A thermodynamic-based approach for the resolution and prediction of protein network structures

As Published: 10.1016/J.CHEMPHYS.2018.03.005

Publisher: Elsevier BV

Persistent URL: https://hdl.handle.net/1721.1/135794

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Chemical Physics xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Chemical Physics

journal homepage: www.elsevier.com/locate/chemphys

A thermodynamic-based approach for the resolution and prediction of protein network structures

Efrat Flashner-Abramson [a], Jonathan Abramson [b], Forest M. White [c], Nataly Kravchenko-Balasha [a,*]

[a] Department of Bio-Medical Research, Institute of Dental Sciences, Hebrew University of Jerusalem, Jerusalem 91120, Israel
[b] Intel, Jerusalem, Israel
[c] Department of Biological Engineering, MIT, Cambridge, MA 02139, United States

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Protein networks
Cell signaling
Cancer-altered signaling
Thermodynamic-based approach
Surprisal analysis
Information theory
Drug response prediction

ABSTRACT

The rapid accumulation of omics data from biological specimens has revolutionized the field of cancer research. The generation of computational techniques attempting to study these masses of data and extract the significant signals is at the forefront.

We suggest studying cancer from a thermodynamic-based point of view. We hypothesize that by modelling biological systems based on physico-chemical laws, highly complex systems can be reduced to a few parameters, and their behavior under varying conditions, including response to therapy, can be predicted.

Here we validate the predictive power of our thermodynamic-based approach, by uncovering the protein network structure that emerges in MCF10a human mammary cells upon exposure to epidermal growth factor (EGF), and anticipating the consequences of treating the cells with the Src family kinase inhibitor, dasatinib.

1. Introduction

The growing realization of the complexity of cancer systems, coupled with the increasing availability of omics data (genomics, transcriptomics, proteomics, etc.), have spurred the development of several computational techniques that aim to unfold the intricacy of cancer and find an underlying order that can be utilized in terms of anti-cancer therapy [1–6]. These techniques include Bayesian methods (based on elucidating the relationships between a few genes at a time [7]), reverse-engineering algorithms (based on chemical kinetic-like differential equations [8]), multivariate statistical methods which include clustering methods [9], principal component analysis [10], singular value decomposition [11], meta-analysis [12], information-theoretical approaches inferring statistical information about the network structure [13–15], and machine learning [16] (additional discussion can be found in references [17,18]). However, despite enormous progress in the fields of data analysis [17] and cancer research, aggressive tumors still respond poorly to the current therapies, suggesting that our understanding of patient-specific signaling networks is incomplete.

We study biological phenomena from a thermodynamic-based point of view, utilizing surprisal analysis [18–24]. We hypothesize that biological phenomena can be modelled based on physico-chemical laws, in order to discover the rules that govern the behavior of these complex systems. Thermodynamic-based approaches have been applied to the analysis of biological systems in a number of cases (for example, see references [25–28]). Understanding the set of rules that drive biological systems should allow prediction of the systems' behavior upon exposure to various environmental conditions, including drugs, for example. We assume that every biological system that is free of environmental and genomic constraints can reach a balanced state, which is minimal in free energy. Upon the application of constraints, the system deviates from its balanced state and reaches a new state, characterized by a higher free energy. We hypothesize that deciphering the complete set of constraints that operate on the system, and the altered molecular processes that have consequently emerged, will allow designing a rationalized method to guide the system back to its normal, balanced state.

We have tested our approach in a number of experimental settings. For example, we have demonstrated our ability to elucidate the molecular processes that govern the directed movement of glioblastoma (GBM) cells, and to intentionally and specifically interfere with their directed motion [23,24]. Based on the knowledge we gained, we predicted that aggressive GBM cells will tend to scatter to larger cell-cell distances, whereas less aggressive

* Corresponding author.
E-mail address: natalyk@ekmd.huji.ac.il (N. Kravchenko-Balasha).

https://doi.org/10.1016/j.chemphys.2018.03.005
0301-0104/

GBM cells will tend to form more compact patterns, consistent with observations of others *in vivo* [23,29].

Cancer is a biological system that has obviously deviated from its balanced state. Every tumor is driven by multiple oncogenic aberrations, and its survival and progression depend on a number of altered signaling pathways [30,31]. Surprisal analysis discovers the complete set of molecular processes gone awry in each tumor. Identifying these aberrations in a patient-specific manner, and simultaneously blocking the entire signaling flux through the altered pathways is key to effective personalized anti-cancer therapy that will destroy the tumor.

We have demonstrated the power of the approach in detecting patient-specific proteomic structures, comprising patient-specific sets of unbalanced molecular processes (Flashner-Abramson et al. submitted for publication and reference [32]).

Here we analyze a dataset obtained from MCF10a human mammary cells that have been stimulated with epidermal growth factor (EGF) in the presence or absence of the Src family kinase inhibitor, dasatinib [33]. We demonstrate the predictive power of surprisal analysis, by studying the protein network structure that emerged in the cells upon EGF stimulation, and then successfully foreseeing the response of the MCF10a protein network to dasatinib treatment.

2. Materials and Methods

2.1. Surprisal analysis

Surprisal analysis is a thermodynamic-based information-theoretic approach [19,20,34]. The analysis is based on the premise that biological systems reach a balanced state when the system is free of constraints [35–37]. However, when under the influence of environmental and genomic constraints, the system is prevented from reaching the state of minimal free energy, and instead reaches a state which is higher in free energy (in biological systems, which are normally under constant temperature and constant pressure, minimal free energy equals maximal entropy).

For example, if the system under study is a living cell, an environmental constraint can be exposure to a drug, which inflicts a change in protein concentrations and activities in the cell. The system can be influenced by genomic constraints as well, such as genomic mutations that in turn affect protein function, often requiring alteration of specific signaling pathways to oppose the functions of the damaged protein.

Surprisal analysis can take as input the expression levels of various macromolecules, e.g. genes, transcripts, or proteins. However, be it environmental or genomic alterations, it is the proteins that constitute the functional output in living systems, therefore we base our analysis on proteomic data. Since the varying forces, or constraints, that act upon living cells ultimately manifest as alterations in the cellular protein network, these constraints can also be viewed as unbalanced molecular processes that emerge in the system.

For every protein, i, surprisal analysis calculates its expected expression level when the system is balanced and free of constraints: X_i^0. This term was shown to be constant, i.e. is independent of time and of the actual state of the system [18,21,38,39]. In terms of information theory, X_i^0 represents the state of maximal entropy, or minimal information. $X_i(t)$ is the actual, experimentally measured expression level of the protein i at the time point t. In cases where $X_i(t) \neq X_i^0$ we assume that the expression level of the protein i was altered due to constraints that operate on the system. Surprisal analysis discovers the complete set of constraints operating on the system at any given time, t, by utilizing the following equation:

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

The term $\sum_{\alpha=1} G_{i\alpha}\lambda_\alpha(t)$ represents the sum of deviations in expression level of the protein i due to the various constraints, or unbalanced processes, that exist in the system at the time t. The processes are indexed $\alpha = 1, 2, 3, \ldots$, such that the significance of the process decreases with increasing index, i.e. unbalanced process 1 acts on the system longer (in more time points) than unbalanced processes 2, 3 etc.

The difference between the balanced state expression level, X_i^0, and the actual expression level, $X_i(t)$, represents the amount of information we have about protein i. A protein that is influenced by constraints, i.e. is influenced by one or more unbalanced processes and is therefore functionally linked to other proteins, cannot take any possible expression level. Rather, its expression level is affected by the expression levels of other proteins.

The term $G_{i\alpha}$ denotes the degree of participation of the protein i in the unbalanced process α, and its sign indicates the correlation or anti-correlation between proteins in the same process. For example, in a certain process α, proteins can be assigned the values: $G_{\text{protein1},\alpha} = -0.50$, $G_{\text{protein 2},\alpha} = 0.24$, and $G_{\text{protein 3},\alpha} = 0.00$, indicating that this process altered proteins 1 and 2 in opposite directions (i.e. protein 1 is upregulated and protein 2 is downregulated, or vice versa due to the process α), while not affecting protein 3. Note that each protein can take part in a number of unbalanced processes at once.

Importantly, not all processes are active all the time. The term $\lambda\alpha(t)$ represents the importance of the unbalanced process α at time point t. Its sign indicates the correlation or anti-correlation between the same processes in different time points. For example, if the process α is assigned the values: $\lambda_\alpha(0\text{ s}) = 3.1$, $\lambda_\alpha(20\text{ s}) = 0.0$, and $\lambda_\alpha(80\text{ s}) = 2.5$, it means that this process influences the sample in the same direction at t = 0 s and t = 80 s, while it is inactive at t = 20 s.

