Supplementary Materials: Multivariate Surprisal Analysis of Gene Expression Levels

Francoise Remacle, Andrew S. Goldstein and Raphael D. Levine

1. Surprisal Analysis and Why the Weights of the Deviations are Ensemble Properties

Surprisal analysis (1,2) of the microarray data (3,4) provides a representation for the logarithm of the expression level of the transcripts. For the array data of patient n we use the form, Equation (1) of the main text:

$$\underbrace{\ln X_i^n(c)}_{\substack{\text{measured expression level}\\ \text{of transcript } i \text{ for cell type } c\\ \text{in data of patient } n}} = \underbrace{\ln X_i^{on}}_{\substack{\text{the base expression level}\\ \text{of transcript } i\\ \text{of patient } n}} - \underbrace{\sum_{\alpha=1} G_{i\alpha}^n \lambda_\alpha^n(c)}_{\substack{\text{deviations from the base line}\\ \text{characteristic of cell type } c\\ \text{of patient } n}} \quad \text{(S1)}$$

Surprisal analysis seeks to keep few terms in the sum in Equation (S1) while still providing an accurate representation of the experimental data. In the data used here there are six patients, $n = 1, 2, .., 6$

The zeroth term, $\ln X_i^{on}$, is the base line part of the level of expression. It is essentially the same (see section III of the SI for all patients). This is definitely not a uniform distribution and different transcripts do have fold differences in their level. It is this variation of the baseline as a function of the gene index, i, that separates our work from studies where entropy is used as a statistical measure of dispersion.

$\alpha = 1, 2, ..$ labels the different possible transcription patterns that cause a deviation of the measured expression level from the base line. $\lambda_\alpha^n(c)$ is the weight of the pattern α for cell type c. The superscript n is the label of the patient for which the microarray data was taken. $G_{i\alpha}^n$ is the weight of gene i in the transcription pattern α.

The purpose of this section is to demonstrate explicitly that the weights $\lambda_\alpha^n(c)$ for different values of the cell type c (and fixed patient index n) are computed as ensemble averages over the expression levels of transcripts i.

To determine the succession of terms for Equation (S1) all the way to an exact representation, we proceed as follows. Taking the microarray data in the form of expression levels $X_i^n(c)$ for transcript i of cell c, we take the logarithm of each entry. Say that there are A cell types that were measured. We form the A by A symmetric connectivity matrix **C** such that its matrix elements are given by

$$C_{c,c'} = \sum_i \ln(X_i^n(c)) \ln(X_i^n(c')) \quad \text{(S2)}$$

Equation (S2) is a central relation in our discussion. It shows the sense in which the elements $C_{c,c'}$ of the matrix **C** are computed as a sum over the index i of the transcripts. Technically speaking, the matrix **C** is (not quite but almost) the covariance matrix of the cell types where the variation is over the different transcripts.

If we introduce the N by A matrix **Y** where $Y_{ic}^n = \ln(X_i^n(c))$ and N is the number of measured transcripts, we can write the covariance matrix as a matrix product

$$\mathbf{C} = \mathbf{Y}^T \mathbf{Y} \quad \text{(S3)}$$

where the superscript T denotes the transpose of the matrix. We drop the superscript n to simplify the notation, but we emphasize that the matrix C is computed for patient n.

C is a symmetric matrix and can be diagonalized. There are A eigenvectors and eigenvalues. We here discuss the common case where all the eigenvalues of **C** are positive. Then the eigenvalue equations as

$$\mathbf{C}\,\mathbf{P}_\alpha = \omega_\alpha^2\,\mathbf{P}_\alpha \quad , \ \alpha = 0,1,2,..,A-1 \tag{S4}$$

The eigenvectors are normalized by the condition $\sum_{c=0}^{A-1}(P_\alpha^n(c))^2 = 1$. The $\alpha = 0$ term is the base line.

The weights $\lambda_\alpha^n(c)$ of the different phenotypes are computed by an equation analogous to that for $\lambda_{\alpha=0}$, namely we write for each term

$$\lambda_\alpha^n(c) = P_\alpha^n(c)\omega_\alpha^n \tag{S5}$$

Equation (S5) is the desired technical conclusion: The cell type dependence of the weights $\lambda_\alpha^n(c)$ is determined, see Equation (S2), by averaging the transcription fold levels over all the transcripts.

2. The Disease Pattern as an Ensemble Average

Two clear disease patterns, healthy vs. diseased and benign vs. cancer, emerge from the analysis of a cohort of patients. Individuality patterns emerge from the analysis of a cohort of cell types. The disease pattern (and other biologically meaningful deviations) appears in terms of the dependence of the weight of the pattern for the different cell types. Individuality appears in terms of the dependence of the weight of the pattern for the different patients. However, if we look not at the weights but at the actual transcription pattern deviations, the $G_{i\alpha}^n$'s , they are not well correlated and more often than not are only poorly correlated, as shown in Tables S1 and S2 below. Why does averaging clearly bring out the ensemble properties?

