Poster H-77

Biological signal prediction using Stochastic Regular Grammar

Authors:
Andre Kashiwabara (*IME-USP*)
Alan Durham (*IME-USP*)

Short Abstract: Two technologies have been widely applied for ab initio splice site prediction, WMM and WAMs. We show that WMMs and WAMs are part of the search space of inference algoritms for Sochastic Regular Grammars and present an algorithm that converge to any of them.

Long Abstract:

Biological signal prediction by computational methods is an important step for an accurate ab initio gene prediction. In particular, splice site prediction is a difficult and relevant problem. Weight Matrix Method (WMM)[11] and Inhomogeneous Markov Model (IHMM or WAM) [16] are two signal sensors that were incorporated in ab initio gene prediction, and are widely used in splice site prediction methods [7, 12, 8, 1].

In the last two decades, various pattern recognition methods were applied in splice site prediction and some of them uses WMM or IHMM, such as decision trees, and Support Vector Machines (SVM) [13]. Maximum Dependence Decomposition [2] is a decision tree approach that makes a partition of the dataset such that the strongest dependencies are captured at the earliest branch points, and WMM or IHMM can be used to represent each subset of the tree. SVM has been employed recently, and it also have applied WMM for feature extraction [5].

Probabilistic Finite Automata (PFA) is a syntactic object which can model and generate the same probability distribution as Hidden Markov Models (HMM) [6] over sets of strings [14]. Deterministic PFA (DPFA) are not as powerful as HMMs, but some problems become tractable: (1) DPFA has a simple and efficient recognizer (called parser) with complexity O(n), where n is the size of the word. (2) DPFA equivalence problem is tractable, because it admits a minimal object. In a previous work, we have shown that a DPFA has a similar performance than NNSPLICE[9] for donor site prediction [15].

In our current work, we show that WMM and IHMM are particular cases of Deterministic Probabilistic Finite Automata (DPFA), and that both signal sensors are in the search space of DPFA inference algorithms that employ prefix tree automata: Learn Acyclic PFA (LAPFA)[10] and ALERGIA[3]. These algorithms work over a search space that can be viewed as a lattice in which the elements are the automatas. The canonical automata of this lattice is the prefix tree automata that recognizes only the training set, and the universal automata has only one state with recursive transitions for every entry symbol, recognizing any string over the alphabet. The DPFA interence algorithms mentioned above modify the prefix tree automata interactively by merging different states. This creates a search space of automata that can be reached from the prefix tree automata by successive state merging operations [4]. The idea

is to search an automata that is a good approximation of a target probabilistic automata.

In spite of IHMMs being in the theoretical search space of the LAPFA algorithm, experiments with splice site prediction have shown us that even the best parameter setting of the algorithm was converging to a DPFA similar to a WMM, in spite of the fact that IHMM presented better prediction performance. We have thus developed a modification of LAPFA that enables it to also converge also to IHMMs, if the training sample supports this convergence.

DPFA could be a good alternative to IHMM and WMM, but more investigation must to be done, in particular trying to find better convergence points in the specific case of splice sites.

In this work we present the proof that IHMM and WMM have equivalent DPFAs, the proof that the DPFAs are in the theoretical search space of prefix-tree-based DPFA inference algorithms, and an extension of the LAPFA inference algorithm, that ensures it can also converge IHMM-equivalent DPFAs. Finally, we show the comparative results of the new DPFA inference algorithm on splice site prediction against WMMs and IHMMs. Future work includes a study of other modifications on the LAPFA algorithm aiming at improving the performance of DPFAs on splice site prediction in particular and signal prediction in general.

References

[1] C. Burge. Identification of genes in human genomic DNA. PhD thesis, Stanford University, 1997.
[2] C. Burge. Modeling dependencies in pre-mRNA splicing signals. Computational Methods in Molecular Biology, 32:129164, 1998.
[3] R. C. Carrasco and J. Oncina. Learning stochastic regular grammars by means of a state merging method. In International Conference on Grammatical Inference, pages 999999. Springer-Verlag, September 1994.
[4] Dupont, Miclet, and Vidal. What is the search space of the regular inference? In ICGI: International Colloquium on Grammatical Inference and Applications, 1994.
[5] Huang, Li, Chen, and Wu. An approach of encoding for prediction of splice sites using svm. Biochimie, Apr 2006. [6] A. Krogh. An introduction to hidden markov models for biological sequences. Computational Methods in Molecular Biology, 32:45--63, 1998.
[7] W. H. Majoros, M. Pertea, and S. L. Salzberg. Tigrscan and glimmerhmm: two open source ab initio eukaryotic gene-finders. Bioinformatics, 20:28782879, Nov 2004.
[8] M. Pertea, X. Lin, and S. L. Salzberg. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res, 29:11851190, Mar 2001.
[9] M. G. Reese, F. H. Eeckman, D. Kulpa, and D. Haussler. Improved splice site detection in Genie. J Comp Biol, 4:311323., 1997.
[10] Ron, Singer, and Tishby. On the learnability and usage of acyclic probabilistic finite automata. JCSS: Journal of Computer and System Sciences, 56, 1998.
[11] R. Staden. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids research, 12:505519, 1984.
[12] Mario Stanke and Stephan Waack. Gene prediction with a hidden markov model and a new intron submodel. Bioinformatics, 19 Suppl 2:II215II215, Oct 2003.
[13] V. Vapnik. Statistical Learning Theory. Wiley, 1998.
[14] E.Vidal, F. Thollard, C. de la Higuera, Casacuberta, and R. C. Carrasco.

Probabilistic finite-state machinespart I. IEEETPAMI: IEEE
Transactions on Pattern Analysis and Machine Intelligence, 27, 2005.
[15] D. da Cruz Vieira, A. Y. Kashiwabara, A. M. Lima, and A. M. Durham. Splice site prediction using stochastic regular grammars. International Conference of Bioinformatics and Computational Biology, 2004.
[16] MQ Zhang and TG Marr. A weight array method for splicing signal analysis. Computer Applied in Bioscience, 9:499--509, 1993.