Metodi kernel in bioinformatica

Giorgio Valentini
DSI – Università degli Studi di Milano

1

Sommario

Applicazione di metodi kernel per:

- Individuazione di omologie remote fra proteine
- Classificazione funzionale di geni e proteine
- Ricerca di pattern in sequenze biologiche
- Analisi supervisionata di dati di espressione genica
- Metodi per l'integrazione di dati bio-molecolari eterogenei
- Altre applicazioni.

2

Approcci classici all'individuazione di omologie remote fra proteine

1. Metodi basati sulla similarità a coppie fra proteine: programmazione dinamica (Smith e Waterman, 1981), metodi euristici: BLAST (Altschul et al., 1990) e FASTA (Pearson, 1990)
2. Metodi basate su statistiche e modelli probabilistici: Profili (Grisbkov et al. 1990); Hidden Markov Model (Baldi et al., 1994)
3. Metodi basati su allineamenti multipli da DB: PSI-BLAST (Altschul et al., 1997) (solo esempi positivi)

3

Metodi basati su kernel per l'individuazione di omologie fra proteine

- SVM Fisher kernel (Jaakkola et al., 1999)
- Composition kernel (Ding e Dubchak, 2001)
- Motif kernel (*Logan et al*., 2001; *Ben-Hur e Brutlag*, 2003)
- Kernel basati sulla comparazione a coppie (*Liao e Noble*, 2003)
- Kernel spettrali (*Leslie et al*. 2002;*Vishwanathan e Smola*, 2003)

4

SVM Fisher kernel

SVM Fisher kernel (Jaakkola et al., 1999):

1. Idea di base: usare anche esempi negativi
2. Accoppiare HMM (modellazione campioni positivi e negativi) e SVM (discriminazione)
3. Addestramento HMM e costruzione di un vettore “gradiente” basato sui parametri dell’HMM → addestramento SVM su vettori positivi e negativi.

5

Composition kernel

- Composition kernel (Ding e Dubchak, 2001):
  - Proteina: vettore di frequenze di “lettere”
  - Sei alfabeti utilizzati (125 feature):
    - Aminoacidi
    - Struttura secondaria
    - Idrofobicità
    - Volume di Van der Waals
    - Polarità
- Kernel simile usato anche per la predizione della struttura secondaria (Cai et al., 2001)

6

Motif kernel

- “Estensione” dei compostion kernel
- Le feature corrispondono a motif di database pre-esistenti
- Es: *Logan et al*., 2001: motif ottenuti dal DB BLOCKS → vettore 10.000 dimensionale (risultati migliori dell’ SVM-Fisher kernel)
- Es: *Ben-Hur e Brutlag*, 2003: eBLOCKS DB (500.000 motivi!): utilizzo di strutture dati speciali per calcolare i kernel.

7

Kernel basati sulla comparazione a coppie

- Basati sull’ assunzione di evoluzione molocolare basata su mutazioni e piccole inserzioni e delezioni
- Algoritmo a due passi (*Liao e Noble*, 2003):
  1. Calcolo dei punteggi di allineamento fra tutte le coppie di proteine (matrice *A* degli allineamenti)
  2. Ciascuna riga di *A* rappresenta la proteina nello spazio delle feature: una funzione kernel standard può essere usata per computare la similarità fra coppie di proteine (empirical kernel map) → calcolo della matrice kernel *K*.
- Problema: complessità di calcolo (riducibile usando BLAST al posto di Smith e Waterman per il passo 1)

8

Kernel per stringhe: spectrum kernel

- I *kernel spettrali* si possono interpretare come una generalizzazione dei *composition kernel*: la composizione è calcolata rispetto a *k-meri*
- Es: per k=5 e un alfabeto di dimensione uguale a 20 → vettori $5^{20} \approx 10^{14}$ elementi!

- Struttura dati trie (*Leslie et al*. 2002)
- Suffix tree (*Vishwanathan e Smola*, 2003)
  - Variante: *mismatch kernel* (*Leslie et al*., 2003) permette *M* mismatch tra *k-meri*

9

Classificazione funzionale di geni e proteine

- Classificazione funzionale basata sull'analisi delle regioni promoter
- Predizione della funzione delle proteine dai profili filogenetici
- Predizione della localizzazione subcellulare delle proteine
- Classificazione tramite feature binarie ottenute da allineamento multiplo

10

Classificazione funzionale basata sull'analisi delle regioni promoter

Analisi delle regioni non codificanti "a monte" del gene da classificare:

1. Applicazione del Fisher kernel con motif-based HMM: predizione dell'appartenenza a classi di geni coregolati nel lievito (*Pavlidis et al*., 2001)
2. Applicazione di kernel gaussiani e sigmoidali a feature estratte con Weeder (*Pavesi et al*., 2001) e metodi statistici: predizione dell'appartenenza a classi di geni coespressi nel lievito (*Pavesi e Valentini*).

