# PROTEIN FINGERPRINTS PAVE THE WAY FOR MICROARRAY COMPARISON

Roger Mulet[§] · Albert Pujol[§‡] · Teresa Sardón[§] · Judith Farrés[§] · José Manuel Mas[§]

[§] *Anaxomics Biotech, c/Balmes 89, 08008 Barcelona, Spain.* [‡] *Institute for Research in Biomedicine and Barcelona Supercomputing Center, c/Baldiri i Reixac 10-12, 08028 Barcelona, Spain.*

ANAXOMICS
the pathway to discovery

*Anaxomics has developed a novel method of analysis for microarray data based on the generation and subsequent comparison of 2D fingerprints for each sample. Microarray results are incorporated into a protein interaction network and embedded with all sorts of biological and medical data. Multidimensional scaling is use to reduce the multidimensional network into a 2D map, generating a specific projection "fingerprint". A repertory of cell-type specific 2D fingerprints based on previous experiments stored in GEO database has been generated. The cell-type fingerprints obtained can be envisaged as the average population. Thus, in addition to objectively quantifying the overall differences between samples, this revolutionary approach can measure the relative distance between any microarray sample and the cell type average.*

## DIFFICULTIES IN MICROARRAY TRANSCRIPTION PROFILES COMPARISON

Microarrays, with their unique ability to create a global picture of gene expression in a cell or a tissue, are a valuable tool. In recent years, manufacturing trends of miniaturization and increased throughput have fostered the development of arrays with higher density, opening the door to the simultaneous monitoring of thousands of genes [1].

The Gene Expression Omnibus (GEO) is a public repository [2] that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. As of today, this database contains more than 500,000 samples from human specimens, and the number of microarray experiments keeps growing at a near-exponential rate [1]. Such a massive amount of information, if correctly managed, might provide a deep insight on the behaviour of biological systems and, hence, help in the creation of virtual models that simulate them. What is more, the sheer number of samples and donors allows us to assume that conclusions extracted from them are, on average, applicable to general population. Regardless of the nature of the study, comparison between samples is at the root of every microarray experiment.

Conventionally, sample groups (control and intervention) are composed of individuals (be them cells, tissues or whole organisms) that should be in the average within their category.

### How do you compare two groupsin a gene-expression assay?

The answer appears to be deceptively simple; the group displaying higher fluorescent intensity for a given gene probe has higher expression, so it is only a matter of confronting every spot of the array. However, the actual procedure is trickier than it looks, when assessing microarrays within the same group discrepancies in expression arise. So there is intra-group variation which significantly hinders interpretation. The solution usually consists of calculating the average for each group, thereby suppressing fluctuations, though at the risk of losing information. Without a reference to guide our adjustment, we may incur in a serious mistake. The quest is even more difficult when trying to compare different microarray experiments.

Classical approaches are fine when the goal of the study is to compare the expression of previously defined sets of genes, but they fail at providing a global account of gene expression in the sample. Performing a comparison that encompasses the entire transcriptional profile would require evaluating the distance between all possible combinations of points in the array, entailing a colossal computational effort. Note that spotted microarrays ordinarily encompass thousands of spots, while some of the more advanced chips in the market feature over a million spots.

# SYSTEMS BIOLOGY

## THE MOST APPROPRIATE APPROACH FOR EXPLOITING MICROARRAY DATA

In the recent years, **systems biology** has arisen as a new discipline that intends to understand cell complexity by considering the available knowledge as a global system where the specific elements are less important than the global relationship between them [3]. Systems biology aims to represent an organism, regardless of its complexity, as a map or graph of interconnected proteins where nodes are proteins and edges are the relations between them. Using this approach it is possible to identify key points over the map that eventually become key proteins in cells [4,5], clarify the mechanisms of action [6,7] and conduct drug repositioning [8,9].

**Anaxomics** has developed a new method for microarray analysis based on the holistic view of living organisms that only systems biology is able to provide [10]. This novel approach highlights the biologically and clinically relevant information that is often hidden in the immensity of gene expression data. Thus, it is particularly appealing when the investigator deals with proteins or pathways that are not familiar or when he simply wants to discover something new.As a part of this mechanistically-driven analysis, we offer an innovative tool for microarray slides comparison that avoids the intrinsic shortcomings of dot-to-dot analysis. Anaxomics has developed a novel method of analysis for microarray data that can quantify the overall differences between groups. In addition, this revolutionary approach can measure the distance relative to previous experiments stored in Gene Expression Omnibus (GEO) database, which can be envisaged as the average population.

