Tailored graph ensembles as proxies or null models for real networks II:

Results on directed graphs

E S Roberts

Randall Division of Cell and Molecular Biology, King's College London

Plan of presentation

Calculating Shannon entropy of random graph ensembles constrained by degree distribution and degree-degree correlation

- Motivation and introduction to the problem
- Previously published results: undirected networks
- New results: directed networks
- Example application to gene regulation networks
- Conclusions, references, acknowledgements and questions.

Motivation

Working with bioinformatics colleagues, it was observed that there is the need for precise approaches to quantify topological properties of large networks.

Aim is to develop rigorous and precise tools which can be applied to studying real biological networks.

Abstracting a network

- A network will be represented as a matrix **c**.
- $c_{ij} = 0$ or 1 .
- Undirected network := symmetric matrix.
- Directed network := unsymmetric matrix.

- Degrees or degree pairs - k_i or $\vec{k}_i = (k_i^{\text{in}}, k_i^{\text{out}}))$ - are for each node i drawn independently from a specified (joint) degree distribution $p(k)$ (or $p(\vec{k})$).
- The average connectivity of the network =: $\bar{k}$.
- The degree-degree correlation is expressed via $W(k,k') = (N\bar{k})^{-1} \sum_{ij} c_{ij} \delta_{k,k_i} \delta_{k',k_j}$ or an analogous expression for the directed case.
- $\pi_{\bar{k}}(k) = e^{-\bar{k}} \bar{k}^k / k!$ (Poisson)
- $W(k)$ is the marginal of $W(k,k')$

Defining the problem

The aim is to calculate the Shannon Entropy $S = -N^{-1}\sum_{\boldsymbol{c}} p(\boldsymbol{c}) \log p(\boldsymbol{c})$ of random graph ensemble defined via tailored topological constraints.

Defining the problem

The aim is to calculate the Shannon Entropy $S = -N^{-1} \sum_{\boldsymbol{c}} p(\boldsymbol{c}) \log p(\boldsymbol{c})$ of random graph ensemble defined via tailored topological constraints.

Additionally, these methods are applied in order to calculate a Kullback-Leibler distance between two ensembles A and B.

$$D_{AB} = \frac{1}{2N} \sum_{\boldsymbol{c}} \left\{ p(\boldsymbol{c}|p_A, Q_A) \log \left[\frac{p(\boldsymbol{c}|p_A, Q_A)}{p(\boldsymbol{c}|p_B, Q_B)} \right] + p(\boldsymbol{c}|p_B, Q_B) \log \left[\frac{p(\boldsymbol{c}|p_B, Q_B)}{p(\boldsymbol{c}|p_A, Q_A)} \right] \right\}$$

Results from *Annibale et. al.*

Complexities of constrained random undirected graph ensembles.

$$S_0[\bar{k}] = \frac{1}{2}\bar{k}\left[\log(N/\bar{k}) + 1\right]$$

Results from *Annibale et. al.*

Complexities of constrained random undirected graph ensembles.

$S_0[\bar{k}] =$	$\frac{1}{2}\bar{k}\left[\log(N/\bar{k})+1\right]$
$\mathcal{C}_{\deg}[p] =$	$\sum_k p(k)\log\left[\frac{p(k)}{\pi_{\bar{k}}(k)}\right]$

Results from *Annibale et. al.*

Complexities of constrained random undirected graph ensembles.

$S_0[\bar{k}] =$	$\frac{1}{2}\bar{k}\left[\log(N/\bar{k}) + 1\right]$
$\mathcal{C}_{\deg}[p] =$	$\sum_k p(k) \log\left[\frac{p(k)}{\pi_{\bar{k}}(k)}\right]$
$\mathcal{C}_{\text{wir}}[p, W] =$	$\frac{1}{2}\bar{k}\sum_{k,k'} W(k,k') \log\left[\frac{W(k,k')}{W(k)W(k')}\right]$

Results from *Annibale et. al.*

Kullback-Leibler distances between constrained ensembles of undirected random graphs.

