Application of Information Theory and Extreme Physical Information to Carcinogenesis

Robert A. Gatenby1 and B. Roy Frieden

Departments of Radiology [R. A. G.], Applied Mathematics [R. A. G.], and Optical Sciences [B. R. F.], University of Arizona, Tucson, Arizona 85724

ABSTRACT

Cellular information dynamics during somatic evolution of the malignant phenotypes are complex and poorly understood. Accumulating, random genetic mutations and, therefore, loss of genomic information appears necessary for carcinogenesis. However, additional control parameters can be inferred because unconstrained mutagenesis would ultimately produce cellular information degradation incompatible with life. Similarly, the stability of some genomic segments, such as those controlling proliferation and metabolism, indicates the presence of selective mutational constraints.

By applying Information Theory and Extreme Physical Information (EPI) analysis, we demonstrate that the phenotypic characteristics and growth pattern of cancer populations are emergent properties resulting from the nonlinear dynamics of accumulating, random genetic mutations and tissue selection factors. Maximum quantitative loss of transgenerational information is demonstrated in genomic segments encoding negative or neutral evolutionary properties. This is most evident in the progressive dedifferentiation observed during carcinogenesis and may terminate in a differentiation “information catastrophe” producing decohherent cellular morphology and function.

INTRODUCTION

The formation of coherent societies of diverse cellular phenotypes from a common inherited genome is dependent on the translation of specific subsets of inherited information and the ability of cells to receive and process information in the environment. The active information content of a cell is a time-dependent summation of translated intracellular and acquired extracellular information. This information state controls the morphology and function of that cell as well as its interaction with the external environment. Abnormalities in cellular information content will result in disease as clearly exemplified by several thousand genetically linked disorders (2).

Less clear is the role of disordered cellular information in cancer. Although carcinogenesis typically requires multiple genetic mutations (3–5), the mechanisms by which this information loss produces the malignant phenotype remain imprecisely defined. This is primarily because of the absence of a prototypical malignant genotype. Cancer cells typically possess hundreds and even thousands of genomic errors, and unique patterns of genetic mutations are found in virtually every different tumor (5, 6). Thus, unlike classical genetic diseases, no single genotype is invariably present in cancer cells, and therefore there is no well-defined correspondence between the genetic mutations present in cancer populations and the cellular characteristics of the malignant phenotype.

Clearly, the large number of genetic mutations found in cancer cells indicates a critical role for genomic information degradation in tumorigenesis. However, although cellular information loss through random accumulating genetic mutations appears necessary for carcinogenesis (7–9), two general observations demonstrate that it is not sufficient:

(a) A constant mutation rate, in the absence of any modifying constraints, will monotonically increase genomic noise while useful information decays until it is quantitatively insufficient to maintain life.

(b) The cancer phenotype exhibits apparent nonrandomness in its genomic degradation such that differentiated cell functions are progressively lost while the gene product necessary for proliferation and energy production remain functional (and often amplified) even in the most advanced cancers.

Thus, cancer, as a “disease of the genes” (10), is a complex disorder of the storage, processing, and propagation of cellular information maintained in the genome. This general class of phenomenon can be addressed through information theory developed by Shannon (11) and Fisher (12) to measure information content and communication. Their seminal contributions have spawned extensive research in the physical sciences (13, 14) and modest application in biology (15).

Both Shannon (11) and Fisher (12) information measures are examples of a more general form of information called cross entropy (Ref. 16; also often called “Kullback-Leibler information”). Although the concepts of Shannon’s information (11) and Fisher’s information (12) that provide the basis for our analyses are but specific cases of one general information measure, we will specifically identify the cross entropy forms as being of the Shannon or Fisher type. We return to this point below.

In this report, we develop mathematical models of cellular and extracellular information dynamics in carcinogenesis. This approach is not meant to be analogous to the linear model of sequential genetic changes in colorectal tumorigenesis described by Fearon and Vogelstein (17). That is, we do not address the role of specific defects in oncogenes and tumor suppressor genes, the sequence of these changes, the causes of tumor genomic instability, or the influence of nonneoplastic cells during carcinogenesis. Rather, we are interested in understanding the dynamic forces including genomic instability and environmental selection factors that control the evolving cellular information content resulting in the malignant phenotype and ultimately invasive cancer growth.

By applying Shannon’s information (11) to carcinogenesis, we...
demonstrate the interactions of stochastic mutation events and environmental selection pressures (18, 19) that produce the cellular characteristics of the malignant phenotype. Following this, we used Fisher’s information to analyze tumor population growth dynamics based on perturbations in environmental information produced by genomic alterations during carcinogenesis. Physical information is the difference between two Fisher information quantities (see below). Using the EPI technique, we accurately predicted tumor growth dynamics in vivo and in vitro, based solely upon changes in the environmental information content.

INTRACELLULAR INFORMATION IN CARCINOGENESIS

The character of information in biological systems can be described as bound or free (20). Bound information is encoded in the genome serving as a reservoir to be passed from one generation to the next. Free information is contained in nonnucleotide organic polymers such as proteins, lipids, fatty acids, and polysaccharides that regulate cellular morphology and function. Free information is dependent in a complex way on the translation of specific subsets of the bound information and their subsequent interactions. Free information is also present in the extracellular space in multicellular organisms. In this case, cellular function is influenced by both the intra- and extracellular free information.

