Ricostruzione di reti di regolazione genica da dati trascrittomici

Progetto Bandiera InterOmics – WP1 CNR-ISTI

Claudia Caudai

Istituto di Scienza e Tecnologie dell'Informazione - CNR
Via G. Moruzzi, 1, I-56124 PISA, Italy
claudia.caudai@isti.cnr.it

Sommario Questo rapporto riassume brevemente alcuni approcci proposti recentemente in letteratura per risolvere il problema dell'identificazione dei meccanismi di regolazione genica dei processi cellulari in base a dati provenienti da esperimenti di sequenziamento genomico o trascrittomico. Poiché il livello di espressione di ogni gene dipende da uno o più fattori di trascrizione e dai livelli di attivazione dei relativi processi biologici, e poiché entrambi i fattori sono largamente incogniti, diversi tentativi di soluzione sono basati su assunzioni statistiche e tecniche di separazione cieca di segnali. Alcuni di questi, tutti derivanti da un modello generativo lineare, sono qui passati in rassegna.

1 Introduzione

Nessun gene in una cellula agisce individualmente: tutti sono collegati da reti di informazione per cui, se uno muta attività, questo si ripercuote sugli altri, che a loro volta modificheranno la loro attività. Da più di 10 anni vengono condotte ricerche tese a separare le sorgenti di regolazione genica in base a uscite ottenute da esperimenti biologici. L'argomento è tuttora oggetto di dibattito. Un certo numero di cose si sanno con ragionevole certezza:

- Molti geni sono attivi in qualunque condizione (in media, il 40%);
- Piccoli cambiamenti (temperatura, nutrimento, ecc.) causano un cambiamento di attività praticamente di tutti i geni;
- L'azione di un gene può essere potenziata o inibita dall'azione di altri geni. Se le variazioni di attività si mantengono sotto una certa soglia, tale effetto non si verifica.

Il problema della ricostruzione delle reti di regolazione da dati trascrittomici consiste nel riuscire a individuare l'azione dei singoli geni all'interno del processo di regolazione avendo a disposizione il risultato di alcuni esperimenti di sequenziamento condotti con le tecniche dei microarray, o con le nuove tecniche di Next Generation Sequencing (NGS), su cellule che si trovano in diverse condizioni, per esempio, per essere state sottoposte a trattamenti diversi o per il diverso

tempo intercorso dalla somministrazione di un determinato farmaco o comunque da una variazione significativa delle condizioni ambientali. Si assume che la vita della cellula sia sottesa da un certo numero di processi biologici, ognuno dei quali regola in qualche modo l'attività di un insieme di geni. Il livello di attività (ovvero, di trascrizione) di ogni gene sarà dunque influenzato da tutti i processi biologici attivi. Alcuni autori assumono che, sotto certe condizioni, possa essere dato da una combinazione lineare dei livelli di attività dei processi [1]. In generale, non è noto né quali siano questi processi né il valore dei coefficienti delle combinazioni lineari che producono i livelli di espressione genica. Una serie di esperimenti di sequenziamento produce una matrice in cui ogni colonna contiene i livelli di espressione rilevati per ogni gene in una data condizione sperimentale. Visto che né i processi biologici né i loro pesi sono noti a priori, una delle strade tentate per separare gli effetti dei diversi processi sull'espressione di ogni singolo gene è di fare delle assunzioni statistiche sulle relazioni reciproche tra le varie quantità in gioco per poi procedere con qualche approccio "cieco" di decomposizione, basato su qualcuna delle tecniche che vanno sotto il nome di BSS (Blind Source Separation). Un'assunzione, arbitraria ma comune, consiste nel supporre che i diversi processi siano scorrelati, o addirittura mutuamente indipendenti statisticamente. Questo tipo di assunzione porta direttamente ai due strumenti classicamente utilizzati per la soluzione dei problemi di BSS:

- La PCA (principal component analysis), che individua componenti scorrelate di massima varianza, non uniche, in contesti in cui la variabilità dei dati è attribuita a un ridotto set di componenti, e i dati sono affetti da errori di vario tipo ed entità;
- L'ICA (independent component analysis), che trova una decomposizione unica dei dati rumorosi in più pattern di espressione genica mutuamente indipendenti.

In questo rapporto, si riassumono le assunzioni di base fatte per la formulazione del problema in termini di BSS, si rende conto di alcune tecniche proposte in letteratura, e si riporta una bibliografia essenziale per l'introduzione all'argomento. Tutte le tecniche esaminate mimano una procedura ICA, e differiscono tra loro per l'interpretazione biologica data al modello matematico lineare e per l'introduzione nel problema di informazione a priori specifica. Sotto certe condizioni, l'informazione introdotta può essere sufficiente a rinunciare alle assunzioni non fondate biologicamente, consentendo di superare diverse delle obiezioni sollevate contro questo tipo di approccio.

