Reliable Editions from Unreliable Components: Estimating Ebooks from Print Editions Using Profile Hidden Markov Models

A. B. Riddell
Indiana University Bloomington
United States

ABSTRACT
A profile hidden Markov model, a popular model in biological sequence analysis, can be used to model related sequences of characters transcribed from books, magazines, and other printed materials. This paper documents one application of a profile HMM: automatically producing an ebook edition from distinct print editions. The resulting ebook has virtually all the desired properties found in a publisher-prepared ebook, including accurate transcription and an absence of print artifacts such as end-of-line hyphenation and running headers. The technique, which has particular benefits for readers and libraries that require books in an accessible format, is demonstrated using seven copies of a nineteenth-century novel.

CCS CONCEPTS
• Information systems → Digital libraries and archives.

KEYWORDS
profile hidden Markov model, digital libraries, ebooks, sequence alignment, digital editions, optical character recognition

ACM Reference Format:
A. B. Riddell. 2022. Reliable Editions from Unreliable Components: Estimating Ebooks from Print Editions Using Profile Hidden Markov Models. In The ACM/IEEE Joint Conference on Digital Libraries in 2022 (JCDL ’22), June 20–24, 2022, Cologne, Germany. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3529372.3533292

1 INTRODUCTION
When publishers bring a book to market today, they offer two versions for sale, a print version and an ebook version. In most cases, a reader encountering a print version and an ebook version at the same time will notice few differences. The sequence of characters visible on the page and on the e-reader will be identical, save for running headers, page numbers, end-of-line hyphenation, and page breaks—features exclusively found in print versions. Although the sequence of characters in each version is virtually identical, the two versions differ in significant ways. As ebooks abide in computer memory, the ebook version costs almost nothing to store, transport, or reproduce. Ebooks are also more accessible to readers with blindness and print disabilities as they can be read using text-to-speech and Braille devices.

This paper describes how to automatically estimate an ebook from several print copies using a profile hidden Markov model (profile HMM), a model introduced in biological sequence analysis in 1993 [2, 3, 5]. Unlike texts produced from page images via optical character recognition (OCR), ebooks estimated using the method described here are potentially indistinguishable from publisher-prepared ebooks. I demonstrate the use of a profile HMM to estimate an ebook version of a chapter of Charles Dickens’ David Copperfield (1850) using seven print copies of the novel. The performance of this method is evaluated by calculating the percentage of matching characters in an alignment between the estimated ebook version and a reference ebook. The profile HMM produces an ebook version that aligns as well with the reference ebook as the Project Gutenberg ebook, a version on which human editors labored for tens if not hundreds of hours.

2 COMPARISON OF TRANSCRIBED BOOK PAGES AND AN EBOOK
Here I introduce terminology and give a concrete example of the differences between transcribed book pages and the machine-readable contents of an ebook.

In October 2020, the London-based publisher Verso made Culture and Materialism by Raymond Williams available for purchase in two formats: paperback (ISBN 9781788738606) and ebook (ISBN 9781789600049). The contents of these two formats, if suitably prepared, can be compared. The paperback version must first be transcribed into a machine-readable form. That is, for each discrete mark on each page of the paperback book, a counterpart must be chosen from the ca. 140,000 characters in the Unicode Standard. For example, a mark resembling “a” would likely be recorded using the character “a”, U+0061 (“Latin Small Letter A”). Horizontal space, line breaks, and page breaks are transcribed using standard conventions: U+0020, U+000A, and U+000C, respectively. Marks on the page, including illustrations, which are not described in the Unicode Standard are ignored. Transcriptions of a page will vary because human and machine transcribers make inadvertent mistakes and enjoy freedom to choose among alternatives. Less work is required to prepare the ebook version for comparison. Because a computer file using the EPUB 3 format consist of files stored in a container (ZIP), the ebook must be uncompressed to get at the enclosed XHTML content documents.

