Universitat
Autònoma
de Barcelona

Obtenció de Jerarquies d'Estats Cel·lulars vía Web

Memòria del Projecte Fi de Carrera
d'Enginyeria en Informàtica
realitzat per Bernat Gispert Pons
i dirigit per Mario Huerta Casado i
Jordi Gonzàlez Sabaté
Bellaterra, 22 de juny de 2010

Universitat Autònoma de Barcelona

etse
Escola Tècnica Superior d'Enginyeria

El sotasignat, ...

Professor/a de l'Escola Tècnica Superior d'Enginyeria de la UAB,

CERTIFICA:

Que el treball a què correspon aquesta memòria ha estat realitzat sota la seva direcció per en

I per tal que consti firma la present.

Signat: .....

Bellaterra, ...de.....de 200.....

Universitat Autònoma de Barcelona

etse
Escola Tècnica Superior d'Enginyeria

El sotasignat, ...

Professor/a de l'Escola Tècnica Superior d'Enginyeria de la UAB,

CERTIFICA:

Que el treball a què correspon aquesta memòria ha estat realitzat sota la seva direcció per en

I per tal que consti firma la present.

Signat: .....

Bellaterra, ...de.....de 200.....

Índex

NOMENCLATURA A TENIR EN COMPTE

Estat fenotípic: Estat en la que es troba la cèl·lula després de l'expressió d'uns gens concrets.

Canvi fenotípic: El pas d'un estat cel·lular a un altre diferent després d'uns gens concrets. Per exemple d'un estat sa a un estat malalt o al revés.

Relació d'expressió: Es la dependència existent entre les expressions de 2 gens. Les expressions dels gens no son arbitràries i estan totes interrelacionades.

Relació d'expressió no lineal: Es la expressió entre 2 gens però que no segueixen una relació de coexpressió o de inhibició. Es a dir que no s'expressen a l'hora (y=mx) o que l'expressió d'un gen no significa que l'altre gen deixa d'expressar-se (y=-mx).

Coexpressió de gens: Gens que s'expressen simultàniament mantenint una relació y=mx en les seves expressions. Els gens que s'expressen simultàniament duran a terme un canvi fenotípic quan s'expressin.

Microarray: Tecnologia que permet obtenir els nivells d'expressió d'un gran numero de gens per un gran nombre de condicions experimentals.

Sample o mostra: En el ciències experimentals les mostres son les dades obtingudes en un experiment. En la tecnologia de microarrays es l'expressió d'un gen donat per una condició experimental concreta.

Condició mostral: Son les condicions a les que es sotmet la cèl·lula per estudiar la expressió dels gens. Es a dir, les condicions experimentals.

PCOP (Principal Curve of Oriented Points): Un mètode d'anàlisi multivariable no paramètric per obtenir els patrons que descriuen la relació entre dues o mes variables a partir d'un conjunt de mostres.

POP(Principal Oriented Points): A partir de una variació del mètode de components principals però només aplicada de forma local a subespais mostrals, s'obtenen aquests POPs que son una discretització d'aquest subespais. Cada POP representa les mostres que discretitza, el conjunt de POPs constitueix la PCOP.

Preprocés: Dins de l'aplicació web, la part de càlculs que generen els fitxers que utilitzarà l'aplicació on-line. El preprocés no es interactiu i s'executa una única vegada per cada conjunt de dades que s'ha d'analitzar.

1 INTRODUCCIÓ

1.1 MOTIVACIONS

Una microarray és un conjunt de massiu de dades, que ens proporciona els graus d'expressió d'un determinat nombre de gens vers a diferents estímuls o condicions mostrals. La tecnologia de microarrays encara que és molt cara, aporta molta informació. És obvi doncs, que cal aprofitar al màxim les dades de les microarrays, i es desenvolupin eines que extreguin la màxima informació rellevant possible. En el IBB-UAB (Institut de Biotecnologia I Biomedicina de la Universitat Autònoma de Barcelona), es desenvolupen eines per a l'anàlisi de les dades obtingudes de les microarrays. Aquestes eines intenten respondre a les necessitats dels investigadors, i els permet la possibilitat de fer proves i formular des de hipòtesis fins a models sencers. Seguint aquesta línia d'investigació s'ha desenvolupat una nova eina per a l'estudi d'interdependències de fenotips, obtenint múltiples distribucions de clusters de les condicions experimentals de la microarray analitzada que representen diferents fenotips així com els gens que duen a terme els canvis fenotípics.

1.2 ESTAT DE L'ART

En l'estudi d'expressions gèniques s'utilitzen conjunts de dades multi-variantss, que son aquells que contenen valors observats de *k* característiques per a *n* individus.
La tecnologia basada en microarrays ens proporciona aquests conjunts massius de dades i les eines matemàtiques i estadístiques resulten molt útils per poder estudiar-les, encara que cada vegada és més necessària una major precisió en els anàlisis per aconseguir resultats que contemplin la complexitat dels models biològics i que, a la vegada, s'ajustin més als comportaments reals.
La majoria d'anàlisi de dades de microarrays estan basats en mètodes de reducció de dimensions com els components principals i mètodes de clustering.
La detecció de components principals permeten representar en una o més dimensions el conjunt de dades de dimensió k. Un dels problemes d'aquests mètodes és la aproximació que fa de la distribució d'aquest conjunt de dades al voltant d'un recta. En el cas dels gens, les relacions d'expressió poden no ser lineals, i per tant aquesta detecció de components principals no és útil en alguns casos. Per aquest motiu nosaltres utilitzarem Corves principals concretament les PCOP[1], que permeten estudiar les relacions no

lineals entre les expressions dels gens[3].

Hi ha diversos mètodes d'anàlisi que duen a terme una agrupació (*clustering*) global (considerant el total de l'espai mostral) de les mostres de les microarrays, com el *Hierarchical Clustering* o el *Self-Organizing Maps*, o d'altres que realitzen agrupaments locals tenint en compte només un subconjunt de gens co-expressats o condicions mostrals, com el *Biclustering*. L'eina desenvolupada pertany als mètodes d'agrupació local, però no considerant subconjunts de gens co-expressats si no tenint en compte només els parells de gens on l'expressió gènica pateix una variació a causa d'un canvi fenotípic. Els fenotips implicats en el canvi fenotípic constitueixen els grups o clusters de mostres que es proporcionaran com a resultat a l'aplicació web. Els parells de gens que pateixen un mateix canvi fenotípic, es proporcionaran juntament amb la seva distribució de clusters, i serviran al usuari per a estudiar la causa-efecte dels canvis fenotípics.

Per obtenir aquest sistema de clustering, es farà us d'un nou mètode de clustering desenvolupat al IBB-UAB a partir del càlcul de les PCOP[2] i que ha demostrat un bon us per fer clusters de samples a partir de les relacions de expressions[4].

L'aplicació web que s'ha desenvolupat s'integrarà a un servidor d'aplicacions web per l'anàlisi de microarrays: http:\\revolutionresearch.uab.es [5][6]. Per desenvolupar l'eina web que permetrà l'anàlisi on-line s'ha fet us de les eines més competitives disponibles actualment: Entorn de desenvolupament en Linux Ubuntu 8.10, Servidor apache 2.0, PHP 5.2.6, la llibreria JavaScript YUI (Yahoo User Interface), entorn de programació eclipse, llibreria gràfica GD de PHP5 per a la generació d'imatges de textos necessàries a la web.

1.3 OBJECTIUS

Desenvolupar una eina per a l'estudi dels canvis fenotípics a partir de dades de microarrays. Aquesta eina crearà els clusters de les condicions mostrals de la microarray que es corresponen amb els diferents canvis fenotípics inherents a les dades de la microarray.

Associats a aquests clusters de condicions mostrals, poder saber quins son els gens involucrats en aquest canvi fenotípic.

A partir de les dades generades per programes que han detectat les relacions no lineals de parelles de gens i que n'obtenen els clusters de les condicions mostrals, desenvolupar el procés per buscar aquelles que son semblants i proporcionar així distribucions de clusters finals que siguin cada una la representació de les diverses distribucions de clusters semblants entre elles.

Alliberar a la interfície web de càlculs costosos, fent-los prèviament un cop es carrega una nova microarray al sistema. Aquests càlculs es faran només 1 cop, on es crearan els fitxers i dades necessàries per poder ser tractades des de la interfície web, sense haver de tornar a fer els mateixos càlculs repetidament.

Crear una interfície web que permeti poder comparar les diferents distribucions de clusters de forma interactiva i fàcil per a l'usuari.

Poder veure gràficament a la interfície web com els clusters d'una distribució de clusters final es distribueixen en la relació d'expressió que separa els estats fenotípics.

Donar la possibilitat a l'usuari de poder comparar distribucions de clusters segons els diferents paràmetres de càlcul i poder estudiar aquells resultats que més li convinguin saben en tot moment amb quins paràmetres han estat creats.

1.4 RESULTATS I ORGANITZACIÓ DE LA MEMÒRIA

En aquest projecte s'ha desenvolupat l'aplicació web: El PCOPSample-cl, una eina que pertany als mètodes d'agrupació (clustering) local, que no busca subconjunts de gens co-expresats (anàlisi de relacions lineals), si no parelles de gens que davant de canvis fenotípics, la seva relació d'expressió pateix fluctuacions. Els fenotips implicats en el mateix canvi fenotípic estaran representats per la mateixa distribució final de clusters. El resultats del PCOPSample-cl seran les diferents distribucions finals de clusters i les parelles de gens involucrades en aquests canvis fenotípics. Aquestes parelles de gens podran ser estudiades per trobar la causa i efecte del canvi fenotípic.

La memòria del projecte està organitzada de la següent manera:
Primer la secció de fonaments teòrics exposa els conceptes biològics bàsics per entendre la rellevància i significació de l'anàlisi de microarrays.

En l'apartat 2 Metodologia, s'explica el mètode utilitzat per obtenir les distribucions de clusters.
En l'apartat 3, Resultats i Discussió, s'avaluaràn els resultats obtinguts. S'avaluaran per un costat la qualitat de les solucions i per l'altre la influència en el resultat final del criteris seguits en el disseny del mètode per obtenir les distribucions de clusters. Després en el mateix apartat 3, s'explica com s'ha estructurat l'aplicació web que permet analitzar i visualitzar els resultats i per quins motius. Posteriorment, es comenta com s'ha realitzat el càlcul de l'error acumulat en els machings de clusters i les repercussions d'aquest error en els matchings. Finalment es mostren diferents tipus d'anàlisis realitzats amb l'us de l'aplicació.
En l'apartat 4, s'exposen les conclusions del projecte.
En l'apartat 5, es dona una visió de com està estructurat el servidor, tant de la part de preprocés com de l'aplicació web online.

1.5 FONAMENTS TEÒRICS

La tecnologia de Microarrays o tecnologia basada en el xip d'ADN, és una de les tècniques usades en la biologia molecular. Aquests monitoritzen els nivells d'expressió de mil·lers de gens del genoma d'un organisme de forma simultànea. Això permet per exemple, estudiar els gens que produceixen certes patologies comparant cel·lules sanes amb cel·lules que desenvolupen certs tipus de malalties.
Com a resultat de la tecnología de microarrays es proporciona una matriu de dades amb l'expressió de tots els gens sota les condicions mostrals de la microarray.
Una de les finalitats de l'anàlisi d'aquestes dades és justament determinar la relació que hi pugui haver entre les expressions dels gens de la microarray involucrats en certs canvis d'estat cel·lulars, per a predir i/o detectar malalties.

<<esto que has puesto sobre microarrays ya está bien, si encuentras algo sobre clusters de microarrays bien, sino puedes centrarte en otras cosas a arreglar, no es prioritario>>

Trobar patrons de relacions d'expressió biologicament significatives, és llavors una de les finalitats de les eines d'anàlisis de les dades obtingudes de les microarrays. Un dels patrons biologicament més significatius son els anomenats switch, on els gens que hi participen fan que altres gens passin a expressarse o deixin d'expressarse, el que comporta un canvi d'estat cel·lular.

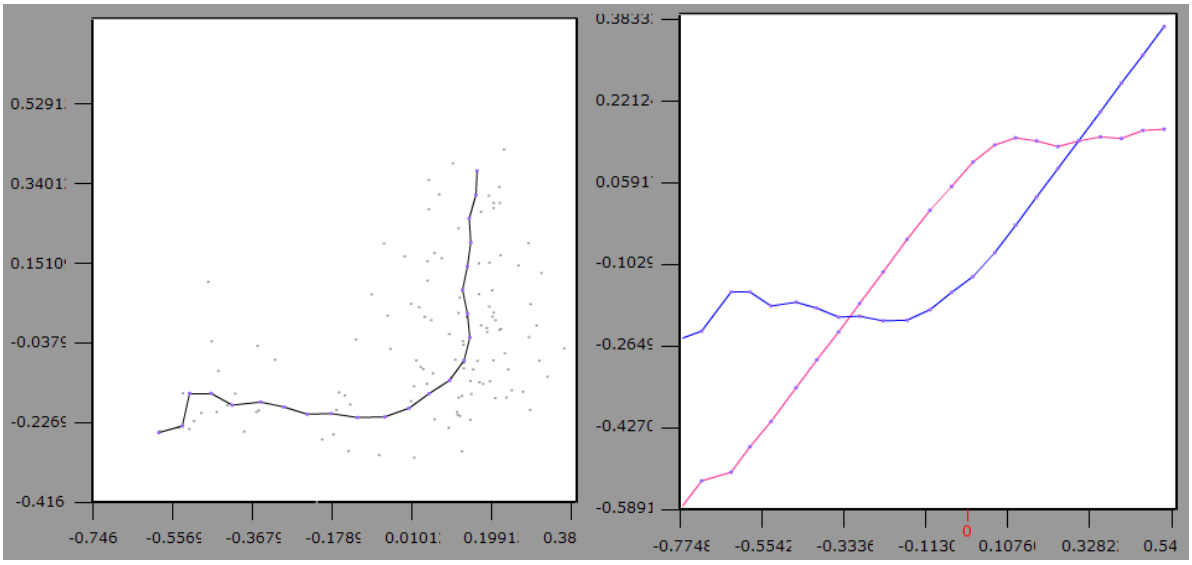


Figura 2.1 i 2.2: Corba PCOP de la relació d'expressió d'una parella de gens.
Protein: mdr1,mrp,p-glycoprotein-log (eix X) i SID 207193, ESTs [5o:H48865, 3o:H48588] (eix Y)

A la imatge de l'esquerra, el núbol de punts son les mostres de la microarray que proporcionen diferents valors d'expressió pels gens comparats(component x i y). La corba que es dibuixa, és la corba PCOP que descriu la relació d'expressió dels 2 gens.
A la figura de la dreta, es veu la relació del nivell d'expressió de cada gen en funció de la corba PCOP de la gràfica esquerra.

La figura 1.a i 1.b mostra una relació d'expressió tipo switch. Com es pot veure a la gráfica 1.b, el gen blau comença a sobrexpresar-se prop del punt 0, que és quan el gen rosa arriba a la seva màxima expressió.

Un altre tipus de patrons interessants biològicament, son els circulars.

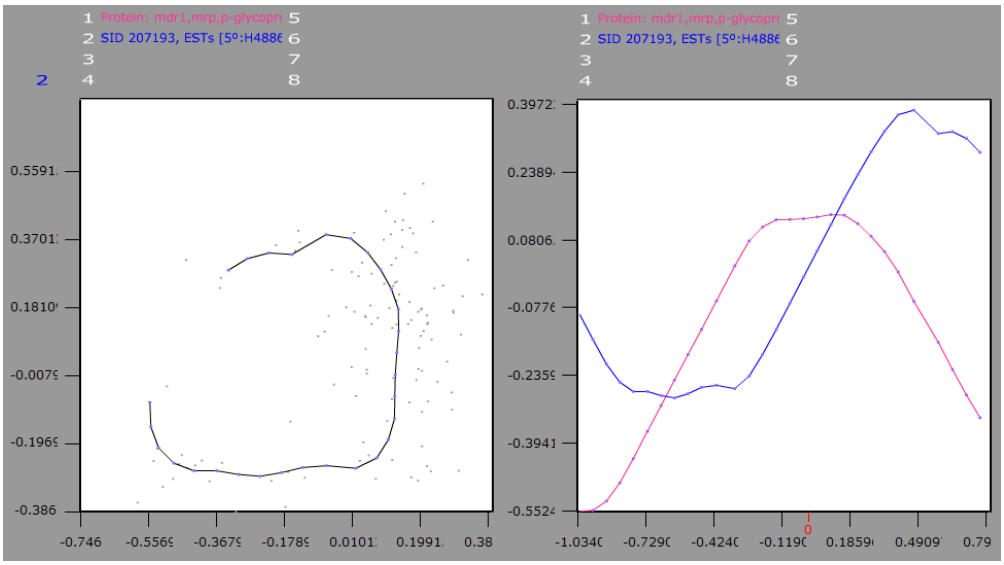


Figura 2.1 i 2.2: *Corba PCOP de la relació d'expressió d'una parella de gens.*
Protein: mdr1,mrp,p-glycoprotein-log (eix X) i SID 207193, ESTs [5o:H48865, 3o:H48588] (eix Y)
A la imatge de l'esquerra, els punts son la mostra que compara els 2 valors d'expressió (component x i y) dels gens relacionats. La corba que es dibuixa, és la corba PCOP que descriu la relació d'expressió dels 2 gens.
A la figura de la dreta, es veu la relació del nivell d'expressió de cada gen, en funció de la corba PCOP de la corba e la'esquerra.

En aquests patrons hi intervenen varis gens encara que només veiem 2. Les interaccions provocades pels canvis de nivells d'expressió d'aquests gens (tant els visibles com els que modulen pero no apareixen) provoquen aquest tipus de patrons.

Els patrons no-linals menys rellevants son aquells en que la relació dels gens és propera a la lineal X=Y o X=-Y . Aquestas corresponen als gens coexpressats (X=Y) o als gens inhividors (X=-Y) i son els estudiats pels mètodes de reducció i clustering clàsics.

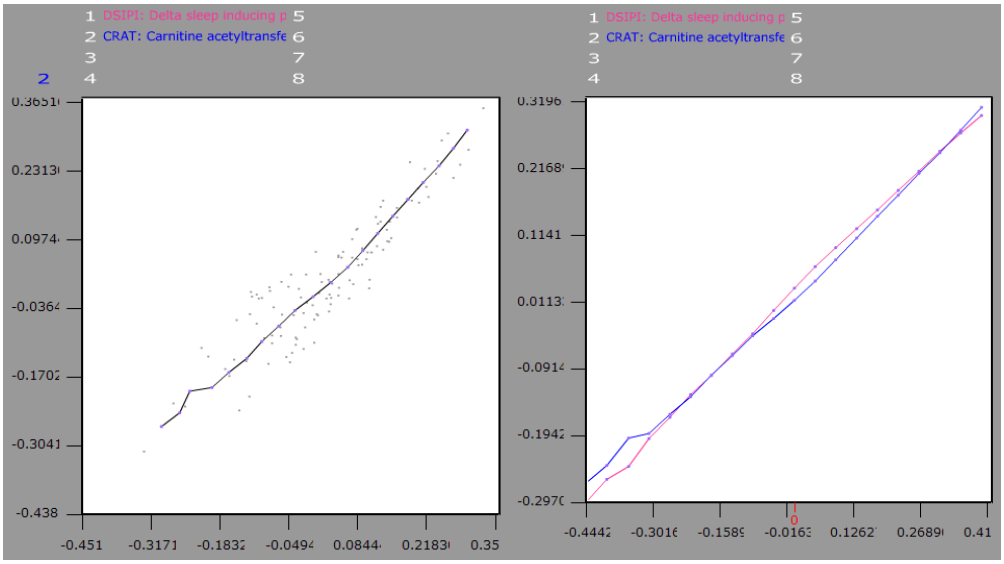


Figura 4.1 i 4.2: *Corba PCOP de la relació d'expressió d'una parella de gens*

GALNT2: UDP-N-acetyl-alpha-D-galactosamine:polypeptide (eix X) y RDH13: retinol dehydrogenase 13 (eix Y)

A la imatge de l'esquerra, el núbol de punts son les mostres de la microarray que proporcionen diferents valors d'expressió pels gens comparats(component x i y). La corba que es dibuixa, és la corba PCOP que descriu la relació d'expressió dels 2 gens.

A la figura de la dreta, es veu la relació del nivell d'expressió de cada gen en funció de la PCOP.

La figura 3 mostra la relació de 2 gens coexpressats. Es pot veure com a mida que un gen s'exrpessa, l'altre també ho fa.

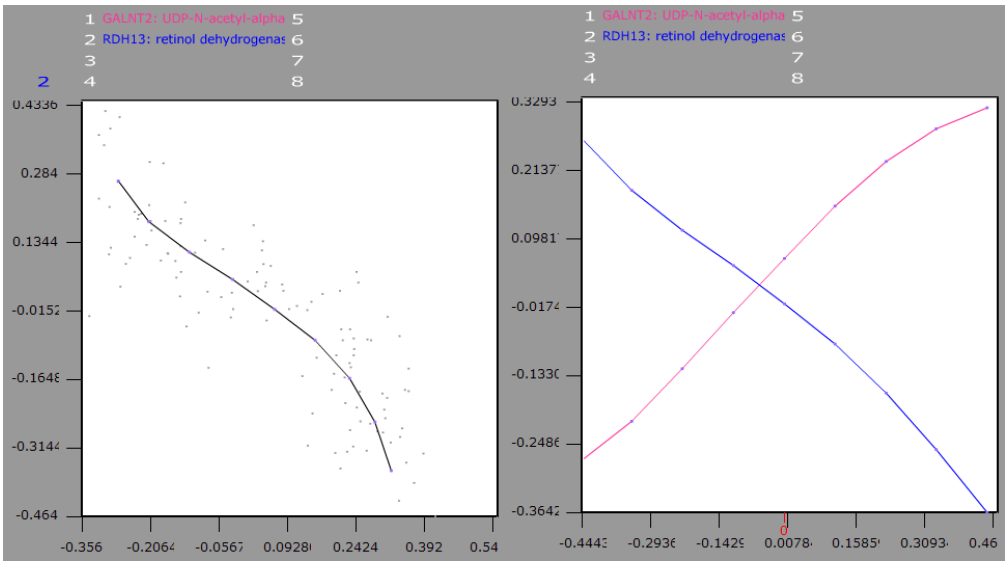


Figura 4.1 i 4.2: *Corba PCOP de la relació d'expressió d'una parella de gens*

GALNT2: UDP-N-acetyl-alpha-D-galactosamine:polypeptide (eix X) y RDH13: retinol dehydrogenase 13 (eix Y)

A la imatge de l'esquerra, els punts son la mostra que compara els 2 valors d'expressió (component x i y) dels gens relacionat. La corba que es dibuixa, és la corba PCOP que descriu la relació d'expressió dels 2 gens.

A la figura de la dreta, es veu la relació del nivell d'expressió de cada gen, en funció de la corba PCOP de la corba de l'esquerra.

En la figura 4 mostra la relació de 2 gens inhividors. Quan un 'expressa, l'atre s'infraexpresa, i a l'inrevés.

2 METODOLOGIA

2.1 MÈTODE D'AGRUPACIÓ DE DISTRIBUCIONS DE CLUSTERS

L'algorisme que s'ha dissenyat per trobar las ditribucions de clusters finals d'una microarray donada està basat en l'**aprenentatge per reforç.**

En aquest Anem construint una llista de solucions a mida que es va agrupant cada distribució de clusters, que a la vegada modifica la distribució arquetip del grup, és a dir la reforça, o en cas de no coincidir amb cap de les ja existents, en crea una de nova per si pot agrupar-la amb alguna posterior. Un cop s'han tractat totes les distribucions de clusters vinculades a una relació d'expressió no lineal, els arquetips son les distribucions de clusters finals, on les que ens interessen, son les que tenen més d'un reforç, es a dir, hi ha mínim 2 distribucions que la representen.

2.2 CÀLCULS PREVIS

Per al procés d'agrupació de clusters i trobar les seves distribucions finals, es necessita com a entrada les distribucions de clusters o subespais vinculats al càlcul de PCOPs.

Per trobar les diferents distribucions de clusters que seràn agrupades en distribucions finals, cal executar previament els diferents procesos:

2.2.1 CÀLCUL DE LES PCOPS

Les PCOP son unes corbes principals, que son corbes continues que passen a través d'un nubol multidimensional de dades de forma no el·líptica.

La Princial Curve of Oriented Points (PCOP) es definida usant la generalització a nivell local de las propietats de la variança en el cálcul per Components Principals. Aquesta generalizació dels CP a nivell local proporciona els Principal Oriented Points (POP). Y son els POPs obtinguts (PC a nivell local) els que constitueixen la PCOP o patroó intern del nubol de mostres [1].

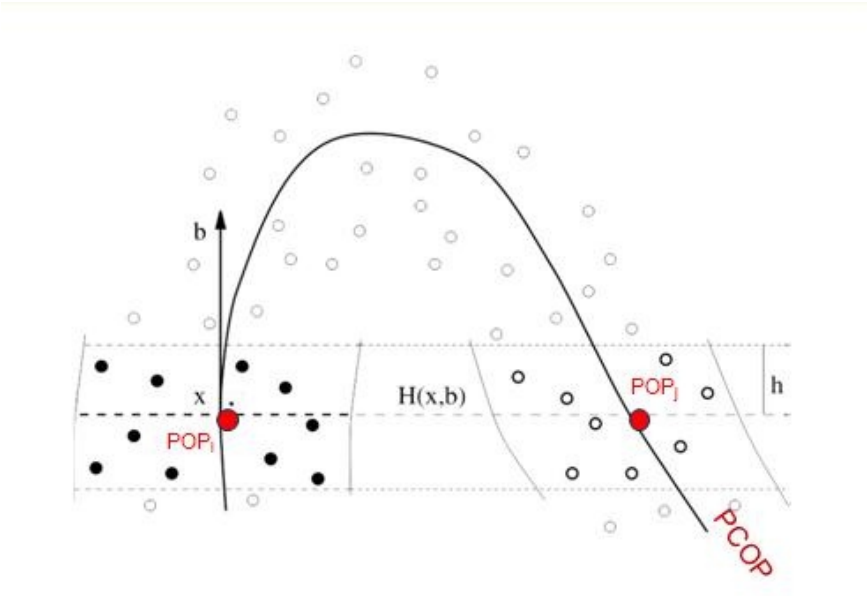


Figura 5. A la gràfica es poden observar el nubol de punts de l'espai mostral, El POPi i POPj calculats i la PCOP que pasa per aquestos dos POPs. Tant el POPi com el POPj son calculats pel seu propi subspai mostral aplicant-hi una generalització dels Components principal per cada respectiu subspai local. Els punts negres corresponen al cluster de mostres representades pel POPi i els punts blancs corresponen el nuvol de mostres representats pel POPj. B, sería la primera component principal del POPi, i H(x,b) n'es la segona [1].

Per cada parella de gens de la microarray, s'aplica llavors el càlcul de les PCOP. D'aquestta manera podem extreure els patrons de comportament no linial de totes les relacions d'expressió génica. Una dels punts forts del càlcul de las PCOP es la seva exactitut en el càlcul tant de la variancia al voltat de la corva com de la correlació entre les variables que compara (en aquest cas gens). Lo que ens serveix per a més d'esbrinar quins gens tenen una relación d'expressió no lineal, quins gens estàn altament correlacionats[3].

2.2.2 Obtenció de la poligonal de la corva

A partir de cada corba principal obtinguda amb el PCOP es detecten els POPs on hi ha un canvi significatiu de la corbatura. A partir d'ells podem obtenir el conjunt de comportaments locals corresponent a cada segment de la poligonal de la corva.

2.2.3 Creació de distribucions clusters

Com hem vist cada comportament local està representat per un segment de la poligonal de la corba. Cada segment de la corba es la recta entre dos punts de curvatura consecutius. I Cada POP es la discretització d'un conjunt de mostres a les que representa. Llavors podem obtenir les distribucions de clusters a partir dels samples associats als POPs de cada segment de la poligonal de la corba, es adir, els samples representats pels POPs entre el punts de curvatura consecutius. Aquest clusters de samples que conformen un comportament local de la relació d'expressió [4].

Notar que les distribucions de més d'un cluster, osigui de més d'un subespai mostral, ens indiquen que els nivells d'expressió dels gens involucrats mantenen una relació no linial.

2.3 AGRUPAMENT DE DISTRIBUCIONS DE CLUSTERS– CADENA DE PROCESSOS

Descrivim a alt nivell l'ordre d'execució de processos utilitzats per a l'agrupació de distribucions de clusters en distribucions de clusters finals.

El procés "lanzadora" és qui controla aquest ordre d'execució de processos.

El procés d'agrupació de distribucions clusters per l'obtenció de les distribucions finals segueix les seguents fases d'anàlisis:

Filtratge de les distribucions de clusters d'entrada en base a la relació d'expressió associada: Aplicació dels diferents filtres per seleccionar les distribucions de clusters a agrupar.

