Università degli Studi di Milano
Dipartimento Scienze dell'Informazione

Draft riassuntivo della tesi

28 Giugno 2001

"On the power of classes of splicing systems"

Dottoranda (XIII ciclo): Rosalba Zizza
Relatori: Prof. Giancarlo Mauri
Prof.ssa Clelia De Felice (Univ. di Salerno)

1 Introduzione

La teoria degli *automi finiti* ha sin dagli inizi abbracciato una grande varietà di aspetti, mostrando legami con campi diversi che vanno dalla teoria algebrica dei semigruppi alla progettazione di algoritmi per la manipolazione di stringhe. E' stata proprio questa diversità che l'ha arricchita e gli aspetti teorici e le potenzialità applicative ne determinano l'ancor vivo interesse. I risultati di questa tesi evidenziano entrambi questi punti di vista, ed utilizzano tecniche formali proprie della teoria, così come della combinatoria delle parole e della teoria dei monoidi finiti.

La prima referenza bibliografica per la teoria degli automi, sia per cronologia che per importanza, è l'articolo di Kleene (1956) in cui si mostra l'equivalenza tra le due nozioni di linguaggio-evento regolare (riconosciuto da un automa finito) e linguaggio razionale (appartenente alla chiusura della famiglia dei linguaggi finiti rispetto alle operazioni razionali). Da allora numerose sono state le investigazioni intraprese, alcune delle quali tese allo studio delle proprietà dei linguaggi regolari (definizione di operazioni tra linguaggi e proprietà di chiusura rispetto ad esse) e altre relative allo studio della interpretazione di tali proprietà sugli automi.

Molto recentemente, tutta la teoria dei linguaggi formali ha ricevuto nuova linfa vitale in seguito all'introduzione di nuovi modelli di computazione. In particolare, è stata recentemente proposta la formalizzazione di un modello matematico per fenomeni biologici di ricombinazione molecolare. Infatti l'analogia tra processi biologici e matematici è evidente, in quanto in entrambi, le strutture biologiche complesse o l'applicazione di funzioni calcolabili ad un argomento sono il risultato della combinazione di operazioni semplici a oggetti-entità di base. Questa analogia è la chiave di interpretazione del *Calcolo Molecolare*, area di collocazione della mia tesi.

Il Calcolo Molecolare viene inteso sia come possibilità di "portare algoritmi in provetta" (simulazione efficiente di problemi NP-completi attraverso algoritmi con operazioni biologiche - esperimento di Adleman et al.), sia come possibilità di trarre ispirazione da operazioni su DNA, proponendo operazioni su stringhe al fine di definire sistemi non convenzionali di calcolo. In tale ambito si inserisce la teoria degli *splicing systems*, oggetto della tesi. Iniziata da Head (1987), la teoria degli splicing systems si è sviluppata successivamente soprattutto con l'obiettivo di proporre sistemi alternativi a quelli classici (automi e grammatiche). Ogni nuova variante di splicing system presentato in letteratura è accompagnato dalla prova della sua universalità (equivalenza con Macchine di Turing). La finalità quindi non è porre in discussione la Tesi di Church-Turing, ma progettare un sistema di calcolo che, in maniera più efficiente rispetto a quelli convenzionali, esegua computazioni difficili.

Lo splicing è una operazione di cut and paste biologico operato su molecole di DNA da enzimi di restrizione (proteine riconoscitori di pattern) e da enzimi di ligasi (proteine che incollano). Un sistema di accoppiamento (splicing system) è un sistema formale che alle sequenze di DNA fa corrispondere un insieme I (iniziale) di parole su un alfabeto finito e agli enzimi fa corrispondere delle parole speciali (dette regole R) in cui è contenuta la specifica sia dei due patterns da riconoscere, sia del punto del taglio sia della ricomposizione dei pezzi tagliati. Dunque, a partire dall'insieme iniziale, per l'applicazione iterata delle regole, si genera un linguaggio, detto linguaggio splicing generato da I e R. In realtà, esistono diverse definizioni di splicing systems che differiscono tra di loro per la diversa maniera di utilizzare le regole e di definire il linguaggio splicing generato.

In maniera naturale, l'introduzione di un modello di calcolo è accompagnata dallo studio delle sue capacità computazionali, cioè dalla caratterizzazione della classe dei linguaggi che da esso vengono generati. Al variare di I, R nella gerarchia di Chomsky, si ottengono splicing sistems che generano una classe della gerarchia, o una classe intermedia tra due di esse o anche non confrontabile con esse. Dunque, in ques'ultimi casi, si pone il problema di caratterizzare la classe generata.

2 Risultati

2.1 Splicing lineare

Una classe di linguaggi che ha storicamente rivestito un ruolo centrale è la classe dei linguaggi regolari, cioè riconosciuti da automi finiti. E' noto che gli splicing systems (considereremo gli splicing systems cosiddetti *non iterati*) con un insieme iniziale I finito ed un insieme R di regole finito generano una sottoclasse propria di linguaggi regolari. Head aveva posto il problema se fosse decidibile stabilire l'appartenenza di un linguaggio regolare a tale sottoclasse e, in caso di risposta affermativa, dare una caratterizzazione di questa famiglia.

