Carels et al. EPJ Nonlinear Biomedical Physics (2015) 3:6
DOI 10.1140/s40366-015-0023-3

EPJ Nonlinear Biomedical Physics
a SpringerOpen Journal

RESEARCH Open Access

Optimization of combination chemotherapy based on the calculation of network entropy for protein-protein interactions in breast cancer cell lines

Nicolas Carels[1], Tatiana Martins Tilli[1] and Jack A. Tuszynski[2,3*]

* Correspondence: jackt@ualberta.ca
[2]Department of Oncology, Faculty of Medicine & Dentistry, University of Alberta, Edmonton T6G 1Z2 AB, Canada
[3]Department of Physics, University of Alberta, Edmonton T6G 2E1 AB, Canada
Full list of author information is available at the end of the article

Abstract

Background: In this report, we show how entropy computation can be applied to the characterization of a protein-protein interaction networks to assist the selection of personalized chemotherapeutic strategy for cancer treatment.

Methods: With seven malignant (BT-20, BT-474, MDA-MB-231, MDA-MB-468, MCF-7, T-47D, ZR-75-1) and one healthy (MCF10A) cell lines, we combined interactome and transcriptome data as well as Shanon entropy computation to classify drugs according to their inhibitory potential and to identify the top-5 protein targets best suited for personalized chemotherapy.

Results: We have investigated breast cancer cell lines and found that the entropy of their protein interaction networks is negatively correlated with their sensitivity to target-specific drugs of high potency. This sensitivity is defined as half cell growth inhibition (GI50) with respect to drug administration. By contrast, we found no correlation for drugs that are either of low potency or with no specific molecular targets (cytotoxic). As a result, drugs can be divided into target specific and generally cytotoxic according to the GI50 they produce in malignant cell lines. By extrapolation, we predict that the inactivation of the top-5 up-regulated protein hubs by specific drugs will reduce the protein network entropy by ~2 %, on average, which is expected to substantially increase the benefit of a personalized chemo-therapeutic strategy for patient survival.

Conclusions: We propose several novel drug combinations using only the approved drugs for the inactivation of the target identified in this study with the purpose of increasing patient survival and lowering the deleterious side effects of cancer chemotherapy.

Keywords: Breast cancer; Entropy; Interaction network; Histological subtype; Chemotherapy

Background

Preamble

The text of this report may appear somewhat specialized to some readers not familiar with drug development or systems biology. In order to improve readability of this paper, we introduce the definitions of the key concepts used in what follows.

Interactome

In molecular biology, an interactome is the whole set of molecular interactions in a particular cell. It refers specifically to physical interactions among proteins, also known as protein-protein interactions, i.e., physical contacts established between two or more proteins as a result of biochemical events and/or biophysical forces. Here, we more particularly refer to transient interactions among proteins in the context of signaling networks, i.e., the protein pathways that connect protein receptors on the cell surface with transcription factors that (up- or down-) regulate gene expression. Mathematically, interactomes are generally displayed as graphs (networks).

Networks

Complex networks are ubiquitous in nature. Mathematically, a network may be described by either a directed or undirected graph $G = (V, E)$ with vertex and edge sets V and E, respectively. An edge appears in the graph if there is a known interaction of the two partners, for example two interacting proteins in a cell, either by direct binding or by enzymatic catalysis. A node is referred to as a node of degree k if it is connected to other nodes by k edges. The connectivity level (or rate) of a network characterizes the average number of interactions (edges) per node. When, a node has a number of interactions (connections or edges) significantly larger than the average, it is called a hub. Top-5 (or 10, or more) refers to the 5 (or 10) best items in a list for a given feature under consideration.

Entropy

In thermodynamics, entropy (usually denoted by symbol S and referred to as the Boltzmann entropy) is a measure of the number of specific ways in which a thermodynamic system may be internally rearranged between its microstates, which is commonly understood as a measure of disorder. In statistical mechanics Boltzmann's equation relates the entropy S of an ideal gas to the quantity W, which is the number of microstates corresponding to a given macrostate, i.e.

$$S = k_B ln\ W \tag{1}$$

where k_B is the Boltzmann constant equal to 1.38065×10^{-23} J/K. For thermodynamic systems where microstates of the system may not have equal probabilities, the appropriate generalization, called the Gibbs entropy, is:

$$S = -k_B \Sigma p_i ln\ p_i \tag{2}$$

Here, the subscript i runs over all microstates and Eq. (2) reduces to Eq. (1) if the probabilities p_i are all equal.

In information theory, entropy (the so-called Shannon entropy) is the negative of the expected value of the information contained in a message received. Mathematically speaking the Shannon entropy, H, of a discrete random variable X is a measure of the amount of *uncertainty* associated with the value of X when only its distribution is known. So, for example, if the distribution associated with a random variable is constant (i.e. equal to some known value with probability 1), then entropy is minimal and equal to 0.

Degree-entropy is computed for a given network as

$$H = -\sum_{k-1}^{N-1} p(k) \ln p(k) \tag{3}$$

where $p(k)$ represents a probability distribution on the nodes of the network, $p(k) = N_k/N$ with N_k the number of nodes with degree k and N is the total number of nodes in the network.

Betweenness-centrality is a measure of the centrality of a node. Given a network graph $G(E,V)$ consisting of nodes V and edges E, the betweenness-centrality c_B is a measure of the centrality of a node, v. Typically it is the sum of the fractions of shortest paths that pass through v and is given by:

$$c_B = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{4}$$

where $\sigma(s,t)$ is the number of shortest paths between two nodes (s,t) and $\sigma(s,t\backslash v)$ is the number of those paths passing through nodes other than v.

Here, the biological system studied represents the interactome structure for a cell, i.e., the number of edges (interactions with neighbor proteins) per node (proteins in the network). The probability distribution of the events (the probability of a given number of edges per node), coupled with the information amount (the probability of a given number of edges for the node considered multiplied by its base 2 logarithm) of every event (node), forms a random variable whose average (also termed expectation value) is the average amount of information. Its inverse is the network entropy generated by this distribution.

Half cell growth inhibition (GI50)

In the context of whole-cell assays, GI50 is the concentration of a drug that is needed to inhibit 50 % of cell proliferation.

Introduction

Breast cancer is a global disease. It is the most common cancer in women (25 % of all cancers), with nearly 281,840 estimated new cases, and 40,290 estimated deaths in 2015 in US population (http://seer.cancer.gov). Breast cancer is also becoming an increasingly urgent problem in low- and middle-income countries.

