

# “From Population Genomes to Global Pedigrees”

Professor Jotun Hein and Professor Mike Steel

## **Part I: Previous Research Track Record of Professor Jotun Hein**

I moved to the Department of Statistics at Oxford University in September 2001, coming from Aarhus University, where I was director of BiRC (Bioinformatics Research Center). Since I have come to Oxford I have worked on a series of issues. Most of my work is focused on developing methodologies that can analyze data arising in molecular genetics and genomics with a comparative/evolutionary focus. The work can be grouped into following headings:

### **A Statistical alignment**

The motivation for studying, modelling and making algorithms for statistical alignment is that most sequence analysis has in the last decade or more benefited tremendously from the use of stochastic models of sequence evolution. This allows parameter estimation, hypothesis testing and more. The alignment problem has not benefited from a similar development. We have already made significant progress: accelerated the basic algorithms enormously, generalized the pair wise algorithm to any number of sequences; we can model longer insertion-deletions. To complete the project, alignment models should allow a heterogeneity among positions, that could then allow these methods to be combined with gene finding and RNA structure prediction algorithms. These projects should be followed up by the development of user-friendly software, which will be quite laborious. These projects have been tremendously successful and our group is now a world leader in this field.

### **B Sequences subject to both RNA structure and protein coding constraints**

I have developed methods in collaboration with Irmtraud Meyer, Roald Forsberg and Jakob Skou Pedersen that modelled constraints from the protein level and also methods that incorporated constraints from RNA structure, but these selection levels have never been combined. This has a very wide range of applications. All protein coding sequences, where mRNA is under some RNA Structure constraints are in need of such analysis. These methods will be especially relevant in the analysis of viral sequences. These methods should allow a quantification and separation of the effects of these two selection levels. Viral evolution at times display an enigmatic mode of evolution, that seem fast on a short time scale, but slow on a long time scale. Either this is a reality or an artefact of unrealistic models of sequence evolution. More realistic models should be able to settle this issue. This project has an interesting corollary – the possibility of gene finding in virus. This have not been done before and finding ignored reading frames in virus would of great importance.

### **C Haplotype block definition via reconstructed evolutionary histories**

In the wake of the determination of the human variation genome sequence, projects have been launched to determine variation in the human genome. A major issue is if recombination occurs uniformly along the chromosome or if it is clustered. It has been postulated that the human genome can be partitioned into large blocks and recombination only occurs between these blocks. Yun Song and I have developed a method that will find a history of a set of sequences that minimizes the number of recombinations plus substitutions and displays a history of the sequences that use this minimal number of events. This history will also find a set of intervals within which there hasn't been any recombinations. This is a rational definition of a haplotype block. It is highly computationally demanding. Our next goal would be a heuristic, but faster version and to allow gene conversions to be found as well.

### **D Pathogen Analysis**

Alexei Drummond, Roald Forsberg and I have been interested in methods used for studying virus evolution for a long time and our residence in the Medawar Building is optimal for research in this field. Due to larger population sizes, shorter generation times and higher mutation rates, pathogens evolve on a faster time scale than larger organisms. Hence, the study of molecular pathogen evolution often requires that existing methods of analysis are modified or that novel methods are developed. We have been involved in both, with main emphasis on the two areas of molecular epidemiology and molecular

evolution. An interesting consequence of the rapid rate of pathogen evolution is that samples of pathogen genetic material obtained over even short periods of time can be shown to be measurably evolving, i.e. show a significant accumulation of new genetic variation during the sampling period. This enables the use of serial pathogen samples to estimate rates and dates of the evolutionary and epidemiological process without the use of external calibration which is extremely useful in elucidating the epidemiological history of emerging diseases. We have developed several new models for the analysis of serial samples and furthermore we have developed novel population genetic models to analyse the epidemiological process of pathogens. Using the above mentioned methods, we are currently studying the molecular evolution of Foot and Mouth disease virus and the HIV virus. In the area of molecular pathogen evolution, we are involved in an ongoing effort to develop new methods for inference in codon-based models. These allow an evaluation of the position-specific selective regime operating on pathogen proteins. The main focus here is to develop models that incorporate environment specific selective pressures and changes in this. We have used these methods to study the evolution of the influenza A virus in different hosts and to identify sites potentially involved in host-specific adaptation and are currently working on the refinement of these methods.

