

# Statistical Physics Approaches to High-Dimensional Inference

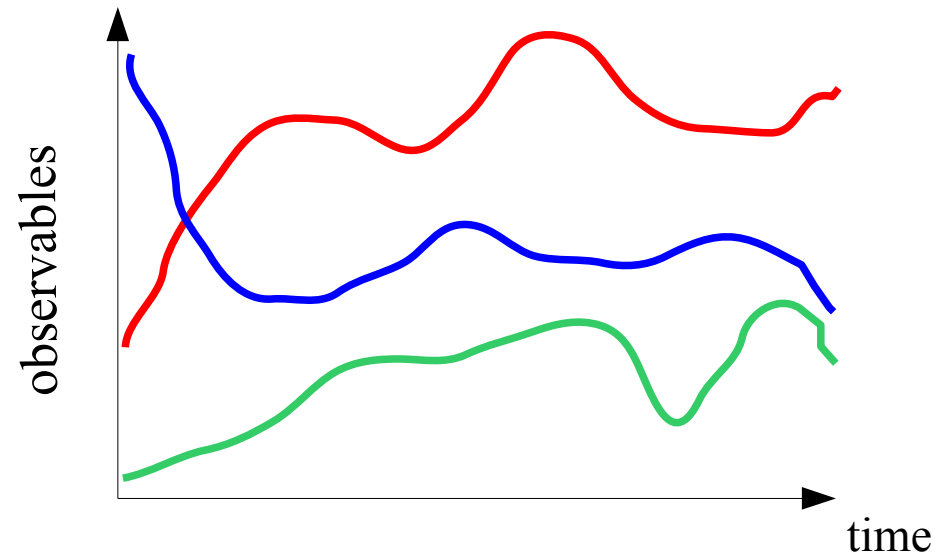
*applications to biological data*

Rémi Monasson

*Laboratory of Theoretical Physics,  
CNRS & Ecole Normale Supérieure, Paris*

Winter School on Quantitative Biology, ICTP, December 2012

# Model inference from observed data



**Many issues :** limitation over temporal and spatial sampling, noise (measurement, dynamics), stationarity, classes of models (number of parameters), computational effort for inference, ...  
(signal/noise < 1, large systems)

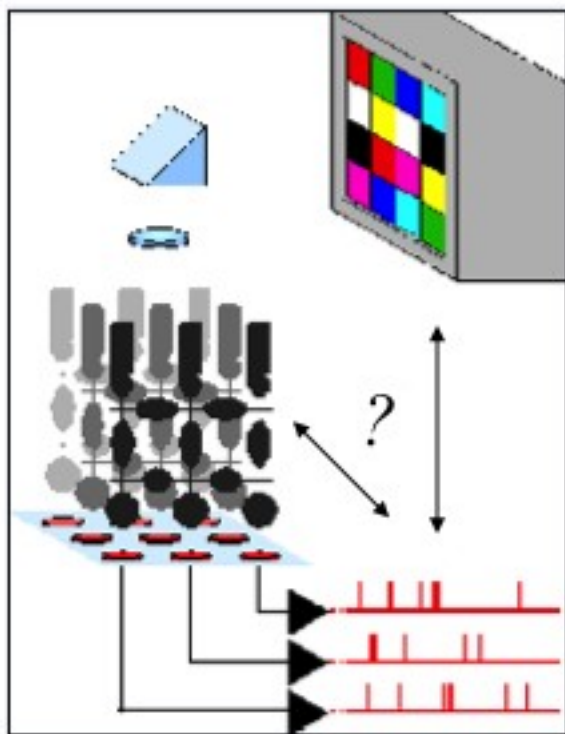
$\neq$  Physics (uniform interactions  $\rightarrow$  low dimensional models, reproducibility  $\rightarrow$  good sampling, thermal equilibrium)

- **Today:** Theoretical framework for model inference  
(special case: many interacting & stationary variables)  
Mean-field inference  
*Applications covariation in protein families (I)*
- **Wednesday:** Issues & advanced statistical physics methods  
Inverse Hopfield-Potts model & Random Matrix Theory  
*Applications to neural data (I), covariation in proteins (II)*
- **Thursday:** Case of interacting & non-stationary variables  
*Applications to neural data (II)*  
*to ecological systems*
- **Now:** Brief overview of the biological/biophysical systems and inference problems

*Example 1:*

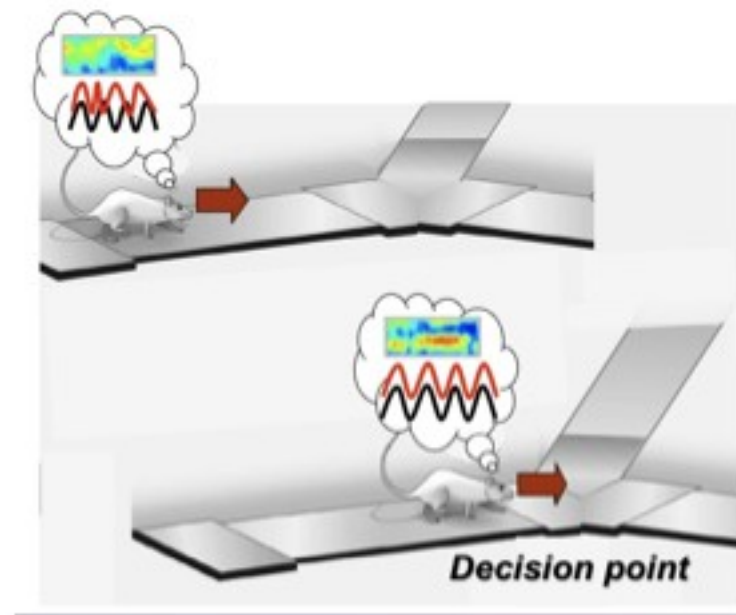
## Concerted activity of a neural population

