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 7, 2015

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 features

- 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

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

-

Discussion

-