Tractament de les interseccions de clusters d'una mateixa distribució: Tractament de les mostres de les regions d'intersecció dels clusters en les distribucions a agrupar.

Càlcul del matching entre les diferents distribucions de clusters: Càlcul de semblança entre distribucions de clusters per determinar la seva equivalencia.

Assignació dels samples a un o altre cluster d'una distribució final: Càlcul utilitzat per determinar el valor mínim de repeticions de les mostres a cada cluster en les distribucions que s'agrupen per poder assignar-la al cluster final.

ALGORISME d'alt nivell utilitzat pel métode d'agrupació de distribucions de clusters per obtenir les distribucions finals:

[Filtratge de les distribucions de clusters d'entrada en base a la relació d'expressió associada]

[Per cada distribució de clusters d'entrada fem:]

[Tractament de les interseccions de clusters d'una mateixa distribució] (Fase ***CL. Intersection***)

[Càlcul del matching entre les distribucions de clusters. Es busca en les distribucions de clusters obtingudes fins al moment, una de semblant a aquesta.] (Fase ***% matching***)

[Si es troba una distribució semblant, l'afegim a la distribució temporal per reforçar-la i informar dels gens que intervenen en aquesta distribució de clusters]

[Si no es troba una distribució semblant, s'afegeix aquesta com a nova distribució temporal a tenir en compte per als següents distribucions de clusters a agrupar]

[Fi del bucle d'agrupació de clusters]

[Càlcul d'errors de cada distribució de clusters]

[Obtenció de les distribucions de clusters finals. Per a cada agrupacio de distribucións de clusters, es fa la assignació dels samples a clusters de la distribució final que representarà al grup] (fase ***d'assignació de samples***)

Guardem a fitxer les distribucions de clusters finals

2.4 FILTRATGE DE LES DISTRIBUCIONS DE CLUSTERS D'ENTRADA EN BASE A LA RELACIÓ D'EXPRESSIÓ GÈNICA ASSOCIADA:

Determina quines son les distribucions de clusters a agrupar. Aquestes distribucions estan separades segons els grau de correlació de les corbes definides per cada parell de gens i pel grau de corvatura de la corva. El grau de correlació de la corba es defineix a partir de la variança entre el patró de la corba i els punts que la defineixen. Una alta correlació té una variança baixa, i les baixes correlacions tenen una variança alta.

En el preprocés es separen les corbes segons el seu factor de correlació i pel grau de corvatura de la corva. Aquesta separació, es fa segons 4 criteris:

Relacions no lineals d'alta correlació (non lineal relationship with high correlation):

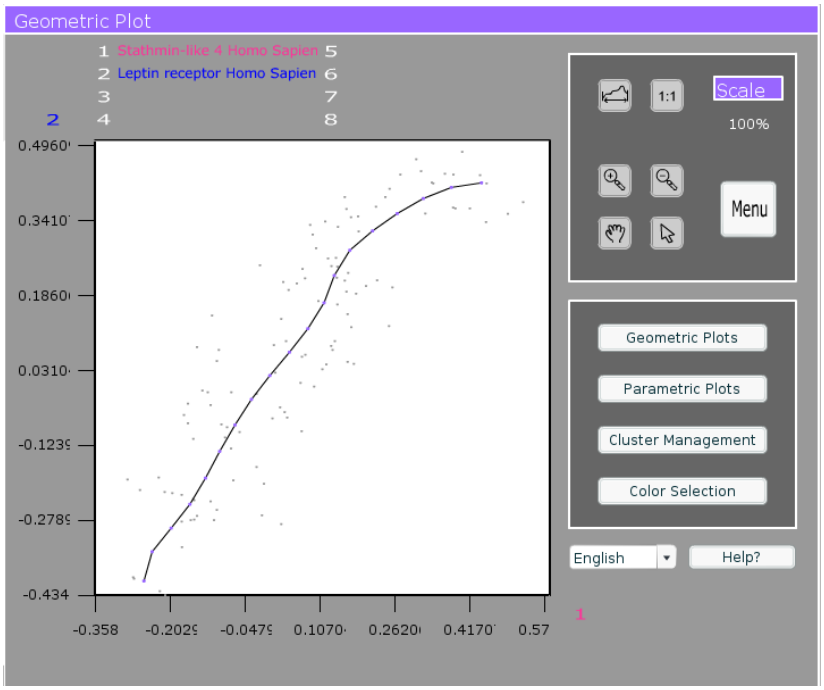


Figura 7 PCOP de la relació d'expressió de 2 gens.

Stathmin-like 4 Homo Sapiens, 153 sequence(s) (eix X) - Leptin receptor Homo Sapiens, 657 sequence(s) (eix Y)

El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens.

A la figura es pot apreciaar que les dades s'ajusten a la PCOP trobada donc existeix una alta correlació entre els gens.

La variança entre la corba i les mostres que la defineixen determina el grau de correlació de les expresions. Es pot veure com en aquesta figura 2.1, les mostres s'ajusten molt be a la trajectòria de la PCOP. No obstant, Com que la curvatura de la corva sol exigir més soroll, la curvatura de les relacions de tan alta correlació no será gaire gran.

Relacions no lineals de correlació mitja (non lineal relationship with medium correlation):

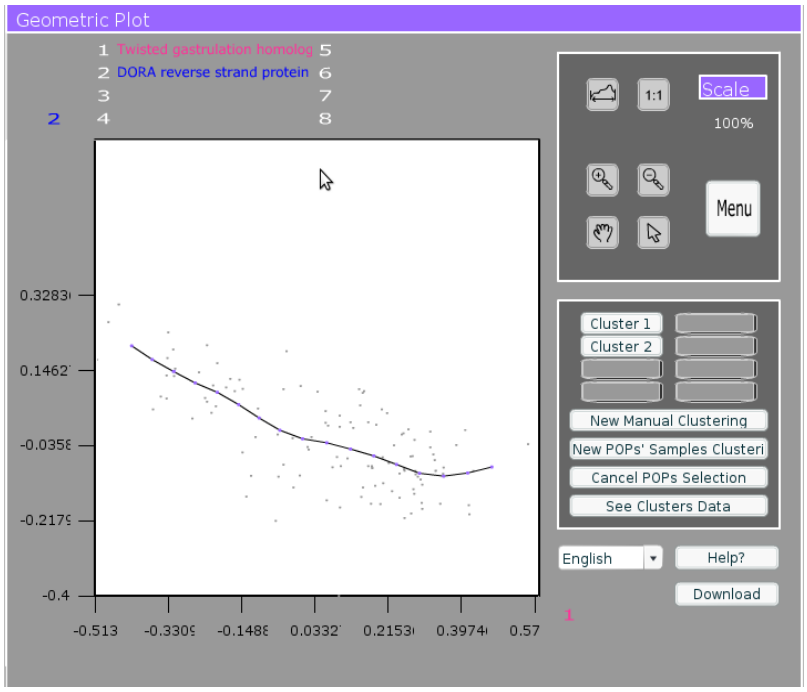


Figura 8 PCOP de la relació d'expressió de 2 gens.

Twisted gastrulation homolog 1 (Drosophila) Homo Sapiens, 254 sequence(s) (eix X)

DORA reverse strand protein 1 Homo Sapiens, 590 sequence(s) (eix Y)

El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens.

A la figura es pot apreciaar que les dades s'ajusten menys a la PCOP que en la fig. 7.

En aquest cas els gens estàn menys correlacionats.

Es pot veure en la figura 8, com els punts que defineixen la corba PCOP tenen més dispersió que a la figura 7 (PCOP d'alta correlació). Aquest tipus de corba, es classifiquen com a corbes de correlació mitja i permeten curvatures més grans.

Relacions no lineals de correlació mitja i alta corbatura (non lineal relationship with medium correlation and high curvature): Aquest tipus de corba és un subconjunt de les corbes PCOP de correlació mitja, però amb una corbatura apreciable.

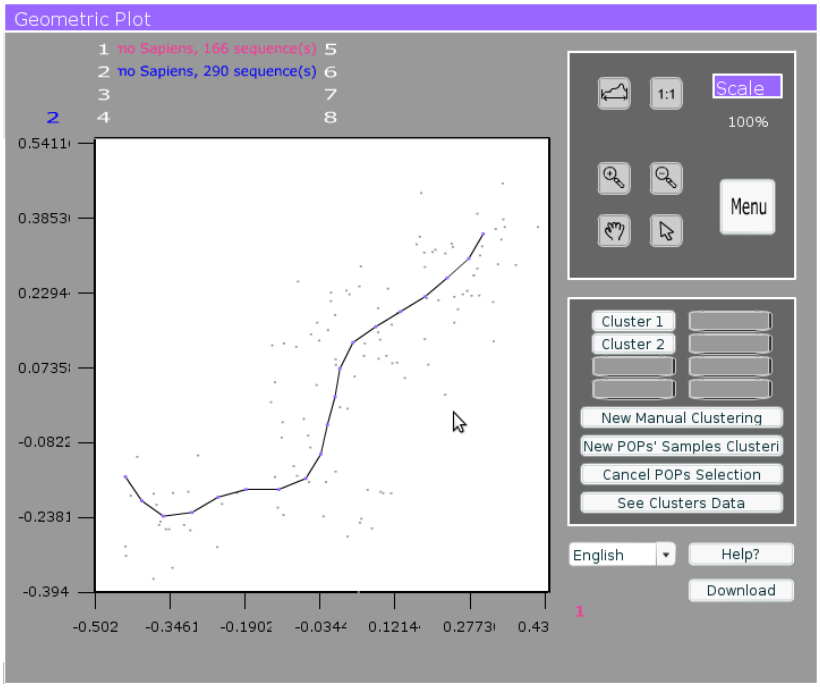


Figura 9 PCOP de la relació d'expressió de 2 gens.
Cryptochrome 1 (photolyase-like) Homo Sapiens, 166 sequence(s) (eix X) -
Pre-B-cell leukemia transcription factor 3 Homo Sapiens, 290 sequence(s) (eix Y)
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens.

En aquest cas les dades s'ajusten a la PCOP com en la figura 8.

Relacions no lineals de baixa correlació i alta corbatura (non lineal relationship with low correlation and high curvature): Com es pot veure en gràfic 9 el núvol de punts és encara més dispers que en la selecció per graus de correlació més alts. Aquest permet grans curvatures, per això només seleccionem les corbes de baixa correlació amb alta corbatura. Perdem significancia en el grau de dependència dels gens, pero ganem significancia en la rellevància en la intensitat del canvi fenotípic que representen.

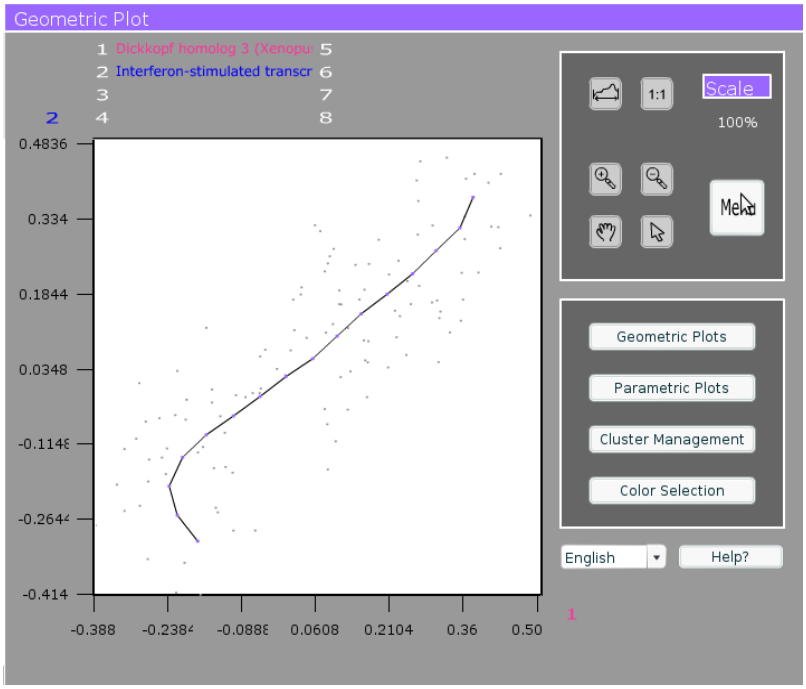


Figura 10 PCOP de la relació d'expressió de 2 gens.
Dickkopf homolog 3 (Xenopus laevis) Homo Sapiens, 520 sequence(s)
Interferon-stimulated transcription factor 3, gamma 48kDa Homo Sapiens, 242 sequence(s)
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens.

A la figura es pot apreciaar que les dades s'ajusten menys a la PCOP que en la fig. 8 y 9 . En aquest cas els gens estàn menys correlacionats.

2.5 TRACTAMENT DE LES INTERSECCIONS DE CLUSTERS A D'UNA MATEIXA DISTRIBUCIÓ

La major part de distribucions de clusters a agrupar, tenen mostres assiganades a més d'un cluster. Això obliga, abans de fer l'agrupament de cada nova distribució de clusters, determinar a quin cluster de la distribució d'entrada assignem aquestes mostres. Una vegada totes les mostres son asignades a un únic cluster ja es pot comparar la distribució d'entrada amb les diferents distribucions temporals trobades.

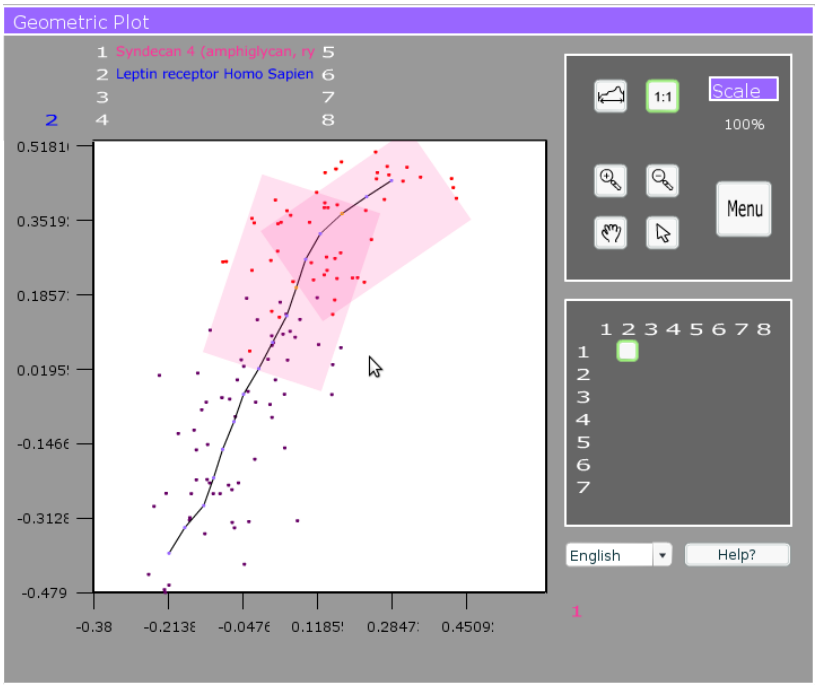


Figura 11 PCOP de la relació d'expressió de 2 gens. Syndecan 4 (amphiglycan, ryudocan) Homo Sapiens, 454 sequence(s)Leptin receptor Homo Sapiens, 657 sequence(s). Les mostres apareixen pintades de dos colors per diferenciar els dos clusters de la distribució de clusters. Les arees roses senyalen els subspais mostrals per a dos POP que pertanyen cadascún a un cluster diferent. Com es pot observar aquestes arees es solapen, amb lo que hi hauràn mostres que poden pertanyer als dos clusters.

A la figura 11 es pot veure com hi ha una zona comú de mostres assigandes a 2 POPs de la corba PCOP definida per les expressions de 2 gens. Aquesta zona és just on es produceix el canvi d'orientació de la corba, que definirà els 2 clusters de la PCOP, i on es definiran quines mostres van assignades a un cluster o un altre. Les mostres assignades a ambdos clusters, l'anomenem zona d'intersecció de clusters.

Les dues interpretacions utilitzades per aquests casos, que generen resultats diferents son:

• Descartat de mostres en zones d'intersecció: Ignorar aquestes mostres en el matching i comparar només aquelles que apareixen a un sol cluster.

• Assignació de les mostres en zones d'intersecció al cluster més petit: Assignar les mostres de les zones d'intersecció al cluster més dèbil o absorbit. D'aquesta manera també tenim totes les mostres només en un únic cluster. Aquest criteri intenta compensar el pes dels clusters amb un major nombre de mostres mitjançant l'assignació de la mostra repetida sempre al cluster més petit on està representada. Aquests clusters els anomenem absorbits.

El tractament sobre les distribucions de clusters temporals que es dona a cada interpretació de la zona d'intersecció, és diferent. Quan es descarten les mostres de la zona d'intersecció, es fa una interpretació per màxims (mètode explicat en el punt 2.4.3) de les distribucions temporals que utilitzem per agrupar . En el cas de fer assignació de mostres al cluster més petit, la interpretació que es fa sobre la distribució de clusters temporals és per T-Student (mètode explicat en el punt 2.4.3). Aquesta interpretació de la distribució temporal (segons la zona d'intersecció) quan comparem amb noves distribucions de clusters és útil per poder avaluar diferents resultats al mètode.

2.6 CÀLCUL DEL MATCHING ENTRE LES DISTRIBUCIONS DE CLUSTERS

Aquest análisi determina el percentatge que existeix entre dues distribucions de clusters (la que s'intenta agrupar i les distribucions intermitja que representa un conjunt d' agrupades) per poder determinar si son semblants o no.

Per veure la semblança de 2 distribucions de clusters, utilitzem una matriu quadrada de dimensions NxN, on N és el número de clusters de cada distribució.

La primera condició per a comparar 2 distribucions de clusters, és que el número de clusters de cada una sigui igual.

Les columnes de la matriu representen els clusters de la distribució que s'intenta agrupar.

Les files de la matriu representen els clusters de la distribució final amb la que es compara.

Cada cel·la de la matriu, conté el nombre de mostres que coincideixen entre C1x i C2y, on C1x és un cluster “x” de la distribució de clusters C1, i C2y és un cluster “y” de la

distribució de clusters C2.

"x" i "y" tenen rang [1..N].

Un cop es té calculada la matriu, es calculen els valors percentuals de les cel·les respecte al total de samples de cada fila, es a dir, s'obtenen els percentatges de correspondència entre la distribució de clusters a agrupar, i la distribució de clusters intermitja amb que es compara.

Si **cada fila** de la matriu té una **única cel·la** que supera el criteri "% matching", i **cada columna** té també una **única cel·la** que supera aquest criteri, es considera que les distribucions de clusters son semblants.

Mostrem a continuació un *exemple* en el que comparem 2 distribucions de clusters.

Donada la distribució de clusters a agrupar a C1:

C11:	X				X	X			X	X	X						X	X		X		X	X				= 11 mostres
C12:		X	X	X			X	X				X	X	X	X	X			X		X			X	X	X	= 15 mostres

la comparem amb la distribució final de clusters a C2:

C21:	X				X	X			X	X	X	X		X			X	X		X		X	X				= 13 mostres
C22:		X	X	X			X	X					X		X	X			X		X			X	X	X	= 13 mostres

La matriu de semblança calculada per aquestes 2 distribucions és:

	C11	C12					C11	C12
C21	11	2	->	13 samples	->	C21	**85%**	15%
C22	0	13	->	13 samples	->	C22	0%	**100%**

Per a un criteri de semblança del 75%, veiem que existeixen 2 columnes i 2 files amb percentatges superiors a aquest valor, per tant les dues distribucions son semblants.

Aquesta matriu de semblança determina que el cluster C11 a agrupar es correspon amb el cluster C21 de la distribució final, i que el cluster C12 es correspon al cluster C22.

Si el criteri fos més exigent, com per exemple un 90% de semblança, veuriem que no podem considerar-los com a semblants, ja que no existeixen 2 columnes i 2 files amb valors percentuals superiors a aquest valor.

Aquest métode és independent al número de clusters a agrupar i al número de samples de l'espai mostral.

2.7 ASIGNACIÓ DELS SAMPLES A UN O ALTRE CLUSTER DE LA DISTRIBUCIÓ FINAL

Un cop s'han agrupat totes le distribucions de clusters, cal crear les distribucions finals tenint en compte que cada representa a distribucions de clusters semblants pero no iguals.

Les diferents distribucions finals s'obtenen de la fussió de les distribucions de clusters semblants que formen cada una. Cal utilitzar un criteri de selecció de samples, segons el número de repeticions de cada una en els diferents clusters als que pot pertànyer, per assignara-la a un cluster o un altra, segons si supera o no aquest número de repeticions.

Els 2 criteris utilitzats per a fer aquesta assignació de samples son:

- Assignació per **màxims** (MAX): Assignem cada sample al cluster on aquesta es repeteix més, i en el cas d'empat, s'assigna al cluster absorvit.

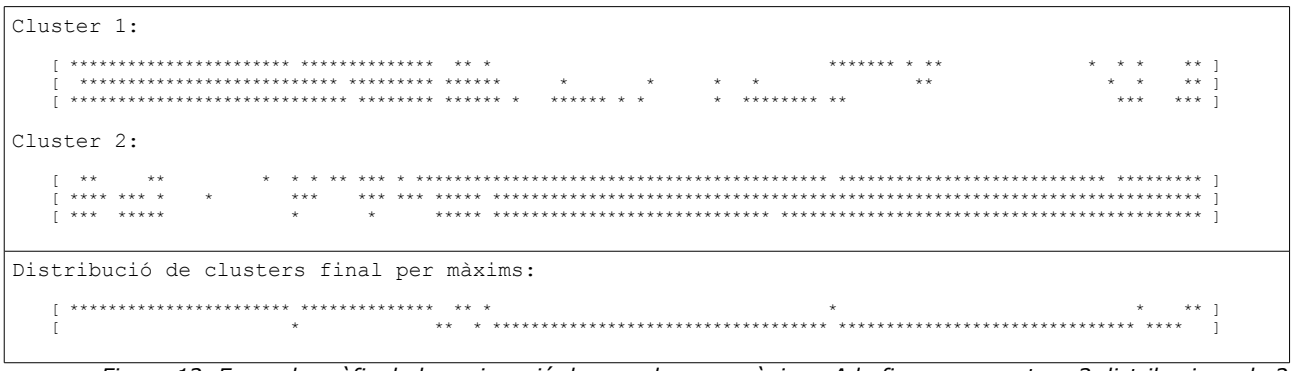


Figura 12. Exemple gràfic de la assignació de samples per màxims. A la figura es mostren 3 distribucions de 2 clusters que s'agrupen per haber sigut considerades similars. Les distribucions es mostren dividides en el cluster 1 i cluster 2, on la mateixa fila al cluster 1 i la mateixa fila al cluster 2 indiquen que els dos clusters pertanyen a la mateixa distribució d'entrada. Les columnes representen les condicions mostrals de la microarray, amb lo que si apareix un '' a una cela, significa que aquesta mostra pertany aquest cluster per a aquesta distribució donada (la fila).Notar que un mateix sample pot apareixer en els dos clusters per a la mateixa distribució. L'ultima cela de la taula mostra la distribució de clusters final obtinguda a partir de les 3 distribucions de clusters que s'han agrupat. L'assignació final dels samples s'ha fet pel nombre d'ocurrencies a cada cluster (l'opció més reforzada). En aquesta distribució final cada fila representa un dels 2 clusters.*

A la figura 12, es veu com 2 clusters d'una agrupació de 3 distribucions de clusters d'entrada assignen les samples a la distribució final per el numero màxim d'ocurrencies. Es a dir, cada sample s'assigna al cluster en el que ha obtingut més reforç.

- Assignació per **T-Student**: L'existencia de mostres que no es defineix clarament quin és el cluster al que pertanyen, degut a que el número de repetcions de la sample en un i l'altre cluster és força representatiu, fa que sigui necessari buscar altres criteris d'assignació. Aquest criteri l'utilitzem per intentar corregir la tendència d'assignació de samples sempre al cluster predominant quan en la majoria de casos, les mostres que realment apareixen numeroses vegades tant en un cluster com en un altre es perque pertanyen a la zona d'intersecció, i la zona d'intersecció del cluster més petit pot arribar a ser tot el cluster.

  La assignació de samples a un cluster per T-Student funciona de la següent manera:

  Un paràmetre determinat en el càlcul de la variança de la T-Student és el valor de α . Degut al desconeixament dels resultats que s'obtindrien, s'ha optat per fer la agrupació per a diferents valors estàndards i poder estudiar els resultats per decidir després quins donen millor resultat.

  L'espai mostral que es crea per poder fer el càlcul de la variança per T-Student, son els diferents valors que tenen els comptadors de repeticions o reforços per a cada cluster.

  Mostrem a continuació un *exemple* en el que utilitzem la sel·lecció per variança per determinar la distribució de clusters final.

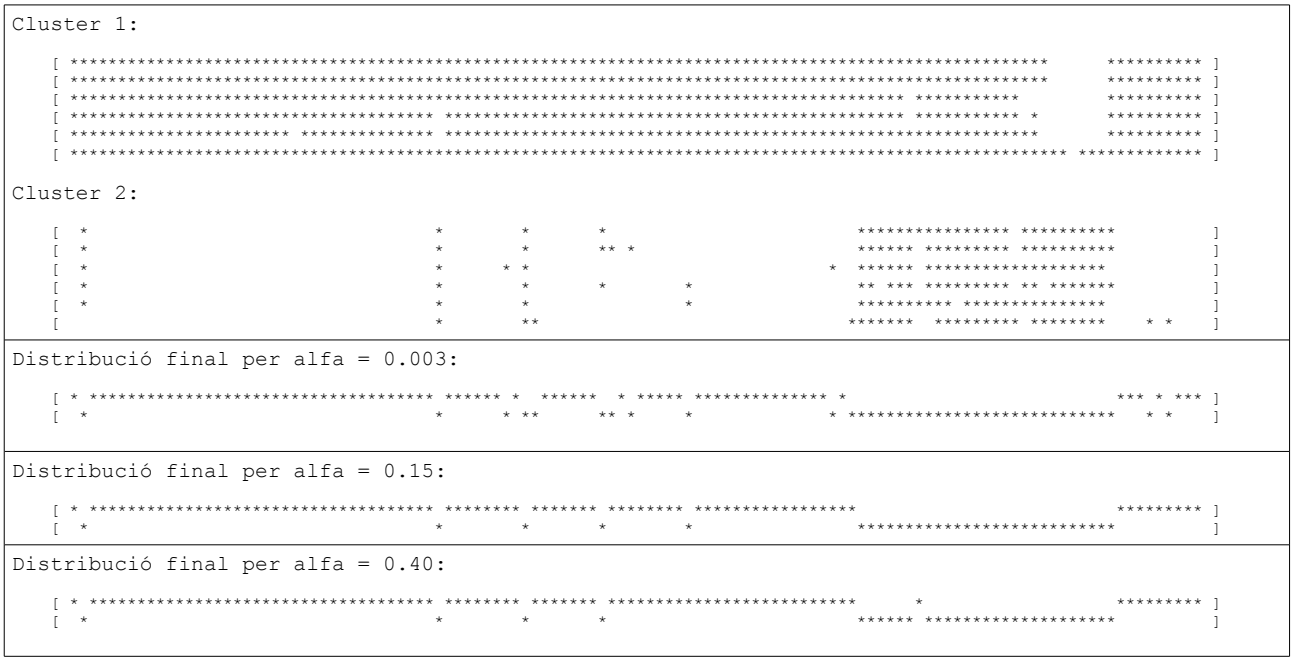


Figura 13. Exemple d'assignació de samples al cluster de la distribució final per a diferents valors d' α .A la figura es mostren 6 distribucions de 2 clusters que s'agrupen per haber sigut considerades similars. La mateixa fila al cluster 1 y al cluster 2 indique que aquests dos clusters pertanyen a la mateixa distribució de

entrada. Les columnes representen les condicions mostrals de la microarray, si apareix un '' vol dir que aquesta mostra pertany aquest cluster. Les caixes inferiors, es corresponen amb la distribucio de clusters final obtinguda de les mateixes distribucions d'entrada pero amb diferents evalor d'* α *.Cada fila representa un cluster diferent.*

A la figura 13 es pot veure com la assignació de samples a clusters per a diferents valors d' α *dona distribucions de clusters finals diferents.*

Els càlculs per a un $\alpha = 0{,}40$ del grup de distribucions de clusters de la figura 13 és el que segueix:

Calculem primer l'interval de confiança per T-Student del cluster més petit, C2:

El número de mostres de l'espai mostral és 5, ja que hi ha repeticions de 1,2,3,5 i 6 samples (no n'hi ha cap de 4). Per tant, els càlculs intermitjos de la T-Student son:

$N = ٥$: Número de mostres de l'espai mostral de la T-Student.

