BIOINFORMATICS

Vol. 14 no. 2 1998
Pages 139–143

Feature selection for genetic sequence classification

Nadia A. Chuzhanova[1,3], Antonia J. Jones[2] and Steve Margetts[2]

[1]Institute of Mathematics, Siberian Branch of Russian Academy of Science, 630090, Novosibirsk, Russia and [2]Department of Computer Science, University of Wales, Cardiff, PO Box 916, Cardiff CF2 3XF, UK

Received on June 16, 1997; accepted on October 7, 1997

Abstract

Motivation: *Most of the existing methods for genetic sequence classification are based on a computer search for homologies in nucleotide or amino acid sequences. The standard sequence alignment programs scale very poorly as the number of sequences increases or the degree of sequence identity is <30%. Some new computationally inexpensive methods based on nucleotide or amino acid compositional analysis have been proposed, but prediction results are still unsatisfactory and depend on the features chosen to represent the sequences.*

Results: *In this paper, a feature selection method based on the Gamma (or near-neighbour) test is proposed. If there is a continuous or smooth map from feature space to the classification target values, the Gamma test gives an estimate for the mean-squared error of the classification, despite the fact that one has no a priori knowledge of the smooth mapping. We can search a large space of possible feature combinations for a combination which gives a smallest estimated mean-squared error using a genetic algorithm. The method was used for feature selection and classification of the large subunits of rRNA according to RDP (Ribosomal Database Project) phylogenetic classes. The sequences were represented by dinucleotide frequency distribution. The nearest-neighbour criterion has been used to estimate the predictive accuracy of the classification based on the selected features. For examples discussed, we found that the classification according to the first nearest neighbour is correct for 80% of the test samples. If we consider the set of the 10 nearest neighbours, then 94% of the test samples are classified correctly.*

Availability: *The principal novel component of this method is the Gamma test and this can be downloaded compiled for Unix Sun 4, Windows 95 and MS-DOS from http://www.cs.cf.ac.uk/ec/*

Contact: *s.margetts@cs.cf.ac.uk*

[3]*Present address: Department of Computer Science, University of Wales, Cardiff, PO Box 916, Cardiff CF2 3XF, UK*

Introduction

The exponential growth of molecular sequencing data requires the development of advanced computational methods for rapid comparison of new sequences with known genetic material in order to make a decision about their taxonomic (evolutionary) and/or functional relatedness.

Most of the existing methods for genetic sequence classification are based on a computer search for homologies in nucleotide or amino acid sequences. The standard sequence alignment programs have been designed to provide a compromise between the speed (it is known that the problem of multiple alignment is NP-hard) and accuracy of the search. As a result, they work well only when there is a reasonably high degree of sequence identity, usually of the order 30% or more.

Some new computationally inexpensive methods based on nucleotide or amino acid compositional analysis with or without biochemical parameters, e.g. molecular weight (if known) and isoelectric point, have been proposed. Different metrics and distances have been used to express the similarity between sequences, but prediction results are still unsatisfactory and depend on the features chosen to represent the sequences.

The frequencies of the oligonucleotides or amino acids are commonly used as classificational features. Nussinov (1984) has found that there are stable and statistically significant asymmetries in some dinucleotide frequency distributions for different taxonomic groups. According to the 'genome hypothesis' formulated by Grantham *et al.* (1980), the choice between synonymous codons varies from one gene to another and depends on the type of genome the gene occurs in. So codon usage may be regarded, to some extent, as a measure of 'genome' similarity and gene expressivity (Cowe and Sharp, 1991). Even the amino acid composition is not random and, as has been shown in Hobohm *et al.* (1994), can be used for protein identification.

On the other hand, as has been shown in Hobohm *et al.* (1994), some classification procedures may perform better if the number of features is reduced according to some biochemical rationale and only a part of the frequency distribu-

© Oxford University Press

tion is used. This also corresponds to current trends in neural networks: the principle of minimal architecture.

It is clear that the ideal feature selection for sequence classification should satisfy the following conditions: similar sequences should be represented by similar feature vectors; similar feature vectors should give similar or equal classification value. Methods for feature selection proposed in Pietrokovski *et al*. (1990) and Wu (1996) satisfied the first of these conditions. They do not incorporate classification values at the stage of feature selection.