The bound genetic information can be calculated using the Shannon’s information (12) function as demonstrated by Kendal (20). This approach measures the quantity of information of a string of symbols such as DNA (21). If there are γ equally probable arrangements for the symbols of a coded message, then the information content is the displacement from randomness defined by k in γ where k is a constant. The information content is often termed the “entropy,” although its relationship to the traditional thermodynamic entropy term is complex and outside the scope of this report. An increase in the Shannon entropy implies an increase in disorder and loss of information. There are several limitations to the application of Shannon’s information to biological systems: (a) because there is redundancy in the genome because of codon degeneracy, Shannon’s information will systematically overestimate the actual information content of a genetic segment; and (b) although Shannon’s information may measure the quantity of information, it does not measure its quality. That is, a small change in the information content of a gene (i.e., a single bp mutation) may catastrophically alter the function of a gene product while multiple errors in a less critical segment of the gene may have no biological consequences.

Despite these limitations, Shannon’s information has been shown to be a useful and reasonably accurate measure of genetic information (20). Furthermore, because Shannon’s information is both additive and conserved, it is well suited to estimating changes in information content over time. Fisher’s information (12) has these properties as well.

Let a cellular genotype consist of g genes, g = 1, 2, . . ., G. If at time t a gene consists of m codons, the entropy H of each codon, in which r’j is the probability of occurrence of the jth configuration of the ith codon is:

\[ H = -k \sum_{j=1}^{N} r'_j \ln r'_j \]  

Limit N is 64 (the number of bp combinations in each triplet). If the m codons are equally probable, then all H’ = H, a single value.

Because there is degeneracy in the third position of the triplet code, this slightly overestimates the actual genetic information content (20). The maximum bound information content of each gene \( I_g \) consisting of m codons is:

\[ I_g = mH \]  

The bound information content of the cell \( I_c \) is estimated as the sum of the information content of all of the genes:

\[ I_c = \sum_{g=1}^{G} I_g \]  

by additivity over the codons.

After reproduction at a time \( t + \Delta t \), the entropy \( H \) of the transgenerational encoded information depends on the probability \( r'_j (t + \Delta t) \) of finding the identical component of j in the ith codon:

\[ \bar{H} = -k \sum_{j=1}^{N} r'_j (t + \Delta t) \ln r'_j (t + \Delta t) \]  

In the absence of mutations, all values of \( r'_j (t + \Delta t) = r'_j (t) \). Then \( \bar{H} = H \), and there is no change in the entropy. However, in the presence of mutations, the two probabilities are not equal, therefore the two entropies are likewise not equal, and an information change for gene g can be defined as:

\[ \Delta I_g = m(H - \bar{H}) \]  

In the evolutionary environment that develops during carcinogenesis (22, 23), the information content of each gene may contribute to the selective growth advantage of the cell which, in turn, determines its ability to proliferate. We designate u as a measure of this growth advantage, \( u_g \) being that for gene g. Cells with the highest values of \( u_g \) in the tissue will proliferate at the expense of those with lower values. We suppose that \( u_g \) is some function f of the the information content \( I_g \):

\[ u_g = f \left( I_g \right) \]  

and after a generation the growth advantage generated by gene g becomes:

\[ u_g (t + \Delta t) = u_g (t) + \Delta u_g = f \left( I_g + \Delta I_g \right) \]  

where \( u_g \) can be 0, positive or negative (i.e., it can confer growth advantage, disadvantage, or have no interaction with evolutionary selection pressures). Similarly, \( \Delta u_g \) can increase, decrease, or leave unchanged the cellular growth advantage. The total growth advantage of a cell with N genes can be estimated as:

\[ u_c = \sum_{g=1}^{N} u_g \]  

This assumes additivity of the growth advantages for each gene, which is undoubtedly an oversimplification because the complex interactions of gene products will likely produce nonlinear dynamics.

Building on Eq. H, we define the proliferative capacity \( P_c \) of each cell population c as some increasing function h of the genetic growth information \( u_c \) of each population as compared with the mean value \( \bar{u} \) over all of the M cellular populations present:

\[ P_c = h \left[ u_c - 1/M \sum_{m=1}^{M} u_m \right] = h(u_c - \bar{u}) \]
EXTRACELLULAR INFORMATION CONTENT IN CARCINOGENESIS

The preceding analyzed the role of perturbations in bound information on intracellular free information. Free extracellular information in the presence of stochastic changes in bound information is now briefly examined.

Consider an arbitrarily small volume of tissue in a multicellular organism that contains subpopulations of cell types \( c \) with \( c = 1, \ldots, R \). There are \( S_c \) members of each cell type \( c \). The total environmental information \( I_c \) over all the cell types can be expressed as a function of the bound intracellular information as follows:

\[
I_c = \sum_{c=1}^{R} \sum_{t=1}^{G} S_c (q^t_{c,t})
\]

where \( q_c \) is the number translated extracellular copies (proteins) of the information content of each gene \( F_c \) in the microenvironment from each local cell type \( c \) (where there are a total of \( G \) genes present as above). These organic molecules are assumed to contribute additively. Also, for simplicity we will for now neglect diffusion of information in and out of the local environment.

Alternatively, \( I_c \) can be simply characterized by a distribution function \( \phi(x,t) \) that describes the local environmental concentration of all organic molecules \( x \) at time \( t \). This distribution function, however, is directly dependent on the number and differentiated phenotype of the local cellular populations. That is \( \phi_{st} \) will differ from \( \phi_{nt} \) and so on. Furthermore, \( \phi_{dnt} \) will vary spatially within the kidney because, for example, the environmental information around the glomerulus will differ from that around the distal tubule. In other words, the environmental free information \( \phi(x,t) \) is a measure of the morphology and function of the constitutive cells in the tissue. We note the possibility of feedback loops such that the environmental information maintains the stability of the cellular distribution that produces it. This may provide a quantitative model for the role of extracellular information in controlling differentiation.

