COMPARISON OF NML AND BAYESIAN SCORING CRITERIA FOR LEARNING PARSIMONIOUS MARKOV MODELS

Ralf Eggeling[1]*, Teemu Roos*[2]*, Petri Myllymäki*[2]*, Ivo Grosse*[1]

[1]Institute for Computer Science, Martin Luther University Halle-Wittenberg,
06099 Halle, GERMANY, {eggeling|grosse}@informatik.uni-halle.de

[2]Helsinki Institute for Information Technology HIIT, University of Helsinki,
P.O.Box 68, FIN-00014 Helsinki, FINLAND, {teemu.roos|petri.myllymaki}@hiit.fi

ABSTRACT

Parsimonious Markov models, a generalization of variable order Markov models, have been recently introduced for modeling biological sequences. Up to now, they have been learned by Bayesian approaches. However, there is not always sufficient prior knowledge available and a fully uninformative prior is difficult to define. In order to avoid cumbersome cross validation procedures for obtaining the optimal prior choice, we here adapt scoring criteria for Bayesian networks that approximate the Normalized Maximum Likelihood (NML) to parsimonious Markov models. We empirically compare their performance with the Bayesian approach by classifying splice sites, an important problem from computational biology.

1. INTRODUCTION

Classifying discrete sequences is an omnipresent task in computational biology, where an additional challenge is limited data. Recently, parsimonious Markov models [1], a generalization of variable order Markov models [2], have been proposed to model complex statistical dependencies among adjacent observations while keeping the parameter space small and thus avoiding overfitting.

Parsimonious Markov models (parsMMs) use parsimonious context trees (PCTs), which differ from traditional context trees [2] in two aspects: (i) a PCT is a balanced tree, i.e. each leaf has the same depth, and (ii) each node represents an arbitrary subset of the alphabet $\mathcal{A}$, with the additional constraint that everywhere in the tree, sibling nodes form together a partition of $\mathcal{A}$. An example PCT, which shows both features, forming a partition of context sequences that can not be represented by a traditional context tree, is shown in Figure 1. A PCT τ of depth d partitions all *context sequences* of length d over alphabet $\mathcal{A}$ into disjoint sets, which are called *context*. We denote all contexts represented by τ as $\mathcal{C}_\tau$. An inhomogeneous parsimonious Markov model of order D for modelling sequences of length L allows using different PCTs at each position in the sequence. The first D positions use PCTs of increasing order $0, \ldots, D-1$, whereas the remaining $L-D$ positions use PCTs of order D. The likelihood function is given by

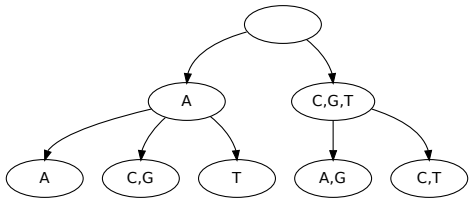

Figure 1. Example PCT of depth 2 over DNA alphabet. It encodes the partitioning of all 16 possible sequences of length 2 into a set of contexts $\mathcal{C}_\tau$ ={{AA},{CA,GA},{TA},{AC,AG,AT,GC,GG,GT}, {CC,CG,CT,TC,TG,TT}}.

$$P(\mathbf{X}|\vec{\Theta}) = \prod_{\ell=1}^{L} \prod_{\mathbf{w}\in\mathcal{C}_{\tau_\ell}} \prod_{a\in\mathcal{A}} (\theta_{\ell\mathbf{w}a}^{\tau_\ell})^{N_{\ell\mathbf{w}a}} . \qquad (1)$$

where $N_{\ell\mathbf{w}a}$ is the number of occurrences of symbol a at position ℓ in all sequences in data set $\mathbf{X}$, whose subsequences from position $\ell - |\mathbf{w}|$ to $\ell - 1$ are an element of context $\mathbf{w}$.

The likelihood is closely related to that of Bayesian networks (BNs), since it factorizes into independent terms for each variable and the number of conditional probability parameters depends on the structure of the model. However, whereas BNs have freedom in choosing the parent nodes of a random variable but always use seperate conditional probability parameters for each possible realization of the parent nodes, parsMMs have fixed parent nodes but freedom in lumping several of their possible realizations together as one context.

There is an efficient dynamic programming (DP) algorithm [3, 1] for finding the PCT that maximizes an arbitrary structure score, which only has to fulfil the property of factorizing into independent leaf scores. In the Bayesian setting, the structure score is usually the local posterior probability of a PCT given data. If the local parameter prior is a symmetric Dirichlet with equivalent sample size (ESS) α, we obtain the BDeu score [4], which

can be used in the DP algorithm since it factorizes along contexts. The conditional probability parameters are estimated by the mean posterior (MP) principle.

