A Computational Framework for Nucleic Acid Sub-Sequence Identification

Scott Mann, Yi-Ping Phoebe Chen and Luke Eaton
School of Information Technology, Deakin University, Melbourne, Australia
Email: phoebe@deakin.edu.au

Abstract

Identification of nucleic acid sub-sequences within larger background sequences is a fundamental need of the biology community. The applicability correlates to research studies looking for homologous regions, diagnostic purposes and many other related activities. This paper serves to detail the approaches taken leading to sub-sequence identification through the use of Hidden Markov Models and associated scoring optimisations. The investigation of techniques for locating conserved basal promoter elements correlates to promoter thus gene identification techniques. The case study centred on the TATA Box basal promoter element, as such the background is a gene sequence with the TATA Box the target.

Outcomes from the research conducted, highlights generic algorithms for sub-sequence identification, as such these generic processes can be transposed to any case study where identification of a target sequence is required. Paths extending from the work conducted in this investigation have led to the development of a generic framework for the future applicability of Hidden Markov Models to biological sequence analysis in a computational context.

1. Introduction

The identification of smaller sequences within the context of a larger sequence is of significant benefit. The goal is to allow the end user to locate the target and infer properties of the host sequence. Implemented in a computational sense, the automation of this process is highly attractive. The outcome serves to accelerate time consuming 'traditional' biological practises, which now can be replaced by a new class of smart computational techniques. Serving as a platform for demonstration purposes toward the goal of sub-sequence identification, initial attention was devoted to the identification of the gene promoter region. This case study forms the central focus of this paper with correlations to the generic applicability of the approaches being discussed frequently. The algorithms and models will be discussed in a generic context allowing for the knowledge imparted to be used in specific contexts.

This paper will demonstrate the incremental progression from an initial model through to a conclusive model accounting for the intricacies of gene sub-sequence identification using Hidden Markov Models. The remainder of the paper introduces related works, the computational approach and includes discussion regarding the scope and scalability of the approaches taken in view of future research paths.

2. Background

Before commencement of an investigation, two principles have to be considered.

2.1 Hidden Markov Models

A dynamic statistical profile built from the analysis of a 'training' set of data. Its major focus centres on states and their transition and can be visualised as a finite state machine. Probabilities are then assigned for each state (emission) and between states (transition). The term 'hidden' arises from the fact that the state of the model at any time is a function of the input string.

The goal of the Hidden Markov Model is to differentiate sequences matching the consensus vs. those which don't match the consensus. The relative difference in candidate sequence scores highlights the applicability of Hidden Markov Models for identification of consensus sequences.

2.2 Suitable Target Sequence

Candidate sequences for identification must be conserved. The level of conservation dictates the success of the procedure. Basal promoter elements were chosen in the case study due to their **conserved**

0-7695-2173-8/04 $ 20.00 © 2004 IEEE

sequence composition and location. The TATA Box fitting these criteria served as the basis for model development. The TATA Box can be seen in relation to other promoter elements and the transcription start site (+1) in Figure 1.

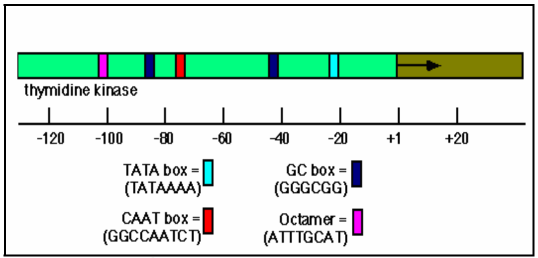

Figure 1. Eukaryote promoter region

With the knowledge that the TATA Box resided approximately 25 nucleotides upstream of the transcription start site and has the consensus sequence of TATAAAA [1], an initial Hidden Markov Model could be developed.

3. Progressive Model Development

On the assumption that TATA Box sequences are highly conserved, the initial model Figure 2 was constructed.