$\sum x = ١٩$: Sumatori del valor de totes les mostres

$\hat{x} = 3.8$: Mitja de les mostres

$\sum x^2 = 87$: Sumatori de les mostres al quadrat

$(\hat{x})^2 = 14.44$ Mitja de les mostres al quadrat

$S = 1.720465$ Desviació

$S^2 = 2.96$ Variança

$\hat{S} = 1.720465$ Desviació típica

$\alpha = 0.40 \rightarrow (1 - (\alpha / 2)) = 0.60$ Valor probabilístic que determinem

$T_{inv(tn-1;1-\alpha/2)} = 0.940965$

$m_1 = \hat{x} - (\frac{T_{inv} * S^2}{\sqrt{N}}) = 3.076007$ Marge inferior o número de mostres mínim que assegura el valor probabilístic de α .

Per tant, seleccionem per al cluster 2 de la figura 13 les samples que estan reforçades més de 3.076 vegades, es a dir, a partir de 4 repeticions.

Amb les samples que queden lliures després de la primera assignació, es torna a calcular aplicant la T-Student el marge inferior per admetre una mostra del següent cluster més

petit, i així successivament fins que no hi hagi més clusters a calcular.

A l'exemple de la figura X, el cluster 1 tindria un marge inferior $m_1 = 1.925065$, es a dir, agrupa les samples en aquest cluster a partir de 2 repeticions.

En l'apartat de Resultats i Discussió es podrà veure com influeix el valor de α en la configuració de la distribució de clusters final.

El tipus d'assignacio per màxims o t-student està vinculat amb el tractament de les mostres de les interseccions dels clusters (secció 2.4) i del reforc que s'aplica a la distribució de clusters temporals que representa cada agrupació de distribucions (secció 2.5). Quan el mètode escollit es descartar les mostres de la zona d'intersecció(secció 2.4), es fa una asignació de les mostres als clusters per màxim nombre de repeticions. Així formem les distribucions temporals que utilitzarem per agrupar. En el cas de fer assignació de les mostres de la intersecció al cluster més petit, l'assignació de mostres triat per asignar les mostres a la distribució final serà el metode empreat per obtindre la distribució de clusters temporal.

3 RESULTATS I DISCUSSIÓ

Tots els resultats d'aquesta memòria estan calculats a partir d'una microarray elaborada pel National Cancer Institute (NCI, USA), amb dades corresponents a les expresions de 9703 cDNAS, que representen aproximadament 800 gens únics expressats en 60 línies cel·lulars després d'administrar-hi 1.400 components químics. D'aquestes dades, s'han seleccionat les expressions ponderades de les 60 línies per a 1416 gens i 200 substàncies anti-tumorals.

3.1 PREPROCÉS

Com ja s'ha explicat a la secció de metodologia, donat un conjunt de distribucions de clusters, obtinguts a partir dels segments de la poligonal de les corves PCOP que representen les relacions d'expressió entre els gens de la microarray, es calculen unes distribucions finals de clusters, que es proporcionen amb tots els gens que participen en el mateix canvi fenotípic descrit per la distribució de clusters final del grup.

El programa d'agrupament de distribucions de clusters i obtencio de les distribucions finals (secció 2) formarà part del preprocés en l'anàlisis de microarrays, lliurant del seu elevat cost computacional a l'interfice web. Amés els seus resultats podràn ser utilitzats per a nous analisis.

L'algorisme utilitzat per al càlcul de distribucions finals es basa en l'aprenentatge per reforç, i consisteix en comparar totes les distribucions de clusters amb una llista de distribucions intermitjes o temporals inicialment buida. Aquesta llista és dinàmica, i creix a mida que es troben noves distribucions intermitjes.

		Non-lineal relationships with high correlation			Non-lineal relationships with medium correlation			Non-lineal relationships with low correlation		
	% Matching	#Distribució clusters	#Distribució clusters final	#distribucions agrupades	#Distribució clusters	#Distribució clusters final	#distribucions agrupades	#Distribució clusters	#Distribució clusters final	#distribucions agrupades
Descarting joint samples	60,00%	146	9	110	4430	69	3503	8861	114	6719
	65,00%	146	10	108	4430	75	3433	8861	143	6580
	75,00%	146	14	99	4430	97	3150	8861	190	5815
	85,00%	146	13	77	4430	114	2900	8861	208	4936
	90,00%	146	13	55	4430	135	2600	8861	213	4389
	95,00%	146	7	20	4430	218	1628	8861	318	2764
Assingning joint samples to the smaller cluster	60,00%	146	11	117	4430	281	3987	8861	714	7812
	65,00%	146	10	112	4430	294	3872	8861	728	7513
	75,00%	146	14	102	4430	274	3550	8861	683	6636
	85,00%	146	22	81	4430	330	2959	8861	657	5187
	90,00%	146	25	66	4430	486	2433	8861	857	3996
	95,00%	146	8	20	4430	386	1073	8861	562	1467

Taula 14: Taula de resultats obtinguts per a 3 filtres diferents en les distribucions de clusters d'entrada diferents. Per a cada filtratge de distribucions d'entrada, mètode de tractament de les zones d'intersecció i percentatge de matching, ens dona el número de distribucions d'entrada que s'han intentat classificar, el número de distribucions de clusters final trobades, i el número de distribucions de clusters d'entrada que s'han aconseguit agrupar amb d'altres distribucions de clusters.

La taula 14 mostra els resultats de les distribucions agrupades amb totes les combinacions possibles de càlcul utilitzades en el procés d'agrupació de clusters, segons els diferents criteris configurables:

Les columnes "#Distribució clusters entr." són el número de distribucions de clusters que s'han agrupat.

Les columnes "#Distribució cluster final" son el número de distribucions finals de clusters que representen a les distribucions de clusters agrupades.

Les columnes "#clusters arrangements" son el número de distribucions de clusters que s'han pogut agrupar en una distribució final (algunes distribucions queden soles sense agrupar, considerant llavors que aquests canvis fenotípics son poc representatius donç involucren poques relacions de gens).

	% Matching	high corr.	meduim corr.	low corr.
Descarting joint samples	60,00%	75,34%	79,07%	75,83%
	65,00%	73,97%	77,49%	74,26%
	75,00%	67,81%	71,11%	65,62%
	85,00%	52,74%	65,46%	55,70%
	90,00%	37,67%	58,69%	49,53%
	95,00%	13,70%	36,75%	31,19%
	% Matching	high corr.	meduim corr.	low corr.
Assingning joint samples to the smaller cluster	60,00%	80,14%	90,00%	88,16%
	65,00%	76,71%	87,40%	84,79%
	75,00%	69,86%	80,14%	74,89%
	85,00%	55,48%	66,79%	58,54%
	90,00%	45,21%	54,92%	45,10%
	95,00%	13,70%	24,22%	16,56%

Taula 15: Taula de percentatge de distribucions agrupades en funció del filtre aplicat a les distribucions d'entrada, mètode de tractament de les zones d'intersecció i percentatge de matching desitjat per agrupar distribucions.

De la taula 15 en podem extreure una comparativa gràfica per tractament de zones d'intersecció. Aquesta comparativa la podem veure a les Taules 16 i 17.

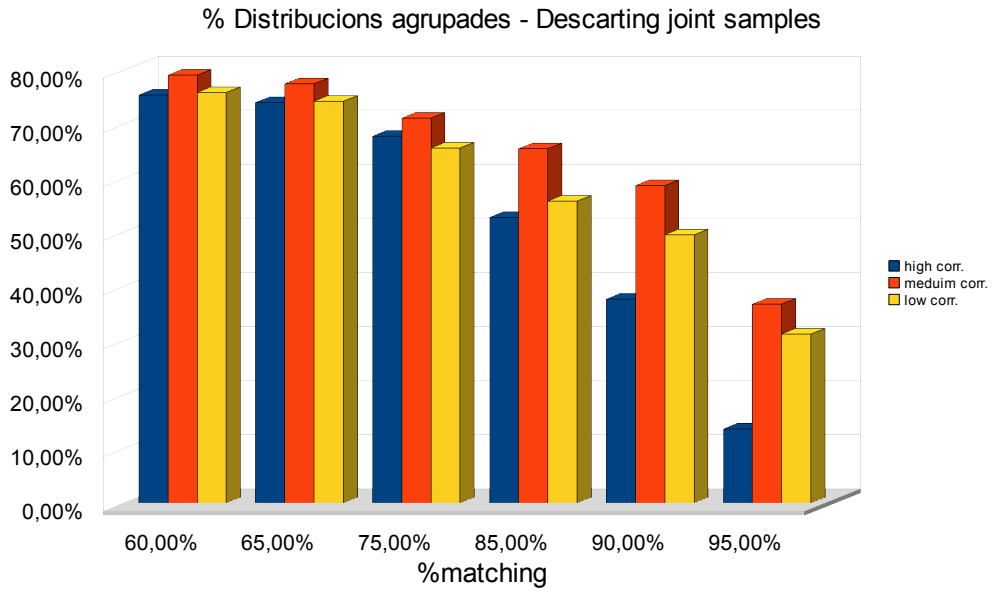


Figura 16: Evolució de percentatges de distribucions de clusters agrupades segons 3 filtres d'entrada diferents, i en funció del %matching. Resultats extrets de la taula 15

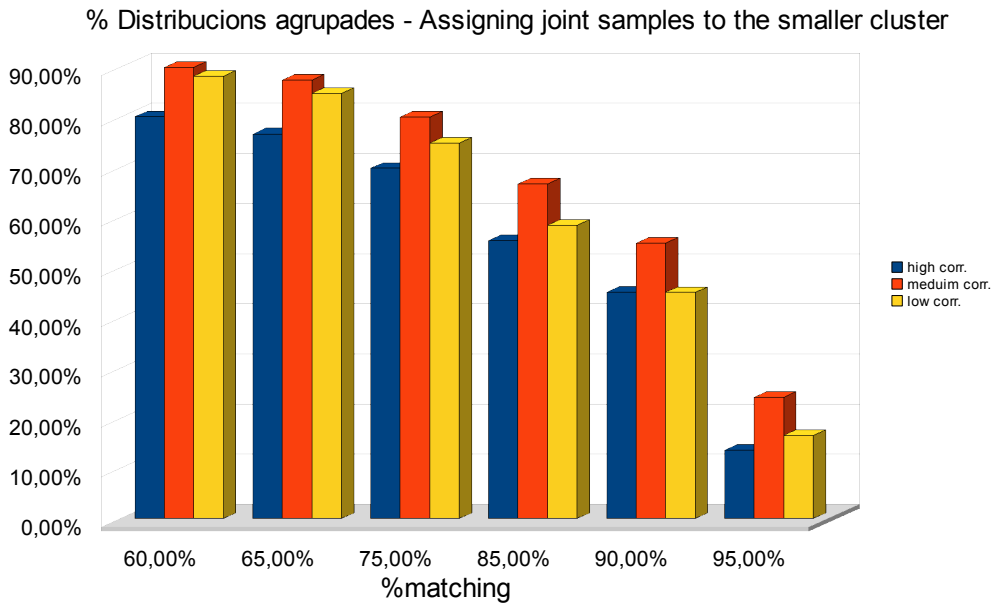


Figura 17. Evolució de percentatges de distribucions de clusters agrupades per a 3 tipologies de distribucions clusters, en funció del %matching. Resultats extrets de la taula 15

Dels resultats de la taula 3.2 representats en la taula 3.3 i 34, podem veure en com el percentatge de matching exigit influeix en el número de distribucions de clusters que s'agrupen. Conforme més alt és el valor, més exigent és el mètode per poder agrupar clusters i per tant, menys distribucions agrupa. Pel que es pot veure, el tipus de tractament de les zones d'intersecció influeix però poc en el número de distribucions classificades.

També es pot veure com el percentage de distribucions de clusters agrupades és manté similar per a les 3 tipologies de distribucions de clusters, encara que sempre el percentatge és més alt per als clusters de corbes PCOP de correlació mitja.

A més, lo realment important es la relevancia biológica de les distribucions de clusters finals. I això depent més de la cuantitat de distribucions agrupades sota una mateixa distribució final de clusters, així com la cuantitat de distribucions finals resultants. Lo primer significa canvis fenotípcs en que es veuen implicats un major número de gens, i lo segón que s'han detectat un major número de canvis fenotípics. A més una distribució de clusters final es torna més estable cuan major hagi sigut el nombre de distribucions d'entrada que han participat en el reforç.

		Non-lineal relationships with high correlation			Non-lineal relationships with medium correlation			Non-lineal relationships with low correlation		
	% Matching	#Distribució clusters	#clusters agrupats	Màxima agrupació	#Distribució clusters	#clusters agrupats	Màxima agrupació	#Distribució clusters	#clusters agrupats	Màxima agrupació
Descarting joint samples	60,00%	146	110	31	4430	3503	1173	8861	6719	966
	65,00%	146	108	28	4430	3433	1077	8861	6580	1585
	75,00%	146	99	36	4430	3150	556	8861	5815	896
	85,00%	146	77	11	4430	2900	388	8861	4936	823
	90,00%	146	55	12	4430	2600	297	8861	4389	826
	95,00%	146	20	7	4430	1628	110	8861	2764	217
Assingning joint samples to the smaller cluster	60,00%	146	117	32	4430	3987	1160	8861	7812	2313
	65,00%	146	112	52	4430	3872	1347	8861	7513	1907
	75,00%	146	102	35	4430	3550	578	8861	6636	1438
	85,00%	146	81	9	4430	2959	250	8861	5187	457
	90,00%	146	66	4	4430	2433	175	8861	3996	220
	95,00%	146	20	6	4430	1073	14	8861	1467	18

Figura 18: Taula de resultats obtinguts per a 3 filtres diferents de distribucions de clusters d'entrada diferents. Per a cada filtratge de distribucions d'entrada, mètode de tractament de les zones d'intersecció i percentatge de matching, ens dona el número de distribucions d'entrada que s'han intentat classificar, el número de distribucions de clusters d'entrada que s'han aconseguit agrupar amb d'altres distribucions de clusters, i el número màxim d'agrupacions aconseguides.

Ens em trobat amb seriosos problemes a l'hora dissenyar el mètode d'agrupació de distribucions degut al gran soroll inherent a les distribucions de clusters d'entrada. En un principi, la agrupació de distribucions de clusters només havia de tenir com a paràmetre el %matching, i la classificació final de samples havia de ser per màxims. Amb els primers resultats obtinguts, ja es va veure que es necessitaven altres criteris en el procés de agrupació, per corregir el fet que les distribucions de clusters grans tinguessin un pes més específic a l'hora de comparar-les amb altres i definir els clusters resultats. La selecció de samples per T-Student ha permès corregir aquest efecte a la sortida. El problema principal ha sigut que les distribucions de clusters d'entrada comptaven amb un gran nombre de mostres amb un cluster no ben definit i que pertenyien a la intersecció de clusters. Per això s'ha agut de disenyar un tractament especial a aquestes dades, tant a l'hora de comparar les distribucions, com a l'hora d'aplicar els reforços a les distribucions intermitjes, com a l'hora de de clasificar les mostres a les distribucions finals de clusters.

3.2 CÀLCUL DE L'ERROR

Poder determinar l'error de cada distribucions de clusters final és important per tenir una idea numèrica que ens indiqui la semblança de les ditribucions de clusters agrupades vers a les seves distribucions finals. Volem d'alguna manera determinar un promig de samples mal assignades entre totes les distribucions de clusters que s'agrupen sota una sola distribució final.

Aquest error el calculem de la següent manera:

Sigui I el nombre total de distribucions de clusters d'entrada que formen la distribució de clusters final.

Sigui C el nombre de clusters que té la distribució.

Sigui M el nombre total de samples de la microarray.

Direm que e_{cm} és 1 si la sample m del cluster c de la distribució d'entrada està mal assignada i 0 si està ben assignada. Valdrà 0 si la sample està esta assiganda als dos clusters o si no està asignada a cap dels dos clusters equivalents, i 1 si està assiganada a un cluster pero a l'altre no. Ho podriem veure com si fos una operació XOR, on el resultat és 1 si els 2 valors de pertanyensa son diferents i 0 si els 2 valors son iguals.

Definim l'error de cada distribució de clusters d'entrada (E_{fi}) com la mitja de mostres mal assigandes per a cada cluster de mostres d'aquesta distribució.

$$E_{fi} = \frac{\sum_{c=1}^{C} \sum_{m=1}^{M} e_{cm}}{C}$$

Per tant, l'error total de la distribució de clusters final obtinguda és la mitja d'errors de tots els errors acumulats entre les distribucions d'entrada agrupades i la distribució final. Diguem I al número de distribucions de clusters agrupades a la mateixa distribució final tenim que l'error acumulat E es:

$$E = \frac{\sum_{f=1}^{I} E_{fi}}{I}$$

Exemple del càlcul de l'error

Podem veure graficament com es calcula aquest error. Agafem com a exemple una distribució de clusters final, agrupada amb els següents paràmetres:

% Matching: 75%
Cl. intersection: Assingning joint samples to the smaller cluster
α : 0.009

Fi: distribució de clusters d'una parella de gens.
SEj: SubEspai-j o cluster-j de la distribució de clusters-i al que pertanyen les samples.

Distribucions de clusters d'entrada agrupades en la mateixa distribució final:

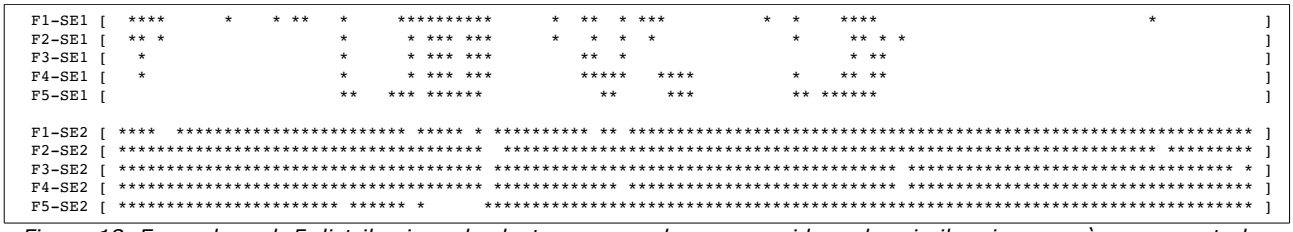


Figura 18 Exemple amb 5 distribucions de clusters agrupades per considerar-les similars i que seràn representades per una única distribució de clusters final. Es mostren 2 caixes: SE1 i SE2 una per cada cluster. La mateixa fila al cluster 1 y al cluster 2 indique que aquests dos clusters pertanyen a la mateixa distribució de entrada. Les columnes representen les condicions mostrals de la microarray, si apareix un '' vol dir que aquesta mostra pertany aquest cluster. Una mostra pot pertanyer als 2 clusters*

Distribució de clusters final d'aquesta agrupació (α = 0.009 -> assignació a partir d'1 sample repetida):

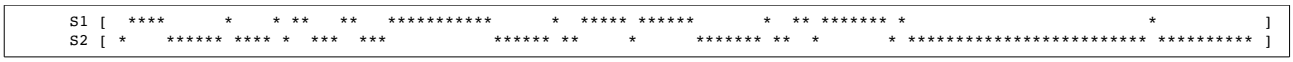


Figura 19 Distribució de clusters final que resumeix les distribucions agrupades a la figura 18. Selecció per T-Student i α = 0.009. El simbol '' indica eu la mostra pertany al cluster.*

Aquesta distribució de clusters final té un error de 31.9 calculat de la següent manera:

Comptem per cada distribució inicial de clusters, quantes samples (de tots 2 els clusters) estan mal assignades a la distribució final. L'error acumulat de cada distribució, és la suma de samples mal assigandes de tots els clusters respecte la distribució final.

Número fitxer	Relació de gens	Mostres Cluster-1	Mostres Cluster-2	Mostres mal assignades
F1	g1167g1177h0.75d0.3.ldom	33	111	52
F2	g1170g1180h0.75d0.3.ldom	20	115	71
F3	g1174g1178h0.75d0.3.ldom	15	115	74
F4	g1175g1178h0.75d0.3.ldom	23	115	64
F5	g1384g1416h0.75d0.3.ldom	24	110	58
	TOTAL:			**319**

Figura 20 Relació de distribucions de clusters que s'agrupen sota la distribució final de la taula 19
La columna "relació de gens", ens mostra parelles de gens amb una relació d'expressió que permet generar distribucions de clusters d'entrada (concretament apareix el nom del fitxer ldom que conté la distribucion de clusters d'entrada).
La columna mostres cluster -1 i cluster-2, indica el nombre de mostres que té cada cluster. La columna mostres mal asignades és la suma de mostres dels 2 clusters asignades a l'altre cluster a la distribució final.

De la taula 20, en podem treure error mig de mostres mal assignades de totes les distribucions inicials entre el número de clusters i les distribucions que la formen és:

Error = 319 / (5*2) = 31.9

3.2.1 Influència d' α de T-Student i sel·lecció per Màxims en l'error acumulat:

Podem veure com influeix el valor de α en l'error acumulat, agafant la mateixa distribució de clusters final però amb diferents valors de α .

El valor de tall per a les zones d'intersecció és més alt quan més alt és α , fent que menys samples siguin assignats a aquest cluster d'aquest samples amb un cluster no ben definit. Com que normalment, a les zones d'intersecció, hi ha una major influència del cluster predominant, la selecció per T-Student intenta corregir que el cluster petit es quedi sense representació. Com que el càlcul de l'error té en compte les diferències de cada distribució de clusters d'entrada respecte la distribució final del grup, l'error acumulat creix quan s'asignen samples a un cluster de la distribució final on no apareixen clasificats mayoritariament a aquest cluster a las distribucions d'entrada.

Distribucions de clusters agrupades sota la mateixa distribució final:

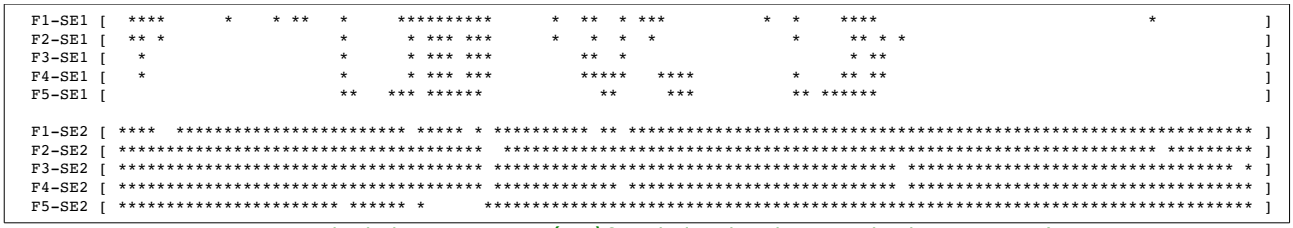


Figura 21 Exemple de la representació gràfica de les distribucions de clusters que s'agrupen per formar una distribució de clusters final. Dades del mateix exemple que la figura 18
*Els * indiquen que la mostra pertany al cluster. Cada columna representa una condició mostral de la microarray.*

Distribució de clusters final d'aquesta agrupació (α = 0.40 -> assignació a partir del reforç de 3 samples**):**

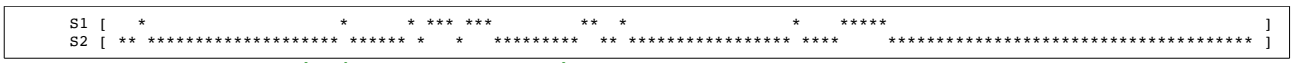


Figura 22 Representació gràfica de la distribució de clusters final per les distribucions agrupades a la figura 21. Assignació de samples per T-Student amb $\alpha = 0{,}40$ *. Els * indiquen que la mostra pertany al cluster.*

Número fitxer	Relació de gens	Mostres cluster-1	Mostres cluster-2	Mostres mal assignades
F1	g1167g1177h0.75d0.3.ldom	33	111	32
F2	g1170g1180h0.75d0.3.ldom	20	115	27
F3	g1174g1178h0.75d0.3.ldom	15	115	22
F4	g1175g1178h0.75d0.3.ldom	23	115	24
F5	g1384g1416h0.75d0.3.ldom	24	110	30
	TOTAL:			**135**

Figura 23 Relació de distribucions de clusters que s'agrupen sota la distribució final de la figura 22.
La columna "relació de gens", ens mostra parelles de gens amb una relació d'expressió que permet generar distribucions de clusters d'entrada (concretament apareix el nom del fitxer ldom que conté la distribucion de clusters d'entrada).
La columna mostres cluster -1 i cluster-2, indica el nombre de mostres que té cada cluster. La columna mostres mal asignades és la suma de mostres dels 2 clusters asignades a l'altre cluster a la distribució final.

Error = 135 / (5*2) = **13.5**

Distribució de clusters final d'aquesta agrupació per Màxims:

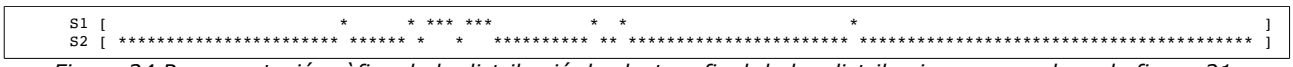


Figura 24 Representació gràfica de la distribució de clusters final de les distribucions agrupades a la figura 21. Assignació per màxims.
*Els * indiquen que la mostra pertany al cluster.*

Número fitxer	Relació de gens	Mostres Cluster-1	Mostres Cluster-2	Mostres mal assignades
F1	g1167g1177h0.75d0.3.ldom	33	111	30
F2	g1170g1180h0.75d0.3.ldom	20	115	21
F3	g1174g1178h0.75d0.3.ldom	15	115	16
F4	g1175g1178h0.75d0.3.ldom	23	115	22
F5	g1384g1416h0.75d0.3.ldom	24	110	24
	TOTAL:			**113**

Figura 25 Relació de distribucions de clusters que s'agrupen sota la distribució final de la figura 24.
La columna "relació de gens", ens mostra parelles de gens amb una relació d'expressió que permet generar distribucions de clusters d'entrada (concretament apareix el nom del fitxer ldom que conté la distribucion de clusters d'entrada).
La columna mostres cluster -1 i cluster-2, indica el nombre de mostres que té cada cluster. La columna mostres mal asignades és la suma de mostres dels 2 clusters asignades a l'altre cluster a la distribució final.

Error = 113 / (5*2) = **11.3**

És evident que la selecció per màxims sempre donarà un error més baix, degut a que clasifica la mostra al cluster que te un reforç més gran.

3.2.2 Selecció de valors d' α representatius

A continuació es fa un estudi de quins son els valors d' α que donen uns resultats variats i representatius, tenint en compte tant l'error acumulat com la rellevància de les distribucions finals resultants.

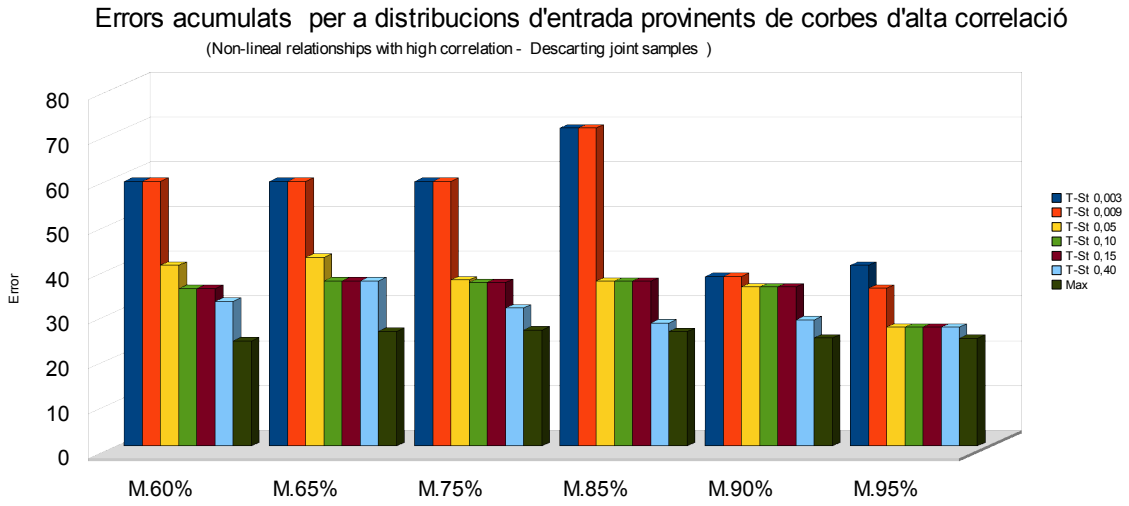


Figura 26 Taula d'errors acumulats per a distribucions d'entrada provinents de corbes d'alta correlació. Assignació de samples per màxims i per diferents valors d' α . Agrupació de distribucions utilitzan diferents percentatges de matching i descartant les zones d'intersecció a tots ells.

