Report Number: WUCSE-2011-78

2011

Optimization of Gene Prediction via More Accurate Phylogenetic Substitution Models

Ezekiel Maier, Randall H. Brown, and Michael R. Brent

Determining the beginning and end positions of each exon in each protein coding gene within a genome can be difficult because the DNA patterns that signal a gene's presence have multiple weakly related alternate forms and the DNA fragments that comprise a gene are generally small in comparison to the size of the genome. In response to this challenge, automated gene predictors were created to generate putative gene structures. N SCAN identifies gene structures in a target DNA sequence and can use conservation patterns learned from alignments between a target and one or more informant DNA sequences. N SCAN... **Read complete abstract on page 2.**

Optimization of Gene Prediction via More Accurate Phylogenetic Substitution Models

Ezekiel Maier, Randall H. Brown, and Michael R. Brent

Complete Abstract:

Determining the beginning and end positions of each exon in each protein coding gene within a genome can be difficult because the DNA patterns that signal a gene's presence have multiple weakly related alternate forms and the DNA fragments that comprise a gene are generally small in comparison to the size of the genome. In response to this challenge, automated gene predictors were created to generate putative gene structures. N SCAN identifies gene structures in a target DNA sequence and can use conservation patterns learned from alignments between a target and one or more informant DNA sequences. N SCAN uses a Bayesian network, generated from a phylogenetic tree, to probabilistically relate the target sequence to the aligned sequence(s). Phylogenetic substitution models are used to estimate substitution likelihood along the branches of the tree. Although N SCAN's predictive accuracy is already a benchmark for de novo HMM based gene predictors, optimizing its use of substitution models will allow for improved conservation pattern estimates leading to even better accuracy. Selecting optimal substitution models requires avoiding overfitting as more detailed models require more free parameters; unfortunately, the number of parameters is limited by the number of known genes available for parameter estimation (training). In order to optimize substitution model selection, we tested eight models on the entire genome including General, Reversible, HKY, Jukes-Cantor, and Kimura. In addition to testing models on the entire genome, genome feature based model selection strategies were investigated by assessing the ability of each model to accurately reflex the unique conservation patterns present in each genome region. Context dependency was examined using zeroth, first, and second order models. All models were tested on the human and D. melanogaster genomes. Analysis of the data suggests that the nucleotide equilibrium frequency assumption (denoted as πi) is the strongest predictor of a model's accuracy, followed by reversibility and transition/transversion inequality. Furthermore, second order models are shown to give an average of 0.6% improvement over first order models, which give an 18% improvement over zeroth order models. Finally, by limiting parameter usage by the number of training examples available for each feature, genome feature based model selection better estimates substitution likelihood leading to a significant improvement in N SCAN's gene annotation accuracy.

2011-78

Optimization of Gene Prediction via More Accurate Phylogenetic Substitution Models

Authors: Ezekiel Maier, Randall H Brown, and Michael R Brent

Abstract: Determining the beginning and end positions of each exon in each protein coding gene within a genome can be difficult because the DNA patterns that signal a gene's presence have multiple weakly related alternate forms and the DNA fragments that comprise a gene are generally small in comparison to the size of the genome. In response to this challenge, automated gene predictors were created to generate putative gene structures. N SCAN identifies gene structures in a target DNA sequence and can use conservation patterns learned from alignments between a target and one or more informant DNA sequences. N SCAN uses a Bayesian network, generated from a phylogenetic tree, to probabilistically relate the target sequence to the aligned sequence(s). Phylogenetic substitution models are used to estimate substitution likelihood along the branches of the tree.

Although N SCAN's predictive accuracy is already a benchmark for de novo HMM based gene predictors, optimizing its use of substitution models will allow for improved conservation pattern estimates leading to even better accuracy. Selecting optimal substitution models requires avoiding overfitting as more detailed models require more free parameters; unfortunately, the number of parameters is limited by the number of known genes available for parameter estimation (training). In order to optimize substitution model selection, we tested eight

Type of Report: Other

Optimization of Gene Prediction via More Accurate Phylogenetic Substitution Models

Ezekiel Maier, Randall H Brown, and Michael R Brent

Department of Computer Science and Engineering,
Washington University, Saint Louis, MO, 63130

Abstract: Determining the beginning and end positions of each exon in each protein coding gene within a genome can be difficult because the DNA patterns that signal a gene's presence have multiple weakly related alternate forms and the DNA fragments that comprise a gene are generally small in comparison to the size of the genome. In response to this challenge, automated gene predictors were created to generate putative gene structures. N-SCAN identifies gene structures in a target DNA sequence and can use conservation patterns learned from alignments between a target and one or more informant DNA sequences. N-SCAN uses a Bayesian network, generated from a phylogenetic tree, to probabilistically relate the target sequence to the aligned sequence(s). Phylogenetic substitution models are used to estimate substitution likelihood along the branches of the tree.

