Clusters, motifs, cross-species correlations in biological networks

Johannes Berg
Institute for Theoretical Physics
University of Cologne, Germany
http://www.uni-koeln.de/∼berg

KITP, March 2007

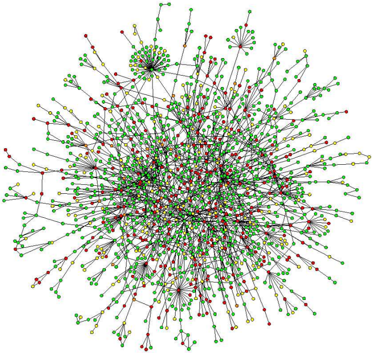
Molecular networks: challenges from experimental data

Protein-protein interactions:

- affinity purification, Co-IP, two-hybrid, structural characterisation

Regulatory interactions:

- mobility shift,ChIP, protein-binding microarrays

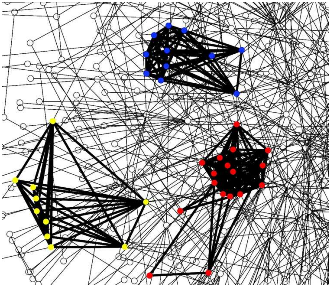


Protein interaction network:
S. cerevisiae, Uetz *et al.* (2000)

Randomness vs. functionality can be addressed beyond sequence data

Analyzing biological networks I:

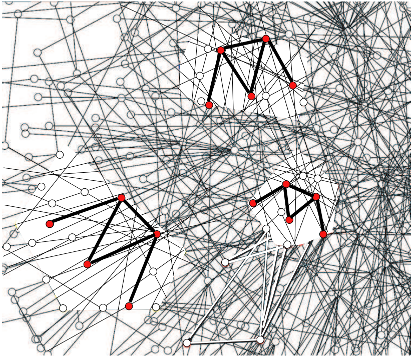
identifying local variations in network statistics: clusters



Domany group
Newman group
Spirin & Mirny

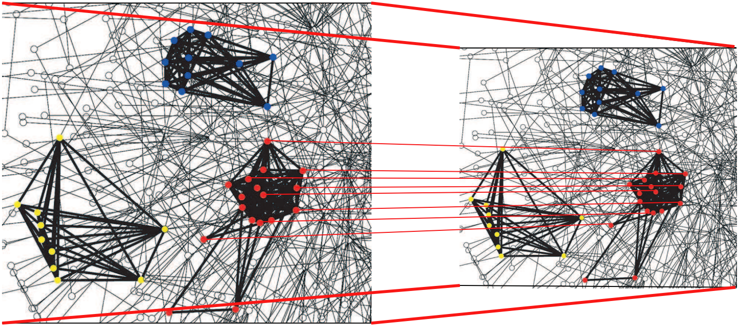
Analyzing biological networks II:

identifying local correlations in networks: network motifs



Alon group
Vergassola and collaborators
Berg & Lässig

Analyzing biological networks III: identifying correlations across species

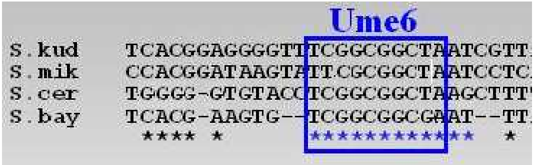


species A

species B

- correlations reflect common evolutionary origin
  - strongly conserved parts: functional core
  - drastic changes: functional innovations
- may be needed to detect orthology if sequence similarity is insignificant

Detour: Sequence analysis



```
S.kud    TCACGGAGGGGTTTCGGCGGCTAATCGTT
S.mik    CCACGGATAAGTATT.CGCGGCTAATCCTC
S.cer    TGGGG-GTGTACCTCGGCGGCTAAGCTTTT
S.bay    TCACG-AAGTG--TCGGCGGCGAAT--TT
           **** *      ************    *
```