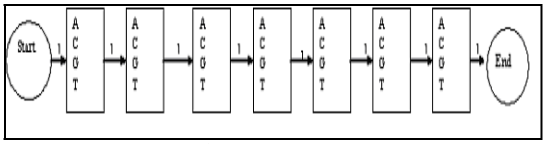

Figure 2. Initial linear HMM

A 7 state HMM was developed; the length of this model was determined by the TATA Box consensus length (7 bases). The emission probabilities ({A,C,T,G} 4 per state) were obtained via a 2 dimensional eukaryote TATA Box probability matrix [2] with the transition probabilities (arrows) set to 1 indicating a linear complete traversal. The assumption made using this model was based on the TATA sequence being fixed at 7 nucleotides.

The ability to quantitatively rank pattern matches based on a trained consensus sequence model, is the key property behind the use of the Hidden Markov Model. The model in Figure 2 was enacted upon TATA Box annotated human gene sequences obtained via the NCBI 'Nucleotide' database located at *http://www.ncbi.nlm.nih.gov*. Initially, a sequence (U07807) was chosen at random to serve as a basis for analysis. Implemented via the JAVA programming language, the core algorithm is shown.

```
START:
        1. Load entire sequence into a string
           variable
        2. For each base extract the next 7 bases
           inclusive
        3. For each of the extracted bases multiply
           their weighted value
                3.1 Retain the highest sequence
                    probability
        4. Go to step 2 until at end of input string
        5. Output the highest scoring sequence
END
```

Figure 3 highlights two primary trends, namely very few sequences scoring above the background, and the incorrect TATA Box element start location being rated the highest.

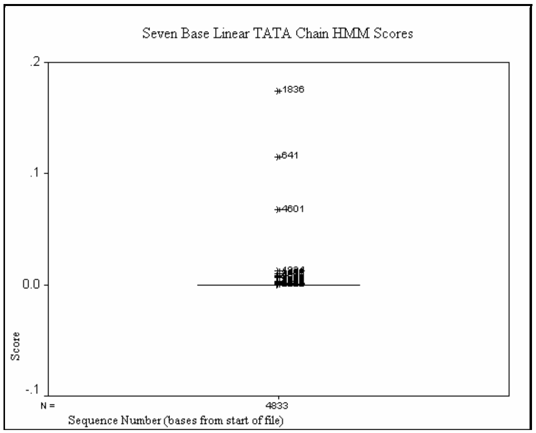

Figure 3. Box plot initial analysis of the U07807 gene sequence

The actual TATA Box was of length 6 nucleotides, as determined from the NCBI 'GenBank' data entry (U07807). In addition, the 7th nucleotide had a sufficiently low score as to mask the correct location of the TATA Box hence allow the incorrect TATA Box sequence at location 1836 to be rated highest.

Six base sequence analysis as per the previous algorithm with appropriate adjustment was considered the next logical step.

*2. For each base extract the next **6** bases inclusive*

The previous trend regarding distribution was preserved, however the **correct** TATA Box promoter element start location (641) has been **ranked the highest**. The modification of the algorithm and repeated analysis on other gene sequences, led to the conclusion that **length dependency dictates correct identification.** Therefore, a static model would not be an effective solution for identifying a dynamic sequence. This issue is addressed later in the paper.

3.1 Scoring Scheme

Implementing a model that scored promoter regions against random nucleotide sequences would lead to more accurate discrimination. Generally, a consensus vs. random nucleotide scheme will provide adequate discrimination. By scoring against a random sequence, the scores would be given meaning relative to a random sequence of nucleotides. For each nucleotide (A, C, G, T) there is a ¼ likelihood of the particular base being present at any given location. Thus a promoter sequence of length n would have a probability of 0.25^{n} random nucleotides.

