

Professor Jotun Hein – Case for Support

Combinatorics, Complexity and Probabilistics of the Ancestral Recombination Graph

Background

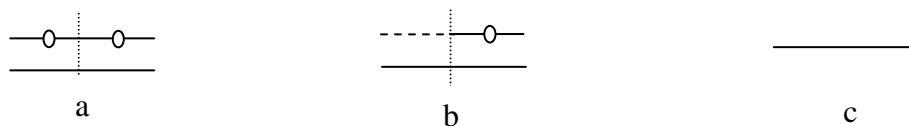
Central to biology is homology and genealogical relationships. Due to the phenomenal growth in sequence data from different species, phylogenetics have risen to prominence and have been put on much firmer statistical ground (Felsenstein, 2004, Semple and Steel, 2002). Similarly intra-population variation has also been catalogued on an unprecedented scale and also led to better characterization of the genealogical relationship of sequence sampled from a population (Hein, Schierup and Wiuf, 2004). The concepts of *phylogeny* and *pedigree* are well known concepts all through the scientific community (even to the general public), and are very old – hundreds and thousands years, respectively. Pedigree could together with *individual*, *life* and *species* be the oldest biological concepts. The genealogical concept describing the relationship of homologous sequences undergoing recombination – the *ancestral recombination graph* (ARG) – is about 2 decades old (Griffiths, 1981 and Hudson, 1983).

The ancestral recombination graph is of tremendous importance and central in population analysis and fine scale mapping of disease genes (Morris et al, 2002). The ARG is central in population analysis since it describes the relationship of sequences sampled from a population and is central in calculating the likelihood function for observed data given a population model. This is necessary for parameter estimation (for instance recombination, mutational and migration rates) and hypothesis testing. Since the likelihood function is calculated via a step summing over all possible genealogical histories, it is possible to pose questions concerning the probability of events in the history of the sequence, like the number of recombinations, the age of specific mutations or time to common ancestors. This is called *ancestral analysis*.

The importance of the ARG has exploded within the last 1-2 years due to the post human genome projects devoted to mapping sequence variation within a population – most notably the HapMap project. The large scale availability of such data, has created the need for methodologies based on the ARG. It is clear that the growth in data easily outgrows the capabilities of methods with the predictable result that data are not analyzed in an optimal fashion.

The ARG and probability of different histories.

The 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. For convenience it will assumed that we know which state is the original. Going backwards sequences can 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 where they carry ancestral material to the data sequences (reducing the sample size by one) and recombination 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. A very simple non trivial data set is shown below.



The ball is a mutant position and absence of a ball represents the original state. We will assume that all recombination events are located exactly between segregating sites, which is illustrated by the dotted vertical line in b. When tracing the history of such a data set back in time, we will eventually get to as single sequence where all the positions of the segregating sites have the original state, not the mutant as shown in c above. One configuration that could be visited from the data set back to a single ancestral sequence is illustrated in b. Counting all possible such configurations to the data set would find 30 configurations. In analogy with recursions for the probability for sequences evolved in absence of recombination found by Griffiths, Ethier and Tavaré in a series of articles from 1987-1995, recursions can be written that calculates the probability of all paths from a data set up to any configuration. A slight complication here is that the ARG can contain cycles, when recombination is allowed in contrast to the classic case where the AC graph is a *directed acyclic graph* (DAG). This is a consequence of

recombination and coalescent events (not involving ancestral material at the same position on two sequences) just redistribute ancestral material and mutations on different sequences, can be reversed. Such a class is characterized by the content in term of mutants and non-mutants at each column of the segregating sites. If the history of the sequences is restricted to a single position at all the sequences, then there can be no recombination and the relationship of the sequences can again be described by a phylogeny – called the *local tree*.

The probability of the data (for instance Figure a above) can be found by following recursion, that sums the probability of all paths from the data to an ancestral configuration, AC:

$$P_{AC} = \sum_{AC'} q_{AC,AC'} P_{AC'} , \quad P_{UA} = 1 \quad (1)$$

the $q_{AC,AC'}$ is the probability that evolution took exactly that that step converting AC to AC' and thus one more step in the path that eventually led to the observed data and this probability is straightforward to calculate. UA abbreviates Universal Ancestors as illustrated in c in the figure. $P_{Data Set}$ can be calculated through this recursion for a fixed set of parameters. The hard problem is summation over of all possible paths or due to the recursion above, over all possible ancestral states.