The partial deviations in expression level of the protein i due to the different constraints sum up to the total change in expression level (relative to the balance state level), $\sum_{\alpha=1} G_{i\alpha}\lambda_\alpha(t)$.

Mathematically, the algorithm is based on the construction of a covariance matrix of the logarithm of protein expression levels. SVD (singular value decomposition) [11] and the method of Lagrange undetermined multipliers are utilized to calculate the maximal entropy and the various constraints that exist in the system under study, i.e. to determine X_i^o, $\lambda_\alpha(t)$ and $G_{i\alpha}$ values. The $\lambda_\alpha(t)$ values represent the Lagrange multipliers. We determine the minimal number of unbalanced processes needed to accurately reconstruct the experimental protein expression levels, as described in the next section. For more details regarding the mathematical analysis see references [21] and [24].

As explained above, the zeroth term $\ln X_i^o(t)$, is the logarithm of the expression level of protein i at the balanced state of the system. This term is utilized as a reference against which the deviation terms are identified. In the current analysis, the dataset contained only measurements of phosphorylated proteins, and therefore the reference state includes an imbalance that reflects a change in general phosphorylation resulting from EGF stimulation, with an importance, $\lambda_0(t)$, that gradually increases with time (see Supp. Tables 2 and 3). See reference [32] for more details regarding this general phosphorylation event.

2.2. Determination of the number of significant unbalanced processes

The analyses of the training and test datasets provided 9×9 matrices of $\lambda_\alpha(t)$ values, such that every row in the matrices contained 9 values of $\lambda_\alpha(t)$ for 9 time points, and each row corresponded to an unbalanced process (see Supp. Tables 2 and 3).

However, not all unbalanced processes are significant. Our goal is to determine how many unbalanced processes are needed in order to reconstruct the experimental data, i.e. for which value of n: $\ln X_i(t) \cong \ln X_i^o(t) - \sum_{\alpha=1}^{n} G_{i\alpha}\lambda_\alpha(t)$. To find n, we performed the following two steps:

(1) *Processes with significant amplitudes were selected:* Three separate analyses were conducted for every dataset: (1) when the input was the experimental data; (2) when the input was the data plus standard deviations; (3) when the input was the data minus standard deviations. The $\lambda_\alpha(t)$ values presented in Supplementary Tables 2 and 3 are the average values obtained from the 3 analyses, with the corresponding standard deviations. An unbalanced process was considered significant only if it was assigned significant $\lambda_\alpha(t)$ values, i.e. if for at least one of the time points its value exceeds the noise threshold (the standard deviations).
Analysis of the training dataset revealed that from $\alpha = 4$, the importance values, $\lambda_\alpha(t)$, become insignificant (i.e. do not exceed the noise threshold), suggesting that 3 unbalanced processes are enough to describe the system. Analysis of the test dataset revealed that from $\alpha = 3$, the importance values, $\lambda_\alpha(t)$, become insignificant, suggesting that 2 unbalanced processes are enough to describe the system.

(2) *Reproduction of the experimental data by the unbalanced processes was verified:* To verify that the numbers of processes identified in step (1) are correct, we plotted $\ln X_i^o(t) - \sum_{\alpha=1}^{n} G_{i\alpha}\lambda_\alpha(t)$ against $\ln X_i(t)$ for different values of n, and examined the correlation between them as n was increased. An unbalanced process, $\alpha = n$, was considered significant if it improved the correlation significantly relative to $\alpha = n - 1$. In any case, a perfect correlation (with a correlation coefficient R = 1) was not expected due to random noise in the biological system.
To verify that 3 unbalanced processes are enough to describe the training dataset, we plotted $\ln X_i^o(t) - \sum_{\alpha=1}^{n} G_{i\alpha}\lambda_\alpha(t)$ for different values of α and compared it to the experimental expression levels, $\ln X_i(t)$. As shown in Supplementary Fig. 1, the quality of the correlation increased until $\alpha = 3$ (reaching R = 0.99). Addition of another unbalanced process, i.e. $\alpha = 4$, did not improve the correlation, strengthening our assumption that following $\alpha = 3$, the rest of the processes represent noise in the biological system.
A similar process was carried out for test dataset, in order to verify that 2 unbalanced processes are enough to describe this system. Supplementary Fig. 2 shows that the quality of the correlation increased until $\alpha = 2$ (reaching R = 0.99), and that addition of a third process did not improve the correlation.

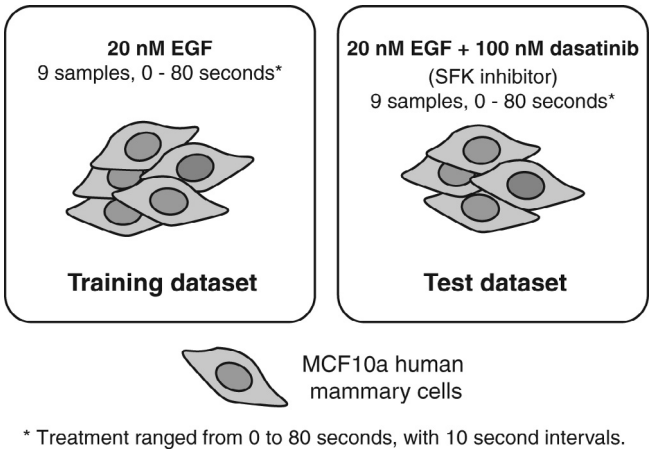


Fig. 1. *The working datasets.* The dataset, obtained in the laboratory of Forest White, was divided into a training dataset (containing proteomic data obtained from MCF10a cells that were stimulated with 20 nM EGF for 0–80 s), and a test dataset (containing the data obtained from MCF10a cells following treatment with 20 nM EGF and 100 nM dasatinib (SFK inhibitor) for 0–80 s). Both datasets contained the same list of 88 phosphoproteins. These datasets will be analyzed independently.

2.3. Determination of proteins with significant weights in every unbalanced process

The analyses of the training and test datasets provided 9×88 matrices of $G_{i\alpha}$ values, such that every row in the matrices contained 9 values of $G_{i\alpha}$ for 9 unbalanced process (not all significant, see previous section), and each row corresponded to a protein (see Supp. Tables 2 and 3). For every unbalanced process considered

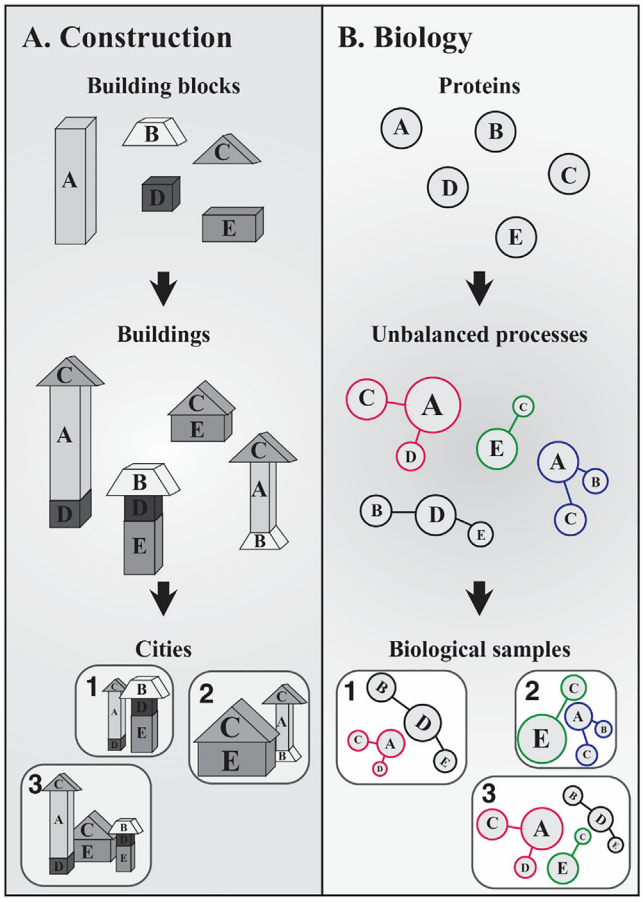


Fig. 2. *An illustrative depiction of our approach.* (A) In construction, a collection of building blocks can be used to assemble different types of buildings. Specific subsets of buildings can exist in different cities. (B) Similarly, in biology, unbalanced molecular processes (buildings) are made up of a collection of proteins (blocks). Each biological sample (city) can contain a unique subset of unbalanced processes that maintain its current state. A certain protein can participate in different unbalanced processes simultaneously. In the figure, the proteins are represented by circles, such that the diameter of the circle denotes its relative weight in the process (analogous to size of blocks in each building), e.g. protein A is important in the red process, less important in the blue process, and does not participate in the green and black processes. The size of the entire process in each sample (or size of building in each city) denotes its sample-specific importance. For example, the red unbalanced process is relatively important in sample 3, less important in sample 1, and insignificant in sample 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

significant, a threshold value was determined for $G_{i\alpha}$, to filter out proteins that were assigned a value of $G_{i\alpha}$ that is insignificant and results from noise in the system. The process is described in detail in reference [21].

