Maria Luisa Chiusano*, Nunzio D'Agostino*,
Alessandra Traini*, Miriam Di Filippo*, Luigi Frusciante*

ISOL@: una piattaforma bioinformatica per l'analisi strutturale e funzionale del genoma del pomodoro

introduzione

La famiglia delle Solanaceae comprende circa 95 generi e almeno 2.400 specie molte delle quali hanno una notevole importanza economica, in quanto largamente coltivate come fonte di cibo (pomodoro, patata, melanzana, peperone), a scopo ornamentale (petunia), a scopo farmaceutico per il loro contenuto in alcaloidi (tabacco, belladonna, stramonio). Le specie appartenenti a questa famiglia mostrano un'ampia variabilità fenotipica e occupano differenti nicchie ecologiche nonostante presentino una elevata conservazione del genoma intesa sia come numero di base dei cromosomi (12) sia come livello di similarità lungo le sequenze genomiche.

La necessità di approfondire la conoscenza sulle caratteristiche molecolari che determinano la variabilità fenotipica e l'adattamento ecologico delle diverse specie di Solanaceae ha portato, nell'anno 2004, a far convergere gli sforzi dei singoli gruppi di ricerca in un progetto organico denominato "International Solanaceae (SOL) Genome Project" (Mueller at al., 2005a).

In questo ambito è stata lanciata l'iniziativa "International Tomato Genome Sequencing Project" (Mueller at al., 2005b), con il fine di sequenziare il genoma del pomodoro a partire dalla definizione della sequenza completa della regione eucromatica. L'obiettivo a lungo termine è quello di sfruttare le informazioni di sequenza generate, per l'analisi dell'organizzazione e della funzionalità del genoma, nonché per comprendere i meccanismi evolutivi alla base della diversificazione dell' intera famiglia delle Solanaceae.

* *Dip. Scienze del Suolo, della Pianta, dell'Ambiente e delle Produzioni animali, Università degli Studi di Napoli Federico II*

Al fine di affrontare le questioni chiave sollevate dalla visione del SOL, sono in corso di generazione grandi quantità di dati derivati da diversi approcci '-omici'. Questi dati vanno ad arricchire quelli resi disponibili per le Solanaceae nel corso degli anni.

Gli approcci di tipo bioinformatico sono gli unici che consentono l'analisi di quantità di dati cospicui ed eterogenei come quelli che la comunità scientifica delle Solanaceae produce allo scopo di identificare caratteristiche strutturali utili a comprendere i processi molecolari che determinano la variabilità fenotipica degli organismi e le loro complesse funzionalità.

La bioinformatica, pertanto, esercita un ruolo chiave nell'interpretazione dei dati grezzi e nella loro conversione in informazione biologicamente significativa. Inoltre consente una visione 'olistica' con un approccio determinante nel rilevare le proprietà emergenti dall'insieme che caratterizza il sistema biologico. La necessità di tale disciplina nella ricerca scientifica moderna è legata allo sviluppo di metodologie che consentono di analizzare, integrare e organizzare grandi quantità di dati al fine di renderli fruibili dalla comunità scientifica mediante risorse idonee. La differenziazione delle risorse, l'eterogeneità dei dati, la qualità delle loro annotazioni e, infine, il livello di dettaglio introdotto nelle specifiche analisi costituiscono il valore aggiunto per un approccio bioinformatico efficace che, di conseguenza, si complica ponendo sfide sempre più consistenti sia dal punto di vista biologico sia dal punto di vista informatico.

Al fine di contribuire in questo ambito, abbiamo progettato e organizzato ISOL@ (Chiusano et al. 2008), una risorsa italiana per la genomica delle Solanaceae. Essa è stata concepita per raccogliere, integrare e riconciliare i risultati generati dai diversi approcci '-omici' all'interno del consorzio SOL al fine di creare una piattaforma utile per sostenere la ricerca su molti degli aspetti delle genomica delle Solanaceae.

IL SEQUENZIAMENTO DEL GENOMA DEL POMODORO

Il sequenziamento del genoma del pomodoro (*Solanum lycopersicum* cv. Heinz 1706) è in corso e la sua prima bozza è stata recentemente completata (Mueller et al., 2009). Il genoma (950 Mb; Arummganathan & Earle 1991) è organizzato in 12 cromosomi ed è costituito in larga parte da eterocromatina che è ricca in sequenze ripetute (Peterson et al., 1996). È stato stimato che la maggior parte dei geni (>90%) risiede nella porzione eucromatica del genoma, la quale corrisponde a circa un quarto (~250 Mb) dell'intera sequenza genomica (Van der Hoeven et al., 2002).

Analisi citogenetiche (De Jong et al., 1999, 2000) hanno mostrato che l'eucromatina si colloca generalmente nella parte distale dei cromosomi (fatta eccezione per i telomeri) e circonda la regione eterocromatica pericentromerica.

Poiché il DNA eterocromatico è ricco in sequenze ripetute (pertanto difficile da sequenziare ed assemblare) e povero in geni, il consorzio ha deciso, in prima istanza, di sequenziare soltanto la porzione eucromatica del genoma. L'attuazione del progetto coinvolge dieci paesi, tra i quali l'Italia, che si sta occupando di determinare la sequenza nucleotidica del cromosoma 12.

