Note on the FGM, data and inferences

1 Introduction

This project is focused on the prediction of fitness of AMR mutants across environments and genotypic backgrounds and the evolution of compensatory mutations. Under this main objective, three main goals have been depicted:

1. Can we use estimates of costs of resistance obtained in one environment to reliably predict costs in a different environment?
2. Are estimates of costs of resistance measured in one genetic background sufficient to predict costs on a different genetic background?
3. Is the route to compensatory evolution similar between environments and genetic backgrounds ?

Before any discussion on these three goals, we need to define some of these notions, which may have multiple definitions in the litterature.

First, the "cost [of resistance] is defined as the selection coefficient of resistance mutations in absence of treatments (or similarly in absence of predator, parasite or pathogens when considering resistance in the context of biotic interactions)" (Lenormand et al. 2018). This cost, as discussed in Lenormand et al. (2018), may be a blurred notion due to the infinite numbers of costs which can be defined through the infinity of different "non-treated" environments. Here, we will focus on the broader idea of the change in fitness effects of a given mutation across environments, with a particular treated environment as a reference.

Second, linked to the previous one, is the notion of environment, which we choose to define here by its measurments. Thus, two different environments can be distinguished by a measurable difference in at least one abiotic and/or biotic variable independent of the genotypes we are focusing on. This means that an environment cannot be defined by the traits (including fitness) of the genotypes in the population(s) that we are studying.

Third, linked to the notions of costs and environments are the estimates of fitness. Multiple estimates (or proxies) of fitness exists, one of the most commonly used is the growth rate, which is defined both at the level of a genotype or at the level of a population. These estimates must be traits highly correlated with the true fitness itself and cannot be defined by a property of the environment. Indeed, the MIC (Minimum Inhibitory Concentration) is sometimes used as a proxy of fitness for the measurements of the capacity to resist to a certain antibiotic, but this estimate is also defining a particular treated environment, which thus cannot be used across environments.

Finally, the notion of compensatory mutations is defined here for a particular scenario. We consider a genotype with a fitness X in environment 1. This genotype faces a new stressing environment 2, adapt to it and has a fitness V in this environment. Then this new genotype (adapted to environment 2) faces environment 1 and as a fitness $Y < X$. In this case, mutations are qualified as compensatory if they increase Y in environment 1 while keeping the fitness in environment 2 equal or higher than V.

With all these notions defined, the method chosen to achieve the 3 main objectives can be separated in two main points:

- Obtain measurements of fitness of multiples genotypes (with a set of known AMR mutations) in multiple environments and obtain replicated time series data of evolution of compensatory mutations.
- Use the fitness landscape theory and more particularly Fisher's Geometric Model (FGM) to predict fitness changes across environments, across genetic backgrounds and through compensatory mutations.

The two methods has to be developed conjointly to be able to use as many information as possible from the data into the model. In the following are detailed the main achievable results depending of the available data and the theoretical assumptions used in the model. The first section depicts a simple version of the FGM, its assumptions and its parametrization from experimental data using these constraints of "simplicity.

The second part shows how this simple form of the FGM can be fitted on data of mutants and used to extrapolate (predict) fitness effects of mutations across environments and genotypes. This section also adresses the perspective of the inclusion of these"static" measurements and predictions of fitnesses into models of eco-evolutionary dynamics. The third part discusses how alternative models (still in the FGM framework), with less restrictive assumptions could be fitted and used for predictions with the appropriate dataset.

