Chapter 1

Introduction

1.1 Preamble

The role of bioinformatics is expanding from one of post-experimental data analysis to include data management and organisation. As a consequence of this, concepts of what is possible in bioinformatics are likely to influence the future design and planning of the grand experiments that will succeed the present genome projects. Researchers who want to access the results of these grand experiments will do so through layers of software that filter, organise and interpret the volumes of available biological data. If the trends apparent in the rest of the software world apply to computational biology then users will be merciless in their demand for tools that are intuitive, powerful, simple and interoperable. To meet this demand, computational biologists will need to be aware of the solid mathematical frameworks that can support apparently simple algorithms; they will need to complement the drive to develop sophisticated techniques with a sensitivity to the changing needs of the scientific community; and they will need to respond quickly and authoritatively to the inevitable developments in technology that are on the way. In short, bioinformatics - and related informatics such as chemoinformatics and clinical informatics - are *at least* as important as they have been hyped to be!

It may be suggested that this view of the future of bioinformatics is influenced by the role of sequence analysis with respect to the genome projects (see e.g. [CSC98, FAW+95, WS98]). These projects are generating megabases of DNA sequence, coding for hundreds of thousands of genes whose structure and function - thanks to the infeasibility of total simulation of the quantum mechanics of peptide molecules [Sha97] - can often only be elucidated by detection of sequence-level homologies to previously characterised proteins. Nevertheless it seems unlikely that the trend of high-throughput data collection will be reversed in the near future, with new technologies like microarrays generating new types of data and new informatics challenges [DIB97, BJVU98]. If this trend does continue, then there are two aspects of bioinformatics research that can be

predicted to survive: the development of self-consistent families of algorithms for finding patterns in data and the presentation of these algorithms to a wider audience of researchers in the form of succinct and intuitive database interfaces. These two activities are closely linked: while it is often true that the program with the better user interface wins, the use of a good interface on a bad algorithm, or a bad interface on a good algorithm, will always result in tension. It is much easier to put a positive spin on a technique that has a good, simple idea at its core than it is to make a complex collection of heuristics seem intuitive. Furthermore, good algorithms stimulate excitement and interest, which is exactly what is needed to motivate the fast-changing world of interface design. It should also be stressed that the meaning of "user interface" here refers to *all* the methods that are used to access large data sets, so that the distinctions between data and methods, interface and algorithm, data-structure and object-model become increasingly blurred. An example of the new class of database is the set of protein family databases (such as BLOCKS [HH91], PRINTS [AB94] and Pfam [SEB+98]), which at one level are simply a clustering of the protein databases, but actually provide considerable added value in the form of annotation, links to other databases and to literature and - crucially - algorithms to make use of the contained information for protein family analyses.

This dissertation addresses some of these issues by example. It is divided into two sections. The first half describes some mathematical and technological developments in the new Bayesian approach to sequence analysis. There is some practical development of the ideas - in particular a software tool for analysing the accuracy of sequence alignments - although the section is essentially theoretical. The second half of the thesis describes construction of a database for the study of gene duplication events in the nematode *Caenorhabditis elegans*, whose genomic sequence was recently completed [CSC98]. This section is more applied than the first and the emphasis is more on the biology than the mathematics, though a number of algorithms and tools are developed that have wider applicability than

the database construction problem described. It is hoped that the results of the first half inform the second half, and that this account demonstrates how sound mathematics can provide a solid foundation for the development of natural and intuitive tools and data sets.

The first half of this introductory chapter begins by outlining the context and history of biological sequence analysis, with particular reference to the use of Bayesian statistics. It proceeds to summarise the work described in the first half of the thesis. The second half gives a brief review of the study of the evolutionary models and mechanisms of gene duplication as a prelude to describing the work in the second half of the thesis. First, sequence analysis.

