Ph.D. Dissertation:

Molecular Evolution and Biological Sequence Analysis

Bjarne Knudsen

BiRC - Bioinformatics Research Center
The Faculty of Science
University of Aarhus
Denmark

July 2002

Contents

Preface

This dissertation represents a collection of the scientific research performed by me during my Ph.D. study at The University of Aarhus.

Acknowledgments

I would like to thank my advisor Jotun Hein for putting me in a very stimulating environment of research. His support has been very valuable to me. In the summer of 2001 he moved to Oxford to further pursue his scientific career, but has continued to advise me.

When Jotun Hein moved to Oxford, Freddy Bugge Christiansen became my advisor at the University of Aarhus. I thank him for taking this responsibility.

My most helpful academic connection apart from my advisors has been Michael M. Miyamoto, whom I had the pleasure of visiting at The University of Florida from January 2001 to July 2002 (interrupted by short stays at the University of Aarhus). He has provided wonderful support and enthusiasm for our research, which I greatly acknowledge.

I would also like to thank the other people in the group in Aarhus and in Florida for all the great times we have spent together, both in science and outside. A special thanks goes to the administrative offices, both of The Department of Ecology and Genetics at The University of Aarhus and The Department of Zoology at The University of Florida. Without the help from the people there, I would never have found my way through my Ph.D. study.

Finally a big thank you to all my friends, and especially my family, for their help throughout this period.

Note

The chapters two, four, and five are copies of previously published articles. They have been included in their original form with page numbers from the journal in which they were published. They also have page numbers relative to this dissertation in the middle bottom of each page.

Bjarne Knudsen
July, 2002

Chapter 1

Introduction

The theme of this dissertation is the analysis of biological sequences using computational methods. Many approaches to such analyses can be chosen, most focusing on the biological processes being modeled, but some focus usually remains on a practical mathematical model. There is always a trade off between biological realism and finding a mathematical model under which calculations can be done efficiently (Baldi and Brunak, 1998; Durbin *et al.*, 1998).

It has been my goal throughout my Ph.D. study to use biology as the main guide to the algorithms being developed, while using appropriate mathematical models. The are two recurring themes in most of the work that I have done: molecular evolution and grammatical models.

Evolution is the basis of most biological research, directly or indirectly (Bull and Wichman, 1998). In most of my work I have tried to model molecular evolution in a very explicit way obtaining methods that closely correspond to the evolutionary processes observed in nature.

The use of grammatical models in molecular biology has become very widespread, particularly in the form of hidden Markov models (HMMs) in the study of protein sequences. More complex grammatical models called stochastic context-free grammars (SCFGs) have become a very useful tool in the study of RNA molecules and their structure (Durbin *et al.*, 1998).

Models of biological phenomena serve a dual purpose. Firstly, they can be tested against each other to find the best model for a certain data set. This choice of model gives information about the underlying biological processes governing the data set in question. Secondly, models can serve as frameworks in which further analyses can be performed. As an example, assume that we have a number of related DNA sequences available. Finding a good nucleotide evolution model can tell us how the DNA has evolved. But the model can also be used to infer a phylogenetic tree of the sequences (Graur and Li, 2000).

This chapter gives a description of methods used in the study of molecular evolution and grammatical models. It is shown how these two areas have provided many useful tools and valuable information by themselves. It is also explained how methods from these two areas can be joined to produce even more powerful methods in the study of biological molecules.

Explicit Models of Molecular Evolution

Many methods studying evolution do not use an explicit evolutionary model, but still implicitly assume that evolution occurs in certain ways. One example is in the alignment of sequences. This is often done by scoring matches and mismatches of residues, while introducing gaps in the alignment at a certain cost (Thompson *et al.*, 1994). An alignment of two sequences is a statement about how an ancestral sequence evolved into the two sequences being aligned. Thus, an alignment method is making assumptions about how evolutionary events occur. In most alignment methods, however, the scoring scheme can not directly be interpreted as a realistic model of evolution, particularly for the insertion and deletion (indel) process (Thorne *et al.*, 1991). This chapter focuses on explicit models for evolution, both for single residues and for whole sequences (as in the alignment problem).

Models of Single Residue Evolution

Jukes and Cantor (1969) introduced a simple explicit model for the evolution of amino acid replacements in proteins, where all types of replacements occur at the same rate over time. Since then more sophisticated amino acid replacement models and nucleotide models have been introduced (Graur and Li, 2000).

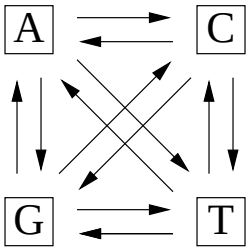
Continuous Time Markov Models

An underlying concept of basic models for individual residue evolution is that the nature of a replacement is only dependent on what residue was present right before the replacement. In particular, the past history of the site in question is irrelevant, given this residue. Such conditional independence of the past is the underlying concept of a group of models referred to as Markov models.

Replacements often happen from generation to generation rather than on a continuous time scale, but generation time is generally very short compared to the evolutionary time scale, so a continuous time model is appropriate. This is another key assumption in the basic evolutionary models that are discussed here.

Lastly, it is often assumed that replacements of a given residue by another given residue occurs at a certain fixed rate, so that a replacement event is equally likely at any point in time (given the starting residue). This is another key element of Markov models. In conclusion, continuous time Markov models are good basic tools to describe the evolutionary process of individual residue replacement (Yang, 1994a).

A continuous time Markov model can be viewed as a directed graph with nodes representing states (in this case residues) and edges representing replacements. Each edge has a rate associated with it, as shown for a nucleotide substitution model here:



Any nucleotide can be replaced by any other nucleotide (all the possible directed edges are included in the graph). Not shown here are the rates associated with each edge.

Calculations in Continuous Time Markov Models

The probability distribution of the residue at a site can be represented by an n dimensional vector p, with entries representing different residues. For amino acid evolution, $n = 20$, while $n = 4$ for nucleotide evolution. The equilibrium distribution is represented by another n dimensional vector $\pi = \{\pi_i\}$. The rate of replacement from residue i to j is denoted r_{ij}, yielding an $n \times n$ rate matrix, $R = \{r_{ij}\}$. Usually this type of evolution is assumed to be time reversible, i.e., the flux of amino acids from type i to j is equal to the flux in the reverse direction:

$$\pi_i r_{ij} = \pi_j r_{ji}, \qquad \text{for } i \neq j.$$

The matrix entry r_{ii} represents the negative total rate by which residue i is replaced:

$$r_{ii} = -\sum_{j \neq i} r_{ij}.$$

This ensures that the relationship between time, t, the probability distribution of residues, and the rate matrix can be expressed as a simple differential equation:

$$\frac{dp(t)}{dt} = p(t)R.$$

This gives the following expression for p as a function of time:

$$p(t) = p(0)e^{-Rt},$$

where e^{-Rt} represents the exponentiation of the matrix $-Rt$, the result of which is a new $n \times n$ matrix. The exponentiation can be carried out mathematically, e.g., by diagonalization of R. Let e_i denote the i'th unit vector. If $p(0) = e_i$, the j'th entry of $p(t)$ represents the probability that residue i becomes residue j in time t. This means that e^{-Rt} is the probability matrix where entry ij

represents the probability of observing residue j after time t, given that we start with residue i (Yang, 1994a).

The rate matrix and the time is represented as a product in the above equation for $p(t)$. This means that the absolute rate of evolution can only be found if the absolute time between two sequences is known, and vice versa. In other words time and rate are confounded, which is the reason evolutionary distances are often measured in units of expected replacements per position, rather than in years. To ensure that time is measured in these units, the overall rate of replacements should be one:

$$\sum_{i \neq j} \pi_i r_{ij} = 1$$

$$\Updownarrow$$

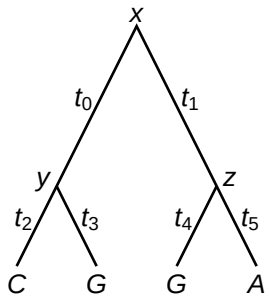
$$\sum_{i} \pi_i r_{ii} = -1.$$

This constraint is often imposed on evolutionary models by scaling the rate matrix by an appropriate factor.

Calculations on Phylogenetic Trees

A typical situation when studying biological sequences is that we obtain a number of sequences from the present, e.g., from different species. It is often assumed that the sequences are related by a tree structure, e.g., via a phylogenetic tree of the species involved. We only know the residues at the leaves of the tree (representing the present), but we can still use the above described Markov model to do calculations on the tree.

To find the probability of observing a given amino acid configuration of a site, given a phylogenetic tree, we can use the above differential equation along the branches from the bottom of the tree, moving upwards to the root. Let us look at an example:



We know the four nucleotides at the bottom of the tree, but not the ones at the internal nodes (x, y, and z). If we assume that y is given, we can calculate the probability of observing the C and G in the first two sequences:

$$P(CG|y) = P(C|y, t = t_2)P(G|y, t = t_3).$$

The terms on the right are entries in the exponentiated matrices as described above. We can now sum over y to obtain an unconditional probability of observing C and G in the first two sequences:

$$P(CG) = \sum_y P(CG|y)\pi_y = \sum_y P(C|y, t = t_2)P(G|y, t = t_3)\pi_y.$$

Such summations can be performed up through the tree to obtain the probability of observing the entire site pattern, given the tree. This method was described by Felsenstein (1981) and is called post order tree traversal.

Now that we are able to calculate the probability of a given site, we can easily extend the method to entire sequences by assuming that sites evolve independently. This allows us to find the maximum likelihood tree, given a number of sequences.

Jukes Cantor Model

The simplest model for the evolution of residues is the model by Jukes and Cantor (1969), which states that all replacements occur at the same rate. Using this model for DNA, the expected number of observed replacements for a site is:

$$E[d] = \frac{3}{4}(1 - e^{-\frac{4}{3}(\mu t)}),$$

where t is the time between the two observations, d is the number of observed replacements, and μ is the replacement rate. Using this we can infer the product of the replacement rate and the time between two sequences. We see that these quantities are confounded as described above.

Nucleotide evolution

Since there are only four standard nucleotides, the nucleotide rate matrix is only a 4×4 matrix compared to the 20×20 matrix of proteins. Furthermore, nucleotides can be split into two groups according to their chemical properties: two pyrimidines (C and T) and two purines (A and G). For these reasons, many simple parameterizations of the rate matrices become possible. The rate matrix for the Jukes Cantor model can be written as:

$$\begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \left[\begin{array}{cccc} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{array}\right] \end{array} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$

Replacements within the chemical groups defined above are called transitions, while replacements between them are called transversions. Kimura (1980) made

a model, where the rates of transitions are allowed to be different from the rates for transversions:

$$\begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \left[\begin{array}{cccc} -2\alpha-\beta & \alpha & \beta & \alpha \\ \alpha & -2\alpha-\beta & \alpha & \beta \\ \beta & \alpha & -2\alpha-\beta & \alpha \\ \alpha & \beta & \alpha & -2\alpha-\beta \end{array}\right] \begin{array}{c} A \\ C \\ G \\ T \end{array} \end{array}$$

This model has two parameters, α and β, and is thus called Kimura's two parameter model. Often the parameters are chosen so that $2\alpha + \beta = 1$, which normalizes the rate matrix to give an overall rate of one. When this is done, time will be measured in expected replacements per position.

The complexity of nucleotide models increase with the HKY model by Hasegawa *et al.* (1985), which allows for different transition and transversion rates at the same time as allowing for any frequency distribution of nucleotides. The most general model for nucleotide evolution is called the general reversible model, where the only constraint is that the evolution is time reversible (Rodriguez *et al.*, 1990). There are numerous nucleotide substitution models with varying degrees of complexity (Felsenstein, 1981; Kimura, 1981; Tamura and Nei, 1993; Zharkikh, 1994).

Typically the parameters of the evolutionary model are maximum likelihood estimated, often at the same time as the phylogenetic tree is estimated.

Protein Evolution

It is not easy to come up with a good parameterization of protein rate matrices, since there are twenty different kinds, which cannot be clearly split up into groups of similar chemical types. For this reason, protein evolution models are usually based on a database of sequences from which a fixed empirical model is built.

The first model to be made in this way was the Dayhoff model (Dayhoff *et al.*, 1978), which was made from closely related proteins. This model can be extrapolated to any protein distance by expressing it as a rate matrix as described above.

Evolutionary rate matrices are often used in database searches for distant homologues, thus matrices for this purpose should be made from distantly related sequences. This was done in the Blosum series of matrices inferred by Henikoff and Henikoff (1992), which were made from conserved blocks of protein sequences.

Another popular protein rate matrix was estimated by Jones *et al.* (1992), who used the Swiss-Prot database of protein sequences (Bairoch and Apweiler, 2000). Most databases are dominated by soluble proteins giving rate matrices estimated from them a bias. Jones *et al.* (1994) derived a rate matrix for membrane spanning segments of proteins, which turned out to be very different from the general matrices produced earlier.

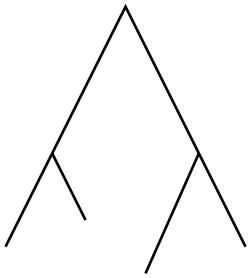
Some efforts to make parameterized models for proteins have been carried out (Yang *et al.*, 1998). These models are based on the underlying DNA codons and a classification of amino acids into groups according to their chemical properties and physical size.

Additional Models for Evolution

The above described models for single residue evolution can be extended and modified in numerous ways. In the following, a few of these are discussed.

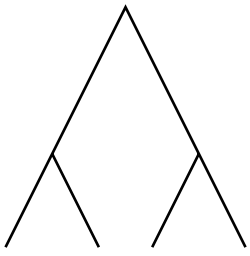
Molecular Clock Hypothesis

When estimating branch lengths for an evolutionary tree, contemporary sequences may have different distances to the root:



This can for example happen when evolutionary rates change over time and between species. As described above, time and rates are confounded, so even though the time is the same from the root to the leaves, the rates can influence the distance measured in expected numbers of replacements and they may not be the same for all sequences.

The molecular clock hypothesis states that the rate of evolution is constant in different species and throughout time (Zuckerkandl and Pauling, 1962). If this assumption holds true, all present sequences will have the same evolutionary distance from the root:



Under the maximum likelihood estimation of the branch lengths, this constraint can be added, enabling a likelihood ratio test for the molecular clock hypothesis to be performed.

If a phylogeny is built from sequences known to adhere to the molecular clock hypothesis, it represents a unique opportunity for dating ancient speciation events. If the time of one such event in the phylogeny is known, the absolute rate can be estimated and the time of all the other events in the tree can be inferred. This approach has proven useful in the dating of human divergence from the great apes (Hasegawa *et al.*, 1985).

The molecular clock hypothesis can also be used for sequences sampled at different points in time. When these time points are known, the absolute rate of evolution can again be estimated (Rambaut, 2000). This has been used in the study of viruses that evolve quickly and where we have older samples available to study (Forsberg *et al.*, 2001).

Rate Variation Between Sites

In biological sequences, there is often a variation in the evolutionary rate from site to site. Some sites are very conserved (e.g., functional residues in enzymes) while others are not very conserved (e.g., surface residues with little function). This variation in evolutionary rates can be incorporated in evolutionary models (Uzzell and Corbin, 1971). This is often done by assuming that the rate variation follows a gamma distribution with mean one and a parameter, $\alpha > 0$ (Yang, 1994b):

$$\phi(x) = \frac{\alpha^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\alpha x}.$$

This is used as a prior distribution of the rates, with a low α resulting in a wide distribution of rates, while a high value of α results in a narrow rate distribution. Rather than estimating the rate at every site, the rate is integrated out, which means that the probability of a site is:

$$P(\text{site}) = \int_{r=0}^{\infty} \phi(r) P(\text{site}|\text{rate} = r)\, dr$$

When inferring a tree for multiple sequences, the α parameter can be maximum likelihood estimated. It has been shown that neglecting to take rate variation into account can have numerous effects on phylogeny estimation, including underestimates of the branch lengths and wrong topologies (Yang, 1996).

Dependence Between Residues

In some situations residues will not evolve in a way that is close to independent, for example of in the evolution of RNA coding genes. In RNA, the secondary structure is often conserved to a higher degree than the individual nucleotides, which means that the nucleotide evolution is under the constraint of having to be consistent with the secondary structure. In this case paired nucleotides can be viewed as residues, yielding 16 different possibilities (Muse, 1995; Schöniger and von Haeseler, 1994). Interestingly double mutations seem to happen at a positive rate (Tillier and Collins, 1998).

Another type of dependence between residues, that has been explored is the co-evolution of amino acids that are close in the three dimensional structure of proteins (Pollock *et al.*, 1999). This has been proposed as a tool aiding in the three dimensional structure prediction of proteins.

Rate Variation over Time

In the work by Fitch and Markowitz (1970) a model of evolution, called the covarion model was introduced. It included an underlying state of a site as being either variable or invariable. The underlying state can change over time, resulting in a rate change from zero to a positive rate or vice versa.

Rates of evolution are an indication of the functional importance of a site. A low evolutionary rate indicates that a site is involved in a function that is important to the protein in question (Kimura, 1983; Graur and Li, 2000). Such functional importance of sites may change over time, for example as a result of a gene duplication or a major speciation event (Gaucher *et al.*, 2001; Wang and Gu, 2001).

Tuffley and Steel (1998) developed a method where sites can change between being variable and invariable at a certain rate. In this model the rate change is treated analogously to the replacement of residues in a Markov model. This was further developed by Galtier (2001) to include numerous different rates, rather than just variable or invariable.

Since there often is a specific point in a phylogeny in which functional divergence is expected, the possible rate shift can be fixed to that specific point (e.g., duplications or major speciation events). This change in evolutionary rate can be modeled by letting predefined parts of a phylogenetic tree have different rates. Then it can be tested whether such rates are significantly different for a given site. Such tests were developed by Knudsen and Miyamoto (2001) (Chapter 5) and shown to be effective in detecting the functional divergence of two related protein families. These methods have been further extended by Knudsen *et al.* (2002) (Chapter 6). In these approaches the rates did not have to be either zero (invariable) or another fixed positive value (variable), rather, rates can attain any non-negative value.

There are other methods of detecting rate variations, e.g., based on Bayesian approaches (Gu, 1999, 2001). Here the rates are assumed to be gamma distributed in the two sub-families and a posterior probability for the rates being independent is developed. The results are generally similar to the results by statistical tests (Knudsen and Miyamoto, 2001, Chapter 5).

Whole Sequence Evolution

Most models of molecular evolution are based on one or two residues (or three in the case of codon models). It is possible to view evolution of whole sequences using a continuous time Markov model. In this view, insertions and deletions become events just like replacements of residues. Exact algorithms have been

developed for indel lengths of one (Thorne *et al.*, 1991; Hein *et al.*, 2000, 2002, Chapter 4)

Evolutionary models for entire sequences allow for the maximum likelihood estimation of new evolutionary parameters, like the insertion and deletion rates. This means that all the evolutionary parameters can be estimated from the data, rather than being specified in advance. This makes statistical alignment much more objective than the predominant alignment methods, where for example gap costs have to be specified by the user (e.g., Thompson *et al.*, 1994).

Another advantage to statistical alignment is that it provides a framework in which statistical tests can be formulated. A very important example is whether two sequences are homologous. This question forms the basis of database searches for biological sequences. Shuffle tests have traditionally been a popular test for homology between sequences (Doolittle, 1986), but is problematic, since it has little theoretical foundation. In the statistical alignment framework, a direct statistical test for homology was developed by Hein *et al.* (2000) (Chapter 4) by testing the hypothesis that two sequences are infinitely far apart in time.

For many databases, fast approximate methods are necessary, because of the massive amounts of data they represent. Therefore fast homology detection methods like BLAST (Altschul *et al.*, 1997) will probably always be preferred for most database searches.

Grammatical Models of Sequences

Grammatical models were introduced by Chomsky (1956) to study the complexity of languages. Later, such models were adopted by the speech recognition community, where they have been extensively used (Baker, 1979; Rabiner and Juang, 1986). Around 1987 their use was introduced to biology (Lander and Green, 1987).

Stochastic grammatical models have turned out to be very useful in the modeling of biological sequences (Durbin *et al.*, 1998). The simplest such model is the stochastic regular grammar, which corresponds to HMMs. HMMs are especially used for proteins, where they for example serve to model individual families (Bateman *et al.*, 2002). More advanced grammatical models called SCFGs (stochastic context-free grammars), have been very valuable in RNA analysis.

A Grammar

A grammar is a method for generating strings of symbols. They work by starting with one symbol, S, and then rewriting it to different symbols according to a set of rules. When no more rewritings are possible, the process stops and we have generated a string. Here is a simple set of rules:

$$S \to aA \mid bS \mid b \qquad A \to aS \mid a$$

This means that S can be rewritten to either aA, bS, or b. A can be rewritten to either aS or a. An example:

$$S \rightarrow bS \rightarrow baA \rightarrow baaS \rightarrow baab$$

The final string has neither A nor S in it, so no more rewritings can be performed. The multiple options for rewriting A and S are what makes it possible to generate many different strings. In fact all strings of a's and b's where there is always an equal number of a's next to each other can be formed. The symbols that can be rewritten are called non-terminals (upper case letters), while the rest of the symbols are terminal symbols (lower case letters).

A Stochastic Grammar

Few biological sequences can be described well by the above type of grammar because biological sequences often do not follow strict rules. A good solution to that problem is to use stochastic grammars, which have probabilities associated with the productions. Thus, the above grammar could be extended to a stochastic grammar:

$$S \rightarrow aA\ (0.8) \mid bS\ (0.1) \mid b\ (0.1) \qquad A \rightarrow aS\ (0.9) \mid a\ (0.1)$$

The numbers in parentheses are probabilities for the different rewriting rules. Now the grammar has a high probability of generating sequences rich in a's.

In sequence analysis, grammars are used to interpret the already existing sequences, rather than to generate sequences. Specifically, it is assumed that the sequence under analysis was generated by a grammar and then a number of questions can be answered: How likely is it that this sequence was generated by a given grammar, how was it most likely generated, and so on.

Regular Grammars

A grammar where all the rules are of the form $N \rightarrow t \mid tM$ is a regular grammar, as in the above example. Analyses of sequences under such a model are computationally quite fast.

HMMs is a slightly different group of models, that consist of a number of states, which typically generate symbols according to probability distributions. The generation of a symbol is followed by a possible state change, again according to a probability distribution.

The difference between stochastic regular grammars and HMMs is that stochastic regular grammars generate symbols at the same time as the non-terminal changes. HMMs generally alternate between forming symbols and changing states. Stochastic regular grammars and HMMs can be used to form the same models, so they are equivalent (Rabiner and Juang, 1986). In biology the term used is generally HMM.

The type of HMMs used for modeling protein families is called profile HMMs (Durbin *et al.*, 1998). They have a start state, a number of match states which

are progressively used to form the amino acids of a typical protein in the family. Insertions and deletions relative to the typical protein are handled by special insertion and deletion states. The process stops with an end state.

Context-Free Grammars

Context-free grammars are grammars, where the rules are of the form $N \rightarrow \alpha$, where α is any string of terminals and non-terminals. The reason for the name 'context-free' is that the rules for transforming N are independent of what surrounds N in the string being rewritten. Any context-free grammar can be put in Chomsky normal form, where all rules are of the form $N \rightarrow M_1 M_2 \mid t$ (Martin, 1991).

Here is an example of a context-free grammar:

$$S \rightarrow aSu \mid uSa \mid cSg \mid gSc \mid gSu \mid uSg$$
$$S \rightarrow aS \mid cS \mid gS \mid uS$$
$$S \rightarrow a \mid c \mid g \mid u.$$

This grammar allows for the generation of multiple non-terminals at the same time. Here the result is a simple grammar that generates RNA sequences with secondary structure in the form of paired nucleotides. This is the basis of the popularity of context-free grammars in biology. SCFGs were introduced in biology as a tool for RNA studies by Eddy and Durbin (1994) and Sakakibara *et al.* (1994).

Beyond Context-Free Grammars

There is a four level nested hierarchy of grammars as described by Chomsky (1959). Regular grammars are the simplest followed by context-free grammars. The next two levels are context-sensitive grammars and unrestricted grammars, respectively. Even context-sensitive grammars are generally too complex to be useful in biology, with a few exceptions. A subclass of context-sensitive grammars, called tree grammars, have been used to model protein structure (Mamitsuka and Abe, 1994). RNA pseudoknots can be modeled with a so called rearrangement grammar, which is also a subclass of context-sensitive grammars (Rivas and Eddy, 2000)

Combining Grammars with Evolutionary Models

While evolutionary models and grammars have proven useful in their own respects, combining these methods give even more powerful tools. Using such an approach, we are able to include information from numerous related sequences, to give improved understanding and prediction of the features of biological sequences.

Protein Secondary Structure

Thorne *et al.* (1996) described a general protein secondary structure HMM with three states: alpha helix, beta sheet, and loops. Instead of just specifying the amino acid frequencies in the three states, they found a full evolutionary rate matrix for each state. This allowed them to construct a HMM that generated amino acids for a number of sequences related by a tree, rather than single sequences. The amino acid configuration at a site, given its state (i.e., secondary structure), is determined from the rate matrix for that state.

Using this HMM it was possible to obtain secondary structure predictions for alignments of protein sequences. It was assumed that the secondary structure was the same for all sequences in the alignment. Predictions improved with more sequences added to the analyses and treating sequences as independent rather than related via a tree worsened the results.

RNA Secondary Structure

An RNA secondary structure prediction method analogous to the protein method by Thorne *et al.* (1996), was developed by Knudsen and Hein (1999) (Chapter 2). Rather than using HMMs, the RNA method uses SCFGs. The evolutionary model for unpaired nucleotides was based on a 4×4 rate matrix, while pairs were modeled using a 16×16 matrix of all possible pairs. This resulted in a very useful method for RNA structure prediction, that is now available through a web server (Knudsen and Hein, 2002, Chapter 3)

It has turned out that this RNA structure prediction method is very useful for researchers working on RNA: In the mRNA of the SCL loci, some hairpins were found by this method (Göttgens *et al.*, 2002) and new structural elements have also been detected in HIV (Damgaard *et al.*, 2002). In the HIV study, variations in evolutionary rates relative to the model of Knudsen and Hein (1999) was a concern, but it was shown by Knudsen *et al.* (2002) that the method is fairly robust towards this effect.

Through the years, many automated methods for RNA structure prediction have been developed. One of the earliest was the method by Nussinov *et al.* (1978), which was based on a very simple scoring scheme for RNA structures. This was soon followed by an energy based method by Zuker and Stiegler (1981). This method has since been developed further and has become the very popular MFOLD algorithm (Zuker, 1989; Zuker *et al.*, 1999; Mathews *et al.*, 1999). All these methods are based on predictions from single sequences and are unable to take into account information from numerous related sequences. Many of the methods that does take numerous sequences into account, fail to deal well with the evolutionary aspect of the sequences in question and with prior expectations to the structure (e.g., Eddy and Durbin, 1994; Tabaska *et al.*, 1998)

Conclusion

The work presented in this dissertation represents advances in a number of areas of bioinformatics, including RNA secondary structure prediction, functional analyses of proteins, and statistical alignment. The unifying theme is molecular evolution, which is not only interesting by itself, but has also been shown to be a great tool for understanding biological molecules.

There are many future directions for this work. The RNA model could be extended to include more sophisticated features, like realistic distributions of stem and loop lengths. A challenging extension would include base stacking interactions, which would require a more complex evolutionary model.

The main problem with the statistical alignment method presented is that it does not include indels of length more than one. This could possibly be remedied through the use of an approximate algorithm for a model with geometric indel lengths. An exact algorithm would most likely be very time consuming and thus impractical. Another important advance in statistical alignment would be a faster approximate way to align multiple sequences according to an explicit evolutionary model.

Evolution and molecular models should be integrated even further to include statistical alignment procedures that take structural models into account. Again time consuming calculations would be a major problem, so fast approximate algorithms would be key to making such procedures useful.

References

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

Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28** (1), 45–48.

Baker, J. K. (1979) Trainable grammars for speech recognition. In Klatt, D. H. and Wolf, J. J. (eds), *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America* pp. 547–550.

Baldi, P. and Brunak, S. (1998) *Bioinformatics: The Machine Learning Approach.* Cambridge, Massachusetts: The MIT Press.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. and Sonnhammer, E. L. (2002) The pfam protein families database. *Nucleic Acids Res.* **30** (1), 276–280.

Bull, J. and Wichman, H. (1998) A revolution in evolution. *Science,* **281** (5385), 1959.

Chomsky, N. (1956) Three models for the description of language. *IRE Transactions on Information Theory,* **2** (3), 113–124.

Chomsky, N. (1959) On certain formal properties of grammars. *Information Control,* **2** (2), 137–167.

Damgaard, C., Andersen, E. S., Knudsen, B., Gorodkin, J. and Kjems, J. (2002). Biochemical and phylogenetic evidence for a higher order structure in the 5'-end of the HIV-1 genome. In preparation.

Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978) A model of evolutionary change in proteins, matrices for detecting distant relationships. In Dayhoff, M. O. (ed), *Atlas of Protein Sequence and Structure* volume 5 pp. 345–352, Washington, D. C.: Cambridge University Press.

Doolittle, R. F. (1986) *Of URFs and ORFs: a Primer on How to Analyze Derived Amino Acid Sequences.* California: University Science Books.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge: Cambridge University Press.

Eddy, S. R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22** (11), 2079–2088.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17** (6), 368–376.

Fitch, W. M. and Markowitz, E. (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4** (5), 579–593.

Forsberg, R., Oleksiewicz, M. B., Petersen, A. M., Hein, J., Botner, A. and Storgaard, T. (2001) A molecular clock dates the common ancestor of european-type porcine reproductive and respiratory syndrome virus at more than 10 years before the emergence of disease. *Virology,* **289** (2), 174–179.

Galtier, N. (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.* **18** (5), 866–873.

Gaucher, E. A., Miyamoto, M. M. and Benner, S. A. (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proc. Natl. Acad. Sci. USA,* **98** (2), 548–552.

Göttgens, B., Barton, L., Chapman, M., Sinclair, A., Knudsen, B., Grafham, D., Gilbert, J., Rogers, J., Bentley, D. and Green, A. (2002) Transcriptional regulation of the stem cell leukaemia gene - comparative analysis of five vertebrate SCL loci. *Genome Res.* **12** (5), 749–759.

Graur, D. and Li, W.-H. (2000) *Fundamentals of Molecular Evolution.* Sunderland, Massachusetts: Sinauer Associates, second edition.

Gu, X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* **16** (12), 1664–1674.

Gu, X. (2001) Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* **18** (4), 453–464.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22** (2), 160–174.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22** (2), 160–174.

Hein, J., Jensen, J. L. and Pedersen, C. N. S. (2002). Recursions for statistical multiple alignment. In preparation.

Hein, J., Wiuf, C., Knudsen, B., Møller, M. and Wibling, G. (2000) Statistical alignment: Computational properties, homology testing and goodnesss-of-fit. *J. Mol. Biol.* **302** (1), 265–279.

Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA,* **89** (22), 10915–10919.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8** (3), 275–282.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1994) A mutation data matrix for transmembrane proteins. *FEBS Lett.* **339** (3), 269–275.

Jukes, T. H. and Cantor, C. R. (1969) Evolution of protein molecules. In Munro, H. N. (ed), *Mammalian Protein Metabolism* pp. 21–123, NY: Academic Press.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16** (2), 111–120.

Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA,* **78** (1), 454–8.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge, U.K.: Cambridge Univ. Press.

Knudsen, B., Andersen, E. S., Damgaard, C., Kjems, J. and Gorodkin, J. (2002). The effect of evolutionary rate variation on the secondary structure prediction of HIV-1 5'-leader RNA. In preparation.

Knudsen, B. and Hein, J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics,* **15** (6), 446–454.

Knudsen, B. and Hein, J. (2002). Practical RNA secondary structure prediction using stochastic context-free grammars. In preparation.

Knudsen, B. and Miyamoto, M. M. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl. Acad. Sci. USA,* **98** (25), 14512–14517.

Knudsen, B., Miyamoto, M. M., Laipis, P. J. and Silverman, D. N. (2002). Functional studies of proteins combining evolutionary rates and structural information. In preparation.

Lander, E. S. and Green, P. (1987) Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA,* **84** (8), 2363–2367.

Mamitsuka, H. and Abe, N. (1994) Predicting location and structure of beta-sheet regions using stochastic tree grammars. In Altman, R., Brutlag, D., Karp, P., Lathrop, R. and Searls, D. (eds), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* pp. 276–284, California: AAAI Press.

Martin, J. C. (1991) *Introduction to Languages and the Theory of Computation.* New York: McGraw-Hill, Inc.

Mathews, D. H., Sabina, J., Zuker, M. and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improve prediction of RNA secondary structure. *J. Mol. Biol.* **288** (5), 911–940.

Muse, S. V. (1995) Evolutionary analyses of DNA sequences subject to con-

straints on secondary structure. *Genetics,* **139** (3), 1429–1439.

Nussinov, R., Pieczenik, G., Griggs, J. R. and Kleitman, D. J. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math.* **35** (1), 68–82.

Pollock, D. D., Taylor, W. R. and Goldman, N. (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.* **287** (1), 187–198.

Rabiner, L. R. and Juang, B. H. (1986) An introduction to hidden Markov models. *IEEE ASSP Mag.* pp. 4–16.

Rambaut, A. (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics,* **16** (4), 395–399.

Rivas, E. and Eddy, S. R. (2000) The language of RNA: A formal grammar that includes pseudoknots. *Bioinformatics,* **16** (4), 334–340.

