doi:10.1016/j.jmb.2003.08.015

J. Mol. Biol. (2003) **333**, 453–460

JMB

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Sequence Alignments and Pair Hidden Markov Models Using Evolutionary History

Bjarne Knudsen* and Michael M. Miyamoto

Department of Zoology, Box 118525, University of Florida Gainesville, FL 32611-8525 USA

**Corresponding author*

This work presents a novel pairwise statistical alignment method based on an explicit evolutionary model of insertions and deletions (indels). Indel events of any length are possible according to a geometric distribution. The geometric distribution parameter, the indel rate, and the evolutionary time are all maximum likelihood estimated from the sequences being aligned. Probability calculations are done using a pair hidden Markov model (HMM) with transition probabilities calculated from the indel parameters. Equations for the transition probabilities make the pair HMM closely approximate the specified indel model. The method provides an optimal alignment, its likelihood, the likelihood of all possible alignments, and the reliability of individual alignment regions. Human α and β-hemoglobin sequences are aligned, as an illustration of the potential utility of this pair HMM approach.

Keywords: evolution; pair hidden Markov models; statistical alignment; insertion and deletion models

Introduction

Algorithms for aligning sequences have been used for decades[1,2] and are mostly based on specified scores for matching residues and penalties for gaps. These approaches have proven very useful but lack an explicit evolutionary model. Furthermore, the relationship between the gap penalties and evolutionary distances between sequences is not entirely clear.

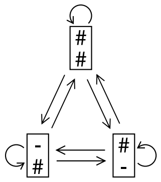
In the area of statistical alignment, there has been a number of efforts to develop explicit models for the insertion and deletion (indel) process.[3–6] However, the main problem with most existing statistical models is that indel lengths are fixed to one, contrary to what is often observed in biological sequences.[7] Consequently, the work presented here specifies an indel process with geometrically distributed indel lengths, which is a significant improvement. The model is approximated with a pair hidden Markov model (HMM) for which efficient calculations can be made.[8] This leads to an approximate statistical alignment algorithm with geometrically distributed indel lengths.

The pair HMM algorithm takes two aligned sequences, finds maximum likelihood (ML) estimates of all substitution and indel parameters by summing over all alignments, and returns the optimal alignment and the reliability of the individual alignment positions. Our HMM algorithm for pairwise sequence alignments with geometrically distributed indels using evolutionary history is provided as a web server†.

Theory

Pair HMMs

A pair HMM can generate two sequences at the same time according to the following model:[8]



The rectangles represent states emitting residues (#) for one or both sequences. The state emitting residues for both sequences is called the match state, while the two other states are called indel

Abbreviations used: HMM, hidden Markov model; ML, maximum likelihood; JTT, Jones, Taylor and Thornton; TKF, Thorne, Kishimo and Felsenstein.

E-mail address of the corresponding author: bk@birc.dk

† http://www.daimi.au.dk/~compbio/pairhmm

states (with "−" referring to a gap). There is also a start and an end state, which are not shown here.

Let us designate the match state 0, and the respective indel states 1 and 2. The start state is denoted s, and the end state e. Each transition has a probability that will depend on the indel process and the evolutionary distance between the two sequences. We can write the transition probabilities as:

$$T = \begin{bmatrix} T_{s0} & T_{s1} & T_{s2} & T_{se} \\ T_{00} & T_{01} & T_{02} & T_{0e} \\ T_{10} & T_{11} & T_{12} & T_{1e} \\ T_{20} & T_{21} & T_{22} & T_{2e} \end{bmatrix}$$

This study relies on a pair HMM to approximate an explicit evolutionary model. This approximation is accomplished by fitting the transition probabilities to an indel model, while the emission probabilities are found using well known nucleotide and amino acid models (see below).

Nucleotide substitutions and amino acid replacements

A standard nucleotide substitution or amino acid replacement model is used for the residues in the sequences. For DNA, this could be (for example) the HKY model[9] or the general reversible model.[10] Here, the widely accepted JTT model is used for proteins.[11]

For an indel state, the emission probability of a residue is given just by the frequency of that residue under the substitution or replacement model. For a match state, the joint probability of the residues is calculated from the model and the evolutionary time, t, between the two sequences. The time is measured in units of expected substitutions or replacements and is ML estimated.

