Phylogenetics and the CBGM @CSNTM

Center for the Study of New Testament Manuscripts

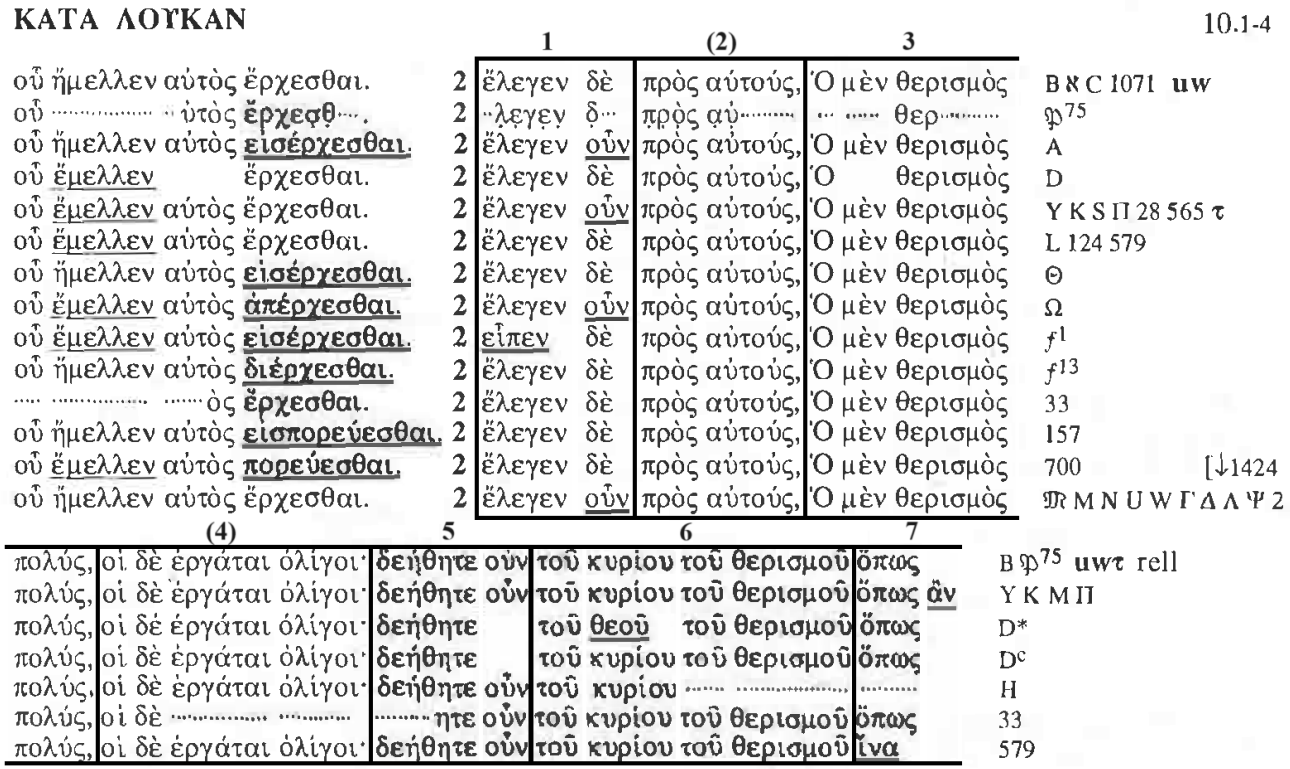
12 February 2024

Joey McCollum
Australian Catholic University
Institute for Religion and Critical Inquiry
james.mccollum@myacu.edu.au
@JoeyMcCollum
jjmccollum

ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

Preliminaries

- To compare textual witnesses, align them at independent *variation units*
- *Variant readings* occur at variation units



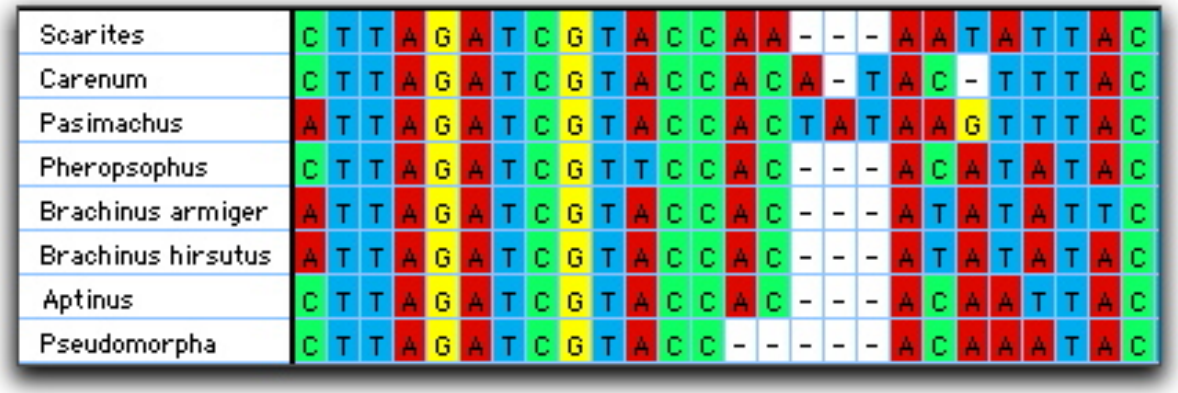
ΚΑΤΑ ΛΟΥΚΑΝ 10.1-4

		1		(2)	3	
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	B ℵ C 1071 **uw**
οὗ ..... ὐτὸς ἔρχεφ̣θ̣...	2	·λ̣ε̣γ̣ε̣ν	δ̣...	πρ̣ὸ̣ς̣ α̣ὐ̣.....	... . θερ.....	𝔓75
οὗ ἤμελλεν αὐτὸς εἰσέρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	A
οὗ ἔμελλεν ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ θερισμὸς	D
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	Y K S Π 28 565 τ
οὗ ἔμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	L 124 579
οὗ ἤμελλεν αὐτὸς εἰσέρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	Θ
οὗ ἔμελλεν αὐτὸς ἀπέρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	Ω
οὗ ἔμελλεν αὐτὸς εἰσέρχεσθαι.	2	εἶπεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	f^1
οὗ ἤμελλεν αὐτὸς διέρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	f^{13}
.... ὸς ἔρχεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	33
οὗ ἤμελλεν αὐτὸς εἰσπορεύεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	157
οὗ ἔμελλεν αὐτὸς πορεύεσθαι.	2	ἔλεγεν	δὲ	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	700 [↓1424
οὗ ἤμελλεν αὐτὸς ἔρχεσθαι.	2	ἔλεγεν	οὖν	πρὸς αὐτούς,	Ὁ μὲν θερισμὸς	𝔐 M N U W Γ Δ Λ Ψ 2

(4)	5	6	7	
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου τοῦ θερισμοῦ	ὅπως	B 𝔓75 **uwτ** rell
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου τοῦ θερισμοῦ	ὅπως ἂν	Y K M Π
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	τοῦ θεοῦ τοῦ θερισμοῦ	ὅπως	D*
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε	τοῦ κυρίου τοῦ θερισμοῦ	ὅπως	D^c
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου	.....	H
πολύς, οἱ δὲ	.....ητε οὖν	τοῦ κυρίου τοῦ θερισμοῦ	ὅπως	33
πολύς, οἱ δὲ ἐργάται ὀλίγοι·	δεήθητε οὖν	τοῦ κυρίου τοῦ θερισμοῦ	ἵνα	579

Collation of Luke 10:2 with variation units numbered above text (Reuben J. Swanson, ed., *New Testament Greek Manuscripts: Variant Readings Arranged in Horizontal Lines against Codex Vaticanus. Luke* [Sheffield: Sheffield Academic Press, 1995], 183)

- Analogous to a DNA sequence alignment



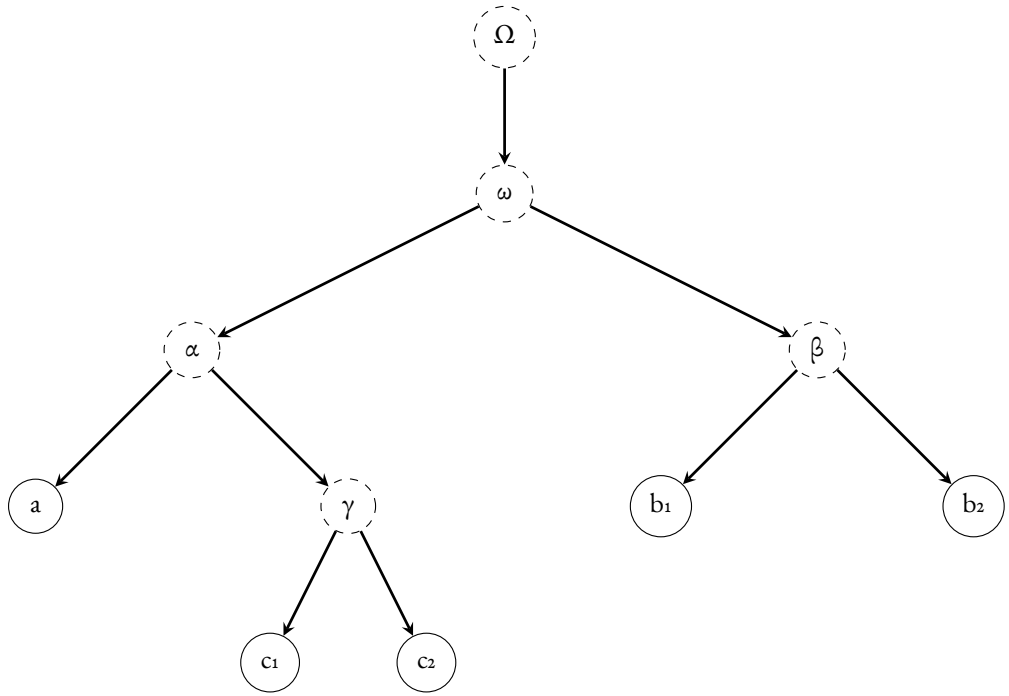
- Rows: *taxa* = witnesses
- Columns: *sites* = variation units
- Cells: *states* = variant readings (including omissions)
  - Lacunae and uncertain retroversions correspond to fully or partially *ambiguous states*

- At the most basic level, a *witness* is just a sequence of readings, a row in the collation

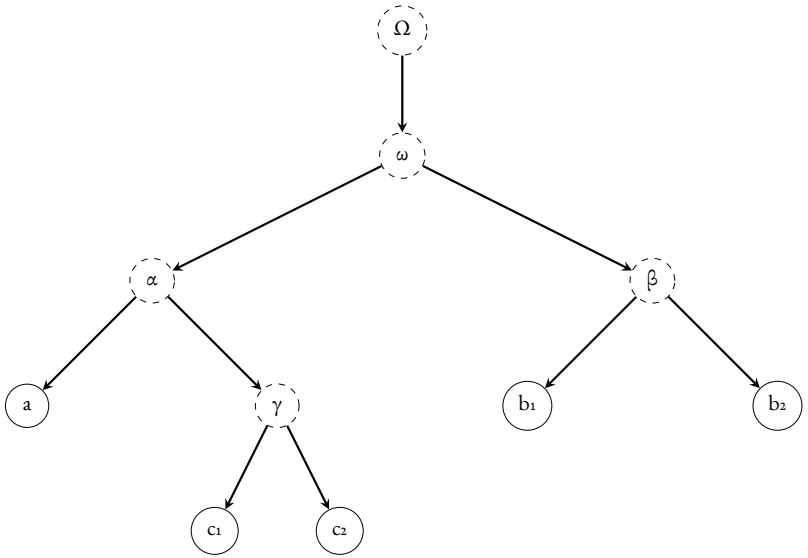
	3Jo 1:1/2	3Jo 1:1/6	3Jo 1:1/8	⋯	3Jo 1:15/23
GA 69	a	af1	a	⋯	a
GA 1739	a	a	b	⋯	a
GA 2243	b	a	a	⋯	a

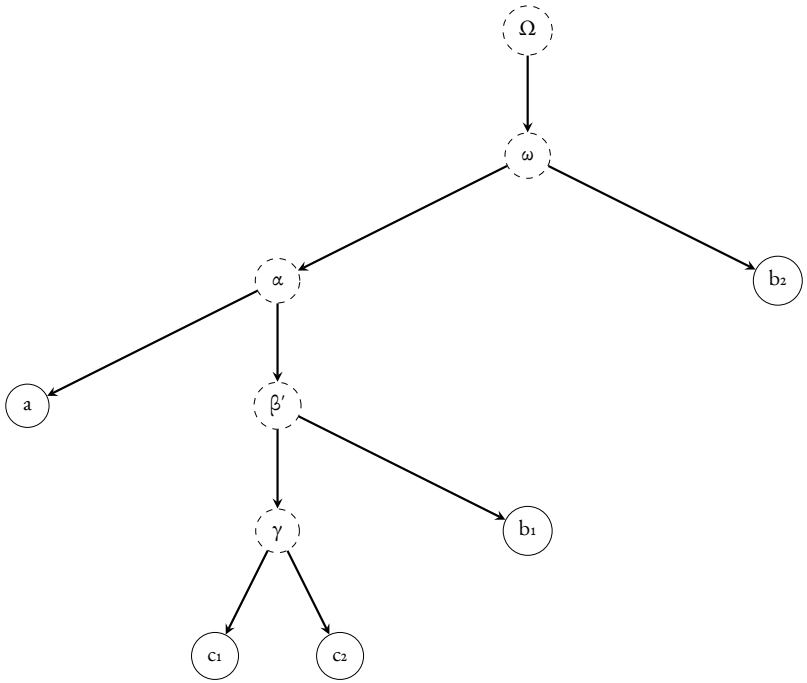
- Paratextual features can be encoded in the same way
- In more complex phylogenetic approaches, age can also be incorporated