Although N-SCAN's predictive accuracy is already a benchmark for *de novo* HMM based gene predictors, optimizing its use of substitution models will allow for improved conservation pattern estimates leading to even better accuracy. Selecting optimal substitution models requires avoiding overfitting as more detailed models require more free parameters; unfortunately, the number of parameters is limited by the number of known genes available for parameter estimation (training). In order to optimize substitution model selection, we tested eight models on the entire genome including General, Reversible, HKY, Jukes-Cantor, and Kimura. In addition to testing models on the entire genome, genome feature based model selection strategies were investigated by assessing the ability of each model to accurately reflex the unique conservation patterns present in each genome region. Context dependency was examined using zeroth, first, and second order models. All models were tested on the human and *D. melanogaster* genomes. Analysis of the data suggests that the nucleotide equilibrium frequency assumption (denoted as π_i) is the strongest predictor of a model's accuracy, followed by reversibility and transition/transversion inequality. Furthermore, second order models are shown to give an average of 0.6% improvement over first order models, which give an 18% improvement over zeroth order models. Finally, by limiting parameter usage by the number of training examples available for each feature, genome feature based model selection better estimates substitution likelihood leading to a significant improvement in N-SCAN's gene annotation accuracy.

1. Introduction

1.1. Background

In the last decade many genomes have been fully sequenced. One of the earliest and most important steps toward understanding a genome is identifying biologically functional stretches of DNA. Embedded within a genome are genes, which encode proteins, used by the cell to mediate the building and operation of a complete organism. Although genes vary in size, there are several conserved features of a gene's structure. At the core of each protein-coding gene is the coding region, comprised of exon DNA sequences, which encode a protein. Intron sequences are interweaved between exons, but are removed by RNA splicing, and consequently are not part of the protein encoding. Marking the beginning and end of the coding region are two special sequences of three nucleotide bases, the start and stop codons. Flanking the coding region are non-coding untranslated regions of sequence.

Experimental approaches used to attain whole genome gene annotation reached a saturation point after determining 10,000-11,000 exact gene structures in the human and mouse genomes[7]. As a result, it was necessary to utilize computational approaches for generating putative gene structures. Automated *de novo* gene prediction tools identify and annotate genes using only genomic sequence as input. These predictors recognize patterns characteristic of coding regions, splice sites, translation initiation, termination sites and other structures. By 2004, the accuracy of gene predictors had reached the point that one half to two-thirds of all genes were predicted correctly in the *Arabidopsis thaliana* genome[10]. With the success of these algorithms, gene annotation has largely become a computational problem.

One of the first *de novo* gene predictors to perform well on eukaryotic genomes was GENSCAN[11]. GENSCAN uses a generalized hidden Markov model (GHMM) to predict gene structures in a target sequence, using only that sequence as input. This system models both intergenic stretches of sequence between genes and conserved features of a gene's structure as GHMM states. Probabilistic models of the GHMM states are used to score the sequence as it is read. The Viterbi algorithm is used to predict gene structures by calculating the most likely series of GHMM states using the scored sequence and transition probabilities between GHMM states. Both the state probability models and state transition probabilities are calculated using known gene structures. Despite its place as a benchmark tool, it became clear that GENSCAN predicts too many genes. For the human genome, GENSCAN predicted 45,000 genes, which is nearly twice the size of later estimates of 20,000-25,000 genes[8,16]. With this result, it became clear that new methods were required to improve gene prediction.

A draft of the mouse genome made it possible for the first time to incorporate a second genome for comparison to improve gene prediction tools[1]. Dual genome *de novo* gene prediction tools use patterns of conservation learned from alignments between a target genome and sequences from an informant genome to improve predictive accuracy on the target genome. The informant genome sequences do not need to be assembled into a genome, nor do the sequences need to be annotated. The informant genome is used to identify conserved regions of the target sequence, through target-informant sequence alignments, because conserved sequence is more likely to be under selective pressure and therefore to have a biological function. One of the first gene predictors to utilize this signal to significantly exceed the GENSCAN benchmark was TWINSCAN[12-14]. This was achieved by adding a model of conservation in each state of the successful GENSCAN GHMM. TWINSCAN measures conservation between sequences by converting local alignments into a conservation sequence in which each target nucleotide either has a matching nucleotide in the aligned informant, is mismatched or not present in the informant, or is not part of a local alignment between the target and informant. This conservation sequence is used to estimate the selective pressure on a given region and influences decisions about the

most likely genomic feature, modeled as a GHMM state, for each base. The success of the dual genome gene prediction tool TWINSCAN motivated the development of generalized approaches which can use multi-genome alignments.

