A Thermodynamic-Based Interpretation of Protein Expression Heterogeneity in Different Glioblastoma Multiforme Tumors Identifies Tumor-Specific Unbalanced Processes

Citation: Kravchenko-Balasha, Nataly; Johnson, Hannah and White, Forest M. “A Thermodynamic-Based Interpretation of Protein Expression Heterogeneity in Different Glioblastoma Multiforme Tumors Identifies Tumor-Specific Unbalanced Processes.” The Journal of Physical Chemistry B 120, 26 (July 2016): 5990–5997 © 2016 American Chemical Society

As Published: http://dx.doi.org/10.1021/acs.jpcb.6b01692

Publisher: American Chemical Society (ACS)

Persistent URL: http://hdl.handle.net/1721.1/110914

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

Published in final edited form as:

J Phys Chem B. 2016 July 7; 120(26): 5990–5997. doi:10.1021/acs.jpcb.6b01692.

A Thermodynamic Based Interpretation of Protein Expression Heterogeneity in Different GBM Tumors Identifies Tumor Specific Unbalanced Processes

Nataly Kravchenko-Balasha[1,2], **Hannah Johnson**[3], **Forest M. White**[4], **James R. Heath**[1,#], and **R. D. Levine**[5,6,*]

[1]NanoSystems Biology Cancer Center, Division of Chemistry, Caltech, Pasadena, CA, United States

[2]Bio-Medical Sciences department, The Faculty of Dental Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel

[3]Signaling Programme, the Babraham Institute, Babraham, Cambridge, United Kingdom

[4]Department of Biological Engineering, MIT, Cambridge, MA, United States

[5]Department of Molecular and Medical Pharmacology, David Geffen School of Medicine and Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA, United States

[6]The Institute of Chemistry, The Hebrew University of Jerusalem, Jerusalem, Israel

Abstract

We describe a thermodynamics-motivated, information theoretic analysis of proteomic data collected from a series of 8 glioblastoma multiforme (GBM) tumors. GBMs are considered here as prototypes of heterogeneous cancers. That heterogeneity is viewed here as manifesting in different unbalanced biological processes that are associated with thermodynamic-like constraints. The analysis yields a molecular description of a stable steady state that is common across all tumors. It also resolves molecular descriptions of unbalanced processes that are shared by several tumors, such as hyperactivated phosphoprotein signaling networks. Further, it resolves unbalanced processes that provide unique classifiers of tumor subgroups. The results of the theoretical interpretation are compared against statistical multivariate methods and are shown to provide a superior level of resolution for identifying unbalanced processes in GBM tumors. The

[#]Corresponding author: heath@caltech.edu, tel: 626 395 6079; fax: 626 395 2355. [*]Corresponding author: rafi@chem.ucla.edu, tel: +1-(310) 206-0476/ +972-2-6585260; fax: 972-2-6513742.

Supporting Information Available

Supporting information includes the following sections. 1. Theory and experimental procedures. 2. Biological classifications of the proteins participating in the unbalanced processes. 3. Principal component analysis of the proteomic data. 4. K-means clustering of the proteomic data. 5. Comparison between CGH datasets from GBM tumors and iTRAQ proteomic datasets. 6. Potential drug targets and pathways as identified by surprisal analysis. 7. Supplementary Figures. 8. Abbreviations used in the text. 9. Supplementary References. There is an additional Excel file containing the identifies of the proteins that are associated with specific constraints.

Competing interests

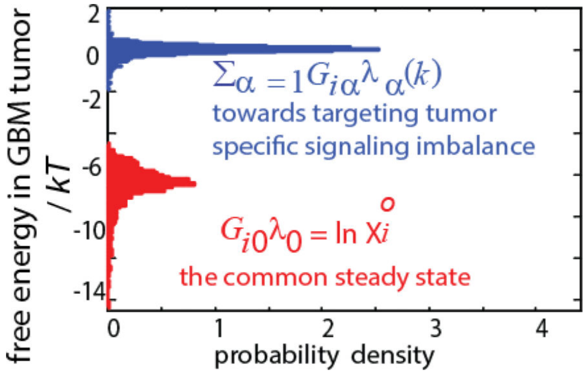
The authors declare that they have no competing interests.

Author contributions:

NKB and RDL designed research; NKB performed research; NKB and RDL contributed analytic tools; NK, FMW, JRH, and RDL analyzed data; NKB, HJ, FMW, JRH, and RDL wrote the paper; HJ designed and performed experimental research; FMW designed experimental research.

identification of specific constraints for each GBM tumor suggests tumor-specific combination therapies that may reverse this imbalance.

Graphical Abstract



Background

GBM is the most common and lethal human brain tumor. GBM tumors exhibit high inter- and intra-tumoral heterogeneity [1, 2], making them one of the most difficult cancers to treat. Despite major efforts to improve GBM patient survival, the majority of patients fail to respond to current standard of care [3]. In fact, certain GBM tumors have been shown to develop resistance to targeted inhibitors via adaptive rather than genetic mechanisms[4]. This highlights the importance of proteomics as a tool for providing an improved understanding of this disease. GBM exhibits high inter-tumor protein expression variability, which is a consequence of both patient specific genetic backgrounds and potentially significant and distinct driver or passenger mutations. To quantitatively analyze how the proteomic heterogeneity of GBM tumors influences functional outcomes, we applied a thermodynamic based theoretical approach that has previously been applied to non-equilibrium chemical and physical systems [5–7]. The goal is to classify GBM tumors within the context of stable, steady states, and unbalanced processes that deviate from those stable states. Our hypothesis is that such a classification might provide guidance for identifying effective therapies and therapy combinations for specific tumors, with the notion that effective therapies are those that remove unbalanced processes.

This paper represents the first application of surprisal analysis to proteomic data. As a proof of principle, we applied this analysis to a previously reported, mass spectrometry-based quantitative protein expression and tyrosine phosphorylation dataset collected from a panel of 8 patient derived GBM xenografts [2]. These tumors variously expressed wild-type (wt) epidermal growth factor receptor (EGFR), overexpressed wtEGFR, or overexpressed the EGFR variant III (EGFRvIII) oncoprotein, and thus reflect some of the dominant molecular signatures that characterize GBM tumors. Quantitative measurements were taken from 4 biological replicates (32 tumors are captured in this dataset) [2].

We base our approach on the premise that biological systems, normal or diseased, reach a state of minimal free energy at the usual conditions of a given temperature and pressure [8–10] subject to environmental and genomic constraints. The stable steady state of the biological system is that state in which the system is unchanging over time, and the inputs and outputs are balanced. An aggressively growing tumor is obviously not at such a stable steady state. Our approach considers that there are environmental and genomic constraints that preclude the system from reaching that stable steady state. This premise implies that tumors with different functional properties, as measured by proteomics, will be subject to the influence of different constraints. Understanding the role of those constraints requires first identifying the stable steady state, which is not influenced by the disease driven constraints.