The indel process

Here, we develop a model of the indel process in biological sequences. For a given sequence of length L, let us assume that an insertion can occur between any two residues. Furthermore, assume that an insertion can occur at the start or at the end of the sequence giving $L+1$ possible sites for insertions. Insertions occur at the same rate, r, throughout the sequence, so rt is the expected number of insertion events between any two neighboring residues over time t.

Let the insertion length follow a geometric distribution with parameter a. This means that the probability of an insertion of length $i \geq 1$ is:

$$P_i = (1-a)a^{i-1}$$

For practical purposes, a time reversible indel process is accepted. Given this insertion process, one now only needs to specify a sequence length distribution to determine the entire model (in the time reversible model, the rate ratio of complementary insertions and deletions is equal to the ratio of the resulting sequence length probabilities).

Given that the goal here is to find a widely applicable indel process, we assume that all sequence length probabilities are the same at equilibrium. Denote this probability δ. Since sequences can be arbitrarily long, all sequence lengths have probability $\delta = 0$ (this value of δ turns out not to be a significant problem; see Transitions out of the start state and to the end state, below).

The rate of any insertion between two given residues is r, so the rate of insertions of length i is rP_i. For each possible deletion of length i, the rate is the same, rP_i, due to the time reversibility. For a sequence of length L, there are $L-i+1$ different deletions of length i. The rate of going from a sequence of length L to a sequence of length $L+i$ is $(L+1)rP_i$, because there are $L+1$ possible sites for insertions. The rate of going the other way is $(L+i-i+1)rP_i$, i.e. the same.

Approximating the indel process with a pair HMM

Since the sequence length distribution is uniform, the transition probabilities to the end state should be zero (or proportional to δ), so they will not be considered before the end of this section.

Assume that we are dealing with two long sequences with an evolutionary time of t between them. Now we will discuss internal residues where no edge effects in the deletion process are apparent from the ends of the sequences.

Indel events

Insertions occur between two given residues at the rate r and deletions start at a given residue at the same rate. Now consider a position in the alignment of the two sequences. The probability that an indel event occurs between this position and the next in time t is $1-\mathrm{e}^{-2rt}$. Given that an indel occurs at this position, the probability that another one occurs in the same time range is approximately:

$$\int_{s=0}^{t} \frac{1}{t}(1-\mathrm{e}^{-2rs})\mathrm{d}s = \frac{1}{t}\left(t - \left[\frac{1}{-2r}\mathrm{e}^{-2rs}\right]_{s=0}^{t}\right)$$

$$= 1 - \frac{1}{2rt}(1-\mathrm{e}^{-2rt})$$

This is an approximation, since it assumes that the time of the first event is uniformly distributed in time between the two sequences (the $1/t$ term). The probability of an indel event can be written as:

$$P_{\mathrm{id}} = 1 - \mathrm{e}^{-2rt}$$

Given one event, the probability of another event can be written as:

$$P'_{\mathrm{id}} = 1 - \frac{1}{2rt}(1 - \mathrm{e}^{-2rt})$$

In the calculations performed here, we will only consider single and double events. Double events are important when approximating the transition probability from one indel state to the other in the pair HMM. The approximation should be sufficiently accurate, since three or more overlapping indels are very rare for biological sequences that are similar enough to be alignable.

Transitions out of the match state

For a given position that is in the match state, there is a probability, P_{id}, that an indel occurs before the next position. If a second indel then does not occur, that state will be recognizable as an indel state. However, if another indel occurs, there is a probability that it will remove the effect of the first event (e.g. an insertion followed by a deletion of the same length in the same place will be invisible). To calculate these probabilities, a tree can be used to keep track of the contributions from different events (see Figure 1).

