entropy

MDPI

Article

Multivariate Surprisal Analysis of Gene Expression Levels

Francoise Remacle [1,2], Andrew S. Goldstein [3] and Raphael D. Levine [2,4,*]

[1] Département de Chimie, B6c, Université de Liège, B4000 Liège, Belgium; fremacle@ulg.ac.be
[2] The Fritz Haber Research Center for Molecular Dynamics, The Institute of Chemistry, The Hebrew University of Jerusalem, Jerusalem 91904, Israel
[3] Department of Urology, David Geffen School of Medicine and Department of Molecular Cell & Developmental Biology, University of California, Los Angeles, CA 90095, USA; AGoldstein@mednet.ucla.edu
[4] Crump Institute for Molecular Imaging and Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, and Department of Chemistry & Biochemistry, University of California, Los Angeles, CA 90095; USA
* Correspondence: rafi@chem.ucla.edu; Tel.: +1-310-206-0476

Academic Editor: Anne Humeau-Heurtier
Received: 14 November 2016; Accepted: 7 December 2016; Published: 11 December 2016

Abstract: We consider here multivariate data which we understand as the problem where each data point *i* is measured for two or more distinct variables. In a typical situation there are many data points *i* while the range of the different variables is more limited. If there is only one variable then the data can be arranged as a rectangular matrix where *i* is the index of the rows while the values of the variable label the columns. We begin here with this case, but then proceed to the more general case with special emphasis on two variables when the data can be organized as a tensor. An analysis of such multivariate data by a maximal entropy approach is discussed and illustrated for gene expressions in four different cell types of six different patients. The different genes are indexed by *i*, and there are 24 (4 by 6) entries for each *i*. We used an unbiased thermodynamic maximal-entropy based approach (surprisal analysis) to analyze the multivariate transcriptional profiles. The measured microarray experimental data is organized as a tensor array where the two minor orthogonal directions are the different patients and the different cell types. The entries are the transcription levels on a logarithmic scale. We identify a disease signature of prostate cancer and determine the degree of variability between individual patients. Surprisal analysis determined a baseline expression level common for all cells and patients. We identify the transcripts in the baseline as the "housekeeping" genes that insure the cell stability. The baseline and two surprisal patterns satisfactorily recover (99.8%) the multivariate data. The two patterns characterize the individuality of the patients and, to a lesser extent, the commonality of the disease. The immune response was identified as the most significant pathway contributing to the cancer disease pattern. Delineating patient variability is a central issue in personalized diagnostics and it remains to be seen if additional data will confirm the power of multivariate analysis to address this key point. The collapsed limits where the data is compacted into two dimensional arrays are contained within the proposed formalism.

Keywords: multivariate analysis; maximal entropy; prostate cancer markers; personalized diagnostics; transcriptomics; high order SVD; tensor data format; ensemble phenotypes

1. Introduction

We derive the extension to multivariate arrays of the thermodynamic-like approach that we have recently proposed for identifying pattern responses in cancerous cell lines and patient data. Our

approach is based on information theory and surprisal analysis. It was earlier introduced to deal with physicochemical systems in disequilibrium [1–3]. It is currently being applied to biophysicochemical systems [4–8]. The essential difference with the methods already proposed [9–13] is that we use the thermodynamic entropy instead of using entropy as a statistical measure of dispersion.

In this paper we discuss the approach to multilevel arrays and apply it to transcription level data. We treat each transcript as a molecule so we have the thermodynamics (of disequilibrium) in a multicomponent system. The thermodynamic-like approach allows for the defining of a base line level for the expression level of each transcript. We show by example that indeed transcripts from different patients and/or different cells do have the same base line level. It is with respect to this baseline that the changes induced by the different conditions are quantified. By using a thermodynamic entropy and surprisal analysis, we have previously shown that the base line can be determined from genomic data. In addition to the partition function of each transcript, the base line includes all the contributions to the entropy that are common to all the conditions: non ideality, environment effects, etc. We have recently argued that the reference base line corresponds to the essential cell machinery and is valid across cell types, patients, and organisms [4]. Such a base line is also recovered in the multivariate case developed here.

One can use entropy in the information theoretic sense as a measure of dispersion from uniformity. Analyses of genetic networks from this point of view have received attention with special reference to the connectivity of the network [10,12,14–21]. It is the use of a thermodynamic-like expression for the entropy that places the present method apart. It is definitely not a minor correction because the base line distribution of expression levels of transcripts is a very non-uniform distribution that accounts for the great majority of an information theoretic entropy (>97%, [4]). The change in expression level induced by the disease is small compared to the base line. In other words, even in diseased patients cells are inherently stable. It is possible to induce a change of the stable state (e.g., [8]), but it is a change from one stable state to another.

Singular value decomposition (SVD) is used here as a mathematical tool to implement a fast and useful surprisal analysis of large data sets [7]. The technical point is that it is the logarithm of the expression level that is being expanded. The invariant part of that expansion is the free energy (in units of the thermal energy) of the transcript and specifies the baseline value.

For 2D arrays, we previously reported on using SVD as a practical tool to implement surprisal analysis and determine the gene patterns in terms of which the data can be analyzed and compacted [4–7,15,22–25]. In the multivariate case derived here, a High Order Singular Valued Decomposition (HOSVD) is used [26]. Compared to other tensor decomposition methods based on the Tucker decomposition [27,28], HOSVD has the advantage of leading to orthogonal patterns which allows it to characterize the deviation of the observed expression level with respect to the base line in a unique way.

We turn next to the background for our work in terms of systems biology [29]. High throughput genomic and proteomic data on various types of organisms are becoming routinely available. The purpose of measuring these massive data sets is to uncover trends and patterns in the genomic and proteomic responses to different conditions and perturbations, with the promise of a better understanding of diseases like cancer or for optimizing energy production biotechnologies based on unicellular organisms like yeast and algae. Typically, the number of conditions that can be engineered is much smaller than the number of gene transcripts that can be measured, where that number ranges in the tens of thousands depending on the organism. Similarly, the number of proteins that can nowadays be quantitatively measured is also quite large. Most of the methods of analysis deal with 2D rectangular arrays where all the conditions are treated on the same level. These can be different points in time [8,23,24,30], different nutrients conditions, or in the case of cancer, different cell types, oncogenes, and/or drug treatments. When all these different conditions are lumped into a 2D array, the correlation between different conditions is not easy to identify. In some cases, a multivariate analysis has been applied [31–33].

Another case that calls for a multivariate analysis is that where the genomic response of cancer is analyzed for different patients and their different cell types. This is the case that we focus on. We

show that identifying the gene response of the 3D tensor of data uncovers correlations between cell type and individual patients that are washed out if the 3D array is collapsed to a 2D one by averaging the gene response over patients and keeping the cell type as a relevant dimension, or averaging over cell type and keeping the different patients as the relevant axis.

The paper is organized as follows. We start by a brief outline of the 2D analysis in Section 2. The extension to a 3D multivariate array is derived in Section 3 where we show how the possible 2D results are limiting cases of the general case. The methodology is applied in Section 4 to a 3D array: the gene response measured for different cell types and different patients in the case of prostate cancer. This example allows us to provide an operative definition of two opposite trends. One the one hand, a disease is defined as common features that affect patients in the same way, but on the other, the analysis of the gene response of individual patients shows that different patients have different disease gene patterns. Correlating cell types and patient responses using a multivariate analysis provides a step towards personalized diagnostics. This is a special advantage of the multivariate analysis.

2. Surprisal Analysis of 2D Arrays

In the bivariate case, the data are arranged as a rectangular matrix **X** of dimension I by J where I, the number of gene transcripts, is typically much larger than J, the number of conditions for which the expression levels of the transcripts have been measured. Each column of **X** is a distribution over the gene transcripts i measured for the given condition j. Each row is the distribution over the conditions j measured for a given transcript i. Each reading X_{ij} is the measured expression level of transcript i under condition j. In a typical experiment involving many cells, X_{ij} is an average value over the fluctuations that can occur from cell to cell because the individual cell is a finite system [15].