Phylogenetics



- A *stemma* is a "family tree" modeling the transmission of the text
- The *leaves* (solid circles) correspond to extant witnesses
- The *hyparchetypes* (dashed circles) are hypothetical (now-lost) ancestors, reconstructed from their descendents along the branches
- The *archetype* (ω) is the earliest reconstructible text
- The *root* (Ω) represents the authorial text



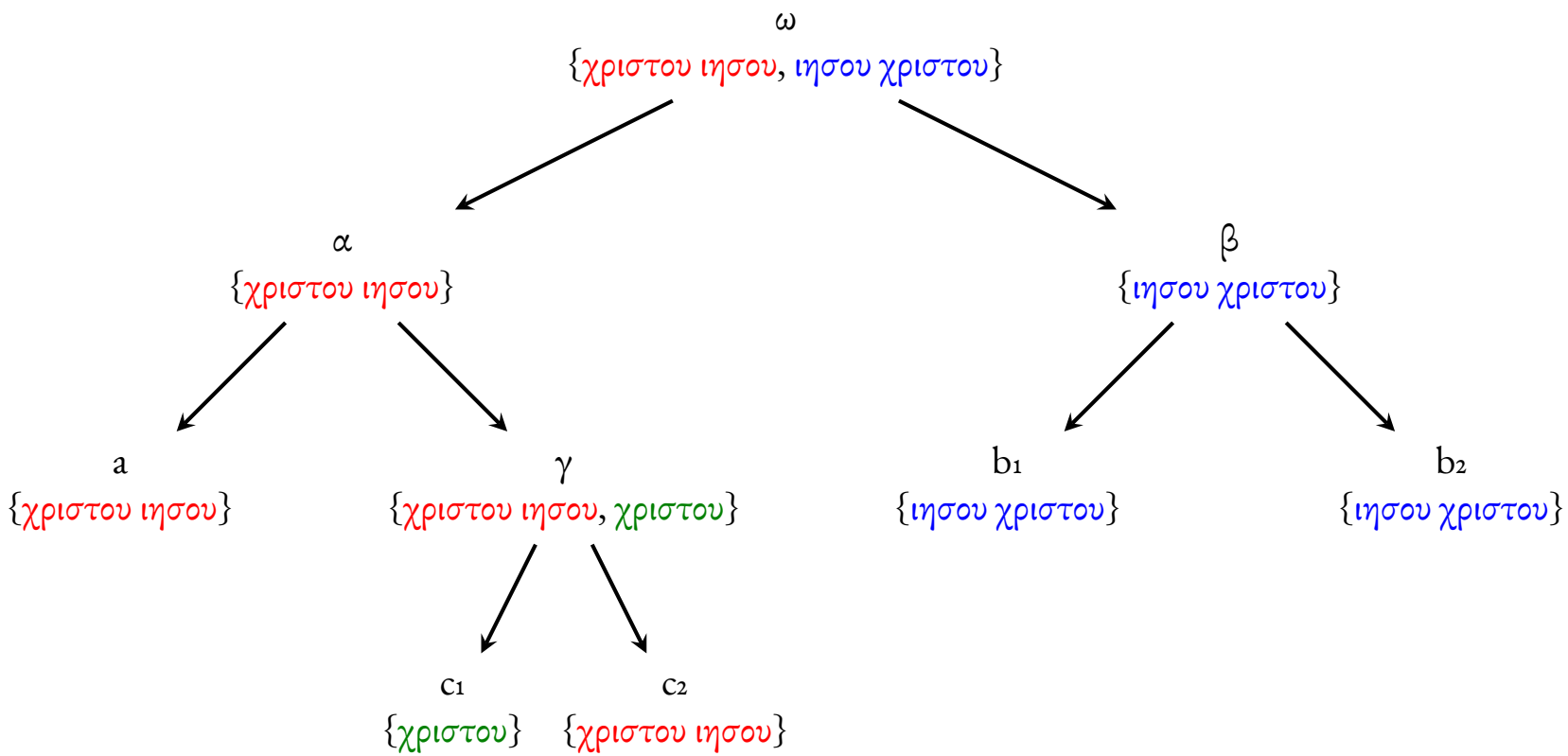


- A stemma represents a hypothesis about transmission history
- The goal is to determine which hypothesis (or hypotheses) best explain the extant data
- To do this, we need a numerical metric for the fitness of a given stemma

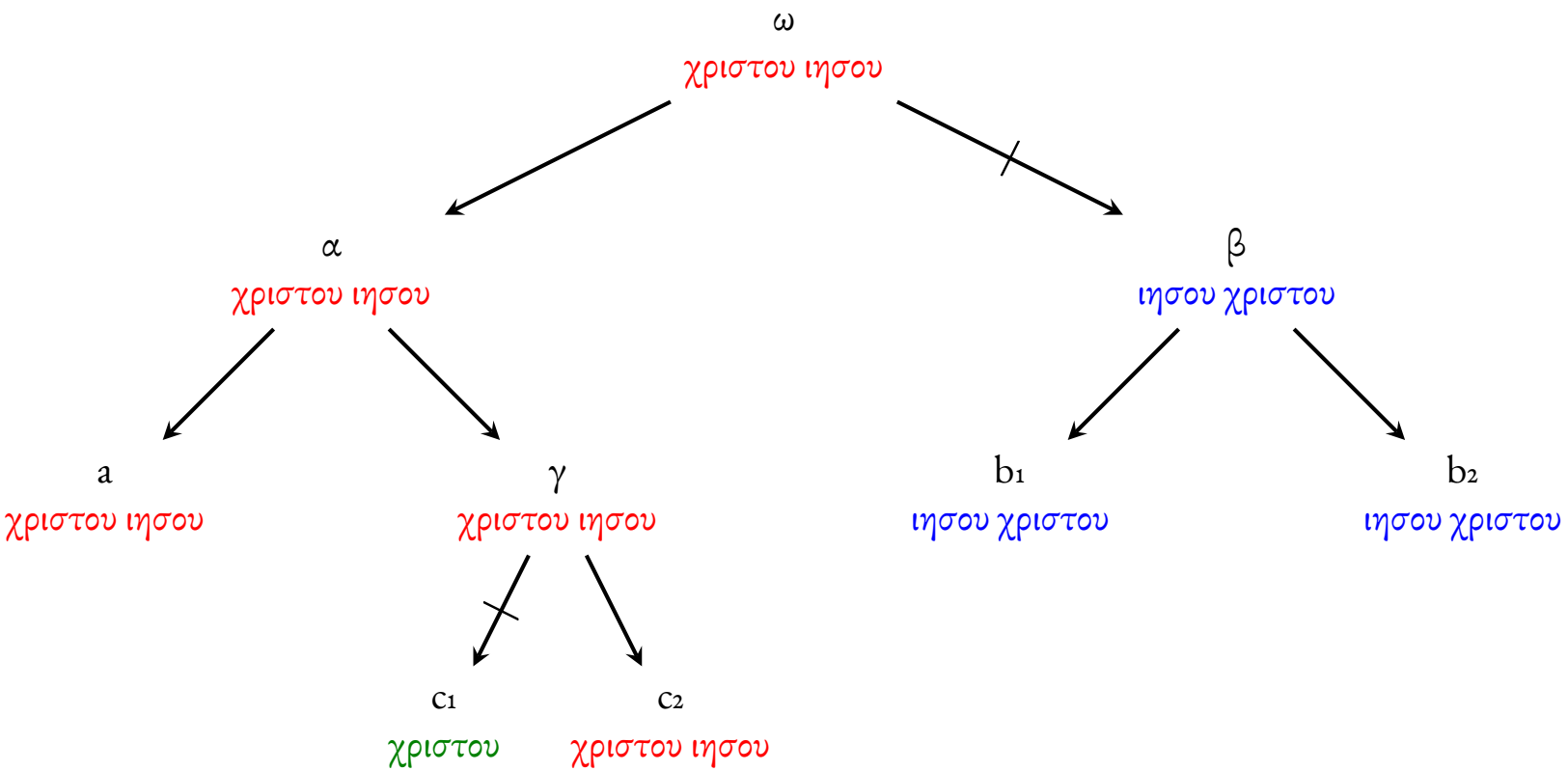
- One such metric is *parsimony*
- Cost = smallest number of changed readings along the branches of the stemma
- Motivated by Ockham's Razor
- Given a candidate stemma, we calculate its cost for each variation unit independently and add up the results
- We calculate it starting at the bottom of the stemma and working our way up

Witness	Reading
a	χριστου ιησου
b_1	ιησου χριστου
b_2	ιησου χριστου
c_1	χριστου
c_2	χριστου ιησου

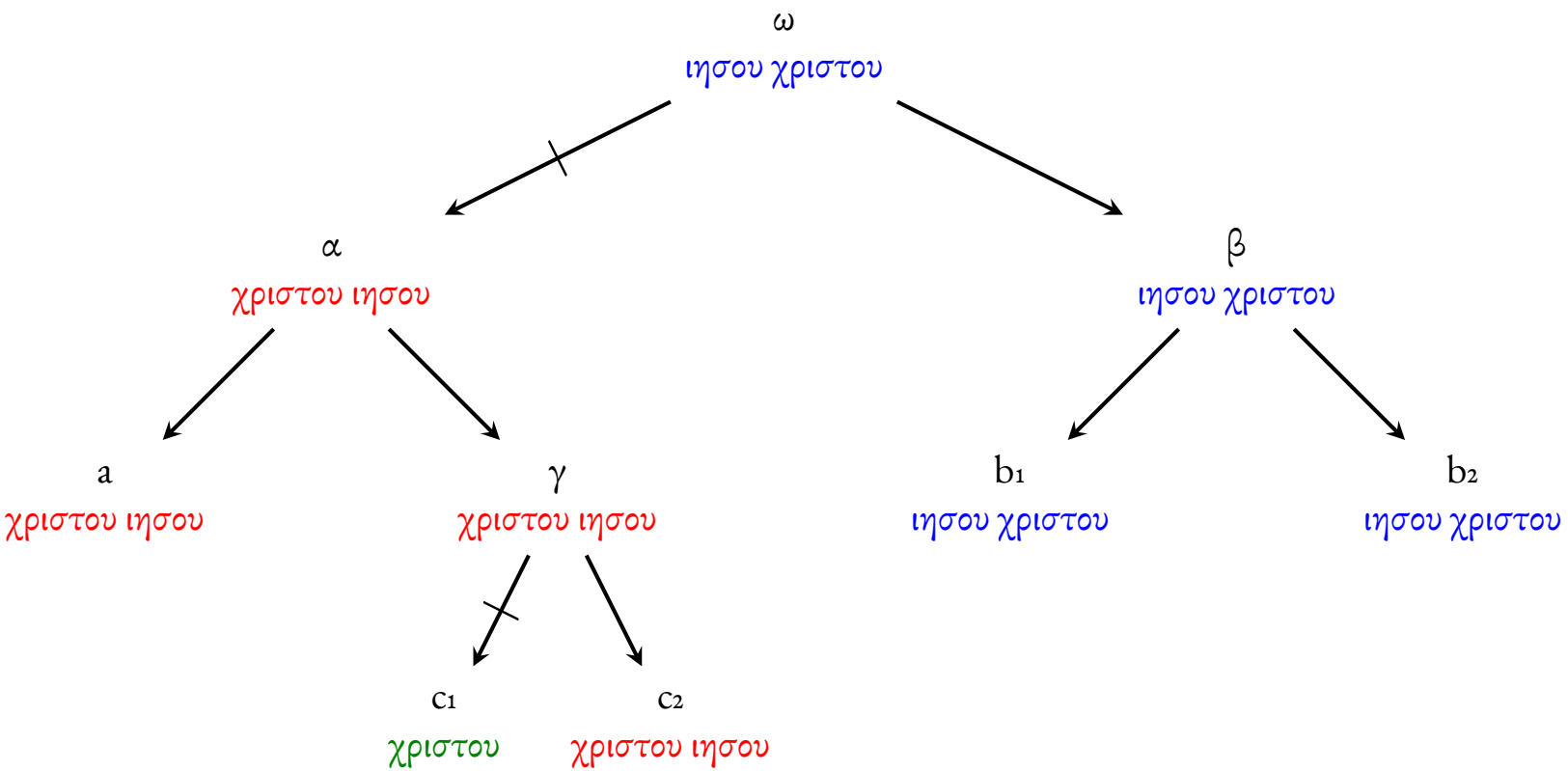
Cost for stemma 1:



