Chapter 6

Intron Clocks: Time-Dependent Models of Intron Evolution

6.1 Introduction

The most widely accepted method of finding divergence times between coding sequences is to exclude all but the silent sites - bases which, due to the redundancy of the genetic code, may be varied without changing the translated protein sequence [LG91]. These are presumed to be free of selective pressures. Under the further assumptions that the average rate of substitutions at a given site is constant (the "molecular clock hypothesis") and that neighbouring-base effects can be safely ignored (see [Bul86] for an evaluation of the error due to this approximation), substitution at silent sites can be modelled with independent continuous-time finite Markov chains and Bayes' theorem applied to yield maximum *a posterior* (MAP) estimates of the divergence time.

There are several ways the "codon clock" method (as it is referred to from now on) can go wrong. Firstly, neighbouring base effects are non-negligible. Unfortunately, modelling these effects in full takes resources that scale exponentially with the sequence length. Secondly, the molecular clock hypothesis has been shown experimentally to be flawed [GWD98]. Thirdly, the assumption that silent sites are unselected ignores the effects of codon bias [SAL$^+$95] as well as the possibility that there are other signals in coding sequences. Finally, the MAP divergence time estimate only represents the maximum of the posterior distribution, which might be very broad.

In an attempt to address the problems of limited data and silent site selection, an alternative method of obtaining molecular clock information from introns has been developed and evaluated in comparison to the silent site approach. The "intron clock" method is straightforward in conception: ancestrally conserved introns are identified from an alignment of coding sequences by looking for aligned pairs of residues that are both on exon boundaries (it is assumed that the probability of deletion and re-insertion of an intron at the same locus is negligible). Time-dependent models of substitution and small indel events, of the sort described in Section 2.4 of Chapter 2, are then used to generate likelihood distributions over the divergence time. These likelihoods can be compared or combined with the likelihoods from codon clocks in a principled Bayesian way.

Section 6.2 of this chapter investigates general patterns of intron evolution, in order to assess the extent to which the patterns of mutations in introns are compatible with the models of Section 2.4; i.e. whether it is legitimate to fit small-indel and single-base substitution models at all. It is found that while many pairwise intron alignments are indicative of infrequent small indels, there is a significant fraction of intron pairs whose lengths are very different. A number of these contain high copy-number repetitive sequences and it is proposed that repeat element insertion is the most plausible cause of large mutations. This conclusion is discussed in the light of recent suggestions that new ***Drosophila melanogaster*** introns originate by duplication [TRTA98]. Following on from this, Section 6.3 suggests a Bayesian methodology for dealing with the problems of large insertions and uses the GFFTools and BayesPerl packages described in sections A.4 and A.3 of Appendix A to estimate molecular evolutionary parameters for intron evolution and to assess the performance of intron clocks relative to codon clocks. Finally, in Section 6.4 the results are summarised and discussed.

6.2 General patterns of intron evolution

The data set of introns was derived from the set of gene duplications described in Section 5.2.6. Conserved intron loci were identified from Smith-Waterman alignments of these coding sequences using the ACeDB annotation. Of the 1035 genes in closely-related families found by the search procedure described above, 46% were found to have at least one ancestrally conserved intron (and on average, two to three), yielding a total of 1142 conserved intron pairs. Visual inspection of the protein alignments suggested that 52 of these pairs contained an intron with a mispredicted splice site, since changing the splice site would radically improve the protein alignment. These 52 introns, and a further 9 that looked as if they might be mispredicted, were removed from the data set, leaving

1081 pairs.

Before investigating this data set further, some ideas are reviewed about the signals that are known to exist in introns and the selection pressures that are expected to apply.

6.2.1 Conserved signals in introns

Splicing - excision of introns from messenger RNA - takes place during the passage of the mRNA to the ribosome. The first stage of splicing is the binding of the U1 and U2 small nuclear ribonucleoproteins (snRNPs) to the splice site consensus sequences which span the 5' and 3' exon-intron boundaries respectively, to form a committment complex. This is followed by the ATP-driven binding of the U4, U5 and U6 snRNPs and subsequently by catalysed intron excision. During the first stage of intron excision, the 5' splice site is cleaved and rejoined in a "lariat" structure to an adenine residue located at the branch point, which is separated from the 3' splice site by a short pyrimidine tract and is also bound by U2 during committment. In the final stage of intron excision the 5' and 3' splice sites are joined and the lariat intron excised [HK94, Bir]. Splicing signals known to be present in introns thus include the 5' consensus (bound by U1) and the 3' consensus, the branch point and the intervening polypyrimidine tract (all bound by U2). The canonical *C.elegans* 5' splice site consensus is thought to extend at least 3 bases upstream and 7 bases downstream of the 5' splice site; the 3' splice site is shorter, but often merges into the polypyrimidine tract [Bir, CLB93, ZB96]. There may well be additional signals subtle enough to have escaped detection. The picture is further clouded by the presence of an alternative splicing system involving U12 snRNPs with a stronger branch-point consensus and a weaker 3' signal, although no U12-type introns have been found in *C.elegans* [BPS98].

