GENETICA, BIOLOGIA MOLECULARĂ ŞI AMELIORAREA

METODOLOGIA DE UTILIZARE A METADATELOR EXPERIENŢELOR MICROARRAY ÎN ELABORAREA IPOTEZELOR ŞTIINŢIFICE

Duca Maria, Leviţchi Alexei, Martea Rodica, Abduşa Daniela, Dragomir Lidia

Laboratorul de Bioinformatică, Centrul universitar Biologie moleculară, Universitatea Academiei de Ştiinţe a Moldovei

Rezumat

S-a efectuat un studiu comparativ al datelor de expresie genică *microarray*, care reprezintă o cale eficientă pentru elaborarea ipotezelor privind funcţia genelor candidate ce stau la baza mecanismelor moleculare a proceselor biologice normale şi/sau a celor patologice. A fost elaborată metodologia de utilizare a metadatelor pentru identificarea experienţelor microarray privind depistarea genelor candidate susceptibile la tratamentul cu gibereline la plante de *Arabidopsis*.

Cuvinte-cheie: gene candidate – gibereline – metadate - microarray

Depus la redacţie 02 noiembrie 2012

Adresa pentru corespondenţă: Dr. Alexei Leviţchi, Laboratorul de Bioinformatică, Centrul universitar Biologie moleculară, Universitatea Academiei de Ştiinţe a Moldovei, str. Academiei 3/2, MD - 2028, Chişinău, R. Moldova, e-mail: *lab.bi.unasm@gmail.com;* tel. (+373) 73 74 15

Introducere

Sistematizarea informaţiei obţinute pentru organismele model reprezintă o importantă sursă de cunoştinţe, cu posibilitate de transfer la diverse specii de interes pentru ştiinţă, dar şi din punct de vedere economic şi strategic, în contextul dezvoltării acestor sfere. Deşi secvenţierea completă a genomului la *Arabidopsis thaliana* a fost finalizată în 2000 [13], pînă în prezent numai 50% dintre cele cca 28 000 gene sunt adnotate funcţional. Se consideră că identificarea şi atribuirea funcţiilor corespunzătoare genelor de interes reprezintă o problemă specifică pentru studiile biologice [10]. Mai mult ca atît, acest lucru implică şi alte aspecte ale cercetării, care necesită o metodologie sistemică în explicarea fenomenelor biologice la nivelul reţelelor proteice sau genice şi a mecanismelor de reglare a acestora.

Actualmente, una dintre căile posibile de analiză a funcţionalităţii genelor este aplicarea studiilor *microarray* [16]. Această tehnologie de performanţă şi înaltă capacitate oferă posibilitatea de estimare a nivelului de expresie a genelor, simultan pentru cîâteva mii de gene, pentru mai multe probe, fiind posibilă analiza complexă a interacţiunilor genice în cadrul reţelelor biologice [8]. Capacitatea de a efectua astfel de experimente a facilitat încheierea secvenţierii genomului uman şi a genomurilor unui şir de organisme model. Studiu comparativ al datelor de expresie microarray reprezintă o cale eficientă pentru elaborarea ipotezelor de explicare a mecanismelor moleculare ce stau la baza proceselor biologice normale şi/sau a celor patologice [11].

Seturile de date microarray sunt disponibile în mai multe resurse electronice (*Gene Expression Omnibus,* GEO (*ncbi.nlm.nih.gov/geo/*); *Gene Expression Atlas,* GEA (*ebi. ac.uk/gxa)*; *Nottingham Arabidopsis Stock Centre Array,* NASCArray (*affymetrix. arabidopsis.info/narrays/experimentbrowse.pl),* etc), care conţin o varietate largă de date referitoare la descrierea completă a informaţiilor privind rezultatele obţinute (*Affymetrix, Illumina,* etc.), condiţiile experienţei, materialul biologic, precum şi alte metadate ce ulterior pot fi utilizate în elaborarea diferitor ipoteze. Sursa principală de date se consideră – GEO, care în majoritatea cazurilor conţine şi datele stocate în *NCBI (National Center for Biotechnology Information).*

