International Journal of Biomedical Data Mining

Ibrahim and Khedr, Int J Biomed Data Min 2018, 7:1
DOI: 10.4172/2090-4924.1000133

Research Article Open Access

Leveraging Pruning Techniques for Improving Generalized HMM Decoding in Gene Classification

Mohamed H Ibrahim[1,3] **and Ahmed M Khedr**[1,2*]

[1]*Department of Mathematics, Zagazig University, Zagazig, Egypt*

[2]*Department of Computer Science, University of Sharjah, Sharjah, UAE*

[3]*Department of Computer Engineering, Polytechnique Montréal, Montreal, Canada*

[*]**Corresponding author:** Khedr AM, Department of Computer Science, University of Sharjah, Sharjah 27272, UAE, Tel: +(971) 6 5053560; E-mail: akhedr@sharjah.ac.ae

Received date: March 6, 2018; **Accepted date:** March 23, 2018; **Published date:** April 2, 2018

Abstract

Gene sequence classification is a well-known problem that impacts several sub-disciplines of Bioinformatics including functional genomics and gene expression data analysis. In gene classification task gene families are frequently formulated using large Generalized Hidden Markov Models (GHMMs) representing a bottleneck for any decoding method and weakening its efficiency. Thus an efficient decoding of such GHMMs remains a key challenge. In this paper, we introduce a new pruned-based strategy for improving the decoding of GHMM using pruning techniques. We focus on viterbi decoding algorithm but the strategy is applicable to GHMM decoding in general. Unlike standard decoding methods, a paradigm shift from screening to-wards recognition is first performed to integrate all considered models into a combined state space. Then the decoding process is limited to the activated states within a beam around the optimal solution to significantly reduce the computational e ort, and thus greatly speeding up the model decoding. Our experiment on Eukaryotic gene demonstrates the e activeness of our approach for speeding up gene classification task.

Keywords: Beam-search; Gene expression; Generalized hidden markov model; Recognition; Pruning techniques; Screening

Introduction

Genomes expression data analysis of organisms has and will continue to produce large quantities of gene sequence data. Only very similar gene sequences are grouped together in databases called Gene Families. The genes in each family would be similar in functions or protein coding. These databases could be searched for other gene members. That is to say, searching similarities of gene sequences with genes databases can also be formulated as a pattern classification problem that is called gene classification. Often classifying the enormous number of genes into relatively small number of groups would be extremely useful to draw valuable information from the data and it is the basis for prediction of the functions of unknown genes.

The gene classification problem can be stated as follows: given a number of gene families and a target gene, how to find the most probable model in which the gene sequence belongs to it. Hidden Markov models (HMMs) [1] are commonly used to formulate gene families. HMMs are probabilistic graphical models that capture the dependencies between random variables in time-series data. They have been successfully applied to several areas of artificial intelligence such as speech recognition e.g. [2] robotics e.g. [3] pattern recognition [4] and several areas of bioinformatics, such as Trans membrane protein classification e.g. [5] to perform predictive and recognitive tasks. The power of HMMs stems from the provision of efficient and intuitive algorithms that grant HMMs their predictive and recognitive capabilities by computing quantities of interest described by the model. For example, given the specifications of the model, there exist efficient algorithms for computing the probability of observed events [6]. HMMs however, remain unexplored in application domains where they can be useful, by virtue of the unavailability of the statistical data necessary for the specification of the parameters of the model. Although overcoming the lack of real data by means of approximation [7] or synthesis [8] is possible for some applications, it is not an option for many types of applications. For example, epidemiological data describing factors influencing the occurrence of illnesses cannot be approximated or synthesized when not sufficient.

According to Bayesian measure, performing a gene classification task using HMMs always endowed with the evaluation process that takes the target gene sequence O and a set of n HMMs $\{\lambda i\}_i^n$ as input and produces the best HMM λ^* model that is suitable to the input gene sequence i.e., the HMM λ^*model that has the maximum value of the probability $P(O|\lambda)$. However, computing $P(O|\lambda)$ is equal to the sum over all maximum possible sequence of hidden states of probability $P(\Pi, O, \lambda)$ of the most likely sequence of hidden stateΠ , the gene sequence O and the HMM λ. Thus, in the evaluation process of HMM, one frequently resorts to apply a decoding algorithm such viterbi or 1-best to obtain the most likely sequence of hidden state Π.

