

UPTEC X 02 004
MAR 2002

ISSN 1401-2138

ANDERS LARSSON

Development of strategy for finding of non-coding RNA pol III genes

Master's degree project



Molecular Biotechnology Programme
Uppsala University School of Engineering

UPTEC X 02 004	Date of issue 2002-03	
Author Anders Larsson		
Title (English) Development of strategy for finding of non-coding RNA pol III genes		
Title (Swedish)		
Abstract <p>A strategy to find novel non-coding pol III gene products was developed. It is based on a hidden semi-Markov model that finds pol III promoter containing sequences given a genomic sequence. These sequences are filtered by searching for secondary structures and conservation between organisms. A project working in parallel has implemented the model and a verification of the program has been done by searching chromosomes 1-10 in <i>Sacharomyces cerevisiae</i> for a certain well-defined class of tRNAs. 24 out of 24 predicted tRNAs were confirmed using a BLAST similarity search. Verification of novel candidates has not been done due to the need for expression analysis <i>in vivo</i>. These experiments will follow this work and will indicate if this is a successful strategy or not.</p>		
Keywords <p>RNA polymerase III, Hidden semi-Markov model, Non-coding RNA</p>		
Supervisors Anders Virtanen ICM, Uppsala universitet		
Examiner Leif Kirsebom ICM, Uppsala universitet		
Project name RNomics	Sponsors	
Language English	Security	
ISSN 1401-2138	Classification	
Supplementary bibliographical information	Pages 26	
Biology Education Centre Box 592 S-75124 Uppsala	Biomedical Center Tel +46 (0)18 4710000	Husargatan 3 Uppsala Fax +46 (0)18 555217

Development of model for finding of non-coding RNA pol III gene products

Anders Larsson

Sammanfattning

För 50 år sedan upptäcktes att all information som krävs för att en cell ska fungera är lagrat i arvsmassan i form av deoxyribonukleinsyra, DNA. Forskning kunde därefter visa hur vissa delar, gener, kopieras till ribonukleinsyra, RNA, som sedan kan översättas till proteiner. Sedan länge visste man att proteiner var viktiga i många cellulära processer medan RNA endast ansågs vara en informationsbärare utan egna funktionella egenskaper. De två senaste decenniernas har dock ändrat på denna syn och allt fler funktioner har kunnat knytas till RNA samtidigt som många organismers arvsmassor blivit avlästa, sekvenserade. Den här sekvenseringen har lett till ett behov av datorprogram som kan leta efter gener i arvsmassan, och tillförlitliga program för prediktion av proteinkodande gener finns redan. Tyvärr hittar de här programmen inte de gener som kodar för funktionella RNA eftersom programmen är baserade på karakteristika hämtade från just proteinkodande gener. Syftet med detta arbete var att ta fram en strategi över hur man kan leta efter gener som kodar för funktionella RNA. Ett program baserat på denna strategi har programmerats i ett angränsande projekt varpå provkörningar på kända och okända delar av arvsmassan har utförts. Testerna har visat att de kända generna plockats upp som önskat medan resultatet av de andra provkörningarna inte har kunnat utvärderas än. Det beror på att en sådan utvärdering kräver biologiska kontrollförsök vilka kommer att utföras i en förlängning av projektet.

Examensarbete 20 p i Molekylär bioteknikprogrammet

Uppsala universitet januari 2002

1. Table of Contents

1. Table of Contents.....	5
2. Background and Motivation.....	6
3. Statement of Problem and Strategy.....	7
4. The Model System.....	8
4.1. History of Gene Modelling	8
4.2. The Markov Model	9
4.2.1 Notation and basic example	9
4.2.2 The forward algorithm and optimal node sequence	11
4.2.3 Hidden semi-Markov models and length distributions	13
5. The Biology.....	16
5.1. Synthesis of Proteins and functional RNA molecules	16
5.2. RNA Polymerases	16
5.3. RNA Polymerase III	17
5.4. Structure and Function	18
6. Modelling and Post Processing of Candidate Genes.....	20
6.1. Mutational Studies and Statistics of box A, box B and box C	20
6.2. HSMM of Pol III Transcription	22
6.3. Post Processing of Gene Candidates	23
7. Discussion.....	26
8. References.....	27

2. Background and Motivation

As a consequence of the rapid sequencing of whole genomes, computer based search methods for gene finding are desirable. Until recently focus has been on genes with open reading frames coding for proteins software for this purpose that is working relatively accurate is now freely available on the Internet. In the past decade evidence has been found that untranslated RNA molecules, called functional RNAs, are essential in several processes such as messenger RNA splicing (mRNA splicing)⁽¹⁾, tRNA processing⁽²⁾, protein secretion⁽³⁾ and gene regulation^(4,5). This is in contrast to the traditional view of RNA as a mere information carrier in the production of proteins. The observation of these important functions has led to a need for new programs that can detect these genes since the existing programs fail to find them. This failure is due to the criteria built into the programs, criteria based on the knowledge gained from the studies of protein production in the cell.

A project aiming at increased understanding of functions and interactions among RNA molecules is at present undertaken at BMC, Uppsala. Gerhart Wagner and Jörg Vogel have as a first step in this project identified *Escherichia coli* RNA gene candidates *in silico*. This was done mainly by searching the genome for prokaryotic promoter and termination motifs. They were then able to verify expression of 17 of these candidates *in vivo*⁽⁶⁾. These results inspired to use a similar approach to search for RNA genes in eukaryotic genomes.

3. Statement of Problem and Strategy

The main objective with this project is to develop a strategy that can be implemented and used as a tool when searching for non-coding genes in eukaryotic genomes. The final result when using this search tool is a list of putative non-coding genes that later can be tested by expression analysis *in vivo*. The implementation will be done in another project working in parallel⁽⁷⁾.

The strategy for reaching this objective is to characterise the transcription machinery with literature studies. This knowledge is then used to build a model of a DNA sequence using a suitable model system. An important part of this model building will be to select the proper criteria and how to define these criteria.

4. The Model System

4.1 History of Gene Modelling

When computer based attempts to find novel genes in genomes started, gene specific features became increasingly important. These features are used to customise gene models to recognise sequences belonging to a certain class. Protein-coding genes, tRNAs, and other groups of genes sharing well defined consensus sequences are suitable for defining a model. These models can then be used for identification of novel genes from uncharacterised sequences. Several different approaches in gene finding have been made^(8,9,10) and like in all pattern recognition problems two different strategies can be utilised⁽¹¹⁾. One way is to make use of requirements known to exist in the modelled system (a priori information). Examples of such requirements when searching for novel genes are conserved promoter sequences, presence of open reading frames, conserved termination sequences *etc.* These criteria can then be exploited and implemented if a proper model structure is chosen^(8,9). The other approach is computer-based analysis of examples known to belong to the modelled system in order to find characteristics that have not been found previously⁽¹¹⁾. By using a training algorithm and training sequences, the model parameters are fine-tuned to capture characteristics from the modelled sequences. One drawback with this approach is the problem of translating the fine-tuned parameters into biologically relevant terms making it difficult to know what the classification of unknown sequences is based on⁽¹¹⁾.