## **E Comparative Genomics**

Irmtraud Meyer, Gerton Lunter and Jakob Skou Pedersen work full time on this, but several other aspects of our work could also fall under this heading due to its comparative aspect. The most basic investigation is to find genes in a pair of aligned genomes. However, this can immediately be extended to many sequences, unaligned sequences, allowing for alternative splicing, RNA gene finding and a stronger focus on specifically motivated biological problems. Stephen McCauley works on extending these models to annotating large sets of viral genomes.

## **F Molecular Population Genetics:**

I have a few running projects, especially with Carsten Wiuf and Mikkel Schierup from Aarhus University. These projects especially involve association mapping and gene conversion. We (Hein, Schierup and Wiuf) have published a 30 page book on molecular population genetics titled "Gene Genealogies, Sequence Variation and Evolution" Oxford University Press. The present project on pedigree combinatorics and reconstruction is an outgrowth of my interest in molecular population genetics.

Bioinformatics is a field strongly in demand and we have contributed to the development of this field. Besides being an active research group, we do much to increase awareness of this field in terms of public courses, seminars and organize lecture series. Additionally, we teach a part time MSc in Bioinformatics and thus educate researcher in this field. Presently, I have 5 D.Phil. students (2 co-supervised with MRC Harwell and Wellcome Centre for Human Genetics).

Being part of the Department of Statistics and physically placed in the Oxford Centre for Gene Function (OCGF) is ideal for developing methodology of Bioinformatics and engaging in collaboration and data analysis.

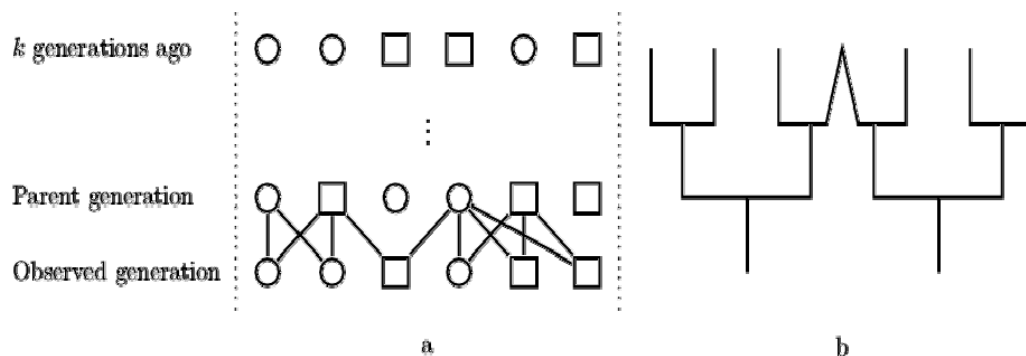
## Part II: Proposed Research

### A Introduction of topic of research and its academic and wider context

Central to biology is homology and genealogical relationships. Due to the phenomenal growth in sequence data from different species, phylogenetics has risen to prominence and been put on much firmer statistical ground (Felsenstein, 2004; Semple and Steel, 2003). Similarly intra-population variation has also been catalogued on an unprecedented scale and also led to better characterization of the genealogical relationship of sequences sampled from a population (Hein, Schierup and Wiuf, 2005). In this case the possibility of modelling the reproduction of a population (in contrast to speciation) has given a natural prior distribution on genealogical histories. The presence of recombination has both made the problem much more complicated (involving the ancestral recombination graph), and shifted the focus from estimating a specific history to inference of the probability distribution on such histories. Besides the genealogical concepts of phylogeny and the ancestral recombination graph and their inference, the deluge of sequence data also herald the possibility of pedigree inference on an unprecedented scale. The concepts of phylogeny and pedigree are well known concepts all throughout the scientific community (even to the general public), and are very old – hundreds and thousands of years, respectively. The term ‘Pedigree’ could together with ‘individual’, ‘life’ and ‘species’ be the oldest biological concepts. But in terms of mathematical and combinatorial studies, pedigrees have received less attention than phylogenies. This project aims to investigate the combinatorial space of pedigree graphs and in particular the connections between a pedigree and its embedded phylogenies, ancestral recombination graphs, and local pedigrees. As a practical outcome, this will further our understanding of the amount of data required to reliably reconstruct a pedigree, and lead to novel algorithms for pedigree reconstruction.

