Surprisal Analysis

To study phenotype changes in carcinogenesis

August 21, 2020

Introduction

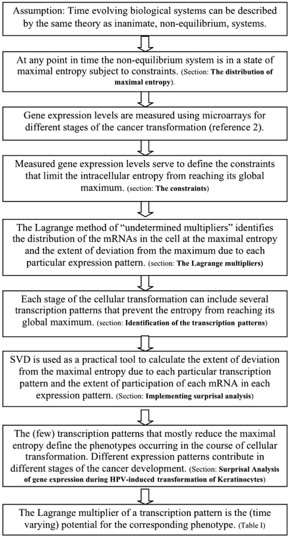
What is surprisal analysis?

Surprisal analysis is an information-theoretical analysis technique that integrates and applies principles of thermodynamics and maximal entropy. Surprisal analysis is capable of relating the underlying microscopic properties to the macroscopic bulk properties of a system.It has already been applied to a spectrum of disciplines including engineering, physics, chemistry and biomedical engineering. Recently, it has been extended to characterize the state of living cells, specifically monitoring and characterizing biological processes in real time using transcriptional data.

Surprisal analysis applied to gene expression

Surprisal Analysis provides a very compact representation for the measured expression levels of many thousands of mRNAs in terms of very few - three, four - transcription patterns.A pattern is a set of genes whose expression levels do not vary with time, as in the steady state, or that vary in concert (i.e., all the transcription levels are time-dependent in the same way).The patterns can be assigned definite biological phenotypic role.Surprisal analysis is hence used to characterize the transcription patterns underlying the process of oncogenic transformation and other changes in biological processes.It seeks to uncover both the changes in expression patterns of different networks and the significance of each altered network in the establishment of a particular phenotype.

Process Outline



Assumption: Time evolving biological systems can be described by the same theory as inanimate, non-equilibrium, systems.

At any point in time the non-equilibrium system is in a state of maximal entropy subject to constraints. (Section: **The distribution of maximal entropy**).

Gene expression levels are measured using microarrays for different stages of the cancer transformation (reference 2).

Measured gene expression levels serve to define the constraints that limit the intracellular entropy from reaching its global maximum. (section: **The constraints**)

The Lagrange method of "undetermined multipliers" identifies the distribution of the mRNAs in the cell at the maximal entropy and the extent of deviation from the maximum due to each particular expression pattern. (section: **The Lagrange multipliers**)

Each stage of the cellular transformation can include several transcription patterns that prevent the entropy from reaching its global maximum. (section: **Identification of the transcription patterns**)

SVD is used as a practical tool to calculate the extent of deviation from the maximal entropy due to each particular transcription pattern and the extent of participation of each mRNA in each expression pattern. (Section: **Implementing surprisal analysis**)

The (few) transcription patterns that mostly reduce the maximal entropy define the phenotypes occurring in the course of cellular transformation. Different expression patterns contribute in different stages of the cancer development. (Section: **Surprisal Analysis of gene expression during HPV-induced transformation of Keratinocytes**)

The Lagrange multiplier of a transcription pattern is the (time varying) potential for the corresponding phenotype. (Table I)

Source: Remacle F, Kravchenko-Balasha N, Levitzki A, Levine RD (2010) Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. Proc Natl Acad Sci USA 107(22):10324–10329

The Method

The main equation:

$$\ln X_i(t) = \ln X_i^o - \sum_{\alpha=1} \lambda_\alpha(t) G_{\alpha i} \quad (1)$$

$$= - \sum_{\alpha=0} \lambda_\alpha(t) G_{\alpha i} \quad (2)$$

$$\ln X_i^o = -\lambda_0 G_{0i} \quad (3)$$

where $X_i(t)$ is the expression level of m-RNA transcript i at time t and X_i^o is its expression level in the steady or balanced state. $\lambda_\alpha(t)$ is the Lagrange multiplier of constraint α at time t. It is the potential that equals the weight of the phenotype α at time t. $G_{\alpha i}$ represents the weight of the gene i in the transcription pattern α, or how much transcript i contributes to the constraint α.

