Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

LETTER

Pathogen genetic variation in small-world host contact structures

Paulo R A Campos[1] **and Isabel Gordo**[2]

[1] Departamento de Física, Universidade Federal Rural de Pernambuco, 52171-900, Dois Irmãos, Recife-PE, Brazil
[2] Instituto Gulbenkian de Ciência, P-2781-901, Oeiras, Portugal
E-mail: paulo.campos@df.ufrpe.br and igordo@igc.gulbenkian.pt

Received 23 October 2006
Accepted 6 December 2006
Published 22 December 2006

Online at stacks.iop.org/JSTAT/2006/L12003
doi:10.1088/1742-5468/2006/12/L12003

Abstract. We introduce a model for assessing the levels and patterns of genetic diversity in pathogen populations, whose epidemiology follows a susceptible–infected–susceptible model. We assume a population which is structured into many small subpopulations (hosts) that exchange migrants (transmission) between their neighbours. We consider that the hosts are connected according to a small-world network topology, and in this way our model interpolates between two classical population genetics models: the stepping-stone and the island model. We have observed that the level of diversity has a maximum at intermediate values of the basic reproductive number R_0. This result is independent of the topology considered, but depends on the relation between parasite load and the rate at which the immune system clears the pathogen. We show that, for a given R_0 of the pathogen, as the host contact structure changes, by increasing the rewiring probability p, the level of pathogen diversity decreases. Its level is higher in regular lattices and smaller in random graphs. The latter topology presents a similar diversity level to the island model (a fully connected network), but also presents a clear pattern of isolation by distance, which is observed in some pathogen populations.

Keywords: models for evolution (theory), mutational and evolutionary processes (theory), population dynamics (theory)

1742-5468/06/L12003+8$30.00

The study of metapopulations has attracted researchers from very distinct fields, from ecology and epidemiology to population genetics. A recent comparison between different metapopulation and epidemiological models has revealed enormous similarities and a basic conclusion can be drawn: they essentially differ in the amount of detail that is included [1]. From the epidemiological side, the focus is to model populations of microbes that can cause important human diseases such as malaria and flu. In the context of population genetics, one of the critical questions is to understand the evolutionary forces responsible for patterns of genetic variation in natural populations. A growing interest of physics in this field has been noticed [4, 5]. Importantly, the consideration of both genetics and epidemiology in modelling infectious diseases, has been increasingly recognized [3].

Understanding genetic variation in pathogens is of extreme importance since it can help in targeting genes under selection pressure created by the immune system [6], and under certain circumstances, it can be used to infer host population history [7]. It also reflects the population evolutionary potential [8]. The standard genetic models of population structure make very simple topological assumptions. The most well studied are Wright's island model [2] and Kimura's stepping-stone model [18]. In these models, a large population (metapopulation) is assumed to be composed of many small subpopulations (named demes). When applying these models to study pathogen variation, each deme corresponds to a host. In the island model, it is assumed that each host interacts with all other hosts in the population. On the other hand, in the stepping-stone model the interactions between hosts are only local, which means that a host only interacts with its nearest neighbours.

Neither the island nor the stepping-stone models, that assume regular networks, are accurate descriptions of the interaction networks in real populations. Particularly, in an epidemiological context we expect that the interactions resemble those of social networks [10]. With the great development in the data processing capability, several recent investigations have probed the topologies of many real systems [11]. In fact, the findings of these studies have shown that real topologies are far from being regular, like the stepping-stone model or the island model, or even completely random, such as random graphs. Some models have been proposed to capture the essential features of actual networks, such as the small-world networks [12] and scale-free networks [13], which are commonly referred to as complex networks. The main motivation of the small-world networks relies on the observation of the small-world effect especially in social networks [14].

The formulation of new models of population structure has provoked a growing interest by the scientific community for understanding the mutual influence between the underlying topology and the intrinsic dynamics of systems. In this context, the study of disease outbreaks have received special attention. For instance, Pastor-Satorras *et al* [15, 16] have shown from the study of the classical susceptible–infected–susceptible (SIS) epidemiological model that scale-free networks are more prone to spreading of diseases than random graphs and regular lattices. As in any standard epidemiological formulation, these models do not concern the role of microbe evolution in the fate of epidemic spreading. Recently, we have demonstrated that this is an essential element to be considered since topology can greatly affect the rate of fixation of deleterious mutations and therefore the evolutionary potential of the populations [17].

