

Professor Jotun Hein “Comparative Virus Annotation”

Comparative Genomics has exploded in recent years and put the techniques of molecular evolution at centre stages in harvesting structural information from sequence data. Originally, sequences/genomes were obtained from a single species at a time, but as more genomes have been determined, it has become apparent that the true benefit of this enterprise lies in comparison between genomes both within and between species. However this is only possible when the evolutionary processes that gave rise to the genomes are adequately modelled. Whilst this evolutionary component continues to rise and will become ever more important within mammals and to some degree vertebrates, comparative genomics will spread to all kingdoms of life as genomes become available to compare. Comparative Genomics is key contributor to annotation especially relating to gene structure, RNA structure, regulatory sequences and eventually also more.

There is one taxonomic domain where many genomes are available, but comparative genomics has yet to be applied to its full extent; and that is that of viruses. Full genomes have here been available for decades more than for instance mammals. The main reason for the small role of comparative genomics in virus analysis is threefold: (i) The biological knowledge of these “organisms” is much higher, proportional to the size of genome, than for instance, mammals, initially reducing the apparent need for annotation. (ii) The techniques of comparative genomics have only been developed recently and lastly (iii) the genomic structures are often complex involving overlapping reading frames and superimposed RNA structure constraints. However, due to improved techniques of analysis and the exponential accumulation of genomes, comparative approaches have the potential for a much larger role. The application proposes specific methods that combine methods and simultaneous use of existing biological knowledge to investigate currently sequenced viruses and integrated retroelements for new reading frames, RNA secondary structure and regulatory signals. These methods will also be invaluable for characterizing newly isolated genomes from emerging viral diseases.

The objectives

- To generate a statistical description of genes, RNA motifs and regulatory signals for viral structures and the evolution of such a virus under these constraints. This will include the incorporation of experimental knowledge of the viruses.
- To develop generally available software. This software will take as **input** aligned (possibly partial) viruses together with a description of genes, RNA motif and regulatory signals. **Output** would be additional genes, RNA motif and regulatory signals and possibly question marks to the given biological knowledge, if the knowledge doesn't accord with the analysis. Additionally, key parameters such as selection strengths, phylogenetic tree and rates of evolution, will be estimated.
- Use the methodology and software to analyze a specific series of data sets of an increasingly challenging nature. At present, over 1100 complete virus genomes have been sequenced ranging in size from less than 1000 nucleotides in length to over 350,000 nucleotides for some pox viruses. For some of most medically or economically important of these, numerous complete genomic sequences are available spanning the entire genetic diversity of the particular virus. Concentrating on the RNA viruses which range up to 35,000 nucleotides in length, we will select a number of well-characterised and well-studied viruses for which numerous genome sequences are available from a genetically diverse range of strains. These sets will include human immunodeficiency virus (HIV), simian immunodeficiency virus (SIV), hepatitis C virus (HCV), influenza virus A (IVA).

Combining Gene Structure, RNA Structure and Regulatory Signals Annotation

At present the central goals of annotation are gene structure, RNA structure and regulatory signals. In creating good annotations of viral genomes the challenge will be to create models that can handle the more compact, complex gene structure of virus and combine this with models of RNA structure and regulatory signals.

i. The best comparative **Gene Finding** methods are based on Hidden Markov Models that describe what we would conceive as legal gene structures in terms of intron/exon structure. One of the first and known such gene finder (non-comparative) gene finder was GENSCAN (Burge and Karlin, 1997), but has since been complemented by many others and reliability has much increased by annotating several genomes at the same time. Real genomes have features that were not incorporated in the first gene finders, such as overlapping and interlaced genes. The main complication in the case of viruses is overlapping genes and this feature has been studied by Stephen McCauley in our group. To formulate the appropriate models is simple, with the central issue being that, instead of toggling between 2 states (coding/non-coding), there now is the possibility of toggling between 8 states if all possible reading frames occur in the same direction corresponding to (coding/non-coding) for up to three possible reading frames. Virus annotation for a single virus becomes more difficult since there are more parameters to estimate and on a much smaller data set. For the human genome, there will 3 billion base pairs and about 24.000 genes to train the models. For HIV, there would be 10.000 base pairs and 10 genes to train more complicated models making a single such genome a very poor training set.

