Relazione avanzamento tesi di dottorato

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

La ricerca nell'ambito dei problemi combinatori derivanti dallo studio di variazioni genetiche si è svolta, nel corso del secondo anno, secondo due linee, in accordo alla proposta di tesi precedentemente sottoposta. La prima linea riguarda la modellazione combinatoria, lo studio delle caratteristiche di complessità e la risoluzione del problema di inferenza di aplotipi secondo il modello di pura parsimonia a partire da una particolare rappresentazione dei dati genotipici, detta a *genotipi Xor*. La seconda linea, invece, si inquadra nell'ambito della predizione degli eventi di splicing alternativo e mi sono interessato, in particolare, al problema computazionale di allineamento di sequenze espresse. Il lavoro svolto in queste due aree verrà descritto nelle sezioni seguenti. Inoltre, per ciascuna di queste frontiere della ricerca, verranno anche delineati gli sviluppi futuri su cui intendo imperniare il lavoro del prossimo anno.

Pure Parsimony Xor Haplotyping

Il problema dell'*inferenza di aplotipi* consiste nel distinguere, a partire dal patrimonio genetico di una popolazione di individui (cioè dai *genotipi* degli individui della popolazione), l'eredità genetica trasmessa dai genitori di ciascun individuo (cioè le coppie di *aplotipi* degli individui in esame). Il processo di inferenza deve essere guidato da un modello biologico di riferimento che consenta di risolvere l'ambiguità che è inerente ai dati in input del problema computazionale generale di inferenza. In particolare, un modello biologico di primaria importanza nella letteratura di questo settore è quello di *pura parsimonia*, ovvero un'applicazione del principio del rasoio di Occam che indica come biologicamente valida la soluzione che consente di generare (in gergo tecnico, spiegare) i genotipi della popolazione utilizzando il minimo numero di aplotipi differenti. La rappresentazione dei dati genotipici è il secondo elemento che gioca un ruolo chiave nella definizione del problema di inferenza.

Recentemente è stato proposto uno schema di codifica dei genotipi, motivato da una nuova tecnica di sequenziamento su larga scala a costi ridotti, chiamato a *genotipi Xor*. In tale schema, il patrimonio genetico dell'individuo è rappresentato da vettori sull'insieme $\{0, 1\}$ in cui gli elementi posti al valore 1 corrispondono a posizioni del genoma dove vi è ambiguità tra il carattere ereditato da padre e il carattere ereditato dalla madre (siti o caratteri eterozigoti). I restanti elementi, invece, presentano uniformità tra le eredità parentali (siti o caratteri omozigoti).

Il problema di inferenza di aplotipi studiato nel corso dell'anno, chiamato *Pure Parsimony Xor Haplotyping* (PPXH), si basa sul modello della pura parsimonia e adotta la rappresentazione Xor dei genotipi. Si è affrontato lo studio di questo problema computazionale sotto diversi aspetti e con metodologie differenti come illustrato di seguito. In particolare l'attività di ricerca si è concentrata sui seguenti punti.

- Modellazione del problema mediante strumenti combinatori e studio delle proprietà del modello.
- Studio della complessità computazionale del problema, con particolare riguardo alle caratteristiche di approssimabilità dello stesso.
- Studio della complessità parametrica del problema.
- Definizione di restrizioni del problema risolvibili in tempo polinomiale.
- Sviluppo di euristiche per la soluzione approssimata del problema.

Le sezioni seguenti andranno a riepilogare, per ciascun punto, l'attività di ricerca che ho svolto e i risultati che ho ottenuto.

Definizione e modellazione del problema Il problema di Pure Parsimony Xor Haplotyping è stato così definito.

Pure Parsimony Xor Haplotyping, PPXH
Istanza: Un insieme G di Xor genotipi.
Soluzione: Un insieme H di aplotipi tale per cui, per ogni $g \in G$, esiste una coppia di aplotipi h_1, h_2 in H che risolve g.
Misura: La cardinalità di H.

Le soluzioni ammissibili del problema sono state modellate attraverso la definizione di una nuova particolare tipologia di grafi etichettati, chiamati *grafi Xor*, riconducendo il problema di inferenza a un problema di ricostruzione di un grafo di dimensione minima che soddisfa particolari vincoli sull'insieme dei suoi cicli.

Il modello delle soluzioni è stato impiegato proficuamente nello studio delle caratteristiche del problema di PPXH che vengono descritte nel seguito.