In vitro recording on retina

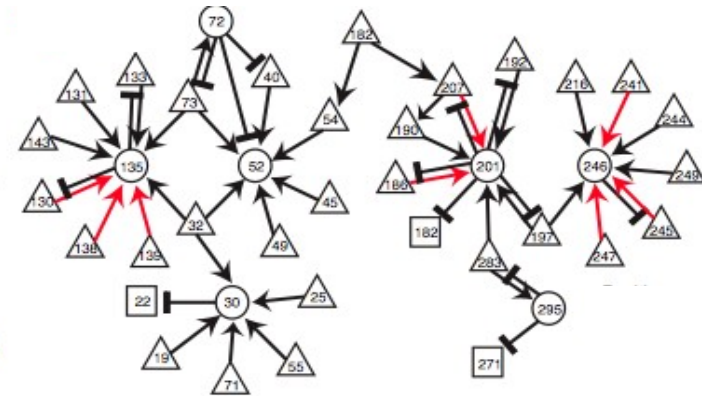
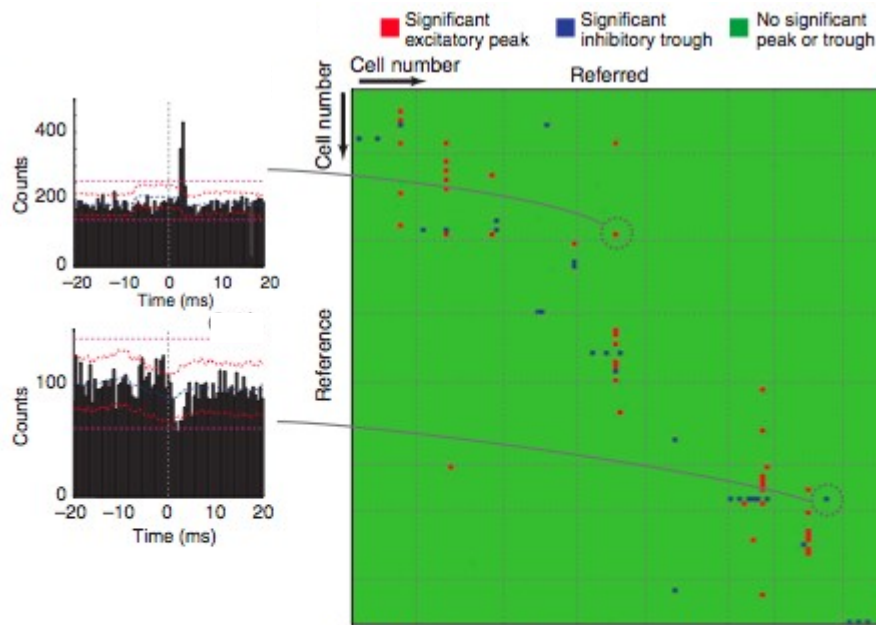


Schnitzer, Meister (2003)  
Schneidman et al. (2006)

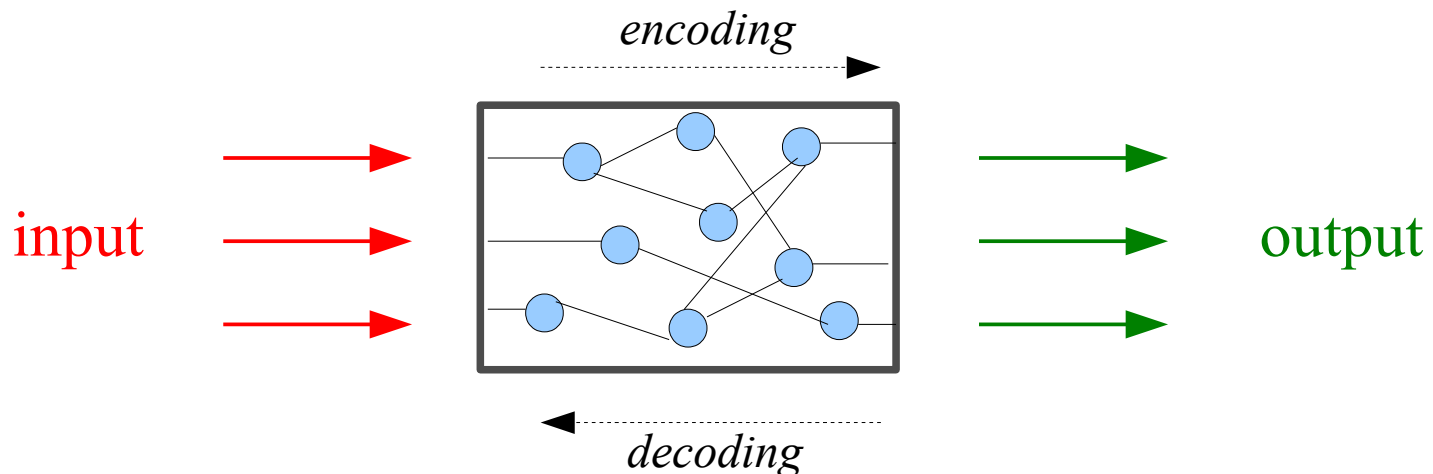
Cortical recording of a behaving rat



Fujisawa, Amarasingham,  
Harrison, Buzsaki (2008)