Statistical models to detect deviations from genomic background

- model for CpG islands
- model for sequence motifs: position-weight matrices
- model for correlated sequences

Comparison with model for the genomic background

- Log-likelihood score defined as $S = \log \frac{Q}{P}$

Background model

Ensemble of uncorrelated networks with the same connectivities as the data

- probability $w_{ii'}$ of having a link depends on connectivities k_i^- and $k_{i'}^+$
- $w_{ii'} \approx k_i^- k_{i'}^+ / K$

$$P_0(\mathbf{a}) = \prod_{i,i'} (1 - w_{ii'})^{1-a_{ii'}} w_{ii'}^{a_{ii'}}$$

Background model

Ensemble of uncorrelated networks with the same connectivities as the data

- probability $w_{ii'}$ of having a link depends on connectivities k_i^- and $k_{i'}^+$
- $w_{ii'} \approx k_i^- k_{i'}^+ / K$

$$P_0(\mathbf{a}) = \prod_{i,i'} (1 - w_{ii'})^{1-a_{ii'}} w_{ii'}^{a_{ii'}}$$

Feature distinguishing clusters: number of internal links

$$L(\hat{\mathbf{a}}) = \sum_{i,i' \in \mathcal{A}}^{n} \hat{a}_{ii'}$$

Statistics describing network clusters

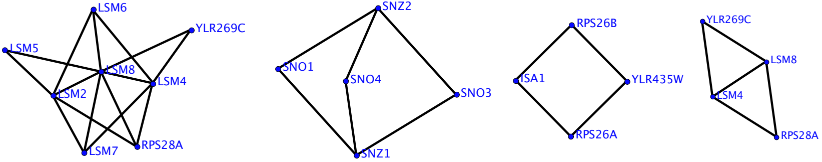
$$Q_\sigma(\hat{\mathbf{a}}) = Z_\sigma^{-1} \exp[\sigma L(\hat{\mathbf{a}})] \, P_0(\hat{\mathbf{a}})$$

Scoring network clusters

Log likelihood score

$$S(\mathcal{A}, \sigma) = \log \left(\frac{Q_\sigma(\hat{\mathbf{a}}(\mathcal{A}))}{P_0(\hat{\mathbf{a}}(\mathcal{A}))} \right) = \sigma L(\hat{\mathbf{a}}(\mathcal{A})) - \log Z_\sigma$$

- positive score indicates likely clusters
- large scores indicate strong deviations from the null model
- compare clusters of different sizes
- determine scoring parameter σ by maximum likelihood



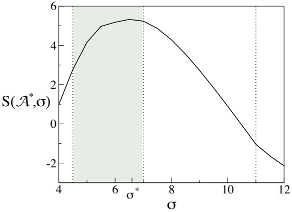
clusters in *S. cerevisiae* Y2H-protein network, data from Uetz *et al.* (2000)

Scoring network clusters

Log likelihood score

$$S(\mathcal{A}, \sigma) = \log \left(\frac{Q_\sigma(\hat{\mathbf{a}}(\mathcal{A}))}{P_0(\hat{\mathbf{a}}(\mathcal{A}))} \right) = \sigma L(\hat{\mathbf{a}}(\mathcal{A})) - \log Z_\sigma$$

- positive score indicates likely clusters
- large scores indicate strong deviations from the null model
- compare clusters of different sizes
- determine scoring parameter σ by maximum likelihood

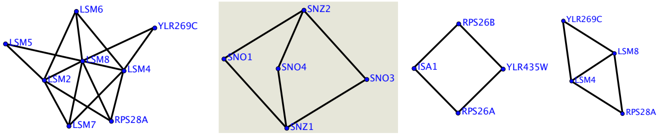


Scoring network clusters

Log likelihood score

$$S(\mathcal{A},\sigma) = \log\left(\frac{Q_\sigma(\hat{\mathbf{a}}(\mathcal{A}))}{P_0(\hat{\mathbf{a}}(\mathcal{A}))}\right) = \sigma L(\hat{\mathbf{a}}(\mathcal{A})) - \log Z_\sigma$$

