Il ruolo della Bioinformatica nella scoperta delle relazioni funzionali tra microRNA di piante e geni umani: dall'ipotesi concettuale alla validazione sperimentale

Flavio Licciulli[1], Arianna Consiglio[1], Faouzi Serroukh[2], Domenica D'Elia[1], Giorgio Grillo[1], Sabino Liuni[1], Elisabetta Sbisà[1], Apollonia Tullo[3], Flaviana Marzano[1,3], Mariano Francesco Caratozzolo[1], Domenico Catalano[1]

[1]Istituto di Tecnologie Biomediche CNR Bari

[2]Campus technique de la Haute Ecole en Hainaut, Mons, Belgio

[3]Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari CNR Bari

La disponibilità di dati di sequenziameto di nuova generazione ha permesso negli ultimi decenni un progresso velocissimo della conoscenza sia in ambito biologico che clinico. Nulla di tutto questo sarebbe stato possibile se in parallelo non si fossero sviluppati strumenti altrettanto sofisticati e avanzati nel campo della Bioinformatica per la loro gestione, analisi e interpretazione.

Il lavoro che presentiamo illustra, attraverso un caso studio sviluppato nel nostro Istituto, l'importanza della Bioinformatica come strumento predittivo per l'elaborazione e la validazione di un'ipotesi funzionale che viene teorizzata *in silico* e validata in laboratorio, per poi essere di nuovo utilizzata per l'analisi dei dati di laboratori e quindi per orientare ulteriori sperimentazioni e produrre nuova conoscenza.

Il caso studio riguarda l'analisi della possibilità che micro RNA (miRNAs) vegetali assunti con la dieta possano regolare l'espressione di geni umani mimando i meccanismi e le funzioni di miRNAs endogeni. Questa ipotesi è stata elaborata sulla base di una serie di ricerche condotte negli ultimi anni che prendono spunto dal primo lavoro pubblicato in letteratura che riporta la dimostrazione di un targeting funzionale di un miRNA di pianta su di un gene umano (1).

Lo studio in silico è stato condotto sulla base di un workflow che ha previsto: 1) la selezione di miRNA di piante edibili, rispetto alle non edibili, collezionate nella banca dati miRBase (2), sulla base delle informazioni tassonomiche in essa contenute; 2) la verifica della possibilità che miRNA di piante e umani potessero avere sequenze "seed", cioè quelle deputate al targeting e localizzate nella regione al 5' del miRNA, identiche; 3) la selezione di un gruppo ristretto di miRNA di piante da usare per la validazione sperimentale sulla base della classificazione funzionale dei potenziali geni target in specifici processi biologici. Per l'estrazione dei dati dalle banche dati pubbliche e la loro integrazione con i risultati delle analisi condotte in loco è stata sviluppata una banca dati relazionale in MySQL. Per l'analisi comparativa al punto 2) è stata sviluppata una pipeline di analisi bioinformatica che ingloba in bash Needleall e Shuffleseq, due programmi inclusi nel pacchetto di analisi bioinformatica EMBOSS (3). Grazie a questa pipeline 2.588 sequenze mature di miRNA umani sono state allineate ciascuna con 4.803 sequenze mature di miRNA di piante edibili, per un totale di 12 milioni di allineamenti. Due script (Needle_parser_CD.pl e Pl_hum.csh) sono stati impiegati per selezionare gli allineamenti in funzione dei valori di similarità osservata. La validità statistica dei risultati ottenuti è stata poi valutata usando il programma Shuffleseq, che permette di ottenere gruppi di sequenze randomizzate di uguale lunghezza e composizione su cui ripetere le stesse analisi. Il confronto dei risultati degli allineamenti delle sequenze randomizzate con quelli ottenuti dal precedente allineamento è stato fatto utilizzando il test del chi-quadro. I risultati di questo test hanno validato i risultati dell'allineamento come significativi. Dei 4.803 miRNA di piante analizzati, 2855 hanno nella regione al 5' una sequenza identica a quella del seed di miRNA umani, ma nessuno di essi è perfettamente identico a un miRNA umano per tutta la sua lunghezza.

