Bayesian Network Classification of Binarized DNA Fingerprinting Patterns

Mats Gyllenberg[1], Jonas Carlsson[2] and Timo Koski[2]

[1] Department of Mathematical Sciences, University of Turku, Finland
[2] Department of Mathematics, University of Linköping, Sweden

Abstract. A Bayesian network [9] based on tree dependent probability distributions for multivariate binary data is applied to clustering of DNA fingerprints of isolates belonging to the genus *Vibrio*. We present an algorithm for unsupervised classification that minimizes stochastic complexity. The same isolates have been previously analyzed in [16] using clustering techniques of numerical taxonomy. The trees are established from data by maximizing mutual information between pairs of nodes using a variant of the Kruskal algorithm that may create even disconnected forests.

1 Introduction

Classifications (taxonomies) of bacteria have been produced by many different techniques [12]. Over the past decade the advancement of molecular biology has led to the development of bacterial taxonomies based on detection of the naturally occurring DNA polymorphisms [8]. These polymorphisms are a result of point mutations or rearrangements in the DNA.

A restriction endonuclease recognizes a specific sequence of nucleotide pairs and cleaves it. The number and locations of restriction sites vary with nucleotide sequence. The higher the similarity of the two DNA sequences compared, the closer the cleavage pattern.

The DNA fingerprinting technique known as AFLP (=Amplified Fragment Length Polymorphism) [1] is based on the selective amplification of genomic restriction fragments by PCR (=Polymerase Chain Reaction) to differentiate bacterial strains at the subgeneric level and consists of three steps: (a) Digestion of total cellular DNA with two restriction enzymes and ligation of restriction halfsite-specific adaptors to al restriction fragments. (b) Selective amplification of these fragments with two PCR primers that have corresponding adaptor- and restriction-site-sequences as their target sites. (c) Electrophoretic separation of the PCR products on a gel.

In electrophoresis, each organism is characterized by a banding pattern, which are thus a direct reflection of the genetic relationship between the bacterial strains examined and therefore these banding patterns can be considered as genomic fingerprints allowing computational algorithms for characterization (typing) and identification purposes. A collection of fingerprints can be translated into a matrix of binary variables, band present (1) or band absent (0), c.f. [10]. In this note we show how a classification of bacteria (suggesting an additional insight into the taxonomy in [16]) is learned using the indicated binary data matrix.

A classification is in [6] viewed as a model of the data and stochastic complexity is computed using statistical representation of the classification. The statistical representation consists of class-conditional probability distributions factorized along a (rooted) tree, whose nodes are the components of a binary vector. The classifiers are trained by minimization of stochastic complexity (SC) [13].

The notion of a classifier augmented by a dependence tree is due to Chow and Liu [2] and has been further developed in [4,7] using the results in [15]. Applied to microbiology this extends and improves probabilistic numerical methods of bacterial taxonomy [6,11]. In terms of machine learning we are dealing with a technique of learning a structure of a Bayesian (tree) network from data [3,9].

2 Likelihood for a dependence tree

Let x denote a vector with d binary components, i.e., $x = (x_i)_{i=1}^{d}$, $x_i \in \{0,1\}$. Let a tree with d nodes topologically ordered be given by $\Pi = (\Pi[1], \Pi[2], \ldots, \Pi[d])$, where $\Pi[j]$ is the parent of the node j, and $\Pi[1] = \emptyset$. The conditional probabilities $P\left(x_i | x_{\Pi(i)}\right)$ are written as

$$P\left(x_i | x_{\Pi(i)}\right) = \left(\theta_i^{x_i}\left(1-\theta_i\right)^{(1-x_i)}\right)^{x_{\Pi(i)}} \cdot \left(\phi_i^{x_i}\left(1-\phi_i\right)^{(1-x_i)}\right)^{1-x_{\Pi(i)}}, \quad (1)$$

where $\theta_i = P\left(x_i = 1 | x_{\Pi(i)} = 1\right)$ and $\phi_i = P\left(x_i = 1 | x_{\Pi(i)} = 0\right)$. The joint distribution factorized along Π is

$$P_{\underline{\theta},\underline{\phi}}(x) = P(x_1) \cdot P\left(x_2 | x_{\Pi(2)}\right) \cdots P\left(x_{d-1} | x_{\Pi(d-1)}\right) \cdot P\left(x_d | x_{\Pi(d)}\right). \quad (2)$$