1.2. N-SCAN

N-SCAN extends the TWINSCAN model to allow an arbitrary number of aligned informant sequences. N-SCAN achieved substantially better performance than other *de novo* systems on both whole gene and exon prediction[1]. Since that time, the tool's standing as a benchmark for GHMM based *de novo* gene predictors has not changed[15-16]. With the use of four plant genome informant sequences, N-SCAN was used to improve the structural annotation of the rice genome[17]. In addition, N-SCAN has been used to annotate many genomes, including the human genome[8] and current work on developing a comprehensive annotation of the Drosophila genome[18].

N-SCAN utilizes a Bayesian network, generated from a phylogenetic tree rooted at the informant genome, to probabilistically relate the target sequence to the aligned sequence(s). The success of N-SCAN can be traced to its ability to learn and exploit any patterns within the aligned sequence(s) that may be useful for gene prediction. An example alignment between human, the target genome, and the mouse and chicken informant genomes is shown below in Fig. 1.

Human	A	A	C	A	G	C	C	T	G	A	C	T	A	G	G	A	C	T
Mouse	●	A	C	-	-	C	C	T	G	A	-	T	A	G	G	A	C	-
Chicken	A	T	G	A	-	C	C	T	G	A	-	T	A	-	G	A	C	●

Fig. 1. An N-SCAN multi-genome alignment with Human as the target and Mouse and Chicken as informants. Dashes indicate a gap within the alignment and dots indicate a region where no alignment can be identified.

N-SCAN recognizes conservation patterns by using substitution models which have a similar form to models of molecular evolution. These matrices describe the substitution patterns of DNA bases or amino acids along a tree structure (phylogenetic tree) which represents the evolutionary relatedness of the species. In Fig. 2 below a DNA base substitution matrix is shown.

	A	C	G	T
A		a	b	c
C	d		e	f
G	g	h		i
T	j	k	l	

Fig. 2. Each lowercase character represents a probability for seeing the labeled substitution along the evolutionary tree. The A→C substitution probability is represented by the character 'a'. The diagonal, or match probability, can be computed from the rest of the row, as the row must sum to one.

Unlike models of molecular evolution, N-SCAN substitution models are not limited to the nucleotide alphabet. In addition to nucleotide substitution, N-SCAN also uses indel (insertion or deletion) mutations and unaligned nucleotide frequencies within the multiple genome alignment to gauge the selective pressure exerted on each genome region. Overall, N-SCAN substitution models track occurrences of target base matches, mismatches, gaps, and unaligned regions in the informant

alignment. The frequency of each substitution is learned (trained) from multi-genome alignments of known gene structures. A trained substitution model is produced for each GHMM state and these models are used to calculate the probability of the conservation pattern observed at the current column of the multi-genome alignment. Similar to TWINSCAN, the probability of each GHMM state for a given alignment column is calculated by multiplying probability of the target model and the conditional probability of the informant base given the target base.

1.3. Substitution Models

There have been many proposed substitution models in molecular evolution literature[4]. These models vary drastically in their assumptions and number of parameters which must be fit to estimate substitution probabilities. One of the simplest models, the Jukes-Cantor model, uses only a single parameter to measure substitution rates and assumes that each base is substituted with equal rate[5]. Hence, this model allows for only two frequencies within the substitution matrix, a measured likelihood for the mutation of any base to another base, and the likelihood of matching nucleotides. Adding a second parameter, the Kimura model incorporates the biologically relevant possibility that mutations involving bases of the same structure (A↔G or C↔T), known as transitions, occur more frequently than mutations involving bases of different structure, known as transversions[6]. At the other end of the parameter usage spectrum, the fully general model (General model) uses a parameter to model each of the 12 nucleotide substitutions, allowing for each substitution to occur at a different rate. The remaining 4 nucleotide match rates can be computed from the rest of the row, as the row must sum to one. Because N-SCAN also measures rates of gaps and unaligned columns within the multi-genome alignment, substitution models must be extended with additional free parameters to measure each rate. The Kimura substitution model implemented in N-SCAN is shown below in Fig. 3.

	A	C	G	T	-	●
A		β	β	α	δ	ε
C	β		α	β	δ	ε
G	β	α		β	δ	ε
T	α	β	β		δ	ε

Fig. 3. The N-SCAN Kimura model. Each row represents the probability distribution on symbols in the aligned informant species, given a symbol from the target species.

