1: Modelling in behavioural ecology

Alan Grafen

1.1 Introduction

The chapter begins with a general justification for using non-genetic models to study adaptation. This is extremely important because the genetics of interesting traits is rarely known. The second section discusses Hamilton's model of social interactions (Hamilton, 1964; 1967), explaining it in a new way. The scope of inclusive fitness theory is discussed, and applications of Hamilton's rule are described. Hamilton's model underlies most current research on social interactions, and shows how a model can become so much part of the framework of our thought that we are unaware of it. A model of signalling forms the basis for the third section, in which the much vilified handicap principle of Zahavi (1975, 1977a) is vindicated. The clarification of Zahavi's ideas is upheld as an example of a cardinal virtue of modelling.

1.2 Population genetics underlies behavioural ecology

The starting point for much behavioural ecology is that animals are maximizers of one sort or another—efficient predators or foragers, or elusive prey. The usual ground for believing this is the presumption that natural selection has made them so. If not now then at some time in the past (Dawkins, 1982, pp. 20–24), there existed heritable variation in hunting and foraging techniques, and in ploys to escape predators. Changes in allele frequencies have made animals good at what they do.

The behavioural ecologist, though, does not usually know the genetics underlying the character she studies. While she would be interested to know this genetic system, it is not of primary importance. Her main aim is to uncover the selective forces that shape the character. The behavioural ecologist has to hope in her ignorance that her method will work almost regardless of which particular genetic system underlies the character (Lloyd, 1977). This hope raises two questions. First, is it justified? Second, is the assumption so powerful and plausible that a whole research strategy should be based on it?

1.2.1 The phenotypic gambit

Let us start with a brief caricature, with examples, of an important method in behavioural ecology. It has two elements.

(i) A strategy set This is a list or set of (perhaps all) possible states of the character of interest. Here are three examples of strategy sets. McGregor *et al.* (1981) studied the song of male great tits, and in particular their repertory size. The strategy set they used was simply every different repertory size they observed: integers from one to five. Brockmann *et al.* (1979) studied the nesting of great golden digger wasps. These wasps sometimes acquire a nest by digging, and sometimes by entering an already existing nest. Brockmann and co-workers were interested in the relative frequency of these two ways of acquiring a nest, and so the strategy set was simply all possible proportions of digging rather than entering—numbers between zero and one. In the hawk–dove game devised by Maynard Smith and Price (1973), the strategy set consists of two strategies, called hawk and dove.

(ii) A rule for determining the success of a strategy The success of a strategy is the number of offspring left by an animal adopting it, or alternatively its inclusive fitness (see section 1.3 below). The rule for determining success may involve the frequencies with which strategies are adopted in the population. One way to determine the rule is to observe it, as McGregor *et al.* They counted how many offspring every male fathered in his lifetime, and averaged across all males sharing the same repertory size. Another way is to model the rule, which Brockmann *et al.* did because they needed to know how the success of the strategies changed as their frequencies changed. They used data to estimate parameters in the rule. When the purpose is to investigate theoretically the consequences of a particular form of frequency dependence, then an appropriate rule is simply assumed: in the hawk–dove game, the rule is represented in the pay-off matrix.

The phenotypic gambit is to examine the evolutionary basis of a character as if the very simplest genetic system controlled it: as if there were a haploid locus at which each distinct strategy was represented by a distinct allele, as if the pay-off rule gave the number of offspring for each allele, and as if enough mutation occurred to allow each strategy the opportunity to invade.

The gambit implies that all strategies occurring in the population are equally successful, and that they are at least as successful as any non-occurring strategy would be if it arose in small numbers. The application of the gambit to a given strategy set and pay-off rule is a powerful way of testing the joint hypothesis that the strategy set and the pay-off function have been correctly identified, and that the gambit is true.

In their first model, Brockmann *et al.* rejected this joint hypothesis when two existing strategies turned out not to be equally successful. They adopted a new strategy set in their second model.

The joint hypothesis might be false because the genetic system underlying the character does not produce the same phenotypic effects as the very simplest genetic system, the one assumed in the gambit. The mere fact that the predictions of equal success of existing strategies and the inferiority of unplayed strategies are rejected does not reveal which element in the joint hypothesis is false. The research strategy implied by the phenotypic gambit is to treat such rejection as evidence that the pay-off function or strategy set is wrong, and not that the genetic system is causing the discrepancy (see also Chapter 4).

1.2.2 *Is it a winning gambit?*

Taken literally, the gambit is usually unjustified: few species studied by behavioural ecologists are haploid. But will the genetic system that does underlie the character produce the same phenotypic effects as the genetic system the gambit assumes?

Two points are important here. First, an example is known in which the gambit would be misleading. In some human populations affected by malaria, there are three distinct phenotypes corresponding to the three possible genotypes at a diploid locus with two alleles (Allison, 1954). One type almost invariably dies before reproducing, of sickle cell anaemia. The other two types differ in their resistance to malaria. The coexistence of these three genotypes with markedly different fitnesses would be very puzzling to a behavioural ecologist applying the phenotypic gambit. The mechanics of Mendelian segregation prevent the whole population from sharing the optimal phenotype, because it is produced by the heterozygous genotype. Here, as undoubtedly elsewhere, it is essential to know the underlying genetics in order to understand the distribution of phenotypes observed in the population.

The second point is that such cases are probably rare. Only certain features of genetic systems, such as over-dominance in the sickle cell case, can sustain dramatic differences in fitness, and these features are not known to be common. Maynard Smith (1982) has analysed how well different genetic systems support the simplification represented by the gambit, and he concludes that by and large they do so very well. The sorts of character studied by behavioural ecologists are likely to be controlled by many loci, and this reduces the scope for the maintenance of large fitness differences.

Genetic systems are themselves subject to evolution. In its simplest form, this is the creation of a new allele by mutation, but more substantial changes could occur. In the sickle cell case, a (functional) gene duplication of the locus would allow one locus to fixate for each allele. Every individual in the population could then have the 'intermediate' genotype that confers malarial protection without the sickle cell anaemia. That this has not happened for sickle cell may be because this intermediate genotype would be disadvantageous where malaria is not a major selective force. The existence of fitness differences between genotypes at equilibrium creates selection for evolution of the genetic system itself.

The behavioural ecologist hopes that genetic systems that do not support the gambit are rare or transient. If the discrepancies produced by genetic systems are smaller than the accuracy of the data, then field-workers can safely ignore them. We know this might not be so, and we should be anxious to find out whether this hope is justified. The soundness of behavioural ecologists' methods depends on arguments concerning population genetics, but our methods are actually designed to avoid doing genetics.

We have seen that the gambit cannot be made with perfect safety. It is a leap of faith. But what would behavioural ecology be like if we refused to use it in our research? It would be very different. Detailed studies in which the precise nature of a character is examined as an adaptation would have to be accompanied by a study in which the genetic mechanism underlying the character was uncovered so precisely that an explicit genetic model could be constructed. There would be no decimal places without genetics. This would reduce drastically the range of characters we can study. Genetically simple and well-studied characters are usually straightforwardly disadvantageous mutants maintained by judicious artificial selection in strains that have spent tens of generations in the laboratory—and so are rarely of evolutionary interest. A behavioural ecological study would have to be very large if genetics were included, and it would be impossible to complete a study on elephants, say, within the lifetime of a scientist.

If the gambit is generally justified, therefore, the genetics is an almost irrelevant complication in understanding the selective forces that shape a character. The gambit makes truly phenotypic explanations possible, and the effort expended in discovering the genetics would be wasted. Better to allocate that effort to studying in an evolutionary ways characters of evolutionary interest, and in a genetic way characters of genetic interest.

These are the reasons why the gambit is so attractive, whether it is justified or not. The advantages seem to me to justify continuing to employ the gambit, always providing we remember that we may be wrong. Theoretical work by Thomas (1985a,b,c,d) and other studies reviewed by Hines (1987) tackle the problem of when the evolutionarily stable strategy approach applied to phenotypes gives the same answer as a more genetical approach would. They lend strong support to the phenotypic gambit.

