Hierarchy of gene expression data is predictive of future breast cancer outcome

Man Chen\textsuperscript{1} and Michael W Deem\textsuperscript{2}

\textsuperscript{1} Department of Physics and Astronomy, Rice University, Houston, TX 77005, USA
\textsuperscript{2} Department of Bioengineering, Rice University, Houston, TX 77005, USA
E-mail: mwdeem@rice.edu

Received 19 April 2013
Accepted for publication 6 September 2013
Published 3 October 2013
Online at stacks.iop.org/PhysBio/10/056006

Abstract
We calculate measures of hierarchy in gene and tissue networks of breast cancer patients. We find that the likelihood of metastasis in the future is correlated with increased values of network hierarchy for expression networks of cancer-associated genes, due to the correlated expression of cancer-specific pathways. Conversely, future metastasis and quick relapse times are negatively correlated with the values of network hierarchy in the expression network of all genes, due to the dedifferentiation of gene pathways and circuits. These results suggest that the hierarchy of gene expression may be useful as an additional biomarker for breast cancer prognosis.

1. Introduction
As cancer develops, there are changes in patterns of gene expression. There are several examples where a defect in a single gene causes a genetic predisposition to developing cancer, for example, the BRCA1 and BRCA2 genes in breast cancer [1–3]. In general, however, the development of cancer is the result of correlated networks of gene expression networks gone awry. That is, cancer is a systemic disease, and changes in the fidelity of gene expression are signs of cancer. In some cases, changes in gene expression networks can determine the disease outcome [4–12]. Thus, structural features of gene expression networks may be ‘biomarkers’ that can predict the probability of a patient developing or surviving cancer.

We here focus on the relation between metastasis and the structure of networks relevant to cancer. Metastasis is the leading cause of cancer mortality [13]. Once metastasis has occurred, the chance of patient survival drops dramatically [14]. Clinicians use prognostic factors such as age or size of tumor at the time of tumor removal to predict the risk of recurrence [14]. Here, we present an analysis of the relation between breast cancer prognosis and the hierarchical structure in correlations of cancer gene expression networks. We will show that these measures of hierarchy in correlations of gene expression distinguish between non-metastatic and metastatic patient populations. We will also show that these measures of hierarchy in gene expression are predictive of the average time of relapse in breast cancer patients.

We are motivated to study the hierarchy of gene expression by a theory that relates hierarchy to environmental stress and variability [15–17]. This theory shows that when a system is placed in a more variable environment, it will tend to become more hierarchical, if it has the ability to do so. This occurs because hierarchy will tend to increase the adaptability of the system. This theory predicts that expression networks of cancer-associated genes may be more hierarchical in more aggressive tumors or during metastasis due to increased correlations in cancer-associated gene pathways. Conversely, measures of hierarchy in the network of all genes will likely decrease for more aggressive tumors or during metastasis, since cancer progression is a dedifferentiation of the entire gene network.

Measures of modularity have been defined for the cancer gene and protein interaction networks. Carro et al identified transcriptional modules in a context-specific regulatory network that controls the expression of the mesenchymal signature associated with metastatic outcome [5, 18]. This result identified a small regulatory module that was part of the mechanism that controlled an important phenotypic state in cancer cells. Chuang et al extracted subnetworks from protein interaction databases and found subnetworks that were significantly enriched with cancer susceptibility genes [5]. A comparison of normal and colon cancer gene
networks identified changes in the network structure. Oslund et al. have ranked cancer gene candidates by local network structures, such as neighbor annotation [19]. Yu et al. have used signature analysis to identify multiple breast cancer modules [20]. Taylor et al. used the co-expression of hub proteins and their partners to identify whether interactions are context specific, i.e., interacting proteins are not always co-expressed, or constitutive, i.e., interacting proteins are always co-expressed [4, 5]. They found that during tumor progression, hub proteins are disorganized by the loss of the coordinated co-expression of components. Thus, changes in the correlation of tumor interactomes were shown to be a prognostic sign in cancer. Other studies have also demonstrated that modularity in the protein–protein interaction network or cell–cell interaction network is an important indicator for cancer prognosis [4] or tumor metastasis [21].

