Modélisation et contrôle statistique de l'analyse cytométrique de la ploïdie en cancérologie

Martial Guillaud

THESE

Présentée par

Martial GUILLAUD

Pour obtenir le titre de

Docteur de l'Université Joseph Fourier - Grenoble 1

Spécialité **Génie Biologique et Médical**

MODELISATION ET CONTRÔLE STATISTIQUE DE L'ANALYSE CYTOMETRIQUE DE LA PLOÏDIE EN CANCEROLOGIE

Thèse soutenue le 7 Mai 1993

Composition du Jury :

DEMONGEOT Jacques	Président
EMPTOZ Hubert	Rapporteur
PALCIC Branko	Rapporteur
REITH Albrecht	Rapporteur
RIGAUT Jean-Paul	Examinateur
CHASSERY Jean-Marc	Examinateur
BRUGAL Gérard	Examinateur
SEIGNEURIN Daniel	Examinateur

Thèse préparée au sein du Laboratoire TIMC-USR B 00690

à mes parents,
à Pascale, Fabrice et yayo

Cette thèse a été réalisée au sein de l'équipe de **R**econnaissance des **F**ormes et de **M**icroscopie **Q**uantitative dirigée par le Professeur Gérard Brugal. Je tiens à le remercier très sincèrement pour son accueil, ses conseils et sa générosité. Mes remerciements s'adressent dans les mêmes termes au Professeur Jean-Marc Chassery, Directeur de cette thèse, et responsable du Groupe INformatique et FOrmes DIScrètes. Il n'est pas si courant de travailler et de s'enrichir avec deux personnalités aussi différentes, mais aussi compétentes et sympathiques.

Monsieur le Professeur Jacques Demongeot m'a fait l'honneur d'être président de ce jury et je l'en remercie profondément.

Doctor Branko Palcic, please accept my warmest thanks for agreeing to be in my thesis jury.

I am extremely grateful to Doctor Albrecht Reith for his participation in this jury.

J'exprime tout ma reconnaissance au Docteur Hubert Emptoz qui a bien voulu juger mon travail.

Monsieur le Professeur Jean-Paul Rigaut m'a fait l'honneur de faire partie de ce jury, qu'il veuille bien trouver ici ma sincère reconnaissance.

Pour avoir apprécier notre collaboration, il m'est particulièrement agréable de remercier le Professeur Daniel Seigneurin d'être membre du jury. Que toute l'équipe de Cytologie Quantitative, en particulier Josette, soit associée à ces remerciements.

Je tiens également à exprimer ma sympathie à tous les membres de l'équipe

Aux "*permanents*", Marie-Paule, Jocelyne, Victoria, Michèle, Françoise, Sylvie, Catherine, Annick, Camille, Gérard, Jean-François, Xavier, Yves et surtout Paulette, Nicole, Jean-Dominique et Guy, ne serait ce que pour leur patience et leur aide *permanentes*.

A tous les thésards, DEA, stagiaires, nouveaux et anciens, collaborateurs ou non

A Marisol, Nationalité espagnole,

Pensées toutes particulières à Isabelle et Loïc

SOMMAIRE

INTRODUCTION GENERALE

Il faut prendre garde au danger de mesurer trop
et de penser trop peu
Jean Bernard (1980)

L'objectif des recherches dans le domaine de l'analyse d'images est la conception de programmes ou systèmes informatiques susceptibles de traiter des images présentées sous forme numérique, c'est à dire d'en extraire certaines composantes ou propriétés, de les décrire, et d'en fournir des éléments d'interprétation. Dans le domaine médical et en particulier l'imagerie microscopique, l'automatisation des examens de laboratoires (lecture de lames au microscope), tache fastidieuse et routinière, a été une des premières applications des techniques d'analyse d'images. La confrontation avec l'expertise humaine apparaît au stade ultime de la justification des moyens mis en oeuvre, selon les critères de coût et de performance atteints. L'échec relatif de ces tentatives ainsi que leur extension à la recherche biomédicale a conduit ensuite à focaliser la recherche sur les tâches descriptives, en réservant à l'expert humain la primauté de la prise de décision. On parle donc d'aide au diagnostic. L'emploi des techniques de l'analyse d'images est justifié par l'apport de descripteurs quantitatifs, fiables, robustes et associés à des informations difficilement mesurables par l'observateur humain. Leur exploitation par l'expert humain implique néanmoins qu'une certaine forme de transposition, de leur forme numérique vers une forme sémantique leur soit applicable.

Les traitements statistiques des données issues d'une phase d'analyse d'images biologiques sont souvent effectués indépendamment de toute la chaîne de l'analyse d'image qui comprend l' acquisition des données, la reconnaissance des formes et la paramétrisation. Les

La deuxième partie concerne le développement d'une méthode d'acquisition de données cytométriques sous contrôle statistique en temps réel. Les outils développés ont permis l'élaboration de tests d'arrêt de l'acquisition et de détection d'individus rares.
Après une introduction sur les méthodes de ré-échantillonage (Méthode du Bootstrap), les résultats obtenus sur le codage des histogrammes d'ADN ainsi qu'un protocole de définition d'un critère mesurant la stabilité des histogrammes d'ADN en cours d'analyse seront donnés sous forme de publications. Nous montrerons également que cette méthode peut se généraliser au contrôle de la stabilité de populations caractérisée par plusieurs paramètres cytométriques, toujours en cours d'analyse.

La troisième partie expose, sous forme de deux publications, les travaux réalisés en vue d'une meilleure compréhension des relations existants entre les caractéristiques biologiques d'une tumeur et le diagnostic effectué sur un échantillon de cette tumeur. Après une brève introduction, nous présenterons le modèle de simulation d'une tumeur au sein d'un tissu différencié et montrerons l'intérêt de ce modèle pour étudier et simuler la majorité des problèmes rencontrés dans l'analyse et l'interprétation du contenu en ADN d'un échantillon tumoral.

En conclusion, nous montrerons comment nos travaux ont abouti à l'élaboration de deux projets de recherches, intégrant d'autres disciplines scientifiques comme l'Intelligence Artificielle Distribuée et la sociologie cellulaire.

ANALYSE DE L'ADN : PRINCIPES, INTERETS ET LIMITES DE LA CYTOMETRIE QUANTITATIVE :

Chapitre I

Cytométrie à Balayage : Généralités et Applications

Depuis plus d'un siècle maintenant, biologistes et médecins ont étudié les caractéristiques qui permettent de discriminer les cellules saines des cellules tumorales. En routine clinique, cette discrimination s'appuie presque exclusivement sur l'analyse visuelle, grâce à la microscopie photonique puis électronique, des préparations biologiques. Les pathologistes, grâce à leur expérience accumulée pendant de nombreuses années, rendent des diagnostics supposés fiables et suffisamment précis pour orienter une décision clinique. Néanmoins, de nouvelles méthodologies ont vu le jour depuis une trentaine d'années. On peut parler par exemple de l'utilisation de sondes ADN comme marqueurs in *situ* de la présence d'oncogènes, ou des marqueurs immunologiques. Ces méthodologies s'inscrivent pour l'instant dans une logique de tout ou rien. La mesure quantitative des caractéristiques cellulaires représente une autre méthodologie pour le diagnostic pathologique. En effet, la microscopie quantitative s'intéresse à des paramètres quantitatifs continus comme la taille, la densité ou la forme. Déjà en 1851, Lebert, et plus tard Virchow, mesuraient le diamètre des cellules et de leurs noyaux et notaient certains changements dans les cancers.

C'est donc tout naturellement que les techniques quantitatives sont devenus des outils indispensables pour la compréhension des mécanismes tumoraux et pour l'aide au diagnostic et au pronostic des cancers.

En outre, la cytophotométrie permet de déterminer la proportion relative de certaines substances dans les cellules ou dans les tissus et, en ce sens elle représente une approche cytométrique indépendante de la morphométrie qui permet l'analyse quantitative des propriétés géométriques et structurales. Progrès technologiques et avancées thérapeutiques ont particulièrement été lié dans ce domaine de la biologie. Les systèmes d'analyse d'images représentent une évolution majeure de ces technologies et sont de plus en plus répandus dans les laboratoires de cytopathologie.

I.Microscopie quantitative : Généralités

Plusieurs raisons expliquent l'augmentation considérable de la microscopie quantitative dans l'analyse d'échantillons cancéreux. Tout d'abord, les techniques quantitatives offrent un degré beaucoup plus élevé de l'objectivité des mesures. De plus, cette objectivité facilite, facilite la communication. Il est par exemple plus facile de juger de la surface nucléaire que de définir l' atypie nucléaire. D'autre part, les mesures quantitatives sont plus reproductibles que des observations subjectives. Si ce constat n'est pas forcément vrai pour discriminer des cellules inflammatoires ou des cellules très fortement malignes, il devient inéluctable quand il s'agit de reconnaître des néoplasmes à petites cellules rondes.

I.1.Les différentes méthodes

Différentes techniques quantitatives, chacune possédant des domaines d'applications précis et spécifiques, sont utilisées en routine clinique.

I.1.1 Comptage

Le comptage, une des plus anciennes et plus simples évaluations quantitatives dans le diagnostic pathologique consiste à dénombrer des événements au sein d'une population cellulaire ou tissulaire. Le comptage de mitoses dans un champ microscopique est par exemple très utilisé pour différencier les tumeurs malignes des muscles lisses utérins.

I.1.2 Morphométrie

La morphométrie peut être définie comme la mesure ou l'estimation de distances, d'aires ou de volumes de certaines entités cellulaires. Ce sont donc des mesures continues, issues de la planimétrie ou de la stéréologie pour traiter des images en deux ou trois dimensions. De nombreux auteurs ont proposé des revues sur le principe et les applications de la morphométrie [Baak 1991].

I.1.3 Photométrie

Les mesures photométriques ou densitométriques sont apparues dès 1930 grâce aux travaux de Caspersson [Caspersson 1960] qui a développé le premier appareil permettant de quantifier l'intensité émise ou absorbée par une substance. Ce cytophotomètre utilisait le rayonnement UV puis plus tard la lumière visible. La photométrie a, dès le début, été consacrée à la mesure du contenu en ADN des noyaux, en utilisant soit des marqueurs fluorescents comme l'acridine orange ou des colorants [Feulgen 1924], tous ayant une relation plus ou moins stoechiométrique avec l'ADN [Santisteban 1992].

Un des principaux inconvénients des cytophotomètres vient de l'utilisation d'une fenêtre d'ouverture fixe pour déterminer l'aire de l'échantillon à mesurer. Comme la géométrie de cette fenêtre d'ouverture ne correspond que rarement avec celle de la surface à mesurer, deux sources d'erreurs de mesures sont introduites.

- l'erreur due à la sous-estimation ou à la surestimation de la surface nucléaire,

- l'erreur introduite par l'hétérogénéité de la distribution de la coloration dans l'objet étudié (noyau par exemple).

La densité étant une fonction du rapport logarithmique entre le flux de photons incidents et le flux de photons incidents, les irrégularités dans la distribution produisent un biais positif dans l'estimation de la densité totale.

I.1.4 Cytophotométrie en flux

La cytophotométrie en flux, apparue en 1965, est devenue une technique de routine tant au niveau de la recherche fondamentale qu'au niveau des applications cliniques.

Cette technique permet d'analyser individuellement et très rapidement les individus d'une population cellulaire ou sub-cellulaire en fonction de plusieurs paramètres. Un faisceau de lumière d'excitation, fourni par un laser ou une lampe à vapeur de mercure interagit sur chacun des éléments séparés de la population. La fluorescence émise peut soit résulter d'un marquage soit être issue d'une auto-fluorescence [Laerum 1981 ; Metezau 1988].

La transformation de cette fluorescence en signaux électriques dont l'analyse de l'amplitude permet, par exemple de construire l'histogramme de la répartition des cellules en fonction du, ou des, paramètres étudiés.

Les avantages de cette technique, par rapport à l'analyse d'image, sont en particulier :

- la possibilité d'analyser de 10 000 à 100 000 cellules par minute et donc d'être relativement sensible à la présence d'événement rares, ou de populations minoritaires,

- la possibilité de trier des sous-populations particulières mises en évidences par l'analyse. Les cellules étant dissociées dans un liquide physiologique, ces sous-populations peuvent être à nouveau remises en culture pour des éventuelles analyses ultérieures.

En contrepartie, cette technique présente certains désavantages comme :

- l'impossibilité de mesurer des paramètres morphologiques ou surfaciques,

- l'impossibilité d'étudier des préparations histologiques,

- l'impossibilité d'éliminer les nombreuses cellules contaminantes, ou débris cellulaires, présentes dans la préparation, provoquant souvent des erreurs d'interprétation.

I.1.5 Cytophotométrie en image

L'imagerie numérique représente l'approche la plus générale pour l'analyse d'images de cellules car la morphométrie, la photométrie et les informations spectrales peuvent être simultanément exploitées pour calculer des caractéristiques de haut niveau de l'objet étudiée. C'est en cela que l'analyse d'images numériques se rapproche sans doute le plus du mécanisme de la vision vis à vis des autres instruments cités précédemment. Les systèmes d'imagerie numériques utilisent une caméra qui produit un signal analogique dont l'amplitude est proportionnelle à l'intensité lumineuse du spot observé. Le signal analogique est échantillonné et l'amplitude convertie en valeurs numériques. Ces nombres sont ensuite enregistrés en mémoire dans l'ordinateur.

La résolution spatiale est une des principales caractéristiques des systèmes d'analyse d'images. Cette résolution spatiale, fonction de l'optique du microscope, de la résolution de la caméra et de la densité d'échantillonnage du digitaliseur, peut être décrite par une fonction de modulation de transfert et constitue une des caractéristiques fondamentales du capteur.
Comparativement aux limites de la cytométrie en flux, l'analyse d'images permet :

- d'analyser des préparations histologiques,
- de sélectionner individuellement les cellules et donc d'éliminer les cellules contaminantes,
- la mesure de paramètres morphologiques, surfaciques et colorimétriques.

I.2. Cytométrie par absorption et par fluorescence

La cytométrie en fluorescence et la cytométrie par absorption sont utilisées en fonction des objets à analyser. Le principe global est commun mais ces deux types d'analyses possèdent des caractéristiques spécifiques que nous présentons dans ce paragraphe.

I.2.1 La cytométrie par fluorescence

Les mesures de fluorescence en biologie fondamentale et médicale [Waggoner 1986] sont de plus en plus pratiquées grâce aux progrès de la chimie des fluorochromes [Mathies 1986] de la diversification des marqueurs immunologiques disponibles et du perfectionnement des techniques de cytométrie [Lansing-Taylor 1986]. La fluorescence est le phénomène physique selon lequel certaines molécules émettent de la lumière pendant une excitation photonique, et ceci dans un laps de temps inférieur à la seconde. Les électrons périphériques libres de la molécule excitée à une longueur donnée, sont portés à un niveau énergétique supérieur au niveau énergétique initial (stable). Après une période définie comme la durée de vie de l'état d'excitation parfois de l'ordre de la nanoseconde, ils retournent à leur niveau énergétique initial en émettant un flux photonique. La diversité des fluorochromes utilisables en cytométrie permet de nombreux typages immuno-fluorescents ou cyto-chimiques ainsi que la révélation de sondes nucléiques après hybridation in situ. La grande spécificité de l'information révélée par les marqueurs fluorescents impliquent naturellement la nécessité d'une quantification très précise [Guilbault 1973 ; Shapiro 1983].

I.2.2 La cytométrie par absorption

Cette méthode est fondée sur le fait que l'intensité de lumière projetée à travers un objet est réduite par la présence de certaines substances dans le noyau. La loi de Beer-Lambert [Atkin 1956] que suivent les mesures d'intensité permet de caractériser le pouvoir absorbant de tel ou tel objet, en fonction de la lumière incidente. La densité optique d'un objet est en principe, dans le cas de colorations stoechiométriques comme celle de Feulgen, proportionnelle à la quantité de substances absorbantes [Piller 1987].

II. Les systèmes d'analyse d'images

L'analyse d'image pour la quantification de phénomènes biologiques microscopiques a débuté il y a une quarantaine d'années par la mesure de l'ADN. [Leuchtenberger 1954, 1960 ; Ojima 1960 ; Sandritter 1966 ; Zetterberg 1976]. Les premiers systèmes de cytométrie utilisés furent essentiellement orientés vers l'automatisation en cytologie gynécologique [Tanaka 1979 ; Ploem 1986 ; Brugal 1984].

Les avancées technologiques récentes (informatiques, optiques, électroniques) ont permis d'aller de plus en plus loin dans l'analyse des éléments constitutifs des cellules. Les systèmes d'analyse d'images sont donc tout naturellement rentrés dans les laboratoires de routines cliniques ainsi que dans les laboratoires de recherche en cancérologie ou en biologie fondamentales. Avant de détailler les différentes étapes de l'analyse d'image, il est utile de rappeler les techniques d'échantillonnage et les caractéristiques de l'instrumentation.

II.1. Echantillonnage et Préparation biologique

II.1.1 Prélèvement

Le prélèvement de l'échantillon (ou biopsie) est la première étape d'une analyse cytométrique (Figure I.1). Le prélèvement dépend du type cellulaire à analyser et est obtenu par méthode endoscopique ou chirurgicale :

- Le grattage est utilisé dans tous les cas de prélèvement de tissus épithéliaux directement accessibles (col de l'utérus, muqueuse vaginale, etc..);
- L'aspiration par aiguille fine (Fine Needle Aspiration Biopsy) est adaptée au prélèvement de cellules en suspension dans des milieux liquides et aux tissus mous : sang, liquide céphalo-rachidien, liquide pleural, moelle osseuse, tumeurs des tissus mous, etc..;
- Les empreintes cytologiques sont obtenues par apposition de la pièce d'exérèse sur une lame microscopique.

Les cellules à analyser sont placées sur une préparation microscopique classique (lame de verre) que l'on traite par balayage (déplacement manuel ou automatique de la lame sous une caméra). La quantification intracellulaire exige la mise en oeuvre de réactions spécifiques et stoechiométriques.

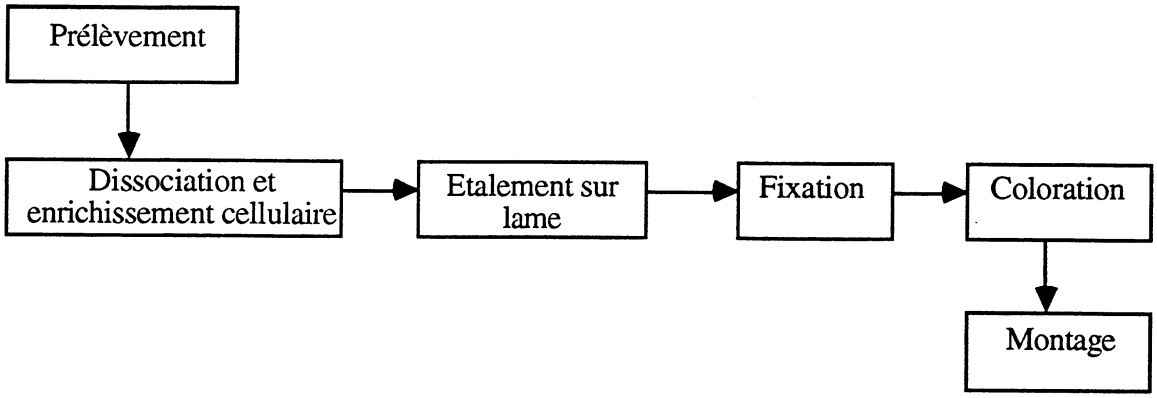


Figure I.1 : Les différentes étapes de l'obtention d'une préparation cellulaire à partir d'un tissu.

II.1.2 Fixation et coloration

Les problèmes liés à la fixation et à la coloration ont fait l'objet d'une revue par Wittekind [Wittekind 1985]. Ces problème sont cruciaux pour une mesure reproductible et objective de cellules en vue d'une reconnaissance optimale par les systèmes automatiques ou semo-automatiques. Selon la situation, la coloration et la fixation optimale n'est pas toujours la même. Aucune règle générale ne peut être définie. La préparation doit être adaptée :

- à l'origine tissulaire du matériel cellulaire analysé
- au mode de prélèvement
- à la technique cytochimique employée

- Séchage à l'air

La vitesse de déshydratation est fonction de la nature du tissu.

- Etalement (richesse cellulaire et déformation)

Deux exigences contradictoires : d'une part l'organisation des cellules est un élément diagnostique et, d'autre part, les cellules doivent être dissociées pour répondre à la fois aux exigences d'homogénéité de réponse vis à vis de la coloration et aussi aux exigences de la segmentation.

- Fixation

Elle est adaptée à la technique cytochimique ou cytoenzymologique employée. Le critère est ici la conservation des molécules et des structures à révéler .

La qualité des différentes étapes de préparation du spécimen est primordiale et nécessite [Scwartz 1983] :

-que les cellules soient déposées régulièrement avec une densité cellulaire optimale,

- que les cellules soient sur un même plan optique,

-que le contraste entre le noyau et le cytoplasme d'une part et le cytoplasme et le fond soit suffisant,

- que le marquage soit reproductible.

De nombreux ouvrages et revues existent sur les différentes techniques de préparations des spécimens en vue de la quantification par analyse d'images [Rosenthal 1987].

II.2. Instrumentation

Un système d'analyse d'images comprend un capteur, lui-même composé d'un microscope et d'une caméra, un convertisseur analogique/numérique, un ordinateur qui gère l'ensemble des opérations et un écran pour visualiser les opérations effectuées sur l'image (Figure I.2).

Le capteur est composé de différentes parties dont la qualité est étroitement corrélée à la précision des mesures effectuées.

Les microscopes à transmission sont conventionnellement utilisés en cytopathologie. Cependant un nombre croissant d'applications basées sur l'utilisation de fluorochromes et de sondes fluorescentes nécessitent la mise en oeuvre des techniques de microscopie à fluorescence. Les microscopes sont classiquement équipés d'objectifs de grandissements X20, X40, X60 et X100. Les plus forts grossissements nécessitent en général l'utilisation d'huile à immersion pour diminuer les effets de "glare" (effets dus à la lumière parasite). Pour la même raison, les ouvertures numériques des condensateurs sont limitées à 0.08.

Pour rehausser le contraste des préparations, les microscopes sont équipés de filtres dont la longueur d'onde dominante est fonction de la coloration employée. On utilisera, par exemple, un filtre de 550 nm, qui correspond à la couleur verte , pour augmenter le contraste de la coloration pourpre de Feulgen.

Lorsque le microscope est intégré à un système d'analyses d'images il est en général doté d'une platine motorisée ce qui permet le déplacement de la préparation cytologique de façon automatique par commande depuis l'unité centrale de l'ordinateur.

II.2.1 La composante électronique

Les caméras utilisées pour travailler dans le domaine du visible emploient essentiellement deux technologies :

- la technologie à tube(type Vidicon), basé sur le principe qu'une résistance de la couche photoconductrice, et donc le courant, varie proportionnellement avec la quantité de lumière.

- la technologie CCD (Charge Coupled Device) qui exploite les propriétés du Silicon. Ces caméras offrent des avantages non négligeables du point de vue de la précision géométrique, de la linéarité de la réponse de la dynamique, d'un grand temps d'intégration possible et surtout de la sensibilité obtenue lorsque l'on travaille dans le rouge.

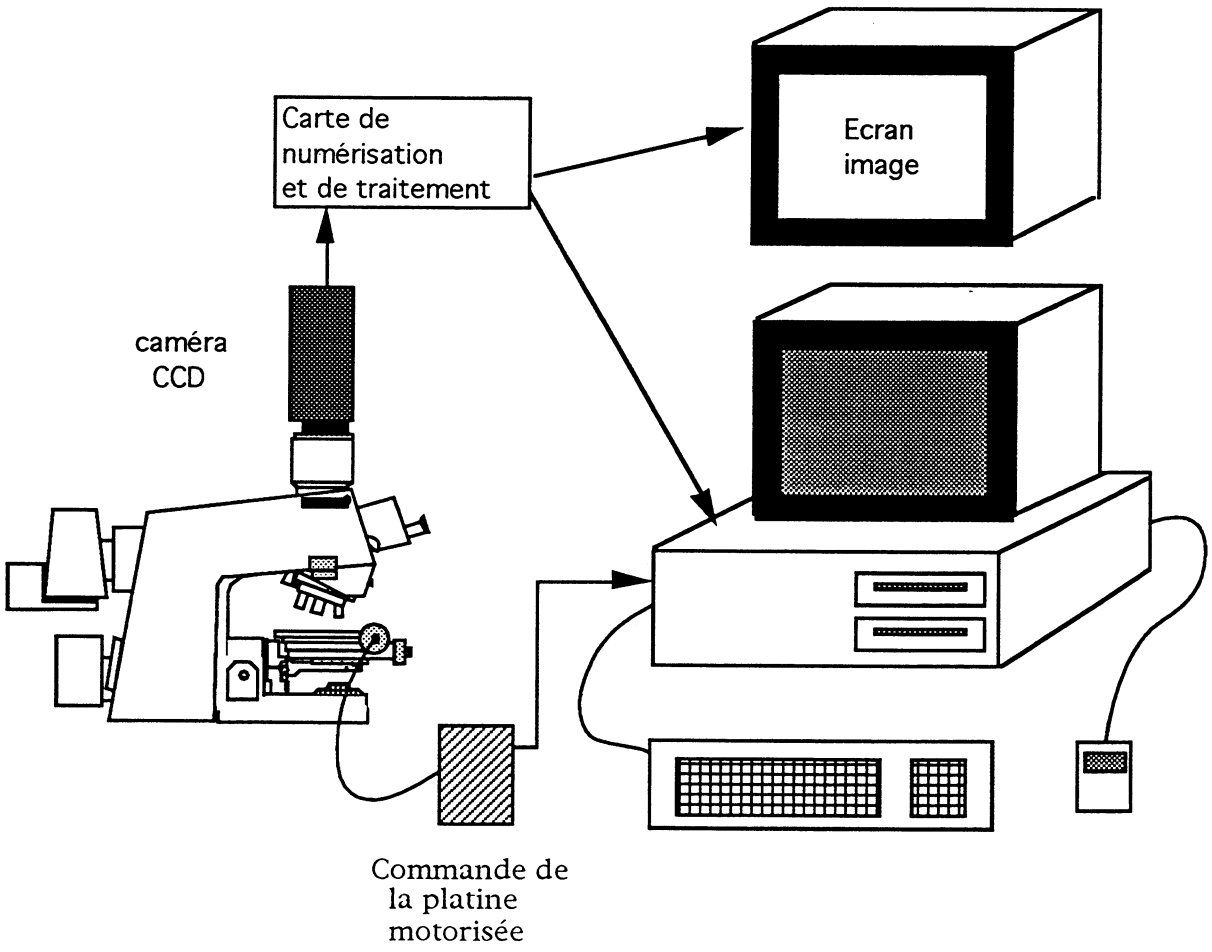


Figure I.2 : Vue général d'un système d'analyse d'image.

II.2.2 Convertisseur Analogique/numérique

Il s'agit de convertir l'information de luminance délivrée par la caméra en image numérique. Pour ce faire le signal est échantillonné puis numérisé. De façon générale, la carte de numérisation comporte trois grandes parties :

- une entrée analogique permettant d'acquérir l'image fournie par la caméra,
- une liaison à l'unité centrale de l'ordinateur permettant à cette dernière d'accéder aux zones mémoires images de la carte,
- une sortie analogique permettant d'afficher sur un moniteur l'image numérisée.

L'entrée et la sortie analogiques sont équipées d'une table de fausses couleurs (ou LUT = Look Up Table) donnant la possibilité de corriger, par exemple, la linéarité du capteur en entrée.

II.3. Analyse d'images

L'analyse d'image exploite la représentation d'une cellule sous forme d'une matrice numérique qui est ensuite traitée par un ordinateur, selon les étapes présentées dans la Figure 3, et comportant successivement une acquisition, un pré-traitement, une segmentation, une paramétrisation et le traitement des données [Brugal 1984].

II.3.1 Acquisition des images

Pour pouvoir être traitée par ordinateur, une image physique doit être digitalisée, c'est à dire mise sous la forme d'une matrice numérique. Chaque valeur numérique représente un niveau d'intensité relative de la lumière transmise ou émise en chaque point de l'image. L'acquisition des images se fait par l'intermédiaire de capteurs (caméra vidéo, scanner photométrique). Ces capteurs doivent être adaptés à la dynamique de l'image et aux besoins en résolution spatiale élevée. Le passage de la scène optique à l'image numérique s'effectue en deux étapes : l'échantillonnage et la quantification. Tout point d'une scène est représentée sous forme spatiale dans un repère cartésien et sous forme spectrale dans un repère colorimétrie associé aux caractéristiques du capteur. La discrétisation du repère spatial fait l'objet de l'échantillonnage : on parlera de pouvoir de résolution spatiale associé au capteur. La quantification consiste en la discrétisation du repère spectral : on parlera alors de dynamique de résolution associée au capteur. Il n'existe pas de dispositif de balayage optimal convenant à tous les types d'images susceptibles d'être analyser. Le système peut varier en fonction de la taille de l'image, la résolution spatiale, la résolution densitométrique et la résolution spectrale du détecteur.

II.3.2 Prétraitement de l'image

Ce prétraitement a pour objectif d'améliorer les contrastes entre éléments à analyser et le reste de l'image (fond). La plupart de ces méthodes sont basées sur l'utilisation de filtres tels que les filtres de Fourier de la moyenne, du gradient ou Laplacien qui ont pour objectifs soit de lisser, soit d'augmenter le contraste de l'image. Si ces méthodes sont adaptées à l'analyse de la texture de certaines scènes, elles ne peuvent être suivies d'une analyse densitométrique, les données numériques de base étant transformées de façon non linéaire.

II.3.3 Segmentation

La segmentation des images consiste à discerner les différents objets qui composent l'image, cette étape est la plus difficile du traitement d'images et la plus spécifique de l'application. Aucune des diverses méthodes utilisés ne conduit néanmoins à un résultat comparable à celui donné par l'ensemble oeil-cerveau. La diversité des outils de segmentation s'impose par la difficulté de définir une méthode performante quelque soit la scène analysée [Chassery 1991]
On regroupe les méthodes de segmentation en deux grandes classes :

+ celles qui consistent à voir un objet comme un ensemble de points d'intensités lumineuses voisines.

+ celles qui s'intéressent uniquement aux contours de l'objet et détectent les frontières entre les différentes plages d'intensité lumineuse similaire. Pour ce faire on utilise des masques de convolution associés à des opérateurs différentiels tels que le gradient de Sobel, de Kirsh, de Prewitt. C'est la première de ces méthodes qui est mise en oeuvre lors de la segmentation des images nucléaires après coloration de l'ADN par la réaction de Feulgen.

Les méthodes de seuillage, basés sur les densités optiques ou sur les gradients locaux sont les plus simples et les plus rapides techniques de segmentation [MacAulay 1988]. Le lissage des images, la modification d'histogramme [Rosenfeld 1978], ou d'autres méthodes [MacAulay 1990 a] récentes sont également utilisées pour améliorer la segmentation des images, étape très délicate de l'analyse d'images.

II.3.4 Paramétrisation

Suite à l'étape de segmentation, les différents objets contenus dans l'image ont été isolés. Ces objets peuvent être superposés à l'image initiale aux erreurs de segmentation près. Mais aucune information n'est connue sur l'objet lui-même. Cependant, ces objets doivent être décrits en vue de les classifier et de les reconnaître. C'est l'objet de l'étape de paramétrisation, ou extraction des caractéristiques, qui permet de décrire l'image par l'intermédiaire des objets qui la composent mais également de réduire les dimensions de l'espace de travail à celle du vecteur unidimensionnel des paramètres calculés. Il est évident que les paramètres que l'on va choisir dépendent de l'application traitée. En cytologie, comme dans de nombreux autres domaines, les paramètres calculables sur des images discrètes peuvent être classés en trois grandes classes :

- Paramètres liés à la distribution spatiale des points constituant un objet de l'image

Ce sont des paramètres de formes, ou de surface ne nécessitant pour leur calcul que l'image des masques.

- Paramètres liés à la distribution des niveaux de gris :

Ces paramètres de type densitométrique ou texturaux nécessitent de calculer l'image initiale masquée avec un masque de segmentation.

Pour l'analyse de l'ADN, on effectue un calcul densitométrique grâce à la réaction de Feulgen. Ce réactif est considéré comme stoechiométrique et spécifique de l'ADN. Elle peut donc permettre la mesure de la quantité de la Densité Optique Intégrée (DOI) qui représente l'absorbance totale du noyau. Ce paramètre répond aux lois de Beer-Lambert de la spectrophotométrie traditionnelle.

Le calcul densitométrique traduit les valeurs intégrées et moyennes ainsi que la distribution des densités optiques des objets.

La DOI qui reflète l'absorbance totale du noyau est donnée par la formule suivante :

$$IOD = \sum_{N} \{(\log I_o(x,y)) - (\log I(x,y))\}$$

avec

N = nombre de pixels du noyau ou surface du noyau,

I = intensité lumineuse de l'image au point (x,y),

I_0 = intensité lumineuse de l'image au point (x,y) si il n'y avait pas d'objet absorbant.

La densité optique moyenne (DOM) qui représente la concentration en ADN par unité de surface nucléaire :

$$MOD = \frac{IOD}{N}$$

La distribution des densités optiques d'un objet traduit l'apparence et le contenu de l'image du point de vue de la distribution des niveaux de gris dans l'objet. Un nombre considérable de travaux a montré que ces paramètres étaient susceptibles de discriminer les cellules bénignes des cellules malignes.

Les paramètres de textures apportent également de précieuses informations sur l'état fonctionnel d'une cellule. Différentes méthodes de mesures de la texture ont été développé dont les plus anciennes sont le calcul des matrices de co-occurence [Pressman 1976] et des longueurs de sections [Galloway 1975].

- Paramètres liés à l'exploitation des couleurs qui nécessitent l'acquisition de trois images élémentaires (Rouge, Vert, Bleu). L'analyse des caractéristiques colorimétriques d'une préparation biologique est relativement récente et encore peu utilisée en cytométrie.

II.4. Analyses des données et classification

L'intérêt de l'analyse d'images réside dans sa capacité à mesurer d'autres informations que la densité optique. Que ce soit en routine clinique ou en recherche fondamentale, la possibilité d'opérer en mode semi-automatique, permet de visualiser chaque cellule analysée et donc de distinguer et d'éliminer d'éventuelles cellules dégénérées, des débris cellulaires ou même des cellules contaminantes qui risquent de "bruiter" l'analyse. Par opposition à la cytométrie en flux qui serait plus quantitative que qualitative, la cytométrie à balayage s'avère plus qualitative que quantitative, puisqu'en général seule deux à trois cents cellules sont analysées. La démarche décrite précédemment est illustrée par la Figure I.3 est commune à la recherche fondamentale et aux applications cliniques.

L'évolution de la cytophotométrie a été suivit par l'évolution des outils statistiques, tout deux étroitement dépendant du développement de l'informatique. Pour étudier et analyser des échantillons cellulaires sur lequel plusieurs dizaines de paramètres sont évalués, l'utilisation d'analyses multifactorielles s'avère parfois nécessaire. A chaque application , des outils statistiques différents s'imposent.

Bartels [Bartels 1979a et b] a écrit une série d'articles traitant de l'évaluation statistique des données cytologiques. Nous donnerons ici sommairement les grands types de traitements statistiques utilisés en cytométrie par analyse d'images.

II.4.1. Statistique classiques

-Analyse mono-paramétrique

Lorsqu'un seul paramètre est mesuré, les outils statistiques classiques sont utilisés dans le but de décrire, de comparer certaines caractéristiques de la variable étudiée :

- *Représentation graphique* : histogramme, diagramme en bâtons, boite à moustache ...

-*Tests statistiques* : Comparaison de moyenne, de proportions, de variances (Test de Student, Kolmogorov-Smirnov, Mann et Whitney, Wicoxon, etc...). Ces tests permettent de comparer par exemple les caractéristiques d'un paramètre mesuré sur deux échantillons différents, dans le but de déceler si tel ou tel paramètre est en moyenne plus fort dans un échantillon que dans un autre.

- Analyse bi-paramétrique

- *Représentation graphique* : les histogrammes bi-dimensionnels, ou les nuages de points sont des représentations très utilisées. Ils permettent de voir comment un paramètre se comporte en fonction d'un autre, comme la surface nucléaire et da quantité d'ADN permettant ainsi de discriminer différents groupes.

- *Corrélations* : les études de corrélations sont très utilisées pour étudier les relations existants entre deux variables mesurées sur un échantillon cellulaire. Mais la significativité du coefficient de corrélation demeure une mesure statistique très délicate à manipuler.

- Analyse multi-paramétrique

De nombreuses activités scientifiques commencent par un recueil de données stockées dans des tableaux numériques de grandes tailles. Ces tableaux sont en général à deux dimensions : ils permettent de décrire un certain nombre d'objets souvent dénommés "individus" à l'aide d'un nombre fini de paramètres. Les besoins et la nécessité de classer, de représenter ou de reconnaître ces objets existent depuis fort longtemps. Cependant, les moyens mis à la disposition des scientifiques étaient limités. Les capacités de calcul de plus en plus grandes des ordinateurs ont permis la mise en oeuvre de méthodes puissantes avec des objectifs et selon des contextes très divers. Certaines de ces méthodes ont pour but d'automatiser et d'étendre notre capacité à appréhender la nature des phénomènes sous-jacents aux données. L'objectif des méthodes d'analyse de données est de ne présenter que les faits bruts en ayant

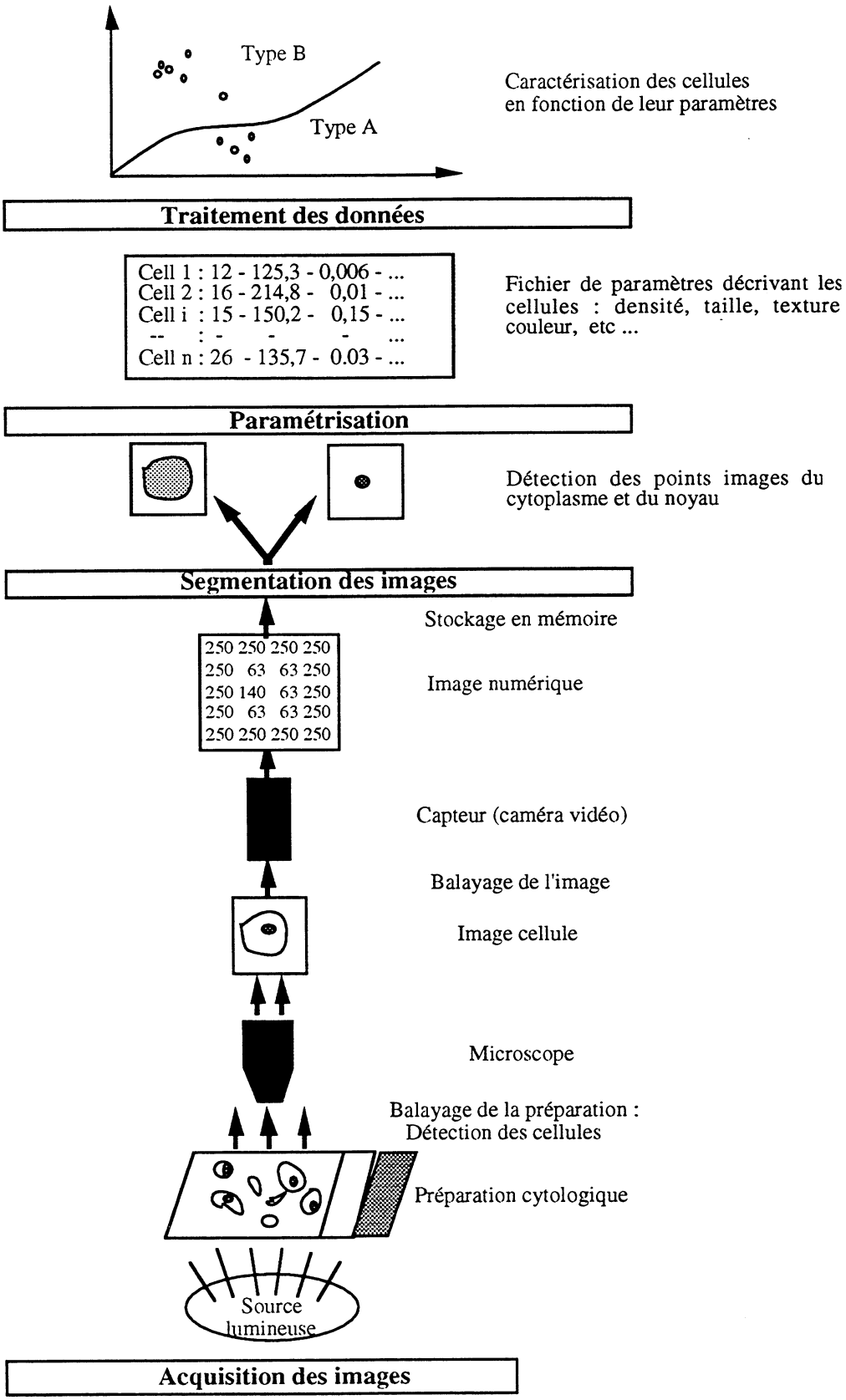


Figure I.3 : Schéma général d'un système d'analyse d'images par microscopie à transmission..

soin de les examiner sous un angle tel que les tendances apparaîtront d'elles-mêmes. Certains analyseurs d'images actuels calculent jusqu'à trente paramètres sur chaque cellule ou sur l'un de ses constituants de sorte qu'une analyse de données multi-dimensionnelles est en général nécessaire à l'exploitation de ces paramètres dans le contexte de l'expérience réalisée ou de la pathologie à diagnostiquer.

On peut séparer ces méthodes (ou méthodes de statistiques multidimensionnelles) en deux grandes familles : les méthodes factorielles et les méthodes de classification.

-Les méthodes factorielles

Les méthodes factorielles fournissent des représentations synthétiques de vastes ensembles de valeurs numériques. Si leurs principes sont anciens (Spearman en 1904), elles ont surtout été développées dans les années 1960-1970. Ces méthodes utilisent des calculs d'ajustement qui font appel à l'algèbre linéaire et produisent des représentations graphiques où les objets à décrire deviennent des points sur un axe ou dans un plan.

A partir des paramètres initiaux caractérisant les individus, les méthodes factorielles cherchent à exprimer de nouveaux paramètres qui permettent de visualiser, de mettre en relief et de mesurer l'information pertinente contenue dans le tableau de données initiales mais qui échappe à l'observateur humain. Lebart [Lebart 1979] et Diday [Diday 1982] décrivent de façon détaillée les principes des différentes méthodes d'analyse factorielle. Parmi celles-ci, les outils de choix les plus utilisés sont l'analyse en composantes principales et l'analyse discriminante :

Analyse en composantes principales (ACP)

L'ACP calcule de nouveaux paramètres ou "composantes principales " qui sont des combinaisons linéaires des paramètres initiaux. La première composante principale correspond à la direction de l'espace multidimensionnel suivant laquelle les individus sont le mieux différenciés. Elle explique la plus grande partie possible de la variance totale, alors qu'inversement, la dernière composante explique la variance la plus faible. Les deux premières composantes principales définissent un "plan principal" où sont représentés à la fois les individus et les paramètres initiaux. Si la part de variance expliquée par le premier plan factoriel est insuffisante, d'autres plans peuvent être définis par combinaison de la première ou de la deuxième composante avec la troisième, etc...

En ayant soin de ne pas confondre proximité sur le plan factoriel, et proximité dans l'espace des paramètres, l'ACP permet alors d'étudier les relations existant entre paramètres, entre individus, et entre paramètres et individus [Manly 1989] .

Analyse discriminante (AD)

L'analyse discriminante permet la mise en évidence des liaisons entre un caractère qualitatif à expliquer et un ensemble de caractères quantitatifs explicatifs [Manly 1989]. Elle existe sous deux formes : l'analyse factorielle discriminante et l'analyse discriminante décisionnelle :

Analyse factorielle discriminante (AFD)

L'AFD permet à l'aide d'une visualisation sur un plan factoriel approprié, de décrire les caractères quantitatifs qui séparent au mieux les individus appartenant à des classes définies par l'utilisateur. Elle cherche à définir des fonctions linéaires discriminantes dont les valeurs pour les individus d'une même classe sont les plus concentrées possibles et les plus dispersées possibles pour les individus appartenant à des classes différentes. Le principe est de maximiser l'inertie inter-classe et de minimiser l'inertie intra-classe du nuage multidimensionnel.
Il existe des méthodes dites "pas à pas" (stepwise) qui consistent à introduire dans la fonction discriminante les variables les unes après les autres suivant l'importance de leur apport conditionnel vis à vis des variables déjà introduites. Elles permettent de répondre à la question suivante : peut-on réduire le nombre de variables quantitatives de manière à ne garder que les variables qui, globalement, différencient le mieux les divers groupes ?

Analyse discriminante décisionnelle (ADD)

A partir d'un ensemble d'apprentissage, l'ADD calcule des "fonctions discriminantes", combinaisons linéaires des caractères initiaux. La qualité de cette analyse est évaluée à l'aide de plusieurs critères comme le "pouvoir discriminant" et la matrice de confusion qui permet de comparer les résultats obtenus par l'ADD avec ceux qu'un observateur expert aurait proposé.
Le but de l'ADD est donc d'établir une règle de décision utilisée, dans un deuxième temps, pour définir l'appartenance d'un individu anonyme à l'une des classes de référence. Différentes méthodes existent :

- les méthodes géométriques basées sur la mesure des distances entre l'individu et les groupes, comme la distance de Mahalanobis [Mahalanobis 1948] : l'individu sera affecté au groupe dont le centre de gravité est le plus proche.

- les méthodes statistiques ou probabilistes qui affectent l'individu au groupe le plus "probable" (méthodes bayésiennes [Dubuisson 1990]).

Un exemple d'application cité par Berthier et Bouroche [Berthier 1975] illustre l'utilisation de l'ADD pour l'aide au diagnostic médical. La discrimination de groupes diagnostiques dans les cancers cervicaux a également été réalisée par cette méthode [Wheeler 1987] utilisait également l'analyse discriminante pour
Notons que l'AFD est en fait une ACP du nuage des centres de gravité des différentes classes étudiées. D'ailleurs, toutes les méthodes d'analyse factorielles peuvent être considérées comme une ACP particulière. Aussi, nombreux sont les mathématiciens qui se sont attachés à en approfondir la connaissance.

-Les méthodes de classification

Le développement des méthodes de classification automatique est plus récent que celui des méthodes factorielles. Si les circonstances de leur utilisation sont sensiblement les mêmes, ces méthodes diffèrent par leurs principes. En effet, les méthodes de classification mettent en

jeu des formulations et des calculs algorithmiques, produisent des classes ou des familles de classes et permettent de grouper et de ranger les objets à décrire. Pour cela, elles nécessitent toutes le choix d'une mesure de ressemblance. Le nombre de méthodes de classification et de variétés d'algorithmes utilisés est considérable. Al Nachawati [Al Nachawati 1985] a proposé un algorithme de classification non arborescent spécifiquement destiné à l'aide au diagnostic. Dans un ouvrage de référence, Jambu [Jambu 1978] décrit les différentes techniques existantes. On peut les regrouper en trois grandes familles :

- Les classifications ascendantes hiérarchiques, comme par exemple la technique dite "de saut minimal " et la technique "d'agrégation selon la variance" [Haroske 1990],

- Les classifications descendantes hiérarchiques, procède par dichotomies, ou scissions, successives. A chaque étape de l'algorithme, deux règles sont à appliquer pour déterminer : le choix de la classe à scinder en 2 et le mode d'affectation des individus à chacune des sous-classes [Fages 1978 ; Hubert 1973].

- Les classifications non hiérarchiques conduisant directement à des partitions comme les méthodes d'agrégation autour de centres mobiles ou les méthodes des nuées dynamiques [Diday 1971], spécialement adaptées à la reconnaissance des formes.

Citons également l'utilisation de classifieurs bayésiens dans la détermination de diagnostic en cyto-histopathologie [Oliver 1977].

L'utilisation d'une méthode de classification est très dépendante du domaine d'application, du problème posé et du type d'information recherchée. On peut obtenir soit une hiérarchie, soit un arbre, soit une partition, soit une typologie, soit des "classes empiétantes" [Dubuisson 1990]

Les caractéristiques obtenues sur chacune des cellules d'un échantillon sont utilisées pour la classification :

- des cellules individuelles,
- et, à un niveau plus élevé, du spécimen étudié.

Les méthodes factorielles et les méthodes de classification sont plus complémentaires que concurrentes et peuvent avec profit être utilisées conjointement sur un même ensemble de données. On commence souvent à positionner les individus à décrire les uns par rapport aux autres à l'aide d'une représentation spatiale continue dans un plan factoriel. On cherche ensuite à les regrouper, et on examine s'il existe des constellations que la procédure précédente n'aurait pas pu mettre en valeur. Le recours aux méthodes factorielles est donc souvent un préalable indispensable pour percevoir la structure de l'ensemble des données. Notons également que de nombreux auteurs ont montré que les différentes méthodes de classifications donnent la plupart du temps des résultats très similaires [Bengtsson 1987]

II.4.2. Les nouvelles méthodes

Le diagnostic et le pronostic médical à partir de données cytométriques a bénéficié, comme d'autres domaines scientifiques, de l'essor considérable des sciences informatiques, des systèmes experts, des réseaux de neurones et des outils de l'Intelligence Artificielle.

- Les réseaux de neurones

Depuis quelques années se sont développées de nouvelles techniques d'analyses de donnée, appelées réseaux de neurones [Dytch 1990 ; Nafe 1992]. Les réseaux de neurones, qui s'appuient également sur la puissance de ordinateurs pour manipuler les données, se rapprochent des méthodes d'analyses factorielles discriminantes [Muller 1990 ; Dawson 1991]. Comme l'Analyse discriminante décisionnelle, un réseaux de neurones permet de classer des individus dans des groupes, pré-définis dans l'étape d'apprentissage. L'efficacité des réseaux de neurones, comme celle de l'ADD, dépend essentiellement de cette étape d'apprentissage. En effet, si les échantillons d'apprentissage ne sont pas suffisamment représentatifs de tout les catégories "diagnostiques" possibles, aucun réseau de neurones, aussi élaboré soit-il, ne peut permettre 100% de bonne classification des individus "test". Que ce soit pour la classification de cellules dans un certain nombre de types (normale, anormale, douteuse) ou pour la classification de spécimens dans des groupes "diagnostiques" (bénin, hyperplasique, malin, etc..), le problème principal reste la bonne caractérisation des classes diagnostiques "rares" [Dytch 1991]. Les recherches actuelles s'orientent d'ailleurs sur le développement d'algorithme d'apprentissage structurel, c'est à dire l'intégration de nouvelles sorties possibles quand de nouveaux modes sont ajoutés aux processus de décision.

-Systèmes experts , intelligence artificielle, et les statistiques incertaines

Dans la dernière décennie, de nombreux systèmes experts d'aide au diagnostic histologique ou cytologique à partir d'informations cliniques et biologiques on été développés [Bartels 1992a et b]. Les outils de l'Intelligence Artificielle [Schaberg 1992] permettent de définir, à partir des informations cliniques et cytologiques, des règles inductives aboutissant au diagnostic et au pronostic [Wied 1990 ; Heddfield 1992]. Les résultats obtenus avec ces systèmes experts s'avèrent toutefois moins intéressant que l'on espérait [Uckhun 1992].

Certains auteurs étudient l'apport de méthodes probabilistes, comme la logique flou ou la théorie des probabilités, dans le diagnostic médical [Bartels 1992b et c ; van Ginneken 1991]

III. Applications en cytopathologie

L'observation microscopique des cellules est indispensable dans des situations biomédicales variées : biologie fondamentale, cancérologie, hématologie, gynécologie, cytogénétique...[Baak 1992]. Les méthodes d'investigation conventionnelles nécessitent de longues et fastidieuses lectures d'un grand nombre de lames microscopiques et l'observateur humain n'appréhende l'information contenue dans une image que de manière subjective, qualitative, et lente. Depuis peu, la diffusion des analyseurs d'images cytologiques permet une description quantitative des cellules et des tissus, d'en extraire des mesures pertinentes et objectives, d'en évaluer la signification statistique afin de permettre une comparaison directe entre les résultats obtenus et la compréhension intuitive des phénomènes biologiques pour la confirmer ou l'infirmer.

Preston et Bartels [Preston 1988] ont présenté les différents aspects de l'analyse d'images microscopiques en cytologie et en histologie. Il ne fait plus aucun doute que la visualisation et la quantification des mécanismes biologiques au niveau cellulaire représentent une approche féconde pour :

- valider les hypothèses proposées par la biologie moléculaire;
- formuler les nouvelles questions posées en biologie fondamentale;
- tenter de préciser les diagnostics et les pronostics médicaux formulés par les cytopathologistes et les histopathologistes.

Les différentes propriétés des cellules ou des constituants cellulaires qu'il est possible de mesurer par cytométrie à balayage sont [Brugal 1987]:

- la taille et les différents paramètres de forme;
- le contenu, la concentration de toutes substances colorables et absorbantes ;
- les antigènes ;
- la plupart des activités enzymatiques ;
- la viabilité cellulaire ;
- les altérations de forme, du rapport nucléo-cytoplasmique, de couleur et de texture dans les situations expérimentales ou pathologiques les plus diverses.

III.1. Applications biologiques.

De nombreux auteurs [Giaretti 1983] ont montré que certains paramètres de texture permettent de baliser de façon précise la progression des cellules au cours du cycle cellulaire [Humbert 1992] . Une analyse factorielle discriminante des données acquises par cytométrie à balayage peut être utilisée pour classer et compter automatiquement les cellules issues d'une culture dans chacune des phases du cycle cellulaire [Giroud 1982] et fournir ainsi une analyse très complète de la cinétique de prolifération de cette population. De nombreux éléments constitutifs de la cellule ou du noyau, comme les récepteurs hormonaux [Charpin 1986], le cytosquelette [Léger 1990] et bien sûr l'ADN sont étudiés par imagerie quantitative.

Le développement des systèmes d'analyse d'images a contribué à une meilleure compréhension des phénomènes biologiques liés à la division cellulaire, et donc au fonctionnement même de la cellule.

III.2. Applications médicales

L'emploi des méthodes de l'imagerie quantitative en cytopathologie est motivé par diverses nécessités :

- améliorer la sensibilité de l'analyse morphologique au microscope ; en effet notre système visuel a certaines limitations physiques et psychophysiologiques qui n'existent pas dans les instruments de mesures microscopiques (par exemple l'observateur ne détecte une différence de surface entre deux cellules que lorsqu'elle dépasse 20%).

- accroître l'objectivité et la reproductibilité de la lecture cytologique : la cytologie est une science cumulative dont les performances sont directement liées à l'apprentissage. Bien que d'importants efforts aient été réalisés pour unifier langage et description, il peut persister certaines imprécisions se traduisant par des divergences de diagnostic ou de classification entre cytologistes.

Le remplacement des critères subjectifs de classification par des mesures objectives devrait entraîner une uniformisation des diagnostics cytologiques et une grande reproductibilité entre cytologistes.

- compléter l'observation visuelle en donnant accès à des informations subvisuelles telles que la quantification de l'ADN, des ARN, de diverses protéines structurales ou fonctionnelles et de divers gènes, mesurés cellules à cellule ; la plupart des dosages biochimiques concernant les cellules sont fait sur des broyats cellulaires qui ne permettent pas de tenir compte de l'hétérogénéité tissulaire ou de la coexistence de différents clones au sein d'une même tumeur :

L'emploi des méthodes de l'analyse d'images microscopiques en cytologie pathologiques aboutit :

- à comprendre plus rationnellement les phénomènes physiologiques ou pathologiques traduits par leur image [MacAulay 1990b] : cinétique cellulaire, ploïdie, différenciation, malignité,

- à établir un diagnostic plus sûr du type cellulaire, de l'échantillon analysé ou le pronostic d'une affection.

Les applications possibles se situent donc au niveau

- du dépistage (exemples : cancers du col de l'utérus [van Driel-kulker 1986], cancers de vessie [Brugal 1986].

- du diagnostic et de la classification [Seigneurin 1984]. L'analyse des leucocytes par des paramètres texturaux est ainsi réalisée de façon automatique [Landeweerd 1981]

- et du pronostic : cancers du sein [Opfermann 1987 ; Auer 1980].

L'analyse quantitative de l'ADN demeure néanmoins l'application la plus répandue dans les laboratoires de cytopathologie [Atkin 1991]. C'est sans doute la seule application à être systématiquement utilisée pour l'aide (ou la confirmation) au diagnostic et au pronostic de certaines tumeurs cancéreuses : vessie, sein, utérus, prostate, etc ...

Chapitre II

Analyse de l'ADN : Principes et Limites

Tous les travaux concernant l'analyse de l'ADN par analyse d'images ou en flux concluent que cette technique devrait être utilisée comme un supplément aux diagnostics cyto-histopathologiques conventionnels dans le but de mieux appréhender le potentiel de malignité des néoplasmes [Bibbo 1985]. Néanmoins, depuis peu, des controverses très vives existent au sujet de la qualité du diagnostic et du pronostic et de la fiabilité de l'analyse de l'ADN d'échantillons tumoraux.

Il faut garder à l'esprit que la quantification de l'ADN est avant tout l'unique moyen, à ce jour "économiquement" possible en routine clinique, d'appréhender et de mesurer le statut ploïdique des cellules normales ou malignes. Le mesure du contenu en ADN ne permettra jamais d'identifier et de dénombrer les hétérosies, les anomalies structurales, les délétions chromosomiques, les amplifications géniques ou les translocations. Un contenu en ADN donné peut correspondre à de nombreuses situations ploïdiques différentes. Une cellule néoplasique peut en outre présenter un contenu en ADN normal.

L'interprétation d'une analyse d'ADN d'un échantillon d'un échantillon, doit donc tenir compte à la fois de cette limite intrinsèque mais également des nombreuses erreurs introduites à plusieurs niveaux de l'analyse cytométrique (du prélèvement à l'analyse des données).

I. Analyse de l'ADN : méthodologie et interprétation

Le pronostic des tumeurs, actuellement défini à l'aide de caractéristiques cliniques, histologiques et biochimiques, ne permet pas de réparer toutes les tumeurs particulièrement agressives. De fait, près de 20 % des malades présentant une tumeur du sein "à bon pronostic"

vont rechuter [Bieche 1992]. L'analyse de l'ADN des échantillons tumoraux fut pendant longtemps considéré comme un marqueur intéressant susceptible de reconnaître ces mauvais diagnostics. L'intérêt de l'analyse de l'ADN varie également considérablement selon le type de tissu : Les résultats les plus significatifs portent sur la vessie, la prostate, l'utérus et surtout le sein ; l'incidence du cancer du sein dans les pays développés explique les nombreuses études destinées à trouver des facteurs pronostiques et diagnostiques de plus en plus performants. Nous nous sommes donc principalement intéressés à ce tissu.

I.1. Intérêt de la quantification de l'ADN

Des études cytogénétiques suggèrent que les changements chromosomiques font partie intégrante de la cancérogénèse. En effet presque toutes les tumeurs malignes contiennent des aberrations chromosomiques structurales et/ou numériques. 60% environ des tumeurs mammaires malignes sont aneuploïdes [Fallienus 1988]. Cependant, même les tumeurs bénignes renferment des anomalies chromosomiques, mettant en évidence que la malignité provient de désordres plus complexes de l'information génétique. Dans le processus néoplasique, l'instabilité génétique initiale est suivie par des processus sélectifs qui mènent à la formation de clones cellulaires. Ces clones sont des populations cellulaires originaires d'une cellule unique [Nowell 1986]. Le terme "lignée souche" définit la constitution chromosomique la plus fréquente pour une tumeur donnée. Le "nombre modal" est le nombre de chromosomes le plus fréquemment rencontré dans la population cellulaire rencontrée. Ce nombre s'exprime par le terme de "ploidie". Les cellules normales possèdent un ensemble dïploide de chromosomes (2c) alors que les cellules malignes montrent des aberrations distinctes.

Analyse chromosomique	**Analyse de l'ADN par cytométrie**
Détection d'aberrations chromosomiques	Mesure du contenu en ADN
- Aberrations numériques (ploidie)	- estimation de la Ploïdie
- Hétérogénéité des lignées souches	- Hétérogénéité de lignées souches d'ADN
- Aberrations structuralles	- Analyse du cycle cellulaire
- "Chromosomes marqueurs"	
- Possible seulement sur 20% des tumeurs solides,	- Possible sur des spécimens frais, fixés ou noyés dans la parrafine
- Technique très longue	- Technique rapide

Tableau II.1 : Possibilités et limites de l'analyse chromosomique et de l'analyse de l'ADN par cytométrie.

L'identification et la désignation de clones cellulaires tumoraux sont possibles grâce à la fois à des analyses chromosomiques détaillée ou grâce à des analyses quantitatives de l'ADN par la cytophotométrie. Le tableau II.1 [Mellin 1990] compare ces deux techniques. Il est

important de distinguer le statut ploïdique et le contenu en ADN d'une cellule. Néanmoins, il est clair que l'intérêt principal de l'analyse de l'ADN par cytométrie réside dans la rapidité de cette technique qui en fait un outil clinique de plus en plus répandu. Par contre, la détection des aberrations chromosomiques numériques ou structurales est encore irréalisable en routine clinique.