In this letter, we propose population genetics models of structured populations, that incorporate epidemiological parameters explicitly, in order to study pathogen genetic

J. Stat. Mech. (2006) L12003

variability under the classical SIS model. This model is one of the simplest classical models in epidemiology and it is commonly used in studies of sexually transmitted diseases [20]. Basically, we aim at investigating how the levels and patterns of sequence variation in pathogens look under this model, and also how the host contact structure influences pathogen diversity.

Because the classical island model fails to produce any correlation between geographical and genetic distances, and as this is observed in several natural pathogen populations (see for example [9]), in our simulations we assume small-world like topologies. This also has the advantage that we can tune the amount of long-range interactions so that our results cover all ranges of topologies from the classical stepping-stone model in population genetics, which corresponds to $p = 0$, to a completely random graph ($p = 1$). In our model, we will explore the relation between genetic and geographical distance by considering a commonly used measure of genetic differentiation (F_{ST}, see below) and a measure of topological distance (the shortest path length). We also compare the results with those of the island model.

The model we will study here is as follows. We consider a metapopulation of hosts (demes). Each host can carry at most N_d pathogens, where each pathogen is represented by an infinitely large sequence $\mathbf{S} = (s_1, s_2, \ldots, s_\infty)$ and the nucleotides s_α can take two distinct values $s_\alpha = 0$ (original state) and $s_\alpha = 1$ (which means that a mutation has occurred). Each node in the network corresponds to a given host, and the total number of hosts is D. Therefore, the maximum number of pathogens in the metapopulation is $N_t = DN_d$. Each host can clear the pathogens it carries with probability e at each generation. When this occurs it becomes empty. However, since we also assume migration, empty hosts can be recolonized. This latter step makes our model distinct from previous metapopulation models where migration and recolonization are distinct processes [19]. At each generation, the number of emigrants n_{e} of a given deme j (which is not empty) is taken from a Poisson distribution of mean $N_d m k_j$, i.e.,

$$P(n_{\mathrm{e}}) = \frac{\mathrm{e}^{-N_d m k_j}(N_d m k_j)^{n_{\mathrm{e}}}}{n_{\mathrm{e}}!} \tag{1}$$

where m is the migration rate and k_j is the connectivity of deme j. The n_{e} pathogens are then sampled at random from deme j and distributed to the k_j recipient demes. This assumption can serve as an approximation to what can occur, for example, in the case of malaria, where the pathogen is transmitted through a mosquito that bites a given host. After transmission, the pathogen grows and can evolve (mutate) inside the host. Each pathogen acquires new mutations according to a Poisson distribution of mean μ. Because all pathogens contribute equally to the next generation, then by random sampling, N_d individuals are chosen to form the new pathogen population of each host. Thus, after each cycle, each infected host has exactly N_d pathogens.

In order to perform the statistical measurements of diversity and differentiation amongst hosts that are infected in the population, every T generations we take a sample of size $n_t = 50$ sequences from the entire population, and samples of size $n_d = 5$ sequences from within each host. The average number of pairwise differences between sequences for the entire population, π_t, or for within subpopulations, π_d, is obtained through

$$\pi_{t(d)} = \frac{\sum_{i<j} \pi_{ij}}{n_t(n_t - 1)} \tag{2}$$

J. Stat. Mech. (2006) L12003

where π_{ij} is the number of differences between two sampled sequences. To study the dependence of genetic differentiation between demes with their topological distance, we introduce a modified estimate of F_{ST}. For each pair of hosts, we estimate the level of diversity for the sample of size $n_{ij} = 10$ sequences, where five sequences are taken from each host. Then, the average number of pairwise differences between sequences for this subpopulation is measured, which we denote π^d_{ij}. The estimate $F_{\mathrm{ST}}(d_{ij})$, which is dependent on the distance d_{ij}, between hosts i and j is further determined by

$$F_{\mathrm{ST}}(d_{ij}) = \frac{1}{2}\left[\frac{\pi^d_{ij} - \pi_i}{\pi^d_{ij}} + \frac{\pi^d_{ij} - \pi_j}{\pi^d_{ij}}\right]. \tag{3}$$

