J Stat Phys (2011) 144:429–442
DOI 10.1007/s10955-011-0215-x

Whose Entropy: A Maximal Entropy Analysis of Phosphorylation Signaling

F. Remacle · T.G. Graeber · R.D. Levine

Received: 9 December 2010 / Accepted: 26 April 2011 / Published online: 13 May 2011

Abstract High throughput experiments, characteristic of studies in systems biology, produce large output data sets often at different time points or under a variety of related conditions or for different patients. In several recent papers the data is modeled by using a distribution of maximal information-theoretic entropy. We pose the question: 'whose entropy' meaning how do we select the variables whose distribution should be compared to that of maximal entropy. The point is that different choices can lead to different answers. Due to the technological advances that allow for the system-wide measurement of hundreds to thousands of events from biological samples, addressing this question is now part of the analysis of systems biology datasets. The analysis of the extent of phosphorylation in reference to the transformation potency of Bcr-Abl fusion oncogene mutants is used as a biological example. The approach taken seeks to use entropy not simply as a statistical measure of dispersion but as a physical, thermodynamic, state function. This highlights the dilemma of what are the variables that describe the state of the signaling network. Is what matters Boolean, spin-like, variables that specify whether a particular phosphorylation site is or is not actually phosphorylated. Or does the actual extent of phosphorylation matter. Last but not least is the possibility that in a signaling network some few specific phosphorylation sites are the key to the signal transduction even though these sites are not at any time abundantly phosphorylated in an absolute sense.

F. Remacle
Département de Chimie, B6c, Université de Liège, 4000 Liège, Belgium

T.G. Graeber · R.D. Levine (✉)
Crump Institute for Molecular Imaging and Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, University of California, Los Angeles 90095, USA
e-mail: rafi@fh.huji.ac.il

R.D. Levine
Department of Chemistry and Biochemistry, University of California, Los Angeles 90095, USA

R.D. Levine
Institute of Chemistry, The Hebrew University of Jerusalem, Jerusalem, Israel

Keywords Information theory · Prior distribution · Systems biology · Signal transduction · High throughput experiments · Phosphoproteomics

1 Introduction

In physics and chemistry the physical entropy as introduced in thermodynamics is the same as the information theoretic, IT, entropy if and only if we are dealing with the distribution of quantum states. (By IT entropy we mean the expression that Shannon introduced on axiomatic grounds). Computing the IT entropy using any other distribution, say the distribution of energy states, will lead to an entropy not equal to the thermodynamic entropy. Another view of the issue is that a measured distribution need not be one of maximal IT entropy. A celebrated example is the Otto Stern experimental verification of the Maxwell-Boltzmann velocity distribution in a gas. The agreement that was obtained was not acceptable to the high standards of Stern. Einstein then pointed out that what Stern measured is the flux distribution of the atoms effusing out of a container whereas a distribution of maximal IT entropy is the velocity distribution. A modern version of this problem, very familiar to people in reaction dynamics [1], is that a mass spectrometer is a flux detector whereas a laser probe is a number density detector. Of course, using a simple Jacobian (that plays the role of a prior distribution) will convert a distribution in one variable into a distribution of another variable. BUT there must be a motivated decision as to which variable to use to obtain that distribution whose IT entropy is maximal. (Or, equivalently, whose prior is uniform.) In physics and chemistry the prior comes from the axiom that the distribution of maximal IT entropy is the distribution of quantum states. This works well when the prior can be computed [2]. But it is a much too fine grained description when we deal with biological systems.

One option is to try to determine the prior from the experimental data, an approach pioneered in [3]. As discussed therein, this requires experimental data measured for several different values of the constraints. The prior is that part which is invariant. In [3] we do not assume that there is an invariant part but let the numerical analysis identify one, if there is. Another option is to try to reason out on biophysicochemical grounds what would be a reasonable prior distribution [4]. One can then seek to validate the choice, but the methodology available for doing so is one of the open questions addressed in this article.

The ultimate answer to the question which are the variable(s) whose distribution is of maximal IT entropy is to be settled by experiment. This is useful if there is additional data that can be used to validate the choice made. This is the course of action followed in [4]. There are intermediate answers such as [5, 6] that the distribution of maximal entropy must be invariant under the symmetry operations that the system admits. A special case, discussed in detail in [7], is that of quantum symmetries that exclude certain outcomes as in the Bose-Einstein or Fermi-Dirac statistics. It is shown in [7] how to write the prior distribution for both cases.

From an information theoretic point of view the distribution of maximal entropy is the distribution that is consistent with the data at hand and is otherwise least informative. The prior distribution is obtained in the same procedure of maximizing the entropy when the only constraints that are imposed are the universal ones, i.e., those that are well established for a given situation. A long recognized example is the conservation of energy for an isolated mechanical system. Sometimes it is loosely stated that the prior distribution is the distribution whose entropy is maximal in the absence of constraints. Strictly speaking, the prior distribution is the distribution whose entropy is maximal in the absence of constraints that are special to the system at hand.

From a statistical point of view a distribution of maximal entropy is the most probable distribution in that it is the distribution that can be realized in the largest number of experiments. This is sometimes called the Boltzmann view. What we would like to achieve is a statistico-mechanical point of view that has also a thermodynamic analog. In doing so we follow the guidance of Gibbs in his treatment of mechanical systems. The reference to the connection to thermodynamics as 'an analogy' follows Gibbs.

When maximizing the entropy but constraining the distribution to be consistent with what we do know, the constraints, there arise 'parameters' that act in analogy to thermodynamic potentials. Technically the parameters arise as Lagrange multipliers that are introduced in the process of seeking a maximum subject to constraints. When the corresponding Lagrange multiplier is not zero the conjugate constraint forces a lower value of the entropy and can be used to specify the direction of response of the system to perturbations (principle of Le Chatelier [8]).