I.2 Méthodologie

L"analyse de l'ADN par cytométrie a fait l'objet de nombreux ouvrages [Falkmer 1989] [Fallienus 1986]. Les différentes étapes de l'analyse de l'ADN sont les suivantes :

I.2.1. Préparation du spécimen

En fonction du tissu étudié, le prélèvement et la préparation sont différents. La nécessité de sécher très rapidement les échantillons après le prélèvement a été démontrée. Une post-fixation avec une solution à 4% de paraformaldéhyde ou avec un mélange FAA (80% méthanol, 15% paraformaldéhyde, 5% acide acétique) est conseillée [Giroud 1988]. La coloration est en général faite par la réaction de Feulgen [Feulgen 1924], qui se lie de façon stochiométrique avec l'ADN. Actuellement, en routine clinique, le protocole de coloration préconisé est le suivant :

- Hydrolyse acide à 20°C avec de l'acide chlorydrique 6N pendant 1 heure,

-Coloration par le réactif de schiff soit à la paraosaniline (coloration rose des groupements aldéhydes) soit à la thionine (coloration bleue) pendant 45 minutes [Schulte 1990].

Une fois la réaction de Feulgen effectuée, la préparation microscopique peut être analysée. Par balayage de la lame, la densité optique de la coloration de l'ADN est mesurée en chaque point du noyau des cellules repérées. Ces valeurs sont ensuite sommées donnant ainsi une valeur de densité optique intégrée (DOI), corrélée avec la quantité en ADN présente dans le noyau.

I.2.2. Calibration

La calibration s'effectue comme suit :

$$DOI(x)\text{recalé} = \frac{DOI(x)}{DOI(\text{témoin}2c) * f} * 2c$$

avec DOI(x) : la valeur densitométrique brute

f : facteur de correction.

La multiplication par 2c, facultative, permet l'expression de la quantité en ADN en "c" (c pour chromatides).

I.2.3. Analyse et représentation

Les valeurs des cellules analysées sont stockées dan un fichier. La représentation graphique de ces valeurs se fait par l'intermédiaire d'un histogramme représentant la répartition

des cellules dans différentes classes de quantité d'ADN. Cet histogramme d'ADN permet d'approcher l'état "ploidique" d'un échantillon tumorale [Atkin 1987 ; Hedley 1989].

I.3.Interprétation des histogrammes d'ADN

Contrairement à la cytométrie en flux où les histogrammes d'ADN sont classés simplement en groupe aneuploïde *versus* diploïde [Wersto 1991], l'interprétation des histogrammes d'ADN obtenus par analyse d'images a suscité de très nombreux travaux. De fait, il existe sans doute une vingtaine de paramètres différents décrivant ces histogrammes : nous ne retiendrons que ceux qui sont les plus fréquemment utilisés dans la littérature :

I.3.1. Interprétations visuelles

Classification d'AUER

Ce fut la première classification a but pronostique proposée à partir de série d'échantillons tumoraux de cancers du sein. La méthode d'Auer permet de distinguer, par observation visuelle, 4 types d'histogrammes, représentés sur la Figure II.1, correspondant à 4 groupes pronostiques [Auer 1980, 1984 ; Ono 1983] :

Type I : un seul mode de contenu en ADN dans la région diploïde. Seul un petit nombre de cellules peut dévier de la valeur normale.

Type II : Soit un mode dans la région tétraploïde, soit deux pics dans les régions 2c et 4c. Présence de peu de cellules dans la région entre 2c et 4c correspondant à la phase de synthèse d'ADN des cellules diploïdes. Seul un un petit nombre de cellules peut se trouver en dehors du domaine des valeurs d'ADN d'une population normale (en général dans la région 8c).

Type III : se distingue des populations de type II par un nombre important de cellules ayant un contenu en ADN intermédiaire entre 2c et 4c. La proportions de ces cellules permet d'évaluer l'activité proliférative.

Type IV: aneuploïdie prononcée et irrégulière avec des contenus en ADN s'étendant de 2c jusqu'à 6c et 8c

Classification en euploïde *versus* aneuploïde

Une tumeur est définie comme euploïde quand la population principale est située dans les régions diploïde, tetraploïde ou octaploïde ainsi que dans toutes les régions correspondant à un multiple entier d'une "unité " de ploïdie, correspondant à un lot entier de chromosomes (23 chez l'homme). Une tumeur triploïde est donc une tumeur aneuploïde.

On parlera de tumeurs aneuploïdes dans tous les autres cas qui correspondent à des variations portant sur une partie seulement d'un lot de chromosomes (exemple : population dans la région 2.5c).

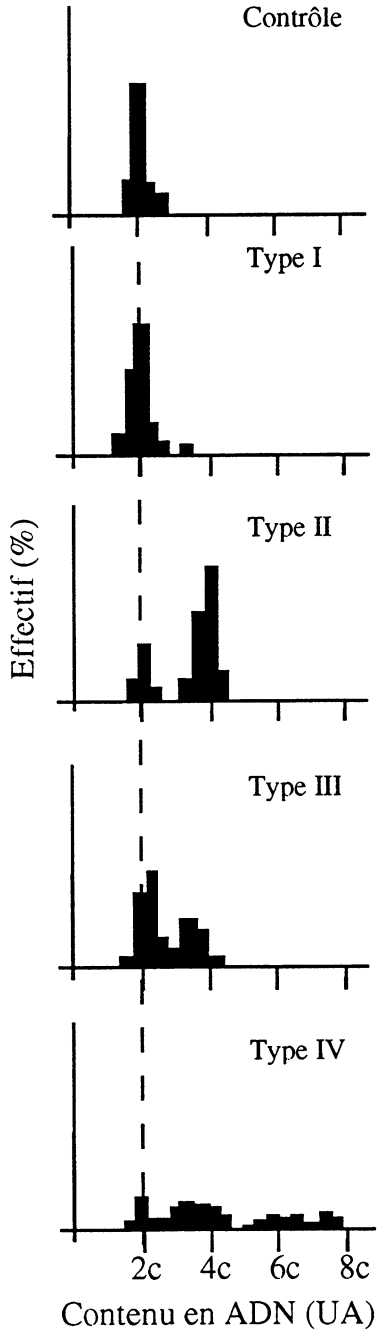


Figure II.1 : Histogramme normal d'un tissu non-proliférant (contrôle) et classification en 4 types des histogrammes d'ADN à partir d'échantillons tumoraux.

Classification en diploïde, aneuploïde, et tétraploïde

Cette classification est motivée par l'observation que les tumeurs tétraploïdes étaient souvent d'un bon pronostic. Par contre les tumeurs aneuploïdes ont un mauvais pronostic. Les tumeurs diploïdes sont en général de bon pronostic : la majorité des tumeurs bénignes sont diploïdes.

I.3.2. Descripteurs quantitatifs des histogrammes d'ADN

Pour pallier la subjectivité des classifications visuelles, des descripteurs numériques ont été proposés pour mesurer la ploïdie tumorale :

Proportions de cellules ayant un contenu en ADN supérieur à une valeur d'ADN donnée

Cette valeur seuil peut varier d'un laboratoire à un autre. Néanmoins, certaines valeurs sont plus utilisées que les autres :

+ Proportion de cellules ayant un contenu en ADN supérieur à 2.5c (2.5c-ER pour 2.5c Exceeding Rate)

Cette valeur peut correspondre à des situations très différentes.

+ Proportion de cellules ayant un contenu en ADN supérieur à 5c (5c-ER pour 5c Exceeding Rate)

Une valeur non-nulle correspond forcément à la présence de cellules n'appartenant pas à une population diploïde (comprise entre 2c et 4c). [Böcking 1984]

+ Proportion de cellules ayant un contenu en ADN supérieur à 8c (8c-ER pour 8c Exceeding Rate)

Cette valeur atteste de la présence de cellules fortement hyperploïdes ou aneuploïdes mais est très rarement différente de 0.

Indices d'ADN

Index d'ADN moyen

Défini comme le rapport de la moyenne de la quantité d'ADN des cellules analysées sur la valeur du témoin 2c.

Index d'ADN du pic modal ou Index de ploïdie

Défini comme le rapport de la valeur de la classe modale du plus grand pic de la distribution sur la valeur du témoin 2c.

Index d'ADN de tous les pics

On peur définir comme précédemment un index pour chaque pic d'une distribution multimodale.

Index de déviation du 2c de Böcking (2cDI)

Cet index représente la distance moyenne de la quantité d'ADN de la population cellulaire par rapport à la valeur de référence 2c [Böcking 1984 ; Auffermann 1987]. il correspond à l'écart-type de la population par rapport à la valeur 2c.

$$2cDI = SQR\ [\ (1/n)SUM(Ci - 2c)^2]$$

avec n = nombre de cellules
Ci = Quantité d'ADN de la cellule i
2c = Quantité d'ADN du témoin 2c.

Balance de Ploïdie et Index de Prolifération d'Opfermann [Opfermann 1987]

Ces deux paramètres sont calculés à partir d'un histogramme codé en 10 classes centrées sur des valeurs de ploïdie pré-définies : 2c, 2.5c, 3c, 3.5c, 4c, 5c, 6c, 7c, 8c, > 8c.

La balance de ploïdie est la différence du pourcentage de cellules dans les classes euploïdes (2c, 4c, 8c) et du pourcentage de cellules dans les autres classes. Ce paramètre varie de -100% (toutes les cellules sont aneuploïdes) à +100% (toutes les cellules sont euploïdes).
L'Index de Prolifération est le pourcentage de cellules situé dans des classes autres que la classe modale, les deux classes adjacentes, les classes multiples de 2 de la classe modale et les classes adjacentes à ces classes.
Ces deux indices ont l'avantage de décrire les deux critères principaux de malignité d'un échantillon tumoral : la prolifération et l'aneuploïdie. La projection des tumeurs, en fonction des valeurs de PB et de IP sur un plan engendré par ces deux paramètres a permis à Opfermann de définir un triangle d'aggressivité. En effet tous les échantillons contenus dans ce triangle étaient de mauvais pronostic.

Grade de malignité basé sur l'ADN selon Böcking (DNA-MG)

Ce paramètre, comme son nom l'indique, permet de discriminer, en fonction d'une valeur seuil, les tumeurs de bon pronostic des tumeurs de mauvais pronostic [Böcking 1989a et b] :

$$\text{DNA-MG} = 3 \times \lg(2\text{cDI} + 1) / \lg 51 = 1.757 \times \lg(2\text{cDI} + 1)$$

Entropie de la distribution [Senkvist 1990].

L'entropie mesure l'hétérogénéité de la répartition des fréquences de la distribution. La valeur est maximale quand toutes les classes ont la même amplitude. Une division de la valeur par le logarithme du nombre de classes rend l'entropie indépendante du nombre de classes :

$$\text{Entropie} = \frac{\sum_{i=1}^{i=k} p_i \times \log2(p_i)}{\log2\ k}$$

avec k = nombre de classes
p(i) = probabilité ou fréquence de la classe i

La méthode de Weber [Weber 1987]

Elle consiste à utiliser les déciles de la distribution, après construction de la fonction de distribution, pour décrire les histogrammes d'ADN. Sur la base de ces nouveaux descripteurs de la distribution, Weber a pu discriminer des échantillons de bon et de mauvais pronostics

II. Les limites et les problèmes le la quantification de l'ADN

Les problèmes liés à l'analyse de l'ADN par imagerie sont communs, excepté au niveau de l'interprétation même des histogrammes, à de nombreuses autres applications cytométriques.

II.1. Les problèmes d'échantillonnage

Que ce soit en flux ou en imagerie, la qualité des analyses obtenues dépendent énormément de la qualité du matériel cellulaire disponible. Cette qualité est évidemment dépendante du tissu étudié et de la technique de prélèvement employée. Pour les tumeurs solides, comme le sein ou la prostate, les problèmes rencontrés sont semblables.

II.1.1 La représentativité des échantillons

Un échantillon tumoral peut être obtenu de diverses façons. Les cellules malignes proviennent parfois de certaines muqueuses, comme l'urine où les cellules tumorales exfoliées montrent certains états de dégénérescences. C'est un exemple typique où le mode de prélèvement affecte certainement les mesures quantitatives.

Les techniques d'aspirations par aiguilles fines (FNAB) sont devenues une technique de plus en plus importante pour le diagnostic cytopathologique. Les frottis obtenus à partir de ces spécimens présentent aussi bien des cellules isolées que des cellules en petits "cluster". Ces échantillons peuvent être considérés comme intermédiaires entre la cytologie traditionnelle et les préparations histologiques. Les premières utilisations de cette technique furent dédiées aux études quantitatives de certaines tumeurs solides comme le cancer du sein [Auer 1980]

De nombreuses études sont basées également sur des sections de tissus imbibées de parraffine, qui offrent de multiples avantages : différentes aires peuvent être analysées et corrélées avec d'autres paramètres morphologiques. De plus, ce matériel permet des études rétrospectives sur la survie des malades opérés d'un cancer.

Les sections histologiques posent également de nombreux autres problèmes techniques ; ainsi contrairement aux cellules obtenues à partir de fluides corporels, d'aspiration par aiguilles ou d'empreintes, qui sont généralement intactes, les spécimens histologiques offrent souvent des cellules sectionnées par le microtome. Cette technique ne sera pas développée ici.

Il reste que quelque soit la technique employée, la représentativité d'un échantillon de quelques centaines de cellules peut être sérieusement mise en cause. Une grande partie des tumeurs est biologiquement hétérogène. La présence de clones cellulaires dans les tumeurs solides est maintenant bien admise. On peut se demander alors quelle est la justification que l'on peut donner sur un diagnostic fait à partir de l'analyse de 200 ou 300 cellules alors qu'une tumeur en contient plusieurs milliards. De plus, même si la ponction par aiguille fine apparait comme la technique la mieux adaptée car elle permet d'aspirer différentes régions dans le nodule, rien ne permet de dire que toutes les sous populations présentes dans la tumeur sont correctement représentées dans l'échantillon [Falkmer 1992]. Ferno [Ferno 1992] a montré que, en cytométrie en flux, que d'importantes différences existaient au niveau de la ploidie et de la fraction de cellules en phase S entre différentes parties d'une même tumeur du sein. L'auteur conclu que ces variations sont plus dus aux difficultés d'interprétation des histogrammes d'ADN plutôt qu'à une réelle hétérogénéité en ADN de la tumeur. Encore une fois, la majorité des erreurs proviennent de la détection de populations faiblement aneuploïdes qui sont parfois considérées comme "near-diploïde". Ces travaux réalisés en flux ont les mêmes résonances en

imagerie. Cependant, de nombreux travaux ont montré que plusieurs ponctions étaient nécessaires pour détecter la présence de lignées aneuploïdes dans une tumeur. Sasaki [Sasaki 1992] a mis en évidence parmi 9 cas de tumeur de la vessie, deux cas où la tumeur était constituée de mélange de deux sous-popupations diploïdes et aneuploïdes, localisées à différents endroits de la tumeur. Dans ce cas, la détection des lignées aneuploïdes implique l'examen de plus d'un échantillon de la tumeur. Beerman [Beerman 1991] indique également qu'il faut plusieurs échantillons de différentes parties de la tumeur pour détecter des sous-populations, oubliées dans un seul échantillon. Il ajoute que plusieurs échantillonnages augmentent l'incidence des tumeurs aneuploïdes allant même jusqu'à préciser que 4 échantillons sont nécessaires pour déterminer de façon fiable le statut ploïdique dans le cancer du sein. Pennes [Pennes 1990], en extrapolant ses résultats, indique que 9 ou 10 aspirations de chaque lésion seraient nécessaires pour atteindre 100% de sensitivité pour le diagnostic d'un cancer. Williams [Williams 1987] avait montré également, pour le cancer du poumon que l'augmentation de la taille de l'échantillon augmentait la précision du diagnostic. D'après Meyer [Meyer 1991], 10% à 20% des erreurs de pronostics peut être expliquée par l'hétérogénéité des carcinomes mammaires.

La détermination d'un nombre d'échantillons nécessaires à prélever pour être sûr à 100% de la représentativité de n'a pas de sens. Nous serions en présence d'un problème classique d'inférence statistique si le prélèvement consistait à tirer aléatoirement n cellules parmi les N cellules de la tumeur. Or nous sommes confrontés à des "tirages" non-aléatoires dans une population biologiquement et topographiquement hétérogène.

Contrairement à ce qui était dit auparavant [Vindelov 1986], l'hétérogénéité intratumorale est beaucoup plus fréquente que l'on ne le supposait auparavant [Sasaki 1992] [Carey 1990] [Vindelov 1980]. De plus certains auteurs ont montré que des différences existaient entre les mesures d'ADN obtenues sur des ponctions ou sur des empreintes [Salmon 1991] [Spyratos 1987].

Les travaux de Mesker [Mesker 1991], sur le cancer du sein, mettent bien en évidence tous ces problèmes d'échantillonnage. Le but de ces travaux était d'accroître la détection d'un faible pourcentage de cellules aneuploïdes par un système d'analyse automatisé. Pour cela, Mesker a utilisé deux techniques différentes dont l'une concerne la technique de la préparation cellulaire.

La figure II.2 montre deux histogrammes d'ADN : le premier correspond à un histogramme obtenu à partir d'une sélection aléatoire des cellules d'une section entière de 50 µm de tissu. Cet histogramme est composé presque entièrement de cellules diploïdes. Parallèlement, les auteurs ont inspectés et sélectionnées visuellement certaines aires de 5 µm. L'histogramme d'ADN correspondant montre une lignée cellulaire indépendante, située dans la région hypotétraploïde, loin d'être évidente dans l'analyse aléatoire d'une section entière.

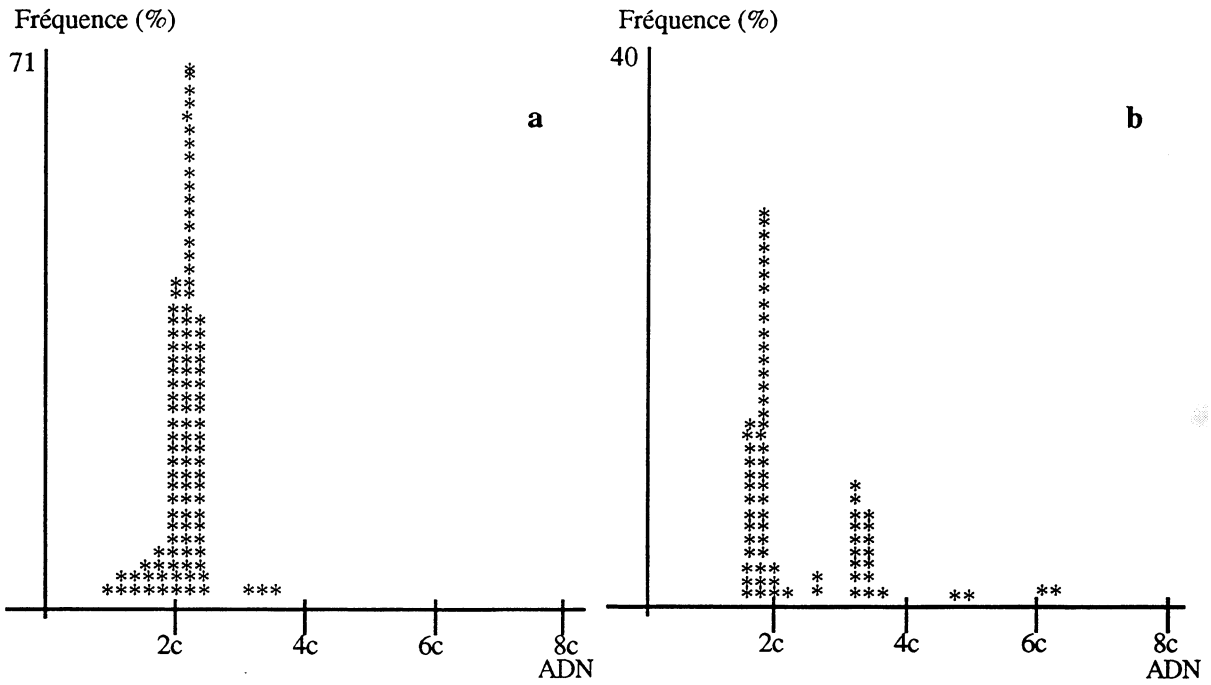


Figure II.2 : Histogrammes d'ADN d'adénomes du sein obtenus par analyse d'images :
a) Sélection aléatoire des cellules à partir d'une section de 50µm
b) Sélection visuelle des cellules à partir de sections de 5µm.

II.2. Les problèmes liés à la préparation biologique

Les différentes étapes de la préparation du matériel biologique apportent elles aussi un certain nombre de biais qui influent à un certain niveau sur la qualité de l'analyse [Goss 1992].

II.3. Les problèmes liés à l'analyse d'images

II.3.1. L'instrumentation

Les différentes erreurs liées à l'instrumentation sont relativement connues. Leur maîtrise et leur correction restent malgré tout très délicates. En ce qui concernent les caméras, plusieurs types d'erreurs peuvent apparaître.

Les caméras à tube souffrent de divers effets : période d'exposition variable pour chaque pixel, sensibilité aux vibrations et aux champs électromagnétiques, rémanence et éblouissement. L'ensemble de ces facteurs défavorables se traduit par une diminution de la précision des mesures densitométriques. Ces limitations des caméras à tube dans le domaine de la mesure densitométrique par comparaison avec les caméras CCD font que ces dernières sont sans doute mieux adaptées pour traiter les applications densitométriques [McEachron 1990]. Ce choix des caméras CCD sur la base de la précision de la mesure densitométrique se trouve conforté par le fait que ces caméras présentent une meilleure sensibilité. En effet, l'expérimentation qui consiste à visualiser la variation des niveaux de gris sur une ligne image comportant un objet est

un test qui permet d'apprécier le degré de sensibilité des caméras. Le profil densitométrique obtenu avec les caméras CCD permet également de localiser un objet avec une précision supérieure au profil obtenu sur des caméras vidéo [Hiraoka 1987]. Cette observation est à mettre en relation non seulement avec la meilleure sensibilité des caméras CCD par rapport aux caméras à tube, mais également avec leur plus grande dynamique.

Les erreurs instrumentales incluent :

(a) les erreurs de distribution dues à l'usage d'un spot lumineux trop large par rapport au champ mesuré ou dues à la mauvaise focalisation des objets. Pour exploiter dans la limite du pouvoir de leur pouvoir de résolution, le diamètre des spots à mesurer ne doit pas dépasser 0.125mm. En pratique cette taille n'est pas respectée car elle délivre une quantité de lumière trop faible et les spots utilisés peuvent atteindre la taille de 1mm. L'erreur de distribution ainsi produite est approximativement proportionnelle au diamètre du spot [Goldstein 1981].

(b) les effets de glare dus aux multiples réflexions dans les parties optiques du microscope. Pour contrebalancer les effets de 'glare', il existe deux grandes méthodes :

- l'une consiste à le minimiser au maximum en nettoyant avec soin toutes les parties optiques, en limitant les interfaces air-verres par l'emploi d'huile à immersion et en fermant le plus possible le diaphragme des objectifs à iris,

- l'autre consiste en une compensation par des moyens électroniques [Goldstein 1981].

Certains auteurs ont élaboré de grandes équations pour corriger les mesures densitométriques du 'glare' en se basant sur le fait que le 'glare' dépend de la forme, de l'absorbance et de la taille de l'objet en relation avec la taille du champ éclairé. Mais ces méthodes ayant été établies pour un système donné sont lourdes à appliquer [Duijndam 1980]

(c) les erreurs chromatiques dues à l'utilisation d'une erreur polychromatique. Le matériel n'absorbe pas de la même façon des longueurs d'ondes différentes bien que proches et des erreurs peuvent être dues à l'emploi d'une lampe insuffisamment monochromatique.

II.3.2. Les problèmes en Reconnaissance des Formes

De nombreuses recherches sont menées par des mathématiciens et des informaticiens dans le but d'améliorer les méthodes de reconnaissances de formes. Des outils de plus en plus sophistiqués existent permettant d'optimiser les différentes techniques de segmentation et de paramétrisation des objets, en mode semi-automatique ou automatique. Les erreurs de segmentation peuvent aussi avoir de grandes conséquences sur le diagnostic final.

Un des dangers de ce fonctionnement est la prise en compte d'objets qui ne sont pas les cellules désirées, par exemple des cellules non urothéliales ou des cellules dégénérées dans un lavage de vessie.

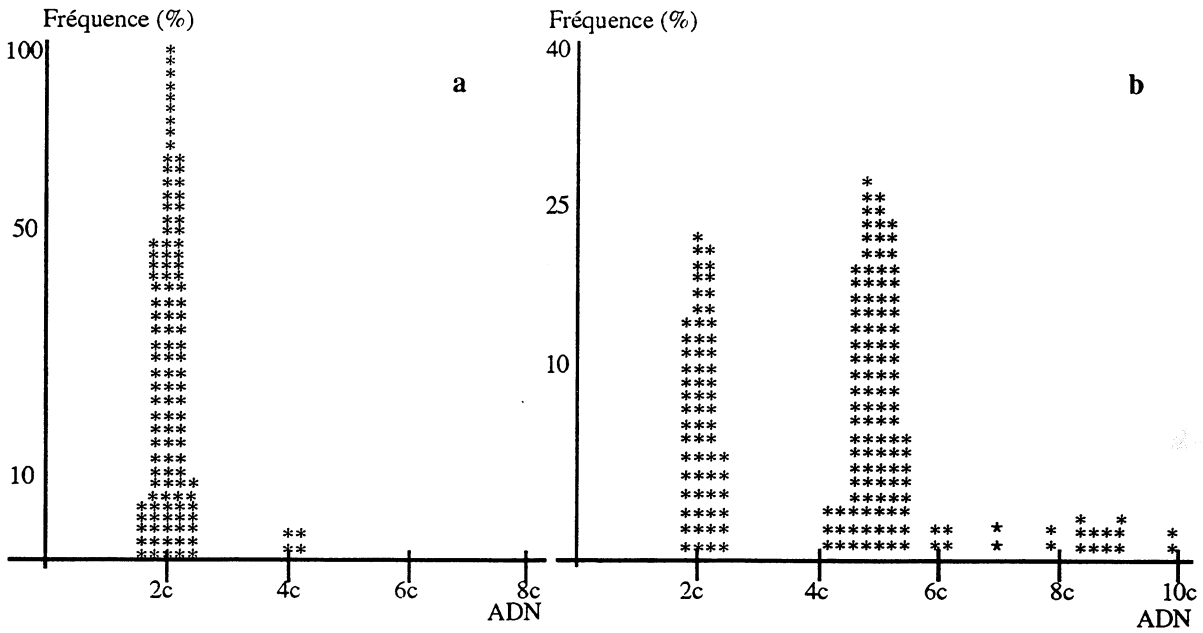


Figure II.3 : Histogrammes d'ADN obtenus à partir d'un carcinome du sein.
a) Sélection aléatoire des cellules : l'histogramme est considéré comme diploïde
b) Sélection des cellules en fonction de leur taille et de l'intensité du marquage. Le nombre relatif de cellules diploïdes a diminué et une population aneuploïde importante apparait vers 5c.

II.3.3. Méthode d'acquisition sélective

Mesker [Mesker 1989] a montré qu'une acquisition sélective des cellules permettait de détecter des sous-populations cellulaires aneuploïdes. Les histogrammes de la Figure II.3a et II.3b ont été obtenus sur le même échantillon : dans le premier cas, l'acquisition des cellules était aléatoire alors que dans le deuxième cas, les cellules ont été sélectionnés en fonction de leur valeur en taille et en intensité de la DOI (définition de seuils). Les deux histogrammes donnent lieu à des diagnostics et des pronostics différents. Cette méthode sélective permet donc de détecter des lignées cellulaires aneuploïdes dans les cas où la population diploïde normale est fortement dominante. Cette méthode peut offrir une solution partielle au problème d'échantillonnage du spécimen étudié.

II.3.4. Rejet des artéfacts

De nombreuses méthodes ont été développées pour rejeter les artefacts lors de l'analyse de préparations biologiques. Une des plus utilisées consiste à éliminer les objets dont les caractéristiques sont en dehors des limites prédéfinies. Mais cette technique n'est pas toujours très performante, en particulier pour distinguer les doublets cellulaires. Des algorithmes plus complexes ont été développés pour résoudre ces problèmes [Tucker 1979 ; Sychra 1978 ; Dytch1983].

II.4. Les problèmes d'analyse de données

II.4.1 Taille de l'échantillon

Les problèmes de tests d'arrêt, quelque soit les domaines d'applications, ont souvent comme fondement des raisons économiques et financières. Une technique, aussi intéressante et performante soit-elle, ne pourra être développée en routine clinique si son coût est trop élevé. Il n'est pas de notre ressort de discuter ce fait. Toujours est-il que le développement des systèmes d'analyse d'images impose un rendement suffisamment élevé. Il est donc tout naturel de déterminer à partir de quel moment une analyse cytométrique, en densitométrie par exemple, peut être stoppée. Ce problème de test d'arrêt est étroitement lié au problème de la taille de l'échantillon requise pour optimiser le diagnostic. En effet, il faut trouver un compromis entre le temps de l'analyse et la qualité voir la fiabilité du diagnostic effectué. Ce problème est réellement spécifique de la cytométrie à balayage. A ma connaissance, aucune recherche n'a jusqu'alors abordé ce problème.

II.4.2 Détection d'individus rares

La détection d'évènements rares dans une population cellulaire est un des aspects très important de la cytométrie. Par rapport à l'analyse d'images, la cytométrie en flux parait beaucoup plus efficace pour détecter des sous-populations faiblement représentées.

Le nombre important de cellules analysées, conjugué à un coefficient de variation relativement faible, permet par exemple d'obtenir des sensibilités de l'ordre de 0.001% [Visser 1986]. La détection de cellules tumorales métastasiques dans la moelle osseuse [Dantas 1983] ou de cellules monoclonales dans le sang périphérique de patients [Ault 1979], la quantification de réticulocytes dans le sang périphérique pour déterminer la prolifération hématopoiétique [Tanke 1983] sont des applications potentielles de la cytométrie en flux. Certains auteurs [Bhattacharya 1971,1973] ont proposé un certain nombre de techniques, basées sur l'approximation des histogrammes par des mélanges de lois théoriques, en vue de détecter des sous-populations minoritaires (lignées aneuploïdes) Le peu de cellules analysées en imagerie ne permet pas de transposer ce type d'approche sur des histogrammes construits à partir de 200 ou 300 cellules.

L'imagerie quantitative peut, en théorie, offrir les moyens de détecter des individus rares. En théorie seulement, car il est clair que parmi un échantillon de 200 ou 300 cellules, on ne peut plus parler d'individus rares. Par contre, il est toujours possible de détecter un individu anormal ou "extraordinaire". Il est toutefois clair que la démarche à suivre est différente selon l'application ; en effet, il y a deux approches :

- soit on cherche une population minoritaire spécifique, et dans ce cas on sélectionne les cellules en fonction d'un critère donné,
- soit on n'en cherche pas et à ce moment là la détection d'individus "anormaux" devient beaucoup plus délicate.

L'analyse de l'ADN reste de nos jours l'application la plus répandue et la plus étudiée en cytométrie en flux comme en imagerie. Contrairement à certaines applications qui cherchent à mesurer les proportions de certaines catégories cellulaires (exemple les différents types de cellules sanguines), l'analyse de l'ADN impose de connaître le plus précisément possible la relation entre le contenu d'ADN mesuré (donné en unité arbitraire) et le statut ploïdique des cellules analysées.

II.5. Les problèmes spécifiques à l'analyse de l'ADN

Contrairement aux applications qui cherchent simplement à discriminer des sous-populations cellulaires dans un échantillon, l'analyse de l'ADN nécessite de connaître la correspondance entre la valeur de Densité optique donnée par le système d'imagerie et la quantité d'ADN correspondante. Un étape de calibration est donc indispensable.

II.5.1. La calibration

L'aspect exemplaire de l'analyse de l'ADN réside dans le fait que tous les problèmes précédemment cités sont dans cette application précise particulièrement évident.

En colorant l'ADN par des techniques de coloration comme le réactif de Feulgen, une analyse par transmission permet de mesurer la quantité d'ADN présente dans les noyaux des cellules étudiées.

Un des facteurs d'incertitude les plus fréquents est la calibration. En effet, la densité optique intégré est un paramètre brut qui doit être ajusté afin de rendre comparable les analyses d'échantillons différents. De plus, cette valeur brute doit être recalée à l'aide de la DOI de cellules ayant une teneur en ADN connue au préalable. Pour permettre cette calibration, des standards doivent être utilisés. Ces standards permettent de mieux appréhender les variations des valeurs densitométriques observées entre cellules du même type, colorées et analysées à différents instants. L'utilisation de ces standards est indispensable pour rendre possible les comparaisons d'une analyse à l'autre au sein d'un même laboratoire et entre laboratoires. Il existe deux grandes classes de standards : les standards internes si ce sont des cellules prises sur un même échantillon et les standards externes si ce sont des cellules provenant d'autres sources biologiques.

Les standard internes

En général, ce sont des cellules non tumorales mais présentes dans l'échantillon analysé, comme les cellules du stroma, les cellules épithéliales, ou surtout les cellules sanguines en particulier les lymphocytes qui ont théoriquement la même quantité d'ADN que les cellules normales de l'échantillon. Opfermann [Opfermann 1987] a montré d'ailleurs que la variation inter-lames des valeurs densitométriques calculées sur des lymphocytes d'un même échantillon était relativement faible (2 à 7 % de variation). Néanmoins, à la décharge des lymphocytes on sait que leur valeur densitométrique est plus faible que celle des autres types de cellules

diploïdes ayant potentiellement le même contenu en ADN [Chatelein 1989]. De plus, leur fréquence peu élevée dans une préparation biologique et la difficulté de les reconnaître rend leur utilisation parfois limitée.

Les standard externes

Contrairement aux références internes, les références externes sont des cellules d'une autre espèce que celle du prélèvement, par exemple, du foie de rat pour l'analyse d'un échantillon humain. Les hépatocytes de foie de souris ou de rat, les érythrocytes de poulet ou de truites sont les références les plus utilisées. Les érythrocytes de poulet ou de truite sont faciles à obtenir en grande quantité et présentent des populations cellulaires d'apparence homogène, mais ils sont de petite taille et possèdent un contenu en ADN bien inférieur aux cellules humaines [Vindelov 1986]. Par contre les hépatocytes de souris ou de rat présentent une surface nucléaire et un contenu en ADN très proches de l'espèce humaine. De plus les empreintes de foie de rat offrent l'avantage de contenir une population de noyaux cellulaires mixtes voir triples (suivant l'âge de l'animal), comptant des noyaux diploïdes (quantité d'ADN 2c), des noyaux tétraploïdes (quantité d'ADN 4c) et des noyaux octoploïdes (quantité d'ADN 8c). Ces trois types de noyaux permettent le contrôle de la linéarité de la réponse densitométrique du système d'analyse d'images.

Procédure de calibration

La taille de l'échantillon de référence doit être suffisamment importante pour que la valeur choisie (moyenne ou médiane dans le cas d'une population gaussienne) soit un bon estimateur de la "vraie valeur". Ainsi comme l'a montré Wied [Wied 1988] sur un échantillon de 20 cellules de référence diploïde présentant un coefficient de variation de 10%, la valeur 2c est définie seulement à +/- 0.1c avec une probabilité de 0. 5.

L'importance de la qualité de la référence diploïde et du choix de la procédure de calibration est parfaitement démontrée par Kiss et ses collaborateurs [Kiss 1992].

II.5.2. Les problèmes de compaction de la chromatines

De nombreux travaux ont montré que la quantité d'ADN mesurée pouvait varier entre deux noyaux ayant la même teneur en ADN, mais ayant une structure différente ou appartenant à des stades différents. Giroud [Giroud 1982] a démontré que la teneur en ADN des cellules en métaphase était inférieur à celle des cellules en anaphase. Giroud [Giroud 1988], encore, a mis en évidence que l'histogramme d'ADN de cellules érythroblastiques variait considérablement en fonction de leur état de maturation (histogramme diploïde unimodal jusqu'à un histogramme multimodal allant de 2c à 4c). Ceci met en évidence les précautions à prendre dans la calibration à partir d'érythrocytes de poulet par exemple. Ces erreurs de mesures sont dues en particulier à deux phénomènes ; tout d'abord, en fonction de l'état de condensation de l'ADN le colorant sera plus ou moins accessible à toute la molécule d'ADN. Enfin, si le noyau est trop condensé, la loi de Berr-Lambert, reliant la quantité de signal reçu par rapport au signal émis, peut ne plus être totalement vérifiée [Schulte 90].

II.5.3. Le problème de l'interprétation des histogrammes d'ADN

Un très grand nombre de publications et de travaux, que ce soit en flux ou en imagerie [Auer 19804 ; Atkin 1979] ont démontré l'intérêt diagnostique et pronostique de l'analyse de l'ADN d'échantillons tumoraux.

La corrélation entre le contenu en ADN des échantillons et des paramètres cliniques, comme les grades histo- et cytologiques ou le statut ganglionnaire a été largement démontré. Néanmoins, la valeur diagnostique et pronostique du contenu en ADN reste parfois contradictoire et varie selon la nature des tissus, la technique employée (cytométrie en flux ou imagerie), et l'approche statistique utilisée. Toutes les études montent que les tumeurs aneuploïdes sont globalement de plus mauvais pronostic les tumeurs diploïdes. Dans certaines tumeurs, les échantillons tétraploïdes se sont même avérés de meilleur pronostic que les tumeurs diploïdes.

La difficulté d'interprétation des histogrammes d'ADN, en cytométrie en flux comme en analyse d'images, explique en partie les résultats contradictoires trouvés dans la littérature.
Les problèmes d'interprétation des histogrammes d'ADN sont différents si l'on travaille en flux ou en image. Dans le cas de la cytométrie en flux, le nombre élevé de cellules analysées permet une approche plus analytique, en modélisant l'histogramme par des distributions connues (loi normale, mésa-normale, etc..). La classification aneuploïde versus euploïde est la plus couramment utilisée. Une très bonne revue sur ce problème d'interprétation des histogrammes d'ADN a été effectuée par Wersto [Wersto 1991].

Les histogrammes obtenus par analyse d'images posent de nombreux problèmes d'interprétation qui seront largement repris dans la suite de cette thèse [Bartels 1985]. Dans le chapitre suivant, deux exemples illustreront les limites de l'analyse de l'ADN comme facteurs diagnostique et pronostique du cancer.

ANALYSE DE l'ADN en CANCEROLOGIE
Qualité du Diagnostic, du Pronostic et de l'analyse

Les limitations du système oeil-cerveau humain aboutissent à un taux d'erreur significatif dans le diagnostic tumoral. Ainsi, dans le dépistage des cancers cervicaux, 20% de faux négatifs et 10% de faux positifs est considéré comme un taux d'erreur moyen. Ces erreurs résultent des erreurs de l'interprétation des images cellulaires (la moitié des faux négatifs) mais aussi de la mauvaise qualité des spécimens analysées (l'autre moitié). Pour les autres organes, le taux d'erreur est approximativement le même.

Les méthodes quantitatives semblent donc à même d'améliorer la reproductibilité et la fiabilité des diagnostics. Malheureusement, définir une corrélation entre les mesures observées et la vérité demeure impossible puisque par définition il n'y a pas de vérité. Même la survie des malades ne peut être considérée comme la "référence" pour définir la qualité du pronostic puisque que la survie d'un malade ne dépend pas seulement de la nature de la tumeur mais de nombreux autres facteurs individuels : génotype, phénotype, psychologie, environnement, éducation du patient ainsi que de la stratégie thérapeutique choisie.

Comme dans tous les domaines où les progrès techniques et technologiques sont très rapides, le danger existe de s'affranchir des limites de la méthodologie et de considérer les données comme exactes et absolues. En recherche clinique, ce problème s'avère évidement beaucoup plus important qu'en recherche fondamentale.

Les résultats exposés dans ce chapitre illustrent les problèmes qui se posent au niveau de l'interprétation des histogrammes d'ADN tant au niveau du diagnostic qu'au niveau du pronostic tumoral.

I. Analyse de l'ADN et diagnostic des cancers

Au deuxième congrès de l'ESACP (Nijmegen, Mai 1992) un workshop intitulé " On histogram Analysis and Ploidy Classification" fut organisé par le Pr. Albrecht Reith. Ce workshop avait pour but de comparer les résultats des classifications d'histogrammes d'ADN et du diagnostic obtenus par 4 laboratoires européens, spécialistes en analyse d'image et en particulier en analyse de l'ADN.

I.1. Méthodologie

Sept échantillons tumoraux, 4 cancers de la prostate et 3 cancers gynécologiques, ont été analysés par analyse d'images dans le laboratoire du Pr.Reith. Les données des références (lymphocytes) et les données brutes ont été ensuite envoyées à chaque laboratoire. La Figure III.1 représente les histogrammes des échantillons de références. La moyenne et le coefficient de variation de ces échantillons ont été calculés et sont donnés pour chacun d'eux.

Voici les caractéristiques concernant l'analyse des échantillons tumoraux:

Cancers de la prostate : Blocs de paraffine, monocouche, coloration de Feulgen, Système d'analyse d'Image Kontron Ibas system, Pathologiste expérimenté

Cancers de l'utérus : Biopsie fraîches, monocouche, coloration de Feulgen, Système d'Analyse d'Image Kontron Ibas, Pathologiste expérimenté.

Chaque laboratoire devait établir un diagnostic en expliquant et en illustrant la démarche suivie. La démarche de notre équipe (laboratoire II) s'est faite en trois étapes;

- Recalage des histogrammes (facteur de correction pour les lymphocytes = 1.15),
- Calcul de très nombreux paramètres (2cDI, Entropie, DNA-MG, etc...),
- Diagnostic ploïdique.

I.2. Résultats

Les fichiers bruts sont recalés en fonction de la valeur moyenne en ADN des fichiers de références et après multiplication par un facteur de correction égal à 1.15 (lymphocytes). Les histogrammes recalés et les histogrammes d'Opfermann correspondant sont représentés sur la Figure III.2 pour les cancers gynécologiques et sur la Figure III.3 pour les fichiers prostatiques.

Seuls les laboratoires I et II ont donné 100% de bons diagnostics. Ces bons diagnostics on été définis à partir des résultats de l'analyse des échantillons par cytométrie en flux.

Les 4 laboratoires ont donné le même diagnostic pour seulement 3 échantillons sur 4 ; les spécimens 1 et 2 de prostate ont été diagnostiqués comme aneuploïdes. Le spécimen 1 de l'utérus a été diagnostiqué aneuploïde.

L'observation visuelle de l'histogramme et les paramètres statistiques (l'indice de ploïdie est à chaque fois supérieur à 1) vont dans le même sens et mettent bien en évidence une aneuploïdie

vers 2.2c avec présence de cellules supérieures à 5c. Tous les paramètres statistiques présentent des valeurs élevées.

L'histogramme d'Opfermann est dans les trois cas très révélateur de cette aneuploïdie puisque la classe centrée sur 2.5c correspond dans les trois cas à la classe modale. Remarquons également que les histogrammes des références sont unimodaux.

Le cas du spécimen 4 de la prostate peut également être considéré comme bien classé par les 4 laboratoires. En effet, cet histogramme ne présente pas de difficulté d'interprétation. Le laboratoire 4 l'a classé comme polyploïde mais nous sommes bien en présence de clones aneuploïdes comme l'atteste en particulier l'histogramme d'Opfermann. De plus cet échantillon manifeste une prolifération très importante (Indice de Prolifération d'Opfermann = 23%) déjà mise en évidence par la bimodalité de l'histogramme de référence.

Les autres spécimens ont donné lieu à des diagnostics différents (diploïde *versus* aneuploïde).

A noter encore une fois que les histogrammes d'Opfermann présentent dans les trois cas une classe modale centrée sur 2c. De plus, les indices de ploïdie sont corrects. Le spécimen 3 de prostate présente cependant quelques variations par rapport à un histogramme diploïde parfait, comme la présence de cellules supérieures à 5c. De plus, son indice de ploidïe est de 1.04, ce qui est confirmé, sur l'histogramme d'Opfermann, par une fréquence élevée d'individus dans la classe centrée sur 2.5c.

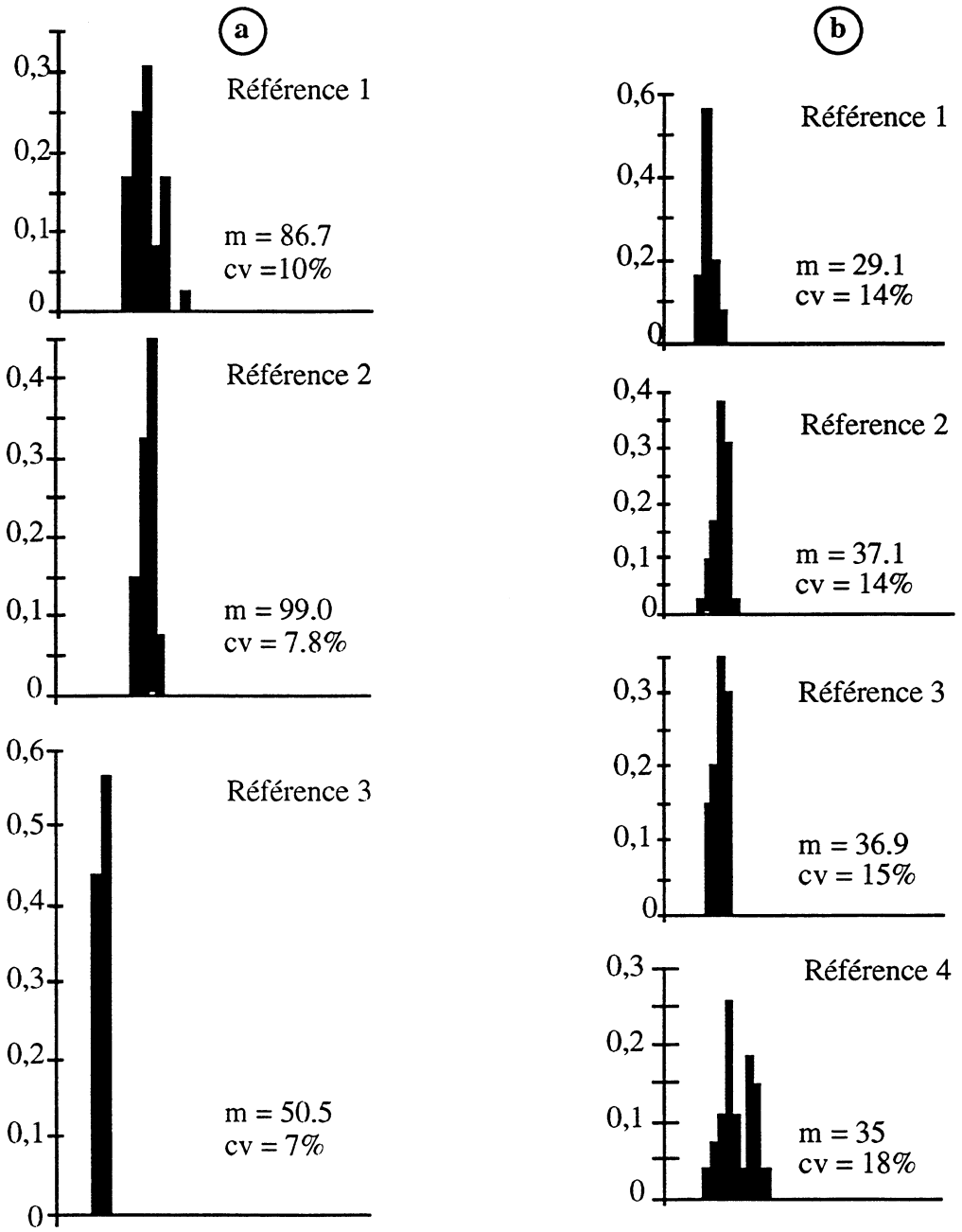


Figure III.1 : Histogrammes des références de lymphocytes :

a) Cancers gynécologiques.

b) cancers prostatiques.

Les moyennes et les coefficients de variation sont calculés à partir de toutes les cellules du fichier de référence.

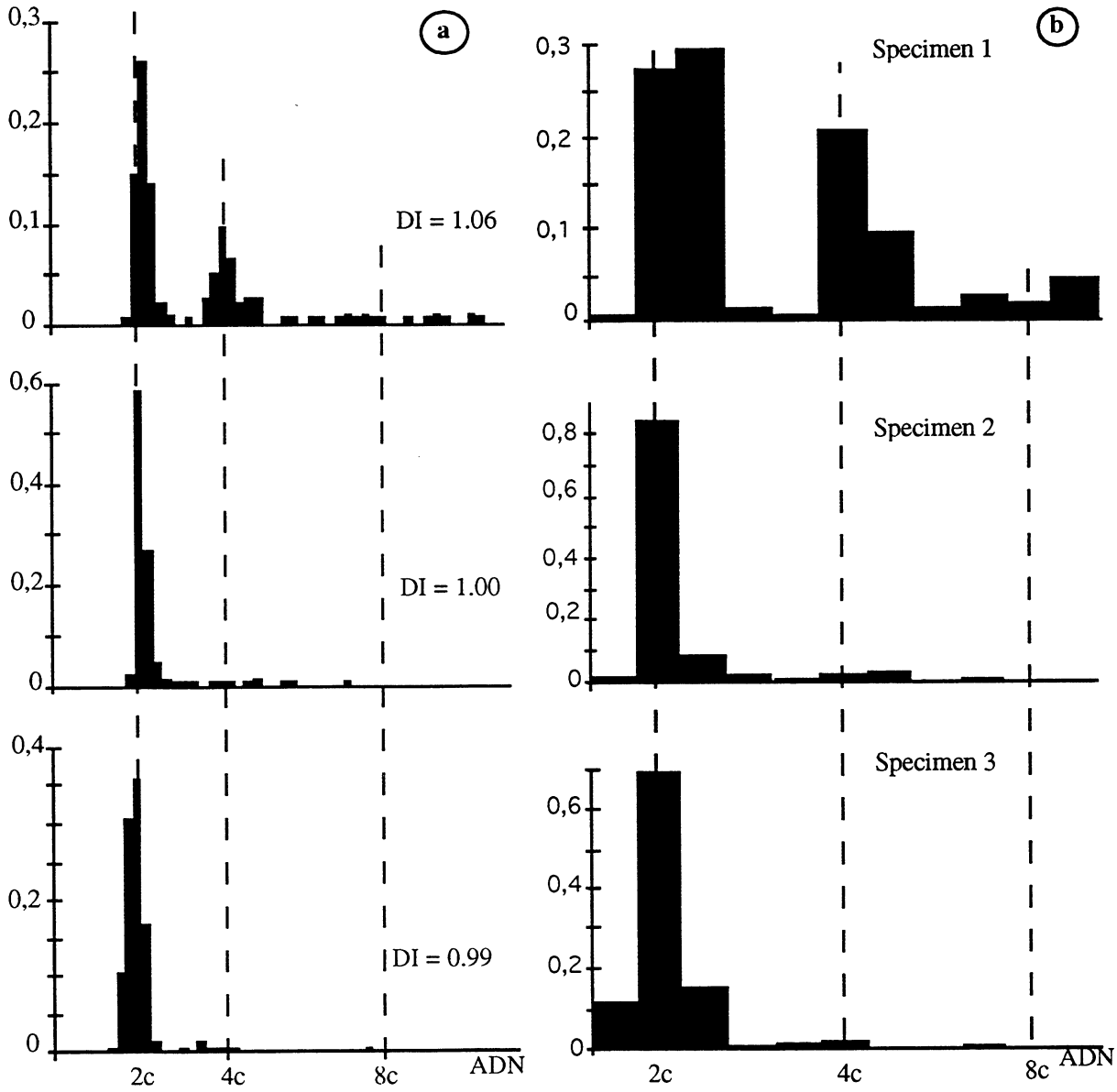


Figure III.2 : Histogrammes d'ADN des trois cancers gynécologiques :

a) Histogrammes recalés par 2c = 1.15 * (valeur d'ADN moyenne des lymphocytes).

b) Histogrammes d'ADN codés en 10 classes selon Opfermann.

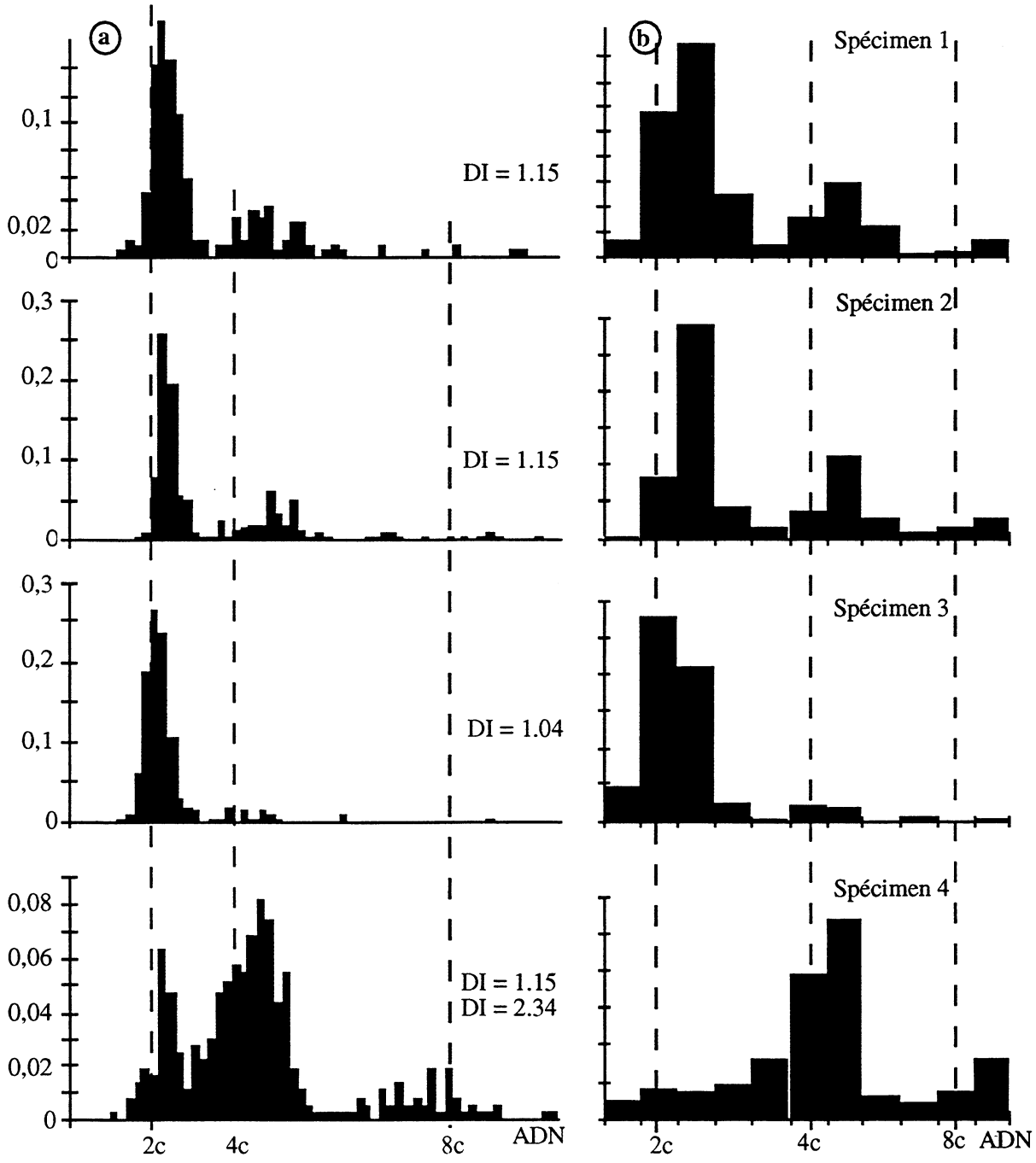


Figure III.3 : Histogrammes d'ADN des 4 cancers de la prostate:

a) Histogrammes recalés par 2c = 1.15 * (valeur d'ADN moyenne des lymphocytes).

b) Histogrammes d'ADN codés en 10 classes selon Opfermann.

	Type d'Auer	> 2.5c	> 5c	2cDI	DNA -MG	Entropie	Balance de Ploïdie	Index de Prolifération
Specimen 1	**II / III**	**0.48**	**0.13**	**6.76**	**1.56**	**3.72**	**0.01**	**0.10**
Specimen 2	**I**	**0.07**	**0.02**	**0.47**	**0.30**	**1.94**	**0.72**	**0.02**
Specimen 3	**I**	**0.04**	**0.00**	**0.25**	**0.17**	**2.28**	**0.42**	**0.00**

Tableau III.1: Paramètres statistiques calculés sur les trois cancers gynécologiques.

	Type d'Auer	> 2.5c	> 5c	2cDI	DNA -MG	Entropie	Balance de Ploïdie	Index de Proliferation
Specimen 1	**III**	**0.50**	**0.11**	**5.05**	**1.37**	**3.96**	**- 0.38**	**0.05**
Specimen 2	**III**	**0.53**	**0.19**	**8.31**	**1.70**	**1.83**	**- 0.59**	**0.09**
Specimen 3	**I / II**	**0.16**	**0.01**	**0.94**	**0.50**	**2.98**	**0.00**	**0.04**
Specimen 4	**I**	**0.86**	**0.28**	**10.97**	**1.89**	**4.86**	**- 0.42**	**0.23**

Tableau III.2: Paramètres statistiques calculés sur les 4 cancers prostatiques.

Spécimen 1	Inflammation et forte suspection d'aneuploïdie
Spécimen 2	Absence d'anomalies ploïdiques
Spécimen 3	Absence d'anomalies ploïdiques

Tableau III.3 : Diagnostics des 3 cancers gynécologiques

Spécimen 1	Inflammation et forte suspection d'aneuploïdie
Spécimen 2	Inflammation et forte suspection d'aneuploïdie
Spécimen 3	Absence d'anomalies ploïdiques
Spécimen 4	Forte aneuploïdie et prolifération

Tableau III.4 : Diagnostics des 4 cancers prostatiques

	Laboratoires			
	I	II	III	IV
Prostate				
spécimen 1	Aneuploid	Aneuploid	Aneuploid	Aneuploid
spécimen 2	Aneuploid	Aneuploid	Aneuploid	Aneuploid
spécimen 3	Diploid	Diploid	Aneuploid	Diploid
spécimen 4	Aneuploid	Aneuploid	Aneuploid	Polyploid
Uterus				
spécimen 1	Aneuploid	Aneuploid	Aneuploid	Aneuploid
spécimen 2	Diploid	Diploid	Aneuploid	Aneuploid
spécimen 3	Diploid	Diploid	Aneuploid	Aneuploid

Tableau III.5: Diagnostics des 4 laboratoires pour les 7 échantillons.

I.3. Discussion

Le but de cette étude était d'observer les différences d'interprétation d'histogrammes d'ADN et du diagnostic entre différents laboratoires. Les résultats mettent bien en évidence les limites inhérentes de ces interprétations.

Il est utile de préciser à nouveau que

- les divergences inter-laboratoires proviennent exclusivement de l'analyse des données, puisque que seuls les fichiers brutes d'ADN étaient disponibles : les étapes d'échantillonnage, de coloration, de préparation, d'acquisition ont été effectués une seule fois par le laboratoire du Pr. A. Reith.

- chaque laboratoire procédait selon son propre savoir-faire (choix de la valeur de recalage, paramètres statistiques, méthode de décision en fonction de la valeur de ces paramètres).

La principale constatation que nous avons faite dans le laboratoire est que nous avions complètement omis d'utiliser les paramètres statistiques. En effet, tous les diagnostics ont été établis sur la simple observation des histogrammes d'ADN et en particulier de l'histogramme d'Opfermann. Ce n'est qu'*à posteriori* que nous avons regardé les paramètres statistiques. Pour les trois échantillons dont les diagnostics inter-laboratoires étaient discordants, aucun de ces paramètres n'apportaient d'informations complémentaires ou nouvelles.

L'histogramme d'Opfermann, plus encore que les deux paramètres qui en dérivent (Index de prolifération et balance de ploïdie), a été l'outil le plus performant dans l'établissement du diagnostic.

En routine clinique, dans la majorité des cas (70%-80%), l'interprétation des histogrammes d'ADN ne pose aucun problème. Il est malheureusement relativement rare que ces histogrammes (comme les 4 histogrammes de cette étude qui ont donné 100% de bonne

classification) apportent une information complémentaire pour le diagnostic clinique. En général, l'analyse de l'ADN ne fait que confirmé le diagnostic clinique.

Il y a plusieurs années, on pensait que l'analyse de l'ADN serait un paramètre intéressant pour l'aide au diagnostic d'échantillons tumoraux pour lesquels les caractéristiques cliniques ne permettaient pas de conclure.

La variabilité inter-laboratoire d'interprétation des 3 échantillons tumoraux reflète la difficulté d'utiliser l'histogramme d'ADN en routine clinique. En effet, la confirmation d'une aneuploïdie est certes intéressante. Mais, on peut supposer que les échantillons dont l'interprétation pose des problème en clinique sont les mêmes que ceux dont l'analyse de l'ADN pose des problèmes.

Le recalage, la construction, et l'interprétation d'un histogramme sont, à l'heure actuelle trop dépendants de l'appareillage, de la méthodologie employée et surtout des paramètres statistiques utilisés.