Per il sequenziamento è stata adottata la strategia BAC-by-BAC (Bacterial artificial chromosome), già impiegata con successo nel sequenziamento di altri genomi quali *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000) e riso (International Rice Genome Sequencing Project, 2005). Questa strategia consiste nel sequenziare alcuni BAC (definti *seed*) che risultano ancorati alla mappa genetica ad alta densità F_2-2000 (vedi Fulton et al., 2002). Tali BAC rappresentano dei punti di partenza per estendere la sequenza mediante BAC *walking* in modo da generare il '*Minimum Tiling Path*' cioè individuare la collezione di frammenti genomici in grado di coprire la regione cromosomica di interesse con il minor grado di sovrapposizione. Nell'ottobre 2008 erano disponibili nella banca dati GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) 1095 sequenze BAC equivalenti a ~120 Mb.

Con l'emergere delle cosiddette "tecnologie di sequenziamento di nuova generazione" (Roche 454, Illumina Solexa, Applied Biosystems SOLiD) si sono rese disponibili valide alternative strategiche e tecnologiche da applicare al sequenziamento del genoma di pomodoro. Pertanto, nell'ottobre 2008 cinque membri del consorzio SOL, tra i quali l'Italia, hanno lanciato una iniziativa che ha orientato la comunità scientifica a considerare fattibile il sequenziamento dell'intero genoma mediante un approccio 'whole genome shotgun' (WGS). L'obiettivo è quello di combinare tale approccio con i risultati già disponibili, ottenuti applicando la strategia BAC-by-BAC, per portare a conclusione il sequenziamento dell'intero genoma di pomodoro entro il 2010.

L'ANNOTAZIONE DEL GENOMA DEL POMODORO

Nell'ambito del progetto internazionale di sequenziamento del genoma del pomodoro è stato fondato un gruppo di lavoro, l'iTAG (international Tomato Annotation Group), che vede coinvolte unità di bioinformatica pro-

venienti da Europa, Asia e Stati Uniti, il cui obbiettivo finale è quello di produrre un'annotazione omogenea e attendibile del genoma mediante uno sforzo condiviso e distribuito tra i vari partecipanti.

Attualmente l'iTAG lavora sull'insieme di sequenze BAC sequenziate nell'ambito del progetto genoma e distribuite via web sia dal *Solanaceae Genomics Network* (SGN; http://solgenomics.net/) sia dalla banca dati GenBank (http://www.ncbi.nlm.nih.gov/Genbank/).

La procedura di annotazione implementata *ad hoc* è anche adatta per analizzare le sequenze genomiche che cominciano a essere prodotte mediante tecnologie di nuova generazione e secondo un approccio WGS.

I passaggi dell'analisi prevista al momento sono descritti di seguito: 1) allineamento di sequenze espresse lungo le sequenze genomiche e definizione di una collezione attendibile di geni modello; 2) predizione di geni mediante la piattaforma integrata EuGene (Foissac et al., 2003; 2008) e annotazione funzionale dei geni predetti (Blake et al., 2002; Mulder et al., 2007; Quevillon et al., 2005); 3) identificazione di geni non codificanti mRNA mediante metodi appropriati. Nell'ambito di questo sforzo distribuito, il nostro gruppo è coinvolto nell'organizzazione e catalogazione di collezioni EST (Expressed Sequence Tag) di pomodoro e di altre specie di Solanaceae da allineare lungo i BAC (D'Agostino et al., 2007a; 2009), e in secondo luogo nella definizione di una collezione attendibile di geni modello utile nell'addestramento dei programmi di predizione genica *ab initio* (Yao et al., 2005; D'Agostino et al., 2007b).

LA BANCA DATI SOLEST: UN APPROCCIO "ONE-STOP-SHOP" PER LO STUDIO DEL TRASCRITTOMA DELLE SOLANACEAE

Sebbene siano al momento in corso sforzi per il sequenziamento del genoma di pomodoro (Mueller et al., 2009), patata (Visser et al., 2009) e tabacco (http://www.tobaccogenome.org/), gran parte dei dati di sequenza disponibili per le Solanaceae sono costituiti da collezioni di EST.

Le EST rappresentano brevi frammenti di sequenze espresse derivate dal sequenziamento *single pass* di librerie di cDNA. Queste ultime sono ottenute a partire da un estratto cellulare di mRNA il quale viene trasformato in DNA a doppio filamento (cDNA) e inserito in un vettore batterico.

Con l'obiettivo di contribuire all'analisi trascrittomica delle Solanaceae, sono state collezionate sequenze EST e mRNA provenienti da differenti specie sia coltivate sia selvatiche.