There is one trivial but useful technical point about constraints. It does no harm to impose too many constraints [9]. The conservation of the number of molecules of each chemical species at chemical equilibrium is a well-known example. At equilibrium these numbers are indeed conserved and unchanging with time. The corresponding Lagrange multipliers are the chemical potential of the different species [10]. However, far fewer constraints, namely the conservation of the number of atoms of each chemical element, are enough to characterize the system by the method of maximal entropy. Moreover, these constraints remain valid when the system is out of equilibrium.

In the recent literature of theoretical biology the principle of entropy maximization has been mostly applied to biological networks [11–21]. Reported examples include the inference of genetic interaction networks from microarray data or to inferring network modularity or to neural networks. In most of these works the constraints are the correlation between the variables. It has therefore been the practice to describe the connectivity in the language developed for the Ising model, with spins being up or down. For neural networks it is natural to think of a neuron having two distinct states, say firing or silent. It is however not completely obvious why a neuron is like a spin in that in the absence of constraints its two states are equally probable. But the distribution of spin orientations that are obtained through the method of maximal entropy necessarily assume this as the prior distribution. What is not clear for neurons is even less clear for genes. It is possible to approximately think of a gene as being on or off. But microarray data and even more so, the deep sequencing method, provide much more details about the expression levels of genes.

Cancer biologists have learned much about the signaling pathways that are disrupted in cancer and disregulate normal growth. Here we specifically address the leukemic transformation to chronic myelogenous leukemia (CML) driven by the oncogenic kinase Bcr-Abl. This mutation activates multiple downstream signaling pathways that combine in a not fully understood manner to contribute to leukemia. The main proximal signaling path is phosphorylation of proteins at the site of the amino acid Tyrosine, Y. A protein can be phosphorylated at one or more Y sites. Mass spectrometry-based experimental proteomic procedures typically enzymatically digest proteins into smaller units called peptides. A peptide fragment can be unphosphorylated or it can carry one or more phosphorylated Y sites. Affinity-based purification techniques allow for the enrichment of phosphorylated peptides from the bulk of unphosphorylated peptides. But these techniques also wash away the unphosphorylated version of the detected phosphopeptides. Thus, many informative phospho mass spectrometry (MS)-based datasets lack data on the unphosphorylated state.

In this paper we discuss an analysis of quantitative MS proteomics data measured for Bcr-Abl. The data was reported in earlier publications [4, 22] and is shown below as a heat

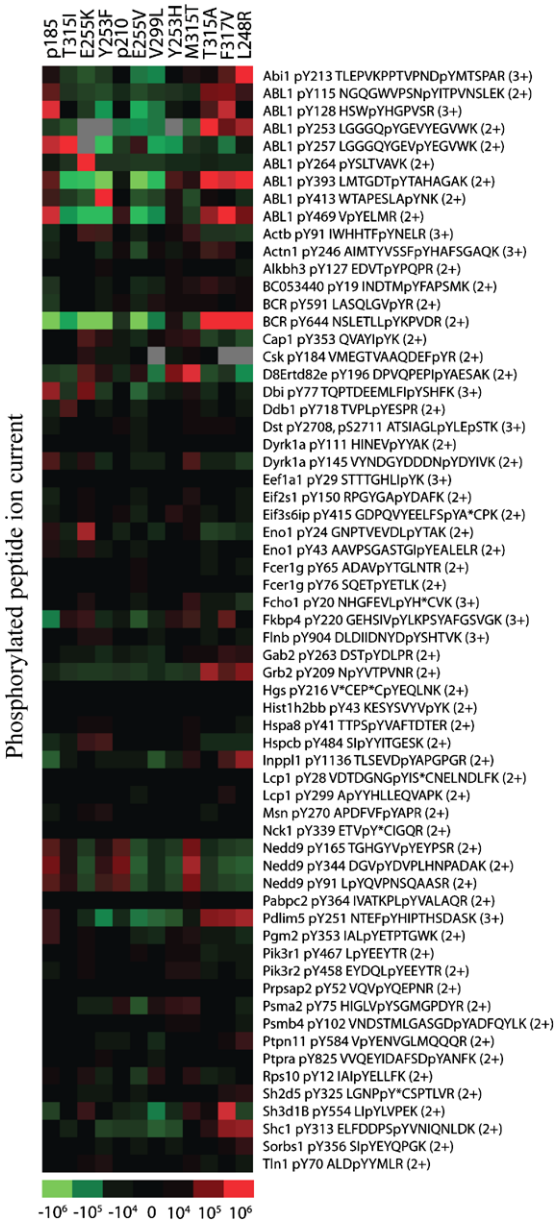

Fig. 1 Heat map representation of the measured phosphorylation data matrix **X** for cells transformed by $N = 12$ different Bcr-Abl oncogene isoforms. The phosphorylation data matrix **X** represents relative levels of each phosphorylation event (*row*) across the isoforms (*columns*). The 73 rows are mean-centered and correspond to the mass spectrometry ion current peak integration values as indicated by the green to *red scale bar* at the *bottom*

map in Fig. 1. This data matrix will be denoted as **X**. It is a rectangular matrix with twelve columns, one for each mutant of Bcr-Abl. The number of rows, seventy three, is the number of phosphorylation events. The entries in the rows of **X** are the steady state phosphorylation level of a given peptide, as identified in the heat map, for the twelve different oncogenes. While we discuss the concept of the appropriate prior distribution here in the context of phosphorylation signaling, parallel issues arise in many other biological contexts such as cytokine ligand-driven paracrine signaling.

2 Maximal Entropy

The first stage in the application of a procedure of maximal entropy is to answer the question 'whose entropy?' In other words we need to specify the limiting situation for which the entropy is at its global maximum. When making this choice is not mentioned explicitly it is done implicitly but the choice *is* being made. It is straightforward to tell what is the choice by inspecting the expression adopted for the entropy. In the phosphorylation signaling analysis of Sect. 5, we explore whether the variables whose entropy is to be maximized is the Boolean state of phosphorylation, the measured values of the phosphorylation sites, or variance-normalized values of the phosphorylation sites.