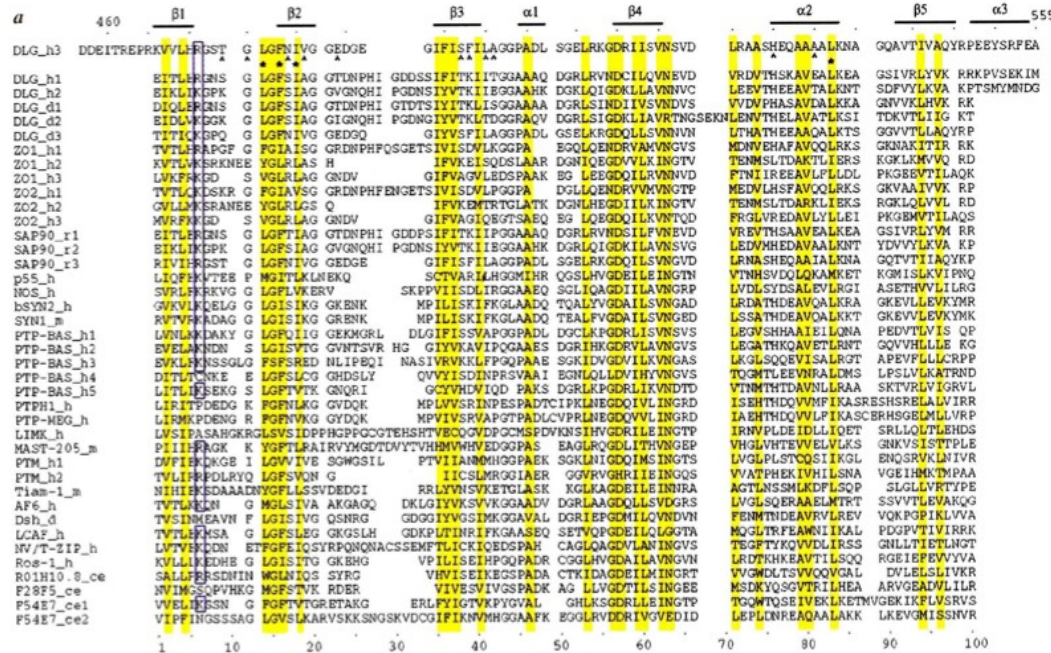


- Network depend on activity (functional connections)
- Connections can be modified through learning ...
- More sophisticated methods to infer effective connections for encoding/decoding:

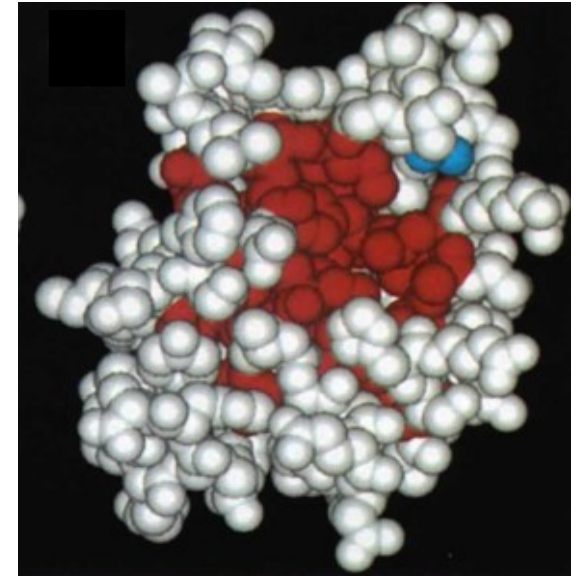


## Example 2:

# Coevolution of residues in protein families



## PDZ domains



Morais Cabral et al. (1996)

- Conservation of residues (used for homology detection, phylogeny reconstruction)
- Two-residue correlations ? could reflect structural and functional constraints ...

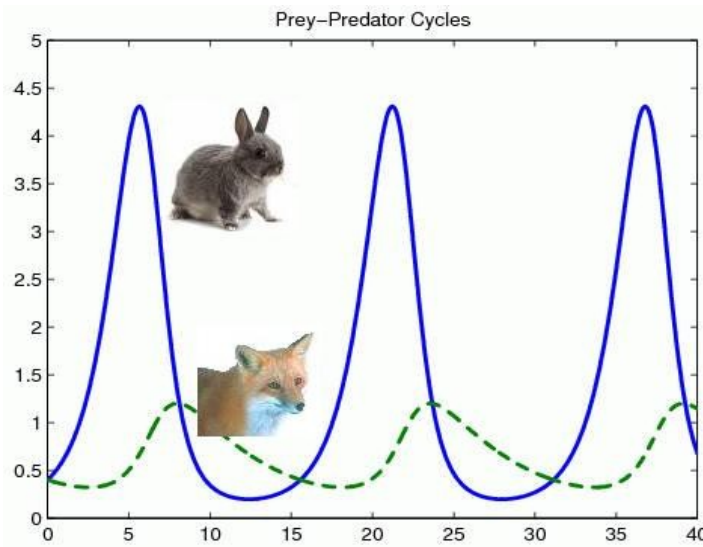


Gobel et al. (1994)

### Example 3:

## Coupled dynamics of species in an ecological system

Population ecology : interactions between species (existence, additivity, ...)



