Drug-Target Interaction Prediction by Integrating Chemical, Genomic, Functional and Pharmacological Data

Fan Yang, Jinbo Xu, Jianyang Zeng

2014, Pacific Symposium on Biocomputing

Marten Heidemeyer

Presentation for Directed Reading Class, Summer 2015

Outline

Motivation

Knowledge of Drug-Target interaction is important for:

- drug development
- predicting drug side effects
- identification of new targets for known drugs

Wet lab experiments for Drug-Target interaction are expensive

Available Resources

- binary/real-value interaction data
  - KEGG, BRENDA, SuperTarget, DrugBank, BindingDB
  - KEGG: 875 Drugs, 249 Proteins, 2596 observations
  - BindingDB: 106527 Ligands, 2133 Proteins, 193603 observations
- KEGG: chemical structure of Drugs
- SIDER: drug side effect database
- GO: functional annotation of Proteins

Available Resources

- binary/real-value interaction data
  - KEGG, BRENDA, SuperTarget, DrugBank, BindingDB
  - KEGG: 875 Drugs, 249 Proteins, 2596 observations
  - BindingDB: 106527 Ligands, 2133 Proteins, 193603 observations
- KEGG: chemical structure of Drugs
- SIDER: drug side effect database
- GO: functional annotation of Proteins
- Goal: integrate genomic, chemical, functional and pharmacological data to predict missing interactions

Conditional Random Field structure

We have given:

$$\text{known drugs: } d_i,\ 1 \leq i \leq n_d$$
$$\text{known targets: } t_j,\ 1 \leq j \leq n_t$$

For each drug d_i, construct CRF over:

$$G = (V_t, E_t), \text{ where } V_t \text{ set of all targets}$$
$$E_t : \text{connect each target to its } k \text{ nearest neighbors}$$

Conditional Random Field structure

We have given:

$$\text{known drugs: } d_i,\ 1 \leq i \leq n_d$$
$$\text{known targets: } t_j,\ 1 \leq j \leq n_t$$

For each drug d_i, construct CRF over:

$$G = (V_t, E_t), \text{ where } V_t \text{ set of all targets}$$
$$E_t : \text{connect each target to its } k \text{ nearest neighbors}$$

Do the same for each target t_i, where nodes are the drugs.

formal definition of CRF

Let $Y = (y_1, y_2, \ldots, y_{n_t})$ denote the prediction of target t_j.
Let X denote known DTIs and similarity scores.
Define the joint probability density function of Y given X:

$$p(Y|X) = \frac{1}{Z(X)} e^{-E(Y|X)}$$

formal definition of CRF

Let $Y = (y_1, y_2, \ldots, y_{n_t})$ denote the prediction of target t_j.
Let X denote known DTIs and similarity scores.
Define the joint probability density function of Y given X:

$$p(Y|X) = \frac{1}{Z(X)} e^{-E(Y|X)}$$
$$Z(X) = \sum_Y e^{-E(Y|X)}$$

formal definition of CRF

Let $Y = (y_1, y_2, \ldots, y_{n_t})$ denote the prediction of target t_j.
Let X denote known DTIs and similarity scores.
Define the joint probability density function of Y given X:

$$p(Y|X) = \frac{1}{Z(X)} e^{-E(Y|X)}$$
$$Z(X) = \sum_Y e^{-E(Y|X)}$$

Definition of CRF from book:

$$P(Y|X) = \frac{1}{Z(X)} \tilde{P}(Y, X)$$
$$\tilde{P}(Y, X) = \prod_i^m \phi_i(D_i)$$
$$Z(X) = \sum_Y \tilde{P}(Y, X)$$

formal definition of CRF 2

For joint configuration Y given X, define energy:

$$E(Y|X) = \sum_i a_i f(y_i|X) + \sum_{i,j} b_{ij} g(y_i, y_j \ |X)$$

where f and g are penalty functions:

$$f(y_i|X) = -(y_i - H_{x_i}(y_i))^2, \text{ where}$$
$$H_{x_i}(y_i) \text{ average number of observed interactions}$$

formal definition of CRF 2

For joint configuration Y given X, define energy:

$$E(Y|X) = \sum_i a_i f(y_i|X) + \sum_{i,j} b_{ij} g(y_i, y_j \ |X)$$

where f and g are penalty functions:

$$f(y_i|X) = -(y_i - H_{x_i}(y_i))^2, \text{ where}$$
$$H_{x_i}(y_i) \text{ average number of observed interactions}$$

and

$$g(y_i, y_j|X) = -H_{x_i,x_j}(y_i - y_j)^2$$

we learn a_i and b_{ij}.

CRF: Parameter Training

learn a_i and b_{ij} by maximizing the conditional log-likelihood of training data.