The probability of a deletion and an insertion having the same length is:

$$P(=) = \sum_{i=1}^{\infty} P_i P_i = \frac{(1-a)^2}{1-a^2} = \frac{1-a}{1+a}$$

The probability that the deletion is shorter or longer, respectively, is:

$$P(<) = P(>) = \frac{1}{2}(1 - P(=)) = \frac{a}{1+a}$$

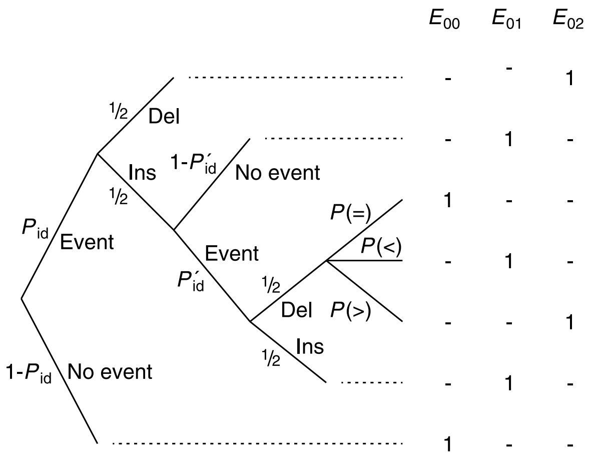


Figure 1. Tree for calculating the transition probabilities out of the match state. The right side represents the expected numbers of different transitions given the events on the branches. In this case, only one transition out of the match state is expected for each line of events. Probabilities for the different cases are to the left of the branches, while a short description is to the right. The probabilities, $P(<)$, $P(=)$, and $P(>)$, represent the probabilities that the length of a second deletion (Del) is shorter than, equal to, or longer than an initial insertion (Ins), respectively.

Using the tree in Figure 1, the expected number of match to match transitions can be determined as:

$$E_{00} = (1 - P_{\mathrm{id}}) + P_{\mathrm{id}} \frac{1}{2} P'_{\mathrm{id}} \frac{1}{2} P(=)$$
$$= 1 - P_{\mathrm{id}} \left(1 - \frac{1-a}{4+4a} P'_{\mathrm{id}}\right)$$

Since the total expected number of transitions is $E_0 = 1$, the transition probability from a match state to a match state is $T_{00} = E_{00}$. From symmetry, we have:

$$T_{01} = T_{02} = \frac{1}{2}(1 - T_{00})$$

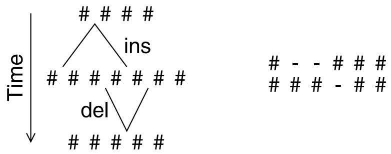
Notice that the numbers for E_{01} and E_{02} from Figure 1 are not exactly identical because evolution is viewed from one sequence to the other. This means that the transition probabilities should be calculated as:

$$T_{01} = T_{02} = \frac{E_{01} + E_{02}}{2E_0}$$

Transitions out of the indel states

When in an insertion state, there is a probability of $1 - a$ that the insertion will stop before the next site. If this happens, a new insertion or deletion at that position can influence the nature of the next state. If the insertion continues, there is still a chance for the new indel to influence the next state (Figure 2).

There are two situations in which T_{12} and T_{21} come into use. The first is an insertion followed by an overlapping or adjacent deletion (resulting alignment to the right):



When in an insertion state, the probability of this occurring is (Figure 2):

$$aP'_{\mathrm{id}} \frac{1}{2} P(>) = \frac{a^2 P'_{\mathrm{id}}}{2(1+a)}$$

The second example is a deletion followed by an insertion in the same place (the three possible

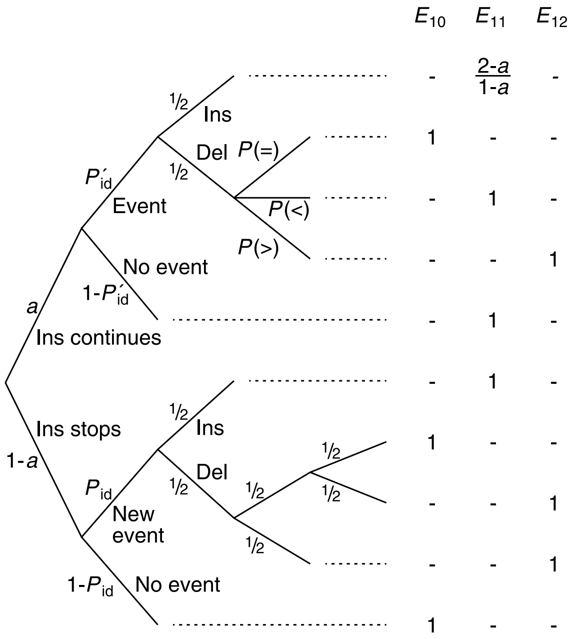
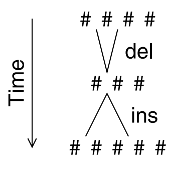