2 The simple version of the FGM:

2.1 What is the Fisher's geometric model ?

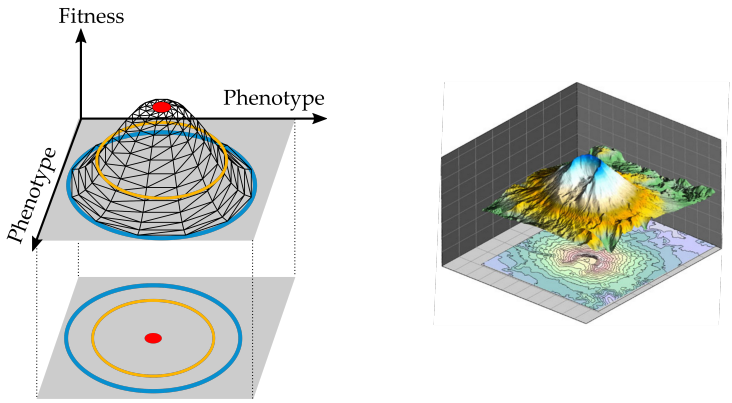

Figure 1: The topography metaphor of fitness landscapes. (Left) FGM with two dimensions: Black lines correspond to the surface of the fitness function and colored lines to iso-fitness lines. (Right) Topographic map.

The FGM is a model of phenotype to fitness landscape, which links the potential phenotypic space of a given organism to the fitness through a "fitness function". This function gives the fitness of any phenotype in the phenotypic space, as a topographic map would give the altitude for any pair of coordinates in the physical space. However, contrary to topographic maps, the FGM is not limited by a two-dimensionnal space and may have a number $n \in \mathbb{N}^+ - \{0\}$ of dimensions (often called complexity) depending on the organism considered. Moreover, its fitness function is dependent on the environment, meaning that for the same phenotype (associated with a genotype) the fitness may vary between environments. Therefore, the FGM is an interesting framework for the modelisation of fitness across environments.

The simplest form of the FGM (whose assumptions are detailed below) can be fully parametrized for a given environment, by the three parameters : n, λ, r_{max}.

- n is the number of uncorrelated phenotypic dimensions of the phenotypic space, often called "complexity".
- λ is the mutational variance per trait (standardized by the strength of selection).
- r_{max} is the growth rate of the optimal phenotype in the environment.

An illustration of a one-dimensionnal FGM ($n = 1$) with two genotypes is given in Figure 2A.

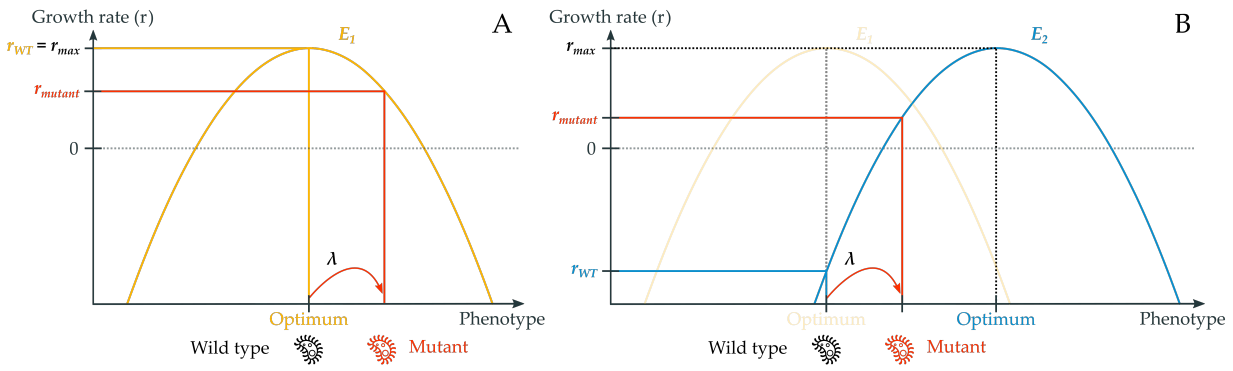

Figure 2: Illustration of two genotypes (wild-type and mutant) in two different environments (E_1 and E_2). (A) The wild type is at the phenotypic optimum and therefore has the maximal growth rate reachable in this environment (r_{max}). λ is proportionnal to the mean mutationnal effect of a mutation. (B) The wild type is no longer at the optimum in the new environment (E_2) due to the shift in the optimum position. Here λ, r_{max} and n are constant across environment.

