

**Professor Jotun Hein**  
**Case for Support BBSRC Grant Application January 2004**  
**Practical Statistical Alignment**

**Background**

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

Although bioinformatics perceived is a new discipline, certain parts have a long history and could be viewed as classical bioinformatics. For example, application of string comparison algorithms to sequence alignment has a history spanning the last three decades, beginning with the pioneering paper by Needleman and Wunsch, 1970. They used dynamic programming to maximize a similarity score based on a cost of insertion-deletions and a score function on matched amino acids.

Independently, Sankoff (Sankoff, 1972) and Sellers (Sellers, 1974) introduced an approach of comparing sequence pairs by minimizing a distance function. Their algorithm is close to identical to the algorithm maximizing similarity.

Sankoff (Sankoff, 1975) generalized the distance minimizing approach to a set of sequences related by a phylogenetic tree. These algorithms have in the following decades received much attention from computer scientists and been accelerated and generalized. Heuristic versions are the core of most of the multiple alignment programs that are being used by researchers. The Clustal family of programs has been used very widely and is based on a series of heuristics. A completely different approach to alignment was introduced in 1994 by Krogh et al. (Krogh et al., 1994), by using Hidden Markov Models (HMMs) to describe a family of homologous proteins. The HMM approach was originally completely free of concepts of phylogeny and evolution.

In 1981 Smith and Waterman, (Smith & Waterman, 1981) introduced a local similarity algorithm, which has remained the gold standard for finding segments within complete sequences that are similar. The main use of this is to search databases and in this context the Smith-Waterman algorithm is too slow and a series of computational accelerations have been proposed, both to the algorithm itself and in the statistical evaluations of observed similarities. The Smith-Waterman algorithm is widely used in the bioinformatics community, for example in the BLAST family (Altschul et al., 1997).

The principle of choosing solutions by minimizing the amount of evolution is also called parsimony and has been widespread in phylogenetic analysis even if there is no alignment problem. Over the last two decades the parsimony method of phylogenetic reconstruction has been severely criticized and has lost terrain to methods based on stochastic modeling of nucleotides, codons or amino acids. The advantages of statistically based methods are many. The weights/costs needed in optimization alignment are replaced by parameters that are readily interpretable and estimation of these parameters naturally gives rise to associated statements about uncertainty. Hypotheses can be tested, and modeling of the underlying evolutionary processes can incorporate increasingly complex features. Such statistical methods and their widespread use represent serious progress and are an indication of the increased statistical awareness of the biological community. The paradoxical situation is that while phylogenetic inference is now fundamentally viewed as a statistical inference problem the corresponding problem of alignment has been very slow in experiencing a similar development. This is despite the fact that users of statistical phylogenetics and similarity/parsimony alignment programs often are the same researches.

This situation is likely to change significantly in the coming years. After a pioneering paper by Bishop and Thompson (Bishop and Thompson, 1986) that introduced and approximated likelihood calculation, Thorne, Kishino and Felsenstein from 1991 proposed a well defined time reversible Markov model for insertion and deletions (denoted more briefly as the TKF91-model), that allowed a proper statistical analysis for two sequences. Such an analysis can be used to provide maximum likelihood (pairwise) sequence alignments, or to estimate the evolutionary distance between two sequences. More recent extensions and improvements

of this model have extended the domain of application considerably. Analyzing more sequences naturally includes phylogeny, alignment and molecular evolution inference. The model can also be used to define a test of homology that does not depend on aligning the sequence. At present, this test corresponds to a test of global similarity as the concept of local alignment that does not yet have an analogue in the statistical alignment framework. In conclusion, although statistical phylogenetic alignment sounds highly specialized it is in reality very general and includes most of the basic concepts behind modeling molecular evolution.

In the last few years this approach to sequence analysis has experienced a renaissance and recent progress has given this methodology the potential for becoming a mainstream tool for empirical researchers. It is statistically better founded than optimization alignment methods (similarity maximization or distance minimization) and allows parameter estimation, hypothesis testing and ancestral analysis not possible by other methods. Recent progress involves better models, the extension of pairwise algorithms to many sequences, faster algorithms and the increased use of MCMC methods.