We here quantify the hierarchical structure in cancer networks, generalizing the concept of modularity. Modularity is one measure of the structure of cancer networks. Hierarchy is a measure of the modularity that exists in cancer networks at different scales. The rest of the paper is organized as follows. In section 2, we describe how we created gene and tissue networks from gene expression data previously collected from a population of metastatic and non-metastatic patients. In section 3, we show that hierarchy in networks of cancer-associated genes is positively correlated with metastasis due to the activation of cancer-specific pathways. Conversely, for the networks of all genes, we show that a measure of hierarchy is negatively correlated with metastasis and early recurrence times due to dedifferentiation. We discuss these results in section 4. We conclude in section 5.

2. Methods

We used gene expression profiles of breast cancer patients to construct the networks. The expression profiles were previously obtained from 286 women with lymph node negative disease who had not received adjuvant systemic treatment [22, 23]. In that experiment, the total RNA of frozen tumor samples was hybridized to Affymetrix Human U133a GeneChips. Expression values were calculated by Affymetrix GeneChip analysis software MAS 5.0. Of the many genes analyzed, 76 cancer-associated genes were identified as the predictive of metastasis. Relapse and metastasis of the patients were examined during follow-up visits within five years [22].

Of these patients, 179 did not relapse and metastasize, and 107 were identified to have developed a distant metastasis during follow-up visits within five years [22]. We seek to distinguish from the data at year 0 the 179 patients who were disease free after extended treatment from the 107 patients who developed distant metastases within five years. We constructed cancer networks with two types of nodes: cancer-associated genes or tissue types.

2.1. Gene networks

To construct the networks of the first type, we defined a network with nodes of cancer-associated genes. A total of 76 genes were previously identified as markers that discriminated patients who developed distant metastases from those remaining metastasis-free for five years [22]. We used these 76 cancer-associated genes as the nodes of our network. The links between pairs of nodes were defined by the Pearson correlation coefficient of the two gene expression values:

$$l_{\alpha,\beta} = \sum_{i}^{n} \frac{(\mu_{i,\alpha} - \mu_{\alpha})(\mu_{i,\beta} - \mu_{\beta})}{\sigma_{\alpha}\sigma_{\beta}}.$$  (1)

From this definition, $l_{\alpha,\beta}$ is symmetric in $\alpha$ and $\beta$, and so the graph is undirected. Here $P_{i,\alpha}$ is the expression data of gene $\alpha$ for patient $i$ from [22], $\mu_{\alpha}$ is the average expression value for gene $\alpha$ for the $n$ patients and $\sigma_{\alpha}$ is the standard deviation of the expression value of gene $\alpha$ for the $n$ patients. The expression value, $P_{i,\alpha}$, is a measure of the abundance of the transcript reported by Affymetrix GeneChip analysis software MAS 5.0, scaled to a standard target intensity [22]. To make comparisons between the non-metastatic and the metastatic groups, which contain a different number of patients, we randomly chose 40 patients, each from the non-metastatic group and the metastatic group. This random selection of patients mitigates bias due to differing patient group sizes. We repeated the procedure 100 times, which gives us 100 networks for the metastatic group and 100 networks for the non-metastatic group. Error bars are calculated from this bootstrapping procedure.

2.2. Tissue networks

Networks of the second type are based upon tissues. We used the tissue expression data previously collected for 79 human tissues [24, 25]. We are motivated to study the tissue network because during metastasis cancer spreads between and through different tissue types. The systemic nature of metastasis suggests that the gene expression in different tissue types may be informative to cancer prognosis. The tissue network is built with tissues as nodes and correlation of the gene expression between different tissues as the link values. Specifically, we treat each tissue as a node and build a 79 × 79 tissue network, where the link value between tissue $i$ and tissue $j$ is weighted by the expression data of patient $k$, $P_{k,\alpha}$ from [22] to calculate a Pearson correlation coefficient:

$$t_{i,j} = \frac{(\mu_{i,\alpha} - \mu_{\alpha})(\mu_{j,\alpha} - \mu_{\alpha})}{\sigma_{i,\alpha}\sigma_{j,\alpha}}.$$  (2)