In recent years, government, academia, industry and foundations have devoted vast resources to research and development in order to identify cancer-related molecular targets (oncotargets) that might help improve both diagnoses and clinical practices. A consensus regarding the definitive prognostic/predictive analysis has yet to be reached, but significant progress continues to be made in the ongoing search for optimized treatment protocols with improved specificity and reproducibility. Basically, breast cancers are classified according to the type of hormone receptor over-expressed either on the surface, in the cytoplasm or in the nucleus of their malignant cells [1]. The three most important receptors for breast tumor classification are:

(i) *Endocrine receptors*, i.e., estrogen (ER) or progesterone (PR) receptors. Breast tumors that grow in response to estrogen are classified as ER+ while those that grow in response to progesterone are classified as PR+. ER+ or/and PR+ tumors (60 % of the cases) are likely to respond to endocrine therapies while ER- and PR- tumors (5 to 10 % of the cases) are not.

(ii) *Human epidermal growth factor receptor 2* (HER2). Malignant cells up-regulate a protein known as HER2/neu in about 20 to 25 % of breast tumors and results in a HER2+ phenotype. These breast tumors tend to be much more aggressive and fast-growing.

(iii) *Triple negative* (TN). About 15–25 % of breast tumors do not over-express any of estrogen, progesterone, or HER2 receptors. TNs are more difficult to treat, since most chemotherapeutic agents target one of the ER, PR or HER receptors and often require combination therapies [2]. The name TN is sometimes used as a surrogate term for basal-like and comprises a very heterogeneous group of cancers. There is no standard classification scheme for TNs, but these malignant cells are frequently defined by cytokeratin 5/6 and EGFR staining. However, no clear criteria or cutoff values have been standardized yet.

The chemotherapy regimens used in breast cancer have a relatively low level of molecular specificity with a wide range of acute and long-term side effects that can be substantially deleterious to patients. In addition, clinicians cannot accurately predict the risk of metastasis development in individual patients. Currently, among about 80 % of patients that received adjuvant chemotherapy, approximately 40 % relapse and ultimately die of metastatic tumors. A further complicating factor in these analyses is that many women who would be cured by local treatment alone, which includes surgery and radiotherapy, will be 'over-treated' and suffer the toxic side effects of chemotherapy needlessly. Based on this context, new strategies, models or paradigms are urgently needed to identify patients, who are at the highest risk for developing metastases, and which might benefit from specific drugs. This approach is at the core of *personalized medicine* (also referred to as *precision medicine*) today.

A tremendous effort is ongoing worldwide to improve treatment success and decrease deleterious side effects in patients. With that concern, cell-lines are very useful models for the identification of clinically relevant molecular determinants of tumor response to drugs. It has been reported that cell lines are, indeed, worthwhile models of primary tumors at both the transcript and genome copy-number levels [3]. The comparative analysis of pathways has shown that the majority of subtype-specific signaling sub-networks are conserved between cell lines and tumors. This similarity is important, given the very different environments between a cell line growing in axenic culture and a primary or metastatic tumor exposed to *in vivo* conditions. This supports the consistency of *in vitro* investigations as relevant inferences for clinical testing [4].

As a fruit of ~30 years of investigations, the interactions between cellular proteins reached a sufficiently high level of description for modeling complex molecular processes such as those involved in cancer. Here, we applied this vast systems biology knowledge base to better understand the behavior of malignant cell lines subjected to drug treatments. We used network entropy as a quantitative measure according to the definition of Shannon [5] to characterize the complexity of protein interaction networks as described by Breitkreutz et al. [6]. We used Eq. (3) to evaluate the network entropy of each of the protein-protein interaction networks considered. This means, we first generated a rank-order distribution function for each network and associated the frequency of a particular number of edges connected to nodes with a probability function, $p(k)$. This was repeated for each particular network with its rearrangement as a result of removing the edges corresponding to the inhibition of a specific protein-

protein interaction due to a targeted pharmacological agent. We have chosen the top 5 protein hubs as predicted best targets for inhibition by drug molecules Our objective has been to quantify the benefit associated to the target inactivation of top-5 protein hubs in up-regulated genes rather than non-hub proteins [7]. We found that the proportion of total entropy represented by top-5 hub proteins is ~2 % of total protein network, on average, which by extrapolation, in the case of breast cancer is expected to bring the 5-year survival in the majority of the cases to 100 % [6, 8], i.e., to improve the 10-year survival expectancy or perhaps even to result in a permanent cure. Consequently, we proposed a few optimized drug combinations based on our inferences using the approved inhibitors available on the market. Each combination is specific to a particular subtype of breast cancer due to their differences in the topology of the corresponding interaction networks.

Methods

Interactome data

The protein connectivity that serves as data for entropy calculation in the present work is based on the protein interactions given in the file intact-micluster.zip available from ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/ (accessed on 04.04.2014). We selected the two columns of UniprotKB identifiers (UID) in the intact-micluster.zip file and eliminated the incomplete pairs (marked as "-", i.e., when an intact access number has no UniprotKB equivalent known). The resulting file contained 308,314 protein pairs. This interaction file was then processed to form a non-redundant UID list used to retrieve the corresponding protein sequences (68,504) by querying UniprotKB at http://www.uniprot.org/help/uniprotkb. Since some UID were obsolete, we substituted them with their current name retrieved by querying the field *search* at UniprotKB using the format 'replaces:*obsolete UID*'. The equivalence between UID and human genes was obtained by homology search (tBLASTn) of protein sequences (68,504) used as queries and human coding sequences (CDS) used as subjects from the dataset (hs37p1.EID.tar.gz) of Fedorov's laboratory [9] available at http://bpg.utoledo.edu/~afedorov/lab/eid.html. Homologous hits were considered significant when their score was ≥120, E-value $\leq 10^{-4}$ and identity rate ≥80 % over ≥50 % of query size (http://mitointeractome.kobic.kr/supplement.php). After elimination of subject redundancy (keeping the hit matching the largest identity rate), the final size of human CDS dataset fully described by protein interactions was 17,301.

Transcriptome data

We recovered transcriptome datasets of cell lines (BT-20, BT-474, MDA-MB-231, MDA-MB-468, MCF-7, MCF10A, T-47D, ZR-75-1, see information at http://www.atcc.org/) from http://www.illumina.com/science/data_library.ilmn. The gene expression profile was evaluated through a homology search with the human CDS sample of Fedorov's laboratory. The fifty bp sequences from transcriptome tags were used as queries in homology searches (BLASTn) in human CDSs. The homology redundancy in the BLASTn output file gave us the tag count per gene, i.e., a profile of human gene expression for the considered sample. Homologous hits were considered significant when covering ≥25 bp (50 % of size).

Each gene expression profile (tag count per gene) was normalized according to CDS size and whole tag count using the formula $(10^9 * C)/(N * L)$, where 10^9 is a correction factor, C is the number of reads that match a gene, N is the total mappable tags in the experiment, and L is the CDS size [10]. When tags were counted for more than one gene isoform (alternative splicing forms), we cumulated counts and allocated them to just one form (the largest one); this strategy means that we looked for gene expression and not isoform expression.

