ARE PHENOTYPE LANDSCAPES OF RNA MOLECULES WELL DESCRIBED BY A CONFIGURATION MODEL?

1. CONFIGURATION MODEL OF RANDOM NETWORKS

Then *configuration model* $\mathrm{CM}_n(\mathbf{d})$ is an ensemble of random multigraphs defined as follows. There is a set of n nodes each of which has a prescribed degree. Degree numbers define the vector $\mathbf{d} \equiv (d_1, \ldots, d_n)$, where d_i is the degree of node i. This defining setup can be thought of as a set of nodes with a number of stubs d_i (see Fig. 1). Each stub is a half-link, so $2\ell = \sum_{i=1}^{n} d_i$, where ℓ is the number of links of the graph. Note that this puts a constraint on $\mathbf{d}$ —acceptable vectors are only those whose sum of components is an even number.

A graph $\mathscr{G} \in \mathrm{CM}_n(\mathbf{d})$ is constructed by randomly joining pairs of stubs, each pair chosen equiprobably, until no free stubs are left. The name of the model stems from its equivalence to a uniform pair matching of 2ℓ elements (the stubs).

Graphs in $\mathrm{CM}_n(\mathbf{d})$ are not all equiprobable. It is not difficult to show that the probability of finding a graph $\mathscr{G} \in \mathrm{CM}_n(\mathbf{d})$ with x_{ij} links joining nodes $i, j = 1, \ldots, n$ (x_{ii} is the number of self-loops) is [1]

$$\Pr(\mathbf{x}|\mathbf{d}) = \frac{\prod_i d_i!}{(2\ell-1)!! \prod_i 2^{x_{ii}} \prod_{j \geqslant i} x_{ij}!}, \tag{1}$$

where $\mathbf{x} \equiv (x_{ij})$ is under the constraints

$$d_i = x_{ii} + \sum_{j=1}^{n} x_{ij}, \quad i = 1, \ldots, n. \tag{2}$$

A variant of this model is the self-loop-free configuration model $\mathrm{SLFCM}_n(\mathbf{d})$, defined as the subset of $\mathrm{CM}_n(\mathbf{d})$ made of those graphs without self-loops. Note that contrary to what happens with $\mathrm{CM}_n(\mathbf{d})$ this new set may be empty. The probability distribution of this

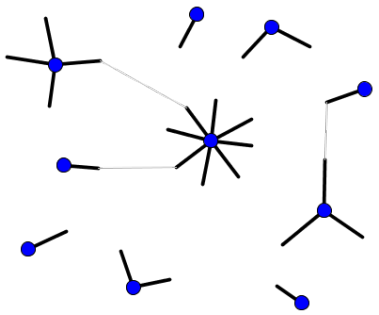

FIGURE 1. Setup of the configuration model: there is a number of nodes with a given number of stubs each.

set is

$$\Pr(\mathbf{x}|\mathbf{d}) = \frac{1}{\mathscr{Q}(\mathbf{d})\prod_i\prod_{j>i}x_{ij}!}, \tag{3}$$

where $\mathbf{x}\in\mathscr{X}$, the set defined by

$$\mathscr{X} \equiv \left\{ \mathbf{x}=(x_{ij}) : x_{ji}=x_{ij},\ d_i=\sum_{j\neq i}x_{ij}, \quad i,j=1,\ldots,n \right\}. \tag{4}$$

The normalisation constant

$$\mathscr{Q}(\mathbf{d}) \equiv \sum_{\mathbf{x}\in\mathscr{X}} \frac{1}{\prod_i\prod_{j>i}x_{ij}!} \tag{5}$$

does not have a simple explicit expresion as a function of $\mathbf{d}$ and n, however the generating function of $\mathbf{z}=(z_1,\ldots,z_n)$ defined as

$$\Xi(\mathbf{z}) \equiv \sum_{\mathbf{d}\in\mathbb{N}_0^n} \mathscr{Q}(\mathbf{d})\prod_{i=1}^n z_i^{d_i} \tag{6}$$

can be shown to have the simple expression (see Appendix)

$$\Xi(\mathbf{z}) = \exp\left(\frac{1}{2}\sum_{i=1}^n\sum_{j\neq i} z_i z_j\right). \tag{7}$$

2. TESTING THE CONFIGURATION MODEL

Suppose we have a multigragh $\mathscr{G}$ with n nodes, degree vector $\mathbf{d}$, and connectivity matrix $\mathbf{x}\in\mathscr{S}_0$. We want to decide whether $\mathscr{G}$ is a 'typical' member of $\mathrm{CM}_n(\mathbf{d})$. It will be if $\mathscr{G}$ has a relatively high probability according to (3). One way to assign a p-value to the hipothesis "$\mathscr{G}$ is a 'typical' member of $\mathrm{CM}_n(\mathbf{d})$" is to compute

$$p = \sum_{P(\mathbf{y}|\mathbf{d})<P(\mathbf{x}|\mathbf{d})} P(\mathbf{y}|\mathbf{d}). \tag{8}$$

Then p is the probability to find a link configuration $\mathbf{y}$ less probable than $\mathbf{x}$. We can reject the hipothesis if, say, $p<0.05$.

