II. Strategii de analiză a datelor microarray în cercetarea biomedicală

Scopul final al oricărui experiment microarray, independent de tehnologia utilizată, este de a furniza o măsură a abundenței relative a fiecărei gene în probele analizate. Analiza de date este considerată cea mai importantă etapă în bioinformatica microarray. Tipurile de analize folosite în microarray depind de abordările studiului respectiv. Există trei categorii majore de studii microarray ale expresiei genice, în funcție de necunoscutele biologice pe care încercă să le elucideze.

Studiile comparative (class comparision) implică identificarea genelor diferit exprimate între două sau mai multe grupuri luate în studiu. De exemplu, identificarea genelor diferit exprimate între subiecți sănătoși și subiecți cu diferite afecțiuni[1] , pacienți tratați vs. netratați[2], pacienți cu supraviețuire pe termen lung vs. pacienți cu supraviețuire pe termen scurt[3] etc.

Studii de predicție (class prediction) implică clasificarea probelor pe baza profilului de expresie genică a acestora (ex. pe bază profilului de expresie genica din sânge se poate previziona dacă un pacient va dezvolta sau nu o anumită formă de cancer). Pornind de la seturi reprezentative de probe cu apartenență de clasă cunoscută (ex. subiecți sănătoși și pacienți cu cancer de prostată) sunt dezvoltate modele matematice de predicție, capabile să analizeze profilul de expresie genică a unei probe și să stabilească apartenența sa la o anumită clasă. Aceste modele matematice sunt apoi aplicate pentru a evalua probabilitatea altor pacienților de a dezvolta cancer, pacienți care nu au fost incluși în construcția modelului.

Studii de clasificare (class discovery) implică analiza unui set dat de profiluri de expresie genică pentru a descoperi subgrupuri cu caracteristici comune. De ex. vor fi analizate profilurile de expresie genică a unui grup de pacienți cu un anumit tip de cancer pentru a identifica subgrupuri de pacienți cu profiluri de expresie similare în cadrul acestui grup. Aceste studii au scopul de a genera o taxonomie moleculară a bolilor. Cu alte cuvinte, câte subtipuri moleculare de cancer de prostată pot fi identificate în rândul pacienților cu această patologie?

În studiile de predicție și clasificare, stabilirea diferențelor între nivelurile de expresie dintre grupuri (ex. cancer vs. normal) este adesea urmată de stabilirea profilului funcțional, *functional profiling*,[4] pentru a obține o perspectivă a proceselor biologice care sunt alterate în patologia studiată.

II.1. Studii comparative

Studiile comparative sunt întreprinse pentru a compara profilurile transcriptomice a două sau mai multe grupuri de pacienți. Obținerea unor concluzii valide ține cont de alegerea corectă a unui design experimental, de formularea explicită a ipotezelor precum și de o dimensiune adecvată a eșantionului de studiu ales.

II.1.1. Design-ul experimentului

Designul experimentului este unul din cele mai importante subiecte în bioinformatica microarray. Un design experimental bun permite obținerea de informații maxime cu minim efort. Design-urile posibile ale studiilor microarray sunt multiple, opțiunea pentru unul sau altul dintre ele trebuie făcută în funcție de problemele biologice cărora le sunt adresate. Câteva dintre cele mai frecvente design-uri sunt detaliate în cele ce urmează.

Design-ul referință este cel mai simplu design experimental și este utilizat pentru array-urile bicolore. Probele provenite de la fiecare pacient sunt marcate cu același fluorocrom (Cy5) și hibridizate împreună cu o probă de referință (Cy3), comună pentru toate array-urile (fig. 2.1). Prin urmare, va exista un array pentru fiecare probă de interes (pacient). Dezavantajul acestui experiment este că proba de referință, cea mai puțin importantă în cadrul experimentului, va fi măsurată de mai multe ori, în timp ce probele de interes (pacienții) vor fi măsurate doar o singură dată[5]. Avantajele acestui experiment sunt reprezentate de simplitate și flexibilitate. Dacă studiul este extins ulterior, în noua analiză pot fi incluse toate array-urile, atât cele noi cât și cele vechi.

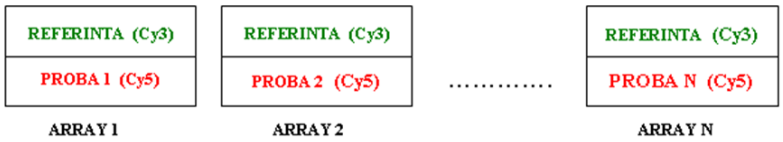


Fig. 2.1. Design referință

Design-ul dye-swap se referă la marcarea dublă, inversată, a două array-uri. Dacă pe un array o probă (pacient) este marcată cu Cy5, în celălalt array proba va fi marcată cu Cy3 (fig. 2.2.). Acest design reduce erorile apărute în procesul de marcare fluorescentă, însă necesită un număr mai mare de array-uri.

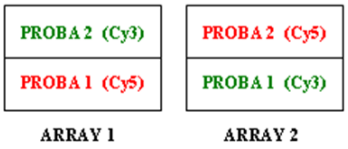


Fig. 2.2. Design dye-swap

Design-ul în buclă (loop) este de asemenea destinat array-urilor bicolore. În acest caz, fiecare probă este hibridizată de două ori, o dată cu fiecare fluorocrom, pe array-uri diferite[6] (fig. 2.3.). Acest design are o putere statistica îmbunătățită, acest aspect fiind, de cele mai multe ori, esențial în rezultatele și concluziile finale. Dezavantajul acestui design constă în complexitatea analizelor și dificultatea de adăugare ulterioară a altor probe.

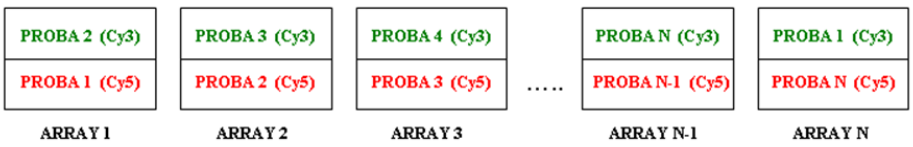


Fig. 2.3. Design în buclă (loop)

În experimentele monocolore, fiecare probă biologică este hibridizată pe un array permițând astfel măsurători independente. Acest design este convenabil deoarece datele pot fi ușor analizate, însă nu se pretează oricărui experiment microarray.

Pentru exemplificare, sunt prezentate două studii cu selecția celui mai potrivit design pentru fiecare studiu în parte.

Studiu 1. Scopul studiului este de a identifica genele supra și sub exprimate în cancerul de prostată față de țesutul normal. În acest scop, au fost recoltate de la pacienți cu cancer de prostată, atât probe de țesut tumoral (TT) cât și probe de țesut normal (TN) pentru a fi hibridizate pe array-uri. În acest caz au fost imaginate două design-uri posibile:

1. cu array-uri bicolore:

a) fiecare TN și TT va fi marcat cu Cy5 și va fi hibridizat pe câte o lamă împreună cu probă de referință marcată cu Cy3 (design referință).
b) fiecare TN (Cy3) și TT (Cy5) de la un pacient vor fi hibridizate împreună pe aceeași lamă.
c) fiecare TN și TT de la un pacient vor fi hibridizate împreună pe două lame, având însă marcaje diferite pe fiecare lamă. Pe o lamă TN va fi marcat cu Cy3 iar TT va fi marcat cu Cy5, pe cealaltă lamă TN va fi marcat cu Cy5 iar TT cu Cy3 (design dye-swap).
d) TN și TT de la același pacient vor fi hibridizate pe aceeași lamă, pe jumătate din numărul lamelor TN fiind marcat cu Cy3 iar TT cu Cy5, pe cealaltă jumătate de lame marcajul făcându-se invers: TN va fi marcat cu Cy5 iar TT cu Cy3.

2. cu array-uri monocolore, fiecare probă este hibridizată pe un array diferit.

Probele luate în studiu în acest experiment sunt pereche (țesut normal și țesut tumoral de la același pacient). Pentru a identifica genele supra și sub exprimate între cele două grupuri, design-ul trebuie să țină cont de această relație dintre probe. Din acest punct de vedere, design-ul cel mai indicat, în acest caz, ar fi 1.b.

Studiu 2. Scopul studiului este de a identifica subgrupuri de pacienți cu carcinom hepato-celular relevante clinic, folosind o analiză de clusterizare și un model de clasificare pentru a diferenția subgrupurile. Probele provenite de la pacienți cu carcinom hepato-celular sunt hibridizate pe aray-uri. În această situație, design-uri posibile ar putea fi:

1. array-uri de tip monocolor, în acest caz fiecare probă este hibridizată pe un array diferit.

2. array-uri de tip bicolor:

a. probele provenite de la jumătate din pacienți sunt marcate cu Cy3, iar probele provenite de la restul pacienților sunt marcate cu Cy5. Două câte două probe diferit marcate sunt hibridizate împreună pe un array.
b. probele provenite de la pacienți sunt marcate cu Cy3 și hibridizate împreună cu o referință comună marcată cu Cy5 (design referință).

În acest studiu design-ul 2a este nepotrivit deoarece pentru a aplică metodele de clusterizare și de clasificare este nevoie că probele să poată fi comparate pe picior de egalitate. Este dificil de comparat două probe hibridizate pe array-uri diferite și marcate cu fluorocromi diferiți.

Cele mai potrivite designuri ar fi, design-ul referință 2b și designul 1. Probele pot fi comparate pe picior de egalitate prin normalizarea relativă la referință (design-ul 2b) sau prin normalizarea între array-uri (design-ul 1).

Un alt aspect important în stabilirea unui design reușit al experimentului microarray, indiferent de tehnologia studiată, este folosirea replicatelor. Există două tipuri de replicate.

Replicatele tehnice implică repetarea unui anumit proces din cadrul experimentului microarray. De exemplu, repetarea procesului de hibridizare pe array a unei anumite probe, replicarea marcării probelor (ex. dye-swap). Replicatele tehnice nu pot fi considerate experimente diferite. În general, se face o medie a acestora care furnizează o singură măsurătoare pentru o probă dată. Replicatele tehnice sunt utile pentru îmbunătățirea calității experimentului, folosirea lor scade variabilitatea experimentală.