ii. The best comparative **RNA structure** prediction methods are based on stochastic context free grammars - SCFGs. These were first introduced for the analysis of single RNA sequences in 1994 by Haussler and colleagues and also by Eddy and Durbin. Hidden Markov Models will scan sequences and generate symbols, where the probability only depends on the last state. SCFGs are a generalization of Hidden Markov Models that allows the simultaneous generation of non-neighborings pairs of symbols in a nested fashion. This technique is quite dominant in the analysis of RNA sequences and genes and is especially powerful, when applied to sets of aligned homologous sequences.

iii. The methods behind **Regulatory Signal** detection are much more heterogeneous (Wasserman, 2004). A successful class of methods are called phylogenetic footprinting and would detect regions of slower evolution (Blanchette, 2003). This field is bound to experience a major growth in coming years and exactly which models will be useful cannot be predicted, but certain general features will be common to RNA/Gene finding: a probabilistic description of classes of signals and a description of how signals evolve. Describing regulatory signals has been done for decades, but is presently very important for genome comparison. In the present framework, a model for the probability of a sequence, when regulatory signals can occur is necessary. If the regulatory signals are not known, then this is close to impossible at present, as there are no universal distinguishing feature analogous to codon usage or base pairing for genes and RNA structures. In this case, comparative approaches are the most promising ways forward. If the regulatory signals are known then there are databases describing these and possible information concerning their interaction with regulatory molecules and a probabilistic description of a signal is often done using a HMM. In this case a probability model can be made by combining this HMM with a description of how it occurs along the sequence.

iv. To extend the original models to allow the simultaneous analysis of a set of closely related virus genomes, **Models of Sequence Evolution** are necessary, where the evolution of a single nucleotide depends on its structural context. Structural context are essentially coding/non-coding, pairing/single and signal position specific evolution in the three classes of models.

- Models of coding versus non-coding regions are very well studied
- Models of single versus base-pairing nucleotides are also well studied.
- Models of positions in regulatory signals are much harder and under much focus presently. Given the position and length of a signal, a natural model is to give each position its own substitution model and then test if a given segment distinguishes itself from the sequence and evolution of background DNA.

These classes of models are all well described, especially so for gene and RNA structure, while regulatory signal evolution is of intense interest at present. Our group has pioneered both comparative gene finding (Petersen and Hein, 2003) and RNA structure prediction (Knudsen and Hein, 1999, 2003).

v. The **Integration of Models** is needed for both description of structure and the structure dependent evolution. Integration of models is needed since a position can be subject to multiple constraints. An easy solution for both descriptions is the assumption of independence. For integrating overlapping reading frames, the assumption would be that a reading frame can occur irrespective of the presence/absence of another reading frame and similarly for any combination of statements involving reading frames, RNA structure and regulatory signals.

For integrating models of evolution combining different constraints this can again be done at different levels of sophistication (Hein and Stoevlbaek, 1995 and Pedersen and Jensen, 2001, Pedersen, Meyer, Forsberg, Simmonds and Hein, 2004a,b). Independence will always be assumed concerning the strength of selection, so for instance the acceleration/deceleration of the evolution of a position is the product of the accelerations/decelerations that the position is subjected to by the two different reading frames. A harder problem is the dependency of a position relative to neighbouring positions. Hein and Støvlbæk (1995) using the “nucleotide independence heuristic” that assumes each nucleotide evolves independently. This simplified calculations considerably. Pedersen and Jensen (2001) formulated a more correct model, that could integrate of over all possible histories of two sequences using a Gibbs sampler. Unfortunately, this method is very slow and harder to implement.

Annotation restricted by given biological knowledge.

It is increasingly important that, when applying bioinformatics techniques, a variety of levels and data types have to be taken into account. Taking a single sequence/genome and annotating it without using available knowledge is highly suboptimal and this is even more the case for virus analysis where the biological “knowledge per nucleotide” is very high in terms of transcripts, base pairing and more. If we were to annotate all existing HIV 1 genomes, the annotations would have to be restricted or even better strongly biased towards annotations that agree with given biological knowledge. This can be done and the resulting annotation would focus on what could be newly inferred by our analysis and which parts of given biological knowledge should be questioned.

Problems ahead: In most evolutionary analysis there are four complicating factors that have to be addressed or data has to be chosen so these factors don't undermine the analysis. These are **alignment, recombination, structure change and local dependencies in substitution process.**

When analyzing very similar sequences, there will be little or no alignment problem, but this will highly constrain the data sets that can be analyzed – especially for fast evolving RNA viruses. As genomes become more diverse the dependence on an uncertain alignment is increasingly problematic. Theoretically there are ways out of this, namely the concept of statistical alignment (a field within which our group is very active (Hein et al.2003, Lunter et al. 2004)).

