Proposal for a Tutorial on Minimal Length Encoding (MLE) in Molecular Biology

Aleksandar Milosavljević
Genome Structure Group
Center for Mechanistic Biology and Biotechnology
Argonne National Laboratory
Argonne, Illinois 60439-4833
e-mail: milosav@anl.gov
tel: (708) 252-7860
fax: (708) 252-3387

March 1994

1 Purpose and intended audience

The main purpose of this tutorial is to introduce the Minimal length encoding (MLE) method to computational biologists who are designing sequence analysis algorithms, to computer scientists who are interested in learning more about macromolecular sequence analysis, and to biologists who are more advanced users of the sequence analysis programs. The first two groups will be fully prepared to grasp all the technical details of the presentation. A small part of the tutorial may not be fully accessible to all the members of the third group, but care will be taken to explain general ideas in less technical terms and to provide illustrative examples. An emphasis of the workshop will be on the use of the MLE method as a tool for comparative analysis of inference programs in computational biology, with an ultimate purpose of providing more methodological coherence to the emerging field of computational biology.

2 Background

The most widely accepted general formal model of inductive inference is based on the assumption that inference is a process of compression of observations. According to this model, a newly discovered theory is preferred over the old ones if it leads to a more concise decription of experimental observations. A theory is represented by a computer program of minimal length that outputs the observations. The process of inference is viewed as the search for the shortest programs.

A whole new field of algorithmic information theory, which deals with minimal length encoding by computer programs, has emerged (the most recent books on the topic include [18, 11, 25, 8]). The concepts of randomness, complexity, structure, specificity, and information have all found their precise definitions in this theory. MLE has been explicitly applied to the problems of image recognition, categorization, supervised learning of decision trees, decision lists, inductive logic programming, grammatical inference, and in many other domains. A sample of the many applications was presented at the 1990 Stanford Spring Symposium on the Theory and Applications of Minimal Length Encoding. The Symposium included a joint session with the AI and Molecular Biology Symposium where several applications of MLE in molecular biology were presented [21, 9, 6, 1].

Biologists applying the parsimony principle (Occam's Razor) for evolutionary reconstructions and the minimal edit distance criterion for macromolecular sequence alignment come closest to using MLE explicitly. Most often, the principle is applied implicitly, e.g., through preference for causal explanations that postulate fewer causes, or for models that have fewer variables. More recently, the sequence alignment [4, 2, 32], sequence categorization [20], evolutionary reconstructions [3, 24, 14], and a variety of DNA and protein pattern discovery problems [23, 19, 28, 16] have all been approached by explicitly applying MLE. The main advantage of an explicit application of MLE is that the inductive assumptions are stated explicitly in terms of a language for encoding the observations and can be modified to suit the application at hand.

The tutorial will have three goals: (1) outline the basic results of classical Shannon information theory that provide the necessary background, and then cover the basic results from algorithmic information theory that are relevant for the application of MLE; (2) provide an overview of the successful applications of MLE in molecular biology; and (3) compare different approaches to the same inference problem by reducing the approaches to their MLE equivalents.

An emphasis of the tutorial will be on demonstrating the potential of MLE to serve as a tool for comparing different approaches to the same inference problem (goal 3). Three standard inference problems will be chosen. For each problem, an MLE method will be presented as well as the methods that are not explicitly stated in terms of MLE. The competing methods will then be compared by rephrasing them in terms of MLE. The assumptions that are hidden in different approaches to particular inference problems will be made explicit by defining languages used for encoding the observations. The assumptions that underlie different approaches will then be compared.

The ultimate goal of the tutorial is not to present the MLE-based methods as yet another class of inference programs, but to demonstrate that the generality of MLE provides a common framework for comparison of the multitude of apparently different programs that are used for similar inference tasks in molecular biology.

3 Plan

The tutorial will consist of four parts. The first part will introduce the basics of Shannon information theory and of algorithmic information theory. The remaining three parts will focus on three standard inference problems in molecular biology: discovering simple sequences, sequence comparison, and evolutionary reconstructions.

