Marchiatura digitale di sequenze video stereoscopiche a disparit coerente

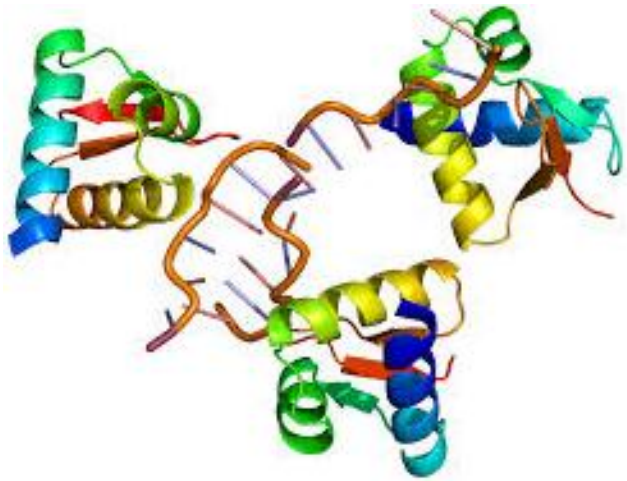
Benedetta Barbetti
Michaela Servi

Università degli studi di Firenze

8 Ottobre 2014

Proteine RNA-binding

- L'RNA interagisce con le proteine per portare a termine molti dei suoi compiti
- Con il termine **RNA-binding protein** si intende una proteina in grado di instaurare legami con l'RNA



- Per capire i meccanismi di questa interazione sono state proposte diverse tecniche che si differenziano sia per algoritmo implementato sia per feature in input

RNA-binding sites prediction methods

Group	Algorithm	Input Info	Category
Jeong et at., 2004	Neural Network	Binary vector + predicted secondary structure	Single sequence
Jeong and Miyano, 2006		PSSM	Multiple sequence
Terribilini et at., 2006	Naive Bayes	Binary vector	Single sequence
	SVM(RBF)	PSSM	Multiple sequence
Wang and Brown, 2006	SVM(RBF)	Hydrophobicity	Single sequence
Kumar et at., 2007	SVM	Binary vector	Single sequence
		PSSM	Multiple sequence

Kumar et al., 2007

- Classificatore basato su sequenze singole e multiple
- Dataset di **86** proteine (Jeong et al., 2006)
- Ogni amino acido è stato codificato come un **vettore binario** o un vettore **PSSM** ottenuto con una ricerca PSI-BLAST sul NCBI nonredundant database
- Come input alla **SVM** è stato utilizzata una sliding window
- La predizione migliore è stata ottenuta utilizzando l'encoding PSSM, che ha prodotto un'accuratezza di **0.81** e un MCC di 0.45

RNA_86

- 86 catene proteiche RNA-binding estratte dalle strutture dei complessi di RNA-proteina
- Le strutture sono state ottenute dalla Protein Data Bank
- La similarità tra coppie di catene è minore del 70%
- Ogni catena proteica ha almeno 4 RNA-interacting residui
- **20071** residui, **4568** dei quali sono RNA-interacting residui

Support Vector Machines

- Per la classificazione dei RNA-interacting residui sono state utilizzate le **SVMs**, implementate utilizzando la libreria `sklearn`
- Applicate con successo a numerosi problemi di regressione e classificazione di pattern, tra cui problemi di bioinformatica