Figure 2. Tree for calculating the transition probabilities out of an insertion state. The right side again represents the expected numbers of different transitions, given the events on the branches. Notice that if an insertion continues and is followed by another insertion, numerous insertion to insertion transitions are expected (the expected insertion length plus one). If the insertion stops and is followed by a deletion, there is a probability of 1/2 that the deletion occurred first (in time), and in this case there is a probability of 1/2 that the deletion is placed before the insertion in the alignment (giving a contribution to E_{10}, rather than E_{12}). This is described further in the text.

resulting alignments are given to the right):



```
# # - - # #
# - # # # #

# - # - # #
# # - # # #

# - - # # #
# # # - # #
```

This hypothetical situation illustrates a limitation of alignments to describe evolutionary history: for a given evolutionary history, there is not always a unique alignment corresponding to it. When choosing the appropriate alignment, the following convention is used here: fragmentation of the indels should be minimized, so either the top or bottom alignment would be chosen for the above example.

The expected numbers of observed transitions can be found using the tree in Figure 2. All possible paths from left to right in the tree are followed and probabilities are found and added up to give the corresponding observations:

$$E_{10} = aP'_{\text{id}}\frac{1}{2}P(=) + (1-a)\left(P_{\text{id}}\frac{1}{2}\frac{1}{2}\frac{1}{2} + (1-P_{\text{id}})\right)$$

$$= (1-a) + \frac{a(1-a)}{2+2a}P'_{\text{id}} - \frac{7-7a}{8}P_{\text{id}}$$

$$E_{11} = a\left(P'_{\text{id}}\left(\frac{1}{2}\frac{2-a}{1-a} + \frac{1}{2}P(<)\right) + (1-P'_{\text{id}})\right)$$

$$+ (1-a)P_{\text{id}}\frac{1}{2} = a\left(1 - P'_{\text{id}}\left(1 - \frac{1}{2}\left(\frac{2-a}{1-a}\right.\right.\right.$$

$$\left.\left.\left. + P(<)\right)\right)\right) + \frac{1-a}{2}P_{\text{id}}$$

$$= a + \frac{a^2}{1-a^2}P'_{\text{id}} + \frac{1-a}{2}P_{\text{id}}$$

$$E_{12} = aP'_{\text{id}}\frac{1}{2}P(>) + (1-a)P_{\text{id}}\frac{1}{2}\left(\frac{1}{2}\frac{1}{2} + \frac{1}{2}\right)$$

$$= \frac{a^2}{2+2a}P'_{\text{id}} + \frac{3-3a}{8}P_{\text{id}}$$

$$E_1 = 1 + \frac{a}{2-2a}P'_{\text{id}}$$

The transition probabilities can now be calculated as:

$$T_{10} = E_{10}/E_1 \qquad T_{11} = E_{11}/E_1 \qquad T_{12} = E_{12}/E_1$$

The transition probabilities out of the other indel state are the same because of symmetry:

$$T_{20} = T_{10} \qquad T_{22} = T_{11} \qquad T_{21} = T_{12}$$

Transitions out of the start state and to the end state

For the start state, the transition probabilities are like the match state. This is because (in the present model) new indels originate at the beginning of the sequence in the same way as they do after a match state:

$$T_{s0} = T_{00} \qquad T_{s1} = T_{01} \qquad T_{s2} = T_{02}$$

The transition probabilities to the end state are proportional to the transition probabilities to the match state:

$$T_{0e} = \delta T_{00} \qquad T_{1e} = \delta T_{10} \qquad T_{2e} = \delta T_{20}$$

The reason for this is that indels have to end before the end of the sequence. Finally, the transition probability from the start to the end state is the same as the transition probability from the match state to the end state:

$$T_{se} = \delta T_{00}$$

When calculating the probability of two sequences generated by the pair HMM, a single δ term will appear in the result, since there is always a single transition to the end state. This means that the relative probabilities of different ways of generating the sequences are independent of δ. Because of this, the value of δ can be set to one rather than zero, resulting in sums of transition probabilities out of the states greater than one. This does not change the results when aligning sequences using the pair HMM.

