Sadie Witkowski [SW]: Personal pet peeve- Did you know that Darwin never actually said, “survival of the fittest”?

IM: Actually, secretly I did! But that’s because I recently read“A short history of nearly everything” by Bill Bryson and he mentioned it!

SW: Yeah! Apparently it was from Spencer Herbert in his book, “The Principles of Biology” in 1864. He was actually arguing against Darwin’s thesis that survival is about being the most adaptable to the environment.

IM: Ugh, society has been duped!

SW: Yeah, Darwin’s work on the finches of the Galapagos Islands was much more interested in how species differentiate and fill niches to take advantage of those resources.

IM: So it’s not about getting a head on competition. It’s really about taking a different approach from others to carve out your own niche.

SW: Exactly! In fact, Darwin drew the first rough draft of an evolutionary tree in his ‘On the Origin of Species.’ Although German zoologist Ernst Haeckel drew a much more comprehensive one a few years later.

IM: Ok, but why are we talking about all this? Evolutionary trees seem pretty *rooted* in biology, not math or stats.

SW: Well, today’s guest sees mathematical analysis as a core component in building accurate evolutionary, ie phylogenetic, trees. Meet Tandy Warnow, from the University of Illinois at Urbana-Champaign, who’ll give us her insights about how to create phylogenetic trees.

Tandy Warnow [TW]: 8:01
It's not just doing analysis, it's discovering weaknesses. And then you address those weaknesses with new methods and new models, and then you get better biological discovery.

IM: I guess you could say, the process of biological discovery is always *evolving*

[laughs]

SW: Ok, but before this conversation *branches* (get it? Evolutionary branches?) into too many topics, let's introduce ourselves!

IM: Right! I'm Ian Martin

SW: And I'm Sadie Witkowski

IM: And you're listening to Carry the Two, a podcast from the Institute for Mathematical and Statistical Innovation, AKA IMSI.

SW: This is the podcast where Ian and I talk about the real world applications of mathematical and statistical research.

IM: We might seem like an odd couple to tackle these topics. Sadie is a cognitive neuroscientist and I'm a high school choir teacher! But it turns out that you don't need a degree in mathematics or statistics to enjoy learning about how they are applied to the world around us.

SW: And today, we're getting into trees! Well, phylogenetic trees at least.

IM: I know you used that word earlier, but I still don't know what it means.

SW: Phylogeny is the relationship between organisms, based on their evolutionary history. So when we talk about evolutionary or phylogenetic trees, it's just a representation of species and the ancestors they evolved from.

IM: Gotcha.

SW: So originally, these trees were drawn based on stuff like the fossil records or looking at the physical structures of the organisms.

IM: But with [dramatic voice] modern science [end voice], don't we make these trees based on genetics?

SW: I mean, yes. Ish. Here, lemme have Tandy take over.

TW: 9:02
But before you get the gene trees, you need the multiple sequence alignments. I mean, there's multiples and before you get the gene sequence alignments, you need to detect homology before you detect homology you need to do annotation. I mean, there's a long process of bioinformatics tests. And you can make mistakes at any one of those stages.

And so there's a lot of like reviewing what you did, and going back and saying, Okay, maybe we made a mistake here and changing it and then looking to see what happens. That's why it's very iterative.

SW: I'm going to head your questions off at the pass, since there's a decent amount of technical language in what Tandy just said.

IM: Thank you! [sounding relieved]

SW: So annotation is identifying the location of genes and their coding regions in the DNA. Just basically labeling where in the genetic sequence a gene's important bits are located.

Then gene sequence alignments basically means taking those genes across multiple samples and comparing them, looking for differences between the samples that can tell us something about evolution.

IM: And then homology?

SW: Homology in this case, is when two genes share a DNA sequence from a common origin or ancestor.

IM: Wait wait wait, so is Tandy saying that because of all these additional steps, evolutionary trees actually aren't settled? Or at least some of them aren't? I thought this was, like, school textbook level complete research!

SW: You would think, right? But turns out we're still trying to understand the evolutionary relationships between all sorts of organisms! That's actually where Tandy's work comes in! While she's in the department of computer science, she straddles this line between lots of fields, trying to help develop better tools.

Specifically, she works with biologists and develops new mathematical models that can take in lots of genetic information and then produce potential evolutionary trees for those specimens.

TW: 32:33
The phylogeny part of biology is very mathematical. And the people who work in this area know this, they know that there are models, they're thinking about the models, they're thinking about what you can guarantee about your method under that model. What, what they don't always think about is how bad the models are. Oh, that's, that's a that's a bigger issue.

