Computational analysis of molecular networks using spectral graph theory, complexity measures and information theory

Chien-Hung Huang[1], Jeffrey J. P. Tsai[2], Nilubon Kurubanjerdjit[3]
Ka-Lok Ng[2,4] *

[1]*Department of Computer Science and Information Engineering, National Formosa University, Yun-Lin, Taiwan 63205*

[2]*Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan 41354*

[3]*SIQ-DIT research group, School of Information Technology, Mae Fah Luang University, Chiang Rai, Thailand 57100*

[4]*Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan 40402*

[1]*chhuang@nfu.edu.tw* [2]*jjptsai@gmail.com* [3]*sendtoopal@gmail.com* [2,4] *ppiddi@gmail.com*

* *corresponding author*

Abstract

Molecular networks are described in terms of directed multigraphs, so-called network motifs. Spectral graph theory, reciprocal link and complexity measures were utilized to quantify network motifs. It was found that graph energy, reciprocal link and cyclomatic complexity can optimally specify network motifs with some degree of degeneracy. Biological networks are built up from a finite number of motif patterns; hence, a graph energy cutoff exists and the Shannon entropy of the motif frequency distribution is not maximal. Also, frequently found motifs are irreducible graphs. Network similarity was quantified by gauging their motif frequency distribution functions using Jensen-Shannon entropy. This method allows us to determine the distance between two networks regardless of their nodes' identities and network sizes.

This study provides a systematic approach to dissect the complex nature of biological networks. Our novel method different from any other approach. The findings support the view that there are organizational principles underlying molecular networks.

Keywords: network motifs, cancer networks, signal transduction networks, cellular processes, graph theory, information theory, graph energy, network complexity, entropy, Jensen-Shannon entropy

Background

Biological networks and network motifs

Molecular biological networks are the basis of biological processes, where the networks exhibit biological functions through interaction among the genetic elements. Each module is expected to perform specific functions, separable from the functions of other modules [1-3]. Such modular networks can be decomposed into smaller clusters, also known as network motifs. These motifs show interesting dynamical behaviors, in which cooperativity effects between the motif components play a critical role in human diseases.

Essentially, network-based analysis falls into the following major categories: (1) motif identification and analysis, (2) global architecture study, (3) local topological properties, and (4) robustness of the network under different types of perturbations.

For the first category, there are a number of publicly available network motif detection tools, namely MFINDER[4], MAVISTO[5], FANMOD[6], NetMatch[7] and SNAVI[8].

For the second category, many studies have employed random graph theory to characterize the global structures, i.e. assortative/disassortative, small world etc., of molecular networks[9-10]. For instance, it has been shown that protein-protein interaction networks are well described by scale-free networks or hierarchical network models[11].

For the third category, topological graph theory has been utilized to compute certain topological parameters, such as, betweenness centrality, closeness centrality, cluster coefficients, and eigenvector centrality, to characterize networks[12-16].

For the last category, it has been shown that molecular networks are robust under random perturbation but fragile under attack perturbation[17]. Further work has demonstrated that molecular networks are also fragile under degree-based, betweenness-based and brokering-coefficient-based perturbations[18].

Although there exist many published works on network analysis, however, many important issues still remain to be investigated. Most previous studies have utilized graph metrics to analyze the network topology, and thus very relevant yet unanswered question arises – do these network topological parameters convey enough knowledge about the networks? The answer seems to be negative. For example, little is known about the architectures or organizational principles of molecular networks. For instance, can we have a unique label for different motifs? Do certain motif patterns embedded in a network occur at a higher frequency? How can we measure the similarity between two networks without referencing the nodes' identities?

Seminal works on the use of the concepts of information content, topology and

entropy in biology were carried out by Dancoff & Quastler[19], Rashvesky[20-21] and Mowshowitz[22]. In this study, it is hypothesized that network motifs are the fundamental building blocks of a network. In other words, motifs are treated as the core components of a network. Thus, we propose that network properties are captured by motifs compose of *N* nodes, which are referred to as *N*-node motifs in the following discussion. In order to systematically characterize a large network, one identifies the 3-node motifs and 4-node motifs up to the *N*-node motifs embedded in the network.

For a digraph, a total of 2, 13, 199, 9364 and 1530843 possible patterns can be defined for the 2-node, 3-node, 4-node, 5-node and 6-node motifs, respectively[23-24]. Since the time complexity of identifying *N*-node motifs in a large network is a NP-complete problem[25], therefore, we worked with 3-node motifs and 4-node motifs only. Motifs composed of five or more nodes are neglected as a first approximation. As we will show below, this approximation could provide useful insights into dissecting the design principles underlying molecular networks. Motifs composed of five or more nodes will be considered if necessary.

An earlier work[26] has shown that certain motifs do not appear at frequencies significantly greater than those appearing in corresponding random graphs; those motifs still play specify functional roles. This justifies our approach because the present work identifies all possible 3-node and 4-node motifs regardless of their frequency of occurrence. In other words, we adopted the notion that motifs are the basic building blocks but do not necessarily occur frequently in a network.

Adami[27] studied un-directed colored graphs (nodes are labelled with different colors) and showed that the relative frequency of the colored motifs embedded in the network can be used to define the information content of the network. In this work, we considered motifs that are *directed* graphs and could possibly be composed of feedback loop(s), also called directed multigraphs.

Spectral graph theory, reciprocal links, complexity measures and information theory

To characterize network motifs, we utilized the following concepts: spectral graph theory (SGT), reciprocal links and complexity measures. SGT is a powerful approach which has been applied in many areas, such as, computer science, computational biology[28-29]. The eigenvalues of a matrix defined on a graph play an essential role in inferring the structural properties of the graph[30]. According to Mowshowitz[31], the characteristic polynomial of the adjacency matrix of a graph distinguishes between non-isomorphic graphs. A reciprocal link is a parameter that quantifies the degree of bi-directional connection of a network motif.

Complexity arises from the interactions among the constitute components. Many complexity measures have been proposed, but there is no standard or formal definition of complexity metrics which can be applied to all circumstances. Each complexity measure has its strength and weakness[32]. Early works on defining complexity for digraphs and infinite graphs can be traced back to Mowshowitz[33]. The concept of graph complexity indices has been applied to infer the hierarchical order of chemical

structures[34]. Given a network motif pattern, we made use of two commonly used complexity measures to characterize the motif.

It is possible that some of the network motifs are associated the same graph energy (degenerated motifs), we seek to find an unique identifier for each motif pattern. Graph descriptors are useful concept to classify complex network[35]. In this study, we used a greedy algorithm to search for an optimal set of parameters that maximize removal of degenerated motifs.

Network comparison has been widely investigated and continues to present challenging problems. Milo[36] proposed the use of a significance profile of subgraphs embedded in a network, to identify superfamilies of evolved and designed networks. Though the null-hypothesis approach is useful, it has limitations[37]. We employed an information-theoretic approach which bypasses the limitations of the null-hypothesis approach.

The concept of entropy has been extensively applied in cancer biology studies. For instance, it was reported that cancer networks exhibit high information entropy[38], as well as increased network entropy[39] and signaling entropy[40]. Recently, Du et al.[41] introduces a decision coefficient to access the impacts of direct and indirect determination factors among different pathways.

Network similarity can be quantified by using the information-theoretic quantity, Kullback–Leibler entropy (H_{KL}). In 2011, Kugler et al.[42] introduced the use of graph

prototyping for network comparison. It was shown that three out of the five utilized graph distance measures, the group of prostate cancer networks is significantly different from the group of benign networks.

The present work utilizes the Jensen-Shannon entropy (H_{JS}), to study the similarity of two molecular networks by comparing the frequency distributions of the two networks' 3-node motifs and 4-node motifs. One of the major characteristics of H_{JS} is that this quantity is symmetric, it allows us to determine the distance between two networks without knowing the nodes' identities and network sizes.

In our previous work[43], we already laid down a foundation for the present study. In another recent work we extended the previous work by developed a motif finding algorithm, *PatternFinder*[44], to identify the 3-node motifs and 4-node motifs in cancer networks, signal transduction networks and cellular processes.

