Lecture 12

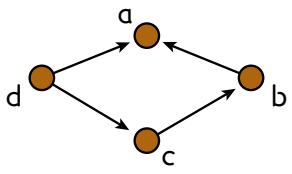
Network Motifs in Gene Regulation

Reading: Here we continue our search for the building-blocks of genetic regulatory networks. Our main focus will be the concept of a network "motif" developed by Uri Alon and his collaborators in a series of papers from the early noughties:

> S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon (2002), Network motifs in the transcriptional regulation network of *Escherichia coli*, *Nature Genetics*, **31**:64–68. DOI: 10.1038/ng881
>
> R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon (2002), Network motifs: simple building blocks of complex networks, *Science*, **298**:824–827. DOI: 10.1126/science.298.5594.824
>
> S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon (2003), Subgraphs in random networks, *Physical Review E*, **68**:026127.
> DOI: 10.1103/PhysRevE.68.026127
>
> S. Mangan and U. Alon (2003), Structure and function of the feed-forward loop network motif, *PNAS*, **100**:11980–11985.
> DOI: 10.1073/pnas.2133841100
>
> S. Mangan, A. Zaslaver, and U. Alon (2003), The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks, *J. Molecular Biology*, **334**:197–204. DOI: 10.1016/j.jmb.2003.09.049
>
> S. Mangan, S. Itzkovitz, A. Zaslaver, and U. Alon (2006), The incoherent feed-forward loop accelerates the response-time of the *gal* system of *Escherichia coli*, *J. Molecular Biology*, **356**:1073–1081.
> DOI: 10.1016/j.jmb.2005.12.003

Alon also has a book that covers some of this material, as well as many other interesting topics

> Uri Alon (2007), *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman & Hall/CRC.



v	$\deg_{in}(v)$	$\deg_{out}(v)$
a	2	0
b	1	1
c	1	1
d	0	2

Figure 12.1: *The in- and out-degrees of the nodes in a small network.*

12.1 Motifs in regulatory networks

In Lecture 11 I mentioned that various strains of domesticated wheat seem to have evolved by duplicating their entire genome and speculated whether similar duplication may have happened on a smaller scale. To address this question, Uri Alon and his collaborators went looking for small sub-networks of the *E. coli* regulatory network[1] that, in a sense they managed to make precise, occur more often than one would expect. Their approach is similar to the reasoning that underpins hypothesis-testing in statistics. They first argued that mutation and genetic recombination—normal processes that are continually, though very slowly, modifying all genomes—can lead to "rewiring" of regulatory networks. For example, mutations may create or destroy the DNA sub-sequences that transcription factors recognise and so create or destroy regulatory interactions.

One can then consider the family of all possible regulatory networks and ask whether a particular sub-network appears more often than one would expect. If so, one might take that as evidence that the sub-network has been selected by evolution. I'll make these ideas concrete with two examples in Section 12.3, but first I'd like to cover a few graph-theoretic preliminaries that will simplify and clarify our analyses.

12.2 A mathematical interlude

Here I'll briefly introduce some mathematical ideas and computational tools that will help us do hypothesis testing in real networks. It will prove helpful to have the following definitions

Definition 12.1. *An arc in a regulatory network is an* ***autoregulatory loop*** *(or* ***loop*** *for short) if it connects a gene to itself.*

Definition 12.2. *Two directed arcs in a regulatory network are* ***parallel*** *if they have the same tip and tail nodes.*

Definition 12.3. *The* ***out-degree*** *of a node v is the number of arcs of the form $v \to u$ that appear in the network. Similarly, the* ***in-degree*** *of v counts the number of arcs of the form $u \to v$. Note that as we permit our networks to have loops, we should include the possibility that $u = v$. Figure 12.1 illustrates these terms.*

[1] They looked at lots of other kinds of networks too, including some in which the nodes represent species and the arcs represent predator-prey interactions: see the 2002 paper by Milo *et al.* in *Science*.

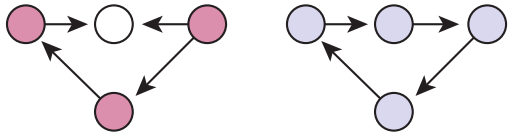


Figure 12.2: *The network at the left is weakly connected, but not strongly connected, because any path from the white node to one of the others involves travelling "backwards" along one of the arcs. The network at the right, however, is both weakly and strongly connected because it is possible to go from any node to any other by tracing over the arcs in the directions in which they point.*

Definition 12.4. *A part of a regulatory network is said to be* ***weakly connected*** *if it is possible to get from any node to any other by travelling—in either direction—along the arcs.*

This notion of connectedness is called "weak" to distinguish it from strong connectedness, which we discussed in Section 10.2. Figure 12.2 offers some examples.

Definition 12.5. *A* ***motif*** *is a small, directed, weakly connected subgraph of a regulatory network that has no parallel arcs and, if it contains more than one node, no loops.*

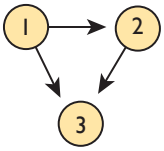
As we'll see, motifs are interesting if they appear more often that one would expect by chance in some suitable family of graphs.

12.2.1 Adjacency matrices

Drawings of large networks such as the one in Figure 11.4 provide one way to represent a network, but they are not especially well-suited to doing computations, so we will instead rely on an alternative representation in terms of *adjacency matrices*. Given a network with N nodes, one builds its adjacency matrix by first numbering the nodes. The adjacency matrix A is then an $N \times N$ matrix whose entries are given by the following rule:

$$A_{ij} = \begin{cases} 1 & \text{if there is an arc from node } i \text{ to node } j \\ 0 & \text{otherwise} \end{cases} \tag{12.1}$$

For example,

if G is 

then $A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$.

12.2.2 Random graphs

Random graphs were invented by the great and highly-prolific Paul Erdős and his collaborator Alfréd Rényi[2]. For our purposes, they come in two main families:

[2] The key reference is P. Erdős and A. Rényi (1959), On random graphs I, *Publ. Math. Debrecen*, **6**:290–297, which introduced $G(N, E)$.

$G(N, E)$ assigns equal probability to each graph that has N distinguishable nodes and E arcs: Figure 12.3 shows the entire family for the case $N = E = 2$. We'll use $G(N, E)$ for the exact calculations in Section 12.3.

