Why would a pure mathematician work in biology?

Andrew Francis

School of Computing and Mathematics
University of Western Sydney

April 2011

1. The seduction of biology
2. Research on TB
3. Genome organization
4. Reflections

The seduction of biology

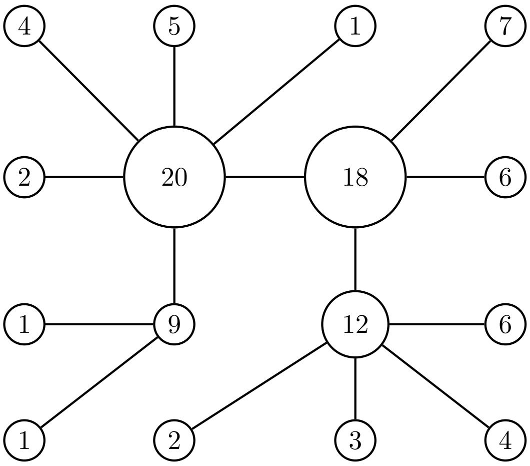
The fateful BBQ

- Mark Tanaka: "how might one assess the severity of an outbreak of tuberculosis using molecular data".
- TB is a major disease:
  - Caused by the bacterium *Mycobacterium tuberculosis*, transmitted through the air;
  - 1.6 million die each year from TB;
  - A third of the world's population carries the pathogen.
- It was exciting because it is clearly so important,

  and
- I thought it sounded like a graph theory problem.

Graph theory?

- One can think of an outbreak as a directed graph.
- I imagined data might be able to be represented something like this (with arrows):



- An index of severity from the graph?

Reflections

Can an algebraist do biology?

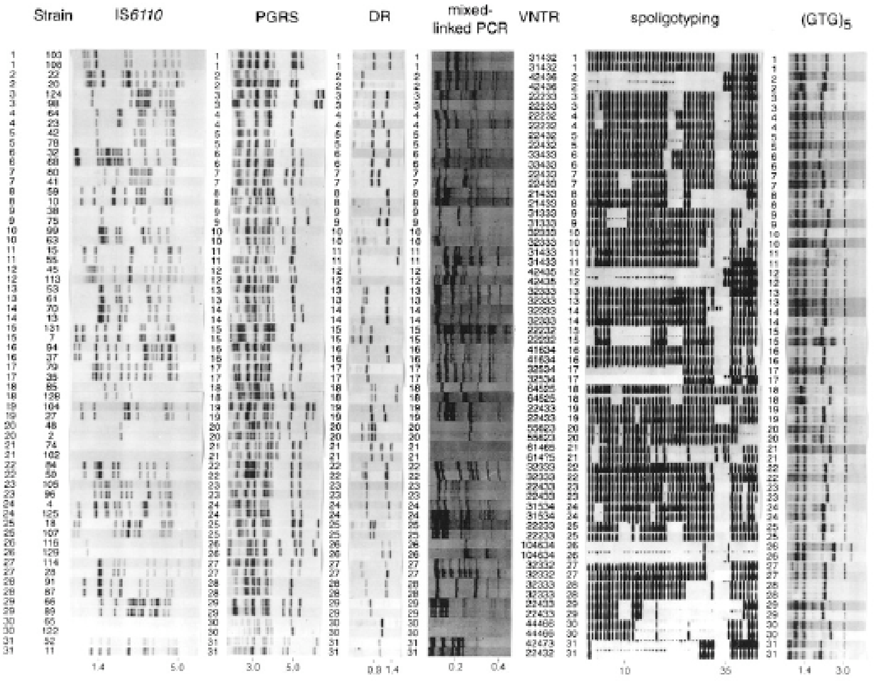
- Work questions:
  1. Am I *allowed* to do this?
  2. What is my job?
  3. Will my employers be disappointed if I stray from the path?
- Moral imperatives?
  1. Did I have a moral obligation to algebra? After all, it seemed I was trained for that.
- Research issues?
  1. Will I lose whatever credibility I have among algebraists?
  2. Will I be able to maintain two disjoint branches of research?

Answers?

