Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

Vers une analyse implicative des données issues de puces à ADN

Gérard Ramstein

LINA, équipe COD
Ecole polytechnique de l'université de Nantes
Rue Christian Pauc BP 50609 44306 Nantes cedex 3
gerard.ramstein@univ-nantes.fr

Résumé. Les puces à ADN permettent l'analyse simultanée de l'expression de plusieurs milliers de gènes dans des conditions expérimentales données. Ce papier introduit l'utilisation des concepts de l'Analyse Statistique Implicative pour ce type novateur de données biologiques. L'application retenue concerne la différenciation de deux formes de leucémie aiguë, les leucémies lymphoblastiques et les leucémies myéloblastiques. Le jeu de données est constitué de 7129 gènes et de 38 patients. Nous montrons à travers le logiciel CHIC que l'analyse implicative met en valeur les deux groupes de patients et qu'elle permet d'extraire les gènes qui différencient le mieux ces derniers. Le concept de contribution conduit à une réduction sensible des gènes marqueurs d'une maladie et pourrait déboucher sur des protocoles de diagnostic efficaces.

1 Présentation du contexte biologique

Les puces à ADN représentent un outil d'analyse de l'expression de gènes dans une cellule. Cette technologie récente doit son émergence à l'essor des technologies des microsystèmes et de la microfluidique ces trente dernières années, et en particulier à la possibilité de greffer des molécules spécifiques sur une surface extrêmement petite, avec une précision micrométrique. Les puces à ADN permettent d'accélérer grandement les expériences biologiques, en assurant un traitement automatique, miniaturisé et parallèle des molécules d'intérêt. Des procédés fastidieux qui exigeaient plusieurs mois de travail se font désormais en routine en quelques heures seulement. Les puces à ADN ont donc révolutionné les pratiques des laboratoires et ouvert de nouveaux champs d'études scientifiques et industrielles. Les domaines d'application vont de la génomique fonctionnelle (étude des gènes et de leurs fonctions) à la recherche pharmaceutique, en passant par le contrôle des produits de l'industrie agro-alimentaire. L'analyse clinique est un champ d'application majeur pour assurer la compréhension et le dépistage de pathologies génétiques.

Les puces à ADN renferment sur des surfaces de quelques cm^2 plusieurs milliers de molécules d'ADN, disposées sur des emplacements précis appelés spots. Ces points disposés en matrice représentent des sondes : un ensemble de gènes que le biologiste désire étudier dans le cadre d'une expérience donnée. Certaines puces, dites pangénomiques, permettent de placer l'ensemble du génome humain (du moins les gènes actuellement connus), sur une seule puce. Les sondes ont la propriété remarquable de pouvoir s'apparier avec des brins complémentaires d'ADN dans une solution biologique. Le procédé consiste donc à apparier des extraits moléculaires inconnus d'une cellule, reflet de son activité à un instant précis (prise de médicaments, stade de maladie, …), avec les sondes dont on connaît l'identité. Rappelons brièvement en effet que le noyau d'une cellule contient l'ensemble des gènes (le **génome**) sous sa forme inactive, et que certains d'entre eux seulement sont appelés à *s'exprimer* (le **transcriptome** : ensemble des gènes actifs) en fonction des besoins de la cellule et de son activité métabolique. Les puces à ADN sont donc une technologie du transcriptome (gènes effectivement transcrits).
L'analyse des puces à ADN conduit au final à la définition d'un niveau d'expression pour chaque gène présent sur une sonde. Ce niveau est sujet à de nombreuses variations dues à la complexité du dispositif expérimental. De nombreux travaux statistiques et des protocoles adéquats ont été mis en place pour assurer une certaine pertinence à cette quantification.

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

2 Spécificité des données de puces à ADN et concept d'implication

Les données des puces à ADN regroupent jusqu'à 30 000 gènes mais ne portent que sur un nombre limité d'expériences en raison du coût élevé du procédé (nombre de l'ordre de la centaine tout au plus). Cette dissymétrie pose un véritable problème pour l'extraction de règles d'association si l'on souhaite définir des associations entre gènes. Nous proposons dans ce papier une méthodologie permettant de réduire le nombre de gènes étudiés pour pallier cette difficulté. Notons cependant que l'analyse des expérimentations est également intéressante pour le biologiste, pour déceler par exemple si l'expression des gènes dans un tissu implique un comportement similaire dans un autre.

Le choix du type de données traitées, numérique, modal ou intervalle, est une question déterminante qui ne sera pas abordée dans ce papier. Par souci de concision, nous nous sommes volontairement limités à une approche simplificatrice, à savoir la binarisation. Différentes méthodes, basées sur des mesures multiples, permettent d'associer à un gène un niveau d'expression qualitatif. Le gène est soit sous-exprimé, normal ou sur-exprimé. Notre étude se bornera au cas binaire de la sur-expression des gènes, bien qu'elle puisse s'étendre à plusieurs modalités. Les biologistes disposent d'outils statistiques ad hoc pour différencier une sur-expression d'un écart non significatif d'un point de vue statistique. Le logiciel SAM [Tusher et al. 2001] permet par exemple une telle analyse. Ce papier ne traitera donc pas cette question et considérera le problème de la binarisation comme résolu.

Le biologiste exploite ces matrices de données par des outils d'analyse classique, la méthode la plus généralement rencontrée dans les publications étant la classification hiérarchique ascendante (CHA). Les gènes sont regroupés en clusters, ce qui permet d'inférer à des gènes non annotés une fonction analogue à celle de ces voisins. Cette analyse repose sur l'hypothèse suivante : des gènes ayant le même profil d'expression (on dit qu'ils sont co-régulés) codent pour la même fonction. Ce principe n'est évidemment qu'une hypothèse de travail pour le biologiste et non une certitude absolue.