To identify the most stable states for the GBM tumors we apply a maximum entropy based [11] surprisal analysis to the experimental data. Surprisal analysis has been recently applied to the analysis of biological systems [12–16] where it has been demonstrated to have a predictive power [17]. By determining the theoretically expected distribution of protein species for each GBM tumor, surprisal analysis identifies the state of the minimal Gibbs free energy G when disease operated constraints are not imposed. A decrease in the free energy is the thermodynamic criterion for spontaneous change. Therefore, a basal biological state at minimal free energy is associated with the most stable distribution of proteins. For each GBM tumor separately we determine this distribution. At the transcription level the commonality of the most stable state in diseased and healthy tissues has been shown for several cancer types [13, 16]. We here examine this commonality for the expression levels of proteins in different tumors.

Surprisal analysis is carried out for every protein from the list of thousands of quantified proteins across each of the 8 GBM tumors. The analysis represents the experimental protein and phosphoprotein expression levels, for all proteins measured, as a sum of two contributions. The larger contribution is a basal expression level, (the level of that protein at the minimal free energy of the system), and the smaller one representing the deviations from the most stable state due to the constraints [13–15, 18]. Utilizing the basal expression level of every protein we obtain the most stable distribution of different protein species, and thus the most stable state of the system. An experimental validation of this theoretically identified most stable state has been recently demonstrated [17].

In the current analysis of the GBM tumors we find that a description of the stable steady state is robust and shared across all 8 tumors. This finding gives us confidence in the approach. Beyond this commonality, the tumors vary significantly from each other, as we identify that different constraints are operating in the different tumors.

A given constraint will influence a subset of proteins in a similar way by causing coordinated deviations of the protein levels (up or down) from the basal level which, in turn, can have a functional outcome, such as increased cellular migration. These tumor specific constraints can be used to characterize the intertumor heterogeneity. The molecular composition of those constraints can be mined to suggest drugging strategies for restoring the stable steady state.

Methods

Theory

Surprisal analysis is a theoretical approach based on the key assumption that a biological system, including a tumor, is in a state of minimal free energy subject to constraints. (For more details see the Supporting Information 1, SI section "Surprisal analysis"). It is the constraints that preclude the system from reaching its most stable state where the free energy is at the lowest possible value. The measured protein abundance in every tumor is resolved as the sum of the protein expression level at the most stable state and the unbalanced processes (constraints). Surprisal analysis numerically represents the logarithm of the measured intensities of each protein *i* from a given tumor as a sum of terms (right hand side of equation (1)) where the leading term is the contribution from the most stable state. This is repeated for every GBM tumor. The mathematical procedure known as singular value decomposition (SVD) is then used as a numerical tool to fit the two sets of parameters in equation (1): the weights, $\lambda_\alpha(k)$ of the constraints that are indexed by α (also called Lagrange multipliers), and the $G_{i\alpha}$ (extent of the influence of the constraint α on the levels of each individual protein *i*). The values of $G_{i\alpha}$ are such that their squares sum to unity, so a given protein abundance is fully represented by the steady state level and the constraints that influence that protein. Those proteins with significant $G_{i\alpha}$ values ($G_{i\alpha} > 0.03$ or $G_{i\alpha} < -0.03$, Fig. S1) were further classified according to the GO biological categories using the DAVID database [19] and further examined for their functional connectivity using the STRING data base [20]. Additional Table *SI* (excel table) contains lists of the proteins and their corresponding $G_{i\alpha}$ values that are influenced significantly by the different constraints α and by the steady state. For more details see Supporting Information, SI section "Surprisal analysis" and [13, 15, 17].

We further compared our thermodynamic analysis with two commonly used statistical approaches, principal component analysis (PCA) and K-Means Clustering.

PCA was used to generate patterns of protein expression levels among GBM patients, using either a correlation matrix that standardizes variables, or directly on the data matrix. All the proteins with significant coefficients (Supporting Information, SI section "PCA analysis of the data" describes how 'significance' was defined) in every principle component were analyzed further using the DAVID database [19]. For more details see the Supporting Information, SI section "PCA analysis of the data".

K-Means Clustering was used to partition the dataset into different numbers of mutually exclusive clusters. An iterative Matlab algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters, was utilized (number of iterations=100). Centroid clusters were computed using squared Euclidean distances. For more details see Supporting Information, SI section "K-means clustering of the data

Availability of supporting data

iTRAQ proteomics datasets used in this study are available in [2]. To examine to what extent genomic changes have a proteomic output in GBM tumors, we used CGH data available in the GEO database (GSE39242) as described in the Supporting Information.

Results and Discussion

Application of Surprisal analysis to GBM proteomics

Surprisal analysis was used to characterize changes in the measured expression levels [2] of over a thousand proteins, plus hundreds of phosphorylation sites, for the 8 GBM tumors. These 8 tumors include two expressing wild-type (wt) epidermal growth factor receptor(EGFR) (GBM k = 10, 12 labeled wt), three overexpressing wtEGFR (GBM k = 8, 15, 26, labeled as wt+), and three expressing the EGFRvIII oncogene (GBM k = 6, 39, 59 labeled EGFRvIII).

For each protein species i we fit the observed protein abundance to the theoretically derived equation (1) (For more details see Supporting Information (SI) section 1a “Surprisal analysis”):

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

The left hand side of this equation is the experimentally reported [2] intensity of protein i in tumor k, $X_i(k)$. On the right hand side the first term is the protein intensity X_i^0 at the stable steady state. In the numerical fit of the right hand side of equation (1) to the data, we allow the most stable state term to be different for different tumors and so it is written as $X_i^0(k)$. The sum of terms $\Sigma_{\alpha=1}G_{i\alpha}\lambda_\alpha(k)$ represents deviation from the reference state due to the constraints labeled by α (α=1,2…) in tumor k. This is also a change in the chemical potential of protein i due to the constraints (SI section 1a “Surprisal analysis”). The weight $G_{i\alpha}$ represents the extent of influence of the specific constraint α on each individual protein i . Proteins with the same sign of $G_{i\alpha}$ (see Fig.S1) have a similar behavior in the process α (deviate in the same direction from the basal state). $\lambda_\alpha(k)$ is the weight, or amplitude, of the constraint α in a particular GBM tumor. We rank the constraints by their weight such that α = 1 is most dominant. A genetic defect or epigenetic perturbation that prevents the GBM tumor cells from reaching the most stable state can be considered as a constraint. For consistency of notation in equation (1) we represent the stable state as $\ln X_i^0(k)=G_{i0}\lambda_0(k)$. Each constraint represents a group of collectively acting proteins. These constraints/groups are mathematically independent. Levels of some proteins can be influenced by more than one constraint (unbalanced process) due to the non-linear organization and high interconnectivity of biochemical pathways. Surprisal analysis allows 1) the resolution of unbalanced modes for every protein and 2) the elucidation of whether zero, one or more unbalanced processes influence a particular protein. Such insight is not achieved through a fold change analysis, or other statistical analyses, such as K-mean clustering.

Experimental noise can affect the fitted weights and we report error bars based on four biological replicates for all quantified proteins, phosphorylation sites and for all tumors.

For the iTRAQ proteomic measurements that comprise our database, protein intensities rather than protein concentrations are generated as an output (Supporting Information: SI section 1b “Implementation of surprisal analysis to the iTRAQ technology”). We interpret this mean intensity to be a semi-quantitative reflection of the protein and phosphorylation levels across the population of cells within each tumor sample.