Les deux laboratoires I et II n'ont pas forcément donné les diagnostics exacts mais simplement les mêmes diagnostics que ceux obtenus après analyse des échantillons tumoraux en cytométrie en flux. La cytométrie en flux, au niveau de la méthodologie et des mesures comme au niveau de l'analyse des histogrammes, connaît également de nombreuses limites. Pour connaître le diagnostic exact il faudrait analyser toute la tumeur. C'est sans doute le seul moyen de confirmer, par exemple, l'absence d'un clone aneuploïde dans l'échantillon prostatique 3.

La variabilité inter-laboratoire, au niveau du diagnostic, mise en évidence dans cette étude, peut expliquer en partie les différences et la variabilité des résultats concernant la valeur pronostique de l'analyse de l'ADN dans certaines tumeurs.

II. Analyse de l'ADN et pronostic des cancers

L'intérêt pronostique de nombreux marqueurs nucléaires et cellulaires a donné lieu depuis une dizaine d'années à une multitude de publications portant sur le rôle pronostique de marqueurs nucléaires, et en particulier de l'ADN. La majorité de ces travaux utilisent la cytométrie en flux. L'analyse d'image, technique plus récemment introduite dans les laboratoires, n'offre encore pas assez de recul pour suivre des cohortes de patients afin d'essayer de trouver des paramètres permettant de prédire la survie individuelle des patients. Les résultats obtenus surtout en cytométrie en flux sont très contradictoires quant à la valeur pronostique des descripteurs des histogrammes d'ADN. Peu de travaux utilisant l'analyse d'image ont été publiés sur ce sujet, à cause sans doute de l'introduction relativement récente des systèmes d'analyse d'image en routine clinique.

Il nous semblait donc intéressant d'étudier la valeur prédictive des descripteurs -ou d'une combinaison de ces descripteurs- des histogrammes d'ADN obtenus par analyse d'images. Nous avons réalisé une étude rétrospective sur une cohorte de 116 patientes, opérées d'un cancer du sein entre 1986 et 1987, et pour lesquelles l'analyse de l'ADN avait été effectuée. Nous nous sommes intéressé également aux valeurs prédictives de certains descripteurs en fonction du statut ganglionnaire des patientes.
Cette étude a fait l'objet d'une publication soumise au Journal CANCER en Février 1993.

DNA Image Cytometry prognosis of Breast carcinomas

Guillaud M.(1), Ph.D, Louis J., Ct, and Seigneurin D., MD-PhD (2)

(1) To whom requests for reprints.
Laboratoire TIMC. Equipe de Reconnaissance des Formes et de Microscopie Quantitative. CERMO. BP 53X 38041 grenoble.
(2) Laboratoire de cytologie quantitative, Faculté de médecine de Grenoble. Domaine de la Merci.38700 La Tronche

Running title : DNA prognosis of breast carcinomas

Acknowledgements

We wish to thank Dr. V. von Hagen for manuscript preparation and Dr. M. Brugal for valuable work in documentation research.

Abstract

The prognosis value of lymph node involvement is now well established but this function does not accurately predict the outcome of every individual case. Thus, some quantitative and objective criteria are needed.

116 cases of breast carcinomas were studied; all were invasive galactophoric carcinomas. They were followed-up for at least five years. Feulgen stained imprints were analysed by SAMBA 200 cell image analyser. DNA histograms were grouped into 4 types according to a new visual classification. Ten parameters were also computed from DNA histograms (2.5c-ER, 5c-ER, 8c-ER, Ploidy Balance, Proliferation Index, Mean DNA Index, Modal DNA index, 2c-DI, DNA-MG and Entropy). The prognostic value of these parameters was studied and related to survival. Only two parameters, Ploidy Balance and the new visual classification, have independent prognostic value in the total population as well in node-positive patients. Furthermore, all axillary node-negative patients and classified as euploid, according to our visual classification, are always alive after five years. This last result has to be confirmed by a longer follow-up, but the visual classification seems to be the most efficient DNA histogram parameter to discriminate, among node-negative patients, which of them require adjuvant chemotherapy.

Key words :

DNA content. Image Cytometry. Breast Cancer. Prognosis. Survival analysis.

INTRODUCTION

Breast carcinoma is the most common malignant neoplasm in women and it extremely variable and unpredictable course is well known. The prognostic value of the lymph node involvement, tumour size, histological grade and hormone receptor status are now well established (1), but these factors do not accurately predict the outcome in every individual case. There still remains wide variation in the survival of patients with tumours of the same clinical status or histological grade.

Axillary node status is the single most important prognostic factor for patients with breast cancer. Nethertheless, although histologically node-negative (Stage I) patients are recognised, in general, to be a good prognostic group, Fisher (2) showed that approximately 30% of patients with axillary node-negative breast cancer will relapse and die of recurrence within 10 years.

Quantitative analysis of cellular DNA content may be clinically useful in the prognostic evaluation of certain types of malignant tumours, including breast cancer.

In earlier studies using flow cytometry, the association of DNA content with prognosis, remission disease free survival, and lymph node involvement have been demonstrated, but the results are not uniform (3,4,5). For instance, Keyhami-Rofagha et al.(5), among others (6,7,8), who defined aneuploid tumour as any additional population excluding the diploid and the tetraploid populations, conclude that simple determination of DNA ploidy fails to indicate prognosis for infiltrative, node-negative breast carcinoma, with a follow-up of 3 to 15 years. Conversely, some authors such Kallioniemi (9), Ellis (10), Hedley (11) and Owainati (12) found that DNA content was indeed a useful prognostic parameters. Recently, Johnson et al (13) showed that DNA ploidy was an independent determinant of survival in a series of 100 patients with node-negative breast cancer and with a follow-up of at least 10 years. Beerman et al (14) showed that a division of tumour ploidy into 5 classes based on DI index improved the discrimination between different prognostic groups of patients, as compared to classical classification (aneuploidy *versus* diploid).

Previous reports based on microspectrophotometry measurements of DNA also suggested that the ploidy of the cells in neoplastic tissue from patients with primary breast cancer may be of prognostic value (15, 16).

Nethertheless, to our knowledge, only two or three recent studies using image cytometry (17, 18) have been published on this subject. The low number of image cytometry studies is due to the only recent placing of image analysis systems in clinical laboratories. Longin et al (19), using image cytometry, found that 80% of the primary breast carcinoma were aneuploid, but conclude that a long-term follow-ups are needed to determine which variable will provide reliable prognostic information. Ten years ago, Auer *et al.*(16), in a retrospective study of material obtained by FNAB, observed a correlation between certain distribution patterns of histograms of cell DNA-content and patient survival.

Recently, Guzman (18) used image cytometry to study 44 cases of invasive ductal breast cancer followed for 10 years. The main results were that all patients with diploid tumours,

tumour size T1 and negative node status survived for >10 years. Nethertheless the independent prognostic value of DNA ploidy was not demonstrated.

During the last decade, numerous indices or parameters were computed from DNA histograms from image cytometry in order to improve the classification or the prognosis of carcinomas. While in flow cytometry, only the diploid *versus* aneuploid distinction is made, in image cytometry, more than ten parameters have been proposed by different authors (15, 16, 17, 20, 21, 22). Hopefully, results given by image cytometry can be compared with results from by flow cytometry. Since comparative studies of DNA ploidy by flow cytometry *versus* image cytometry generally show strongly concordant results (23). Nethertheless, some authors (24) have showed that image analysis had significant advantages over flow cytometry, such as the capability of detecting rare high ploidy cells, often determinant in DNA histogram classification.

Despite the enormous amount of literature on the prognostic value of DNA cytometry in breast cancer, prediction of the outcome for a particular patient is still impossible.

In this paper, we report on the predictive value of 10 well-known DNA parameters and of a new DNA pattern visual classification on a series of intraductal breast carcinomas analysed by image cytometry.

Characteristics	Age (years)		Cytologic Grade			Tumor Size (cm)			Node status			
	<50	≥50	1	2	3	<2	2< x < 5	>5	0	1	2	3
N° of Patients	31	85	17	37	58	40	59	13	57	18	26	10

MATERIAL AND METHODS

Patients and sample collection

One hundred sixteen patients included in this prospective study underwent surgery from the beginning of 1986 to the end of 1987. They had all galactophoric breast carcinomas. Patients range in age from 25 to 88 with median age of 59. All patients underwent tumorectomy or radical mastectomy procedures. Axillary lymph node status, tumour size, histological grade and cytological grade are also available (Table 1). These patients were followed up for a minimum of five years. During the five years which followed the diagnosis, 14 deaths, 6 metastases and 4 local recurrences were observed among the 116 patients.

Table 1: Clinical characteristics of the patients.

Cell preparation and staining.

For DNA assessment, imprint smears were made from gently scraped fresh-cut surface of the breast tumour. The imprints were immediately fixed in a phosphate buffer with formol (9%) and acetone (45%) for at least 2 minutes and stained using a Feulgen staining procedure. Tumour fragments used for touch imprints were fixed in Bouin's fixative and embedded in paraffin for histological diagnosis and grading. The histological examination of tumour imprints were performed by experienced pathologists.

Image cytometry DNA determination

The image analyser used was the SAMBA 200 (System for Analytical Microscopy in Biological Applications) cell image processor, purchased from Alcatel TITN, FRANCE. The organisation and operation of this analyser have already been described (25).
Between 100 and 300 tumour cell nuclei were measured in each case. To determine nuclear DNA content of the tumour cells, trout erythrocytes stained simultaneously with tumour samples, were use as external diploid DNA control reference. At least 30 trout erythrocytes were analysed for each case.

Histogram analysis

Because the Feulgen reaction is considered to be stochiometric and specific for DNA, the stereologically corrected extinction values obtained can be interpreted as the DNA content of cell nuclei. Table 2 gives the 10 DNA parameters computed on each DNA histogram.

Parameters	**Abbreviation**	**References**
2.5c-Exceeding rate	2.5c-ER	Böcking (20)
5c-Exceeding rate	5c-ER	Böcking (20)
8c-Exceeding rate	8c-ER	Böcking (20)
Ploidy Balance	PB	Opfermann (17)
Proliferation Index	PI	Opfermann (17)
Mean DNA Index	DI-mean	Atkin (15)
Modal DNA value	DI-mod	Atkin (15)
2c-Deviation Index	2cDI	Böcking (20)
DNA malignancy Grade	DNA-MG	Böcking (20)
Entropy	Entropy	Stenkvist (21)

Table 2 : List of computed DNA parameters (see text).

In addition, the DNA histograms from the 116 breast carcinomas have been grouped into four classes or types (Figure 1) which are reasonably well defined because few cases fell between classes. Type I is characterised as having a single distinct modal value in the diploid or near-diploid region of normal cells. Type II is characterised as having a major population in the aneuploid region (> 2.3c). Type III populations show either a distinct modal value in the tetraploid or near-tetraploid region with or without a well-defined peak in the 2c region. Type

IV populations show a very pronounced and irregular aneuploidy, with DNA amounts per cell ranging from levels near 2c to values beyond 6c or even 8c.

In comparison with AUER's classification, the main differences in our visual classification lie in the type II and type III. Our classification takes into account the location of the histogram major peak (aneuploid *versus* tetraploid) while Auer's classification takes into account the percentage of cells between 2c and 4c (i.e. proliferating cells) to be classified as type II or type III.

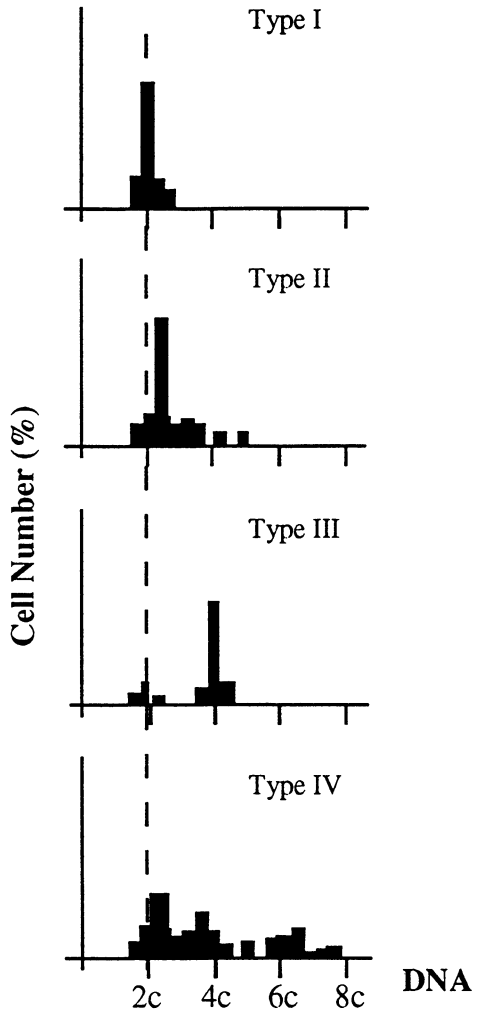


Figure 1: Visual classification of DNA distribution patterns, determined by Feulgen cytophotometry from a series of mammary carcinomas (types I to IV).

Statistical methods

Principal components and linear discriminant analysis were used to study the capacity of the DNA parameters to discriminate disease free-survival patients from patients who relapse or die. Survival curves were estimated using the Kaplan-Meier technique and were compared using log-rank statistics.

RESULTS

Multivariate analysis of DNA parameters

Correlations study

Table 3 shows the correlation matrix of the ten DNA parameters. Some of these correlations were relatively strong and are indicated in bold type. In particular, we observed, the absence of a significant correlation between the two Opferman's parameters, PB and PI, and the other parameters. Furthermore, the correlation coefficient between the Proliferation Index and all the others is always negative.

	2.5c-ER									
2.5c-ER	1	**5c-ER**								
5c-ER	0.31	1	**8c-ER**							
8c-ER	0.25	0.5	1	**PI**						
PI	0.37	0.16	0.11	1	**PB**					
PB	-0.30	-0.23	-0.12	-0.15	1	**DImean**				
DImean	**0.87**	0.64	0.50	0;32	-0.27	1	**DI mod**			
DImod	**0.80**	0.41	0.25	0.14	-0.24	**0.80**	1	**2cDI**		
2cDI	0.67	**0.83**	0.62	0.26	-0.20	**0.93**	0.66	1	**DNA-MG**	
DNA-MG	**0.84**	0.57	0.54	0.37	-.024	**0.97**	0.67	**0.90**	1	**Entropy**
Entropy	0.65	0.31	0.38	0.55	-0.30	0.65	0.36	0.57	0.76	1

Table 3 : Correlations Matrix of the 10 DNA parameters.

Multivariate analysis

Figure 2a shows the first factorial plan of principal components analysis. No immediate structure can be observed. In particular, patients who relapse (or die) and disease-free patients cannot be separated in this first projection plane

In order to find a combination of the initial DNA parameters (Table 2) which allows discrimination between favourable outcome patients and unfavourable outcome patients, we isolated two groups of 22 patients : the first includes all patients with an unfavourable outcome, the second group includes 22 patients with favourable outcome (selected from the left side of the cluster of the Figure 2A). A linear discriminant analysis was performed on the both groups, as learning step.

The purpose of this analysis was limited to the eventual detection of the parameters - or combination of these parameters - which permit these two groups to be separated. In the learning step, only one parameter, Opferman's Ploidy Balance (13) is needed to obtain 100% good classification (Figure 2B).

The use of the previous linear discriminant function to classify the 116 patients, in decisional analysis, gave only 50% good classification.

Thus we chose to consider only the Ploidy Balance parameter in the following survival study.

Survival study

Dead patients, patients with metastases and patients with local recurrences are called unfavourable outcome patients (or recurence free-survival patients).

Visual Classification : The types I and III on the one hand and types II and IV on the other have been regrouped in two classes, which can be considered as euploid *versus* aneuploid classification.

In the first step, the predictive value of Ploidy Balance, node status and visual classification for the total population was studied (Figure 3).

These three factors predicted, independently, the number of recurrence free-survival patients (PB : $p = 0.005$; Node status, $p = 0.009$; visual classification: $p = 0.002$).

The total population was then subdivided as function of the node status.

In axillary lymph node-positive patients, PB (Log-rank value = -2.45, $p = 0.014$) and visual classification (Log-Rank value = -2.05, $p = 0.04$) predicted the recurrence free-survival patients.

In axillary node-negative patients (n = 56), neither of these two parameters predicted the recurrence free-survival patients (Figure 5). Nethertheless, as Figure 5B shows, there are no pejorative cases in the first group of patients which correspond to types I and III of our classification. The Log-Rank test cannot be used here because there are no pejorative events in this "euploid" class.

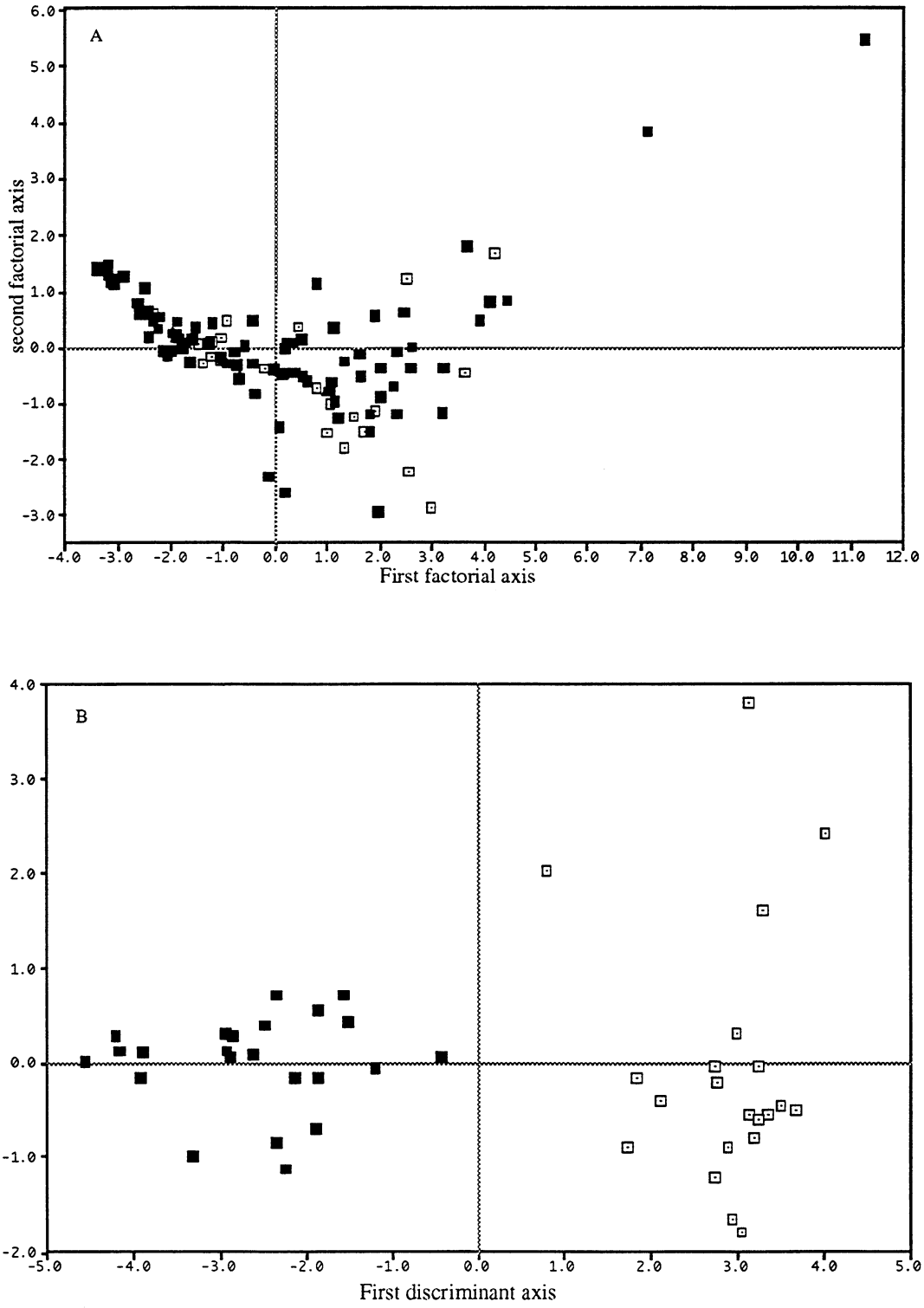


Figure 2: Factorial analysis.
A : Plot of the 116 patients against values for the first two principal components of a PCA.
B: Plot of the two groups of 22 patients (learning groups) against their values for linear discriminant function
Open squares indicate patients with unfavourable outcomes, Closed squares indicate patients with favourable outcomes.

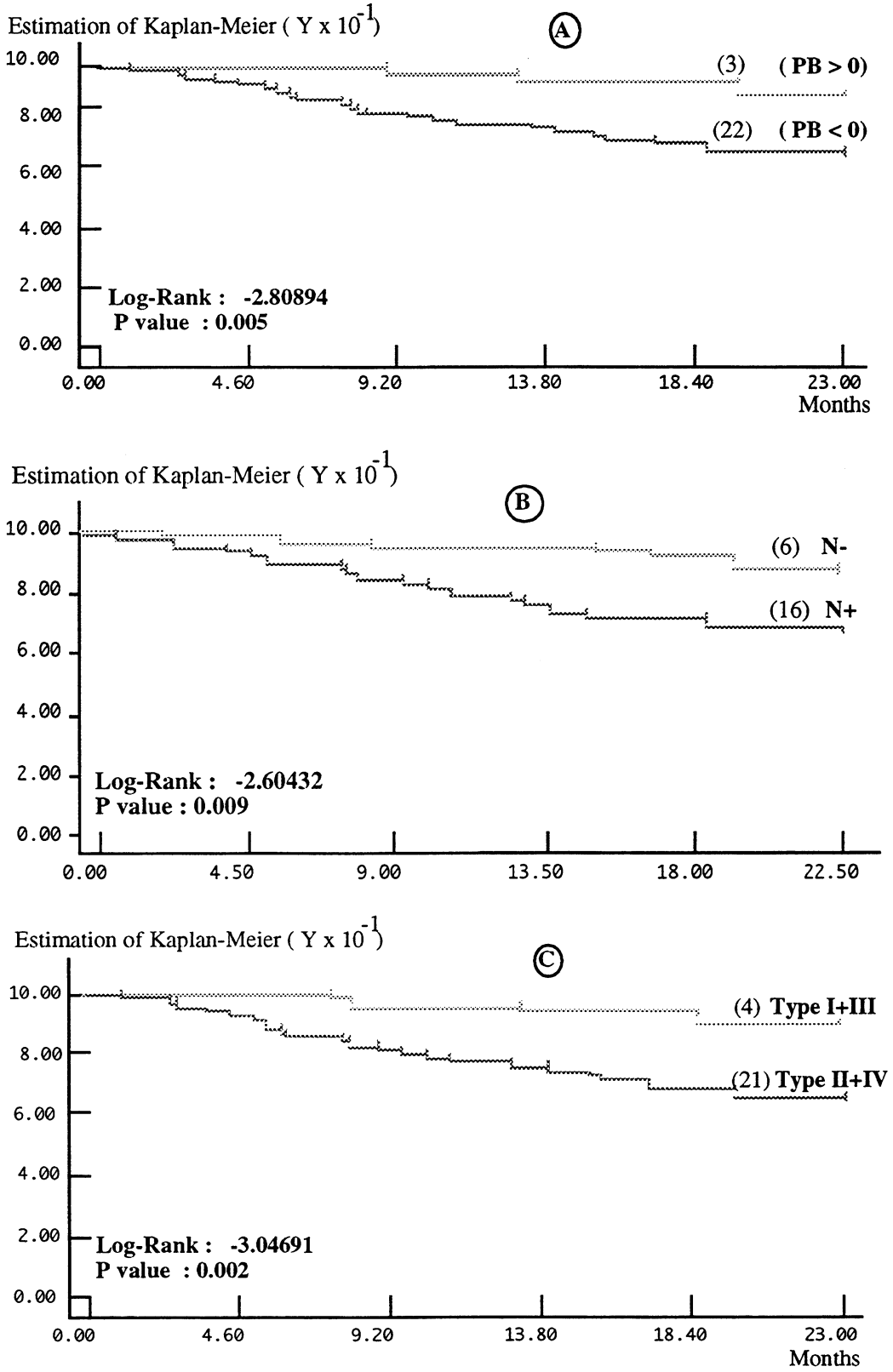


Figure 3 : Desease-free survival by Ploidy Balance status (**A**), by Node status (**B**) and by visual classification (**C**) for the total population (n =116).

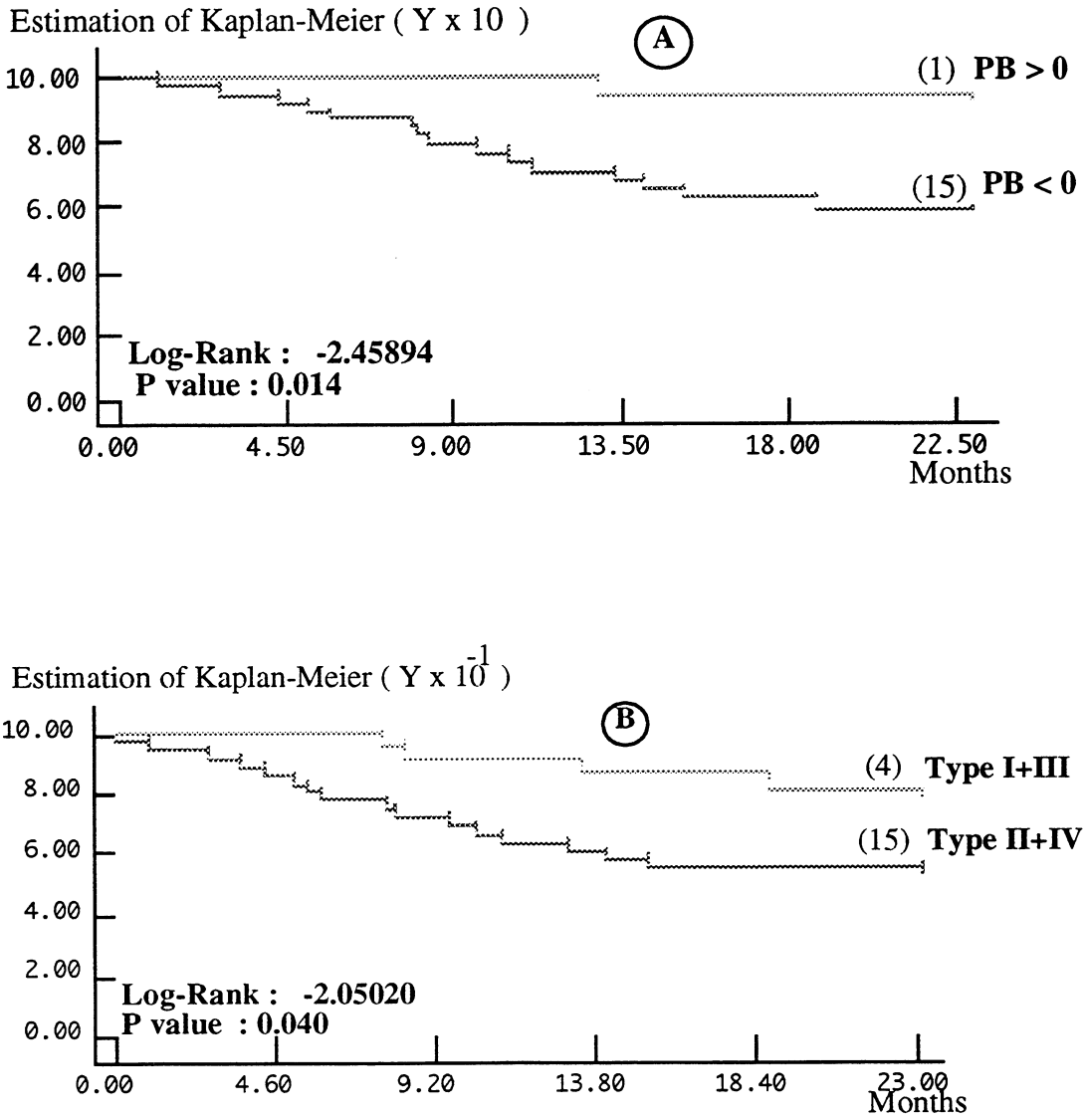


Figure 4. Desease-free survival by Ploidy Balance status (**A**, n = 55) and by Visual Classification (**B**,n = 59) for node-positive patients.

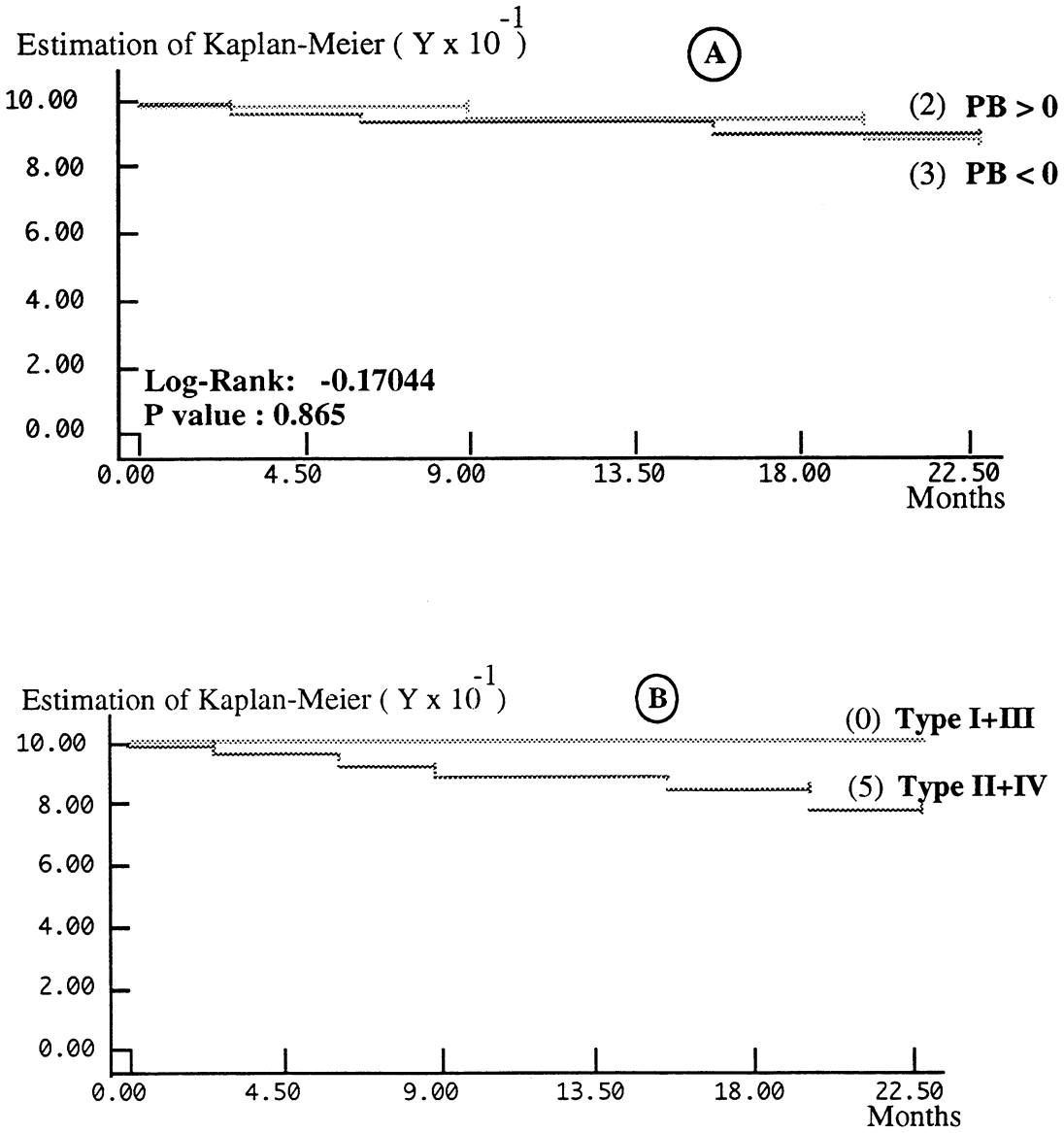


Figure 5 : Desease-free survival by Ploidy Balance status (**A**, n = 59) and by Visual classification (**B**, n =56) for node-negative patients.

DISCUSSION

In our study, we examined 10 parameters computed from DNA histograms in order to find which parameter or combination of parameters provide non-redundant prognostic information for breast cancer. The prognostic value of the combination of DNA parameters with other clinical parameters such as tumour size or cytological grade gave no statistically significant results.

One of the major conclusion of this study is that "classical" parameters derived from DNA histograms, such as DNA-MG, entropy or 2cDI, fail to characterise DNA profile.
Only two parameters, PB and visual classification, have independent prognostic value in the total population as well in node-positive patients.

The advantage of the PB, as compared with the others (Entropy or 2cDI, for instance), is that the coding of this parameter into two groups (required for survival analysis) can be made a *priori*, independently of the sample studied; PB corresponds, in fact, to the classical classification of aneuploid *versus* euploid. Euploid histograms have a positive PB (which varies from 0 % to 100%) while aneuploid histograms have a negative PB (-100% to 0%). Nethertheless, the value of PB is strongly dependant on the range of the 10 classes proposed by Opfermann (17). The class range has to be computed as a function of the value of the coefficient of variation of the reference diploid population.

The prognostic value of this new histogram visual classification is not surprising because it corresponds, after grouping into two classes, to the classical aneuploid *versus* euploid classification. The advantage of this classification, as Auer's, is that it takes into account all the information contained in the DNA histogram.

Neither PB parameter nor visual classification are of independent prognostic value in the node-negative patients population. Nethertheless, all axillary node-negative patients and classified as type I or III are always alive after five years (Figure 5B). This result has to be confirmed by a longer follow-up, but the visual classification seems to be the most efficient parameter to discriminate, among node-negative patients, which of them require adjuvant chemotherapy.

The restriction of DNA profile to a single parameter results in a loss of the all information contained in the histogram, and explain the low prognostic values of classical DNA parameters. Any indices -or combinations- can replace the general and global assessment of the histograms by an expert pathologist. We considered that the DNA histogram should be analysed as a pattern. Thus the DNA histogram interpretation can be treated as a pattern recognition problem. The prognostic values of DNA histogram visual classification by human experts, which corresponds to a pattern recognition procedure, lends weight to this approach..
Published results are often contradictory and difficult to compare because too many different parameters or approaches have been used (26,18). The DNA histogram classification into euploid *versus* aneuploid classes can be obtained from so many different methods that, in our opinion, no inter-laboratory comparison is possible. Furthermore, there are many other explanations for the lack of the reproducibility of published results : differences in sampling

heterogeneity, selection bias, differences in the quantitative parameter computation (such as Ploidy Index), differences in histogram rescaling, reference choice and measurements methods.

However, a long-term follow-up of at least 10 years are needed to confirm the prognostic value of the DNA visual classification. The small number of cases does not permit definitive statistical conclusions. It is well known that DNA histogram interpretation couldn't alone predict survival of node-negative patients. We thought, nethertheless, that inter-laboratory standardisation of the DNA analysis, at the methodological as well as the statistical interpretation level, would be of great interest to identify, especially among node-negative patients, those with unfavourable outcomes.

References

1 Merkel D.E. and Osborne CK. Prognostic factors in breast cancer. Hematol. Oncol Clin North Am. 1989; 3:641-652.

2 Fischer B., Bauer M., Wickerman L. et al. Relationship of the number of positive axillary nodes to the prognosis of patients with primary breast cancer. Cancer 1983; 52: 1551-1557.

3 Coulson PB., Thornthwaite JT., Wooley TW., Sugarbaker EV., Seckinger D. Prognostic indicators including DNA histograms type, receptor content, and staging related tohuman breats cancer patient survival. Cancer Res. 1984; 44 : 4187-4196.

4 Cornelisse CJ., Van de velde CJ., Caspera RJ., Moolenar AJ., Hermans J. DNA ploidy and survival in breast cancer patients. Cytometry 1987; 8:225-234.

5 Keyhani-Rofagha S., O'toole R.V., Farrar W.B., Sickle-santanello B., Decenzo J. and Young D. Is DNA ploidy an independant prognostic indicator in infiltrative node-negative breast carcinoma ? *Cancer* 1990: 65: 1577-1582.

6 Merkel DE., Mc Guire WL. Ploidy, proliferative activity and prognosis. DNA flow cytometry of solid tumors. Cancer 1990; 65: 1194-1205.

7 Dowle CS., Owainati A., Robins A., Burns K., Ellis IO., Elson CW., Blamey RW. Prognostic significance of the DNA content of human braest cancer. BR. J. Surg. 1987; 87: 133-136

8 Ewers SB., Baldetorp B., Killander D., Langström E. Flow cytometry DNA ploidy and number of cell populations in the primary breast cancer and their correlation to the prognosis. Acta Oncol. 1989; 6: 913-918.

9 Kallioniemi O. Blanco G., Allavaiko M., Hietanen T., Mattila J., Lauslahti M.and Koivola T. Tumour DNA ploidy as an independant prognostic factor in breast cancer. Br. J. Cancer 1987; 56:637-642.

10 Ellis CN., Frey ES., Burnette JJ., et al. The content of tumor DNA as an indicator of pronosis in patients with T1N0M0 andT2N0M0 carcinoma of the breast. Surgery 1989; 106:133-138.

11 Hedley DW., Rugg CA. and Gelberg RD. Assoociation of DNA index ans S-phase fraction, with prognosis of nodes-positive early breast cancer. Cancer Research,1987 47, 479-4735.

12 Owainati AA., Robins RA., Hinton C et al. Tumour aneuploidy, prognostic parameters and survival in primary breast cancer. Br. J. Cancer 1987; 55:449-454.

13 Johnson H., Masood S., Belluco C., Abdou-Azama A., dee S., Kahn L. and Wise L. Prognostic factors in Node-Negative breast cancer. Arch.Surg. 1992; 127:1386-1391.

14 Beerman H., Kluin PM., Hermans J., van de Velde C.J.H. and Cornelisse C.J. Prognostic significance of DNA-ploidy in a series of 690 primry breast cancer patients. Int. J. Cancer: 1990; 45:34-39.

15 Atkin NB. Modal DNA value and survival in carcinoma of the breast. Br Med J. 1972; i: 271-272.

16 Auer G.U., Caspersson T.O. and Wallgren A.S. DNA Content and Survival in Mammary Carcinoma. Anal. Quant. Cytol. Histol. 1980 ; 2: 161- 165.

17 Opfermann M., Brugal G., Vassilakos P. Cytometry of breast carcinoma : Significance of ploidie Balance and Proliferation index. Cytometry 1987; 8:217-224.

18 Guzman J., Rückmann A., Glaser A., Wittekind C., schönfeld B. and Kiefer G. DNA cytophotometric Analysis of Breast Cancer. Follow-up for 10 years. Anal. Quant. Cytol. Histol. 1992 ; 14 : 427-432.

19 Longin A., Fontaniere B., Pinzani V., Catimel G., Souchier C., Clavel M. and Chauvin F. An image cytometric DNA_analysis in Breast neoplasm. *Path. Res. Pract.* 1992; 188:466-472.

20 Böcking A., Adler C-P., Common H.H., Hilgarth M., Granzen B., Auffermann W. Algorithm for a DNA-Cytophotometric diagnosis and grading of malignancy. Anal. Quant. Cytol. Histol. 1984; 6: 1-8.

21 Stenkvist B. and Strande G. Entropy as an algorithm for the statistical description of DNA cytometric data obtained by image analysis microscopy. ACP 1990; 2:159-165.

22 Weber J.F., Bartels P.H., Bartels H.C. and Bibbo M. Discrimination of DNA ploidy patterns by order statistics. Anal. Quant. Cytol. Histol. 1987; 9, 60-68

23 Wingren S., Hatschek, Stal O., Boeryd B. and Nordenskyold B. Comparison of static and flow cytofluorometry for estimation of DNA index ans S-phase fraction in fresh and paraffin-embedded breast carcinoma tissue. Acta Oncologica 1988; 27:793-797

24 Ghali V.S., Liau S., Teplitz C. and Prudente R. A comparative study of DNA Ploidy in 115 Fresh-Frozen Breast carcinomas by image Analysis versus Flow cytometry. Cancer 1992; 70: 2668-2672.

25 Adelh D., signor G. Un analyseur d'images pour la cytologie. *In* Reconnaissance de Formes et Intelligence Artificielle. Paris, IRIA, 1981, pp 203-210.

26 Haroske G., Kunze K.D. and Theissig F. Prognosis significance of image cytometry DNA parameters in tissue sections from breast and gastric cancers. Anal.Cell Pathol. 1991; 3:11-24.

III. Qualité des analyses cytométriques

Les diagnostics et les pronostics de la majorité des cancers requièrent l'examen d'une biopsie de l'organe affecté dans le but d'évaluer les déviations des cellule ou des tissus par rapport à l'organe normal. La décision finale de la présence ou non d'un cancer et de son degré de malignité est réalisée par un pathologiste en observant un échantillon de l'organe suspect. C'est une procédure qualitative, subjective et faiblement reproductible. La caractérisation d'une lésion peut être discrète ou continue. La cytologie quantitative avait pour but de rendre les analyses objectives, quantitatives et reproductibles. L'analyse des échantillons tumoraux, caractérisés par de nombreux paramètres quantitatifs, a pu s'appuyer sur des outils statistiques très puissants. Ces techniques ont permis des progrès considérables dans le diagnostic et le pronostic des cancers. Malgré tout, le pourcentage de mauvais diagnostic, en particulier pour l'analyse de l'ADN, est toujours trop élevé. Une des solutions passe par l'amélioration, à tous les niveaux de la *qualité* des analyses.

III.1. Les limites des statistiques

L'augmentation considérable de la puissance des ordinateurs, couplée au développement d'outils d'analyse des données de plus en plus complexes et de plus en plus abordable pour les non-spécialistes a donné lieu à des situations parfois paradoxales. En effet, cette course effrénée à la quantification, ces flux de données semblent s'effectuer en oubliant certaines notions essentielles concernant l'échantillonnage.

Le développement d'outils statistiques comme les réseaux de neurones, laissaient supposer qu'il suffirait de fournir une multitude de données à un ordinateur pour que celui-ci fournisse, par exemple, une estimation de l'espérance de vie d'un patient, en fonction d'un certain nombre de paramètres quantitatifs.

Or on pensait que la quantification serait moins sujette aux erreurs que les jugements qualitatifs. A la subjectivité de l'observation humaine, la cytométrie a opposé l'objectivité des mesures. A des adjectifs comme grand, gros, petit, allongé, polylobé, la cytométrie oppose de nombreuses données numériques. Mais la cytométrie a également apporté l'imprécision, la variabilité et l' imperfection des mesures et des instruments constituant les systèmes d'analyse d'images.

L'analyse de l'ADN, comme nous venons de le voir, illustre bien les limitations des outils de l'analyse statistique. La qualité d'un diagnostic et d'un pronostic sont dépendant de la qualité de l'analyse proprement dite. Quelle que soit la méthode statistique ou analytique utilisée, si les données sont bruites ou imprécises, le diagnostic et la décision finale seront entachés eux aussi d'une certaine erreur et d'une imprécision plus ou moins grande.

Les tests statistiques employés pour détecter une aneuploïdie ou une hyperploïdie sont fondés sur l'hypothèse que les distributions des différentes sous-populations de l'échantillon, caractérisées par des quantités d'ADN différentes, peuvent être approximées par des lois

normales ou mésa-normales [Weber 1985a]. Bhattacharya [Bhattacharya 1971, 1973] utilisait également ce type d'approximations pour détecter la présence de cellules anormales dans un échantillon. Ces auteurs ont montré également l'insuffisance et les limites de tels tests dans certaines conditions [Weber 1985b] où les hypothèses de départ sont trop contraignantes.

Ces techniques de la statistique classique semblent peu adaptées à l'analyse et au contrôle de grands tableaux de données et ceci pour deux raisons principales :

- l' hétérogénéité des échantillons, situation très fréquente en cytologie [Ferno 1992], n'est pas réellement prise en compte et de ce fait l'interprétation des résultats statistiques est faussée ou biaisée.

- la formulation d'hypothèses et l'ajustement des distributions des variables par des lois de probabilité connues sont des procédés trop lourds et peu compatibles avec une analyse multivariée, dont l'intérêt est justement d'étudier l'information globale contenue dans un tableau de données multidimensionnelles.

Un des principaux problèmes inhérent en cytopathologie quantitative réside dans la difficulté de classifier et de discriminer des spécimens. C'est un niveau supérieur à celui de la classification d'une cellule dans un type ou dans un autre. Deux sortes d'erreurs, au minimum, se produisent dans la classification de spécimens comme des individus. Les faux-positif sont des cellules bénignes classées comme malignes et les faux-négatifs sont des cellules malignes classées comme bénignes. Les courbes ROC [Langley 1985] permettent de visualiser sur un même graphique ces deux types d'erreurs et, en fonction de l'erreur que l'on veut minimiser, modifier les valeurs seuils des facteurs diagnostiques.

A chaque application cytométrique correspond une analyse statistique et des exigences différentes. Par contre toute application exige une qualité optimale de l'analyse proprement dite. Les recherches se sont surtout focalisées, depuis une dizaine d'années, sur le développement et l'utilisation d'outils statistiques divers et de plus en plus sophistiqués. Malgré cela, les résultats sont relativement décevants. Nous venons de voir que le cas de l'analyse de l'ADN, sur lequel de nombreux espoirs s'étaient fondés en est un parfait exemple.

Devant un tel constat, la communauté scientifique a compris la nécessité d'améliorer la qualité et la fiabilité des analyses cytométriques, mais en amont de l'analyse des données.

III.2. Mesure de la qualité

La qualité d'un diagnostic conventionnel est appréhendée en étudiant la correspondance entre le diagnostic donné par un cytopathologiste et la vérité. Définir la qualité d'un diagnostic portant sur l'analyse de l'ADN d'un échantillon tumoral est beaucoup moins évident.

Contrairement à un test antiviral, pour lequel la sensibilité et la spécificité sont facilement mesurables, mesurer la qualité d'un diagnostic d'une analyse cytométrique est impossible. L'analyse de l'ADN ne permet pas des conclusions binaires, comme l'est par exemple la détection d'anticorps. De plus, seule une analyse de toute la tumeur (non seulement de l'ADN

mais également du caryotype) permet de dire si un échantillon diagnostiqué comme aneuploïde l'est vraiment.

La qualité du diagnostic ou du pronostic d'une analyse cytométrique se situe donc au niveau de la fiabilité, de la reproductibilité et de la standardisation de ces analyses. La Figure III.4 illustre comment les différentes étapes de l'analyse influent les unes sur les autres, et en particulier sur le diagnostic et le pronostic final. La "vérité" ne pouvant être connue, il s'agit donc d'optimiser la fiabilité et la qualité des différentes étapes de l'analyse d'un échantillon tumoral, du prélèvement jusqu'aux méthodes de segmentation.

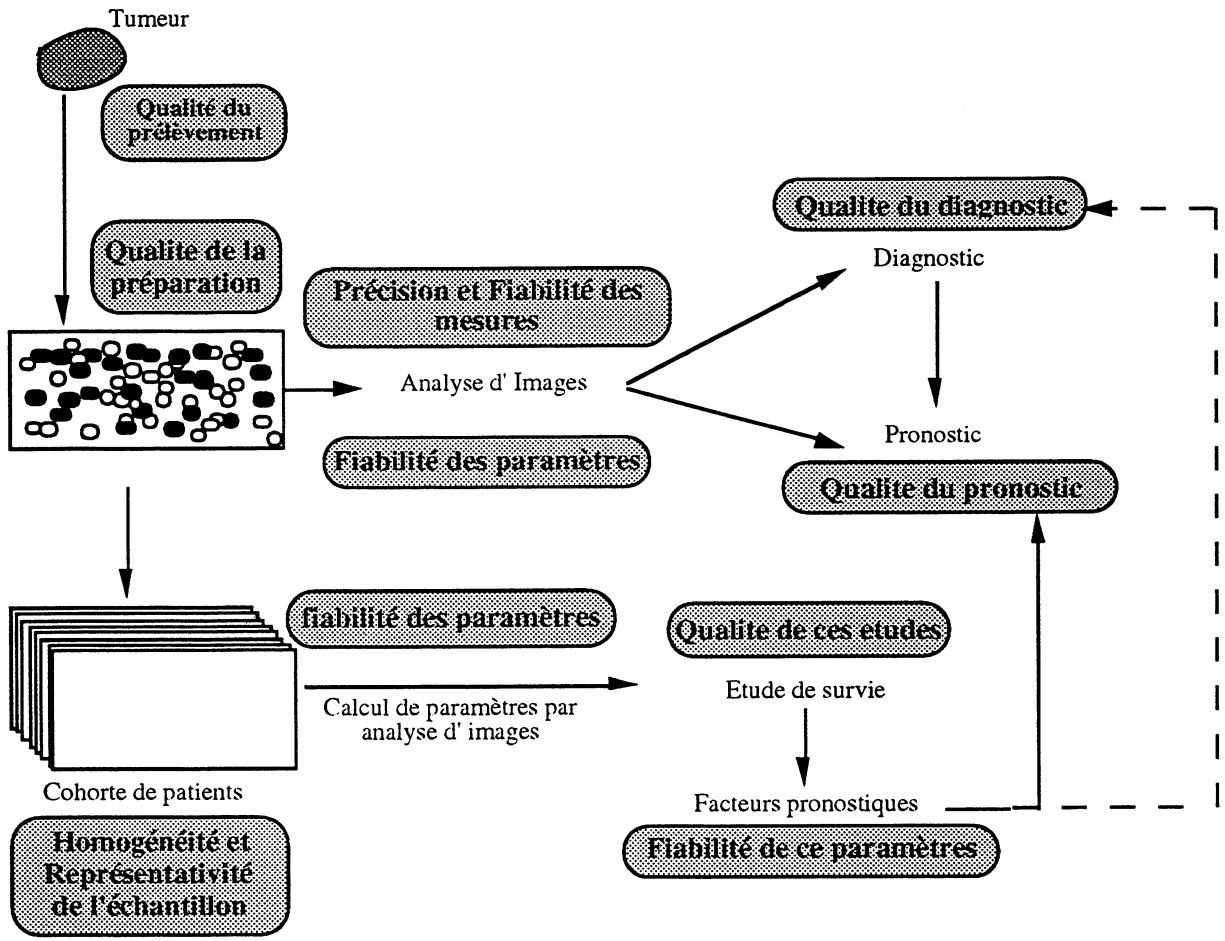


Figure III.4 : Influences des différentes étapes de l'analyse d'images d'échantillons cytologiques sut la qualité du diagnostic et du pronostic des cancers.

III.3. Amélioration de la qualité des analyses : quelques exemples

Le projet PRESS (Prototype Reference Standard Slides) a pour objectif de standardiser les différentes étapes de l'analyse de l'ADN d'un échantillon tumoral par analyse d'images [Giroud 1992] : préparations cytologiques, contrôle de qualité des instruments, normalisation de la calibration des instruments, standardisation des lames microscopiques, harmonisation de la décision médicale.

Wittekind [Wittekind 1985] a proposé une revue très complète les techniques de standardisation des colorations et des marquages des éléments cellulaires pour faciliter la reconnaissance des cellules

Certains ont proposé des tests pour effectuer la calibration de systèmes d'analyse d'images [Lavia 1989 ; Mc Eachron 1990 ; Sanchez 1989]. Les test développés concernent en général le contrôle de :

- la linéarité de la réponse du système par utilisation de filtres neutres de densités connus,

- la reproductibilité de la mesure en effectuent des mesures répétitives sur le même échantillon communément placé au centre du champ,

- le "shading" en mesurant la densité optique d'un objet positionné en différents endroits du champ, ou en mesurant la densité optique de filtres neutres à travers des fenêtres centrales et périphériques,

- la stabilité de la réponse du système en déterminant l'évolution de la densité de filtres de densités connues au cours du temps,

- la distorsion géométrique en mesurant le diamètre apparent de billes en latex parfaitement rondes positionnées à différents endroits du champ,

- les erreurs dues au 'glare' estimées grâce à des billes de diamètre très homogène ou grâce à une préparation de billes à base de colorant Oil Red O.

PRESENTATION DES TRAVAUX

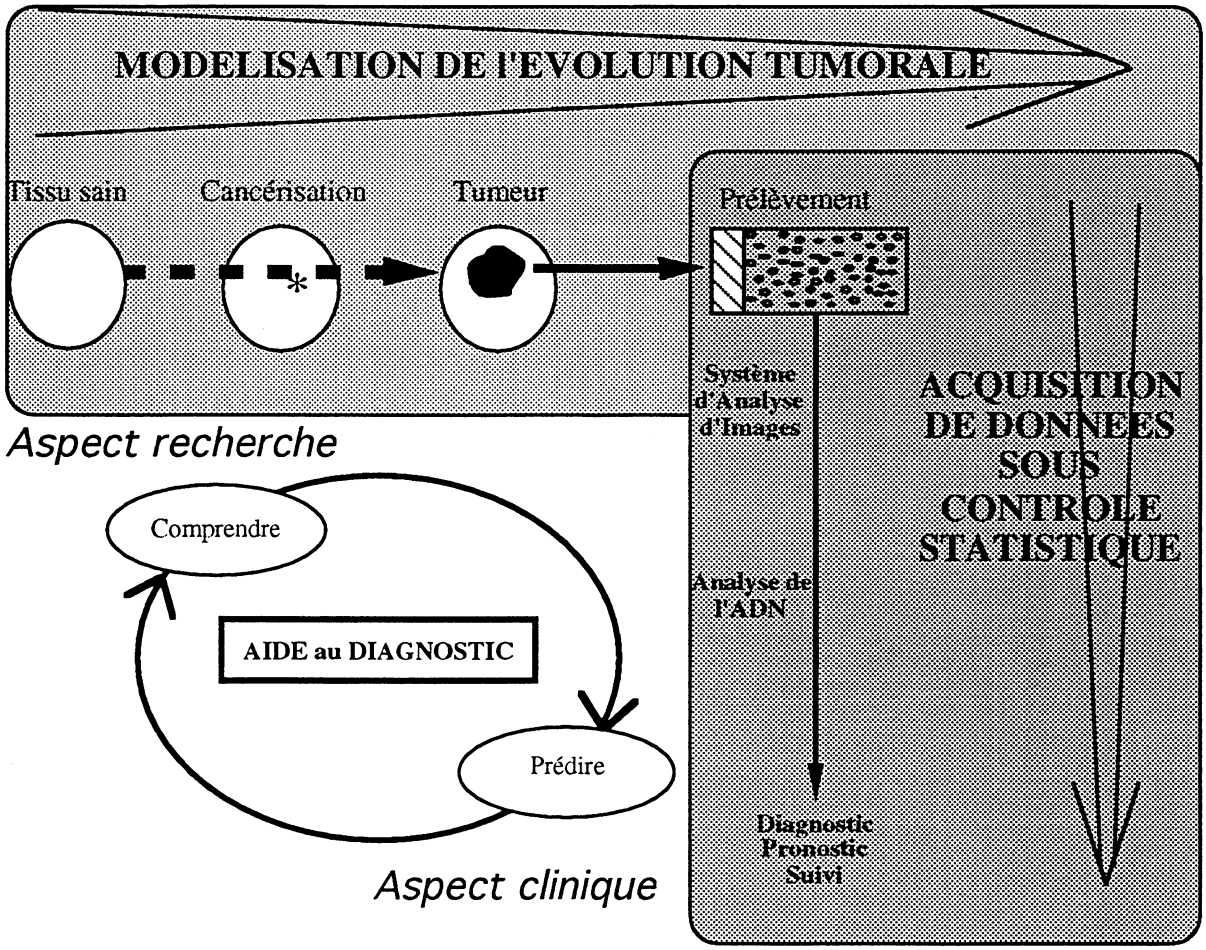
Confrontés aux différents axes de recherches possibles, mis en évidence dans la première partie, en vue d'améliorer la qualité des analyses cytométriques et en particulier celle de l'analyse de l'ADN, nous avons privilégié deux approches bien distinctes, schématisées sur la figure ci-dessous.

La première approche a consisté a mettre au point une méthode d'acquisition de données sous contrôle statistique. Une telle méthode a pour but, non pas de conclure si un échantillon est représentatif ou non de la population tumorale mais de déterminer si un échantillon de n cellules peut être considéré comme stable, lorsque l'on accroît le nombre d'observations, au niveau de l'information qu'il contient. La qualité de l'analyse cytométrique sera donc fonction de la fiabilité de l'analyse de l'échantillon de chaque spécimen analysé. L'application principale concerne l'analyse de l'ADN mais la méthode développée peut également s'appliquer à des problèmes de classification et de reconnaissance de groupes cellulaires.

La deuxième approche se veut plus fondamentale et analytique. Dans le but d'appréhender plus facilement les problèmes de la représentativité d'un échantillon par rapport à la tumeur totale, un modèle d'émergence et de croissance tumorale a été développé. Ce modèle permet comparer les caractéristiques biologiques d'un échantillon par rapport à la population totale (dont les caractéristiques sont, contrairement à la réalité, connues). Les nombreux problèmes de représentativité, de variabilité des échantillons, de significativité des paramètres statistiques des histogrammes d'ADN, et des erreurs introduites par les systèmes d'analyse d'images peuvent être facilement étudiés. La comparaison entre le contenu en ADN et la ploïdie est également possible.

La confrontation des résultats de simulations avec les résultats expérimentaux offrira sans doute des informations pertinentes sur les processus mis en jeu dans l'émergence et l'évolution des cancers.

Un des aspects les plus intéressants du modèle est la possibilité de suivre, d'étudier et de comparer l'évolution biologique d'une tumeur avec l'évolution du contenu en ADN d'un échantillon de cette tumeur. Une meilleure connaissance des différences entre les caractéristiques réelles de toute la tumeur et l'image partielle et imparfaite qu'en donne les histogrammes d'ADN, est en effet indispensable pour améliorer le diagnostic et le pronostic tumoral.



ACQUISITION DE DONNEES CYTOMETRIQUES SOUS CONTROLE STATISTIQUE

Travaux

- **Publication 1**
  Histograms Analysis by use of L-moments linear functions of order statistics
  Statistique et Analyse des données ; 16 : 85-106, 1991
- **Publication 2**
  Cytometric data acquisition under statistical control : use of L-moments variances as stability parameters of DNA histograms.
  Soumis à Analytical Cellular Pathology
- **Publication 3**
  Apport de la méthode du bootstrap pour l'élaboration d'un critère d'arrêt en reconnaissance des formes.
  Innov. tech. Biol. Med ; 11 : 544-558, 1990

Comme nous l'avons démontré dans la première partie, aux problèmes classiques de l'analyse des données s'ajoutent, en routine clinique, le problème de coût et ceci indépendamment de l'analyse effectuée. Le problème de la qualité de l'analyse dans ce contexte précis est d'autant plus considérable. En recherche fondamentale, les expériences peuvent être multipliés et reproduites plusieurs fois. En routine clinique, un seul échantillon du tissu ou de la tumeur est disponible. Il est donc nécessaire de développer un outil permettant de mesurer la qualité d'une analyse. La méthode employée pour suivre en temps réel l'évolution d'histogrammes d'ADN s'applique également au contrôle de la stabilité d'une population multidimensionnelle caractérisée par les valeurs propres d'une Analyse en Composantes Principales. Ce contrôle statistique est basé sur la méthode du bootstrap, une des nombreuses techniques de ré-échantillonage qui tirent profit de la puissance des ordinateurs.

I. Les techniques de ré-échantillonage : la méthode du Bootstrap

I.1 Contexte

Une grande partie des méthodes statistiques utilisées actuellement a été élaborée entre 1800 et 1930, à une époque où les calculs étaient longs et coûteux. Comme ces calculs sont aujourd'hui des millions de fois plus rapides et moins coûteux, de nombreuses méthodes (et de théories) mettant à profit l'accroissemnt des vitesses de calcul des ordinateurs, ont vu le jour récemment sous le terme générique de calculs "intensifs".

La question posée par toute théorie statistique est la suivante : comment déterminer la véracité d'un résultat ? A cette question les outils statistiques donnent une indication quantitative de la fiabilité d'une estimation. Comme les observations empiriques sont toujours sujettes à erreur, les conclusions scientifiques s'appuient généralement sur des évaluations statistiques de la vérité. Par conséquent, toutes méthodes donnant des évaluations statistiques plus précises ou portant sur des données plus variées, méritent d'être développées.

La Figure 1 illustre les avantages des nouvelles méthodes statistiques, fondées sur des calculs massifs, par rapport aux anciennes. Avec les anciennes méthodes, il fallait émettre un certain nombre d'hypothèses sur les données avant de les analyser comme par exemple l'hypothèse de normalité de la distribution des données. L'expérience montre que la théorie de la loi Normale (ou théorie de Gauss) donne des estimations fiables même lorsque la distribution ne suit qu'approximativement une loi normale. Par contre, lorsque ces données ne satisfont réellement pas aux conditions de normalité, les résultats obtenus sont beaucoup moins sûrs.

