Maximal entropy inference of oncogenicity from phosphorylation signaling

T. G. Graeber[a], J. R. Heath[a,b], B. J. Skaggs[a], M. E. Phelps[a,1], F. Remacle[a,2], and R. D. Levine[a]

[a]Crump Institute for Molecular Imaging and Department of Molecular and Medical Pharmacology, University of California, Los Angeles, CA 90095; and [b]NanoSystems Biology Cancer Center and Department of Chemistry, California Institute of Technology, Pasadena, CA 91125

Contributed by Michael E. Phelps, January 29, 2010 (sent for review December 7, 2009)

Point mutations in the phosphorylation domain of the Bcr-Abl fusion oncogene give rise to drug resistance in chronic myelogenous leukemia patients. These mutations alter kinase-mediated signaling function and phenotypic outcome. An information theoretic analysis of the correlation of phosphoproteomic profiling and transformation potency of the oncogene in different mutants is presented. The theory seeks to predict the leukemic transformation potency from the observed signaling by constructing a distribution of maximal entropy of site-specific phosphorylation events. The theory is developed with special reference to systems biology where high throughput measurements are typical. We seek sets of phosphorylation events most contributory to predicting the phenotype by determining the constraints on the signaling system. The relevance of a constraint is measured by how much it reduces the value of the entropy from its global maximum, where all events are equally likely. Application to experimental phospho-proteomics data for kinase inhibitor-resistant mutants shows that there is one dominant constraint and that other constraints are not relevant to a similar extent. This single constraint accounts for much of the correlation of phosphorylation events with the oncogenic potency and thereby usefully predicts the trends in the phenotypic output. An additional constraint possibly accounts for biological fine structure.

high-throughput measurements | information theory | phospho proteomics | signal transduction networks | systems biology

Biological systems use complex networks of molecular events, such as signaling and transcription, to regulate cellular outcome. Experimental biology is now able to measure up to thousands of these events from individual samples, inspiring efforts to identify both the events most contributory to cellular phenotypes and the nature of how multiple regulatory mechanisms are coordinated. Such understanding is crucial to the next generation of single agent and mixture molecularly targeted therapeutics. Kinase-mediated signaling is central to the execution of cellular programs and to communication with the environment and other cells. Aberrant signaling is implicated in many diseases and is a hallmark of cancer (1). Mass spectrometry- and antibody-based phospho proteomics allow the global profiling of the state of the signaling network (2–5). These approaches permit site-specific monitoring of phosphorylation and, thus, provide discrimination of the sometimes multiple and differential regulatory phosphorylation events on individual proteins. Here, we develop and apply an information theoretic maximal entropy analysis of the correlation of signaling events to cellular outcome to identify the dominant constraints that describe and predict the phenotypic output.

There are equivalent ways of motivating the choice of a distribution of maximal entropy (6). From an information theoretic point of view it is the distribution that is consistent with the data at hand and is otherwise least informative. From a statistical point of view, a distribution of maximal entropy is the most probable distribution in that it is the distribution that is observed in the largest number of experiments (the Boltzmann view; see also ref. 6). For our purpose the statistical, or strictly speaking statistico-mechanical, point of view has also a thermodynamic analog (7). When maximizing the entropy but constraining the distribution to be consistent with what we do know, there arise "parameters" that act as thermodynamic potentials. A well known example is the chemical potentials (8) that ensure the conservation, at equilibrium, of the number of molecules of a given species. Technically the thermodynamic potentials arise as Lagrange multipliers that are introduced in the process of seeking a maximum of the entropy subject to constraints. We use the terms parameters and "multipliers" interchangeably.

The principle of entropy maximization has been applied as an approach toward understanding biological networks. Examples include the extraction of genetic interaction networks from microarray data, the inference of the modularity of genomic networks, or for information processing in neural networks (9–18). The present application differs in two ways. First, our approach has a thermodynamic flavor: We seek to identify constraints that force a lower value of the (maximal) entropy and, thereby, allow us to specify the directions of response of the signaling system to perturbations (principle of Le Chatelier; ref. 19). Second, our approach directly addresses an essential characteristic of systems biology experiments—namely that the number of experiments is large compared with the number of phenotypic outputs, and so the system is statistically sampled, or, in other words, it is overdetermined. This usage contrasts with the more common situation in which the approach of maximal entropy is applied to an underdetermined system. An additional motivation for the present application is the partial least squares regression-based seminal contributions to systems biology analysis by Janes et al. (4, 20). Although our general approach and various technical details differ, the spirit is the same. We want to predict phenotypic outcome from signaling measurements. We report below that one (or possibly two) constraints, computed only from the data, suffices to semi-quantitatively predict the relative trends in the phenotypic output. We also report results for the "leave one out" cross-validation procedure, LOOCV (21), where the data, taken together with N-1 outputs, are used to predict the missing N'th output.

The method introduced in this paper is applied toward understanding the phenotypic implications of phospho proteomics data measured at steady state for point mutations (equals isoforms) of the Bcr-Abl oncogenic kinase (Fig. 1). These isoforms were discovered because they confer Abl inhibitor drug resistance in chronic myelogenous leukemia (CML) patients (22) and cause gain or loss in transformation potency (5). The transformation potency, also known as the growth rate, is the inverse of the doubling time for cells containing the oncogene. The signaling network was profiled (5) by using tyrosine phospho-peptide enrichment followed by mass spectrometry, yielding quantitative measurements of site-specific phosphorylation events on Bcr-Abl and downstream signaling proteins as shown in Fig. 1.