Recombination is a complicating factor in that the underlying assumption of comparative genomics is that there is a shared phylogeny among all analyzed positions. Again this can in principle be addressed (and again our group is active in this (Hein, 1993, Song and Hein, 2003)), but the reality is that at present, this problem leads to such computational challenges that this should be addressed at a later stage.

Most comparative methods for detecting hidden structure assume that the structure of what is being compared remains unchanged through their shared evolutionary history. This is normally true for very similar genomes, but, again, can be increasingly questioned for more distantly related genomes. If the gene/RNA/regulatory structure is different between different genomes, then it will pose problems in the analysis if not properly addressed. For instance HIV-1 and HIV-2 have slightly different gene sets and comparative gene finding would only work for the genes they have in common.

The last problem of dependency of nucleotide substitution process on its neighbour nucleotides arises automatically, when there are overlapping functional constraint like overlapping reading frames or paring for RNA structures surely? (Pedersen et al, 2004). How to solve it is known, but will make everything computationally slower and programming more difficult.

Test data and biological analysis.

The methodological developments proposed in this application will be tied to the analysis of a series of pre-defined sets of data. The ideal data set will have the following properties:

i. It should have a large and growing number of known genomes, at different degrees of similarity. The larger number of genomes the more evolution will have been observed and correspondingly the

- comparative approach will be more powerful. It is also advantageous to have the possibility of including more distant genomes if they can be properly aligned and the biology of the virus hasn't changed too much.
- ii. The biology of the virus should be well studied so much information about gene structure, RNA structure and regulation is readily available.
 - iii. The virus should be of general interest and still pose challenging analysis problems.

These criteria immediately imply that the chosen data will be from a generally known virus and we have chosen to focus on retroviruses, influenza and SARS. However, it should be emphasized that if the general methodological problems are solved as proposed, it will be easy to venture into new virus families.

Application and collaboration 1: Retroviruses – HIV1 with Prof. Kjems (<http://www.rna.dk/jk/>) Due to the HIV pandemic, HIV and retroviruses are extremely well studied with a huge and growing data set available. The HIV Sequence Database is mainly aligned and annotated according to the open reading frames and contains very limited RNA structural annotation and almost no RNA structural alignment (Korber et al., 2002). However, many new RNA structure-functional elements are on a regular basis being identified throughout the HIV-1 genome e.g. splicing regulatory elements in coding regions (Tange et al., 2001) and a new intermolecular dimerization sites (Andersen et al., 2004). The complex biology of HIV is contained in a relative short genome of approximately 10.000 nucleotides (Coffin, 1997), and HIV thus represents a challenging case of structural and functional overlap at the level of DNA, RNA and protein. The many overlapping elements require an extremely careful annotation of the *cis*- and *trans* acting elements of the viral replication cycle to account for all the evolutionary pressures exerted on the sequence. Application of the RNA structure prediction methods presented above have been used to study the untranslated parts of the HIV genome and combined with biochemical data (Knudsen et al., 2004, Damgaard et al., 2004). Given the recent development of algorithms for prediction of RNA structures in coding regions (Pedersen, 2004a, b) the analysis will be extended to RNA structure prediction in the coding regions that comprise 95% of the HIV-1 genome. Some HIV sequence regions are known to contain RNA structures on top of both one, two and even three reading frames. In the analysis of full-length HIV genomes recombination is a common feature that will be annotated for the given sequence set and incorporated in the analysis. We will start with groups of similar (40-80% matched nucleotides identical) sequences (only HIV-1 for instance) annotated using the proposed methods and then expand to include more distantly related sequences from lentiviruses towards other retroviral groups.

Application and collaboration 2: Endogenous Retroviruses: with Prof. Finn Skou Pedersen (<http://www.mbio.au.dk/~fsp/>). The most exciting aspect of effective comparative virus annotation would obviously be to find new reading frames. One of the reasons virus annotation has not caught on earlier is that biological knowledge is so rich for these small genomes and thus it is unlikely that any reading frames will have been overlooked. However, there is one important case where this fails: ancient integrated retroviruses in human (and other) genomes (Hughes et al. (2004), Belshaw et al. (2004)). These provide interesting cases of ancient viruses that are much further back in time than what could be reconstructed from extant viruses and they are very difficult to study biologically. One annotation source is comparison to the known present virus proteins, but that has limitations, since such distant viruses are likely to contain new reading frames.