1.2 Sequence analysis

Sequence analysis of peptides can be seen as the biologist's practical response to the intractability of predicting a protein's higher-level structure and function from its sequence [Sha97]. Nature uses a limited number of structural motifs to construct its cellular machinery - over 40% of known protein sequence belongs to under 1500 families or "domains" [Cho92, SEB+98] - and these structural homologies often correspond to sequence-level homologies, representing as they do an evolutionary connection by means of the gradual accumulation of mutations [Kim83]. The idea of protein sequence analysis is to search for these homologies and to exploit them in order to make inferences about the structure and function of novel proteins based on experimentally determined properties of well-characterised proteins [DEKM98]. The mutations that accumulate during evolution typically include residue substitutions and small insertions and deletions; there are thus a number of ways in which two sequences could be related and the exact nature of a homology between a set of sequences is usually represented in diagrammatic form as an "alignment". Part of a multiple alignment of protein sequences from the rhodopsin-like G-protein-coupled receptor family [AF94] is shown in Figure 1.1: sequences are laid out in rows with gap

characters inserted so that homologous residues are aligned in columns. The time-intensive task of working out where to put the gap characters in order to get the best alignment is ripe for automation; algorithms that do this task are called *sequence alignment algorithms.* Parallel problems are encountered in DNA sequence analysis; the precise mechanisms by which nature recognises which parts of the DNA of a cell are genes, and which parts are regulatory sequences controlling the expression or splicing of the genes, are either unknown or too difficult to model completely. However, statistical comparisons with well-characterised sequences can answer some of these questions and are regularly used to locate genes in newly sequenced DNA [Hau98].

In recent years, the theory of sequence analysis has benefited considerably from the discovery [KBM+94] that many of the alignment algorithms it had been using - which had been classified under the broad umbrella of "dynamic programming" - could be related to a mathematical framework that had been used very successfully in other fields, notably speech recognition [Rab89]. The framework is that of hidden Markov models (HMMs). The probabilistic nature of the HMM formulation provides strong links to the field of Bayesian statistics, a powerful revision of statistical ideas that has enthusiastic support in the machine learning community where HMMs were primarily in use. A brief discussion of the context of Bayesian methods and of pre-HMM sequence analysis may illustrate the happy significance of their combination.

1.2.1 Bayesian methods

At the most fundamental level it can be difficult to pin down the difference between the Bayesian approach to statistics and the classical or "frequentist" approach against which the Bayesians set themselves in opposition. All statistics essentially involves postulating probabilistic models and seeing how well these models fit observed data. Perhaps the definitive mark of Bayesian analysis is the emphasis on the application of Bayes' rule to likelihoods and priors

Figure 1.1: Part of a multiple alignment of protein sequences from the rhodopsin-like 7-transmembrane receptor family [AF94], displayed by the BELVU alignment browser [SD94]. Each row represents a sequence in the family. Gap characters (in this case, dots ".") are inserted into the sequences so that homologous residues are aligned in columns. In this figure, residues are shaded by column conservation.

in order to obtain a *posterior* probability denoting a level of belief in a hypothesis [Mac92a]. In the author's personal experience, Bayesians tend to be convincingly more open to discussion of the nature of the models they use and of the fundamental ideas in probability that underpin their methods. Counterbalancing this refreshing openness there is often a lack of interest in heuristic approaches that smack of frequentism, even though such approaches are often later found to have good theoretical justification.

Many biologists find this skepticism unnecessarily pedantic. Nonetheless, it is from the Bayesian camp that the theory of hidden Markov models - one of the most exciting developments in bioinformatics in recent years - has emerged. To understand the significance of HMMs it is useful to put them in context, by delving a little deeper into the history of sequence analysis techniques. The following brief sketch draws on the accounts in [Wat95] and [DEKM98].

In 1970, Needleman and Wunsch published the first dynamic programming alignment algorithm for aligning pairs of sequences [NW70]. Their algorithm finds the highest-scoring path through a dynamic programming matrix by exploring out from the top-left corner (see Figure 1.2). A path through the matrix corresponds to an alignment of the sequences, which lie on the horizontal and vertical axes of the matrix. The Needleman-Wunsch paper was succeeded by a number of further algorithms for pairwise alignment, e.g. [SW81, Got82, WE87, AGM+90]. A range of publically available programs implementing these algorithms are available, two of the most widely used being SSEARCH [PL88] and BLAST [AG96]. A considerable amount of research has been directed towards the problem of assigning statistical significance to the scores obtained by these methods, most notably by Karlin and Altschul [KA90, KA93, AG96].