Rodriguez, F., Oliver, J. L., Marin, A. and Medina, J. R. (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142** (4), 485–501.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. C. and Haussler, D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.* **22** (23), 5112–5120.

Schöniger, M. and von Haeseler, A. (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.* **3** (3), 240–247.

Tabaska, J. E., Cary, R. B., Gabow, H. N. and Stormo, G. D. (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics,* **14** (8), 691–699.

Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10** (3), 512–526.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22** (22), 4673–4680.

Thorne, J. L., Goldman, N. and Jones, D. T. (1996) Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13** (5), 666–673.

Thorne, J. L., Kishino, H. and Felsenstein, J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33** (2), 114–124.

Tillier, E. R. and Collins, R. A. (1998) High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics,* **148** (4), 1993–2002.

Tuffley, C. and Steel, M. (1998) Modeling the covarion hypthesis of nucleotide substitution. *Mathematical Biosciences,* **147** (1), 63–91.

Uzzell, T. and Corbin, K. W. (1971) Fitting discrete probability distribution to evolutionary events. *Science,* **172** (988), 1089–1096.

Wang, Y. and Gu, X. (2001) Functional divergence in the caspase gene family and altered functional constraints: Statistical analysis and prediction. *Genetics,* **158** (3), 1311–1320.

Yang, Z. (1994a) Estimating the pattern on nucleotide substitution. *J. Mol.*

Evol. **39**, 105–111.

Yang, Z. (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39** (3), 306–314.

Yang, Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11** (9), 367–372.

Yang, Z., Nielsen, R. and Hasegawa, M. (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol,* **15** (12), 1600–1611.

Zharkikh, A. (1994) Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* **39** (3), 315–329.

Zuckerkandl, E. and Pauling, L. (1962) Molecular disease, evolution and genetic heterzygosity. In Kasha, M. and Pullman, B. (eds), *Horizons in Biochemistry* pp. 189–225, Academic Press.

Zuker, M. (1989) Computer prediction of RNA structure. *Methods in Enzymology,* **180**, 262–288.

Zuker, M., Mathews, D. H. and Turner, D. H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In Barciszewski, J. and Clark, B. F. C. (eds), *RNA Biochemistry and Biotechnology* pp. 11–43, Dordrecht, the Netherlands: Kluwer Academic Publishers.

Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res.* **9** (1), 133–148.

Chapter 2

RNA Secondary Structure Prediction Using Stochastic Context-Free Grammars and Evolutionary History

Article by Bjarne Knudsen and Jotun Hein.

Appeared in *Bioinformatics*, **15** (6), 446–454, 1999.

BIOINFORMATICS

Vol. 15 no. 6 1999
Pages 446–454

RNA secondary structure prediction using stochastic context-free grammars and evolutionary history

B. Knudsen and J. Hein

Department of Genetics and Ecology, The Institute of Biological Sciences, University of Aarhus, Building 550, Ny Munkegade, 8000 Aarhus C, Denmark

Received on December 22, 1998; revised and accepted on February 22, 1999

Abstract

Motivation: *Many computerized methods for RNA secondary structure prediction have been developed. Few of these methods, however, employ an evolutionary model, thus relevant information is often left out from the structure determination. This paper introduces a method which incorporates evolutionary history into RNA secondary structure prediction. The method reported here is based on stochastic context-free grammars (SCFGs) to give a prior probability distribution of structures.*

Results: *The phylogenetic tree relating the sequences can be found by maximum likelihood (ML) estimation from the model introduced here. The tree is shown to reveal information about the structure, due to mutation patterns. The inclusion of a prior distribution of RNA structures ensures good structure predictions even for a small number of related sequences. Prediction is carried out using maximum a posteriori estimation (MAP) estimation in a Bayesian approach. For small sequence sets, the method performs very well compared to current automated methods.*

Contact: *bk@imf.au.dk*

Introduction

Computerized methods have been used for RNA secondary structure prediction for a number of years (e.g. Nussinov *et al*., 1978; Zuker and Stiegler, 1981). During the last 10 years, further methods have been developed (e.g. Zuker, 1989; Eddy and Durbin, 1994; Sakakibara *et al*., 1994; Cary and Stormo, 1995; Tabaska *et al*., 1998). Some methods use single sequences, which take advantage of prior information on RNA structures, usually through energy functions, e.g. Zuker (1989). No knowledge concerning related sequences is used, so these methods are not ideal when estimating structures of sequences with known homologs.

Covariance methods (Eddy and Durbin, 1994) and profile stochastic context-free grammars (SCFGs) (Sakakibara *et al*., 1994), on the other hand, do use information from more than one sequence, but do not explicitly take phylogeny into account, and do not use a prior probability distribution of structures. Maximum weighted matching methods (Cary and Stormo, 1995; Tabaska *et al*., 1998) share these characteristics.

The method introduced here uses prior knowledge about RNA structure in making a maximum a posteriori (MAP) estimation of the secondary structure. This is performed on an alignment of sequences assumed to have identical secondary structures, i.e. the alignment is assumed to be a structural alignment. The method takes the phylogenetic tree of the sequences into account, including branch lengths, using a model of mutation processes in RNA. Furthermore, the tree can be estimated by a maximum likelihood (ML) method.

The idea for this work originates in work by Goldman *et al*. (1996), who developed a method for predicting protein secondary structure using hidden Markov models (HMMs) and including phylogenetic information. This method uses 20×20 rate matrices for amino acid replacements. Three matrices are employed: one for α-helices, one for β-sheets and one for coils (the rest). These matrices are estimated from sequences of known structure. An HMM with three states, corresponding to the structure types, models the structures along sequences. This HMM is then used in conjunction with the rate matrices to find the ML estimate of the tree relating sequences in an alignment and to predict their secondary structures. The method described here is an extension of this model to RNA secondary structure.

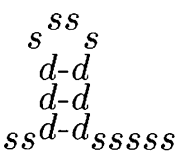
Secondary structures in RNA are not local, like in proteins, thus it is necessary to use a more complex model than an HMM for modelling these. SCFGs, which are used here, can describe some long-range interactions, including most of the ones in RNA secondary structure. SCFGs are unable to model crossing interactions, thus pseudoknots cannot be predicted by this method.

Algorithms

The input for this analysis is an alignment of RNA sequences, while the output is a single common structure for the sequences. The model consists of two distinct parts: the SCFG and the evolutionary model.

a) $S \rightarrow LS \rightarrow LLLLLLLS \rightarrow LLLLLLLL$
$\rightarrow ssLssssss \rightarrow ssdFdssssss$
$\rightarrow ssdddFddddssssss$
$\rightarrow ssdddLSddddssssss$
$\rightarrow ssdddLLLLddddssssss$
$\rightarrow ssddddssssddddssssss$

b)



c) $F \rightarrow dFd \rightarrow ddFdd \rightarrow ddLSdd$
$\rightarrow ddLLdd \rightarrow ddLsdd \rightarrow dddFdsdd$

Fig. 1. Production of RNA structures by the grammar. (**a**) The rules being used, starting from *S*. (**b**) The corresponding structure. (**c**) The production of a stem with a bulge.

The grammar

A grammar consists of a set of variables, some terminal and some non-terminal. Specifically, a starting non-terminal called *S* is contained in a grammar. The non-terminals are rewritten according to a set of production rules, which (for an SCFG) specifies a single non-terminal and a string of variables, that it should be changed to. Successive production rules are applied until only a string of terminals are left (Chomsky, 1959; Lari and Young, 1990).

The basis of the model is a simple SCFG with the following production rules:

$$S \rightarrow LS \mid L$$
$$F \rightarrow dFd \mid LS$$
$$L \rightarrow s \mid dFd$$

Here, *s* symbolizes a base in a single string and *d*s symbolize bases that pair up in a stem. The non-terminal *S* produces loops and *F* produces stems, while *L* decides whether a specific loop position should be a single base or the start of a new stem. An illustration is given in Figure 1. Stems can have any length, whereas loops have lengths of at least two positions, due to the fact that *F* produces *LS* instead of just *S*. Here, positions should be understood in the broad interpretation that the start of a new stem is also a position. This means that the two-position loops can either be two bases, one base and a new stem (a bulge), or two new stems (a bifurcation).

The probabilities of the production rules determine the prior distribution of secondary structures, in that each structure has a certain probability given by the SCFG. The SCFG production probabilities are estimated from a training set of folded RNAs.

Most literature on SCFGs assumes the grammar to be in Chomsky normal form for the algorithms to be used (Chomsky, 1959; Baker, 1979; Lari and Young, 1990). The algorithms are easily adapted to other forms, which has been done here. The number of computations needed for solving problems with these algorithms is proportional to the cube of the sequence length.

Probabilities of columns

First, we look at the columns of the alignment one at a time. Denote the number of sequences in the alignment by *n*, which gives the height of the columns. The probability of a column of non-pairing bases is assumed to be independent of the other columns, given the tree relating the sequences. Likewise, the probability of two pairing columns is assumed to be independent of any other columns, again given the tree.

Let $p = (p_A, p_U, p_G, p_C)$ be the distribution of bases in loop regions of RNA sequences. Furthermore, for $X \neq Y$, let r_{XY} denote the rate of mutation from base *X* to base *Y*, while r_{XX} is the negative of the rate by which base *X* mutates to other bases. The rate matrix can be written as:

$$R = \begin{bmatrix} r_{AA} & r_{AC} & r_{AG} & r_{AU} \\ r_{CA} & r_{CC} & r_{CG} & r_{CU} \\ r_{GA} & r_{GC} & r_{GG} & r_{GU} \\ r_{UA} & r_{UC} & r_{UG} & r_{UU} \end{bmatrix}$$

The rates are assumed to satisfy:

$$p_X r_{XY} = p_Y r_{YX} \text{ for } X \neq Y$$

which means that the overall flow from base *X* to *Y* equals the flow from base *Y* to *X*, i.e. reversibility of mutations. The r_{XX} values are calculated as

$$r_{XX} = -\sum_{Y \neq X} r_{XY}$$

This is known as a general reversible model (Tavaré, 1986). Given a tree, including branch lengths, the column probabilities are calculated using post-order traversal as described by Felsenstein (1981).

For base pairs, the probability of the two columns, given that they form a pair, is calculated using a similar rate matrix, except that base pairs are used instead of single bases. We thus have a distribution of the 16 base pairs and a 16×16 rate matrix, i.e. a general reversible model for base pairs. The reason for including all 16 base pairs is to make it possible to model rare non-standard base pairs. Parameterized rate matrices for base pairs were described by Muse (1994) and Schöniger and von Haeseler (1994). This way of looking at base pairs means that any

base pair change, like AU to GC, is regarded as a single mutation. Even very closely related sequences show these 'double' mutations. Pairs of the rRNAs described below, with sequence identity of 98% or more, were analysed (again as described below). This showed that the base pair mutations between them consisted of 22% 'double' mutations, justifying using a full 16 × 16 matrix for the mutation model. This makes it possible to exploit the differences in base distribution and mutation patterns between loops and stems to obtain good structure predictions.

If a gap is present in one of the sequences, it is handled by treating it as an unknown base, according to the overall base distribution in the model.

Probability of an alignment

Now the entire alignment is taken into consideration. The columns are numbered C_1, C_2, …, C_l, where l denotes the total length of the alignment. The input data, D, are then given as the ordered set of columns: $D = (C_1, C2, \ldots, C_l)$. By M, denote the model including the mutational model and the SCFG. Assuming that the tree is known and the model given, the probability of the alignment can be found. This is done by summing over all possible secondary structures, σ:

$$\begin{aligned} & P(D|T,M) \\ &= \sum_{\sigma} P(D,\sigma|T,M) \\ &= \sum_{\sigma} P(D|\sigma,T,M)P(\sigma|T,M) \\ &= \sum_{\sigma} P(D|\sigma,T,M)P(\sigma|M) \end{aligned}$$

The last equality stems from the fact that the secondary structure only is dependent on the tree through the data. The terms $P(\sigma|M)$ are probabilities of secondary structures, given the model. These are the prior probabilities from the grammar previously described.

The terms $P(D|\sigma, T, M)$, i.e. the alignment probabilities, given the secondary structure and the tree, are products of the column probabilities. This results from the assumption that columns which do not pair are independent:

$$\begin{aligned} & P(D|\sigma,TM) \\ &= P(C_1 \cdot \cdot \cdot C_n|\sigma,T,M) \\ &= \prod_{s} P(C_s|\sigma,T,M) \prod_{\mathrm{d}} P(C_d C_{d^c}|\sigma,T,M) \end{aligned}$$

The product over s is over the columns of single bases, while the product over d is over left columns of pairs, while the d^cs are the corresponding right columns of the pairs.

The sum can be calculated using a dynamical programming approach (Baker, 1979), by extending the view of the grammar described above to include productions of columns as follows. When an s is used in a production rule, it corresponds to a column in the alignment of sequences. Such a column has a probability, given the tree, which is multiplied to the production probability each time an s is produced. Likewise, probabilities for rules producing base pairs, like $F \rightarrow dFd$, are multiplied to the probability of the two columns, given that they form a pair. This makes the grammar equivalent to a grammar that generates columns in alignments instead of just secondary structure, meaning that for a two-sequence alignment, the production rule $L \rightarrow s$ covers the following rules:

$$L \rightarrow \begin{bmatrix} X \\ Y \end{bmatrix} \text{ for } X, Y \in \{\mathrm{A}, \mathrm{U}, \mathrm{G}, \mathrm{C}\}$$

with $\begin{bmatrix} X \\ Y \end{bmatrix}$ denoting a column with the base X in the first sequence and the base Y in the second sequence. Thus, for n aligned sequences a rule like $L \rightarrow s$ covers 4^n rules, while a rule like $F \rightarrow dFd$ covers 4^{2n} rules (some being unlikely, with rare base pairings).

The full model

If the phylogenetic tree relating the sequences is not given, it must be estimated from the model. For a given tree, T, $P(D|T, M)$ can be calculated as above. The ML estimate of the tree, given the model, can then be obtained by:

$$T^{ML} = \underset{T}{\operatorname{argmax}} \; P(D|T,M)$$

which can be found by using numerical optimization: given a tree topology, the branch lengths can be obtained by maximizing the probability of the alignment, $P(D|\text{topology}, M)$. This is a $2n - 3$ dimensional search for a maximum, which can be done using standard methods (e.g. Press *et al*., 1992). Estimating tree topology can, for example, be done by an exhaustive search, a branch and bound method or a heuristic method (Swofford *et al*., 1996). The choice will be highly dependent upon the number of sequences in the alignment, considering the fast rate of growth in the number of trees with respect to the number of sequences. The maximum likelihood estimate of the tree is used in the MAP estimation of the structure. It would be better to integrate over all trees during the structure determination, but the above described approach is simpler.

The alignment of sequences is the data to be used in the secondary structure estimation. To perform a MAP estimation, we need to maximize $P(\sigma|D, T, M)$, which means to find the most likely secondary structure, given what we know. Using Bayes theorem, while conditioning on T and M, we obtain:

$$\begin{aligned} & P(\sigma|D,\ T,\ M) \\ &= \frac{P(D|\sigma,T,M)P(\sigma|T,M)}{P(D|T,M)} \\ &= \frac{P(D|\sigma,T,M)P(\sigma|M)}{P(D|T,M)} \end{aligned}$$

$P(\sigma|M)$ is the prior distribution of structures given by the SCFG. $P(D|T, M)$ is independent of the structure, and thus constant over all structures. The MAP estimate of the structure is then given by:

$$\sigma^{MAP} = \underset{\sigma}{\text{argmax}} \; P(D|\sigma, T^{ML}, M)P(\sigma|M)$$

which is found using the CYK algorithm (Durbin *et al.*, 1998) on the extended grammar, producing alignments.

From the posterior secondary structure prediction, various questions regarding the structure can be answered, including the most probable overall secondary structure (MAP estimate), the certainty of the prediction in each position and probabilities of the pairing of specific bases.

Implementation

The model was estimated in a number of steps:

1. A suitable database of sequences with known structures was made.
2. Single base and base pair frequencies were estimated.
3. Mutation rates were estimated.
4. The grammar parameters were estimated.

The database

The database used for estimating this model should represent RNA secondary structures in general, because it is attempted here to model RNA structures as a whole. For this reason, the database should be composed of various types of RNA. tRNAs and large subunit ribosomal RNAs (LSU rRNAs) were chosen. These are publically available and have well-established structures. The database made here consists of RNA sequences along with their entire secondary structures.

The tRNAs are from the database by Sprinzl *et al.* (1998). Part one of this database contains 2146 aligned tRNA gene sequences with corresponding RNA structures. This database was reduced by removing sequences with unknown bases and the like. Furthermore, interior loops, having one unpaired base on each side, were changed into stems [e.g. structures like '(.(.....).)' were changed to '(((.....)))']. [Parenthesis notation is used for describing the structures in this article. Matching parentheses (or later, brackets) denote positions that form a pair.] This pairs the non-standard pairs that the structures imply, which are assumed to bond (this is sometimes true, sometimes not). Allowing for non-standard base pairs gives the algorithm more robustness towards sequencing and alignment errors. Before this operation, the database only contained AU, GC and GU base pairs. The revised database had 1968 tRNA sequences with corresponding secondary structures.

The LSU rRNAs are from a database by De Rijk *et al.* (1998), which contains 709 sequences. A reduction was performed as above, resulting in 305 remaining sequences. This database contains a number of non-standard base pairs.

The training was carried out with a weighting of the sequences to represent the two RNA families equally.

Frequencies

The single base frequencies were estimated from counts of the bases in the single base positions of the sequences. Overall base frequencies were also determined. Base pair frequencies were estimated by counting base pairs. Interestingly, tRNAs show more GC than CG base pairs, meaning that in GC/CG base pairs the G tend to be nearer the 5′ end of the RNA than the C. This might have to do with functional constraints on evolution. As this model aims to be general, unique characteristics of the training sequences should not be modelled. Therefore, to obtain equal frequencies of *XY* and *YX* base pairs, each occurrence of an *XY* base pair also counted as a *YX* base pair (the rarely occurring pairs of identical bases were counted twice).

The obtained frequencies are shown in Table 1, which shows that the overall base frequencies are approximately equal. In stems, there are a significant majority of GC/CG base pairs, which probably has to do with the high binding energy associated with this pair.

Table 1. Base frequencies, showing nearly equal overall distribution of bases, with a slight underrepresentation of Cs. Stems have high GC/CG base pair frequencies, while loops have low content of Cs and Gs. The lowest row shows the distribution of bases between loops and stems

Stem		Loop		Overall	
AU/UA	35.6%	A	36.4%	A	26.8%
GC/CG	53.4%	C	15.1%	C	21.4%
UG/GU	9.8%	G	21.2%	G	26.7%
Other	1.2%	U	27.3%	U	25.1%
Total: 52.6%		Total:	47.4%		

Mutation rates

For estimating mutation rates, a number of sequences from the above-described database were paired. All possible ordered pairs were made of sequences having at least 85% identical base sequences. The 85% limit makes it reasonable to assume that only single mutations, in the sense of the mutation mechanisms described above, have occurred between the sequences. The single base positions in these sequence pairs were examined and all differences between the sequences counted. Thus, if a given position had base *X* in one sequence and base *Y* in the other, the counters c_{XY} and c_{YX} were incremented. Columns, in the pairs, having a gap, were not used. For a given pair, *P*, define t_P as the time between

sequences, N_P as the number of columns in the two-sequence alignment and P_S as the probability of a base being in a single base position. Because of the single mutation assumption, we have for $X \neq Y$:

$$E(c_{XY}) \approx \sum_P P_s N_p (p_x t_p r_{XY} + P_Y t_p r_{YX})$$
$$\Rightarrow E(c_{XY}) \approx 2 \sum_P P_s N_P t_P p_x r_{XY}$$
$$\Rightarrow r_{XY} \approx \frac{c_{XY}}{2 p_X P_s \sum_P t_P N_P} = K \frac{c_{XY}}{p_X P_s}$$

where the sums over P are over all pairs. K is a constant that is independent of X and Y, implying that $r_{XY} \propto c_{XY}/(p_X P_s)$ for all bases X and Y with $X \neq Y$. To ensure equal weighting of information from different sequences, the count from pairs having the same first sequence was divided by the number of pairs having this first sequence. This should decrease the variance of the estimates and only affect the constant K, ensuring that we still have $r_{XY} \propto c_{XY}/(p_X P_s)$. The rates were normalized so that the total rate of mutations in single base positions was one, making the rate matrix uniquely determined. In the article by Goldman *et al.* (1996), the rates were found in a similar way, except that the constant K was divided out. This meant that they had to estimate amino acid frequencies from the rate matrix. In this work, it is viewed as essential to have the best possible estimates of base frequencies, thus the rate matrix is estimated using these.

Pairs were counted using symmetry, both in position and time. The counts were dealt with in a similar fashion as the single-base counters. The normalization was performed relative to the single base rates, which shows that the mutation rate, considered on a single base level, for stem regions is 0.90 times the rate for single bases.

The mutation rates for single bases are shown in Table 2. Variations between mutation rates are observed. It is obvious that transitions (A–G and T–C mutations in DNA) are more frequent than transversions (the rest), which agrees with established knowledge (e.g. Gojobori *et al.*, 1982).

Table 2. The entries, r_{XY}, for the loop rate matrix. Transitions are more frequent than transversions

$X\backslash Y$	A	C	G	U
A	–0.75	0.16	0.32	0.26
C	0.40	–1.57	0.24	0.93
G	0.55	0.17	–0.96	0.24
U	0.35	0.51	0.19	–1.05

Mutation rates for the most frequent base pairs are shown in Table 3. This table shows that the pair mutations requiring only a single base change are the most frequent, while mutations requiring two transversions are very rare. This is what should be expected. Table 4 shows the mutation rates for single bases in stem regions. This table shows that the transition/transversion ratio is higher in stem regions than in loop regions. This is because single transversions disrupt pairing, while transitions may conserve pairing (e.g. both A and G can pair with U).

Table 3. Some of the entries for the stem rate matrix. Only rates between the six most frequent base pairs are included

$X\backslash Y$	AU	UA	GC	CG	UG	GU
AU	–1.16	0.18	0.50	0.12	0.02	0.27
UA	0.18	–1.16	0.12	0.50	0.27	0.02
GC	0.33	0.08	–0.82	0.13	0.02	0.23
CG	0.08	0.33	0.13	–0.82	0.23	0.02
UG	0.08	1.00	0.10	1.26	–2.56	0.04
GU	1.00	0.08	1.26	0.10	0.04	–2.56

Table 4. Marginal rate matrix for stems. This matrix is similar to the above matrix for loops, except that this one was estimated from stem regions. Notice the high transition/transversion ratio relative to loops

$X\backslash Y$	A	C	G	U
A	–1.15	0.13	0.79	0.23
C	0.09	–0.84	0.16	0.59
G	0.45	0.13	–0.70	0.11
U	0.18	0.70	0.16	–1.03

Grammar parameters

The production probabilities of the grammar reflect the way secondary structures behave: loop lengths, stem lengths, bifurcations, etc. For estimating these probabilities, secondary structures from the database were used. This estimation can be done using the inside–outside algorithm (an expectation maximization procedure) on this training set of secondary structures (Baker, 1979; Lari and Young, 1990). In the case of the simple grammar described here, the number of times each rule is used is uniquely determined by the training set, meaning that only one iteration had to be performed. Furthermore, the counting was performed in a simple way, which made it possible to analyse the long LSU rRNA sequences. The production probabilities obtained were the following:

$$S \rightarrow LS\ (86.9\%) \mid L\ (13.1\%)$$
$$F \rightarrow dFd\ (78.8\%) \mid LS\ (21.2\%)$$
$$L \rightarrow s\ (89.5\%) \mid dFd\ (10.5\%)$$

Probabilities are written in parentheses.

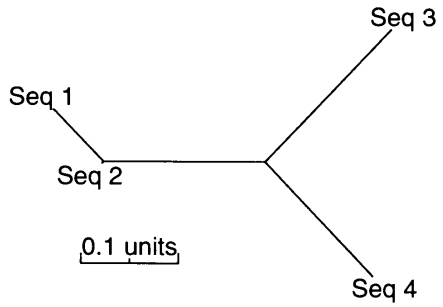


Fig. 2. The phylogenetic tree relating the four analysed sequences, as calculated using the ML estimation described above. The length units correspond to the rate matrices of the model.

Results

The test sequences

To test the method described here, four representative bacterial RNase P RNA sequences were chosen from the database by Brown (1998) and analysed:

Sequence 1	*Klebsiella pneumoniae*
Sequence 2	*Serratia marcescens*
Sequence 3	*Pseudomonas fluorescens*
Sequence 4	*Thiobacillus ferrooxidans*

The structures and alignment of the sequences are known. The sequences have lengths ranging from 344–383 bases, while their structural alignment has a total of 385 columns. The pairwise sequence identities range from 65–92%. The relationships between the sequences are shown in Figure 2, while the alignment is shown in Figure 3. The pseudoknot denoted by square brackets from positions 68–76 and 368–361 could be written using parentheses in these positions and square brackets in position 18–12 and 370–364. This is a stem of seven positions, while the other pseudoknot has four pairs. This means that a structure prediction of this type will have at least 22 positions wrongly predicted in each sequence.

Using related sequences

A number of predictions were made from the four RNase P RNA sequences by the method described above. The accuracy of a prediction is here defined as the total number of non-gap positions in each sequence having the correct assignment, divided by the total number of non-gap positions. A base pair is only considered correct if both base positions are correct. The alignments used were the structural alignments from the database by Brown (1998). Firstly, all sequences were analysed one by one, then all six pairs of sequences were used, then all four triples, and finally all the sequences were used. The results in the top of Table 5 show very significant improvement of prediction accuracy when sequences are added, especially going from one to two sequences. This exemplifies the large potential of methods using several sequences and their phylogeny, in making RNA secondary structure predictions. The pseudoknot stems, denoted by brackets in Figure 3, are invariant throughout these four sequences, which makes them hard to predict when using mutational patterns.

Table 5. Structural alignment, no phylogeny

	Structural alignment			
No. of sequences	1	2	3	4
Min result	41.2%	65.2%	73.9%	79.2%
Max result	57.7%	82.1%	79.6%	79.2%
Average	48.3%	73.6%	77.8%	79.2%
	CLUSTAL W alignment			
No. of sequences	1	2	3	4
Min result	41.2%	54.9%	60.1%	73.8%
Max result	57.7%	69.1%	76.9%	73.8%
Average	48.3%	64.4%	68.5%	73.8%
	Structural alignment, no phylogeny			
No. of sequences	1	2	3	4
Min result	41.2%	59.9%	67.7%	76.2%
Max result	57.7%	76.6%	76.6%	76.2%
Average	48.3%	68.9%	72.2%	76.2%

Table 6. What happens when a limit of certainty is imposed on the results. Each row shows how many positions have a certainty above a given limit and how many of these are correctly predicted. There is a high correlation between the accuracy of prediction and the certainty that the model predicts

Limit	No. of positions	Correct positions	Accuracy
0%	1459	1156	79.2%
50%	1314	1146	87.2%
70%	1150	1064	92.5%
80%	1068	1014	94.9%
90%	932	890	95.5%
95%	825	799	96.8%

In many situations, the structural alignment is not available. Therefore, it is necessary to assess the results using an alignment algorithm. For this, the same analyses as above were made, but each subset of the four sequences was aligned using CLUSTAL W (Thompson *et al.*, 1994). The results of this are shown in the middle of Table 5 with the column of one sequence identical to the earlier analysis. This gave lower accuracies than using structural alignments, which is not surprising. The rise in accuracy, when using more sequences, now arises both from better alignments and more data. Good results are still obtained using four sequences.

Neglecting phylogeny

If the phylogeny of the sequences is not taken into account, some information is lost and poorer prediction results. Such results are shown at the bottom of Table 5. This table was made like the top of Table 5, but using long branch lengths to simulate independent sequences. For two and three sequences, the phylogenetic information improves the result by ~5%. This shows that the tree conveys information about the structure. Results are compared in Figure 4.

Weight of results

The algorithm allows for a calculation of the probability that each position is correctly predicted. This is done using the inside and outside variables. It can give the user of the method an impression of how certain the predictions are, assuming that the model is correct. This can be considered as an equivalent to the partition function for energy calculations (McCaskill, 1990).

Taking the analysis of the structural alignment of all four sequences with a phylogenetic tree as an example, results from choosing only to believe regions of high certainty are shown in Table 6. This shows that discarding, for example, positions having a certainty of <70% means that 309 positions are discarded, of which only 92 were correctly predicted. This results in an accuracy of prediction for the remaining positions of 92.5%. This, of course, does not improve overall accuracy, but shows that badly predicted areas can be pointed out. Other methods, perhaps experimental, can then be used for these areas.

Comparison with other methods

To give an impression of the performance of this method relative to other methods, some comparisons have been made. The folding program, MFOLD Version 3.0 (Web server URL: http://www.ibc.wustl.edu/~zuker/rna/form1.cgi), by Zuker (1989) and Walter *et al*. (1994), using energy minimization was used for folding the four sequences one by one. Standard parameters were used, resulting in predictions ranging from 36 to 68%, with an average of 51% (see Table 7). This is comparable to the above-described method applied to single sequences, but does not suggest that this method is always as good as Zuker's for single sequences. The energy minimization method has more parameters than the above-described model, in the case of one sequence, where evolution does not come into consideration. This gives Zuker's method a potential for better results. Varying the parameters for the method might improve results; furthermore, results will be different for different families of RNA.

The method of maximum weighted matching was used on the four sequences, with the structural alignment. The scoring schemes used here are the ones described by Tabaska *et al*. (1998). Both helix-plot and mutual information were in-

```
     1                                                                                                                  100
1:   GAAGCUGACC AGACAGUCGC CGCUUCGUCG UCGUCCUCCU UCGGGGGGAG ACGGGCGGAG GGGAGGAAAG UCCGGGCUCC AUAGGGCAAG GUGCCAGGUA
2:   GGAGUUGACC AGACAGUCGC CGCUUCAUUG CCGUCCUC-U UCG-GGGGAG ACAGAUGGAG GGGAGGAAAG UCCGGGCUCC AUAGGGCAGG GUGCCAGGUA
3:   AGAGUCGAUU GGACAGUCGC UGCCCUCUAU ---------G AAA------- -AUUAGGGGG GGGAGGAAAG UCCGGGCUCC AUAGGGCGAA GUGCCAGGUA
4:   GGAGUGGGCC AGGCGACCGC CGCGGA---- ---------G CAA------- ------UCCG GGGAGGAAAG UCCGGGCUCC AUAGGGCAAG GCGCCGGUUA
s:   ((((((((((  (((((((((.( (.(((((((( {(.((((((. ...))))))) })).)))))) .)).....[[[ .[[[[[(((( ...[[[[.{( ((((((((..
p:   ((((((((((  .(((((((((  ..(((((((( (((((((((. ..)))))))) ))).)))))) ....(((... ))))..((((( ........(( ((((((((..

     101                                                                                                                200
1:   ACGCCUGGGG GGUGUCACGA CCCACGACCA GUGCAACAGA GAGCAAACCG CCGA-UGGCC CGCGCAAGCG GGAUCA-GGU AAGGGUGAAA GGGUGCGGUA
2:   ACGCCUGGGA GGC-GCAA-G CCUACGACUA GUGCAACAGA GAGCAAACCG CCGA-UGGCC CGCGCAAGCG GGAUCA-GGU AAGGGUGAAA GGGUGCGGUA
3:   AUGCCUGGGG GGC-GUGA-G CCUACGGAAA GUGCCACAGA AAAUA-ACCG CCUAAGCAC- ---UUCG--- -G-UGCCGGU AAGGGUGAAA AGGUGCGGUA
4:   ACGGCCGGGG GGC-GUGA-G CCUACGGAAA GUGCCACAGA AAAUAUACCG CCAA-GCGC- ---GUAA--- -G-CGC-GGU AAGGGUGAAA AGGUGCGGUA
s:   ...))))))( (((......) )))..)((... ((........ .......((( ((...((((( (((....))) )).))).))) ..)).....  .((((((...
p:   ...))))))..( (((......) ))).....  .........  ......(((( .....((((( (((....))) )).)))))))  ..((((.... .((((((...

     201                                                                                                                300
1:   AGAGCGCACC GCGCGGCUGG UAACAGUCCG CGGCACGGUA AACUCCACCC GGAGCAAGGC CAAAUAGGGG UUCAUAAGGU ACGGCCCGUA CUGAACCCGG
2:   AGAGCGCACC GCGCGGCUGG UAACAGUUCG UGGCACGGUA AACUCCACCC GGAGCAAGGC CAAAUAGGGG UUCACAUGGU ACGGCCCGUA CUGAACCCGG
3:   AGAGCGCACC GCACGACUGG CAACAGUUCG UGGCUAGGUA AACCCCACUU GGAGCAAGAC CAAAUAGGGU UCCA--AGGC GUGGCCCGCG CUGGAACCGG
4:   AGAGCGCACC GCAUUUCCGG UAACGG-AAA UGGCAGGGAA AACCCCGCCU GGAGCAAGAC CAAAUAGGCG UGCGA-UACC GUGGCCCGCG GUGCACGCGG
s:   ....))))))  .(((((((( ...)))).)))  )).....  .)))))))}  ))))...(((  (......(((  (((....(((  ((.]]]]))))  ))))))))..
p:   ....))))))  .(((((((( ...)))).)))  )).....  .)))))))))  )))))...((  (......(((  ((((....((  (((...))))  ))))))))..

     301                                                                                                  385
1:   GUAGGCUGCU UGAGCCAGUG AGCGAUUGCU GGCCUAGAUG AAUGACUGUC CACGACAGAA CCCGGCUUAU CGGUCAGUUU CACCU
2:   GUAGGCUGCU UGAGCCAGUG AGCGAUUGCU GGCCUAGAGG AAUGACUGUC CACGACAGAA CCCGGCUUAU CGGUCAACUC CCUC-
3:   GUAGGUUGCU AAAGAUGUCC AGUGAUGGCC AUCGUAGACG AAUGACUGUU CAAGACAGAA CCCGGCUUAU AGAUCGACUC UCCAC
4:   GUAGGUUGCU GGAGCCUGUG CGUAAGUGCA GGCCUAGAGG AAUGGUCGUC CACGACAGAA CCCGGCUUAU CGGCCCACUC CAAUU
s:   ...))))...  ...((((.{(( (....)))}) )))......  ...)))))))  .........  ]]]]]]]].)  .)))))))))  ).....
p:   ...))).)).  ...(((((((  (....)))))  )))......  ...)))))))  .........  .........  .)))))))))  ).....
```

Fig. 3. The alignment of the four RNase P RNA sequences. The predicted structure, using all four sequences, is denoted p. The structur e from the database is denoted s, with square brackets denoting parts of pseudoknots. The square brackets used here match the structure description in the database. The curly brackets denote positions where the structure differs: the sequences that have a non-standard pair in these positions have loop regions or bulges, the rest have pairs.