Questo risultato supporta l'ipotesi che il meccanismo attraverso il quale i miRNA di piante potrebbero modulare o inibire l'espressione dei geni umani deve necessariamente essere molto simile a quello usato dai miRNA endogeni. Questi risultati hanno permesso di passare al punto 3) del workflow. La selezione dei miRNA di piante da utilizzare è stata orientata da un'analisi funzionale dei gruppi di potenziali geni target così come riportati in miRTarBase (4), una banca dati che contiene dati di validazione sperimentale di targeting di miRNA su geni umani. I dati contenuti in questa banca dati sono stati estratti e riversati nella nostra banca dati, il che ci ha permesso di mettere in correlazione diretta i miRNA di piante con seed identico ai miRNA umani con i geni target sperimentalmente validati di questi ultimi in miRTarBase. L'analisi funzionale dei geni umani, potenziali target dei miRNA di piante edibili con seed identico a quelli umani validati sperimentalemente, è stata condotta utilizzando diversi strumenti come g:Profiler (5) e DAVID (6, 7) che forniscono dati circa l'arricchimento statistico delle liste di geni analizzate, attraverso tipi diversi di test, in diversi processi biologici (Gene Ontology) e pathways (Reactome, KEGG) utilizzando oltre che specifiche banche dati anche tecnologie di text e data mining per il clustering dei geni in specifiche categorie funzionali. Gruppi funzionali specifici di geni sono stati anche analizzanti con strumenti che consentono di rilevare network di interazione funzionale, come StringDb (8) e Cluego (9). Il risultato di questa analisi ha permesso la selezione 7 miRNA di piante, con seed identico (e quindi potenziali omologhi funzionali) di 35 miRNA umani, per la loro predetta capacità di regolare geni coinvolti nel ciclo cellulare e quindi nella progressione tumorale e nella formazione delle metastasi. Questi miRNA sono stati usati in singolo e in mix per trasfettare colture cellulari immortalizzate di cellule di cancro al colon. I risultati dei saggi di proliferazione cellulare effettuati in laboratorio a tempi e concentrazioni diverse dei miRNA selezionati (vedi abstract Marzano et al.), hanno evidenziato una significativa riduzione della proliferazione cellulare, supportando la validità dell'ipotesi di partenza e dell'analisi bioinformatica predittiva. L'analisi di trascrittomica su campioni selezionati di cellule trasfettate a tempi diversi è stata condotta utilizzando procedure standardizzate di analisi di dati NGS e di nuovo, le liste di geni differenzialmente espressi ottenute dall'analisi statistica dei dati di trascrittomica con i su citati strumenti di analisi funzionale, ci hanno permesso di individuare potenziali geni target che sono stati poi validati sperimentalmente attraverso saggi di Luciferasi.
I risultati di questo studio, oltre che essere estremamente incoraggianti per il valore scientifico che tali dati possono avere per ulteriori indagini sperimentali o lo sviluppo di interessanti applicazioni per la ricerca scientifica nel campo della nutrigenomica, rappresentano un valido modello di approccio multidisciplinare integrato tra la ricerca biologica e quella bioinformatica, un requisito oramai indispensabile per lo studio delle scienze omiche.

Referenze

1. Zhang L, Hou D, Chen X, Li D, Zhu L, Zhang Y, et al. Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Res*. 2012; 22:107-126. doi: 10.1038/cr.2011.158. Epub 2011 Sep 20.
2. Griffiths-Jones S. miRBase: microRNA sequences and annotation. Curr Protoc Bioinformatics. 2010 Mar;Chapter 12:Unit 12.9.1-10. doi: 10.1002/0471250953.bi1209s29.
3. Rice P, Longden I and Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics.* 2000; 16:276—277.

4. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res*. 2018;46(D1):D296-D302. doi: 10.1093/nar/gkx1067.
5. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J. g:Profiler -- a web server for functional interpretation of gene lists (2016 update) *Nucleic Acids Research.* 2016; doi: 10.1093/nar/gkw199.
6. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* 2009;4:44-57.
7. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009;37:1-13.
8. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015;43(Database issue): D447–D452. doi: 10.1093/nar/gku1003.
9. Bindea G and Mlecnik B. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009;25:1091-1093. doi: 10.1093/bioinformatics/btp101. Epub 2009 Feb 23.