Aalto University
School of Science
and Technology

Transporter protein classification by structured prediction and multiple kernel

Hongyu Su

Helsinki Institute for Information Technology HIIT
Department of Computer Science
Aalto University

October 8, 2015

Motivation

- Membrane transporter proteins cover 10% proteins in a cell.
- An accurate classification model can help in
  - studying comparative and functional genomics
  - probing metabolic processes
  - developing new therapeutic targets
  - identifying pharmacologically proteins
- Transporter protein classification TC is a hierarchical system of thousands of classes.
- By predicting the whole TC system, we implicitly predict
  - mode of actions
  - energy coupling mechanism
  - phylogenetic group
  - substrate specificity
- Additionally, we would also benefit from the relationship between classes (class-subclasses, different level of granularities).
- The prediction task: input is a transporter protein sequence; output is the corresponding TC.

Notations

- Training examples come in pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$.
- $\mathcal{X}$ is an arbitrary input space.
- $\mathcal{Y}$ is an output space of a collection of ℓ-dimensional *multilabels*.

$$\mathbf{y} = (y_1, \cdots, y_\ell) \in \mathcal{Y}.$$

- y_i is a *microlabel* and $y_i \in \{+1, -1\}$.
- $\mathbf{x}$ is a protein sequence, y_i is a function class, $\ell = 3145$.
- We are given a set of m training examples $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{m}$.
- Each example $(\mathbf{x}, \mathbf{y})$ is mapped into a joint feature space $\phi(\mathbf{x}, \mathbf{y})$.
- $\mathbf{w}$ is the weight vector operates in the joint feature space.
- Define a linear score function $F(\mathbf{w}, \mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle$.
- $\mathbf{w}$ ensures that example $\mathbf{x}_i$ with correct multilabel $\mathbf{y}_i$ achieves higher score than with any other incorrect multilabel $\mathbf{y}' \in \mathcal{Y}$.

Prediction

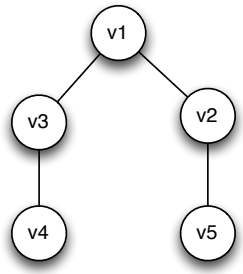
- The prediction $\mathbf{y}_{\mathbf{w}}(\mathbf{x})$ of an input $\mathbf{x}$ is the multilabel $\mathbf{y}$ that maximizes the score function

$$\mathbf{y}_{\mathbf{w}}(\mathbf{x}) = \operatorname*{\mathbf{argmax}}_{\mathbf{y} \in \boldsymbol{\mathcal{Y}}} \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \rangle. \tag{1}$$

- Search space is exponential in size, $|\boldsymbol{\mathcal{Y}}| = 2^{\ell}$.
- (??) is called *inference* problem which is $\mathcal{NP}$-hard for most output feature maps.
- Often, we want a feature map in which the inference can be solved with a polynomial algorithm, e.g., dynamic programming.

TC hierarchy as the output graph

- Transporter classification (TC) system is a hierarchical system.

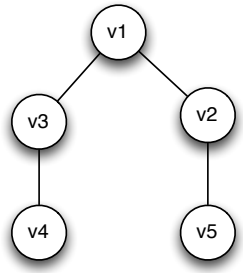


- A vertex $v_i \in V$ corresponds to a microlabel y_i (function class).
- An edge $(v_i, v_j) \in E$ corresponds to class-subclass of the microlabel y_i and y_j.
- We assume that the joint feature map ϕ is a potential function on the TC hierarchy $G = (E, V)$.
- $\varphi(\mathbf{x}) \in \mathbb{R}^d$ is the input feature map, e.g., N-gram feature of a sequence.
- $\psi(\mathbf{y}) \in \mathbb{R}^{4|E|}$ is the output feature map which maps the multilabel $\mathbf{y}$ into a collection of edges and labels

$$\varphi(\mathbf{y}) = (\mathbf{1}_{(\mathbf{y}_e = u_e)})_{e,u_e}, e \in E, u_e \in \{-1, +1\}^2.$$