Author contributions: T.G.G., F.R., and R.D.L. designed research; T.G.G., J.R.H., B.J.S., M.E.P., F.R., and R.D.L. analyzed data; F.R. and R.D.L. performed research; and T.G.G., J.R.H., M.E.P., F.R., and R.D.L. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: mphelps@mednet.ucla.edu.

[2]Permanent address: Département de Chimie, Université de Liège, B4000 Liège, Belgium.

This article contains supporting information online at www.pnas.org/cgi/content/full/1001149107/DCSupplemental.

The site phosphorylation data as measured by mass spectrometry (see *SI Appendix* and ref. 5), is represented by a matrix denoted as **X** and shown in Fig. 1. It is a rectangular matrix with $n = 12$ columns, one column for each mutant of Bcr-Abl. Each row is the intensities of a site-specific phosphorylation event measured for the N different mutants. The identity of the peptide that is shown in each row is identified in the table in *SI Appendix*, section VIII of the SI. The number of rows, $P = 63$ in our example, is the number of quantitative measurements of such events that have been assessed (5) to be the most reliable. Many signaling proteins have more than a single phosphorylation site. For the 63 phosphorylation events measured at high confidence there correspond 49 proteins.

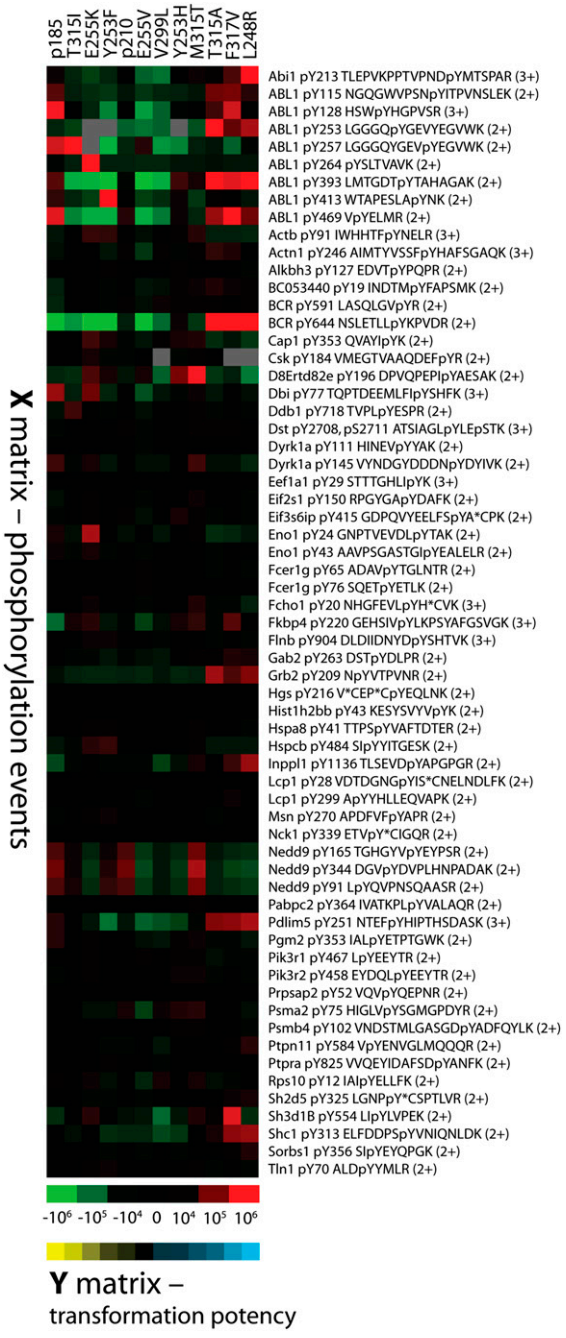

Fig. 1. Heat map representation of the measured phosphorylation data matrix **X**. Each one of the $n = 12$ columns of **X** is for a particular Bcr-Abl oncogene isoforms as identified. Each row of the phosphorylation data **X** represents relative levels of a phosphorylation event of a particular signaling protein. Several proteins have multiple phosphorylation sites (see *SI Appendix*). The row-wise mean-centered mass spectrometry ion current peak integration values are as indicated by the green to red scale bar. The relative transformation potency, **Y**, for the same 12 oncogenes ranges from 1 (yellow) to 0.42 (blue). The experimental determination of transformation potency is described in ref. 5 and in *SI Appendix*, and the numerical data for **X** and **Y** is also provided. See Fig. S9 for a row variance normalized heat map representation of the **X** matrix.

In addition to the phosphorylation data matrix **X** there is an output row vector **Y**, with 12 components. Each entry in **Y** is the transformation potency (outcome phenotype) for the corresponding phosphorylated Bcr-Abl isoform (Fig. 1). The theory that we set up aims to recover the values of the transformation potency by identifying the relevant thermodynamic, or Lagrange, parameters. Lagrange parameters that have relatively small numerical values are those that fine-tune the fit of the noise—i.e., they contain little or no biological information. By comparing the influence of a Lagrange multiplier against experimental uncertainties it is possible to test when small Lagrange multipliers are de facto zero (23).

