1. It's a pleasure to be giving my second talk at the CSNTM! The first time was two years ago, when I led a Q&A with interns about the CBGM. This year's talk is more of a lecture. I will get into the CBGM in the second half, but the first half will be an introduction to the related subject of phylogenetics, which is the methodology I'm using for my PhD research. [0:25]

2. So let's jump right in! We'll start with some basic concepts that are common to both methods. [0:05]

3. When a textual tradition consists of multiple surviving witnesses, like manuscripts, the first thing we need is a basis for comparing them. The traditional approach is to line up their texts in a *collation*, highlighting their *variant readings* in segments of text called *variation units*. Here you can see Reuben Swanson's collation of Luke, with the variation units highlighted. The witnesses are listed in the right margin. These units with numbers in parentheses are *constant*, meaning that all the witnesses agree at them. They're usually not included in analyses, but they can help us estimate how often scribes and readers *didn't* change the text. [0:45]

4. Powerful techniques from biology can be applied to textual criticism thanks to the insight that textual collations are analogous to DNA sequence alignments. The rows for species or taxa correspond to witnesses, the columns for sites or characteristics correspond to variation units, and the cells containing states like DNA nucleobases and gaps correspond to variant readings, including omissions. As with missing DNA data, lacunae and things like retroversions from other languages can be encoded as fully partially ambiguous states. [0:45]

5. This means that witnesses like manuscripts, at the most basic level, are just rows of the collation, or sequences of readings. But paratextual features of the physical artifacts themselves could be encoded as states in a similar way, and as I will explain later, the age ranges for witness can be incorporated in more complex phylogenetic methods. [0:30]

6. But we'll get to that later. For now, I'll start by explaining how phylogenetics can take witnesses encoded in this way and evaluate hypotheses about where they belong in textual history. [0:15]

7. The underlying model for textual history in phylogenetics is a family tree of the textual witnesses, commonly called a *stemma*. The family tree, as you'll notice, is upside-down. This is because unlike people, who typically have two parents, a manuscript is assumed to be copied primarily from one source. The surviving witnesses to the text are located at the bottom of the tree, at the "leaves," which are depicted here as solid circles with Latin letters. These witnesses are traced back along their branches to common ancestors that are no longer extant; these are referred to as "hypothetical ancestors" or *hyparchetypes*, and they are portrayed here as dashed circles with Greek letters. The appeal of phylogenetics is that it can reconstruct the readings of these lost ancestors. The earliest common ancestor of all the surviving witnesses is called the *archetype*, and it is often denoted by a lowercase omega. It represents the earliest text that can be reconstructed from the surviving part of the tradition. If the tradition is well-sampled by its surviving witnesses, then this could correspond to the text as it entered circulation. But if entire early branches of the tradition have gone extinct, then this authorial text will correspond to a separate ancestor called the *root* of the stemma, denoted by a capital omega, and we may have to reconstruct it by conjecturally emending the text of the archetype. [01:30]

8. In short, a stemma represents a hypothesis about how witnesses are related and how close they are to the earliest text. The problem is that many different hypotheses are possible. In the examples here, the witnesses b_1 and b_2 could be siblings with equal textual value, or b_2 alone could be a significant witness to a very early text. Our goal is to determine which hypotheses best explain the collation data that we have. And to do this, we need a way to measure how good a candidate stemma is. [0:30]

9. The traditional metric used in phylogenetic textual criticism is *parsimony*. Under this metric, the cost of a stemma is the smallest number of changed readings along its branches. This is basically the same idea as what the CBGM calls *coherence*. Both ideas are based on Ockham's Razor: under the assumption that scribes copied faithfully more often than they erred or innovated, the hypothesis that requires the fewest violations of this assumption is best. We can conveniently calculate this cost for a given stemma at each variation unit and get the total cost by adding up these individual costs. Even more conveniently, there's an efficient way to do this by starting from the leaves of the stemma and working our way up to the top.