1.3 Inclusive fitness and Hamilton's rule

It is all very well to say that animals are maximizers, but what do they maximize? Is it number of offspring? This section displays and discusses one of the most important models in modern evolutionary theory, Hamilton's social interactions model (Hamilton, 1964), made more elegant by Hamilton (1970). It is the basis for the powerful principle that animals act as if maximizing their inclusive fitness.

Inclusive fitness is a fundamental concept of evolutionary biology. A widespread misconception is that the point of inclusive fitness is to help us understand interactions between relatives. The model will make clear that the scope of inclusive fitness covers all interactions in which the genotype of one individual affects the fitness of conspecifics. The special role of relatives is a powerful result of the theory, not a restricting assumption. Indeed, the most important case on which inclusive fitness theory sheds light is where interactants are unrelated. Here, they should behave so as to maximize their own number of offspring, and have no regard for the effect of their actions on the number of offspring of the other individual. This will be the case in many interactions, perhaps the majority.

1.3.1 Hamilton's model of social interactions

Inclusive fitness is based on Hamilton's model of social interactions. We begin by illustrating in Fig. 1.1 a simpler, conventional model used in population genetics. Each individual in the population is represented by a square with a rectangular addition. The combined area represents the individual's number of offspring. The idea is that the square is a standard unit of fitness, while the rectangular addition represents the effect of the individual's own genotype on its number of offspring. The rectangle will always be drawn outside the square, but, for convenience of illustration and for generality, let us agree that its area may count positively or negatively, depending on whether the individual has an advantageous or disadvantageous genotype. The sign of the numerical value attached to the rectangle will specify where necessary whether an area counts positively or negatively.

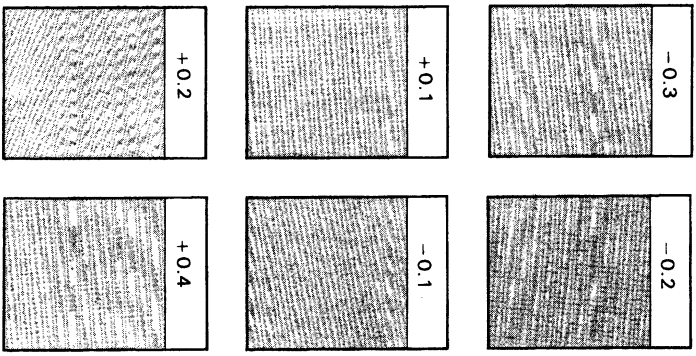

Fig. 1.1 The figure shows in schematic form the numbers of offspring of six individuals. The square represents a standard unit of one offspring, and the rectangular sections represent differences from one caused by the individual's own genotype. These differences can be positive or negative. The six individuals therefore have 1.2, 1.1, 0.7, 1.4, 0.9 and 0.8 offspring.

Consider any locus, say the *G* locus, and focus on one allele *G* and amalgamate all other alleles at the locus under the name *g*. The first use for our model will be to compute the frequency of *G* in the next generation, on the assumption of Mendelian segregation. Each individual will have a genotype at the *G* locus, which for diploids will be either *GG*, *Gg* or *gg*. The number of *G* alleles in offspring is the combined area of all *GG* individuals and half the area of *Gg* individuals. The number of *G* alleles in offspring is the combined area of all *gg* individuals and the remaining half of the area of *Gg* individuals. The frequency of *G* among the offspring is just the number of *G* alleles divided by the sum of the number of *G* and *g* alleles.

This elementary calculation has an important consequence. *G* will increase in frequency if it is associated with larger number of offspring. But we have not specified which allele *G* is, nor which locus it is at, and neither have we made any detailed assumptions about how the fitness effects arise. Therefore, every allele will increase in frequency if it is associated with a larger number of offspring. If every allele at every locus is under selection of this sort, then it is reasonable to say that the organism is under selection to maximize its number of offspring. This kind of statement about selection on the organism transcends the picky details of genetics and justifies the application of 'selection thinking' by organismal biologists.

This crucial organismal conclusion from a genetic model for non-social traits attracted Hamilton to try to devise a parallel model for traits in which one individual's genotype was allowed to affect the fitness of others. The model is altered by introducing two new kinds of areas. An elementary social interaction is represented by a triangle attached to an actor and a circle attached to a recipient. The triangle represents the action's effect on the actor's number of offspring, and the circle represents the effect on the recipient's number of offspring. A dotted line can be drawn between corresponding triangles and circles. It is possible to have more than one recipient, so one triangle may connect to more than one circle. Figure 1.2 therefore represents Hamilton's model of social interactions.

The next step is to notice that number of offspring will do some things in the same way as in the simpler model, but fails to provide the organismal conclusion. The areas of *GG*, *Gg* and *gg* individuals can still be appropriately combined to compute the frequency of *G* among the offspring. (For this purpose, the areas of the triangles and circles are simply added to the individual's number of offspring; all that matters is how many offspring an individual has, not why.) It is still true that an allele *G* will be selected if it is associated with greater number of offspring, in just the same way as before. The trouble is that number of offspring is no longer under the exclusive control of the individual. So we cannot say that organisms are selected to maximize the number of offspring, because one component of their offspring is not under their genetic control, but someone else's. Also, an individual controls the number of offspring other individuals have, and if social behaviour can be selected, then this effect on others must also be

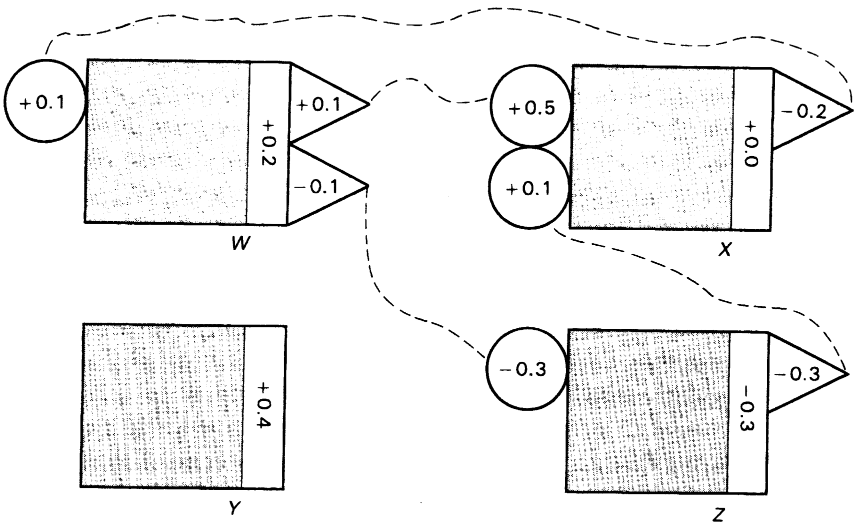

Fig. 1.2 The figure illustrates Hamilton's model of social interactions. In addition to the square and rectangle of Fig 1.1, there are four social actions represented here, each requiring a triangle and a circle. The triangle represents the effect on the actor, while the circle connected by a dashed line represents the effect of the action on the recipient. Each effect can be positive or negative. The letters *W*, *X*, *Y* and *Z* name the individuals. The topmost line, for example, shows *X* paying a cost of 0.2 offspring with a net gain of 0.1 for *W*. In this arrangement, the areas representing offspring are drawn in contact with the parent of the offspring. *W* individual has 1.3 offspring, *X* has 1.4 offspring, *Y* has 1.4, while *Z* has 0.1.

included in any measure that selection might cause organisms to maximize. In other words, we need a measure based on control, not one based on results.

Hamilton produced a different accounting system for computing the frequency of *G* among the offspring. His measure is illustrated in Fig. 1.3. It includes the square (the standard unit), the rectangle (effect of own genotype through non-social traits), the triangles (effect of own genotype through social traits) and the circles linked to those triangles (effect of own genotype on the number of offspring of others). Only a fraction of these circles counts towards the new measure of fitness, and that fraction is the relatedness to the recipient. In recognition of the fact that the effects on others are included, Hamilton called this new measure 'inclusive fitness'. Also important is the fact that it *excludes* the effects of others on the individual's own offspring (the circles that were directly attached to an individual in Fig. 1.2).