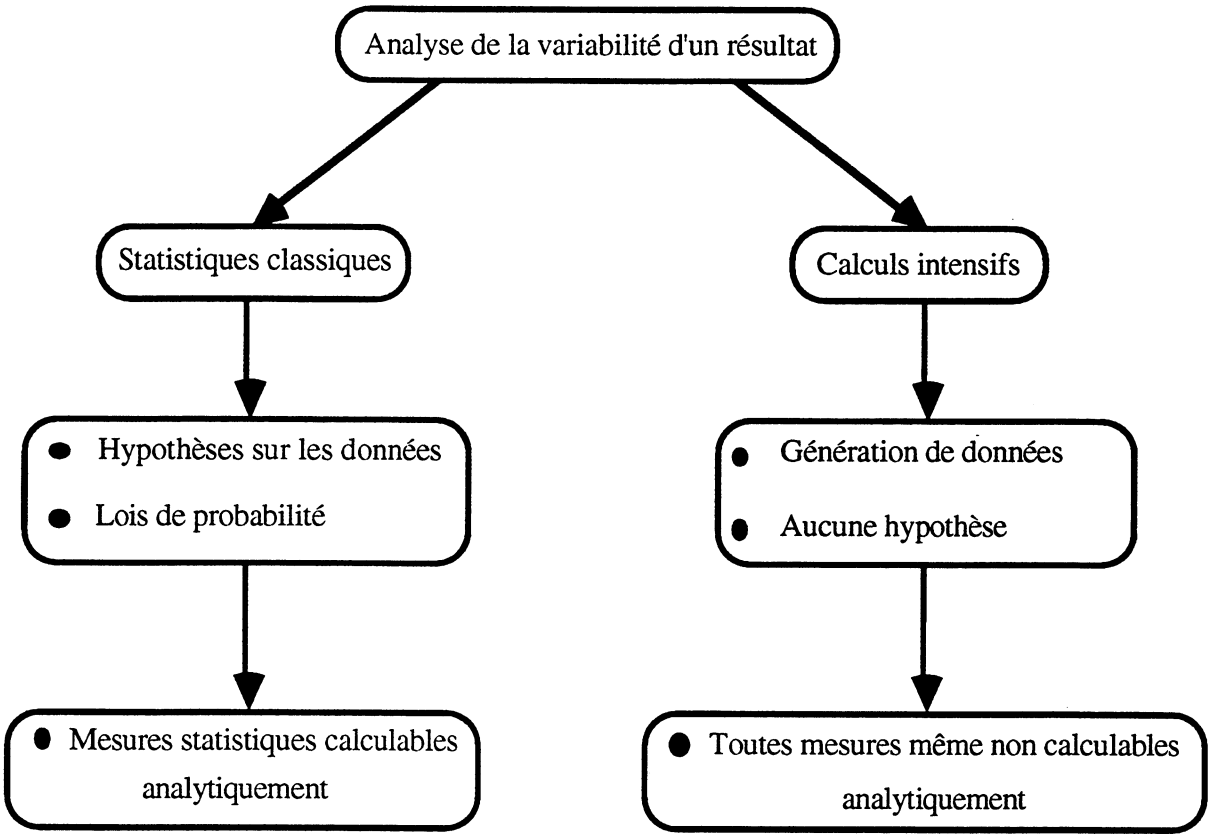


Figure 1 : Comparaison entre les méthodes statistiques classiques et les méthodes de calculs statistiques intensifs sur ordinateur.

Les méthodes de calculs intensifs résolvent de nombreux problèmes sans qu'il soit nécessaire de supposer que la distribution soit normale.

Cette indépendance vis à vis de la loi de Gauss est un progrès considérable, mais le second avantage des nouvelles techniques donne encore plus de liberté. En effet, la théorie statistique classique ne porte que sur quelques propriétés des échantillons comme la moyenne, ou l'écart-type facilement manipulables analytiquement. Les méthodes de calculs intensifs permettent d'aborder et d'étudier des estimations de fiabilité de conclusions scientifiques beaucoup plus complexes et hors de portée d'une analyse mathématique classique. Une de ces méthodes, la méthode Bootstrap permet d'estimer, par exemple la variabilité et la précision [Efron 83] :

- des lignes de contour sur une carte;
- des composantes principales issues d'une ACP;
- d'un modèle énergétique obtenu par la méthode des moindres carrés;
- de certaines prévisions médicales.

Ces méthodes de calculs intensifs permettent donc de "penser l'impensable" comme l'exprime Efron [Efron 1979a] ; l' impensable étant d'effectuer des milliers de fois une analyse plutôt que d'émettre une quelconque hypothèse.

I.2 Les différentes techniques de ré-échantillonnage

De nombreuses méthodes, fondées sur la puissance des ordinateurs sont actuellement disponibles : chacune d'elles engendre des séries de données artificielles à partir des données originelles et évalue la variabilité réelle d'un paramètre statistique par sa variabilité pour l'ensemble de ces séries de données. Ces méthodes diffèrent entre elles par leur façon d'engendrer les échantillons artificiels.

Dans deux articles de référence [Efron 1979 a,b], Bradley Efron, qui est l'un des pionniers du développement de ces nouvelles méthodes statistiques, décrit et compare certaines d'entre elles : le jacknife, la validation croisée, la duplication par blocs et le bootstrap.

I.2.1 La méthode du Jacknife ou "canif suisse"

Cette technique introduite par Quenouille (1949) et Tukey (1958), joue en fait un double rôle : réduire ou supprimer le biais d'un estimateur biaisé et fournir une estimation approchée mais robuste d'un intervalle de confiance autour de toute estimation. A la suite de travaux de Miller [Miller 1974] qui a résumé les différents développements et a cité de nombreuses applications de cette méthode, il est utile de décrire le principe de base du Jacknife, que l'on peut considérer comme l'ancêtre des nouvelles méthodes statistiques.

Soit ($X_1, X_2, ..., X_n$) un échantillon aléatoire de la variable parente X. Imaginons que l'on ait à estimer un paramètre θ pour lequel on utilise l'estimateur noté :

$$Y_{tot} = f(X_1, X_2, ..., X_n)$$

L'échantillon de taille n est divisé arbitrairement en k parties disjointes, chacune d'elles contenant m = n / k individus. On appelle Y_{-j} le résultat du calcul effectué sur l'échantillon amputé de la j-ième partie (c'est-à-dire constitué des k - 1 parties autres que la j-ième) :

$$Y_{-j} = f(\text{ les } X_i \text{ sauf le groupe j })$$

(En général, le groupe j n'est constitué que d'un seul individu)

On appelle pseudo-valeurs, d'après TUKEY, les k différences pondérées :

$$\tilde{Y}j = k\,Y_{tot} - (k-1)Y_{-j} \qquad (j = 1, 2, ..., k).$$

L'estimateur de Quenouille-Tukey est la moyenne des pseudos-valeurs :

$$\tilde{Y} = \frac{1}{k}\sum \tilde{Y}_j = kY_{tot} - (k-1)\left\{\frac{1}{k}\sum Y_{-j}\right\}$$

La variance de Quenouille-Tukey est un estimateur de $\tilde{Y}$ (et de Y_{tot}) :

$$\tilde{S}^2 = \frac{1}{k}\left\{\frac{\Sigma(\tilde{Y}_j - \tilde{Y})^2}{k-1}\right\}$$

La méthode du Jacknife s'est avérée efficace dans de nombreux contextes. Il est maintenant reconnu cependant qu'il faut éviter de l'utiliser pour des statistiques dont la distribution présente une "extrémité abrupte" ainsi que pour les distributions fortement dissymétriques ou ayant des extrémités très dispersées.

I.2.2 Les méthodes de validation croisée et de duplication par blocs

La méthode de validation croisée par blocs est surtout employée pour le calcul de courbes ou l'ajustement d'un polynôme par la méthode des moindres carrées. Le principe de cette méthode est simple : on divise l'ensemble des données en deux moitiés dont l'une est laissée de côté ; on calcule les courbes pour la première moitié et on les teste successivement dans la deuxième pour obtenir la meilleure approximation. On peut faire des blocs très inégaux et effectuer la validation croisée plusieurs fois en séparant aléatoirement les données. Le test final qui est la validation croisée, indique de façon fiable dans quelle mesure la courbe optimisée peut prédire les valeurs associées à de nouvelles données.

La méthode des duplications successives par blocs fait appel à des techniques de division plus systématiques, la taille des blocs étant déterminée de façon à calculer la variabilité d'échantillons d'enquêtes et de recensements.

I.2.3 La méthode du BOOTSTRAP

La méthodesdu bootstrap a été mise au point en 1977 par Bradley Efron [Efron 1979b]. Cette méthode, très simple dans son principe, est si dépendante de la puissance de l'ordinateur qu'elle était inapplicable il y a 30 ans. Cette procédure, qu'on appelle également "test du Bootstrap", permet d'évaluer la précision statistique d'un paramètre ou de tout autre résultat scientifique à partir des données d'un seul échantillon. Le principe consiste à créer, à partir de l'échantillon original, un grand nombre d'échantillons par tirages aléatoires. Sur chacun de ces échantillons, le paramètre étudié est calculé afin de fournir une estimation de la valeur calculée pour l'échantillon original. Cette estimation est souvent donnée par un intervalle correspondant à la distribution simulée par bootstrap.

La méthode bootstrap s'est révélée très performante et la précision statistique qu'elle donne est fiable dans un grand nombre de situations. On peut par exemple citer le cas exemplaire décrit par Diaconis [Diaconis 1983] sur l'estimation d'un coefficient de corrélation mesuré entre deux séries de notes de 15 facultés de droit ; la précision obtenue par bootstrap s'est avérée meilleure que la précision obtenue par la méthode du jacknife et presque aussi

bonne que celle obtenue par approximation par la loi normale. De plus, dans cet exemple, l'hypothèse de normalité de la distribution des coefficients de corrélation est certainement fausse.

S'il existe toujours quelques situations pour lesquelles cette méthode ne donne pas de bons résultats, ceci n'est qu'un rappel de l'incertitude qui obère toute analyse statistique.

La méthode du bootstrap a été largement développée ces dernières années. De nombreux mathématiciens se sont intéressés à cette méthode et ont montré que, pour un problème tel que l'estimation de la précision de paramètres simples (comme le coefficient de corrélation) et pour lequel la réponse exacte était connue, les estimations se sont révélées correctes.

De nombreuses publications traitent des propriétés théoriques de la méthodologie du Bootstrap. Banks [Banks 1988b] et Hinkley [Hinkley 1988] présentent une comparaison des variantes de la méthode bootstrap comme par exemple (liste non-exhaustive] :

- le "smoothed " bootstrap qui génère des échantillons à partir d'une transformation "lissée " de la distribution empirique initiale,

- le "bayesian " bootstrap [Rubin 1981] qui génère les échantillons par des tirages non-aléatoires mais en affectant aux individus de l'échantillon initial des poids issus d'une distribution de Dirichlet,

Toutes ces variantes cherchent à réduire les biais éventuels et les erreurs statistiques des estimateurs calculés par la méthode classique du Bootstrap. Ces biais peuvent avoir diverses origines selon la nature des statistiques étudiées. On peut citer entre autre les travaux de Lo [Lo 1987, 1988], de Davidson [Davidson 1986], de Silverman [Silverman 1987] et de Banks [Banks 1988a]. L'importance du nombre d'échantillons Bootstrap à générer dans la précision des estimations Bootstrap a fait également l'objet de nombreuses études [Johns 1988 ; Do 1991]. Ce nombre, d'après Efron et Tibishirani, doit être au moins supérieur à 250 [Efron 1985].

D'autres auteurs ont démontré également la validité de la méthode du Bootstrap appliquée à des problèmes plus complexes. Béran [Beran 1985], par exemple, illustre l'intérêt du bootstrap pour estimer les valeurs propres d'une matrice de covariance. La stabilité du modèle de régression de Cox a également été étudié par le Bootstrap [Altman 1989].

Efron [Efron 1979a et b] a décrit en détail les fondements et les relations théoriques existant entre les différentes techniques de ré-échantillonnage qui viennent d'être décrites.

Parmi toutes les variantes de la méthode du Bootstrap existantes, nous avons volontairement opté pour la méthode de base, dite classique, dont le principe est décrit en détail dans les publications 2 et 3. L'étude de ces différentes variantes fera l'objet de travaux ultérieurs à cette thèse.

II. Acquisition d'histogrammes d'ADN sous contrôle statistique

La démarche choisie pour le contrôle statistique de l'acquisition des histogrammes d'ADN est la suivante :

- Recherche des meilleurs descripteurs de l'histogrammes (codage des histogrammes),
- Application de la méthode du bootstrap sur ces descripteurs dans le but de mesurer la variabilité de ces descripteurs pour un échantillon de n cellules. La valeur des variances Bootstrap des descripteurs, donne une estimation de la stabilité des histogrammes d'ADN.
- La stabilisation des variances Bootstrap, quand le nombre d'individus analysés augmente, est utilisée comme critère de stabilité des histogrammes d'ADN.

II.1 Codage des histogrammes d'ADN

Les résultats obtenus précédemment ont mis en évidence les limites des descripteurs classiques des histogrammes d'ADN, comme l'entropie, le 2cDI, l' Index d'ADN moyen, ou l' Index de Prolifération pour décrire un histogramme. Un histogramme ne peut être résumé par un seul de ces paramètres, ni même par une combinaison de ces paramètres. Par contre, nos études ont mis en évidence le rôle de l'observation visuelle des histogrammes tant au niveau du diagnostic qu'au niveau du pronostic.

Nous nous sommes donc orientés vers des descripteurs représentatifs de la forme de l'histogramme. Il s'agissait de décrire le plus fidèlement possible la distribution de la variable étudiée (quantité d'ADN) tout en réduisant le nombre de données. Pour cela, nous nous sommes inspirés des travaux de Weber [Weber 1987] qui avait utilisé les déciles de la fonction de distribution pour coder les histogrammes d'ADN. Une des grandes limitations de cette méthode réside dans la significativité des déciles pour décrire parfaitement l'histogramme d'ADN.

Les moments statistiques classiques (moyenne, variance, moment d'ordre 3, 4 etc...) étaient également des candidats potentiels au codage des histogrammes. Cependant, la difficulté d'interpréter les moment d'ordres élevés, le problème du choix du nombre de moments nécessaires pour coder des histogrammes souvent très irréguliers et leur trop forte sensibilité à l'égard des valeurs extrêmes n'ont pas pesé en leur faveur.

Ce sont d'ailleurs ces mêmes critères qui ont dicté le choix des L-moments, comme descripteurs des histogrammes. Les L-moments sont des combinaisons linéaires de statistiques d'ordre. Hosking [Hosking 1986, 1989, 1990], qui a considérablement étudié les L-moments, tant au niveau théorique qu'au niveau de leurs utilisations pratiques, décrit en détail leur avantages par rapport aux moments classiques, en particulier pour décrire des distributions dont les fonctions analytiques n'existent pas.

Le tableau 1 résume les avantages et les limites des trois types de descripteurs candidats potentiels au codage des histogammes d'ADN.

Descripteurs	Avantages	Limites
Moments classiques	Simples à calculer Réduction des données	Nombre nécessaire pour décrire l'histogrammes ? Trop sensibles aux valeurs extrèmes
Descripteurs des histogrammes	Simples à calculer Réduction des données	Significativité ? Fiabilité? Perte d'information Représentativité de l'histogramme ?
L-moments	Réduction des données Peu sensibles saux valeurs extrèmes Caractéristiques de la distribution Reconstruction possible de l'histogramme Proches de l'interprétation visuelle	Temps de calcul ? Interprétation des moments d'ordres élevés ? Approximations discrètes.

Tableau 1 : Caractéristiques des différents candidats au codage des histogrammes d'ADN

Le principe et les résultats du codage des histogrammes d'ADN par les L-moments a fait l'objet de la publication suivante :

Histogram analysis by use of L-moments, linear functions of order statistics Statistique et Analyse des données ; 16 : 85-106, 1991

HISTOGRAM ANALYSIS BY USE OF L-MOMENTS, LINEAR FUNCTIONS OF ORDER STATISTICS.

Martial Guillaud & Jean-Marc Chassery

Equipe RFMQ. Laboratoire TIM3 - IMAG - USR B 00690
Université Joseph Fourier, CERMO, BP 53X, 38041 Grenoble Cedex - France
Phone: (33) 76514813 / Fax : (33) 76514948

Abstract

It is current statistical practice to summarize observed data by the moments or cumulants of the distribution. In comparison with conventional moments, L-moments, linear combinations of an ordered data set, are of considerable interest in defining an experimental data set because no assumption is made concerning the probability distribution of the data. The aim of this publication is to demonstrate the advantages as well as to explain the properties of L-moments as features of histograms. We use a pattern recognition approach in order to interpret and analyse histograms which are defined in a new vectorial space which is determined by Legendre polynomials.

Résumé

La description de données expérimentales par les moments ou les cumulants de la distribution sous-jacente est une pratique statistique classique. Par rapport aux moments conventionnels, les L-moments, combinaisons linéaires de l'ensemble des données ordonnées, sont d'un intérêt certain pour décrire les données sans faire d'hypothèses sur la loi de distribution. Dans cet article, nous montrons l'intérêt et les propriétés des L-moments comme descripteurs d'un histogramme, représentation graphique de données expérimentales. L'analyse et l'interprétation de ces histogrammes sont abordées comme un problème de reconnaissance des formes, dans le but de décrire ces histogrammes dans un nouvel espace de primitives engendré par les polynômes orthogonaux de Legendre.

Key words :
Pattern recognition. Histogram. Legendre polynomials. L-moments. High order statistics.

Classification AMS : 62 G 05
Classification STMA : 04 180, 04 080

HISTOGRAM ANALYSIS USING L-MOMENTS, LINEAR FUNCTIONS OF THE ORDER STATISTICS.

It is current statistical practice to summarize observed data by moments or cumulants. In comparison with conventional statistical moments, L-moments are of considerable interest to define an experimental data set, because no assumption need be made on the probability distribution of the data [HOS90]. Furthermore, vectorial data is normally defined in reference to a polynomial transformation. This publication reports on the use of L-moments and certain properties of Legendre orthogonal polynomials for the analysis of histograms. Several examples will be given.

Image analysis of DNA ploidy pattern has become of vital diagnostic interest since the discovery in the 1950s that tumour cells possess elevated quantities of DNA. DNA histograms are classical representations of the cellular DNA quantities in a tumour sample. The development of numerous statistical methods to interpret these histograms has progressed since 1980. The majority of these methods are based on the computation of a single parameter such as the index or percentage of cells containing a given quantity of DNA [OPF87]. To date, however, the analysis of biological samples by the statistical methods presently available does not necessarily correlate with the pathological status. This paper proposes a new statistical approach, based on the theoretical results of Hosking [HOS90], in order to improve DNA histogram interpretation. We propose to replace univariate distribution by the corresponding set of L-moments, which can subsequently be used in factorial analysis

I.INTRODUCTION

The L-moments, analogous to conventional moments, can be estimated by linear combinations of an ordered data set, i.e by L-statistics. These L-moments have the theoretical advantage over conventional statistical moments in that they characterize a wider range of distributions. Moreover, they are robust when outliers are present in the sample. Greenwood et al.[GRE79] defined the probability of weighted moments of a variable in terms of order statistics from a random sample of size n. However, these moments are less adapted for use in the pattern recognition approach than L-moments, and they have therefore not been used in this study. On the other hand, Hosking [HOS86] developed a unified theory covering the characterization of statistical distributions, the representation of data samples, the fitting of the data to probability distributions with data and an hypothesis for testing the fitted distributions. We use some of his theoretical results to extract certain features from the DNA distributions and to study their properties.

By definition, a histogram is dimensioned by a certain number of classes. In this study histogram analysis is treated as a problem of pattern recognition. Since it is usual to define vectorial data by a polynomial transformation, we use Legendre orthogonal polynomials to define the histogram in a new vectorial space.

Pattern recognition can be divided into two steps [YOU74]. In the first step the most prominent features of the histogram are extracted. The second step concerns the classification of this histogram based on the new features. With q L-moments, it is possible to reconstruct the original histogram using the reduction property of the Legendre polynomials [SIL69]. The quality of the reconstructed histogram is then compared with the original. The two first conventional moments are sufficient to describe a Normal distribution pattern, but when dealing with an unknown probability density, the conventional moments are of limited value in defining the distribution.

II.DEFINITIONS AND ALGORITHMS

II.1 Preliminary Notations

Let X be a continuous random variable with a distribution function F. Let x_i be the value of the variable X for different objects i, for i = 1, ..., N. Let x(F) be the quantile function or inverse distribution function of X. x(F) is defined by N points. Let f be the notation of the original histogram. The component f(i) is the relative frequency of objects belonging to class i, for i = 1, ..., K, where K is the number of classes by which the histogram is defined. g(i) is defined as the relative frequency of objects belonging to class i for a reconstructed histogram.

II.2 L-moments : definitions and properties

This method (illustrated in flow diagram 1) parametrizes an original distribution function by its L-moments and this distribution function is reconstructed using only a few of these L-moments.

* first, the L-moments are computed [HOS89] :

$$\lambda_r = \int_0^1 x(F)\ P^*_{r-1}(F)\ dF, \quad r = 1, 2, \ldots, \quad (1)$$

$P^*_r(x)$ is the *rth* shifted Legendre polynomial, derived from the Legendre polynomial $P_r(x)$,and given by $P^*_r(x) = P_r(\frac{x+1}{2})$. The Legendre polynomials are orthogonal for the interval [0,1] with unit weight function. These polynomials can be used to define vectorial data by a polynomial transform. The selection of the number of L-moments will be developed below.

** then using the inversion theorem, proposed by Sillitto [SIL69], x(F) is approximated by a quantile function x*(F) given in terms of L-moments by equation (2) :

$$x^*(F) = \sum_{r=1}^{\infty} (2r-1)\lambda_r P^*_{r-1}(F) \quad 0<F<1 \quad (2)$$

x*(F) is convergent to x(F) in mean square sense, i.e.

$$R_s(F) \equiv x(F) - \sum_{r=1}^{s} (2r-1)\lambda_r P^*_{r-1}(F)$$

the remainder after stopping the infinite sum after s terms, satisfies

$$\int_0^1 \{R_s\}^2 dF \rightarrow 0 \text{ as } s \rightarrow \infty \qquad (3)$$

This theorem is valid for discrete random variables, provided that the quantile function is normalized, i.e. that

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{2}\{(x(F+\varepsilon) + x(F- \varepsilon)\} = x(F) \text{ for all } F \in (0, 1)$$

*** finally a new distribution function F* can be generated from the quantile function x*(F). The relative frequency of each class i can be estimated from this distribution by the integration between two values a and b, which are extremities of the class i. Since a density function is the derivative of the distribution function, it follows that :

$$g(i) = F^*(a) - F^*(b) \quad \text{for } i = 1 \rightarrow K$$

II.3 Specific algorithm

The L-moment properties described above are adapted to the present problem; but since the analytical formula of the quantile functions are unknown, we use numerical computing algorithms to resolve the different equations.

- Let X be a continuous random variable, with N observations $x_1, x_2, \ldots, x_N$.
- Let F be the distribution function of the variable X, defined on N points.
- Let f be a histogram, a graphical representation of the variable X, represented by f(i) for i = 1,..., K, with K classes.
- empirical distribution function F can be plotted

$$F(x) = Pr(X \leq x)$$

$$F(x_j) = \sum_{z=1}^{j} Pr(x_z), \text{ for } j = 1 \text{ to } N$$

- and q L-moments (Simpson method) be computed:

$$\lambda_r = \int_0^1 x(F)\ P^*_{r-1}(F)\ dF, \quad r = 1, 2, \ldots, \quad (1)$$

From the N values of the empirical distribution function F, this integral is computed by means of the Simpson's algorithm.

- Restoration phase :
  - compute $x^*_q(F)$ the reconstructed quantile function by using the q first L-moments :

$$x^*_q(F) = \sum_{r=1}^{q} (2r-1)\lambda_r P^*_{r-1}(F) \quad 0<F<1 \qquad (4)$$

- plot the approximated distribution function curve $\mathbf{F_q^*}$ from $\mathbf{x_q^*(F)}$,

- compute $\mathbf{g_q(i)}$ the restored histogram with the q first L-moments, from $\mathbf{F_q^*}$

$$\mathbf{g_q(i) = F_q^*(a) - F_q^*(b) \quad for\ i = 1 \rightarrow K} \quad \mathbf{(5)}$$

with a and b extremities of the class i.

F*(a) and F*(b) are computed by dichotomy.

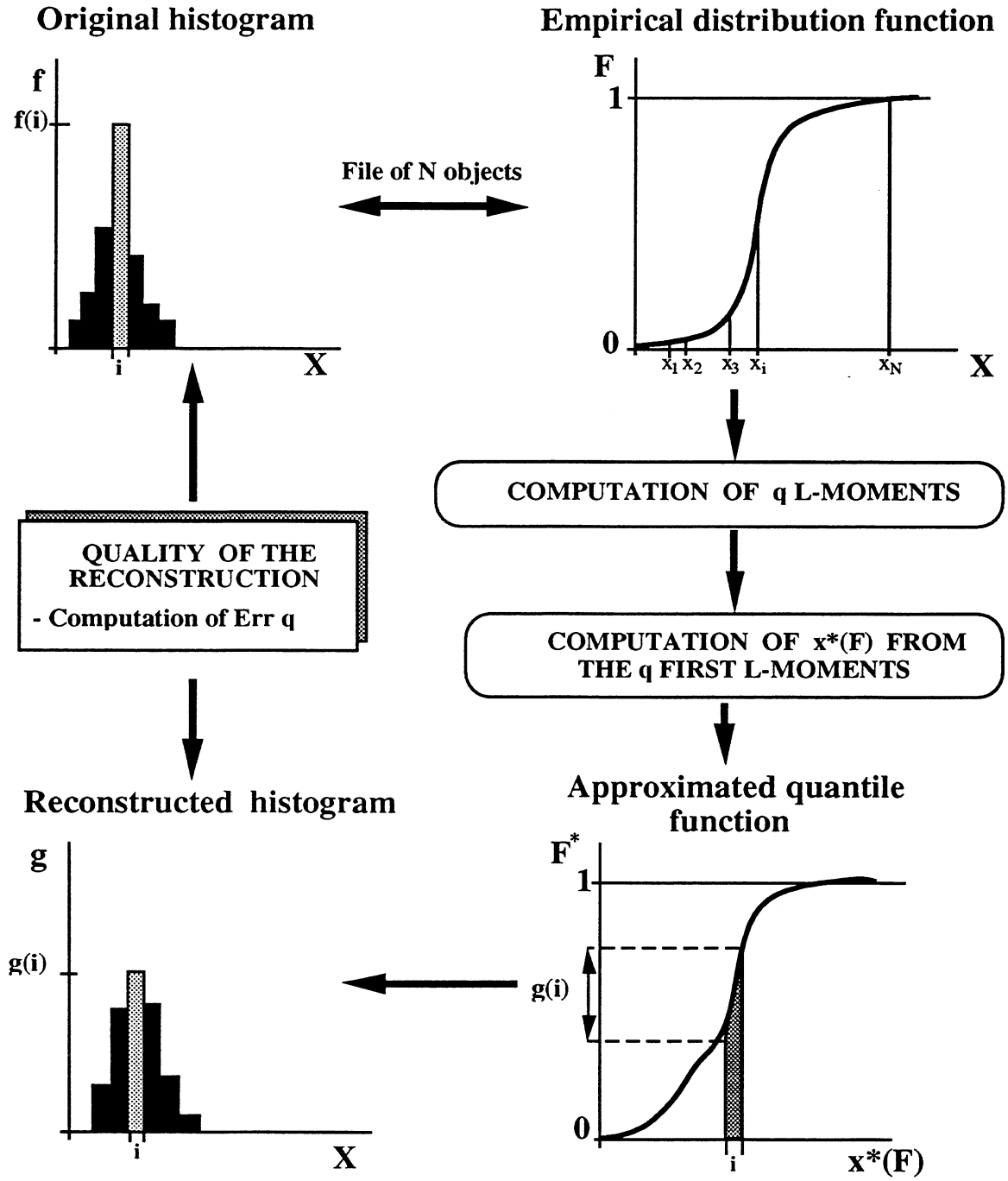
Remarks : This algorithm permits the reconstruction of the original distribution function F, from the formula (4). However, we have chosen to study the reconstruction of histograms using formula (5). In the present biological situation, distributional analysis is less informative than the form of the histogram itself. In this respect, the quality of histogram reconstruction depends entirely on the characteristics of the original distribution function F : a histogram is merely a graphical representation of the sample.

II.4 Choice of the number of L-moments

The choice of the number of L-moments needed to describe the distribution F depends on the quality of the inversion. The process is terminated when a criterion error, called Err, reaches a minimum. The most obvious criterion is given by relation (3). However, Hosking leaves a number of questions unanswered about the convergence of Rs(F) to 0. Thus we prefer to compute a stop criterion based directly on the quality of the reconstructed histogram. We note Err_q as the quadratic error associated with the reconstruction of the original histogram f(i) by $g_q(i)$ using the q first L-moments. For a histogram defined by K classes, Err_q is defined by the relation :

$$\mathbf{Err_q} = \sum_{\mathbf{i = 0}}^{\mathbf{K}} \frac{\mathbf{(f(i) - g_q(i))^2}}{\mathbf{f(i)^2}}$$

The best reconstruction is obtained with q L-moments when Err_q is a minimum. Nevertheless, we are also interested in the evolution rate of the reconstructed histogram in order to study the contribution and the significance of each L-moment. Oja [Oja81] studied this problem for the four first conventional moments using a theoretical approach.



FLOW DIAGRAM 1 : Illustration of the algorithm : The original histogram may be reconstructed by using the q first l-moments in order to study the quality of these features.

III. RESULTS

Some results of using L-moments in synthetic and real situations will now be given. L-moments will be calculated for synthetic distributions with a mixture of normal sub-populations and real distribution of biological populations. Some histograms will be reconstructed by using an ascending number of the L-moments.

III.1 Analysis of normal distribution by L-moments

In the case of a few known continuous distributions Hosking [HOS90] has given some analytical relations between the first four L-moments and the first two conventional moments. These relations are listed in table 1 for a simulated normal distribution of 300 objects. We can verify that the analytical relations remain correct through the use of numeric algorithms. The differences between the computed results and expected results are due to approximation errors of the successive integrations. As expected, these differences decrease as the sample size increases (results not shown).

In Figure 1, the histogram of the former distribution has been reconstructed according to our algorithm. The histogram A is the original, represented on 30 classes. We used successively 1, 2, 3, 4, 5, 7, 11 and 16 L-moments to reconstruct the histograms B, C, D, E, F, G, H and I respectively.

The first-order L-moment gives the location of the distribution, i.e the mean or the median in the case of a symmetric distribution. The second-order L-moment gives the shape of the distribution while the third-order L-moment gives the degree of skewness. In a normal distribution, the third-order L-moment is approximately equal to 0, so the reconstructed histogram is also approximately the same as the former reconstructed with two L-moments. The fourth-order L-moment gives the degree of kurtosis of the distribution. The interpretation of these four L-moments is identical to that obtained from conventional moments. The reconstructed histogram using the first four L-moments is already a good approximation to the original histogram. The contribution of the fifth to the sixteenth-order L-moment (which gives the best reconstruction according to our criterion) is not significant and it is difficult to interpret each one for this normal distribution. Indeed, in this case, for which two first conventional moments are sufficient to define the probability density function, the use of L-moments (16 L-moments !!!) seems to be of little interest. But, when the distribution law is unknown, the use of L-moments to define any kind of distribution is, on the contrary, of prime interest.

Conventionnal moments	**Computed L-Moments**	**Expected L-moments from analytical relations**	
μ = 2007,97	λ_1 = 2006,12	2007,97	($\lambda_1 = \mu$)
σ^2 = 43689	λ_2 = 118,20	117.04	($\lambda_2 = 0.56\sigma$)
	λ_3 = 0.06	0	($\lambda_3 = 0$)
	λ_4 = 13.6	14.42	($\lambda_4 = 0.069\sigma$)

Table1. Comparison between the two first conventional moments (mean μ and variance σ^2) and the two first computed L-moments λ_1 and λ_2 for a simulated normal distribution (mean μ = 2007, standard deviation σ = 209, objects = 300). The first four expected L-moments from Hosking's analytical relations, given in the third column, are similar with the four first computed L-moments.

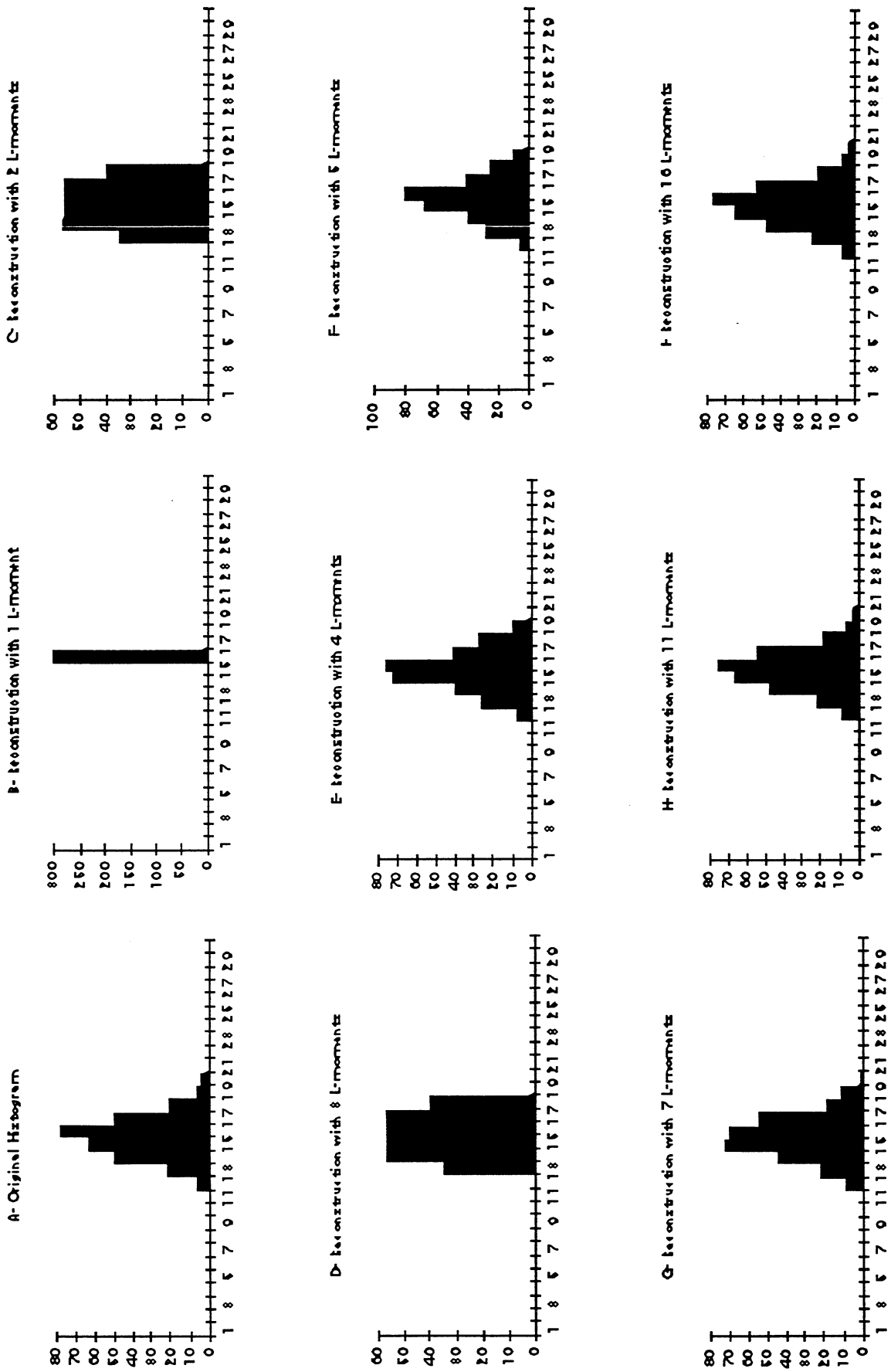


Figure 1 : Reconstruction of an original histogram (A) of 30 classes and 300 objects (Normal distribution, mean = 2000, CV = 10%) with respectively 1, 2, 3, 4, 5, 7, 11 and 16 L-moments. The best reconstruction is obtained by using 16 L-moments (I).

III.2. Detection of sub-populations in some multimodal distributions

Figures 2, 3, 4 and 5 illustrate the analysis of simulated bimodal histograms in which the two sub-populations are weighted differently. The proportion of the two subpopulations, are respectively 0.9/0.1, 0.75/0.25, 0.5/0.5, 0.25/0.75.

In Figure 2, where the initial histogram (Figure 2A) corresponds to the sum of two normal distributions with probability 0.90 and 0.10, the first-order L-moment corresponds to the central tendancy and the second-order L-moment corresponds to the dispersion of the distribution. The third-order L-moment represents the skewness of the distribution whereas degree of kurtosis of the distribution is represented by the fourth-order L-moment. After the sixth-order L-moment, which is not represented here, the principal sub-population is well represented. With eight L-moments (Figure 2G), reconstructed histogram shows a normal population with a long tail to the right. Progressively, with other L-moments up to twelfth-order L-moment, an increasing number of classes to the right of the principal population (which remains stable) are represented. It is necessary to use 19 L-moments (Figure 2I) to obtain the best reconstructed histogram with the second sub-population correctly represented.

In Figures 3, 4 and 5, the evolution of reconstructed histogram is approximatively the same as for the former histogram. Nevertheless, it should be noted that the number of L-moments required to detect the small sub-population varies inversely with the weight of this sub-population. Indeed the subpopulation of probability 0.10, 0.25, 0.50 are detected respectively by using 19, 10, and 5 L-moments (Figures 2I, 3H, and 4F).

Moreover, the original histogram of Figure 5 can be considered to be symmetric when compared to that of Figure 2. The difference between the two is the result of differences in distributions. Indeed the coefficient of variation is the same for the two distributions so the standard deviation varies as a function of the subpopulation mean value. It should be noted that both of these reconstructions evolve similarly, i.e the main population is detected first then the minor population is detected.

The bimodal histogram of Figure 4 indicates that the fifth-order L-moment characterizes the bimodality of the distribution. This observation has also been made by other authors [HOS90].

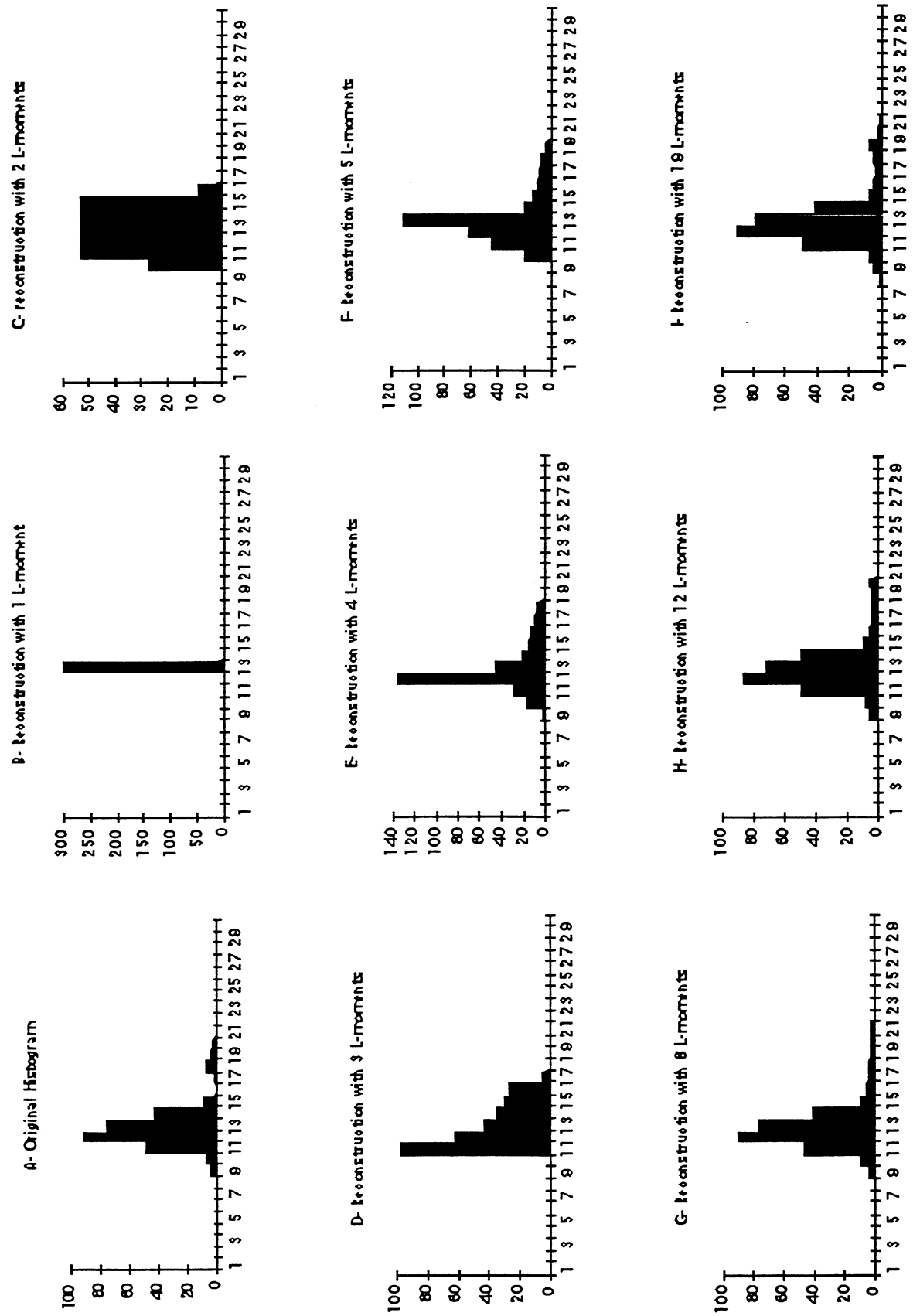


Figure 2 : Reconstruction of an original histogram (A) of 30 classes and 300 objects (Normal distribution 1, mean1 = 2000, CV1 = 10%, P1 = 0,90; Normal distribution 2, mean2 = 4000, CV2 = 10%, P2 = 0,10) with respectively 1, 2, 3, 4, 5, 8, 12 and 19 L-moments. The best reconstruction is obtained by using 19 L-moments (I).

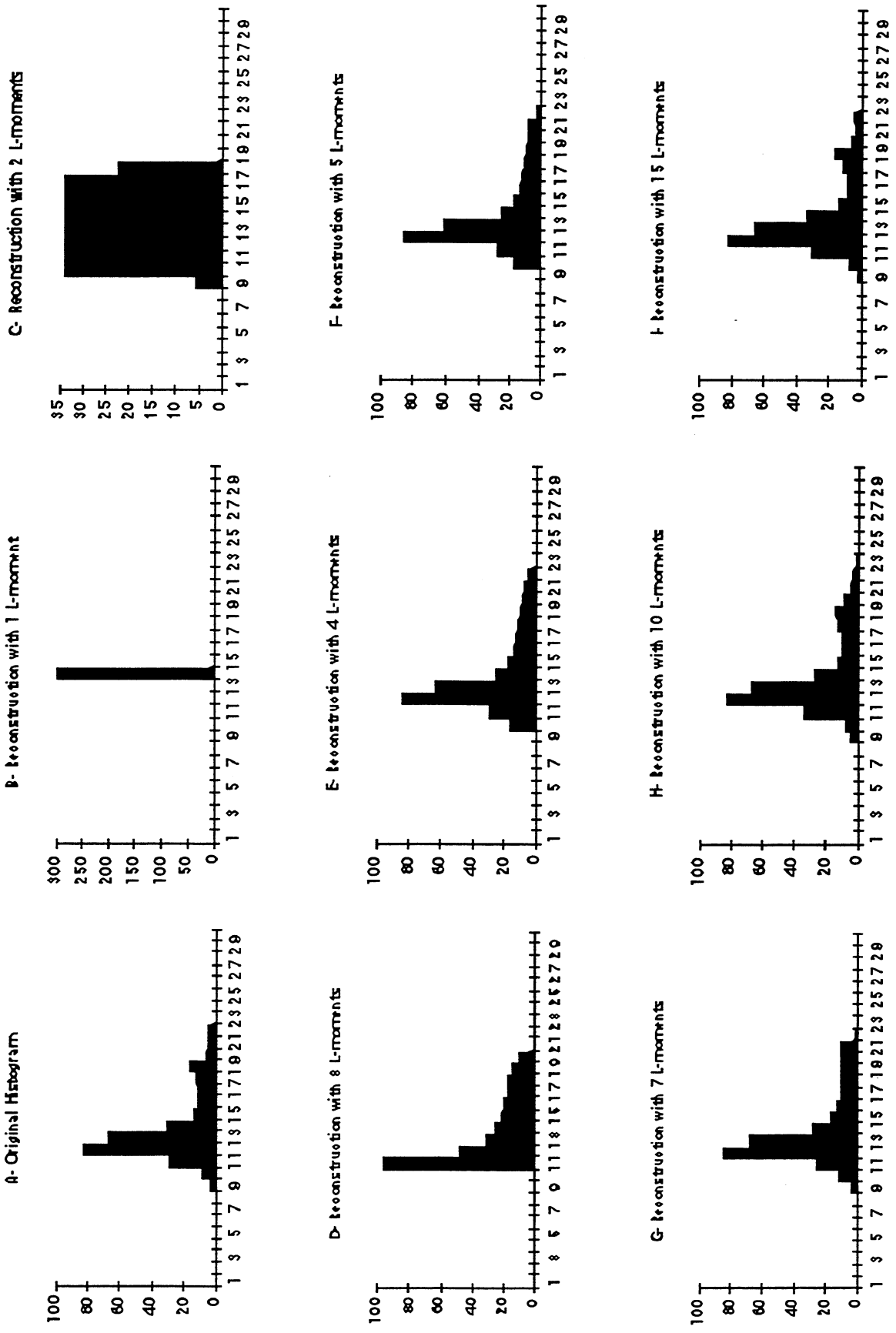


Figure 3 : Reconstruction of an original histogram (A) of 30 classes and 300 objects (Normal distribution 1, mean1 = 2000, CV1 = 10%, P1 = 0,75; Normal distribution 2, mean2 = 4000, CV2 = 10%, P2 = 0,25) with respectively 1, 2, 3, 4, 5, 7, 10 and 15 L-moments. The best reconstruction is obtained by using 15 L-moments (I).

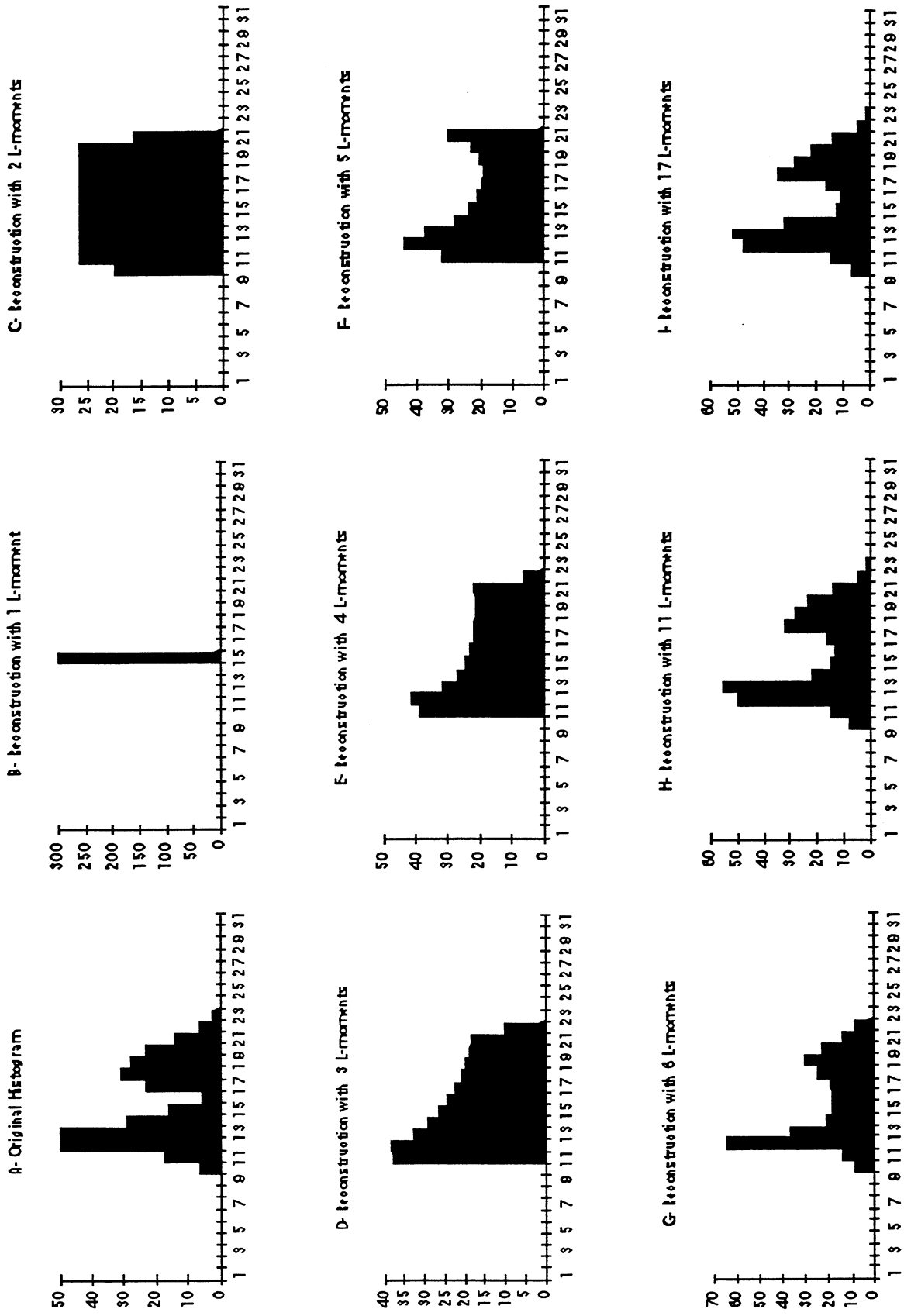


Figure 4 : Reconstruction of an original histogram (A) of 30 classes and 300 objects (Normal distribution 1, mean1 = 2000, CV1 = 10%, P1 = 0,50; Normal distribution 2, mean2 = 4000, CV2 = 10%, P2 = 0,50) with respectively 1, 2, 3, 4, 5, 6, 11 and 17 L-moments. The best reconstruction is obtained by using 17 L-moments (I).

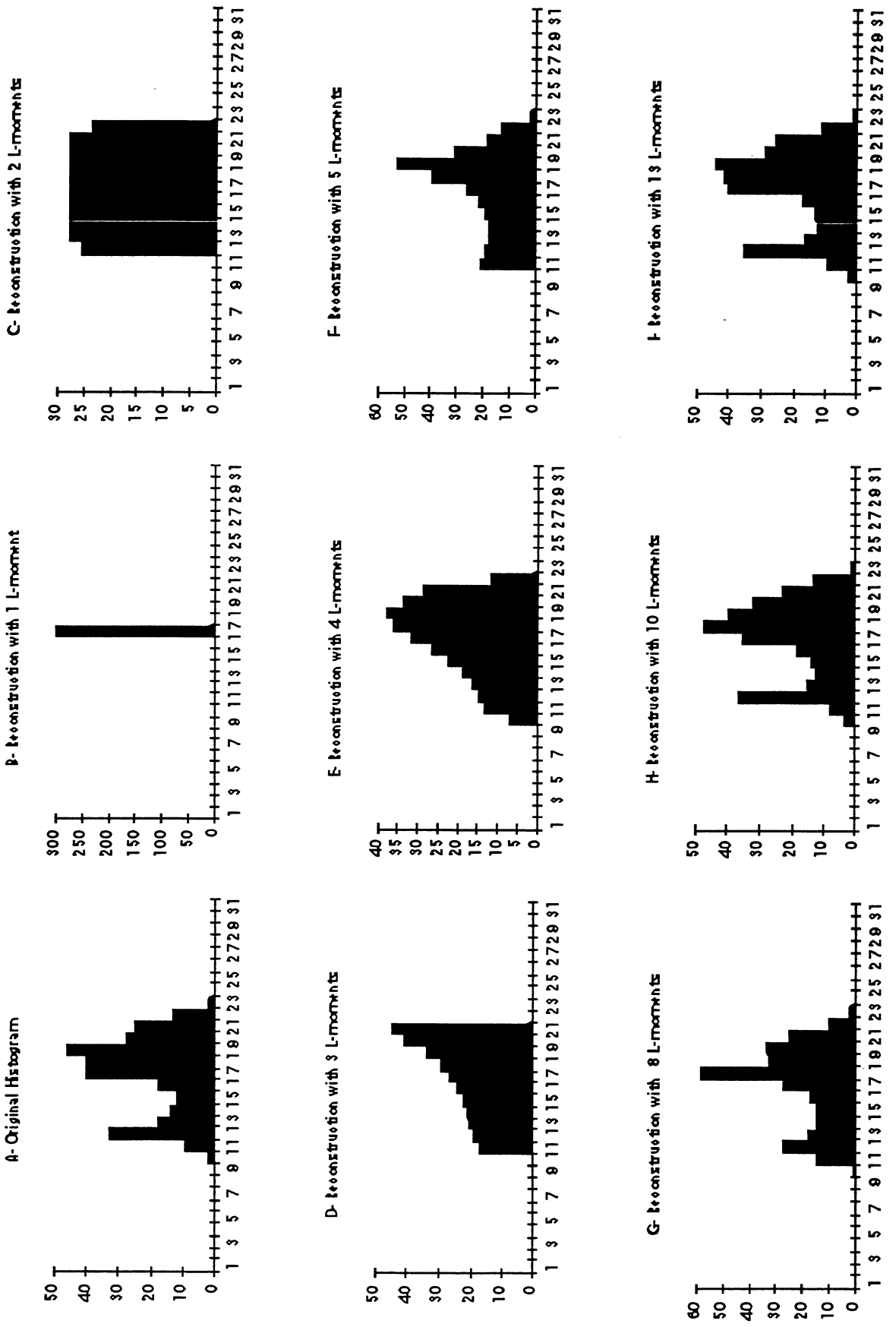


Figure 5 : Reconstruction of an original histogram (A) of 30 classes and 300 objects (Normal distribution 1, mean1 = 2000, CV1 = 10%, P1 = 0,25; Normal distribution 2, mean2 = 4000, CV2 = 10%, P2 = 0,75) with respectively 1, 2, 3, 4, 5, 8, 10 and 13 L-moments. The best reconstruction is obtained by using 13 L-moments (I).

III.3. Problem of outliers in real histogram

In a mixture of two normal distributions, the first detected distribution is the one which is more weighted, of course. The speed of reconstruction convergence depends on the size of the minor distribution. But in real distributions as DNA histograms, histograms are not so simple.

In Figure 6, a real DNA histogram built on 30 classes is shown which has been generated from 30 classes (with 346 objects). A major sub population can be easily detected. Moreover this histogram shows some successive irregularities, due to very low weighted sub-populations, and especially an isolated class (one object) at the tail of the distribution. The best reconstruction is obtained with 16 L-moments. The approximation is more or less accurate but the reconstruction "forgets" the smallest classes at the end of the distribution. The presence of the outlier in class 29 is responsible for the smooth tail that can be seen in Figure 6I at the right of the distribution. Indeed, even if the outlier is not detected, the polynomial approximation is sensitive to this value and tends to show it. Thus the tail of the distribution is not well represented.

We are faced with a fundamental problem. In order to solve it, we propose to modify this histogram. Indeed, the existence of empty classes between the majority of the distribution and the outlier is an element of the unsatisfactory representation of the extremely low size classes. What is proposed here is to transform the initial histogram of Figure 6 by adding a same number of objects in each of the original histogram classes. This new histogram is represented by Figure 7A. It has no more empty classes. In comparison with the six first reconstructed histograms of Figure 6, the six first reconstructed histograms (Figure 7B-7F) evolve in the same way. The difference concerns a stronger skewness of the reconstructed histograms of Figure 7, in response to the increase of objects number in the distribution tail. 21 L-moments are needed to obtain the best reconstruction (Figure 7I) and even if a class has been created at the end of the distribution, the rise in number of L-moments needed to improve the quality of reconstruction seems to be relatively high.

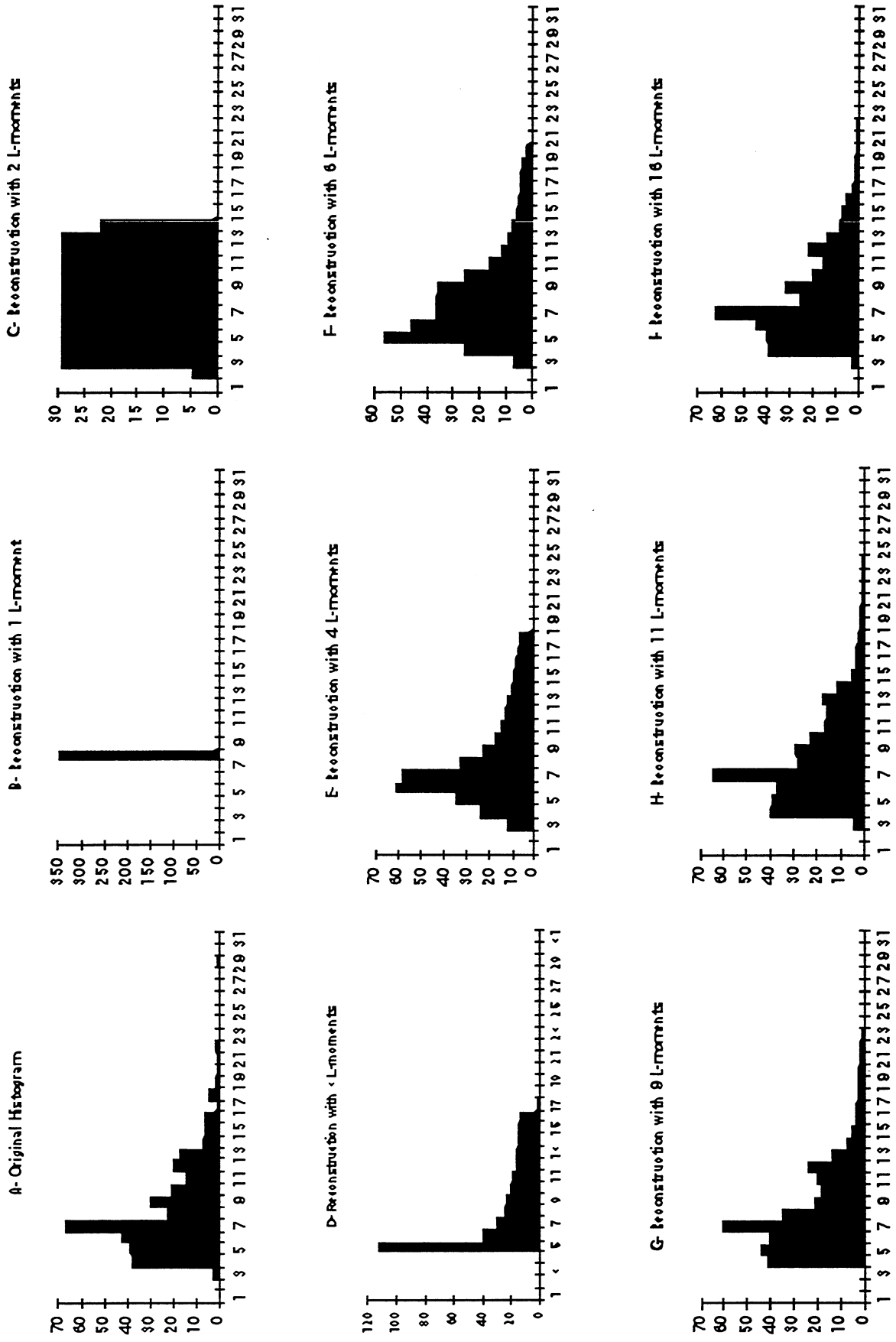


Figure 6 : Reconstruction of an original real ploidy histogram (A) of 30 classes and 300 objects with respectively 1, 2, 3, 4, 6, 9, 11 and 16 L-moments. The best reconstruction is obtained by using 16 L-moments (I).