To facilitate comparison among ebook versions of similar works and ebooks of the same work in different formats (e.g., EPUB 2, Mobipocket, HTML, CHM), the XHTML content documents are mechanically converted to HTML5 and stripped of redundant markup and whitespace. At this point we have two sequences of characters that can be compared. I will refer to a machine-readable transcribed...
sequence as a "print sequence". I will refer to an HTML5 sequence extracted from an ebook as an "ebook sequence".

If we align a print sequence obtained via OCR and an ebook sequence of *Culture and Materialism*, we find considerable agreement. Characters in the two sequences match most of the time and virtually all of the differences are predictable. Relative to the ebook sequence, the print sequence inserts running headers, page numbers, end-of-line hyphenation, line breaks, and page breaks. Relative to the print sequence, the ebook sequence inserts HTML tags indicating paragraph breaks (<p>) and emphasis (<em>).

### 3 PROFILE HIDDEN MARKOV MODEL

To estimate an ebook sequence, homologous print sequences are collected and then used to estimate a profile hidden Markov model (profile HMM). (*Homology* is similarity owing to common descent.) The fitted profile HMM’s most probable sequence—its mode or modal sequence—is then used as the estimated ebook sequence.

The profile HMM has been a staple of biological sequence analysis since 1993 [2]. The procedure described here is a trivial extension of prior art.

#### 3.1 Background and model architecture

The profile HMM provides a probabilistic approach to the problem of multiple alignment. In the context in which it was introduced, the sequences being aligned were proteins and RNAs. (The "profile" in the model’s name is a biology-specific reference.) A typical use of a profile HMM in biology is identifying potentially homologous proteins, such as oxygen-carrying (globin) proteins in different mammals. These proteins—sequences of amino acids—tend to resemble each other. They are, however, far from identical. A profile HMM provides the needed method for characterizing differences and commonalities among sequences.

The architecture of the profile HMM (Figure 2) is best understood by reference to a multiple alignment, such as the one shown in Figure 1. The profile HMM is an inhomogeneous HMM—its hidden states are position-specific. In essence, each state is associated with a column of the multiple alignment. Amino acid residues (or, in our case, characters) associated with consensus columns are emitted by *match* states. *Insert* states at each position allow for the possibility that a particular sequence may have one or more idiosyncratic insertions relative to the sequence of *match* states. *Delete* states allow for a particular sequence to lack a residue (character) at a position where other sequences tend to have one. *Delete* states do not emit a residue (character)—they are "silent states".

Transitions between states are constrained to flow in one direction, as arrows in Figure 2 indicate. A *match* or *delete* state, once visited, is never re-visited. An *insert* state can be revisited, but only from the *insert* state itself. Dwelling in an *insert* state allows for a series of idiosyncratic residues (characters) to be emitted—e.g., a running header found in only one print sequence.

Figure 2 shows the profile HMM architecture for a model with three *match* states. The three additional special states (labeled "b", "i0", and "e") are used to model the beginning and ending of an observed sequence. The first *insert* state ("i0") allows for the observation of idiosyncratic insertions before observing an emission associated with the first *match* state.

Beyond expanding the alphabet to accommodate characters that occur in print sequences, no modification of the profile HMM is required to use it in the context of print sequences.

For those coming to profile HMMs from computer science or computational linguistics [4, 6], a distinctive feature of the profile HMM is its numerous hidden states. The large number of states creates no theoretical issues but it does pose a computational problem. In general, when using the forward algorithm to calculate the probability of an observation of length $T$ in an HMM with $N$ hidden states, $T N^2$ calculations are required [4, p. 255]. In the case of the profile HMM, $T \approx N/3$.

Those familiar with biological sequence analysis will note that there are other variants of the original profile HMM—a model introduced in Haussler et al. [3] and Krogh et al. [5]. In this paper, I only consider the problem of global alignment and "profile HMM" only refers to the original profile HMM or to its streamlined version, the "Plan 7" profile HMM introduced by Eddy [2].