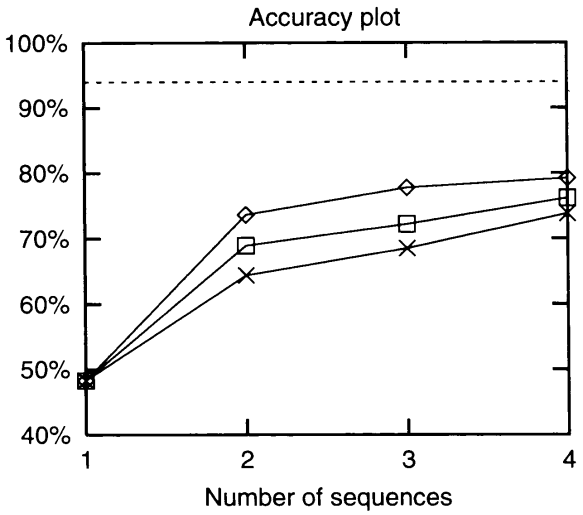


Fig. 4. A comparison of results with and without phylogeny. Diamonds (◇) denote the curve for predictions with phylogeny, while boxes (□) denote the one without. Crosses (×) denote results using CLUSTAL W alignments and phylogeny estimation. The dotted line at 94% represents the maximum possible prediction accuracy with regard to the pseudoknots.

corporated, giving a maximum of 60% accuracy. The covariance method, COVE Version 2.4.4., by Eddy and Durbin (1994), was also tried on the sequences, with lower accuracy. These methods were developed for larger numbers of sequences, and should not be expected to give optimal results using only four sequences. This shows the significance of the method described here in situations where only a few sequences are known.

Table 7. Accuracy table, showing comparisons of single sequence predictions using the method described in this paper and MFOLD Version 3.0, by Zuker (1989) and Walter *et al.* (1994). Predictions of secondary structures were made on single sequences, which is the only possibility using MFOLD. The average results are comparable

Sequence	SCFG method	MFOLD
Seq 1	57.7%	67.1%
Seq 2	48.2%	54.0%
Seq 3	41.2%	35.6%
Seq 4	46.2%	50.3%
Average	48.3%	51.7%

Conclusion

The limitations of this method include:

- Inability to predict pseudoknots.
- Loop and stem lengths are assumed to be geometrically distributed.
- A good alignment is needed.
- The dynamical programming algorithms are relatively slow. [They have a time complexity of $O(N^3)$ with respect to the length of the alignment.]

The problem with pseudoknots is shared by many algorithms (e.g. Zuker, 1989; Eddy and Durbin, 1994), although some algorithms can predict pseudoknots (e.g. Tabaska *et al.*, 1998).

The problem relating to the length distributions has to do with the nature of the specific SCFG used here, and can be solved by making different non-terminals producing stems or loops of various lengths. Special non-terminals describing small bulges will probably improve results. This introduces some extra computation time, but can definitely be carried out.

The problem of the alignment is not easily solved, because making an alignment without knowing the structure is unlikely to produce a structural alignment. It might be possible to realign sequences once a structure prediction has been made. This approach would probably be prone to local maxima in the likelihood function for alignments. One possible way of avoiding this would be to use Gibbs sampling in a Markov chain Monte Carlo method, sampling from alignments and summing over structures (Gilks *et al.*, 1996).

An alignment method which simultaneously folds and aligns a set of RNA sequences to find common structural elements locally has been implemented by Gorodkin *et al.* (1997). The algorithm has a computational complexity of $O(N^4)$, relative to the sequence length. The method has proven useful for relatively short sequences, and an alignment produced by such a method would be a good starting point for SCFG methods (Gorodkin *et al.*, 1997), including the one described here.

Profile SCFGs and covariance models predict secondary structure at the same time as making alignments, but seem to need a large number of sequences (Eddy and Durbin, 1994; Sakakibara *et al.*, 1994). Further work in making algorithms for simultaneous RNA folding and alignment will probably show up in the future because of the importance of solving this problem.

Further improvements to the method introduced here include modelling base stacking, which is not very difficult. It consists of conditioning the probability of two pairing columns on the neighbouring columns. Thus, in estimating the model, neighbour base pairs should be counted to indicate the conditional distributions of base pairs. This information could then be used in the calculations to give improved results.

Finally, it would be interesting to look into the evolutionary model proposed here. Statistical tests of its ability to describe RNA evolution would be enlightening. It would also

be useful to reduce the number of parameters for the rate matrices, especially the base pair rate matrix, e.g. as done by Muse (1994).

It is the hope of the authors that this method can be made available to the public via the World Wide Web.

Acknowledgements

We would like to thank Jan Gorodkin, Carsten Wiuf and the anonymous reviewers for critically reviewing the manuscript and suggesting improvements. Furthermore, J.H. acknowledges generous support by the Newton Institute for Mathematical Sciences in Cambridge, UK.

References

Baker,J.K. (1979) Trainable grammars for speech recognition. In Klatt,D.H. and Wolf,J.J. (eds), *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*. Acoustal Society of America, New York.

Brown,J.W. (1998) The ribonuclease P database. *Nucleic Acids Res*., **26**, 351–352.

Cary,R.B. and Stormo,G.D. (1995) Graph-theoretic approach to RNA modeling using comparative data. In Rawlings,C., Clark,D., Altman,R., Hunter,L., Lengauer,T. and Wodak,S. (eds), *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*.

Chomsky,N. (1959) On certain formal properties of grammars. *Info. Control*, **2**, 137–167.

De Rijk,P., Caers,A., Van de Peer,Y. and De Wachter,R. (1998) Database on the structure of large ribosomal subunit RNA. *Nucleic Acids Res*., **26**, 183–186.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*., **22**, 2079–2088.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*., **17**, 368–376.

Gilks,W.R., Richardson,S. and Spiegelhalter,D.J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Gojobori,T., Wen-Hsiung,L. and Graur,D. (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol*., **18**, 360–369.

Goldman,N., Thorne,J.L. and Jones,D.T. (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol*., **263**, 196–208.

Gorodkin,J., Heyer,L.J. and Stormo,G.D. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*., **25**, 3724–3732.

Lari,K. and Young,S.J. (1990) The estimation of stochastic context-free grammars using the inside-outside algorithm. *Comput. Speech Lang*., **4**, 35–56.

McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

Muse,S.V. (1994) Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics*, **139**, 1429–1439.

Nussinov,R., Pieczenik,G., Griggs,J.R. and Kleitman,D.J. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math*., **35**, 68–82.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) *Numerical Recipes in C*, 2nd edn. Cambridge University Press, Cambridge.

Sakakibara,Y., Brown,M., Hughey,R., Mian,I.S., Sjölander,K., Underwood,R.C. and Haussler,D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res*., **22**, 5112–5120.

Schöniger,M. and von Haeseler,A. (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol*., **3**, 240–247.

Sprinzl,M., Horn,C., Brown,M., Ioudovitch,A. and Steinberg,S. (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res*., **26**, 148–153.

Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M. (1996) Phylogenetic inference. In Hillis,D.M. and Moritz,C. (eds), *Molecular Systematics*, 2nd edn. Sinauer Associates, pp. 407–514.

Tabaska,J.E., Cary,R.B., Gabow,H.N. and Stormo,G.D. (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, **14**, 691–699.

Tavaré,S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In *Lectures on Mathematics in the Life Sciences*. American Mathematical Society, Vol. 17, pp. 57–86.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Walter,A.E., Turner,D.H., Kim,J., Lyttle,M.H., Mueller,P., Mathews,D.H. and Zuker,M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA*, **91**, 9218–9222.

Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res*., **9**, 133–148.

Chapter 3

Practical RNA Secondary Structure Prediction Using Stochastic Context-Free Grammars

Manuscript by Bjarne Knudsen and Jotun Hein.

Submitted to *Bioinformatics*.

Abstract

Motivation: RNA secondary structures are important in many biological processes, and efficient structure prediction can give vital directions for experimental investigations. Most available programs for RNA secondary structure prediction only predict structures for a single sequence at a time. This may be sufficient in some applications, but often it is possible to obtain related RNA sequences with conserved secondary structure. These should be included in structural analyses to give improved results.
Results: This work presents a practical way of predicting RNA secondary structure that is especially useful when related sequences can be obtained. It is shown that including these sequences in the analysis greatly improves prediction accuracy.
Availability: RNA secondary structures can be predicted on a web server at http://www.daimi.au.dk/~compbio/rnafold.
Contact: bk@birc.dk

Introduction

Many computational methods for RNA structure prediction have been developed. Early algorithms were made by Nussinov *et al.* (1978) and Zuker and Stiegler (1981). Zuker's energy calculations have been further improved (Zuker, 1989; Walter *et al.*, 1994; Mathews *et al.*, 1999), and are probably the most widely used RNA secondary structure prediction method today.

This work improves the basic algorithm of Knudsen and Hein (1999) (here, denoted the $\mathcal{KH}$-99 algorithm), and makes it publicly available.

Algorithm

The algorithm presented in this article is based on the $\mathcal{KH}$-99 algorithm. The original algorithm was only useful for a limited number of sequences, due to its large computation time. For the same reasons, it was not made available to the public. This work makes the algorithm practically useful for larger amounts of sequences. The main concerns are the treatment of gaps, computational speed, and robustness.

The $\mathcal{KH}$-99 algorithm

The $\mathcal{KH}$-99 algorithm uses a stochastic context-free grammar (SCFG) to produce a prior distribution of RNA structures. Given an alignment and a phylogenetic tree relating the aligned sequences, the posterior probabilities of the structures can be calculated. The posterior probability is based on individual probabilities for the columns, or pairs of columns, in the case of a base-pair. Column probabilities are calculated using the likelihood approach by Felsenstein (1981). The evolution of column pairs is modeled using a rate matrix for

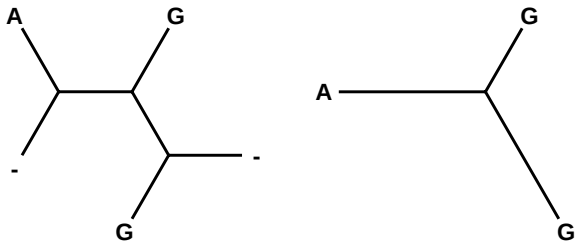


Figure 1: The effect of treating gaps as unknown nucleotides. Only a single column from the alignment is considered, with the nucleotides put at the leaves of the phylogenetic tree. When gaps are treated as unknown nucleotides the two trees have identical probabilities since leaves with gaps can be removed.

base-pairs, i.e. a 16 by 16 matrix. The most likely structure is found using the inside-outside algorithm (Lari and Young, 1990). The tree is estimated using a maximum likelihood approach in the SCFG model described.

Gaps

Treating gaps in a sensible way is a returning problem in biological sequence analysis. The best way to deal with gaps would probably be to make an explicit model for insertions and deletions, and use that in the sequence analysis. Unfortunately, such calculations are often complicated, as in statistical alignment (Bishop and Thompson, 1986; Thorne *et al.*, 1991; Hein *et al.*, 2000).

Two simpler ideas are to treat gaps as a fifth nucleotide or to treat them as unknown nucleotides. When using an evolutionary model, many problems arise from treating a gap as a fifth nucleotide. First of all, the frequency of gaps will depend on how many sequences are being analysed and of their evolutionary distance. Furthermore, it cannot be viewed as evolving in the same way as a nucleotide, thus the rates of evolution are difficult to specify.

When treating gaps as unknown nucleotides, the gapped sequence position should have probability one for any nucleotide. This has the advantage that the probability of a column with gaps is equal to the probability of the same column in an alignment without the gapped sequences, and the tree correspondingly pruned, see Figure 1.

From this, it seems that gaps should be treated as unknown nucleotides. This is indeed what is often done in situations where different alignment columns are looked at individually (e.g., Thorne *et al.*, 1996). The problem arises when pairs of columns are analysed together, as is the case for RNA structure prediction. When treating gaps as unknowns, there is no immediate way of controlling whether gaps form pairs with nucleotides, the result of which is shown in the top of Figure 2.

The problem was handled by removing columns where less than 75% of

```
Seq 1       CGAC-----AGCUGAGUGUGACUUUAGAAU
Seq 2       UGACGGUCUAGCUGACUGAUACUUCAGAGU
Seq 3       CGAC-----AGCUGAAUGAGACUUCAGAAU
Structure   ...(((((.....)))))............

Seq 1       CGAC-----AGCUGAGUGUGACUUUAGAAU
Seq 2       UGACGGUCUAGCUGACUGAUACUUCAGAGU
Seq 3       CGAC-----AGCUGAAUGAGACUUCAGAAU
Structure   ..........((((........))))...
```

Figure 2: A structure prediction for three hypothetical sequences. In the top alignment, gaps are treated as unknown nucleotides. The structure, shown as parentheses, include pairs between nucleotides and gaps. In the parenthesis notation, corresponding parentheses indicate positions forming base-pairs. In the bottom alignment, the columns with gaps have been left out of the prediction, because less than 75% of the sequences have nucleotides in these positions.

the sequences have nucleotides. The result of this is shown in the bottom of Figure 2. This seems to be a reasonable way to deal with gaps in most cases and was therefore adopted for this algorithm.

Unknown nucleotides

In biological sequences, some nucleotides may be unknown, or only partial information may be available. All these situations can be treated by letting the unknown nucleotide have a probability of one for each of the possible nucleotides. This means, for example, that if a given position is known to be a pyrimidine, its probability of being a U is set to one, and its probability of being a C is also set to one. Using this method, any symbols of the extended nucleotide alphabet can be treated correctly by the algorithm. This is in accordance with Felsenstein (1981).

Tree estimation

In the $\mathcal{KH}$-99 algorithm, the tree was estimated through a maximum likelihood method using the SCFG model. While this gave good results and was interesting with respect to phylogenetic analysis, it was slow. A much faster method is to estimate the tree first. This can be done using standard methods.

In this implementation, pairwise distances between sequences are calculated using maximum likelihood. The rate matrix used should correspond as close as possible to what is used in the $\mathcal{KH}$-99 algorithm. Since the tree is calculated before the structure has been estimated, a single rate matrix has to be used for this purpose. It was made from the $\mathcal{KH}$-99 algorithm by summing the loop rate matrix and a reduction of the base-pair rate matrix to single positions. The rate matrices were weighted with the probabilities that a given position is in

```
Seq 1       UGGCG - - CUAGCCAUCUGAUACUUCAGAUU
Seq 2       UGACGGACUAGCCAACUGAUACUUCAGAUA
Seq 3       U - ACGUAGUAGCCAUCUGAUACUUCAGAU -
Structure   . . . . . . . . . . . . . . ( ( ( ( . . . . ) ) ) ) . . .

Seq 1       UGGCG - - CUAGCCAUCUGAUACUUCAGAUU
Seq 2       UGACGGACUAGCCAACUGAUACUUCAGAUA
Seq 3       U - ACGUAGUAGCCAUCUGAUACUUCAGAU -
Structure   . . . . . . . . . . . . ( ( ( ( ( ( . . . . ) ) ) ) ) ) .
```

Figure 3: In the top alignment, the standard result is given. The bottom alignment shows the structure, when all nucleotides have a 1% chance of being any other nucleotide. The result is a longer stem, which includes one non-standard pair.

a loop region or a stem region, respectively. The tree was calculated from the pairwise distances using the Neighbor Joining algorithm (Saitou and Nei, 1987) and adjusting branch lengths to maximum likelihood estimates. This gave a large increase in speed, since the stochastic context-free grammar calculations only needed to be done once, as opposed to the multiple iterations used in the previous method for estimating the tree.

Robustness

The algorithm assumes that all sequences have exactly the same structure. This means that a single sequence with a slightly different structure might ruin a structure prediction. The same situation applies for alignment errors and sequencing errors. When a single error might change a prediction significantly, the method is not robust. An example is given at the top of Figure 3.

A way to make the algorithm more robust is to let any nucleotide have a small probability of being any other nucleotide. The interpretation of this is most obvious in terms of sequencing errors, but the method works for alignment errors and structure differences, too. Figure 3 shows how introduction of this probability changes the results. In this implementation of the algorithm, the probability was set to 1%.

Partially known structure

It is often the case that something is known about the structure being predicted. Including this knowledge in the analysis can give improved results. Different kinds of knowledge can be included in the analysis:

- That two given columns form a pair together.
- That a given column is involved in a pair.
- That a given column is not involved in a pair.

This can all be included in the calculations by letting the structures that does not satisfy the previous knowledge have a probability of zero. This approach does not change relative prior distributions of allowed structures.

As a side note, no loops of length two is allowed in this implementation, as opposed to the $\mathcal{KH}$-99 algorithm. This was implemented by disallowing pairs between positions of distance less than four.

What structure should be chosen?

A prediction program should of course report a single prediction as the best. The CYK algorithm for finding the most likely parsing from the grammar is often used (Durbin *et al.*, 1998). An alternative to this has been chosen here. The goal of this method is to give the nested structure with the highest expected number of correct predictions. The appendix describes how this structure can be found. Notice that this approach removes some of the problems associated with using CYK with ambiguous grammars.

Once the best nested structure has been chosen, the reliability of the prediction in each position is evaluated relative to the model. Knowing which parts of the prediction are reliable is very important when using the prediction in further work.

Finding a single best structure might not always give all the information that would be useful. To give an overview of the prediction, a dot plot is produced as well. It is a square plot of pairing probabilities for all different pairs. Each probability is represented by the size of a dot in the appropriate position. Probabilities of not pairing are shown on the sides of the plot, see Figure 4. These calculations resemble the work by McCaskill (1990).

Making the obvious individual structure changes

Sometimes, a single sequence will have a slightly longer stem than its relatives. This can be incorporated in the prediction by extending it, if immediate neighbors can form a base pair. Another obvious change is to remove non-standard base pairs from individual sequences. This is done after the structure predictions given by this algorithm.

Implementation

In the implementation of this algorithm, speed and practical usefulness was important.

Calculations

The algorithm has two primary time consuming elements. With an alignment length of n, they are:

- Doing the inside-outside calculations, which has a running time of $O(n^3)$.

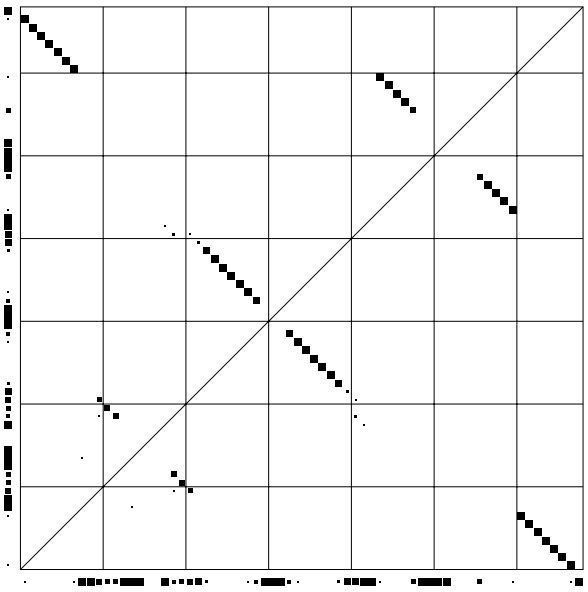


Figure 4: A dot plot made from GCA-tRNA sequences from rat, chicken, mouse and cow (Sprinzl *et al.*, 1998). The lower left corner represents the beginning of the alignment. Imagining the alignment laid out upwards, from here, and toward the right, the dots inside the square represents pairing probabilities between positions. The dots outside the square represents probabilities of not pairing. The tRNA structure turns up quite clearly.

- Calculating the column pair probabilities, which has a running time of $O(n^2)$.

It seems that the inside-outside calculations might consume the most time. It turns out, however, that the constant in the column pair probability calculation can be quite large. This part can be the most time consuming of the calculations, when short sequences or many aligned sequences are analysed. The reason is that the column pair probabilities are calculated by multiplying vectors with 16 by 16 matrices in a post-order traversal of the tree. For large trees, many such matrix multiplications has to be done.

The first step in saving time is to avoid doing calculations of column pairs with the same nucleotides more than once. This is effective for trees in the middle size range, because different columns often have the same nucleotides, e.g. all identical. When analysing larger trees, a few mutations have happened in most columns, thus often making them slightly different. To cope with this situation, partial column calculations can be stored for later use.

This has been implemented in the following way: when a calculation is made, the results are stored for the sub-tree below each node. When a new calculation is made, the results up to a given node are re-used if the sequence positions are the same for the sequences in the sub-tree below the node. This drastically reduces the computation time, since no identical matrix multiplications are done twice. It has the cost of a high memory usage and in some cases, a compromise is useful. This was implemented by only saving the results from a sub-tree if it was likely that it would be encountered again. Ideally, this likelihood for nucleotides in a sub-tree should be calculated using the same matrix as in the tree estimation described above. Probabilities for column pairs are closely approximated as the products of the individual columns, since two given columns has a probability of less than $1/n$ of forming a pair and non-pairing columns are treated as independent. Instead of using the matrix from the tree calculations, the matrix for single positions was used, since these calculations are already used in the algorithm.

Output

When a prediction is done, the web-server returns an e-mail to the user. This e-mail contains a link to a web page with the results. On the web page, the following is available in multiple formats:

- A summary of the input.
- The calculated tree.
- The predicted structure.
- Reliabilities of individual predictions.
- A dot plot.

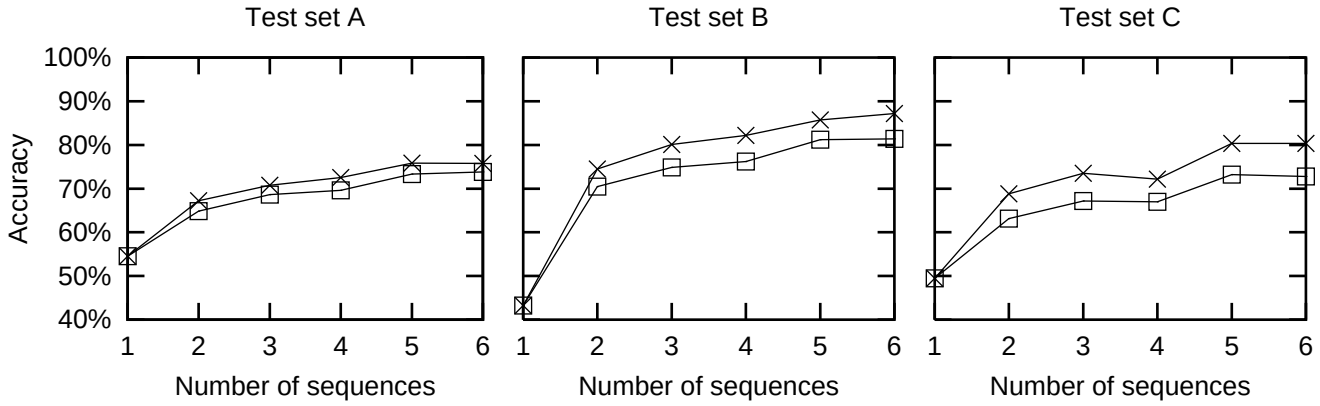


Figure 5: Accuracy as a function of the number of sequences used in the prediction. Crosses are from results using 'correct' alignments, while boxes are from CLUSTAL W alignments. Each point represents average results for either all possible combinations of the relevant number of sequences or 50 random combinations, whatever is the lowest number.

Results

For all predictions in this section, two versions were made. The first version was made using the alignment from the database of the sequences being analysed. This alignment was assumed to be 'correct'. In the second version, sequences from databases were aligned using the CLUSTAL W alignment program by Thompson *et al.* (1994) to imitate a realistic scenario of RNA structure prediction.

Test sets

A number of test sets were made, as described in Table 1. The sets A, B and C are used in evaluating the prediction accuracy as a function of the number of sequences used in the analysis. Test set D is used to show how prediction accuracy varies as a function of evolutionary distance.

Prediction accuracy

An evaluation of the prediction accuracy is shown in Figure 5. The accuracy was calculated as the percentage of positions for which the secondary structure were correctly predicted. For a pairing position to be counted as correct, the position of the predicted pair had to be correct. Sequences in the test sets used here have a maximum pairwise distance of 0.50 units in the Jukes-Cantor model (Jukes and Cantor, 1969). This means that the sequences are quite diverse, but still possible to align without too many errors.

Table 1: Test sets from Gorodkin *et al.* (2001) and Knudsen *et al.* (2001). Sets A, B and C were chosen, so that no pairwise distances within each set is more than 0.5 units in the Jukes-Cantor model (Jukes and Cantor, 1969). Set D was chosen as all unique sequences of length greater than 250 in the eukaryotic SRP RNA database. No two sequences are identical within any of the sets. The average sequence length of each test set is written in parentheses.

Test Set A: 9 tmRNAs (363.8)
act.act., hae.inf., kle.pne., pas.mul., sal.par., sal.typ., she.put., vib.cho., yer.pes.

Test Set B: 13 Bacterial SRP RNAs (270.5)
bac.alc., bac.bre., bac.cer., bac.cir., bac.mac., bac.meg., bac.pol., bac.pum., bac.sph., bac.ste., bac.thu., bre.bre., clo.per.

Test Set C: 10 Eukaryotic SRP RNAs (300.9)
ory.sat., tri.ae-a, tri.ae-b, zea.ma-a, zea.ma-b, zea.ma-c, zea.ma-d, zea.ma-e, zea.ma-f, zea.ma-h

Test Set D: 51 Eukaryotic SRP RNAs (297.4)
ara.th-a, ara.th-b, cae.el-a, cae.el-b, cae.el-c, cae.el-d, can.spe., cin.hyb., dro.mel., fug.rub., hom.sa-a, hom.sa-b, hom.sa-c, hum.ja-a, hum.ja-b, hum.lu-a, hum.lu-b, hum.lu-c, hum.lu-d, lep.col., lyc.es-a, lyc.es-b, lyc.es-c, lyc.es-e, lyc.es-f, lyc.es-g, lyc.es-h, lyc.es-i, lyc.es-j, lyc.es-k, lyc.es-m, lyc.es-n, lyc.es-o, ory.sat., rat.rat., sch.pom., tet.ros., tet.the., tri.ae-a, tri.ae-b, try.br-a, try.br-b, xen.lae., yar.li-a, yar.li-b, zea.ma-a, zea.ma-b, zea.ma-c, zea.ma-d, zea.ma-e, zea.ma-f

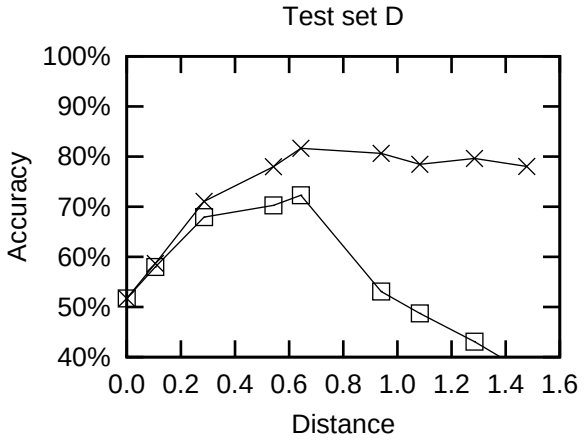


Figure 6: Accuracy as a function of pairwise distances between sequences being analysed. As in Figure 5, crosses are from results using 'correct' alignments, while boxes are from CLUSTAL W alignments. The pairs were grouped according to their Jukes-Cantor distances, in the intervals [0;0.2), [0.2;0.4), [0.4;0.6) etc. The points represent average results for 50 random sequence combinations from a specific range of distances. The x-value of a point is the average of the 50 distances.

As expected, prediction accuracy rises with the number of sequences used, as more covariance information becomes available. This shows that whenever related sequences are available, they should be used in the structure prediction, if alignable. It also shows that an accuracy of around 75% is obtainable with six sequences. The algorithm can cope with many more sequences, so even higher accuracies could be expected in this case.

Evolutionary distance effect

Two effects of evolutionary distance on prediction accuracy are that large distances implies much covariation information, but it also means that sequences are difficult to align.

These effects are illustrated in Figure 6. When using 'correct' alignments, the accuracy rises with distance, as the evolutionary information increases. The accuracy levels off at around 80%, which seems to be the maximum obtainable average, for two sequences of this type, using this algorithm. The graph made from the CLUSTAL W alignments show how accuracy increases until an evolutionary distance of around 0.6. After this, the accuracy drops due to alignment errors. At a distance of around 0.9, the quality of alignments become so low, that the structures might as well be predicted individually.

Speed

The result of all this is an algorithm that is fast enough to do a structure prediction on a 25 sequence alignment of length 600 in around five minutes on a fast desktop computer. This makes a web server version of the program feasible.

Comparison with other methods

The accuracies obtained by using MFOLD (Zuker *et al.*, 1999; Mathews *et al.*, 1999) with standard parameters are 47.2% for test set A, 68.8% for test set B, 57.0% for test set C and 69.3% for test set D. These results are generally better than for the algorithm described here, using single sequences. When using multiple sequences, accuracies from this algorithm generally become better than for the MFOLD results.

The RNAGA method by Chen *et al.* (2000) also works on a limited set of sequences, rather than single sequences, but has no web-site, on which structures are easily predicted. The method is relatively slow, but seems very promising.

Methods like COVE (Eddy and Durbin, 1994) and Maximum Weighted Matching (Tabaska *et al.*, 1998) depend on more than a few sequences for reliable results (Knudsen and Hein, 1999).

Discussion

Some aspects of this method remain to be explored, as described by Knudsen and Hein (1999). These include: base stacking interactions, a grammar more closely describing real RNA structures and other models for base-pair evolution.

If base-stacking interactions and a better grammar is incorporated in the algorithm described here, the prediction accuracies should become close to the MFOLD results for single sequences, since the methods resemble each other closely, in that situation. For multiple sequences, this algorithm should still be able to perform even better.

As emphasized by this work, an important aspect of RNA structure prediction is the alignment problem. In methods that depend on a sequence alignment, the success of the method is closely linked to the quality of the alignment. Some work has been done in the field of RNA structural sequence alignment (e.g., Gorodkin *et al.*, 1997). The RNAGA method by Chen *et al.* (2000) predicts consensus structures without trying to align the sequences, which might be a useful approach to avoiding the alignment problem.

With this, a new and efficient method of RNA secondary structure prediction is made available via the World Wide Web.

Acknowledgments

I would like to thank Christian N. S. Pedersen for his great help in answering questions about UNIX and C programming. I would also like to acknowledge

the help of Ebbe S. Andersen in making this method practically useful and in illustrating the web server. Finally, I thank Jakob S. Pedersen for helpful comments on the manuscript.

Appendix: Finding the best structure

The best structure for the prediction is here defined as the nested structure with the highest expected number of correctly predicted positions. This structure is found by a dynamical programming approach similar to the CYK algorithm, using pairing probabilities which can be calculated from the inside-outside variables. Denote the sub-alignment from position i to j as $C_{i,j}$ (i.e. columns i through j in the alignment) and let C_i represents column i by itself. Write the inside and outside variables as:

$$\begin{array}{ll} \text{Inside:} & e_{i,j}(N) = P(N \rightarrow C_{i,j}) \\ \text{Outside:} & f_{i,j}(N) = P(S \rightarrow C_{1,i-1} N C_{j+1,n}) \end{array}$$

where N is a non-terminal, S is the starting non-terminal, and n is the alignment length. Assume that all pairs are formed by production rules of the type: $N_1 \rightarrow dN_2d$, where the d's represent a pair. The probability (under the model) that position i pairs with j is:

$$P_d(i,j) = \sum_{N_1,N_2} f_{i,j}(N_1) P(N_1 \rightarrow C_i N_2 C_j) e_{i+1,j-1}(N_2)$$

The sum is effectively only over the rules forming pairs. The probability that position i does not form a pair is:

$$P_s(i) = 1 - \sum_{j \neq i} P_d(i,j)$$

Define $E_{i,j}$ as the maximum possible number of expected correct non-nested predictions in the window from position i to position j in the alignment. Setting $E_{i,j} = 0$, for $i > j$, the $E_{i,j}$'s can be calculated for $i \leq j$:

$$E_{i,j} = \max \begin{cases} E_{i+1,j} + P_s(i) & \text{(Unpaired)} \\ E_{i+1,j-1} + P_d(i,j) & \text{(Paired)} \\ E_{i,j-1} + P_s(j) & \text{(Unpaired)} \\ E_{i,k} + E_{k+1,j} \quad i \leq k < j & \text{(Bifurcation)} \end{cases}$$

The $E_{i,j}$'s can now be calculated in order of increasing $j - i$. Actually, the unpaired part of the recursion could be done by the bifurcation step, by initializing the $E_{i,i}$'s to $P_s(i)$, but the above is easier to interpret. $E_{1,n}$ is the maximum expected number of correctly predicted positions in the alignment. The prediction that gives this maximum can be found by backtracking the above recursion.