NCBI-GEO reprezintă o resursă publică, care cuprinde rezultatele obţinute prin metode contemporane de cercetare [4], cu indicarea protocoalelor şi a tabelelor cu valori de expresie, elaborate conform standardului MIAME (***M***inimum ***I***nformation ***A***bout a ***M***icroarray ***E***xperiment, Conţinutului Minim de Informaţii despre o Experienţă Microarray) [5].

De asemenea, GEO stochează şi date despre expresia genică obtinuţe printr-un şir de tehnici şi metode, precum *tilling array, high throughput sequencing, SAGE, MPSS, RT-PC*R, datele despre ARN necodificator, SNP, metilare, etc. (Tabelul 1).

Se evidenţiază faptul că tipul cel mai răspândit de date sunt cele referitoare la studierea profilurilor de expresie genică bazată pe ARNm, aspecte epigenetice legate de interacţiunea ADN şi proteinele de reglare a activităţii genelor şi a ARN-ului necodificator, precum şi aspectele structurale ale genomului. Dintre tehnicile pe larg utilizate, pot fi enumerate cele bazate pe chip-uri microarray, genome tilling array şi secvenţiere. În acelaşi timp se atestă un număr scăzut de date referitoaare la polimorfismul mononucleotidic (acestea concentrîndu-se în bazele de date specializate) şi analiza expresiei bazată pe profiluri proteice [2].

Tabelul 1. Tipuri de date stocate în NCBI GEO [2, modificat]

Tehnologia / Tipul studiului	array	genome tilling array	high throughput sequencing	SNP array	SAGE	MPSS	RT-PCR	protein array	Total
Profilarea expresiei	17812	303	131	-	206	21	25	-	18498
Profilarea ARN ne-codificator	341	81	233	-	-	-	-	-	655
Profilarea legării/ fixării temporale ale proteinelor pe genome*	70	835	238	-	-	-	-	-	1143
Profilarea variaţiei genomului	309	406	269	-	-	-	-	-	984
Profilarea metilării	46	115	30	-	-	-		-	191
Genotiparea SNP	-	-	-	149	-	-	-	-	149
Profilarea proteinelor	-	-	-	-	-	-	-	31	31
TOTAL	18578	1740	901	149	206	21	25	31	

* ***Genome binding/occupancy profiling***

Integrarea cunoştinţelor obţinute prin diverse tehnologii, în diferite laboratoare, prin aplicarea de metode variate de extragere a probelor de ADN/ARN, induc o serie de erori, care anterior nu puteau fi depistate. În această perspectivă aplicarea tehnicilor *high-throughput* asigură măsurarea semnalului de hibridizare pentru estimarea expresiei genelor studiate. Din acest considerent, la înregistrarea seturilor în resursa *GEO* este importantă indicarea tuturor informaţiilor suplimentare care reprezintă *metadatele,* importante pentru analizele *explorative* [12].

În studiul expresiei genelor metadatele sunt indispensabile pentru identificarea de date specifice, descărcarea seturilor de date de interes şi interpretarea rezultatelor, contribuind la obţinerea cunoştinţelor ce conduc la generarea de rezultate şi ipoteze. Utilizarea metadatelor reprezintă unul dintre obiectele de studiu în *bioinformatică*, în contextul în care aceste metode conduc la avansarea conceptelor biologice. Toate aceste informaţii pot fi considerate şi valorificate în momentul cînd se cunoaşte structura lor, instrumentele prin care ele pot fi integrate şi analizate şi în cazul în care se aplică o prelucrare statistică corespunzătoare pentru selectarea rezultatelor.