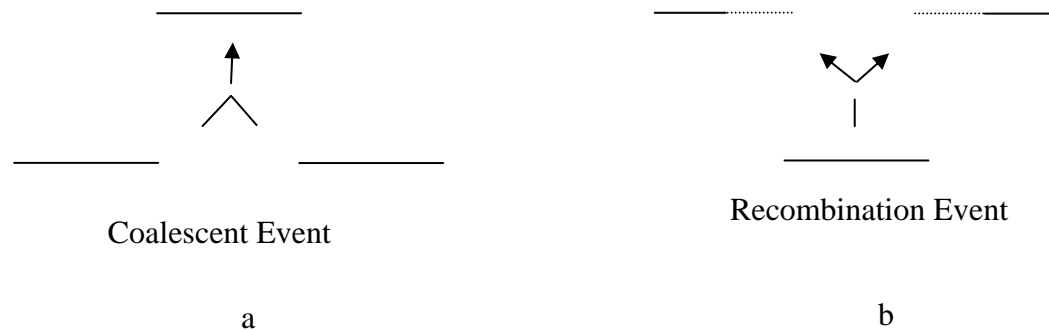
In the current setting, a *Pedigree* refers to a graph with extant individuals labelled with distinct names and unlabelled ancestors. As a simplifying assumption we require that ancestors occur at discrete generations going back in time, at least initially, though counterexamples abound in natural pedigrees as well as pedigrees in animal and plant studies. The individuals will be nodes and may or may not be labelled by gender. Individuals will have two edges that point to their parents (father and mother) in the previous generation. In the figure, the leftmost illustration (a) shows part of a global pedigree. At the present we find the extant individuals, and in the generation before their (gender labelled) parents; this can in principle be extended arbitrarily far back. In the middle illustration (b) we have extracted the pedigree (without gender) of two individuals two generations back, and they have a single grandparent in common.

Tracing the history of sequences back in time could in principle yield information on the pedigree of the individuals carrying them. Any sufficiently small segment could be traced back to a common ancestor segment through a specific path up within the pedigree. Under generous assumptions about rates of mutation and recombination the length of this could be recovered directly by comparing homologous segments.



The *Ancestral Recombination Graph* (ARG) relating a set of sequences can be described by starting in the present and going backwards in time until all positions of the sequences have found one single ancestor. Going back in time sequences encounter mutations, coalescences and recombinations. Mutations (backwards in time) will change a single position in a single sequence from the mutant state to the ancestral state. Coalescent events will merge sequences that are identical, reducing the sample size by one. Recombinations will redistribute a single sequence to two sequences, where one sequence will carry the material to the left of the recombination point and the other the material to the right of that point. Mutations do not influence the genealogical relationship of the sequences, but their actual content in terms

of differences or actual observed nucleotides. Below is illustrated the effect of a coalescent event and a recombination event on the distribution of ancestral material. These two event classes relate to pedigrees as follows: sequences that coalesce must occur in siblings (possibly half-siblings) and the merged sequence exists in their (shared) parent. A recombinant sequence can be found in the eggs or sperm of an individual and thus in its offspring, and the two molecules creating the recombinant is on homologous chromosomes of that individual. One is inherited from its mother and one from its father. Tracing backwards the ancestry of the diploid genome of an individual will, for obvious reasons, have to be within the pedigree of that individual.