Despite this optimistic view of these developments, major challenges and work remain to make this approach dominant. The size of data sets will become very large indeed and statistical alignment programs will need to be based on increasingly realistic models. Nevertheless applications to biological structure predictions must be developed to achieve the admiration of the biologist community. To date, statistical alignment has got positive echo from evolutionary biologists for its success in phylogenetic inference. With statistical alignment, alignment-free protein structure and gene structure prediction is possible which are free from artifacts might arise from misalignments. Models for statistical alignment must be extended to be able to model gene structure (Pedersen & Hein, 2003), RNA structure (Knudsen & Hein, 2003) and events that create dependence among positions. Comparative Genomics makes great use of substitution models of evolution that is either dependent on an unobserved gene structure or RNA structure. Inference can then be made about this structure. Some of these projects will need theoretical and algorithmic developments, while much is a question allocating resources to software development.

## **B Past and current work in the subject area both in the US and abroad.**

In the years following the TKF91 paper, there were developments towards statistical approaches to alignments, by for instance Jun Liu and Charles Lawrence (Zhu et al., 1998) and Durbin and Mitchison, but their approaches were not based on explicit evolutionary process. After the first introduction of statistical alignment, there has been very little work for almost a decade until around 2000, when Mike Steel, Jotun Hein, Ian Holmes and a few others have taken interest to the topic (Hein et al. 2000; Steel & Hein, 2001; Hein 2001; Holmes & Bruno 2001). István Miklós wrote his PhD on statistical alignment at Eötvös University, Budapest in 2001 (Miklós & Toroczka, 2001; Miklós 2002, 2003), and since then several postdocs (including IM) around Jotun Hein have made great progress (Lunter et al. 2003a, 2003b, 2003c, Lunter et al., 2004). Bjarne Knudsen (Florida), an earlier student of Jotun Hein, has developed interesting approximation algorithms to complex models based on Hidden Markov Models (Knudsen & Miyamoto, 2003). The group around Arndt von Haeseler (including Metzler, Wakolbinger and others) has also started to be interested in this topic (Metzler et al., 2001; Metzler, 2003).

### **Programme and methodology**

## **C Main aims of the project and individual measurable objectives**

There are a series of criteria for success of the project: package, tutorial, applications to data, methodology testing, publications and generation of activities like recruitment of students and collaborations.

**i. Software package:** A central aim of the project is to develop a package doing statistical alignment for many sequences and performs a series of essential associated analyses: hypothesis testing for molecular clock, phylogeny, mutation rate, selection estimation, assessing ancestral sequences, and structure prediction. It should produce nice graphical outputs of different marginalization (tree, alignment, structure, ancestral sequences, posterior distribution of mutation rates) for easy visualization. Prior parameters will be allowed to be set by users. This package of easy to use, biologist-friendly software should be of wide use in the community. A web based server will be developed that researchers can submit jobs to.

**ii. Development of better biological models and corresponding algorithms:** The TKF91 model allows only single insertions and deletions which is biologically irrelevant. Modeling long insertions and deletions is necessary to make statistical alignment practical. Site heterogeneity and residue-dependent indel rates would also increase the strength of the method. Fast algorithms able to handle such models must be developed, too.

**iii. Structure prediction:** Most of the comparative structure prediction methods are based on alignments, which might be misleading when the alignment is wrong. Much effort has been made on finding good alignments, and many researchers think that the central question in structure prediction is how to find good alignments. Statistical alignment predicts a structure obtained considering all the possible alignments weighted by their posterior distribution. Posterior probabilities can tell how reliable each part of the predicted structure is, an important piece of information that other prediction methods cannot provide.

**iv. Comparative genomics:** At present the capabilities of our statistical alignment methods matches the growth of mammalian genomes. Several questions are of interest here. Since statistical alignment is based on a general model of molecular evolution including both insertions and deletions, issues related to the nature of these events can be addressed and parameterized. What is the nature of insertions and deletions, their length distribution and rates? There is no reason to assume that insertion process is the time reversed process of deletion as assumed by many models. Quantifying the departure from this assumption, would be of great interest. The removal of the intermediate alignment step makes statistical alignment very suited for large scale genomic analysis and question such as departure from the molecular clock and time reversibility of the underlying substitution process can be tested. A good model of neutral molecular evolution also increases the power to detect selection which is central in functional interpretation of individual genes. An especially interesting question at present is the enigma of the amount of selection outside genes.