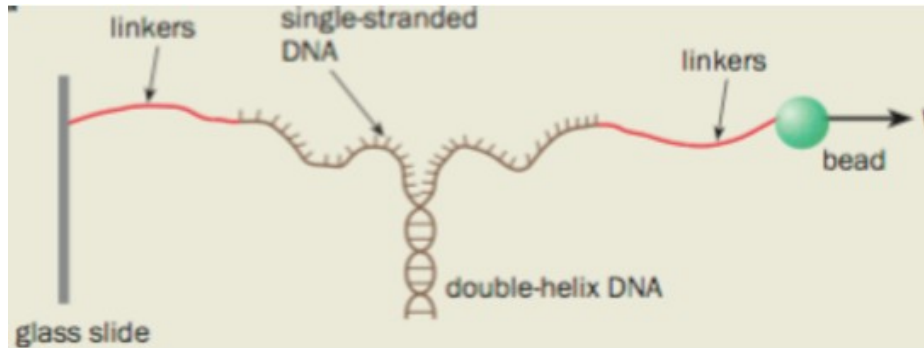
**Lotka-Volterra equations:**

$$\frac{dN_i}{dt} = N_i \left( r_i - \sum_j A_{ij} N_j \right)$$

**Issues :** Measurement noise, dynamical noise,  
limited number of samples, unknown (not measured) species, ...

Is there any reliable signal about species-species 'interactions'? (additivity?)  
Exploit interactions to predict dynamics, extinction ?

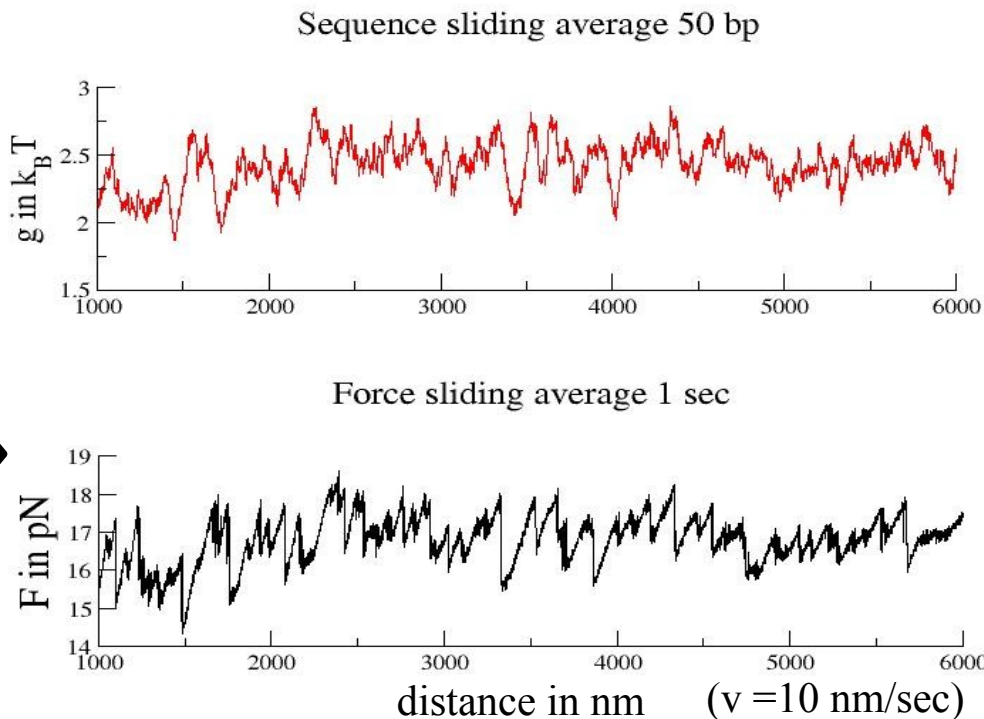
# Example 4: Unzipping dynamics of a single DNA molecule



Bockelmann, Heslot (1996)

Huguet,  
Ritort (2010)