Let next $X^t = \left\{x^{(l)}\right\}_{l=1}^{t}$ be a set of binarized band patterns, which are seen as independent instantiations of variables with a dependence tree. The likelihood for the tree topology Π is obtained by integrating the joint distribution with respect to a prior density $g\left(\underline{\theta}, \underline{\phi}\right)$, i.e.,

$$P\left(X^t | \Pi\right) = \int_{\Theta} \int_{\Phi} \prod_{l=1}^{t} P_{\underline{\theta},\underline{\phi}}\left(x^{(l)}\right) g\left(\underline{\theta}, \underline{\phi}\right) d\underline{\theta} d\underline{\phi}. \quad (3)$$

Here $\Theta = \{\underline{\theta} = (\theta_i)_{i=1}^{d}; 0 \leq \theta_i \leq 1\}$ and Φ is analogous. Let $\widehat{P}_{i,\Pi(i)}(u,v) = \frac{n_i(u,v)}{t}$, $u, v \in \{0,1\}$ be the relative frequency of a pair of bits in positions $i, \Pi(i)$ in X^t and $\widehat{P}_i(u)$ and $n_i(u)$ be margins and let

$$I_{i,\Pi(i)} = \sum_{u=0}^{1} \sum_{v=0}^{1} \widehat{P}_{i,\Pi(i)}(u,v) \log_2 \frac{\widehat{P}_{i,\Pi(i)}(u,v)}{\widehat{P}_i(u) \cdot \widehat{P}_{\Pi(i)}(v)}. \quad (4)$$

For local meta independence of parameters and Jeffreys' prior quoted below we refer to [3] and [13], respectively.

Theorem 1. *If* (3) *is evaluated assuming both local meta independence of parameters and Jeffreys' prior for each of the parameters, then*

$$-\log_2 P\left(X^t|\Pi\right) = t\cdot\sum_{i=2}^{d} h\left(n_i(1)/t\right) - t\cdot\sum_{i=2}^{d} I_{i,\Pi(i)} + \frac{1}{2}\sum_{i=2}^{d}\log_2\left(n_{\Pi(i)}(1)\right) + \log_2\left(n_{\Pi(i)}(0)\right) + C,$$

where $h(x)$ *is the binary entropy function and* C *contains terms of order* $o(t)$.

A proof of the theorem can be found in [7].

3 Classification by minimizing SC

By a classification of a set of binary vectors X^t we mean the subdivision of X^t into k disjoint classes $c_j \subseteq X^t$, $j = 1,\dots,k$. The number of classes, k, is *not* given in advance but is also learned from data [5]. We define for any given classification the class membership functions $u_j^{(l)} := 1$ if $x^{(l)} \in c_j$ and 0 otherwise, and incorporate these in the matrix $U^t = \left\{u_j^{(l)}\right\}_{l=1,j=1}^{t,k}$. The prevalence of class is $\lambda_j = P\left(u_j^{(l)} = 1\right)$ and $V = \{\underline{\lambda} = (\lambda_j)_{j=1}^{k}; \sum \lambda_j = 1, 0 \leq \lambda_j \leq 1\}$. Putting $t_j = \sum_{l=1}^{t} u_j^{(l)}$ we set $Q^*\left(U^t\right) = \int_V \prod_{j=1}^{k} \lambda_j^{t_j}\psi\left(\underline{\lambda}\right)d\underline{\lambda}$, where $\psi\left(\underline{\lambda}\right)$ is Jeffreys' prior. For each j let $\Theta_j = \{\underline{\theta}_j = (\theta_{ij})_{i=1}^{d}\}$ (ditto for Φ_j), and let $P_{\underline{\theta}_j,\underline{\phi}_j}(x)$ be defined as corresponding to (2), and

$$P^*\left(X^t|U^t,\Pi\right) = \prod_{j=1}^{k}\int_{\Theta_j}\int_{\Phi_j}\prod_{l=1}^{t}\left[P_{\underline{\theta}_j,\underline{\phi}_j}\left(x^{(l)}\right)\right]^{u_j^{(l)}} g\left(\underline{\theta}_j,\underline{\phi}_j\right)d\underline{\theta}_j d\underline{\phi}_j.$$