Surprisal analysis seeks to characterize the level distribution X_{ij} as a distribution of maximal entropy subject to constraints. We assume that the same constraints operate for all the conditions but it can be that different constraints are more important for different conditions. The number of constraints, C, is typically smaller than the number of conditions, J, and therefore the constraints by themselves are not sufficient to determine a unique distribution. Among all the distributions that are consistent with the constraints, surprisal analysis determines that unique distribution whose entropy is maximal. When such a distribution exists it can be analytically constructed by the method of Lagrange undetermined multipliers [4,5,22]. Surprisal analysis employs the following expression for the logarithm of the expression level of gene i under condition j:

$$\ln\left(X_{ij}\right) = \ln\left(X_{i0}\right) - \sum_{\alpha=1}^{C} G_{i\alpha}\lambda_{\alpha j} \tag{1}$$

where $\ln(X_{i0})$ is the base line level for transcript i, that is the expression level of transcript i at maximum entropy without any constraints imposed by the conditions:

$$\ln\left(X_{i0}\right) = -G_{i0}\lambda_0 \tag{2}$$

The additional terms in Equation (1) reduce the value of the entropy from its thermodynamic maximum. $\lambda_{\alpha j}$ is the Lagrange multiplier for the constraint α in condition j. For the condition j, each constraint α lowers the value of the entropy by an extent given by the Lagrange multiplier $\lambda_{\alpha j}$ The Lagrange multiplier $\lambda_{\alpha j}$ is common to all genes but does depend on the condition. By definition of a base line, λ_0 in Equation (2) does not depend on the condition j. This invariance is not imposed in the numerical procedure described below and we show that up to numerical noise, the values of λ_0 computed for each condition are indeed the same. The value of the constraint α for the transcript i is $G_{i\alpha}$. These values for the transcripts are common for all the different conditions.

A key property of Equation (1) that is retained also in the higher dimensional case is the separability. Each deviation term in Equation (1) is made up as a product of two factors, so that all the genes that contribute to pattern α are changing in the same way (i.e., coherently) as the condition

is being changed. The "weight" $\lambda_{\alpha j}$ of condition j is the same for all the genes that participate in pattern α.

The aim of surprisal analysis is to retain as few as possible terms in Equation (1), that is, to identify the minimum set of constraints α that will give a good fit of $\ln(X_{ij})$ simultaneously for all conditions j. The constraints can be given a biophysical interpretation in terms of the processes that restrain the value of the entropy of the distribution to reach its maximal value. Unlike in the univariate case typically encountered in physical chemistry where the form of the constraints can be known a priori on physicochemical grounds [2], in the bivariate case, the constraints are typically not known. One efficient computational way to determine the constraints is to apply singular valued decomposition, SVD, of the logarithm of data, as follows. Construct the matrix **Y** such that its elements are the logarithms of the expression levels $Y_{ij} = \ln(X_{ij})$. The range of the indices i and j are quite different, $1 \le i \le I$, $1 \le j \le J$ and so Y is a rectangular I by J matrix, where typically $I >> J$. Therefore **Y** is a singular matrix but we can 'diagonalize' it in the manner of SVD as

$$\mathbf{Y} = \mathbf{G} \cdot \mathbf{\Omega} \cdot \mathbf{V}^T \tag{3}$$

where the columns of the matrices **G** and **V** are orthonormal. The matrix **G** is an I by J rectangular matrix and its rank is therefore J or lower. The matrix **Ω** is a diagonal J by J matrix made up of the J eigenvalues of **V** (or of the non-zero eigenvalues of **G**). If we want to express **Ω** as a function of **Y**, we get

$$\mathbf{\Omega} = \mathbf{G}^T \cdot \mathbf{Y} \cdot \mathbf{V} \tag{4}$$

which we can call "the most compact form" of the data.

The matrices **G** and **V** in Equation (3) are respectively the left and right eigenvectors of **Y**. The number of non-zero eigenvalues of **Y** is limited by its smallest dimension, which for us is the number of conditions J. The left and right eigenvectors of **Y** can be obtained as the normalized eigenvectors of the two covariance matrices that can be built from **Y**: the "small" $J \times J$ matrix $\mathbf{Y}^T\mathbf{Y}$ and the "large" $I \times I$ matrix $\mathbf{Y}\mathbf{Y}^T$. The matrix **Ω** in Equation (3) is obtained from the eigenvalue equation

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{V}^T \cdot \mathbf{\Omega}^2 \cdot \mathbf{V} \tag{5}$$

where $\mathbf{\Omega}^2$ is the diagonal matrix of the eigenvalues ω_α^2 and the vector $\mathbf{V}_\alpha$ the corresponding eigenvector. The maximum rank of $\mathbf{Y}^T\mathbf{Y}$ is its dimension J. The large I by I matrix $\mathbf{Y}\mathbf{Y}^T$ is also maximum rank J and has the same non-zero eigenvalues as $\mathbf{Y}^T\mathbf{Y}$. The eigenvectors $\mathbf{G}_\alpha$ correspond to the same eigenvalue as the eigenvector $\mathbf{V}_\alpha$ of the $\mathbf{Y}^T\mathbf{Y}$ matrix. Knowing the eigenvalues ω_α and the normalized eigenvectors $\mathbf{V}_\alpha$ is sufficient to determine the vectors $\mathbf{G}_\alpha$, $\alpha = 0,\ldots, J-1$, from the data:

$$\mathbf{G} = \mathbf{Y} \cdot \mathbf{V} \cdot \mathbf{\Omega}^{-1} \tag{6}$$

The SV decomposition given in Equation (3) can be recast into the surprisal form of Equation (1)

$$Y_{ij} = \ln\left(X_{ij}\right) = \sum_{\alpha=0}^{J-1} G_{i\alpha}\omega_\alpha \left(\mathbf{V}^T\right)_{\alpha j} = \sum_{\alpha=0}^{J-1} G_{i\alpha}\lambda_{\alpha j} \tag{7}$$

where the Lagrange multiplier $\lambda_{\alpha j}$ is thereby expressed as

$$\lambda_{\alpha j} = \omega_\alpha \left(\mathbf{V}^T\right)_{\alpha j} = \omega_\alpha V_{j\alpha} \tag{8}$$

The Lagrange multiplier $\lambda_{\alpha j}$ determines the importance of the constraint α for condition j. In Equation (8), this importance is shown to be a product of two terms. One factor, ω_α, is the overall importance of constraint α. The other is the (normalized) importance of this constraint for condition j. It is a normalized weight because the vector $\mathbf{V}_\alpha$ is normalized.

For a 2D data array, the Lagrange multipliers can be written in a matrix form

$$\mathbf{\Lambda} = \mathbf{\Omega} \cdot \mathbf{V}^T \tag{9}$$

For a 3D data array the Lagrange multipliers will form a 3D tensor.

From Equation (9) above, we see that the matrix $\mathbf{\Lambda}$ is a partial de-diagonalization of the diagonal most compact form of the data, $\mathbf{\Omega}$.

Using Equations (3) and (9), the surprisal given in Equation (1) can be written as matrix product

$$\mathbf{Y} = \mathbf{G} \cdot \mathbf{\Lambda} \tag{10}$$

The SVD approach ensures that the constraints are linearly independent since they are orthogonal. Both the eigenvectors $\mathbf{G}_\alpha$ and $\mathbf{V}_\alpha$ are orthonormal. Therefore, also the rows of the matrix of the Lagrange multipliers $\mathbf{\Lambda}$ are orthogonal.

The 2D approach was successfully implemented on microarrays data of gene transcripts for patient data, for several cancer cell models, and for unicellular organisms [4–7,15,22–25,30,34] subject to various conditions *j*. In particular, the values of the Lagrange multipliers that define the base line, λ_0, are found numerically to be independent of the condition index *j* in agreement with Equation (2). From Equation (8), one can see that the only way to ensure that λ_0 does not depend on *j* is that the eigenvector corresponding to the largest eigenvalue is uniform and from the condition that the eigenvectors are normalized the amplitudes in the vector have the value $1/\sqrt{J}$. Note that the rows of the matrix $\mathbf{Y}$ are not mean centered. When the rows of the matrix $\mathbf{Y}$ are mean centered, the matrix $\mathbf{Y}^T\mathbf{Y}$ becomes rank deficient and has one zero eigenvalue with the same corresponding uniform eigenvector. The remaining *J*−1 eigenvalues are unchanged compared to the non-mean centered case. The 2D arrays analysis provides a clear meaning to the base line term of Equation (2). The analysis of the gene transcripts that dominate the G_{i0} constraint show that the pattern of the base line corresponds to the essential machinery of the cell [4]. For $\alpha > 0$, the biological interpretation of the transcription patterns are defined by the analysis of the values of the Lagrange multipliers $\lambda_{\alpha j}$ on the different conditions *j*. Each constra $G_{i\alpha}$ corresponds to a pattern in the gene expression associated with the phenotype α. We will therefore use the terms transcription pattern and phenotype interchangeably where by phenotype we mean a biological process where all the participating genes act coherently (see Equation (1) or (7)). The analysis of the phenotypes $\alpha > 0$ and of the gene expression patterns was shown to characterize the progression of cancer in cell lines [5,24].