A generic logarithmic function [3] was applied to the weights obtained in the frequency matrix [2] as per the below equation to implement scoring against a random model.

$$\text{Positional Score} = \ln \frac{weighting}{0.25} \qquad \text{(i)}$$

This function 'log-odds' essentially defines a ratio where the log function is applied for computational efficiency and to achieve an additive scoring scheme. Another benefit of using the logarithm function helps avoid underflow issues arising from the comp utation of very small probabilities.

Preserving the previous trends, resultant plots display the highest scoring outlier as the **correct** TATA Box start location. The logarithm function in combination with box plots allowed us to conclude that the majority of scores fit the null model considerably better than the TATA box consensus. An equivalent statement suggests the majority of sequences behave like a string of random bases as opposed to matching the TATA Box consensus.

The benefits of using the null model logarithm function verse the serial multiplication of probabilities is that it adds biological significance, solves underflow issues and is computationally more efficient especially for larger sequences where an additive model is more beneficial than serial multiplication. These advantages apply generally to HMM analysis.

3.2 Modelling Length Variability

Length dependence is an undesirable property detracting from the identification of the correct TATA Box start location. Observation of 'GenBank' data entries for human TATA Box annotated gene sequences has led to the realisation that the TATA Box sequence has forms 5, 6, 7, 8 nucleotides in length. This level of variance from the consensus of 7 nucleotides would render the HMM model Figure 2 ineffective. To allow for the varying forms of the TATA Box sequence an adaptive HMM was developed Figure 4.

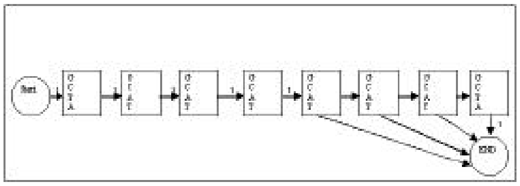

Figure 4. Length independent hidden Markov model

Figure 4 above, is similar to that shown in Figure 2, the commonality exists as the first 5 nucleotides have a '1' transition probability. From the fifth base onwards, the probability of progressing to the next nucleotide is dependent on the frequency obtained from a set of training data. The decision to allow alternative paths through the model from the 5th base onwards was a direct result from viewing human gene DNA sequences and noting TATA Box sequence lengths. The sequences observed were obtained from the NCBI website searching the 'Nucleotide' database with the search string 'TATA_signal "Homo Sapiens"'. Results returned from the NCBI database comprised our training set of 116 sequences. By allowing premature termination paths, length variability can be effectively modelled. With length independency was introduced, candidate sequences are scored according to their sequence composition, independent of length. The presence of 6, 7, 8 base sequences verifies the effectiveness of the length independency optimisation, resultant trends are displayed in Figure 5.

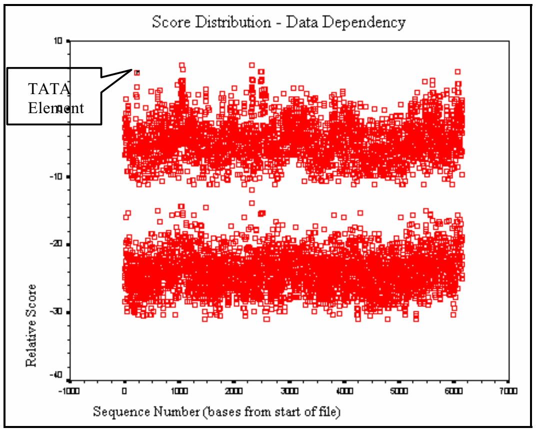

Figure 5. Candidate score distribution gene sequence (M24543)

Figure 5 shows three populations of scores. Beginning at the top, there are very few scores which represent the most likely TATA Box sequences. Next follows a large population of scores that match a random string of nucleotides, lastly a population that deviates significantly from the consensus sequence. The key to greater discrimination of the correct sequence against the high background scores represents the next stage of optimisation.

3.3 Background Reduction

To reduce the number of false positives (i.e. other higher than background scores) it can be hypothesised that placing the scores into a biological context would be more appropriate.