Our Ancestral Recombination Graph related research.

I had the privilege to do a postdoc in 1985-7 in the group most active and pioneering in this kind of research (Richard Hudson, Norman Kaplan, Martin Kreitman and Charles Langley, NIEHS, NC). I worked on sequence comparison in this period and it was only later that I was able to do more focused research on this topic with Carsten Wiuf, Mikkel Schierup and Yun Song.

Yun Song and I have worked on finding the histories of a set of sequences that need the least number of recombinations. Finding bounds on the number of recombinations needed for a set of sequences has been pursued since 1985 with a focus on the data published by Martin Kreitman, which consisted of 11 sequences with 43 segregating sites. Hudson and Kaplan (1985) proved that at least 5 recombinations were needed in the history of this data set and also that the real number of recombinations could be much higher, so many recombinations are not reconstructable from the data. Myers and Griffiths (2002) proved at least 6 were necessary and by a different method Song and Hein (2003) proved that at least 7 were necessary. Later, the same year Song and Hein (2004) proved that an actual history could be constructed with 7 recombinations. We have enumerated the number of strings of local trees yielding a minimal solution, but not enumerated the number of ARGs, which would entail describing how all the local trees are integrated into an ARG. And we have not enumerated ACs that could be visited in a minimal history. These are non-statistical approaches, but in collaboration with Carsten Wiuf, we have also investigated properties of the coalescent process when including recombination. Wiuf and Hein (1997,1999a) described the number of ancestors to a single sequence and the time back to the time, when all positions along sequences have found a common ancestor. Wiuf and Hein (1999b) proposed a way to characterize this process that did not go from the present towards the past, but started at the leftmost position of the sequences and scanned them towards the right. Recombination would here be an operator on phylogenies instead of on sequences.

The challenge of the Ancestral Recombination Graph.

An ultimate investigation of the properties of the ARG should incorporate complex models of population structure and mutation processes, To make progress it is essential to limit investigations to the basic population model and simple mutation processes. The simplification in the population structure would mean: discrete generations, constant population size, no geographical structure or selection. Relaxing these assumptions are easy. In the mutation process, the infinite sites model where a mutation always is in a new position is highly attractive and will be assumed, but precludes the direct application of to viral data, where the ARG is also a central genealogical structure. The infinite sites assumption is accomplished by representing a sequence with a segment of the real line. However, the introduction of the real line implies that recombination could occur at any position along such a sequence. This gives the problem a continuous component and introduces integration in the key recursions. This is complicating and we will use what could be called the mid-point heuristic: Recombinations are assumed to occur at the mid-points between segregating sites. As a consequence of this all questions relating to the ARG will now only involve discrete sets and can be addressed by combinatorial methods.

Griffiths and Tavaré (1994) introduced an MCMC method that could calculate the probability of data by sampling over possible histories creating the data. Griffiths and Marjoram (1996) applied a stochastic integration technique to evaluate the likelihood function for a data set. However, this method was very slow. Subsequent attempts (Fearnhead and Donnelly, 2001) have accelerated this some, but very substantial improvements are still needed if these methods are to handle data sets that presently exist.

Understanding the size of the data is important in setting realistic goals. The size is described by sequence length (**L**), number of columns with variation (**K**) and by number of sampled sequences (**N**). What are the prospects for likelihood methods as function of N and L and how large data sets do we need to analyze. The relevant parameter here is not L directly, but a rescaling (called ρ) of this, that can be rephrased as the expected number recombination event in two sequences until the expected coalescent event for two sequences. For humans, $\rho=1$ roughly corresponds to 5.0 kb. Θ is defined as the rate of the mutation process, when time is measured in units corresponding to the expected time for two sequences to find a common ancestor in the coalescent process.

Proposed Research

The overall motivation for this research and the research of most other theoretical groups working on the analysis of population variation data that involves recombination is to find good approximations and accelerations to the original Griffiths-Marjoram approach, since this is central in calculating the probability of a set of sequences sampled from a population. There is a staggering growth in this kind of data and computational methods cannot keep up at present. These methods are necessary in understanding population dynamics, human evolution and important in an increasing number of fine scale mapping of disease genes analyses. Our methodological approaches will focus on combinatorics and complexity theory. We will analyse small cases in great detail, where cases can either be exhaustively enumerated or bounds be obtained on how much error is made by confining our interest to a subgraph of the full ARG. Dependent on our success we will move towards larger data sets.

