

Relazione Annuale - A.A. 2007/08

Yuri Pirola

Dottorato di Ricerca in Informatica - XXII Ciclo

Dipartimento di Informatica Sistemistica e Comunicazione

Università degli studi di Milano–Bicocca

11 settembre 2008

Titolo: “Problemi combinatori nello studio di variazioni genetiche”

Relatore: Prof. Paola Bonizzoni

Tutor: Prof. Lucia Pomello

1 Attività Formativa

Corsi di dottorato Durante l’anno accademico 2007/08 ho partecipato ai seguenti corsi di dottorato e ne ho sostenuto la relativa prova di esame.

- *Multilevel Models*, 1-5 ottobre 2007, tenuto dalla Dott.ssa Marta Blangiardo e organizzato nell’ambito della Scuola di Dottorato di Scienze.
- *Reti Bayesiane*, 19 ottobre - 14 novembre 2007, tenuto dal Prof. Enrico Fagioli.
- *Teoria e Applicazioni del Calcolo Evoluzionistico*, 13 febbraio - 7 marzo 2008, tenuto dal Dott. Leonardo Vanneschi.
- *Biomolecular Computing: Theory and Experiments*, 31 marzo - 14 aprile 2008, tenuto dalla Prof.ssa Natasha Jonoska.

Ho inoltre partecipato alle due edizioni del corso di inglese organizzato dal dipartimento tenutesi nei mesi Ottobre-Dicembre 2007 (1a edizione) e Gennaio-Aprile 2008 (2a edizione).

Scuole Ho partecipato alla Lipari International Summer School on Bioinformatics and Computational Biology intitolata “Biological Networks: Evolution, Interaction and Computation” organizzata dal Prof. Alfredo Ferro del dipartimento di Matematica e Informatica dell’università di Catania. La scuola prevedeva un esame finale che ho sostenuto con successo.

2 Attività Didattica

L’attività didattica dell’anno accademico si è così articolata:

- esercitazioni del corso di *Bioinformatica*, laurea di I livello in informatica, 12 ore;
- lezioni frontali del corso di *Informatica Generale 2*, laurea in lingue e letterature straniere, università degli studi di Bergamo, 30 ore.
- esercitazioni del modulo di *Analisi di Algoritmi* del corso *Tecniche di Analisi e Verifica*, laurea magistrale in informatica, 12 ore.

Inoltre sono stato correlatore di due tesi di primo livello in informatica e sono attualmente correlatore di una tesi di primo livello e di una tesi magistrale in informatica presso il nostro dipartimento.

3 Attività di Ricerca

3.1 Descrizione dell’attività di ricerca

L’attività di ricerca svolta durante l’anno accademico è stata sviluppata su due fronti: lo studio del problema di inferenza di aplotipi secondo il modello di pura parsimonia e basato su genotipi Xor e lo sviluppo di un metodo per l’allineamento di sequenze espresse.

Pure Parsimony Xor Haplotyping Obiettivo dei problemi di inferenza di aplotipi è l’ottenimento mediante strumenti computazionali delle informazioni sull’eredità genetica che gli individui hanno ottenuto dai genitori a partire da informazioni riguardanti il loro patrimonio genetico. La mia attività in quest’ambito ha riguardato la definizione, la modellazione, l’analisi delle proprietà e la risoluzione di un problema di inferenza di aplotipi basato sul modello biologico della pura parsimonia (un’applicazione del principio del rasoio di Occam) e che considera i soli siti eterozigoti di un individuo (ovvero le posizioni del patrimonio genetico in cui l’eredità materna differisce da quella paterna). Il

problema computazionale che ne deriva è stato chiamato *Pure Parsimony Xor Haplotyping (PPXH)*.

Lo studio del problema di PPXH ha previsto una fase iniziale di definizione e modellazione del problema stesso mediante strumenti combinatori. A partire da questa fase preliminare, lo studio del problema si è articolato nelle seguenti attività di ricerca:

- studio della complessità computazionale;
- studio della complessità parametrica;
- definizione di restrizioni del problema risolvibili in tempo polinomiale;
- disegno di tecniche euristiche per la soluzione approssimata di istanze generali del problema.

Allo stato attuale si sta completando l'analisi della complessità computazionale del problema, mentre si sono già ottenuti risultati interessanti nelle altre tre direzioni. Si prevede di sottomettere una pubblicazione al completamento dello studio.

Allineamento di Sequenze Espresse Le sequenze espresse (EST) sono frammenti di sequenze nucleotidiche che codificano una proteina. Esse contengono solo le parti codificanti del patrimonio genetico di un individuo e, per tale motivo, possono essere utilizzate per inferire computazionalmente la struttura di un gene e gli eventi di Splicing Alternativo che lo interessano, informazioni importanti nello studio, in particolare, di malattie genetiche e di resistenze ai farmaci. Tuttavia il problema di allineamento di queste sequenze ha alcune peculiarità che lo distinguono fortemente dai problemi classici di allineamento dell'era pre-genomica.

Durante quest'anno ho partecipato in prima persona all'ideazione, alla realizzazione e alla sperimentazione su dati reali di un nuovo metodo di allineamento di sequenze espresse. I risultati della sperimentazione hanno evidenziato un'accuratezza comparabile ad altri metodi presenti in letteratura richiedendo, però, una quantità generalmente inferiore di risorse computazionali. Nonostante i buoni risultati ottenuti (sui quali si sta preparando una pubblicazione da sottomettere), si è ritenuto di poter sfruttare l'esperienza ottenuta nello sviluppo di questa applicazione per ideare un secondo algoritmo di allineamento che, in particolare, possa sfruttare l'elevata ridondanza delle librerie di sequenze espresse per migliorare ulteriormente l'accuratezza degli allineamenti. Correntemente sono stati studiati i problemi che derivano da questo nuovo metodo e se ne è iniziata l'implementazione.

3.2 Collaborazioni a progetti di ricerca

Collaborazione al progetto italiano FIRB – Bioinformatica per la Genomica e la Proteomica. All'interno di questo progetto, nell'unità organizzativa locale, è stato sviluppato il metodo di allineamento di sequenze espresse che è stato descritto precedentemente.

3.3 Piano di confronto nazionale e internazionale

Il piano di confronto internazionale prevede la sottomissione dei risultati iniziali dei lavori a conferenze e workshop che trattano tematiche di tipo algoritmico. Sulla base del feedback ricevuto e con la maturazione del contributo si prevede la preparazione e la sottomissione dei lavori a riviste di settore.

L'algoritmica è un campo di ricerca maturo e consolidato, quindi esistono diverse conferenze e riviste di rilievo internazionale che trattano tali argomenti. La scelta concreta della conferenza e della rivista a cui sottomettere i lavori sarà quindi dettata dalla tipologia dei risultati ottenuti e delle tecniche impiegate per ottenerli.

3.4 Pubblicazioni

Sottomissione dell'articolo intitolato "Detecting alternative gene structures from spliced ESTs: a computational approach" di P. Bonizzoni, G. Mauri, G. Pesole, E. Picardi, Y. Pirola e R. Rizzi alla rivista "Journal of Computational Biology" (under review).

Candidato: Yuri Pirola

Relatore: Prof. Paola Bonizzoni

Tutor: Prof. Lucia Pomello