For the general case, the output **Y** will be a rectangular matrix in which each row represents a different possible outcome (i.e., expression rate of a particular protein, a parameter related to apoptosis, cell motility, etc.) (4, 20) and not a single row vector as is the case here. It can then be that even when **X** is given, residual uncertainty about the output **Y** remains. In the Supplemental Information, we show that also for the underdetermined case one can use the method of maximal entropy. We do so by relating it to the measure of the information that **X** provides about **Y** when **Y** is not fully determined (see *SI Appendix*, Fig. S1).

The statement "**X** over determines **Y**" has both a mathematical and a biological/experimental meaning. The mathematical requirement is a necessary condition. For a full rank matrix **X** to fully determine **Y** it is necessary that the number, P, of events that are measured is larger than the number, N, of different initial conditions in the experiment (here the number of phosphorylated isoforms of Bcr-Abl). There is also a biological/experimental requirement that ensures sufficiency. This condition is determined by the nature of the biological system and the signal-to-noise quality of the experimental results: Given the data matrix **X** of P rows and N columns, we require that the columns of **X** are linearly independent. At the current state of characterization of phosphorylation networks, this condition is to be checked by examining concrete experimental results. The simplest test is that the N by N matrix $\mathbf{X}^T\mathbf{X}$ has N nonzero eigenvalues. Here the superscript T means the transpose of a matrix with the implication that the nonzero eigenvalues have to be positive. A (near) zero eigenvalue indicates (near) linear dependence of the N experimental conditions. Even for a well-designed experiment it is worthwhile to verify explicitly that the rank of **X** is N and not less.

We find the rank by diagonalizing $\mathbf{X}^T\mathbf{X}$and sorting the eigenvectors by the size of the corresponding eigenvalues. For the phosphorylation data matrix of Fig. 1, there is one dominant eigenvalue, an order of magnitude larger than the others. For the Gaussian distribution that we will use and provided that the data **X** is mean centered on the rows, the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are the Lagrange multipliers of the constraints in the method of maximal entropy. When only one eigenvalue is sufficient to provide a realistic approximation we say that there is one dominant constraint. The dominant constraint is the eigenvector corresponding to the largest eigenvalue. For the data of Fig. 1, the dominant constraint represents almost 60% of the norm of the phenotypic output as measured for 12 different oncogenes, with two additional constraints ≈85% is recovered and the balance is within the error bars of the measured output.

That one constraint is so dominant suggests that there may be one leading control (or regulatory) mechanism toward the phenotypic output. As shown in Fig. S2, the dominant constraint is quite localized about one particular phosphorylation site, BCRpY644, in the BCR protein. This site is clearly suggested as a target for further intervention. At the same time, the second and third most important constraints are localized on other peptides of the Bcr-Abl proteins and, in particular, on peptides that belong to the Bcr-Abl kinase ATP binding loop. If, by hand, we reduce the measured intensity of the BCRpY644 peptide, then these other peptides, and in particular Abl1pY264, Abl1pY257, and Abl1pY253, become dominant (see *SI Appendix*). It is then these sites whose phosphorylation best predicts the phenotype. We have experimentally confirmed contributions to Bcr-Abl-mediated transformation by the Abl1pY257 and pY253 phosphorylation events (5).

A declaration about the background assumptions that are made in the procedure of maximal entropy is discussed in section II. The analysis of the data matrix **X** is presented in section III. Prediction of the output by using constraints derived from the data are in section IV. The results of the Leave One Out Cross Validation, LOOCV, procedure are mentioned in section V and reported in *SI Appendix*. Technical comments on the notion of the information that the data provides about the output is discussed in section VI. Our summary is section VII.

II. Declaration Regarding Maximal Entropy of Biological Signaling States

The first stage in an application of a procedure of maximal entropy is to specify the limiting, and not necessarily biologically pertinent, situation for which the entropy is at its global maximum. Then, as relevant constraints are imposed on the system, the entropy value will be lowered, because relevant constraints, by definition, limit the range of possible results. Experimental noise can limit the ability to distinguish between relevant and irrelevant constraints, in that it can preclude deciding whether the addition of yet another constraint is warranted, because the accompanying decrease in entropy may be below the experimental noise (23).

Each entry in the data matrix **X** is a measurement of a signaling event for particular conditions. We declare that in the absence of constraints that require otherwise, we take all these measured values to be equally probable. In this case, a heat matrix map of **X** should be uniform in color at maximal entropy. The experimental heat map (Fig. 1) is clearly not uniform in color. It therefore contains biological and chemical information. The role of this theory is to determine how many constraints are necessary to capture this information. A secondary outcome is to determine how much of the nonuniformity is possibly due to noise. Technically, the expression for the entropy (in dimensionless units) is

$$H(\mathbf{X}) = -\sum P(\mathbf{X}) \ln P(\mathbf{X}), \qquad [1]$$

where the summation is over all of the possible values that **X** can assume. Here, $P(\mathbf{X})$ is the probability of a particular data matrix **X** (24) with details to follow in section III.