Replicatele biologice implică utilizarea în studiu a mai multor probe biologice independente, pentru fiecare categorie de interes în parte. De ex., fiecare pacient cu o anumită patologie este o replicată biologică pentru patologia respectivă. Aceste replicate sunt cele mai importante deoarece cresc tăria

statistica a experimentului. Includerea unui număr suficient de replicate oferă garanția că efectele observate pot fi generalizate.

În unele experimente microarray sunt hibridizate pe lamă amestecuri de probe de la subiecți aparținând unei anumite categorii de interes. Acest amestec este considerat ca o probă singulară, deoarece în acest proces este eliminată informația legată de variabilitatea dintre indivizi. Aceste amestecuri pot fi folosite pentru probele de referință utilizate în anumite design-uri, însă trebuie evitate în experimentele în care această variabilitate este importantă.

II.1.2. Testarea ipotezelor statistice

În studiile comparative, scopul este de a identifica genele diferit exprimate între două sau mai multe grupuri luate în studiu pe baza ipotezei nule și a ipotezei alternative. În aceste studii *ipoteza alternativă* susține că nivelul de expresie al unei gene este diferit între grupurile luate în studiu, iar *ipoteza nulă* neagă ipoteza alternativă și susține inexistența efectelor biologice, adică o anumită genă de pe array nu este diferit exprimată în cele două grupuri. Dacă ipoteză nulă este adevărată atunci diferența între nivelele de expresie nu se datorează efectului biologic luat în studiu, ci doar unei variabilități existente între probe sau a unor erori de măsurare. Dacă ipoteză nulă este respinsă ipoteză alternativă este confirmată.

Folosind modele probabilistice, se calculează un parametru statistic (ex. *t*) care se compară cu un prag calculat cu modelul statistic respectiv, obținându-se un nivel de semnificație *p*. Cu cât acest *p* este mai mic, probabilitatea ca rezultatul obținut să se datoreze hazardului este mai mică, deci rezultatul este mai semnificativ.

Testarea ipotezelor statistice introduce două tipuri de erori: *eroarea de tip I* și *eroarea de tip II*. Semnificația erorilor de tip I și II depinde de felul în care este definită ipoteză nulă.

Eroarea de tip I respinge ipoteză nulă atunci când ea este adevărată. În cazul studiilor comparative, această s-ar traduce prin identificarea unei anumite gene ca fiind diferit exprimată între cele două grupuri, când de fapt nu este. În consecință, eroarea de tip I duce la obținerea unor rezultate fals pozitive.

Eroarea de tip II acceptă ipoteză nulă când de fapt ea este falsă. În cazul studiilor comparative, acest lucru s-ar traduce prin faptul că nu sunt identificate diferențe între nivelul de expresie al unei anumite gene într-un grup față de celălalt grup, când de fapt, aceste diferențe există. În consecință, eroarea de tip II duce la obținerea unui rezultat fals negativ.

Nivelul de semnificație ($\alpha = p$) se alege la începutul experimentului și reprezintă procentajul din eroarea de tip I pe care investigatorul este pregătit să îl accepte. De ex., un nivel de semnificație de 1% arată că există 1 genă fals pozitivă la fiecare 100 de gene identificate ca diferit exprimate. Altfel spus, pentru un $p = 0{,}01$ există 1% șansa că diferență observată între nivelele de expresie să fie datorată hazardului. Prin convenție statistică, un $p < 0{,}05$ este considerat că oferă un rezultat semnificativ statistic, însă, în microarray, uneori este nevoie de valori ale lui *p* mult mai stringente.

II.1.3. Metode de selecție a genelor diferit exprimate

În studiile timpurii, metoda de detecție a genelor diferit exprimate a fost valoarea *fold change* (Fc). Valoarea Fc a unei anumite gene măsurate în două probe este exprimată prin raportul intensităților corespunzătoare genei în proba de interes vs. proba de referință.

$$Fc = I_i/I_r$$

În cazul particular al marcajului bicolor, în care proba de interes este marcată cu Cy5 (**R**oșu), iar proba de referință este marcată cu Cy3 (**V**erde), valoarea Fc reprezintă valoarea raportului:

$$Fc = R/V$$

Genele selectate ca semnificative sunt cele pentru care -2 < Fc > 2. Cu alte cuvinte, sunt selectate genele al căror nivel de expresie este crescut de 2 ori în proba de interes față de referință, respectiv genele al căror nivel de expresie este scăzut de 2 ori în proba de interes față de referință.[7] [8] În cazul în care se optează pentru o valoare logaritmică a lui Fc, adică

$$Fc = \log_2 \cdot I_i/I_r$$

genele considerate ca semnificative sunt cele pentru care -1 < Fc > 1, ceea ce indică o sub sau a supra exprimare de 2 ori a genelor ($\log_2 2 = 1$ respectiv $\log_2 1/2 = -1$).

Acest prag este însă ales arbitrar, ceea ce determină creșterea riscului de apariție a erorilor de tip I și II. Un alt punct slab al acestui criteriu de selecție este supraestimarea genelor cu nivel scăzut de expresie în proba de referință (când numitorul unui raport tinde spre zero, valoarea raportului tinde să crească) și tendința de a omite modificări minore dar importante ale nivelelor de expresie a anumitor gene. De exemplu, unele gene, cum sunt factorii de transcripție, pot avea efecte biologice importante chiar dacă nivelul lor de expresie nu este modificat de 2 ori. Din aceste considerente, metoda de selecție pe baza valorii Fc nu este recomandată decât în combinație cu alte metode statistice.

Pentru o selecție corespunzătoare a genelor diferit exprimate este indicată testarea ipotezelor statistice[9] [10]. Modelele statistice folosite în acest scop sunt multiple, în funcție de datele ce urmează a fi analizate:

teste parametrice pentru date cu distribuții normale (se distribuie pe o curbă Gauss)

- *testul t pentru probe pereche*. Este folosit în cazul în care probele analizate sunt dependente una de cealaltă (probe recoltate de la același pacient înainte și după tratament, țesut tumoral și normal recoltate de la același pacient etc.).
- *testul t pentru probe nepereche*. Se aplică pentru probe independente una de cealaltă

teste neparametrice pentru date ale căror distribuții nu sunt normale:

- *testul Wilcoxon* similar testului t pentru probe pereche.
- *testul Mann-Whitney* similar testului t pentru probe nepereche.

Testele t compară diferența mediilor nivelelor de expresie a celor două grupuri studiate luând în considerare variabilitatea datelor (deviația standard). Există situații în care valoarea foarte mică (tinde spre 0) a deviației standard este datorată hazardului sau erorilor în determinarea nivelelor de expresie ale genelor. În acest caz, deoarece numitorul expresiei tinde spre 0, valoarea parametrului statistic t devine foarte mare, genele par a fi foarte semnificative iar, în realitate, nu este așa. Pentru a elimina acest risc au fost dezvoltate teste t îmbunătățite: **testul t moderat**[11] și **statistica S**[12]. Diferența între testele t clasice și testele t îmbunătățite constă în faptul că, acestea din urmă, estimează variabilitatea luând în considerare nu doar informația genelor testate, ci și a altor gene care prezintă o variabilitate similară. Acest mecanism este echivalent cu o estimare globală a variațiilor, metoda având cele mai bune rezultate când se lucrează cu un număr mic de array-uri[13].

Pentru compararea a mai mult de două grupuri sunt folosite metode ca ANOVA simplă, ANOVA multifactorială sau modele liniare.

II.1.4. Problema testărilor multiple

O problemă majoră în analiza microarray este testarea simultană a unui număr extrem de mare de ipoteze statistice. Testarea ipotezei statistice implică, așa cum s-a specificat anterior, testarea diferenței de expresie a fiecărei gene de pe array între grupurile luate în studiu. Deoarece array-urile moderne conțin zeci de mii de gene, aceste comparații multiple vor genera sute de rezultate fals pozitive, proporționale cu pragul de semnificație ales. De exemplu, pentru testarea expresiei diferențiale a 10.000 de gene la un prag de semnificație de 1% ($p < 0,01$) vor exista 100 de valori fals pozitive, adică 100 de gene a căror diferență de expresie poate să fie datorată hazardului. Ținând cont de acest aspect, se impune folosirea unor corecții pentru a reduce rezultatele fals pozitive (erorile de tip I). Există multe metode de corecție[14] [15] [16] [17] [18] însă, unele dintre ele nu sunt potrivite pentru analizele microarray, fie pentru că presupun independența variabilelor, fie pentru că sunt considerate prea restrictive.

Independența variabilelor este o condiție care nu poate fi îndeplinită în studiile microarray deoarece între genele și implicit între nivelele de expresie ale acestora, există o interdependență, fiind implicate în mecanisme reglatoare și căi moleculare complexe[19]. Interacțiunile complexe dintre gene determină homeostazia țesuturilor și sunt o caracteristică a proceselor biologice și patologice.

Corecția Bonferoni este o metodă care s-a dovedit a fi extrem de restrictivă. Această corecție consideră semnificative doar acele gene pentru care valoarea ajustată a lui p este $p^* < p/$*numărul de gene testate*.

În cazul exemplului descris mai sus, corecția Bonferoni implică, pentru compararea multiplă a celor 10.000 de gene la un prag $p < 0,01$, ajustarea pentru fiecare genă a valorii lui p^* sub un prag de 0,01/10.000. Cu alte cuvinte, pentru ca rezultatul să fie semnificativ, corecția Bonferoni impune un $p^* < 0,000001$. Această valoare este extrem de restrictivă, și, cu toate că reduce drastic rata de apariție a erorilor de tip I, elimină rezultatele care ar putea fi semnificative din punct de vedere biologic. În prezent, metoda nu este prea agreată de comunitatea științifică deoarece se consideră că este preferabil de acceptat un anumit grad de eroare atât timp cât aceasta poate duce la descoperirea unor rezultate relevante[20].