Scopul prezentei lucrări a constat în elaborarea metodologiei de analiză a metadatelor experienţelor microarray şi determinarea etapelor prealabile de lucru cu seturile de datele microarray.

Materiale şi metode

În cercetări au fost utilizate profilurile de expresie a genelor la *Arabidopsis thaliana,* ca rezultat al tratamentului plantelor cu giberelină, în perspectiva identificării ulterioare a factorilor ereditari susceptibili la acţiunea acestui fitohormon.

Pentru extragerea şi analiza datelor a fost utilizat mediul de programare R [1, 9]. Seturile de date microarray au fost descărcate cu ajutorul pachetului *GEOquery* [6]. Experienţa este bazată pe platformă de tip *Affymetrix*. Pachetul de adnotare folosit pentru chipul respectiv este *ATH1-121501 Affymetrix Arabidopsis ATH1 Genome Array,* stocat

pe Bioconductor v.2.9 (*bioconductor.org)* [7], „ath1121501.db" versiunea recentă 2.6.3 (*bioconductor.org/packages/release/data/annotation/html/ath 1121501.db.html),* care include date despre sondele moleculare a genelor, poziţia lor pe cromosom, simbolul şi denumirea completă, proteina corespunzătoare, apartenenţa la căile metabolice şi ontologia genică etc.

Suplimentar, pentru a completa informaţiile obţinute privind sondele de pe chip, a fost utilizat instrumentul *NetAffy Analysis Center* (Affymetrix) (*affymetrix.com/ analysis/index.affx*). Cu ajutorul acestei surse au fost extrase informaţii privitor la secvenţele sondei şi a fragmentului cu care aceasta poate hibridiza şi poziţia acestuia pe cromosom, sensul catenei, asocierea cu caractere cantitative, codurile de adnotare suplimentare din alte baze de date (de ex., de pe portalul ExPaSy), care nu sunt prezente în „*ath1121501.db*".

Din considerentul că fişierele cu date microarray sunt voluminoase, iar conexiunea la Internet uneori este instabilă şi cu viteză joasă, rapiditatea analizei este de obicei într-o situaţie critică. Astfel fişierele au fost identificate şi stocate local pe calculatorul de lucru.

Rezultate şi discuţii

În baza de date NCBI GEO în categoria *DataSets* sunt stocate următoarele categorii de date (**xxx** – reprezintă un cod unic atribuit) [3]:

a) înregistrare **GPLxxx** – reprezintă o ***Platforma*** cu mai multe seturi de probe, care corespunde anumitui tip de chip produs de un anumit producător. În această înregistrare se conţin următoarele categorii de informaţii: *Status, Title, Technology type, Distribution, Organism, Manufacturer, Manufacture protocol, Description, Web link , Submission date, Last update date, Organization, E-mail(s), Phone, URL, Street address, Samples, GSM, Data table header descriptions, ID, GB_ACC, SPOT_ID, Species Scientific Name, Annotation Date, Sequence Type, Sequence Source, Target Description, Public ID, Gene Title, Gene Symbol, ENTREZ_GENE_ID, RefSeq Transcript ID, Gene Ontology Biological Process, Gene Ontology Cellular Component, Gene Ontology Molecular Function, Supplementary data files not provided.*

b) înregistrare **GSExxx** - ***Serie*** combină mai multe probe ce au fost cercetate conform unui scop comun de studiu, aici se prezintă următoarele informaţii: *Status, Title, Organism, Experiment type, Summary, Genotypes, Time points, Keywords, Contributor(s), Citation(s), Submission date, Last update date, Contact name, E-mail(s), Phone, Organization name, Department, Lab, Address, Platforms, Samples, Download family, SOFT formatted family file(s), MINiML formatted family file(s), Series Matrix File(s), Supplementary data files not provided.*

c) înregistrare **GSMxxx** - ***Sample*** este echivalentă cu o probă biologică şi se consideră a fi un studiu individual, aici se poate găsi sumarul experienţei, cu următoarele categorii de informaţii: *Status, Title, Sample type, Source Name, Organism, Extracted molecule, Description, Time point, Alleles, Keywords, Submission date, Last update date, Contact name, Organization name, AddresS, Platform ID, Series, GSE, Data table header descriptions, Supplementary data files not provided*.

