Support Vector Machine Approach

for Protein Subcellular Localization Prediction

Sujun Hua & Zhirong Sun

Tsinghua University

Rohit Bhattacharya

JOHNS HOPKINS
BIOMEDICAL ENGINEERING

TISSUE ANALYTICS

Outline

- Project Recap
- Motivation
- Biological Background
- Mathematical Background
- Methods
- Validation
- Discussion of Findings
- Relevance
- Critique
- Questions/Comments

Project Recap

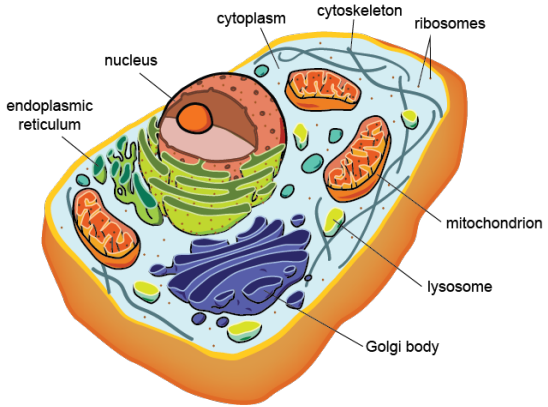
Mobile perfusion analysis - An integrated software and hardware solution that uses mobile-captured images or video data to present a measure of local blood flow to the clinician.

Motivation

- Convert raw genomic sequence data into biological knowledge
- Automate this process
- Previous work computationally expensive and inadequate results
  - Neural nets
  - Covariant discrimination
  - Markov model
- Robustness of solution

Biological Background

- Major subcellular locations:
  - Prokaryotes: Cytoplasm, Periplasm, Extracellular
  - Eukaryotes: Nucleus, Cytoplasm, Mitochondria, Extracellular
- Amino acid composition of proteins is a key functional characteristic and might specify their localization



Mathematical Background - SVMs

A (traditionally) binary classifier that separates classes by a hyperplane given

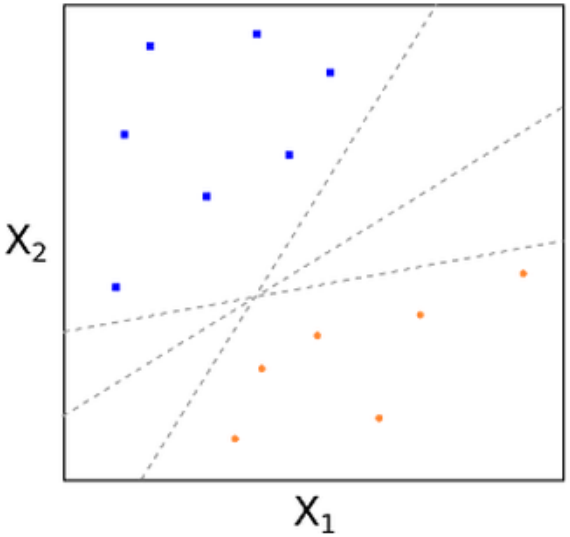


Fig 4: Multiple separating hyperplanes;

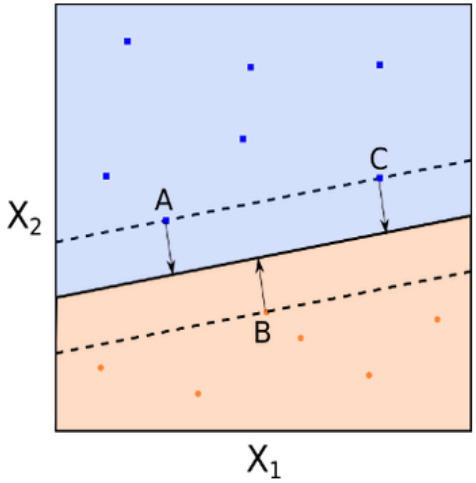


Fig 6: Maximal margin hyperplane with support vectors (A, B and C)

Mathematical Background - SVMs (cotd)

The primal problem

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ & \xi_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}$$

