Hamilton's rule, inclusive fitness, and the evolution of altruism

Anand Natarajan
Center for Theoretical Physics, Massachusetts Institute of Technology

(Dated: May 18, 2018)

We study the evolution of cooperative and altruistic traits, starting with Hamilton's rule and the theory of inclusive fitness. The key insight of this theory is that genes encoding cooperative behaviors can be favored if they cause an individual to help others that share copies of the same gene. We derive a general version of Hamilton's rule study the dynamics of a simple example based on the prisoner's dilemma.

INTRODUCTION

According to basic natural selection, populations will evolve to maximize the reproductive fitness of individuals in the population. However, in the real world, it is common to see organisms behave in altruistic ways, benefiting other organisms at some personal cost. For instance, in many species of animals, upon sighting a predator, one member of a herd or flock may make an alarm call, a behavior which benefits the group but increases the chance of the actor drawing the attention of the predator. As a more extreme example, several species of insects in the order Hymenoptera exhibit a type of highly cooperative behavior called *eusociality*, often including the presence of "worker" or "soldier" castes that do not reproduce and sacrifice themselves for the good of the colony.

How can such behaviors arise under natural selection? The received answer to this question, developed largely by W. D. Hamilton [3], is that genes promoting cooperative behaviors can be favored if they result in an organism helping closely related organisms, which are likely to share copies of the same gene. This insight was pithily expressed by J. B. S. Haldane, who is reported to have said that he "was prepared to lay down his life for eight cousins or two brothers" [6]. A more formal version of this idea is given by *Hamilton's rule*, which states that an allele promoting altruistic behavior will be favored by natural selection if

$$r > \frac{c}{b}, \tag{1}$$

where r is the average relatedness between the actor and the recipient of the altruistic behavior, b is the benefit gained by the recipient, and c is the cost to the bearer of the altruistic allele.

PRICE'S EQUATION AND HAMILTON'S RULE

Of course, it is impossible to interpret Hamilton's rule without a definition of the quantities r, b, and c, and different authors have used various definitions, resulting in different versions of the rule. The two main variants, as elucidated by Birch [1], are the general Hamilton's rule (HRG), which is a mathematical consequence of natural selection that can be proven to *always*, and the specialized Hamilton's rule (HRS), which only holds for a special class of systems but for which the coefficients b and c have a natural causal interpretation. In passing, we note that, as pointed out by Birch, much controversy surrounding a recent article by Nowak et al. [4] attacking Hamilton's rule is due to a conflation between these two versions, with the detractors attacking HRS and the defenders arguing for HRG. In this paper we will show how to derive HRG in the infinite population limit, following the approach of Queller [5]. We will then consider special cases where the coefficients in HRG have an intuitive interpretation.

To start, we need to introduce Price's equation, which describes how average values of traits change over time in a large population under selection. Here, by a "trait," we refer to any characteristic of individuals which we can represent using a real number. The most general form of this equation handles both mutation and selection, but in our setting we will ignore mutation, and assume that reproduction is perfect. To derive the equation, following [2] we will assume the following model. Imagine that we have a parental population, consisting of n individuals, with each individual $i \in \{1, \dots n\}$ having a trait value of z_i and a fitness value of w_i. This population has an average trait value of

$$\overline{z}_{\text{parent}} = \frac{1}{n} \sum_i z_i. \tag{2}$$

We now suppose that reproduction occurs, with each individual producing offspring at a rate proportional to its fitness. Then after reproduction, the new average trait value will be

$$\overline{z}_{\text{child}} = \frac{1}{n} \sum_i z_i \frac{w_i}{\overline{w}}, \tag{3}$$

where $\overline{w} = \frac{1}{n} \sum_i w_i$ is the mean fitness of the parental population. (We have assumed that the population size n is very large, so that statistical fluctuations in reproduction can be ignored.) We see that the change in the average trait value is given by