Sfruttando una procedura automatizzata da noi progettata *ad hoc*, (ParPEST; D'Agostino et al., 2005) le EST in ciascuna collezione sono processate per rimuovere eventuali sequenze contaminanti e individuare e mascherare sequenze ripetute; sono quindi raggruppate in *cluster*, ognuno dei quali rappresenta un gene; infine le sequenze EST in ciascun *cluster* sono opportunamente assemblate nel tentativo di ricostruire l'intera sequenza dell'mRNA dal quale le EST derivano. La procedura di '*assemblaggio*' può produrre *tentative consensus sequences* (TCs), ossia assemblati di due o più sequenze, e singoletti, ovvero sequenze non incluse in alcun TC. Il numero risultante di trascritti (TCs + singoletti) definiti dalla strategia descritta è, quindi, sottoposto ad annotazione funzionale, ossia a una procedura per l'identificazione della regione codificante per la proteina, mediante confronto con la banca dati proteica UniProt/Swiss-prot (http://www.uniprot.org). La descrizione della funzione di ciascun trascritto include, laddove possibile, le ontologie geniche, che sono ricavate da un vocabolario controllato e strutturato per la descrizione dei prodotti genici (Gene Ontology, http://www.geneontology.org/); la classificazione e nomenclatura degli enzimi ricavate dalla banca dati ENZYME (http://www.geneontology.org/) e l'associazione di ciascun trascritto codificante un enzima alle mappe metaboliche già organizzate nella banca dati KEGG (http://www.genome.jp/kegg/), da noi considerata come riferimento.

Le sequenze EST grezze, i dati intermedi e le informazioni relative al raggruppamento e assemblaggio delle sequenze e all'annotazione funzionale sono stati raccolti in una banca dati relazionale. È stata implementata una interfaccia grafica per consultare il database (http://biosrv.cab.unina.it/solestdb) che comprende anche menù ad albero per una facile interrogazione da parte dell'utente.

Le sequenze EST di Solanaceae sono state collezionate ed analizzate con lo scopo di campionare il trascrittoma delle Solanaceae fornendo un ampio catalogo di trascritti e con lo scopo di contribuire in maniera determinante all'annotazione delle sequenze genomiche in corso di produzione per pomodoro e per altre Solanaceae evidenziandone le caratteristiche strutturali e funzionali.

ANNOTAZIONE DI SEQUENZE GENOMICHE

In qualità di membri dell'iTAG, abbiamo implementato una procedura automatizzata, su piattaforma di calcolo parallelo, per l'annotazione delle sequenze BAC rese disponibili alla comunità scientifica attraverso la banca dati GenBank. Tale procedura consente quotidianamente di scaricare e analizzare

ogni nuova sequenza BAC disponibile e allineare lungo di essa le sequenze EST collezionate nella banca dati SolEST (D'Agostino et al., 2009) e i corrispondenti *tentative consensus* derivanti dall'assemblaggio di sequenze EST mediante la procedura ParPEST (D'Agostino et al., 2005). Sono allineate, peraltro, lungo il genoma: i) differenti collezioni di sequenze ripetute caratteristiche della famiglia delle Solanaceae (Plant Repeat Databases at Michigan State University, Ouyang and Buell, 2004; the RepBase.13.06, Jurka et al., 2005; SGN tomato UniRepeats (ftp://ftp.sgn.cornell.edu/tomato_genome/repeats/); ii) le sequenze proteiche collezionate dalla banca dati UniProt; iii) le sequenze di RNA appartenenti alla pianta modello *Arabidopsis thaliana*.

Tutte le sequenze BAC di pomodoro, annotate con i dati descritti, sono raccolte e catalogate in una banca dati progettata *ad hoc* e sono visualizzabili tramite il software Genome Browser (Stein et al., 2002).

Con l'intento di rendere pubblico e maggiormente fruibile il servizio non solo ai partecipanti al progetto genoma, ma anche a tutta la comunità scientifica interessata alla genomica della famiglia delle Solanaceae, è stata implementata un'interfaccia grafica per consentire la consultazione dei risultati da noi prodotti accessibile al seguente indirizzo http://biosrv.cab.unina.it/GBrowse.

La costruzione di pagine web dinamiche consente un accesso guidato ai dati genomici catalogati. Nella pagina iniziale (http://biosrv.cab.unina.it/GBrowse) è possibile consultare l'elenco delle sequenze BAC annotate divise per cromosoma di appartenenza, e organizzate in una struttura ad albero di facile navigazione. Per ciascun BAC è possibile visualizzare: i) lo stato di sequenziamento, tramite l'indice HTGS (High Throughput Genome Sequence; http://www.ncbi.nlm.nih.gov/HTGS/examples.html), compreso tra 1 (stadio preliminare) e 3 (stadio finale); ii) l'accesso diretto alla pagina descrittiva nella banca dati GenBank; iii) alcune statistiche generali riguardo le informazioni allineate su ciascun BAC, ad esempio numero di EST e TC, per ogni specie, e calcolo della densità genica (espressa come percentuale di nucleotidi della sequenza genomica che risultano allineati alle EST).

Il contributo alla definizione dei geni modello necessari per l'addestramento dei programmi di predizione genica *ab initio* è stato quello di concepire e implementare un programma, GeneModelEST (D'Agostino et al., 2007b), che automatizzasse la procedura, tradizionalmente "manuale", della valutazione delle regioni geniche (la struttura esoni, introni, CDS, UTR) del genoma del pomodoro a partire dai dati di espressione genica e dell'identificazione di trascritti alternativi in una regione genomica.