3. Generalization of the Surprisal Analysis for a Multivariate Array

The need for a generalization of surprisal analysis arises from the availability of gene transcript levels measured, for example, for individual patients and for different conditions or for different cell types. Keeping the different patients as a separate dimension in addition to cell type leads to data in the format of a 3D array. Another case where one can have more than 2D arrays of data is that where different cell types are submitted to different conditions [31–33]. Of course, a 4D data or even higher is also possible.

To bring the 3D tensor into a surprisal form, we use HOSVD [26]. This approach ensures that the constraints remain linearly independent and reverts to the 2D case discussed above as a limit. As an illustration of the formalism we will use data arrays where the gene expression level of different cell types is measured for several individual patients. The 3D analysis allows us to propose an interpretation of the notion of disease and of patient diversity.

We consider a 3D array, $\mathbf{T}$, for which the three dimensions are the number of gene transcripts, *I*, the number of cell types, *J*, and the number of patients, *K*. The elements of the tensor are defined as $T_{ijk} = \ln X_{ijk}$ where X_{ijk} is the measured expression level for transcript *i* of cell type *j* of patient *k*. The number of transcripts, of the order of 20,000, is typically much larger than the numbers of cell types or the number of patients. We have therefore a rectangular 3D $I \times J \times K$ tensor where $I >> J \approx K$. This inequality, typical of gene expression data where the number *I* of transcripts is very large, plus the

fact that we typically want the expression levels to remain an axis, dictates the compaction of the 3D data to a 2D form as discussed next.

*3.1. Representing the Tensor Surprisal **T** in Matrix Form*

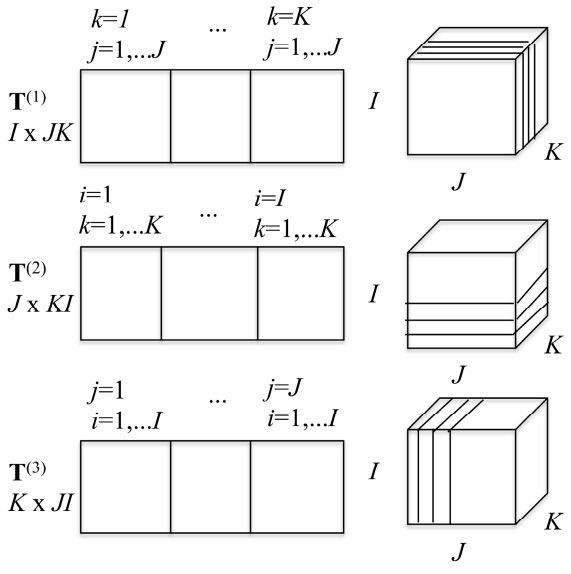
2D slices can be made from the 3D array in three ways. Keeping i fixed, one has I different $J \times K$ matrices, keeping j fixed, one has $JK \times I$ matrices, and keeping k fixed generates $KI \times J$ matrices. These three possible decompositions of the 3D tensor are arranged as shown in Scheme 1. From these three decompositions, one can build three matrices that possess the same information as the tensor itself, as is also shown in Scheme 1. The matrix $\mathbf{T}^{(1)}$ is a 2D $I \times (J \times K)$ matrix. $\mathbf{T}^{(1)}$ is made of K side by side $I \times J$ slices. The matrix $\mathbf{T}^{(2)}$ is of dimension $J \times (K \times I)$ and is made of I side by side $(J \times K)$ matrices and the matrix $\mathbf{T}^{(3)}$ of dimension $K \times (I \times J)$ that is made of J side by side $(K \times I)$ matrices. The three matrices $\mathbf{T}^{(1)}$, $\mathbf{T}^{(2)}$, and $\mathbf{T}^{(3)}$ are called the flattened matrices and correspond to the three ways to represent the tensor as a 2D matrix. These three flattened matrices are the basis for the high order SV decomposition of the tensor [26]. Each of these flattened rectangular matrices can be decomposed using the 2D SVD procedure as explained in Section 2 above. We get

$$\mathbf{T}^{(1)} = \mathbf{U} \cdot \mathbf{\Omega}^{(1)} \cdot \mathbf{M}^T \tag{11}$$

$$\mathbf{T}^{(2)} = \mathbf{V} \cdot \mathbf{\Omega}^{(2)} \cdot \mathbf{N}^T \tag{12}$$

$$\mathbf{T}^{(3)} = \mathbf{W} \cdot \mathbf{\Omega}^{(3)} \cdot \mathbf{L}^T \tag{13}$$

The SV decomposition of each of the three flattened matrices $\mathbf{T}^{(1)}$, $\mathbf{T}^{(2)}$, and $\mathbf{T}^{(3)}$ is algebraically the same procedure as discussed in Section 2 above for the 2D surprisal matrix. In particular, since the rows of the three matrices are not mean centered, for each of them, the normalized eigenvector of the small covariance matrix that corresponds to the largest eigenvalue has uniform amplitudes given by $1/\sqrt{\dim}$ where dim is the smallest dimension of each matrix.



Scheme 1. The three ways of representing the tensor **T** as two dimensional, Equations (11)–(13).

The left-hand side matrices in Equations (11)–(13) provide a SV decomposition for the elements of the 3D tensor [26]:

$$T_{ijk} = \sum_{r=0}^{JK-1} \sum_{s=0}^{J-1} \sum_{t=0}^{K-1} \Omega_{rst} U_{ir} V_{js} W_{kt} \tag{14}$$

where the matrix **U** is the $I \times JK$ matrix of the left eigenvectors of $\mathbf{T}^{(1)}$, **V** the $J \times J$ matrix of the left eigenvectors of $\mathbf{T}^{(2)}$, and **W** the $K \times K$ matrix of the left eigenvectors of $\mathbf{T}^{(3)}$. $\boldsymbol{\Omega}$ is the core tensor. Its dimensions are $JK \times J \times K$ and correspond to the number of non-zero eigenvalues of the flattened matrices $\mathbf{T}^{(1)}$, $\mathbf{T}^{(2)}$, and $\mathbf{T}^{(3)}$, respectively. The elements of the core tensor can be computed from its $JK \times JK$ flattened matrix $\boldsymbol{\Omega}^{(1)}$ [26] by

$$\begin{aligned} \boldsymbol{\Omega}^{(1)} &= \mathbf{U}^T \cdot \mathbf{T}^{(1)} \cdot (\mathbf{V} \otimes \mathbf{W}) \\ \left(\boldsymbol{\Omega}^{(1)}\right)_{rq} &= \sum_{i=0}^{I-1} U_{ir} \sum_{l=0}^{JK-1} \left(\mathbf{T}^{(1)}\right)_{il} (\mathbf{V} \otimes \mathbf{W})_{lq} \end{aligned} \tag{15}$$

In Equation (15), the symbol ⊗ stands for the direct product of the **V** and **W** matrices (see Equations (12) and (13)), which leads to a square $JK \times JK$ $\boldsymbol{\Omega}^{(1)}$ matrix. The matrices **U** and $\mathbf{T}^{(1)}$ are defined in Equation (11) above. Equation (15) is the practical form that can be used to compute the core tensor as a flattened matrix. Formally, the core tensor $\boldsymbol{\Omega}$ is defined from the n mode multiplication [26,27] of the input tensor **T** by the three matrices **U**, **V**, and **W**:

$$\boldsymbol{\Omega} = \mathbf{T}\ {}_{\times 1}\mathbf{U}\ {}_{\times 2}\mathbf{V}\ {}_{\times 3}\mathbf{W} \tag{16}$$

where the symbol "×1" means to multiply the input tensor from the gene direction I by the $I \times KJ$ **U** matrix of the left eigenvectors of $\mathbf{T}^{(1)}$ (Equation (11)), "×2" means to multiply **T** from the cell type direction J by the matrix **V** of the left eigenvectors of $\mathbf{T}^{(2)}$ (Equation (12)), and "×3"means to multiply **T** from the patient direction K by **W** (Equation (13)). By doing so, one obtains the core tensor $\boldsymbol{\Omega}$ of dimensions $JK \times J \times K$ whose flattened expression is given by Equation (15).