For a demonstration of how this works, I'll consider a single variation unit where five extant witnesses attest to common variations on a common name. [1:00]

10. The process of calculating a given stemma's cost at a given variation unit consists of a forward pass and a backward pass. We only work up to the archetype, as this is the earliest ancestor whose reading can be reconstructed. We start with the lowest witnesses, c_1 and c_2. They have different readings, so we say that their common ancestor γ could have either of these readings. Next, we have the extant witness a and the hyparchetype γ: a reads χριστου ιησου, and γ also has this as one of its potential readings, so their common ancestor α reads χριστου ιησου. On the same level, we also have b_1 and b_2: they agree on the reading ιησου χριστου, so their common ancestor β also has this reading. That means that we have a tie at the archetype: it could have either of these readings. [Click to next slide]

The backward pass starts from the archetype and goes down to the leaves. We can pick any reading that the archetype could have based on the forward pass. If we choose χριστου ιησου and then resolve the ambiguities from the forward pass as we go down the branches, we get these ancestral readings. The stemma therefore features two changes in reading, resulting in a cost of 2. [Click to next slide]

If instead we choose ιησου χριστου as the reading of the archetype, then we get these ancestral readings, but the cost of the stemma remains the same. [1:30]

11. Now let's see what cost we get for a different stemma. We'll work through the forward pass again. Like before, c_1 and c_2 have different readings, so their ancestor γ could have either of these readings. At the next level, we see that γ and b_1 also have no readings in common, so their ancestor β could have one of three readings. We finally can break the tie once we get to the next level: a and β have the reading χριστου ιησου in common, so we set this as the reading of their ancestor α. Finally, we have another tie at the hyparchetype between the reading of α and the reading of b_2. [Click to next slide]

We can start with either reading in the backward pass. If we start with χριστου ιησου, then we get these ancestral readings down the branches, with three changes of reading in total. [Click to next slide]

And if we start with ιησου χριστου, then we get these ancestral readings, also with a cost of 3. [1:00]

12. In fact, it's been proven that if the cost of a stemma is calculated this way, it will be the same no matter how we break ties at the archetype. Even more importantly, the cost of the stemma will be the same *no matter where the archetype and root are located*! This handy fact has enabled textual critics to separate their work into an automated part and a human part: the computer can calculate the costs of millions of stemmata without needing to know where their roots are, and afterwards, the textual critic can identify where the root belongs based on human judgments about the internal evidence of readings. This, in fact, is how computer-assisted textual criticism has traditionally been done. [0:45]

13. But what if we want to incorporate internal evidence of readings into the calculation of a stemma's cost? One benefit of doing this is that we get a clean separation of concerns between different types of internal evidence. There are two main types of internal evidence: intrinsic probabilities, which concern what the author most likely wrote, and transcriptional probabilities, which concern what later scribes were most likely to do. If we just lump these two types of evidence together to determine where the root of the stemma is, it's not clear how we're supposed to decide between them when they point in opposite directions. But if we introduce intrinsic and transcriptional evidence earlier in the process, then we can use them in distinct ways. With intrinsic evidence, we can include the root in the stemma (like we have in this picture) and specify which reading or readings it could have. This will affect how we calculate the cost of a stemma in the backward pass. With transcriptional evidence, some types of scribal changes—like skips of the eye, confusions involving similar sounds, and harmonizations—may be more likely than others, and a change from one reading to another may not be as easy to make in the opposite direction. We'd like for this to figure into our calculation of a stemma's cost.

The cool thing is that we can extend the traditional approach to do this. The main trade-off, of course, is that the cost of the stemma will now depend on where its root is. [1:30]

14. Let's start by incorporating just intrinsic evidence. We'll pretend that we're being particularly bold and we've conjectured that the author originally wrote ιησου, a reading not attested in the surviving tradition. Then this would be the reading at the root, and the forward pass up to the archetype would go as it normally would. [Click to next slide]

But in the backward pass, we'll have an extra change of reading, and there are multiple ways to explain it. One possibility is that an early scribe expanded the author's ιησου to χριστου ιησου, and the scribe behind the β branch then transposed this to ιησου χριστου. [Click to next slide]

Another possibility is that an early scribe expanded ιησου to ιησου χριστου, which the scribe behind the α branch transposed to χριστου ιησου. [Click to the next slide]

The third possibility is probably the most transcriptionally plausible: two branches of the tradition independently expanded ιησου in different ways, with the α branch adding χριστου before it and the β branch adding χριστου after it. [1:15]