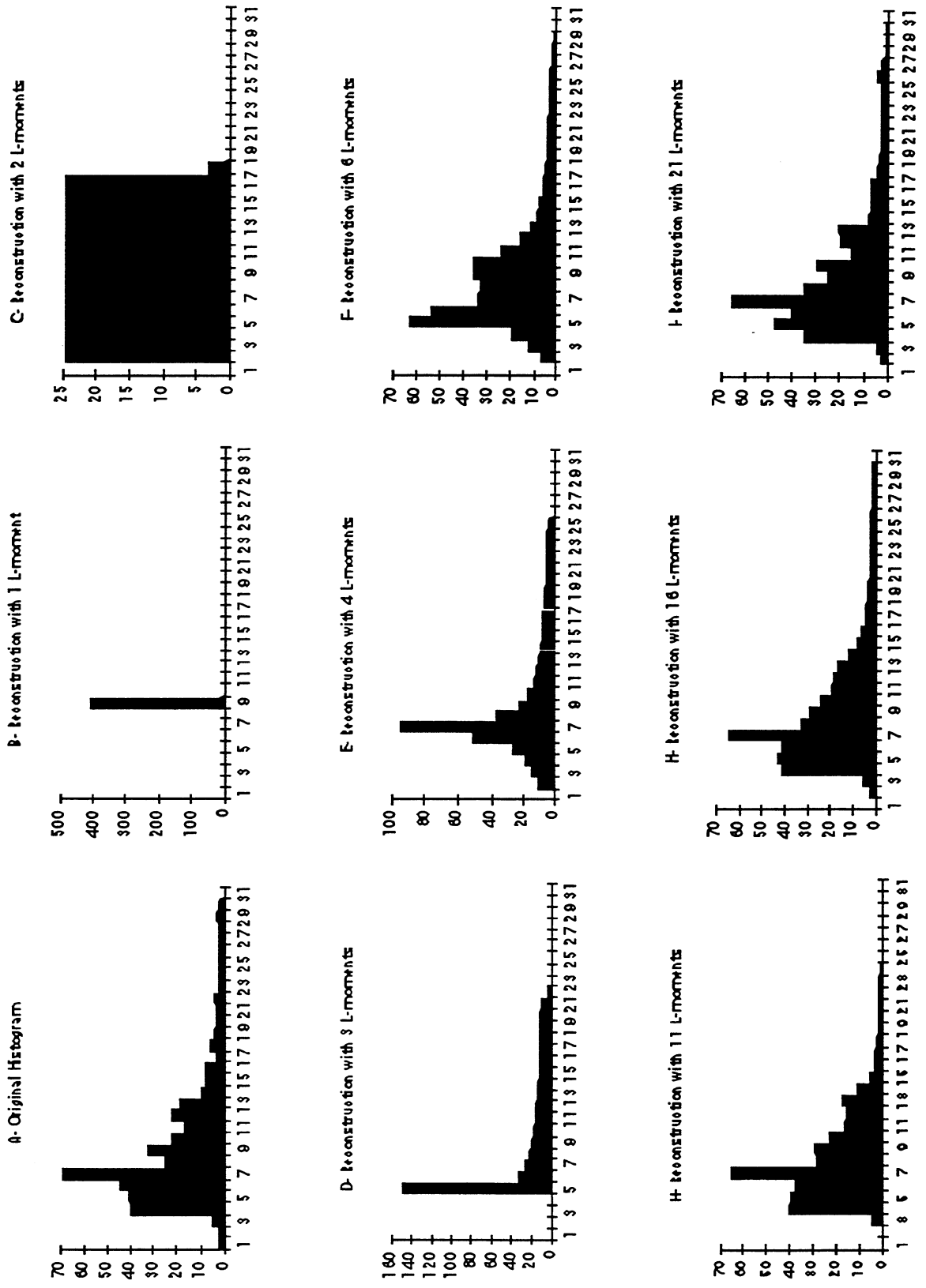


Figure 7 : Reconstruction of an original modified ploidy histogram (A) of 30 classes and 300 objects with respectively 1, 2, 3, 4, 6, 11, 16 and 21 L-moments.The best reconstruction is obtained by using 21 L-moments (I).

III.4.Problem of degradation

When the number of Legendre polynomials is too high, a degradation of the reconstruction appears due to the non-monotonic increment in the approximated quantile function. In fact, the problem of degeneration is common with all polynomial approximations encountered in many inverse problems such as image restoration [AND77]. In Figure 8, the evolution of criterion Err in terms of number of L-moments used for the reconstruction is shown. This degradation does not appear when the distribution has a regular and smooth tail, as in the case of a normal distribution (Figure 8A). But when the distribution is irregular (heterogeneous), the phenomenon appears (Figure 8B).

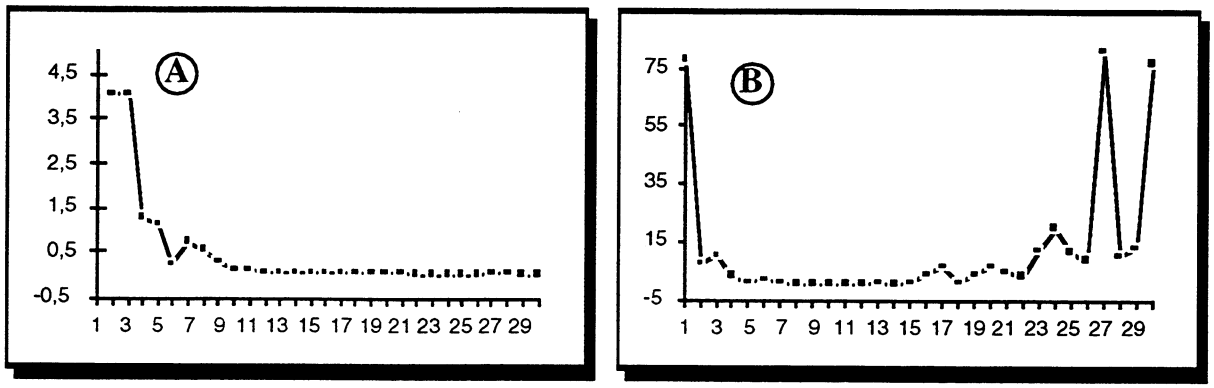


Figure 8. Evolution of stop criterion Err (ordinate) as a function of the number of L-moments (abscissa) used to approximate the original histogram. A : simulated normal distribution, B: real heterogeneous distribution.

Complete reconstruction remains an open problem and we only propose to detect the appearance of degradation in order to stop the reconstruction.

III.5. Dependency on class number

The dependency on number of classes associated with the best representation has also to be considered. Depending on the application, this choice is often arbitrary. In our problem we have chosen to keep the same number of classes regardless of the range of the sample. Figure 9 shows three histograms A, B and C of the same distribution but generated from respectively 10 classes (Figure 9-A1), 20 classes (Figure 9-B1) and 30 classes (Figure 9-C1). The best representation is obtained with 10 L-moments. It is satisfactory for the 10 classes histogram (Figure 9-A3) but not for the 30 classes histogram (Figure 9-C3). Here it is clear that the lower the number of classes, the better the reconstruction. But, the ratio of the number of L-moments associated with the best reconstruction to the number of classes is better for histogram C (1:3) as compared with histogram A (1:1). Thus a compromise has to be found between the quantity of information present in a histogram and the quality of the reconstruction.

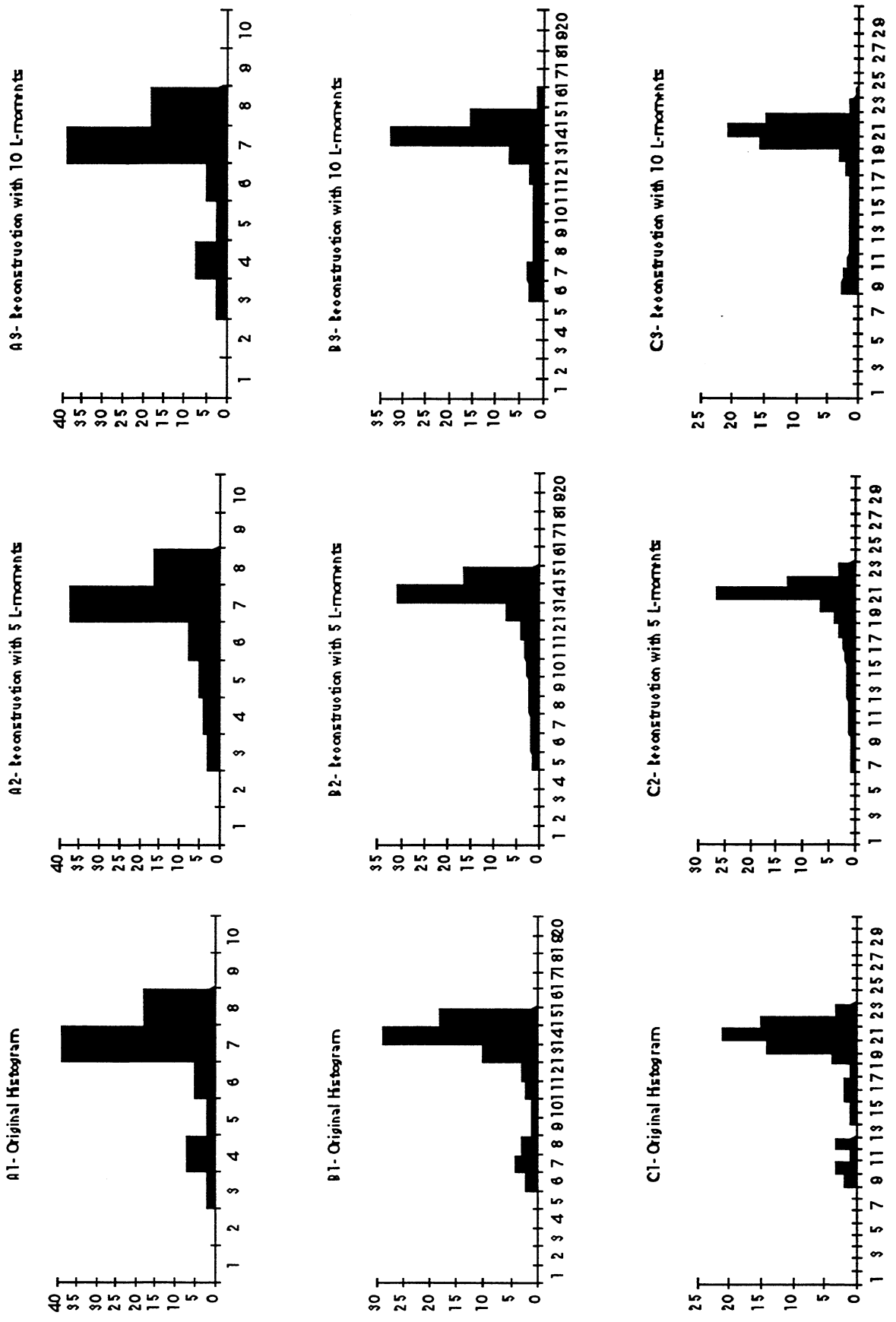


Figure 9 : Reconstruction of the same real ploidy histogram defined on 30 classes (A1), 20 classes (B1), 10 classes (C1) and 300 objects using respectively 5 (A2, B2, C2) and 10 L-moments (A3, B3, C3). The best reconstructions are obtained by using 10 L-moments.

IV.DISCUSSION

In each pattern recognition problem , feature extraction is a crucial step. In this paper, we have shown that L-moments are the principal features of a histogram. Two important points can be deduced from these results.

Firstly, a histogram initially represented by K classes has been transformed into a Q point (Q L-moments) representation, with Q inferior to K. Thus the dimensions of the representation space has been reduced with little or no loss of information. Furthermore, this new features are independent contrary to histogram classes.

Secondly, it follows from the inversion theorem that the quality of the reconstruction can be is consistent.

These two main points show that the method of analysis of any distributional pattern can be demonstrated without the need to have defined a probability law. These results demonstrate therefore the limitations to the inverse quantile theorem in pattern reconstruction. The lack of quality in pattern reconstruction when too many L-moments are used is not a problem encountered only when Legendre polynomials are used. Many studies have demonstrated that this problem is still unresolved [AND77]. The detection of rare objects, in relation to empty classes, is limited by the large number of polynomial approximations needed. Nevertheless, these problems can be handled as long as there are few histogram classes.

The number of L-moments satisfying the stop criterion is sometimes high. In our opinion, this criterion can be more efficiently replaced by a different test. Indeed, it would be sufficient to retain the number of L-moments which provide a reconstructed histogram "biologically equivalent" to the original histogram. In this respect it is possible to compare the distributions using, for instance, the Kolmogorov test or the Khi2 test. The first kind error can thus be fixed at a value depending on the appropriate quality of reconstruction.

It is usually sufficient to determine 20 to 30 classes in order to correctly reproduce a set of data in the form of a DNA histogram. Moreover, it is unlikely that the presence of rare objects at the extremities of the distribution will alter the resulting diagnosis. DNA histograms can be grouped into a limited number of diagnostic groups [AUE80]. As far as this application is concerned, it is apparent that the quality of pattern reconstruction is not hindered by a low number of L-moments. These results demonstrate a novel application of L-moments in DNA histogram analysis. Future work should lead to the classification of histograms from biological samples at different pathological stages as a function of L-moment values.

Acknowlegments

The authors thank Victoria von Hagen for the critical revision of the manuscript.

References

[AND77] Andrews and Hunt
Digital Image restoration
Prentice-Hall, New York, 1977.

[AUE80] Auer, G.U. et AL.
DNA Content and Survival in Mammary Carcinoma.
AQCH 2 : 161- 165 (1980)

[GRE79] Grenwood, J.A, Landwehr, J.M., Matalas, N.C. and Wallis, J.R.
Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form.
Water Resour. Res., 15, 1049-1054, 1979.

[HOS86] Hosking J.R.M.
The theory of probability weighted moments.
Research report RC12210.
IBM Research, Yorktown Heights.

[HOS89] Hosking J.R.M.
Some theoretical results concerning L-moments
Research report RC14492.
IBM Research, Yorktown Heights.

[HOS90] Hosking J.R.M.
L-moments: Analysis and estimation of distributions using Linear combinations of order statistics.
J.R. Statist. Soc. B (1990), 52, N° 1, pp. 105-124.

[OJA81] Oja H.
On location, scale, skewness and kurtosis of univariate distributions
Scand.J. Statis, 8, 154-168, 1981.

[OPF87] OPFERMAN M., BRUGAL G. and VASSILAKOS P.
Cytometry of breast carcinoma: Significance of ploidy balance and proliferation index.
Cytometry 8:217-224 (1987)

[SIL69] Sillitto G.P.
Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample.
Biometrika (1969), 56, 3, p.641.

[YOU74] Young., T.Y. and Calvert T.W.
Classification, estimation and pattern recognition
pp 1-11.
American Elsevier Publishing Company, INC. 1974

II.2 Acquisition d' histogrammes d'ADN sous contrôle statistique

Les L-moments s'avèrent de très bons descripteurs des histogrammes d'ADN. Notre problème de reconnaissance de formes, évoluants en cours d'analyse, est abordé en considérant non pas les caractéristiques propres de l'histogramme (fréquences des classes) mais les L-moments. La variabilité de ces L-moments, en fonction de la taille des échantillons, est utilisée comme critère de stabilité des histogrammes. Ces travaux sont exposés sous forme de publication :

Cytometric Data acquisition under statistical control
Soumis à Analytical Cellular Pathology en Avril 1993

CYTOMETRIC DATA ACQUISITION UNDER STATISTICAL CONTROL

Martial GUILLAUD and Jean-Marc CHASSERY
Laboratoire TIMC - IMAG - URA D 1618
Université Joseph Fourier, CERMO, BP 53X, 38041 Grenoble Cedex - France
Phone: (33) 76514813 / Fax : (33) 76514948

Abstract

Numerous problems occur in DNA analysis and mainly in DNA histograms interpretation We propose that each sample be considered independently, as a function of the tumour type, the number of cells and DNA histogram. This stratagem is to find an arrest test for data acquisition. This approach, which turns out to be a pattern recognition problem, can be subdivided into three steps. Firstly, the histograms can be replaced by the corresponding set of L-moments, combinations of order statistics. Secondly, the stability of an histogram is investigated by measuring the variances of its features, i.e. L-moments, by the Bootstrap method. Finally, both the previous steps are repeated on samples of increasing size. One hundred real DNA histograms, and synthetic data have been analysed by simulating real time data acquisition. Despite the variability of the values of a rare event, our model is able to detect the advent of a cell with a DNA content which was not still represented. An histogram is said to be stabilised if firstly, the higher order L-moment variances are stabilised, and secondly, at the same time, the lower order L-moments variances remain stable or slightly decrease.Our strategy, based on stability of the L-moment variances, gave uniform and homogeneous results. We propose then to use it as *stop analysis criterion* .

Key words:
DNA image cytometry. Bootstrap. L-moments. Statistics. Stop criterion. Statistical control.

Corresponding author:
Martial Guillaud.
Laboratoire TIMC - IMAG - URA D 1618. Université Joseph Fourier, CERMO, BP 53X, 38041 Grenoble Cedex - France. Phone: (33) 76514813 / Fax : (33) 76514948

I. INTRODUCTION

DNA analysis is the most used cytometric application in clinical laboratories [1]. Nethertheless, numerous problems occur in DNA analysis and mainly in DNA histograms interpretation [4, 2]. One of these problems is well known as sample representativity, i.e. the sample size required and quality of the analysis [3,4]. Several statistical studies have proposed that an analysis of 300 cells is sufficient for DNA histogram interpretation [5]. Other authors suggest that not less that 2 000 to 4 000 cells have to be analysed [24].

In clinical routine analysis, when the quality of the preparation is satisfactory, 250 cells, on the average, are analysed. But, frequently, the quality of biological preparations only permit the analysis of 100 cells. The graphic representation of the DNA histogram distribution can be made regardless of the size of the sample. The diagnostic established by the cytopathologist based on this histogram does not take into account the sample size, and uses only the DNA histogram pattern.

The biological and topographical heterogeneity of some solid tumours [19, 22] and the intrinsic limits of sampling methods (Imprints, FNAB, etc..) [20,21] do not allow to define some general and theoretical rules about the sample size required in DNA analysis.

Based on this consideration, we propose that each sample be considered independently, as a function of the tumour type, the number of cells and DNA histogram. Even if question of the representativity of a sample with respect to the whole tumour has not yet been solved, the representativity of a cells sample comparing to the cell population of the slide, can be investigated.

We propose here a new tool to examine in real time the DNA histogram evolution as the number of analysed cells increases.

This stratagem is to find an arrest test for data acquisition. The data in this case is DNA histograms rather than raw data. This approach, which turns out to be a pattern recognition problem, can be subdivided into three steps.

Firstly, L-moments, combinations of order statistic, introduced and studied by Hosking [14,15,16] are used to characterise DNA histograms. In a recent publication [12], we demonstrated the usefulness interest of the first L-moments as features for histograms. When considered as univariate distribution, the histogram can be replaced by the corresponding set of L-moments. These L-moments will be used to follow the DNA histogram evolution as sample size increases.

Secondly, the stability of an histogram is investigated by measuring the variances of its features, i.e. L-moments. To this end, we used the Bootstrap method [10]. This method is based on the following observations : if a sample of size n contains all the information about the population , then proceed **as if** the sample was the population for purposes of estimating confidence intervals (and variances) of the statistics used (L-moments). The term "information", in our context, concerns the presence of cells with different DNA values. In this paper, we use the Bootstrap method to determine the stability of the L-moments [13].

Finally, both the previous steps are repeated on samples of increasing size, the purpose being to study the relationship between DNA histogram evolution and the stability of the corresponding ands relevant L-moments.

II. MATERIALS AND METHODOLOGY

Because the Feulgen reaction is stochiometric, our ploidy analysis software makes it possible to measure the DNA content of cell population. This DNA content of a sample is represented by an histogram, coded on K classes (where K varies from 10 to 100) and from which the diagnosis is made.

II.1 L-moments : Descriptors for histograms

The L-moments are analogous to the conventional moments but can be estimated by linear combinations of order statistics. L-moments have theoretical advantages over conventional moments :

(i) they characterise a wider range of distributions

(ii) when estimated from a sample, they are more robust to the presence of the outliers in the data,

(iii) they are less subject to estimation bias and they approximate their asymptotic normal distribution more closely in finite samples.

Hosking has shown [16] that L-moments form the basis for the statistical analysis of univariate probability distributions, since they cover the characterisation of probability distributions, the summarisation of observed data samples and the fitting of probability distributions to the data.

II.1.1 Preliminary Notations

Let X be a real-valued random variable with a cumulative distribution function F(x).

Let x_i be the value assumed by the variable X for different elements of a random sample of size N, indexed by i, for i = 1, ..., N.

Let x(F) be the quantile function or inverse distribution function of X. x(F) is defined by N points.

Let f be the notation of the initial histogram. The component f(j) is the relative frequency of objects belonging to class j, for j = 1, ..., K, where K is the number of classes referring the histogram.

II.1.2 L-moments : definitions and properties

This method (illustrated in Figure 1) codes an initial distribution function using the L-moments. We generate a reconstructed approximation of the initial distribution defined as projections on the basis of Legendre polynomials. By comparison between the initial and the reconstructed distribution, we propose to determine the number of L-moments that need to be preserved.

Definitions :

Let $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ be the order statistics of a random sample of size n drawn from the distribution of X. Define the L-moments of X to be the quantities

$$\lambda_r \equiv r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r-k:r}, \qquad r = 1, 2, \ldots \qquad (1)$$

The L in 'L-moments' emphasises that λ_r is a *linear* function of the expected order statistics. Furthermore, the natural estimator of λ_r based on an observed data sample is a linear combination of the ordered data values, i.e. an L-statistic. The expectation of an order statistic may be written as [7] :

$$EX_{j-r} = \frac{r!}{(j-1)!\,(r-j)!} \int x \{F(x)\}^{j-1} \{1 - F(x)\}^{r-j} \, dF(x)$$

Substituting this expression in definition 1, expanding the binomials in F(x) and summing the coefficients of each power of F(x) gives the following definition :
The L-moments of a quantile function x(F) are [15] :

$$\lambda_r = \int_0^1 x(F) \, P^*_{r-1}(F) dF, \text{ with } r = 1, 2, \ldots . \quad (2)$$

where $P^*_r(x)$ is the **r***th* shifted Legendre polynomial, derived from the Legendre polynomials $P_r(x)$, and given by the relation : $P^*_r(x) = P_r(\frac{x+1}{2})$.

The Legendre polynomials generate an orthogonal base for polynomials with unit weight function [18]. These polynomials can be used to define vectorial data by a polynomial transform.

The first 4 L-moments are :

$$\lambda_1 = EX = \int_0^1 x(F) dF \, ,$$

$$\lambda_2 = \frac{1}{2} E(X_{2:2} - X_{1:2}) = \int_0^1 x(F)(2F-1) \, dF \, ,$$

$$\lambda_3 = \frac{1}{3} E(X_{3:3} - 2X_{2:3} + X_{1:3}) = \int_0^1 x(F)(6F^2 - 6F + 1) \, dF \, ,$$

$$\lambda_4 = \frac{1}{4} E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}) = \int_0^1 x(F)(20F^3 - 30F^2 + 12F - 1) \, dF$$

The use of L-moments for probability distributions has been widely justified [7,17].

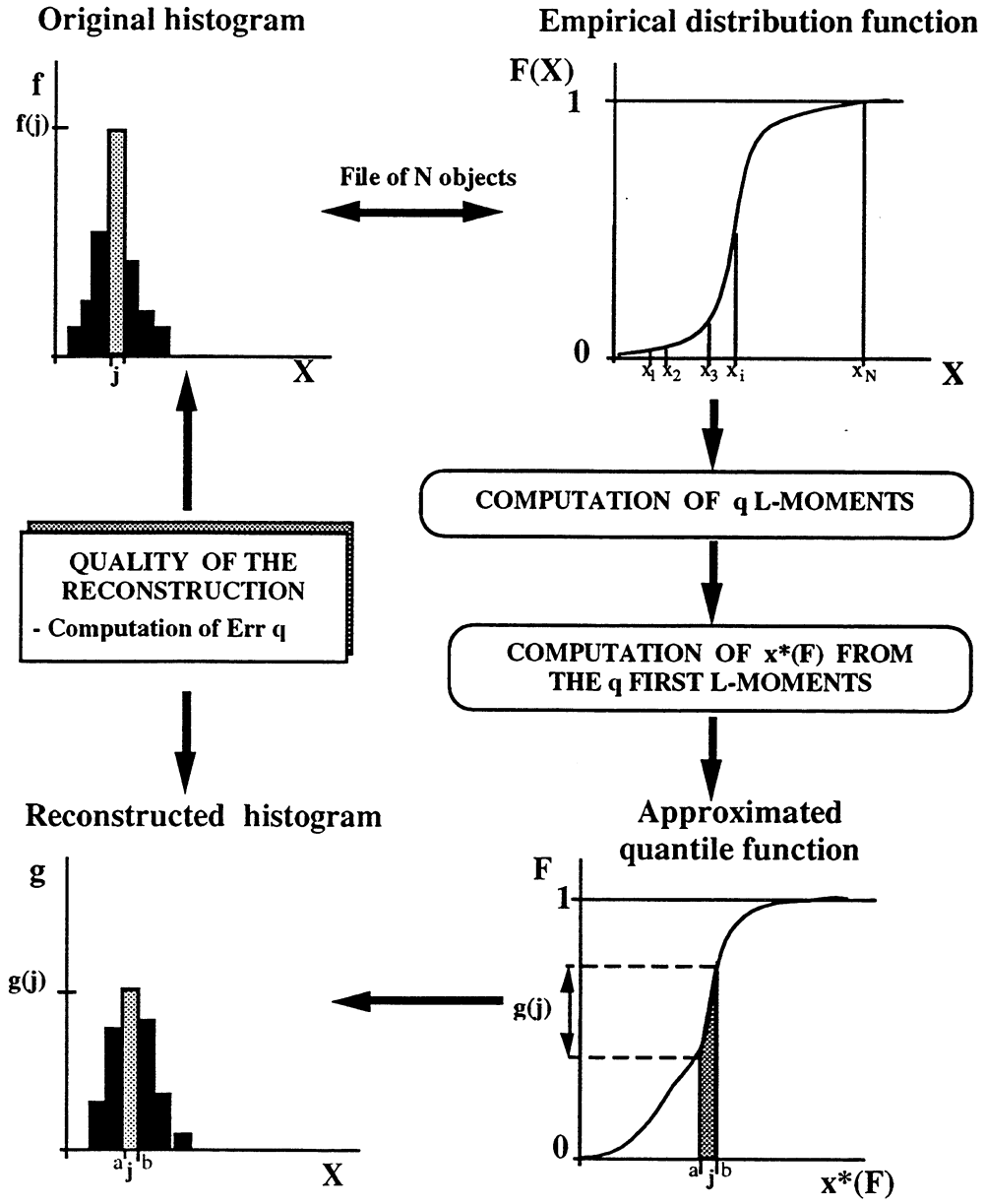


Figure 1 : Illustration of the algorithm of histogram coding. The original histogram is coded on the basis of L-moments representation. Only the q first l-moments are retained in order to generate the "best" approximated histogram.

A distribution may be specified by its L-moments even if some of its conventional moments do not exist.

The Probability Weighted Moments, defined by Grenwood [11], can be expressed as linear combinations of L-moments. These L-moments are more convenient, however, because they are more directly interpretable as measures of the scale and shape of probability distributions.

The selection of the number of L-moments will be developed below.

Reversibility properties

Using the inversion theorem, proposed by Sillitto [23], x(F) can be approximated by a quantile function x*(F) given in terms of L-moments by equation (3) :

$$x^*(F) = \sum_{r=1}^{s} (2r-1)\lambda_r P^*_{r-1}, \text{ with } 0 < F < 1 \quad (3)$$

The convergence of such a sequence toward x(F) is satisfied in a mean square sense i.e. :

$$\lim_{s\to\infty} \int_0^1 (x(F) - x^*(F))^2 \, dF \quad \text{tends to zero.} \quad (4)$$

Approximative coding and reconstructed histogram.

A new distribution function F* can be generated from the approximated quantile function x*(F). The relative frequency of each class j of the corresponding reconstructed histogram, noted by g, can be estimated from this distribution by integration between two limit values a and b of the class j. Therefore

$$g(j) = F^*(b) - F^*(a) \text{ for } j = 1 \text{ to } K.$$

II.1.3 Algorithm

Since the analytical formulation of the quantile functions is unknown, we use the numerical vector representation and propose the following algorithm :

- Let X be a continuous random variable given by N observations $x_1, x_2, \ldots, x_N$.
- Let F be the distribution function of the variable X defined on N points .
- Let f be the histogram, representation of the variable X on K classes by its components f(j) for j = 1 to K.
- The empirical distribution function F can be plotted as

$$F(x) = Pr(X \le x) \text{ and } F(xi) = \sum_{z=1}^{i} Pr(X \le x_z) \text{ for } j = 1 \text{ to } N.$$

- The L-moments can be computed as:

$$\lambda_r = \int_0^1 x(F) \, P^*_{r-1}(F) dF, \text{ with } r = 1 \text{ to } q.$$

From the N values of the empirical distribution function F, this integral is approximated by means of Simpson's algorithm.

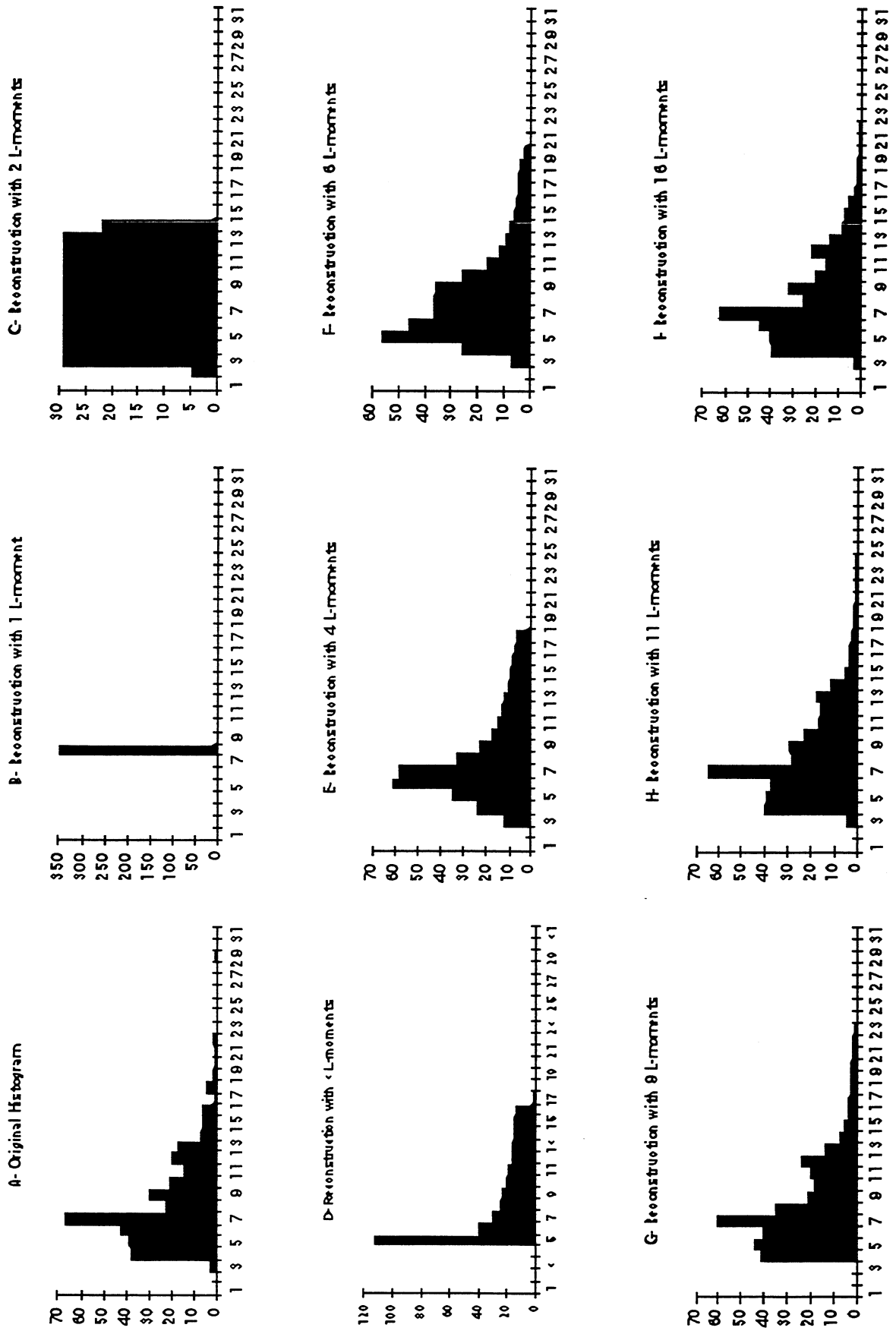


Figure 2 : Reconstruction of an original DNA ploidy histogram (A) of 30 classes and 300 objects with respectively 1, 2, 3, 4, 6, 9, 11 and 16 L-moments. The best reconstruction is obtained by using 16 L-moments (I).

II.1.4 Determination of the best reconstruction

The choice of the number of L-moments needed to describe the distribution F depends on the quality of the inversion. The process is stopped when a criterion error, called Err, reaches a minimum. The most obvious criterion is given by relation (4). However, Hosking [14] leaves a number of unanswered questions about this convergence. Thus we prefer to compute a convergence criterion based directly on the reconstructed histogram. We note Err_q the quadratic error between the initial histogram f, the reconstructed one noted by g_q using the first q L-moments. If K is the number of classes of the histogram, Err_q is defined by the relation :

$$Err_q = \sum_{j=0}^{K} \frac{(f(j) - g_j(j))^2}{f(j)^2} \qquad (5)$$

The best reconstruction is obtained with q L-moments when Err_q reaches a minimum.
An example of real histogram reconstruction is given in Figure 2.

II.2 Bootstrap method

The bootstrap method was originally developed to compute confidence intervals in statistics, such correlation coefficients [9], for which no theoretical analysis is available. The variance of the L-moments can be computed by this method.

The idea of the Bootstrap (illustrated in Figure 3) can be presented as following. Suppose we have a random sample of data (x_1, x_2, x_i, ..., x_n), which is used to estimate a parameter θ , resulting in the estimate $\hat{\theta}$. The indication of the precision of this estimation is a property of the sampling distribution of $\hat{\theta}$. The idea of the bootstrap method is to sample, repeatedly, T times, uniformly at random, and with replacement, from the one sample of real data (x_1, x_2, x_i, ..., x_n).
Successive T bootstrap samples will therefore contain the elements x_1,..., x_n , but some may be repeated several times for any sample, and then others will be missing. For each bootstrap sample, the estimate $\hat{\theta}$ of θ may be calculated, allowing a form of sampling distribution to be estimated, which may then be used to gage the precision of $\hat{\theta}$.

Preliminary results show that the sampling distribution of L-moments estimates noted by $\hat{\lambda}_r$ are a good approximation of a Gaussian distribution. Using such results, the variance of these q L-moment estimates will be used as estimations of the variability of the q L-moments.

Let us extract a sample of n cells drawn from a population of N cells. From the T "Bootstrap samples", referring to the q L-moments, we evaluate the Bootstrap mean and the Bootstrap variance as :

$$\text{Mean } \hat{\lambda}_i = \frac{\sum_{j=0}^{T} \hat{\lambda}_i^j}{T}, \qquad \text{Var } \hat{\lambda}_i = \frac{\sum_{j=0}^{T} (\hat{\lambda}_i^j - \text{Mean } \hat{\lambda}_i)^2}{T} \text{ for i = 0 to q.}$$

In practice, the Bootstrap samples are obtained by random drawings using computer intensive simulations.

II.3 Real time DNA data acquisition under statistical control

In the two previous sections, we proposed computational evaluation of the variability of the L-moments, features of the DNA histograms. This operation is repeated on samples of increasing size. The Figure 4 illustrates the principle of the real time data acquisition under statistical control. We will examine the evolution of the Bootstrap variances for all L-moments, as sample size increases and we are waiting for the detection of new classes.

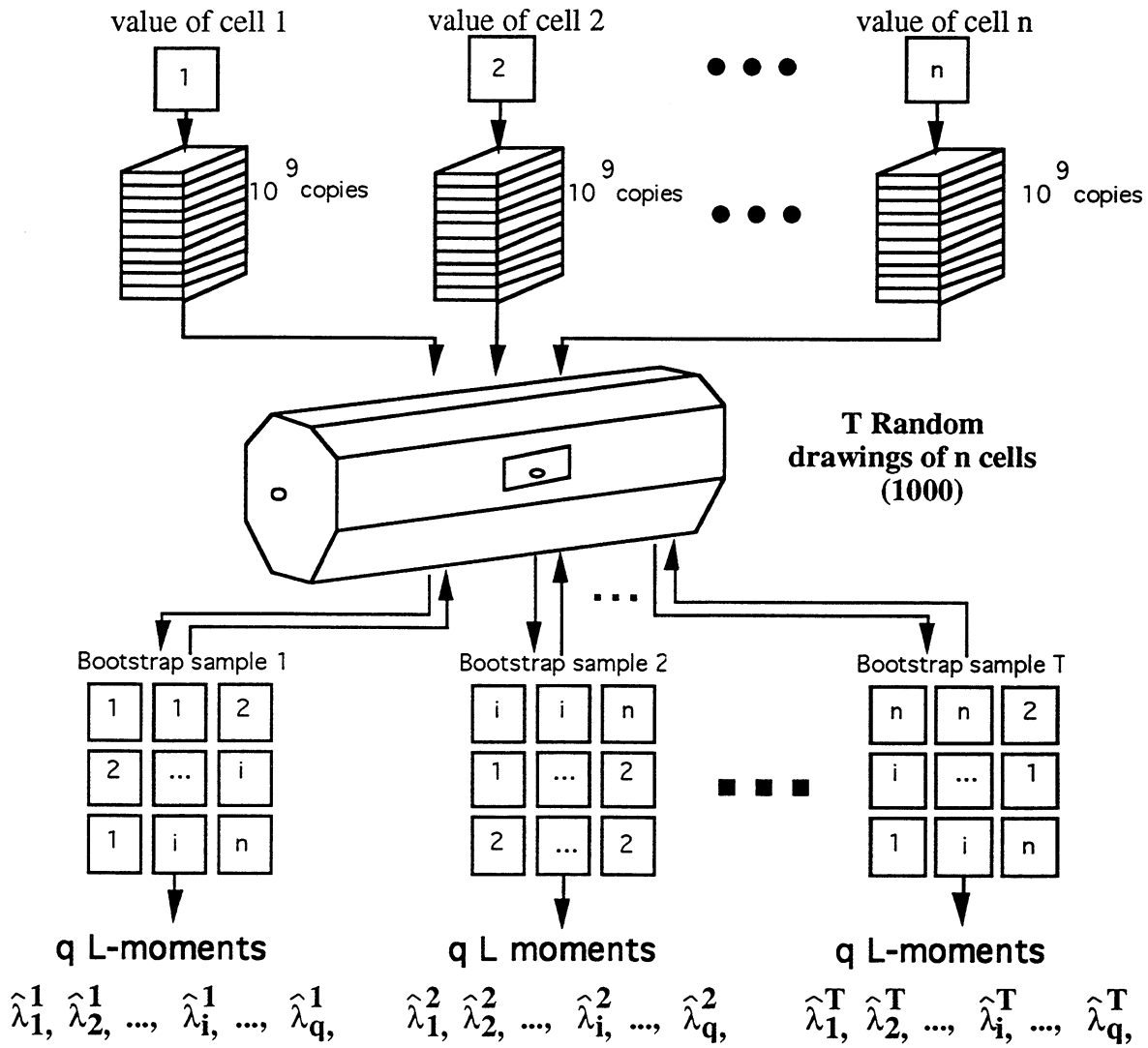


Figure 3 (from Efron (8)) : Bootstrap method to compute the variances of q L-moments from an initial sample of n cells drawn from a cell population of N cells.

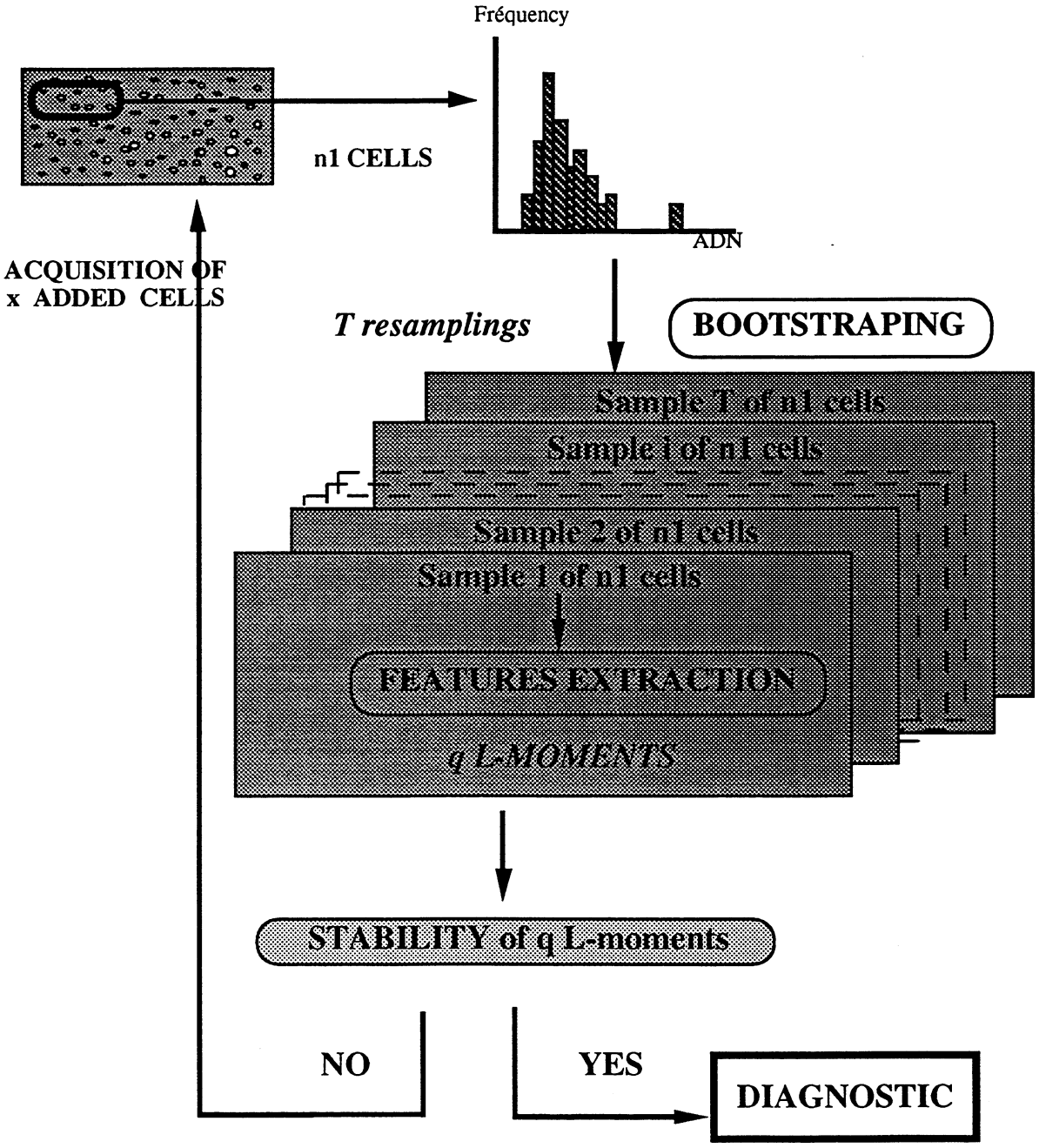


Figure 4 : Principle of the real time DNA acquisition under statistical control.

III. RESULTS

The proposed method has been tested on two different sets of data ;

- Synthetic data issued from a random sample generated by a probabilistic law,
- Real data from the DNA image analysis of breast cancer samples.

Synthetic data

Figure 5 shows the evolution of the Bootstrap variances of the first 10 L-moments computed from a histogram obtained from a Gaussian distribution (mean 2000, standard deviation 200). This histogram corresponds to an idealised diploid histogram. All the information contained in the whole population of 300 cells is already present in the first sample of 50 cells. The curves show that all bootstrap variances of the first 10 L-moments follow an exponential decrease. Stability is reached from a sample size of 200 cells.

Figure 6 shows the evolution of Bootstrap variances of the first 10 L-moments when a new datum appears during data acquisition. This new information will appear at order 110. A cell is analysed with a DNA content different from the other value of the homogeneous diploid population. Different simulations have been made corresponding to different values of this new information comparing with the already established population. Figures 6a, 6b, 6c, 6d, 6e, and 6f corresponds respectively to the advent of the cell in 2.5c, 3c, 3.5c, 4c, 5c and 8c. A few observations can be made. First, the greater the difference between the new information and the established one, the greater the increase of the bootstrap variances of each of the L-moments. Nethertheless, the advent cell in 2.5c does not induce any increase in the variances. An advent of a 3c cell is detected by a slight increase of the variances of all L-moments, excepting variances of L-moments number 1, 2, 9 and 10. The variance Bootstrap of the first L-moments is only sensitive to the advent of a 4c cell or more. The low increase of some variances around the sample of 200 cells come from the advent of cells with very low DNA content (near 1.6c) just at the limit of the extremity of a diploid Gaussian distribution.

Figure 7 shows the evolution of the Bootstrap variances of the first 10 L-moments when the representativity of a sub-population changes during acquisition. In a sample size of 50 cells, a 4c cell is already present. When a new 4c cell appears at order 110, only the first 6 L-moments increase. The 300^{th} cell, which is also a 4c cell, involves a real increases in the first 6 L-moment variances. This phenomenon is characteristic ; only the first L-moment variances are sensitive to the advent of some new information, already present but in low proportions, as compared to the principal one.

Real data

The following results have been obtained from an image workstation SAMBA 200. The number of cells in each file is determined either by the quality of the preparation (in this case the cell number is generally very low, for example 90) or by an arbitrary decision by the cytotechnicien.

Figure 8 shows the evolution of the first L-moments variances on a real file of 278 cells. This histogram shows a pronounced aneuploidy corresponding to a bad prognosis. We observe that all L-moments variances follow a regular exponential decrease. A stabilisation is observed at 200 cells. The new cells which appear between 2c and 4c cannot be considered as representative of new data (proliferating or aneuploid cells).

Figure 9 shows the evolution of the first 10 L-moments variances on a real file of 212 cells. At order 70, a new cell appears, near the 4c position, and is detected by an increase of the first 4 L-moments variances, but not by the others. These results confirm observations made with the synthetic file of Figure 3. Furthermore, the variances of the same first L-moments increase, in response to the new advent of a 4c cell of order 130, but with a very low amplitude.

Figure 10 shows the evolution of the first 10 L-moments bootstrap variances on a real file of 100 cells. As it is evident from this figure, no stabilisation was observed. Even if the 10 Bootstrap variances decrease and seem to enter into a stabilisation phase, such stabilisation is not enough to order the arrest of acquisition. This histogram is not stable and is not sufficient for the purposes of diagnosis or prognosis.

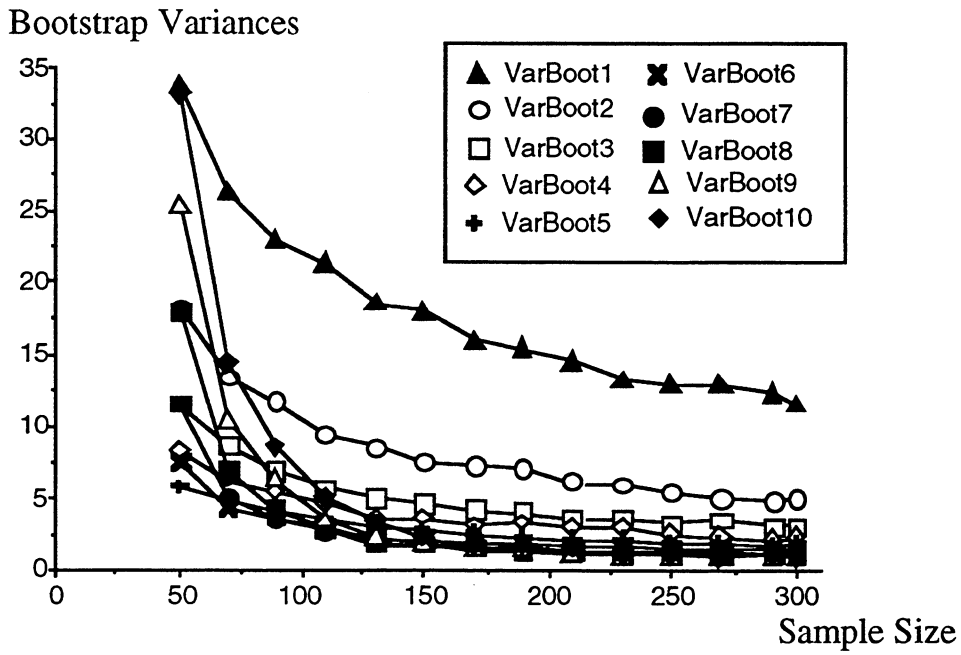


Figure 5 : Evolution of the first 10 L-moments Bootstrap Variances as sample size increases for an histogram which exhibits a diploid normal distribution (Mean = 2000, CV = 10%).

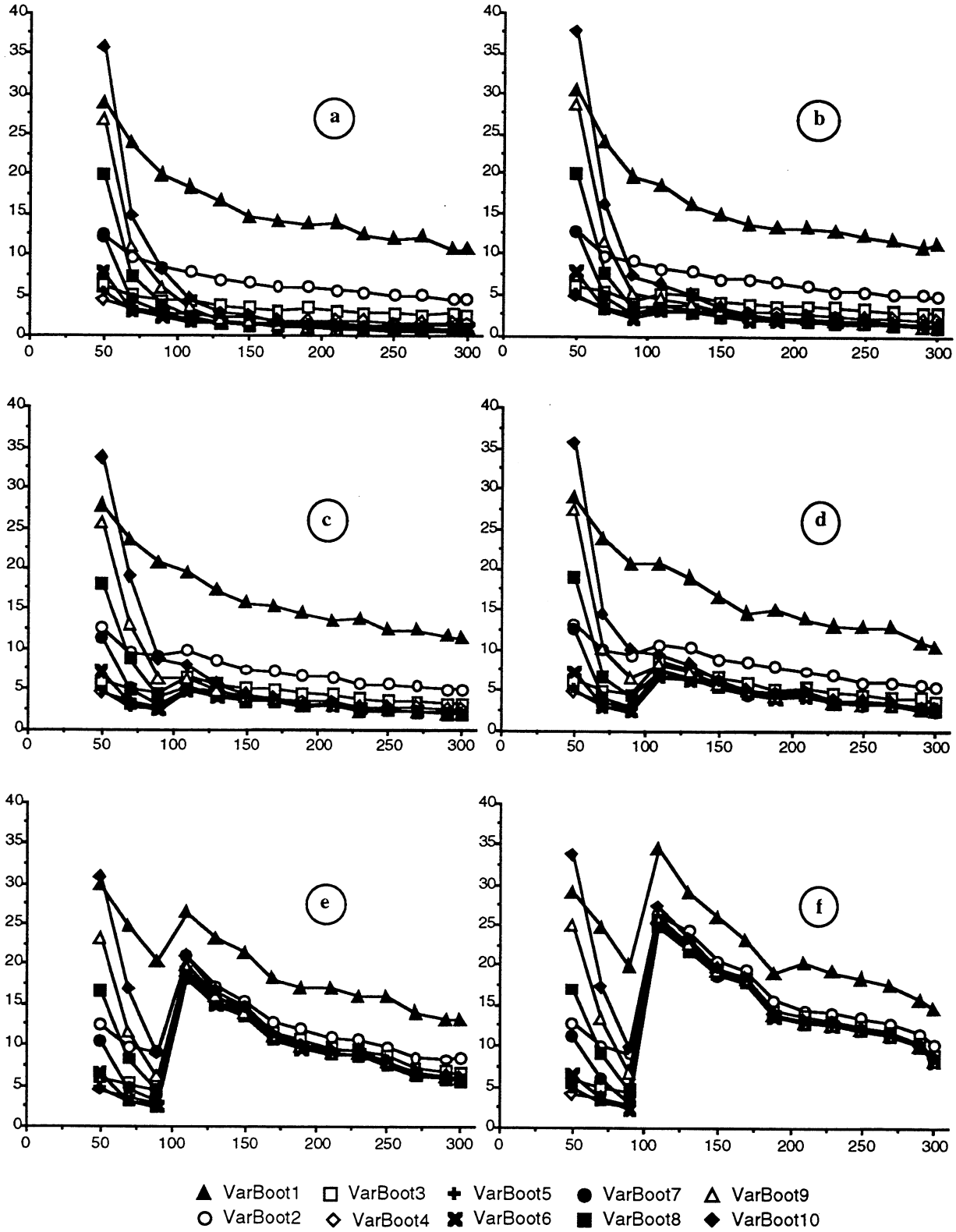


Figure 6 : Evolution of the first 10 L-moments Bootstrap Variances as sample size increases. Until sample of 109 cells, DNA histogram exhibits a normal diploïd distribution (mean = 2000, CV= 10%). In 110^{th} position, a cell appears with a respectively value of : 2500 (**a**), 3000 (**b**), 3500 (**c**), 4000 (**d**), 5000 (**e**) and 8000 (**f**).

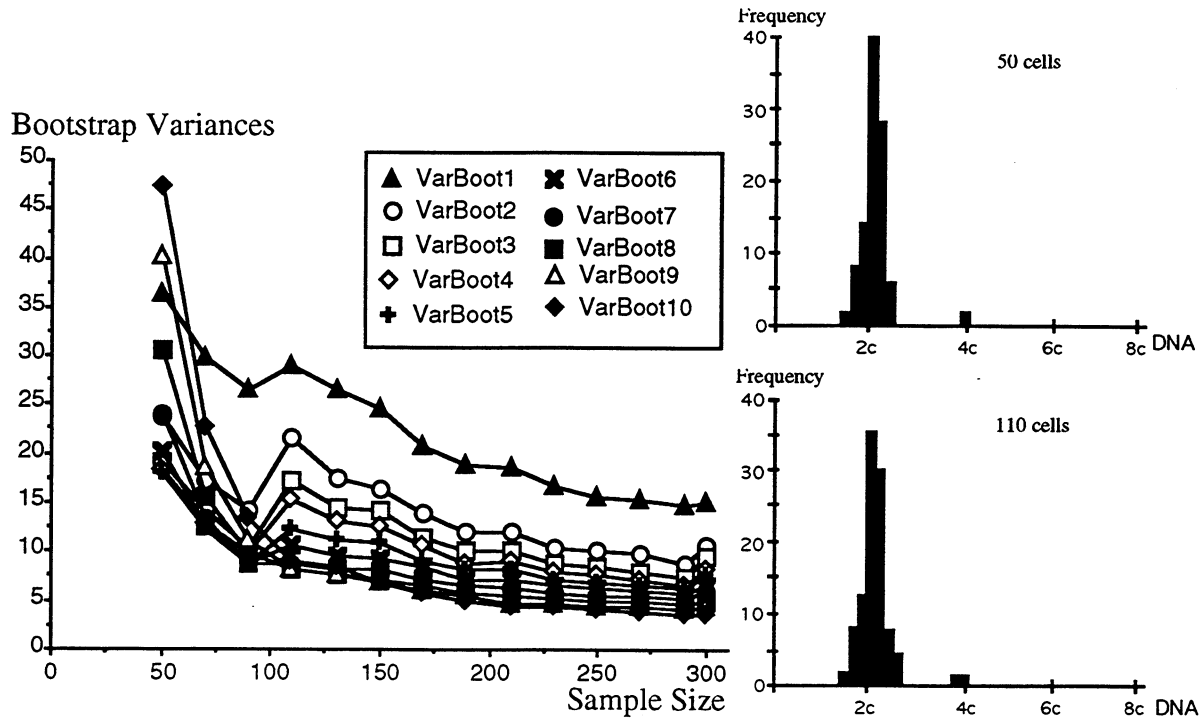


Figure 7 : Evolution of the first 10 L-moments Bootstrap Variances as sample size increases. DNA histogram of 50 cells and 110 cells are illustrated. At 110th position, a 4c cell appears. Such advent can be detected by an increasing of the first 4 L-moments Bootstrap variances.

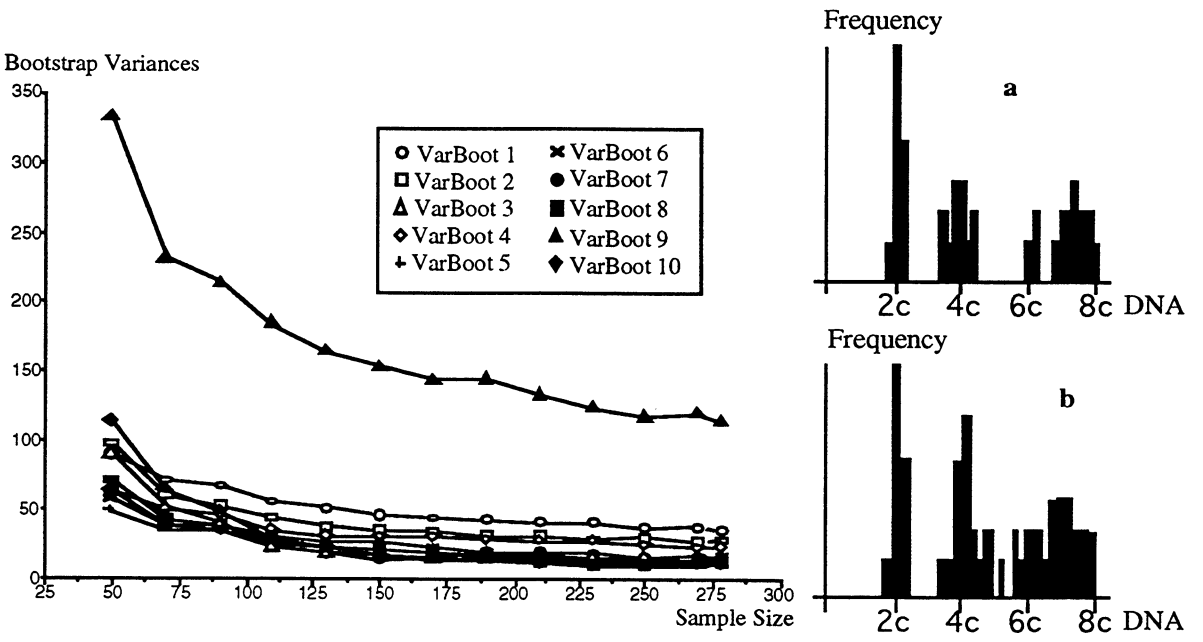


Figure 8 : Evolution of the first 10 L-moments Bootstrap variances as sample size increases. DNA histograms of the sample of first 50 analysed cells (a) and of the total analized sample (278 cells) (**b**) are illustrated.

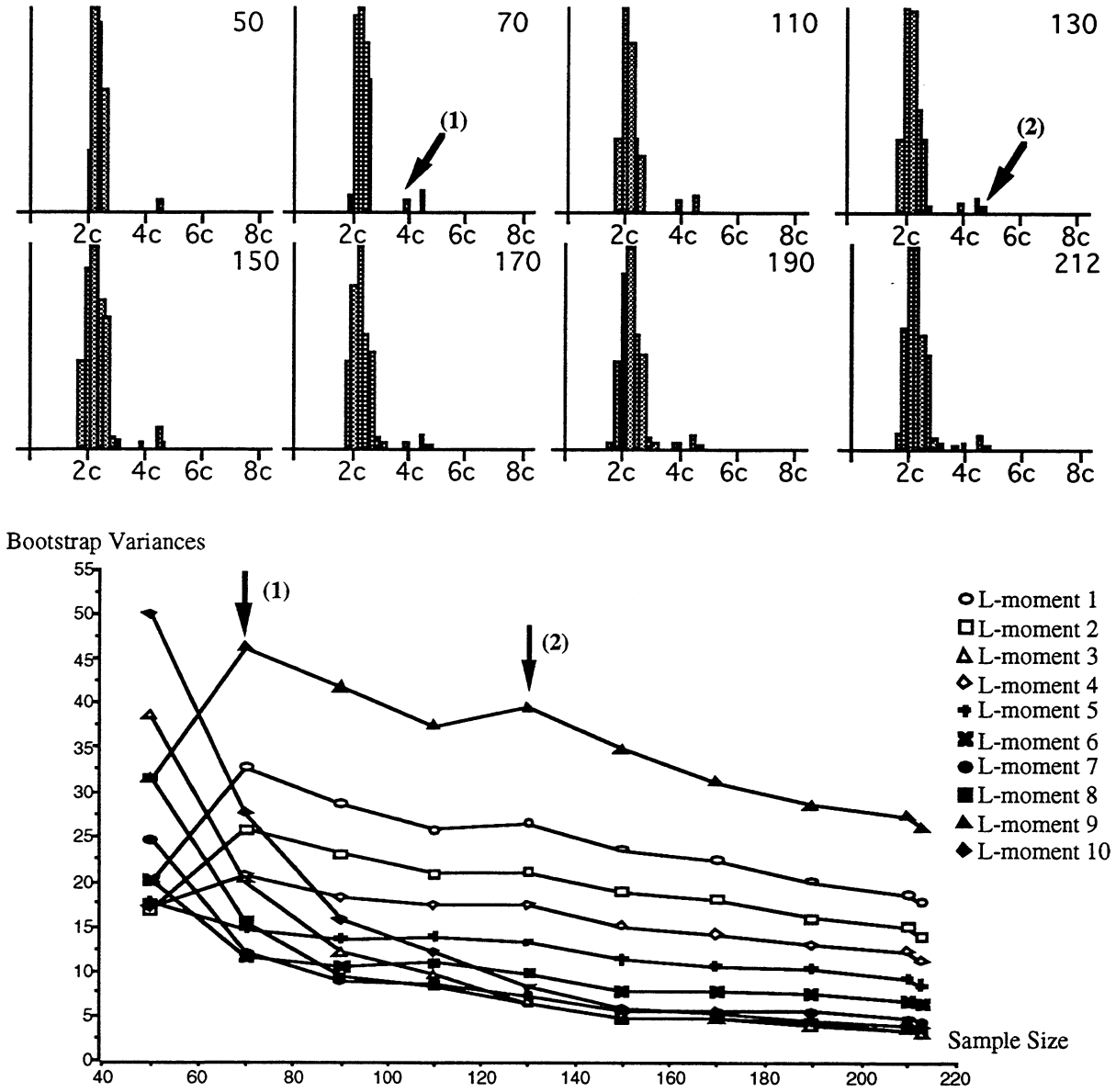


Figure 9 : Evolution of the first 10 L-moments Bootstrap Variances as sample size increases. DNA histograms from samples of 50 cells to 212 cells are illustrated. At 70 cells sample, a 4c cell appears (**1**). Such advent can be detected by an increasing of the first 4 L-moments Bootstrap variances. At 130 cells sample, a 4.5c cell appears (**2**). Such advent can be detected by a lowest increase of the first 4 L-moments Bootstrap variances than the previous advent.