Some non-nested structures may exist, which has a higher expected number of correctly predicted positions than the nested one found above. This could

happen even though the SCFG used to find pairing probabilities does not take non-nested structures into account. This will be a rare occurrence and would probably not give any practical information. The structure could be found by a maximal weighted matching approach on the pairing probabilities. The Maximal Weighted Matching algorithm is described by Tabaska *et al.* (1998).

References

Bishop, M. J. and Thompson, E. A. (1986) Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* **190** (2), 159–165.

Chen, J.-H., Len, S.-Y. and Maizel, J. V. (2000) Prediction of common secondary structures of RNAs: a genetic algorithm approach. *Nucleic Acids Res.* **28** (4), 991–999.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge.

Eddy, S. R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22** (11), 2079–2088.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17** (6), 368–376.

Gorodkin, J., Heyer, L. J. and Stormo, G. D. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.* **25** (18), 3724–3732.

Gorodkin, J., Knudsen, B., Zwieb, C. and Samuelsson, T. (2001) SRPDB (signal recognition particle database). *Nucleic Acids Res.* **29** (1), 169–170.

Hein, J., Wiuf, C., Knudsen, B., Møller, M. and Wibling, G. (2000) Statistical alignment: Computational properties, homology testing and goodnesss-of-fit. *J. Mol. Biol.* **302** (1), 265–279.

Jukes, T. H. and Cantor, C. R. (1969) Evolution of protein molecules. In Munro, H. N. (ed), *Mammalian Protein Metabolism.* Academic Press., New York, pp. 21–123.

Knudsen, B. and Hein, J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics,* **15** (6), 446–454.

Knudsen, B., Wower, J., Zwieb, C. and Gorodkin, J. (2001) tmRDB (tmRNA database). *Nucleic Acids Res.* **29** (1), 171–172.

Lari, K. and Young, S. J. (1990) The estimation of stochastic context-free grammars using the inside-outside algorithm. *Comput. Speech Lang.* **4** (1), 35–56.

Mathews, D. H., Sabina, J., Zuker, M. and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improve prediction of RNA secondary structure. *J. Mol. Biol.* **288** (5), 911–940.

McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers,* **29**, 1105–1119.

Nussinov, R., Pieczenik, G., Griggs, J. R. and Kleitman, D. J. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math.* **35** (1), 68–82.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4** (4), 406–425.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. and Steinberg, S. (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26** (1), 148–153.

Tabaska, J. E., Cary, R. B., Gabow, H. N. and Stormo, G. D. (1998) An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics,* **14** (8), 691–699.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22** (22), 4673–4680.

Thorne, J. L., Goldman, N. and Jones, D. T. (1996) Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13** (5), 666–673.

Thorne, J. L., Kishino, H. and Felsenstein, J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33** (2), 114–124.

Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Mueller, P., Mathews, D. H. and Zuker, M. (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl. Acad. Sci. USA,* **91** (20), 9218–9222.

Zuker, M. (1989) Computer prediction of RNA structure. *Methods in Enzymology,* **180**, 262–288.

Zuker, M., Mathews, D. H. and Turner, D. H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In Barciszewski, J. and Clark, B. F. C. (eds), *RNA Biochemistry and Biotechnology.* Kluwer Academic Publishers, Dordrecht, the Netherlands, pp. 11–43.

Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res.* **9** (1), 133–148.

Chapter 4

Statistical Alignment: Computational Properties, Homology Testing and Goodnesss-of-Fit

Article by Jotun Hein, Carsten Wiuf, Bjarne Knudsen, Morten Møller, and Gustav Wibling.

Appeared in *J. Mol. Biol.*, **302** (1), 265–279, 2000.

doi:10.1006/jmbi.2000.4061 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 302, 265–279

JMB

Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit

J. Hein[1]*, C. Wiuf[2], B. Knudsen[1], M. B. Møller[3] and G. Wibling[3]

[1]*Department of Genetics and Ecology The Institute of Biological Science, University of Aarhus, Building 540, Ny Munkegade, 8000 Århus C Denmark*

[2]*Department of Statistics University of Oxford, 1 South Parks Road, Oxford, OX1 3TG UK*

[3]*The Institute of Computer Science, University of Aarhus Building 550, Ny Munkegade 8000 Århus C, Denmark*

**Corresponding author*

The model of insertions and deletions in biological sequences, first formulated by Thorne, Kishino, and Felsenstein in 1991 (the TKF91 model), provides a basis for performing alignment within a statistical framework. Here we investigate this model.

Firstly, we show how to accelerate the statistical alignment algorithms several orders of magnitude. The main innovations are to confine likelihood calculations to a band close to the similarity based alignment, to get good initial guesses of the evolutionary parameters and to apply an efficient numerical optimisation algorithm for finding the maximum likelihood estimate. In addition, the recursions originally presented by Thorne, Kishino and Felsenstein can be simplified. Two proteins, about 1500 amino acids long, can be analysed with this method in less than five seconds on a fast desktop computer, which makes this method practical for actual data analysis.

Secondly, we propose a new homology test based on this model, where homology means that an ancestor to a sequence pair can be found finitely far back in time. This test has statistical advantages relative to the traditional shuffle test for proteins.

Finally, we describe a goodness-of-fit test, that allows testing the proposed insertion-deletion (indel) process inherent to this model and find that real sequences (here globins) probably experience indels longer than one, contrary to what is assumed by the model.

Keywords: statistical alignment; homology testing; goodness-of-fit

Introduction

Statistically well founded methods have become increasingly used over the last decade in the analysis of biological sequences. This has not been the case in the alignment of sequences, partly because of the general conception that the statistical approach to alignment is computationally too slow and partly due to the lack of user-friendly programs. Often, the sequences are aligned using parsimony or similarity based methods (optimisation alignments). The alignment is subsequently treated as a series of columns that are independent realizations of a substitution process on a phylogeny that is to be estimated. It is an inconsistent approach to first use parsimony/similarity and then halfway in the analysis switch to a statistical approach. In addition, the alignment created by parsimony/similarity can create unknown biases in the estimation of substitutional parameters, as these procedures will align to create as much identity within each column as possible.

The first attempt to do statistical alignment was done by Bishop & Thompson (1986), with approximate likelihood calculations. Thorne *et al*. (1991) introduced an exact method (TKF91). In this framework, there will not be one alignment, but all possible alignments will contribute to the likelihood of the two observed sequences. Should one alignment be highlighted, it could be the alignment that contributed the most to the likelihood. Alignments can have runs of gap signs, which in a parsimony/similarity setting would be interpreted as a longer insertion/deletion (indel), but here it would be the consequence of several neighbouring indels of single nucleotides. Most optimisation alignment methods interpret runs of gap signs as a single event, even when they may be the result of multiple independent insertions and deletions. Nevertheless, indels longer than one nucleotide or amino acid must occur biologically and the TKF91 model does not allow for that. Thorne *et al*. (1992) tried to incorporate this in the model by letting insertions

E-mail address of the corresponding author: jotun.hein@biology.au.dk

0022-2836/00/010265–15 $35.00/0

and deletions involve fragments that each had a geometrical distribution. For computational reasons each fragment would have to be treated as an unbreakable unit, which is not a biologically well founded assumption.

An alternative approach to statistical alignment has been taken by Allison & Wallace (1994), Zhu *et al*. (1998), and Mitchison (1999). These are Bayesian approaches and also differ from the TKF91 model in not being based on an evolutionary process, but on a probability measure on alignments directly.

Pairwise sequence alignment, homology testing and multiple alignment has great importance in sequence analysis and there is an increasing awareness of the advantages of statistical approaches within the bioinformatics community. Therefore, statistical alignment and its ramifications deserve to be pursued with much greater intensity.

Theory

The TKF91 statistical model of DNA evolution is a continuous time model with a state space consisting of all sequences over an alphabet of nucleotides (yielding DNA sequences) or amino acids (yielding proteins) that includes the empty sequence. If we, for any possible sequence, can define the waiting time to the first event (insertion, deletion or change of single element) to occur and the probability of all the possible events, the model has been characterised.

Modelling substitutions

Since the novelty of the TKF91 model is in the modelling of the indel process, the substitution process will receive little attention here. Almost all substitution models are continuous time Markov models on the state space of nucleotides or amino acids. To define such a model, the rate matrix, Q, must be specified. This matrix is 4 by 4 for nucleotides and 20 by 20 for amino acids. Off diagonal elements are non-negative and the sum of each row is zero. This matrix describes the intensity of different substitution events over an infinitesimal time period. The transition probabilities over a longer time interval, t, can be obtained by:

$$P(t) = e^{tQ} = \sum_{k \geqslant 0} \frac{(tQ)^k}{k!}.$$

$P_{i,j}(t)$ refers to the probability that i has changed to j after a time t. Normally, it is also assumed that the process is time reversible, i.e. that $\pi_i P_{i,j}(t) = \pi_j P_{j,i}(t)$, where the π_i terms are the equilibrium frequencies of the nucleotides/amino acids in the process. In this case, the evolutionary process can be viewed from any time perspective, the total probabilities involved will be the same. This has computational advantages, since the evolutionary history can be rooted anywhere.

Since the time, t, and the rate matrix, Q, always appear together as a product in these calculations, it is not possible to estimate them individually. It is often convenient to scale them, so one event is expected per unit time per position in the equilibrium process. This is equivalent to placing the restriction $\Sigma_i \pi_i Q_{ii} = -1$ on the rate matrix.

The main difference between substitution processes on nucleotides and on amino acids, stems from the larger set of amino acids. In principle a 20 by 20 rate matrix with many parameters could be inferred, but there are not such clear distinctions as the transversion/transition bias in the case of nucleotides. This means that more crude ways of choosing a Q matrix, than maximum likelihood estimates of individual entries are used, since there are too many parameters to be estimated. Dayhoff *et al*. (1978) pioneered the use of Q-matrices obtained from counting mutations in comparisons of similar protein sequences. Due to the inability to distinguish where A has mutated to B or *vice versa*, such matrices will be symmetric and give rise to equilibrium distributions where all the amino acids are equally likely. This is in conflict with the frequencies that can be observed in real sequences. Fortunately, it is possible to modify a symmetric matrix, so it will give rise to the observed frequencies of the 20 amino acids. In addition, the resulting matrix will yield a mutation process that is time reversible. The rate matrix used here was obtained from Ziheng Yang (personal communication).

The TKF91 model of the indel process

The basic model

The statistical model of sequence evolution incorporating insertions and deletions can be viewed as a Markov model with all sequences as possible states. The indel part of this model can be illustrated by the use of links connecting the letters (nucleotides or amino acids) of the sequences. Each letter has a mortal link associated to it, on its right. The left end of the sequence has an immortal link. Consider an example, the DNA sequence *AGG*:

$$\bullet\, A \star G \star G \star$$

Mortal links are symbolised by ★, while immortal links are symbolised by •. Links can give birth to new links to their right. Along with the birth of such a link, comes a letter drawn from the equilibrium distribution. The mortal links can, as the name suggests, also die. When a link dies, it takes its letter (to the left) with it. The transition in the Markov model, when the mortal link between the two G residues gives birth to a new link and a C, is shown here:

$$\bullet\, A \star G \star G \star \rightarrow \bullet\, A \star G \star C \star G \star$$

The new link is the one to the right of the C. If the first mortal link dies, it looks like this:

$$\bullet\ A\ \star\ G\ \star\ G\ \star \rightarrow \bullet\ G\ \star\ G\ \star$$

This process can be visualised in the following way:

$\bullet$	A	$\star$	G	$\star$	G	$\star$
I_0	S_1	I_1/D_1	S_2	I_2/D_2	S_3	I_3/D_3

I, D and S terms are all independent processes that describe insertions, deletions and substitutions, respectively. For each of these, there will be an exponential waiting time for an event to occur. Whichever fires first, determines the next event.

I_i has intensity parameter λ. When I fires first, a nucleotide will be inserted, according to the equilibrium frequency of the substitutional process, with a mortal link on its left side. The newborn nucleotide and associated mortal link will be inserted to the right of the I, that fired.

D_i has intensity parameter μ. When a D fires first, the link and its nucleotide (to the left) will be removed. The immortal link will never be deleted, since it is not associated with a deletion process. The deletion rate has to be bigger than the insertion rate ($\mu > \lambda$), otherwise the sequence length would grow towards infinity. This process will have a stationary distribution on sequences:

$$P(s) = \gamma_l \pi_A^{\#A} \pi_C^{\#C} \pi_G^{\#G} \pi_T^{\#T},$$

where, for $l \geqslant 0$:

$$\gamma_l = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^l.$$

l is the length of the sequence and #A (#C, #G, #T) indicates the number of A residues (C, G, T residues) in the sequence. This indel process is time reversible, thus, if the substitution process is also time reversible, the process on full length sequences is reversible. This is not a pure birth-death process because of the immortal link, but can be viewed as a birth-death process with immigration.

Evolution over a time period

Above, it was described how a sequence would evolve in an infinitesimal time interval. Thorne *et al*. (1991) also described the transition probabilities for a fixed time interval (Figure 1). Due to the independence between links, it is sufficient to describe what happens to a single link. There is an insertion process associated with the immortal link, with the transition probability $p_n''(t)$, the probability that the immortal link has left itself and $n-1$ mortal descendants after time t. There is an indel process associated with the mortal links, with two sets of transition probabilities, depending on whether the link has survived or not. The probability is $p_n(t)$ if the link has survived and left n (mortal) descendants, including itself. If the link has not survived, the probability is $p_n'(t)$, again with n being equal to the number if descendants. This last distinction between survival and non-survival is necessary, since only in the first case will a nucleotide exist both at time zero and time t and the probability of going from one nucleotide to n nucleotides will involve a substitutional probability. The surviving children of a nucleotide will be to the right of the parent nucleotide. In the following, the evolutionary process parameters, including t, will often be suppressed.

The functions p_k, p_k' and p_k'' are modified geometric distributions. The function describing immortal link (p_k'') is the geometric function shifted so that it starts in one instead of zero. The function describing the case where a mortal link survives (p_k) is again shifted to start in one and every position has been multiplied with the probability of survival ($e^{-\mu t}$). The probability of the nucleotide not surviving is $1 - e^{-\mu t}$. In this case, there will be a probability for having zero surviving offspring ($\mu\beta(t)$), with:

$$\beta(t) = \frac{1 - e^{(\lambda-\mu)t}}{\mu - \lambda e^{(\lambda-\mu)t}}$$

and the remaining distribution (no survival but surviving descendants) has again the same geo-

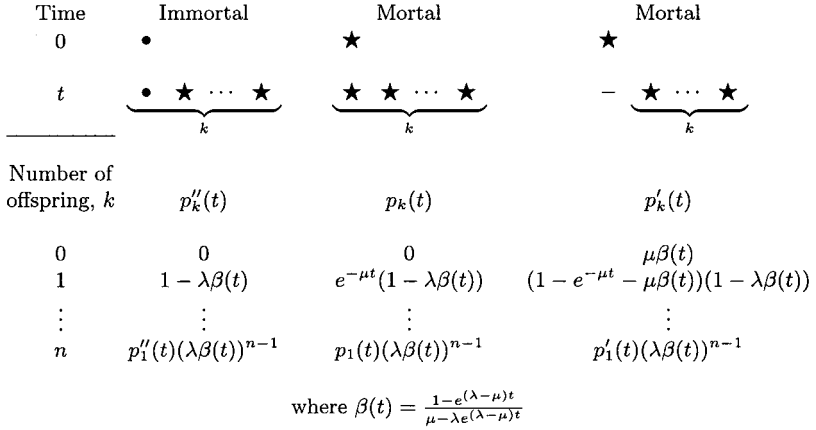


Figure 1. The probability distributions of different link configurations after a period t. The fate of a mortal link has to be split into two cases to accommodate the possibility of substitutional evolution. Since $\lambda < \mu$, $\beta(t)$ is always smaller than one.

metric tail, but with total mass adjusted to give total probability of one to all possible fates of a mortal link.

These transition probabilities of substitutions and of the fates of mortal and immortal links, allow a dynamic programming algorithm calculating $P(s^{(1)}, s^{(2)})$ to be formulated, where $s^{(1)}$ and $s^{(2)}$ are the two complete sequences.

The probability of the sequences and one specific alignment is easily described in terms of the substitution probabilities and p functions. Regard the following example:

$$\begin{array}{ccccccc} \bullet & A & \star & T & \star & - & \\ \bullet & C & \star & T & \star & G & \star \end{array}$$

Here, $P(s^{(1)}, s^{(2)}, \text{alignment}) = (p_1'')(\pi_{Ap1}P_{AC})(\pi_T P_{TTp2}\pi_G)$. Each parenthesis describes the probability of a link, the associated nucleotide (in the case of mortal links) and their fates. The first parenthesis is the probability that the immortal link survives with one descendant — itself. The last parenthesis says that a T was chosen, the T evolved into a T, the link associated to the T had two descendants including itself and the extra descendant is G. To calculate the probability of two sequences without conditioning on the alignment, it is necessary to sum over all alignments weighted with their probabilities according to the TKF91 process.

A simpler recursion

In the following, $s_i^{(1)}$ is the prefix of length i in $s^{(1)}$ and $s_j^{(2)}$ is the prefix of length j in $s^{(2)}$. $s^{(k)}[i]$ refers to the ith nucleotide in the kth sequence. $\pi_{s^{(k)}[i:j]}$ is the probability of the elements from i to j in sequence k in the equilibrium distribution of the substitution process. The probability of the complete sequences, $P(s^{(1)}, s^{(2)})$, can be written as $P(s^{(1)})P(s^{(2)}|s^{(1)})$. The first factor is straightforward to calculate and we will focus on calculating the second only. The reformulation of the algorithm below represents a simplification and acceleration relative to the original TKF91 formulation. Firstly, we only make a recursion for $P(s^{(2)}|s^{(1)})$, while they incorporated the probability $P(s^{(1)})$. Secondly, we only need one or two quantities per entry (i,j) while they needed three quantities. The resulting recursion (5) is as simple as the most basic optimisation alignment algorithm. The basic recursions are summarised in Table 1.

It is possible to decompose the probability $P(s_j^{(2)}|s_i^{(1)})$ by partitioning the conversion of $s_i^{(1)}$ to $s_j^{(2)}$, into the fate of $s_{i-1}^{(1)}$ and the fate of $s^{(1)}[i]$, since these fates are independent (Figure 2).

This example illustrates why it is necessary to distinguish whether a nucleotide survives or not. Only in the former case has substitutional evolution been observed. If the first sequence is empty, i is zero and the immortal link must have evolved into $s_j^{(2)}$, which has probability $p_j''\pi_{s^{(2)}[1:j]}$.

The above illustration can be summarised in following recursion:

$$P(s_j^{(2)}|s_i^{(1)}) = p_0' P(s_j^{(2)}|s_{i-1}^{(1)}) + \sum_{1\leqslant k\leqslant j} P(s_{j-k}^{(2)}|s_{i-1}^{(1)}) \times (p_k P_{s^{(2)}[i],s^{(2)}[j-k+1]}\pi_{s^{(2)}[j-k+2:j]} + p_k'\pi_{s^{(2)}[j-k+1:j]}) \quad (1)$$

$$P(s_j^{(2)}|s_0^{(1)}) = p_j''\pi_{s^{(2)}[1:j]} \quad (2)$$

This recursion allows an $O(l^3)$ (l denoting average sequence length) algorithm to be formulated for calculating $P(s^{(1)}, s^{(2)})$.

Due to the geometric tails of the p functions, this formulation can be changed, resulting in an $O(l^2)$ algorithm. The trick applied here is highly reminiscent of the method used by Gotoh (1982) in reducing the computational complexity of an optimisation alignment algorithm from $O(l^3)$ to $O(l^2)$.

Define $R_{i,j} = P(s_j^{(2)}, s^{(2)}[j]$ is a descendant of $s^{(1)}[i]|s_i^{(1)})$. This will be the sum on the right side of (1). The first recursion above can now be written as:

$$R_{i,j} = (p_1 P_{s^{(1)}[i],s^{(2)}[j]} + p_1'\pi_{s^{(2)}[j]})P(s_{j-1}^{(2)}|s_{i-1}^{(1)}) + \lambda\beta\pi_{s^{(2)}[j]}R_{i,j-1} \quad (3)$$

$$P(s_j^{(2)}|s_i^{(2)}) = R_{i,j} + p_0' P(s_j^{(2)}|s_{i-1}^{(1)}) \quad (4)$$

The functions $p_1 P_{s^{(1)}[i],s[j]}^{(2)} + p_1'\pi_{s^{(2)}[j]}$ and $\lambda\beta\pi_{s^{(2)}[j]}$ are functions in two sequence elements that can be tabulated. Equation (4) asserts that either $s^{(2)}[j]$ is a descendant of $s^{(1)}[i]$ or it is not. Recursion (3) can be verified using the recursive relationships $p'1 = \lambda\beta p_k'$, $p_{k+1} = \lambda\beta p_k$, for $k \geqslant 1$, and $\pi_{s^{(n)}[i:j]} = \pi_{s^{(n)}[i:j-1]}\pi_{s^{(n)}[j]}$. $R_{i,j}$ is subject to the initial condition $R_{i,j} = 0$ if i or j is zero.

Insertion of (4) into (3) and simplification yields:

$$P(s_j^{(2)}|s_i^{(1)}) = P_0' P(s_j^{(2)}|s_{i-1}^{(1)}) + \lambda\beta\pi_{s^{(2)}[j]}P(s_{j-1}^{(2)}|s_i^{(2)}) + (p_1 P_{s^{(1)}[i],s^{(2)}[j]} + p_1'\pi_{s^{(2)}[j]} - \lambda\beta\pi_{s^{(2)}[j]}p_0')P(s_{j-1}^{(2)}|s_{i-1}^{(1)}) \quad (5)$$

Again $p_1 P_{s^{(1)}[i],s[j]}^{(2)} + p_1'\pi_{s^{(2)}[j]} - \lambda\beta\pi_{s^{(2)}[j]}$ p_0' and $\lambda\beta\pi_{s^{(2)}[j]}$ can be tabulated, simplifying and accelerating the recursion.

Recursion (5) allows an efficient summation over all alignments of $s^{(1)}$ with $s^{(2)}$. In this context there are two additional quantities of interest, as follows below. To cope with these, it is advantageous to continue with recursions (3) and (4).

First, it is of interest to find the alignment that contributes the most to $P(s^{(2)}|s^{(1)})$, and which, given a set of parameters, will be the most probable. Secondly, it is of interest to generate alignments in

A) Parent nucleotide survives

Number of offspring from $s^{(1)}[i]$	Prefix	Tail	Substitutional evolution	Insertion
1	$P(s^{(2)}_{j-1}\|s^{(1)}_{i-1})$	p_1	$P_{s^{(1)}[i],s^{(2)}[j]}$	1
2	$P(s^{(2)}_{j-2}\|s^{(1)}_{i-1})$	p_2	$P_{s^{(1)}[i],s^{(2)}[j-1]}$	$\pi_{s^{(2)}[j]}$
⋮	⋮	⋮	⋮	⋮
j	$P(\emptyset\|s^{(1)}_{i-1})$	p_j	$P_{s^{(1)}[i],s^{(2)}[1]}$	$\pi_{s^{(2)}[2:j]}$

The product of the factors in the second row corresponds to the probability of the alignment:

$$\begin{array}{ccccccc} s^{(1)}_{i-1} & \bigstar & s^{(1)}[i] & \bigstar & - & & \\ s^{(1)}_{j-2} & \bigstar & s^{(2)}[j-1] & \bigstar & s^{(2)}[j] & \bigstar & \end{array}$$

B) Parent nucleotide dies

Number of offspring from $s^{(1)}[i]$	Prefix	Tail	Substitutional evolution
0	$P(s^{(2)}_{j}\|s^{(1)}_{i-1})$	p'_0	1
1	$P(s^{(2)}_{j-1}\|s^{(1)}_{i-1})$	p'_1	$\pi_{s^{(2)}[j]}$
2	$P(s^{(2)}_{j-2}\|s^{(1)}_{i-1})$	p'_2	$\pi_{s^{(2)}[j-1:j]}$
⋮	⋮	⋮	⋮
j	$P(\emptyset\|s^{(1)}_{i-1})$	p'_j	$\pi_{s^{(2)}[1:j]}$

The product of the factors in the third row corresponds to the probability of the alignment:

$$\begin{array}{ccccccccc} s^{(1)}_{i-1} & \bigstar & s^{(1)}[i] & \bigstar & - & & - & & \\ s^{(1)}_{j-2} & \bigstar & - & & s^{(2)}[j-1] & \bigstar & s^{(2)}[j] & \bigstar & \end{array}$$

Figure 2. The independence of the indel process and the substitution process allows the two to be combined easily. (a) The possible fates of $s^{(1)}[i]$ in $s^{(2)}_j$, given that it survives. (b) The possible fates of $s^{(1)}[i]$ in $s^{(2)}_j$, given that it dies.

proportion to their probability. Methods for this are shown in an Appendix I.

Results

The method is illustrated on human α globin and β globin, that are 141 and 146 amino acids long, respectively (Figure 3). The expected length of a sequence in the equilibrium process is 143.5024, very close to the average length of the two proteins. The asymptotic variances and covariances of the parameters can be obtained from the inverse of the matrix of second derivatives of the likelihood with respect to the parameters (not shown) (Edwards, 1972).

The expected number of events in the evolution from α globin to β globin is difficult to compute, as computation of the expected number of events for any small alignment block is difficult. Take for example an amino acid aligned with another. The expected number of events would involve summing over cases where amino acids were inserted, experienced mutations and were then deleted.

However, the analogous quantity for a randomly chosen sequence in the equilibrium distribution for the estimated parameters can be calculated. Let π_i and Q_{ij} have the same meanings as in the section on substitutional models, except that the i and j terms refer to entire sequences instead of nucleotides and amino acids. For estimated time and rates, the expected number of events is $-t\ \Sigma_{i \in S}\ \pi_i Q_{ii}$. The summation is here over the complete sequence space, and π_i is the probability of i in the stationary distribution on the complete sequences, not single elements. $-Q_{ii}$ is the rate of events leaving i. The sum is equal to:

$$\frac{\mu}{\mu-\lambda}\lambda + (\mu+s)\frac{\lambda}{\mu-\lambda} = (2\mu+s)\frac{\lambda}{\mu-\lambda}.$$

This is identified as the expected number of I terms times their intensity parameters, plus the expected number of D terms times their intensity parameters, plus the expected number of S terms times their intensity parameters. It is intuitively reasonable that the expected number of insertions must equal the expected number of deletions. For the maximum likelihood parameters of this example, the expected number of insertions and deletions in the equilibrium process is 10.74. The expected number of substitution events (in the equilibrium process) is 131.59, slightly less than one event per position. The maximally contributing

Table 1. Summary of recursions

Elementary parsimony algorithm:

$$D_{i,j} = \min \begin{cases} D_{i-1,j} + g \\ D_{i-1,j-1} + d(s^{(1)}[i],\ s^{(2)}[j]) \\ D_{i,j-1} + g \end{cases}$$

Original TKF91 recursion:

$$L^0(s_i^{(1)}, s_j^{(2)}) = \frac{\lambda}{\mu}\pi_{s^{(1)}[i]}p_0' \sum_{k=0}^{2} L^k(s_{i-1}^{(1)}, s_j^{(2)})$$

$$L^1(s_i^{(1)}, s_j^{(2)}) = \frac{\lambda}{\mu}\pi_{s^{(1)}[i]}(P_{s^{(1)}[i],s^{(2)}[j]}p_1 + \pi_{s^{(2)}[j]}p_1') \sum_{k=0}^{2} L^k(s_{i-1}^{(1)}, s_{j-1}^{(2)})$$

$$L^2(s_i^{(1)}, s_j^{(2)}) = \pi_{s^{(2)}[j]}\lambda\beta \sum_{k=0}^{2} L^k(s_i^{(1)}, s_{j-1}^{(2)})$$

Simpler recursion:

$$P(s_j^{(2)}|s_i^{(1)}) = p_0' P(s_j^{(2)}|s_{i-1}^{(1)}) + \lambda\beta\pi_{s^{(2)}[j]} P(s_{j-1}^{(2)}|s_i^{(1)}) + g(s^{(1)}[i], s[j]) P(s_{j-1}^{(2)}|s_{i-1}^{(1)})$$

with

$$g(i,j) = p_1 P_{s^{(1)}[i],s^{(2)}[j]+(p_1'-\lambda\beta)\pi_{s^{(2)}[j]}}$$

The first recursion is the simplest parsimony algorithm. $D_{i,j}$ is the distance between $s_i^{(1)}$ and $s_j^{(2)}$, $d(\ ,\ )$ is a distance function on single elements and g is the gap penalty for a single element. All the involved quantities are integers. The second set of recursions are from the original TKF91 paper. The last recursions are from the present paper.

alignment has nine gap signs, 104 mismatches and 37 matches. Just inspecting this alignment for events would probably underestimate the number of these events in the true history of the sequences. Obviously, indels are much rarer than substitutions.

It is now possible to evaluate whether the length of sequence $s^{(2)}$, has evolved more or less than expected. The TKF91 model assumes that insertion and deletion rates are independent of sequence length. Figure 4 illustrates this distribution and β globin, for instance, is not extreme. If t goes toward

TKF91 analysis of α and β globins

$-ln(L_{\text{tot}}) = 730.428$
$L_{\max}/L_{\text{tot}} = 0.0057$

	Estimate
λ	0.03718
μ	0.03744
s	0.91618

	Expected	Observed[a]
Length	143.5024	141/146
Indels	10.74	9
Substitutions	131.59	104

[a] In maximally contributing alignment

Maximally contributing alignment to the likelihood:

```
V-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS--H---GSAQVKGHGKKVADALT
VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFS

NAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
DGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
```

Figure 3. α Globin and β globin analysis using the TKF91 model.

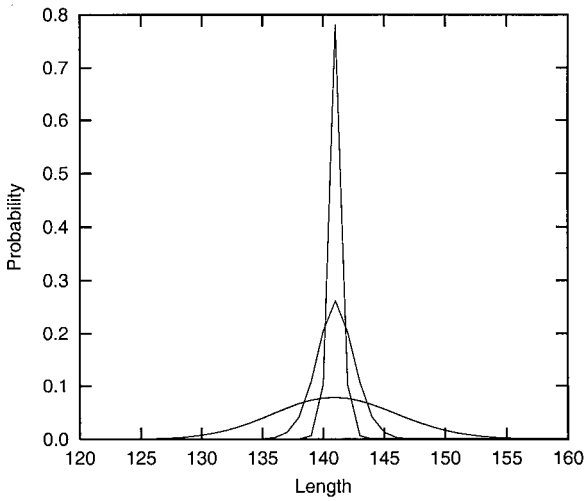


Figure 4. The length distribution of a protein that evolves from α globin with the estimated parameters of globin evolution (for 20, 200, and 2000 million years). For derivations see Appendeses. For instance, the length of the β globin (146) is not very extreme in the distribution of distance lengths, if starting with 141 amino acids and evolving for 800 million years. This is twice a resonable guess of the time to the most recent common ancestor of α and β globin.

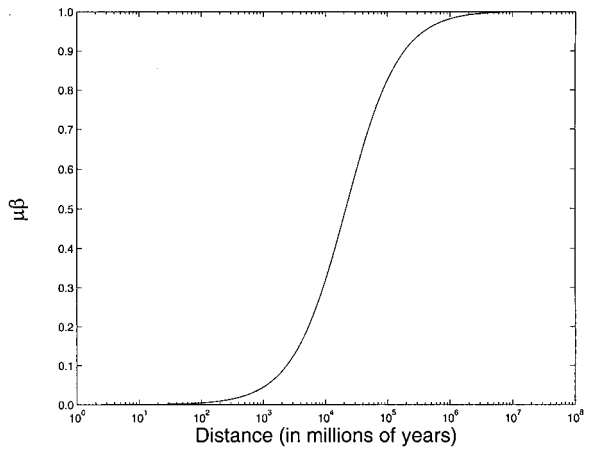


Figure 5. The β function plotted with parametersestimated from the α, β globin analysis. At the time when μ β is 0.5, the descendants of the immortal link are expected to contribute half of the complete sequence. This time is seen to bearound 20 billion years. The time taken for it to contribute 5% is around one billion years. The effect of the immortal link on realsequences is vanishing over realistic time periods.

infinity, this distribution will go towards the equilibrium length distribution for this process, but very slowly. Specifically, the mortals and their descendants will go extinct, while the descendants of the immortal link will be dominating the complete sequence. This means that $p''_k(t)$ is a geometric distribution with parameter $\beta\lambda$, when $t \to \infty$. As shown in Figure 4, even if the most recent common ancestor were 2000 million years back in time, the length is still distributed as a bell around the initial length. If a sequence is chosen from the equilibrium distribution and observed for a period of time, the descendants of the immortal links will be expected to constitute $\mu\beta(t)$ of the complete sequence. $\mu\beta(t)$ is plotted in Figure 5 and it is clear that it converges very, very slowly to one.