It is possible to argue that even when we do not know otherwise, it is not reasonable to take all signaling events to be equally probable. The reasoning is that often only relative phosphorylation values are measured and, thus, unnormalized comparisons can be misleading, e.g., amino acid composition influences ionization efficiency in mass spectrometry and binding affinity influences antibody-based results. Until absolute quantitation is common, it can be argued that one should scale the distinct measurements in the raw data matrix; for example, by operating on each row to make it mean centered and scaling each entry in the row by the variance of the row as shown in *SI Appendix* and refs. 4 and 25. Another argument for scaling encountered in systems biology is that low abundance events can have notably high biological relevance. For example, many regulatory proteins (transcription factors, kinases) are expressed at relatively low levels (as little as a few copies per cell) compared with structural proteins (25, 26). The full implication of these views awaits the further development of absolute quantitation methodologies. See *SI Appendix* for how the concept of a prior distribution (27, 28) allows one to discuss these alternative views.

III. The Distribution of Maximal Entropy

X is a matrix whose dimensions are the number, P, of different phosphorylation events, the row labels, times the number, N, of different isoforms, the column labels. One can also regard **X** not as a matrix but as a sample of N readings of the column vector $\mathbf{X}_n$ of P components, $\mathbf{X}_n \equiv (X_{1n}, X_{2n}, \ldots, X_{Pn})^T$. n is a label of the oncogenes, $n = 1, 2, \ldots, N$, and the different oncogenes differ in their phosphorylation strength and specificity. To characterize the distribution of phosphorylation events we make the assumption that it is of maximal entropy subject to constraints. The simplest constraints are the given means and variances of the rows (equals the phosphorylation events induced by different oncogenes) and the covariances between rows of the **X** matrix as defined in Eq. **2**.

We therefore seek a distribution of maximal entropy constrained by the values of the means and covariances (including the variances) computed for the measured data matrix **X**. This last sentence specifies how we derive the Gaussian distribution that is given in Eq. **3** below.

We arrange the means to be zero by centering each row of the data matrix **X**. For the covariances we encounter the basic reality of data matrices provided by systems biology, namely that there are more measurements than phenotypic outputs, $n < P$. Therefore, the P by P covariance matrix $\mathbf{X}\mathbf{X}^T$ with elements indexed by the phosphorylation events:

$$\left(\mathbf{X}\mathbf{X}^T\right)_{pq} = \sum_{n=1}^{N} (\mathbf{X})_{pn} (\mathbf{X})_{qn} \qquad [2]$$

cannot be inverted. $\mathbf{X}\mathbf{X}^T$ is necessarily singular because its rank cannot be higher than the smaller dimension of **X**, namely N. In other words, $\mathbf{X}\mathbf{X}^T$ can have no more than N nonzero eigenvalues.

As is well known (see e.g., refs. 6 and 29) the multivariate distribution of maximal entropy subject to given means and covariances is Gaussian. To explicitly write down a multivariate Gaussian distribution $P(\mathbf{X})$ for the matrix **X** we need to invert the covariance matrix $\mathbf{X}\mathbf{X}^T$ as given in Eq. **2**. This inversion cannot be done because this matrix has at least N-P eigenvalues that equal zero. A mathematical examination given in *SI Appendix* (see also ref. 24) shows that under such circumstances, the relevant covariance matrix is the N by N matrix $\mathbf{X}^T\mathbf{X}$. This matrix has the same nonzero eigenvalues as $\mathbf{X}\mathbf{X}^T$, the covariance matrix of the data, and this result is central to our technical discussion. Introducing the N column vectors of N components $\mathbf{Z}_i$, $i = 1, 2, \ldots, N$ that are computed as the eigenvectors of the $\mathbf{X}^T\mathbf{X}$ matrix, $\mathbf{X}^T\mathbf{X}\mathbf{Z}_i = \lambda_i \mathbf{Z}_i$, we have as a final result the normalized form of the multivariate Gaussian distribution

$$P(\mathbf{X}) = \left(1/(2\pi)^{N/2} |\mathbf{\Sigma}|^{1/2}\right)^N \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \mathbf{Z}_i^T \mathbf{\Sigma}^{-1} \mathbf{Z}_i\right]. \qquad [3]$$

$\mathbf{\Sigma}$ is the diagonal $N \times N$ variance matrix, meaning that the entries along its diagonal are the inverse of the eigenvalues $\lambda_i, i = 1, 2, .., N$ of the variance matrix $\mathbf{\Sigma}$. $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$, and because $\mathbf{\Sigma}$ is diagonal, $|\mathbf{\Sigma}|$ is the product of the eigenvalues $|\mathbf{\Sigma}| = \prod_i \lambda_i$. The eigenvalues are shown, vs. the running index i in Fig. 2.