CRF: Parameter Training

learn a_i and b_{ij} by maximizing the conditional log-likelihood of training data.
conditional probability was defined as:

$$p(Y|X) = \frac{1}{Z(X)} e^{-E(Y|X)} \Rightarrow p_\theta(Y|X) = \frac{1}{Z_\theta(X)} e^{\theta h}$$

$\Rightarrow$ log-likelihood:

$$L_\theta = \sum_{i=1}^{n_t} log(p(y_i|X))$$

CRF: Parameter Training

learn a_i and b_{ij} by maximizing the conditional log-likelihood of training data.
conditional probability was defined as:

$$p(Y|X) = \frac{1}{Z(X)} e^{-E(Y|X)} \Rightarrow p_\theta(Y|X) = \frac{1}{Z_\theta(X)} e^{\theta h}$$

$\Rightarrow$ log-likelihood:

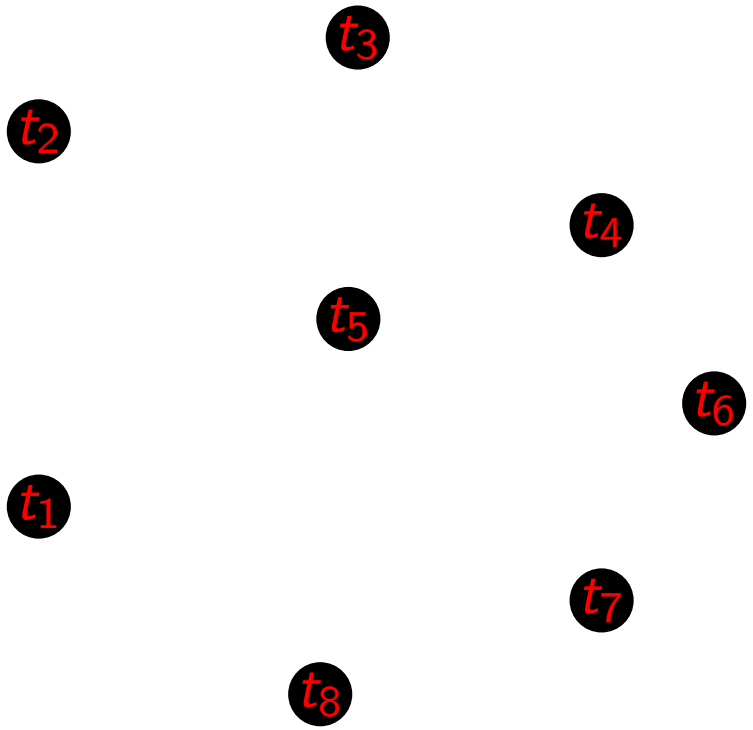
$$L_\theta = \sum_{i=1}^{n_t} log(p(y_i|X)) \Big| \theta = (e^{\theta'_1}, \ldots, e^{\theta'_{n_t}})$$

$\Rightarrow$ derivative of log-likelihood:

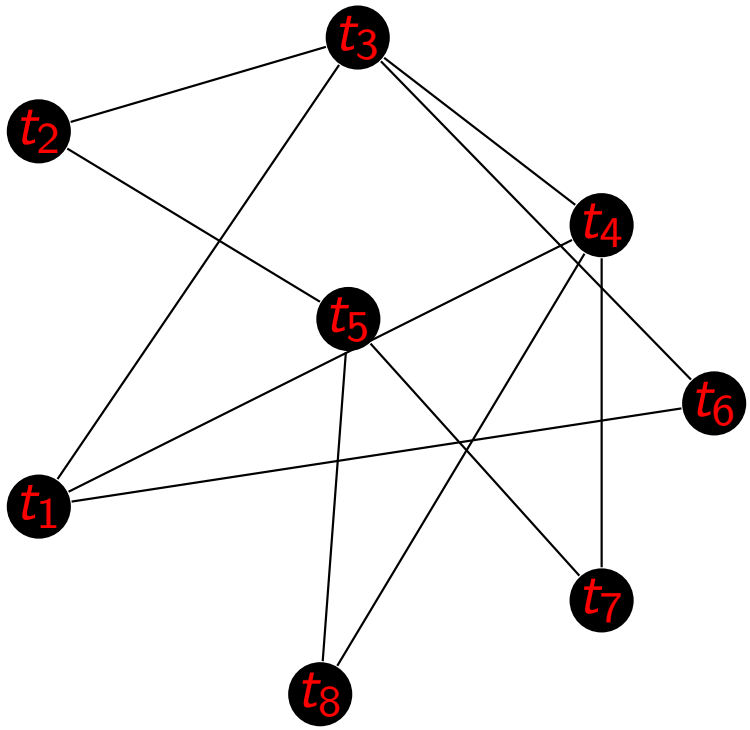
$$\frac{\delta L_\theta}{\delta \theta'} = \theta \sum_{i=1}^{n_t} h(y_i|X) - E_\theta(h(Y|X))$$