#### 3.2 Estimating a profile HMM’s parameters

A profile HMM is an ordinary hidden Markov model whose parameters may be estimated using standard Bayesian methods. Typically maximum a posteriori parameter estimates are obtained using expectation maximization (aka Baum-Welch) or gradient ascent. The number of hidden states in the profile HMM—proportional to "model length" (the number of *match* states)—must also be picked in advance.
The same procedures for fixing model length and choosing initial parameters can be used for character sequences as for amino acid sequences. Durbin et al. [1] make several recommendations and describes in detail how to estimate a profile HMM’s parameters.

4 EXPERIMENT: ESTIMATING AN EBOOK EDITION OF DAVID COPPERFIELD

In order to verify that the procedure works as expected, I estimate a profile HMM for a randomly-chosen chapter of David Copperfield using seven homologous print sequences. The model’s modal sequence is then compared with a reference ebook version using the usual pairwise sequence identity metric.

I estimate the parameters of the profile HMM by using statistics derived from a guess at a multiple alignment. I use the non-probabilistic Barton–Sternberg iterative multiple alignment algorithm to arrive at the initial guess [1, p. 149–150]. I then use a simple heuristic to identify match states: any column that has less than 50% gap symbols is associated with a match state [1, p. 123]. Given a multiple alignment with match states marked, serviceable approximations of the maximum a posteriori estimates of all the profile HMM’s parameters—the emission and transition probability distributions—can be obtained [1, p. 123-124].

The length of model (16,783) makes using Baum-Welch or gradient ascent prohibitively expensive for a naive implementation. So, for the moment, I do not perform any gradient ascent steps. I use the initial parameter estimates without any further fitting. More careful estimation of model parameters will certainly be required when only two different print copies are available as observations. Here, with seven copies, it seems unlikely to make a significant difference. To check this claim, I estimate a smaller model of the initial paragraph (length 625) using the same seven copies, allowing gradient ascent to run for 10 epochs. No change to the modal sequence is observed.

I use a 107-character alphabet for all emission distributions. This alphabet includes characters frequently found in publisher-prepared ebook editions and an opportunistic sample of Internet Archive print sequences: other David Copperfield chapters as well as chapters from print sequences from George Eliot’s The Mill on the Floss (1860). When calculating empirical distributions from sequences that contain out-of-alphabet characters, such characters are replaced with spaces. Code and data are available at https://doi.org/10.5281/zenodo.6412406.

4.1 Seven copies of David Copperfield

In this experiment, I work with seven copies of Charles Dickens’ David Copperfield (1850). I use a single chapter—Chapter 29—selected uniformly at random, rather than the entire book to reduce computation time. David Copperfield was selected because it has many distinct public domain editions.

In the experiment I use character sequences from seven distinct copies of David Copperfield (Table 1). These seven sequences are associated with six editions of David Copperfield. The two sequences from first edition copies come from distinct items located in different university libraries. In all cases I use the machine-transcribed plain text sequences obtained from the digital library that hosts images of the item’s pages (Internet Archive or Hathi Trust).

| Year | Publisher | Identifier | Short Name |
|------|-----------|------------|------------|
| 1850 | Bradbury & Evans | personalhistoryo001850dick | 1850-LON-1 |
| 1850 | Bradbury & Evans | duliark:/13960/t9pc47v32 | 1850-LON-2 |
| 1858 | Chapman & Hall | personalhistoryo00dickala | 1858-LON |
| 1866 | Chapman & Hall | coo:319240313074142 | 1866-LON |
| 1867 | Ticknor & Fields | aa:4786.0001.001.umiich.edu | 1867-BOS |
| 1880 | T. B. Peterson | davidcopperfield01dicke | 1880-PHL |
| 1910 | Musson Book Co. | personalhistoryo00dickuod | 1910-TOR |

All day, she seemed to pervade the whole house. If All day, she seemed to pervade the whole house. If

I talked to the “304 THE PERSONAL HISTORY AND EXPERIENCE Steerforth in his room, I heard her dre Steerforth in his room, I heard her dre

Figure 3: Alignment of “1850-LON-2” with “1866-LON” (excerpt). Lines in the figure are wrapped for display purposes. “ ” is used as the “gap” character.

