Methodology article

**Mixture models for analysis of melting temperature data**
Christoffer Nellåker*1, Fredrik Uhrzander2, Joanna Tyrcha2 and Håkan Karlsson1

Address: 1Department of Neuroscience, Karolinska Institutet, Retzius Väg 8 B2:5, 17177 Stockholm, Sweden and 2Mathematical Statistics, Stockholm University, Kräftriket Hus 6, 106 91 Stockholm, Sweden

Email: Christoffer Nellåker* - christoffer.nellaker@ki.se; Fredrik Uhrzander - fredrik.uhrzander@gmail.com; Joanna Tyrcha - joanna@math.su.se; Håkan Karlsson - hakkar@ki.se
* Corresponding author

**Abstract**

**Background:** In addition to their use in detecting undesired real-time PCR products, melting temperatures are useful for detecting variations in the desired target sequences. Methodological improvements in recent years allow the generation of high-resolution melting-temperature (Tm) data. However, there is currently no convention on how to statistically analyze such high-resolution Tm data.

**Results:** Mixture model analysis was applied to Tm data. Models were selected based on Akaike's information criterion. Mixture model analysis correctly identified categories in Tm data obtained for known plasmid targets. Using simulated data, we investigated the number of observations required for model construction. The precision of the reported mixing proportions from data fitted to a preconstructed model was also evaluated.

**Conclusion:** Mixture model analysis of Tm data allows the minimum number of different sequences in a set of amplicons and their relative frequencies to be determined. This approach allows Tm data to be analyzed, classified, and compared in an unbiased manner.

**Background**
Real-time PCR or semiquantitative PCR is widely used to detect and quantify specific target sequences. The exponential amplification of a sequence is monitored in real time by fluorescence. Commonly, a nonspecific fluorescent dye is used, such as SYBR Green I or LCGreen, which only reports the presence of double-stranded DNA. These dyes do not distinguish sequences and can thus report the amplification of undesired targets. Undesired sequences are normally detected during a dissociation step after thermocycling is complete. During dissociation, the double-stranded PCR products melt into single strands, so fluorescence is diminished. A curve can be produced by plotting the loss of fluorescence against a gradual increase in temperature. The temperature at which the rate of signal loss is the greatest can be defined as the melting temperature (Tm) of the PCR product. Although the Tm is sequence dependent, different sequences do not necessarily have different Tm. However, the converse is true. The detection of different Tm does imply the presence of different sequences. Therefore, by monitoring Tm, we can distinguish different targets for one set of primers. This technique has been used for the detection of single-nucleotide polymorphisms [1], allelic discrimination [2], and strain typing of microorganisms [3-5]. We previously reported the use of Tm analysis to detect the expression patterns of

Published: 11 September 2008
BMC Bioinformatics 2008, 9:370 doi:10.1186/1471-2105-9-370
Received: 5 March 2008
Accepted: 11 September 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/370

© 2008 Nellåker et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
transcripts containing different members of the W family of human endogenous retrovirus (HERV) elements in vitro and in vivo [6,7].

The precision of the Tm measurements determines the sensitivity with which different sequences can be distinguished. The instrument used to obtain the Tm recordings is the principal factor limiting the amount of information that can be extracted from the data. We recently reported a method that allows improved resolution, reduced spatial bias, and automated data collection for Tm detection in an ABI Prism 7000 Sequence Detection System (Applied Biosystems, Palo Alto, CA) [8]. Using a temperature indicator probe (Tmprobe) and an algorithm (GcTm) to interpolate more-precise Tm measurements from multiple data points, the standard deviation of the measurement error (σ) of the Tm recordings was improved from 0.19°C to 0.06°C [8].