There are three fundamental problems when applying this second approach⁽¹¹⁾. Is the model structure able to represent the modelled class? Are there enough training examples to fine-tune the parameters? Are the training examples representative of the modelled class? An example illustrating the first question is when trying to represent a banana with a circular model. No matter how the value of the radius is adjusted, the model will be bad. Increasing model flexibility, for example by representing the banana with an ellipse instead of a circle, improves the model but also increases the number of model parameters. A circle is described with only the radius whereas an ellipse must be described with two parameters. This increase in model parameters may lead to the problem addressed in question number two which can be easiest understood by considering the effect if the number of parameters is far more than the number of known examples. There will simply not be enough examples to fine-tune the parameters resulting in almost arbitrary parameters. The third problem concerns the quality of the examples. If the model is trained to recognise long and yellow bananas it will not find any brown-spotted or green bananas. The examples must therefore represent all variations within the modelled class.

As will be stated and motivated later, this project will focus on RNA polymerase III transcribed genes and considering that only a few types of pol III transcripts have been identified (*e.g.* tRNA, 5SRNA, U6RNA)⁽¹²⁾ it is unwise to use the machine learning approach and try to fine-tune a great number of parameters. Rather, the model must be based on a priori information that later can be implemented in the model as parameter values⁽¹³⁾. Such implementations can be done in Markov models, a model that is presented below.

4.2 The Markov Model

A Markov model (MM) has a structure designed for modelling of discrete sequence families and can be viewed as a box generating sequences of symbols according to the parameters the model contains⁽¹⁴⁾. It was originally used for speech recognition problems and is still the major model system in such applications, but other pattern recognition problems are also solved using MMs. The model consists of probability parameters that are chosen or trained to capture characteristics of the modelled sequence family. By calculating the probability that the model has generated an unknown sequence it is possible to classify it as a member of the modelled sequence family.

A powerful feature of MMs compared to other models is the ability to model sequence families of varying lengths⁽¹⁴⁾. This is also a necessary feature when gene families are modelled and one of the reasons why geneticists have found MMs useful. Another valuable feature is the straightforward way of implementing a priori information such as conserved sequences.

4.2.1 Notation and a basic example

A Markov Model (MM) consists of a number of nodes (N), including a start and a stop node. Each node has two lists of probabilities consisting of $Tr_{nodeX \rightarrow nodeY}$ and $Em^{nodeX}_{symbolY}$, P(transition from node X to node Y) and P(emission of symbol Y at node X) respectively. A model generates a sequence by making transitions from node to node, emitting a symbol at every node until the stop node is reached. The transitions between nodes and the emissions of symbols are determined by the probabilities corresponding to each node, making each model stochastic. In this paper an arbitrary sequence is denoted O, has a length T and the symbol at position x is denoted O(x).

To fix ideas, consider this simple example. Assume that a particular gene family has an unusual high percentage of purines (A and G), say 80% relative to the expected 50%. Assume further that a model of this family is of interest for classification of unknown gene sequences such as ATGGT (denote this sequence Ω).

A MM with three nodes, including a start and a stop node modelling this gene family can be constructed as in figure 1. To generate a sequence from this model you start at the start node and follow its transition probabilities. Since there is only one transition probability ($Tr_{Start \rightarrow 1}=1.0$), node 1 will be the next node. This transition is followed by an emission of a symbol according to the emission probabilities of the new node ($Em^1_{symbolY}$). A purine or a pyrimidine is emitted with probability 0.8 ($Em^1_{pur}=0.8$) and 0.2 ($Em^1_{pyr}=0.2$) respectively. The next node is given by the transition probabilities and since only transitions to the stop node or back to node 1 is allowed, this model will continue to emit symbols according to the emission probabilities at node 1 until a transition to the stop node occurs. Hence, a sequence from this model will have a purine at any position with probability 0.8 in accordance with the priori information.

MM of hypothetical gene family

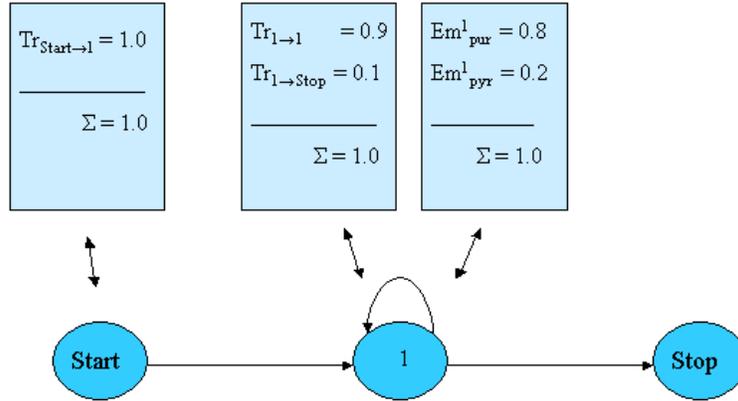


Figure 1. Structure and parameter values of a MM constructed to model a gene family with a purine content of 80 %. Starting at the start node, obeying transition and emission probability lists at the corresponding node, a growing sequence is generated until the stop node is reached.

After a little thought, it can be seen that this model can generate any possible nucleotide sequence. In order to discriminate between model sequences and other sequences, a new definition must be introduced. $P(O \mid \lambda)$ is the probability that a model with given parameter values (λ) generates sequence O and it can then be used in several ways. In the given example, an alternative model can be designed to match general sequences and the analysed sequence is classified as belonging to the MM producing the highest probability. Great care has to be taken when a competing model is made. If this competing model is based on irrelevant assumptions it will end up with extremely low or zero probability and thus falsely indicate a positive result from the first model.

In this particular example the second model is identical with the first model except the emission probabilities at node 1. They are here changed to suit the purine and pyrimidine frequencies associated with a general gene sequence, *i.e.* $Em_{pur}^1=0.5$ and $Em_{pyr}^1=0.5$.

In this example there is only one possible way of generating the sequence Ω in both models making the calculations intuitive.

$$P(\Omega \mid \lambda) = Tr_{START \rightarrow 1} \cdot Em_{pur}^1 \cdot Tr_{1 \rightarrow 1} \cdot Em_{pyr}^1 \cdot Tr_{1 \rightarrow 1} \cdot Em_{pur}^1 \cdot Tr_{1 \rightarrow 1} \cdot Em_{pur}^1 \cdot Tr_{1 \rightarrow 1} \cdot Em_{pyr}^1 \cdot Tr_{1 \rightarrow STOP}$$

This gives the following probabilities for model λ_1 and model λ_2 .

$$P(\Omega \mid \lambda_1) = 2.1 \cdot 10^{-3}$$

$$P(\Omega \mid \lambda_2) = 1.3 \cdot 10^{-3}$$

According to these models, the sequence O is classified as not belonging to the gene family. This seems reasonable since the only consideration in the model is purine frequency and Ω 's frequency (60 %) is closer to a hypothetical general gene than to the gene family.

4.2.2 The forward algorithm and optimal node sequence

When more complicated structures are considered different problems arise. As mentioned above, MMs can be constructed which generates a particular sequence in many ways, *i.e.* you cannot tell in which order the nodes are visited by looking at the sequence. These models are called hidden Markov models, HMMs, due to this characteristic⁽¹⁴⁾. This lack of absolute connection between sequence and path through the model makes the calculation of $P(O \mid \lambda)$ a much more challenging problem. One way to solve this problem is to examine every possible path through the model, calculate the corresponding probability and sum all probabilities corresponding to a sequence matching O . In models allowing transition between any nodes this leads to $2T \cdot N^T$ calculations and will thus be infeasible considering that genomic searches include sequences with hundreds of nucleotides. Fortunately, there is a general algorithm based on induction, the forward algorithm, reducing the number of calculations considerably. The basic idea is to examine the sequence step-by-step and make use of previous calculations, thereby lowering the number of calculations. This is best understood by viewing the algorithm.