An example of $\psi(\mathbf{y})$

- Given TC hierarchy $G = (E, V)$



- Multilabel $\mathbf{y}$

$$\mathbf{y} = (y_1, y_2, y_3, y_4, y_5) = (+1, -1, +1, +1, -1)$$

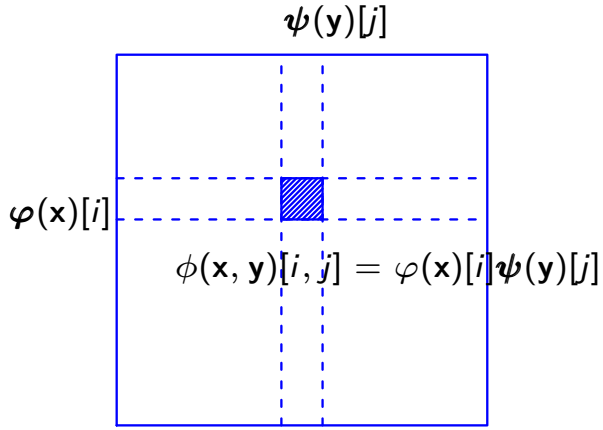
- Output feature map $\psi(\mathbf{y})$

$$\psi(\mathbf{y}) = (\underbrace{\underbrace{0}_{--}, \underbrace{0}_{-+}, \underbrace{0}_{+-}, \underbrace{1}_{++}}_{(v_1, v_3)}, \underbrace{\underbrace{0}_{--}, \underbrace{0}_{-+}, \underbrace{1}_{+-}, \underbrace{0}_{++}}_{(v_1, v_2)}, \cdots)$$

Joint feature map $\phi(\mathbf{x}, \mathbf{y})$

- The joint feature is the Kronecker product of $\varphi(\mathbf{x})$ and $\psi(\mathbf{y})$

$$\phi(\mathbf{x}, \mathbf{y}) = (\phi_e(\mathbf{x}, \mathbf{y}))_{e \in E} = (\varphi(\mathbf{x}) \otimes \psi_e(\mathbf{y}_e))_{e \in E}.$$



- The score function can be factorized by the output graph G

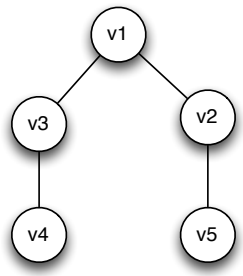
$$F(\mathbf{w}, \mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle = \sum_{e \in E} \langle \mathbf{w}_e, \phi_e(\mathbf{x}, \mathbf{y}_e) \rangle.$$

Search space reduction

- Recall the inference problem

$$\mathbf{y}_{\mathbf{w}}(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle.$$

- Search space $|\mathcal{Y}| = 2^{\ell}$.
- Given TC hierarchy $G = (E, V)$



- Valid multilabels defined by the TC hierarchy

$$\mathbf{y} = (y_1, y_2, y_3, y_4, y_5) = (+1, -1, +1, +1, -1),$$
$$\mathbf{y} = (y_1, y_2, y_3, y_4, y_5) = (+1, +1, -1, -1, +1).$$

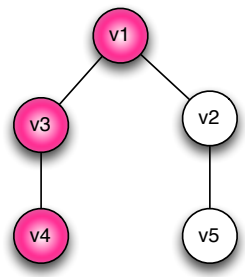
- Search space reduction $|\mathcal{Y}| = |V_{\text{leave}}|$.

Search space reduction

- Recall the inference problem

$$\mathbf{y}_{\mathbf{w}}(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle.$$

- Search space $|\mathcal{Y}| = 2^{\ell}$.
- Given TC hierarchy $G = (E, V)$



- Valid multilabels defined by the TC hierarchy

$$\mathbf{y} = (y_1, y_2, y_3, y_4, y_5) = (+1, -1, +1, +1, -1),$$
$$\mathbf{y} = (y_1, y_2, y_3, y_4, y_5) = (+1, +1, -1, -1, +1).$$

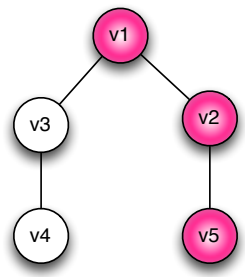
- Search space reduction $|\mathcal{Y}| = |V_{\text{leave}}|$.

Search space reduction

- Recall the inference problem

$$\mathbf{y}_{\mathbf{w}}(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle.$$

- Search space $|\mathcal{Y}| = 2^{\ell}$.
- Given TC hierarchy $G = (E, V)$



- Valid multilabels defined by the TC hierarchy

$$\mathbf{y} = (y_1, y_2, y_3, y_4, y_5) = (+1, -1, +1, +1, -1),$$
$$\mathbf{y} = (y_1, y_2, y_3, y_4, y_5) = (+1, +1, -1, -1, +1).$$

- Search space reduction $|\mathcal{Y}| = |V_{\text{leave}}|$.

