USING DATA MINING TO FIND MARKERS RELATED TO THE ROLE OF OSTEOCLASTS IN BONE AND JOINT DESTRUCTION IN RHEUMATOID ARTHRITIS

R. Rak[1], G. Boire[2], A.J. de Brum-Fernandes[2], S.J. Dixon[3], R. Harison[4], S.V. Komarova[5], M.F. Manolson[4], S.M. Sims[3], and L. Kurgan[1]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada; [2]Division of Rheumatology, Department of Medicine, Faculty of Medicine and Health Sciences, Sherbrooke University, Sherbrooke, QC, Canada; [3]Department of Physiology and Pharmacology, University of Western Ontario, London, ON, Canada; [4]Dental Research Institute, Faculty of Dentistry, University of Toronto, Toronto, ON, Canada; [5]Faculty of Dentistry, McGill University, Montreal, QC, Canada.

Aim of the study: To test the hypothesis, using data mining methods, that variations in the capacity to generate osteoclasts (OCs) in a normal population correlate with joint destruction in RA patients.

Methods: A cohort of patients satisfying the American College of Rheumatology (ACR) criteria for RA has been recruited from the outpatient rheumatology clinic and has been augmented by a control population. Data collected so far from the ongoing tests consists of various test results for 122 patients satisfying the RA criteria and 14 controls. In this preliminary report we selected 80 RA patients and 14 controls narrowing the number of attributes to 22, which are related to the analysis of peripheral blood mononuclear cells (PBMCs), namely osteoclastogenesis, OC resorptive activity, and OC survival, together with demographical and general health information.

Results: We performed a series of experiments trying a variety of *descriptive* data mining techniques such as production rules, decision trees, and decision lists, imposing five-fold cross validation to ensure that the generated model is not biased to a given pair of training and testing sets. The best results were obtained using the alternating decision tree learning algorithm with custom cost functions that was able to predict the presence of RA with 94% accuracy, 93% precision, and 100% recall. In contrast to other black-box predictors that could potentially give even more accurate predictions, the selected method generated a human-readable, easy to interpret model. The model, in the form of a small decision tree with 20 branches, allowed us to find several markers (features) that are related to presence of RA together with their associated confidence (strength). The model shows that the higher numbers of OCs and the lower values of OC apoptosis significantly contribute to the presence of RA, whereas the lower resorption values indicate lower risk of RA.

Conclusions: The preliminary data-mining experiments seem to confirm the hypothesis that (1) the osteoclastogenic ability of PBMCs correlate with joint destruction in RA, (2) decreased OC apoptosis contributes to enhanced OC activity and joint destruction, and (3) OC resorptive activity is enhanced in samples generated from PBMCs of patients with RA.