Chapter 6

SUMMARY AND DISCUSSION

6.1 SUMMARY

Software was produced which took data about TFBSs known to exist, and outputted examples of TFBSs that appear to have not been conserved during evolution. To do this, the software retrieved genomic DNA sequence around the TFBS, and orthologous sequence for other vertebrate species, which was then aligned. A PWM was used to assess if, in other species, the TF would bind the DNA orthologous to the known TFBS.

A TFBS would be classed as "diverged" in a species only if there was a good case for this, and nothing was found to suggest otherwise. For instance, failure to obtain an alignment was not regarded as proof that the TFBS was not conserved; instead it produced the conclusion "uncertain". "Uncertain" was a frequent conclusion.

To check what would happen if incorrect data was used, a collection of fictional TFBSs was created and analysed.

Starting with TransFac data, cases of "conserved" TFBSs were produced and compared with "diverged" TFBSs. The two were similar in some respects (table 2.15). There were a few measures for which a difference was found, yet the same difference was found for fictional TFBSs, raising doubts about

the reliability of the results (table 2.16), so examination of these was not pursued. Most interesting were the measures which not only differed between "conserved" and "diverged" TFBSs, but also differed between real and fictional data (table 2.17).

Of these, the most interesting was felt to involve "neighbourhood" DNA (defined as 50 bases on each side of the TFBS). If a TFBS was "conserved", the sequence of the neighbourhood DNA was also more highly conserved than was the sequence of the neighbourhood around "diverged" TFBSs. This was chosen for further investigation (table 2.20).

Also, the explanation for this observation was sought. More than one hypothesis was produced. The "correlated evolution" hypothesis supposed that several adjacent TFBSs would be lost at the same time. The "site-density" hypothesis supposed that the probability that a TFBS would be gained/lost varied with the number of TFBSs nearby, being lower if the neighbourhood contained a particularly large number of TFBSs. Another hypothesis supposed the explanation to be based on variation in the mutation rate, affecting TFBS that were evolving neutrally.

An investigation of which hypothesis was correct might require more data than was available from TransFac. Therefore, and to confirm the earlier work, additional data produced by the chip-chip method was used, based on a publicly available list of targets for the transcription factor HNF4. This did not confirm the earlier work; although the experimental data gave a similar result, the fictional-TFBS analysis gave a result similar to that from the experimental data (and larger than the fictional-TFBS result in the earlier work). Therefore, the possibility remained that the "experimental TFBS" result was the consequence of faulty data.

Detailed investigations failed to produce a complete explanation for why the chip-chip results differed from the earlier work, but one clue was produced by an exercise with the TransFac data. TransFac data generally includes the sequence of the TFBS, from which its exact location can be found; but chip-chip data does not, so the TFBS was found by using a PWM to search 500bp upstream of a gene, with some risk of selecting the wrong location.

The exercise treated TransFac-based fictional data like chip-chip data (with the TFBS being located by a 500bp search); this produced a result that differed from that obtained when the same data had been analysed by the earlier method (that did not involve a 500bp search); the difference made it much more similar to that obtained in the analysis of fictional chip-chip data. Therefore, the reason the chip-chip analysis failed to confirm the original result may well be related to the 500bp search.

A discussion was given that considered what changes in sequence could be expected on the basis of the different models. One interesting suggestion was a prediction that "lost TFBSs" should differ from "gained TFBSs" in a particular way. This was that, according to the "correlated-evolution" hypothesis, the "neighbourhood" DNA flanking a TFBS that has been lost should be more highly conserved than the "neighbourhood" DNA flanking a TFBS that has been gained. In contrast, the "site-density" hypothesis does not predict that any difference should be observed between "gains" and "losses", in respect of the conservation of neighbourhood DNA. This therefore provided a method of deciding which hypothesis to reject, providing data could be produced that distinguished "gains" from "losses".

To distinguish "gains" from "losses", an analysis was based on the phylogenetic tree of vertebrates. Provided the existence or non-existence of a TFBS was known in at least three species, in some cases it was evident that the most parsimonious explanation was that the TFBS was an ancestral TFBS that had been lost in one species; in other cases, it similarly appeared that the TFBS had been recently gained. In many cases, no clear conclusion was possible, so the sample of TFBSs was small (table 5.1). It did, however, provide some evidence against the idea that "gains" are many times more frequent than "losses". A fictional-TFBS analysis produced far more "gains" than "losses", suggesting that incorrect data will tend to bias results in the direction of suggesting "gains" are far more frequent than "losses".