2 Modello di espressione genica con il metodo ICA

Supponendo che l'attività di una cellula sia guidata da M processi biologici, ciascuno indipendente dall'altro, che ognuno di essi regoli in diversa misura ciascuno dei K geni, e che tali effetti di regolazione si compongano linearmente, si può costruire un modello matematico in base al quale ogni processo latente

possa essere separato dagli altri, e se ne possa nel contempo valutare il peso in ognuna delle diverse condizioni sperimentali considerate.

Partiamo definendo, per ogni processo, un vettore di regolazione genica:

$$\mathbf{s}_i = (s_i(1), s_i(2), ..., s_i(K)), \qquad i = 1, 2, ..., M \tag{1}$$

dove il valore $s_i(k)$ indica in che misura il processo biologico i influenza il livello di espressione del gene k. A loro volta, in ognuno di N diversi esperimenti, i vari processi biologici possono essere più o meno attivi nella cellula. Il livello totale di espressione del generico gene k-esimo al j-esimo esperimento, $x_j(k)$, sarà dato dalla somma dei contributi dovuti ai vari processi biologici, ciascuno pesato dal livello di attivazione del processo stesso all'esperimento j-esimo, a_{ji}, per $i = 1, ..., M$:

$$x_j(k) = a_{j1}s_1(k) + a_{j2}s_2(k) + ... + a_{jM}s_M(k). \qquad k = 1, 2, ..., K; j = 1, 2, ..., N \tag{2}$$

In forma matriciale, il modello (2) può essere espresso come

$$\mathbf{X} = \mathbf{AS}, \tag{3}$$

dove $\mathbf{A}$ riunisce i livelli di attivazione, ed è detta *matrice di mixing*, $\mathbf{X}$ è la matrice che rappresenta l'espressione dei vari geni nei vari esperimenti e $\mathbf{S}$ è la matrice che rappresenta le sorgenti di regolazione incognite. Quando la matrice $\mathbf{X}$ rappresenta il logaritmo dei rapporti fra i valori degli esperimenti e valori di riferimento, tale impostazione matriciale rappresenta un modello di interazione moltiplicativa tra processi biologici (vedi p.es. [1–3]). Tramite l'assunzione di mutua indipendenza, il modello (3) assicura l'esistenza di una matrice di separazione $\mathbf{W}$ e di una decomposizione $\hat{\mathbf{S}} = \mathbf{WX}$, uniche a meno di permutazioni e fattori di scala e stimabili per mezzo di una tra le varie tecniche ICA disponibili.

Una volta che le sorgenti di regolazione sono state stimate, occorre procedere alla valutazione del loro significato biologico; in altre parole, occorre individuare i processi biologici putativi che esse rappresentano. Il processo biologico corrispondente ad una data componente viene ricostruito sulla base della funzionalità dei geni in essa predominanti (sovra- o sotto-regolati). Ciascun gene è rappresentato mediante un codice identificativo (nome e simbolo del gene, specie, cromosoma e posizione nel cromosoma), e queste informazioni sono reperibili nelle numerose banche dati disponibili in rete (ad es. Gene[1]). I codici identificativi vengono annotati con attributi funzionali e strutturali, mediante opportuni vocabolari controllati, detti *Gene Ontologies*, che caratterizzano il coinvolgimento del gene nelle varie funzioni molecolari e nei vari processi biologici, e la sua localizzazione nelle varie componenti cellulari. Quindi, l'individuazione del processo biologico associato a una data componente indipendente avviene studiando il totale delle annotazioni dei geni in essa predominanti, e analizzando tali annotazioni per individuare le categorie strutturali e funzionali più rilevanti nelle condizioni in esame.