Identification of the most stable state

We first seek to identify the stable steady state as a protein expression level baseline. There is considerable heterogeneity amongst the different tumors, and so each tumor is mathematically allowed to have its own stable state expression levels. It is therefore remarkable that a common stable state ($X_i^0(k)$) was found across all 8 GBM tumors. Mathematically this commonality is represented by the result that the value of $\lambda_\alpha(k)$,the weight of the stable state in tumor k, is the same, within the error bars, for all eight tumors (Fig. 1a).

The levels of more than 500 proteins are well fitted by the stable state term and were not influenced significantly by any unbalanced process. These proteins are identified within the red box of Fig. S1 of the Supporting Information. These proteins participate in the homeostatic functions of the cell, such as protein and RNA metabolism, and the cell cycle (Supporting Information, Table S1). It was previously shown using surprisal analysis of a diverse set of transcriptomic data sets that the most stable state of different organisms, different cell lines and healthy/cancer patients was associated with similar groups of transcripts involved in cellular homeostasis [13, 14, 16]. This suggests that this stable state is also present in normal tissues.

Unbalanced processes operating in GBM tumors

For every observed protein, surprisal analysis identifies which, if any, unbalanced processes influence the observed level of that given protein. When the level of a particular protein is influenced by only a single unbalanced process (i.e. $\alpha = 1$), the corresponding experimental protein expression level will be well approximated by the steady state term ($\ln X_i^0(k)$) and the single deviation term $G_{i1}\lambda_1(k)$:$\ln X_i^0(k) - G_{i1}\lambda_1(k)$. Likewise, if the level of a given protein is influenced by more than one unbalanced processes, more than one deviation term is required to accurately fit the experimental level of that protein.

Regulatory constraint

Equation 1 lists the constraints as $\alpha = 1,2\ldots$ and the corresponding weights $\lambda_1(k)$, $\lambda_2\ (k)\ldots$, with lower valued indices implying more dominant constraints. Our analysis resolved 7 constraints in the data [2] which is the maximal number of constraints allowed by the data for 8 tumors [15]. We use the weight of the $\alpha = 1$ constraint, $\lambda_1(k)$, to illustrate how to interpret the results. Unlike the most stable state term (Fig. 1a), the $\alpha = 1$ constraint has different amplitudes in different tumors. It is particularly active in GBMs k = 10, 15 and 26 (Fig. 1b). Fig. 1c presents a heat map of the $\alpha = 1$ induced deviations $G_{i1}\lambda_1(k)$ of the protein levels from the stable steady state for each tumor k. Fig. 1d is an expanded heat map of the proteins influenced most significantly by $\alpha = 1$. Phosphotyrosine events by far dominate this group (Additional Table *S1* (excel table)), implying that the influence of this constraint is the activation of phosphotyrosine signaling pathways (Supporting Information Fig.S2, Table S2). Note that for GBM10 (a wtEGFR expressing tumor), the constraint represses this activity (λ_1 is negative-valued in Fig 1b, and the heat maps of Fig. 1c and 1d are color-coded blue to indicate proteins with repressed levels relative to the steady state. In contrast, the first constraint leads to increased phosphoprotein signaling in GBM15, GBM26 and, to a lesser extent, in GBM59. This first constraint can be called a phosphorylation constraint.

We further validated the interpretation of the role of the first constraint by performing a surprisal analysis on just the *subgroup* of phosphoproteins. Our interpretation suggests that all members of this group must be subjected to this constraint. Indeed we find that the most stable state distribution for this subgroup that we label as pX_i^0 (Supporting Information, Fig. S3d) is quantitatively similar to the distribution of those same phosphoproteins identified through the first constraint via analysis of the full dataset (Fig. S3b, labeled as $X_i^o + (\alpha=1)$).

To further illustrate the influence of the 1st constraint, we showed that the measured levels of the two phosphoproteins pSTAT3 and pAbi2 were well described by summing the contributions from the stable state term and 1st constraint, $\alpha = 1$ (Fig. 1e and 1f). This summation reproduced the measured levels for all tumors, except for pSTAT3 in GBM59 and GBM12, suggesting that this phophoprotein may be influenced by additional constraints in those two tumors (Supporting Information, SI sections 1a “Surprisal analysis of the data” and “error determination”).

Significant unbalanced processes in the GBM tumors

Higher-index constraints—In Figure 2a we plot the amplitude, of constraints α = 2,3,4 for the 8 GBM tumors. These constraints exhibit a highly variable amplitude across the tumors. For example, the α = 2 constraint exerts a strong influence in two of the three EGFRvIII tumors (GBM39 and 59, Fig. 2a).

GO functional analysis of the proteins associated with this constraint indicates that changes in cell motility/cell adhesion are a significant functional consequence (Supporting Information, Table S3). The relevant proteins include key regulators of cell motion and adhesion, such as the proteins pLyn, pPLCγ, CD44, and pPxn [21–24] (Supporting Information, Tables S3 and Fig. S4). These results correspond with the results obtained from previous studies in which increased cell motility and invasion were detected in GBM EGFRvIII tumors [24, 25]. Similarly, constraints indexed by α = 3–7 (Fig 2a, Fig. S5,S6 and Supporting information, SI sections 1d “*Minor unbalanced processes*”) exhibit appreciable amplitude on only specific tumors, and the identities of the proteins influenced by those unbalanced processes can inform a GO functional analysis for identifying the functional consequences of the specific constraints (Tables S4–S8). Figure 2b highlights the fact that the tumor-specific levels of individual proteins (p-Lyn and p-Pxn are shown) can be influenced by multiple unbalanced processes.

A summary of the unbalanced processes and genomic relationships

A summary of the resolved unbalanced processes, their functional consequences, and their amplitude on the individual tumors, is provided in Fig. 3. Each tumor is influenced by a combination of 2–3 distinct constraints. Molecular and biological analysis of those constraints may provide guidance for combination therapies as discussed below.

The unbalanced processes identified by the information theory analysis of proteomic data should be consistent with genomic information. To explore this, we analyzed available Comparative Genomic Hybridization (CGH) datasets (GSE39242) from the Gene Expression Omnibus (GEO) database [26] for GBM 12, 8, 15, 26 and 59 tumors. This analysis is described in detail in the Supporting Information (SI section 5 "Comparison between CGH datasets from GBM tumors and iTRAQ proteomic datasets"). We identified unbalanced processes and associated biological categories in the CGH data, and compared those to the unbalanced processes from the proteomic data. About 25–30% of the genomic changes correlate with the proteomic changes for a given constraint. In other words, a given constraint will positively influence the levels of around 200 proteins. That same constraint positively influences a similar number of genes, and there is a 25–30% overlap between these two data sets. This correspondence is consistent with literature findings regarding the extent to which variations in protein concentration can be explained through knowledge of variations in mRNA abundances [27]. Interestingly, the robust core of the genomic dataset has significantly larger overlap (~50%) with the proteomic robust core (the stable steady state) and is classified according to similar biological functions (i.e. protein synthesis, mRNA metabolism, etc. (Supporting Information, Table S16)). This result suggests that the homeostatic steady state core possesses significant robustness at both the genomic and proteomic level. It also the validity of the information theoretic analysis of the GBM proteomic data sets.

