

Relazione sull'attività svolta nell'ambito del programma di mobilità breve CNR 2009

Proponente: *Domenico Talia*

Fruitore: *Mario Rosario Guarracino*

Istituto di afferenza del Fruitore: *Istituto di Calcolo e Reti ad Alte Prestazioni del Consiglio Nazionale delle Ricerche*

Descrizione dettagliata dell'Istituzione ospitante: *Center for Applied Optimization, Department of Industrial and System Engineering, University of Florida*

Numero di codice del dipartimento (v. All.2) 9

Titolo del programma: *Metodi di apprendimento automatico per la predizione dell'interazione tra le proteine*

Obiettivi

Sviluppo di algoritmi, metodi e strumenti software dell'apprendimento automatico (supervisionato e non supervisionato) ad alte prestazioni per la predizione delle interazioni tra proteine umane. Validazione e test delle soluzioni sviluppate.

Attività da svolgere

- 1. Analisi delle soluzioni esistenti per la predizione in silicio dell'interazione delle proteine mediante metodi di apprendimento automatico.*
- 2. Individuazione di casi di studio riferiti al proteoma umano per la validazione delle soluzioni prodotte.*
- 3. Disegno ed implementazione prototipale di algoritmi di apprendimento per la predizione delle interazioni.*
- 4. Analisi dei risultati mediante l'utilizzo di metodologie di validazione standard, confronto con le altre soluzioni esistenti e casi di studio.*
- 5. Disseminazione dei risultati mediante pubblicazioni scientifiche.*

Descrizione del progetto

Gli studi di proteomica sono usati come mezzi di indagine sui processi biologici. L'individuazione di marcatori e la loro caratterizzazione ha numerosi vantaggi, tra cui l'opportunità di uso diagnostico e l'aumento della conoscenza sugli effetti patologici a livello molecolare, richieste per nuovi farmaci e terapie. Recentemente sono stati proposti metodi di apprendimento supervisionato per lo studio del proteoma, che lasciano intravedere la possibilità di analizzare in maniera efficiente e veloce l'elevato numero di proteine presenti nel proteoma umano. Il presente progetto si propone dunque di superare tali limiti mediante l'individuazione di nuove metodologie in collaborazione con i ricercatori del CAO, che hanno forti competenze in materia. L'attività si inquadra nella collaborazione esistente su temi di interesse del progetto Bioinformatica del Dipartimento ICT del CNR, come attestato dal 2005 da pubblicazioni JCR.



Risultati

Si sono analizzate le soluzioni esistenti in letteratura per la predizione delle interazioni basate sull'utilizzo della struttura primaria, e si sono scelti *k-Nearest Neighbor* (Mitchel, 2005) e *ReGEC* (Guarracino et al, OMS 2007) come algoritmi di apprendimento supervisionato. Tra i network di interazione presenti in letteratura, è stato scelto come caso di studio quello di Bell et al (Plos Genet., 2009) sui meccanismi di invecchiamento.

Si è quindi implementata una soluzione algoritmica che permettesse di predire l'interazione tra proteine, a partire dai dati del database *HPRD* (www.hprd.org) e da una rappresentazione delle proteine basata sul metodo delle triplette di Shen et al (Pnas, 2007).

Si è quindi valutata la qualità della soluzione algoritmica mediante la validazione statistica dei risultati. Poichè i primi risultati ottenuti migliorano quelli esistenti in letteratura, sono stati sottomessi per la pubblicazione negli atti del IEEE International Conference on Complex, Intelligent and Software Intensive Systems, che si terrà nel 2010 a Cracovia (Polonia).

Nel corso del soggiorno si è presentata l'opportunità di lavorare anche su dati di spettrometria Raman, per lo studio del cancro al polmone. Anche in questo caso, l'applicazione dell'algoritmo *ReGEC* ha dato risultati nella discriminazione di cellule apoptotiche, necrotiche e sane. I risultati sono stati sottomessi per pubblicazione su rivista JCR.

Firma del Proponente

Firma del Fruitore

.....

.....