The difference between "gains" and "losses" predicted by the "correlated-evolution" model was looked for (table 5.4). There was some evidence that the predicted difference does exist; however, the evidence was on the bor-

derline of the 5% statistical significance level, and therefore is not of strong reliability. It is tentative evidence for preferring the "correlated-evolution" model to the "site-density" model.

6.2 COMPARISON WITH WORK IN THE LITERATURE

6.2.1 The turnover model

The "turnover model", as it has been called (Doniger and Fay, 2007), supposes that mutations create a surplus TFBS and later destroy another TFBS in the same regulatory region. The TFBS that has been gained acts as a replacement for the TFBS that has been lost, the implication being that there is no significant change in the expression profile, or that there is no difference in the phenotype. Much of the interest in "turnover" originated with a study that demonstrated a good example in *Drosophila* (Ludwig et al., 2000) and many examples could be give of subsequent papers that mention it (Gasch et al., 2004).

In particular, a number of the most interesting papers in the field devote considerable space to examining the "turnover model". One *Drosophila* study unsuccessfully looked for evidence of frequent turnover (Moses et al., 2006). However, another study found evidence of widespread turnover in mammals (Odom et al., 2007), whilst another found evidence of widespread turnover in yeast (Doniger and Fay, 2007). (Fuller details were given on page 51). This shows that the "turnover model" is one of the most commonly used models of TFBS evolution. It is therefore worth examining how the turnover model links with (i) the site-density model, (ii) the correlated-evolution model, and (iii) certain evidence, that have been presented in this thesis.

(i) The site-density model presented in this thesis does not have a clear link with the turnover model. This is because the site-density model predicts that the rate of gain/loss will depend on site-density, whereas the turnover model

predicts that a gain will be accompanied by a loss. These two predictions do not support each other, nor do they contradict each other. Both models could be correct, or just one of them could be correct. Particular cases of TFBS evolution might conform to both models. Hence there is no clear link.

(ii) Turning to the correlated-evolution model, it is noticeable that the turnover model is usually described as if the gain of an *individual TFBS* is followed by the loss of an *individual TFBS*. Admittedly, this is often not stated explicitly, but most of the evidence in the papers just cited is evidence of individual TFBSs being gained or lost. This is different to the simultaneous loss of several TFBSs that is an essential part of the correlated-evolution model. However, it is still possible that both models are correct (it may be that some examples of TFBS loss conform to the turnover model, whereas other examples of TFBS loss conform to the correlated-evolution model). But it does imply that a particular example of TFBS loss cannot conform to the correlated-evolution model *and* also conform to the model of turnover of individual TFBSs.

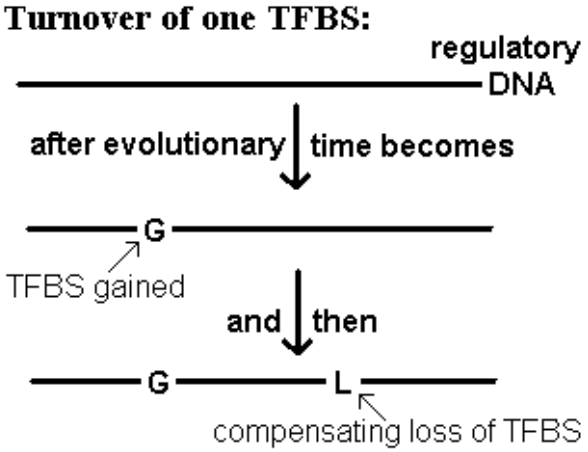
One might attempt to combine the correlated-evolution model with the turnover model, by proposing a modified turnover model in which the rapid gain of many TFBSs is followed by the simultaneous loss of many other TFBSs from the same regulatory region, so that the net effect of these gains and losses causes no change in the expression profile (of the gene controlled by that regulatory region). But that would be a substantial modification to the turnover model currently in use.

(iii) Table 5.4 presented evidence that the DNA flanking "losses" is more highly conserved than the DNA flanking "gains", as shown by the 4.3% difference at the bottom of that table. If the turnover model were used to predict the outcome of that analysis, what would it predict? Figure 6.1 addresses that question. This shows that the model of turnover of individual TFBSs would predict no difference; hence, the 4.3% difference actually observed is evidence of TFBS evolution being affected by some phenomenon other than turnover of individual TFBSs.

