Information-theoretic analysis of phenotype changes in early stages of carcinogenesis

F. Remacle[a,b,1], Nataly Kravchenko-Balasha[c,1], Alexander Levitzki[c,2], and R. D. Levine[b,d,2]

[a]Fonds National de la Recherche Scientifique, Département de Chimie, Université de Liège, B4000 Liège, Belgium; [b]The Fritz Haber Research Center for Molecular Dynamics, Institute of Chemistry, Hebrew University of Jerusalem, Jerusalem 91904, Israel; [c]Unit of Cellular Signaling, Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, Hebrew University of Jerusalem, Jerusalem 91904, Israel; and [d]Crump Institute for Molecular Imaging and Department of Molecular and Medical Pharmacology, University of California, Los Angeles, CA 90095

Contributed by R. D. Levine, April 20, 2010 (sent for review March 3, 2010)

Cancer is a multistep process characterized by altered signal transduction, cell growth, and metabolism. To identify such processes in early carcinogenesis we use an information theoretic approach to characterize gene expression quantified as mRNA levels in primary keratinocytes (*K*) and human papillomavirus 16 (HPV16)-transformed keratinocytes (HF1 cells) from early (*E*) and late (*L*) passages and from benzo(*a*)pyrene-treated (BP) L cells. Our starting point is that biological signaling processes are subjected to the same quantitative laws as inanimate, nonequilibrium chemical systems. Environmental and genomic constraints thereby limit the maximal thermodynamic entropy that the biological system can reach. The procedure uncovers the changes in gene expression patterns in different networks and defines the significance of each altered network in the establishment of a particular phenotype. The development of transformed HF1 cells is shown to be represented by one major transcription pattern that is important at all times. Two minor transcription patterns are also identified, one that contributes at early times and a distinguishably different pattern that contributes at later times. All three transcription patterns defined by our analysis were validated by gene expression values and biochemical means. The major transcription pattern includes reduced transcripts participating in the apoptotic network and enhanced transcripts participating in cell cycle, glycolysis, and oxidative phosphorylation. The two minor patterns identify genes that are mainly involved in lipid or carbohydrate metabolism.

microarray analysis | oncogenic transformation | surprisal analysis | maximal entropy | gene transcription patterns

Gene expression profiling describes the transcription patterns of thousands of mRNAs at the same time point, allowing insight into or comparison of different cellular conditions. Regulation of gene expression is relevant to many areas of biology and medicine, including the study of different diseases and specifically cancer. To cope with the massive amount of available microarray data [see, for example, the Gene Expression Omnibus (GEO) database], many software packages have been developed (1). These techniques identify a list of "interesting" genes and search for their biological relevance. The techniques used for analysis of microarray data can identify networks that have been changed at each condition. However, it is not possible to delineate the significance of such overall changes to the different transcription patterns that are associated with the different phenotypes. We here propose and apply a physically motivated global method of gene expression analysis that seeks to uncover both the changes in expression patterns of different networks and the significance of each altered network in the establishment of each particular phenotype.

Cancer is an evolving, complex system, which goes through several stages before full malignancy. To demonstrate the application of our method we compare gene expression between different stages of human papillomavirus (HPV)16-induced transformation of keratinocytes. The levels of >22,000 gene probes have been monitored at four stages of early cancer development using the HG-U133A2 array (2). We validated the conclusions of the present study by comparison with the results of bioinformatics and biochemical analysis (2).

Our theoretical approach is based on the proposition that the process of gene expression and signal transduction is subject to the same quantitative laws as inanimate nonequilibrium systems in physics and chemistry. For some time we have explored the application of the procedure of maximal entropy to describe such systems (3).This approach, called "surprisal analysis," was initially introduced to characterize the selectivity of energy requirements and specificity of energy disposal in elementary chemical reactions (4, 5). We use the same method here considering mRNAs, proteins, and metabolites present inside the cell as large molecules and therefore subject to thermodynamic-like considerations.

The theoretical analysis is summarized in Scheme 1 and in the section *Summary of Principles of Data Analysis*. A summary of the terminology is given in Scheme 2.

Distribution of Maximal Entropy

In chemical equilibrium and disequilibrium the system contains several species. The value of entropy depends on the distribution of species and on the internal structure of each. The molecules of chemistry have a structure. The quantum state is the most refined technical expression of the structure of the molecule. In other words the entropy of the system is determined by the distribution of species *and* by the distribution of their quantum states. So even a single species has an entropy associated with its spreading over quantum states. The multitude of the possible quantum states is much more extreme in biology both because the biomolecules have a wider amplitude of structural fluctuations and because they have a large number of low energy conformers. See ref. 6 for a promising approach in that direction that is applied to the toggle switch. When we deal with a mixture of species we therefore talk about two contributions to the entropy. One term, sometimes called the entropy of mixing, is the entropy of a mixture of species. The other contribution is the weighted sum of the entropy of the distribution of quantum states for each of the different species. The weight is the number of molecules of that species.