Reviewing the Bioinformatics publications beyond these principle aspects, there is hardly any work discussing the optimization of the model evaluation itself. Performing database screening implies computing the classification scores of multiple models for the target sequences. Usually the evaluation is performed sequentially i.e., the scores for each model are calculated separately and the computational e ort is substantial. Even the widely used method of parallelization using specialized hardware does not solve the principle problem since it only heals the symptoms.

In order to improve the sequential search procedure using all known models, we reorganize the classical model evaluation. This offers the opportunity of integrated classification with these models. Pruning the search space of a HMM literally means reducing the number of explored states during the decoding process. In this paper, using the well-known beam-search approach we prune the search space of GHMMs which greatly reduces the computational e ort during model evaluation. Therefore, our proposed technique speeds up the gene classification by doing the following:

- Integrate all considered HMMs into a combined state space, so a paradigm shift from screening towards recognition is performed.
- Incorporate beam-search as a pruning technique [9] into the evaluation of HMMs to reduce the computations.

The rest of the paper is organized as follows: Section 2 covers the related work of our proposed problem. In section 3 we present a description of HMMs for pattern classification. Section 4 describes the proposed approach to speed up gene classification. In Section 5, the performance evaluation of the proposed technique is introduced. We conclude our work in Section 6.

Related Research

The challenge of gene finding in the genome is an important and di cult task. Finding genes will lead to more progress like knowing the functions of gene, predicting diseases, and determining spurious genes, also considered an important step in understanding the genome of a species and more. Using computational techniques, much e orts have been done in working with genes such as classify genes, identify common phenomena in known genes, describe the common phenomena, and scan uncharacteristic sequence to identify regions that match the model which become putative genes. The most two common computational approaches are:

Direct

In direct approaches, the exact or near-exact matches of DNA, or proteins from the same or closely related organism are used, there are many tools for finding gene based on direct approach such as FASTA [10] BLAST, and ESTGENOME [11].

Indirect

The indirect approaches are classified into the following:

- Comparative homology approaches: As the entire genomes of many different species are sequenced, a promising direction in current research on gene finding is a comparative genomics approach. This is based on the principle that the forces of natural selection cause genes and other functional elements to undergo mutation at a slower rate than the rest of the genome, since mutations in functional elements are more likely to negatively impact the organism than mutations elsewhere. Genes can thus be detected by comparing the genomes of related species to detect this evolutionary pressure for conservation. This approach was first applied to the mouse and human genomes, using programs such as SLAM, SGP and Twinscan/N-SCAN [12]. Parra et al. applied Comparative homology approach on Mouse and Human genomes [13]. Marina et al. applied Comparative approach on Human genome using SLAM program [14]. Wiehe et al. proposed a program called SGP based on homology approach [15]. Birney et al. proposed high quality annotations project from one genome to another for comparative gene finding by using GeneWise which is based on HMM [16]. In our proposed approach, we use Gene-Mark for nding genes, which is different from GeneWise, where GeneWise is a pair-HMM style with strong similarities to the more recent dual genome predictors called Double Scan, but Gene-Mark is a pro le-HMM style. However, there has been no work on the theory by which the GeneWise was developed, and also there is no details on the precise implementation of aspects of the GeneWise, but for Gene-Mark there have been many works discussed the precise implementation of it. Such techniques play the central role in the annotation of all genomes [17].
- *Ab initio* approaches: Because of the inherent expense and difficulty in obtaining extrinsic evidence for many genes, it is also necessary to resort to *Ab initio* gene finding, in which genomic DNA sequence alone is systematically searched for certain tell-tale signs of protein-coding genes. These signs can be broadly categorized as either signals, specific sequences that indicate the presence of a gene nearby, or content, statistical properties of protein-coding sequence itself. *Ab initio* gene finding might be more accurately characterized as gene prediction, since extrinsic evidence is generally required to conclusively establish that a putative gene is functional. [18,19]. Burge et al. proposed one of the most common program called Gene Scan for finding genes based on *Ab initio* approach which we use in our proposed approach [20].
- Hybrid approaches: The hybrid approach is the combination of homology comparative and *Ab initio* approaches. Yeh et al. proposed Genome Scan which is based on hybrid approach. GenomeScan is very similar to Gen Scan which is based on *Ab initio* approach, the only di erence between them is that Genome Scan finds genes in a single strand of DNA, but GENSCAN find genes in double strands of DNA sequence [21]. Krof et al. proposed Twinscan [22].