To allow the comparison between independent gene expression profiles, we further applied Quantil-normalization (Q-norm) considering the eight samples of this study [11]. Then, we measured the entropy by summing the product of edge probability per node and its base two logarithm over nodes whose expression level was different from zero according to formula 1 where Pv is the probability of having v edges per node, n is the number of nodes and $i \in \{i, \ldots, n\}$.

$$\mathrm{H(G)} = -\sum_{i=1}^{n} Pv \log(Pv), \tag{5}$$

Net entropy differences occur between cell lines because of a combination of the interactome (network of protein interactions) that define in a fixed way the number of interactions between a protein and its neighbors in the network and the transcriptome that shows whether a gene (corresponding to a node in the protein network) is expressed or not. The interactome does not change from one cell line to another in our computational experiments because it is the product of ~30 years of wet lab experimentation. By contrast, the transcriptome (gene expression) is relatively easy to measure by high throughput sequencing techniques, which allow the identification whether a gene for a protein of the network is expressed or not according to the cell line under consideration. If the gene is not expressed, the corresponding node in the network does not exist in the cell line in the expression state considered and its entropy is not included. In the other cases where the expression is larger than zero, the entropy is computed with the consequence that the network is Boolean in essence, which is an approximation in the sense that each node could be modulated by its level of expression to compute the entropy. However, the Shannon entropy does not account for relative statistical weights and hence this level of information has been neglected.

Classification of genes according to expression rates

Since genes with a low expression rate are the most numerous, the distribution of gene frequency according to normalized tag counts is Poisson like. To classify genes into down- or up-regulated, a symmetrical distribution is necessary in order to estimate a p-value on a Gaussian curve resulting from the best fit with the observed distribution.

To obtain a symmetrical distribution, we subtracted the normalized (according to size and number) data from the transcriptome of a malignant cell line from the non-tumoral cell line (MCF10A). After normalization using Q-norm, the distribution's mean was close to zero for any comparison between a malignant cell line and the control. The $\log_{10}(x_i + 1)$ transformation brought the observed distribution closer to a Gaussian distribution. We used PRISM to perform the best fit (95 %) with a Gaussian distribution of $\log_{10}(x_i + 1)$ data classified by increasing values from the largest negative

number to the largest positive number. In this investigation, we only considered up-regulated genes since it is those genes that encode proteins targeted under the classical concept of protein inhibition by drug binding. The boundaries corresponding to *p*-values of 1 % considering a one-tails *p*-value (up-regulated side) on the best fit of a Gaussian distribution were used to calculate the classification threshold of down- and up-regulated genes on the observed distribution using the inverse function, i.e., $10^{\log 10(xi+1)}$ and subtracting 1 from the result of the exponential. The up-regulated genes at *p*-values <0.001 were those with positive values higher than the classification thresholds of +150.

To calculate the entropy correlation between *potentially up-regulated genes* (≥150 reads per gene) and the total gene sample, we identified the list of genes for which up-regulation occurred at least once over all seven malignant cell lines and summed the entropy per cell line over these gene subsets for the eight cell lines (including the reference MCF10A). We did the same calculation for the set of genes corresponding to the total set minus the potentially up-regulated genes referred to as the *complement* and verified that the total entropy was indeed the sum of those of potentially up-regulated genes and the complement over the eight cell lines. Finally, we calculated correlations of entropies by pairs considering the potentially up-regulated, complement and whole sets of genes.

Drug data

The GI50 were derived from the sd02 datasheet (http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1018854108/-/DCSupplemental/sd02.xlsx) from Heiser et al. [4] and the target annotation associated to these drugs from the sd04 datasheet of the same source. The 74 drugs selected for screening cover a wide range of targets and processes implicated in cancer biology and progression and can be classified into two major groups: (i) agents that target specific receptors (n =54) and (ii) general cytotoxic chemotherapeutics (n = 20), defined as *various.* Thus, we analyzed the correlation between the $-\log_{10}$(GI50), associated to both target-specific and broadly cytotoxic drugs, and the corresponding entropy per node of the protein network in the control MCF10A cell line as well as in luminal A (MCF-7, T-47D, ZR-75-1), luminal B (BT-474) and triple-negative (BT-20, MDA-MB-231, MDA-MB-468) malignant cell lines.

Benefit of targeting up-regulated protein hubs as therapeutic targets

Since the selection of up-regulated protein targets is expected to reduce as much as possible the incidence of adverse side effects for the patient, we calculated the benefit, in terms of entropy per node, that could be associated to the inactivation of top-5 most connected proteins (hubs). To do this, we simply computed the entropy per node associated to top-5 most connected proteins in the context of the up-regulated sub-network and computed the relative difference of entropy per node of this sub-network with and without these top-5 hub proteins. According to Cheang et al. [12], the average probability of 5-year survival of patients with luminal breast cancer is ~90 % while that of the patients with triple-negatives is ~70 % and the control is of course 100 %. Thus, we measured the benefit of protein inactivation by the reduction of entropy of its protein network with the consequence that it comes closer to that of the non-tumoral cell, which can be predicted in terms of benefit (%) to the patient by interpolation using the

orthogonal regression line through the average protein network entropy of luminal, triple-negative and control cell lines as well as their associated 5-year patient survival. Finally, we searched the database of clinically approved drugs that potentially inhibit top-5 hub targets and proposed their optimized combination for new cocktails in the treatment of breast cancer.

Results

The size of the human interactome used in our computational experiment was 9724 proteins common to all eight particular breast cancer cell lines, which represents about one third of the whole human set of expressed genes [13] and half of the human proteome [14]. However, much of the interactome proteins were not expressed in our cell line sample and the number of expressed genes per cell line was 7207, on average (Table 1).

As expected, from the very similar number of expressed genes in the eight cell lines, the entropies per node were also very similar, only differing in the third to fourth decimal place. Consequently, the total entropies cumulated over all nodes differed in the second decimal place and were included in a range of 0.06 units between 2.356 and 2.416 for the sample of up-regulated genes (Fig. 1) and in a range of 0.03 units between 11.415 and 11.441 for the total gene sample (Figs. 1, 2 and 3). This is simply due to the fact that the large majority (6913 genes, on average; $\sigma = 217.6$) of genes encode proteins with low connectivity levels (11 edges, on average; $\sigma = 0.2$), while the sample size of up-regulated genes was much smaller (311 genes, on average; $\sigma = 14.7$) and the connectivity level (22 edges, on average; $\sigma = 1.6$) of their proteins was higher. The features of down-regulated genes were very similar concerning sample size and connectivity to those of up-regulated genes (Fig. 2).