- use *stochastic gradient ascent* to find maximizing θ
- use *contrastive divergence* to deal with $E_\theta(h(Y|X))$

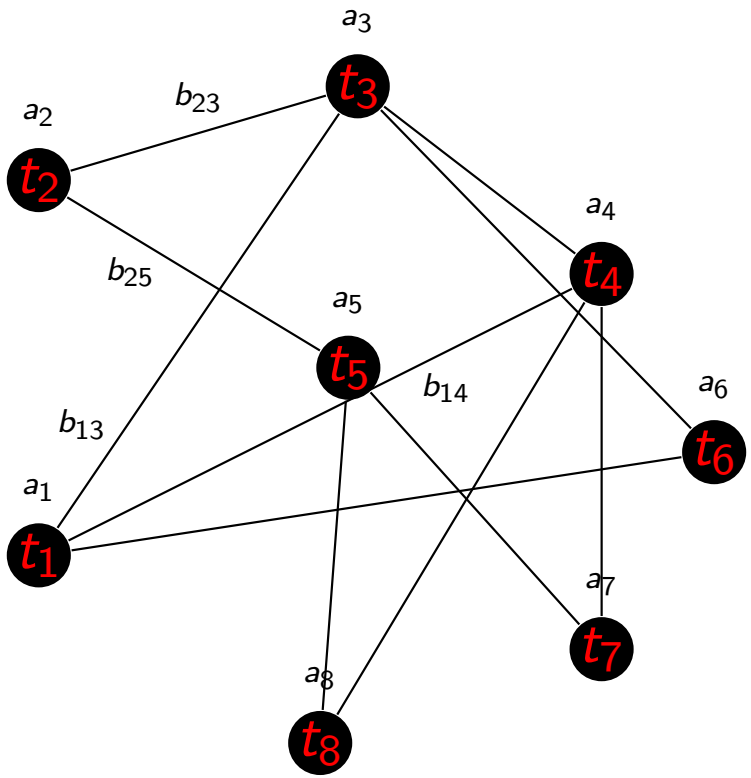
Example: Target-Based CRF



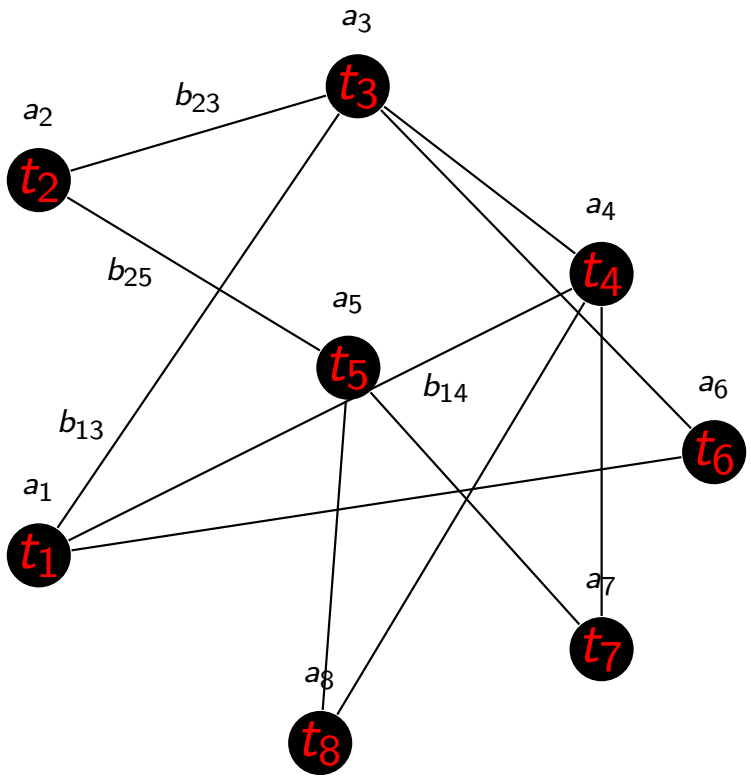
Example: Target-Based CRF



Example: Target-Based CRF

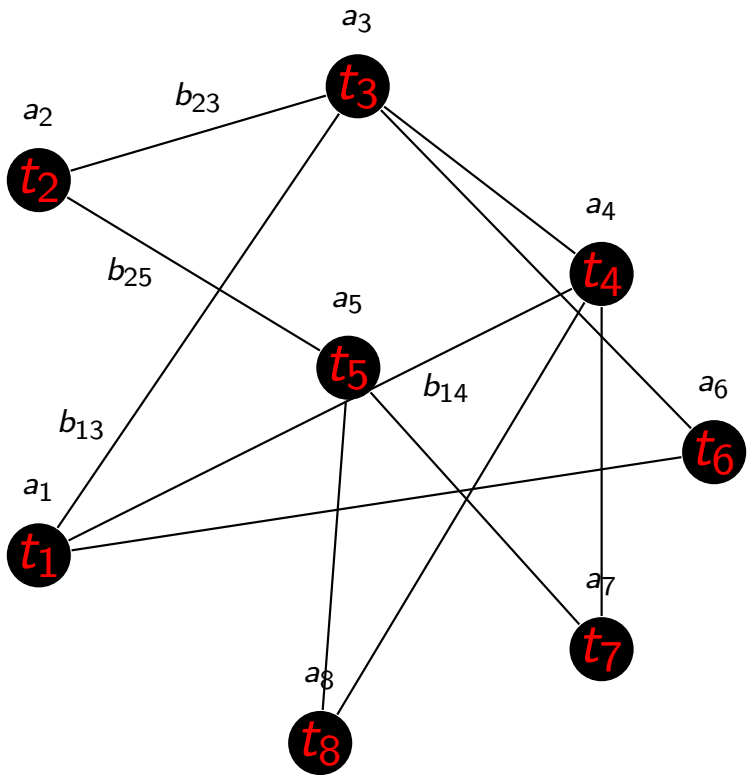


Example: Target-Based CRF



- all Target-Based CRFs share the same a_i and b_{ij}.

Example: Target-Based CRF



- all Target-Based CRFs share the same a_i and b_{ij}.
- exact same procedure for construction of Drug-Based CRFs.

Predicting New Drug-Target Interactions

prediction for target t_k:

- previously we learned $P(Y|X)$
- compute conditional probability distribution $p(y_k|y_{-k}, X)$
  - y_{-k}: all other targets except t_k, set this value to 1 if target is known to interact with query drug, and 0 otherwise
  - $p(y_k|y_{-k}, X) = \frac{p(y_k|X)}{p(y_{-k}|X)}$
- prediction score: conditional expectation of y_k

Construction of CRF

Different approaches to define edges:

- target-based CRF: sequence similarity measure (Genomic approach)
- target-based CRF: functional similarity measure (Functional approach)
- target-based CRF: OR of Genomic and Functional measure (Integrated Genomic-Functional approach)

Construction of CRF

Different approaches to define edges:

- target-based CRF: sequence similarity measure (Genomic approach)
- target-based CRF: functional similarity measure (Functional approach)
- target-based CRF: OR of Genomic and Functional measure (Integrated Genomic-Functional approach)
- drug-based CRF: chemical similarity measure (Chemical approach)
- drug-based CRF: side effect similarity measure (Pharmacological approach)