15. We can compare scenarios like these in a quantifiable way if we assign different costs to different types of changes between readings. It's helpful to think of this in terms of a graph like the one on the right: the nodes are the readings, and the edges between them correspond to changes with different costs. It's a way of modeling the transcriptional behavior of the average scribe. The costs on the edges can be conveniently encoded in a table called a *cost matrix*, which is pictured to the left here. The starting readings are along the rows, and the resulting readings are along the columns, so the cost of starting at one reading and changing to another is located at the cell in that row and column. Naturally, the cost of copying the same reading faithfully is just 0. A skip of the eye, which would explain the loss of χριστου after ιησου or the loss of ιησου after χριστου, would be very common, so we'll assign it a low cost, say, 2. An expansion of ιησου or χριστου to ιησου χριστου or χριστου ιησου would be almost as common, so we'll assign it a slightly higher cost at 3. Substitutions between ιησου and χριστου and transpositions between ιησου χριστου and χριστου ιησου are even less common, so we'll assign all of them a cost of 5. Finally, for the changes we can't explain, like the shortening of χριστου ιησου to ιησου and the shortening of ιησου χριστου to χριστου, we'll assign a high cost, like 10. [1:30]

16. We can then use the cost matrix to calculate the cost of a stemma while taking all transcriptional possibilities into consideration. This is a bit more involved than the forward and backward passes we did earlier, so I'll walk you through this one from the start. The first thing you'll notice is that we don't just keep track of which reading or readings the witnesses and their ancestors have. This time, for every witness or ancestor, we keep track of its lowest total cost for every reading it could have. Thus, every witness or ancestor has four entries for costs—one for each of the four variant readings in this unit. Every extant witness has a cost of 0 for its known reading and infinity for every reading it is known not to have.

So let's figure out what the cost entries for γ should be based on its children, c_1 and c_2. If γ reads ιησου, the first reading, then its minimum cost would be the cost of ιησου changing to χριστου in c_1 plus the cost of ιησου being expanded to χριστου ιησου in c_2. Checking the cost matrix, we see that these costs are 5 and 3, respectively. We would then add these values to the costs of c_1 reading χριστου and c_2 reading χριστου ιησου, but since those values are 0 they don't change anything. Adding up the transition costs of 5 and 3 from the matrix gives us a total cost of 8 for γ reading ιησου. [Click to next slide]

What if γ reads χριστου ιησου? Then χριστου ιησου is shortened to χριστου in c_1, which has a cost of 2, and χριστου ιησου stays the same in c_2, which has a cost of 0. So the total cost of γ reading χριστου ιησου is 2. [Click to next slide]

Meanwhile, if γ reads ιησου χριστου, then ιησου χριστου gets shortened to χριστου in c_1, which has a cost of 10, and ιησου χριστου gets transposed to χριστου ιησου in c_2, which has a cost of 5. So the total cost of γ reading ιησου χριστου is 15. [Click to next slide]

And if γ reads χριστου, then χριστου stays the same in c_1, which has a cost of 0, and it gets expanded to χριστου ιησου in c_2, which has a cost of 3. So the total cost of γ reading χριστου is 3. [2:45]

17. Now let's go up a level and calculate the costs for α. If α reads ιησου, then we have a cost of 3 for it expanding to χριστου ιησου in witness a, plus 0, since a has a cost of 0 for that reading. But what about the cost that γ contributes? To calculate this, we have to consider the cost of each reading that γ could have, plus the cost of ιησου changing to that reading, and we pick the smallest of these costs. So, if γ reads ιησου, then we add its cost of 8 to the cost of 0 for α's reading not changing. If γ reads χριστου ιησου, then we add its cost of 2 to the cost of 3 for ιησου being expanded to χριστου ιησου. If γ reads ιησου χριστου, then we add its cost of 15 to the cost of 3 for ιησου being expanded to ιησου χριστου. And if γ reads χριστου, then we add its cost of 3 to the cost of 5 for ιησου being changed to χριστου. The lowest of these costs is 5, which occurs if γ reads χριστου ιησου. So we add this cost of 5 to the cost of 3 for ιησου changing to χριστου ιησου in a, and we get a total cost of 8.