The labels of matrix elements as used in Equation (14) are, of course, arbitrary. However, once we made a particular choice it carried through. Specifically, because we use the index i to label transcripts, the notation U_{ir} implies that r will index the transcription patterns where **U** is the left matrix of $\mathbf{T}^{(1)}$ (Equation (11)). Similarly j is a cell type index so s will label cell type patterns where V_{js} is an element of the left matrix of $\mathbf{T}^{(2)}$ (Equation (12)) while r is a patient type index.

3.2. The Tensor Form of the Surprisal

Equation (14) generates a fully separable expression for the tensor form of the surprisal for a 3D data array

$$T_{ijk} = \sum_{r=0}^{JK-1} \sum_{s=0}^{J-1} \sum_{t=0}^{K-1} \Omega_{rst} U_{ir} V_{js} W_{kt}$$

It takes $JK \times J \times K$ terms to get an equality. The remarkable practical observation is that, in all the examples that we analyzed, very few terms need to be retained for this equation to provide an excellent approximation.

Equation (14) is analogous to Equation (3) of the 2D case. As in the 2D case it is possible to rewrite Equation (14) in a surprisal form. Here we discuss the surprisal of the expression level of the different transcripts, and since these are enumerated by the index i we pull the vector **U** up front in the summation. Rearrangement of the order of summations one has the 3D surprisal form as a summation over the different constraints labeled by r

$$\begin{aligned} T_{ijk} &= \sum_{r=0}^{JK-1} \sum_{s=0}^{J-1} \sum_{t=0}^{K-1} \Omega_{rst} U_{ir} V_{js} W_{kt} = \\ &= \sum_{r=0}^{JK-1} \mathbf{U}_{ir} \left(\sum_{s=0}^{J-1} \sum_{t=0}^{K-1} \Omega_{rst} V_{js} W_{kt} \right) = \sum_{r=0}^{JK-1} U_{ir} \lambda_{r(jk)} \end{aligned} \tag{17}$$

The Lagrange multiplier of the r'th constraint, $\lambda_{r(jk)}$, is thereby expressed as the term in the big bracket

$$\lambda_{r(jk)} = \sum_{s=0}^{J-1} \sum_{t=0}^{K-1} \Omega_{rst} V_{js} W_{kt} \tag{18}$$

In the 2D case the Lagrange multipliers for a given constraint (a given value of α) form a vector. The entries of this vector are labeled by the different conditions j. In the 3D case the Lagrange multipliers are matrices, one matrix for each value of the index r of the constraints. (We use a different letter for the index of the constraints in 2 and 3D so as to explicitly indicate the dimension). The elements of each matrix are labeled by j and k, the indices of the different cell types and the different patients.

The entire set of Lagrange multipliers that were a matrix in the 2D case (Equation (9)) are now a 3D tensor $\mathbf{\Lambda}$ of dimension $JK \times J \times K$, $\mathbf{\Lambda} = \mathbf{\Omega}_{\ x2}\mathbf{V}^T_{\ x3}\mathbf{W}^T$ whose flattened matrix form is given by

$$\mathbf{\Lambda}^{(1)} = \mathbf{\Omega}^{(1)} \cdot \left(\mathbf{V} \otimes \mathbf{W}\right)^T = \mathbf{\Omega}^{(1)} \cdot \left(\mathbf{V}^T \otimes \mathbf{W}^T\right) \tag{19}$$

By analogy to the matrix product representation, Equation (10), of the 2D case, one can also rewrite the surprisal as given by Equation (17) as an n mode product of the tensor representing the Lagrange multipliers, $\mathbf{\Lambda}$, by a matrix representing the constraints, $\mathbf{U}$:

$$\mathbf{T} = \mathbf{\Lambda}_{\ x1}\mathbf{U} \tag{20}$$

3.3. The Lagrange Multipliers

Numbering the eigenvalues by decreasing order, we can rewrite the surprisal form (17) as

$$T_{ijk} = \frac{1}{\sqrt{KJ}} U_{i0} \Omega_{000} + \sum_{r=1}^{JK-1} U_{ir} \lambda_{r(jk)} \tag{21}$$

where the first term is the base line for the 3D case, taking into account that

$$\begin{aligned} V_{0j} &= \frac{1}{\sqrt{J}},\ j = 0,\ldots,J-1 \\ W_{0k} &= \frac{1}{\sqrt{K}}, k = 0,\ldots,K-1 \end{aligned} \tag{22}$$

We stress that due to the condition on the dimensions of the tensor, $I >> J \approx K$, the I components of the vector $\mathbf{U}_0$ are not uniform. They define the gene transcript pattern that corresponds to the base line. The vector $\mathbf{U}_0$ plays the role of the vector $\mathbf{G}_0$ in the 2D case.

The elements of the core tensor, Ω_{rst} in Equation (17), can be ranked in decreasing order. Their weight is defined as

$$w_{rst} = \frac{\Omega_{rst}^2}{\sum_{r=0}^{KJ-1} \sum_{s=0}^{J-1} \sum_{k=0}^{K-1} \omega_{rst}^2} \tag{23}$$

The Ω_{000} term has always the highest weight. This is the term that defines the baseline. As we show in the illustrative example below, this term alone usually provides a semiquantitatively acceptable fit of the data. Often, fewer than a dozen terms suffice to recover most of the information contained in the data.

In Equation (21), each element of the tensor of the Lagrange multiplier $\lambda_{r(jk)}$

$$\lambda_{r(jk)} = \sum_{s=0}^{J-1} \sum_{t=0}^{K-1} \Omega_{rst} V_{js} W_{kt} \tag{24}$$

can be represented as a $J \times K$ heat map. Among the $J \times K$ terms in the sum typically very few ones of the highest weight dominate. We show an example of such heat maps of the elements of the tensor of the Lagrange multipliers in the analysis of the numerical example below. Here we draw attention to a 2D view that also leads to this form for the surprisal. In this 2D view we arrange the 3D data as a 2D array by giving each data column a double index *jk*. On reflection this is exactly the matrix $\mathbf{T}^{(1)}$ of dimension $I \times JK$ as discussed above. This matrix has JK non-zero eigenvalues. Its SV decomposition is given by Equation (11), where $\mathbf{M}$ is the matrix on the right:

$$T^{(1)}_{i(jk)} = \sum_{r=0}^{JK-1} G_{ir}\omega_r M_{(jk)r} = \sum_{\alpha=0}^{JK-1} G_{ir}\lambda_{r(jk)} \tag{25}$$

From the surprisal form (21), we have the equality

$$\lambda_{r(jk)} = \omega_r M_{(jk)r} = \sum_{s=0}^{J-1}\sum_{t=0}^{K-1} \Omega_{rst} V_{js} W_{kt} \tag{26}$$