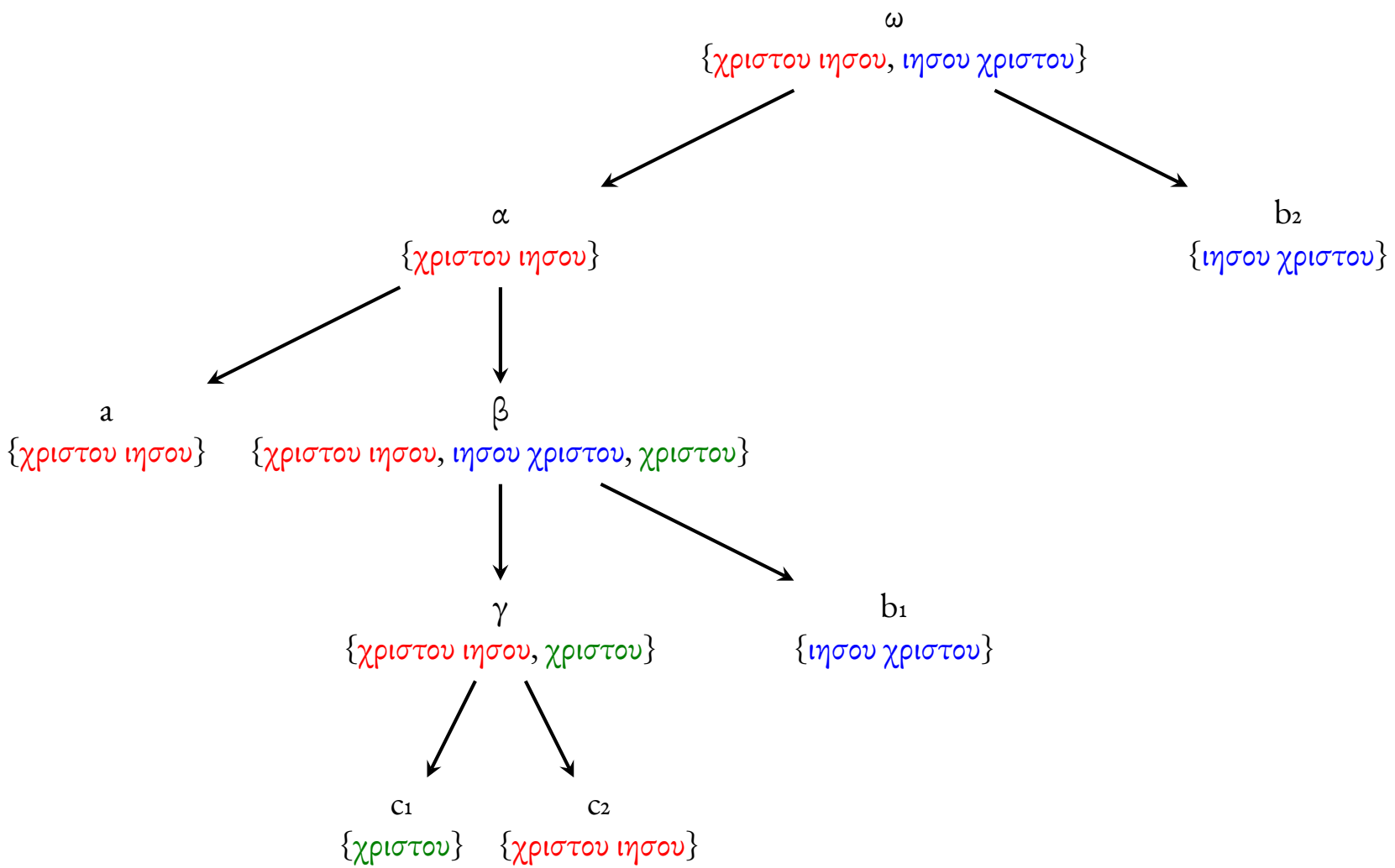
Cost for stemma 1: **2**



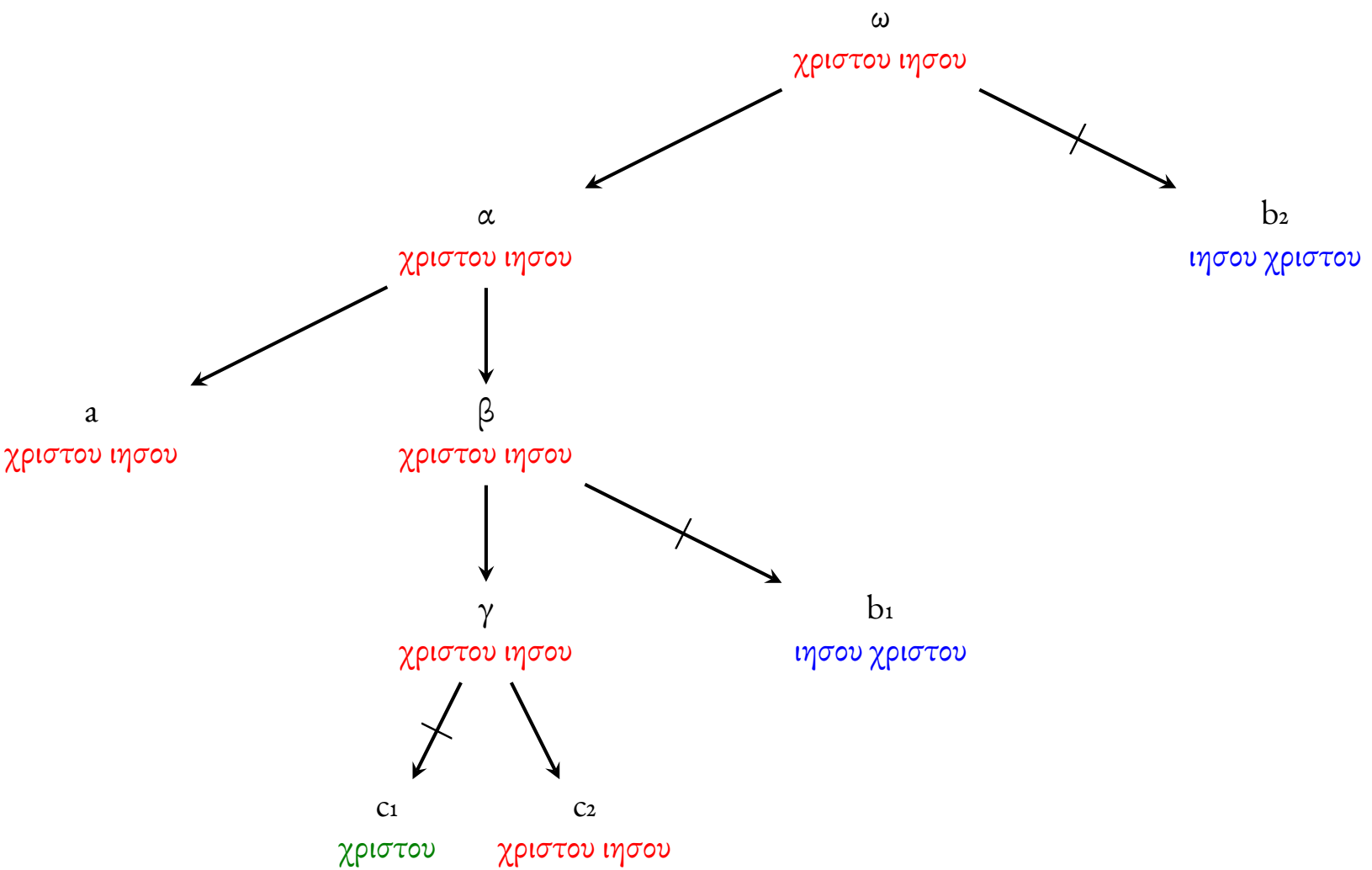
Cost for stemma 1: **2**



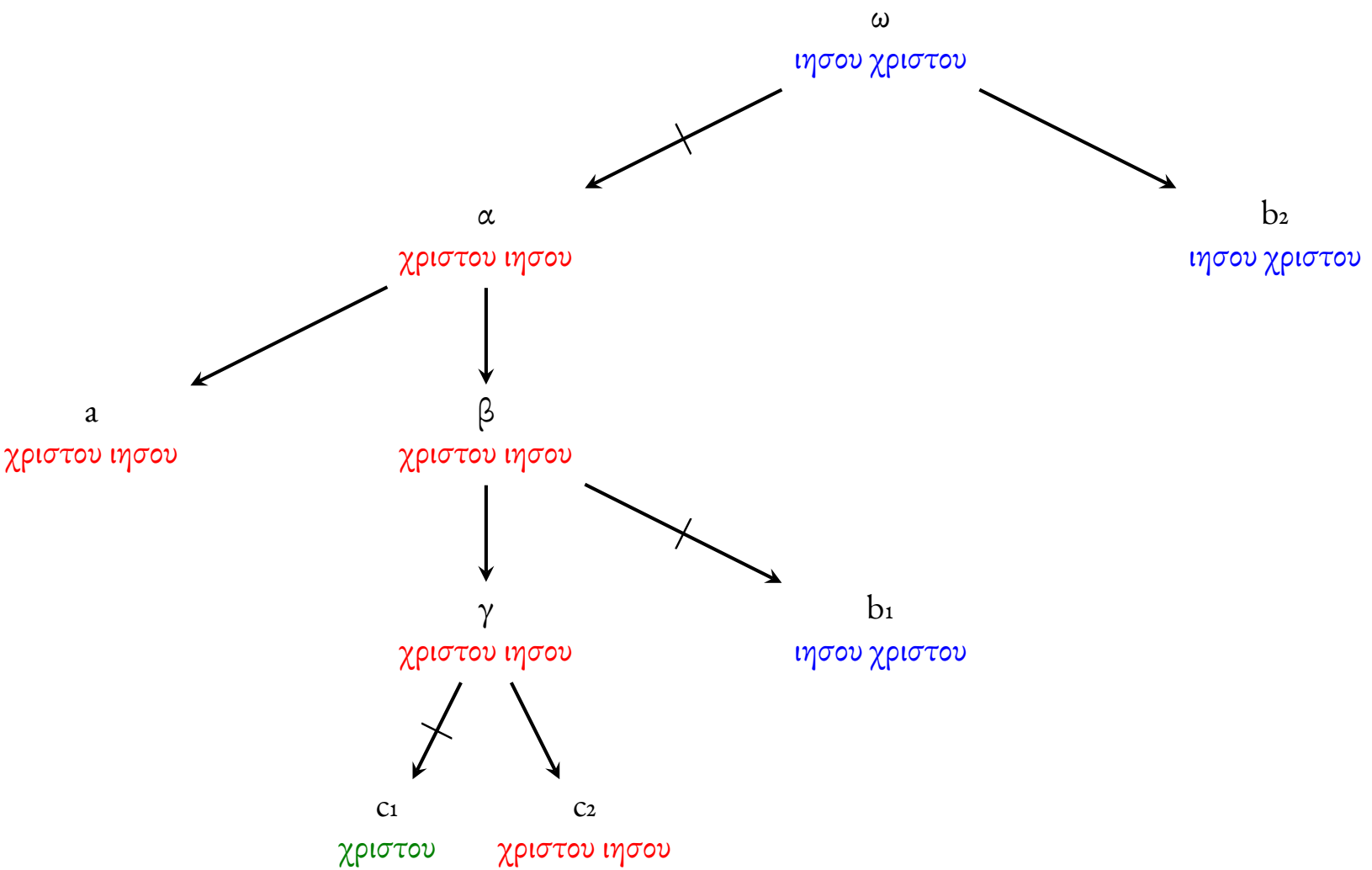
Cost for stemma 2:



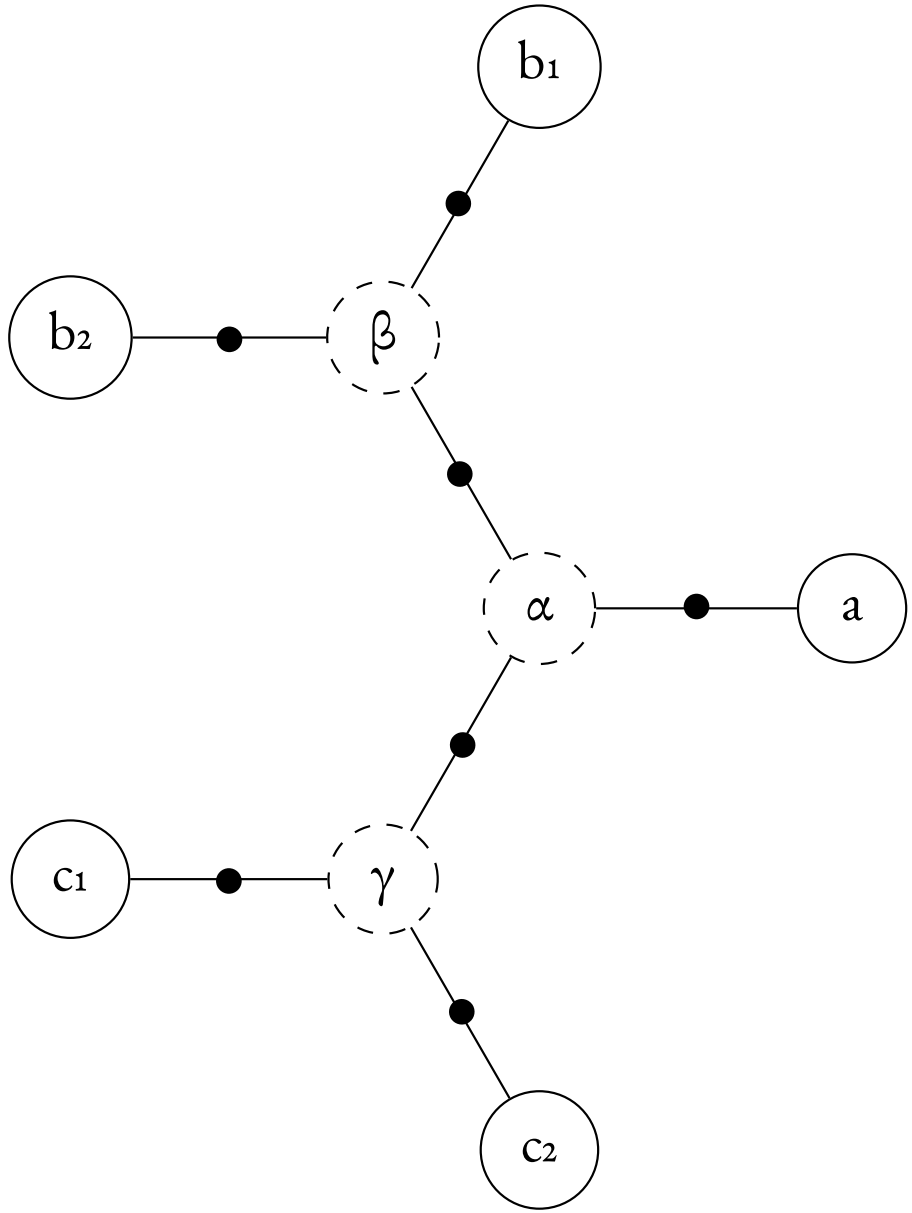
Cost for stemma 2: **3**



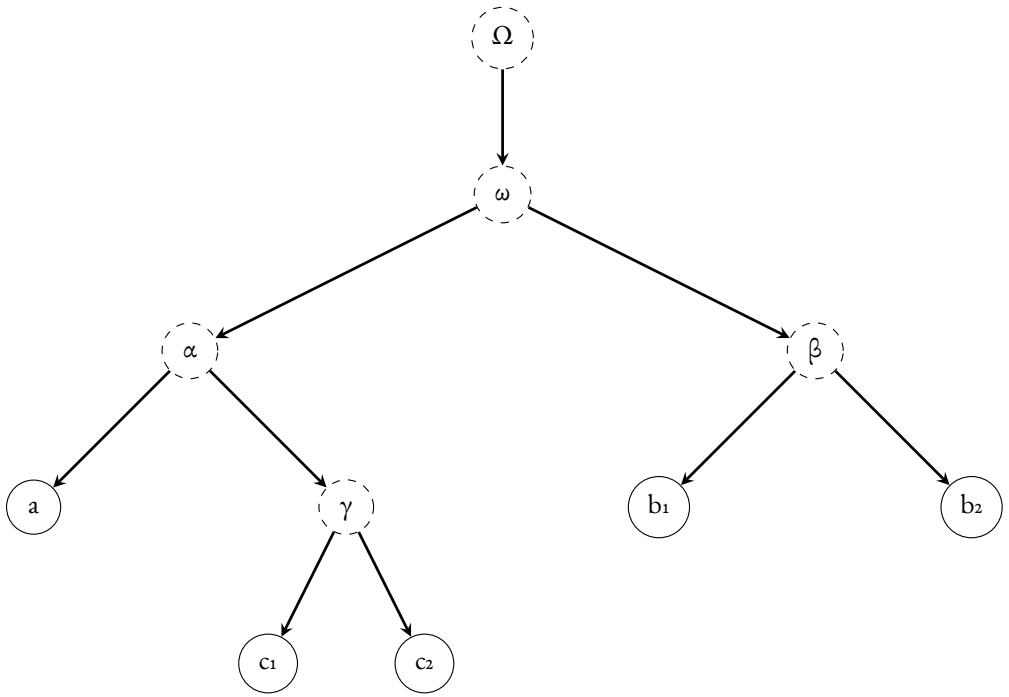
Cost for stemma 2: **3**



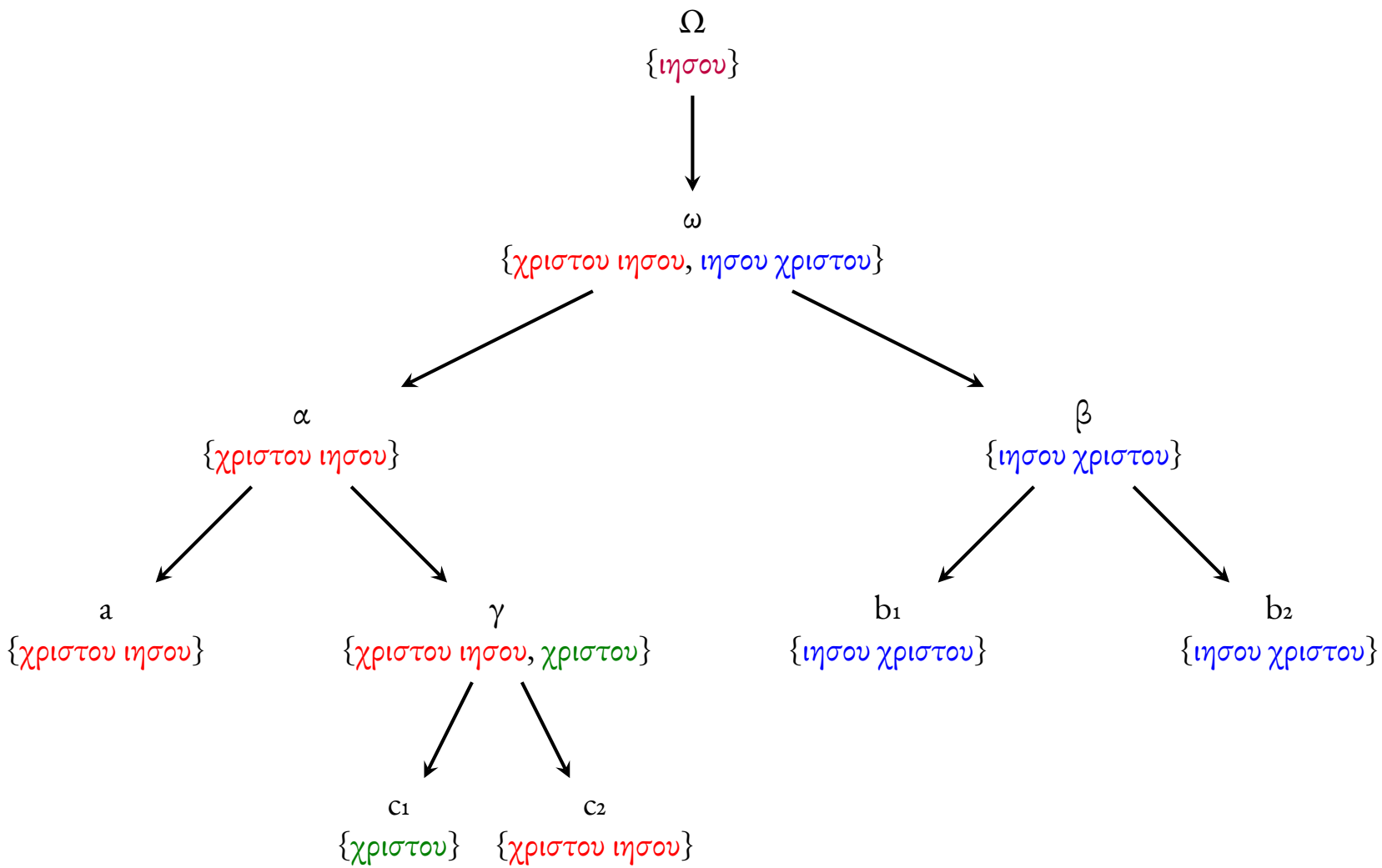
- Minimum number of changes is the same regardless of what the archetype reads
- A stemma's cost *does not depend on where its root is*
- This means the computer can calculate the costs of stemmata without knowledge about the root, and we can postpone the assessment of internal evidence of readings to the end, when we want to determine where the root of the tradition is
- The traditional approach for computer-assisted stemmatics



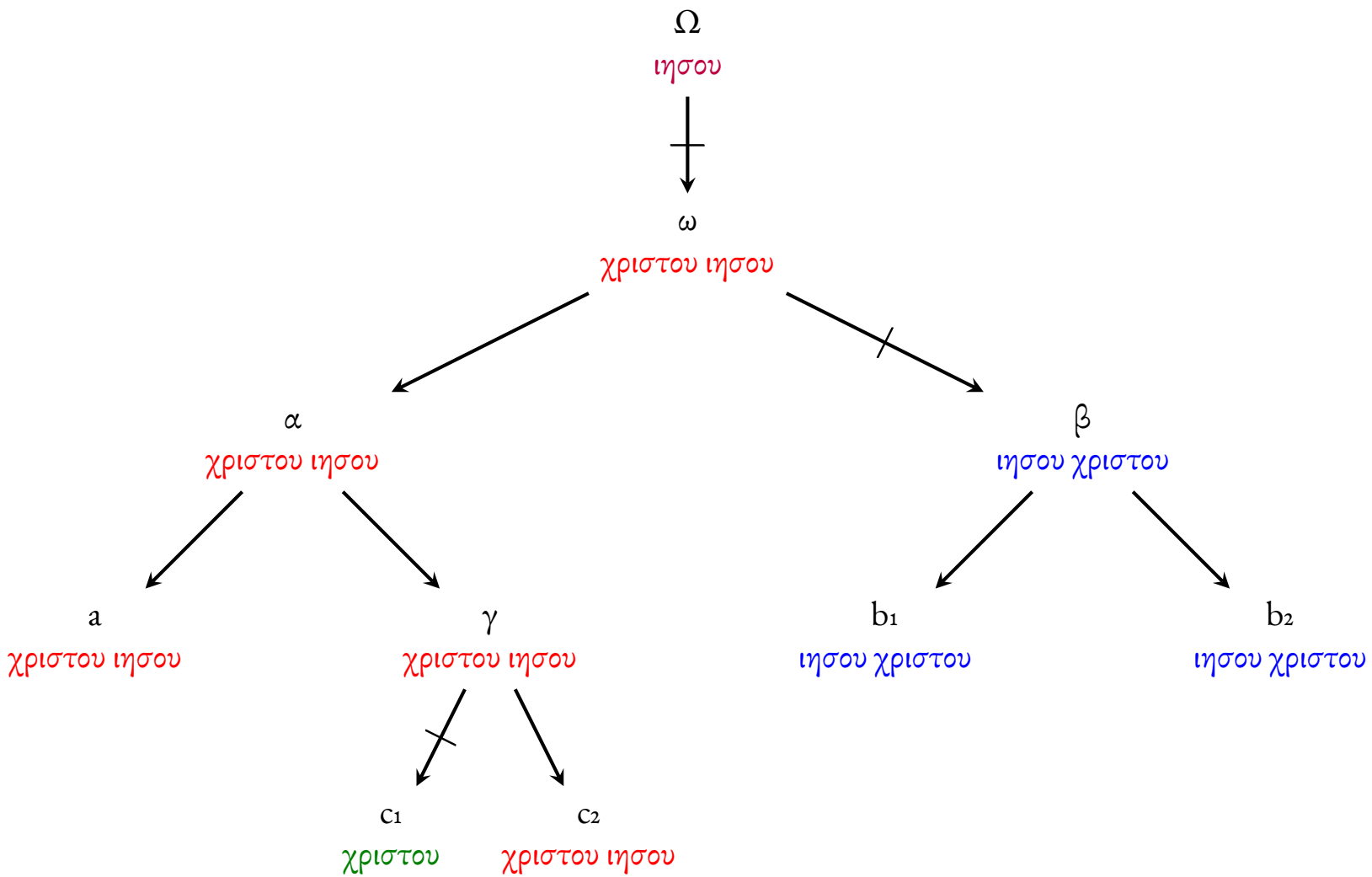
- But we may want to incorporate internal evidence up front:
  - Separation of concerns between *intrinsic probabilities* ("what would the author write?") and *transcriptional probabilities* ("how would later scribes/readers change it?")
  - Intrinsic probabilities can affect the backward pass
  - Transcriptionally, some changes are more likely than others or irreversible
- We can extend the approach to incorporate this evidence, but the stemma costs will now depend on where the root of the tradition is located



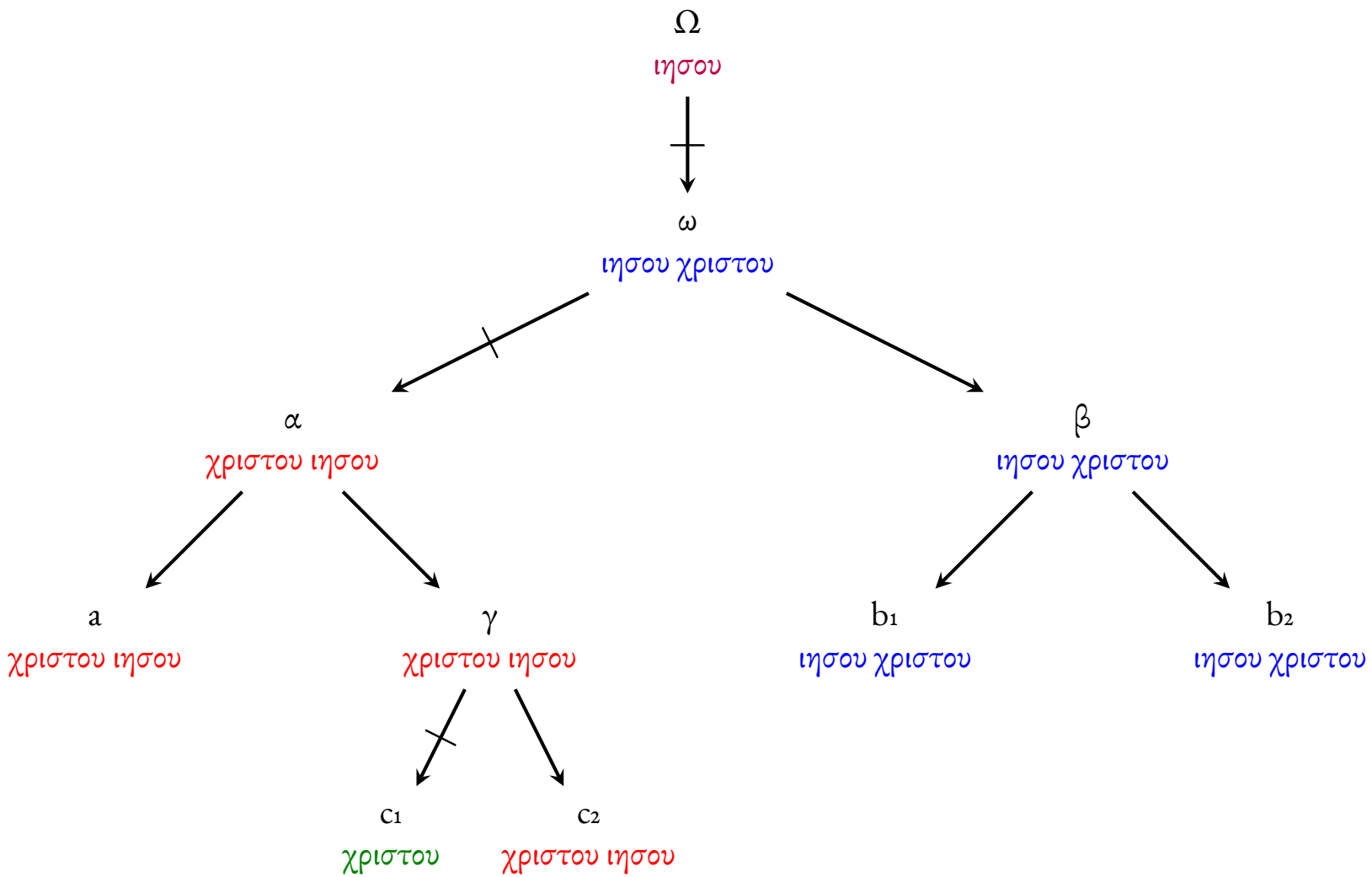
Cost for stemma:



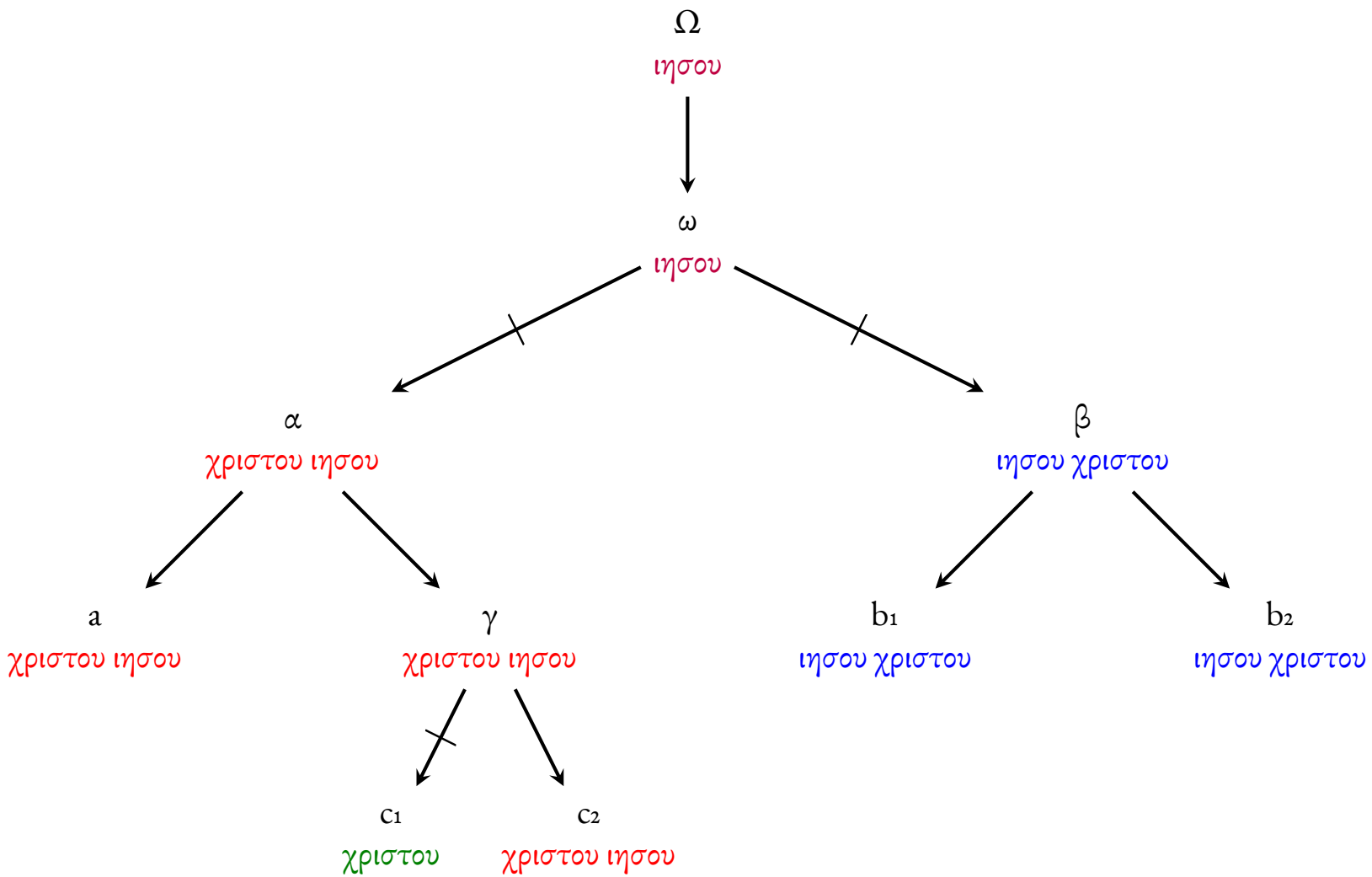
Cost for stemma: **3**



Cost for stemma: **3**

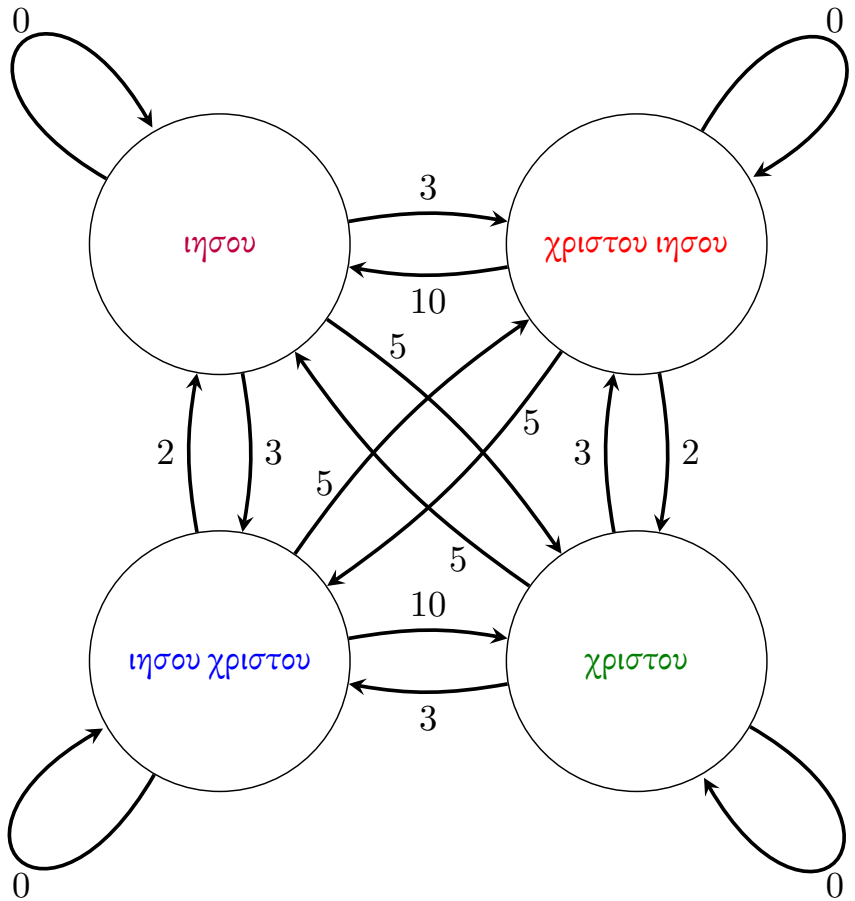


Cost for stemma: **3**



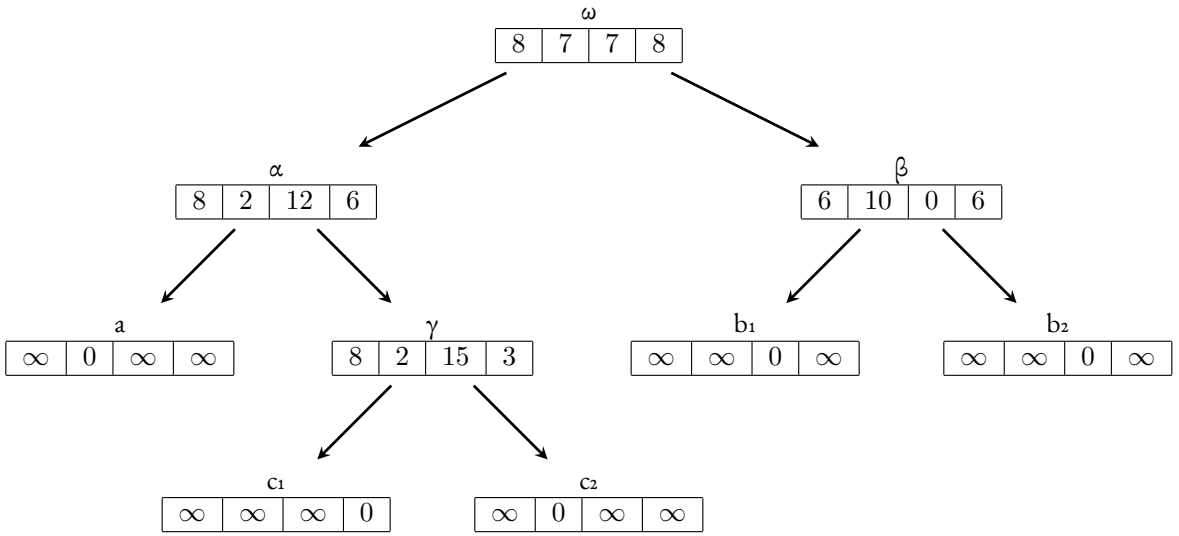
- We can likewise assign weighted costs to different transitions between readings using a *cost matrix*
- A model of the average scribe's behavior

	ιησου	χριστου ιησου	ιησου χριστου	χριστου
ιησου	0	3	3	5
χριστου ιησου	10	0	5	2
ιησου χριστου	2	5	0	10
χριστου	5	3	3	0



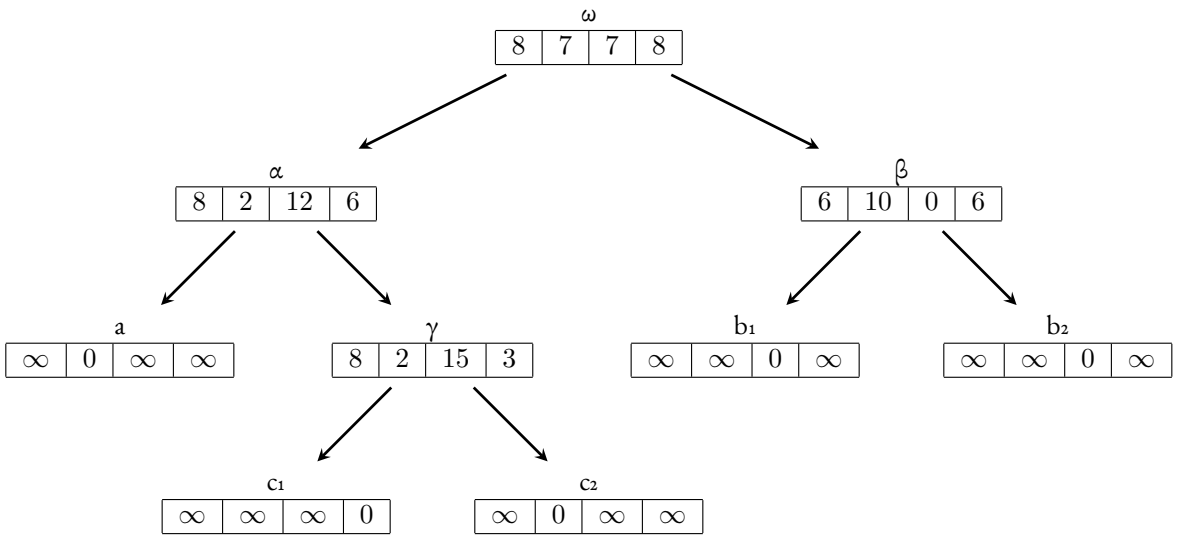
- For each hypearchetype, compute the current minimum cost for each reading it could have based on the minimum costs of its children
- For example, the minimum cost of γ reading *ιησου* is the cost of the transition from *ιησου* to *χριστου* (in c_1) plus the cost of the transition from *ιησου* to *χριστου ιησου* (in c_2): $(5 + 0) + (3 + 0) = \mathbf{8}$

	ιησου	χριστου ιησου	ιησου χριστου	χριστου
ιησου	0	3	3	5
χριστου ιησου	10	0	5	2
ιησου χριστου	2	5	0	10
χριστου	5	3	3	0