1. Initiation:

$$\alpha_1(i) = Tr_{start \rightarrow i} \cdot Em_{O(1)}^i, \quad 1 \leq i \leq N$$

2. Induction:

$$\alpha_{t+1}(j) = \sum_{i=1}^N (\alpha_t(i) \cdot Tr_{i \rightarrow j} \cdot Em_{O(t+1)}^j), \quad 1 \leq t \leq T-1$$

$$1 \leq j \leq N$$

3. Termination:

$$P(O \mid \lambda) = \sum_{i=1}^N (\alpha_T(i) \cdot Tr_{i \rightarrow Stop}), \quad 1 \leq i \leq N$$

As can be seen, the matrix α has dimension $[N, T]$ and $\alpha_t(i)$ contains the probability of being at node i when the model has generated the sequence from $O(1)$ to $O(t)$, *i.e.* $P(O(1) \rightarrow O(t), \text{standing at } i \mid \lambda)$. At initiation, $\alpha_1(i)$ is the probability to go from start node to node i and emitting the first symbol in the sequence (symbol $O(1)$). To calculate $\alpha_{t+1}(j)$ in the induction step, every individual $\alpha_t(i)$ is multiplied with the probabilities of going to node j and emitting the next symbol, *i.e.* $\alpha_t(i) \cdot Tr_{i \rightarrow j} \cdot Em_{O(t+1)}^j$. These probabilities are then summed over all i to give the total probability of $\alpha_{t+1}(j)$. The induction ends by calculating $\alpha_T(i)$ which is the probability of generating sequence O and emitting the last symbol at node i . $P(O \mid \lambda)$ is finally calculated simply by multiplying $\alpha_T(i)$ with the probability of going to the stop node and adding these probabilities over all nodes.

Two interesting questions to answer are: If there are multiple paths through the model, which node sequence contributes the most to $P(O \mid \lambda)$? Or in other words: which is the most probable node sequence? The second question is; what is the probability of this node sequence? These questions can be illustrated by redesigning the HMMs in the previous example. This is done in figure 2, where model λ_1 and model λ_2 are fused into one single HMM.

Alternative solution

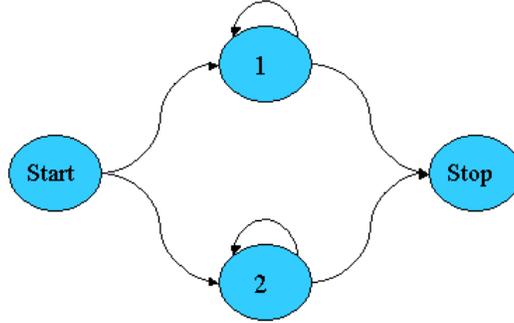


Figure 2. Model λ_1 and λ_2 fused into one single HMM. The path from start to stop via node 1 corresponds to model λ_1 and the path via node 2 corresponds to model λ_2 . Both paths can generate any possible sequence raising the question of the most likely path.

The sequence Ω can be generated by going from start node to the stop node either via node 1 or via node 2. These two paths represent a gene family member and a non gene family member, respectively. If only $P(O | \lambda)$ is computed, there is no way to determine which path is optimal since $P(O | \lambda)$ will be the sum of both alternatives. A general way of solving this kind of problem is to store some of the computations made in the forward algorithm. Two more variables must be introduced to solve this problem, $\beta_t(i)$ and $\psi_t(i)$.

1. Initiation:

$$\beta_1(i) = Tr_{start \rightarrow i} \cdot Em_{O(1)}^i$$

$$\psi_1(i) = 0, \quad 1 \leq i \leq N$$

2. Induction:

$$\beta_{t+1}(j) = \max_{1 \leq i \leq N} (\beta_t(i) \cdot Tr_{i \rightarrow j} \cdot Em_{O(t+1)}^j), \quad 1 \leq t \leq T-1$$

$$1 \leq j \leq N$$

$$\psi_{t+1}(j) = \arg \max_{1 \leq i \leq N} (\beta_t(i) \cdot Tr_{i \rightarrow j} \cdot Em_{O(t+1)}^j), \quad 1 \leq t \leq T-1$$

3. Termination:

$$P(\text{most probable path}) = \max_{1 \leq i \leq N} P(\text{path}_i | O, \lambda) = \max_{1 \leq i \leq N} (\beta_T \cdot Tr_{i \rightarrow \text{Stop}})$$

$$\Psi_{\text{Stop}} = \arg \max_{1 \leq i \leq N} (\psi_T(i) \cdot Tr_{i \rightarrow \text{Stop}})$$

As described above, $\alpha_t(i)$ contain $P(O(1) \rightarrow O(t), \text{ending at } i | \lambda)$, *i.e.* the accumulated probability of all paths generating sequence $O(1) \rightarrow O(t)$ and ending at node i . The only difference between $\beta_t(i)$ and $\alpha_t(i)$ is the use of maximisation instead of summing. As a consequence of this, $\beta_t(i)$ contains the probability of the most probable path generating $O(1) \rightarrow O(t)$ and ending at node i instead of the accumulated probability of all paths generating $O(1) \rightarrow O(t)$ and ending at node i . At each maximisation the index corresponding to the highest term is stored in matrix ψ at element $\psi_t(i)$, making it possible to trace the most probable path backwards. Finally, $P(\text{most probable path})$ and Ψ_{Stop} are calculated. Ψ_{Stop} contains the index of the node last visited by the optimal path and hence serves as a start when the optimal path is back tracked through the model. In the example illustrated in figure 2, Ψ_{Stop} is node 2, showing that the most probable path was via node number

two and thus classifying the sequence Ω as a general gene. This is the same result as when two competing models were used.

4.2.3 Hidden semi-Markov models and length distributions

Sometimes it is known that a modelled sequence family has boundaries for possible lengths and that some lengths are more probable than others. In order to implement such a priori information, the length distribution of sequences from a HMM has to be controlled. In a simple HMM such as the one in figure 1, the distribution will decrease exponentially since there is a fix possibility to end the sequence at every transition from node 1. Generally the length distribution from a HMM will have an exponentially decreasing form. The MM of the hypothetical gene family is used to exemplify this model feature.

$$P(L) = \text{Tr}_{\text{START} \rightarrow 1} \cdot (\text{Tr}_{1 \rightarrow 1})^{L-1} \cdot \text{Tr}_{1 \rightarrow \text{STOP}}$$

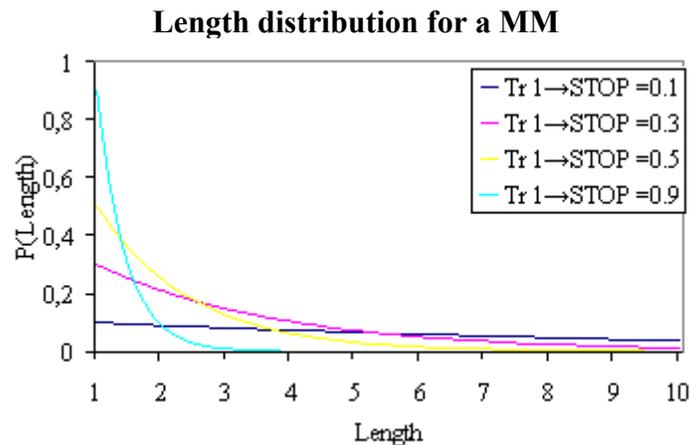


Diagram 1. The length distribution for the MM modelling a hypothetical gene family using four different values on $\text{Tr}_{1 \rightarrow \text{STOP}}$.

