Chapter 8

Conclusion

The evolution of genetic material is fundamental to the whole of biology; the sequencing and analysis of whole genomes has begun a revolution in scientific understanding. As we gather more and more data about how cellular processes work, we will need more powerful software agents to cluster, filter, organise and digest the information so that we can get on with the creative process of interpreting, describing and acting upon it. Designing these tools - and being the first to use them - is what bioinformatics is all about.

This thesis has presented work on the theory of biological sequence alignment and the evolution of the first sequenced animal genome, *Caenorhabditis elegans*. The sequence alignment theory has looked at a number of issues, addressing the question of accuracy - how accurate can an alignment be? - and discussing how to side-step this issue by summing over all alignments (and over a range of scoring schemes). The evolutionary investigations have looked at questions associated with genomic duplication in the nematode worm *C.elegans*, including causes and characteristics of duplications and the random divergence of sequences following duplication. The two approaches have informed each other closely: e.g. the measurement of substitution and indel rates motivated the development of parameter-estimation algorithms and the development of a Bayesian framework for model comparison enabled the evaluation of a new molecular clock based on introns.

The fusion of ideas from computer science and molecular biology is one of the things that make bioinformatics an exciting field. The development of the Bayesian view of sequence alignment is a shining example of this fusion. Sequence alignment is the king pin of homology analysis which, at a structural, functional and evolutionary level, shapes our understanding of protein families and the patterns of process that nature has used. With the data that fund this analysis multiplying rapidly and the study of protein families blossoming, now is the right time to build solid foundations for sequence analysis so that issues like uncertainty and accuracy are not awkward embarassments but robust

parameters of the theory. Bayesian statistics, with its concept of a probability as a level of belief in a hypothesis, is an ideal framework in which this process can work. It is hoped that this dissertation has pointed out some of the fronts on which progress can be made. It has been said that Biologists stole Statistics from Physicists [Jay86]; they now have a chance to steal it again.

Flourishing technologies such as expression analysis by microarrays and ESTs provide yet more opportunities. Computational biology has an important role as interpreter for data - such as these - that are so voluminous they can only be visualised at a statistical remove. The basic kinds of operation that one tends to want to do on these data include clustering, pattern identification, construction of models for these patterns, classification of the data according to these models, construction of new models and collection of new data - the iterative loop of refining models and, behind the models, biological ideas. The particular algorithms for analysing each class of data will be different; finding the right approach demands mathematical skill and intuition together with a broad biological awareness so that the practical issues may be separated from the mathematical curiosities. However, bioinformatics has matured and is ready for the challenge. A network of computational biologists exists, with the background and the abilities to respond to these kinds of technological advances.

One of the things that can drive new algorithm development is a convincing format or specification. As an example, the GFF processing tools developed for this project would probably have remained a collection of throwaway scripts were it not for the stabilisation of a good, simple format that not only represented annotative data well, but was seen to gain enough support in the bioinformatics community that it seemed inevitable that it would catch on. This has a lot in common with the history of HMMs and Bayesian methods in sequence analysis. With the increasing heterogeneity of data and the diversity of sequenced animal genomes, standardisation will tend to become the path of least resistance as organisations start to seek common ways of handling and sharing their data

without wasting time on different protocols for each organism or experiment.

This brings us to another exciting aspect of bioinformatics which is, of course, its proximity to the genome projects. As more complete animal genomes become available we may hope to move beyond individual, anecdotal accounts of positive selection or selective sweep and towards an understanding of evolutionary dynamics that may be increasingly quantitative. The ecology of transposons and viral elements and their role in stimulating animal evolution is a fascinating topic that has led to much speculation. This speculation is now confronting hard data produced by the sequencing effort. Soon we can hope to start to develop a rounded account of evolutionary history that speaks directly of molecular mechanisms. This will be a major humanistic triumph and a direct consequence of the genome projects.