LA BIOINFORMATICA COME STRUMENTO PER LO STUDIO DELL'ESPRESSIONE GENICA DURANTE LO SVILUPPO EMBRIONALE

di Francesca Amati

La regolazione dell'espressione genica durante l'embriogenesi dei mammiferi è fondamentale per una normale anatomia e fisiologia. In questo articolo si discuteranno le più recenti applicazioni della tecnologia dei microarrays allo studio della regolazione dell'espressione genica embrionale e la collaborazione tra la Sezione di Genetica Medica dell'Università Tor Vergata ed il CASPUR per sviluppare un "ambiente informatico" adatto ad interpretare la grande mole di dati prodotta da queste nuove tecnologie.

Dott.ssa Francesca Amati
Università "Tor Vergata", Roma
Sezione di Genetica Medica
amati@med.uniroma2.it

• Abstract

Regulation of gene expression during embryogenesis and development is a crucial clue for a normal anatomy and physiology [1]. The combined use of experimental high-throughput methods, such as DNA microarrays, and bioinformatic methods has innovated the analysis of temporal patterns of gene expression in embryos. Microarray analysis, in fact, provides a large amount of data -at molecular level- that once acquired, must be functionally integrated in order to find common patterns within a defined group of biological samples.

Through the use of cDNA microarrays, investigators can measure mRNA [1] levels for thousands of genes simultaneously, rather than one gene at a time. In fact, DNA microarrays are constituted by small glass or filter matrix that contain arrays of DNA sequences (each highly specific to a single gene) and by means hybridization of fluorescent cDNA, they are capable of simultaneously quantifying the expression of thousands of genes in a single experiment. The results of these experiments are spots whose brightness varies from gene to gene corresponding to the transcriptional activity of the examined genes (Fig. 1). This analysis requires sophisticated bioinformatics tool [2]. An interestingly application of DNA microarrays is the analysis of mRNA expression (such as the **transcriptome**) in embryos. A systematic genomic approach to analyze global gene expression patterns and functions during embryogenesis has recently been named developmental genomics or **embryogenomics** [3].

Functional genomics of embryo development requires the integration of information from genome sequence and structure, gene and protein expression, and metabolite profiles with knowledge databases by using computational and bioinformatics tools [1,4].

[1] mRNA (messanger RNA o RNA messaggero): la popolazione di acidi nucleici presente all'interno di una cellula e che hanno la funzione di determinare la formazione di proteine funzionali alla cellula stessa. L'RNA messaggero ha la funzione di trasferire l'informazione genetica contenuta nel DNA.

Il 14 Aprile 2003 l'American Institute for the Human Genome Project annunciò agli scienziati di avere appena trovato l'ultimo tassello del complesso puzzle costituito dalla sequenza del DNA umano. Uno dei risultati più sorprendenti del completamento del Progetto Genoma Umano è stato quello di determinare un numero dei geni umani (circa 20.00) sensibilmente ridotto rispetto a quanto stimato in precedenza (oltre 100.000). Questo dato ha decisamente avvalorato l'idea che gli organismi più complessi (come i mammiferi) estendano la funzionalità dei loro genomi attraverso differenti meccanismi molecolari come l'attivazione di differenti promotori[2], lo *splicing* alternativo[3], le modificazioni post-trascrizionali dell'RNA[4], l'espressione di RNA non codificanti[5], di RNA antisenso[6] e di micro RNA[7] [5,6]. Molti di questi meccanismi operano insieme per costituire quello che viene chiamato il **trascrittoma** di un organismo. Trascrittoma è il termine che definisce la collezione di tutti gli RNAm (RNA messaggeri) presenti in una cellula o tessuto in un determinato tempo o situazione fisiologica. L'analisi del trascrittoma mediante l'utilizzo della tecnica dei *microarrays* è sempre più utilizzata per investigare ad esempio le basi molecolari di una malattia genetica umana [7,8]. Infatti la determinazione di un'alterazione dei livelli di espressione genica in un tessuto patologico rispetto ad un controllo fornisce importanti informazioni sulle possibili cause molecolari della malattia. La tecnologia dei *microarrays* permette ai ricercatori di analizzare il livello di espressione di migliaia di geni contemporaneamente. Dal 1995, anno in cui è stato pubblicato per la prima volta il termine "microarray [9], tale tecnica ha trovato largo impiego in diversi campi, come dimostrato dal numero elevato di articoli (circa 30.000) pubblicati fino ad oggi (www.pubmed.gov). Il grande sviluppo e l'ampia applicazione di questa tecnica sono giustificati dalla sua potenzialità di consentire lo studio dei livelli di espressione di migliaia di geni simultaneamente in un unico esperimento attraverso l'ibridazione di acidi nucleici su un supporto solido. Infatti i metodi classici di analisi dell'espressione genica, come il Northern blotting, permettono di analizzare solo pochi geni alla volta. In sostanza i *microarrays* rappresentano modelli miniaturizzati d'ibridazione. Nella biologia molecolare, lo studio degli acidi nucleici mediante **ibridazione** è un sistema universalmente adottato e si basa sulle caratteristiche peculiari della doppia elica del DNA, ovvero la natura complementare delle due catene e la specificità dell'accoppiamento delle basi (Figura 1).