Results

Given the 3-node motifs and 4-node motifs, we used *PatternFinder* to identify their subgraphs (all possible functional motifs). For the 3-node motifs, it was found that motif 'id_6' (SIM), motif 'id_12' (cascade) and motif 'id_36' (MIM) are not composed of any 3-node functional motifs. For the 4-node motifs, there are eight motifs that are not composed of any 4-node functional motifs, i.e. motif 'id_14' (SIM), motif 'id_28', motif 'id_74', motif 'id_76' (MIM), motif 'id_280', motif 'id_328' (cascade), motif 'id_392' and motif 'id_2184'. These eight motifs exhibit the property of *irreducibility.* However, each one of the eight motifs are embedded with exactly one 3-node functional

motif. In other words, given the 4-node motifs, the *irreducible* property does not apply if we consider motifs composed of three nodes. Supplementary file 2 summarizes the functional motifs for 3-node motifs, 4-node motifs and 3-node motifs embedded in 4-node motifs, where integers '1'/'0' denotes the presence/absence of a functional motif, respectively.

Spectral graph theory, reciprocal links, complexity measures

Table 1 summarizes the results of the nine graph energies and edge information for the 3-node motifs. Firstly, since some of the matrices, such as, L and Q, are asymmetric, their eigenvalues are complex in general. In fact, among the 3-node motifs, motif 'id_98' has a pair of complex conjugate eigenvalues, and their associated eigenvectors are composed of complex components.

Secondly, we noted that graph energy is correlated with the number of edges of a motif. For instance, graph energies of fully connected 3-node motifs and 4-node motifs are maximal, despite having different energy definitions.

Thirdly, it is rather common for certain motifs to have the same graph energy; that is energy degenerated motifs are rather common. Two motifs are said to be *equi-energetic* if they have the same total energy. For instance, two pairs of motifs, i.e. 'id_6' and 'id_36' and 'id_14' and 'id_74', are *equi-energetic* regardless of the graph energy definition. The results of the 4-node motif graph energies and eigenvectors are given in Supplementary file 3.

Fourthly, although the results of the nine graph energies are rather similar, there are difference among them; i.e., the multiplicity of the energy levels is somewhat different. For the 3-node motifs, the multiplicity of graph energy *E,* i.e., 0, 2 and 2.83 are 4, 4 and 1 respectively. For *QE*, there are three energy values; i.e., 2.67, 4.00 and 5.33, that are associated with the multiplicity of 3, 3 and 3 respectively.

Fifthly, energy degenerated motifs may or may not have identical spectrums, *Sp(G)*. This suggests that the use of *Sp(G)* could allow for further distinction between the motifs. More details are given below in the 'Unique identifiers for network motifs' section.

In Supplementary file 1- Supplementary Table S6 summarizes the lower (E_{min}) and upper (E_{max}) graph energy bounds and ratios for the 3-node motifs and 4-node motifs. For the 3-node motifs, the ratios are bounded between 2 and 4.91. These ratios are a bit larger for 4-node motifs, i.e., they are bounded between 3.00 to 6.88. Later in our study, it was found that most of the molecular biological networks are not composed of motifs with large graph energies; therefore, the maximum ratio cannot be achieved. Details are reported below in the 'Network motifs absent from the network' section.

Reciprocity of motifs

Table 2 depicts the traditional reciprocity, *R*, reciprocity *r* and $\overline{a}$ for the 3-node motifs. Most of the *R* values are zero, which indicates that there is no edge pointing in both directions. Positive and negative values of *r* denote the presence of feedback loops. Among the 13 reciprocity values, nine of them are negative, meaning that the majority

of the 3-node motifs have either in or out links only. We note that motifs composed of one or two feedback loops can still have negative reciprocity values. The complete sets of R, r and $\overline{a}$ values of the 4-node motifs are given in Supplementary file 4.

Graph complexity - cyclomatic complexity and Kolmogorov complexity

For the 3-node motifs, Table 3 summarizes the results of the cyclomatic complexity (*CC*) and the Komologrov complexity (*KC*) and their rankings. The ranges of *CC* and *KC* values are 0~3 and 23.34~25.50, respectively. The complete sets of *CC* and *KC* values of the 4-node motifs are given in Supplementary file 5, where the ranges of *CC* and *KC* values are 0~8 and 33.80~43.74, respectively. These findings are compatible with the notation that motifs compose of more nodes have higher complexity.

A network motif with a large *CC* value implies a more complex decision structure. From Table 3, it is noted that *KC* can serve as a parameter in distinguishing motif patterns without any degeneracy. In other words, no two motifs have the same *KC*, which is also true for the 4-node motifs. Motif 'id_238' is a complete graph that is described by the binary string '011101110', and this string corresponds to lower algorithmic complexity (5th rank).

Next, we examined the correlations among the two complexity measures. We ranked *CC* and *KC* in ascending order and computed their Spearman Rank Correlation Coefficients (*SRCC)*. First, we noticed that the correlation is not perfect; this is because motif 'id_238' is associated with the largest *CC* values (ranks). However, this is not the case for *KC* (5th rank). It was found that *CC* and *KC* show a relative weak correlation,

i.e., 0.083 and 0.381, at the 3-node and 4-node levels respectively. It is known that *CC* and *KC* have different meanings. *CC* is a measure of the complexity of a motif's decision structure. It is the number of independent gene regulation paths. *KC* is an algorithmic measure which characterizes the randomness and compressibility of a bit string.

Lastly, we investigated whether graph energy is proportional with graph complexity. The results are listed in Supplementary file 1 – Supplementary Table S7. Compared to *CC*, *KC* has a much higher correlation with graph energy. *CC* exhibits relatively weak and modest correlations with the graph energy at the 3-node motif level and 4-node motif level, respectively. *KC* exhibits a modest correlation with all the graph energies at the 3-node and 4-node levels.

Supplementary file 1 – Supplementary Table S8 summarizes the results of strength of *SRC* (including minimum, maximum and ranges) between graph complexity and graph energy for 3-node motifs and 4-node motifs. Our results suggest that there are relative weak (3-node *CC*) and modest correlations (3-node motif *KC* and 4-node motif *KC*) between graph complexity and graph energy.

Unique identifiers for network motifs

This section reports the results of determining an optimal parameter combination that maximize removal of degenerated motifs. As shown in Table 4, three cases are considered. 'Case a' made use of the graph energy only, 'case b' utilized graph energy, *r* and *CC*, and 'case c' employed energy, *r*, *CC* and the eigenvalue spectrum. After

including r and CC, we can distinguish more motifs. The use of AL^t, r, CC and energy spectrum can fully distinguish the 3-node motifs. For 4-node motifs, the use of LL^t, QQ^t and LQ^t achieves the best level of distinguishability, i.e., 136 out of 199 (68.3%). In comparison to E, LE and QE, both *symmetric* and *asymmetric* generalized energies serve as superior measures for distinguishing different motif patterns.

Frequently found network motifs

Among the 17 cancer networks, 46 STN and nine cellular processes, there are 15 (88.2%), 40 (87.0%) and 7 (77.8%) networks, respectively, where more than 70% of the nodes are embedded in both 3-node motifs and 4-node motifs. Thus, motif-associated nodes account for a major portion of a network.

To determine frequently occurring motifs, we tabulated the frequency of occurrence of each motif pattern, then normalized the frequency to one, and computed the average normalized frequency (probability) across the studied networks. Table 5 summarizes the top seven most frequently found 3-node motifs and 4-node motifs in cancer networks.

For cancer networks, the top three most frequently found 3-node motifs are id_6, id_12 and id_36. By examining the top three motifs, we found the following common features; (i) they do not compose of any subgraphs (*irreducible*), (ii) they are composed of a minimal number of edges (n-1 edges for a N-node motif), (iii) the reciprocity r values are negative, i.e., -0.5 and those motifs have *no* feedback loops, (iv) they account

for at least 8% of the total number of motifs, and (v) they are associated with the lowest or the second lowest graph energies, regardless of the graph energy definition.