Computing p *brute force* may be a daunting task, so we can resort to a Monte Carlo simulation to estimate it. To this purpose we need to generate a Markov chain $\mathbf{x}_t$, $t=0,1,2,\ldots$, with the stationary probability distribution $P(\mathbf{x}|\mathbf{d})$, and then estimate p as

$$p \approx \frac{1}{T}\sum_{t=t_0}^{T+t_0} \Theta\Big(P(\mathbf{x}|\mathbf{d}) - P(\mathbf{x}_n|\mathbf{d})\Big), \tag{9}$$

where $\Theta(x)=1$ if $x\geqslant 0$ and 0 otherwise.

Let $\mathbf{x}$ and $\mathbf{x}'$ respectively be the connection matrices of two graphs $\mathscr{G},\mathscr{G}'\in\mathrm{SLFCM}_(\mathbf{d})$. We seek for the transition probability $T(\mathbf{x}\to\mathbf{x}')$ of an ergodic Markov chain. If we assume it to satisfy detailed balance, the condition that these transition probabilities must fulfill is (for simplicity we drop the dependence on $\mathbf{d}$ in the probability distribution)

$$P(\mathbf{x})T(\mathbf{x}\to\mathbf{x}') = P(\mathbf{x}')T(\mathbf{x}'\to\mathbf{x}). \tag{10}$$

Let us factor the transition probabilities as $T(\mathbf{x}\to\mathbf{x}') = M(\mathbf{x}\to\mathbf{x}')A(\mathbf{x}\to\mathbf{x}')$, where $M(\mathbf{x}\to\mathbf{x}')$ is a proposed transformation from $\mathbf{x}$ to $\mathbf{x}'$ (a 'Monte Carlo move') and $A(\mathbf{x}\to\mathbf{x}')$ is the probability to accept such a transformation. $M(\mathbf{x}\to\mathbf{x}')$ must take into account that $\mathbf{x}$ and $\mathbf{x}'$ have to be valid configurations, i.e., the cannot change the degrees of the nodes. A simple proposal for $M(\mathbf{x}\to\mathbf{x}')$ is illustrated in Fig. 2. It amounts to random and uniformly

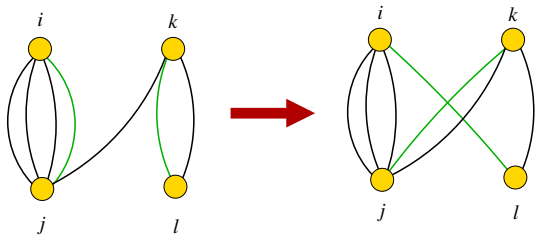

FIGURE 2. Monte Carlo move: two randomly chosen pairs of connected nodes, (i,j) and (k,l), exchange one link.

choosing two pairs of connected nodes and exchanging one of their links (also random and uniformly chosen). In other words,

$$x'_{ij} = x_{ij} - 1, \quad x'_{kl} = x_{kl} - 1, \quad x'_{il} = x_{il} + 1, \quad x'_{kj} = x_{kj} + 1, \tag{11}$$

and $x'_{pq} = x_{pq}$ for any other pair of indices $p \neq q$.

This choice clearly maintains the degrees of all nodes and is reversible, i.e., $M(\mathbf{x} \to \mathbf{x}') = M(\mathbf{x}' \to \mathbf{x})$. Therefore (10) becomes

$$\frac{A(\mathbf{x} \to \mathbf{x}')}{A(\mathbf{x}' \to \mathbf{x})} = \frac{P(\mathbf{x}')}{P(\mathbf{x})}. \tag{12}$$

Substituting (3) in this equation and taking into account (11) we end up with

$$\frac{A(\mathbf{x} \to \mathbf{x}')}{A(\mathbf{x}' \to \mathbf{x})} = \frac{x_{ij}x_{kl}}{(x_{il}+1)(x_{kj}+1)}. \tag{13}$$