- Work:
  1. Nothing in my contract said I had to do only algebra research.
- Moral imperatives?
  1. I did feel a certain betrayal.
- Research issues?
  1. Credibility?
  2. Maintaining output in algebra?
     These were serious: it is hard enough trying to stay at the cutting edge of algebra when it's one's only research area.
- Conclusions:
  - Hedonism? It was pleasant and satisfying research to do, and people to do it with, so I did it.
  - I would do my best to continue to work in both.

What the TB data look like

- The molecular data in general breaks the sample into clusters with an identical genotype.
- Genotyping data look like this:



from Kremer *et al.*1999

Interpreting data

- How can you use data like these to tell how “bad” an outbreak is?

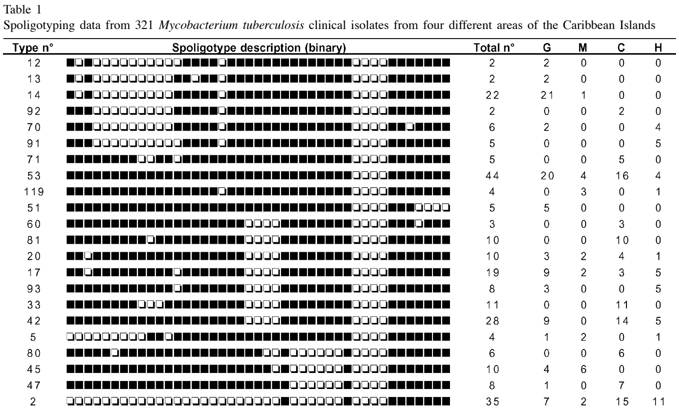


Table 1
Spoligotyping data from 321 *Mycobacterium tuberculosis* clinical isolates from four different areas of the Caribbean Islands

Type n°	Spoligotype description (binary)	Total n°	G	M	C	H
12	■□■□□□□□□□□□□■■■■□■■■■■■■■■■■■■■■□□□□■■■■■■■	2	2	0	0	0
13	■□■□□□□□□□□□■■□■■□■■■■■■■■■■■■■■□□□□■■■■■■■	2	2	0	0	0
14	■□■□□□□□□□□□■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	22	21	1	0	0
92	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	2	0	0	2	0
70	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■■■□□□□■■□■■■■	6	2	0	0	4
91	■■■□□□□□□□□□□■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	5	0	0	0	5
71	■■■■■■■■□□■■□■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	5	0	0	5	0
53	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	44	20	4	16	4
119	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	4	0	3	0	1
51	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□□□□	5	5	0	0	0
60	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■□■■■	3	0	0	3	0
81	■■■■■■■■■□■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	10	0	0	10	0
20	■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	10	3	2	4	1
17	■■□■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	19	9	2	3	5
93	■■■■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	8	3	0	0	5
33	■■■■■■■■□□□■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	11	0	0	11	0
42	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	28	9	0	14	5
5	□□□□□□□□□■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	4	1	2	0	1
80	■■■■■□■■■■■■■■■■■■■■■■□□■□□□□□□■□□□□■■■■■■■	6	0	0	6	0
45	■■■■■■■■■■■■■■■■■■■■■■■□■□□□□□□■□□□□■■■■■■■	10	4	6	0	0
47	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	8	1	0	7	0
2	□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□□■□□□□■■■■■■■	35	7	2	15	11

Some of the spoligotypes from Duchene et al., 2003

- Some assumptions have been common, for example,
  - homoplasy is rare enough to ignore,
  - the mutation process is slow enough to ignore.
- These imply each “cluster” arises from a single re-activated source, and so remaining cases result from recent transmission.

A recent transmission index

- The proportion of recent cases according to this model is often used as a measure of the severity of the outbreak.
- If there are n isolates, g genotypes and n_i cases of genotype i, this proportion is

$$\frac{1}{n}\sum_{1\leq i\leq g}(n_i-1)=\frac{n-g}{n}=1-\frac{g}{n}.$$

- Is it really likely that this reflects anything like the "severity" of an outbreak?
  - It is sensitive to additional re-activated cases (singleton clusters), and
  - it does not account for mutation within the outbreak:
    a high mutation rate may produce many small clusters.

What is severity?