In normal tissue, the number and phenotypic distribution of cells will remain constant maintaining a stable \( I_c \). However, in the presence of transformed cells, \( S_c, R, \) and \( f \) will be subject to stochastic perturbations from random genetic mutations as well as local selection forces. This will produce a complex dynamics in which, by Eq. 1, \( I_c \) will vary with time.

Thus, in the absence of severe injury or infection, the information distribution function is highly robust: \( \phi(x,t) = \phi(x,t + \Delta t) \), where \( \Delta t \) represents a time frame on the scale of a cellular generation (i.e., days or weeks), thus smoothing diurnal and day-to-day stochastic perturbations. The presence of tumor will alter the distribution function depending on the number of tumor cells present (e.g., the presence of tumor typically decreases glucose concentration and increase lactic acid concentrations). This time-dependent change in the environmental information content will, thus, be a measure of tumor growth within the tissue.

We will now examine the consequences of this intra- and extracellular information degradation in the development of the malignant phenotype and tumor formation. The former follows immediately and focuses on cellular morphological and functional effects of random genomic mutations constrained by environmental selection factors. Subsequently, we explore the role of transformation-induced changes in the extracellular information content in tumor growth dynamics in vivo and in vitro using EPI.

LIMITS OF INFORMATION DEGRADATION IN CARCINOGENESIS

We view the interaction of mutation rate, evolutionary competition, and transmitted information in carcinogenesis by applying the limit developed by Eigen and Schuster (24). In our previous notation, this is:

\[
I_{suc} = \ln(u_c - \bar{u}) / \alpha
\]

where \( I_{suc} \) is the maximum number of “signals” or genetic information that can be passed successfully through each generation, \( u_c - \bar{u} \) is the competitive advantage of the population compared with the average of all of the local populations, and \( \alpha \) is the mutation rate. Rearranging and adding the result to Eq. 1 gives:

\[
P_e = h(u_c - \bar{u}) = h(e^{I_{suc}})
\]

We now have explicitly linked the proliferation of transformed populations to mutation rate and maximum information transfer. Environmental parameters are indirectly included in Eq. L because they modify the specific competitive advantage of one or all of the populations (i.e., change \( u \) and \( \bar{u} \)). Again, note the time frame is one or more cellular generations, thus smoothing stochastic events that might transiently reverse the otherwise monotonic-appearing multistep cellular progression from normal to cancer.

The functional relationship of information loss, mutation rate, and selective growth advantage can be applied to individual genetic segments via Eq. F. Genes with small \( u_c \) contribute minimally to the overall fitness of the cell and are, by Eq. I, subject to the maximal information degradation, whereas genes with large \( u_c \) will demonstrate the least drift from the initial information content.

If genes encoding specific differentiated traits contribute minimally to the proliferative advantage of the cell (i.e., have a small value for \( u_c \)), then \( I_{max} \) will be monotonically degraded by unconstrained mutational forces. This is reasonable because differentiated morphology and function are specialized traits found only in multicellular organisms and can be expected to influence the survival of the organism but not the competitive advantage of an individual cell. The time-dependent decline in \( I_{max} \) is apparent clinically in progressive morphological and functional drift away from the original differentiated phenotype observed during carcinogenesis (Ref. 25; Fig. 1). Interestingly, the unconstrained accumulation of mutations that produces dedifferentiation may eventually produce sufficient genetic noise that a differentiation phase transition (24) results. This new cell state, in the absence of sufficient transgenerational information to retain a “memory” of the original differentiated state, will exhibit nonphysiological combinations of phenotypic traits, such as expression of endothelial cell traits in melanoma and breast cancer (26, 27). These cells will, in turn, produce morphologically disorganized and poorly functioning tissue typical of invasive cancers (Ref. 28; Fig. 1).

Similarly, mutations in genes with large \( u_c \) will decrease the cellular competitive advantage, resulting in nonproliferation or cell death. Although this is catastrophic for the individual cell, it eliminates the entry of the mutation into the evolving neoplastic population. Thus, genes controlling phenotypic traits critical for proliferation (such as oncogenes, glucose transporters, enzymes in the glycolytic pathway, angiogenesis, promoters, and others) are predicted to exhibit only gain of function or amplification genetic defects, whereas loss of function defects will dominate in noncritical genes.

This prediction is supported by carcinogenesis literature showing gain of function defects in oncogenes such as K-RAS and amplification of glucose transport genes (29–31). An interesting test of this model is observed when environmental selection forces are perturbed...
by the administration of chemotherapy. This new environmental selection parameter will increase in the value of $u$ for genes controlling the multidrug resistance phenotype. The predicted amplification and gain of function mutations in these genes during the subsequent evolution of the cancer population has been observed (32–34).

The development of angiogenesis probably represents a similar evolutionary phenomenon. Early cancers rely on diffusion and the native blood supply for substrate delivery. As the tumor grows, angiogenesis, becomes a critical survival parameter. It has been shown that, as predicted by this model, the angiogenic phenotype is found far more frequently in “older” tumors and develops far more quickly in the presence of a high mutation rate (35).

Specific environmental selection pressures can be inferred by common phenotypic traits found in transformed populations. Thus, the ability of most transformed cells to withstand severe conditions of nutrient deprivation suggests that enhanced ability to obtain and use substrate efficiently is strongly selected in the Darwinian environment of carcinogenesis (36).