Now we turn to the correspondence between the aforementioned model and the susceptible–infected–susceptible model (SIS). In the SIS model, the individuals can be in one of two states: susceptible or infected. A susceptible individual can become infected when in contact with infected individuals. The transmission occurs at rate β. On the other hand, infected individuals return to the susceptible class at rate α. In the deterministic limit and not considering any level of structuring, the evolution of the system is described by the following set of equations:

$$\frac{\mathrm{d}S}{\mathrm{d}T} = -\beta SI + \alpha I \tag{4}$$

$$\frac{\mathrm{d}I}{\mathrm{d}T} = \beta SI - \alpha I \tag{5}$$

where S is the fraction of susceptible individuals and I corresponds to the fraction of infected individuals. Nevertheless, by measuring the time in units of duration of infection $t = T/\alpha$ and recalling $S + I = 1$, the time evolution of the population is obtained through

$$\frac{\mathrm{d}I}{\mathrm{d}t} = R_0 I(1 - I) - I. \tag{6}$$

The parameter $R_0 = \beta/\alpha$ is the relevant parameter of the model and it is known as the basic reproductive number. The basic reproductive number of an infection is the mean number of secondary cases a typical single infected individual can infect in a completely susceptible population. The solutions of equation (6) are $I = 0$ and $I = 1 - 1/R_0$. The solution $I = 0$ is only stable when $R_0 < 1$, while the solution $I = 1 - 1/R_0$ is stable for $R_0 > 1$.

As we consider that a host is depicted as a deme in our model, an empty deme corresponds to a host which is in the susceptible state, whereas a deme which is full corresponds to an infected host. A deme that is currently full can become empty with probability e, which means that e corresponds to α. A deme that is currently empty can become full through the migrants it receives from nearby demes. This implies that β is proportional to m. Given that the average connectivity of a deme is K and that the number of migrants per link is $N_d m$, then β corresponds to $N_d mK$. In analogy with the deterministic description of the SIS model, the relevant parameter R_0 equals $R_0 = N_d mK/e$ and so the average frequency of infected individuals in the metapopulation can be approximated by:

$$I = 1 - e/N_d mK. \tag{7}$$

J. Stat. Mech. (2006) L12003

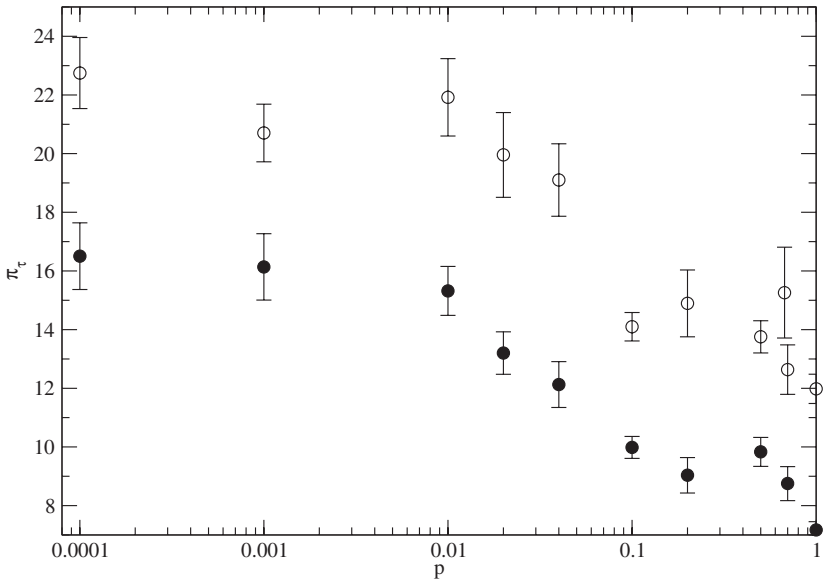

Figure 1. The level of diversity π_t as a function of p. The parameters are $D = 900$, $N_d = 10$, $e = 0.01$, $n_t = 50$ and $\mu = 0.0002$. The open symbols represent $R_0 = 2$ and the full symbols represent $R_0 = 10$.