The 4th to 7th ranked motifs, i.e. 'id_38', 'id_74', 'id_14' and 'id_98', are composed of three edges. Motif 'id_38' is the so-called feed-forward loop (FFL) that does not consist of a feedback loop, whereas both 'id_74' and 'id_14' consist of feedback loops. The 7th ranked motif, 'id_98', is the so-called 3-cycle. Motif 'id_12' is a subgraph of 'id_38', 'id_74' and 'id_98'; whereas, 'id_36' (MIM) is a subgraph of FFL; and SIM is a subgraph of 'id_14'. In other words, the frequently found motifs are the simplest motifs, which are subgraphs of more complex motifs.

For 4-node motifs, except for feature (v), the above-mentioned features (i) to (iv) are also valid for the top seven most frequently found motifs. It is interesting to note that the *irreducible* and negative reciprocity value (that is -1/3) features hold at the 4-node level. Also, feature (v) holds if we consider graph energies, *E, LE* and *QE* but not the other six graph energy definitions. Furthermore, the above five features also hold for STN and cellular processes (see Supplementary file 1 - Supplementary Tables S9 and S10).

The ranges of the average probability for the top three 3-node motifs and top seven 4-node motifs are depicted Supplementary file 1 - Supplementary Tables S11. Other than the 4-node motifs of cellular processes, which are a bit small, the ranges of the average probability are rather similar. We noted that cellular processes may be different from

the other two families of networks, but it is not clear whether this is because a relatively small numbers of networks are available or not, i.e., nine networks.

In Table 6, we summarized the top three most frequently found motifs and the top seven motifs identified among different networks. Cancer networks and STNs exhibit very similar results, which suggests that the underlying architectures are highly similar. It indicates that molecular networks are composed of a *finite* number of motif patterns, i.e., around seven patterns, with an upper *graph energy limit*. We conjecture that other molecular networks, such as cell cycles, may demonstrate similar properties. We will examine this issue in a later work.

Next, we examined the association of frequently found motifs and complexity measures. From Table 6, we noted that frequently found motifs have a lower *KC* ranking (smaller *KC* value). A smaller *KC* value implies a lower degree of randomness, less information and higher compressibility. However, this observation is not exactly the same at the 4-node level, i.e., there are three instances where the ranks of *KC* are larger; for instance, the rank of *KC* is as high as 16, 24 and 33, for id_2184, id_280 and id_392, respectively. No obvious association exists between frequently found motifs and cyclomatic complexity measures.

In summary, according to the above findings, we hypothesize that there are underlying organizing principles that lead to network structures.

Network motifs absent from the network

We enumerated all possible 3-node motifs and 4-node motifs for the 17 cancer networks, 46 STN and nine cellular processes. In Supplementary file 1 - Supplementary Tables S12 summarizes the number of 3-node motifs and 4-node motifs with non-zero frequency of occurrence in cancer networks, STN and cellular processes. It is interesting to see that certain motifs are never present in the three families of networks, except for in two of the cellular processes. In fact, the 'adherens junction' network consists of a 3-node motif (id_110) composed of three genes: *CTNNA1, ACTB* and *AFDN*. Another network is 'Signaling pathways regulating pluripotency of stem cells (hsa04550)'. We have identified a fully connected 3-node motif (id_238) with three feedback loops connecting three genes: *Oct4, Sox2* and *Nanog*. It is well-known that with the *LIN28* genes, these four genes are liable to reprogram human somatic cells into pluripotent stem cells[45].

As long as there are motifs absent in the cancer networks, STN and cellular processes, there is a graph energy cutoff associated with these three families of networks. Supplementary file 1 - Supplementary Tables S13 depicts the results of the energy cutoffs, maximum graph energies and their ratios for the cancer networks. Among the nine graph energy definitions, the ratios may be as high as 0.750 and 0.667 for 3-node motifs and 4-node motifs, respectively. However, they can be as low as 0.536 ($LQ^{'}$ energy) and 0.481 ($AQ^{'}$ energy) for the 3-node motifs and 4-node motifs, respectively. The results of the graph energy cutoffs and ratios for the STN and cellular processes are given in Supplementary file 1 - Supplementary Tables S14 and S15. For cellular

processes, the cutoff ratio may be as high as 1.00 for the 3-node motifs, because we have identified a fully connected 3-node motif (id_238). At the 3-node level, two of the cellular processes exhibit peculiar network structures; this is an open issue which remains to be investigated.

Nevertheless, our results suggested that there is an existing energy cutoff or ratio which constrains the presence of certain motifs embedded in a molecular network. Also, the data indicate that the cutoff ratio for 3-node motifs is a bit higher than that for 4-node motifs. Furthermore, the motif probability distribution and the graph energy of a motif obeys an inverse relation; that is, the smaller the probability the higher the graph energy.

Charactering the frequency distributions of motifs using entropy

We utilized the entropy-based quantity, normalized Shannon entropy, H_n, to quantify the frequency distributions of the occurrence of motifs for the cancer networks. For randomized distribution, H achieves the maximal values, 3.700 [$\log_2$ (13)] and 7.637 [$\log_2$ (199)], for 3-node motifs and 4-node motifs, respectively. Table 7 lists the total number of 3-node motifs, N_3, normalized Shannon entropy, H_{3R}, for the 3-node motifs, total number of 4-node motifs, N_4, and normalized Shannon entropy, H_{4R}, for the 4-node motifs.

For all the cancer networks we studied, the frequency distributions of the motifs are not uniformly distributed among the motif patterns; hence, H_{3R} and H_{4R} are different

from one another. The results of N_3, N_4 H_{3R} and H_{4R} for STN and cellular processes are given in Supplementary file 1 - Supplementary Tables S16 and S17, respectively.

Supplementary file 1 - Supplementary Tables S18 depicts the ranges of H_{3R} and H_{4R} for cancer networks, STN and cellular processes. The ranges of H_{3R} and H_{4R} for cancer networks and cellular processes are rather similar, whereas STN has relative large ranges, i.e. 0.532 and 0.622. We also noted that the HIF-1 signaling pathway [hsa04066] has very small H_{3R} and H_{4R}. This is because the transcription factor, HIF-1β, functions as a master regulator of many genes, therefore, the SIM motif is the dominant motif at both 3-node and 4-node levels.

Network similarity - Jensen-Shannon entropy

Network similarity can be quantified by using the information-theoretic quantity, such as Jensen-Shannon entropy (H_{JS}). This quantity allows us to determine the distance between two frequency distributions of motifs regardless of the nodes' identities and network sizes. The advantage of this approach is that one does not have to consider the common genetic components shared between two networks. The comparison is purely based on the regulatory structure of the two networks. Table 8 list the top five most similar pairs of cancer networks based on the H_{JS} distance of the 3-node motif and 4-node motif distribution functions.

At the 3-node level, the acute myeloid leukemia (AML) network and the pancreatic cancer network is highly similar, because the H_{JS} distance is the smallest. That is, the difference between two networks' probability distributions is minimal (among the four

motifs, i.e. SIM, cascade, MIM and FFL). Chronic myeloid leukemia (CML) and gastric cancer are associated with the second smallest H_{JS} distance. These two networks have rather similar probability distributions for the following three motifs (FFL motif is not included): SIM, cascade and MIM. In other words, given two pairs of networks, they have different H_{JS} because of one of the following reasons: (i) different H_{JS} among the same set of motifs, or (ii) different H_{JS} among different set of motifs.

It has been reported that AML and pancreatic disease are both observed in the same patients in clinical diagnosis[46], pancreatic masses develop during or after AML[47], AML can rarely mimic the clinical picture of pancreatic cancer[48], and pancreatitis is a sign of appearance of AML[49].

For the second pair of networks, previous works have shown that (i) both CML and gastric cancer co-exist in a single patient[50], which may be due to decreased immunity[51], and (ii) medical doctors are recommended to pay attention to the association of CML and gastric cancer[52].