În ultimii ani, există un tot mai mare consens pentru folosirea ca metodă de corecție în analizele microarray a metodei **FDR (False Discovery Rate)** dezvoltată de Benjamini și Hochberg[21]. Pentru reducerea rezultatelor fals pozitive, această metodă impune o valoare ajustată a lui p, denumită q. Pentru a calcula această valoare, se întocmește o lista în care valorile p ale fiecărei gene sunt sortate descendent. Cea mai mare valoare a lui p rămâne neschimbată, celelalte valori fiind ajustate după formula:

$$q = p^*n/(n-(k-1))$$

unde n este numărul de gene testate, n-(k-1) este poziția în listă (în funcție de valoarea lui p) a genei k. Un $q < 0,01$ (1%) detectează ca fals pozitive 1% din genele considerate statistic semnificative care au trecut restricțiile corecției. După cum se observă, această corecție este mai puțin conservativă decât metoda Bonferoni. Conform datelor din literatură, chiar și o valoare a lui q între 10-20% se poate considera acceptabilă.

II.1.5. Stabilirea dimensiunii lotului de studiu

Determinarea dimensiunii lotului implică stabilirea numărului de replicate biologice necesare pentru obținerea unor rezultate valide. Alegerea numărului de replicate biologice ține cont de: nivelul *fold change* (Fc) stabilit de cercetător, deviația standard a nivelelor de expresie din cadrul grupurilor experimentale, puterea statistică dorită a experimentului. Cu cât diferența dintre nivelele de expresie

ale unei gene este mai mică (Fc mic), cu atât este nevoie de o dimensiune mai mare a lotului de studiu pentru a surprinde această diferență. În cazul în care nivelele de expresie prezintă deviații standard mari, se impune folosirea unui lot de studiu mare pentru a avea certitudinea că diferențele detectate între nivelele de expresie sunt reale. Cu alte cuvinte, valori mici ale pragurilor *fold change* și deviațiile standard mari ale nivelelor de expresie impun folosirea unui număr mare de replicate biologice. Chiar și pentru un număr mic de replicate biologice este posibilă detectarea câtorva gene diferit exprimate între grupurile luate în studiu, însă, în general, în cadrul studiilor comparative, scopul este de a detecta majoritatea genelor diferit exprimate. Astfel, un lot mare de studiu este necesar pentru obținerea unei puteri crescute a testului. În concluzie, pentru a surprinde întreaga gamă de modificări la nivel molecular asociate unei patologii, se impune folosirea unui număr mare de replicate biologice.

Recent, au fost dezvoltate numeroase metode[22] [23] [24] și instrumente de calcul (aplicația online Power Atlas) pentru determinarea numărului optim de replicate biologice într-un experiment microarray. În practică, costul experimentului și numărul probelor clinice disponibile sunt determinante majore în alegerea dimensiunii lotului. Din aceste considerente, mulți cercetători folosesc un număr minim acceptat prin convenție de 5 replicate per grup[25] [26]. Însă, aceste studii cu număr mic de replicate, nu au întotdeauna o putere suficientă de a detecta diferențele între nivelele de expresie și pot fi complet inadecvate în cazul în care aceste nivele de expresie prezintă deviații standard mari.

II.2. Studii de predicție (studii supervizate)

Una dintre cele mai interesante aplicații ale tehnologiei microarray în cercetarea medicală este identificarea unor grupuri de gene (semnături moleculare) care pot fi folosite pentru diagnostic sau prognostic.

Aceste studii, denumite studii de predicție, folosesc grupuri predefinite de probe cu apartenență cunoscută, provenite de la pacienți cu patologii diverse, cu răspuns la terapie cunoscut, de la subiecți sănătoși etc. Scopul studiilor (folosind aceste grupuri predefinite) este orientat spre identificarea unui număr de gene pe baza cărora să se poată face o predicție asupra apartenenței la un grup sau altul al unui anumit individ, pe baza profilului său de expresie genică. Pentru îndeplinirea acestui scop, este necesară construirea unui model de predicție (clasificator) capabil să separe grupurile predefinite pe baza profilului de expresie genică. Un clasificator este un model matematic de tipul:

$$CL = a_1g_1 + a_2g_2 + a_3g_3 + a_4g_4 + \ldots + a_ng_n$$

unde g_1, g_2, ... gn sunt valorile de expresie ale n potențiale gene marker pentru patologia respectivă, a_1, a_2, ... a_n sunt parametrii inițiali necunoscuți dar identificați prin folosirea grupurilor predefinite iar CL este o variabila care indică dacă pacientul este sau nu încadrat în patologia respectivă.

Dezvoltarea unui model predictiv implică trei etape importante: stabilirea metodelor de clasificare capabile să separe grupurile predefinite (identificarea parametrilor a_1, a_2, ... a_n), reducerea dimensiunilor datelor prin limitarea numărului de gene (identificarea g_1, g_2, ... g_n de interes) și validarea modelului.

II.2.1. Metode de clasificare

Modelele predictive consideră că fiecare probă ocupă o locație într-un spațiu *n* dimensional, fiecare axă din acest spațiu reprezentând expresia unei anumite gene. De exemplu, în cazul în care sunt alese 2 gene pentru construirea clasificatorului, proba va fi localizată într-un spațiu bidimensional, dacă sunt alese 3 gene proba va fi localizată într-un spațiu tridimensional ș.a.m.d. În cele mai multe cazuri grupurile de probe sunt doar parțial separabile în acest spațiu, cu unele suprapuneri în zonele de graniță.

Scopul metodelor de clasificare este de a găsi o cale pentru a împărți spațiul, astfel încât fiecare grup de probe să ocupe o regiune diferită și bine definită a spațiului, ca apoi să poată fi stabilită, fără echivoc, apartenența unei noi probe la un grup sau altul. În experimentele microarray nivelul de separabilitate al datelor este determinat de grupul de gene ales. În acest spațiu probele pot fi separate într-un mod liniar (printr-o linie dreaptă) sau neliniar (linia de demarcație nu este o linie dreaptă). Datele microarray sunt de cele mai multe ori neliniare, de aceea, este adesea recomandată atât testarea metodelor liniare cât și neliniare. Pentru o mai ușoară vizualizare a felului în care funcționează metodele de clasificare, va fi folosită o reprezentare bidimensională care implică utilizarea nivelelor de expresie a doar două gene pentru separarea claselor. În cele ce urmează, sunt enumerate și detaliate cele mai importante metodele de clasificare.

k-nearest neighbours (KNN)

Algoritmul de calcul al acestei metode implică următoarele etape:

1. sunt reprezentate probele cu apartenență de clasă cunoscută într-un spațiu ale cărui axe reprezintă nivelele de expresie ale genelor marker (ex. g_1, g_2).
2. proba care se dorește a fi clasificată este reprezentată în acest spațiu pe baza nivelelor sale de expresie corespunzătoare genelor implicate în modelul de predicție.
3. se identifică probele aflate în imediata vecinătate a probei de interes prin măsurarea distanțelor euclidiene între aceasta și vecinii cei mai apropiați. Se definește un parametru *k* (în general se folosește un k = 3) a cărui valoare va indica numărul probelor din vecinătatea probei de interes (probe care vor fi luate în calcul) și un parametru *l* care determină numărul minim al vecinilor ce trebuie să fie în aceeași clasă pentru a putea clasifica proba.
4. apartenența probei la o clasă se va stabili în funcție de apartenența celor mai apropiate probe din vecinătatea acesteia ținând cont de valorile *k* și *l*.

De exemplu, dacă se alege un *k* = 3 și *l* = 3, pentru clasificare sunt luate în considerare cele mai apropiate 3 probe din vecinătate. Proba de interes va fi clasificată dacă toate 3 probele vecine se vor afla în aceeași clasă, în caz contrar, proba va rămâne neclasificată. În cazul exemplului din fig. 2.4, proba este clasificată în categoria cancer de prostată (CP).

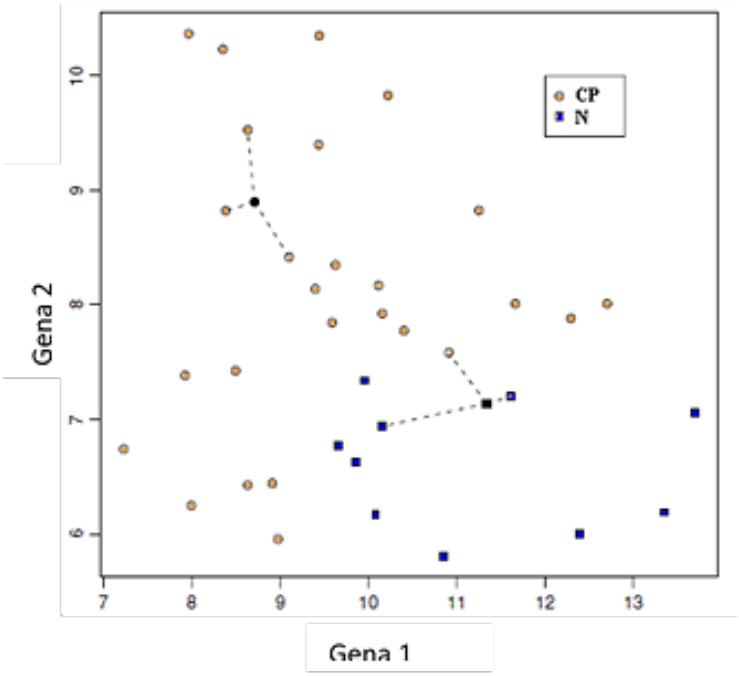


Fig. 2.4. Clasificator KNN

Pentru un *k* = 3 și *l* = 2, din cele 3 probe vecine luate în considerare, este suficient ca doar două să aparțină aceleiași clase pentru ca proba să fie integrată în clasa respectivă. În cazul exemplului din fig.2.4, proba va fi clasificată în categoria normal (N).

Metoda KNN este o metodă de separare neliniară, însă ea nu este considerată o metodă robustă la valori extreme, valori care pot determina rezultate false sau neclasificări ale probelor. Modelul se poate extinde ușor la mai mult de două clase.

Clasificarea centroidă (centroid classification)
În cazul acestei metode algoritmul are următoarele etape:

1. pentru fiecare clasă se calculează centrul de masă al celor două grupe de probe cu apartenență cunoscută (fig. 2.5).
2. se calculează distanța dintre poziția probei ce urmează a fi clasificată și fiecare dintre centrele de masă ale claselor predefinite.
3. se stabilește apartenența probei la clasa cu cel mai apropiat centru de masă.