By combining the score for the TATA box sequence with the associated INR scores (~25 bases downstream) it was theorised to help reduce the number of false positives in areas that have no biological relevance to the TATA Box sequence. In order to implement this approach, 'INR' sequences had to be identified given their consensus (Py Py A N T/A Py Py).

3.4 Sequence Scoring Optimisation

The training set used to obtain parameters for the HMM showed a distinct trend. The first 5 bases show a close sequence match to the consensus as does the 6th base, however variation from consensus was clearly apparent beginning at the 7th base. Given the length variability beginning at the 6th base and sequence composition variability beginning at the 7th base, it was theorised that these regions would score poorly, perhaps masking the correct TATA Box sequence.

To negate the effects of scoring the latter portion of the TATA sequence poorly, thus overall reducing the sequences' score, the 'Cumulative/Single Calculation' algorithm was applied.

Giving both regions equal scoring significance would see the masking of correct promoter element sites, the solution was to treat the first 5 bases differently to the last 3 bases (Figure 6).

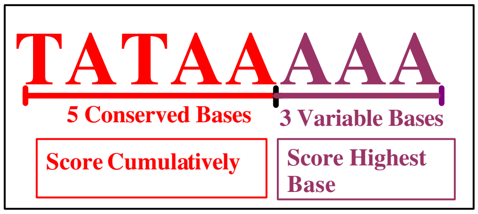

Figure 6. Graphical representation of cumulative-single calculation algorithm

Greater significance was given to the first 5 bases which principally when correct score highly, lesser significance was attributed to the last 3 bases as the variability incurred in scoring these regions may mask the potentially correct score from the first five bases. To score these regions appropriately, **cumulative** scoring was applied to the first 5 bases. Variability within this region will impact greatly on the cumulative score, which is a desirable attribute given that this region matches the consensus closely.

For the remaining 3 bases which represent considerable variability, only the **single** highest score was used to represent the score for this region. This allowed for the observed variation but without penalising the prior 5 base score.

The cumulative score for the first 5 bases are added to the highest transition score from the remaining 3 bases to equal the final score for the potential TATA sequence.

3.5 Dataset Representativeness

Using a broader dataset was predicted to give more representative results. When training a HMM, performance is based largely on the quality of the training data. This was partly accomplished when our trained emission probability matrix was substituted with the generalised eukaryote matrix [2] obtained via 60 different vertebrate protein encoding genes.

3.6 Consensus Blurring

The majority (~91%) of the correctly identified TATA Box sequences matched the TATA Box consensus. Therefore sequences not strictly conforming to the consensus were bypassed. This is a property of the parameters to the Hidden Markov Model. If a consensus base had a very high frequency the model became very intolerant to variation at that state.

The frequency data from [2] indicated that the TATA Box column 2 (nucleotide 'A') had a ~97% match to the consensus at the second base position. This constituted the largest match to the consensus over the TATA Box sequence, therefore any variation from this base would severely effect the overall score for the sequence.

Of particular note, this observation in the frequency matrix concurred with experimental outcomes, as none of the correctly identified sequences contained variation of the 2nd base. To rectify this issue, the log-odds score for the 2nd consensus 'A' base was inverted (1.35 to –1.35) effectively allowing TATA Box sequences varying in the 2nd base to have a greater likelihood of scoring higher.

4. Experimental Design

The acquisition of results was derived from processing involving two datasets. The training set was composed of 116 human TATA Box annotated gene sequences obtained from the NCBI 'Nucleotide' database. This dataset served as the basis for training the Hidden Markov Models discussed previously. Result generation required obtaining 100 additional sequences which formed the testing set.

Repeated execution of our JAVA application entitled 'GeneProFinder' over the testing dataset comprised the data collection procedure. Output was collected via text file redirection. The data collected from incremental optimisation stages was stored in spreadsheet form for comparative purposes.

5. Results and Discussion

A series of optimisation stages has produced the following trend in correct recognition, Figure 7.