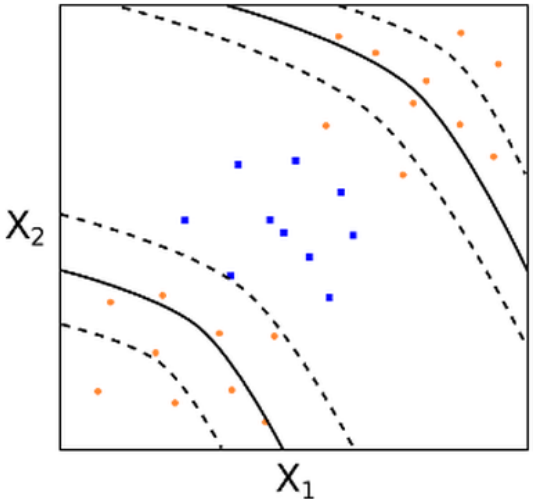
The dual problem

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j \cdot y_i y_j \cdot K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) \\ \text{subject to} \quad & 0 \leqslant \alpha_i \leqslant C \qquad (2) \\ & \sum_{i=1}^{N}\alpha_i y_i = 0 \quad i = 1, 2, \ldots, N. \end{aligned}$$

The decision function

$$f(\vec{\mathbf{x}}) = \text{sgn}\left(\sum_{i=1}^{N} y_i\alpha_i \cdot K(\vec{\mathbf{x}}, \vec{\mathbf{x}}_i) + b\right) \qquad (1)$$

Mathematical Background - Kernels



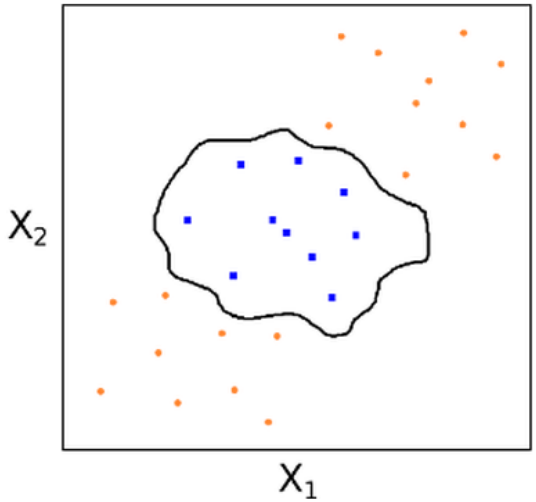


Fig 16: A d-degree polynomial kernel; Fig 17: A radial kernel

$$K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = (\vec{\mathbf{x}}_i \bullet \vec{\mathbf{x}}_j + 1)^d, \quad (3)$$

$$K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \exp(-\gamma \|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j\|^2), \quad (4)$$

Methods

- Data set: Reinhardt and Hubbard (1998)
- Classes: The major subcellular locations
- Feature space: Amino acid composition of proteins
- Kernels: Linear, d-polynomial, and radial
- k-class SVMs
  - ‘1-v-r’ technique
- Jackknife method of validation
- Robustness test: removal of segment of N-terminal sequence

Discussion of Findings

Effect of using different Kernels

Table 2. Prediction accuracies for prokaryotic sequences with different type of kernel functions

Location	Linear		Polynomial*		RBF	
	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	98.1	0.83	97.5	0.86	97.5	0.86
Periplasmic	66.8	0.68	78.7	0.78	78.2	0.78
Extracellular	74.8	0.76	75.7	0.77	76.6	0.77
Total accuracy	89.3	–	91.4	–	91.4	–

Linear: polynomial kernel with $d = 1$; Polynomial*: polynomial kernel with $d = 9$ which is finally used in our prediction system; RBF: RBF kernel with $C = 1000$ was used for each SVM. The results were given by the jackknife test.

Table 3. Prediction accuracies for eukaryotic sequences with different type of kernel functions

Location	Polynomial		RBF*	
	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	78.4	0.63	76.9	0.64
Extracellular	70.2	0.71	80.0	0.78
Mitochondrial	46.1	0.53	56.7	0.58
Nuclear	88.0	0.72	87.4	0.75
Total accuracy	77.3	–	79.4	–

Polynomial: polynomial kernel with $d = 9$; RBF*: RBF kernel with $\gamma = 16.0$ which is finally used in our prediction system. $C = 500$ was used for each SVM. The results were given by the jackknife test.