Where pedigrees and ARGs can represent complex networks of descendancies, a phylogeny describes a simple tree-like relationship between individuals. A phylogeny assumes a hierarchical clustering of relationships, where any individual can trace only a single lineage back to the common ancestor.

The three basic genealogical structures are closely related: The ARG for a point (not an interval) is a phylogeny, if a pedigree is pruned by tracing only one parent for each individual a phylogeny will be obtained, and the ARG is embedded within a pedigree. The pedigree of a population thus constrains the ARGs and phylogenies observable for the population. Conversely, this means that the ARGs and phylogenies of observable in a population will be informative about the pedigree of the population. This pedigree-focused proposal will repeatedly refer to properties of the phylogeny and the ARG.

As described above, the genealogical structures are combinatorial objects, but there are natural probability measures induced by the evolutionary processes of reproduction, recombination and mutation. There are questions that do not rely on probability measures, such as "Do sequences under idealised models characterize the global pedigree?" Similarly, there are questions that do rely on probability measures, such as "What is the probability that the pedigrees are characterized  $k$  generations back by all the genomes of the population?" Both classes of questions are interesting in their own right.

The fundamental models defining the probability distributions on pedigrees are almost inherent in the models underlying coalescent theory. The missing component is marriage formation (possibly including an infidelity parameter) and progeny distribution for a given marriage. These have been explored and can easily be incorporated in population reproduction models. This naturally defines a prior distribution on pedigrees.

The main question is how reliable is pedigree inference as a function of generations back in time. It should be expected that reliability is good for a few generations but then tails rapidly off, but if this is 3-4 generations or more than 15 is presently unknown. If it was beyond the latter number, large scale sequencing would eventually have great potential to aid historical demography going back many centuries. For example, if it is 15 generations, one could reconstruct the pedigree of all European descended New Zealanders based upon a large sequencing project.

## **B Past and current work in the subject area**

Steel and Hein (2005) showed how to construct global pedigrees from pair wise pedigrees and from two classes of paths embedded in the global pedigree. The authors also provided a lower bound on the number of global pedigrees and the minimum number of segregating sites needed to reconstruct the global pedigree. The main conclusion from this paper is that reconstructing the global pedigree is much easier if the information on ancestors include their gender. This is an important practical issue as the gender information of ancestors will be difficult to obtain from genomes. The paper only considers the problem of reconstructing pedigrees from certain embedded pedigrees, however. It remains an open

problem whether knowledge of the genomes of individuals determines the global pedigree under idealised models of mutation, recombination and reproduction.

## C Main aims of the project

**I. Counting Pedigrees.** Understanding the underlying structure and size of the set of pedigrees is essential to evaluate the hardness of pedigree inference and to formulate appropriate algorithms operating on the set of pedigrees. The asymptotic growth of the number of pedigrees lower bounds the minimum amount of data required for pedigree reconstruction, and is thus highly informative about the feasibility of pedigree inference. Moreover, an improved understanding of the combinatorial space of pedigrees is essential for developing and analysing efficient algorithms for pedigree inference and analysis. The number of pedigrees of interest will only have extant individuals labelled and there are series of enumeration problems that will be pursued with different numbers of extant individuals from 1 to the whole population. The two extreme cases are of key interest, and a series of variants will be pursued for enumerating pedigrees with ancestors gender-labelled or not. Questions to be addressed include:

- How many ways can an unlabelled pedigree be gender labelled?
- What is the asymptotic growth of these different numbers?

The enumeration of phylogenies has been attracting attention for years. Recent work has also started to address the enumeration problem for ARGs. Surprisingly, the corresponding problems for pedigrees remain to be explored. Chen, Lyngsø and Hein (2005) found a simple recursion for gender labelled pedigrees. If  $L_i(g, p)$  denotes the number of distinct pedigrees with  $i$  extant individuals,  $g$  generations, and  $p$  ancestors in generation  $g$ , then

$$L_i(g+1, p) = \sum_{1 \leq c \leq 2^g} L_i(g, c) \sum_{\substack{1 \leq m, f \leq c \\ m+f=p}} S_c^{(m)} S_c^{(f)},$$

where  $S_i^{(j)}$  is the Stirling number of the second kind, i.e. the number of ways to separate  $i$  distinct elements into  $j$  non-empty sets. The number of distinct gender labelled pedigrees with  $i$  extant individuals and  $g$  generations can now be calculated as  $L_i(g) = \sum_{1 \leq p \leq 2^g} L_i(g, p)$ .