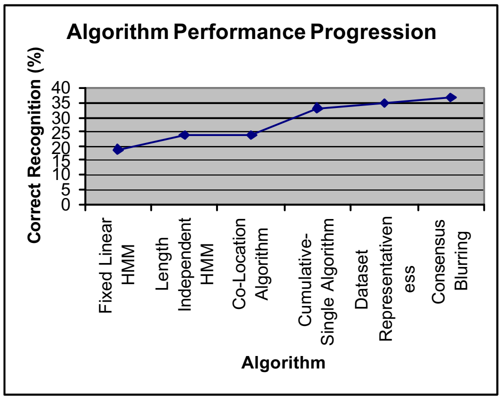

Figure 7. Algorithm performance progression over optimisation cycles

From the initial non-optimised model, the results returned were non representative of actual TATA Box sequence distributions. The static 7 base model would score 7 bases regardless, hence if encountering a 5 or 6 base TATA Box sequence the 7th base would always be scored which had the potential to drop the cumulative score for the candidate sequence. Results obtained from the execution of the fixed length (7 nucleotide) linear HMM over the testing set resulted in a 19% correct TATA Box recognition value over the testing set.

At the second optimisation stage, the length independent model became very sensitive to the values used to train the HMM. Frequency denominator adjustments for the 6th, 7th, 8th bases were required inline with proportions from the training set. This model represented a 5% gain over the static model and served as the basis for further optimisation. Results obtained from the execution of the variable length HMM over the testing set resulted in a 24% correct TATA Box recognition value. The flexibility introduced by the modified HMM affords the ability of candidate sequences to be scored dependent on sequence composition, allowing for length independence.

Building upon the length independent model, the Co-Location algorithm was implemented with no apparent increase in correct recognition. Our search for consensus matching 'INR' sequences suffered from the degenerate consensus sequence of the INR. As a consequence a high degree of variability was encountered. The assertion that in some genes the INR is not present in the consensus form [4], added to the difficulty of this approach. Our findings supported this observation. Readings from [5] indicate that the presence of a TATA Box may diminish the strength (consensus matching) of the INR sequence, this would account for our findings. Lodish [2] suggests the INR sequence replaces the TATA Box sequence, which experimentally conforms to our findings. Serving as an internal control, our model behaves inline with current biological knowledge. Results obtained from the execution of the variable length Co-Location HMM over the testing set resulted in a 24% correct TATA Box recognition value. The implementation of this algorithm has produced no improvement in correct recognition, due to the biologically defined assertions. Given the opportunity to model a Co-Location on two different conserved sequences, background levels are predicted to fall dramatically.

Affording candidate sequences greater flexibility was the goal of the 'Cumulative-Single Scoring' algorithm The outcome led to the capturing of strong TATA Box sequences with the ability to compensate for a degree of variability. This algorithm represented a **9%** increase, the **largest single increment in correct recognition** to a value of 33%, Figure 7. Results obtained from the execution of the variable length HMM using the Cumulative-Single Scoring Algorithm over the testing set resulted in a 33% correct TATA Box recognition value. **The introduction of this algorithm produced the largest increase in correct TATA Box recognition.**

Additional optimisations included expanding the dataset, here a generalised eukaryote TATA Box Frequency Matrix [2] was employed on the assumption that its basis was derived from a more representative dataset, than our training set. By training the HMM model with the most representative data, results would be more conclusive, in our experiments a 2% improvement was achieved. Results obtained from the execution of the variable length HMM using the Cumulative-Single Scoring Algorithm (representative dataset) over the testing set resulted in a 35% correct TATA Box recognition value. A 2% increase was attributed to the more representative dataset.

At the conclusion of the development-optimisation process a 37% correct recognition value was achieved. Consensus blurring was employed to reach this value by capturing sequences not strictly conforming to the consensus. This strategy is a tightly controlled balance between flexibility and consensus matching. The decision to invert **only** the most highly conserved base score was to limit the allowance for wide variability off the consensus, while allowing for limited flexibility in the most conserved region. Inverting or lowering the score of other bases would not necessarily lead to an improvement in correct identification. Other than the highest consensus matching base, others should not have their scores modified as the allowance for variability off consensus would be too wide resulting in a high degree of false positive identification.