It follows readily that $\log_2 P^*\left(X^t|U^t,\Pi\right) = -\sum_{j=1}^{k}\log_2 P\left(X^t|\Pi_j\right)$, where Π_j is the tree for class c_j. Each term $\log_2 P\left(X^t|\Pi_j\right)$ can be expanded as in the theorem above. The *stochastic complexity* per sample is defined as

$$\mathrm{SC} = -\frac{1}{t}\log_2 P^*\left(X^t|U^t,\Pi\right) + \left(-\frac{1}{t}\log_2 Q^*\left(U^t\right)\right). \tag{5}$$

SC is thus the sum of the length of the description of the classes with respect to the classification model chosen, and the length of the description of the items within the cluster. We minimize SC by the following alternating algorithm based on the theorem above:

1. Fix k, the number of classes, set $w = 0$ and store an arbitrary (random) matrix of class memberships $U_{(w)}$.
2. Find the structure $\widehat{\Pi}_{(w)}$ maximizing

$$\sum_{j=1}^{k} \frac{t_j}{t} \sum_{i=2}^{d} I_{i,\Pi_j(i)} - \frac{1}{2} k \cdot (2d) \frac{\log_2 t}{t}$$

using a version of the Kruskal algorithm.
3. For $U_{(w)}$ and $\widehat{\Pi}_{(w)}$ compute the maximum likelihood estimates $\widehat{\underline{\theta}}_j$, $\widehat{\underline{\phi}}_j$ and $\widehat{\lambda}_j$ for each class c_j.
4. Given $\widehat{\Theta}_{(w)}$, $\widehat{\lambda}_{(w)}$, and $\widehat{\Pi}_{(w)}$ set

$$c_*^{(l)} = \arg \max_{1 \leq j \leq k} P_{\widehat{\underline{\theta}}_j, \widehat{\underline{\phi}}_j} \left(x^{(l)} \mid c_j \right) \widehat{\lambda}_j, \tag{6}$$

and determine $U_{(w+1)} = \left\{ (u_j^{(l)})_{(w+1)} \right\}_{j,l=1}^{t,k}$ using $(u_j^{(l)})_{(w+1)} = 1$ if $c_*^{(l)} = j$ and 0 otherwise.
5. If $U_{(w+1)} = U_{(w)}$, then stop, otherwise set $w = w + 1$ and go to **2.**.

A tree Π_j found in **2.** may very well be disconnected, which is clearly a natural consequence of minimization SC. The assignment in (6) extends the rules of *probabilistic identification* as known in microbiology [11,10]. In order to establish k, the number of classes, this algorithm is used repeatedly for different k.

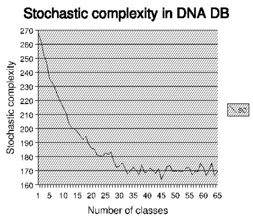

Fig. 1. SC as function of the number of classes

4 A summary of classification of genus *Vibrio* strains

The 505 FAFLP (=Fluorescent Amplified Fragment Length Polymorphism, AFLP without radioactive materials) fingerprints of genomes of *Vibrio* strains listed in [16] were binarized by Dr. Peter Dawyndt (Laboratory of microbiology, University of Ghent). The preceding algorithm was then applied to this set of data with several randomized initial $U_{(w)}$. The minimum value of SC in (5) is 163.8 (binary information units per sample). This is obtained at $k = 45$. The value of SC as a function of k is seen in Figure 1. This improves by 28 % the optimum value of SC achieved by minimizing SC under a statistical model with independent binary features ('naive Bayes model'). Because classification by minimization of SC compresses data, it comes as no surprise that the optimum number of clusters (classes) found by the algorithm turns out to be smaller than the number of clusters (=69) established in [16]. The last mentioned clustering is obtained from a dendrogram, based on Dice's coefficient of similarity and Ward's method of agglomeration [14], cut at a subjectively chosen threshold value.

The classification minimizing SC should in this context be evaluated with respect to its biological relevance. By inspection of the results one finds that 49 of the 69 clusters in [16] have remained intact or have all but one or two of their strains in the same larger class in the classification minimizing SC. However, the two clusters with the type strains *Vibrio penaeicida* and *Vibrio tubiashi*, respectively, have been completely disintegrated in the classification minimizing SC.