In this paper, the formal feature selection method based on the Gamma (or near-neighbour) test (Koncar, 1997; Stefánsson *et al*., 1997) is proposed. This procedure appears to give accurate (probabilistic) estimates for the mean-squared error of the classification variable, for a wide class of feature vectors (not especially for oligonucleotide or amino acid frequency distribution), independently of any detailed knowledge of the function from input feature space to output classification target except that it should be smooth (bounded first- and second-order partial derivatives). The process of finding the best subset of the features is speeded up by using genetic algorithms (Holland, 1975) and a *kd*-tree technique for the construction of the nearest-neighbour lists (Friedman *et al*., 1977).

The method was used for feature selection and classification of the large subunit of rRNA (LSU rRNA) according to RDP (Ribosomal Database Project) phylogenetic classes (Maidak *et al.*, 1994). The sequences were represented by dinucleotide frequency distribution. The nearest-neighbour criterion has been used to estimate the predictive accuracy of the classification based on selected features.

System and methods

Data representation

Denote by P a finite alphabet of cardinality |P|. Let T be a string $a_1\, a_2 \ldots a_N$ of length N over the alphabet P. A substring of length l such as $a_i\, a_{i+1} \ldots a_{i+l-1}$ is called an oligonucleotide or l-gram. There are $|P|^l$ possible l-grams over the alphabet P and $N - l + 1$ l-grams in the text of length N.

Any text T may be represented by its l-gram frequency distribution or, in other words, by the l-gram spectrum (Chuzhanova, 1989) for fixed l which consists of pairs {l-gram, the number of occurrences of this l-gram in text T}. When the order of l-grams is fixed, the spectrum contains only the frequencies of corresponding l-grams.

Feature selection via the Γ-test

Let a data sample be represented by $((x_1, \ldots, x_m), y) = (\mathbf{x}, y)$ in which we think of the vector $\mathbf{x} = (x_1, \ldots, x_m)$ as the input, confined to a closed bounded set $C \subseteq \mathbb{R}^m$ and the scalar y as the output. In the present case, the input is the l-gram spectrum of a data sample and the output is the class number to which the data sample belongs.

We assume that training and testing data are different sample sets in which:

(i) the training set inputs are non-sparse in input space;
(ii) each output is determined from the inputs by a deterministic process which is the same for both training and test sets;
(iii) each output is subjected to statistical noise with finite variance whose distribution may be different for different outputs, but which is the same in both training and test sets for corresponding outputs.

We focus on the case where samples are generated by a suitably continuous but unknown function $f: C \subseteq \mathbb{R}^m \to \mathbb{R}$ and $y = f(x_1, \ldots, x_m) + r$, where r represents an indeterminable part, which may be due to real noise or might be due to lack of functional determination in the posited input/output relationship, i.e. an element of 'one → many-ness' present in the data. The Gamma test is designed to give a data-derived estimate for the variance Var(r).

Suppose $(\mathbf{x}, y)$ is a data sample. Let $(\mathbf{x}', y)$ be a data sample such that Euclidean distance $|\mathbf{x}' - \mathbf{x}| > 0$ is minimal and the minimum is taken over the set of all sample points different from $\mathbf{x}$. Thus, $\mathbf{x}'$ is the nearest neighbour to $\mathbf{x}$ (in any ambiguous case we just pick one of the several equidistant points arbitrarily). The Gamma test (or near-neighbour technique) is based on the statistic:

$$\gamma = \frac{1}{2M}\sum_{i=1}^{M}(y'(i)-y(i))^2$$

where M is the number of input/output training pairs.

It can be shown that $\gamma \to \mathrm{Var}(r)$ in probability as the nearest-neighbour distances approach zero. In a finite data set, we cannot have nearest-neighbour distances arbitrarily small so the Gamma test is designed to estimate this limit by means of a linear correlation.

Given data samples $(\mathbf{x}(i), y(i))$, where $\mathbf{x}(i) = (x_1(i), \ldots, x_m(i))$, $1 \le i \le M$, let $\mathbf{x}(N(i, p))$ be the pth nearest neighbour to $\mathbf{x}(i)$. Nearest-neighbour lists for $p \le 20$ (say) nearest neighbours can be found in O(Mlog M) time using a *kd*-tree technique developed by Bentley *et al*. (1977).