**v. Methodology testing:** All software will have to be thoroughly tested. Additionally, MCMC algorithms should be tested relative to the exact dynamical programming solutions for low sequence numbers, tests of homology based on statistical alignment should be compared to for instance the local alignment algorithm (Smith-Waterman) for sensitivity and the improved performance of phylogeny estimation should be tested relative to methods that presupposes an alignment.

**vi. Publications:** We have had a good publication record and I am convinced that this project would lead to at least 3-4 good publications per year. Potentially, it could be considerably more if we get into much good empirical collaboration.

**vii. General activities:** Since statistical alignment is a very general technique, it has the potential to generate research for many years to come. We will do our best to define new projects and goals that would both recruit students and engage us in collaborations that would continue beyond the funding period.

## **D Details of methodology of the research**

Methods have been used to address statistical alignment are standard but still challenging. In the period from 1986 to 2001 only 2 sequences could be analyzed by algorithms that were highly reminiscent of the algorithms used to analyze 2 sequences minimizing distance or maximizing similarity (invented 1966-1972). The method is dynamic programming and is based on the multiplicativity of probability of evolutionary events for different nucleotides for statistical alignment. These pairwise algorithms were generalized to an arbitrary number of sequences related by a phylogeny (Hein, 2001; Steel & Hein, 2001; Miklós, 2002; Lunter et al., 2003b; Hein et al., 2003). These algorithms generally have running time as the length of the sequences powered to the number of sequences. For a small number of sequences, these dynamic programming algorithms for exact likelihood calculations are very good, but for more than 4-5 sequences they are too slow and likelihood can be quickly calculated only on augmented data. Markov Chain Monte Carlo (Liu, 2001) is applied to sample from this augmented data. We have already developed several sampling techniques (Holmes & Bruno, 2001; Lunter et al., 2003c; Jensen & Hein, 2004; Lunter et al., 2004) that define different data augmentations. We want to use these sampling techniques in Bayesian frameworks co-estimating alignment, phylogeny and evolutionary parameters, where prior information is given using biological knowledge (like coalescent prior, molecular clock prior, prior information on

mutation rates along the sequences, etc.). Biological questions (like which position is homologous to which one, what is the posterior distribution of tree topologies, what is the posterior distribution of ancestral sequences, what is the expected number of mutations happened, what is the most plausible structure of a given biological sequence, is deviance from the molecular clock significant, etc.) can be answered with marginalization of the sampled data.

Although a rule of thumb is that one should carry out analytical computations as much as possible (coined as Rao-Blackwellisation, Liu, 2001), there is no ‘golden solution’, since we want to apply statistical alignment for several biological problems, and some data augmentation makes several marginalization impossible. For example, a data augmentation developed by Lunter and Miklós does not need data on internal sequences at all, and this makes sampling tree topology very cheap (Lunter et al., 2003c). However, sampling from this data augmentation does not allow inferring ancestral sequences. Another example is the technique developed by Jensen and Hein which yields the best mixing amongst techniques having augmented data at internal nodes of the evolutionary trees (Jensen & Hein, 2004) and thus, it is the best method for inferring ancestral sequences, up to date. On the other hand, Jensen and Hein’s method cannot be used for homology modeling since it sums out all the evolutionary relationships along the branches so one cannot tell which position is related to which.

In summary, at present there are only 2 techniques that can solve these problems – dynamic programming and MCMC – so that is where our focus lies. Over time we will also shift emphasis from the former to the latter. This does not mean dynamic programming will be without value. Dynamic programming provides exact solution and MCMC provides only approximation. The former can be used to test the validity of the later on small data sets.

## **E Timeliness**

The amount of sequence data and completely determined genomes grows and will reach – by previous standards – enormous levels. Statistical alignment is a general method that addresses almost all of the issues involved in sequence analysis: alignment, phylogeny estimation, molecular evolution and homology. Additionally, there is an increased demand for statistically valid methods by the biomedical community. Lastly, a large number of statisticians have turned their attention to the problems arising in comparative genomics and they will also start to work on statistical alignment.

Statistical alignment will become dominant in the coming decade. Our group at present have a lead in this development and funding for this project would allow us to maintain this and contribute seriously to improved analysis of the flood of sequence data.

## **F Milestones and management of the project**

Our project has two main tasks: developing biologically more reliable models and implementing feasible, user-friendly programs. The two tasks run parallel, as described below.

