



Università degli Studi di Milano–Bicocca
Dipartimento di Informatica, Sistemistica
e Comunicazione



Proposta tesi di Dottorato

Yuri Pirola

Dottorato di Ricerca in Informatica - XXII Ciclo

Titolo proposto: “Problemi combinatori nello studio di variazioni genetiche”

Relatore: Prof. Paola Bonizzoni

Tutor: Prof. Lucia Pomello

Obiettivi

Il lavoro di tesi proposto consiste nello studio della complessità computazionale e, conseguentemente, nel disegno di algoritmi risolutivi per problemi combinatori derivanti dall'analisi di variazioni genetiche. In particolare, l'interesse della presente proposta è focalizzato allo studio di metodi combinatori per l'inferenza di aplotipi rivolti alla individuazione di polimorfismi (cioè singole variazioni nucleotidiche).

Un'analisi preliminare di questi problemi combinatori ha evidenziato che la loro soluzione algoritmica è strettamente collegata alla soluzione di alcuni problemi su grafi di ampio interesse applicativo.

Per questo motivo, il lavoro di tesi avrebbe diverse ricadute. Innanzitutto è volto a disegnare algoritmi risolutivi per problemi di rilevante interesse nella ricerca biologica dell'era post-genomica. Inoltre, si vogliono evidenziare le connessioni tra questi problemi di interesse biologico e problemi di interesse interdisciplinare al fine di disegnare nuovi algoritmi che consentano di risolvere problemi aperti o che migliorino la complessità computazionale di quelli esistenti. Sulla base dei risultati precedenti, si vogliono infine estendere i modelli di inferenza in modo che le assunzioni che ne stanno alla base si avvicinino ulteriormente alla realtà biologica.

Il resto del documento è così strutturato. Inizialmente (sez. 1) verrà introdotta e definita la terminologia specifica necessaria per la comprensione delle restanti sezioni nonché verranno illustrate le motivazioni di carattere biologico del problema. La sezione 2 cercherà di presentare lo stato dell'arte riguardante i problemi combinatori derivanti dai modelli per l'inferenza di aplotipi. In sez. 3 si illustreranno alcuni dei principali problemi ancora aperti che si vogliono affrontare durante il lavoro di tesi. Inoltre, verranno brevemente descritte le strategie che si vogliono perseguire e i primi risultati che sono stati ottenuti in merito.

1 Inquadramento biologico del problema e terminologia

Da un punto di vista biologico, gli studi della varietà genetica degli individui di una popolazione rientrano in un recente ambito di ricerca, nato nell'era post-genomica, che ha consentito e che tuttora consente di determinare nel patrimonio genetico di un individuo nuovi fattori di rischio per l'insorgenza di malattie e resistenza a farmaci oppure di ricostruire la storia evolutiva di popolazioni. Questi studi di carattere biologico necessitano di dati genetici provenienti da specifiche posizioni del DNA dell'organismo chiamate Single Nucleotide Polymorphism (SNP), ovvero delle posizioni in cui si riscontrano differenze genetiche tra gli individui di una popolazione dovute a mutazioni di una singola base. La sequenza delle basi (chiamate, in questo caso, *alleli*) estratte nelle posizioni, o siti, di SNP costituisce l'*aplotipo* di un individuo. In generale, inoltre, i polimorfismi si manifestano in una popolazione in solo due stati, chiamati varianti bialleliche.