IM: Oh fun, back to modeling, like where I belong… So how do researchers like Tandy even go about putting together a tree? Just compare a bunch of DNA?

SW: Not just a bunch of DNA, but specific sections of the genome, aka the entire map of all the DNA. And oh man, it's so much more complicated than you'd like to think…

TW: 10:42
But what a gene tree is, is just an evolutionary tree of a small genomic region. That's all it is. A small enough genomic region that you don't have any recombination. What that really just means, simply speaking, is that all of the DNA in that region evolves down one tree. The problem is, as you go from region to region to region, you actually find different trees. This is the problem.

IM: Uhhhhhhh…. I might need you to rewind a bit here. I'm feeling well and truly lost.

SW: Totally fair. I like to think that I have a solid biological science background and the complexities of this research still throw me too. Like I had to look up what recombination means to jog my memory.

IM: And google search said….?

SW: Recombination is just what it sounds like. It's a process by which chunks of DNA are broken apart and recombined to produce new combinations.

IM: Ooooo, that's the thing with parents each giving a version of a series of genes and then the two get mixed up so the next generation might have Dad's gene for brown eyes but Mom's gene for red hair.

SW: Yeah, it gets more technical than that in real life, but that's the basic premise. But before we get lost in technical jargon, how about we start with how Tandy starts her research projects?

IM: And how's that?

SW: With collaboration!

TW: 2:19
The problem of trying to understand evolutionary histories, whether it's birds, or its plants, or its other organisms, very similar questions arise in all of these cases. And yes, there's mathematics there. And it's very interesting mathematics. But in those projects, I'm generally involved with helping the biologists figure out how to analyze their data, which means actually, what kinds of of mathematical models to use to infer evolutionary trees under. So it's basically choosing models and then choosing methods for those models.

IM: So is Tandy like a hired gun? She gets called in to help with the heavy lifting of analyzing data that the biologists gather?

SW: That's what I originally thought as well. Of course, seeing as I spend 6 years in graduate school conducting my own research, I should have known better. You have to think about your analyses right at the beginning of the project so that you actually know that you're collecting the right kind of data.

IM: So she's actually in the planning room from the get-go?

TW: 3:12
It usually starts out earlier with, we'd like to figure this out, do you want to get on board with us? And then I say, Sure. And then they say, Okay, now we've got to go get the genomes, and then they spend a year getting the genome sequenced and assembled, and then they spend time analyzing the genome. So it actually starts with a social engagement. It starts with people who know each other and like each other and like working together.

SW: So even though Tandy is part of the team as the resident math-expert, they have to have these conversations about what the research goals are, really early on.

IM: I remember when it was a huge deal to sequence the human genome, aka the entire set of genes and genetic material that makes up a person. But that was a long time ago and I assume technology has advanced since then. How long do these kinds of projects take?

SW: Considering how long science takes, not even just writing up the results, it can be years before research is complete!

TW: 5:44
I mean, these things take so long. These are very long term projects, you know, you start by figuring out which species you want and why. And then you're going and get the genomes and you're getting them assembled. Sometimes you're going and trying to get new samples. I mean, it's a long, long process.

IM: So is it the data collection that takes forever? Or the mathematical modeling?

SW: Both? Just like evolution itself, there's a lot of trial and error to see what data they can actually gather and what tools will ultimately work the best. Even when you have a strong game plan at the start, there are all sorts of complications that can crop up.

IM: Such as…?

SW: Well, why don't we take a short ad break and I'll tell you when we come back.

IM: [sigh] ugh, fine.

[music starts]

[ad break for Entitled]

[music fades]

SW: So before the break, we were talking about how creating genetic trees takes a whole village (of researchers) but that there are some unique challenges. The first challenges is, well, pretty obvious

TW: 31:39
One of the problems with evolution is you can't do an experiment, you know, it's in the past is so unlike some things, you know, you can set up an experiment to test what's the function of this protein, right? But you really can't set up an experiment to say, When did birds diverge? So that's why math is so important

IM: Right, but this is't a new challenge. Like, Darwin had the same problem when he was trying to figure out which finch species diverged from the others and along which line.

SW: True, but when you're building trees based on genetics, there are even more factors to think about. For example, did you know that it isn't just species that are evolving and changing? Even within a species, not all genes are perfectly copied over from our parents. Individual genes are also evolving across generations.