- positive score indicates likely clusters
- large scores indicate strong deviations from the null model
- compare clusters of different sizes
- determine scoring parameter σ by maximum likelihood

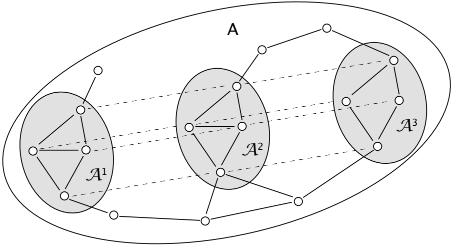


⇑

Cluster of enzymes involved in the metabolism of pyridoxine and thiamin synthesis.

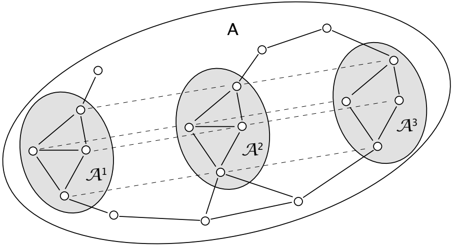
BERG & LÄSSIG, IN PRESS (2007)

Search for network motifs



- patterns occurring repeatedly in the network
- building blocks of information processing [Alon lab]
  - counting of identical patterns: subgraph census
- alignment of topologically similar regions of a network
  - allow for mismatches

Search for network motifs



Consensus motif

$$\bar{\mathbf{a}} = \frac{1}{p} \sum_{\alpha=1}^{p} \hat{\mathbf{a}}^{\alpha}(\mathcal{A})$$

Pairwise pattern mismatch

$$M(\hat{\mathbf{a}}^{\alpha}, \hat{\mathbf{a}}^{\beta}) = \sum_{i,i'=1}^{n} [\hat{a}_{ii'}^{\alpha}(1 - \hat{a}_{ii'}^{\beta}) + (1 - \hat{a}_{ii'}^{\alpha})\hat{a}_{ii'}^{\beta}]$$

Scoring network motifs

- enhanced correlation of subgraphs
- ensemble with enhanced number of links

$$
\begin{aligned}
Q_{\mu,\sigma}(\hat{\mathbf{a}}^1, \ldots, \hat{\mathbf{a}}^p) = {} & Z_{\mu,\sigma}^{-1} \prod_{\alpha=1}^{p} P_0(\hat{\mathbf{a}}^\alpha) \\
& \times \exp\left[-\frac{\mu}{2p} \sum_{\alpha,\beta=1}^{p} M(\hat{\mathbf{a}}^\alpha, \hat{\mathbf{a}}^\beta) + \sigma \sum_{\alpha=1}^{p} L(\hat{\mathbf{a}}^\alpha) \right] .
\end{aligned}
$$

Scoring network motifs

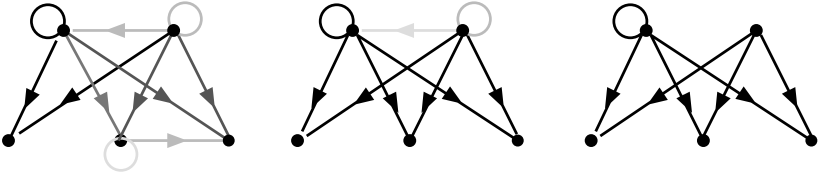
- enhanced correlation of subgraphs
- ensemble with enhanced number of links

$$
\begin{aligned}
S(\hat{\mathbf{a}}^1, \ldots, \hat{\mathbf{a}}^p) &= \log \left(\frac{Q(\hat{\mathbf{a}}^1, \ldots, \hat{\mathbf{a}}^p)}{\prod_{\alpha=1}^{p} P_\sigma(\hat{\mathbf{a}}^\alpha)} \right) \\
&= -\frac{\mu}{2p} \sum_{\alpha,\beta=1}^{p} M(\hat{\mathbf{a}}^\alpha, \hat{\mathbf{a}}^\beta) + (\sigma - \sigma_0) \sum_{\alpha=1}^{p} L(\hat{\mathbf{a}}^\alpha) - \log Z
\end{aligned}
$$

