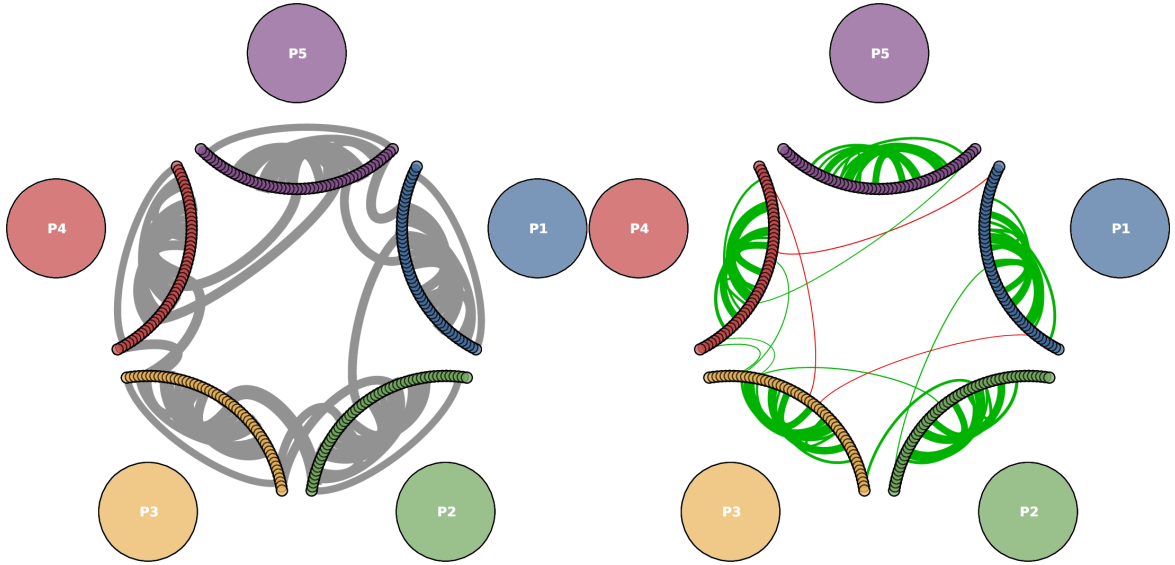
Protein Interaction Networks and Community Link Detection



Christoph Feinauer

Politecnico di Torino

07/07/2014 Cortona

This Talk

Review of Protein Contact Inference

- The Data
- The Potts Model for Proteins
- Inference with Pseudo-Likelihood

Extending to Protein Interaction Networks

- Data Generation by Sequence Matching
- The Contact Score

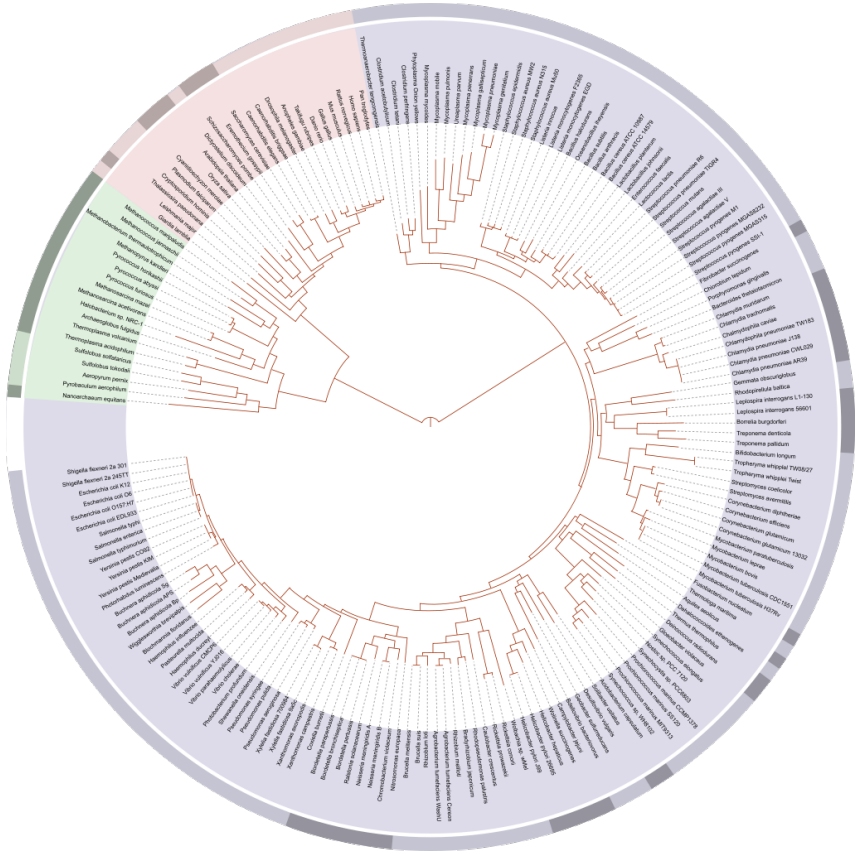
Application to the Ribosome and Artificial Data

