UNIVERSITÀ DEGLI STUDI DI PADOVA
FACOLTÀ DI INGEGNERIA

Corso di Laurea Specialistica in Ingegneria Informatica

Ragionamento Qualitativo e Apprendimento Automatico per l'Analisi di Dati di Genomica

LAUREANDO: Matteo Zanini

RELATORE: prof. Silvana Badaloni
CORRELATORE: dott. Francesco Sambo

INDICE

1. Obiettivi della Tesi;
2. Problema Biologico;
3. Rappresentazione Simbolica e Algoritmi di Classificazione;
4. Risultati Sperimentali;
5. Conclusioni.

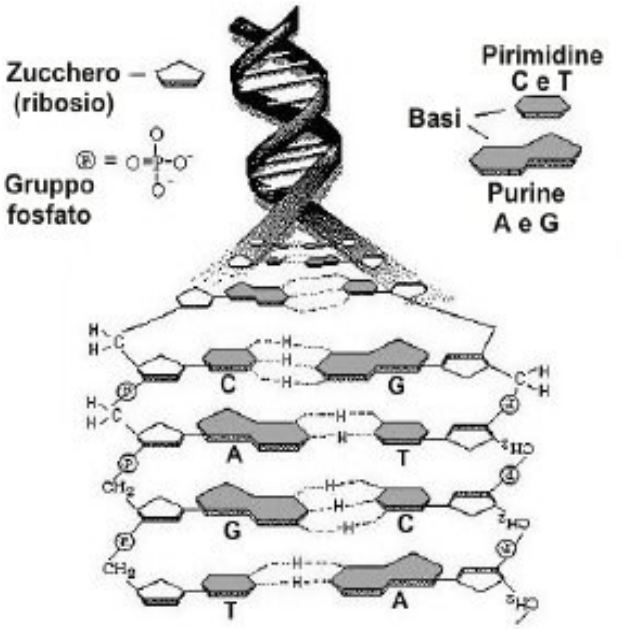
1.0 - OBIETTIVI DELLA TESI

Lo scopo di questa tesi è stato quello di elaborare dati di genomica, tramite algoritmi d'**intelligenza artificiale** per il **ragionamento qualitativo** e per l'**apprendimento automatico**.

Nello specifico, sono stati riconosciuti due obiettivi:

- Ricostruire la rete di regolazione genica;
- Identificare opportune sottostrutture significative, presenti all'interno della rete di regolazione.

2.0 - PROBLEMA BIOLOGICO (TECNOLOGIA DNA-MICROARRAY)



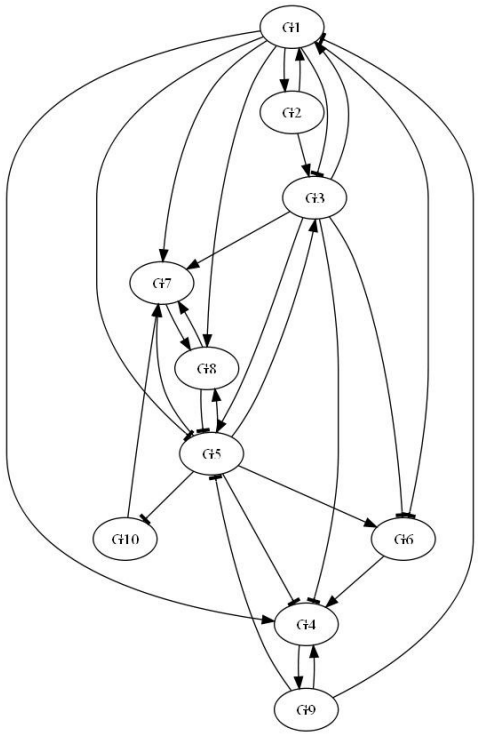
Il DNA (*Acido DesossiriboNucleico*) è costituito da filamenti di nucleotidi (*adenina, timina, citosina, guanina*).