As in physics and chemistry we introduce the entropy from the fundamental axiom that at the global maximum of the physical entropy all quantum states are equally probable. We then use the grouping property (*SI Appendix*) to assign a weight to each mRNA, a weight that is the number of effectively occupied quantum states. The view of the partition function as the effective number of populated quantum states (*SI Appendix*) allows the

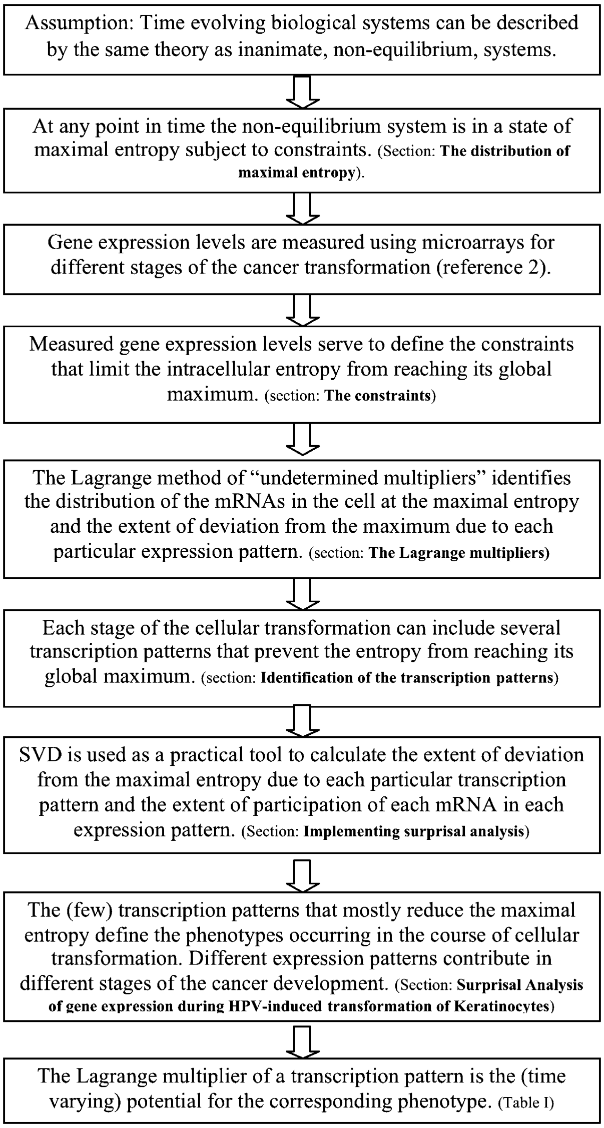
Author contributions: F.R. and R.D.L. designed research; F.R., N.K.-B., and R.D.L. performed research; F.R., N.K.-B., A.L., and R.D.L. analyzed data; and F.R., N.K.-B., A.L., and R.D.L. wrote the paper.

The authors declare no conflict of interest.

[1]F.R. and N.K.-B. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: levitzki@vms.huji.ac.il or rafi@fh.huji.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1005283107/-/DCSupplemental.



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)

Scheme 1. Outline of the theoretical procedure with references to the sections of the text (see also Scheme 2).

introduction of a simple practical working expression for the entropy of the system composed of many species. Methods for computing, or at least estimating, the partition function were developed and very successfully applied [the book by Mayer and Mayer (7) is a good representation of this state of knowledge]. To use here the maximum entropy formalism as a practical tool for a cell at room temperature we need the partition function for the biomolecules. At the present state of knowledge this function is not available. It is a central feature of the information-theoretic surprisal analysis as applied in this study that it uses the mRNA array data to determine the effective number of populated quantum states for each transcript. In the cellular gene expression system there are many other actors besides the mRNAs. In this study we include in the physical entropy only the distribution of the mRNAs and their states.

Scheme 2. A summary of the terminology

Quantity	Maximal entropy formalism	Biological context
α	Labels the constraints and their associated Lagrange multipliers	Labels the transcription patterns and their associated phenotypes
$G_{\alpha i}$	The value of the constraint α in the state *i*	The weight of gene *i* in the transcription pattern α
$\lambda_\alpha(t)$	The Lagrange multiplier equals the weight of constraint α at time *t*	The potential equals the weight of phenotype α at time *t*

Constraints

The entropy of a system tends to its global maximum. Nonetheless the entropy cannot increase as much as it could because the system is subject to a constraint or several constraints. As an example imagine a gas in a cylinder that is confined by a piston to a fraction of the available volume. If the constraint is removed, the gas will expand to fill the entire volume. The entropy of the gas will thereby increase. This is a general property: Removing a constraint allows the entropy to increase. This increase is a general property because we are seeking a maximum of the entropy over all possible distributions. Removing a constraint enlarges the set of distributions over which we search for the maximum. The search can either deliver a higher maximum or not change the value of the already known maximum. The method that we follow determines the constraints that prevent the entropy from increasing as much as it could. In other words, a system that is not in equilibrium is here considered as a system that is in equilibrium but an equilibrium that is subject to constraints whose weights are time dependent. The time evolution of the system is reflected in the evolving weights of the constraints. From an information theoretic point of view a constraint is as important as the extent to which the entropy would increase if this constraint is removed. The information-theoretic analysis automatically delivers this quantitative measure. In this paper we provide a corresponding biological view of the importance of a constraint. For a development of another physical approach to the nonequilibrium biological systems see ref. 8.