Pair HMM algorithms

The sum of the probabilities of all alignments for two given sequences is calculated using a method similar to the forward algorithm for standard HMMs.[12] The idea is to recursively calculate the probability of partial alignments, building them up from left to right. This gives a time and memory complexity on the order of L^2, where L is the average sequence length. An analogue to the backward algorithm is used to find the probabilities that specific residues match or form part of an indel. The most likely alignment is found by a method resembling the Viterbi algorithm. A good review of these algorithms is provided by Durbin *et al.*[8]

Results

Evolutionary simulations

Figure 3 shows a comparison of the results for the equations derived here and from evolutionary simulations. Each point represents the results for the simulations of 10,000 sequences of length 1000, ignoring the first and last 100 positions of the resulting alignments (to avoid edge effects). The simulations were done by applying indels using exponential waiting times as dictated by the indel model parameters r and a. The indel process was stopped when the time period t had elapsed. Notice that the particular residues are of no importance, since only the nature of indels is studied here. There is a high similarity between the curves and the simulated points, illustrating that the equations provide a good approximation to the indel process.

The transition probabilities from the match state to an indel state (T_{01} and T_{02}) increases with rt, but is almost independent of a. The length of the indels does not directly affect the transitions out of the match state. On the other hand, the transition from an indel state to the same state (T_{11} and T_{22})