- Is "severity"
  - the reproductive number (the average number of cases each individual infects)?
  - the rate of growth of the population?
- The former of these is the average out-degree of the (unknown) directed graph representing the outbreak, whose nodes are the individual cases and where edges represent transmission.
- The latter is difficult to establish with flat data (without a time reference).
- We developed an alternative index incorporating mutation, based on the MLE for the reciprocal of the generation time.
  - It did better in tests, but not well enough.
    (for instance, it was sensitive to sample size).

[Tanaka & Francis, *Infection, Genetics and Evolution*, 2005]

Alternatives?

Alternatives to indices:

1. parameter estimation using approximate Bayesian computation (ABC).
2. *relative* growth rates of individual clusters.

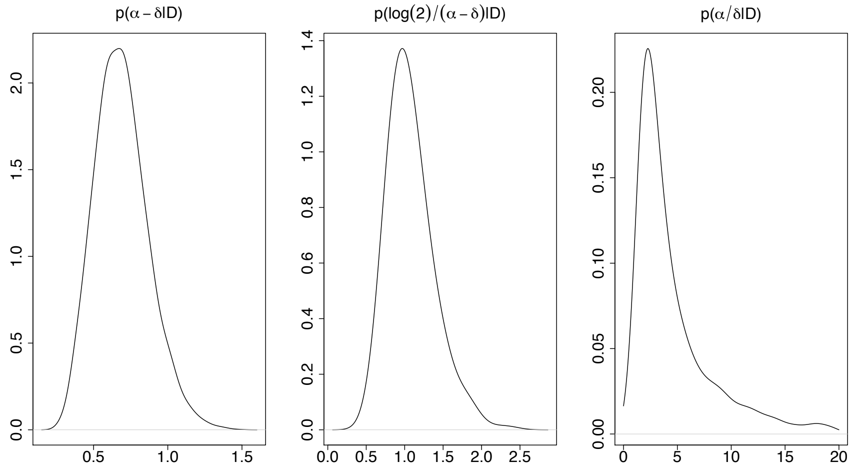
1. ABC

How it works

- Construct a simple model of an outbreak that incorporates both transmission and mutation: an extension of the linear birth-death process.
- Input parameters are the mutation rate μ, birth rate α, death rate δ per case per year.
  - Events occur stochastically.
  - Assume some things – mutation rate, transmission rate – are constant during an outbreak.
- Run the model simulation with different parameter values, selecting those parameters that better fit the observed data (according to some chosen summary statistics).
- (involve an ABC expert: Scott Sisson, UNSW).

1. ABC parameter estimates

- Compound parameters estimated:
  - nett transmission rate $\alpha - \delta$: 0.68 per case per year (median)
  - doubling time $\ln 2/(\alpha - \delta)$: 1.02 years
  - reproductive value α/δ: 3.43

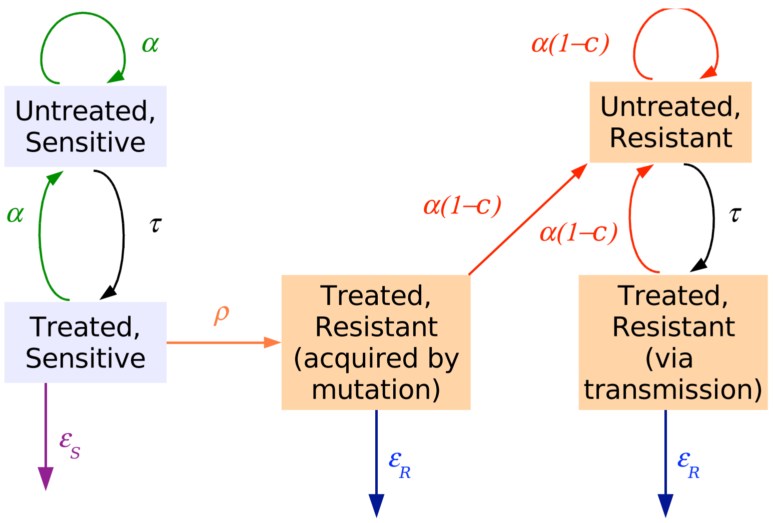


[Tanaka, Francis, Luciani, Sisson, *Genetics*, 2006]

- These are consistent with other estimates obtained with different (epidemiological) methods.