$G(N, p)$ In this family, each possible arc has probability p of being present—and hence probability $(1 - p)$ of being absent—independently of all the others. For a given N this family is larger (contains more graphs) than $G(N, E)$ and, further, these graphs are not all equally likely. Figure 12.4 shows the entire family $G(N = 2, p)$, along with the probability of each member. It's not hard to see that in $G(N, p)$, the probability of finding a graph with exactly E arcs is

$$P(E \text{ arcs in } G(N,p)) = p^E (1-p)^{N^2-E} \left(\begin{array}{c} N^2 \\ E \end{array} \right) \tag{12.2}$$

where the factor p^E is the probability that E arcs are present, the factor $(1 - p)^{N^2-E}$ is the probability that the remaining $N^2 - E$ possible arcs are absent and

$$\left(\begin{array}{c} N^2 \\ E \end{array} \right) = \frac{(N^2)!}{E!(N^2 - E)!}$$

is the binomial coefficient N^2-choose-E, which counts the number of ways to choose the E arcs from the N^2 possibilities.

There are many, many other families of random graphs besides these two, a theme to which we'll return to in Section 12.6, when we'll discuss critiques of the Alon group's work. But $G(N, E)$ and $G(N, p)$ are arguably the best studied families and they are easy to use, both in the sorts of exact calculations we'll do below and in numerical experiments. To generate the adjacency matrix for a random member of $G(N, E)$, one simply draws E elements from the set $\{1, \ldots, N^2\}$ and then sets the corresponding elements (number them, say, left-to-right across rows and top-to-bottom in the matrix) to one. It's even simpler to generate a member of $G(N, p)$: one simply makes N^2 independent tosses of a biased coin that says "arc is present" with probability p and "arc is absent" with probability $(1 - p)$.

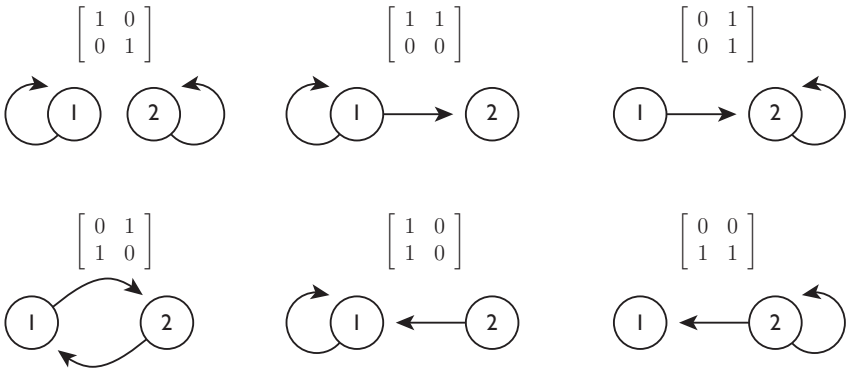


Figure 12.3: *These graphs (shown with their adjacency matrices) are the members of $G(N, E)$ for the case $N = E = 2$: all are equally likely and so have probability 1/6.*

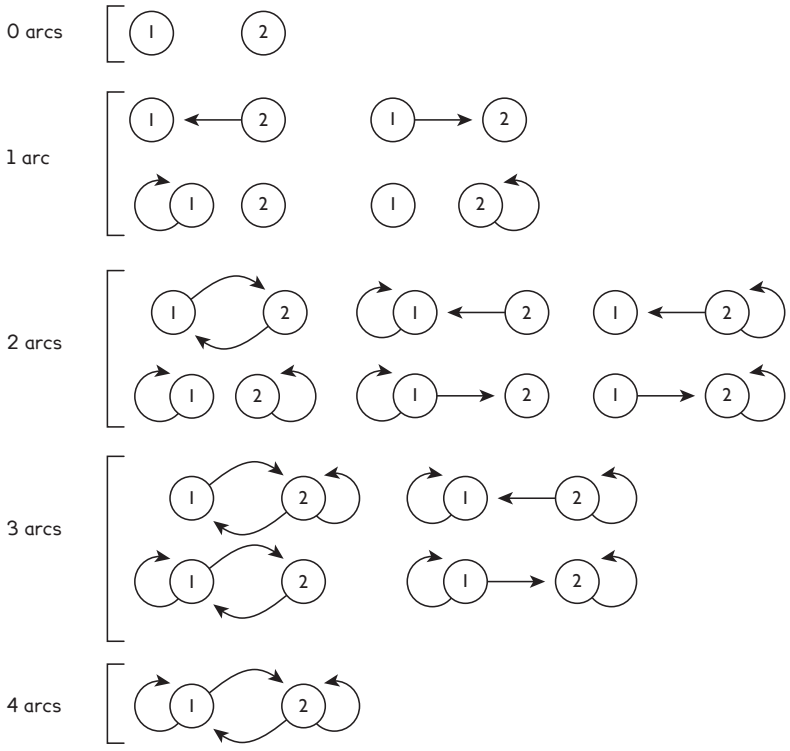


Figure 12.4: *All the regulatory networks that can be made with $N = 2$ genes. $G(N, p)$ assigns them probabilities that depend on the number of arcs E, as given by Eqn.* (12.2)

.

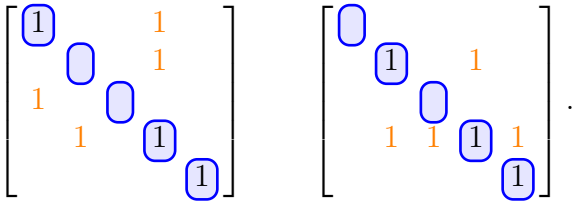


Table 12.1: *Two examples of adjacency matrices for graphs in $G(N = 5, E = 7)$ that contain $k = 3$ autoregulatory loops—entries that are zero have been suppressed for clarity. The 1's representing the autoregulatory loops appear on the diagonal, while the remaining $E - k = 4$ arcs are represented by the four orange off-diagonal 1's that appear in each matrix.*

12.3 Two exact calculations and an application

Here we illustrate the Alon group's strategy by doing exact calculations of the probability of finding k copies of a certain motifs in a random network drawn from $G(N, E)$. As we remarked above, the adjacency matrix of a graph is $G(N, E)$ has exactly E entries equal to 1 and 0's everywhere else. Thus there are exactly N^2-choose-E or

$$\text{Number of graphs in } G(N,E) = \left(\begin{array}{c} N^2 \\ E \end{array} \right) = \frac{(N^2)!}{E!(N^2-E)!} \tag{12.3}$$

equally likely graphs in the family.

