

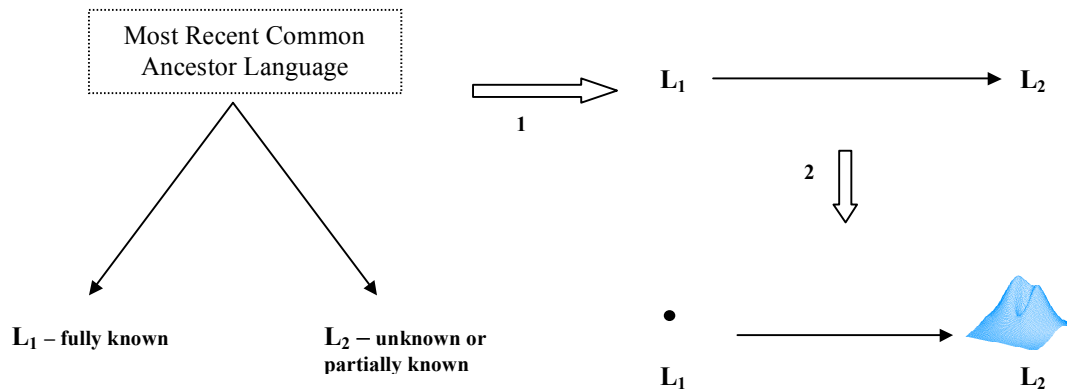
# *Models of Grammar Evolution I*

## *Biosequence Applications*

20.6.07

A language evolves by modification and *cultural* inheritance. This legitimizes the attempt to reconstruct language history and describe the processes governing language evolution. The techniques developed in linguistics are also used in sequence analysis, where inheritance is *genetic* instead of cultural.

These problems are exciting and hard. For sake of realism the project will entail major simplifications to allow focus on basic conceptual issues.



In this simple illustration of the underlying problem, two languages are given. The first language  $L_1$  is fully known, while the second,  $L_2$ , is either totally unknown or partially known. They are both descendants of some unknown language and has evolved according to an evolutionary process on the left and right branches. If the evolutionary process is time reversible the setup can be simplified (arrow 1) to assuming the one language is the ancestor of the other. In this situation (arrow 2)  $L_1$  can be viewed as a point as all parameters and rules are known, while the evolutionary process will create a set of probabilities/densities for  $L_2$ .

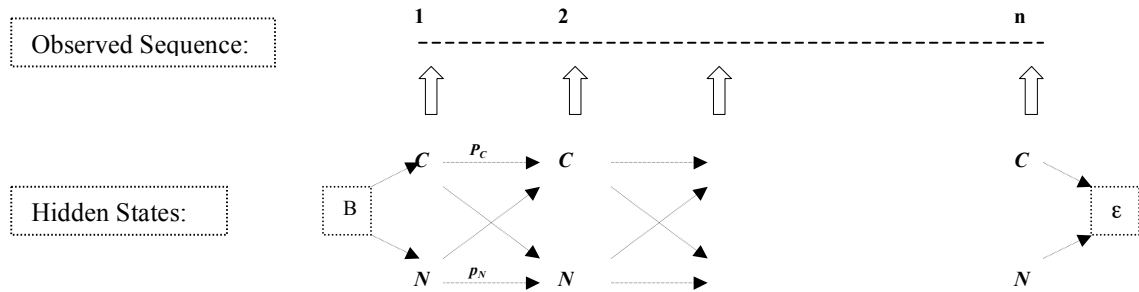
This setup and can be generalized in two directions. Firstly, you might have more languages surrounding the badly characterized language. This is conceptually straightforward and completely analogous the problems studied in phylogenetics (Felsenstein, 2002). From an analysis point of view it will have advantages in more information about the immediate ancestor the language and also that the process of language evolution would be better characterized. However, going from two to many languages could imply a major jump up in computational complexity. Secondly, one could assume symmetry between the two languages. Ie not perfect knowledge of one language, that is then transferred to the second language, but partial knowledge for both languages and information would go both ways.