ISOL@: UNIONE, INTEGRAZIONE E CONVERGENZA COME ELEMENTI FONDANTI PER MUOVERE I PRIMI PASSI VERSO LA BIOLOGIA DEI SISTEMI

ISOL@ nasce con l'obbiettivo di sviluppare una piattaforma che integri dati da diversi livelli informativi riguardanti la funzionalità cellulare. È stata pertanto progettata come un ambiente computazionale multi-livello e si compone di diverse risorse dati nonché degli strumenti necessari a migliorarne la qualità, estrarne il contenuto informativo e sfruttarne la loro integrazione in modo efficiente.

ISOL@ è al momento costituita da due livelli principali: genoma e trascrittoma ed è predisposta per accogliere dati di proteomica e metabolomica. L'elemento fondante del '*livello genoma*' è rappresentato dalle sequenze genomiche di Solanum lycopersicum prodotte nell'ambito del consorzio internazionale per il sequenziamento del genoma del pomodoro. Quest'ultimo è predisposto per integrare genomi di altre Solanaceae (Di Filippo et al., 2009). Invece, l'elemento base del '*livello trascrittoma*' è costituto dalle collezioni complete di EST da diverse specie di Solanaceae.

È possibile interrogare la piattaforma per ottenere informazioni da ciascun livello mediante punti di accesso indipendenti, consentendo di esplorare in via preliminare il genoma di pomodoro e le annotazioni ottenute o di indagare il trascrittoma esaminando le collezioni EST. Un *cross-link* tra i due livelli garantisce la condivisione delle risorse dati.

La creazione di una piattaforma come ISOL@ non può prescindere dalla necessità di processare i dati grezzi e riconciliarli in modo da aumentarne il contenuto informativo e renderne possibile l'integrazione. Pertanto, sono parte integrante della piattaforma diversi strumenti bioinformatici di base deputati a questo scopo e anche strumenti da noi definiti '*ausiliari*' i quali, sfruttando la sinergia tra i livelli, producono informazione di valore aggiunto. La costruzione di un catalogo attendibile di geni modello mediante la progettazione del software GeneModelEST (D'Agostino et al., 2007b), ad esempio, è proprio dovuta allo sfruttamento della effettiva integrazione dei principali livelli della piattaforma ISOL@.

ISOL@ è concepita come una piattaforma in evoluzione, flessibile per adeguarsi allo sviluppo di nuove tecnologie. Infatti si tratta di una piattaforma che tiene conto della continua crescita dei dati, nonché dei nuovi metodi sperimentali *high-troughput*. A tal fine, è necessaria una continua implementazione di nuovi metodi computazionali in grado di rendere disponibili al meglio informazioni utili alla comunità scientifica interessata in attesa del rilascio di un'annotazione ufficiale da parte dell'iTAG. Una bozza preliminare e incompleta del genoma di pomodoro ne rivela una struttura tipica dei cromosomi.

Oltre a fornire una annotazione del genoma aggiornata quotidianamente, ISOL@ consente di effettuare analisi preliminari sulla struttura del genoma nascente. Infatti, l'integrazione delle collezioni di dati di genomica e trascrittomica nella piattaforma ISOL@ è stata sfruttata per esaminare l'organizzazione strutturale del genoma di pomodoro, andando ad analizzarne la distribuzione dei geni (definiti grazie all'allineamento delle EST lungo il genoma) e della frazione di DNA ripetuto intersperso. Il *genome browser* (http://biosrv.cab.unina.it/GBrowse) consente, infatti, anche la visualizzazione delle sequenze di DNA ripetuto lungo il genoma, grazie all'utilizzo di collezioni di riferimento.

In particolare, sono stati analizzati cinque dei dodici cromosomi per cui è stato ricostruita una ossatura preliminare ('*Minimum Tiling Path*'). In figura 1 sono mostrati gli istogrammi che indicano, per ognuno dei BAC ordinati lungo i cinque cromosomi suddetti, la distribuzione della percentuale di copertura nucleotidica di geni (in nero) e di sequenze ripetute (in grigio). In questi casi, infatti, è possibile seguire la variazione della composizione lungo i cromosomi, andando a distinguere regioni relativamente più ricche in sequenze ripetute e altre più ricche in geni. Grazie anche al supporto di riferimenti sperimentali, si possono individuare porzioni eterocromatiche ed eucromatiche e caratterizzarle mediante l'analisi della sequenza. È possibile riconoscere anche la regione pericentromerica, che è rappresentata dal blocco più ampio di sequenze ricche in DNA ripetuto che si ritrova lungo i cromosomi ed è circoscritta, in figura 1, da rettangoli con linea discontinua. Per i cromosomi più densi di informazioni, ossia quelli per i quali i BAC sequenziati sono molti, è possibile discriminare regioni ricche in DNA ripetuto localizzate esternamente al pericentromero, le quali non sempre si evincono attraverso le comuni analisi sperimentali.

Le analisi, al momento in corso, delle regioni più ricche di sequenze ripetute indicano che esse sono composte principalmente da sequenze retrotrasponibili. La variazione delle varie classi di dette sequenze lungo i cromosomi sembra essere conservata lungo ciascun cromosoma, delineando un andamento che potrebbe risultare tipico dell'intero genoma del pomodoro.