In parallel with pairwise alignment, many algorithms for simultaneous or progressive alignment of multiple sequences were also developed [SK83, WP84, CL88, FD87, Tay87, BS87, HS89]. Without going into excessive detail, these algorithms attempt to find shortcuts to calculating the full multi-dimensional

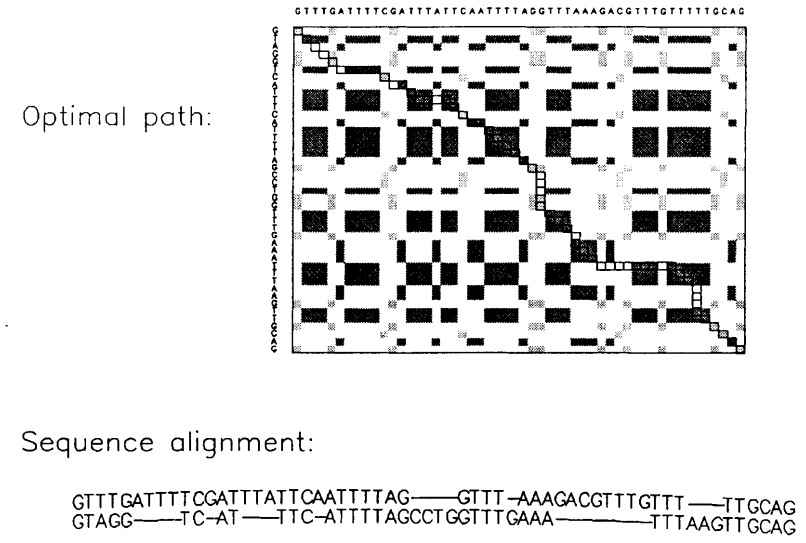

Figure 1.2: The dynamic programming matrix for Needleman-Wunsch alignment of two intron sequences from *C.elegans*.

analogue of the two-dimensional dynamic programming matrix in Figure 1.2 (a task too lengthy for practical analysis of large numbers of sequences). Typically, the algorithm estimates a "guide tree" and progressively aligns subgroups of sequences. Some programs implementing these algorithms are `PILEUP` [But98] and `CLUSTALW` [THG94a]. The latter makes use of sequence profiles during the progressive alignment. Sequence profiles are structures that model the position dependence of the gap and substitution propensities; see e.g. [GV96, THG94b].

In 1993 and 1994, Anders Krogh and co-workers from the Haussler group at UC Santa Cruz published several papers on the use of hidden Markov models (HMMs) for modelling protein and DNA sequences [HKMS93, Kro94, KBM+94, KMH94]. Hidden Markov models were previously widely used in speech recognition [Rab89]. Krogh and co-workers realised they provided a formal probabilistic model with all the behaviour of sequence profiles [DEKM98]. The Santa Cruz group published numerous subsequent papers on the use of HMMs in sequence

analysis (e.g. [BHK+93, SKB+96, HKMS93, KHRE96, REKH97]) and HMM ideas were also quickly latched onto and extended by workers at the MRC Laboratory of Molecular Biology in Cambridge, England [EMD95, MD95, Edd95, KM95] and other places [BC94, BH96, PBBC96]. The probabilistic view of sequence alignment had also been described by analogy with statistical mechanics [LAB+93, Miy94].

Hidden Markov models did not directly solve the problem that was perceived by many to be of chief mathematical importance in sequence alignment: that of assigning statistical significance levels to the alignment scores. What HMMs did do was to form a solid connection between biological sequence alignment algorithms and Bayesian machine learning. The Bayesian framework yielded insight into how sequence profiles and other alignment "machines" should be "trained" on data; some of the immediate dividends (there were many) included the principled justification of pseudocounts for sparse data sets [SKB+96], the discriminative training framework [EMD95] and the use of the Baum-Welch algorithm for finding the optimal scoring scheme [Kro94, HK96]. The probabilistic formulation has been extended to pairwise alignment [BH96, ZLL98] and it has also been widely used for gene prediction (see [Hau98] for a review) as well as more eclectic HMM architectures [BD97]. The connection to Bayesian machine learning research continues to generate interesting new ideas such as probability density networks [Mac96a, Pov98] and Fisher kernels [JH98].

With this context, Part I of this thesis can be summarised.