12.3.1 The autoregulatory motif

Arguably the simplest possible motif is the autoregulatory loop. Such loops appear as ones on the diagonal of the graph's adjacency matrix: if the graph contains an arc from vertex j to itself then Eqn. (12.1) tells us that $A_{jj} = 1$. So, for example, the adjacency matrices shown in Table 12.1—each of which contains seven ones, three of which are on the diagonal—correspond to graphs in $G(N = 5, E = 7)$ that have exactly three autoregulatory loops.

This observation makes it easy to count those graphs in $G(N, E)$ that contain exactly k autoregulatory loops, as it is the same as counting the number of adjacency matrices that have exactly k ones on the diagonal and $E - k$ ones in off-diagonal positions. That's

$$\text{Number of } k\text{-loop graphs in } G(N,E) = \left(\begin{array}{c} N \\ k \end{array} \right) \left(\begin{array}{c} N^2-N \\ E-k \end{array} \right), \tag{12.4}$$

where the first factor, N-choose-k, accounts for the number of ways to choose the k genes that have autoregulatory loops (or, equivalently, to place k ones on the diagonal of the adjacency matrix) while the second factor, (N^2-N)-choose-$(E-k)$ counts the number of ways to choose the remaining $E - k$ interactions or, equivalently, to distribute the remaining $E-k$ ones over the the N^2-N off-diagonal entries. Bearing

k	$P(k \text{ loops } \mid N=2, E=2)$
0	$\dfrac{\binom{2}{0}\binom{4-2}{2-0}}{\binom{4}{2}} = \dfrac{1\times 1}{6} = \dfrac{1}{6}$
1	$\dfrac{\binom{2}{1}\binom{4-2}{2-1}}{\binom{4}{2}} = \dfrac{2\times 2}{6} = \dfrac{4}{6}$
2	$\dfrac{\binom{2}{2}\binom{4-2}{2-2}}{\binom{4}{2}} = \dfrac{1\times 1}{6} = \dfrac{1}{6}$

Table 12.2: *The results of applying Eqn. (12.5) to the family $G(N=2, E=2)$, which contains 6 graphs. The probabilities computed here agree with those obtained from direct counting of the graphs in Figure 12.3.*

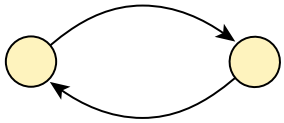
in mind that all graphs in $G(N, E)$ are equally likely, we can combine Eqns. (12.3) and (12.4) to obtain:

$$P(k \text{ loops } \mid N, E) = \frac{\binom{N}{k}\binom{N^2-N}{E-k}}{\binom{N^2}{E}}. \tag{12.5}$$

Table 12.2 lists the results of applying this formula to $G(N=2, E=2)$.

12.3.2 Mutually regulating pairs

Our next example is a mutually-regulating pair such as that illustrated below:



To count graphs that contain exactly k such mutually-regulating pairs we'll generalise the approach form the previous section by thinking carefully about the form of the adjacency matrix.

If gene r and gene s form a mutually-regulating pair, then $A_{r,s} = 1 = A_{s,r}$. That is, a mutually-regulating pair appears as a pair of ones placed symmetrically above and below the diagonal of the adjacency matrix. Thus each a mutually-regulating gives rise to a 1 in the sub-diagonal part of the matrix. This part of the matrix

contains a total of $1+2+\cdots+(N-1) = N(N-1)/2$ entries and so we can count the number of ways to choose our k pairs with the factor

$$\text{Ways to choose } k \text{ mutually-regulating pairs} = \left(\begin{array}{c} N(N-1)/2 \\ k \end{array} \right). \tag{12.6}$$

Eqn. (12.6) accounts for the number of ways to choose the k pairs and thus accounts for $2k$ arcs—two per pair—so we now need to count the ways to choose the remaining $E-2k$ arcs. Equivalently, we want to scatter $E-2k$ ones over the adjacency matrix in such a way that we don't create any new mutually-regulating pairs. One way to accomplish this is to place some (or all) of the remaining ones on the diagonal, which is the same as adding, say, $0 \leq j$ autoregulatory loops to the graph. This clearly can't create any new mutually-regulating pairs, but still leaves us with a final $E-2k-j$ arcs to be accounted for.

The corresponding ones need to appear in off-diagonal positions, subject to the condition that if $A_{r,s} = 1$, then $A_{s,r} = 0$. These patterns of entries are illustrated in Table 12.3 and lead to the result:

$$\left\{ \begin{array}{c} \text{Ways to place remaining } E-2k \\ \text{arcs, including } j \text{ loops} \end{array} \right\} = \left(\begin{array}{c} N \\ j \end{array} \right) \left(\begin{array}{c} N(N-1)/2-k \\ E-2k-j \end{array} \right) 2^{E-2k-j}$$

where:

- the first factor, N-choose-j, accounts for the number of different possible choices of j autoregulatory loops;
- the second factor, $(N(N-1)/2-k)$-choose-$(E-2k-j)$, accounts for the number of ways to choose the sub-diagonal matrix entries associated with the remaining, non-loop arcs;
- the third factor, 2^{E-2k-j}, accounts for the fact that the remaining $E-2k-j$ arcs can appear in two ways: as a sub-diagonal 1 and a symmetric, supra-diagonal 0 or as a sub-diagonal 0 and a symmetric, supra-diagonal 1.

If we now note that the number of loops satisfies $0 \leq j \leq N$ we arrive at the following result,

$$\left\{ \begin{array}{c} \text{Ways to place} \\ \text{remaining} \\ E-2k \text{ arcs} \end{array} \right\} = \sum_{j=0}^{\min(E-2k,N)} \left(\begin{array}{c} N \\ j \end{array} \right) \left(\begin{array}{c} N(N-1)/2-k \\ E-2k-j \end{array} \right) 2^{E-2k-j}, \tag{12.7}$$

where the upper limit on the sum follows from the observation that if the number of extra arcs $E-2k$ is small enough, then all of them can potentially appear as autoregulatory loops, but if $E-2k > N$, then at most N of them can.