Computational results

Computationally, maximum likelihood alignment is inherently more expensive than parsimony/similarity. There are three areas of importance to the time of performing a likelihood alignment: the number of entries in the matrix needing to be calculated, the number of evaluations needed to find the maximum likelihood estimate and the time used in calculating the basic recursion

Matrix entries necessary

Figure 6 illustrates the alignment path of α globin and β globin including boundaries defined by suboptimal alignments. (In these investigations PAM250 was used for similarity alignments and a gap penalty cost of 4.5 was used per amino acid.) For a given ε, a boundary corresponding to the suboptimal paths with a score of $1 - \varepsilon$ of the optimal score can be found. This defines an area of the matrix. If ε is less than zero, the area is empty, if it is zero, it will be the entries that are on optimal alignments of the sequences. As ε increases, the defined area will converge toward the complete matrix. Figure 7 shows how much of the likelihood function is found within the area as a function of ε. The sequences involved were derived from a sequence of length 1500 amino acids, that experienced evolution corresponding to the difference between α globin and myoglobin. It can be observed that the relative underestimation of the likelihood is less than $e^{-13} = 2.3 \times 10^{-6}$ if an ε of 0.01 is used. An ε value of 0.01 corresponds to 1.8×10^{-3} of the area of the complete matrix. This gives rise to a very significant acceleration. This is a favourable case, but in general the acceleration is considerable and the area containing all significantly contributing paths is very narrow. The closer related the sequences are, the narrower the band will typically be.

If ε is too small, alignments that contribute significantly to the alignment will be discarded, resulting in a bias in the estimated parameter values. Since a low ε value will discard alignments with many indels, this is expected to create a bias towards smaller values of μ and λ, which was also observed (results not shown).

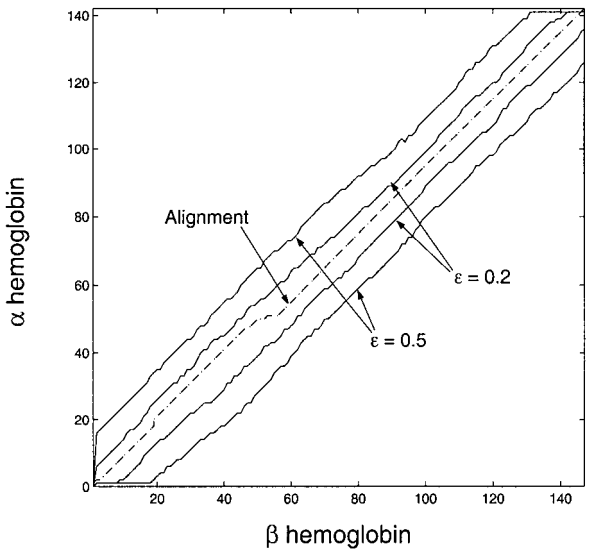


Figure 6. Illustration of the similarity alignment as a path in the matrix. The maximally contributing statistical alignment and the similarity alignment is identical in this simple case. The area containing nodes that could be on a suboptimal solution, better than $(1-\varepsilon)$ of the optimal similarity score, is also illustrated for ε equal to 0.2 and 0.5. For practical purposes a band much narrower can be used, typically less than 0.005.

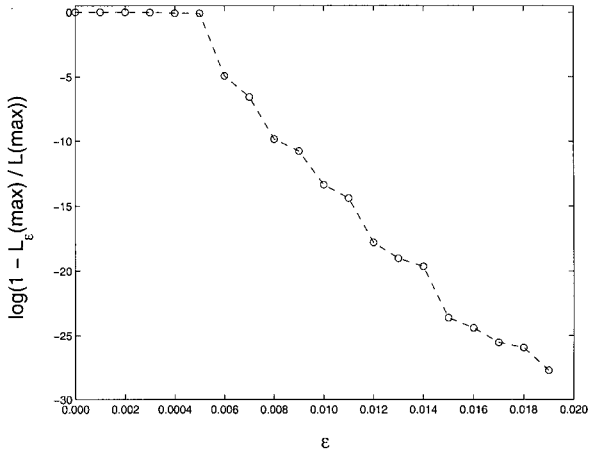


Figure 7. A plot of $\log(1-L_\varepsilon/L)$ as a function of ε for simulated globins of length 1500, with evolutionary distance like myoglobin (153 amino acids) and α globin (141 amino acids). This maps [0, 1] into $[0,-\infty]$, and illustrates how much of the contributions to the likelihood function is within ε of the similarity optimum alignment solution. It is obvious that most contributing paths to the likelihood functions are within a very narrow band. This points to an obvious speedup of the likelihood method.

The number of evaluations

This will consist of three parts: an initial guess of parameters, an algorithm searching for the minimum, and a stopping criteria determining whether the current parameter estimates are sufficiently close to the maximum likelihood values. The problem is illustrated in Figure 8, where the $L(\mu,s)/L_{\max}$ is plotted in an area close to the maximum likelihood estimate for α globin and myoglobin. The method used is to make a good guess that is within the through containing the maximum likelihood point, crawl in few steps to the bottom of the basin and stop, when iterations does not improve the estimates significantly.

Initial guess. We only consider the protein model and the strategy will be to assume that the parsimony/similarity alignment is the correct alignment. We calculate how many gap signs, matches, and mismatches would be expected and choose the parameters that give these expectations. In the protein substitution model, the expected number of mismatches, $(1-\Sigma_i \pi_i P_{ii})n_{\text{align}}$, were calculated and equated to the observed number, giving a guess for s (n_{align} is the number of columns without gap signs in the alignment). λ and μ was guessed by first observing that the expected length of a sequence is $\lambda/(\mu-\lambda)$. This would fix the ratio of λ/μ to:

$$\frac{\lambda}{\mu}=\frac{l_{\text{ave}}}{l_{\text{ave}}+1}$$

where l_{ave} is the average sequence length. In addition, it is possible, when knowing the length of one sequence and the p functions, to calculate the expected number of gap signs in an alignment to (see Appendices):

$$\#\text{gap}=\frac{\lambda\beta}{1-\lambda\beta}+s\left(e^{-\mu t}\frac{\lambda\beta}{(1-\lambda\beta)}+\mu\beta+(1-e^{-\mu t}-\mu\beta)^2\frac{2-\lambda\beta}{1-\lambda\beta}\right)$$

Using this, a guess of $\mu=0.0316$ and $s=0.9500$ is obtained from the α globin *versus* β globin similarity alignment. A slightly inferior guess of μ can be obtained by assuming that the observed gap signs are the events that actually have happened in the evolutionary history. This gives $2\mu L=\#\text{gap}$ and will give a lower estimate than using the p functions. Using this method μ is estimated to 0.0311. This difference is probably larger for more distantly related sequences. The maximum likelihood results are shown in Figure 3.

Optimisation. Several numerical optimisation methods were tried (e.g. simplex and Powell), but given a good initial guess, the best was BFGS (Broyden-Fletcher-Goldfarb-Shanno) (Press *et al.*, 1992). An example of the search for the maximum likelihood estimate is illustrated in Figure 8, for simulated sequences approximately of length 1500,

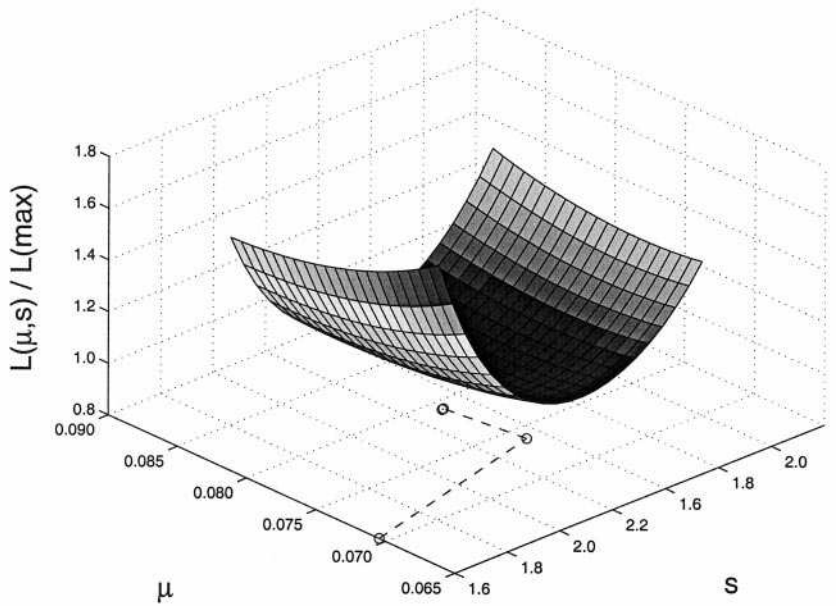


Figure 8. Likelihood surface for human α globin aligned with human myoglobin. Only two parameters are allowed to vary, since the ratio of λ/μ has been fixed to $l_{ave}/(l_{ave}+1)$. l_{ave} is the average length of the two proteins. The numerical optimisation part of the statistical alignment problems is to find an initial point within this valley, as close as possible to the bottom, and then through a series of iterations get close to the minimum. In the floor of the diagram the search for the minimum is shown. $(s, \mu)_0$ is the initial guess obtained from analysing the similarity alignment. After three iterations the improvements in the likelihood was negligible. BFGS (see the text) had then used 28 evaluations of the likelihood function. Each iteration needs several evaluations to determine first and second derivatives of the likelihood function (in our implementation derivatives were found numerically, but could in principle be found by dynamical programming).

with an evolutionary distance like α globin and myoglobin. In this case, four iterations and 28 likelihood evaluations were needed. Each iteration needs a series of likelihood evaluations to determine first and second derivatives of the likelihood function. The total number of likelihood evaluations is typically less than 50.

Stopping condition. When an iteration produced changes in the likelihood estimates, that was less than 10^{-3}, the iterations stopped. Figure 9 shows $L_{tot}(k)$ as function of iteration number, k. It can be seen that major improvements are obtained in the first few jumps. After three to four iterations $L_{tot}(k)$ is very close to the likelihood function taken in the maximum likelihood estimate.

The basic recursion

The likelihood recursions (three to four) of TKF91 are a bit more complicated than the recursion in the optimisation alignment algorithm (parsimony/similarity). Comparison between the two indicated that the likelihood recursion was 50-70 times slower than the optimisation recursion. The main reason for this large difference is that multiplication of reals is slower than addition of integers on most computers.

Summary

The above improvements yield a method that is significantly faster than the one described by Thorne *et al*. (1991). It seems probable that a further increase in speed can be obtained from focusing on the last two factors. In absolute terms, two proteins of length 1500, can be analysed in less than five seconds on a fast desktop computer (Silicon Graphics Octane with a 300 MHz R12000 processor), which makes statistical alignment a fully practical method for two sequences.

Homology test

Consider the α globin and myoglobin. Are they homologous? Homology must here be the answer to the question, whether the value of t is finite or infinite. A value of infinity implies that they could both have been drawn independently from the stationary distribution of the evolutionary process. Statistical alignment can contribute considerably to this question.

Parsimony/similarity alignment based test

Most tests in a parsimony/similarity alignment setting presuppose an alignment and regard the matched positions as independent realizations of the same distribution.

Most homology tests are based on a similarity scoring function, for each position in the alignment, of the form: $W_{i,j} = \ln(\pi_i P^{2.5}_{i,j}/(\pi_i \pi_j))$ (Altschul, 1993). In this expression, $P^{2.5}_{i,j}$ is the transition probabilities, when 2.5 units of time has passed. This amounts to choosing among the competing hypothesis that two sequences are 2.5 events apart *versus* infinitely far apart. It only handles substitutions ''correctly''. The rationale for indel cost is more arbitrary.

In a frequently used test, the shuffle test, the two sequences are aligned and a score is obtained (Doolittle, 1986). The significance of this score is

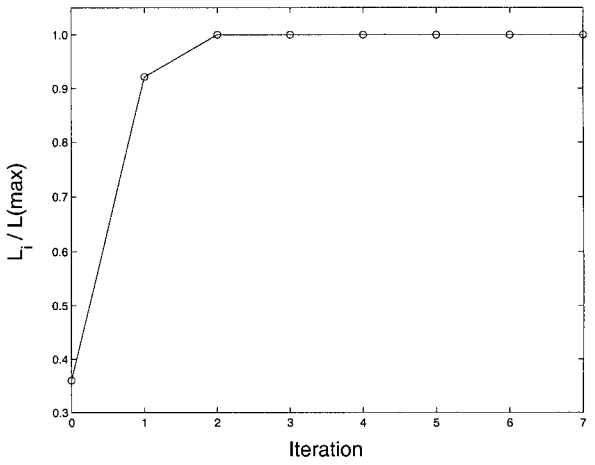


Figure 9. This Figure shows the likelihood values in different iterations. After three to four steps, a guess very close to the maximum likelihood has been achieved.

evaluated by permuting the order of the amino acids and aligning the permuted proteins:

```
          Real                 Random
s(1) = ATWYFCAK-AC   s(1) = ATWYFC-AKAC
s(2) = ETWYKCALLAD   s(2) = LTAYKADCWLE
       *** **  *             **
```

This is done many times and the real score is compared to the score of the permuted proteins. This amounts to sampling in the observed amino acid distributions without replacement. If the score for the real sequences are much better (high for similarity alignment or low for parsimony alignment), the proteins are assumed homologous. An illustration of this test for α globin and myoglobin is shown in Figure 10 (top).

This approach has several drawbacks. Firstly, it is dependent on having the correct alignment, which is unlikely for distant sequences. Secondly, it must fix a time back to a common ancestor that is unknown, since any substitution matrix assumes a distance between sequences. Thirdly, it is hard to introduce more realistic models of sequence evolution in this test. Statistical alignment has the potential of solving these problems.

Statistical alignment based test

In testing homology we are asking if two sequences have a common ancestor finitely far back in time. Here we try to distinguish two competing hypotheses. Are they independent sequences from the equilibrium distribution on the set of sequences? Or are they related by a tree with a root finitely far back in time? The test will be parametric bootstrap as described by Cox (1962).

(1) All parameters, (λ_{real}, μ_{real}, s_{real}), are estimated by maximum likelihood for the given sequences.

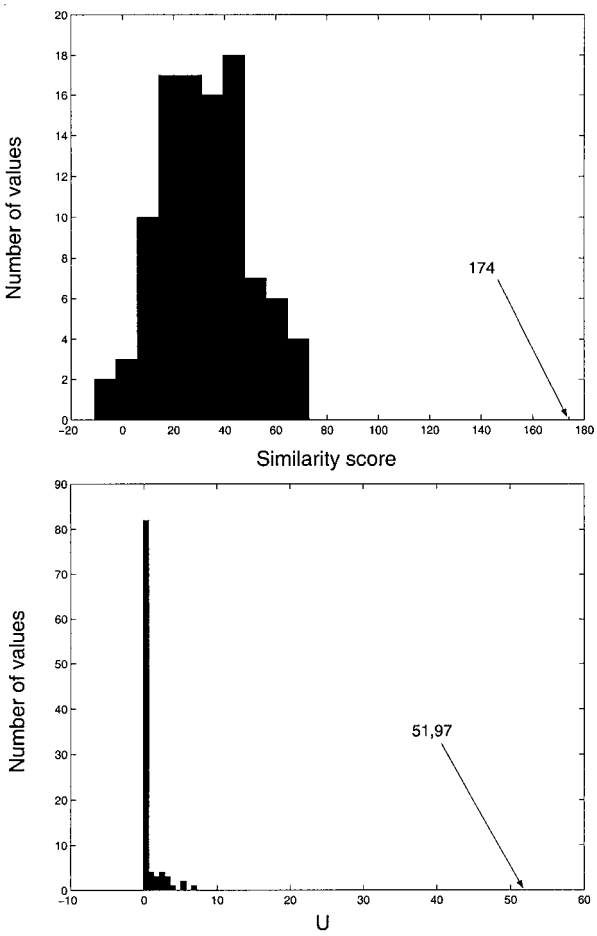


Figure 10. Top: shuffle testing of the homology between myoglobin and α globin. The arrow to the right is the score of the real sequences. Bottom: testing the homology between myoglobin and α globin, using statistical alignment. The arrow to the right is the score of the real sequences.

(2) Pairs of independent sequences, $(s^{(1)}, s^{(2)})_i$, are simulated using these parameters.

(3) These simulated sequences are analysed using statistical alignment and parameters are reestimated, (λ_i, μ_i, s_i).

The following statistic, U, is calculated for the real sequences and for the simulated sequences:

$$U = -2 \ln \frac{P(s^{(1)}, s^{(2)})}{P(s^{(1)})P(s^{(2)})}$$

Now, the value for the real sequences can be compared to the distribution of the values for the simulated sequences. If the value for the real sequences is extreme for the distribution, the sequences are homologous.

An illustration of this test for α globin and myoglobin is shown in Figure 10 (bottom). This approach has solid potential, but at present it has a number of drawbacks that prevent it from broad use. Firstly, it is much slower than database scan-

ning programs. Secondly, the TKF91 process is not a realistic model for sequence evolution. Especially, the geometric equilibrium distribution of sequence lengths is not believable and should be improved.

A method for homology testing, that also sums over all alignments, but without an evolutionary model has been made by Bucher & Hofmann (1996).

Goodness of fit-testing the TKF process

It is obviously of importance to test the proposed model when using it to analyse real data. Tests have been developed (especially due to Goldman (1993)) for testing substitution models, when an alignment is given. Since the new aspect of the TKF91 process is the indel process, we will focus on testing whether this aspect of sequence evolution is well modelled. The TKF91 model assumes that insertions and deletions occur in steps of one.

Many optimisation alignment methods put much emphasis on having the correct gap penalty function and assume that indels can involve longer segments. The alignments obtained by the TKF91 method can also have consecutive runs of gaps signs, but they would all have been inserted or deleted individually. If longer indels occur, it should nonetheless be reflected in longer runs of gap signs in the alignments proposed by the TKF91 method. It is therefore very natural to compare the p functions with the corresponding configurations of survival and number of descendant obtained from the TKF91 method.

Again, consider human α globin and β globin. If their true alignment could be observed, the fate of 141 amino acids and one immortal link had been observed. Table 2B shows which numbers would be obtained if the alignment in Figure 3 were used. Given the maximum likelihood parameters and the two sequences, Table 2C can be filled by a dynamic programming algorithm, using recursion (1-2) to assign probabilities of the fate of $s^{(1)}[i]$ in $s^{(2)}$ given that $\{s_i^{(1)} \to s_j^{(2)}\}$. We chose to sample alignments (100) according to their probability, using the stochastic backtracking procedure, since this was easier to program. This is not an alignment chosen uniformly among all possible alignments, but chosen randomly in proportion to how much they contribute to the likelihood function in the maximum likelihood point.

The difference between Table 2A and C is measured by the $X^2 = (\text{obs} - \text{exp})^2/\text{exp}$ statistic. This is 532.17 in this case. The cells contributing to this are shown in Table 2D. It is obvious from alignments of real sequences, that longer runs of gap signs occur, that are not in accordance with the model.

It is possible to get longer series of gap signs in alignments of real sequences, that contribute massively to the X^2_{real} statistic. If this contribution is statistically significant, a natural interpretation would be that longer indels had occurred. Nonetheless, alternative explanation cannot be ruled out. For instance, the indel rate could be unevenly distributed along the sequence, so many single indels occurring next to each other were actually quite probable.

To assess significance between Table 2A and C, 100 sequences (s_i terms) were simulated starting from α globin and evolving according to the estimated evolutionary parameters. The X^2 statistics were calculated from these by making analogues

Table 2. Results from the goodness of fit test for the indel model in the evolution of α globin to β globin

A. *Expected according to model and length of α globin*							
No. of descendants	0	1	2	3	4	5	6
Immortal link	-	0.9642	0.0346	0.0012	0.0000	0.0000	0.0000
Mortal links - survived	-	130.95	4.6937	0.1682	0.0060	0.0002	0.0000
Mortal links - died	5.0891	0.0890	0.0032	0.0004	0.0000	0.0000	0.0000
B. *Observed in optimal alignment*							
No. of descendants	0	1	2	3	4	5	6
Immortal link	-	1	0	0	0	0	0
Mortal links - survived	-	135	2	1	1	0	0
Mortal links - died	2	0	0	0	0	0	0
C. *Expected according to model and sequences*							
No. of descendants	0	1	2	3	4	5	6
Immortal Link	-	0.95	0.05	0.00	0.00	0.00	0.00
Mortal links - survived	-	132.77	4.14	0.67	0.35	0.21	0.05
Mortal links - died	2.68	0.12	0.01	0.00	0.00	0.00	0.00
D. X^2 *difference between A and C*							
No. of descendants	0	1	2	3	4	5	6
Immortal Link	-	0.0004	0.0069	0.0012	0.0000	0.0000	0.0000
Mortal links - survived	-	0.0253	0.0653	1.496	19.62	203.62	311.04
Mortal links - died	1.1404	0.0108	0.0145	0.0001	0.0000	0.0000	0.0000

Maximum likelihood parameter estimates were obtained from analysis of α globin and β globin and then regarded as fixed. Each section (A to D) tabulates the quantities relating to the three p functions. A. The expectation from the different p functions. For the mortal links, these expectations are $141p_k$ (survived) and $141p'_k$ (died). For the immortal link it is simply p_k'', since there is only one. B. The result, if the optimal alignment (Figure 3) were used in filling out the Table. C. Sampling 100 random alignments in proportion to their probability using equation (11). D. Contributions to the X^2 statistic form the difference between A and C.

to Table 2A and C. If X^2_{real} is extreme in this distribution, the indel process does not fit well with the real sequences. The distribution of X^2 is shown in Figure 11 together with X^2_{real}. Obviously, the TKF91 needs modification to be a satisfactory model.

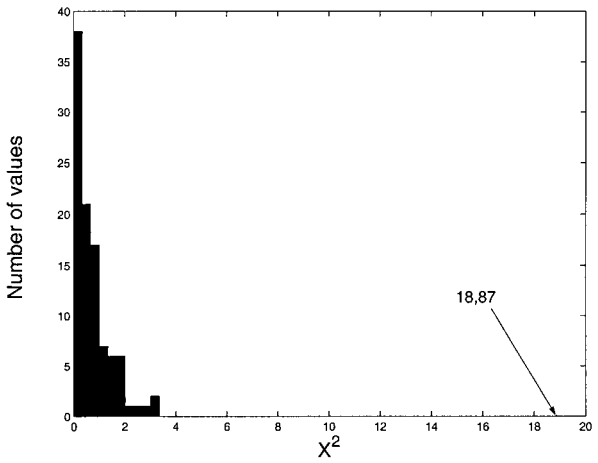


Figure 11. Goodness of fit testing indel lengths in the TKF model. The X^2 value of the real data (marked by an arrow) is very extreme in the distribution of the simulated X^2 values.

Discussion

This article has highlighted some of the advantages of a more statistical approach to alignment, the main ones being:

(1) It is explicitly founded on a description of molecular evolution.
(2) Parameters are estimated and biologically meaningful.
(3) Different evolutionary events can be assigned probabilities.

However, the present model is unrealistic, thus, many generalisations and improvements are of immediate practical interest.

Since it is clear that indels longer than one nucleotide or amino acid do occur, incorporating this into a model would be a significant step towards biological realism. However, it is not straightforward to do this. Allowing for longer insertions is simple and such a longer insertion could be associated to a single link, as in the TKF91 process. Longer deletions remove intervals of sequences, and should be modelled so that the whole process is time reversible, since this has computational advantages. There is no biological reason for believing that the insertion process should be the time-reversed image of the deletion process. It remains to be explored how seriously this assumption is violated in real data.

A second extension would be to generalise the TKF91 dynamic programming algorithm, calculating the likelihood for a set of homologous sequences. Steel & Hein (2000) have done this for k sequences, related by a star-shaped tree. J.H. (unpublished results), has generalised this further to k sequences, related by a binary tree, in an algorithm that has $O(l^k)$ running time in the sequence length. This is analogous to the parsimony algorithm relating k sequences devised by Sankoff (1975). However, an implementation of the likelihood method would be much slower than the parsimony/similarity method, due to its more complicated algorithm, parameter optimisation, etc. To yield a practical statistical multiple alignment method, other methods than the dynamic programming algorithm would have to be used, e.g. Markov chain Monte Carlo methods.

Modelling substitutions and indels that are unevenly distributed along the sequences could also be improved. Real sequences will have different probabilities of insertion/deletion for different regions. For proteins, it is well known that insertion/deletions are more frequent in loop regions than in sheets and helices, and it would give a more realistic model to take advantage of this knowledge. Incorporating the hidden Markov model for different structural categories as done by Goldman *et al*. (1996) seems especially relevant. Using hidden Markov models will pose a problem, since the indel process would make the Markov model longer and shorter at stochastic times.

The TKF91 model is simple and tractable, but it would be of interest to explore alternatives. The view of a sequence being tagged by an immortal link does not conform to biological intuition. A possibility would be to let a sequence be born from a given equilibrium distribution and to be killed according to some process. Whether this could lead to a tractable process remains to be explored, but it would conform better to biological intuition and could give a better equilibrium distribution of sequence lengths than the geometric distribution of the TKF91 model. In this context, it would also be of interest to formulate how subsequences can be homologous to subsequences. The tests described here were solely addressed in terms of global comparisons and to devise a practical competitive test, it would be necessary to formulate an analogue of local alignment for statistical alignment.

More realistic models of sequence evolution and methods for aligning more sequences would automatically lead to better homology tests. In this context, it should be noted that when molecular biologists perform homology tests (or database searches), their prime objective is not homology, but rather inferences about function. It might be advantageous to model this explicitly, i.e. to model not only the sequence but also the probability that a sequence with one function obtains another function. The approach taken here might also unify the contending approaches of Dayhoff *et al*. (1978) *versus* Henikoff & Henikoff (1992), in making score matrices. Dayhoff constructs matrices from closely

related sequences that will define log-odd scores for distantly related sequences. Henikoff & Henikoff use conserved blocks in distantly related sequences to define log-odd scores directly. These two approaches seem to focus on quickly and slowly evolving positions, respectively. A statistical alignment model directly incorporating quickly and slowly evolving positions would unify the two approaches.

The concept of homology in sequence comparison is not crystal clear. Since the earliest organism probably contained very few sequences (possibly only one), maybe all sequences are homologous in the strict sense. There have been assertions about the number of different protein families appearing in life on earth (Chothia, 1992).

Statistical approaches to alignment have many advantages relative to parsimony/similarity approaches, but the latter methods have a large lead in software developments. Even if statistical approaches were developed to a stage where it was better at the conceptual and good at the algorithmic level, there would still be a huge software gap for many years to come.

Comment

The programs and tests developed in this paper can all be accessed at the web-site: www.brics.dk/ ~ compbio. The program contains the following parameters to be set of the user: the narrowness of the band where dynamical programming is performed and the level of precision in parameters, when iterations are to be stopped.

Acknowledgements

Anne-Mette Krabbe Pedersen, Christian Nørgård Storm Pedersen, Jan Gorodkin, Mikkel Schierup, Jakob Skou Pedersen, Joseph Felsenstein, and the anonymous reviewer are thanked for useful comments on the manuscript. C.W. and J.H. were generously supported by the Newton Institute of Mathematical Sciences. G.W. was supported by BRICS and Danish Agricultural and Veterinary Research Council (grant 9901522). C.W. was supported by grant BBSRC 43/MMI09788 and by the Carlsberg Foundation.

References

Allison, L. & Wallace, C. S. (1994). The posterior probability distribution of alignments and its application to parameter estimation of evolutionary trees and to optimization of multiple alignments. *J. Mol. Evol.* **39 (4)**, 418-430.

Altschul, S. F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* **36 (3)**, 290-300.

Bishop, M. J. & Thompson, E. A. (1986). Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* **190 (2)**, 159-165.

Bucher, P. & Hofmann, K. (1996). A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology* (States, D. J., Agarwal, P., Gaasterland, T., Hunter, L. & Smith, R. F., eds), pp. 44-51, AAAI Press, California.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature,* **357 (6379)**, 543-544.

Cox, D. (1962). Further results on tests of separate families of hypotheses. *J. R. Stat. Soc. ser. B,* **24**, 406-424.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins, matrices for detecting distant relationships. In *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, pp. 345-352, Cambridge University Press, Washington, DC.

Doolittle, R. F. (1986). *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*, University Science Books, California.

Edwards, A. W. F. (1972). *Likelihood*, Cambridge University Press, Cambridge.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36 (2)**, 182-198.

Goldman, N., Thorne, J. L. & Jones, D. T. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.* **263**, **(2)**, 196-208.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705-708.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89 (22)**, 10915-10919.

Mitchison, G. J. (1999). A probabilistic treatment of phylogeny and sequence alignment. *J. Mol. Evol.* **49 (1)**, 11-22.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipies in C*, 2nd edit., Cambridge University Press, Cambridge.

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

Steel, M. & Hein, J. J. (2000). A generalisation of the Thorne-Kishino-Felsenstein model of statistical alignment to k sequences related by a star tree. *Appl. Math. Letters,* **in the press**.

Thorne, J. L., Kishino, H. & Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33 (2)**, 114-124.

Thorne, J. L., Kishino, H. & Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* **34 (1)**, 3-16.

Zhu, J., Liu, J. S. & Lawrence, C. E. (1998). Bayesian adaptive sequence alignment algorithms. *Bioinformatics,* **14 (1)**, 25-39.

Appendix I: Finding Alignments

The most probable alignment for a given set of parameters, can be found. Since the number of alignments is large, this probability will be small relative to $P(s^{(2)}|s^{(1)})$. The reasoning behind equations (3)-(4) can be applied again. Let $s^{(1)} \to s^{(2)}$ denote all evolutionary paths from $s^{(1)}$ to $s^{(2)}$. Instead of keeping track of the set $\{x \in s_i^{(1)} \to s_j^{(2)}\}$ and $\{x \in s_i^{(1)} \to s_j^{(2)} | s^{(2)}[j]$ is a descendant of $s^{(1)}[i]$ in $x\}$ with many alignments in them,

only keep the most probable alignment in these sets (indicated with a $\{\ \}_{\max}$):

$$R^{\max}_{i,j} = \max\{p_1 P_{s^{(1)}[i],s^{(2)}[j]} P(\{s^{(1)}_{i-1} \to s^{(2)}_{j-1}\}_{\max}),$$
$$p'_1 \pi_{s^{(2)}[j]} P(\{s^{(1)}_{i-1} \to s^{(2)}_{j-1}\}_{\max}),\ \lambda\beta\pi_{s^{(2)}[j]} R^{\max}_{i,j-1}\} \quad (6)$$

$$P(\{s^{(1)}_i \to s^{(2)}_j\}_{\max}) = \max\{R^{\max}_{i,j}, p'_0 P(\{s^{(1)}_{i-1} \to s^{(2)}_j\}_{\max})\} \quad (7)$$

Using backtracking, the most probable alignment can be found. This alignment is of little interest and is mainly calculated to be able to generate one alignment for illustration. As shown by Thorne *et al*. (1991), this alignment is not representative of the actual history of $s^{(1)}$ and $s^{(2)}$, but without it, this method would be an alignment method that did not produce any alignment.

Alignments can be generated in proportion to their probability. This can be done by the following procedure starting in ($l1$, $l2$) (the lengths of the two sequences) and going down to (0, 0):

Step	Probability	Alignment block
$R_{i,j} \to P(s^{(2)}_{j-1} \mid s^{(1)}_{i-1})$	$p_1 P_{s^{(1)}[i],s^{(2)}[j]} P(s^{(2)}_{j-1} \mid s^{(1)}_i)/R_{i,j}$	$s^{(1)}[i]$ / $s^{(2)}[j]$
$R_{i,j} \to P(s^{(2)}_{j-1} \mid s^{(1)}_{i-1})$	$p'_1 \pi_{s^{(2)}[j]} P(s^{(2)}_{j-1} \mid s^{(1)}_i)/R_{i,j}$	$s^{(1)}[i]$ / $-$ / $-$ / $s^{(2)}[j]$
$R_{i,j} \to R_{i,j-1}$	$\lambda\beta\pi_{s^{(2)}[j]} P(s^{(2)}_{j-1} \mid s^{(1)}_i)/R_{i,j}$	$-$ / $s^{(2)}[j]$
$P(s^{(2)}_j \mid s^{(1)}_i) \to R_{i,j}$	$R_{i,j}/P(s^{(2)}_j \mid s^{(1)}_i)$	Nothing
$P(s^{(2)}_j \mid s^{(1)}_i) \to P(s^{(2)}_j \mid s^{(1)}_{i-1})$	$p'_0/P(s^{(2)}_j \mid s^{(1)}_{i-1})/P(s^{(2)}_j \mid s^{(1)}_i)$	$s^{(1)}[i]$ / $-$

It is also possible to sample random alignments, using an analogue to equation (5), but it seems difficult to formulate a maximum analogue to equation (5) in the style of equations (6) and (7).

Reference

Thorne, J. L., Kishino, H. & Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Enol.* **33(2)**, 114-124.

Appendix II: Calculating Length Distributions

Take a sequence of length n. Letting this sequence evolve over time, makes the length change according to the λ and μ parameters of the model. It is possible to calculate the distribution of lengths for any given time t that passes.