The project will start with implementing a good MCMC sampling method for Bayesian phylogenetic investigation under the TKF91 model. In fact, we already developed an algorithm for fast likelihood calculations, which was published in the WABI’03 proceedings (Lunter et al., 2003c). This calculator is the ‘heart’ of the java code co-sampling alignments and trees. The method should be tested on large scale, on structural alignment databases like HOMSTRAD. These investigations will be carried out in 2004 in the frame of student projects. For this time I. Miklós will visit Oxford to help supervising students and discuss the performance of the proposed sampling method.

Several sampling methods as well as other MCMC tricks like simulated tempering, parallel tempering, population based Monte Carlo methods, etc. (Liu, 2001) might be useful at accelerating computations, and hence make programs more practical. It is practically impossible to predict the performance of these tricks on a particular problem; therefore all of them must be implemented and tested. This needs a large amount of work, which will be one of the postdocs’ duty (PD1). PD1 must have good programming skills, and beyond the previously mentioned work, PD1 will be put in charge of designing the web-based versions of

the developed methods, namely making programs user-friendly.

The reason to test the performance of sampling methods on the relatively simple TKF91 model is that all the computations are about one order faster. We want to finish this by the end of 2004; nevertheless, we want to continue developing better models which are biologically more reliable, for example allow long insertions and deletions, site heterogeneity, etc. We have already made lot of progress in this area (Lunter et al., 2003a; Miklós and Toroczka, 2001; Miklós 2003); however, new models were tested only on pairwise sequence comparisons. We want to extend methods based on these more sophisticated models to be able to analyze many sequences. This needs lots of work, too, for which we would like to apply another postdoc (PD2).

By early 2005, we will have an implementation of sampling methods having best performance on the TKF91 model. In 2005, we would like to use these sampling methods as a core of a user-friendly software package which allows several analyses of biological sequences (molecular clock testing, inferring ancestral sequence, evolutionary parameter estimation, structure prediction, etc.). This will be done by PD1; however, other members of the group will help PD1. Indeed this needs a large amount of work, we should provide graphical tools to visualize several marginalizations easily and allow different priors to use: some of the biologists will be interested in evolutionary trees, others in alignments or evolutionary rates, etc. They might want to introduce prior knowledge in case of comparative structure prediction when the structure of one of the sequences is known.

The other main aim in 2005 will be to study biologically better models and understand how to extend calculations under these models to many sequences, which will be the main duty of PD2 in 2005. To date, we have several ideas (Holmes & Bruno, 2001; Lunter et al., 2003c; Jensen & Hein, 2003; Holmes 2003), but it is not clear now which idea will be the most fruitful. Definitely we will need lots of discussions, for that purpose I. Miklós will visit Oxford twice in 2005. He will also help making programs user-friendly that year.

By the end of 2005, we will have user-friendly software of statistical alignment based on the TKF91 model. Using this software, we can start collaborations with other groups who need statistical investigations of biological sequences hard to be aligned in 2006. This will be the major duty of PD1. By early 2006, we will also have several empirical results on the performance of sampling methods, and we will start developing methods for multiple statistical alignments based on comprehensive models. This needs a tight collaboration of all the group (so this will be the minor duty of PD1, as well as this will be the duty of PD2).

We would like to finish the first implementation of the computer code that makes a full Bayesian statistics under a biologically sophisticated model by the end of 2006. This model will definitely involve long insertions and deletions, if we are able to make enough progress in theory, it will involve rate heterogeneity, as well.

In 2007, we will make the above mentioned code user-friendly, which will be the duty of PD1. PD2 will work on coupling these sequence evolution models with other models, like SCFGs.

We summarize the duties of each participant below (see also Tables I. and II.):

**Post.Doc.1:** will test sampling strategies in 2004, will make user-friendly software for statistical alignment under the TKF91 model in 2005, will engage collaborations with other groups to use the previous program in 2006, will implement software for statistical alignment under a more sophisticated model in 2006, and will make this model user-friendly in 2007. In general, Post.Doc.1 must have good programming skills and also be able to understand biological problems. An ideal applicant would have computer science background and also have experience in bioinformatics; however we do not exclude the possibility to apply a biologist who has enough experience in programming.

**Post.Doc.2:** will work on developing biologically more reliable models in 2004, will extends these models to multiple sequence comparisons in 2005, will implement software based on these models in 2006, will

work on higher level models in 2007. An ideal applicant would have serious mathematical background and also have programming skills.