By making the choice of 'whose entropy' we mean the follows. Take $\mathbf{n}$ as a, possibly vector-like, specification of the states of the system and $P(\mathbf{n})$ as the probability of the state and say that the entropy is written in its IT form

$$S = -\sum_{\mathbf{n}} P(\mathbf{n}) \ln(P(\mathbf{n})) \tag{1}$$

By the very definition of the entropy through (1) the choice has been made that at the global maximum of the entropy of the states labeled $\mathbf{n}$ are equally probable. The specification $\mathbf{n}$ thus defines the prior that is the states that in the absence of reasons to the contrary are equally likely.

Once the global maximum is defined one can begin to think of constraints as the conditions that can cause the entropy to reduce in value. When constraints are imposed the maximal value of the entropy is necessarily either lower than or equal to its global maximum. The value is lower if the constraints that are assumed actually do restrict the range of possible distributions. In this case at the maximal value we select from all those distributions that do satisfy the constraints the one distribution whose entropy is maximal. It is a familiar result that if there is such a distribution, it is unique. If the constraints do not lower the value of the entropy at its maximum then the constraints are not relevant.

In practice, and notably so for the high throughput data sets of systems biology, experimental noise can limit the ability to distinguish between relevant and irrelevant constraints. Adding a constraint can reduce the entropy but one has to assess if the decrease in the value of the maximal entropy due to the additional constraints warrants the introduction of this constraint [23, 24]. This issue arises when we are aiming for a too perfect match to the observed data because the experimental values are only known to a finite precision. It is the Lagrange multipliers whose numerical value are small that cause the fit to be too perfect. In other words, from the point of view of the actual data these multipliers are fitting the noise and not any experimentally resolvable reality. The value of these small Lagrange multipliers should be put to zero implying that the corresponding constraints are not relevant.

3 Distribution of Maximal Entropy: The Analogue of a Grand Canonical Ensemble

In the phosphorylation signaling data of Sect. 5, different phosphoprofilings taken for different mutations of the oncogene can differ in the number, N_i, of molecules of species i where i is a label for phosphorylation of a particular peptide on a particular Y site. The distribution that we have is of the number N_i, of readings of the phosphorylated peptide i. Note that this is a mean over a series of independent measurements. We take the simplest possible choice for the constraints over the distribution of phosphorylation profiles. We take as the constraint that the mean number $\bar{N}_i$ of copies of each species i is fixed. The other conserved quantity is

the energy. Searching amongst all the distribution of given number means and mean energy, we determine that distribution whose entropy is maximal. Then the probability of a system in a particular composition is written in the notation used in textbooks to describe the grand ensemble as

$$P(N_1, N_2, \ldots) = \exp\{\beta(\Sigma_i \mu_i N_i - E)\}/\Xi \tag{2}$$

To preclude the analogy with the textbook case to be too close note that here the distribution is over an ensemble of independent and not interacting systems. This means that for us each measurement of a set of numbers $N_1, N_2, \ldots$ of phosphorylation events is a separate experiment measured for a different oncogene. However, as in textbooks, β is the Lagrange multiplier that is determined by the conservation of energy and, in the textbooks context, can be related to a temperature T as $\beta = 1/kT$ where k is Boltzmann's constant. The μ_i's are the chemical potentials as introduced in the thermodynamics of systems of more than one component. In a statistical approach the μ_i's are the Lagrange multipliers that correspond to the given mean number of species i. As Planck already noted, it is actually more convenient to work with $\alpha_i = \beta\mu_i$ sometimes called the Planck potentials. Ξ is the grand partition function. It is a function of all the Lagrange multipliers and its role is to insure that the sum of the probability over all possible compositions yields one,

$$\Xi = \sum \exp\{\beta(\Sigma_i \mu_i N_i - E)\} \tag{3}$$

Sometimes one writes $\Xi = \exp(\lambda_0)$ where λ_0 is the Lagrange multiplier that insures the conservation of probability. Equation (3) defines the partition function as a function of the other Lagrange multipliers.

Say now that we make a small change in the value of the chemical potential μ_i from its current equilibrium value to some new value $\mu_i + \delta\mu_i$. We have defined the Planck potential (which is the Lagrange multiplier of the constraint of the conservation of concentration) as $\alpha_i = \beta\mu_i$ and so we need to specify that the change in μ_i is made isothermally. This change in the chemical potential can change the equilibrium mean concentration of all species from $\bar{N}_j$ to $\bar{N}_j + \delta\bar{N}_j$, all j. To compute the change in concentrations we need to consider the change in the probability as given in (2). To do so we make use of the definition of the mean concentration

$$\bar{N}_j = \sum N_j P(N_1, N_2, \ldots) \tag{4}$$

The summation in (4) is over all the possible compositions, each weighted by its probability $P(N_1, N_2, \ldots)$. We denote this averaging by an over bar. We will also need the corresponding change of the grand partition function computed from (3), $\delta \ln(\Xi)/\delta(\beta\mu_i) = \bar{N}_i$.

Taking the variation of the probability when the particular chemical potential μ_i is changed $\delta P(N_1, N_2, \ldots) = \beta\delta\mu_i(N_i - \bar{N}_i)P(N_1, N_2, \ldots)$ we can write:

$$\begin{aligned}
\delta\bar{N}_j &= \sum N_j \delta P(N_1, N_2, \ldots) = \sum (N_j - \bar{N}_j)\delta P(N_1, N_2, \ldots) \\
&= \beta\delta\mu_i \sum (N_j - \bar{N}_j)(N_i - \bar{N}_i) P(N_1, N_2, \ldots) \\
&= \beta\delta\mu_i \overline{(N_j - \bar{N}_j)(N_i - \bar{N}_i)} \equiv \Sigma_{ji}\beta\delta\mu_i
\end{aligned} \tag{5}$$

Note that the conservation of normalization implies that the average change in the probability must be zero, $0 = \sum \delta P(N_1, N_2, \ldots)$ and we have used this result in the derivation

above. In the last line in (5) we have avoided writing the summation over all compositions by the use of the over bar to designate an average over the probability $P(N_1, N_2, \ldots)$, which is the notation introduced in (4). The last line defines the covariance matrix $\sum$.