This definition is symmetric in $i$ and $j$, and so the network is undirected. Here $\alpha$ is the gene, $T_{i,\alpha}$ is the expression level of gene $\alpha$ in tissue $i$ from [24], $\mu_{\alpha}$ is the average value of $T_{i,\alpha}P_{k,\alpha}$ over all the available genes and $\sigma_{\alpha}$ is the standard deviation of $T_{i,\alpha}P_{k,\alpha}$. The expression value, $T_{i,\alpha}$, is a measure of the abundance of the transcript reported by Affymetrix GeneChip analysis software MAS 5.0, normalized by a global median setting [24]. We set a cutoff for using tissue expression data, from 10% to 90%. Values of $T_{i,\alpha}$ falling below the cutoff are set to zero. BioGPS was used to map the gene names from the tissue dataset [24] and the breast cancer dataset [22] into NCBI IDs. Equation (2) is an approximation to an ideal of a dataset with expression data for each tissue type from each breast cancer patient. We will show that the structure in the network defined by equation (2) has predictive power for the probability of metastasis within five years.
distances between all nodes of the network, a metric, called the Euclidean commute time distance. This method takes the matrix of distances between each pair of nodes in the network are calculated by the Euclidean commute time distance, i.e. the square root of the average round trip time from one node to the other node and back. For each network, a tree best representing the network is constructed by the average linkage hierarchical clustering algorithm. The distance between two nodes in the tree (y-axis above) is the height above the baseline at which two nodes are joined in the tree topology. To quantify the match between the tree and the original network, we calculated the correlation between the distances in the network and the distances in the tree. This correlation is termed the CCC. The more hierarchical the network, the greater the value of the CCC.

2.3. The CCC

To quantify the structure in these networks, we define a measure of the hierarchy in the networks. Since a tree topology is the archetypal hierarchical structure, we use a measure of hierarchy that quantifies how tree-like the network is. To calculate this measure of hierarchy, we first computed the distance matrix defined by the network. We defined the distance between nodes \(i\) and \(j\), \(d_{ij}\), by the square root of the commute time. The commute time is the expected time a random walker takes to travel from one of the nodes to the other and back [26]. The commute time between nodes \(i\) and \(j\) depends not only on the link value but also on all the other possible paths between nodes \(i\) and \(j\). Note that the commute time between two nodes of a weighted graph decreases when the number of paths connecting the two nodes increases. The commute time between two nodes also decreases when the length of any path connecting the nodes decreases. These properties make the commute time well suited for clustering tasks. To define the commute time, we let \(L\) denote the graph Laplacian, defined as \(L = D - A\), where \(A\) is the matrix of links, \(A = I\) in equation (1) or (2), and the diagonal matrix \(D = \text{diag}(A_i)\), with \(A_i = \sum_j A_{ij}\). The commute time is obtained from \(L_+\), the Moore–Penrose pseudoinverse of the graph Laplacian \(L\) by [27]

\[
n(i, j) = V_G(e_i - e_j)^T L_+ (e_i - e_j)
\]

(3)

Here \(e_i\) and \(V_G = \sum_j a_{ij}\). Since \(L_+\) is symmetric and positive semidefinite, \(d_{ij} = \sqrt{n(i,j)}\) is a Euclidean distance metric, called the Euclidean commute time distance.

