Clustering Techniques in Biological Sequence Analysis*

A.M. Manning[1], A. Brass[2], C.A. Goble[3], J.A. Keane[1]

[1]Department of Computation, UMIST, Manchester, M60 1QD, UK.
[2]School of Biological Sciences, University of Manchester, M13 9PL, UK.
[3]Department of Computer Science, University of Manchester, M13 9PL, UK.

Abstract. In biological sequence analysis many DNA and RNA sequences discovered in laboratory experiments are not properly identified. Here the focus is on using clustering algorithms to provide a structure to the data. The approach is inter-disciplinary using domain knowledge to identify such sequences. The enormous volume and high dimensionality of unidentified biological sequence data presents a challenge. Nonetheless useful and interesting results have been obtained, both directly and indirectly, by applying clustering to the data.

1 Introduction

Data mining can be defined as the nontrivial deduction of implicit, previously unknown and potentially useful information from data [7]. Approaches can be divided into *supervised* and *un-supervised* learning. Supervised learning is when known cases are used to indicate well defined patterns so that generalisations can then be made about the data as a whole. In un-supervised learning patterns are found by starting not from known examples but rather from a logical characterisation of regularities within the data space.

Clustering is based on un-supervised learning. It is the process of sorting objects into clusters such that the degree of association between members of a group is high and that between members of different groups is relatively low. Clusters may be distinct in some cases and overlapping in others [4]. Clustering can help develop a class structure against which new cases can be judged and can detect previously unperceived structures[1].

In this paper we consider the application of clustering techniques in biological sequence analysis. In the Human Genome Project [6] a significant problem is the amount of unrecognised DNA and RNA sequences produced. It is probable that important information about the human genome is hidden in these unrecognised sequences. Any method that can aid their identification is thus extremely valuable.

The investigation is based on an inter-disciplinary approach using domain expertise. This expertise is vital for valid interpretation of un-supervised learning classification. We establish that two Bayesian clustering algorithms are efficient in dividing a sample data set of known DNA/RNA sequences into biologically sensible clusters. This, however, provides limited information for use in individual sequence identification. Further

* Work supported by UK EPSRC MRes studentship and ESPRIT HPCN Project No 22693.

[1] NB clustering and classification are used interchangeably here although classification usually refers to a process where the classes are predefined [7].

applications of the algorithms to different manipulations of the same data set were necessary. Such algorithms require data reduction and sampling processes to be performed before they can be applied. “Biologically interesting” clusters have been derived.

The structure of the paper is as follows: §2 gives biological background; §3 and §4 discuss clustering and its application in this area; §5 presents results; §6 concludes and discusses further work.

2 Biological Background

The human genome consists of a complete set of instructions for creating and maintaining human life. This is made up of tightly coiled threads of DNA and associated protein molecules. A DNA molecule is made up of two strands, each being a linear arrangement of units consisting of one sugar, one phosphate and a nitrogenous base. There are four bases present in DNA, encoded *A*, *T*, *C* and *G*. These strands are held together with weak bonds consisting of pairs of bases (referred to as base pairs, *bp*). The order of the bases along each strand in any particular case is called the DNA sequence.

A DNA molecule contains many genes, each comprising a specific sequence of bases, which hold the necessary information for producing proteins. These proteins provide the structural components of cells and tissues as well as enzymes for essential biochemical reactions. The human genome is estimated to comprise at least 100,000 genes which vary considerably in length. The protein-coding instructions from the genes are transmitted indirectly through messenger ribonucleic acid (mRNA), a molecule similar to a single strand of DNA.

It is possible for the mRNA molecule to be isolated and used as a template to synthesize a 'coded' DNA (cDNA) strand. The cDNA's are believed to identify the parts of the genome with the most significant biological and medical characteristics because they represent expressed genomic regions - i.e. those that transcribe into mRNA. A common reference system consisting of a partial sequence of 300-500bp of the complete cDNA sequences is used for reasons of convenience. The partial sequences (termed *Expressed Sequence Tags* or ESTs) serve as markers but can also identify expressed genes. Such a system therefore gives an efficient method of identifying most human genes [1, 8].

Pharmaceutical companies have enormous databases of ESTs of which about 30% have been identified. Computational tools exist that match unidentified ESTs against known sequences with certain similar characteristics. Nevertheless, these techniques do not give a clear picture of where the EST in question fits into the database as a whole, i.e. groups of sequences that are related in varying degrees. Such information not only helps with EST identification, it is also useful, for example, when developing a new drug aimed at one particular protein as information is provided about its effect on related proteins thus preventing cross reactivity. To provide such information the EST data needs to be given some structure (by, for example, clustering).

2.1 Characteristics of the EST data

The EST data is stored in a flat file made up of reference codes and sequences of bases. For identification purposes one way to represent this data is to have each record represent one sequence and each attribute within the record represent one base from the sequence (with additional attributes for references). For example, the data in Figure 1 (which refers to a particular human RNA sequence) consists of two reference codes, a number representing the length of the sequence (i.e. 381bp) and 381 bases. It would, therefore, need a record of 384 attributes in which to store it.

```
X89055 HSIGAF2H6 381  gaggtgcagc tggtggagtc tgggggaggc ttggtccagc ctggggggtc
cctcaaactc tcctgtgcag cctctgggtt caccttcagt ggctctacta tgcactgggt ccgccaggct
tccgggaaag ggctggagtg ggttggccgt ataagaaaca aagacaacag ttatgcgaca gcatatgctg
cgtcggtgaa aggcaggttc accatctcca gagatgattc agagaacacg gcgtatttgc aaatgaacag
cctgaaaatc gaggacacgg ccgtctatta ctgtactagg gggtctagta tggttcgggg agtaaacggc
tactacggca tggacgtctg gggccaaggg accacggtca ccgtctcctc a
```

Figure 1. Representative EST record

The main characteristics of the EST data when stored in the above way are:

1. there are no labelled (classified) examples;
2. each record is generally independent of any other: dependencies can occur between records if they contain sequences from the same gene but these are minimal;
3. the number of attributes within a record is not uniform due to differing EST lengths;
4. there are an insignificant number of missing values;
5. the database contains around a million records, each with 300-500 attributes;
6. the attributes within the records of the database are not independent.

3 Clustering

Un-supervised learning is appropriate as no prior information is known about the structure of the EST data. For identification, information is needed about relationships between clusters within the data. Two un-supervised clustering systems, *AutoClass C* and *Snob* (both based on Bayesian statistics), were used on the data.

AutoClass C seeks a *maximum posterior* probability classification [3]. Inputs consist of a database of attribute vectors (cases), either real or discrete valued, and a cluster model. AutoClass C finds the maximally probable set of clusters with respect to the data and model. The output is a set of cluster descriptions and partial membership of the cases in the clusters. Each instance has a probability of being a member of each different cluster and consequently such clusters are "fuzzy". AutoClass C has been used in biological sequence analysis to classify DNA intron data [3].

Like AutoClass C, Snob regards the best clustering of the data as that with the highest posterior probability [10]. To achieve this it uses the *Minimum Message Length* principle. The input requirements are similar to AutoClass C and cases are given a probabilistic class membership.

AutoClass C and Snob determine classes automatically. They can use both discrete and real valued data and their processing time is roughly linear to the volume of data.

Their main limitations, in relation to the EST problem, is inability to manage varying record lengths and relationships between attributes within a record. Both also have a limit to the amount of data that they can handle in a 'reasonable' amount of time. For AutoClass C the number of cases multiplied by the number of attributes should be no more than `10**6`.

4 Applying AutoClass C and Snob to the EST data

Representative biological sequences (54,129 in total) were selected from the EMBL Nucleotide Sequence Database [5]. This is a comprehensive database of DNA and RNA sequences which are all identified. All records chosen for this work were from primates. The data set records were formed from these sequences, using each base as an attribute. Each record, therefore, contained a varying number of attributes as the number of bases per sequence can range from under 100 to over 1000. The following preliminary transformation was necessary:

1. Occurrences of triplets of A, C, G, T (i.e. the number of times a set of 3 bases, e.g. GCA, occur together) were recorded for each sequence to produce uniform records. The percentage frequency of each triplet was then stored in a record with 64 attributes (one for each possible triplet) with 3 further attributes for references.
2. Although step 1 resulted in uniform records the 64 percentage triplet frequency attributes were clearly not independent. This was achieved by applying a *Principle Component Analysis* which reduced the dimension of the data by selecting the base triplets which had the most influence. This resulted in records containing 48 attributes together with 3 for references.

The altered data set still contained 54,129 cases each with 51 independent attributes. AutoClass C was initially applied to a 10% sample of this data set (using a Pentium 90, with 16 Mbytes memory) and ten attempts at finding the most probable classification were made, the best containing 60 clusters. The two largest clusters contained very long and very short sequences, both of which are unhelpful: very long sequences often provide information on more than one gene, and very short sequences usually do not provide enough information on a single gene. All sequences less than 300bp or more than 800bp were removed together with all sequences of human DNA of type 'CpG Island repeat' which were felt to be unimportant. Approximately 10,000 sequences remained and one in three of these were selected. This modified set is referred to as *Data Set 2*. AutoClass C was applied to Data Set 2 and after 93 tries (i.e. 93 tests for a new number of clusters) a classification of 50 clusters was found to be most probable.

Further applications of AutoClass C were made to Data Set 2 specifying a larger margin of error (0.01 instead of 0.001) for the data. The aim was to force the clusters in the data to overlap more so that additional information could be gained about relationships between them. AutoClass C was also applied to a data set that contained the frequency of groups of four bases, as opposed to the triplets already considered, because the groups of four give more information about the structure of DNA/RNA sequences. The two best classifications from these were considered and common overlaps recorded (i.e. cases that appeared together within clusters in all four classifications) - this is referred to below as the *four way overlap* data.

Snob was allowed to run for 19 hours (using data set 2) on a Sun Workstation (10 MIPs processor and 24 Mbytes of memory) This resulted in 62 classes.

5 Results

The most useful AutoClass C report resulted from sorting the cases into their respective clusters and tabulating their allotted case number (which corresponded to the position of the case in the original data set) and reference codes for each. Software was also written to match each case with its description from the EBML database [5]. The descriptions for the cases in each cluster were then studied and, where possible, a label was given to each cluster representing the general trend of these descriptions.

Figure 2 illustrates the general pattern of clusters. Several distinctive clusters can be seen in the diagram. AutoClass C produces a heuristic measure of cluster strength - the approximate geometric mean probability for instances belonging to each cluster, computed from the cluster parameters and statistics. This approximates the contribution made, by any one instance of the cluster, to the log probability of the data set with respect to the classification. It thus provides a heuristic measure of how strongly each cluster predicts its instances. In a 'strong' class the entries were found to have very similar descriptions. The descriptions of the cases in the strongest cluster, cluster 47, were almost identical whereas in a 'weaker' cluster general characteristics were less evident.

The reports also provided details of how the clusters overlapped with each other; when a case had a probability of 98% or less of being in a particular cluster alternatives were given. This was used to locate strong overlaps between the clusters (see Figure 2). Groups of overlapping clusters are referred to as batches.

Relatively 'strong' clusters are generally those which do not overlap (e.g. 38, 47), or appear at the edge of a batch (e.g. 10 and 29 which are not part of the rather vague IG group indicated by the irregular shaded area in Figure 2). Weaker clusters tend to exist in batches which are difficult to label, e.g. the unlabelled batch at the bottom left of Figure 2.

Various clusters of IG light chain have been distinguished, together with a cluster of high Alu content. A cluster of high tandem repeat (number 23) together with one of alphoids and DNA alphas (number 31) were also isolated.

AutoClass C appears to group together cases that it finds difficult to identify. This was evident from the first application (discussed in §4) which separated very long and very short sequences. In this, the second application, clusters that are 'weak' and are not characterised by one 'strong' feature are grouped together. These are clusters which were too vague to do anything with. This can be influential in analysis, allowing attention to the cases which give most important information and indicating where more attention is needed.

AutoClass C only appears to have confidence in clusters whose cases are well represented. The data used was dominated by IGs and thus the clusters containing IGs were

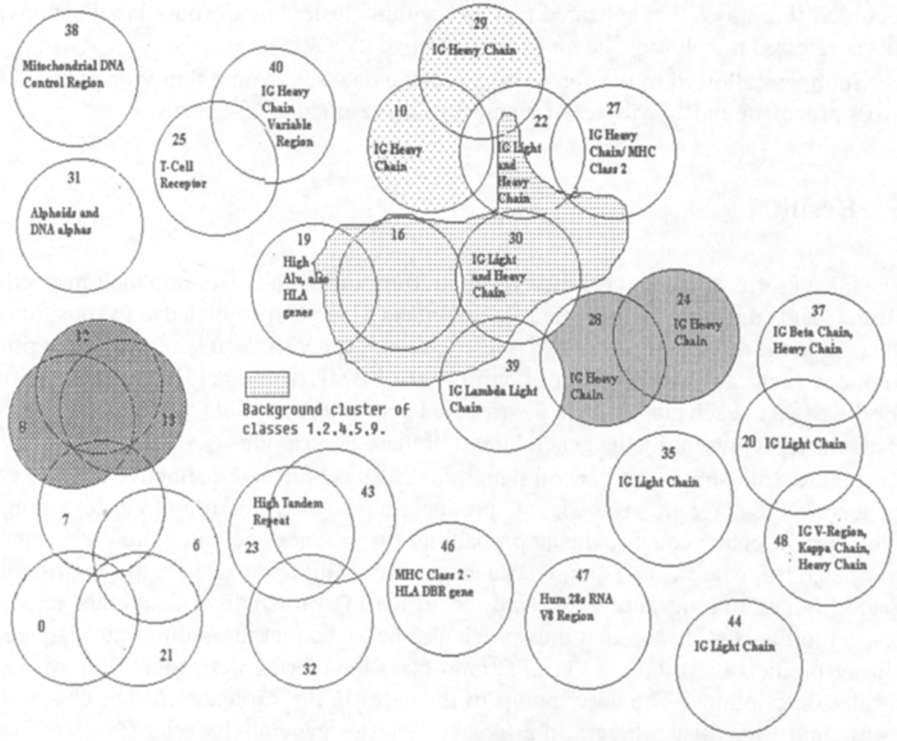

Figure 2. Auto Class C Clusters

relatively 'strong'. This can be misleading as it suggests that cases in 'weaker' clusters are less important whereas their 'weakness' may be a result of bias in the data.

The clusters do not appear to be those that would result from conventional methods. For example, one method for forming clusters in biology is to randomly select a case and then, using the results of Principle Component Analysis [9], include all cases which fall within a certain distance from it. This tends to produce classes of a neat circular shape whereas AutoClass C is inclined to create a less regular figure. This suggests that AutoClass C is not merely forming clusters on the basis of measurement.

5.1 Comparisons between the AutoClass C and Snob Classifications

Snob was used mainly for comparative purposes. The Snob clusters overlapped to some extent with those from AutoClass C. The percentage overlap of each cluster of AutoClass C was taken with its closest match in Snob (cases which had a 95% or more probability of belonging to the cluster in either classification were considered). The average overlap was 41.6%, although this was far higher for 'strong' clusters. The correlation coefficient [2] between class 'strength' and overlap was 0.6 which indicates a significant relationship.

The clusters in AutoClass C and Snob were often found to contain cases from corresponding batches in the two classifications; for example, this was true in 99.5% of the cases contained in the 'strongest' 35 AutoClass C clusters. This indicates that the classifications produce similar patterns of batches even though the individual clusters might themselves be different.

In summary, there was little significant difference between the classifications.

5.2 The nature and importance of the four way overlap data

For EST identification purposes the above classifications were useful, particularly in the incidence of 'strong' clusters where cases are similar. With 'weaker' clusters EST identification would not be so straightforward as the contents tend to vary widely. The general size of a cluster in the AutoClass C classification was also quite large, an average of 67.56 cases per class, which clearly exacerbates identification.

Useful results came from studying the *four way overlap* data. In this two of the initial clusters, 38 and 47, remained complete whereas other 'strong' clusters (e.g. 40) were split into a small number of subsets. Weaker clusters, such as 6, were broken up into numerous very small subsets which were often mixed with cases from other clusters. Most groups recorded contained less than 10 members - frequently groups of only 2 or 3 cases were found. These small groups usually exhibited some common trend: for example, case 17 of cluster 1 (*human gene for growth hormone*) and case 2648 of cluster 2 (*human gene for somatomammotropin hormone*) were grouped together (i.e. they had occurred together in clusters in each of the four classifications considered) and were found to have almost identical structures. Without the four way overlap data this would not have been detected: clusters 1 and 2 are large and 'weak' and this information would not have been apparent. When considering EST identification such information could be invaluable: for example, if case 17 was unknown, the knowledge that it was likely to be similar in some way to case 2648 could greatly speed up its detection.

For AutoClass C to be really useful for identifying ESTs information from the 'four way overlap' should be used with the results from a single classification.

6 Conclusions and Future Work

Clustering algorithms have been applied to EST identification with some success, producing "biologically interesting" data, although, as yet, there are no new biological discoveries. Domain expertise for validation of the results at each phase has been extremely important, providing useful guidance for what to focus upon in the next phase.

Clusters were formed (with no prior information) which often exhibited a strong biological trend. However, such clusters were often too large (unless very 'strong') to aid EST identification significantly. Only after applying AutoClass C (in this case) to different manipulations of the same data (e.g. taking base quadruples instead of triplets) and noting where cases group together under each classification was it clear that clustering techniques could be of benefit in the EST identification process.

Two areas to improve the investigation have been noted. Firstly, data reduction and projection techniques had to be applied to the data in order for the algorithms to accept it when ideally the data should be considered in as raw a state as possible.

Secondly, it was not possible with the available platforms to apply either algorithm to the whole database - a sample of less than 1% was the largest considered. Working with a sample can create false bias so that some cases are not classified properly on a preliminary analysis. The database had almost 1 Million records each with about 50 attribute (each attribute containing up to 9 characters). 5,000 records have been tackled at any one time; each try takes 2 hours (Pentium 90, 24 Mbytes). Usually between 100 and 150 tries are necessary and thus the process takes several days. Parallel systems with more processors and larger memory allied to the parallelising of the algorithms may address this issue. A preliminary analysis of the process of clustering in AutoClass C has identified potential parallelism: the generation of the number of classes (call this c) to be sought in each try depends on the success of previous values of c. An asynchronous parallel algorithm may be appropriate, with each value of c in each block of tries (i.e. each parallel set) being generated only from the results of previous blocks. It may also be possible to stagger tries within each block so that the results from tries at the beginning of the block can be used to generate values of c for tries nearer the end.

References

1. Australian Biotechnology Association (ABA), What is genetic engineering, Educational leaflet, http://www.aba.asn.au/leaf2.html, 1996.
2. M. Bland, *An Introduction to Medical Statistics*, Oxford Medical Publications, 1994.
3. P. Cheeseman and J. Stutz, Bayesian Classification (AutoClass): Theory and Results, *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), AAAI Press, pp. 153-181, 1995.
4. M.J. Currie and Q.A. Parker, Clustan - A Cluster-Analysis Package, Science and Engineering Research Council, Rutherford Appleton Laboratory, Starlink Project, User Note 26.6, 1993.
5. EMBL Nucleotide Sequence Database: http://www.ebi.ac.uk, 1997.
6. K.H. Fasman, A.J. Cuticchia and D.T. Kingsbury, The GDB (TM) Human Genome *Database Anno, Nucl. Acid. R.* 22 (17), pp. 3462-3469, 1994.
7. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.) *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1995.
8. D. Jacobson, Mapping and sequencing the human genome, http:/www.gdb.org/Dan/DOE/prim2.html, 1995.
9. I.T. Jolliffe, *Principle Component Analysis*, Springer Series in Statistics, 1986.
10. C.S. Wallace and D.L. Dowe, Intrinsic classification by MML - the Snob program, *Proc. 7th Australian Joint Conference on Artificial Intelligence* World Scientific, pp. 37-44, 1994.