Most (62%) *C.elegans* introns are between 40 and 60bp in length, so that selection pressures due to the need for the above splicing signals may be expected to act on around 15% to 25% of bases, most of which will be close to the splice sites. Although this will retard the effective substitution rate near the splice sites, one can also expect to see a higher substitution rate in the less selected regions relative to coding sequences, since DNA damage is often not confined to a single base and there are correlations in the local probability of substitution [LKW97, SO95, KB95].

6.2.2 Sizes of indels in introns

Apart from substitutions, the effects of insertions and deletions must be considered. Standard dynamic programming algorithms typically assume an exponential prior distribution over gap lengths [DEKM98], though a study of processed pseudogenes suggests a power-law distribution to be more accurate [GL95b] and algorithms implementing alternative gap penalties have been described [MM88, ZLL97]. One can get an overview of the sizes of indel events by plotting the log-frequency distribution of percentage differences in length between paired introns Figure 6.1. (The number of indels is expected to be proportional to the length of the sequence, so percentage differences may be more informative than absolute differences.)

Figure 6.1 shows that while the frequency distribution is reasonably well-approximated by an exponential fit up to around a 15% difference in length, there is a long tail that is not well described by an exponential or, indeed, by a power-law distribution. 23% of the intron pairs in the data set lie in this long tail region.

In an attempt to explain the observed elevated frequency of large indels in introns, the `gffintersect.pl` program described in Appendix A was used together with the HMMER [Edd95] and GCG suites, the published *C.elegans* annotation and the CeRep database of *C.elegans* repeat families to look for repetitive elements present in one but not both members of an intron pair. Of the 1081 pairs of introns in the data set, 7% contain repetitive elements in at

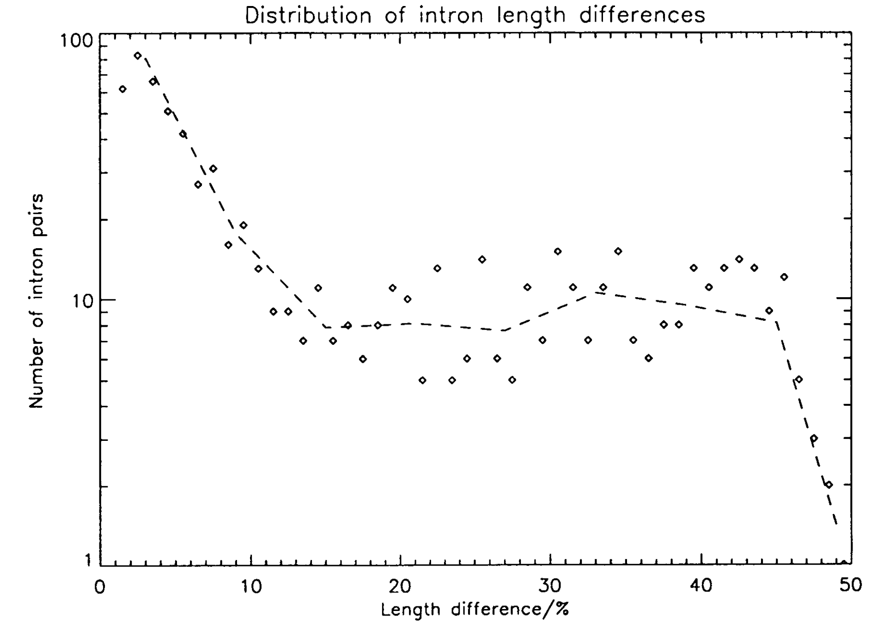

Figure 6.1: Distribution of fractional differences in conserved intron lengths. The dotted line is obtained by averaging nearby points. A large number of introns have length differences not easily explained by small indel models.

least one of the introns. The frequency of length mismatches between repeat-containing intron pairs is over 60% - considerably higher than the frequency for the whole data set. In all cases of length mismatch involving repeats, a repeat is found in one of the introns but not the other. Repeat insertion often appears to be associated with the formation of inverted and tandem repeats. These results are strongly suggestive that repetitive element insertion is a major cause of large indels in introns.