Un reproche que l'on peut faire à la classification non supervisée est qu'un gène ne se retrouve que dans un seul cluster, alors qu'il peut potentiellement être présent dans deux processus biologiques distincts (gène apparaissant dans plusieurs voies métaboliques par exemple). De plus, la découverte de réseaux géniques est délicate, dans la mesure où des gènes co-régulés peuvent ne pas être adjacents, ce qui rend difficile leur identification. Ces critiques [Liping et al. 2004] plaident pour une approche alternative, basée sur l'extraction de règles d'association (pour des articles traitant de règles d'association sur les données de puces à ADN, voir notamment [Oyama et al. 2002], [Cong et al. 2004]). Ce paradigme possède l'avantage d'exprimer directement l'implication de certains gènes sur d'autres. Une règle comme « si g1 et g3 sont sur-exprimés, alors g2 est sur-exprimé » est en effet une information pertinente pour le biologiste. Contrairement à la CHA, ces règles ne s'appliquent pas à l'ensemble des expériences, mais à un sous-ensemble de celles-ci. Un des attraits en effet de cette approche est de faire émerger des classes multiples d'expériences, ce qui fournit une vue plus fine (mais aussi plus complexe) des phénomènes incriminés. On pourra ainsi découvrir des règles qui ne sont vérifiées que pour un groupe précis, comme par exemple le groupe des patients victimes d'un antécédent cardiaque.

Un enjeu majeur des puces à ADN concerne la médecine prédictive, dont l'objectif ne se résumera plus uniquement à guérir des patients, mais à diagnostiquer des risques futurs de maladies. Il est envisageable de pratiquer des tests cliniques sur des sujets à risques, à condition de limiter pour des raisons économiques le nombre de gènes. Il est donc important de trouver les gènes les plus pertinents pour discriminer telle ou telle classe de maladie. L'apport de l'analyse statistique implicative [Gras 1996] est ici manifeste, puisqu'on peut définir les gènes qui contribuent le mieux à une classe donnée d'expériences. Notons par ailleurs que l'intensité d'implication diffère fortement des mesures de dépendance statistique classiquement utilisées en analyse de puces à ADN, telles que le coefficient de corrélation. L'intensité d'implication est en effet orientée (c'est une mesure non symétrique). Elle exprime la qualité d'une règle d'association en prenant en considération les ensembles en présence : on montre notamment que l'intensité d'implication et le coefficient

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

de corrélation peuvent évoluer de façon opposée selon la taille respective des populations concernées (voir [Gras 1996] p. 40 pour une comparaison des deux mesures).

3 Application à la différenciation de deux formes de leucémie

Nous allons considérer l'intérêt du logiciel d'analyse de données CHIC (Classification Hiérarchique Implicative et Cohésitive [Couturier 2001]) pour l'identification de groupes de patients atteints de certaines formes de leucémie et des gènes impliqués dans ces pathologies. La leucémie se caractérise par une prolifération maligne de cellules d'origines hématopoïétiques peu matures et rapidement diffusantes. Différents lymphomes malins se développent, aboutissant à une atteinte massive de la moelle osseuse. On distingue les leucémies aiguës lymphoblastiques (notées ALL par la suite pour *Acute Lymphoblasic Leukemia*) des leucémies aiguës myéloblastiques (notées AML pour *Acute Myeloid Leukemia*). La distinction entre ces deux formes est essentielle pour le succès des thérapies envisagées : le traitement diffère selon l'une ou l'autre de ces deux classes de leucémie. Nos données proviennent d'une expérimentation [Golub et al. 1999] visant à montrer que l'analyse des puces à ADN peut conduire à une identification de ces deux formes ainsi qu'à une prédiction du type de leucémie dont souffre le patient. Les données sont accessibles publiquement sur le site du centre de génomique du MIT, le Whitehead Institute Center for Genome Research [WICGH]. Le jeu de données correspond à une analyse d'échantillons de moelle osseuse sur 38 patients dont 27 ALL et 11 AML. Des sondes Affymetrix contenant 7129 gènes humains ont été utilisées à cette fin. Comme le suggèrent les auteurs, nous avons procédé à une normalisation des valeurs. Pour chaque patient, nous avons centré et normé les valeurs d'expression. Nous avons ensuite effectué une discrétisation en valeurs binaires par simple seuillage. Les valeurs d'expression supérieures à la valeur 0.5 ont été considérées comme sur-exprimées. Ce seuil a été défini par analyse visuelle des données en fausses couleurs.
Le tableau 1 présente un extrait de notre jeu de données. Les quatre premières lignes (catégorie A) indiquent des gènes dont la sur-expression n'est pas liée à un groupe de patients particulier. Les lignes suivantes (catégories B et C) indiquent les gènes que nous avons pu extraire grâce au logiciel CHIC et dont nous avons pu vérifier par la littérature qu'ils jouent effectivement un rôle prépondérant dans les deux formes de leucémies étudiées (catégorie B (resp. C) : prépondérance de sur-expression pour le groupe ALL (resp. AML)).