1.2.2 Summary of Part I of the thesis

Part I of this dissertation comprises three chapters exploring issues that have arisen during the development of the HMM theory of sequence analysis. Chapter 2 is in the nature of a mathematical introduction to the remainder of the dissertation; it reviews some elementary terms, definitions and algorithms that are central to HMM theory and introduces certain ideas that will be used later

on. The algorithms described can be divided into two categories: those pertaining to alignments of sequences to Markov models, and those pertaining to parameterisation of the model. Many of these algorithms are already well-explored, but there is some new material. Some of the new material introduces the elements of alignment accuracy and Bayesian decision theory that will be used in the second and third chapters. The rest outlines some possible improvements to training and model comparison algorithms for HMMs. A statistical mechanical view of HMM score distributions is also outlined; this will also be useful for Chapter 3. The idea of the generalised HMM is touched upon, and finally some molecular evolution models are introduced.

Chapter 3 of the dissertation describes investigations into the issue of how accurate a sequence alignment algorithm is. In the HMM view, a dynamic programming alignment is essentially trying to reconstruct an evolutionary history from what can be seen as noisy data. The question addressed in this chapter is: to what extent does this reconstruction procedure give an accurate result when the assumed evolutionary model is the correct one? (Of course, in reality the model is not actually expected to be precisely correct - it is, after all, just a way of finding homologies between sequences - but this analysis gives us insight into what kind of errors the algorithm would make even in the best of circumstances.) The question is explored using computer simulations and a Bayesian technique for extracting very weak alignments from sequences is presented and evaluated.

Chapter 4 of the dissertation - the final chapter of Part I - describes a practical tool `postal` that was designed with the aim of applying some of the more useful results from Chapter 3 to the analysis of protein sequences using HMM profiles. Given a multiple protein sequence alignment, `postal` uses posterior probability techniques (described in Chapters 1 and 2) to identify which sequences (and which parts of those sequences) may be poorly aligned. While `postal` does not identify every misaligned sequence, it can pick up some ob-

vious errors and flag low-information-content sections of alignments. It is thus suitable for use as a semi-automatic quality control tool for curators of large databases of gapped multiple alignments such as the Pfam database [SEB+98]. `postal` is constructed using parts of the HMMER package [Edd96].

1.3 Gene duplications

Part II of the dissertation deals with the application of some of these techniques, along with other bioinformatics methods, to a specific problem of interest in molecular evolution: the study of gene duplications in the nematode *Caenhorabditis elegans*.

Why is it interesting to study gene duplications? The study of evolution is of central academic interest because of the attractive - if often elusive - idea of a driving principle underlying biology. Evolutionary frameworks can lend meaning and context to descriptions of biological mechanisms and hence organise knowledge. One such framework is the hypothesis of gene evolution via duplication and genetic redundancy, which suggests that new genes evolve by duplication of existing genes. When a gene is copied, the selective constraints on each copy are proposed to be relaxed, so that each copy is free to evolve (slightly) modified function [Ohn70, Oht89]. With international sequencing collaborations generating data for organisms' entire genomes rather than isolated genes or fragments of chromosomes, the prospect of analysing the long-timescale dynamics of genes in genomes is a realistic option. The completion of the genome sequence of the yeast *Saccharomyces cerevisiae*, long known to contain a number of gene duplications [Smi87, Ols91, Kab95], enabled the beautiful demonstration by Wolfe and Shields [WS97] that the chromosomal positioning and orientation of the duplicate genes were consistent with duplication on a large scale by the hypothesised mechanism of whole-genome (polyploid) duplication proposed by Ohno [Ohn70].

As the largest eukaryotic genome to be completely sequenced (and the first

animal genome), the nematode *C.elegans* is a natural candidate for further systematic study of gene duplications [CSC98]. The progress of the nematode genome project has seen a rapid increase in the number of gene families to be characterised in this organism, with functions ranging through neuronal development [NKML98, BH97], chemosensation [Rob98, TCD+95], miscellaneous cell signalling and development [PS98, TW96, SDF+96, SLP+96, BR93] and other categories [AKW+98, DGPB98, LC95]. The resolution of standards and file formats in the annotation phases of the genome project and improvements in the technology of protein family databases have enabled automatic classification of many gene families [SEB+98, CSC98].

A large number of *C.elegans* gene families are found to comprise genes that are located close together on the chromosomes. These gene clusters are of particular interest since *C.elegans* was the first eukaryotic genome found to contain *operons* [BS97]. Operons in *C.elegans* consist of clusters of genes typically separated by around 100bp, that are transcribed together as a single strand of pre-mRNA and subsequently separated by *trans*-splicing. Co-transcription implies shared modes of transcriptional regulation. Many operons contain groups of genes that are not homologous and therefore cannot be said to have arisen from duplication; some of these non-homologous clusters are known to code for genes whose functions demand co-expression and, in these cases, a clear argument can be made for the evolutionary importance of the transcriptional co-regulation experienced by these genes [Blu98]. Operons were well-known in bacteria and archaea before their discovery in *C.elegans*; they have been found in other orders of the nematode phylum [EZM+97] and in flatworms [DH97].