Now Hamilton proved an important fact about inclusive fitness. We can calculate the combined inclusive fitness of *GG* individuals, and half of the inclusive fitness of *Gg* individuals, and call this the summed inclusive fitness of *G*. We can calculate similarly the summed inclusive fitness of *g*. If we then divide the summed inclusive fitness of *G* by the total summed inclusive fitnesses of *G* and *g*, we might hope to get a frequency of *G* among the offspring. We do not. But the answer we get has a very important property. It is in the same direction from the frequency of *G* among the parents as the frequency of *G* among the offspring. It therefore gets the direction of gene frequency range right, even if the magnitude is wrong. Thus, a gene will increase in frequency if it is associated with higher inclusive fitness and decrease if it is associated with a lower inclusive fitness.

In return for this weakening so far as computing gene frequencies is concerned, we obtain an enormous strengthening on the organismal side. All the effects of an individual's genotype are now included in its inclusive fitness. *G* increases in frequency if it is associated with an increased inclusive fitness. Remember that *G* can be any allele at any locus. We may conclude that all alleles are selected to increase the inclusive fitness of the individual bearing them, and so we may reach another organismal conclusion that transcends the details of genetics: organisms are under selection to maximize their inclusive fitness. This was Hamilton's goal, and it is a result of fundamental importance in the study of social behaviour.

Two implications can be noted here. First, it is common to specify a

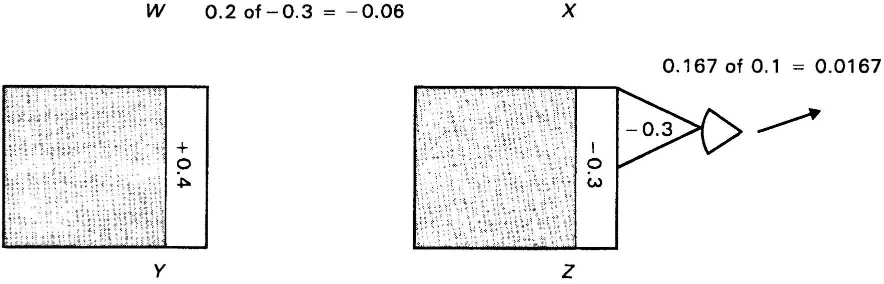

Fig. 1.3 This figure shows the areas of Fig. 1.2 rearranged. An area is now attached to the individual whose genotype caused to exist the offspring represented by the area. Each individual has lost its own circles, but has gained the circles it gave to others. Figure 1.2 showed accounting by results; here we see accounting by control. To calculate inclusive fitness, we need to add up the areas attached to an individual but diminishing the area of a circle. Only a fraction of the circles is counted, and that fraction is the relatedness of the actor to the recipient. Assume that the relatednesses are as follows: *W* to *X*, 0.5; *W* to *Z*, 0.2; *X* to *W*, 0.5; *Z* to *X*, 0.167. The inclusive fitness of *W* is therefore 1 (square, standard unit)+0.2 (rectangle, non-social effect of own genotype)+0.1−0.1 (the effects on his own number of offspring of his social actions)+0.5×0.5−0.3×0.2 (the effects of his social actions on others (0.5, −0.3) weighted by the appropriate relatednesses (0.5, 0.2), coming to 1.39 in total. The other inclusive fitnesses are 0.85 for *X*, 1.4 for *Y* and 0.4167 for *Z*.

social action by saying that it loses the actor one offspring and gains its full sibling three offspring. Inclusive fitness immediately allows us to calculate that this action would be favoured by selection. The loss of one to the actor's inclusive fitness is outweighed by the gain of a half times three. The second implication is the powerful idea of a rate of exchange between own offspring and the offspring of others. An individual acts as if it valued each other individual's offspring as worth a fraction of one of its own—and that fraction is the relatedness between the individuals.

The organismal level of the conclusions reached by inclusive fitness theory is important first of all because it achieves a radical simplification if genetics can be passed over. It is also important because in social traits, even more than in non-social traits, virtually nothing is ever known of the genetics of evolutionarily interesting characters. We observe organisms and interesting morphology and behaviour, but we rarely observe interesting genes. Practical applications therefore require a principle at the level of the organism.

In making use of inclusive fitness, it is often helpful to go back to the model of social interactions on which it is based. The model has been examined in some detail because a full understanding of inclusive fitness is important for students of behavioural ecology. No proof of Hamilton's central result has been given, but in the next section a logically parallel result is proved about what has come to be known as 'Hamilton's rule'.

1.3.2 Relatedness defined and Hamilton's rule deduced

The notion of relatedness was left unexamined in the previous section, with the hope that the reader had some loose notion that identical twins have a relatedness of one, full sibs have a relatedness of one half, and cousins have a relatedness of one eighth. It turns out that relatedness is quite a subtle notion, and it is convenient here to explain the meaning of relatedness at the same time as giving a convincing if informal proof of Hamilton's rule.

The idea of inclusive fitness is that helping relatives is a bit like helping yourself, because they share your genes. This approach will now be made more precise, by finding a definition of relatedness with the 'design requirement' that the relatedness of a potential actor A to the potential recipient R measures the extent to which A helping R is like A helping itself. Relatedness is usually introduced in connection with common ancestry: full sibs share both parents, half sibs share one parent, and cousins share two grandparents. But for the moment we are interested only in measuring genetic similarity, which can be caused by common ancestry but can also be caused by other processes.

There are many senses of genetic similarity, and correspondingly many ways to measure it. For example, we could concentrate on only one locus, and assign a similarity of 1 if two individuals shared both alleles at that locus, 0.5 if they shared only one allele, and 0 if they shared no allele. Or we could ask what fraction of all alleles at all loci are shared between two individuals. But neither of these suggestions satisfies our design requirements.

The measure of genetic similarity that does the trick is illustrated in Fig. 1.4. A line indicates possible gene frequencies, numbers from zero to one. Three points are marked on it, whose relative positions define relatedness. We concentrate on just one allele at one locus. The first point, μ, is the frequency of that allele in the population whose gene frequencies we are tracking. The second point, *A*, is the frequency of that allele in the potential actor, and the third point, *R*, is the frequency of that allele in the potential recipient.

We now trace the consequence for the spread of the special allele of altruism performed by *A* to *R*. In two special cases illustrated in Fig. 1.4, our design requirement specifies the relatedness right away. If *R* and μ coincide, then the recipient is the same as the population average (Fig. 1.4b). When *A* helps *R*, he adds alleles in the existing proportion, and so does not help to change the population gene frequency at all. From the point of view of the allele *G* that controls the action, *A* may as well throw his help away, because that does not change the population gene frequency either. The relatedness should therefore be zero in the case where *R* and μ coincide. On the other hand, when *A* and *R* coincide, helping *R* has just the same consequence for the changing gene frequency as if *A* helped himself to the

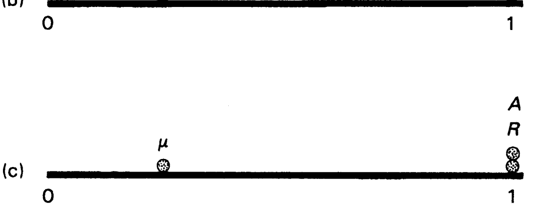

Fig. 1.4 This figure illustrates the meaning of relatedness. The line between zero and one represents the gene frequency of a particular allele at a particular locus. *A* represents the actor's gene frequency, μ represents the population average gene frequency, and *R* represents the average position of the recipient. Relatedness is the fraction of distance from μ to *A* at which *R* lies—in case (a) about one quarter. In case (b), the recipient has the same gene frequency as the population average and so *A* may as well throw his help away on a random recipient because relatedness is zero. In case (c), *R* has the same gene frequency as *A*, and so *A* helping *R* is as good as *A* helping himself. The relatedness in this case is therefore one.

same extent (Fig. 1.4c). When *A* and *R* coincide, therefore, the relatedness must be one.