Finally, by combining Eqns. (12.6) and (12.7), we obtain

$$\left\{ \begin{array}{c} \text{Number of graphs containing} \\ k \text{ mutually-regulating pairs} \end{array} \right\} =$$

$$\left(\begin{array}{c} N(N-1)/2 \\ k \end{array} \right) \sum_{j=0}^{\min(E-2k,N)} \left(\begin{array}{c} N \\ j \end{array} \right) \left(\begin{array}{c} N(N-1)/2-k \\ E-2k-j \end{array} \right) 2^{E-2k-j},$$

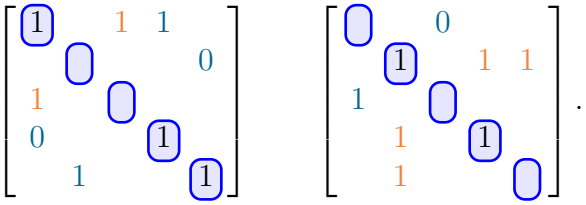


Table 12.3: *Two examples of adjacency matrices for graphs in $G(N = 5,\, E = 7)$ that contain mutually regulating pairs—blank entries are zero. Each mutually-regulating pair gives rise to a pair of entries $A_{r,s} = A_{s,r} = 1$ with $r \neq s$ and these are shown in orange. The remaining arcs are either self-loops that appear as 1's on the diagonal or as as parts of certain symmetrically-placed off-diagonal pairs (shown in blue) in which one entry is 0 and the other is 1.*

and so

$$P\,(k \text{ mutually-regulating pairs } |\, N,\, E) = \frac{\binom{N(N-1)/2}{k} \sum_{j=0}^{\min(E-2k,N)} \binom{N}{j} \binom{N(N-1)/2-k}{E-2k-j} 2^{E-2k-j}}{\binom{N^2}{E}}. \quad (12.8)$$

12.3.3 Application to a network from *E. coli*

Figure 12.5 shows a recent version of the regulatory network for *E. coli*. It includes $N = 424$ genes with $E = 578$ regulatory interactions of which 335 are enhancing, 214 are repressing and a further 29 can—depending the cell's state—be either enhancing or repressing. This network contains 59 autoregulatory loops, which is far more than one would expect in an arbitrary graph drawn from $G(N = 424, E = 578)$. Indeed, exact calculations with *Mathematica* establish that

$$P(k \geq 10 \text{ autoregulatory loops} \,|\, N = 424,\, E = 578) < 1.6 \times 10^{-6},$$

so it is overwhelmingly unexpected to see as many loops as we do. This suggests that natural selection may have favoured the motif. On the other hand, although the *E. coli* network doesn't contain any mutually-regulating pairs, it's hard to draw much of a conclusion as one wouldn't expect to see very many:

$$P(k \geq 5 \text{ mutually-regulating pairs} \,|\, N = 424,\, E = 578) < 3.7 \times 10^{-4}.$$

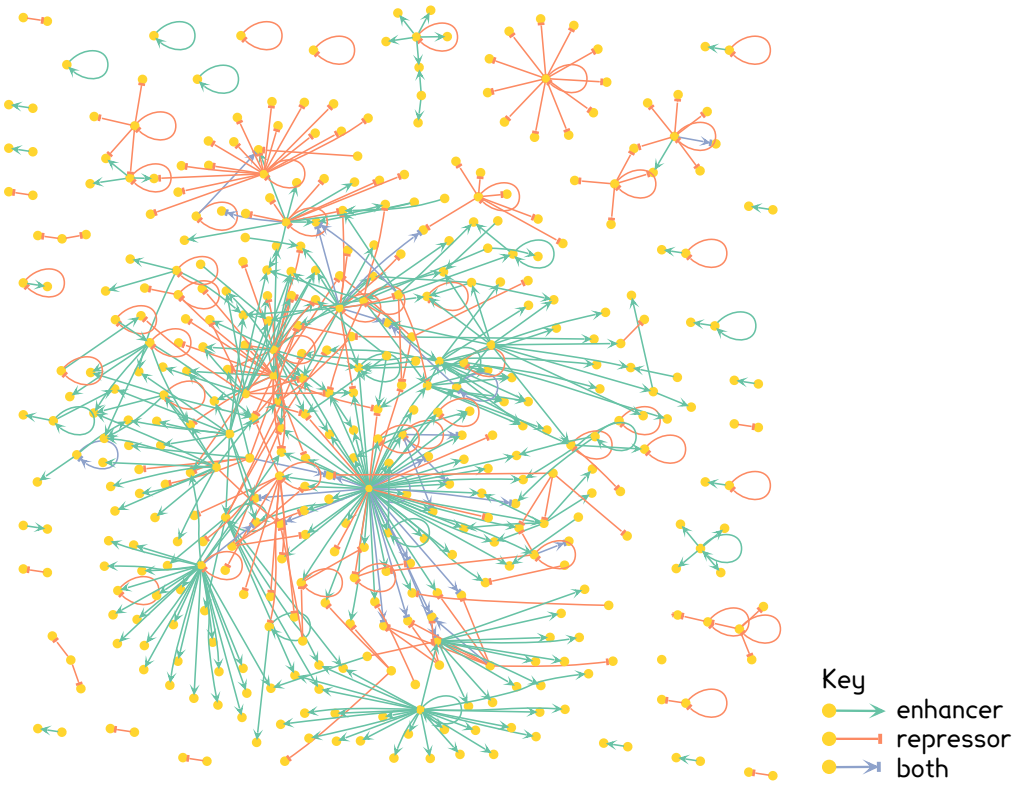


Figure 12.5: *A complete version of the transcriptional regulatory network of the bacterium* ***E. coli****, based on* data *provided by the Alon lab: it is an updated version of the one used in the* 2003 paper *by Mangan* et al.*. Genes are shown as yellow circles and and regulatory interactions are illustrated by three kinds of arcs: green arrows for enhancers; orange, blunt-tipped arcs for repressors and blue arcs with both arrows and blunt tips for transcription factors that both enhance and repress their target.*

Autoregulatory loops		Mutually-regulating pairs	
k	$P(k \mid N = 424, E = 578)$	k	$P(k \mid N = 424, E = 578)$
0	0.255	0	0.395
1	0.349	1	0.368
2	0.238	2	0.170
3	0.108	3	0.052
4	0.037	4	0.012
5	0.010	5	0.002
⋮	⋮	⋮	⋮
10	1.37×10^{-6}	10	3.9×10^{-8}