There are a series of competing ways to represent *grammars*. Transformational grammars initiated by Noam Chomsky ( ) have an intellectual appeal and has further attractiveness in being central in the definition of computer languages and has further applications in the interpretation in the analysis of biomolecules (Durbin et al., 1998).

This description is very general and needs specification to be applicable. Again this can be done in two directions. Firstly, by specification of which language is evolving (regular, CFG,..) and more specific description within a class of the specific rules. Secondly, by application area, which in our case would be within linguistics or biology. We suggest the following application areas and languages:

*i. Stochastic Regular Grammars (SRG) and Protein Gene Finding.* Making Regular Grammars stochastic turns them into Markov Models and if the Stochastic Regular Grammars are used to describe a hidden structure, they become Hidden Markov Models. The first use of Hidden Markov Models in biology is found in Lander and Green (1987), where it was applied to pedigree analysis. In the following years their use became ubiquitous in Bioinformatics. A large class of applications now involves comparative genomics and was initiated independently by a series of groups. Below is illustrated a simplified version of the approach taken by Pedersen and Hein (2003). It is simplified as we ignore that coding regions are elongated by triples and can have introns. Assuming this we only need 2 kinds of hidden states – Coding and Non-coding. Dependent on the hidden state (C,N) an

observable nucleotide is emitted with different probabilities. Additionally, we know the length of the string that is to be generated which is solved by having a counter on the hidden states. Only two probabilities are necessary to describe the probability of the hidden states. The probability to stay coding,  $p_C$ , and the probability of staying non-coding,  $p_N$ . The probability of emitting the nucleotides would need 3 more parameters for each of the two hidden states.

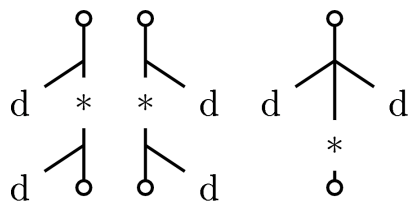


ii. *Stochastic Context Free Grammars (SCFG) and RNA Structure Prediction/Gene Finding.* The main use of SCFG is in the analysis of RNA structure and first publications are from 1993-4 ( Sakakibara et al.; Eddy and Durbin). Here we illustrate by the grammar found in Knudsen and Hein (1999):

$$\begin{aligned}
 S &\rightarrow LS (.87) \quad L (.13) \\
 F &\rightarrow dFd (.79) \quad LS (.21) \\
 L &\rightarrow s (.89) \quad dFd (.11)
 \end{aligned}$$

Where large letters are variables and small letter nucleotides.  $s$  is chosen among the 4 nucleotides according to a probability distribution (3 par) and two  $d$ 's are chosen according to a probability distribution on dinucleotides (15 par). Alternatively to choosing nucleotides and di-nucleotides probabilities, rate parameters on continuous time markov chain can be chosen, which is more suitable when many sequences have been observed on the leaves of a phylogeny.

iii. *Stochastic Tree Adjoining Grammars (STAG).* The class of languages defined by tree adjoining grammars sits in-between two of the original classes defined by Chomsky, the context free languages and the context sensitive languages. Where variables in the current string are expanded according to production rules in context free grammars, in tree adjoining grammars variable branches in the current tree can be expanded according to production rules in tree adjoining grammars. An example is the grammar used by Uemura et al. (1999) to model some RNA pseudoknots from which the following three replacement trees is an excerpt.



The variable branches are indicated by \*, and again the two  $d$ 's need to be chosen according to a probability distribution on dinucleotides (15 par). Where context free grammars can only model nested dependencies, tree adjoining grammars allow modelling of some non-nested dependencies. This is critical for RNA pseudoknots and complex protein structures.