Now suppose that α reads χριστου ιησου. Then it would have a total cost of 0 along the branch to a, because their readings would agree. That just leaves the cost along the branch to γ. If γ reads ιησου, then we add its cost of 8 to the cost of 10 for χριστου ιησου changing to ιησου. If γ reads χριστου ιησου, then we add its cost of 2 to the cost of 0 for χριστου ιησου remaining unchanged. If γ reads ιησου χριστου, then we add its cost of 15 to the cost of 5 for χριστου ιησου being transposed to ιησου χριστου. And if γ reads χριστου, then we add its cost of 3 to the cost of 2 for χριστου ιησου being shortened to χριστου. The smallest of these costs is 2, so the cost of α reading χριστου ιησου is 0 + 2, or 2.

Now suppose that α reads ιησου χριστου. Then it would have a total cost of 5 along the branch to a for the transposition of ιησου χριστου to χριστου ιησου. Let's see what the cost along the branch to γ would be. If γ reads ιησου, then we add its cost of 8 to the cost of 2 for ιησου χριστου losing its second word, for a total of 10. If γ reads χριστου ιησου, then we add its cost of 2 to the cost of 5 for ιησου χριστου being transposed to it, for a total of 7. If γ reads ιησου χριστου, then we add its cost of 15 to a cost of 0 for ιησου χριστου remaining unchanged, for a total of 15. And if γ reads χριστου, then we add its cost of 3 to a cost of 10 for ιησου χριστου being changed to χριστου, for a total of 13. The lowest combined cost along this branch, then, is 7. Once we add this to the cost of 5 for the branch to a, we get that the total cost of α reading ιησου χριστου is 12.

Finally, let's suppose α reads χριστου. Then it would have a total cost of 3 along the branch to a for the expansion of χριστου to χριστου ιησου. We then calculate the cost along the branch to γ in the usual way. If γ reads ιησου, then we add its cost of 8 to the cost of 5 for χριστου being changed to it, for a total of 13. If γ reads χριστου ιησου, then we add its cost of 2 to the cost of 3 for χριστου being expanded to it, for a total of 5. If γ reads ιησου χριστου, then we add its cost of 15 to a cost of 3 for χριστου being expanded to ιησου χριστου, for a total of 18. And if γ reads χριστου, then we add its cost of 3 to a cost of 0 for χριστου remaining unchanged, for a total of 3. This last total of 3 turns out to be the lowest cost for the branch to γ. Once we add this to the cost of 3 for the branch to a, we get that the total cost of α reading χριστου is 3 + 3, or 6.

And that's how it's done! If you'd like to get a better grasp of the procedure, I would encourage you to work it out for yourself in the other parts of the stemma. It takes a lot of small calculations, but thankfully, it's stuff that's easy for computers to do. [5:00]

18. If, on the basis of intrinsic evidence, we narrow the reading of the root down to one or more options, then can we use the same process to determine the cost of the stemma from the root. So if the root reading is ιησου, then the smallest cost is achieved if the archetype also reads ιησου: the cost of the archetype reading ιησου is 8, and the cost of the reading staying the same from the root to the archetype is 0, resulting in a total cost of 8.

But, if the root reading is χριστου ιησου, then the smallest cost is achieved if the archetype also reads χριστου ιησου: the cost of the archetype reading χριστου ιησου is 7, and the cost of the reading staying the same is 0, resulting in a total cost of 7.

This also means that if we were undecided between these two readings for the authorial text, χριστου ιησου would be preferable, because it results in a slightly lower cost for the stemma. [1:00]

19. The use of variable costs is in fact a stepping stone to an even more substantial improvement: we could incorporate intrinsic and transcriptional evidence as probabilities between 0 and 1, and in this way, we could evaluate stemmata based on their probabilities rather than their costs. Based on intrinsic evidence, the root could have different probabilities for different readings based on how well they fit the author's argument and style. Likewise, transitions between different readings could be assigned probabilities based on how likely scribes were to make certain types of changes. Probabilistic models of states and transitions between them are well-understood in mathematics, where they're called *Markov chains*. We don't have time to get into the technical details of this, but it should suffice to say that phylogenetic methods that work with probabilities have been developed and used successfully in biology and in textual criticism, including that of the New Testament. [1:00]