11

Classificazione funzionale di geni e proteine: altri esempi

- Predizione della funzione delle proteine dai profili filogenetici (Vert, 2002)
  - Profilo filogenetico di una proteina rappresentato tramite stringa di bit: ciascuna bit rappresenta se esiste un omologo in una determinata specie → rappresenta (parte) della storia evolutiva della proteina
  - Proteine con profili simili hanno funzioni simili
  - Un tree kernel rappresenta la somma pesata sulle possibili storie evolutive
- Predizione della localizzazione subcellulare delle proteine (Hua e Sun, 2002):
  - 20-feature composition kernel
- Classificazione tramite feature binarie ottenute da allineamento multiplo (Zavaljevski et al. 2002)
  - Dall'allineamento multiplo si ottiene una vettorizzazione: feature binarie che rappresentano l'occorrenza di un particolare aa in una particolare posizione dell'allineamento. Compressione ad un alfabeto di dimensione 7.

12

Ricerca di pattern in sequenze biologiche

- Predizione dei siti di inizio della traduzione
  - Importante per il processo di riconoscimento dei geni.
  - Zien et al., 2000: Finestra di lunghezza fissa con codifica a 4 bit per base.
- Predizione dei siti di splicing
  - Ricerca di siti di inzio degli introni
  - Degroeve et al, 2002: Finestra di lunghezza fissa con codifica a 4 bit per base e ricerca delle posizioni più significative
- Predizione della localizzazione subcellulare
  - Ottenibile dall'analisi dei peptidi-segnale
  - Vert, 2002: riconoscimento della posizione in cui il peptide è tagliato tramite kernel derivati da modelli probabilistici bayesiani.
- Predizione della struttura secondaria
  - Predizione di struttura ad elica, a foglietto o a "gomitolo"
  - Hua e Sun, 2001: kernel gaussiano con finestra a scorrimento di 11 aa.

13

Analisi supervisionata dei dati di espressione genica

- Classificazione dei geni
- Classificazione degli esperimenti /pazienti
- Selezione dei geni

	Array1	*Array2*	*...*	*ArrayK$_1$*	*ArrayK$_1$+1*	*...*	*ArrayK*
Gene 1	X_{11}	X_{12}	...	X_{1K1}	X_{1K1+1}	...	X_{1K}
Gene 2	X_{21}	X_{22}	...	X_{2K1}	X_{2K1+1}	...	X_{2K}
...	...	...	...	...	...	...	...
Gene n	X_{N1}	X_{N2}	...	X_{NK1}	X_{NK1+1}	...	X_{NK}

14

Classificazione funzionale dei geni del lievito (Brown et al., 2000)

- Dati: 79 esperimenti di microarray relativi a circa 6000 geni di Saccharomices Cerevisiae.
- 5 classi funzionali (MIPS database)
- Risultati:
  - SVM con kernel gaussiano e polinomiale risultati migliori
  - Possibile utilizzo per predizione di geni non annotati
  - Identificazione di geni mis-etichettati
  - Falsi negativi ciclo TCA dovuti a modificazioni post-traduzionali

15

Diagnosi bio-molecolare

- Classificazione degli esperimenti (trasposta della matrice di espressione)
- Diagnosi basata sulle caratteristiche bio-molecolari dei pazienti

Esempi:

- *Leucemie ALL-AML*
  (Golub et al., 1999), primo esempio applicazione algoritmi supervisionati: 72 campioni, 7129 geni. (Migliori risultati ottenuti in seguito con SVM).
- *Tumore del colon*.
  (Moler at al. 2000): 40 campioni tumorali e 22 sani, 2000 geni. Classificazione con feature selection NBR+SVM
- *Tumore delle ovaie* (Furey et al., 2000): 31 campioni. Risultati comparabili con algoritmi basati sul percettrone.
- *Sarcoma dei tessuti molli* (Segal et al. 2003): 2 classi (melanoma e STS) 76 campioni. T-test ed SVM.
- …

16

Selezione dei geni

Quali geni sono correlati/rilevanti per un particolare stato funzionale?