However, there is no convention on how to analyze Tm data to objectively distinguish sequences by Tm. The need for such a tool becomes apparent when the Tm data are: i) not easily stratified because of overlapping clusters of Tm observations, and/or ii) if the number of different sequences and possible Tm categories are unknown. In this report, we use mixture model analysis to construct a model for a particular set of primer targets, to classify Tm data, and to calculate the mixing proportions of the amplicons within these categories. The mixture model technique allows Tm analysis to be applied to any set of primers to determine the minimum number of Tm categories (i.e., the number of different sequences detected) and the mixing proportions (frequency distributions) of the detected categories. Thus, mixture model analysis of Tm data is an objective method with which more refined Tm assays can be established.

Results

In a Tm analysis using the Tm probe and GcTm program, described previously [8], we demonstrated, using plasmids containing known sequences, that it was possible to distinguish some but not all sequences based on their Tm. In the present report, we applied the mixture models and the ρ established in the previous publication [8] to determine the Tm categories and mixing proportions of these data (Figure 1). Akaike’s information criterion (AIC), a measure of how well a model explains the data, with a penalty for the number of parameters estimated, determined that the Tm of the four sequences were best represented by a three-category mixture model. This model precisely estimated the mixing proportions of the Tm into the categories, attributing the correct number of Tm recordings to each of the four sequences (where two of them shared a category). For an overview of the procedure for using mixture models to analyze Tm data, see the Methods section.

Figure 1

Tm profiles for four individual plasmid targets with known sequences and their analysis in a mixed population by mixture model analysis. Box and whisker plots represent the Tm (median and range) of the four known sequences amplified separately. Dot plot shows the mixed population of the four Tm groups. Horizontal bar graph and Gaussian curve plot represent the grouped raw data of the Tm and the categories, respectively, with mixing proportions determined by mixture model analysis. The mixture models determined the mixing proportions of the three categories to be 15, 24, and 16, which exactly matched the proportions of the different sequences used.
We next assessed the performance of the mixture model analysis in constructing models for categories of $T_m$ with varying separations. Therefore, we generated simulated data points mimicking the $T_m$ of four sequences separated by multiples of $\sigma$. These data were used to identify the model that best explained the data according to AIC (see an example of the AIC plot in Figure 2) for a range of $T_m$ separations and numbers of data points (Figure 3). A large separation of $T_m$, $10 \times \sigma (0.6 ^\circ C)$, allow the mixture model analysis to close in on four separate categories with only 10 data points. Smaller separations of $T_m$ require larger numbers of data points to determine the correct number of $T_m$ categories. The distinction of categories with a separation of $1 \times \sigma$ required approximately 2000 data points to model the correct number of $T_m$ categories.

Next, we evaluated the fit of the data points to preestablished models. For this purpose, we generated data points corresponding to a sample containing three of four possible $T_m$ represented in a model. We compared the mixing proportions reported by the mixture model analysis with the mixing proportions in which all four $T_m$ were present at equal frequencies. In Figure 4, the $P$ values obtained from $\chi^2$ analyses for various separations of the $T_m$ are plotted against the numbers of data points used. The $P$ values for the $\chi^2$ test drop rapidly with increasing sample numbers for any $T_m$ separation of more than $1 \times \sigma$. With smaller separations of the $T_m$ categories, the mixture model analysis is unable to reliably establish the differences in the mixing proportions.

**Discussion**

We report the application of mixture models to the analysis of high-resolution $T_m$ data. Whereas the plasmid $T_m$ data reported are sufficiently separated to be stratified manually, we use these data to demonstrate the principle that can be applied to analyze more complex $T_m$ data.

Mixture model analysis of $T_m$ data entails the construction of a model based on the $T_m$ data for a set of primers. With such a model established, it is possible to fit smaller subsets of data to calculate the mixing proportions of the $T_m$ categories of the model. This gives a proxy marker for the frequency distributions of different amplicon sequences in the analyzed data. This approach requires no prior knowledge of how many different amplicons are present and there is no limit to the number of different $T_m$ that can be distinguished. However, the $T_m$ analysis method with mixture models only reports the minimum number of different sequences required to explain the $T_m$ data because different sequences can have the same $T_m$.