20. As a final note on phylogenetics, it is worth mentioning that when we work in probabilities, we can estimate various other parameters of interest while we evaluate stemmata. If we don't know how likely different types of scribal changes are up front, we can estimate their probabilities as parameters in our transmissional model. This means that we can estimate average scribal habits automatically. Here you can see a histogram approximating how the rates for common classes of scribal changes are distributed. We can also use branches of varying lengths to model more copying events or copyists who made more errors on average. We can also use *clock models* so that these branch lengths can reflect durations of time. Using clock models allow us to include information about the date ranges of witnesses, too, which is something traditional phylogenetics and other methods can't do. Perhaps most importantly, working with probabilities allows us to quantify how certain we can be about different hypotheses about the history of the text. While working with all of these additional parameters is much more complex than the basic forward and backward pass I described at the start, it makes for a much more robust model, and modern computers are powerful enough to handle the complexity. [1:30]

[24:30 for the phylogenetics half]

21. That brings us to the second half of today's lecture, which concerns the Coherence-Based Genealogical Method, or CBGM. Once again, we'll start with the basic ideas behind the method, and then we'll take a closer look under the hood. As we do this, we'll note comparisons to phylogenetics where they come up. [0:15]

22. The CBGM was developed over the course of two decades by Gerd Mink. One of his goals was to find a way to manage the problem of *contamination*, or mixture of sources. The idea is illustrated here: the witness b_1 inherits different readings from the two β ancestors in different branches. This is common in the tradition of the New Testament. In fact, New Testament textual critics have considered phylogenetic methods unusable for this very reason. [0:30]

23. Gerd Mink based the CBGM on four fundamental principles or operating assumptions about textual transmission. The first one here is just the assumption that faithful copying is more likely than error or innovation, as I mentioned a while back. The second basically states that variant readings that aren't original don't arise out of thin air; they have to be explained from other readings. The third is that scribes typically used fewer sources rather than many. Intuitively, we'd expect most scribes to do what is easiest and copy from just one exemplar. There were surely readers and emendators who did the extra work of consulting other sources or making corrections against them, but they were the exception rather than the rule. The fourth is that scribes typically used closely related sources rather than distant ones. This would also require less effort, and most manuscripts in the same scriptorium or monastery were likely of a similar character. Certainly, manuscripts of diverse geographic origin and textual character traveled great distances and sometimes did end up together, but again, this was more exceptional than normal. All of these assumptions are also suitable for phylogenetics, although the last two don't come into play as much, because traditional phylogenetics doesn't model contamination.

There are a couple other functional assumptions that the CBGM makes. One is that witnesses represent abstract states of the text rather than physical artifacts bound by age and other features. The CBGM must stress this point because it sometimes has to treat younger witnesses as ancestors to older ones. The point is that the relationship of ancestry is not between two manuscripts—which would be impossible due to their ages—but between a "better" text and a "worse" text.

Such situations can arise because of a more crucial assumption the CBGM makes—namely, that there are no hypothetical ancestors allowed. The only exception is the initial text, also called the *Ausgangstext*. In phylogenetic terms, the CBGM doesn't work with hyparchetypes; it just keeps the earliest text at the root of the stemma and uses the extant witnesses at the leaves as substitutes for the lost parts in the middle. The justification for this is that contamination makes the construction of a stemma, including hyparchetypes, impossible. But this comes with a functional advantage: the problem becomes simpler, because we now only have to relate witnesses whose readings are fully known. [3:00]

24. In the CBGM, relationships between readings in a variation unit are encoded in a structure called the *local stemma*. The nodes represent readings, and an edge from one to another indicates that the former reading gave rise to the latter. In describing relationships between witnesses, local stemmata function in the same way that cost graphs and Markov chains do. But in principle, they represent something different. A cost graph or Markov chain represents what scribes *might* do, while a local stemma reflects a text-critical judgment of what they *did* do. This is why the local stemma looks a lot more like a standard stemma, with branches that point in one direction only and don't form cycles. [0:45]

25. Using local stemmata, we can calculate the overall relationship that two witnesses have. At every variation unit where both witnesses are extant, their readings can have one of five relationships in the corresponding local stemma. They can agree, as they do in this first unit. One reading can be prior to the other, if there is a path in the local stemma from it to the other. So in this second unit, W1's reading is prior to W2's, and in this third unit, W2's is prior to W1's. The two readings could have risen independently from some other reading, in which case they have no relationship. And finally, if we don't know where one reading fits in the local stemma, then the relationship between it and other readings is unclear. This last type of relationship is really only tentative, as every local stemma should account for all of its readings once textual critics have revised it.

