Computational Systems Biology: lecture 11

Tommi S. Jaakkola
tommi at csail dot mit dot edu
Outline

• Examples of Markov Random Fields (MRFs)
• MRFs and data integration
• Inferring protein-protein interactions from multiple data sources
Example: side-chain placement

- Consider a fixed backbone structure for a protein. Our goal is to find energetically favorable rotamer angles for the residues.
- Each $x_i$ is a vector of 2-3 discretized angles (rotamers). The settings of these variables are energetically coupled.
Example: side-chain placement

- Consider a fixed backbone structure for a protein. Our goal is to find energetically favorable rotamer angles for the residues.
- Each $x_i$ is a vector of 2-3 discretized angles (rotamers). The settings of these variables are energetically coupled.

![Diagram of a protein structure and a graph representing the energy landscape.](image)
Example: side-chain placement

- Consider a fixed backbone structure for a protein. Our goal is to find energetically favorable rotamer angles for the residues.
- Each $x_i$ is a vector of 2-3 discretized angles (rotamers). The settings of these variables are energetically coupled.

Each pair of residues that could interact (edge $(i, j) \in E$) has an associated potential term $\phi_{ij}(x_i, x_j)$.
Example: side-chain placement

- Consider a fixed backbone structure for a protein. Our goal is to find energetically favorable rotamer angles for the residues
- Each $x_i$ is a vector of 2-3 discretized angles (rotamers). The settings of these variables are energetically coupled

![Diagram of a protein structure with variables $x_1, x_2, x_3, x_4, x_5$]

Each pair of residues that could interact (edge $(i, j) \in E$) has an associated potential term $\phi_{ij}(x_i, x_j)$

The resulting Gibbs' distribution over rotamer angles is

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \exp \left( \sum_{(i,j) \in E} \phi_{ij}(x_i, x_j) \right)$$
Example: clustering

- Suppose each item to be clustered (gene, miRNA, tissue, etc.) can be represented by a profile \( x_i \)
- One way to cluster the set of profiles is to assign a label \( y_i \in \{1, \ldots, k\} \) to each profile \( x_i \) (max \( k \) clusters)

\[
y_i = 1 \quad x_i \\
y_j = 1 \quad x_j
\]

- The label (cluster) assignments should be biased based on the similarity \( S(i, j) \) between the profiles.

\[
S(i, j) = -\|x_i - x_j\|^2
\]
The label (cluster) assignments can be biased to reflect profile similarities

\[ P(y_1, \ldots, y_m; \beta) = \frac{1}{Z(\beta)} \exp \left( \beta \sum_{(i,j)} S(i, j) \delta(y_i, y_j) \right) \]

where the “same label” indicator \( \delta(y_i, y_j) = 1 \) if \( y_i = y_j \) and zero otherwise.
Example: clustering
tissues

\[ y_i = 1 \quad x_i \]
\[ y_j = 1 \quad x_j \]

miRNAs

- The label (cluster) assignments can be biased to reflect profile similarities

\[
P(y_1, \ldots, y_m; \beta) = \frac{1}{Z(\beta)} \exp \left( \beta \sum_{(i,j)} S(i, j) \delta(y_i, y_j) \right)
\]

where the “same label” indicator \( \delta(y_i, y_j) = 1 \) if \( y_i = y_j \) and zero otherwise.

high \( S(i, j) \) \( \Rightarrow \) \( \delta(y_i, y_j) = 1 \) (same cluster)
Example: clustering tissues

\[ y_i = 1 \quad x_i \]
\[ y_j = 1 \quad x_j \]

- The label (cluster) assignments can be biased to reflect profile similarities

\[
P(y_1, \ldots, y_m; \beta) = \frac{1}{Z(\beta)} \exp \left( \beta \sum_{(i,j)} S(i, j) \delta(y_i, y_j) \right)
\]

where the "same label" indicator \( \delta(y_i, y_j) = 1 \) if \( y_i = y_j \) and zero otherwise.

high \( S(i, j) \) \( \Rightarrow \) \( \delta(y_i, y_j) = 1 \) (same cluster)

low \( S(i, j) \) \( \Rightarrow \) \( \delta(y_i, y_j) = 0 \) (different cluster)
Example: clustering

- The label (cluster) assignments can be biased to reflect profile similarities

\[ P(y_1, \ldots, y_m; \beta) = \frac{1}{Z(\beta)} \exp \left( \beta \sum_{(i,j)} S(i, j) \delta(y_i, y_j) \right) \]

where the “same label” indicator \( \delta(y_i, y_j) = 1 \) if \( y_i = y_j \) and zero otherwise.