This number grows extremely fast, e.g. there are more than  $10^{1700}$  distinct pedigrees for one extant individual if we go ten generations back in time. The above formulas only provide recursions for computing the number of distinct pedigrees. Some boundary cases allow simple expressions, e.g.  $L_i(g, i2^g) = 1$ ,  $L_i(g+1, 2) = L_i(g)$ , and  $L_i(g, i2^g - 1) = i(2^{2g-2} - 2^{g-1}) + \binom{i}{2} 2^{2g}$ . However, no closed form expression is known for the general case. For pedigrees with no gender labelling, Steel and Hein (2005) provide a lower bound for global pedigrees with  $g$  generations and fixed population size  $n$  of  $\Omega\left(\frac{n!}{2^{n/3} \lfloor n/3 \rfloor!}\right)^g$ .

Interestingly, also for pedigree enumeration the unlabelled case seems to be much more difficult than the gender labelled case. This even though there are far fewer distinct unlabelled pedigrees than distinct labelled pedigrees for a fixed number of extant individuals and generations: the fact that the number of unlabelled pedigrees of  $i$  distinct extant individuals,  $g$  generations, and one less than the maximum number of ancestors in generation  $g$  is  $U_i(g, i2^g - 1) = g + \binom{i}{2}$  suggests at least an exponential separation between  $L_i(g)$  and  $U_i(g)$  in terms of  $g$ . Thomas and Cannings (2003, 2004) has also counted the number of pedigrees obeying certain constraints on extant and ancestral individuals.

**II. Identifiability and Reconstruction Principles.** Whether sequences under idealised models determine pedigrees or not is a major open problem and deserves serious attention. Idealised models will here assume arbitrarily high recombination and mutation rates such that i) the phylogeny relating a sufficiently small segment for a set of sequences is readily obtainable by the pair wise differences between the segments, and ii) the full set of phylogenies  $\mathbf{T}$  that could be obtained by each individual choosing only one of the parents would be known – for instance, for a set of extant individuals, only tracing paternal ancestry would lead to one phylogeny, but there are many other possibilities depending on which parent was

chosen for each individual. Additionally, one would know not only  $T$  but which  $p \in \Pi$  relates the sequences at any sequence point (called the *local phylogeny*), i.e. the mapping.

Although these assumptions are strong, it is not clear that different pedigrees couldn't be used to explain the same sequences. However, Steel and Hein (2005) gave an example of two distinct pedigrees with the same embedded phylogenies. Characterizing the set of pedigrees with the same embedded phylogenies would be of interest. There are other restrictions on the set of pedigrees compatible with the series of local trees, which could be used to further strengthen the pedigree inference. For instance, two neighbour local trees must be related by one recombination with associated coalescent event. As stated in the introduction, recombinations must occur in an individual and the individuals carrying the sequences just after a coalescent event must be siblings or half-siblings. The recombination part of this statement is wholly reliant on the relationship between neighbour local trees and this additional constraint relative to knowing all the embedded phylogenies. A complicating factor here is that the recombination point necessary to explain to neighbour trees is only rarely unique. More generally, the sequence of local phylogenies does not characterize the ARG (Hein, 1993). Had this been the case, it would be easy to prove that sequences characterize the pedigree.