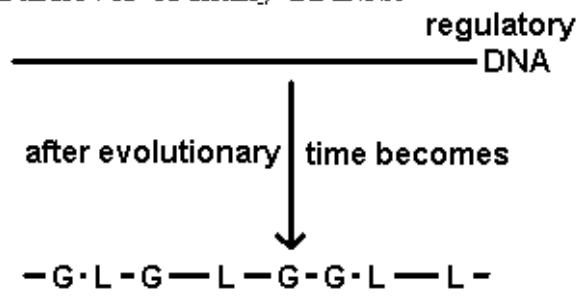
However, it should be remembered that the 4.3% difference was not a partic-

ularly strong result (it was on the borderline of the 5% statistical significance level), so it can only be claimed as tentative evidence for a phenomenon that cannot be explained by the existing turnover model. But it does illustrate how the methods of this thesis have the potential to produce evidence that cannot be explained by the model of turnover of individual TFBSs. Such evidence would show up the limitations of the existing turnover model, and

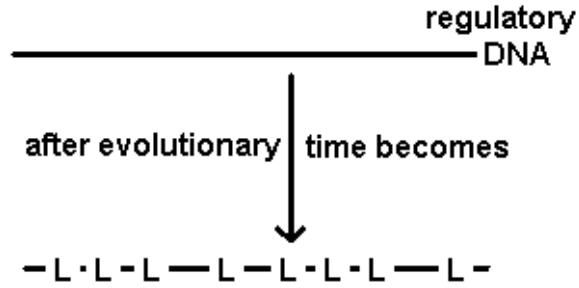
Figure 6.1: This, and the following three drawings compare the "turnover" model with other models. Here, "L" marks a TFBS that has been lost (compared with the ancestral DNA sequence shown near the top of each diagram), and "G" represents a gain of TFBS. The first drawing shows a "turnover" model in which a gain-of-TFBS makes an existing TFBS redundant, which is lost afterwards. The next drawing, "Turnover of many TFBSs", shows several such turnover events happening in a regulatory region, thus resulting in equal numbers of gains and losses. Therefore, if we pick any particular lost-TFBS, its flanking DNA will contain *equal numbers* of gains and losses (approximately). Similarly, if we pick any particular gained-TFBS, its flanking DNA will also contain approximately equal numbers of gains and losses. Hence, in respect of flanking DNA, this "turnover" model does not predict any difference between lost-TFBSs and gained-TFBSs. Whilst turnover of individual TFBSs cannot explain a difference of this kind, it might be explained by other models of TFBS evolution, such as shown in the drawings illustrating simultaneous-loss, or multiple-gain.



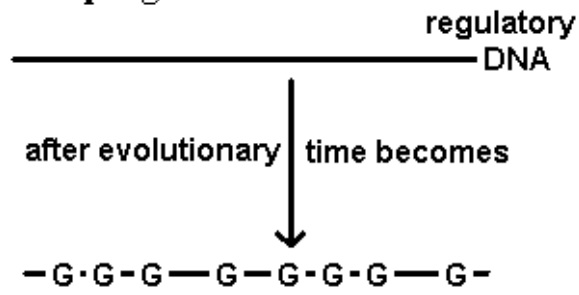
Turnover of many TFBSs:



Simultaneous loss of TFBSs:



Multiple gains of TFBSs:



so be a useful step towards generating a model of TFBS evolution that was more sophisticated and more realistic.

6.2.2 Other studies

There do not seem to have been any papers which have produced conclusions exactly the same (or exactly opposite) to those of this thesis, regarding the conservation of DNA around TFBSs that have been gained or lost. However, some published work overlaps closely with that described here, and merits comparison.

Other studies (O'Lone et al., 2004),(Sauer et al., 2006) also took TFBSs that were known to exist in one species of mammal and used *in-silico* comparisons to determine if they were conserved, or not, in other mammals, which is the same strategy as used in this thesis.