Thus, for genes with small values of $u_g$:

$$\frac{dI_g}{dt} < 0 \quad \text{for all time } t$$  \hspace{1cm} (M)

and genes with large values of $u_g$:

$$\frac{dI_g}{dt} < 0 \quad \text{if } I_g > I_{\text{max}}$$  \hspace{1cm} (N)

or

$$\frac{dI_g}{dt} = 0 \quad \text{if } I_g = I_{\text{max}}$$  \hspace{1cm} (O)

---

**Fig. 1.** Macrographs illustrating the information degradation associated with the somatic evolution of colon cancer. A, normal colon mucosa with regular, highly predictable cellular morphology and well-organized, functional tissue indicating a full expression of the encoded information necessary for this differentiated phenotype. B and C, colon adenomatous polyps with low-grade (B) and high-grade (C) dysplasia. The heterogeneous cellular morphology and tissue disorganization exhibit progressive drift from the fully differentiated state. Both polyps probably contain a mosaic of competing populations, each with partial degradation of genomic information controlling the differentiated phenotype. The degradation is greater in C, which is less able to recapitulate normal colonic mucosa than B. However, despite the degradation of differentiation information, both sets of populations evidently retain the ability to proliferate successfully. D, adenocarcinoma of the colon (well differentiated, minimally invasive) demonstrating further cellular heterogeneity and degradation of tissue morphology and function while maintaining the proliferative phenotype. E, undifferentiated, invasive colon cancer with highly disordered, heterogeneous cellular morphology and complete loss of normal colon tissue organization and function. This demonstrates an error catastrophe (see text) in genomic information controlling differentiation with chaotic, unpredictable cellular and tissue morphology and function. Genetic information necessary for proliferation is obviously retained despite the unconstrained loss in segments controlling differentiation.
where $I_u$ is the information content of the gene and $I_{\text{gen}}$ represents some minimal quantity that allows the gene product to remain functional. In other words for high $u_c$ genes, the time-dependent function describing the information loss has an extremum at $I_{\text{gen}}$ because of the selection pressures. This allows the cells to maintain a minimal information content necessary for life (i.e., they continue to proliferate and generate energy) despite the marked drift in morphology and function that result from genomic instability.

**THE EVOLVING MICROENVIRONMENT AND MUTATION RATE**

We briefly address the critical role of an unstable microenvironment in carcinogenesis. From Eqs. K and L, it appears that although cellular mutations will produce perturbations in the microenvironment, these changes, in turn, will alter the competitive advantage of individual cell population $u_c$, and the average of all of the populations $\bar{u}$. For example, if the competitive microenvironment of carcinogenesis results in $\frac{du_c}{dt} > 0$, clonal expansion will be restricted to populations with $du_c/dt > \frac{du}{dt}$. This accelerating rate of competitive phenotypes is consistent with observations of malignant progression in which increasing aggressive cellular populations emerge over time during carcinogenesis (37). Other perturbations that may be stimulated by the presence of tumor cells include host immune response. If, for example, cytotoxic T cells detect a specific surface antigen expressed on some of the tumor cells, the $u_c$ of that population will decline. However, as noted below, the forces of mutation and selection may simply result in new resistant populations as the total tumor population continues to evolve into progressively more diverse phenotypes (38).

In an evolving microenvironment, from Eq. K and the increasing nature of function $h$ with its argument, populations with $\frac{du_c}{dt} < 0$ will not survive (an obvious exception being the situation in which $\bar{u}$ is decreasing at a greater rate) and those with $\frac{du_c}{dt} > 0$ will proliferate. In this way, the high mutation rate found in cancer is favored because it will produce a greater cellular diversity and increased probability of a proliferative phenotype. This is demonstrated in bacterial studies that have found a 5000-fold increase in the mutation rate when culture conditions became more restrictive and made to evolve over time (39). Clinical studies have demonstrated mutation rate increases during tumor progression (40). Loeb (8) has hypothesized that acquisition of the mutator phenotype is a critical and necessary event in carcinogenesis. Our analysis supports this hypothesis.

**APPLICATION OF EPI TO TUMOR GROWTH**

We now turn our attention to the dynamics of altered environmental information caused by the degradation of the genome in transformed cells. As noted earlier, normal tissue will contain an extracellular distribution of organic molecules $\phi(x,t)$ that will remain stable with time under physiological conditions. However, the presence of tumor will perturb this distribution, depending on the number of tumor cells present. Thus, changes in $\phi(x,t)$ will be dependent on the number of tumor cells present and the deviation of their internal information content from the tissue of origin. This new (cancer) distribution function $\phi_c(x,t)$ will increasingly differ from $\phi(x,t)$ with time. This difference between the distribution functions conveys information regarding the onset and progression of carcinogenesis.

This model prediction of time-dependent environmental change as a consequence of cellular information degradation during carcinogenesis can be tested by applying EPI techniques to derive estimates of tumor growth kinetics. This calculation is carried through next.

The problem we address is the form taken by the marginal probability law $p(t)$ of $p_c(x,t)$. This marginal law also represents, by the law of large numbers, the relative number of cancer cells anywhere in the space $x$ of points within an affected organ. The time dependence of $p(t)$ then defines as well the growth with time of the cancer.

The EPI approach (14, 41, 42) uses Fisher information (12) rather than Shannon information (11) $H$ or $I$ (as in the preceding). However, there is a tie-in between the two. Both are manifestations of a cross entropy measure of information (16);

$$I_c = \int dt \ p(t) \log \frac{p(t)}{r(t)} \quad \text{(P)}$$

The probability law $r(t)$ is called the “reference” law. Mathematically, the Shannon information $H$ is the cross entropy between a probability law $p(t)$ and a constant, or uniform, reference law $r(t) = \text{const}$. By comparison, Fisher’s information is the the cross entropy between a probability law $p(t)$ and its slightly shifted version $r(t) = p(t) + \Delta t$.