For the third pair of networks, it has been reported that (i) gastrointestinal metastases from lung cancer are diagnosed at a late stage and are thus life threatening[53], (ii) gastric cancer metastasis results from lung cancer[54], (iii) gastric cancer is associated with lung cancer[55-58], and (iv) both diseases can be successfully treated with carboplatin[59].

The results of the top five most similar H_{JS} distances for the 4-node case are shown in the bottom half of Table 8. Many works reported that, (i) both gastric cancer and hepatocellular carcinoma (HCC) can co-exist in the same patient[60-61], (ii) optimal

surgical strategies have been developed for treating synchronous gastric cancer and HCC[62-63] and (iii) it is necessary to closely follow up patients with gastric cancer or HCC for early diagnosis[64].

The difference between gastric cancer and HCC probability distribution is minimal among the following motifs, i.e. id_74, id_76 (MIM), id_280 and id_2184; whereas, the difference between CML and melanoma is minimal among a different set of motifs, i.e. id_14 (SIM), id_28, id_328 (cascade) and id_392.

Lists of the top five most similar H_{JS} distance for STN and cellular processes for the 3-node motifs and 4-node motifs are given in Supplementary file 1 - Supplementary Tables S19, S20, S21 and S22. Supplementary Table S23 summarizes the ranges of the H_{JS} distances of the 3-node motifs and 4-node motifs for cancer networks, STN and cellular processes. Relatively speaking, cellular processes have larger low end H_{JS} distances, i.e. 0.0697 and 0.125.

At the present moment, we do not know why pairs of networks with small H_{JS} interact; the exact mechanism is not clear and further work is needed to describe the explicit mechanism of the cross-talk behavior.

It is noted that the H_{JS} distance is non-zero, and the value may be as small as 0.0214. In other words, no two networks have exactly the same underlying motif frequency distributions. The H_{JS} distances of the 4-node motifs has a larger value than 3-node motifs, i.e. 0.804. The range of H_{JS} distances for 3-node motifs is large/small for

STN/cellular processes, i.e. 0.689/0.216. The reason is because we have collected a relatively large/small number of STN/cellular processes; i.e. 46/9.

Discussions and conclusions

Network motifs play an important role in biological networks. We made use of a rigorous mathematical and systematic approach, i.e. spectral graph theory, to characterize 3-node and 4-node network motifs. A total of nine graph energies were introduced to characterize network motifs. Also, we characterized network motifs' complexity by using two widely-accepted complexity measures, i.e. *CC* and *KC*.

The concept of a unique identifier was introduced to label network motifs. This novel idea combines four parameters; i.e. graph energy, reciprocal links, *CC* and eigenvalue spectrum, to characterize a network motif.

A foreseeable application of the identifier is to examine the transition between different motifs. It is possible that the regulatory interactions among the genetic elements may be disrupted or activated due to genetic mutation or epigenetic modification.

We conjecture that driver mutations are enriched or depleted in certain motif positions, such as, the source node. A source node is a node that has an out-degree only. In other words, a mutation driver gene acts as an upstreaming regulator. Several works have reported that certain motif positions, such as, the source nodes and target nodes, are enriched in cancer-associated genes[65-66].

Some of the 3-node motifs and 4-node motifs are interconnected through shared genetic elements. These types of modules are called coupled motif structures (CMS) in our previous work[52]. One can merge interconnected motifs to form higher order network structures.

We have extended a developed algorithm to identify complete sets of 3-node motifs and 4-node motifs for 17 cancer networks, 46 STN and nine cellular processes. Except for a few networks, 3-node motifs and 4-node motifs account for more than 70% of the nodes in the studied networks. Furthermore, this study discovered the following features: (i) the relative entropies of the motif distributions are not equal or close to one, indicating that the identified motifs are not distributed uniformly among the 13 and 199 patterns, (ii) our study suggested that biological networks are composed of a finite number of motif patterns; hence, graph energy cutoffs exist, (iii) irreducible motif patterns are the most frequently found motifs; for instance, the cascade pattern is the most frequently found motif, followed by the SIM and MIM motifs, (iv) all of the three families of networks exhibit the above features, which suggest that there is an universal organization principle determining the underlying network architecture, and (v) we utilized information-theoretic entropy (H_{JS}) to infer pairs of networks that exhibit similar regulatory interaction topologies.

In conclusion, this study provides a systematic and rigorous approach to dissecting the underlying structures of biological molecular networks. SGT serves as a powerful approach in distinguishing different motif topologies or connectivity patterns. Whether

or not there is acceptance of our hypothesis that network structures can be understood in terms of the 3-node motifs and 4-node motifs. The next step is to test our hypothesis by conducting the 5-node motifs analysis. We expect that our efforts may shed light on elucidating the complex nature of molecular networks.

Methods

Input datasets

Network information was obtained from the KEGG database (August 2017)[67]. Four families of networks were employed, including: (i) Environmental Information Processing, (ii) Cellular Processes, (iii) Organismal Systems and (iv) Human Cancers.

Not every network recorded by KEGG was imported. After inspection, we disregarded networks composed of several separate modules, such as, the 'Two-component system', 'MicroRNAs in cancer', 'Chemical carcinogenesis' and 'Viral carcinogenesis'. Also, we put together networks labeled with the name 'signaling pathway', and called them 'signal transduction networks (STN)'. We noted that STN ranges across different families in KEGG classification, such as, 'Signal transduction', 'Immune system' and 'Endocrine system'.

In total we have collected 17 cancer networks, 46 STN and nine cellular processes. We downloaded KEGG pathway KGML files, and then made use of the KEGGScape[68] and KEGGparser[69] packages to visualize and save the node and edge information for the given network.

In Supplementary File 1, Supplementary Table S1 summarizes the nodes, edges and motif-associated node information for the 17 cancer networks. The complete list of node and edge information of the 46 STN and nine cell processes can be found in Supplementary File 1 - Supplementary Table S2 and Supplementary Table S3, respectively.

Adjacency matrix

By analyzing the connectivity of each gene one constructs an adjacency matrix, ***A,*** to represent the interaction network. In general, there are 13 and 199 3-node and 4-node network motifs respectively[3,70].

It is possible that some motifs are subgraphs of other motifs (structural motifs). In a previous work[71], such subgraphs are called functional motifs. In a brain network, a structural motif and functional motif represent anatomical building blocks and the elementary processing mode of a network, respectively.

We have developed an algorithm named *PatterFinder* to enumerate all possible functional motifs embedded in the 3-node motifs and 4-node motifs. Details of *PatterFinder* are given in Supplementary File 1 – Supplementary Table S4.

Characterization of network motifs - motif graph energy, reciprocal links, graph complexity

The energy of a graph is an invariant[72-74], which is equal to the sum of the absolute value of the eigenvalues of ***A***. Originally, the concept of graph energy introduced by

Gutman was applied to study an undirected graph and the concept has been applied to estimate the π–electron energy of hydrocarbons.

The adjacency matrix ***A*** of a graph can be expressed in terms of its eigenvalues and eigenvectors and eigenvalues. Since ***A*** is a non-symmetric matrix in general, eigenvalues may be complex, and all of its eigenvectors are non-orthogonal. Let *n, e* and d_i denote the nodes, edges and the degree of the *i*th node of graph *G*, respectively, where *G* is called an (*n, e*)-graph. The energy of a graph, *E(G)*, is defined by

$$E(G) = \sum_{i=1}^{n} |\alpha_i| \tag{2}$$

where α_i denotes the *i*th eigenvalue of ***A***. It is also known that the sum of all the eigenvalues, α_i is equal to zero.

Assume that the graph energy eigenvalues are labeled in descending order, i.e. $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_n$, while the whole spectrum is denoted by $Sp(G) = [\alpha_1 , \alpha_2 , \ldots \alpha_n]$. The largest eigenvalue is referred to the spectral radius of graph *G*[75].