- The Small Ribosomal Subunit
- Artificial Data

- Proteins are Sequences of Amino Acids

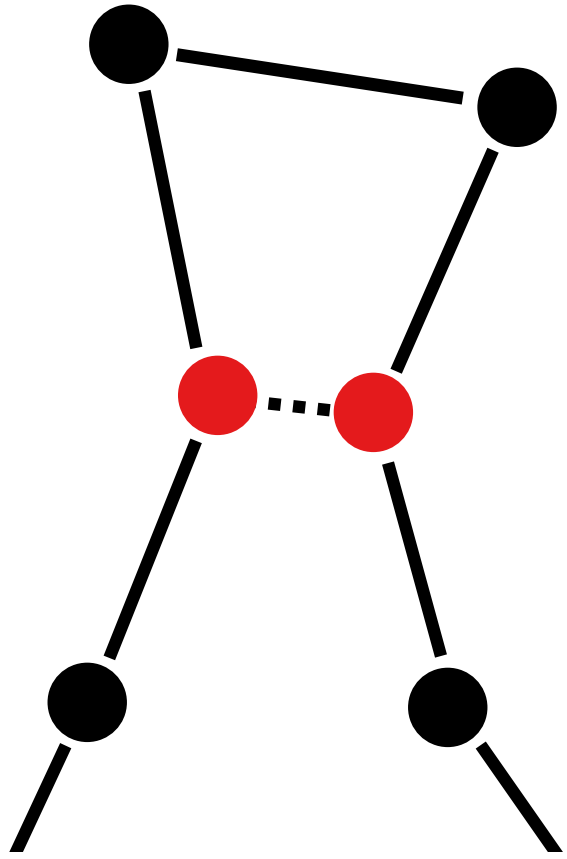
```
        10         20         30         40         50
MAHIEKQAGE LQEKLIAVNR VSKTVKGGRI FSFTALTVVG DGNGRVGFGY

        70         80         90        100        110
QKAMEKARRN MINVALNNGT LQHPVKGVHT GSRVFMQPAS EGTGIIAGGA

       130        140        150        160
HNVLAKAYGS TNPINVVRAT IDGLENMNSP EMVAAKRGKS VEEILGK
```

- Almost (?) every cell on earth has ribosomes



Source: Letunic I and Bork P (2006) Bioinformatics 23(1):127-8 Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation

...NNFKTFNECETKCC...
...NRFNTKSECEKTCV...
...NRFNTKSECKSECV...
...NRFNTKSECEKTCV...
...NNFVHKKHCIKMCM...
...NNFDTQEDCEASCK...
...NNFDTQEDCEASCK...
...NNFDTQEDCEASCK...
...NNFATREDCEGYCG...
...NNFASREECISVCC...
...NNFKNLEECEQQCG...