direct  
problem

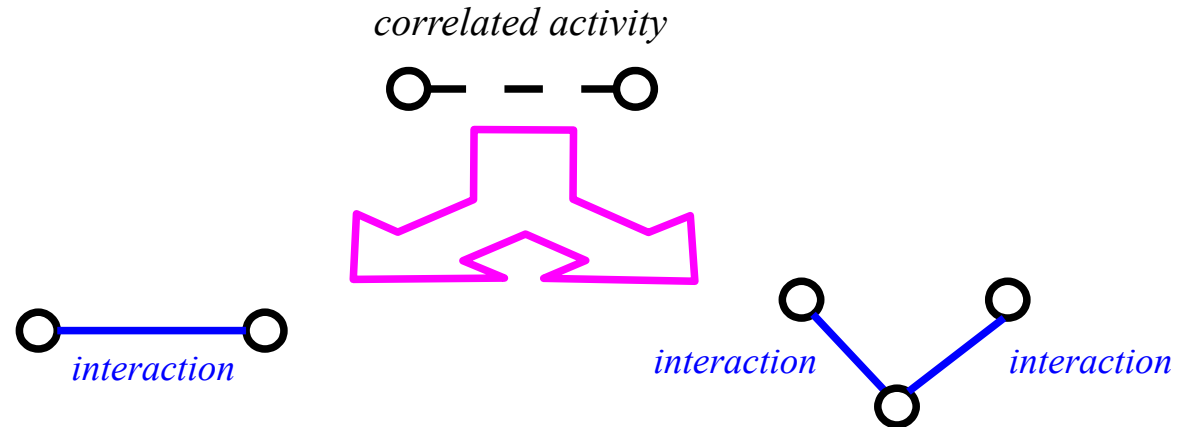


inverse  
problem

- **Today:** Theoretical framework for model inference  
(special case: many interacting & stationary variables)  
Mean-field inference  
*Applications to neural data (I)*
- Wednesday: Advanced issues & statistical physics methods  
Inverse Hopfield-Potts model & random matrix theory  
*Applications to covariation in protein families*
- Thursday: Case of interacting & non-stationary variables  
*Applications to neural data (II)*  
*to ecological systems*

# Goals

- Compression of data (eliminate indirect correlations, sparser representation?)



- Find effective interaction network  
(ex: contact map in protein residue case)
- Obtain predictive, generative models  
(ex : model for artificial protein sequences)  
Could be used to test effect of perturbation ...  
to define 'energy' landscape and probe configuration space ...
- Feedback with experiments : design of optimal, maximally informative protocols

# Microscopic model for the data (1)

(here, stationary and discrete data)

**Data**

(1,0,0,0,1,0,1,1, ..., 1,0,0,1,1,0)  
(0,1,0,0,0,1,1,1, ..., 0,1,1,0,0,0)  
(1,1,0,1,0,1,1,0, ..., 1,1,0,0,1,1)  
...  
(0,1,0,0,1,0,0,0, ..., 0,0,0,1,0,0)

$\Rightarrow$  Probability  $p(\sigma_1, \sigma_2, \dots, \sigma_N)$  ?

**Constraints**

$$m_i = \langle \sigma_i \rangle, \quad c_{ij} = \langle \sigma_i \sigma_j \rangle, \quad \dots$$

(constraints are realizable)

**Maximum entropy principle** (Jaynes, 1957)

Find  $p(\boldsymbol{\sigma})$  maximizing the entropy  $S[p] = - \sum_{\boldsymbol{\sigma}} p(\boldsymbol{\sigma}) \ln p(\boldsymbol{\sigma})$   
under the selected constraints

# Microscopic model for the data (2)

## Analogy with Thermodynamics and Ensembles in Statistical Physics

- System with energy  $E$ , volume  $V$ ,  $N$  particles, ...
- Fix average value of volume  $V \leftrightarrow$  impose pressure  $p : E \rightarrow E + pV$   
number of particles  $N \leftrightarrow$  impose chemical potential  $\mu : E \rightarrow E - \mu N$

**Model** •  $E(\sigma; J, h) = - \sum_{i < j} J_{ij} \sigma_i \sigma_j - \sum_i h_i \sigma_i$

Ising model!

•  $p^{\text{MAXENT}}(\sigma; J, h) = \exp(-E(\sigma; J, h)) / Z[J, h]$

where  $Z[J, h] = \sum_{\sigma} \exp(-E(\sigma; J, h))$

- find couplings and fields such that all  $N + N(N-1)/2$  constraints are fulfilled

$$\frac{\partial \ln Z[J, h]}{\partial h_i} = m_i, \quad \frac{\partial \ln Z[J, h]}{\partial J_{ij}} = c_{ij}$$

# Boltzmann Machine Learning

Ackley, Hinton, Sejnowski (1985)

- Start from random  $J_{ij}$  and  $h_i$
- Calculate  $\langle \sigma_i \sigma_j \rangle$  and  $\langle \sigma_k \rangle$  using Monte Carlo simulations
- Compare to  $c_{ij}$  and  $m_k$  (data) and update  $J_{ij} \rightarrow J_{ij} - a (\langle \sigma_i \sigma_j \rangle - c_{ij})$   
 $h_k \rightarrow h_k - a (\langle \sigma_k \rangle - m_k)$