A common assumption of more recent substitution models is that each nucleotide is expected to make up the same fraction of all nucleotides before and after the substitutions. The nucleotide equilibrium frequency assumption (denoted as π_i) is enforced by multiplying each substitution probability by the fraction of the target species nucleotide present in the columns of substitution model. Since each π_i can be estimated from the genome, the nucleotide equilibrium assumption does not add any additional parameters. Hasegawa and coworkers (HKY model) modified the Kimura model by adding this assumption[19]. Hence, the HKY model allows for differing match, transition, and transversion frequencies, while enforcing nucleotide equilibrium. A more general model using half the number of parameters as the fully generalized model is the General Reversible model, shown below in Fig. 4 (Reversible model)[20-21]. This approach assumes that each substitution and its reverse occur at the same rate, and incorporates the nucleotide equilibrium assumption.

	A	C	G	T	-	●
A		$\alpha\pi_C$	$\beta\pi_G$	$\gamma\pi_T$	δ	ε
C	$\alpha\pi_A$		$\zeta\pi_G$	$\eta\pi_T$	δ	ε
G	$\beta\pi_A$	$\zeta\pi_C$		$\theta\pi_T$	δ	ε
T	$\gamma\pi_A$	$\eta\pi_C$	$\theta\pi_G$		δ	ε

Fig. 4. The N-SCAN Reversible model uses six parameters to measure nucleotide substitution, and a parameter for gaps and unaligned columns.

Also affecting the quality of the substitution probability estimations and the number of parameters is the substitution model order. The order of the model refers to the number of previous columns to consider in the target/informant alignment as context when evaluating the substitution probability of the current alignment column. Increasing the model order allows a substitution model to better estimate substitution frequencies, because substitution rates are context dependent in that they depend on the identity of neighboring bases. Nucleotides that flank a site have a large effect on substitution rate, but effects lessen as distance increases[26-27]. In coding regions, selection of bases acts primarily on the level of codons[19,22], a three base unit of DNA which codes for a specific amino acid. Therefore, incorporating the effects of flanking bases allows the model to capture codon substitution rates. Order 0 substitution models unrealistically assume that the probability of a substitution occurring is not affected by previous alignment columns. Figures 3 and 4 present order 0 substitution models. Each of these models could be extended to order 1 models by expanding the table to include all 16 possible dimers along both axis. Order 2 models are of particular interest because the additional context allows the model to capture codon substitution patterns.

Considering only nucleotide substitutions, a fully general model of order o estimates the substitution of any (o+1)-mer for any other (o+1)-mer. As the DNA alphabet size is 4, there are $4^{(o+1)}$ different (o+1)-mers. Therefore, allowing for all possible (o+1)-mer substitutions requires $4^{2(o+1)}$ fitted parameters, but since the match probabilities can be computed from the rest of the probabilities, $4^{2(o+1)} - 4^{(o+1)}$ fitted parameters are required. The number of fitted parameters required for each model at each order is shown below in Fig. 5. In general, the number of parameters which must be fit increases exponentially with the order of the substitution model. Thus, a higher model order can more accurately capture the effect of context dependence on nucleotide substitution, but more parameters must be fit to achieve this accuracy.

Model Order				
	0	**1**	**2**	**3**
General	12	240	4032	65280
Reversible	6	120	2016	32640
Kimura	2	8	26	80
Jukes-Cantor	1	1	1	1

Fig. 5. The number of free parameters used to learn nucleotide only substitution probabilities for various models of order 0-3.

2. Optimizing N-SCAN Substitution Models

2.1. Parameter count limitations

There is a tradeoff in selecting substitution models for use in gene prediction. More accurate models

generally require more parameters, but the accuracy of parameter estimations can be reduced as the number of parameters is increased. This occurs because there are a limited number of annotated genes available for training. Obtaining maximal N-SCAN predictive accuracy requires optimizing the selection of substitution models based on the number and use of parameters to provide accurate substitution likelihood estimates with the available training data.

The parameters of each model which are used to fit substitution frequencies are learned from training data. Because each substitution model is specific to a GHMM state, training examples for each model must come from an annotated example of the genome feature being modeled by the state. Hence, a training example for an order 0 coding sequence substitution model consists of a single column of a target-informant alignment in coding sequence. Annotated genes, necessary for training a model, can be severely lacking in many genomes, resulting in sparse training data. The models described above can be estimated by a fast EM procedure. Thus, the main limitation on model parameter count is the availability of training data. Because each training example is used in the estimation of one parameter, models with too many parameters are susceptible to overfitting.