2.4. Generation of functional networks

The functional networks presented in Figs. 5 and 6 were generated using a python script. The goal was to generate a functional network according to STRING database, where proteins with negative G values are marked blue and proteins with positive G values are marked red, in order to easily identify the correlations and anti-correlations between the proteins in each network. The script takes as an input the names of the proteins in the network and their G values, obtains the functional connections from STRING database (string-db.org), and then plots the functional network.

2.5. Generation of barycentric plots

Proteins with significant positive G values were selected, and their G values for unbalanced processes 1^-, 2^- and 3^- were normalized such that for every protein, i, $G_{i1^-} + G_{i2^-} + G_{i3^-} = 1$. These values were then projected onto a triangle by calculating the dot product of the G vectors and the following array: [[0,0],[1,0],[cos(60),sin(60)]]. The size of each circle was set in proportion to the sum $G_{i1^-} + G_{i2^-} + G_{i3^-}$ before normalization to 1, thus representing the relative weight of the protein i in the three unbalanced processes.

3. Results

3.1. Obtaining the working datasets

We utilized surprisal analysis to study the protein network structure that emerged in MCF10a human mammary cells in response to epidermal growth factor (EGF) stimulation, and to predict the effect of the Src family kinase (SFK) inhibitor, dasatinib, on this network structure [33]. The dataset, which was obtained in the laboratory of Forest White [33], contained the mass-spectrometry-based proteomic measurements of the expression levels of 88 tyrosine-phosphorylated proteins found in serum-starved MCF10a cells following stimulation with 20 nM EGF for 9 different periods of time (ranging from 0 to 80 s with 10 s intervals), in the presence or absence of 100 nM dasatinib (a SFK inhibitor) (Fig. 1 and Supp. Table 1).

We divided the dataset into 2 parts: (1) A training dataset, containing the data obtained from EGF-stimulated cells, in the absence of dasatinib; (2) A test dataset, containing the data obtained from cells that were stimulated with EGF in the presence of dasatinib (Fig. 1).

3.2. A brief overview of the theoretical approach

A detailed explanation of the approach can be found in the Materials and Methods section. Here we will provide a brief description of the method, beginning with an illustrative explanation (Fig. 2).

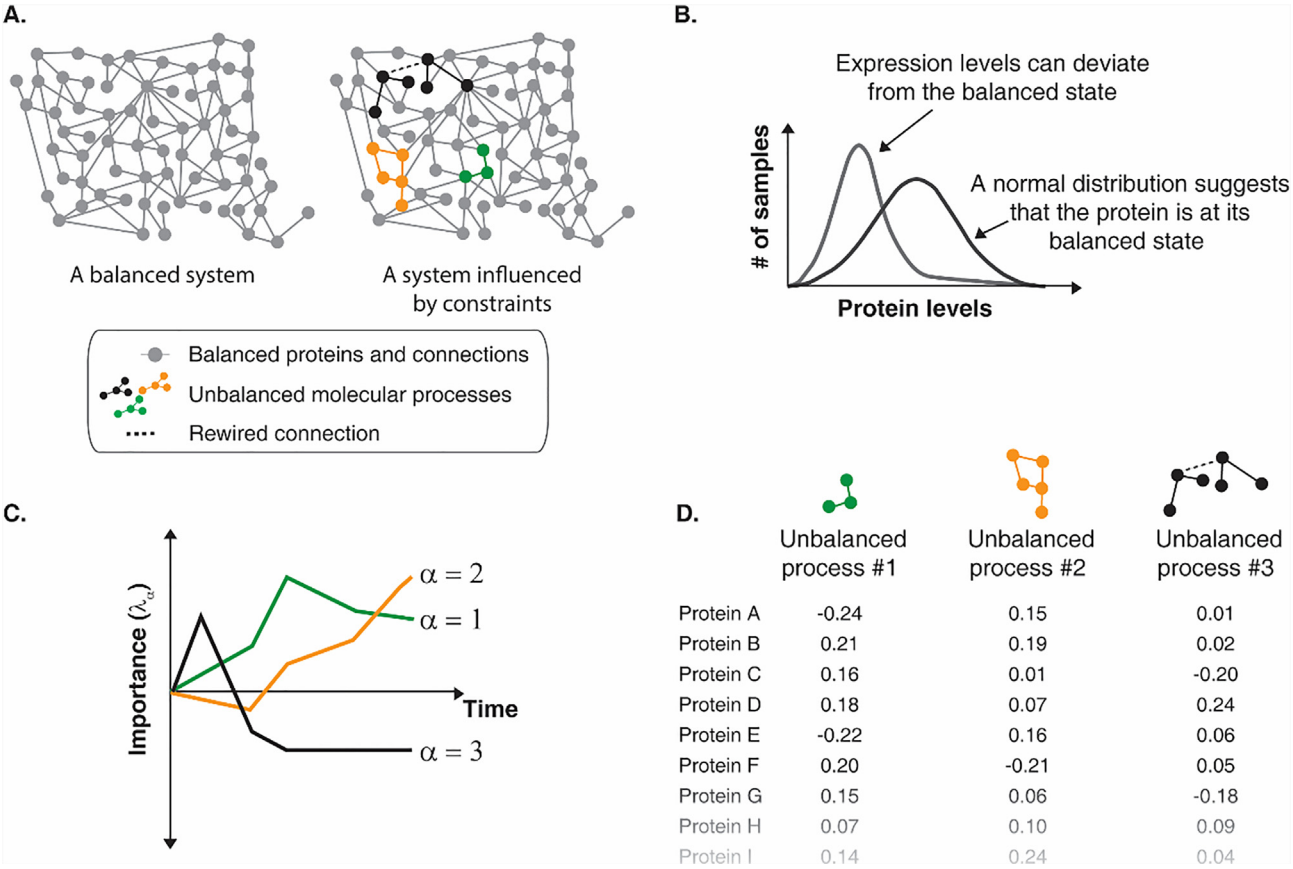


Fig. 3. *The thermodynamic principle at the base of surprisal analysis.* (A) We base our approach on the premise that every biological system can reach a balanced, steady state, which is minimal in free energy, when it is free of constraints. The application of environmental or genomic constraints on the system deviates it from its balanced state, and drives the emergence of unbalanced processes in the system. (B) Surprisal analysis discovers for every protein whether it is at its balanced state level, or rather has it deviated from its balanced level due to the influence of constraints, or unbalanced molecular processes. (C) The analysis provides the time-dependent importance ($\lambda_\alpha(t)$) of every unbalanced process in the system. (D) The relative weights of the proteins in every unbalanced process ($G_{i\alpha}$) are discovered by the analysis. (The reader is referred to the web version of this article to view the colored version of the figure.)

An overview of the thermodynamic principles on which we base the algorithm will follow (Fig. 3 below).

In our approach, biological samples are analogous of construction sites (Fig. 2). The algorithm makes use of surprisal analysis [21], a theoretical approach originally developed for use in chemistry and physics by Levine et al. [19,20]. We have previously demonstrated the application of surprisal analysis to biological systems [18,21–24]. Surprisal analysis takes as input the experimental expression levels of proteins (can also be genes/transcripts) in different samples, and identifies the basic building blocks that are needed to reconstruct the entire dataset. Consider a city containing different types of buildings (Fig. 2A). A finite number of types of building blocks is enough to rebuild the buildings in this city. In biological terms, our building blocks are the proteins, and these proteins assemble the processes (buildings) that exist in a specific sample (city) (Fig. 2B). Different sizes of a specific building block can be used in construction (Fig. 2A, middle). Similarly, in a specific biological process, certain proteins play key roles, while other proteins contribute only modestly to the process. Accordingly, surprisal analysis provides the relative weights of the different proteins in every process (denoted by the sizes of the circles in Fig. 2B, middle). A certain protein (block) can participate in different unbalanced processes (buildings) simultaneously (Fig. 2A and B, middle panels). This is an important attribute of surprisal analysis, enabling to address the complexity of biological protein networks, which frequently demonstrate non-linearity and rewired protein connections [31,40]. The collection of buildings that can be built using the existing building blocks enables the construction of different cities, each containing a specific subset of buildings (Fig. 2A, bottom). Correspondingly, different biological samples can harbor different subsets of unbalanced processes out of the collection of processes identified by the analysis (Fig. 2B, bottom). While two cities can both harbor the same types of buildings, the sizes of the buildings may differ (Fig. 2A, bottom). Similarly, a specific unbalanced molecular process can exist in different samples, but its importance may vary, e.g. it can be very significant in one sample, while it can be secondary in significance or insignificant in another sample (Fig. 2B, bottom).