PCA analysis and k-means clustering

Our use of surprisal analysis is motivated by the notion of creating a physicochemical based framework for understanding biological processes. As in physics and chemistry, a free energy framework has the potential to predict a direction of biological behaviors [17, 28]. Thus, the use of thermodynamics is motivated by pragmatism. It is, however, useful to compare our approach against purely statistical methods that are often applied to large biomolecular data sets, such as Principal Component Analysis (PCA) and K-means clustering [29, 30].

The covariance matrix used in surprisal analysis serves as initial input for further thermodynamic based analysis and theoretical interpretation of the results. That matrix has a mathematical form dictated by the theory and it plays the role of bridging between the proteomic experimental data and the theory (Supporting Information, section 1a "Surprisal analysis of the data" and [15]). It is *NOT* the same matrix that is used in PCA analysis.

PCA analysis is a statistical approach that concentrates on variations relative to the mean (Supporting Information, SI section 3 "PCA analysis"), and so identifies groups of the proteins with similar behavior relative to the mean. The covariance matrix of the PCA analysis is a covariance matrix of the experimental data, and the PCA eigenvectors (the principle components) are analyzed to extract further biological meaning. As an intermediate numerical step, surprisal analysis uses a covariance matrix of the surprisals, meaning the covariance of the natural logarithms. This has a physical, rather than purely statistical, significance.

PCA divided the proteomic data into 7 main patterns of proteomic alterations (Supporting Information, SI section 3 "PCA analysis of the data" and Fig. S7). The top four first

Principal components (PC) accounted for 92% of the data variance (Supporting Information, SI section 3 “PCA analysis of the data”). PCA was extremely sensitive to the phosphoproteins due to higher variance of those proteins. The eigenvectors of all 7 PCs were almost exclusively associated with phosphoproteins (Additional Table *S1* (excel table), and Supporting Information, SI section 3 “PCA analysis of the data”), and thus had limited value beyond classifying the tumors according to phosphoprotein activity.

PCA did resolve significant statistical differences between the tumors (Supporting Information, Fig. S7), but gave little guidance to the biological interpretation of those differences. PC1, for example, resolves a difference between GBM10 and GBM39 (Fig. S7). This difference appears to be associated with the unbalanced processes $\alpha = 1$ and $\alpha = 2$ identified by surprisal analysis. In fact, most proteins associated with PC1 (as identified in Fig. S7d) were part of the lists generated by $\alpha = 1$ and $\alpha = 2$ from surprisal analysis (Fig. S7e). Attempts to classify the proteins associated with PC1 to biological processes did not reveal any significant enriched biological categories. However, central phosphorylated proteins, such as pLyn and pPxn that are known from the literature to participate in cell motility pathways, were picked up by both analyses (Additional Table *S1* (excel table)). Similarly PC2 resolved a difference between GBM15 and GBM6. The proteins associated with PC2 formed an almost complete subset of those associated with $\alpha = 1$ from surprisal analysis (Fig. S7f). However, it is those additional proteins that are captured by surprisal analysis that aid in the biological interpretation. The only PC that showed a pattern similar to the results of surprisal analysis was PC6, which resolved a difference between the GBM 8 and GBM26 in a similar manner to $\alpha = 6$ (Supporting Information, Fig. S7c, Table S12). A summary of the biological categories associated with the different PCs is presented in the Supporting Information, Fig. S9. This table is quite different from that generated by surprisal analysis. In part this is because PCA is more sensitive to the higher variance of the phosphoproteins. As a result, some of the important biological categories, such as glycolysis through oxidative phosphorylation or DNA packaging, were missed in the PCA analysis.

K-means clustering algorithm was able to generate significant clusters (with more than >95% of the analyzed proteins) only when just the phosphorylated proteins were included in the analysis (Supporting Information, SI section 4” K-means clustering of the data”, Fig. S10,S11) and thus had limited biological resolution of the GBM proteomic signatures (Supporting Information, Table S14).

These results suggest that surprisal analysis led to improved resolution of the biological tumor heterogeneity due to mathematical differences in the approaches.

Towards Tumor Specific Drug Targets

Our hypothesis is that drug-targeting an unbalanced process will repress that process and help restore the robust steady state. In other words, an appropriate therapy will reduce the weights, $\lambda_\alpha(k)$, of tumor specific unbalanced processes. In thermodynamic terms, $\lambda_\alpha(k)$ is the measure of how far the constraint α increases the free energy of the tumor.

In order to target a particular unbalanced process, we searched for extensively correlated hub proteins that are influenced by that process. To this end, we appeal to the theory [13]. Given the extent of the influence of the constraint α on the levels of each individual protein *i* ($G_{i\alpha}$), the theory gives the pairwise correlation between proteins *i* and *j* as $G_{i\alpha}$ $G_{j\alpha}$[13]. Typical results for such protein correlations are shown as a heat map in Fig. 4a for the α = 2 process in GBM59.

The group of the proteins influenced the most by the same unbalanced process (significant $G_{i\alpha}$) should exhibit the highest correlation (upper left hand corner of the heat map, Fig. 4a). These proteins identified by surprisal analysis, generate a highly connected protein-protein interaction map, according to the STRING database [20] (Fig. 4b). These proteins generate a network highly enriched in interactions (p-value ~ 0, 140 proteins have more than 360 interactions according to the STRING database). This implies that a protein influenced significantly by an unbalanced process and independently confirmed by STRING to be highly connected, will likely influence the entire unbalanced process. Additional examples drawn from the analysis of other constraints are provided in the SI materials (Figs. S2,S4 and S14,S15- with additional comments in SI section 6). We suggest that targeting of a few proteins with significant $G_{i\alpha}$ values from distinct tumor specific constraints, rather than proteins from the same unbalanced process, will be advantageous for reducing the tumor specific imbalances.

Fig. 4c compares protein intensities, in logarithmic scale, as contributed by the steady state term and the intensities due to the unbalanced processes. Fig. 4c shows that for most proteins in any particular unbalanced process α their intensities are centered about zero. Therefore only a limited number of proteins have significant $G_{i\alpha}$ values and so need to be examined.

Fig. 4d suggests that targeting a combination of highly correlated proteins from α =2,4 unbalanced processes that specifically characterize GBM59 could effectively reduce the tumor imbalance and shift the tumor towards the balanced state. In GBM59 the largest unbalanced process, α = 1, is less specific.

To illustrate a proposed strategy for choosing potential protein candidates for tumor specific combination therapy we consider the pair of EGFRvIII tumors, GBM 59 and 39. These are similar in the constraints α = 1, 2 (Fig.3). They possess known GBM targets such as EGFR and the Lyn/Src family kinases [31–33] which are, in fact, influenced by the constraints. However these tumors differ in the α = 4 constraint. GBM 39 tumor has an additional induced migration module with PDGFR as a potential druggable target [31, 33] while analysis of the GBM59 tumor points to enhancement of aerobic glycolysis through pPKM2 (y105, Supporting Information, Fig. S14,S15). This provides the hypothesis that a PDGFR inhibitor (e.g. Imatinib), if used in combination with an EGFR (e.g. Erlotinib) and Lyn/Src inhibitor (e.g. Dasatinib), would be more effective in treating GBM 39 (reducing GBM39 specific imbalances) than GBM 59. Note that in addition to a very high functional connectivity, pEGFRy992 and pLyn are influenced by more than one unbalanced process (Additional Table *SI*, (excel table)), thus targeting those hub proteins can be particularly advantageous for pushing the GBM39 tumor towards the stable steady state. In general, we look for hub proteins that are influenced by an important constraint and preferably more than one.