D^{deg}_{AB}	$\frac{1}{2}\sum_k p_A(k)\log\left[\frac{p_A(k)}{p_B(k)}\right] + \frac{1}{2}\sum_k p_B(k)\log\left[\frac{p_B(k)}{p_A(k)}\right] +$
$+$	
D^{wir}_{AB}	$\frac{1}{4}\bar{k}_A\sum_{k,k'} W_A(k,k')\log\left[\frac{\Pi_A(k,k')}{\Pi_B(k,k')}\right]$
	$+\frac{1}{4}\bar{k}_B\sum_{k,k'} W_B(k,k')\log\left[\frac{\Pi_B(k,k')}{\Pi_A(k,k')}\right] +$
$+$	
D^{int}_{AB}	$\frac{1}{2}\bar{k}_A\sum_k W_A(k)\log\rho_{AB}(k) + \frac{1}{2}\bar{k}_B\sum_k W_B(k)\log\rho_{BA}(k)$

$$p(\boldsymbol{c}|\vec{k}_1 \dots \vec{k}_N) = \frac{\prod_i \delta_{\vec{k}_i, \vec{k}_i(\boldsymbol{c})}}{Z(\vec{k}_1 \dots \vec{k}_N)}, \qquad Z(\vec{k}_1 \dots \vec{k}_N) = \sum_{\boldsymbol{c}} \prod_i \delta_{\vec{k}_i, \vec{k}_i(\boldsymbol{c})}$$

$$p(\boldsymbol{c}|\vec{k}_1 \ldots \vec{k}_N) = \frac{\prod_i \delta_{\vec{k}_i, \vec{k}_i(\boldsymbol{c})}}{Z(\vec{k}_1 \ldots \vec{k}_N)}, \qquad Z(\vec{k}_1 \ldots \vec{k}_N) = \sum_{\boldsymbol{c}} \prod_i \delta_{\vec{k}_i, \vec{k}_i(\boldsymbol{c})}$$

Manipulate the expression for Shannon entropy to show that the key term involves the logarithm of the associated partition function

$$p(\boldsymbol{c}|\vec{k}_1 \dots \vec{k}_N) = \frac{\prod_i \delta_{\vec{k}_i, \vec{k}_i(\boldsymbol{c})}}{Z(\vec{k}_1 \dots \vec{k}_N)}, \qquad Z(\vec{k}_1 \dots \vec{k}_N) = \sum_{\boldsymbol{c}} \prod_i \delta_{\vec{k}_i, \vec{k}_i(\boldsymbol{c})}$$

Manipulate the expression for Shannon entropy to show that the key term involves the logarithm of the associated partition function

This expression is more tractable when we transform $Z(\vec{k}_1 \dots \vec{k}_N)$ into an average involving an alternative measure.

$$w(\boldsymbol{c}|\bar{k}) = \prod_{ij} \left[\frac{\bar{k}}{N} \delta_{c_{ij},1} + \left(1 - \frac{\bar{k}}{N}\right) \delta_{c_{ij},0} \right]$$

$$p(\boldsymbol{c}|\vec{k}_1 \dots \vec{k}_N) = \frac{\prod_i \delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})}}{Z(\vec{k}_1 \dots \vec{k}_N)}, \qquad Z(\vec{k}_1 \dots \vec{k}_N) = \sum_{\boldsymbol{c}} \prod_i \delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})}$$

Manipulate the expression for Shannon entropy to show that the key term involves the logarithm of the associated partition function

This expression is more tractable when we transform $Z(\vec{k}_1 \dots \vec{k}_N)$ into an average involving an alternative measure.

$$w(\boldsymbol{c}|\bar{k}) = \prod_{ij} \left[\frac{\bar{k}}{N} \delta_{c_{ij},1} + \left(1 - \frac{\bar{k}}{N}\right) \delta_{c_{ij},0} \right]$$

$$S = \frac{1}{N} \sum_{\vec{k}_1 \dots \vec{k}_N} \left[\prod_i p(\vec{k}_i) \right] \log \langle \prod_i \delta_{\vec{k}_i,\vec{k}_i(\boldsymbol{c})} \rangle_{\bar{k}} - \sum_{\vec{k}} p(\vec{k}) \log p(\vec{k})$$
$$+ \langle k \rangle \left[\log(N/\langle k \rangle) + 1 \right] + \varepsilon_N$$

Using Fourier representations of the Kronecker deltas and some straightforward manipulations brings us to

$$\phi = \frac{1}{N} \sum_{\vec{k}_1 \ldots \vec{k}_N} \Big[\prod_i p(\vec{k}_i) \Big] \log \int_{-\pi}^{\pi} \prod_i \Big[\frac{\mathrm{d}\omega_i \mathrm{d}\psi_i}{4\pi^2} e^{\mathrm{i}[\omega_i k_i^{\mathrm{in}} + \psi_i k_i^{\mathrm{out}}]} \Big] L(\boldsymbol{\omega}, \boldsymbol{\psi})$$

$$L(\boldsymbol{\omega}, \boldsymbol{\psi}) = \exp \Big[\bar{k} N \Big(\frac{1}{N} \sum_i e^{-\mathrm{i}\omega_i} \Big) \Big(\frac{1}{N} \sum_j e^{-\mathrm{i}\psi_j} \Big) - \bar{k} N + \mathcal{O}(N^0) \Big]$$