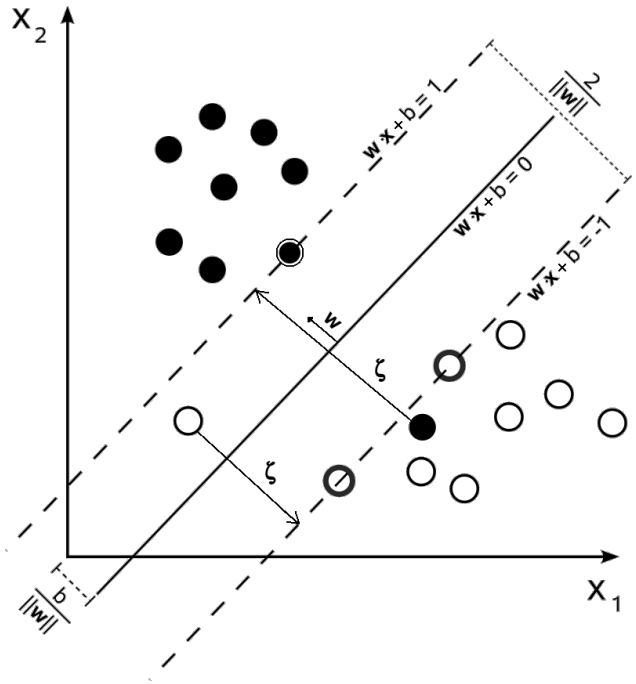
Definizione

Dati $\mathbf{x_i} \in R^p, i = 1, ..., n$, in 2 classi, e un vettore $\mathbf{y} \in R^p$ tale che $y \in \{-1, 1\}$, SVC risolve il seguente problema:

$$\min_{\omega, b, \zeta} \frac{1}{2} \omega^T \omega + C \sum_{i=1,n} \zeta_i$$

$$\text{subject to } y_i(\omega^T \phi(x_i) + b) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0, i = 1, ..., n$$



Problema duale

Il problema in forma duale:

$$\min_{\alpha} \ \frac{1}{2}\,\alpha^T Q\alpha - e^T\alpha$$
$$\text{subject to } y^T\alpha = 0$$
$$0 \le \alpha_i \le C,\ i = 1, ..., l$$

La funzione di decisione è

$$sgn(\sum_{i=1}^{N} y_i\alpha_i K(x_i, x) + b)$$

dove Q è una matrice *nxn* semidefinita positiva, $Q_{ij} \equiv K(x_i, x_j)$ e K è la funzione di kernel

Funzione di kernel

La funzione di kernel può essere una delle seguenti:

- lineare $\langle x, x' \rangle$
- polinomiale $(\gamma \langle x, x' \rangle + r)^d$,
- rbf $exp(-\gamma |x, x'|^2)$
- sigmoide $tanh(\gamma \langle x, x' \rangle + r)$

Generazione dei pattern

Sono stati generati due diversi input per il modello:

- Pattern binari
- PSSM

Pattern binari

- Pattern sovrapposti di lunghezza fissata N dove ogni amino acido è rappresentato come un vettore *one hot* lungo 21
- Pattern corrispondenti ai residui terminali nella catena proteica sono stati generati aggiungendo $(N-1)/2$ "dummy" residui X agli estremi della proteina
- Se il residuo centrale del pattern è un RNA-interacting residuo, il pattern viene classificato come positivo
- Viene generato un pattern per ogni residuo della sequenza proteica