The best approaches for finding genes are the following: BLAST or FASTA as a direct approach, Gene-Mark based on HMM as a comparative homology approach, and Neural Networks (NNs) and GENSCAN based on HMMs *as Ab initio* and Hybrid approaches. Guig et al. showed that HMM is the best suited model for the problems that need grammatical structure such as gene finding and gene classification. All of the previous approaches are relied on HMMs [23].

HMMs for Pattern Classification

In the pattern recognition domain, memberships of sequences to patterns are recognized, but in Bioinformatics applications, sequence databases are screened for new targets of more or less abstract pattern families e.g. gene families. There is an enormous volume of literature on the application of HMMs to a broad range of pattern recognition tasks. The suitability and decay of HMMs is undeniable and so they are established as one of the major workhorses of the pattern recognition community.

Performing pattern classification with HMMs requires the robust estimation of their parameters, namely the transition and emission probabilities as well as the model structure. Using representative samples the probability values are usually estimated by means of variants of the EM algorithm called Baum-Welch. Once HMMs are established, they can serve as models for distinct pattern families e.g. gene families, and after that they have to be evaluated when classifying sequences of observations e.g. genes sequences. This evaluation is performed using the Forward algorithm in which the general

probability $P(O \mid \lambda_K)$of a HMM λ_K that produce the sequence O. This evaluation can be approximately done by using the common Viterbi algorithm. Viterbi algorithm additionally decodes the most probable sequence of chosen states Π producing the sequence of observations using Bayes rule and thus obtain the evaluation value as:

$$P(\Pi \mid O, \lambda) = \frac{P(\Pi, O \mid \lambda)}{P(O \mid \lambda)} \Rightarrow P(\Pi^*, O \mid \lambda) = \max_{\Pi} P(\Pi, O \mid \lambda) = P^*(O \mid \lambda) \quad (1)$$

When the evaluation of HMMs is performed via Viterbi algorithm, the states are arranged in a N M dimensional matrix V, where N is the number of states in HMM and M is the length of the sequence *O*.

HMMs evaluation aspects: screening vs. recognition

Besides the theoretical framework there are several aspects to be mentioned regarding the evaluation of HMMs in pattern classification applications.

Screening: In Screening a specific HMM λ_k represents model for a single pattern family K e.g. gene family. Traditionally, the databases are screened for more or less similar occurrences of such models. Therefore, for each part of the database O, some kind of score is computed representing the probability of λ_k producing the requested sequence O. As stated above, Viterbi algorithm is used to deliver the score while additionally respecting the most probable state sequence Π*. Generally the probability $P(O, \prod * \mid \lambda_K)$is defined as in Equation 1.

Recognition: Recognition tasks usually consist of the discrimination of multiple pattern e.g genes families, so several HMMs are established. In terms of Bioinformatics domain, the search for homologies of multiple different sequence families is performed on a database of sequences. The evaluations of each HMM λ_k are compared for all relevant sequences. The higher evaluation probability of the appropriate HMM is a stronger homology.

Accelerating the Evaluation Model

Since the number of sequences is enormous, homology detection as well as homology classification became very challenging task in terms of computational e ort dedicated to the search and classification problem. Usually, research in molecular biology is performed in a more or less iterative manner, i.e., once new insights are obtained new questions are to be answered. Mostly this implies new database searches in order to find gene sequences belonging to a particular gene family. Thus, efficient model evaluation techniques are mandatory when applying gene family models to the task of gene classification for huge amounts of data.

In Bioinformatics domain, the acceleration of the evaluation is performed algorithmically. Currently, biological sequence analysis is usually accelerated by increasing the computational power, i.e., by the deployment of more computers. Unfortunately, the general problem of extraordinary computational e ort still remains. Brute Force methods for example using specialized hardware for massive parallel model evaluation only treat the symptoms whereas the reasons for computationally expensive data base searches when applying GHMMs are usually not addressed. The dimensionality of the problem even increases when applying new methods for improving the general detection and classification performance as discussed before.