A 2% increase comprising the final 37% score was attributed to the blurring of the second consensus location in the TATA Box sequence.

6. Related Work

Related work correlates distantly to the application and processes undertaken in this study, therefore representing significant originality.

The simplified static models proposed by Krogh [3], were rejected early in development. The lack of work specifically relating to our investigation made fair comparisons unreasonable. The closest available research in this area was the work done by Ohler [6] with 'MCPromoter' Version 1.1 using a 5th order interpolated Markov Chain resulting in **28.2%** target promoter sequence recognition with 1/2633 base false positive rate. This work was conducted on the **genomic** scale using *D. melanogaster* as the target organism, as such direct comparison with our **gene** scale investigations are not accurate, however the recognition percentages serve as a basis for this type of analysis. As such our research and optimisations could not be directly compared with work conducted by others in this field of study.

Two properties of the HMM made our work unique and more challenging. These two limitations of the Hidden Markov Model include the 'Limited Alphabet' and 'Short Sequence Length'.

7. Limitations of HMMs for Subsequence Identification

DNA sequence analysis of the type conducted in this investigation is restricted to a very small alphabet {A, C, T, G}. A property of the Hidden Markov Model is, the greater number of possible paths, the more accurate recognition of target matching sequences will become. Given only four bases there is a ¼ likelihood of a random occurrence per state in the models used throughout this paper. As a consequence, a consensus base per state has a comparable background likelihood of 25% at each state. This value represents a very high background to distinguish a consensus score against.

When applied to protein analysis the alphabet is expanded by default to 20 accounting for each amino acid. This increase in the size of the alphabet significantly lowers the random likelihood of a unit matching the target at each state of the HMM.

To summarise, the larger the alphabet, the lower the likelihood random occurrences will match the consensus sequence under analysis. This is especially pertinent for sequences of considerable length. Protein analysis having a 5 fold larger alphabet (20 vs. 4) is better suited to HMM processing in this form.

Our analysis was concerned with the basal promoter sequence 'TATA Box' which has a consensus sequence length of 7 bases. When compared to the number of bases within the genes under analysis, the 7 base TATA sequence had a $\frac{1}{4^7} = \frac{1}{16384}$ nucleotide probability of occurring, considering the typical length of human genes the likelihood of encountering more than one matching sequence is quite high. This assumption is based on a GC content of 50%. Previous work using HMMs has centred around modelling large sequences of proteins. As an analogous example, a small polypeptide chain of 7 amino acids has an $\frac{1}{20^7} = \frac{1}{1.28 * 10^9}$ amino acid chance of occurring in a given sequence of amino acids, significantly smaller than 7 base DNA sequence analysis.

In summary the greater the targets' length, the less likely the target sequence will occur in the data under analysis. When combined with a larger alphabet, the relative occurrence values for our TATA Box sequence vs. the hypothetical protein scenario is staggering. These two limitations restricted our investigations to the gene level, genomic scale study however has been discussed later in this paper.

8. Software Engineering Integration

Implementations of the models presented in this paper were conducted through the use of the JAVA programming language. This language is well suited to biological sequence analysis. The learning curve associated with this programming language is less steep than with others. The ability to rapidly uptake this now mature language will lessen the learning period for the researcher and allow them to concentration on the key tasks. The benefits of the JAVA language include cross-platform interoperability and the wealth of pre-compiled libraries. Both these attributes afford the user the ability to transport their work and spend less time coding.

Applying object oriented techniques to the problem of sub-sequence identification, the requirements phase is critical to the projects success. Step one in this phase required the simplistic use of Unified Modelling Language (UML) use-case diagrams, [7].