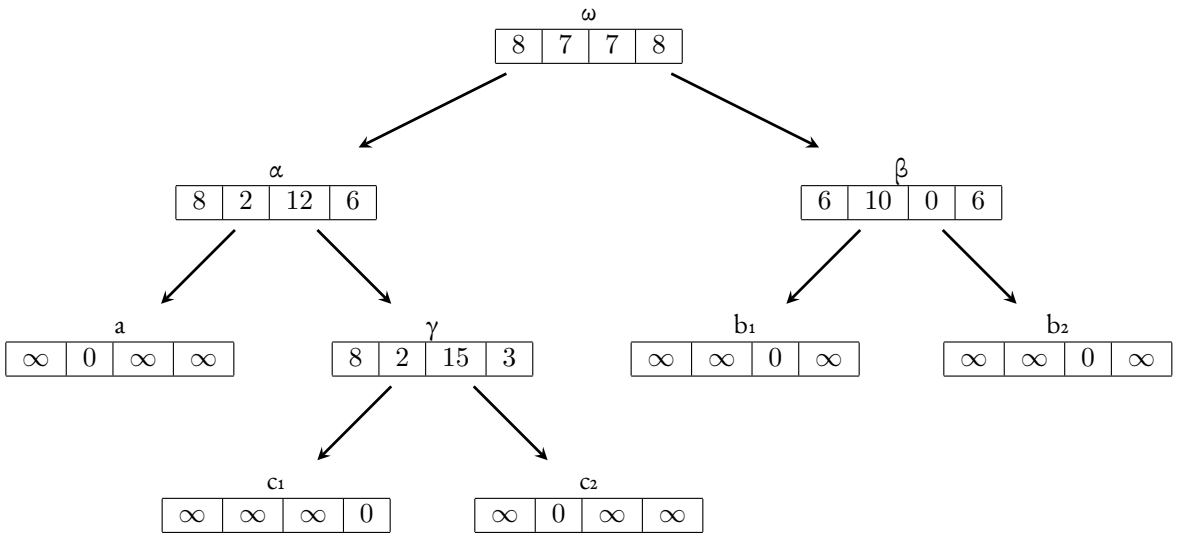
- For α reading $\iota\eta\sigma o\upsilon$, it is $3 + 0$ (for the transition to $\chi\rho\iota\sigma\tau o\upsilon\ \iota\eta\sigma o\upsilon$ in a) plus $\min(0 + 8, 3 + 2, 3 + 15, 5 + 3) = \min(8, 5, 18, 8) = 5$ (for transitions to any of the readings in γ) $\Rightarrow$ **8**
- Meanwhile, for α reading $\chi\rho\iota\sigma\tau o\upsilon\ \iota\eta\sigma o\upsilon$, this is $(0 + 0)$ for a plus $\min(10 + 8, 0 + 2, 5 + 15, 2 + 3) = \min(18, 2, 20, 5) = 2$ for $\gamma \Rightarrow$ **2**

	ιησου	χριστου ιησου	ιησου χριστου	χριστου
ιησου	0	3	3	5
χριστου ιησου	10	0	5	2
ιησου χριστου	2	5	0	10
χριστου	5	3	3	0

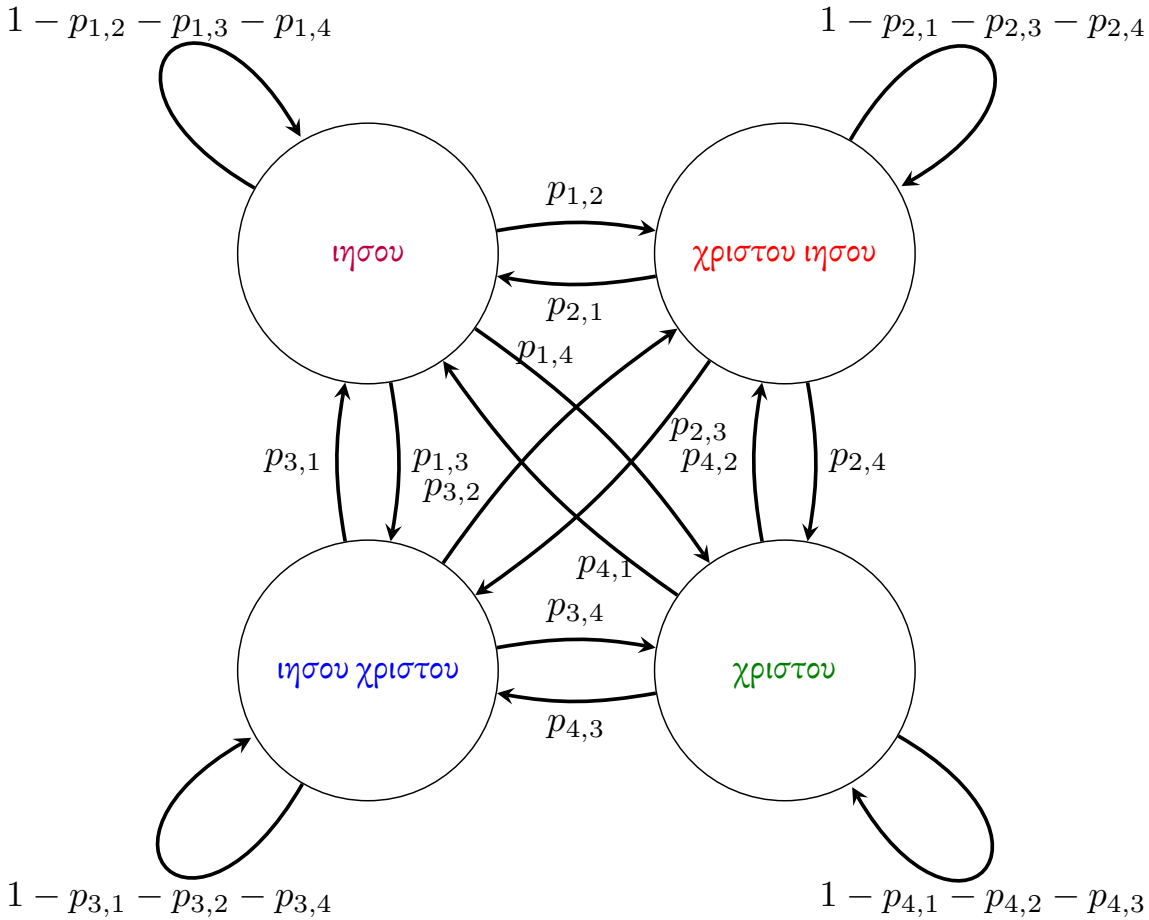


- We can also incorporate intrinsic evidence
- If the root reading is *ιησου*, then the minimum cost of the stemma is $\min(0+8, 3+7, 3+7, 5+8) = \min(8, 10, 10, 13) = \mathbf{8}$
- If the root reading is *χριστου ιησου*, then the minimum cost is $\min(10+8, 0+7, 5+7, 2+8) = \min(18, 7, 12, 10) = \mathbf{7}$

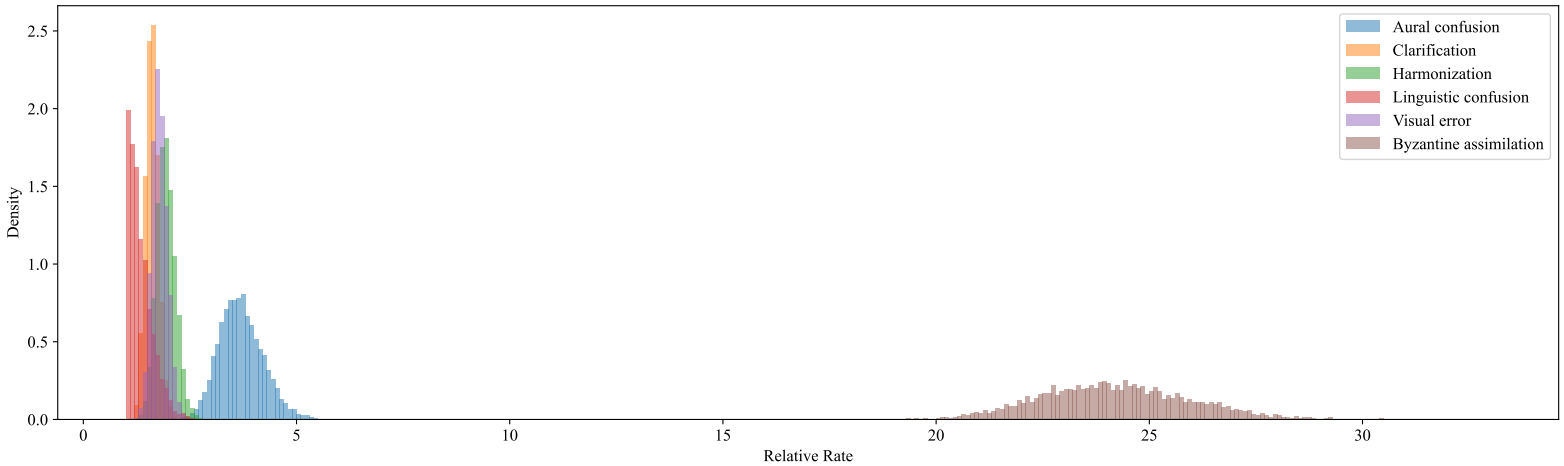
	ιησου	χριστου ιησου	ιησου χριστου	χριστου
ιησου	0	3	3	5
χριστου ιησου	10	0	5	2
ιησου χριστου	2	5	0	10
χριστου	5	3	3	0



- We can even use probabilities for intrinsic and transcriptional evidence
- Intrinsic probabilities = *prior probabilities* at root
- Transcriptional probabilities = probabilities of copying the same reading or changing it to another reading, modeled as a *Markov chain*



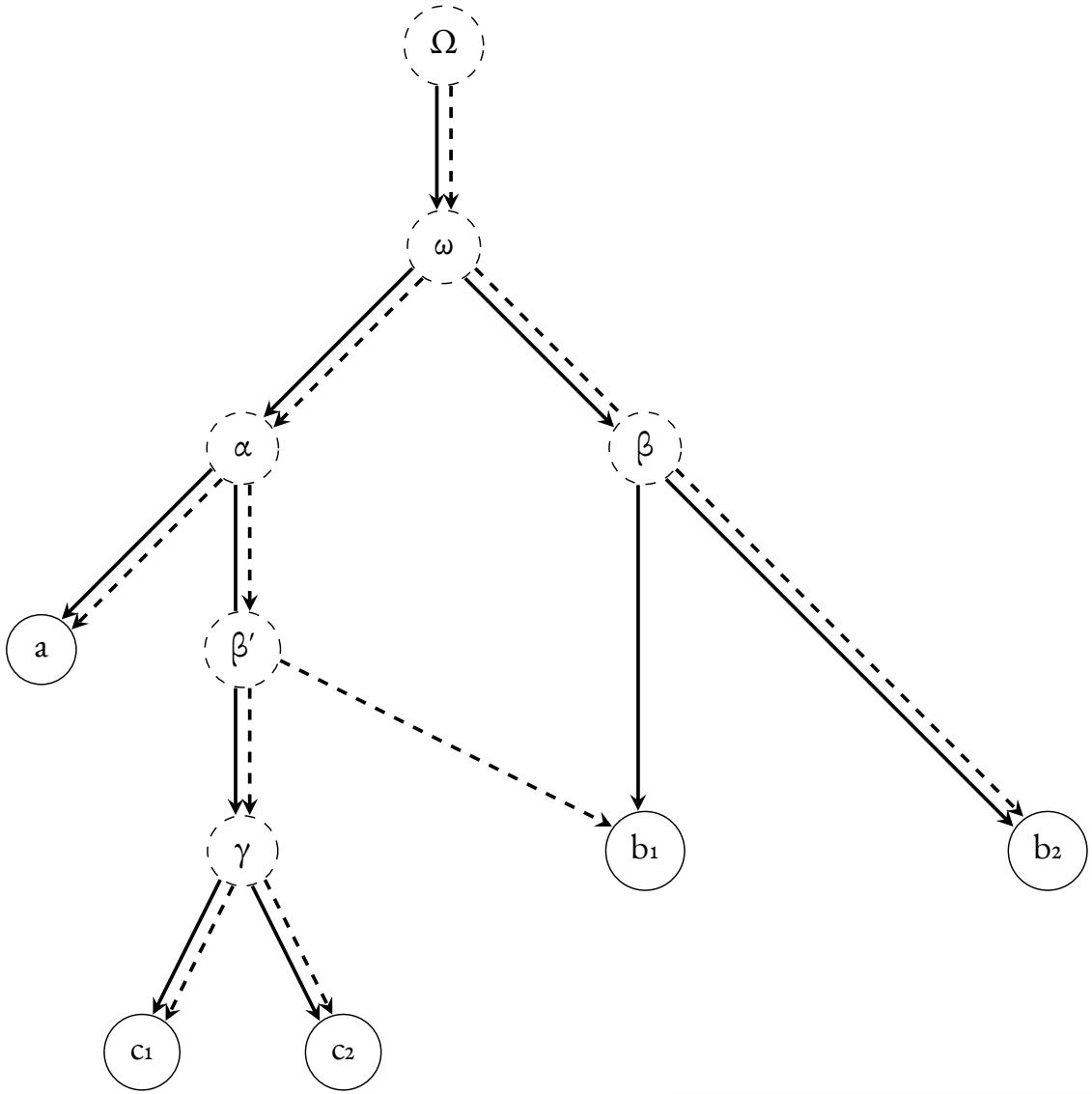
- In a probabilistic setting, we can incorporate and estimate other parameters of interest:
  - Probabilities for classes of transitions between readings (scribal habits)
  - Lengths of branches (how many copying events separated an ancestor from a descendent, and how error-prone were the scribes involved?)
  - Dates of witnesses (using clock models)
  - Measurements of how certain we can be about the best stemmata found in the process
- More complex, but feasible with modern computers!