A path integral form is achieved by manipulating δ functions in order to isolate the site specific terms.

$$\phi((\mathbf{k}^{\mathbf{in}},\mathbf{k}^{\mathbf{out}})) = \frac{1}{N}\sum_{(\mathbf{k}^{\mathbf{in}},\mathbf{k}^{\mathbf{out}})} p(\mathbf{k}^{\mathbf{in}},\mathbf{k}^{\mathbf{out}}) \log \int \{\mathrm{d}P(\omega)\mathrm{d}\hat{P}(\omega)\}\{\mathrm{d}Q(\psi)\mathrm{d}\hat{Q}(\psi)\}e^{N\Psi}$$

A path integral form is achieved by manipulating δ functions in order to isolate the site specific terms.

$$\phi((\mathbf{k^{in}},\mathbf{k^{out}})) = \frac{1}{N}\sum_{(\mathbf{k^{in}},\mathbf{k^{out}})} p(\mathbf{k^{in}},\mathbf{k^{out}}) \log \int \{\mathrm{d}P(\omega)\mathrm{d}\hat{P}(\omega)\}\{\mathrm{d}Q(\psi)\mathrm{d}\hat{Q}(\psi)\}e^{N\Psi}$$

$$\begin{aligned}\Psi[P,\hat{P},Q,\hat{Q}] = \sum_{k^{in}} p(k^{in}) log \int_{-\pi}^{\pi}\frac{\mathrm{d}\omega}{2\pi}\mathrm{e}^{\mathrm{i}[\omega k^{in}-\hat{P}(\omega)]} \\ +\sum_{k^{out}} p(k^{out}) log \int_{-\pi}^{\pi}\frac{\mathrm{d}\psi}{2\pi}\mathrm{e}^{\mathrm{i}[\psi k^{out}-\hat{Q}(\psi)]} \\ +i\int_{-\pi}^{\pi}\mathrm{d}\omega\hat{P}(\omega)P(\omega) + i\int_{-\pi}^{\pi}\mathrm{d}\psi\hat{Q}(\psi)Q(\psi) \\ +\int_{-\pi}^{\pi}\mathrm{d}\omega\mathrm{d}\psi P(\omega)Q(\psi)\bar{k}(\mathrm{e}^{-\mathrm{i}(\omega+\psi)}) - \bar{k} + \mathcal{O}(N^{-1})\end{aligned}$$

Final form of the directed result: complexity

$$S_0[\bar{k}] = \bar{k}\left[\log(N/\bar{k}) + 1\right]$$

Final form of the directed result: complexity

$S_0[\bar{k}] =$	$\bar{k}\left[\log(N/\bar{k}) + 1\right]$
$\mathcal{C}_{\text{deg}}[p] =$	$\sum_{\vec{k}} p(\vec{k}) \log\left[\frac{p(\vec{k})}{\pi_{\bar{k}}(k^{\text{in}})\pi_{\bar{k}}(k^{\text{out}})}\right]$

Final form of the directed result: complexity

$S_0[\bar{k}] =$	$\bar{k}\left[\log(N/\bar{k})+1\right]$
$\mathcal{C}_{\mathrm{deg}}[p] =$	$\sum_{\vec{k}} p(\vec{k}) \log\left[\frac{p(\vec{k})}{\pi_{\bar{k}}(k^{\mathrm{in}})\pi_{\bar{k}}(k^{\mathrm{out}})}\right]$
$\mathcal{C}_{\mathrm{wir}}[p,W] =$	$\bar{k}\sum_{\vec{k},\vec{k}'} W(\vec{k},\vec{k}') \log\left[\frac{W(\vec{k},\vec{k}')}{W_1(\vec{k})W_2(\vec{k}')}\right]$