Lagrange Multipliers

We search for a maximum of the entropy in the presence of constraints. Lagrange developed the technique of "undetermined multipliers" as a practical way to perform such a search by the introduction of what we now call the Lagrangian, £ : £ = Entropy $- \sum_\alpha \lambda_\alpha \text{Constraint}_\alpha$. The sum over α is the sum over all of the constraints as required for the analysis of the data. The numbers λ_α are the Lagrange undetermined multipliers. Seeking the unconstrained maximum of £ is the practical route to maximization of the entropy subject to constraints. The multiplier λ_α expresses the extent of deviation from the maximal value of the entropy in a system subject to constraint α. If the value of λ_α equals zero, then that constraint is not relevant. The analysis identifies one or more relevant constraints as dictated by the problem. In this paper we also provide a biological interpretation of the relevant constraints.

We show below that our approach, sometimes called surprisal analysis (3, 4), determines the values of the Lagrange multipliers *and also identifies the constraints* from the given experimental data. When the data are provided at several points in time, the values of the Lagrange multipliers can be different at the different times. The changing value of a Lagrange multiplier reflects the evolution of the cellular phenotype. Some Lagrange multipliers can become effectively zero at later times in the history of the cell whereas others can rise in importance. We write "effectively zero" because the values of the Lagrange multipliers are determined by the real data and real data are often subject to inevitable noise. Therefore we need to determine error bars on the values of the Lagrange multipliers (9). If the error bar spans the value zero, then the constraint is not relevant (10).

Constraints prevent the entropy from reaching its global maximal value. A constraint means that the number X_i of transcript molecules of type *i* is not completely free to vary but is limited by the quantity $\langle G_\alpha \rangle$ being constant, where

$$\langle G_\alpha \rangle = \sum_i G_{\alpha i} X_i \qquad [1]$$

and α is the label of the constraint. The numbers $G_{\alpha i}$ are the value of the constraint α for the transcript i. With a good understanding of the biophysical chemistry one can tell beforehand what the constraints are. Here we determine the $G_{\alpha i}$'s by a surprisal analysis of the transcription profiling. We then seek to explain the biochemical significance of the constraints. The constancy of $\langle G_\alpha \rangle$, Eq. **1**, implies that the X_i 's cannot be varied independently of one another because any variation must be consistent with the quantity $\langle G_\alpha \rangle$ remaining unchanged. Lagrange's method of undetermined multipliers solves this mathematical problem by the introduction of (Lagrange's undetermined) multipliers λ_α, one multiplier for each constraint (Eq. **1**). With the values of the multipliers not yet determined one introduces the Lagrangian £:

$$\begin{aligned} £ &= \text{Entropy} - \sum_{\alpha=1} \lambda_\alpha \langle G_\alpha \rangle \\ &= -\sum_i X_i[\ln(X_i) - 1 - \ln(Q_i)] - \textstyle\sum_{\alpha=1} \lambda_\alpha \langle G_\alpha \rangle. \end{aligned} \qquad [2]$$

In the second line of Eq. **2** we used Eq. S8 of the *SI Appendix* for the expression of the entropy where Q_i is the partition function of transcript i, which acts as the "effective number of quantum states" (*SI Appendix*). To determine an extremum of the Lagrangian one can vary each X_i independently of one another. It is precisely because we use the assumption that at the global maximum entropy all quantum states are equally probable that it is not the case that all genes are equally probable. Rather, at the unconstrained global maximum of the entropy the population of each gene is proportional to its partition function. A summary of the terminology is given in Scheme 2.

Identification of the Transcription Patterns That Determine the Process of Transformation

To quantify the changes occurring during the different stages in the process of transformation we determine first the gene expression levels at a given point in time. As shown in detail in *SI Appendix*, Eq. S16 in particular, the distribution of mRNAs at the maximal entropy is of the form

$$X_i(t) = X_i^{\text{o}} \cdot \underbrace{\exp\left(-\textstyle\sum_{\alpha=1} \lambda_\alpha(t) G_{\alpha i}\right)}_{\text{deviation from the global extremum}}. \qquad [3]$$

The time-independent concentration X_i^{o} is determined by the effective number of quantum states. It is derived in *SI Appendix* as the limit of secular variation, namely, the limit where on the timescale relevant to the experiment the expression levels, the X_i 's, are not changing. The second factor in Eq. **3** is the time-varying part. Our derivation of Eq. **3** shows that the index α lists the constraints that stop the entropy from climbing as high as it could. In this paper α lists the transcription patterns and their corresponding phenotypes. $G_{\alpha i}$ is the time-independent extent of participation of mRNA i in the transcription pattern α. A particular pattern of gene expression is associated with a corresponding phenotype. As shown in *SI Appendix*, the process of seeking the maximal entropy identifies α as enumerating the constraints that govern the time evolution of the system. $\lambda_\alpha(t)$ is the Lagrange multiplier that is conjugate to the constraint α at the time t. $G_{\alpha i}$ is the value of the constraint α for the gene i. $\lambda_\alpha(t)$ is the extent of deviation from the maximal entropy due to the particular gene expression phenotype α. We therefore propose to identify $\lambda_\alpha(t)$ as the measure of the contribution of phenotype α at the time t. We call it "the potential."