6.2.3 Intron mobility

It has been claimed that some *Drosophila melanogaster* introns at unaligned positions show significant homology, and that this is evidence for autonomous intron replication [TRTA98]. As part of the present study of *C.elegans* introns, a search for homologous introns was conducted using a BLAST search as a prefilter to the the Probabilistic Smith-Waterman (PSW) algorithm [BH96]. Gaussian distributions were fitted to PSW score frequency-distributions for random length-separated samples of the intron database to estimate "significant" (to 4 standard deviations) score thresholds. This is not a Bayesian approach, although a plausibility argument on Bayesian grounds is given in Section 2.7 of Chapter 2. The reason this approach was used was the woeful inadequacy of the "null" Bayesian generative model for introns: Figure 6.2 shows score-frequency curves for a pseudo-data set of introns generated from 4-mer frequencies (as might be sampled from a naive null model) and a real data set taken from the *C.elegans* intron database; there is a difference of 11 nats (= 16 bits) in the mean scores, though the variances are similar. To assess the significance of comparisons between introns of different lengths A and B (with $A < B$), a Gaussian score distribution with mean μ and variance σ^2 is used, where $\mu = \mu_A$ is the mean of a Gaussian fit to the score distribution of a random sample of sequences of length A, and $\sigma^2 = \sigma_B^2$ is the variance of a Gaussian fit to the score distribution of a random sample of sequences of length B. Roughly speaking, the

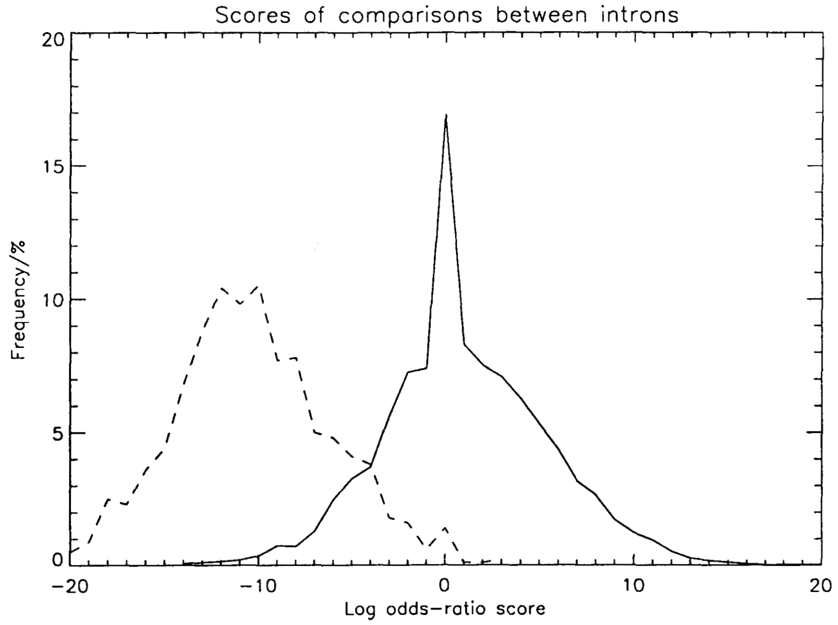

Figure 6.2: Log odds-ratio scores (relative to a null model of single-base composition) obtained by summing over alignments between randomly selected pairs of introns from *C.elegans* (solid line) and between pseudo-introns generated using 4-mer frequencies from real introns (dotted line). The sharp peak at zero is a fixed-precision rounding error.

average score can be expected to decrease with sequence length, but the variance to increase with length; so using μ_A and σ_B^2 gives the most conservative significance estimate.

The above search procedure yielded 110 intron pairs with "significant" homology. 61 of these were at aligned coding sites and 49 were unaligned. However, further analysis revealed that 34 of the 49 potential "mobile introns" in fact contained repeat sequences, accounting for their surprising homology by enclosed mobile elements, rather than intrinsic mobility. Three more of the homologies are between introns on the same protein, suggesting a tandem duplication. The remaining 12 significant intron homologies are listed in Table 6.1. Given the observed diversity of transposon species in the worm genome, and the low copy

Intron #1	Intron #2
AC3.6i2	C33D9.4i1
AC3.6i2	K03H1.5i7
B0024.6i16	C35A5.2i1
C05C10.2i6	E02H4.4i5
C05C10.2i6	F55E10.5i4
C09B9.1i2	K04C2.4i7
C03B8.4i2	W04D2.4i3
C05G5.6i4	T27B1.1i11
C05C10.2i6	PAR2.2i3
C05D9.6i3	F40E10.5i2
C11H1.4i11	F17A2.3i4
C18H2.2i5	K06B9.4i2

Table 6.1: Unexplained intron homologies in *C.elegans.*

number of some of these species, it seems entirely possible that many or all of the homologies in Table 6.1 correspond to uncharacterised mobile elements that happen to have landed in these introns.