Since it may seem unwarranted to draw conclusions from (i) data that only differ in the second decimal place and (ii) the size of the cell sample addressed here is too small to conduct statistical testing based on the variance, we analyzed entropy patterns among cell lines to detect whether some internal consistency may justify the general trends reported here. We found a positive correlation ($r = 0.72$, $P = 0.04$) between the entropies of the subsets of potentially up-regulated genes ($n = 923$) taking the eight cell lines into account and the

Table 1 Statistics of sample size and entropy per node of cell line samples

Cell line	Histological subtype	N[a]	Entropy	ER/PR[b]	HER2[c]	EGFR[d]	CK5-6[e]
MCF10A	Control	7200	11.7390910	0	0–1+	2+	+
MCF-7	LA[f]	7209	11.7484749	6	0–1+	1+	-
T-47D	LA	7205	11.7567033	Positive	2+		
ZR-75-1	LA	7215	11.7660540	3–4	2+	1+	-
BT-474	LB[g]	7208	11.7443749	0/8	3+	1+	-
BT-20	TN[h]	7205	11.7649257	0	0–1+	2+	-
MDA-MB-231	TN	7208	11.7577692	0	0–1+	1+	-
MDA-MB-468	TN	7207	11.7505671	0	0	3+	-

[a]N: Sample size of expressed genes on a total gene sample of 9724
[b]ER/PR: Estrogen/Progesterone receptor
[c]HER2: Human epidermal growth factor receptor
[d]EGFR: Epidermal growth factor
[e]CK5-6: Cytokeratin 5/6
[f]LA: Luminal A
[g]LB: Luminal B
[h]TN: Triple-negative

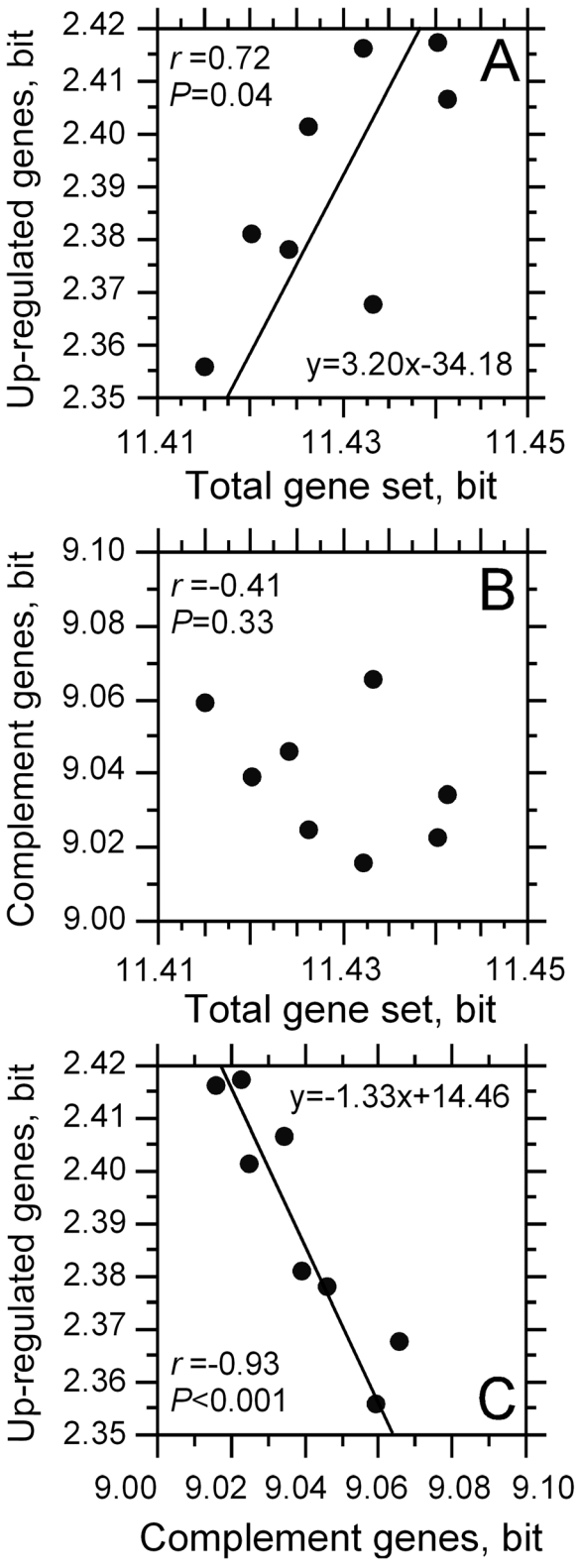

Fig. 1 Correlation of entropy level between (**a**): the up-regulated (the entropy value is computed over 923 genes) and total gene sets (the entropy value is computed over 9724 genes); (**b**): the complement (the entropy value is computed over 8801 genes) and total gene sets; (**c**): the up-regulated and complement gene sets

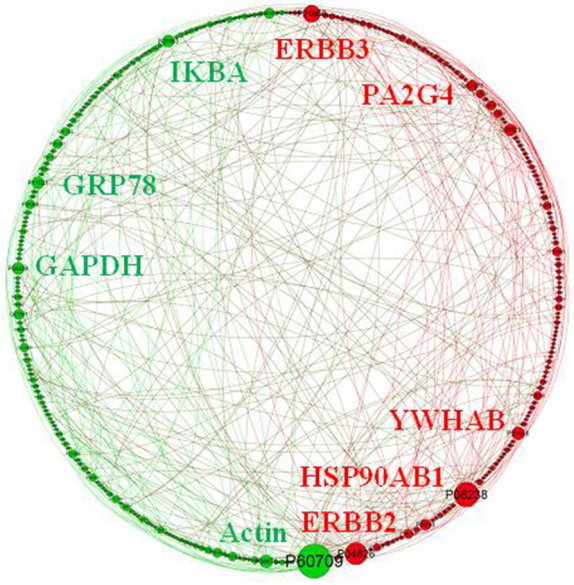

Fig. 2 Sub-network of differentially expressed genes between BT-474 (Luminal B) and MCF10A, represented in a circular layout. Nodes represent genes while links represent interactions between the proteins of these genes. The size of nodes indicates the connectivity rate in the total network (not in the sub-network of down- and up-regulated genes) and color represents an expression pattern (green: down-regulated, red: up-regulated) between tumoral versus non-tumoral breast cell line at $p < 0.01$. Gephi was used to produce the network image

entropies of the total gene set ($n = 9724$) (Fig. 1a). By contrast, we did not find any significant correlation ($r = -0.29$, $P = 0.51$) when comparing entropies of the gene set ($n = 8801$) corresponding to the total sample minus the potentially up-regulated ones (referred to as the *complement*) with the total sample ($n = 9724$) (Fig. 1b). However, the comparison of potentially up-regulated genes ($n = 923$) to the complement ($n = 8801$) demonstrated a correlation of $r = -0.93$ ($P = 0.0002$) (Fig. 1c). This pattern is unlikely to occur just by chance or by some sample bias and demonstrates that the general conclusions drawn in this paper are consistent.

