New Mathematical Challenges from Molecular Biology

Richard Durrett (Cornell U.),
Ed Perkins (U. British Columbia)

September 6–11, 2009

1 Introduction

This meeting brought together a broad spectrum of researchers across the continuum from mathematics to molecular biology. The coalescent and other genealogical or *dual* processes were a common tool for the study of the effects of natural selection, population subdivision, large family size, etc. on genetic diversity. The motivation for these investigations is, of course, to use various statistics to infer which forces have acted in the evolution of genetic loci. Rather than try to recount all of the many developments, we will highlight some.

2 Six Exciting Research Directions

Next generation sequencing methods produce a huge number of short DNA reads. Andy Clark described problems in the use of these methods to study gene conversion in tandem arrays of duplicated genes. Mathematical methods need to be developed if we are going to make optimal use of this type of data.

In some bacteria the genic content varies widely between individuals. Peter Pfaffelhuber discussed methods for inferring rates of evolution, which generalize the infinite alleles model but lead to a number of new mathematical and biological questions.

Much of genetics studies the evolution of neutral loci, which have no consequences for the reproduction of individuals. However, in many cases when the mean number of offspring is constant the variance of the number of offspring varies. Jay Taylor investigated the consequences of this fecundity variance polymorphism for genealogies. His model was biologically motivated but also had some fascinating mathematical properties. The

models typically produce differences in the (backward) genealogies but not in the forward frequency diffusion models.

It is important to understand the mechanisms of regulatory complex assembly to understand how genes are described. Steve Evans showed that "simple models" with only a dozen states lead to serious mathematical complexities when one wants to symbolically compute formulas rather than just produce simulation curves.

Paul Joyce works with experimentalists studying the evolution of viruses in the laboratory. He showed how extreme value theory could give insights into the distribution of fitness improvements in this system. More specifically although past work assumes the fitness distribution belongs to the more standard Gumbel domain of attraction he considered two other domains from extreme value theory (Weibull and Frechet corresponding to truncated and fat tails, respectively). The lab results suggest that in some cases the theoretical results predicted by assuming it belongs to the Weibull domain of attraction provide a better fit. The combination of elegant mathematics with experimental results was particularly attractive.

Mueller's ratchet refers to the steady accumulation of deleterious mutations in systems, especially those without recombination. Anton Wakolbinger asked "When does the ratchet click rarely?" (in the large population limit). His conjectured answer (with Alison Etheridge and Peter Pfaffelhuber) in terms of the selection and mutation parameters is an interesting open problem which kept some of the participants busy through the meeting. The Fleming-Viot model used to model the ratchet in the large population limit is a discrete type version of a continuous branching model being studied by one of the students (Hardeep Gill) at the meeting.

3 Substantial progress on a recent topic

Six talks on the second day concerned the Λ-coalescent which occurs when individuals have widely variable number of offspring. Talks discussed how to use the model for inference, compute likelihoods for this model and its site frequency spectrum, investigate its properties when selection acts or individuals are distributed in space. At the biological end, Ori Sargsyan proposed a coalescent model (actually a mixture of Λ coalescents) to describe the multiple mergers of the genealogies of marine species such as the Pacific oyster. At the mathematical end, Jason Schweinsberg showed how this process and the Bolthausen-Sznitzman coalescent in particular arose in a branching Brownian motion with absorption. The latter is proposed as substitute for a fixed population model with selection proposed by Brunet et al who conjectured the Bolthausen-Sznitman coalescent should describe its genealogies.

This meeting allowed individuals who work on this topic to exchange ideas and to explain the workings and consequences of this model to biologists, and also allowed the biologists to present modeling situations where it may arise.

4 Scientific Progress Made

It is now three weeks since our meeting and it is unrealistic to point to immediate scientific breakthroughs which have been worked out since the meeting. However, a number of new insights were communicated at the meeting and new collaborations have been launched. The best evidence for the benefits of a meeting bringing together biologists and mathematicians is in the letters we have received from the participants (which include at least one "Eureka" after all):

Senior Scientists

Dear Ed and Rick,

Thanks so much for taking the time to set up such a successful week. The BIRS workshop will be helpful to me and my research group in a variety of ways.