We discuss a simple toy model that both mathematically and intuitively exhibits the properties we wish to demonstrate. It is a toy model which means that it is just a caricature of the very very much more complex biological reality. The tradeoff is that it is rather clear how an ensemble can exhibit average properties that are not easily discerned in any of its individual members.

Consider tossing a die. It is a regular die with six faces labeled 1 to 6. We make a list of the number showing up upon each toss. We consider an experiment that is 100 tosses. It is certain that in any experiment the frequency of the different faces coming up will not be the same. (It just cannot be the same, 6 × 15 = 96, 6 × 16 = 102). The multinomial distribution can readily be manipulated to show that when we average over all possible experiments the average frequency of any one face is exactly 1/6. The theorem states that the *average* frequency exactly equals the probability. The theorem holds for any number of tosses and, as we shortly prove, the theorem also holds even if the die is biased. Conclusion: A result that does not hold for any particular experiment holds exactly for the ensemble average.

Lastly, we make a partial average as follows. Do not ask for the value of the face, ask only if it is even or odd. In other words, lump faces 1, 3, and 5 together and ditto for faces 2, 4, and 6. In any one experiment (with one experiment representing a hundred tosses) the frequency of even faces showing up will be quite close to 1/2. This is so even though the frequency of individual faces can be not very similar. Conclusion: a partial average goes some significant way towards recovering an ensemble average.

It is also possible to lump faces 1, 2, and 3 together and ditto for faces 4, 5, and 6. Same conclusion.

To prove the theorem for any number, N, of tosses we consider a biased die where the probability of showing face i, i = 1,2,..6, is p_i. For an unbiased die $p_i = 1/6$. The multinomial distribution states that the probability to perform an experiment where face i is recorded N_i times, i = 1,2,..6, is

$$P(N_1, N_2, .., N_6) = \frac{N!}{N_1! N_2! ... N_6!} \prod_{i=1}^{6} p_i^{N_i}$$

Note that this is already a partial average. The probability of one particular set of N tosses where face i is recorded N_i times, i = 1,2,..6, is

$$\prod_{i=1}^{6} p_i^{N_i}$$

There are

$$W(N_1, N_2, .., N_6) = \frac{N!}{N_1! N_2! ... N_6!}$$

different possible experiments all of which have face i show up N_i times, i = 1,2,..6. These experiments differ in the order of the faces that show up.

Multiplying the probability of a particular experiment by the number of equivalent experiments we obtain $P(N_1, N_2, .., N_6)$.

The mean or expected number of face i to show up

$$\langle N_i \rangle = \sum_{\{N_1, N_2, .., N_6\}} N_i P(N_1, N_2, .., N_6)$$

is the sum over all sets of numbers $\{N_1, N_2, .., N_6\}$ such that their sum is N. To compute this number note that by the binomial theorem

$$\sum_{\{N_1, N_2, .., N_6\}} P(N_1, N_2, .., N_6) = (p_1 + p_2 + .. + p_6)^N$$

Taking the derivative of both sides with respect to p_i we have

$$\sum_{\{N_1, N_2, .., N_6\}} \frac{N_i}{p_i} P(N_1, N_2, .., N_6) = N(p_1 + p_2 + .. + p_6)^{N-1} = N$$

Or

$$\boxed{\langle N_i \rangle = N p_i}$$

If all faces are equally probable the proof will yield $\langle N_i \rangle = N/6$. For any N that is not a multiple of 6 we have that $\langle N_i \rangle$ is not an integer even though any observed value of N_i must be integers.

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.

Cell Phenotype	Eigenvalue of $\mathbf{T}^{(2)}$	Eigenvalues of Data Averaged over Patients
1	4607.10	1890.94
2	157.30	54.19
3	92.46	19.60
4	74.59	12.71

Table S2. Gene analysis of the cell phenotype.

Kegg Pathways	***p* Value**
1. Intestinal immune network for IgA production	1.9×10^{-3}
2. Type I diabetes mellitus	5.2×10^{-3}
3. Antigen processing and presentation	6.9×10^{-3}
4. Asthma	7.3×10^{-3}
5. Small cell lung cancer	7.4×10^{-3}
Gene Ontology	***p* Value**
1. Regulation of apoptosis/anti-apoptosis	2.0×10^{-7}
2. Response to organic substance	6.0×10^{-7}
3. Response to hypoxia/oxidative stress	2.6×10^{-6}
4. Adaptive immune response	1.0×10^{-5}
5. Lymphocyte mediated immunity	3.1×10^{-5}
Functional Annotation Chart	***p* Value**
1. Immune response	2.4×10^{-5}
2. Tumor suppressor	6.8×10^{-4}
3. Disease mutation	1.1×10^{-3}
4. Proto-oncogene	1.3×10^{-3}
5. Innate immunity	1.3×10^{-3}