La natura *diploide*¹ degli organismi più evoluti, quale l'uomo, comporta la presenza in ciascun individuo di una coppia di aplotipi, ognuno ereditato, eventualmente mutato, da uno dei due genitori. Questioni di carattere tecnologico limitano la possibilità di ottenere su larga scala gli aplotipi degli individui di una popolazione. Risulta meno dispendioso, invece, ottenere il *genotipo* degli individui, cioè la sequenza delle coppie non ordinate formate dagli alleli dei due aplotipi che compongono (o *risolvono*) il genotipo. Si sono rese quindi necessarie delle tecniche computazionali che consentano di ricostruire, o *inferire*, gli aplotipi di una popolazione a partire dai dati genotipici degli individui. Il problema dell'inferenza di aplotipi in popolazioni consiste, quindi, nella risoluzione dell'ambiguità derivata dalla presenza nei genotipi di siti eterozigoti, ovvero di coppie formate da alleli differenti tra loro. Per questi siti, infatti, non è immediatamente desumibile a quale aplotipo appartiene ciascuno dei due alleli. Le tecniche computazionali necessitano, chiaramente, di un modello biologico di riferimento in modo da risolvere questa ambiguità, cioè di ulteriori assunzioni che consentano di determinare un insieme biologicamente plausibile di aplotipi che generano i genotipi dati. In assenza di ulteriori assunzioni, infatti, la soluzione del problema di inferenza consisterebbe nella mera enumerazione delle combinazioni di aplotipi che risolvono i genotipi in ingresso. Ovviamente tale soluzione avrebbe una quasi nulla rilevanza biologica.

I modelli per l'inferenza di aplotipi proposti in letteratura sono principalmente due: il modello *coalescente* e il modello di parsimonia. Nel modello *coalescente* si suppone che gli aplotipi si debbano disporre secondo una struttura ad albero, chiamata filogenesi perfetta, che indica l'evoluzione dei caratteri e la comparsa delle mutazioni. Il modello di *parsimonia*, invece, rispecchia il principio del rasoio di Occam, secondo il quale la soluzione più verosimile risulta essere quella più semplice. In questo caso, il rasoio di Occam si realizza considerando come soluzione del problema di inferenza il minimo insieme di aplotipi differenti che risolvono i genotipi della popolazione.

¹ Un organismo si dice diploide se il suo patrimonio genetico è composto da due set di cromosomi omologhi, cioè che controllano gli stessi caratteri, e che vengono ereditati, ciascuno, da un genitore.

2 Modelli per l'inferenza e problemi combinatori: stato dell'arte

L'ambito di ricerca riguardante i metodi di inferenza di aplotipi, nonostante sia relativamente recente, ha riscosso molto interesse e ha indagato il problema secondo diversi approcci; fra questi i tre principali sono quello statistico, quello combinatorio e quello basato sulla teoria dell'informazione. Il lavoro di tesi che si vuole sviluppare si concentrerà esclusivamente sulle tecniche combinatorie per l'aplotipizzazione e, per questo motivo, nel resto della sezione verranno presentati i modelli principali per l'inferenza di aplotipi e i problemi combinatori che da essi ne derivano.

Dal punto di vista combinatorio, il problema di inferenza di aplotipi consiste nel determinare un insieme H di aplotipi che risolve l'insieme G dei genotipi dato in input, cioè nel determinare un insieme H tale che, per ogni genotipo $g \in G$, esistono due aplotipi $h^1, h^2 \in H$ che compongono (o risolvono) g .

Modello di parsimonia I primi algoritmi combinatori per la ricostruzione di aplotipi si basavano sulla costruzione incrementale della soluzione H mediante l'applicazione iterata di una regola di inferenza, detta regola di Clark [3]. Si noti che non tutte le sequenze di applicazione della regola di inferenza consentono di raggiungere una soluzione del problema. A seconda dell'ordine scelto, infatti, è possibile che l'algoritmo termini con successo, cioè con tutti i genotipi risolti, oppure con insuccesso, cioè con alcuni genotipi non risolti. Il problema di ottimizzazione che è stato formulato sulla base della regola di Clark richiede, quindi, di trovare il più grande sottoinsieme G' dell'insieme G dei genotipi in input che è possibile risolvere mediante l'applicazione della regola di inferenza. Gusfield [6] ha poi dimostrato che questo problema e alcune sue restrizioni sono NP-hard. Di questo problema è disponibile una formulazione di programmazione lineare intera che ne consente la risoluzione tramite rilassamento.