Studiul informaţiilor din acestă resursă au demonstrat o varietate mare de seturi de rezultate microarray. Conform datelor statistice prezentate pe portalul GEO (*ncbi.nlm. nih.gov/geo/*), resursa de faţă stochează peste 30 000 de studii cu cca 800 000 de probe

analizate pentru 1 000 de organisme, obţinute în 8 000 diferite laboratoare de cercetare cu circa 10 000 referinţe bibliografice [3]. Acest volum impunător de cunoştinţe oferă o oportunitate deosebită pentru analiza datelor la nivel de gene individuale sau de experienţe complexe [2].

Pentru facilitarea analizei explorative şi valorificarea eficientă a datelor identificate am elaborat un tabel analitic, în care am inclus informaţii specifice referitoare la fiecare experienţă selectată pentru *Arabidopsis thaliana.* Analiza metadatelor pentru aceste experienţe a permis să stabilim prezenţa informaţiei complete la nivel de *specie, condiţie, durată, ţesut* etc., acestea au fost stocate local într-un tabel analitic şi sunt considerate cuvinte cheie în evidenţierea ulterioară a seturilor microarray de interes.

Ulterior s-au analizat seturile de date propuse de toate tipurile de tehnologii (producători) microarray (platforme) disponibile. Au fost identificate 50 *platforme* microarray, ce corespund chip-urilor de la trei tipuri de producători: **Affymetrix**, **Agilent Technologies** şi **NimbleGen**. Pentru fiecare dintre platformele microarray identificată s-a calculat numărul de seturi de date (GSE) şi numărul de probe (GSM) analizate. Investigările referitoare la tipurile de platforme stocate în GEO privind *Arabidopsis thaliana* au permis evidenţierea faptului că dintre tipurile cercetate, cele mai multe informaţii (29 platforme – 58 %) sunt obţinute în baza chip-urilor *Affymetrix*, 11 (21%) platforme revine producătorului Agilent Technologies, iar 10 platforme - NimbleGen (Figura 1):

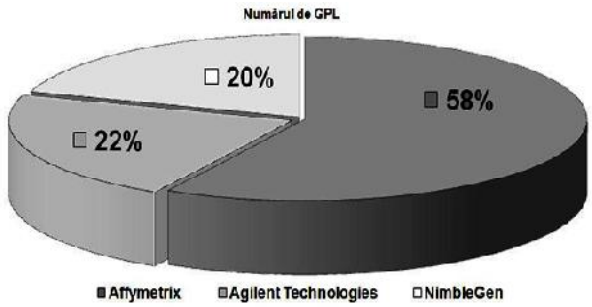

Figura1. Tipurile de platforme microarray pentru *Arabidopsis*

Studiul comparativ al numărului total de seturi de date microarray pentru platformele identificate a evidenţiat că 93,7%, din cele 714 serii, corespund platformelor produse de *Affymetrix* (Figura 2).

De asemenea, a fost evidenţiată platforma **GPL198** - *ATH1-121501 Affymetrix Arabidopsis ATH1 Genome Array,* aceasta conţine un număr record de experienţe 548 GSE-uri, cuprinzînd 6 992 probe testate.

Ca rezultat, în cadrul studiului au fost analizate datele produse în baza chip-urilor microarray *Affymetrix*. Seturile extrase s-au inclus în tabelul analitic elaborat în persectiva stocării datelor de expresie într-o formă accesibilă. În continuare a urmat extragerea propriu-zisă a datelor din *GEO* şi au fost incluse în tabel informaţiile corespunzătoare, după cum urmează: *codul_GSE, numele_GSE, numele_GPL, codul_ GPL, codul_experientei, sumarul_experientei, cuvintele_cheie, referinta, codul_GSM, numele_GSM, codul_GSM, sumarul_GSM şi bioconductor_annotation.*

Astfel, odată fiind determinată structura şi componenţa tabelului analitic, se stabilesc etapele ulterioare de lucru (Figura 3).