Then the 'Metropolis' choice for $A(\mathbf{x} \to \mathbf{x}')$ is

$$A(\mathbf{x} \to \mathbf{x}') \equiv \min\left\{1, \frac{x_{ij}x_{kl}}{(x_{il}+1)(x_{kj}+1)}\right\}. \tag{14}$$

3. CONFIGURATION OF MAXIMUM LIKELIHOOD

In order to find the most probable configuration we can maximise $\log P(\mathbf{x}|\mathbf{d})$ for the distribution (3) within the set (4). Assuming $x_{ij} \gg 1$ for all $1 \leqslant i < j \leqslant n$, this amounts to maximising

$$\Phi(\mathbf{x}|\mathbf{d}) = \sum_{i=1}^{n}\left\{\sum_{j>i}(x_{ij} - x_{ij}\log x_{ij}) + \lambda_i\left(\sum_{k<i}x_{ki} + \sum_{k>i}x_{ik}\right)\right\}, \tag{15}$$

where λ_i, $i = 1,\ldots,n$, are the Lagrange multipliers associated to the constraints on the degrees. Then, for all $1 \leqslant i < j \leqslant n$,

$$\frac{\partial\Phi}{\partial x_{ij}} = -\log x_{ij} + \lambda_i + \lambda_j = 0, \tag{16}$$

so the minimum is reached when $x_{ij} = \xi_i\xi_j$ (where $\xi_i \equiv e^{\lambda_i}$). Imposing the constraints,

$$d_i = \xi_i\sum_{j\neq i}x_j = \xi_i(\Xi - \xi_i), \qquad \Xi \equiv \sum_{i=1}^{n}\xi_i, \tag{17}$$

The case where $d_i = d$ (hence $\xi_i = \xi$) for all $i = 1,\ldots,n$ yields $\Xi = n\xi$ and the constraint becomes $d = (n-1)\xi^2$. Accordingly $\xi = \sqrt{d/(n-1)}$.

In the general case we have

$$\xi_i^2 - \Xi\xi_i + d_i = 0, \tag{18}$$

with the two solutions

$$\xi_i = \frac{\Xi}{2} \pm \sqrt{\frac{\Xi^2}{4} - d_i}.$$

From them two, that with the minus sign is the one that recovers the solution of the homogeneous degree case. Therefore

$$\xi_i = \frac{\Xi}{2} - \sqrt{\frac{\Xi^2}{4} - d_i}. \tag{19}$$

Adding up equation (19) we obtain an implicit equation for Ξ, namely

$$\frac{1}{n-2}\sum_{i=1}^{n}\sqrt{1-\frac{4d_i}{\Xi^2}} = 1. \tag{20}$$

Now, $\sqrt{1-x}$ is a concave function of x, therefore

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{1-\frac{4d_i}{\Xi^2}} \leqslant \sqrt{1-\frac{4}{\Xi^2}\left(\frac{1}{n}\sum_{i=1}^{n}d_i\right)} = \sqrt{1-\frac{8\ell}{n\Xi^2}}.$$

Accordingly,

$$\frac{n-2}{n} \leqslant \sqrt{1-\frac{8\ell}{n\Xi^2}} \quad \Leftrightarrow \quad \frac{(n-2)^2}{n^2} \leqslant 1-\frac{8\ell}{n\Xi^2},$$

so we obtain the lower bound

$$\Xi \geqslant \sqrt{\frac{2\ell n}{n-1}}. \tag{21}$$

3.1. **Approximation.** Given the lower bound (21) we have

$$\frac{4d_i}{\Xi^2} \leqslant \frac{2d_i}{\ell}\left(1-\frac{1}{n}\right) \ll 1$$

except for extremely biased degree distributions where one of the nodes takes a large fraction of links. Using this approximation

$$\xi_i = \frac{\Xi}{2}\left(1-\sqrt{1-\frac{4d_i}{\Xi^2}}\right) \approx \frac{d_i}{\Xi}. \tag{22}$$

On the other hand, from this approximation $\Xi \approx 2\ell/\Xi$, so

$$\Xi \approx \sqrt{2\ell}, \tag{23}$$

and we are left with the result

$$\xi_i \approx \frac{d_i}{\sqrt{2\ell}}, \tag{24}$$

from which we obtain the approximation

$$x_{ij} \approx \frac{d_i d_j}{2\ell}. \tag{25}$$

This is an estimate of the expected number of links of the most probable configuration graph.