CBGM

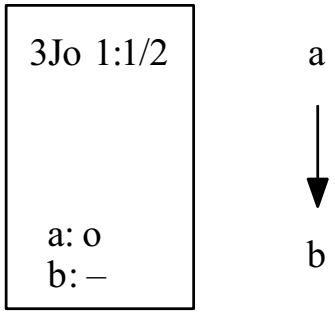
ACU INSTITUTE FOR RELIGION & CRITICAL INQUIRY

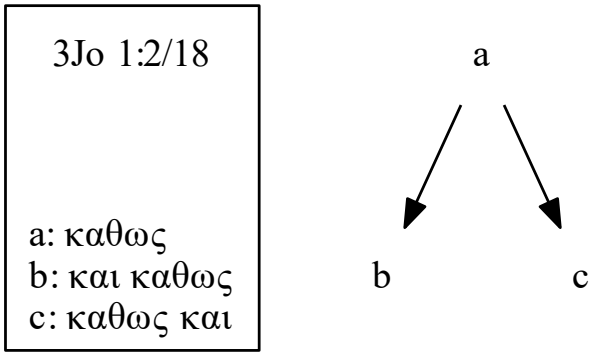
- Developed to solve *contamination*, or mixture across branches of the textual tradition



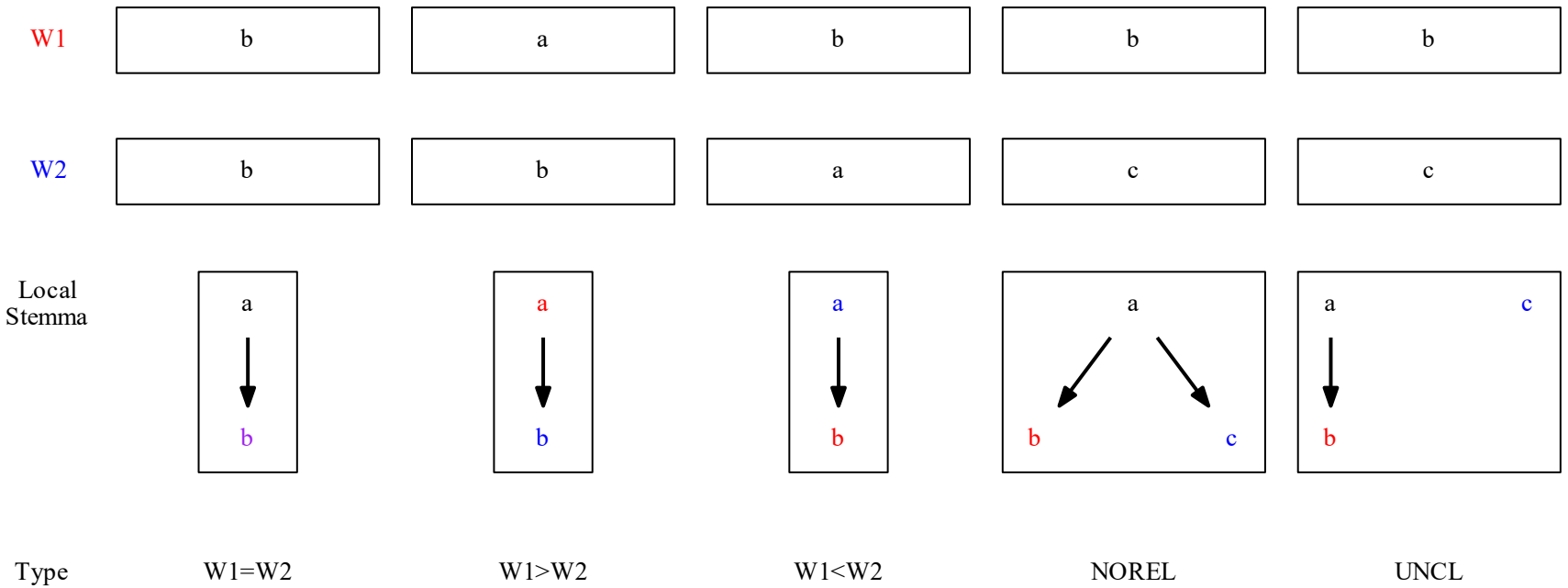
- Foundational principles:
  1. Scribes typically copied their exemplars with fidelity.
  2. If a scribe introduced a variant, then it came from some other reading.
  3. Scribes typically used fewer sources rather than many.
  4. Scribes typically used closely related sources rather than distant ones.
- Witnesses are *texts* (sequences of readings), minus the material baggage (date, provenance, etc.)
  - "How texts relate" ≠ "How manuscripts relate"
- *No hypothetical ancestors* (except for the *Ausgangstext* A)
  - Contamination would (presumably) hinder their reconstruction
  - Instead, we use extant witnesses as proxies for different states of the text
  - This yields a much smaller and more manageable space of solutions compared to the space of all possible stemmata for a given set of witnesses

- The basic unit of comparison
- One for each variation unit
- A graphical representation of our judgments of readings
- Similar to cost graphs in function, but in principle, represents a judgment about what *did* happen, not what *could* happen
- Thus, no bidirectional edges or cycles like we have in cost graphs



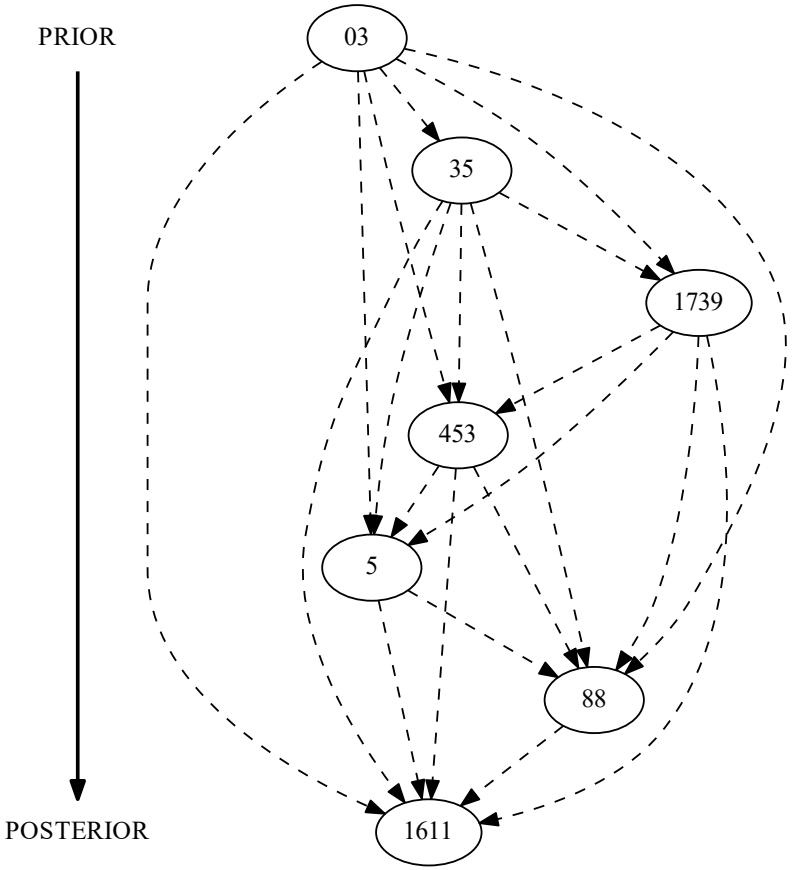


- The relationship of two witnesses is the overall pattern *of the relationships of their readings* at all variation units where both are extant

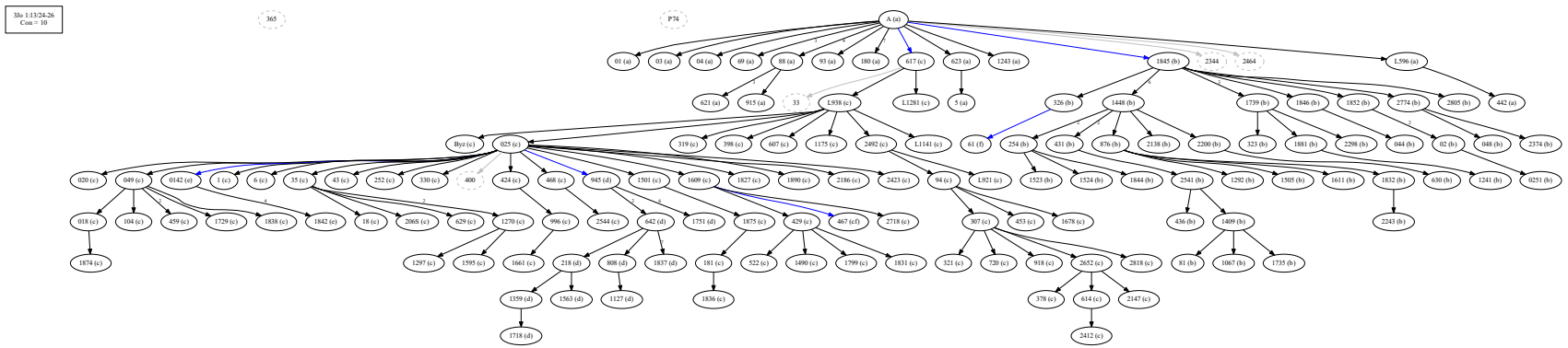


- The first three are the most important

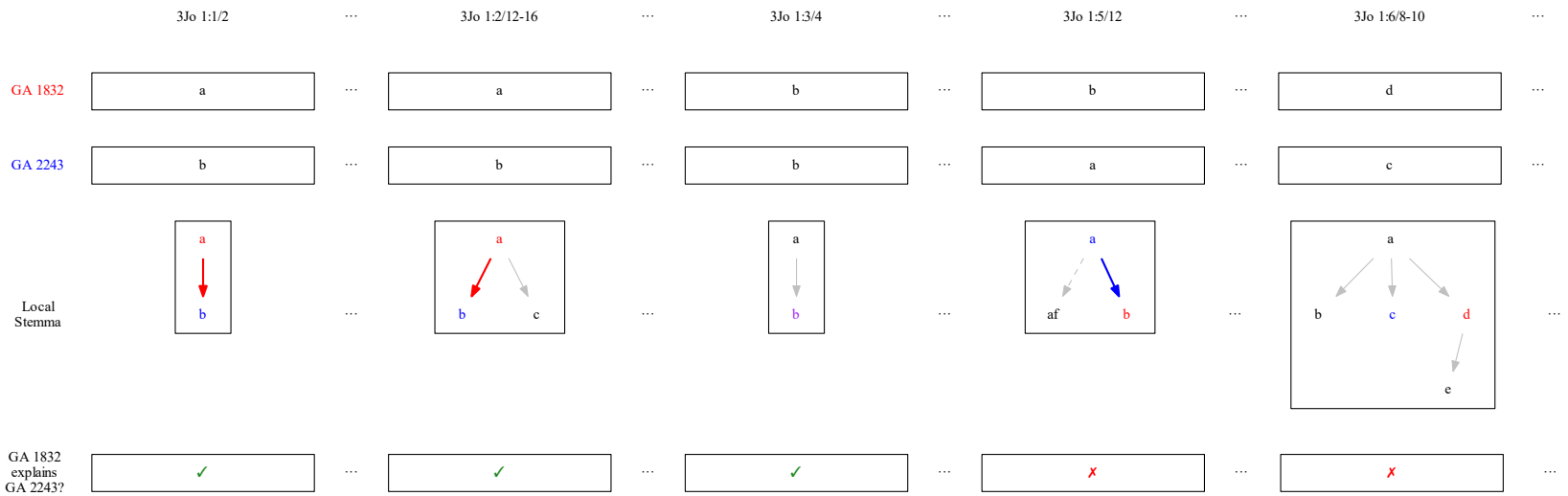
- Potential ancestor = “more prior than posterior readings”



- Textual flow diagram = a tree that relates each witness to its closest potential ancestor, with as few changes in reading as possible
- Similar to a most-parsimonious stemma for a specific variation unit
- We specify a *connectivity limit* κ (i.e., a radius of "close-enough" neighbors)
- Then, for each witness:
  1. List its potential ancestors, sorted from most agreement to least
  2. If one of the first κ has the same reading at this unit, pick the first that does
  3. If not, pick the first (non-lacunose) potential ancestor
- Core idea: use *general relationships* between witnesses to find *specific relationships* between readings, so local stemmata can be refined



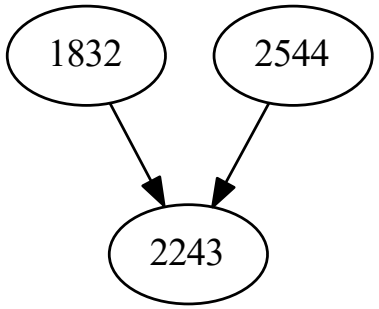
- We say that one reading *explains* another if
  - it is the same reading (“explanation by agreement”), or
  - there is an edge in the local stemma from it to the other reading (“explanation by descent”)



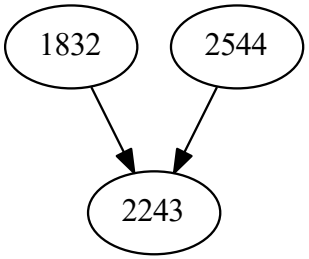
- Lacunae do not have to be explained, and they cannot explain readings

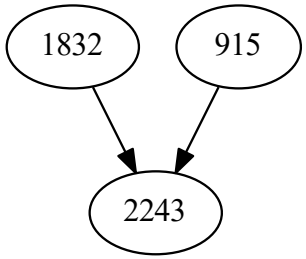
- The *substemma* of a witness is the portion of the global stemma consisting of the witness and its ancestors in the stemma
- Requirement: *every* extant reading in the witness must be explained by a reading in at least one of its ancestors