For small isothermal variations in all the chemical potentials we have the general equation of change that a form of (5) extended to all possible small variations of the distribution

$$\delta \bar{N}_j = \beta \sum_i \overline{(N_j - \bar{N}_j)(N_i - \bar{N}_i)} \delta \mu_i \equiv \beta \sum_i \Sigma_{ij} \delta \mu_i \tag{6}$$

It is a linear sum as expected in general for a combined effect of small perturbations.

4 Distribution of Maximal Entropy: The Gaussian Distribution

For small deviations from the distribution of maximal entropy one can rewrite the exponential distribution (2) as a Gaussian. In a thermodynamic context this was pointed out by Einstein when he noted that thermodynamics describes not only a stable equilibrium but also small fluctuations about the equilibrium. By expanding the logarithms of the exponential distribution we have

$$\begin{aligned} \ln P(N_1, N_2, \ldots) &= \ln P(\bar{N}_1, \bar{N}_2, \ldots) + \beta \sum_{i,j} \frac{\partial \mu_i}{\partial \bar{N}_j} (N_i - \bar{N}_i)(N_j - \bar{N}_j) \\ &= \ln P(\bar{N}_1, \bar{N}_2, \ldots) + \beta (\delta \mathbf{N})^T \Sigma^{-1} \delta \mathbf{N} \end{aligned} \tag{7}$$

There is no linear term because the expansion is around equilibrium. The quadratic form for the deviation from equilibrium is known as the stability of equilibrium because the covariance matrix Σ is, from its definition, positive definite. It can however be that the covariance matrix is only semi-positive definite and then the constraints are not all linearly independent [23, 25].

In (7) the covariance matrix is the covariance matrix that can be computed from the distribution given by (2). This is the distribution where the only constraints are the means. It is however possible to also impose the covariance matrix as a constraint. This means that the constraints are not only the $\bar{N}_i$'s but also the second moments $\overline{N_i N_j}$. If the second moments are or are not needed as independent constraints has to be determined by the data. It does not harm to impose the $\bar{N}_i$'s and also the covariances $\Sigma_{ij} = \overline{N_i N_j} - \bar{N}_i \bar{N}_j$ as constraints. This is since, as discussed above, unnecessary constraints do not affect the distribution whose entropy is maximal. Deviations from the linear response, as defined through (6), means that the covariance cannot be predicted using only the conservation of the mean numbers. It is then necessary to specifically include the elements of the covariance matrix as constraints. This must be done with care because it also covers cases when the distribution does not have a single peak. The problem can arise even for a distribution of just one variable when the distribution is specified to be too broad as compared to a distribution of maximal entropy for a given mean.

5 The Entropy of the Distribution of Phosphorylation Event

Turning to the case of phosphorylation signaling data, each entry in the data matrix $\mathbf{X}$ is a measurement of a phosphorylation event for particular conditions, [4, 22]. We begin by

taking it that in the absence of constraints that require otherwise, all the measured phosphorylation events are equally probable. In this case a heat map of matrix $\mathbf{X}$ should be uniform in color at maximal entropy. But the heat map is clearly not uniform in color, see Fig. 1. It contains biological and chemical information. We will seek to determine how many constraints are necessary to capture this information. Then, we will compare how such results of a maximal entropy analysis compare when differently summarized variables of the measured phsophorylation values are used as the variables for the maximal entropy prior.

It is possible to argue that even when we do not know otherwise it is not reasonable to take all phosphorylation events to be equally probable. This is because often only relative phosphorylation values are measured, and thus unnormalized comparisons can be misleading. In particular, the amino acid sequence of any given peptide can influence its ionization efficiency in the mass spectrometer. It can also be argued that the same peptide, singly and doubly ionized should be counted as two distinct species. But note that this is an MS experimental state difference and not a biological difference. On the other hand, two peptides of the same mass but each one singly phosphorylated on a different Tyrosine residue cannot in all cases be resolved by MS but are biologically distinct. It is therefore possible to argue that the measurements should be scaled as part of the answer to 'whose entropy', for example by operating on each row of the data matrix to make it mean centered and scaling each entry in the row by the variance of the row [26, 27]. Another argument for scaling encountered in systems biology is that low abundance events can be highly biological relevant. For example, regulatory proteins (transcription factors, kinases) are often expressed at relatively low levels (as little as a few copies per cell) compared to structural proteins [27, 28].

$\mathbf{X}$ is a matrix whose dimensions are the number, P, of different phosphorylation events, the row labels times the number, N, of different mutants of the oncogene, the column labels. Therefore we can also regard $\mathbf{X}$ not as a matrix but as a sample of N readings of the 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 they are Gaussian distributed. This means that the distribution is characterized by the means and covariances. 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 matrix of second moments $\mathbf{X}\mathbf{X}^T$ with elements indexed by the phosphorylation events

$$(\mathbf{X}\mathbf{X}^T)_{pq} = \sum_{n=1}^{N} (\mathbf{X})_{pn} (\mathbf{X})_{qn} \tag{8}$$

cannot be inverted. It is necessarily singular since its rank cannot be higher than the smaller dimension of the matrix $\mathbf{X}$ namely N. In other words, $\mathbf{X}\mathbf{X}^T$ can have no more than N non-zero eigenvalues. In matrix algebra it is shown that these are the same as the non-zero eigenvalues of the N by N matrix $\mathbf{X}^T\mathbf{X}$. This is a matrix whose elements are indexed by the oncogenes. The rank of the matrix $\mathbf{X}^T\mathbf{X}$ can be N but it can also be lower if the N columns of the data are not linearly independent. This is not a gratuitous warning, if, as is often done in practice, the rows of the data matrix $\mathbf{X}$ are mean centered so that the sum of entries is zero. Then the N'th column of $\mathbf{X}$ is fully determined by the other $N-1$ columns and the rank of $\mathbf{X}^T\mathbf{X}$ will be no more than $N-1$. In the experiments that we report on, without mean centering, the columns of the raw $\mathbf{X}$ are linearly independent also in a practical sense meaning that no eigenvalue is dangerously close to zero.