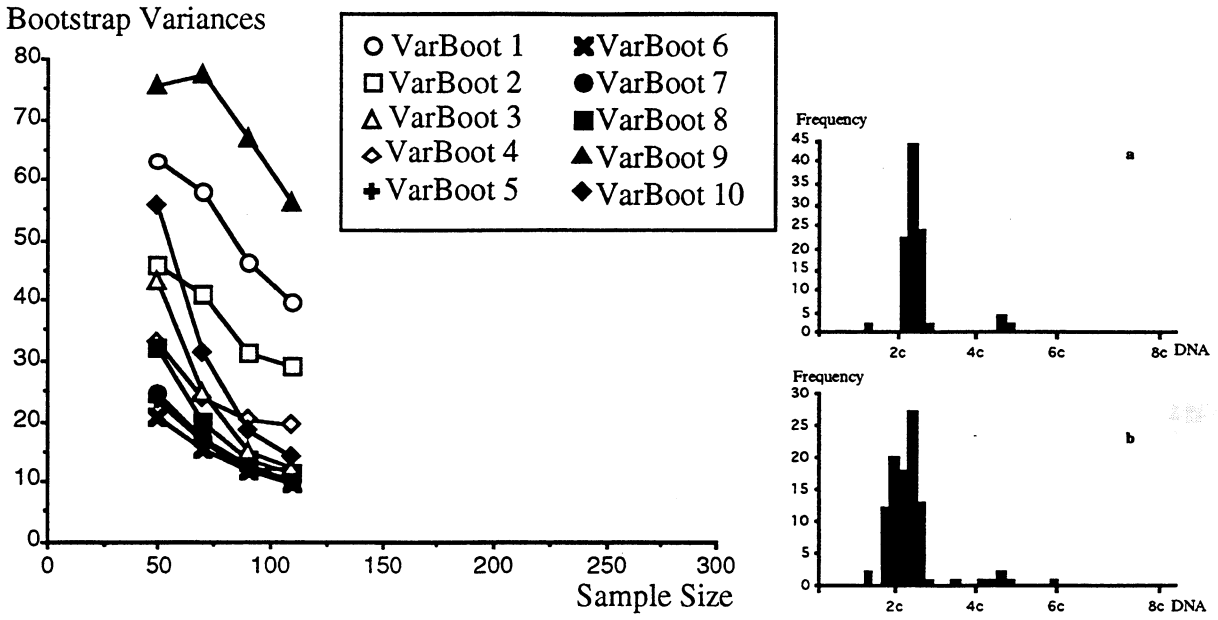


Figure 10 : Evolution of the first 10 L-moments Bootstrap Variances as sample size increases. The DNA histograms of the first 50 analysed cells (**a**) and of the total analysed sample (100 cells) (**b**) are illustrated.

IV. DISCUSSION OF STRATEGIES

The approach proposed in this paper offers the possibility to observing the stability phase of the histogram evolution as sample size increases. We do not conclude that the sample of n cells is representative of the total tumour, considering its DNA content. More modestly, we proposed an efficient method, which may answer the question : does a DNA histogram, obtained from a sample of **n** cells, reach stability or not ?

We then considered the DNA histogram as a pattern. Pattern recognition has been investigated by means of L-moments. These L-moments permit the reduction of the number of features in the distribution without too much loss of information. In general, the number of L-moments needed to code histograms varies from 10 to 15 [12], while the number of classes commonly used to code DNA histogram varies on average from 30 to 100.

The bootstrap variance of L-moments permits us to follow the stability of DNA histograms in which the pattern changes during cell acquisition. The use of the bootstrap method to generate variance of L-moments allows a good solution to the problem of sampling, which is essential in DNA histogram interpretation. Furthermore, the bootstrap method allows one to drop the L-moments distribution hypothesis[8].

The results, obtained both on synthetic data and real data, seems to answer positively to our initial requests : rare event detection, stop analysis criteria, DNA histogram stabilisation control.

Rare event detection

A rare event, in DNA histogram analysis, is defined as the advent of a cell or a few cells in a new position in comparison with the DNA values of the cells already analysed. If in a given sample, the DNA histogram exhibits diploid peak with few cells in synthesis or in 4c region, a rare event could be a cell with a DNA content higher than 5c, for instance. A rare event can also be a cell with a DNA value in 3c position, while the established DNA histogram shows a diploid peak and a few cells in 4c region.
Despite the variability of the values of a rare event, our model is able to detect the advent of this cell.

DNA histogram stabilisation and *stop analysis criterion*

An histogram is said to be stabilised if firstly, the higher order L-moment variances (from order 6 to 10) are stabilised, and secondly, at the same time, the lower order L-moments variances (from order 1to 5) remain stable or slightly decrease. One hundred real DNA histograms have then been analysed by simulating real time data acquisition. Our strategy, based on stability of the L-moment variances, gave uniform and homogeneous results. We propose then to use it also as *stop analysis criterion* : acquisition of data can be halted when DNA histogram satisfies the above definition of stability.

The proposed method showed its capacity to propose a solution to the problem of representativity of a DNA histogram at an individual level and not at a general level. Nethertheless, we have keep in our mind that the Bootstrap method is known to be unreliable and should be used with caution. Future theoretical studies will be necessary to determine the theoretical validity of this method.
A quality control of DNA analysis in the clinical laboratory, based on our methodology, is in the process of investigations. Other parameters, such as clinical data of the patient under consideration or the type of the tumour, may be integrated into the data acquisition algorithm under statistical control. For this purpose, tools of Distributed Artificial Intelligence, and the use of new statistical theories, such as possibility theory, [6], will be studied to formulate an efficient quality control algorithm.

REFERENCES

1 Auer G.U., Caspersson T.O. and Wallgren A.S. DNA Content and Survival in Mammary Carcinoma. Analyt Quant Cytol Histol 1980; 2 : 161- 165.

2 Auer G.U., Falkmer UG and Zetterberg AD. Image cytometric nuclear DNAanalysis in clinical tumour material. pp213-231. *In* Manual of Quantitative pathology in cancer diagnosis and prognosis. JPA Baak, Springer-Verlag Berlin Heidelberg 1991.

3 Bartels P.H. Numerical evaluation of cytological data. I. Description of profil. Analyt Quant Cytol Histol; 1979; 1: 20-28

4 Bartels P.H. Numerical evaluation of cytological data. II. Comparison of profiles Analyt Quant Cytol Histol 1979; 1 : 77-83.

5 Bartels P.H, Weber J.E and Bibbo M. Ploidy Pattern Analysis, Statistical Considerations.Analyt Quant Cytol Histol 1985 ; 7 : 126-130

6 Bartels P.H., Thompson D. and Weber J.E. Expert systems in histtopathology V. DS theory, certainty factors and Possibility theory. Analyt Quant Cytol Histol 1992 ; 14 : 165-174.

7 David H.A. Order Statistics, 2nd edn. New York:Wiley, 1981

8 Efron B."The Jacknife, the Bootstrap and other Resampling Plans". S.I.A.M., Philadelphia, Pennsylvania, 1982.

9 Efron B. " Computers and the theory of Statistics ; Thinkink the unthinkable." SIAM REVIEW 1979 ; 21 : 460-480.

10 Efron B."Bootstrap methods : Another look at the jacknife". The annals of statistics 1979; 7 : 1-26.

11 Grenwood, J.A, Landwehr, J.M., Matalas, N.C. and Wallis, J.R. Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. Water Resour. Res. 1979 ; 15 :1049-1054.

12 Guillaud M. and Chassery J.M. Histograms analysis by use of L-moments, linear functions of order statistics. Statistique et Analyse des Données 1991 ; 16 : 85-106

13 Guillaud M. and Chassery J.M. Apport de la méthode du Bootstrap pour l'élaboration d'un critère d'arrêt en reconnaissance des formes. Innov. Tech. biol. Med 1990; 11: 544-558.

14 Hosking J.R.M.The theory of probability weighted moments. Research report RC12210. IBM Research, Yorktown Heights.1986

15 Hosking J.R.M. Some theoretical results concerning L-moments. Research report RC14492. IBM Research, Yorktown Heights.1989

16 Hosking J.R.M. L-moments: Analysis and estimation of distributions using Linear combinations of order statistics. J.R. Statist. Soc. B 1990 ; 52 : 105-124.

17 Konheim A.G. A note on order statistics. Ann. Math. Mthly 1971 ; 71 : 524-530, .

18 Lanczos C. Applied Analysis.286-295. *In* : London: Pitman, 1957

19 Meyer J.S., Wittliff J.L.Regional heterogeneity in breast carcinoma: thymidine labelling index, steroid hormones receptors, DNA ploidy. Int. J. Cancer 1991 ; 47 : 213-220.

20 Pennes D.R., Naylor B. and Rebner M.Fine Needle Aspiration Biopsy of the Breast. Influence of the number of passes and the sample size on thediagnostic Yield.

Acta cytologica 1990 ; 34 : 673-676.

21 Salmon I., Coibon M., Larsimont D., Badr-El-Din A., Verhest A., Pasteels J.L and Kiss R. Comparison of Fine needle Aspirates of Breast Cancer to Imprint Smears by Means of digital Cell image Analysis. Analyt Quant Cytol Histol 1991 ; 13 : 193-200.

22 Sara A. and El-Naggar A.K. Intratumoral DNA content Variability. A study of non-small cell Lung Cancer. American Journal of Clinical Pathology 1991 ; 96 : 311- 317.

23 Silitto GP.Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample. Biometrika, 1969 ; 56 : 641-650.

24 Wied GL., Bartels PH., Bibbo M. and Dytch HE. Image Analysis and quantitative cyto- and histopathology. Techn. Report 1988 ; 2, The internationnal academy of cytology.

III. Acquisition de données multi-dimensionnelles sous contrôle statistique

Nous avons vu comment la méthode du Bootstrap pouvait être utilisée pour étudier la stabilité d'histogrammes d'ADN, codés sous formes vectorielles. Certaines applications biologiques et médicales nécessitent le calcul de plusieurs paramètres pour décrire une population cellulaire. Dans la majorité des cas, des analyses multifactorielles sont nécessaires pour représenter ces populations et discriminer les éventuelles sous-populations. L'établissement de formules leucocytaires, par une discrimination des groupes sanguins basée sur des paramètres denstiométriques et texturaux, en est une parfaite illustration.

Nous nous sommes donc intéressé à l'intérêt de la méthode du bootstrap dans ce genre de situations. La démarche suivit est la même que celle utilisée pour les histogrammes d'ADN

- Recherche des meilleurs descripteurs de la population. La représentation de données multidimensionnelles dans un espace engendré par des combinaisons linéaires des variables initiales, permet de réduire l'espace de représentation. Les valeurs propres issues d'une Analyse en Composantes Principales représentent le pourcentage de l'inertie totale expliquée par chacune des nouvelles composantes principales. Nous utiliserons donc ces valeurs propres comme "descripteurs " du nouvel espace de représentation.

- Mesure de la variabilité de ces valeurs propres par la méthode du Bootstrap. La valeur des variances Bootstrap donne une estimation de la stabilité du nouvel espace de représentation engendré par les différentes composantes principales. Cette technique a déjà été étudié entre autre par Holmes-Junca [Holmes-Junca 1985] et Besse [Besse 1989].

- La stabilisation des variances Bootstrap des valeurs propres, quand le nombre d'individus analysés augmente, est utilisée comme critère de stabilité de la population.

Ces travaux ont donné lieu à la publication suivante :

Apport de la méthode du bootstrap pour l'élaboration d'un critère d'arrêt en reconnaissance des formes
Innov. Tech. Biol. Med., 11, 544-558, 1990

APPORT DE LA METHODE DU BOOTSTRAP POUR L'ELABORATION D'UN CRITERE D'ARRET EN RECONNAISSANCE DES FORMES

GUILLAUD M., CHASSERY J.M.

RESUME

Le comportement du test du bootstrap en acquisition de données sous contrôle statistique dynamique a été étudié. Nous avons démontré la capacité de ce test à détecter l'homogénéité ou l'hétérogénéité d'un échantillon en estimant la stabilité des valeurs propres issues d'une analyse en composantes principales. Ce test a été appliqué à des données réelles et a permis de déterminer un seuil de stabilité de la population (diminution de l'hétérogénéité de l'échantillon en fonction de la taille) qui pourrait être utilisé comme critère d'arrêt de l'acquisition des données. Le test du bootstrap, associé à un test d'arrêt, sera intégré dans un système de prédiction afin d'être implémenté sur un système d'analyse d'image en microscopie quantitative.

ABSTRACT

This report is concerned with the bootstrap test for data acquisition under statistic control. We shown that this test detects the homogeneity or heterogeneity of a sample by estimation of stability of eigenvalues using principal components analysis. This test was applied to real data and permitted the evaluation of a population stability threshold wich can be used to stop data acquisition when the heterogeneity of the sample declines (as the size increases). The bootstrap test will be integrated into an image analysis system software in quantitative microscopy for the purpose of introducting an automated data acquisition termination.

MOTS CLES :

TEST DU BOOTSTRAP. ANALYSE EN COMPOSANTES PRINCIPALES. TEST D'ARRET. CYTOMETRIE A BALAYAGE. ECHANTILLONAGE.

APPORT DE LA METHODE DU BOOTSTRAP POUR L'ELABORATION D'UN CRITERE D'ARRET EN RECONNAISSANCE DES FORMES.

GUILLAUD M.*, CHASSERY J.M.*
*Equipe de Reconnaissance des Formes et Microscopie Quantitative, Laboratoire TIM3-IMAG, U.A CNRS N°397, Université Joseph Fourier, CERMO, BP 53X. 38041 Grenoble Cedex.

INTRODUCTION

Les analyseurs d'image au microscope trouvent dès maintenant leur place dans les laboratoires d'anatomo-pathologie pour l'aide au diagnostic et au pronostic des tumeurs solides (1). La fiabilité des résultats d'une analyse par cytométrie à balayage est une exigence médicale primordiale. Cependant, les contraintes économiques et médicales imposent aux pathologistes d'analyser le plus rapidement possible les prélèvements effectués sur les patients. Or le temps de ces analyses dépend du nombre d'objets examinés sur une lame, nombre généralement associé à l'expérience du pathologiste. Cette détermination empirique dépend de la nature de l'application (2).

D'autre part, pour une même application, le nombre de cellules nécessaire à l'obtention d'un résultat fiable varie en fonction de la nature de l'échantillon, de la qualité du prélèvement, et il peut varier, par conséquent, d'une lame à l'autre.

Le but du travail présenté est d'élaborer un protocole statistique permettant de définir, pour chaque lame analysée, un nombre minimum de cellules satisfaisant la représentativité de la population totale de la lame étudiée et respectant les proportions des différents groupes éventuellement présents dans la population cellulaire. Il s'agit donc d'offrir à l'utilisateur du système d'analyse d'image, le moyen de suivre l'évolution des caractéristiques de l'échantillon au fur et à mesure de l'acquisition des données et de stopper l'analyse dès que le nombre de cellules acquises sera statistiquement et biologiquement représentatif de toute la population cellulaire.

Les systèmes d'analyse d'image appliqués à la cytométrie à balayage créent des fichiers contenant les valeurs de p paramètres quantitatifs mesurés sur n objets (cellules ou noyaux). Pour traiter de tels tableaux de données, il est souvent nécessaire de faire appel aux outils de l'analyse de données multi-dimensionnelles comme l'analyse en composantes principales (ACP). L'étude de la variabilité des valeurs propres extraites d'une ACP revient à étudier la stabilité des résultats en reconnaissance de formes, la forme étant une représentation vectorielle paramètrée.

La méthode du bootstrap, une des nouvelles méthodes de calcul statistique intensif sur ordinateur (3,4,5) permet de mesurer la précision de certains paramètres statistiques à partir d'un seul échantillon. Il parait donc intéréssant de l'adopter et de l'évaluer en cytométrie.

Dans un premier temps, il s'agit de tester la capacité de la méthode bootstrap à différencier un échantillon représentatif d'un échantillon non représentatif d'une population cellulaire. Dans

un deuxième temps, il s'agit d'appliquer cette méthode en acquisition de données sous contrôle statistique dynamique, en étudiant son comportement pour des échantillons de taille variable.

MATERIEL ET METHODE

I. Présentation des fichiers étudiés : étude de cas exemplaires

I.1. Nature des fichiers

Les fichiers utilisés pour cette étude sont tous issus du système d'analyse d'image SAMBA 2005 (TITN ALCATEL). Les différents fichiers sont caractérisés par une structure bien définie, mise en évidence par le plan factoriel principal issu d'une Analyse en Composantes Principales (figure 1).

L'ordre des individus dans ces fichiers correspond à l'ordre dans lequel ils ont été observés au cours de l'acquisition par le système d'analyse d'image. En général, l'utilisateur procède à un balayage systématique de la lame de gauche à droite et de haut en bas. En considérant que la répartition des individus sur la lame est homogène, les n premiers individus du fichier peuvent donc être considérés comme un tirage aléatoire de n individus parmi les N individus présents sur toute la lame.

- Fichier AMDTOT

Ce fichier contient 204 cellules sur lesquelles 6 paramètres ont été mesurés. Une ACP effectuée sur ce fichier a permis de séparer deux groupes d'effectifs sensiblement égaux, correspondant à une réalité biologique (figure 1A) :

- groupe 1 : 100 individus
- groupe 2 : 104 individus

- Fichier FOIE

Ce fichier contient 334 cellules sur lesquelles 8 paramètres ont été mesurés. Une ACP effectuée sur ce fichier a permis de séparer quatre groupes d'effectifs inégaux, correspondant à une réalité biologique (figure 1B) :

- groupe 1 : 156 individus
- groupe 2 : 107 individus
- groupe 3 : 67 individus
- groupe 4 : 4 individus

- Fichier FORNB

Ce fichier contient 273 individus sur lesquelles 15 paramètres ont été mesurés. Une ACP effectuée sur ce fichier ne dévoile aucune structure caractéristique (figure 1C) : ce fichier n'est donc constitué que d'un seul groupe d'individus, mais possédant une grande variabilité.

I.2. Création de fichiers "synthétiques"

En vue d'étudier la capacité du test du bootstrap à détecter une éventuelle répartition hétérogène de différents groupes d'individus sur la lame, nous avons créé des fichiers synthétiques correspondant à des situations extrêmes où chaque groupe d'individus serait acquis l'un après l'autre.

A partir du fichier **foie**, nous avons créé deux fichiers "synthétiques" :

- le fichier **foie1234** qui contient dans l'ordre les individus du groupe 1, puis ceux du groupe 2, puis ceux du groupe 3, et enfin ceux du groupe 4.
- le fichier **foie1432** qui contient dans l'ordre les individus du groupe 1, puis ceux du groupe 4, puis ceux du groupe 3, et enfin ceux du groupe 2.

II. La méthode du Bootstrap

La méthode du bootstrap a été mise au point en 1977 par Bradley Efron (5). Cette méthode, très simple dans son principe, est si dépendante de l'ordinateur qu'elle eut été inapplicable il y a 30 ans. Cette procédure permet d'évaluer la précision statistique d'un paramètre ou de tout autre résultat scientifique à partir des données d'un seul échantillon. Le principe consiste à créer, à partir de l'échantillon original, un grand nombre d'échantillons par tirages aléatoires. Sur chacun de ces échantillons, le paramètre étudié est calculé afin de fournir une estimation de la valeur calculée pour l'échantillon original. Cette estimation est souvent donnée par un intervalle correspondant à la distribution simulée par Bootstrap.

L'application du test du Bootstrap à l'étude de la variabilité des valeurs propres extraites d'une ACP (shéma 1) a été effectuée selon l'algorithme suivant :

Soient les n premiers individus des fichiers totaux contenant N individus. Une ACP est effectuée sur cet échantillon et donne p' nouveaux paramètres (p'=p). Soient $\hat{\lambda}_1$, $\hat{\lambda}_2$, ..., $\hat{\lambda}_{p'}$, les valeurs propres extraites de l'ACP, variances de ces nouveaux paramètres. On s'intéressera aux q premières valeurs propres $\hat{\lambda}_1$, $\hat{\lambda}_2$, ..., $\hat{\lambda}_q$ dont la somme, notée $\hat{\lambda}_{som}$, correspondant aux q premières composantes principales, explique au moins 80% de l'inertie totale du nuage de point ($q < p'$).

1 - Soit $\hat{F}$ la distribution empirique des données des n objets observés c'est à dire la distribution de probabilité donnant un poids de $1/n$ pour chaque objet **Xi** observé (Xi1, Xi2, ..., Xip).

2 - On extrait n nouveaux points ($X^*i1, X^*i2, ..., X^*ip$) par un générateur de nombres aléatoires indépendants à partir de $\hat{F}$, chaque nouveau point étant une sélection aléatoire d'un des n points originaux. Ces nouveaux points, qu'on appellera "échantillon Bootstrap", forment

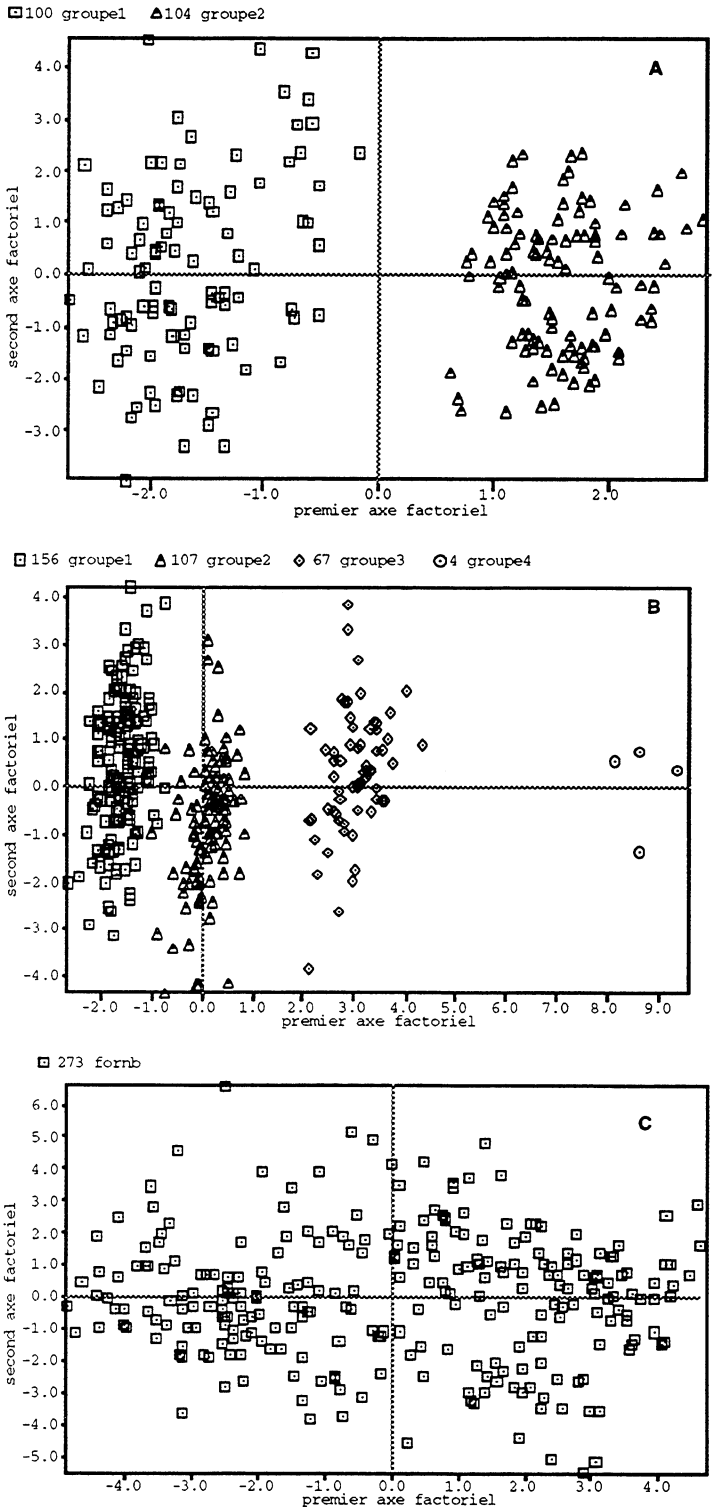


Figure 1 : **A)** Projection des individus du fichier **amdtot** dans le plan factoriel défini par les deux premières composantes principales. Deux groupes peuvent être séparés : le groupe 1 contenant 100 individus et le groupe 2 contenant 104 individus. **B)** Projection des individus du fichier **foie** dans le plan factoriel défini par les deux premières composantes principales. Quatre groupes peuvent être séparés : le groupe 1 contenant 156 individus, le groupe 2 contenant 107 individus, le groupe 3 contenant 67 individus et le groupe 4 contenant 4 individus. **C)** Projection des individus du fichier **fornb** dans le plan factoriel défini par les deux premières composantes principales. Ce fichier ne contient qu'un seul groupe d'individus.

un sous-groupe de l'échantillon des objets originaux. Certains d'entre eux seront "tirés" 0 fois, d'autres 1 fois, certains 2 fois, etc...

3 - On calcule les q premières valeurs propres pour cet échantillon Bootstrap (signalé par *) : $\hat{\lambda}^*_1, \hat{\lambda}^*_2, \ldots, \hat{\lambda}^*_q$ ainsi que la somme de ces q valeurs $\hat{\lambda}^*_{som.}$.

4 - On répète les étapes (**2**) et (**3**) un grand nombre de fois, disons T fois, pour chacune desquelles on utilise un nouvel ensemble de nombres aléatoires pour engendrer le nouvel échantillon Bootstrap. On obtient, après T "tirages", la série de valeurs suivantes :

- $\hat{\lambda}^{*1}_1, \hat{\lambda}^{*1}_2, \ldots, \hat{\lambda}^{*1}_q$ et $\hat{\lambda}^{*1}_{som.}$
- $\hat{\lambda}^{*2}_1, \hat{\lambda}^{*2}_2, \ldots, \hat{\lambda}^{*2}_q$ et $\hat{\lambda}^{*2}_{som.}$
- ..., ..., ..., ...
- $\hat{\lambda}^{*k}_1, \hat{\lambda}^{*k}_2, \ldots, \hat{\lambda}^{*k}_q$ et $\hat{\lambda}^{*k}_{som.}$
- ..., ..., ..., ...
- $\hat{\lambda}^{*T}_1, \hat{\lambda}^{*T}_2, \ldots, \hat{\lambda}^{*T}_q$ et $\hat{\lambda}^{*T}_{som.}$

5 - On définit les écarts Bootstrap $\hat{\sigma}^B_j$, estimateurs Bootstrap des écart-types σ de la distribution des valeurs de $\hat{\lambda}^*_1, \hat{\lambda}^*_2, \ldots, \hat{\lambda}^*_q$ et de $\hat{\lambda}^*_{som.}$:

Pour j = 1, 2, ..., p et j = som. :

$$\hat{\sigma}^B_j = \sqrt{\frac{\sum_{k=1}^{k=T} \left(\hat{\lambda}^{*k}_j - \bar{\lambda}^B_j \right)}{T-1}}$$

où $\bar{\lambda}^B_j$ est défini comme l'estimateur Bootstrap de la moyenne des valeurs de $\hat{\lambda}^*_1, \hat{\lambda}^*_2, \ldots, \hat{\lambda}^*_q$ et de $\hat{\lambda}^*_{som.}$:

Pour j = 1, 2, ..., q et j = som :

$$\bar{\lambda}^B_j = \frac{\sum_{k=1}^{k=T} \hat{\lambda}^{*k}_j}{T}$$

Cet écart-type Bootstrap mesure la variabilité des valeurs propres Bootstrap. En effet, la distribution des valeurs propres Bootstrap peut être approximée par une loi normale : 68% des

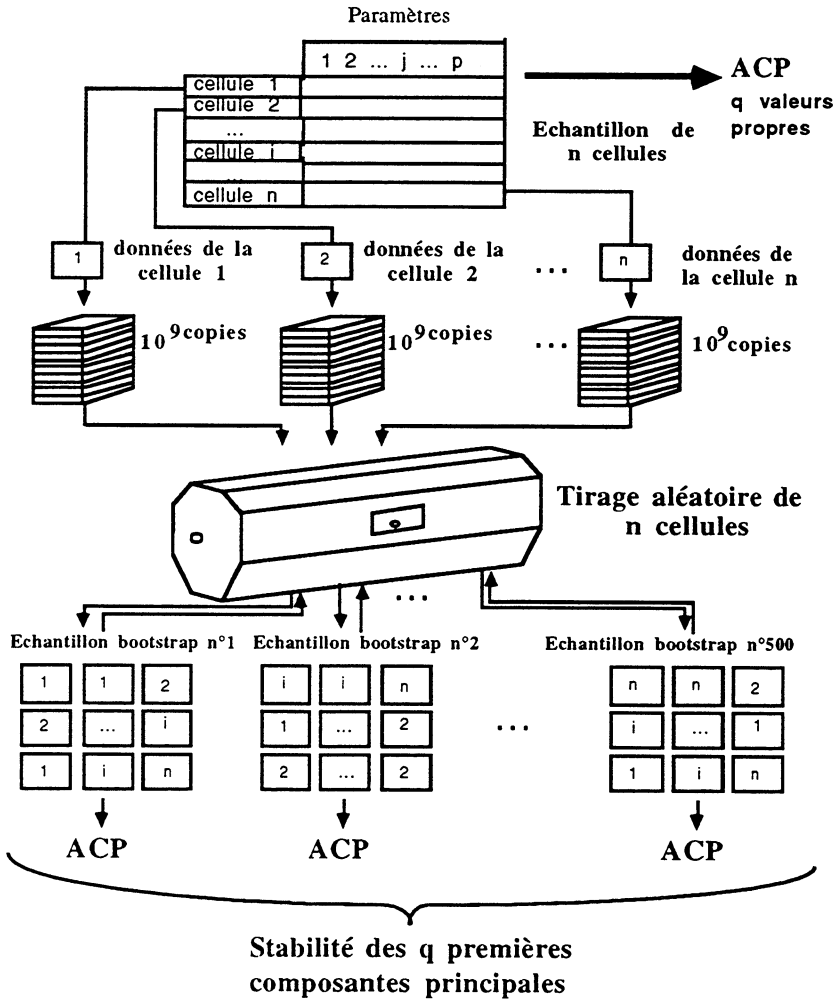
valeurs propres Bootstrap sont donc situées à moins d'un écart Bootstrap de la moyenne Bootstrap.

PRINCIPE

On effectue l'algorithme précédent pour des échantillons de taille variable : aux n premiers individus de départ on ajoute les x individus suivants dans l'ordre d'acquisition.

On calcule ainsi les écarts Bootstrap pour des échantillons de taille n, n+x, n+2x,... jusqu'à l'échantillon de taille N correspondant à la population totale. L'écart Bootstrap donne une estimation de l'écart-moyen entre les valeurs des valeurs propres Bootstrap pour des échantillons simulés et la valeur réelle.

Les écarts Bootstrap calculés sur la population totale mesurent en fait la variabilité "intrinsèque" de la population. En effet toute l'information est contenue dans ce fichier total et les valeurs propres correspondantes sont les valeurs propres réelles du fichier.



Shéma 1 (6). La méthode Bootstrap est appliquée ici à un échantillon de n cellules afin d'évaluer la précision des valeurs propres calculées sur cet échantillon. Ce procédé consiste à recopier les données, un milliard de fois par exemple, à les mélanger pour créer des échantillons artificiels de n unités tirées au hasard. On calcule les valeurs propres pour chacun des échantillons ainsi simulés ce qui permet d'évaluer leurs variations statistiques.

RESULTATS

I. Fiabilité de la méthode Bootstrap

Les premiers travaux traitant de la fiabilité du test du Bootstrap ont montré que cette méthode donnait, dans la plupart des cas, une bonne estimation de la variabilité de paramètres calculés sur un échantillon, comme par exemple les composantes principales issues d'une ACP (5). Toutes ces études ont confirmé que la méthode Bootstrap était statistiquement fiable, mais l'origine récente de cette méthode fait, qu'à ce jour, aucune application du Bootstrap à un problème d'acquisition de données en temps réel, n'a été proposée.

La mesure de la variabilité des valeurs propres pour un échantillon de taille n donné était-elle suffisamment précise pour donner une idée de la représentativité de cet échantillon ? Les résultats obtenus sur les fichiers "synthétiques" apportent des éléments de réponse intéressants.

I.1. Fichier **foie1234** (Graphique 1)

On constate pour les cinq premiers échantillons du fichier **foie1234** une diminution exponentielle des trois écarts Bootstrap. L'apparition des individus du groupe 2 correspond à une augmentation nette de l'écart Bootstrap 1(**a**) alors que l'écart Bootstrap 2 n'augmente qu'à partir de l'échantillon 190. A partir de l'échantillon 210 les écarts Bootstrap 1 et 2 décroissent jusqu'à l'échantillon 263 (qui contient tous les individus du groupe 1 et 2). L'apparition des individus du groupe 3 correspond à une légère augmentation de ces deux écarts Bootstrap (**b**). L'écart Bootstrap 1 continue à augmenter jusqu'à l'échantillon 334. L'augmentation plus accentuée des écarts Bootstrap 1, 2 et som. de l'échantillon 330 à l'échantillon 334 correspond à l'apparition des individus du groupe 4 (**c**).

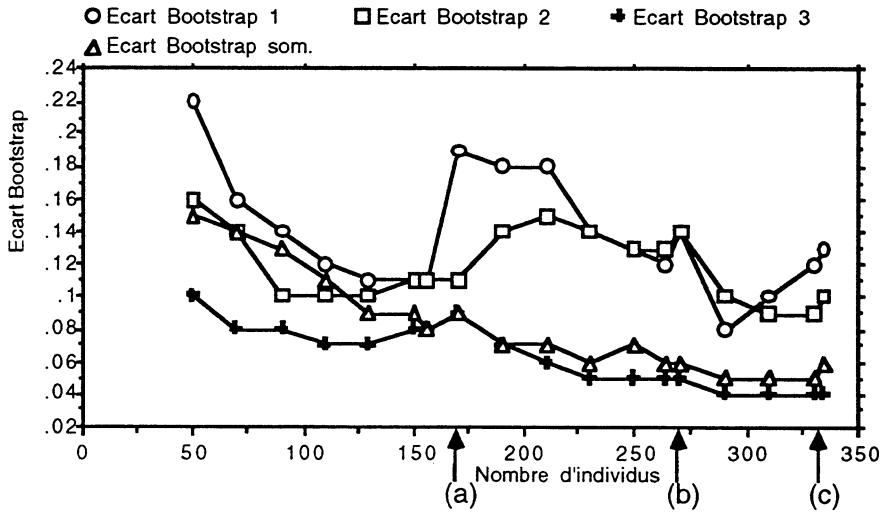
I.2. Fichier **foie1432** (Graphique 2)

On constate pour les cinq premiers échantillons du fichier **foie1432** une diminution exponentielle des trois écarts Bootstrap. Ceci confirme les observations faites sur le fichier **foie1234**. L'apparition des individus du groupe 4 et de 10 individus du groupe 3 (échantillon170) correspond à une augmentation très nette de l'écart Bootstrap 1, et également de l'écart Bootstrap 2 et som (**a**). Les quatre écarts Bootstrap suivent ensuite une décroissance similaire, entrecoupée de légères augmentations (écarts Bootstrap 1 et 2) au moment de l'apparition des individus du groupe 2 (**b**).

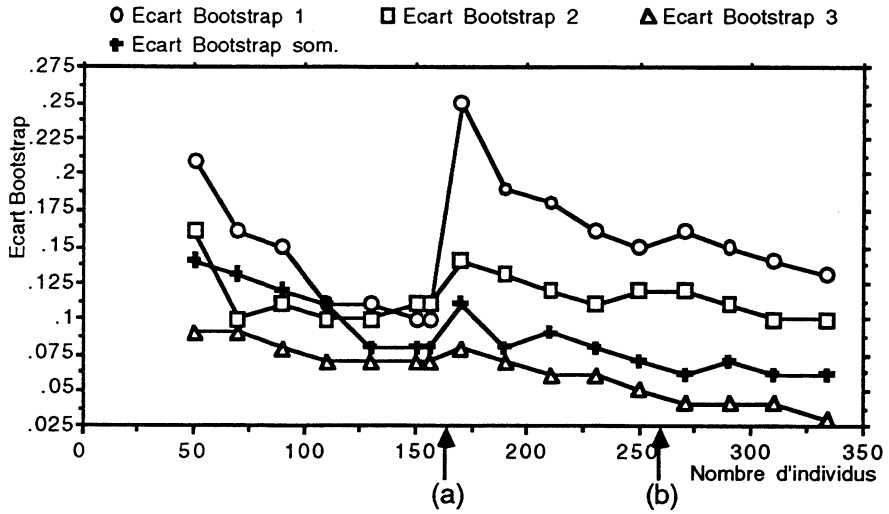
L'étude de ces fichiers "synthétiques" montre, tout d'abord, que toute apparition d'une information nouvelle dans un échantillon est détectée par l'augmentation des écarts Bootstrap, en particulier des écarts Bootstrap 1 et 2 qui correspondent aux valeurs propres définissant le plan factoriel principal. Une information nouvelle correspond à l'apparition d'individus d'un type cellulaire différent des individus précédemment acquis. On peut donc parler de **fiabilité** de la méthode Bootstrap.

Par ailleurs, la comparaison des graphiques 1 et 2 est riche d'autres enseignements. En effet, l'apparition des individus du groupe 4 du fichier **foie1432** provoque une augmentation plus grande de l'écart Bootstrap 1 (0,15) que l'apparition des individus du groupe 2 du fichier **foie1234** (0,08). Or si l'on examine la projection des individus du fichier **foie** dans le plan factoriel (figure 1B), on remarque que le groupe 4 est plus "différencié" du groupe 1 que ne l'est le groupe 2. Nous avons vérifié, en répétant le test du Bootstrap sur ces mêmes fichiers, que *l'importance de l'augmentation de l'écart Bootstrap 1 était toujours proportionnelle à la nature de l'information nouvellement acquise.*

Ces résultats (confirmés sur d'autres fichiers) démontrent clairement que les écarts Bootstrap donnent de bonnes estimations de la variabilité des composantes principales, correspondant à la variabilité à l'intérieur des échantillon étudiés. La robustesse et la précision de ces estimateurs Bootstrap sont fiables.



Graphique 1 : Evolution des écarts Bootstrap 1, 2, 3 et som. en fonction de la taille de l'échantillon (en nombre d'individus) du fichier **foie1234**. 300 tirages Bootstrap ont été effectués.



Graphique 2 : Evolution des écarts Bootstrap 1, 2, 3 et som. en fonction de la taille de l'échantillon (en nombre d'individus) du fichier **foie1432.** 300 tirages Bootstrap ont été effectués.

II. Simulation d'acquisition de données sous contrôle statistique: comportement des écarts Bootstrap sur des fichiers réels

L'étude de cas réels a pour but d'étudier la capacité de la méthode Bootstrap à détecter une stabilité des formes obtenues par ACP.

II.1. Fichier **amdtot** (Graphique 3)

On suit l'évolution des écarts Bootstrap 1, 2 et som. en fonction de la taille de l'échantillon. Le fichier **amdtot** contient deux groupes d'effectifs sensiblement égaux.

L'hypothèse d'une distribution homogène de ces deux groupes sur la lame étudiée est confirmée par l'évolution des différents écarts Bootstrap, malgré la présence de deux "accidents" : on observe en effet une augmentation de l'écart Bootstrap 1 et som. pour l'échantillon 60 **(a)** et une augmentation de l'écart Bootstrap 2 pour l'échantillon 180 **(b)**. Ceci peut s'expliquer par les variations, mêmes minimes, du pourcentage d'individus des deux groupes d'un échantillon à l'autre (voir tableau ci-dessous).

L'écart Bootstrap som., qui augmente nettement pour l'échantillon 60, suit ensuite une décroissance régulière et se stabilise à partir de l'échantillon 160.

échantillon	20	40	60	80	100	120	140	160	180	204
Groupe1(%)	55	55	56	58	53	55	51	50	49	49
Groupe2(%)	45	45	44	42	47	45	49	50	51	51

II.2. Fichier **foie** (Graphique 4)

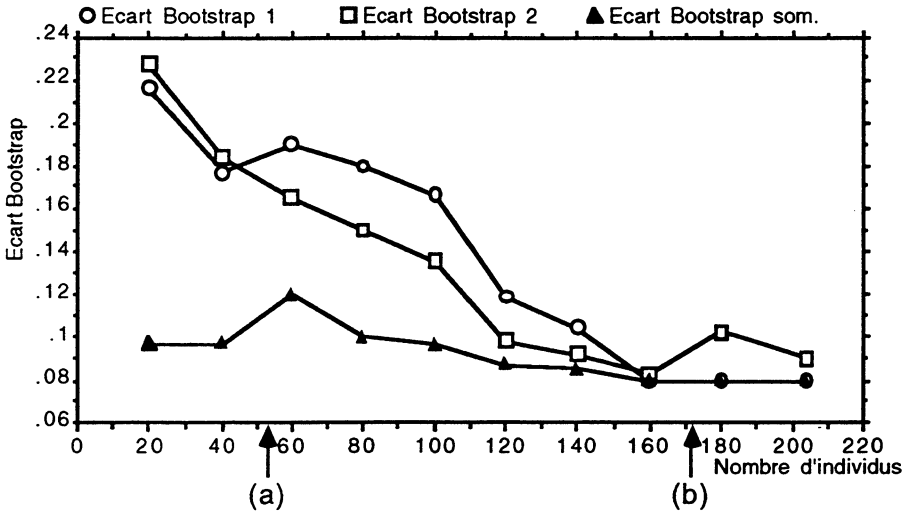
On suit l'évolution des écarts Bootstrap 1, 2, 3 et som. en fonction de la taille de l'échantillon. Le fichier **foie** contient quatre groupes d'effectifs inégaux (le groupe 4 ne contient que quatre individus soit 1,2 % de la population totale).

L'hypothèse d'une répartition homogène de ces quatres groupes sur la lame étudiée est confirmée par l'évolution des différents écarts Bootstrap. On observe en effet que les écarts Bootstrap 3 et som., après une décroissance continue, semble déjà se stabiliser à partir de l'échantillon 130. Par contre les écarts Bootstrap 1 et 2, qui ont de très fortes valeurs pour le premier échantillon, connaissent une décroissance en "dents de scie". Ce phénomène s'explique par les variations des pourcentages d'individus des différents groupes d'un échantillon à l'autre (voir tableau ci-dessous). On remarque d'ailleurs que la plus forte augmentation de l'écart Bootstrap 1 correspond à l'apparition du premier individu du groupe 4 (a). Nous avons vérifié, en introduisant cet individu dans des échantillons différents, que ceci correspondait toujours à l'augmentation la plus importante de l'écart Bootstrap 1 (par rapport aux fluctuations moyennes). La détection de l'apparition d'un seul individu, très différent des autres, confirme la **sensibilité** du test du Bootstrap.

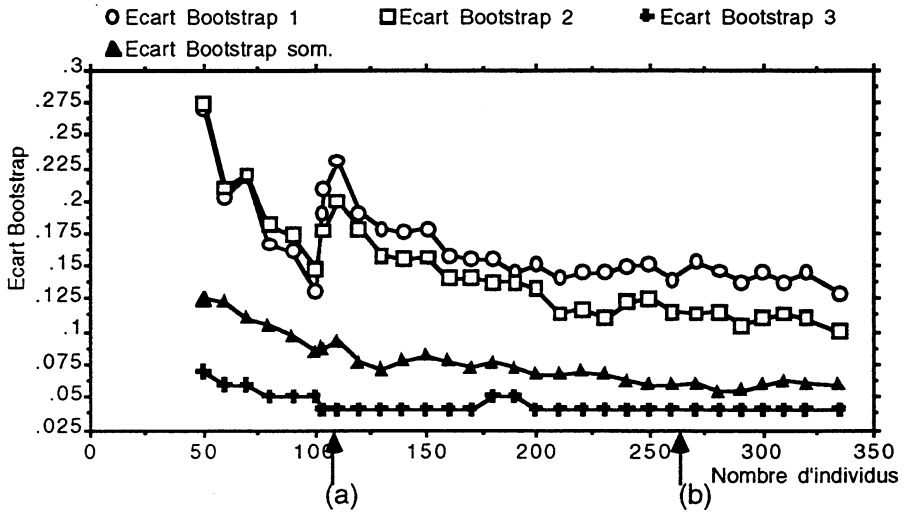
L'apparition de deux autres individus de ce groupe provoque également une augmentation de l'écart Bootstrap 1 (**b**). Si les fluctuations des deux écarts Bootstrap mettent en évidence la présence des différents groupes dans les échantillons, l'évolution de l'écart Bootstrap som., qui mesure la stabilité du plan principal, présente également un intérêt considérable. En effet, cet écart Bootstrap, après une phase de décroissance exponentielle se stabilise à partir de l'échantillon 190.

Echantillon	50	70	90	110	130	150	170	190	210	230	250	270	290	310	334
Groupe1(%)	50.0	47.2	45.6	47.3	47.7	47.4	49.5	50.5	51.9	50.5	48.0	47.4	46.5	46.7	46.7
Groupe2(%)	36.0	40.0	37.8	34.5	35.4	35.4	32.9	31.6	30.0	30.9	30.8	30.7	31.8	32.1	32.0
Groupe3(%)	14.0	12.8	16.6	17.3	16.2	16.6	17.1	17.4	17.6	18.2	20.4	20.8	20.7	20.0	20.1
Groupe4(%)	0.0	0.0	0.0	0.9	0.7	0.6	0.5	0.5	0.5	0.4	0.8	1.1	1.0	1.3	1.2

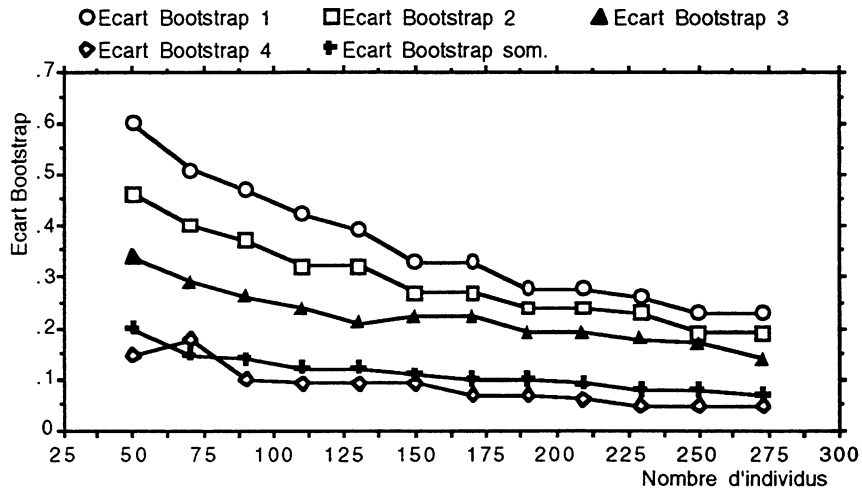
(a) (b)



Graphique 3. Fichier **amdtot** : Evolution des écarts Bootstrap 1, 2 et som. en fonction de la taille de l'échantillon (en nombre d'individus). 500 tirages Bootstrap ont été effectués.



Graphique 4. Fichier **foie** : Evolution des écarts Bootstrap 1, 2, 3 et som. en fonction de la taille de l'échantillon (en nombre d'individus). 300 tirages Bootstrap ont été effectués.



Graphique 5. Fichier **fornb** : Evolution des écarts Bootstrap 1, 2, 3, 4 et som. en fonction de la taille de l'échantillon (en nombre d'individus). 500 tirages Bootstrap ont été effectués.

II.3. Fichier **fornb** (Graphique 5)

On suit l'évolution des écarts Bootstrap 1, 2, 3, 4 et som. en fonction de la taille de l'échantillon. Le fichier **fornb** ne contient qu'un seul type d'individus. On peut donc s'attendre par rapport aux deux fichiers précédents à une évolution plus régulière des écarts Bootstrap : on observe de manière frappante la décroissance exponentielle comparable de tous les écarts Bootstrap qui tendent à se stabiliser à partir de l'échantillon 170. La dispersion des individus de ce fichier sur le plan principal (figure 1C) rend compte de la variabilité importante existant au sein de cette population. Cette variabilité peut expliquer les quelques faibles fluctuations observées chez certains écarts Bootstrap.

Application du test du Bootstrap en acquisition de données en temps réel : notion de test d'arrêt.

Pour tous les fichiers réels étudiés, nous avons pu observer une stabilisation des écarts Bootstrap pour des échantillons de taille bien inférieure à la population totale. On peut donc penser que de tels échantillons sont suffisamment représentatifs de la population totale pour stopper l'acquisition.

L'application du test du Bootstrap pour rechercher le nombre minimum d'individus nécessaires à acquérir et représentatif de la population totale doit être associée à un test d'arrêt. N'oublions pas qu'en situation réelle d'acquisition de données sous contrôle statistique, la population totale est inconnue. Il reste maintenant à intégrer cette méthode à un système de prédiction fondé sur l'évolution des écarts Bootstrap.

Ceci nécessite dans un premier temps de modéliser et de préciser le comportement des écarts Bootstrap. En effet, il reste à définir dans quelle mesure l'augmentation d'un écart Bootstrap peut être considérée comme significative (détection d'une information nouvelle) ou simplement révélatrice de variations dues à l'échantillonnage.

L'implémentation d'un test d'arrêt, comparable aux tests d'adéquations, nécessite de déterminer une valeur critique de l'écart Bootstrap en dessous de laquelle le cytopathologiste pourra stopper l'acquisition de ces données. Une telle valeur critique sera fonction de nombreux facteurs : nombre d'individus de l'échantillon, nombre de paramètres, nombre de composantes principales significatives, évolution des valeurs des écarts Bootstrap, etc..

Cette valeur critique devra tenir compte également de la nature de la préparation biologique, du phénomène biologique étudié, et en particulier du nombre de groupes biologiques qu'il est possible de rencontrer sur la préparation cellulaire.

L'intégration de tous ces paramètres est indispensable à la conception d'un test d'arrêt intégré à un système de prédiction. Béran et Srivastava (7) ont appliqué les propriétés asymptotiques du Bootstrap afin d'obtenir des tests sur les valeurs propres, utilisant des cônes

de confiance sur les composantes principales. Ce principe de test sur les valeurs propres, fondé sur l'établissement d'intervalles de confiance autour des valeurs propres et sur le calcul de valeurs critiques, devra être développé et étudié dans un contexte de contrôle en temps réel ; en effet, un système de prédiction exige une précision et une fiabilité maximales du test utilisé.

CONCLUSION

Dans le domaine de la reconnaissance des formes, l'analyse des données offre des outils performants d'évaluation d'homogénéité ou d'hétérogénéité d'une population d'individus. A cet égard, l'un des indices essentiels caractérisant une population est son effectif ; on est donc confronté à la définition d'un critère permettant de limiter les dimensions de la population à la séquence obtenue.

Les résultats obtenus ont permis de mettre en évidence la possibilité de "mesurer" la représentativité d'un échantillon. Le test du Bootstrap, en effet, offre une bonne estimation de la variabilité d'un échantillon en mesurant la stabilité des composantes principales issues d'une analyse en composantes principales. Cette méthode permet également de déterminer une taille d'échantillon représentatif de la population totale. Son utilisation comme critère d'arrêt en reconnaissance des formes est donc envisageable.

Cependant il reste indispensable de confirmer ces premiers résultats par d'autres expériences sur des fichiers ayant plus de paramètres ou possédant des structures différentes. Les travaux récents sur certaines variantes de la méthode Bootstrap (8,9) et l'engouement actuel de la recherche fondamentale en statistique pour cette méthode permettront d'espérer obtenir des estimations encore plus précises et donc des résultats encore plus significatifs.

BIBLIOGRAPHIE

1 - BRUGAL G. "Place de la microscopie quantitative" in "Les marqueurs tumoraux", edt. Bolla (M), Martin (P). Masson. Paris, 111-128, (1989).

2 - WILLIAMS M.A. "Quantitative Methods in Biology": chap.2, Vol.6, Pt II of Practical Methods in Electron Microscopy, ed. Glauert A.M, North-Holland, Amsterdam, (1977).

3 - EFRON B. "The Jacknife, the Bootstrap and other Resampling Plans". S.I.A.M., Philadelphia, Pennsylvania, (1982).

4 - EFRON B. " Computers and the theory of Statistics ; Thinkink the unthinkable."SIAM REVIEW, 21, n°4, 460-480, (1979).

5 - EFRON B. "Bootstrap methods : Another look at the jacknife". The annals of statistics, 7, 1-26, (1979).

6 - EFRON B., DIACONIS P. " Computer intensive methods in statistic". Scientific American, 248, 116-130, (1983).

7 - BERAN R., SRIVASTAVA M.S. " Bootstrap tests and confidence regions for functions of a Covariance matrix". The Annals of Statistics, 13, 95-115, (1985).

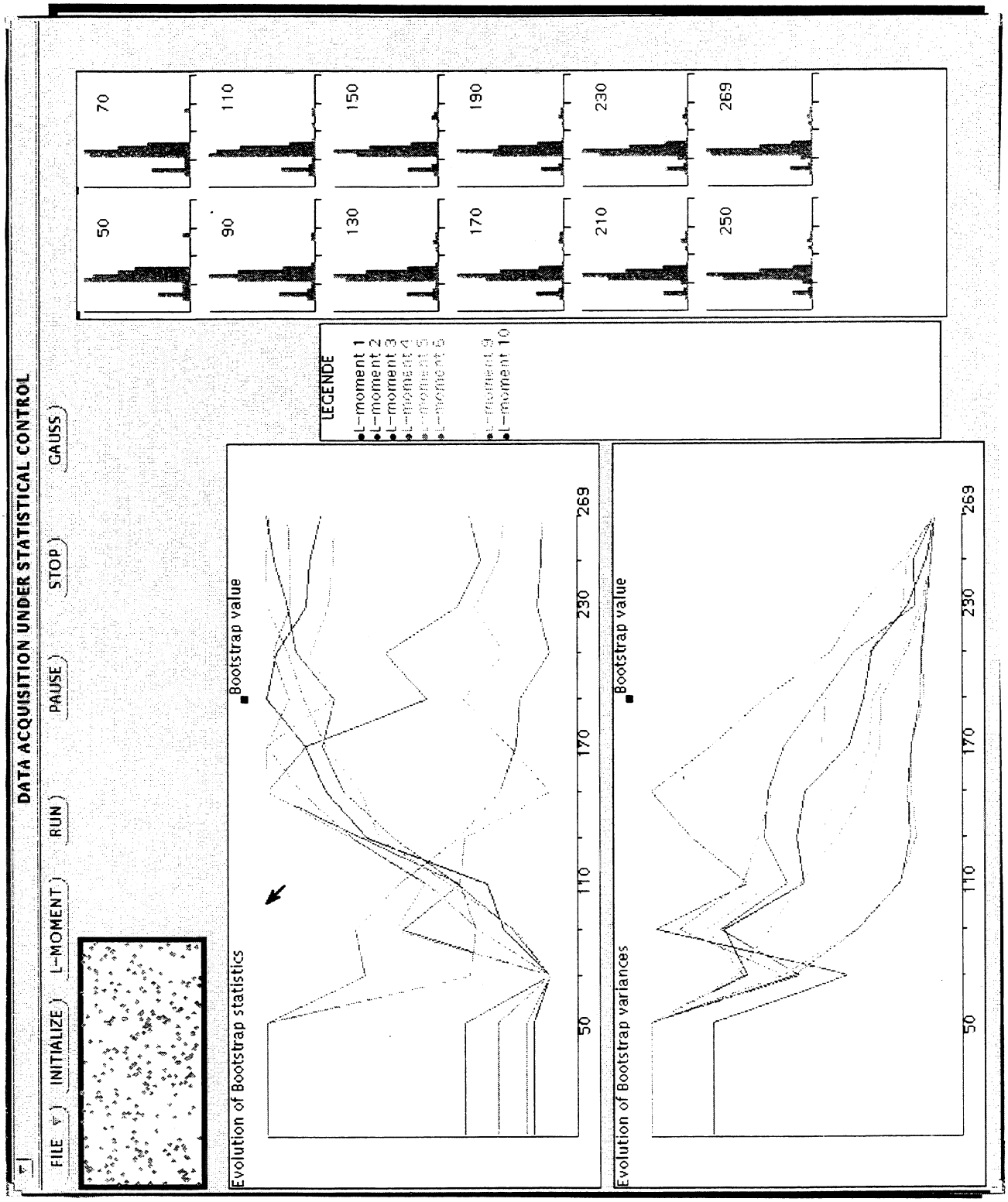
8 - BANKS D.L. "Smoothing the bayesian Bootstrap". Technical Report 415, Departement of statistics, Carnegie Mellon University, Pittsburgh, Pennsylvannia, (1988).

9 - LO A.Y. "A Bayesian Bootstrap for a finite population". The Annals of Statistics, 16, 1684-1695, (1988).

IV. Conclusions

Un des avantages de la méthode du bootstrap, comme toutes les méthodes basées sur le calcul intensif sur ordinateur, est qu'elle n'exige aucune hypothèse statistique sur la distribution de la statistique mesurée.

Une étude théorique plus approfondie est cependant nécessaire; il est en effet indispensable de montrer que les variances estimées par la méthode du bootstrap sur des échantillons de taille variable sont comparables. Nous avons effectué certaines simulations sur des distributions normales montrant que les estimations données par la méthodes du boostrap étaient correctes quelque soit la taille de l'échantillon. Ces résulats empiriques doivent être confirmer de façon plus formelle. L'étude des variantes de la méthode du Bootstrap, en particulier du Bootstrap bayésien, est en cours. Des estimations plus précises et plus robustes des variances des L-moments comme celles des valeurs propres sont espérées.



INTERFACE DU LOGICIEL D'ACQUISITION D'HISTOGRAMMES D'ADN SOUS CONTROLE STATISTIQUE

MODELISATION DE L'EMERGENCE ET DE LA CROISSANCE D'UNE TUMEUR DANS UN TISSU SAIN DIFFERENCIE : UN OUTIL DE SIMULATION ET D'ETUDE DE L'ANALYSE CYTOMETRIQUE DE l'ADN

Travaux

- **Publication 1**

Computer Model for the Emergence of Neoplasia in Growing Cell Populations
Part I : Simulation of Growth Parameters
Soumis à Cancer Research en Février 1993

- **Publication 2**

Computer Model for the Emergence of Neoplasia in Growing Cell Populations
Part II : a tool to simulate and study DNA analysis by image cytometry.
Soumis à Cancer Research en Mars 1993

Nous venons de voir l'importance des problèmes d'échantillonnage dans la fiabilité des diagnostics cytologiques, et en particulier en cytologie quantitative. En vue d'une meilleure compréhension des divergences entre un échantillon et la population dont il est issu, un modèle de simulation de l'évolution tumorale a été développé.

De nombreux modèles de croissance tumorale sont proposés dans la littérature. Ces modèles déterministes, basés sur des équations différentielles, s'intéressent à la croissance tumorale au niveau populationnel.

Brièvement, cette croissance tumorale est modélisée le plus souvent par [Hecquet 1985] :

-*une évolution exponentielle*. L'évolution cancéreuse reste occulte jusqu'à ce que la tumeur atteigne une taille détectable cliniquement. La durée de cette phase occulte est très variable selon les cancers. Dans le cancer du sein cette phase durerait environ 9 ans alors que les métastases se développeraient 1 à 3 ans après l'émergence de la tumeur primaire.

-*une évolution gombertzienne*. L'évolution de la tumeur, correspondant au temps de doublement, diminue progressivement.

-*une évolution stochastique*. La croissance de la tumeur serait irrégulière et caractérisée par des phases de croissance exponentielle entrecoupées de périodes de quiescence.

Ces trois types d'évolutions sont représentés sur la Figure 1.

Malgré la prise en compte de facteurs de plus en plus nombreux et divers, comme la concentration en hormones de croissance, la nourriture disponible dans le milieu, les échanges inter-cellulaires ou la vascularisation du tissu, les modèles dits déterministes présentent deux inconvénients majeurs :

- Ils s'intéressent à la population cellulaire, c'est à dire au nombre de cellules ou à la masse de la tumeur. La tumeur est modélisée globalement sans connaissance ni modélisation des entités qui la composent.
- Ils s'intéressent à l'évolution tumorale mais pas à la naissance ni à l'émergence de la tumeur à partir d'un tissu sain différencié.

Ces modèles déterministes sont donc inadaptés à notre problème. Une approche radicalement différente s'avère nécessaire. Elle repose sur trois concepts fondamentaux :

- *la modélisation cellulaire*. L'unité de base du modèle est la cellule, avec ses propriétés biologiques et des comportements spécifiques et déterminés. Le comportement de la population n'est pas déterminé mais est la résultante des comportements de chacune des cellules qui la composent.
- *La modélisation tissulaire*. Nous nous intéressons à la modélisation d'un tissu et non pas d'un groupe de cellules. La population de départ est constituée de cellules différenciées et de cellules non-différenciées, dont les relations et les échanges sont responsables de la cohésion et du fonctionnement du tissu.

- L'émergence tumorale.

La modélisation de l'homéostasie tissulaire permet d'étudier l'émergence de la tumeur au sein d'un tissu différencié normal, en simulant, stochastiquement, différents types de mutations.