### Converting complex protein maps into 2D pictures

The maps or graphs attempting to represent all the myriad of proteins and interactions in human physiology face a daunting challenge. A graph is a mathematical function where every variable corresponds to a link between nodes, so the pattern of connections of each protein to the rest of proteins can be seen as an “n” dimensional space with as many dimensions as number of links [11]. With tens of thousands of protein-protein interactions already characterized, graphs depicting biological networks are difficult to interpret and require intensive calculations. One way of overcoming this difficulty is the application of dimensionality reduction techniques that can provide structure-preserving mappings of the data into lower-dimensional spaces. The resulting 2D images are easy to visualize and may uncover relevant conclusions that remain hidden in the abstruse complexity of the intricate protein network. Importantly, since the topology of the protein in map is conserved during this transformation, it is possible to travel back and forward without losing valuable information (Figure 1). Anaxomics uses for the generation of cell type fingerprints **Multidimensional Scaling** (MDS), a dimensionality reduction procedure to visualize high-dimensional data in 2D [8]. Multidimensional scaling refers to a set of statistical techniques for data visualization and exploratory data analysis that interactively reduce n dimensions of the map into two. The 2D version obtained after the MDS analysis preserves the notion of “nearness” and therefore locates together those proteins that in the initial network are also near i.e., they have less number of nodes between each other. The dimensional reduction of the data allows the application of **clustering analysis**, which leads to the identification of regions enriched in proteins involved in certain biological processes when compared with the rest of the network or proteins differentially expressed when compared with the rest of the microarray data.

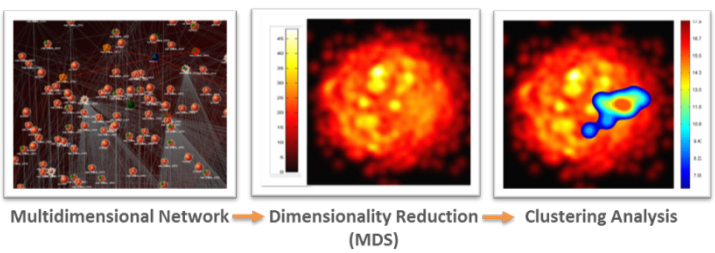


**FIGURE 1 - PROTEIN NETWORK CLUSTERING ANALYSIS.** A MULTIDIMENSIONAL PROTEIN NETWORK IS REDUCED TO 2D USING A MULTIDIMENSIONAL SCALING PROCEDURE. SUBSEQUENTLY, CLUSTERS OF PROTEINS OF INTEREST IN THE 2D NETWORK ARE IDENTIFIED. THE COLOUR GRADING OF THE IMAGES REFLECT PROTEIN DENSITY ACCORDING TO ADJUNCT SCALES (PROTEIN/PIXEL2).

## GENERATION OF BIOLOGICAL-CONCEPT SPECIFIC FINGERPRINTS

The microarray data submitted to GEO database comes from all kinds of sources, so we have samples ranging from the hundreds to the thousands for most biological concepts: cell types, tissues, pathologies, … It is a number high enough to consider the average gene profiling data for each concept as a reasonable representation of the normal expression levels for that concept.

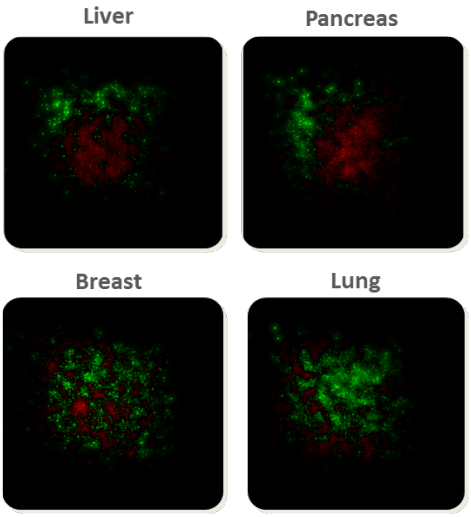


**FIGURE 2 – TISSUE-SPECIFIC 2D-FINGERPRINTS FROM MICROARRAY DATA.** GREEN DOTS CORRESPOND TO OVEREXPRESSED PROTEINS, WHILE RED DENOTES UNDER-EXPRESSED PROTEINS FOR THE CORRESPONDING TISSUE