For low values of t, the length distribution will be very narrow around n. For larger values of t, the distribution becomes a skewed bell shape around n. With very large t values the distribution will become geometric as dictated by the λ and μ parameters.

The generating functions (GFs) can be found for the number of children of mortal and immortal links. Multiplying an appropriate number of these can give the GF for the entire length of a sequence. Thus, given an initial length and an amount of time, the length distribution can be calculated as (P_m being the probability of having length m at time t, starting at length n):

$$P_m = \sum_{i=0}^{\min(m,n)} \binom{n+m-i}{m-i\quad n-i\quad i} (-a)^{n-i} d^{n-i} c^i b^{-n-m+i-1}$$

with:

$$a = a(t) = -\frac{\lambda}{\mu-\lambda}\gamma(t)$$

$$b = b(t) = 1 - a(t) = 1 + \frac{\lambda}{\mu-\lambda}\gamma(t)$$

$$c = c(t) = 1 - \frac{\lambda}{\mu-\lambda}\gamma(t)$$

$$d = d(t) = 1 - c(t) = \frac{\lambda}{\mu-\lambda}\gamma(t)$$

$$\gamma(t) = 1 - e^{(\lambda-\mu)t}$$

The average and variance of the length distribution is:

$$E(S_t) = n + \left(\frac{\lambda}{\mu - \lambda} - n\right)\gamma(t)$$

$$Var(S_t) = nc(1-c) - (n+1)a(1-a)$$

Appendix III: Expected Number of Gaps

The expected number of gaps produced by a mortal link in time t, times its probability of survival is:

$$g(t) = \sum_{n=1}^{\infty} p_n(t)(n-1)$$

since $n-1$ gaps are produced when a mortal link has n children and it survives. This can be written as:

$$g(t) = e^{-\mu t}(1-\lambda\beta)\sum_{n=0}^{\infty} n(\lambda\beta)^n = e^{-\mu t}\frac{\lambda\beta}{(1-\lambda\beta)}.$$

The same calculations can be done for mortal links that die and for immortal links. The total expected number of gaps is the sum of an appropriate number of each of these g functions:

$$\#gap = \frac{\lambda\beta}{1-\lambda\beta} + s\left(e^{0\mu t}\frac{\lambda\beta}{(1-\lambda\beta)} + \mu\beta + (1-e^{-\mu t}-\mu\beta)\frac{2-\lambda\beta}{1-\lambda\beta}\right)$$

$$\beta = \frac{1-e^{(\lambda-\mu)t}}{\mu - \lambda e^{(\lambda-\mu)t}}$$

$$\frac{\mu t}{\lambda t} = \frac{s+1}{s}$$

Edited by J. Karn

(Received 17 April 2000; received in revised form 21 July 2000; accepted 25 July 2000)

Chapter 5

A Likelihood Ratio Test for Evolutionary Rate Shifts and Functional Divergence among Proteins

Article by Bjarne Knudsen and Michael M. Miyamoto.

Appeared in *Proc. Natl. Acad. Sci. USA*, **98** (25), 14512–14517, 2001.

A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins

Bjarne Knudsen*† and Michael M. Miyamoto‡

*Bioinformatics Research Center, University of Aarhus, Høegh Guldbergsgade 10, Building 090, DK-8000 Århus C, Denmark; and ‡Department of Zoology, Box 118525, University of Florida, Gainesville, FL 32611-8525

Communicated by Walter M. Fitch, University of California, Irvine, CA, September 28, 2001 (received for review July 9, 2001)

Changes in protein function can lead to changes in the selection acting on specific residues. This can often be detected as evolutionary rate changes at the sites in question. A maximum-likelihood method for detecting evolutionary rate shifts at specific protein positions is presented. The method determines significance values of the rate differences to give a sound statistical foundation for the conclusions drawn from the analyses. A statistical test for detecting slowly evolving sites is also described. The methods are applied to a set of Myc proteins for the identification of both conserved sites and those with changing evolutionary rates. Those positions with conserved and changing rates are related to the structures and functions of their proteins. The results are compared with an earlier Bayesian method, thereby highlighting the advantages of the new likelihood ratio tests.

The explosive growth of available sequence data has necessitated the development of new computerized methods for the functional analysis of proteins. A number of methods have been developed for studying the functions of proteins from their sequences and protein-coding DNAs (1–3). Some of these methods estimate the ratio between nonsynonymous and synonymous rates within protein-coding genes, with ratios >1 and <1 indicating positive versus negative selection, respectively (4). Methods for performing these analyses on a site-specific level also have been developed (5). Along these lines, other methods have focused on amino acid conservation as an indication of protein function (6, 7). This approach is founded on the assumption of functional constraint (i.e., that functionally important residues and sequences are under stronger selective constraints that lower their evolutionary rates).

The concept of amino acid conservation can be taken one step further to yield insights about changes in function over time. This divergence of protein function often is revealed by a rate change in those amino acid residues of the protein that are most directly responsible for its new function (8, 9). To investigate this change in evolution, a likelihood ratio test (LRT) is developed for detecting significant rate shifts at specific sites in proteins.

Such rate changes at a site over evolutionary time trace back to the covarion model of Fitch and Markowitz (10). In this model, the state of a site can change between variable and invariable. Such changes can also occur anywhere in the evolutionary tree relating the sequences under analysis. Furthermore, as the acronym implies (concomitantly variable codons), these rate shifts are tied to sites whose evolution is correlated and is not independent (11). The LRT method assumes that changes occur at a specific point in evolution and that these changes are independent. Here, change is not limited to variable versus invariable, but involves shifts between any two rates. For this reason, we are not dealing with a true covarion model (12, 13). Thus, a site showing a significant rate change will from here on be called a rate shift site, rather than a covarion site.

The reason for focusing on a specific evolutionary point is that gene duplications can create opportunities for functional divergence as one copy of the gene can divergently evolve, whereas the other fulfills the original function (2, 7). Other points in gene evolution where functional change is most likely reflect speciation events that lead to the origins of new major groups [e.g., ciliates versus other eukaryotes in the divergence of their elongation factors (14)].

A slow evolutionary rate at a given site would indicate that this position is functionally important for the protein. Conversely, a high evolutionary rate would indicate that the position is not involved in an important protein function. A significant rate difference between two subfamilies at a given site would thereby mean that the function of this position is probably different in the two groups.

Some work has been done in this area before (2, 8, 15). The approach developed here is unique in that it uses an LRT to determine the significance of the rate differences at specific positions. A test is also developed for deciding whether a given site is evolving slower than the average for the entire protein being analyzed.

Tests for detecting whether two subfamilies have undergone functional divergence have been developed before (15) and will not be the focus of this work. Instead, it is assumed that the subfamilies are known to be functionally divergent, either from biochemical knowledge or previous statistical tests. This work aims to pinpoint the protein positions responsible for this divergence.

The methods are illustrated with a set of Myc proteins and the biochemical significance of these results is discussed. The results are compared with those using the Bayesian method of Gu (15), which calculates the posterior probability of a rate shift. The reasons for the differences between the two approaches are explained.

The Model

The LRT is used as the basis for detecting rate shift sites. The basic idea behind this test, as used in an evolutionary context, was reviewed by Huelsenbeck and Rannala (16).

Position-Specific Rate Shift Test. To test whether a site from two related groups of sequences is evolving differently, the positions are analyzed individually. An outline of the method is shown in Fig. 1 *Left* and *Center*.

The test used is as follows. The null hypothesis, H_0, states that a given position evolves with different rates in the two sequence subfamilies. The likelihood under this model is calculated by using the method of Felsenstein (17). The rate matrix used is the JTT matrix of Jones *et al.* (18). The two rates of evolution are varied to obtain the maximum-likelihood (ML) value under this model, L_0.

In contrast, hypothesis one (H_1) states that the position evolves at the same rate in the two subfamilies. Again, calculations are done according to Felsenstein (17) and with the JTT matrix, but with a single rate used for the two subfamilies. The optimal rate is found, giving the ML value under this model, L_1.

Abbreviations: LRT, likelihood ratio test; ML, maximum likelihood; bHLHZip, basic helix–loop–helix leucine zipper.

†To whom reprint requests should be addressed. E-mail: bk@birc.dk.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

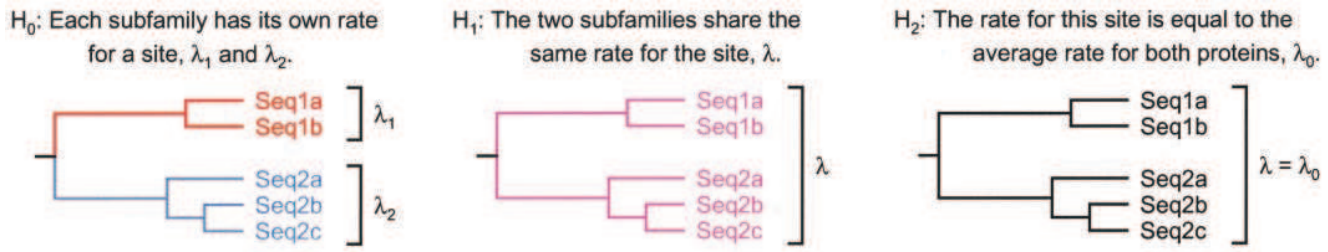


Fig. 1. Assume that a gene duplication has resulted in two protein subfamilies. The first consists of sequences Seq1a and Seq1b, whereas the second includes sequences Seq2a, Seq2b, and Seq2c. (*Left*) H_0, where the rates for a site may differ from one protein subfamily to the other. This rate divergence occurs at the root of the tree, where the duplication event occurred. (*Center*) The situation under H_1. The evolutionary rate for a site remains the same throughout the entire tree. If H_1 is rejected, rate shift behavior is present at the position under inspection. If H_1 is retained, then one can test whether the rate for this site is equal to the average for both proteins. (*Right*) The testing of this hypothesis (H_2). If H_2 is rejected, the evolutionary rate for the site is significantly different from the average for all positions.

Using an LRT statistic, we can evaluate H_1. The test statistic can be written as:

$$U = -2log\frac{L_1}{L_0}.$$

Because H_1 is a special case of H_0 (the hypotheses are nested), the likelihoods will always obey the relationship that $L_1 \leq L_0$. This means that U will never be negative.

There are two degrees of freedom under H_0, whereas there is only one under H_1. This could indicate that under H_1, the distribution of U is approximately χ^2 with one degree of freedom, here denoted $\chi^2(1)$. To investigate how close the distribution of U is to $\chi^2(1)$, a number of simulations were conducted (Fig. 2). The simulated distributions were quite close to the $\chi^2(1)$ distribution, so a $\chi^2(1)$ test can be used with some caution.

Unknown and partially known amino acids are treated as described by Felsenstein (17). This means that unknown amino acids have the effect of pruning the tree to remove the sequences containing them. Gaps are treated like unknown amino acids. This means that all columns in the alignment can be used in the analysis, even though some sequences are unknown or gapped in that region. For moderate numbers of sequences with a gap at a specific position, the test statistic is not influenced much, because this corresponds to using a smaller tree (which would ideally have the same distribution of U).

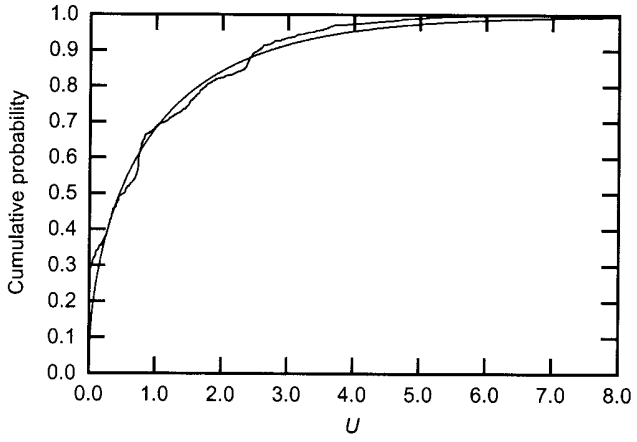


Fig. 2. The χ^2 distribution with one degree of freedom (smooth curve) compared with a simulation study of U (ragged curve). The simulations consisted of 1,000 samples generated under H_1, with rates drawn from a gamma distribution. The calculations are based on the phylogeny and ML conditions used in the Myc protein example. The distribution of U approximately follows that of the $\chi^2(1)$ statistic, especially in the upper part. Other follow-up simulations indicate that this distribution generally conforms more closely to the $\chi^2(1)$ curve as the two subfamilies increase, both in terms of their branch lengths and numbers of sequences (Figs. 4–7, which are published as supporting information on the PNAS web site).

Advantages and Disadvantages of the Method. The LRT method has the advantage that it is simple and direct. It answers exactly the question of interest: Does a given position evolve at different rates in different protein subfamilies? The Bayesian method is indirect, because it only uses the JTT matrix to count the expected number of replacements within each subfamily, before comparing these counts (15). The problem is that some replacements are rare (e.g., lysine to cysteine), whereas others are more common (e.g., valine to isoleucine). The JTT rate matrix is fully incorporated in the likelihood calculations presented here to accommodate this fact.

Another advantage of this method, compared with some earlier ones (e.g., ref. 8), is that it acknowledges that the subfamilies are related to each other and are not independent. To illustrate this, consider a given position in the sequences of Fig. 1. Assume that Seq1a and Seq1b have an isoleucine and a leucine, respectively, at this position, whereas Seq2a, Seq2b, and Seq2c have alanines at this site. We know that at least two replacements have occurred. The ML estimations of the individual rates for the two subfamilies give a slow rate to subfamily 2, because there is no direct evidence of a replacement there. Subfamily 1, on the other hand, requires one replacement, and its rate of evolution is estimated to be fast. This means that the model will tend to assume that both of the replacements occurred in subfamily 1. This gives a more significant difference than methods that do not take the relationship between the two subfamilies into account, because they only use the single replacement. Here, then, this hypothetical site would be significant according to our test with the two subfamilies considered together ($U = 4.07, P \approx 0.044$), but barely insignificant if the two were analyzed separately ($U = 3.17, P \approx 0.075$).

The obvious next question is: Which significance value should be used in these tests? Often a value of $P = 0.05$ is chosen. The problem here is that multiple tests are being performed. For an alignment of length l, this means that $\approx 0.05l$ sites will be significant just by chance when $P = 0.05$ is used. To correct for this multiple testing, a stricter P value should be chosen, depending on the number of sequences under analysis. For small data sets with relatively few sequences, power is low, so a very strict significance level will yield few results.

Taking all of this into account, we recommend that $0.05l$ be used to estimate the expected number of sites with $P \leq 0.05$ by chance alone. This expectation can then be compared with the

observed number of such sites to assess the number of positions with significantly different rates. Those sites with very significant rate changes may stand out among the others. In turn, the entire set of potentially significant sites can be further evaluated for their importance against independent structural and functional data for their proteins (8, 19). A combined approach that uses both perspectives is illustrated below for Myc proteins.

LRT for Conserved Sites. As outlined in Fig. 1, one can also test whether the rate for a site is different from the average for all protein positions. Such a test is done if H_1 is accepted (i.e., both subfamilies have the same rate). The test is done exactly like the rate shift test described above. The test statistic, U, again approximates a $\chi^2(1)$ distribution according to simulations under this hypothesis (H_2) (Fig. 8, which is published as supporting information on the PNAS web site, www.pnas.org). In many cases, the most interesting sites will be those that have a significantly slower rate than the average, because these positions are most likely to be those under the strongest selective constraints and of greatest functional importance.

Analysis of a Set of Myc Sequences

To illustrate the utility of these methods, a set of 38 proteins for c-Myc, N-Myc, and L-Myc (27, seven, and four sequences, respectively) was analyzed for sites with rate shifts and slower rates. These protein sequences included all of those used by Miyamoto and Freire (20), except for those of the intron-less retrogenes, viruses, and nonvertebrates. In addition, these 38 Myc sequences included the five new ones for eutherian mammals reported by Miyamoto *et al.* (21).

The alignment of the 38 Myc proteins was based on the conserved regions used by Miyamoto and Freire (20). The areas between their conserved regions (including the common boundary between exons 2 and 3) were aligned by using CLUSTAL W (22). The final length of the alignment was 583 positions, of which 285 had no gaps. In turn, 440 aligned positions did not have a gap in at least one sequence in both the c-Myc and N-Myc subfamilies. Thus, 440 positions were considered in our analysis of rate changes among sites (see below). All of the position numbers discussed in the following are relative to human c-Myc (23).

The phylogenetic tree used was that of Miyamoto and Freire (20), except that the interordinal relationships of eutherian mammals were fixed according to recent phylogenetic syntheses of both their molecular and morphological data (24–26). The branch lengths of this final phylogeny were optimized by ML using the JTT matrix and gamma distribution for rate heterogeneity among sites (27). The two protein subfamilies compared in our example were c-Myc and N-Myc, whereas L-Myc was used as their outgroup.

The final set of Myc sequences (with accession numbers), multiple sequence alignment, and phylogenetic tree are shown in Table 3, Fig. 9, and Fig. 10, respectively, which are published as supporting information on the PNAS web site.

Results and Interpretation. The c-*myc*, N-*myc*, and L-*myc* genes encode transcription factors that are important in the regulation of cell proliferation and differentiation (23, 28, 29). The c-*myc* gene is expressed in many tissues and developmental stages, whereas the expression of both N-*myc* and L-*myc* is reduced spatially and temporally. Mutations in these genes have been implicated in many human cancers (30). The proteins of all three genes can be divided into three primary regions: (*i*) the N-terminal domain (positions 1–144 of human c-Myc); (*ii*) the central region (positions 145–354); and (*iii*) the basic helix–loop–helix leucine zipper (bHLHZip) (positions 355–439) (Fig. 3). The N-terminal domain is essential for transcriptional regulation through both transactivation and repression, whereas the bHLHZip is critical for specific DNA binding. The central region includes sites for nonspecific DNA binding, nuclear localization, and additional phosphorylation.

Ninety one sites in our evolutionary analyses were defined by rates that were the same in c-Myc and N-Myc, but that were slower than the average for both proteins (Fig. 3). These positions map to different boxes and regions that are of known functional importance to Myc proteins (e.g., Myc boxes 1 and 2 that are critical for the modulation and integration of transcriptional regulation and for transcriptional repression, respectively) (28, 29). Furthermore, these 91 sites with slower rates are not randomly distributed across the three primary regions of Myc proteins (Table 1). This nonrandom pattern identifies the N-terminal domain and bHLHZip as conserved relative to the more variable central region. This greater conservation for the N-terminal domain and bHLHZip is not surprising, given that the primary functions of Myc proteins (transcriptional regulation and specific DNA binding) depend on these two regions.

Our LRTs identify 49 sites with significant rate differences at the level of 5% (Table 2). Because the alignment has 440 positions that could show rate shifts, ≈22 such sites are expected by chance alone (440×0.05). This indicates that there are ≈27 more sites with significant rate differences than expected. At the 1% level, there are 16 sites with significant rate changes, which again is more than expected by chance (4 or 440×0.01). This illustrates the value of using significance levels that are easy to interpret.

These 49 sites are not randomly distributed across the three primary regions of c-Myc and N-Myc (Table 1). Rather, there are relatively too many and too few sites with significant rate changes in the N-terminal domain versus bHLHZip, respectively (Fig. 3). These results agree with those of Dermitzakis and Clark (31), who showed that the transactivation domains (but not the DNA binding regions) of different transcription factors from the MyoD and Mef2 gene families were characterized by variable rates between duplicate genes. Thus, these results are consistent with their hypothesis that the domains for transcriptional regulation may be more important for the functional differences among transcription factors than their DNA binding regions.

Furthermore, these 49 sites pinpoint more specific boxes and other regions, as of greatest potential importance for the known functional differences between c-Myc and N-Myc. For example, Prendergast (28) hypothesized that positions 107–130 may underlie the functional differences in transactivation and transformation that distinguish c-Myc from N-Myc. Our results identify nine sites with significant rate differences that map to this region (Fig. 3). These nine sites can now serve as specific targets in experiments with site directed mutagenesis for their effects on transactivation and transformation (32).

Comparison to Earlier Work. The ranking of sites by their significance values differs from that derived from the Bayesian method (15) (Table 2). This is primarily because all replacements are effectively equally weighted in this method. Even though the JTT matrix is used to infer expected numbers of replacements, all replacements are treated equally thereafter. Any method based on comparisons of replacement counts will suffer from this problem. It is not only the number of replacements, but also the nature of those replacements that is important in estimating the significance of an observation.

To illustrate this point, consider position 414 (Fig. 3). It has leucine in all N-Myc sequences, whereas the c-Myc sequences have isoleucine, leucine, threonine, and valine. The latter four amino acids can quickly change between each other, as indicated in the JTT matrix. This means that a relatively slow evolutionary rate can explain the variation at this position in c-Myc. This reason is why this site has a low rank (47 overall and 25 among ungapped sites), compared with the Bayesian method (five among ungapped positions) (Table 2). The latter considers the

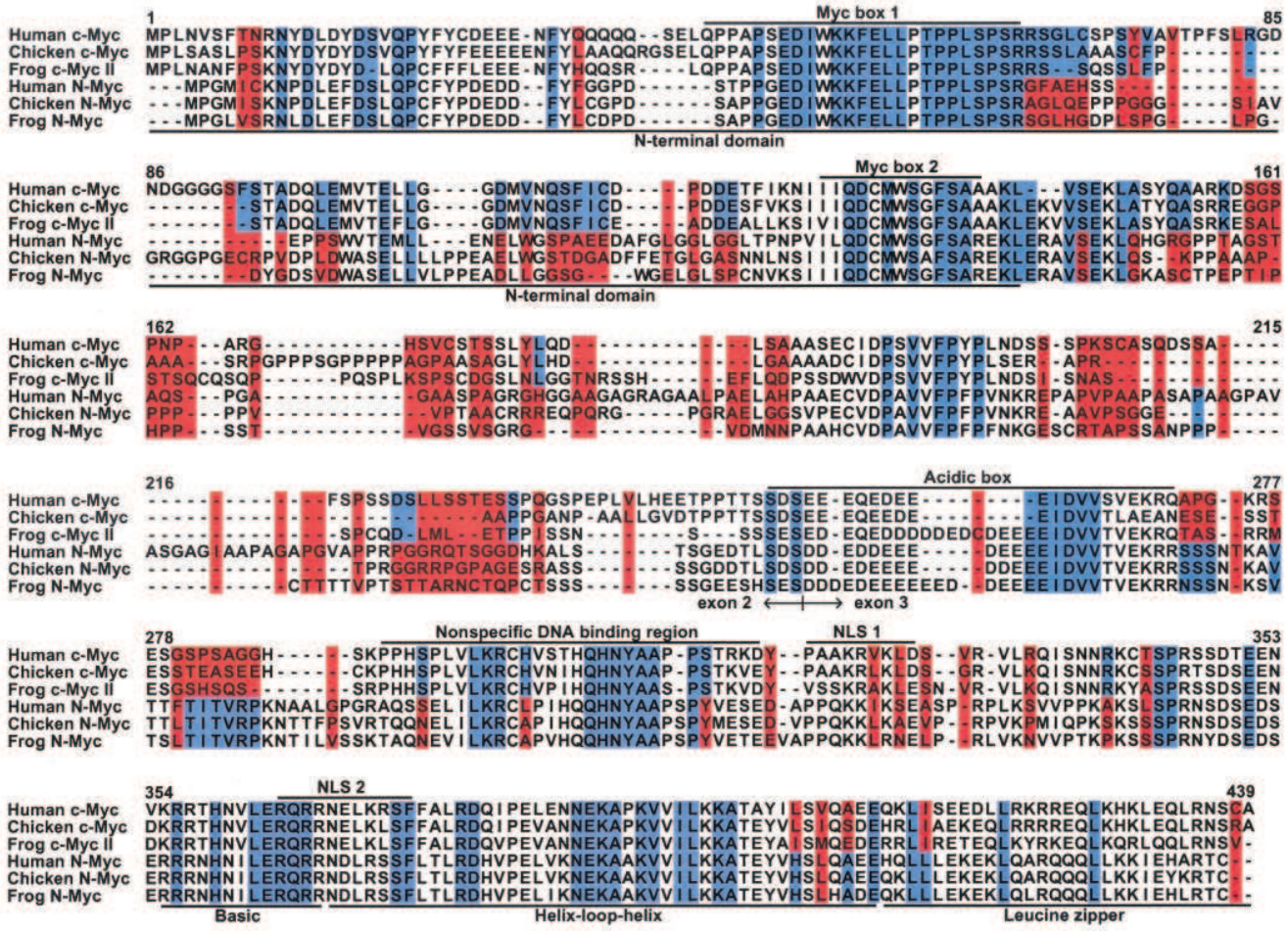


Fig. 3. Summary of results for the 38 Myc proteins, as represented by the c-Myc and N-Myc sequences for human (*Homo sapiens*), chicken (*Gallus gallus*), and frog (*Xenopus laevis*). The full alignment for all 38 Myc sequences is provided in Fig. 9, which is published as supporting information on the PNAS web site. Sites with both blue and red highlighting correspond to those with significant rate differences between the two subfamilies. In these cases, the blue and red distinguish the subfamily with the slower rate from the one with the faster rate, respectively. In turn, sites that are entirely blue or red highlight those with the same rate in the two subfamilies, but with significantly slower or faster rates than the average for all positions, respectively. In all cases, significance refers to the 5% level. Key structural and functional regions of the Myc proteins are labeled above and below the multiple sequence alignment (23, 28, 29). NLS, nuclear localization signal.

high number of replacements, but fails to acknowledge that these changes are all very common ones.

Another distinction between our LRT and the Bayesian method (15) is that ours does not assume anything about the distribution of rates across sites. Our distribution-free approach stands in contrast to the latter's reliance on the gamma distribution for the accommodation of rate heterogeneity among sites. Our approach also addresses a different question than the one asked by the Bayesian method. In our approach, the question is: Are the rates for a site the same in two subfamilies (Fig. 1)? In the alternative method, the question is instead: Are the rates for a site independent between two subfamilies? This latter distinction becomes particularly important, when the rates for a site are both fast but different in two subfamilies. Here, our approach is more likely to identify this site as significant, because it tests for rate differences, rather than rate correlations.

Table 1. Distributions of sites with significant rate shifts and equal, but significantly slower rates among the three primary regions of Myc proteins (Fig. 3)

	Rate shift sites			Equal, but slow rates		
Myc region	Significant sites	Other sites	Totals	Significant sites	Other sites	Totals
N-terminal domain	22 (15.0)	113 (120.0)	135	36 (26.3)	77 (86.7)	113
Central region	24 (24.7)	198 (197.3)	222	30 (46.1)	168 (151.9)	198
bHLHZip	3 (9.2)	80 (73.8)	83	25 (18.6)	55 (61.4)	80
Totals	49	391	440	91	300	391

These summaries are for the 440 positions that could show rate changes between c-Myc and N-Myc. The chi-square test for rate shift sites is significant at the level of 1.5% ($X^2 = 8.4$). The chi-square test for equal, but slow, rates is also significant ($X^2 = 14.8$, $P = 0.001$%). Expected counts are given in parentheses.

Table 2. The 49 positions with significant rate shifts between the c-Myc and N-Myc subfamilies

Rank	Position in human c-Myc	Slower subfamily	Significance	Bayesian rank	Rank	Position in human c-Myc	Slower subfamily	Significance	Bayesian rank	Rank	Position in human c-Myc	Slower subfamily	Significance	Bayesian rank
1	94	c-Myc	0.000040	—	18	284	N-Myc	0.010	4	35	146	c-Myc	0.027	11
2	221	c-Myc	0.00054	—	19	230	c-Myc	0.012	12	36	285	N-Myc	0.028	16
3	96	c-Myc	0.00062	—	20	157	c-Myc	0.012	—	37	153	c-Myc	0.031	—
4	272	N-Myc	0.00081	—	21	117	c-Myc	0.013	—	38	408	N-Myc	0.031	15
5	113	c-Myc	0.00085	2	22	283	N-Myc	0.014	6	39	121	c-Myc	0.035	13
6	68	c-Myc	0.0020	—	23	111	c-Myc	0.014	39	40	150	c-Myc	0.035	41
7	99	c-Myc	0.0025	1	24	277	N-Myc	0.015	3	41	83	c-Myc	0.036	—
8	114	c-Myc	0.0030	—	25	154	c-Myc	0.015	80	42	73	c-Myc	0.037	—
9	286	N-Myc	0.0040	—	26	293	c-Myc	0.017	10	43	282	N-Myc	0.037	24
10	66	c-Myc	0.0044	—	27	222	c-Myc	0.017	—	44	340	c-Myc	0.040	17
11	178	c-Myc	0.0071	—	28	301	c-Myc	0.017	9	45	404	N-Myc	0.041	21
12	273	N-Myc	0.0072	—	29	100	c-Myc	0.018	19	46	281	N-Myc	0.043	8
13	75	c-Myc	0.0077	—	30	122	c-Myc	0.022	14	47	414	N-Myc	0.048	5
14	69	c-Myc	0.0082	—	31	214	N-Myc	0.023	—	48	67	c-Myc	0.049	—
15	274	N-Myc	0.0084	30	32	109	c-Myc	0.024	18	49	93	c-Myc	0.050	—
16	116	c-Myc	0.0091	—	33	314	c-Myc	0.026	47					
17	70	c-Myc	0.0099	—	34	115	c-Myc	0.027	—					

These 49 positions are ranked according to their *P* values. At a significance level of 0.05, approximately 22 significant sites are expected by chance. Thus, approximately 22 of these sites may be random occurrences. Bayesian rank refers to the results from the Bayesian analysis of these Myc sequences (15). As this method cannot accommodate sites with any gaps or unknown positions, several sites in our analysis (marked by dashes) were excluded by the former.

The above analysis of rate shift sites by the Bayesian method is based on the fast approximate procedure that is now available in the DIVERGE (version 1.04) computer program (ref. 15; http://xgu1.zool.iastate.edu/doc.html). Recently, Gu (2) presented a full Bayesian alternative for such analyses under the JTT model, thereby correcting for the differences in replacement rates among amino acids. Currently, a finished computer program for general distribution is not available for this alternative, although one is expected soon (X. Gu, personal communication). Furthermore, this alternative still differs from our LRT in its dependence on the gamma distribution to model rate heterogeneity among sites and in its testing of rate correlations, rather than rate differences. It also relies on an indirect procedure for its likelihood calculations of the whole tree, whereby these determinations are made for two extreme lengths of the internal branch that connect its two subtrees. These separate calculations are then linearly combined to obtain the final likelihood of the whole tree. In the *Appendix*, we present a direct procedure for the calculation of this likelihood.

Power Analysis. The power of the LRT for rate shift sites was examined with evolutionary simulations using the Myc phylogenetic tree (Fig. 10, which is published as supporting information on the PNAS web site). When the same rates were used at each site between the c-Myc and N-Myc subfamilies, 3.9% of the positions (of 1,000) were significant at the level of 5%. This number should ideally be 5%, but the χ^2 distribution of the test statistic is not exact as shown in Fig. 2. When the N-Myc rate at each position was doubled, but halved in c-Myc, for 500 sites, then vice versa for 500 additional sites, the percentage of significant positions of 1,000 increased to 10.4% for this rate ratio of four. When the rate ratio was then increased in this fashion to 16, 34% of the 1,000 sites were now significant. These power analyses indicate that quite high rate ratios are needed to detect rate shift sites between c-Myc and N-Myc. Because of the limited power of the test, it is particularly important to use as many sequences as possible for each subfamily. Furthermore, when using few sequences, evolutionary simulations are recommended, instead of the χ^2 approximation, for determining significance levels.

Phylogenetic Errors. To examine the effects of phylogenetic error on the detection of rate shift sites, the LRTs for the Myc sequences were repeated by using five additional phylogenies. The first two phylogenies were obtained from the neighbor-joining and protein parsimony analyses of the Myc sequences (33), whereas the next two were produced by rerooting the accepted tree at the basal nodes of the c-Myc and N-Myc subfamilies, respectively (20, 24–26). The fifth tree was generated by randomly rearranging the sequences within each subfamily of the accepted phylogeny.

The first two trees were relatively similar topologically to the accepted phylogeny, as they differed from the latter by symmetric differences of 20 and 21, respectively (33). In turn, the two rerooted trees varied from the accepted phylogeny only by their minimized versus maximized basal branches for the c-Myc versus N-Myc subfamilies (and vice versa), respectively. Forty one to 57 sites were significant according to these four alternatives, with 38 to 46 of these positions overlapping with the 49 for the accepted phylogeny (Tables 4 and 5, which are published as supporting information on the PNAS web site). These results indicate that the LRT for rate shift sites is relatively insensitive to rearrangements within the gene tree.

In contrast, the "random" alternative was quite different from the accepted phylogeny, as it varied from the latter by a symmetric difference of 60. One hundred and sixteen sites were significant according to this random alternative, with 41 of these positions overlapping with the 49 for the accepted phylogeny (Tables 4 and 5). These 116 sites document an increase in the frequency of false positives as valid groups are fragmented and additional parallel and back replacements are introduced into one subfamily versus another. This situation becomes most acute when one subfamily is varied for a site, but another is not. In this case, the addition of parallel and back replacements in the first subfamily exaggerates its rate for the site relative to that of the second. Correspondingly, the

chance of a significant rate difference between them (i.e., a false positive) becomes exaggerated, too.