1.3.1 Agents of change

There are a number of ways that duplications of sections of genomic DNA can occur in nature. Duplications in *C.elegans* can be induced in the laboratory by irradiation or by the introduction either of mutagenic chemical agents

(formaldehyde, for example) or of mutator loci that up-regulate transposable element activity such as *mut*-2 [PvL97, JB97]. All these types of duplication can occur in the wild, with the additional possibility of random errors in replication and recombination. In some strains the principle cause of spontaneous duplications and other mutation appears to be transposable element activity [PvL97]. These elements are worth describing in more detail.

Transposable elements (or *transposons*) are well-defined (albeit fast-evolving) sequences that are found to spontaneously copy themselves (or have themselves copied) into genomic DNA [Jur98, PvL97, Smi96]. They are distinguishable from viruses - their closest relatives [GL95a, Jur98] - by not normally crossing cell boundaries invasively; they are not infectious. For this reason they are generally less destructive to their hosts and more restrained in their rate of proliferation. Transposons may be primarily subdivided according to whether the intermediate genetic component in the transposition cycle is DNA or RNA. DNA-mediated transposition proceeds by a "cut-and-paste" mechanism, whereby transposase proteins excise the mobile sequence and re-integrate it at the target site by DNA cleavage [PvL97, HLL97]. While the cut-and-paste mechanism is not itself replicative (it merely moves the sequence around), repair of the double-strand breakage from where the sequence was excised can generate a copy of the transposable element from the homologous chromosome, which the repair machinery uses as a template [vLCP94]. Characteristics of DNA transposons include flanking inverted repeat sequences and site-specific integration, both consequences of the transposition mechanism [HLL97]. RNA-mediated transposition, on the other hand, involves first transcription by RNA polymerase, then reverse transcription back to DNA by reverse transcriptase [Jur98]. Transposons may be further subdivided by whether the genes coding for the transposase proteins that catalyse transpositon are contained within the transposon sequence itself (in which case the transposon is said to be *autonomous*) or elsewhere (in which case the transposon is a *non-autonomous*

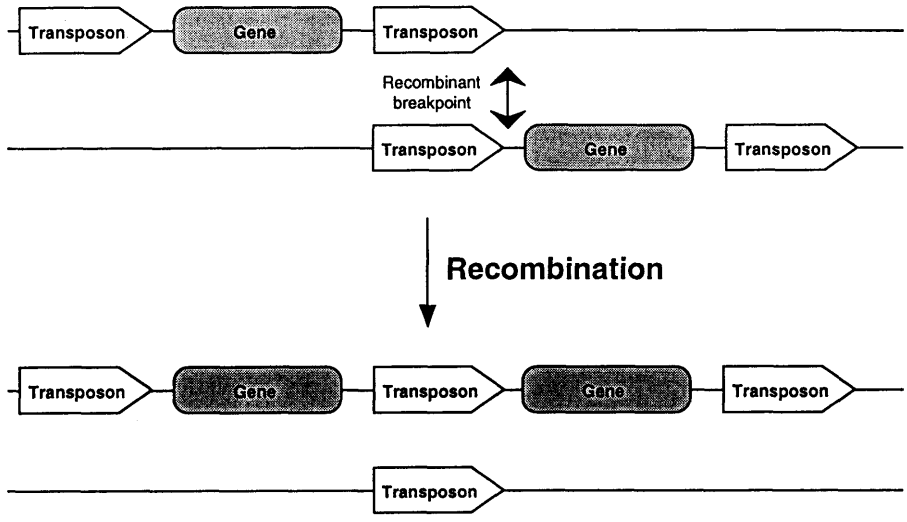

Figure 1.3: Transposons can precipitate gene duplication during recombination if pairing between adjacent copies leads to unequal crossing-over.

"hitch-hiker") [Smi96]. All of the above kinds of transposon are found in the sequenced Bristol N2 strain of *C.elegans*. The completion of the genomic sequence has led to a rise in the number and variety of transposons and other repetitive sequences reported in *C.elegans* [OGB95, OGB96, RvLDP97].