Il processo attraverso il quale il DNA viene tradotto in proteina consta di due fasi fondamentali:

- *Trascrizione*;
- *Traduzione*.

La tecnologia **DNA-microarray**, consente di quantificare centinaia/migliaia di mRNA presenti in un unico esperimento.

2.1 - PROBLEMA BIOLOGICO (RETE DI REGOLAZIONE)



Per descrivere gli stati, dell'ambiente interno di una cellula, è possibile ricorrere ad una rappresentazione, basata sul comportamento reciproco dei geni, detta **rete di regolazione**.

Un modo per rappresentarla è tramite un grafo orientato, in cui i nodi simboleggiano i geni e gli archi i rapporti di regolazione.

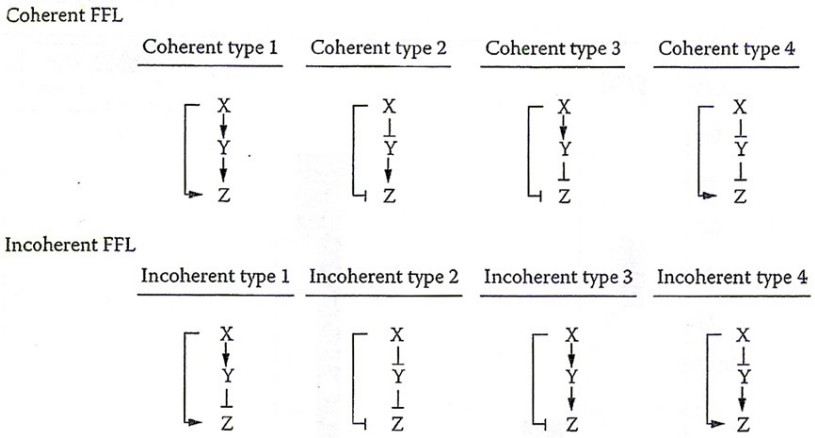
A seconda del tipo di regolazione, i geni sono suddivisi in **promotori** e in **repressori**.

2.2 - PROBLEMA BIOLOGICO (NETWORK MOTIF)

Analizzando diverse reti di regolazione è possibile domandarsi se esistano delle particolari configurazioni che siano ricorrenti e abbiano una certa importanza, nella dinamica della cellula.

Nel caso sia appurato che una certa sottorete occorra frequentemente si parla di **network motif**.

I **Feed-Forward Loop** (*FFL*) risultano statisticamente le configurazioni più rilevanti, tra quelle costituite da tre nodi.



3.0 - RAPPRESENTAZIONE SIMBOLICA

I dati ottenuti tramite DNA-microarray sono serie temporali campionate.
Per elaborarli in maniera efficiente è stata introdotta una rappresentazione, denominata simbolica.

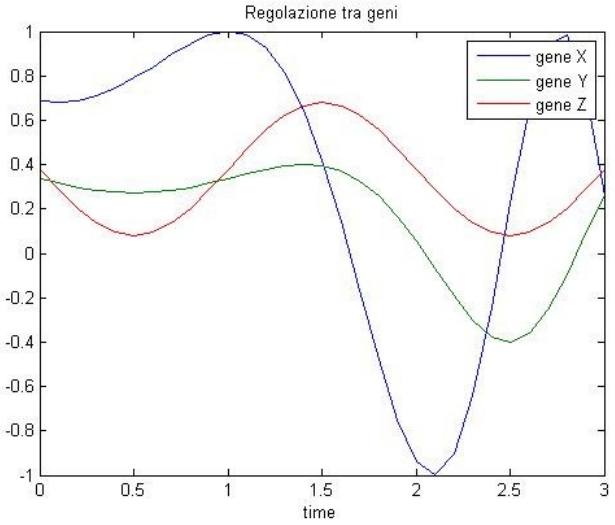
I passaggi principali per determinare tale rappresentazione sono:

- Determinazione dei punti dominanti, tramite algoritmo **CAL** (*Chord and Length*);
- Determinazione delle astrazioni temporali.