We have checked that equation (7) provides a good approximation to the results of the level of infection in the simulations with the various topologies (data not shown). We have also seen that $R_0 = 1$ is the critical value to have a non-null probability of an outbreak occurrence.

These conditions show that we have an exact correspondence between the classical formulation of the SIS model and the approach following standard models in population genetics. We can now study the levels and patterns of genetic diversity in the pathogen populations whose demography follows an SIS epidemiological model.

In figure 1, we plot the level of diversity in samples taken from the whole population, π_t, as a function of p for different values of R_0. It is clear from the figure that when p achieves the value where the average path length drops, i.e., when the topology is small-world, the level of pathogen genetic diversity drops abruptly. This drop is independent of the value of R_0, although the level of diversity maintained in the pathogen metapopulation depends on its basic reproductive number.

In figure 2, we plot the level of diversity, π_t, as a function of R_0 for different topologies. We have considered the stepping-stone model ($p = 0$), small-world networks and the island model. From figure 2(A), we observe a pattern that is independent of the topology: the level of diversity peaks around intermediate values of R_0. This result establishes the existence of two distinct regimes: one for small values of R_0 and a second regime for large values of R_0. In the former regime, where $mK \ll e$, the level of infection is very low and consequently it bounds the level of diversity in the population, since it is expected that diversity will be higher when the total number of pathogens in the metapopulation is bigger. With the augment of R_0, by raising the migration rate m, the level of infection shows a steep growth, and so enables the population to reach a higher level of genetic diversity. The second regime comes about at large R_0 values, which means

J. Stat. Mech. (2006) L12003

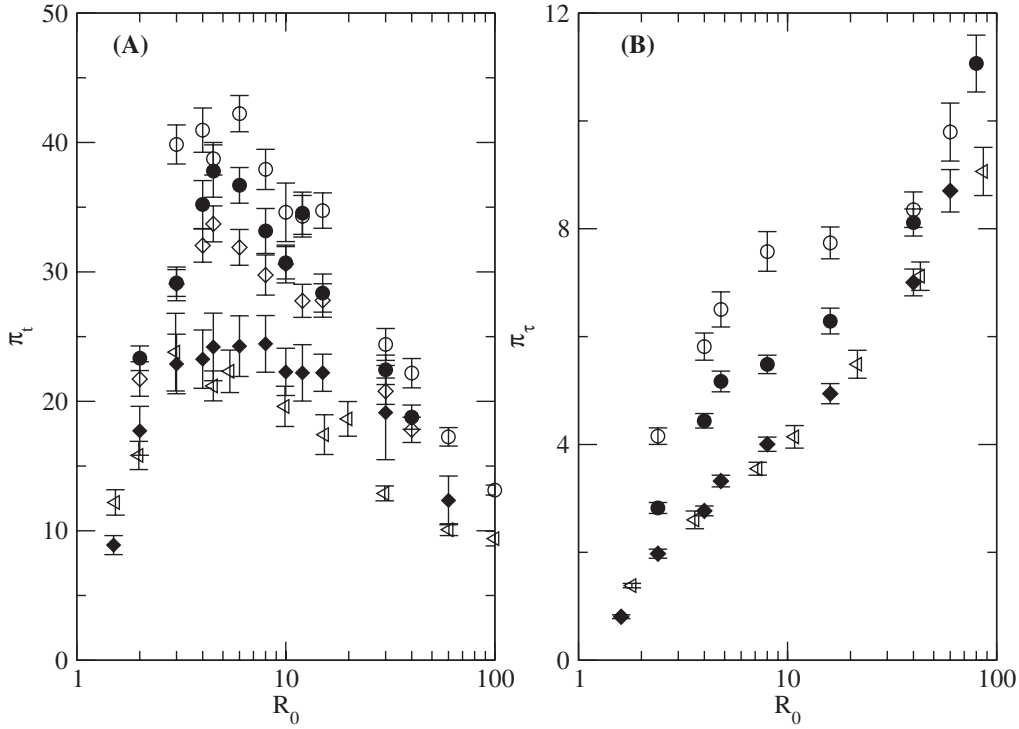

Figure 2. The level of diversity π_t as a function of R_0. The parameters for (A) are $D = 900$, $N_d = 10$, $e = 0.01$, $n_t = 50$ and $\mu = 0.0004$. The parameters for (B) are $D = 900$, $N_d = 20$, $e = 0.1$, $n_t = 50$ and $\mu = 0.0004$. The open circles correspond to $p = 0$ (stepping-stone model), filled circles to $p = 0.05$, open diamonds to $p = 0.1$, filled diamonds to $p = 1.0$ (random-graphs) and left triangles correspond to the island model.