To summarize, at the molecular level a particular transcription pattern corresponds to each phenotype (equal to a relevant constraint). For each constraint α we generate the list and the extent, $G_{\alpha i}$, of participation of mRNA i. It is therefore possible to refer to the set of numbers $G_{\alpha i}$as the transcription pattern α. To each such transcription pattern there is a conjugate phenotype with a potential $\lambda_\alpha(t)$. The correspondence between the terminology of the maximum entropy analysis parameters and their biological meaning is provided in Scheme 2.

Summary of Principles of Data Analysis

Gene expression was measured at four discrete time points, called K, E, L, and BP in the data (2). We use the subscript T to enumerate these points in time: $T = K$, the keratinocytes (normal cells untransformed by the papilloma virus); $T = E$, for early (HPV16 transformed cells from early stage of transformation); T = L, for late (transformed cells from late stage of transformation); and T = BP [the cells from the late stage that were treated by benzo(*a*)pyrene]. Because the expression levels are nonnegative numbers, we can take the logarithm of Eq. **3** and then combine the two terms together:

$$\begin{aligned} \ln(X_i(t_T)) &= \ln\left(X_i^{\text{o}}\right) - \sum_{\alpha=1} \lambda_\alpha(t_T) G_{\alpha i} \\ &= -\sum_{\alpha=0} \lambda_\alpha(t_T) G_{\alpha i}. \end{aligned} \qquad [4]$$

To obtain the second line we wrote X_i^{o} as

$$\ln\left(X_i^{\text{o}}\right) = -\lambda_0 G_{0i}. \qquad [5]$$

With this change the logarithm of the expression level is a sum of terms all having the same form as shown in the second line of Eq. **4**.

Eq. **5** defines a "zeroth" genotype G_{0i} that represents the gene expression at the global maximum of the entropy. We provide a practical method to determine all of the values $G_{\alpha i}$ of all of the constraints including the G_{0i} 's. Unlike the other Lagrange multipliers, Eq. **5** specifies that the zeroth multiplier $\lambda_0(t_T)$ should *not* depend on the time t_T of measurement. In practice we do not impose the condition that $\lambda_0(t_T)$ is constant but instead allow it to freely vary between different points in time. The values of λ_0 for different points in time are then examined to see to what extent they are constant. The constancy of the value of $\lambda_0(t_T)$ at different times offers a numerical validation of the analysis of the data.

The procedure of the maximal physical entropy determines a nonnegative (see the exponential form in Eq. **3**) gene expression level, for each gene i and at each time point t_T. The logarithm of the expression level is a sum (over α) of terms reflecting the different phenotypes. The expression level itself, Eq. **6**, is not a sum but a product of terms, one term for each phenotype:

$$X_i(t_T) = X_i^{\text{o}} \prod_{\alpha=1} \exp(\lambda_\alpha(t_T) G_{\alpha i}) = \prod_{\alpha=0} \exp(\lambda_\alpha(t_T) G_{\alpha i}). \qquad [6]$$

This product form highlights the difference of what we do here from statistical data analysis that typically seeks linear models where the data are resolved as a sum of terms.

Implementing Surprisal Analysis

By surprisal analysis we mean fitting the sum of terms as shown on the right-hand side of Eq. **4** to the logarithm of the measured expression level of transcript i at the given time t. This is repeated for every time point. A suitable general method has been published (11) and extensively applied mRNA microarrays monitor the expression levels of thousands of genes at a time. The analysis of the data needs therefore to be handled by numerical procedures that are compatible with such a large dataset. For this purpose we have adapted the well-documented singular value decomposition (SVD) method (12–14) that has been extensively applied in biology. It is important to note that we do *not* apply SVD to the measured gene expression data as is usually done, (see ref. 12 and references therein). We use SVD as a mathematical tool for "diagonalizing" (*SI Appendix*) a not square matrix and thereby to determine the two sets of parameters that are needed in surprisal analysis: the Lagrange multipliers $\lambda_\alpha(t)$ for all α's at a given time point and for all times and the (time-independent) transcription patterns $G_{\alpha i}$ for all transcripts i at each pattern α. We discuss the principle and the key steps in the implementation in *SI Appendix*. We show therein that we can explicitly numerically identify the transcription patterns and determine their potential.