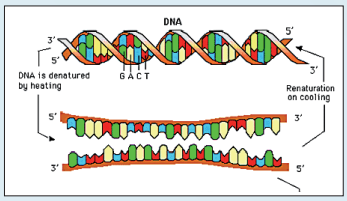


Figura 1
Appaiamento degli acidi nucleici.

[2] Sequenza di nucleotidi posta a monte della sequenza codificante di un gene e che ne regola l'espressione, attraverso la presenza di specifici elementi (le sequenze di riconoscimento dei fattori di trascrizione).

[3] Processo cellulare in seguito al quale a partire da un gene si originano differenti trascritti (ovvero mRNA) che determinano la formazione di isoforme proteiche che possono essere anche tessuto-specifiche.

[4] Tutta quella serie di modificazioni del trascritto primario di RNA che portano alla formazione di una molecola di RNA messaggero maturo.

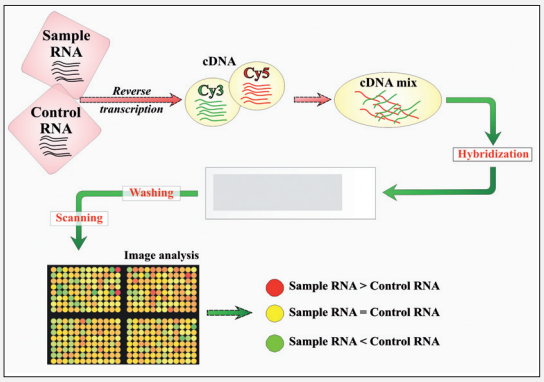
[5] Trascritto genico che non produce una proteina; ovvero l'RNA non codificante è una molecola di RNA (e quindi non un mRNA, perché altrimenti andrebbe incontro a traduzione) che può svolgere diverse funzioni: può essere un RNA ribosomiale (rRNA), RNA transfer (tRNA) oppure può essere un componente di complessi enzimatici implicati nei processi di trascrizione, replicazione, *splicing* e in altri processi riguardanti l'espressione genica.

[6] Gli RNA antisenso sono frammenti di RNA a doppio filamento in grado di interferire (e spegnere) l'espressione di specifici mRNA. Il processo tramite cui gli RNA antisenso svolgono questa importante funzione di regolazione dell'espressione genica è chiamata RNA interference (dall'inglese interferenza dell'RNA, abbreviata comunemente come RNAi).

[7] I microRNAs sono piccole molecole di RNA, a singolo filamento di 20-22 nucleotidi, che svolgono diverse funzioni, la più nota attualmente è una regolazione post-trascrizionale dell'RNAm.

Un DNA *microarray* è costituito da una collezione di migliaia di specifiche sequenze di acidi nucleici, chiamate sonde o "probe", immobilizzate su una superficie solida (vetro o nylon), dove ogni sequenza unica forma uno spot. Il campione di acido nucleico da testare, chiamato bersaglio o "target", viene marcato e poi fatto ibridare in condizioni altamente stringenti (Figura 2). L'ibridazione tra probe e target è generalmente individuata e quantificata tramite un sistema di rilevazione a fluorescenza.

Figura 2
Schema riassuntivo di un esperimento di microarray.