These two special cases suggest that relatedness be defined more generally as 'the fraction of the distance from μ to *A* at which *R* lies'. This definition corresponds to the regression definition of relatedness first put forward by Hamilton (1970), and which is now gaining popularity in modern theory. Regression relatednesses are not always the same as Sewall Wright's correlation coefficients of relatedness (Wright, 1969) or as Hamilton's life-for-life coefficients (Hamilton, 1972). Table 1.1 shows the life-for-life and regression coefficients for some relationships under haplodiploidy. All three values are the same under out-breeding diploidy with no selection in an infinite population.

The terms *correlation* and *regression* arise because the definitions can be formulated to look exactly like statistical formulae for correlation and regression coefficients (see Hamilton, 1975). If we code a 'score' for each individual in the population as the fraction of alleles at the *G* locus that are *G*, then Wright's correlation coefficient is simply the correlation between the scores of potential altruists and potential recipients. It is symmetric, and can range from -1 to 1. The regression coefficient is the slope in a regression of the potential recipients' scores on the potential altruists'. It can in principle take any value and, as Table 1.1 shows, need not be symmetric.

With the regression definition of relatedness it is now very easy to provide a convincing if informal proof of Hamilton's rule. For the sake of definiteness, suppose *R* is one quarter of the way from μ to *A*, perhaps because *A* and *R* are half-sibs. Imagine that *A* has resources that he can convert into four offspring for *R*. These are shown at *R* in Fig. 1.5. These offspring will have the same effect on the gene frequency as the combination of one offspring at *A* and three offspring at μ. This is so because the groups are the same size and because their gene frequencies are the same. This shows that *A* helping *R* to have four offspring has the same effect as if he were able to have one offspring himself, and produce three with the population gene frequency.

In deciding in what direction the population gene frequency will change, we can ignore offspring at μ and sum up the effects of the other offspring produced by all the individuals in the population. But the offspring at μ are not altogether irrelevant. The more offspring there are at μ, the more the next generation will resemble the current generation. Hence offspring at μ slow down the magnitude of the change without affecting its direction. Hamilton (1964) called this the 'diluting effect'.

We can say that *A*'s creation of four children for *R* has the same effect on the direction of gene frequency change as producing one offspring for itself. This is simply a verbal formulation of Hamilton's rule. It is easy to see that the same argument applies for any relatedness, not just for one quarter.

The regression definition of relatedness therefore makes Hamilton's rule work, whether the gene in question is rare or common. What, though,

Table 1.1 The relatednesses under haplodiploidy between some categories of relatives under two definitions of relatedness: Hamilton's (1972) life-for-life definition as employed by Trivers and Hare (1976) and Hamilton's (1970; 1972) regression definition as used in this chapter.

Sex of donor	Relationship of recipient to donor	Life-for-life	Regression
Female	Mother	0.5	0.5
	Father	0.5	1.0
	Sister	0.75	0.75
	Brother	0.25	0.5
	Daughter	0.5	0.5
	Son	0.5	1.0
Male	Mother	1.0	0.5
	'Father' (mother's mate)	0.0	0.0
	Sister	0.5	0.25
	Brother	0.5	0.5
	Daughter	1.0	0.5
	'Son' (mate's son)	0.0	0.0

Only outbred relationships are considered. Quick methods of calculating these values are (i) life-for-life: the fraction of the donor's genes that are identical by descent with any of the recipient's; (ii) regression: the fraction of the recipient's genes that are identical by descent with any of the donor's. For example, all of a male's genes are identical by descent with genes in his mother, so the life-for-life coefficient is 1. On the other hand, only one half of his mother's genes are identical by descent with genes in him, so the regression coefficient is one half. These methods do not work under inbreeding. Note that within-sex values are the same, and between-sex values are converted by multiplying by the ratio of ploidies of donor to recipient. Neither the life-for-life nor regression coefficient is *symmetric*. The regression relatedness of a son to his mother is one half, and of a mother to her son is one; these values are reversed for life-for-life coefficients.

does it have to do with common ancestry? An argument used by Charnov (1977) tells us. Consider the simple case of diploidy and random mating, with an infinite population at Hardy–Weinberg equilibrium not undergoing selection. Suppose we are concentrating on an allele *M*, and we agree to class all other alleles at the locus under the name *N*. A homozygote has genotype *MM*. What is the genotype of a full sib? The parents must have genotypes *M?* and *M?*, where '*?*' indicates that the allele is unspecified by knowledge of the focal individual's genotype. Because we have assumed random mating, Hardy–Weinberg equilibrium and no selection, the chance that '*?*' will be any particular allele is proportional to its frequency in the population. So '*?*' equals *M* with probability *p*, say, and *N* with

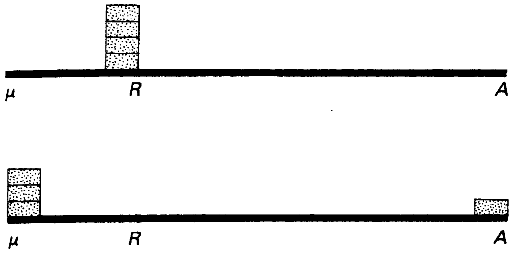

Fig. 1.5 The recipient is shown on the upper line as having four offspring. These are redistributed on the lower line, one to A and three to μ. The two sets of offspring will have the same expected effect on the gene frequency because they contain on average the same number of genes, and the same gene frequency. Any rearrangement that preserves the number of offspring and their 'centre of gravity' would have this property. The special point about this rearrangement is that offspring that contain the population average gene frequency do not contribute to a change in that frequency in the next generation. From the point of view of the direction of the change in gene frequency, therefore, the upper and lower arrangements have identical effects. Thus four children for R have the same effect as one child for A. This is Hamilton's rule. If R is a fraction λ of the distance from μ to A, then one offspring at R has the same centre of gravity as the combination of $1-\lambda$ offspring at μ, and λ offspring at A. Hence, in general, as well as in the special case of one quarter illustrated here, the factor by which A must discount offspring given to R is the fraction of the distance from μ to A at which R lies. Thus that fraction is the desired measure of relatedness.

probability $1-p$. A sib has a one quarter chance of being *MM*, half a chance of being *M?* (one quarter each from one parent having *M* and the other having *?*), and a quarter chance of being *??*. The frequency of *M* on average is therefore

$$\frac{1}{4}\times 1+\frac{1}{2}\times\frac{1}{2}(1+p)+\frac{1}{4}\times p=\frac{1}{2}(1+p)$$

The average gene frequency in a sib is therefore $R=\frac{1}{2}(1+p)$, which is halfway between the gene frequency of the focus individual ($A=1$) and the population average ($\mu=p$). The same result would apply if we had started with a heterozygote ($R=\frac{1}{2}(\frac{1}{2}+p)$; $A=\frac{1}{2}$; $\mu=p$) or the opposite homozygote ($R=\frac{1}{2}(0+p)$; $A=0$; $\mu=p$). The same agreement between common ancestry and the regression relatedness works for all other relationships under the assumptions we have made.

This example makes clear an initially puzzling feature of the definition of relatedness. Under diploidy, every individual's gene frequency is either 0, $\frac{1}{2}$ or 1. Yet R is often assumed to take other values. For example, when μ is $\frac{1}{3}$ and A is 1, a full sib will be taken as having a gene frequency of $\frac{2}{3}$, which it is impossible for any one individual to have. This is because when A makes his decision, he is assumed to have only certain information available to him—in this case, of sibship—and this means that A has to assume an average position for *R*. The average position can be anywhere between 0 and 1, in just the same way as the population average can be.

This Charnovian calculation relies on selection being weak at the locus. To see why, consider an individual of genotype *MM*, whose parents are inferred to be *M?*, *M?*. The interpretation of the *?* as *M* with probability *p*, *N* with probability $1-p$ depends on an assumption of Hardy–Weinberg equilibrium. For example, if *MN* is lethal, then we should infer that *MM*'s parents are both *MM*. Even granted the *M?*, *M?* parents, selection on the offspring of those parents may have altered the *MM* : *MN* : *NN* genotypes among surviving sibs from the 1 : 2 : 1 ratio needed for our computation.