However, both those studies put considerable emphasis on estimating the percentage of TFBSs that were not conserved in a human-rodent comparison, one (O'Lone et al., 2004) estimating 60-80% for ER sites, but the other (Sauer et al., 2006) estimated 72% being *conserved* as the average over a variety of TFs. So at first there appears to be a large discrepancy. But the latter paper also claimed that the rate depended on the TF being studied, the "pattern matching" conservation rate (C_{pat}) for ER being only 52.9%, which is lower than for all but 2 of the 22 TFs considered (however, another estimate of the ER conservation rate - C_{seq} - was rather different at 82.3%). So possibly the discrepancy was because one study (O'Lone et al., 2004) only considered one type of TF, whose TFBSs evolve rapidly.

This thesis, in contrast, has not put so much emphasis on estimating the percentage of TFBSs that are non-conserved. This is because it was thought that there would be some cases which could not be classified reliably as "conserved" or "non-conserved", so those were placed in an "uncertain" category. This resulted in a large number of cases for which no definite conclusion was drawn (table 2.14), far exceeding the number of examples of non-conservation. Consequently, an estimate of the percentage of TFBSs that were not conserved would be subject to very large error limits, so that it would only be an order-of-magnitude estimate. So this aspect was not emphasised.

Unlike my study, a comparison of nine different alignment algorithms was

inclued in one of these studies (Sauer et al., 2006). There were no great differences between their performances.

The conservation of "neighbourhood" sequence around a TFBS has formed an important part of my thesis, and very similar observations have been made (O'Lone et al., 2004). They plot how "conservation score" changes as one goes along a DNA sequence and note that conserved TFBSs often occur within a peak - "visual examination suggests that the estrogen-responsive sites are located as anticipated within peaks of similarity between mouse and human sequences ... Sites that are not functionally conserved (by matrix comparisons) are predominantly located in regions of low conservation or in unalignable regions". (They note exceptions to both these rules.) This observation is qualitatively the same as the "%id of neighbourhood" observations I have made in tables 2.17, 2.18 and 2.20. They do not elaborate the "visual observation" by giving actual statistics, or suggest any explanations. They do, however, give some statistics relating to "the average sequence conservation of 100 bp of flanking sequences (50 bp to either side)", but these address a slightly different question - is the TFBS sequence conserved more strongly than the flanking sequence? For conserved TFBSs the answer is yes - conservation is 6.9-fold higher in the TFBS than in the flanking sequence - but for non-conserved TFBSs it is much less so, only 2.3-fold. This does not address whether the flanking sequence around the conserved TFBSs is more strongly conserved than the flanking sequence around the non-conserved TFBSs.

Evidence has been presented that (in yeast) gain-of-TFBSs is more likely to occur in promoters with large numbers of TFBSs (Bilu and Barkai, 2005). That is, a site-density effect occurs, but in the *opposite direction* to the site-density effect considered in this thesis. Their observed site-densities are relatively low; thus, in their study, a high-density promoter would contain 15 TFBS with an average length of 7.5 bases in a promoter 1000 bases long, which works out as only 11% of the bases being part of a TFBS. Of course, undiscovered TFBSs may mean that the true site-density is much higher than that given by the current yeast data.

6.3 DISCUSSION

6.3.1 Possible mechanisms

Earlier chapters have described the correlated-evolution and site-density hypotheses, but without suggesting the possible mechanisms behind them. This will now be done.

For the correlated-evolution hypothesis, one can imagine a regulatory region performing a particular task, such as activating a gene within a particular tissue type, so that if, say, an environmental change made this task no longer useful, then all the TFBSs within that regulatory region would become useless at the same time. This would be correlated-evolution loss of TFBSs.

However, it is more difficult to imagine correlated-evolution *gain* of TFBSs, as this would require a number of simultaneous mutations that created a number of TFBSs within the 100-base region, which seems extremely unlikely. Predictions have been calculated (Stone and Wray, 2001) of the length of time required to produce TFBSs by chance by point mutations, under certain assumptions. For example, they predict a TFBS will take 752 years to appear in a 200-bp region in a population of one million mice; a pair of TFBSs would take 180,810 years; but "the likelihood of a dozen binding sites evolving simultaneously without selection is infinitesimally small", but they might be produced by stepwise assembly instead. These estimates refer only to the times required for the TFBS to appear in one individual; it would take additional time to spread throughout the whole population. Thus the stepwise addition of 12 TFBSs would require 12*752 years, plus the unestimated time required for each TFBS to spread through the population before the next addition was likely to occur. One suspects the unestimated time would greatly increase the time required. Still, this line of thought does not rule out the possibility of the additional of a number of TFBSs within 100-200 bases of DNA within, say, a million years. Although not strictly simultaneous, that would be such a short time compared with the human-mouse divergence ($\sim 10^8$ years ago) that, using data used for this project, it would

be indistinguishable from simultaneous addition of many TFBSs.

