced side there can be constructed a corresponding refinement on the reducing side. Such a weak notion involving mere possibilities can indeed be shown to follow logically from a formal correspondence of the respective core models on either side. In order to have reduction at the level of theory-nets for each refinement on the reduced side there has to be a corresponding refinement on the reducing side which has to exist and to be acknowledged before the reduction relation is at issue. This condition is not met for the case before us.

Chapter 8

Conclusion and Perspectives

One central theme of this book has been to abstract unifying features in genetics. We have used metatheoretical models which proved fruitful in many other disciplines in order to give more inner structure to the representation of the whole discipline. This has allowed a broad account of genetics to be rigorously treated. Our way of showing the unity of the field was to start with one common, basic model from which all other models subsuming the different areas of application could be obtained by successive refinement. In the net structure of the discipline thus becoming apparent, we have taken transmission genetics and molecular genetics as central. In fact, a marked similarity in the basic or 'core' structures of each was found, although there were many differences at the level of specialisation or refinement. It is these differences in specialisations which account for the obvious differences between the subjects apparent in textbooks and scientific papers. Indeed, if this distinction between the core structure and refinements is maintained, it is possible to show a unification of molecular and transmission genetics.

The overall net structure of the discipline's models deserves further recognition. Besides showing that the field is unified, and precisely how its parts are interrelated it may also be regarded as an instrument for guiding research. Research is basically conservative in that new hypotheses which have to be advanced in order to keep theory consistent with the data are usually chosen such as to only deviate from the established core of models as much as is really necessary. This conservative strategy can be substantiated to some extent using the net picture of a discipline. In the case of contradiction one has to give up one node in the net but one wants to keep as many other nodes as possible. If the nodes are related by refinement only, this may be achieved simply by going ' upward' in the net to the 'next coarser' node, and by trying to refine this in a new way such that fit to the data can be achieved. Only if this is not possible will one have to go further 'upward', but examples of such more comprehensive moves are rare. We would mention that such a distinction between core models, refined models, and the corresponding net structure is not readily made in any other account of which we are aware.

We see another achievement of the present work in the simple fact that we did succeed in axiomatising the various genetic subfields. Roughly, this amounts to choosing a fixed list of terms, or primitives, and a definite list of axioms or basic assumptions such that all, or at least most, of the statements and claims made in the field can be derived from those either directly, or by intermediate

steps of further refinement of the assumptions. In actually presenting such sets of assumptions we are in direct opposition to claims that in genetics no set of basic axioms can be made out.[69] By looking at the textbooks, or by looking into experimental research this impression might perhaps be created because geneticists, as any other scientists, are usually concerned with genetic research, and not with questions as to whether this or that statement might be more central or allow for the derivation of many or fewer other statements. However, as stated already in Chap.7, once a discipline becomes comprehensive there is a drive towards clarification and simplification, be it for purposes of review, teaching, planning and design of research strategy or indeed of scientific argument in the context of scientific progress. We think that our analysis shows that axiomatisation is possible, and we also think that some substantial steps in this direction have been made. Of course, we do not deny the possibility of further improvement.

Another way of looking at the issue just addressed is in terms of application. In the process of application as described in some detail before, usually a great number of additional ad hoc assumption are made which are not all covered by the theory's axioms, even if the theory used is already very special. In order to do empirical research, it is advantageous to make as many such special assumptions as necessary. This yields one successful application. However, there is a problem with this strategy as soon as we come into a more advanced stage where there are many different applications. How can these be systematised if each of them is dependent on a large number of assumptions particular to it but not to other systems? Developing theoretical models involves abstracting from the features peculiar to only single, or few, systems. We think that we have achieved some degree of such abstraction. This also holds perhaps with even more weight in attempts at comparing different models and theories. The analysis underlines the problems which will result when relations between differing areas of genetics are sought at a 'local' level of particular applications rather than at the level of comprehensive theory-cores. Not that such relations are impossible to demonstrate, but that they rely on special features of the particular application considered and do not add clarity to the overall situation.

On the other hand, we have tried to keep things legible. We have avoided formalism whenever possible. In chapter 7 the limitations of this have become visible. When things become somewhat complicated symbolic notation is unavoidable in order not to lose sight of some of the details.

Although our underlying thesis has been one of unification, and in this we have not been disappointed, the problem of intertheoretic relation between molecular genetics and transmission genetics when seen as separate entities is appreciated. We feel that in order to carry out such an analysis, an adequate formalisation of the areas concerned is necessary. When this is done, it is evident that forming direct relations between fully fledged theories is indeed very complicated, and there is some difficulty in finding a fully satisfactory result.

