

Fuzzy Cluster and Rule Based Modelling to Classify Rheumatoid Arthritis and Osteoarthritis Patients Based on Gene Expression Data

Michael Pfaff¹, Dirk Woetzel¹, Dominik Driesch¹, Susanne Toepfer¹,
René Huber², Dirk Pohlers², Ute Staufenbiel², Dirk Koczan³,
Hans-Juergen Thiesen³, Reinhard Guthke⁴, Raimund W. Kinne²
¹BioControl Jena GmbH

Wildenbruchstr. 15, D-07745 Jena, Germany
Phone: +49-3641-527831, Fax: +49-3641-527832
email: biocontrol@t-online.de

²Experimental Rheumatology Unit, Department of Orthopaedics,
Friedrich Schiller University Jena, Germany

³Institute for Immunology, University of Rostock, Germany

⁴Leibniz Institute for Natural Product Research and Infection Biology –
Hans Knoell Institute, Jena, Germany

ABSTRACT: Rheumatoid Arthritis (RA) and Osteoarthritis (OA) both belong to the chronic rheumatic diseases. Whereas RA represents an inflammatory joint disease with an aggressive, joint destructive character (affecting approximately 1% of the population in industrialised countries), OA is a degenerative rheumatic disease with superimposed inflammatory flares. The purpose of the current investigation was to identify gene expression patterns to distinguish between RA and OA patients. Current medical diagnosis of the two diseases is primarily based on clinical data. In the future, this diagnosis may be supported by a gene expression based (automated) diagnosis. Such an approach may not only aid future differential diagnosis, but also provide clues to disease mechanisms and an improved therapy. The initial study described here focused on the application of fuzzy cluster and rule based modelling methods to reveal interrelations between gene expression in synovial tissue and the clinical diagnosis of RA and OA as well as of joint trauma (JT) as control. The original database was assembled at the Rudolf Elle Hospital Eisenberg, Germany, and covered 207 patients (80 RA, 107 OA, 13 JT and 7 other). Gene expression data were available for 28 of these patients (13 RA, 10 OA, 5 JT) and 22,283 Affymetrix® U133A gene fragments each. The modelling approach applied consists of five consecutive steps: data pre-processing, clustering, rule extraction, rulebase construction as well as cluster and rule based classification with validation. Data pre-processing included logarithm calculation and median based normalisation of the gene expression intensities as well as outlier detection and removal. Clustering of the data into two clusters (low and high expression of the respective gene) was performed applying a modified fuzzy c-means algorithm. Extraction of uni-conditional rules from the clustered data was carried out using a modified Kiendl relevance index to rate and rank the relevant rules with the conclusions RA, OA, JT. The highest ranked rules were used to construct a rulebase to classify patients with respect to the diseases/control. Rule extraction yielded three ranked rule lists consisting of 19, 7, 6 rules with the conclusion RA, OA, JT, respectively. Of the 32 genes contained in the conditional parts of these rules, 17 (53%) are considered by experts to be of potential pathogenetic relevance in rheumatic diseases. The selective and balanced use of the 6 top ranked rules each for RA, OA, JT in the rulebase for classification resulted in a modelling error of 0% and a generalised classification error after leave-one-out cross-validation of just 14% (4 errors out of 28). Following this initial study, additional gene expression data for a larger number of patients have to be analysed to further validate the obtained results and to arrive at gene expression based support systems for the differential diagnosis of rheumatic diseases in the future.

KEYWORDS: Gene Expression Data Analysis, Rheumatoid Arthritis, Osteoarthritis, Fuzzy Clustering, Fuzzy Rule Extraction, Classification