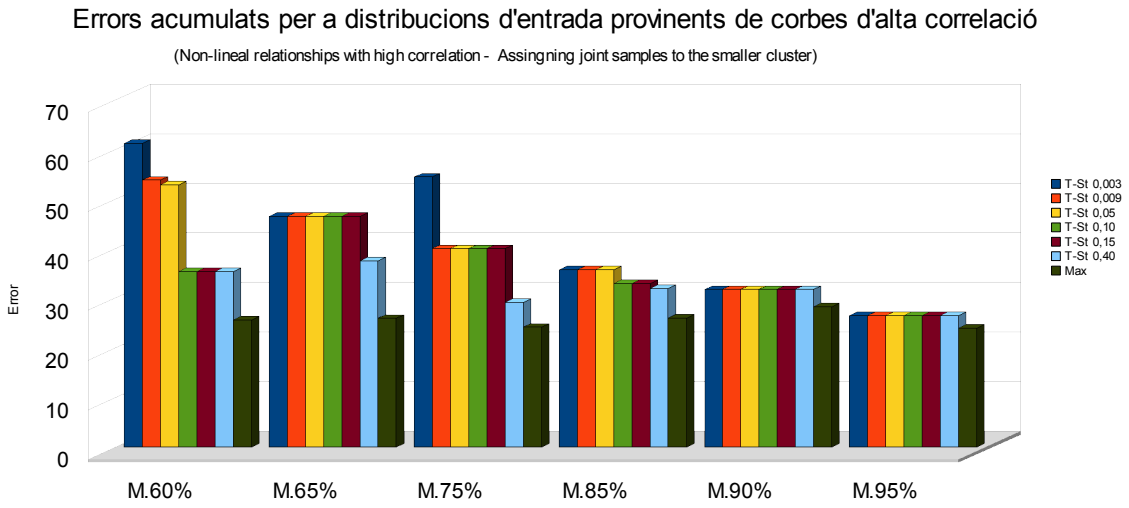


Figura 27 Taula d'errors acumulats per a distribucions d'entrada provinents de de corbes d'alta correlació. Assignació de samples per màxims i per diferents valors d' α . Agrupació de distribucions per als diferents percentatges de matching i assignant les zones d'intersecció al cluster més petit.

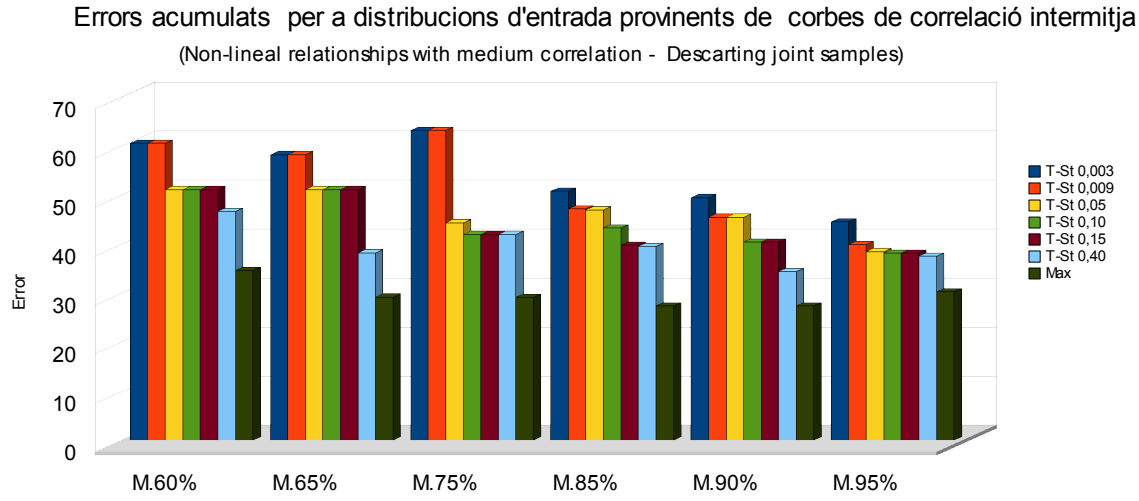


Figura 28 Taula d'errors acumulats per a distribucions d'entrada provinents de corbes de correlació intermitja. Assignació de samples per màxims i per diferents valors d' α . Agrupació de distribucions per als diferents percentatges de matching i descartant les zones d'intersecció.

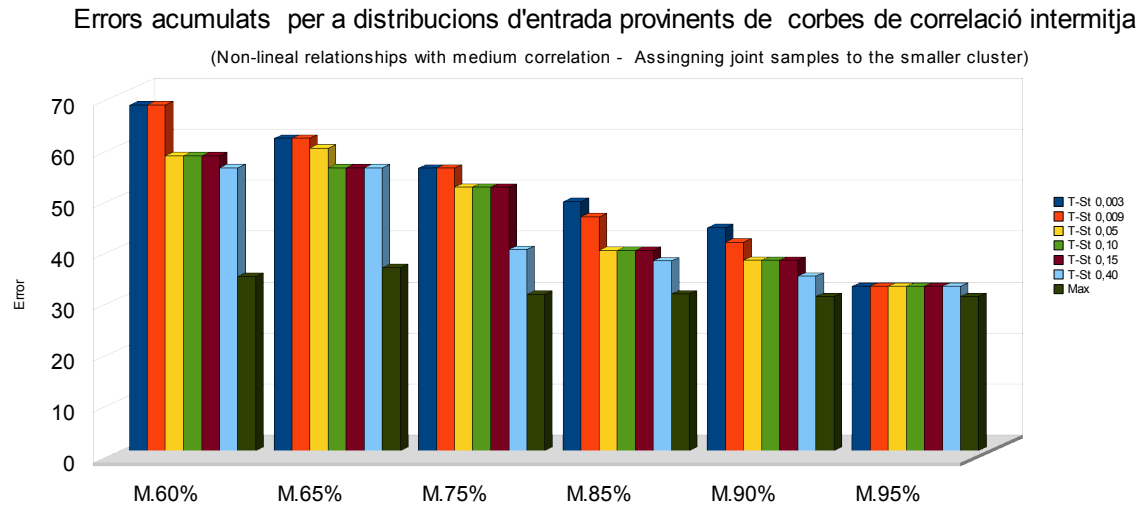


Figura 29 Taula d'errors acumulats per a distribucions d'entrada provinents de corbes de correlació intermitja. Assignació de samples per màxims i per diferents valors d' α . Agrupació de distribucions per als diferents percentatges de matching i assignant les zones d'intersecció al cluster més petit.

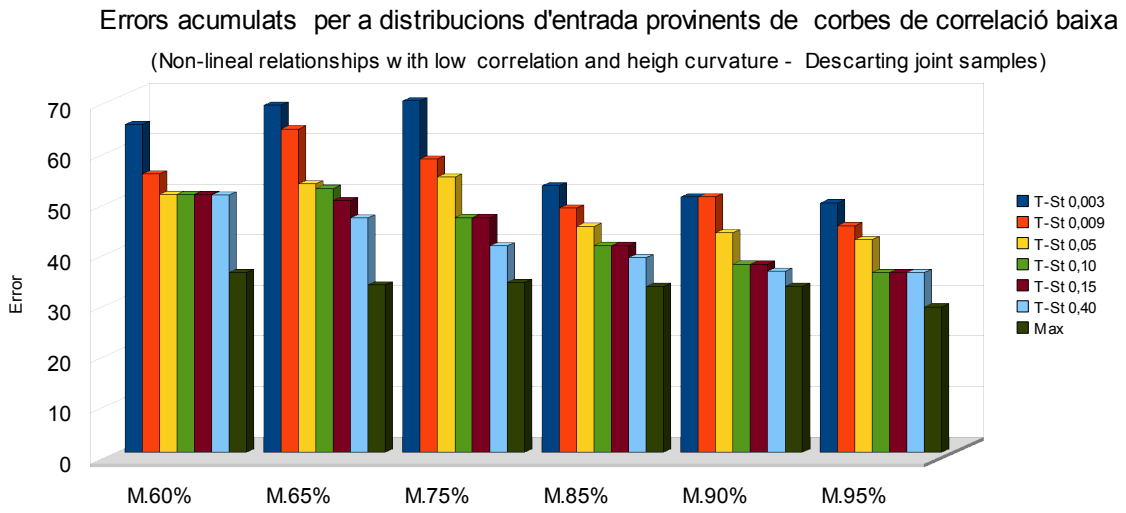


Figura 30 Taula d'errors acumulats per a distribucions d'entrada provinents de corbes de correlació baixa. Assignació de samples per màxims i per diferents valors d' α . Agrupació de distribucions per als diferents percentatges de matching i descartant les zones d'intersecció.

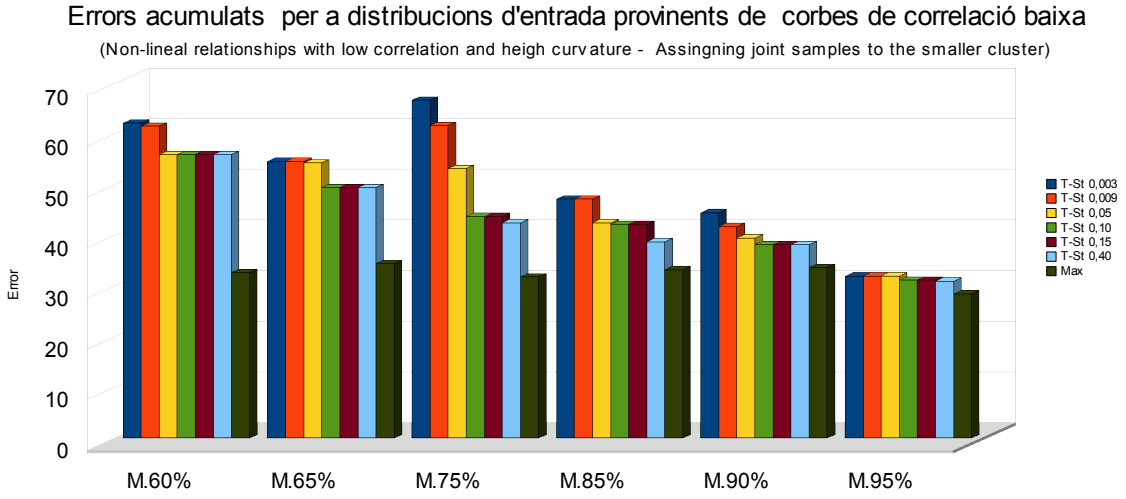


Figura 31 Taula d'errors acumulats per a distribucions d'entrada provinents de correlació baixa. Assignació de samples per màxims i per diferents valors d' α . Agrupació de distribucions per als diferents percentatges de matching i assignant les zones d'intersecció al cluster més petit.

A les figures 26, 27, 28, 29, 30 i 31, es poden veure els errors acumulats en cada agrupació per filtre de distribucions d'entrada i tipus de mètode de tractament de les zones d'intersecció. Es pot observar que en la majoria de gràfiques els valors d' α més influents en l'error respecte a la resta son el 0,003 i 0,40, ja que tenen valors d'error molt diferenciats. La resta de valors (α = 0.009, 0.05, 0.10, 0.15), obtenen resultats semblants o poc variants respecte a aquests dos.

Es pot observar també que l'error acumulat per al valor d' α = 0.40 s'aproxima al de la

clasificació per màxims, però evitant que en zones d'intersecció el cluster dominant sigui molt influent en l'assignació de samples.

3.3 INTERFACE ONLINE VIA WEB

Per poder analitzar els resultats obtinguts, s'ha creat un conjunt de pàgines web interactives, que s'han integrat al servidor d'aplicacions web al portal del IBB-UAB per a l'anàlisi de microarrays. L'usuari podrà veure a partir d'ara les distribucions finals de clusters que representen els diferents canvis fenotípics i els gens involucrats directament en aquests canvi per cada microarray que ell o el seu grup de treball vulgui analitzar.

La pàgina web s'ha desenvolupat en llenguatge PHP per la configuració actual del servidor web. La seva fàcil portabilitat a altres plataformes, el fa ser un llenguatge a tenir en compte per a qualsevol desenvolupament web. Les aplicacions ja instalades en el servidor web estan escrites en PHP. Tot i ser un llenguatge interpretat, la seva velocitat en el tractament de dades és prou alta com per tenir-lo en compte a l'hora de treballar-hi. El PHP es caracteritza per ser un paquet fàcil d'instal·lar en plataformes web com Apache. Té un repertori d'instruccions prou ampli com per cobrir la majoria de necessitats de qualsevol programa. Les pàgines web fetes en PHP, permeten fer canvis molt ràpidament als programes sense haver d'implicar a altres mòduls ja instal·lats, com podrien ser les aplicacions escrites en JAVA en plataformes també web.
El ràpid tractament que té el PHP sobre objectes de tipus llista, és una de les opcions a tenir en compte a l'hora de treballar dades com les que es manipulen en aquest projecte.

3.3.1 Visualització de distribucions finals de clusters

En aquest apartat s'explica com l'usuari pot interactuar amb la web per analitzar els resultats de la agrupació de distribucions de clusters.

Per poder treballar les distribucions finals de clusters, s'ha incorporat l'accés a la nova eina d'nalisis desde la página de gestio de les microarrays pujades al servidor. Desde allà s'accedeix directament a PCOPSample-cl per la microarray donada.

Per defecte, sempre s'entra a aquesta pàgina amb els parámetres següents:

PCOP Input data: Non-lineal relationships with high correlation

Cl. Intersection: Assingning joint samples to the smaller cluster

% matching: 75%

Sample assign to cl. by: t-student amb α = 0,003

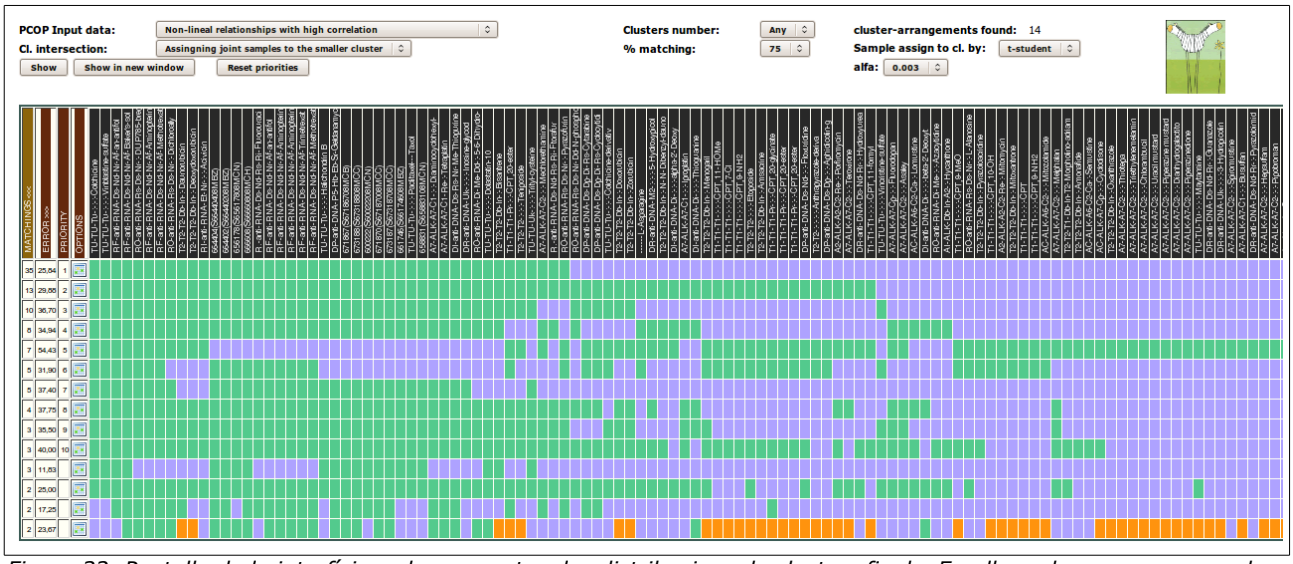


Figura 32. Pantalla de la interfície web per mostrar les distribucions de clusters finals. En ella podem veure que podem seleccionar entre els diferents criteris d'agrupació de distribucions de clusters per mostrar els resultats obtinguts en el procés d'agrupació. Les mostres de les distribucions de son les columnes, amb els seus nom abreviats dibuixats en format vertical. Les distribucions de clusters finals son les file. Les mostres de cada distribució es pinten en un color difernciat, indicant el cluster al que està assignat.Per cada distribució de clusters finals es mostra el número de agrupacions fetes i el seu error. Les columnes (mostres), s'ordenen cada vegada en funció de de l'ordre en que prioritzem la visualització de les distribucions de clusters. La més prioritària s'ordena primer. La segona s'ordena respecta la primera, dibuixant els subclusters que en depenen, i així successivament fina a la darrera distribució. Tots els elements de la pantalla tenen l'ajuda alternativa.

Aquesta pantalla mostra les distribucions de clusters finals obtingudes per els parámetres seleccionats.

La distribució d'elements visuals a la pantalla s'ha dissenyat de forma lo més còmode possible per a l'usuari, els paràmetres a la part superior i la representació de les dades a sota.

A la part superior dreta, hi ha la icona d'accés a l'ajuda, que explica la funcionalitat i opertivilitat de l'eina i el significat dels diferents parámetres.

Els paràmetres d'entrada son:

PCOP Input data: filtre de les distribucions de clusters d'entrada en base al grau de correlació de la relació d'expressió no lineal associada i el grau de corvatura d'aquesta relació (veure secció 2.3 - Filtratge de les distribucions de clusters d'entrada en base a la relació d'expressió gènica associada).

Clusters number: Selecció de les distribucions finals que es volen veure en funció del número de clusters de cada una

Cl. Intersection: Tractament que s'ha donat a les interseccions de clusters d'una mateixa distribució abans de compararles amb les distribució de clusters temporals. (veure secció 2.4 - tractament de les interseccions de clusters d'una mateixa distribució).

% matching: Percentatge mínim exigit de matching per considerar equivalents les distribucions de clusters d'entrada. (Veure secció 2.5 - càlcul del matching entre les distribucions de clusters)

Sample assign to cl. by: Criteri d'assignació de samples a les distribucions de clusters finals. Pot ser ver "T-Student" escollint un valor d' α o per "màxims". (Veure secció 2.6 - Assignació dels samples a un o altre cluster de la distribució final)

Cada fila representa una distribució de clusters final trobada en funció dels filtres possibles del procés PCOPSample-cl.

Les columnes representen les mostres de la microarray, dibuixades en colors, en funció del cluster al que s'ha assignat de cada distribució de clusters final.

El nom de la sample, sempre es mostra amb un text alternatiu quan es situa el mousse sobre de cada casella, sigui en el títol o en qualsevol sample d'una distribució de clusters final. Això s'ha fet degut al poc espai disponible per visualitzar tanta informació. Com que la informació dibuixada és tant petita que per alguns és il·legible, s'ha optat per ajudar al màxim en la compressió dels textos que hi apareixen.

3.3.2 Ordenació de les distribucions de clusters i ordenació de les mostres.

Les condicions mostrals (les columnes) s'ordenen de forma qeu els samples que pertanyin al mateix cluster apareixin junts. Això no es pot fer per totes les distribucions de clusters a l'hora , per aquest motiu s'estableixen prioritats. L'usuari pot canviar les prioritats, seleccionant les distribucions de clusters que vol que siguin les primeres en ordenar-se, de forma que, la resta de distribucions s'ordenen respectant els clusters d'aquesta distribució(com es mostra a la fig 33 per 3 distribucions ordenades per prioritat). D'aquesta forma es poden comparar les distribucions de clusters, trobar intersecciónes entre una distribució i una altre i establir una jerarquía de distribucions.

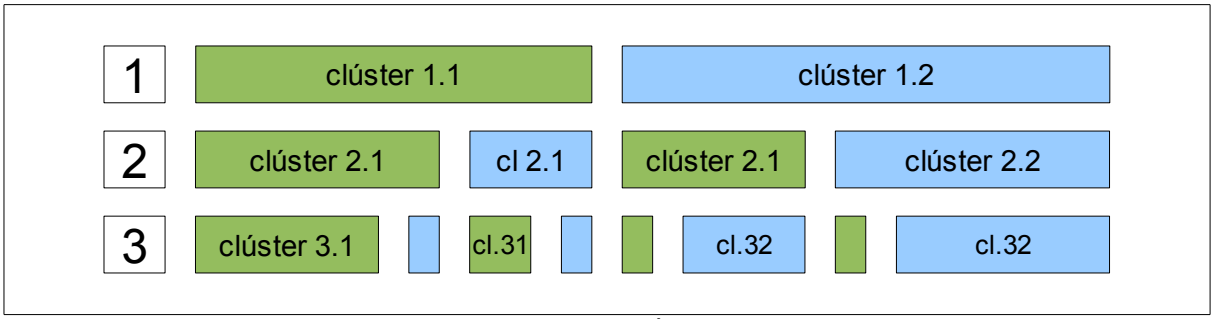


Figura 33. Exemple de com es visualitzen els clusters en funció de la prioritat de les distribucions de clusters finals. Cada fila representa a una distribució de clusters, Cada color representa un cluster diferent de la distribució. A la distribució de prioritat 1, les mostres de cada cluster es mostren totalment separades, a la distribució de prioritat n, els clusters es mostren separats pero com a subclusters de la distribució de prioritat n-1 .Això ens indica que un fenotip pot formar part de dos fenotips descrits per una altre distribució.

Per a cada distribució de clusters final és mostren els sgüents valors de possible interès:

Número de "matchings" o distribucions de clusters que s'han agrupat per considerarse similars segons els parámetres d'agrupació sol·licitats, i que corresponen a la mateixa distribució de clusters final, o sigui que participen del mateix canvi fenotípic que descriu aquesta distribució final. L'interès científic d'aquest número de matchings, és que quan més alt sigui, més gens estaran implicats en el canvi fenotípic que representa la distribució de clusters.

Error de la distribució de clusters final: Per cada distribució de clusters final, es mostra l'error d'aquesta en el procés d'agrupació (secció 3.3).

La ordenació de la llista de distribucions finals de clusters (les files), es pot fer per 2 critreris: Una llista decreixent en funció del número de distribucions de clusters agrupats en una mateixa distribució final, o una llista creixent en funció de l'error acumulat a les distribucions finals visualitzades (la distribució amb menys error acumulat amunt de tot). En tot moment, es pot saber quin és el criteri d'ordenació de files i columnes que es

mostra.

Es pot veure a les següents imatges com es reordenen les columnes (samples) en funció de la prioritat marcada a cada distribució.

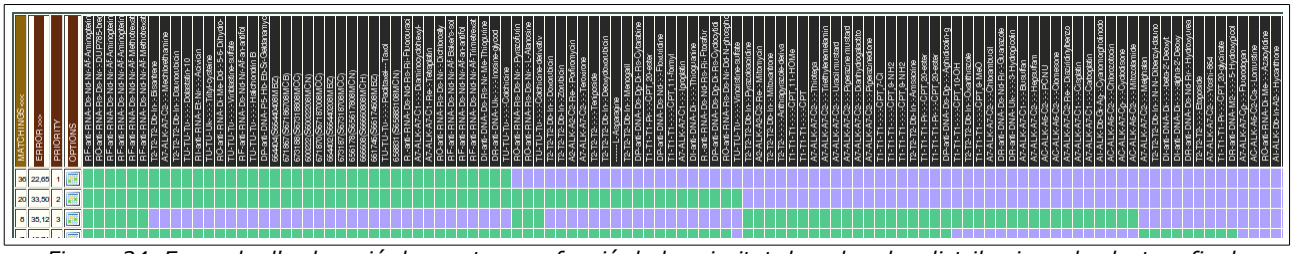


Figura 34. Exemple d'ordenació de mostres en funció de la prioritat donada a les distribucions de clusters finals

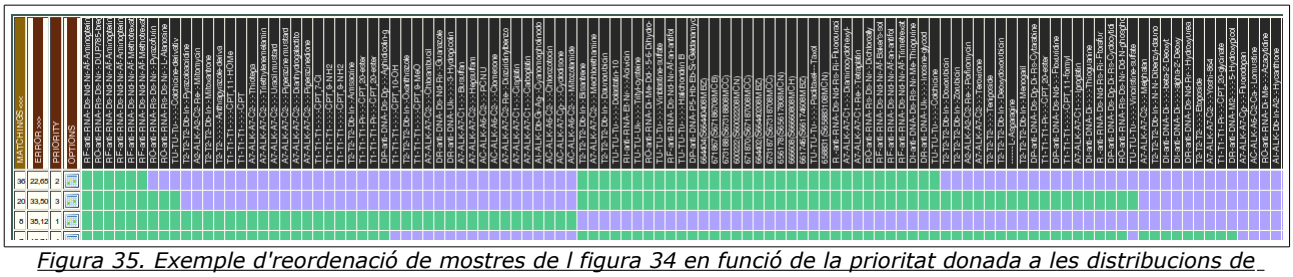


Figura 35. Exemple d'reordenació de mostres de l figura 34 en funció de la prioritat donada a les distribucions de clusters finals

3.3.3 PARELLS DE GENS ASSOCIATS A CADA DISTRIBUCIÓ DE CLUSTERS FINAL

Per poder veure les relacions d'expresió que s'an utilitzat per obtindre les distribucions finals, s'ha desenvolupat una pantalla que mostra aquesta informació. A partir d'una distribució de clusters final de la llista, s'accedeix al menú que permet accedir a la pantalla que mostra les diferents parelles de gens i la seva relació amb els clusters descrits per aquesta distribució de clusters final.

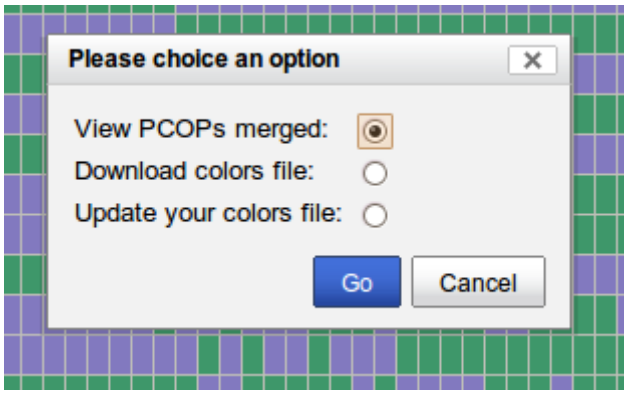


Figura 36. Pantalla d'opcions d'una distribució de clusters final.

La figura 36 mostra la pantalla d'opcions, també permet descarregar el ficher de clusters d'aquesta distribució de clusters final, o canviar en el servidor, la distribució de clusters que l'usuari té assignada per utilitzar en altres eines del servidor com buscar gens marcador d'un cluster, etc.

Relació de parelles de gens involucrades en el mateix canvi d'estat fenotípic.

PCOPs merged analysis

Results

	Id	Gene-1	f: uncorrelation	Id	Gene-2	Graphical interface
	696	This term is	0.040146	766	This term is	
	764	LOX	0.041097	767	PARP3	
	732	Error en la	0.043317	735	COL4A2	
	645	AKT3	0.046119	735	COL4A2	
	698	Hypothetical	0.046597	742	NNMT	
	700	DKK3	0.047458	721	This term is	
	698	Hypothetical	0.048484	715	Transcribed	
	786	SLC36A4	0.050618	928	This term is	
	734	Error en la	0.051203	735	COL4A2	

Number of merges: 9

Figura 37. Relació de parelles de gens que pertanyen a la mateixa distribució de clusters finals.
(Característiques de la distribució de clusters final: Paràmetres de selecció per trobar-la: 85% matching, merges 9, errror 23.39, T-St 0,003)
Totes elles estan involucrades en el mateix canvi fenotípic

La figura 37 mostra les diferents relacions no lineals de les parelles de gens. És a dir, relaciona tots els parells de gens, que participen del mateix canvi fenotípic. El llistat està ordenat pel factor de correlació de les corbes PCOP que descriuen la relació d'expressio de cada parella de gens.

Des d'aquesta interficie, es pot accedir a la descripció de cada gen, veure per a cada gen, quins son els altres gens de la microarray amb els que està relacionat. També es pot accedir a la informació de la BBDD del NCBI (National Center for Biotechnology Information), que indica si aquest gen es un gen marcador per a les microarrays daqueesta BBDD.

La finestra ve identificada per el %matching utilitzat, el nº de distribucions agrupades, el error acumulat, i la clasificació de les mostres utilitzades dins de cada cluster. D'aquesta forma els parells de gens implicats en una distribució de clusters final o en una altra son fàcilment comparables.

Finalment, es pot veure un icone que descriu la topologia de la corva, i es pot obrir la pantalla de visió en detall de la PCOP de la parella de gens. En aquesta interfice apareixen colorejades les mostres finalment clasificades en un cluster o un altre a la distribució final de clusters. D'aquesta manera es pot veure i comparar com les corbes PCOP de cada parella de gens d'una mateixa distribució de clusters final, descriuen el mateix canvi fenotipic.