Although every print sequence is distinct, sequences tend to use the same characters in at least 90% of positions. Differences tend to involve the insertion or deletion of whitespace, typically due to edition-specific pagination or use of different transcription software. For example, although the aligned 1850 first edition sequences (“1850-LON-1” and “1850-LON-2”) differ in 528 positions, 369 (70%) of these differences involve whitespace characters.

4.2 A reference ebook sequence of the first edition of David Copperfield

Dickens’ publisher, Bradbury & Evans, did not publish an ebook version alongside the print version of David Copperfield in 1850. Here I describe the construction of an ebook that approximates what Bradbury & Evans would have published had the practice of publishing ebook editions existed in 1850. This reference ebook sequence serves as the benchmark or “gold standard” against which an estimated ebook sequence will be judged.

Creating such a reference ebook edition of David Copperfield is made easy by the fact that ebook editions of David Copperfield already exist that have been painstakingly prepared by established publishers such as Penguin Classics, Oxford World’s Classics, and Project Gutenberg. Given that these ebook editions are intended for use as substitutes for the first edition, they give us considerable information about what characters should appear in the reference ebook sequence. Starting from any one of these ebooks, the only step required to arrive at the reference ebook sequence is to verify that the characters in the ebook sequence have corresponding characters in the 1850 first edition. With page images of the first edition in hand, doing this for a single chapter is not onerous. (As we are only estimating a single chapter, Chapter 29, the reference ebook need only contain that chapter). Typically only a handful of changes are required. To construct the reference ebook I begin
Table 2: Pairwise sequence identity with the reference ebook sequence. A near miss is a mismatch without semantic consequence (see text).

| Short name    | % match ignoring HTML tags | % match ignoring HTML, near misses |
|---------------|-----------------------------|-----------------------------------|
| Profile HMM mode | 95.45                      | 97.11                             |
| Oxford ebook  | 99.63                      | 99.96                             |
| Penguin ebook | 97.87                      | 99.13                             |
| PG ebook      | 98.26                      | 99.64                             |
| 1850-LON-1    | 92.67                      | 94.23                             |

with the Project Gutenberg edition and make edits such that the ebook sequence aligns with the 1850 first edition.

5 RESULTS

Table 2 shows measurements of pairwise sequence identity between the reference ebook sequence and other sequences, including the modal sequence. The third column of Table 2 reports pairwise sequence identity ignoring HTML tags and near misses. A near miss is a mismatch that is not semantically consequential involving an inserted space, typesetter’s apostrophe, or typographic ("curly") quotation marks. For example, a mismatch between a typewriter apostrophe ’ (U+0027) and a typesetter’s apostrophe ’ (U+2019) is a near miss. So is the insertion of a space (U+0020) in one of the sequences. The first column in Table 2 reports pairwise sequence identity without any adjustments.

When considering raw mismatch statistics (vs. the reference ebook) it is essential to ask what characters are involved. A mismatch involving two letters tends to be more consequential for a reader’s experience than a mismatch involving an extra space inserted after a period. Even letter mismatches can vary in their impact.

The most important thing to note when comparing the modal sequence with the reference ebook sequence is that, with one exception, no mismatches involve letters. The mismatches all involve punctuation (quotation marks, em-dashes, commas). And the single exception is telling. It involves a correction to the 1850 first edition that appears in a majority of the print sequences from later editions.

A shortcoming of the method when used on plain text print transcriptions (rather than transcriptions that attempt to mark paragraph boundaries) is that the profile HMM will occasionally preserve a line break if a majority of print sequences use an intra-paragraph line break at the same position. Across hundreds of lines, this can happen by chance. It happens 16 times in the modal sequence. If these 16 mismatches are excluded—and HTML tags and near-misses ignored—the pairwise sequence identity between the modal sequence and the reference ebook is essentially the same as that between the Project Gutenberg ebook and the reference ebook.