We write:

$$\Delta(p) = \frac{1}{p}\sum_{h=1}^{p}\frac{1}{M}\sum_{i=1}^{M}|\mathbf{x}(N(i,h))-\mathbf{x}(i)|^2$$

and

$$\Gamma(p) = \frac{1}{p}\sum_{h=1}^{p}\frac{1}{2M}\sum_{i=1}^{M}(y(N(i,h))-y(i))^2$$

then $\Delta(p)$ is the mean square distance of the $h \le p$ nearest neighbours and $\Gamma(p)$ is an estimate for the statistic γ based on

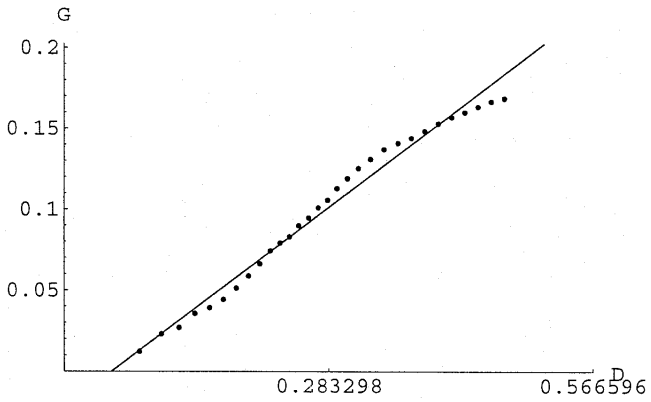

Fig. 1. Gamma test regression line on the LSU rRNA dataset based on 2-grams, where selected features are included, see Koncar (1997).

the $h \leq p$ nearest neighbours. Both $\Delta(p)$ and $\Gamma(p)$ are easily computed from the data set.

As p increases, we might expect to find that $\Gamma(p)$ grows approximately linearly with $\Delta(p)$ for small p. We certainly found this correlation to hold when function f was known to be smooth. The linear correlation of $\Gamma(p)$ with $\Delta(p)$ suggests a possible method for dealing with the fact that in a finite data set we cannot have $max\ |\mathbf{x}' - \mathbf{x}|$ arbitrarily small. We can estimate $\Gamma(p)$ and $\Delta(p)$ for the first several values of p, and extrapolate the regression line to $\Delta = 0$. The intercept $\overline{\Gamma} = \lim \Gamma$ (see Figure 1) will usually give an improved estimate for Var(r).

The Gamma test can be used to estimate the best embedding dimension: the minimal $\overline{\Gamma}$ corresponding to the more important and informative subset of features. On data sets which are not excessively large and the number of features m is less then 20, the test is sufficiently fast to be run on a complete examination of all possible $2^m - 1$ subsets of features. For a larger number of features, we use a genetic algorithm (in the sense of Holland, 1975). Every feature selection S_j, $1 \leq j \leq 2^m - 1$ is represented by a binary string of length m, where '0' in the ith position means that the ith feature is excluded from current selection and '1' indicates its presence otherwise. The fitness of each member (individual) S_j of the population $\{S_1, \ldots, S_K\}$ ($K \leq 2^m - 1$) is a function of the $\overline{\Gamma}$-value: $Fitness(S_j) = 1/(1 + e^{\overline{\Gamma}(j)/\tau})$, where $\tau > 0$ is a constant which controls the shape of the fitness curve, and $\overline{\Gamma}(j)$ is the $\overline{\Gamma}$-value found by running the Gamma test with features selected according to S_j.

An initial population of unique random bit strings is generated and the fitness of each is found as above. In each generation, a set number of breeding events take place. Two parents P_1 and P_2 are randomly selected from the current population: P_1 with a probability proportional to its fitness and P_2 uniformly from all individuals. These two parents are combined using a one-point crossover operation, by picking a random cut point c, where $1 \leq c \leq m$. A 'child' is formed from the first c bits of P_1 and the last $m - c$ bits of P_2. This child is then subjected to a mutation operator which flips each bit in the child according to a small mutation probability (the experiments performed used a probability of 0.01). If the resulting child is unique and not the all-zero string, its fitness is calculated as above and it is inserted into the population by overwriting the least fit individual. If not, another cut-point is chosen and a new child generated. If no unique child can be formed after 10 iterations, a new pair of parents is chosen. Testing the uniqueness of the children in this way takes very little effort and ensures that the population remains diverse.

Accuracy of feature selection

To estimate the accuracy of feature selection, the proposed method was tested on a set of sequences with known feature preferences.

Let us consider Markov chains of order 0 where the symbols from the alphabet $\{A, C, G, U\}$ occur with probabilities p_A, p_C, p_G and p_U. It is clear that the probabilities of corresponding bigrams will be $p_{AA} = p_A \cdot p_A$, $p_{AC} = p_A \cdot p_C$ and so on. Suppose that the training set includes two classes of the sequences over the alphabet $\{A, C, G, U\}$ generated with the probabilities $p_A = 0.25$, $p_C = 0.25$, $p_G = 0.25$, $p_U = 0.25$ and $p_A = 0.2$, $p_C = 0.2$, $p_G = 0.3$, $p_U = 0.3$. In other words, the sequences from the second class have more letters G or U and, correspondingly, more bigrams GU or UG ($p_{GU} = p_{UG} = 0.09$) than the sequences from the first class ($p_{GU} = p_{UG} = 0.0625$). Thus, the frequency of bigrams GU or UG might be a good classification feature for these particular classes.

We were therefore encouraged to find that the feature selection method proposed herein, when applied to this problem, does indeed select the frequency of bigram UG as the most informative feature.

The analysis of the Markov chains of order 0 with other symbol distributions has shown that if there are asymmetries in bigram frequencies then the feature selection obtained by the method discussed here is the same as expected according to the probabilities of bigrams.

Implementation

Ribosomal LSU RNA classification

To illustrate its possibilities, the method was used for feature selection and LSU rRNA classification according to RDP phylogenetic classes (Maidak *et al.*, 1994). The RDP database was chosen for the following reasons. First, the RDP database is one of the few databases that organize their entries according to phylogenetic relationships with a high level of accuracy. Second, its size is relatively small. Third, classification and annotation of unknown rRNA sequences is now very important as a method of biosphere monitoring.

Table 1. LSU rRNA phylogenetic classes

Class no.	No. of sequences	Class name
1	3	Archaea.Crenarchaeota
2	11	Archaesta:Euryarchaeotata:Archaeoglobales
3	2	Bacteria:Flavobacteria and relatives
4	10	Bacteria:Gram Positives and relatives, High G+C
5	23	Bacteria:Gram Positives and relatives, Low G+C
6	5	Bacteria:Proteobacteria Alpha
7	8	Bacteria:Proteobacteria Beta
8	2	Bacteria:Proteobacteria Epsilon
9	7	Bacteria:Proteobacteria Gamma
10	2	Bacteria:Spirochetes
11	9	Eukarya:Animalia:Arthropoda:Uniramia:Insecta
12	5	Eukarya:Fungi:Eumycota:Ascomycotina Hemiascomycetes
13	6	Eukarya:Plantae:Magnoliophyta::Magnoliopsida
14	15	Eukarya:Protoctista:Zoomastigina::Kinetoplastida
15	69	Mitochondria:Animalia:Arthropoda:Uniramia:Insecta
16	4	Mitochoneria:Fungi:Eumycota: Ascomycotina:Plectomycetes
17	3	Mitochondria:Plantae:Bryophyta::Marchantiopsida
18	21	Mitochondria:Protoctista:Rhizopoda::Lobosea
19	8	Plastids:Plantae:Magnoliophyta::Magnoliopsida
20	23	Plastids:Protoctista:Chlorophyta::Chlorophyceae

Table 2. Training, embedding and prediction results

Group ID	No. of training sequences	No. of selected features	$\overline{\Gamma}$ value ($P \leq 20$)	No. of testing sequences	Prediction accuracy (%)		
					First fit	5-fits	10-fits
1	165	8	1.92×10^{-7}	70	70	84.3	93
2	160	10	5.2×10^{-7}	75	74.7	92	94.6
3	161	8	2.8×10^{-6}	74	64	75	88
4	236	13	5.1×10^{-9}	34	80	91	94

The data used for the Gamma test contained 236 LSU rRNA from 20 phylogenetic classes (Table 1) derived from the RDP database (Release 5.0, December 13, 1996). Five classes were rejected because each had only one sequence present in the database. The alphabet of RNA includes four symbols {A, C, G, U}. The approximate length of the sequences is 3200–5000 nucleotides. The sequences were represented by their bigram spectrum or, in other words, by 16 features. To make the sequence representation length invariant, all values were scaled by $N_i - l + 1$ where N_i is the length of the ith sequence. Note that the choice of the length l is very important. Values $l = 2, 3$ are preferable both from the biological point of view (see the Introduction) and for the robustness of prediction: with $l > 3$, the training set is separated better but the robustness of prediction falls down (Gusev and Chuzhanova, 1990).

To estimate the predictive accuracy of the proposed method, a set of sequences was randomly divided into three groups of approximately equal size. Two of them were used for feature selection (or training) and one for prediction. All three combinations have been employed. In the fourth case, all 236 LSU rRNA were used for feature selection and 34 other sequences from 10 phylogenetic classes for predicting (i.e. testing).