3.1 Basics of Algorithmic Information Theory

The basic concepts of Shannon information theory will be covered as a necessary background: the definitions of entropy, relative and mutual information, as well as the source coding theorem will be explained at the level of an introductory textbook (e.g., [11]).

The basic concepts of algorithmic information theory [7, 18, 11, 25] will be covered next. The concept of randomness (complexity) of an individual object, as opposed to the concept of the randomness of a set of objects (as defined in Shannon's theory), will be explained.

The concept of universal inference will be introduced by explaining the meaning of the universal coding theorem. The relationship of MLE and the standard statistical method of hypcthesis testing, as well as Bayesian inference, will be discussed. Computational limitations of universal inference and the consequent need for inductive bias will be explained. Algorithmic significance will be introduced as a practical MLE method for pattern recognition.

3.2 Simple sequences

Selected methods for identifying "simple" DNA and protein sequences [31, 10, 26, 27, 29]) will be surveyed, as well as the MLE-based method [22]. Selected non-MLE methods will be defined in terms of MLE and the underlying assumptions will be made explicit by specifying encoding schemes. The validity of the hidden assumptions will be examined to determine the domains of applicability of individual methods. Finally, we apply the concept of algorithmic mutual information to determine the effect of complexity of two compared sequences on the significance of their mutual similarity.

3.3 Sequence similarity

Starting from Altschul's information-theoretic analysis [5], we will study the BLAST scoring scheme. By rephrasing the BLAST score in terms of encoding length, we will demonstrate that the significance of BLAST scores can be established via algorithmic significance and without any sophisticated statistical analysis [15]. We then apply the MLE method to get further insight into the encoding scheme that is implicit in the BLAST scoring scheme: we compute optimal PAM matrices (or sets thereof), and we show that the relationship between the minimal length of a significant matching segment and the evolutionary distance can be explicitly determined. Following a recently proposed

MLE method for sequence alignment [2, 32] we show how insertions and deletions can be combined with PAM matrices while still guaranteeing the algorithmic significance of matches (something not achieved in the standard statistical significance framework). Finally we show that the apparent great difference between global and local alignment methods [30] can be reduced to a very small difference in the encoding schemes that are implicit in them.

3.4 Evolutionary reconstructions

MLE methods for evolutionary reconstructions based on non-aligned [3] as well as aligned [24] sequences will be described. It will be shown that standard weighted parsimony [12, 13] and compatibility [17] methods for evolutionary reconstructions correspond to two opposite variants of a single encoding scheme. The reconstruction of Alu evolution [14] will be used as a case study to demonstrate the advantages that are offered by MLE.

4 Acknowledgements

Thanks to David E. Nadziejka for editorial assistance. This work was supported in part by U.S. Department of Energy, Office of Health and Environmental Research, under Contract W-31-109-Eng-38.

References

[1] L. Allison, C.S. Wallace, and C.N. Yee. Inductive inference over macromolecules. In *Working Notes of the AAAI Spring Symposium on Theory and Applications of Minimal-Length Encoding*, Stanford, 1990.

[2] L. Allison, C.S. Wallace, and C.N. Yee. Finite-state models in the alignment of macromolecules. *Journal of Molecular Evolution*, 35:77–89, 1992.

[3] L. Allison, C.S. Wallace, and C.N. Yee. Minimum message length encoding, evolutionary trees, and multiple alignment. In *Proceedings of the 25th Hawaii International Conference on System Sciences*, 1992.

[4] L. Allison and C.N. Yee. Minimum message length encoding and the comparison of macromolecules. *Bulletin of Mathematical Biology*, 52:431–453, 1990.

[5] S.F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology*, 219:555–565, 1991.

[6] M.S. Babcock, W.K. Olson, and E.P.D. Pednault. The use of the minimum description length principle to segment DNA into structural and functional domains. In *Working Notes of the AAAI Spring Symposium on Theory and Applications of Minimal-Length Encoding*, Stanford, 1990.

[7] G.J. Chaitin. On the length of programs for computing finite binary sequences. *Journal of the Association for Computing Machinery*, 13:547–569, 1966.