This simple version of the FGM assumes :

- Isotropy: all traits are equivalent with respect to selection or mutation.
- Constant dimensionality n across environments.
- Constant strength of selection and mutational variance (summarized by λ) across environments.
- Constant height of the fitness peaks across environments.
- No phenotypic plasticity.
- Universal pleiotropy (no modularity or partial pleiotropy).

Assuming constant n, λ and r_{max} across environments constrain the change of environment to be modelled in the FGM by a shift of the position of the optimum (see Figure 2B). The limits of theses assumptions are discussed in section IV.

2.2 Parametrization:

Under the model detailed above, we need to estimate the three parameters n, λ and r_{max} for a given environment and estimate the distance between the optima of the different environments. (i) To parametrize this simple version, the easiest way is to use random mutations stemmed from a genotype at the optimum (for an example see Martin & Lenormand 2006a). A genotype at the optimum can be obtained from a population at mutation selection balance which is characterized by a negligible rate of adaptation. Note that for organisms with a very large mutation rates, the genotype sampled may not be at the optimum, even if the population is at mutation-selection balance, due to the mutation load. A mutation accumulation experiment started from this optimal genotype can then give a cloud of single step random mutants. From the mean $\mathbb{E}(s)$ and the variance $V(s)$ of the selection coefficients obtained from these mutants, $n = 2\mathbb{E}(s)^2/V(s)$ and $\lambda = -V(s)/\mathbb{E}(s)$ can be directly determined (Eq.(4), Martin and Lenormand 2006a).
Moreover, the measure of the growth rate of the optimal phenotype directly gives the fitness at the optimum r_{max}.

(ii) It is also possible to parametrize the model using mutants stemmed from a non-optimal genotype, as would give a screen at a given antibiotic dose. In this case it is necessary to fit the full distribution of fitness effects of mutations (not just the first two moments as in the previous case) to get an estimates of n and λ. This can be done by fitting the distribution of fitness effects given in Martin and Lenormand (2015) Eq.(3) using maximum likelihood (for an example see Harmand et al. 2017). This also gives the distance of the wild type genotype to the optimum but do not allow to estimate r_{max}.

Note that engineered mutants can also be used as in Hietpas et al. (2013), but these mutations are only available for a given gene, which would certainly violate the hypothesis of isotropy.

In the case where the selection coefficients of a cloud of random mutants is available from both an optimal (case (i)) and a non-optimal genotype (case (ii)), we can estimate the three parameters n, λ and r_{max}, plus the distance to the optimum of the non-optimal genotype. This method allows to check that n and λ are constant across the genotypes in the given environment and can be used to map the position of environments in a landscape as detailed in the next section.

3 Predictions using the simple version of the FGM

The aim of the parametrization of a fitness landscape is to position genotypes in the landscape and use these positions to extrapolate the fitnesses, DFEs, and probabilities of resistance/compensatory mutations emergence that these genotypes would show in a new environment.

3.1 Mapping of the genotypes and the environments in the landscape:

In Figure 2B the phenotypic landscape as a single dimensions and two environments are illustrated. From a data point of view, E_1 would be parametrized using the method (i) and E_2 the method (ii) (even if in this simple version of the FGM the distance between E_2 and E_1 would be sufficient, as n, λ and r_{max} remain constant across environments). Knowing the position of these two optima, we can map the position of any genotypes in this landscape by mesuring its fitness in each environment. Then, if we consider a new environment E_{new}, for which we only know the distance to E_1 and E_2, we can predict its fitness and its DFE in E_{new}. This simple example is valid only for a one-dimensionnal lanscape but can be extended to landscapes with higher dimensions and in the following we consider the case $n = 2$ as an example.