**Problems:**

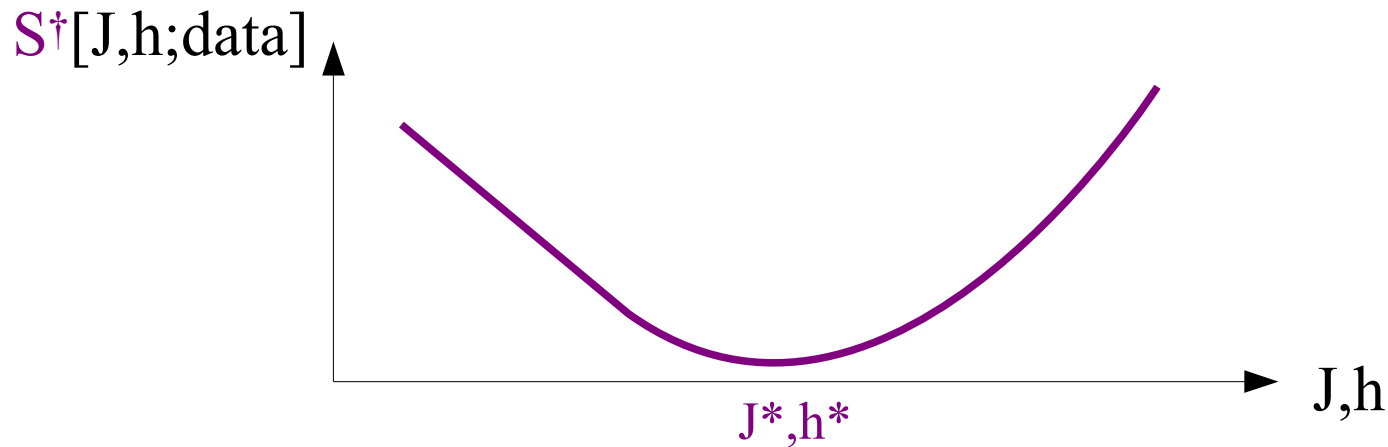
1. issue of thermalization (critical point ?  
may take exponential-in-N time ...)
2. convergence (yes, but flat modes ?)  
  
→ slow

# Microscopic model for the data (3)

Cross-entropy of data ( $= \sigma^1, \dots, \sigma^B$ )

$$S^\dagger [J, h] = \sum_{b=1}^B -\ln p(\sigma^b; J, h)$$

$$= B \left( \ln Z[J, h] - \sum_{i < j} J_{ij} c_{ij} - \sum_i h_i m_i \right)$$



- The minimum of  $S$  is the Ising model we are looking for :

$$S[p] \longrightarrow \quad \longleftarrow S^\dagger [J, h]$$


---


$$S[p^{\text{MAXENT}}] = S^\dagger [J^*, h^*]$$

- The hessian of  $S$  is positive semi-definite, hence  $S$  is convex

# Microscopic model for the data (4)

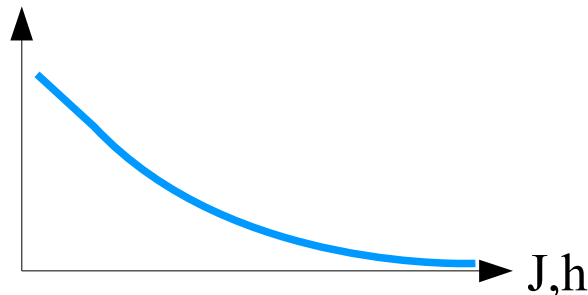
Hessian of the cross-entropy

$$\frac{\partial^2 S(J,h;\text{data})}{\partial (J,h) \partial (J,h)} = \begin{pmatrix} \langle \sigma_i \sigma_j \sigma_k \sigma_l \rangle - \langle \sigma_i \sigma_j \rangle \langle \sigma_k \sigma_l \rangle & \langle \sigma_i \sigma_j \sigma_k \rangle - \langle \sigma_i \rangle \langle \sigma_j \sigma_k \rangle \\ \langle \sigma_i \sigma_j \sigma_k \rangle - \langle \sigma_i \sigma_j \rangle \langle \sigma_k \rangle & \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle \end{pmatrix}$$

where  $\langle \rangle$  = Gibbs average with the Ising model

$$X = \begin{pmatrix} X_{ij} \\ X_k \end{pmatrix} \rightarrow X^T X = \langle \left( \sum_{i < j} X_{ij} (\sigma_i \sigma_j - \langle \sigma_i \sigma_j \rangle) + \sum_k X_k (\sigma_k - \langle \sigma_k \rangle) \right)^2 \rangle$$

Zero modes ?



Non-realizable cases...

# Bayesian inference framework (1)

Data = set of configurations  $\sigma^b$ ,  $b = 1, 2, \dots$ ,  $B = \text{nb. of configs}$

**Bayes formula** 
$$P[ \mathbf{J}, \mathbf{h} \mid \text{Data} ] \propto \prod_b P[ \sigma^b \mid \mathbf{J}, \mathbf{h} ] \times P_0[ \mathbf{J}, \mathbf{h} ]$$

**Prior** 
$$P_0[ \mathbf{J}, \mathbf{h} ] \quad (\text{useful in case of undersampling ...})$$

For instance : 
$$P_0 \propto \exp\left( - \sum_{i < j} J_{ij} / (2J_0^2) \right)$$

**Likelihood** 
$$P[ \sigma \mid \mathbf{J}, \mathbf{h} ] = \exp\left( \sum_{i < j} J_{ij} \sigma_i \sigma_j + \sum_i h_i \sigma_i \right) / Z[\mathbf{J}, \mathbf{h}]$$

## Bayesian inference framework (2)

Posterior Proba  
of  $\mathbf{J}, \mathbf{h}$

$$P[ \mathbf{J}, \mathbf{h} \mid \text{Data} ] \propto \exp( B[ \sum_{i < j} J_{ij} \mathbf{c}_{ij} + \sum_i h_i \mathbf{m}_i ] ) / Z[\mathbf{J}, \mathbf{h}]^B \\ \times P_0[ \mathbf{J}, \mathbf{h} ]$$