Within the human mouse whole-genome alignment, the substitution probability estimated by each parameter can be calculated with an error of less than ±1% by 25 training examples per parameter. However, more training examples per parameter will result in better substitution probability estimations, until diminishing returns are reached when using more than 50 training examples per parameter. Therefore, an order 0 fully general coding sequence substitution model requires at minimum an alignment sequence of 300 non-match substitution examples, split evenly among the 12 parameters. But, the types of substitutions will not be spread evenly throughout the whole-genome alignment and matches in many cases will be much more likely than mismatches. Thus, to accurately attain substitution probability estimations, sequences longer than the total number of training examples for a model are necessary. A sampling of the human mouse whole-genome alignment revealed that sequences 5 times as large as the number of parameters are satisfactory. Sufficient training examples for fitting these parameters are easily attainable in almost any genome, but the availability of training examples can easily be exhausted through a combination of modeling minimally annotated gene features and model context order increases. The human genome, which represents the best case for the amount of training data, contains annotations for approximately 150,000 exons, 130,000 introns, and 19,000 start and stop codons (hg18, RefSeq)[9,28]. Figure 6, shown below, presents the maximum fully general substitution model order which can be effectively trained by annotated examples of selected gene features in the human genome.

	Exon	**Intron**	**Start/Stop Codon**
Feature count	1.5×10^5	1.3×10^5	1.9×10^4
Training Examples (Sequence length)	2.5×10^7	7.1×10^8	5.7×10^4
Maximum model parameters	$\sim 10^5$	$\sim 10^6$	$\sim 10^2$
Maximum model order	2 or 3	3 or 4	0 or 1

Fig. 6. The implications for the maximum number of parameters and hence maximum order of a fully generalized substitution model on the human genome.

2.2. Substitution model selection investigation

To improve gene prediction accuracy, we investigated the space of nucleotide substitution models. Originally, only the 1st order reversible model was implemented in N-SCAN. Although this model aided N-SCAN in achieving much success[1, 15-16], it is possible that a different substitution model parameterization or larger context order could result in better predictions. To address this we considered eight nucleotide substitution model formalisms: General, General with nucleotide equilibrium, Reversible, Blaisdell[24] (assumes opposite strand mutations occur at the same rate), HKY, Kimura, Felsenstein[23] (rate of substitution depends only on nucleotide equilibrium frequency), and Jukes-Cantor. For each of these eight substitution models, we implemented their zeroth, first and second context order parameterizations in N-SCAN. For each of the 24 substitution model implementations we tracked the number of model parameters against available training examples, to test for potential cases of poor substitution probability estimations caused by too few training examples. Finally, we evaluated each model implementation for use in vertebrate and invertebrate gene prediction by generating computational annotations of the human and *D. melanogaster* genomes.

2.3. Extending N-SCAN for state based model selection

From our initial investigation, we realized that choosing a single best substitution model for gene prediction was not appropriate. Instances of the chosen model are fitted using available training examples from each gene feature modeled by a GHMM state. However, as Figure 6 demonstrates, the number of training examples can vary by several orders of magnitude between gene features. The fluctuation in available training examples creates a scenario in which the context order of the model and hence the number of parameters could be too large, smaller than necessary, and proper depending on the gene feature being fitted. In addition, the substitution processes of similar genome features often have common properties. These common properties can be due to similar evolutionary pressure exerted on the features.[31] However, different genome features often have different substitution properties due to different evolutionarily pressures. Therefore, we modified the default behavior of N-SCAN so that a separate substitution model and context could be chosen for each gene feature. Using this new behavior, we searched for the best model parameterization to fit each gene feature. In the process, we developed a multi-model approach which avoids under-training and improves gene prediction.

2.4. Experimental Design

Because we are using these models within the N-SCAN architecture, our evaluations are based on overall gene prediction accuracy rather than substitution probability estimations. Testing gene prediction accuracy requires predictions to be made on genomes with known gene structures. To evaluate both vertebrate and invertebrate gene predictions, we made predictions on both the human and *D. melanogaster* genomes. For human gene prediction, the mouse genome was used as an aligned informant, and for *D. melanogaster* prediction, the *D. ananassae* genome was used as an aligned informant. A cross validation approach was used to evaluate predictions, in which the set of known genes was partitioned into four groups of equal size. One group of genes was held out as a test set, while the other three groups were used for parameter estimation. For each substitution model implementation, whole genome gene prediction was repeated four times, holding out each test group of genes once. The results were averaged over the four test sets to give an overall accuracy measurement for gene prediction on each genome using each substitution model implementation.

Gene prediction accuracy is typically measured in terms of sensitivity and specificity[25]. Gene

sensitivity (Sn) is defined as proportion of known genes which are correctly predicted and gene specificity (Sp) is defined as the proportion of predicted genes which are known genes. In both cases, a gene prediction is correct if and only if all exon boundaries within the gene correctly match a known annotated transcript. We used the Eval package[3] to generate sensitivity and specificity measurements by analyzing each computationally generated genome annotation against the held out annotation set. Since specificity measurements can be penalized by correct novel gene predictions, we scaled each specificity measurement (Sp*) by the number of expected genes over the number of annotated genes in each genome. The scaling factor used for human gene prediction evaluation was 1.588, which is calculated by dividing the 23,000 expected genes by 14,482 annotated genes in the RefSeq annotation[28] of the hg18 genome build[9]. For Drosophila evaluation, the scaling factor was 15,000 expected genes over 9,229 known genes in the RefSeq annotation of the dm3 genome build[29-30], or 1.625.