Table 12.4: *Probabilities of finding k copies of the motifs studied in Sections 12.3.1 and 12.3.2 in the* E. coli *network of Figure 12.5.*

12.4 Approximations in $G(N, p)$

The calculations for motifs in $G(N, E)$ had, by the end of Section 12.3, become rather intricate and so here I'll introduce an approximate approach based on calculating probabilities in the family $G(N, p)$. Although we will no longer require the number E of arcs in our random graphs to be exactly equal to the observed value, we will still want to constrain the graphs in the family to be similar to the observed one: we'll do this via the *expected out-degree*. Consider, for example the vertex v_1 in a graph drawn from $G(N, p)$. It has N possible outgoing arcs, each of the form (v_1, v_j) with $1 \leq j \leq N$, and each of them is present with probability p. This means that v_1's number of outgoing arcs has a binomial distribution:

$$P\left(k \text{ arcs of the form } (v_1, v_j)\right) = p^k (1-p)^{N-k} \left(\begin{array}{c} N \\ k \end{array} \right).$$

The mean of this distribution is

$$\lambda = pN \qquad \text{or} \qquad p = \frac{\lambda}{N} \tag{12.9}$$

and we'll use the quantities N and λ—the number of nodes and the expected number of outgoing arcs per node—to characterise observed networks.

12.4.1 Approximate probabilities for motifs in $G(N, p)$

The arguments in this section are similar to those in the 2003 paper in *Phys. Rev. E* by Itzkovitz *et al.* or in Alon's book. The key result[3] is

Proposition 12.6 (Expected motif counts in $G(N, p)$)**.** *In $G(N, p)$, the expected number of appearances of a motif H that has n nodes and g arcs is*

$$\text{Expected number of appearances of } H \equiv \mathbb{E}(N_H) = \alpha(H) \left(\begin{array}{c} N \\ n \end{array} \right) p^g.$$

where we have introduced the notation $\mathbb{E}(N_H)$ for the expected number of appearances of H. The factor $\alpha(H)$ depends on the symmetries of H and satisfies $1 \leq \alpha(H) \leq n!$.

For those who are interested, the factor $\alpha(H)$ is discussed more thoroughly in Section 12.4.2, but the details are not important for our main argument and may be skipped.

[3] Mathematically intrepid readers might like to look at Noga Alon (no relation to Uri, as far as I know) and Joel E. Spencer (2000), *The Probabilistic Method*, 2nd edition, Wiley Interscience. ISBN 0-471-37046-0.

If we combine Prop. 12.6 with Eqn. (12.9) we obtain

$$\begin{aligned}
\mathbb{E}(N_H) &= \alpha(H) \left(\begin{array}{c} N \\ n \end{array} \right) p^g \\
&= \alpha(H) \frac{N!}{n!(N-n)!} \left(\frac{\lambda}{N} \right)^g \\
&= \lambda^g N^{-g} \left(\frac{\alpha(H)}{n!} \right) \frac{N!}{(N-n)!} \\
&= \lambda^g \left(\frac{\alpha(H)}{n!} \right) \frac{N \times (N-1) \times \cdots \times (N-n+1)}{N^g}.
\end{aligned}$$

And if, as is typically the case, we have that $n \ll N$, we can approximate the product $N \times (N-1) \times \cdots \times (N-n+1)$ by N^n so that

$$\left(\frac{N \times (N-1) \times \cdots \times (N-n+1)}{N^g} \right) \approx \frac{N^n}{N^g} = N^{n-g}$$

and

$$\mathbb{E}(N_H) \approx \lambda^g \left(\frac{\alpha(H)}{n!} \right) N^{n-g}. \tag{12.10}$$

In light of this approximation and the observation that typical regulatory networks involve hundreds or even thousands of genes, we can conclude that if

$g > n$: (the motif of interest has more arcs than nodes) then we expect to find very few or no copies in a typical member of $G(N, p)$;

$g = n$: (the motif has exactly as many arcs as nodes) then we expect to find a few copies in a typical member of $G(N, p)$;

$g < n$: (the motif has more arcs than nodes) then we expect to find many, many copies.

This means that the style of motif-hunting introduced in Section 12.3, which relies on a kind of hypothesis-testing argument to say "over-represented motifs may have been favoured by evolution", is most likely to succeed for motifs with $g \leq n$. In Section 12.5 we'll study for one final motif, the feed-forward loop, that has $n = g = 3$.

12.4.2 Graph automorphisms and the factor $\alpha(H)$

This short section is somewhat technical and not necessary for the main discussion of motifs, but I include it for completeness. Figure 12.6 illustrates the reason that we need the factor $\alpha(H)$ in Prop. 12.6: once we've chosen n nodes to use in a motif, we may be able to arrange them into several distinct versions of the motif. If so, we need to count each version separately, as each makes a contribution of p^g (where g is the number of arcs) to $\mathbb{E}(N_H)$.

Another way to describe this phenomenon is to say that we need to account for the motif's symmetry. For the motif H in Figure 12.6, if we swap the genes assigned to the nodes v_1 and v_2 we get a graph that looks different, but actually has exactly

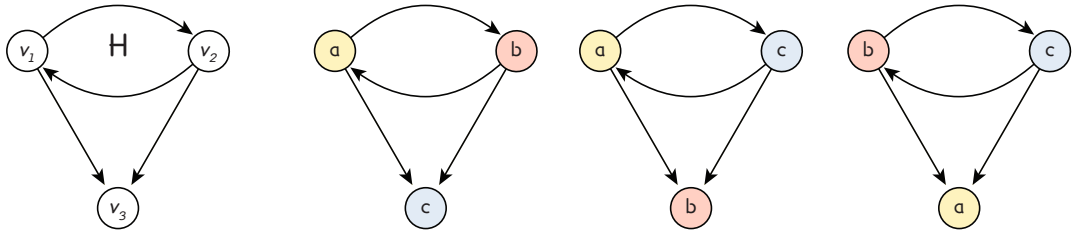


Figure 12.6: *Building motifs with named genes. Once we have chosen 3 genes—call them a, b and c—we can use them to make three distinct versions of the triangular motif H at left and each will contribute $p^g = p^4$ to the expectation $\mathbb{E}(N_H)$.*

the same regulatory interactions. In fact, each of the three distinct versions of H that appears in Figure 12.6 has a symmetric partner produced by interchanging the genes assigned to v_1 and v_2. The natural mathematical language for talking about symmetry is group theory[4], but before we can apply it, we'll need a bit of notation and a pair of definitions.