This implies that the most probable length always is one, making it impossible to implement information saying that other lengths are more probable. It would also be useful if boundaries for accepted lengths could be implemented. The solution to this problem is to give a length distribution associated to the node determining the number of nucleotides emitted from that particular node. This makes it possible to implement criteria such as lower length limit and upper length limit. A HMM with this feature is called a hidden semi-Markov model (HSMM).

If genes belonging to the hypothetical gene family in figure 1 were known to be around five nucleotides long, a normal distribution with mean five and standard deviation one could be added to node 1. This would yield a model able to consider the length of the sequences.

Markov model with length distribution (SMM)

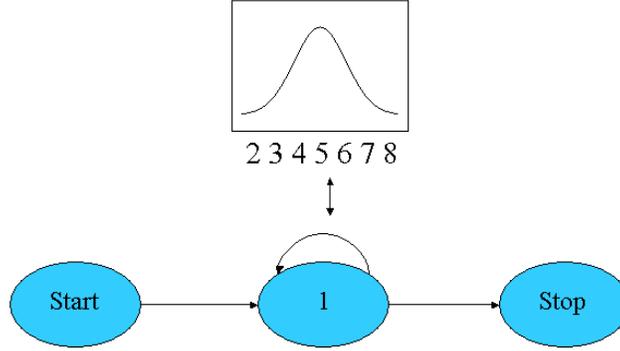


Figure 3. Markov model with a gaussian length distribution, $N(5,1)$, associated with node 1.

Modifications need to be done in the algorithms described above when length distributions are introduced. The major difference is that $\alpha_{t+1}(i)$ and $\beta_{t+1}(i)$ will be determined not only by $\alpha_t(i)$ and $\beta_t(i)$, but also by α_s and β_s with indexes lower than t . This is due to the possibility of multiple emissions at a single node. A third variable, $\kappa_t(i)$, is also introduced for storing of the optimal duration at each step. The final algorithms will be as follows:

1. Initiation:

$$\beta_1(i) = Tr_{start \rightarrow i} \cdot Em_{O(1)}^i$$

$$\psi_1(i) = 0, \quad 1 \leq i \leq N$$

$$\kappa_1(i) = 1$$

2. Induction:

$$\beta_{t+1}(j) = \max_{\substack{1 \leq i \leq N \\ d_{\min} \leq d \leq d_{\max}}} \left(\beta_{t-d}(i) \cdot Tr_{i \rightarrow j} \cdot P(O_{t-d} \rightarrow O_{t+1} | j) \cdot P_j(d) \right),$$

$$1 \leq t \leq T-1 \\ 1 \leq j \leq N$$

$$\psi_{t+1}(j) = \arg \max_{1 \leq i \leq N} \left(\beta_{t-d}(i) \cdot Tr_{i \rightarrow j} \cdot P(O_{t-d} \rightarrow O_{t+1} | j) \cdot P_j(d) \right)$$

$$1 \leq t \leq T-1 \\ 1 \leq j \leq N \\ d_{\min} \leq j \leq d_{\max}$$

$$\kappa_{t+1}(j) = \arg \max_{d_{\min} \leq d \leq d_{\max}} \left(\beta_{t-d}(i) \cdot Tr_{i \rightarrow j} \cdot P(O_{t-d} \rightarrow O_{t+1} | j) \cdot P_j(d) \right)$$

$$1 \leq t \leq T-1 \\ 1 \leq j \leq N \\ 1 \leq i \leq N$$

3. Termination:

$$P(\text{most Probable Path}) = \max_{1 \leq i \leq N} P(\text{path}_i | O, \lambda) = \max_{1 \leq i \leq N} (\beta_T \cdot Tr_{i \rightarrow Stop})$$

$$\Psi_{Stop} = \arg \max_{1 \leq i \leq N} (\psi_T(i) \cdot Tr_{i \rightarrow Stop})$$

With this general description of Markov models in mind it is possible to build a HSMM modelling DNA sequences and identifying possible functional RNA molecules transcribed by polymerase III. The only thing missing is criteria to implement and they are found by studying the biology. For a more thorough review of MMs and the concept of associated length distributions I refer to Rabinier⁽¹⁴⁾.

5. The Biology

5.1 Synthesis of Proteins and Functional RNA Molecules

The synthesis and processing of mRNAs coding for proteins are well characterised and makes a solid foundation from which gene finding motifs can be extracted. These motifs are mostly identification signals for catalysing units such as polymerases, spliceosomes and ribosomes, all participating in the transcription, modification and translation of the gene into a functional protein. Other characteristics and constraints for protein coding sequences are increased GC-content, an open reading frame and species specific codon usage. Unfortunately, most of these motifs are not present in functional non-coding RNAs and can therefore not be used in programs designed to find them⁽¹³⁾. The only motif common for proteins and functional RNAs is the need of a promoter motif to initiate the transcription.

The work previously undertaken to identify functional RNA genes in the *E.coli* bacteria was based on promoter and termination motifs known to exist in bacteria. The reason why this was such a success is the organisations of genes into operons having a strong termination motif. This motif proved to be the single most important criterion for predicting operons (personal communication)⁽¹⁶⁾. Unfortunately, this termination motif is not present in eukaryotes disqualifying it to be used in eukaryotic models. The most useful idea from this approach that can be applied in this work is the characterisation and usage of conserved promoter sequences.

5.2 RNA Polymerases

Eukaryotic cells contain three RNA polymerases (pol I-III) each transcribing a different set of genes. Pol I transcribes large ribosomal RNAs (rRNA) whereas pol III transcribes transfer RNAs (tRNAs), 5SRNAs and a few other small RNAs. These two polymerases generate more than 80% of the total transcription activity in a growing cell. In spite of this most research has been focused on polymerase II since it transcribes mRNAs, has the biggest variety of transcripts and shows a great number of associated transcription factors and are therefore thought to be important in gene regulation^(12,17).

By using deletion experiments *in vitro* and *in vivo*, the core promoter sequences of every class have been revealed and it is defined as the minimal promoter structure necessary for transcription to occur. These conserved sequences have led to the identification of transcription-associated factors but are also useful in gene finding programs as argued previously. It is important to know that the core promoter sequences alone are not sufficient for functionality *in vivo*. Sequences for enhancer elements, regulator binding motifs and proper spacing in-between are all necessary criteria. Unfortunately, these latter criteria are useless when searching for novel genes since they tend to be gene specific⁽¹²⁾.

Contrary to the old opinion that RNA polymerase II is the most important polymerase for transcription of gene regulatory factors, this work will focus on polymerase III transcription and this is based on two arguments. The first is a practical consideration; it has the most conserved

promoter structure among the polymerases making it the easiest to model. The second is the more interesting one; indications show that RNA polymerase III is a major transcriber of regulatory RNAs.