	ALL																											AML									
A																																					
0039	**1**	0	**1**	**1**	0	**1**	**1**	**1**	**1**	0	0	**1**	0	0	**1**	0	0	0	0	0	0	**1**	**1**	0	**1**	0	**1**	0	**1**	0	1	0	**1**	0	0	0	0	**1**	**1**	0
0041	**1**	0	**1**	**1**	0	**1**	**1**	0	**1**	0	0	**1**	0	**1**	**1**	0	0	0	0	0	0	**1**	**1**	0	**1**	0	**1**	0	0	0	0	**1**	**1**	0	0	**1**	0	0
0052	**1**	**1**	**1**	**1**	0	**1**	**1**	**1**	**1**	0	**1**	**1**	0	0	**1**	0	0	**1**	0	0	0	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	0	**1**	**1**	0
0067	0	**1**	**1**	**1**	0	0	0	**1**	**1**	0	**1**	0	**1**	0	0	0	**1**	0	0	**1**	0	**1**	0	**1**	0	0	1	1	0	0	**1**	0	0	**1**	0	**1**	0	**1**
B																																						
2642	**1**	0	0	**1**	**1**	0	**1**	**1**	0	0	0	**1**	**1**	0	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	0	**1**	**1**	**1**	**1**	0	0	0	0	0	0	0	0	0	0	0
5772	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	0	**1**	0	**1**	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	0	0	0	0	0	0	0	0	0	0
2354	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**	0	0	0	0	0	0	0	0	0	0	0
C																																						
4373	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	0	0	0	**1**	**1**	0	0	0	**1**	**1**	**1**	**1**
5976	0	0	0	0	**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	**1**	**1**	0	0	0	**1**	**1**	**1**	**1**
1882	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	0	0	0	**1**	**1**	0	0	**1**	**1**	**1**	**1**	**1**
4847	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	**1**	**1**	**1**	0	**1**	**1**	**1**	0	**1**	**1**
6200	0	**1**	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	**1**	0	**1**	**1**	**1**	**1**	0	**1**	0	**1**	**1**
6201	0	**1**	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
1249	0	**1**	0	0	1	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

TAB 1 – *Valeurs binaires des gènes mentionnés par la suite dans le texte et qui ont été découverts par le logiciel CHIC. Les 27 premières colonnes sont relatives à des patients ALL ; les 11 restantes à des patients AML.*

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

Nous avons utilisé deux filtres pour définir l'ensemble des gènes d'intérêt. Le premier filtre consiste à ne retenir que les gènes ayant une sur-expression notable. Nous avons retenu que les gènes présentant au moins 5 valeurs à 1 sur 38 et au plus 33 valeurs à 1 sur 38. Ce filtre consiste à éliminer les gènes « inactifs » chez l'ensemble des patients ainsi ceux qui sont « sur-actifs » dans la mesure où ils ne différencient pas suffisamment des groupes potentiels de patients. Nous obtenons 416 gènes sur lesquels nous appliquons une analyse par le logiciel CHIC. L'arbre des similarités [Lerman 1981] permet d'identifier trois groupes de patients (fig. 1). Les patients AML sont bien regroupés ensemble tandis que les patients ALL sont séparés en deux parties. Il semblerait que ce deuxième ensemble de patients soit moins homogène que le premier, sans qu'on puisse trancher véritablement sur la cause de cette différence (patients au profil plus disparate ou mise en évidence de deux sous-classes de pathologie : la première des deux hypothèses étant probablement la plus plausible, étant donné le faible nombre de patients étudiés).

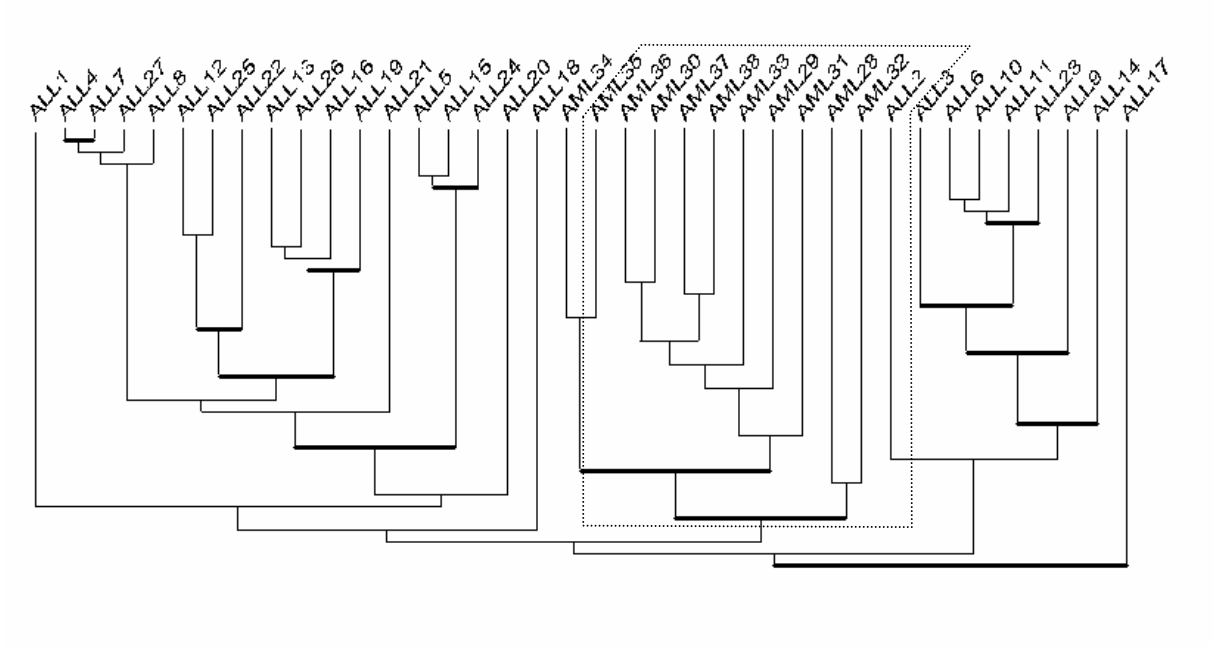

FIG. 1 - *Arbre des similarités sur le jeu de 416 gènes. Le groupe AML apparaît nettement comme un nœud significatif de l'arbre.*

Le premier filtre ayant permis de dégager le groupe AML, nous pratiquons une deuxième sélection des gènes en ne considérant que ceux qui marquent une sur-expression plus importante dans un groupe que dans l'autre. Nous avons ainsi retenu les gènes ayant au moins une sur-expression auprès de la moitié d'un des groupes et au plus trois patients du deuxième groupe présentant une sur-expression. Nous obtenons ainsi 93 gènes au total. Le nouvel arbre des similarités est présenté sur la figure 2.