[8] G.J. Chaitin. *Algorithmic Information Theory.* Cambridge University Press, 1987.

[9] P. Cheeseman and B. Kanefsky. Evolutionary tree reconstruction. In *Working Notes of the AAAI Spring Symposium on Theory and Applications of Minimal-Length Encoding*, Stanford, 1990.

[10] J.-M. Claverie and D.J. States. Information enhancement methods for large scale sequence analysis. *Computers in Chemistry*, 17:191–201, 1993.

[11] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley, 1991.

[12] J.S. Farris. A successive approximations approach to character weighting. *Systematics and Zoology*, 18:374–385, 1969.

[13] J. Felsenstein. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society*, 16:183–196, 1981.

[14] J. Jurka and A. Milosavljević. Reconstruction and analysis of human Alu genes. *Journal of Molecular Evolution*, 32:105–121, 1991.

[15] S. Karlin and V. Brendel. Chance and statistical significance in protein and DNA sequence analysis. *Science*, 257:39–49, 1992.

[16] A. Konagaya and H. Kondou. Stochastic motif extraction using a genetic algorithm with the MDL principle. In *Proceedings of the 26th Hawaii International Conference on System Sciences*, 1993.

[17] W.J. LeQuesne. A method of selection of characters in numerical taxonomy. *Systematic Zoology*, 18:201, 1969.

[18] M. Li and P.M.B. Vitanyi. *An Introduction to Kolmogorov Complexity and its Applications.* Springer Verlag, 1993.

[19] H. Mamitsuka and K. Yamanishi. Protein alpha-helix region prediction based on stochastic-rule learning. In *Proceedings of the 26th Hawaii International Conference on System Sciences*, 1993.

[20] A. Milosavljević. *Categorization of Macromolecular Sequences by Minimal Length Encoding.* PhD thesis, Computer Science Department, University of California at Santa Cruz, 1990.

[21] A. Milosavljevic, D. Haussler, and J. Jurka. Clustering of macromolecular sequences by minimal length encoding. In *Working Notes of the AAAI Spring Symposium on Theory and Applications of Minimal-Length Encoding*, 1990.

[22] A. Milosavljević and J. Jurka. Discovering simple DNA sequences. *Working Notes of the Workshop on AI Approaches to Classification and Pattern Recognition in Molecular Biology*, 1991.

[23] A. Milosavljević and J. Jurka. Discovering simple DNA sequences by the algorithmic significance method. *Computer Applications in Biosciences*, 9(4):407–411, 1993.

[24] A. Milosavljević and J. Jurka. Discovery by minimal length encoding: A case study in molecular evolution. *Machine Learning Journal, Special Issue on Machine Discovery*, 12(1,2,3):69–87, 1993.

[25] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, 1989.

[26] P. Salamon and A.K. Konopka. A maximum entropy principle for distribution of local complexity in naturally occurring nucleotide sequences. *Computers in Chemistry*, 16:117–124, 1992.

[27] P. Salamon, J.C. Wootton, A.K. Konopka, and Hansen L. On the robustness of maximum entropy relationships for complexity distributions of nucleotide sequences. *Computers in Chemistry*, 17:135–148, 1993.

[28] S. Shimozono, A. Shinohara, T. Shinohara, S. Miyano, S. Kuhara, and S. Arikawa. Finding alphabet indexing for decision trees over regular patterns: an approach to bioinformatical knowledge acquisition. In *Proceedings of the 26th Hawaii International Conference on System Sciences*, 1993.

[29] E.N. Trifonov. Making sense of the human genome. In R.H. Sarma and M.H. Sarma, editors, *Human Genome Initiative and DNA Recombination*, volume 1. Adenine Press, 1990.