In questo modo se la quantità di RNA espressa da un gene nelle cellule di interesse è aumentata (up regolata) rispetto a quella del campione di riferimento, lo spot che ne risulta sarà del colore del primo fluorescente. Viceversa se l'espressione del gene è diminuita (down regolata) rispetto al campione di riferimento lo spot risulterà colorato dal secondo fluorescente (Figura 2). Il segnale fluorescente rilevato dallo scanner viene poi misurato nei due canali e confrontato con il segnale di background. La normalizzazione è la procedura con la quale si eliminano dai dati di *microarray* le variazioni sistematiche dovute ad effetti quali la differente efficienza di marcatura, la differente efficienza fluorometrica, gli effetti spaziali sulla superficie del *microarray*, etc. Vari metodi bioinformatici e statistici sono stati sviluppati sia per la fase di normalizzazione che per l'analisi finale, cioè la definizione dei geni differenzialmente espressi. L'incrocio dei dati dell'esperimento con database bioinfomatici quali il Gene Ontology e i Kegg pathway , permettono infatti di caratterizzare il profilo trascrittomico del campione e di trarre conclusioni biologicamente significative.

Le domande fondamentali a cui si può rispondere analizzando l'espressione genica negli embrioni (*embryogenomics*) [4] sono:

- quanti geni sono espressi, quali sono questi geni e quale è il loro livello di espressione?
- Come questi "programmi" di espressione cambiano in base ad alterazioni funzionali o strutturali di geni?

L'applicazione della tecnica dei *microarray* allo studio dello sviluppo embrionale è recente, ma sta sempre più diffondendosi. Questa diffusione porterà sempre più informazioni riguardo la regolazione dell'espressione genica durante lo sviluppo embrionale [10].

In generale, possiamo affermare che la tecnologia dei *microarray* permette l'identificazione di geni co-regolati in un determinato sistema biologico. Una volta

identificati tali geni si può procedere ad esempio all'analisi di specifici domini proteici. Combinando tali dati con quelli provenienti dallo studio sistematico di regioni di regolazione genica (i promotori) si può iniziare a ricostruire il complesso sistema di regolazione della cellula [11].

La bioinformatica e la genomica sono discipline strettamente connesse che hanno permesso lo sviluppo di ricerche complesse nell'ambito della sanità pubblica, della farmacologia, della genomica comparativa, ecc. L'avvento dei *microarrays* per l'analisi globale dell'espressione genica [9] ha generato la necessità di interpretare l'enorme mole di dati ottenuti e di riordinarli in base ad esempio all'espressione temporale. In effetti è ragionevole pensare che geni con una funzione simile abbiano un pattern di espressione simile, presumibilmente dovuto ad una comune regolazione trascrizionale, o che appartengano a vie metaboliche o regolatrici comuni. Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) e ArrayExpress (http://www. ebi.ac.uk/microarray-as/ae/) sono due dei principali database per il deposito dei dati ottenuti da esperimenti di espressione genica di *microarray*. La bioinformatica è uno strumento fondamentale per estrarre da questa enorme mole di dati e di informazioni utili sulla regolazione di un gene, sulla sua espressione temporale e tissutale, sul suo ruolo funzionale. A tale scopo sono stati sviluppati algoritmi come l'*hierarchical clustering* ed il SOM [12,14]. Oltre queste analisi, si è originata una parallela richiesta di metodi bioinformatici per l'interpretazione dei risultati, ovvero metodi utili per tradurre i dati sperimentali in conoscenza biologicamente utile [15]. Uno dei siti bioinformatici più utilizzati per questo fine è il GO database (Gene Ontology Consortium http://www. geneontology.org), in cui vengono annotate le informazioni funzionali di ogni gene eucariotico[8] e procariotico[9]. La missione primaria del GO database è di facilitare la comprensione delle relazioni funzionali del genoma (geni e proteine). I prodotti funzionali di un gene (proteine) sono classificati in base alla funzione molecolare, al processo biologico e alla localizzazione cellulare, utilizzando soprattutto informazioni ottenute dalla revisione della letteratura scientifica e/o, se queste non sono disponibili, dall'omologia con proteine di altre specie. Un'altra banca dati importante per la comprensione funzionale di un gene è KEGG (http://www.genome.ad.jp/kegg/pathway.html). Questo database contiene una collezione di mappe rappresentanti l'attuale conoscenza sulle reti molecolari di reazione e di interazione delle vie metaboliche e di alcune vie regolatrici conosciute. Oltre al deposito dell'enorme quantità di dati ottenuti da esperimenti di *microarray*, alla loro catalogazione e classificazione ed alla ricostruzione di vie metaboliche o di regolazione note, in questi ultimi anni, la bioinformatica è stata applicata allo studio comparativo di sequenze nucleotidiche al fine di identificare le sequenze regolatrici (*transcription factor binding sites*, TFBS e microRNA, miRNA) responsabili dell'espressione di un dato gene. Tuttavia, predire le sequenze regolatrici all'in-

[8] Gene presente negli organismi viventi mono- o pluricellulari costituiti da cellule dotate di nucleo.