Transposons can promote gene duplication during meiotic recombination if homologous pairing between adjacent copies of a repetitive element leads to unequal crossing-over as in Figure 1.3 [NCC+92, YWB97, FBT+91]. (A similar mechanism underlies gene conversion, the phenomenon of expansion or contraction of existing clusters of homologous genes [LG91, Rob98].) Double-strand breakage repair following transposon excision can lead to local duplications [MKW91] and it is also possible that transposons may carry flanking sequence with them when they transpose [GL95b]. Transposon insertions into all kinds of host-important sequences - including introns, exons and promoter regions -

have been observed [Wes89, OGB95, BHP89]. Indeed the potential for transposons to have quite drastic effects on the evolution of their host organism is considerable [HLNL97] and the mechanisms by which their transpositional activity is regulated so as not to over-burden their hosts' replicative machinery have been topics of some interesting research [LC97]. One of the best-studied of transposon families is the relatively simple Tc1-*mariner* group, of which several variants are known in *C.elegans* [PvL97]. The canonical Tc1-*mariner* transposon in *C.elegans*, Tc1, is a DNA-based transposon consisting of an inverted repeat flanking a two-exon gene that codes for a single transposase protein. The protein catalyses the entire "cut-and-paste" transposition process [VBP96]. The ecology of *mariner*-like transposons is better understood than most; there are a number of ways in which these elements can interact to reduce transpositional activity [HLL97], including burdening of active transposase by non-autonomous and autonomous-but-defective transposons ("transposase titration" and "subunit poisoning" respectively) and poorly-understood interactions between active transposase proteins ("overproduction inhibition").

The aspects of genome duplication described above have been observed in the laboratory or in isolation, but an overview of their relative importances to the dynamics of gene duplication is lacking. With the completion of the *C.elegans* DNA sequence, the timing is ideal for a genome-wide systematic study of gene duplication in a model animal.

1.3.2 Summary of Part II of the thesis

Part II of this thesis dissertation describes the construction of a database, named Wormdup, to facilitate the study of genome duplications in *C.elegans*.

Chapter 5 (the first chapter of Part II) describes the construction and preliminary analysis of the main portion of the Wormdup database. The purpose of Wormdup is to provide a resource for answering questions about the sizes, frequencies, locations, causes and other aspects of genomic duplications. The

chapter describes how the data in Wormdup are used to calculate various molecular evolutionary parameters for *C.elegans* such as the transition/transversion ratio, the rate of small indels and the rate and size distribution of fixation of non-coding and coding duplications. The fixation rates for non-coding and coding duplications are compared and the results are discussed with reference to the reliability of molecular clocks in general and the selective pressures that may act on duplicated DNA.

Chapter 6 (the second chapter of Part II) looks at a new method of dating gene duplications. Many of the interesting questions in molecular evolution rely on dates and times of speciation and divergence events. To answer these questions, it is useful to have an accurate "molecular clock" that can be used, for example, to compare rates of duplication for coding and non-coding DNA as in the previous chapter. The most commonly used kind of clock counts the number of nucleotide substitutions that have occurred at synonymous codon positions in the sequences and uses this to estimate a maximum-likelihood divergence time. This kind of clock is called a "codon clock". In this chapter a new kind of "intron clock" that counts substitutions and indels within conserved introns is proposed and evaluated using Bayesian techniques described in the first part of the thesis. The results suggest that intron evolution can be fitted to time-dependent models but that the clocks, as proposed, do not synchronise well with codon clocks. Possible reasons for this and potential improvements to the model are discussed.

Chapter 7, the final chapter of Part II and of the dissertation, looks at a specific class of DNA-based transposon - the Tc1-*mariner* group - that has been well-studied in nematodes. Using hidden Markov model and other dynamic programming techniques, statistics are obtained for the representation of previously characterised *mariner* families in the sequenced strain of *C.elegans*, as well as for six previously uncharacterised families. The new families are analysed using protein sequence analysis and phylogenetic techniques and found to be more

closely related to one another than the previously characterised families. These results are discussed in the context of transposon ecology and evolution.

The dissertation ends with an appendix which describes the software tools that were developed specifically for this project (but with re-useability in mind). It is hoped that these may prove useful to other projects.

1.4 Statement of originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration.