MADAP

Mauro C. Delorenzi

March 29, 2007

1 Introduction. MADAPs specific aim.

MADAP is an annotation tool. Given a collection of experimental points that represent the 5' ends of cDNA clones believed to belong to the same gene, extract the likely position of true transcriptional start sites (TSS). The aim was to have a rapid and convenient way to process automatically vast amounts of data and obtain an annotation that corresponds well to the annotation, that a biological expert would propose for the same set of data. It is recommended, that an expert annotator validates the annotation produced by MADAP and tunes its parameters to improve annotation quality.

It takes as input one-dimensional data, a list of integer numbers representing positions, or a frequency table that summarizes such data. It outputs a proposed model for the position of the TSS. It does so by interpreting the positions as being due to a small number of TSS around which the data cluster. The positions are divided into clusters, a position for the TSS inside each cluster is proposed, clusters which are insufficiently supported by the data are excluded and clusters that are too close to each other to represent different TSS elicit a reassessment that typically results in the fusion of the two clusters but, depending on the parameters used, can also result in the exclusion of the one represented by less points.

The positions of the 5' ends of cDNA clones tend to form fairly tight clusters. A small portion of the points is typically far away from any other points. These islands of one or a few points can be experimental artifacts or can be due to a rarely used bona-fide TSS. This cannot be decided until there is more data. Such points cannot be used to propose the position of an experimentally well-supported TSS and are therefore operationally considered as background noise.

The user can specify a number of options and parameters when running MADAP. For the selection of the TSS, we found it helpful to use a set of restrictive constraints based on expert knowledge, which have a fair impact on the characteristics of the final model. For this reason, the development of a better defined and more complex probabilistic model for the data, or a more advanced criterion to choose the best data model were outside the scope of the project.

The default output consists in six lists of the selected model. There are three different maximization criteria (which tend to give the same result) and two different output formats for each of them.

2 Description

2.1 MADAPs data model selection.

MADAP builds, to a set of one-dimensional input data, a model for the position of clusters and cluster centers, which are sufficiently supported from the data and are compatible with users defined constraints. Data that do not fall into these clusters are considered to represent background noise. The probability of observing a position in a cluster of points is modeled by the superposition (mixture) of normal (Gaussian) distributions. Each cluster of points is modeled by a Gaussian distributions with a centre and a standard deviation. Such a model is called a mixture model. The standard deviation of the Gaussians can be imposed to be equal or even be numerically fixed by the user.

The number, the location of the centers and the relative frequency of the normal distributions are deduced from the data. The location and relative frequency of the normal distributions are parameters that are estimated from the data using a modified version of the standard EM algorithm for the fitting of a mixture of Gaussian distributions, which can be easily found in textbooks. The number of components that is clusters of data is decided by selection among a set of generated models.

MADAP differs from the standard algorithm, because of the addition of a background distribution and of a set of side conditions explained below. A drawback in the parameter fitting and model selection with a mixture of Gaussian components, is that isolated points that are far away from the cluster centers can distort the parameter fitting and have an undesirable high impact on on the model likelihood. This can disturb the selection of the model that best describes the clusters of points. For our application, the aim is to identify the clusters, while the isolated points are a disturbing factor that has to be controlled for. To do so, we add to the mixture a constant background probability for spurious isolated points or small islands of points in the form of a uniform distribution over the whole range of positions. This depends only on the data and is the same for all the models on the same data. The likelihood of the isolated point is dominated and stabilized by this non-gaussian ”background” component, reducing the negative influence of these points on the model selection. The relative frequency of the uniform background distribution is a user parameter.

The analysis can be summarized by the following steps:

1. The users chooses a set of values for the initial number of clusters n $(m <= n <= M)$

2. The program creates for each initial number of clusters n a few sets of initial models for the position of the centers of the clusters (heuristic model initialisation)

3. The model is evolved by fitting a mixture of n iteratively recalculated Gaussian components with initial centers obtained by the model initialisation. Iterations are performed with the EM algorithm until stabilization of the data likelihood or until a maximal number of iterations (Gaussian model fitting). A particularity is that also a constant background noise density is used in the E step of the EM algorithm, when points are probabilistically assigned to components.

4. The data model obtained in 3 is compared to the constraints. If the model does not satisfy the set of specified constraints, the number of components in the model is reduced. If there is at least one Gaussian components left, the steps 3 and 4 are repeated, otherwise the model is rejected. If the model satisfies constraints, it is stored and the program continues at 2 with the next initial model, until all initial models to all values of n are processed.

5. The final model is chosen among the stored models by taking the one with the highest data likelihood. Two variants are offered: the usual likelihood under a mixture model and a likelihood that takes calculated after attributing each point to the component with the highest density at that position.

The implementation of the constraints can reduce but not increase the number of components in the model. For this reason the algorithm starts with an assumed maximal number of initial components. It is sometimes useful to start with more components than the number of clusters expected in the data, because some clusters might be missed by the model initialisation that generates the initial values for the mixture model and the parameter fitting is unlikely to find clusters that were missed at that stage. Conversely, the elimination of superfluous components is reliable, but will increase the memory requirement and the running time of the program.

The model fitting always converges, because the number of model modifications that can be introduced by the constraints is finite and the EM algorithm converges to a local maximum of the likelihood, for a fixed the number of components.

There are two simple constraints currently implemented. One consists in requiring that a component has attributed to it a minimal number of data points, when every point is attributed to the distribution with the highest density in the position of the point. The second consists in a minimal distance between components peaks in the model (a peak here is the local modus of the distribution, not necessarily coinciding with the mean) . The algorithm checks the components in order of decreasing number of data points. A new component is accepted if it has enough data points and if it is not too close to one of the already accepted components.

Parameters of the Gaussian components are fitted with a standard unconstrained Expectation-Maximization (EM) algorithm except that the probabilistic attribution of points to components includes attribution to the background model, so that points far from any center have very low impact on the estimated mean and standard deviation of any component.

2.2 MADAPs heuristic, explanation

The likelihood of a data point under a Gaussian distribution is given by the product of the prior probability of the distribution component in the mixture and the density of the Gaussian at that location. It therefore increases with the density, a function of its distance to the center and the standard deviation of the component. This can be increased by adding a component to the mixture, which has higher density at the location of the point. The more components there are, the higher the average density that can be reached. But the likelihood of a point can increase also, when the number of components is reduced: Then the prior probabilities of the remaining components are higher. There is literature on the criteria to choose the best Gaussian mixture model for a set of data including selection of the number of components. In our case, experience has shown that model selection based on the simple principle of maximization of the data likelihood generated adequate solutions for the specific purpose of annotating collections of 5' ends of cDNA clones to reproduce the annotation done by a human expert.

The actual form of the parametric distribution used for the clusters does not have a strong impact, the Gaussian was chosen because of its universality and the simplicity with which a model can be fitted to the data. The core algorithm fits a normal mixture model to the data using the iterative expectation maximization (EM) algorithm, adjusted to the presence of the background distribution. This has the effect that points far from the centers are explained by the background distribution and therefore do not disturb the fitting of the Gaussian distributions. These points are then also interpreted as not supporting the presence of a bona-fide cluster of points and a TSS in their close environment.

The user can specify a number of options and parameters when running MADAP. For the selection of the TSS, we found it helpful to use a set of restrictive constraints based on expert knowledge, which have a fair impact on the characteristics of the final model. In this case, the algorithm is mainly just a rapid and convenient way to process vast amounts of data and obtain an automatic annotation that corresponds well to the annotation, that a biological expert would propose for the same set of data.

The same program could prove helpful also for another similar problems, although users might have to find a set of parameters that fits well their particular problem.

3 MADAP, some details

3.1 Data Likelihood

Let l and r be the start and end positions of the genomic sequence region, n_i the number of EST 5ends starting at sequence position i, K the number of Gaussian components (centers, clusters) of the model, m_j, σ_j, and π_j the mean, standard deviation, and mixing proportion of the model component j. Accordingly, the log-likelihood L of the data given the mixture model is:

$$log\ \ likelihood\ \ L = \sum_{l<=i<=r} n_i * log\left(bgrerrorprior * bgrdensity + \sum_{1<=j<=K} \pi_j * p_{ij} \right) \tag{1}$$

Note that this log-likelihood is insensitive to missing points for particular positions. The p_{ij} is the estimated probability for the interval (density at that position times a width of 1) under the Gaussian component j, the *bgrerrorprior* is set by the user, the sum over the Gaussian priors will be $(1 - bgrerrorprior)$ and the *bgrdensity* is uniform over the set of positions represented in the data, that is:

$$bgrdensity = \frac{bgrerrorprior}{numberofpositionsinthedata} \tag{2}$$

By identifying background noise model with a (K+1)th component, we can write more compactly:

$$log\ \ likelihood\ \ L = \sum_{l<=i<=r} n_i * log\left(\sum_{1<=j<=(K+1)} \pi_j * p_{ij}\right) \tag{3}$$

MADAP proposes two best models, which frequently coincide. A second selection is based on a variation of the log likelihood to be maximized, in which every data point is attributed to the distribution (any of the Gaussians or the background error distribution) with the highest density at its position. The internal sum of the log likelihood collapses to just one term (the prior probability is retained in the formula).

$$log\ \ likelihood\ \ L_{alt} = \sum_{l<=i<=r} n_i * log\left(\pi_{j(i)} * p_{ij(i)}\right) \tag{4}$$

where

$$j(i) = \underset{(1<=j<=K+1)}{\operatorname{argmax}} \quad p_{ij} \tag{5}$$

3.2 Model Initialisation

The initial values for the centers m_j are obtained by four different heuristic methods. At the end, the model with the highest data is retained, independent of the initialisation method that was used to generate it.

First, taking the appropriate number of peaks, that is ordering the single positions by their frequency in the data. Second, taking the appropriate number of peaks after having attributed progressively to the positions ordered by frequency also all the other counts inside a window of half-width specified by the user (option c, parameter fusionsdistance). The third and fourth method take the appropriate number of positions resulting from two variants of a progressive agglomerative clustering method based on the location of the points.

3.3 Typical specifications used for application to determine TSS

Only model components that explain at least 10 5' ends (input parameter n) and that have a minimal distance of 50 bp between the selected TSS inside the Gaussian component (input parameter p) are accepted. The search for the best data model alternates between a parameter estimation and a filtering step in which components that fail to fulfill the above-mentioned conditions are eliminated. A constant standard deviation of 20 (input parameter d) is used.

3.4 Program availability

The program is available under a open-source license from the FTP site of the Swiss Institute of Bioinformatics (ftp://ftp.isrec.isb-sib.ch/pub/software/unix/madap) and from the developer's website (http://www.isrec.isb-sib.ch/BCF/Delorenzi/Madap.html).

MADAP can also be accessed with user-supplied data on the MADAP server (http://www.isrec.isb-sib.ch/madap/). The web version allows for specification of the options and parameters that are most frequently of interest for users.