The Potts Model for Proteins

$$P(\underline{a}) = \frac{\exp\left(\sum_{i<j} J_{ij}(a_i, a_j) + \sum_i h_i(a_i)\right)}{Z}$$

- Maximum Entropy Model
- Sufficient Statistics, frequencies $f_i(a)$ and $f_{ij}(a, b)$

- $J_{ij}(:,:)$ is a 21×21 matrix describing the interaction of site i and j
- Empirically best interaction score: $\sqrt{\sum_{a,b} J_{ij}^2(a,b)}$

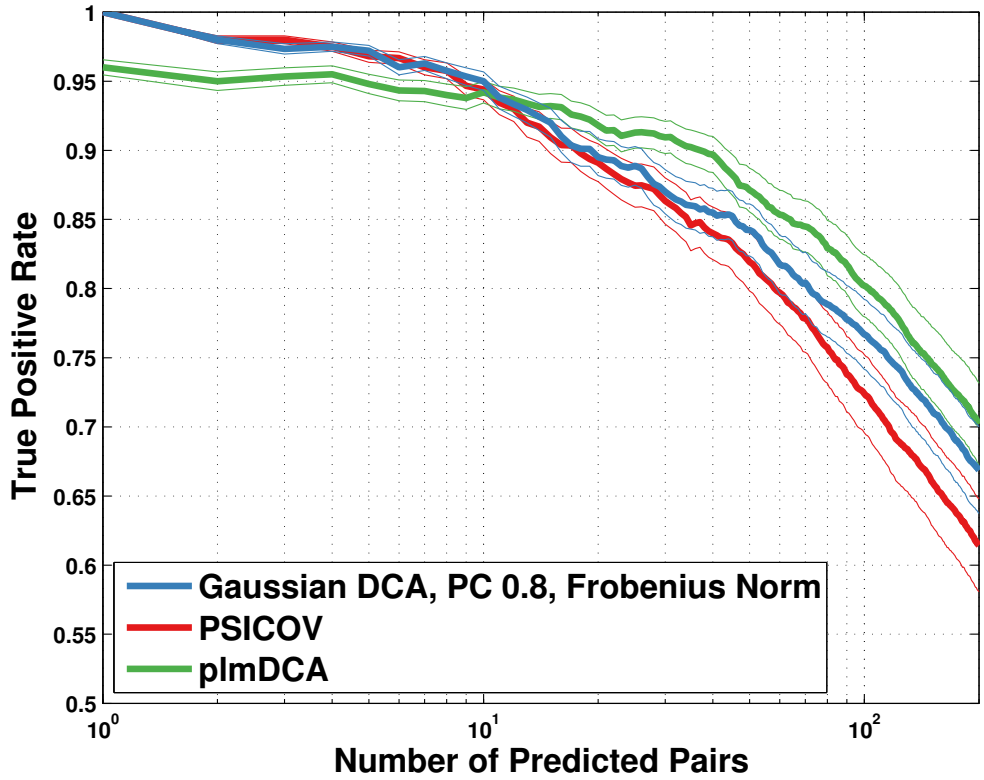


Figure: *Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, et al. (2014) Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. PLoS ONE 9(3)*

Maximum-Likelihood

What you would actually like to do:

$$\{J^*, h^*\} = \underset{h,J}{argmin} \left[-\frac{1}{M} \log P(D|\{J,h\} \right]$$

$$= \underset{h,J}{argmin} \left[-\sum_{i<j,a,b} J_{ij}(a,b) f_{ij}(a,b) - \sum_{i,a} h_i(a) f_i(a) + \ln Z \right] \quad (1)$$

- Intractable for any reasonable system size

Pseudo-Likelihood Maximization

$$\underset{h_r, J_r}{argmin} \left[-\frac{1}{M} \sum_{m=1}^{M} \log P_r \left(a_r^m | \{a_{i \neq r}^m\}, \{J_r, h_r\} \right) + \text{ Prior } \right], \quad (2)$$

- Computationally efficient
- Consistent
- Uses all the data
- In Protein Contact Inference an $l2$-prior is used

M. Ekeberg, C. Lvkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models, Phys. Rev. E 87, 012707 (2013)

Work in Progress: How to measure interaction?