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

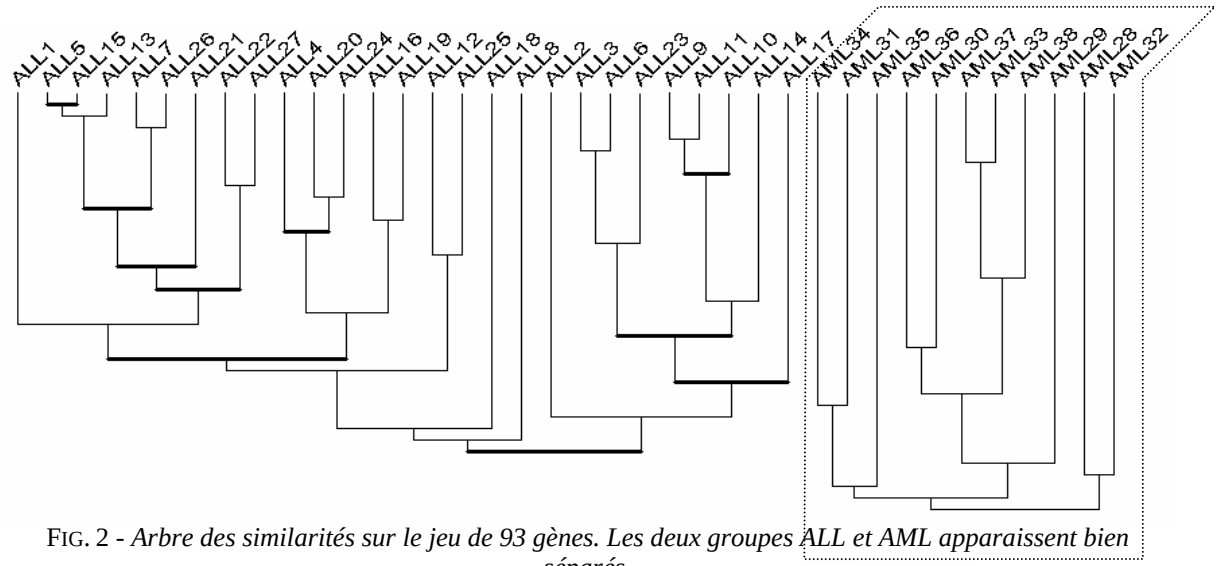

FIG. 2 - *Arbre des similarités sur le jeu de 93 gènes. Les deux groupes ALL et AML apparaissent bien séparés.*

L'arbre cohésitif (fig. 3) permet de rechercher des groupes optimaux au sein de classes d'observations. Nous retrouvons par cette analyse deux arbres (A1 et A2) parmi le groupe ALL. A chaque classe peut être associé un individu optimal théorique (cf. conférence de R.Gras, Actes ASI05) qui satisferait les maxima des intensités d'implication pour tous les couples sous-tendant la classe. La responsabilité, appelée contribution, d'un individu relativement à la classe est une mesure qui exprime le degré de rapprochement entre cet individu et l'individu optimal. En cherchant l'individu le plus contributif dans chacune des classes A1 et A2, nous avons obtenu deux gènes qui sont explicitement mentionnés dans [Golub et al. 1999] comme étant discriminants sur les deux types de leucémie.

