UNIVERSITA' DEGLI STUDI DI MILANO – BICOCCA

Dottorato di Ricerca in Informatica

Dottorato di Ricerca in Informatica XXII ciclo

Relazione annuale, A.A. 2006/07

Dottorando: Yuri PIROLA

Supervisore: Prof. BONIZZONI Paola

Tutor: Prof. POMELLO Lucia

Attività formativa

Corsi e Scuole

Durante l'anno 2006-2007 ho frequentato i corsi, organizzati internamente al dipartimento, di "Informatica Teorica", tenuto dalla Prof. Pomello, e di "Web Services: concetti, applicazioni e problemi aperti", tenuto dal Prof. De Paoli. Di quest'ultimo ho sostenuto con successo l'esame finale.

Attività didattica

Supporto all'attività didattica del corso "Laboratorio di Linguaggi di Programmazione" tenuto, nel corso del primo semestre dell'a.a. 2006/07, dal Prof. Antoniotti Marco (Bando LAE – 24 ore).

Attività di Ricerca

Descrizione dell'attività di ricerca

I miei interessi di ricerca riguardano lo studio della complessità computazionale di problemi combinatori in bioinformatica e il disegno di algoritmi (esatti o approssimanti) per la loro soluzione. In particolare la mia attività si è focalizzata sullo studio di tecniche algoritmiche e strutture dati per trattare problemi di allineamento e confronto di sequenze genomiche e per la ricostruzione di reti filogenetiche.

In tale ambito mi sono occupato della soluzione di due differenti problemi relativi ai miei interessi di ricerca: (1) inferenza di aplotipi secondo il modello coalescente e (2) fattorizzazione e clusterizzazione di sequenze di EST (Expressed Sequence Tag).
I risultati ottenuti nelle ricerche svolte su queste ultime due tematiche sono in fase di stesura per essere sottoposti a pubblicazione.

Il problema dell'inferenza di aplotipi è biologicamente motivato da considerazioni di carattere tecnologico. Il sequenziamento del patrimonio genetico degli individui di un'intera popolazione è utile per poter determinare associazioni fra mutazioni genetiche e predisposizione a determinate malattie. Per contro, il sequenziamento è anche un'operazione lenta e molto costosa. Ai fini degli studi di associazione, è però possibile restringere l'attenzione solamente a determinati nucleotidi delle sequenze genetiche dei cromosomi di un individuo, chiamati *single nucleotide polymorphism* o SNP. Negli organismi evoluti, il patrimonio genetico è rappresentato da un insieme di coppie di cromosomi, detti omologhi, che provengono, eventualmente ricombinati, dall'eredità genetica di entrambi i genitori. Attualmente esiste una tecnologia a larga scala e a basso costo che consente di ottenere la coppia di basi presenti in una data posizione su una coppia di cromosomi omologhi ma che non permette di stabilire l'effettivo cromosoma di provenienza di ciascuna delle due basi. Di conseguenza per ogni individuo di una popolazione è possibile ottenere una lista, chiamata genotipo dell'individuo, di coppie non ordinate di basi. Un aplotipo di un individuo, invece, è una singola sequenza di basi azotate presenti in corrispondenza degli SNP di un singolo cromosoma. Si dice che una coppia di aplotipi risolvono un dato genotipo se la lista delle coppie di basi prese dai due aplotipi in posizioni corrispondenti coincide con il genotipo. Il problema dell'*inferenza di aplotipi* consiste, dato l'insieme dei genotipi di n individui di una popolazione, nel determinare un insieme di $2n$ aplotipi tale per cui per ogni individuo esiste una coppia di aplotipi che risolvono il suo genotipo e che l'intero insieme degli aplotipi soddisfa un modello biologico di riferimento. In letteratura i modelli biologici comunemente utilizzati sono quello coalescente e quello della massima parsimonia.
In collaborazione con Prof. Bonizzoni, ci si è interessati alla risoluzione algoritmica del problema dell'inferenza di aplotipi sotto l'assunzione del modello coalescente, anche chiamato Perfect Phylogeny Haplotyping Problem (PPHP). In questa formulazione, l'insieme degli aplotipi corrisponde all'insieme delle foglie di un albero etichettato che rappresenta l'evoluzione degli SNP presi in esame. In particolare si è affrontato il problema di disegnare ed implementare un algoritmo lineare per il problema. In letteratura è stato recentemente pubblicato un algoritmo lineare, ma quello da noi individuato si basa su una idea originale che mostra interessanti connessioni del problema con la teoria dei grafi: l'insieme dei caratteri forma un insieme parzialmente ordinato secondo una particolare relazione d'ordine. Inoltre tutte le soluzioni di un'istanza del problema rappresentano particolari sottoinsiemi di questa relazione d'ordine che soddisfano determinate proprietà. Il PPHP è stato quindi risolto attraverso la costruzione e la manipolazione di sottografi del diagramma della relazione d'ordine sui caratteri. Il contributo personale a questa collaborazione riguarda (a) l'assistenza nella definizione teorica