We next applied the average linkage hierarchical clustering algorithm to construct the tree topology that best fits the cancer network [28]. This method takes the matrix of distances between all nodes of the network, \(d_{ij}\), and produces a tree topology that best reproduces those distances. The construction of the tree from the network by this algorithm is unique, and approximately optimal in reproducing the distances. The distances of the tree topology are denoted by \(c_{ij}\). The tree topology has the same nodes as the original network, but different links. We calculate the correlation between the original data and the best-fitting tree, which gave the cophenetic correlation coefficient (CCC). The greater the correlation, the more hierarchical the data. The nodes and links of the cancer network define the original data. The tree that best fits the data defines an approximation to the original network, termed the cophenetic matrix. The elements of the cophenetic matrix are the heights where two network nodes become the members of the same cluster in the tree, see figure 1. The distance between the nodes in the best-fitting tree is obtained from the height of the common bifurcation point between those nodes. This height is the cophenetic element of these two nodes, \(c_{ij}\). The correlation between this cophenetic matrix constructed from the best-fitting tree and the original data distances is the CCC. The CCC is a measure of similarity between the original data and the cophenetic matrix. The CCC is defined as

\[
\text{CCC} = \frac{\sum_{i < j}(d_{ij} - d)(c_{ij} - c)}{\sqrt{(\sum_{i < j}(d_{ij} - d)^2)(\sum_{i < j}(c_{ij} - c)^2)}}.
\]

Here \(d\) is the average of the distances in the original network, \(d_{ij}\), and \(c\) is the average of the tree distances, \(c_{ij}\).

3. Results

3.1. Cancer-associated gene network

For the network of cancer-associated genes, we took 76 cancer-associated genes as the nodes, with the link values from equation (1). We built a network by constructing a bootstrap sample of patients from either the non-metastatic outcome or the metastatic outcome groups. We calculated the average CCC value for many 40-person networks extracted from the bootstrap sample. The bootstrap method was then used to calculate the overall average CCC and the standard error of this average. Figure 2(a) shows the result: the hierarchy of the cancer-associated gene network is greater in the metastatic group than that in the non-metastatic group.
Figure 2. (a) The CCC measure of hierarchy in the network of cancer-associated genes for the metastatic and non-metastatic patient groups. We randomly choose 40 patients from each group to construct the two networks. The bootstrap method was used to calculate the averages and standard errors, shown by the error bars. Data are from [22]. This result shows that the network of cancer-associated genes is more hierarchical in the metastatic group. (b) The normalized CCC in the network of cancer-associated genes for the metastatic and non-metastatic patient groups.

Figure 3. The tissue networks with nodes as tissues and links calculated from gene expression values. (a) Link value from tissue–tissue database. (b) Link value weighted by the gene expression data from a patient from the non-metastatic outcome group. (c) Link value weighted by the gene expression data from the metastatic outcome group. Here, only genes with the \( t_c = 0.1 \) highest expression values are used.

We compared these results to those from random networks. We built 100 random networks of the same size and the total number of edges as the cancer-associated network, then randomly reassigned the link values in the network. We define a normalized CCC as \( \text{CCC}_{\text{norm}} = (\text{CCC} - \text{CCC}_{\text{rand}})/(1 - \text{CCC}_{\text{rand}}) \), where CCC is the value of the real cancer-associated network and \( \text{CCC}_{\text{rand}} \) is the average CCC value of the randomized network. We computed the z-score of the cancer-associated network CCC relative to the distribution of the CCC values of the random networks of the same size and sparsity, \( Z_{\text{CCC}} = (\text{CCC} - \text{CCC}_{\text{rand}})/\sigma_{\text{rand}} \). We found \( Z_{\text{CCC}} = 1.64 \) and \( Z_{\text{CCC}} = 2.14 \) for the cancer-associated gene networks in the non-metastatic and metastatic patient groups compared to randomly rewired networks.

To compare the network structure between 76 cancer-associated genes with the network structure of the other genes, we randomly chose 76 genes from a total of 12,926 genes and calculated the CCC for two groups, equation (1). In particular, we constructed a bootstrap sample of all the 12,926 genes, and then calculated the average CCC for the networks of 76 randomly chosen genes from this bootstrap sample. The link value is the Pearson correlation coefficient for each pair of 76 genes for the two groups. We repeated the procedure 1,000 times. The bootstrap method was then used to calculate the average CCC for all the genes and the standard error of this average. Compared to the CCC for the metastatic patient group of 0.918 with the standard deviation 0.0148, the CCC for the non-metastatic patient group with the standard deviation 0.0125 is 0.925. The difference for a student’s t-test is significant, \( p \)-value = \( 1.05 \times 10^{-3} \).