The computation of the entropy of a multivariate Gaussian distribution is well known (see, for example, ref. 29). The deri-

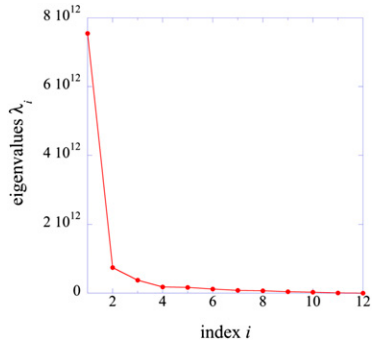

Fig. 2. The Lagrange parameters for the constraints on the phosphorylation data arranged in decreasing order starting at $i = 1$. (i is a running index that sorts the eigenvalues of the covariance matrix). The Lagrange parameters are computed as the eigenvalues of the 12 by 12 matrix$\mathbf{X}^T\mathbf{X}$. The data are not centered by rows so that all 12 eigenvalues are distinctly above zero but one eigenvalue, i =1, is well above the others.

vation is also given in ref. 24, where it is proven that the entropy of the entire data matrix $\mathbf{X}$ is given by Eq. **4**:

$$H(\mathbf{X}) = \frac{1}{2}\ln|\mathbf{\Sigma}| + \frac{1}{2}N\ln(2\pi e) = \tfrac{1}{2}\sum_{i=1}^{N}\ln(2\pi e/\lambda_i). \quad [4]$$

The explicit form for the entropy, Eq. **4** shows that one can rank the eigenvalues of $\mathbf{X}^T\mathbf{X}$ by their size. Eq. **4** further shows that the largest eigenvalue decreases the entropy the most. The ranking by size is the order for the Lagrange multipliers of the constraints, the largest being the most relevant. If the rows of $\mathbf{X}$ are mean-centered, the lowest eigenvalue for the 12 by 12 matrix will be zero and it does not contribute to lowering the entropy. So we keep only $n = 11$ terms in Eqs. **3** or **4**. If there is a dominant constraint, keeping just one term, $N = 1$, is sufficient.

Using the data matrix provided in *SI Appendix* and shown as a heat map in Fig. 1, the Lagrange multipliers λ_i, $i = 1, 2, \ldots, N$ are determined as the eigenvalues of the N by N matrix$\mathbf{X}^T\mathbf{X}$. The results for the multipliers exhibit a separation in scale with one large eigenvalue, one eigenvalue smaller by an order of magnitude and the rest even smaller (Fig. 2). The smaller eigenvalues are very small compared with the large, $i = 1$, eigenvalue but are sufficiently removed from zero to conclude that the N columns of $\mathbf{X}$ are linearly independent. This result allows us to exactly predict the phenotypic data (see Eq. **5** below). That there is a dominant constraint allows us to approximately predict the phenotype with one data vector as shown in section IV.

We note without proof that if we mean center the rows of $\mathbf{X}$ and also scale the entries of each row by the variance and only then seek a distribution of maximal entropy, this procedure leads to the same result as seeking a maximum of the entropy of the distribution $P(\mathbf{X})$ relative to a nonflat prior distribution $P^\circ(\mathbf{X})$ (27, 28) given as a product of independent Gaussian (or "normal") distributions, each for the given mean and variance. This factorized product is different from the correlated distribution as given by Eq. **3** above. The origin of the difference is the covariance of different phosphorylation events.

IV. Example: Predicting the Potencies from the Site-Specific Phosphorylation Data

This section predicts the potencies from the data and then shows that even just the one leading constraint already provides a realistic inference. The inference can be improved by adding the more marginal constraints. For each constraint i that we use, we need one parameter (coefficient). The number of parameters equals the number of constraints that we use (Eq. **4**). When the output is mean-centered, only the relative values of the potencies are predicted. This prediction requires one fewer parameter meaning that no parameter is needed when only the dominant constraint is used in the prediction.

The diagonalization of the covariance matrix $\mathbf{X}^T\mathbf{X}$ specifies a set of N (orthogonal and, by our choice, normalized) eigenvectors $\mathbf{Z}_i$ that we label by the same running index i as the eigenvalues. Each such eigenvector has N components. For a well defined experiment, as long as the number, N, of conditions is smaller than the number, P, of phosphorylation events, the rank of the data matrix $\mathbf{X}$ can be N and there will be N linearly independent eigenvectors of$\mathbf{X}^TX$.

On algebraic grounds, when the rank of the covariance matrix $\mathbf{X}^T\mathbf{X}$ is N, the N component vector $\mathbf{Y}$ of potencies (phenotypic output) can be exactly represented, noise and all, as a linear combination of the N linearly independent (orthogonal) vectors $\mathbf{Z}_i$, $i = 1, 2, \ldots, N$. Explicitly

$$\mathbf{Y} = \textstyle\sum_{i=1}^{12}\alpha_i\,\mathbf{Z}_i. \quad [5]$$

As a practical matter, it is our intention to reduce the upper limit of the sum to a lower value so that there are fewer terms in Eq. **5** than the maximal value $n = 12$. The truncated sum need not be an exact representation for $\mathbf{Y}$. It is well known in linear algebra that the error in such a truncation is minimized if we evaluate the expansion coefficients as scalar products

$$\alpha_i = \mathbf{Z}_i^T\cdot\mathbf{Y} = \textstyle\sum_{n=1}^{12} Z_{in}Y_n. \quad [6]$$