- Rasmus Nielsen's work on probabilities of identity by descent and some of Jeff Jensen's approaches using summary statistics in an approximate Bayesian computation may very well be helpful to the researchers in Michael Hammer's lab

- Paul Joyce's small scale, laboratory based, evolutionary models is close in scientific perspective to Joanna Masel, an evolutionary biologist at the University of Arizona. I plan to present some of these results to the Masel group. Joanna and I share a doctoral student, Grant Peterson. Idaho looks like a good postdoctoral opportunity for Grant.

- Jay Taylor and Bob Griffiths have made advances in the work on ancestor selection graphs under a variety of backgrounds. This will be helpful in our group's work on the nature of modeling of genetic resistance to disease for our study area in the Indonesian archipelago.

- Steve Evans approach to regulatory complex assemblies is very similar to an approach used by some of my colleagues and me in the study of allosteric proteins in the presence of a heme. I plan to maintain contact with Steve as we try to add to the list of biochemical systems that will be amenable to this Marlov chain, pinch point methodology.

- Our group is presently working with Ryan Gutenkunst. However, it was a good opportunity to catch up.

Finally, I enjoyed the advances seen on work on the lambda coalescent and on sampling formulae. It was also very helpful to me to make personal connections with collaborators of my collaborators and delightful to have the opportunity to catch up with a few individuals that I have now known for more than a couple of decades.

Best, Joe Watkins, U. Arizona.

Rick and Ed, Great thanks to both of you for organizing a wonderful meeting in a fabulous setting. ... The mathematics of the lambda-coalescent is very interesting, but at this meeting I finally got some examples of marine organisms where this type of model might be useful. All of the talks were interesting, but a few stood out for me because they had a profound effect on my thinking. These talks either provided a new perspective on topics I am or have worked on, or they got me interested in topics that I had yet work on. Andy Clark's talk is a good example of the later. At Idaho, I am involved in a center grant from NIH where, as part of that grant, we invested in the relatively new 454 sequencing. So I have been trying to wrap my head around the implications of this next generation sequencing technology on population genetics. I had never thought about how these new short read sequence data

could allow for a much more comprehensive study of gene conversion. It has been a long time since I have thought about this early work of Nagylaki and Ohta, so I am planning to go back and read those old papers as well as Andy's work. Rasmus Nielsen's talk was very insightful...I have been developing methods for detecting overdominance selection, where the primary data for my methods is the HLA region. However Rasmus's clever identity by descent approach shows that overdominance cannot explain the significant increase in IBD. His paper is now a must read for me. Yun Song's talk was really impressive and it made me wonder if the same logic used to get a large recombination approximation to the sampling distribution for the Ancestral Recombination Graph (ARG) could be used to get a similar approximation to the Ancestral Selection Graph (ASG) that was devised by Krone and Neuhauser in the late 90's. This could have a major impact on making ASG useable as an inference tool. I plan to be in contact with Yun to get his views on this problem. I always learn so much from these smaller conferences with a more specific agenda than the big conferences. So I again, thanks for putting this together. Also, BIRS and the Banff Center facilities were great.

Paul Joyce, U. Idaho

Dear Rick and Ed,

Thanks again for hosting the BIRS meeting on Mathematical Challenges from Molecular Biology. I learned a great deal at the meeting, and got an especially concrete take-home from the meeting. We have been working with a group that sequenced two genes in 15,000 people, with the goal to understand the nature of rare variation in humans. Rare variation is all the rage now in human genetics, because people think that it might be responsible for the "missing heritability" – the gap between estimated heritability and the variation explained by genome-wide association studies. Our sample of 15,000 actually exceeds the typical estimates of the human effective population size (10,000), violating assumptions of the Kingman coalescent. I did some scratching and simulating and saw that this should result in some multiple mergers, possibly inflating rare variants. Wakeley and Takahashi wrote a paper on the problem, but I had no idea that the lambda coalescent was generating so much excitement among mathematical population geneticists. I got a great deal from the talks and discussions with Bob Griffiths, Matthias Birkner, Ori Sarasayan and Nathaniel Berestycki. This is work that we are deeply engaged in now, so the timeliness of finding this literature and this gang of experts was perfect.