3.2. Tissue–tissue network

We built a tissue–tissue network for each patient, as a function of the expression level cutoff. Nodes are tissues, and link values are given by equation (2). Figure 3 shows an example of this network for a patient from the non-metastatic outcome group and a patient from the metastatic outcome group. Figure 3(a) shows the values of the links in the tissue–tissue network, before scaling by the expression data from the breast cancer patients, i.e. \( l_{i,j} \) from equation (2) with \( P_{k,a} = 1 \). Figure 3(b) shows the values of the links in the tissue–tissue network for a patient in the non-metastatic group, i.e. \( l_{i,j} \) from equation (2), where \( k \) is in the non-metastatic group. Figure 3(c) shows the values of the links in the tissue–tissue network for a patient in the metastatic group, i.e. \( l_{i,j} \) from equation (2), where \( k \) is in the metastatic group. Figure 4(a) shows the amount of hierarchy in the tissue–tissue network with the expression level cutoff ranging from 0.1 to 0.9 for the metastatic and non-metastatic patient groups. For each patient, we determined the time of cancer recurrence. In typical cancer analysis, recurrence within five years of surgery indicates ‘non-cure’. More rapid relapse times are interpreted as more aggressive cancer recurrence.
Figure 4. (a) The average normalized CCC calculated from the tissue–tissue network for patients in metastatic (dashed) and non-metastatic groups (solid). The CCC for metastatic patients is less than that for non-metastatic patients. The error bars are one standard error. The p-value for the student-t test is 0.0295 for the $t_c = 0.1$ data point and less than 0.05 for all three $t_c = 0.3$–0.1 data points. The insert shows the distribution of expression levels within the tissue–tissue network. We use the highest $t_c$ fraction of these data in our analysis. (b) The correlation between CCC and relapse time in the metastatic patients. Cancer appears a dedifferentiation on the set of all gene values, and we observe a correlation between shorter relapse times and lower CCC values.

The structure of the tissue–tissue network is different from that of a random network. As with the gene network, we compared the real tissue network to a randomized network, in which link values are randomly reassigned. The average CCC of the randomized network in the range $t_c = 0.1$–0.9 is 0.864 with the standard deviation 0.017. The corresponding average value of the real network from which figure 4(a) is derived is 0.955 with the standard deviation 0.011. The z-score is, therefore, $Z_{\text{CCC}} = 5.35$, which indicates that there is statistically significantly more hierarchy in the real tissue network than in the randomized network.

4. Discussion

The classical view of cancer is that it is a dedifferentiation of the host. A disruption of the structure of the tissue–tissue network is, indeed, observed in figure 4. The structure of the network in patients with more aggressive, metastatic cancers is more disrupted than in patients with no metastasis. Furthermore, among the metastatic patients, the structure of the network was more disrupted in the patients with the more aggressive cancers that recurred earlier, as shown in figure 4(b).

Both of these results are consistent with the picture of cancer as a general dedifferentiation of the host tissue network. From the point of view of the host, cancer is a disruption. The structure in the host network, which endows the host with robust functioning, is destroyed by the cancer.

Thus, we expect the values of the CCC for the tissue–tissue network to be lower for patients in the metastatic group. Figure 5(a) shows the distribution of the normalized CCC for each patient in the non-metastatic or metastatic outcome groups. We use $t_c = 0.1$ because for this value, there is the greatest discrimination between the metastatic and non-metastatic populations in figure 4(a). These distributions, when averaged, give the $t_c = 0.1$ values in figure 4(a). The distribution of the metastatic outcome group is shifted to lower average normalized CCC values. In addition, the width of the metastatic population distribution is slightly larger. Figure 5(b) shows the probability of metastasis for a patient with a given normalized CCC value, according to the equation

$$p(\text{metastatic}) = \frac{N_{\text{metastatic}} f_{\text{metastatic}}}{N_{\text{metastatic}} f_{\text{metastatic}} + N_{\text{non-metastatic}} f_{\text{non-metastatic}}}.$$  