Acest model este mai simplu din punct de vedere computațional, este foarte rapid și se poate extinde ușor la mai mult de două clase, însă rezultatele pot fi complet eronate dacă datele nu sunt separate liniar. Clasificarea centroidă este, de asemenea, o metodă liniară de separare. În fig. 2.5 se observă că una dintre probele de cancer de prostată (CP) nu este clasificată corect.

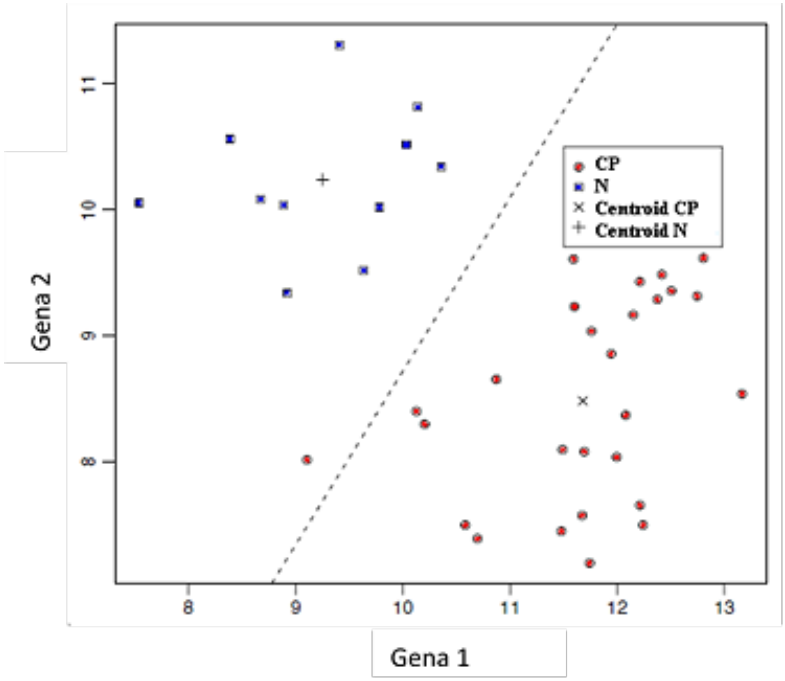


Fig. 2.5. Clasificator centroid

Analiza discriminantului liniar (Linear Discriminant Analisys - LDA)
Această metodă construiește un model statistic pe baza datelor cunoscute. Modelul implică două etape:

1. linia (în cazul în care se lucrează în două dimensiuni) sau hiperplanul (pentru mai mult de 2 dimensiuni) care separă cele 2 clase cunoscute este aleasă în așa fel încât să minimizeze variația intra-grup a datelor de ambele părți ale dreptei/hiperplanului și să maximizeze variația inter-grup a datelor.

2. apartenența de clasă a probei luate în studiu este determinată de poziția acesteia față de dreaptă/hiperplan.

LDA este construită pe un model statistic foarte robust. Deoarece ia în considerare variabilitatea datelor, LDA funcționează mai bine decât clasificarea centroidă. LDA nu poate fi extinsă ușor la mai mult de 2 clase și funcționează numai dacă datele sunt separabile liniar. LDA clasifică probele mult mai corect decât clasificarea centroidă.

Rețele neuronale (Neural Networks - NN)
Rețelele neuronale sunt modele de separare a spațiului într-un mod neliniar și pot fi extinse ușor la mai mult de 2 clase. Funcționarea rețelelor neuronale este similară cu funcționarea creierului. Rețeaua este organizată dintr-o serie de noduri (simulând neuronii) cu intrări și ieșiri. Ieșirile nodale depind de intrările nodale. Rețelele neuronale necesită etape de testare și optimizare și din această cauză este o metodă consumatoare de timp și putere de calcul.

Support Vector Machines (SVM)
Este o metodă similară cu LDA. Hiperplanul este ales în așa fel încât să minimizeze erorile de neclasificare și este delimitat de câteva puncte denumite vectori suport. Nu este o metodă ușor de extins la mai multe clase. Această metodă funcționează cu date neliniare și nu poate fi extinsă ușor la mai mult de 2 clase.

II.2.2. Reducerea dimensiunilor datelor

Experimentele microarray generează o cantitate extrem de mare de date. Construirea unor modele predictive robuste impune reducerea dimensiunilor acestor date prin selecția variabilelor. Această selecție implică identificarea genelor relevante pentru patologia studiată, gene care să poată fi folosite în diferențierea și clasificarea probelor. Nu toate genele dintr-un experiment microarray sunt relevante în studiile de predicție, informațiile conținute de majoritatea genelor nu sunt utile în diferențierea probelor. Selecția setului adecvat de gene este o problemă dificilă, aflată încă în curs de cercetare. Metodele cele mai frecvent folosite și descrise în literatură sunt prezentate mai jos.

Selecția individuală a genelor
Metoda identifică genele care separă cel mai bine clasele în mod individual. Genele sunt selectate folosind testul t, apoi ierarhizate după valoarea lui *p*. Genele cu cele mai mici valori ale lui *p* sunt folosite în modelele de predicție. Această metodă nu este însă foarte robustă în separarea claselor deoarece, deși probele pot fi separate foarte bine de genele individuale, când aceste gene sunt luate împreună, ele pot să nu mai separe probele. De aceea, pentru separarea grupurilor este recomandată utilizarea seturilor de gene.

Selecția seturilor de gene
Această metodă identifică perechile de gene care separă cel mai bine grupurile de probe, după care, aceste gene sunt combinate pentru a obține setul folosit pentru predicție. Se testează toate perechile de gene dar sunt alese doar acelea care separă cel mai bine clasele. Numărul perechilor alese depinde de dimensiunea dorită a setului de gene. În cazul în care se dorește construirea unui predictor cu 20 de gene, se aleg primele 10 perechi care separă cel mai bine clasele. Metoda este superioară selecției individuale a genelor, însă este consumatoare de timp și putere de calcul.

II.2.3. Validarea modelului

În studiile de predicție se pun două probleme:

- dacă metoda de clasificare testată pe grupurile de probe cu apartenență cunoscută este validă și se poate generaliza pentru alte probe neclasificate.
- dacă grupul de probe ales este reprezentativ pentru clasa pe care o reprezintă.

Pentru a rezolva aceste probleme sunt folosite metode de validare:

- metoda loturilor de testare și validare.
- validarea încrucișată.

Validarea prin folosirea loturilor de test și validare este mult mai eficientă, însă funcționează mai bine cu seturi mari de date. Validarea încrucișată funcționează mai bine cu seturi mici de date și este în general folosită în etapa de testare pentru a optimiza parametrii algoritmului.

În cazul metodei de testare/validare, aproximativ 2/3 din probe sunt folosite pentru testarea algoritmului, astfel, acesta este optimizat pentru clasificarea datelor. După testare, algoritmul este validat pe treimea de date rămase pentru a demonstra reușita algoritmului. În general grupul de testare este mai mare decât grupul de validare.

În cazul metodei de validare încrucișată, probele sunt împărțite aleatoriu în *n* grupuri egale sau aproape egale (*n* este fixat de utilizator), apoi algoritmul este rulat de *n* ori folosind *n-1* seturi pentru testarea algoritmului. La fiecare rulare a algoritmului este folosit un alt set pentru validare astfel încât la finalul procesului toate seturile vor fi folosite atât pentru testare cât și pentru validare. Succesul algoritmului este dat de suma clasificărilor corecte la fiecare rulare. Validarea încrucișată prezintă dezavantajul că rezultatele generate nu sunt independente.

II.3. Studii de clasificare (studii nesupervizate)

O altă problemă importantă în studiile de microarray este identificarea subgrupurilor cu trăsături comune în cadrul unui set de probe. Acest aspect implică măsurarea profilurilor de expresie genică provenite de la pacienți cu o anumită patologie cu scopul de a-i clasifica în subgrupuri de pacienți cu profiluri de expresie similare. Interesul medical și biologic în studiile de clasificare este de a înțelege mecanismele care stau la baza apariției anumitor boli, de a identifica diferite stadii ale unei patologii și grupuri de gene care au un comportament similar într-un anumit stadiu al bolii.

În prezent, clusterizarea este una dintre cele mai frecvent folosite tehnici în studiile nesupervizate[27] [28]. Metodele de clusterizare cele mai utilizate sunt: gruparea ierarhică și gruparea iterativă. Scopul metodelor de clusterizare este de a divide genele sau probele în grupuri (clustere) pe bază similarității dintre acestea folosind corelații sau distanțe euclidiene. Aceste analize confirmă că probele cu proprietăți biologice similare tind să aibă un pofil molecular similar, însă profilul molecular al probelor reflectă și eterogenitatea tumorala, utilă în descoperirea subgrupurilor unei patologii.

II.3.1. Clusterizarea ierarhică aglomerată.

Această metodă grupează succesiv genele sau probele într-o structură arborescentă, asemănătoare unui arbore filogenetic, astfel încât, genele/probele cu profiluri similare apar foarte apropiate iar pe măsură ce similaritatea scade, probele sunt tot mai îndepărtate. Această structură arborescentă poartă numele de dendrogramă. Similaritatea dintre gene sau probe este cuantificată prin distanța dintre acestea într-un spațiu multidimensional. Dacă distanța dintre două profiluri (gene sau probe) este zero, profilurile sunt identice. Cu cât această distanță crește, profilurile prezintă o similaritate scăzută. Calcularea distanțelor care indică gradul de similaritate se face fie prin corelație fie prin măsurarea distanțelor[29].

II.3.1.1. Alegerea metodei de calcul a similarității

Măsurarea similarității pe baza coeficientului de corelație

Coeficientul de corelație este dat de următoarea formulă:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

Unde x_i , y_i , $1 < i < n$, sunt două seturi de măsurători corespunzătoare genelor/probelor, considerate ca vectori într-un spațiu multidimensional. Coeficientul de corelație ia valori între -1 și +1. O valoare de -1 indică o corelație negativă perfectă (când o variabilă este mare cealaltă este scăzută). O valoare de +1 indică o corelație pozitivă perfectă, iar valoarea 0 indică absența totală a corelației. Acest coeficient este o măsură a similarității și este convertit în distanțe, după următoarea formulă:

$$d(x, y) = 1 - \text{abs}(r(x, y))$$

unde d(*x, y*) reprezintă distanța dintre profilul *x* și profilul *y*, iar abs(*r*(*x, y*)) reprezintă valoarea absolută a coeficientului de corelație. În cazul în care profilurile (genele/probele) sunt corelate negativ sau pozitiv (*r* = -1 respectiv *r* = +1) distanța dintre ele este *d* = 0, în consecință, probele sunt identice. În cazul în care sunt necorelate (*r* = 0), distanța dintre ele este *d* =1, probele prezintă diferențe destul de mari.