5.3 RNA Polymerase III

RNA polymerase III consists of 18 or possibly 19 subunits encoded by 17 unique genes. Five of these are present as homologues in bacteria and archeobacteria. The remaining eukaryotic subunits have no homologues in bacteria. When comparing to other eukaryotic polymerases five subunits are common to all and eight are pol III specific. Three of these eight are homologous to subunits required for specific transcription. Pol III does not show specific transcription activity in itself but it is gained through the associated transcription initiation factors (TFIIIA-C). Sequence recognition can be achieved in three different ways, all with TFIIIB ending up at the start site, melting the DNA, recruiting pol III and thereby initiating transcription⁽¹²⁾.

The simplest initiation starts with start site recognition directly by TFIIIB without any help from the other two TFs. This gene promoter family has one essential upstream enhancer element (PSE) and one quantitatively dominant upstream enhancer element (DSE) and can be seen in *e.g.* human U6 snRNA genes. This type of promoter structure is seen in both pol III and pol II transcribed genes. How this distinction between polymerases is done is not fully understood but the position of the PSE element and possible presence of a TATA box seems to be important. So far, only short and weak consensus sequences have been reported for the PSE and the DSE making them poor indicators of RNA genes. This is the main reason why this promoter structure is neglected and not modelled in this work. It is also the smallest pol III promoter family identified^(12,17).

The second promoter family has two internal promoter sequences recognized by TFIIIC, box A and box B. TFIIIC is a 520 kDa complex consisting of six subunits organized in two globular domains. These domains contact box A and box B respectively. Sequence recognition by TFIIIC is followed by recruitment of TFIIIB resulting in transcription initiation. The most studied genes belonging to this promoter family are the tRNA genes^(12,17). Homology, deletion and mutational studies have revealed conserved gene sequences and identified essential nucleotides in both box A and box B. These results are presented and compared with statistics below (see *Mutational Studies and Statistics of box A, box B and box C*).

Two large and internal sequences characterise the third promoter family. One of the large sequences is a box A homologous to the one found in the tRNA promoter family. The other is called box C and is recognized by TFIIIA. TFIIIA is later interacting with and recruiting TFIIIC and TFIIIB to form a preinitiation complex. This is a well-studied promoter structure and 5S rRNA genes contain this structure^(9,10). The essential motifs in this promoter family have been characterized with the same approach as the second promoter family and are also presented below (see *Mutational Studies and Statistics of box A, box B and box C*).

Two well defined promoter structures for RNA pol III

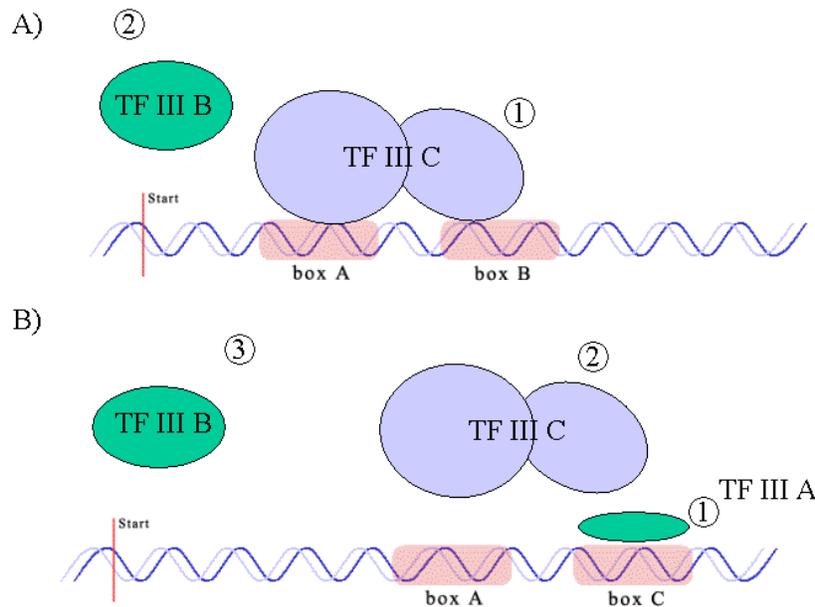


Figure 4. The two most common promoter structures associated with RNA pol III. They are both internal promoters as can be seen by the downstream location relative the gene start. Shaded parts of the DNA helix marks conserved sequences. Pol III is always recruited by TFIIIB making it an essential member of the initiation complex. A) TFIIIC recognises box A and box B and recruits TFIIIB. This promoter structure is found in tRNA genes. B) In this promoter structure, TFIIIC is not capable of promoter recognition but are guided by TFIIIA, which binds to box C. This is a structure found in e.g. 5S rRNA.

5.4 Structure and Function

To be able to make an accurate prediction of the function of an unknown molecule it is useful to know its 3-D structure since that is how they exert their activity. This is a knowledge that was quickly adopted in the study of proteins. Sometimes accurate predictions of structure can be made from amino acid sequence data, but it is computationally demanding and always based on assumptions and simplifications in order to make it possible. The major problem is the enormous number of alternative foldings and internal interactions possible for an amino acid sequence. In spite of fewer variants of building blocks making an RNA molecule, it is still too difficult to make an accurate structure prediction, mainly because of increased difficulties in modelling interactions with the solvent⁽¹⁸⁾. Due to this, attempts of using the stability of a molecule solely, i.e. the Gibbs free energy associated to a conformation, have failed to predict the 3-D structure of RNAs. These are also the reasons why gene finding programs based on such stability calculations are not able to detect novel genes⁽¹⁰⁾. In spite of this setback, the known importance of stability is still a criterion worth to consider in the process to design new programs.

The recent interest in structure and function of RNAs has lead to the identification of many conserved motifs such as alternative base pairing, U-turns, GNRA-loops, UNCG-loops and hydrogen-bonding patterns^(19,20,21). These motifs all have different characteristics and appear to be modules used and conserved in different functional RNAs. As long stretches of single stranded

RNA are unstable, foldings into loops are common while they stabilize RNA sequences, making it the primary criteria in this work when evaluating potential RNA sequences. An interesting feature frequently observed in the stem of these loops is the presence of extra nucleotides in one of the strands resulting in unpaired nucleotides referred to as bulges. These extra nucleotides provide unique recognition sites both directly and by distorting the RNA backbone to increase access to base pairs. These mismatches and kinks in a double stranded backbone are common in functional RNAs, and is used for both structural and functional interactions⁽²²⁾. Another interesting feature seen in both bulges and loops is the importance of adenosine residues in structural framework⁽²³⁾. Based on these findings, bulges containing adenosines will later be allowed in the stem of the loops.

Different loop structures

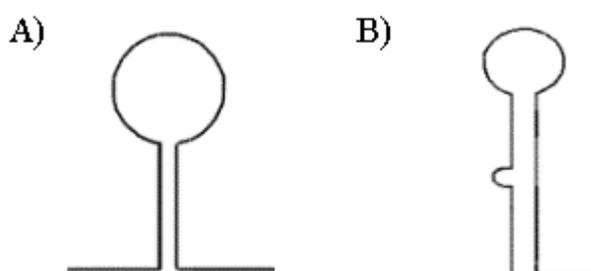


Fig 5. Examples of different loop structures. A) Ordinary loop consisting of stem and loop. B) Loop with extra nucleotides in the left strand of the stem creating a bulge.