In questa tesi abbiamo fornito risposta (positiva) al problema posto da Head. Più precisamente abbiamo provato che dato un linguaggio L regolare, descritto attraverso l'automa minimale che lo riconosce, è decidibile se L è generabile da splicing finito, ossia se esiste un sistema finito che genera L. Inoltre, in quest'ultimo caso, viene descritto il sistema che genera L.

Il problema è quello classico di fornire una descrizione finita (nel nostro caso un insieme iniziale finito ed un insieme finito di regole) di un linguaggio infinito. Si intuisce allora che il compito delle regole sarà quello di allungare le parole del linguaggio iniziale e quelle via via ottenute, simulando così l'attraversamento illimitato dei cammini chiusi (cicli) dell'automa.

Riassumiamo brevemente le tecniche usate per la soluzione. Il primo passo fatto è stato quello di dimostrare l'esistenza, per ogni linguaggio razionale, di una particolare espressione razionale ad esso associata. Inoltre è noto che le parole di un linguaggio razionale sono identificabili con i cammini di successo di un automa. Grazie alla particolare espressione razionale sopra citata, abbiamo mostrato che è possibile estrarre dall'insieme dei cammini di successo un particolare insieme di cammini "canonici". Su questi cammini canonici abbiamo definito delle proprietà che devono essere soddisfatte da un linguaggio razionale affinché sia un linguaggio splicing.

Infine, abbiamo descritto come definire il sistema, cioè come definire due insiemi finiti I, R, a partire da questi cammini canonici, tali che il sistema genera L. Le regole consentiranno di ampliare i cicli, ma in maniera consistente: se una regola è applicata a due parole del linguaggio, le parole ottenute appartengono ancora al linguaggio. I vincoli sui cammini, garantiranno tale situazione.

Attualmente stiamo modificando un articolo in preparazione sull'argomento per rendere più chiara la presentazione dei risultati.

Oggetto di un'altra ricerca in corso è lo studio di un corrispondente problema di complessità descrizionale: se un linguaggio razionale è generabile mediante un sistema splicing finito, fornire una limitazione alla lunghezza delle regole di un sistema che lo genera.

Per questi risultati si è usata la definizione di Paun, una delle tre diverse definizioni formali di splicing, citate in letteratura (le altre due sono state proposte da Head e Pixton). Una naturale domanda è il confronto tra le capacità computazionali dei sistemi splicing definiti corrispondentemente, ossia se esistono linguaggi generabili da un sistema e non da un altro. Sempre restringendoci al caso finito, abbiamo confrontato i tre sistemi fornendo esempi di linguaggi che mostrano la differenza del potere computazionale delle tre definizioni di sistemi splicing. Questo risultato, in corso di stampa su una rivista internazionale, prova una inclusione stretta tra le classi dei linguaggi splicing generati dai sistemi di Head, Paun, Pixton (nell'ordine).

2.2 Splicing circolare

Un altro aspetto della problematica degli splicing systems fa intervenire la nozione di relazione di coniugazione. Si tratta di una relazione di equivalenza ben nota nella combinatoria delle parole, legata anche a problemi da fattorizzazione di linguaggi. Infatti, in letteratura è stato formalizzato un fenomeno di cut and paste biologico operato dagli enzimi su DNA circolare (come i plasmidi o, in alcuni casi, i batteri). Questo ha stimolato l'interesse verso lo studio delle cosiddette parole circolari. Esse si possono definire come le classi di equivalenza rispetto alla relazione di coniugazione. Infatti, due parole coniugate sono due parole che differiscono per una permutazione ciclica delle lettere, ossia esse sono identificabili se disponiamo le lettere delle parole su un cerchio. Per estensione si possono definire i linguaggi circolari come linguaggi i cui elementi sono parole circolari. Anche il concetto di regolarità può essere introdotto: un linguaggio circolare C è regolare se e solo se è regolare il linguaggio formale ottenuto prendendo tutte le parole nelle classi di equivalenza che costituiscono gli elementi di C (linearizzazione piena di C). Allora, studiare linguaggi circolari regolari equivale a studiare linguaggi formali regolari chiusi rispetto alla relazione di coniugazione, cioè linguaggi regolari L che contengono tutti gli elementi della classe di coniugazione di ogni elemento in L.

Come per il caso lineare, sono note tre definizioni di splicing circolare (Head, Paun, Pixton), sul cui potere computazionale la letteratura esistente è scarna. Questa è una delle motivazioni del nostro studio dei sistemi circolari splicing.