1. ABC: studying drug resistance

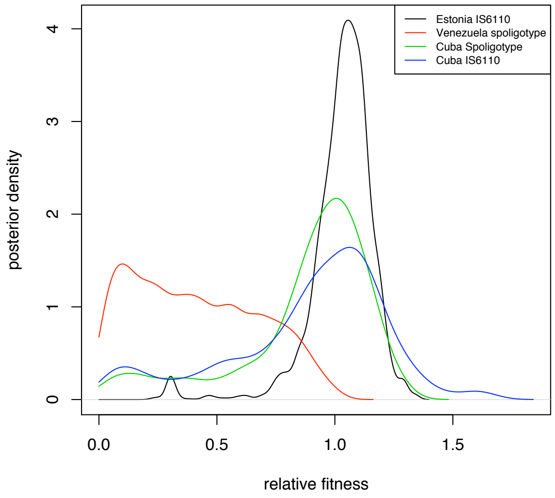
This idea can be extended to model the evolution of drug resistance.



Not shown: δ = rate of death or natural recovery; μ = marker mutation rate.
[Luciani, Sisson, Jiang, Francis, Tanaka, *PNAS*, 2009]

1. ABC: drug resistance conclusions

- We used eleven summary statistics, and studied three data sets.
- The data were consistent with the drug resistant strains being at least as fit as the sensitive strains.



- Over 90% of drug resistant cases arose from transmission rather than treatment failure.

Aside: impact of results

- These ABC papers (*Genetics* 2006, *PNAS* 2009) have had greater impact, in terms of citations, than anything else I've been involved with.
- In pure mathematics, such "impact" is usually the result of the solution of a difficult and important problem.
- In this case, while we did address ("solve" is not the right word) difficult and significant problems, many citations have been to do with our use of ABC in epidemiology.
- In other words, our *methods* generated as much buzz as the results.

  (the methods were a result of the cross-disciplinary collaboration: biology, mathematics, statistics).
- [Disclaimer: this is an observation, not an endorsement of these measures of impact.]

2. Relative growth rates

- With *spoligotyping*, more information about the mutation process is available.
- Spoligotypes (spacer oligo-nucleotide types) are essentially binary strings of length 43.

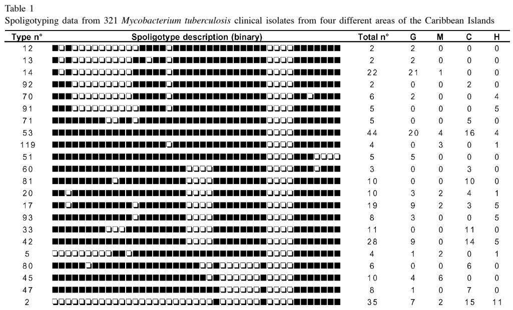


Table 1
Spoligotyping data from 321 *Mycobacterium tuberculosis* clinical isolates from four different areas of the Caribbean Islands

Type n°	Spoligotype description (binary)	Total n°	G	M	C	H
12	■□■□□□□□□□□□□■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	2	2	0	0	0
13	■□■□□□□□□□□□■■□■■□■■■■■■■■■■■■■■□□□□■■■■■■■	2	2	0	0	0
14	■□■□□□□□□□□□■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	22	21	1	0	0
92	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	2	0	0	2	0
70	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■■■□□□□■■□■■■■	6	2	0	0	4
91	■■■□□□□□□□□□□■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	5	0	0	0	5
71	■■■■■■■■□□■■□■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	5	0	0	5	0
53	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	44	20	4	16	4
119	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■	4	0	3	0	1
51	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□□□□	5	5	0	0	0
60	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■□■■■	3	0	0	3	0
81	■■■■■■■■■□■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	10	0	0	10	0
20	■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	10	3	2	4	1
17	■■□■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	19	9	2	3	5
93	■■■■■■■■■■■■□■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	8	3	0	0	5
33	■■■■■■■■□□□■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	11	0	0	11	0
42	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■	28	9	0	14	5
5	□□□□□□□□□■■□■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	4	1	2	0	1
80	■■■■■□■■■■■■■■■■■■■■■■□□■□□□□□□■□□□□■■■■■■■	6	0	0	6	0
45	■■■■■■■■■■■■■■■■■■■■■■■□■□□□□□□■□□□□■■■■■■■	10	4	6	0	0
47	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■	8	1	0	7	0
2	□□□□□□□□□□□□□□□□□□□□□□□□■□□□□□□■□□□□■■■■■■■	35	7	2	15	11