6. Modelling and Post Processing of Candidate Genes

6.1 Mutational Studies and Statistics of box A, box B and box C

In the literature studies of this work, papers identifying and characterising the core promoter sequences have been studied^(24,25,26). By stepwise deleting neighbouring nucleotides and observing the effect on transcription the borders of the promoters have been identified. In order to obtain better understanding of the mechanism and essential base pairs in the promoter, the effect on transcription after point mutations have been studied and reported^(25,26). These results have been the starting point for this work when consensus sequences of the promoters have been defined. Statistics of every promoter structure have been generated to complement experimental data and to aid when experiments are missing. This has been done by using ClustalW to align genes known to contain the observed promoter, and making positional statistics for each nucleotide. All sequences used for generation of statistics have been downloaded from IMB-JENA's website.

The promoter structure found in *e.g.* tRNAs including box A and box B has been studied *in vitro* by Brow *et al.*⁽²⁵⁾ using a subcellular minimal transcription model derived from yeast. Point mutations were inserted in box B and the number of transcripts was used as the measure of transcription activity. They also inserted point mutations *in vivo* and measured the steady state level of transcripts. By using their results a box B consensus was defined by allowing any nucleotide at a specific position that does not decrease the transcription *in vivo* by more than 20 % (a cut off value of 0.8). The obtained consensus is, G(A/T)(A/T)C(A/G)ANN(C/G).

Tables of mutational studies of box B

In vivo, box B

	G1	T2	T3	C4	G5	A6	A7	C8	C9
A	0.50±0.04	0.92±0.06	1.00±0.14	0.02±0.01	1.17±0.10			0.86±0.18	0.75±0.16
C	0.53±0.09	0.57±0.16	0.02±0.02		0.10±0.01	0.44±0.06	0.95±0.35		
G		0.69±0.04	0.75±0.08	0.02±0.01		0.77±0.29	1.03±0.51	0.92±0.36	0.92±0.10
T	0.76±0.08			0.06±0.01	0.33±0.06	0.57±0.17	0.73±0.11	0.80±0.27	0.73±0.14

In vitro, box B

	G1	T2	T3	C4	G5	A6	A7	C8	C9
A	0.17±0.02	0.34±0.07	0.64±0.07	<0.01	0.94±0.10			1.11±0.16	0.53±0.03
C	0.13±0.02	0.25±0.03	<0.01		<0.01	0.17±0.01	1.07±0.10		
G		0.21±0.04	0.18±0.05	<0.01		0.28±0.02	0.91±0.09	1.06±0.15	0.85±0.09
T	0.19±0.03			0.04±0.01	0.08±0.01	0.22±0.01	1.07±0.10	1.07±0.13	0.74±0.03

Table 1. Experimental results from Brow *et al.* showing the effects of point mutations in box B. The upper table contains the relative amount of transcripts found when a minimal subcellular transcription system derived from yeast is used. The lower table contains relative steady state values of transcripts *in vivo*. The symbols above each column indicate original nucleotide and position whereas the left column indicates the nucleotide after mutation.

Romaniuk *et al.*⁽²⁶⁾ have performed similar point mutation experiments on box C using the equilibrium constant of TFIIIA's binding to box C as an assay for transcription activity. This choice of assay is due to TFIIIA's promoter recognition property. Since no *in vivo* experiments have been found on box C, the *in vitro* results have been compared with statistics derived from 5SRNA and U6RNA which both contain box C. A trend seen in Brow *et al.*'s experiments is the smaller decrease of transcription for *in vivo* mutations than for *in vitro* mutations. With this in

mind the box C consensus sequence was defined by allowing nucleotides that decrease transcription as much as 60 %. Statistics of U6 RNA and 5S rRNA sequences was compared with this consensus and nucleotides existing in more than 10 % of the sequences were added to obtain the final box C promoter, (A/C/G)(A/G/T)G(A/G)(A/C/G)(A/G/T)G(A/C)(C/T).

Table of mutational study in box C

In vitro, box C

	G1	G2	A3	T4	G5	G6	G7	A8	G9
A	0.30±0.06	0.40±0.01		0.57±0.13	0.23±0.07	0.75±0.18	0.49±0.08		ND
C	0.13±0.03	0.55±0.08	0.70±0.07	0.39±0.10	0.23±0.15	0.35±0.14	0.61±0.19	ND	0.19±0.06
G			0.58±0.04	0.82±0.14				0.70±0.06	
T	0.22±0.05	0.90±0.10	0.20±0.09		0.19±0.07	0.43±0.22	0.56±0.07	0.39±0.07	0.25±0.08

	A10	C11
A		0.48±0.06
C	0.90±0.10	
G	0.40±0.02	0.24±0.05
T	0.50±0.06	0.39±0.13

Table 2. Experimental results from Romaniuk *et al.* showing the effects of point mutations to the equilibrium constant of TFIIB's binding to box C. The symbols above each column indicate original nucleotide and position whereas the left column indicates the nucleotide after mutation.

Statistics of aligned 5S rRNA sequences

Statistics from 299 5S rRNA sequences, box C

	1	2	3	4	5	6	7	8	9	10	11
% A	12.4	7.7	43.0	4.0	1.3	6.7	0.3	18.7	0.01	100	0
% C	20.7	5.7	0.7	0	5.4	1.0	0	1.7	0	0	92.3
% G	50.5	86.3	53.7	0.7	91.6	92.3	99.3	36.8	99.0	0	1.0
% T	16.4	0.3	2.7	95.3	1.7	0	0.3	42.8	0	0	6.7

Statistics from 26 U6 rRNA sequences, box C

	1	2	3	4	5	6	7	8	9	10	11
% A	6.1	3.0	97.0	100	0	0	81.8	0	0	97.0	6.1
% C	0	94.0	0	0	0.3	0	18.2	3.0	3.0	0	78.8
% G	93.9	0	3.0	0	97.0	100	0	0	97.0	3.0	0
% T	16.4	3.0	0	0	0	0	0	97.0	0	0	15.2

Table 3. Tables containing positional statistics from aligned sequences. All sequences downloaded from IMB-JENA's website. The numbers above each column indicate position.

When the box A consensus was defined, no mutation studies was available forcing us to rely on statistics solely. These were gathered from tRNA, 5S rRNA and U6 RNA and ended in this consensus (A/G)G(C/T)(C/G/T)NA(A/G)(C/G/T).

Statistics of aligned sequences containing box A

Statistics from 161 tRNA sequences, box A

	1	2	3	4	5	6	7	8	9
% A	0	84.5	3.8	0	5.0	11.9	98.8	49.1	12.8
% C	1.3	0	1.3	75.2	8.1	31.9	0	2.5	3.0
% G	0	13.0	93.1	1.2	3.1	3.8	1.3	42.2	3.0
% T	98.8	2.5	1.9	23.6	83.9	52.5	0	6.2	81.2

Statistics from 299 5S rRNA sequences, box A

	1	2	3	4	5	6	7	8	9
% A	*	99.7	1.7	5.4	0.3	99.3	99.0	6.4	0.3
% C	*	0	3.7	13.4	21.1	0	0	9.7	94.3
% G	*	0	93.3	0	1.0	0	0.3	82.6	1.7
% T	*	0.3	1.3	81.3	77.6	0.7	0.7	1.3	3.7

Statistics from 26 U6 RNA sequences, box A

	1	2	3	4	5	6	7	8	9
% A	3.9	30.8	0	0	7.7	11.5	96.2	26.9	3.9
% C	3.9	0	3.9	34.6	30.8	34.6	0	0	7.7
% G	0	69.2	92.3	0	26.9	3.9	0	73.1	19.2
% T	92.3	0	3.9	65.4	34.6	50.0	3.9	0	69.2

Tables 4. Tables containing positional statistics of aligned sequences. All sequences downloaded from IMB-JENA's website. The symbols above each column indicate position. The stars in the column of position 1 is due to gaps in the alignment, *i.e.* half of the sequences contained a gap at position 1 whereas the other half did not.