One important point is that common ancestry turns out to cause the same relatedness for all alleles, at all autosomal loci, under our assumptions of random mating and weak selection in an infinite population. But under inbreeding, or strong selection, this definition can give different relatednesses, even for different alleles at the same locus! Causes of genetic similarity other than common ancestry are likely to bring about different relatednesses at different loci, even under random mating. Hamilton's rule still applies to each allele when relatednesses vary, but to each with its own appropriate relatedness. The behaviour of individuals may therefore be selected in different directions by the selection on different alleles. Different relatednesses are therefore likely to cause intragenomic conflict of the kind first considered by Hamilton (1967) in the context of sex ratios.

This possibility has an implication for Hamilton's model of social interactions. When relatednesses are the same for all alleles and loci, then the inclusive fitness of an individual is uniquely defined, and the organismal conclusion that individuals are under selection to maximize their inclusive fitness holds. Once complications arise that cause relatednesses to vary across alleles and loci, this simple picture breaks down.

The usual form of Hamilton's rule incorporates the equivalence illustrated in Fig. 1.5 into a rule about when a gene is selected for, that causes a potential actor to incur a cost, *c*, to itself, while conferring a benefit, *b*, to a potential recipient with relatedness *r*. The direct effect on own offspring is minus *c*, and the equivalent effect from helping *R* is *rb*. Selection favours the allele if $rb-c > 0$, which is Hamilton's rule.

Relatedness can help us to glimpse something of the more complicated world in which relatednesses vary between alleles and loci. For any given social action, for example helping a nest-mate, suppose we know the benefit to a recipient, *b*, and the cost to the actor, *c*. Then the critical relatedness above which the action will be favoured, that is, $rb-c > 0$, is c/b. Now if a large majority of alleles have a relatedness greater then c/b, we can expect the action to be favoured by selection; while if the large majority of relatednesses are less than c/b, then it will be selected out. So in an important class of social actions, the variability in relatedness will not matter. In intermediate cases, the results will be a complicated mess. The conclusion to be stressed is that the organismal level conclusions justified by inclusive fitness theory are threatened only for these intermediate cases. Most social actions will be unequivocally favoured or disfavoured: only a few will bring about genomic conflict.

The main purpose of this section was to provide a demonstration of Hamilton's rule. It was necessary at the same time to give an account of relatedness as a measure of genetic similarity that has peculiar significance for social behaviour, and to discuss the possibility that relatedness might vary between alleles and loci.

1.3.3 How not to measure inclusive fitness

Hamilton's (1964) verbal definition of inclusive fitness is:

> 'the animal's production of adult offspring . . . stripped of all components . . . due to the individual's social environment, leaving the fitness he would express if not exposed to any of the harms or benefits of that environment, . . . and augmented by certain fractions of the harm and benefit the individual himself causes to the fitnesses of his neighbours. The fractions in question are simply the coefficients of relationship . . .'

The picture of social interactions shown in Fig. 1.2 shows immediately that measuring inclusive fitness is a subtle business. It is not enough to measure how many offspring an individual produces, it is also necessary to be able to partition them. How many are the result of actions by others, and how many were the results of actions taken towards others?

There have been in the past various flawed attempts to measure inclusive fitness in nature that have ignored the partitioning of an individual's number of offspring. This leads to an exaggeration of the 'own offspring' component, by inclusion of help from others (the individual's own circles in Fig. 1.2), and it can also lead to a gross distortion of the 'others' offspring' component of inclusive fitness if *all* of the offspring of a relative are included instead of only the additional help the focal individual supplied (i.e. only the other individual's circles in Fig. 1.2 should be included, and only those caused by the focal individual).

Some readers may be now tempted to despair—it is hard enough to measure the number of offspring of an individual in the field, without managing the almost metaphysical task of deciding who was really responsible for them (the 'causal parent'). Part of this feeling is justified. Measuring the inclusive fitness of an individual is indeed a tall order, if it is even possible. Remember that the scheme of Fig. 1.2 has a mainly conceptual purpose. The right way to apply inclusive fitness theory to data is to apply Hamilton's rule, and this has been achieved to good effect with real data. For examples and references see Grafen (1984, section 3.3.3).

1.3.4 *How to use Hamilton's rule*

Hamilton's rule was derived in section 1.3.2. The purpose of applying it is to ask: would a given social action be favoured by selection? Hamilton's rule implies that only three quantities need be known to answer the question: the benefit to the recipient, the cost to the actor, and the relatedness. Under the assumption that the only cause of genetic similarity is recent common ancestry, the regression coefficients of relatedness from Table 1.1 are the right ones to use to make Hamilton's rule work.

There have been various quibbles about the validity of Hamilton's rule based on misunderstandings of the meaning of cost and benefit. The conceptual scheme of Fig. 1.2 can help here. Cost is the area of the triangle, the number of offspring the actor loses through performing the action. One way to estimate this from data would be to consider otherwise similar individuals, one of which performs the act and the other of which does not. Then the difference between their total number of offspring will represent the effect of the action. This method will be used below. More generally it is important to realize that the theory says the cost is a difference in numbers of offspring and is not, for example, a ratio. Inclusive fitness theory is a strong theory, in the sense that it dictates how its terms should be measured. It is not just a casually thrown together collection of symbols.

The same considerations can be applied to the benefit to the recipient, which could therefore be estimated by considering two otherwise identical individuals, one of which had the action performed to it while the other had not. Of course, in any application of this method, serious attention will need to be paid to the phrase 'otherwise identical'.

Having established what all three terms in Hamilton's rule mean, we move on to the first step in applying it in practice. This is to choose the decision we are interested in, being as explicit as possible about the alternative courses of action. Let us call performing the action *Y*, and not performing it *X*. To calculate *b* and *c*, we will need to estimate the difference it makes to lifetime number of offspring to do *Y* rather than *X*. If the actions are only vaguely defined, then we cannot estimate those differences.

A difference of *c* in the animal's lifetime number of offspring results from doing *Y* rather than *X*. Any consequences that would follow from doing *Y* and not *X* should be taken into account—decreased longevity, retribution and so on. It may seem at first sight that a simple way to estimate *c* from data is to take the difference in lifetime number of offspring between animals that do *X* and animals that do *Y*. However, this seemingly reasonable procedure may give the wrong answer. The reason is that animals that do *X* will have relatives who tend to do *X*, and animals that do *Y* will have relatives that tend to do *Y*. The simple difference in number of offspring between *X*-doers and *Y*-doers will therefore include the extra *b*s that a *Y*-doer can expect to receive from his *Y*-performing relatives, and therefore does not give a proper estimate of *c*. Of course, the same caution applies to measuring *b*.

The value of *r* has been assessed in various ways. Bertram (1976) modelled the structure of lion prides to arrive at relatednesses; Brown (1975) used simple ancestry; and Metcalf and Whitt (1977) used electrophoresis. These are all approximations to what *r* must be in principle, which is discussed in section 1.3.2.

Finally, Hamilton's rule in the form $rb-c > 0$ has definite advantages over the more popular form $b/c > 1/r$. For one thing, the second form is wrong if *c* or *r* is negative (though correct if both are). For another, in the common case that *r* is known on *a priori* grounds, then the sampling variance of $rb-c$ is calculable simply from the sampling variances of *b* and *c*, allowing confidence intervals to be constructed. In contrast, the ratio b/c has a sampling variance that depends in a more complicated way on the sampling distributions of *b* and *c*. Examples of applying Hamilton's rule are given in Chapters 10 and 11.

1.3.5 The validity of Hamilton's rule

The scope of inclusive fitness theory depends on the validity of the model of Fig. 1.2. We shall first look at an example of Charlesworth (1978a) and then draw some general conclusions.

In Charlesworth's example, a dominant allele causes its bearer to kill itself and feed itself to its sibs. An application of Hamilton's rule implies that this will be advantageous provided the sib gains more than twice as many offspring from the extra food as the bearer would itself have had. Yet a moment's reflection shows that if all bearers of the dominant allele kill themselves, then the allele will be extinguished immediately, no matter what advantage may be gained by the sibs of suicidal nest-mates. No bearers of the allele will survive to reproduce.