Future Directions

A direct statistical test for rate shift sites is presented. It takes the known replacement pattern of amino acids into account through a suitable rate matrix and provides significance values that are easy to interpret. The method is shown to perform well on a protein family that has been studied before for its rate shift positions (15). These comparisons now await further analyses of this protein family with a new ML method (2).

One interesting area of future research is to study the heterogeneity of amino acid frequencies between subfamilies, in addition to their rate shift sites (15). This can be done both on a position-specific level and the whole sequence level. Such investigations would complement the use of rate changes to identify sites of potential functional significance (2).

To compensate for the limited power of the LRT, one can analyze groups of sites rather than individual positions. These groups should be defined *a priori* according to the structural and functional properties of the protein (e.g., the bHLHZip of Myc). By analyzing positions together, one can increase the power of the test, but at the cost of site specificity. Furthermore, the χ^2 method for testing significance becomes questionable in this case, as the small deviations at each site of the group will lead to a large overall departure from this idealized distribution. Thus, when sites are grouped, evolutionary simulations will provide a superior test of significance. Finally, by considering the entire protein as the group, one can test for rate shifts at the whole sequence level in a manner that is analogous to the θ coefficient in the Bayesian method (2, 15).

Availability of Computer Programs

The programs of this study are available at www.daimi.au.dk/~compbio/rateshift. These programs can analyze both protein and nucleic acid sequences for rate shift sites and conserved positions.

Appendix: A Bayesian Approach for the Identification of Rate Shift Sites

If the rates among sites are assumed or known to follow some distribution, e.g., a gamma distribution, this information can be used as a prior in a Bayesian analysis of rate shift positions.

The whole tree is designated T, whereas the subtrees for the two subfamilies under investigation are denoted T_1 and T_2, respectively. Note that T_1 and T_2 include the branches that connect their most recent common ancestors to the root of the whole tree. Thus, T can be formed directly by joining T_1 and T_2. The inclusion of these basal branches with their subtrees eliminates the need for separate likelihood calculations, as in the whole tree procedure of the new Bayesian method (2). For a given site, let X then denote the amino acid configuration for all sequences, whereas X_1 and X_2 represent the configurations in the two respective subfamilies.

We can calculate the probability of the data, $P_0(X)$, given that the rates for a site are independent between the two subfamilies. Here, $\lambda_1 \perp \lambda_2$ is used to symbolize that the two rates are independent, with ϕ referring to their prior distributions. In the equations below, x represents the amino acids at the root of the whole tree:

$$\begin{aligned} P_0(X) &= P(X|T, \lambda_1 \perp \lambda_2) \\ &= \int_{\lambda_1=0}^{\infty}\int_{\lambda_2=0}^{\infty} P(X|\lambda_1, \lambda_2, T)\phi(\lambda_1)\phi(\lambda_2)d\lambda_1 d\lambda_2 \\ &= \int_{\lambda_1=0}^{\infty}\int_{\lambda_2=0}^{\infty} \sum_x [P(X_1|x, \lambda_1, T_1)P(X_2|x, \lambda_2, T_2)P(x)] \\ &\quad \cdot \phi(\lambda_1)\phi(\lambda_2)d\lambda_1 d\lambda_2 \\ &= \sum_x P(X_1|T_1, x)P(X_2|T_2, x)P(x). \end{aligned}$$

Notice that no two-dimensional integration is necessary. The integrals can be computed numerically.

We can also calculate the probability of the data, $P_1(X)$, given that the two rates are equal.

$$P_1(X) = P(X|T, \lambda_1 = \lambda_2) = \int_{\lambda=0}^{\infty} P(X|\lambda_1 = \lambda_2 = \lambda, T)\phi(\lambda)d\lambda.$$

The two hypotheses can now be compared by comparing their two probabilities. This can be expressed as the posterior probability that the rates are independent at the site under investigation.

$$\begin{aligned} P(\lambda_1 \perp \lambda_2|T, X) &= \frac{P(X|T, \lambda_1 \perp \lambda_2)P(\lambda_1 \perp \lambda_2|T)}{P(X|T)} \\ &= \frac{P_0(X)P(\lambda_1 \perp \lambda_2)}{P_1(X)(1 - P(\lambda_1 \perp \lambda_2)) + P_0(X)P(\lambda_1 \perp \lambda_2)}. \end{aligned}$$

A prior probability for the rates being independent, $P(\lambda_1 \perp \lambda_2)$, is needed. This probability can be estimated as by Gu (15), who uses θ (the coefficient of functional divergence) as the prior. Using this, we can obtain the probability that the rates at a given site are independent between the two subfamilies.

We thank X. Gu, M. R. Tennant, and an anonymous reviewer for their comments about our research and X. Gu for the use of his program. This research was supported by funds from the Hede Nielsen Family Foundation to B.K. and by the assistance of the Department of Zoology, University of Florida.

1. Bork, P. & Koonin, E. V. (1998) *Nat. Genet.* **18,** 313–318.
2. Gu, X. (2001) *Mol. Biol. Evol.* **18,** 453–464.
3. Thornton, J. M. (2001) *Science* **292,** 2095–2097.
4. Yang, Z. & Bielawski, J. P. (2000) *Trends Ecol. Evol.* **15,** 496–503.
5. Suzuki, Y., Gojobori, T. & Nei, M. (2001) *Bioinformatics* **17,** 660–661.
6. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.).
7. Graur, D. & Li, W.-H. (2000) *Fundamentals of Molecular Evolution* (Sinauer, Sunderland, MA), 2nd Ed.
8. Gaucher, E. A., Miyamoto, M. M. & Benner, S. A. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 548–552.
9. Wang, Y. & Gu, X. (2001) *Genetics* **158,** 1311–1320.
10. Fitch, W. M. & Markowitz, E. (1970) *Biochem. Genet.* **4,** 579–593.
11. Pollock, D. D., Taylor, W. R. & Goldman, N. (1999) *J. Mol. Biol.* **287,** 187–198.
12. Tuffley, C. & Steel, M. (1998) *Math. Biosci.* **147,** 63–91.
13. Galtier, N. (2001) *Mol. Biol. Evol.* **18,** 866–873.
14. Moreira, D., Le Guyader, H. & Philippe, H. (1999) *Mol. Biol. Evol.* **16,** 234–245.
15. Gu, X. (1999) *Mol. Biol. Evol.* **16,** 1664–1674.
16. Huelsenbeck, J. P. & Rannala, B. (1997) *Science* **276,** 227–232.
17. Felsenstein, J. (1981) *J. Mol. Evol.* **17,** 368–376.
18. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Comput. Appl. Biosci.* **8,** 275–282.
19. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001) *J. Mol. Biol.* **307,** 1487–1502.
20. Miyamoto, M. M. & Freire, N. P. (2000) *Mol. Phylogenet. Evol.* **16,** 475–481.
21. Miyamoto, M. M., Porter, C. A. & Goodman, M. (2000) *Syst. Biol.* **49,** 501–514.
22. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
23. Hesketh, R. (1997) *The Oncogene and Tumor Supressor Gene Factsbook* (Academic, San Diego), 2nd Ed.
24. Liu, F.-G. R., Miyamoto, M. M., Freire, N. P., Ong, P. Q., Tennant, M. R., Young, T. S. & Gugel, K. F. (2001) *Science* **291,** 1786–1789.
25. Madsen, O., Scally, M., Douady, C. J., Kao, D. J., DeBry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., de Jong, W. W. & Springer, M. S. (2001) *Nature (London)* **409,** 610–614.
26. Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A. & O'Brien, S. J. (2001) *Nature (London)* **409,** 614–618.
27. Yang, Z. (1997) *Comput. Appl. Biosci.* **13,** 555–556.
28. Prendergast, G. C. (1997) in *Oncogenes as Transcriptional Regulators: Volume 1, Retroviral Oncogenes*, eds. Yaniv, M. & Ghysdael, J. (Birkhäuser, Basel), pp. 1–28.
29. Facchini, L. M. & Penn, L. Z. (1998) *FASEB J.* **12,** 633–651.
30. Nesbit, C. E., Tersak, J. M. & Prochownik, E. V. (1999) *Oncogene* **18,** 3004–3016.
31. Dermitzakis, E. T. & Clark, A. G. (2001) *Mol. Biol. Evol.* **18,** 557–562.
32. Golding, G. B. & Dean, A. M. (1998) *Mol. Biol. Evol.* **15,** 355–369.
33. Swofford, D. L. (1998) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)* (Sinauer, Sunderland, MA), Version 4.0.

Chapter 6

Evolutionary Rate Analyses of Functional Divergence and Conservation among Proteins

Manuscript by Bjarne Knudsen, Michael M. Miyamoto, Philip J. Laipis, and David N. Silverman.

Submitted to *Proc. Natl. Acad. Sci. USA*.

Figures and tables that are reffered to as being published as supporting information on the PNAS web site are included in the end of the manuscript.

Abbreviations: AE1, Cl^-/HCO_3^- anion exchanger; CA, carbonic anhydrase; GSH, glutathione; JTT, Jones, Taylor, and Thornton; LRT, likelihood ratio test; ML, maximum likelihood.

Abstract

Selective and functional constraints on proteins limit their evolutionary rates at specific sites. These constraints allow for the interpretation of conserved residues and sites with a rate change as those most likely underlying the functional similarities and differences among protein subfamilies, respectively. This study describes new likelihood ratio tests (LRTs) that complement existing ones for the identification of both conserved and rate change sites. These identifications are validated by the recovery of residues that are known from existing biochemical and structural information to be critical for the functional similarities and differences among carbonic anhydrases (CAs). In combination with this other information, these LRTs also support an antioxidant defense role for the enigmatic CA III. As illustrated by the CAs, these LRTs, in combination with other biological evidence, offer a powerful and cost effective approach for testing hypotheses, making predictions, and designing experiments in protein functional studies.

Introduction

Functionally important sites and regions of biological sequences are under strong purifying selection and therefore evolve slowly according to the rule of functional constraint in molecular evolution (Kimura, 1983; Li, 1997). This widely acknowledged rule forms the foundation of many comparative approaches for the functional analysis of protein and nucleic acid sequences (Gaucher *et al.*, 2002; Hughes, 1999; Landgraf *et al.*, 2001; Nei and Kumar, 2000). For example, conserved amino acids are routinely interpreted as those that are most likely critical for an enzyme's function. Furthermore, site-specific rate changes among different proteins are often taken as evidence that these positions most likely underlie the functional differences among their subfamilies. When integrated with biochemical, structural, and other biological information, these rate tests of functionally important sites offer a powerful and cost effective way to generate new hypotheses and experiments for testing protein function (Golding and Dean, 1998).

Protein functional divergence is related to gene duplications and major speciations (Ohno, 1970; Nei *et al.*, 1997; Hughes, 1999; Lynch and Force, 2000; Gaucher *et al.*, 2002). In particular, gene duplications provide the additional coding and regulatory sequences for the origins of new protein functions and sub-specializations of their ancestral roles. Correspondingly, most rate tests of functional divergence focus on the subfamilies from duplications and major speciations (Gu, 1999, 2001; Knudsen and Miyamoto, 2001). For relatively recent events, these tests usually rely on comparisons of the nonsynonymous (replacement) to synonymous (silent) substitution rates for coding DNAs (Hughes, 1999; Nei and Kumar, 2000). However, this approach is limited by the relatively rapid saturation of the synonymous substitutions by multiple hits. Thus, studies of older protein subfamilies usually rely on the replacement rates alone to iden-

tify sites that are most likely responsible for their divergent versus conserved functions (Gaucher *et al.*, 2002).

In his studies of protein functional divergence, Gu (1999, 2001) recognized two patterns of rate change following a gene duplication or major speciation. In type I divergence, the site-specific rate increases after the duplication or speciation, then decreases to a new level different from that prior to the event. This new rate is a reflection of varying functional constraints on the site and is therefore expected to differ between the subfamilies. In type II divergence, the site-specific rate returns to its original value following its initial increase. Here, the rate change is restricted to the stems of the protein subfamilies, since similar functional constraints are re-imposed on the site after the initial increase. A site that is fixed between subfamilies for chemically dissimilar amino acids provides the best example of type II divergence.

Type I and II divergences belong to a series of five nested hypotheses for rate change and conserved sites (Fig. 1). Recently, Knudsen and Miyamoto (2001) described new likelihood ratio tests (LRTs) for type I and conserved sites (H_{1a}, H_2, and H_3). They then compared their type I LRT to the approximate and full Bayesian approaches of Gu (1999, 2001) and quantile-based method of Gaucher *et al.* (2001). This study complements theirs by describing new LRTs for type II sites (H_0 and H_{1b}). As an illustration of its utility, this extended series of LRTs is applied to a set of carbonic anhydrases (CAs). These LRTs recover known sites of functional importance to CAs and support a distinct biological role for their enigmatic CA III.

LRTs for Rate Change and Conserved Sites

Type II Model. In type II divergence, the site-specific rate is accelerated in one or both of the subfamily stems (Fig. 1). This stem acceleration can be modeled either by multiplying the rates for the stems by a common factor, $a > 1$, or by inserting an additional branch of positive length at the root. These two approaches are identical when there are no prior constraints on the stem acceleration. In this study, the stem acceleration is modeled with the new factor, thereby yielding three parameters for the type I and/or II tests (a for the basal increase and r_I versus r_{II} for the site-specific rates in subfamilies I versus II, respectively).

Testing the Hypotheses. The maximum likelihood (ML) scores for the three rate change hypotheses [$L(H_0)$, $L(H_{1a})$, and $L(H_{1b})$] are each tested against the ML score for the hypothesis with a single rate for the entire tree [$L(H_2)$]. These evaluations are quantified by the U values of their LRTs (Knudsen and

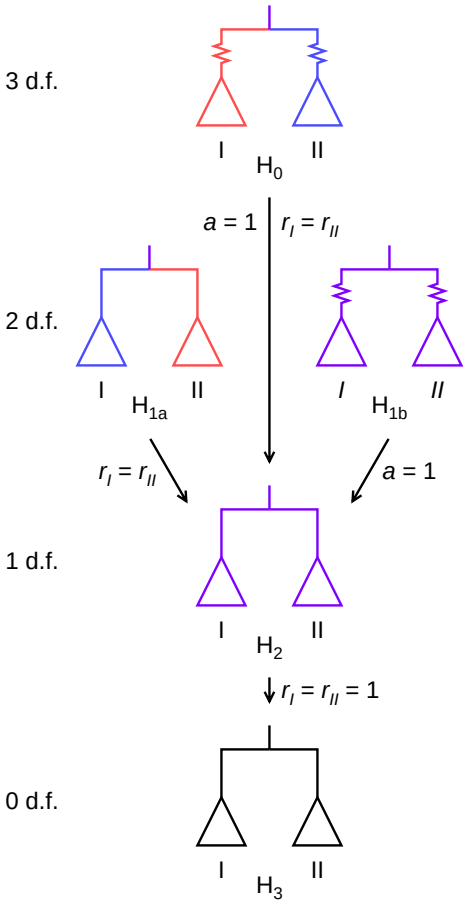


Figure 1: Nested hypotheses and LRTs for rate change and conserved sites. Triangles represent the two protein subfamilies from a gene duplication or major speciation. Red and blue signify different subfamily rates and therefore a type I site. Zigzags reflect an accelerated rate at the base of the tree and therefore a type II site. Black and purple denote a site with a single rate that is equal or unequal to the overall average for the protein, respectively. The five hypotheses are for type I and II sites (H_0), type I or II positions (H_{1a} and H_{1b}, respectively), and those with a single rate that is unequal or equal to the overall average (H_2 versus H_3). The numbers of free parameters for each hypothesis (degrees of freedom) are listed to the left. These parameters include the stem acceleration factor (a) and separate rates for the two subfamilies (r_I versus r_{II}). Arrows indicate which nested hypotheses are directly compared in the LRTs and which parameters are reduced from the more complex to simpler models.

Miyamoto, 2001):

$$U_0 = -2 \log \frac{L(H_2)}{L(H_0)}$$
$$U_{1a} = -2 \log \frac{L(H_2)}{L(H_{1a})}$$
$$U_{1b} = -2 \log \frac{L(H_2)}{L(H_{1b})}.$$

U_0 and U_{1b} are strongly influenced by a stem replacement. Thus, neither statistic closely follows a χ^2 or related distribution, since neither approximates a sum of squared normally distributed values. Consequently, the 5% significance levels for U_0, U_{1a}, and U_{1b} ($U_0^{5\%}$,$U_{1a}^{5\%}$, and $U_{1b}^{5\%}$, respectively) are found with simulations (see below). The 5% cutoffs from the simulations are compared to the observed U values (Huelsenbeck and Rannala, 1997):

$$\Delta U_0 = U_0 - U_0^{5\%}$$
$$\Delta U_{1a} = U_{1a} - U_{1a}^{5\%}$$
$$\Delta U_{1b} = U_{1b} - U_{1b}^{5\%}.$$

A positive ΔU indicates that the corresponding rate change hypothesis is a significantly better explanation of the data than is H_2. If ΔU is positive for more than one rate change hypothesis, then the one with the greatest difference is retained for the site in question. If no ΔU is positive, then the rate for this site is accepted as constant throughout the tree. The constant rate can then be tested against the average for the entire protein to determine whether this site is evolving significantly slow or fast. This test is done with the following LRT (Knudsen and Miyamoto, 2001):

$$U_2 = -2 \log \frac{L(H_3)}{L(H_2)}.$$

Although U_2 approximately follows a χ^2 distribution, simulations are again recommended for the determination of its 5% cutoffs, since they are more reliable.

In this series of LRTs, H_0 is directly compared to H_2, even though H_{1a} and H_{1b} are also nested in the former hypothesis (Fig. 1). Thus, alternatively, H_0 could be directly evaluated against H_{1a} and H_{1b}, rather than H_2. However, this alternative sequence is not preferred, since their 5% cutoffs are determined with simulations. Direct testing of H_0 against H_{1a} and H_{1b} requires the specification of r_I and r_{II} or a in their respective simulations. By comparing instead H_0 to H_2, these extra parameterizations are avoided.

Evaluating Multiple Subfamilies. The type I and/or II LRTs specifically test for rate changes at the root of the phylogeny for two subfamilies (Fig. 1). By analogy, these LRTs can be extended to the stems of multiple subfamilies (Fig. 2). Given multiple subfamilies, ML scores are separately calculated under

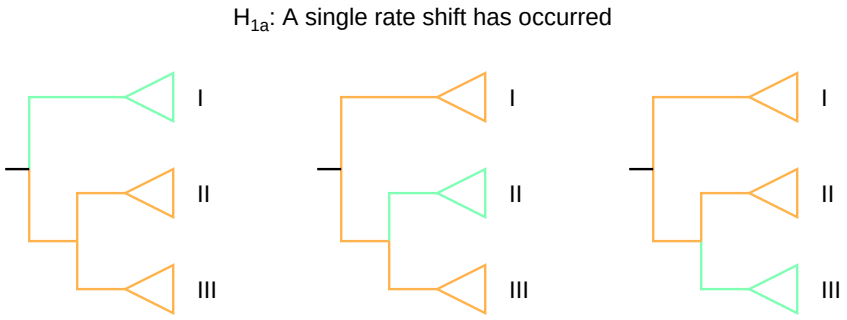


Figure 2: Choosing which stem to test for a rate change when more than two subfamilies are represented. This selection is illustrated for three protein subfamilies and H_{1a} (Fig. 1). There are three stems to test for a type I change given three subfamilies. The one stem with the greatest ML score is chosen for the LRT of H_{1a} to H_2. In the same way, single stems are selected for the LRTs of H_0 and H_{1b}.

H_0, H_{1a}, and H_{1b} for a type I and/or II change along each stem. The one stem with the greatest ML score for H_0, H_{1a}, or H_{1b} is retained for further testing of that rate change hypothesis with U and ΔU. As before, if ΔU is positive for more than one rate change hypothesis, then the one with the greatest difference is accepted as the best explanation for the site in question.

The final selection of H_0, H_{1a}, or H_{1b} for a site with more than one significant ΔU does not inflate the overall significance of the accepted rate change hypothesis, since this decision is made after the LRTs are completed. In contrast, the selection of which stem to test given more than two subfamilies forms the basis of the LRTs themselves and is therefore vulnerable to the effects of multiple testing. This source of inflated significance can be readily corrected by establishing the 5% cutoffs in the simulations with only the best ML scores for the multiple subfamilies.

Type II LRTs: Power Analyses and Phylogenetic Errors. A site with a fixed amino acid difference between two subfamilies provides the clearest evidence of type II divergence (Gu, 2001). The probability of obtaining this fixed difference by chance (P) can be approximated under the Jukes-Cantor model that assumes equal replacement rates among all amino acids (Appendix). A highly significant P will reflect a large difference in the probability of an amino acid change within versus between the subfamilies. Thus, P can serve as a measure of the power available to type II LRTs.

The expected power of the type II LRTs is illustrated with the CAs (Table 1). By varying the relative lengths of the stem for one CA subfamily (l_0) versus their total phylogeny (l_t), these tests show that power increases with l_t and decreases with l_0. Thus, power is maximized when the opportunity for a stem

Table 1: Power analyses of the type II LRTs

(A)	$2l_t$(CA)	l_t(CA)	$0.5l_t$(CA)
$2l_0$(CA I)	0.72%	2.51%	7.32%
l_0(CA I)	0.35%	1.19%	3.34%
$0.5l_0$(CA I)	0.17%	0.58%	1.60%

(B)	l_0	$\alpha = 1.15$	$\alpha = \infty$
CA I	0.2140	1.19%	0.82%
CA II	0.1332	0.72%	0.49%
CA III	0.2953	1.67%	1.18%

Percentages are the probabilities of observing by chance (P) a fixed amino acid difference between one CA subfamily and the two others (Appendix). (A) Power analysis for CA I versus II and III given varying lengths of the stem for the former subfamily [l_0(CA I) = 0.2140 replacements/site] versus total tree [l_t(CA) = 3.3730 replacements/site]. (B) Power analysis for each CA subfamily, with and without a gamma distribution (α = 1.15 and ∞, respectively). The 1.15 estimate is the ML value for the CAs under the JTT model with a gamma correction.

replacement is small, but that for a change within subfamilies is large. This conclusion becomes important when many sites of the protein have evolved as type II positions. In these cases, phylogenetic methods will overestimate the lengths of the stems and thereby lead to underestimates of the actual numbers of type II sites. In contrast, this conclusion also indicates that an obvious strategy to reduce l_0 and thereby increase power is to sample species from near the stems of each subfamily.

An alternative strategy to increase power in the type II LRTs is to include an appropriate replacement matrix for unequal rates among amino acids [e.g., the Jones, Taylor, and Thornton (JTT) model] (Jones *et al.*, 1992). For CAs, P under the Jukes-Cantor model is $\sim$ 1.19% for a site with any fixed amino acid difference between CA I versus II and III (Appendix). In contrast, P under the JTT model will vary from $<$ 0.01% for P versus K to $\sim$ 1.79% for H versus Y of CA I versus II and III, respectively. These extremes agree with the premise that radical amino acid differences are more informative about the functional divergence of proteins than are conservative ones (Gu, 2001; Livingstone and Barton, 1996). By incorporating an appropriate unequal rate matrix in their LRTs, fixed radical differences can contribute even stronger evidence to the recognition of type II sites.

The power analyses further illustrate that rate heterogeneity among sites increases the chances of a type II position (Table 1). Under a gamma process with α = 1.15, a relatively large proportion of sites is slowly evolving (Yang, 1996). For these slow sites, any chance replacement in the stems is less likely to be followed by a subsequent change within the subfamilies relative to a position

evolving at or faster than the average rate. Thus, a stem replacement for a slow site is more likely to be preserved as a fixed difference among subfamilies. This conclusion reinforces the overall conservative nature of type II sites and their corresponding potential as indicators of protein function (Gu, 2001).

In general, phylogenetic errors are not expected to diminish greatly the power of the type II LRTs, since their strongest support is obtained from fixed amino acid differences among subfamilies. By definition, these fixed differences will remain, even if lineages are shifted within subfamilies and the latter are rearranged (Knudsen and Miyamoto, 2001). However, the power analysis with l_0 and l_t serves as a reminder of the importance of accurate branch lengths, particularly for the stems (Table 1). In this regard, phylogenetic errors may indirectly affect the type II LRTs by influencing the branch length estimations.

Type II Bayesian Test. The only other ML test for type II sites is the Bayesian method that begins in effect with the separation of stems and subfamilies (Gu, 2001). Rates are then estimated for both under a gamma distribution, followed by the calculations of the site-specific posterior probabilities that the separate estimates for the stems versus subfamilies are independent. Thus, the Bayesian method differs from the type II LRTs that test for $a > 1$ given no prior assumptions about the distributions of this parameter or the rates. Consequently, the type II LRTs are less likely to be sensitive to sampling effects, since they test for a stem acceleration with an extra parameter, rather than with separate rates for different phylogenetic regions that are defined by the available sequences.

The original article provided a description of the underlying theory and algorithm for type II Bayesian analysis (Gu, 2001). However, this method remains unavailable in DIVERGE or any other computer program for general use (3). Furthermore, its performance has not yet been evaluated nor has it been applied to real or simulated data. For these reasons, further comparisons of the type II LRTs and Bayesian method await the implementation, testing, and application of an available computer program for the latter.

Rate and Functional Analyses of CAs

CA I, II, and III. The CA family of ubiquitous enzymes catalyzes the reversible hydration of CO_2 to bicarbonate and protons in many fundamental biological processes (e.g., respiration and photosynthesis) (Lindskog, 1997; Chegwidden and Carter, 2000). At least 15 CAs are known in mammals, with each encoded by a different duplicate gene (Hewett-Emmett, 2000). Their diverse biological significance, expression patterns, and catalytic efficiencies, coupled with the successful development of a CA glaucoma drug, ensures that this family will remain a primary target for biochemical, physiological, structural, and pharmacological research.

Phylogenetic and linkage analyses indicate that CA I, II, and III are mono-

phyletic, even though their tissue expression patterns and CO_2 hydration rates vary almost as much as for the entire family (Hewett-Emmett and Tashian, 1996; Lindskog, 1997; Hewett-Emmett, 2000). CA II remains the primary reference for the family, because of its high catalytic efficiency and broad tissue expression. Its high CO_2 hydration rate is related to its conserved H64 that functions as a highly effective intra-molecular shuttle for proton transfer from the zinc catalytic center to the surrounding medium. CA II is also characterized by a set of five or six H and K residues at its N-terminus for Cl^-/HCO_3^- anion exchanger (AE1) binding for bicarbonate channeling from inside to outside the cell (Vince *et al.*, 2000). This set of basic residues may also contribute to the transfer of protons from H64 to the bulk solvent (Briganti *et al.*, 1997).

CA I and III are more restricted in their tissue expressions and their catalytic rates are $\sim 20\%$ and $< 1\%$ of that for CA II, respectively (Lindskog, 1997). In CA I, H64 is also conserved, but its set of basic residues at the N-terminus is greatly diminished (Briganti *et al.*, 1997; Vince *et al.*, 2000). Thus, CA I cannot bind to AE1 and its N-terminus probably does not participate in proton shuttling. In light of its reduced, but significant CO_2 hydration rate, CA I is thought to be a backup to CA II. Conversely, the physiological role of CA III remains unresolved. Its active site shows several important changes (e.g., K and R at position 64) and its N-terminal set of basic residues is reduced. In contrast, CA III evolves slowly and comprises $\sim 8\%$ and $\sim 25\%$ of the total soluble proteins in red skeletal muscle and adipose tissue, respectively. Collectively, these characteristics suggest a major biological role for CA III, which is distinct from the standard CA function of reversible CO_2 hydration.

CA Sequences, Phylogeny, and LRTs. To evaluate further their functional similarities and differences, all available sequences of CA I, II, and III were compiled (Tashian *et al.*, 1980; Eriksson and Liljas, 1993; Hewett-Emmett and Tashian, 1996; Benson *et al.*, 1999; Bairoch and Apweiler, 2000), aligned, and analyzed with the LRTs for rate change and conserved sites (Table 3 and Fig. 5, which are published as supporting information on the PNAS web site). The final alignment consisted of 260 positions for 11 CA I, 8 CA II, 6 CA III, and 5 CA Va and Vb (outgroup) sequences. The accepted phylogeny combined the CA gene tree from phylogenetic and linkage analyses with the eutherian mammal phylogeny from a recent molecular synthesis (Fig. 6, which is published as supporting information on the PNAS web site) (Hewett-Emmett and Tashian, 1996; Hewett-Emmett, 2000; Murphy *et al.*, 2001).

The standard approach in the LRTs is to measure the site-specific rates for subfamilies and stems against the local averages for their regions of the phylogeny (Knudsen and Miyamoto, 2001). By relying on relative rates, changes in the local averages due to varying demographic (e.g., population size) and mutation/repair factors are compensated, thereby allowing for functional interpretations of the significant sites (Gaucher *et al.*, 2002). However, this compensation also precludes the detection of significant sites after a local overall functional shift in a subfamily. In the case of duplicate genes, this limitation can be addressed by focusing on the shared species and nodes of the different subfamilies.

By emphasizing common lineages, the site-specific rates can be compared among subfamilies on a more absolute basis without the complication that they reflect different demographic and mutation/ repair factors. In this way, the sensitivity of the LRTs is enhanced, along with the functional interpretability of their significant sites.

In the case of the CAs, this advantage of shared lineages was accommodated by a constraint that required the distances from the root to each common species and node to be equal across subfamilies (Fig. 6). The branch lengths of the phylogeny were then estimated with ML under the JTT model with the gamma distribution (JTT + Γ). As illustrated by the phylogeny, this constraint did not impose a molecular clock on the analysis, since rates remained free to vary across different lineages.

Ten thousand sites were simulated under H_2 to establish the 5% cutoffs for H_0, H_{1a}, and H_{1b}. These simulations relied on the JTT + Γ model with α set to its ML value of 1.15 for the CAs. Ten thousand sites were also simulated under H_3 (i.e., with a single rate equal to the average for the entire protein) to determine the 5% significance for H_2. Finally, a set of 42 functionally important sites for CAs was defined according to the 36 positions of the active site and 6 basic residues of the N-terminus for AE1 binding and/or proton shuttling (Briganti *et al.*, 1997; Hewett-Emmett and Tashian, 1996; Vince *et al.*, 2000).

Significant Sites and Functional Interpretations. The LRTs for conserved sites recovered 47 positions that were evolving significantly slower than the overall average for the entire protein (Table 4, which is published as supporting information on the PNAS web site). These 47 conserved sites were over-represented among the 42 functionally important positions (Table 2) and included the 7 direct and indirect ligands to the zinc catalytic center of the active site (Q92, H94, H96, E117, H119, T199, and N244) (Hewett-Emmett and Tashian, 1996; Lindskog, 1997). Collectively, these results reconfirmed the rule of functional constraint that the biologically important sites of proteins are under the strongest purifying selection and thereby evolve the slowest.

The LRTs for rate change sites identified 32, 10, and two type I, II, and I/II positions, respectively (Fig. 3 and Table 4). The expected numbers of type I, II, and I/II sites were 11.8, 11.7, and 2.9 according to the simulations, respectively. Thus, almost three times as many type I sites were recovered as expected by chance. The 32 type I sites included position 64, with its fixed H in CA I and II versus variable R and K in CA III (Hewett-Emmett and Tashian, 1996; Lindskog, 1997).

Despite their near equal observed to expected frequencies, further analyses validated the importance of the type II sites to the greater understanding of CA functional divergence. The 44 type I and/or II sites were over-represented among the 42 functionally important positions (Table 2). However, this significance depended on the recognition of both divergences, since P became ~ 0.20 when the 10 type II sites were instead counted among the "other positions." Thus, type II divergence complements type I change and both processes must be considered in evolutionary studies of protein function (Gu, 2001).

Table 2: Frequency distributions of conserved (A) and type I and/or II (B) sites between the functionally important residues and all other positions of the CAs

(A)	Functionally important positions	Other positions	Totals
Conserved sites	13 (6.7)	34 (40.3)	47
Other sites	18 (24.3)	151 (144.7)	169
Totals	31	185	216

(B)	Functionally important positions	Other positions	Totals
Type I and/or II sites	11 (6.0)	33 (38.0)	44
Other sites	18 (23.0)	151 (146.0)	169
Totals	29	184	213

Type I and/or II sites are not included in Table (A), whereas conserved positions are conversely excluded from Table (B). Expected counts are given in parentheses. Functionally important positions refer to the 42 residues of the active site and N-terminus for AE1 binding and/or proton shuttling. The Chi-square tests for both contingency tables are significant ($P = 0.005$ and 0.017, respectively). Complete lists of the type I, II, and conserved sites are presented in Table 4, which is published as supporting information on the PNAS web site.