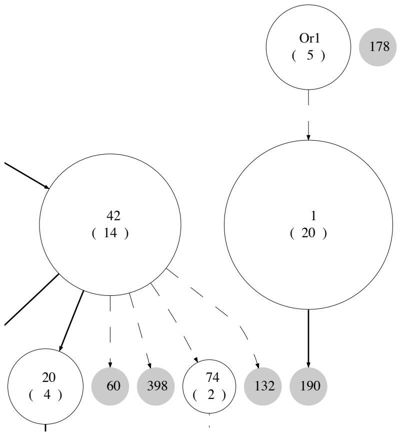
Some of the spoligotypes from Duchene et al., 2003

- Mutation occurs through deletions of adjacent blocks →
  - can identify genotypes (possibly) related by a single mutation, and
  - can infer the direction of the mutation.

2. Relative growth rates

Mutations mark time

- Large clusters could be a result of rapid spread, or simply age.
- Heuristic:
  - given the mutation rate is fixed, the number of mutations a genotype has experienced indicates time.
  - therefore, the ratio of cluster-size to out-degree indicates rate of growth.

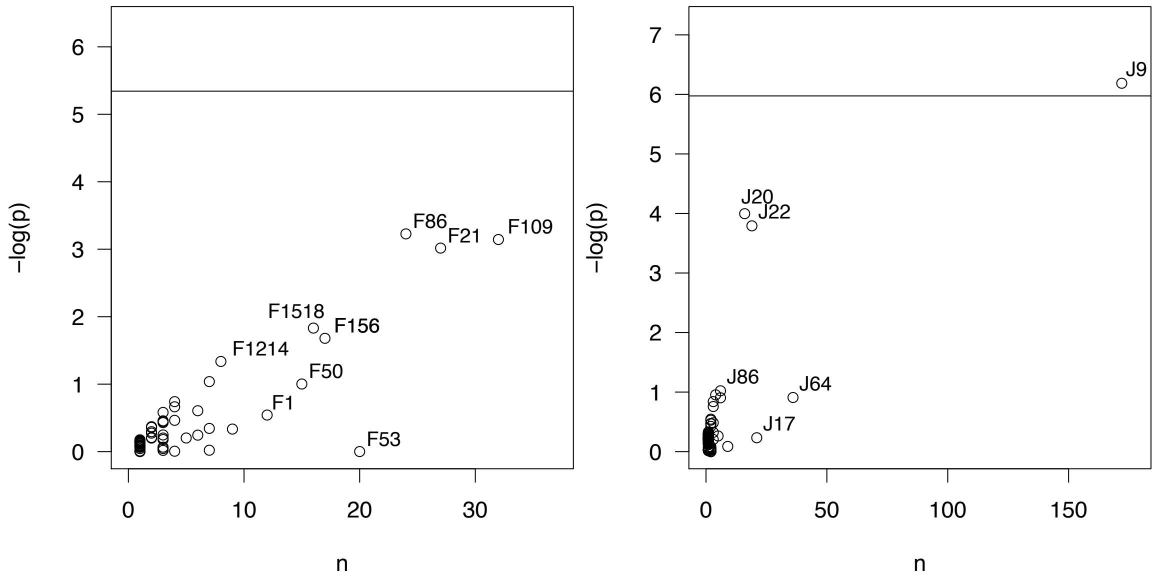


- We developed a method that identifies relatively fast spreading strains, correcting for multiple testing.

[Tanaka & Francis, *PNAS*, 2006]

Emerging strains

Applying this technique to some published data sets, several strains were identified as "emerging", including the W-Beijing strain.