Pattern binari

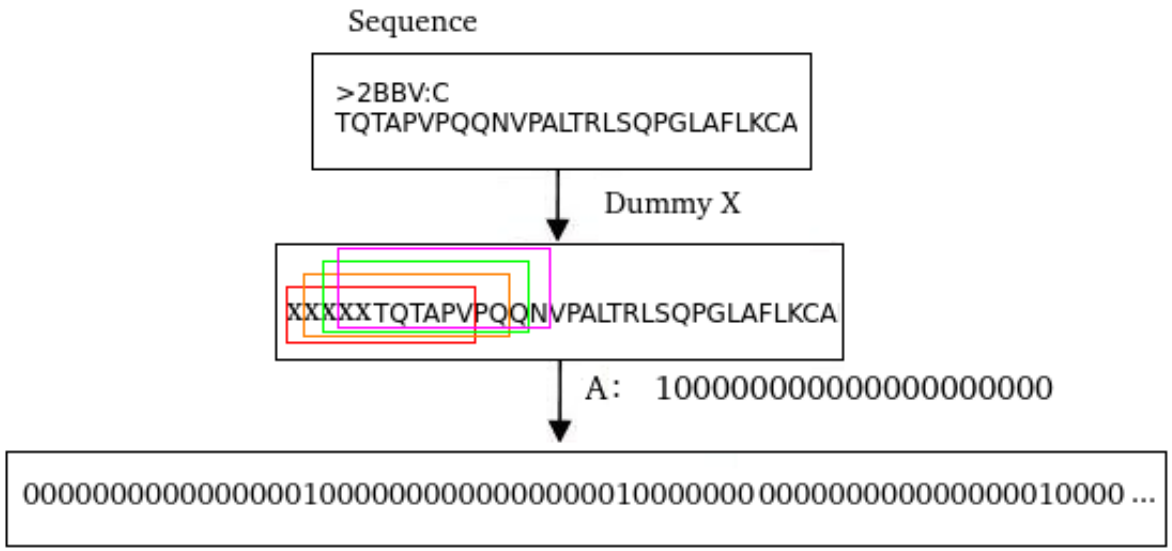


Figure : Generazione pattern binari

PSSM

- L'informazione evoluzionistica ottenuta dall'allineamento multiplo di sequenze fornisce maggiore informazione sulla proteina rispetto all'allineamento singolo
- Position specific scoring matrix generata da PSI-BLAST contro NCBI nonredundant database
- Pattern soprapposti ottenuti facendo scorrere una finestra di dimensione $N \; x \; 21$ sulla matrice

PSSM

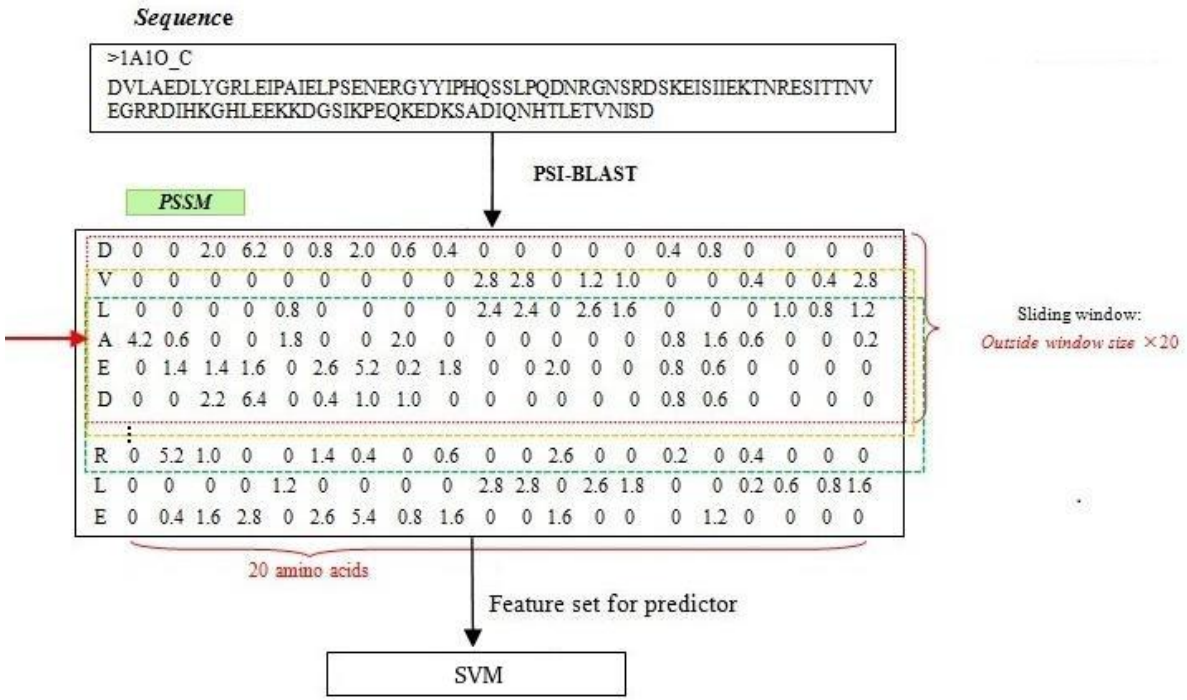


Figure : Generazione feature PSSM

Valutazione delle prestazioni

- Per valutare ler performance del modello sono stati calcolati i seguenti parametri:
  - Accuratezza $= \frac{TP + TN}{TP + FN + TN + FP} \; X \; 100$
  - Sensitività $= \frac{TP}{TP + FN} \; X \; 100$
  - Specificità $= \frac{TN}{TN + FP} \; X \; 100$
  - MCC $= \frac{TP \; x \; TN - FP \; x \; FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \; X \; 100$
- É stata utilizzata la tecnica del five-fold cross-validation