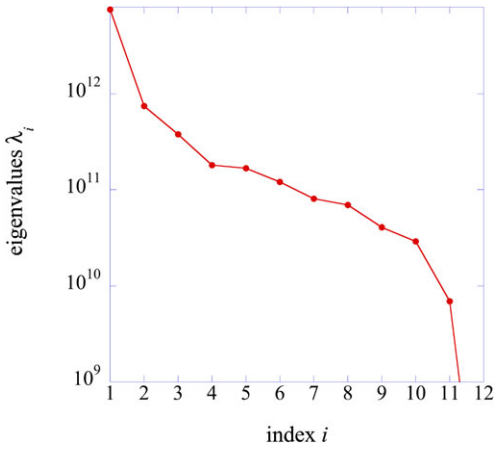

Fig. 2 The eigenvalues of the matrix $\mathbf{X}^T\mathbf{X}$ for mean centered rows of the matrix $\mathbf{X}$. In the mean-centered case, there are $N-1$ non-zero eigenvalues, $N = 12$

6 The Constraints

Using the data reported in [4, 22] and shown as a heat map in Fig. 1 we diagonalize the N by N matrix $\mathbf{X}^T\mathbf{X}$. The diagonalization determines a set of N (orthogonal and normalized) eigenvectors $\mathbf{z}_i$, $i = 1, 2, \ldots, N$. Each such eigenvector has N components, one for each oncogene. As long as the number of outputs is smaller than the number, P, of phosphorylation events the rank of the data matrix $\mathbf{X}$ is N and there will be N linearly independent eigenvectors of $\mathbf{X}^T\mathbf{X}$. The corresponding eigenvalues are denoted λ_i

$$(\mathbf{X}^T\mathbf{X})\mathbf{z}_i = \lambda_i \mathbf{z}_i, \quad i = 1, 2, \ldots, N \tag{9}$$

To each one of the N eigenvalues of the N by N matrix $\mathbf{X}^T\mathbf{X}$ there corresponds an eigenvector of the P by P matrix $\mathbf{X}\mathbf{X}^T$ with the same eigenvalue

$$(\mathbf{X}\mathbf{X}^T)\mathbf{Z}_i = \lambda_i \mathbf{Z}_i, \quad i = 1, 2, \ldots, N \tag{10}$$

All the remaining $P - N$ eigenvectors of $\mathbf{X}\mathbf{X}^T$ correspond to a zero eigenvalue. One shows, [4, 9] that in the Gaussian approximation the eigenvalues λ_i, $i = 1, 2, \ldots, N$ are the Lagrange multipliers. The p'th component of each eigenvector $\mathbf{Z}_i$, $p = 1, 2, \ldots, P$, is the value of the p'th phosphorylation event in the i'th constraint. For more on the eigenvectors of $\mathbf{X}\mathbf{X}^T$ see the Appendix.

Figure 2 is a plot of the eigenvalues in decreasing order plotted on a logarithmic scale to emphasize that the largest eigenvalue is largest by far etc. Note that because the rows are mean centered there are 11 eigenvalues that are non-zero.

Figure 3 shows the weight of the P different phosphorylation events in the vector $\mathbf{Z}_i$, $i = 1$, that has the largest Lagrange multiplier. It is seen that the vector is very much localized about one particular peptide, Bcr pY 644. The result is not typical of the other eigenvectors. Mostly they are not so localized. Figure 4 compares eigenvectors 1 and 11 where the latter mostly represents noise. Figure 5 shows the components for the second and third eigenvectors and these are localized on the Bcr and Abl peptides as one might expect since Bcr-Abl is the driving force of the signaling cascade that results in the outcome phenotype of transformation.

A different view of Fig. 3 is shown in Fig. 6. The motivation is that the dominant event in Fig. 3, is the phosphorylation of the Bcr pY 644 peptide. This peptide contains two basic

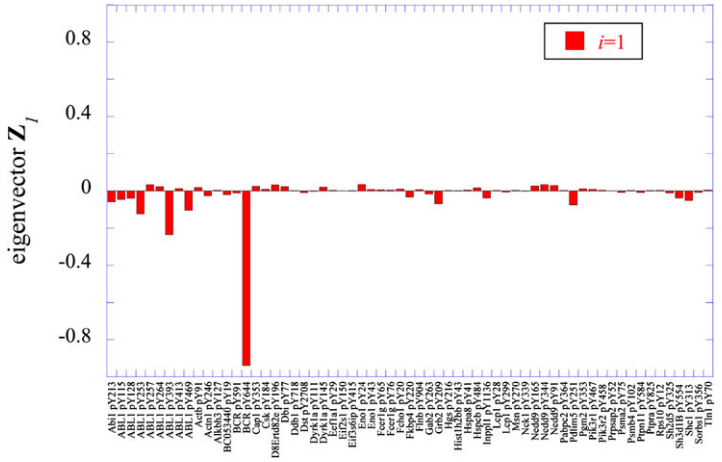

Fig. 3 The amplitudes of the eigenvector $\mathbf{Z}_1$ of the $\mathbf{X}\mathbf{X}^T$ matrix for the largest eigenvalue $i = 1$. Note that this vector is mainly localized on the Bcr pY 644 peptide