The two measures also differ qualitatively. Shannon’s information measures the degree to which a random variable is uniformly random (i.e., resembles its uniform reference function, see above). This in turn measures the ability of a system to distinguish signals, as in Eq. K, where it is used to distinguish genotypes. By comparison, Fisher’s information measures the degree to which a required parameter (in our case, the time) may be known (14). Knowledge of the time is particularly apt to the analysis of cancer growth, as is discussed below.

The working hypothesis of the EPI approach (14, 41, 42) is that the data in any measurements result from a flow of information, specifically Fisher’s information, that proceeds from an information source to a sink in data space. This follows a model of Brillouin (16). The information source is the phenomenon that underlies the data. Its information level is denoted as $J$. The data information is called $I$, and the information flow is symbolized by the transition $J \rightarrow I$. The flow physically consists of material particles that carry the information. Each phenomenon generally uses different messenger particles. (For example, in an optical microscope, the particles are photons.) The source information $J$ represents the Fisher information that is intrinsic to the measured phenomenon. Assuming that we are dealing with an isolated, passive system, the received information can at best equal the source level, i.e., $I \leq J$. The level $J$ corresponds as well to the bound, as opposed to free, information that was referred to earlier.

In our analysis $J$ represents the total extracellular Fisher information that is produced by a cancer cell. The information is about the time, in particular the time $\theta$, at which the carcinogenesis process began. This information may be carried by several biological intermediates. Of these, protons appear to be the dominant messengers because increased glucose uptake and excessive secretion of acid into the extracellular spaces are observed in the vast majority of clinical cancers (43, 44). Furthermore, experimental demonstration of the proton as a biological messenger between macromolecules has been published recently (45). Other carriers of information may exist as well, such as increased interstitial pressure or decreased oxygen or glucose concentrations.

The totality of such information that reaches a normal cell because of all neighboring cancer cells is what we call $I$. Thus, $I$ is the total information that is provided the functioning cell by all its neighboring cells about the onset time $\theta$ of carcinogenesis. Empirically, the received information tends to suppress the receiver, i.e., the normal cell. We now apply EPI to predict tumor growth dynamics expected from this model.

Downloaded from cancerres.aacrjournals.org on December 28, 2017. © 2002 American Association for Cancer Research.
If $\theta$ is the age of the individual at which carcinogenesis began, $t$ is the time since it began (i.e., the age of the tumor). At the time $(\theta + t) = \tau$, some quantity of information (e.g., lactate concentration) is present in the environment because of the presence of the developing tumor. In the usual clinical situation, when an observer detects the information (the presence of a cancer), he has knowledge of the value of $\tau$ but not the value of $\theta$. Of course, by the above definition of $\tau$, the detected information (i.e., the size of the tumor) carries information about $\theta$. For example, if $t$ (and, therefore, the tumor size) is small then the value of $\tau$ is very close to the value of $\theta$. We seek to find $p(\tau|\theta)$, the probability of observing the existence of a tumor at time $\tau$ in the presence of the fixed value of $\theta$. In other words, we want the probability of detecting environmental information resulting from the presence of a tumor that originated at time $\theta$ at some later time $\tau$. This is called the “likelihood law” for the measurement process. This likelihood law is equivalently $p(t)$, because it will not change shape with a shift in the absolute position $\theta$.

Thus, we define time $\tau$ as simply the time at which a cancer is observed in a given organ. $t$ is its unknown (random) component since carcinogenesis began, and $p(t)$ is the probability density for the latter.

By the law of large numbers, $p(t)$ is also proportional to the relative mass of the cancer, i.e., the total mass of cancer cells within the organ divided by the total organ mass. This assumes the cancer cells and functioning cells to have about the same mass density.

The EPI approach establishes an answer $p(t)$ by first determining its real probability amplitude function $q(t)$, where:

$$p(t) = |q(t)|^2 \quad (Q)$$

Here we seek an EPI self-consistent solution for $q(t)$. This is to satisfy the basic condition that the information loss $(I - J)$ in transit from the phenomenon to the data is an extreme value:

$$I - J = \text{extrem.} \quad (R)$$

through choice of the function $q(t)$. The solution $q(t)$ to the problem obeys the Euler-Lagrange condition (14):

$$\frac{d}{dt} \left( \frac{\partial(i - j)}{\partial q} \right) = \frac{\partial(i - j)}{\partial q} \quad (S)$$

The indicated differentiations in Eq. S result is a differential equation in the unknown function $q(t)$. This can be linear or nonlinear, depending upon case. A linear example from physics is the Schroedinger wave equation, a linear, second-order differential equation (14, 42). A nonlinear example from genetics is the law of genetic change (41). Biological growth processes, including cancer growth, are generally nonlinear (41).

The Eq. S is to be supplemented by a zero-condition:

$$i - \kappa j = 0, \quad \kappa \geq 0 \quad (T)$$

as a constraint. This merely states that $I = \kappa J$, or $I \leq J$ as mentioned before. The solution to the combined problem Eqs. S and T is called a self-consistent EPI solution.