Source: Remacle F, Kravchenko-Balasha N, Levitzki A, Levine RD (2010) Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. Proc Natl Acad Sci USA 107(22):10324–10329

The Method

$$Surprisal = -\ln[\frac{X_i(t)}{X_i^o}] \tag{4}$$

$$= -\sum_{\alpha=1} \lambda_\alpha(t) G_{\alpha i} \tag{5}$$

The fold difference is known as the surprisal. Surprisal analysis is the act of fitting of the surprisal by a sum of terms as shown in equation 2. SVD can be used to approximate this sum as a sum with possibly fewer terms:

$$\ln X_i(t) \cong -\sum_{\alpha=0}^{A} \lambda_\alpha(t) G_{\alpha i} \tag{6}$$

The lowest value of A that provides a good approximation for $\ln X_i(t)$, i and T given, is the number of relevant constraints needed to account for the gene expression data at that time.

Source:Gross A, Levine RD (2013) Surprisal analysis of transcripts expression levels in the presence of noise: A reliable determination of the onset of a tumor phenotype, PLoS One 8(4):e61554

Using SVD to compute the constraints

Singular Value Decomposition is a mathematical procedure that extends the notion of a diagonalization to any matrix, square or not. Then, in addition, SVD generalizes the notion of ordering the set of eigenvalues so as to obtain a series of successively better approximations to the original matrix. When all the eigenvalues are used the SVD approximation becomes an exact reproduction.

SVD is applied on the I X T rectangular matrix Y, whose entries are the logarithms of the transcript expression levels $Y_{iT} = \ln X_i(t_T)$. I is the number of transcripts and T is the number of time points measured, where in general I»T. Y is very much not a square matrix.

Source: Remacle F, Kravchenko-Balasha N, Levitzki A, Levine RD (2010) Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. Proc Natl Acad Sci USA 107(22):10324–10329

Using SVD to compute the constraints

A T X T matrix Y^TY is constructed. The rank r of this matrix is the number of eigenvectors of this matrix with non-zero eigen values. The rank r represents the number of relevant constraints. In general $1 \leq r \leq T$. The equation that determines the constraints is:

$$Y^TYP_\alpha = \omega_\alpha^2 P_\alpha, \alpha = 0, 1, 2, ..., T-1 \tag{7}$$

where ω_α^2 are the eigen values and P_α are the eigen vectors of the matrix Y^TY. The Lagrange multipliers are given by:

$$\lambda_\alpha(t_T) = \omega_\alpha P_{\alpha T} \tag{8}$$

where $P_{\alpha T}$ is the T'th component of the eigen vector P_α. Similarly, $G_{\alpha i}$'s can be computed from the I X I matrix YY^T using the equation:

$$YY^TG_\alpha = \omega_\alpha^2 G_\alpha, \alpha = 0, 1, 2, ..., T-1 \tag{9}$$

To determine the few eigen vectors G_α of interest, the following equation can be used:

$$G_\alpha = \omega_\alpha^{-1} YP_\alpha \tag{10}$$

Source: Remacle F, Kravchenko-Balasha N, Levitzki A, Levine RD (2010) Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. Proc Natl Acad Sci USA 107(22):10324–10329

Zeroth multiplier and the zeroth genotype

$$X_i^o = e^{-\lambda_0(t) G_{0i}} \tag{11}$$

Equation 11 defines a zeroth genotype G_{0i} that represents the gene expression at the global maximum of entropy.The zeroth multiplier $\lambda_0(t)$ should not depend on the time point of the measurement. In practice,$\lambda_0(t)$ is allowed to vary freely but it comes out to be constant at different times, which serves as a numerical validation of the analysis of the data.$\lambda_0(t) G_{0i}$ is the free energy(in thermal units,RT) of transcript i. Consequently, transcripts with the lowest i.e. most negative free energy are those with the highest levels in the balance state. m-RNAs identified with the lowest free energy are those most correlated with cellular networks that regulate and maintain cellular homeostatis machinery.

