CutoffFinder - Cutoff optimization for diagnostic variables and correlation with clinical outcome and survival data

Jan Budczies, Frederick Klauschen, Wolfgang D Schmitt, Carsten Denkert
Institut für Pathologie, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin

Introduction

Unlike sequencing data, gene and protein expression data are usually represented by continuous or at least ordinal variables. Gene expression can be quantified using PCR and hybridization based methods, for example by single gene assays, such as TaqMan, or on global scale using microarrays. Gene expression data are often represented on log2 scale where an increment by 1 corresponds to a doubling of the gene expression level. In histopathological routine diagnostics, protein expression is usually evaluated by immunohistochemistry (IHC) and quantified on ordinal scale, for example using the percentage of stained cells, staining intensity or combinations thereof, such as the immunoreactive score (IRS). In order to translate a continuous diagnostic variable into a clinical decision, it is necessary to determine a cutoff point and to stratify patients into two groups each requiring a different kind of treatment. Determination of cutoff points can be done based on the distribution of the continuous or ordinal variable alone or by including information on clinical outcome or survival. Currently, no comprehensive and easy-to-use software is available for cutoff determination. Therefore, we present a bundle of optimization and visualization methods that can be accessed through a web-based interface. *CutoffFinder* implements three methods for cutoff optimization and enables users to study the correlation between cutoff-point selection and outcome or survival variables for an optimal patient stratification.

Material and Methods

CutoffFinder is implemented as Java Server Pages (JSPs) that connect to the statistical engine R using the package Rserve. All data analysis and visualization is performed with R (www.r-project.org). The software offers three different methods for cutoff determination: The first method fits a mixture model of two Gaussian distribution to the distribution of the variable using the R package flexmix. The optimal cutoff is the value of the variable where both probability density functions coincide. For the other two methods, several cutoff points are chosen and the patients are stratified into one stratum where the variable is above the cutoff and another stratum where the variable is below the cutoff. The second method correlates the variable under investigation with a binary outcome variable using logistic regression. The optimal cutoff is then defined as the point with the most significant (Fisher's exact test) split.

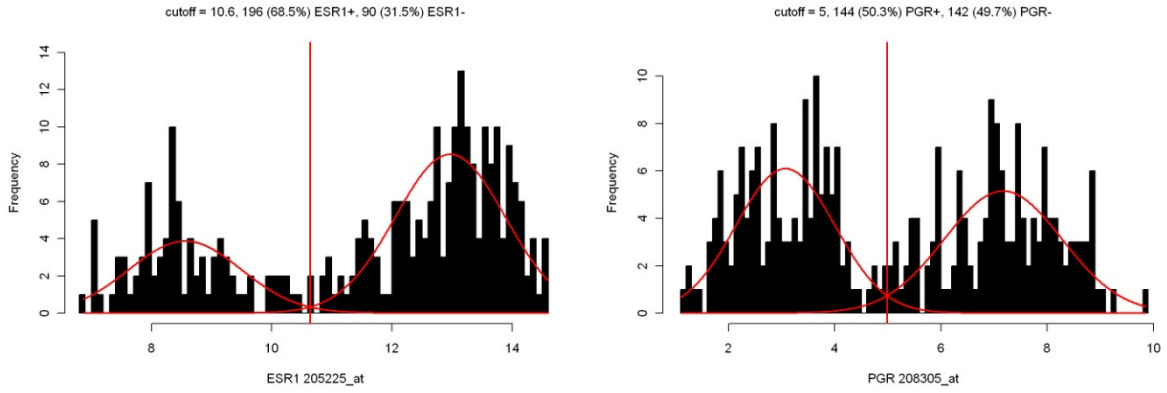

Figure 1: Distribution of ESR1 and PGR gene expression in 286 lymph-node negative breast cancers. A mixture model of two Gaussian distributions is fitted to each of the histograms (red lines). Vertical lines designate the optimal cutoffs derived from the distributions.