In spectral graph theory, there are two other matrices, i.e., Laplacian[76] and signless Laplacian[77-78], that can be defined to characterize the graphs. The Laplacian matrix ***L*** and signless Laplacian matrix ***Q*** of a graph *G* are defined as ***L*** = ***D*** – ***A***, and ***Q*** = ***D*** + ***A*** respectively; where ***D*** is the diagonal matrix where the diagonal elements are the vertex degrees. The Laplacian energy of graph *G*, *LE(G),* is defined by

$$LE(G) = \sum_{i=1}^{n} |\beta_i| - \frac{2e}{n} \tag{4}$$

where $|\beta_i|$ denotes the absolute value of the *i*th eigenvalues ***L***. There is an analogy between the properties of *E*(*G*) and *LE*(*G*), but some significant differences remain between these two quantities[79].

The signless Laplacian energy of graph *G*, *QE(G),* is defined by

$$QE(G) = \sum_{i=1}^{n} |\gamma_i| - \frac{2e}{n} \tag{5}$$

where $|\gamma_i|$ denotes the absolute value of the *i*th eigenvalues ***QE***.

A more general definition of graph energy was suggested by Nikiforov[80-81]. Let ***M*** be a *n* by *n* real matrix and the singular values are denoted by $s_1, s_2, \ldots s_n$, the total energy, ***ME,*** obtained from ***M*** is defined by

$$ME(G) = \sum_{i=1}^{n} |s_i| \tag{6}$$

The singular values of ***M*** are equal to the positive square roots of the eigenvalues of MM^t, where ***t*** denotes matrix transpose. Let ***M*** equals to ***A, L*** or ***Q*** and consider the eigenvalues of AA^t, LL^t and QQ^t, which are called generalized energy, ***ME(G)***. We extend the definition to consider MN^t, and therefore, three additional energies, i.e., AL^t, AQ^t and LQ^t, are computed; we call them *asymmetric* generalized energies. The sums of the absolute values of the eigenvalues of MM^t and M^tM are the same. The same holds for MN^t and NM^t. Therefore, one needs to compute MM^t and MN^t only. The advantages of using *asymmetric* generalized energies will be demonstrated later in this article. To the best of our knowledge, no or few previous works have made use of *asymmetric* generalized energies to characterize network motifs. In total, we have devised nine

graph energies to describe the motifs. We also note that Adiga et al.[82] proposed a parameter named skew energy, obtained from the skew-adjacency matrix, to characterize digraphs; however, the parameter does not apply to graphs consisting of multiple arcs.

Several studies[83-84] have suggested that reciprocal links in directed networks play an important role in dynamical processes and network growth. The traditional definition of reciprocity is $R = L^{\leftrightarrow}/L$, where $L^{\leftrightarrow}$ and L denote the number of edges pointing in both directions and the total number of edges respectively. This definition of reciprocity was modified by Garlaschelli and Loffredo[83], who defined reciprocity as the correlation coefficient between the entries of the adjacency matrix, ***A***, of a directed graph with ***N*** nodes. The reciprocity, *r*, is given by

$$r = \frac{\sum_{i \neq j}(a_{ij} - \bar{a})(a_{ji} - \bar{a})}{\sum_{i \neq j}(a_{ij} - \bar{a})^2} \tag{8}$$

where a_{ij} equals one if there is a link from node *i* to node *j*; and the average, $\bar{a}$, is defined by,

$$\bar{a} = \frac{\sum_{i \neq j} a_{ij}}{N(N-1)} \tag{9}$$

A positive value of *r* indicates that the motif has bi-directional connections, whereas negative *r* implies the motif has either an in or out connection.

To further understand the connectivity structure of the network motifs, we seek metrics that can be used to measure graph complexity. In software engineering, the cyclomatic complexity (*CC*) is a metric developed by McCabe[85] to measure the

complexity of a program by using the control flow graph. *CC* is defined by the expression $CC = e - N + 2P$, where e and N denote the number of edges and number of nodes of the graph, and P denotes the number of predicate/exit nodes[32,85]. Node and edge denote a program unit and the execution order of the program. *CC* depends only on the global decision structure (the number of edges and nodes) of a program.

In addition to *CC*, we utilized the algorithmic complexity measure, the Kolmogorov complexity (*KC*), to characterize graph complexity. Essentially, *KC* of a bit string is given by the smallest computer program that can generate the string. Given the adjacency matrix (or the equivalent bit string), we used the 'Block Decomposition Method (BDM)'[86] to determine the *KC* for both 3-node[87] and 4-node motifs. Bit string with high *KC* implies that it has a higher degree of randomness, contains more information and it is less compressible. The complete graph has a smaller *KC* value, whereas the random graph has higher *KC*, which is less compressible.

Unique identifiers for network motifs

Every 3-node motif and 4-node motif has a different *KC* value, so *KC* can be used as a unique identifier. However, given the graph energy, asymmetric graph energies, graph energy spectrum, reciprocal links, and *CC*, we seek to determine a minimal set of parameters that serve as a label of the network motifs. This set of parameters describes certain aspects of the motifs differently than the algorithmic complexity measure. Although this set of parameters does not uniquely label each motif, that is, some motifs share the same parameters, those parameters conserve or degenerate quantities which

may suggest useful implications in network motif analysis. To the best of our knowledge, the concept of introducing the use of energy, reciprocal links and *CC* to labeling network motifs is a novel idea. The pseudocode for determining the minimal set is based on greedy strategy are given in Supplementary File 1 – Supplementary Table S5.

Frequently found motifs, network entropy and network similarity

Given a molecular network, *PatternFinder* identifies both sets of 3-node motifs and 4-node motifs. Two motifs could have the same ID may partially embed with the same genetic element(s). These two motifs were counted twice in our calculation. We expect that certain motif patterns which occur with higher probabilities are the dominant underlying network structure. Let $p_3^{(k)}$ denotes the frequency (probability) distribution of a 3-node network motif, where k denotes one of the 13 patterns. The Shannon entropy for three-node motifs and 4-node motifs H_3 and H_4, of a molecular network are computed. The normalized Shannon entropies for the 3-node motifs and 4-node motifs are given by $H_{3R} = H_3 / \log_2(13)$ and $H_{4R} = H_4 / \log_2(199)$ respectively.

In information theory, a number of quantities can be used to characterize the distance between two probability distributions. For instance, one can use the relative entropy, also known as Kullback–Leibler entropy (H_{KL}), and the cross-entropy[88]. Given two discrete probability distributions, X and Y, the relative entropy of X with respect to Y is defined by

$$H_{KL}(X \| Y) = -\sum_i X_i \log \frac{X_i}{Y_i} \tag{11}$$

H_{KL} is asymmetric under the interchange of X and Y. One can symmetrize H_{KL} by adding the term $H_{KL}(Y||X)$.

We utilized a similar quantity, named the Jensen–Shannon entropy (H_{JS}) to gauge network similarity. H_{JS} has been applied in many applications, including: (i) predicting functionally important amino acids from sequence conservation[88], (ii) pattern recognition in bioinformatics[89], (iii) predicting important non-conserved residues in protein sequences[90], (iv) analyzing DNA sequences[91] and (v) measuring the distance between random graphs[92]. Wong & You proposed a distance metric between two random graphs based on the smallest change of Shannon entropy before and after merging the two random graphs.

H_{JS} is a symmetric function, it was used to measure the distance between the frequency distributions of the 3-node motifs (or 4-node motifs) X and Y in network N_X and network N_Y. H_{JS} is defined by

$$\boldsymbol{H_{JS} = \frac{1}{2}[H_{KL}(X||Z) + H_{KL}(Y||Z)]} \tag{14}$$

where $Z = \frac{1}{2}(X + Y)$.

The unique feature of H_{JS} is that it measures the similarity between two networks in terms of the networks' underlying architecture rather than the nodes' identities. In other words, without referring to the genetic identities of a motif, network similarity is measured. The square root of H_{JS} is a metric called Jensen-Shannon distance[93]. To the

best our knowledge, the present study is the first to make use of H_{JS} to address the molecular network similarity problem.