Here $N_{\text{metastatic}} = 107$ and $N_{\text{non-metastatic}} = 179$. The values of $f_{\text{metastatic}}$ and $f_{\text{non-metastatic}}$ are represented by the dashed and solid curves in figure 4(a), respectively. The quantity $p(\text{metastatic})$ is a biomarker. The biomarker is highly discriminating for low values of the CCC, although only a small fraction of the patients have such low values of the CCC. For example, 5% of the patients values are below $\text{CCC}^\text{norm} = 0.634$, here for which $p(\text{metastatic}) = 0.5$.

We note that the standard BRCA-1 and BRCA-2 biomarkers for cancer apply to roughly 5%–10% of women [29, 30]. Thus, the biomarker in equation (5), which applies to only 5% of the patients, is perhaps of more significance than it may initially appear. Female subjects with BRCA-1 biomarkers have a cumulative lifetime risk of breast cancer in the range of 50%–80%, versus a background risk of 12.5% [29, 30]. The predictive power in figure 5(b), also > 50% for the 5% of the population with CCC < CCC*, is, therefore, perhaps also of greater significance than one might initially think. The CCC may be combined with other genetic biomarkers to achieve increased predictive power [4–12].

In figure 4, we find that highly expressed genes contribute more to the distinction between structures of tissue–tissue networks in patients with metastatic and non-metastatic outcomes. That is, highly expressed genes may have more impact on cancer outcome than lowly expressed genes. We do find, for example, the average expression level of cancer-associated genes is 15.25% higher than the average expression level of all genes.

One view of cancer is that it is an activation of cancer-specific pathways, perhaps hijacked atavistic host pathways [31]. Therefore, by examining cancer-specific pathways, we should see the development of structure
hierarchical than in non-metastatic patients. This increased structure often allows for greater evolvability and is often induced by environmental stress [15]. Networks of cancer-associated genes in metastatic patients are more hierarchical than in non-metastatic patients.

5. Conclusion

We have defined a measure of hierarchy in cancer networks. We found a correlation between the CCC and the clinical outcome. In our study, the CCC of the cancer-associated gene network was higher for the metastatic outcome group than for the non-metastatic outcome group. We anticipated this result, partly because physics of evolution in changing environments [15, 32] suggests that increased hierarchical structure helps cancer to better adapt to the changing environments encountered in metastasis and to overcome the natural barriers to metastasis in the body.

We found that highly expressed genes play a particularly important role in predicting the metastasis of breast cancer. We also found that the disruption of the tissue–tissue network is correlated with both metastatic potential and shorter time of recurrence. Cancer is a complex disease involving genetic, epigenetic and environmental perturbations. Furthermore, cancer operates within and between tissues. Our study of the tissue–tissue network provides additional insights and a possible additional biomarker for breast cancer metastasis and recurrence.

Acknowledgment

This work was supported in part by the US National Institutes of Health under grant 1 R01 GM 100468–01.

References

[1] Balkwill P A, Liu O, Shattuck-Eidens D, Cochran C, Harshman K, Tavtigian S, Bennett L M, Haugen-Strano A, Swensen J and Miki Y 1994 BRCA1 mutations in primary breast and ovarian carcinomas Science 266 120–2
[2] Miki Y et al 1994 A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1 Science 266 66–71
[3] Wooster R, Bignell G, Lancaster J, Swift S, Seal S S, Mangion J, Collins N N, Gregory S, Gumbs C and Micklem G 1995 Identification of the breast cancer susceptibility gene brca2 Nature 378 21–28
[4] Taylor I W, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q and Wrana J L 2009 Dynamic modularity in protein interaction networks predicts breast cancer outcome Nature Biotechnol. 27 199–204
[5] Chuang H Y, Lee E, Liu Y-T, Lee D and Ideker T 2007 Network-based classification of breast cancer metastasis Mol. Syst. Biol. 3 140
[6] Pavlidis P, Lewis D P and Noble W S 2002 Exploring gene expression data with class scores Pac. Symp. Biocomput. 7 474–85
[7] Doniger S W, Salomonis N, Dahlquist K D, Vranizan K, Lawlor S C and Conklin B R 2003 MAPPFinder: using gene
ontology and GenMAPP to create a global gene-expression profile from microarray data Genome Biol. 4 R7