With a general plan for the functionality of the application, the architectural structure for the application was developed. The goal was to capture the simplest features and then add the details. Firstly our application 'GeneProFinder' required a user interface, and an analysis capability. This formed the 2 class basis of the application. Next, details were added, namely what are we trying to analyse? To build a biological framework into the application the modelling of the promoter region was implemented via a JAVA class entitled 'Promoter'.

It was essential to retain a JAVA class modelling the biology of the target as detailed as possible. The 'Promoter' class contained all the attributes of a biological promoter including sequence composition, length, and problem domain specific fields including HMM scores. Lastly, to add the functionally of multiple sequence analysis a further class was added to mediate this functionality.

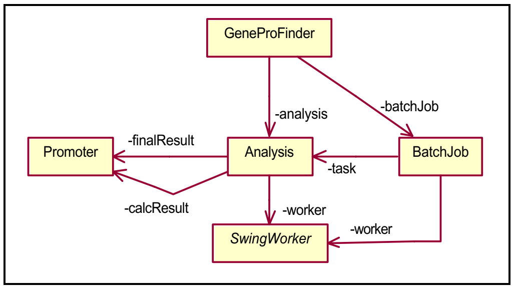

Figure 8. Class Diagram

The class diagram Figure 8, describes the general structure. 'GeneProFinder' displays the GUI and handles user interaction, 'Analysis' mediates the analysis tasks, 'BatchJob' handles via the 'SwingWorker' class multiple sequence analysis. Finally 'Promoter' is regraded as a data store for processed candidate sequences and to model the domain biology.

Commonality between tasks in sequence identification allows for the development of generic components. The consensus format of eukaryote and prokaryote genes represents entities for modelling. In this paper the promoter region has been modelled as a re-useable JAVA class. The bounds for reusable components are limited only by biological constraints.

9. Generic Outcomes

Presented are the series of considerations that make sub-sequence identification possible via HMM modelling. The steps shown in Figure 9 represent the principles developed over the model development process. These principles are generic and can be applied to any conserved sequence amongst a background of biological data. These series of steps with respect to the limitations of the HMM previously discussed offer a concrete approach to model development.

The boxed regions represent decision points with the branches and relevant text describing the appropriate action to take. The decision points and actions are a direct result of the model development process described previously.

10. Conclusions and Future Work

The development process has concluded with a preliminary correct human gene TATA Box recognition of **37%** arising from a series of optimisations shown in Figure 7. The problem of sub-sequence identification has been addressed using Hidden Markov Models in conjunction with **novel** optimisation schemes.

The originality of the optimisation approaches taken, stem from firstly analysing human genomic DNA sequences and secondly matching for very short target sequences. The development process conducted as serial optimisations has led to the development of a generic framework for the computational implementation of sub-sequence identification. The development of generic biological components would serve as re-usable entities for the biological community (perhaps in the JAVA programming language). The promoter region modelled in this paper can be decomposed down into its elemental composition. Similarly it can be scaled up for integration into a re-usable gene entity mapping all the way to the genomic level.

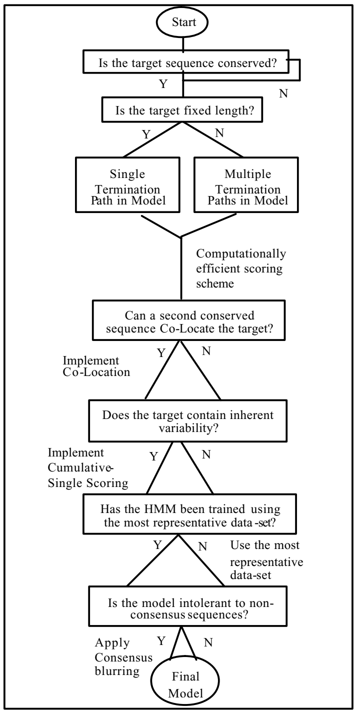
Start
Is the target sequence conserved?
Y
N
Is the target fixed length?
Y
N
Single Termination Path in Model
Multiple Termination Paths in Model
Computationally efficient scoring scheme
Can a second conserved sequence Co-Locate the target?
Implement Co-Location
Y
N
Does the target contain inherent variability?
Implement Cumulative-Single Scoring
Y
N
Has the HMM been trained using the most representative data-set?
Y
N
Use the most representative data-set
Is the model intolerant to non-consensus sequences?
Apply Consensus blurring
Y
N
Final Model