In practice, there is rarely reliable a priori knowledge available for specifying α. Since it is known that the choice of α influences the model complexity in the case of Bayesian networks [5], it is safe to assume that a similar effect may be observed for parsimonious Markov models. Often a cross validation (CV) on the training data is used to obtain a reasonable choice for this external parameter. However, CV is a time consuming procedure and there is no guarantee that a useful prior on a subset of the training data will also yield optimal results when learning from the complete training data for classifying previously unseen test instances.

In order to avoid CV, we propose using NML approximating methods for structure and parameter learning, which have been initially proposed for BNs, for parsimonious Markov models. The fNML score [6] has been suggested as score for structure learning of BNs, whereas the corresponding conditional probability parameters have been obtained in the same setting by using fsNML estimates [7]. Due to the structural similarity of the likelihood function of parsMMs and that of BNs, both methods can be adapted without modification.

2. RESULTS

We compare two different scores for the PCT structures, BDeu and fNML, and two different methods for estimating conditional probability parameters of each PCT, MP and fsNML. In order to determine whether structure or parameter learning is dominating the results, we do not only compare MP parameter estimates for a BDeu optimal structure with fsNML parameter estimates for an fNML optimal structure, but also consider the other two possibilities (Table 1).

We perform two seperate case studies. The first study is a standard classification experiment for short symbolic sequences, which uses labeled training data and involves structure and parameter learning for both classes. In computational biology, this an abundant task, when experimentally verified training data is available.

The second study is inspired by the computational problem of de novo motif discovery [8, 9]. Motif discovery usually involves latent variables, hence it cannot be solved exactly, and approximate algorithms, such as the expectation-maximization (EM) algorithm [10] have to be resorted to. Formulating fNML and fsNML in a setting with latent variables, i.e. utilizing weighted data inside the EM algorithm is not straightforward, but a slight modification of the classification problem resembles the task that typically arises in those iterative algorithms. In the modified classification, the structure and parameters of the background class are fixed and there is much more background training data available. Hence the prior in the Bayesian setting only affects the foreground model. This resembles the problem of motif discovery, where only structure and parameters of a motif model (foreground) are to be estimated, whereas the structure and parameters of the background model remain fixed.

Table 1. The two combinations in the major diagonal are the obvious ways of learning parsMMs in the Bayesian and NML setting respectively, whereas the minor diagonal contains rather artificial combinations, which we mainly investigate for academic purposes.

	BDeu	fNML
MP	BDeu-MP	fNML-MP
fsNML	BDeu-fsNML	fNML-fsNML

2.1. Standard classification

In the first experiment, we perform a standard classification on the benchmark data set of Yeo and Burge [11]. It consists of 12,623 experimentally verified splice donor sites (foreground data) and 269,157 non splice sites (background data). Both data sets, consisting of sequences of length 7 over the quarternary DNA alphabet, were already split by Yeo and Burge into training and test data at the ratio of 2:1 [11], and we use the same partitioning.

Since we are interested in situations with limited data, we randomly pick 500 sequences from each of the training data sets for learning foreground and background model, both being second order inhomogeneous parsimonious Markov models. We learn – for each possible combination of scores – structure and parameters of two parsimonious Markov models. For the Bayesian scores, we learn models for a large variety of possible ESS values, ranging from 10^{-5} to 10^{8}. We repeat the procedure 10^{3} times with different training samples.

In Figure 2, we compare the average complexities of the learned models. For the BDeu score, we observe with increasing ESS an increase in model complexity, which is a behaviour that is already known from Bayesian networks [5]. The fNML score has the advantage of not being affected by the ESS at all. However, it yields a comparatively low model complexity for the foreground model, which is surprising since the foreground data set is known to contain strong statistical dependencies. The background model is surprisingly complex, given the fact that the background data shows much less dependencies.

Additional studies have have shown that the difference in model complexity of fNML estimated foreground and background model decreases when both samples sizes are reduced. The BDeu score, however, retains a certain difference in model complexity, even when sample sizes are very small.