In summary, surprisal analysis represents the data as a sum of terms shown in the second line of Eq. **4** and repeated here: $\ln(X_i(t_T)) = \sum_{\alpha=0}\lambda_\alpha(t_T)G_{\alpha i}$. We use SVD (see *SI Appendix* for mathematical details) to approximate this sum as a sum with possibly fewer terms

$$\ln(X_i(t_T)) \cong -\sum_{\alpha=0}^{A} \lambda_\alpha(t_T) G_{\alpha i} \qquad [7]$$

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

Eq. **7** is the working equation. The logarithm of the expression level of a gene at a given time is a sum over a few types. For each type there is a product of $\lambda_\alpha(t_T)$, the potential of the phenotype α at the time T and of $G_{\alpha i}$, the participation of gene i in the transcription pattern α.

It is important for a system biologist to observe that Eq. **7** exhibits separation of variables: The gene participations $G_{\alpha i}$ are independent of time and the phenotype potentials $\lambda_\alpha(t)$ are independent of the genes. Additional concerns for the theorist about the sense of convergence in Eq. **7** are provided in *SI Appendix*.

Surprisal Analysis of Gene Expression During HPV-Induced Transformation of Keratinocytes

The data analysis follows the requirements of Eq. **7**. We take the input raw data as $X_i(t_T)$, the level of transcript i at time $t = T$. T assumes four values, K, E, L, and BP. We take the (natural) logarithms of the levels of transcripts $\ln(X_i(t_T)) = Y_{iT}$. Using the data reported in ref. 2 and also the raw experimental data for I = 22,277 probes, we applied this analysis as just summarized, and the result for the potentials is reported in Table 1. The data have been measured in triplicates for each time point. The results of the three triplicates are similar but not identical, meaning that the data have some experimental noise. The data that we analyze are the mean of the three triplicates, keeping all 22,277 genes. We also performed the analysis using only those genes that are filtered by the requirement that the variation between the triplicates is small (as judged by a t test) or that the variation between the K and the other time points is large (as judged by a paired t test) and other tests as discussed below. No reasonable filtering altered any of our qualitative conclusions. In summary, the noise in the data allows us to determine the weights only to within an interval of 10%. The results for the four time points in the process of HF1 cells transformation are shown in Table 1.

The global term (α = 0) that theory requires to be independent of time is indeed found to be constant to well within the noise (it is constant to ~1.5%). The major phenotype, α = 1, contributes significantly at all time points. This major phenotype switches its role between the two earlier and the two later time points as indicated by the change in sign between the E and L time points. Finally we identify two more minor phenotypes. The α = 2 type is important at the earlier times and it switches roles between the K cells and the E cells. The α = 2 type has a borderline importance at later times. The α = 3 type is important only at the later times and it switches roles between the L cells and the BP cells.

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

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

Given the transcription patterns $G_{\alpha i}$ as determined from the data and the weights in Table 1, there are three graphical ways to represent our algebraic statements. Fig. 1 shows the quality of the reproduction of the experimental data using just one or using the sum of two phenotypes. The α = 0 term in Eq. **7**, the global reference, is included in both. We show data at early and late times. At either time just two phenotypes and their associated transcription patterns clearly represent the data well. Note that phenotypes α = 1 and 2 are needed for a high-quality fit at time E whereas phenotypes 1 and 3 are needed at time L.

The entries in Table 1 are the values of the potentials, see Scheme 2, at four time points rounded off as discussed, because of the noise in the data. Not rounded values are reported in *SI Appendix*. The zeroth potential should not depend on time; see Eq. **5**. We do not impose this condition and it inherently comes from the analysis of the data.

The other two plots show the value of the (positive) entropy deficiency (15), DS, which is how far the entropy is below its maximal value (when gene expression is at a steady state). Fig. 2 shows DS vs. time in the physical sequence K, E, L, BP. Clearly DS decreases and therefore the entropy increases toward its maximal value during the progression from K to L. This result correlates with the monotonic increase in entropy in the course of time that is expected to occur in a system with a constant level of nutrients and constant temperature. However, in the time point BP we detected the drop in the entropy. It is important to recall that the BP cells were treated by the carcinogen Benzo[a] pyrene that was added from the outside. The system was opened, and therefore the entropy can go down, as it does.

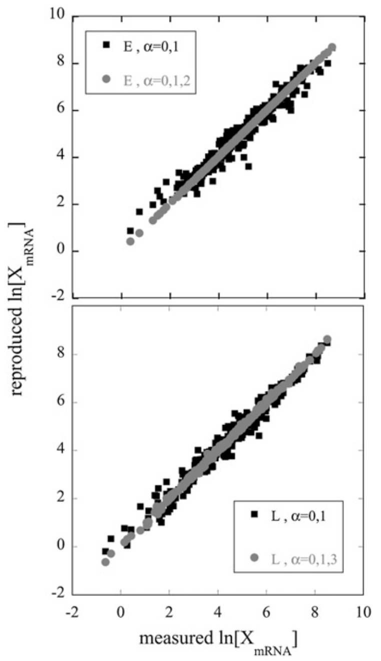


Fig. 1. The quality of reproducing the (logarithm of the) data (22,277 genes), with just one (squares, black on line) and just two (circles, red on line) of the three phenotypes. *Upper*, early; *Lower*, late. A good fit requires including the main phenotype α = 1 and one minor phenotype α = 2 at the *E* time point and α = 1 and the other minor phenotype, α = 3, at the *L* time point. If we represent the (logarithm of the) data as a sum of all three phenotypes (plot not shown), we get a mathematically exact reproduction.