TW: 11:27
So different genes have different trees, different genomic regions have different trees. And even just human, Chimp, gorilla is hard to figure out. A given gene could look like human and gorilla are siblings. Another one might be gorilla and chimp are siblings. But in the species tree, we think that human and chimp are siblings, how is that figured out? In fact, that's where you do it with mathematics, there is a mathematical model called the multi species coalescent. And with the multi species coalescent, you can predict the how different genes will have different trees. And looking at the frequency of each tree, you can actually say the true evolutionary history of the species is this.

IM: Woah, jargon alert. Multi species coalescent? Can you give me the sparknotes version? [sarcastic voice] You know, that thing my students are using instead of doing the actual reading assignment…

SW: Sure. So multi species coalescent is basically a method that compares samples of DNA from different genes in multiple species and tries to identify the true evolutionary tree of the species from these samples. Each gene comparison might make a slightly different tree, but by comparing across them and finding the most frequent relationships, you can find the true tree.

IM: Sounds..complicated. And hard.

SW: It is! Or at least, I was blown away by the complexity of it all. And we've only scratched the surface! In smaller organisms like bacteria, you get what's called horizontal gene transfer!

IM: There's an inappropriate 'horizontal gene transfer' joke somewhere in there...

SW: It's mostly seen in single celled organisms. But essentially instead of only passing genetic material from parent to offspring, genetic data can be passed on to other members of the species horizontally, ie to others in the same generation.

IM: Does that mean that bacteria can transfer genes to other living bacteria? I think I remember reading about that in that Bill Bryson book but that's wild!

SW: Yeah, basically.

IM: And I'm sure that's just suuuper easy to factor into their mathematical modeling…

TW: 33:28
The problem is, when you have horizontal gene transfer, you don't really get a tree. Okay? The Evolution looks much more web-like. But this also, this kind of problem also occurs when you have hybrids. So two organisms having viable offspring. I think lions and tigers can make ligers. Um, so they're, you know, there's lots of examples of that in plants especially. And that also creates situations where you can't get a tree. So the problem is, what do you do when you're trying to understand the evolution of your group? When there are things going on that don't fit trees? That's a really big problem. And it turns out to be harder to deal with than I expected.

SW: Oh, and one more challenge I wanted to bring up: gene duplications!

IM: There's *more*?!

SW: [salesman voice] But wait, there's more! So we have genes evolving and changing within species at different rates, then we have the passing of genetic material between cohorts through horizontal gene transfer. But we would be remiss not to mention that there's also a thing called gene duplication. Essentially, gene duplication is when the DNA machinery messes up and creates two or more copies of a gene back to back, lengthening that DNA sequence. For evolutionary processes, this redundancy allows one copy to potentially accumulate mutations that might be beneficial to the organism.

TW: 15:10
And then there's also gene duplications. And that's a big problem in plants. And then there's also hybridization, which is a big problem in plants. So basically, evolution is extremely complicated. And when we do, when we try to construct evolutionary trees, we're typically trying to ignore all that complication. And just hope that everything works

out. And it turns out in on a high level, yes, it works out, even if you make very strong simplifying assumptions, most of what you're going to come up with is going to be close to correct. But when you care about the details, it really matters how you do the analysis.

SW: So before your eyes glaze over, I just wanted to take this side adventure to show all the complexity that's required to create a more accurate understanding of how species evolve, whether they be bacteria, plants, or mammals like us!

IM: A fun reminder of the trillions of processes and events that led to you and me being human people.

SW: I know we've focused a lot on the challenges, but let's talk a bit about the mathematical tools and methods that are helping us make headway on this research.

IM: Right. So how does Tandy deal with these challenges?

SW: So the very short answer is, she builds models.

TW: 45:15
the main thing is just to realize the kind of math that that happens here is not as much, you know, equations and stuff like that. It's actually a lot of it's more about graph theory

IM: Oh we've talked about graph theory before, haven't we?

SW: Yup, that was a main focus of our first ever episode with Carrie Diaz Eaton!

IM: But I can explain it to our non-diehard fans who missed that episode. Graph theory is, obviously, the study of graphs. [laughs] But really, it's a mathematical approach to looking at relationships between variables or objects. So the variables are the nodes, or points, and the relationships are vertices, or lines drawn between them.

SW: Exactly correct! And graph theory gets more complicated with weighted vertices or directional flow, but that's the basics. So that's a major tool that folks in Tandy's line of work use. In fact you can think of evolutionary trees as a sort of graph. At least, that's how mathematicians tend to treat them.