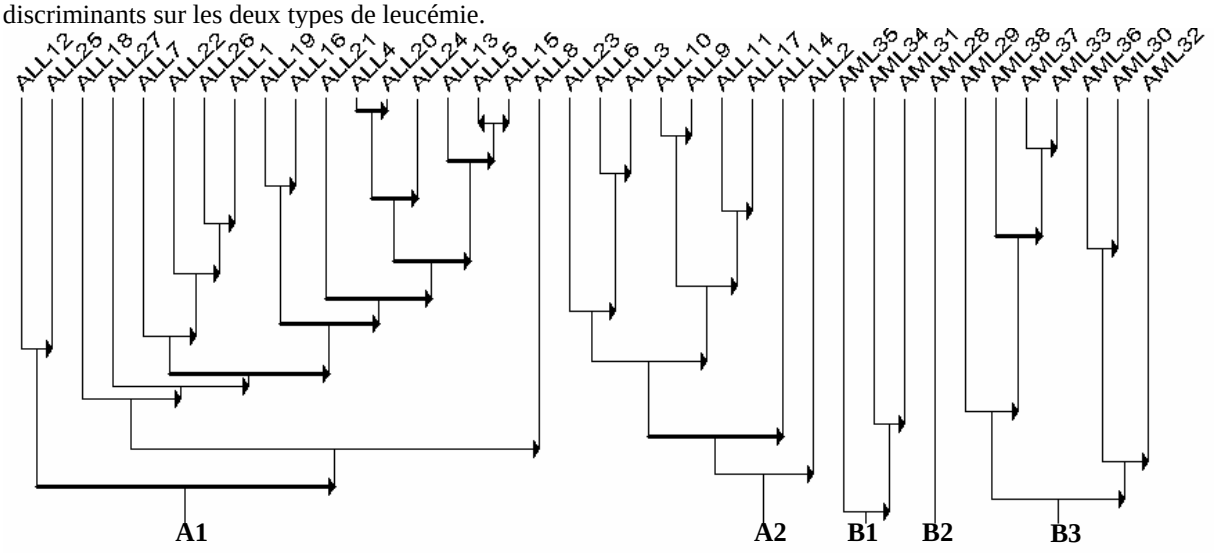

FIG. 3 - *Arbre cohésitif sur le jeu de 93 gènes.*

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

Le gène *MB-1*(#2642) trouvé par notre analyse de l'arbre A1 code pour un anticorps monoclonal qui identifie l'antigène *CD37*[1]. Il a été démontré que ce gène favorise la distinction entre les formes AML et ALL. Le gène *Cyclin D3* (#2354) , obtenu sur l'arbre A2, est cité par Golub comme permettant lui aussi cette classification et d'être prometteur pour d'autres applications, comme la parthénogenèse et la pharmacologie. Nous avons relevé un autre gène cité dans le même article. Il s'agit du gène *c-MyB* (#5772), un oncogène connu dont on sait qu'il favorise le développement de la leucémie de type AML [Slamon et al. 1986].
Le groupe AML est, quant à lui, éclaté en deux classes et un singleton (arbres B1, B2 et B3). Dans le groupe optimal de l'arbre B3, nous retrouvons notamment le gène *zyxin* (#4847), explicitement cité par les auteurs. Ce gène est également cité par Alex Smith comme un des plus discriminants [Smith 2000] sur le même jeu de données. Il est remarquable de découvrir comme gène le plus contributif le gène *MCL1* (#1249). Ce dernier code en effet pour une protéine au nom explicite : *induced myeloid leukemia cell differentiation protein*.
Nous avons inversé la matrice pour établir le graphe implicatif relatif aux gènes. La figure 4 met en évidence deux ensembles de gènes pour un seuil d'intensité d'implication de 0,99. Les gènes et les groupes contributifs recoupent dans une grande part les résultats obtenus précédemment. Les deux nœuds terminaux correspondent à des gènes déjà mentionnés et particulièrement spécifiques de chacune des deux classes. Les chemins du graphe expriment les implications entre gènes. Le chemin `4373 – 1882 – 6201 – 1249` de seuil d'intensité 0,98 dénote par exemple un lien fort de sur-expressions. Les valeurs du tableau 1 montrent que toutes les sur-expressions du gène de départ se vérifient sur les mêmes patients pour tous les gènes appartenant au chemin. Il en est de même pour le gène suivant (`1882`). Plus on avance dans le chemin, plus on étend évidemment le domaine des patients concernés par ce phénomène de sur-expression. Les nœuds terminaux du graphe sont donc caractérisés par une couverture plus grande des patients.

1 Rappelons brièvement qu'un anticorps est une molécule naturelle qui reconnaît spécifiquement une autre molécule, l'antigène. Notre système immunitaire produit des anticorps dirigés contre les bactéries et les virus. Certaines cellules tumorales portent également des antigènes.

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

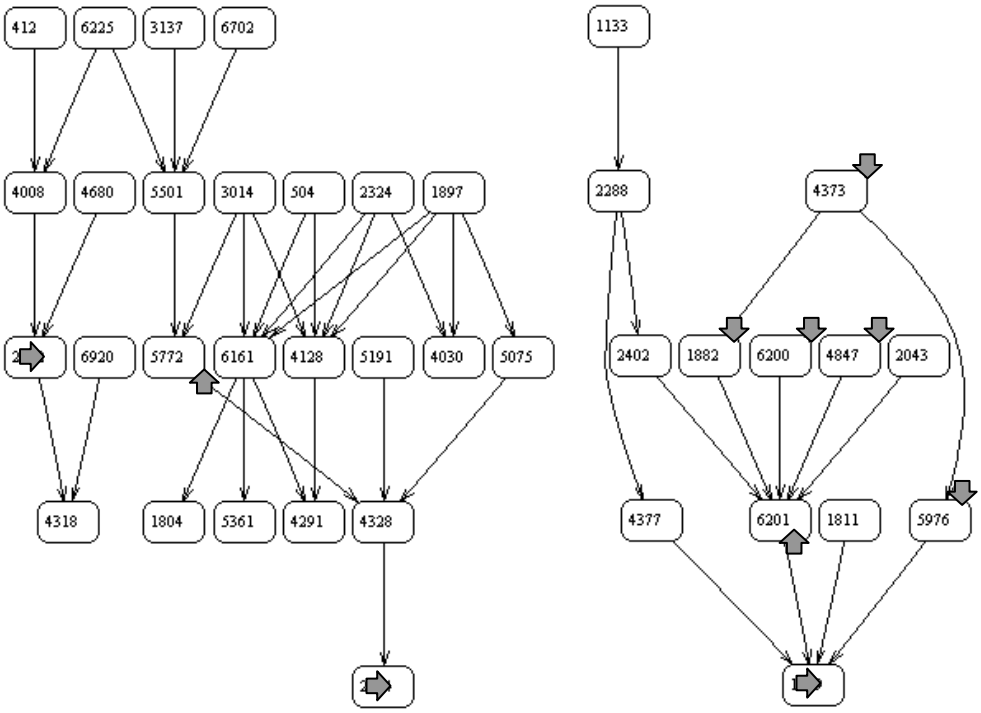

Fig. 4 - *Graphe implicatif relatif aux gènes. Mise en évidence de gènes spécifiques au groupe ALL (graphe de gauche) et ceux spécifiques du groupe AML (graphe de droite). Les flèches en grisé indiquent les gènes mentionnés dans le texte.*

Pour illustrer la cohésion entre les deux classes, nous avons représenté les dix gènes mentionnés sur la figure 5. Cette dernière met en valeur les différents niveaux de cohésion, ce qui permet de considérer des classes de profils de gènes. Les sondes #6200 et #6201 (resp. *IL8* et *précurseur de l'IL8*) sont ainsi deux échantillons du même gène et sont donc très proches au niveau de leur profil d'expression. Les membres de cette classe se retrouvaient également dans l'analyse sur les 93 gènes de départ. Le nœud le plus significatif correspondait au niveau 76 et était formé par la hiérarchie de cohésion 0,85 suivante :`((4461 ((((1745 (4373 5976)) 1882) 6041) (3341 1784))) (4682 (((922 5884) 4407) ((2659 (2043 (1092 (4847 ((6200 6201) 1249)))) 1811))))`. Les profils d'expression du tableau 1 indiquent effectivement deux sous-groupes de profil, le premier `(4373 5976)) 1882)` étant moins dense en sur-expression que le deuxième :`(4847 ((6200 6201) 1249))`.