Complessità computazionale Lo studio della complessità computazionale del problema è un'attività ancora in corso e, nonostante si congetturi che il problema sia NP-hard, la dimostrazione non è stata ancora completata. In particolare ci si sta concentrando sulla costruzione di una L-riduzione (una riduzione fra problemi di ottimizzazione che preserva il fattore di approssimazione) a partire dal problema di Vertex Cover su grafi cubici e

triangle-free (ovvero in cui tutti i vertici hanno grado 3 e che non contengono cicli di lunghezza 3). Se si riuscisse a completare tale riduzione, si arriverebbe a dimostrare che, nell'ipotesi che $\mathsf{P} \neq \mathsf{NP}$, PPXH è APX-hard e, di conseguenza, si potrebbe concludere che non esiste un schema di approssimazione per PPXH. In altre parole si riuscirebbe a dimostrare l'esistenza di un qualche $\varepsilon > 1$ per cui PPXH non può essere ε-approssimato in tempo polinomiale nella dimensione dell'istanza. Si noti che il risultato che si sta cercando di ottenere sia una caratterizzazione più forte della dimostrazione che PPXH è NP-hard, da cui deriverebbe solo l'impossibilità (nell'ipotesi ragionevole di $\mathsf{P} \neq \mathsf{NP}$) di risolvere esattamente PPXH in tempo polinomiale.

Complessità parametrica La complessità parametrica del problema è uno strumento che consente di confrontare, utilizzando le parole di Fellows e Downey, i fondatori di tale teoria, l'intrattabilità computazionale di problemi diversi, ovvero la relativa "difficoltà" dell'essere risolti esattamente. Oltre ad essere uno studio di interesse teorico perché completa la caratterizzazione dell'intrattabilità del problema, lo studio della complessità parametrica ha importanti applicazioni di carattere pratico perché consente il disegno di algoritmi esatti ed efficienti (ma non, ovviamente, polinomiali) per la risoluzione del problema.

Lo studio della complessità parametrica di PPXH mi ha già portato a dimostrare la trattabilità del problema fissato un parametro, ovvero che PPXH $\in \mathsf{FPT}$. Questa dimostrazione è avvenuta tramite la tecnica di riduzione del problema al nucleo (*kernelization*), in cui si riduce l'istanza in ingresso a un'istanza equivalente la cui dimensione è limitata superiormente da una funzione polinomiale del costo della soluzione ottima del problema. È importante notare come la dimostrazione di questo risultato è stata resa possibile dall'utilizzo delle proprietà delle soluzioni che sono state evidenziate nella fase di definizione e modellazione del problema.

Algoritmi esatti polinomiali per restrizioni del problema Di interesse applicativo è stato il disegno di tre algoritmi esatti e polinomiali per la risoluzione di altrettante restrizioni del problema che si potrebbero verificare considerando dati reali. In particolare sono state risolte le restrizioni del problema in cui:

- ciascun genotipo ha al più 2 caratteri eterozigoti (ovvero ciascun vettore non contiene più di due elementi con valore 1);
- ciascun carattere appare come eterozigote in al più due genotipi (ovvero al più due vettori hanno il valore 1 in ciascuna posizione);
- l'insieme dei genotipi ammette una "graph realization", cioè possiede una proprietà studiata nell'ambito della ricerca operativa e della teoria dei matroidi.

Gli algoritmi polinomiali disegnati sono in grado di fornire una soluzione esatta in tempo polinomiale alle restrizioni descritte. Inoltre, essi potrebbero essere utilizzati come routine efficienti all'interno di strategie euristiche per la soluzione approssimata di istanze generali del problema. Tuttavia, allo stato attuale, la possibilità di impiegarli a tal fine è stata solo ipotizzata e non è stata ancora indagata nel dettaglio.

Euristiche Il disegno di tecniche euristiche per la risoluzione approssimata del problema PPXH si è concentrato sull'utilizzo di tecniche di tipo evolutivo basate su algoritmi genetici. Questa attività è stata svolta all'interno del corso di dottorato di "Teoria e Applicazione del Calcolo Evoluzionistico". In particolare è stata confrontata la performance degli algoritmi genetici applicati a questo problema adottando tre diversi schemi di codifica delle soluzioni. I primi due schemi di codifica erano basati su due rappresentazioni dirette delle soluzioni in stringhe binarie di lunghezza fissa. Il terzo schema, invece, ha rappresentato mediante stringhe binarie il grafo Xor definito durante l'attività di modellazione del problema. I risultati dello studio hanno chiaramente evidenziato una miglior performance degli algoritmi genetici che utilizzavano la rappresentazione basata sui grafi Xor a discapito degli algoritmi genetici che utilizzavano le due rappresentazioni dirette delle soluzioni. Inoltre, la rappresentazione basata su grafi Xor ha mostrato un fattore di approssimazione pressoché costante al variare della dimensione delle istanze, mentre gli altri due schemi di codifica hanno mostrato un deciso peggioramento del fattore di approssimazione all'aumentare della dimensione dell'input. Questi risultati sono un'ulteriore prova dell'importanza dell'attività preliminare di studio e di modellazione del problema.

