

Algebraic Statistics for Computational Biology  
Errata

*Edited by*  
Lior Pachter and Bernd Sturmfels



- p. 32, line -1: missing space in “64states”
- p. 47 (11 lines up from bottom) “and that” could be “implies” and assume less.
- p. 51 (12 lines up from bottom) ‘The “I” and “D” need to be switched.
- p. 54 The “I” and “D” need to be switched in the extended alignment graph.
- p.57  $g_{\sigma^1, \sigma^2}$  is a polynomial in the 24 unknowns  $\theta_{\dots}$  and  $\theta'_{\dots}$ , and is obtained from  $f_{\sigma^1, \sigma^2}$  by setting the 9 unknowns  $\theta'_{\dots}$  to 1.
- p. 60 line -3: This is the first use of “irredundant” and there are several more. It doesn’t seem to be defined anywhere; add a definition.
- p. 65, Theorem 2.26: “If all coefficients of f and g are positive ...” change “positive” to “negative”, or change “all coefficients” to “all nonzero coefficients”
- p. 71 (12 lines up from the bottom) (A,C,G,C)
- p. 83 PHYLIP is freeware not open source.
- p. 83 The example of the phylip file should be 5.6, not 5 10.
- p. 87 bottom. The definition of  $V_S(\mathcal{F})$  is correct, but takes awhile to unravel. Could you add a more direct definition

$$V_S(\mathcal{F}) = \{(z_1, \dots, z_m) \in S : f(z_1, \dots, z_m) = 0 \text{ for all } f \in \mathcal{F}\}$$

- p. 93  $m$  is used to denote a monomial and is also used to count the polynomials  $p_1, \dots, p_m$ . Both of these are in the same problem and the usage conflicts.
- p. 109 Tropical monomial notation in (3.30)  $q_1^2 q_2^3 q_3^2 q_4$  is confusing. On pages 45-46 you use  $D_G^{\odot n}$  to denote tropical powers (yeah, that was for an operator instead of a scalar variable, but it could still be used for (3.30) and what follows from it)
- p. 112 Figure 3.4: This is the picture of the tropical curve obtained using ‘max’. You want the one obtained from ‘min’. (Note that the written description of the curve in Example 3.37 is correct but does not match the figure)
- p. 118 Proof of Thm 2.45. Second line reads “thef our point condition holds.” Should be “the four point...”
- p. 120, line (3.43): Subscript on  $x$  should be  $j_r$ , not  $r$ . (doublecheck)
- p. 121, 2/3 of way down: “from Theorem 2.41. See Chapter 18.” 2.41 was an Algorithm, not a Theorem.
- p.123 Last line before Prop 3.49, reads “complete subgroup  $K_{2m}$  has...”. Should read “complete subgraph...”
- p. 128, in Table 4.1 and the text that follows, you left out start codons.
- p. 129: You need to define  $u_{+j}$ ,  $u_{i+}$ ,  $u_{++}$ . Using “+” to sum over an index instead of a dot is not standard (but even summing with a dot should be explained). You do define it much later, on page 286.
- p. 131: “In [Tesler, 2002], it is shown that ... can be computed in time polynomial” It was shown that distance (as opposed to exhibiting an actual sequence of rearrangements) can be computed in \*linear\* time. Hannenhalli and Pevzner had already shown polynomial time for the distance.

- p. 139, line -1: missing space in “thatis”
- p. 140, line -1: missing space in “ofcomparative”
- p. 154 and p. 293: Different notations for the Felsenstein hierarchy
- p. 155, line -9: “(Lemma 2.32)”: 2.32 is an “Example”. I think you might mean “(Lemma 2.33)”, since Lemma 2.33 shows that the number of tree types is huge. If that’s it, correct it to say “Lemma 2.33” and add an explanation of why Lemma 2.33 is relevant to the claim that it is infeasible to inspect all trees.
- p. 168 13 lines down from the top... both words are indexed to have n letters.
- p. 169 first line... MAXIMUM negative log probabilitiy?
- p. 171 in  $P^D(i, 0)$  the first theta has bad indices.
- p. 172, line -1: Just the opposite of the usual last line of page problem! Here there is a SUPERFLUOUS SPACE between the “)” and “,” in “deletion) , the”
- p. 175, 18 lines down from the top... there is a big D index (in the image of phi) that should be a little d.
- p. 197, last line: missing space in “algebrafrom”
- p. 198, line -1: missing space in “ofthe”
- p. 220: “2-parameter model for sequence alignment. Note that it actually comes from a 3-parameter model...” It seems to me that you’ve covered models with many more parameters and so it doesn’t specifically come from a 3-parameter model, but could be any of the other models with more parameters as well.
- p. 261: The superscript notation  $\theta^t$  ( $t$  an index, not a power) is potentially confusing, here and elsewhere. It’s not the first place it happens in the book, but it’s potentially more confusing here than before because you talk about the Taylor expansion (in which powers normally appear). I think the places you use sub and superscripts both are probably ok, but things like this should possibly be done with different notation. Either subscripts, or adorned superscripts like  $\theta^{(t)}$   $\theta^{[t]}$  or something.
- p. 261, 2nd displayed equation: It’s a big leap from the EM explanation given on pages 17- to the formula

$$\operatorname{argmax}_{\theta} E[\ell_{obs}(\theta|\tau, \sigma)|\sigma, \theta^t]$$

Fill in some steps, either in the discussion on pages 17- or else here.