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

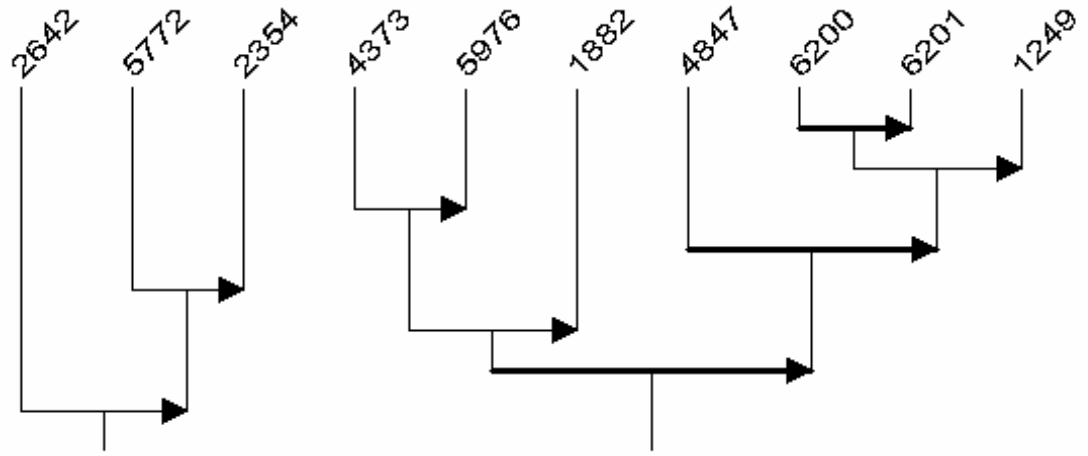

FIG. 5 - *Arbre cohésitif sur les dix gènes sélectionnés par l'analyse. On retrouve les deux classes étudiées (ALL et AML sur les sous-arbres respectifs de gauche à droite) et les implications relatives des gènes.*

Ces différents résultats nous confortent dans la capacité de ce type d'analyse à mettre en évidence les gènes les plus pertinents pour le biologiste. Nous allons maintenant examiner comment il est possible d'interpréter les gènes découverts par une analyse implicative. Pour cette étude, nous allons considérer les 16 gènes les plus contributifs de la classe AML obtenus à travers l'arbre cohésitif. Deux gènes, non identifiés par les nomenclatures imposées par les logiciels utilisés ont été enlevés de cette liste (le précurseur de l'IL8 ainsi qu'un transcrit non identifié). Nous proposons deux approches : la première utilise une ontologie sur les gènes et la deuxième s'appuie sur les connaissances enfouies dans la littérature.

Gene Ontology [GO] est une ontologie permettant de catégoriser les gènes. FatiGO [Al-Shahrour et al. 2004] est un outil qui détermine les concepts de Gene Ontology qui sont communs à un groupe de gènes donné. Sur notre liste, nous avons obtenu 75% des gènes classés au niveau de la *communication cellulaire* et plus particulièrement de la *transduction du signal* (ce dernier concept recense environ 8% de la totalité des entités connues à ce jour). Notons également la présence de trois gènes participant de la réponse immunitaire (*IL8* (#6200, #6201), *IFI30* et *CD1A* (#1882)).
Une autre manière de relever des faits communs consiste à explorer la base de textes *MEDLINE* contenant plus de 12 millions d'articles biomédicaux. *PubGene*, un outil disponible publiquement [Jenssen et al. 2001], permet de relever les co-citations de gènes dans les articles, ce qui, sans constituer en soi une validation formelle, apporte l'indice d'un niveau de similarité dans l'observation et l'étude de ces gènes. Cinq gènes se sont retrouvés isolés sur le réseau de voisinage de *PubGene* et un gène n'a pas pu être identifié. Les 8 gènes restants sont reliés entre eux par le graphe représenté en figure 6.