We found a tendency in malignant cell lines to be more sensitive, on the average, to target-specific drugs than to broadly cytotoxic ones (Tables 2, 3 and Additional file 1: Table S1). Cytotoxic drugs with no specific molecular targets performed poorly since their associated $-\log_{10}(\text{GI50})$ was never larger than 5.3, on average, which is considered as a rule of thumb by the state of the art in drug development as the minimal GI50 necessary at a 10 μM concentration to select a candidate molecule to become a potential candidate compound for lead optimization (Fig. 3). By contrast, target specific drugs showed, on the average, a $-\log_{10}(\text{GI50})$ larger than the 5.3 threshold. When comparing the entropy per node of the total protein network of a cell line to its value of $-\log_{10}(\text{GI50})$ for different drugs (Tables 2, 3 and Additional file 1: Table S1), we found a noisy pattern in cytotoxic drugs, while a clear negative correlation ($r = -0.859$, Fig. 3a) could be found, on average, for target specific drugs. The negative correlation was especially convincing for luminal ($r = -0.923$, Fig. 3b) and triple-negative cell lines ($r = -0.725$, Fig. 3c).

Given that target inactivation by specific drugs appeared as a more productive strategy than therapies based on cytotoxic compounds, we listed the top-5 most connected proteins encoded by up-regulated genes according to a p-value of 0.001 (Additional

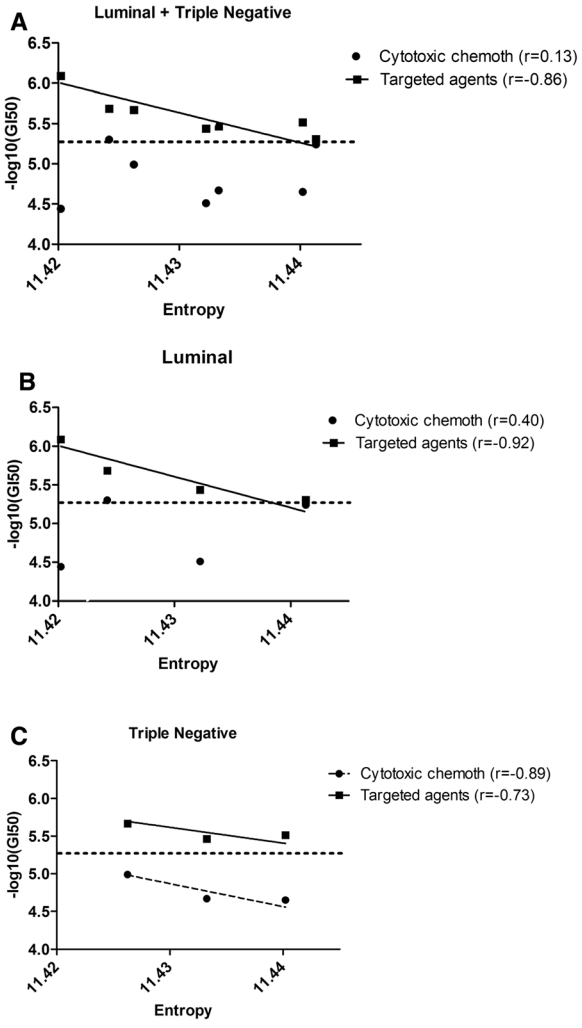

Fig. 3 Relationship between the entropy of the protein interactome of malignant cell lines and their sensitivity to cytotoxic (●) and target specific agents (▌) drugs. Average drug sensitivity across all cell lines (**a**), luminal (**b**) and triple negative (**c**). "chemoth" stands for chemotherapy

file 2: Table S2). The subtraction of the entropy contribution by each top-5 target from the total protein network entropy (PNE) corresponding to the cell line under consideration yields a net entropy corresponding to that network, i.e., the benefit that can be expected from the inactivation of these targets. By interpolation with the orthogonal regression line through average network entropies of triple-negative, luminal and control cell lines, on the one hand, and patient 5-year survival (70, 90, 100, respectively), on the other hand, the inactivation of most top-5 targets resulted in decreasing the total entropy of malignant cell lines to values close to or lower than the entropy of the control. As a consequence of the significant effect of top-5 target inactivation on network entropy reduction in malignant cells, top-5 targets are recommended for drug development in a combination therapy context (Fig. 4 and Additional file 2: Table S2). Actually, we believe that top-5 target inactivation potentially offers a complete 5-year survival of the patient population under consideration provided no serious overlapping toxicities of these drugs exist. The analysis of data from Additional file 2: Table S2 and

Table 2 The panel of cytotoxic drugs classified according to their therapeutic targets, primary effector pathways, or signaling pathway; and the sensitivity for each malignant cell lines

Drugs group	NuclSynt[a]	Metab[b]	DNA[c]	Various[d]	Average
Targets	I[e]	II[f]	III[g]	IV[h]	
L[i]					
MCF-7	4.72	NA	5.07	6.10	5.30
T-47D	3.01	4.46	5.19	5.39	4.51
ZR-75-1	5.16	4.16	5.99	5.65	5.24
BT-474	3.12	3.98	4.76	5.89	4.44
Correl.[j]	0.500[j]	0.433	0.961	−0.623	0.401
TN[k]					
BT-20	3.48	4.69	4.93	5.50	4.65
MDA-MB-231	3.44	4.13	5.71	5.40	4.67
MDA-MB-468	4.31	4.03	5.59	6.04	4.99
Correl.	−0.845	0.926	−0.785	−0.783	−0.891
Tot. correl.[l]	0.197	0.633	0.486	−0.683	0.126

[a]NuclSynt: DNA synthesis
[b]Metab: antimetabolites
[c]DNA: alter DNA structure
[d]Various: diverse spectrum of biological activities
[e]I: TYMS, DNA, RNA, DHFR, GART
[f]II: FDPS
[g]III: DNA cross-linker, TOP1, TOP2A, TOP2BA, TOP2AB, pyrimidine anti-metabolite
[h]IV: PSMD2, PSMB1, PSMB5, PSMB2, PSMD1, IKBKB, SRC, MDM2, FLT3, NTRK1
[i]Luminal: Luminal A and B
[j]Correl: Correlation of drug 50 with entropy per node (see Table 1)
[k]TN: Triple-negative
[l]Tot. correl: correlation for L and TN together

Fig. 4 shows that some approved drugs are already clinically available for inactivation of the hub targets considered here (Table 4). However, it should be noted that the availability of existing target-specific compounds in drug libraries is limited. Clearly, the combination of chemotherapy drugs increases the expected benefit compared to single-drug treatments. Table 4 shows that the expected efficacy of a drug combination varies with the target cell line according to the methodology used in this study.