Hamilton's model of social interactions assumes that because the recipient is a sib, it has a relatedness of one half to the actor, and in effect computes the gene frequency of the created offspring on that basis. But because of the nature of its action, the suicidal allele cannot be present in a receiving sib. Hamilton's model fails to represent the social interactions correctly, and so the inclusive fitness principle does not apply.

This example lies at the intersection of three distinct general classes of exception, and can be looked at in three illuminating ways. We can look at Charlesworth's example as a case of multiplicative interactions of fitness, in place of the additive interactions the model requires. The action multiplies the actor's fitness by zero, and the recipients fitness by, say, three. Then multiplying a recipient's fitness by three when he is himself about to multiply it by zero is not a very effective way to help him.

The second general class of exception is where the benefits are genotype specific. In this case, the benefit to the non-suicidal genotypes is much greater than the benefit to suicidal genotypes. Hamilton's rule, on the other hand, assumes that the benefit is not systematically different for different recipient genotypes.

We move on to a third way in which Charlesworth's versatile example can be understood, as the result of strong selection at the controlling loci. The suicidal genotype helps its sibs, but owing to strong selection on the sibs (namely some of them have killed or are about to kill themselves), the relatedness to the recipients is less than common ancestry alone would predict.

What should we conclude from these three kinds of exception? Take, first, non-additivity of fitness interactions. During rapid change in a character there may well be favourable alleles with large effects. But during periods of equilibrium, when the character is close to its optimum, we can expect that evolution is slow, perhaps mainly stabilizing. In these circumstances, the relevant alleles have small effects because alleles of large effect are strongly disadvantageous. For alleles of small effect, fitness interactions will be additive to a sufficient approximation.

Let us see how this might work. The effects of different costs and benefits may not add up. For example, the first donated food item may save an animal's life; the second may enable it to have five offspring; and the third, even though of the same size, may raise the number of offspring only to six. This is a *prima facie* contradiction of the model. However, things look different when we consider an allele of small effect that tinkers with this system. Suppose the first and second donations are always made, and that the variability under selection is the third. Then the model will reflect correctly the existing variation, and so make the correct predictions about selection. Hamilton's rule works even for alleles of large effect if the decision to give the first, second or third donation is distinguished and separately subject to natural selection.

Thus we must admit that circumstances can be imagined in which the inclusive fitness principle breaks down because of non-additive fitness interactions. But if we accept that for most characters the relevant allelic effects are small most of the time, then the inclusive fitness principle will be upheld in the circumstances that matter most to us.

True genotype specificity of the magnitudes of benefits is hard to imagine if the action simply supplies food, or saves from a predator. If the benefit were a blood transfusion between vertebrates, on the other hand, then the benefit would be very different depending on the compatibility between donor and recipient. Accepting this as an exception in principle, it is probably fair to assume that genotype specificity will rarely be a problem in practice.

The last type of exception arose from strong selection, and indeed we noted in section 1.3.2 that the calculations linking relatedness to common ancestry break down then. The assumption of weak selection is therefore necessary to make Hamilton's rule work with ancestral relatednesses. But, as with additivity, even if sometimes strong selection does occur, it is reasonable to suppose that a character is perfected by natural selection under conditions of weak selection. The assumption of weak selection is therefore acceptable.

These three classes of exception are important in principle, particularly if you want to use Hamilton's rule to predict gene frequency changes in models with strong fitness effects, but do not threaten the practical value of Hamilton's rule. That value is to allow us to find organismal accounts of adaptations in social behaviour without doing genetics. If in any case we did need to do genetics, the study would almost certainly be too costly in time and effort to be worthwhile.

A good example of the value of inclusive fitness in modelling is provided by the disagreement between O'Connor (1978) and Stinson (1979) about the evolution of siblicide in birds. O'Connor used a naïve inclusive fitness argument and came to one set of conclusions, while Stinson used a population genetics model and came to another. Godfray and Harper (1990) present a more sophisticated population genetics model than Stinson's, incorporating the assumption of small genetic effects, and obtain O'Connor's 'naïve' results. This gives confidence that inclusive fitness arguments should be treated seriously, even when the actions in question involve large effects, and even when biologically naïve population genetics models cast doubt on them.

1.3.6 Conclusions on Hamilton's model of social interactions

Hamilton's model of social interactions has been outstandingly successful. Hamilton's rule has provided a practical tool with which social behaviour can be studied. The exchange rate concept of valuing others' offspring against one's own according to relatedness is invaluable in thinking about social traits. It is unrivalled in its scope as a model of social behaviour because it captures the biological essentials. At one time, it was a favourite pastime of theoretical biologists to prove Hamilton's rule incorrect, by setting up a model that broke one of the assumptions discussed above (reviewed by Grafen, 1985). Now, however, the value of Hamilton's model of social interactions is widely recognized.

1.4 A model of biological signals

One of the great virtues of evolutionarily stable strategy (ESS) modelling (see Maynard Smith, 1982) is that it helps to clarify muddy verbal arguments. Of the many topics discussed in previous editions of this book (Krebs & Davies, 1978; 1984), one in particular has since benefited from ESS modelling. This section describes the application of the phenotypic gambit of section 1.2.1 to biological signalling. Previous approaches will be discussed once the new ESS model has been described. (Handicap models of signalling are also discussed in section 7.3 and in sections 12.3 and 12.4.)

We will consider a simple imaginary species of beetle. Males vary in how much food they happen to find as larvae. For energetic and nutritional reasons, this affects various aspects of their ability to function as adults. In particular, better-fed males have more viable sperm. Males block the female reproductive tract with a plug after mating, and so females mate only once. Females therefore have a great interest in mating with a male with more viable sperm.

If females could perceive directly the viability of a male's sperm, there would be no signalling problem. Females would 'see' the sperm viability and choose a mate accordingly. Instead, let us make the reasonable assumption that females cannot perceive sperm viability.

Does a male know his own sperm viability? Translated out of metaphor, can there exist developmental rules that are flexibly expressed in male bodies, such that some other character can be made to co-vary with the viability of a male's sperm? We shall assume the answer is yes. This is reasonable because we have assumed that differences in sperm viability are caused by differences in larval nutrition. It is easy to imagine that some structures will be more fully developed in adults that were well fed as larvae than in those that were ill fed.

We have a set of females all wanting to know the sperm viabilities of the males they encounter, and a set of males each of whom knows his own sperm viability. The evolutionary signalling problem is this: is there an evolutionarily stable way in which males can convey the information females would like to have?

This might seem straightforward: the males have the information; the females want it. The problem is that each mate benefits if he is judged to have more viable sperm, and so, considered loosely, all want to signal that they are the best male. If all males do make the same signal, then females cannot distinguish between males at all, and so will not attend to it. At first sight, therefore, stable signalling cannot evolve. But our argument has been rather rough and ready. This is the kind of verbal impasse at which we should turn to ESS modelling. Here we will deliberately exclude mathematical symbols—but they were necessary for rigour in the papers on which the discussion is based (Grafen, 1990a,b).

We begin constructing the ESS model by assuming, for concreteness, that males have horns and that any communication about sperm viability takes place through the size of the male horns (although of course any other trait would have done if females could perceive it). We can now follow the phenotypic gambit of section 1.2.1. What is the strategy set for males? Let us assume that developmental rules exist that can establish any relationship between sperm viability and horn size. This means that, in terms of Fig. 1.6, a male strategy is any rule that specifies for each sperm viability a particular horn size.