Optimization problem

- Solve the following optimization problem to learn $\mathbf{w}$

$$\begin{aligned} \min_{\mathbf{w},\xi_i} \quad & \frac{1}{2}\,||\mathbf{w}||_2^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & \langle \mathbf{w}, \phi(x_i, \mathbf{y}_i)\rangle - \max_{\mathbf{y}\in\mathcal{Y}/\mathbf{y}_i} \langle \mathbf{w}, \phi(x_i, \mathbf{y})\rangle \geq \ell(\mathbf{y}_i, \mathbf{y}) - \xi_k, \\ & \xi_i \geq 0\,, \forall\ i \in \{1, \ldots, m\}, \end{aligned}$$

- The impact of the constraints of the above optimization problem is to push the score of input $\mathbf{x}_i$ with correct multilabel $\mathbf{y}_i$ above the scores of all competing multilabels $\mathbf{y} \in \mathcal{Y}/\mathbf{y}_i$.
- The slack parameters ξ_i is used to relax the constraints so that a feasible solution can always be found.
- C is the margin slack parameter that controls the amount of regularization in the model.
- The objective minimizes the L_2-norm of the weights and the slacks allocated to the training data which is equivalent to maximizing the margin subject to allowing some data to be outliers.

Feature extraction

- BLAST features
  - Identify database sequences that are similar to a query sequence via BLAST search.
  - 'Similarity' is defined as BLAST score (statistically significant matches).
  - In practice, we build a sequence database with all TCBB sequences.
- Features computed from INTERPROSCAN
  - Identify sequence signatures via scanning many signature databases.
  - 18 databases of 3 'big categories': Families, domains, sites, repeats; structural domain; other sequence features.
- Position-specific-score-matrix (PSSM)
  - RPSBLAST: identify sequence profiles via scanning 8 profile database, profile is defined by PSSM.
  - PSIBLAST: search a profile against a sequence database
    (a) build a PSSM profile for a query sequence.
    (b) search a profile against the database or compute similarity of two profiles.

Summary of data

- Number of proteins 12515, number of classes 3145.
- BLAST features: similarity matrix in which elements are BLAST scores.
- INTERPROSCAN features

Type	Dim	Type	Dim	Type	Dim
Protein Domain	145	Hapmap	209	SMART	240
Protein Family	512	PRINTS	579	Panther	4070
Gene3D	611	PIRSF	283	PfamA	2025
Prosite Profile	282	TIGRFAM	769	Prosite Patterns	285
Coil	1	TMHMM	1	Phobius	7
SignalP1	2	SignalP2	2	SignalP3	1

- PSSM features (size of DB)

Type	Dim	Type	Dim
CDD	47363	Pfam	14837
COG	4825	KOG	4875
SMART	1013	PRK	10885
TiGRFAM	4488	CDD NCBI	11273

Multiple kernel learning (MKL)

- A collection of p feature maps $\{\varphi_k(\mathbf{x})\}_{k=1}^{p}$ is on hand.
- We use MKL to combine these feature maps.
- Input kernels $\{K_1, \cdots, K_p\}$ and target kernel $K_{\mathbf{y}} = YY^{\intercal}$.
- **Uniform kernel combination (UNIF)**

$$K_{\text{UNIF}} = \sum_{k=1}^{p} \frac{1}{p} K_k^c,$$

- **Centred kernel alignment (ALIGN)**

$$K_{\text{ALIGN}} = \sum_{k=1}^{p} \alpha_k K_k^c, \quad \alpha_k = \frac{\langle K_k^c, K_{\mathbf{y}}^c \rangle_F}{||K_k^c||_F \, ||K_{\mathbf{y}}^c||_F}$$

- **Two stage MKL (ALIGNF)**

$$K_{\text{ALIGNF}} = \sum_{k=1}^{p} \beta_k K_k^c, \quad \max_{\beta} \frac{\langle K_{\text{ALIGNF}}^c, K_{\mathbf{y}}^c \rangle_F}{||K_{\text{ALIGNF}}||_F \, ||K_{\mathbf{y}}^c||_F}, \quad \text{s.t.} \sum_{k=1}^{p} \beta_k^2 = 1, \beta_k \geq 0, \forall k.$$