Măsurarea similarității pe baza distantelor euclidiene

Distanțele euclidiene sunt calculate pe baza teoremei lui Pitagora. Pentru un spațiu bidimensional, distanța dintre doua puncte X, Y este data de relația:

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Unde x_1, y_1 sunt abscisele celor două puncte, x_2, y_2 sunt ordonatele acestora. Pentru un spațiu multidimensional (n-dimensional), distanța euclidiană dintre două profiluri X și Y este dată de formula:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Măsurarea similarității pe baza distanțelor euclidiene pătratice

Această metodă accentuează distanța dintre profiluri. Profilurile mai apropiate sunt aduse mai aproape, iar cele mai îndepărtate sunt separate mai mult. Distanța dintre profiluri este pătratul distanței euclidiene:

$$d(X,Y) = \sum_{i} (x_i - y_i)^2$$

Măsurarea similarității pe baza distanțelor Manhattan

Distanța este suma diferențelor absolute dintre cele două profiluri:

$$d(X,Y) = \sum_{i} |x_i - y_i|$$

Măsurarea similarității pe baza distanțelor Cebisev

Distanța este diferența maximă absolută dintre valorile celor două profiluri:

$$d(X,Y) = \max_i |x_i - y_i|$$

Algoritmul de clusterizare implică următorii pași:

1. calcularea distanțelor dintre gene/probe și alcătuirea unei matrice a distanțelor.
2. identificarea celor mai apropiate gene.
3. formarea unui cluster din aceste gene.
4. calcularea distanțelor între clusterul nou format și celelalte gene sau clustere.
5. se reiau pașii 2, 3, 4 până la integrarea completă a tuturor genelor și clusterelor.

II.3.1.2. Alegerea metodei de grupare

La formarea unui nou cluster se impune recalcularea distanței dintre acesta și restul genelor (clusterelor). Există mai multe metode, fiecare va produce clusterizări diferite, în consecință este important să se aleagă cu atenție metoda ce urmează să fie folosită. Aceste metode sunt implementate în aplicații software de analiză cum ar fi: GeneSpring GX, R, J-Express.

Metoda legăturii simple (single linkage)

Această metodă definește distanta dintre două clustere ca fiind distanța minimă dintre membrii celor două clustere (gene). Metoda poate fi utilă în cazul în care datele prezintă o tendință de clusterizare naturală dar au profiluri neregulate. Se recomandă cu o oarecare rezervă pentru clusterizarea datelelor microarray (fig. 2.6).

Metoda legăturii complete (complete linkage)

Această metodă definește distanța dintre clustere ca fiind distanța maximă dintre membrii acestora. Produce clustere compacte și bine definite, funcționează mult mai bine cu date în care există o tendința naturală de clusterizare decât cu date neomogene (fig. 2.7).

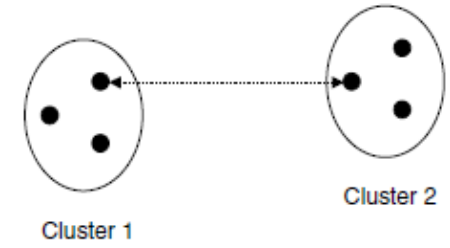


Fig. 2.6. Metoda legăturii simple

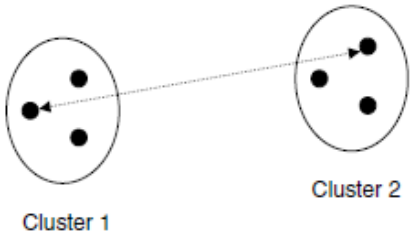


Fig. 2.7. Metoda legăturii complete

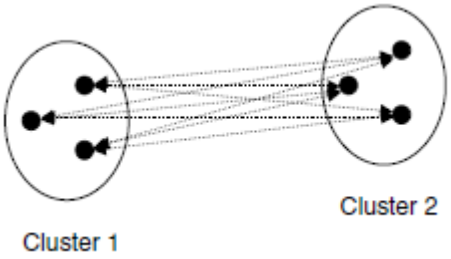


Fig. 2.8. Metoda mediei

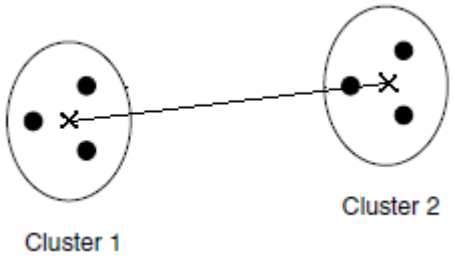


Fig. 2.9. Metoda legăturii centroidelor

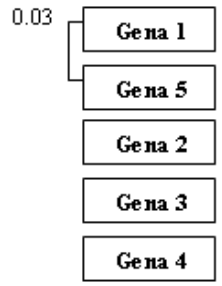
Metoda mediei (average linkage)

Această metodă definește distanța dintre clustere ca fiind distanța medie între toate perechile care se pot forma între membrii celor două clase. Metoda produce rezultate bune în aplicațiile microarray (fig. 2.8).

Metoda legăturii centroidelor (centroid linkage)
Această metodă definește distanța dintre clustere ca fiind distanța dintre centrele de masă ale membrilor celor două clase. Este o metodă des preferată în studiile de microarray (fig. 2.9).

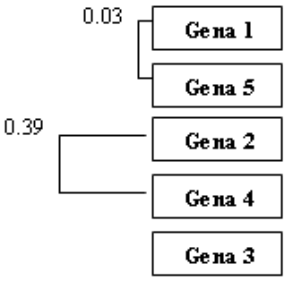
Exemplificare a algoritmului pe un set de 5 gene.
Este alcătuită matricea distanțelor pentru aceste gene și este identificată cea mai mică distanță dintre gene (gena1-gena5). Aceste gene, cu cel mai mare grad de similaritate, sunt cuplate între ele formând un cluster.

	Gena1	**Gena2**	**Gena3**	**Gena4**	**Gena5**
Gena1	0,00				
Gena2	0,57	0,00			
Gena3	0,79	0,46	0,00		
Gena4	0,39	0,39	0.51	0,00	
Gena5	0,03	0,62	0.79	0.43	0,00



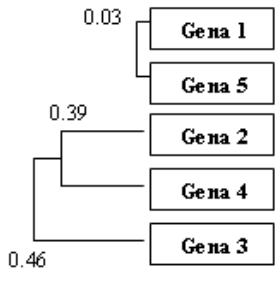
Distanțele dintre clusterul nou format și celelalte gene sunt recalculate și înlocuite în matricea distanțelor. Sunt identificate din nou cele mai mici distanțe (gena2-gena4) și este alcătuit un nou cluster.

	G1-G5	**Gena2**	**Gena3**	**Gena4**
G1-G5	0,00			
Gena2	0,59	0,00		
Gena3	0,79	0,46	0,00	
Gena4	0,41	0,39	0.51	0,00

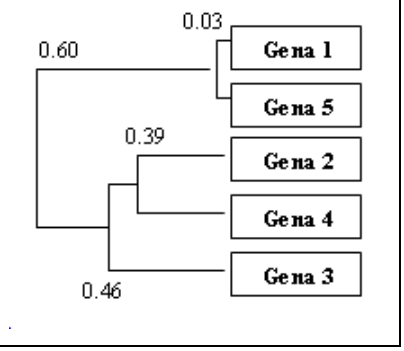


Distanțele sunt recalculate, similaritatea cea mai mare fiind între clusterul G2-G4 și gena3.

	G1-G5	**G2-G4**	**Gena3**
G1-G5	0,00		
G2-G4	0,50	0,00	
Gena3	0,79	0,46	0,00



În final, cele două clustere sunt unite, obținându-se dendrograma finală.

	G1-G5	G2-G4-G3	
G1-G5	0,00		
G2-G4-G3	0,60	0,00	

Rezultatul clusterizării este, în general, afișat sub formă unei *heat map* în care intensitățile probelor sunt reprezentate într-o hartă a culorilor, combinată cu una sau două dendrograme (fig. 2.10). Genele sunt reprezentate pe rânduri, iar probele pe coloane. Culorile reprezintă nivelul de expresie al genelor: roșu înseamnă nivele înalte de expresie, verde înseamnă nivele scăzute de expresie, negru reprezintă lipsa diferenței de expresie.

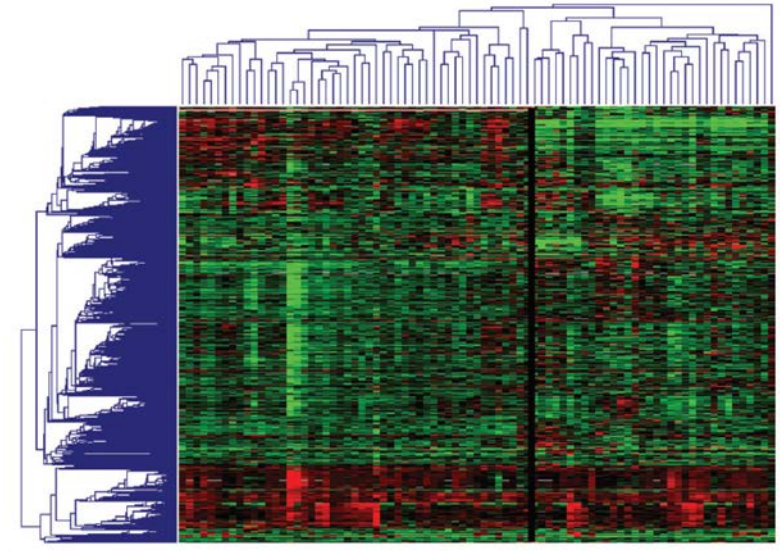


Fig. 2.10. Clusterizare ierarhică pentru un experiment bicolor.