Mixture model analysis is a modern type of cluster analysis. The purpose of cluster analysis is to group data that have properties in common. When constructing the mixture model for a set of primers, the number of categories in the model that most appropriately explains the $T_m$ data is determined by AIC. Other information criteria exist, such as the Bayesian information criterion, but this penalizes free parameters more harshly than does the AIC.

By empirical testing with simulated data, we found that smaller separations of $T_m$ require exponentially larger numbers of data points to distinguish the correct number of categories in a mixture model. Insufficient numbers of observations yield an underestimation of the numbers of unique $T_m$ represented by the data, erring on the side of safety. In other words, with insufficient data, the number of unique sequences in the data is underestimated by the optimal model.

In an established model, based on a large number of $T_m$ observations, a smaller number of observations can be fitted to calculate the mixing proportions in the $T_m$ categories. These proportions can then be compared between sets of $T_m$ data as frequency distributions of sequences and analyzed with $\chi^2$ tests. We observed that, whereas a large number of $T_m$ observations are required to establish a model with a small separation between categories (e.g., 1000 data points are required with $2 \times \sigma$ separation), far fewer are sufficient for comparisons once the model is established (e.g., 100 data points for $P < 0.001$). A separation of the $T_m$ categories in the model of less than $1 \times \sigma$ results in unreliable mixing proportions. However, this should rarely be a problem in practice, because constructing the models puts a larger constraint on $T_m$ separation.
by AIC. In other words, models constructed with mixture model analysis will consist of $T_m$ categories separated by more than $1 \times \sigma$.

Not all dissociation curves are easily defined by a single $T_m$, as in the case of multiple domain transitions in longer sequences [9] (generally longer than those generated in real-time PCR assays) and for heterodimers. Using the GcTm approach to curve fitting and SYBR Green I chemistry, such melting profiles will be assigned a single $T_m$ value. Although some additional information is therefore lost, mixture model analysis still validly identifies clusters of $T_m$ and sequences. There is an established high-resolution amplicon melting analysis (usually denoted HRM) using LCGreen, primarily based on differences in the profiles of melting curves rather than on absolute $T_m$ [10]. Although this method is superior to mixture model analysis in identifying heterodimers, absolute $T_m$ values are required to identify homodimers. Recently, a method with sufficient resolution to distinguish base-pair neutral homozygotes was reported [11]. Mixture model analysis of $T_m$ can be used in all cases where the $T_m$ can be denoted as a single value, but primarily for homodimer discrimination.

**Conclusion**

In conclusion, the mixture model analysis of $T_m$ presented here allows the unbiased analysis of high-resolution $T_m$ data. This analysis is applicable to the identification of sequences in $T_m$ data regardless of the method by which the $T_m$ are acquired, provided the measurement error is known. Mixture models allow $T_m$ analyses to be performed on more complex and varied sequence targets than hitherto possible. Possible applications include typing microbial strains and their relative abundances in a population and the analysis of transcripts containing repetitive elements [3,4,6,12].

**Methods**

**Finite mixture models**

Mixture models are useful for describing complex populations with observed or unobserved heterogeneity. The
We apply the theory of finite mixture models to $T_m$ data consisting of normally distributed components in a mixture model, where each component has a standard deviation of $\sigma^2 \mathcal{C}$. The finite mixture density function is then as follows:

$$f(x | \psi) = \sum_{i=1}^{k} \pi_i \frac{1}{\sigma \sqrt{2\pi}} \exp \left( \frac{(x-\mu_i)^2}{2\sigma^2} \right)$$

where $\psi = (\pi_1, ..., \pi_k, \mu_1, ..., \mu_k, \sigma)^T$.