Data from Ferdinand et al (2005, Madagascar); Jou et al. (2005, Taiwan)
p is the probability of observing fewer out-edges than actually observed.

Horizontal line represents the threshold under the Dunn-Sidak correction for multiple testing, significance level 0.25

Reflections on doing TB research

The positives

- This research programme:
  - has been very satisfying, contributing to an important area (especially the drug resistance work)
  - contained some neat ideas (especially the emerging strains)
  - was a lot of fun because of the people I was working with.
- According to some modern measures, it was also successful — probably more so than my algebra work:
  - Papers in well-ranked journals,
  - Grants.

Reflections on doing TB research

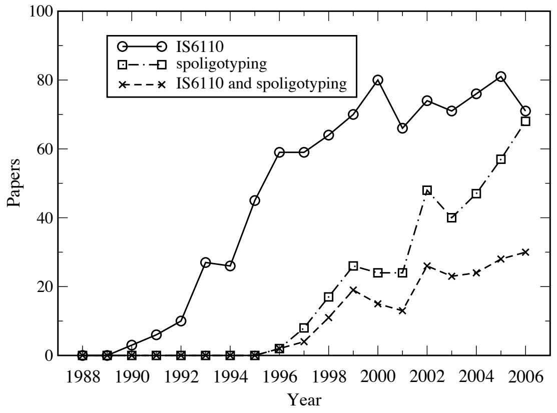
The negatives

There were some downsides. . .

- While I was pleased with some clever ideas, the mathematical theory was not very deep.
  (There was some tricky stats, and we did use some ideas from graph theory and combinatorics).
- Opportunity cost: difficult to raise the depth and breadth of my algebra output.
  I have stayed in a algebraic research groove.
- Results that relate to a specific molecular marker are rather impermanent, because technologies are improved and new markers are developed.

Impermanence

- In 2008 we wrote
  The number of tuberculosis papers referring to IS6110 or spoligotyping is growing.
  (Numbers from PubMed)



Luciani, Francis, Tanaka, *Infection, Genetics & Evolution*, 2008

- But since then:

	2007	2008	2009	2010
IS*6110*	64	73	58	59
Spoligotypes	64	77	72	79
both	23	26	17	21

- Quite a lot of the current papers referring to spoligotypes are using an extension involving *VNTR*.
- Some of our best ideas will soon be redundant.

Ontology

- There is a trade-off between immediacy and permanence:
  - On the one hand, one can develop ideas that help resolve immediate topical questions for a particular purpose.
  - On the other, one can address fundamental questions about the nature of living organisms.
- The latter are more permanent, and closer ontologically to mathematics.
- We are now studying processes giving rise to genome structures observed in bacteria.
- Evaluating hypotheses explaining such structure might be a more lasting contribution.

Genome organisation

Understanding what we see

- All bacteria have their DNA on a circular genome
- We are learning more about the structure of this DNA all the time.
- For instance, we know that genes on the same pathway are often located in the same region of the genome.
- We also know some of the evolutionary mechanisms that occur:
  - segments of DNA can be moved around (often "inverted")
  - segments can be "horizontally transferred" from a neighbouring organism
  - segments can be deleted, or duplicated.

Genome organization

Competing hypotheses

- One hypothesis asserts that *horizontal transfer* explains pathway clustering
- Another describes *cryptic variation*, in which a pathway is acquired despite a cost to the organism in carrying a partial pathway.

- These cannot be tested in the lab because of the timeframe, but can through models and simulations.
- These models generally involve quite a bit of combinatorics, and often simulations (deterministic or stochastic).

Bringing in algebra

- Note that in none of the problems I have described has algebra reared its head.
- I will briefly describe some significant problems in bacterial evolution that *do* involve algebra.