Conclusion

GBM is a prototypical heterogeneous tumor, yet surprisal analysis was able to demonstrate a reference level (the steady state) *common* for *all* the quantitatively measured proteins across a panel of GBM tumors. Some of the experimentally measured protein levels are close to their steady state value while others deviate significantly. For each protein in a given tumor, surprisal analysis resolves the deviation from the steady state into a few, two or three, distinct ongoing processes that reflect biological constraints. These constraints help provide a biological differentiation of the various tumors (Fig. 3) that could not be achieved through PCA or K-means clustering. The definition of a highly robust, stable steady state that is common across all tumors was further supported by a joint genomic/proteomic analysis. This suggests that tumor evolution is driven by unbalanced pathways that vary significantly. The identification of distinct unbalanced processes, the associated biological constraints, and their differential influence across the various tumors, identifies a potential combination of tumor specific pathways from distinct constraints, and highlight a few proteins within each pathway. The unbalanced processes provide targets towards reducing the tumor driving unbalanced processes with the purpose of restoring the stable state. A further validation of our conclusions in different cancers and patients is required.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by an EMBO postdoctoral fellowship to N.K.B. and European Commission FP7 Future and Emerging Technologies–Open Project BAMBI 618024 (to R.D.L), and by the National Cancer Institute (1U54 CA199090-01 JRH PI), the Ben and Catherine Ivy Foundation (JRH), the Jean Perkins Foundation (JRH PI). We are also delighted to submit this paper to this Special Issue honoring our longtime colleague, Bill Gelbart.

References

1. Bonavia R, Inda MM, Cavenee WK, Furnari FB. Heterogeneity Maintenance in Glioblastoma: a Social Network. Cancer Res. 2011; 71:4055–4060. [PubMed: 21628493]
2. Johnson H, Del Rosario AM, Bryson BD, Schroeder MA, Sarkaria JN, White FM. Molecular Characterization of EGFR and EGFRvIII Signaling Networks in Human Glioblastoma Tumor Xenografts. Mol. Cell. Prot. 2012; 11:1724–1740.
3. Mrugala MM. Advances and challenges in the treatment of glioblastoma: a clinician's perspective. Discovery Medicine. 2013; 15:221–230. [PubMed: 23636139]
4. Nathanson DA, Gini B, Mottahedeh J, Visnyei K, Koga T, Gomez G, Eskin A, Hwang K, Wang J, Masui K, et al. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. Science. 2014; 343:72–76. [PubMed: 24310612]
5. Levine RD. An information theoretical approach to inversion problems. J. Phys. A: Gen. Phys. 1980; 13:91.
6. Levine, RD. Molecular Reaction Dynamics. Cambridge, U.K.: The University Press; 2005.
7. Levine RD, Bernstein RB. Energy Disposal Energy Consumption in Elementary Chemical Reactions Information Theoretic Approach. Acc. Chem. Res. 1974; 7:393–400.
8. Mayer, JE.; Mayer, MG. Statistical Mechanics. New York: Wiley; 1966.
9. McMillan WG, Mayer JE. The Statistical Thermodynamics of Multicomponent Systems. J. Chem. Phys. 1945; 13:276–305.

10. McQuarrie, DA. Statistical Mechanics. 1. Boston: Addison Wesley; 1976.
11. Levine, RD.; Tribus, M. The Maximum Entropy Formalism. Cambridge, MA: MIT Press; 1979.
12. Facciotti MT. Thermodynamically Inspired Classifier for Molecular Phenotypes of Health and Disease. Proc. Nat. Acad. Sci. U. S. A. 2013; 110:19181–19182.
13. Kravchenko-Balasha N, Levitzki A, Goldstein A, Rotter V, Gross A, Remacle F, Levine RD. On a Fundamental Structure of Gene Networks in Living Cells. Proc. Nat. Acad. Sci. U. S. A. 2012; 109:4702–4707.
14. Kravchenko-Balasha N, Remacle F, Gross A, Rotter V, Levitzki A, Levine RD. Convergence of Logic of Cellular Regulation in Different Premalignant Cells by an Information Theoretic Approach. BMC Syst. Biol. 2011; 5:42. [PubMed: 21410932]
15. Remacle F, Kravchenko-Balasha N, Levitzki A, Levine RD. Information-Theoretic Analysis of Phenotype Changes in Early Stages of Carcinogenesis. Proc. Nat. Acad. Sci. U. S. A. 2010; 107:10324–10329.
16. Zadran S, Remacle F, Levine RD. miRNA and mRNA Cancer Signatures Determined by Analysis of Expression Levels in Large Cohorts of Patients. Proc. Nat. Acad. Sci. U. S. A. 2013; 110:19160–19165.
17. Kravchenko-Balasha N, Wang J, Remacle F, Levine RD, Heath JR. Glioblastoma Cellular Architectures are Predicted through the Characterization of Two-Cell Interactions. Proc. Nat. Acad. Sci. U. S. A. 2014; 111:6521–6526.
18. Poovathingal SK, Kravchenko-Balasha N, Shin YS, Levine RD, Heath JR. Critical Points in Tumorigenesis: A Carcinogen-Initiated Phase Transition Analyzed via Single-Cell Proteomics. Small. 2016; 12:1425–1431. [PubMed: 26780498]
19. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Gen. Biol. 2003; 4:P3.
20. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, et al. The STRING Database in 2011: Functional Interaction Networks of Proteins, Globally Integrated and Scored. Nucleic Acids Res. 2011; 39(Database issue):D561–D568. [PubMed: 21045058]
21. Bourguignon LY, Zhu H, Shao L, Chen YW. CD44 Interaction with c-Src Kinase Promotes Cortactin-Mediated Cytoskeleton Function and Hyaluronic Acid-Dependent Ovarian Tumor Cell Migration. J. Biol. Chem. 2001; 276:7327–7336. [PubMed: 11084024]
22. Huang C, Jacobson K, Schaller MD. MAP Kinases and Cell Migration. J. Cell. Sci. 2004; 117:4619–4628. [PubMed: 15371522]
23. Ishibe S, Joly D, Liu ZX, Cantley LG. Paxillin Serves as an ERK-Regulated Scaffold for Coordinating FAK and Rac Activation in Epithelial Morphogenesis. Mol. Cell. 2004; 16:257–267. [PubMed: 15494312]
24. Kalluri R, Weinberg RA. The Basics of Epithelial-Mesenchymal Transition. J.Clin. Invest. 2009; 119:1420–1428. [PubMed: 19487818]
25. Lal A, Glazer CA, Martinson HM, Friedman HS, Archer GE, Sampson JH, Riggins GJ. Mutant Epidermal Growth Factor Receptor Up-Regulates Molecular Effectors of Tumor Invasion. Cancer Res. 2002; 62:3335–3339. [PubMed: 12067969]
26. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. Nucleic Acids Res. 2002; 30:207–210. [PubMed: 11752295]
27. Vogel C, Marcotte EM. Insights Into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses. Nat. Rev. Genet. 2012; 13:227–232. [PubMed: 22411467]
28. Shin YS, Remacle F, Fan R, Hwang K, Wei W, Ahmad H, Levine RD, Heath JR. Protein Signaling Networks from Single Cell Fluctuations and Information Theory Profiling. Biophys. J. 2011; 100:2378–2386. [PubMed: 21575571]
29. Jolliffe, IT. Principal component analysis. 2nd. New York: Springer; 2002.
30. Seber, GAF. Multivariate observations. New York: Wiley; 1984.
31. Mao H, Lebrun DG, Yang J, Zhu VF, Li M. Deregulated Signaling Pathways in Glioblastoma Multiforme: Molecular Mechanisms and Therapeutic Targets. Cancer Invest. 2012; 30:48–56. [PubMed: 22236189]