Implementing surprisal analysis

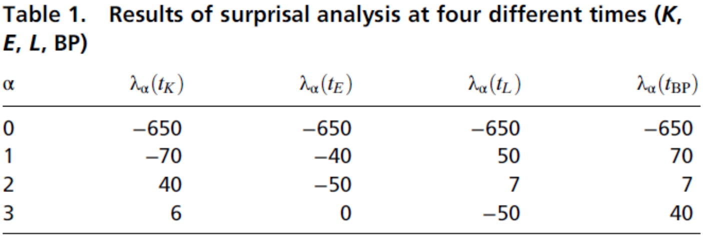


Table 1. Results of surprisal analysis at four different times (*K*, *E*, *L*, BP)

α	$\lambda_\alpha(t_K)$	$\lambda_\alpha(t_E)$	$\lambda_\alpha(t_L)$	$\lambda_\alpha(t_{\mathrm{BP}})$
0	−650	−650	−650	−650
1	−70	−40	50	70
2	40	−50	7	7
3	6	0	−50	40

Figure: Lagrange multipliers identified for 4 different time points in HPV-induced transformation of keratinocytes

Source: Remacle F, Kravchenko-Balasha N, Levitzki A, Levine RD (2010) Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. Proc Natl Acad Sci USA 107(22):10324–10329

Reproduced results

Lagrange multipliers identified for 4 different time points(without rounding) in HPV-induced transformation of keratinocytes:

α	$\lambda_\alpha(t_K)$	$\lambda_\alpha(t_E)$	$\lambda_\alpha(t_L)$	$\lambda_\alpha(t_{BP})$
0	-664.33	-659.3	-645.9	-648.04
1	-76.53	-41.97	49.77	71.54
2	38.62	-51.65	6.43	6.54
3	4.68	2.63	-46.56	38.94

Implementing surprisal analysis

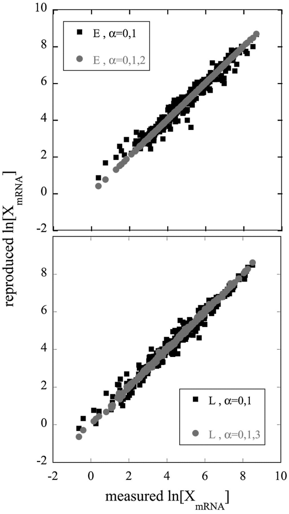
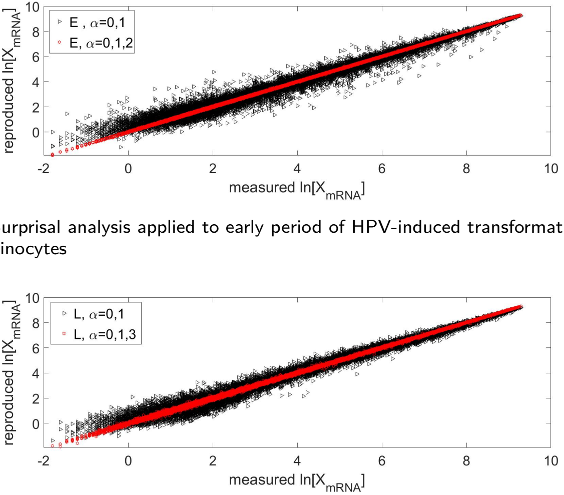


Figure: Surprisal analysis applied to HPV-induced transformation of keratinocytes

Fig. Source: Remacle F, Kravchenko-Balasha N, Levitzki A, Levine RD (2010) Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. Proc Natl Acad Sci USA 107(22):10324–10329