3.4. 2D Limiting Forms of the Tensor Form of the Surprisal

2D limits of the 3D surprisal enable us, for example, to analyze separately the correlations between cell types, irrespective of the patients or the correlations between patients irrespective of cell types. To do so one needs to average the data over patients or over cell type respectively. In the first case, one will get a 2D array of dimension $I \times J$, and in the second one, a 2D array of dimension $I \times K$. Using an overbar to designate an average we write when we average the surprisal

$$T_{ij\bar{k}} = \frac{1}{K}\sum_{k=0}^{K-1} T_{ijk} \tag{27}$$

$$T_{i\bar{j}\,k} = \frac{1}{J}\sum_{j=0}^{J-1} T_{ijk} \tag{28}$$

These two 2D matrices can be obtained by starting from the general surprisal form given by Equation (17) and, as we next show, retain only terms $t = 0$ (average over patients) or only terms $s = 0$ (average over cell type). From averaging over patients in the surprisal form of the tensor, Equation (27), takes the form:

$$\begin{aligned} T_{ij\bar{k}} &= \frac{1}{K}\sum_{k=0}^{K-1} T_{ijk} = \frac{1}{\sqrt{K}}\sum_{r=0}^{KJ-1} U_{ir}\sum_{s=0}^{J-1} V_{js}\sum_{t=0}^{K-1}\Omega_{rst}\sum_{k=0}^{K-1} W_{kt}W_{k0} \\ &= \frac{1}{\sqrt{K}}\sum_{r=0}^{KJ-1} U_{ir}\left(\sum_{s=0}^{J-1}\Omega_{rs0}V_{js}\right) = \frac{1}{\sqrt{K}}\sum_{r=0}^{KJ-1} U_{ir}\lambda_{rj0} \end{aligned} \tag{29}$$

where Equation (22) and the orthogonality between the eigenvectors of $\mathbf{T}^{(3)}$ have been taken into account. Similarly, when averaging over cell types, one obtains

$$\begin{aligned} T_{i\bar{j}\,k} &= \frac{1}{J}\sum_{j=0}^{J-1} T_{ijk} = \frac{1}{\sqrt{J}}\sum_{r=0}^{KJ-1} U_{ir}\sum_{t=0}^{K-1} W_{kt}\sum_{s=0}^{J-1}\Omega_{rst}\sum_{j=0}^{J-1} V_{js}V_{j0} \\ &= \frac{1}{\sqrt{J}}\sum_{r=0}^{KJ-1} U_{ir}\left(\sum_{t=0}^{K-1} W_{kt}\Omega_{r0t}\right) \\ &= \frac{1}{\sqrt{J}}\sum_{r=0}^{KJ-1} U_{ir}\lambda_{r0k} \end{aligned} \tag{30}$$

The expressions of Equations (29) and (30) are not fully identical to what one would get by applying 2D SVD directly to averaged $I \times J$ and $I \times K$ data matrices where the difference is that the

logarithm of the average is not the same as the average of the logarithm (see also Section S1 of the supplementary materials).

By performing a double average over both patients and cell types, the 1D limit is given by

$$\begin{aligned} T_{i\overline{jk}} &= \frac{1}{JK}\sum_{j=0}^{K-1}\sum_{k=0}^{K-1} T_{ijk} \\ &= \frac{1}{\sqrt{JK}}\sum_{r=0}^{KJ-1} U_{ir}\sum_{s=0}^{J-1}\sum_{t=0}^{K-1}\Omega_{rst}\sum_{j=0}^{J-1} V_{js}V_{j0}\sum_{k=0}^{K-1} W_{kt}W_{k0} \\ &= \frac{1}{\sqrt{JK}}\sum_{r=0}^{KJ-1} U_{ir}\Omega_{r00} \end{aligned} \tag{31}$$

The 2D and even more so the 1D limit show how information is lost upon averaging. When 3D data is available the Lagrange multiplier for constraint r is given by, cf. Equation (17)

$$\lambda_{r(jk)} = \left(\sum_{s=0}^{J-1}\sum_{t=0}^{K-1}\Omega_{rst}V_{js}W_{kt}\right) \tag{32}$$

When there is an averaging on either patient or cell type data

$$\lambda_{r0k} = \sum_{t=0}^{K-1}\Omega_{r0t}W_{kt} \quad , \quad \lambda_{rj0} = \sum_{s=0}^{J-1}\Omega_{rs0}V_{js}$$

while in the 1D case when averaging is over both

$$\lambda_{r00} = \Omega_{r00} \tag{33}$$

4. Analysis of the Correlations between Cell Type and Patient in a Cohort of 6 Patients

Section 4 illustrates multivariate analysis of a data tensor where the gene level expressions are recorded for two cell types (basal or luminal) of six patients with a benign or advanced stage of prostate cancer. For each patient, there are therefore four cell types: benign luminal (BeLu), cancerous luminal (CaLu), benign basal (BeBa), and cancerous basal (CaBa). Each measurement was done in three replicates and only the genes that satisfied the confidence Student test were kept in the analysis. This leads to a data tensor of dimensions 18220 × 4 × 6.

4.1. Multivariate Surprisal Analysis

The starting point of the multivariate analysis of the surprisal is to compute the weights of the elements of the core tensor Ω_{rst} given by Equation (23). The highest ones are plotted in Figure 1. The base line element, Ω_{000}, dominates and with a weight (Equation (23)) equal to 0.99721 the base line accounts for 99.7% of the data. Also, in 2D expression level vs. time data the base line dominates [4] and it typically accounts for about 98% of the data. The next highest two elements are three orders of magnitude smaller and correspond to averaging over patient (1-1-0) and averaging over cell type (2-0-1). The three highest weight terms together represent 99.8% of the data. To interpret these results in terms of cell and patient phenotypes we plot the amplitudes $\Omega_{rst}V_{js}W_{kt}$ of the terms 1-1-0 and 2-0-1 as a function of the combined cell type-patient index, Figure 2. Thereby it is seen that the 1-1-0 element of the tensor is a cell type phenotype and provides the distinction between basal and luminal cells, while the 2-0-1 is a patient phenotype which distinguishes between patient P6 and the rest of the patient. The amplitudes plotted in Figure 2 are the terms in the tensor expansion, Equation (14), and are the main contributors to tensor of the Lagrange multiplier $\lambda_{r(jk)}$ defined in Equation (24). To ascertain that these amplitudes dominate we show the heat map of the *s* and *t* terms that contribute to a given $\lambda_{r(jk)}$ shown in Figure 3. It is seen that indeed the largest terms are the 1-1-0 and 2-0-1 terms. The averaging over cell types term 1-0-1 is also shown in Figure 2. The pattern of its amplitude on the *jk* cell-type-patient index is very similar to that of the 2-0-1 term. Its weight amounts to

0.01% of the data. We also show in Figure 2 the amplitudes of a disease phenotype, the 3-2-0 term. This term has opposite amplitudes on the benign and cancer cells. As we further discuss in Section 4.2 below, a disease phenotype needs to have the value 2 for the index *s*. The 3-2-0 term is the highest weight term of this type that appears in the highest weight list. It is not dominant (see Figure 1). This term corresponds to the disease cell phenotype ($s = 2$) and an averaging over patients ($t = 0$).

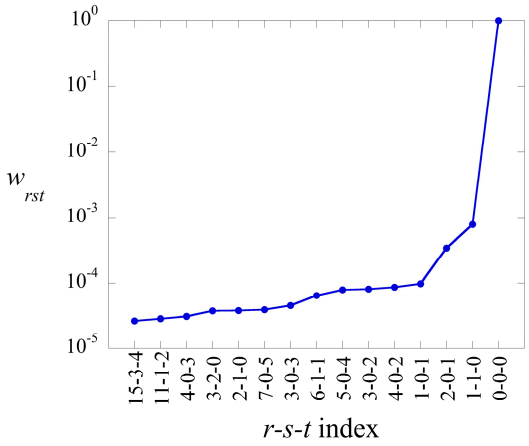


Figure 1. The 15 highest weights w_{rst} (Equation (23)) of the core tensor plotted on a logarithmic scale in decreasing order. Note the sheer drop beyond 0-0-0. Including terms on or beyond 4-0-3 only serves to fit the experimental noise in the data. See also Figure 4.

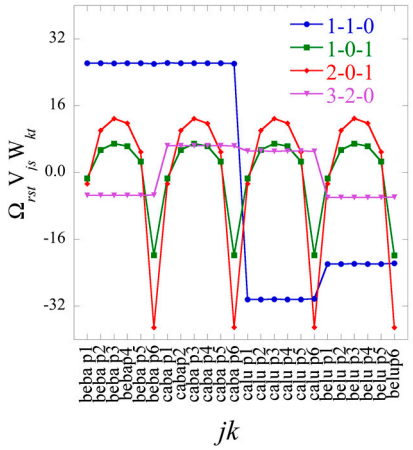


Figure 2. The magnitude of the two highest weight terms (*r-s-t* =1-1-0, 2-0-1) of the tensor form of the surprisal as a function of the *jk* index, conditionpatient. The first two highest terms, 1-1-0 and 2-0-1, already provide a very good fit for the surprisal. The phenotypes $r = 1$ and $r = 2$ are shown in Figures 3 and 4 below. While $r = 1$ mainly corresponds to the distinction between basal and luminal cells with a small modulation due to the patients, $r = 2$ corresponds to the distinction between P6 and P1 and the rest of the patients with a small modulation due to the cell type. In this sense, 2-0-1 is more clearly a patient phenotype than 1-0-1. The term 3-2-0 is the first disease term. As shown, it distinguishes between benign and cancer cells in that it changes its sign. It is the 12th term in the tensor expansion (see Figure 1 for the weights, w_{rst}) of the different terms and so it does not contribute much to the recovery of the input data.