Quantity $\kappa$ is a positive constant that is to be found. Quantities $i$ and $j$ are information densities, i.e., the integrands of information functionals:

$$I = 4 \int dt \, i(t), \quad \text{and} \quad J = 4 \int dt \, j(q,t) \quad (U)$$

where

$$i(t) = q^2 \quad (V)$$

Eqs. U and V together define the Fisher information level $I$ in received data (see above) for a temporally shift-invariant system. From Eq. T, the constant $\kappa = i/j$ represents the efficiency with which information $j$ is transformed into data information $i$. We next find the self-consistent solution to the problem Eqs. R and T.

A POWER-LAW SOLUTION

The constant $\kappa$ and the functional $j(q,t)$ are the effective unknowns of the problem. They are solved for by carrying through the self-consistency approach. The solution is found in Eq (A21) of Appendix A to be:

$$q(t) = \left( \frac{1 + \kappa}{A} \right) \left( \frac{A t}{\kappa} \right)^{i(1+\kappa)} \quad (W)$$

where $A = \text{const.}$ This is a power-law solution. The power $\kappa/(1 + \kappa)$ is as yet unknown. We find it next.

DETERMINING THE POWER

Without loss of generality, Eq. W may be placed in the form:

$$q(t) = Et^{\alpha}, \quad \alpha = \kappa/(1 + \kappa), \quad E = \text{const.} \quad (X)$$

The power is now $\alpha$. Our overall aim is to find this power.

Because form Eq. X grows unlimitedly with $t$, for the corresponding $p(t)$ to be normalized the cancer growth must be presumed to exist over a finite time interval $t = (0, T)$. This is of course the survey time over which cancers are observed. The normalization requirement then gives a requirement:

$$\int_0^T dt \, p(t) = E^2 \int_0^T t^{2\alpha} = 1, \quad \text{or} \quad E^2 = (2\alpha + 1) T^{(2\alpha + 1)} \quad (Y)$$

after the integration.

The Fisher information conveyed by the cancer over the entire survey period $(0, T)$ is:

$$I = 4 \int_0^T dt \, \dot{q}^2 \quad (Z)$$

EPI theory (14) generally requires that a “free field” solution for $q(t)$ convey minimum Fisher information. The field would in general be some source or force that is imposed from outside that modifies the evolution of $q(t)$. Examples of external fields might include clinical treatment such as some regimen of chemotherapy, tissue constraints such as inadequate blood supply producing tumor necrosis, host immune response, and physical barriers such as adjacent bone or cartilage. However, here we are studying a freely growing tumor. This then defines our free-field scenario. Thus, we require the Fisher information for the solution $q(t)$ to be a minimum. The minimum will be through choice of the free parameters of the process. The observation interval $T$ is already fixed, and $E$ is fixed in terms of it by the second Eq. Y. Hence, there is only one free parameter left, the power $\alpha$.

Differentiating Eq. X, substituting it into Eq. Z and using Eq. Y gives, after an integration:

$$I = \frac{4}{T^2} \left( \frac{2\alpha + 1}{2\alpha - 1} \right) q^2 \quad (AA)$$

3680
This may be minimized through choice of \( \alpha \) by ordinary differentiation. Setting \( \partial I / \partial \alpha = 0 \) results in the algebraic equation:

\[
4\alpha^2 - 2\alpha - 1 = 0, \quad \text{or} \quad \alpha = \frac{1}{4} \left(1 \pm \sqrt{5}\right) \tag{BB}
\]
as roots. However the negative square-root choice when placed in Eq. AA gives a negative value for \( I \), which is inconsistent because \( I \) must be positive [see Eqs. U and V]. Thus, the answer is a power-law solution:

\[
p(t) = E t^\gamma = F t^I, \quad \gamma = -\frac{1}{2} \left(1 + \sqrt{5}\right) = 1.62, \quad F = \text{const.} \tag{CC}
\]

This value of the power \( \gamma \) can be compared with published data on growth rate determined by sequential mammographic measurements of tumor size in human breast cancers (46–52). In the six available studies (46–51), the value of \( \gamma \) was: 1.72, 1.69, 1.47, 1.75, 2.17, and 1.61 (mean 1.73 with standard deviation 0.23). The tumors in these studies seem well suited to test our prediction because they were small (<2 cm in diameter), untreated cancers growing within the breast, i.e., the tumors were not subject to any apparent clinical or tissue constraints so that the conditions for a “free field” solution are probably met.

**EXPERIMENTAL VERIFICATION**

The power-law result Eq. CC was experimentally confirmed as follows. The most convenient clinical data for purposes of confirming Eq. CC would directly observed values of \( p(t) \), i.e., cancer mass versus time. On the other hand, the clinical data in Refs. 46–48 determine a straight-line relation with a slope of \( \alpha \). Thus, the answer is a power-law solution:

\[
p(v) = B v^{I-\gamma}, \quad B = \text{const.} \tag{DD}
\]

Given the value of \( \gamma \) in Eq. CC, this becomes

\[
p(v) = B v^{\frac{1}{2} - \frac{\sqrt{5}}{2}}, \quad (EE)
\]

The exponent is approximately value \(-0.382\). Hence, on a log-log basis, Eq. DD predicts a straight-line relation with a slope of \(-0.382\). The data in Refs. 46–51 are values of \( p(v) \) versus tumor size (apparent diameter). We cubed the diameters to give equivalent volumes \( v \).

Fig. 2 shows, on a log-log basis, the theoretical solution Eq. EE with, for comparison, all points of the clinical studies in Refs. 46–51 that are quantitatively tabulated. [One point in each of the three studies (46–48) is ignored because the indicated tumor sizes for these points is the indefinite value 2.+] The agreement between the theoretical curve and this experimental data is visually good. Moreover, as we saw above, the theoretical curve is well within 1 SD of the experimentally determined values.