The thermodynamic principle that underlies the approach is that every biological system can reach a steady, balanced state, and that it can deviate from this balanced state upon the application of environmental or biological constraints (Fig. 3A). A

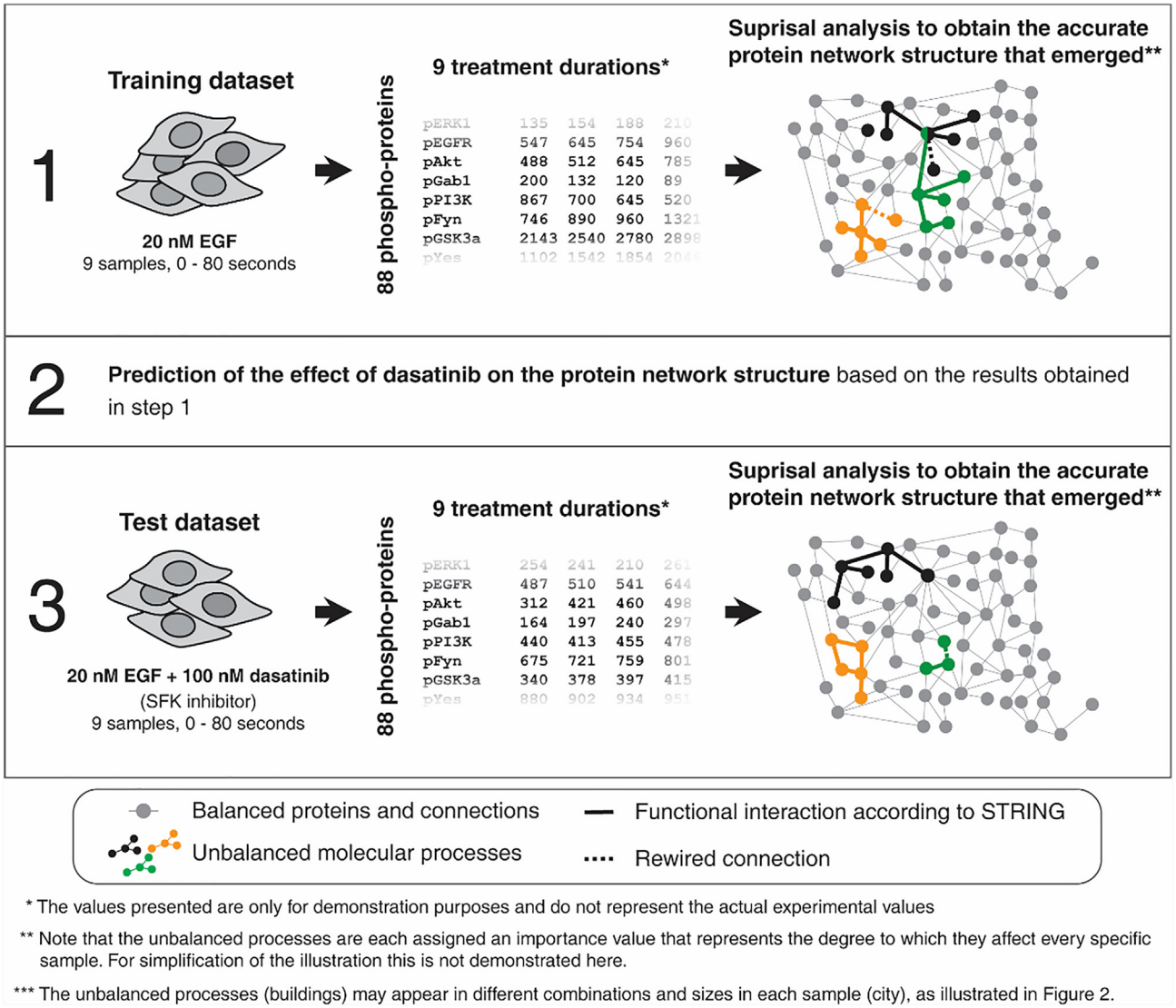


Fig. 4. *A schematic workflow.* We will begin by analyzing the training dataset, aiming to decipher the accurate structure of the protein network in MCF10a cells upon EGF stimulation (step 1). Based on the results of the analysis, we will predict the effect of addition of dasatinib, a SFK inhibitor, to the system (step 2). Finally, we will analyze the test dataset (step 3), and verify whether the structure of the protein network that emerged in the cells agrees with our prediction. (The reader is referred to the web version of this article to view the colored version of the figure.)

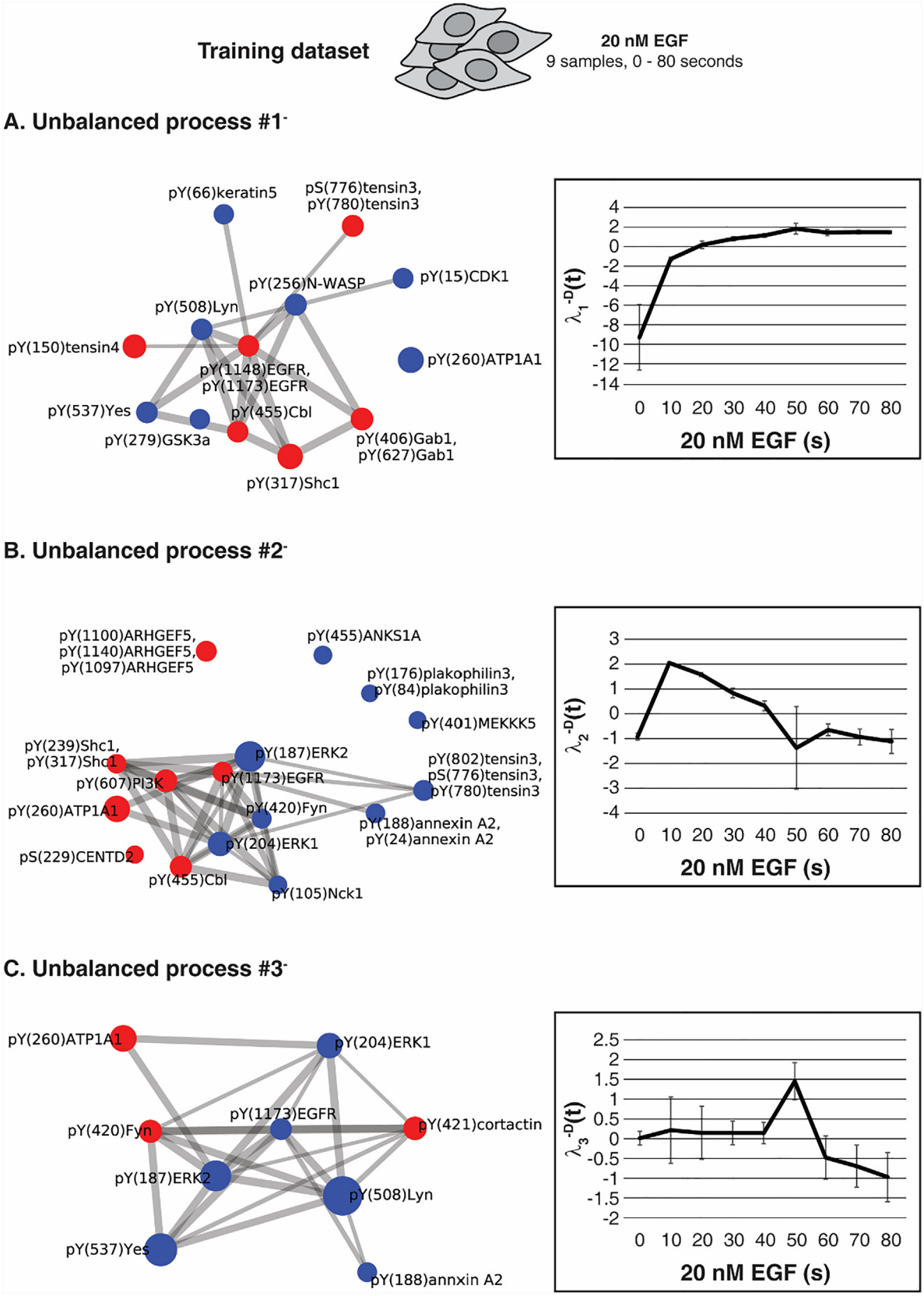


Fig. 5. *3 unbalanced processes emerged in MCF10a cells upon EGF stimulation.* Analysis of the training dataset, containing the data obtained from MCF10a cells that were stimulated with 20 nM EGF for 0–80 s, revealed that 3 unbalanced processes emerged in the system. The processes are indexed according to their general significance, such that unbalanced process 1^- is the most significant and unbalanced process 3^- is the least significant. The superscript $(-)$ indicates the absence of dasatinib. For each process, the proteins that participate significantly in the process (see Material and Methods) were assembled into networks with functional connections between the proteins according to String database (left panels). Red proteins are upregulated by the process and blue proteins are downregulated by the process (assuming $\lambda_\alpha(t) > 0$). The time-dependent importance of every process is presented in the right panels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