La presenza di una considerevole porzione di DNA ripetuto nelle sequenze genomiche ha messo inoltre in luce la difficoltà nel selezionare in modo affidabile la regione eucromatica, sulla quale si era focalizzato inizialmente il sequenziamento. Ciò ha dato il via all'iniziativa mirata al completamento del sequenziamento del genoma attraverso un approccio WGS, che dovrà di sicuro avvalersi delle informazioni provenienti dall'approccio '*BAC-by BAC*'

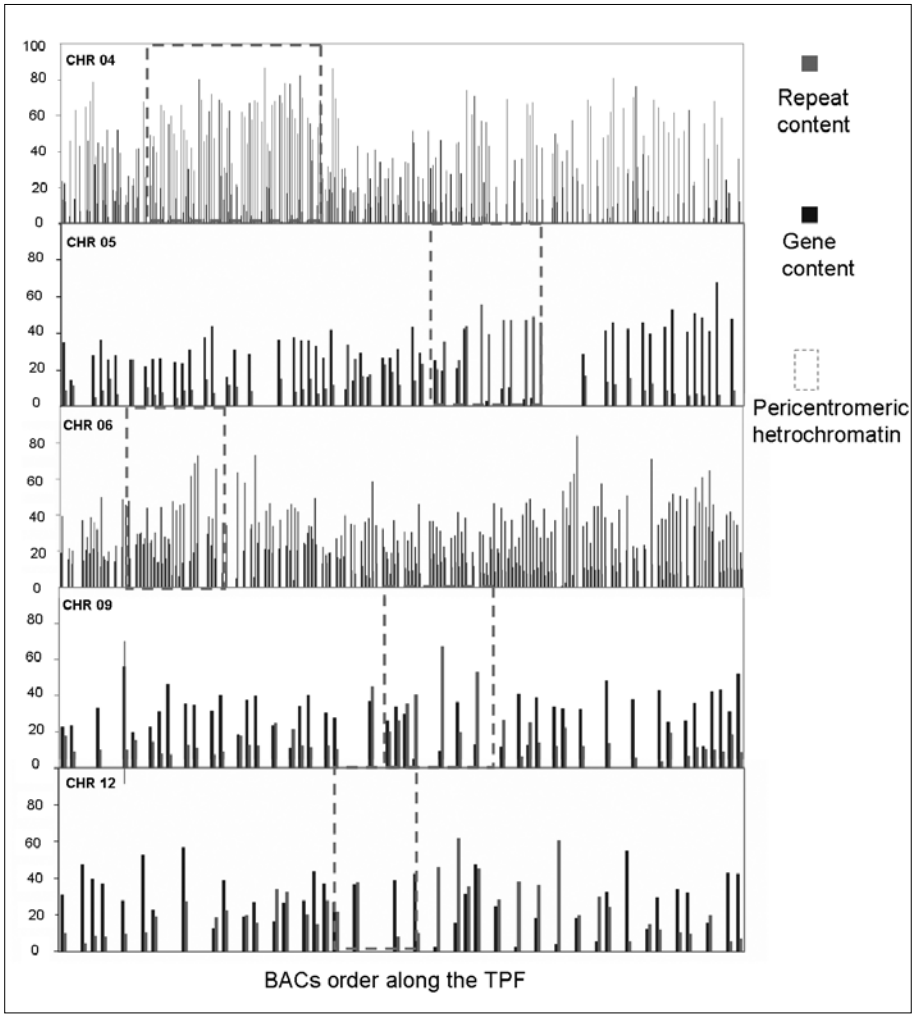

Fig. 1 *Analisi della composizione in geni e sequenze ripetute lungo cinque cromosomi di pomodoro. Ciascun istogramma riporta la percentuale di copertura nucleotidica delle EST (barre nere) e delle sequenze ripetute (barre grigie) per BAC. L'analisi è stata effettuata per i cromosomi di pomodoro (4, 5, 6, 9 e 12) per i quali è stato assemblato un 'Minimum Tiling Path'. I rettangoli selezionati in grigio includono la regione eterocromatica pericentromerica. Essa è stata determinata utilizzando le informazioni disponibili presso il sito del SGN ftp://ftp.sgn.cornell.edu/tomato_genome/seedbacs/eu-hetero-limit-markers.txt, riguardanti la localizzazione dei marcatori molecolari e corrispondenti BAC che individuano le regioni limite tra eucromatina ed eterocromatina*

per una corretta ricostruzione della sequenza di ciascun cromosoma e per assemblare in maniera attendibile le sequenze ripetute.

CONCLUSIONI

Nell'era delle scienze '-omiche' è necessario allestire metodologie bioinformatiche idonee a consentire un'analisi integrata dei dati. Queste devono essere mirate a gestire l'insieme dei dati al fine di definire proprietà emergenti da una visione 'olistica' altrimenti non evidenziabili mediante analisi particolareggiate e relative a specifici aspetti della funzionalità cellulare.