- high \( S(i, j) \Rightarrow \delta(y_i, y_j) = 1 \) (same cluster)
- low \( S(i, j) \Rightarrow \delta(y_i, y_j) = 0 \) (different cluster)
- high \( \beta > 0 \Rightarrow \max \text{ similarity assignments (subject to } k \text{ labels) } \)
Example: clustering tissues

\[ y_i = 1 \quad x_i \]
\[ y_j = 1 \quad x_j \]

miRNAs

- The label (cluster) assignments can be biased to reflect profile similarities

\[
P(y_1, \ldots, y_m; \beta) = \frac{1}{Z(\beta)} \exp \left( \beta \sum_{(i, j)} S(i, j) \delta(y_i, y_j) \right)
\]

where the “same label” indicator \( \delta(y_i, y_j) = 1 \) if \( y_i = y_j \) and zero otherwise.

- high \( S(i, j) \Rightarrow \delta(y_i, y_j) = 1 \) (same cluster)

- low \( S(i, j) \Rightarrow \delta(y_i, y_j) = 0 \) (different cluster)

- high \( \beta > 0 \Rightarrow \text{max similarity assignments (subject to } k \text{ labels)}

- low \( \beta > 0 \Rightarrow \text{no similarity bias, uniform labeling} \)
Example: clustering

- The label (cluster) assignments can be biased to reflect profile similarities

\[ P(y_1, \ldots, y_m; \beta) = \frac{1}{Z(\beta)} \exp \left( \beta \sum_{(i,j)} S(i,j) \delta(y_i, y_j) \right) \]

where the “same label” indicator \( \delta(y_i, y_j) = 1 \) if \( y_i = y_j \) and zero otherwise.

- “Typical cut” clustering result: place \( i \) and \( j \) in the same cluster if

\[ E_y \{ \delta(y_i, y_j) \} \geq 1/2 \]
Example: data integration

Labeling based on expression profiles

\[ P(y_1, \ldots, y_m; \beta) = \frac{1}{Z(\beta)} \exp \left( \beta \sum_{(i,j)} S(i, j) \delta(y_i, y_j) \right) \]

Labeling based on protein-protein interactions

\[ P(y_1, \ldots, y_n; \eta) = \frac{1}{Z(\eta)} \exp \left( \eta \sum_{(i,j)} I_{ij} \delta(y_i, y_j) \right) \]

\[ I_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ interact} \\ 0, & \text{otherwise} \end{cases} \]

How can we combine these?
Example: data integration

\[
P(y_1, \ldots, y_n; \eta, \beta) = \frac{1}{Z(\eta, \beta)} \exp \left( \eta \sum_{(i,j)} I_{ij} \delta(y_i, y_j) + \beta \sum_{(i,j)} S(i, j) \delta(y_i, y_j) \right)
\]

- It remains to set the parameter values \( \eta \) and \( \beta \)
Example: data integration

- If we know some pairs, e.g., \((l, k)\), should belong to the same cluster, we can maximize the log-likelihood of getting these pairs correctly assigned

\[
J(\eta, \beta) = \log P(y_l = y_k; \eta, \beta) = \log P(\delta(y_l, y_k) = 1; \eta, \beta)
\]

We would also need pairs that belong to different clusters.
Example: data integration

- If we know some pairs, e.g., \((l, k)\), should belong to the same cluster, we can maximize the log-likelihood of getting these pairs correctly assigned

\[
J(\eta, \beta) = \log P(y_l = y_k; \eta, \beta) = \log P(\delta(y_l, y_k) = 1; \eta, \beta)
\]

We would also need pairs that belong to different clusters.