The human genome harbour large amounts of repetitive DNA related to retroviruses (Bannert and Kurth, 2004, Boeke and Stoye, 1997, Lower et al., 1996). These so-called human endogenous retroviruses (HERVs) most likely represent provirus insertions into the germ-line of our ancestors 10-100 Mio years ago. This „fossils record” presents a tremendous amount of retroviral diversity that extends far beyond extant retroviral species (Herniou et al., 1998). Recent research at BiRC (www.birc.dk) has resulted in identification of ~8000 such HERV sequences (Villesen et al 2004). These data provide access to a specialized and previously poorly annotated part of the human genome that most likely harbours regulatory *cis*-acting elements or encodes *trans*-acting factors that are engaged in (viral) RNA processing (as known from exogenous retroviruses like HIV). Examples include the small Rec protein from the HERV-K group that partially overlap with the envelope gene (Yang et al., 1999). As a good starting point for genome annotation analysis of HERV sequences we have delineated the open reading frame structure for the major retroviral genes (Villesen et al., 2004). One important caveat of examining endogenous retroviral genomes is that lack of selection for functionality and they are often degenerate. Hence, in order to deduce functional active regions or genes one has to reconstruct the ancestral sequence. This may be solved by the presence

of multiple related copies (10-100) of each HERV group in the human genome (parelogs) as well as the comparative sequence analysis of the chimpanzee orthologs. Importantly, a model trained on thousands of retroviral genomes may likely out-compete existing similarity-based methods for increased possibility in detection of endogenous viral elements.

Application and collaboration 3. Hepatitis C virus: (HCV) with Peter Simmonds (<http://www.virology.ed.ac.uk/research/Simmonds.htm>) is a leading cause of chronic liver disease, infecting more than 170 million people (~2.5% of the world population). Persistent infection with HCV leads to irreversible cirrhosis, can cause hepatocellular carcinoma (>100,000 cases/annum), and is the principal indication for liver transplantation in US and European adults (Hoofnagle, 2002). HCV is a member of the family *Flaviviridae*, which includes the arthropod-borne flaviviruses, such as yellow fever virus, and the pestiviruses, such as bovine viral diarrhoea virus. The 9.6kb positive-sense RNA genome of the virus encodes a single open reading frame (ORF) of ~3000 aa., processed co- and post-translationally into at least 10 proteins.

There is a growing realization that despite the apparent simplicity in its genome organization (a feature shared with many other positive stranded RNA viruses), HCV replication is very tightly regulated both at the level of replication and translation functions, but also anatomically within the cell. RNA folding through internal base-pairing play important roles in many steps in the replication of HCV; translation is mediated through an internal ribosomal entry sites containing RNA structures that interact directly with the ribosome (Pestova *et al.*, 1998). It has also been shown that HCV replication is dependent on stem-loop structures embedded in coding sequences at the 3' end of the genome, potentially analogous to the *cis*-replicating elements of picornaviruses (You *et al.*, 2004). We have recently discovered that the genome of HCV also shows evidence for extensive RNA secondary structure formation throughout the genome, a feature shared with other positive-stranded RNA viruses in the *Picornaviridae* and *Caliciviridae* that establish persistent interactions with their host, and which may function as some kind of blocking mechanism of host-cell defences triggered by dsRNA (Simmonds *et al.*, 2004).

For HCV and other viruses in the flavivirus family (including the non-pathogenic GBV-C virus, which is also capable of host persistence, and shares with HCV, extensive RNA secondary structure), there is a clear need to annotate the RNA secondary structures in their genomes to understand more about their replication and interaction with the host cell. This task will be assisted by the large amount of comparative sequence information for viruses within each of the flavivirus genera, and the contrasts in replication mechanism and ability to cause persistence between members of this virus family.