A more sensitive search was carried out for examples of the proposed phenomenon of "intron drift" (slight dislocations in the positions of conserved introns relative to the coding sequences). No trace of this phenomenon was found.

6.3 Fitting time-dependent models to pairs of introns

Using the data set of conserved intron pairs, the hypothesis that introns are informative molecular clocks can be tested. Throughout this section the time-dependent coupled HMM with time-independent exponentially distributed gap lengths and Hasegawa substitution matrix described in Chapter 2 will be used.

To work with the time-dependent model, two pieces of software designed for this project were used. The first was a set of C++ classes designed to evaluate log-likelihoods of the form $\log \Pr[D|M,\Theta]$ where M is a (pairwise or single-sequence) hidden Markov model and Θ is a point in the parameter space of M.

To work with the likelihood data generated by the first program, a second piece of software that was designed to perform common manipulations on tables of log-likelihood values in multi-dimensional subspaces (including addition, multiplication, integration, marginalisation *et cetera*) was used; this software was written in Perl 5.0. Both these pieces of software are described in Appendix A.

6.3.1 Down-weighting uninformative pairs

Given the high number of introns disrupted by mobile elements or other kinds of mutation blitz, it would be useful to have a way of weighting intron pairs according to whether they look useful or not. A general, Bayesian way of doing this is as follows: Suppose that d is an element of data (in this case, a pair of introns) and that $D = \{d_i\}$ is an entire data set, and that it is desired to estimate a parameter Θ (or even a set of parameters, such as the divergence time of each pair). Suppose further that each data point d_i has an associated missing boolean variable $s_i \in \{0,1\}$ indicating whether it is of relevance or not. More specifically, say that the data point d_i was generated by one of two models, M_0 or M_1, where M_0 is a null model that is independent of Θ (i.e. $\Pr[d|\Theta, M_0] = \Pr[d|M_0]$), and that the choice of model is determined by s_i; so that if $s_i = 0$, then the data point was generated by the null model and is uninformative for the estimation of Θ. To make this work, the posterior probability of Θ is marginalised over S as follows:

$$\begin{aligned}
\Pr[\Theta|D] &= \sum_{\text{all } S} \frac{\Pr[D,S,\Theta]}{\Pr[D]} \\
&= \frac{\prod_i \sum_{s_i \in \{0,1\}} \Pr[d_i, s_i|\Theta] \Pr[\Theta]}{\Pr[D]} \\
&= \prod_i \sum_{s_i \in \{0,1\}} \Pr[d_i|s_i,\Theta] \Pr[s_i|\Theta] \frac{\Pr[\Theta]}{\Pr[D]}
\end{aligned}$$

This approach weights contributions to the MAP estimate of Θ according to the posterior probability that the paired sequences are alignable (i.e. that they have not been disrupted by a transposon insertion). The prior probabilities $\Pr[s_i|\Theta]$ determine the weighting bias towards "alignable" or "unalignable". In theory, a time-dependent prior could be used, but there are too many different types of transposon-induced disruption to estimate a meaningful transposon insertion rate from the present data. For the present work, the dependence of the s-prior on Θ was dropped and a score cutoff of 10 bits (6.9 nats) was used, corresponding approximately to a 1000 : 1 weighting against informative pairs ($\Pr[s_1] \simeq 0.001$).

6.3.2 Testing intron clocks

To test the intron clock hypothesis, the likelihoods of four different models were evaluated:

- Model $\mathcal{M}_0$: All introns mutate at different rates.
- Model $\mathcal{M}_1$: All introns mutate at the same rate, but this rate is not correlated to the rate of synonymous substitutions in coding sequences.
- Model $\mathcal{M}_2$: All introns mutate at the same rate, which is exactly identical to the rate of synonymous substitutions in coding sequences.
- Model $\mathcal{M}_3$: All introns mutate at the same rate, which is advanced or retarded by a constant factor, relative to the rate of synonymous codon substitutions.

Implicit in each model are the assumptions that (i) the molecular clock hypothesis is valid for synonymous substitutions in coding sequences and (ii) the previously described time-dependent gap HMM is a valid model for neutral intron evolution.