We'll write $H(V, E)$ to mean a motif H whose nodes form a set V and whose arcs, represented as $u \to v$, (with $u, v \in V$) form a set E.

Definition 12.7. *Two motifs $H_1(V_1, E_1)$ and $H_2(V_2, E_2)$ are said to be **isomorphic** if there exists a bijection[5] $\alpha : V_1 \to V_2$ such that the arc $\alpha(a) \to \alpha(b) \in E_2$ if and only if $a \to b \in E_1$. If such a bijection α exists, it is called an **isomorphism** between G_1 and G_2.*

This is a natural notion of "sameness" in graphs and means, informally, that one can convert G_1 into G_2 simply by relabelling its vertices.

Definition 12.8. *An **automorphism** of a motif $H(V, E)$ is a bijection $\alpha : V \to V$ such that $\alpha(v_j) \to \alpha(v_k) \in E$ if and only if $v_j \to v_k \in E$.*

That is, an automorphism is an isomorphism between a graph and itself.

Given a motif $H(V, E)$, the bijections $\beta : V \to V$ are essentially permutations of the genes assigned to the nodes, so they form a group $\mathcal{B}_H$ with $|\mathcal{B}_H| = n!$ elements whose group operation is composition of bijections. The automorphisms of H form a subgroup of this group (scrupulous readers should check this) that we'll call $\mathcal{A}_H$. For the motif in Figure 12.6, $\mathcal{A}_H$ has just two elements: the identity and the permutation that swaps the genes assigned to v_1 and v_2.

Finally, once we have $\mathcal{B}_H$ and $\mathcal{A}_H$ we can define $\alpha(H)$ as follows:

$$\alpha(H) \equiv |\mathcal{B}_H \backslash \mathcal{A}_H| \;=\; \frac{|\mathcal{B}_H|}{|\mathcal{A}_H|} \;=\; \frac{n!}{|\mathcal{A}_H|},$$

which is the number of distinct versions of H that can be formed with a given set of n named genes. Our approximation Eqn. (12.10) then becomes

$$\mathbb{E}(N_H) \approx \lambda^g \left(\frac{\alpha(H)}{n!}\right) N^{n-g} = \frac{\lambda^g N^{n-g}}{|A_H|}.$$

[4]Those wishing to become fluent in this sort of talk might consider the 3rd year module MATH35081, Symmetry in Nature.

[5]Recall that a bijection is a mapping that's one-to-one and onto.

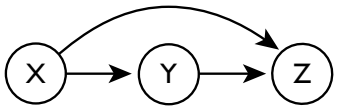


Figure 12.7: *A feed-forward loop involves three genes: X regulates Z both directly and indirectly, via Y.*

12.5 The feed-forward loop motif

The last motif we'll study is the *feed-forward loop* (FFL), in which a gene X regulates a gene Z in two ways: directly, and indirectly via a third gene Y, as illustrated in Figure 12.7. An FFL has $n = 3$ nodes and $g = 3$ arcs, so the arguments from Section 12.4.1 suggest that we should expect a modest, but nonzero number of copies in a real network.

12.5.1 Coherent and incoherent FFL's

In order to study the *E. coli.* network represented in Figure 12.5, it's helpful to make a distinction between two types—enhancing and repressing—of regulatory interactions. If we have a motif with g arcs and we specify the type of each arc, then we get a total of 2^g variants of the motif: Figure 12.9 illustrates the $2^3 = 8$ possible types of feed-forward loops.

These eight FFL's fall into two families according to whether the direct and indirect influences of X on Z are the same or not.

Definition 12.9. *A feed-forward loop is **coherent** if the direct and indirect influences of X on Z are of the same type (enhancing or repressing). If the direct and indirect influences are of opposite type, then the FFL is **incoherent**.*

To compute the net influence of a path through an FFL (or, indeed, through any motif or network), one forms a product with a factor for each arc in the path, using $+1$ for enhancing arcs and a -1 for a repressing arc. Thus for the FFL at left in Figure 12.8, we have

$$\text{direct influence} = +1, \qquad \text{indirect influence} = (-1) \times (-1) = +1, \tag{12.11}$$

so this loop is coherent (repressing a repressor has a net enhancing effect). For the FFL at right the relevant calculations are

$$\text{direct influence} = +1, \qquad \text{indirect influence} = (-1) \times (+1) = -1, \tag{12.12}$$

so this FFL is *incoherent*.

Figure 12.8: *A coherent and an incoherent FFL: arcs are shown with their associated signs. The feed-forward loop at left is coherent, while the one at right is incoherent: see Eqns.* (12.11) *and* (12.12) *for details.*

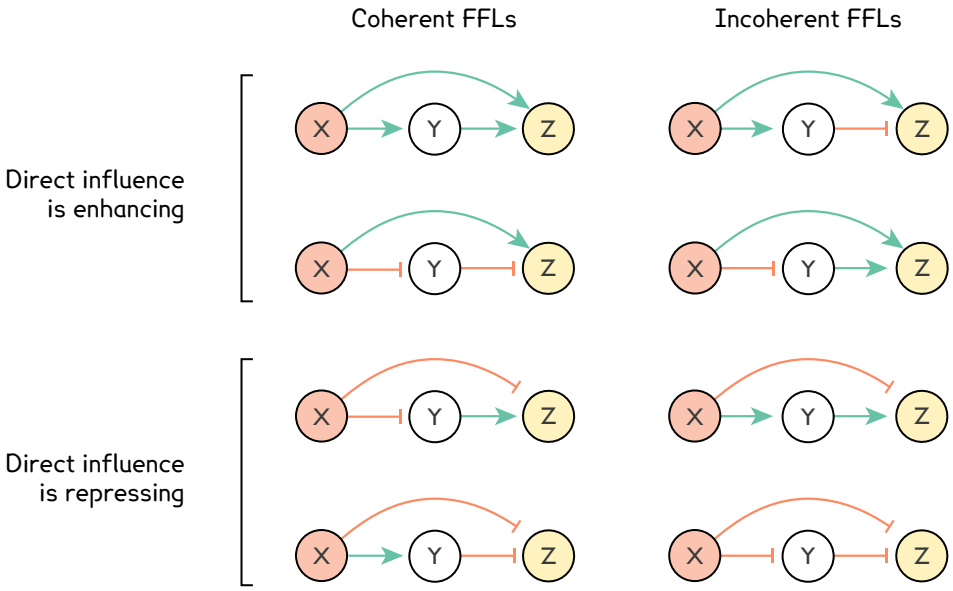


Figure 12.9: *The 8 kinds of feed-forward loop. Here, as in Figure 12.5, enhancing interactions are shown with green, sharp-tipped arcs, while repressing interactions are shown in orange, with blunt tips. The FFL's fall into two classes:* ***coherent*** *FFL's (left column), in which the direct and indirect influences of X on Z are have the same net effect and* ***incoherent*** *ones (right column), in which the direct and indirect influences differ.*