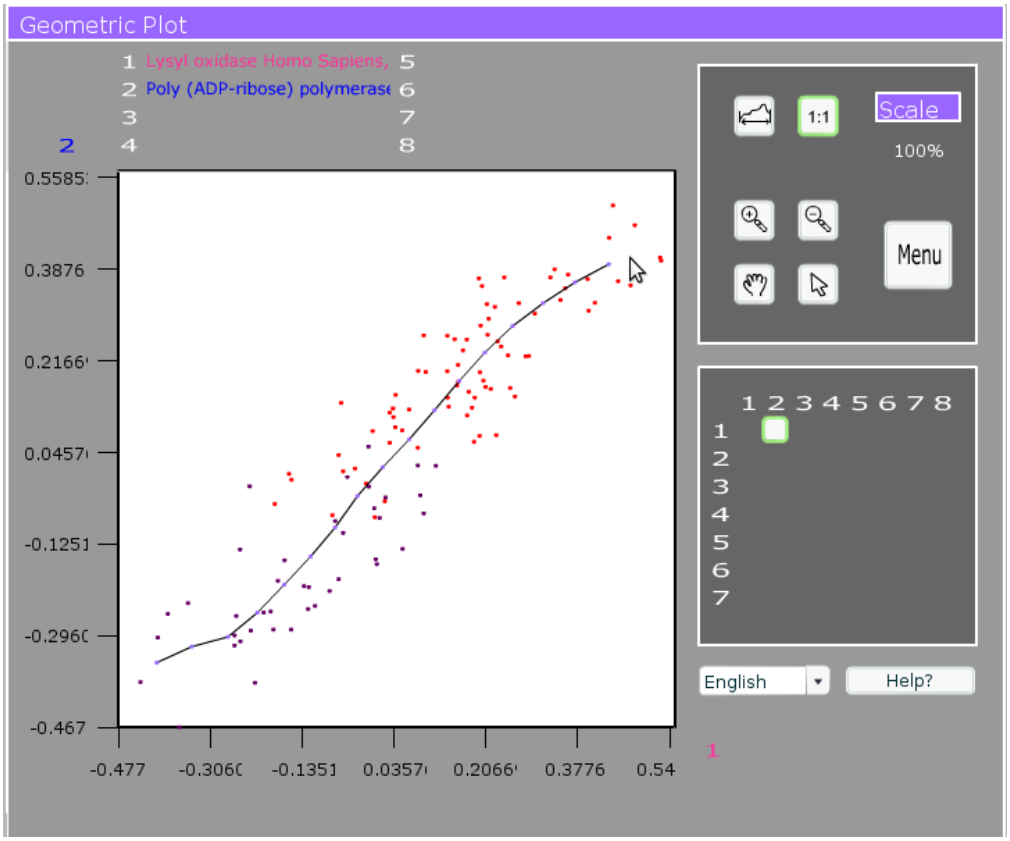


Figura 38. Representació d'un canvi fenotípic marcat pels clusters de la distribució final on estan representats totes les parelles de gens de la figura 37
Lysyl oxidase Homo Sapiens, 404 sequence(s) -
Poly (ADP-ribose) polymerase family, member 3 Homo Sapiens, 118 sequence(s)
A la imatge els punts son la mostra que compara els 2 valors d'expressió (component x i y) dels gens relacionat. La corba que es dibuixa, és la corba PCOP que descriu la relació d'expressió dels 2 gens.
A la figura es veu la relació del nivell d'expressió de cada gen, en funció de la corba PCOP.

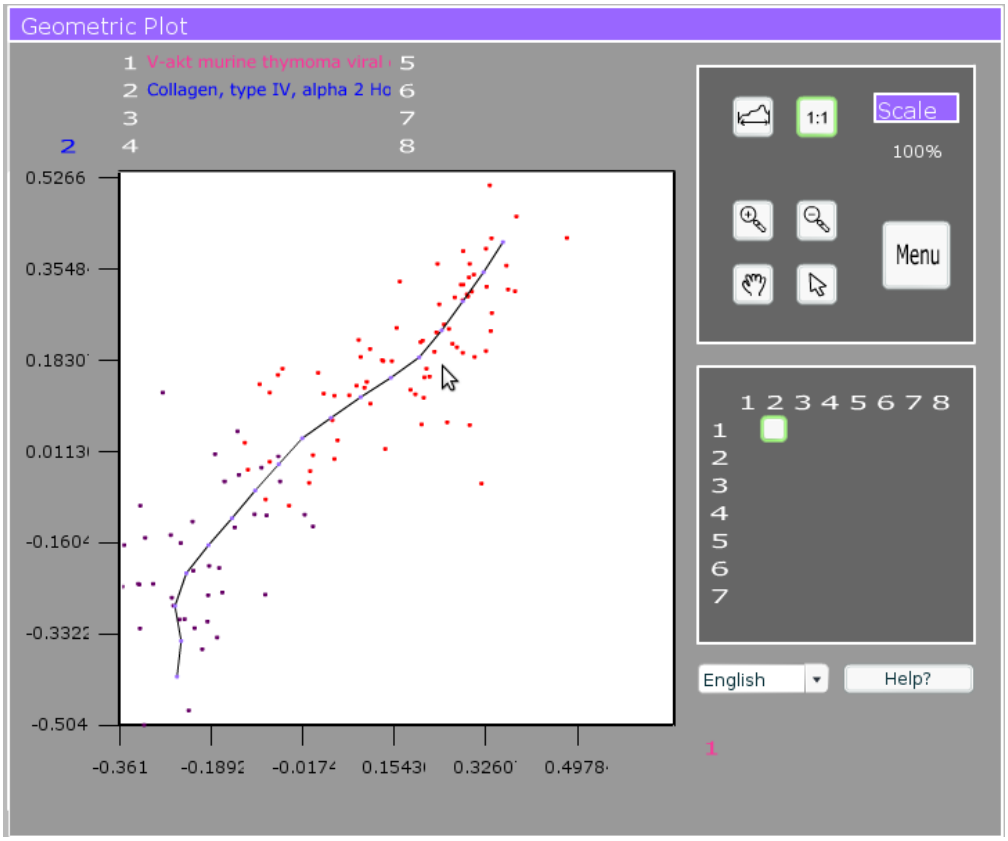


Figura 39. Representació d'un canvi fenotípic marcat pels clusters de la distribució final en que hi estan representades totes les parelles de gens de la figura 37

Lysyl oxidase Homo Sapiens, 404 sequence(s) -

Poly (ADP-ribose) polymerase family, member 3 Homo Sapiens, 118 sequence(s)

A la imatge els punts son la mostra que compara els 2 valors d'expressió (component x i y) dels gens relacionat. La corba que es dibuixa, és la corba PCOP que descriu la relació d'expressió dels 2 gens.

A la figura es veu la relació del nivell d'expressió de cada gen, en funció de la corba PCOP.

Les imatges de les figures 38 i 39 mostren les corbes de 2 relacions no linias de gens que participen del mateix canvi fenotípic. Totes les parelles de gens de la figura 2.23 també estan involucrades en aquest canvi fenotípic.

Els grups de punts d'un mateix color (un mateix cluster a la distribució final), marquen les mostres d'un mateix fenotip. Com que les mostres de totes les relacions estan asignades als mateixos clusters, podem dir que aquests gens estan involucrats en el mateix canvi fenotípic.

3.4 Ejemplos de anàlisis.

A continuació es mostren exemples d'us de l'eina desenvolupada, on es pot observar la significancia dels resultats oferts.

3.4.1 EXEMPLE 1. CANVI FENOTÌPIC AMB LA IMPLICACIÓ DE VARIS GENS ALTAMENT CORRELACIONS.

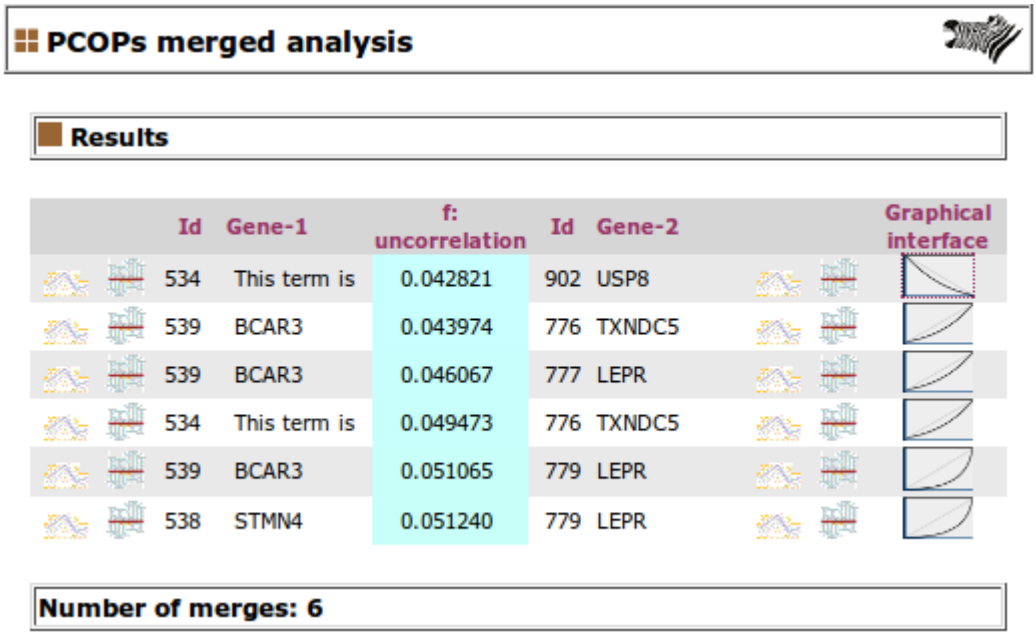


Figura 40 Exemple de parelles de gens altament correlacionats, agrupades sota una mateixa distribució de clusters final. Característiques de la distribució de clusters final: Agrupació de distribucions de clusters per assignació de les mostres en zones d'intersecció al cluster més petit, amb %matching 85%, assignació per T-Student amb $\alpha = 0{,}003$ *Error acumulat =23,39*

Podem veure a la Figura 40, un exemple de parelles de gens que formen part del mateix canvi fenotípic, obtingudes gràcies al procés d'agrupació de distribucions. Podem veure el tipus de corba o relació de cada parella de gens a l'icona que hi ha a la dreta de cada un d'ells. Es pot veure com totes les relacions tendeixen a la coexpressió menys una que és inhibidora. No obstant, les 2 últimes parelles de gens mantenen una relació d'expressió amb una corba mes pronunciada. La primera relació de la taula és de tipus X=-Y. Això ens indica que mentre per a la resta de gens el canvi fenotìpic els fa expressar-se, per aquest gen el que fa el mateix canvi fenotípic és infraexpresar-lo. Com que tenim el punt exacte en que es produeix el canvi de tendencia en la relació, podem saber exactament a quin nivell d'expressió comença el canvi fenotipic (una cosa imposible d'esbrinar amb els mètodes clàsics de clustering, ja siguin locals o globals)

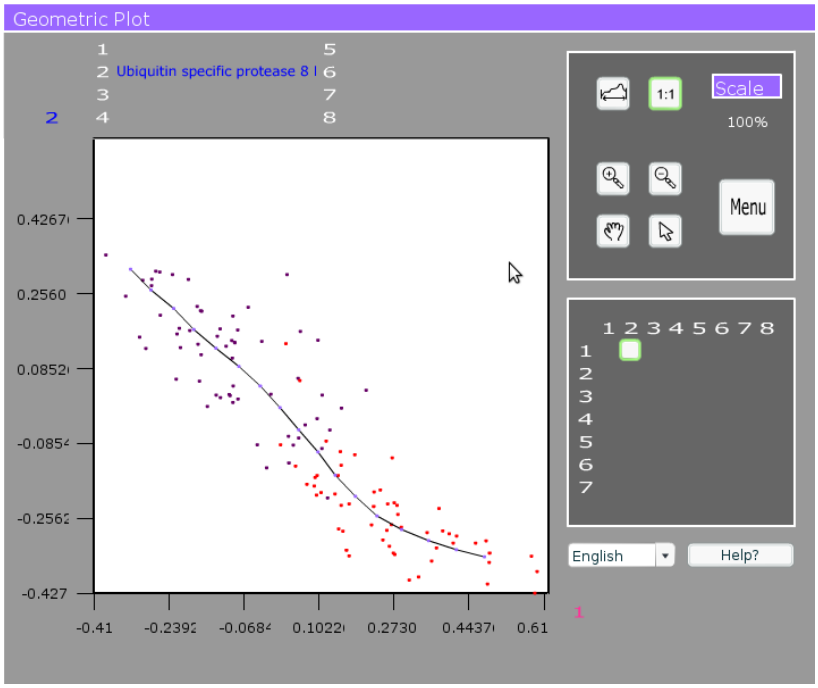


Figura 41 relació d'expressió de gens inhibidors.

gen 534 de la microarray (sense nom o desconegut) - Ubiquitin specific protease 8 Homo Sapiens, 373 sequence(s)
Primera parella de gens de la distribució final de clusters descrita a la figura 40
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens. Els mostres apareixen pintades d'un color o un altre depenent al cluster al que pertanyin. La distribució de clusters es la mateixa que la de les figures 42 i 43.

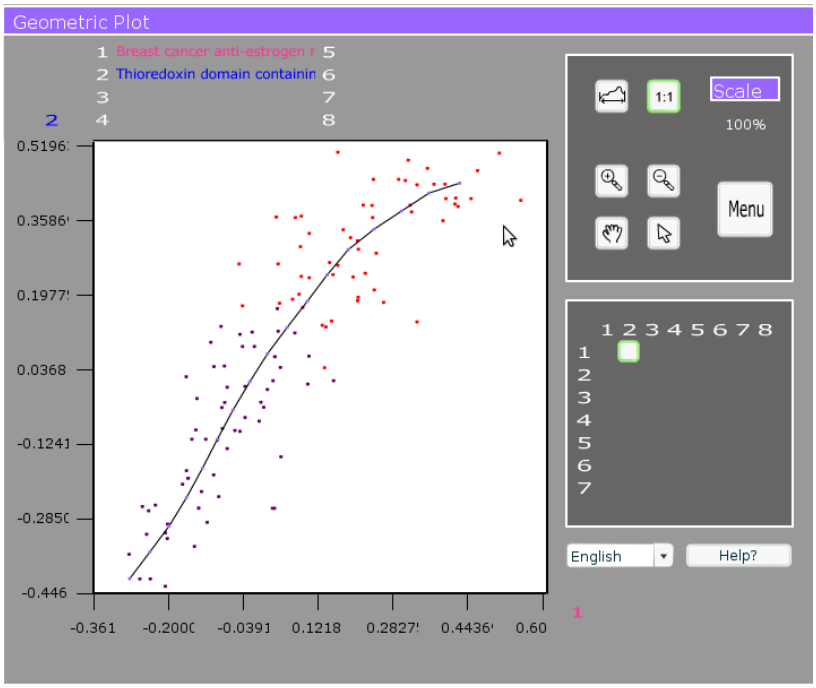


Figura 42 relació d'expressió de 2 gens gairabé coexpresats.
Breast cancer anti-estrogen resistance 3 Homo Sapiens, 272 sequence(s) - Thioredoxin domain containing 5 Homo Sapiens, 1410 sequence(s)
Segona parella de gens de la distribució de clusters final descrita a la figura 40
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens. Els mostres apareixen pintades d'un color o un altre depenent al cluster al que pertanyin. La distribució de clusters es la mateixa

que la de les figures 41 i 43

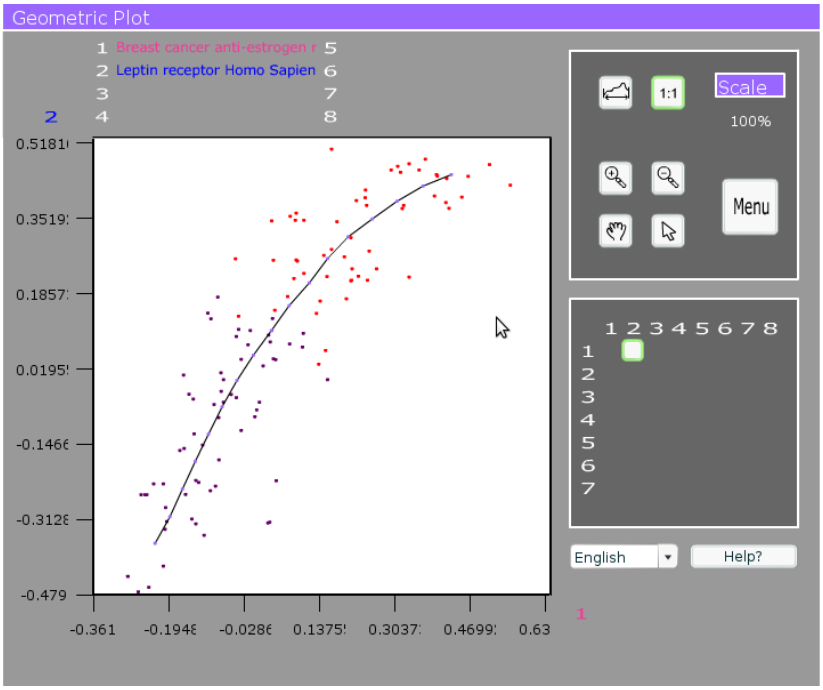


Figura 43
Breast cancer anti-estrogen resistance 3 Homo Sapiens, 272 sequence(s) - Leptin receptor Homo Sapiens, 657 sequence(s)
Tercera parella de gens de la distribució final de clusters descrita a la figura 40
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens. Els mostres apareixen pintades d'un color o un altre depenent al cluster al que pertanyin. La distribució de clusters es la mateixa que la de les figures 41 i 42.

De les figures 41, 42 i 43, podem observar com la mateixa distribució de clusters final separa per igual les corbes de les diferents relacions de gens. Això indica que el mateix canvi fenotípc afecta al nivell d'expressió de tots ells. Quants més gens afecti un canvi fenotípic, més gran serà el seur efecte.

3.4.2 EXEMPLE 2. DISTRIBUCIONS FINALS DE 3 CLUSTERS

PCOPs merged analysis

Results

Id	Gene-1	f: uncorrelation	Id	Gene-2	Graphical interface
784	CGI-49	0.063468	966	MARCKSL1	
784	CGI-49	0.066690	965	PDXK	
770	TRAM1	0.071968	964	UQCRH	
784	CGI-49	0.073339	961	FLJ20232	
522	EPHA2	0.078381	745	ITGA3	
732	Error en la	0.079141	951	CKMT1	
745	ITGA3	0.079353	749	NID2	
541	PDLIM7	0.080523	738	NYD-SP21	
784	CGI-49	0.081759	963	LKAP	
768	QSCN6	0.081811	788	GNAS	
716	C6orf108	0.081973	762	FNDC3B	
784	CGI-49	0.082277	962	EIF3S6IP	

Number of merges: 12

Figura 44 Exemple de parelles de gens de correlació mitja, agrupades sota una mateixa distribució de clusters final. Característiques de la distribució de clusters final: Agrupació de distribucions de clusters descartant les mostres en zones d'intersecció, amb %matching 85%, assignació per T-Student amb $\alpha = 0{,}40$ *Error acumulat=32,44*

A la figura 44 apareix el llistat de parelles de gens que s'agrupen sota la mateixa distribució de clusters final. La particularitat d'aquestes corbes és que a més de estar involucrades en el mateix canvi fenotípc, aquest implica 3 fenotips diferents, es a dir, la distribució de clusters te 3 clusters. Com es pot veure, les corbes que descriuen son diferents les unes amb les altres, però no obstant totes obeixen segueixen el mateix canvi d'estats pel fet d'haver estat agrupades en la mateixa distribució de clusters final.

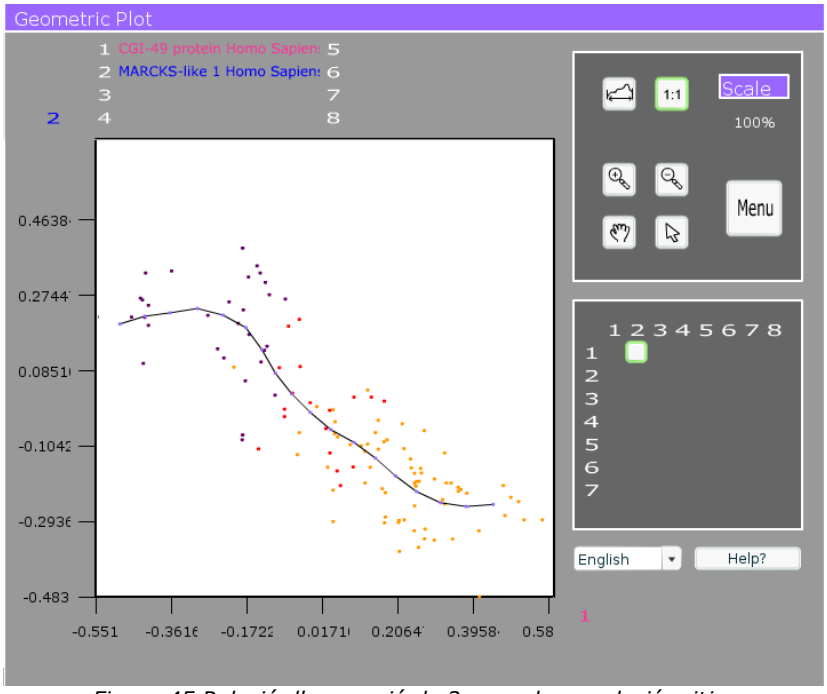


Figura 45 Relació d'expressió de 2 gens de correlació mitja.
CGI-49 protein Homo Sapiens, 456 sequence(s) - MARCKS-like 1 Homo Sapiens, 1021 sequence(s)
Primera parella de gens de la distribució de final de clusters descrita a la figura 44
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens. Els mostres apareixen pintades d'un color o un altre depenent al cluster al que pertanyin. La distribució de clusters es la mateixa que la de les figures 46 i 47.

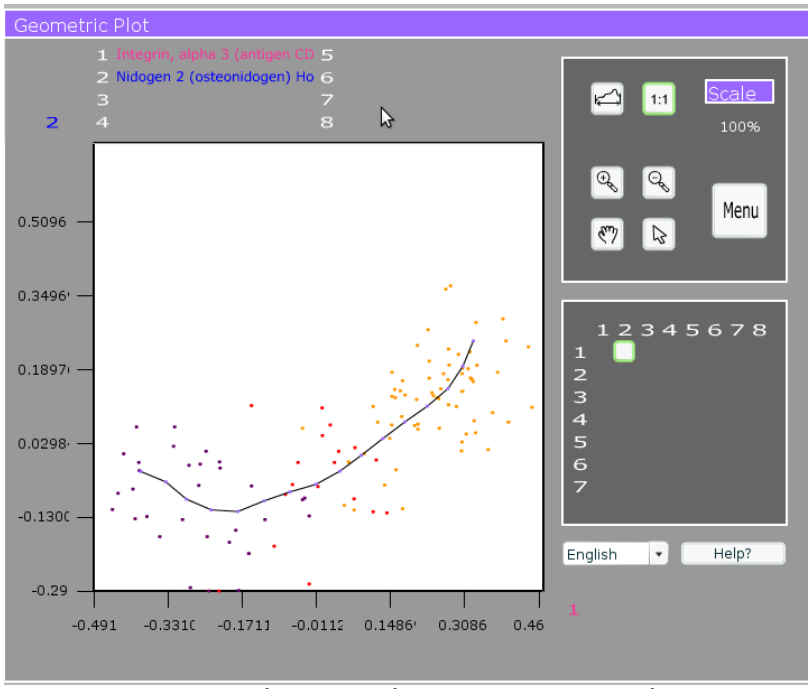


Figura 46 Relació d'expressió de 2 gens de correlació mitja.
Integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor) Homo Sapiens, 720 sequence(s) - Nidogen 2 (osteonidogen) Homo Sapiens, 200 sequence(s)
Setena parella de gens de la distribució final de clusters descrita a la figura 44
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens. Els mostres apareixen pintades d'un color o un altre depenent al cluster al que pertanyin. La distribució de clusters es la mateixa que la de les figures 45 i 47

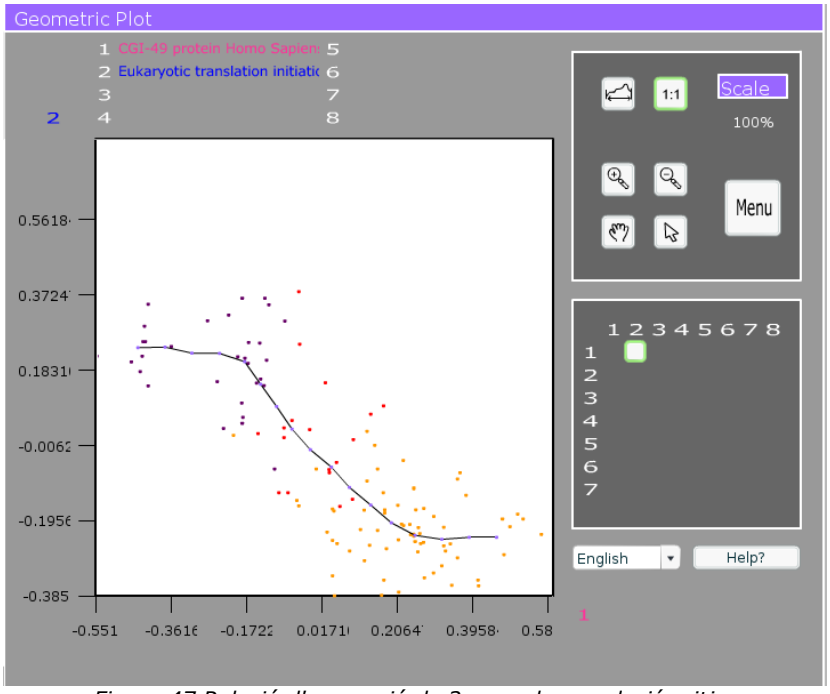
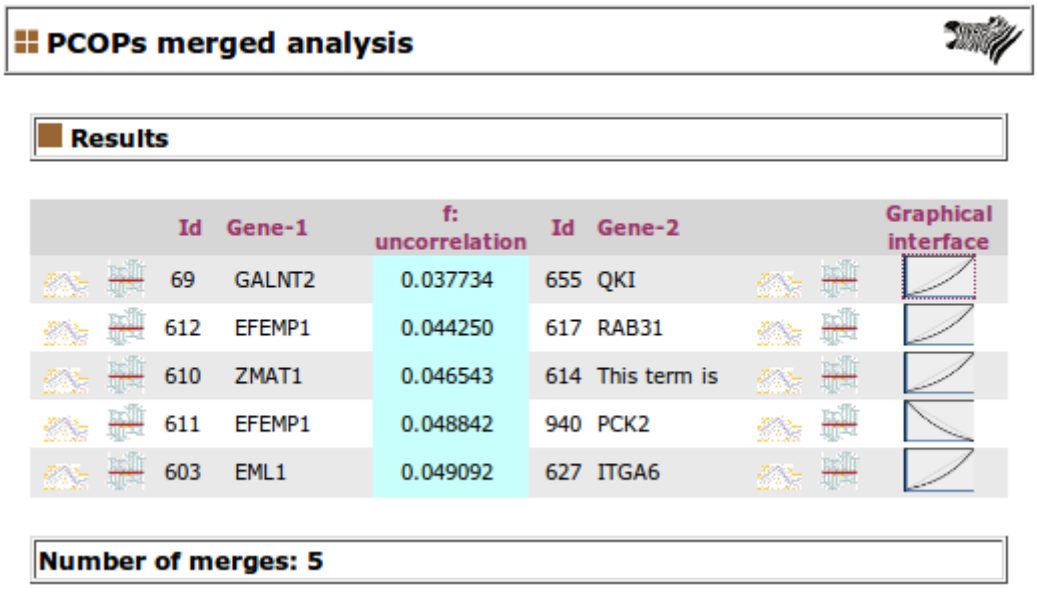


Figura 47 Relació d'expressió de 2 gens de correlació mitja.
CGI-49 protein Homo Sapiens, 456 sequence(s) -
Eukaryotic translation initiation factor 3, subunit 6 interacting protein Homo Sapiens, 3425 sequence(s)
Dotzena parella de gens de la distribució final de clusters descrita a la figura 44
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens. Els mostres apareixen pintades d'un color o un altre depenent al cluster al que pertanyin. La distribució de clusters es la mateixa que la de les figures 45 i 46

Podem veure de les figures 45, 46 i 47, com la distribució de clusters finals separa de la mateixa manera els diferents fenotips en les 3 relacions d'expressió, encara que les corbes dels diferents parells de gens siguin diferents.

3.4.3 EXEMPLE 3 SOROLL A LES DISTRIBUCIONS DE CLUSTERS FINALS



	Id	Gene-1	f: uncorrelation	Id	Gene-2	Graphical interface
	69	GALNT2	0.037734	655	QKI	
	612	EFEMP1	0.044250	617	RAB31	
	610	ZMAT1	0.046543	614	This term is	
	611	EFEMP1	0.048842	940	PCK2	
	603	EML1	0.049092	627	ITGA6	

Figura X.3.1 Exemple de parelles de gens d'alta correlació, agrupades sota una mateixa distribució de clusters final. Característiques de la distribució de clusters final: Agrupació de distribucions de clusters assignant les mostres en zones d'intersecció al cluster més petit, amb %matching 85%, assignació màxims. Error de la distribució de cluster final=17,6

A la figura X.3.1 apareix un llistat de parelles de gens que s'agrupen sota la mateixa distribució de clusters final. Aquesta distribució final, tot i tenir un error acumulat dels més baixos, comporta una distorsió apreciable en la zona d'intersecció o canvi de tendencia en la relació d'expressió gènica, degut al tipus d'assignació de samples aplicat (selecció per màxims) en l'hora de constituir la distribució final dels clusters.

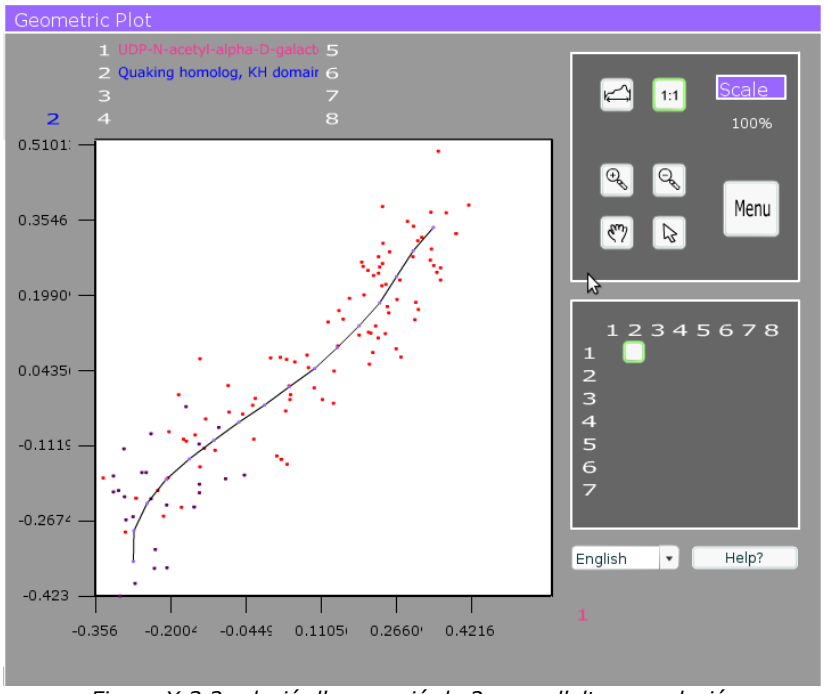


Figura X.3.2 relació d'expressió de 2 gens d'alta correlació.
UDP-N-acetyl-alpha-D-galactosamine - Quaking homolog, KH domain RNA binding (mouse) Homo Sapiens, 784 sequence(s)
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens. Els mostres apareixen pintades d'un color o un altre depenent al cluster al que pertanyin. La distribució de clusters es la mateixa que la de les figures X.3.3 i X.3.4

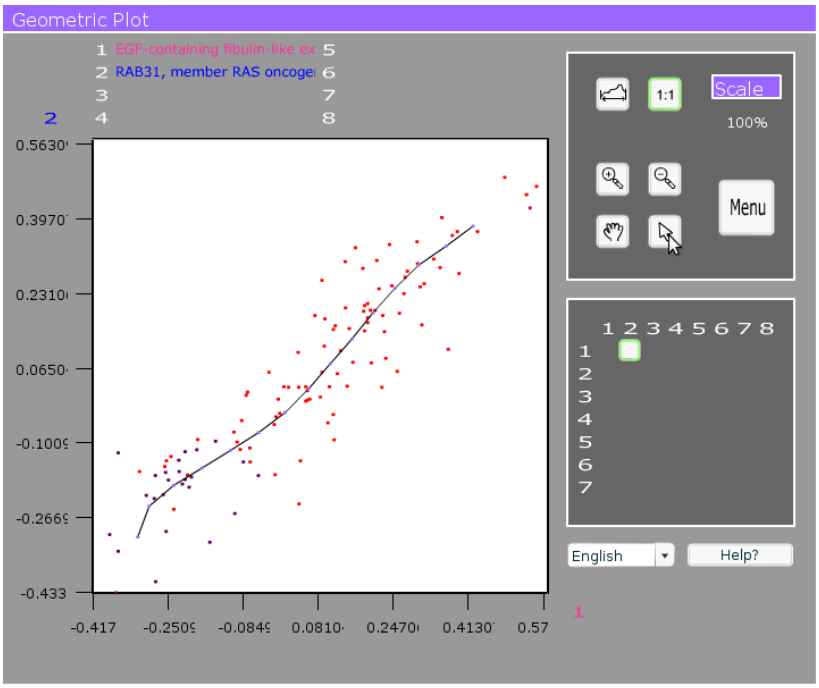


Figura X.3.3 Relació d'expressió de 2 gens d'alta correlació.
EGF-containing fibulin-like extracellular matrix protein 1 Homo Sapiens, 597 sequence(s) - RAB31, member RAS oncogene family Homo Sapiens, 775 sequence(s)
final de clusters descrita a la figura 40
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens. Els mostres apareixen pintades d'un color o un altre depenent al cluster al que pertanyin. La distribució de clusters es la mateixa que la de les figures X.3.1 i X.3.2

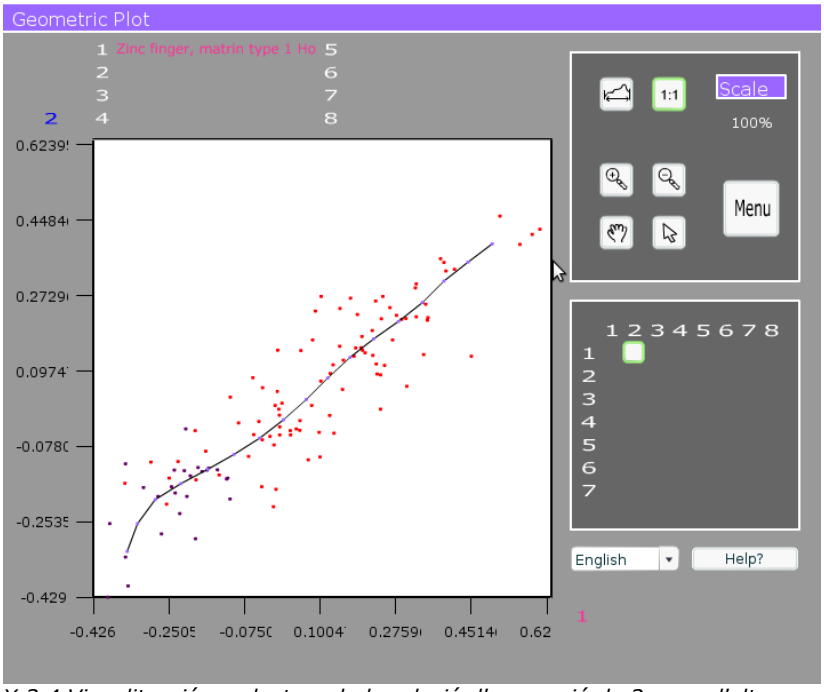


Figura X.3.4 Visualització en clusters de la relació d'expressió de 2 gens d'alta correlació.
EGF-containing fibulin-like extracellular matrix protein 1 Homo Sapiens, 597 sequence(s) - RAB31, member RAS oncogene family Homo Sapiens, 775 sequence(s)
A la imatge els punts son la mostra que compara els 2 valors d'expressió (component x i y) dels gens relacionat. La corba que es dibuixa, és la corba PCOP que descriu la relació d'expressió dels 2 gens. A la figura es veu la relació del nivell d'expressió de cada gen, en funció de la corba PCOP. La visualització en clusters segons la distribució de clusters final a la que pertany (figura X.3.1), es pot diferenciar pels colors de les mostres.

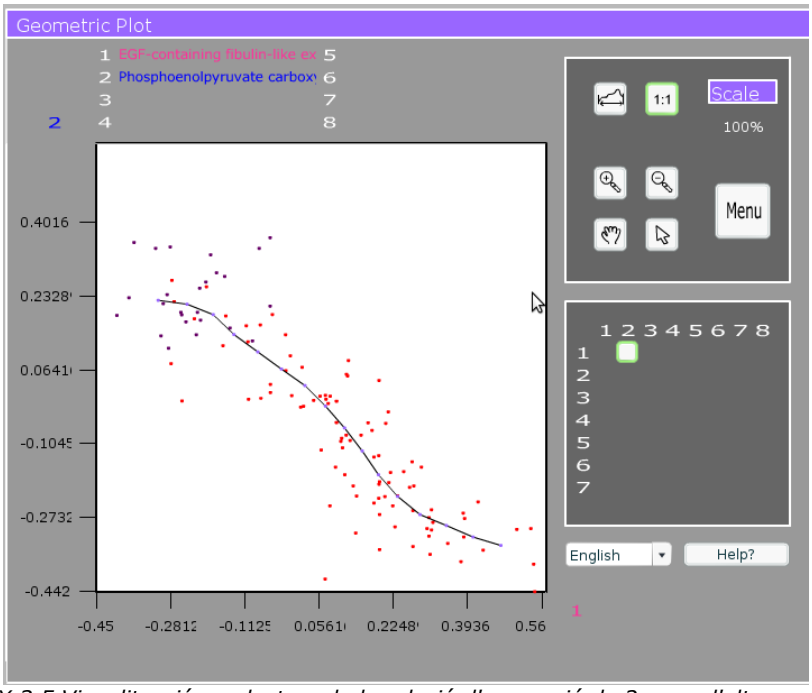


Figura X.3.5 Visualització en clusters de la relació d'expressió de 2 gens d'alta correlació.
EGF-containing fibulin-like extracellular matrix protein 1 Homo Sapiens, 597 sequence(s) - RAB31, member RAS oncogene family Homo Sapiens, 775 sequence(s)
A la imatge els punts son la mostra que compara els 2 valors d'expressió (component x i y) dels gens relacionat. La corba que es dibuixa, és la corba PCOP que descriu la relació d'expressió dels 2 gens.
A la figura es veu la relació del nivell d'expressió de cada gen, en funció de la corba PCOP.
La visualització en clusters segons la distribució de clusters final a la que pertany (figura X.3.1), es pot diferenciar pels

colors de les mostres.

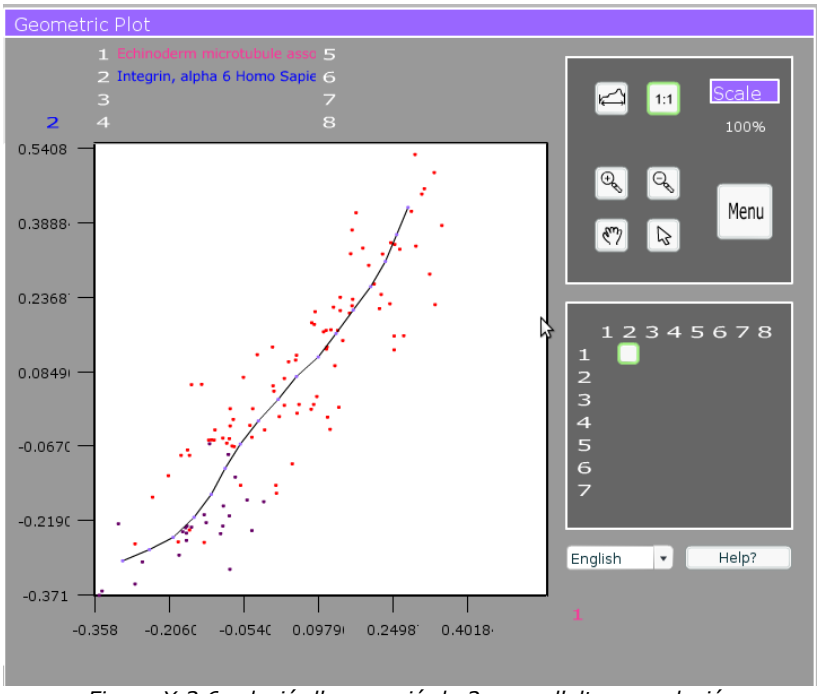


Figura X.3.6 relació d'expressió de 2 gens d'alta correlació.
EGF-containing fibulin-like extracellular matrix protein 1 Homo Sapiens, 597 sequence(s) - RAB31, member RAS oncogene family Homo Sapiens, 775 sequence(s)
El núbol de punts son les mostres i la corba es la PCOP que descriu la relació d'expressió dels 2 gens. Els mostres apareixen pintades d'un color o un altre depenent al cluster al que pertanyin. La distribució de clusters es la mateixa que la de les figures X.3.4 i X.3.5

Es pot apreciar a les figures X.3.2 a X.3.6, on es dibuixen les diferents relacions d'expressió gènica, com a totes hi ha bastant soroll a la zona d'intersecció de clusters, però que tot i així, es pot apreciar bé els canvis comuns que hi ha entre les diferents parelles de gens, degut a que totes elles continuen realitzant el mateix canvi fenotípic. El problema es que aquestes parelles de gens separen un fenotip molt gran (un cluster amb un gran nombre de mostres) d'un fenotip marginal (amb un nombre petit de mostres i nivells d'expressió molt extrems), pero s'ens compensa amb l'oportunitat d'estudiar aquestos fenotips excepcionals. El fet que ha provocat aquest soroll, es el mateix que fa interesant el seu estudi, si com en aquest cas, l'eina es capaç de detectar el canvi fenotípic i construir els clusters. Per un altre costat al seleccionar asignació per maxims el que fem es que al cluster petit només apareguin les mostres que realment apareixen més a aquest cluster que al cluster gran. Això pot resultar interesant per estudiar les mostres que realment pertanyen a aquest fenotip, pero si es vol estudiar fins on s'estèn aquest fenotip dona millors resultats fer asignació per t-student i així s'asignen a aquest fenotip les mostres que apareguin un suficient número de vegades.

4 CONCLUSIONS

Els objectius del projecte han estat complerts molt satisfactòriament. S'ha aconseguit disenyar un mètode per agrupar les distribucions de clusters obtingudes de l'anàlisi de microarrays (secció 2). També s'ha desenvolupat un mòdul d'una aplicació web per poder analitzar i visualitzar els resultats obtinguts (secció 3.3). Ambdós objectius s'han complert i resten totalment operatius.

S'ha fet una integració del programa C++ al entorn de processos d'anàlisis de microarrays del IBB-UAB. També s'ha pogut integrar el mòdul web al entorn web ja existent també en el IBB-UAB.

Els objectius d'aplicació en el camp de l'investigació biomèdic han complert les expectatives que motivaven desenvolupar un mètode i la corresponents eina d'anàlisi. Gràcies a l'eina desenvolupada, ara es poden estudiar grups de mostres corresponents a diferents fenotips juntament amb els gens que hi participen en el canvi d'un fenotip a un altre. El gran número de parelles de gens involucrades en les diferents distribucions de clusters finals i que l'eina ha sapigut trobar, dona validesa als clusters obtinguts, pero sobretot facilita molt el poder entrendre la naturalesa biológica d'aquests clusters.

Les futures lines de desenvolupament s'enamarcarien en 2 linies d'actuació: La primera clasificar les distribucions de clusters d'entrada entre les distribucions finals trobades i comprovar si les distribucions agrupades sota una mateixa distribució final continuen sent representades per la distribució final. Com que l'algorisme d'agrupació de distribucions de clusters està basat en l'aprenentatge per reforç, les primeres distribucions agrupades s'han fet sobre distribucions amb menys reforç que les últimes. La segona, millorar el métoda de clustering per segments de la poligonal de la corva de forma que les distribucions de clusters d'entrada tinguessin menys soroll, amb clusters més ben definits i amb menys mostres a la zona d'intersecció de clusters.

5 INFORME TÈCNIC

5.1 ESTRUCTURA DEL SERVIDOR.

A continuació es descriu com està configurat el servidor en quan als programes que s'utilitzen per al càlcul de PCOPs i agrupació de distribucions de clusters, com la estructura de directoris i fitxers d'emmagatzemament de dades utilitzats pels diferents programes.

5.1.1 ESTRUCTURA DE DIRECTORIS

L'estructura de directoris definida al servidor, unifica els criteris d'accés a les dades per a tots els processos que es puguin executar, donant així una imatge compacte de procés global en els processos que executa el servidor. D'aquesta manera, tots els processos que intervenen en els diferents càlculs saben on han de llegir les dades que necessitin i on han de deixar els resultats per a altres processos.

L'estructura de directoris al servidor tenen una configuració fixe, encara que no s'utilitza mai directoris absoluts per a l'accés de dades i execució de processos, si no que tots els directoris son relatius respecte al directori arrel de l'aplicació. D'aquesta manera, podem moure tots els processos de directori o màquina sense que s'hagi de tornar a configurar res.

Donat un directori arrel on s'instal·la els paquet d'aplicacions, tots els directoris de processos, dades i fitxers anexos son relatius a aquest. Per tant, els programes accedeixen a les dades a un directori relatiu i no absolut per facilitar aquesta movilitat.

```
... / (directori arrel on a partir d'aquest instal·lem els programes)
    |
    |__ microarray (directori arrel de dades de microarrays)
    |   |
    |   |__ mXX  (directori arrel de les dades de la microarray XX)
    |   |   |
    |   |   |__ factors
    |   |   |
    |   |   |__ nonlineal (directori arrel dels fitxers de les relacions no lineals de gens)
    |   |       |
    |   |       |__ normal
    |   |       |
    |   |       |__ class
    |   |       |
    |   |       |__ HeighF
    |   |       |
    |   |       |__ HeighFfiltered
    |   |       |
    |   |       |__ classHeighF
    |   |       |
    |   |       |__ RH_F10filtered
    |   |       |
    |   |       |__ classRH_F10
    |   |       |
    |   |       |__ arquetypes (directori de fitxers de configuracions finals de clusters)
    |   |           |
    |   |           |__ normal
    |   |           |   |
    |   |           |   |_ colors
    |   |           |
    |   |           |__ HeighF
    |   |           |   |
    |   |           |   |_ colors
    |   |           |
    |   |           |__ HeighFfiltered
    |   |           |   |
    |   |           |   |_ colors
    |   |           |
    |   |           |__ RH_F10filtered
    |   |           |   |
    |   |           |   |_ colors
    |   |           |
    |   |           |__ images
    |   |
    |   |__ mYY  (directori arrel de les dades de la microarray YY)
    |   |   |
    |   .   |...
    |   .
    |
    |
    |__ fullcorrelations (directori del programa "lanzadora" i tots els executables del servidor)
        |
        |__ compile (directiori arrels dels fonts dels diferents programes del servidor)
            |
            |__ factors (directiori dels fonts del programa factors2samples)
            |
            |__ samplestopop (directori dels fonts del programa pcop)
            |
            |__ localdomains (directori dels fonts del programa pop2domains)
            |
            |__ HeighFfilter (directori dels fonts del programa HeighFfilter)
            |
            |__ gencluster (directori dels fonts del programa gencluster)
            |
            |__ curv2class (directori dels fonts del programa curv2class)
            |
            |__ pcop_clustering (directori dels fonts del programa pcop_clustering)
```

Figura 5.1 Estructura de directoris de les dades de les microarrays i els programes d'anàlisi del preprocés.

Els directoris marcats en "**negreta**" son els generats per al procés d'agrupació de distribucions de clusters.

El directori arquetypes, conté els resultats de les distribucions finals de clusters. Els 4 directoris que hi ha, a banda del directori images que és exclusi de la web, pertanyen a les diferents agrupacions de corbes PCOP analitzades en funció de la seva correlació. S'hi poden afegir més directoris si els processos anteriors detecten altres tipologies de corbes .

Estructura de directoris del servidor web:

Tots els fitxers que s'han creat al servidor, han estat anomenats amb el prefix "arq_",

per poder detectar-los i saber que el mòdul web al que pertanyen és el de visualització de distribucions finals de clusters. En negreta es marquen els directoris específics del mòdul web programat en aquest projecte.

```
... / (directori arrel del servidor WEB)
    |
    |__ applic
    |   |
    |   |__ gexp
    |   |   |
    |   |   |__ microarrays (directori arrel de les pàgines web del nostre aplicatiu)
    |   |   |   |
    |   |   |   |__ arquetipes (directori de les extensions del programa. Classes,funcions,config)
    |   |   |   |   |
    |   |   |   |   |__ config (directori de fitxers de configuració i diccionari dels labels web)
    |   |   |   |   |
    |   |   |   |   |__ includes (directori de funcions i classes per al tractament de la pàgina)
    |   |   |   |   |   |
    |   |   |   |   |   |__ classes
    |   |   |   |   |   |
    |   |   |   |   |   |__ funcions
    |   |   |   |   |
    |   |   |   |   |__ templates (directori de plantilles HTML per no barrejar-lo amb el PHP)
    |   |   |   |   |
    |   |   |   |   |__ js (directori per guardar els fitxers JS necessaris)
    |   |   |   |   |   |
    |   |   |   |   |   |__ yui (llibreria gratuita de yahoo per a tractar elements visuals HTML)
    |   |   |   |   |
    |   |   |   |   |__ images (directori de les imatges del programa web)
    |   |   |   |   |
    |   |   |   |__ microarrays (ULL !!! Soft link al directori arrel de les dades de microarrays)
    |   |   |   |   |
    |   |   |   |   |__ mXX (directori amb les dades de la microarrays XX).
...
...
```

Figura 5.2 Estructura de directoris del mòdul php programat per visualitzar les distribucions de clusters finals

5.1.2 PROGRAMA "LANZADORA"

El programa "lanzadora" és l'enacarregat d'executar ordenadament tots els processos de càlcul i agrupació de cada microarray. Aquest programa s'executa sota demanda des de la web cada vegada que es carrega una microarray al sistema. Degut a la gran magnitud del projecte (en processos i dades), aquest programa és l'encarregat de mantenir el control i ordre d'execució dels diferents processos. S'encarrega també de crear els directoris que necessita cada procés per als càlculs i als resultats.

S'ha modificat el programa, per afegir al final la execució del propgrama pcop_clustering.

(Veure ANEX-IV)

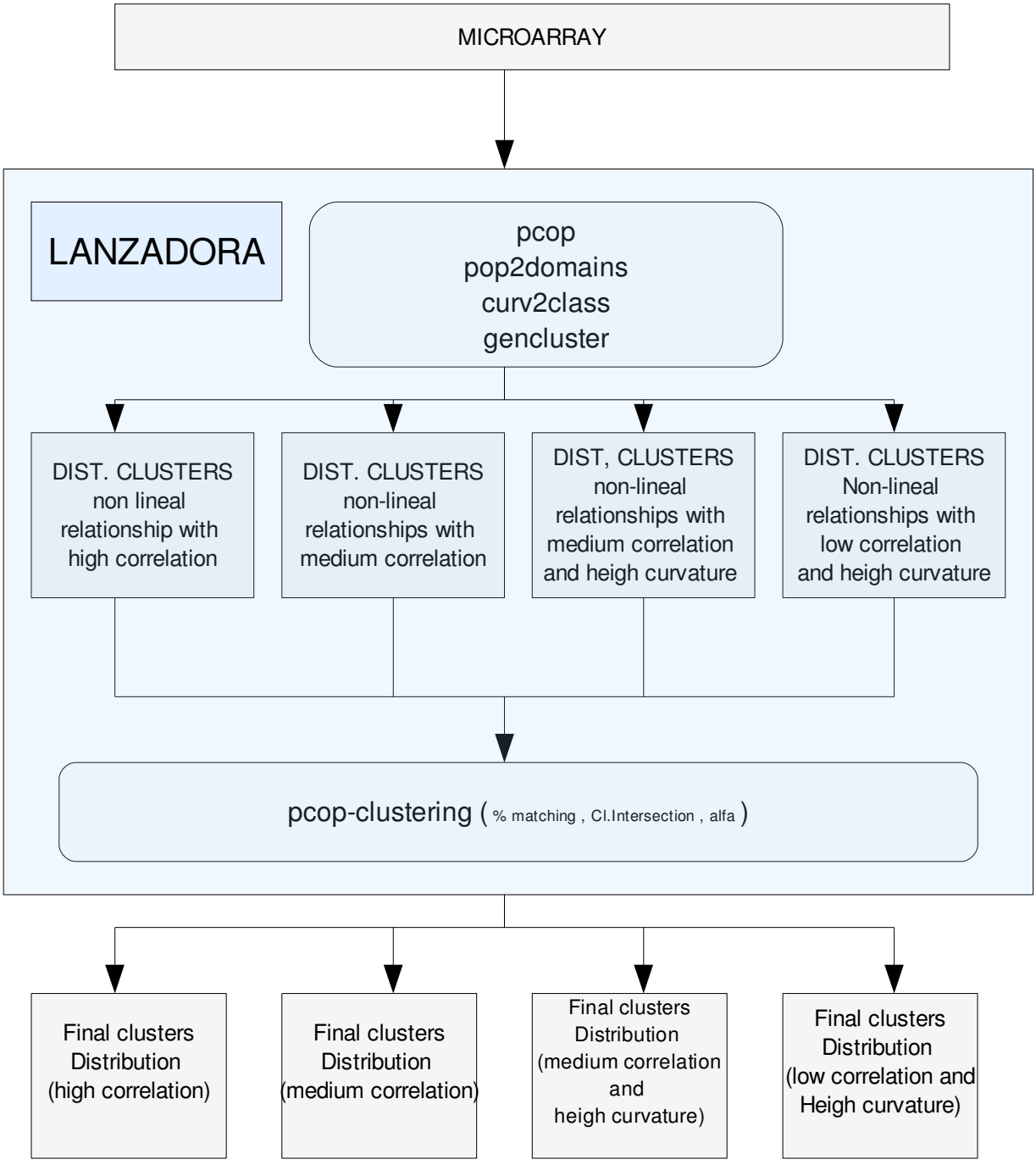


Figura 5.3 Esquema de processos executats en el servidor per a la obtenció de distribucions de clusters finals.

5.1.3 PROGRAMA "PCOP_CLUSTERING"

El programa d'agrupació de distribucions de clusters "pcop_clustering" és un programa escrit en C++. Es podria haver utilitzat un altre llenguatge de programació orientat a objectes com JAVA o d'altres, però s'ha optat per aquest, pels següents motius:

Tots els programes del servidor ja existents estan escrits en C o C++. Per tant, és manté unificat el criteri, i no cal conèixer més llenguatges si es volen fer canvis a qualsevol d'ells.

És un llenguatge compilat, i genera codi màquina optimitzat, característica molt necessària per la quantitat de càlculs que es fan.

És un llenguatge prou conegut i molt utilitzat en el món científic.

El servidor on es compila i s'executa el programa és un entorn Linux, que té com a llenguatge natiu el C i C++.

Un risc del llenguatge és que el programa ha d'estar ben escrit i estructurat, i el programador ha de tenir cura de les estructures de dades que crea i allibera en tot moment.

Execució del programa:

El programa "pcop_clustering" desenvolupat en aquest projecte només necessita 2 paràmetres d'execució. El número de la microarray on estan les distribucions de clusters a agrupar, i el nom del directori específic on es troben aquestes, en funció de la correlació de les seves corbes PCOPs. L'encarregat d'executar-lo és el programa "lanzadora" que executa ordenament els processos per a l'anàlisi de la microarray.

Amb aquests 2 paràmetres, el programa accedeix al directori de la microarray, i agrupa les distribucions de clusters del directori seleccionat en distribucions finals de clusters. Els resultats, els deixa en un directori amb el mateix nom, però dins del directori arquetypes. El directori "colors" de cada tipologia de corbes agrupades, conté tots els fitxers "colors" que pertanyen a totes les distribucions finals de clusters trobades. Cada fitxer colors, pertany només a una distribució de clusters final. Aquest fitxer descriu per totes les samples de la microarray, a quin clúster queda assignada cada una, en funció de la

distribució de clusters final a la que pertany el fitxer. S'ha pogut veure a les pantalles web on es mostren les corbes de les relacions de gens, l'ús d'aquest fitxer, que permet per cada corba, dibuixar el punts de les samples en funció del clúster al que pertanyen.

5.1.4 Classes i Estructures de dades del programa de agrupació de clusters

Les classes i estructures de dades utilitzades en el proceś d'agrupació de distribucions de clusters PCOP (pcop_clustering) son dinàmiques i adaptades a les necessitats del problema.

El programa escrit en C++, té com a classes principals les següents:

Classe **ldomains**: Aquesta classe s'encarrega d'agrupar ordenadament les distribucions de clusters, i guarda

Classe **ll_q**: Classe que guarda la llista de distribucions finals de clusters agrupades, i mètodes necessaris per comparar distribucions de clusters i generació de resultats o distribucions finals.

El procés permet tractar les distribucions de clusters extretes de microarrays d'un número indeterminat de condicions mostrals i de gens, és a dir, està preparat per analitzar i extreure les distribucions finals de clusters qualsevol microarray.

Les estructures de dades que s'utilitzen es creen de forma dinàmica, i per tant només hi ha la limitació de capacitat de memòria en el servidor. El programa "pcop_clustering" està lliure de memory-leaks.