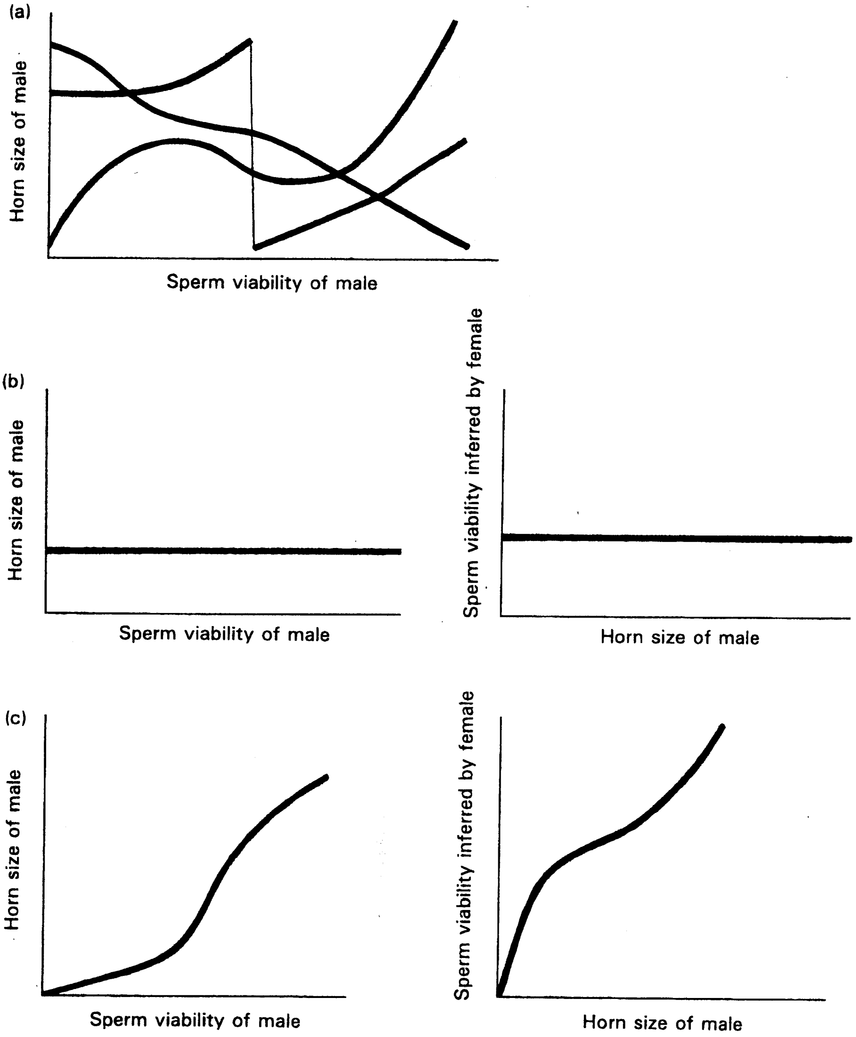

Fig. 1.6 (a) shows three possible rules relating sperm viability to horn size in males. One rule slopes downwards throughout, while another goes upwards overall but with a central downwards section. The third rule has a discontinuity half-way along the horizontal axis. Any rule is allowed in the model. (b) illustrates the non-signalling equilibrium. Males produce the minimum cost horn size, whatever their sperm viability. Females assess males, whatever their horn size, as having average sperm viability. No deviator can gain in this situation. (c) illustrates a signalling equilibrium. Males with greater sperm viabilities produce larger horns. Females assess males with larger horns as having greater sperm viabilities. Each rule is the reverse of the other. Start with any sperm viability, and look up the horn size from the male rule; then look up from the female rule what sperm viability is inferred from that horn size. The inferred sperm viability is the same as the orginal sperm viability, so that females make the correct assessment of sperm viability. Zahavi's handicap theory is developed from the supposition that an equilibrium like this exists. Grafen (1990a,b) shows that in a wide class of signalling models, an equilibrium like this does indeed exist.

Now what about the female strategies? A female has to infer from a male's horn size what his sperm viability is. Thus we can represent a female strategy as a rule specifying for each horn size an inferred sperm viability.

With the strategy sets in place, the next step is to consider the pay-off functions for males and females. For present purposes, we do not need to be precise. Assume that the size of a male's horns affects his fitness in some way, by using energy and so reducing survival from larval to adult stages, or by increasing the risk of predation as an adult. How the cost relates to horn size and sperm viability will turn out to be of the highest importance, but for the moment we leave it unspecified. So far as the signalling consequences of their horns are concerned, it will be enough to specify that a male is fitter, the higher his sperm viability is inferred to be. A female is assumed to be fitter the more accurately she infers males' sperm viabilities.

With these rather vague assumptions, we cannot prove that an equilibrium exists, which is anyway a rather technical affair. But we can ask: suppose there was an equilibrium, what would it be like?

The simplest equilibrium is the non-signalling one, illustrated in Fig. 1.6b. A male produces the same, cheapest, horn size irrespective of sperm viability. Females, with no information to go on, assess males as having the average sperm viability. Crucially, in the figure, the female rule assesses any rare mutant males as having the same average sperm viability.

It is an equilibrium because no rare mutant male or female can spread. A male that produces any other horn size will still be assessed as having average sperm viability, and so will gain no mating advantage. But other males are producing the cheapest horn size, and so the mutant must lose out in that respect, and so lose out overall. (Notice that a situation in which all males produce a horn size that is not the cheapest could not be an equilibrium.)

A female whose rule differs from the common rule in its assessment of the cheapest horn size will suffer, because she will think all males are better than they are, and so be too ready to mate; or think they are worse than they really are, and be too reluctant. We can imagine the female taking risks to mate, or mating at a suboptimal time because she assesses an encountered male as above average. Female mutants cannot gain either.

This discussion of the non-signalling equilibrium has shown the kinds of conclusion we can draw from the supposition of equilibrium. What can be inferred about a signalling equilibrium in which horn size increases with sperm viability?

We may immediately conclude what the female rule must be. As each horn size is produced only by males of a particular sperm viability, the best female rule is to draw the correct inferences and assess a given horn size as the sperm viability of the males that actually produce it. The equilibrium female rule must therefore be just the reverse of the equilibrium male rule, as illustrated in Fig. 1.6c. This is the first conclusion we reach on the supposition of equilibrium.

Now we turn to the males. Why does a male not have a uniformly high horn size, at the level played by the best males, no matter what his sperm viability? This would certainly be more advantageous from the point of view of obtaining matings. If the rule illustrated in Fig. 1.6c really is an equilibrium, then the other costs of having a higher horn size must outweigh the mating advantage. It must therefore be more costly in fitness to have larger horns. Perhaps the horns make the male clumsy, rendering him at a disadvantage in fighting, or less able to avoid predators. The theory cannot say what the disadvantage is, but it does say that there must be one. Otherwise there can be no signalling equilibrium.

There is an additional conclusion that we can draw about the cost of horns. The differential cost of having larger horns must be greater for worse males. This follows because the assumption of equilibrium tells us that each quality of male produces its own optimal horn size. We have assumed that horn size alone signals quality. A male of any quality will therefore gain the same mating advantage from the same horn size. If good males and poor males have different optimal horn sizes, then it must be because the other relevant component of male fitness, the cost of the horns, is different for good and poor males. And if good males have a larger optimal horn size than poor males, it must be because the cost of having larger horns is less for good males.

These arguments are all based on the assumption that a signalling equilibrium exists, and are therefore worthless if such equilibria are impossible. In fact, I have shown (Grafen, 1990a,b) that signalling equilibria do indeed exist in a wide range of models. Unfortunately, the arguments are too technical to give here, and we shall simply assume that signalling equilibria exist. From the supposition of a signalling equilibrium, we have therefore concluded that: (i) females make the correct deductions about male sperm viability from horn size; (ii) larger horns are costlier to males than smaller horns; and (iii) the differential cost of larger horns is greater for worse males than for good males.

These conclusions are the same as those of Zahavi (1975; 1977a). Zahavi claimed that signals were honest, our conclusion (i), that signals must be costly (ii), and that they were more costly for worse males (iii). We have just followed through, therefore, a vindication and clarification of Zahavi's handicap theory.

The discussion was made in terms of horns and sperm viability, to make it easier to follow. But the arguments hold more generally and are not restricted to mate choice. For example, consider a fight in red deer between a harem-holder and a challenger. Harem-holders eat very little and so lose strength. We can view the remaining strength of the harem-master as quality, known to the harem-master and unknown but of great interest to the challenger. In the early stages of a fight, we can expect signals to convey information from the harem-master to the challenger about his strength. We could reach the analogous conclusions about these signals, assuming them to be in evolutionary equilibrium, as we did about the hypothetical insects' horn size. The signal will be correctly interpreted by the challenger; the signal must be costly; and the signal must be more costly to harem-masters with lower remaining strength. A signal requiring strength or stamina might well satisfy the second two conditions. There are complications in assessing the cost of a signal in cases of this sort (Grafen, 1990a), but the principle is just the same.