deviation from the balanced state can manifest as a change in protein concentrations as well a change in the protein network structure [18] (Fig. 3A and B). Surprisal analysis identifies which proteins are at their balanced levels, and which proteins have deviated from their balanced levels. Mathematically, this is done by utilizing the following equation for every protein i: $\ln X_i(t) = \ln X_i^0 - \sum_{\alpha=1}^{n} G_{i\alpha}\lambda_\alpha(t)$ where $X_i(t)$ is the experimental expression level of the protein i at time t, X_i^0 is its expression level at the balanced state, and $\sum_{\alpha=1}^{n} G_{i\alpha}\lambda_\alpha(t)$ is the sum of the deviations in expression level from the balanced state level, due to the various constraints, α, that operate on the system at time t. Each constraint represents an unbalanced process that emerged in the system. The unbalanced processes represent groups of proteins that exhibit deviations (or partial deviations) from the balanced state in a correlated manner. Every constraint, or unbalanced process, is assigned an importance, $\lambda_\alpha(t)$, which can change with time in a specific sample, or vary in different given samples (Fig. 3C; the importance of the process is represented by the size of the process/building in Fig. 2). The unbalanced processes are indexed such that their importance decreases with increasing index. For every unbalanced process, α, each protein i is assigned a weight, $G_{i\alpha}$, which denotes the degree to which it participates in the process (Fig. 3D; see size of protein/block in Fig. 2).

3.3. Three unbalanced processes emerged in MCF10a cells following stimulation with 20 nM EGF

Our overarching goal in this study was to demonstrate the ability of our approach to predict the response of EGF-stimulated MCF10a cells to the SFK inhibitor, dasatinib (Fig. 4). To make such a prediction, we first set out to unravel the protein network structure that emerged in the cells in response to EGF stimulation. This was achieved by applying surprisal analysis to the training dataset (Fig. 4, step 1).

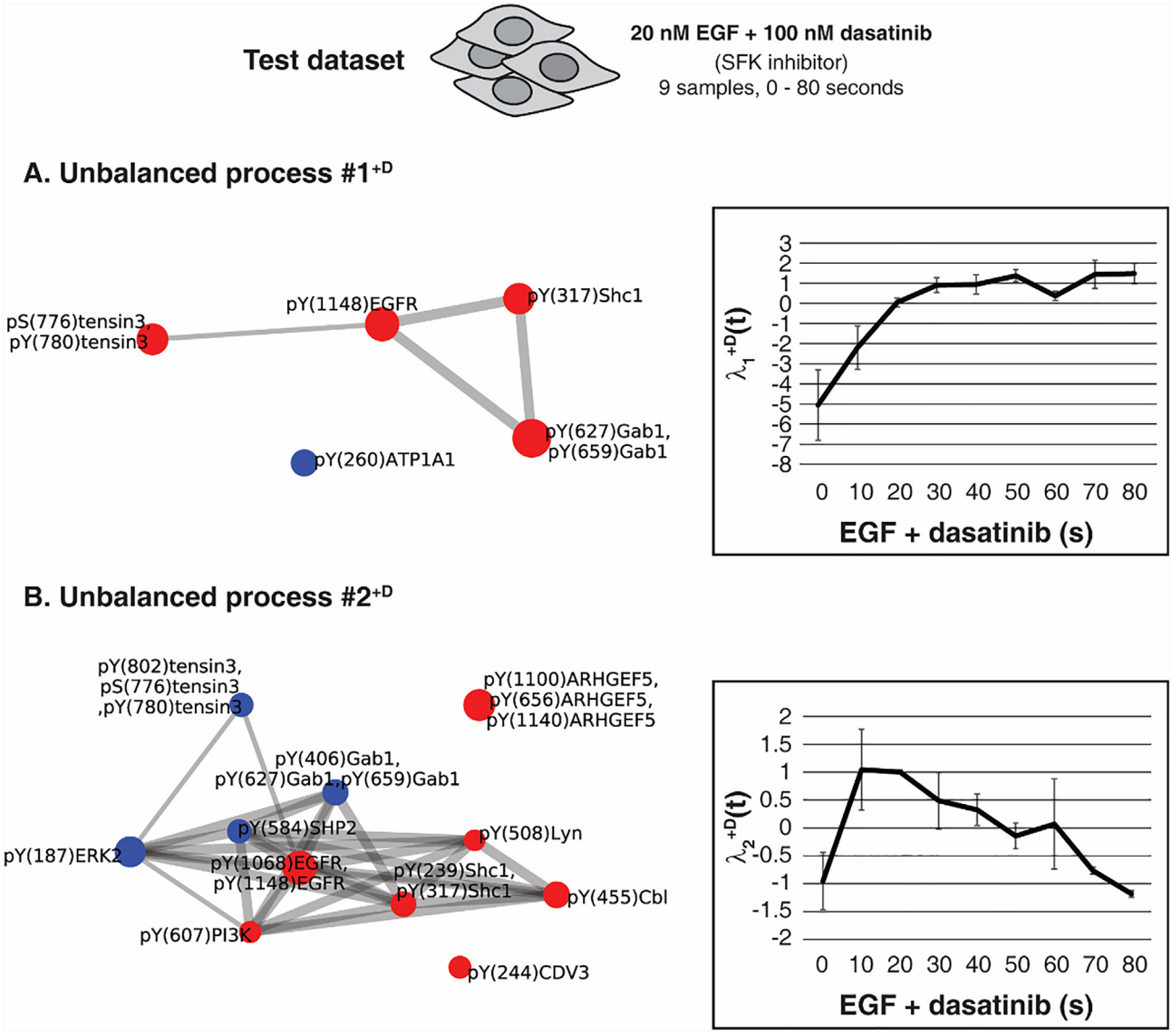


Fig. 6. *Inhibition of SFK proteins induced 2 unbalanced processes in MCF10a cells.* Analysis of the test dataset, containing the data obtained from MCF10a cells that were stimulated with 20 nM EGF for 0–80 s in the presence of dasatinib, revealed that 2 unbalanced processes emerged in the system. +D indicates the presence of dasatinib. For each process, the proteins that participate significantly in the process (see Material and Methods) were assembled into networks with functional connections between the proteins according to String database (left panels). Red proteins are upregulated by the process and blue proteins are downregulated by the process (assuming $\lambda_\alpha(t) > 0$). The time-dependent importance of every process is presented in the right panels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The analysis of the training dataset revealed that following EGF stimulation, the unbalanced protein network in MCF10a cells consisted of 3 distinct unbalanced processes (Supp. Table 2 and Fig. 5). For clarification, these processes will be indexed 1^-, 2^- and 3^-, denoting that these processes emerged in the system before the addition of dasatinib. These 3 unbalanced processes are enough to reproduce the experimental training data (Supp. Fig. 1 and Material and Methods). The proteins marked red are proteins that were assigned a positive $G_{i\alpha}$ value (i.e. are upregulated by the process when $\lambda_\alpha(t) > 0$ and downregulated by the process when $\lambda_\alpha(t) < 0$), and the proteins marked blue are proteins that were assigned a negative $G_{i\alpha}$ value (i.e. are downregulated by the process when $\lambda_\alpha(t) > 0$ and upregulated by the process when $\lambda_\alpha(t) < 0$) (colors are indicated in the online version). The connections between the proteins denote known functional interactions according to String database [41]. The importance of each of the 3 unbalanced processes changes over time (Fig. 5A–C, right). For example, the importance of unbalanced process 3^- is negligible until t = 40 s, and then peaks after 50 s of stimulation with 20 nM EGF (Fig. 5C, right). Therefore, in each time point, the system is characterized by a specific subset of these 3 unbalanced processes (see Figs. 2 and 3).