A drug such as fusicoccin is expected to be effective against most breast cell lines because its protein target is almost always among the top differentially expressed hub proteins (except in MCF-7). By contrast, to increase the patient's 10-year survival, one should complement fusicoccin with some other drugs according to the cell line under consideration. Of course, due to tumor heterogeneity, several cellular phenotypes can be identified at the same time complicating the issue of optimal drug selection immensely. However, the same reasoning should be applied to all of the phenotypes represented, perhaps with a statistical weight applied, but an in-depth discussion on this issue is beyond the scope of the present study so we only address isolated cell lines in what follows. To complement fusicoccin, one may consider the following panel of drugs from the theoretically most efficacious to the less efficacious according to the entropy of their respective target in cell signaling networks: gefitinib, erlotinib, cetuximab, lapatinib, panitumumab, vandetanib, trastuzumab, pertuzumab, afatinib, neratinib, AZD9291 or CLO-1686 > difopein or R18 (triple-negative); CGP78850 or C90 > HDGF-H3 or NSC348884 (Luminal A); and difopein or R18 > trastuzumab, pertuzumab or NeuVax vaccine (Luminal B).

Table 3 The panel of targeted drugs classified according to their therapeutic targets, primary effector pathways, or signaling pathway; and the sensitivity for each malignant cell lines

Drugs group	HSP90*	PI3K-AKT¥	mTOR†	Angiog.‡	EGFR Retal£	MT/Cyts**	RasRaf***	Cell cycle ◦	HDAC◦◦	Average
Targets	I◦◦◦	II††	III‡‡	IV££	V[a]	VI[b]	VII[c]	VIII[d]	IX[e]	
L[f]										
MCF-7	6.43	5.92	6.32	4.47	4.65	7.27	4.99	5.57	5.54	5.69
T-47D	NA	6.21	6.08	4.56	4.62	6.33	4.63	5.49	5.57	5.44
ZR-75-1	6.74	5.66	4.18	4.41	4.57	6.91	4.95	5.13	5.22	5.31
BT-474	7.77	6.42	7.84	4.58	5.73	6.93	4.55	5.14	5.82	6.09
Correl.[g]	−0.462	−0.746	−0.953	−0.699	−0.704	−0.327	0.420	−0.222	−0.918	−0.923
TN[h]										
BT-20	NA	5.38	6.99	4.47	4.99	7.00	4.81	5.40	5.09	5.52
MDA-MB-231	6.82	5.06	5.45	4.42	4.65	7.70	5.22	5.18	4.68	5.47
MDA-MB-468	6.59	5.70	5.40	4.38	4.90	7.78	4.78	6.31	5.19	5.67
Correl.	1.000	−0.504	0.877	0.998	0.245	−0.905	0.064	−0.764	−0.189	−0.725
Tot. correl.[i]	−0.391	−0.605	−0.528	−0.411	−0.546	−0.205	0.351	−0.298	−0.588	−0.859

*HSP90: Heat shock protein 90; ¥PI3K-AKT: Phosphatidylinositol 3' -kinase(PI3K)-Akt signaling pathway; †mTOR: Mammalian target of rapamycin; ‡Angiog.: Angiogenesis ; £EGFR Retal: EGFR/FGFR/HER2/IGFR pathway; **MT/Cyt: microtubule/cytoskeleton; ***RasRaf: Ras-Raf-MEK-MAPK-ERK pathway; ◦Cell cycle: cell cycle involved proteins ; ◦◦HDAC: histone deacetylases; ◦◦◦I: HSP90AA1; ††II: AKT, AKT1, AKT2, PIK3CA, PIK3CB, PIK3CD, PIK3CG, ZNF217; ‡‡III: mTOR; ££IV: MMP2, MMP9, VEGFR2

[a]V: EGFR, ERBB2, ESR1, FGFR3, IGF1R

[b]VI: BCL2, CENPE, Kinesin, ROCK2, TUBB, TUBB1, TUBB3

[c]VII: BRAF, ELK3, MAP2K1, MAP2K2, MAPK9, MAPK10

[d]VIII: AURKA, AURKB, AURKC, CCNB1, CDC25A, CDC25B, CDC25C, CDK1, CDK4, CHEK1, NAE1, PLK1, polyamine analogue

[e]IX:HDAC

[f]L: Luminal

[g]Correlation of drug GI50 with entropy per node (see Table 1)

[h]TN: Triple-negative

[i]Tot. correl: correlation for L and TN together

Discussion

The protein network representation used in this study can be considered very large since from a total of 9724 genes, an average of 7207 were included at the same time for all eight cell lines. In spite of the fact that all conclusions drawn in this paper rely on entropy differences in the second decimal place, we believe that they are significant because an internal pattern was found in the computational experiment in the form of

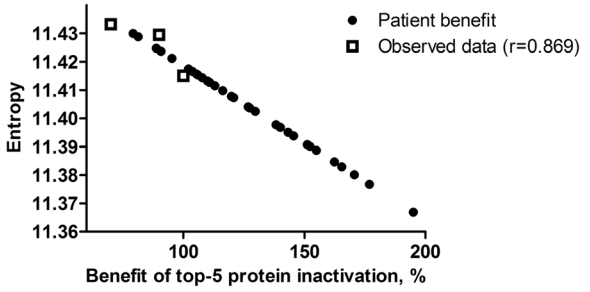

Fig. 4 Expected benefit of top-5 most connected protein of up-regulated genes for 5-year of patient survival. (□): observed data ($r = -0.869$); (●): expected patient benefit derived from $y = -0.0005*x + 11.4732$, where y is the entropy of top-5 protein hubs and x is the 5-year survival benefit expected from target inactivation (see Additional file 2: Table S2) by the corresponding specific drug

Table 4 Expected benefit of drug combination in breast cancer therapy using entropy data from Additional file 2: Table S2 and the relationship $y = -0.0005*x + 11.4732$

	Targets	Drug combination	Benefit, %
BT-20			157
P00533	EGFR	Gefitinib, erlotinib, cetuximab, lapatinib, panitumumab, vandetanib, trastuzumab, pertuzumab, afatinib, neratinib, AZD9291, CLO-1686[a] [15]	120
P62258	YWHAE	Fusicoccin[c] [16]	91
MDA-MB-231			157
P31946	YWHAB	Difopein[b] [17], R18[c] [18]	127
P62258	YWHAE	Fusicoccin[c] [16]	104
MDA-MB-468			229
P00533	EGFR	Gefitinib, erlotinib, cetuximab, lapatinib, panitumumab, vandetanib, trastuzumab, pertuzumab, afatinib, neratinib, AZD9291, CLO-1686[a] [15]	146
P31946	YWHAB	Difopein[b] [17], R18[c] [18]	140
P62258	YWHAE	Fusicoccin[c] [16]	116
MCF-7			216
P62993	GRB2	CGP78850[c] [19], C90[b] [20]	195
P06748	NPM1	NSC348884[c] [21]	111
T-47D			105
P62258	YWHAE	Fusicoccin[c] [16]	
ZR-75-1			236
P62993	GRB2	CGP78850[c] [19], C90[b] [20]	162
P62258	YWHAE	Fusicoccin[c] [16]	89
P51858	HDGF	HDGF-H3[b] [22]	81
P06748	NPM1	NSC348884[c] [21]	79
BT-474			214
P31946	YWHAB	Difopein[b] [17], R18[c] [18]	151
P04626	ERBB2	Trastuzumab, pertuzumab, NeuVax vaccine[a] [23]	130
P62258	YWHAE	Fusicoccin[c] [16]	127

[a]Clinical trials
[b]Pre-clinical animal models
[c]*In vitro* assays

negative and positive correlations that cannot be explained by chance. Of course, the pattern is not only due to up-regulated genes since, for entropy comparison, the same sample size must be taken for all seven malignant cell lines and the number of significantly up-regulated genes is not the same in each of these lines. Thus, according to the cell line a sizeable number of non up-regulated genes may contaminate that sample. However, it is in the potentially up-regulated genes that one must look for an increase of entropy correlating with cell malignancy, which is also consistent with the increased metabolism of malignant cells.