State space pruning

Analyzing the state-of-the-art in HMM based on gene sequence analysis and reconsidering the basic theory of HMMs, it becomes clear that the evaluation of gene family models is usually rather straightforward. This means that no optimizations either heuristic or theoretic are applied. In 1970, Bruce Lowered proposed the so-called Beam-Search algorithm for heuristic state space pruning during model evaluation [9]. Using this technique, substantial accelerations in HMM evaluation become possible. In this paper, using Beam-Search and Viterbi algorithms a new technique to prune the state space in evaluation process is proposed which greatly speeds up the computation.

Search space pruning

The Viterbi algorithm is widely used to find the most probable path Π* through the whole state space V of a HMM K producing the observation sequence O. The main idea for each step l of the incremental algorithm consists of the calculation of maximally achievable probabilities $\rho_l(i)$ for partial emission sequences $O_{1\ldots\ldots}O_l$ and state sequences$\pi_{1\ldots\ldots}\pi_l$.

$$\rho_l(i) = \max_{\Pi_1 \ldots \Pi_l - 1} \left\{ P(o_1, \ldots, o_l, \pi_1 \ldots \pi_l \mid \lambda_K) \,\middle|\, \pi_l = S_i \right\} \quad (2)$$

Since the dependencies of the HMM states are restricted to their immediate predecessors (Markov Assumption) the calculation of $\rho l + 1(j)$ is limited to the estimation of the maximum of the product of the preceding $\rho l(i)$ and the appropriate transition probability. Additionally the local contribution of the emission probability $e_j(ol + 1)$is considered. Stepping through the state space all $\rho l + 1(j)$ are recursively calculated using the following rule:

$$\rho_l + 1\,(j) = \max_i \left\{ \rho_l\,(i)\, \mathrm{t}_{ij} \right\} \cdot e_j(o_l + 1) \quad (3)$$

Finally the globally most probable state sequence is created by tracing back the local maxima. Considering the necessary computational e orts for Viterbi algorithm where a large number of possible paths needs to be considered. The more states which have to be explored at each step, the more continuations of all possible paths so far become reasonable. Thus, the number of traced paths dramatically increases and as a consequence the processing time for model evaluation will be increased. In order to reduce the processing time, Lowerre in [9] introduced the beam-search algorithm that establishes a dynamic criterion for search space pruning based on the relative differences of the partial path scores. The state space is pruned and the search is restricted to the promising paths only.

Some of HMMs are used in order to sufficiently represent the applications domain, many of the HMM states cover mutually quite different pattern families. Some parts of the state space of each HMM are quasi irrelevant for one particular final solution and the remaining states are activated regarding to the most probable path. Formally all states $\{\pi l\}$ that have $\rho l\,(i)$ are more or less similar to the locally optimal solution $\rho_l^* = \max_j \rho l\,(j)$are activated. The threshold of acceptable differences in the local probabilities is proportional to ρ_l^* by

the parameter B. So the set of activated states at a given time is located in a beam around the optimal solution and determined by:

$$\Lambda = \left\{ \pi_i \middle| \rho_l(i) \geq B \cdot \rho_l^* \right\} s.t\ \rho_l^* = \max_j \rho_l(j)\ and\ 0 < B << 1 \qquad (4)$$

The only parameter of this optimization technique is the beam-width B. Exploring the Viterbi matrix V at the next step l + 1, the new activated states will be the states that are treated as possible predecessors for the estimation of local path probabilities. Thus, at every Viterbi step the following modified rule is used for the estimation of $\rho l + 1\ (j)$ as:

$$\rho_l + 1\ (j) = \max_{i \in \Lambda_l} \left\{ \rho_l(i) t_{ij} \right\} e_j(o_l + 1) \qquad (5)$$

Note that the Beam-Search algorithm represents a heuristical approximation of the standard Viterbi procedure.

Combined state spaces

Pure classification tasks (in terms of classical pattern recognition) e.g. performing target classification for gene sequences, are calculated separately by every HMM finally the highest scoring.