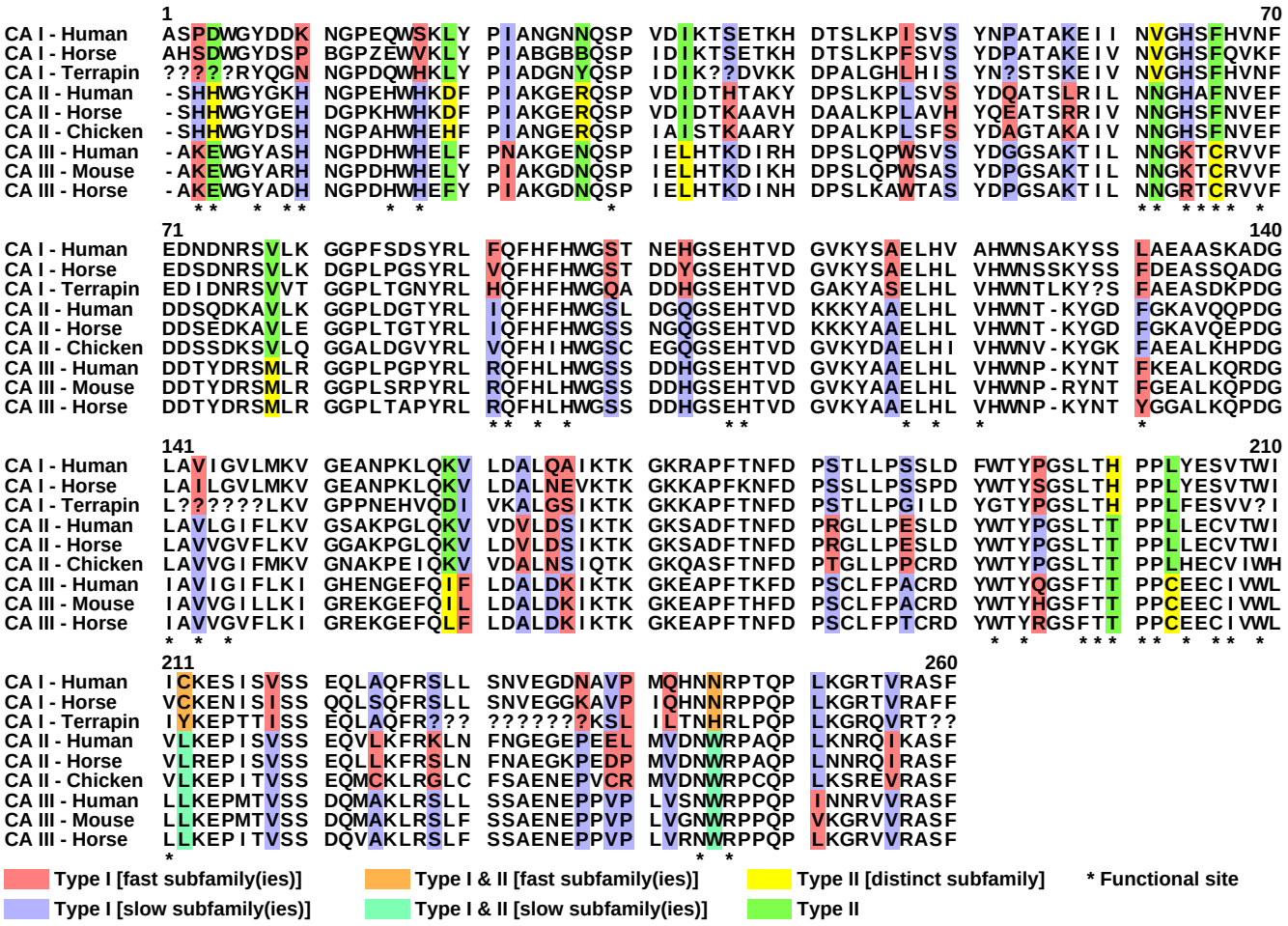


Figure 3: Multiple sequence alignment for representative CAs. The complete alignment for all CAs is presented in Fig. 5, which is published as supporting information on the PNAS web site. The 42 functionally important positions are marked. Conserved sites are not colored here, but are instead listed in Table 4, which is published as supporting information on the PNAS web site.

Of the 10 type II sites, four map to functionally important positions (Fig. 3). These four type II sites emphasize fixed radical differences among the subfamilies within two primary functional regions of CAs. For example, type II site 4 highlights the fixed radical difference of H in CA II against the acidic D and E in CA I and III. H4 of CA II is one of the 5 or 6 basic residues at its N-terminus for AE1 binding. A truncation mutant of CA II, which is missing its first five residues (and therefore H4), shows a measurable decrease in AE1 binding (Vince *et al.*, 2000). One obvious follow-up experiment is to re-test the AE1 binding of a site-directed CA II mutant after the replacement of its H4 with acidic D or E (Golding and Dean, 1998).

Available biochemical, mutagenic, and structural information define a series of sites that are of known importance to the common and unique functions of CAs. The ability of the LRTs to detect these known sites, as demonstrated both collectively (Table 2) and individually (e.g., H4, H64, and the 7 direct and indirect ligands to the zinc catalytic center), validates their utility for both testing existing hypotheses and generating new ones. In the case of CA III, these LRTs, in combination with biochemical, structural, and other bioinformatic information, support a distinct role for this engimatic isozyme.

In CA III, C183 and C188 are unique surface residues that are known binding targets for glutathione (GSH) (Figs. 3 and 4. CA III is among the first proteins to be glutathiolated during oxidative stress and a mutant cell line that is deficient for this isozyme is particularly sensitive to oxyradical insults (Chai *et al.*, 1994; Räisänen *et al.*, 1999). Thus, CA III is hypothesized to function as an oxyradical scavenger, whereby glutathiolation protects its C183 and C188 from irreversible oxidation. In support of an antioxidant defense role, the LRTs recover three rate change sites that lie next to or directly underneath C183 and C188 (positions 182/187 and 212, respectively). The conserved or nearly conserved residues of CA III at these rate change sites may contribute to the greater surface exposure and weaker acidic surroundings that enhance GSH binding at C188 over that at C183 (Mallis *et al.*, 2000).

Interestingly, S259, which lies close to C188 at the surface (Fig. 4), is a potential phosphorylation site according to NetPhos (an artificial neural network algorithm) (Blom *et al.*, 1999). The score for conserved S259 being a phosphorylation site is 0.995 out of 1.000 for every CA III, except for that of bovine (0.928). Thus, S259 phosphorylation/dephosphorylation may affect C188 glutathiolation or vice versa. In these ways, CA III may then function as a sensor of oxidative stress, whose activity is tied to the signalling pathways for antioxidant defense (Räisänen *et al.*, 1999; Chegwidden and Carter, 2000)

Availability of Computer Program

A computer program for the LRTs of rate change and conserved sites is available as a web server at www.daimi.au.dk/~compbio/LRTs.

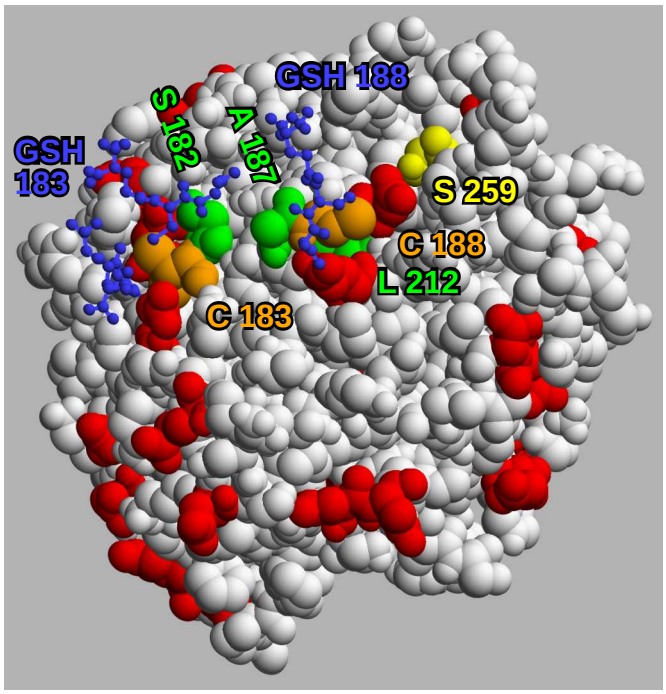


Figure 4: Spacefill model of the tertiary structure for rat CA III with bound GSH (PDB accession number 1FLJ), as rendered with RasMol (Sayle and Milner-White, 1995; Mallis *et al.*, 2000). This view focuses on the key residues around C183 and C188, with both alternative conformations of GSH183 shown. Acidic D and E residues are colored red.

Appendix: Power Calculations for the Type II LRTs

Assuming that all replacements occur with equal frequency and that none are hidden due to multiple changes, P can be approximated for a given rate (r) under the Jukes-Cantor model by the following equation (Jukes and Cantor, 1969):

$$P(r) \approx (1 - e^{-rl_0})e^{-r(l_t - l_0)} = e^{-r(l_t - l_0)} - e^{-rl_t}$$

The gamma distribution can be incorporated to accomodate site-to-site variation in rates (Yang, 1996):

$$P = \int_{r=0}^{\infty} \phi(r)P(r)\,dr \approx \int_{r=0}^{\infty} \phi(r)\left(e^{-r(l_t - l_0)} - e^{-rl_t}\right)dr.$$

The gamma density function, with parameter α, is calculated by:

$$\phi(x) = \frac{\alpha^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\alpha x},$$

thereby leading to:

$$P \approx \frac{\alpha^\alpha}{\Gamma(\alpha)} \int_{r=0}^{\infty} r^{\alpha-1}\left(e^{-r(l_t - l_0 + \alpha)} - e^{-r(l_t + \alpha)}\right)dr.$$

The integral can be calculated by:

$$\int_{x=0}^{\infty} x^a e^{-bx}\,dx = \Gamma(a+1)/b^{a+1},$$

thereby resulting in:

$$\begin{aligned} P &\approx \frac{\alpha^\alpha}{\Gamma(\alpha)}\left(\frac{\Gamma(\alpha)}{(l_t - l_0 + \alpha)^\alpha} - \frac{\Gamma(\alpha)}{(l_t + \alpha)^\alpha}\right) \\ &= \frac{\alpha^\alpha}{(l_t - l_0 + \alpha)^\alpha} - \frac{\alpha^\alpha}{(l_t + \alpha)^\alpha}. \end{aligned}$$

The inclusion of an appropriate replacement matrix for unequal rates among amino acids (e.g., the JTT model) greatly complicates these equations and thus their incorporation is not shown here. More importantly, the CA simulations illustrate that the above equations with their equal replacement rates among all amino acids are more than sufficiently accurate to justify the major conclusions of this study about power (Table 1).

Acknowledgements

We thank A. C. Harmon, R. L. Levine and M. R. Tennant for their comments about our research. This study was supported by grants from the National Institutes of Health [P.J.L. and D.N.S. (GM25154)] and by funds from the Department of Zoology, University of Florida.

References

Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28** (1), 45–48.

Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., Ouellette, B. F., Rapp, B. A. and Wheeler, D. L. (1999) Genbank. *Nucleic Acids Res.* **27** (1), 12–17.

Blom, N., Gammeltoft, S. and Brunak, S. (1999) Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294**, 1351–1362.

Briganti, F., Mangani, S., Orioli, P., Scozzafava, A., Vernaglione, G. and Supuran, C. T. (1997) Carbonic anhydrase activators: X-ray crystallographic and spectroscopic investigations for the interaction of isozymes I and II with histamine. *Biochemistry,* **36**, 10384–10392.

Chai, Y. C., Hendrich, S. and Thomas, J. A. (1994) Protein S-thiolation in hepatocytes stimulated by t-butyl hydroperoxide, menadione, and neutrophils. *Arch. Biochem. Biophys.* **310**, 264–272.

Chegwidden, W. R. and Carter, N. D. (2000) Introduction to the carbonic anhydrases. In Chegwidden, W. R., Carter, N. D. and Edwards, Y. H. (eds), *The Carbonic Anhydrases: New Horizons* pp. 13–28, Basel, Switzerland: Birkhäuser Verlag.

Eriksson, A. E. and Liljas, A. (1993) Refined structure of human carbonic anhydrase II at 2.0 Å resolution. *Proteins,* **16**, 29–42.

Gaucher, E. A., Gu, X., Miyamoto, M. M. and Benner, S. A. (2002) Detecting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* **27**, 315–321.

Gaucher, E. A., Miyamoto, M. M. and Benner, S. A. (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proc. Natl. Acad. Sci. USA,* **98** (2), 548–552.

Golding, G. B. and Dean, A. M. (1998) The structural basis of molecular adaptation. *Mol. Biol. Evol.* **15** (4), 355–369.

Gu, X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* **16** (12), 1664–1674.

Gu, X. (2001) Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* **18** (4), 453–464.

Hewett-Emmett, D. (2000) Evolution and distribution of the carbonic anhydrase gene families. In Chegwidden, W. R., Carter, N. D. and Edwards, Y. H. (eds), *The Carbonic Anhydrases: New Horizons* pp. 29–76, Basel, Switzerland: Birkhäuser Verlag.

Hewett-Emmett, D. and Tashian, R. E. (1996) Functional diversity, conservation, and convergence in the evolution of the alpha-, beta-, and gamma-carbonic anhydrase gene families. *Mol. Phylogenet. Evol.* **5** (1), 50–77.

Huelsenbeck, J. P. and Rannala, B. (1997) Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science,* **276** (5310), 227–232.

Hughes, A. L. (1999) *Adaptive Evolution of Genes and Genomes.* NY: Oxford Univ. Press.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8** (3), 275–282.

Jukes, T. H. and Cantor, C. R. (1969) Evolution of protein molecules. In Munro, H. N. (ed), *Mammalian Protein Metabolism* pp. 21–123, NY: Academic Press.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge, U.K.: Cambridge Univ. Press.

Knudsen, B. and Miyamoto, M. M. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl. Acad. Sci. USA,* **98** (25), 14512–14517.

Landgraf, R., Xenarios, I. and Eisenberg, D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307** (5), 1487–1502.

Li, W.-H. (1997) *Molecular Evolution.* Sunderland, MA: Sinauer.

Lindskog, S. (1997) Structure and mechanism of carbonic anhydrases. *Pharmacol. Ther.* **74** (1), 1–20.

Livingstone, C. D. and Barton, G. J. (1996) Identification of functional residues and secondary structure from protein multiple sequence alignment. *Methods Enzymol.* **266**, 497–512.

Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics,* **154**, 459–473.

Mallis, R. J., Poland, B. W., Chatterjee, T. K., Fisher, R. A., Darmawan, S., Honzatko, R. B. and Thomas, J. A. (2000) Crystal structure of S-glutathiolated carbonic anhydrase III. *FEBS Lett.* **482** (3), 237–241.

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W. and Springer, M. S. (2001) Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science,* **294** (5550), 2348–2351.

Nei, M., Gu, X. and Sitnikova, T. (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune systems. *Proc. Natl. Acad. Sci. USA,* **94**, 7799–7806.

Nei, M. and Kumar, S. (2000) *Molecular Evolution and Phylogenetics.* NY: Oxford Univ. Press.

Ohno, S. (1970) *Evolution by Gene Duplication.* Berlin: Springer-Verlag.

Räisänen, S. R., Lehenkari, P., Tasanen, M., Rahkila, P., Härkönen, P. L. and Väänänen, H. K. (1999) Carbonic anhydrase III protects cells from hydrogen peroxide-induced apoptosis. *FASEB J.* **13**, 513–522.

Sayle, R. and Milner-White, E. J. (1995) Rasmol: Biomolecular graphics for all. *Trends Biochem. Sci.* **20** (9), 374.

Tashian, R. E., Hewett-Emmet, D., Stoup, S. K., Goodman, M. and Yu, Y.-S. L. (1980) Evolution of structure and function in the carbonic anhydrase isozymes of mammals. In Bauer, C., Gros, G. and Bartels, H. (eds), *Biophysics and Physiology of Carbon Dioxide* pp. 165–176, Berlin: Springer-Verlag.

Vince, J. W., Carlsson, U. and Reithmeier, R. A. (2000) Localization of the $Cl^-/HCO3^-$ anion exchanger binding site to the amino-terminal region of carbonic anhydrase II. *Biochemistry,* **39** (44), 13344–13349.

Yang, Z. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11** (9), 367–372.

Web Table 3: List of species and accession numbers for the CAs

Sequence	Species	Source
CA I – Human	*Homo sapiens*	SwissProt (P00915)
CA I – Chimpanzee	*Pan troglodytes*	GenBank (JN0835)
CA I – Gorilla	*Gorilla gorilla*	GenBank (JN0836)
CA I – Rhesus macaque	*Macaca mulatta*	SwissProt (P00916)
CA I – Pig-tailed macaque	*Macaca nemestrina*	SwissProt (P35217)
CA I – Mouse	*Mus musculus*	SwissProt (P13634)
CA I – Rabbit	*Oryctolagus cuniculus*	SwissProt (P07452)
CA I – Horse	*Equus caballus*	SwissProt (P00917)
CA I – Cow	*Bos taurus*	Tashian *et al.* (1980)
CA I – Sheep	*Ovis aries*	SwissProt (P48282)
CA I – Diamondback terrapin	*Malaclemys terrapin*	Hewett-Emmett and Tashian (1996)
CA II – Human	*Homo sapiens*	SwissProt (P00918)
CA II – Mouse	*Mus musculus*	SwissProt (P00920)
CA II – Rat	*Rattus norvegicus*	SwissProt (P27139)
CA II – Rabbit	*Oryctolagus cuniculus*	SwissProt (P00919)
CA II – Horse	*Equus caballus*	GenBank (223999)
CA II – Cow	*Bos taurus*	SwissProt (P00921)
CA II – Sheep	*Ovis aries*	SwissProt (P00922)
CA II – Chicken	*Gallus gallus*	SwissProt (P07630)
CA III – Human	*Homo sapiens*	SwissProt (P07451)
CA III – Mouse	*Mus musculus*	SwissProt (P16015)
CA III – Rat	*Rattus norvegicus*	SwissProt (P14141)
CA III – Horse	*Equus caballus*	SwissProt (P07450)
CA III – Cow	*Bos taurus*	Eriksson and Liljas (1993)
CA III – Pig	*Sus scrofa*	From 7 EST records (see below)
CA Va – Human	*Homo sapiens*	SwissProt (P35218)
CA Va – Mouse	*Mus musculus*	SwissProt (P23589)
CA Va – Rat	*Rattus norvegicus*	SwissProt (P43165)
CA Vb – Human	*Homo sapiens*	SwissProt (Q9Y2D0)
CA Vb – Mouse	*Mus musculus*	SwissProt (Q9QZA0)

The CA III – Pig sequence is derived from an analysis of seven overlapping ESTs (GenBank accession numbers AJ301094, BF074991, AJ301337, AU059476, BI360558, AJ301207, and AJ301290). BF074991 varied from AJ301337 and AJ301094 by one silent difference. In turn, the two terminal nucleotides of AJ301290 were ignored, since they differed from the corresponding identical bases of BI360558 and AJ301290.

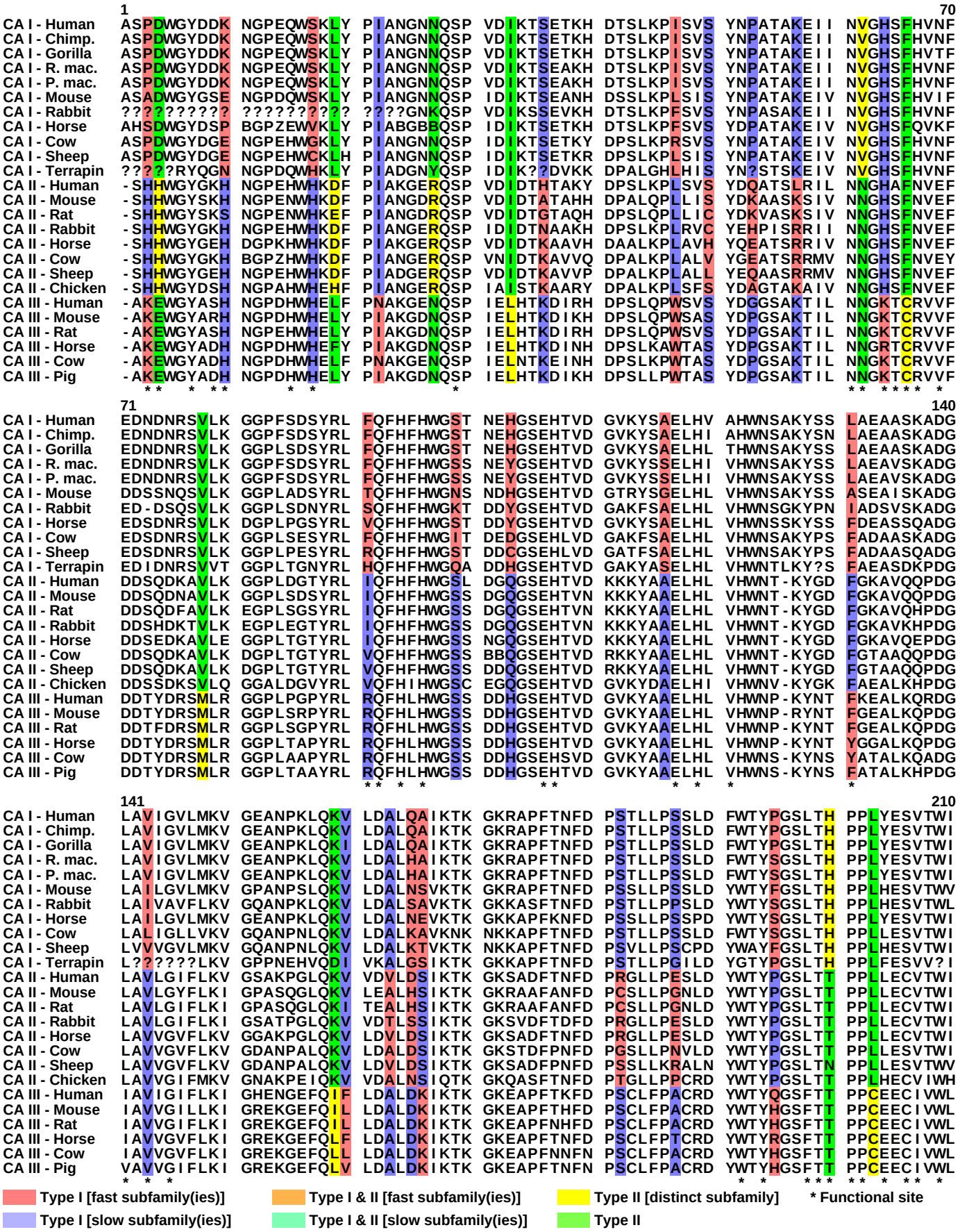
Web Table 4: Significant type I, II, and conserved sites for the CAs

(A) Type I and/or type II sites

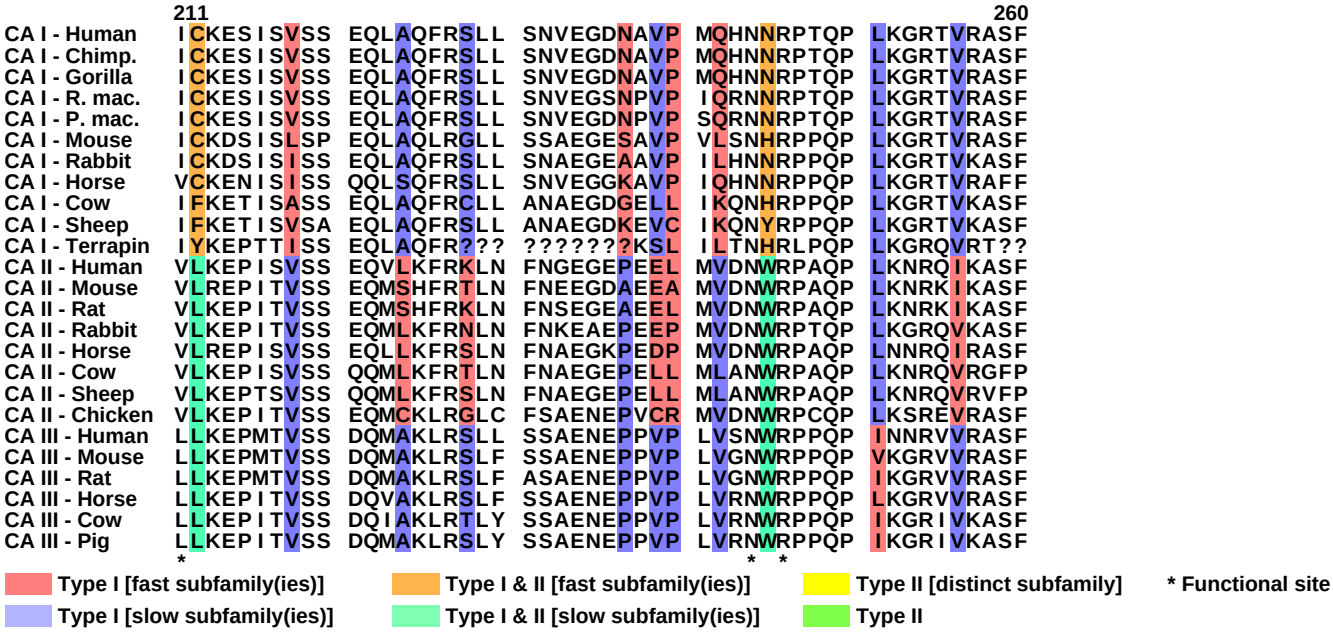
Position	Type	ΔU value	Slow or distinct CA subfamily	Position	Type	ΔU value	Slow or distinct CA subfamily	Position	Type	ΔU value	Slow or distinct CA subfamily
3	I	0.46	II	66	II	1.47	III	195	I	0.12	II
4	II	0.96	II	78	II	0.82	III	200	II	1.20	I
10	I	3.15	II, III	91	I	5.61	II, III	203	II	3.95	III
17	I	4.89	II, III	99	I	5.42	II, III	212	I & II	3.02	II, III
19	II	0.18	II	103	I	2.25	II, III	218	I	5.52	II, III
22	I	1.09	I, II	116	I	1.17	II, III	224	I	1.56	I, III
27	II	0.69	II	131	I	0.06	II	228	I	0.54	I, III
33	II	1.03	III	143	I	3.81	II, III	237	I	3.75	II, III
36	I	0.14	I, III	159	II	0.08	III	239	I	0.38	I, III
47	I	0.86	II	160	I	3.18	I, II	240	I	0.66	III
50	I	12.15	I, III	163	I	4.26	I, III	242	I	1.43	II, III
53	I	1.05	I, III	165	I	0.28	III	245	I & II	1.53	II, III
57	I	1.62	I, III	166	I	1.24	II	251	I	2.55	I, II
62	II	5.97	I	182	I	9.81	I, III	256	I	1.71	I, III
64	I	3.25	I, II	187	I	1.30	I, III				

(B) Conserved sites

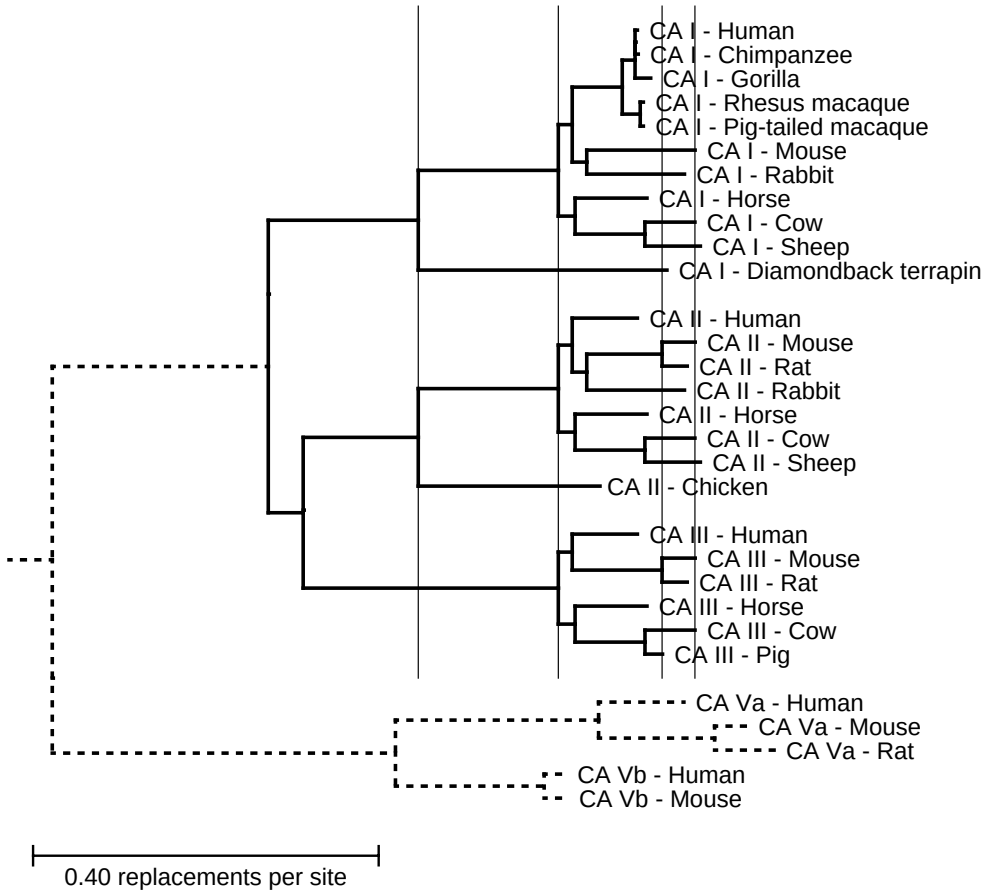
13, 21, 23, 28, 29, 30, 41, 44, 61, 68, 72, 89, 90, 92, 94, 96, 105, 106, 107, 109, 117, 118, 119, 122, 124, 139, 149, 158, 164, 170, 172, 181, 184, 197, 199, 201, 202, 205, 219, 222, 227, 234, 244, 246, 249, 250, 254



Web Figure 5: (First page) Caption, see next page.



Web Figure 5: (Second page) Multiple sequence alignment for all 30 CAs. This alignment follows that of Hewett-Emmett and Tashian (Hewett-Emmett and Tashian, 1996). Sources for these sequences and the scientific names of their species are given in Table 3, which is published as supporting information on the PNAS web site. See Fig. 3 for other details.



Web Figure 6: Accepted phylogeny for the CAs. The distances from the root to identical species and ancestors are fixed across subfamilies, as highlighted by the thin vertical lines. This constraint allows for the more direct interpretation of the site-specific rates among subfamilies in terms of their absolute, rather than relative differences (see text). Sites 1 and 121 are excluded from the branch length estimations, because of their gaps in more than 25% of the sequences. Furthermore, the CA Va and Vb outgroups are not constrained in these estimations, since they are included only to root the phylogeny.

Appendix A

List of Published Articles and Articles in Preparation

Poster

Knudsen, B., Andersen, E. S., Damgaard, C., Kjems, J. and Gorodkin, J. (2002) The impact of evolutionary rate variation on RNA secondary structure prediction. *Bioinformatics 2002*, Bergen.

Articles

Gorodkin, J. and **Knudsen, B.** (2000) RNA informatik. *Naturens Verden,* **11/12**, 2–9. Danish popular science.

Gorodkin, J., **Knudsen, B.**, Zwieb, C. and Samuelsson, T. (2001) SRPDB (signal recognition particle database). *Nucleic Acids Res.* **29** (1), 169–170.

Göttgens, B., Barton, L., Chapman, M., Sinclair, A., **Knudsen, B.**, Grafham, D., Gilbert, J., Rogers, J., Bentley, D. and Green, A. (2002) Transcriptional regulation of the stem cell leukaemia gene - comparative analysis of five vertebrate SCL loci. *Genome Res.* **12** (5), 749–759.

Hein, J., Wiuf, C., **Knudsen, B.**, Møller, M. and Wibling, G. (2000) Statistical alignment: Computational properties, homology testing and goodnesss-of-fit. *J. Mol. Biol.* **302** (1), 265–279.

Knudsen, B. and Gorodkin, J. (2001) Semi-automated update and cleanup of structural RNA alignment databases. *Bioinformatics,* **17** (7), 642–645.

Knudsen, B. and Hein, J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics,* **15** (6), 446–454.

Knudsen, B. and Miyamoto, M. M. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl. Acad. Sci. USA,* **98** (25), 14512–14517.

Knudsen, B., Wower, J., Zwieb, C. and Gorodkin, J. (2001) tmRDB (tmRNA database). *Nucleic Acids Res.* **29** (1), 171–172.

In preparation

Damgaard, C., Andersen, E. S., **Knudsen, B.**, Gorodkin, J. and Kjems, J. (2002) Biochemical and phylogenetic evidence for a higher order structure in the 5'-end of the HIV-1 genome.

Knudsen, B., Andersen, E. S., Damgaard, C., Kjems, J. and Gorodkin, J. (2002) The effect of evolutionary rate variation on the secondary structure prediction of HIV-1 5'-leader RNA.

Knudsen, B. and Hein, J. (2002) Practical RNA secondary structure prediction using stochastic context-free grammars. (Submitted to *Bioinformatics*).

Knudsen, B., Miyamoto, M. M., Laipis, P. J. and Silverman, D. N. (2002) Evolutionary Rate Analyses of Functional Divergence and Conservation among Proteins. (Submitted to *Proc. Natl. Acad. Sci. USA*).

Appendix B

Summary in Danish

Ph.D. afhandlingen 'Molecular Evolution and Biological Sequence Analysis' beskriver en række resultater opnået ved hjælp af matematiske modeller for evolution. Et af hovedresultaterne er en serie af nye metoder til at studere udviklingen af proteiners funktion. Disse metoder har vist sig yderst anvendelige til at forudsige proteiners funktionelt vigtige områder.

Afhandlingen beskriver også en ny metode til at analysere RNA strukturer. Metoden er baseret på en kombination af kendte evolutionære modeller og en model for RNA struktur. Ved at udnytte information fra beslægtede RNA molekyler tillader denne kombination præcise strukturforudsigelser og metoden har blandt andet været anvendt til forudsigelse af strukturelle elementer i HIV virus genomet.