In particolare, nell'ambito della genomica vegetale, la spinta alla caratterizzazione molecolare di differenze e similarità tra specie distinte, ha portato alla organizzazione di consorzi internazionali focalizzati sulla definizione della struttura genomica. Tali progetti, mirati all'analisi di una o più specie, generalmente di interesse agroalimentare, come nel caso del progetto SOL, contribuiscono alla definizione di informazioni utili per applicazioni immediate, come l'identificazione di geni di interesse o di marcatori molecolari utili alla diagnostica e al monitoraggio, nonché al miglioramento genetico, e rappresentano il punto di partenza per approcci ancora più ambiziosi. Ad esempio, la disponibilità della struttura del genoma è fondamentale per la comprensione dei complessi meccanismi che contribuiscono all'affermazione di fenotipi estremamente eterogenei in specie con genomi conservati. Inoltre, è utile per comprendere i meccanismi che conferiscono plasticità al genoma vegetale, e che si riflettono nella sopravvivenza di ibridi e nella tolleranza di fenomeni di temporanea poliploidizzazione. Inoltre, uno studio di genomi che si avvale di tecnologie sempre più avanzate è il punto di partenza per la comprensione di fenomeni di estremo interesse nell'ambito della genomica agraria come ad esempio l'eterosi.

Abbiamo progettato ISOL@ per rispondere alle esigenze del progetto di sequenziamento del genoma di pomodoro, contribuendo agli sforzi internazionali con una piattaforma bioinformatica che consentisse, da subito, l'accesso all'annotazione delle sequenze prodotte. Tale strategia si è rivelata utile a livello intenzionale dato che si è ancora in attesa dell'annotazione ufficiale da parte del gruppo iTAG. Peraltro, il nostro scopo era quello di progettare e testare l'efficienza di un sistema bioinformatico complesso utile per le attività che la ricerca genomica implica.

In particolare il progetto SOL rappresenta una sfida in quanto richiede la progettazione di una piattaforma flessibile ed espandibile, in grado di gestire dati da più di un genoma. A tal fine ISOL@ è già predisposta per poter in-

cludere altri genomi nell'ambito della famiglia delle Solanaceae, come patata e tabacco. Inoltre, una piattaforma multilivello, in grado di integrare e gestire dati da genomica, trascrittomica, proteomica e metabolomica, rappresenta un primo risultato per la risoluzione di problemi tipici della biologia dei sistemi.

Necessaria sarà quindi l'evoluzione della piattaforma verso l'integrazione di dati di fenotipizzazione, di cui al momento esistono già risorse per pomodoro e per altre Solanaceae. Anche in tal caso, la gestione integrata di queste informazioni e di quelle raccolte in una piattaforma come ISOL@, con dati da molteplici livelli della funzionalità cellulare, sarà utile per descrivere adeguatamente le relazione fenotipo-genotipo, ed evidenziare l'associazione tra caratteristiche di interesse ed i processi molecolari che le determinano.

RIASSUNTO

L'esigenza di ampliare le attuali conoscenze sui meccanismi genetici che determinano la variabilità fenotipica e l'adattamento a differenti nicchie ecologiche della famiglia delle Solanaceae ha portato alla realizzazione del Progetto Internazionale Genoma delle Solanaceae (SOL). A tal fine, il pomodoro (*S. lycopersicum*) è stato scelto come sistema modello e attualmente ne è in corso il sequenziamento della regione eucromatica del genoma. I dati provenienti dalle analisi del trascrittoma delle Solanaceae sono stato collezionati in una piattaforma multilivello (ISOL@) per permettere la loro integrazione con le sequenze genomiche di pomodoro al momento disponibili. Sebbene la sequenza della regione eucromatica del genoma di pomodoro sia meno di un quarto del totale, la realizzazione di ISOL@ ha permesso di effettuare una prima esplorazione del genoma di pomodoro analizzando la composizione in geni ed elementi ripetuti e delineando una tipica organizzazione strutturale. ISOL@ è stata predisposta per collezionare altri genomi e per integrare dati ottenuti con differenti tipi di tecnologie, aprendo la prospettiva a una analisi comparativa tra i genomi di Solanaceae. Inoltre, una piattaforma multilivello, in grado di integrare e gestire dati da genomica, trascrittomica, proteomica e metabolomica, rappresenta un primo risultato per la risoluzione di problemi tipici della biologia dei sistemi offrendo nuovi metodi per l'analisi genomica di organismi di interesse agrario.

ABSTRACT

The need to enhance our knowledge on the genetic mechanisms which determine Solanaceae diversification and adaptation to extremely different environments has led scientific efforts to be gathered under the International Solanaceae (SOL) Genome Project. Tomato (*S. lycopersicum*) has been chosen as the reference genome and its sequencing, which is mainly focused on the euchromatin region, is currently ongoing. The Solanaceae transcriptome data have been collected within a multilevel platform (ISOL@), to integrate them with the tomato genome sequences. Although the tomato genome sequences currently available are only one quarter of the total DNA amount, the integration of

genomics and transcriptomics data in ISOL@ permitted a preliminary investigation of the genome in terms of gene and repeat content, revealing a typical design of the genome structure. ISOL@ was set up to include data from other sources, also obtained by next generation sequencing, paving the way for comparative analyses among emerging Solanaceae genomes. In addition, such a platform, built to integrate and manage genomics, transcriptomics, proteomic and metabolomic data, represents a first step to approach a system biology view offering novel methodologies for genome analyses of species of agriculture interest.

REFERENZE

Arabidopsis Genome Initiative (AGI) (2000): *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana*, «Nature», 408, pp. 796-815.