Model determines the classification result. It is clear that, it will be more appropriate if all relevant models are treated combined instead of fully evaluating every model separately. Technically, this implies that the states of all HMMs are integrated into a global state-space which is conceptually segmented into the particular models. In Figure 1, the accelerated model evaluation process for gene family HMMs is illustrated. All known models of gene families (1,...,K) are integrated into a single combined state space (the grey shaded box). The evaluation process is shown in the diagram of the state space V on the right side. Large number of HMM states do not need to be activated (pruned states). The sequence O is classified using the combined state space by finding the Viterbi path (dashed line) through all models. The effect of state space pruning can be noticed via the ratio of activated states (black circles) to the overall number of states. For every Viterbi step only a moderate "Beam" of states around the Viterbi path is activated. Arranging all relevant gene families models (Figure 1), we can note the following:

At the beginning of a particular sequence classification process the initial states of all models are activated, i.e., they are treated as starting points for Viterbi path-search. These paths including their extensions of the remaining Viterbi steps need to be considered in parallel which is basically no advantage compared to the usual separate evaluation. However, applying the Beam-Search algorithm further substantial savings of necessary computations can be obtained in addition to those implied by local state-space pruning for single models.

Huge number of HMM states can be skipped for further evaluation. Reasoned by the avoidance of the exploration of devious paths within the combined state space where it is not necessarily to completely evaluate all known GHMMs. Contrary to this, when performing serialized evaluation of multiple models, at least one complete path through all particular models needs to be evaluated.

The lower model will not be evaluated since all successor states are pruned (global pruning) and for the remaining models a certain number of assigned states is pruned as well. At the end of combined model evaluation the index of the best fitting model is delivered including its score for the requested sequence. Compared to the conventional approach multiple repeats of this procedure are not necessary since a global classification is performed.

Due to the deployment of the beam-search pruning algorithm a substantial reduction of the computational complexity can be achieved for single model evaluations since irrelevant parts of the models need not to be explored by transfer recognition into screened. So for screening applications all considered models are integrated into a combined search space yielding further savings of necessary computations. Reasoned by the avoidance of the exploration of devious paths within the combined state space not necessarily all known GHMMs need to be evaluated completely.

Figure 1 illustrates the proposed model evaluation process for GHMMs. We choose the search for homologies of gene families. All known GHMMs of gene families (λ_1, λ2,...λKon the left side of the sketch) are integrated into a single combined state space (the grey shaded box). The evaluation process is performed on the state space V on the right side. Following our approach lots of HMM states do not need to be activated-they are pruned (empty circles). The sequence O is classified using the combined state space by finding the Viterbi path (dashed line) through all models. The effect of state space pruning can be noticed via the ratio of activated states (black solid circles) to the overall number of states (all circles). For every Viterbi step only a moderate "beam" of states around the Viterbi path is activated.

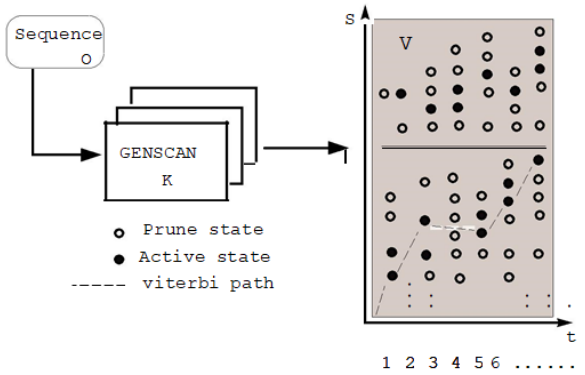

Figure 1: The explored states in case of combined and pruned state space.

Results and Discussion

The plain Viterbi algorithm always ensures that the most probable path through the state space for a given observation sequence is found because the whole Viterbi matrix is examined. The existing pruning techniques like the beam-search are in principle suboptimal since only a subset of HMM states are explored. Therefore, the possibility to miss the optimal solution exists. Contrary to the existing pruning techniques, the capabilities of our proposed pruning approach are assessed concerning the maximally possible search space reduction together with the increasing classification error rate. We performed baseline experiments evaluating the Viterbi matrices in the conventional way without pruning. Adjusting the beam-width B according to this constraint then we evaluate our approach for a given corpus comparing to the baseline results in two stages:

- All known models are subsumed in order to create a combined state space allowing integrated search for the most probable path.
- For all sequences, GHMMs of our working domain are evaluated sequentially including the beam-search pruning. The main idea here is to demonstrate the local search space reduction for single model. Thus mutual influences of the different search processes are neglected by keeping the sequential processing methodology.