6 LIMITATIONS

The chief limitation is that the method requires at least two print editions with distinct pagination. Other limitations involve HTML tags that are present in the reference ebook but are missing in the modal sequence. Not only are these HTML tags missing in the estimated ebook version, the fitted profile HMM will assign lower probability to a sequence with these HTML tags than to one without them. Fortunately, these limitations are either easy to resolve or of bounded consequence.

The main limitation of the method, that it requires at least two distinct editions, does not render the method useless. Books that were popular around or after the time they were published frequently have at least two editions. Most of these (once) popular books lack editions in an accessible format. So the procedure using a profile HMM can still facilitate the creation of many accessible ebooks from print copies.

The failure of the profile HMM to assign higher probability to ebook sequences with the correct HTML tags (e.g., <p>, <em>, <h2>) has a ready solution: use a more descriptive (machine) transcription. Using a transcription of the printed page that includes additional information about line spacing and typeface changes would solve the problem of missing paragraph, section heading, and emphasis tags. For example, had the experiment described above used a more detailed transcription instead of plain text, paragraph tags would already have been present in the print sequences because the relevant format (hOCR) uses XHTML, which marks paragraph boundaries in the same manner as the reference ebook sequence. So the profile HMM would have included three match states at the beginning of each paragraph capturing the observed regularity of the three-character sequence (<, p, >) tending to occur in the same position across several editions.

Illustrations, figures, and bespoke glyphs pose no theoretical problem for an approach using a profile HMM because these elements of the text can be "transcribed" using PETSCII, SVG, or embedded using a data URI. As a practical matter, however, each such transcription is likely to be unique and will never be associated with a match state—only with an insert state. Some sort of special handling is likely to be required.

7 DISCUSSION

This paper offers an eye-catching demonstration of one potential use of a profile HMM. Due to space constraints, I can only gesture at the myriad potential uses of profile HMMs in the study of transcribed sequences of books and periodicals. Since their introduction in 1993, profile HMMs have been used in numerous different tasks in biology [2]. For every task in biology, it is easy to imagine an analogous task of interest to researchers in information retrieval, book history, and media studies. For example, the task profile HMMs initially addressed was search [3, 5]. Profile HMMs proved capable of identifying closely- and distantly-related proteins and RNAs. Translated into the context of information retrieval, profile HMMs used in this way would identify (inexact) copying of particular paragraphs in other works.
8 CONCLUSION

A profile hidden Markov model of character sequences can be used to estimate an ebook edition from several print copies of a given work. The resulting ebook has the desired properties we anticipate in a publisher-prepared ebook. This technique is successful because the ebook edition features only those elements shared by a majority of the print copies. The profile HMM models these shared elements using match states. The model uses insert states to model elements found in only one or two print editions, elements such as running headers, page numbers, end-of-line hyphenation, and page breaks. As the profile HMM assigns higher probability to sequences that tend to feature only shared elements, high probability sequences under the model tend to resemble an ebook sequence. The paper verifies this through an experiment using seven copies of a nineteenth-century novel.

REFERENCES

[1] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press.
[2] Sean R. Eddy. 1998. Profile Hidden Markov Models. Bioinformatics 14, 9 (1998), 755–763.
[3] David Haussler, Anders Krogh, I Saira Mian, and Kimmen Sjolander. 1993. Protein modeling using hidden Markov models: Analysis of globins. In Proceedings of the Twenty-sixth Hawaii International Conference on System Sciences, Vol. 1. IEEE, 792–802.
[4] Biing Hwang Juang and Laurence R Rabiner. 1991. Hidden Markov models for speech recognition. Technometrics 33, 3 (1991), 251–272.
[5] Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sjölander, and David Haussler. 1994. Hidden Markov models in computational biology: Applications to protein modeling. Journal of molecular biology 235, 5 (1994), 1501–1531.
[6] Wesley Mackay and Grzegorz Kondrak. 2005. Computing word similarity and identifying cognates with Pair Hidden Markov Models. In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). 40–47.