Scoring network motifs

- enhanced correlation of subgraphs
- ensemble with enhanced number of links

Motifs in the *E. coli* regulatory network (Alon data)



$\mu = \mu^* = 2.25$

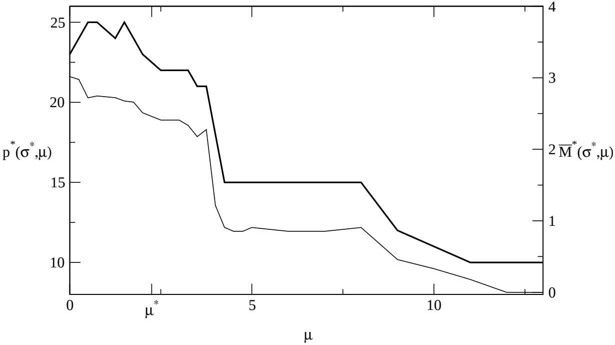
$\mu = 5$

$\mu = 12$

Berg & Lässig, PNAS (2004)

Scoring network motifs

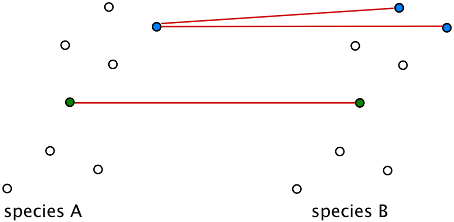
- enhanced correlation of subgraphs
- ensemble with enhanced number of links



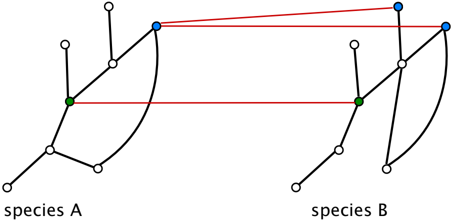
Number of non-overlapping subgraphs and their mismatch versus mismatch penalty μ

BERG & LÄSSIG, PNAS (2004)

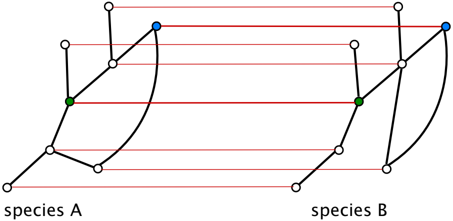
Cross-species network comparison



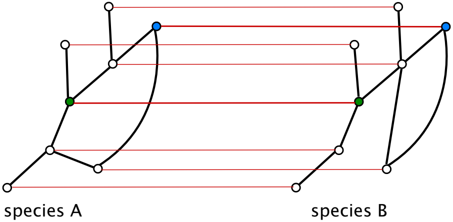
Cross-species network comparison



Cross-species network comparison



Cross-species network comparison



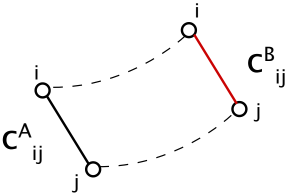
Network alignment: Pairwise association of nodes across species

- networks represented by adjacency matrices A_{ij} and B_{IJ}
- network alignment is a mapping $\pi : i \longrightarrow I$ between nodes in the two networks
- non-trivial interplay between sequence of a gene and position in the network:
  - topology may be conserved even if sequences have diverged [functional constraints]
  - function and position in the network may change with small sequence changes [binding sites]