III. Reconstructability of global pedigrees from local pedigree knowledge. Global pedigrees can be reconstructed (up to isomorphism) from local pedigrees on pairs of extant individuals, provided these local pedigrees are gender labelled (Steel and Hein, 2005). However, in the case where this gender information is not available for the local pedigrees, the authors showed that the reconstruction result from pairs can fail. Nevertheless this suggests a tantalizing question, posed in that paper: Do local pedigrees (without gender) on  $k$ -tuples of individuals (for some fixed  $k$ , independent of the size of the population) suffice for reconstructing a global pedigree? Thatté (2006) provides a counterexample; however, that counterexample requires a pedigree of depth at least  $(n - 2)$  and  $2^{n-1}$  ancestors at that depth. This result leaves unanswered whether this represents a minimal counterexample, how restrictions on the population size affect global pedigree reconstructions, and fails to provide a constructive algorithm for building global pedigrees from sub-pedigrees on restricted populations.

A further question is how to handle errors in local pedigrees (either in the gender-given or gender-free case). Ultimately one would like to provide a unified statistical approach whereby the confidence in alternative local pedigrees can be translated into confidence values for global pedigrees.

IV. Sharpening of bounds on needed segregating sites. Steel and Hein (2005) showed that the number of segregating sites required to accurately reconstruct a pedigree up to depth  $d$  (generations into the past) for an extant population of size  $n$ , must grow with a rate that is  $\Omega(d \log(n))$ . This bound, although general and robust to model specification, was obtained by a primitive counting argument based on a particular subclass of pedigree graphs. An interesting question is to determine how close this bound is to the actual number of segregating sites that might be needed. There are two questions here – one is purely combinatorial, and involves enumerating pedigrees of depth  $d$  on a population of size  $n$ . This alone may enable the asymptotics of the rate function to be strengthened from  $\Omega(d \log(n))$  to e.g.  $\Omega(dn)$  or  $\Omega(d^2 \log(n))$ . The second question is statistical – under standard models of mutation and recombination, how fast must the number of segregating sites grow as a function of  $d$  and  $n$ ? We know already it must grow at least as fast as  $d \log(n)$ , but it is conceivable it could grow much faster – for example exponentially with  $d$  – than any bound obtained from combinatorial considerations. Clearly this is an area that has much scope for further analysis.

Although little work appears to have been done on this question so far (beyond (Steel and Hein, 2005)), the parallel field of phylogenetics provides a case study in how enumeration results can guide statistical insights and conjectures. In particular, the direct enumeration based bounds that were obtained on the sequence length required for accurate phylogenetic tree reconstruction turned out to be generally the same (up to a constant) as the actual number required when sequences evolve under the sorts of Markov processes that are used for modelling the substitution process of aligned DNA sequences. Some of the techniques employed in establishing this equivalence (particularly for locating 'deep' resolutions in a phylogenetic tree from short sequences), may well be useful in the pedigree setting also.

V. Combining the individual and sequence genealogical processes exactly in small populations. We will program a simulation tool that can simulate the combined global pedigree and ARG for small toy populations, like population size 50 ( $N$ ), 50 generations ( $k$ ) and chromosomes with less than 100

expected recombinations per generations ( $R$ ). To simulate one generation, *i.e.* the genealogical history linking the present and previous generation, requires  $O(RN)$  time and this must be done for like  $3k$  generations. Simulating populations would need  $N > 10^4$  and we would be interested in letting  $R$  become large. Additionally, a large number of replications would be needed to map the distribution of required quantities. In summary, simulating real populations shouldn't be done with exact methods. But such a program can still be very useful to test theoretical results, to gain intuition about the processes, and finally to disseminate understanding of these problem to other researchers (like the demonstration software found at [www.coalescent.dk](http://www.coalescent.dk)). We would build a comprehensive, extensible, and open Java API (application programming interface) that would provide all the necessary tools for simulating and analysing pedigrees and pedigree data.

#### **D Future research and possible extensions**

Since both Jotun Hein and Mike Steel teach and attract students at their respective Universities to research, it is conceivable that additional research could be initiated and a few comments on further problems are in place. A large family of counting problems can be proposed – discrete generations, continuous time births, pedigrees with and without infidelity or close interbreeding. Different construction principles could be implemented in programs and applied to simulated data. How does the ancestral material distribute itself in individuals? For instance how much genetic material did the Universal common ancestors (Rhode et al., 2004) have? How far back in time can pedigrees be reconstructed reliably could be evaluated by simulations, but even beyond this limit there would be pedigree information although it might not allow the extraction of specific relationships and it would be relevant to visualize or map distributions on the set of global pedigrees.