General applications: A general virus annotator: The methodology proposed should be built into a general virus annotator, that could take new viral sequence and possible associated biological knowledge or even better a set of such viral sequences and annotate accordingly. Obviously biological knowledge will still be of prime importance, but such an annotator would both a good tool to immediately obtain a rough guess of the biology of the new virus and a useful tool to handle established knowledge. Make such an annotator starting de Novo would be very hard, but we can build on existing expertise and ongoing projects, which would include simple overlapping gene finder (Stephen McCauley), models of sequence evolution (2 postdocs to be hired, Gerton Lunter (OCGF), Istvan Miklos (Hungary), RNA secondary structure (R. Lyngsoe, OCGF), regulatory signals (Gerton Lunter and Asger Hobolt, Aarhus, Dk), the evolution of gene structure (Saskia DeGroot) and combining selective constraints (Jotun Hein).

Our group has close ties to the evolutionary virology group around Andrew Rambaut, Alexei Drummond, Oliver Pybus and others in Zoology, Oxford. We can draw upon their expertise and collaborations could well emerge.

References

- Andersen, E.S., Contera, S.A., Knudsen, B., Damgaard, C.K., Besenbacher, F. and Kjems, J. (2004) Role of the Trans-activation Response Element in Dimerization of HIV-1 RNA. *J Biol Chem*, 279, 22243-22249.
- Bannert N, Kurth R. (2004) Retroelements and the human genome: New perspectives on an old relation. *Proc Natl Acad Sci U S A*. 2004
- Belshaw, R., V.Pereira, A.Katzourakis, G. Talbot. J.Pates, A.Burt and M.Tristem (2004) "Long-term reinfection of the human genome by endogenous retroviruses" *PNAS* . vol.101.14.4894-99
- Blanchette, M. (2003) "A comparative analysis method for detecting binding sites coding regions" In Miller, W, Vingron, M., Istrail, S, Pevzner, P. and Waterman, M. (eds) *Recomb-03*) p57-66.