32. Nam S, Kim D, Cheng JQ, Zhang S, Lee JH, Buettner R, Mirosevich J, Lee FY, Jove R. Action of the Src Family Kinase Inhibitor, Dasatinib (BMS-354825), on Human Prostate Cancer Cells. Cancer Res. 2005; 65:9185–9189. [PubMed: 16230377]

33. Ohka F, Natsume A, Wakabayashi T. Current Trends in Targeted Therapies for Glioblastoma Multiforme. Neur. Res. Int. 2012; 2012:878425.

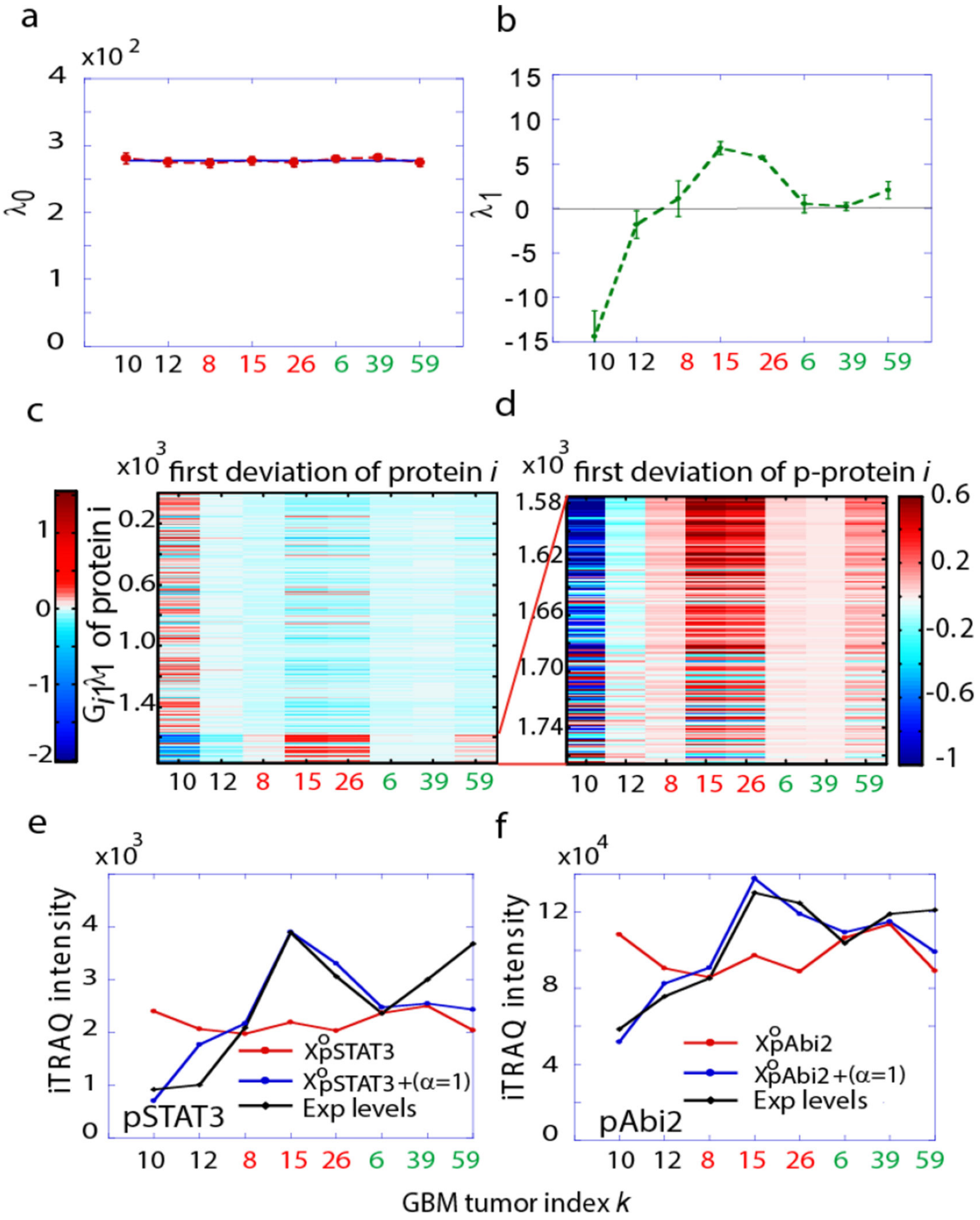


Fig. 1. Identification of the steady state and regulatory constraint in GBM tumors

(**a**) The amplitude of $\lambda_0(k)$., the k'th GBM tumor steady state term α= 0 and error bars for the different tumors. To within the small error bars the steady state is found to be invariant across all tumors. (**b)** The amplitude $\lambda_1(k)$.)of the unbalanced process α = 1 reflects the extent of the deviation from the steady state due to the first constraint. The role of constraint $\alpha$is similar in tumors that have the same sign of $\lambda_\alpha(k)$. The error bars for the $\lambda_\alpha(k)$. were calculated from the errors associated with the mean values of measured proteins as a function of GBM tumor (Supporting Information 1, SI section 1c "error determination"). **(c)**

Heat map representing deviations in protein expression levels from the steady state due to the unbalanced process $\alpha = 1(G_{i1}\lambda_1(k))$. The heat map includes the entire dataset of unmodified and phosphorylated proteins. **(d)** Heat map representing deviations from the steady state only in the subset of phosphorylated proteins due to the α= 1 unbalanced process. **(e, f)** For every measured protein the importance of the unbalanced process α= 1 is determined by comparison of the $X_i^0+(\alpha=1)$ (blue curve), sum of the stable state and the α= 1 deviation term, to the experimental protein expression levels (black curves).