Acknowledgement

Dr. Chien-Hung Huang and Mr. Alice, Hsieh work are supported by the Ministry of Science and Technology (MOST) under the grants of MOST 106-2221-E-150-056 and 107-2221-E-150-038. Dr. Ka-Lok Ng work is supported by the grants of MOST 106-2221-E-468-017, MOST 107-2632-E-468-002, and also supported under the grant from Asia University, 106-asia-06. Dr. Jeffrey J. P. Tsai work is supported by the grant of MOST 107-2632-E-468-002. We would like to thank He-Xing Li, Ci-Jun Peng and I-Lun Hsieh, who spent efforts on developing the codes.

Conflict of interest

None.

Authors' Contributions

Chien-Hung Huang conducted the algorithm development, review and edited the manuscript. Jeffrey J. P. Tsai provided interpretation of the results and review the manuscript. Nilubon Kurubanjerdjit conducted the complexity analysis, literature search and participated in discussion. Ka-Lok Ng is the corresponding author, who designed the study, review and drafted the manuscript. All authors read and approved the manuscript.

Availability of Data and Material

All the network information are publicly available, which were obtained from the KEGG database, see https://www.genome.jp/kegg/pathway.html.

References

1. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. Nature. 402(6761 Suppl), C47-52 (1999).
2. Lauffenburger, D. A. Cell signaling pathways as control modules: complexity for simplicity? Proc. Natl. Acad. Sci. USA. 97, 5031-5033 (2000).
3. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli.* Nature Genetics. 31, 64 (2002).

4. Milo, R. et al. Network motifs: simple building blocks of complex networks. Science. 298, 824-827 (2002).
5. Schreiber, F. & Schwobbermeyer, H. MAVISTO: a tool for the exploration of network motifs. Bioinformatics. 21, 3572-3574 (2005).
6. Wernicke, S. & Rasche, F. FANMOD: a tool for fast network motif detection. Bioinformatics. 22, 1152-1153 (2006).
7. Ferro, A. et al. NetMatch: a Cytoscape plugin for searching biological networks. Bioinformatics. 23, 910-912 (2007).
8. Ma'ayan, A. Insights into the organization of biochemical regulatory networks using graph theory analyses. The Journal of biological chemistry. 284, 5451-5455 (2009).
9. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. Nature. 407(6804), 651-654 (2000).
10. Lee, S. H., Kim, P. J. & Jeong, H. Global organization of protein complexome in the yeast Saccharomyces cerevisiae, BMC Syst. Biol. 5,126 (2011).
11. Lee, P. H., Huang, C. H., Fang, J. F., Liu, H. C. & Ng, K. L. Hierarchical and topological study of the protein-protein interaction networks. Advances in Complex Systems. 8(4), 383-397 (2005).
12. Bonacich, P. & Lloyd, P. Eigenvector-like measures of centrality for asymmetric relations. Social Networks. 23, 191-201 (2001).
13. Newman, M. E. J. The Mathematics of networks, The New Palgrave Encyclopedia of Economics, 2nd edition, L. E. Blume and S. N. Durlauf (eds.), Palgrave Macmillan, Basingstoke (2008).
14. Pavlopoulos, G. A. et al. Using graph theory to analyze biological networks. BioData Min. 4, 10 (2010).
15. Konganti, K., Wang, G., Yang, E., & Cai, J. J. SBEToolbox: a MATLAB toolbox for biological network analysis. Evol. Bioinform. Online. 9, 355-362 (2013).
16. Bloch, F., Jackson, M. O., & Tebaldi, P. Centrality Measures in Networks. arXiv:1608.05845v3 [physics.soc-ph] (2016).
17. Albert, R., Jeong, H. & Barabasi, A. L. Error and attack tolerance of complex networks. Nature. 406 (6794), 378–382 (2000).
18. Huang, C. H., Chen, T. H. & Ng, K. L. Graph theory and stability analysis of protein complex interaction networks. IET Systems Biology. 10(2), 64-75 (2016).
19. Dancoff, S. M. & Quastler, H. Information content and error rate of living things. In: Essays on the Use of Information Theory in Biology, eds. H. Quastler, 263–274. Urbana: University of Illinois Press (1953).
20. Rashevsky, N. Topology and life: in search of general mathematical principles in biology and sociology. Bull. Math. Biophys. 16(4), 317–348 (1954).
21. Rashevsky, N. Life, information theory, and topology. Bull. Math. Biophys. 17, 229 (1955).
22. Mowshowitz, A. Entropy and the complexity of the graphs I: An index of the relative complexity of a graph. Bull. Math. Biophys. 30, 175–204 (1968).
23. Harary, F. & Palmer, E. M. Graphical Enumeration. Academic Press, NY (1973).
24. Sloane, N. J. A. & Plouffe, S. The Encyclopedia of Integer Sequences. Academic Press, San Diego (1995).
25. Kim, W., Diko, M. & Rawson, K. Network motif detection: algorithms, parallel and cloud computing, and related tools. Tsinghua science and technology. 18(5), 469-489 (2013).
26. Konagurthu, A. S. & Lesk, A. M. On the origin of distribution patterns of motifs in biological networks. BMC System Biology. 2(73), 1-8 (2008).

27. Adami, C., Qian, J., Rupp, M. & Hintze, A. Information content of colored motifs in complex networks. Artificial Life. 17(4), 375-390 (2011).
28. Cvetkovic D. & Simic, S. Graph spectra in computer science. Linear Algebra and its Applications. 434(6),1545-1562 (2011).
29. Banerjee, A. & Jost, J. Graph spectra as a systematic tool in computational biology. Discrete Applied Mathematics. 157(10), 2425-2431 (2009).
30. Brualdi, R. A. Spectra of digraphs. Linear Algebra and its Applications. 432(9), 2181-2213 (2010).
31. Mowshowitz, A. The characteristic polynomial of a graph. Journal of Combinatorial Theory, Series B, 12(2), 177-193 (1972).
32. Weyuker, E. J. Evaluating software complexity measures. IEEE transactions on Software Engineering. 14(9), 1357-1365 (1988).
33. Mowshowitz, A. Entropy and the complexity of the graphs II. The information content of digraphs and infinite graphs. Bull. Math. Biophys. 30, 225–240 (1968).
34. Minoli, D. Combinatorial graph complexity. Atti. Accad. Naz. Lincei, VIII. Ser., Rend., Cl. Sci. Fis. Mat. Nat. 59, 651–661 (1975).
35. Mueller, L. A., Kugler, K. G., Dander, A., Graber, A. & Dehmer, M. QuACN: an R package for analyzing complex biological networks quantitatively. Bioinformatics, 27(1), 140-1 (2011).
36. Milo, R. et al. Superfamilies of designed and evolved networks. Science. 303(5663), 1538-1542 (2004).
37. Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N. & Stone, L. Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks. Science. 305(5687), 1107 (2004).
38. Schramm, G, Kannabiran, N. & König, R. Regulation patterns in signaling networks of cancer. BMC Systems Biology. 4, 162 (2010).
39. West, J., Bianconi, G., Severini, S. & Teschendorff, A. E. Differential network entropy reveals cancer system hallmarks. Scientific Report. 2, 802 (2012).
40. Teschendorff, A. E., Banerji, C. R., Severini, S., Kuehn, R. & Sollich, P. Increased signaling entropy in cancer requires the scale-free property of protein interaction networks. Scientific Report. 5, 9646 (2015).
41. Du, J. et al. A decision analysis model for KEGG pathway analysis. BMC Bioinformatics. 17(1), 407 (2016).
42. Kugler, K. G., Mueller, L. A., Graber, A. & Dehmer, M. Integrative network biology: graph prototyping for co-expression cancer networks. PLoS One, 6(7), e22843 (2011).
43. Hsieh, W. T. et al. Transcription factor and microRNA-regulated network motifs for cancer and signal transduction networks. BMC Systems Biology. 9(1), S5 (2015).
44. Li, H. X. In silico study of significant network motifs in the cancer networks. Master Thesis, National Formosa University, Taiwan (2016). Advisor: Huang, C. H., co-advisor: Ng, K. L.
45. Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. Science. 318(5858), 1917-1920 (2007).
46. Cascetta, K, et al. A 60-year-old male with synchronous acute myeloid leukemia and metastatic adenocarcinoma of the pancreas. Semin. Oncol. 41(6), e51-59 (2014).
47. Messager, M., Amielh, D., Chevallier, C. & Mariette, C. Isolated granulocytic sarcoma of the pancreas: A tricky diagnostic for primary pancreatic extramedullary acute myeloid leukemia. World Journal of Surgical Oncology. 10, 13 (2012).