Un'altra formulazione del problema di aplotipizzazione si basa sul principio di *pura parsimonia* (proposto da Gusfield [8]). In questo caso si ricerca l'insieme di aplotipi di minima cardinalità che risolve l'insieme dei genotipi dati. Nonostante si sia dimostrato che questo problema è APX-hard (Lancia *et al.* [10]), in letteratura sono stati proposti numerosi algoritmi risolutivi, sia esatti (tramite branch-and-bound, [12]), che approssimati (con fattori non costanti, [10] e [9]), che euristici ([13]).

Modello coalescente Il secondo modello di ricostruzione di aplotipi, il modello coalescente (proposto in [7]), si basa su assunzioni biologiche relative alla storia evolutiva delle mutazioni. In particolare, secondo il modello coalescente, le mutazioni (cioè le transizioni da un allele ad un altro in alcuni polimorfismi) avvengono una sola volta e, per questo, sono condivise da un insieme di individui con il medesimo antenato. In questo caso gli aplotipi degli individui di una popolazione si dispongono secondo un albero etichettato, chiamato filogenesi perfetta, in cui le foglie sono etichettate dagli aplotipi mentre le mutazioni ne etichettano gli archi. In una filogenesi perfetta, inoltre, l'insieme delle etichette poste su ciascun percorso dalla radice a una foglia coincide con le mutazioni presenti nell'aplotipo

che etichetta la foglia stessa. Il problema dell'inferenza di aplotipi secondo il modello coalescente (*Perfect Phylogeny Haplotyping*, *PPH*) richiede che, dato un insieme di genotipi G , venga determinato, se esiste, un insieme di aplotipi H che ammette filogenesi perfetta. Il primo algoritmo con complessità computazionale quasi lineare è dovuto a Gusfield [7] che lo riduce a un problema classico su grafi, chiamato *Graph Realization (GR)*. La riduzione mostrata richiede tempo lineare nella dimensione dell'input ($O(nm)$, dove n è il numero di genotipi in input e m il numero di SNP considerati) mentre i due migliori algoritmi risolutivi per GR (Fujishige [5] e Bixby e Wagner [2]) richiedono tempo $O(nm\alpha(nm))$, con α funzione inversa di Ackerman. L'implementazione dell'algoritmo di Gusfield non ha però impiegato gli algoritmi quasi-lineari per la loro elevata complessità di funzionamento e di realizzazione ma si è basata su un altro algoritmo per GR di complessità $O(nm^2)$. In seguito, Gusfield [4] ha disegnato un algoritmo diretto e lineare per la soluzione del problema del PPH, seguito poi da Vijaya Satya [11].

Xor-genotipi Le tecniche sperimentali di genotipizzazione di una popolazione riescono a discriminare efficacemente i siti eterozigoti dai siti omozigoti ma, per questi ultimi, risulta essere complesso determinare quale dei due alleli presentano. Questa considerazione ha portato a definire gli *Xor-genotipi*, cioè vettori binari in cui ciascuna componente, che corrisponde a uno SNP o sito genomico, può essere solo nello stato omozigote o eterozigote. I modelli visti precedentemente possono essere impiegati anche su questa nuova rappresentazione dei genotipi e danno luogo alle relative varianti dei problemi formulati su genotipi tradizionali. In particolare, in [1], è stato introdotto il problema di inferenza di aplotipi mediante filogenesi perfetta su Xor-genotipi, XPPH. Nello stesso lavoro è stato dimostrato che il problema XPPH è equivalente al problema della Graph Realization e, di conseguenza, ha complessità quasi-lineare.

Centri di ricerca In ambito internazionale si possono principalmente distinguere tre distinti centri di eccellenza nella ricerca su problemi di inferenza di aplotipi: il gruppo di ricerca di Dan Gusfield, università della California a Davis, il gruppo di Ron Shamir, università di Tel Aviv, e il gruppo di Richard Karp, università della California a Berkeley.