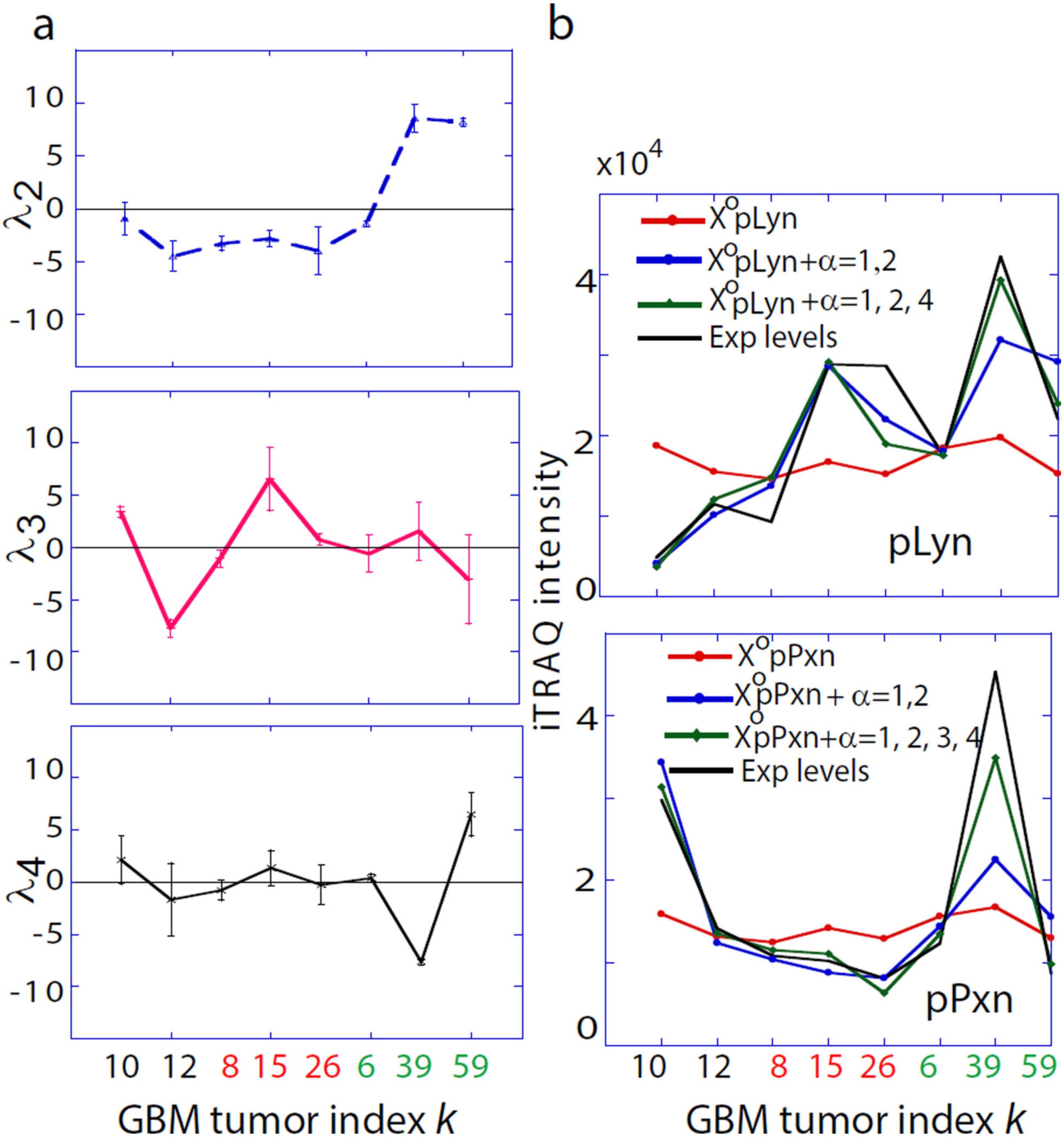
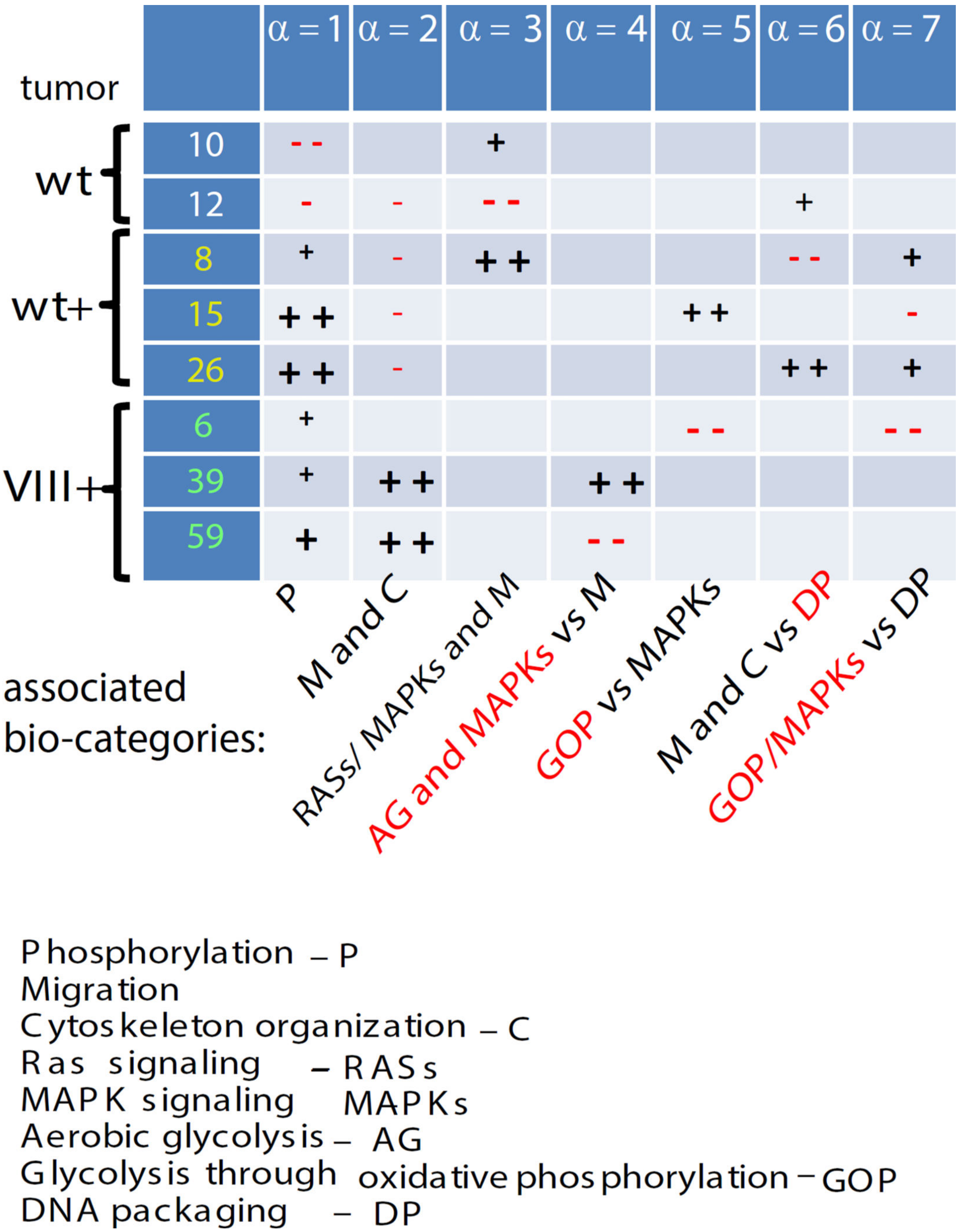


Fig. 2. Unbalanced processes (α = 2,3,4)

(**a**) Amplitudes of the unbalanced processes α = 2, 3, 4 as represented by $\lambda_2(k)$, $\lambda_3(k)$, $\lambda_4(k)$.