3. Results

3.1. Substitution model performance comparison

To test the effect of each substitution model and context order on N-SCAN's predictive accuracy, we generated cross validated gene predictions for the human and *D. melanogaster* genomes. Predictions were generated for the entire hg18 release of the human genome[9] and the dm3 release of the *D. melanogaster* genome[29-30]. The results of the *D. melanogaster* predictions are shown in Fig. 7. Human results are not shown, but conclusions made are supported by predictions in both genomes.

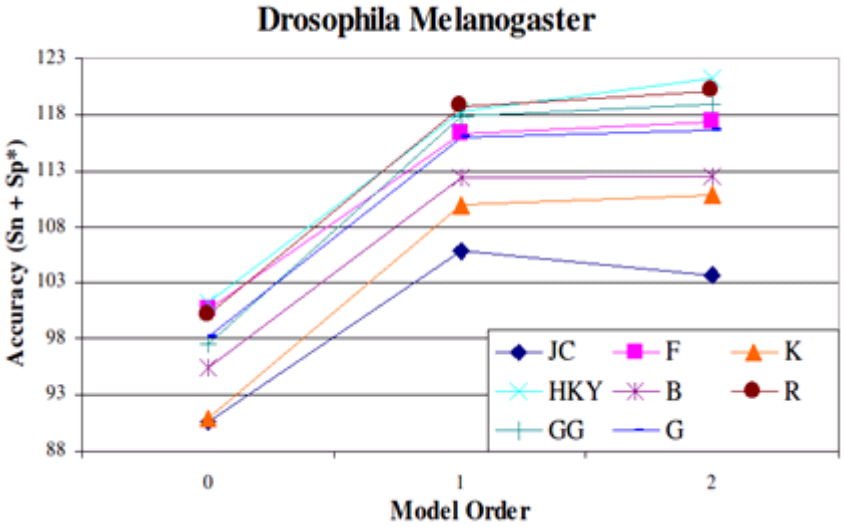

Fig. 7. N-SCAN prediction performance on the *D. melanogaster* genome using each substitution model at context orders 0-2.

These results show that N-SCAN can be improved over its Reversible(R) order 1 accuracy. The HKY and Reversible models are top models at all three context levels, and when using these models at context order 2, N-SCAN makes its best predictions. Although these two models are not similar in the number of parameters, they do make similar assumptions. Both the HKY and Reversible models are reversible, meaning that both models assume that a substitution and its reverse occur at the same rate. Therefore, both models allow for transitions and transversion substitutions to occur at different rates. In addition, both models incorporate the nucleotide equilibrium frequency assumption. When comparing the HKY and Kimura(K) models, it becomes apparent that the nucleotide equilibrium frequency assumption, an assumption taken by the HKY which distinguishes the models, has a strong effect on

prediction accuracy. The nucleotide equilibrium benefit is also apparent when comparing the results of the Felsenstein(F) and Jukes-Cantor(JC) single parameter models. As expected, the addition of context for evaluating substitution probabilities was clearly beneficial. Averaged over all models, an 18% improvement in gene prediction accuracy is seen when moving from an order 0 to order 1 substitution model. A further 0.6% improvement in prediction accuracy is seen when using an order 2 model over an order 1 model.

3.2. State based model performance

Using our knowledge of parameter training limitations and the general use substitution model results, we searched for a best substitution model for each feature modeled by a GHMM state. Our feature based substitution model is shown in Fig. 8.

Model	Feature
1st order Reversible	Conserved Non-coding Regions
2nd order Reversible	Coding Regions
2nd order HKY	Start/Stop Codons, Donor/Acceptor Sites
2nd order Blaisdell	Intergenic Regions

Fig. 8. The state based substitution model. Each model is only applied to its own features.

Previous studies of substitution patterns in well conserved genome regions have shown that the Reversible model more accurately captures substitution patterns than the HKY model[21], which has been shown to adequate for most purposes[32]. Because the Reversible model is more general than the HKY model, it can take advantage of different rates of A↔G versus C↔T substitutions[33], which the HKY model groups under a single transition substitution rate parameter. In addition, many training examples exist for coding regions, which allow the highly parameterized 2nd order Reversible model to be appropriately trained. Due to fewer training examples than Coding Regions, the 1st order Reversible model is optimal for Conserved Non-coding Regions. In addition, we believe the 2nd order HKY model is optimal for Start/Stop Codons and Donor Acceptor Sites, because it is able to incorporate 2nd order context while still being easily trainable by smaller training datasets. The 2nd order Blaisdell model is also highly parameterized and was created to model freely mutating sequence, fitting the specifications of intergenic regions.