We establish GHMMs as stated before using a GENSCAN package. The domain for our experiments is finding homology members of gene families. Thus, we trained several GHMMs using multiple alignments of the likewise publicly available gene family library [24]. The recognition domain consists of 4 models. Randomly separating sample sequences of each gene family delivers a test set used for the experiments. We separated 80 sequences for every family yielding 320 test sequences. For assessing the efficiency of our approach, we initially performed experiments without any pruning of the state space using the recognizer of the GENSCAN package. Since the HMMs used in both experiments are identical, similar recognition rates were achieved. Performing the GENSCAN baseline experiment results in 95:1% classification accuracy.

In Figure 2, the efficiency of the pruning techniques are illustrated, where the observed slight improvements over the baseline results are caused by artifacts of the beam-search algorithm and are in general not significant. Yielding evidences for the efficiency within the proposed combined state space as well as for the traditional sequential model evaluation two test series are plotted in the diagram.

- The recognition rate for experiments using the combined search space depending on the appropriate beam-widths is shown using the dot line.
- The dash line indicates the average recognition rates of experiments performed in the traditional way of classifying sequentially. All 4 models are evaluated separately for all 320 sequences with varying beam-widths reducing the local search space of the appropriate model.

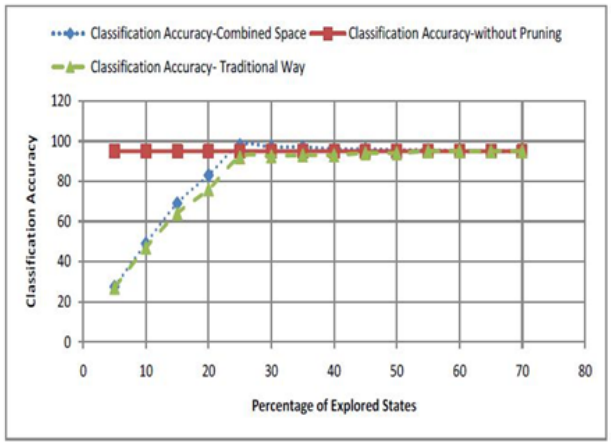

Figure 2: Efficiency of search space pruning for combined and local search using various beam-widths.

The varying beam-widths B for all experiments are not shown directly but by the percentage of overall explored states. Additionally, the classification accuracy of the baseline experiment (without pruning) is shown using the sold line.

It can be seen that only a small fraction of the states actually need to be explored while keeping the classification accuracy as high as if the complete state space would be considered. As a consequence of these great reductions of computational effort can be achieved. Using the classification paradigm with combined search space only about 25% of the HMM states needs to be explored. Comparing to the standard experiment without pruning on a CPU of 2:8 MGz and 512 MB of RAM, the run time of the model evaluation for the test set could be reduced by approximately a factor of 4. The absolute run times for the standard experiment were 204 seconds whereas the pruning experiment takes about 51 seconds.

Furthermore, even when evaluating the models sequentially, as in most present GHMM applications where databases are screened using a single target HMM, a large number of states does not need to be explored. Here local reductions of up to 50% of explored states can be achieved. On the same computational equipment as described above the model evaluation could be accelerated by a factor of two. For this type of experiments the absolute run times are 220 seconds for the standard experiment and 112 seconds in pruning case. With small modifications even on highly specialized parallel hardware can immediately save the computational e ort and so speed up the evaluation of GHMMs two times.

The combined state space only makes sense when actually deploying the pruning method. Since the states of all models are integrated into a global state space without pruning more alternative successors for each state become possible. Deploying the beam-search algorithm to this global state space after the first evaluation step most of the states not belonging to the most probable model are deactivated. Besides the local pruning within the most probable model in summary, we have stronger reduction rates of the global state space than the local state spaces (factor of 4 vs. factor of 2). Using the combined state space in our implementation the absolute run time is only half of the sequential case. Thus the overall run time can be reduced to a third of the standard GHMM evaluation.