Arumuganathan K., Slattery J.P., Tanksley S.D., Earle E.D. (1991): *Preparation and flow cytometric analysis of metaphase chromosomes of tomato*, «Theor Appl Genet», 82, pp. 101.

Blake J.A Harris M.A. (2002): *The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and ex-pression analysis*, «Curr Protoc Bioinformatics» Chapter 7, Unit 7.2.

Chiusano M.L., D'Agostino N., Traini A., Licciardello C., Raimondo E., Aversano M. (2008): *ISOL@: an Italian SOLAnaceae genomics resource*, «BMC Bioinformatics», 9, Suppl 2, S7.

D'Agostino N., Aversano M., Chiusano M.L. (2005): *ParPEST: a pipeline for EST data analysis based on parallel computing*, «BMC Bioinformatics», 6 Suppl 4, S9.

D'Agostino N. Chiusano M.L., Aversano M. (2007a): *TomatEST database: in silico exploitation of EST data to explore expression patterns in tomato species*, «Nucleic Acids Res.», 35 (Database issue), pp. D901-D905.

D'Agostino N., Traini A., Frusciante L., Chiusano M.L. (2007b): *Gene models from ESTs (GeneModelEST): an application on the Solanum lycopersicum genome*, «BMC Bioinformatics», 8 (Suppl 1), pp. S9.

D'Agostino N., Traini A., Frusciante L., Chiusano M.L. (2009): *SolEST database: a "one-stop shop" approach to the study of Solanaceae transcriptomes*, «BMC Plant Biology», in press.

De Jong J.H. (1998): *High resolution FISH reveals the molecular and chromosomal organization of repetitive sequences in tomato*, «Cytogenet Cell Genet.», 81, pp. 104.

De Jong J.H., Zhong X.B., Fransz P.F., Wennekes-van Eden J., Jacobsen E., Zabel P.E. (2000): *High resolution FISH reveals the molecular and chromosomal organisation of repetitive sequences of individual tomato chromosomes*, in Chromosomes Today, 13 voll., Edited by Olmo E and Redi, CA. Basel Switzerland: Birkha¨ user Verlag, pp. 267-275.

Di Filippo M., Maselli V., Traini A., D'Agostino N., Frusciante L., Chiusano M.L. (2009): *Solanaceae genomics: maybe we can.* 53rd SIGA Annual Congress. (Turin) Italy, September 16-19, 2009.

Foissac S., Bardou P., Moisan A., Cros M.J., Schiex T. (2003): *EUGENE'HOM: a generic similarity-based gene finder using multiple homologous sequences*, «Nucleic Acids Res.», 31(13), pp. 3742-3745.

Foissac S., Gouzy J., Rombauts S., Mathe C., Amselem J., Sterck L., de Peer Y.V.,

ROUZE P., SCHIEX T. (2008): *Genome Annotation in Plants and Fungi: EuGene as a model platform*, «Curr Bioinformatics», 3, pp. 87-97.

FULTON T.M., VAN DER HOEVEN R., EANNETTA N.T., TANKSLEY S.D. (2002): *Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants*, «Plant Cell», 14, pp. 1457-1467.

JURKA J., KOHANY O., ADAM PAVLICEK A., KAPITONOV V.V., JURKA M.V. (2004): *Duplication, coclustering, and selection of human Alu retrotransposons*, «Proc Natl Acad Sci USA», 101(5), pp. 1268-72.

INTERNATIONAL RICE GENOME SEQUENCING PROJECT (2005): *The map-based sequence of the rice genome*, «Nature», 436, pp. 793-800.

MULDER N.J., APWEILER R., ATTWOOD T.K., BAIROCH A., BATEMAN A., BINNS D., BORK P., BUILLARD V., CERUTTI L., COPLEY R., COURCELLE E., DAS U., DAUGHERTY L., DIBLEY M., FINN R., FLEISCHMANN W., GOUGH J., HAFT D., HULO N., HUNTER S., KAHN D., KANAPIN A., KEJARIWAL A., LABARGA A., LANGENDIJK-GENEVAUX P.S., LONSDALE D., LOPEZ R., LETUNIC I., MADERA M., MASLEN J., MCANULLA C., MCDOWALL J., MISTRY J., MITCHELL A., NIKOLSKAYA A.N., ORCHARD S., ORENGO C., PETRYSZAK R., SELENGUT J.D., SIGRIST C.J., THOMAS P.D., VALENTIN F., WILSON D., WU C.H., YEATS C. (2007): *New developments in the InterPro database*, «Nucleic Acids Res.», 35 (Database issue), pp. D224-D228.

MUELLER L.A., SOLOW T.H., TAYLOR N., SKWARECKI B., BUELS R., BINNS J., LIN C., WRIGHT M.H., AHRENS R., WANG Y., HERBST E.V., KEYDER E.R., MENDA N., ZAMIR D., TANKSLEY S.D. (2005a): *The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond*, «Plant Physiology», 138(3), pp. 1310-1317.