- Boeke, J. D., and Stoye, J. P. (1997). Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In "Retroviruses" (J.M. Coffin, D. C. Hughes, and H. E. Varmus, Eds.), pp. 343-436. Cold Spring Harbor Laboratory Press, New York.
- Burge and Karlin (1997) "Prediction of Complete gene structures in human genomic DNA" *J.Mol.Bio.*, 268.1.78-94.
- Coffin, J.M., Hughes, S.H. and Varmus, H.E. (eds.) (1997) *Retroviruses*. Cold Spring Harbor Laboratory Press, New York.
- Damgaard, C.K., Andersen, E.S., Knudsen, B., Gorodkin, J. and Kjems, J. (2004) RNA interactions in the 5'-region of the HIV-1 genome. *J Mol Biol*, 336, 307-568.
- Eddy and Durbin (1994) "RNA Sequence Analysis Using Covariance Models" *Nucleic Acid Research* 22.2079-88
- Gifford and Tristem (2003) "The Evolution, Distribution and Diversity of Endogenous Retroviruses" *Virus Genes* 26.3.291-315.
- Guan *et al.* (2004) Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature*. 430: 209-213.
- 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. and J. Støvlbæk. (1994). "A Maximum Likelihood Approach to the Evolution of Overlapping Reading Frames." *J.Mol.Evol.* 40.181-190.
- J.Hein, J.Jensen and C.Storm (2003) "Recursions for Multiple Statistical Alignment" (PNAS 100(25):14960-14965.)
- Herniou, E., Martin, J., Miller, K., Cook, J., Wilkinson, M., and Tristem, M. (1998). Retroviral diversity and distribution in vertebrates. *J Virol* 72(7), 5955-66.
- Hoofnagle, J. H. (2002). Course and outcome of hepatitis C. *Hepatology* 36, S21-S29.
- Lower, R., Lower, J., and Kurth, R. (1996). The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc Natl Acad Sci U S A* 93(11), 5177-84.
- Knudsen, B. and J.J.Hein (1999) "Using stochastic context free grammars and molecular evolution to predict RNA secondary structure (Bioinformatics vol 15.5 15.6.446-454
- Knudsen,B. and J. Hein Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* 2003 Jul 1; 31(13): 3423-8.
- Knudsen, B., Andersen, E.S., Damgaard, C., Kjems, J. and Gorodkin, J. (2004) Evolutionary rate variation and RNA secondary structure prediction. *Comput Biol Chem*, 28, 219-226.
- Kuiken CL, Foley B, Freed E, Hahn B, Korber B, Marx PA, McCutchan F, Mellors JW, and Wolinsky S, Eds. HIV Sequence Compendium 2002. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 03-3564.
- Lunter, Miklos, Drummond, & Hein (2004) "Alignment, Statistics and Evolution" (in "Statistical Methods in Molecular Evolution" ed. Rasmus Nielsen, in Press)
- M.Mar Albà, M., Lee, D., Pearl, F.M.G., Shepherd, A.J., Martin, N., Orenge, C.A. and Kellam, P. VIDA: a virus database system for the organisation of virus genome open reading frames. *Nucleic Acids Research* 2001; 29(1), 133-136
- Pedersen, JS, IM Meyer, R Forsberg, P. Simmonds and JJ Hein (2004) "A comparative method for predicting and folding RNA structures within protein coding regions" (in press) *Molecular Biology & Evolution*
- Pedersen, J.S. and J.J. Hein (2003) "Gene finding with as hidden Markov model of genome structure and evolution" *Bioinformatics* 19.2.219-227.
- Pedersen, JS, IM Meyer, R Forsberg, P. Simmonds and JJ Hein (2004) "A stochastic model for the evolution of nucleotides subject to selective constraints from coding and RNA folding criteria" submitted to *Nucleic Acid Research*
- Pestova, T. V., Shatsky, I. N., Fletcher, S. P., Jackson, R. T. & Hellen, C. U. T. (1998). A prokaryotic-like mode of cytoplasmic eukaryotic ribosome binding to the initiation codon during internal translation initiation of hepatitis C and classical swine fever virus RNAs. *Gene Develop* 12, 67-83.
- Petersen, A.-M. K. and J.L.Jensen (2001) "A Dependent-Rates Model and an MCMC-Based Methodology for the Maximum-Likelihood Analysis of Sequences with Overlapping Reading Frames." *Mol.Biol.Evol.* 18.5.763-776.
- Sakakibara, Brown, Hughey, Mian, Sjolander, Underwood and Haussler (1994) "Using Stochastic Context Free Grammars for tRNA Analysis" *Nucleic Acid Research* 22.5112-20
- Simmonds, P., Tuplin, A. & Evans, D. J. (2004). Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *RNA* 10, 1337-1351.
- Song, Y and J.J. Hein (2003) "Parsimonious Reconstruction of Evolution and Haplotype Blocks" (WABI03, Hungary. Lecture Notes in Bioinformatics vol.2812. p287-302)
- Tange, T.O., Damgaard, C.K., Guth, S., Valcarcel, J. and Kjems, J. (2001) The hnRNP A1 protein regulates HIV-1 tat splicing via a novel intron silencer element. *Embo J*, 20, 5748-5758.
- Villesen, P., Aagaard, L., Wiuf, C. and Pedersen FS. (2004) The coding potential of human endogenous retroviruses. Submitted
- Yang, J., Bogerd, H. P., Peng, S., Wiegand, H., Truant, R., and Cullen, B. R. (1999). An ancient family of human endogenous retroviruses encodes a functional homolog of the HIV-1 Rev protein. *Proc Natl Acad Sci U S A* 96(23), 13404-8.
- You, S., Stump, D. D., Branch, A. D. & Rice, C. M. (2004). A cis-acting replication element in the sequence encoding the NS5B RNA-dependent RNA polymerase is required for hepatitis C virus RNA replication. *J Virol* 78 , 1352-1366.

	Year 1	Year 2	Year 3
Post. Doc.	The articles and programs of related to comparative genefinding/RNA structure prediction, superimposed constraints modelling are studied and integration planned. Implementation of toy versions of final program done.	Program that can simulated viral structures and their evolution is implemented. More advanced analysis program that can read alignments and annotations are implemented. Data analysis is started.	Full scale data analysis and article writing. Polishing of programs are performed and made publically available.
Jotun Hein	The Post.Doc. is directed and experience of other projects are used maximally. Stephen McCauley is presently implementing algorithms to find overlapping reading frames. Saskia DeGroot studies models of RNA and gene structure change. A major program development for statistical alignment is relevant both for algorithms and program development with 2 postdocs here and with collaborations with Istvan Miklos (http://ramet.elte.hu/~miklos/).		
Kjems Group HIV analysis	3 month stay of Ebbe Sloth (PhD from Kjems Group)		
Finn Skou Pedersen Endogenous retrovirus analysis	3 month stay of Anders Kjeldbjerg (PhD from Finn Skou Pedersen Group)		
Peter Simmonds Hepatitis C Analysis	We have ongoing collaborations and we will have a string of smaller continuous visits.		

Table. Summation of the programme of work by contributors.