L'alfabeto della rappresentazione simbolica utilizzato è il seguente: massimo (**M**), minimo (**m**), crescente (**c**), decrescente (**d**), flesso (**f**), stazionario (**s**), zero (**z**), saturazione(**t**).

Per il confronto tra le stringhe, risultanti dalla rappresentazione simbolica, sono stati usati gli algoritmi **LCString** (*Longest Common String*) e **LCS** (*Longest Common Sequence*).

3.1 - RAPPRESENTAZIONE SIMBOLICA



Gene X: (d, m, f, M, f, m, f, M, d).
Geni Y, Z: (d, m, f, M, f, m, f, M).

Introducendo concetti della **logica fuzzy** è stata modificata la rappresentazione simbolica, per sfruttare al meglio la descrizione dei profili genici.

Con questa modifica si utilizzano delle **funzioni d'appartenenza**, adattate alle caratteristiche principali degli andamenti genici, per esprimere una misura d'appartenenza di un punto dominante ad un simbolo.

Queste modifica è stata incorporata come una variante degli algoritmi LCString e LCS, in modo da renderli più efficaci.

4.0 - RISULTATI SPERIMENTALI

Per l'obiettivo di ricostruire la rete di regolazione genica sono stati utilizzati due dataset:

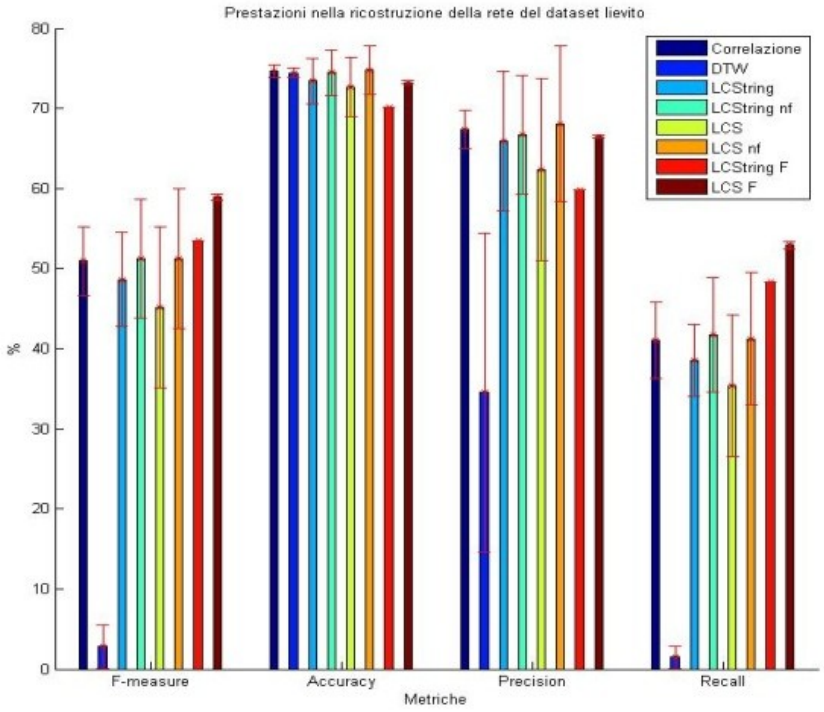
- Dati reali(*ciclo cellulare del lievito Saccharomyces cerevisiae*), strutturati in un'unica rete, con 24 geni, campionati in 14 punti;
- Dati simulati, costituiti da 5 reti, con 10 geni, campionati in 11 punti.

Del secondo dataset sono state usate due versioni: una priva e una con l'aggiunta di una componente d'errore nel processo di creazione dei dati.

I metodi confrontati tra loro, per ricavare la rete di regolazione, sono i seguenti:

- Correlazione, ricavata dalla distanza euclidea tra i profili genici;
- DTW (Dynamic Time Warping);
- LCString-CAL (Longest Common String CAL);
- LCS-CAL (Longest Common Sequence CAL).