Figure 12.10: *The most common coherent (left) and incoherent (right) FFL's in the* E. coli *network of Figure 12.5 and Table 12.5.*

12.5.2 FFL's in the *E. coli* network

The regulatory network in Figure 12.5 includes 424 nodes with 335 enhancing interactions, 214 repressing interactions and 29 interactions that are both enhancing and repressing, leading to a total of 578 regulatory interactions. This means that the mean out-degree is

$$\lambda = \frac{E}{N} = \frac{578}{424} \approx 1.36$$

and so Prop. 12.6 leads us to expect

$$\begin{aligned}
\mathbb{E}(N_{FFL}) &= \alpha(\text{FFL}) \binom{N}{n} \left(\frac{\lambda}{N}\right)^g \\
&= 1 \times \binom{424}{3} \left(\frac{\lambda}{424}\right)^3 \\
&= \frac{718116815319}{1712928739328} \\
&\approx 0.42
\end{aligned}$$

copies of the FFL motif. In passing from the first line to the second, I have used the facts that feed-forward loops involve $n = 3$ genes and $g = 3$ arcs and that $\alpha(\text{FFL}) = 1$, as the FFL motif has no symmetry: any permutation of the labels X, Y and Z in Figure 12.7 gives rise to a distinct FFL that has different regulatory interactions than the one pictured. Another way to see that $\alpha(\text{FFL}) = 1$ is to note that all three genes in an FFL play distinct roles.

The actual network has 42 FFLs, many more than one would expect in $G(N, p)$, so it seems that evolution may have favoured this motif. Table 12.5 provides further detail, giving the counts for each of the eight kinds of FFL shown in Figure 12.9. Here there may be further evidence of selection in that two FFL's—one coherent and one incoherent—occur much more frequently than others of the same type. The most frequently occurring FFLs are illustrated in Figure 12.10.

Pattern of Regulation ($X \to Z$, $X \to Y$, $Y \to Z$)	Type	Count
$(+,+,+)$	Coherent	28
$(+,-,-)$	Coherent	1
$(-,-,+)$	Coherent	2
$(-,+,-)$	Coherent	4
$(+,+,-)$	Incoherent	5
$(+,-,+)$	Incoherent	1
$(-,+,+)$	Incoherent	1
$(-,-,-)$	Incoherent	0

Table 12.5: *Feed-forward loops in the regulatory network of* E. coli. *The pattern of regulation is specified with a triple of* $\pm$ *signs where, as in Figure 12.8, a* $+$ *indicates an enhancing interaction and a* $-$ *indicates a repressing one. The first sign refers to the direct influence of* X *on* Z*, while the second described the influence of* X *on* Y *and the third the influence of* Y *on* Z*. In cases where one or more of the arcs in an FFL has both enhancing and repressing activity, I have followed the 2003 paper of Mangan et al. and counted it as an enhancer only.*

12.6 Afterword

The work described above stimulated a lot of research and raised quite a few questions, some of which I'll discuss briefly.

What does a motif do for the organism?

This is a difficult question to answer as it's not possible to say much about regulatory networks in the deep past: we can only observe organisms alive now and, though it is sometimes possible to see evolution happening in real time (viruses, for example, evolve very rapidly when they expand into a new host species), this doesn't necessarily tell us much about what happened many hundreds of millions of years ago. Alon and his collaborators did, nonetheless, attempt to address this question in the some of the more recent papers among those listed at the beginning of this chapter.

Their main idea is that a motif will be selected-for if it does some useful job for the cell: if it is a useful sub-circuit that performs some consistent, recognisable function. Thus they built ODE models for gene expression as regulated by, for example, the feed-forward motif and analysed them, writing papers with such titles as "The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks" or "The incoherent feed-forward loop accelerates the response-time of the *gal* system of *Escherichia coli*".

But giving after-the-fact explanations for why one feature survived through millennia of evolution while another didn't is a fraught business: such accounts are sometimes dismissed as "just-so stories" because present-day scientists have no real way to check whether the explanation is true or not. Alon's group has faced similar

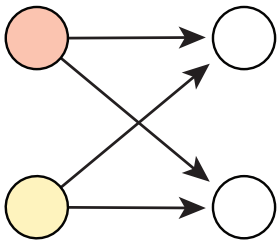


Figure 12.11: *The bi-fan motif:. two genes mutually regulate two others.*

criticism, notably in

> Piers J Ingram, Michael PH Stumpf and Jaroslav Stark (2006), Network motifs: structure does not determine function, *BMC Genomics*, **7**:108. DOI: 10.1186/1471-2164-7-108

which studies a particular motif—the bi-fan, see Figure 12.11—and demonstrates that there is

> ... no *characteristic* behaviour for the motif, and with the correct choice of parameters and of internal structure, very different, indeed even opposite behaviours may be obtained.

Which family of random graphs?

In Sections 12.3–12.5 we used two well-studied families of random graphs, $G(N, E)$ and $G(N, p)$, as the basis of a hypothesis test to determine whether a given motif was over-represented. In both cases, we used the properties of the graph to fix the parameters of the family, but as there are only two parameters in these families, the control this affords us is limited. One can ask whether, once we've fixed N and E in $G(N, E)$ or N and p in $G(N, p)$, a graph drawn at random from the resulting distribution "looks like" the observed regulatory network.