Reductions in the number of ACs summed over in recursion (1) without discarding significant probability mass is an attractive path forward. The main idea is to assume that recombination events only occur in the midpoint between segregating positions. Hopefully, this is an innocent assumption, but transforms the problem into a problem only involving a finite number of ancestral configurations (ACs), since the decision on where recombinations have occurred has been fixed. Recursions can be written involving this finite number of ACs. The number of ACs grows very quickly as a function of number of sequences and segregating positions. Finding a set of ACs that carries the bulk of the probability mass would be major step forward. A series of obvious reduced sets have been investigated and the one that holds the most promise are ACs that can be found as nodes in a minimal or close-to-minimal ARG. However, it is doubtful that any computational trick can allow us to handle large data sets. Nevertheless, exact treatment of smaller data sets/segments could be useful building blocks for approximate methods.

Combinatorics

- The number of AC equivalence classes can easily be described. Two ACs belong to the same equivalence class if one can be converted to the other by a series of recombination and coalescent events and visa versa. The number is $\prod_j \{k_j(n - k_j + 1) + 1\}$ where k_j is the number of mutants in the j 'th column. It is easily verified by remembering that mutants can coalesce with mutants until a single mutant is present that will be converted to a non-mutant by a mutation. Non-mutants can coalesce with non-mutants and there must always be at least one non-mutant in the column of a segregating site as it is traced back in time.
- What is the number of ACs for a given data set and what is the number when the content of each column is fixed? Both these numbers can be enumerated by recursions that scan the sequences and leads to very large numbers.
- What is the growth in the largest number of ACs as a function of any data set as K and L grow? This will depend on the distribution of mutants and non-mutants and the number is largest when mutants is $\lfloor N/2 \rfloor$. The number of ACs grows incredibly fast. For 7 sequences and 5 segregating sites there are more than 10^{17} ACs! So for real size data, this number will grow to extraordinary size and asymptotic methods would have be used to obtain lower and upper bounds on these numbers.

- Number of ACs under constraints. A very natural approach to finding an important subset of AC of smaller cardinality would be to constrain it. An obvious ideas to focus ACs that could be visited on a history of the sequence using a minimal number of recombinations (AC_{min} s). Enumeration of the number of AC_{min} as a function of ρ , Θ and N . Obviously, when $\rho=0$ then we are in the classical case. As a function of ρ , how much of the probability is found in histories of the data that only visits ACs in histories using k recombinations (AC_k)? How much probability is found in histories needing k recombinations as a function of k ? It is straightforward to realized that any AC can be visited using less than $2N*(K-1)$.
- Finding a directed acyclic graph that is equivalent to the ARG. This can be done switching from AC as states to all possible non-decomposable cycles that can be created from the ACs for each equivalence class. This allows the distributions of the times spent in each cycle to be calculated and additionally all probabilities can be calculated without solving linear equation systems.
- Internal isomorphisms in the ARG: The ARG has a large component of self-similarity since strings of identical events occur in different contexts. To exemplify, two sequences with ancestral material at their left and right ends respectively could coalesce, then recombine etc. The ancestral configurations involved could involve other sequences not involved in these events and would thus be in different parts of the ARG. The probability of the events of this example would depend on the complete configuration, although it would be the same events. Being able to find all such isomorphisms would be essential in describing which events dominate the history of the data set.

Complexity

- Complexity Theory from computer science can make statements about the computational time and space needed to calculate certain combinatorial quantities or to approximate certain quantities. Such questions have not yet been asked relative to the ARG. The complexity of finding the minimal history and finding the probability of a data set, i.e. finding the minimal path and the sum over all paths in the Ancestral Recombination Graph respectively are central. We guess that the first problem is NP-Complete, while nothing is known about the second question. Complexity results on the second problem would be most intriguing as this would prove generally that finding the probability of a data set is fundamentally hard and there is no efficient algorithm to be discovered.

Probability Issues

- For how large data sets can the equation (1) above be analyzed fully? How large a set of AC's must be included if all histories carrying at least $1-\epsilon$ of the full probability. How can it be verified that it contains at least $1-\epsilon$ of the probability?
- How does the probability of the sum of paths through a subgraph defined by certain classes of AC grow toward the total probability as the number of ACs grow? In the case before (7 sequences, 5 segregating sites), it is clearly serious if a good subset of ACs of size less than 10^6 within the 10^{17} cannot be found.
- What is the distribution of contribution of different cycles to the probability of the data?
- What is the total length of paths and time spent in individual equivalence classes? How much does this vary conditioned on entrance and exit chosen for these classes?