The third method fits Cox proportional hazard models using the implementation in the R package *survival*. The capabilities of *CutoffFinder* are demonstrated using gene expression data of estrogen receptor (ESR1) and progesterone receptor (PGR) from a publicly available microarray data set of 286 breast cancer samples (GSE2034 at www.ncbi.nlm.nih.gov/geo).

Results

Histograms of the distribution of estrogen (ESR1) and progesterone receptor (PGR) are shown in Fig. 1. Both distributions were clearly bimodal with optimal cutoffs at 10.6 (ESR1) and 5.0 (PGR). Immunohistology (IHC) is the gold standard for determination of hormone receptor status and IHC data were available for all 286 tumors of the microarray data set. Fig. 2A shows the odds ratio (OR) for correlation ERS1 expression with ER status in dependence of all possible cutoffs. The optimal cutoff was determined as 10.1 with OR = 67.8 (30.2 - 152.1). Fig. 2B shows the hazard ratio (HR) for correlation of PGR expression with distance-metastasis-free survival. The optimal cutoff was determined as 2.5 with HR = 0.46 (0.30 – 0.71).

Discussion

A limitation of the current implementation concerns the calculation of p-values and confidence intervals that are not corrected for multiple testing. A multiple testing problem occurs because several cutoff values are tested leading to an overestimation of the significance at the optimal cutoff. Correction of p-values and confidence intervals in this context is discussed in the literature [1, 2] and can be included in future versions of *CutoffFinder*. In summary, we presented a comprehensive and easy-to-use software for cutoff determination. Additionally to the data shown here, *CutoffFinder* produces a plot of the result for the optimal cutoff point: a plot of the correct and incorrect classifications in case of a binary outcome data and a Kaplan-Meier plot in case of survival data.

References

[1] Koziol, J.A. (1991). Biometrics **47**, 1557-1561.
[2] Bonetti, M. & Gelber, R.D. (2004). Biostatistics **5**, 465-481.

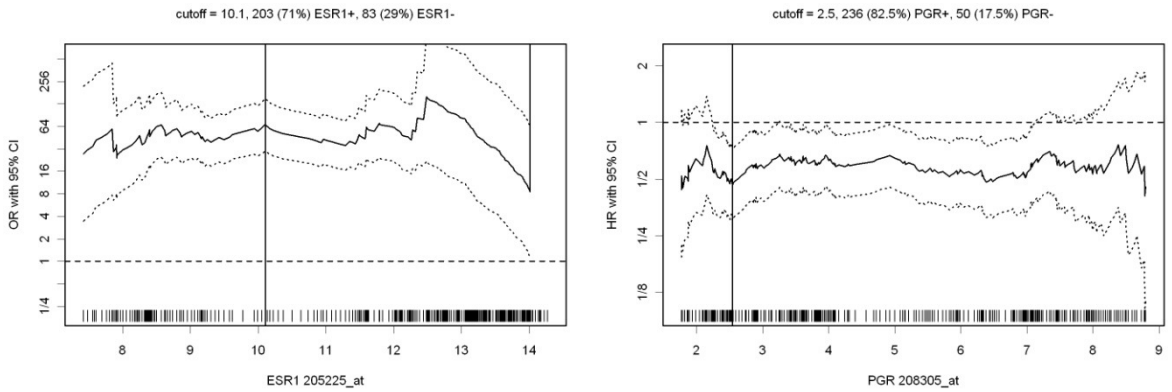

Figure 2: Cutoff determination by correlation with a binary variable or a survival variable. **Left panel:** For each possible cutoff, ESR1 gene expression is correlated with the immunohistological determinedestrogen receptor status. A horizontal line designates the cutoff with the most significant odds ratio (OR). **Right panel:** For each possible cutoff, PGR gene expression is correlated with distance metastasis free survival. A horizontal line designates the cutoff with the most significant hazard ratio (HR). The track at the bottom of the graphics shows the distribution of the gene expression values in the 286 tumors.