that $mK \gg e$. At this point, the level of infection is close to one and so it is not a limiting factor for diversity growth anymore. But further increasing the migration rate causes a drastic reduction in the isolation between nodes and now the diversity decreases with m. In the limit of very high m, the migration steers the diversity towards the value $\pi_t = \pi_d = 2N_dD\mu$, which corresponds to $\pi_t = 7.2$ for the simulations in figure 2(A).

In figure 2(B), we no longer observe a peak of diversity at intermediate R_0. The main difference here is that the parameter values of figure 2(B) are such that $e > 1/N_d$, i.e., the rate of drift within the host is lower than the rate of extinction of the within host pathogen population. With these parameter values, the expected level of π_t in a non-structured neutral model is 14.4. In a structured population, the level of diversity is always lower than this value.

Another important observation from both figures is that, by continuously adjusting the probability of rewiring p, we drive the level of genetic diversity from that expected for a stepping-stone model ($p = 0$) to that under the island model. This latter model maintains the same level of diversity as a random graph ($p = 1$).

Simple structured models, such as the island model, do not enable us to capture an essential property of real populations, including ecological systems or even pathogen populations, which is the correlation between genetic and geographical distance [19, 9].

J. Stat. Mech. (2006) L12003

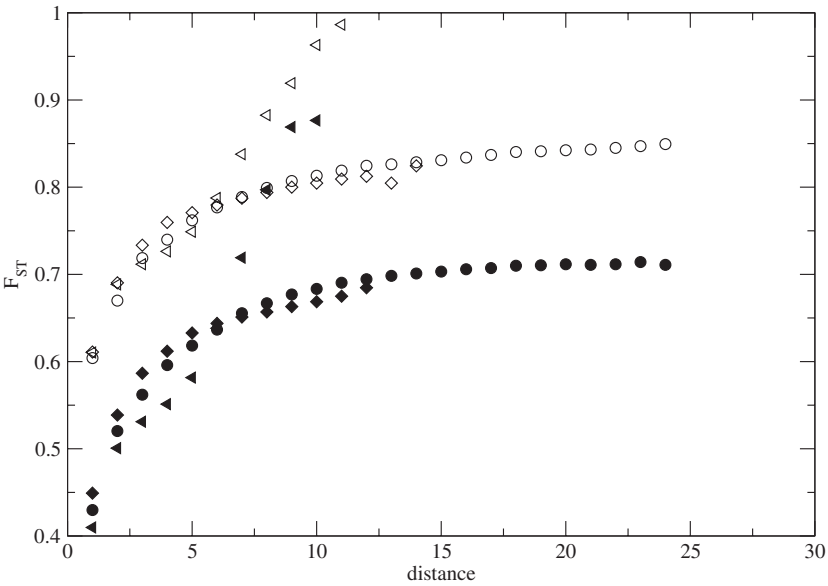

Figure 3. The level of differentiation between hosts, F_{ST}, as a function of their topological distance. The parameters are $D = 625$, $N_d = 10$, $\mu = 0.004$ and $e = 0.01$. The circles are for $p = 0$, diamonds for $p = 0.05$ and left triangles for $p = 1.0$. The open symbols correspond to $R_0 = 15$ and the filled symbols correspond to $R_0 = 30$.