3.4. Unbalanced processes 1 and 3 are predicted to be most affected by addition of the SFK inhibitor, dasatinib

Our next step was to predict the effect of addition of dasatinib to MCF10a cells, based on the analysis of the training dataset (Fig. 4, step 2). Inspection of the unbalanced subnetworks that emerged in MCF10a cells upon EGF stimulation reveals that Src family proteins – Fyn, Yes and Lyn – are activated in processes 1^- and 3^-, and not in process 2^-. Note that the tyrosine phosphorylation sites pY(5 0 8)Lyn and pY(5 3 7)Yes are inhibitory phosphorylation sites, and therefore their downregulation indicates an increase in protein activity (see unbalanced processes 1^- and 3^- in Fig. 5A and C, respectively). In contrast, pY(4 2 0)Fyn is an activating phosphorylation site, and therefore its downregulation in process 2^- indicates a decrease in protein activity (Fig. 5B). Hence, we predict that addition of the SFK inhibitor, dasatinib, to the system will not completely reduce the imbalance that was created in the system upon EGF stimulation, but rather will mainly affect the proteins involved processes 1^- and 3^-.

3.5. Two unbalanced processes emerged in MCF10a cells that were treated with EGF in the presence of the SFK inhibitor, dasatinib

To test our prediction, we conducted an additional computational analysis, independent of the previous analysis. This time we analyzed the test dataset, i.e. only the measurements obtained when the cells were exposed to 20 nM EGF and 100 nM dasatinib together (Fig. 4, step 3).

The analysis revealed that only 2 unbalanced processes emerged in the system when dasatinib was present (Fig. 6, Supp. Table 2 and Supp. Fig. 2). These processes will be indexed 1^{+D} and 2^{+D}, denoting that these processes emerged in the system following the addition of dasatinib.

The first important question that arises is whether the processes that emerged upon addition of dasatinib are the processes that emerged before its addition, e.g., is unbalanced process 1^{+D} the same as unbalanced process 1^-? Or rather are these new processes that were induced as a result of the actions of dasatinib in the system? To answer this, we examined the weights of the 88 phospho-proteins in each of the unbalanced processes, reasoning that if two processes are the same process, the relative weights of the proteins should be highly correlative with each other. In other words, the relative weights of the proteins in a specific process can be viewed as the structure of the process: some proteins are key proteins in this process and will therefore be assigned a significant $G_{i\alpha}$ value (significantly higher or lower than 0), and some do not participate significantly and will be thus assigned an insignificant $G_{i\alpha}$ value. Two processes possessing a highly similar protein structure are considered to be the same process. We note that we do not expect a perfect correlation due to possible protein-protein rewiring in the system in the presence of dasatinib, as well as random fluctuations in the system. We found that the relative weights of the proteins in unbalanced process 1^- highly correlated with the weights of the proteins in unbalanced process 1^{+D} (Fig. 7A). Similarly, the relative weights of the proteins in unbalanced process 2^- highly correlated with the weights of the proteins in unbalanced process 2^{+D} (Fig. 7B). In addition, the time-dependent importance of processes 1^{+D} and 2^{+D} (Fig. 6A and B, right panels) markedly resembles the time-dependent variation of process 1^- and 2^-, respectively (Fig. 5A and B, right panels). In contrast, the weights of the proteins in unbalanced process 3^- did not correlate with any of the unbalanced processes that emerged following addition of dasatinib (Supp. Fig. 3). This was expected, considering that unbalanced processes 1^{+D} and 2^{+D} sufficed to reproduce the experimental data (Supp. Fig. 2). The correlations between all pairs of processes can be found in Supplementary Fig. 3. From these results, we deduce that unbalanced processes 1^- and 2^- persisted in the system, while unbalanced process 3^- was completely eliminated by dasatinib treatment. It is important to recognize that the fact that processes 1^- and 2^- persisted in the system does not necessarily indicate the degree to which these processes were affected by dasatinib. For example, the degree of participation of specific proteins within

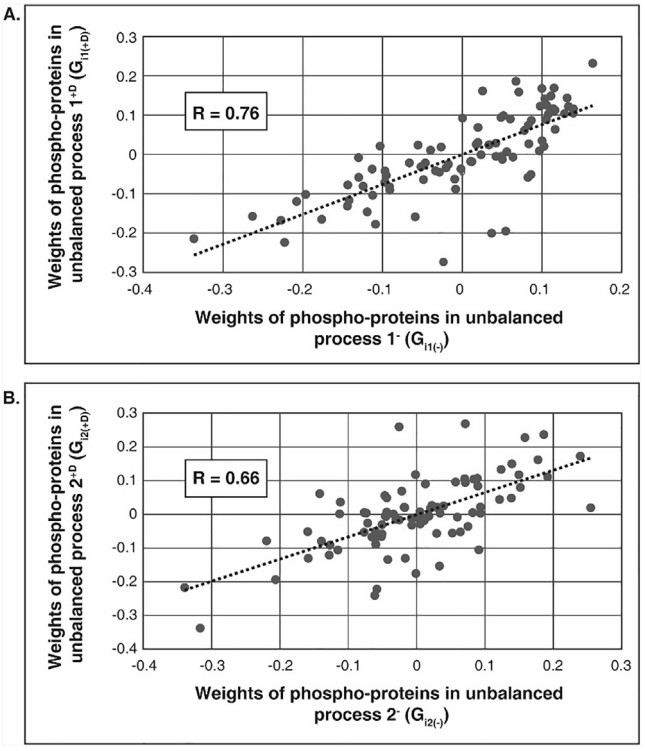
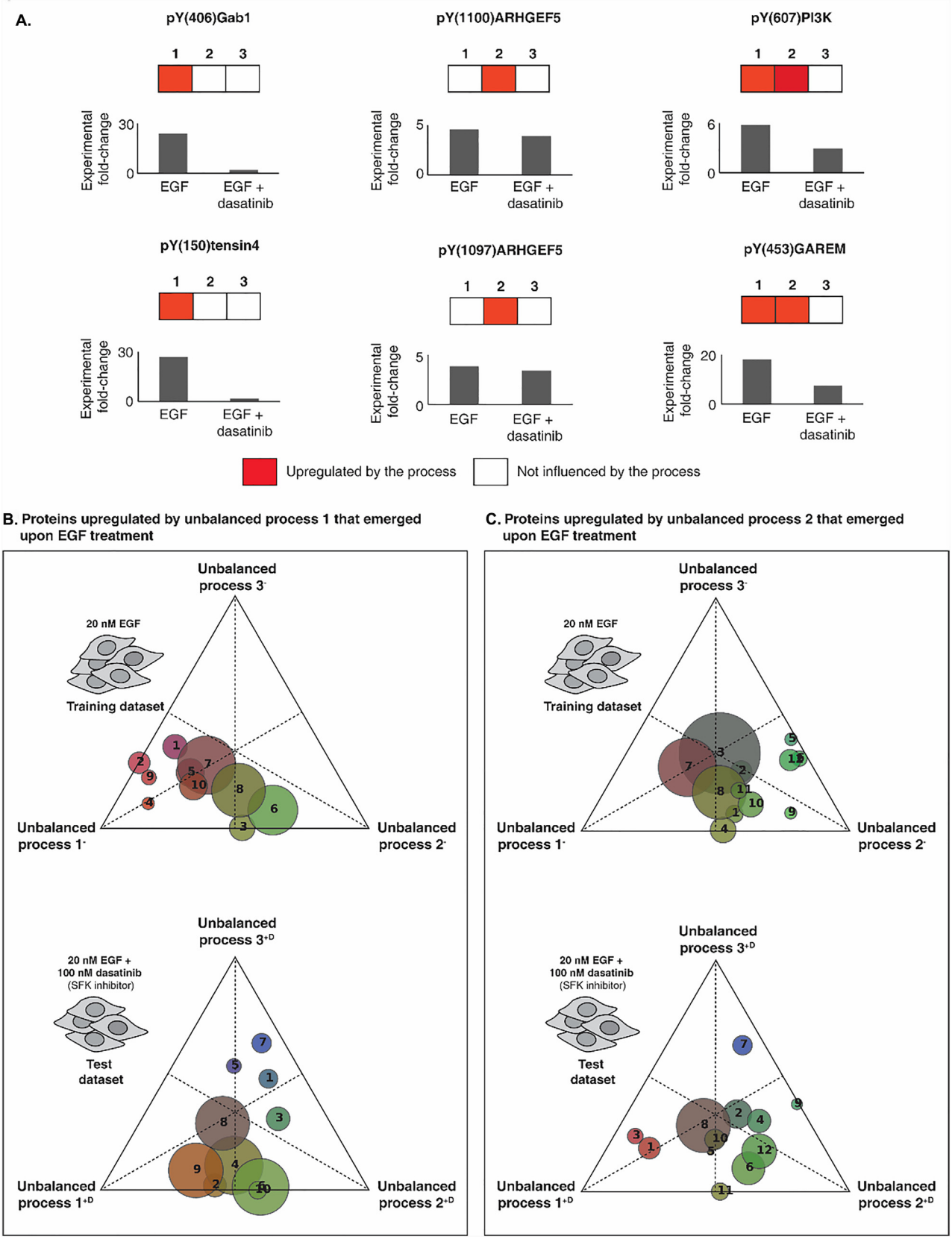


Fig. 7. *The same unbalanced processes 1 and 2 were independently identified in MCF10a before and after addition of dasatinib.* The relative weights of the proteins in the unbalanced processes that emerged in the system before and after addition of dasatinib were compared, attempting to find out whether they are the same processes that persisted in the system. (A) The relative weights of process 1^{+D} highly correlated with the weights of process 1^-. (B) The weights of process 2^{+D} highly correlated with the weights of process 2^-. See Supplementary Fig. 3 for the reciprocal correlation plots.

the process may be affected, as well as the general importance of the process in the system. To infer whether our prediction regarding the effect of dasatinib on EGF-stimulated MCF10a cells was correct, the protein network structure that emerged in the test dataset will be explored in depth below.