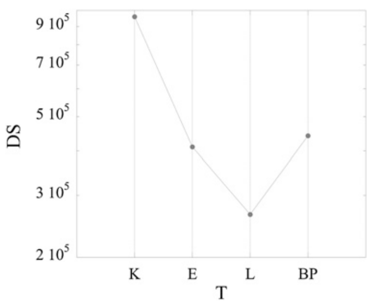


Fig. 2. The entropy deficiency DS, logarithmic scale, vs. time in the physical sequence *K*, *E*, *L*, BP. The increase of DS at the last point is due to the experimental system being open. The value of DS is per the entire cellular transcripts and not per mRNA.

Another plot of DS is in Fig. S1. This view shows how much only the first, only the second, and only the third phenotype contribute to the lowering of the entropy from its maximal value. Evidently the first pattern is the most important one and it is most important at all times. The second pattern is somewhat important at the earlier times (*K* and *E*) but not at the later times and vice versa for the least important third pattern.

In exploring the biological significance of the results of the analysis (see next section), it is interesting to examine the change in the expression level between two time points. From Eq. **2** and noting that , the weight of a transcript in the pattern α is not dependent on time, we calculate that

$$\ln(X_i(t_T)) - \ln(X_i(t_{T'})) = -\sum_{\alpha=1}\left(\lambda_\alpha(t_T) - \lambda_\alpha(t_{T'})\right)G_{\alpha i}. \qquad [8]$$

The sign of $G_{\alpha i}$, the extent of participation of the transcript *i* in the pattern α can be either positive or negative as shown in Fig. 3 in the next section. Therefore in the same phenotype some transcripts in the pattern can be reduced whereas others can be induced and the surprisal analysis predicts which one does what.

Discussion of Phenotypic Changes During HPV-Induced Transformation of Keratinocytes

The theoretical analysis revealed three transcription patterns involved in the process of cellular transformation. This section discusses the identification of groups of transcripts participating in the three phenotypes that correspond to the three transcription patterns.

The main phenotype, as identified by the transcription pattern with the highest weight (Table 1) and by its significant contribution at all stages of transformation, includes a group of transcripts with gradual induction in expression during the transformation, a group with gradual reduction in expression, and a large group of transcripts whose level hardly changes. The range of change in contributions between times *K* and *E* is shown in Fig. 3.

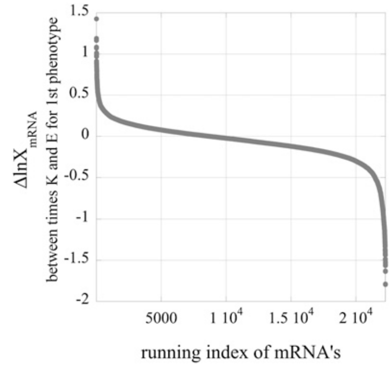


Fig. 3. The changes in the expression level of mRNAs between times *K* and *E* for the first transcription pattern, logarithmic scale. For ease of visualization the values are sorted by descending magnitude of change. See Eq. **8**. The same analysis was performed for other times and for the two other patterns. Those mRNAs that change significantly up or down are sorted as two separate groups and subjected to a bioinformatics analysis.

Using one term from Eq. **8**, we can identify those transcripts that change in phenotype α = 1 with respect to the earliest, $T = K$ time. Keeping only transcripts whose fractional level changes by at least ±0.7, 2,064 transcripts are gradually induced. On the other hand, 2,184 transcripts are repressed compared to the earliest, $T = K$ time. Bioinformatic analysis of the two sets of transcripts led to the remarkable insight that the major phenotype is the same phenotype as the shrinkage in the pathways controlling apoptosis and enhancement in the metabolic and cell cycle networks.

Using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to categorize transcripts that were significantly induced/repressed during the transformation, we previously found contraction of MAPK, TGF-β-Smad4, and JAK-STAT pathways. GO analysis highlighted reduction in the NFκB pathway. On the other hand we identified large groups of induced transcripts participating in glycolysis, TCA cycle, oxidative phosphorylation, and cell cycle (2).

The shrinkage of the signaling pathways participating in regulation of apoptosis and cell survival was validated by detailed biochemical analysis and revealed substantial contraction of both pro- and anti-apoptotic networks during transformation and subsequent switch from apoptotic to necrotic cell death. Also, *L* and BP cells display enhanced growth rate that correlates with increased transcription of the metabolic and cell cycle genes. The theoretical analysis performed here identifies the changes above as the major phenotype, α = 1.