MUELLER L.A., TANKSLEY S.D., GIOVANNONI J.J., VAN ECK J., STACK S., CHOI D., KIM B.D., CHEN M., CHENG Z., LI C., LING H., XUE Y., SEYMOUR G., BISHOP G., BRYAN G., SHARMA R., KHURANA J., TYAGI A., CHATTOPADHYAY D., SINGH N.K., STIEKEMA W., LINDHOUT P., JESSE T., LANKHORST R.K., BOUZAYEN M., SHIBATA D., TABATA S., GRANELL A., BOTELLA M.A., GIULIANO G., FRUSCIANTE L., CAUSSE M., ZAMIR D. (2005b): *The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL)*, «Comparative and Functional Genomics», 6, pp. 153-158.

MUELLER L.A., LANKHORST R.K., TANKSLEY S.D., GIOVANNONI J.J., WHITE R., VREBALOV J., FEI Z., VAN ECK J., BUELS R., MILLS A.A., MENDA N., TECLE I.Y., BOMBARELY A., STACK S., ROYER S.M., CHANG S.B., SHEARER L.A., KIM B.D., JO S.H., HUR C.G., CHOI D., LI C.B., ZHAO J., JIANG H., GENG Y., DAI Y., FAN H., CHEN J., LU F., SHI J., SUN S., CHEN J., YANG X., LU C., CHEN M., CHENG Z., LI C., LING H., XUE Y., WANG Y., SEYMOUR G.B., BISHOP G.J., BRYAN G., ROGERS J., SIMS S., BUTCHER S., BUCHAN D., ABBOTT J., BEASLEY H., NICHOLSON C., RIDDLE C., HUMPHRAY S., MCLAREN K., MATHUR S., VYAS S., SOLANKE A.U., KUMAR R., GUPTA V., SHARMA A.K., KHURANA P., KHURANA J.P., TYAGI A., SARITA, CHOWDHURY P., SHRIDHAR S., CHATTOPADHYAY D., PANDIT A., SINGH P., KUMAR A., DIXIT R., SINGH A., PRAVEEN S., DALAL V., YADAV M., GHAZI I.A., GAIKWAD K., SHARMA T.R., MOHAPATRA T., SINGH N.K., SZINAY D., DE JONG H., PETERS S., VAN STAVEREN M., DATEMA E., FIERS M.W.E.J., VAN HAM R.C.H.J., LINDHOUT P., PHILIPPOT M., FRASSE P., REGAD F., ZOUINE M., BOUZAYEN M., ASAMIZU E., SATO S., FUKUOKA H., TABATA S., SHIBATA D., BOTELLA M.A., PEREZ-ALONSO M., FERNANDEZ-PEDROSA V., OSORIO S., MICO A., GRANELL A., ZHANG Z., HE J., HUANG S., DU Y., QU D., LIU L., LIU D., WANG J., YE Z., YANG W., WANG G., VEZZI A., TODESCO S., VALLE G., FALCONE G., PIETRELLA M., GIULIANO G., GRANDILLO S., TRAINI A., D'AGOSTINO N., CHIUSANO M.L., ERCOLANO M.R., BARONE

A., Frusciante L., Schoof H., Jöcker A., Bruggmann R., Spannagl M., Mayer K.X.F., Guigó R., Camara F., Rombauts S., Fawcett J.A., Van de Peer Y., Knapp S., Zamir D. & Stiekema W.A. (2009): *A Snapshot of the Emerging Tomato Genome Sequence*, «The Plant Genome», 2(1), pp. 78-92.

Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N., Apweiler R., Lopez R. (2005): *InterProScan: protein domains identifier*, «Nucleic Acids Res.», 33 (Web Server Issue):W116-W120.

Ouyang S. And Buell R.C. (2004): *The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants*, «Nucleic Acids Res.», 32(Database issue), D360-3.

Peterson D.G., Price H.J., Johnson J.S., Stack S.M. (1996): *DNA content of heterochromatin and euchromatin in tomato (Lycopersicon esculentum) pachytene chromosomes*, «Genome», 39, pp. 77-82.

Stein L.D., Mungall C., Shu S., Caudy M., Mangone M., Day A., Nickerson E., Stajich J.E., Harris T.W., Arva A., Lewis S. (2002): *The generic genome browser: a building block for a model organism system database*, «Genome Res.», 12, pp. 1599-1610.

Van der Hoeven R., Ronning C., Giovannoni J., Gregory Martin G., Tanksley S. (2002): *Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing*, «The Plant Cell», 14, pp. 1441-1456.

Visser R.G.F., Bachem C.W.B., de Boer J.M., Bryan G.J., Chakrabati S.K., Feingold S., Gromadka R., van Ham R.C.H.J., Huang S., Jacobs J.M.E., Kuznetsov B., de Melo P.E., Milbourne D., Orjeda G., Sagredo B., Tang X. (2009): *Sequencing the Potato Genome: Outline and First Results to Come from the Elucidation of the Sequence of the World's Third Most Important Food Crop*, «American Journal of Potato Res.», DOI 10.1007/s12230-009-9097-8.

Yao H., Guo L., Fu Y., Borsuk L.A., Wen T.J., Skibbe D.S., Cui X.Q., Scheffler B.E., Cao J., Emrich S.J., Ashlock D.A., Schnable P.S. (2005): *Evaluation of five ab initio gene prediction programs for the discovery of maize genes*, «Plant Mol Biol.», 57(3), pp. 445-60.