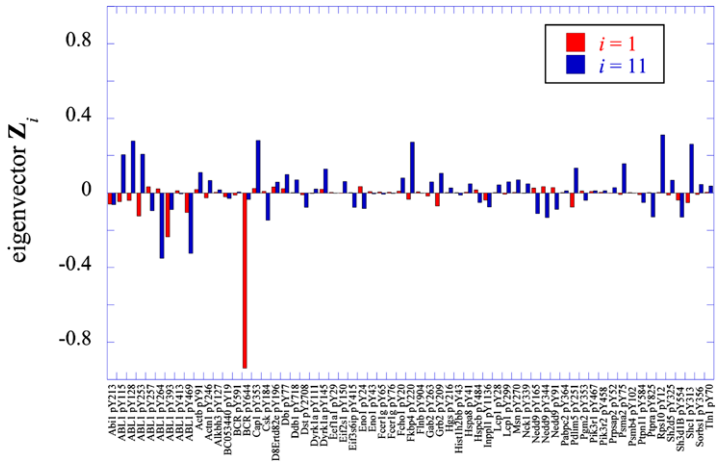

Fig. 4 The amplitudes of the eigenvectors $\mathbf{Z}_1$ and $\mathbf{Z}_{11}$ of the $\mathbf{X}\mathbf{X}^T$ matrix. The eigenvector $\mathbf{Z}_{11}$ corresponds to the smallest non-zero eigenvalue, $\lambda_{11} = 6.9 \times 10^{10}$, see Fig. 2

residues, one more than most other peptides. Typically, ionization efficiency in MS positive ion mode increases with increasing number of basic residues because the basic residues more readily accept positive charge. Thus it is possible that this peptide ionizes more efficiently than the other phosphopeptides of the data set resulting in elevated measured intensity value. In Fig. 6 we therefore use a different data matrix $\mathbf{X}$. In this data file we do not group together ions of the same peptide that have the same mass but different charge. Such ions are detected separately since mass spectrometry measures the mass/charge ratio. It is seen that the vector that has the largest Lagrange multiplier is localized about two events but it is the same peptide, Bcr pY 644, in two different charge states.

We emphasize that the localization seen in Fig. 6 is dependent on our *not* scaling the observations for the same phosphorylation event for different oncogenes to have the same

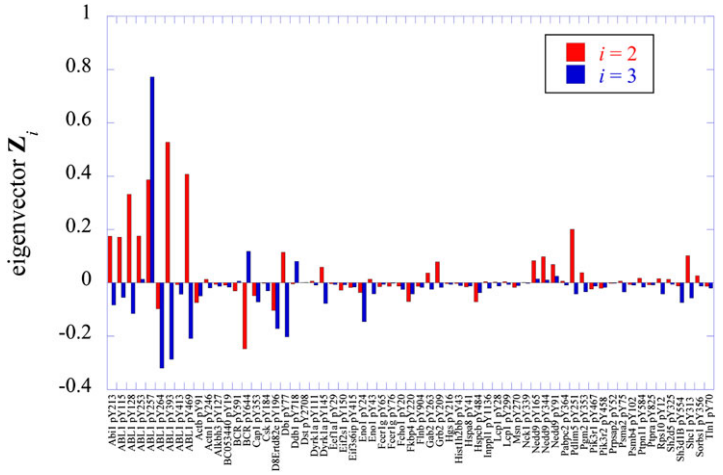

Fig. 5 The amplitudes of the eigenvectors $\mathbf{Z}_2$ and $\mathbf{Z}_3$ of the $\mathbf{X}\mathbf{X}^T$ matrix. The corresponding eigenvalues are $\lambda_2 = 7.4 \times 10^{12}$ and $\lambda_3 = 3.8 \times 10^{12}$. Note how the eigenvectors are mainly localized on the Abl peptides

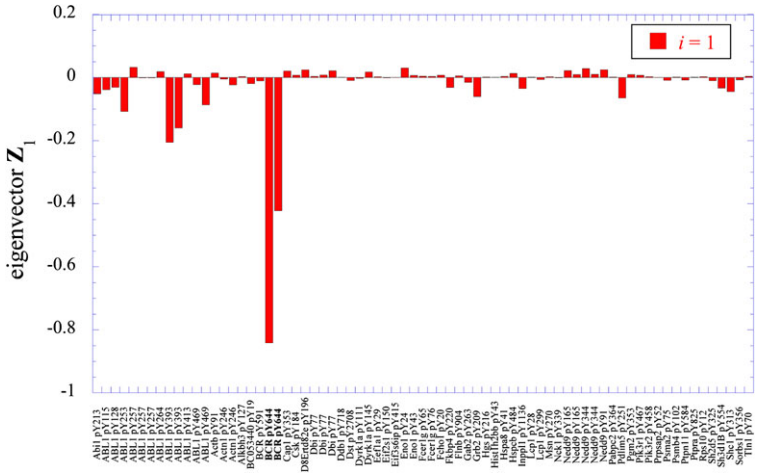

Fig. 6 The amplitudes of the eigenvector of the largest eigenvalue for the $\mathbf{X}\mathbf{X}^T$ where the different charge state of a given peptide are distinguished as different row in $\mathbf{X}$. One can see that the largest amplitudes are on the two charge states of the Bcr pY 644 peptide

variance. If one does insure that all the rows of $\mathbf{X}$ have a variance of unity the resulting matrix is fairly grey and devoid of much structure. It is possible to diagonalize $\mathbf{X}^T\mathbf{X}$ using a data matrix $\mathbf{X}$ with 'standardized' entries. The eigenvectors are far less localized on the phosphorylation events and it requires independent validation before concluding that the residual structure seen corresponds to real biology beyond noise in the data. Additional comments on the use of standardized values for the variables are offered in an Appendix.