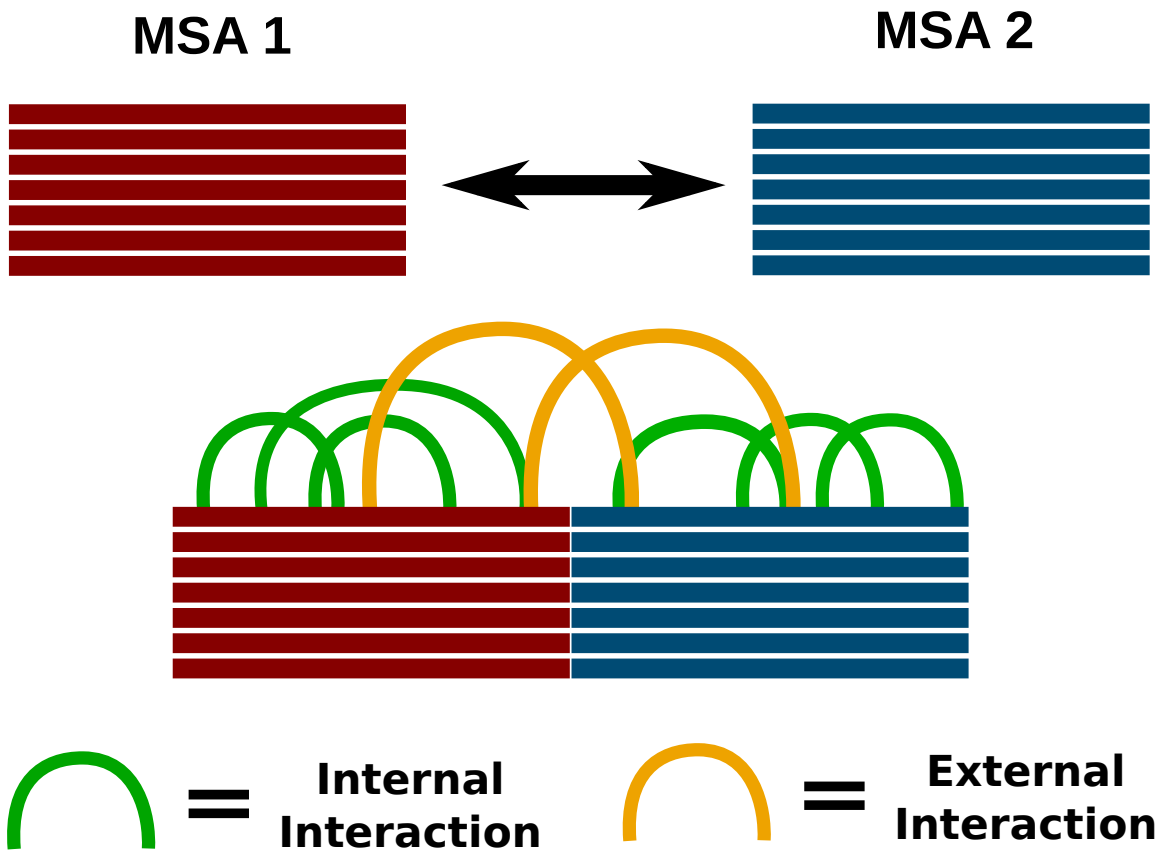
- Gauge Freedom: $\{h, J\} \rightarrow \{p(a)\}_a \leftarrow \{h', J'\}$
- Couplings related by a gauge transformation are physically equivalent
- PL is gauge invariant
- PL $+ l2$ is not gauge invariant

An gauge-invariant Prior

$$\boxed{J_{ij}(a,b) - J_{ij}(a,\eta) - J_{ij}(\xi,b) + J_{ij}(\xi,\eta)}$$

- This quantity is zero if and only if the model is non-interacting
- Works better or worse than $l2$ dependent on the choice of ξ and η
- We have no method of choosing the best one

Data Generation by Sequence Matching



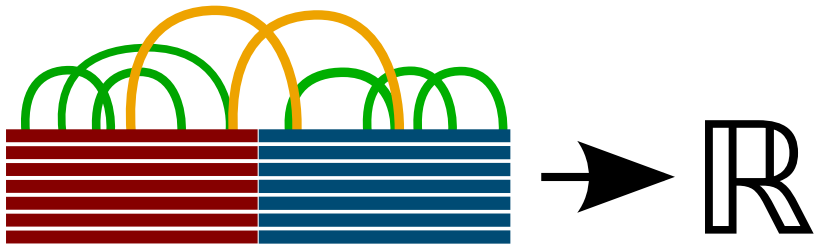
Data Generation by Sequence Matching

- If possible: Matching by Uniqueness
- Many species have several members of a family