48. Schäfer, H. S., Becker, H., Schmitt-Gräff, A. & Lübbert, M. Granulocytic sarcoma of Core-binding Factor (CBF) acute myeloid leukemia mimicking pancreatic cancer. Leuk Res. 32(9):1472-1475 (2008).
49. de Castro, J. A., Vencer, L. & Espinosa W. Acute Myelogenous Leukemia Presenting as Acute Pancreatitis: A Case of Primary Pancreatic Extramedullary Acute Myeloid Leukemia. Clinical Gastroenterology and Hepatology. 15(1), e30-e31 (2017).
50. Butala, A., Kalra, J. & Rosner, F. Chronic myelocytic leukemia and gastric cancer in the same patient. Journal of the National Medical Association. 81(4), 457-459 (1989).
51. Mangal, M., Sudharsanan, S., Elamurugan, T. & Jagdish, S. Gastric carcinoma in a patient with chronic lymphocytic leukemia: coincidence or consequence? Cureus. 10(4), e2405 (2018).
52. Mokhtarifard, A., Kooshyar, M. M., Nassiri, M. R., Mohammadipour, A. & Edalatianalipour Z. Gastric cancer and CML: a literature review and case report. World Cancer Research Journal. 3(2), e709 (2016).
53. Li, X., et al. Multiple gastrointestinal metastases of squamous-cell lung cancer: A case report. Medicine. 9(24), e11027 (2018).
54. Gao, S. et al. Gastric metastasis from small cell lung cancer: a case report. World J Gastroenterol. 21(5), 1684-1688 (2015).
55. Nitipir C, et al. A rare case of advanced lung cancer presenting as a symptomatic gastric tumor. Mol Clin Oncol. 8(4), 600-602 (2018).
56. Koh, S. A. & Lee, K. H. Adenocarcinoma of lung cancer with solitary metastasis to the stomach. Korean J Gastroenterol. 64(3):154-157 (2014).
57. Snyder, E. L. et al. Nkx2-1 represses a latent gastric differentiation program in lung adenocarcinoma. Mol. Cell. 50(2), 185-99 (2013).
58. Park, M. K. et al.. A Case of Early Gastric Cancer Associated with Small Cell Lung Cancer. Journal of the Korean Cancer Association. 30(2), 414-420 (1998).
59. Sano, T. et al. A case report of synchronous small cell lung cancer and gastric cancer successfully treated with carboplatin. Jpn. J. Cancer Res. 77(8),790-798 (1986).
60. Hu, Z. R. et al. Synchronous Gastric Cancer and Hepatocellular Carcinoma. Journal of Cancer Research and Practice. 1(3), 226-232 (2014).
61. Sakumura, M., Tajiri, K. & Sugiyama, T. Gastric metastasis of hepatocellular carcinoma mimicking early gastric cancer. Clin Gastroenterol Hepatol. 16(10), e99-e100 (2018).
62. Uenishi, T. et al. Surgical management of synchronous hepatocellular carcinoma and gastric cancer. Dig. Surg. 20(2), 133-140 (2003).
63. Tawada, A. et al. Intracranial metastasis in a patient with hepatocellular carcinoma and gastric cancer. Case Rep Oncol. 7, 199-203 (2014).
64. Chen, P. D., Chen, C. N., Hu, R. H. & Lai, H. S. Clinical experience of double primary gastric cancer and hepatocellular carcinoma. Formos. J. Surg. 50, 10-15 (2017).
65. Awan, A. et al. Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network. IET Syst. Biol. 1(5), 292-297 (2007).
66. Carson, M. B., Gu, J., Yu, G. & Lu, H. Identification of cancer-related genes and motifs in the human gene regulatory network. IET Syst. Biol. 9(4), 128-134 (2015).
67. Nakaya, A. et al. KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. Nucleic Acids Res. 41, D353-357 (2013).
68. Nishida, K., Ono, K., Kanaya, S. & Takahashi, K. KEGGscape: a Cytoscape app for pathway data integration. F1000Research. 3, 144 (2014).

69. Arakelyan, A. & Nersisyan, L. KEGGParser: parsing and editing KEGG pathway maps in Matlab. Bioinformatics. 29(4), 518-519 (2013).
70. Alon, U. An introduction to systems biology: design principles of biological circuits. Boca Raton, FL: Chapman & Hall/CRC (2007).
71. Sporns, O. & Kotter, R. Motifs in brain networks. PLoS Biol. 2(11), e369 (2004).
72. Gutman, I. The Energy of a Graph. Ber. Math-Statist. Sekt. Forschungsz. Graz, 103, 1-22 (1978).
73. Gutman, I. Total π–electron energy of benzenoid hydrocarbons. Topics Curr. Chem. 162, 29-63 (1992).
74. Li, X. L., Shi, Y. T. & Gutman, I. Graph Energy. New York: Springer Verlag (2012).
75. Cvetkovic, D., Doob, M. & Sachs, H. Spectra of Graphs: Theory and Applications. 3rd edition, Huthig Pub Ltd (1997).
76. Chung, F. R. K. Spectral Graph Theory. American Mathematical Soc., USA (1997).
77. Cvetkovic D. & Simic, S. K. Towards a spectral theory of graphs based on signless Laplacian I. PUBLICATIONS DE L'INSTITUT MATHÉMATIQUE, Nouvelle série, tome. 85(99), 19-33, DOI:10.2298/PIM0999019C (2009).
78. Cvetkovic D. & Simic S. Towards a spectral theory of graphs based on signless Laplacian II. Linear Algebra and its Applications. 432, 2257-2272 (2010).
79. Gutman, I. & Zhou, B. Laplacian energy of a graph. Linear Algebra and its Applications. 414, 29-37 (2006).
80. Nikiforov, V. The energy of graphs and matrices. J. Math. Anal. Appl. 326, 1472-1475 (2007)
81. Nikiforov, V. Beyond graph energy: Norms of graphs and matrices. Linear Algebra and its Applications. 506, 82-138 (2016).
82. Adiga, C., Balakrishnan, R. & So, W. The skew energy of a digraph. Linear Algebra Appl. 432,1825-1835 (2010).
83. Garlaschelli, D. & Loffredo, M. I. Patterns of Link Reciprocity in Directed Networks. Phys. Rev. Lett. 93, 268701 (2004).
84. Squartini, T., Picciolo, F., Ruzzenenti, F. & Garlaschelli, D. Reciprocity of weighted networks. Scientific Reports. 3, 2729 (2013).
85. McCabe, T. J. A Complexity Measure. IEEE Transactions on Software Engineering. 4, 308-320 (1976).
86. Soler-Toscano, F., Zenil, H., Delahaye, J. P. & Gauvrit, N. Calculating kolmogorov complexity from the output frequency distributions of small turing machines. PLoS ONE. 9(5), e96223 (2014).
87. Zenil, H., Kiani, N. A. & Tegnér, J. Methods of information theory and algorithmic complexity for network biology. Seminars in Cell and Developmental Biology. 51, 32-43 (2016). [an earlier version appeared in arXiv:1401.3604 [q-bio.MN, 15 Oct 2014].
88. Capra J. A. & Singh, M. Predicting functionally important residues from sequence conservation. Bioinformatics. 23(15), 1875-1882 (2007).
89. Loog, M., Wessels, L., Reinders, M. J. T. & de Ridder, D. Pattern Recognition in Bioinformatics: 6th IAPR International Conference PRIB 2011, Springer-Verlag Berlin Heidelberg (2011).
90. Gultas, M. et al. Quantum coupled mutation finder: predicting functionally or structurally important sites in proteins using quantum Jensen-Shannon divergence and CUDA programming. BMC Bioinformatics. 15, 96 (2014).
91. Grosse, I. et al. Analysis of symbolic sequences using the Jensen-Shannon divergence. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 65(4 Pt 1), 041905 (2002).