Use of proposed research

- **Evaluation of existing sampling schemes.** The importance sampling procedures of Fearnhead-Donnelly and Marjoram-Griffiths and the MCMC method of Kuhner-Yamoto-Felsenstein, what are their advantages and failings? An in-depth understanding of this could lead to proposals of more efficient methods.
- **Fine-scale mapping.** A factor in the rise of coalescent thinking has been the possibility of their use in fine scale mapping (Larribe et al., 2002, Morris et al., 2002 and others). Again these methods are based on important approximations. Often a "simplified ARG" is used. It could be very illuminating to investigate these approximations in cases that could be analyzed exhaustively.
- **Aid to pseudolikelihood methods.** Although the focus of this proposal is to start with cases much smaller than presently existing data sets, our research could be of use. Many of the methods that researchers are forced to use, define a pseudolikelihood function, that use likelihood functions of what can be calculated, such as the conditional likelihood of pairs of segregating sites (Hudson, 2000). This gives the pseudo-likelihood function a shape dominated by long-distance information. The methods in

this proposal would be very strong in local information, because that is what we propose to solve exactly.

- **Visualisation.** Although the methods proposed by early research (1996-2003) are of high quality and pioneering, it is difficult to get intuition about what is “going on”. It could be of great help for the general community, if tools were developed that could visualize these and other techniques. We have already been involved in developing one such tool – www.coalescent.dk - but this should be extended.

Future Projects.

There are natural follow-ups to the research described in this proposal and we will pursue this by collaborations or definition of DPhil projects. Even given the most optimistic fulfilment of this project would leave several good projects for further work. *Gene conversion* is relevant for the history for small regions of the genome and thus for fine scale mapping. The issues arising will be similar, but have not been explored in this context (Wiuf and Hein, 2000). *Recombination* is also relevant for viral evolution, but the infinite sites assumption does not hold in the presence of such high mutation rates. Removing the infinite sites assumption is bound to make the problem computationally harder. If we have very large segments, then the underlying *assumption of low sample size* will be violated, since recombination alone will create many segments that have to be traced. A large number of ancestral sequences will make the rate of multiple and simultaneous coalescent events important and would explicitly have to be taken into account, which is not done presently. The approximations proposed in this proposal all uses that ρ is *small* (close to zero). If ρ is infinite the problem also becomes easier, since all genealogies now are independent. It could be of interest if it was possible to find approximations for ρ large (close to infinity). I.e. is it possible to take a series of independent genealogies and add “a little dependence”?

Timeliness and relation to other research on the ARG.

Genomic sequence data from a population are becoming available at an increasing rate and such data are related by ARGs. This has created the need for improved methods and understanding of the ARG.

There is research directed towards the ARG headed by groups of Donnelly, Griffiths and McVean in Oxford, but also other groups in the world headed by researchers such as R. Hudson, D. Balding, J. Felsenstein and S. Tavaré. These groups have mainly focussed on the application of statistical sampling methods such as MCMC and Importance Sampling and then attempting the analysis of large data sets. Our approach will focus on smaller data sets and attempt a thorough combinatorial understanding of the ARG in this context. Additionally, we will also address purely computational issues such as complexity and algorithmic questions. This has not been done at all so far, which is surprising, given the important contributions from computer science to phylogenetics and sequence analysis.

This research will be conducted at the Oxford Centre of Gene Function in the Bioinformatics group headed by Jotun Hein. There has been long term involvement in ARG related problems in this group (Yun Song, Rune Lyngsoe) and Jotun Hein’s previous group (www.birc.dk), so it is an ideal setting for this research. We are part of the statistics department, that also host the groups of Prof. Peter Donnelly, Prof. Bob Griffiths and Dr. Gillean McVean, that have expertises relevant for the project and with whom we have much interaction.