In order to quantify these correlations, we have performed measurements of the level of differentiation between hosts, F_{ST}, as a function of their topological distance, which we take as a surrogate for geographical distance. To the topological distance of any pair of nodes, we attribute the shortest path length between the nodes in the network. The results of our simulations in figure 3 show that, as intuitively expected, pathogens with low R_0 have higher levels of between host genetic differentiation. This qualitative result is independent of the topology. Figure 3 also shows an isolation by distance pattern which depends on the topology. We observe that whereas in random networks the relation between genetic and geographical distance is close to linear, in regular networks it is logarithmic. In topologies with $p = 0.05$, the relation is closer to that in regular than to that in random networks. Furthermore at intermediate distances, F_{ST} is higher in random graphs than in the other networks, with smaller p. This can be explained by the following reasoning: in random graphs the distribution of deme connectivities is Poisson with both a low clustering coefficient and average short path length. The demes that are at the highest distance are those that are more isolated and therefore exchange less migrants, which leads to a higher F_{ST}. With smaller p, the clustering coefficient is larger which leads to a smaller F_{ST}. Albeit we have seen that the level of diversity of random graphs is very similar to that obtained for the island model, importantly in random graphs we also see clear evidence of isolation by distance, as observed in natural populations.

One of the most important roles of studying neutral genetic diversity in pathogen populations, whose demography follows classical epidemiological models, is that such information can be directly tested against sequence information. In particular, deviations

doi:10.1088/1742-5468/2006/12/L12003

J. Stat. Mech. (2006) L12003

from neutral expectations in certain genes can suggest relevant alternative hypothesis, such as the presence of selection shaping pathogen sequences [6].

This work was supported by project POCTI/BSE/46856/2002 through Fund. para a Ciência e Tecnologia (FCT). IG is supported by a FCT/FEDER fellowship. PRAC is supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

[1] Dobson A, *Ecology: Metalifes!*, 2003 *Science* **301** 1488
[2] Wright S, *Evolution in Mendelian populations*, 1931 *Genetics* **16** 97
[3] Grenfell B T, Pybus O G, Gog J R, Wood J L N, Daly J M, Mumford J A and Holmes E C, *Unifying the epidemiological and evolutionary dynamics of pathogens*, 2004 *Science* **303** 327
[4] Sella G and Hirsh A, *The application of statistical physics to evolutionary biology*, 2005 *Proc. Nat. Acad. Sci.* **102** 9541
[5] Simon D and Derrida B, *Evolution of the most recent common ancestor of a population with no selection*, 2006 *J. Stat. Mech.* P05002
[6] Connway D J *et al*, *A principal target of human immunity to malaria identified by molecular population genetics and immunological analyses*, 2000 *Nat. Med.* **6** 689
[7] Falush D *et al*, *Traces of human migrations in Helicobacter pylori populations*, 2003 *Science* **299** 1582
[8] McDonald B A and Linde C, *Pathogen population genetics, evolutionary potential, and durable resistance*, 2002 *Annu. Rev. Phytopathol.* **40** 349
[9] Real L A *et al*, *Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus*, 2005 *Proc. Nat. Acad. Sci.* **102** 12107
[10] Keeling M J and Eames K T D, *Networks and epidemic models*, 2005 *J. R. Soc. Interface* **2** 295
[11] Newman M E J, *The structure and function of complex networks*, 2003 *SIAM* **45** 167
[12] Watts D J and Strogatz S H, *Collective dynamics of small world networks*, 1998 *Nature* **393** 440
[13] Albert R and Barabási A-L, *Statistical mechanics of complex networks*, 2002 *Rev. Mod. Phys.* **74** 47
[14] Milgram S, *The small world problem*, 1967 *Phychol. Today* **2** 60
[15] Pastor-Satorras R and Vespignani A, *Epidemic spreading in scale-free networks*, 2001 *Phys. Rev. Lett.* **86** 3200
[16] Barthélemy M, Barrat A, Pastor-Satorras R and Vespignani A, 2004 *Phys. Rev. Lett.* **92** 178701
[17] Campos P R A, Combadão J, Dionísio F and Gordo I, *Muller's ratchet in random graphs and scale-free networks*, 2006 *Phys. Rev.* E **74** 042901
[18] Kimura M, *The stepping stone model of population*, 1953 *Annu. Rep. Natl Inst. Genet. Japan* **3** 62
[19] Pannell J R and Charlesworth B, *Effects of metapopulation processes on measures of genetic diversity*, 2000 *Phil. Trans. R. Soc.* B **355** 1851
[20] Lloyd A L and May R M, *How viruses spread among computers and people*, 2001 *Science* **292** 1316

J. Stat. Mech. (2006) L12003