Reproduced results

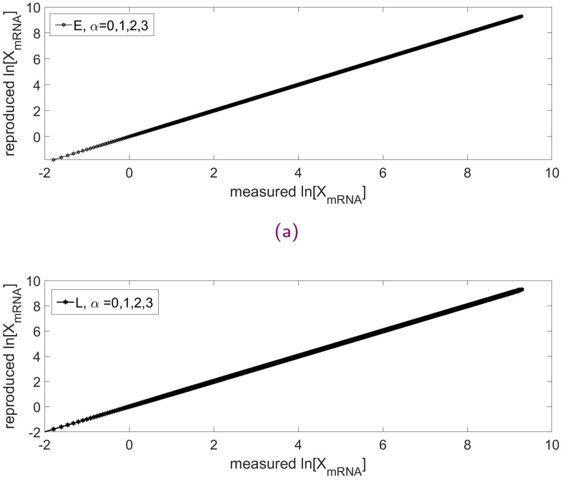


(a) Surprisal analysis applied to early period of HPV-induced transformation of keratinocytes

(b) Surprisal analysis applied to late period of HPV-induced transformation of keratinocytes

Figure

Reproduced results

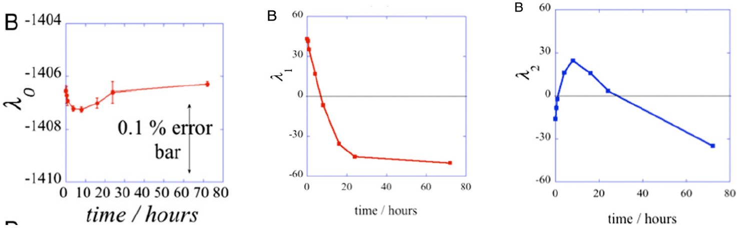


(a)

(b)

Figure: Surprisal analysis applied to early and late periods of HPV-induced transformation of keratinocytes with all transcription patterns included gives an exact fit

Surprisal analysis applied to EMT

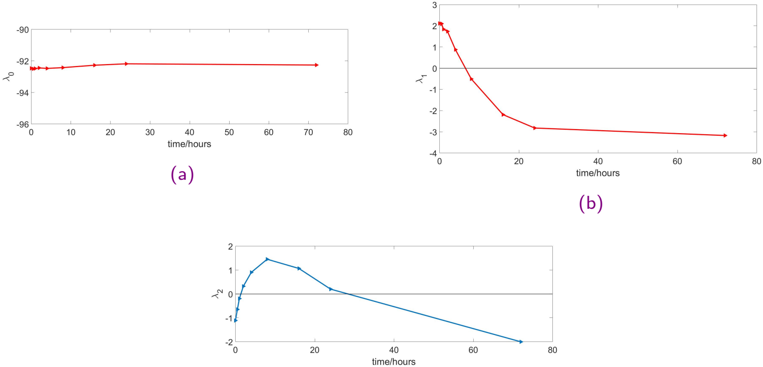


(a) (b) (c)

Figure: Two major phenotypes identified during surprisal analysis of m-RNA time course data for TGF-β1–Treated A549 Human Lung Cancer Cells.The time dependence of the second state variable, $\lambda_2(t)$, exhibits a change in sign (c) that coincides with the epithelial state transitioning (at about 2 h) to what we call a "maturation state." This state is a state of low free energy. The maturation state is followed by a second change (at about 24 h) out of the maturation stage and into the mesenchymal state.

Fig. Source:Zadran S, Arumugam R, Herschman H, Phelps ME and Levine RD (2014) Surprisal analysis characterizes the free energy time course of cancer cells undergoing epithelial-to-mesenchymal transition. Proc Natl Acad Sci U S A 111, 13235– 13240.

Reproduced results



(a)

(b)

(c)

Figure: Two major transcription patterns identified during surprisal analysis of m-RNA time course data for TGF-β1–Treated A549 Human Lung Cancer Cells.

Surprisal analysis applied to EMT

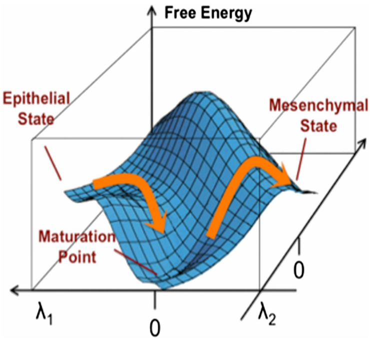


Figure: The free energy landscape for the TGF-β1–induced EMT in A549 cells. At every point in time, the energetic state of the cell undergoing the EMT may be characterized by two time-dependent state variables, $\lambda_1(t)$ and $\lambda_2(t)$. The path of the transition is along the orange arrows and proceeds through a stable intermediate, the maturation point. Energy is released in transition from the epithelial state to the maturation point and consumed from the maturation point to the higher-in-energy mesenchymal state.