Risultati

- $C = 70$
- $\gamma = 0.07$

Input Pattern	N	Sensitivity(%)	Specificity(%)	Accuracy(%)	MCC
Binary Pattern	11	33.06	90.17	76.27	0.27
	13	31.92	92.27	77.58	0.30
	15	30.59	93.72	78.35	0.31
	17	29.71	94.60	78.80	0.32
	19	**27.95**	**95.54**	**79.08**	**0.33**
PSSM	11	29.27	96.56	80.18	0.37
	13	31.92	96.50	80.78	0.40
	15	33.33	96.19	80.88	0.40
	17	**34.65**	**95.88**	**80.97**	**0.41**
	19	36.06	95.28	80.86	0.40

Idea

- Partendo dal modello di Kumar si è cercato di migliorarne l'accuratezza usando come nuove feature rappresentazioni non supervisionate delle sequenze proteiche
- Tecnica già utilizzata nell'ambito del Natural Language Processing e dell' Handwritten Digit Recognition

Estrazione delle nuove feature

Utilizzando la libreria `pylearn2`

- Implementazione di un Denoising Autoencoder (DAE)
  - Implementazione di una funzione di corruzione dei dati in ingresso all'autoencoder
  - Utilizzo della funzione softmax per aiutare l'autoencoder a ricostruire l'output in modo che rappresenti una sequenza di amino acidi
- Implementazione di una Restricted Boltzmann Machine (RBM)

RNA_439

439 catene proteiche RNA-binding ottenute da quattro diversi dataset

- Dataset di Kumar (86 catene proteiche)
- RB198 compilato da Lewis et al. 2010 (198 catene proteiche che condividono meno del 30 % di similarità)
- RB44 costruito da Puton et al. 2012 (44 catene proteiche che condividono meno del 40 % di similarità)
- RB111 composto da 111 catene proteiche che condividono meno del 30 % di similarità e meno del 40 % con quelle di RB198 e RB44

Generazione dei pattern

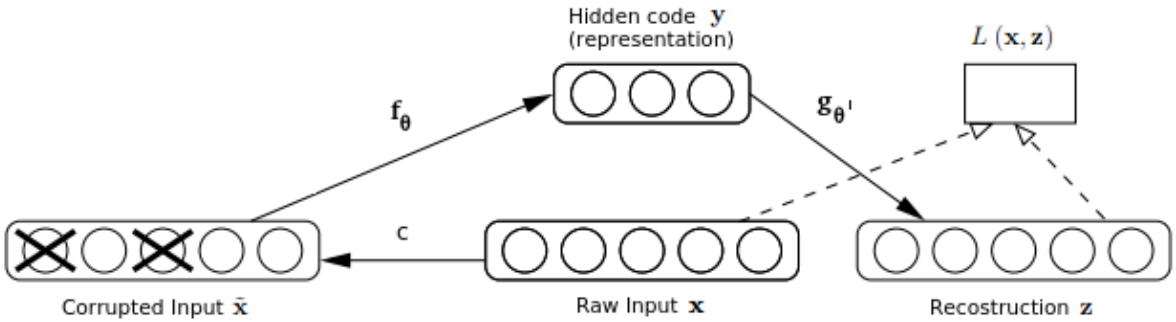
Sono stati generati due diversi input per il modello:

- Pattern binari di lunghezza N
- Pattern binari di lunghezza 5, ottenuti con una sliding window dai pattern di lunghezza N

Denoising Autoencoders

- Estensione di un classico autoencoder
- L'autoencoder viene allenato a riscostruire l'input a partire da una versione corrotta di questo, per forzare i livelli nascosti a scoprire feature più robuste
- Il processo di corruzione è di tipo stocastico
  - Corruttore gaussiano: aggiunge rumore gaussiano con media nulla all'input
  - Corrutore one-hot: cambia l'elemento attivo con una certa probabilità