3.6. Comparison of the unbalanced processes that emerged before and after addition of dasatinib to the cells

By inspecting the unbalanced processes as they appear in Fig. 5 (EGF alone) and Fig. 6 (in the presence of dasatinib), it is difficult to



assess whether dasatinib indeed affected unbalanced process 1^- more than it affected process 2^-, as we predicted. This is because some proteins participate in a number of processes at once, and therefore, in biological terms, the unbalanced processes are not entirely independent of each other. For example, our analysis determined that pY(4 0 6)Gab1 and pY(1 5 0)tensin4 both participated only in unbalanced process 1^- (Fig. 8A, left). According to our prediction, since SFK proteins are upregulated by process 1^- (Fig. 5A), pY(4 0 6)Gab1 and pY(1 5 0)tensin4 should be significantly affected by dasatinib treatment. Inspection of the experimental expression levels of these proteins before and after addition of dasatinib revealed that, indeed, they were both significantly inhibited upon the addition of dasatinib (Fig. 8A, left). pY(1100)ARHGEF5 and pY(1097)ARHGEF5 participate only in unbalanced process 2^-, and according to our prediction should not be affected by dasatinib. Correspondingly, their expression levels were not significantly changed upon the addition of dasatinib (Fig. 8A, middle). In the above examples, the prediction regarding the effect of dasatinib treatment is relatively straight forward. In contrast, in the case of proteins that participate in a number of unbalanced processes the prediction becomes more complex. For example, pY(6 0 7)PI3K and pY(4 5 3)GAREM were assigned by our analysis to both processes, 1^- and 2^-. Their expression levels demonstrate a limited response to dasatinib treatment (Fig. 8A, right).

Therefore, to analyze the response of the protein network to dasatinib in a more rigorous manner, we took the weights of the proteins that were upregulated upon EGF stimulation by unbalanced processes 1^- and 2^-, and projected them onto triangular barycentric coordinates (Fig. 8B and C). Each circle represents a protein. The weights of the proteins were normalized such that for each protein its weights in unbalanced processes 1^-, 2^- and 3^- sum up to 1. The location of each protein demonstrates its relative participation in the 3 unbalanced processes. For example, protein 2 in Fig. 8B, top panel, participates the most in unbalanced process 1^-, to a lesser extent in unbalanced process 3^-, and not at all in unbalanced process 2^-, while protein 3 in the top panel of Fig. 8C participates equally in all 3 processes – 1^{+D}, 2^{+D}, and 3^{+D}. For additional clarity, each circle was colored according to its location (colors are indicated in the online version). The closer to process 1, the more red color; the closer to process 2, the more green color; the closer to process 3, the more blue color. The size of the circle indicates the general importance of the protein in all 3 processes (i.e. according to the actual weights before normalization). See Materials and Methods for a detailed explanation of the construction of the barycentric plots.

Fig. 8B shows the proteins that upon EGF stimulation were upregulated by unbalanced process 1^- (as shown in Fig. 5A). As explained above, we predicted that these proteins should be highly affected by SFK inhibition, due to the participation of SFK proteins in this process. The top panel presents the distribution of these proteins before the addition of dasatinib, according to their weights as obtained from analysis of the training dataset (G_{i1}^{-}). The bottom panel shows the distribution of these proteins following the addition of dasatinib, according to their weights as obtained from analysis of the test dataset (G_{i1}^{+D}). A general shift of these proteins *away* from unbalanced process 1 is clearly shown, as indicated by the disappearance of the red color.

Similar barycentric plots were generated for the proteins that were upregulated upon EGF stimulation by unbalanced process 2^- (Fig. 8C; as shown in Fig. 5B). According to our analysis of the training dataset, these proteins should not be significantly affected by dasatinib, because SFK proteins were not upregulated by this process. Fig. 8C demonstrates that, indeed, the distribution of these proteins (especially the green proteins associated significantly with process 2^-) remains similar, regardless of the presence of dasatinib in the system.

4. Discussion and conclusions

The accurate resolution of protein networks in biological samples is crucial for devising effective personalized drug combinations [1,42]. In terms of diagnosis and treatment, it is essential to identify not only the key oncogenic proteins upregulated in a specific tumor sample, but rather to resolve the complete protein network, as well as the structure of the network. For example, a number of oncoproteins may be upregulated in a particular tumor, but participate in *distinct* unbalanced processes (as shown in this study). Inhibition of the protein targets from one process will not necessarily lead to the inhibition of the proteins from another process. Targeting of the entire unbalanced signaling flux, by means of directing drugs to central proteins in each unbalanced process, is essential for the treatment to be effective. Therefore, an in depth understanding of the protein network structure, including the division into distinct subnetworks, or unbalanced processes, should allow for improved design of combination therapy.

In this work we demonstrated the power of the resolution of the complete protein network structure, by analyzing the protein network that develops in MCF10a human mammary cells following exposure to EGF for different periods of time, and using this knowledge to foresee the response of the system to inhibition of Src family proteins by dasatinib. Based on our analysis, we anticipated that treatment with dasatinib alone will not collapse the entire unbalanced signaling network in MCF10a cells, but rather will target 2 out of the 3 distinct unbalanced processes that were induced by EGF stimulation of the cells. Indeed, analysis of the protein expression levels in MCF10a cells that were exposed to EGF in the presence of dasatinib, revealed that unbalanced process 1^- was significantly harmed by dasatinib, and unbalanced process 3^-

Fig. 8. A general shift away from unbalanced process 1 is evident upon inhibition of SFK proteins, while not from unbalanced process 2. (A) 6 representative proteins were selected to demonstrate the independent influences of unbalanced process 1^- and 2^-. pY(4 0 6)Gab1 and pY(1 5 0)tensin4, which are influenced only by unbalanced process 1^-, were significantly downregulated upon addition of dasatinib. In contrast, pY(1100)ARHGEF5 and pY(1097)ARHGEF5, which are influenced only by unbalanced process 2^-, demonstrated no significant response to dasatinib. pY(6 0 7)PI3K and pY(4 5 3)GAREM, which participate in both processes, exhibited a partial response to dasatinib addition. (B, C) The weights of the proteins upregulated by processes 1^- and 2^- were normalized such that for every protein $G_{i1^-} + G_{i2^-} + G_{i3^-} = 1$ (top panels) and $G_{i1^{+D}} + G_{i2^{+D}} + G_{i3^{+D}} = 1$ (bottom panels), and plotted on triangular barycentric graphs. Each circle represents a protein, and its location and color denote its distribution among the three processes – bottom left corner and red color denote high participation in process 1^-; bottom right corner and green color denote high participation in process 2^-; top corner and blue color denote high participation in process 3^-. The plots show that the proteins upregulated by process 1^- (B, top panel), shifted away from process 1 upon addition of dasatinib (B, lower panel), while the proteins upregulated by process 2^- (C, top panel), remained localized to the same area of the graph (C, lower panel). Proteins in panel (B): 1 – pY(4 0 6)Gab1 (KDASSQDc(pY)DIPR), 2 – pY(6 2 7)Gab1 (GDKQVE(pY)LDLDLDSGK), 3 – pY(4 5 5)Cbl, 4 – pY(1148)EGFR, 5 – pY(1 5 0)tensin4, 6 – pS(7 7 6)Y(7 8 0)tensin3, 7 – pY(1173)EGFR, 8 – pY(3 1 7)Shc, 9 – pY(6 2 7)Gab1 (QVE(pY)LDLDLDSGK), 10 – pY(4 0 6)Gab1 (DASSQDc(pY)DIPR). Proteins in panel (C): 1 – pY(4 5 3)GAREM, 2 – pY(2 3 9)Shc, 3 – pY(2 6 0)ATP1A1, 4 – pY(4 5 5)Cbl, 5 – pS(2 2 9)Y(2 3 1)CENTD2, 6 – pY(1100)ARHGEF5, 7 – pY(1173)EGFR, 8 – pY(3 1 7)Shc, 9 – pY(1140)ARHGEF5, 10 – pY(6 0 7)PI3K, 11 – pY(7 6 7)GAREM, 12 – pY(1097)ARHGEF5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

was completely abolished by the treatment. As per our prediction, unbalanced process 2^- was not significantly affected by dasatinib, underscoring the importance of targeting the complete set of unbalanced processes in order to restore the balance in the system.