92. Wong, A. K. &You, M. Entropy and distance of random graphs with application to structural pattern recognition. IEEE Trans. Pattern Anal. Mach. Intell. 7(5), 599-609 (1985).
93. Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. IEEE Trans. Inf. Theory. 49(7), 1858-1860 (2003).

Table 1. The results of the nine graph energies and edge information for the 3-node motifs

ID	*E*	*LE*	*QE*	AA^{+}	LL^{+}	QQ^{+}	AL^{+}	AQ^{+}	LQ^{+}	*e*
6	0.00	2.67	2.67	1.41	4.32	4.32	1.41	1.41	3.83	***2***
12	0.00	2.67	2.67	2.00	4.34	4.34	2.00	2.00	3.93	***2***
14	2.00	4.00	4.00	2.41	6.13	6.13	3.00	3.00	5.45	3
36	0.00	2.67	2.67	1.41	4.32	4.32	1.41	1.41	3.83	***2***
38	0.00	4.00	4.00	2.24	6.39	6.34	2.63	2.37	6.01	3
46	2.00	5.33	5.33	2.73	8.24	8.16	2.00	3.86	7.59	***4***
74	2.00	4.00	4.00	2.41	6.13	6.13	3.00	3.00	5.45	3
78	2.83	5.33	5.33	2.83	8.00	8.00	3.86	3.86	7.29	4
98	3.00	4.29	4.46	3.00	6.29	6.46	4.25	4.36	6.01	3
102	3.06	5.33	5.56	3.24	8.17	8.25	4.76	5.02	7.57	4
108	2.00	5.33	5.33	2.73	8.24	8.16	2.00	3.86	7.59	***4***
110	3.24	6.67	6.72	3.49	10.09	10.09	5.38	5.70	9.40	5
238	4.00	8.00	8.00	4.00	12.00	12.00	6.47	6.93	11.21	6

Table 2. The results of the traditional reciprocity (*R*), reciprocity (*r*) and the average reciprocity ($\overline{a}$) of the 3-node motifs.

ID	*R*	*r*	*e*	$\overline{a}$
6	0	-0.5	2	1/3
12	0	-0.5	2	1/3
14	0	1/3	3	0.5
36	0	-0.5	2	1/3
38	0	-1	3	0.5
46	0	-0.5	4	2/3
74	0	1/3	3	0.5
78	1	1	4	2/3
98	0	-1	3	0.5
102	0	-0.5	4	2/3
108	0	-0.5	4	2/3
110	0	-0.2	5	5/6
238	1	1	6	1

Table 3. The results of the cyclomatic complexity (*CC*) and Komologrov complexity (*KC*) and their ranking for the 3-node motifs.

ID	*CC*	*KC*	*Rank of CC*	*Rank of KC*
6	3	23.34	11	1
12	1	23.83	3	3
14	2	24.30	8	6
36	1	23.55	3	2
38	2	24.87	8	8
46	3	25.50	11	13
74	0	23.85	1	4
78	1	25.00	3	9
98	0	24.82	1	7
102	1	25.01	3	10

108	1	25.11	3	11
110	2	25.25	8	12
238	3	24.14	11	5

Table 4. The number of distinguishable motifs using optimal parameter combination of graph energy, *r, CC* and energy spectrum.

	3-node motifs			4-node motifs		
	case a	case b	case c	case a	case b	case c
E	7	11	11	42	57	60
LE	6	10	11	35	51	96
QE	9	11	11	51	67	72
AA^t	10	12	12	74	86	92
LL^t	10	12	12	94	103	136
QQ^t	10	12	12	88	96	136
AL^t	10	12	13	117	128	130
AQ^t	10	12	12	120	129	131
LQ^t	10	12	12	109	117	136

'case a' uses graph energy only, 'case b' uses graph energy, *r* and *CC*, and 'case c' uses energy, *r*, *CC* and graph energy spectrum

Table 5. The top seven most frequently found 3-node motifs and 4-node motifs in cancer networks.

	ID	average probability	*r*	Name, embedded motif ID
3-node				
1	6	0.421	-1/2	SIM
2	12	0.414	-1/2	Cascade
3	36	0.152	-1/2	MIM
4	38	0.0091	-1	FFL, 12, 36
5	74	0.0022	1/3	12
6	14	0.0016	1/3	6
7	98	0.00092	-1	3-cycle, 12
4-node				
1	14	0.224	-1/3	SIM
2	328	0.158	-1/3	Cascade
3	28	0.148	-1/3	-
4	74	0.137	-1/3	-
5	76	0.100	-1/3	MIM
6	392	0.099	-1/3	-
7	280	0.0864	-1/3	-

SIM denotes simple input module, MIM denotes multiple input module, FFL denotes feed-forward loop

Table 6. Comparison of frequently found motifs identified in cancer networks, STN and cellular processes

	Cancer networks	STN	Cellular processes
	Top three most frequently found motifs		
3-node ID	6, 12, 36	12, 6, 36	12, 36, 6
Rank of *KC*	1, 3, 2	3, 1, 2	3, 2, 1
	Top seven most frequently found motifs		
4-node ID	14, 328, 28, 74, 76, 392, 280	14, 28, 74, 328, 280, 76, 392	392, 328, 76, 280, 2184, 74, 28
Rank of *KC*	1, 7, 3, 5, 6, 33, 24	1, 3, 5, 24, 7, 24, 6, 33	33, 7, 6, 24, 7, 16, 5, 3

Table 7. The results of N_3, N_4, H_{3R} and H_{4R} for cancer networks.

Cancer networks	N_3	H_{3R}	N_4	H_{4R}
Acute_myeloid_leukemia_[hsa05221]	160	0.428	577	0.385
Basal_cell_carcinoma_[hsa05217]	34	***0.323***	51	***0.273***
Breast_cancer_[hsa05224]	145	0.449	445	0.437
Choline_metabolism_in_cancer_[hsa05231]	70	0.465	193	0.406
Chronic_myeloid_leukemia_[hsa05220]	71	0.368	145	0.323
Colorectal_cancer_[hsa05210]	71	0.467	124	0.403
Endometrial_cancer_[hsa05213]	45	0.329	60	0.295
Gastric_cancer_[hsa05226]	87	0.355	153	0.344
Glioma_[hsa05214]	80	0.410	183	0.390
Hepatocellular_carcinoma_[hsa05225]	65	0.355	80	0.350
Melanoma_[hsa05218]	46	0.374	82	0.337
Non-small_cell_lung_cancer_[hsa05223]	103	***0.493***	284	***0.483***
Pancreatic_cancer_[hsa05212]	74	0.422	131	0.371
Pathways_in_cancer_[hsa05200]	640	0.473	2795	0.450
Prostate_cancer_[hsa05215]	102	0.372	357	0.335
Renal_cell_carcinoma_[hsa05211]	58	0.385	114	0.353
Small_cell_lung_cancer_[hsa05222]	61	0.362	96	0.318

Table 8. The top five most similar pairs of cancer networks based on the H_{JS} distance of the 3-node motifs and 4-node motifs.

Network A	Network B	H_{JS} distance
3-node motifs		
Acute myeloid leukemia	Pancreatic cancer	0.0214
Chronic myeloid leukemia	Gastric cancer	0.0309
Gastric cancer	Small cell lung cancer	0.0339
Hepatocellular carcinoma	Melanoma	0.0411
Chronic myeloid leukemia	Hepatocellular carcinoma	0.0422
4-node motifs		
Gastric cancer	Hepatocellular carcinoma	0.0998
Chronic myeloid leukemia	Melanoma	0.138
Hepatocellular carcinoma	Small cell lung cancer	0.138
Basal cell carcinoma	Small cell lung cancer	0.150
Breast cancer	Pathways in cancer	0.151