The length of $\mathbf{Y}$ is$\mathbf{Y}^T\cdot\mathbf{Y} = \sum_{i=1}^{12}|\alpha_i|^2$. Because the eigenvectors are normalized, $\mathbf{Z}_i^T\mathbf{Z}_i = 1$, we can interpret $|\alpha_i|^2 = (\mathbf{Z}_i^T\cdot\mathbf{Y})^2/\mathbf{Y}^T\cdot\mathbf{Y}$ as the correlation coefficient between the output vector $\mathbf{Y}$ and the N eigenvectors $\mathbf{Z}_i$ of the N by N matrix $\mathbf{X}^T\mathbf{X}$. In general, we expect that there are fewer than N, say N_{eff} in number, eigenvectors that are relevant, $\mathbf{Y} = \sum_{i=1}^{N_{eff}}\alpha_i\,\mathbf{Z}_i$. We find that for the experimental result shown in Fig. 1, keeping even just one term, $N_{eff} = 1$, is realistic, as seen in Fig. 3, because for the eigenvector that is associated with the largest eigenvalue (equals the most relevant constraint equals the constraint with the largest Lagrange multiplier, eigenvalue i =1) $\alpha_1/(\mathbf{Y}^T\cdot\mathbf{Y})^{1/2} = 0.76$. The values of the other contributions are also shown in Fig. 3, plotted such that the sum of all terms adds to unity, namely the plot shows the sum of $(\mathbf{Z}_j^T\cdot\mathbf{Y})^2/(\mathbf{Y}^T\cdot\mathbf{Y})$ from j =1 to i vs. the index i. Note that $\sum_{j=1}^{N}(\mathbf{Z}_j^T\cdot\mathbf{Y})^2/(\mathbf{Y}^T\cdot\mathbf{Y}) = 1$. Of this unity, 58% ($0.76^2$) is contributed by the eigenvector of the largest eigenvalue, i =1. The $i = 2$ term contributes 14%. The next eigenvector makes an essentially negligible contribution. There is a small but finite contribution from the $i = 8$ eigenvector. The Lagrange parameter of this constraint is already fully two orders of magnitude smaller than that for $i = 1$ (Fig. 2). It is marginal to conclude that the contribution of the eighth eigenvector, 12%, is above the noise level. Three constraints, $i = 1$, 2, and 8, specify the output essentially to within its experimental (5) error bars. One constraint, the one that is dominant in describing the phosphorylation data, $i = 1$, is dominant also in accounting for the phenotypic output as shown in Fig. 4.

In section VI, we show that the procedure as discussed here is a direct conclusion from a maximal entropy consideration.

Fig. S2 shows the weight of the P different phosphorylation events in the vector $\mathbf{Z}_i$, $i = 1$, that has the largest Lagrange multiplier. The vector is localized about one particular phosphoevent, Bcr pY 644. The result is not typical of the other eigenvectors. Mostly they are not so localized as seen in Figs. S3 and S4. Fig. S3 compares eigenvectors 1 and 11 where the latter primarily represents noise (see also *SI Appendix*, Fig. S5). Fig. S4 shows the components for the three eigenvectors most correlated with the output. Also, the second and third are localized on the Bcr and Abl phosphopeptides and, in particular, on the P

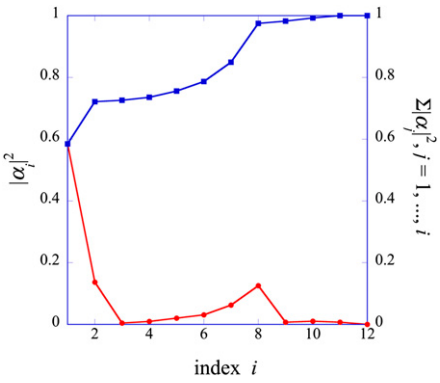

Fig. 3. The correlation coefficient between the output phenotype vector and the different eigenvectors (equals the constraints) (Eq. **6**), arranged by decreasing order of their corresponding Lagrange multipliers, λ_i. The same index labeling is as in Fig. 2. See section IV for a discussion of the correlation coefficients. The cumulative contribution of the constraints, $\sum_{j=1}^{i}|\alpha_j|^2$, is shown on the right ordinate. The dominant constraint, i = 1, accounts for 58% of the total. The first two constraints already account for 71% of the total and, as seen in Fig. 4, they almost suffice to predict the output. Adding the third fourth,... constraints does not add much to the accuracy of the description of the output. A possibly significant increase occurs when adding the eighth constraint (possibly because it is just above the noise level).

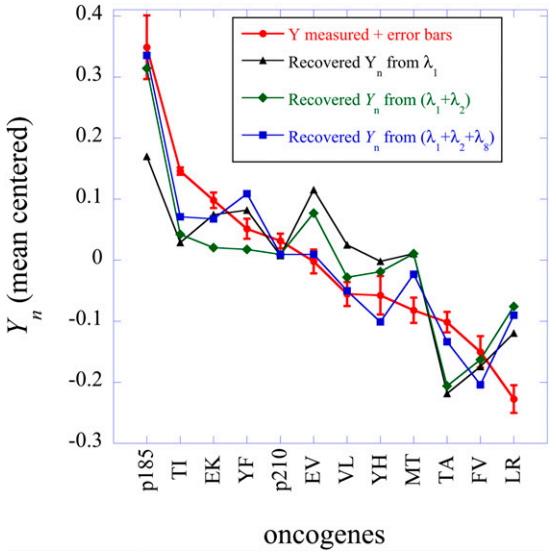

Fig. 4. The oncogenic potencies (Y_n; mean centered) and the error bars (SEM) for the different oncogenes. See *SI Appendix* and ref. 5 for details on how the potency phenotypes are defined in terms of directly measured values. Also shown are the values of the potencies predicted by using one, two, or three constraints. (The third constraint is indexed 8; Fig. 3). Using all 12 constraints will exactly recover the potencies. Note, however, that including just the two leading constraints already is rather adequate. Because the potencies are mean centered no additional input is required for the prediction as shown. To predict the absolute values we need to know one potency or an equivalent input that sets the scale.

phosphorylation loop as one might expect because Bcr-Abl ATP binding is the driving force of the signaling cascade that results in the outcome phenotype of transformation.