$$\Delta \overline{z} = \frac{1}{n}\sum_i z_i \left(\frac{w_i}{\overline{w}} - 1\right) \tag{4}$$

$$= \frac{1}{n}\sum_i z_i \left(\frac{w_i - \overline{w}}{\overline{w}}\right) \tag{5}$$

$$= \frac{\mathrm{E}_i[z_i w_i] - \mathrm{E}_i[z_i]\,\mathrm{E}_i[w_i]}{\overline{w}} \tag{6}$$

$$= \frac{1}{\overline{w}}\operatorname{cov}(z, w), \tag{7}$$

where the notation $\mathrm{E}_i[\cdot]$ denotes the expectation over i, and the quantity cov is the covariance between the trait value z and the fitness value w over the population. Equation (7) is Price's equation in the case of no mutations. This equation encompasses Fisher's fundamental theorem, which is obtained when the trait z_i is taken to be the fitness w_i itself.

For the purpose of deriving Hamilton's rule, we would like the trait value z to reflect the presence of the altruistic gene. For instance, suppose we are considering a single locus with two possible alleles, where one allele codes for cooperative behavior and the other does not. In this case, we could take z_i to be equal to the *breeding value*: fraction of an individuals alleles at a particular locus that are the cooperative variant. With this definition, the population average of z tells us the fraction of the gene pool that consists of the cooperative allele. We would like to use Equation (7) to determine whether a cooperative allele will be favored by natural selection, and in order to do this, we must evaluate the covariance $\operatorname{cov}(z, w)$. In general, the fitness w_i of an individual i will be a function of all of its genetic traits, as well as the genetic traits of any individual j that interacts with i. Let us denote the average of z_j over all j interacting with i by z'_i. Then generically, we can write

$$w_i = \alpha + \beta_{w,z|z'} z + \beta_{w,z'|z} z' + \epsilon_i, \tag{8}$$

where $\alpha, \beta_{w,z|z'}$, and $\beta_{w,z'|z}$ are as yet arbitrary coefficients that are independent of i, and ϵ_i is a residual term that is allowed to depend on i. So far, this decomposition is completely arbitrary; however, if we choose that α, β such that the sum of squared residuals $\sum_i \epsilon_i^2$ over the population is minimized, it turns out that $\operatorname{cov}(z, \epsilon) = \operatorname{cov}(z', \epsilon) = 0$. This implies that the covariance $\operatorname{cov}(w, z)$ can be written as

$$\operatorname{cov}(w, z) = \beta_{w,z|z'} \operatorname{cov}(z, z) + \beta_{w,z'|z} \operatorname{cov}(z, z'). \tag{9}$$

Following Hamilton, we define the *relatedness coefficient*

$$r = \frac{\operatorname{cov}(z, z')}{\operatorname{var}(z)}. \tag{10}$$

Substituting (9) into (7) and using the definition (10) of r, we obtain

$$\Delta \overline{z} = \frac{1}{\overline{w}}(\beta_{w,z|z'} + \beta_{w,z'|z} r)\operatorname{var}(z). \tag{11}$$

As $\overline{w}$ and $\operatorname{var}(z)$ are both positive, Equation 11 implies that a trait will increase in frequency if

$$r \geq -\frac{\beta_{w,z|z'}}{\beta_{w,z'|z}} \tag{12}$$

Equation (12) is the generalized Hamilton's rule (HRG), where we interpret r as the relatedness coefficient and identify $b = \beta_{w,z'|z}$ as the benefit granted by the trait to others and $c = -\beta_{w,z|z'}$ as the cost incurred by the individual. But do these identifications make sense? We will explore this with some simple examples.