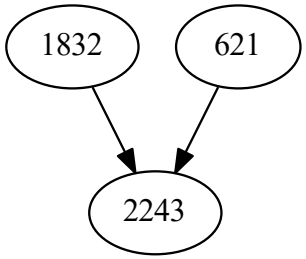
Explained by GA 1832	⋯	✗	✓	✓	✓	⋯
Explained by GA 2544	⋯	✓	✗	✗	✓	⋯
Explained by Either	⋯	✓	✓	✓	✓	⋯

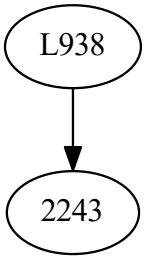


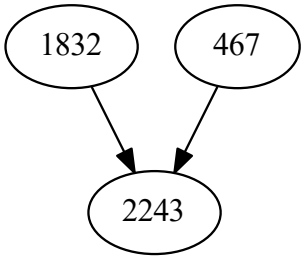
- A witness may have multiple valid substemma (i.e., ones that explain all of its readings), but some are better than others
- Two of the CBGM's methodological assumptions are important here:
  3. Scribes typically used fewer sources rather than many.
  4. Scribes typically used closely related sources rather than distant ones.
- Thus, we need a cost function to distinguish between candidate substemmata



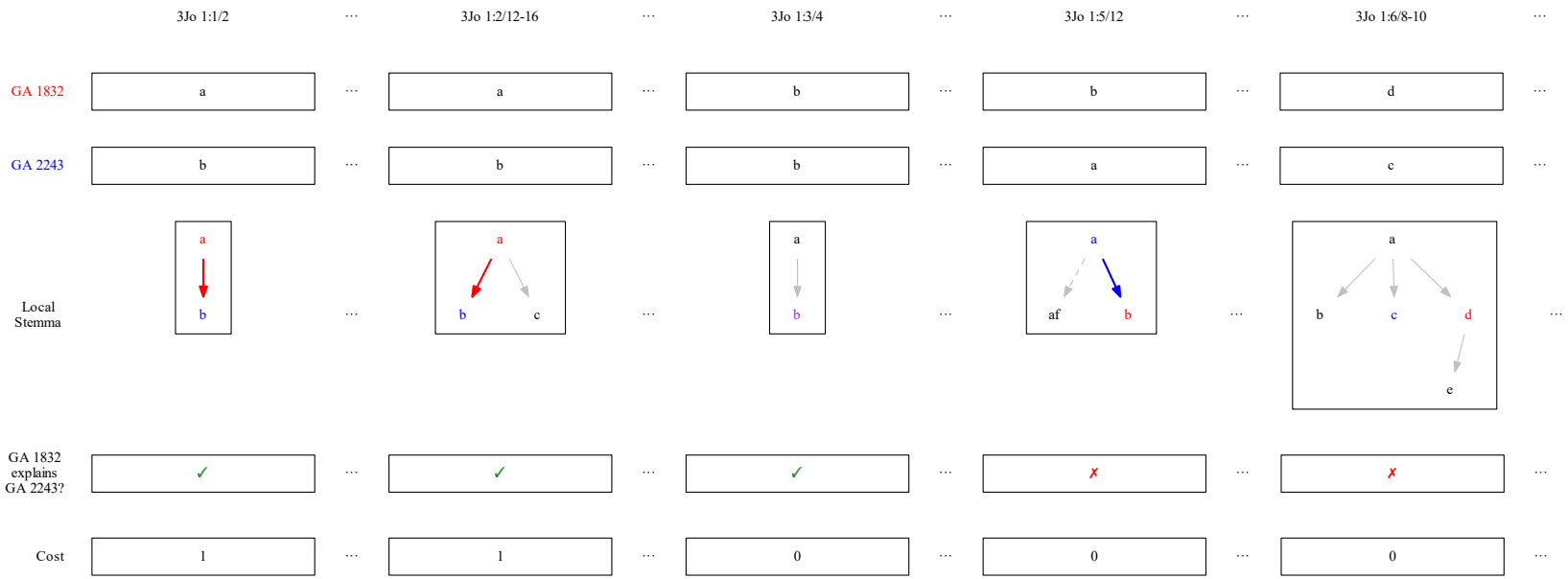








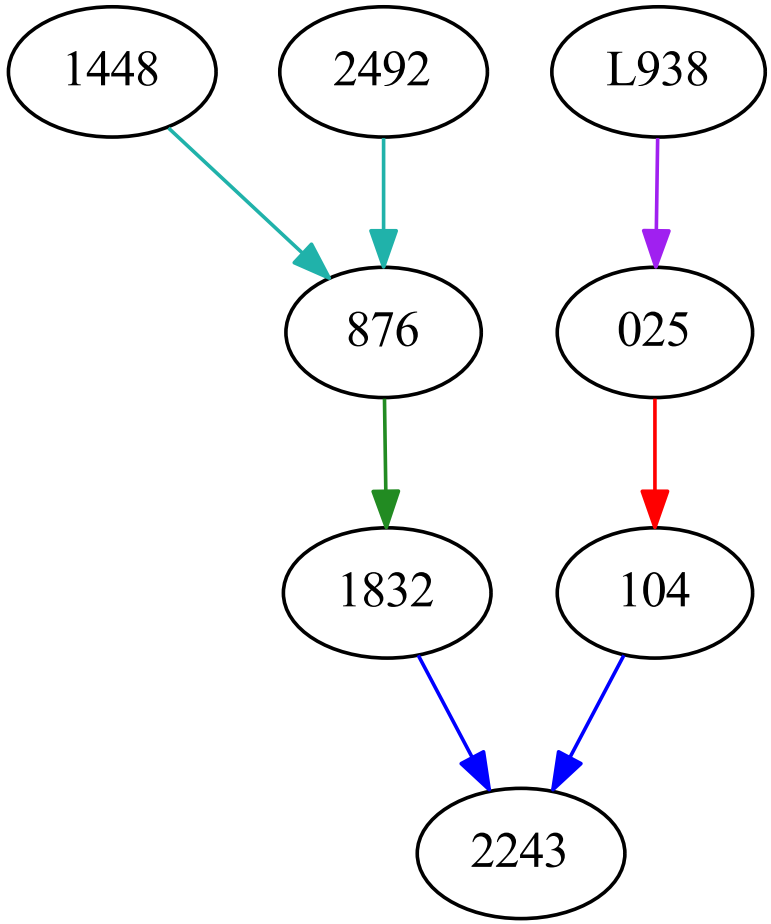
- A simple cost function: “the number of variation units where the ancestor explains the witness by descent and not agreement”
- Thus, in the example below, GA 1832 is a stemmatic ancestor of 2243 with a cost of 2 (but it cannot be its only stemmatic ancestor)



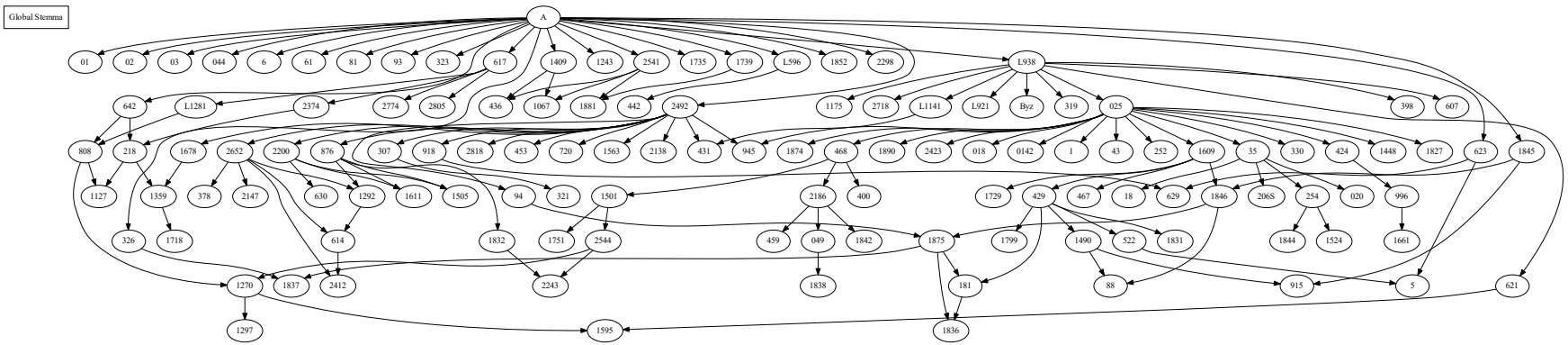
- The process of finding the best substemma for a witness
- For n potential ancestors, a *weighted set cover* problem with n sets
- $2^n - 1$ possible combinations to check (!), but fast heuristics exist

Substemma	Variation Units Explained				Cost
{A}	✓	✓	✓	✓	4
{B}	✓	✓	✗	✗	1
{C}	✗	✓	✓	✓	2
{A, B}	✓	✓	✓	✓	4+1=5
{A, C}	✓	✓	✓	✓	4+2=6
{B, C}	✓	✓	✓	✓	1+2=3
{A, B, C}	✓	✓	✓	✓	1+2+4=7

- Just as the local stemma relates readings, the *global stemma* relates witnesses
- Combination of all substemmata into a single graph
- Analogous to a phylogenetic stemma

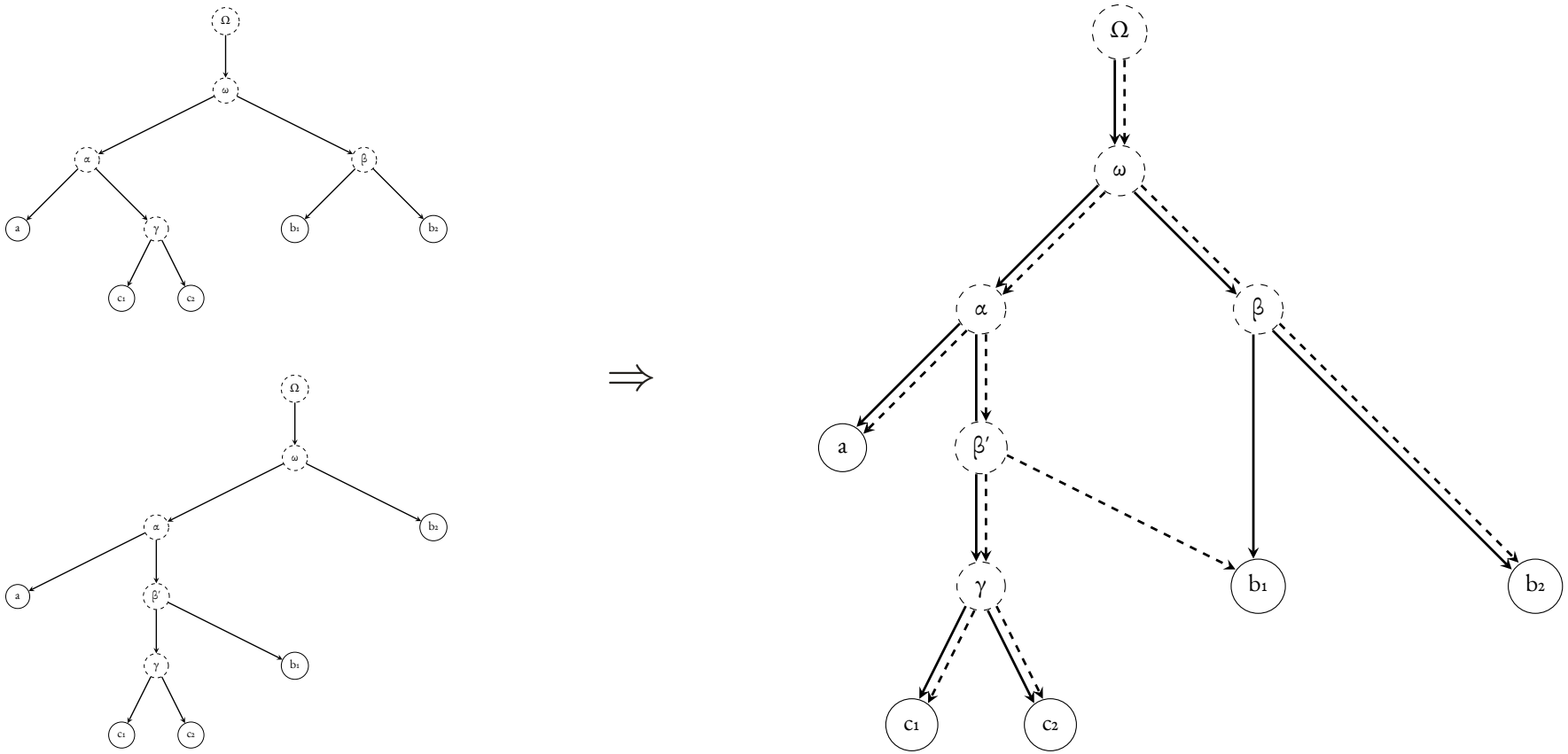


- How does this differ from a phylogenetic stemma?
  - Converging branches (reflecting contamination) are allowed
  - In practice, it takes much less time to produce a global stemma (minutes) than it does to do a satisfactory phylogenetic search for promising stemmata (> a day)
  - No hyparchetypes, and texts found in later manuscripts can be ancestors to texts found in earlier manuscripts
  - Can it model a history of the text?



Conclusions

- Can we get the advantages of both approaches?
- The CBGM could use cost graphs instead of local stemmata
- Phylogenetics could use the *local-genealogical principle* to model contamination
  - Different low-cost stemmata at different variation units



Questions?