- drug-based CRF: OR of Chemical and Pharmacological measure (Integrated Chemical-Pharmacological approach)

Full Integration approach

For a given drug-target pair:

- let S_d denote prediction score using the drug-based CRF
- let S_t denote prediction score using the target-based CRF

Compute score for this query drug-target pair as

$$S = \alpha S_d + (1 - \alpha) S_t$$

Testdata and similarity metrics

- experimentally-verified drug-target interactions from *KEGG* database.
- 875 drugs, 249 proteins, 2596 tested interactions $\Rightarrow$ 0.4%
- graph kernel approach to compute chemical similarities between drugs.
- local alignment kernel approach to compute sequence similarities between targets.
- *FunSimMat* to compute functional similarities between targets.
- pharmacological information from *SIDER* database.

Performance Evaluation

Approach		Evaluation Criterion	
		AUC	AUPR
Target-based CRF	GEN	97.3	80.7
	FUN	97.7	80.9
	IGF	98.0	83.9
Drug-based CRF	CHEM	96.0	81.5
	PHAR	96.6	79.9
	ICP	98.1	85.9
Full Integration Approach (FI)		**99.2**	**94.9**

Table 1: Prediction results using 10-fold cross validation

Comparison with existing approaches

- *KEGG* dataset, where all drugs have records in drug side-effects databases *SIDER*, *JAPIC* and *AERS*
- 359 drugs, 226 targets, 1188 drug-target interactions $\Rightarrow$ 1.4%

Approach	**AUPR**	
	CRF	PKR
AERS-freq	85.7	80.6
AERS-bit	85.4	81.3
SIDER	87.3	76.8
JAPIC	91.2	87.7
CHEM	87.7	79.7
INTEG-P	90.7	87.4
INTEG-PC	90.4	88.5
INTEG-ALL	**91.5**	\

Table 2: comparison with existing Pairwise Kernel Regression model

Future work

- incorporate other data such as drug-drug interaction and protein-protein interaction

Thank You!