Thus we should consider that it may, after all, be possible to have correlated-evolution which consisted of the almost simultaneous gains of a number of TFBSs in one regulatory region. Nonetheless, correlated-evolution by gain-of-TFBSs still seems less plausible, compared to correlated-evolution by loss-of-TFBSs.

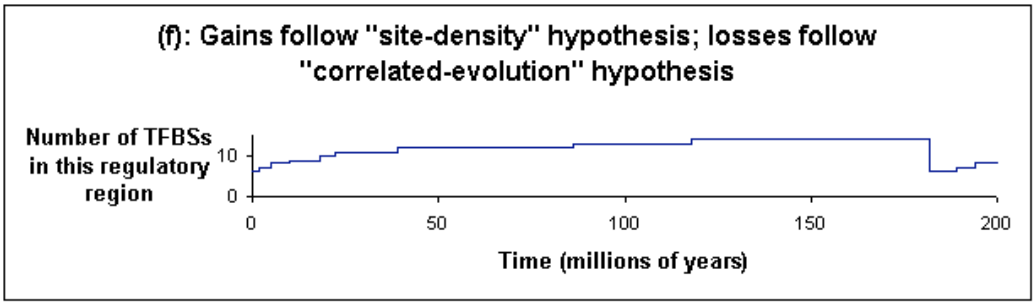
Turning to the site-density hypothesis, one possible explanation would be that regulatory regions with low site-density may have a lot of "spacer" DNA between each TFBS which could easily mutate to create a new TFBS. Conversely, regulatory regions with high site-density could have great difficulty gaining a new TFBS simply because virtually all the DNA is part of a TFBS, and hence cannot form part of a new TFBS (except for overlapping TFBSs). With this explanation, one might suppose that the probability of gaining a new TFBS was proportional to the amount of DNA that did not make up existing TFBSs.

That explanation applied to gain-of-TFBS, but it is not obvious that the same explanation would apply to loss-of-TFBS. It could apply if a gain-of-TFBS made an existing TFBS redundant, causing it to be lost; so if most losses of TFBS were caused in this manner, then the site-density hypothesis could apply to losses. But it seems rather speculative to suppose that most losses of TFBS occur in that manner, so the site-density hypothesis looks more plausible when applied to gains rather than losses.

So far the hypotheses have been discussed as if either one or the other is correct. However, it is also possible that both hypotheses are correct and form part of the explanation. Notice that the correlated-evolution hypothesis seemed more plausible when applied to losses of TFBS than it did when it was applied to gains of TFBS; whereas, for the site-density hypothesis, it was the other way round. Thus, one idea that should be considered is that the correlated-evolution hypothesis is correct but only applies to losses, whilst the site-density hypothesis is correct but only applies to gains. That idea suggests an interesting picture of the evolution of a regulatory region. If we imagine a regulatory region with low site-density, TFBSs will be gained

step by step over a long period of evolutionary time, the rate of gain slowing down as the amount of spacer DNA reduces, followed by the sudden loss of a considerable number of TFBSs, followed (if the regulatory region has not been destroyed completely) by another prolonged step-by-step gain of TFBSs. Figure 6.2 illustrates this idea, using the same format as figure 1.6 in the Introduction. Figure 6.2 was inspired by knowledge of the data collected during this project, but obviously goes beyond what can be proven by that data, so like figure 1.6 it should be considered as speculative.

Figure 6.2: Another way a regulatory region might evolve
Another speculative scenario for the evolution of a regulatory region, which uses both of the hypotheses considered in this thesis



It is also possible that the main explanation is not any of the hypotheses that have been considered so far. Here is another hypothesis, inspired by an idea in the literature (Bilu and Barkai, 2005). That paper suggests that some genes need to maintain a precise expression pattern, but others do not; genes in the latter category will be much more tolerant of TFBSs being gained or lost in their regulatory regions. Presumably, a consequence of this would be that these regulatory regions will not have a highly conserved DNA sequence, even if they are filled with a high density of TFBSs, because so many of the TFBSs will be gained or lost. Thus, in principle, this could result in non-conserved TFBSs tending to be surrounded by weakly conserved DNA, as has been observed in an earlier chapter of this thesis.

