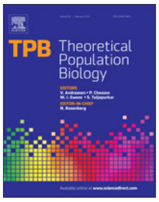
Theoretical Population Biology 122 (2018) 3–4

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb



Commentary

Paul Joyce and the infinite alleles model

Stephen M. Krone

Department of Mathematics, University of Idaho, Moscow, Idaho 83844-1103, United States

ARTICLE INFO

Article history:
Received 26 September 2017
Accepted 3 October 2017
Available online 16 October 2017

Keywords:
Infinite alleles model
Poisson–Dirichlet
GEM
Neutral
Selection

ABSTRACT

Paul Joyce's work touched on a variety of topics in population genetics—from mathematical models of idealized systems to working closely with biologists on experimental evolution and landscape genetics. I will focus on his earlier mathematical/statistical work that centered on the infinite alleles model.

One day during the spring semester of 1998, Paul came to my office at the University of Idaho to let me know that Tom Kurtz had agreed to visit and give a talk. We shared the news with our colleague Dick Gomulkiewicz at Washington State University, just 8 miles down the road. Dick had some good news of his own: John Gillespie would be coming to give a talk at WSU that same semester. As (bad) luck would have it, though, their talks had been scheduled on the same day at the same time. It was too late to change the schedules, but Dick suggested that he could bring John to Moscow for breakfast and a group of us would at least have time for a nice chat. After breakfast, Paul, Tom, John and I settled down for a serious mathematical discussion. Gillespie began by telling us about some curious simulation results he had obtained using a model of selective overdominance that had selection intensity and mutation increasing together with population size (Gillespie, 1999). As usual, he had very clear insights about why this model was a reasonable one to consider, and he wondered if there was a mathematical explanation for the results. The simulations suggested that one would not be able to detect selection, from a sample of genetic data, when selection intensity and the scaled mutation rate increased together at the same rate. Paul had some rough ideas about what might be going on and suggested a few mathematical approaches. Before long, Tom had filled several pages in his pad of paper with equations from the powerful mathematical machinery he always has ready for action. By the time we left breakfast, we had the beginnings of what would become two papers (Joyce et al., 2002, 2003 to be discussed below). Paul always liked telling this story. He had lots of stories he liked to tell, so it seemed fitting to begin this remembrance with this one.

Much of Paul's work in the early phase (before he began active collaborations with biologists) involved the *infinite alleles model*, and it is this work that I will focus on. This model has a rich history in mathematical population genetics. It carries a good blend of biological realism and mathematical abstraction, allowing one to ask biologically important questions that can be addressed with mathematical precision. The infinite alleles model and its close cousin, the infinite sites model, ignore the details of DNA sequences that have become so ubiquitous since advances in sequencing technology revolutionized evolutionary biology. Nevertheless, much of our understanding of what population genetics is about comes from the analysis of these simpler models. More complicated, and less tractable, DNA sequence models have become important because they lead to computational/statistical analyses of sequence data, and Paul worked on these as well.

One of the signature results of the *neutral* infinite alleles model is the Ewens Sampling Formula, which describes the equilibrium probability of observing a given set of allele frequencies in a sample. It encompasses much of the combinatorial flavor of sampling distributions, and Paul was always drawn to this. The Poisson–Dirichlet distribution arises as the equilibrium distribution of the allele frequencies, presented in descending order, of the infinite alleles model. The single parameter, θ, in the Poisson–Dirichlet distribution is the scaled mutation rate. In an influential paper, Donnelly and Joyce (1989) showed that the size-biased permutation of the Poisson–Dirichlet distribution with parameter θ is the GEM distribution with parameter θ and, similarly, the ranked permutation of the GEM distribution is the Poisson–Dirichlet distribution. This means that calculations which are invariant under permutations can be carried out with either distribution. This is useful since the GEM distribution is much easier to work with than the Poisson–Dirichlet, and it has a nice "stick-breaking" construction. Donnelly and Joyce also pointed out how these connections could be used to advantage in calculating limiting behavior.

E-mail address: krone@uidaho.edu.

https://doi.org/10.1016/j.tpb.2017.10.001
0040-5809/© 2017 Elsevier Inc. All rights reserved.

The infinite alleles model with *selection* also has a nicely characterized stationary distribution; it is absolutely continuous with respect to the neutral stationary distribution (Poisson–Dirichlet) and the Radon–Nikodym derivative serves as a likelihood ratio when testing selection versus neutrality. In the case of overdominance (heterozygote advantage), this likelihood ratio is proportional to $\exp(-\sigma H(x))$, where σ is the scaled selection intensity and $H(x)$ is the population homozygosity when x is the vector of allele frequencies. One of the keys to our treatment of Gillespie's problem was proving the asymptotic normality of the homozygosity as $\theta \to \infty$. While we were working on this, Paul found out that Bob Griffiths had proved a special case 20 years earlier—not the first time he had anticipated a result. Ever after, when Paul was working on a promising mathematical idea he would joke that he had better take a look at Bob's publications to see if he had already done it.

Working with Paul, one always heard him tell stories that weaved together math or science with the people who did the work. Although he could be found pacing alone and muttering to himself for hours (J. Joyce, personal communication) as he tried to crack a difficult problem, for Paul science was a very social endeavor and he truly enjoyed the social tapestry that lay beneath the process of doing science. Years later, Paul would become a very effective and beloved Dean of the College of Science at the University of Idaho. When I would go to meetings on mathematical genetics (often with Paul), colleagues would initially express disbelief that Paul had gone over to the Dark Side. However, after a few moments, they would come to realize how he could be really effective in such a role. The personal side of science, which he enjoyed so much, found many expressions during his career.