While it was not our intention, we also found that one major concept from

[69] (Kitcher, 1982).

genetics cannot be formulated unambiguously, ironically, the 'gene' itself. Other writers have studied the variation found in this concept non-formally with similar findings. Essentially, each 'gene' has to be seen against its theoretical backdrop. Thus, it is difficult to take the 'gene' as a primitive concept, and we do not. If the relations between different areas of genetics are studied in a manner which takes the 'gene' as a primitive, difficulties are likely to arise. One particular problem here turns up when we try to identify the 'gene' in transmission genetics with 'DNA' in molecular genetics, a popular identity. We then overlook that the DNA molecule only gains genetic status by virtue of the Watson-Crick model, that is, by means of a corresponding theory. Therefore, the identification cannot be made unless it is shown that the two versions of the concept as determined in the two corresponding theories, in fact are compatible. As far as we know, no such attempt has been made up to now.

Important though the question of the relation between molecular and transmission genetics is, and its bearing on the unity of genetics, we have been able to generalise further. Our models are compatible with those of the genetic algebraist, providing access to powerful tools of analysis in population genetics, and to multiple generations. As shown in Chap.6 our models may be taken as providing the empirical genetic basis of the mathematical formalisms of genetic algebras as well as of pedigrees. We made some effort to show how our unified models would operate in contemporary studies of pedigrees.

Our models point towards two further applications, one of which we anticipated, while the other one became clear only when the overall structure was apparent.

Certainly from the beginning we were aware of the great potential which formalisation of the kind undertaken has for computer application. Indeed, we even adjusted our notation to this, at least to some extent. Thus turning all the expressions written in capital letters into lower case letters we have all the concepts ready to be typed into the computer in PROLOG, which at the moment is *the* language for AI applications.[70] On the basis of our definitions it is fairly easy to formulate rules for expert systems for the areas covered by our different models. The only thing we did not do, is to transform our axioms into production rules. Though this may not be entirely trivial it seems not too difficult for an experienced programmer. Thus our models may form the basis of several expert systems for the areas of genetics treated.[71] In contrast to the way expert systems are usually developed, the way open on the basis of our models is more permeated by theory. In order to arrive at useful applications of the computer along these lines, further special assumptions (in the form of further production rules) will have to be added to the basis provided by our models. We believe we may claim to have gone a long way towards the end of computerisation of global models of genetics, and we hope to go further in that direction in the future.

Another perspective less evident in the beginning of our work which has

[70] A legible introduction to PROLOG is (Clocksin and Mellish, 1984).
[71] The best known expert system in genetics is MOLGEN, see (Stefik, 1981a,b).

now become apparent is this. One line for future investigation by geneticists undoubtedly concerns the relation between molecular processes and evolution. It is increasingly accepted that crossing over and recombination are not random processes, as had once been imagined. Furthermore, they may be under genetic control. Certainly, there are 'hot spots' for crossing over, making those progeny with certain combinations of characters more likely to appear. If this is considered against the backdrop of natural and sexual selection, there arises an interesting interplay between molecular genetics and evolution. The study of evolution has historically more in common with that of transmission genetics, however. Conceivably, a better understanding of the overall structure of genetics may assist in research into this interplay between evolution and molecular genetics.

REFERENCES

Anderson, E.G., 1925: Crossing over in a case of attached X chromosomes in Drosophila Melanogaster, *Genetics* 10, 403-17.

Balzer,W., Moulines, C.U. and Sneed, J.D.: 1987: *An Architectonic for Science*, Dordrecht: Reidel.

Balzer,W. and Dawe,C.M., 1986a: Structure and Comparison of Genetic Theories: I Classical Genetics, *BJPS* 37, 55-69.

Balzer,W. and Dawe,C.M., 1986b: Structure and Comparison of Genetic Theories: II The Reduction of Character Factor to Molecular Genetics, *BJPS* 37, 177-91.

Balzer, W., and Sneed, J.D., 1977: Generalized Net Structures of Empirical Theories, *Studia Logica* 36, 195-211, and *Studia Logica* 37, 167-94.

Bassett,H.L., 1931: see (Lawrence, 1950).

Bateson, W. and R.C.Punnett, 1905: Experimental Methods in the Physiology of Heredity, Report to the Evolution Committee Royal Society II, London: Harrison and Sons.

Bauer, H., 1974: *Wahrscheinlichkeitstheorie und Masstheorie*, Berlin-New York: de Gruyter.

Beet, E.A., 1949: The Genetics of the Sickle-Cell Trait in a Bantu Tribe, *Ann.Eugenics* 14, 279-84.

Cannings, C., Thompson, E.A. and Skolnick,M.H., 1978: Probability Functions on Complex Pedigrees, *Adv.Appl.Prob.* 10, 26-61.