Relatedness coefficient First, let us start with r. We can explore several limits of this parameter. In one limit, if each individual interacts with members of the population chosen uniformly at random, then z and z' are independent and $\operatorname{cov}(z, z') = 0$, and hence $r = 0$. In another limit, suppose that we are dealing with k-ploid individuals, and the behavior coded by the trait under consideration results only in interactions between individuals and their siblings. Let us write the random variable z_i as a sum $\frac{1}{k}\sum_k y_{i,k}$ where each $y_{i,k}$ is an indicator variable for the presence of the cooperative allele in the kth copy of the gene. Likewise, we may write the population average as $z = \frac{1}{k}\sum_k y_k$ where $y_k = \mathrm{E}_i[y_{i,k}]$. If the cooperative allele is present in the gene pool with a frequency p, then we can evaluate the covariance:

$$\operatorname{cov}(z, z') = \frac{1}{k^2}\sum_{k,\ell} \operatorname{cov}(y_k, y'_\ell) \tag{13}$$

$$= \frac{1}{k}\operatorname{cov}(y_1, y'_1) \tag{14}$$

$$= \frac{1}{k}\mathop{\mathrm{E}}_{i}\left[y_{i,1}\left(\mathop{\mathrm{E}}_{j\in\text{siblings}(i)}[y_{j,1}]\right)\right] - y_1^2 \tag{15}$$

$$= \frac{1}{k}p\left(\frac{1}{k} + \left(1 - \frac{1}{k}\right)p\right) - p^2 \tag{16}$$

$$= \frac{1}{k^2}p(1-p) = \frac{1}{k}\operatorname{var}(z). \tag{17}$$

Hence, the relatedness coefficient r is $1/k$. A similar calculation can be performed for cousins, etc., justifying the quip of Haldane cited in the introduction.

Cost and benefit We will now consider the interpretation of the regression coefficients in (12). To do so, let us suppose that each individual's fitness consists of a baseline value 1 plus a term arising from interactions, where these are modeled by a simple game of the type of the prisoner's dilemma. In an interaction, each organism has a choice between two strategies, labeled "cooperate" and "defect," with a payoff matrix as shown in Table II. We continue to consider a diploid population, and make the further assumption that the fraction z_i of an individual's alleles that are of the cooperative variant is equal to the probability that the individual executes the cooperative strategy. Thus, an individual with $z_i = 0$ always defects, and one with $z_i = 1/2$ will cooperate with probability

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	$b-c$	$-c$
	Defect	b	0

TABLE I. Payoff matrix for a two-player prisoner's dilemma. The entries are the payoffs for player 1; the payoff matrix for player 2 is assumed to be identical. The parameter b is the benefit to being a recipient of a cooperative behavior, and c is the cost to performing such a behavior.

1/2 and defect with probability 1/2. With this assumption, the fitness w_i of individual i can be written exactly as

$$w_i = 1 - cz_i + bz_i', \tag{18}$$

and thus in this case the regression coefficient $\beta_{w,z|z'} = -c$ and $\beta_{w,z'|z} = b$ have the interpretations we claimed.

The payoff matrix in Table II possesses a special structure, known as *equal gains from switching* [7]: the sum of the two diagonal entries is equal to the sum of the two off-diagonal entries. When this fails to hold, the regression coefficients are harder to interpret.

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	$b-c+d$	$-c$
	Defect	b	0

TABLE II. Payoff matrix for a two-player prisoner's dilemma with synergy. The entries are the payoffs for player 1; the payoff matrix for player 2 is assumed to be identical. The parameters b and c are as in Table II, and d is an additional synergistic benefit gained when both players cooperate.