Anyway, we tally up these relationships at all of the variation units to get the overall genealogical relationship between two witnesses. For the purposes of the CBGM, the first three types of relationships—agreements, prior readings, and posterior readings—are the most important. [1:30]

26. If we do this for every pair of witnesses, then we can establish a convenient order between them. If one witness has more prior readings relative to another, then it is considered a *potential ancestor* to that witness. This is a one-way relationship, so witnesses with more posterior readings will have more potential ancestors, while witnesses with more prior readings will tend to have fewer potential ancestors. [0:30]

27. This allows us to describe relationships between witnesses like we do for readings in the local stemma. The structure the CBGM uses to do this is called a *textual flow diagram*, and it's the main tool that the method offers for textual critics to evaluate and revise their judgments about how variant readings arose. The diagram is a tree with as few changed readings along its branches as possible, so it's like a most-parsimonious stemma. But it's primarily concerned with readings at a specific variation unit.

Here's how we build it. We start by specifying a *connectivity limit*, denoted κ. This is just a way of saying how distantly related a witness's ancestor can be in this diagram. Then, for every witness, we take three steps. First, we list out its potential ancestors, starting with those that agree with it most. If any of the closest κ has the same reading in this unit as the witness in question, then we pick the first one that does as its textual flow ancestor. If none of them has the same reading as the witness in question, then we pick the closest extant potential ancestor. The idea is that we want as few changes as possible within the connectivity limit. Where this isn't possible, we have branches where the reading changes, which are highlighted in blue.

This allows us to check and revise the local stemma for the specified variation unit. Since we know the readings of the textual flow ancestors and descendants, the blue edges can tell us which readings might have given rise to others and how many times they might have done so. In short, *general* relationships between witnesses shed light on *specific* relationships between readings. [2:00]

28. It's important to note here that despite their similarities to stemmata, textual flow diagrams aren't actually the CBGM's counterpart to stemmata of witnesses. For that, the CBGM has a structure called the *global stemma*, which relates the witnesses in a way that accounts for contamination. But before we can get to that, we have to understand a couple other things first.

The first thing is the concept of "explained readings." In the CBGM, one reading in a local stemma is said to *explain* another if one of two things happens. The readings could agree, if which case the explanation is by agreement. In other words, the ancestor's reading was just copied faithfully in the descendent. Alternatively, the first reading could have an edge pointing to the other, in which case we say the explanation is by descent: the reading in the ancestor was changed to the reading in the descendant. We extend this idea to witnesses by comparing them at all their shared variation units. In this example, GA 1832 explains 2243 by descent the first two times and by agreement the third time. In the fourth unit, 1832 doesn't explain 2243, because 2243's reading is prior to its reading. And in the fifth unit, 1832 doesn't explain 2243, because their readings aren't related. Explanation is only possible where both witnesses are extant, so lacunae don't have to be explained, and they can't explain readings. [1:30]

29. The other thing we have to understand is the idea of a witness's *substemma*. This refers to the part of the global stemma that concerns that witness. A witness's substemma connects it to one or more of its potential ancestors so that each of its readings is explained by one of their readings. A witness's substemma is valid if all of its readings are "covered" by its stemmatic ancestors, as this figure illustrates. [0:30]

30. In practice, most witnesses will have more than one valid substemma, so we need a way to choose the best one. Two of the CBGM's operating principles can help us here. Principle 3 suggests that we should prefer substemmata with fewer ancestors, while principle 4 suggests that we should prefer substemmata with ancestors closer to the target witness. As in phylogenetics, it's helpful to have a cost function that implements these criteria in a quantifiable way. [0:30]

31. A simple option is just to count the units where an ancestor explains the target witness by descent and not by agreement. This naturally favors closer ancestors over more distant ones. But since we apply it to every ancestor separately and then add up the costs for the whole substemma, it also favors substemmata with fewer ancestors, because they have fewer sources of additional costs. This example shows how we get the cost of GA 1832 as a stemmatic ancestor to 2243. The cost is 2 because 1832's reading is prior to 2243's reading in the first two units. Note that the only units that count towards the cost are the ones where 1832 explains 2243; those with red *x* marks are places where 1832 can't explain 2243, and so another ancestor would be needed to explain 2243 there. [1:00]