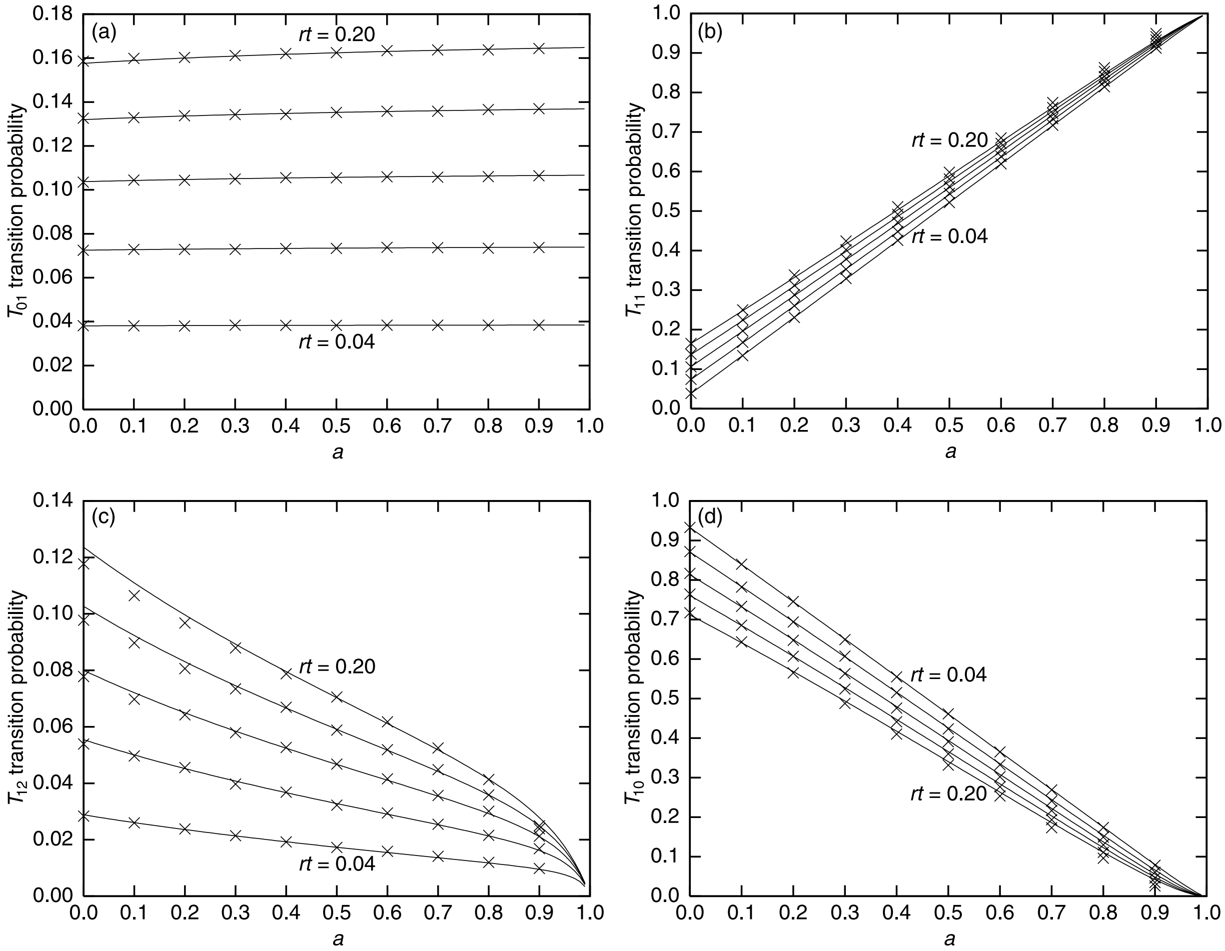


Figure 3. Various transition probabilities as functions of a. Data are shown for indel rates (multiplied by time) of rt = 0.04, 0.08, 0.12, 0.16, and 0.20. Crosses are for the simulated results, while the curves are calculated according to the equations presented herein.

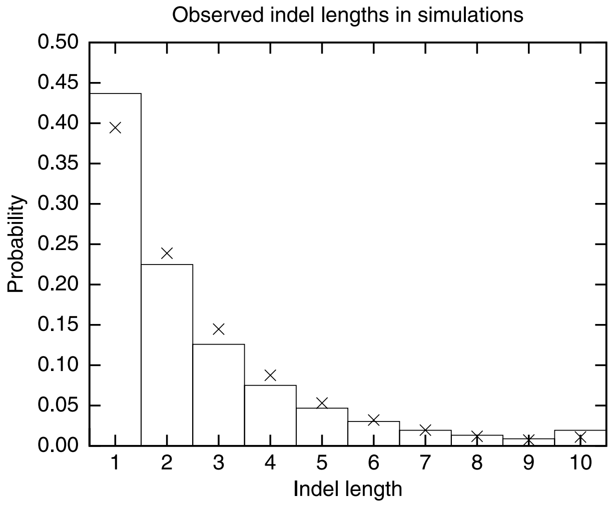


Figure 4. Observed indel lengths from simulations with $rt = 0.2$ and $a = 0.5$. The histogram summarizes the length distribution for 394,062 simulated indels. Crosses represent the geometric distribution with the same average as the simulation results. Short and long observed indels are over represented due to overlapping indel events.

is largely dependent on a, since this determines the indel length. The smaller dependence on rt stems from contiguous indels of the same type. The most complex relationship is for the transitions between one indel state and the other (T_{12} and T_{21}), which has a strong dependence on both a and rt.

Significant discrepancies between the results for the simulations *versus* equations do not occur for a and rt values small enough to be applicable to most alignable sequences. However, even if the curves coincided perfectly, the process described by the pair HMM and the indel process still could be slightly different. An example of why such a difference would occur is the accumulation of overlapping indels to give a length distribution of observed indels different from a geometric one. The pair HMMs developed here can only generate geometrically distributed observed indel lengths (see Figure 4).

Human hemoglobin protein sequences

To illustrate the potential utility of this new alignment method, human α and β-hemoglobin