However, the PCT structure itself is not sufficient to compare scoring criteria, since we are mainly interested in the classification performance of the learned models. In order to evaluate the classification performance of a set of PCTs, we estimate conditional probability parameters, build a likelihood ratio classifier, compute probabilities for each sequence in both test data sets and compute

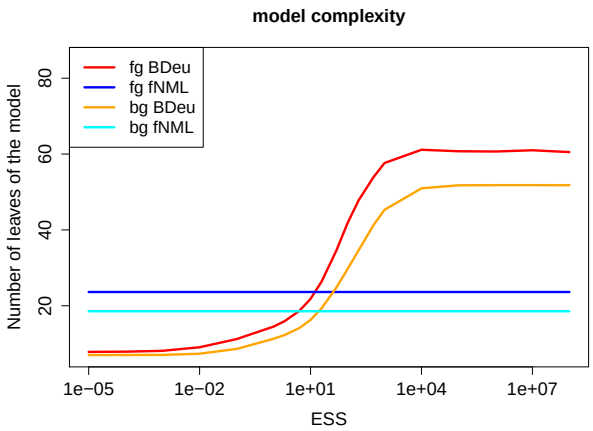

Figure 2. Averaged model complexities (measured as the total number of leaves in the model) for foreground and background model are plotted against the equivalent sample size. Since the fNML criterion does not use the ESS parameter, the model complexities is constant. Standard errors are 0.1 at most, hence error bars are omitted from the plot.

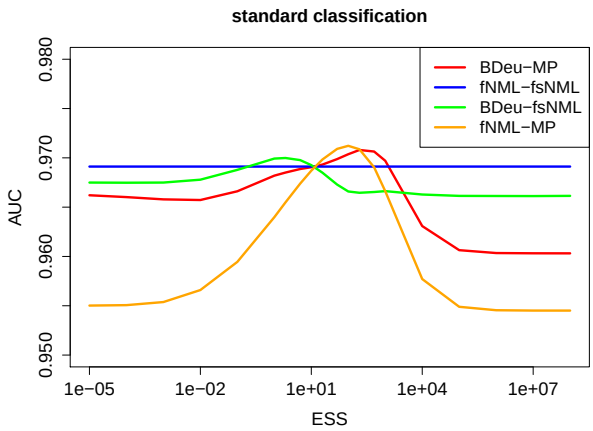

Figure 3. Averaged AUC values for the standard classification experiment plotted against the equivalent sample size. In the BDeu-MP setting, the same ESS is used for structure and parameter learning. For BDeu-fsNML, the ESS only affects structure learning, whereas for fNML-MP is only affects parameter learning. Standard errors are 10^{-4} at most, hence error bars are omitted.

the area under the ROC curve (AUC) [12]. When combining the Bayesian structure and parameter learning, we apply the same prior to both problems.

For each of the four possible score combinations, we repeat the entire study with 10^3 different training samples and average the resulting AUC values. The results are shown in Figure 3. We observe an AUC of 0.9691 for the fNML-fsNML method. For an ESS ranging from 10^1 to 10^3, the Bayesian approach outperforms fNML-fsNML method, obtaining a maximal AUC of 0.9708 for an ESS of 200. Interestingly, an ESS of 1, which is often considered to be the most uninformative choice, is obviously not optimal, since performs significantly worse than larger ESS values and even slightly worse than the NML approach.

The mixed approach of combining fNML structure learning with MP parameter estimates also yields a good classification, if the ESS is chosen correctly. For ESS values between 10 and 500, it outperforms the pure NML method, and its absolute maximum with an AUC of 0.9712 at ESS of 100 even outperforms the pure Bayesian method, even though the difference is quite small.

The BDeu-fsNML method does not show strong over- or underfitting, but it is even with perfectly chosen ESS only slightly better than the pure NML method. In general, the parameter learning seems to dominate the experiment, since the methods using the same parameter estimate resemble each other more than the methods using the same structure score.

2.2. Fixed background

In the second experiment, we consider a different setting. Now fix the background model to a simple independence model and estimate its parameters once from the entire background training data set according to the maximum likelihood (ML) principle. Since the complete background data contains over 10^5 data points, the ML estimator is basically identical to fsNML and MP estimates. The repeated holdout experiment as described in the previous section is only carried out for the foreground model. This situation resembles the problem de novo motif discovery [8, 9], where there is orders of magnitude more data available for learning the parameters of the background compared to the foreground, and where learning the background model does not contain a model selection step.

The results of this modified classification are shown in Figure 4. We observe the fNML-fsNML approach in comparison with the BDeu-MP approach to be almost optimal. There is only a tiny improvement that the Bayesian approach may achieve if the ESS would have been chosen perfectly at a value of approximately 20. Interestingly, both mixed approaches perform better than the pure Bayesian approach, since the range of good ESS values and the maximal improvement in AUC are increased.