Cannings, C. and Thompson, E.A., 1977: Ascertainment in the sequential sampling of pedigrees, *Clinical Genetics* 12, 208-12.

Carlson, E.A., 1966: *The Gene: A Critical History*, Philadelpia: Saunders & Co.

Clocksin,W.F. and Mellish, C.S., 1984: *Programming in PROLOG*, (2.ed.), Berlin-Heidelberg-New York-Tokyo: Springer.

Crick, F.H.C. and Watson, J.D., 1953: The Structure of DNA, *Cold Spring Harbour Symposia for Quantitative Biology* 18, 123-31.

Culp, S. and Kitcher, P., 1989: Theory Structure and Theory Change in Contemporary Molecular Biology, *British Journal for the Philosophy of Science* 40: 459-483.

Dawe, C.M., 1982: The Structure of Genetics, PHD Dissertation, London University.

Dawe, M.S. and Dawe, C.M., 1994, Prolog for Computer Science, London: Springer.

Devoret, R., 1988: Molecular Aspects of Recombination, in Michod,R.E. and Levin,B.R.: *The Evolution of Sex: An Examination of Current Ideas*, Massachusetts: Sinauer Associates Inc.

DeVries, H., 1900: The Law of Segregation of Characters in Grosses, *Journ.Royal Horticultural Society* 25, 243-48.

Dobzhansky, T., 1932: Cytological Map of the X chromosome of

Drosophila Melanogaster, *Biologisches Zentralblatt* 52, 493-509.
Edwards, A.W.F., 1986: Are Mendel's Results really too close?, *Biological Reviews* 61, 295-312.
Etherington, I.M.H., 1939: Genetic Algebras, *Proc.Royal Soc. Edinburgh* 59, 153-62.
Elandt-Johnson, R.C., 1971: *Probabilistic Models and Statistical Methods in Genetics*, New York: Wiley and Sons.
Garrod, A.E., 1902: The Incidence of Alkaptonuria: A Study in Chemical Individuality, *Lancet* 2, 1616-20.
Garrod, A.E., 1909: *Inborn Errors of Metabolism*, Oxford: UP.
Goodenough, U. and Levine, R.P., 1974: *Genetics*, London-New York-Sydney-Toronto: Holt Rinehart & Winston.
Holgate, P., 1968: The Genetic Algebra of k-linked Loci, *Proceedings London Mathematical Society* 18-3, 315-27.
Hull, D., 1969: What Philosophy of Biology is not, *Journal of the History of Biology* 2,, 241-68.
Hull, D., 1972: Reduction in Genetics -Biology or Philosophy?, *Philosophy of Science* 39, 491-499.
Hull, D., 1974: *Philosophy of Biological Sciences*, Englewood Cliffs, NJ: Prentice-Hall Inc.
Ingram,V.M., 1957: Gene Mutations in Human Haemoglobin: The Chemical Difference between Normal and Sickle-Cell Haemoglobin, *Nature* 180, 326-328.
Ingram, V.M., 1965: *The Biosynthesis of Macromolecules*, New York: W.A.Benjamin.
Jervis, G.A., 1954: *Research Publication of the Association of Research into Nervous Mental Diseases* 33, 719.
Kitcher, P., 1982: Genes, *BJPS* 33, 337-59.
Kitcher, P., 1984: 1953 and All That: A Tale of Two Sciences, *The Philosophical Review* 93: 335-373.
Krantz,D.H., Luce, R.D., Suppes, P. and Tversky,A., 1971: *Foundations of Measurement*, New York-London: Academic Press.
Kuhn, T., 1962: *The Structure of Scientific Revolutions*, Chicago: UP.
Kyburg, H.E.jr., 1968: *Philosophy of Science: A Formal Approach*, New York: Collier-Macmillan.
Langley,P., Simon, H.A., Bradshaw;G.L. and Zytkow, J.M., 1987: *Scientific Discovery*, Cambridge/Mass.: MIT Press.
Lawrence, W.J.C., 1950: Genetic Control of Biochemical Synthesis as Exemplified by Plant Genetics-Flower Colours, *Biochemical Society Symposia* 4, 3-9.
Mackie, J.L., 1974: *The Cement of the Universe*, Oxford: UP.
Mayr, E., 1967: *Artbegriff und Evolution*, Hamburg-Berlin: Parey.
McCusick, V.A., 1969: *Human Genetics*, Englewood Cliffs NJ: Prentice-Hall.Inc.
McCusick, V.A., 1989: Mapping and Sequencing the Human Genome, *New England Journal of Medicine* 320, 910.