**Jotun Hein** will supervise the project.

**István Miklós** will be involved in testing sampling methods in 2004, will work on sophisticated evolutionary models in 2004 and 2005, and will help at implementing computer codes from 2004 till 2007. He will also be put in charge to establish collaborations in Hungary (for example Mediago sequencing group, Gödöllő).

**Gerton Lunter** will develop new evolutionary models and related algorithms for fast likelihood calculations, will help at programming from 2004 till 2007.

**Alexei Drummond** will contribute knowledge of MCMC and development of user friendly software programs from 2004 till end of 2006.

### **Justification of resources**

The activity in our group has involved 6 people (estimated % of their time in parenthesis) and has lead to a lot of progress: Jotun Hein (50%), Gerton Lunter (70%), Alexei Drummond (35%), Istvan Miklós (80%), Yun Song (25%) and Ian Holmes (50%) and for Jotun Hein this activity goes back to about 1997. This means that in Oxford alone more than 3 man-equivalents have been working on statistical alignment. We have made several progresses both in theory and practice, and especially the BEAST software package developed by Alexei Drummond and others (Drummond et al., 2002) seems very promising. We already used BEAST interfaced with statistical alignment codes (Lunter et al., 2003c) and the success of it encourages us to do that in the future, as well.

In addition to this, come the collaborations with Aarhus University, Aarhus and Eötvös University, Budapest. In short, progress on statistical alignment demands a solid activity, both in terms of quality and quantity. The algorithms used are a generalization of algorithms used for phylogeny and alignment estimation and uses algorithms that involves techniques from computer science and probability theory. The only advantage this project has over the development of earlier packages (PHYLIP (<http://evolution.genetics.washington.edu/phylip.html>), PAUP (<http://paup.csit.fsu.edu/>), etc.) is that it can efficiently build on earlier experience.

### **Our relevant publications:**

Drummond, A.J., G.K. Nicholls, A.G. Rodrigo & W. Solomon. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161(3): 1307-20.

Hein, J., C. Wiuf, B. Knudsen, Møller, M., and G. Wibling (2000): Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit. *J. Mol. Biol.* 302:265-279.

Hein J.J. (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by Relevant a binary tree. (Pac.Symp.Biocomp. 2001 p179-190. (eds RB Altman et al.)

Hein, J., Jensen, J.L. and Storm, C.S.N. (2003) "Algorithms for Multiple Statistical Alignment" *PNAS* 100(25):14960-14965

Holmes, I. & William J. Bruno (2001) "Evolutionary HMMs: a Bayesian approach to multiple alignment" *Bioinformatics*, 17::803-820.

Holmes, I. (2003) Using Guide Trees to Construct Multiple-Sequence Evolutionary HMMs.

*Bioinformatics*, special issue for ISMB2003, 19:147i-157i.

Jensen, J. and J. Hein (2004) "A Gibbs Sampler for Multiple Statistical Alignment" (*Statistica Sinica*, in press)

Knudsen, B. & Miyamoto, M.M. (2003) Sequence alignments and pair hidden Markov models using evolutionary history. *J Mol Biol.* 333(2):453-60.

Lunter, G. A., Miklós and Holmes (2003a) A 'long indel' model for evolutionary sequence alignment. *Mol. Biol. Evol.*, accepted.

Lunter, G.A., Miklós, I., Song, Y.S. & Hein, J (2003b) An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol.*, 10(6):869-889.

Lunter, G.A., Miklós, I., Drummond, A., Jensen, J. L. & Hein, J. (2003c) Bayesian Phylogenetic Inference under a statistical insertion-deletion model. *Lecture Notes on Bioinformatics, Proceedings of WABI'03* vol.2812. 228-244)

Lunter, G. A., Drummond, A., Miklós, I. & Hein, J. (2004) "Alignment, Statistics and Evolution" (in press , "Statistical Methods for Molecular Evolution" ed. Rasmus Nielsen)

Miklós I. (2002) An improved algorithm for statistical alignment of sequences related by a star tree. *Bul Math. Biol.* 64(4):771-779.

Miklós I. (2003) Algorithm for statistical alignment of sequences derived from a Poisson sequence length distribution. *Disc. Appl. Math.* 127(1):79-84.

