Meta-analysis of six Thyroid Tumor Microarray Datasets: a one-gene Classifier (SERPINA1) for Papillary Thyroid Carcinoma

Klemens Vierlinger[1], Martin Lauss[1], Christa Nöhammer[1], Klaus Kaserer[2], Bruno Niederle[3], and Friedrich Leisch[4]

[1]Austrian Research Centers GmbH - ARC, Division of Life Sciences, A-2444 Seibersdorf, Austria
[2]Department of Clinical Pathology, University of Vienna Medical School, A-1090 Vienna, Austria
[3]Section of Endocrine Surgery, Department of Surgery, Medical University of Vienna, A-1090 Vienna, Austria
[4]Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 München, Germany

June 13, 2007

Background: Thyroid nodules are endemic in iodine deficient areas, like Europes alpine regions, where they have a prevalence of 10-20 %. They are classified by their histology into the 2 benign types Struma nodosa (SN) and Follicular Thyroid Adenoma (FTA) and the malignant entities Follicular Thyroid Carcinoma (FTC), Papillary Thyroid Carcinoma (PTC), Medullary Thyroid Carcinoma (MTC) and Anaplastic Thyroid Carcinoma (ATC). Conventionally, discrimination between benign and malignant thyroid nodules is done by scintigraphy and fine needle aspiration followed by histology. Despite many advances in the diagnosis and therapy of thyroid nodules and thyroid cancer, these methods have a well known lack of specificity. Several DNA microarray based expression classifiers for the different clinically relevant thyroid tumor entities have been described over the past few years. However, reproducibility of these classifiers is generally low, mainly due to study biases, small sample sizes and the highly multivariate nature of microarrays.

Methods: Therefore we adopted a meta analysis approach for six publicly available microarray datasets on thyroid carcinoma. Four of those studies tackled the discrimination between papillary thyroid carcinoma and non-carcinoma tissue and two studies tackled the distinction between benign and malignant follicular thyroid disease. Merging of these datasets yielded a matrix of 146 samples on 5757 shared genes. Study bias was removed using DWD and classification / crossvalidation was performed using PAM.

Results: After Data-integration to remove study specific biases, we identified a one-gene classifier (SERPINA1) for papillary thyroid carcinoma and a 110-gene classifier for follicular thyroid disease. Classification of papillary thyroid disease was achieved with 99% accuracy (in leave-one-out crossvalidation), follicular disease was correctly identified in 87% of cases (in loocv). Identification of papillary thyroid disease was further validated by rigorous study-crossvalidation, which gave similar results (97.9% weighted average accuracy). Removing the SERPINA1-gene from the dataset yielded a 9-gene classifier with a similar performance. Removing these 9 genes still doesn't hamper classification.

Conclusions: The presented work represents the largest cohort of thyroid carcinoma microarray data analysed to date. It makes use of the latest methods for microarray data integration and classification. The results indicate a huge potential for future diagnostic applications. We also show that there is a large abundance of molecular markers for papillary thyroid carcinoma. Unfortunately, as yet the number of follicular datasets is low and to our knowledge a comprehensive molecular analysis of medullary and anaplastic thyroid carcinoma is still largely missing. The approach demonstrated here will also be applicable to multiclass classification once more data becomes available for all thyroid tumor entities.