Scoring network alignments

Evolutionary dynamics of links

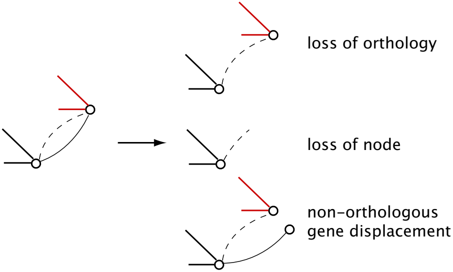
- model of correlated networks: $Q(a, b)$
- model of uncorrelated networks: $P_A(a)\ P_B(b)$



Scoring network alignments

Evolutionary dynamics of nodes

- Gain and loss of genes
- Loss of mutual sequence similarity
- Recruitment of a gene into a new function



Scoring network alignments

Evolutionary dynamics of nodes

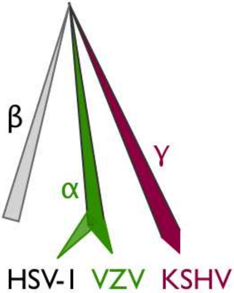
- Gain and loss of genes
- Loss of mutual sequence similarity
- Recruitment of a gene into a new function

Log-likelihood score

$$S(\mathcal{A}) = S^{\text{topo}}(\mathcal{A}) + S^{\text{node}}(\mathcal{A})$$

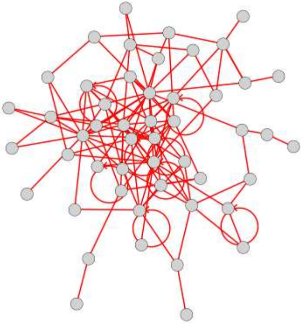
BERG & LÄSSIG, PNAS (2006)

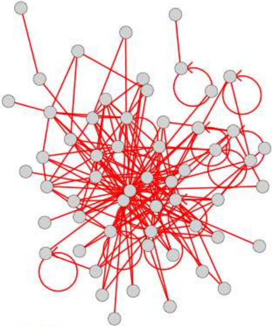
Herpes genomics



- high rates of mutation and gene turnover
- ORFs are short ($\sim$ 100*aa*)
- sequence homologs have $\sim$ 20% aa sequence identity

Herpes interactomics



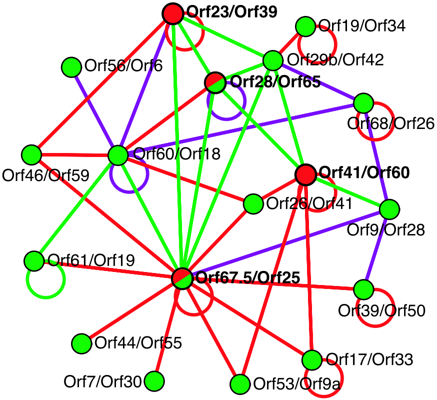


KSHV
84 ORFs, 50 in the graph
124 links

VZV
76 ORFs, 57 in the graph
173 links

(Uetz et al., Science 2006)

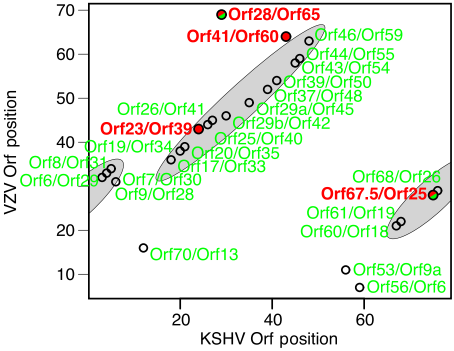
Herpes interactomics



Alignment of KSHV/VZV: matching links and sequence homologs are shown green

KOLÁŘ, LÄSSIG & BERG (2007)