(**b**) pLyn and pPxn are shown as examples for the proteins influenced by multiple unbalanced processes. Experimental protein expression levels (black curve) could be closely reproduced for the majority of the tumors (except GBM 26) only when for pLyn the α = 4 deviation term was added to the sum $X_i^o+(\alpha=1,2)$ and for pPxn both α = 3 and α = 4 terms were added to the sum $X_i^o+(\alpha=1,2)$ (green curves), pointing to the significant influence of the constraints α = 3 and α = 4 on the protein expression levels of these proteins.



	tumor	α = 1	α = 2	α = 3	α = 4	α = 5	α = 6	α = 7
wt	10	- -		+				
	12	-	-	- -			+	
wt+	8	+	-	+ +			- -	+
	15	+ +	-			+ +		-
	26	+ +	-				+ +	+
VIII+	6	+				- -		- -
	39	+	+ +		+ +			
	59	+	+ +		- -			
associated bio-categories:		P	M and C	RASs/ MAPKs and M	AG and MAPKs vs M	GOP vs MAPKs	M and C vs DP	GOP/MAPKs vs DP

Phosphorylation – P
Migration
Cytoskeleton organization – C
Ras signaling – RASs
MAPK signaling MAPKs
Aerobic glycolysis – AG
Glycolysis through oxidative phosphorylation – GOP
DNA packaging – DP

Fig. 3. Summary of the unbalanced processes operating in the GBM tumors

Surprisal analysis identified several distinct unbalanced processes for each GBM tumor k. The table summarizes $\lambda_\alpha(k)$ values denoted by different sizes and numbers of + or − symbols. The size and number of +/− signs reflect the relative importance of the particular unbalanced process to that tumor. For example, constraint $\alpha = 4$ includes an enhanced migration network (M) and decreased aerobic glycolysis (AG) and MAPK networks in GBM39, whereas GBM59 exhibits a decreased M network and enhanced AG and MAPK pathways (note the opposite signs for these tumors).

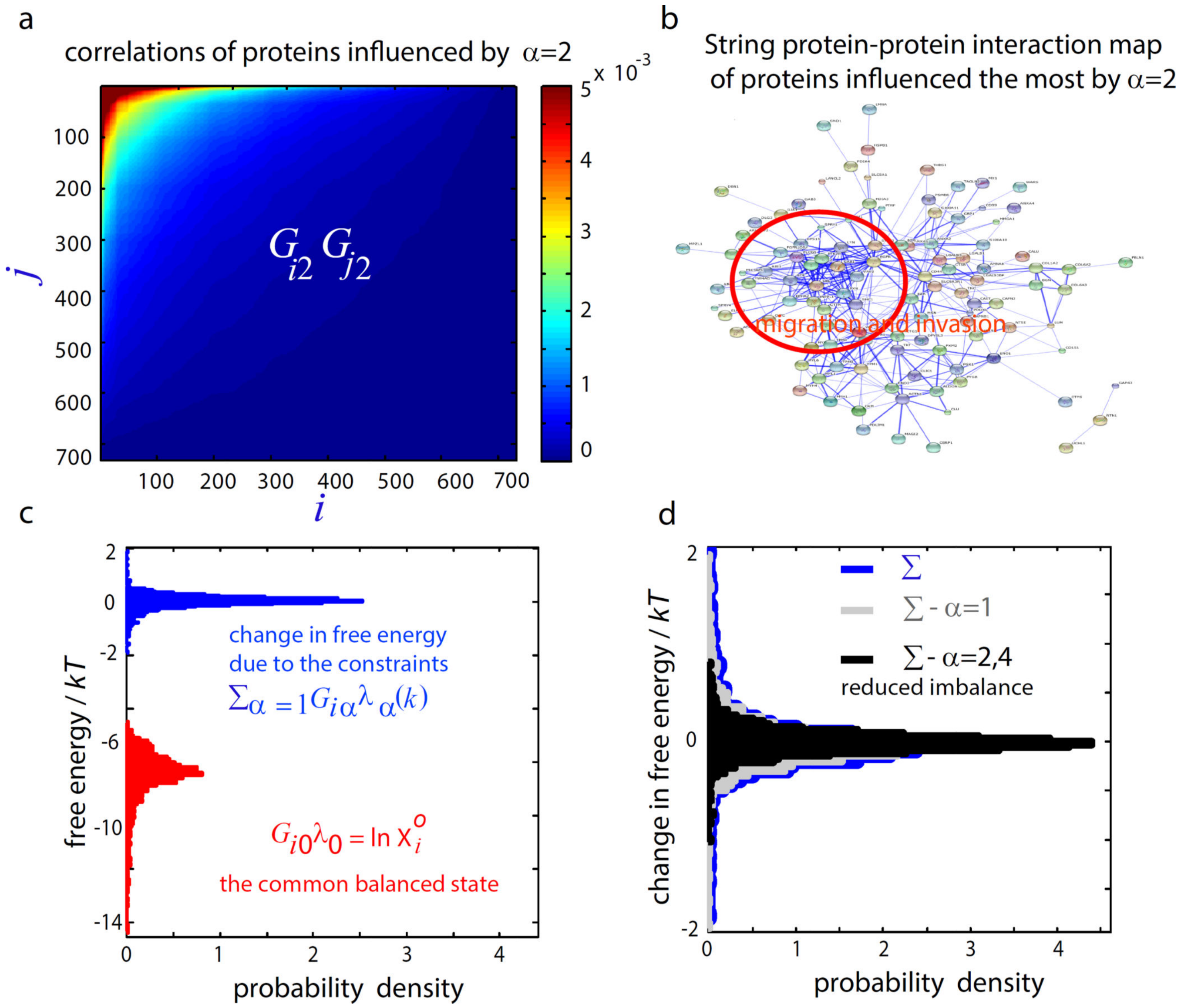


Fig. 4. Towards Tumor Specific Drug Targets

(a) A heat map of the correlation of proteins in the second, α = 2, unbalanced process in GBM59. Shown are all the proteins with $G_{i2} > 0$. Proteins with the highest G_{i2} values (and consequently highest $G_{i2}G_{j2}$ values, values shown in red) are influenced significantly by the unbalanced process. These proteins are affected in the same way by the process. (**b)** 140 of the proteins with the most positive values of G_{i2} (values shown in red in A) were used as an input for generation protein-protein network. Only 115 *connected* proteins are shown. **(c)** Histogram of the protein intensities at the steady state and the deviations thereof in logarithmic scale: $G_{i0}\lambda_0$, that is the minimal value of the free energy and the distribution of the deviations, $\Sigma_{\alpha=1}G_{i\alpha}\lambda_a(k)(\alpha=1,2,..7)$, correspondingly in the tumor GBM59 for every protein *i*. The values of $G_{i0}\lambda_0$ and $\Sigma_{\alpha=1}G_{i\alpha}\lambda_a(k)$ are distributed in a bell-shaped manner about a finite negative number and about zero respectively. **(d)** The effect of a 4-fold decrease in the weights of the unbalanced processes (α = 2, 4) operating in the GBM59 on the free energy. The notation Σ in the figure means $\Sigma_{\alpha=1} G_{i\alpha} \lambda_a(k) (\alpha=1,2,..7)$ and is the same data shown in the upper part of panel **c**. The notation $\Sigma - \alpha = 2,4$ in the figure means the sum $\sum_{\alpha=1}^{7} G_{i\alpha} \lambda_a(k)$ with λ_2 and λ_4 decreased by 4-fold.