5.1.5 llibreries utilitzades

Per al programa escrit en C++, s'ha necessitat un llibreria per fer els càlculs de la Inversa de T-Student. S'ha utilitzat una llibreria externa descarregada de la web "http://www.alglib.net/". Aquesta llibreria s'ha testejat abans de ser afegida al programa d'agrupació de distribucions de clusters, assegurantt que el valor calculat de la taula T-Student inversa era correcta per a un volum alt deparàmetress de la T-Student.

A banda de les funcions necessàries per al càlcul de la T-Student, no s'ha necessitat cap més llibreria per al programa que les normals de qualsevol programa escrit en C o C++.

5.1.6 Distribucions de clusters d'entrada - fitxers ldom

Els fitxers d'entrada al procés d'agrupació "pcop_clustering", tenen la següent estructura:

La extensió del fitxer és "ldom".

El nom del fitxer ha de contenir els 2 gens de la microarray que formen les corbes o clusters a agrupar, com per exemple "g10g786h0.75d0.3.ldom", on g10 és el gen 10 (IFITM2: Interferon induced transmembrane protein 2 (1-8D) Homo Sapiens, 650 sequence(s)) de la microarray, i el g786 és el gen 786 (SLC36A4: Solute carrier family 36 (proton/amino acid symporter), member 4 Homo Sapiens, 102 sequence(s)).

El contingut del fitxer "ldom" conté la següent informació:

```
popIni 0 2 5 8 12 15

corbes

       C 0 2 5 8 12 15

popFi 0 6 15

       popFiCorbes1 0 6 15

segmentsCorbes1

       SEGMENT: 1 - 5 (55 mostres)
       DOMINI: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 25 26 27 28
29 30 31 32 33 34 35 36 37 38 41 42 44 80 81 82 83 84 85 86 88 90 91 107 110 112 117 118

       SEGMENT: 7 - 14 (93 mostres)
       DOMINI: 2 3 9 10 21 24 26 28 29 31 32 33 35 37 38 39 40 41 42 43 44 45 46 47 48
49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77
78 79 81 82 83 84 85 86 8
7 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 110 111 112
113 114 115 116 117 118
```

La seva representació gràfica de subespais és:

Figura 5.4 Exemple del contingut d'un fitxer LDOM i com és la seva representació de la distribució de clusters

El primer apartat "segmentCorbes" (hi ha fitxers amb més d'una distribució de clusters, però la primera és la més representativa) és el que sempre el llegeix el programa pcop_clustering, i ens diu quines samples estan assignades al cada cluster (valors de les samples després del subapartar DOMINI), on DOMINI representa el subespai de cada segment. Dels valors de les samples de cada DOMINI, el programa extreu la seva distribució de clusters.

5.1.7 Fitxers resultats

Els procés d'agrupació de distribucions de clusters genera diferents tipus de fitxers, segons la informació que hi guarda, i en diferents directoris.

En el directori "arquetypes", dins de cada subdirectori per factor de correlació (filtre PCOP input data), el procés "pcop_clustering" hi deixa els fitxers amb les distribucions finals de clusters trobades, segons els diferents criteris d'agrupació aplicats.

A cada un d'aquests directoris, el programa també deixa tots els fitxers colors associats a totes les distribucions finals de clusters.

Els fitxers que s'hi deixen son els d'extenció "max" i "var".

Els fitxers "max" contenen les distribucions de clusters finals agrupades per "màxims".

Els fitxers "var" contenen les distribucions de clusters finals agrupades per "T-Student".

En el nom de cada fitxer hi ha explícit els criteris de la agrupació de les distribucions finals de clusters que s'hi defineixen.

Els fitxers "max" tenen el següent patró de nom:

m<microarray>_n_<%matching>_h_<cl.intersection>.max

on:

<microarray>: Número de la microarray que es consulta

<%matching>: Percentatge de semblança mínima de clusters exigida(60, 65, 75, 85, 90, 95)

<cl.intersection>: Tractament que es dona a la zona d'intersecció de clusters. Aquest té 2 valors possibles.

1: Descarting joint samples

2: Assingning joint samples to the smaller cluster

Per exemple, el fitxer de distribucions finals de clusters de la microarray 17, amb %matching = 75%, "Descarting joint samples" i selecció per màxims és:

m17_n_75_h_1.max

Els fitxers "var" tenen el següent patró de nom:

m<microarray>_n_<%matching>_a_<alfa>h_<cl.intersection>.var

on els paràmetres son els mateixos que per als fitxers "max" menys:

<alfa>: És el valor de alfa representat amb 1 enter i 3 decimals després d'un punt.

Per exemple, el fitxer de distribucions finals de clusters de la microarray 17, amb %matching = 75%, "Assingning joint samples to the smaller cluster" i selecció per T-Student amb $\alpha = 0.003$ és:

m17_n_75_a_0.003_h_2.var

Es pot veure un exemple del contingut del fitxer "var" en el ANEX-I.

5.2 INTERFÍCIE GRÀFICA

5.2.1 Configuració de l'entorn

El programa PHP desenvolupat en aquest projecte, ha estat dissenyat i programat de forma independent a la resta de mòduls del servidor, però fàcil d'integrar-lo a ell, ja que les referències a tots els objectes que es fan des de l'aplicació, son sempre relatives al directori arrel de l'aplicació del servidor.

Hi ha un fitxer de configuració on es defineixen les referències relatives als directoris del programa, els diferents valors i textos de les combos del programa, i l'accés a pàgines d'altres servidors com la del "www.ncbi.nlm.nih.gov". (Veure ANEX-III)

5.2.2 Parser del fitxer de distribucions finals de clusters.

En els fitxers "var" i "max" generats pel programa pcop_clustering, estan definides les distribucions finals de clusters trobades per a cada un dels criteris utilitzats en el procés d'agrupació. El PHP és un llenguatge que es caracteritza per la seva velocitat en la lectura de fitxers, el tractament de cadenes de text i en la ordenació d'estructures de dades grans, per tant, s'ha pogut programar sense gaires dificultats un lector de fitxers de text, per poder parsejar les diferents distribucions finals de clusters definides en els fitxers "var" i "max" (Veure *ANEX II EXEMPLE FITXER VAR DE DITRIBUCIONS FINALS DE CLUSTERS*)

5.2.3 Llibreries PHP necessàries

Per al programa escrit en PHP, és necessari que el servidor web tingui instal·lat i configurat el mòdul GD de la llibreria gràfica de PHP. Aquest mòdul conté funcions gràfiques, que han permès concretament visualitzar elements text en pantalla en format vertical.

5.2.4 *IMATGES VERTICALS DELS NOMS DE LES SAMPLES*

S'han programat una sèrie de funcions PHP (utilitzant les llibreries gràfiques del mòdul GD de PHP) que permeten a la web dibuixar textos en format vertical. Només cal fer la crida al mòdul php "arq_getimagetext.php" que retorna la imatge en format GIF en funció del text passat per paràmetre i els seus atributs com color de fons, color de text, orientació del text en graus a la imatge, tamany de lletra, ajustament del text en píxels a la esquerra i bottom de la imatge.

La crida HTML específica que es fa per exemple per dibuixar el nom d'una condició mostral o sample és:

```
<img src="arq_getimagetext.php?text="271674-A7-ALK-A7-C1-:-:-Diaminocyclohexyl"
        title="271674-A7-ALK-A7-C1-:-:-Diaminocyclohexyl-Pt-II(S27167408MCN)">
```

Per dibuixar en vertical els títols de les columnes "Matchings" i "Error" es fa amb la mateixa crida però diferents paràmetres de formateig de la imatge:

```
<img title="Click to order by Matching"
        onclick="document.formulari.column_order.value=1; document.formulari.submit();"
src="arq_getimagetext.php?text=MATCHINGS <<< &amp;bgred=100&amp;bgblue=12"/>


<img title="Click to order by Error"
        onclick="document.formulari.column_order.value=2;document.formulari.submit();"
src="arq_getimagetext.php?text=ERROR >>> &amp;
        bgred=140&amp;bgblue=10&amp;bggreen=100"/>
```

Aquesta utilitat, ha permès visualitzar en un espai reduït totes les mostres de la microarray (118 mostres) i facilitar la comprensió de la distribució de clusters en caixes relativament petites com es pot veure en la figura 5.5.

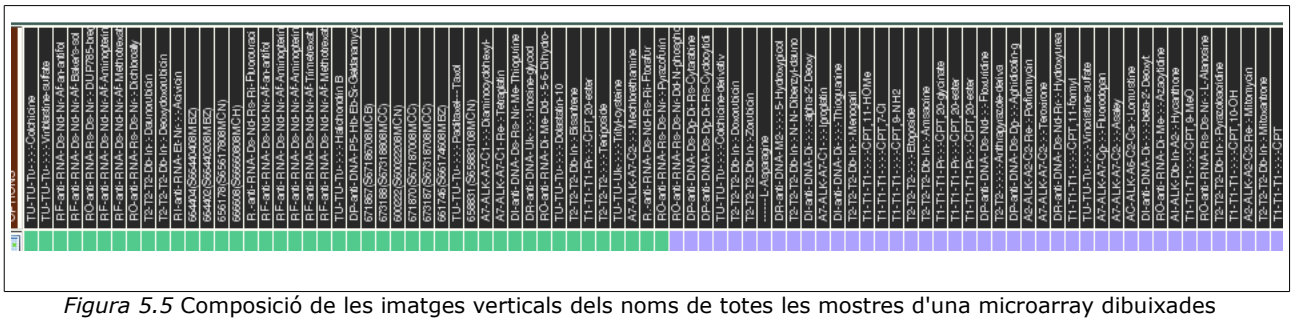


Figura 5.5 Composició de les imatges verticals dels noms de totes les mostres d'una microarray dibuixades

5.2.5 *Imatges iconogràfiques de corbes per representar les configuracions.*

En la pantalla web on es mostra el llistat de parelles de gens relacionats en el mateix canvi d'estat fenotípic, o que pertanyen a la mateixa distribució de clusters finals, s'hi mostra una imatge aproximada de la orientació de la corba PCOP que descriu cada parella de gens. El tipus de corba PCOP i la seva orientació, que descriuen cada parella de gens, ja està calculada en el Preprocés (curv2class). Aquest deixa un fitxer "relbyGen.txt" en el directori "class" de les dades consultades. En aquest fitxer hi ha la definició de la orientació de totes les corbes de tots els parells de gens del directori. Com la definició del tipus de corba en el fitxer "relbyGen.txt" és mnemotècnica, cal interpretar-la i transformar-ho en el nom de la imatge GIF, que estan guardades en directori "curveimages" ja esmentat abans en la estructura del servidor.

S'ha programat una classe en PHP per parsejar aquest fitxer i trobar la definició de les corbes de qualsevol parella de gens consultada. La classe permet a partir de 2 gens retornar el nom del fitxer imatge "GIF", que pertany a la corba de la relació d'aquests 2 gens.

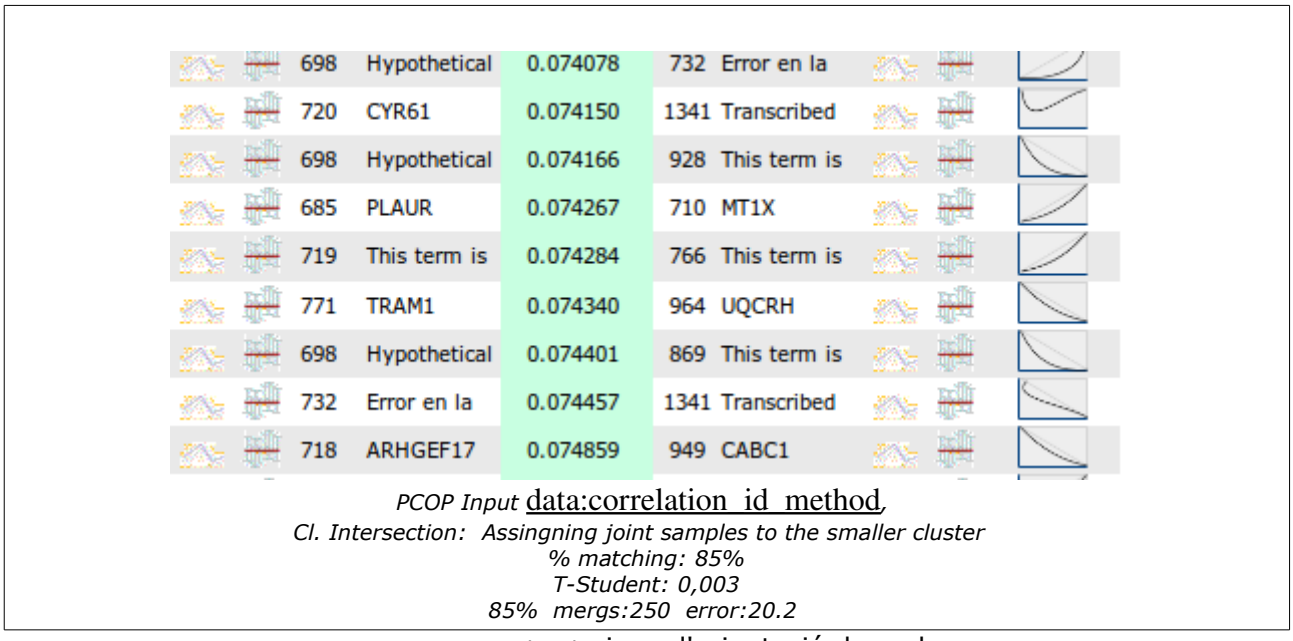


Figura 5.6 Exemples de tipus d'orientació de corbes

5.3 Entorn i eines de desenvolupament

El desenvolupament del projecte s'ha fet tot amb eines open-source, menys l'entorn de programació PHP.

Tot el desenvolupament s'ha fet i provat en entorns Linux.
El servidor web utilitzat és un LAMP (Linux+Apache+MySql+Php).
Per programar en C++ s'ha utilitzat l'entorn Eclipse Galileo, que ja porta integrat el compilador de C++.
L'entorn de desenvolupament en PHP que s'ha utilitzat és el ZendStudio 7 Unregisterd, basat en eclipse. Si no es té la llicència, l'entorn queda restringit, i no deixa utilitzar algunes de les seves utilitats, però que no han estat requerides per al desenvolupament del projecte. Només s'ha utilitzat l'entorn per a programar.
Per fer la memòria del projecte s'ha utilitzat el OpenOffice 3.0.
Per al control de versions, s'ha utilitzat el Subversion amb 2 repositoris. El del programa C++ i el PHP.

6 ANEX I – TAULES DE RESULTATS ESTADÍSTICS

			Non-lineal relationships with high correlation					
	% Matching	Samples classification	#Distribució clusters	#Distribució clusters final	#clusters Arrangement	Max Arrangement	Min Error	Max Error
Descarting joint samples	60,00%	T-St 0,003	146	9	110	31	15,5	59,12
		T-St 0,009	146	9	110	31	15,5	59,12
		T-St 0,05	146	9	110	31	15,5	40,33
		T-St 0,10	146	9	110	31	15,5	35,12
		T-St 0,15	146	9	110	31	15,5	35,12
		T-St 0,40	146	9	110	31	15,5	32,33
		Max	146	9	110	31	13,4	23,5
	65,00%	T-St 0,003	146	10	108	28	15,5	59,12
		T-St 0,009	146	10	108	28	15,5	59,12
		T-St 0,05	146	10	108	28	15,5	42,12
		T-St 0,10	146	10	108	28	15,5	36,88
		T-St 0,15	146	10	108	28	15,5	36,88
		T-St 0,40	146	10	108	28	15,5	36,88
		Max	146	10	108	28	12,38	25,62
	75,00%	T-St 0,003	146	14	99	36	15,5	59,12
		T-St 0,009	146	14	99	36	15,5	59,12
		T-St 0,05	146	14	99	36	15,5	37,12
		T-St 0,10	146	14	99	36	15,5	36,5
		T-St 0,15	146	14	99	36	15,5	36,5
		T-St 0,40	146	14	99	36	15,5	30,75
		Max	146	14	99	36	10,75	25,9
	85,00%	T-St 0,003	146	13	77	11	15,5	71
		T-St 0,009	146	13	77	11	15,5	71
		T-St 0,05	146	13	77	11	15,5	36,83
		T-St 0,10	146	13	77	11	15,5	36,83
		T-St 0,15	146	13	77	11	15,5	36,83
		T-St 0,40	146	13	77	11	15,5	27,5
		Max	146	13	77	11	10,75	25,5
	90,00%	T-St 0,003	146	13	55	12	13,25	37,9
		T-St 0,009	146	13	55	12	13,25	37,9
		T-St 0,05	146	13	55	12	13,25	35,5
		T-St 0,10	146	13	55	12	13,25	35,5
		T-St 0,15	146	13	55	12	13,25	35,5
		T-St 0,40	146	13	55	12	13,25	28,17
		Max	146	13	55	12	9,25	24,17
	95,00%	T-St 0,003	146	7	20	7	14	40,36
		T-St 0,009	146	7	20	7	14	35,21
		T-St 0,05	146	7	20	7	14	26,5
		T-St 0,10	146	7	20	7	14	26,5
		T-St 0,15	146	7	20	7	14	26,5
		T-St 0,40	146	7	20	7	14	26,5
		Max	146	7	20	7	10	24

			Non-lineal relationships with high correlation					
	% Matching	Samples classification	#Distribució clusters	#Distribució clusters final	#clusters Arrangement	Max Arrangement	Min Error	Max Error
Assingning joint samples to the smaller cluster	60,00%	T-St 0,003	146	11	117	32	22,83	61,19
		T-St 0,009	146	11	117	32	22,83	53,75
		T-St 0,05	146	11	117	32	19,74	52,88
		T-St 0,10	146	11	117	32	19,74	35,38
		T-St 0,15	146	11	117	32	19,74	35,38
		T-St 0,40	146	11	117	32	13,83	35,38
		Max	146	11	117	32	11,17	25,62
	65,00%	T-St 0,003	146	10	112	52	23,67	46,5
		T-St 0,009	146	10	112	52	21,38	46,5
		T-St 0,05	146	10	112	52	18,38	46,5
		T-St 0,10	146	10	112	52	15,38	46,5
		T-St 0,15	146	10	112	52	15,38	46,5
		T-St 0,40	146	10	112	52	15,38	37,5
		Max	146	10	112	52	11,12	25,95
	75,00%	T-St 0,003	146	14	102	35	11,83	54,43
		T-St 0,009	146	14	102	35	11,83	40
		T-St 0,05	146	14	102	35	11,83	40
		T-St 0,10	146	14	102	35	11,83	40
		T-St 0,15	146	14	102	35	11,83	40
		T-St 0,40	146	14	102	35	11,83	29,2
		Max	146	14	102	35	8,5	24,17
	85,00%	T-St 0,003	146	22	81	9	13,25	35,75
		T-St 0,009	146	22	81	9	13,25	35,75
		T-St 0,05	146	22	81	9	13,25	35,75
		T-St 0,10	146	22	81	9	13,25	33
		T-St 0,15	146	22	81	9	13,25	33
		T-St 0,40	146	22	81	9	11,8	32
		Max	146	22	81	9	9,25	26
	90,00%	T-St 0,003	146	25	66	6	11,83	31,75
		T-St 0,009	146	25	66	6	11,83	31,75
		T-St 0,05	146	25	66	6	11,83	31,75
		T-St 0,10	146	25	66	6	11,83	31,75
		T-St 0,15	146	25	66	6	11,83	31,75
		T-St 0,40	146	25	66	6	11,83	31,75
		Max	146	25	66	6	8,5	28,25
	95,00%	T-St 0,003	146	8	20	4	10,25	26,5
		T-St 0,009	146	8	20	4	10,25	26,5
		T-St 0,05	146	8	20	4	10,25	26,5
		T-St 0,10	146	8	20	4	10,25	26,5
		T-St 0,15	146	8	20	4	10,25	26,5
		T-St 0,40	146	8	20	4	10,25	26,5
		Max	146	8	20	4	8,25	24

			Non-lineal relationships with medium correlation					
	% Matching	Samples classification	#Distribució clusters	#Distribució clusters final	#clusters Arrangement	Max Arrangement	Min Error	Max Error
Assingning joint samples to the smaller cluster	60,00%	T-St 0,003	4430	69	3503	1173	19	60,6
		T-St 0,009	4430	69	3503	1173	19	60,6
		T-St 0,05	4430	69	3503	1173	19	51,11
		T-St 0,10	4430	69	3503	1173	19	51,11
		T-St 0,15	4430	69	3503	1173	19	51,11
		T-St 0,40	4430	69	3503	1173	19	46,67
		Max	4430	69	3503	1173	12,5	34,67
	65,00%	T-St 0,003	4430	75	3433	1077	19	58,33
		T-St 0,009	4430	75	3433	1077	19	58,33
		T-St 0,05	4430	75	3433	1077	19	51,11
		T-St 0,10	4430	75	3433	1077	18,31	51,11
		T-St 0,15	4430	75	3433	1077	18,31	51,11
		T-St 0,40	4430	75	3433	1077	16,77	38,25
		Max	4430	75	3433	1077	12,46	29,25
	75,00%	T-St 0,003	4430	97	3150	556	19	63,25
		T-St 0,009	4430	97	3150	556	19	63,25
		T-St 0,05	4430	97	3150	556	19	44,42
		T-St 0,10	4430	97	3150	556	18,25	42
		T-St 0,15	4430	97	3150	556	18,25	42
		T-St 0,40	4430	97	3150	556	15,67	42
		Max	4430	97	3150	556	12	29,17
	85,00%	T-St 0,003	4430	114	2900	388	17,25	50,86
		T-St 0,009	4430	114	2900	388	17,25	47,17
		T-St 0,05	4430	114	2900	388	17,25	47
		T-St 0,10	4430	114	2900	388	13,58	43,31
		T-St 0,15	4430	114	2900	388	13,58	39,83
		T-St 0,40	4430	114	2900	388	13,58	39,56
		Max	4430	114	2900	388	8,92	27,38
	90,00%	T-St 0,003	4430	135	2600	297	13	49,5
		T-St 0,009	4430	135	2600	297	13	45,5
		T-St 0,05	4430	135	2600	297	13	45,5
		T-St 0,10	4430	135	2600	297	13	40,4
		T-St 0,15	4430	135	2600	297	13	40,4
		T-St 0,40	4430	135	2600	297	13	34,4
		Max	4430	135	2600	297	8,92	27,4
	95,00%	T-St 0,003	4430	218	1628	110	7,75	44,57
		T-St 0,009	4430	218	1628	110	7,75	39,92
		T-St 0,05	4430	218	1628	110	7,75	38,5
		T-St 0,10	4430	218	1628	110	7,75	38,17
		T-St 0,15	4430	218	1628	110	7,75	38,17
		T-St 0,40	4430	218	1628	110	7,75	37,5
		Max	4430	218	1628	110	6,25	30,25

			Non-lineal relationships with medium correlation					
	% Matching	Samples classification	#Distribució clusters	#Distribució clusters final	#clusters Arrangement	Max Arrangement	Min Error	Max Error
Assingning joint samples to the smaller cluster	60,00%	T-St 0,003	4430	281	3987	1160	19	67,83
		T-St 0,009	4430	281	3987	1160	19	67,83
		T-St 0,05	4430	281	3987	1160	18	57,83
		T-St 0,10	4430	281	3987	1160	18	57,83
		T-St 0,15	4430	281	3987	1160	18	57,83
		T-St 0,40	4430	281	3987	1160	12,6	55,5
		Max	4430	281	3987	1160	11	34,17
	65,00%	T-St 0,003	4430	294	3872	1347	19	61,3
		T-St 0,009	4430	294	3872	1347	17,81	61,3
		T-St 0,05	4430	294	3872	1347	13,81	59,38
		T-St 0,10	4430	294	3872	1347	13,81	55,5
		T-St 0,15	4430	294	3872	1347	11,94	55,5
		T-St 0,40	4430	294	3872	1347	11,94	55,5
		Max	4430	294	3872	1347	9,19	35,91
	75,00%	T-St 0,003	4430	274	3550	578	11,25	55,42
		T-St 0,009	4430	274	3550	578	11,25	55,42
		T-St 0,05	4430	274	3550	578	11,25	51,67
		T-St 0,10	4430	274	3550	578	11,25	51,67
		T-St 0,15	4430	274	3550	578	11,25	51,67
		T-St 0,40	4430	274	3550	578	9,86	39,5
		Max	4430	274	3550	578	7,33	30,67
	85,00%	T-St 0,003	4430	330	2959	250	7	48,86
		T-St 0,009	4430	330	2959	250	7	45,9
		T-St 0,05	4430	330	2959	250	7	39,29
		T-St 0,10	4430	330	2959	250	7	39,29
		T-St 0,15	4430	330	2959	250	7	39,29
		T-St 0,40	4430	330	2959	250	7	37,25
		Max	4430	330	2959	250	7	30,75
	90,00%	T-St 0,003	4430	486	2433	175	7	43,75
		T-St 0,009	4430	486	2433	175	7	40,92
		T-St 0,05	4430	486	2433	175	7	37,33
		T-St 0,10	4430	486	2433	175	7	37,33
		T-St 0,15	4430	486	2433	175	7	37,33
		T-St 0,40	4430	486	2433	175	7	34,25
		Max	4430	486	2433	175	6,33	30,25
	95,00%	T-St 0,003	4430	386	1073	14	7	32,25
		T-St 0,009	4430	386	1073	14	7	32,25
		T-St 0,05	4430	386	1073	14	7	32,25
		T-St 0,10	4430	386	1073	14	7	32,25
		T-St 0,15	4430	386	1073	14	7	32,25
		T-St 0,40	4430	386	1073	14	7	32,25
		Max	4430	386	1073	14	6,75	30,25

			Non-lineal relationships with medium correlation					
	% Matching	Samples classification	#Distribució clusters	#Distribució clusters final	#clusters Arrangement	Max Arrangement	Min Error	Max Error
Assingning joint samples to the smaller cluster	60,00%	T-St 0,003	8861	114	6719	966	24,67	64,81
		T-St 0,009	8861	114	6719	966	24,67	54,94
		T-St 0,05	8861	114	6719	966	23,27	51
		T-St 0,10	8861	114	6719	966	23,27	50,96
		T-St 0,15	8861	114	6719	966	23,27	50,96
		T-St 0,40	8861	114	6719	966	14	50,83
		Max	8861	114	6719	966	10,67	35,5
	65,00%	T-St 0,003	8861	143	6580	1585	24,5	68,57
		T-St 0,009	8861	143	6580	1585	23,27	63,88
		T-St 0,05	8861	143	6580	1585	21,11	53
		T-St 0,10	8861	143	6580	1585	21,11	52,12
		T-St 0,15	8861	143	6580	1585	21,11	49,67
		T-St 0,40	8861	143	6580	1585	14	46,25
		Max	8861	143	6580	1585	10,67	33,08
	75,00%	T-St 0,003	8861	190	5815	896	12,25	69,44
		T-St 0,009	8861	190	5815	896	12,25	57,89
		T-St 0,05	8861	190	5815	896	12,25	54,38
		T-St 0,10	8861	190	5815	896	12,25	46,33
		T-St 0,15	8861	190	5815	896	12,25	46,33
		T-St 0,40	8861	190	5815	896	12,25	40,75
		Max	8861	190	5815	896	9,83	33,58
	85,00%	T-St 0,003	8861	208	4936	823	12,25	52,69
		T-St 0,009	8861	208	4936	823	12,25	48,3
		T-St 0,05	8861	208	4936	823	12,25	44,62
		T-St 0,10	8861	208	4936	823	12,25	40,83
		T-St 0,15	8861	208	4936	823	12,25	40,83
		T-St 0,40	8861	208	4936	823	12,25	38,5
		Max	8861	208	4936	823	8,67	32,75
	90,00%	T-St 0,003	8861	213	4389	826	12,25	50,4
		T-St 0,009	8861	213	4389	826	12,25	50,4
		T-St 0,05	8861	213	4389	826	12,25	43,38
		T-St 0,10	8861	213	4389	826	12,25	37,17
		T-St 0,15	8861	213	4389	826	12,25	37,17
		T-St 0,40	8861	213	4389	826	11,17	35,75
		Max	8861	213	4389	826	8,25	32,75
	95,00%	T-St 0,003	8861	318	2764	217	6,75	49,29
		T-St 0,009	8861	318	2764	217	6,75	44,75
		T-St 0,05	8861	318	2764	217	6,75	42
		T-St 0,10	8861	318	2764	217	6,75	35,5
		T-St 0,15	8861	318	2764	217	6,75	35,5
		T-St 0,40	8861	318	2764	217	6,75	35,5
		Max	8861	318	2764	217	6,75	28,75

