Convergence of the proteomic pattern in cancer

Ute Müller, Günther Ernst, Christian Melle, Reinhard Guthke and Ferdinand von Eggeling

1Core Unit Chip Application (CUCA), Institute of Human Genetics and Anthropology, Friedrich-Schiller-University, 07740 Jena, Germany, 2aura optik gmbh, Wildenbruchstrasse 15, 07745 Jena, Germany and 3Leibniz Institute for Natural Product Research and Infection Biology, Hans Knoell Institute (HKI), 07745 Jena, Germany

Received on January 13, 2006; revised on February 10, 2006; accepted on February 27, 2006
Advance Access publication March 7, 2006
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: On the histological level the differentiation of normal epithelial tissues is well known. The phenomenon of dedifferentiation, which occurs as cells develop towards malignancy is also well described. To identify an epithelial tumor-specific proteomic profile as well as to measure the proximities between we used data from tumor tissue and adjacent normal tissue microdissected from head and neck and colon cancer samples which were analyzed using ProteinChip technology and performed a bioinformatic meta-analysis on the resulting four complex datasets.

Results: All four groups could be identified based on their proteomic signatures and the tumor tissues were found to be more similar to one another than to the normal epithelial tissue from which they progressed. This study shows at the proteomic level that changes in the histological features of tumors as compared to the tissues from which they arise are reflected in the convergence of proteomic pattern during the development to cancer.

Contact: leegg@mti.uni-jena.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

An epithelium is an assembly of polarized cells with defined apical and basolateral domains that line the inner and outer organ surfaces. There are many types of epidermis specialized to carry out functions, including protection, secretion, nutrient resorption and polarized transport between tissue compartments. The morphological reorientation of cellular differentiation during the progression to cancer reflects the functional properties of malignant cells as increased proliferation, invasion and metastasis. However, their way to ‘simplicity’ was not revealed for the expression of genes and proteins.

The ProteinChip technology (surface-enhanced laser desorption/ionization mass spectrometry; SELDI) which utilizes chromatographic surfaces able to retain proteins depending on their physicochemical properties followed by direct analysis via time of flight mass spectrometry (von Eggeling et al., 2001; Hutchens and Yip, 1993) has been predominantly used to analyze bodily fluids including serum (Zhang et al., 2004), urine (Vlahou et al., 2001), nipple aspirates (Paweletz et al., 2001) and pancreatic juice (Rosty et al., 2002). We employed microdissected tumor tissue free of adjacent, non-malignant tissue for ProteinChip analysis to improve the chances of identifying reliable biomarkers for cancer diagnostics (Melle et al., 2005; Melle et al., 2004a). The principal task of studies utilizing ProteinChip technology to date has been tumor marker identification, and not the general analysis of the differences in the tumor and normal tissue proteomes.

From histological prospective, epithelial cells appear to undergo a dedifferentiation process during their progression towards malignant neoplasias. These changes should also be apparent at the biochemical level of the tissue proteome. Bioinformatic analysis methods that are capable of analyzing and comparing the entire proteomes of tumor and normal tissue on a global level should make it possible to identify a potential tumor-specific proteomic signature as well as to show the convergence to a more common protein signature (Fig. 1).

We combined the proteomic profiles generated in previous studies of microdissected tissue areas from head and neck cancers and the adjacent normal pharyngeal epithelium with the proteomic profiles from non-hereditary colorectal cancer and the adjacent normal colonic epithelium in a meta-analysis. We show that these groups can be separated by specific peaks or signatures, and that the two different cancer types are more similar to each other than to the epithelia from which they developed.

2 METHODS

2.1 Tumor specimens and analysis of microdissected tissue by ProteinChip arrays

Tumor samples were obtained after surgical resection in the Department of General and Visceral Surgery and the ENT Department of the Friedrich Schiller University in Jena. Staging and grading information can be found in former studies (Melle et al., 2005, 2004a, 2003). Laser microdissection of normal tumor epithelium was performed with a laser microdissection and pressure catapulting microscope (LMPC; Palm, Bernried, Germany) (Melle et al., 2005, 2003). Protein lysates were prepared from microdissected tissues, and lysates were analyzed on strong anion exchange arrays (SAX2; Ciphergen Biosystems Inc, Fremont, CA) as described (Melle et al., 2003). Mass analysis was performed in a ProteinChip Reader (model PBS IIc, Ciphergen Biosystems Inc, Fremont, CA). Spectra with at least 10 signals in the range between 2 and 20 kDa exhibiting a signal-to-noise ratio (S/N) of at least 5 were selected and exported for further analysis.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
3 RESULTS

3.1 Classification and search for characteristic features

Classification was performed using different methods for six pairs of sample groups which are listed in Supplementary Table 1. As results, the following information was derived: (1) features distinguishing the two groups (classifiers), their number as well as their significance and (2) quality of class prediction/error rate.