Matching 1			**Matching 2**		
Sequence		Partner	Sequence		Partner
A_1	$\rightarrow$	B_1	A_1	$\rightarrow$	B_2
A_2	$\rightarrow$	B_2	A_2	$\rightarrow$	B_1

- Biological approach: Interaction partners are often close on the genome

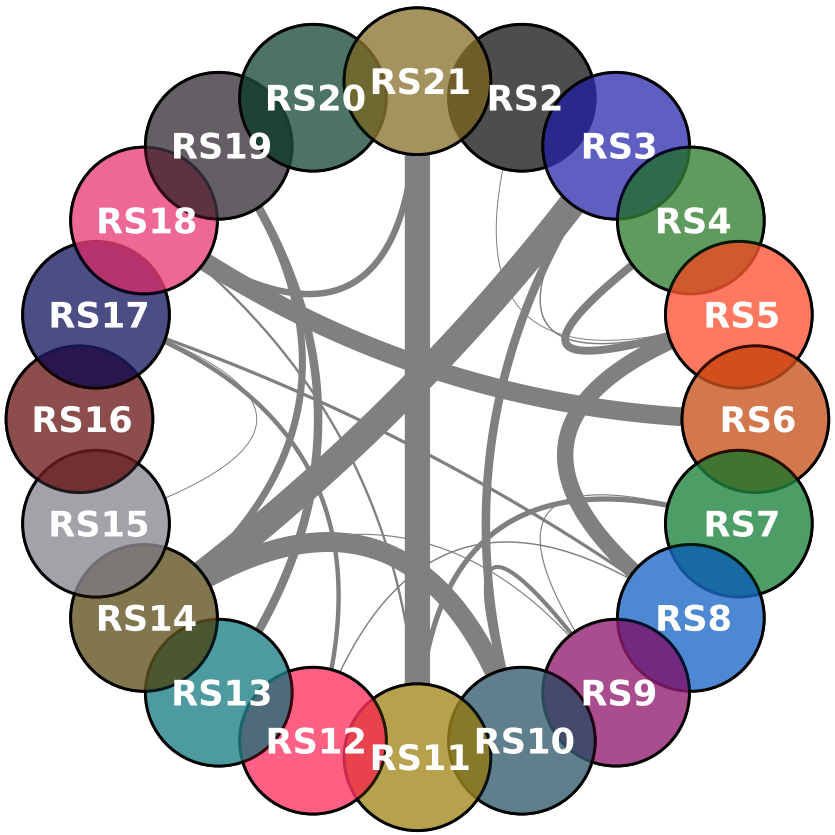
Contact Score



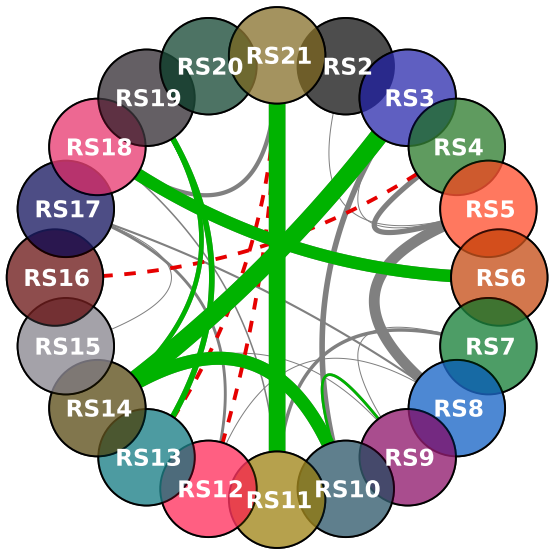
- Contact score between proteins, not positions
- Empirically: Take the mean of the 4 largest scores between the alignments

