Poster G-19

Identification of Thermostability in Microorganisms using Compression Algorithms

Authors:
Van der Linden, M. G. (*Laboratório Nacional de Computação Científica*)
Batista, L. V. (*Universidade Federal da Paraíba*)
Farias, S. T. (*Universidade Federal de Minas Gerais*)
Marques, J. R. T. (*Universidade Federal da Paraíba*)

Short Abstract: Popular data compression algorithms were applied to the problem of classifying microbial proteomes according to their optimal growth temperatures. The implications of the results are discussed for the problem of understanding patterns of protein sequences that relate to resistance of microbial proteins to extreme temperatures.

Long Abstract:

Microorganisms can be classified according to their optimal growth temperature (OGT) into psychrophiles (OGT < 20°C), mesophiles (20°C < OGT < 45°C), thermophiles (45°C < OGT < 80°C) and hyperthermophiles (80°C < OGT) (Madigan et al., 2003). Macromolecules must be stable and functional within the temperature range in which each organism lives. The thermal resistance of some proteins is noteworthy not only for the scientific curiosity they inspire, but also because of the biotechnological interest they may feature.

The search for patterns in protein sequences that could help explain the resistance of some microorganisms to extreme temperatures has been an active topic of research in molecular biology. Comparisons between selected proteins from organisms that live in extreme temperatures to proteins from organisms that grow in moderate temperatures has been a common methodology in studies that aim to understand how proteins can maintain their properties in extreme environments.

Data classification or categorization is the automatic process of attributing a given piece of data X to one class Ci, i=1,2,...N. Supervised learning methods involve the use of training or learning sets, i.e. groups of selected elements whose correspondent classes are already known. The learning process consists in using distinctive attributes or structural/stochastic models to characterize each class according to the corresponding training set. Categorization is performed on basis of similarity between X and the models or attributes discovered in the learning phase

Lossless data compression is the process of re-encoding a message composed of a sequence of symbols in a more efficient way. It involves two tasks:

(1) The generation of models that describe the source data. Simple models might involve only a computation of the probability of occurrence of each symbol. More elaborated algorithms generate models that also take into account the context in which symbols appear in the source data.

(2) The encoding of the data based on the model. In general, better models correspond to better compression rates.

The ability of lossless data compression algorithms to generate and apply data models can be employed for data classification with supervised learning. This method has already been successfully applied to categorization of images, natural-language texts and music pieces,

often achieving rates of success higher than those of neural networks and Bayesian classifiers (Batista et al., 2004; Begleiter et al., 2004; Coutinho et al., 2005; Frank et al., 2000).).

The Lempel-Ziv (LZ) family of algorithms (Ziv & Lempel, 1978) are used in many of the most popular free and commercial compression applications. The Prediction by Partial Matching (PPM) algorithm (Cleary and Witten ,1984) is regarded as the state of the art algorithm for lossless data compression. A Lempel-Ziv-based algorithm, however, has been demonstrated to be superior to PPM for a number of protein classification problems (Begleiter et al., 2004) and for this reason was also included in this study.

The applicability of the Lempel-Ziv and PPM algorithms was tested in this work for the problem of classifying microbial proteomes according to their OGTs. In the learning stage, samples from complete microbial proteomes were used to generate data models (LZ dictionaries or PPM tables of conditional probabilities) for the four OGT classes (psychrophiles, mesophiles, thermophiles and hyperthermophiles). In the classification stage, the models were used in static mode, i.e., they were not updated during compression. A distinct set of organisms was used to test the classification. For each one of the two data compression algorithms used, each organism proteome was compressed four times, one for each data model, and categorized in the class whose model provided the best compression. The results were compared for both algorithms and for different configurations of each algorithm.

Data models generated by compression algorithms describe patterns in the sequences of symbols (in this case, amino acids) that were used as data sources. Current knowledge of primary structure patterns for thermostability is generally limited to statistical analysis of amino acid frequency (Das & Gerstein, 2000; Chakravarty & Varadarajan, 2002; Farias & Bonato, 2003). Therefore, besides the applicability in categorization, the study of the models themselves may also have important implications for understanding protein thermostability.

References

Batista L. V. ; Meira M. M. Texture Classification Using the Lempel-Ziv-Welch Algorithm. Lecture Notes in Computer Science, Berlin (Qualis 2004 Int. A), v. 3171, p. 444-453, 2004

Begleiter R., El-Yaniv R., Yona G. (2004) On Prediction Using Variable Order Markov Models. Journal of Artificial Intelligence Research 22 385-421.

Chakravarty S., Varadarajan R. (2002). Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry. 41(25):8152-61.

Cleary J., Witten I. (1984). Data compression using adaptive coding and partial string matching. IEEE Transactions on Communications, COM-32(4), 396-402.

Coutinho B., Macedo J., Rique A., Batista L.V. (2005) Atribuição de Autoria usando PPM. In: III Workshop em Tecnologia da Informação e da Linguagem Humana, 2005, São Leopoldo. Anais do XXV Congresso da Sociedade Brasileira de Computação, 2005. v. 1. p. 2208-2217.

Das R., Gerstein M. (2000). The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics. 1(1):76-88.

Farias S.T., Bonato M.C.M. (2003) Preferred amino acids and thermostability. Genet. Mol. Res., 2, 383-393.

Frank E., Chi C., Witten I.H. (2000) Text Categorization Using Compresssion Models. Proceedings of the Data Compression Conference, Salt Lake City, pp. 500

Madigan M.T., Martinko J.M., Parker J. (2003). Brock Biology of Microorganisms, 10th edn.

Pearson Education Inc., NJ, USA, pp. 151-156.
Moffat A. (1990). Implementing the PPM data compression scheme. IEEE Transactions on Communications, 38 (11), 1917-1921.
Ziv J., Lempel, A. (1978). Compression of individual sequences via variable-rate coding. IEEE Transactions on Information Theory, 24, 530-536.