Sviluppi futuri

Lo studio di questo problema prevede, nel corso del prossimo anno, il completamento dell'attività ancora in corso di studio della complessità computazionale del problema. Inoltre, è di sicuro interesse teorico e applicativo, il disegno di un algoritmo approssimato polinomiale (con fattore di approssimazione garantito, a differenza della tecnica euristica fin qui disegnata). Nuovi algoritmi polinomiali esatti per ulteriori restrizioni del problema, nonché nuovi algoritmi parametrici, sono elementi di interesse che potrebbero essere considerati nella pianificazione delle attività.

Nell'ambito più generale dei problemi di inferenza di aplotipi, invece, si potrebbero studiare i problemi combinatori derivanti dall'adozione di nuovi modelli biologici di riferimento, come il modello dei *caratteri persistenti* o dei *galled tree* e delle *reti filogenetiche*. L'opportunità di considerare i problemi che scaturiscono da nuovi modelli biologici è ancora da verificare rispetto la letteratura scientifica più recente e il confronto con esperti del settore (in particolare con specifiche competenze di carattere biologico).

Allineamento di Sequenze Espresse

Le sequenze espresse (EST) sono frammenti di sequenze nucleotidiche che contengono solamente le parti del patrimonio genetico di un individuo che codificano una proteina. L'allineamento di sequenze espresse rispetto a sequenze genomiche consente di inferire computazionalmente alcune informazioni riguardanti la struttura del patrimonio genetico degli individui, informazioni che vengono utilizzate negli studi di associazione per determinare correlazioni fra differenze genetiche tra gli individui e la presenza di caratteristiche osservabili quali malattie genetiche e resistenze ai farmaci. Il problema di allineamento di sequenze espresse rispetto a sequenze genomiche, tuttavia, ha importanti peculiarità che lo differenziano dai problemi classici di allineamento di sequenze.

Durante quest'anno ho approfondito lo studio di questo problema di allineamento, che

ho modellato come un problema computazionale di *fattorizzazione* di una sequenza rispetto a un'altra. Formalmente il problema di fattorizzazione è stato così definito.

(l, k)-fattorizzazione
Istanza: Due sequenze P e T e due interi l e k.
Soluzione: Una suddivisione della sequenza P in al più k fattori contigui e lunghi, ciascuno, almeno l caratteri tale che la sequenza dei fattori occorra in ordine all'interno della sequenza T.

Accanto alle applicazioni biologiche che ne hanno motivato la formulazione, questo problema è di interesse più generale in quanto può essere considerato come una generalizzazione del problema algoritmico classico di estrarre la più lunga sottosequenza comune a due sequenze (LCS).

Si è quindi disegnato un algoritmo efficiente per la risoluzione del problema di (l, k)-fattorizzazione che fa uso degli "alberi di suffisso generalizzati", una struttura dati per l'indicizzazione dei fattori di una sequenza.

Questo algoritmo è stato in seguito esteso in modo da individuare e rappresentare in modo compatto tutte le (l, k)-fattorizzazioni massimali di una sequenza rispetto a un'altra (ovvero, informalmente, le (l, k)-fattorizzazioni che non sono "contenute" in altre (l, k)-fattorizzazioni). Infatti la soluzione al problema della (l, k)-fattorizzazione non è, generalmente, unica e, nell'applicazione biologica, non è a priori sempre chiaro quale soluzione sia da preferire.

L'elevata ridondanza delle banche dati di sequenze espresse mi ha permesso di definire un problema di ottimizzazione che, ispirato dal principio del rasoio di Occam, potesse determinare la fattorizzazione plausibilmente più significativa (nel seguito verrà usato il termine "fattorizzazione" in luogo di "(l, k)-fattorizzazione" per semplicità espositiva). In particolare il problema di *riconciliazione* delle fattorizzazioni di un insieme S di sequenze rispetto a una sequenza fissata T consiste nello scegliere, per ogni sequenza s appartenente a S, una fattorizzazione di s rispetto a T tale che l'insieme dei fattori di T indotti da tutte le fattorizzazioni delle sequenze in S che sono state scelte sia di cardinalità minima.

Durante l'attività di ricerca ho studiato la complessità computazionale del problema di riconciliazione delle fattorizzazioni e ho dimostrato che esso è NP-hard mediante riduzione da un problema classico di ottimizzazione, Set Cover. A seguito di questo risultato di intrattabilità del problema su istanze generali, mi sono concentrato nel disegno di algoritmi in grado di risolvere il problema in modo efficiente su istanze che soddisfano alcune assunzioni ragionevoli e che si incontrano frequentemente in casi reali.

La fase di implementazione di questi algoritmi è tuttora in corso. All'implementazione seguirà una fase di sperimentazione volta a confrontare l'accuratezza e il consumo di risorse computazionali degli algoritmi disegnati rispetto a un metodo di allineamento da noi precedentemente sviluppato e rispetto ai metodi allo stato dell'arte presenti in letteratura.

Candidato: Yuri Pirola **Relatore:** Prof. Paola Bonizzoni **Tutor:** Prof. Lucia Pomello