Mendel, G., 1901: Experiments in Pea Hybridization, *Journal of the Royal Horticultural Society* 26, 1-32.
Meselson, and Radding, , 1975: A Genetic Model for Recombination, *Proc.Nat.Acad.Sci.USA* 72, 358-61.
Morgan, T.H., 1911: An Attempt to Analyse the Constitution of the Chromosomes on the Basis of Sex-limited Inheritance in Drosophila, *Journal of Experimental Zoology* 11, 365-414.
Morgan, T.H. and Bridges, 1916: Carnegie Institute of Washington 237, 21.
Neel, J.V., 1949: The Inheritance of Sickle Cell Anaemia, *Science* 110, 64-66.
Nickles, T., 1973: Two Concepts of Intertheoretic Reduction, *The Journal of Philosophy* 70, 181-201.
Onslow, M.W. (nee Wheldale) and Bassett, H.L., 1913: The Flower Pigments of Antirrhinum Majus. 2.The Pale Yellow or Ivory Pigment. *Biochemical Journal* 7, 441-4.
Pauling, L.H., M.A.Itano, S.J.Singer and I.C.Wells 1949: Sickle Cell Anaemia, a Molecular Disease, *Science* 110, 543-548.
Pearce, D., 1987: *Roads to Commensurability*, Dordrecht: Reidel.
Rasmussen, J., 1935: Studies on the Inheritance of Quantitative Characters in Pisum I, *Hereditas* 20, 161-80.
Rizzotti, M. and Zanardo, A., 1986: Axiomatization of Genetics I & II, *Journal of Theoretical Biology* 118: 61-71, 145-152.
Rupp, W.D., C.E.Wilde, D.L.Reno and P. Howard Flaunders, 1971: Exchanges between DNA Strands in UV irradiated E.Coli, *Journal of Molecular Biology* 61, 25-44.
Schafer, R.D., 1949: Structure of Genetic Algebras, *Amer.J.Math.* 71, 121-35.
Schaffner, K., 1967: Approaches to Reduction, *Philosophy of Science* 34, 137-47.
Schaffner, K., 1969a: The Watson-Crick Model and Reductionism, *BJPS* 20, 325-48.
Schaffner K., 1969b: Correspondence Rules, *Philosophy of Science* 36, 280-90.
Sinnot, E. W. and Dunn, L. C., 1925: *Principles of Genetics: An Elementary Text, with Problems*, New York: McGraw-Hill.
Sklar, L., 1967: Types of Intertheoretic Reduction, *BJPS* 18, 109-24.
Sneed, J.D., 1971: *The Logical Structure of Mathematical Physics* Dordrecht: Reidel.
Stefik, M., 1981a: Planning with Constraints, *Artificial Intelligence* 16, 111-40.
Stefik, M., 1981b: Planning and Meta-Planning, *Artificial Intelligence* 16, 141-70.
Stegmueller, W., 1976: *The Structure and Dynamics of Theories*, Berlin-Heidelberg-New York: Springer.
Stegmueller, W., 1986: *Theorie und Erfahrung, Dritter Teilband*,

Berlin-Heidelberg-New York: Springer.
Strickberger, M.W., 1985: *Genetics*, (3.ed.), New York-London: Macmillan.
Stryer, L., 1981: *Biochemistry*, New York: Freeman & Co.
Sturtevant, A.H. and Morgan, T.H., 1923: Reverse Mutation of the Bar Gene Correlated with Crossing Over, *Science* 57, 746-747.
Sturtevant, A.H., 1925: The Effects of Unequal Crossing Over at the Bar Locus in Drosophila, *Genetics* 10, 117-147.
Suppes, P., 1970: *A Probabilistic Theory of Causality*, Acta Philosophica Fennica, Amsterdam: North Holland.
Szostak, J.W., T.L.Orr-Weaver and R.J.Rothstein, 1983: The Double Strand Break Repair Model for Recombination, *Cell* 33, 25-44.
Taliaferro, W.H. and Huck, J.G., 1923: *Genetics* 8, 594-98.
Wheldale,M., 1907: The Inheritance of Flower Colour in Antirrhinum Majus, *Proceedings of the Royal Society of London* 79, 288-305.
Woodger, J.H., 1959: Studies in the Foundations of Genetics, in Henkin, L. et al. (eds.), *The Axiomatic Method*, Amsterdam: North Holland, 408-28.
Woerz-Busekros, A., 1980: *Algebras in Genetics* (Lecture Notes in Biomathematics 36), Berlin-Heidelberg-New York: Springer.
Yanofsky, C., B.C.Carlton, J.R.Guest, D.R.Helinski and V.Henning,1964: On the Collinearity of Gene Structure and Protein Structure, *Proc. Nat. Acad. Sci. USA* 51, 266-72.

AUTHORS INDEX

SUBJECT INDEX

LIST OF SYMBOLS