And it makes sense, since Tandy wouldn't be interested in evolutionary trees if her work wasn't also pushing the boundaries of our mathematical knowledge.

TW: 7:03
We were trying to do what's called a species tree. And the methods that were available to do species trees, which means you're using like all the genomic data, the ones that met the requirements of the mathematics didn't have good accuracy. So there's this big difference between mathematical correctness, so to speak, and actual performance on data. And that gap between how you think something is going to perform and how it

actually performs is one of the interesting things in mathematics, it doesn't mean that the math is incorrect, it just means that math hasn't told you what you need yet. And what we had learned in that first study was that the current methods didn't work well. And we needed to design new methods.

IM: Wait, wait, wait. She's inventing new math just to deal with building evolutionary trees like the one Darwin doodled in the 1800s?

SW: Well, yeah. I mean Darwin was just making guesses based on how the finches looked. Tandy is working with waaaaaay more complex data here. Turns out, older methods worked because there was an assumption that you'd only have a small amount of data to work with.

TW: 16:07
the availability of good methods is increasing because people are trying to do this, but it's still lagging behind the data, in the sense that some of the best methods can only run on small datasets. And so getting methods that can run on your data sets can be hard. And even the best methods don't deal with all of the complexity. So it's this race between, you know, methods trying to catch up with the data and appropriate choice of models. And it just it's an interesting, it's an interesting challenge.

IM: To be honest, this wasn't the challenge I was envisioning when you said we were gonna be talking about evolution.

SW: Me neither. Turns out, you can have pretty mathematical methods that don't hold up when confronted with real-world data.

TW: 21:00
so there are there are many methods that are designed that have really, really, really beautiful theory and beautiful performance on data. But you can't run them except under small datasets. So Bayesian methods, in my experience, fall into that category. They're wonderful, but they just don't converge.

SW: Just jumping in to remind you that Bayesian methods here refers to a field of statistics. In this case, Tandy was basically using statistical methods in a computer program to put together lots of data to try to converge on the true, evolutionary tree.

TW: 21:36
And when I say that, it doesn't converge. What I mean is, you can run it for a few months, and it still won't have converged. There are. So there are data, there are methods that just have computational limitations, either memory or runtime. Okay, then there are methods that have really good performance if you have enough data. But for that, there's two dimensions to data. So if we go back to the idea of tree estimation, they're essentially there's two dimensions, there's how many leaves in your tree? Like how many species, think of it that way. And the other one is, how much genomic data do you have for each species? Okay, so two different dimensions. And the two

dimensions affect methods differently. When you have a small number of species, like just for human, Chimp, Gorilla and orangutan, but you have genome scale data. Then you have a small number of species, but a large number of sites are long sequences.

IM: All of this is kinda making my head spin.

SW: I think the big takeaway I want to make here is, that even with modern technology, we still face challenges because of the complexity in the amount of data in the genome and how genes interact. You might have what seems like a great method, but it turns out that even with modern computing, it takes *forever* for it to run, and it *still* might not converge on an answer.

IM: How long do some of these run?

SW: Well, if the program isn't properly designed for the data they have, literally the program can run forever and never come up with a solution. Some that do have solutions might still need months of computing time.

IM: Wow. So we actually need to be doing even more tool development to answer these kinds of questions of inheritance and build better models?

SW: Yeah, turns out it's not a settled field at all! And Tandy really wants to push for more and earlier collaboration between folks like her and the biologists collecting the data.

IM: Are you saying she wants them to co-evolve?

SW: [laughs] Sure.

TW: 39:47
from the perspective of a practitioner, if I were a biologist, if I were talking to a biologist, the thing that I would say is often biologists think that the problem is the data, okay, they think they have to get rid of some of their data because the data are just can't be analyzed. And I think often the problem is not the data. The problem is the methods. And what they should be doing is pushing people to develop better methods.

SW: This feels a bit like one person's trash is another's treasure. The data all being collected is so rich that previous methods couldn't handle it and researchers would just throw out all this great information. But now all the data is meaningful, assuming we can develop methods that take full advantage of it.

IM: I know we've been doing this podcast for a while, but it still surprises me when we still have unanswered math questions. Math courses are always taught as enclosed systems.

SW: Yeah, basic math classes are often taught as like, just a series of known equations that you need to plug-and-chug. But math researchers, whether they're interested in applied or theoretical work, are still constantly pushing the boundaries.