- Maximum likelihood parameter estimation in MRFs corresponds to matching statistics. For example,

\[
\frac{\partial}{\partial \eta} J(\eta, \beta) = \sum_{(i, j)} I_{ij} \left[ E_y \{ \delta(y_i, y_j) | \delta(y_k, y_l) = 1 \} - E_y \{ \delta(y_i, y_j) \} \right] = 0
\]

\[
\text{probability that } i \text{ and } j \text{ are in the same cluster given that } k \text{ and } l \text{ are in the same cluster}
\]

\[
\text{probability that } i \text{ and } j \text{ are in the same cluster}
\]

(\text{the statistics have to be evaluated approximately})
Outline

- Examples of Markov random fields
- Data integration
- Inferring protein-protein interactions from multiple data sources
Inferring physical interactions

- Our focus here will be on binary interactions, seeking to identify which pairs of proteins interact.
- There are a number of measurement technologies for identifying binary interactions:
  - yeast two hybrid
  - protein-chip
  - co-precipitation/mass-spec,
  - etc.
Protein-protein data

• Inherent challenges:
  - the interaction partners may change from one condition to another
  - proteins can be localized to different cellular compartments (e.g., the nucleus)

• Measurement challenges:
  - noisy measurements (e.g., large fraction of false positives)
  - little overlap across measurements of different type
Localization and interactions

- Sub-cellular localization of proteins provides information about interactions (Huh et al., 2003):
How to find interaction graphs

• We can estimate a Markov Random Field model for the purpose of inferring which interactions are real and where they might be localized within the cell (Jaimovich et al., 2006)
How to find interaction graphs

- We can estimate a Markov Random Field model for the purpose of inferring which interactions are real and where they might be localized within the cell (Jaimovich et al., 2006)

- There are several effects to include
  - overall bias of proteins to interact (how often randomly chosen pair of proteins would interact)
  - overall localization biases of each protein
  - how localization co-varies with interaction
  - measurement (noise) models

- The probability model is a product of terms corresponding to these effects
Basic variables

• Interaction variables

\[ I_{ij} = \begin{cases} 
1, & \text{if proteins } i \text{ and } j \text{ interact} \\
0, & \text{otherwise} 
\end{cases} \]
Basic variables

• Interaction variables

\[ I_{ij} = \begin{cases} 
1, & \text{if proteins } i \text{ and } j \text{ interact} \\
0, & \text{otherwise} 
\end{cases} \]

• Localization variables

\[ L_{il} = \begin{cases} 
1, & \text{if proteins } i \text{ appears in location } l \\
0, & \text{otherwise} 
\end{cases} \]
Basic variables

• Interaction variables

\[ I_{ij} = \begin{cases} 
1, & \text{if proteins } i \text{ and } j \text{ interact} \\
0, & \text{otherwise}
\end{cases} \]

• Localization variables

\[ L_{il} = \begin{cases} 
1, & \text{if proteins } i \text{ appears in location } l \\
0, & \text{otherwise}
\end{cases} \]

• Measurement variables
  - protein-protein assay

\[ A_{ij} = \begin{cases} 
1, & \text{if protein } i \text{ and } j \text{ were found interacting} \\
0, & \text{otherwise}
\end{cases} \]
Basic variables

• Interaction variables

\[ I_{ij} = \begin{cases} 1, & \text{if proteins } i \text{ and } j \text{ interact} \\ 0, & \text{otherwise} \end{cases} \]

• Localization variables

\[ L_{il} = \begin{cases} 1, & \text{if proteins } i \text{ appears in location } l \\ 0, & \text{otherwise} \end{cases} \]

• Measurement variables
  
  - protein-protein assay

\[ A_{ij} = \begin{cases} 1, & \text{if protein } i \text{ and } j \text{ were found interacting} \\ 0, & \text{otherwise} \end{cases} \]

  - localization assay

\[ A_{il}^L = \begin{cases} 1, & \text{if protein } i \text{ was found in location } l \\ 0, & \text{otherwise} \end{cases} \]
Potential functions

- Overall bias of proteins to (not) interact

\[ \phi(I_{ij}) = \begin{cases} 
\phi(1), & I_{ij} = 1 \\
0, & \text{otherwise} 
\end{cases} \]

- Terms to include in the probability model:

\[ \prod_{(i,j)} \exp(\phi(I_{ij})) \]
Potential functions

- Overall bias of proteins to be found in a specific location (location, not protein specific)

\[ \phi_l(L_{il}) = \begin{cases} 
\phi_l(1), & \text{if } L_{il} = 1 \\
0, & \text{otherwise} 
\end{cases} \]