4.1 - RISULTATI SPERIMENTALI



Dai risultati si nota che:

- I metodi basati sulla rappresentazione numerica risultano peggiori rispetto a quelli basati sulla rappresentazione simbolica;
- Le varianti fuzzy sono più efficienti delle altre.

4.2 - RISULTATI SPERIMENTALI

Per il riconoscimento dei FFL sono state utilizzate le **SVM** (*Support Vector Machine*) per il processo di classificazione.

Obiettivo: distinguere la tipologia di FFL, a partire da delle terne (X,Y,Z) di profili di espressione genica.

Si è proceduto per gradi:

- Analizzare i problemi disgiunti di classificazione binaria, considerando un solo tipo alla volta;
- Affrontare il problema più complesso di un classificazione multiclasse.

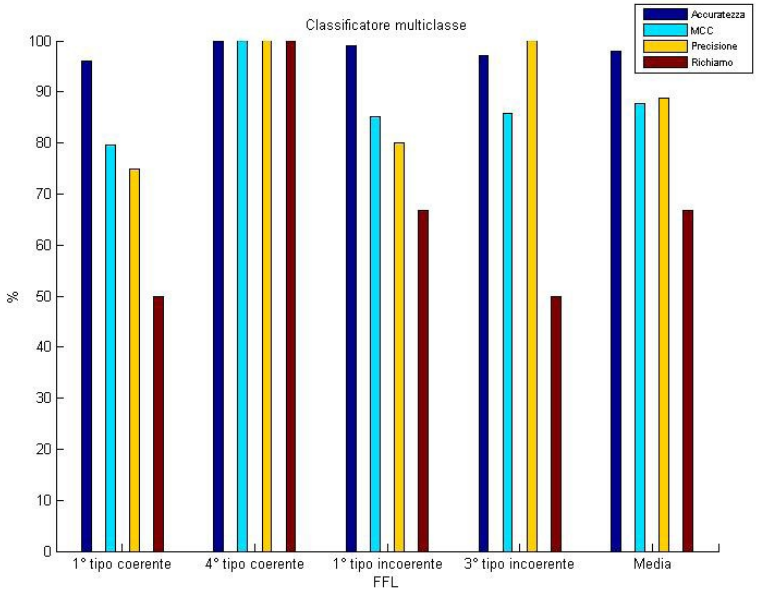
Il Dataset è stato ottenuto estraendo tutti i FFL da 4 reti note ed è stato così suddiviso:

- Tre set per il tuning dei parametri, con un 3-fold cross validation, e poi per allenare il classificatore;
- Un set come test indipendente.

4.3 - RISULTATI SPERIMENTALI

Dai risultati si ha, per il problema del classificatore monoclasse:

- Accuratezza 83,333% del classificatore di maggioranza;
- Accuratezza 83,333% con rappresentazione simbolica;
- Accuratezza del 100% per i FFL del primo e del quarto tipo coerente, con il kernel RBF;
- Accuratezza media, con kernel RBF, del 93,953%.



Per il classificatore multiclasse si ottiene un'accuratezza del 98,039% con la rappresentazione numerica, mentre con quella simbolica si mantiene l'accuratezza di 83,333%.

5.0 - CONCLUSIONI

Obiettivo 1: Ricostruzione della rete di regolazione genica.

- La rappresentazione simbolica dà risultati migliori, rispetto a quella numerica;
- Il miglior algoritmo risultante è LCString, nella sua variante fuzzy;
- Le performance migliori ottenute sono confrontabili con lo stato dell'arte (algoritmi ARACNE e REVEAL).

Obiettivo 2: Classificazione delle tipologie di FFL a partire da terne di espressione genica.

- La rappresentazione numerica dà risultati migliori, rispetto a quella simbolica;
- Si ottiene un'elevata accuratezza media (93,953%), con il classificazione monoclasse con kernel RBF;
- Si ottiene un'elevata accuratezza con il classificazione multiclasse (98,039%).

Grazie