3.1.1 The environments of reference:

In a landscape with two dimensions, in order to geometrically map genotypes, we need to have 3 points of reference ($n + 1$ in a landscape with n dimensions). These points of reference will be used for trilateration of any new point added in the landscape. The best option is to consider one genotype adapted per environment as reference points. The relative coordinates in the phenotypic space of these points can be obtained from the measure of the fitness of each optimal phenotype in each of the 3 environments (the fitness measured need to be converted in phenotypic distance through the inverse of the fitness function). These fitness also gives the maximal growth rate (r_{max}) for each environment. Thus, these 3 points define a plane (for $n = 2$) in which we assume that every other environments or genotypes will lie.

3.1.2 Trilateration of the genotypes' positions:

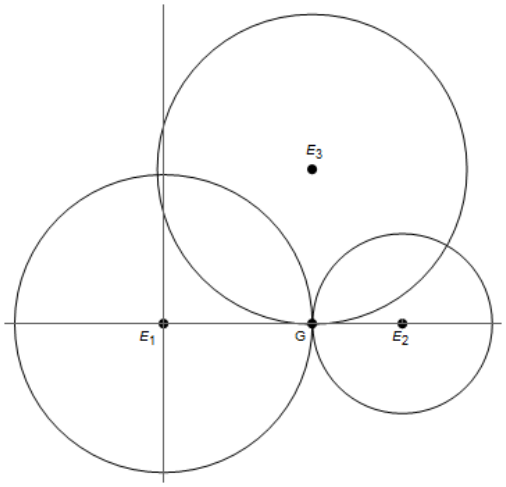

Figure 3: Example of trilateration

Using the three points of reference defined above, the coordinates of any genotype can be obtained by measuring its fitness in the 3 environments, which gives the phenotypic distance to each optimum. These distances can then be used to derive the coordinates of the genotype using trilateration as shown in Figure 3. Note that equations of trilaterations can be solve numerically for any $n \in \mathbb{N}^+ - \{0\}$, considering with have $n+1$ points of reference.

With a collection of phenotype mapped in the landscape, it is then possible to predict the effect of mutations across environments and genotypes. To do so, mutants emerged from one given genotype among the previously mapped ones, have to be mapped in the same methods as previously. From these data, it is possible to assess the fitness of one of these mutations in any genotype and/or any environment (previously mapped). It is also possible to predict epistasis between them. Thus it is possible to predict the cost of mutations across environments and genotypes without directly measuring it, which correspond to goals (1) and (2).

3.2 Probability of emergence of a resistant mutant:

Once a genotype is mapped into the landscape, if we know the distance of this genotype to a new optimum corresponding to a new antibiotic, it is possible to predict the probability of emergence of a resistant mutant using the model of evolutionary rescue from Anciaux et al. (2018). Indeed, this method use the geometric property of the FGM coupled with a stochastic eco-evolutionary dynamic to predict the probability of emergence of at least one resistant mutant in a population facing a new stressing environment. The parameters needed are the initial population size, the parameters from the landscape (n, λ, r_{max} and the fitness of the genotype in the new environment), and the mutation rate. If the landscape has been parametrized from mutation accumulation, the mutation rate may also be available from the data.

3.3 Probability of evolving a compensatory mutation:

The FGM allows to predict the proportion of mutations which are beneficial in multiple environment (see Martin & Lenormand 2015). However, the FGM does not include a genotype to phenotype map. Thus, it cannot predict the probability of emergence of a given mutation, but can predict the probability of emergence of mutations of a given effect s. Therefore the ability of the FGM to predict the emergence of compensatory mutations depends on the definition of compensation.

- If defined as "the adaptation to a second environment while keeping the same mutations of resistance adapted to the first environment", it is not possible to give a prediction on the probability of such events.
- However, if defined as the adaptation to a second environment while keeping the same (or higher) fitness in the first environment, the FGM may give some predictions. Indeed in this definition the focus is not on a particular genotypes but on fitnesses in both environment as defined in the introduction section.

Using the geometry of the landscape, it is possible to give the probability that a mutant will have the n-uplet of growth rates $(r_1, r_2, ..., r_n)$ in the n environments $E_1, E_2, ..., E_n$ (has been done for two environements in Martin & Lenormand 2015 and can be approximated for K environments but only for $n \gg 1$). Combined to an evolutionary model, it is possible to predict the probability of fixation of such a mutation.