Figure 9. Generic Model Development

Future optimisation of the framework (both model and scalable JAVA componentry) suggests the integration into a hierarchical scheme Figure 10. This figure serves many purposes, the Hidden Markov Model in conjunction with our scoring algorithms (dotted box) can be reused for identification of **other promoter elements**. Further integration with HMM promoter identification may lead to improved **promoter element** recognition. Furthermore **promoter identification** would be enhanced when integrated with **gene identification**. The hierarchical integration scheme benefits recognition by placing the HMM results in biological context. At each level, the range of the pattern **recognition domain is shortened hence recognition improved**. This addresses the limited length of the target sequence, a limitation of the HMM.

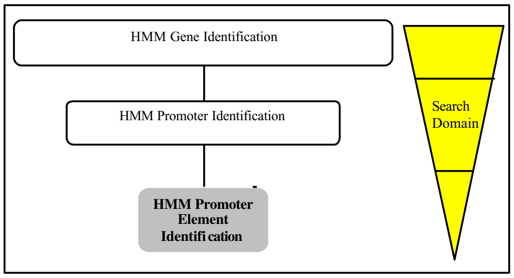
HMM Gene Identification
HMM Promoter Identification
HMM Promoter Element Identification
Search Domain

Figure 10. Future Hierarchical Framework

While identification benefits from the reduction of the search domain, a clean, efficient object-oriented solution becomes apparent.

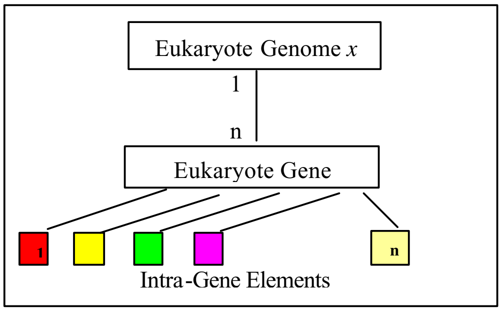

Figure 11. Scalable Model and JAVA component diagram

Figure 11 is represented as 1 genome having many genes with each gene having many intra-gene elements. All of these entities can be modelled as per object-oriented principles.

In conclusion, this paper has proposed an innovative framework for model optimisation leading to subsequence identification. Software engineering aspects have been discussed with their integration and future extensibility defined. The framework and associated software engineering techniques represent a scalable base for future applicability toward sequence identification.

11. Acknowledgement

Part of this work was supported by an Australian Research Council (ARC) Discovery grant DP0344488, and CE0348221 ARC Centre for Bioinformatics.

12. References

[1] N.Campbell, J.Reece, L.Mitchel, *Biology.* 5th Ed. Addison Wesley Longman, New York, 1999.

[2] H. Lodish, *Molecular Cell Biology*. 4th Ed. Houndsmills, New York, 2000.

[3] A. Krogh, "An Introduction to Hidden Markov Models for Biological Sequences", *Computational Methods in Molecular Biology*, 1998, pp. 45-63.

[4] M. Hewlett, http://www.blc.arizona.edu/marty/411/Modules/eukproms.html,October 2003.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell.* Garland Science, New York, 2002.

[6] U.Ohler, "Promoter Prediction on a Genomic Scale – The Adh Experience," *Genome Research* **10**, 2000, pp. 539-542.

[7] M. Fowler and K. Scott, *UML Distilled: A Brief Guide to the Standard Object Modeling Language.* Addison Wesley, 1999.