Although that is another hypothesis, it still fits into the broad idea considered in this thesis, that the probability that a TFBS will be gained/lost is corre-

lated with what occurs in the regulatory region that immediately surrounds it.

6.3.2 Questions raised and implications

Some possible modifications to the analysis used here require better scientific understanding before they can be implemented.

One is the question of what happens when a useful TFBS becomes, as a result of evolution, no longer useful. Does it (i) in most cases become useless (and so unaffected by natural selection), or does it (ii) in most cases become harmful (and so mutations that destroy it are positively selected for)? During the literature review it was noted (page 44) that, as a broad generalisation, natural selection does not remove unnecessary TFBSs. Blindly applying this rule indicates that (i) is the correct answer, so (i) was adopted as the correct assumption for this thesis - and this assumption has also been adopted by other groups modelling TFBS evolution (see the more detailed discussion on page 163). But that assumption may not be correct. If we ask why natural selection does not remove unnecessary TFBSs, one possible reason is that many TFs cannot activate transcription on their own, but only by synergistic cooperation with other TFs; hence, if a spurious TFBS is not near other TFBSs of the right kind, it will unable to activate and so harmless. It seems likely that many spurious TFBSs will be in this situation, not near other TFBSs of the right kind, but there will be exceptions (which may therefore be subject to counterselection). But, in the case of a once useful TFBS that is being lost, we are dealing with a situation which is just that described in the exception. So it seems to be an open question whether (i) or (ii) correctly describes the fate of lost TFBSs - which is a possible question for future research. Indeed, some papers admit that counterselection of losses is a possibility (Berg et al., 2004), (Doniger and Fay, 2007).

There is a somewhat similar question about gains of TFBSs: When a new, useful TFBS is created as a result of a chance mutation, which is the most common route: (i) The mutation initially creates a useless TFBS (which can

bind but does not affect the fitness of the animal), and the TFBS becomes useful during subsequent evolution; or (ii) The mutation creates a TFBS which is useful immediately.

Turning to the correlated-evolution hypothesis, the literature review earlier referred to two enhancer models (Arnosti and Kulkarni, 2005), the "enhanceosome" and the "billboard", and I noted that the "enhanceosome" model seems to imply that the loss of one TFBS would make the entire enhancer useless. That, of course, matches what happens when a TFBS is lost under the correlated-evolution hypothesis. In other words, if most enhancers conform to the "enhanceosome" model, then "correlated-evolution" is predicted.

However, the reverse does not follow; that is, it does not follow that correlated-evolution implies that the "enhanceosome" model must be correct. The rival "billboard" model does permit the enhancer to lose a TFBSs and still retain partial function, but it does not follow that TFBS loss always occurs in that fashion. If, say, an environmental change suddenly causes an enhancer to become entirely useless, then simultaneous loss of all TFBSs will occur irrespective of whether the "enhanceosome" or "billboard" model applies.

This thesis reports work on vertebrates, which raise the question of whether similar results would be obtained for other organisms. On page 198 it was noted that a site-density effect in yeast (Bilu and Barkai, 2005) appears to go in the opposite direction to the site-density effect reported in this thesis. That may suggest a fundamental difference between yeast and vertebrates, but it may also be that the site-density hypothesis presented in this thesis is simply not true at all - which emphasises the importance of getting an increased sample of data to decide which of the alternative hypotheses is most plausible (possibly by making better use of existing data).

That, indeed, would be the most important extension of the current project. This thesis indicates that existing data can provide evidence for models relating the gain/loss of TFBSs to the surrounding regulatory region. It also indicated how data could in principle be used to decide between different models; this approach produced tentative evidence for preferring the "correlated-evolution" model to the "site-density" model (page 186), but given

the significance level of that result, it would clearly be desirable to obtain a larger sample, thus reducing the sampling errors and possibly providing a more reliable answer as to which hypothesis was the best.

Declaration: no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning, except for the reuse of some Perl modules (altered as required), originally written during the candidate's Master in BioInformatics course (University of Exeter)