The exercise of correlating potentially up-regulated genes to the gene complement demonstrates the internal consistency of our sample according to the entropy calculation made here. The pattern of entropy distribution found recapitulates the notion that the more malignant a cell line is, the larger is the associated entropy of its network. Interestingly the protein network being finite by nature, if the entropy of up-regulated genes increases, a compensation effect occurs at a cost represented by the entropy of the total network. However, still a higher level of significance exists when considering

entropy differences over the whole sample because this also takes into account genes that are not necessarily *significantly* up-regulated, but still over-expressed compared to the reference. The sum total of entropies over these genes makes a difference at the whole sample level. Therefore, a sample of potentially up-regulated genes cannot be taken into account to calculate the benefit of drug treatment to the patient; the right choice involves the full set of genes.

A correlation between malignancy and PNE was first shown by Breitkreutz et al. [6] by considering different types of tumors, which incidentally did not include breast cancer. Here, we studied several breast cancer cell lines and a similar trend has been observed (Table 1). It is conceivable that the small entropy differences observed here are at least in part due to this specific situation of dealing with a single cancer type, but also due to the different methodology used by Breitkreutz et al. [6].

The boundary between both types of targets (broadly cytotoxic and target-specific) is not always clear because some *specific* targeting drugs may have unintended *off-target* effects leading to inhibition of DNA synthesis, for instance, such as is the case with methotrexate, which is specific for the folate receptor, hence inhibiting purine and pyrimidine base biosynthesis and ultimately blocking DNA synthesis. When grouping drugs by more narrowly defined activities, the noise in the data tended to be reduced and a positive correlation appeared for cytotoxic drugs (data not shown), but the negative correlation associated to specific targets remained. Here, we took a conservative position and presented the data without additional potentially confounding filtering operations. However, it is interesting to note that if the positive correlation between $-\log_{10}(\mathrm{GI50})$ of cytotoxic drugs and PNE were to be confirmed in the future, it would mean that cytotoxic drugs are involved in another type of relationship regarding cell sensitivity compared to target-specific drugs. Rather, it means that the system relies much more on the mechanism that is inactivated when the entropy of the system is high than when it is low. As a metaphor for this concept, the consequences of a central power plant destruction are much greater for an industrialized country than for a developing one, simply because the entire system depends on it due to strong interconnectedness.

A negative correlation between $-\log_{10}(\mathrm{GI50})$ and PNE has the consequence that cells with more complex protein networks (higher entropy) have more options to explore as alternative pathways in order to cope with target inhibition, which is not the case with cells characterized by less complex protein networks (lower entropy) which, as a consequence, are expected to take more time to adapt (or eventually die). This notion is reminiscent of the *gene-for-gene concept* described by Harold Henry Flor [15] in plant pathology. In plants, the gene-for-gene relationship is generally seen as the collapse of a host's resistance to a parasite that may occur as a response to a mutation in a parasite's gene of virulence that allows it to overcome the host's resistance and invade its tissues. For this reason, plant breeding has traditionally coped with resistance collapse through the accumulation of genes encoding host resistance. This process of gene accumulation, which has been called *gene pyramidation* [16], lowers the probability of parasite adaptation by increasing its virulence because the accumulation of virulence genes in a parasite generally decreases its fitness in its environment and, as a consequence, decreases its likelihood.

Conversely, in the case of cancer cell resistance to drugs, the more complex the protein network, the more alternative escape routes/pathways it has compared to a specific

target inhibition. The fact that susceptibility of malignant cell lines is always higher for target-specific than cytotoxic drugs suggests that drug development efforts should be concentrated on target-specific drugs and combinations thereof. Thus, it is natural to expect the development of resistance to specific drugs by malignant cell according to the gene-for-gene or, here, the *gene-for-inhibitor* concept. Following this logic, one may consider a gene-for-inhibitor relationship in the case of malignant cell lines *vis-à-vis* drugs. Accordingly, formulating drugs into a cocktail should overcome malignant cell's resistance, which is actually one of the modern therapeutic trends [6, 7]. However, formulating a drug combination should also account for the dose-limiting negative side effects for normal cells and to protect the immune system's integrity. Thus, we first seek the most probable protein targets (top-5) for drug inhibition in order to maximize the patient benefit from such a therapeutic combination. Top-5 is justified by the fact that more than five drugs cannot be realistically fit within one drug capsule. Of course, drugs could be administrated in several capsules or through intravenous injections. However, such developments should be seen in the scope of clinical trials that we do not address here.

The concept of patient benefit maximization is closely related to the choice of protein targets that act as connectivity hubs in the signaling pathway, but are up-regulated in malignant cells compared to normal cells in order to minimize deleterious side effects for the patient's health. We found that in the majority of cases, one hub-specific drug would be enough to bring the 5-year survival expectancy close to that of normal cells, based on entropy calculations. When the benefit of 5-year survival is estimated to exceed 100 %, it simply means that the benefit should be seen in more than 5-year survival expectancy, i.e., 10-year survival expectancy, which is now the state of the art in breast cancer statistical evaluation.

In general, it is hard to determine what cocktail should be applied to maximize the 10-year survival expectancy and minimize deleterious side effects on patients. Currently, the only way to shed light on this issue is through a trial-and-error experimentation. However, drug combinations from Table 4 could be good starting points based on rational arguments and they can be evaluated immediately in clinical trials since most of the drugs involved are already approved. Interestingly, without taking clinical considerations into account, our investigation shows that fusicoccin should be a basic cocktail component, which should be complemented with other drugs according to the specific breast cancer type developed by the patient in order to maximize the 10-year survival expectancy, which opens an important avenue for personalized medicine.