E' opportuno precisare immediatamente che le definizioni di splicing circolare date da Head e Paun modellano un meccanismo diverso rispetto a quello osservato da Pixton. Il meccanismo biologico alla base della definizione di Head e Paun consiste nel riconoscimento di due patterns in due parole circolari, nel taglio (apertura) delle stringhe all'interno dei patterns riconosciuti, nella concatenazione delle due stringhe lineari così ottenute e nella ricircolarizzazione della stringa risultante. Nella definizione data da Pixton invece, l'ultima operazione è preceduta da una sostituzione dei patterns individuati.

Sempre ponendoci nelle ipotesi di finitezza del linguaggio (circolare) iniziale e dell'insieme delle regole, abbiamo confrontato le tre definizioni. Attraverso un'ampia collezione di esempi di linguaggi generati e non, abbiamo provato un'inclusione tra le classi di linguaggi generati dai tre modelli descritti da Head, Paun e Pixton (nell'ordine). Inoltre abbiamo mostrato che i sistemi di Pixton sono computazionalmente più potenti rispetto agli altri due tipi di sistemi.

Non è nota una caratterizzazione della classe dei linguaggi circolari generati da sistemi splicing circolari finiti (insieme iniziale finito e insieme finito di regole). Abbiamo affrontato il problema di caratterizzare i linguaggi regolari in essa contenuti, limitandoci ad usare la definizione di Paun. Le ragioni di questa scelta sono da ricercarsi nel fatto che essa è più vicina alla classica definizione di concatenazione.

Abbiamo iniziato a studiare una classe di linguaggi regolari, da noi chiamata *star languages* (linguaggi chiusi sotto coniugazione e della forma X^*, con X linguaggio regolare) ed i linguaggi circolari corrispondenti. La motivazione è legata proprio al comportamento dello splicing (secondo Paun): se il linguaggio iniziale è un sottoinsieme del linguaggio circolare corrispondente a X^*, il sistema genera ancora un sottoinsieme di X^*.

Abbiamo dimostrato che la circolarizzazione di star languages che soddisfano una proprietà aggiuntiva (linguaggi fingerprint closed) sono generabili da splicing circolare finito. Un'importante sottoclasse degli star languages fingerprint closed è costituita dai linguaggi della forma X^*, con X codice a gruppo regolare. In questo contesto, la nozione di codice è quella di codice a lunghezza variabile, come introdotta da Schützenberger. Un'altra sottoclasse è costituita dai linguaggi chiusi rispetto alla relazione di coniugazione e della forma X^*, con X insieme finito.

L'esistenza della classe dei *linguaggi ciclici*, da noi definita, mostra che la classe dei linguaggi circolari generabili da splicing finito non coincide con la classe degli star languages fingerprint closed. La prova di questo risultato è basata su tecniche di combinatoria delle parole e utilizza un'idea del tutto diversa di ampliamento dei cicli, rispetto a quella usata per gli star languages, dettata dalla struttura stessa dei linguaggi ciclici.

Infine abbiamo studiato il caso dei linguaggi unari (su un alfabeto $\{a\}$ ad una lettera), linguaggi che sono evidentemente chiusi sotto coniugazione. Abbiamo dato una caratterizzazione dei linguaggi unari generabili con sistemi circolari finiti di Paun. Essi sono l'unione di un insieme finito di parole "corte" e di un linguaggio della forma X^+. Inoltre X deve essere un insieme finito e soddisfacente una ipotesi particolare che fa intervenire la nozione di sottogruppo di un gruppo ciclico finito. I risultati descritti in questa sezione sono stati oggetto di una comunicazione ad un convegno internazionale e sono contenuti in un lavoro sottomesso per pubblicazione in una rivista.

3 Conclusioni e sviluppi futuri

Lo studio dello splicing circolare ha validità sia nell'ambito dei modelli teorici per il Calcolo Molecolare (vista la scarsezza di risultati esistenti), sia in ambiti più enerali, applicativi e teorici. Infatti, da un punto di vista applicativo, molto recentemente i plasmidi sono stati utilizzati per simulare algoritmi in provetta. Essi hanno sostituito il DNA lineare, per consentire di sfruttare i vantaggi che hanno questi vettori di DNA.

Dunque, le problematiche per ricerche future ispirate da quest'area si differenziano per ambiti e grado di difficoltà. Tra i problemi più complessi relativi ai sistemi splicing resta aperto quello della descrizione del potere computazionale dei sistemi circolari splicing finiti, mentre quello della caratterizzazione dei linguaggi razionali chiusi per coniugazione appartiene più in generale al contesto della teoria dei linguaggi formali. Infine un punto di unione tra splicing circolare e splicing lineare potrebbe essere proprio lo studio delle due operazioni nel caso dell'alfabeto unario. Finora non esistono risultati e proposte in tal senso. Sarebbe realmente interessante stabilire connessioni tra due contesti che finora sono stati trattati solo distintamente. In dettaglio, se un linguaggio razionale unario non è generabile usando una definizione di splicing lineare, potrebbe esserlo usando una circolare, e viceversa.