Discussion of Findings (cotd)

Comparison against other methods

Table 4. Performance comparisons for the prokaryotic sequences. The neural network results were given by cross validation. The covariant discrimination, the Markov model and SVM method results were given by the jackknife test

Location	Neural network	Covariant discrimination	Markov model		SVM	
	Accuracy (%)	Accuracy (%)	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	80	91.6	93.6	0.83	97.5	0.86
Periplasmic	85	72.3	79.7	0.69	78.7	0.78
Extracellular	77	80.4	77.6	0.77	75.7	0.77
Total accuracy	81	86.5	89.1	–	91.4	–

Table 5. Performance comparisons for the eukaryotic sequences. The neural network results were given by cross validation. The Markov model and SVM method results were given by the jackknife test

Location	Neural network	Markov model		SVM	
	Accuracy (%)	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	55	78.1	0.60	76.9	0.64
Extracellular	75	62.2	0.63	80.0	0.78
Mitochondrial	61	69.2	0.53	56.7	0.58
Nuclear	72	74.1	0.68	87.4	0.75
Total accuracy	66	73.0	–	79.4	–

Discussion of Findings (cotd)

- Robustness of SubLoc SVM

Table 6. Performance comparisons for the prokaryotic sequences with one segment of N-terminal sequence removed

	Accuracy (%)				MCC		
	Total	Cyto	Peri	Extra	Cyto	Peri	Extra
COMPLETE	91.3	97.8	76.2	77.6	0.85	0.77	0.78
CUT-10	91.5	90.6	77.3	78.6	0.86	0.78	0.78
CUT-20	90.6	96.5	77.2	77.6	0.85	0.75	0.76
CUT-30	91.1	97.0	77.8	78.5	0.86	0.76	0.77
CUT-40	90.1	96.4	74.8	78.5	0.84	0.73	0.77

COMPLETE: prediction performance for the complete sequences; CUT-10: prediction performance for the remaining sequence parts when 10 N-terminal amino acids were removed; CUT-20, CUT-30 and CUT-40 have similar meanings. Cyto, Peri and Extra are short for Cytoplasmic, Periplasmic and Extracellular, respectively.

- Reliability index

$$\mathrm{RI} = \begin{cases} \mathrm{INTEGER}\,(\mathrm{diff}) + 1 & \text{if} \quad 0 \leqslant \mathrm{diff} < 9.0 \\ 10 & \text{if} \quad \mathrm{diff} \geqslant 9.0. \end{cases} \tag{8}$$

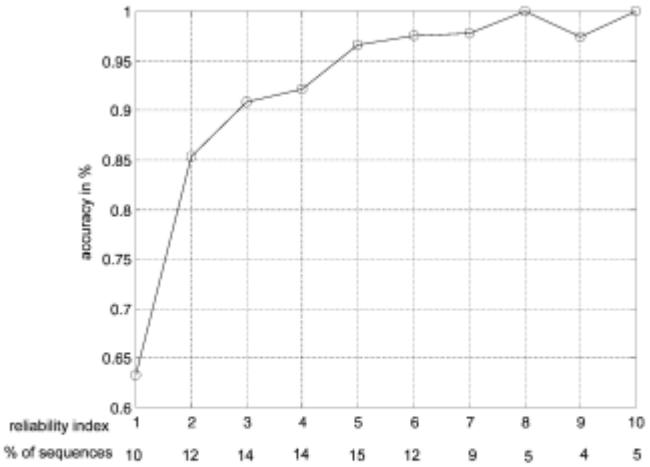


Fig. 2. Expected prediction accuracy with a reliability index equal to a given value. The fractions of sequences that are predicted with $\mathrm{RI} = n$, $n = 1, 2, \ldots, 10$ are also given.

Relevance To Our Project

- SVM classifier for perfusion (either multi-class or binary high/low)
- Possible feature space from Eulerian Video Magnification:
  - Peak-peak distance
  - Zero-crossings
  - Characteristic frequency from Fast Fourier Transform
  - Average intensity
  - Rate of change of intensity

Critique

Pros

- Good applications paper
- Good validation methods
- Links to finished software (as well as tools used to build it)

Cons

- No explicit mention of features used
- No mention of cost parameter tuning
- Why the drop in accuracy between prokaryotes and eukaryotes

Questions/Comments