Neural Nets WIRN11
B. Apolloni et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-972-1-162

Identification of promoter regions in genomic sequences by 1-dimensional constraint clustering

A. BERTONI [a,1], M. RÈ [a], F. SACCÀ [a], G. VALENTINI [a]
[a] *DSI - Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano*

Abstract
Size constrained clustering has been recently proposed to embed "a priori" knowledge in clustering methods. By exploiting the "string property" we propose an exact and efficient algorithm based on dynamic programming techniques to solve size-constrained one-dimensional clustering problems. We show the applicability of the proposed method in a difficult computational biology problem: the prediction of the transcription start sites of genes. The obtained experimental results clearly show the potential of the proposed approach when compared with previously published methods.

Keywords. Monodimensional clustering, constrained clustering, computational biology, TSS prediction

Introduction

Clustering is one of the most used technique in statistical data analysis. A cluster is a set of data points that are similar in some sense, and clustering is a process of partitioning a data set into disjoint clusters. In distance clustering , "similarity" is interpreted in terms of a distance function. Distance clustering is a difficult problem: it is NP-hard even if the number of clusters is 2 (for arbitrary dimension) [1], or if the dimension of data is 2 (for arbitrary arbitrary number of clusters) [5]. For the Euclidean distance, a well-known heuristics is the Lloyd's algorithm [4]; as it is a heuristic algorithm, there is no guarantee that it will converge to the global optimum. In real-world problems, often people has some information on the clusters: incorporating this information can increase the clustering performance. Models that incorporate instance-level background information are called constrained clustering [2,9,10]. Typically, clustering problems with cluster-based constraints incorporate constraints concerning the size of the possible clusters [2,8,9]: in [7] clustering with cluster size constraints has been studied in the 1-dimensional case. In this setting, it can be proved that the optimal solution verifies the so called "String-Property", i. e. every cluster is composed by consecutive examples. This fact allows to design an exact dynamical programming techniques. In particular, we present an algo-

[1]Corresponding Author: A.Bertoni, Dipartimento di Scienze dell'Informazione, Università di Milano; E-mail: bertoni@dsi.unimi.it.

rithm (1SCC) in the Euclidean case, in time $O(m \cdot n^2)$, where *n* is the size of data and *m* is the number of clusters

Mono dimensional clustering techniques can be applied to the problem of the identification of promoter regions in genomic sequences [12]. Promoters are regions in the genomic sequence, that act as finely regulated switchers enabling the cell to activate or silence the genes; mutations occurring in the promoters regions are involved in the pathogenesis of many diseases. In this paper we apply the 1SCC algorithm to detect the transcription start sites of known genes in a genomic region. Experiments were realized using public data [11] associated with the previously published experiments [12], in order to allow a fair evaluation of performances achievable by the proposed method. Results show the effectiveness of the method.

1. 1-Dimensional size constraint clustering

In distance clustering , "similarity" is interpreted in terms of a distance function. A popular distance is the Minkowski metric $d(x, y) = ||x - y||_p$, derived by the norm L_p that, for d-dimensional data, is:

$$||(t_1, ..., t_d)||_p := \sqrt[p]{|t_1|^p + ... + |t_d|^p}$$

The Euclidean distance and the Manhattan metric are special cases with *p* = 2 and *p* = 1. In this setting, given n points $x_1, ..., x_n \in R^d$ and an integer *m*, Clustering Problem requires to determine a m-partition $(A_1, A_2, ..., A_m)$ of $\{1, 2, ..., n\}$, that minimizes the objective function:

$$W(B_1, B_2, ..., B_m) = \sum_{i=1}^{m} \sum_{j \in B_i} ||x_j - C_{B_i}||_p^p \tag{1}$$

where C_{B_i} is the p-centroid of B_i, i.e.

$$C_{B_i} = arg \min_{c \in R^d} \sum_{j \in B_i} ||x_j - c||^p$$

In real-world problems, often people has some information on the clusters: incorporating this information into traditional clustering algorithms can increase the clustering performance. For instance, clustering problems with cluster-based constraints incorporate constraints concerning the size the possible clusters. Here we are interested in efficiently solving clustering with cluster size constraints in 1-dimensional case.

The problem of 1-Dimensional Size Constraint Clustering (**1SCC**) consists, given n reals $x_1 \leq x_2, ... \leq x_n)$, an integer *m* and a set $F = \{k_1, .., k_s\}$ of size constraints, in finding a m-partition $(A_1, A_2, ..., A_m)$ of $\{1, 2, ..., n\}$, with $|A_1|, |A_2|, ..., |A_m| \in F$ that minimizes (1).

An important condition verified by the optimal partition $(A_1, A_2, ..., A_m)$ is the "'**String-Property**"': $(A_1, A_2, ..., A_m)$ verifies the string-property if every class A_j is composed by consecutive examples.

This result has been proved for 1-dimensional clustering with Euclidean distance in [3]; it has been extended to the case of arbitrary norm $|| \cdot ||_p$ with $p > 1$ in [6] and

furthermore proved in the case of 1-Dimensional Size Constraint Clustering in [7] for all $p \geq 1$.

The fact that the optimal partition verifies the string property allows us to design algorithms for **1SCC** by means of dynamical programming techniques: the algorithm is particularly simple and efficient in case of Euclidean distance. In this case, in fact, by simple manipulations the objective function $W(B_1, B_2, ..., B_m)$ can be rewritten as:

$$W(B_1, B_2, ..., B_m) = \sum_{i=1}^{n} x_i^2 - \sum_{k=1}^{m} \frac{(S_k)^2}{|B_k|}$$

where, for $1 \leq k \leq m$, $S_k = \sum_{i \in B_k} x_i$.

Since $\sum_{i=1}^{n} x_i^2$ is constant, the problem is equivalent to find the partition $(A_1, A_2, ..., A_m)$ of $\{1, 2, ..., n\}$, with $|A_1|, |A_2|, ..., |A_m| \in F$ that maximizes the function:

$$G(B_1, B_2, ..., B_m) = \sum_{k=1}^{m} \frac{(S_k)^2}{|B_k|}$$

Let $D(j, k)$ be the optimal value of G on the instance $\langle (x_1, ..., x_n), j, F = \{k_1, k_2, ..., k_s\} \rangle$. $D(j, k)$ verifies the recurrence:

$$\begin{cases} D(1,k) = 0 \text{ if } k \notin F \\ D(1,k) = \frac{(S_k)^2}{|B_k|} \text{ if } k \in F \\ D(j,k) = Max_{1 \leq g \leq s, k_g \leq k} D(j-1, k-k_g) + \frac{(S_k - S_{(k-k_g)})^2}{k_g} \qquad (1 < j < m) \end{cases}$$

$D(m, n)$ may be computed by the following program:

Algorithm for 1SCC:

Input: $\begin{cases} \text{Reals: } x_1 \leq x_2 \leq ... \leq x_n \\ \text{Integer } m \leq n \\ \text{Constraint set } F = \{k_1, k_2, ..., k_s\} \end{cases}$

Begin procedure
$S(1) \leftarrow x_1$
for $k = 2 : n$ **do** $S(k) \leftarrow S(k-1) + x_k$
for $j = 1 : m$, $k = 1 : n$ **do** $D(j, k) \leftarrow 0$
for $t \in F$ **do** $D(1, t) \leftarrow S_t^2 / t$
for $j = 2 : m$, $k = 1 : n$, $t \in F$ **do**
if $t < k$ **then** $\begin{cases} cand = D(j-1, k-t) + \frac{(S_k - S_{k-t})^2}{t} \\ \textbf{if } cand \geq D(j,k) \textbf{ then } D(j,k) \leftarrow cand. \end{cases}$
return $D(m, n)$
End procedure

The number of arithmetical operations in the previous algorithm is $O(m \cdot n^2)$. By an empirical analysis, the computational time T has been estimated as $T = 3.2 \cdot 10^{-8} \cdot n^{1.95} \cdot m^{1.05}$ sec.

2. Prediction of transcription start sites (TSSs) in the human genome using 1SCC clustering

Mono dimensional clustering techniques can be easily applied to problems of great interest in computational biology. A typical example is the identification of promoter regions in genomic sequences [12]. In this section we introduce briefly some basic notions describing the transcriptional regulation and the role played by the promoter regions in this process.

2.1. Transcription and translation

The information required to drive the synthesis of proteins is encoded in the DNA, a long polymer constituted by nucleotides. The synthesis of proteins occurs in two main steps: transcription and translation. In the first one (transcription) the information encoded in the DNA by mean of regions known as genes is copied into the RNA. Once completed the transcription step, the RNA leaves the nucleus and moves to the cytoplasm. In the second phase (translation) a ribosome read the information encoded into the RNA by means of a set of rules known as genetic code and translate them into a protein.

2.2. Promoter regions

Both the transcription and translation processes are finely regulated in order to ensure the presence of specific proteins only in case of need and in response to specific intra and extracellular signals. The complex network of regulatory events governing the expression of thousands of proteins is of crucial importance because the presence of some proteins in the cell can trig signaling cascades leading to irreversible processes that, if not coordinated, can literally kill the cell. The regulation of transcription initiation is realized by mean of signals located immediately before the genes in the DNA molecule. These signals are bound by regulatory proteins and only if specific combination of signals are covered by specific set of regulatory proteins the transcription process can start. The regions containing the signals located upstream the genes are called promoters. Promoters are finely regulated switchers enabling the cell to activate or silence the genes. Mutations occurring in the promoters regions can disrupt the ability of the cell to control the genes and are involved in the pathogenesis of many diseases.

2.3. Automated identification of promoter regions

The automated identification of promoter regions is an important and active research area in computational biology. The easiest approach for the identification of the promoters is to produce the sequence of all the DNA of an organism and the sequences of all the transcripts. The DNA is a very long molecule and only a little fraction of its information is transcribed. Once obtained the sequences of both the DNA and the transcripts, the transcripts can be mapped onto the DNA producing a series of points in genomics coordinates corresponding to the transcription starting points (TSS). These points can be clustered to identify the initial position of each gene. This allows investigators to infer the location of the proximal promoters because the promoters are defined relatively to the TSS: a proximal promoter is, indeed, a region comprising the 1000 nucleotides upstream the TSS.

3. Experimental setup

The goal of the presented experiments is to verify the extent at which the proposed 1SCC algorithm is able to detect the transcription start sites of known genes in a genomic region. All the experiments were realized using public data [11] associated with previously published experiments in order to allow a fair evaluation of the performances achievable by the proposed method.

3.1. Datasets

In [12] Shmid et al. analyzed a 250 kb region of the human chromosome 12. This region is characterized by a relatively high gene density. Raw data were collected by using a wet lab technique named ChIP-chip, that combines chromatin immunoprecipitation (ChIP) with microarray technology (chip).

The usage of this method enables the investigator to filter out the noise present in the output of canonical microarray experiments, mainly due to the low rate transcription events physiologically occurring outside the promoter regions. ChIP-on-chip is thus a very effective and specific filter enabling molecular biologist to record the transcriptional activity only in regions of the genome known to be compatible with the promoters regions.

The raw output of a ChIP-on-chip experiment is constituted by a collection of peaks, each characterized by volume, height and center and directly related with the frequency of the detected transcriptional events and to a specific coordinate along the genomic sequence. In [12] the authors proposed a mono dimensional clustering method,

filter	τ	**n. cluster**	**TP**
10	750	1	1
10	750	2	2
10	750	3	3
10	750	4	3
10	750	5	3
10	750	6	3
10	750	7	4
10	750	8	5
10	750	9	6
10	750	10	7
10	750	11	7
10	750	12	8
10	750	13	9
10	750	14	9
10	750	15	10
10	750	16	11
10	750	17	12
10	750	18	13
10	750	19	13
10	750	20	14

Table 1. Results obtained in the analysis of filtered data. $\tau = 750$

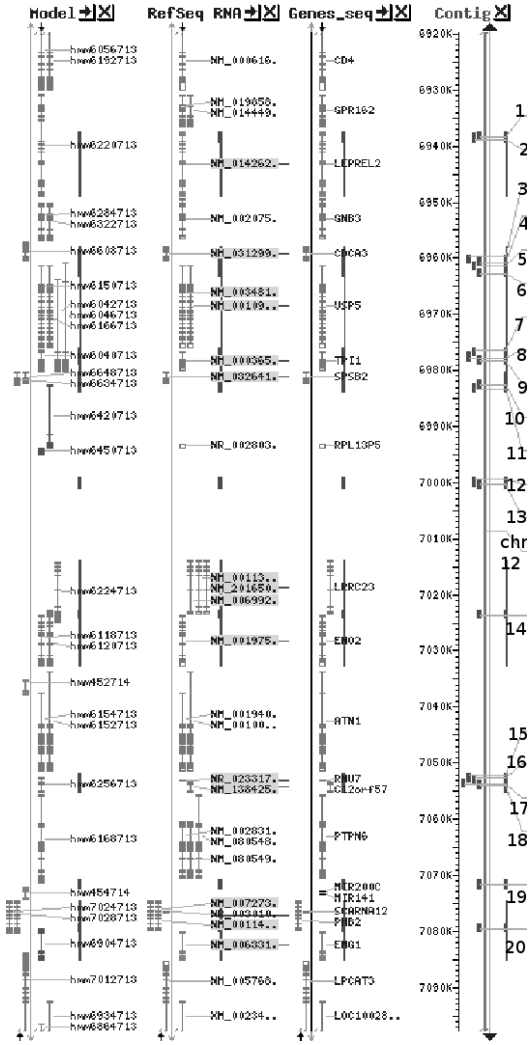

Figure 1. Evaluation of the predicted TSS.

MADAP, based on mixtures of Gaussians and EM algorithms, able to detect the positions in a genomic sequence associated with high transcriptional activity and thus suggesting the presence (and the position) of the genes that are transcriptionally active under the investigated experimental conditions. The expression obtained by evaluation via ChIP-on-CHIP of the genomic region analyzed in [12] (human chromosome 12, 6767416-6990346) and involved in the presented test were downloaded from `www.isrec.isb-sib.ch/madap/`. The usage of the data provided by the MADAP web site ensures the absence of biases in the comparison of the collected experimental results. The downloaded data were preprocessed according to the experimental setup guidelines reported in [12]. In particular we applied a filter aimed to remove all the peaks with an expression events count lower than 10. We then projected the starting nucleotides of the sequences transcripts onto the genome and we finally duplicated each genomic

coordinate times the number of the transcription events. The application of this protocol produced two mono dimensional datasets: the first obtained by projecting the unfiltered data and the second obtained by projecting the filtered data. In [12] the authors identified, in the evaluated genomic regions, 8 clusters corresponding to 8 transcriptionally active sites. Their conclusions were supported by the a priori known content (in terms of genes) of the genomic region. We empirically decided to set the parameters of the proposed method in order to force it to retrieve solutions comprising at most 20 clusters. This choice is not too restrictive because the algorithm return all the optimal solutions composed by a number of clusters ranging from 1 to 20 and is also motivated by the interest in the detection of potentially unannotated TSS.

3.2. Performance evaluation

In this test the method was evaluated only w.r.t. it's ability to detect experimentally supported TSS located in the considered genomic region. To this end we extracted the coordinates of the first elements of each clusters and we used them as centers of windows of 2τ size (τ nt upstream and τ nt downstream each center). We then evaluated the overlap of the identified genomic regions with existing TSS annotations. In the reported experiments we set $\tau = 750$ nt which leads to genomic windows of 1500 nt, a value compatible with the expected size of the proximal promoter regions, commonly accepted to be comprised between 1000 and 2000 nt. Each time a windows corresponding to a cluster was found to be in overlap with at least one gene TSS the prediction was counted as true positive. The genomic annotations containing the locations of the TSSs were obtained by mean of in house written MySQL queries to the public UCSC Genome Browser database (genome-mysql.cse.ucsc.edu, database: hg19). In order to provide a fair comparison of the performances achieved by MADAP and by the described method we used the annotations available at the moment of the publication of [12].

4. Results

In absence of the transcription level filter, the best solution, among the 20 available, was the one composed by 20 clusters. In this test, when using a τ value of 750 nt we found only 5 true positives and thus we failed to overcome the results produced by MADAP (8 true positive). In order to raise the sensitivity of the method to 8 true positives we were forced to increase the value of the τ parameter to 2000. This way we obtained the same predictions produced by MADAP (data not shown). Using the filtering policies reported in [12], the optimal solution obtaining the best performances was the one composed by 20 clusters. In this test we used a τ value of 750 and 14 out of 20 clusters were found to be true positives. Results are reported in Table.1.

These results are quite promising because we detected an amount of TSSs which is near twice the one obtained by MADAP in the same test. Despite these encouraging results, we are left with 6 false positives. In order to better characterize the obtained false positives we compared their genomic coordinates with updated genomic annotations obtained by direct SQL query to the UCSC database. This final step was performed by comparing the false positives against all the available annotations included in the UCSC genome database until February 2011. Results are reported in Figure.1.

Of the 6 false positives, the clusters 3 and 6 are located near to the TSS of the CDCA3 gene while the cluster 9 maps immediately downstream the TSS of the TPI1 gene. It was not possible to support the predictions associated with clusters 12 and 13 but the last false positive, cluster 19 is located very close to two microRNA annotations (MIR200C and MIR141). We thus found a total of 18 out of 20 supported predictions.

5. Conclusions

In this work we presented a novel 1-dimensional constrained clustering method. The proposed method was applied to a complex computational biology problem and was found to be more sensitive than previously published approaches in the identification of putative TSS. When applied on expression data preprocessed via standard filtering techniques, the proposed method was not only able to correctly predict the TSS of more genes than the compared MADAP method but also able to predict the location of the TSS of genes that were unknown at the moment of the experiment reported in [12]. All the collected results clearly demonstrated the potential of the proposed research line in the solution of complex computational biology problems.

References

[1] Aloise D., Deshpande A., Hansen P., Popat P. *NP-hardness of Euclidean sum-of-squares clustering*. Machine learning , 245-248, 2009.
[2] [13] Bradley, P.S., Bennett, K.P., Demiriz, A.: Constrained K-Means Clustering. Miscrosoft Research publication, MSR-TR-2000-65 (May 2000)
[3] Edwards A.W.F., Cavalli-Sforza L.L. A method for cluster analysis, *Biometrics* 21, 362-375, 1965.
[4] J. B. MacQueen. Some method for the classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Structures, pages 281-297, 1967.
[5] Mahajan, M.; Nimbhorkar, P.; Varadarajan, K. "The Planar k-Means Problem is NP-Hard". Lecture Notes in Computer Science 5431: 274-285, 2009.
[6] Novick B. Norm statistics and the complexity of clustering problems *Discrete Applied Mathematics*, v.157 n.8, 1831-1839, 2009.
[7] Saccà, F., Problemi di clustering con vincoli: algoritmi e complessità. Doctorate Ph.D. Tesis, University of Milan, 2010.
[8] Shunzi Zhu, Dingding Wang, Tao Li, Data Clustering with size constraints, Knowledge-Based Systems, Vol. 23 Issue 8, 2010.
[9] [18] Tung, A.K.H., Ng, R.T., Lakshmanan, L.V.S., Han, J.: Constraint-based clustering in large databases. In: Proc. of the 8th Intl. Conf. on Database Theory, pp. 405-419, 2001.
[10] Vattani, A., "'K-means requires exponentially many iterations even in the plane"'. Proceedings of the 25th Symposium on Computational Geometry (SoCG), 2009.
[11] T.H. Kim et al. A high-resolution map of active promoters in the human genome. *Nature*, 436:876–880, 2005.
[12] D.C. Shmid et al. MADAP, a flexible clustering tool for the interpretation of one-dimensional genome annotation data. *Nucleic Acids Research*, 35, W201–W205, 2007