A more extreme test is to completely delete the entry for Bcr pY 644 from the data matrix. Applying our maximal entropy approach to the $\mathbf{X}$ matrix so adjusted, we find that the new first and second constraints are localized on the Bcr and Abl peptides, similar but

to greater extent than on the original second and third eigenvectors. The first eigenvector in the original analysis is therefore a meaningful representative of the real biology in the data. It furthermore suggests that phosphorylation events are a meaningful set of events and that their distribution of maximal entropy, with a uniform prior, provides a robust representation of reality.

A yet different kind of test, is by compaction of the data matrix. Throughout we emphasized that the fine structure namely the site-specific phosphorylation is the carrier of the information. One can however identify each site as determined by mass spectrometry with the protein that is being phosphorylated. Therefore it is possible to average over the intensities of all the phosphorylations of a given protein. Thereby one has a data matrix in the basis of proteins. Unlike the results shown in Fig. 2, here the eigenvalues of the matrix $\mathbf{X}^T\mathbf{X}$ are not well separated in magnitude. As is to be expected, this eigenvector has a high component on Bcr-Abl but it also has not small weight on a few other proteins that are downstream.

7 Discussion

For the phosphorylation signaling data of Fig. 1, we explored several variable choices for 'whose entropy' in the context of maximal entropy analysis: a Boolean-like representation, directly measured phosphorylation site data, variance-normalized phosphorylation site data, or protein-collapsed phosphorylation data. It is not the case that a Boolean like representation of the signaling network, meaning a Tyrosine site is or is not phosphorylated, represents the data well. Elsewhere we provided an additional criterion: can the data predict the potency of the different mutants of the oncogene. Also by this test the Boolean approximation is not a reasonable choice [4]. Using standardized variables, meaning centering the phosphorylation levels of a peptide for different oncogenes about zero and dividing by the variance is more realistic than a Boolean representation. For these variables two constraints are needed to account for the most of the deviations from a grey data matrix (data not shown). Moreover these constraints are centered on the Bcr-Abl peptides. Using the data matrix where the levels of peptides that are fragments of the same protein are grouped together leads to a dominant constraint but one that has weight on quite a few proteins. Only the directly measured data matrix has one eigenvalue that is overwhelmingly larger, see Fig. 2, and therefore has one clearly dominant constraint. This constraint centers attention of a particular phosphorylation site, Bcr pY 644, see Fig. 3. It remains to be seen if this is biologically meaningful. What is already shown to be biologically relevant is that the data matrix $\mathbf{X}$ does quite well by the criterion of the ability to predict the potency of the different mutants of the oncogene [4].

Empirically we therefore favor the use of actual phosphorylation events as the variables. Can we do better? How can one determine a prior that is the distribution at the global maximum of the entropy? One possible solution was explored in Ref. [3]. This searches for how the different columns in the data deviate from a mean. In the present context this theoretical prescription implies the prior to be the phosphorylation level averaged over all *oncogenes*. When formulated such that at the prior all oncogenes are equally effective the constraints on the real system highlight the role of individual oncogenes. A prior distribution chosen such that all oncogenes are equally potent is *not* uniformly distributed over the phosphorylation events. For such a prior distribution the biological information provided by the experiment, [22] is provided by the differences between the oncogenes.

Acknowledgements FR is director of research of FNRS, Belgium.

Appendix: Using Variance Scaled Variables

$\mathbf{X}$ is a matrix whose dimensions are the number, P, of different phosphorylation events, the row labels times the number, N, of different mutants of the oncogene, the column labels. Therefore we can also regard $\mathbf{X}$ not as a matrix but as a sample of P readings of the row vector $\mathbf{X}_p$ of N components, $\mathbf{X}_p \equiv (X_{p1}, X_{p2}, \ldots, X_{pN})$. p is a label of the phosphorylation events, $p = 1, 2, \ldots, P$. The different components of $\mathbf{X}_p$ differ because oncogenes differ in their phosphorylation strength and specificity. We can therefore try to standardize the variables by making the entries for each row mean centered and by dividing by the standard deviation of the entries of the row so that each row has unit variance. We call the new data matrix $\tilde{\mathbf{X}}$ where on can write

$$\tilde{\mathbf{X}} = \boldsymbol{\sigma}^{-1}\mathbf{X}$$

where σ is a square matrix with non-zero entries only along the diagonal, each entry being the standard deviation of a row of the original matrix $\mathbf{X}$.

The matrix $\mathbf{X}\mathbf{X}^T$ is non-negative and can therefore be written in its spectral form

$$\mathbf{X}\mathbf{X}^T = \sum_i \omega_i^2 \mathbf{Z}_i \mathbf{Z}_i^T$$

where $\mathbf{Z}_i$ is an eigenvector with the non-negative eigenvalue ω_i^2, $\mathbf{X}\mathbf{X}^T\mathbf{Z}_i = \omega_i^2\mathbf{Z}_i$, see (10) where here we write the non-negative eigenvalues as $\lambda_i = \omega_i^2$. We take the eigenvectors to be orthonormal. Each row of $\mathbf{X}$ is first mean centered so the rank of $\mathbf{X}\mathbf{X}^T$ is at most $P - 1$ and many of the eigenvalues will be zero. We rank the eigenvalues by their size. The largest corresponds to the most important constraint in the procedure of maximal entropy [3, 9].