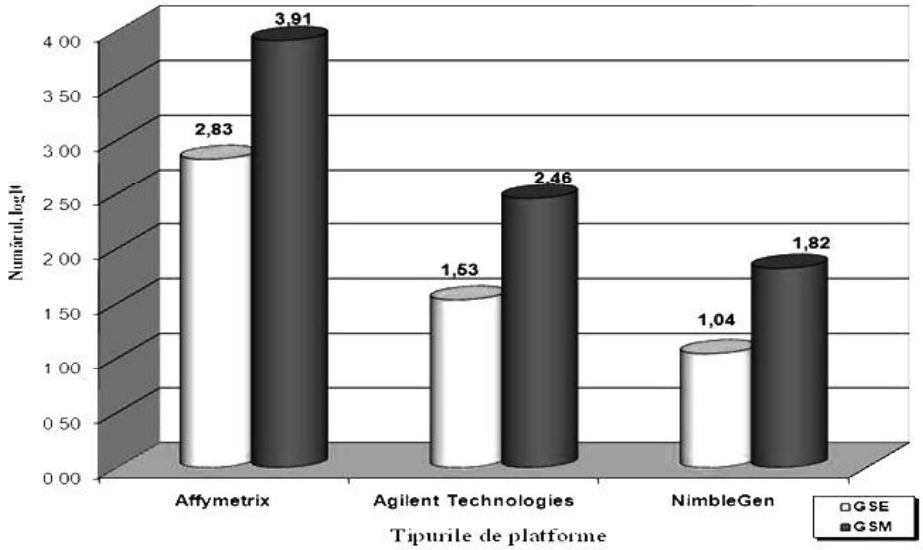

Figura 2. Repartizarea GSE şi GSM pentru Arabidopsis

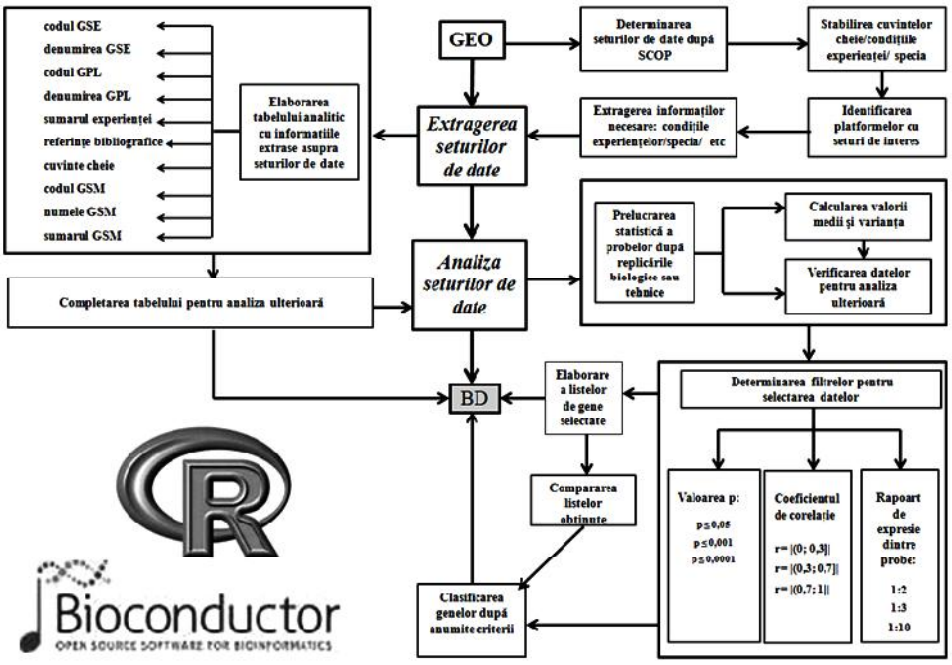

Figura 3. Metodologia de lucru elaborată pentru analiza datelor microarray.