The likelihood function corresponding to the data $(x_1, ..., x_n)$ is as follows:

$$L(\psi; x_1, ..., x_n) = \prod_{j=1}^{n} f(x_j | \psi).$$

The logarithm of the likelihood function is

$$\ln L(\psi) = \sum_{j=1}^{n} \ln f(x_j | \psi).$$

We attempt to find the particular $\psi$ that maximizes the likelihood function. This maximization can be undertaken in the traditional way by differentiating $L(\psi; x)$ with respect to the components of $\psi$ and equating the derivatives to zero to give the likelihood equation:

$$\frac{\partial L(\psi)}{\partial \psi} = 0,$$

or equivalently

$$\frac{\partial \ln L(\psi)}{\partial \psi} = 0.$$

Quite often, the log likelihood function cannot be maximized analytically, i.e., the likelihood equation has no explicit solutions. In such cases, it is possible to compute the maximum likelihood of $\psi$ iteratively. To calculate maximum likelihood estimates, we use the expectation maximization (EM) method in combination with the Newton-Raphson algorithm. Iterations of the EM algorithm consist of two steps: the expectation step or the $E$-step and the maximization step or the $M$-step [13,14]. The Newton-Raphson algorithm for solving the likelihood equation approximates the gradient vector of the log likelihood function by a linear Taylor series expansion [15]. We use the Newton-Raphson algorithm in the $M$-step of the EM method.

We developed an algorithm that allows the automated estimation, in parallel, of a finite number of normally distributed components. The number of components can be assessed by several different methods, although none of them is optimal. We chose the AIC [16,17]. AIC is a relative score between different models where the selection of
the optimal model is made by considering the number of
data points and categories and the separation of the $T_m$
categories. AIC is defined as $-2\ln(L_m) + 2m$, where $L_m$ is
the maximized log likelihood and $m$ is the number of
parameters.

**Acquisition of HERV-W gag $T_m$**

$T_m$ data were generated with GcTm, as previously
described [8], on dissociation data obtained from the
amplification of plasmids containing known HERV-W gag
sequences.

Simulated $T_m$ data recordings and GcTm analysis were
performed in MATLAB™ (The MathWorks) version
7.0.1.24704 with the Optimization Toolbox. Mixture
model analysis was performed in R 2.6.0 [18] with the
MIX software [19,20].

**Overview of mixture model analysis of $T_m$**

A mixture model is constructed for a set of primers. The
model should be constructed on a large enough sample of
$T_m$ data to expect all possible sequences to be represented.
The $T_m$ data are then stratified into small-interval groups
and the frequency distributions of these arbitrary catego-
ries are used to construct and compare the mixture mod-
els. AIC is used to evaluate which model best explains the
data, while a minimum number of different categories is
used. Lower values of AIC indicate the preferred model,
i.e., the one with the fewest parameters. Once a model is
selected, $T_m$ data from different samples can be fitted to
the model and the mixing proportions compared between
differences. Between samples can be evaluated with $\chi^2$
tests if a conservative stance is taken, depending on the
separation between the $T_m$ categories and the num-
bers of data points.

**Abbreviations**

$T_m$: Melting temperature; AIC: Akaike’s information crite-
rian; HERV: human endogenous retrovirus; EM: expecta-
tion maximization.

**Authors’ contributions**

CN conceived the study, tested and prepared the manu-
script; FU developed the method and critically revised the
manuscript; JT developed the method and prepared the
manuscript; HK conceived the study and prepared the
manuscript.

**Acknowledgements**

This study was generously supported by the Stanley Medical Research Insti-
tute, Bethesda, MD, and the Swedish Research Council (21X-20047).