Conclusions

The response rate to a chemotherapeutic drug treatment may be relatively low in a population of unselected patients. To improve the effectiveness of cancer therapies, a repurposing strategy should include tumor phenotype characterization by molecular techniques in order to design a treatment regimen optimal for the patient outcome. We found that the susceptibility of malignant cells to drugs that are specific for their target is negatively correlated with the entropy of their protein-protein interaction network, which implicitly means that malignant cell resistance to specific drugs is due to the larger number of potential alternative routes in their signaling network. The consequence of the positive correlation between protein network entropy and malignant cell

resistance to specific drugs is that drug cocktails addressing a large number of protein targets are expected to be more effective for the treatment of malignant cell lines with high entropy levels than cocktails targeting only a few proteins. We show that the best protein targets to be addressed for drug development are those (i) whose entropy is large (interaction hubs) and (ii) that are up-regulated in malignant cells compared to normal cells. It is easy to understand that the larger the interaction rate of a protein hub is, the greater its inactivation effect will be on the protein network. It is also readily appreciated that the inactivation of up-regulated hub targets in malignant cells compared to normal ones is more beneficial to the patient because this will minimize negative side effects of a drug treatment. Since specific drugs are more potent and potentially safer than cytotoxic ones, we propose a rational methodology based on protein network entropy to choose the best cocktail of specific drugs according to the protein profile of malignant cells for a given tumor. Our approach differs from the traditional drug repurposing since it allows the application of personalized therapies that should affect essential breast cancer pathways resulting in malignant cell death with minimal side effects for normal cells. In addition, the strategy outlined here should be easy to extend to the personalized therapy of other cancer types.

Additional files

Additional file 1: Table S1. Relationship sensitivity ($-\log_{10}$(GI50)) to drugs with protein network entropy in cell lines of breast cancer. (DOC 167 kb)

Additional file 2: Table S2. Entropy benefit of inactivating top-5 up-regulated hub proteins. (DOC 75 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
NC conceived the study. NC did the scripting and data formatting. NC and TT analyzed the data and wrote the manuscript. JT performed critical reading and improved the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This research was supported by a fellowship from CAPES-Fiocruz (cooperation term 001/2012 CAPES-Fiocruz) to T. M. Tilli, the National Institute for Science and Technology on Innovation on Neglected Diseases (INCT/IDN, CNPq, 573642/2008-7), the Canadian Breast Cancer Foundation, the Allard Foundation and the Alberta Cancer Foundation.

Author details
[1]Laboratório de Modelagem de Sistemas Biológicos, National Institute for Science and Technology on Innovation in Neglected Diseases (INCT/IDN), Center for Technological Development in Health (CDTS), Oswaldo Cruz Foundation (Fiocruz), Rio de Janeiro, Brazil. [2]Department of Oncology, Faculty of Medicine & Dentistry, University of Alberta, Edmonton T6G 1Z2 AB, Canada. [3]Department of Physics, University of Alberta, Edmonton T6G 2E1 AB, Canada.

Received: 16 April 2015 Accepted: 20 July 2015
Published online: 11 August 2015

References
1. Onitilo AA, Engel JM, Greenlee RT, Mukesh BN. Breast cancer subtypes based on ER/PR and Her2 expression. Comparison of clinicopathologic features and survival. Clin Med Res. 2009;7:4–13.
2. Hudis CA, Gianni L. Triple-negative breast cancer: an unmet medical need. Oncologist. 2011;16:1–11.
3. Neves RM, Chin K, Fridlyand J, Yeh J, Baehner FL, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell. 2006;10:515–27.
4. Heiser LM, Sadanandam A, Kuo W-L, Benz SC, Goldstein TC, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proc Natl Sci Acad USA. 2012;109:2724–9.
5. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.
6. Breitkreutz D, Hlatky L, Rietman EA, Tuszynski JA. Molecular signaling network complexity is correlated with cancer patient survivability. Proc Natl Acad Sci U S A. 2012;109:9209–12.
7. Carels N, Tilli T, Tuszynski JA. A computational strategy to select optimized protein targets for drug development toward the control of cancer diseases. PLoS One. 2015;10:e0115054.

8. Breitkreutz D, Rietman EA, Hinow P, Healey L, Tuszynski JA. Complexity of molecular signaling networks for various types of cancer and neurological diseases correlates with patient survivability. In: BIOMAT 2013. Singapore: World Scientific; 2014. p. 250–62.
9. Shepelev V, Fedorov A. Advances in the exon-intron database. Brief Bioinform. 2006;7:178–85.
10. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.
11. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19:185–93.
12. Cheang MCU, Voduc D, Bajdik C, Leung S, McKinney S, et al. Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. Clin Cancer Res. 2008;14:1368–76.
13. Carels N, Frias D. A statistical method without training step for the classification of coding frame in transcriptome sequences. Bioinformatics Biol Insights. 2013;7:35–54.
14. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. A draft map of the human proteome. Nature. 2014. doi:10.1038/nature13302.
15. Arteaga CL, Engelman JA. ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. Cancer Cell. 2014;25:282–303.
16. Bury M, Andolfi A, Rogister B, Cimmino A, Mégalizzi V, et al. Fusicoccin a, a phytotoxic carbotricyclic diterpene glucoside of fungal origin, reduces proliferation and invasion of glioblastoma cells by targeting multiple tyrosine kinases. Transl Oncol. 2013;6:112–23.
17. Cao W, Yang X, Zhou J, Teng Z, Cao L, et al. Targeting 14-3-3 protein, difopein induces apoptosis of human glioma cells and suppresses tumor growth in mice. Apoptosis. 2010;15:230–41.
18. Dong S, Kang S, Lonial S, Khoury HJ, Viallet J, Chen J. Targeting 14-3-3 sensitizes native and mutant BCR-ABL to inhibition with U0126, rapamycin and Bcl-2 inhibitor GX15-070. Leukemia. 2008;22:572–7.
19. Gay B, Suarez S, Caravatti G, Furet P, Meyer T, Schoepfer J. Selective GRB2 SH2 inhibitors as anti-Ras therapy. Int J Cancer. 1999;83:235–41.
20. Giubellino A, Gao Y, Lee S, Lee MJ, Vasselli JR, et al. Inhibition of tumor metastasis by a growth factor receptor bound protein 2 Src homology 2 domain-binding antagonist. Cancer Res. 2007;67:6012–6.
21. Qi W, Shakalya K, Stejskal A, Goldman A, Beeck S, et al. NSC348884, a nucleophosmin inhibitor disrupts oligomer formation and induces apoptosis in human cancer cells. Oncogene. 2008;27:4210–20.
22. Ren H, Chu Z, Mao L. Antibodies targeting hepatoma-derived growth factor as a novel strategy in treating lung cancer. Mol Cancer Ther. 2009;8:1106–12.
23. Schneble EJ, Berry JS, Trappey FA, Clifton GT, Ponniah S, et al. The HER2 peptide nelipepimut-S (E75) vaccine (NeuVax™) in breast cancer patients at risk for recurrence: correlation of immunologic data with clinical response. Immunotherapy. 2014;6:519–31.
24. Flor HH. Current status of the gene-for-gene concept. Annu Rev Phytopathol. 1971;9:275–96.
25. Patocchi A, Walser M, Tartarini S, Broggini GAL, Gennari F, Sansavini S, et al. Identification by genome scanning approach (GSA) of a microsatellite tightly associated with the apple scab resistance gene Vm. Genome. 2005;48:630–6.