**Evolutionary Model:** Above we have describes some grammars and their biological applications, but not how a grammar could evolve. Concerning evolutionary process there is a natural process to be used. Since we have probabilities, which are real numbers, Brownian Motion (Karlin and Taylor, 1975 chapt. 8) is the obvious choice. We also want an equilibrium distribution on languages, so the use of rules is not allowed to be lost. Ie either reflecting, sticky or absorbing boundaries (some probability is zero) should be used. Sticky boundaries seems the natural choice here. Reflecting boundaries will make all languages have the same rules available, although with different probabilities. Absorbing

boundaries will make the languages more and more trivial as time passes. Sticky boundaries are for our purposes intermediate between absorbing and reflecting. If very sticky, then only few rules will be active at a given time point. If the boundary is not very sticky, then almost all rules will be active at a given time point. If such a model was applied to the Pedersen model there would be 4 probability vectors, that could evolve: 2 describing the HMM and 2 describing the emit probabilities. Assuming that each position in the vector evolves at the same rate and all boundaries are equally sticky such a model would need 2 times 2 parameters (infinitesimal variance and stickiness). If applied to the Knudsen model, there would be 5 probability vectors, that could evolve giving 5 times 2 parameters.

The grammar models above were formulated to be the full description of the phenomena, so applying an evolutionary model that can remove rules, will most likely give deficient models. It will here be natural to define a considerably richer model, where even sub-models are rich enough to describe the phenomena.

### **The key questions**

An immediate application of modelling evolution of grammatical rules will be to transfer information about one sublanguage of a larger language family to another sublanguage. For example, knowledge of one family of proteins can be used to assist in modelling a related family of proteins. An important aspect of this project will thus be to explore and develop methods for applying models of grammar evolution to this end. Assume that an HMM describing the encoding of one protein family and a number of sequences from a related protein family is given. The HMM, coupled with a model of how it evolves, gives us a prior on the HMMs that can be used to model sequences from the related family. Combining this with how well each HMM explains the sequences, this will yield a posterior probability on the HMMs taking the relationship into account.

A further step will be to try to infer trajectories that could lead from the known grammar to a suitable grammar for the related sublanguage. Just using the known grammar to define a prior on suitable grammars ignores the question of whether all intermediate steps also represent reasonable models for the larger language family. A key problem here will be how to define reasonable modelling of the larger language family. One approach could be to use a weighted average of how well the known sublanguage is modelled and how well the target sublanguage is modelled. This weight would then be required to progressively shift from the known sublanguage to the target sublanguage. A further benefit, compared to the increased realism, will be that modelling of trajectories will allow statements about possible ancestors of the two sublanguages.

### **Project**

The reference list is too large to read, so trying to read the following ... with much reading in week 1-2 and less in the remaining time could be advisable.

Week 1-3: Implement the Knudsen Grammar and simulate strings and structures from this.

Week 4: Implement a general grammar evolution.

### **References**

- Brown, D. et al. (2005) "Subfamily HMMS in Functional Genomics" Pacific Symposium on Biocomputing 10:322-333 Eddy, SR and R. Durbin (1994) "Covariance models of RNA" Nuc. Ac. Res.
- Felsenstein (2002) "Inferring Phylogenies" Sinauer
- Chomsky, N ()
- Durbin et al. (1998) "Biological Sequence Analysis" CUP
- Karlin and Taylor (1975) "A First Course in Stochastic Processes" Academic Press
- Knudsen, B. and J.J. Hein (1999) "Using stochastic context free grammars and molecular evolution to predict RNA secondary structure (Bioinformatics vol 15.5 15.6.446-454)
- Lander, E.S. and Green, P. (1987) "Construction of multi-locus genetic linkage maps in humans" PNAS 84: 2363-2367
- Meyer IM, Durbin R. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res* 32(2):776-83. (2004)
- Mitkov, R. (2005) "Oxford Handbook of Computational Linguistics" OUP
- Pedersen, J.S. and J.J. Hein (2003) "Gene finding with a hidden Markov model of genome structure and evolution" Bioinformatics 19.2.219-227.
- Sakakibara, Y. et al. (1993) Stochastic Context-Free Grammars for Modeling RNA" Proceedings of the 27th Hawaii International Conference on System Sciences
- Uemura et al. (1999) "Tree adjoining grammars for RNA structure prediction" Theor. Compu. Sci. 210.2.277 - 303