Al Consiglio Nazionale delle Ricerche
Direzione Generale
Ufficio Paesi Industrializzati- Organismi Int.li
P.le Aldo Moro, 7
00185 ROMA

Oggetto: programma di scambi internazionali per la mobilità di breve durata (Short-term mobility – Anno 2008)

Relazione scientifica sull'attività di ricerca svolta

Durante la permanenza presso l'Istituto per le Applicazioni del Calcolo "M. Picone" di Napoli della Prof.ssa Marianna Pensky (Dep. of Mathematics of the University of Central Florida. Orlando - USA) è stato studiato il problema del clustering di profili di espressione genica in esperimenti di time-course con microarray. Lo studio intrapreso costituisce un aspetto molto importante per la comprensione dei meccanismi di regolazione genica all'interno di una cellula.

L'attività di ricerca si è svolta in due fasi ed ha coinvolto, oltre alla Prof. Pensky, le dott.sse Claudia Angelini (IAC-Napoli) e Daniela De Canditiis (IAC-Roma). In un primo momento sono state studiate le metodologie presenti in letteratura per il suddetto problema mettendo in luce i vantaggi e gli svantaggi dei diversi approcci; in un secondo momento è stato sviluppato un approccio statistico innovativo per la soluzione del problema in esame.

La metodologia elaborata prende spunto da alcuni recenti lavori pubblicati dalla prof.ssa Pensky in collaborazione con le due ricercatrici Angelini e De Canditiis *[C. Angelini et al. (2007)* and *C. Angelini et al. (2008)]* nei quali vengono introdotti modelli funzionali Bayesiani per l'individuazione automatica di profili temporali di espressione genica significativi, rispettivamente per il disegno sperimentale 'one-sample' e 'two-sample'.
Durante la presente visita della Prof. Pensky il modello Bayesiano funzionale è stato esteso al contesto del clustering di profili temporali seguendo un approccio `model based', dove i profili temporali di espressione genica vengono rappresentati con una mistura di infinite distribuzioni (processo Dirichlet) raggruppando nello stesso cluster i profili temporali provenienti dalla stessa componente della mistura, vedi *[Medvedovic and Sivaganesan (2002)]*.
La peculiarità del metodo elaborato è l'incorporazione formale della variabile temporale all'interno del modello statistico attraverso un'espansione in serie dei profili. Inoltre, il vantaggio dell'utilizzo del modello di mistura infinita rispetto a quello di mistura finita risiede nella non necessità di dover indicare esplicitamente un numero di clusters rispetto al quale partizionare i campioni e/o elicitare una distribuzione a priori su tale numero. In questo modo l'incertezza sul numero effettivo di clusters presenti in un data-set viene incorporata nel modello e, quindi, costituisce una parte integrante del processo inferenziale.
Purtroppo, per questo tipo di modelli la dimensionalità del problema in esame (i.e., analisi dati da microarray) rende impossibile la determinazione in forma analitica della soluzione ottimale. Pertanto, la possibilità di fare inferenza è

strettamente collegata alla possibilità di utilizzare un approccio MCMC, ovvero generare campioni casuali dalla distribuzione a posteriori e successivamente determinare una stima della soluzione ottimale. Il problema computazionale non è affatto secondario per il modello di mistura del processo di Dirichlet. In letteratura la proposta più interessante in tal senso è data dall'algoritmo Split-Merge MCMC di *[Jain and Neal (2004)]* e dalle sue successive variazioni proposte in *[Dahl (2006)].* Tali idee sono state incorporate nel modello da noi studiato. In pratica, l'introduzione di una variabile latente per ogni profilo temporale (contenente l'informazione sulla mistura da cui proviene il profilo e quindi codificante la partizione sottostante i dati) permette di esplorare in modo relativamente economico la distribuzione a posteriori di questa variabile e quindi della clusterizzazione. I vantaggi principali di tale metodologia esplorativa della catena sono i seguenti: ad ogni passo un cluster può essere diviso in due oppure due clusters possono essere accorpati (permettendo cosi di esplorare un maggior numero di configurazioni possibili), inoltre ogni movimento è giudicato in base ad una probabilità di accettazione in cui i parametri che distinguono le diverse misture vengono integrati fuori (permettendo un giudizio meno distorto dal particolare valore assunto dai parametri delle misture al passo considerato).

La validazione e l'implementazione della metodologia qui proposta saranno oggetto di futuro lavoro da parte della Prof. Pensky e delle Dott.sse Angelini e De Canditiis.

- C. Angelini, D. De Canditiis, M. Mutarelli and M. Pensky "A Bayesian approach to estimation and testing in time-course microarray experiments" **Statistical Applications in Genetics and Molecular Biology** Vol 6 n°4 (2007).
- C. Angelini, D. De Canditiis and M. Pensky "Bayesian models for the two-sample time-course microarray experiments" to appear on **Computational statistics and Data Analysis** (2008).
- D.B. Dahl "Model-Based clustering for expression data via a dirichlet process mixture model" **in Bayesian Inference for Gene Expression and Proteomics**, Kim-Anh Do, Peter Müller, Marina Vannucci (Eds.), Cambridge University Press (2006).
- S. Jain and R. M. Neal "A Split-Merge Markov Chain Monte Carlo procedure for the Dirichlet process mixture models" **Journal of Computation & Graphical Statistics** Vol 13 n°1, pag 158-182, (2004).
- M. Medvedovic and S. Sivaganesan "Bayesian infinite mixture model based clustering of gene expression profiles" **Bioinformatics** Vol 18 n°9, pag 1194-1206, (2002).

Firma

14/07/2008