Detailed investigation of the major transcription pattern, α = 1, using KEGG identifies 57 gene products in BP cells with large deviations from the stage K that belong to the MAPK pathway. Of these, the expression of 38 gene products is reduced, compared with 19 that are induced. In the JAK-STAT pathway, 20 genes are reduced in expression, compared with 5 whose expression is increased. Among genes with significantly reduced expression, KEGG analysis identifies 16 members of the TGF-β-Smad4 pathway. Seven members of the TGF-β-Smad4 pathway, as defined by KEGG, show induced expression. Using GO analysis, 34 gene products belonging to the NFκB pathway are induced and 9 reduced.

The surprisal analysis therefore identifies components of at least four major pathways controlling apoptosis that exhibit strongly reduced expression in transformed *L* and BP cells. This contraction of the pathways controlling apoptosis is consistent with our previous conclusion (2) based on a biochemical analysis.

When performed on genes with enhanced expression in the major transcription pattern, KEGG analysis identifies 58 transcripts with high weights participating in the cell cycle network. Forty-six of them are induced during the transformation and only 12 transcripts are reduced. Similarly, we find 28 induced transcripts participating in oxidative phosphorylation and 11 induced transcripts participating in glycolysis compared with 9 reduced transcripts involved in oxidative phosphorylation and 4 in glycolysis.

In summary, the main transcription pattern identified by surprisal analysis is the overall shrinkage in the apoptotic network and enhancement in the cell cycle, as was validated experimentally (2, 16). In addition, the bioinformatic analysis shows that the main phenotype contains transcripts that participate in the metabolism as noted above.

An advantage of surprisal analysis lies in its ability to identify secondary phenotypes that are not necessarily significant at all of the stages of the transformation. The second phenotype, α = 2 (Table 1), exhibits high and opposite contributions to the *K* and *E* cells. Using the same analysis as for phenotype 1, that is, keeping only transcripts that change by at least ±0.7 in deviation from stage *K*, phenotype 2 comprises 510 significant transcripts that are induced in *E* cells in comparison with *K* cells and 562 significant transcripts that are reduced in *E* cells.

KEGG analysis was also applied to those transcripts defining phenotype 2 that undergo a significant change. We identified 16 induced transcripts participating in lipid metabolism of 37 induced metabolic transcripts. Another group of 53 reduced metabolic transcripts included 22 reduced transcripts that participate in lipid metabolism. We also identified induction in Akt3 and reduction in PTEN transcripts in *E* cells. Activation of Akt3 occurs in many forms of cancers, for example, refs. 17 and 18. It has previously been shown that selective activation of Akt3 occurs as a result of increased expression of Akt3 at the mRNA level and decreased PTEN protein activity due to loss or haploinsufficiency (single functional copy) of the PTEN gene (18). The surprisal data analysis indicates that the changes in these two transcripts at the mRNA level could significantly alter the cellular behavior and lead to cell survival and resistance to apoptosis. Indeed, in our previous study we noted that *E* cells were much less sensitive to CDDP treatment than *K*, *L*, and BP cells and hardly underwent apoptotic cell death (2). Moreover, *E* cells had enhanced activity of the PI3K/Akt pathway compared to the *K* cells. Additionally, these cells had elevated ATP levels and enhanced rate of oxygen consumption that might be explained by significant changes in lipid metabolism. The reorganization of the lipid metabolism network, which is identified by the surprisal analysis, points to a modification of the process of fatty acid oxidation and an enhancement of the energy production inside the cell.

In summary, the minor phenotype 2 identified by the theoretical analysis contributes to secondary features and particularly to the lipid metabolism in the behavior of the *E* cells.

The theoretical analysis identifies an additional minor phenotype, $\alpha = 3$, that contributes significantly and in an opposite manner to the *L* and BP cells (Table 1). This phenotype includes ∼560 transcripts that are reduced significantly in *L* cells in comparison with *K* cells and induced in BP cells compared with *L*. The group includes Akt1, NFκB, and 14 transcripts participating in the MAPK pathway. The transcripts of the group that is reduced in BP cells compared with *L* are essentially all within the noise limit. All these pathways contribute to anchorage-independent growth of transformed cells (19–21). As we have shown previously, BP cells are able to form colonies in soft agar. We also noted that the process of cell death after CDDP treatment lasts longer in BP cells than in *L* cells. The beginning of the cell death has a delay of several hours in BP cells in comparison with *L* cells. All these phenomena may be accounted for by the here identified enhanced expression of these pathways in BP cells. We do, however, note that it is in the nature of matrix diagonalization that the third and least weighted eigenvector is the one most susceptible to the effect of experimental noise in the data.

Akt regulates fatty acid biosynthesis through transcriptional factor SREBP1 (22). Phenotype 3 includes a group of 41 metabolic genes that are classified as mainly involved in lipid or carbohydrate metabolism. Nine members of the group were previously reported to be regulated by Akt through SREBP1 (22). Moreover, KEGG analysis of the significant transcripts of surprisal analysis identifies two additional categories: pathways in cancer and the MAPK pathway. These categories include 8 of 33 of the genes known to be regulated by Akt (22). This finding is consistent with the increased expression of the Akt1 gene in BP cells characterized by surprisal analysis. Thus, the surprisal analysis of phenotype 3 provides insight into differences that have been experimentally observed between *L* and BP cells.