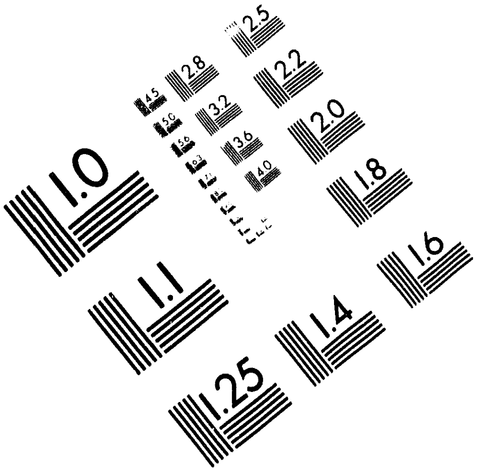
[30] M.S. Waterman. Sequence alignments. In *Mathematical Methods for DNA Sequences.* Boca Raton, Florida, 1989.

[31] J.C. Wootton and S. Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Computers in Chemistry*, 17:149–163, 1993.

[32] C.N. Yee and L. Allison. Reconstruction of strings past. *Computer Applications in Biosciences*, 9(1):1–7, 1993.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

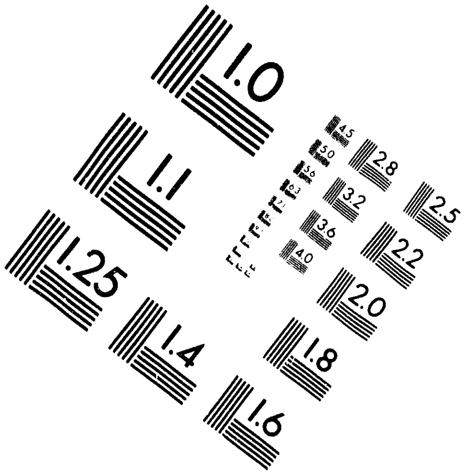


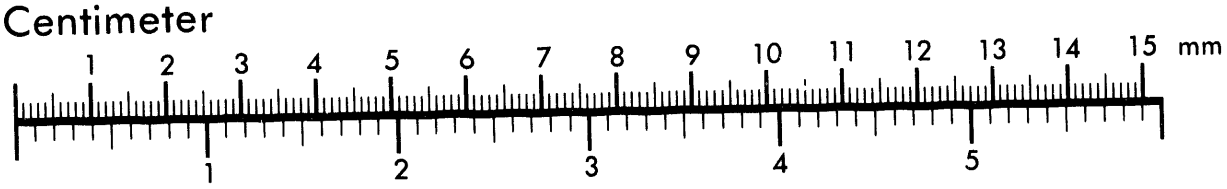
AIIM

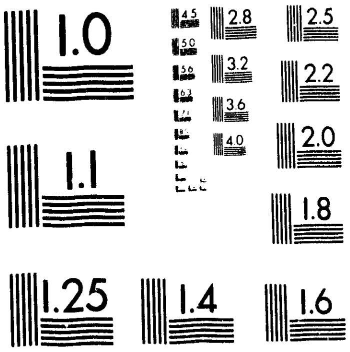
Association for Information and Image Management

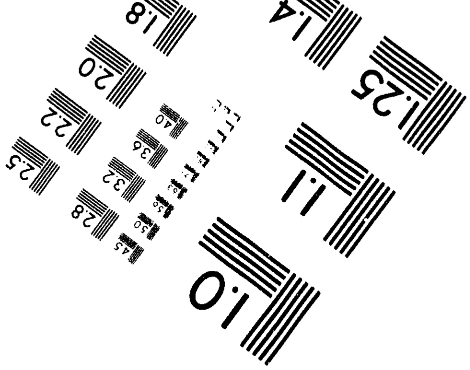
1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

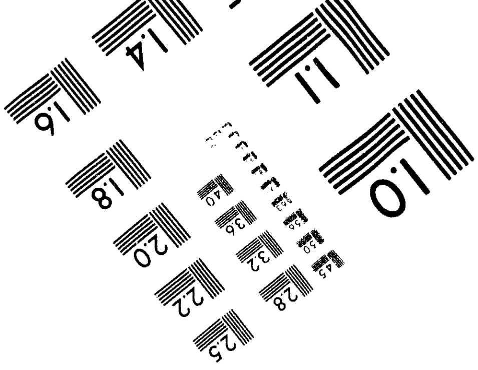
301/587-8202











MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

DATE
FILMED
12/15/94

END