3.1.1 t-test

The number of proteins whose mean expression data averaged over two groups of tissues differ significantly (Supplementary Table 1). Many proteins, i.e. 61 with a significance level of 5% and 33 with a significance of 0.01% of the 172 investigated proteins are differentially expressed in tissues head and neck versus colon. However, the number of proteins which are differentially expressed in normal versus tumor cells are only few. Proteins identified by t-test on the significance level 5% are shown in Supplementary Table 2. Supplementary Figure 2 (Supplementary Data) illustrates that the mean values of expression signals for the protein at 9645 Da is different, averaged over normal and tumor samples.

3.1.2 Decision trees

Also here all the six comparision were carried out. For the comparison NCOL–HNH, the most clear decision tree was found: there was just one split point corresponding to the peak at 10 848 Da, which was identified in a former study as calgranulin A (Melle et al., 2004a). Here intensities smaller than 1 exclusively belong to the group NCOL and such bigger one to HNH. For the other five comparisons more complex trees were found, indicating that these pairs are less different.

3.1.3 Random forest classification

The RF for the whole dataset (four classes at once) gave a classification with an error rate close to 20%. This result was attained after parameter optimization. The confusion matrix (Supplementary Table 3) shows for each tissue class the number of samples and the class labels RF assigned to them. The tissue type [colon or head and neck (pharynx)] was almost correctly assigned for both N-groups, except from two cases where TuCOL samples are classified as TuHN. On the contrary, frequent confusion occurred between normal and tumor samples of the same tissue type. From the above classification, the importance of the features making up the forest was derived. These calculations yielded about 50 features which were more important than the background. The 20 most important peaks are listed in Supplementary Table 4 in order of decreasing importance. Among all these classifiers, it is worth noting two outstanding features, the peaks corresponding to and at 13 245.3 Da and 10 848.5 Da are clearly more important than the rest. The importance value of the latter is not so apart from the others, however, this feature has special properties: m/z 10 848.5 was found to play an important role in the identification of each of the four classes. After that, RF for two classes was run. RF performed a perfect classification of normal tissue samples (NCOL versus HNH). In contrast, in all other cases (TuCOL versus TuHN, NCOL versus TuCOL) misclassification of samples occurred (Supplementary Table 5).
3.1.4 Support vector machines SVM were used to calculate the prediction accuracy as a quotient $Q$ dividing the number of true predicted observations by the total number of tests. A total of 50 pairs of proteins were found whose expression signals predict the samples NCOL versus NHN with the prediction accuracy of 100% (Supplementary Table 6). The expression signals of one of these pairs are shown in Supplementary Figure 3. In opposition to the error free prediction of colon versus head and neck for normal cells (NCOL versus NHN) the prediction accuracy is $<94\%$ for tumor cells (TuCOL versus TuHN) (Supplementary Table 6). A total of eight protein pairs were found that allow a prediction accuracy $>90\%$ but $<94\%$. The pairs of proteins whose expression signals allow an accuracy $>85\%$ for the prediction TuCOL versus NCOL are shown in Supplementary Table 7. The prediction accuracy for tumor versus normal head and neck tissue is $<85\%$. A total of 19 pairs of proteins were found whose expression signals allow a prediction accuracy between 80 and 85%, no pair exists with a prediction accuracy $>85\%$ (Supplementary Table 6).

3.2 Similarity measures

3.2.1 RF-proximities The average proximities were derived for the same six pairs of sample groups that were compared in the classifications (Supplementary Figure 4). Sorted by magnitude, the proximity values increase almost evenly from 0.06 to 0.16. As a conclusion, it can be stated that there is higher proximity between normal and tumor samples (from one tissue type as well as from both tissue types together) than between samples of different tissue types. It is noted that the lowest proximity is the one between NCOL and NHN, these two classes are the most dissimilar classes.

3.2.2 Mean square distances for classifiers found with random forest For the most important peaks from the RF classification NCOL–NHN, the mean square distances for the six pairs of sample groups were calculated. For the peak at 10,848.5 Da, the result is shown in Figure 2. The mean square distance between colon, head and neck tissue is significantly larger in the normal state than in the tumor state. This should be taken into account because it is not an isolated fact considering 10 of the most important peaks, in 6 cases, the COL–HN distances are larger in the normal state than in the tumor state. The difference between the two values is mainly not so striking as in the above example. However, these data allow to ascertain a tendency. From the overall RF classification, the 20 most important features (listed in Supplementary Table 8) were investigated. For nine among them, mean square distances between COL and HN behave according to the tendency stated in Supplementary Figure 4 (data not shown).

4 DISCUSSION

The combination of different chip-based genomic and proteomic techniques with bioinformatic analysis methods is indispensable for the search of tumor-specific signatures. In addition, the relatedness of different tissues to the tumors arising from these tissues is important to completely understand the biology of cancer. Expression studies have been carried out for many different tumor types. Most of these studies compared homogenates of whole tumor biopsies that contain enough cells of the normal tissue in order to deliver a mixed tumor and normal signature. This provides problems with the analysis, because the tumor signature is not clearly definable.