While I do a certain amount of work in theoretical population genetics, I am not a mathematician, and most of the mathematicians at the meeting could have given talks that would be totally incomprehensible to me. But they found a good compromise level that kept all well engaged. I deeply appreciate this kind of stretch-your-boundaries meeting, and the way the BIRS series is run is just superb.

sincere thanks, Andy

Andrew G. Clark Molecular Biology and Genetics 227 Biotechnology Building Cornell University

Ed and Rick,

1) I had some discussions with Rick Durrett and John Mayberry about cancer modeling, which could help to guide the direction of my future research in this area.

2) I had a short discussion with Jay Taylor about the model that I studied in my talk,

which might enable me to formulate more biologically realistic models for future related work.

3) I had a brief discussion with Nathanael Berestycki concerning how to finalize the write-up of our joint paper. We also briefly discussed possible follow-up work.

4) I learned about some intriguing open problems, including a problem about Muller's ratchet from Anton Wakolbinger's talk and a problem about characterizing the possible coalescent processes dual to a given diffusion from Jay Taylor's talk.

5) From the conference overall, I got the impression that coalescent processes with multiple mergers were being taken more seriously by biologists than I had previously thought.

Jason Schweinsberg, U. California, San Diego.

Junior Researchers

As a young researcher, I found the dialogue between mathematicians and biologists at this meeting a refreshing example of what cross disciplinary research can be like. It was a great opportunity to learn more about what type of questions researchers in population genetics and evolutionary biology really care about and I came away from the meeting feeling less timid about actually discussing my work with people in other disciplines. The last two papers I worked on claimed to have biological motivation, but were criticized by referees (and rightly so) for lacking biological relevance. I realize now that any future projects I work on related to the analysis of biological models could be very much improved by increased discussion with actual biologists and hopefully made a couple of contacts this trip that I can write to in the future with questions.

John Mayberry, Cornell U.

I have been working pretty hard on the project I spoke about in Banff for the whole summer, and was fairly happy with how it came together. Only one thing was missing: a theorem describing the behavior for larger distance matrices. Although it's clear that the level of understanding we have of the system for small cases won't be possible in the larger case, one should be able to prove something about it. I was very much hoping to prove such a theorem in the weeks leading up the conference, and had even left a "gap" in the slides for such a theorem. Unfortunately, no such theorem presented itself. However, in the afternoon after giving my talk I had the insight I needed to prove the missing theorem. In fact, the new theorem is quite a bit more general than I had hoped for, and will be one of the main results for this work. I'm not sure if it was a product of the discussions I had at BIRS, the quiet time away from the university, or the stimulating mountain air, but I'm definitely going home with a souvenir!

Thank you again for inviting me!

Erick Matsen, U. California, Berkeley.

Hi Ed and Rick,

Thanks very much for organising the workshop – I really enjoyed it.

This is still an area I'm learning about, so it was incredibly useful to me to have such an array of experts, and many excellent talks, to learn from.

More particularly:

I learnt from Jason's beautiful talk that the Bolthausen-Sznitman coalescent (which I have spent a lot of time studying) might have some biological relevance after all!

It looks like Nathanael, Alison and I might have got started on a collaboration based on a something which came up in discussion following Ori's talk.

Nathanael and I had some interesting conversations about exceptional times for coalescent processes. Likewise, there's work for us to do here!

Best wishes, Christina Goldschmidt, U. Warwick.