dell'algoritmo, (b) l'implementazione dell'algoritmo in linguaggio di programmazione e (c) la relativa sperimentazione.

Il problema della *k*-fattorizzazione di un frammento di una sequenza espressa (EST) rispetto ad una sequenza genomica di riferimento consiste nel determinare una partizione del frammento in al più *k* fattori tale per cui i singoli fattori possono essere identificati nella genomica di riferimento nello stesso ordine con cui compaiono all'interno dell'EST e con una distanza di edit limitata da una costante *e*. Gli EST hanno un ruolo fondamentale nell'identificazione di trascritti di geni e sono molto utili per la comprensione dell'espressione-tessuto specifica di determinati geni. La crescente attenzione che la comunità biologica ha rivolto verso le potenzialità degli EST ha portato a un aumento esponenziale del numero di EST individuati e memorizzati in banche dati specializzate. Di conseguenza, si rendono necessarie nuove tecniche algoritmiche in grado di trattare in modo rapido ed efficiente i nuovi dati. In questo filone ci si è concentrati sullo sviluppo di un algoritmo approssimato di *k*-fattorizzazione attraverso l'utilizzo di una struttura dati nota in letteratura come alberi suffisso. A partire dall'idea iniziale della Prof. Bonizzoni e con la sua supervisione, il contributo personale a questa ricerca ha riguardato la definizione dettagliata del metodo di fattorizzazione e la sua sperimentazione su dati reali. I risultati preliminari della sperimentazione mostrano un effettivo incremento prestazionale rispetto a programmi simili comunemente utilizzati mantenendo, tuttavia, un'adeguata precisione dei risultati forniti. Attualmente si sta lavorando alla realizzazione di un'applicazione dell'algoritmo di *k*-fattorizzazione per ridurre la ridondanza informativa delle banche dati di EST attraverso un'opportuna clusterizzazione degli stessi. Il vantaggio di questo processo di clusterizzazione consiste principalmente nella possibilità di ridurre la dimensione dell'input di programmi di analisi di EST e, di conseguenza, il loro tempo di esecuzione.

Durante la prima parte dell'anno di dottorato, in collaborazione con Dott. Vanneschi, ho completato la ricerca iniziata con il lavoro di tesi della laurea specialistica riguardante lo studio della difficoltà dei paesaggi di fitness booleani in programmazione genetica. Durante questo lavoro, sono state definite alcune misure relative alle caratteristiche di un modello di funzionamento della programmazione genetica, chiamato paesaggio di fitness, applicata a problemi di sintesi automatica di funzioni booleane. Si è quindi studiato la distribuzione dei valori di queste misure rispetto ai problemi classici della letteratura della sintesi delle funzioni binarie di parità e multiplexer. I risultati ottenuti, pubblicati in [1], mostrano effettivamente che esiste una relazione fra certe caratteristiche del paesaggio di fitness e la difficoltà della programmazione genetica di determinare una soluzione ottima al problema.

Collaborazioni a progetti di ricerca

Collaborazione al progetto italiano FIRB – Bioinformatica per la Genomica e la Proteomica. All'interno di questo progetto si è sviluppato il problema della fattorizzazione e clusterizzazione di EST descritto precedentemente.

Pubblicazioni

[1] Leonardo Vanneschi, Marco Tomassini, Philippe Collard, Sébastien Verel, *Yuri Pirola* and Giancarlo Mauri. A comprehensive view of fitness landscapes with neutrality and fitness clouds. In Marc Ebner, et al., editors, Proceedings of the 10th european conference on genetic programming, volume 4445 of Lecture Notes in Computer Science, pages 241--250. Valencia, Spain, 2007. Springer.