În prezent, nu există un consens pentru alegerea unei metode de clusterizare, această opțiune rămâne la latitudinea cercetătorului. În funcție de metodă de clusterizare folosită se pot produce diferite dendrograme. Astfel, un anumit *pattern* al datelor poate fi doar rezultatul unei anumite metodologii și nu al informației biologice propriu zise. De aceea, informațiile furnizate de metodele de clusterizare trebuie validate.

II.3.2. Validarea structurii de cluster. Arbore de consens

Validarea structurii de cluster se realizează prin generarea unor seturi de date imaginare cu structură similară cu setul inițial. Datele sunt obținute dintr-o distribuție normală având fixate deviația standard și eroarea ca și în experimentul inițial. Aceste date imaginare sunt adăugate nivelelor de expresie inițiale ale genelor introducând un nivel de zgomot în cadrul datelor care practic simulează erorile experimentale. În studiile de validare sunt generate peste 1000 de astfel de seturi de date.

Unele gene pot fi clusterizate diferit în noua dendrogramă. Acele clusterele care rămân neschimbate la fiecare clusterizare sunt considerate robuste la erorile experimentale și pot fi folosite pentru extragerea informației biologice. Celelalte clustere care apar doar tranzitoriu, sunt afectate de erorile experimentale, folosirea lor pentru trasarea unor concluzii biologice fiind riscantă. Matematicienii consideră clusterele care apar cu regularitate în mai mult de jumătate din clusterizări, un model consecvent denumit arbore de consens.

II.3.3. Gruparea iterativa (*k*-means clustering)

Această metodă diferă de clusterizarea ierarhică prin faptul că numărul de clustere trebuie stabilite apriori și nu există nici o relație ierarhică între clustere și nici între genele sau probele din cadrul unui cluster. Clusterele sunt doar grupuri cu profiluri de expresie genică similare.

Algoritmul de clusterizare *k*-means:

1. stabilirea numărului de clustere (*k*).
2. alocarea aleatorie a profilurilor de expresie genică fiecăruia dintre cele *k* clustere
3. calcularea centroidelor fiecărui cluster.
4. evaluarea individuală a fiecărui profil de expresie genică prin calcularea distanței (similarității) dintre acesta și centroidul fiecăruia dintre cele *k* clustere.
5. dacă profilul este mai apropiat de un alt cluster decât cel în care se află, acest profil este mutat în noul cluster după care centroidele ambelor clustere sunt recalculate.
6. algoritmul continuă până când fiecare profil este corect încadrat.

Rezultatul clusterizării este influențat de numărul *k* de clustere alese (fig. 2.11). Alegerea numărului de clustere se poate face printr-o metodă de reducere a dimensiunilor care permite vizualizarea genelor și probelor într-un spațiu bidimensional pe baza matricei distanțelor, astfel încât distanțele din această matrice să fie cât mai bine aproximate în spațiul bidimensional. Vizualizarea într-un spațiu bi- sau tri- dimensional permite observarea numărului de clustere *naturale* care se formează, putând astfel alege numărul *k* în funcție de clusterele observate. Este posibilă folosirea unui *k* arbitrar ales de către cercetător, aici experiența cercetătorului și interpretarea rezultatelor biologice are un rol determinant.

II.3.4. Validarea structurii de cluster

Este similară cu validarea clusterizării ierarhice. O altă metodă de validare poate fi repetarea procesului pentru un același *k*.

În medicină, asocierea factorilor de prognostic cu supraviețuirea aduce extrem de multe beneficii. Identificarea, prin analize nesupervizate, a subgrupurilor de probe în cadrul unei patologii considerate în prealabil *omogene* permite analize de supraviețuire care să testeze dacă aceste subgrupuri prezintă rezultate clinice diferite. Legătură dintre nivelul de expresie al unui *pattern* de gene (factor de prognostic) și perioada de supraviețuire poate furniza instrumente utile în stabilirea diagnosticului și terapiilor precoce, prompte și personalizate.

O altă aplicație a setului de gene obținute în analizele nesupervizate este corelarea, prin analize de regresie liniară, a nivelelor de expresie ale acestor gene cu un *marker surogat* care indică o măsură a progresiei bolii. De exemplu, marker surogat pot fi considerate proteinele serice, poate fi considerat un anumit răspuns la tratament, sau orice alte indicații clinice care pot fi corelate cu statusul bolii.

II.4. Profilul funcțional

Independent de platforma și analizele folosite, rezultatele experimentelor microarray sunt, în cele mai multe cazuri, o lista de gene diferit exprimate. Provocarea în studiile de expresie genică este de a

transforma această listă de gene diferit exprimate într-un profil funcțional, capabil să ofere o mai bună înțelegere a fenomenelor biologice care stau la baza apariției și progresiei bolilor. Datorită numărului mare de gene diferit exprimate, căutarea informațiilor disponibile, genă cu genă, este un proces consumptiv de timp și complicat de efectuat. Abordarea modernă implică gruparea genelor pe procese și apoi urmărirea proceselor ce sunt reprezentate în experimentul de interes și identificarea *pattern*-urilor de expresie a genelor implicate în același proces. În prezent, există numeroase instrumente de analiză computerizată care folosesc datele din *Gene Ontology* pentru a rezolva această problemă[30].

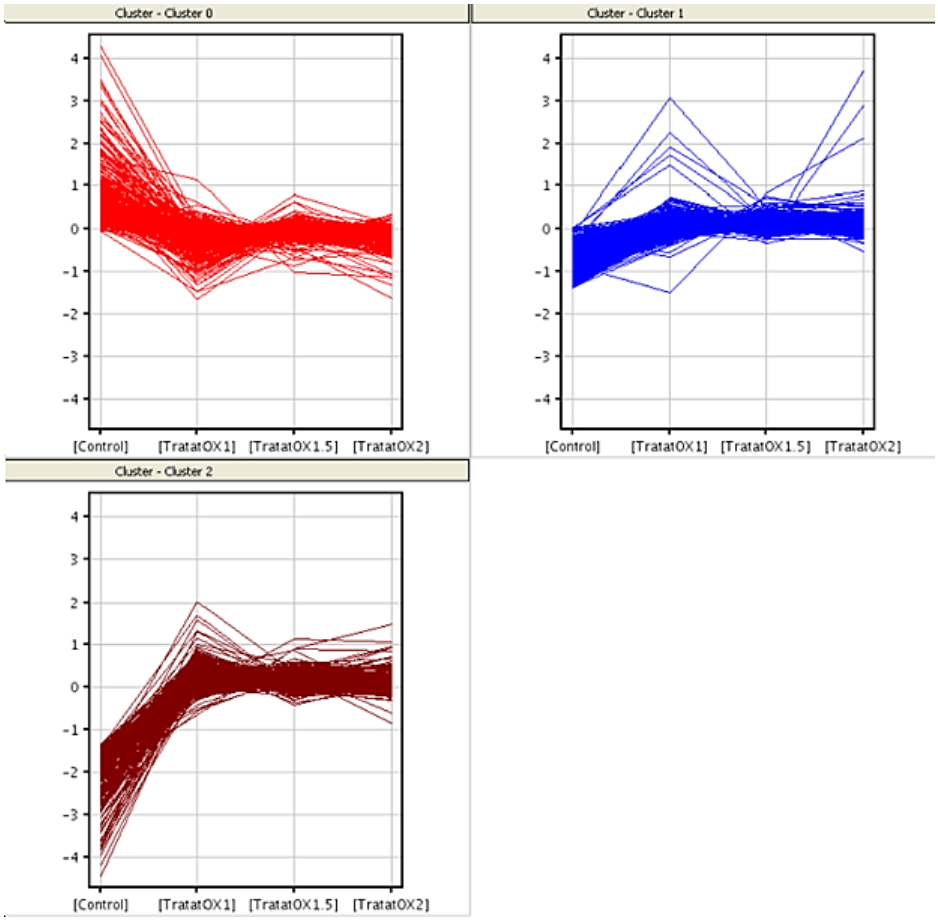


Fig. 2.11. Clusterizare iterativă pentru un k = 3.

Consorțiul *Gene Ontology* (http://www.geneontology.org), înființat în 1998, menține o bază de date cu vocabular controlat pentru a descrie atributele genelor și a produșilor genici pentru orice organism. *Gene Ontology* (GO) folosește aceste vocabulare pentru a descrie genele la trei niveluri:

- la nivelul funcțiilor moleculare (ex. activitate catalitică, reglarea transcripției etc.)
- la nivelul proceselor biologice (ex. proliferare celulară, apoptoză etc.)
- la nivelul componentelor celulare (ex. nucleu, citoschelet etc.)

În momentul de față, există 32992 termeni GO din care: 19867 termeni definesc procesele biologice, 2771 definesc componente celulare și 8900 definesc funcții moleculare. Fiecare categorie este definită de un cod unic de înregistrare în format GO:xxxxxxx, genele fiind asociate acestor termeni printr-un proces manual sau automat de adnotare (fig. 2.12). Fiecare genă sau produs de expresie poate avea una sau mai multe funcții moleculare, poate lua parte în unul sau mai multe procese biologice sau poate acționa în una sau mai multe componente celulare. Cele trei vocabulare sunt independente și fiecare este reprezentat într-o structură DAG (Directed Acyclic Graph) care stabilește

relații de tip ierarhic între acești termeni, cu posibilitatea ca un descendent să aibă mai mulți strămoși.

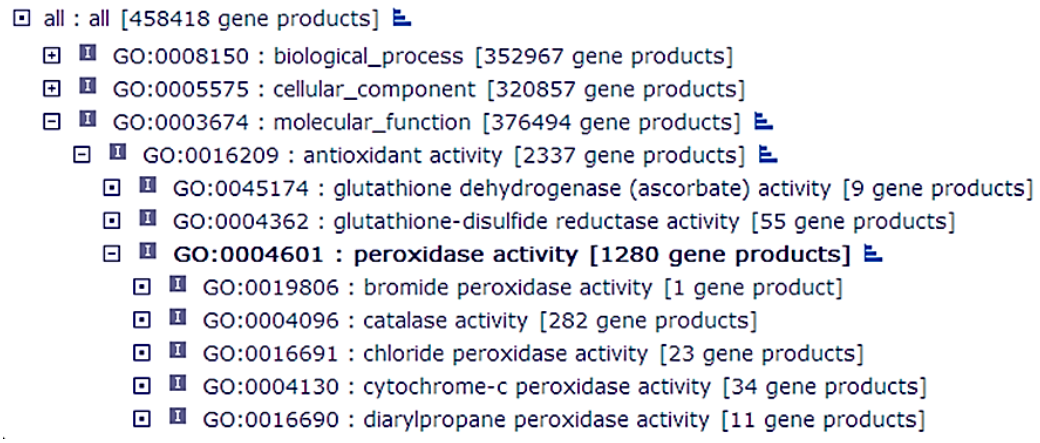


Fig. 2.12. Gene Ontology AmiGO Tree Browser.