Back to Gillespie's problem. Gillespie's simulations suggested that when selection intensity scales like mutation rate, $\sigma = c\theta$, in the limit as $\theta \to \infty$, selection and neutrality would be indistinguishable. The reason for the limit involved the usual diffusion scaling of mutation $u = 4N\theta$ and selection $s = 4N\sigma$, with population size N becoming large (while s and u remain fixed). In Joyce et al. (2003), we were able to demonstrate this and extend it to show that overdominant selection is indistinguishable from neutrality when $\sigma = c\theta^{\gamma}$ for any $\gamma < 3/2$, while selection can be detected when $\gamma > 3/2$. This was a case of starting with a guiding conjecture and then following the mathematics wherever it led. We needed to calculate the likelihood ratio for the stationary distributions under both neutrality and selection. Due to the relation between these two distributions, mentioned above, one of the key ingredients was to demonstrate a Gaussian limit theorem for the population homozygosity under neutrality. This ended up being quite technical and resulted in a separate paper (Joyce et al., 2002). It involved weak convergence of stochastic integrals à la Kurtz and Protter, as well as many other approximations and convergence results. There were so many places for the results we needed to go wrong, but Paul was undaunted as we struggled to deal with all sorts of technical calculations. "It has to work", he kept saying, and it did. Once Paul had acquired the "scent" of a problem and felt in his bones what must be true, he would go at it relentlessly–chewing through multiple pens in the process. At times like these, Paul's wife Jana would chuckle and say "The man is driving me crazy!"

The difficulty of detecting selection based on genetic data was also a theme in a number of Paul's other publications. Joyce and Tavaré (1995) developed Poisson approximations to the Ewens Sampling Formula to show that rare alleles cannot be used to determine if selection is acting. In Joyce (1994), Paul showed that any parameter distinguishing an infinite alleles model with selection from the neutral infinite alleles model cannot be consistently estimated based on gene frequencies at a single locus. In other words, increasing the sample size is not enough.

Without going into details, other notable papers that Paul wrote on the infinite alleles model include (Tavaré et al., 1995), which showed that the consistent estimator (K=number of alleles in a sample) for the scaled mutation rate, θ, under neutrality is consistent and asymptotically normal even under selection; (Joyce, 1991), which used calculations with the GEM distribution to calculate Bayes estimators for the frequency of the oldest allele in a sample; and (Joyce, 1989), which showed that adding information on age order would not improve the power of Watterson's (sample homozygosity) test statistic for selective neutrality.

Paul's thinking about mathematical and statistical approaches to population genetics was strongly influenced by Simon Tavaré, Peter Donnelly, and Warren Ewens. Simon was Paul's dissertation advisor at the University of Utah and Paul later took a visiting position at USC after Simon had moved there. Various visits with Peter in London and Oxford, and with Warren at occasional meetings, were also essential. One could see the influence of this dynamic troika in Paul's publications, his approach to mathematical and statistical aspects of populations genetics, and in the stories and jokes he liked to tell. He occasionally talked of typing up a list of "Simonisms"—somewhat "colorful" quotes from one of his mentors that found application in many situations. When he visited Peter, the long train rides from the university back to Peter's house were frequently filled with lively discussion and brilliant calculations. After getting off the train, however, the brilliant calculations were sometimes found to have holes. Paul would say "If we'd just stayed on the train, everything would have been fine!"

Just as Paul benefited greatly from the mentors who helped to guide and inspire him, he was very fond of his own Ph.D. students and post docs (including Kathrine Johnson, Zaid Abdo, José Miguel Ponciano, Erkan Buzbas, Andrzej Wojtowicz, Hua Feng, Craig Miller, Darin Rokyta, Roland Fleissner), and many others (including my Ph.D. student Grant Guan). They found Paul to be accessible, friendly, and full of good ideas, and he played a big role in their own budding careers. He, in turn, found great joy interacting with these bright minds and he never tired of bragging about them.

References

Donnelly, P., Joyce, P., 1989. Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex. Stochastic Process. Appl. 31, 89–103.

Gillespie, J.H., 1999. The role of population size in molecular evolution. Theoret. Popul. Biol. 55 (2), 145–156. http://dx.doi.org/10.1006/tpbi.1998.1391.

Joyce, P., 1989. Is knowing the age-order of alleles in a sample useful in testing for selective neutrality?. Genetics 122, 705–711.

Joyce, P., 1991. Estimating the frequency of the oldest allele: A Bayesian approach. Adv. Appl. Probab. 23 (3), 456–475. URL http://www.jstor.org/stable/1427617.

Joyce, P., 1994. Likelihood ratios for the infinite alleles model. J. Appl. Probab. 31, 595–605.

Joyce, P., Krone, S.M., Kurtz, T.G., 2002. Gaussian limits associated with the Poisson-Dirichlet distribution and the Ewens sampling formula. Ann. Appl. Probab. 12 (1), 101–124.

Joyce, P., Krone, S.M., Kurtz, T.G., 2003. When can one detect overdominant selection in the infinite-alleles model?. Ann. Appl. Probab. 13 (1), 181–212.

Joyce, P., Tavaré, S., 1995. The distribution of rare alleles. J. Math. Biol. 33, 602–618.

Tavaré, S., Ewens, W., Joyce, P., 1995. Robustness of the Ewens sampling formula. J. Appl. Probab. 32, 609–622.