			Non-lineal relationships with medium correlation					
	% Matching	Samples classification	#Distribució clusters	#Distribució clusters final	#clusters Arrangement	Max Arrangement	Min Error	Max Error
Assingning joint samples to the smaller cluster	60,00%	T-St 0,003	8861	714	7812	2313	12,25	62,27
		T-St 0,009	8861	714	7812	2313	12,25	61,61
		T-St 0,05	8861	714	7812	2313	12,25	56
		T-St 0,10	8861	714	7812	2313	12,25	56
		T-St 0,15	8861	714	7812	2313	12,25	56
		T-St 0,40	8861	714	7812	2313	12,25	56
		Max	8861	714	7812	2313	10,75	32,67
	65,00%	T-St 0,003	8861	728	7513	1907	15	54,67
		T-St 0,009	8861	728	7513	1907	15	54,67
		T-St 0,05	8861	728	7513	1907	14	54,38
		T-St 0,10	8861	728	7513	1907	14	49,44
		T-St 0,15	8861	728	7513	1907	14	49,44
		T-St 0,40	8861	728	7513	1907	11,8	49,44
		Max	8861	728	7513	1907	8,5	34,44
	75,00%	T-St 0,003	8861	683	6636	1438	11,25	66,75
		T-St 0,009	8861	683	6636	1438	11,25	61,75
		T-St 0,05	8861	683	6636	1438	11,25	53,25
		T-St 0,10	8861	683	6636	1438	11,25	43,78
		T-St 0,15	8861	683	6636	1438	11,25	43,78
		T-St 0,40	8861	683	6636	1438	11,25	42,5
		Max	8861	683	6636	1438	6,75	31,83
	85,00%	T-St 0,003	8861	657	5187	457	10,5	47,2
		T-St 0,009	8861	657	5187	457	10,5	47,2
		T-St 0,05	8861	657	5187	457	10,5	42,5
		T-St 0,10	8861	657	5187	457	10,5	42,17
		T-St 0,15	8861	657	5187	457	8,35	42,17
		T-St 0,40	8861	657	5187	457	8,35	38,75
		Max	8861	657	5187	457	7,25	33,25
	90,00%	T-St 0,003	8861	857	3996	220	9,25	44,5
		T-St 0,009	8861	857	3996	220	9,25	41,75
		T-St 0,05	8861	857	3996	220	9,25	39,5
		T-St 0,10	8861	857	3996	220	9,25	38,25
		T-St 0,15	8861	857	3996	220	6	38,25
		T-St 0,40	8861	857	3996	220	6	38,25
		Max	8861	857	3996	220	5,75	33,75
	95,00%	T-St 0,003	8861	562	1467	18	7,62	32
		T-St 0,009	8861	562	1467	18	7,62	32
		T-St 0,05	8861	562	1467	18	7,62	32
		T-St 0,10	8861	562	1467	18	7,62	31,17
		T-St 0,15	8861	562	1467	18	7,62	31,17
		T-St 0,40	8861	562	1467	18	7,62	31
		Max	8861	562	1467	18	5,38	28,5

7 ANEX II EXEMPLE FITXER VAR DE DITRIBUCIONS FINALS DE CLUSTERS

```
 ---------------- N O D E  -  2 ----------------
NODE: 2
FILES: 8
DOMAINS: 2
PERCENTAGE MATCHING = 60.00
MATCH METHOD: 1 - (DESCARTING REPEAT INPUT SAMPLES)
ABSORVING LIST: [ 0 (223) , 1 (925)  ]
ERROR: 59.12
COUNTERS:
  DOMAIN- 0:  [  0  1  4  1  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  3
0  0  0  0  0  0  3  0  3  3  3  0  3  3  8  0  1  1  1  1  0  3  0  5  6  6  2  2  6  1
1  3  5  4  5  4  0  0
4  4  5  4  4  1  4  3  5  0  2  0  2  4  5  1  4  4  0  3  1  1  1  1  2  2  0  0  1  3
3  2  3  1  4  2  2  4  5  5  5  4  3  5  3  4  1  0  1  0  1  0  0  0  0  0  ] (sum =
223)
  DOMAIN- 1:  [  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  7
8  8  8  8  8  8  8  8  8  8  8  8  8  8  0  7  8  8  8  8  8  8  8  8  8  8  8  5  8
8  8  8  8  8  8  8  8
8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  6  8  8  8  8  8  8  8  8  8  8
8  8  8  8  8  8  8  7  8  8  8  8  7  8  8  8  7  8  8  8  8  8  8  8  7  8  ] (sum =
925)
INTERVAL CONFIANÇA:
   Domain- 0 -
      N = 7  -  Mitja = 4.142857  -  Mitja² = 17.163265 - sumcounters = 29  -  sumcounters2 = 155
      Variança (S²) = 4.979592  -  Desviacio (S) = 2.231500  -  Desviacio (S^) = 2.231500
      TStudent Inverse (Tn-1;1-alfa/2) = 4.800243  -  alfa = 0.003000  -  1-(alfa/2) = 0.998500
      m1 = 0.094199  -  m2 = 8.191515  -  Interval +/- [4.048658]
    Domain- 1 -
      N = 4  -  Mitja = 6.500000  -  Mitja² = 42.250000 - sumcounters = 26  -  sumcounters2 = 174
      Variança (S²) = 1.250000  -  Desviacio (S) = 1.118034  -  Desviacio (S^) = 1.118034
      TStudent Inverse (Tn-1;1-alfa/2) = 8.891456  -  alfa = 0.003000  -  1-(alfa/2) = 0.998500
      m1 = 1.529525  -  m2 = 11.470475  -  Interval +/- [4.970475]
  DOMAIN- 0:                [  ****            *    * *** *** **** * ************  ********* * ****** *******  ****************** * *     ]
- TALL = 0.09 - Samples = 74
  DOMAIN- 1:                [ *   ****************** ****** *   *  *   * *          **       * *    *     **              * * ***** ]
- TALL = 1.53 - Samples = 44
MAP:
  DOMAIN- 0:
    g1114g1121h0.75d0.3.ldom      [   ***                          * **** * ******  *****  ********* * ***                  ****  *  *        ]
- Samples 39
    g1170g1180h0.75d0.3.ldom      [  ** *              *     * *** ***     *   *  *  *          *    ** * *                              ]
- Samples 20
    g1174g1178h0.75d0.3.ldom      [   *                *     * *** ***       **  *                 * **                               ]
- Samples 15
    g1175g1178h0.75d0.3.ldom      [   *                *     * *** ***       *****  ****        *   ** **                              ]
- Samples 23
    g690g720h0.75d0.3.ldom        [                              *      *           *            *   **  *** * *  *********         ]
- Samples 20
    g696g950h0.75d0.3.ldom        [                              *      **  *  * *  ***** **  *              *  * ************         ]
- Samples 28
    g699g707h0.75d0.3.ldom        [                              *      ***     ****  ***** ***   * *          **  *   ****** *         ]
- Samples 28
    g715g727h0.75d0.3.ldom        [                              *     * ***  ********  ***** * *  *** ** *******   ************ ***  * *     ]
- Samples 50
  DOMAIN- 1:
    g1114g1121h0.75d0.3.ldom      [ ********************************** ************ ********************************************** ***************** ]
- Samples 115
    g1170g1180h0.75d0.3.ldom      [ **********************************  ********************************************************** ********* ]
- Samples 115
    g1174g1178h0.75d0.3.ldom      [ ********************************** ********************************************************** ********** * ]
- Samples 115
    g1175g1178h0.75d0.3.ldom      [ ********************************** ************ ************************ ************************************ ]
- Samples 115
    g690g720h0.75d0.3.ldom        [ ********************************** ************ ************************************************************ ]
- Samples 116
    g696g950h0.75d0.3.ldom        [ ********************************** ************************************************************************ ]
- Samples 117
    g699g707h0.75d0.3.ldom        [ ********************** ************ ************************************************************************ ]
- Samples 116
    g715g727h0.75d0.3.ldom        [ ********************************** ************************************************************** ************ ]
- Samples 116
ERRORS - FILES: 8 - SUM_ERROR: 946
   g1114g1121h0.75d0.3.ldom: 106
   g1170g1180h0.75d0.3.ldom: 127
   g1174g1178h0.75d0.3.ldom: 132
   g1175g1178h0.75d0.3.ldom: 122
   g690g720h0.75d0.3.ldom: 126
   g696g950h0.75d0.3.ldom: 119
   g699g707h0.75d0.3.ldom: 118
```

```
   g715g727h0.75d0.3.ldom: 96
 ---------------- N O D E  -  4 ----------------
NODE: 4
FILES: 5
DOMAINS: 2
PERCENTAGE MATCHING = 60.00
MATCH METHOD: 1 - (DESCARTING REPEAT INPUT SAMPLES)
ABSORVING LIST: [ 0 (134) , 1 (572)  ]
ERROR: 46.40
COUNTERS:
  DOMAIN- 0:  [  1  1  2  2  1  0  1  1  0  0  3  4  3  3  1  1  2  3  2  3  1  1  1  5
2  3  3  4  4  4  5  2  3  5  5  4  5  5  1  1  0  0  1  3  2  1  0  0  2  1  0  0  1  0
2  1  1  0  0  0  0  0
0  0  0  0  0  1  0  0  2  2  1  2  1  2  2  2  1  1  1  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  2  0  ] (sum =
134)
  DOMAIN- 1:  [  4  5  5  5  4  3  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  2
5  5  5  5  5  5  1  5  5  5  5  4  4  4  4  5  5  5  5  5  5  5  5  5  5  4  5  5  4  5
5  5  5  5  5  5  5  5
5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5
5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  5  4  5  5  5  5  5  5  ] (sum =
572)
INTERVAL CONFIANÇA:
    Domain- 0 -
       N = 5  -  Mitja = 3.000000  -  Mitja² = 9.000000  -  sumcounters = 15  -  sumcounters2 = 55
       Variança (S²) = 2.000000  -  Desviacio (S) = 1.414214  -  Desviacio (S^) = 1.414214
       TStudent Inverse (Tn-1;1-alfa/2) = 6.434848  -  alfa = 0.003000  -  1-(alfa/2) = 0.998500
       m1 = -1.069755  -  m2 = 7.069755  -  Interval +/- [4.069755]
    Domain- 1 -
       N = 5  -  Mitja = 3.000000  -  Mitja² = 9.000000  -  sumcounters = 15  -  sumcounters2 = 55
       Variança (S²) = 2.000000  -  Desviacio (S) = 1.414214  -  Desviacio (S^) = 1.414214
       TStudent Inverse (Tn-1;1-alfa/2) = 6.434848  -  alfa = 0.003000  -  1-(alfa/2) = 0.998500
       m1 = -1.069755  -  m2 = 7.069755  -  Interval +/- [4.069755]
  DOMAIN- 0:                [ ***** **  *************************** ****  **  * ***       * **********                    *      * ]
- TALL = -1.07 - Samples = 61
  DOMAIN- 1:                [     * **                        **   **  ** *  ********** **        ************************ ******* * ]
- TALL = -1.07 - Samples = 57
MAP:
  DOMAIN- 0:
    g1167g1177h0.75d0.3.ldom      [  ****     *   * **  *    **********     *  **  * ***        * *   ****                        *        ]
- Samples 33
    g1211g1401h0.75d0.3.ldom      [                   *  ** *  ** **    **  *                 ** *****                                  ]
- Samples 18
    g496g544h0.75d0.3.ldom        [         ****  *    ********  ***** *  ***       *            ***                                 *  ]
- Samples 27
    g673g761h0.75d0.3.ldom        [         ****  * *  * ****** ******    *                                                          ]
- Samples 20
    g696g766h0.75d0.3.ldom        [ * **  **  ***************************                                 **                          *  ]
- Samples 36
  DOMAIN- 1:
    g1167g1177h0.75d0.3.ldom      [ ****  ************************ ***** * ********** ** ************************************************************ ]
- Samples 111
    g1211g1401h0.75d0.3.ldom      [  ********************** ****** ************************************************************************************ ]
- Samples 115
    g496g544h0.75d0.3.ldom        [ ********************** ************************************************************************************** ]
- Samples 117
    g673g761h0.75d0.3.ldom        [ ***** ********************** ************************************************************************************** ]
- Samples 116
    g696g766h0.75d0.3.ldom        [ ********************** ****** **** * ********************************************************************************** ****** ]
- Samples 113
ERRORS - FILES: 5 - SUM_ERROR: 464
   g1167g1177h0.75d0.3.ldom: 84
   g1211g1401h0.75d0.3.ldom: 101
   g496g544h0.75d0.3.ldom: 94
   g673g761h0.75d0.3.ldom: 102
   g696g766h0.75d0.3.ldom: 83
```

8 ANEX-III FITXER DE CONFIGURACIÓ DEL MÒDUL PHP

Fitxer de configuració PHP en el directori arquetips/config/arq_config.php

```
<?php

define('LOG_DIRECTORY', '/home/bernat/uab/projecte/log/');
define('LOG_FILENAME', 'arquetips.log');
$_log_level = 'ALL';  // INFO, DEBUG, ALL

define('DIRECTORI_MICROARRAYS','./microarray/');
define('DIRECTORI_PREFIX_MICROARRAYS','m');
define('DIRECTORI_MICROARRAYS_POST','nonlineal/');
define('DIRECTORI_NAME_NONLINEAL','nonlineal/');
define('DIRECTORI_NAME_ARQUETIPS','arquetypes/');
define('DIRECTORI_COLORS','colors/');
define('DIRECTORI_IMAGES_SAMPLES','images/');
define('DIRECTORI_ARQUETIPS_CLASS','nonlineal/class/');
define('FILE_ARQUETIPS_CLASS','relbyGen.txt');

define('DIRECTORI_INCLUDES','arquetips/includes/');
define('DIRECTORI_CLASSES','arquetips/includes/classes/');
define('DIRECTORI_FUNCTIONS','arquetips/includes/funcions/');
define('DIRECTORI_TEMPLATES','arquetips/templates/');
define('DIRECTORI_CSS','arquetips/css/');
define('DIRECTORI_IMATGES','arquetips/images/');
define('DIRECTORI_CURVEIMATGES','curveimages/');

define('DRAW_SAMPLE_WIDTH', 10);
define('DRAW_SAMPLE_HEIGHT', 19);
define('NODE_EXTRA_WIDTH', 200);

define('CORRELATION_ID_1', 1);
define('CORRELATION_ID_1_LABEL', ' Non-lineal relationships with high correlation ');
define('CORRELATION_ID_1_DIRECTORY', 'normal/');
define('CORRELATION_ID_1_DIRECTORY_CLASS', 'class/');

define('CORRELATION_ID_2', 2);
define('CORRELATION_ID_2_LABEL', ' Non-lineal relationships with medium correlation ');
define('CORRELATION_ID_2_DIRECTORY', 'HeighF/');
define('CORRELATION_ID_2_DIRECTORY_CLASS', 'classHeighF/');

define('CORRELATION_ID_3', 3);
define('CORRELATION_ID_3_LABEL', ' Non-lineal relationships with medium correlation and heigh
curvature ');
define('CORRELATION_ID_3_DIRECTORY', 'HeighFfiltered/');
define('CORRELATION_ID_3_DIRECTORY_CLASS', 'classHeighF/');

define('CORRELATION_ID_4', 4);
define('CORRELATION_ID_4_LABEL', ' Non-lineal relationships with low correlation and heigh
curvature ');
define('CORRELATION_ID_4_DIRECTORY', 'RH_F10filtered/');
define('CORRELATION_ID_4_DIRECTORY_CLASS', 'classRH_F10/');

define('CORRELATION_ID_DEFAULT', CORRELATION_ID_1);

$_arr_correlations = array();
$_arr_correlations[CORRELATION_ID_1] = array('id' => CORRELATION_ID_1, 'label' =>
CORRELATION_ID_1_LABEL, 'directory' => CORRELATION_ID_1_DIRECTORY, 'directory_class' =>
CORRELATION_ID_1_DIRECTORY_CLASS);
$_arr_correlations[CORRELATION_ID_2] = array('id' => CORRELATION_ID_2, 'label' =>
CORRELATION_ID_2_LABEL, 'directory' => CORRELATION_ID_2_DIRECTORY, 'directory_class' =>
CORRELATION_ID_2_DIRECTORY_CLASS);
$_arr_correlations[CORRELATION_ID_3] = array('id' => CORRELATION_ID_3, 'label' =>
CORRELATION_ID_3_LABEL, 'directory' => CORRELATION_ID_3_DIRECTORY, 'directory_class' =>
CORRELATION_ID_3_DIRECTORY_CLASS);
$_arr_correlations[CORRELATION_ID_4] = array('id' => CORRELATION_ID_4, 'label' =>
CORRELATION_ID_4_LABEL, 'directory' => CORRELATION_ID_4_DIRECTORY, 'directory_class' =>
CORRELATION_ID_4_DIRECTORY_CLASS);

$_arr_config_microarrays = array();
$_arr_config_microarrays['matching'] = array(60 => '60', 65 => '65', 75 => '75', 85 => '85', 90 =>
```

```
'90', 95 => '95');
$_arr_config_microarrays['alfa'] = array('0.003' => '0.003', '0.009' => '0.009', '0.050' => '0.05',
'0.100' => '0.10', '0.150' => '0.15', '0.200' => '0.20', '0.400' => '0.400');
$_arr_config_microarrays['method'] = array(1 => ' Descarting joint samples ', 2 => ' Assingning
joint samples to the smaller cluster ');
$_arr_config_microarrays['classification'] = array('var' => 't-student', 'max' => 'maximum');
$_arr_config_microarrays['correlation_analisis_id'] = array();
foreach ($_arr_correlations as $key => $valors) {
       $_arr_config_microarrays['correlation_analisis_id'][$valors['id']] = $valors['label'];
}


define('ORDER_MATCHINGS', 1);
define('ORDER_ERROR', 2);

define('DEFAULT_USER', 313);

define('GUI_HOST_PATH', 'http://revresearch.phpwebhosting.com/wwwprueba/bernat/gexp/microarray/');
define('GUI_PAGE','InterfazFlash.phtml');
define('GUI_FACTORS_PAGE',
'http://revolutionresearch.uab.es/applic/gexp/microarray/factors.phtml');
define('GUI_NCBI_PAGE', 'http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=geo');

?>
```

9 ANEX-IV MODIFICACIONS AL PROGRAMA LANZADORA

// BERNAT GISPERT - INICI CODIFICACIO DE pcop_clustering PER AL lanzadora

```
        pcop_clustering:   // BERNAT GISPERT 6/6/2010 - ETIQUETA DE GOTO PER SALTAR EL PRE-PROCES
AL pcop_clustering
                           // m17/nonlineal/normal : 147 fitxers "ldom" -> 147*1.5=220,5 == 221
fitxers ldom de threshold per processar fitxers de baixa correlacio.

        int _pcop_clustering_th = 221;  // CONSTANTE PARA SABER EL LIMITE DE FICHEROS "ldom" DE
DIRECTORIO "normal" A PROCESAR POR LA MICROARRAY.
        int _ldom_files = 0;            // PER DEFECTE, SEMPRE ANALITZEM TOTS ELS DIRECTORIS.

        sprintf(aux,"find ../microarray/m%s/nonlineal/normal/ -type f -name \"*.ldom\" | wc -l > m
%s_th.tmp", argv[1], argv[1]);  // FICHEROS "ldom" DE "normal".
        system(aux);
        sprintf(aux,"m%s_th.tmp", argv[1]);
        _ldom_files = llegeix_valor_fitxer(aux);
        remove(aux);

        if (_ldom_files <= _pcop_clustering_th) printf("Creacio de TOTA la estructura de directoris
per a la sortida de pcop_clustering (normal, HeighF, HeighFfiltered, RH_F10filtered)\n");
        else printf("Creació estructura de directoris per a la sortida de pcop_clusterin (normal i
HeighF)\n");
        sprintf(aux,"if [ ! -d ./log ] ; then mkdir ./log; fi");
        system(aux);
        // CREAR DIRECTORI ARQUETYPES
        sprintf(aux,"if [ ! -d ../microarray/m%s/nonlineal/arquetypes ] ; then mkdir ../microarray/
m%s/nonlineal/arquetypes; fi", argv[1], argv[1]);
        system(aux);
        // CREAR DIRECTORI IMATGES (per guardar les imatges dels noms de les samples en format gif
i vertical)
        sprintf(aux,"if [ ! -d ../microarray/m%s/nonlineal/arquetypes/images ] ; then mkdir
../microarray/m%s/nonlineal/arquetypes/images; fi", argv[1], argv[1]);
        system(aux);
        sprintf(aux,"rm -f ../microarray/m%s/nonlineal/arquetypes/images/*", argv[1]);
        system(aux);
        // CREAR DIRECTORIS DE SORTIDA
        // DIRECTORI "normal"
        sprintf(aux,"if [ -d ../microarray/m%s/nonlineal/arquetypes/normal ] ; then rm -fR
../microarray/m%s/nonlineal/arquetypes/normal; fi", argv[1], argv[1]);
        system(aux);
        sprintf(aux,"mkdir ../microarray/m%s/nonlineal/arquetypes/normal", argv[1], argv[1]);
        system(aux);
        sprintf(aux,"mkdir ../microarray/m%s/nonlineal/arquetypes/normal/colors", argv[1],
argv[1]);
        system(aux);
        // DIRECTORI "HeighFfiltered" Alta correlacio filtrats
        sprintf(aux,"if [ -d ../microarray/m%s/nonlineal/arquetypes/HeighFfiltered ] ; then rm
-fR ../microarray/m%s/nonlineal/arquetypes/HeighFfiltered; fi", argv[1], argv[1]);
        system(aux);
        sprintf(aux,"mkdir ../microarray/m%s/nonlineal/arquetypes/HeighFfiltered", argv[1]);
        system(aux);
        sprintf(aux,"mkdir ../microarray/m%s/nonlineal/arquetypes/HeighFfiltered/colors", argv[1],
argv[1]);
        system(aux);
        if (_ldom_files <= _pcop_clustering_th) {  // SI NO ES SUPERA EL treshold TAMBE ANALITZEM
ELS DIRECTORIS DE CORRELACIO INTERMITJA
          // DIRECTORI "HeighFNOfiltered" Alta correlacio sense filtrar
          sprintf(aux,"if [ -d ../microarray/m%s/nonlineal/arquetypes/HeighF ] ; then rm -fR
../microarray/m%s/nonlineal/arquetypes/HeighF; fi", argv[1], argv[1]);

          system(aux);
          sprintf(aux,"mkdir ../microarray/m%s/nonlineal/arquetypes/HeighF", argv[1]);
          system(aux);
          sprintf(aux,"mkdir ../microarray/m%s/nonlineal/arquetypes/HeighF/colors", argv[1],
argv[1]);
          system(aux);

          // DIRECTORI "RH_F10filtered" Alta correlacio sense filtrar
          sprintf(aux,"if [ -d ../microarray/m%s/nonlineal/arquetypes/RH_F10filtered ] ; then rm
-fR ../microarray/m%s/nonlineal/arquetypes/RH_F10filtered; fi", argv[1], argv[1]);
          system(aux);
```

```
          sprintf(aux,"mkdir ../microarray/m%s/nonlineal/arquetypes/RH_F10filtered", argv[1]);
          system(aux);
          sprintf(aux,"mkdir ../microarray/m%s/nonlineal/arquetypes/RH_F10filtered/colors",
argv[1], argv[1]);
          system(aux);
        }
        printf("Fi de la creaciód'estructura de directoris per a la sortida de pcop_clustering\n");

        // EXECUTAR EL CLUSTERING PER AL DIRECTORI "normal"
        sprintf(aux, "%s %s", "./pcop_clustering", argv[1]);
        printf("%s\n",aux);
        system(aux);

        // EXECUTAR EL CLUSTERING PER AL DIRECTORI "HeighFfiltered"
        sprintf(aux, "%s %s %s", "./pcop_clustering", argv[1], "HeighFfiltered");   // Alta
correlacióFiltrat
        printf("%s\n",aux);
        // #### OJO !!!  BERNAT : (descomentar cuando tengamos terminado el directorio de alta
correlacion filtrado)
        // system(aux);    //   <---- ######
        // #### FIN OJO !!!!  BERNAT

        if (_ldom_files <= _pcop_clustering_th) {  // SI NO ES SUPERA EL treshold TAMBE ANALITZEM
ELS DIRECTORIS DE CORRELACIO INTERMITJA
          // EXECUTAR EL CLUSTERING PER AL DIRECTORI "HeighF"
          sprintf(aux, "%s %s %s", "./pcop_clustering", argv[1], "HeighF");    // Alta correlació
sense filtrar
          printf("%s\n",aux);
          system(aux);
          // EXECUTAR EL CLUSTERING PER AL DIRECTORI "HeighF10filtered"
          sprintf(aux, "%s %s %s", "./pcop_clustering", argv[1], "RH_F10filtered");    //
Correlacióintermitja
          printf("%s\n",aux);
          system(aux);
        }
        // BERNAT GISPERT - FINAL DEL PROCES DE pcop_clustering
```

10 BIBLIOGRAFIA

[1] Delicado, P. and Huerta, M. (2003): 'Principal Curves of Oriented Points: Theoretical and computational improvements'. Computational Statistics 18, 293-315.

[2] Cedano, J., Huerta, M., Estrada, I., Ballllosera, F., Conchillo, O., Delicado, P., and Querol, E. (2007). A web server for automatic analysis and extraction of relevant biological knowledge. Comput. Biol. Med. 37, 11 (Nov. 2007), 1672-1675.

[3] Huerta M, Cedano J, Querol E. (2008) Analysis of nonlinear relations between expression profiles by the principal curves of oriented-points approach. *J Bioinform Comput Biol*. 6:367-386.

[4] Cedano J, Huerta M, Querol E. (2008) NCR-PCOPGene: An Exploratory Tool for Analysis of Sample-Classes Effect on Gene-Expression Relationships *Advances in Bioinformatics*, vol. 2008

[5] Huerta M, Cedano J, Peña D, Rodriguez A, Querol E. (2009) PCOPGene-Net: holistic characterisation of cellular states from microarray data based on continuous and non-continuous analysis of gene-expression relationships. *BMC Bioinformatics.*, 9;10:138

[6] http://revolutionresearch.uab.es : Web server for on line microarray analysis supported by the Institute of Biotechnology and Biomedicine of the Autonomous University of Barcelona (IBB-UAB).

Hi ha diversos mètodes d'anàlisi que duen a terme una agrupació global de la sèries de mostres de microarrays, com Self-Organizing Maps, o que realitzen agrupaments locals tenint en compte només un subconjunt de gens co-expressats, com Biclustering, entre d'altres. En aquest projecte s'ha desenvolupat una aplicació web: El PCOPSample-cl, és una eina que pertany als mètodes d'agrupació (clustering) local, que no busca subconjunts de gens co-expresats (anàlisi de relacions linials), si no parelles de gens que davant canvis fenotípics, la seva relació d'expressió pateix fluctuacions. El resultats del PCOPSample-cl seràn les diferents distribucions finals de clusters i les parelles de gens involucrades en aquests canvis fenotípics. Aquestes parelles de gens podràn ser estudiades per trobar la causa i efecte del canvi fenotípic. A més, l'eina facilita l'estudi de les dependencies entre les diferents distribucions de clusters que proporciona l'aplicació per poder estudiar la intersecció entre clusters o l'aparició de subclusters (2 clusters d'una mateixa agrupació de clusters poden ser subclusters d'altres clusters de diferents distribucions de clusters) . L'eina es disponible al servidor: http://revolutionresearch.uab.es/

Hay varios métodes d'análisis que llevan a una agrupació global de la serie de muestras de microarrays, como Self-Organizing Maps, o que realizan agrupaciones locales considerando sólo un subconjunto de genes coexpressados, como Biclustering, entre otros. En este proyecto se ha desarrollado una aplicación web: El PCOPSample-cl, és una herramienta que pertenece a los métodos de agrupación (clustering) local, que no busca subconjuntos de genes coexpresados (anàlisis de relaciones lineales), sino pares de genes que ante cambios fenotípicos su relación de expresión sufre fluctuaciones. El resultado de PCOPSample-cl seran la distintas distribuciones finales de clusters y los pares de genes involucrados en estos cambios fenotípicos. Estos pares de genes podran ser estudiados para encontrar la causa i efecto del cambio fenotípico. Además, la herramienta facilita el estudio de las dependencias entre las diferentes distribuciones de clusters que proporciona la aplicación para poder estudiar la intersección entre clusters o la aparición de subclusters (2 clusters de una misma agrupación de clusters pueden ser subclusters de diferentes distribuciones de clusters). La herramienta está disponible en el servidor: http://revolutionresearch.uab.es/

There are several analytical methods that perform a global clustering of the microarray sample series, such as Self-Organizing Maps, or which perform local clusterings considering only a subset of co-expressed genes, such as Biclustering, and so on. The PCOPSample-cl belongs to the local-clustering methods, but not considering subsets of co-expressed genes if not considering only pairs of genes whose expression dependence suffers a fluctuation due to a phenotype change. The phenotypes involved in the phenotype change will constitute de final sample clusters. The pairs of genes whose expression

dependence suffers the phenotype change are provided together with the sample-clusters arrangement they belongs. These pairs of genes can be used to study the cause or the effect of the phenotype changes. The different sample-culster arrangements can be compared to determine their dependence: cluster-intersection or subclusters (two clusters of one cluster arrangement can be subclusters of other cluster of a different cluster arrangement). The webtool is available at the server: http://revolutionresearch.uab.es/