6.2 HSMM of Pol III Transcription

The model used in this work to model pol III transcribed genes interspersed with random DNA sequences is based on the conserved promoter sequences found both in literature studies and in statistics. As can be seen in figure 6, the model consists of six nodes making up two loops. Each node is a MM itself and represents either a conserved sequence or a random DNA sequence with defined length distribution. The loop structure allows the model to deal with sequences of arbitrary length and is a nice way of handling very long sequences such as genomes. The upper loop represents a stretch of DNA containing an A box followed by a spacer sequence and a B box or a C box. The lower loop represents random DNA sequence without any constraints. The general idea is to associate the higher loop, corresponding to interesting sequences, with a much higher probability than the lower loop. The consequence of this is that the most probable node sequence will go through the upper loop whenever possible and thus identifying every possible stretch of sequence containing the conserved sequences and having the proper spacing in-between.

HSMM modelling RNA pol III genes in genomic DNA

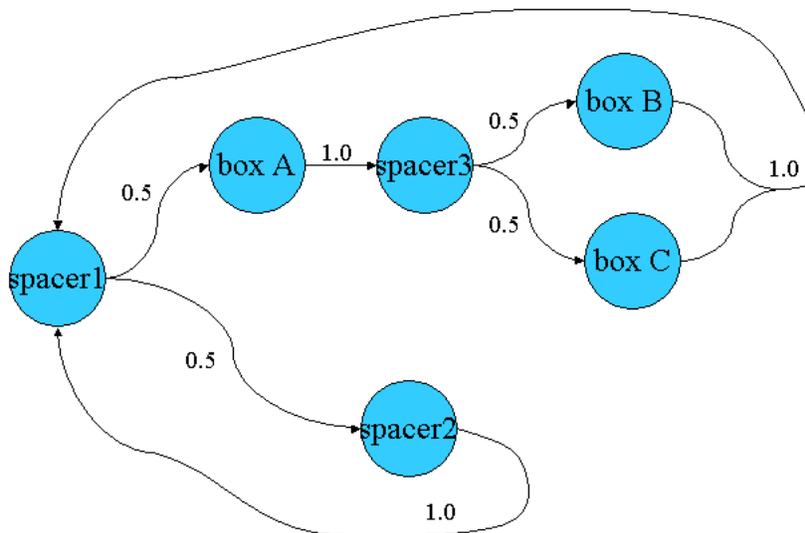


Figure 6. Structure of HSMM designed to model pol III interspersed with random DNA sequences. Box A, box B and box C is linear MM of conserved sequences whereas spacers are SMM of random DNA sequences. The numbers on the arrows indicate the transition probabilities between nodes.

Figure 7 shows that nodes corresponding to conserved sequences are linear in structure while nodes corresponding to spacer sequences consist of one node that allows transitions back to itself. Spacer models also have associated length distributions determining minimum and maximum length. In order to force the most probable path to go through the upper loop when possible, all transition and emission probabilities are put to one for conserved sequences while these probabilities are put to 0.25 for spacer sequences. The length distributions associated with spacer sequences are set to allow lengths of 1-30, 1-200 and 1-150 nucleotides for spacer 1,2 and 3. These parameter values make it possible, although probabilistically expensive to alternate between spacer 1 and spacer 2 making the model behave as desired.

Structures of nodes in pol III HSMM

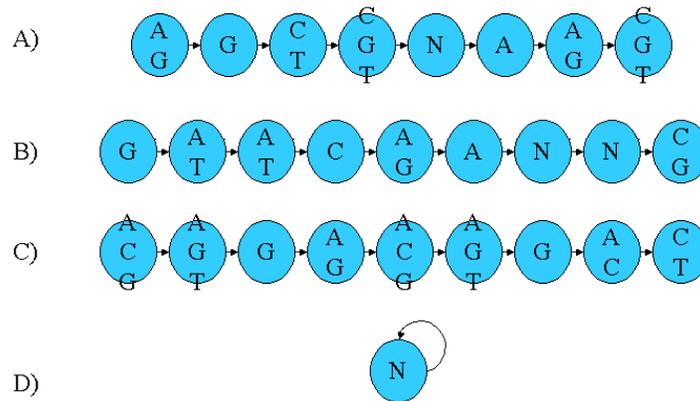


Figure 7. The structures of nodes present in HSMM of pol III genes in genomic DNA. The nucleotides possible to emit are written in each node. N means any nucleotide. In the models corresponding to box A, box B and box C, all transition and emission probabilities are set to one while models corresponding to spacers have transition and emission probabilities set to 0.25. A) A linear MM corresponding to box A consensus. B) A Linear MM corresponding to box B consensus. C) A Linear MM corresponding to box C consensus. D) SMM over spacer sequence. Has an associated length distribution that determines the intervals the length can vary inbetween.

6.3 Post Processing of Gene Candidates

When scanning a genome sequence, the output from the HSMM described above is a list presenting the optimal path through the model. By processing this information it is possible to obtain files where sequences passing through the upper loop in the HSMM is listed, *i.e.* sequences matching the conserved promoter structures. If this is applied to large sequences such as the human X-chromosome the resulting files are huge and impossible to evaluate manually. A phylogenetic analysis is required in order to reduce the number of gene candidates and this is based on loops, especially GNRA and UNCG tetra loops.

The idea is to make local predictions of the secondary structure by searching for possible stem loops in a sequence and listing them together with the analysed sequence. Unfortunately, the number of possible stem loops in a sequence is high even if it is only one hundred nucleotides long, making the existence of possible loops a poor indicator of real functional RNA molecules. The lists are therefore searched for GNRA and UNCG loops, which have been proven to be

present and have functional and structural properties in many functional RNA molecules⁽²¹⁾. This filtering of sequences is not enough making additional criteria necessary. These criteria are made by exploiting the structural homology that exists in homologues across species, a homology that does not have to be based on primary sequence conservation. Since the most important thing for a molecule to survive evolution is to preserve its functional property, conservation of the functional site including 3-D structure and participating residues is most important. This has led to the criterion that if two neighbouring stem loops are found in sequences from two different organisms and the sequences belong to the same pol III promoter family, it is a strong indication of homology. The exact criteria for considering two loops as identical can be altered but it can be based on loop length, loop sequence, stem lengths, bulge length, bulge position relative loop and distance between loops.