One example given in [7] is the payoff matrix for a prisoner's dilemma with synergy in Table II. In this case, the fitness of an individual i is

$$w_i = 1 - cz_i + bz_i' + dz_iz_i'. \tag{19}$$

We will see that the regression coefficients will not be simply b and c in this case. To simply the calculation, let us switch to a haploid model, so that $z_i^2 = z_i$. We need to minimize the sum of the squares of the residuals.

$$\begin{aligned}
\epsilon_i &= (1-\alpha) - (c + \beta_{w,z|z'})z_i \\
&\qquad + (b - \beta_{w,z'|z})z_i' + dz_iz_i' && (20) \\
\frac{\partial}{\partial\alpha} \operatorname*{E}_i \epsilon_i^2 &= -\operatorname*{E}_i \epsilon_i && (21) \\
&= 0 && (22) \\
\frac{\partial}{\partial\beta_{w,z|z'}} \operatorname*{E}_i \epsilon_i^2 &= \operatorname*{E}_i \epsilon_i \cdot (-z_i) && (23) \\
&= (-(1-\alpha) + c + \beta_{w,z|z'}) \operatorname*{E}_i[z_i] \\
&\qquad - (b - \beta_{w,z'|z} + d) \operatorname*{E}_i[z_iz_i'] && (24) \\
&= 0 && (25) \\
\frac{\partial}{\partial\beta_{w,z'|z}} \operatorname*{E}_i \epsilon_i^2 &= \operatorname*{E}_i \epsilon_i \cdot (-z_i') && (26) \\
&= (-(1-\alpha) - b + \beta_{w,z'|z}) \operatorname*{E}_i[z_i'] \\
&\qquad + (c + \beta_{w,z|z'} - d) \operatorname*{E}_i[z_iz_i'] && (27) \\
&= 0 && (28)
\end{aligned}$$

Solving this system of equations yields

$$\begin{aligned}
\beta_{w,z|z'} &= -c + \frac{d(rz(1-z) + z^2)}{(r+1)z} && (29) \\
\beta_{w,z'|z} &= b + \frac{d(rz(1-z) + z^2)}{(r+1)z}. && (30)
\end{aligned}$$

So we see that here the regression coefficients have a correction that depends on z.

DYNAMICS

Price's equation has its limitations. First, it assumes that reproduction is deterministic and as such implicitly assumes that the population size is very large. Second, following [7], we note that this equation cannot give us the full evolutionary dynamics even in the infinite population limit as it does not enable us to compute the evolution of $\operatorname{cov}(z, w)$ with time.

Thus, in order to describe the dynamics, we will have to specify further information. Here, we will consider a diploid population undergoing rounds of a Fisher-Wright process followed by Hardy-Weinberg random mating. The probability of survival in the Fisher-Wright process is taken to be proportional to the fitness function in the prisoner's dilemma model of Table II, with the further stipulation that interactions are with siblings, so that $r = 1/2$ at all times. In this case, we can calculate the mean fitness $\overline{w}$ as a function of the population average z of the breeding value. To do so, introduce variables $y_{i,1}$ and $y_{i,2}$ describing the two copies of the gene as in the derivation of (17). We compute that

$$\Pr_{j\in\text{siblings}(i)}[y_{j,\ell}=1|y_{i,k}=1]=\frac{1}{2}(1+z) \tag{31}$$

$$\Pr_{j\in\text{siblings}(i)}[y_{j,\ell}=1|y_{i,k}=0]=\frac{1}{2}z. \tag{32}$$

Using this, we see that if an individual i is homozygous cooperative, then $z_i'=\frac{1}{2}(1+z)$; if it is heterozygous, then $z_i'=\frac{1}{4}+\frac{1}{2}z$, and if it is homozygous defect, then $z_i'=\frac{1}{2}z$. Using these results, we can now write the mean fitness of the population:

$$\begin{aligned}\overline{w}(z) &= z^2\cdot\left(1-c+\frac{1}{2}(1+z)b\right)\\ &\quad + 2z(1-z)\cdot\left(1-\frac{c}{2}+\left(\frac{1}{4}+\frac{1}{2}z\right)b\right)\\ &\quad + (1-z)^2\cdot\left(1+\frac{1}{2}zb\right) \qquad (33)\\ &= 1-zc+zb. \qquad (34)\end{aligned}$$

Substituting this into Hamilton's rule (11), we have the dynamics

$$\Delta z=\left(\frac{b}{2}-c\right)\cdot\frac{z(1-z)}{2}\cdot\frac{1}{1-zc+zb}. \tag{35}$$

This no longer has the form $\Delta z\propto\frac{d}{dz}\ln\overline{w}$ that we found in class in the case of no interaction, suggesting that the dynamics do not optimize the average fitness function. Indeed, this is easy to see for the prisoner's dilemma: suppose that $c<b<2c$. Then, the optimum average fitness would be achieved if all individuals cooperated, yet the dynamics will drive towards the all defecting state as no individual has an incentive to switch to cooperating.