Two basic questions arose during the course of our studies of different microdissected epithelial tumor entities using ProteinChip technology (Melle et al., 2005, 2004a,b, 2003; von Eggeling et al., 2000). The central question whether it is possible to identify a single set of individual peaks differentiating between normal and tumor tissues has been affirmatively answered in several studies by our own group as well as others (Melle et al., 2005; Zhang et al., 2004; Wilson et al., 2004; Melle et al., 2004a,b). We became curious as to whether a common signature for malignancies arising from various epithelial tissues could be identified. Pathologists have long described that during tumorogenesis a morphological dedifferentiation in comparison with the tissue of tumor origin takes place. However, this process of dedifferentiation has only partly been shown on a molecular level. Proteomic profiling using purely defined tissue material presents an ideal method to show the molecular equivalents of this process as well as to investigate what changes dedifferentiation entails. To approach these questions, we analyzed the data produced from proteomic profiling of microdissected normal and tumor tissues from head and neck and colon cancers using two different bioinformatic approaches.

The different bioinformatic algorithms applied for the meta-analysis of the proteomic profiles of the four tissue groups focused on two different questions. First, to identify characteristic features differentiating between the four tissue groups, and second, to analyze how similar the groups were to one another. Classification of the group using characteristic features was possible using all the bioinformatic methods applied to a certain extent. The comparison NCOL–NHN was the only case where sample classes could be predicted perfectly, whereas in all other comparisons classification errors occurred. We conclude, that these two groups appear to be the most different from one another. Similar features can be found in all the resulting peak lists generated using different methods, indicating that the patterns generated using different methods were quite similar (see Supplementary Tables 2, 4, 7 and 8). That these features were identified by more than one method underlines the significance of these classifiers. The similarity measures calculated in this study agree with the interpretation of the group classification results. Both proximities as well as mean square distances between groups revealed that the tumorous tissues (COL and HN) were more similar than the normal tissues from which they were derived. This is
especially true for the peak at 10 848 Da, which could be identified as a classifier in a former study (Melle et al., 2004a). The results from both bioinformatic approaches concluded that the differences between normal epithelial tissues are larger than between the tumor tissues arising from these epithelial tissues. These results support the hypothesis that epithelial tissues become more similar during their progression towards a tumors state at the proteomic level (Fig. 1). This conclusion is consistent with the biological features of these tissues. The epithelia from colon and pharynx develop namely from the endoderm, but are functionally reflected in their proteomic distance. The hypo- and mesopharynx is lined by a multi-layered, squamous epithelium that assures a protective function. The colon, however, is lined by a prismatic, single-layered epithelium with predominantly secretory (mucine) und absorptive (water) capabilities in addition to a function for the excretion of heavy metals. In contrast, TuCOL and TuNH display a more closely related common tumor signature that is recognized in bioinformatic analysis through the closer similarity or proximity of these groups. For the tumor, functions such as proliferation, invasion and metastasis become more important for tumor survival than the maintenance of the normal epithelial cell functions. Therefore, structural and functional dedifferentiation from the functionality of the tissue of origin is the result also reflected in the histological appearance of the tumor tissue. Interestingly the dedifferentiation is also known in the histopathological grading which forms the basis for the classification of tumors. Additionally, it could be shown that common signatures are present for the two normal epithelial tissues and for the two tumor tissues. Specific classifiers could be identified using both bioinformatics methods employed here to distinguish among all four tissue groups. For the groups with the lowest proximity (NHN versus NCOL) only one peak (10.8 kDa) is necessary, but for groups with a high proximity the classifier becomes more complex. Especially for the classification of clinically relevant groups (e.g. NHN versus TuHN), it becomes clear that a reliable diagnosis can only be achieved using a multi-marker classifier or a specific signature of expressed genes or proteins.

Our analyses support the notion that the identification of a general tumor signature including similarities indicating features, such as malignancy, will be feasible in future. This proteomic signature may further contribute, especially in combination with additional methods, mainly an optimized tissue microdissection and analysis [e.g. Ernst et al. (2006)], to reveal the heterogeneity of the tumor and to identify metastasizing clones inside the malignant tissue.

In conclusion, we provide proof-of-principle at the proteomic level that changes in the histological features of tumors as compared with the tissues from which they arise are reflected in the convergence of proteomic pattern during the development to cancer.

ACKNOWLEDGEMENTS

The authors would like to thank Kathy Astrahantseff and Susanne Michel for stimulating discussions and critical reading of the manuscript. This work was supported by the IZKF Jena and the BMBF (NBL3 and FKZ 0313652A).

Conflict of Interest: none declared.

REFERENCES