We used this feature based multi-substitution model for human and *D. melanogaster* N-SCAN gene prediction. Shown in Fig. 9, this approach gave a 5.0% average improvement over order 1 single substitution models and a 2.4% average improvement over order 2 models. Hence, the use of this approach gives a significant improvement in prediction accuracy compared to single models.

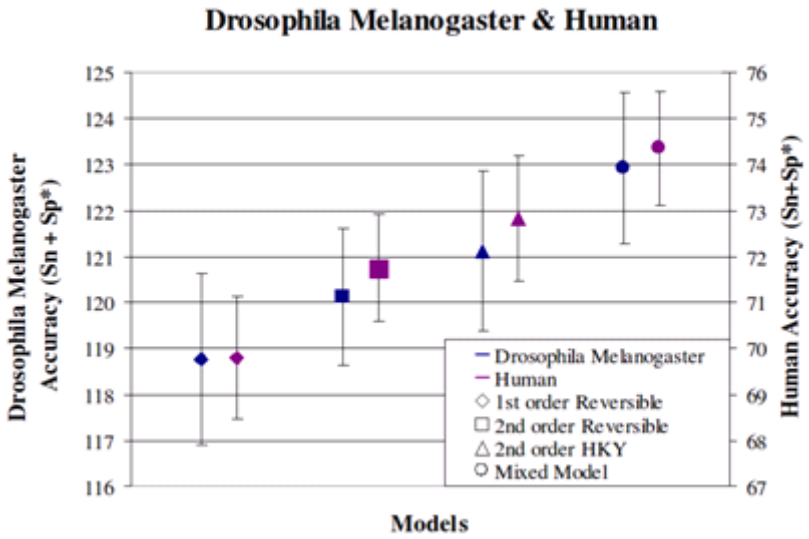

Fig. 9. Comparing N-SCAN prediction performance using a 1st order Reversible model, 2nd order Reversible model, 2nd order HKY model, and the state based multi-substitution model.

4. Discussion

We have investigated the space of possible substitution models for use in N-SCAN gene prediction and have found that proper substitution model selection can result in substantial accuracy gains. In most instances, the HKY and Reversible models should be considered the top general use substitution models. When choosing between these models, the HKY model should be used in instances when few training examples are available, while the reversible model requires many additional training examples.

Gene prediction accuracy is hurt by inaccurate substitution model parameter estimates caused by overfitting available training examples. Our analysis indicates that the fully general model results were diminished due to the large number of parameters in comparison to training examples in many features. Hence, when selecting a model and context order, optimizing the parameter demands against the available training data should be the first factor considered.

We have found that a greater context order can result in more accurate modeling of substitution patterns and improved gene prediction accuracy. These improved results are clearly visible between orders 0, 1, and 2 models. However, we believe the use of order 3 or larger models will not result in improved gene prediction. When comparing orders 1 and 2 models, a diminished effect is visible with increasing context order. In addition, for several models used in this study, the number of training examples was insufficient for accurate parameter estimation. Because two well annotated genomes were used in this study, more training examples were present than would be in many other genomes. Therefore, in many cases higher order substitution models cannot be effectively trained given the available data.

Because the number of training examples can vary by several orders of magnitude between gene features, selecting a single substitution model for use in all gene features can be difficult. To solve this problem, we have extended N-SCAN to allow for state based substitution model selection. Using this approach, we have shown that a significant improvement in predictive accuracy is achieved by avoiding under-training substitution models.