Fig. Source:Zadran S, Arumugam R, Herschman H, Phelps ME and Levine RD (2014) Surprisal analysis characterizes the free energy time course of cancer cells undergoing epithelial-to-mesenchymal transition. Proc Natl Acad Sci U S A 111, 13235– 13240.

Surprisal analysis applied to EMT

Gene ontogeny (GO) analysis is applied separately to each of the phenotypes that surprisal analysis identifies as relevant for the transition and to the gene list of the balance state. GO analysis of the second constraint suggests that the critical cellular functions that dominate this phenotypic state are functions involved in receptor binding mechanisms. We suggest that this phenotype integrates information from the cell environment, through receptors and subsequent ligand-dependent signaling pathways, into its cellular decision-making processes. It is expected that during an EMT, the ability of cells to assimilate environmental cues is critical as they take on a mesenchymal-like state and acquire migratory properties.

Surprisal analysis to study robustness of steady state

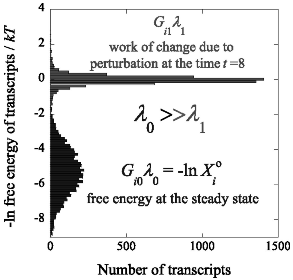


Figure: Histogram of free energy values in units of the thermal energy kT of the transcripts at the steady-state $G_{0i}\lambda_0$ and the work done by the major transcription pattern (=1, $G_{1i}\lambda_1$) at time point 8 in the trajectory 15781012 (in the middle of the transformation process) for the WI-38 cancer model. The values of $G_{0i}\lambda_0$ and $G_{1i}\lambda_1$ are distributed in a bell-shaped manner around a finite negative number and around zero, respectively. Examination of the $G_{0i}\lambda_0$ and $G_{\alpha i}\lambda_\alpha$ values of different trajectories in the WI-38 system revealed that for most transcripts the values of the free energy $G_{0i}\lambda_0$ are lower than the free energy changes, $G_{i\alpha=1}\lambda_{\alpha=1}$,because of the main pattern of the disease. This suggests that the steady state is very stable and that the process of transformation changes the free energy balance of the steady state only slightly.

Source: Kravchenko-Balasha N, et al. (2012) On a fundamental structure of gene networks in living cells. Proc Natl Acad Sci USA 109(12):4702–4707

Robustness of the steady state

The essence of the robustness of the steady state is the inequality in the weights of the transcripts:

$$\lambda_0 >> \lambda_1, \lambda_2, ... \qquad (12)$$

The separation of fold magnitudes implies that the free energy of the transcripts in the steady state is more than an order of magnitude larger than the perturbations of the free energy of the transcripts caused by disease.Robustness also has an implication for the expression levels of particular genes.The weight of the steady transcription pattern is significantly larger than the weights of the disease-induced patterns. The second manifestation applies at the level of the individual transcripts and defines the set of more durable core transcripts that have the lowest of the low G_{i0} values. For such exceptionally stable transcripts i one has the inequality $G_{i0}\lambda_0 >> G_{i\alpha}\lambda_\alpha, \alpha = 1, 2, \ldots$. On the other hand, for steady-state transcripts with a weight near zero (i.e, transcripts that have higher free energy), the change in the free energy caused by the disease can be significant.Depending on the sign of the work done by the disease, $G_{i\alpha}\lambda_\alpha$, the fold change in these more adaptable transcripts can either destabilize the transcript completely or make it significantly more stable.