[1] http://www.ncbi.nlm.nih.gov/gene/

3 Modello di espressione genica con il metodo NCA

La Network Component Analysis (NCA) è un metodo introdotto in [4], che introduce una variante nell'approccio ICA, consentendo di utilizzare informazione a priori. La forza del metodo NCA rispetto all'ICA è la possibilità di introdurre conoscenza a priori nella matrice di mixing. Ciò permette di avere delle linee guida importanti nella risoluzione del problema. Il valore aggiunto (e anche la difficoltà) di tale approccio sta nel fatto che esso richiede la stretta collaborazione di ricercatori che appartengono a sfere disciplinari differenti, come biologi, bioinformatici, matematici, e statistici. La decomposizione fornita dall'ICA senza l'aggiunta di informazione specifica fornisce un modello fenomenologico dei dati osservati, ma non assicura che i segnali di regolazione stimati abbiano un significato biologico. La NCA consente almeno di sfruttare la parziale conoscenza disponibile sulle reti di regolazione genica per ricavare risultati biologicamente consistenti. Sebbene tutto somigli molto a una procedura ICA, la presenza di informazione a priori consente di non ricorrere ad assunzioni statistiche per giungere alla soluzione, a patto di soddisfare a un insieme di condizioni di identificabilità, qui di seguito elencate

1. La matrice di mixing deve avere rango pieno (assumendo ovviamente che il numero di geni sia maggiore del numero di esperimenti);
2. Se si elimina una colonna nella matrice di mixing e le righe corrispondenti ai valori non nulli di quella colonna (in pratica si trascura un esperimento e tutti i geni attivi in esso), la matrice che ne deriva deve essere ancora a rango pieno;
3. La matrice di regolazione deve avere rango pieno (assumendo che il numero di esperimenti sia maggiore del numero di sorgenti di regolazione).

L'implementazione pratica del metodo consiste nel risolvere il sistema (3) imponendo che $\mathbf{A}$ soddisfi a queste condizioni, e che rappresenti correttamente l'informazione biologica disponibile. Una prima stima di $\mathbf{A}$ si forma dunque ponendo a zero, per ogni gene, gli elementi corrispondenti a fattori di regolazione che si sanno non influire, e a valori arbitrari tutti gli altri elementi. La soluzione procede poi per minimizzazione della norma $||\mathbf{X} - \mathbf{AS}||$, vincolata, ad ogni iterazione, a soddisfare alle condizioni di identificabilità e alla informazione a priori, consistente nella presenza di zeri in posizioni note o anche al segno di alcuni degli elementi, nei casi in cui è noto l'effetto di promozione o inibizione di un fattore di regolazione sull'attività di un gene.

Questo approccio non include esplicitamente il rumore nei dati osservati. Sebbene rinunciando a sfruttare elaborati modelli statistici per l'errore di misura commesso negli esperimenti di sequenziamento, Jacklin et al. [5] introducono esplicitamente un processo di rumore, assumendo che una distribuzione additiva, gaussiana e bianca sia comunque una sufficiente approssimazione della realtà. Il loro approccio divide la soluzione del problema della separazione in due fasi: nella prima si stima $\mathbf{A}$, soggetta alle consuete condizioni e all'informazione a priori disponibile, mentre nella seconda si estraggono le colonne di $\mathbf{S}$ soggette alla conoscenza incompleta di $\mathbf{A}$. Riconoscendo che la forzatura di alcuni elementi di

A rende non convesso il problema di minima norma risolto iterativamente in [4], Jacklin et al. propongono una tecnica non iterativa, basata sull'identificazione della dimensione effettiva dello spazio dei dati, per la stima robusta di **A**, e un metodo ai minimi quadrati totali per la stima di **S** dato **A**.

4 Modello di espressione genica con il metodo RCA

Wang et al. [3] definiscono un'evoluzione del metodo NCA che consente di rendere meno rigidi i vincoli derivanti dalle conoscenze biologiche disponibili, che sono incomplete, come accennato, ma possono anche essere errate. Infatti, l'informazione biologica deriva in ogni caso dall'interpretazione di risultati sperimentali e, da un lato, questi sono sempre soggetti a errore, dall'altro, proprio la loro interpretazione può essere errata. Ne consegue che 1) un fattore di trascrizione ritenuto ininfluente sull'attività di un gene può invece influenzarla, per cui il corrispondente elemento della matrice **A** può essere erroneamente forzato a zero, oppure 2) il valore di qualche elemento di **A** può essere erroneamente lasciato libero, se si ritiene, erroneamente, che il corrispondente fattore regoli la trascrizione del gene cui l'elemento si riferisce. Il metodo RCA (Regulatory Component Analysis) proposto da Wang et al. introduce una plasticità nella matrice di mixing, dando la possibilità di modificare i suoi elementi durante l'esecuzione dell'algoritmo di risoluzione del sistema, e rimediando così all'eventuale presenza di falsi positivi o negativi nella conoscenza a priori. Un altro problema che la RCA affronta esplicitamente è che l'incompletezza dell'informazione biologica non consiste solo nel fatto che non di tutti i fattori di trascrizione si conosce la maggiore o minore influenza sui geni bersaglio, ma anche nel fatto di non conoscere esplicitamente quali e quanti fattori di trascrizione agiscono nella rete di regolazione. In altre parole, non solo gli esperimenti possibili non consentono di identificare quali elementi possono essere forzati a zero nella matrice di mixing, ma non si sa nemmeno quante siano davvero le sue colonne. Sotto condizioni meno stringenti di quelle richieste dalla NCA, la RCA è in grado di estrarre le componenti di regolazione ricorrendo alla decomposizione agli autovalori generalizzati di due sottomatrici estratte opportunamente dalla matrice **X**. Questo metodo consente anche di valutare la consistenza della conoscenza biologica a priori con i dati, permettendo di aggiornarla in base ai falsi positivi o negativi scoperti. Nell'investigare il problema della separazione delle sorgenti di regolazione genica, sembra che la conoscenza a priori rivesta un ruolo sempre più importante e si ponga alla base delle attuali tecniche risolutive. Oltre a offrire un metodo flessibile ed efficiente computazionalmente, la tecnica RCA lascia anche aperta la possibilità di includere ulteriori funzionali di regolarizzazione della soluzione. Se tali funzionali possono essere espressi come forme quadratiche, la soluzione del problema può ancora essere trovata sfruttando la decomposizione agli autovalori generalizzati su opportune coppie di matrici.