**References**

1. Germser S, Higuchi R: Single-tube genotyping without oligonu-
cleotide probes. Genome Res 1999, 9(1):72-78.
2. Graziano C, Giorgi M, Malentauchi C, Matiusz PL, Porfirio B:
Sequence diversity within the HA-1 gene as detected by
melting temperature assay without oligonucleotide probes.
BMC Med Genet 2005, 6:36.
3. Pham HM, Konnai S, Usui T, Chang KS, Murata S, Mase M, Ohashi K,
Onuma M: Rapid detection and differentiation of Newcastle
disease virus by real-time PCR with melting-curve analysis.
Arch Virol 2005, 150(12):2429-2438.
4. Waku-Kouomou D, Alla A, Blanquier B, Jeanet D, Caïd H, Rguig A,
Freymuth F, Wild FT: Genotyping measles virus by real-time
amplification refractory mutation system PCR represents a
rapid approach for measles outbreak investigations. J Clin
Microbiol 2006, 44(2):487-494.
5. Harasawa R, Mizusawa H, Fuji M, Yamamoto J, Mukai H, Uemori T,
Asaka K, Kato I: Rapid detection and differentiation of the
major Mycoplasma contaminants in cell cultures using real-
time PCR with SYBR Green I and melting curve analysis.
Microbiol Immunol 2005, 49(9):859-863.
6. Neßler C, Yao Y, Jones-Brando L, Mallet F, Yolken RH, Karlsson H:
Transactivation of elements in the human endogenous ret-
rovirus W family by viral infection. Retrovirology 2006, 3(1):44.
7. Yao Y, Schröder J, Neßler C, Bottmer C, Bachmann S, Yolken RH,
Karlsson H: Elevated levels of human endogenous retrovirus-
W transcripts in blood cells from patients with first episode
schizophrenia. Genes Brain Behav 2007, 7:103-112.
8. Neßler C, Wallgren U, Karlsson H: Molecular beacon-based
temperature control and automated analyses for improved
resolution of melting temperature analysis using SYBR Green
chemistry. Clin Chem 2007, 53(1):98-103.
9. Volker J, Blake RD, Delcourt SG, Breslauer KJ: High-resolution
calorimetric and optical melting profiles of DNA plasmids:
resolving contributions from intrinsic melting domains and
specifically designed inserts. Biopolymers 1999, 50(3):303-318.
10. Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ:
High-resolution genotyping by amplicon melting analysis
using LCGreen. Clin Chem 2003, 49(6 Pt 1):853-860.
11. Gundry CN, Dobrowolski SF, Martin YR, Robbins TC, Nay LM, Boyd
N, Coyne T, Wall MD, Wittwer CT, Teng DH: Base-pair neutral
hyomozygotes can be discriminated by calibrated high-reso-
lution melting of small amplicons. Nucleic Acids Res 2008,
36(10):3401-3408.
12. Slinger R, Bellföy D, Desjardins M, Chan F: High-resolution melt-
ing assay for the detection of gyrA mutations causing qui-
none resistance in Salmonella enterica serovars Typhi and
Paratyphi. Diagn Microbiol Infect Dis 2007, 57(4):455-458.
13. Dempster AP, Laird NM, Rubin DB: Maximum likelihood from
incomplete data via the EM algorithm. J Roy Stat Soc B 1977,
39(1):1-38.
14. McLachlan GJ, Krishnan T: The EM Algorithm and Extensions.
New York: Wiley; 1997.
15. Dennis JJE, Schnabel RB: Numerical Methods for Unconstrained
Optimization and Nonlinear Equations. New Jersey: Prentice
Hall; 1983.
16. Akaike H: A new look at the statistical model identification.
IEEE Trans Automat Control 1974, 19(6):716-723.
17. Akaike H. (ed.): Information Theory and an Extension of the
Maximum Likelihood Principle. Budapest: Akademiai Kiado;
1973.
18. Team RDC: R: A Language and Environment for Statistical
Computing. 2.6.0 edition. Vienna, Austria: R Foundation for
Statistical Computing; 2008.
19. Macdonald P: MIX Software for Mixture Distributions. 2.3th
edition. Ontario, Canada: Ichthus Data Systems; 1988.
20. Du J: Combined algorithms for fitting finite mixture distribu-
tions. In Masters thesis Hamilton, Ontario: McMaster University;
2002.