Summary

A major question is which altered networks play a central role in different stages of carcinogenesis. High-throughput screening methods were used to follow gene expression patterns in the course of cellular transformation. This study showed that it is possible to determine the contribution of different molecular-level phenotypes (or transcription patterns) to the establishment of particular stages during carcinogenesis. We used an information theoretic approach based on the assumption that biological systems, similarly to inanimate, nonequilibrium systems, tend to the maximal possible entropy subject to environmental and genomic constraints. The distribution of the mRNAs in the cell at the maximal entropy subject to constraints was derived and a method was developed for explicitly identifying the constraints (the transcription patterns and their corresponding phenotypes). The contribution of each gene to a given constraint is determined as is the weight of each constraint at every stage of the cell development. By analysis of experimental data the extent of deviation from the maximal entropy in every stage of transformation was identified as due to specific transcription patterns. The method determined one major and two minor phenotypes whose significance was validated by biochemical means. The major constraint contributes throughout the carcinogenesis. One minor constraint is important in the early stages of the transformation and the other at the late stages.

ACKNOWLEDGMENTS. The experiments of N.K.-B. and A.L. were supported by a grant from Deutsches Krebsforschungszentrum, Heidelberg, Germany.

1. Weniger M, Engelmann JC, Schultz J (2007) Genome Expression Pathway Analysis Tool—analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics* 8:179.
2. Kravchenko-Balasha N, Mizrachy-Schwartz S, Klein S, Levitzki A (2009) Shift from apoptotic to necrotic cell death during human papillomavirus-induced transformation of keratinocytes. *J Biol Chem* 284:11717–11727.
3. Levine RD (1978) Information theory approach to molecular reaction dynamics. *Annu Rev Phys Chem* 29:59–92.
4. Levine RD (2005) *Molecular Reaction Dynamics* (Cambridge Univ Press, Cambridge, UK).
5. Levine RD, Bernstein RB (1974) Energy disposal and energy consumption in elementary chemical-reactions. Information theoretic approach. *Acc Chem Res* 7: 393–400.
6. Kim K-Y, Wang J (2007) Potential energy landscape and robustness of a gene regulatory network: Toggle switch. *PLoS Comput Biol* 3:e60.
7. Mayer JE, Mayer MG (1966) *Statistical Mechanics* (Wiley, New York).
8. Wang J, Xu L, Wang EK (2008) Potential landscape and flux framework of non-equilibrium networks: Robustness, dissipation and coherence of biochemical oscillations. *Proc Natl Acad Sci USA* 105:12271–12276.
9. Alhassid Y, Levine RD (1980) Experimental and inherent uncertainties in the information theoretic approach. *Chem Phys Lett* 73:16–20.
10. Kinsey JL, Levine RD (1979) A performance criterion for information theoretic data analysis. *Chem Phys Lett* 65:413–416.
11. Agmon N, Alhassid Y, Levine RD (1979) An algorithm for finding the distribution of maximal entropy. *J Comput Phys* 30:250–258.
12. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* 97:10101–10106.
13. Wall ME, Rechtsteiner A, Rochas LM (2003) *A Practical Approach to Microarray Data Analysis*, eds Berrar DP, Dubitzky W, Granzow M (Kluwer, Norwell, MA), pp 91–109.
14. Holter NS, et al. (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc Natl Acad Sci USA* 97:8409–8414.
15. Bernstein RB, Levine RD (1972) Entropy and chemical change. 1. Characterization of product (and reactant) energy distributions in reactive molecular collisions: Information and entropy deficiency. *J Chem Phys* 57:434–449.
16. Mizrachy-Schwartz S, Kravchenko-Balasha N, Ben-Bassat H, Klein S, Levitzki A (2007) Optimization of energy-consuming pathways towards rapid growth in HPV-transformed cells. *PLoS ONE* 2:e628.
17. Cristiano BE, et al. (2006) A specific role for AKT3 in the genesis of ovarian cancer through modulation of G(2)-M phase transition. *Cancer Res* 66:11718–11725.
18. Stahl JM, et al. (2004) Deregulated Akt3 activity promotes development of malignant melanoma. *Cancer Res* 64:7002–7010.
19. Ishino K, et al. (2002) Enhancement of anchorage-independent growth of human pancreatic carcinoma MIA PaCa-2 cells by signaling from protein kinase C to mitogen-activated protein kinase. *Mol Carcinog* 34:180–186.
20. Liu X, et al. (2001) Downregulation of Akt1 inhibits anchorage-independent cell growth and induces apoptosis in cancer cells. *Neoplasia* 3:278–286.
21. Zahir N, et al. (2003) Autocrine laminin-5 ligates alpha6beta4 integrin and activates RAC and NFkappaB to mediate anchorage-independent survival of mammary tumors. *J Cell Biol* 163:1397–1407.
22. Rome S, et al. (2008) Microarray analyses of SREBP-1a and SREBP-1c target genes identify new regulatory pathways in muscle. *Physiol Genomics* 34:327–337.