(a)

```
                                                     *
α V-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GSAQVKGHGKKVADALT
β VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFS

α NAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
β DGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

Reliability ■ 0% - 40% ■ 40% - 60% ■ 60% - 80% ■ 80% - 90% ■ 90% - 100%

Parameters	
t – Time	0.876
r – Indel rate	0.0165
a – Indel length parameter	0.554

Negative log likelihood values[a]	
All alignments	726.68
Best alignment	728.63

Summary statistics	
Expected number of indels[b]	4.19
Average indel length (sites)	2.24
Expected number of replacements[b]	125.8

[a] Using $\delta = 1$.
[b] Based on the average sequence length.

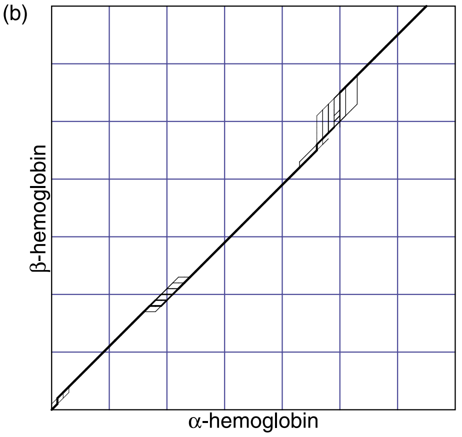


Figure 5. The pair HMM alignment and statistics for human α and β-hemoglobin protein sequences (HBA_HUMAN and HBB_HUMAN, respectively, in the SWISS-PROT database[24]). This alignment and associated statistics were determined with a new computer program available as a web server†. The top shows the optimal pair HMM alignment with the reliabilities of the matches and indels coded as colors. The asterisk highlights the single difference that distinguishes this alignment from the published TKF one by Hein *et al.*[14] (see the text). In the TKF alignment, this asterisked histidine of α-hemoglobin is shifted two positions downstream, thereby fragmenting the adjacent region of five contiguous gapped sites into two smaller ones of two and three dashed positions. The graph shows the pair HMM alignment of the first 70 amino acids from each sequence, starting from the lower left corner. The line width is proportional to the probability of an indel or match at that specific position. The graph highlights the uncertainties in the positioning of the gaps (horizontal and vertical lines).

† http://www.daimi.au.dk/~compbio/pairhmm

sequences were aligned using the pair HMM procedure with the JTT model of protein evolution[11] (Figure 5). The evolutionary replacement distance (time) between the α and β-hemoglobins is estimated as 0.876 expected replacements per site. Notice that this value is calculated from all possible alignments (weighted by their probabilities), rather than from the best one. The indel rate, *r*, is estimated to be 0.0165 expected insertions per expected replacement per position, so an insertion (or a deletion) occurs at a given point at a rate of about 60-fold less than the replacement rate. The main uncertainties in the alignment lie around the gaps where the specific configurations could vary.

CLUSTAL W[13] is a well-known program for global alignment, which uses the affine gap cost method by Gotoh[2] along with site specific gap penalties and other more advanced features. Although the same pairwise alignment is obtained by CLUSTAL W in this case, the results for the pair HMM procedure for human α and β-hemoglobins highlights several distinct advantages of the latter method: (1) the pair HMM alignment is less subjective, since all parameters are estimated by the method rather than being pre-specified; (2) evolutionary distances are incorporated in a probabilistic way rather than by a scoring scheme; and (3) the reliability of different regions of the alignment is readily estimated by this procedure.

Hein *et al.*[14] published an alignment of human α and β-hemoglobins using the model of Thorne, Kishino, & Felsenstein[4] (the TKF model). Under this model, every indel event has a length of only one residue, making all indel positions independent of each other. Their TKF alignment shows a single difference for the human α and β-hemoglobins compared to that for the method presented here (Figure 5). The asterisked histidine of α-hemoglobin in the pair HMM alignment is moved two positions downstream in their TKF result, thereby fragmenting the immediately adjacent region of five contiguous gapped positions into two smaller ones of two and three gapped sites. This TKF alignment matches the asterisked histidine with aspartic acid, rather than threonine in β-hemoglobin. Since indels are treated as independent events of single amino acids, pairing this histidine with either the aspartic acid or threonine results in the same indel contribution to the likelihood under the TKF model. Furthermore, the TKF alignment was done with the Dayhoff[15] replacement model, where histidine is more likely to align with asparatic acid than threonine. In contrast, in the pair HMM method, a single longer indel is preferred over the two smaller fragmented ones of the TKF alignment, as the former allows for indels of varying length.

The evolutionary time separating the two sequences is estimated as 0.916 according to the TKF model, compared to 0.876 given the pair HMM approach. This difference is most likely due to the use of different replacement models by the two approaches (the Dayhoff *versus* JTT matrices, respectively). The insertion and deletion rates for the TKF alignment are estimated as 0.03718 and 0.03744, respectively. These estimates are 2.251 and 2.267 times greater than the estimated indel rate for the pair HMM alignment. These ratios are consistent with the different ways that the pair HMM *versus* TKF methods view the four separate indels of nine total residues in the final optimal alignment for the former (Figure 5). The observed ratio of nine to four (=2.25) emphasizes once again that all gapped positions are viewed as separate independent indels in the TKF approach.