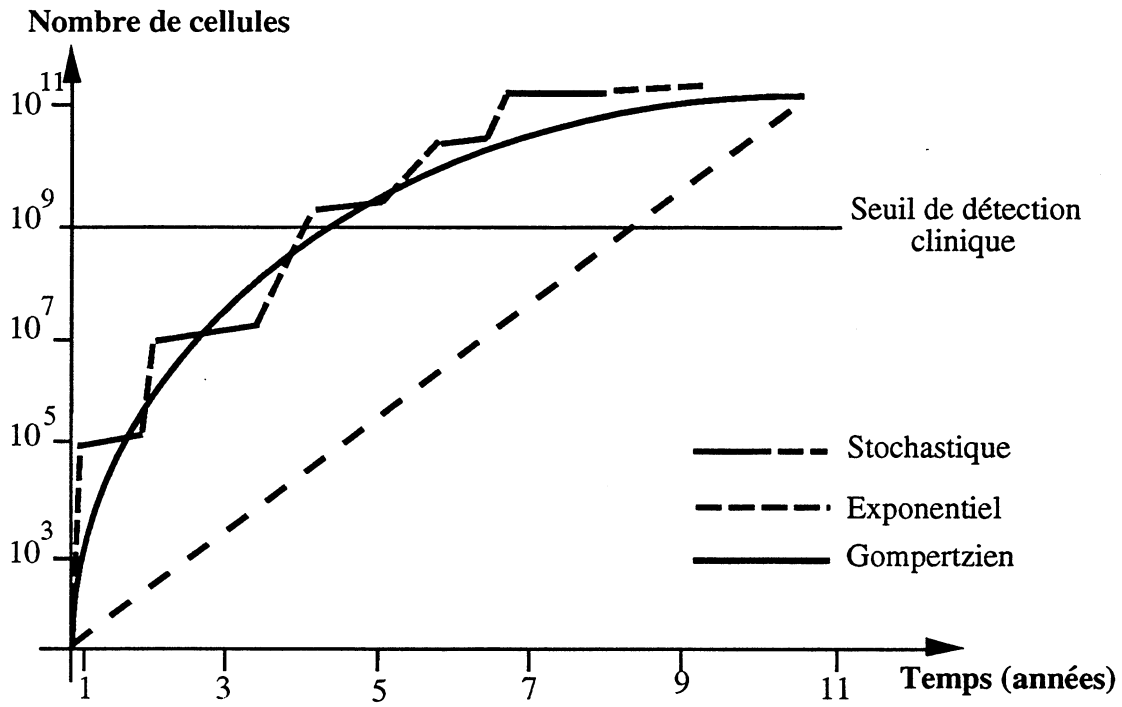


Figure 1 : Comparaison des trois modèles de croissance tumorale les plus courants [Hecquet 1985].

Les principes biologiques du modèle sont expliqués en détail dans la première publication. Ils reposent sur des données biologiques récentes concernant les mécanismes de différenciation et de prolifération cellulaire. La cancérisation et l'émergence de clones tumoraux dans un tissu sain sont modélisées au niveau cellulaire par le jeu de mutations affectant certains gènes responsables du contrôle de la prolifération cellulaire.

La deuxième publication illustre l'intérêt de ce modèle pour l'étude des problèmes d'échantillonnage en particulier pour l'analyse de l'ADN. La connaissance de la genèse, de la progression de la tumeur au sein d'un tissu différencié, des propriétés biologiques de chacune des cellules de toute la population permet d'étudier l'évolution biologique (ploïdie, ADN) de la tumeur. L'analyse de l'ADN d'échantillons tumoraux est possible à tous moments de la croissance de la tumeur. Les caractéristiques biologiques de la tumeur sont accessibles avant que la taille cliniquement détectable soit atteinte. La structure de ce modèle constitue une plate-forme idéale d'étude et de compréhension des problèmes liés à l'interprétation des histogrammes d'ADN, décrits précédemment, et des éventuelles erreurs de diagnostics et de pronostics qui en résultent.

I. Modélisation de l'émergence et de la croissance d'une tumeur dans un tissu sain différencié

Cette publication présente le principe, l'implémentation d'un modèle de simulations de l'émergence et de croissance d'une tumeur solide à partir d'un tissu différencié stable. Ce modèle peut être considéré comme un modèle de troisième génération basé sur les principaux mécanismes biologiques identifiés comme responsables de l'initiation et de la promotion tumorale. Il s'agit d'un modèle plus explicatif (il se rapproche en cela du modèle proposé par Shackney [Shackney 1989] que déterministe dans le sens où il s'appuie sur une approche stochastique plutôt que sur des équations différentielles [Calderon 1991] [Pollack 1991]. La compétition entre les facteurs de croissance et les facteurs de différentiation est au centre du processus d'occurrences des anomalies génétiques pouvant entraîner un déséquilibre entre la différentiation et la prolifération cellulaire.
Ce modèle simule les événements néoplasiques intervenant dans une population cellulaire stable et différenciée. Il permet d'étudier de nombreuses caractéristiques, biologiques et génétiques, de la progression tumorale qui peuvent ensuite être comparées avec les observations cliniques.

Ces travaux ont fait l'objet d'une publication soumise à Cancer Research en Février 1993.

Computer Model for the Emergence of Neoplasia in Growing Cell Populations. Part I : Simulation of Growth Parameters

Martial Guillaud [1] and Gérard Brugal

Laboratoire TIMC. ERFMQ. CERMO. Université Joseph Fourier. BP 53 X. 38 041 Grenoble Cedex.

Running title : Computer model for the emergence of neoplasia

Keywords : Modelization. Neoplasia. Differentiation factors. Growth factors. cell cycle.

[1] To whom requests for reprints should be addressed at ERFMQ, CERMO, Université Joseph Fourier, BP 53X, 38041 Grenoble Cedex.

ABSTRACT

This paper presents the design, implementation and testing of a new computer model simulating the occurrence of a solid tumour within a steady state differentiated normal cell population. This model initiates a "third generation" of models based on the major biological mechanisms identified as responsible for tumour initiation and promotion. It is explicative rather than deterministic and thus based on a stochastic approach rather than differential equations. It considers the competition between differentiation and growth factors as responsible for the likelihood of the occurrence of genetic abnormalities which may shift the differentiation to proliferation equilibrium. In this respect, this model can simulate the neoplastic events occurring in a steady state differentiated normal cell population. It is thus uniquely suited to investigate a variety of biological and clinical progression features which can be further compared with clinical observations.

Acknowledgments

We wish to thank DR. V. von Hagen for manuscript preparation and Dr. M. Brugal for valuable work in documentation research.

Introduction

During the last decade, numerous authors have proposed a variety of computer models to simulate proliferation and growth of tumours. Depending on model, considered different aspects of cancer progression were taken into account, such as tumour mass, food supply, immune system, vascularization, cell to cell diffusion (1,2), autocrine and paracrine growth factors production (3). These models described the evolution in the size of a tumour that originates either from progeny of a transformed cell or from a multicellular plant (4) but did not consider the normal tissular environment in which the tumour arises. The majority of these models, here referred to as "first generation models", was deterministic and involved differential equations to mimic the empirical data observed on clinical cases (2, 5, 6).

Recently, Pollack (7) and Sainsbury (8) proposed a model based on the hypothesis that overexpression of genes encoding normal growth factor receptors can contribute to the abnormal proliferation of neoplastic cells. This model was based on recent clinical data suggesting that some human neoplasms exhibit unusually high levels of cell-surface receptors for epidermal growth factors, and that this abnormality is associated with rapid cell proliferation and poor prognosis (9,10). These recent models, here referred to as "second generation models", were more explicative than the first generation models for the biological mechanisms involved in neoplasia, but they are still deterministic and based on differential calculations.

Both the first and second generation models were derived from the key concept of clonal evolution of neoplasms which is based on biochemical, cytogenetic, molecular genetic and immunological evidences indicating that most tumours arise from a single altered cell, the progeny of which expands as a neoplastic clone (11,12,13). Unfortunately, these models did not take into account the steady state differentiated normal cell population where such a single altered neoplastic cell appears, and may appear repeatedly, as the consequence of genetic instability. These models also did not consider the proliferation *versus* differentiation antagonism which is another key concept in tumour progression. As a matter of fact, there is a tendency in neoplasms to increase their growth rate with time and to escape from local growth control mechanisms. Usually, this does not result from a shortening of the cell cycle time, but rather from an increase of the "growth fraction". This means that cells belonging to the neoplastic population continue to proliferate actively instead of progressing toward terminal differentiation and subsequent cell death (4). Moreover, as tumours become more malignant, they show morphological and metabolic alterations that are globally interpreted as loss of differentiation. It is still not clear to which extent the tumoural phenotype results from either an actual "block" of the differentiation process or from an increased proportion of cells remaining in an "undifferentiated" or "poorly differentiated" state compatible with proliferation.

This paper presents the design, implementation and testing of a new computer model simulating the occurrence of a solid tumour within a steady state differentiated normal cell population. This model initiates a "third generation" of models based on the major biological mechanisms identified as responsible for tumour initiation and promotion. It is explicative rather than

deterministic and thus based on a stochastic approach rather than differential equations. Although Shackney (14) already presented a model which used a stochastic approach to simulate genetic instability and tumoural cell cycle progression, the model presented here is basically different. It considers the competition between differentiation and growth factors as responsible for the likelihood of the occurrence of genetic abnormalities which may shift the differentiation back to proliferation equilibrium. In this respect, this model can simulate the neoplastic events occurring in a steady state differentiated normal cell population. It is thus uniquely suited to investigate a variety of biological and clinical progression features which can be further compared with clinical observations.

Biological rationale of the model (Figure 1)

The model proposed here is based on four main biological concepts arising from the literature.

Firstly, it is now well established that growth factors (GF) are not only involved in the cell renewal of adult tissues but also in carcinogenesis since some alterations of the mitogenic signal transduction mechanisms can result in, or contribute to, tumour promotion (15). Moreover, Sporn (16,17), among others, has shown that some transformed cells can produce, in an autocrine way, those GF required for cycling and for which they bear the corresponding receptors (GFR). Alternatively, non autocrine malignant cells may have an abnormally high level of mitogenic receptors, particularly for epidermal growth factor, which makes them more reactive to systemic growth promotors (18,19). Local GF level as well as GFR cellular concentration are parameters to handled by the model since they account for the probability of GF to GFR binding at the individual cell level and the subsequent start of the cell cycle.

Secondly, it is important to point out that sister cells arising from mitosis do not have identical phenotypes and consequently do not have the same differentiation *versus* proliferation capabilities. Recently, Brooks and Riddle (20) studying the variability of cell cycle duration in 3T3 cell, showed that sister cells having considerable differences in capacity for further proliferation appeared at high frequency in culture. Many years ago, Holtzer (21) demonstrated that in cell cultures of normal and cancerous cells two different types of cell cycles do exist : "quantal cell cycle" in which cells are sensitive to growth factors (GF), and differentiation factors (DF) and a "proliferation cell cycle" which only increases the number of cells. Analysing, in MCF7 breast cancer cell line, the intra-cellular distribution of Ki-67 antigen, a cell cycle dependent protein, we also showed the differential behaviour of cells with respect to mitotic cycle. A small proportion of G_1 new-born cells was keeping a high level of Ki-67 antigen and passed directly to S phase in local optimum conditions of growth while other cells were successively loosing and then synthesising the Ki-67 before moving to S phase. The last were susceptible of regulation by extracellular factors to either continue or quit the cycle (22). Cells in the quiescent state marked by low Ki-67 level (and which is not yet a deep G_0 state) are sensitive to growth and differentiation factors (23). Differential competence for proliferation

versus differentiation is a parameter to handled by the model since it accounts for the individual cell likelihood of passing into successive mitotic cycles.

Thirdly, renewable adult tissues are comprised of stem cells, cells committed to differentiation, and fully differentiated cells whose life span is a cell lineage characteristic (24). Under normal conditions, the mitotic cycle of any stem cell yields another stem cell for self-replacement and a cell committed to differentiate in order to replace any differentiated cell that has either died or left the tissue. In 1975, Bullough (25), among others, demonstrated that in renewable adult tissue one new cell is produced for each old cell that is lost, which implies that the average outcome of each mitosis is one daughter cell able to cycle again and one non-cycling but ageing daughter cell. The precision of this homeostatic mechanism together with the observation of tissue mass recovery after cell lost (e.g. healing), argue in favour of a tissue specific negative feedback from differentiated cells to stem cells. Several diffusible mitotic inhibitors produced by differentiated cells, have been described in skin, bone marrow, liver and intestine (26). Although the existence of tissue specific mitotic inhibitors is still questionable, many inhibitory effects have been described even for some growth stimulating substances such as the mesenchymal TGFß which is a growth inhibitor for epithelial and other cell types (27). In addition to their anti-mitotic role, the chalones or chalone-like substances obviously trigger differentiation both *in vivo* and *in vitro* (28). More recently, retinoic acid has been characterised as a mitotic inhibitor inducing or repressing growth factors and growth factors receptors (29,30). The capacity of cells to be prevented from cycling by mitotic inhibitors, and to consequently differentiate, is a parameter to handled by the model since it accounts for the individual cell likelihood of definitely quitting the stem cell compartment.

Fourthly, it is now generally accepted that initiation of malignancy results from genetic disorders at the individual cell level triggering either over-expression of oncogenes (coding for the proteins involved in the cascade of mitotic stimulations from activation of growth factors receptors to DNA replication) and/or lack of expression of anti-oncogenes (coding for all proteins involved in mitotic inhibition and thus promoting differentiation). The mechanisms usually involved are point mutation, amplification, translocation, inversion and deletion which occur at particularly weak points in the chromatids. The probability of such disorders increases as a cell lineage repeatedly replicates its DNA and divides through successive generations. Many of these disorders finally result in numerical and/or structural chromosomal abnormalities appearing during the promotion of malignancy (31). Aneuploïdy has actually been reported in many solid tumours and is usually related to increased growth fraction in neoplasia (32). The cell's risk of developing numerical and/or structural abnormalities as a consequence of its mitotic history is a parameter to handle by the model since it accounts for the initiation of cancer and the extent of the growth advantage provided by genetic abnormalities eventually incurred during tumour promotion.

Modelling of normal cell population (Figure 1)

The computer model presented here uses the Monte-Carlo simulation techniques to stochastically represent genetic mutation probabilities and any factor to its corresponding receptor binding probabilities as well. The model was written in C++ and implemented on a SUN SPARK2 workstation.

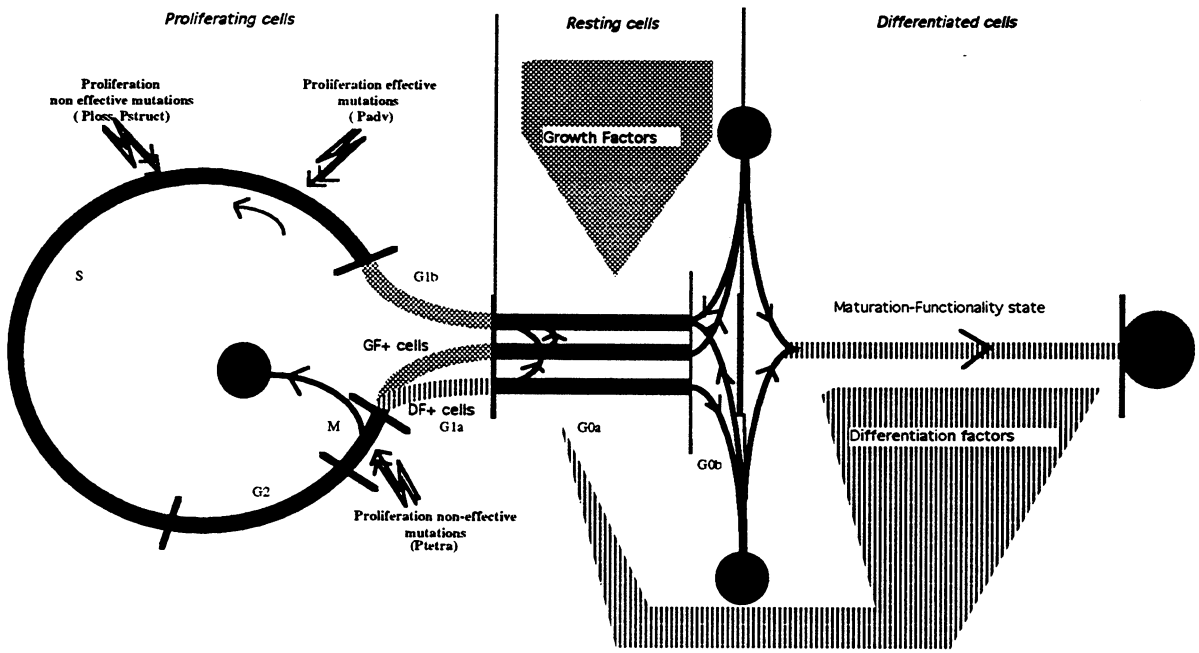


Figure 1 : Biological rationale for cell proliferation model (see text). **D**ifferentiation and **G**rowth **F**actors are in competition for the maturation of undifferentiated cells in G0 compartment. Black circles represent death compartments.

Cell status with respect to cycle

The initial cell population handle by the model is composed of three compartments ;

- the proliferating cells (anywhere in the cycle),
- the differentiated cells (either maturating or functional)
- the resting cells (or in G0 state).

The time spent in any of these compartment is expressed in hours and the initial proportions of each can be specified. To simulate cell growth, each cell progresses sequentially through the five steps : G1a->G1b->S->G2->M or leaves the cycle between the G1a and G1b steps toward G0 state. These step durations are kept the same for all cells at any time but, depending how long a cell is resting in G0, the generation time separating two successive mitoses may vary markedly.

Growth and differentiation factors and receptors

The model involves two factors available to resting cells : the growth factor (GF) and the differentiation factor (DF) which bind with - and compete for - the same cell receptors. The number of GF available in the medium at any time is kept constant and equal to the number of receptors to occupy for committing one single cell to growth. DF on the other hand are synthesised by differentiated cells themselves. During its life, any differentiated cells synthesises and releases as much DF as is required to commit one G0 resting cell to differentiation.

Two factor to receptor affinity coefficients x and y can be adjusted so that the two daughter cells arising from a mitosis do not have the same receptors to GF *versus* DF affinity. The receptors of one cell are x-fold more affine to GF than to DF (GF+ cells), while the receptors of its sister are y-fold more affine to DF than to GF (DF+ cells). It should be noted that the factor to receptor binding efficiency and subsequent G0 transition thus depends on (i) the concentration of the factors, (ii) the number of receptors per cell, and (iii) their differential affinity for GF *versus* DF. Since both the number of receptors and the GF and DF concentrations can change as a result of genetic disorders ; the factor to receptor binding efficiency may vary from cell to cell.

Resting behaviour

The daughter cells arising from a mitosis go through the G1a step during which they synthesise a number of receptors which become equal to that of their mother and then move to G0 compartment. The G0 cells are found in two different states : G0a and G0b. The G0a cells increase their receptor number until the initial number is doubled. Should the GF and or DF concentration be below the effective threshold, the G0a cells enter into the G0b deep resting state. Should this situation obtain for the allowed cell life span, the G0b cells then die. Alternatively, should the concentration of the DF or GF become effective before death, the G0b cell will move to growth or differentiation accordingly. A resting cell thus has three possible behaviours :

- staying in G0a and then G0b state until death,
- leaving the G0a or G0b state to go back to the cell cycle after a time which depends on the GF to receptor binding status,
- leaving the G0a or G0b state to become differentiated and functional until death.

The time spent in resting state before leaving to proliferate or differentiate depends on the GF or DF to receptors binding status.

Cell committment toward either growth or differentiation

A resting cell is "GF committed" when the proportion g of its initial receptors occupied by GF is sufficient to step backward to cycle thus entering the G1b phase after a time equal to the time already spent in G0a (i.e. time to return to the cycle = time spent in G0a). Alternatively, a resting cell is "DF committed" when the proportion d of its initial receptors occupied by DF is

sufficient to go forward to maturation after the time necessary to complete G0a. The proportion thresholds g and d are fixed during the initialisation of the model. If the committment occurs during the G0b state, then the time to go back to G1b is equal to the time spent in G0a (in case of growth commitment) while the maturation starts immediately (in case of differentiation committment).

Generation time

The cell generation time is the sum of 4 elementary time durations :

a) G1a,

b) the time spent in resting state to become GF "committed" and which varies from 0 hr to G0a + G0b duration at the maximum,

c) the time spent in resting state to go back to G1b and which varies from 0 hr to G0a duration at the maximum,

d) the time to complete the cycle (G1b + S + G2 + M).

After each time interval of 1 hr, the modelized cell population is updated with respect to :

a) the age distribution,

b) the proportion and number of cells in the various states or cell cycle phases,

c) the respective proportion of DF and GF in the medium as well as the number of free *versus* occupied receptors per cell. If the total number of occupied receptors (all cells committed) is lower than that of available factors then the unbound factors in excess are preserved for the next one hour time interval (cumulative effect).

Proliferation effective and non-effective mutations (Figure 1)

The model makes possible the occurrence of two types of genetic abnormalities during the cell cycle.

The first type of abnormality, which occurs during S phase with a probability *Padv*, is said to be proliferation effective since it is a mutation of oncogenes which actually modifies the proliferating capability of cells.

- A cell undergoing a first proliferation effective mutation doubles its receptors number while it still requires the same number of factors as a normal cell to exit G0 phase either backward to cycle or forward to differentiate. Therefore, such a cell is likely to become committed faster than a normal cell.

- A cell undergoing a second proliferation effective mutation, synthesises and releases, during its cycle, as much GF as needed to leave G0 phase. This cell thus becomes GF autocrine.

Since these mutations are hereditary, progeny of mutant cells keeps the same characteristics.

The second type of abnormality, which occurs during S or M phase, is said to be proliferation non-effective since it modifies the genome of the cell but not the cell proliferating capabilities. These abnormalities, such as chromosome loss, structural reorganisation and tetraploidization can occur with different probabilities. Cells with chromosome numbers exceeding 200 or

inferior to 40 are considered non viable and are thus discarded from the population as are the cells with structural abnormalities number exceeding 20.

Results

The results presented here have been obtained with the following parameters set up by default unless otherwise specified :

- Initial cell population = 1000 cells,
- Compartments respective size : Proliferating cells = 5%, Resting cells percent = 5%, Differentiated cells = 90%;
- Phase durations : G1a = 2 hr, G1b = 2 hr, S = 8 hr, G2 = 2 hr, M = 2 hr, G0a = 10 hr, G0b = 1000 hr, Differentiated and resting cells life span = 1000 hr;
- Receptor status : Number per cell = 100, Proportion thresholds : g = d = 50%, Affinity coefficient : x = y = 100,
- Proliferation effective abnormality likelihood (Padv) = 0.
- Proliferation non effective abnormality likelihoods = 0.

Simulation of normal cell population

A first series of experiments aimed at testing how a steady state cell population can be established by the model. For this purpose, the population was set up as above. The evolution of the population was followed during 30 000 hours (about 3.5 years). The results are presented in Figure 2.

Figure 2a shows that during a first period of adjustment, the total cell number markedly increased and decreased. This adjustment wave lasted 1 000 hr and obviously resulted from the progression of new-born cells toward resting state and the transition of resting toward differentiated state while the number of proliferating cells decreased accordingly. The model thus self-regulates the proportion of cells in the three compartments as a consequence of the respective life span fixed by the initialisation. As expected, fixing the differentiated cells lifetime to 2 000 hr (instead of 1 000 hr), made the adjustment wave to last 2000 hr.

After the adjustment period, the proportions of cells in the three compartments (proliferating, resting, differentiated) remained stable as did the total cell number. Namely, the cell proportions stabilised at 1% proliferating cells, 14 % resting cells and 85 % differentiated cells. It should be noted that if the initial parameters were too far from this equilibrium (for example : 40% proliferating, 10% resting, and 50% differentiated cells), then the model was unable to achieve the adjustment and the simulated cell population escaped from steady state to become undifferentiated and exponentially growing.

Figure 2b shows the variations of the relative proportion of GF and DF available during the experiment. Given the default set-up, the proportions of GF is slightly superior to DF most of

the time . Moreover, some abrupt variations occur from time to time with very limited amplitude (± 15%) thus showing how the regulation operates.

Consequences of tissue damage

A second series of experiments to test how the model would recover from a significant lost of differentiated cells. For this purpose, a cell population was set up as above. After 1 000 hr the steady state was achieved as in the previous experiment. Then, 10%, 20%, 30% or 45 % of the differentiated cells were depleted at 2 000 hr and the cell population was followed during 10 000 hr. The results are shown in Figure 3.

A depletion of 10% (Fig. 3a), 20% (Fig. 3b) and 30% (Fig. 3c) of the differentiated cells was regulated by the model. The time to return to the steady state compartment sizes was proportional to the extent of the depletion ; namely, 250 hr after a 10% depletion, 1200 after a 20% depletion and 5000 hr after a 30% differentiated cell loss.

When 45% of differentiated cells was depleted (Fig. 3d), the model proved to be unable to recover. As a result of the damage, the proportion of DF produced by the remaining differentiated cells became too low to trigger the differentiation of the increasing proportion of resting cells issuing from the cycle. Consequently, a majority of DF+ resting cells spent their life span in the resting compartment and the differentiated compartment drastically decreased in size. After 5 000 hr, the population total cell number followed a damped sinusoidal curve and then converged to a size which depends on the resting phase duration. This new cell population is undifferentiated.

Emergence and Growth of a tumour

Continuous stochastic mutations

A third series of experiments were carried out to test how the model would generate tumour cells. For this purpose, a cell population was set up as above except that the proliferation effective mutation likelihood (Padv) was set different from 0 after completion of the 1000 hr adjustment wave. Each cell passing through the S phase was thus susceptible to be mutated at any of the successive cell cycle with a Padv probability. The cell population was followed during 20 000 hr. The results are shown in Figures 4, 5 and 6.

As shown in Fig. 4a, a mutation likelihood Padv = 0.001 resulted in a slight increase of the population cell number at 18 000 hr. The evolution of the respective proportions of single and double mutated cells illustrated in Fig. 4b shows that the second mutation appeared a long time (13 000 hr, i.e. about 1.5 years) after the first mutation. The increased proportion of double mutated cell resulted in significant release of autocrine GF after 15 000 hr as shown in Fig. 4c.

As shown in Fig. 5a, a mutation likelihood Padv = 0.005 resulted in a slight increase of the population cell number at 12 000 hours. The evolution of the respective proportions of single and double mutated cells illustrated in Fig. 5b shows that the second mutation appeared a short time (500 hr) after the first mutation, i.e. a shortening by a multiple of 26 results from a multiple of 5 increase of Padv likelihood when compared to the previous experiment. At 20 000

hr the proportion of single and double mutated cells reached 70% as compared to 20% in the previous experiment. The increased proportion of double mutated cells resulted in a significant release of autocrine GF after 2 000 hr as shown in Fig. 5c.

As shown in Fig. 6a, a mutation likelihood Padv = 0.01 resulted in a drastic increase of the population cell number at 10 000 hr. In contrast with the previous experiments, the respective proportions of the three cell compartments became unbalanced with fast predominance of the undifferentiated cells. The evolution of the respective proportions of single and double mutated cells illustrated in Fig. 6b shows that the second mutation appeared a very short time (100 hr) after the first mutation. By 12 000 hr all cells have either a single or a double mutation and the DF concentration collapsed while the autocrine GF reached a maximum as shown in Fig. 6c.

Individual cell's growth advantage

The evolution of the cell population was simulated where either a single or a double mutation occurred in only one cell at about 1000 hours. This did not induce any detectable change of the population cell number even for a simulation lasting 50 000 hr (i.e. 6 years)).

The proliferation rate of a unique mutated cell has been tracked and is reported in Figure 7. In case of a single mutation which doubles the receptors number (Fig. 7a), the size of the mutated progeny became significantly higher from 1 200 hr on than the size of a normal cell progeny tracked in parallel. The advantage reached 20% at 18 000 hr. In case of a double mutation which doubles the receptor number and makes the cell autocrine for GF (Fig. 7b), the growth advantage of the mutated progeny was similar to that observed after a single mutation. It is thus obvious that the autocrine capability, when acquired by an individual cell, does not provide additional growth advantage since the GF in excess are shared within the whole population unless they were not diffusible.

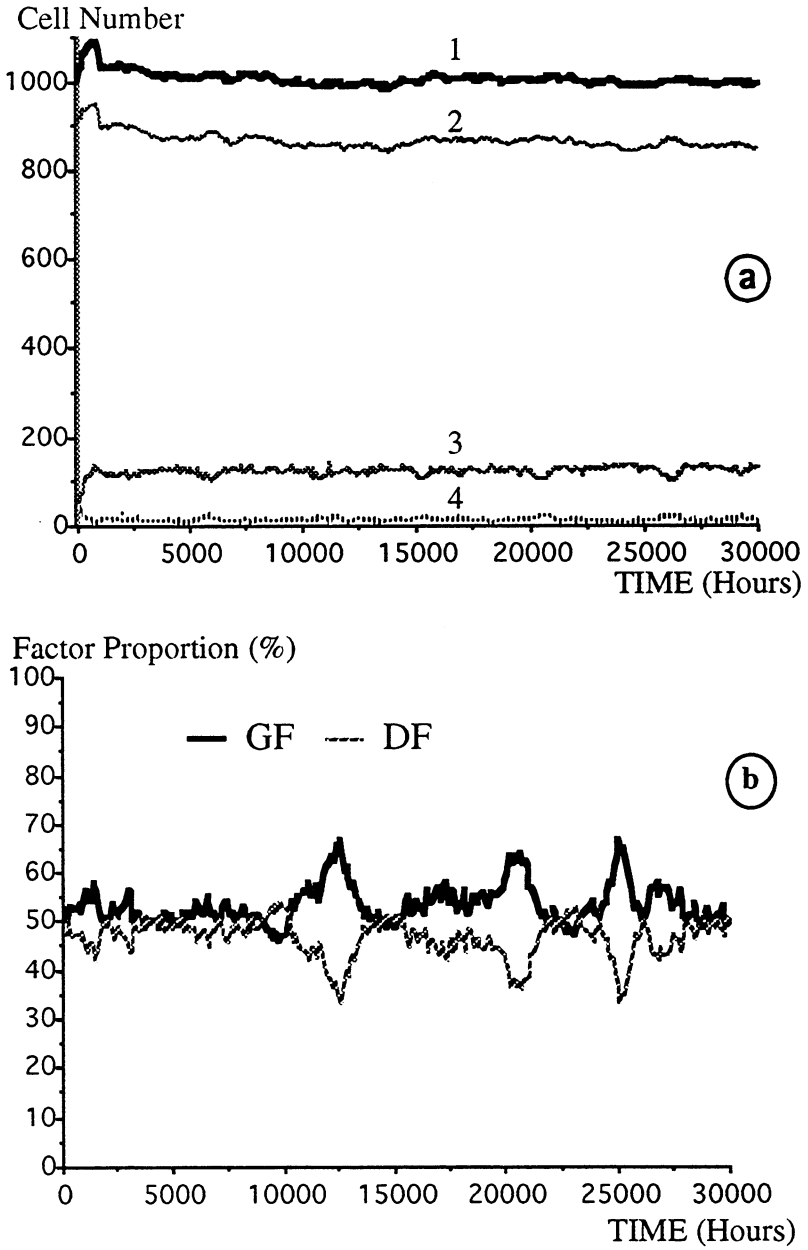


Figure 2 : Simulation of a normal differentiated cell population.
a : Evolution of the total cell number (1), of the differentiated cell number (2), of the resting cell number (3) and of the proliferating cell number (4) as time increases.
b : Evolution of proportions of Growth Factors and Differentiation Factors in environment as time increases.

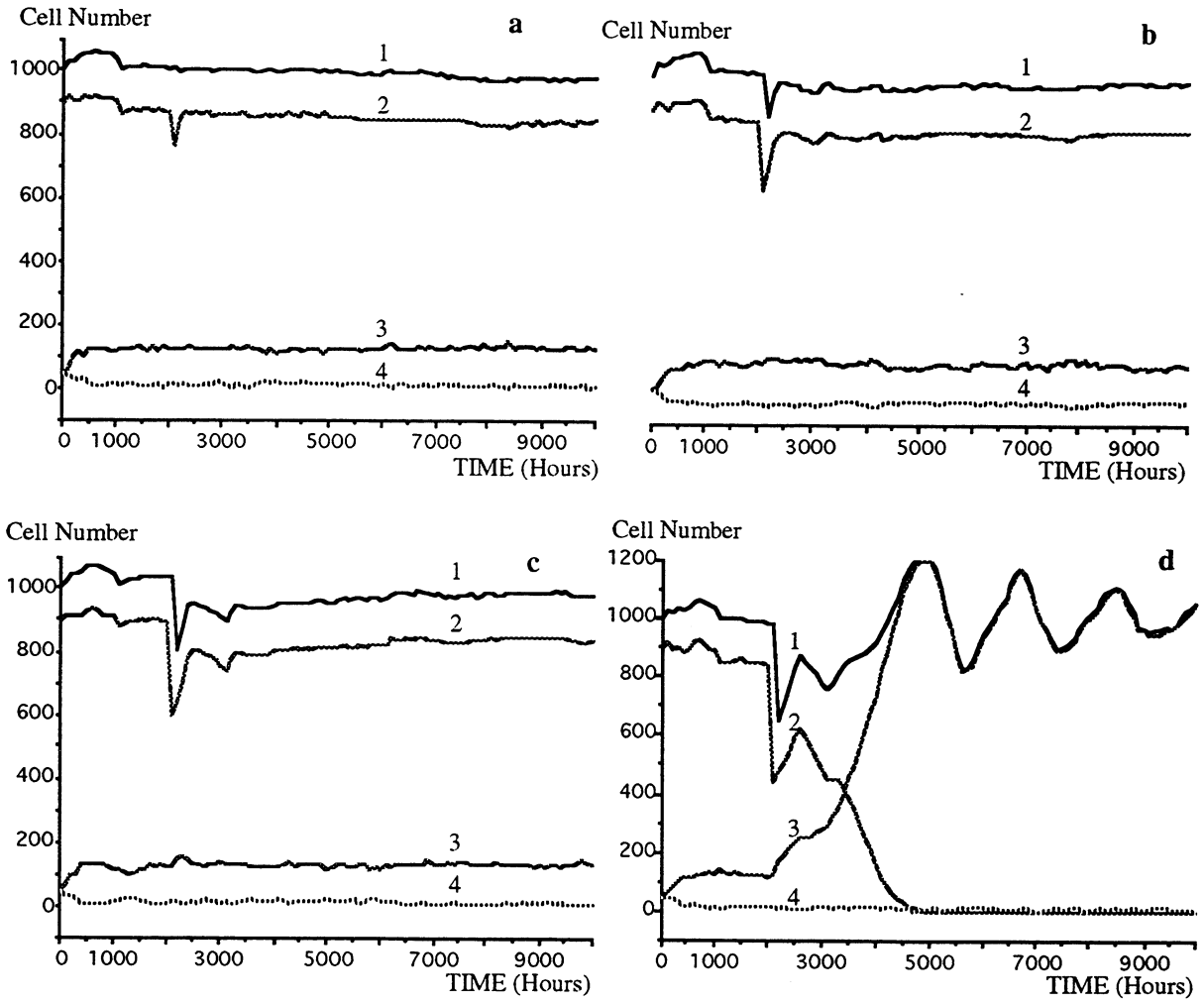


Figure 3 : Consequences of tissue damage. Evolution of the total cell number (1), of the differentiated cell number (2), of the resting cell number (3) and of proliferating cell number (4) as time increases.
a : 10% of the differentiated cells are cleared at 2000 Hr,
b : 20% of the differentiated cells are cleared at 2000 Hr,
c : 30% of the differentiated cells are cleared at 2000 Hr,
d : 40% of the differentiated cells are cleared at 2000 Hr.

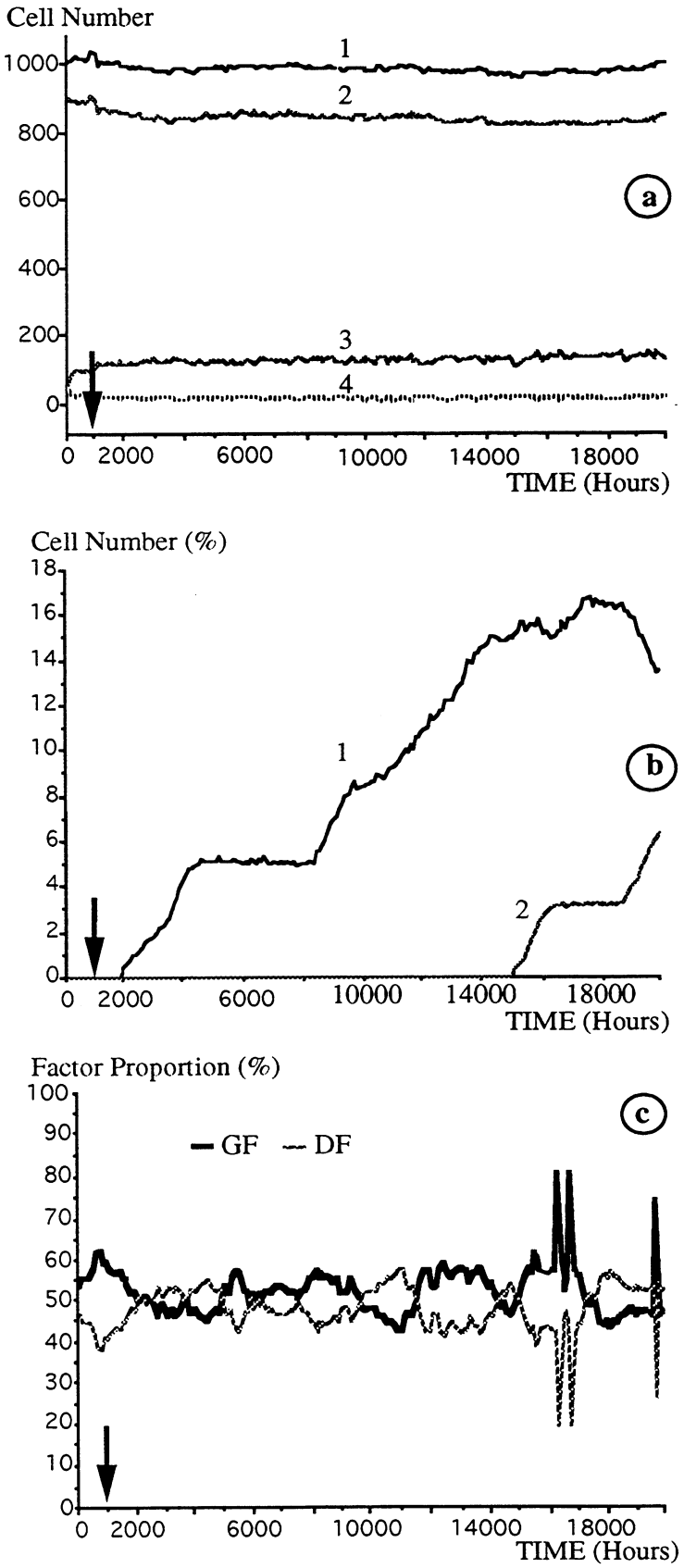


Figure 4 : Emergence and Growth of tumour. Arrows indicates the beginning of occuring mutations : Mutation probability, Padv = 0.001. **a :** Evolution of the total cell number (1), of the differentiated cels number (2), of the resting cell number (3) and of proliferating cell number (4) as time increases, **b :** Evolution of single (1) and double mutated (2) cell number as time increases, **c :** Evolution of Factor proportions as time increases.

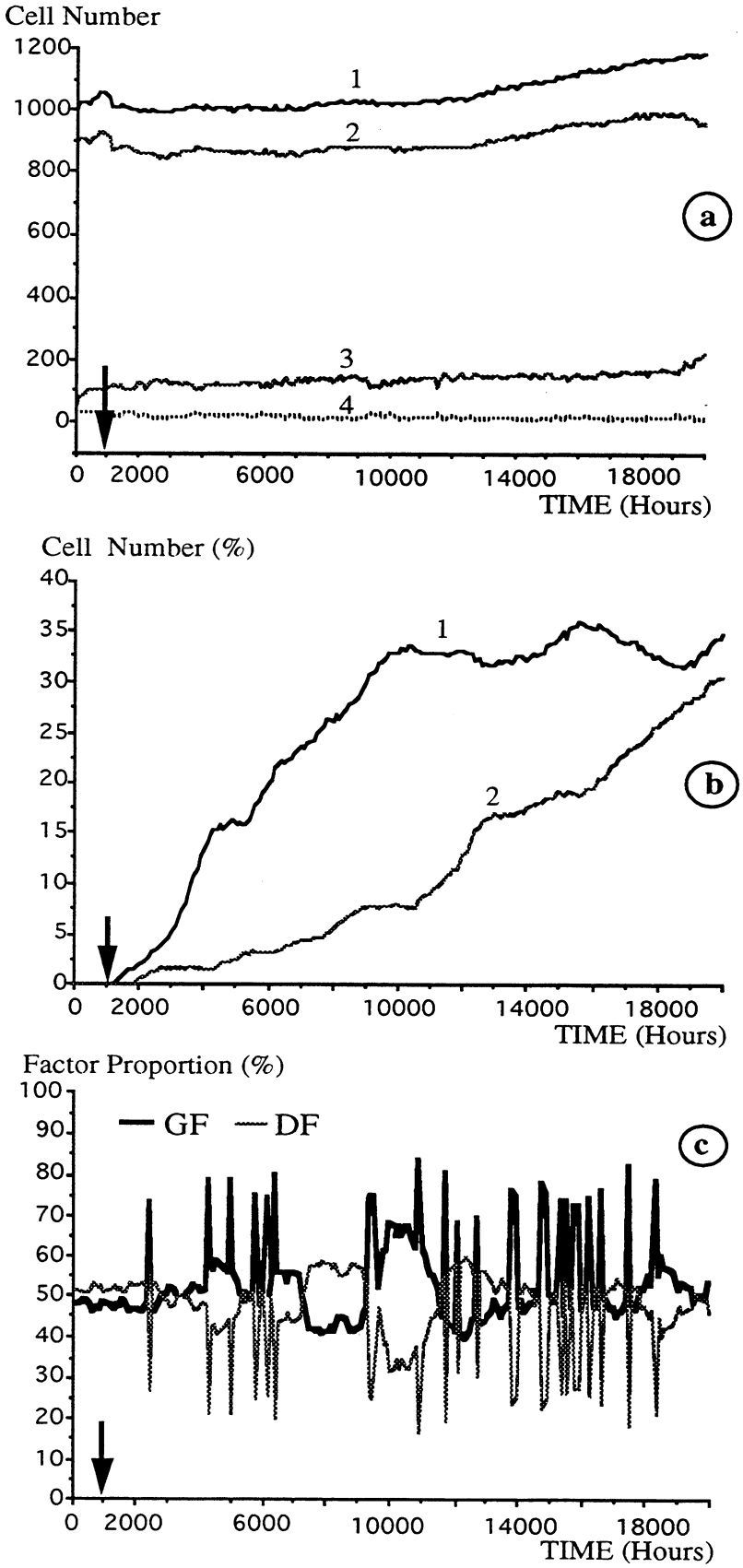


Figure 5 : Emergence and Growth of tumour. Arrows indicates the beginning of occurring mutations : Mutation probability, Padv = 0.005. **a :** Evolution of the total cell number (1), of the differentiated cell number (2), of the resting cell number (3) and of proliferating cell number (4) as time increases, **b :** Evolution of single (1) and double mutated (2) cell number as time increases, **c :** Evolution of Factor proportions as time increases.

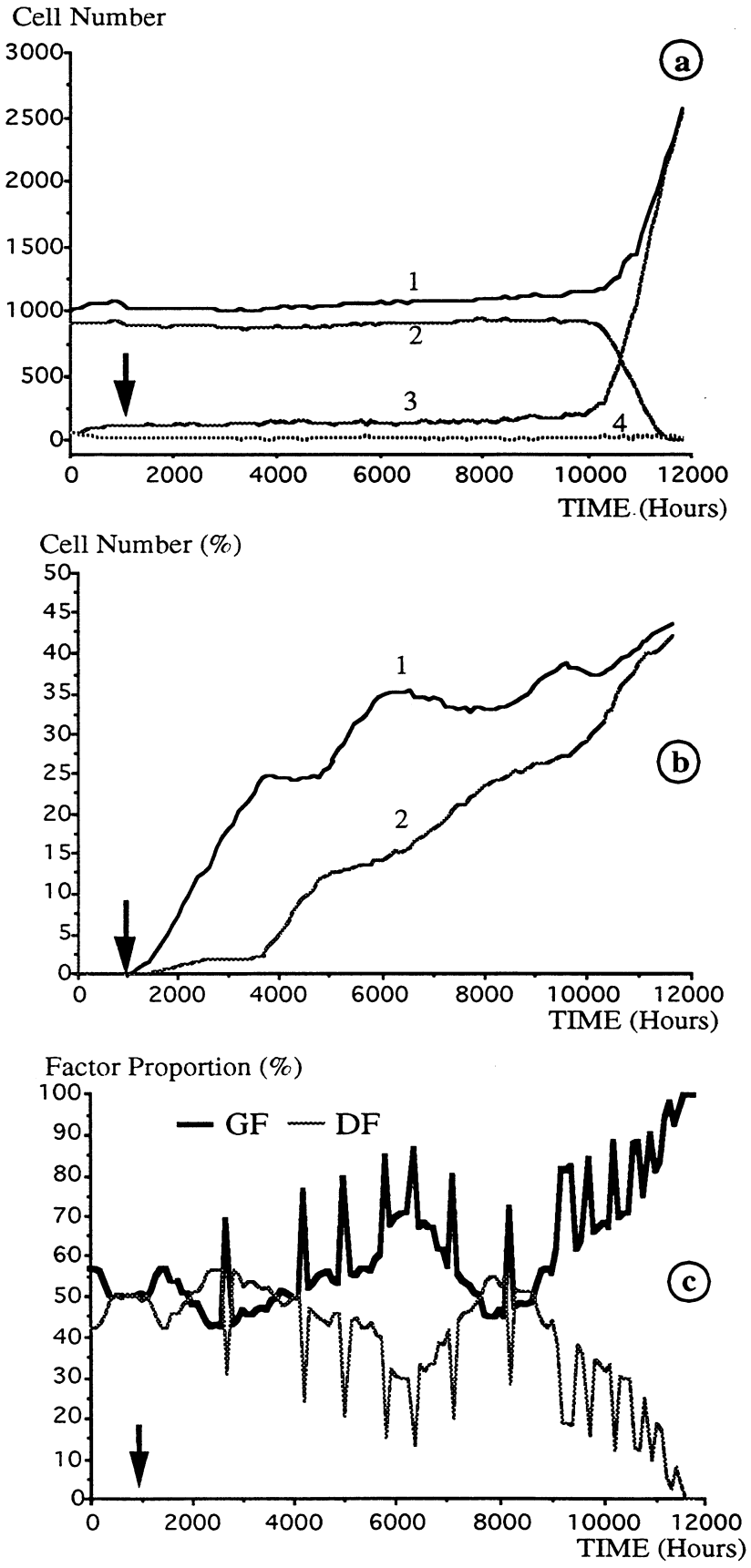


Figure 6 : Emergence and Growth of tumour. Arrows indicates the beginning of occurring mutations : Mutation probability, Padv = 0.01. **a :** Evolution of the total cell number (1), of the differentiated cell number (2), of the resting cell number (3) and of proliferating cell number (4) as time increases, **b :** Evolution of single (1) and double mutated (2) cell number as time increases, **c** : Evolution of Factor proportions as time increases.

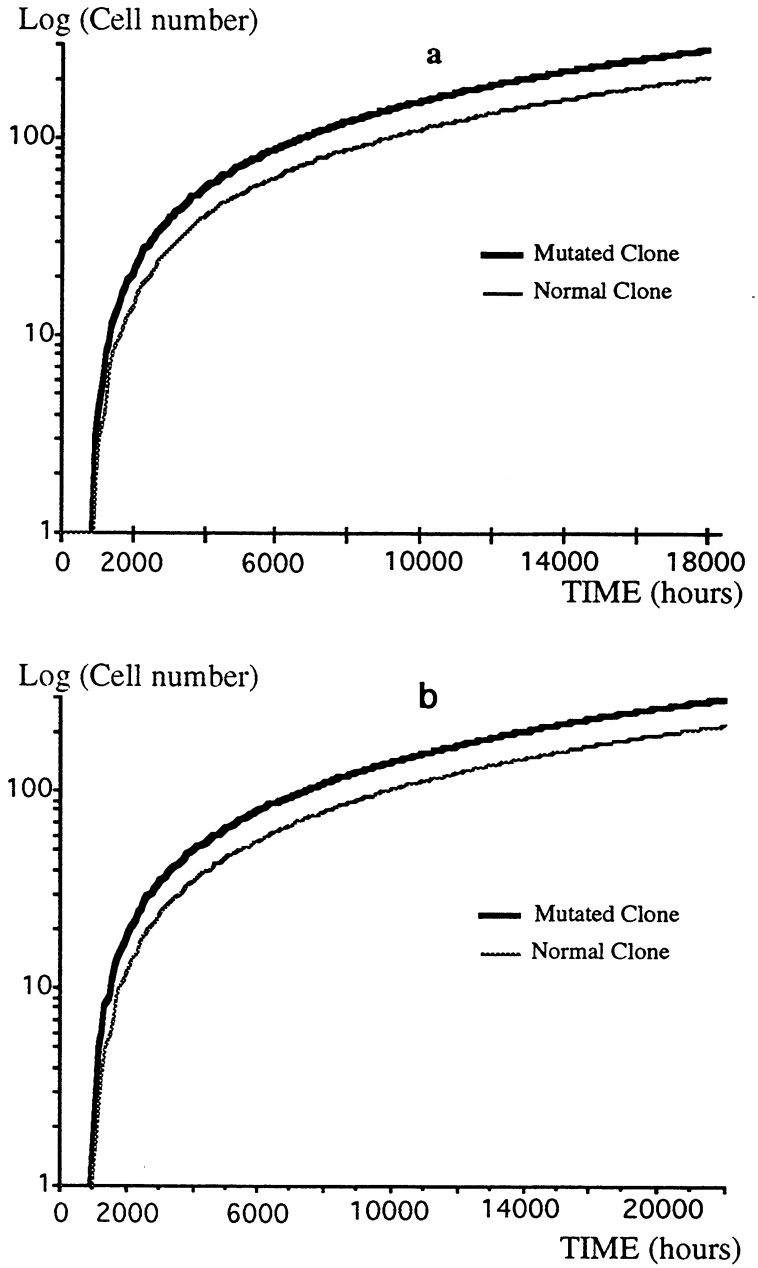


Figure 7 : Individual cell's growth advantage. Growth difference between a normal cell and a mutated cell.
a : in case of a single mutation (receptor number doubling)
b : in case of a double mutation (receptor number doubling and releasing of Growth Factors).
The single or double mutation occurs in only one cell at about 1000 hours.

Discussion

For the purpose of analysing neoplasm occurrence and growth as a consequence of proliferating effective mutations, we have developed a stochastic model that assumes that the homeostasis of a differentiated tissue depends on the equilibrium between two categories of factors : growth factors (GF) and differentiation factors (DF).

GF Vs DF competition - A number of authors have already proposed a variety of models taking into account GF concentration in the development of neoplasms (7,3) but none considered the competition between GF and DF. While GF involvement in the control of cell proliferation has been extensively studied recently, the action of DF remains poorly documented. Consequently, the precise mechanisms involved in the competition between GF and DF are still hypothetical but, whatever they may be, it is now obvious that different extracellular messages trigger the cell to either proliferate (as the result of oncogene expression) or differentiate (as the result of anti-oncogene expression). The model presented here assumes that if the proportion of occupied receptors reaches a 50% threshold, cell division or differentiation occurs. This threshold has been chosen arbitrarily but is consistent with the observation of Rozengart (33) that EGF triggers maximal mitotic activity when only a fraction of the EGF-receptors is occupied.

In a steady-state differentiated cell population, a dead cell has to be replaced by a new born cell, a condition which is statistically fulfilled by the model presented here which takes advantage of a simple concept : the number of GF in the medium has to be equal, on an average, to the number of DF released by differentiated cells. The simulated populations were very tolerant with respect to the initial conditions and the emergence of the steady-state finally involved a very low number of cycling cells as in real renewable tissues.

Growth advantage - In the model presented here, two proliferating effective mutations are required for neoplastic emergence. The first mutation triggers the doubling of the receptor number which provides a growth advantage. Such mutations resulting in an abnormally high level of functionnal cell surface EGF-receptors have actually been observed in solid tumours (18). The second mutation permits the cell to produce and release the GF it requires for cycling. The autocrine production of growth factor is also well known in malignant tumours (34) and has been modelized by Michelson (3). The growth advantage provided by the first mutation was clearly demonstrated by our model and very similar to the results already obtained by Pollack (7). Interestingly, our results show that the second mutation did not increase the growth, the reason is that our model is more suitable for autocrine than for paracrine mechanisms. "Paracrine" refers to a control mechanism in which the local release of growth factors affects cells of a type different from the secreting ones, while "autocrine" refers to secreting cells having the corresponding surface receptor so that a positive feed-forward control loop may be established. So any growth advantage conferred by the growth factor upon a target cell would also act upon the producer cell itself (16,17,35).

Structural abnormalities - Proliferating non effective mutations, also defined by Shackney (14), have been implemented in this model but their specific effect has not been investigated here because they make sense in neoplasm growth only if they induce proliferative effective mutations. Although some tumours are obviously diploïd (from both flow cytometry and cytogenetics), the majority of malignant lesions show aneuploïd and/or polyploïd cells. Any of these structural abnormalities could generate either amplification and activation of oncogenes or deletion and mutation of anti-oncogenes. It thus seems more consistent to consider both proliferating and non proliferating mutations in the same series of stochastic events (14). Once cell has acquired a growth advantage from a first mutation, it thus becomes more susceptible to further mutations activating growth promoting genes which probability is increasing with cell cycle number as already demonstrated (36,37). Sequential genetic abnormalities in tumours confer incremental proliferative capability upon successive clonal overgrowth (12).

Local environmental conditions - The model presented here does not take into consideration a variety of factors which determine cell proliferation, such as tumour mass (38), various mitogenic factors *Vs* growth inhibitors, nutritional factors, cell density, vascularization, cell to cell relationships, diffusion of growth and differentiation factors, cell topography with respect to receptor equipment and paracrine *Vs* autocrine capabilities in a tumour. All these factors, already modelised in a few first and second generation models (3,7), will be progressively integrated into our model as biological evidence becomes available.

In the present form of the model, growth factors released in the medium by mutated cells becomes available for each cell of the population. This is at variance with recent biological findings that local concentration of various factors may play an important role in the development of certain cell clone in tumours thus showing the limitation of a model based only on a stochastic approach. Including the spatial relationships and clonal heterogeneity in the model will soon become feasible because of the recent development of topographical simulation of cell populations using the Voronoi diagrams as proposed by Marcelpoil et al. (39). The combination between our stochastic approach to follow, to modify, to keep track of the mitotic history of individual cells on the one hand, and the spatial behaviour of cell progeny given the topographical organisation of the cell population on the other hand, opens the way for a fourth generation of models simulating the cell sociology.

References

1 Vaidya P.G., Vayda V.G. and Martin D.G. An application of the non-linear bifurcation theory to tumor growth application. Int J.Biomed Comput, *27* : 27-46, 1991.

2 Volkov E.I., Stolyarov M.N and Brooks R.F. The modelling of heterogeneity in proliferative capacity during clonal growth. J. Non Linear Biol. (in press)

3 Michelson S. and Leith J. Autocrine and paracrine growth factors in tumor growth : a mathematical model. Bulletin of mathematical biology, *53,* 639-656, 1991.

4. Atkinson E.N., Bartoszynski R., Brown B.W. and Thomson J.R. On estimating the growth function of tumors. Math. Biosci. *67*: 145-166, 1983.

5 Calderon Calixto P. Modeling tumor growth. Mathematical Biosciences, *103*, 97-114, 1991.

6 Demicheli R., Pratesa G. and Foroni R. The exponential-gompertzian tunor growth model : Data from six tumr cell lines in vitro and in vivo estimate of the transition point from exponential to gompertzian growth and potential clinical implications. Tunori, *77* : 189-195, 1991.

7 Pollak M., Boyarsky A. and Gora P. A mathematical model describing consequences of abnormally high levels of epidermial receptor on the proliferation of neoplastic cells Cancer investigation, *9*: 513-520, 1991,

8 Sainsbury J.R.C, Ramdon J.R., Needham GR et al. Epidermal -growth-factor receptor status as predictor of early recurrence od and death from breast cancer. Lancet, *1*: 1398-1402, 1987.

9 Filmus J., Pollack MN., Cailleau R., Builk RN. MDA - a human breast cancer cell line with a high number of epidermal growth factor (EGF) receptor, has an amplified EGF recepteur gene and is growth inhibited by EGF. Biochem Biophys Res Comm,*128* : 898-905, 1985.

10 Filmus J., Pollack MN., Caincross JG., Builk RN. Amplified, overexpressed and rearranged epidermal growth factor receptor gene in a human astrocyma cell line. Biochem Biophys Res Comm,*131*: 207-215, 1985.

11 Nowell, P. The clonal evolution of tumor cell population. Science (Wash. DC), *194*: 197-200, 1975.

12 Nowell P.C. Mechanisms of tumor Progression. CANCER RESEARCH, *46*: 2203-2207,1986.

13 Buick R.N. and Pollack M.N. Perpectives on clonogenic tumor cells, stem cells, and oncogenes. CANCER RESEARCH, *44* :4909-4918, 1984.

14 Shackney S.E., Smith C.A, Miller B.W., Burholt D.R., Murtha K., Giles H.R., Ketterer D.M. and Pollice A.A. Model for the genetic evolution of human solid tumors. CANCER RESEARCH, *49* : 3344-3354, 1989.

15 Goustin A.S., Leof E.B., Shipley G.D., and Moses H.L. Growth factors and cancer CANCER RESEARCH *46* : 1015-1029, 1986.

16 Sporn M.B. and Roberts A.B. Autocrine growth factors and cancer. Nature, *313*, 1985.

17 Sporn M.B. and Todaro G.J. Autocrine secretion and malignat transformation of cells New Engl. J. Med, *303* : 878-880, 1980.

18 Dong X, Berthois Y. and P.M Martin. Effect of epidermal growth factor on the proliferation of human epithelial cancer cell lines: correlation with the level of occupied EGF receptor. Anticancer Research *11*: 737-744, 1991.

19 Colomb E., Berthon P., Dussert C., Calvo F. and P.M. Martin. Estradiol and EGF requirements for cell-cycle progression of normal human mammary epithelial cells in culture. Int. J. Cancer, *49* : 932-937, 1991.

20 Brooks R.F. and P.N. Riddle. Diffreences in growth factor sensitivity between individual 3T3 cells arise at high frequency : possible relevance to cell senescence Exp.Cell.Res., *174* : 378-387, 1988.

21 Holtzer H., Biehl J., Antin P.,Tokunaka S., Sasse J., Pacifici M. and Holtzer S. Quantal and proliferative cell cycles : how lineages generate cell diversity and maintain fidelity. In: Calobin gene Expression and hematopoietic Differenciation, Stamatoyannopoulos G, Nienhuis AW(eds). Alan R. Liss, Inc., New York, 213-227.1983.

22 du Manoir S., Guillaud P., Camus E., Seigneurin D. and Brugal G. Ki-67 labeling in postmitotic cells defines different Ki-67 pathways within the 2c compartment Cytometry *12*:455-463, 1991.

23 Ferrari S, Calabretta B., Battini R., Stephen C., Cosenza SC., Owen TA, Soprano KJ. and Baserga R. Expression of c-myc and induction of DNA synthesis by Platelet-poor plasma. Exp Cell Res *174*: 25-33,1988.

24 Hall PA, Watt FM. Stem cells : the generation and maintenance of cellular diversity. Development, 106 : 619-633, 1989.

25 Bullough W.S. Mitotic control in adult mammalian tissues. Biol.Rev;, *50* : 99-127, 1975.

26 Laurence E.B. The significance of chalones in epidermal growth.*In* : R.I.C. Spearman and Riley P.A. (eds). The Skin of vertebrates. Linnean Society Symposium Series N° 9, pp. 139-150, 1980.

27 Sonnenschein N.C. and Soto A.M., Symposium on the control of cell proliferation and cancer. Cancer Res.*49* : 6161,1989

28 Brugal G. Presence of Intestinal halones. *In* : STEM CELLS of Renewing Cell populations. Academic Press New York , Inc.1976.

29 Zheng Z., Polakowska R., Johnson A. and Golsmith L.A. Transcriptional control of epidermal growth factor receptor by retinoic acid. Cell growth and differentiation, *3*: : 225-232,1992.

30 Strickland S., Smith K. and Marotti K. Hormonal inductionn of differentiation in teratocarcinoma stem cells : generation of parietal endoderm by retinoic acid and dibutyryl cAMP. Cell. *21* : 347, 1980.

31 Dutrillaux B., Gerbault-Seureau M., Remvikos Y., Zafrani B. and Prieur M. Breast Cancer genetic evolution: I. Data from cytogenetics and DNA content. Breast Cancer Research and Treatment,*19* : 245-255, 1991

32 Friedlander M.L., Hedley D.W. and Taylor I.W. Clinical and biological significance of aneuploidy in human tumours. J. Clin. Pathol ; *37* : 961-974, 1984

33 Rozengart E. and Collins M. Molecular aspects of growth factor action : Receptors and intracellular signals. J. Pathol., *141* : 309-310, 1983.