Control analyses for the possibility that the Bcr pY 644 ionizes more efficiently than other phosphopeptides because of its two basic residues are described in *SI Appendix* (section "control analyses for ionization efficiency"). Here, we say that even when this peptide is reduced to 1/5 of its presence, the first two constraints predict the oncogenicity with the same fidelity as was done with one constraint in the unadjusted case. These constraints are centered about a few, a minority, of the possible Bcr phosphorylation sites (see *SI Appendix*, Figs. S4 and S8). What is quite significant is that these few peptides include the ATP phosphorylation loop of the Abl kinase. We categorically state that the variance in the rows contains biological information because if we mean-center each row and scale by the variance, we remove a good deal of the structure in the data, retaining only the signal from the off-diagonal covariance. Another test is to use a data matrix where the rows are proteins and not phosphorylation sites. This test is possible by representing the intensity of each protein as an evenly weighted average of the contributions from all its phosphorylation sites. Such a compaction of the data matrix leads to a covariance matrix that needs three constraints for its characterization and to predict the phenotype. These tests lend further support to the suggestion that the fine structure in the cellular choice of phosphorylation sites carries essential biological information.

V. Relation to Other Methods

An important goal of systems biology, especially in applications toward medicine, is to be able to take a set of signaling measurements and infer something about the phenotype of a new sample. Cross-validation is a standard method for validating such inference (21). Leave one out cross-validation, LOOCV, requires leaving one sample out from the known phenotypic output and trying to infer the phenotypic output for the one sample that is left out. In *SI Appendix*, Figs. S6 and S7, we show that LOOCV works well for predicting the potency of an oncogene that is left out, using the measured potencies of the other 11 oncogenes. In comparison, Fig. 4 shows that the entire output can be predicted given not more than three output vectors and that even just one output provides a reasonable inference. These vectors are generated only from the phosphorylation data and do not use any measured potency in the prediction of the relative potencies.

In the method of principal component analysis (PCA), one seeks to find an effective reduced representation for the matrix $\mathbf{X}^T\mathbf{X}$ by diagonalizing it and sorting the eigenvectors by the size of the corresponding eigenvalues. One then uses one or more of these eigenvectors to provide a reduced rank approximation for both the matrix $\mathbf{X}^T\mathbf{X}$ and for the data matrix itself. The principal components that are retained in PCA are exactly our constraints. From the constraints we get the entire distribution (Eq. **3**). The results of PCA are only the averages over the distribution.

VI. The Information That the Data X Conveys About the Phenotypic Output Y

Shannon has argued that the information conveyed by $\mathbf{X}$ about $\mathbf{Y}$ and often denoted as $I(\mathbf{Y};\mathbf{X})$ must be given by the uncertainty about $\mathbf{Y}$ when the data $\mathbf{X}$ is not known minus any remaining uncertainty about the output $\mathbf{Y}$ once $\mathbf{X}$ is known (30):

$$I(\mathbf{Y};\mathbf{X}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}). \quad [7]$$

As in section II, H denotes the entropy and $H(\mathbf{Y}|\mathbf{X})$ denotes the entropy of $\mathbf{Y}$ when $\mathbf{X}$ is given. The entropy $H(\mathbf{Y}|\mathbf{X})$ cannot exceed $H(\mathbf{Y})$ because acquiring the data cannot increase our uncertainty about the output. Therefore, the information that $\mathbf{X}$ provides about $\mathbf{Y}$ is always positive or zero.

For the typical system biology experiment in steady state, we have argued that it is to be expected that mathematically $\mathbf{X}$ fully determines $\mathbf{Y}$. Then, when $\mathbf{X}$ is given, there is no residual uncertainty about $\mathbf{Y}$, so that $H(\mathbf{Y}|\mathbf{X}) = 0$. From Eq. **7** the information that $\mathbf{X}$ provides about $\mathbf{Y}$ is $H(\mathbf{Y})$, the entropy of $\mathbf{Y}$. In other words, knowing $\mathbf{X}$ removes any uncertainty about $\mathbf{Y}$. In *SI Appendix*, we discuss the more general case when $H(\mathbf{Y}|\mathbf{X})$ is not zero because some uncertainty about the output $\mathbf{Y}$ remains even when the data are given. This case is discussed in *SI Appendix*.

To predict, we seek a linear transformation $\mathbf{A}$ from the data to the output where $\mathbf{A}$ is to be computed from the data only. When the output is a vector, $\mathbf{A}$ is also a vector and we write

$$\mathbf{Y} = \mathbf{X}^T\mathbf{A} + \mathbf{e}_\mathbf{y}, \quad [8]$$

where the yet unknown vector $\mathbf{A}$ has P components and $\langle \mathbf{e}_\mathbf{y} \rangle = 0$ when the output is mean-centered. The transpose data matrix $\mathbf{X}^T$ has the dimensions N by P so that the output $\mathbf{Y}$ has N components, as expected.