#### **E Timeliness**

Due to the increased availability of complete genomes, the increased interest in human evolution and demography and finally the widespread attempts to put genealogical inference in all its forms on a solid statistical footing, the proposed project will address problems that are not only exciting but also will be of great practical importance a few years from now. The resources asked for are modest, but since both applicants teach and recruit students at both their universities this could lead to much more activity than the single researcher that will be recruited for the project.

#### **F Milestones, management of the project and justification for resources**

This proposals stems from collaboration between Jotun Hein (University of Oxford, Oxford, UK) and Mike Steel (University of Canterbury, Christchurch, NZ) and it is essential that the research assistant (RA) will have the opportunity to interact with both. Of the 2 years, the RA will be expected to spend a period of 6 months in New Zealand. Additionally, Mike Steel will visit the Oxford group once for 2 weeks and Jotun Hein will visit the NZ group once for 2 weeks. In the Oxford group at least 2 postdocs (Thomas Mailund starting February 2006 and Rune Lyngsø) are heavily involved in projects that would lend expertise to the present proposal (combinatorics and complexity of the ancestral recombination graph), besides being very interested in the present project. Both have strong expertise in combinatorics, complexity theory and software development. There is a natural progression of tasks to be addressed by the RA and also a natural division of labour between the NZ and Oxford groups. Mike Steel has developed many of the techniques needed for combining local pedigrees into global pedigrees and estimating the necessary number of segregating sites needed to reconstruct global pedigrees (III and IV above). The Oxford group has a larger software development experience (V) and research activity relating to the ancestral recombination graph (II). Both groups are equally interested in the enumeration problems (I). These considerations lead to the following work plan included in appendix A.

The expenses associated with two years hire of the postdoc is 2 years salary. The postdoc will need a computer, e.g. a top end laptop with docking station (2500£), to carry out the proposed research. Software, e.g. Microsoft Office, and essential books, e.g. Statistical Inference from Genetic Data on Pedigrees by E.A. Thompson and Gene Genealogies, Variation and Evolution by J. Hein et al. will incur costs of 300£ per year. During the eighteen months the postdoc is spending in the UK he will be expected to go to one national conference, e.g. MASAMB (750£), and one international conference, e.g. RECOMB (1500£), and to publish in leading journals, e.g. Bioinformatics, that have (excess) page charges. A strength of this project is the collaboration between the groups of Prof. Hein and Prof. Steel and this induces some additional expenses, in particular three return trips between UK and New Zealand (total cost

6000£). Two of these trips will be for Mike Steel and Jotun Hein, which will further incur hotel expenses of an estimated 2200£.