Of course, we must also emphasize the many factors that may produce clinical variation from the predicted growth dynamics. These include:

(a) Altered tumor cell volume. We assumed that the tumor cell volume is exactly equal to the volume of the normal cells, but some heterogeneity in tumor cell size is frequent.

(b) Power law growth is predicted only for tumors unconstrained by external factors (i.e., a “free field”) and thus represents a relative maximum value, i.e., the predicted dynamics are only applicable to tumor growth governed by unimpeded in vivo replication of the tumor cells. If tumor cell expansion is constrained by external factors such as inadequate vascular supply, host response, spatial considerations or clinical therapy, slower growth will be observed.

(c) Tumors may contain significant volumes of noncellular regions such as necrosis and blood vessels or large numbers of nonneoplastic cells such as endothelium, fibroblasts, macrophages, and lymphocytes.

It is encouraging, then, that despite these possible complications, the predicted tumor growth law agrees well with the experimental data.

**EFFICIENCY \( \kappa \)**

By the second Eq. X, \( \kappa = \alpha/(1 - \alpha) \). Then by the positive root of Eq. BB:

\[
\kappa = 4.27 \tag{FF}
\]

This is the total efficiency attributable to the unknown number \( S \) (see Eq. I) cancer cells that are communicating with the functioning cell. Because the efficiency for any single cell can be at most unity, this implies that there are at least \( S = 4 \) neighboring cancer cells that are interacting with the functioning cell. That \( S \) is rather large means that a relatively large amount of lactic acid bathes the healthy cell. This forms a toxic environment that may eventually kill the normal cell, leading to compromise of functioning tissue and death of the host. This mechanism of tumor invasion has been proposed and modeled previously (43).

**FIBONACCI CONSTANT**

We note that the power \( \gamma \) of the growth process Eq. CC is precisely the Fibonacci “golden mean.” This number occurs naturally as the relative increase in an ideally breeding population \( p_f(t) \) from one generation to the next:

\[
\frac{p_f(t + \Delta t)}{p_f(t)} = \gamma \tag{GG}
\]
Comparing Eqs. AA and JJ shows that, in a complex system such as carcinogenesis, is necessarily limited by a theoretical limit on the value of screening if the development of phenotype and the first observation of the cancer. This may place a large degree of uncertainty could place a fundamental limit on the efficacy of cancer screening if the onset of metastases correlates with the age of the cancer rather than its size.

As with most nonlinear dynamic systems, the genetic pathways in carcinogenesis are complex and unpredictable. Thus, fixed, linear genetic sequences leading to cancer (17) will likely be uncommon in sporadic human cancers (exceptions may be congenital neoplastic syndromes such as retinoblastoma, neurofibromatosis, and familial polyposis syndrome). This is consistent with studies demonstrating heterogenous genetic pathways in the pathogenesis of breast and renal cancers (57, 58). Finally, our model adds a cautionary note to a wide range of strategies in cancer therapy because the same mutational and selection forces that produced the initial malignant population remain available to foil these therapies through rapid evolution of resistant phenotypes.

**APPENDIX A. DERIVATION OF EPI POWER LAW SOLUTION**

General Solution. By the use of Eq. V, Eqs. S and T become, respectively:

\[
\frac{d}{dt} \left( \frac{\partial(q^2 - J(q,t))}{\partial q} \right) = \frac{\partial(q^2 - J(q,t))}{\partial q}
\]  

(A1)
and
\[ q^2 - k[q, t] = 0 \]  
(A2)

Explicitly carrying through the differentiations in Eq. A1 gives:
\[ 2 \frac{d^2 q}{dt^2} = -\frac{\partial j}{\partial q} \]  
(A3)

Eqs. A2 and A3 must be simultaneously solved for \( j \) and \( q \). Notice that each equation involves both \( j \) and \( q \). The aim is to get one equation in one of the unknown functions \( q \) or \( j \).

Differentiating \( \frac{d^2 q}{dt^2} \) Eq. A2 gives:
\[ 2q \frac{d^2 q}{dt^2} = \kappa \left[ \frac{\partial j}{\partial t} + \frac{\partial j}{\partial q} \right] \]  
(A4)

Eqs. A3 and A4 must be solved simultaneously.

Using Eq. A3 on the left side of Eq. A4 gives a requirement:
\[ -\frac{\partial j}{\partial q} = \kappa \left[ \frac{\partial j}{\partial t} + \frac{\partial j}{\partial q} \right] \]  
(A5)

Taking the square root of Eq. A2 gives:
\[ j = \sqrt{\kappa q} \]  
(A6)

where the + sign was chosen because we are seeking a solution that grows with the time.

Substituting Eq. A6 into Eq. A5 gives directly:
\[ -\sqrt{\kappa} \sqrt{\frac{\partial j}{\partial q}} = \kappa \left[ \frac{\partial j}{\partial t} + \frac{\partial j}{\partial q} \right] \]  
(A7)

After a rearrangement of terms and division by \( \sqrt{\kappa} \), this becomes:
\[ \sqrt{\frac{\partial j}{\partial q}} (1 + \kappa) + \sqrt{\kappa} \frac{\partial j}{\partial t} = 0 \]  
(A8)

This accomplishes our immediate aim of obtaining one equation in the single unknown function \( j \).