5 Conclusione

Lo studio delle reti di regolazione genica sulla base di dati di sequenziamento può portare a un significativo aumento della comprensione biologica di questi fenomeni. La classe di approcci qui esaminati viene ormai seguita da più di dieci anni. Come sempre, le prime applicazioni ricalcano fedelmente tecniche di supposto uso generale, senza preoccuparsi di introdurre specificità legate al problema concreto che si tenta di risolvere. Dal punto di vista teorico e dell'esperienza necessaria, il nostro laboratorio dispone di tutti gli strumenti per contribuire all'avanzamento di queste ricerche. Tradurre le conoscenze scientifiche in modelli adatti a descrivere fedelmente i fenomeni, o in vincoli che stabilizzino le soluzioni entro spazi consistenti, richiede però una stretta collaborazione e una comunanza di linguaggio con gli specialisti della materia.

Riferimenti bibliografici

1. Su-In Lee and S. Batzoglou, "Application of independent component analysis to microarrays", Genome Biology, 2003, 4:R76.
2. W. Liebermeister, "Linear modes of gene expression determined by independent component analysis", Bioinformatics, 2002, 18:51-60.
3. C. Wang, J. Xuan, I.-M. Shih, R. Clarke and Y. Wang, "Regulatory component analysis: A semiblind extraction approach to infer gene regulatory networks with imperfect biological knowledge", Signal Processing, Volume 92, Issue 8, 2012, pp. 1902-1915.
4. J.C. Liao, R. Boscolo, Y.L. Yang, L.M. Tran, C. Sabatti, and V.P. Roychowdhury, "Network Component Analysis: Reconstruction of Regulatory Signals in Biological Systems," Proc. Natl Academy of Sciences USA, vol. 100, no. 26, pp. 15522-15527, 2003.
5. N. Jacklin, Z. Ding, W. Chen, C. Chang, "Noniterative Convex Optimization Methods for Network Component Analysis", IEEE/ACM Trans. Comput. Biology Bioinform., Volume 9, Issue 5, 2012, pp. 1472-1481.
6. C.Q. Chang, Y.S. Hung, P.C.W. Fung, and Z. Ding, "Network component analysis for blind source separation," in Proc. 2006 International Conference on Communications, Circuits and Systems (ICCCAS2006), Guilin, China, June 2006, Vol. 1, pp. 323326.
7. C.Q. Chang, Y.S. Hung, and Z. Ding, "A New Optimization Algorithm for Network Component Analysis Based on Convex Programming," Proc. IEEE Intl Conf. Audio Speech and Signal Processing, Mar. 2009.
8. A.V. Camargo-Rodriguez, and J.T. Kim, "DoGeNetS: using optimisation to discriminate regulatory network topologies based on gene expression data", IET Syst. Biol. 6: pp.1-8, 2012.
9. J. Liu, M.M. Ghassemi, A.M. Michael, D. Boutte, W. Wells, N. Perrone-Bizzozero, F. Macciardi, D.H. Mathalon, J.M. Ford, S.G. Potkin, J.A. Turner, and V.D. Calhoun, "An ICA with reference approach in identification of genetic variation and associated brain networks", Front. Hum. Neurosci. 6: 21, 2012.