Conclusion

The well-known beam-search algorithm is used to reduce the number of activated explored states in Viterbi decoding process. The reduction of the number of activated states significantly speeds up the process of evaluation model. Performing a paradigm shift from screening towards recognition, where we subsumed all considered models into a combined state space then the Viterbi-path through all models is determined by an integrated evaluation approach. For a representative test, we perform several experiments on a set of 320 sequences belonging to 4 different gene families. Following our recognition approach, the combined state space could be significantly pruned and only about 25% of all states need to be explored. Deploying the pruning approach to the conventional process of sequentially evaluating the appropriate model yields average reductions of about 50% of the states explored while decoding. Comparing to the classical evaluation of GHMMs in our implementation within the traditionally framework the absolute run times depending on the state space organization could be accelerated by factors of 4 (for the combined approach) and 2 (for the conventional screening approach) respectively. The proposed approach is generally applicable to a wide variety of Bioinformatics tasks. Even high throughput applications

which are presently often performed on massive parallel architectures or specialized hardware can be done using the proposed approach of local pruning.

References

1. Rabiner R (1989) tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77: 257-289.
2. Rosti AVI, Gales MJF (2003) Factor analyzed Hidden Markov Models for speech recognition. Cambridge University, England.
3. Fox M, Ghallab M, Infantes G, Long D (2006) Robot introspection through learned Hidden Markov Models. Artificial Intelligence 170: 59-113.
4. Lovell BC (2003) Hidden Markov Models for spatio-temporal pattern recognition and image segmentation. International Conference on Advances in Pattern Recognition, Calcutta.
5. Kahsay RY, Gao G, Liao L (2005) An improved Hidden Markov Model for trans-membrane protein detection and topology prediction and its applications to complete genomes. Bioinformatics 21: 853-1858.
6. Hand DJ (2006) Protection or privacy? data mining and personal data. Advances in Knowledge Discovery and Data Mining (PAKDD), UK.
7. Smyth P (1997) Belief Networks, Hidden Markov Models, and Markov Random Fields: A unifying view. University of California, USA.
8. Ramezani V, Marcus S (2002) Estimation of Hidden Markov Models: Risk-sensitive filter banks and qualitative analysis of their sample paths. IEEE Transactios on Automatic Control 47: 1000-2009.
9. Lowerre BT (1976) The Harpy speech recognition system. Carnegie-Mellon Universit, Pittsburgh.
10. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85: 2444-2448.
11. Paquola ACM, Nishyiama MY, Reis EM, da Silva AMD, Verjovski-Almeida S, et al. (2003) ESTWeb: Bioinformatics services for EST sequencing projects. Bioinformatics 19: 1587-1588.
12. Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16: 512-524.
13. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, et al. (2003) Comparative gene prediction in human and mouse. Genome Res 13: 108-117.
14. Alexandersson M, Cawley S, Pachter L (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res 13: 496-502.
15. Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guig R (2001) Comparative genomics: At the crossroads of evolutionary biology and genome sequence analysis. Genome Res 11: 1574-1583.
16. Birney E, Durbin R (2000) Using GeneWise in the drosophila annotation experiment. Genome Res 10: 547-548.
17. Siepel A, Haussler D (2004) Computational identification of evolutionarily conserved exons. Proceedings of 8th International Conference on Research in Computational Molecular Biology, University of California, Santa Cruz.
18. Nekrutenko A, Chung WY, Li WH (2003) An evolutionary approach reveals a high protein-coding capacity of the human genome. Trends Genet 19: 306-310.
19. Salamov AA, Solovyev VV (2000) Ab initio gene finding in drosophila genomic DNA. Genome Res 10: 516-522.
20. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78-94.
21. Yeh RF, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. Genome Res 11: 803-816.
22. Korf I, Flicek P, Duan D, Brent MR (2001) Integrating genomic homology into gene structure prediction. Bioinformatics 17: 140-148.
23. Guigó R, Agarwal P, Abril JF, Burset M, Fickett JW (2000) An assessment of gene prediction accuracy in large DNA sequences. Genome Res 10: 1631-1642.
24. https://en.wikipedia.org/wiki/List_of_24_characters