Small Ribosomal Subunit

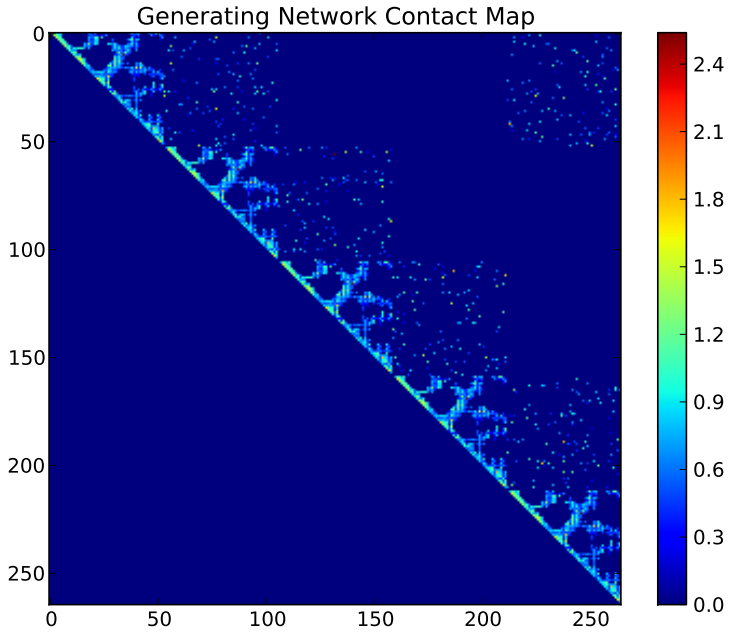
Small Ribosomal Subunit, Full Network



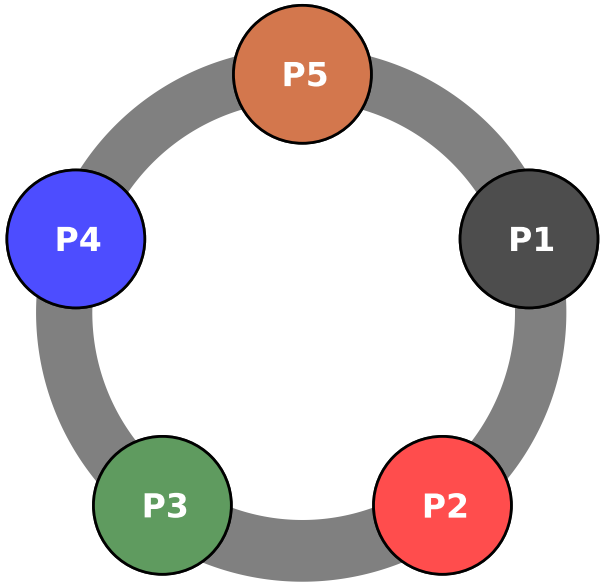
Small Ribosomal Subunit, 10 First Predictions Colored Network



Generating our own data



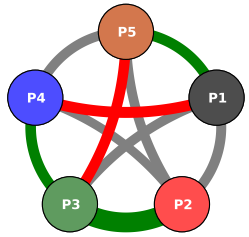
Generating Network



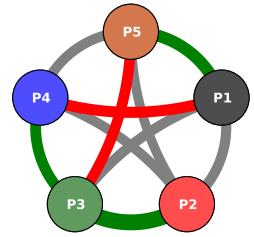
- 5 proteins concatenated
- Random subset of internal couplings as interaction couplings

Combined and Paired Analysis

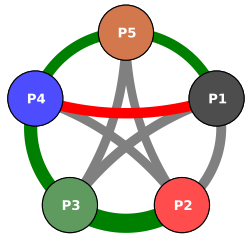
Inferred Network, Combined Analysis, 2000 Sequences



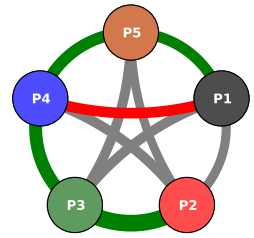
Inferred Network, Paired Analysis, 2000 Sequences



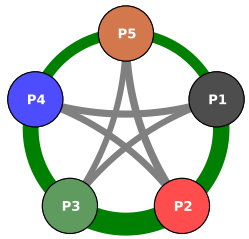
Inferred Network, Combined Analysis, 4000 Sequences



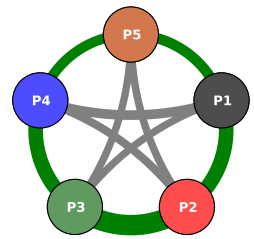
Inferred Network, Paired Analysis, 4000 Sequences