Final form of the directed result: distance

D_{AB}^{deg}	$\frac{1}{2}\sum_{\vec{k}} p_A(\vec{k}) \log\left[\frac{p_A(\vec{k})}{p_B(\vec{k})}\right] + \frac{1}{2}\sum_{\vec{k}} p_B(\vec{k}) \log\left[\frac{p_B(\vec{k})}{p_A(\vec{k})}\right]$
$+$	$+$
D_{AB}^{wir}	$\frac{1}{2}\bar{k}_A \sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}') \log\left[\frac{\Pi_A(\vec{k},\vec{k}')}{\Pi_B(\vec{k},\vec{k}')}\right]$
$+$	$+ \frac{1}{2}\bar{k}_B \sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}') \log\left[\frac{\Pi_B(\vec{k},\vec{k}')}{\Pi_A(\vec{k},\vec{k}')}\right]$
	$+$
D_{AB}^{int}	$\frac{1}{2}\bar{k}_A \sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}') \log[\rho_{AB}(\vec{k})\sigma_{AB}(\vec{k}')]$
	$+\frac{1}{2}\bar{k}_B \sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}') \log[\rho_{BA}(\vec{k})\sigma_{BA}(\vec{k}')]$

Final form of the directed result: distance

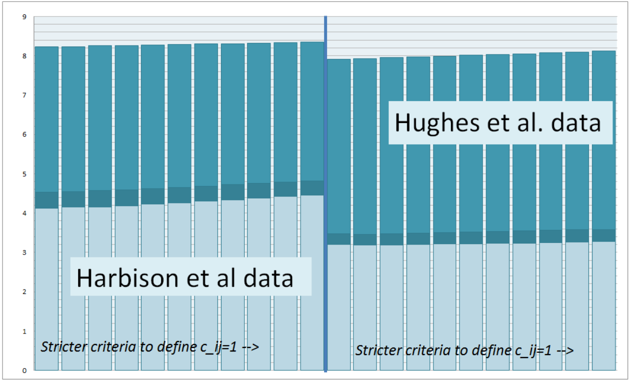
D^{deg}_{AB}	$\frac{1}{2}\sum_{\vec{k}} p_A(\vec{k}) \log\left[\frac{p_A(\vec{k})}{p_B(\vec{k})}\right] + \frac{1}{2}\sum_{\vec{k}} p_B(\vec{k}) \log\left[\frac{p_B(\vec{k})}{p_A(\vec{k})}\right]$
$+$	$+$
D^{wir}_{AB}	$\frac{1}{2}\bar{k}_A \sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}') \log\left[\frac{\Pi_A(\vec{k},\vec{k}')}{\Pi_B(\vec{k},\vec{k}')}\right]$
$+$	$+\frac{1}{2}\bar{k}_B \sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}') \log\left[\frac{\Pi_B(\vec{k},\vec{k}')}{\Pi_A(\vec{k},\vec{k}')}\right]$
	$+$
D^{int}_{AB}	$\frac{1}{2}\bar{k}_A \sum_{\vec{k},\vec{k}'} W_A(\vec{k},\vec{k}') \log[\rho_{AB}(\vec{k})\sigma_{AB}(\vec{k}')]$
	$+\frac{1}{2}\bar{k}_B \sum_{\vec{k},\vec{k}'} W_B(\vec{k},\vec{k}') \log[\rho_{BA}(\vec{k})\sigma_{BA}(\vec{k}')]$

Self consistency relation to be satisfied by the interference term

$$\rho_{AB}(\vec{k}) = \sum_{\vec{k}'} \Pi_B(\vec{k}, \vec{k}') W_{2A}(\vec{k}') \sigma_{AB}^{-1}(\vec{k}')$$

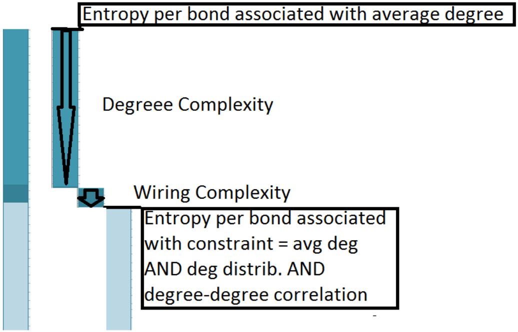
$$\sigma_{AB}(\vec{k}) = \sum_{\vec{k}'} \Pi_B(\vec{k}', \vec{k}) W_{1A}(\vec{k}') \rho_{AB}^{-1}(\vec{k}')$$