3 Linee di ricerca per il lavoro di tesi

La mia attività di ricerca nell'ambito della tesi riguarderà inizialmente lo studio di problematiche aperte relative all'inferenza di aplotipi la cui soluzione richiede l'approfondimento di specifici problemi su grafi. Più precisamente sarà focalizzata sullo studio del problema dell'inferenza di aplotipi a partire da Xor-genotipi in presenza dei modelli coalescente e di pura parsimonia. La letteratura e alcuni studi preliminari hanno evidenziato importanti connessioni fra questi problemi e il problema di Graph Realization. Pertanto la realizzazione di un algoritmo efficiente per la Graph Realization costituisce un obiettivo primario del lavoro di tesi. In un secondo momento l'attenzione sarà rivolta allo studio di estensioni del modello coalescente che introducono nuove tipologie di mutazione, come la possibilità

di ricombinazioni e duplicazioni di parti degli aplotipi. In entrambe le direzioni di ricerca sono stati già ottenuti alcuni risultati parziali che verranno di seguito presentati.

Complessità computazionale di Xor Parsimony Haplotyping Il problema di inferenza di aplotipi secondo il modello di pura parsimonia su Xor-genotipi (XPH) è un problema ancora aperto. Esso consiste nel determinare l'insieme degli aplotipi di cardinalità minima che risolve l'insieme degli Xor-genotipi dato in input.

La complessità computazionale di questo problema non è ancora conosciuta ma si congettura che esso sia intrattabile nel caso generale. Pertanto è di interesse lo studio di alcune delle sue restrizioni, denotate con $XPH[k,l]$ dove k è il numero massimo di siti eterozigoti in un genotipo mentre l è il numero massimo di genotipi in cui un sito eterozigote può comparire. Studiare la complessità computazionale del problema a partire da sue restrizioni è una strategia già utilizzata in letteratura per questa tipologia di problemi perché permette, in primo luogo, di determinare quali restrizioni sono risolvibili in tempo polinomiale se il caso generale si dimostra essere non trattabile e, inoltre, consente di individuare su istanze semplici le proprietà utili alla soluzione del problema, proprietà che si possono poi cercare di estendere anche alle istanze più generali.

Lo studio di due istanze ristrette, $XPH[* ,2]$ e $XPH[2,*]$, ha portato al disegno di un algoritmo polinomiale per la loro risoluzione. Attualmente si sta procedendo alla formalizzazione della dimostrazione di correttezza degli algoritmi e all'estensione del lavoro su istanze quali $XPH[* ,3]$ e $XPH[3,*]$.

Graph Realization e inferenza di aplotipi mediante filogenesi perfetta La letteratura riguardante i problemi di inferenza tramite filogenesi perfetta ha inizialmente risolto questi problemi mediante riduzione al problema di Graph Realization, di cui si conoscono algoritmi risolutivi quasi-lineari. L'attività del primo anno di dottorato è stata in parte orientata al disegno di un algoritmo lineare diretto per la risoluzione del problema di aplotipizzazione su genotipi tradizionali. Durante il lavoro di tesi si vuole continuare questa linea di ricerca determinando un algoritmo risolutivo lineare per il problema di aplotipizzazione basato su Xor-genotipi. Un tale algoritmo avrebbe importanti ricadute anche in altre discipline perché potrebbe essere esteso al problema della Graph Realization, per il quale, nonostante la sua rilevanza applicativa, non si hanno miglioramenti significativi da molti anni.

Estensioni del modello coalescente Durante l'attività di tesi si vuole dare particolare rilievo anche allo studio dei problemi computazionali derivanti da estensioni dei modelli biologici fin qui presentati. In particolare, il modello coalescente si basa su due assunzioni principali: (1) assenza di ricombinazioni e (2) irripetibilità delle mutazioni durante la storia evolutiva. Seppur siano assunzioni di carattere abbastanza generale, esse, nella pratica, limitano la lunghezza degli aplotipi che possono essere ricostruiti con il modello coalescente. Per questo motivo in letteratura sono state introdotte delle estensioni del modello coalescente che rilassano queste assunzioni. Un'importante estensione che è stata definita è il *modello di filogenesi con caratteri persistenti*, che si vuole analizzare durante