Justification of resources

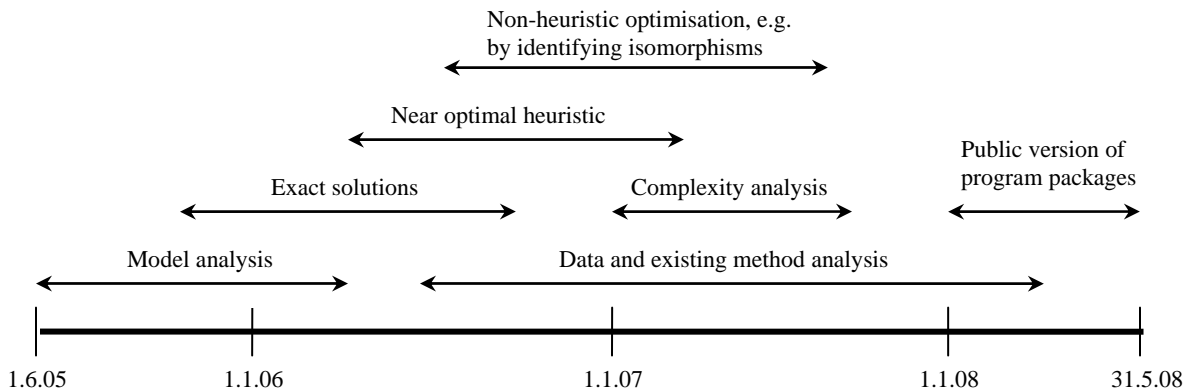
The bulk of the resources is allocated in the postdoc that will be in Oxford Centre for Gene Function in the Bioinformatics group directed by Jotun Hein. Since it is researchers doing computational biology and bioinformatics is in strong demand and hard to attract, we have as for the possibility to give very good salary. Since the research proposed will involve much collaboration and is based on previous work by Jotun Hein and co-workers, we have asked for travel to visit Yun Song and to allow Yun Song to visit Oxford. Jotun Hein has continued much collaboration with his earlier group (www.birc.dk - especially Carsten Wiuf, Mikkel Schierup, Thomas Mailund, but also several others). The collaboration will in this case be focused on Mikkel Schierup and his graduate student (possibly Bo Eskerod), that will come for 6 months 1.7.06-31.12.06 to work with the postdoc. Lastly, Rune Lyngsoe (with Thomas Mailund, BiRC) is presently implementing pilot versions of some ARG algorithms and will most likely not be in Oxford for the full period, but we will continue to collaborate with him due to his complexity theory expertise. The collaboration with a person in California, some people in Denmark creates the need for a series of visits to and from these countries. The post.doc. will need good laptop and desktop in such a computationally intensive project. We will need support in being serviced from the Statistics Department and when we

transport software to other computers, makes it available at a www-page so we have asked for 10% of a computer officer. We expect to submit papers to international conferences and presentations at these will of course incur travel and conference attendance costs.