În conformitate cu aceste două criterii de filtrare au fost identificate şi extrase informaţii pentru ***147 seturi microarray,*** repartizate în ***8 serii*** ce se referă la **18 platforme,** care descriu experienţe referitoare la tratamentul cu giberelină (5 GSE), auxină, etilenă, citochinină şi acid abscizic.

Filtrarea ulterioară, efectuată în corespundere cu scopul propus, ne-a permis să determinăm 5 seturi de date (GSE8739, GSE8785, GSE8741, GSE18985, GSE7353),

care pun în evidenţă experienţe de tratament cu giberelină. Dintre acestea pentru cercetare a fost selectat setul **GSE8739**, caracterizat prin următoarele metadate [15]:

1. Denumire *- Early gibberellin responses in Arabidopsis*

2. Organism *- Arabidopsis thaliana*

3. Tipul experienţei *- Expression profiling by array*

4. Sumar *- The aim is to identify early gibberellin responsive genes in a gibberellin deficient strain such as ga1-3. Such genes are likely regulated by DELLA proteins which are master gibberellin repressors. DELLA proteins are rapidly degraded after gibberellin treatment, but their direct target genes still need to be elucidated.//*

5. Design-ul experienţei *- A set of 4 biological replicates was generated for each treatment. Arabidopsis seedlings were treated with water or 2 µM* GA_4 *and whole shoots collected after 1h. A comparison of water vs.* GA_4 *treated samples should render a list of early gibberellin responsive genes.*

GSE8739 a fost elaborat pentru determinarea genelor cu răspuns precoce la tratamentul giberelinic, reglate de proteinele DELLA, ce degradează rapid după tratament. În acest context, datele pot fi utilizate pentru identificarea ulterioară a genelor candidate, susceptibile la acţiunea acestui fitohormon.

Setul studiat include 8 probe biologice, reprezentate în 4 repetiţii tehnice şi două biologice, dependente de condiţia *tratamentul cu apa* şi *tratamentul cu* GA_4*:*

1. GSM216888 - ga1-3_shoots_1h_water_repl1
2. GSM216893 - ga1-3_shoots_1h_water_repl2
3. GSM216895 - ga1-3_shoots_1h_water_repl3
4. GSM216896 - ga1-3_shoots_1h_water_repl4
5. GSM216899 - ga1-3_shoots_1h_GA4_repl1
6. GSM216901 - ga1-3_shoots_1h_GA4_repl2
7. GSM216904 - ga1-3_shoots_1h_GA4_repl3
8. GSM216906 - ga1-3_shoots_1h_GA4_repl4

Mutantul utilizat *ga1-3* se caracterizează prin mutaţie în gena *GA1*, care codifică enzima *ent*-cauren sintaza A, care stopează sinteza formelor active a GA şi respectiv determină timpul înfloririi plantei, similar cu tipul sălbatic [14], iar tratamentul extern cu GA restabileşte complet tipul sălbatic. Astfel, experienţa dată se propune a fi utilizată pentru identificarea prin analiza explorativă a *genelor candidate care îşi modifică expresia la tratamentul cu giberelina* GA_4, şi confirmarea ulterioară a acestei ipoteze prin metode de laborator.

Concluzii

Prin extragerea şi analiza informaţiilor de interes din baza de date GEO a fost elaborată o metodologie de cercetare a metadatelor referitoare la experienţele microarray. S-a creat un tabel analitic care cuprinde descrierea seturilor de interes.