Defining gene regulation networks in term of their observed topological features

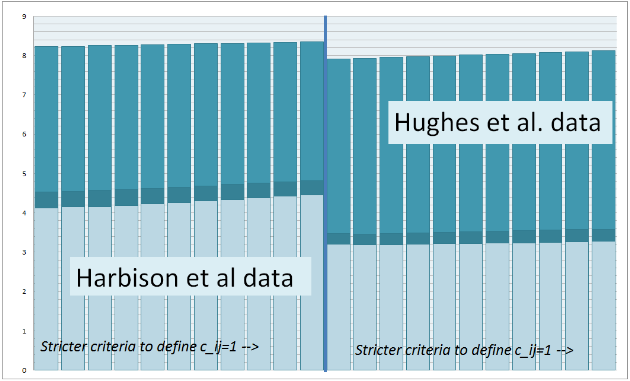


Defining gene regulation networks in term of their observed topological features

Break down of what each bar on the graphic represents



Defining gene regulation networks in term of their observed topological features



Summary

What has been achieved is:

- Exact and explicit formulae for the leading orders in the system size of
    - Shannon entropies and complexities of these ensembles
    - Information-theoretic distances
- for random graph ensembles constrained by a prescribed degree distribution and a prescribed degree degree correlation.

Summary

What has been achieved is:

- Exact and explicit formulae for the leading orders in the system size of
    - Shannon entropies and complexities of these ensembles
    - Information-theoretic distances
- for random graph ensembles constrained by a prescribed degree distribution and a prescribed degree degree correlation.
- Software implementation and initial applications to gene regulation data, in order to present the result as a bioinformatics tool, promoting cross-disciplinary research and dialogue.

Next steps

- Extending the range of properties that can be analysed
  - Generalised degrees
  - Loops
- Finding interesting ways in which to apply the tools developed to real biological problems

Acknowledgements

Thanks to King's College London colleagues.

Mathematics Department: Disordered Systems Group

- Ton Coolen

Randall Division of Cell and Molecular Biophysics

- Franca Fraternali

BBSRC bioscience for the future

KING'S College LONDON

Acknowledgements

Thanks to King's College London colleagues.

Mathematics Department: Disordered Systems Group

- Ton Coolen
- Alessia Annibale

Randall Division of Cell and Molecular Biophysics

- Franca Fraternali
- Luis Fernandes

BBSRC bioscience for the future

KING'S College LONDON

Acknowledgements

Thanks to King's College London colleagues.

Mathematics Department: Disordered Systems Group

- Ton Coolen
- Alessia Annibale

Randall Division of Cell and Molecular Biophysics

- Franca Fraternali
- Luis Fernandes

- Also Thomas Schlitt and Benjamin Lehne (Department of Medical and Molecular Genetics).

BBSRC bioscience for the future

KING'S College LONDON

References

DIRECTED RESULT:
Roberts, E S , Coolen, A C C, Schlitt, T 2011
http://arxiv.org/abs/1101.6022

UNDIRECTED RESULT:
Annibale A, Coolen A C C, Fernandes L P, Fraternali F and Kleinjung J 2009 *J. Phys. A* **42**(48):485001

APPLICATION:
Fernandes L P *et al.* 2010 *PLoS ONE* **5**(8):e12083

OVERVIEW:
Coolen A C C, Fraternali F, Annibale A, Fernandes L and Kleinjung J 2011 *Handbook of Statistical Systems Biology*

References

Directed result:
Roberts, E S , Coolen, A C C, Schlitt, T 2011
http://arxiv.org/abs/1101.6022

Undirected result:
Annibale A, Coolen A C C, Fernandes L P, Fraternali F and Kleinjung J 2009 *J. Phys. A* **42**(48):485001

Application:
Fernandes L P *et al.* 2010 *PLoS ONE* **5**(8):e12083

Overview:
Coolen A C C, Fraternali F, Annibale A, Fernandes L and Kleinjung J 2011 *Handbook of Statistical Systems Biology*

Questions?

References

DIRECTED RESULT:
Roberts, E S , Coolen, A C C, Schlitt, T 2011
http://arxiv.org/abs/1101.6022

UNDIRECTED RESULT:
Annibale A, Coolen A C C, Fernandes L P, Fraternali F and Kleinjung J 2009 *J. Phys. A* **42**(48):485001

APPLICATION:
Fernandes L P *et al.* 2010 *PLoS ONE* **5**(8):e12083

OVERVIEW:
Coolen A C C, Fraternali F, Annibale A, Fernandes L and Kleinjung J 2011 *Handbook of Statistical Systems Biology*

QUESTIONS? *ekaterina.roberts@kcl.ac.uk*