Dottorato di Ricerca in Informatica - Ciclo XXII
Dipartimento di Informatica, Sistemistica e Comunicazione
Facoltà di Scienze Matematiche, Fisiche e Naturali
Università degli Studi di Milano–Bicocca

Problemi combinatori nello studio di variazioni genetiche

Presentazione proposta di tesi

29 gennaio 2008

Candidato: Yuri Pirola

Supervisore: Prof. Paola Bonizzoni

Tutor: Prof. Lucia Pomello

Outline

Obiettivi e Motivazioni

Obiettivo

Analisi della complessità computazionale e disegno di algoritmi per lo studio delle variazioni genetiche in vaste popolazioni.

Contributo:

- Modellazione di problemi reali mediante strumenti combinatori
- Formulazione e risoluzione di problemi combinatori di interesse generale
- Realizzazione e sperimentazione degli algoritmi disegnati

Studio delle variazioni genetiche: Problematiche

Problema di Inferenza di Aplotipi

Determinare per ciascun individuo di una popolazione il patrimonio genetico ereditato da ciascuno dei genitori coerentemente con un modello genetico di riferimento.

Studi di Associazione

Individuare associazioni fra variazioni genetiche fra gli individui di una popolazione e caratteristiche osservabili (ad es. malattie e resistenze ai farmaci).

Inferenza di Aplotipi

Problema di Inferenza di Aplotipi

Determinare per ciascun individuo di una popolazione il patrimonio genetico ereditato da ciascuno dei genitori *coerentemente con un modello genetico di riferimento*.

Il modello genetico caratterizza il problema!

In letteratura:

- Modello di Pura Parsimonia
- Modello Coalescente
- Modelli Evolutivi Generalizzati
- Genotipizzazioni XOR

Pure Parsimony Haplotyping (Gusfield, CPM, 2003)

Modello genetico: Modello di Pura Parsimonia

Problema combinatorio: Minimizzazione degli aplotipi differenti nella popolazione

Stato dell'arte:

APX-hard (Lancia et al., INFORMS J. on Comp., 2004)
Algoritmi esatti (branch-and-bound, (Wang et al., Bioinf., 2003))
Algoritmi approssimati (Lancia et al., INFORMS J. on Comp., 2004)
Algoritmi euristici (Wang et al., Bioinf., 2005)

Perfect Phylogeny Haplotyping (Gusfield, RECOMB, 2002)

Modello genetico: Modello coalescente

Problema combinatorio: Ricostruzione dell'*albero filogenetico* dell'evoluzione dei caratteri

Stato dell'arte:
Modello restrittivo ma algoritmo esatto lineare (Ding et al., RECOMB, 2005)

Adatto come parte di modelli più generali (Karp et al., ICALP, 2004)

Importante connessione con problema classico su grafi (l'albero è una realizzazione particolare di un insieme di percorsi, *Graph Realization Problem*)

Phylogenetic Network Haplotyping (Gusfield, J. of Comp. and Syst. Sci., 2005)

Modello genetico: Modello evolutivo generalizzato (ad es. ricombinazioni, ...)

Problema combinatorio: Ricostruzione del *grafo diretto aciclico* dell'evoluzione dei caratteri

Stato dell'arte:
NP-hard nel caso generale (Wang et al., SAC, 2001)
Trattabile con alcune limitazioni (Song et al., WABI, 2006)

Questioni aperte:
Formulazione di restrizioni (=problemi combinatori) per i nuovi modelli biologici (caratteri persistenti)

XOR Genotypes (Shamir et al., IEEE/ACM Trans. Comp. Biol. Bioinf., to appear)

Modello genetico:
Distinzione fra siti eterozigoti e siti omozigoti

Problema combinatorio: Filogenesi Perfetta e Pura Parsimonia

Stato dell'arte:
Quasi-lineare su Filogenesi Perfetta (equivalente a *Graph Realization*)

Risultati preliminari e questioni aperte:
NP-hard su massimo sottoinsieme che ammette Filogenesi Perfetta (con riduzione da Max-3-SAT). Approssimabile? FPT?