A fost identificat setul de interes, GSE8739, care poate fi utilizat pentru elaborarea ipotezei de cercetare privind *genele candidate care îşi modifică expresia la tratamentul cu giberelie* GA_4*.*

Cercetările au fost realizate în cadrul proiectului instituţional 11.817.04.19F ***Aspecte funcţionale şi genetico - moleculare ale genomului la floarea-soarelui (Helianthus annuus L.),*** etapa *Identificarea genelor cu expresia indusă de semnalul giberelinic prin analiza profilurilor microarray existente în bazele de date.*

Bibiliografie

1. Arppe A., Milin R. P., Baayen H. Package Naive Discriminative Learning Version 0.1.1, 2011, http://cran.r-project.org/web/packages/ndl/ndl.pdf.

2. Barrett T. NCBI GEO: archive for functional genomics data sets - 10 years on. // Nucleic Acids Research, 2011, 39:1005–1010.

3. Barrett T. NCBI GEO: mining millions expression profiles- database and tools. // Nucleic Acids Research, 2005, 33: 562–566.

4. Barrett T. NCBI GEO: mining tens of millions of expression profiles - database and tools update. // Nucleic Acids Research, 2007, 35:760–765.

5. Brazma A., Hingamp P., Quackenbush J., et al. Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. // Nat. Genet., 2001, 29, (4):365–371.

6. *Davis S., Meltzer P. S.* GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. // Bioinformatics, 2007, 23, (14):1846–1847.

7. Gentleman R. C., Carey V. J., Bates D. M. Bioconductor: open software development for computational biology and bioinformatics. // Genome Biol, 2004, 5(10):80.

8. Giorgi F. M., Bolger A. M., Lohse M., et al. Algorithm-driven Artifacts in median polish summarization of Microarray data. // BMC Bioinformatics, 2010, 11:553.

9. *R Development Core Team.* R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Version 2.11.1 (2010-05-31), ISBN 3-900051-07-0.

10. Rhee S. Bioinformatic resources, challenges, and opportunities using Arabidopsis as a model organism in a postgenomic era. // Plant Physiology, 2000, 124(4):1460–1464.

11. Ron E., Barrett T. NCBI GEO standards and services for microarray data. // Nat. Biotechnol., 2006, 24(12):1471–1472.

12. Tukey J. W. We need both exploratory and confirmatory. // The American Statistician, 1980, 34(1):23–25.

13. Walbot V. A green chapter in the book of life. // Nature, 2000, vol. 408, p. 794–795.

14. Wilson R., Heckman J., Somerville C. Gibberellin is required for flowering in Arabidopsis thaliana under short days. // Plant Physiol., 1992, 100:403–408.

15. Zentella R., Zhang Z. L, Park M., et al. Global analysis of DELLA direct targets in early gibberellin signaling in Arabidopsis. // Plant Cell, 2007, 19(10):3037–3057.

16. Zhang A. Advanced analysis of gene expression microarray data. // World Scientific Publishing Co. Pte. Ltd., 2006, 4:51–82.

PRIMERII OMOLOGI TRANSPOZONULUI *ACTIVATOR* ÎN EVIDENŢIEREA POLIMORFISMULUI MOLECULAR AL GENOMURILOR VEGETALE

Paşa Lilia

Institutul de Genetică şi Fiziologie a Plantelor al Academiei de Ştiinţe a Moldovei

Rezumat

În lucrare, este arătată posibilitatea utilizării primerilor omologi elementului transpozabil *Activator* în evidenţierea polimorfismului molecular al speciilor de plante distanţate filogenetic – *Asparagus officinalis* L. (sparanghel), *Allium cepa* L. (ceapă), *Magnolia sp.* (magnolia), *Buxus sempervirens* L. (cimişir), *Anethum graveolens* L. (mărar).

Cuvinte cheie: element transpozabil – *Activator* - polimorfism molecular - *Asparagus*