In Figure 3 one can also see what are the other terms of the tensor form of the surprisal, (Equation 17), that mostly contribute to the value of the Lagrange multiplier $\lambda_{r(jk)}$ for a given cell type *j* and a given patient *k*. For example, it is clear from the heat map that more terms are needed to reproduce the cell type "benign basal" of patient P4 than the cell type "cancer luminal" of P6. This is

in part due to the fact that the most important patient phenotype, $t = 1$, that emerges from our analysis is one that has an overwhelming weight in P6. The description of the individuality of P4 involves more patient phenotypes. This is the reason why most of the highest weight terms beyond the terms 1-1-0 and 2-0-1 involve a patient phenotype and averaging over cell type. The heat map of the terms that contribute to a given Lagrange multiplier $\lambda_{r(jk)}$ provides a quantitative analysis of the correlations between cell type and patient and opens the way to personalized medicine.

The detailed structure shown in Figure 3 can also be viewed as when the Lagrange multiplier $\lambda_{r(jk)} = \omega_r M_{(jk)r}$ is plotted vs. the combined index jk. As discussed in connection with the derivation of Equation (26), this multiplier can also be viewed as a result of analysis of a 2D data matrix where each data column corresponds to a pair of conditions jk. Such a plot is shown for $r = 1,2$ in Figure 4. To plot two terms on a common scale we plot the amplitude $M_{(jk)r}$.

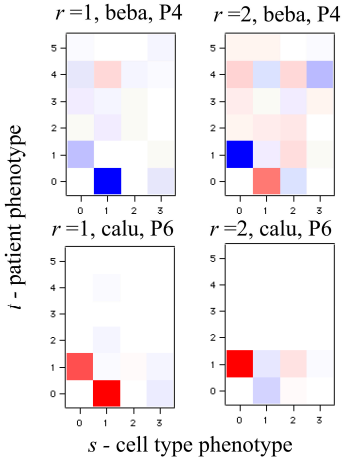


Figure 3. The magnitudes of the 24 terms $\Omega_{rst}V_{js}W_{kt}$ (see Equations (17) and (18)) that contribute to the value of a Lagrange multiplier $\lambda_{r(jk)}$ plotted as a heat map as a function of the cell phenotype index s and the patient phenotype t, r is the index of the constraint. Analyzing the relative importance of the different terms for a given constraint r and a given cell type and patient gives access to personalized medicine. Top left: $r = 1$, j = beba, k = P4, top right $r = 2$, j = beba, k = P4, bottom left: $r = 1$, j = calu, k = P6, bottom right: $r = 2$, j = beba, k = P4. Note that the terms $s = 1$, $t = 0$, and $s = 0$ and $t = 1$ dominates all panels. In addition, the heat maps clearly indicate that more terms significantly contribute to a given cell type for patient P4 than P6. This can be understood because several patient phenotypes contribute to the description of P4.

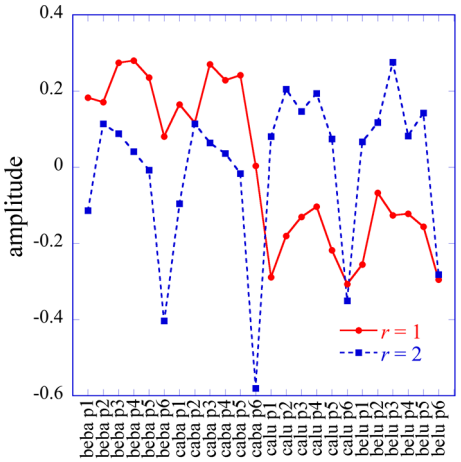


Figure 4. Amplitudes $M_{(jk)r}$ of the two dominant patterns, $r = 1,2$, of the bivariate transcription data matrix with 24 columns labeled by both cell type and patient. The $r = 1$ pattern is high on basal cells and low on luminal for each patient. The $r = 2$ pattern distinguishes patient 6 from the rest.

The base line and the two highest terms in the expansion of the tensor surprisal suffice to provide a very acceptable fit of the data (see Figure 5). The next five terms that are shown in Figure 1 serve mainly to distinguish among patients and correspond to an averaging over cell type. Terms that show correlations between cell type and patient, meaning both values of s and t are different from zero, have smaller weights.

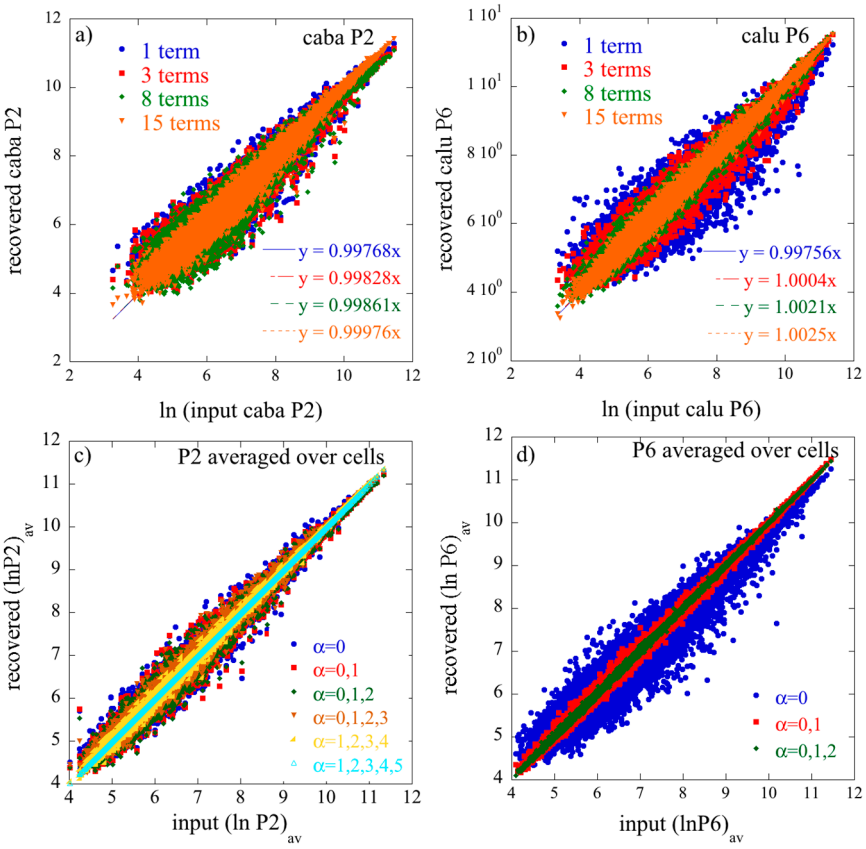


Figure 5. Scatter plots of the recovered data vs. the input data for given cell type and patient and for a 2D computation that is done for data averaged over cell types. Shown in panels (**a**) and (**b**) are the cell type caba for P2 and calu for P6, respectively. The recovered data are computed for an increasing number of terms in the surprisal expansion. In the tensor analysis, panels (**a**) and (**b**), one (only the base line), three, eight, and fifteen terms are kept. There are 24 combinations of patients and cell types and there are thousands of transcription levels for each pair. For all 24 pairs the three terms suffice to characterize the data. Eight terms are needed for representation of details and from twelve terms on the fit is to the noise in the data. The RNA profile calu of P6 which is characterized by the first patient phenotype in the tensor analysis is better fitted than the caba profile of P2 that is spread over several patient phenotypes. In panels (**c**) and (**d**) where the data is averaged over cell types, two terms beyond the base line $\alpha = 0$ capture most of P6 (**d**), which is characterized by the first phenotype patient. Many more terms are need for P2, because the RNA profile is described by several patient phenotypes.