Discussion

This study has shown that there are many advantages to the statistical alignment of sequences using geometrically distributed indel lengths. A good approximation for such a model was made using a pair HMM. There are many applications for a method like this, not only in sequence alignment, but also in evolutionary studies. One advantage is that all alignments are taken into account when finding the evolutionary distance between two sequences. This may prove useful in the estimation of phylogenetic trees.[16]

There are a number of ways in which the method could be improved, including the incorporation of more states in the pair HMM, for example to represent conserved regions. Extra states could also be introduced to more closely approximate the observed indel length distribution, which is not precisely geometric.[7] However, the gain from this is likely to be limited, since the geometric approximation is quite good.

Another possible improvement is to put a prior distribution on the indel length parameter, *a*, and the indel rate, *r*. These distributions could be estimated from databases of trusted alignments. Having a prior on the indel parameters would improve prediction results by making the indel evolution conform to what is normally observed in the databases. On the other hand, it would make the method a little less objective.

The present method provides a framework for a number of other possible extensions. Multiple alignment methods based on this pair HMM approach would be useful. Such methods could be based on an optimal multiple alignment procedure, which is slow,[17] or progressive alignment, which is much faster.[18,19] Ideally, an intermediate could be found that provides fast multiple alignments that are close to optimal.

The pair HMM approach also provides a good connection to grammatical models, which have been applied to given alignments in comparative gene finding,[20,21] protein structure prediction[22] and RNA structure prediction.[23] The present alignment method could be combined with these methods to exploit the predictive capabilities of the grammars, while allowing for multiple alignments to be considered. This could be done through the use of the sub-optimal alignments generated from the pair HMM.

Acknowledgements

We thank the University of Florida and the Carlsberg Foundation (B.K.) for support. We also thank Sean Eddy for his helpful comments on the manuscript.

References

1. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
2. Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705–708.
3. Bishop, M. J. & Thompson, E. A. (1986). Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* **190**, 159–165.
4. Thorne, J. L., Kishino, H. & Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**, 114–124.
5. Holmes, I. & Bruno, W. J. (2001). Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**, 803–820.
6. Lunter, G. A., Miklós, I., Song, Y. S. & Hein, J. (2003). An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol.* In the press.
7. Benner, S. A., Cohen, M. A. & Gonnet, G. H. (1993). Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.* **229**, 1065–1082.
8. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.
9. Hasegawa, M., Kishino, H. & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.
10. Rodriguez, F., Oliver, J. L., Marin, A. & Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**, 485–501.
11. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282.
12. Rabiner, L. R. & Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Mag.* **3**, 4–16.
13. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
14. Hein, J., Wiuf, C., Knudsen, B., Møller, M. & Wibling, G. (2000). Statistical alignment: computational properties, homology testing and goodnesss-of-fit. *J. Mol. Biol.* **302**, 265–279.
15. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins, matrices for detecting distant relationships. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, pp. 345–352, Cambridge University Press, Washington, DC.
16. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996). Phylogenetic inference. In *Molecular Systematics* (Hillis, D. M., Moritz, C. & Mable, B., eds), pp. 407–514, Sinauer Associates, Sunderland, MA.
17. Sankoff, D. & Cedergren, R. J. (1983). Simultaneous comparison of three or more sequences related by a tree. In *Time Warps, String Edits, and Macromolecules: The Theory Arid Practice of Sequence Comparison* (Sankoff, D. & Kruskal, J. B., eds), pp. 253–264, Addison-Wesle, Reading, MA.
18. Feng, D. F. & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360.
19. Hein, J. (1989). A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. *Mol. Biol. Evol.* **6**, 649–668.
20. Rivas, E. & Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.
21. Pedersen, J. S. & Hein, J. (2003). Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, **19**, 219–227.
22. Thorne, J. L., Goldman, N. & Jones, D. T. (1996). Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13**, 666–673.
23. Knudsen, B. & Hein, J. (1999). RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, **15**, 446–454.
24. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E. *et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL. *Nucl. Acids Res.* **31**, 365–370.

Edited by J. Thornton

(Received 30 April 2003; received in revised form 1 August 2003; accepted 1 August 2003)