Both methods using the MP parameter estimates break down if the ESS is larger than 100, which might be explained as follows. If the foreground parameters are computed by using a large ESS, resulting large pseudocounts, they get concentrated around the uniform distribution. This is not a problem as long as the same applies to the background parameters, since even small differences between foreground and background parameters are sufficient to classify a test sequence correctly. However, if the background parameters are fixed to certain values, only smoothing the foreground parameters creates an imbalance which prevents a fair comparison of foreground and background likelihood for a test sequence, resulting in

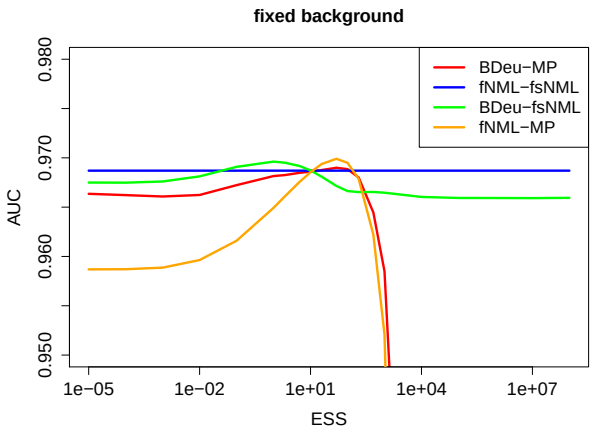

Figure 4. Averaged AUC values for the classification experiment with fixed background model. The standard errors are below 10^{-5}, hence error bars are omitted.

many classification errors. This situation however, typically occurs in the problem of de novo motif discovery, where a motif model is estimated from small data samples, and where and the background model, it is compared with, has fixed parameters that may have been estimated from a much larger amount of data.

3. CONCLUSIONS

We have compared NML with Bayesian criteria for structure and parameter learning of parsimonious Markov models with application to the classification of DNA sequences. In a standard classification, we found the Bayesian approach to perform well, outperforming the NML approach for a comparatively large range of ESS values. We also found the optimal ESS parameter for classification purposes to be larger than 1, which is often an intuitive choice, but smaller than 500. In a classification with fixed background model structure and parameters, we found the NML approach to be as good as the optimal Bayesian approach. The latter does not yield a significant improvement in AUC, even if the optimal value of the ESS would have been guessed. Moreover, we find the Bayesian approach in this setting to be very sensitive towards very large ESS values. This makes it tempting to speculate that the NML learning approach might be also of use in the problem of de novo motif discovery, which includes a classification step with fixed background parameters.

4. ACKNOWLEDGMENTS

This work was funded by *Reisestipendium des allg. Stiftungsfonds der MLU Halle-Wittenberg* and the Academy of Finland (PRIME and MODEST).

5. REFERENCES

[1] Pierre-Yves Bourguignon, *Parcimonie dans les modèles markoviens et applications à l'analyse des séquences biologiques*, Ph.D. thesis, Université Evry Val d'Essonne, 2008.

[2] Jorma Rissanen, "A universal data compression system," *IEEE Trans. Inform. Theory*, vol. 29, no. 5, pp. 656–664, 1983.

[3] P. Bühlmann and A.J. Wyner, "Variable length Markov chains," *Annals of Statistics*, vol. 27, pp. 480–513, 1999.

[4] G. Heckerman, D. Geiger, and D. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data," *Machine Learning*, vol. 20, pp. 197–243, 1995.

[5] T. Silander, P. Kontkanen, and P. Myllymäki, "On Sensitivity of the MAP Bayesian Network Structure to the Equivalent Sample Size Parameter," in *Proceedings of the The 23rd Conference on Uncertainty in Artificial Intelligence (UAI-2007)*, 2007, pp. 360–367.

[6] T. Silander, T. Roos, P. Kontkanen, and P. Myllymäki, "Factorized NML Criterion for Learning Bayesian Network Structures," in *Proceedings of the 4th European Workshop on Probabilistic Graphical Models (PGM-08)*, 2008.

[7] T. Silander, T. Roos, and P. Myllymäki, "Locally Minimax Optimal Predictive Modeling with Bayesian Networks," in *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2009, pp. 504–511.

[8] C.E. Lawrence and A.A. Reilly, "An Expectation Maximization Algorithm for the Identification and Characterization of Common Sites in Unaligned Biopolymer Sequences.," *Proteins: Structure, Function and Genetics*, vol. 7, pp. 41–51, 1990.

[9] T.L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 1994, pp. 28–36.

[10] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[11] G. Yeo and C.B. Burge, "Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals," *Journal of Computational Biology*, vol. 11(2/3), pp. 377–394, 2004.

[12] Kent A. Spackman, "Signal detection theory: Valuable tools for evaluating inductive learning," in *Proceedings of the Sixth International Workshop on Machine Learning*, San Mateo, CA, 1989, pp. 160–163.