4. RESULTS

We have apply this to the genotype networks obtained for *RNA* of length $L = 12$, both based con phenotypes and based on connected components. In each case, a phenotype or a component is considered a node. The number of outer links (those pointing to other phenotypes/components, excluding those pointing to the open chain) of each node define the degrees d_i, and the links connecting nodes i and j correspond to the actual realisation of variables x_{ij} of the corresponding configuration multigraph (excluding self-links).

Running a Monte Carlo leads to the conclusion that the probability of the RNA12 connection multigraph as a realisation of the configuration model is very low (several orders of magnitude below that of the most probable multigraph). In order to gain further insight we have computed, for each pair of connected nodes i, j, the maximum likelihood estimate (25), and then we have made an histogram of the number of pairs with a given value of the variable $\log(x_{ij}^{\text{estimate}}/x_{ij})$. This histogram, for the component-based analysis, is shown in Figure 3.

This figure shows that the estimate is often off by 2 orders of magnitud with respect to the real value (which explains the outcome of the Monte Carlo process). Nonetheless it also shows a strong peak around 0, suggesting that the configuration model partly captures the connectivity of these graphs.

The question that we need to answer is whether RNA12 is made of much too short molecules for the model to hold, and therefore whether longer chains would produce narrower peaks around 0.

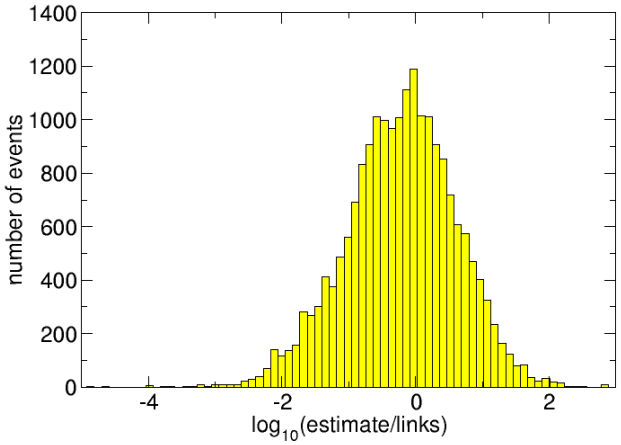

FIGURE 3. Histogram of the number of pairs with a given value of the variable $\log(x_{ij}^{\text{estimate}}/x_{ij})$

APPENDIX A. GENERATING FUNCTION OF THE NORMALIZATION FACTOR

By definition

$$\Xi(\mathbf{z}) = \sum_{\mathbf{d}\in\mathbb{N}_0^n} \left(\prod_k z_k^{d_k}\right) \sum_{\mathbf{x}\in\mathscr{X}} \frac{1}{\prod_i \prod_{j>i} x_{ij}!} = \sum_{\mathbf{d}\in\mathbb{N}_0^n} \sum_{\mathbf{x}\in\mathscr{X}} \frac{\prod_i z_i^{\sum_{k<i} x_{ki} + \sum_{k>i} x_{ik}}}{\prod_i \prod_{j>i} x_{ij}!}$$
$$= \sum_{\mathbf{x}\in\mathscr{S}_0} \frac{\left(\prod_i \prod_{k<i} z_i^{x_{ki}}\right)\left(\prod_i \prod_{k>i} z_i^{x_{ik}}\right)}{\prod_i \prod_{j>i} x_{ij}!},$$

where

$$\mathscr{S}_0 \equiv \left\{\mathbf{x} : \mathbf{x} = \mathbf{x}^\mathsf{T},\ x_{ii} = 0,\ i = 1,\ldots,n\right\}.$$

On the other hand,

$$\prod_i \prod_{k<i} z_i^{x_{ki}} = \prod_k \prod_{i<k} z_k^{x_{ik}} = \prod_i \prod_{k>i} z_k^{x_{ik}},$$

thus

$$\Xi(\mathbf{z}) = \sum_{\mathbf{x}\in\mathscr{S}_0} \frac{\prod_i \prod_{k>i} (z_k z_i)^{x_{ik}}}{\prod_i \prod_{j>i} x_{ij}!} = \sum_{\mathbf{x}\in\mathscr{S}_0} \prod_i \prod_{j>i} \frac{(z_i z_j)^{x_{ij}}}{x_{ij}!} = \prod_i \prod_{j>i} \sum_{x=0}^{\infty} \frac{(z_i z_j)^x}{x!}$$
$$= \exp\left(\frac{1}{2} \sum_{i=1}^n \sum_{j\neq i} z_i z_j\right).$$

REFERENCES

[1] van der Hofstad, R. (2014) *Random Graphs and Complex Networks,* Vol. I (`http://www.win.tue.nl/~rhofstad/NotesRGCN.html`).