We note that in the dataset analyzed, only a single dose of dasatinib was tested. A higher dose of the drug could possibly abrogate processes 1^- and 3^- completely.

According to our analysis, combining dasatinib with an EGFR inhibitor, such as erlotinib, should bring about the collapse of the entire unbalanced network in EGF-stimulated MCF10a cells.

Efforts are underway in our laboratory to test the approach in additional experimental systems, and to pave the way towards the development of a computational approach that can be used by clinicians to analyze patient samples and assign smart, personalized drug combinations.

Acknowledgements

We are thankful to Mr. Raven J. Reddy for providing us with the experimental dataset that he gathered.

Funding

The funding sources for this work were from Dr. Nataly Kravchenko-Balasha's research endowments from the Hebrew University of Jerusalem, and from Yissum Research Development Company for the Hebrew University of Jerusalem.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.chemphys.2018.03.005.

References

[1] A. Alyass et al., From big data analysis to personalized medicine for all: challenges and opportunities, BMC Med. Genomics 8 (2015) 33.
[2] L. Hood, Systems biology and p4 medicine: past, present, and future, Rambam Maimonides Med. J. 4 (2013) e0012.
[3] J.J. Zhu, E.T. Wong, Personalized medicine for glioblastoma: current challenges and future opportunities, Curr. Mol. Med. 13 (2013) 358–367.
[4] S.H. Cho, J. Jeon, S.I. Kim, Personalized medicine in breast cancer: a systematic review, J. Breast Cancer 15 (2012) 265–272.
[5] J.M. Drake et al., Phosphoproteome integration reveals patient- specific networks in prostate cancer resource phosphoproteome integration reveals patient-specific networks in prostate cancer, Cell 166 (2016) 1–14.
[6] B. Regierer, V. Zazzu, R. Sudbrak, A. Kühn, H. Lehrach, Future of medicine: models in predictive diagnostics and personalized medicine, Adv. Biochem. Eng. Biotechnol. 133 (2013) 15–33.
[7] N.M. Friedman, I. Linial, D. NachmanPe'er, Using bayesian networks to analyze expression data, J. Comput. Biol. 7 (2000) 601–620.
[8] M. Bansal, D. di Bernardo, Inference of gene networks from temporal gene expression profiles, IET Syst. Biol. 1 (2007) 306–312.
[9] J.C. Mar, C.A. Wells, J. Quackenbush, Defining an informativeness metric for clustering gene expression data, Bioinformatics 27 (2011) 1094–1100.
[10] Jolliffe, I. T. T. Principal component analysis. Springer series in statistics (Springer, 2002). doi:10.1007/b98835
[11] O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 10101–10106.
[12] D.R. Rhodes et al., Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 9309–9314.
[13] M. Nykter et al., Critical networks exhibit maximal information diversity in structure-dynamics relationships, Phys. Rev. Lett. 100 (2008).
[14] A. Levchenko, I. Nemenman, Cellular noise and information transmission, Curr. Opin. Biotechnol. 28 (2014) 156–164.
[15] C. Waltermann, E. Klipp, Information theory based approaches to cellular signaling, Biochim. Biophys. Acta 1810 (2011) 924–932.
[16] A.D.J. van Dijk, H. Lähdesmäki, D. de Ridder, J. Rousu, Selected proceedings of machine learning in systems biology: MLSB 2016, BMC Bioinform. 17 (2016) 437.
[17] P. Creixell et al., Pathway and network analysis of cancer genomes, Nat. Methods 12 (2015) 615–621.
[18] N. Kravchenko-Balasha et al., On a fundamental structure of gene networks in living cells, Proc. Natl. Acad. Sci. U.S.A. 109 (2012) 4702–4707.
[19] R.D. Levine, R.B. Bernstein, Energy disposal and energy consumption in elementary chemical reactions Information theoretic approach, Acc. Chem. Res. 7 (1974) 393–400.
[20] R.D. Levine, Molecular Reaction Dynamics, The University Press (The University Press, Cambridge, 2005.
[21] F. Remacle, N. Kravchenko-Balasha, A. Levitzki, R.D. Levine, Information-theoretic analysis of phenotype changes in early stages of carcinogenesis, Proc. Natl. Acad. Sci. U.S.A. 107 (2010) 10324–10329.
[22] N. Kravchenko-Balasha et al., Convergence of logic of cellular regulation in different premalignant cells by an information theoretic approach, BMC Syst. Biol. 5 (2011) 42.
[23] N. Kravchenko-Balasha, Y.S. Shin, A. Sutherland, R.D. Levine, J.R. Heath, Intercellular signaling through secreted proteins induces free-energy gradient-directed cell movement, Proc. Natl. Acad. Sci.U.S.A. 113 (2016) 5520–5525.
[24] N. Kravchenko-Balasha, J. Wang, F. Remacle, R.D. Levine, J.R. Heath, Glioblastoma cellular architectures are predicted through the characterization of two-cell interactions, Proc. Natl. Acad. Sci. U.S.A. 111 (2014) 6521–6526.
[25] U. Lucia, A. Ponzetto, T.S. Deisboeck, A thermodynamic approach to the 'mitosis/apoptosis' ratio in cancer, Phys. A Stat. Mech. Appl. 436 (2015) 246–255.
[26] U. Lucia, Thermodynamics and cancer stationary states, Phys. A Stat. Mech. Appl. 392 (2013) 3648–3653.
[27] D.T. Haynie, Biological Thermodynamics, Cambridge University Press, 2008, 10.1017/CBO9780511802690.
[28] U. Lucia et al., Constructal thermodynamics combined with infrared experiments to evaluate temperature differences in cells, Sci. Rep. 5 (2015) 11587.
[29] M.M. Inda et al., Tumor heterogeneity is an active process maintained by a mutant EGFR-induced cytokine circuit in glioblastoma, Genes Dev. 24 (2010) 1731–1745.
[30] I.B. Weinstein, Cancer. Addiction to oncogenes–the Achilles heal of cancer, Science 297 (2002) 63–64.
[31] M.J. Lee et al., Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks, Cell 149 (2012) 780–794.
[32] N. Kravchenko-Balasha, H. Johnson, F.M. White, J.R. Heath, R.D. Levine, A thermodynamic-based interpretation of protein expression heterogeneity in different glioblastoma multiforme tumors identifies tumor-specific unbalanced processes, J. Phys. Chem. B (2016), https://doi.org/10.1021/acs.jpcb.6b01692.
[33] R.J. Reddy et al., Early signaling dynamics of the epidermal growth factor receptor, Proc. Natl. Acad. Sci. U.S.A. 113 (2016) 3114–3119.
[34] R.D. Levine, An information theoretical approach to inversion problems, J. Phys. A Math. Gen. 13 (1980) 91.
[35] W.G. McMillan, J.E. Mayer, The statistical thermodynamics of multicomponent systems, J. Chem. Phys. 13 (1945) 276–305.
[36] J.E. Mayer, M.G. Mayer, Statistical Mechanics, Wiley, 1977.
[37] D.A. McQuarrie Statistical Mechanics, first ed. University science books (2000). http://www.uscibooks.com/mcqstatm.htm. (accessed 10.12.15)
[38] N. Kravchenko-Balasha, H. Johnson, F.M. White, J.R. Heath, R.D. Levine, A thermodynamic based interpretation of protein expression heterogeneity in different GBM tumors identifies tumor specific unbalanced processes, J. Phys. Chem. B 120 (2016) 5990–5997.
[39] S. Zadran, F. Remacle, R.D. Levine, miRNA and mRNA cancer signatures determined by analysis of expression levels in large cohorts of patients, Proc. Natl. Acad. Sci. U.S.A. 110 (2013) 19160–19165.
[40] T. Pawson, N. Warner, Oncogenic re-wiring of cellular signaling pathways, Oncogene 26 (2007) 1268–1275.
[41] STRING database, http://string-db.org/.
[42] F.M. White, A. Wolf-Yadlin, Methods for the analysis of protein phosphorylation-mediated cellular signaling networks, Annu. Rev. Anal. Chem. 9 (2016) 295–315.