Figure 5a shows the fit to the data for a given cell type and patient computed for an increasing number of terms in the expansion of the surprisal, Equation (17). The number of terms kept are one term, that corresponding to the 0-0-0 base line term, three terms (0-0-0, 1-1-0, and 2-0-1), eight terms, and 15 terms (see Figure 1 for the values of the indices). Keeping only three terms, terms that are just three entries of the core tensor, is already sufficient to provide a good fit for all the 24 combinations between cell type and patients. We also note that retaining three and eight terms does not significantly improve the fit obtained by retaining the base line only. Also shown in Figure 5b is the fit to the data, but from a 2D analysis where the patient data has been averaged over cell types. The point is to show how the tensor representation that keeps the cell-patient correlation explicitly has a

richer and finer structure. Figure 5c,d show the 2D surprisal analysis of the input data. The 2D input matrix is obtained by averaging over cell type for each of the six patients. In such an analysis, the information over cell type is washed out. The Lagrange multiplier, $\lambda_{\alpha j}$, of a given constraint, α, is a vector that depends on the index j of the patient only (see Equation (1) and Section 2). For such averaging, the constraints correspond to a patient phenotype, irrespective of the cell type. One needs several patient phentoypes to satisfactorily recover the data for patient P2 (Figure 5c) while the data of patient P6 are well fitted with only two patient phenotypes.

4.2. Dominant Phenotypes Along Each Tensor Direction

We discuss the phenotypes as determined by the 3D tensor data and as approximated by the ensemble averaging approach discussed in detail in the supplementary materials file.

The "cell type" phenotypes are defined by the left eigenvectors of the matrix **V** of the flatten matrix $\mathbf{T}^{(2)}$, Equation (12). These eigenvectors are also those defining the phenotypes in the 2D limit given by Equations (27) and (28). Their importance is ranked by the magnitude of the eigenvalues. The eigenvalues of $\mathbf{T}^{(2)}$ are reported in Table S1 of the supplementary materials. As discussed, the largest eigenvalue corresponds to a uniform eigenvector whose amplitudes are given by $1/\sqrt{J}$ and this defines the base line. The amplitudes of the eigenvector corresponding to the next largest eigenvalue reflects the distinction between basal and luminal cells (see Figure 6a) while the next largest phenotype discriminates between benign and cancerous cells, Figure 6b. Beyond that, the phenotypes are becoming not so secure because they are increasingly more contaminated by noise. To assess the biology associated with disease, genes that discriminate between benign and cancerous cells were characterized into pathways using DAVID Bioinformatics [35]. Of note, the most significant pathways included immune response and several immune-related terms, supporting a role for the immune system in tumor development (Table S2).

To assess the robustness of these three phenotypes, we have also carried out the 2D SVD decomposition of the "gene transcript" vs "cell type" matrices for each of the six patients separately, as shown in Figure 6. The corresponding eigenvectors are reported in dotted lines in the figures and the eigenvalues in Table S1. As a last check, we analyzed the $I \times J$ matrix obtained by averaging the logarithm of the data over patients. For the off diagonal elements of the 2D covariance matrix there is a difference between averaging first and analyzing next or analyzing first and averaging the surprisal next.

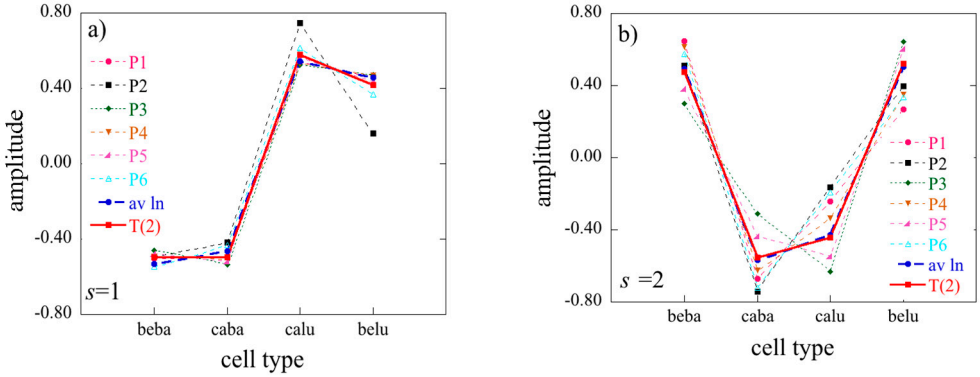


Figure 6. Amplitudes of the dominant cell type eigenvector V_{js}, (**a**) for $s = 1$ and (**b**) for $s = 2$ on the four cell types, $j = 1,2,3,4$. Bold full line (red, color on line), computed for the diagonalization of the $\mathbf{T}^{(2)}$ matrix, cf. Equation (12). Bold dashes (blue, color on line), computed from the diagonalization of the 2D matrix obtained by averaging the logarithm of the data over patients (see Section S1 of the supplementary materials)**.** Thin lines, computed by 2D SVD analysis of the surprisal for each one of the six patients. For the dominant phenotype, which corresponds to the distinction between basal and luminal cells, all methods of analysis concur. $s = 2$ is the "disease" phenotype and distinguishes between benign and cancer cells. While there is a wider dispersion from one patient to the next, the tensor analysis and the average over patients give similar results.

For the dominant phenotype, that discriminates between basal and luminal cells, cf. Figure 6a, eigenvectors computed for each patient individually almost superimpose with those defined from the tensor analysis and by averaging over patients. For the next phenotypes that are related to the disease, there is somewhat more dispersion among the eigenvectors computed for individual patients, but the tensor analysis and the average over patient lead to very similar results. This indicates that the "cell phenotypes" are robust and common to all patients.

We turn next to the complementary aspect, that of the analysis of the "patient" phenotypes. Those are defined by the left eigenvectors, $\mathbf{W}$, of the flattened $\mathbf{T}^{(3)}$ matrix (Equation (13)). Those eigenvectors also define the patient phenotype in the 2D analysis given by Equation (29). Similarly to the analysis of the "cell type" phenotypes, the "patient" phenotypes obtained from the tensor decomposition can be compared to those defined by the 2D SVD analysis of the average of logarithm of the data over cell types as well as those obtained for each cell type individually.

Unlike for the cell type analysis, the variations from patient to patient are more correlated to the cell type. Only two patient phenotypes, $t = 1$ (see Figure 7) and $t = 2$ are common to all cell types and even for those, there is more dispersion between the results for each cell type than between the results of each patient in the cell type analysis. The patient phenotype $t = 1$ indicates that patient 6 and to a certain extent patient 1 are different from the rest of the patients.

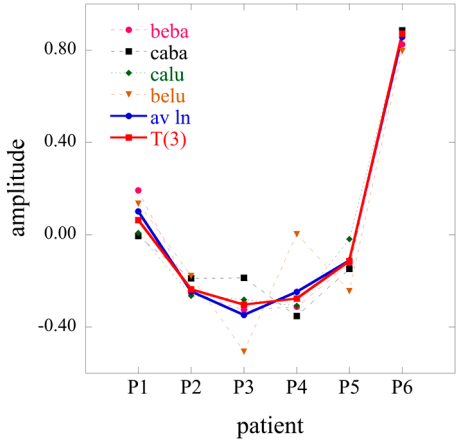


Figure 7. Amplitudes W_{kt} of the patient eigenvector, $t = 1$, on the six patients, $k = 1,2,..,6$. Bold full line (red, color on line), computed for the diagonalization of the $\mathbf{T}^{(3)}$ matrix (Equation (13)). Bold dashes (blue, color on line), computed from the diagonalization of the 2D matrix obtained by averaging first over patients and then performing surprisal analysis. Thin lines, computed for each of the four cell types individually. The patient-cell type correlations are manifested strongly in the patient phenotype analysis which leads to a larger dispersion among the patient analysis for given cell types. Note, however, that the tensor analysis and the averaging over patients give similar results.

From $t = 3$ on, the patient-cell type correlations are too strong to be able to define a patient phenotype across all cell types. In that case, the tensor phenotype is different from that obtained by averaging over cell type.