Analize de ontologie genică (analize GO)

Utilizarea adnotărilor de ontologie genică permite obținerea informațiilor de ansamblu despre un număr mare de gene fără a fi nevoie să fie accesată fiecare genă în mod individual. Pe baza acestor informații au fost dezvoltate instrumente de analiză computerizată pentru a identifica funcțiile biologice *cel mai bine* reprezentate într-o listă de gene. Există multe pachete care implementează astfel de analize: *Onto-Express*, *GoMiner*, *DAVID*, *FatiGO*, *CLENCH*, *GOToolBox*, *GeneSpingGX* (Agilent), *NetAffx Analysis Center* (Affymetrix). Toate aceste aplicații software lucrează similar, diferența făcând-o în principal metoda statistică pe care o implementează fiecare.

Primul pas într-o analiză de ontologie genică este adnotarea funcțională a genelor de interes prin descărcarea de pe site-ul *Gene Ontology Consortium* a fișierelor OBO care conțin termenii GO. Pornind de la o listă de gene diferit exprimate și folosind un test statistic pot fi identificate categoriile GO (procese biologice, funcții moleculare) care sunt supra sau sub reprezentate în condițiile studiului. Pentru un set dat, această metodă compară numărul de gene diferit exprimate găsite în fiecare categorie GO de interes cu numărul de gene care se găsesc în aceea categorie datorită întâmplării. Dacă numărul de gene observat este semnificativ mai mare decât numărul de gene datorate hazardului, pe baza unui test statistic (ex. distribuția hipergeometrică, chi-pătrat, testul Fisher), categoria este considerată semnificativă, raportându-se o valoare *p* [31] [32] (fig. 2.13). Așa cum se observă din figură, semnificația nu este dată de numărul de gene existente într-o anumită categorie, ci de valoarea lui *p*.

Toate aceste pachete generează o lista de categorii funcționale în care sunt implicate genele inițiale. Categoriile funcționale care sunt reprezentate statistic semnificativ în lista de gene diferit exprimate sunt considerate a fi semnificative pentru studiul respectiv.

II.5. Rețele moleculare

Rețelele moleculare oferă o imagine intuitivă asupra genelor identificate ca fiind importante în studiul respectiv prin vizualizarea relațiilor dintre ele și a interacțiunilor cu alte gene.

Ingenuity Pathways Analysis (IPA) este o aplicație care permite generarea de *novo* a rețelelor moleculare pentru un set de gene de interes, folosind una dintre cele mai mari baze de date de rețele biologice, și anume, *Ingenuity Pathways Knowledge Base* (IPKB). Această bază de date stochează

relații între proteine, gene, celule, componente celulare și milioane de căi de semnalizare moleculară extrase din literatură. Listele de gene diferit exprimate sunt încărcate ca fișiere *Excel* în care este specificat nivelul *fold change* al genelor. Din această listă se pot adnota gene de interes care vor fi evidențiate în studiu. Aceste gene sunt selectate ca gene *focus* (focus gene) dacă îndeplinesc două criterii: sunt desemnate de cercetător că fiind de interes și au interacțiuni directe cu alte gene din IPKB. Genele focus sunt folosite în IPA pentru a crea rețele moleculare.

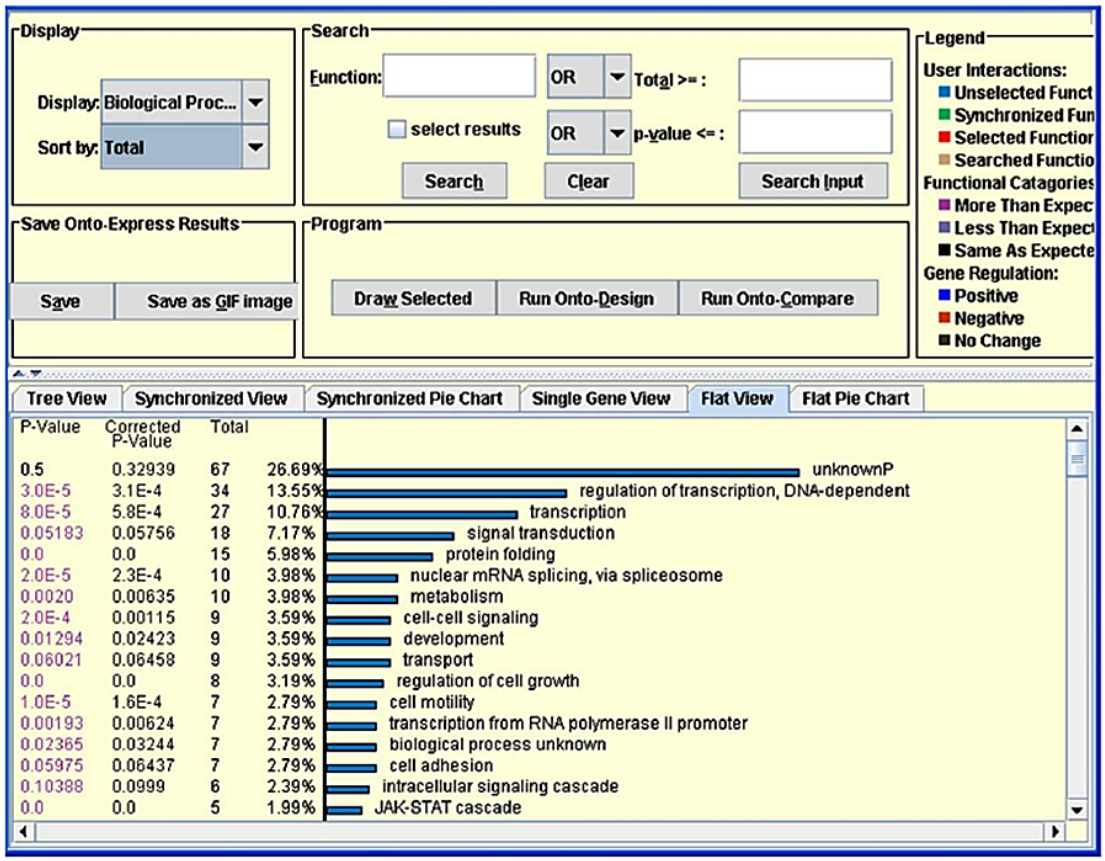


Fig. 2.13. Obținerea profilului funcțional folosind aplicația Onto-Express. (http://vortex.cs.wayne.edu/projects.htm). Figura evidențiază procesele biologice semnificative reprezentate în setul de gene de interes. Coloanele în albastru închis reprezintă procesele biologice, lungimea acestora fiind proporțională cu numărul de gene implicate în fiecare proces.

O rețea moleculară globală este generată pe baza a sute de mii de interacțiuni fizice și funcționale, directe și indirecte între genele din baza de date IPKB. Algoritmul dezvoltat de IPA permite selectarea unor subrețele ale acestei rețele globale, centrate pe genele focus și care includ interacțiuni cu alte gene. Într-o anumită rețea locală, sunt integrate cu predilecție genele focus care interacționează cu cât mai multe gene din această subrețea, în detrimentul celor care interacționează cu un număr mare de alte gene din toată rețeaua globală. Ulterior, IPA permite extinderea interacțiunilor directe și indirecte ale genelor la alte subrețele prin intermediul punctelor nodale.

Subrețelele selectate, care conțin genele focus, sunt prezentate sub formă unor tabele în ordinea semnificației lor. Genele care compun rețeaua sunt listate în una dintre coloane (gene focus - îngroșat). Rețelele locale sunt alcătuite din maximum 35 de gene. IPA atribuie scoruri fiecărei rețele. Scorul este stabilit în funcție de numărul de gene focus și dimensiunea rețelei, fiind cu atât mai mare cu cât numărul genelor focus este mai mare. De asemenea, sunt specificate și primele trei dintre cele mai reprezentative funcții pentru fiecare rețea (fig. 2.14).

ID	Genes	Score	Focus Genes	Top Functions
1	ABCC2, ARG1, ↓**CAPN2,** CD48, ↑**CLU,** COL5A1, CTCF, CYP3A2, ↓**DCN,** ↑**DDX17,** ↑**DNAJC7,** ↓**FOXC1,** ↓**GAPD*,** GBP2, GDI2, GSTA2, HPGD, ↓**ID2,** ↑**IL6,** ↑**LMNB1,** LY6A, ↑**MMP10,** ↓**MYC,** NR1I2, PAX4, PLS3, PON1, ↑**SFRS5,** SLCO1A4, SRM, ↑**TFF3,** TGFB1, ↑**TGFBI,** TM4SF2, UBE2C	25	15	Cell Cycle, Cellular Growth and Proliferation, Skeletal and Muscular System Development and
2	↑**AKR1C1,** ↑**AKR1C3,** AURKB, BBC3, BMX, CASP3, CTSD, EDN1, EEF2, EP400, ↓**FZD1,** ↓**GAPD*,** GPI, ↓**GSTM1,** ↑**HSPH1,** IGF1, MAD2L1, MKI67, MT2A, NDRG1, ↑**NGFRAP1,** ↓**NMB,** PDCD8, ↑**PLCB4,** ↓**PODXL,** PTEN, PTK2, RHOA, ROCK1, ↑**S100A2,** ↑**TMOD1,** TP53, ↓**TPM2,** ↓**TRPV2,** WEE1	23	14	Cellular Assembly and Organization, Cell Death, Connective Tissue
3	ADORA2B, ↓**ARL3,** ↑**CA12,** CD209, CDKN1A, CLEC4E, CRLF1, CX3CR1, DIO1, EGR2, ↓**F11R,** FCGR1B, GBP2, HMGN2, HOXA9, ↓**HRASLS3,** ↓**ICAM2,** ↑**IFITM1,** IFNG, IGF2, ↓**IGFALS,** IL15, IL18RAP, IL1F9, ↓**IMP-3,** INDO, MNT, ↑**MSX1,** PENK1, ↓**PIGR,** ↓**PMP22,** ↑**TFPI2,** TNF, ↓**TPR,** VHL	21	13	Cell-To-Cell Signaling and Interaction, Cellular Growth and Proliferation, Infectious Disease
4	AGT, ALB, ↑**AREG,** C5ORF13, COL1A2, ↓**COL3A1,** COX2, COX4I1, COX5A, COX5B, COX6C, ↓**COX7A1,** COX7A2, COX7C, ↑**CXCL14,** EREG, FASN, FGF2, FTH1, GFAP, GH1, ↓**HBA1,** ↓**ID2,** ID3, IGFBP4, ↑**ITM2C,** KLKB1, NCL, PDGFA, PTMA, RBL2, REN, TF, TFDP1, ↓**TNFRSF12A**	11	8	Cell Cycle, Cellular Growth and Proliferation, Cellular Development

Fig. 2.14. Detaliu de rețele generate cu pachetul computațional *Ingenuity Pathways Analysis*.