References

1. Gross S., Brent MR.. 2006. “Using Multiple Alignments to Improve Gene Prediction”. *J. Comput. Biol.* **13**:379-393.
2. Brent MR.. 2007. “How does eukaryotic gene prediction work?”. *Nature Biotechnology* **25**(8):883-885.
3. Keibler E., Brent MR.. 2003. “Eval: A software package for analysis of genome annotations”. *BMC Bioinformatics* **4**:50.
4. Li´o P. , Goldman N.. 1998. “Models of molecular evolution and phylogeny”. *Genome Research* **8**:1233-1244.
5. Jukes H, Cantor C.. 1969. “Evolution of Protein Molecules”. *New York: Academic Press*. 21-132.
6. Kimura M.. *1980. “*A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences”. *J. Mol. Evol.* **6**:111–120.
7. The MGC Project Team. 2004. “The Status, quality, and Expansion of the NIH Full-Length cDNA Project: The Mammalian Gene Collection (MGC)”. *Genome Research.* **14**: 2121-2127.
8. The MGC Project Team. 2009. “The completion of the Mammalian Gene Collection (MGC)”. *Genome Res*. **19**:2324-2333.
9. International Human Genome Sequencing Consortium. 2004. “Finishing the euchromatic sequence of the human genome”. *Nature* **431**:931–945.
10. Brent MR., Guigo R.. 2004. “Recent advances in gene structure prediction”. *Current Opinion of Structural Biology* **14**:264-272.
11. Burge C., Karlin S.. 1997. “Prediction of complete gene structures in human genomic DNA”. *J. Mol. Biol.* **268**:78-94.
12. Korf I, Flicek P, Duan D, Brent MR.. 2001. “Integrating genomic homology into gene structure prediction”. *Bioinformatics* **17**:140-148.
13. Flicek P, Keibler E, Hu P, Korf I, Brent MR.. 2003. “Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map”. *Genome Res.* **13**:46-54.
14. Tenney AE., Brown RH., Vaske C., Lodge JK., Doering TL., and Brent MR.. 2004. “Gene prediction and verification in a compact genome with numerous small introns”. *Genome Res.***14**:2330-2335.
15. Gross S., Do C., Sirota M., Batzoglou S.. 2007. “CONTRAST: a discriminative, phylogeny-free approach to multiple informant *de novo* gene prediction”. *Genome Biology* **8**:R269.
16. Brent MR.. 2008. “Steady progress and recent breakthroughs in the accuracy of automated genome annotation”. *Nat. Rev. Genet.* **9**:62-73.
17. Zhu W. Buell C.R.. 2007. “Improvement of whole-genome annotation of cereals through comparative analyses”. *Genome Res*. **17**:299-310.
18. The modENCODE Consortium. 2010. “Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE”. *Science*. **330**:1787-1797.
19. Hasegawa M., Kishino H., and Yano T.. 1985. “Dating of the human-ape splitting by a molecular clock of mitochondrial”. *J. Mol. Evol.* **22**:160-174.
20. Tavare, S.. 1986. “Some probabilistic and statistical problems in the analysis of DNA sequences”. *Lectures in mathematics in the life sciences* **17**: 57–86.
21. Yang Z.. 1994. “Estimating the pattern of nucleotide substitution”. J. *Mol. Evol.* **39**:105-111.
22. Siepel A., Haussler D.. 2004. “Phylogenetic estimation of context dependent substitution rates by maximum likelihood”. *Mol. Biol. Evol.* **21**:468-488.
23. Felsenstein J.. 1981. “Evolutionary trees from DNA sequences: A maximum likelihood approach”. *J. Mol. Evol.* **17**:368-376.
24. Blaisdell B.E.. 1985. “A method for estimating from two aligned present day DNA sequences their ancestral composition and subsequent rates of composition and subsequent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitutions at the same site”. *J. Mol. Evol.* **22**:69–81.
25. Burge C., Karlin S.. 1998. “Finding the genes in genomic DNA”. *Current Opinion in Structural Biology* **8**:346-354.

26. Blake R. D.,Hess S. T., Nicholson-Tuell. J.. 1992. “The influence of nearest neighbors on the rate and pattern of spontaneous point mutations”. *J. Mol. Evol.* **34**:189-200.
27. Hess, S. T., Blake J. D., Blake R. D.. 1994. “Wide variations in neighbor-dependent substitution rates”. *J. Mol. Biol.* **236**:1022-1033.
28. Pruitt KD, Tatusova T, Maglott DR. 2005 “NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins”. *Nucleic Acids Res.* 33:D501-504.
29. Celniker S.E. and Rubin G.M.. 2003. “The Drosophilia melanogaster genome”. *Annual Reviews of Genomics and Human Genetics* **4**:89-117.
30. Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Mendez-Lago M, Rossi F, Villasante A, Dimitri P, Karpen GH, Celniker SE.. 2007. “Sequence finishing and mapping of Drosophila melanogaster heterochromatin”. *Science* **316**:1625-1628.
31. Pedersen J.S., Forsberg R., Meyer I.M., and Hein J.. 2004. “An Evolutionary Model for Protein-Coding Regions with Conserved RNA Structure”. *Mol. Biol. Evol.* **21**:1913-1922.
32. Yap V.B., Speed T.P.. 2004. “Modeling DNA Base Substitution in Large Genomic Regions from Two Organisms”. *J. Mol. Evol.* **58**:12-18.
33. Tamura K., Nei M.. 1993. “Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees”. *Mol. Bio. Evol.* **10**:512-526