For finite population sizes n, using the evolution equation (35) and in addition now assuming forward and reverse mutation rates μ_1,μ_2, we can write a drift-diffusion equation describing the evolution of the average z, in analogy with the analysis in Lecture 4 of the course notes. The velocity and drift parameters will be given by

$$v(z)=\mu_1(1-z)+\mu_2 z+\left(\frac{b}{2}-c\right)\frac{z(1-z)}{2(1+z(b-c))} \tag{36}$$

$$D(z)=\frac{z(1-z)}{4n} \tag{37}$$

To calculate the steady state distribution $p^*(z)$, we write

$$\begin{aligned}D(z)p^*(z) &= \int^z dz'\,\frac{v(z')}{D(z')} \qquad (38)\\ &= 4n\int^z dz'\,\Big[\frac{\mu_1}{z'}-\frac{\mu_2}{1-z'}\\ &\quad +(b/2-c)\frac{1}{2(1+z(b-c))}\Big] \qquad (39)\\ &= 4n\Big[\mu_1\ln z+\mu_2\ln(1-z)\\ &\quad +\frac{b/2-c}{2(b-c)}\ln(1+z(b-c))\Big]\\ &\quad +\text{constant}. \qquad (40)\end{aligned}$$

This yields a stationary distribution

$$\begin{aligned}p^*(z)\propto{}&\frac{1}{z(1-z)}z^{4n\mu_1}(1-z)^{4n\mu_2}\\ &\overline{w}(z)^{2n(b/2-c)/(b-c)}. \qquad (41)\end{aligned}$$

There are three realistic cases to consider. If $b<c$, then defection is always favored and the fittest states win. If $b>2c$, the cooperation is favored, and likewise the fittest states win. If $c<b<2c$, then defection is favored even though cooperation would lead to a globally better outcome: fitter states are *disfavored.*

CONCLUSION

In this paper, we derived a general form of Hamilton's rule, and used it to study the dynamics and stationary state of a simple system. In general, one could imagine much more complicated interactions than those considered here. In particular, in our example we assumed that the relatedness r was a constant, whereas in reality it will evolve with time just like the other parameters in the model. A more realistic model would perhaps be to consider organisms that undergo spatially local mating and interaction processes. Of course, more general models are likely to no longer be analytically tractable and will require computer simulations.

ACKNOWLEDGEMENTS

Thanks to Prof. Kardar and Prof. Mirny for teaching the course and providing an excellent set of lecture notes.

[1] J. Birch. Hamilton's rule and its discontents. *The British Journal for the Philosophy of Science*, 65(2):381–411, 2013.

[2] A. Gardner. The Price equation. *Current Biology*, 18(5):R198–R202, 2008.

[3] W. D. Hamilton. The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1):1–16, 1964.

[4] M. A. Nowak, C. E. Tarnita, and E. O. Wilson. The evolution of eusociality. *Nature*, 466(7310):1057, 2010.

[5] D. C. Queller. A general model for kin selection. *Evolution*, 46(2):376–380, 1992.

[6] J. M. Smith. Survival through suicide (review of "Sociobiology—the new synthesis" by E. O. Wilson). *New Scientist*, 67(964):496–497, 1975.

[7] A. Traulsen. Mathematics of kin-and group-selection: Formally equivalent? *Evolution*, 64(2):316–323, 2010.