[8] Draghici S S, Khatri P, Martins R P, Ostermeier G C and Krawetz S A 2003 Global functional profiling of gene expression Genomics 81 98–104

[9] Subramanian A et al 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles Proc. Natl Acad. Sci. USA 102 15545–50

[10] Tian L, Greenberg S A, Kong S W, Altshuler J, Kohane I S and Park P J 2005 Discovering statistically significant pathways in expression profiling studies Proc. Natl Acad. Sci. USA 102 13544–9

[11] Wei Z and Li H 2007 A Markov random field model for network-based analysis of genomic data Bioinformatics 23 1537–44

[12] Rapaport F, Zinovyev A, Dutreix M, Barillot E and Vert J P 2007 Classification of microarray data using gene networks BMC Bioinformatics 8 35

[13] Mehlen P and Puisieux A 2006 Metastasis: a question of life or death Nature Rev. Cancer 6 449–58

[14] DeMatteo R P, Lewis J J, Leung D, Mudan S S, Woodruff J M and Brennan M F 2005 Two hundred gastrointestinal stromal tumors: recurrence patterns and prognostic factors for survival Ann. Surg. 231 51–8

[15] Sun J and Deem M W 2007 Spontaneous emergence of modularity in a model of evolving individuals Phys. Rev. Lett. 99 228107

[16] Lorenz D, Jeng A and Deem M W 2011 Modularity in biological system Phys. Life Rev. 8 129–60

[17] Deem M W 2013 Statistical mechanics of modularity and horizontal gene transfer Annu. Rev. Condens. Matter Phys. 4 1–4.25

[18] Carro M S et al 2010 The transcriptional network for mesenchymal transformation of brain tumours Nature 463 318–25

[19] Ostlund G, Lindskog M and Sonnhammer E L 2010 Network-based identification of novel cancer genes Mol. Cell. Proteomics 9 648–55

[20] Yu K, Ganesan Miller K, Lance D M and Tan P 2006 A modular analysis of breast cancer reveals a novel low-grade molecular signature in estrogen receptor-positive tumors Clin. Cancer Res. 12 3288–96

[21] Balkwill F 2004 The significance of cancer cell expression of the chemokine receptor CXCR4 Semin. Cancer Biol. 14 171–9

[22] Wang Y et al 2005 Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer Lancet 365 671–9

[23] www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2034 (accessed 15 July 2013)

[24] Su A I et al 2004 A gene atlas of the mouse and human protein-encoding transcriptomes Proc. Natl Acad. Sci. USA 101 6062–7

[25] www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1133 (accessed 15 July 2013)

[26] Saerens M, Fouss F, Yen L and Dupont P 2004 The principal components analysis of a graph, and its relationships to spectral clustering ECML: Proc. 15th Eur. Conf. on Machine Learning pp 371–83

[27] Barnett S 1990 Matrices: Methods and Applications (Oxford: Clarendon)

[28] Sokal R R and Michener C D 1958 A statistical method for evaluating systematic relationships Univ. Kans. Sci. Bull. 38 1409–38

[29] National Cancer Institute 2008 SEER Cancer Statistics Review 1975–2005 (Bethesda, MD: National Cancer Institute)

[30] Jardines L, Goyal S, Fisher P, Weitzel J, Royce M and Goldfarb S B 2013 Breast cancer overview: risk factors, screening, genetic testing, and prevention Cancer Management: A Multidisciplinary Approach

[31] Davies P C W and Lineweaver C H 2011 Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors Phys. Biol. 8 015001

[32] Kashtan N, Noor E and Alon U 2007 Varying environments can speed up evolution Proc. Natl Acad. Sci. USA 104 13711–6