DNA from near and afar

Local and topological evolution

- DNA up close is just a sequence of paired nucleotides $\{A, C, G, T\}$ on a double helix.
- From afar, it is a circle, but often *knotted*.

$$\begin{tikzcd}
& \text{DNA Configurations} \arrow[dl, "\textit{forget knotting}"'] \arrow[dr, "\textit{forget sequence data}"] & \\
\text{Sequences} \arrow[loop left, "\textit{inversions, etc}"] & & \text{Topologies} \arrow[loop right, "\textit{topoisomerase}"]
\end{tikzcd}$$

- *Locally*, *inversions* are a major player in bacterial evolution.
  - These cut a segment of DNA and re-insert it with the opposite orientation.
- *Topologically*, the actions of some enzymes known as *topoisomerase* cause knotting through cutting and rejoining at crossings of the strands.
  - These have been successfully studied using *tangle algebras*, even successfully predicting distributions of knots.

Local evolution

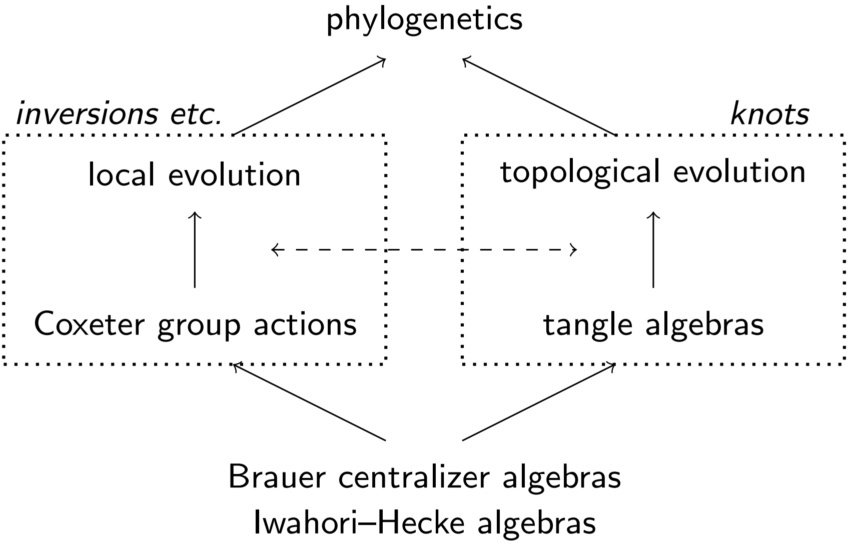
Inversions and Weyl groups

- Inversions are not just permutations of regions of DNA.
- Because of the orientation on DNA nucleotides, they are *signed* permutations, and hence can be thought of as the action of a type B Weyl group.
- Would like to better understand evolutionary distance based on inversions, and reconstruct phylogenies based on inversions.
- Incorporating other evolutionary processes such as deletion gives rise to other algebraic models involving monoids.

Topological evolution

Hecke algebras in biology?

- Tangle algebras, used to study knotting in DNA, are closely related to the family of diagram algebras including the Temperley-Lieb algebras and the Brauer centralizer algebras.
- These are connected to braid group algebras and Iwahori-Hecke algebras.



phylogenetics

inversions etc.

local evolution

Coxeter group actions

knots

topological evolution

tangle algebras

Brauer centralizer algebras
Iwahori–Hecke algebras

Reflections

Pure and applied mathematics

- What distinguishes “pure” and “applied” mathematics?
  - Motivation?
- “Applied mathematics is mathematics that is applied”

  (Jacqui Ramagge, Wollongong)
  - She also said “I find these distinctions frustrating and divisive”.
- My fellowship was listed by the ARC as “pure”. I thought it was applied.
- Is there a distinction between what one is doing, and what one’s aim is?
  - If one solves a mathematical problem that is motivated by an application, is that solution pure or applied?
    (Does it have to be “of independent interest” to be pure?)

Reflections

Why do we do mathematical research?

- Why research?
  - Curiosity?
  - Desire to prove oneself?
  - To be significant or make a significant contribution?
  - To be immortal?
  - To solve problems (for the satisfaction)?
  - To fulfil an obligation to make use of our abilities?
- "I want to find neat math problems"

  (Seth Sullivant 2008, Harvard/NC State).
  - I wanted to find problems of real importance to biology.
- I reflect on my move to mathematical biology.
  - Does the Universe miss my contribution to algebra?
  - Or poetry? Politics?
- Aside from those considerations, if one doesn't feel a hunger to solve something, one is unlikely to succeed, or be satisfied.

Thank you

More questions than answers. . .