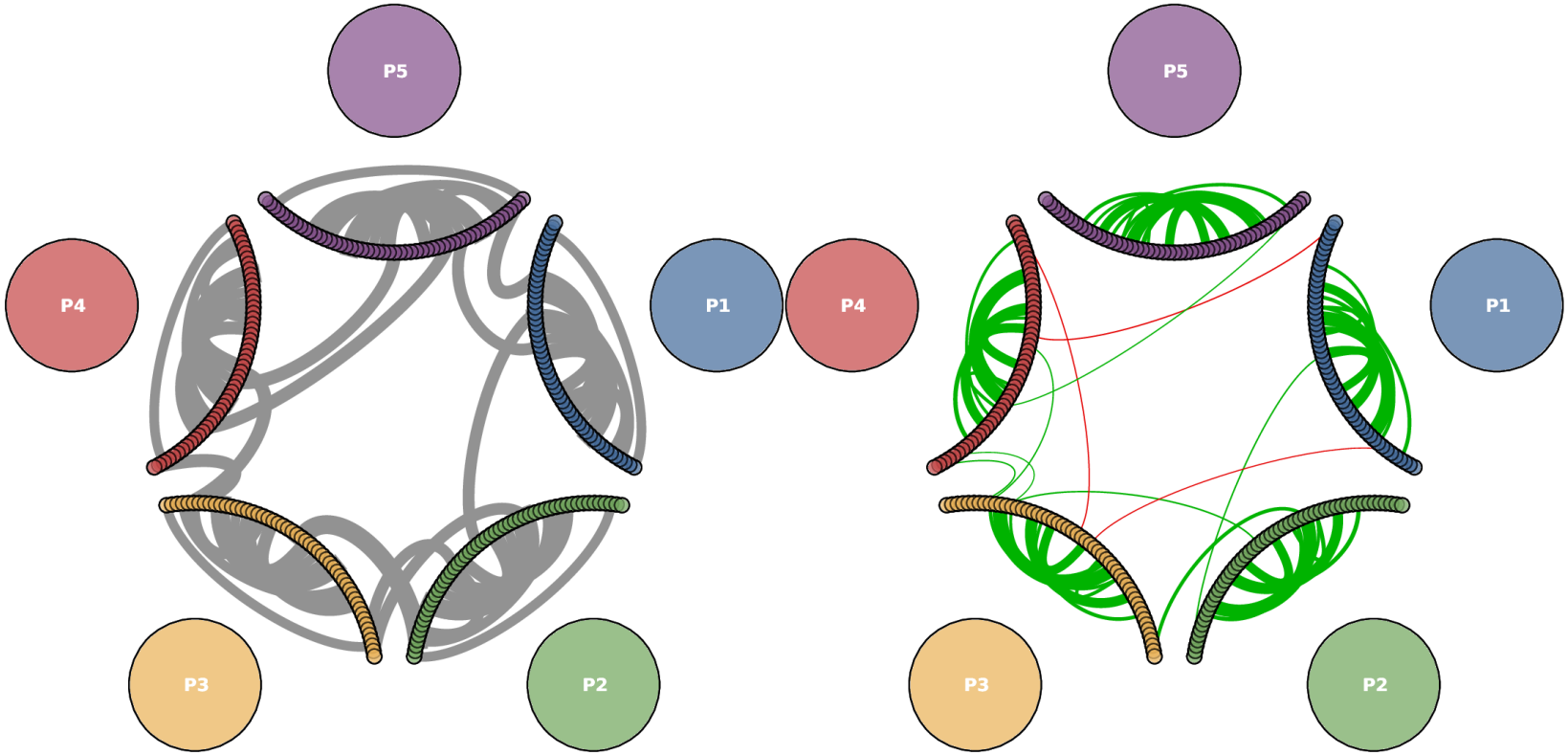
Inferred Network, Combined Analysis, 8000 Sequences



Inferred Network, Paired Analysis, 8000 Sequences



Internal and External Analysis



- The predicted scales for the external and internal contacts are different

Work in Progress: Randomizing a System

- Data seems to indicate that randomizing the links in the generating system makes inference harder