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

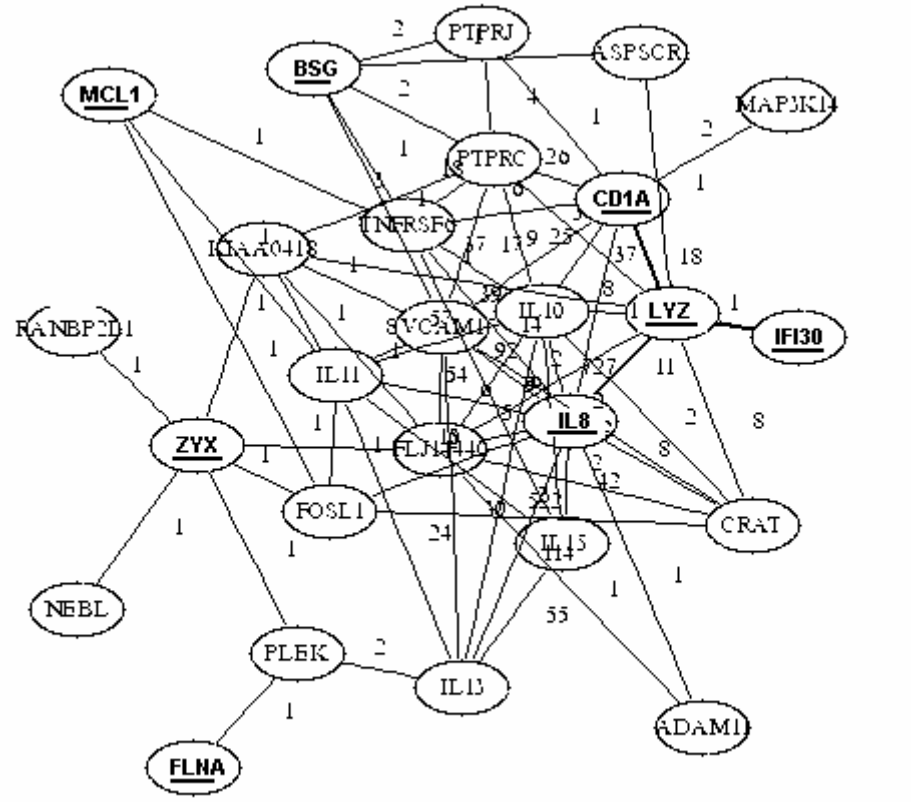

FIG. 6 - *Réseau de co-citations dans la littérature biomédicale par l'outil Network Browser de PubGene. Les nœuds représentent des gènes et les arcs des co-citations (l'étiquetage porte sur le nombre d'articles trouvés). Les gènes soulignés sont ceux de notre étude.*

Ce graphe montre le réseau d'articles qui les lient. Notons que seuls certains gènes sont directement co-cités avec d'autres dans notre liste (il existe cependant un chemin *IL8* (#6200, #6201), *CD1A* (#1882), *LYZ* (#5976), *IFI30*). Restons cependant prudent sur l'interprétation de ce type de graphe, qui ne reflète qu'une vision partielle de l'état de la connaissance en matière d'annotation. De même, nous avons surtout insisté sur les gènes connus pour jouer un rôle dans l'une ou l'autre des deux formes de leucémies. Les gènes présentant un lien d'implication avec ces derniers mais dont la fonction est mal connue (comme le transcrit mentionné précédemment) sont des bons candidats pour une validation de nature expérimentale.

4 Autres travaux

La démarche que nous avons poursuivie vise à identifier des gènes marqueurs d'un certain phénotype. Cet objectif se rapproche des méthodes de sélection de variables où l'on cherche à déterminer un sous-ensemble de variables permettant de discriminer au mieux deux ou plusieurs classes. Deux types de techniques sont généralement employées, celles qui sont uniquement fondées sur les données (*filter*) et celles reposant sur un algorithme de classification (*wrapper*). Les méthodes de type *filter* sont généralement basées sur un ordonnancement des gènes (*ranking*) [Pardo et al. 2004] et des tests statistiques, notamment le *t-test* [Golub et al. 99]. Un problème récurrent de ces démarches est la redondance liée à la forte corrélation des gènes

Troisièmes Rencontres Internationales – Terzo Convegno Internazionale - Third International Conference
A.S.I. Analyse Statistique Implicative – Analisi Statistica Implicativa – Implicative Statistic Analysis
Palermo (Italy) 6-8 Octobre 2005

obtenus : une approche plus satisfaisante consiste à imposer une contrainte de minimisation de la redondance [Ding et al. 2003]. Les techniques de type *wrapper* reposent, quant à elles, sur une optimisation des performances d'apprentissage d'un algorithme par rapport à une sélection donnée. Ces méthodes utilisent des heuristiques qui consistent à rajouter (*Sequential Forward Selection*) ou à éliminer (*Sequential Backward Selection*) itérativement des gènes. Des techniques plus élaborées ont été mises en œuvre pour ne pas être tributaires d'un seul algorithme de classification [Lau et al. 2003].
Dans une approche non-supervisée, il est possible de remplacer la classification par une clusterisation. Cette dernière apporte cependant un biais et une réitération de la clusterisation n'apporte pas nécessairement de meilleurs résultats (problème du maximum local). Il est préférable d'éliminer les gènes les moins discriminants par analyse de variance [Hastie et al. 2000], par analyse par composantes principales ou par analyse des correspondances [Ding 2003]. Le classifieur bayésien naïf a également été appliqué pour déterminer des scores de dépendances entre variables [Søndberg-Madsen 2003].
Notre approche complète ces différents travaux en proposant un outil statistique qui tient compte du contexte global d'expression des gènes au sein d'une méthodologie essentiellement interactive : l'utilisateur sélectionne une classe cohésitive ou à un chemin d'implication particuliers. Cette démarche s'inscrit donc dans une étude locale du comportement de certains gènes et de leurs relations et ne se résume pas à la seule sélection de variables.