- Terms to include in the probability model

\[ \prod_{i,l} \exp (\phi_l(L_{il})) \]
Potential functions

- Dependence between co-localization and interaction (location, not protein specific)

\[
\psi_l(I_{ij}, L_{il}, L_{jl}) = \begin{cases} 
\psi_l(1, 1, 1), & \text{if } I_{ij} = L_{il} = L_{jl} = 1 \\
\psi_l(1, 1, 0), & \text{if } I_{ij} = 1 \text{ and } L_{il} + L_{jl} = 1 \\
\psi_l(1, 0, 0), & \text{if } I_{ij} = 1 \text{ and } L_{il} = L_{jl} = 0 \\
0, & \text{otherwise}
\end{cases}
\]

- Terms to include in the probability model

\[
\prod_{(i,j), l} \exp \left( \psi_l(I_{ij}, L_{il}, L_{jl}) \right)
\]
Potential functions

• Protein-protein interaction measurements (e.g., from yeast two hybrid)

\[ P(A_{ij} = 1 | I_{ij} = 1) \quad \text{true positives} \]
\[ P(A_{ij} = 1 | I_{ij} = 0) \quad \text{false positives} \]

• Terms to include in the probability model

\[ \prod_{(i,j) \in PP} P(A_{ij} | I_{ij}) \]
Potential functions

• Localization measurements (noise model)

\[ P_L(A_{il}^L = 1|I_{il} = 1) \quad \text{true positives} \]
\[ P_L(A_{il}^L = 1|L_{il} = 0) \quad \text{false positives} \]

• Terms to include in the probability model

\[ \prod_{(i,l) \in L} P_L(A_{il}^L|L_{il}) \]
The probability model

- The resulting probability model is over interaction and localization variables:

\[
P(\{I_{ij}\}, \{L_{il}\}) \propto \\
\prod_{(i,j)} \exp \left( \phi(I_{ij}) \right) \times \prod_{i,l} \exp \left( \phi_l(L_{il}) \right) \times \\
\prod_{(i,j),l} \exp \left( \psi_l(I_{ij}, L_{il}, L_{jl}) \right) \times \prod_{(i,j) \in PP} P(A_{ij} \mid I_{ij}) \times \\
\prod_{(i,l) \in L} P_L(A_{il}^L \mid L_{il})
\]

- Lots of variables, few adjustable parameters
About the estimated model

- The parameters estimated on the basis of the available data (cross-validation) reflect the co-localization bias

$$\psi(I_{ij} = 1, L_{il} = 1, L_{jl} = 0) \quad \psi(I_{ij} = 1, L_{il} = 1, L_{jl} = 1)$$

<table>
<thead>
<tr>
<th>localization</th>
<th>Basic model</th>
<th>Noise model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L_{l,p_i} = 1$</td>
<td>$L_{l,p_i} = 1$</td>
</tr>
<tr>
<td></td>
<td>$L_{l,p_j} = 0$</td>
<td>$L_{l,p_j} = 1$</td>
</tr>
<tr>
<td>Nucleus</td>
<td>-0.47</td>
<td>0.66</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>-0.66</td>
<td>-0.02</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>-0.71</td>
<td>1.26</td>
</tr>
<tr>
<td>ER</td>
<td>-0.82</td>
<td>1.18</td>
</tr>
</tbody>
</table>

(b) Localization to interaction
Localization and interaction

- Strong interaction implies co-localization

\[ L_{ij} = 1 \]

\[ I_{ij} = 1 \]

\[ L_{jl} = ? \]

- Inference: \[ L_{jl} = 1 \]
Localization and interaction

- Interactions can be inferred via localization

unknown localization

predicted interaction (computational)

measured interaction

localization in the nucleus
Localization and interaction

- Interactions can be inferred via localization

  ![Diagram](image)

  - unknown localization
  - predicted interaction (computational)
  - measured interaction
  - localization in the nucleus

- Inference: pre7 also localized in the nucleus, interacts with Pre9
Model components

![ROC curve diagram]

- **Full**: Complete model
- **Triplets**: Model with triplets
- **Noise**: Noise model for localization labels
- **Basic**: Localization labels directly from measurements

Note: The ROC curve compares the test set performance of these four models. The advantage of using an integrative model that allows propagation of influence between interactions and protein attributes is evident in Section 2 (see Fig. 3).