Distance-based profiling aids in evaluation of ageing-related phenomena

Lei Wang[1], Tiange Cui[1], Boris Veytsman[1], Alexey Moskalev[1, 2, 4, 5], Ancha Baranova[1,2,3]

[1] *School of Systems Biology, George Mason University, 4400 University drive, Fairfax, VA USA 22003*
abaranov@gmu.edu
[2] *Moscow Institute of Physics and Technology, 9 Instituitsky per., Dolgoprudny, Moscow Region, 141700, Russia*
[3] *Research Centre for Medical Genetics (RCMG) of RAMS, 1 Moskvorechie str, Moscow, Russia*
[4] *Institute of Biology, Komi Science Center of RAS, 28 Kommunisticheskaya st., Syktyvkar, 167982, Russia*
[5] *Syktyvkar State University, 55 Oktyabrsky ave., Syktyvkar, 167001, Russia*

In typical biological assay performed in a high-throughput mode, either expression levels for individual genes or other quantifiable variables are assessed in parallel. These variables could be represented as dimensions of the information space that we study. In high dimensional space, the data become sparse. In other words, when a data set contains a large number of attributes, we are faced with a choice of either completely suppressing most of the data or losing the desired level of statistical significance for any possible finding. The problem outlined above is known as the "curse of dimensionality" [1]. There is a need to develop integrative approaches, capable of combining data from multiple high-throughput experiments to increase sample size [2, 3] or statistically sound and robust techniques to reduce the data to the most informative features.

In our previous studies, we developed a novel approach based on the "distances" in the multidimensional space of gene expression values. As a proof-of-principle, we showed that this approach produces surprisingly good results in separation of normal and affected samples both for analysis of human malignancies and for chronic progressive conditions like psoriasis [4]. In current work, we applied distance-based metrics to the problem of quantification of ageing and age-related phenotypes.

Aging has been an intriguing field of study for biologists for decades. As cells experience stress and damages from internal and external factors, they normally progress toward cellular senescence at which point they cease to replicate, but acquire pro-inflammatory features. This process comes with significant changes in gene expression profile (GEP) of the cell.

Here we performed a systematic classification of gene expression profiles from 12

microarray dataset. Samples from multiple disorders and healthy controls that were taken from various tissues were included. The array data were grouped and analyzed by the age of the donor. Pearson and Kolmogorov-Smirnov and correlation coefficient were used to compare GEPs between different groups. In such way, we built a holistic marker taking into account the quantifiable expression levels of all genes assayed, rather than extracting top ranked features as markers. In our analysis, the cumulative gene expression pattern of an individual patient is considered as a whole and is represented as a data point in a multidimensional space formed by all gene expression features assayed in the given system. The degree of separation between samples indicates the drift of the testing samples away from the cellular stable state in the process of cellular senescence. The classifiers showed clear separation between different age groups, as verified by k-fold cross validation. The holistic marker was further compared with specific markers extracted based on the ranking of statistical significance. The performance of the classifiers was evaluated by receiver operating characteristic curve (ROC curve).

As an example of analysis, here we show linear distance plots for datasets GSE13330. In respective experiment, human foreskin BJ fibroblasts were mock or Bleomycin sulfate-treated (100ug/ml, Sigma, St. Louis, MO) for 24 hrs, while replicatively senescent fibroblasts were obtained by continuous passage. After 72 hr serum-starvation, RNA was collected and biotinylated cRNA was hybridized to Affymetrix Human Genome U133 Plus 2.0 GeneChips (Affymetrix, Santa Clara, CA) in the Washington University Microarray Facility [5]. There were 4, 6, and 6 samples for Stress-Induced Prematurely Senescent (SIPS), Replicative Senescent (RS), and Young respectively.

Our distance-based marker demonstrates the predictive power of global signatures is as good as specific markers, yet with better robustness and reproducibility. The classifiers may be used to identify the aging status of tissues and verify whether disease-based aging models resemble normal aging process.

We are grateful to Dr. Ganiraju Manyam (MD Anderson Cancer Center, TX, USA) and Prof. Alessandro Giuliani (Istituto Superiori de Sanita, Italy) for the discussions that greatly contributed to initial stages of the development of the holistic analysis of gene expression and to the concept of distance metric. This project was partially supported by "Human Proteome" program of the Ministry of Education and Science of the Russian Federation.

1. Bellman RE (1961) Adaptive Control Processes. In A Guided Tour. Princeton University Press, Princeton, NJ.

2. Waldron L, Coller HA, Huttenhower C (2102) **Integrative approaches for microarray data analysis.** *Methods Mol Biol*, **802:**157-182.

3. Michiels S, Kramar A, Koscielny S (2011): **Multidimensionality of microarrays: Statistical challenges and (im)possible solutions.** *Mol Oncol*, **5**(2)**:**190-196.

4. Veytsman B, Wang L, Cui T, Bruskin S, Baranova A (2014) Distance-based classifiers as potential diagnostic and prediction tools for human diseases. *BMC Genomics*, 15 Suppl 12:S10.

5. Pazolli E, Luo X, Brehm S, Carbery K, Chung JJ, Prior JL, Doherty J, Demehri S, Salavaggione L, Piwnica-Worms D, Stewart SA (2009) Senescent stromal-derived osteopontin promotes preneoplastic cell growth. *Cancer Res*, 69(3):1230-9.