Another way to look at the biological meaning of the result is by means of the class memberships of the *type strains*. The clusters in [16] can in broad terms be described as falling in three categories: (a) isolates with genomes related to only one known type strain, (b) isolates clustering to one or more type strains, and (c) isolates with genome unrelated to any type strain (less than 45 % similarity). It is clear that the identification rule (6) does not measure similarity in the same explicit biological sense as Dice's coefficient. Hence it is interesting to note that in the classification minimizing SC those 49 classes that have remained intact have the corresponding type strains in the same cluster in the classification minimizing SC or are clusters that do not have any type strain either in [16] or in the classification minimizing SC. There are, however, exceptions: two large untyped clusters in [16] with isolates from the gut of turbot larvae *Scophthalmus maximus* have been completely joined in a class with the type strains *Listonella pelagia, Vibrio nigripulchritudo*, and *Vibrio navarrensis*. The cluster *Vibrio diabolicus* has been completely torn up in pieces and even the type strain *Vibrio diabolicus* forms a singleton class in the the classification minimizing SC. The biological meaning (or lack of it) of these phenomena is not clear for the moment.

5 Acknowledgement

A more extensive analysis of this data with different binarizations using the methods outlined above will be the topic of joint work with Brian Austin, Fabiano Thompson, Peter Dawyndt and Jean Swings.

References

1. Applied Biosystems (2002): *AFLP™ Microbial Fingerprinting. Protocol.* http://docs.appliedbiosystems.com/search.taf
2. C.K. Chow and C.N. Liu (1968): Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory*, IT-14, pp. 462–467.
3. R.G. Cowell, A.P. Dawid, S.L. Lauritzen and D.J. Spiegelhalter (1999): *Probabilistic Networks and Expert Systems*, Springer Verlag, New York.
4. N. Friedman, D. Geiger and M. Goldszmidt (1997): Bayesian Network Classifiers. *Machine Learning*, 29, pp. 1–36.
5. M. Gyllenberg, T. Koski and M. Verlaan (1997): Classification of Binary Vectors by Stochastic Complexity. *Journal of Multivariate Analysis*, 63, pp. 47–72.
6. M. Gyllenberg and T. Koski (2001): Probabilistic Models for Bacterial Taxonomy. *International Statistical Review*, 69, pp. 249–276.
7. M. Gyllenberg and T. Koski (2002): Tree Augmented Classification of Binary Data Minimizing Stochastic Complexity. *Technical Report LiTH-MAT-R-2002-04*, http://www.mai.liu.se/~tikos/papers.html
8. P. Janssen, R. Coopmans, G. Huys, J. Swings, M. Bleeker, P. Vos, M. Zabeuau, K. Kersters (1996): Evaluation of the DNA fingerprinting method AFLP as a new tool in bacterial taxonomy. *Microbiology*, 142, pp. 1881–1891.
9. F.V. Jensen (2001): *Bayesian Networks and Decision Graphs.* Springer, New York.
10. Y. Kassama, P.J. Rooney, and R. Goodacre (2002): An amplified fragment length polymorphism probabilistic database for the identification of bacterial isolates from urinary tract infections. *Journal of Clinical Microbiology*, 40, pp.2795–2800.
11. J.T. Magee (1998): Taxonomy and Nomenclature of Bacteria. *Topley's and Wilson's Microbiology and Microbial Infection. Vol.2., Systematic Bacteriology.* Edward Arnold, London, pp. 44-64.
12. F. Priest and B. Austin (1993): *Modern Bacterial Taxonomy. Second Edition.* Chapman and Hall, London.
13. J. Rissanen (1997). Stochastic Complexity in Learning. *Journal of Computer and System Sciences*, 55, 89–95.
14. P.H.A. Sneath and R.R. Sokal (1973): *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* W.H. Freeman and Co., San Francisco.
15. J. Suzuki (1999): Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: Basic Properties. *IEICE Transactions on Fundamentals*, vol. E82, No. 10., pp. 2237–2245
16. F.L. Thompson, B. Hoste, K. Vandenmeulebroecke and J. Swings (2001): Genomic Diversity Amongst *Vibrio* Isolates from Different Sources Determined by Fluorescent Amplified Fragment Length Polymorphism. *Systematic and Applied Microbiology*, 24, pp. 520–538.