Numărul maxim de rețele afișate pentru o analiză sunt plafonate la 200. Astfel, s-ar putea ca un număr de gene de interes să nu fie afișate în aceste analize dacă nu se încadrează în topul celor 200 de rețele. De asemenea, se poate vizualiza legătura cu alte rețele, procese celulare, localizări subcelulare, rețele canonice, pentru a avea o vedere de ansamblu asupra proceselor biologice (fig. 2.15).

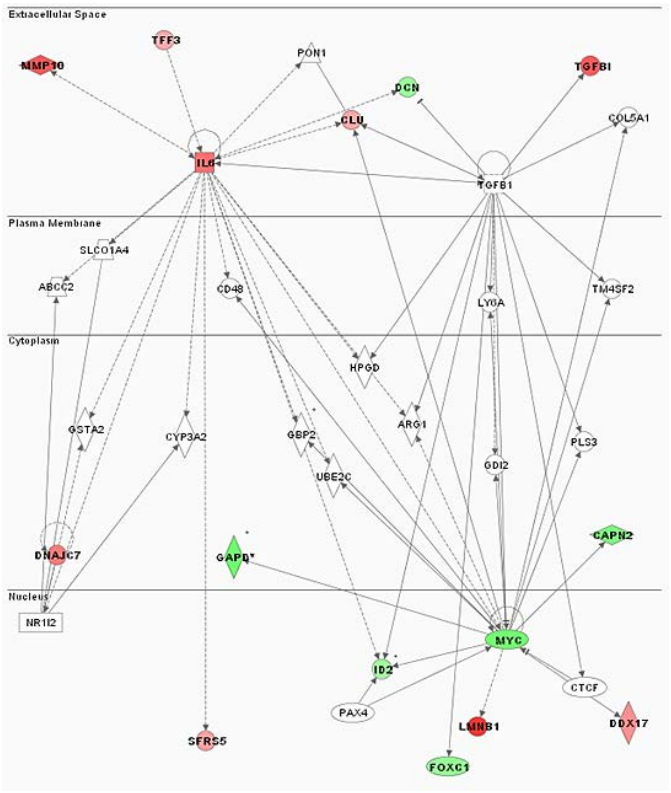


Fig. 2.15. Rețea moleculară generată în IPA.

II.6. Concluzii

Studiile de genomică funcțională bazate pe tehnologia microarray permit cuantificarea simultană a nivelelor de expresie a zeci de mii de gene. Aceste studii generează cantități enorme de date care furnizează informații cu rol deosebit de important în cercetarea biomedicală, de la înțelegerea proceselor biologice, la identificarea de noi biomarkeri, identificarea mecanismelor de toxicitate,

predicția răspunsului la terapie, prognostic și supraviețuire. Alături de folosirea unor probe de calitate adaptate condițiilor de studiu, strategiile în stabilirea design-ului, alegerea celor mai potrivite metode și validările repetate ale acestora, sunt esențiale pentru a garanta obținerea unor informații de calitate.

II.7. Referințe bibliografice

[1] **Beer D.G., Kardia S.L., Huang C.C., Giordano T.J., Levin A.M., Misek D.E.** et al, 2002, *Gene-expression profiles predict survival of patients with lung adenocarcinoma*. Nat. Med. 2002, 8:816-24, [PubMed: 12118244]

[2] **Swagell C.D., Henly D.C., Morris C.P.**, 2005, *Expression analysis of a human hepatic cell line in response to palmitate*, Biochem. Biophys. Res. Commun. 2005; 328:432-41.

[3] **Pass H.I., Liu Z., Wali A., Bueno R., Land S., Lott D.** et al, 2004, *Gene expression profiles predict survival and progression of pleural mesothelioma*, Clin. Cancer Res. 2004, 10:849-59.

[4] **Khatri P., Draghici S.**, 2005, *Ontological analysis of gene expression data*: *current tools*, *limitations and open problems*, Bioinformatics 2005, 21:3587-95.

[5] **Kerr M.K., Churchill G.A.**, 2001, *Statistical design and the analysis of gene expression microarray data*, Genet. Res. 2001, 77:123-8.

[6] **Kerr M.K., Afshari C.A., Bennett L., Bushel P., Martinez J., Walker N.J.** et al., 2001, *Statistical analysis of a gene expression microarray experiment with replication*. Statistica Sinica 2001, 12:203-18.

[7] **DeRisi J., Penland L., Brown P.O., Bittner M.L., Meltzer P.S., Ray M.** et al., 1996, *Use of a cDNA microarray to analyse gene expression patterns in human cancer*, Nat Genet 1996, 14:457-60.

[8] **Wellmann A., Thieblemont C., Pittaluga S., Sakai A., Jaffe E.S., Siebert P.** et al., 2000, *Detection of differentially expressed genes in lymphomas using cDNA arrays: identification of clusterin as a new diagnostic marker for anaplastic large-cell lymphomas*. Blood 2000, 96:398-404.

[9] **Draghici S.**,2003, *Data analysis tools for DNA microarrays*. Chapman and Hall/CRC Press, Boca Raton (FL)

[10] **Sebastiani P., Gussoni E.K.I.**, 2003, *M.F.R. Statistical challenges in functional genomics*. Stat Sci. 2003, 18:33-70.

[11] **Smyth G.K.**, 2005, *Limma: linear models for microarray data*, Springer, New York, NY:2005.

[12] **Tusher V.G., Tibshirani R., Chu G.**, 2001, *Significance analysis of microarrays applied to the ionizing radiation response*, Proc Nat Acad Sci. USA 2001, 98:5116-21.

[13] **Smyth G.K., Yang Y.H., Speed T.**, 2003, *Statistical issues in cDNA microarray data analysis*, Methods Mol Biol 2003, 224:111-36.

[14] **Dudoit S., Shaffer J., Boldrick J**., 2003, *Multiple hypothesis testing in microarray experiments*. Stat Sci. 2003, 18:71-103.

[15] **Holm S**., 1979, *A simple sequentially rejective multiple test procedure*, Scand J Stat 1979, 6:65-70.

[16] **Westfall, P.H., Young S.S**., 1993, *Resampling-based multiple testing: examples and methods for p-value adjustment*, Wiley, New York, NY:1993.

[17] **Benjamini Y., Hochberg Y**., 1995, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J Royal Stat Soc B 1995, 57:289-300.

[18] **Shaffer J.P**., 1995, *Multiple hypothesis testing*, Ann Rev Psych 1995, 46:561-84.

[19] **Quackenbush J**., 2006, *Microarray analysis and tumor classification*. N Engl J Med 2006, 354:2463-72.

[20] **Allison D.B., Cui X.Q., Page G.P., Sabripour M**., 2006, *Microarray data analysis: from disarray to consolidation and consensus*, Nat Rev Genet 2006, 1:55-65.

[21] **Dupuy A., Simon R.M**., 2007, *Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting*. J Natl Cancer Inst 2007, 99:147-57.

[22] **Muller P., Parmigiani G., Robert C., Rousseau J**., 2004, *Optimal sample size for multiple testing: The case of gene expression microarrays*. J. Am. Stat. Assoc 2004, 99, 990-1001.

[23] **Pawitan Y., Michiels S., Koscielny S., Gusnanto A., Ploner A**., 2005, *False discovery rate, sensitivity and sample size for microarray studies*, Bioinformatics 2005, 21, 3017-3024.

[24] **Gadbury G.L**. et al., 2004, *Power analysis and sample size estimation in the age of high dimensional biology: a parametric bootstrap approach and examples from microarray research*. Stat. Methods Med. Res. 2004, 13, 325-338.

[25] **Tsai C.A., Hsueh H.M., Chen J. J**., 2003, *Estimation of false discovery rates in multiple testing: application to gene microarray data*, Biometrics 2003, 59, 1071-1081.

[26] **Pavlidis P., Li Q., Noble W.S**., 2003, *The effect of replication on gene expression microarray experiments*, Bioinformatics 2003, 19, 1620-1627.

[27] **Lee J.K.,** et al, 2003, *Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells*, Genome Biol 2003, 4(12):R82.

[28] **Weinstein J.N.**, et al., 2002, *The bioinformatics of microarray gene expression profiling*, Cytometry 2002, 47 (1):46-9.

[29] **Bethin K.E., Nagai Y., Sladek R., Asada M., Sadovsky Y., Hudson T.J.** et al., 2003, *Microarray analysis of uterine gene expression in mouse and human pregnancy*. Mol Endocrinol 2003, 17:1454-69.

[30] **Draghici S., Khatri P., Martins R.P., Ostermeier G.C., Krawetz S.A**., 2003, *Global functional profiling of gene expression*, Genomics 2003, 81:98-104.
[31] **Draghici S., Khatri P., Bhavsar P., Shah A., Krawetz S., Tainsky M.A**., 2003, *Onto-Tools, The toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate*, Nucleic Acid Research 2003, 31:3775-81.
[32] **Khatri P., Bhavsar P., Bawa G., Draghici S**., 2004, *Onto-Tools: an ensemble of web-accessible, ontologybased tools for the functional design and interpretation of high-throughput gene expression experiments*. Nucleic Acids Research 2004, 32:W449-56.