Miklós I. & Toroczka, Z. (2001) An improved model for statistical alignment, in: *WABI2001, Lecture Notes in Computer Science*, (O.Gascuel & B.M.E.Moret, eds.) 2149:1-10. Springer, Berlin.

Pedersen, J.S. and J.J. Hein (2003) "Gene finding with a hidden Markov model of genome structure and evolution" *Bioinformatics* 19.2:219-227.

Song, Y. (2003) "Reducing multi-state recursions in hidden Markov models to single-state recursions", submitted to *Mathematical Bioscience* for publication.

Steel, M. & J.J. Hein (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by a star tree. *Appl. Math. Lett.* 14:679-684.

### **Relevant Publications of others.**

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389-3402.

Bishop, M.J. and Thompson, E.A. (1986).Maximum likelihood alignment of DNA sequences, *J. Mol. Biol.* 190:159-165.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 1981;17(6):368-76.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. In H. N. Munro, ed., *Mammalian Protein Metabolism*, pp. 21-132, Academic Press, New York.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., & Haussler, D. (1994a). Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235, 1501-

1531.

Liu, J.S. (2001) Monte Carlo strategies in scientific computing. Springer Series in Statistics.

Metzler, D. (2003) Statistical Alignment based on fragment insertion and deletion models. *Bioinformatics* 19(4):490-499.

Metzler, D., Fleißner, R., Wakolbinger, A., von Haeseler, A. (2001) Assessing variability by joint sampling of alignments and mutation rates. *J. Mol. Evol.* 53(6):660-669.

Sankoff, D. (1972) Matching sequences under deletion/insertion constraints. In *Proc. Nat. Acad. of Sci. U.S.A.* 69:4-6.

Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, 28(35 - 42).

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* 26:787-793.

Smith TF, Waterman MS. (1981) Identification of common molecular subsequences. *J Mol Biol.* 147(1):195-197.

Thorne, J.L., Kishino, H. & Felsenstein, J. (1991) An evolutionary model for Maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33**:114-124.

J.L. Thorne, H. Kishino and J. Felsenstein, Erratum, An evolutionary model for maximum likelihood alignment of DNA sequences, *J. Mol. Evol.* 34, 91-92 (1992).

Zhu, J. Liu, J.S. and Lawrence, C.E. (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics* 14:25-39.

	2004	2005	2006	2007
<b>Post.Doc.1</b>	tests sampling strategies	makes user-friendly software for statistical alignment under the TKF91	Engages collaborations with other groups to use the previous program; implements software for statistical alignment under a more sophisticated model	makes previous model user-friendly
<b>Post.Doc.2</b>	works on developing biologically more reliable models	extends these models to multiple sequence comparisons	implements software based on these models	works on higher level models
<b>Jotun Hein</b>	supervise the project			
<b>István Miklós</b>	contributes to testing sampling methods; helps at implementing computer codes	works on sophisticated evolutionary models; helps at implementing computer codes	helps at implementing computer codes, engage collaborations (especially in Budapest, Hungary)	helps at implementing computer codes
<b>Gerton Lunter</b>	develops new evolutionary models and related algorithms for fast likelihood calculations, helps at programming, especially in 2005.			
<b>Alexei Drummond</b>	contributes knowledge of MCMC and development of user friendly software programs			

**Table I.** Summation of the programme of work by contributors.

	2004	2005	2006	2007
<b>Software writing</b>	Bayesian phylogenetic estimation using the TKF91 model	User friendly statistical alignment package based on the TKF91 model	Bayesian phylogenetic estimation using “long indel” model, probably site heterogeneity	User friendly statistical alignment package based on comprehensive models
<b>Methodology</b>	Algorithms for comprehensive evolutionary models (long indels, site heterogeneity, site dependencies, etc.)	Extension of comprehensive models to many sequences, statistical alignment for gene prediction	Sampling methods using comprehensive evolutionary models	Higher level models (combining statistical alignment with SCFGs and other models)
<b>Empirical investigations</b>	Testing performance of MCMC methods, testing TKF91 on structural database HOMSTRAD	Testing comprehensive models on HOMSTRAD structural database	Gene finding, protein structure prediction, evolutionary hypothesis testing with TKF91	Gene finding, protein structure prediction, evolutionary hypothesis testing with comprehensive models

**Table II.** Summation of the programme of work by main tasks.