For standardized variables we have a corresponding expression

$$\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \boldsymbol{\sigma}^{-1}\mathbf{X}\mathbf{X}^T\boldsymbol{\sigma}^{-1} = \sum_i \omega_i^2 (\boldsymbol{\sigma}^{-1}\mathbf{Z}_i)(\boldsymbol{\sigma}^{-1}\mathbf{Z}_i)^T$$

The eigenvalues are unchanged but the vectors are scaled. The p'th component in each vector $\mathbf{Z}_i$ is scaled by the variance of row p in $\mathbf{X}$. It follows that the entries in the new vectors $\boldsymbol{\sigma}^{-1}\mathbf{Z}_i$ are more uniform because they have a unit variance. The matrix element $|Z_{ip}|^2$ is the weight of phosphorylation event p in the i'th constraint. In the standardized variables, Z_{ip} is replaced by $\sigma_p^{-1}Z_{ip}$ so that the constraints are much more uniformly distributed over the phosphorylation events. Most constraints are associated with a zero eigenvalue and so are not informative. But even the informative constraints will be more uniform. In the unnormalized case the weight over the different phosphorylation events of the dominant constraint is shown in Fig. 3. The event, Bcr pY 644, with the by far highest weight also has the highest variance. By dividing by the variance, the dominant constraint will be much more uniformly distributed.

Lastly we turn to comment on the prediction of the potency of the oncogenes [4]. The role of the N oncogenes is described by the vectors $\mathbf{z}_i$ that are the eigenvectors of the N by N covariance matrix $\mathbf{X}^T\mathbf{X}$, see (9). In terms of the standardized data matrix $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathbf{X}^T\boldsymbol{\sigma}^{-2}\mathbf{X}$. The scaling by the variance will therefore alter how the constraints account for the potency of the oncogenes. Ultimately it is the superior performance of $\mathbf{X}$ over $\tilde{\mathbf{X}}$ in predicting potency that makes us prefer it as defining the nature of the variables to be used for the prior distribution.

References

1. Levine, R.D.: Molecular Reaction Dynamics. Cambridge University Press, Cambridge (2005)
2. Levine, R.D., Bernstein, R.B.: Energy disposal and energy consumption in elementary chemical-reactions—information theoretic approach. Acc. Chem. Res. **7**, 393–400 (1974)
3. Remacle, F., et al.: Information-theoretic analysis of phenotype changes in early stages of carcinogenesis. Proc. Natl. Acad. Sci. USA **107**(22), 10324–10329 (2010)
4. Graeber, T.G., et al.: Maximal entropy inference of oncogenicity from phosphorylation signaling. Proc. Natl. Acad. Sci. USA **107**(13), 6112–6117 (2010)
5. Levine, R.D.: Invariance and the distribution of maximal entropy. Kinam **3**, 403 (1981)
6. Levine, R.D.: Dynamical symmetries. J. Phys. Chem. **89**, 2122 (1985)
7. Levine, R.D.: Information theoretical approach to inversion problems. J. Phys. A **13**, 91–108 (1980)
8. Callen, H.B.: Thermodynamics and an Introduction to Thermostatics. Wiley, New York (1985)
9. Remacle, F., Levine, R.D.: The elimination of redundant constraints in surprisal analysis of unimolecular dissociation and other endothermic processes. J. Phys. Chem. A **113**(16), 4658–4664 (2009)
10. Mayer, J.E., Mayer, M.G.: Statistical Mechanics. Wiley, New York (1966)
11. Margolin, A.A., Califano, A.: Theory and limitations of genetic network inference from microarray data. Ann. N.Y. Acad. Sci. **1115**, 51–72 (2007)
12. Ziv, E., Nemenman, I., Wiggins, C.H.: Optimal signal processing in small stochastic biochemical networks. PLoS ONE **2**(10), e1077 (2007)
13. Banavar, J.R., Maritan, A., Volkov, I.: Applications of the principle of maximum entropy: from physics to ecology. J. Phys., Condens. Matter **22**(6) (2010)
14. Krawitz, P., Shmulevich, I.: Entropy of complex relevant components of Boolean networks. Phys. Rev. E **76** (2007)
15. Lezon, T.R., et al.: Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. Proc. Natl. Acad. Sci. USA **103**(50), 19033–19038 (2006)
16. Locasale, J.W., Wolf-Yadlin, A.: Maximum entropy reconstructions of dynamic signaling networks from quantitative proteomics data. PLoS ONE **4**(8) (2009)
17. Mora, T., et al.: Maximum entropy models for antibody diversity. Proc. Natl. Acad. Sci. USA **107**(12), 5405–5410 (2010)
18. Roudi, Y., Nirenberg, S., Latham, P.E.: Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't. PLoS Comput. Biol. **5**(5) (2009)
19. Theis, F.J., Bauer, C., Lang, E.W.: Comparison of maximum entropy and minimal mutual information in a nonlinear setting. Signal Process. **82**(7), 971–980 (2002)
20. Schneidman, E., et al.: Network information and connected correlations. Phys. Rev. Lett. **91**, 238701 (2003)
21. Tkacik, G., Calan, C.G., Jr., Bialek, W.: Information flow and optimization in transcriptional regulation. Proc. Natl. Acad. Sci. USA **105**, 12265–12270 (2008)
22. Skaggs, B.J., et al.: Phosphorylation of the ATP-binding loop directs oncogenicity of drug-resistant BCR-ABL mutants. Proc. Natl. Acad. Sci. USA **103**(51), 19466–19471 (2006)
23. Alhassid, Y., Levine, R.D.: Experimental and inherent uncertainties in the information theoretic approach. Chem. Phys. Lett. **73**(1), 16–20 (1980)
24. Kinsey, J.L., Levine, R.D.: Performance criterion for information theoretic data-analysis. Chem. Phys. Lett. **65**(3), 413–416 (1979)
25. Agmon, N., Alhassid, Y., Levine, R.D.: Algorithm for finding the distribution of maximal entropy. J. Comput. Phys. **30**(2), 250–258 (1979)
26. Janes, K.A., Lauffenburger, D.A.: A biological approach to computational models of proteomic networks. Curr. Opin. Chem. Biol. **10**(1), 73–80 (2006)
27. van den Berg, R.A., et al.: Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics **7**, 142 (2006)
28. Bar-Even, A., et al.: Noise in protein expression scales with natural protein abundance. Nat. Genet. **38**(6), 636–643 (2006)