[9] Gene presente negli organismi viventi unicellulari (o, al più, coloniali) di dimensioni dell'ordine del micrometro, senza membrana nucleare o altre suddivisioni interne.

terno di una lunga sequenza genomica è molto più difficile dell'identificazione di un gene. Le sequenze del genoma che contengono tali sequenze regolatrici sembrano essere conservate fra le varie specie. Quindi l'allineamento delle sequenze genomiche di parecchie specie può facilitare molto l'identificazione degli elementi regolatori della trascrizione. Questo processo è chiamato "impronta filogenetica" [16]. Molti algoritmi bioinformatici sono stati quindi progettati per questo scopo; MULAN è uno di questi [17]. MULAN permette l'allineamento e la visualizzazione di più sequenze ed è stato sviluppato per l'identificazione di elementi di regolazione evolutivamente conservati. Il codice MatInspector identifica TBFS usando *weight matrices* e può essere applicato su altri programmi di analisi di sequenza, quali DiAlignTF, FrameWorker o SequenceShaper [18,19]. Una procedura originale per l'identificazione di TBFS conservati è stata realizzata recentemente nel codice WeederH [20].

La regolazione dell'espressione di gene in un genoma eucariotico è determinata da un'attività cooperativa complessa che coinvolge i promotori prossimali ed elementi regolatori distanti. L'identificazione di geni (ottenuta da esperimenti di *microarrays*) e delle loro vie di regolazione necessita di analisi bioinformatiche *high-throughput* che comprendono il raggruppamento dei geni co-regolati (*clustering*) e la ricerca di elementi funzionali comuni. È però importante rammentare che la conservazione di un TFBS non è una prova che questo elemento sia realmente funzionale, tuttavia queste informazioni sono utili per una migliore comprensione della regolazione dell'espressione di un gene.

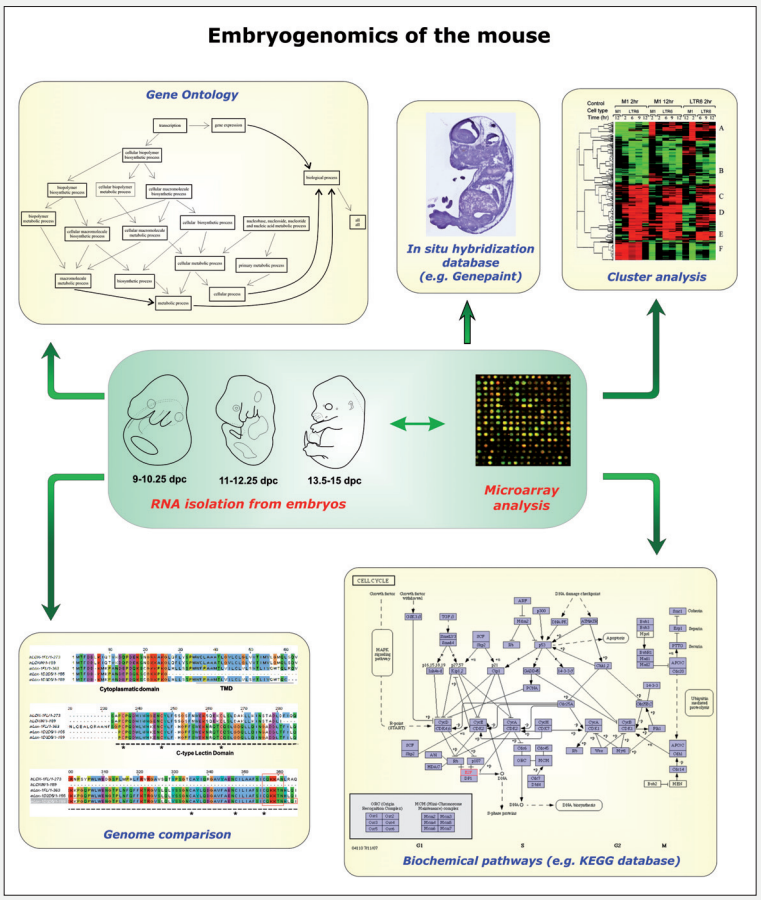


Figura 3
Interazioni tra tecniche di analisi del trascrittoma, proteoma e metaboloma.