5. Conclusions

The multivariate surprisal analysis of the gene transcript expression levels measured for several cell types of different patients allows us to characterize the gene expression of dominant cell type and patient phenotypes. In the case of the data that we have analyzed, the dominant cell phenotype distinguishes between basal and luminal cells, irrespective of the benign or cancer disease character. The next cell phenotype is the disease phenotype, that discriminates between benign and cancer, irrespective of the cell type. The patient analysis shows that patient P6 is different from the rest of the five patients. A second patient phenotype indicates that P1 and P6 are different from P2, P3, P4, and P5.

The tensor analysis shows that beyond the distinction between basal and luminal cells, the data are dominated by the patient to patient variations. The tensor analysis thereby opens the way to a quantification of the correlations between cell types and patients and could therefore contribute towards personalized diagnostics and intervention.

Supplementary Materials: The following are available online at www.mdpi.com/1099-4300/18/12/445/s1. Table S1: Eigenvalues of the flatten matrix $\mathbf{T}^{(2)}$ that defines the cell phenotype in the tensor analysis and of the 2D surprisal analysis carried out on average over patients of the ln of the input data, Table S2. Gene analysis of the cell phenotype.

Acknowledgments: This work was supported by the prostate cancer foundation grants to AG and RDL and by the EC FP7-funded BAMBI Project 618024. FR is a director of research with FNRS (Fonds National de la Recherche Scientifique), Belgium.

Author Contributions: All authors participated in all aspects of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alhassid, Y.; Levine, R.D. Connection between maximal entropy and scattering theoretic analyses of collision processes. *Phys. Rev. A* **1978**, *18*, 89–116.
2. Levine, R.D. Information theory approach to molecular reaction dynamics. *Annu. Rev. Phys. Chem.* **1978**, *29*, 59–92.
3. Levine, R.D.; Bernstein, R.B. Energy disposal and energy consumption in elementary chemical reactions. Information theoretic approach. *Acc. Chem. Res.* **1974**, *7*, 393–400.
4. Kravchenko-Balasha, N.; Levitzki, A.; Goldstein, A.; Rotter, V.; Gross, A.; Remacle, F.; Levine, R.D. On a fundamental structure of gene networks in living cells. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 4702–4707.
5. Remacle, F.; Kravchenko-Balasha, N.; Levitzki, A.; Levine, R.D. Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 10324–10329.
6. Zadran, S.; Remacle, F.; Levine, R.D. miRNA and mRNA cancer signatures determined by analysis of expression levels in large cohorts of patients. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19160–19165.
7. Remacle, F.; Levine, R.D. Statistical thermodynamics of transcription profiles in normal development and tumorigeneses in cohorts of patients. *Eur. Biophys. J.* **2015**, *44*, 709–726.
8. Zadran, S.; Arumugam, R.; Herschman, H.; Phelps, M.E.; Levine, R.D. Surprisal analysis characterizes the free energy time course of cancer cells undergoing epithelial-to-mesenchymal transition. *Proc. Natl. Acad. Sci. USA* **2014**, *109*, 4702–4707.
9. Mora, T.; Walczak, A.; Bialek, W.; Callan, C.G., Jr. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA* **2009**, *107*, 5405–5410.
10. Lezon, T.R.; Banavar, J.R.; Cieplak, M.; Maritan, A.; Fedoroff, N.V. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 19033–19038.
11. Aghagolzadeh, M.; Soltanian-Zadeh, H.; Araabi, B.N. Information theoretic hierarchical clustering. *Entropy* **2011**, *13*, 450–465.
12. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Dalla Favera, R.; Califano, A. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **2006**, *7* (Suppl. S1), S7, doi: 10.1186/1471-2105-7-S1-S7.
13. Margolin, A.A.; Califano, A. Theory and limitations of genetic network inference from microarray data. *Ann. N.Y. Acad. Sci.* **2007**, *1115*, 51–72.
14. Yeung, M.K.; Tegner, J.; Collins, J.J. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6163–6168.
15. Shin, Y.S.; Remacle, F.; Fan, R.; Hwang, K.; Wei, W.; Ahmad, H.; Levine, R.D.; Heath, J.R. Protein signaling networks from single cell fluctuations and information theory profiling. *Biophys. J.* **2011**, *100*, 2378–2386.
16. Schneidman, E.; Still, S.; Berry, M.J.; Bialek, W. Network information and connected correlations. *Phys. Rev. Lett.* **2003**, *91*, 238701, doi: 10.1103/PhysRevLett.91.238701.
17. Rosvall, M.; Bergstrom, C.T. An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7327–7331.

18. Quigley, D.A.; To, M.D.; Kim, I.J.; Lin, K.K.; Albertson, D.G.; Sjolund, J.; Pérez-Losada, J.; Balmain, A. Network analysis of skin tumor progression identifies a rewired genetic architecture affecting inflammation and tumor susceptibility. *Genome Biol.* **2011**, *12*, R5, doi: 10.1186/gb-2011-12-1-r5.
19. Nykter, M.; Price, N.D.; Larjo, A.; Aho, T.; Kauffman, S.A.; Yli-Harja, O.; Shmulevich, I. Critical networks exhibit maximal information diversity in structure-dynamics relationships. *Phys. Rev. Lett.* **2008**, *100*, 058702.
20. Alter, O. Genomic signal processing: From matrix algebra to genetic networks. In *Microarray Data Analysis: Methods and Applications*, Korenberg, M.J., Ed. Humana Press: Totowa, NJ, USA, 2007.
21. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531–537.
22. Gross, A.; Levine, R.D. Surprisal analysis of transcripts expression levels in the presence of noise: A reliable determination of the onset of a tumor phenotype. *PloS ONE* **2013**, *8*, e61554.
23. Gross, A.; Li, C.M.; Remacle, F.; Levine, R.D. Free energy rhythms in saccharomyces cerevisiae: A dynamic perspective with implications for ribosomal biogenesis. *Biochemistry* **2013**, *52*, 1641–1648
24. Kravchenko-Balashaa, N.; Remacle, F.; Gross, A.; Rotter, V.; Levitzki, A.; Levine, R.D. Convergence of logic of cellular regulation in different premalignant cells by an information theoretic approach. *BMC Syst. Biol.* **2011**, *5*, 42.
25. Wei, W.; Shi, Q.H.; Remacle, F.; Qin, L.D.; Shackelford, D.B.; Shin, Y.S.; Mischel, P.S.; Levine, R.D.; Heath, J.R. Hypoxia induces a phase transition within a kinase signaling network in cancer cells. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, E1352–E1360.
26. De Lathauwer, L.; De Moor, B.; Vandewalle, J. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **2000**, *21*, 1253–1278.
27. Kolda, T.G.; Bader, B.W. Tensor decompositions and applications. *SIAM Rev.* **2009**, *51*, 455–500.
28. Tucker, L.R. Some mathematical notes on three-mode factor analysis. *Psychometrika* **1966**, *31*, 279–311.
29. Alon, U. *An Introduction to Systems Biology*. CRC Press: Boca Raton, FL, USA, 2007.
30. Willamme, R.; Alsafra, Z.; Arumugam, R.; Eppe, G.; Remacle, F.; Levine, R.D.; Remacle, C. Metabolomic analysis of the green microalga chlamydomonas reinhardtii cultivated under day/night conditions. *J. Biotechnol.* **2015**, *215*, 20–26.
31. Omberg, L.; Golub, G.H.; Alter, O. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 18371–18376.
32. Ponnapalli, S.P.; Saunders, M.A.; Van Loan, C.F.; Alter, O. A higher-order generalized singular value decomposition for comparison of global mrna expression from multiple organisms. *PLoS ONE* **2011**, *6*, e28072.
33. Sankaranarayanan, P.; Schomay, T.E.; Aiello, K.A.; Alter, O. Tensor GSVD of patient- and platform-matched tumor and normal DNA copy-number profiles uncovers chromosome arm-wide patterns of tumor-exclusive platform-consistent alterations encoding for cell transformation and predicting ovarian cancer survival. *PLoS ONE* **2015**, *10*, e0121396.
34. Zadran, S.; Remacle, F.; Levine, R.D. Microfluidic chip with molecular beacons detects miRNAs in human CSF to reliably characterize CNS-specific disorders. *RNA Dis.* **2016**, *3*, e1183.
35. Huang, D.W.; Sherman, B.T.; Lempicki, R.A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc.* **2008**, *4*, 44–57.