Regularized  
Cross-entropy

$$S = - \ln P[ \mathbf{J}, \mathbf{h} \mid \text{Data} ] \\ = B ( \ln Z[\mathbf{J}, \mathbf{h}] - \sum_{i < j} J_{ij} \mathbf{c}_{ij} - \sum_i h_i \mathbf{m}_i ) - \ln P_0[ \mathbf{J}, \mathbf{h} ] \\ = B ( \ln Z[\mathbf{J}, \mathbf{h}] - \sum_{i < j} J_{ij} \mathbf{c}_{ij} - \sum_i h_i \mathbf{m}_i ) + \sum_{i < j} J_{ij} / (2J_0^2 )$$

(with Gaussian prior)

# Analytical approaches

- Mean field inference
- 

- Importance of prior(s)
- Pseudo-likelihood algorithms
- Advanced statistical physics methods
- Inverse Hopfield-Potts model

# Analytical approaches

- Mean field inference
  - Importance of prior(s)
- 
- Pseudo-likelihood algorithms
  - Advanced statistical physics methods
  - Inverse Hopfield-Potts model

# Applications to protein residue covariation (1)

Potts model with 20 (amino-acids) +1 (gap) symbols

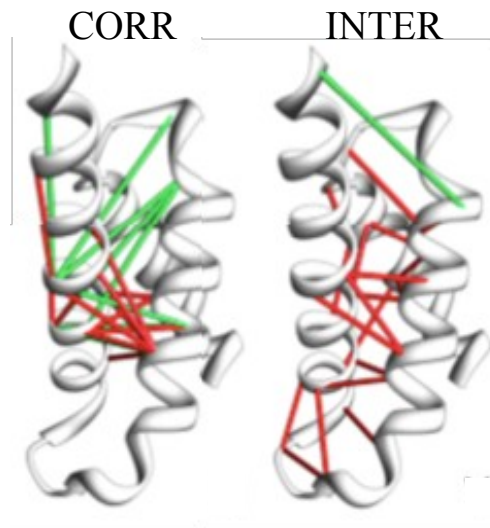
Compute 1- and 2- residues frequencies,  $f_{ia}$  and  $f_{ia,jb}$

Regularization = pseudo-counts ...

Find couplings  $J_{ia,jb}$  from the inversion of correlation matrix  $c_{ia,jb} = f_{ia,jb} - f_{ia} f_{jb}$

Weigt et al. (2009)

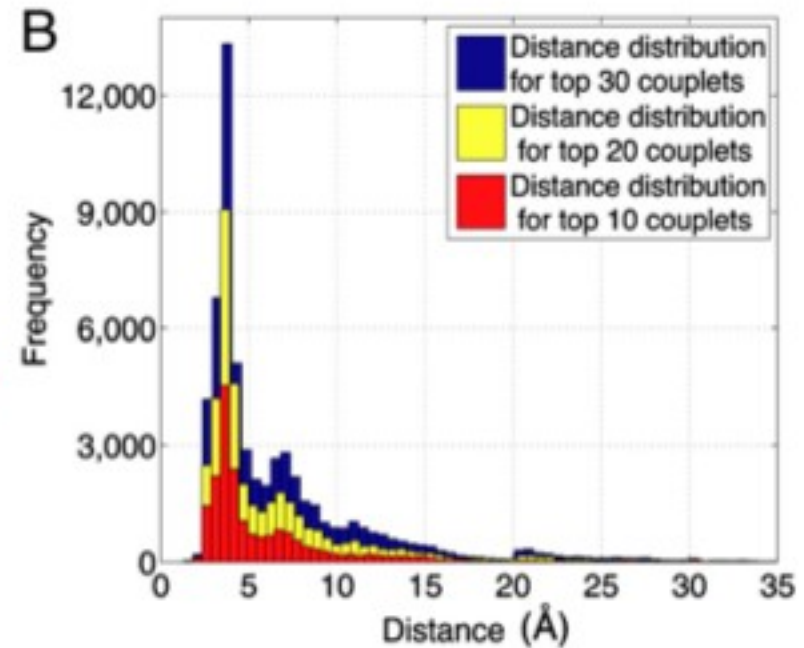
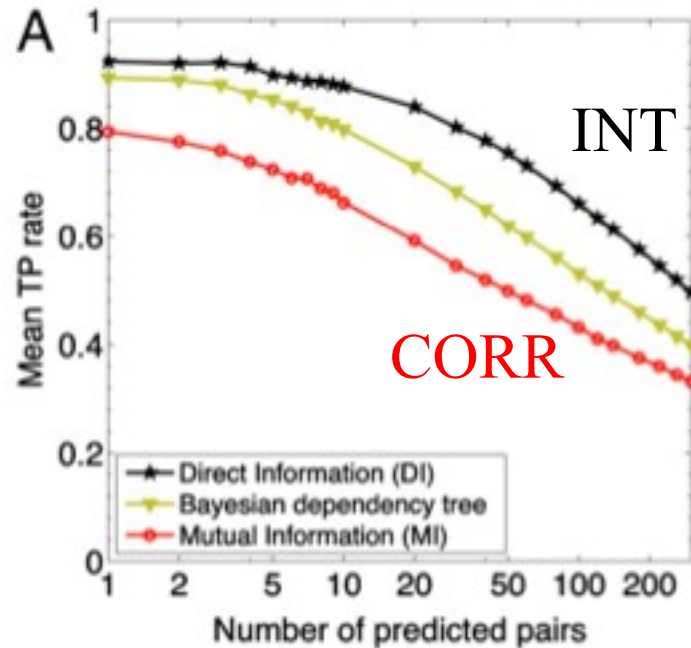
Mapping of  
20 top  
correlations  
and  
interactions