- p. 268, middle: “Equation 1.66” → “Equation (1.66)”
- p. 273, middle: “Equation 13.3” → “Equation (13.3)”
- p. 283, middle: “Lemma 14.6” → “Theorem 14.6”
- p. 292: “r” denotes the root vertex and ALSO denotes the root distribution vector. These are incompatible uses.
- p. 305, strand symmetry: The two strands are related by reverse complement, not just complement. So I am confused as to why it’s “ $\theta_{AC} = \theta_{TG}$ ” etc. rather than “ $\theta_{AC} = \theta_{GT}$ ” (since the reverse complement of AC is GT,

and the plain complement is TG). Does strand symmetry only take into account the complementation, not the reversal?

- p. 306 line 1: HUGE space in “i.e. Jukes” It was formatted as though it’s the end of the sentence. Use “i.e.slash Jukes” for a normal sized space (breakable) or “i.e.tildeJukes” (unbreakable).
- pp. 307, 308, 309, “IntV(T)” notation: sometimes it’s in roman, sometimes it’s in math italic, and the spacing between “Int” and “V” is poor. Define a macro (or some other name if you’re concerned about a conflict with the integral symbol slash int) and use “Int V(T)”. That should force it to roman and put decent spacing between “Int” and “V(T)”.
- p. 339, line -1: missing space in “Inparticular”
- p. 343, Theorem 18.9: Use big [] (with slash left[ ... slash right] or slash Bigl[... slash Bigr] or something along those lines)
- p. 343, line -2: “equation (18.1)” → “Equation (18.1)”
- p. 344, line -1: missing space in “theapplication”
- p. 346, line 5: “equation (18.1)” → “Equation (18.1)” Again, if you used a consistent scheme for labelling equations, you can make a perl script that scans the .tex files to make sure equations are referenced in a consistent format.
- p. 371, after bulleted “Iteration:”: missing space in “Step1”
- Notation varies chapter to chapter (# matches, # mismatches, # spaces):  
p. 195:  $(m_h, x_h, s_h)$  p. 207:  $(z, x, 2y)$  p. 220:  $(z, x, \tilde{y})$  (at least this one explains relation to p. 207)
- Chapter 21 has HUGE spaces before citations: p. 385: “e.g., [Huelsenbeck ...]” p. 386: “in [Yoder ...]” p. 386: “in [Douzery ...] I suspect that they misused “tildeslash cite”, something like this: “in tildeslash citeHuelsenbeck” (both a space AND a “tilde”) OR they put “in” at the end of one line and “ slash cite” at the start of the next line. Again, make a perl script to scan for misformatting of slash cite’s. Make sure it’s always XXXtildeslash citeYYY and flag all occurrences of slash cite that do not appear in that fashion.
- References, pp. 403-417: When there are multiple pages, it says “p. a, b, c” Instead, it should say “pp. a, b, c.” (use of “pp.” instead of “p.”, and also for both one or many pages it should terminate the sentence with a period) Also, if you wrote it yourselves, perhaps when it’s a range of pages, it could be abbreviated “pp. a–b, c, d–e.” or whatever.
- The reference [Garcia et al. 2004] should be [Garcia et al. 2005].
- There is an inconsistency in the definitions of “combinatorial tree” throughout the book.

(1) Trees as graphical models, in Chapter 1. A tree is a directed graph which, if the directions of the arrows are erased and it is viewed as an undirected graph, is connected and acyclic, and has its degree-1 nodes labelled (though we must in some sense also label interior nodes here to deal with their states). So interior nodes may have arbitrary degree  $\leq 3$ , and if there are  $n$  leaves, each is labelled by one element of  $[n]$ .

(2) Tree metrics: Example 2.32, Thm 2.36 (Four-Point Condition), Proposition 2.37, Lemma 2.38, Thm 2.39 (Cherry-Picking), Cor 2.40, Algorithm 2.41, Thm 2.42, all of section 3.5. A tree is a \*trivalent\* tree with labelled leaves, and if there are  $n$  leaves, each is labelled by one element of  $[n]$ . If we allow trees that have degree 2 nodes as in (1), the dimensionality is not the number of edges, there are infinitely many combinatorial types on  $[n]$  nodes, etc. If we allow multiple labels, we get extra combinatorial types and some cones with smaller dimension.

(3) Neither of these work for the Splits Equivalence Theorem (Thm 2.35). For example, the collection of 2 splits on 5 nodes 12,345,1,2345 is pairwise compatible, and there exists a trivalent tree  $T$  such that  $S$  is contained  $\text{Splits}(T)$ , but it is not unique, and there is no trivalent tree such that  $S = \text{Splits}(T)$ . Thus the theorem seems to require that we use  $X$ -Trees as in Semple and Steel, p. 16. (connected acyclic undirected graphs where nodes of degree less than three are labelled by disjoint nonempty subsets of  $[n]$ ).

Corrections have been submitted by William Felder, Joe Felsenstein, Luis Garcia, Jason Morton, Glenn Tesler and Debbie Yuster. Special thanks to Glenn Tesler for his careful reading of the book.