TW: 40:52
From a method developer perspective, the gap between what we think we know from math, and what we actually see is big. And it's interesting. And even if all you are is a pure mathematician, that gap has interesting questions. And if you're an empiricist, or someone who cares about like method development, not just understanding mathematical properties, there's really interesting stuff to do there. So it's an interesting research area, the biologists should not take, take us for granted that the methods are adequate, they should not. They should push, they should push the method developers to get on board. And the method developers should appreciate the fact that future datasets will be more complex, and they will be bigger, and prepare for that complexity prepare for that data set size.

IM: Yeah, what she said. But seriously, I don't think I realized the mathematical work required just to create the tree diagram of when chimps and gorillas diverged.

SW: Me neither, but we really need these tools to understand evolution.

TW: 28:50
So many things you can understand in biology by understanding evolution, you know, even just understanding like how species adapt to their environments, you can't do that. Without understanding the evolutionary context. You can't understand how they co evolve how genes interact with each other, you can't understand anything without an evolutionary perspective. So there are lots of reasons that biologists are interested in evolutionary trees. Even timing the dates at which things happen, like when did humans leave Africa? You know, when did these things happen? Right, those questions you can't answer without very good understanding about evolution. So evolution is a fundamental thing for biologists. And I'm fascinated by the questions about how hard it is to do those estimations.

IM: So then does Tandy consider herself to be more on the applied side of things?

SW: I think she's more in between applied and theoretical, according to her own words.

TW: 28:21
So it's like it's a spectrum. And I'm somewhere in the middle. I'm not purely on the math side, and not purely on the biology side. But I'm more driven by the methods than I am by the biology. So if I were really a biologist, I would actually really care what the answer was, and I don't actually care.

[Laughs]

IM: Well, that's refreshingly honest!

SW: It seems that even if Tandy's work is mostly applied, she definitely falls in the "loves math for math's sake" category.

TW: 0:33
[Sadie] So what got you interested in mathematical research to begin with?
TW: Believe it or not, I found it beautiful. For me, what was wonderful about math was it's just elegance, its clarity, its perfection. It was just for me something about beauty. It was not about its utility. It was not about its relationship to science. It was just how pretty it was.

IM: Oh I love that. There is something beautiful and satisfying about the clarity of math. Like when you get a perfectly fitting tetris piece. And, it's always a plus to find more beauty in the world.

SW: But to recap, it turns out we need to continue pushing method development in mathematics in order to better build accurate phylogenetic trees.

IM: And these new methods can hopefully help deal with the complexities of genetic data, like horizontal gene transfer or gene duplication.

SW: Or how not all genes within a species evolve at the same rate! That's still blowing my mind.

IM: And so, we have collaborations between biologists and folks like Tandy to thank, when it comes to creating these trees.

SW: Buuuuuut, as data sets become more rich or complex, we need conversations between mathematicians and biologists to happen earlier in the process so that they can keep developing new methods to answer evolutionary questions.

FINAL QUOTE
TW: 16:52
What keeps me going is how interesting it is. So I now see big gaps between what math might suggest and what actually happens in practice. And I can approach that from a purely mathematical perspective and be fascinated as a mathematician. I can approach it as an empiricist. Who cares about method development, whether or not I have theorems. So mathematicians tend to want theorems. And I look at the theorems, and I say, that's very pretty. But what does it actually tell me? And what it actually tells me is not everything I need. And so then I have to go and think about it empirically. So it's it's back and forth between theory and empiricism

[Music starts]

SW: And don’t forget to check out our show notes in the podcast description for more on evolutionary research and further resources on the mathematics and statistics behind building phylogenetic trees.

IM: And if you like the show, give us a review on apple podcast or spotify or wherever you listen. By rating and reviewing the show, you really help us spread the word about Carry the Two so that other listeners can discover us.

SW: And for more on the math research being shared at IMSI, be sure to check us out online at our homepage: IMSI dot institute. We’re also on twitter at IMSI underscore institute, as well as instagram at IMSI dot institute! That’s IMSI, spelled I M S I.

IM: And do you have a burning math question? Maybe you have an idea for a story on how mathematics and statistics connect with the world around us. Send us an email with your idea!

SW: You can send your feedback, ideas, and more to sadiewit AT IMSI dot institute. That’s S A D I E W I T at I M S I dot institute.

IM: We’d also like to thank our audio engineer, Tyler Damme for his production on the show. And music is from Blue Dot Sessions.

SW: Lastly, Carry the Two is made possible by the Institute for Mathematical and Statistical Innovation, located on the gorgeous campus of the University of Chicago. We are supported by the National Science Foundation and the University of Chicago.