Homology criterion

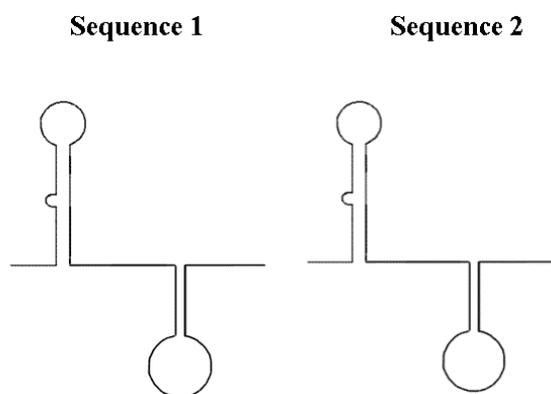


Figure 8. Schematic picture of how the homology criterion is defined. Two sequences from different organisms containing the same promoter structure are analysed according to loops. If both sequences contain two structural similar loops they are considered as homologues. The definition of similar can be changed but it is based on loop length, loop sequence, bulge length, stem lengths and distance between loops.

Steps in extracting gene candidates from genomes

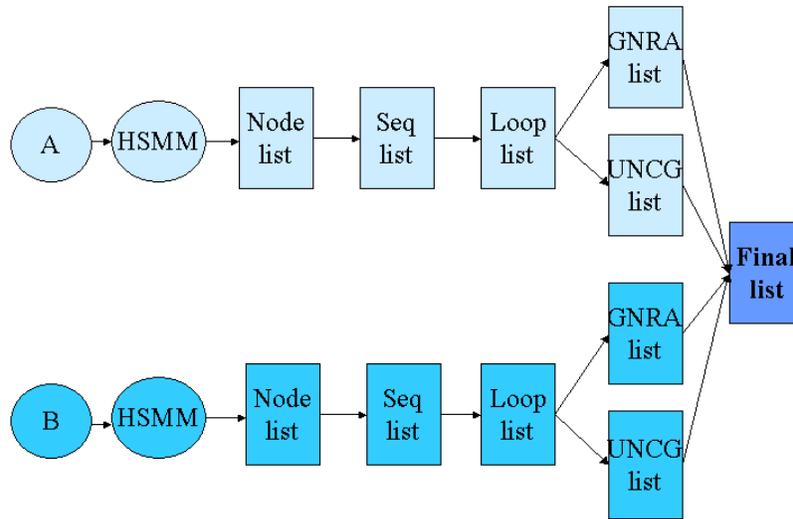


Figure 9. A flow scheme of how a genomic sequence can be processed and how interesting sequences are extracted. The genomic sequence from one organism is run through the HSMM to generate a list containing the node sequence and nucleotides emitted at each node. This list is filtered and reorganised to build a list with sequences matching criteria of conserved promoter sequences, i.e. sequences passing through the upper loop in the HSMM. All sequences in this list are then searched for stretches possible to generate loops. These stretches are stored next to the sequences resulting in a loop list. This list is then searched for specific loops, for example tetra loops such as GNRS and UNCG loops. In order to find homologous sequences, these filtered lists from two different genomes (genomes indicated by A and B) are compared in search of homologues structures.

7. Discussion

As the alert reader has noticed, this is not a work that tries to find *every* functional RNA gene in a genome. Since general criteria allowing for all possible features characterised will yield a vast amount of potential RNA genes, most of them false positives, some sort of specialised search must be done. This specialisation will exclude a lot of the false positives but, hopefully, still include some of the sequences that are interesting. The general idea has been that the HSMM will generate a list with numerous candidates that later can be searched for specific motifs known or suspected to be present in functional RNAs. These additional criteria can be changed and searches can be modified to suit different classes of genes as new information continues to accumulate. So far, the strategy has been implemented in another project⁽⁷⁾ and the resulting program has been verified to behave as desired. This was done by searching *Sacharomyces cerevisiae* chromosome 1-10 for a well-defined subclass of tRNAs. 24 out of 24 predicted tRNAs were confirmed using a BLAST⁽²⁷⁾ similarity search. No evaluation of novel candidates has been possible to do due to the need for expression analysis *in vivo* in order to verify candidates. These experiments will follow this work and will show if this is a successful strategy or not.

8. References

- (1) Madhani, H. D., Guthrie, C. (1994) *Annu. Rev. Genet.* **28**, 1-26
- (2) Kirsebom, L. A. (1995) *Mol. Microbiol.* **17**, 411-20
- (3) Luirink, J., Dobberstein, B. (1994) *Mol. Microbiol.* **11**, 9-13
- (4) Hildebrandt, M., Nellen, W. (1992) *Cell* **69**, 197-204
- (5) Clemens, M. J., Sharp, T. V., Schwemmler, M., Jeffrey, I., Laing, K., Mellor, H. G., Proud, G., Hilse, K. (1993) *Nucl. Acids Res.* **21**, 4483-90
- (6) Wagner, E. G., Margalit, H., Altuvia, S., Argaman, L., Hershberg, R., Vogel, J. (2001) *Curr Biol.* **11**(12) 941-950
- (7) Larsson, P. (2002) *Development of application for predicting non-coding RNA genes*. Master thesis in Molecular Biotechnology, Uppsala university.
- (8) Borodovsky, M., Shmatkov, A. M., Melikyan, A. A., Chernousko, F. L. (1999) *Bioinformatics* **15**, 874-886
- (9) Eddy, S. R., Rivas, E. (1999) *J. Mol. Biol.* **285**, 2053-2068
- (10) Eddy, S. R., Rivas, E. *BMC Bioinformatics* **2**(1):8
- (11) Bishop, C. M. (2000) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.
- (12) Geiduscheck, E. P., Kassavetis, G. A. (2001) *J. Mol. Biol.* **310**, 1-26
- (13) Burge, C. B., Karlin, S. (1998) *Curr. Opin. In Struc. Biol.* **8** 346-354
- (14) Rabinier, L. R. *Proc. of the IEEE* **77**, 257-286
- (15) Howard, R. A. (1971) *Dynamic Probabilistic Systems Vol. II: Semi-Markov and Decision Processes*. John Wiley & Sons, New York.
- (16) Wagner, Gerhart ICM, Uppsala University.
- (17) White, R. J., Paule, M. R. (2000) *Nuc. Acids Res.* **28**, 1283-1298
- (18) Moore, P. B. (1998) The RNA folding problem. In *The RNA World 2* ed., pp. 381-401. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, New York.
- (19) Hermann, T., Patel, D. J. (1999) *J. Mol. Biol.* **294**, 829-849
- (20) Gutell, R. R., Cannone, J. J., Konings, D., Gautheret, D. (2000) *J. Mol. Biol.* **300**, 791-803
- (21) Pardi, A., Heus, H. A. (1991) *Science* **253**(5016), 191-19
- (22) Hermann, T., Patel, D. J. (2000) *Structure* **8**, R47-R54
- (23) Steitz, T. A., Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B. (2001) *Proc. Natl Acad. Sci.* **98**, 4899-4903
- (24) Erdmann, V. A., Pieler, T., Appel, B., Li Oei, s., Mentzel, H. (1985) *EMBO J.* **4**(7) 1847-1853
- (25) Brow, D. A., Kaiser, M. W. (1995) *J. Biol. Chem.* **270**, 11398-11405
- (26) Romaniuk, P. J., Veldhoen, N., You, Q., Setzer, D. R. (1994) *Biochemistry* **33**, 7568-7575
- (27) National centre for biotechnology information (NCBI) <http://www.ncbi.nlm.nih.gov/blast/> 2002-01-20