il lavoro di tesi. La strategia che si vuole adottare anche in questo caso, per gli stessi motivi illustrati in precedenza, è quella di iniziare a studiare alcune restrizioni relativamente semplici del problema e, in seguito, generalizzare i risultati ottenuti a istanze più complesse. In questa direzione ho partecipato alla supervisione di uno stage che ha portato a disegnare un algoritmo polinomiale per la ricostruzione di filogenesi con caratteri persistenti in cui ciascun aplotipo non possiede più di due siti mutati.

Pianificazione temporale L'esposizione delle linee di ricerca fatta in questa sezione rispecchia la pianificazione e le priorità delle tematiche che si vogliono affrontare durante il lavoro di tesi. A livello indicativo, ci si attende di concludere il lavoro sul problema di aplotipizzazione con pura parsimonia su Xor-genotipi entro tre o quattro mesi dal suo inizio. Il disegno di un algoritmo per l'aplotipizzazione con filogenesi perfetta su Xor-genotipi, e quindi la possibile estensione a Graph Realization, è sicuramente più impegnativa e potrebbe richiedere 5 o 6 mesi, concludendosi così all'inizio dell'estate ventura. Si prevede, invece, di impiegare i mesi successivi per approfondire le basi relative all'estensione dei modelli biologici e preparare così il lavoro per la parte finale del corso di dottorato.

Riferimenti bibliografici

- [1] T. Barzuza, J. S. Beckmann, R. Shamir, and I. Pe'er. Computational problems in perfect phylogeny haplotyping: Xor-genotypes and tag SNPs. In *Proc. 13th Symposium on Combinatorial Pattern Matching (CPM)*, pages 14–31, 2004.
- [2] R. E. Bixby and D. K. Wagner. An almost linear-time algorithm for graph realization. *Mathematics of Operations Research*, 13:99–123, 1988.
- [3] A. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, 7(2):111–122, 1990.
- [4] Z. Ding, V. Filkov, and D. Gusfield. A linear-time algorithm for the perfect phylogeny haplotyping (PPH) problem. In *Proc. 9th Annual Conference on Research in Computational Molecular Biology (RECOMB)*, pages 585–600, 2005.
- [5] S. Fujishige. An efficient PQ-graph algorithm for solving the graph realization problem. *Journal of Computer and System Science*, 21:63–68, 1980.
- [6] D. Gusfield. Inference of haplotypes from samples of diploid populations: Complexity and algorithms. *Journal of Computational Biology*, 8(3):305–323, 2001.
- [7] D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Proc. 6th Annual Conference on Research in Computational Molecular Biology (RECOMB)*, pages 166–175, 2002.
- [8] D. Gusfield. Haplotyping by pure parsimony. In *Proc. 14th Symposium on Combinatorial Pattern Matching (CPM)*, pages 144–155, 2003.
- [9] Y.-T. Huang, K.-M. Chao, and T. Chen. An approximation algorithm for haplotype inference by maximum parsimony. *Journal of Computational Biology*, 12(10):1261–1274, 2005.
- [10] G. Lancia, M. C. Pinotti, and R. Rizzi. Haplotyping populations by pure parsimony: Complexity of exact and approximation algorithms. *INFORMS Journal on Computing*, 16(4):348–359, 2004.
- [11] R. V. Satya and A. Mukherjee. An efficient algorithm for perfect phylogeny haplotyping. In *Proc. 4th International IEEE Computer Society Computational Systems Bioinformatics Conference (CSB 2005)*, pages 103–110, 2005.
- [12] L. Wang and Y. Xu. Haplotype inference by maximum parsimony. *Bioinformatics*, 19(14):1773–1780, 2003.
- [13] L. Wang and Y. Xu. A parsimonious tree-grow method for haplotype inference. *Bioinformatics*, 21(17):3475–3481, 2005.