To find $\mathbf{A}$ we appeal to the result above that the information that we can extract from the data is maximal when the entropy of $\mathbf{Y}$ is maximal. In *SI Appendix*, we show that this choice of $\mathbf{A}$ is equivalent to the procedure discussed in section IV.

VII. Concluding Remarks

Our final aim is to predict the oncogenic potency from the measured data of the intensity of phosphorylation events for each oncogene. The data can be organized as a matrix $\mathbf{X}$ where the rows are the phosphorylation events and the data for each oncogene is a column (see Fig. 1 and *SI Appendix*, Fig. S9). We show that for the common case of high throughput, more events than oncogenes, the prescription is to diagonalize the square, nonnegative, matrix $\mathbf{X}^T\mathbf{X}$. Arrange the eigenvalues by their size. If there is one eigenvalue that is by far larger, as is the case for the data examined here, there is one dominant constraint that is obtained as the corresponding eigenvector. This constraint predicts the phenotype with realistic accuracy. Including additional constraints, arranged by the decreasing size of the eigenvalues, improves the prediction.

ACKNOWLEDGMENTS. We are grateful to Professors Amos Golan and Nathan Price who acted as the internal referees of this paper. Professor Michael Fisher kindly commented on the draft manuscript. Funding for the experimental research was provided by National Institutes of Health National Human Genome Research Institute Grant HG002807 (to T.G.G.) and National Cancer Institute Grant 5U54 CA119347 (to T.G.G., J.R.H., and M.E.P.; J.R.H., principal investigator).

1. Hunter T (1998) The Croonian Lecture 1997. The phosphorylation of proteins on tyrosine: its role in cell growth and disease. *Philos Trans R Soc Lond B Biol Sci* 353: 583–605.
2. Irish JM, et al. (2004) Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* 118:217–228.
3. Rikova K, et al. (2007) Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* 131:1190–1203.
4. Janes KA, et al. (2005) A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* 310:1646–1653.
5. Skaggs BJ, et al. (2006) Phosphorylation of the ATP-binding loop directs oncogenicity of drug-resistant BCR-ABL mutants. *Proc Natl Acad Sci USA* 103:19466–19471.
6. Jaynes ET (2004) *Probability Theory: The Logic of Science* (Cambridge Univ Press, Cambridge).
7. Gibbs JW (1961) *The Scientific Papers of J. W. Gibbs* (Dover, New York).
8. Mayer JE, Mayer MG (1966) *Statistical mechanics* (Wiley, New York).
9. Ivakhno S, Armstrong JD (2007) Non-linear dimensionality reduction of signaling networks. *BMC Syst Biol* 1:27.
10. Janes KA, Lauffenburger DA (2006) A biological approach to computational models of proteomic networks. *Curr Opin Chem Biol* 10:73–80.
11. Krawitz P, Shmulevich I (2007) Entropy of complex relevant components of Boolean networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 76:036115.
12. Lezon TR, Banavar JR, Cieplak M, Maritan A, Fedoroff NV (2006) Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc Natl Acad Sci USA* 103:19033–19038.
13. Margolin AA, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 (Suppl 1):S7.
14. Nykter M, et al. (2008) Critical networks exhibit maximal information diversity in structure-dynamics relationships. *Phys Rev Lett* 100:058702.
15. Rosvall M, Bergstrom CT (2007) An information-theoretic framework for resolving community structure in complex networks. *Proc Natl Acad Sci USA* 104:7327–7331.
16. Slonim N, Atwal GS, Tkacik G, Bialek W (2005) Information-based clustering. *Proc Natl Acad Sci USA* 102:18297–18302.
17. Tkacik G, Callan CG, Jr, Bialek W (2008) Information flow and optimization in transcriptional regulation. *Proc Natl Acad Sci USA* 105:12265–12270.
18. Ziv E, Nemenman I, Wiggins CH (2007) Optimal signal processing in small stochastic biochemical networks. *PLoS One* 2:e1077.
19. Callen HB (1985) *Thermodynamics and an Introduction to Thermostatics* (Wiley, NY).
20. Janes KA, et al. (2004) Cue-signal-response analysis of TNF-induced apoptosis by partial least squares regression of dynamic multivariate data. *J Comput Biol* 11: 544–561.
21. Picard RR, Cook RD (1997) Cross-Validation of Regression Models. *J Am Stat Assoc* 79: 575–583.
22. Gorre ME, et al. (2001) Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* 293:876–880.
23. Kinsey JL, Levine RD (1979) A Performance Criterion for Information Theoretic Data Analysis. *Chem Phys Lett* 65:413–416.
24. Remacle F, Levine RD (2009) The elimination of redundant constraints in surprisal analysis of unimolecular dissociation and other endothermic processes. *J Phys Chem A* 113:4658–4664.
25. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 7:142.
26. Bar-Even A, et al. (2006) Noise in protein expression scales with natural protein abundance. *Nat Genet* 38:636–643.
27. 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.
28. Dinur U, Levine RD (1975) On the entropy of a continuous distribution. *Chem Phys* 9: 17–27.
29. Ahmed NA, Gokhale DV (1989) Entropy expressions and their estimators for multivariate distributions. *IEEE Trans Inf Theory* 35:688–692.
30. Ash RB (1990) *Information Theory* (Dover, Mineola, NY).