Definizione



- L'input corretto $\mathbf{x} \in [0,1]^d$ viene parzialmente corrotto,$\widetilde{\mathbf{x}}$
- $\widetilde{\mathbf{x}}$ viene mappato in una rappresentazione nascosta $\mathbf{y} = \mathit{sigmoid}(\mathbf{W}\widetilde{\mathbf{x}} + \mathbf{b})$
- A partire da y viene ricostruito $\mathbf{z} = \mathit{sigmoid}(\mathbf{W'y} + \mathbf{b'})$
- Il modello viene trainato in modo che i parametri minimizzino l'errore di ricostruzione $\mathbf{L(x,z)} = ||x - z||^2$

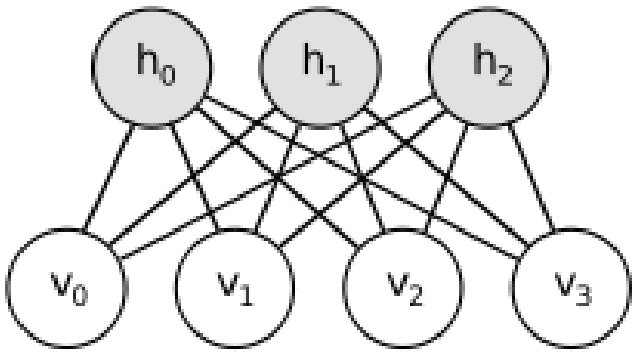
Risultati

- #unità nascoste = 500
- #epoche = 100

Input Pattern	N	Sensitivity(%)	Specificity(%)	Accuracy(%)	MCC
Binary Pattern	**5**	**28.13**	**93.21**	**77.36**	**0.28**
	11	30.68	89.49	75.17	0.23

Definizione

- Markov Random Field associate con un grafo non orientato bipartito
- Valori binari per unità visibili e unità nascoste



- La probabilità congiunta è data dalla distribuzione di Gibbs:

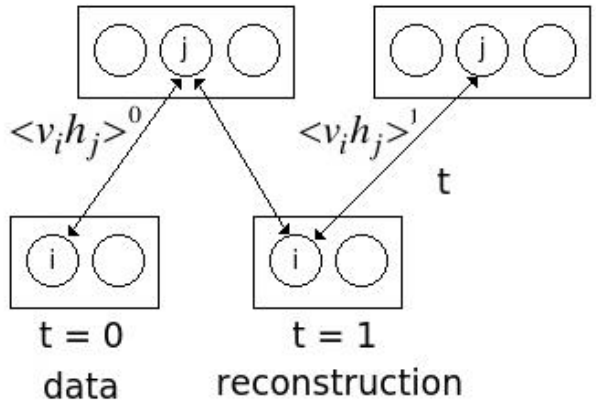
$$p(v,h) = \frac{1}{Z} e^{-E(v,h)}$$
$$Z = \sum_v \sum_h e^{E(v,h)}$$
$$E(v,h) = -\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} h_i v_j - \sum_{j=1}^{m} b_j v_j - \sum_{i=1}^{n} c_i h_i$$

Training

- $$\max_{w,b,c} \log p(v)$$

$$p(v) = \frac{1}{Z}\sum_h p(v,h) = \frac{1}{Z}e^{-F(v)}$$
$$F(v) = -log\sum_h e^{-E(v,h)} - \sum_{j=1}^{m} b_j v_j$$

- Contrastive Divergence



Risultati

- #unità nascoste = 500
- #epoche = 100

Input Pattern	N	Sensitivity(%)	Specificity(%)	Accuracy(%)	MCC
Binary Pattern	5	9.2	98.92	78.73	0.20
	11	**22.32**	**96.99**	**80.19**	**0.30**

Confronto

Metodo	Input Pattern	N	Sensitivity(%)	Specificity(%)	Accuracy(%)	MCC
SVM	Binary Pattern	**17**	**29.71**	**94.60**	**78.80**	**0.32**
	PSSM	**17**	**34.65**	**95.88**	**80.97**	**0.41**
DAE + SVM	Binary Pattern	**5**	**28.13**	**93.21**	**77.36**	**0.28**
RBM + SVM	Binary Pattern	**11**	**22.32**	**96.99**	**80.19**	**0.30**