?? su Pura Parsimonia

Studi di Associazione

Studi di Associazione: (Zelikovsky et al., WABI, 2006)

- *Input:* Aplotipi e una partizione degli individui della popolazione (ad es. sani e malati)
- *Output:* Stabilire il minimo insieme di caratteri che permette di distinguere gli elementi della partizione

Metodi combinatori efficienti per lo studio su larga-scala.

Algoritmo greedy con buoni risultati sperimentali (Zelikovsky et al., BIBE, 2007).

Necessita di maggiore formalizzazione e studio.

Linee di ricerca

- Formulazione di modelli combinatori per problemi reali
  - Inferenza di aplotipi: caratteri persistenti e genotipi XOR
  - Studi di associazione
- Studio della complessità computazionale e utilizzo di tecniche algoritmiche innovative per la loro soluzione
  - Algoritmi di approssimazione
  - Algoritmi parametrizzati
- Analisi teorica e sperimentale del comportamento degli algoritmi introdotti

Panorama internazionale

Conferenze:

- SODA, ACM-SIAM Symp. on Discrete Algorithms
- ICALP, Int. Colloquium on Automata, Languages, and Programming
- ESA, European Symposium on Algorithms
- WABI, Work. on Algorithms in Bioinformatics
- RECOMB, Int. conf. on Research in Computationl Molecular Biology
- CPM, Ann. Symp. on Combinatorial Pattern Matching

Riviste:

- ACM Transactions on Algorithms
- Algorithmica
- ACM Journal on Experimental Algorithms
- Journal of Algorithms
- SIAM Journal on Computing
- Theoretical Computer Science
- Discrete Applied Mathematics
- ACM/IEEE Transactions on Bioinformatics and Computational Biology
- Journal on Computational Biology

Riferimenti essenziali (1)

Pure Parsimony Haplotyping

Gusfield. Haplotyping by pure parsimony. In *Proc. CPM'03*, 144–155, 2003.

Lancia, Pinotti, and Rizzi. Haplotyping populations by pure parsimony: Complexity of exact and approximation algorithms. *INFORMS J. on Computing*, 16(4):348–359, 2004.

Wang and Xu. Haplotype inference by maximum parsimony. *Bioinformatics*, 19 (14):1773–1780, 2003.

Wang and Xu. A parsimonious tree-grow method for haplotype inference. *Bioinformatics*, 21(17):3475–3481, 2005.

Perfect Phylogeny Haplotyping

Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Proc. RECOMB'02*, 166–175, 2002.

Ding, Filkov, and Gusfield. A linear-time algorithm for the perfect phylogeny haplotyping (PPH) problem. In *Proc. RECOMB'05*, 585–600, 2005.

Halperin and Karp. The minimum-entropy set cover problem. In *Proc. ICALP '04*, 733-744, 2004

Riferimenti essenziali (2)

Phylogenetic Network Haplotyping

Gusfield. Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination. *J. of Computer and System Sciences*, 70(3):381–398, 2005.

Wang, Zhang, and Zhang. Perfect phylogenetic networks with recombination. In *Proc. SAC'01*, 46–50, 2001.

Song, Liu, Malmberg, and Cai. Phylogenetic network inferences through efficient haplotyping. In *Proc. WABI'06*, 68–79, 2006.

Genotipizzazione XOR

Shamir, Barzuza, et al.. Computational Problems in Perfect Phylogeny Haplotyping: Typing without calling the allele. *IEEE/ACM Transaction on Computational Biology and Bionformatics*, to appear.

Studi di associazione

Zelikovsky and Brinza. Combinatorial methods for disease association search and susceptibility prediction. In *Proc. WABI'06*, 286–297, 2006.

Zelikovsky and Brinza. Discrete methods for association search and status prediction in genotype case-control studies. In *Proc. BIBE'07*, 2007.