Source: Kravchenko-Balasha N, et al. (2012) On a fundamental structure of gene networks in living cells. Proc Natl Acad Sci USA 109(12):4702–4707

Most stable transcripts(lowest free energies) are strongly connected

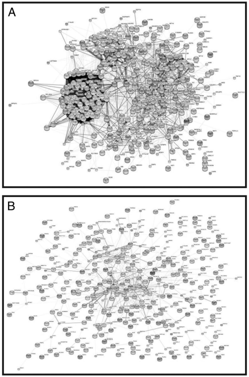


Figure: Examination of the experimentally determined networks obtained from the G_{0i} values in the trajectory 15781012. (A) The 66 transcripts with the lowest (i.e., most stable) G_{0i} values (less than -0.018) are most strongly connected. (B) The 354 transcripts with the highest G_{0i} values (greater than -0.008) are sparsely connected.

Source: Kravchenko-Balasha N, et al. (2012) On a fundamental structure of gene networks in living cells. Proc Natl Acad Sci USA 109(12):4702–4707

Error estimate in surprisal analysis to act as cancer forewarning

The weight of each transcription pattern is determined by surprisal analysis. The weight of this pattern changes with time; it is never strictly zero but it is very low at early times and then rises rather suddenly. It is suggested that the low weights at early time points are primarily due to experimental noise. A necessary formalism is developed to determine at what point in time the value of that pattern becomes reliable. Beyond the point in time when a pattern is deemed reliable the data shows that the pattern remains reliable.This allows a determination of the presence of a cancer forewarning. Similarly, the transcription patterns that account for healthy cell pathways, such as apoptosis, need to be switched off in cancer cells and their weight eventually falls below the threshold. The principle of error estimation in surprisal analysis is determining how many terms need to be included in the sum to fit the surprisal in equation 2.

Source: Gross A, Levine RD (2013) Surprisal analysis of transcripts expression levels in the presence of noise: A reliable determination of the onset of a tumor phenotype. PLoS One 8(4):e61554.

Error estimate in surprisal analysis

At each time t the importance of each term in the sum in the surprisal is determined by the value of the Lagrange multiplier $\lambda_\alpha(t)$ at that time.If the value of the Lagrange multiplier is zero, $\lambda_\alpha(t) = 0$, then that term is unimportant at that time and can be omitted.It means that the constraint α does not lower the entropy at that time. In other words, at the time t constraint α does not provide information on the state of the transcription system.In the presence of noise, when there is an error range associated with each Lagrange multiplier, a Lagrange multiplier provides no new information when zero is a possible value. If $\delta\lambda_\alpha(t)$ is the error range of the Lagrange multiplier for pattern α at the time t, then it is not informative at that time if:

$$\delta\lambda_\alpha(t) \geq \lambda_\alpha(t) \tag{13}$$

Source: Gross A, Levine RD (2013) Surprisal analysis of transcripts expression levels in the presence of noise: A reliable determination of the onset of a tumor phenotype. PLoS One 8(4):e61554.

Constraints significant at later times act as cancer signature pattern

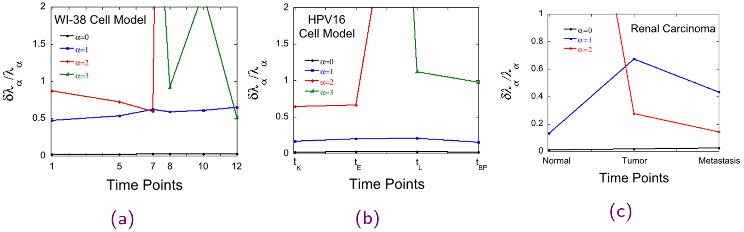


Figure: a) $\alpha = 3$ is the tumour signature and it is seen that it is only valid in later times but well before the cell is cancerous that is observed at time point 12. b) Reliability of weights of phenotypes in tumorigenesis c) Reliability of weights of phenotypes in a renal cancer patient

Fig. Source: Gross A, Levine RD (2013) Surprisal analysis of transcripts expression levels in the presence of noise: A reliable determination of the onset of a tumor phenotype. PLoS One 8(4):e61554.

The End