#### **Relevant publications of applicants:**

- Chen, T., Lyngsø, R. and Hein, J.J. (2005) Enumeration of Pedigrees. Report available at <http://www.stats.ox.ac.uk/mathgen/bioinformatics/projects/>
- Erdős, P.L., Steel, M.A., Székely, L.A. and Warnow, T. (1999) A few logs suffice to build (almost) all trees (Part I). *Random Structures and Algorithms* **14**(2): 153–184.
- Erdős, P.L., Steel, M.A., Székely, L.A. and Warnow, T. A few logs suffice to build (almost) all trees (Part II). *Theoretical Computer Science* **221**: 77–118.
- Hein, J.J. (1993) A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination. *J.Mol.Evol.* **20**: 402–411.
- Hein, J.J. (2004) Human evolution: Pedigrees for all humanity. *Nature* **431**: 518–519
- Hein, J.J., Schierup, M.H. and Wiuf, C.H. (2005) *Gene Genealogies, Variation and Evolution*. Oxford University Press, 296 pages.
- Semple, C. and Steel, M (2003) *Phylogenetics*. Oxford University Press, 254 pages.
- Steel, M.A. and Székely, L.A. (1999) Inverting random functions. *Annals of Combinatorics* **3**: 103–113.
- Steel, M.A. and Székely, L.A. (2002) Inverting Random Functions II: Explicit Bounds for Discrete Maximum Likelihood Estimation, with Applications, *SIAM J. Discr. Math.* **15**(4): 562–575.
- Steel, M.A. and Hein, J.J. (2005) Reconstructing Pedigrees: A combinatorial perspective. *Journal of Theoretical Biology*, accepted.
- Wiuf, C and Hein, J.J. (1997) On the Number of Ancestors to a DNA Sequence. *Genetics* **147**: 1459–1468.
- Wiuf, C and Hein, J.J. (1999) The Coalescent with Recombination as a point process moving along sequences. *Theoretical Population Biology* **55**: 248–259.
- Wiuf, C. and Hein, J.J. (1999) A contribution to the discussion of J. Chang's paper "Recent common ancestors of all present-day individuals". *Adv. Appl. Prob.* **31**(4): 1029–1030.

#### **Other relevant publications of others:**

- Chang, J.T. (1999) Recent common ancestors of all present-day individuals. *Adv. Appl. Prob.* **31**(4): 1002–1026.
- Derrida, B. and Jung-Muller, B. (1999) The Genealogical Tree of a Chromosome. *J. Stat. Phys.* **94**(3): 277–298.
- Derrida, B., Manrubia, C.M. and Zanette, D.H. (1999) Statistical Properties of Genealogical Trees. *Phys.Rev.Lett.* **82**(9): 1987–1990.
- Derrida, B., Manrubia, C.M. and Zanette, D.H. (2000) Distribution of repetitions of ancestors in genealogical trees. *Physica A* **281**: 1–16.
- Derrida, B., Manrubia, C.M. and Zanette, D.H. (2000) On the Genealogy of a Population of Biparental Individuals. *J. Theor. Biol.* **203**(3): 303–315.
- Donnelly, K.P. (1983) The Probability that Related Individuals Share Some Section Genome Identical by Descent. *Theoretical Population Biology* **23**: 34–63
- Elston, R.C. and Stewart, J. (1971) A general model for the genetic analysis of pedigree data. *Human Heredity* **21**: 523–543.
- Helgason, A. et al. (2003) A population-wide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: Evidence for a faster evolutionary rate of mtDNA lineages than Y-chromosomes. *American Journal Human Genetics* **72**(6): 1370–1388.
- Kammerle, K. (1989) Looking Forward and Backwards in a Bisexual Moran Model. *J. Appl. Prob.* **27**: 880–85.
- Kammerle, K. (1991) The Extinction Probability of Descendants in Bisexual Models of Fixed Population Size. *J. Appl. Prob.* **28**: 489–502.
- Lander, E.S. and Green, P. (1987) Construction of multilocus genetic linkage maps in humans. *PNAS* **84**: 2363–2367
- Moehle, M. (1994) Forward and Backward Processes in Bisexual Models with Fixed Population Sizes. *J. Appl. Prob.* **31**: 309–322.
- Rhode, D.L.T., Olson, S. and Chang, J.T. (2004) Modelling the recent common ancestry of all living humans". *Nature* **431**: 562–566.
- Thatte, B. (2006) *Combinatorics of pedigrees*. <http://arxiv.org/abs/math/0609264>.
- Thomas, A. and Cannings, C. (2003) Enumeration and simulation of marriage node graphs on zero loop pedigrees. *Math. Med. Biol.* **20**: 261–275.
- Thomas, A. and Cannings, C. (2004) Simulating realistic zero loop pedigrees using a bipartite Prüfer code and graphical modelling. *Math. Med. Biol.* **21**: 335–345.

Thompson, E.A. (2001) *Statistical Inference from Genetic Data on Pedigrees*. Inst. of Math. Stat., 169 pages.

## Appendix A Milestones and work plan