References

- Ethier, S.N. and Griffiths, R.C. (1987) "The infinite-many-site model as a measure valued diffusion" *Ann. Probab.* 15.2.515-545
- Fearnhead, P. and Donnelly P, (2001) "Estimating recombination rates from population genetic data", *Genetics* 159.3.1299-1318.
- Felsenstein (2003) "Inferring Phylogenies" Sinauer
- Griffiths, R.C. (1981). Neutral two-locus multiple allele models with recombination. *Theor. Popul. Biol.* 19, 169-186
- Griffiths, R.C. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* 46, 131-159.
- Griffiths, R.C. and Tavaré, S. (1995) "Unrooted genealogical tree probabilities in the infinite sites model" *Math. Biosc.* 127.1.77-98
- Griffiths, R.C. and Marjoram, P. (1996). "Ancestral inference from samples of DNA sequences with recombination" *Journal of Computational Biology* 3, 479-502.
- The International HapMap Project, Richard A. Gibbs et al. *Nature* 426, 789 - 796 (18 Dec 2003)
- J.J.Hein: A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination. *J.Mol.Evol.* 20.402-411. 1993
- Hein,J.J., T.Jiang, L.Wang & K.Zhang (1996): "On the complexity of comparing evolutionary trees" *Discrete Applied Mathematics* 71.153-169.
- Hein, Schierup and Wiuf (2004) "Gene Genealogies, Variation and Evolution" Oxford University Press
- Hudson, R.R. (1983) "Properties of the neutral model with intragenic recombination" *Theor.Pop.Biol.* 23.2.213-201.
- Hudson and Kaplan (1985) "Statistical properties of the number of recombination events in the history of a sample of DNA sequences" *Genetics* 111.1.147-64.
- Jobling, Hurler and Tyler-Smith (2004) "Human Evolutionary Genetics" Garland Science
- Kreitman, M. (1983) "Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*" *Nature* 304.412-417.
- Kuhner, Yamoto and Felsenstein (2000) "Maximum likelihood estimation of recombination rates from population data" *Genetics* 156.3.1393-1401
- Larribe, F., Lessard S. and Schork, N.J. (2002) "Gene mapping via the ancestral recombination graph" *Theor. Pop. Biol.* 62.2.215-229.
- Morris, Whittaker and Balding (2002) "Fine-scale mapping of disease loci via shattered coalescent modelling of genealogies" *Am. J. Hum. Genet.* 70.3.686-707.
- Myers, S and Griffiths, R.C. (2002) "Bounds on the minimum number of recombination events in a sample history" *Genetics* 163.1.375-394.
- Ohta and kimura (1971)
- Papadimitriou, C(1992) "Computational Complexity" Addison-Wesley
- Schierup, M. and J.Hein (2000 October): Consequences of Recombination on Traditional Phylogenetic Analysis. (*Genetics* 156.897-91).
- Semple and Steel (2002) *Phylogenetics* OUP
- Song, Y and J.J. Hein (2003) "Parsimonious Reconstruction of Evolution and Haplotype Blocks" (WABI03, Hungary. Lecture Notes in Bioinformatics vol.2812. p287-302)
- Song, Y and J.J. Hein (2004) On the Minimum Number of Recombination Events in the Evolutionary History of DNA Sequences (*J. Math.Biol.* 48.160-86)
- Song and Hein (2004) "Constructing minimal ancestral recombination graphs" (in press *J. Compu.Biol.*)
- Wiuf, C and J.J.Hein (1999): The Ancestry of a Sample of Sequences Subject to Recombination. (*Genetics* 151.1217-1228 March 1999).
- Wiuf, C and J.J.Hein (1999): "The Coalescent with Recombination as a point process moving along sequences" *Theor. Popul. Biol.* (55.248-259).
- Wiuf,C. and J.J.Hein (May 2000): The Coalescent with Gene Conversion (*Genetics* 155.451-462).

Milestones and Work Plan:



During the first three months of the project, the postdoctoral researcher (PR) will analyse the statistical model underpinning the project. The PR cannot be expected to join the project with a comprehensive background in coalescent analysis including recombination. During this initial phase the PR is expected to acquire comprehensive knowledge of the area, and concurrently develop a computational framework for developing an exact solution to the ancestral analysis problem under the model. Jotun Hein (JH), Yun Song (YS) and Rune Lyngsoe (RL) will assist with reading suggestions and discussions.

The first task of the project is to develop a method for exact analysis under the model. This should be at a stage where we can start using it to analyse (small) data sets in July 2006. The implementation of this method will be the main duty of the PR during the first year of the project, with assistance on modelling aspects from YS and JH and on computational aspects from RL.

The main tasks of the second year of the project are to further develop methods to allow larger data sets to be handled, and to apply the software developed to real data. Both heuristic and non-heuristic approaches to allow for larger data sets will be pursued. As part of pursuing non-heuristic optimisations, the inherent computational complexity of analysis under the model will be investigated. The PR, RL, and JH will collaborate on developing these improvements, with the PR responsible for their implementation. This work should be finished by September 2007. By January 2007 we should have software allowing us to analyse practical examples. The first aim is to assess the potential of the methods developed in fine scale mapping. This work will be carried out by the PR, Mikkel Schierup (MS) (possibly Bo Eskerod (BE)), and JH.

Further analyses using the developed methods and software will be the main task of the third year of the project. The aim of this phase is to apply the methods to real data to acquire new biological knowledge, and to apply the methods to real and simulated data to analyse the performance of existing heuristic methods on various types of data. The PR will focus on analysing practical examples, and will, together with JH and YS, use the developed software to analyse the performance of existing methods. It is the duty of the PR, before the end of the project, to make sure that the developed software can easily be distributed and is well documented.

| | |
|--|--|
| Jotun Hein | Directs the project and provides coalescent theory expertise |
| Postdoctoral researcher | Perform combinatorial and probabilistic analysis, implements and tests methods, and assist in analysing practical examples |
| Yun Song | Provides combinatorics and coalescent theory expertise |
| Mikkel Schierup/ Bo Eskerod | Applies developed software to practical examples and assess its performance in conjunction with fine scale mapping |
| Rune Lyngsoe | Provides theoretical computer science and programming expertise |

Table 1: Summation of Division of Labour