Relazione scientifica dell'attività svolta dalla Dott.ssa Claudia Angelini nell'ambito del Programma Short Term Mobility 2010.

Il programma Short Term Mobility 2010 è stato svolto nel periodo dal 19 Aprile al 11 Maggio 2010.

Durante il periodo di permanenza presso l'University of Central Forida, in collaborazione con la Prof. Marianna Pensky, è stato iniziato uno studio sui modelli probabilistici e statistici che sono alla base delle analisi dei dati di RNA-Seq.

Tali tipologia di dati –ottenuti tramite l'utilizzo di sequenziatori massivi di nuova generazione- costituiscono uno degli strumenti di indagine genetica più innovativi presenti attualmente sul mercato ed, al momento, stanno rivoluzionando il panorama della ricerca in campo bio-medico, offrendo la possibilità di guardare alla variabilità umana ed alle malattie mendeliane e complesse da una diversa prospettiva. Al contempo, l'utilizzo di queste nuove piattaforme di sequenziamento massivo sta ponendo alla comunità scientifica nuove sfide sia dal punto di vista sperimentale che computazionale e modellistico.

La collaborazione ha riguardato lo sviluppo di un modello statistico alla base della generazione di tali dati.

Si è partiti dalla descrizione del problema biologico e del contesto sperimentale mettendo in luce le differenze esistenti (sia da un punto di vista biologico che computazionale) tra la precedente tecnologia dei microarray e le moderne tecnologie del sequenziamento massivo. Si è concluso che il trascrittoma costituisce l'insieme completo dei trascritti in una cellula, pertanto risulta di particolare rilievo dal punto di vista prognostico la ricostruzione e la quantificazione del profilo di espressione individuale e la sua associazione ad una specifica fase di sviluppo o condizione fisiologica. Comprendere il trascrittoma è indispensabile per individuare ed interpretare correttamente gli elementi funzionali del genoma, che rivelano le componenti molecolari di cellule e tessuti, ed anche per la comprensione dello sviluppo delle patologie. Fino ad oggi erano state sviluppate diverse tecnologie atte ad individuare e quantificare trascritti specifici, compresi i microarray sui quali precedenti collaborazioni avevano portato a diverse pubblicazioni scientifiche. Tuttavia, studi recenti dimostrano che il sequenziamento diretto del RNA (RNA-Sequencing) mediante i moderni sequenziatori può fornire una misura dei livelli di espressione dei trascritti ed un'individuazione di forme di splicing alternativo di gran lunga più precisa rispetto ad altri metodi esistenti. Si è passati quindi ad uno studio approfondito della letteratura ed all'esplorazione dei dati di alcuni esperimenti reali prodotti dal sistema SOLiD 3 disponibile presso l'IGB-CNR con il quale è attiva una collaborazione. E' emerso che l'analisi statistica di dati di RNA-Seq è particolarmente interessante: i valori di espressione sono infatti ottenuti in termini di contatori di brevi sequenze opportunamente allineate in specifici loci e che, di conseguenza, tutti i metodi statistici disponibili per l'analisi di microarray non risultano applicabili in tale contesto. Inoltre è stato notato che la maggior parte dei metodi proposti per le analisi non parte da uno studio modellistico, ma si basa su considerazioni empiriche. Si è pertanto intrapreso uno studio modellistico del processo alla basa della generazione dei dati di RNA-Seq che tenga in considerazione il fatto che il dato effettivamente prodotto è costituito da decine/centinaia di milioni di brevi sequenze (i.e., frammenti di trascritto) che devono essere allineate al genoma di riferimento prima di procedere alla quantificazione. La produzione delle sequenze è un processo stocastico, mentre il successivo allineamento è un processo inferenziale non esente da incertezza (vedi ad esempio il caso delle sequenze che possono allineare in posizioni multiple a causa della ripetitività del genoma pur essendo state generate da una singola posizione, o il caso di varianti strutturali presenti nel campione o di errore di lettura introdotto dai sequanziatori). L'intero processo di produzione delle sequenze è stato modellato con un processo di Poisson con funzione di intensità costante a tratti (con valori strettamente positivi su le zone esoniche espresse e con intensità nulla altrove). Mentre i dati osservati in termini di ricoprimento

lungo il genoma risultano ulteriormente contaminati da un errore fortemente (localmente) correlato la cui forma dipende dalle specifiche dell'algoritmo di allineamento (i.e., da come vengono trattati gli allineamenti multipli, in numero di mismatches consentiti e lo splicing delle sequenze). In ogni caso il rumore risulta essere asintoticamente gaussiano.
La collaborazione ha anche riguardato la sviluppo di un modello computazionale per la quantificazione di isoforme annotate e per l'individuazione automatica della presenza di forme di splicing non annotate. Tale modello è al momento in fase di implementazione.

La collaborazione è proseguita successivamente con la visita della Prof. Pensky presso l'IAC di Napoli dal 26 Giugno al 2 Luglio. La collaborazione ha, in questo caso, coinvolto anche la Dott.ssa I. De Feis (IAC-CNR).

Durante questa fase sono state apportate ulteriori correttivi al modello (in particolare relativi alla determinazione della struttura di correlazione) ed è stato quindi proposto un algoritmo computazionale per la segmentazione ed il denoising dei dati di RNA-Seq. La collaborazione riprenderà a settembre con la costruzione di un modello simulato di generazione dati di RNA-Seq e con l'implementazione dell'algoritmo proposto. Superata tale fase si procederà successivamente all'applicazione della metodologia proposta per l'analisi di dati reali.

Napoli, 19 Luglio 2010

Dott.ssa Claudia Angelini