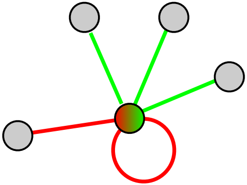
Herpes interactomics



Genomic position supports alignment

Functional predictions from network alignment

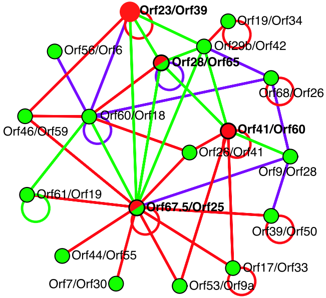
```
Query= KSHV-Orf28, Length= 102
Sbjct= VZV-Orf65,  Length= 102

Global alignment: Score = -0.36, Length = 101, Identity = 10%

Q: SMTSPSPVTGGMVDGSVLVRMATKPPVIGLITVLFLLVIGACVYCCIRVFLAARLWR...
   +  +  +    ++   ++V   + P    + ++        C+Y      +A+ + R...
S: AGQNTMEGEAVALLMEAVVTPRAQPNNTTITAIQPSRSAEKCYYSDSENETADEFLR...
```



Alignment of KSHV Orf 28 – VZV Orf 65

- low sequence similarity of 10% over 100aa
- proteins aligned though link overlap: 3 matching links (p-value 10^{-3})
- functional prediction of KSHV Orf 28 as a virion protein
- prediction consistent with gene expression+mass spectroscopy experiments

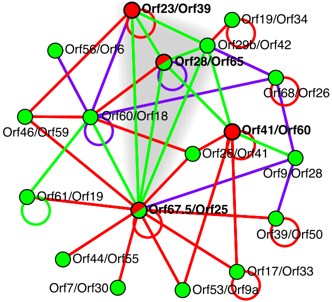
Functional predictions from network alignment



Alignment of KSHV Orf 23 – VZV Orf 39

- no significant sequence similarity
- 3 matching links: alignment due to link similarity (p-value 2×10^{-2})
- functional prediction of KSHV Orf 23 as a membrane glycoprotein
- prediction consistent with gene expression experiments

Functional predictions from network alignment

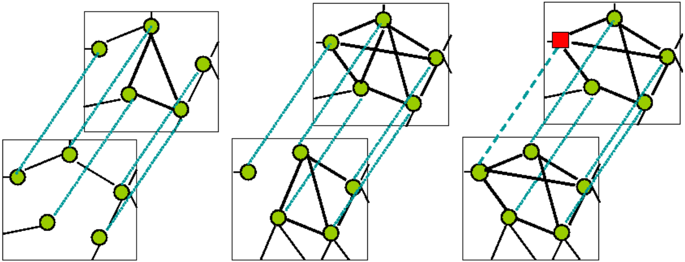


Conserved cluster of interactions between structural genes (virion assembly)

- guilt-by-conserved association: identify modules from selection pressure
- slower link dynamics between nodes of the same function

Outlook

- compare biological networks across a range of evolutionary distances
- application to different types of networks: co-expression, metabolic, protein interaction, regulatory
- infer evolutionary modes



Local moves: Link Dynamics node loss node replacement

Acknowledgements

People

- Michael Lässig
- Michal Kolář
- Jörn Meier

Publications

- J. Berg and M. Lässig, ”Local graph alignment and motif search in biological networks”, *Proc. Natl. Acad. Sci. USA*, **101** (41) 14689-14694 (2004)
- J. Berg and M. Lässig, ”Alignment of biological networks”, *Proc. Natl. Acad. Sci. USA*, **103** (29), 10967-10972 (2006)
- J. Berg and M. Lässig, “Bayesian analysis of biological networks: clusters, motifs, cross-species correlations”, in Statistical and Evolutionary Analysis of Biological Network Data, M. Stumpf and C. Wiuf (Eds.), Imperial College Press, in press.
- M. Kolář, M. Lässig, J. Berg, “Detecting functional and evolutionary relationships by aligning protein interaction networks”, in preparation

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft
DFG

SFB 680
Molecular Basis of
Evolutionary Innovations

STIFTUNG MERCATOR