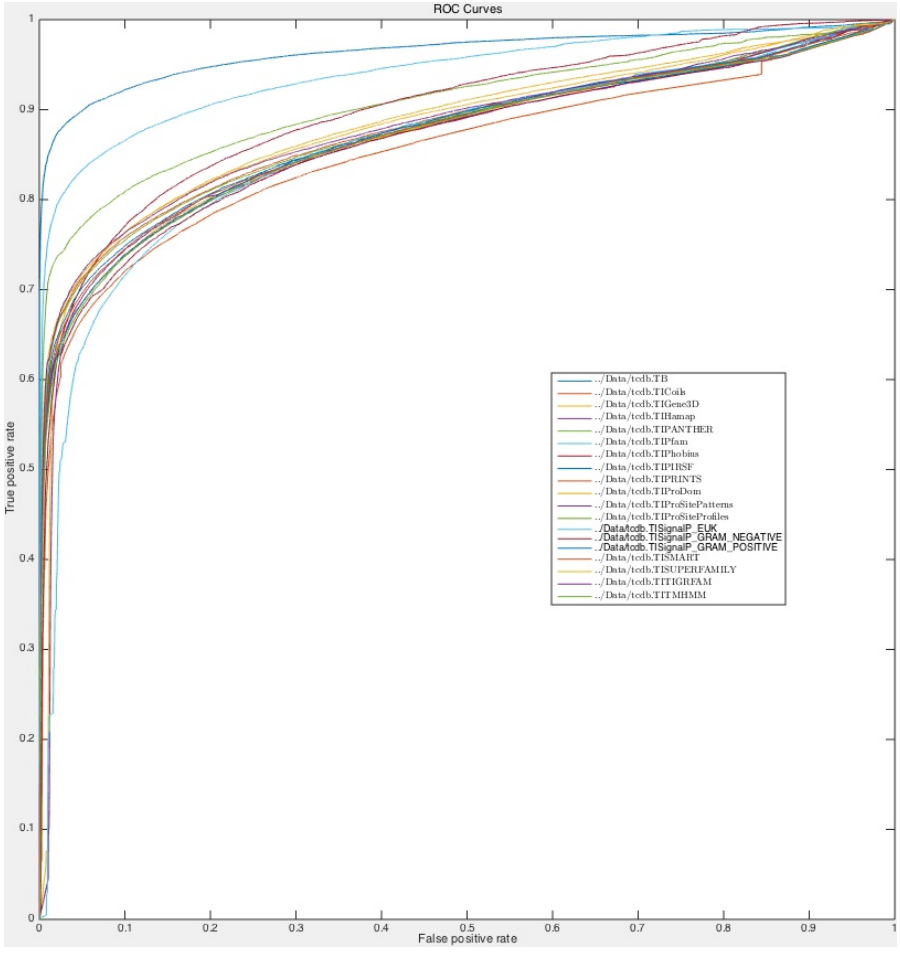
Experiments

SVM + single feature map + linear kernel

	F_1		F_1		F_1
Blast	**74.5**	PIRSF	11.2	SignalP2	3.6
Coils	03.6	PRINTS	13.7	SignalP3	4.1
Gene3D	04.0	ProDom	03.5	SMART	7.2
Hamap	05.5	ProSite Patterns	03.0	SUPERFAMILY	14.9
PANTHER	42.4	ProSite Profiles	15.6	GRFAM	22.3
Pfam	38.2	SignalP1	1.1	TMHMM	06.8
Phobius	11.7				

Experiments (cont.)

SVM + single feature map + linear kernel



Experiments

(SVM, MMR, SOP) + MKL + (linear, Gaussian)

	F_1					0/1			
	Linear			*Gaussian*		*Linear*		*Gaussian*	
	SVM	MMR	SOP	MMR	SOP	MMR	SOP	MMR	SOP
UNIF	68.3	35.1	71.7	79.9	79.9	06.9	55.1	64.1	64.3
ALIGN	74.6	33.9	76.9	83.0	82.8	05.9	58.4	68.3	68.6
ALIGNF	79.2	50.1	80.0	**85.4**	*85.2*	21.0	62.9	*72.7*	**72.8**

Discussion

- Transporter classification TC system is a hierarchical system.
- We developed structured output prediction model SOP to predict TC given a protein sequence.
- Based on TC, we are able to dramatically reduce the search space from exponential to linear which allows
  - Feasible training
  - Accurate prediction
- For each protein sequence, we generate various types of features and apply MKL to merge multiple input feature maps.
- The experiment shows that structured prediction approach significantly improves single label classification approach.