RNA polymerase sigma-70 region 2

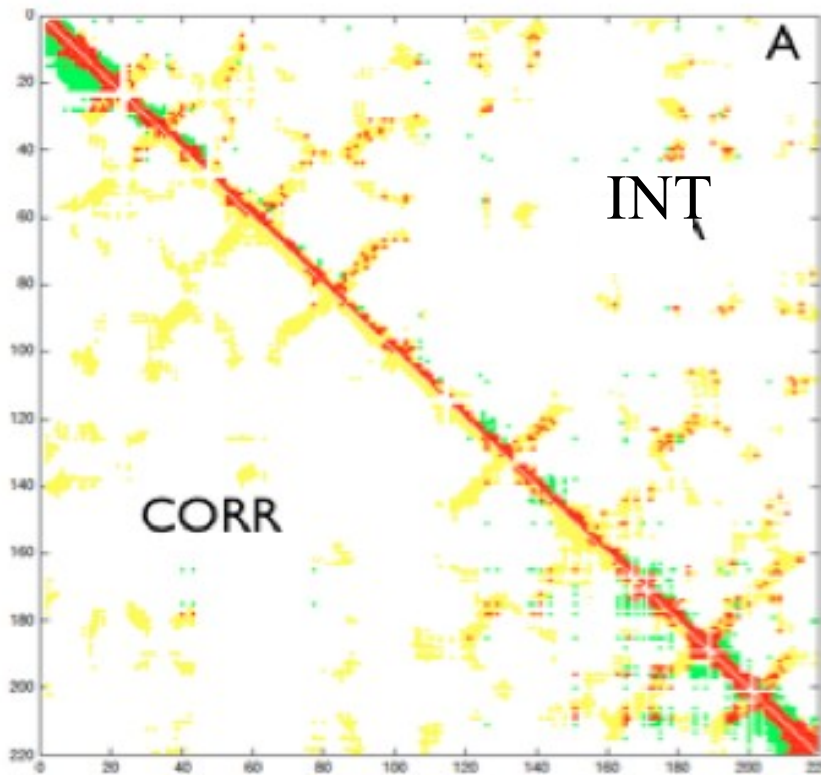
# Applications to protein residue covariation (2)

Morcos et al. (2011)

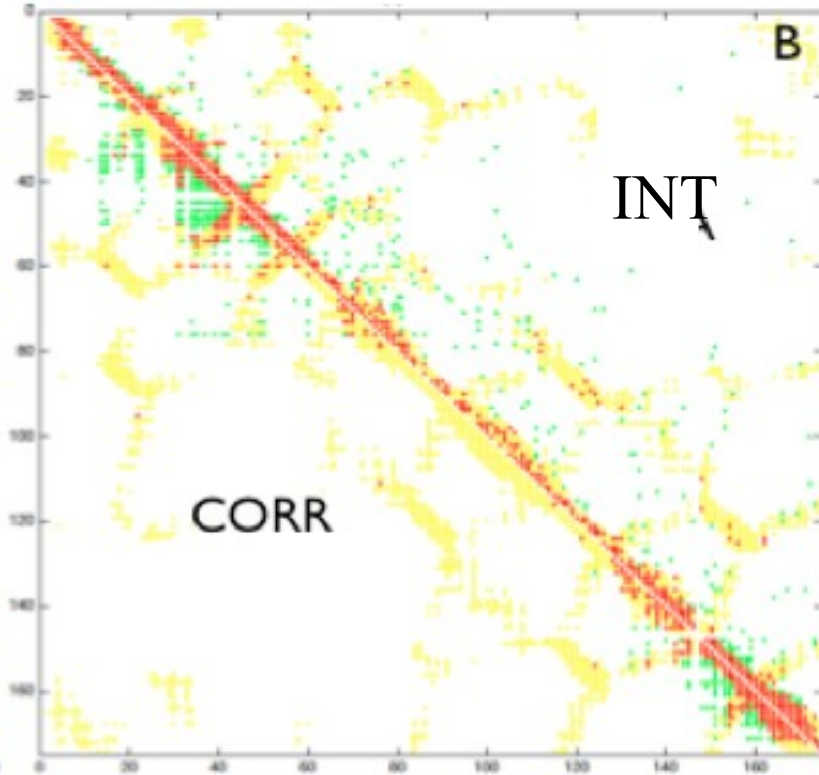


True Positive rate ( $< 8 \text{ \AA}$ )  
over 131 protein families

# Applications to protein residue covariation (3)



Trypsin family PF00089, PDB 3tgi



ADP-ribosylation factor family PF00025, PDB 1fzq

- Contact ( $< 8\text{\AA}$ )
- No contact ( $> 8\text{\AA}$ )

- Issues :
- large number of parameters to be inferred ( $\sim (20 L)^2$ )
  - not always successful (mean field?)
  - not accurate enough to be a generative model