Since each of these models has a different number of parameters, it is necessary to integrate the likelihood across the entire parameter space of each (see e.g. [Mac92a] for a readable explanation of why integrating across the whole parameter space penalises models with more parameters). An approximation to this

Model	Brief description of model	$\log_2[\Pr[\text{data}\vert\text{model}]]$ $-\log_2[\Pr[\text{data}\vert\mathcal{M}_0]]$
$\mathcal{M}_1$	Unsynchronised w/codons	655
$\mathcal{M}_2$	Perfectly synchronised w/codons	91
$\mathcal{M}_3$	Imperfectly synchronised w/codons	485

Table 6.2: Log-odds-ratios of synchronisation hypotheses for intron and codon clocks, relative to the null hypothesis that introns do not show clock-like behaviour at all ($\mathcal{M}_0$).

integral was found using the trapezium rule with a finite range for divergence times of $0 \leq t \leq 10$ with a time-step $\Delta t = 0.05$. Model $\mathcal{M}_3$ has an extra parameter $r = t_i/t_c$ determining the relative rates of the intron and codon clocks; this was integrated over $0 \leq r \leq 2$ with $\Delta r = 0.1$. Uninformative (flat) priors were used for r and t. Strictly speaking, parameters such as the gap-open rate g, the mean gap length l and the transversion/transition ratio k should be integrated over as well, but these were also approximated by $g = 0.039$, $l = 1.2$ and $k = 0.53$, values which were obtained by a crude approximate Viterbi-likelihood method. Uninformative intron pairs were down-weighted, as described in the previous section. The likelihood calculations and the numerical integration were performed with the aid of the LogSpace and BayesPerl packages described in sections A.2 and A.3 of Appendix A.

The log odds ratios (in bits) of models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ to the null model $\mathcal{M}_0$ are shown in Table 6.2. The clear winner is model $\mathcal{M}_1$: intron clocks are synchronised between introns, but do not bear any relation to the synonymous codon substitution clock. A clue as to why this might be is offered by the superior performance of model $\mathcal{M}_3$ over model $\mathcal{M}_2$; recall that model $\mathcal{M}_3$ allowed intron and codon clocks to be out of sync by a constant ratio, whereas $\mathcal{M}_2$ required that they stay in exact step. In fact the maximum-likelihood value of the relative clock-rate parameter r was $\hat{r} = 0.1$, suggesting that to make intron clocks work under the present model, they would have to run significantly slower than the codon clocks. When the analysis was repeated without alignment weighting (which will tend to introduce a negative bias to the intron clock rate) the ML value for r rose to $\hat{r} = 0.5$ but the ranking of the four models remained unchanged.

It is possible that a time-dependent prior for whether sequences were alignable would improve the performance of the clock-like models. The limited range over which the time parameter was integrated may also be a source of error.

These results suggest that while there is hope of fitting time-dependent models to non-coding DNA (and, in particular, to introns), the current models are far from perfect and are not yet suitable sources of clock information. The use of codon clocks is itself known to be a flawed technique (see, for example, [GWD98]). With the increasing availability of non-coding DNA sequence, it might be a good idea for studies of non-coding DNA evolution to consider unpredictable, traumatic mutations as well as the tractable single-base substitutions and small indel events that are more typical of coding DNA.

6.4 Discussion

Introns within the same gene evolve at the same rate in *C.elegans*, but this rate does not correspond well to the rate of syonymous substitution in the coding sequence. If the correspondence is made, however, it is better to allow the introns to evolve at a slower rate than the synonymous sites. This suggests that the selection pressure on introns is greater than on synonymous codons. One reason for this could be that *C.elegans* introns are rather short and around 20% of the average intron sequence length is taken up by splicing signals. Another reason could be small genes (for e.g. snRNAs) in *C.elegans* introns; these will be subject to selection.

The distribution of length differences between introns in homologous positions suggest that while 77% of intron pairs have diverged according to the kind of time-dependent stochastic model of small-indel accumulation proposed by Thorne *et al* [TKF92], the remaining fraction of pairs have been subject to large insertions or deletions. Introns containing repetitive elements are strongly associated with this effect, suggesting that repetitive element insertion is a primary cause of large mutations in intron sequences. It is proposed that repetitive elements also account for some of the surprising homologies that are found to exist between intron sequences.

The inadequacy of the default null model for introns has implications for the design of genefinding algorithms. A more sophisticated model should not only take account of the known splicing signals within introns, but also reflect the empirically observed propensity of unselected sequence for low-complexity regions such as poly-AT tracts. Modelling these kinds of features with HMMs demands large state spaces since the lengths of the features are not geometrically distributed. To avoid the training problem, the effective number of parameters can be reduced. An outline of the derivation of constraints on HMM parameters for modelling complex length distributions is given in Section 2.7 of Chapter 2.

6.4.1 Availability

The gene duplication data described here are available with the rest of Wormdup at the following URL:

`http://www.sanger.ac.uk/Users/ihh/Wormdup/`