1.4.1 A comparison with previous approaches

How does this ESS model advance our understanding of signalling compared with previous approaches? The first comparison to be made must be made with Zahavi's work (Zahavi, 1975; 1977a; 1978). Here the advance is one of clarification. The ESS model takes Zahavi's ideas and, by placing them in a more formal context, makes explicit the assumptions and arguments. The conclusions remain the same.

The second comparison is with the work of Dawkins and Krebs (1978) and Krebs and Dawkins (1984). Krebs and Dawkins provide a general view of signalling that is very congenial, and in effect forms the background for the model given above. Their view about how signals involved in assessment can be stable is that the signals must be 'reliable indicators of RHP' (resource holding potentional), which must be 'too costly to fake'. Zahavi's handicap principle is one way of making signals reliable, by making them too costly to be worth faking. The ESS models of Grafen (1990a,b) are fairly general models of signalling, and lend support to the view that the handicap principle may be the only way of making signals reliable. (It should be noted that this strong conclusion relies on what is meant by the term 'signal', which is also discussed in those papers.)

Krebs and Dawkins correctly identify selective forces. Signallers do wish to mislead receivers. Receivers do not wish to be tricked by signallers. Each party is concerned only about its own fitness. This inherent conflict of interest leads Krebs and Dawkins to expect continuing change and spiralling strategies. Pre-ESS thinking leads to similar expectations in other games (well discussed for the hawk–dove game, for example, by Dawkins (1989)). An ESS approach leads to meaningful and interesting conditions that emerge at equilibrium. The model of the previous section complies with all the requirements of Krebs and Dawkins' discussion, but when we insist on asking about what must be true in an equilibrium, we find the emergence of Zahavi's handicap principle. I believe that had Zahavi's principle been properly understood at the time, it would have taken a central part in Krebs and Dawkins' discussion.

Let us take apparently conflicting conclusions of the two approaches. At equilibrium, there is honesty. Why do males not lie as recommended by Krebs and Dawkins? One answer is that if they did lie, we could not be at equilibrium: females would have learned not to believe the signal. This just pushes the question one stage further back: how can it not be advantageous for males to lie? Because the signal they would have to make in order to lie would cost more than the mating benefit they would receive. If signals cannot be costly, then there can be no equilibrium. But the extravagance of signals is one of the facts that all the theories under consideration set out to explain, and the ESS approach brings out the vital conclusion about honesty at equilibrium.

There are other differences of emphasis between the approaches. The ESS model allows females not to use the information from signalling if it is unreliable, and takes this reluctance into account in studying the evolution of signals. Krebs and Dawkins assume that the females' only defence is to require even stronger signals. Simply ignoring signals is an option that would tend to defuse arms races. The signals must continue to convey important information if females are to be selected to continue to attend to them. What drives the ESS model is always the variability between males that matters to females, and signals are selected only according to relevant correlations. The pure arms race, sales resistance phenomenon that Dawkins and Krebs particularly stressed relies on the next difference we note between the approaches.

This lies in how signals acquire meaning. In the ESS model, a signal acquires a meaning only when receivers learn empirically (and possibly only over evolutionary time) that a certain signal is given by signallers of a certain quality. The signals have no inherent meaning. Female response to a 2 cm horn is determined only by previous experience of males with 2 cm horns. Krebs and Dawkins sometimes seem to assume that signals already possess a meaning before they are used. For example, if a man asserts that he is entirely free of parasites, this has meaning to a female even if she (or her ancestors) have never encountered such an assertion before, and even if there has never been a parasite-free man. This difference is in a way quite fundamental. Humans can continue to lie in some situations because their statements are given meaning by the context of language. It seems likely that peacocks are less likely to have meanings pre-assigned to different tail lengths or styles. Where signals are inherently meaningless, that is, they acquire meaning only empirically through the experiences of receivers, this too is likely to limit the scope for spiralling arms races. This difference may arise because Krebs and Dawkins focus on signals about intentions, while the model above concerns signals of quality.

The handicap principle does not assert that any possible signalling system is honest, costly, and has greater costs for poorer quality signallers. It merely asserts that any stable signalling system must have these properties. What happens to a signalling system without them? The system may cycle endlessly, with booms and busts, in a way that resembles an arms race much of the time, leading to a world like that envisaged by Krebs and Dawkins in which 'in actor–reactor coevolution both sides may gain the upper hand'. On the other hand it is also possible, and in view of earlier paragraphs I think more likely, that such a system would collapse. Females would be selected not to attend to the males' signals. The outcome could be settled for any hypothesized signalling system only with a model.

The model of the previous section is not intended as a template for every example of signalling, but to illustrate that the handicap principle can work. Even in our hypothetical example, we can see one omitted complication. We assumed that females gain information about male quality only through horn size. But if quality is dictated by success at larval feeding, then good males are likely to be simply bigger than worse males. Good and bad males will differ in many ways, and not just in horn size. Females will have many possible traits to detect that may correlate with quality. This would complicate our model of the handicap principle, as it would any model of female choice. The idea that quality shows through in many ways is the basis of the argument in the second paragraph of Zahavi (1975), that when real mate choice is going on, Fisher's runaway process is unlikely to occur. He argued that females probably use multiple cues to detect important variation in male quality, so if any one trait gets out of step it is likely to be ignored (see Chapters 7 and 12).

This discussion is based on models of Grafen (1990a,b) as these claim to provide a full and explicit justification of Zahavi's handicap principle. Previous relevant work on signals includes the first model of biological signals, by Enquist (1985), who refuted the still popular notion that animals cannot be selected to signal their intentions; a graphical exposition of the handicap principle by Nur and Hasson (1984) very close in spirit to Grafen's model; and Andersson's (1986) model of conditional handicaps.

1.4.2 Conclusion

The ESS signalling model clarifies Zahavi's original arguments and makes explicit a number of assumptions about how signals operate. The requirement of equilibrium made in the ESS model turns the deceit recommended by Dawkins and Krebs into the honesty of Zahavi. We can expect similar transmutations, counter-intuitive at first hearing but in reality quite reasonable, to arise from the application of ESS models in other areas, particularly other topics in signalling.

The model has various artificial restrictions, made to simplify the argument. The chief restriction is that the variation in male quality is assumed to be purely environmental. Greenough and Grafen (in preparation) present a model in which male quality is genetic, and study it by computer simulation. They show that advertising and preference work in the same way as when male quality is environmental. Other simplifications are discussed by Grafen (1990a,b).

The ESS signalling model described here is fairly complicated. A strategy is a function, loosely any curve drawn from left to right on the plane, allowing vertical jumps. The strategy set therefore has an infinite number of dimensions, a rather daunting fact that requires careful technique. Further, there are two sets of players, males and females, with different pay-off structures. These complications may help account for the delay between Zahavi's verbal proposal and an analytical justification.

I do not see these complexities as marking the boundary of the usefulness of ESS theory. Quite the reverse. The long arguments over the handicap principle show that verbal reasoning is even more adversely affected than analytical reasoning in complicated cases. If we are to understand communication, and other sophisticated games played by organisms, then we will need to become familiar with this kind of ESS model, grapple with its complexities and welcome its illuminations.

In broader terms, the modelling exercise shows that it is necessary to be wary of failures to model an idea. The handicap principle has been confidently refuted many times by modellers who did not understand it. The model described here shows that, as Zahavi maintains, the handicap principle is about the strategy of communication. It applies to human communication between governments and to interspecific communication in just the same way as it applies to sexual selection. It has therefore nothing to do with genetics. Zahavi (1987) goes so far as to claim that the distinction Darwin drew between natural and sexual selection is properly understood as a distinction between the selection of ordinary traits on the one hand and the selection of signalling traits on the other. Ordinary traits are selected for efficiency, while an essential part of the selection of signalling traits is that they are wasteful, the waste being the self-inflicted costs of the signallers. The model of the handicap principle offered above implies that Zahavi's far-reaching claim deserves serious attention.