• Conclusioni

I metodi di analisi su larga scala descritti in questo articolo, facilitano enormemente lo studio delle relazioni funzionali tra i nostri geni (genoma), i loro trascritti (trascrittoma) e i prodotti proteici (proteoma) (Figura 3). È però evidente come sia necessaria una sempre maggiore integrazione dei differenti algoritmi bioinformatici applicati all'enorme mole di dati che si ottengono da esperimenti di DNA *microarrays*. Ne consegue che la genomica funzionale raggiungerà l'obiettivo di comprendere la complessa funzionalità del trascrittoma embrionale solo quando la bioinformatica fornirà il supporto essenziale per la completa integrazione dei dati ottenuti (Figura 3).

• Bibliografia

[1] Amati, F., Chillemi, G., & Novelli, G. (2009). Gene Expression Analysis during Development by High-Throughput Methods in Development Gene Expression Regulation, Nova Science Publishers.

[2] Hovatta, I., Kimppa, K., Lehmussola, A. *et al.* (2005). Picaset Oy, Helsinki, The authors and CSC Scientific Computing Ltd.

[3] Ko, M.S. (2001). *Trends Biotechnol,* ***19***, 511-18.

[4] Amati, F., Biancolella, M., Farcomeni, A., Giallonardi, S., Bueno, S., Minella, D., Vecchione, L., Chillemi, G., Desideri, A., & Novelli, G. (2007). *Gene,* ***391***, 91-102.

[5] Rassoulzadegan, M., Grandjean, V., Gounon, P., Vincent, S., Gillot, I., & Cuzin, F. (2006). *Nature,* ***441***, 469–474.

[6] Heintzman, N. D. & Ren, B. (2007). *Cell Mol Life Sci,* ***64***, 386-400.

[7] Amati, F., Biancolella, M., D'Apice, M. R., Gambardella, S., Mango, R., Sbraccia, P., D'Adamo, M., Margotti, K., Nardone, A., Lewis, M., & Novelli, G. (2004). *Gene Expr.,* ***12***(1), 39-47

[8] Botta, A., Vallo, L., Rinaldi, F., Bonifazi, E., Amati, F., Biancolella, M., Gambardella, S., Mancinelli, E., Angelini, C., Meola, G., & Novelli, G. (2007). *Gene Expr.,* ***13***(6), 339-51.

[9] Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). *Science,* ***270***, 467-70.

[10] Niemann, H., Carnwath, J. W., & Kues, W. (2007). *Theriogenology,* ***68S***, S165–S177.

[11] Kendall, S. L., Movahedzadeh, F., Wietzorrek, A., & Stoker, N. G. (2002). *Comp Funct Genomics,* ***3***, 352-54.

[12] Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). *Proc. Natl. Acad. Sci.,* ***95***, 14863-68.

[13] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., & Golub, T. R. (1999). *Proc. Natl. Acad. Sci.,* ***96***, 2907-12.

[14] Törönen, P., Kolehmainen, M., Wong, G., & Castrén, E. (1999). *FEBS Lett,* ***451***, 142-46.

[15] Al-Sharour, F., Carbonell, J., Minguez, P., Goetz, S., Conesa, A., Tárraga, J., Medina, I., Alloza, E., Montaner, D., & Dopazo, J. (2008). *Nucleic Acids Res,* ***36*** (Web Server issue) W341-46.

[16] Hooghe, B., Hulpiau, P., van Roy, F., & De Bleser, P. (2008). *Nucleic Acids Res,* ***36*** (Web Server issue), W128-32.

[17] Ovcharenko, I. & Nobrega, M. A. (2005). *Nucleic Acids Res,* ***33***, W403-07.

[18] Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., & Werner, T. (2005). *Bioinformatics,* ***21***, 2933-42.

[19] Morgenstern, B. (2007). *Methods Mol Biol,* ***395***, 195-204.

[20] Pavesi, G., Zambelli, F., & Pesole, G. (2007). *BMC Bioinformatics,* ***8***, 46.