- Metodi “filtro”
- Metodi “wrapper”
- Metodi “embedded”

(Vedi *Guyon ed Ellisseeff*, 2003 per una review dei metodi di feature selection)

Esempio: *RFE* Recursive Feature Elimination (*Guyon et al*. 2002).

1. Inizializzazione del data set (contiene tutte le feature)
2. Training della SVM : $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
3. Ordinamento delle feature (geni) in base a $|w_i|$
4. Eliminazione della feature in coda al’ordinamento
5. Ritorna al passo 2 finchè rimane una feature.

17

Il problema del selection bias

- L’analisi delle prestazioni del classificatore viene effettuata con dati utilizzati per selezionare le feature (geni) (Ambroise and McLachlan, 2002)

- Con tecniche di hold-out multiplo o cross validation bisogna ripetere la selezione per ogni training set.
- Come selezionare sottoinsiemi di geni rilevanti ?

18

Classificazione multi-classe di dati di espressione genica

- One per class (OPC) e All pairs (PWC): OPC senza feature selection ottiene migliori risultati (*Ramaswamy et al.* 2001)
- Error Correcting Output Coding (ECOC): gli ensemble ECOC ottengono migliori risultati degli OPC e comparabili con CC (Correcting Classifiers, variante di PWC) (*Valentini*,2002)

Alcuni problemi aperti:

• Quale approccio per la classificazione multiclasse dei dati di espressione genica?

• Come valutare l'affidabilità delle predizioni?

• Come valutare l'affidabilità della etichettatura multiclasse ?

• Come classificare classi strutturate e campioni multi-etichetta? (Es: gene ontology)

19

Integrazione di dati biologici eterogenei

Un esempio:

Dati associabili ad un gene:

- La sequenza del gene
- La sequenza della proteina codificata
- La struttura della proteina
- La similarità con altre proteine
- I livelli di mRNA associati al gene in diverse condizioni
- Le occorrenze dei siti di legame dei fattori di trascrizione corrispondenti
- La mappa delle interazioni con altre proteine

Come integrare dati strutturalmente differenti per fini predittivi?

20

Tecniche di integrazione basate su kernel

Tramite i kernel si possono rappresentare dati diversi in modo omogeneo (*matrice dei kernel*)

Tipologie di integrazione basate su kernel:

- Integrazione “iniziale”: semplice concatenazione dei vettori
- Integrazione “intermedia”: kernel computati separatamente e successivamente sommati (*Pavlidis et al.* 2002)
- Integrazione “tardiva”: kernel e funzione discriminante computate separatamente succesivamente integrate
- Integrazione intermedia “pesata”: le matrice sommate sono pesate (*Lanckriet et al*., 2004)

21

Altri esempi di applicazioni

- Classificazione dei tumori basata su dati di metilazione: classificazione basata sui pattern di metilazione della citosina nelle regoni di regolazione dell’espressione dei geni (Model et al., 2001)
- Predizione delle interazioni proteina-proteina: data una coppia di proteine una SVM predice se interagiscono o no. Dati: insieme di feature per ciascun aa.; coppia di proteine come concatenazione dei vettori corrispondenti (Boch e Gough, 2001)
- Identificazione di peptidi con dati di spettrometria di massa: 1) Digestione enzimatica di proteine → peptidi → selezione tramite spettrometro di massa → frammentazione tramite ionizzazione → misura dei frammenti tramite un secondo spettrometro (Tandem mass spectrometry): lo spettro finale contiene picchi che corrispondono alle sottostringhe dei peptidi → inferenza dei peptidi tramite comparazione con DB (algoritmo SEQUEST). SVM utilizzata per ridurre i falsi positivi (Anderson et al., 2003).

22

Perchè i kernel per la bioinformatica

Caratteristiche dei dati e problemi bio-molecolari

- I dati bio-tecnologici sono caratterizzati da elevata dimensionalità
- Molti dati bio-molecolari hanno struttura non vettoriale
- Conoscenza biologica sui dati è a volte disponibile
- Disponibilità di fonti diverse di dati relative ad un medesimo fenomeno biologico

Caratteristiche dei metodi kernel

- Possono trattare dati di elevata dimensionalità
- Possono trattare dati non vettoriali (stringhe, alberi, grafi)
- Possono incorporare conoscenza a priori
- Possono integrare dati eterogenei

23