Perhaps unsurprisingly, the short answer is "No" and one way to see this is to consider the out-degrees of the vertices. In $G(N, p)$ it's not hard to prove that for a given node v,

$$P(\deg_{out}(v) = k) = \left(\frac{\lambda^k}{k!}\right) e^{-\lambda}.$$

That is, the out-degrees of vertices in $G(N, p)$ have a Poisson distribution: in-degrees have the same distribution. And although it's fiddlier to find the distribution of $\deg_{out}(v)$ for $G(N, E)$, the basic principle remains the same, low degrees are most likely and high degrees are extremely unlikely.

Observed regulatory networks are not like this at all, as is illustrated in Figure 12.12, which shows the distribution of out-degrees for the *E. coli* network we studied. As is typical of real regulatory networks, the one in Figure 12.5 has a few "hubs" or "master regulators": genes with very high out-degree. To address this problem, researchers in many fields have invented families of random graphs that

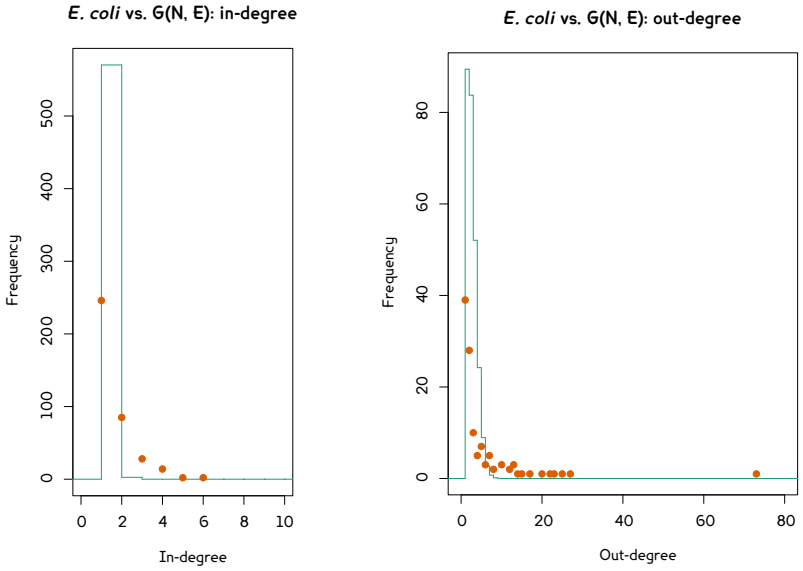


Figure 12.12: *The in- and out-degrees of vertices in the* E. coli *regulatory network differ from those for* $G(N, E)$*. The distribution of in- (left) and out-degrees (right) found in a random sample of 5000 graphs drawn from* $G(N = 424, E = 578)$ *is shown as a green, piecewise constant function, while those observed in the network of Figure 12.5 are shown as orange dots.*

have more parameters and so can better approximate the properties of an observed network. A particularly simple and attractive family invented by Fan Chung and Linyuan Lu is similar to $G(N, p)$, but has non-uniform probabilities chosen to ensure that vertices have their observed in- and out-degrees in expectation: the directed arc from v_i to v_j exists with probability

$$p_{ij} = \frac{\deg_{out}(v_i) \times \deg_{in}(v_j)}{E}$$

where E is the total number of edges. If you find this interesting, you might want to look at some of the following:

- Fan Chung and Linyuan Lu (2002), Connected components in a random graph with given degree sequences, *Annals of Combinatorics*, **6**:125–145. DOI: 0218-0006/02/020125-21

- Mark E Newman, Steve H Strogatz, and Duncan J Watts (2001), Random graphs with arbitrary degree distributions and their applications, *Physical Review E*, **64**:026118. DOI: 10.1103/PhysRevE.64.026118

- E. S. Roberts and A. C. C. Coolen (2012), Unbiased degree-preserving randomization of directed binary networks,*Physical Review E*, **85**:046103. DOI: 10.1103/PhysRevE.85.046103

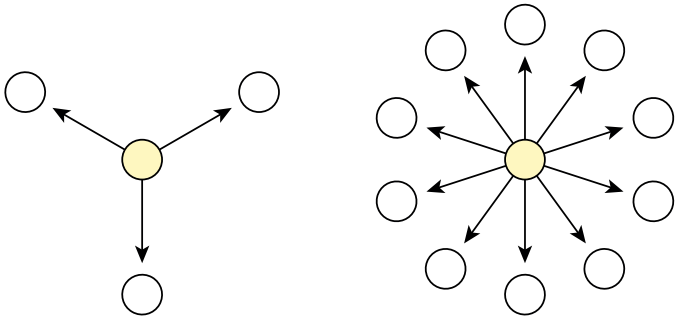


Figure 12.13: *At left, a four-node motif and at right, an eleven-node subgraph that, depending on how one counts, way contain many copies of the motif.*

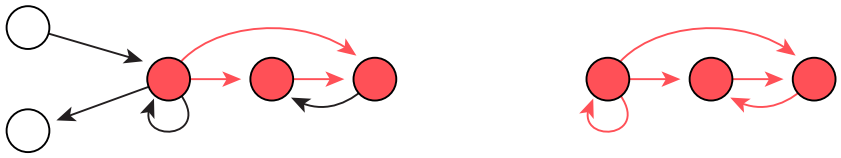


Figure 12.14: *The difference between a motif and a graphlet: the small regulatory network at left includes what Alon's group would regard as a feed-forward loop (vertices and arcs both in red), but this ignores two of the edges (shown in black) that run between the red vertices. In Pržulj's work we would include these two edges and so the network at left would contain an example of the graphlet at right.*

Why motifs and not some other subgraphs?

Finally, one might find the notion of motif itself somewhat odd and unsatisfying. Suppose, for example, that we were studying the four-node motif at left in Figure 12.13 and decided to analyse a network that contained the 11-node structure illustrated at right in the same figure: should we ignore it, or count it as 10-choose-3 = 120 examples of the motif? Alon and his collaborators would do the latter, but other authors, notably Nataša Pržulj, take a different approach.

The basic objects of her analysis are small, weakly-connected graphs that she calls *graphlets*. One says that a network contains a particular n-node graphlet if, somewhere in the network, one can find a set of n nodes such that they—*along with* ***all*** *the arcs that run between them*—look like the graphlet: Figure 12.14 illustrates the distinction between a graphlet and a motif. Dr. Pržulj has had considerable success with this approach and you can read about some of it in:

> Nataša Pržulj (2007), Biological network comparison using graphlet degree distribution, *Bioinformatics*, **23**:177–183.
> DOI: 10.1093/bioinformatics/btl301