5 Conclusion et perspectives

Les données de puces à ADN, par la dissymétrie des matrices, nécessitent un filtrage des gènes. Nous avons utilisé un critère simple qui consiste à ne retenir que les gènes qui s'expriment au moins k fois sur les n expériences, en partant du principe qu'un grand nombre de gènes ne bougent pas vis-à-vis d'une pathologie donnée. Il serait plus judicieux de faire une sélection sur une base statistique solide. La même remarque prévaut sur la binarisation : il existe des méthodes statistiques permettant d'affirmer qu'un gène est bien sur-exprimé et qu'il ne s'agit pas d'un artefact du procédé de mesure. Nous n'avons pas non plus traité d'autres options, comme de considérer des valeurs modales, de type intervalle ou flou.
Cette étude montre que l'analyse implicative est pertinente dans les deux espaces complémentaires que représentent les gènes et les expériences. Un des intérêts de la méthode est d'ailleurs d'établir des ponts entre ces deux vues. En effet, quand on expose à un biologiste une implication et qu'on lui annonce qu'elle est valide avec une mesure de qualité de 0.8, son premier réflexe est de demander quelles entités sont réellement concernées par ce résultat. Le concept de contribution apparaît donc très prometteur dans ce cadre. Outre cette propriété de représentativité d'une règle ou d'une classe de règles, la contribution possède la remarquable faculté de réduire le nombre de gènes d'intérêt, ce qui demeure un des principaux soucis de l'expert biologiste lorsqu'il est confronté à des mesures concernant des dizaines de milliers de gènes.
Nous avons proposé de confirmer la pertinence des résultats obtenus en recoupant les informations extraites avec celles enfouies dans la littérature scientifique. Nous avons également utilisé une technique d'investigation basée sur les ontologies et qui met en évidence des propriétés communes à un ensemble de gènes.

La présente étude est tributaire du jeu de données utilisé et demande à être étendue à d'autres expériences. Une analyse plus fine reste à entreprendre, pour explorer notamment l'adéquation du concept de chemin dans le domaine du transcriptome. Il conviendrait pour ce faire d'enrichir les données d'expression, en considérant par exemple des séries temporelles caractérisées par l'apparition d'un phénomène précis (stress induit par l'expérimentateur sur un organisme cellulaire, prise de médicaments, …). La richesse du contexte sémantique permettrait de mieux appréhender la notion de causalité à travers les outils de l'Analyse Statistique Implicative.

Références

Al-Shahrour, F., Díaz-Uriarte, R. & Dopazo, J. (2004), "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes", Bioinformatics 2004 20: 578-580.
Cong G., Tung A., Xu X., Pan F., Yang J. (2004), "FARMER: Fining Interesting Association Rule Groups by Row Enumeration in Biological Datasets" , Proceedings of the 2004 ACM SIGMOD international conference on Management of data, Paris, France, pp 143 – 154, 2004, ISBN:1-58113-859-8.
Couturier R. (2001), Traitement de l'analyse statistique implicative dans CHIC , Actes des Journées sur la Fouille dans les données par la méthode d'analyse implicative, IUFM Caen, 2001, 33-50.
Ding C., Peung H. (2003), "Minimum Redundancy Feature Selection for Gene Expression Data", Proc. IEEE Computer Society Bioinformatics Conference (CSB 2003), pp.523-529, August 2003. Stanford, CA..
Ding C. (2003), "Unsupervised feature selection via two-way ordering in gene expression analysis", Bioinformatics. v.19, pp.1259-1266, 2003.
[GO] http://www.geneontology.org
Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M., Downing J.R., Caligiuri M.A, Bloomfield C.D. and Lander E.S.(1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science 286:531-537., 15 octobre 1999.
Gras R(égis). et coll. (1996), L'implication Statistique, Grenoble, La Pensée Sauvage
Hastie T., Tibshirani R. et al. (2000), "Gene Shaving: a New Class of Clustering Methods for Expression Arrays", Tech. report. Janvier 2000, Department of Statistics, Standford University.
Jenssen TK, Lægreid A, Komorowski J, Hovig E (2001), "A literature network of human genes for high-throughput analysis of gene expression", Nature Genetics 28:21-8, May 2001.
Lerman I.C. (1981), Classification et analyse ordinale des données, Paris, Dunod.
Liping Ji, Kian-Lee Tan (2004), "Mining Gene Expression Data for Positive and Negative Co-regulated Gene Clusters", Bioinformatics, Oxford University Press, Vol. 20, No. 16, pp. 2711-2718, 2004.
Lau M., Schultz M. (2003), "A Feature Selection Method for Gene Expression Data with Thousands of Features"; internal report, Yale University, 2003.
Oyama T, Kitano K, Satou K, Ito T. (2002), "Extraction of knowledge on protein-protein interaction by association rule discovery", Bioinformatics. 2002 May;18(5):705-14.
Slamon DJ, Boone TC, Murdock DC, Keith DE, Press MF, Larson RA, Souza LM (1986) , "Studies of the human c-myb gene and its product in human acute leukemias", Science. 1986 Jul 18;233(4761):347-51.
Pardo M., Sberveglieri G., Wold B. (2004), "Yet another Feature Selection Study for Microarrays", Bioinformatics Italian Society Meeting (BITS) , Padova, 26-27 mars 2004.
Smith A. (2000), "Exploring Class Prediction for Leukemia Gene Expression Data", Critical Assessment of Microarray Data Analysis (CAMDA), Caroline du Nord, USA, December 18-19, 2000.
Søndberg-Madsen N., Thomsen C. and Pena J. M. (2003). "Unsupervised Feature Subset Selection". In Proceedings on the Workshop on Probabilistic Graphical Models for Classification (within ECML/PKDD 2003), 71-82.
Tusher, Tibshirani and Chu (2003), "Significance analysis of microarrays applied to the ionizing radiation response". PNAS 2001 98: 5116-5121, 24 avril 2001.

Summary

Microarrays or DNA chips monitor expression of tens of thousand of genes simultaneously in one single experiment. This paper introduces the application of Implicative Statistic Analysis to these challenging type of biological data. We apply our study to the differentiation of two forms of leukaemia, the acute lymphoblasic leukaemia and the acute myeloid leukaemia. The dataset comprises 7129 genes and concerns 38 patients. We show that the CHIC software reveals the presence of the two groups of patients and that it extracts a discriminative set of genes. The concept of contribution yields to a sensible reduction of the marker genes and may lead to new diagnostic protocols.