32. Now that we have a way to measure the cost of a substemma, we can use it to identify the best candidates. This process is called *substemma optimization*. The example here illustrates a brute-force approach. The substemmata in red aren't valid because they don't cover all of the target witness's readings. But if we checked all the substemmata, we would eventually find that the one highlighted in green, consisting of the potential ancestors B and C together, is the lowest-cost option that is also valid. This task is equivalent to a well-known problem in computer science called the *weighted set cover* problem. It's possible that in theory, we cannot do substantially better than the brute-forcing approach of checking all of the exponentially many candidate solutions. But thankfully, in practice, heuristics that are guaranteed to find the optimal solution tend to solve the problem in seconds. [1:00]

33. We're now ready to get into the global stemma. The global stemma is the result of linking together all of the substemmata we find for the witnesses. Like a phylogenetic stemma, it offers a big-picture view that accounts for all textual relationships between all witnesses. It is meant to be the final step of the CBGM workflow, although to my knowledge, the INTF has not yet generated a complete global stemma for any book of the New Testament. [0:30]

34. Using software that I developed, I was able to produce a complete global stemma for 3 John using the collation data of the *Editio Critica Maior*. I've included it here, although scaling it to fit the page might have made it hard to read.

As I conclude this section of the talk, it's helpful to compare the global stemma and a phylogenetic stemma at a few points. The first point is that in practice, a global stemma can be constructed much more quickly. With parallel processing, we can build one for over 500 witnesses in less than fifteen minutes. By comparison, a satisfactory phylogenetic stemma search with 200 witnesses can take thirty hours or more. A second and more obvious point of comparison is that witnesses in the global stemma can and often do have multiple ancestors, while this is not allowed in standard phylogenetic approaches. As I've explained, the CBGM uses this feature to account for contamination. But hypothetical ancestors aren't allowed in the CBGM, and multiple ancestors are sometimes its only way to deal with missing ancestral data, so we have to be careful in how we interpret them. And as I've also noted, the CBGM divorces texts from the manuscripts that bear them, so texts in the CBGM can have relationships that contradict known historical relationships between their manuscripts. These issues have raised the important question of whether the CBGM's global stemma can even model textual history like a phylogenetic stemma can. [1:45]

35. We're nearly out of time, so I'm going to wrap this up with some concluding considerations. [0:05]

36. We’ve reviewed how phylogenetics and the CBGM work, and we’ve touched on their strengths and weaknesses. In light of what we’ve learned about them both, we might ask ourselves, can we improve either approach with techniques from the other to get the best of both worlds? The way I see it, straightforward improvements can be made on either front.

The CBGM, for its part, does not have to constrain local stemmata to avoid cycles. It could in principle use cost graphs instead, and this would better accommodate uncertainty about how some variant readings arose. The determination of potential ancestors would probably have to be made using total costs rather than counts of prior and posterior readings, but otherwise, it would work as it usually does.

Phylogenetics, for its part, could benefit from the same principle that the CBGM uses to deal with contamination. The *local-genealogical principle*, developed by Barbara and Kurt Aland, states that we should consider the history of each variation unit independently of the other variation units. This is why we construct a different local stemma and textual flow diagram for each variation unit, and it is why a witness’s substemma has to explain every one of its readings. Phylogenetics can do the same thing with stemmata of witnesses. It already calculates the cost of a stemma one variation unit at a time, so there nothing in principle that prevents us from combining different stemmata that are best for different variation units, as this image illustrates. In fact, this idea has a biological precursor in *gene trees*. In biological settings, it has long been known that basic phylogenetics can be augmented with techniques for modeling *horizontal gene transfer*, which is the biological equivalent of contamination.

An important takeaway from all of this is that advances in biology and textual criticism have historically inspired developments in the other field and can continue to do so. The CBGM is advertised as an approach to deal with contamination, but it is certainly not the first, and it may not be any better than techniques that have already been developed. Ultimately, with respect to what has already been accomplished, we might only find a better path forward by first looking backwards. [2:30]