The full embedding has been carried out for each case and subsets of features were selected using the Gamma test. During the prediction phase for each sequence from the test set n nearest neighbours in this selected feature space are computed. The output values of these n nearest-neighbour sequences are then used as outputs for the new sequences. For 'first fit' $n = 1$ and the class number is predicted as simply the class of the first nearest-neighbour sequence. For 'five fits' $n = 5$ and correct classification lies within the set of the first five nearest neighbours, etc.

The information about training and testing sets, the number of selected features and prediction results according

to one (first fit), five (5-fits) and 10 nearest neighbours (10-fits) are shown in Table 2.

As can be seen from Table 2, the predictive accuracy according to the 10 nearest neighbours is ~94%. Of course, once the search has been narrowed with high probability to 10 possibilities, then a more detailed examination of these 10 sequences can feasibly complete the classification. For the sequences which are not recognized correctly according to n nearest neighbours, where $1 \leq n \leq 10$, the predicted n classes, although incorrect, are nevertheless closely located on the phylogenetic tree. It would also be possible to train a neural network using these features and we shall report neural network classification results in a later paper. Here we are more interested in the feature selection procedure.

Conclusions

The method of feature selection and classification described here does not require searches for homologies, common subsequences or specific patterns, all of which are very time consuming. It gives the opportunity to classify the new sequences into predefined classes on the phylogenetic tree without sequence alignment. The method is robust in the sense that, when errors occur, the incorrect classification is phylogenetically close to the correct classification.

Experiments with neural networks on the 72 LSU rRNA from 15 phylogenetic classes derived from the same database have been reported (Wu, 1996). The sequences were represented by their octagram spectrum ($l = 8$). The singular value decomposition method was used to reduce the number of octagrams to 40. In comparison, the predictive accuracies are higher, from 92 to 100%. Nevertheless, the results of Gusev and Chuzhanova (1990) suggest that methods based on the frequency of longer l-grams (in fact $l > 3$) are much less likely to be robust, in the sense that given a particular example of a class one can readily find an extended l-gram whose presence characterizes this example uniquely. However, the same l-gram is relatively unlikely to occur without any mutations in another example of the same class. Thus, frequencies of a diverse range of shorter l-grams taken together are likely to provide a much more robust classification.

This is a preliminary account of a new technique and is principally designed to show that the method has promise. One can envisage many improvements, e.g. the Euclidean metric may not be the most appropriate given the context, and the weighting of features could be continuous rather than discrete, but these are questions that can reasonably be addressed in future studies. Of course, it is also possible to improve the prediction accuracy by enlarging the training set and by increasing the quality of the samples.

References

Chuzhanova,N. (1989) Inductive method of program synthesis for symbolic sequence processing. In *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Berlin, Vol. 397, pp. 317–327.

Cowe,E. and Sharp,P.M. (1991) Molecular evolution of bacteriophages: discrete patterns of codon usage in T4 genes are related to the time of gene expression. *J. Mol. Evol.*, **33**, 13–22.

Friedman,J.H., Bentley,J.L. and Finkel,R.A. (1977) An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software*, **3**, 200–226.

Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pave,A. (1980) Codon catalog usage and genome hypothesis. *Nucleic Acids Res.*, **8**, 49–62.

Gusev,V.D. and Chuzhanova,N.A. (1990) The algorithms of recognition of the functional sites in genetic texts. In *Proceedings of the First International Workshop on Algorithmic Learning Theory*. Japanese Society for Artificial Intelligence, Tokyo, pp. 109–119.

Hobohm,U., Houthaeve,T. and Sander,C. (1994) Amino acid analysis and protein database compositional search as a rapid and inexpensive method to identify proteins. *Anal. Biochem.*, **222**, 202–209.

Holland,J. (1975) *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA.

Koncar,N. (1997) Optimisation methodologies for direct inverse neurocontrol. PhD Thesis, Department of Computing, Imperial College, London.

Maidak,B. *et al.* (1994) The ribosomal database project. *Nucleic Acids Res.*, **22**, 3485–3487.

Nussinov,R. (1984) Strong duplet preferences in nucleotide sequences and DNA geometry. *J. Mol. Evol.*, **20**, 111–119.

Pietrokovski,S., Hirshon,J. and Trifonov,E.N. (1990) Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. *J. Biomol. Struct. Dyn.*, **7**, 1251–1268.

Stefánsson,A., Koncar,N. and Jones,A.J. (1997) A note on the Gamma test. *Neural Comput. Applic.*, **5**, 131–133.

Wu,C.H. (1996) Gene classification artificial neural system. *Methods Enzymol.*, **266**, 71–88.