The complementary contribution of various hundreds of microarrays ends up giving a valid, generalizable panorama of gene expression profiles for a particular biological concept. Taking advantage of this, we can add microarray transcription profiles for each biological concept to the global map, thereby generating a wide collection of specific maps for any biological concept contained in GEO database. The information coming from GEO is filtered, normalized and analysed and the key proteins differentially expressed for a particular biological concept respect all microarray data identified. The location of those proteins in an MDS projection of the human protein interaction map gives a specific image, so we have coined the term "**biological-concept protein fingerprint**" to reflect the fact that each projection univocally identifies a biological concept (In summary, feeding the global protein map with microarray data and applying clustering analysis allows us to generate biological-concept specific 2D-fingerprints. This achievement is by no means trivial; it provides an effective and simple method to characterize microarray data as whole taking into account interactions between proteins and gene expression profiles. Otherwise, we are limited to evaluating certain proteins, pathways or even regions of the map, at the expense of losing valuable information in the process.

Since the distance between 2D fingerprints is quantifiable the overall differences between control and intervention group can be measured. Thus, we can accomplish a global comparison between microarray slides, gaining extra information derived from systems biology. Furthermore, having a repertory of biological-concept 2D-fingerprints based on previous experiments stored in GEO, which can be envisaged as the average population, allows us to measure the distance between any newly generated data and the average population. This information could be useful to guide the adjustment intended to minimize the variations within sample groups, instead of simply relying on a statistical mean. Moreover, it highlights the key differentially expressed proteins characterizing a specific biological concept that are different from the rest of biological concepts and worth analysing further.

In summary, feeding the global protein map with microarray data and applying clustering analysis allows us to generate biological-concept specific 2D-fingerprints. This achievement is by no means trivial; it provides an effective and simple method to characterize microarray data as whole taking into account interactions between proteins and gene expression profiles. Otherwise, we are limited to evaluating certain proteins, pathways or even regions of the map, at the expense of losing valuable information in the process.

Since the distance between 2D fingerprints is quantifiable the overall differences between control and intervention group can be measured. Thus, we can accomplish a global comparison between microarray slides, gaining extra information derived from systems biology. Furthermore, having a repertory of biological-concept 2D-fingerprints based on previous experiments stored in GEO, which can be envisaged as the average population, allows us to measure the distance between any newly generated data and the average population. This information could be useful to guide the adjustment intended to minimize the variations within sample groups, instead of simply relying on a statistical mean. Moreover, it highlights the key differentially expressed proteins characterizing a specific biological concept that are different from the rest of biological concepts and worth analysing further.

## ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers HEALTH-F4-2012-305869 (SysMalVac) and HEALTH-F2-2012-306240 (SyStemAge).

## BIBLIOGRAPHY

1. Wheelan SJ, Martinez Murillo F and Boeke JD, The incredible shrinking world of DNA microarrays, *Mol Biosyst*, 4(7): 726-732, 2008
2. Edgar R, Domrachev M and Lash AE, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res*, 30(1): 207-210, 2002
3. Kitano H, Systems biology: a brief overview, *Science*, 295(5560): 1662-1664, 2002
4. Hase T, Tanaka H, Suzuki Y, Nakagawa S and Kitano H, Structure of protein interaction networks and their implications on drug design, *PLoS Comput Biol*, 5(10): e1000550, 2009
5. Sambourg L and Thierry-Mieg N, New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size, *BMC Bioinformatics*, 11: 605, 2010
6. Farrés J, Pujol A, Coma M, Ruiz JL, Naval J, Mas JM, Molins A, Fondevila J and Aloy P, Revealing the molecular relationship between type 2 diabetes and the metabolic changes induced by a very-low-carbohydrate low-fat ketogenic diet, *Nutr Metab (Lond)*, 7: 88, 2010
7. Yoshida M, Hatano N, Nishiumi S, Irino Y, Izumi Y, Takenawa T and Azuma T, Diagnosis of gastroenterological diseases by metabolome analysis using gas chromatography-mass spectrometry, *J Gastroenterol*, 47(1): 9-20, 2011
8. Ashburn TT and Thor KB, Drug repositioning: identifying and developing new uses for existing drugs, *Nat Rev Drug Discov*, 3(8): 673-683, 2004
9. Dudley JT, Deshpande T and Butte AJ, Exploiting drug-disease relationships for computational drug repositioning, *Brief Bioinform*, 12(4): 303-311, 2011
10. Mas JM, Pujol A, Aloy P and Farrés J. Methods and systems for identifying molecules or processes of biological interest by using knowledge discovery in biological data, US Patent Application Nº. 12/912,535, 2010
11. Godsil CD and Mckay BD, The Dimension of a Graph, *Quarterly Journal of Mathematics*, 31(124): 423-427, 1980