Let us seek a separable solution:
\[ j[q, t] = j_q(q) \cdot j_t(t) \]  
(A9)

where functions \( j_q(q) \) and \( j_t(t) \) are to be found. Substituting Eq. A9 into Eq. A8 gives a requirement:
\[ \sqrt{\frac{\partial j}{\partial q}} j_t(1 + \kappa) + \sqrt{\kappa} j_q = 0 \]  
(A10)

where the derivatives indicated by primes are with respect to the indicated subscripts (and arguments) \( q \) or \( t \). Multiplying the equation by \( j_q^{-1/2} j_t^{-3/2} \) gives:
\[ j_q^{-1/2} j_t(1 + \kappa) + j_t^{-3/2} = 0 \]  
(A11)

The first term on the left side is only a function of \( q \) and the second is only a function of \( t \). The only way they can add up to zero (the right side) for all \( q \) and \( t \) is for the first to equal a constant and the second to equal the negative of that constant. Thus:
\[ j_q^{-1/2} j_t(1 + \kappa) = A; \quad j_t^{-3/2} = -A \]  
(A12)

The first equation Eq. A12 can be placed in the integrable form:
\[ j_q^{-1/2} dq = A(1 + \kappa)^{-1/2} dq \]  
(A13)

Integrating and then squaring both sides gives:
\[ j_q(q) = \frac{1}{4} \left( \frac{Aq}{1 + \kappa} + B \right)^2, \ A, B = const. \]  
(A14)

The second Eq. A12 can be placed in the integrable form:
\[ j_t^{-3/2} dt = -A\kappa^{-1/2} dt \]  
(A15)

Integrating this, then solving algebraically for \( j_t \), gives:
\[ j_t(t) = 4 \left( \frac{At}{\kappa} + C \right)^{-2}, \ C = const. \]  
(A16)

Using solutions Eqs. A14 and A16 in Eq. A9 gives:
\[ j[q, t] = \frac{\left( \frac{Aq}{1 + \kappa} + B \right)^2}{\left( \frac{At}{\kappa} + C \right)^2} \]  
(A17)

Having found \( j \), we can now use it to find \( q \). Using Eq. A17 in Eq. A6 gives directly:
\[ \frac{dq}{1 + \kappa} = \frac{dt}{A\kappa + D} \]  
(A19)

This may be readily integrated, giving logarithms of functions on each side. After taking antilogarithms the solution is:
\[ q(t) = \left( \frac{1 + \kappa}{A} \right) \left( \frac{At}{\kappa} + D \right)^{(1 + \kappa)} - (1 + \kappa) \frac{B}{A} \]  
(A20)

This is a general power-law solution in the time \( t \). Two of the constants \( A, B, D \) may be fixed from boundary value conditions, as follows.

**Boundary-Value Conditions.** At \( t = 0 \) the cell is presumed to not be cancerous. Cancer has its onset at \( t = \epsilon, \epsilon > 0 \), a small increment beyond 0. This has two boundary-value effects: (a) At \( t = 0 \), the information \( j \) about cancer in the given tissue is zero, \( j(0) = 0 \), and (b) At \( t = 0 \) the cancer has zero mass, \( q(0) = 0 \).

Using these conditions in Eq. A17 gives the requirement \( B = 0 \). Using the latter and condition \( b \) in Eq. A20 gives either \( D = 0 \) or \( \kappa = -1 \). But the latter is ruled out a priori by EPI, because \( \kappa \approx 0 \) represents an efficiency measure of information transmission. Hence, \( D = 0 \). With this choice of constants, Eq. A20 becomes:
\[ q(t) = \left( \frac{1 + \kappa}{A} \right) \left( \frac{At}{\kappa} + D \right)^{(1 + \kappa)} - (1 + \kappa) \frac{B}{A} \]  
(A21)

This is the power-law we sought. Its square, the probability law \( p(t) \), is therefore also a power law. Eq. A21 is the result \( D = 7 \) formulation that assumes growth with time, whereas accrual growth occurs at discrete generations. The discrete effects are more important at small values, so that Eq. A21 becomes inaccurate near \( t = 0 \). Equivalently, the constant time offset \( D \) in Eq. A20 is some unknown, small constant.

**Error in the Estimated Duration of the Cancer.** The time \( t \) is the time since the unknown onset time \( \theta \) of the cancer condition. However, the time of observation of the tumor is not \( t \) but rather \( \tau \), where \( \tau = \theta + t \). The question is, how well can \( \theta \) be estimated based upon knowledge of \( \tau \)? This amounts to determining the position of the origin of the growth curve \( p(t) \) based upon one observed event at the randomly displaced time value \( \theta + t \).

The Cramer-Rao inequality (12) states that any estimate of \( \theta \) that is correct on average (unbiased) suffers a root-mean-square error \( e \) that can be no smaller than:
\[ e_{\text{min}} = \frac{1}{\sqrt{\tau}} \]  
(A22)
In our case I has a value given by Eqs. AA and BB:

\[ I = \frac{11 + 5 \sqrt{5}}{27} \]  

(A23)

Using this in Eq. A22 gives a minimum rms error of size:

\[ \epsilon_{\text{rms}} = 0.3T \]  

(A24)

As a relative error, this is quite large, amounting to 30% of the total time interval \( T \) over which cancer events are detected. Thus, cancer has an innate ability to mask its onset time. This contributes to its prevalence, and hence might be considered an evolutionary tactic.

REFERENCES

Application of Information Theory and Extreme Physical Information to Carcinogenesis

Robert A. Gatenby and B. Roy Frieden


Updated version
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/62/13/3675

Cited articles
This article cites 49 articles, 16 of which you can access for free at:
http://cancerres.aacrjournals.org/content/62/13/3675.full#ref-list-1

Citing articles
This article has been cited by 4 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/62/13/3675.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://cancerres.aacrjournals.org/content/62/13/3675.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.