These results could be combined with the probability of emergence of a resistant mutant to assess the joint probability of evolving a resistance for a certain antibiotic which is costly in an antibiotic-free environment and then evolve a compensatory mutation, which correspond to goal (3).

These results and especially those on compensation are highly dependent on the assumptions of the model. The next section is an attempt at relaxing some of these assumptions.

4 Alternative FGM

From the model described in the previous section, some assumptions are less "realistic" (i.e. validated by empirical results) than others and could greatly influence the "validity" of the results (i.e. ability of accurately

predict the fitness of mutations across environments and genotypes). However, relaxing these assumptions imply to increase the number of parameters and thus the number/type of data needed.

4.1 Constant strength of selection and mutational variance across environments:

Considering variations of in the mutational variance across environments could potentally be introduced in the model, as long as the model stay isotropic, but requires to fit distribtion of fitness effects in each environments. However, varying the strength of selection would be more difficult as the transformation of the phenotypic space used to scale the mutationnal variance by the strenght of selection would also vary across environments, leading to a hange in the geometric frame of the landscape between environments.

4.2 Constant peak height across environments:

Relaxing this hypothesis is straightforward as the fitness of the optimal genotypes in each environment should be measured in any case (however, this implies to measure absolute fitnesses, i.e. growth rates).

4.3 Constant / fixed dimensionality across environments:

First, the change of dimensionality across environments could be implemented but seems a priori less pertinent biologically, as the complexity" is more an attribute of the organism and not of the environment. The dimensionality estimated by a fit of a distribution of fitness effects may vary across environments, however this may reflect the variation of effective dimensions across environments rather than "true complexity" (as discussed in Lourenço et al. 2011). Thus, considering partial pleiotropy would capture these variation while being more intuitive biologically, as discussed further below. Second, considering higher dimensionality than 2 is possible and common in the FGM. However, this would increase linearly the number of reference points needed for positioning genotypes in the space and may be not be analytically tractable in high dimensions (numerical optimization of "the distance geometry problem"). One way to overcome this problem would be a dimensionality reduction by projecting the points in higher dimensions in the hyperspace of selected dimensionality. This could be achieved by multidimensional scaling methods (MDS).

4.4 Isotropy and universal pleiotropy:

Considering anisotropy is the most difficult assumption to relax, has the geometry of the landscape could be greatly affected. Anisotropy, "can be introduced by mutational or selective correlations, and/or heterogeneity in mutational variance or strength of stabilizing selection across traits. These effects may in addition differ between environments, for example if different combinations of traits are favored in different environments ("syndromes") or if mutational variances and covariances change with the environment ("phenotypic plasticity")." (from Martin & Lenormand 2015).

First, anisotropy can reflect a certain directionality in the mutation process in the phenotypic space. This means that mutation or selection have a stronger effect in certain directions of the phenotypic space. Thus, the "orientation" and the "magnitude" of selection and mutation can change across environments which can be a way of modelling parallel evolution. This form of anisotropy can be introduced in the FGM with analytical tractability for one environment but has only been done by simulations for multiple environments. It seems that mild anisotropy could be absorbed by the isotropic model even in a case with multiple optima but for biological scenario different from the ones we are interested in (genotypes maladapted to every optimum or well adapted to one).

Second, if we relax the assumption of universal pleiotropy, mutations may affect only a certain subset of the phenotypic traits. These subsets can be random (partial/restricted pleiotropy) or form independent clusters (modular pleiotropy). This form of anisotropy is not exclusive with the first one and can be a way to model the parallel evolution often observed in AMR evolution. Different models taking into account anisotropy and partial pleiotropy have been derived but none of them consider the case of multiple optima. In terms of parametrization, such models would require much more parameters, all the more so that the number of

dimensions is large. Time series data of replicated evolution toward the different optima would be necessary to assess the level of anisotropy or partial pleiotropy.