34 Lang R.A. and Burgess A.W. Autocrine growth fcators and tumourigenic transformation. Immunology Today, *11* : 244-249, 1990.

35 Loef E.B., Proper J.A.., Goustin A.S., Shipley G.D., Dicorleto P.E. and Moses H.L. Induction of c-sis mRNA and activity simular to platelet-derived growth factor by transforming growth factor beta : a proposed model for indirect mitogenesis involving autocrine activity. Proc.Natn.Acad. Sci. USA, *83* : 2453-2457, 1986.

36 Rowley J.D. Human oncogenes locations and chromosomes aberrations. Nature (Lond.), *301*: 290-291,1983.

37 Yunis J.J. Chromosomal rearrangements, genes and fragile sites in cancer (clinical and biologic implications. In: V.T. Devita, S.Hellman, and S.A. Rosenberg (eds), Important Advances in Oncology 1986 , pp 93-128. Philadelphia: J.B. Lippincott Company, 1986

38 Prehn R.T. The inhibition of Tumour growth by tumor mass. CANCER RESEARCH *51* : 2-4, 1991.

39 Marcelpoil R. and Usson Y. Methods for the study of cellular sociology: Voronoï diagrams and parametrization of the spatial relationships. J. theor. Biol., *154* : 359-369, 1992.

II. Modélisation de l'analyse de l'ADN et de la ploïdie d'un échantillon tumoral par cytométrie à balayage.

La progression tumorale est définie comme un phénomène d'acquisition progressive des capacités métastatiques et invasives accompagnées d'altérations de l'expression de certains gènes [Fould 1957] [Nowell 1986] . Les anomalies chromosomiques prononcées [Mitelman 1976] et les changements quantitatifs d'ADN [Sugihara 1990] représentent les bases génétiques de la progression tumorale. L'identification des anomalies génétiques, grâce aux récents développement en particulier de l'hybridation *in situ*, est possible sur des cellules en mitoses. Cependant, certains problèmes méthodologiques, comme l'accessibilité aux chromosomes, en limitent l'utilisation en routine clinique [Harden 1979]. L'analyse du contenu en ADN, technique plus rapide, a l'avantage d'être applicable sur des cellules en interphase et sur de nombreuses cellules, permettant ainsi une analyse statistique du statut "ploïdique" de la tumeur. En revanche, cette technique n'offre aucune information sur la structure individuelle des chromosome.

A partir du modèle de simulation de l'émergence et de la croissance d'une tumeur, décrit précédemment, l'évolution du contenu en ADN de la tumeur et des anomalies génétiques au cours de la progression tumorale peut être étudiée. La structure du modèle permet, en outre, d'appréhender les différents problèmes statistiques liés à la méthodologie de l'analyse d'ADN par cytométrie en image : prélèvement, échantillonnage, taille de l'échantillon, représentation des histogrammes d'ADN, variabilité instrumentale ...

Ces travaux ont fait l'objet d'une publication soumise à Cancer Research en Avril 1993.

Computer Model for the Emergence of Neoplasia in Growing Cell Populations. Part II : a tool to simulate and study DNA analysis by image analysis

Martial Guillaud [1] and Gérard Brugal

Laboratoire TIMC. ERFMQ. CERMO. Université Joseph Fourier. BP 53 X. 38 041 Grenoble Cedex.

Running title : Computer model for the emergence of neoplasia

Keywords : Modelization. Neoplasia. Statistics. DNA cytometry. Sampling.

[1] To whom requests for reprints should be addressed at ERFMQ, CERMO, Université Joseph Fourier, BP 53X, 38041 Grenoble Cedex.

ABSTRACT

In a related paper (Part I), a simulation model of neoplastic growth was proposed. This new computer model simulates the occurrence of a solid tumour within a steady state differentiated normal cell population. One of the main interests of this model, as compared to other deterministic models based on differential equations, is its ability to recognise and to follow any cell of the simulated population. One of our motivations in designing a computer model for the emergence of neoplasm in growing cell population was to offer the pathologists with a tool to investigate the possible major drawbacks of DNA histograms interpretation. The numerous DNA histograms simulated here illustrate the possible variability and misinterpretation as far as cancer diagnosis and prognosis are concerned.

For this purpose, the model differentiates between the ploidy and DNA histograms of the whole reference cell population and a number of FNAB-like samples for which the bias introduced by sampling, sample size, class number and coefficient of variation are illustrated. Moreover, some of the statistical features commonly used to support diagnosis and prognosis were calculated and objectively showed the frequent unreliability of the overall approach to caryological disorders based on image cytrometry of cell DNA content.

Acknowledgments

We wish to thank DR. V. von Hagen for manuscript preparation and Dr. M. Brugal for valuable work in documentation research.

Introduction

In a related paper (1), a simulation model of neoplastic growth was proposed. This new computer model simulates the occurrence of a solid tumour within a steady state differentiated normal cell population. This model initiates a "third generation" of models based on the major biological mechanisms identified as responsible for tumour initiation and promotion. It is explicative rather than deterministic and thus based on a stochastic approach rather than differential equations. This model is basically different than the one of Shackney (2) since it considers the competition between differentiation and growth factors as responsible for the likelihood of the occurrence of genetic abnormalities which may change differentiation to proliferation equilibrium. In this respect, this model can simulate the neoplastic events occurring in a steady state differentiated normal cell population. It is thus uniquely suited to investigate a variety of biological and clinical progression features which can be further compared with clinical observations. One of the main interests of this model, as compared to other deterministic models based on differential equations, is its ability to recognise and to follow any cell of the simulated population. This model was initially developed to study and analyse statistical aspects of DNA histogram interpretation in the framework of image cytometry.

Technological advances have made DNA ploidy assessment a clinically and economically feasible procedure, 30 years after its medical potential was established (3, 4). The distribution of tumour cells according to their total DNA content - often referred to as ploidy pattern, ploidy profile, DNA distribution pattern or DNA histogram - is actually a set of proportions indicating occurrence frequency of nuclei having a DNA content within predefined ploidy classes. The availability of automated, rapid recording, image cytometry has led to greatly increased interest in the diagnostic and prognostic indices provided by the ploidy pattern. Whatever the method used to feature the DNA histogram, there is an overall consensus that the majority of human malignant tumours exhibit aneuploidy.

Nethertheless, evaluation of DNA ploidy patterns is typical of the practical problems encountered in applying statistical methods to cytological data. Many authors classify DNA patterns by only visual inspection. The first and most famous classification was given by Auer (5) who classified DNA histograms into four types. Although visual inspection methods may be adequate for a large number of situations, for others they can be criticised as being too subjective or inadequate for borderline DNA histogram profiles. Since DNA histograms do not take into consideration any other cell feature, the ploïdy measurement does not bear the full force of quantitative techniques (6). For 15 years, DNA histograms have been studied extensively. Numerous authors have proposed several methods and parameters to quantitatively interpret DNA histograms. But the reduction to a single number, such as a "malignancy grade" or an " aneuploidy index " may be too restricted. Numerous studies have shown that parameters such as entropy (7) or malignancy grade (8) were correlated either with histological grade of the disease or with other clinical indices such as node involvement, tumour size and survival time

analysed by the reference Cox regression model. Nethertheless, such correlations are of little assistance to the pathologist who has to give a prognosis for the one patent under consideration. As is the case for any data simplification scheme, there is a trade-off between data reduction and overall loss of information that is, to limit the loss of diagnostically significant information. Whatever the parameters or methods used, either for diagnostic or prognostic purposes, DNA histogram interpretation is likely to be the most relevant among the non-clinical information. Nevertheless, all trade-offs have not been solved, particularly with respect to the considerable variability of ploidy patterns in any given clinical group of patients. Many sources contribute to this variability ; true differences between diagnostic categories, differences among patients in the same category, differences in ploidy patterns recorded for the same patient from different sites of the tumour or different cell samples size (9). Moreover, differences in the staining procedures, specimen preparation techniques and measuring system variability, introduce even more variability (6).

The purpose of this study is to simulate and analyse the significance of DNA histograms taking advantage of the tumour cell population model previously developed (1) and which stochastically generates mutations responsible for malignant transformation. A series of experiments exemplified how this model could generate tumour cells. For DNA pattern analysis purposes, the model has been extended to simulate some mutations of tumorous cells which induce ploidy abnormalities.

Modelisation and simulation principles

Description of computer model

The reference computer model (1) involved the Monte-Carlo simulation techniques to stochastically represent the genetic mutations probabilities and the binding likelihood of any growth versus differentiation factors to their corresponding receptors. The model was written in C++ and implemented on a SUN SPARK2 workstation.

As described in part I, the initial population handled by the model is composed of three compartment (Fig.1);

- proliferating cells (anywhere in the cycle),
- differentiated cells (either maturating or functional)
- resting cells (or in G0 state).

The time spent in any of these compartment is expressed in hours and the initial respective proportions can be set-up. To simulate cell growth, each cell progresses sequentially through the five steps : G1a->G1b->S->G2->M or leaves the cycle between the G1a and G1b steps toward G0 state. These steps respective durations are kept the same for all cells at any time but, depending how long a cell is resting in G0, the generation time separating two successive mitoses may vary drastically.

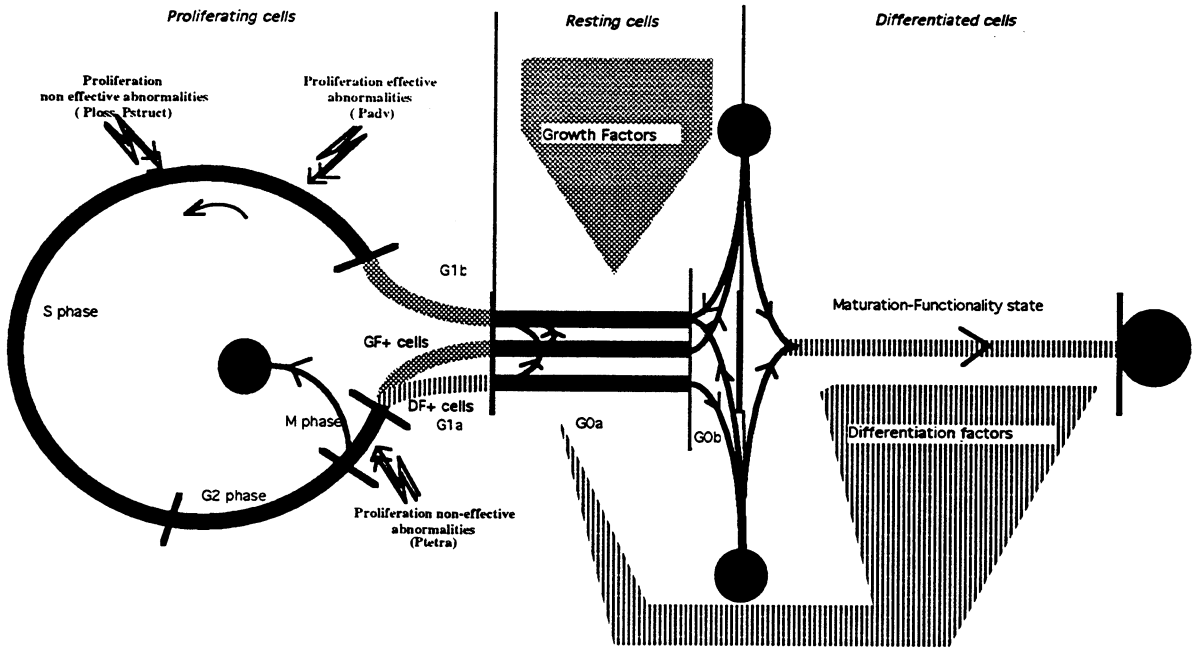


Figure 1 : Biological rationale for cell proliferation model. Growth and Differentiation factors are in competition to drive the G0 cell toward cycle or differentiation respectively. Black circles represent compartment exit by cell death.

Proliferating effective and non-effective mutations (Fig.1)

The model makes possible the occurrence of two types of genetic abnormalities during the cell cycle.

The first type of abnormality, which occurs during S phase with a probability *Padv*, is said to be proliferation effective since it is a mutation of oncogenes which actually modifies the proliferating capability of the mutated cells.

- A cell undergoing a first proliferating effective mutation doubles its receptors number while it still requires the same number of factors as any normal cell to exit G0 state either backward to cycle or forward to differentiation. Therefore, such a cell is likely to become committed faster than a normal cell.

- A cell undergoing a second proliferation effective mutation, synthesises and releases, during its cycle, as much GF as needed by a single cell to leave the G0 state. This cell thus becomes GF autocrine.

Since these mutations are hereditary, the progeny of a mutant cell keeps the same characteristics.

The second type of abnormality, which occurs during S or M phase, is said to be proliferating non-effective since it modifies the genome of the cell but not the cell proliferating capabilities. These abnormalities, such as chromosome loss, structural reorganisation and tetraploidization can occur with the different probabilities *Ploss*, *Pstruct* and *Ptetra* respectively. Cells with chromosome numbers exceeding or inferior to fixable thresholds (200 and 40 are currently used) are considered non viable by the model and thus discarded as are the cells with structural abnormalities number exceeding an arbitrary threshold (20 is currently used).

DNA Histograms and sampling simulation

During the simulated growth of a tumour, the user can temporarily halt or definitely the growth of the process in order to study the features of the current cell population as well as those of a sample of cells thus simulating a Fine Needle Aspiration Biopsy (FNAB). Indeed, every cell, even dead cells, are recorded in an array. Each cell is an instance of class CELL, with the following characteristics :

class CELL

{
- cell number,
- mother cell number ,
- sister cell number ,
- daughter cell numbers ,
- number of generation,
- time spent in the all states,
- number of receptors,
- number of occupied receptors ,
- number of chromosomes,
- number of structural abnormalities,
- age of the cell,
- cell cycle phase of the cell

};

The user clicks on the tumour as illustrated in Figure 2. The user can draw as many samples (i.e. pick up as many biopsies) he wishes. For each new sample, a dialogue box (Figure 3) appears allowing the user select the type of drawing ; the style of histogram ; the sample size ; and the histogram features, as follows.

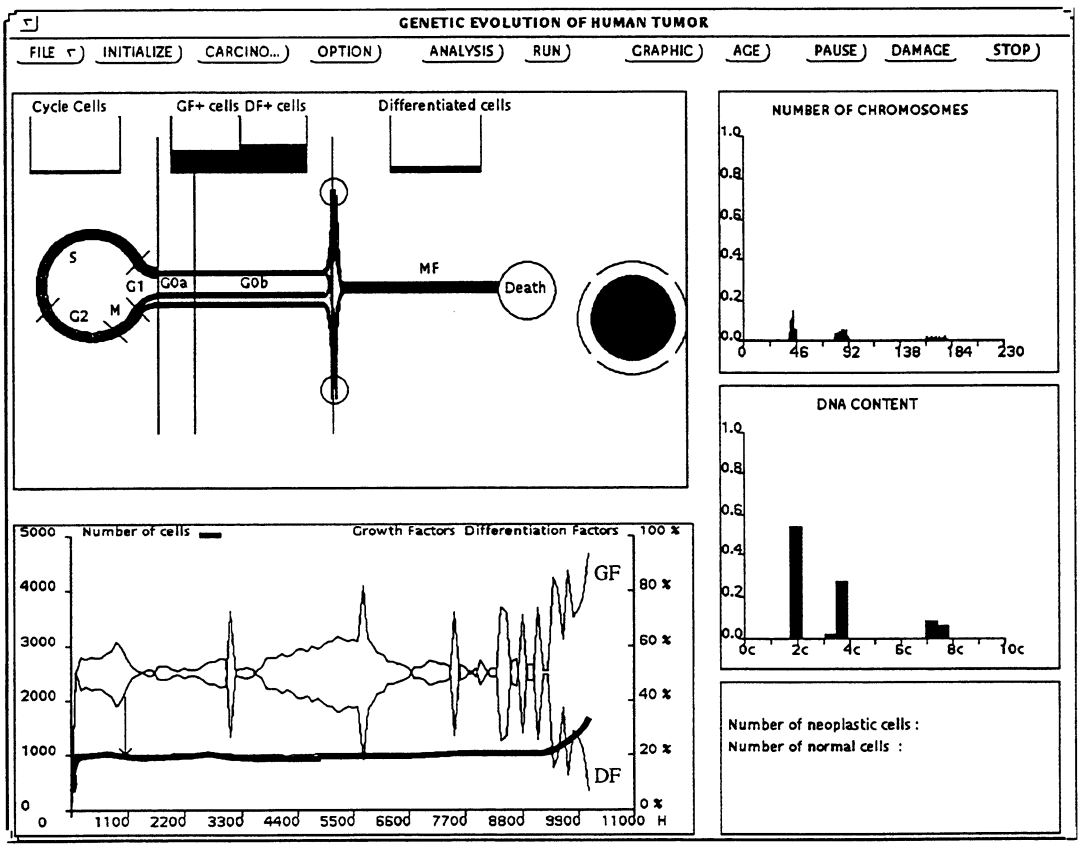


Figure 2 : Interface of the computer model. At the top , a series of buttoms permits to initialize and or to access to the differents options. The scheme of the model is given in the top-left window, with the updated corresponding proportions in the different cells compatment (open boxes). The black circle represent the proportions of mutated cells with regard to the total population (white circle). By clicking in this circle, the user simulates a Fine Needle Aspiration Biopsy. The number of cells as well as the proportion of Growth and Differentiation factors are represented in the left-bottom windows. DNA and ploidy histograms of the population are given in the right windows. All these characteristics are updated according to the set-up .

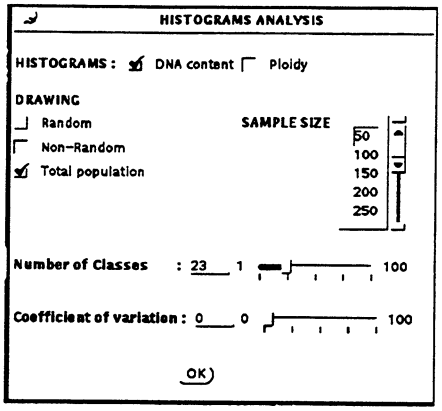


Figure 3 : Dialog Box for sample analysis. The user can studied some non-random or random drawing samples of various size. The ploidy histogram and DNA histogram can be displayed at anay time and the user can choose the number of classes and the coefficient of variation.

Drawing types

Intratumural heterogeneity of DNA ploidy are handled by the model since the user can draw as many different samples as he wishes.

Moreover, samples can be drawn randomly or non-randomly. The cells drawn randomly have the same probability of beloning to any of the cell progenies of the simulated tumour. On the contrary; the cells drawn non-randomly are selected on the basis of their neighbourhood on the assumption that the cells of the same progeny stay close to one another in the population (local clone) as shown in the Figure 4. This selection is more realistic than a random selection since it approaches, although imperfectly, the most likely topolographical behaviour of cell clones during tumour growth.

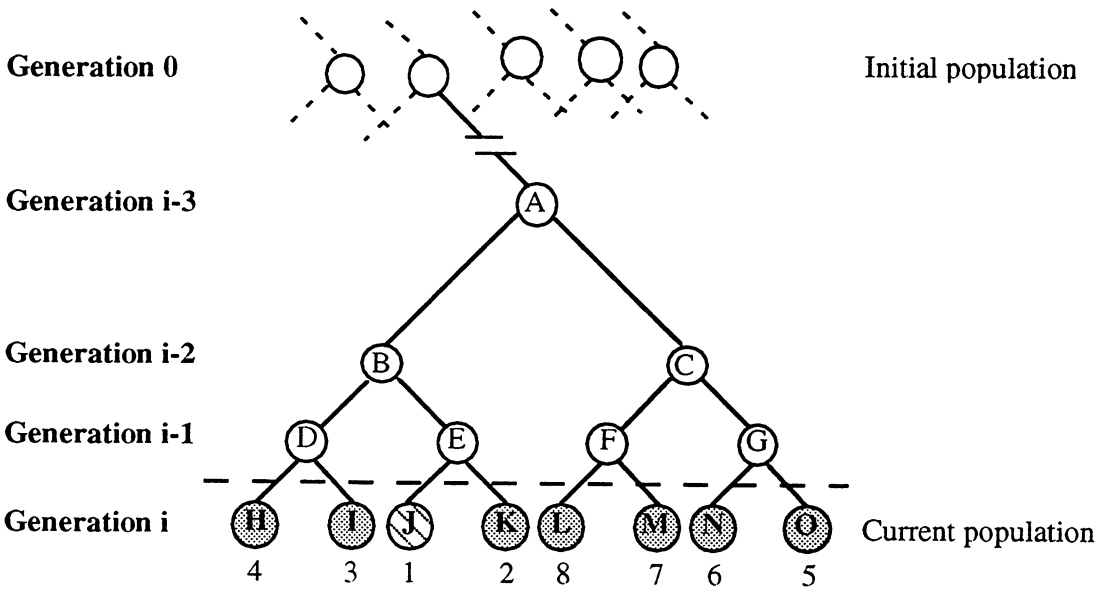


Figure 4 : Non-random drawing. The process starts with selecting randomly the first cell (J) and adding new cells (successively K, I, H, O, N, M, L...) according to generation proximity supposed to represent topographical neighbourhood since new cells usually keep close to one another. Increasing the sample size thus increases the clonal diversity.

Histogram

Distinction between the ploidy (number of chromosomes) and DNA pattern (DNA content) of a cell sample is handled by the model which provides both DNA histogram and ploidy histogram on request for a given sample. It should be noted that a normal G2 cell is in the diploid state (46 chromosomes) but its DNA content (4c) is twice that of a diploid cell in G0/G1 phase (2c) and a tetraploid G0/G1 cell has the same amount of DNA as a normal G2 cell (4c) but twice the number of chromosomes (92) ! The predicted DNA content corresponding to several ploidy status are given in table 1. These calculations assume an average arbitrary DNA content of 2c/46 per chromosome. As a consequence of chromosome decreasing size (from 1 to 22 in Human), the aneuploïdy changes the DNA content in according to which chromosomes are involved. Such inaccuracy of the model can be neglected with regard to the image cytometry DNA measurement errors.

Ploidy status	**Number of Chromosomes**	**DNA content**
diploïd (G0 or G1)	46	2c
diploïd (S)	46	between 2c and 4c
diploïd (G2, M)	46	4c
tetraploïd	92	4c
triploïd	69 (46+23)	3c (2c + c)
aneuploïd	51(46+5)	2.2c(2c+0.2c)

Table 1: Relation between number of chromosomes and DNA content for various ploidy status.

DNA histogram features

Number of classes

The number of classes can be chosen by the user from 1 to 100 classes within the interval 0c to 10c DNA content. The comparison between several histograms, for the same sample but with different number of classes, allows visualization of the influence of this choice on the visual interpretation of the DNA histogram and thus on the diagnosis.

Sample size

The sample size can be chosen by the user who can draw as many samples of different sizes as he wishes, from 50 cells to the total cells population.

Coefficient of variation of cytometric system

The coefficient of variation (CV) is one of the major limitations of image cytometry when compared to flow cytometry where a CV of 3% to 4% is usually obtained on a reference population. The CV of image analysis systems is more sensitive to chromatin differential condensation between cells and ranges from 5% to 6%. The problem of the influence of the CV

on DNA histogram interpretation is handled by the model since DNA histograms can be represented according to any chosen CV.
For example, if a CV of 10% is set-up, a cell containing 46 chromosomes will be considered as belonging to a normal population of mean 46 and of standard deviation of 4.6. Then, considering the Gaussian probability density, a value is randomly drawn between (46 - 3*4.6) and (46 + 3*4.6). The higher the drawing sample size, the more satisfying the Monte-carlo simulation is. This method is still imperfect however because the measurement error varies according to ploidy status, DNA content and chromatin condensation as a consequence of the position of cell in the cycle.

DNA histogram descriptors

The major DNA histogram descriptors already reported in the literature are calculated for any sample and refer to :

- Mean DNA Index
- 2cDI (8)
- Ploidy Balance and Proliferation Index (10)
- Entropy computed from 80 classes (7) .
- Malignancy Grade (8,11)

Results

The results presented here were obtained with the following parameters set by default unless otherwise specified :

- Initial population = 1000 cells,
- Compartments respective size : Proliferating cells = 5%, Resting cells = 5%, Differentiated cells = 90%,
- Phase durations : G1a = 2hrs, G1b = 2hrs, S = 8hrs, G2 = 2hrs, M = 2hrs, G0a = 10hrs,
- Differentiated and resting cells life span = 1000hrs,
- Receptor status : Number per cell = 100, GF and DF proportion threshold : g = d = 50%, Corresponding affinity coefficient : x = y = 100,
- Proliferating effective abnormality likelihood : Padv = 0.01
- Proliferating non effective abnormality likelihood : Pstruct = 0.001 , Ploss = 0.0008, Ptetra = 0.005

As shown in Fig. 2, a proliferating effective mutation probability (Padv) of 0.01, occurring from 1000 hr, when steady state is reached, resulted in a drastic increase in the population cell number up to 1612 cells from 9500 hr to the end of experiment at 10100 hr. The total number of cells generated by this simulation was of 20 056 cells.

The three cell compartments respective proportions became unbalanced with fast predominance of 81% undifferentiated resting cells compared to 2% proliferating and 11% differentiated cells. This reference population was used to analyse various aspects of DNA histograms and sampling conditions by using the interface dialog box shown in Figure 3.

Ploidy histogram versus DNA histogram of the reference population

Figure 5 shows ploidy histogram, DNA histogram and derived statistical figures of the total simulated population. Such a population is comprised of 3 cell cohorts, near-diploid, near tetraploid and near octaploid, each one having lost a few chromosomes. None of the cells was euploid in this population, i.e. they are all aneuploid. It should be noted that given a representation in 20 or 30 classes, the hypodiploid population would be considered as diploid. All statistical parameters have a high value, excepted the Proliferation Index of Opfermann which was equal to 0, as expected. The Ploidy Balance was positive despite the fact that all cells were aneuploid.

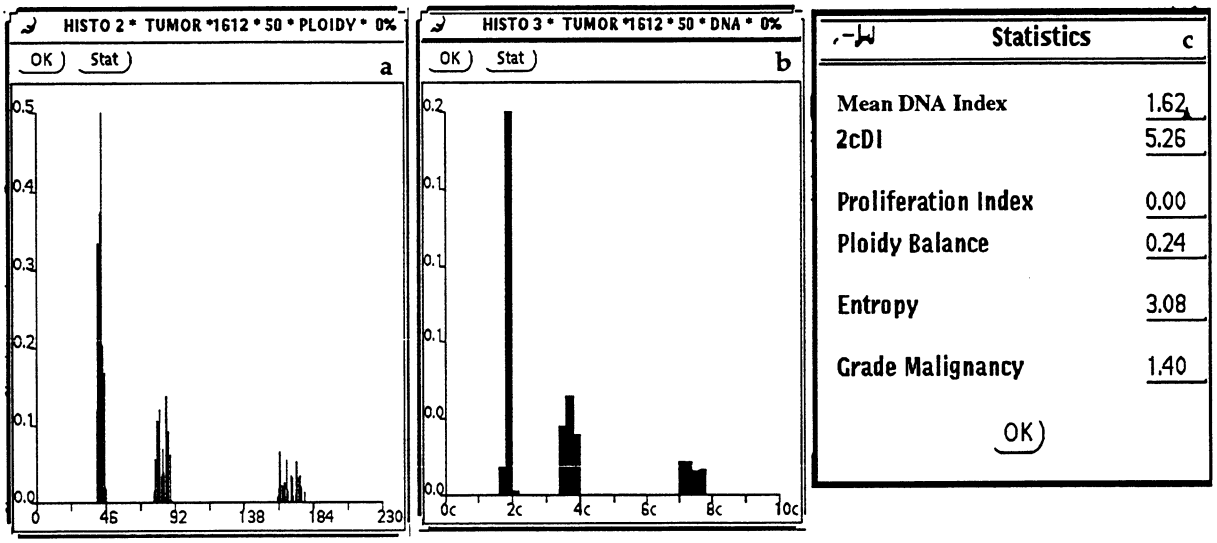


Figure 5 : Characteristics of the reference population : DNA histogram (a), ploidy histogram (b) and statistical features (c).

Effects of sample size

The first 50, 100, 200 and 300 cells of non-random drawing samples were analysed. DNA histograms and statistical features are given in Fig 6. The sample of 50 cells exhibited only the near-diploid sub-population. The three other samples showed the three sub-populations, but only the sample of 300 cells, contained the three sub-populations in about the same proportions and derived statistical features as the reference population of Fig 5b. This case illustrated one of the most biased drawing since the majority of non-random drawing picked up the three sub-populations, even in the 50 cell sample, but rarely in proportions close to that of the reference population.

Effects of drawing

Four samples of 150 cells were drawn independently. As shown in Fig 7., none was perfectly representative of the reference population. The DNA histograms 7a and 7b showed only the near-diploid and the near-diploid plus the near-tetraploid sub-populations respectively. Moreover, these DNA histograms and corresponding statistical features might be interpreted as normal, given some rescaling inaccuracy.

The DNA histograms 7c and 7d showed the three sub-populations but in incorrect proportions and highly variable statistical features.

Effect of the Coefficient of Variation

A sample of 200 cells, only slightly different from the reference population, was analyzed by applying an increasing CV of the DNA content measurements. DNA histograms and statistical features of this sample, were built with a CV of 0%, 2%, 5%, 8%, 10 % and 15 % respectively. As shown in Fig 8., an increasing CV from 8a (no measurement error) to 8f broadened the peaks and raised the DNA content of a cell ; the distribution is much larger and the probability of cells appearing in between the modal peaks increased. These cells might be wrongly interpreted as proliferating cells. In contradistinction to 2cDI, DNA mean index and DNA-MG, which did not vary significantly as CV increased, the Ploidy Balance, Proliferation Index and, entropy were extremely sensitive to DNA measurement errors.

Cumulative effect

To investigate the cumulated effect of sample size, selection, CV and histogram classes number, a series independent samples, illustrated in Fig 9., clearly demonstrated that the ploidy characteristics of the reference population can not be ascertained.

Discussion

One of our motivations in designing a computer model for the emergence of neoplasm in growing cell population was to offer the pathologists with a tool to investigate the possible major drawbacks of DNA histograms interpretation. The numerous DNA histograms simulated here illustrate the possible variability and misinterpretation as far as cancer diagnosis and prognosis are concerned.

For this purpose, the model differentiates between the ploidy and DNA histograms of the whole reference cell population and a number of FNAB-like samples for which the bias introduced by sampling, sample size, class number and coefficient of variation are illustrated. Moreover, some of the statistical features commonly used to support diagnosis and prognosis were calculated and objectively showed the frequent unreliability of the overall approach to caryological disorders based on image cytrometry of cell DNA content.

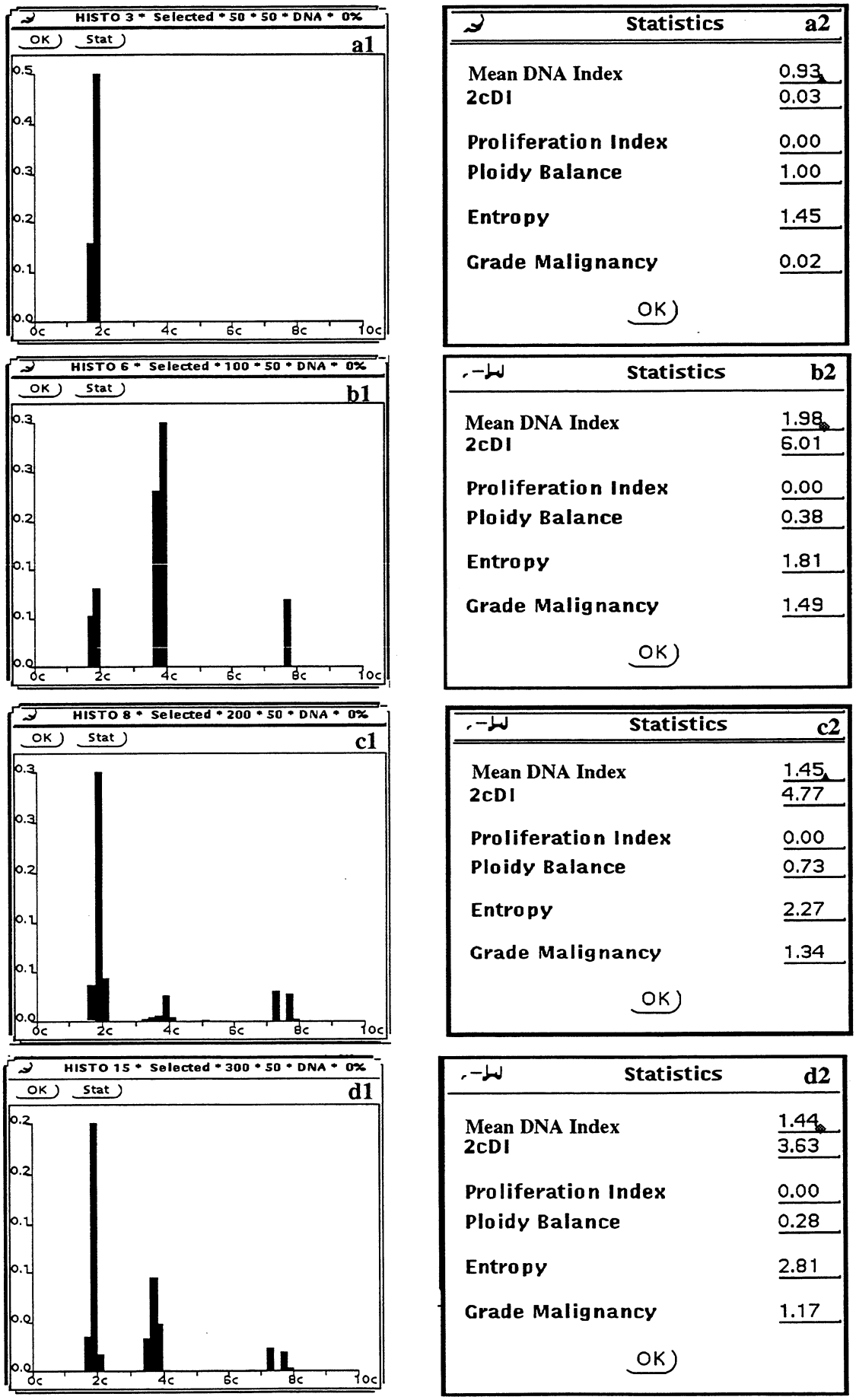


Figure 6 : Effect of sample size on DNA histograms and statistical features. From the same drawing, samples of 50 cells (a1-a2), 100 cells (b1-b2), 200 cells (c1-c2) and 300 cells (d1-d2) are representated.

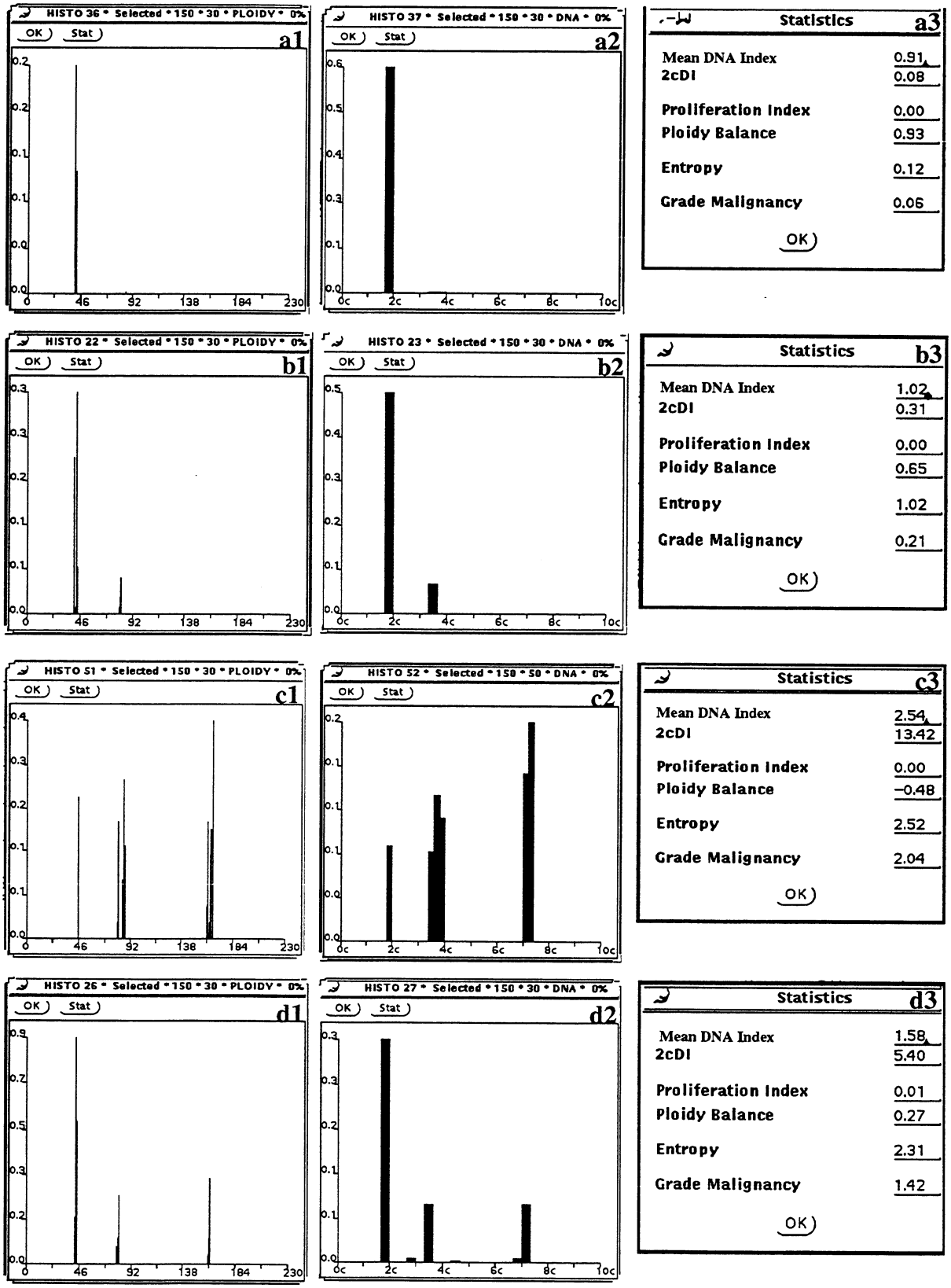


Figure 7 : Effect of non-random drawing on Ploidy histogram (a1-d1), DNA histogram (a2-d2) and statistical features (a3-d3) of 4 independant samples of 150 cells.

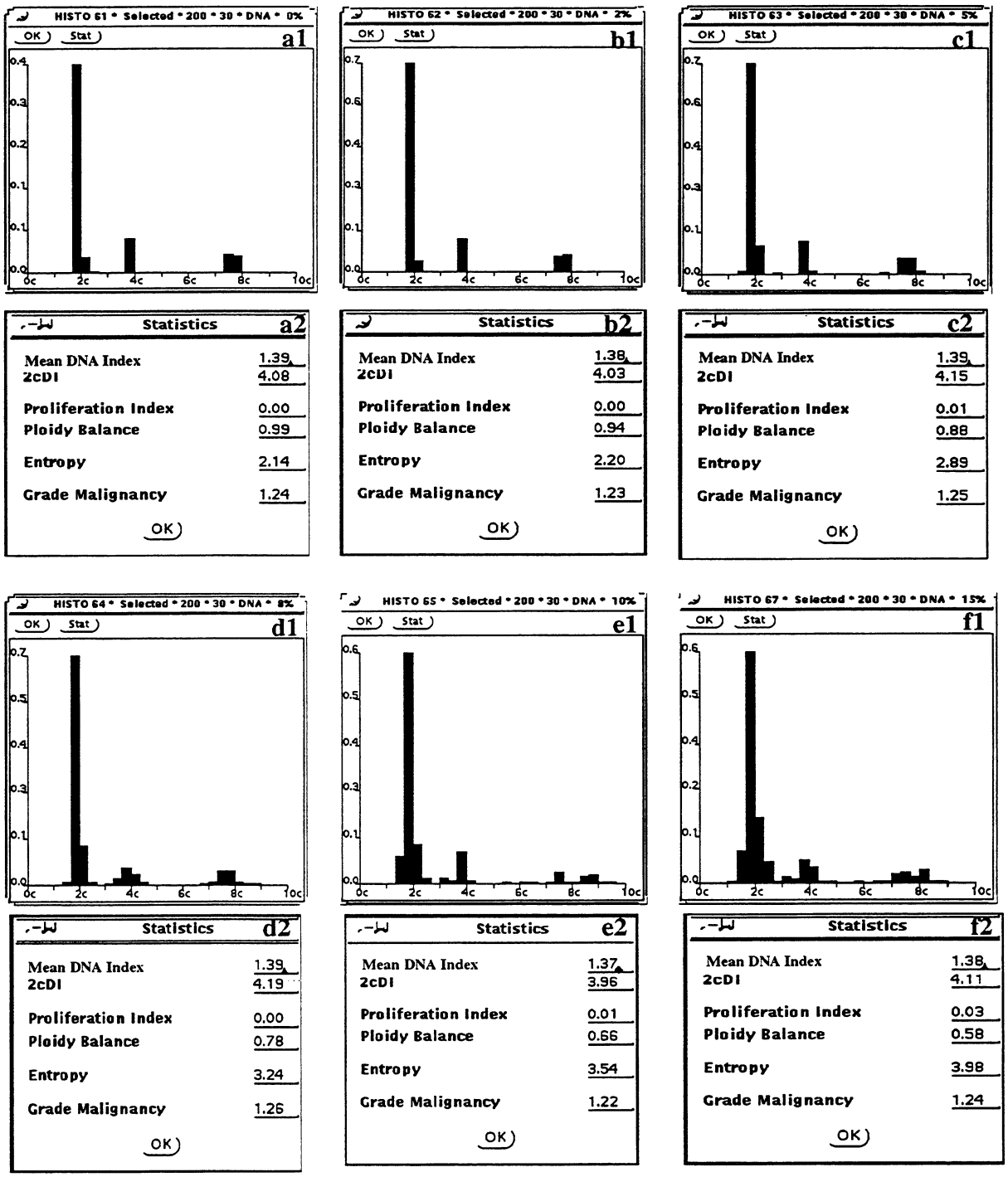


Figure 8 : Effect of Coefficient of Variation on DNA histogram (a1-f1) and statistical features (a2-f2) of a non-random sample of 200 cells. The value of the CV was respectively 0% (a), 2% (b), 5% (c), 8% (d), 10% (e) and 15% (f).

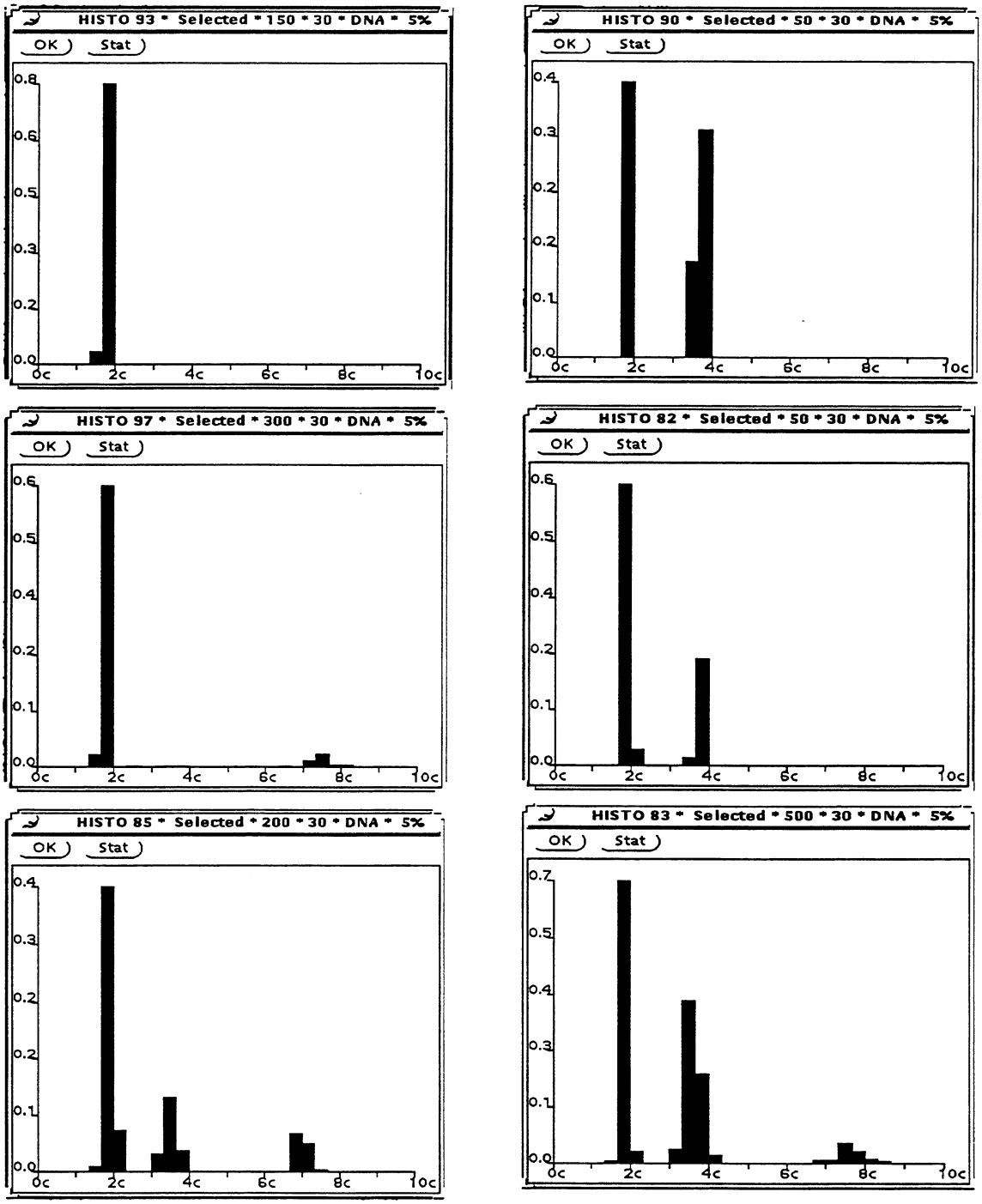


Figure 9 : Galery of DNA histograms of independant samples of different size drawn from the reference population. The applied CV (5%) and the number of classes (50) were the same for each histogram.

A pre-requisite for the reliability of the results of cytometric data from clinical materials is that the DNA histogram obtained from a random biopsy taken from a solid tumour actually represents the DNA pattern of the entire lesion. It is well known since the work of Sasaki (9) on bladder carcinomas, that some aneuploid lines could not have been detected in a single sample of the tumours. Whether or not the biopsies are representativeness is rarely demonstrated in the considerable number of papers already published. It is difficult in any case to acertain the representativeness of DNA histograms since it depends on intratumoural heterogeneity in addition to sampling and measurement bias. In carcinoma of several organs, the frequency of DNA aneuploid histograms increased as the number of samples increased (12, 13, 14). These results suggested that intratumoural heterogeneity with regard to DNA ploidy is more frequent than previously estimated. In this respect many authors argue that aspirated cellular material is more representative of the cellular composition of a tumour (12) than imprints and smears. This may be explained by the fact that the needle passes through the entire tumour mass and collects cells from many parts of the nodule. Credence in this assumption has, never the less, to be modified given the results of the simulation presented here. In view of our results, the representativeness of a simulated single FNAB is always questionable except when the DNA histogram has a very irregular profile like AUER's type IV. But, for all other histogram profiles, poor representativeness may frequently lead to underestimating the spectrum of ploidy abnormalities possibly present among the tumoural clones, or at least lead to identify the clones but in abnormal proportions.

Sample size

The variability of the histograms observed in our results with respect to the sample size, demonstrates the danger of interpretating the DNA histograms obtained from a low number of cells. It is obvious that the higher number, the greater is the probability of obtaining a representative sample. But a large sample may be poor in diagnostic and prognostic relevant cells as encountered in some flow cytometry analyses, although the total number of analysed cells reaches several thousands (15). There is no fixed rule to define the sample size. While 300 cells is considered to be sufficient for DNA histogram interpretation by some authors (6), others suggest a sample no less than 1000 or 3000 cells (16). An approach based on bootstrap analysis applied to DNA histogram outline stability has been recently proposed to overcome the difficulties of setting up simple rules for the sample size (17). This opens the way to image cytometry DNA measurement under statistical control until an objective arresting criterion is satisfied (18).

Histogram features

The Coefficient of Variation, which results from the instrumental error and biological variations (chromatin condensation) may also result in interpretation errors especially if the CV is not taken into account in the construction of histogram construction (number and size of

classes). A near-diploid population can be mistakenly considered to be diploid if the number of classes is too small. When the a coefficient of variation is too high, a non-proliferating euploid or aneuploid population can be interpreted as being proliferating from visual and statistical features (Entropy, Proliferation Index) as well. The consequence of these two errors on the diagnosis and prognosis are well known, especially if proliferating activity (SPF) is considered as a major argument in favor of adjuvant chemotherapy.

Diploid reference and histogram rescaling

A major but difficult problem that arises in ploidy analysis is that the rescaling of the DNA histogram may introduce additional inaccuracy (19) not demonstrated by the simulations presented here. Indeed, the choice of a diploid reference standard is not straightforward. Internal standards, such as lymphocytes, neutrophiles, stroma cells or normal epithelial cells, present in the cell sample can be used. Moreover, the ratio between the non-epithelial and malignant cells in carcinomas depends on the cell type, the preparation techniques and the image analysis system used (glare). Coen (20) proposed a protocol for setting up a standard to rescale DNA histograms which makes use of intra- and inter-imprint variations. In spite of a well defined rescaling protocol a modal population centered on 2.2c is usually interpreted as diploid to inperfect rescaling (Fig 7a).

Statistical features

Since the visual classification proposed by Auer (5), numerous authors have proposed other indices to characterize the distribution of cells according to their measured DNA content. These indices, more objective and quantitative than visual assessment, were expected to provide a better discrimination between , for example type I and type II of Auer. To be useful, these statistical features require that thresholds be defined for the purposes of discriminating, patients with short versus long survival (7, 8, 10), for example. Our model clearly shows that such parameters are too dependent on the sampling conditions. In spite of published significant correlations between some of these parameters and clinical features or patient outcome (21, 22, 23), they are of little assistance to the pathologist who has to formulate a diagnosis and a prognosis for an individual patient. Our results showed that the statistical features provide no more accuracy than the DNA histograms, especially when the CV is high. Our results thus demonstrated that Entropy, Ploidy Balance and Proliferation Index are too dependant on the number of classes and CV, to be secure enough in clinical routine. In daily interpretation of ploidy histograms of breast cancers, and as far as survival after the first diagnosis is concerned, the visual classification limited to euploid versus aneuploid histograms proved to be the most reliable (24).

Obviously, statistical features are only helpful to assist visual interpretation provided the DNA histograms are constructed after taking into account the coefficient of variation, the number of cells, and the representativeness of the sample versus the apparent ploidy abnormalities.

Recent developments in expert systems embedded statistics such as DS theory, Certainty Factors or Possibility theory (25) could be of considerable assistance to improve the interpretation of DNA histogram (26,27). The model presented here will make it possible to develop and test such new approaches taking into consideration operational features such as the CV and the sample size and clinical data as well to provide a ranking of alternative interpretations according to 'real-numbered' score.

References

1 Guillaud M. and Brugal Gérard. Computer Model for the Emergence of Neoplasia in Growing Cell Populations. Part I : Simulation of Growth Parameters. Submitted to Cancer Research

2 Shackney S.E., Smith C.A, Miller B.W., Burholt D.R., Murtha K., Giles H.R., Ketterer D.M. and Pollice A.A. Model for the genetic evolution of human solid tumors. Cancer Research, *49* : 3344-3354, 1989.

3 Atkin NB, Richards BM. Desoxyribonucleic acid in human tumours as measured by microspectrophotometry of feulgen stain: A comparison of tumours arising at different sites. Br J Cancer 10:769-786, 1956.

4 Caspersson T., Lomakka G., Casspersson O. Quantitative cytochemical methods for the study of tumor cell populations. Biochem Pharmacol 4:113-127, 1960.

5 Auer G.U., Caspersson T.O. and Wallgren A.S. DNA Content and Survival in Mammary Carcinoma. Analyt Quant Cytol Histol ; *2* : 161- 165, 1980.

6 Bartels P.H, Weber J.E and Bibbo M. Ploidy Pattern Analysis, Statistical Considerations. Analyt Quant Cytol Histol ;*7* : 126-130, 1985

7 Stenkvist B. and Strande G.Entropy as an algorithm for the statistical description of DNA cytometric data obtained by image analysis microscopy. Anal Cell Pathol ; *2* : 159-165, 1990

8 Bocking A., Adler C-P., Common H.H., Hilgarth M., Granzen B., Auffermann W. Algorithm for a DNA-Cytophotometric diagnosis and grading of malignancy. Analyt Quant Cytol Histol ; 6 : 1-8, 1984

9 Sasaki K., Hamano K., Kinjo M. and Hara S. Intratumoral heterogeneity in DNA ploidy of bladder carcinomas. Oncology ; *49*: 219-222, 1992

10 Opfermann M., Brugal G., Vassilakos P.Cytometry of breast carcinoma : Significance of ploidie Balance and Proliferation index.Cytometry ;*8* : 217-224, 1987.

11 Böcking A., Chatelain R., Bisterfield S., Noll E., Biesterfelf D., Wohltmann C., Goecke C.DNA grading of malignancy in breast cancer. Prognostic validity, reproductibility and comparison with other classifications.Analyt Quant Cytol Histol ; 11: 73-80, 1989

12 Wersto R.P., Liblit R.L., Deitch D., Koss L.G.Variability in DNA measurements in multiple tumor samples of human colonic carcinoma.Cancer ; *67*: 106-115, 1991

13 Beerman H., Smith V.T.H.B.M., Kluin PM., Bonsing B.A., Hermans J. and Cornelisses C.J.Flow cytometric analysis of DNA stemline heterogeneity in primary and metastatic breast cancer. Cytometry ; *12*: 147-154, 1991

14 Pennes DR., Naylor B. and Rebenr M. Fine Needle Aspiration Biopsy of the Breast. Influence of the number of passes and the sample size on the diagnostic yiels. Acta Cytologica; *34* : 673-676, 1990

15 Friedlander M. Flow cytometric analysis in Quantitative pathology. *In* : Manual of Quantitative Pathology in cancer diéagnosis and prognosis. Jan P.A.Baak, Springer--Verlag Berlin heidelberg New Yorrk, p 245-268, 1991

16 Wied GL, Bartels PH., Bibbo M., Dytch HE. Image analysis and quantitative cyto- and histopathology. Techn Report, N°2, The international academy of cytology, 1988.

17 Guillaud M. and Chassery J-M. Histogram analysis by use of L-moments, Linear functions of order statistics. Statistiques et Analyse des données; 16, 85-106, 1991

18 Guillaud M and Chassery J-M. Cytometric data acquisition under statistical control. *In press*

19 Kiss R., Gasperin P., Verhest A. and Pasteels J.-L. Modification of tumor ploidy level via the choice of tissue taken as diploid reference in the digital image analysis of Feulgen-stained nuclei.Modern Pathology; *5*: 655-660, 1992.

20 Cohen H., Pauwels M. and Roels F. The rat liver cell nuclear imprint as a standard for DNA measurements. ACP ;*4* , 276-285, 1992.

21 Merkel D.E. and Osborne CK. Prognostic factors in breast cancer. Hematol. Oncol Clin North Am. ; *3*:641-652, 1989.

22 Coulson PB., Thornthwaite JT., Wooley TW., Sugarbaker EV., Seckinger D. Prognostic indicators including DNA histograms type, receptor content, and staging related to human breats cancer patient survival. Cancer Res. ; *44* : 4187-4196, 1984.

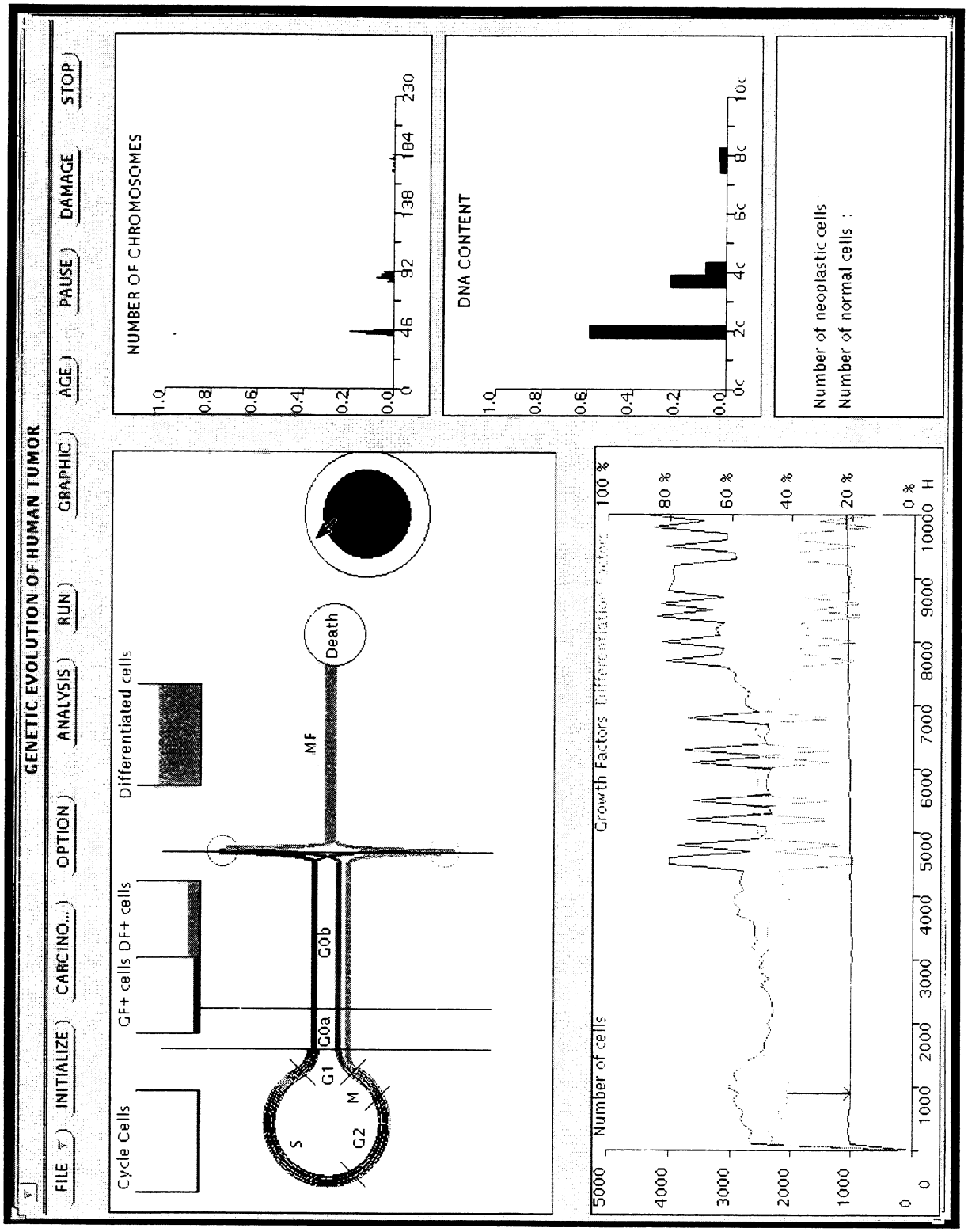
23 Cornelisse CJ., Van de velde CJ., Caspera RJ. , Moolenar AJ., Hermans J. DNA ploidy and survival in breast cancer patients. Cytometry ; *8* : 225-234, 1987.

24 Guillaud M., Louis J. and Seigneurin D. DNA image cytometry of breast carcinomas. *In press*

25 Van Ginneken A.M. and Smeulders A.W.M. Reasoning in uncertainties. An analysis of five strategies and their suitability in pathology.Analyt Quant Cytol Histol ; *13* : 93-109, 1991

26 Bartels P.H., Thompson D. and Weber J.E. Expert systems in histopathology. IV. The management of Uncertainty. Analyt Quant Cytol Histol ; *14* : 1-13, 1992

27 Bartels P.H., Thompson D. and Weber J.E. Expert systems in histopathology. V. DS theory, certainty factors and Possibility theory. Analyt Quant Cytol Histol ; *14*, 165-174 1992



INTERFACE DU LOGICIEL DE MODELISATION DE L'EMERGENCE ET DE LA CROISSANCE D'UNE TUMEUR DANS UN TISSU SAIN DIFFERENCIE

CONCLUSIONS ET PERSPECTIVES

Durant cette thèse, deux grands axes de recherches ont été développés. Ils ont aboutit à la conception et au développement de deux applications dinstinctes mais proches quant à leur objectifs. En effet, l'intérêt de ces deux méthodes est d'améliorer la qualité de l'analyse, de l'interprétation et de la compréhension de phénomènes mis en évidence par les données obtenues en analyse d'images .

La première approche, plus orientée à la recherche clinique, nous a permis de développer une méthode de suivi d'acquisiton de données cytométriques en temps réel. Le contrôle de qualité des analyses cytométriques s'impose comme indispensable pour l'améloration du diagnostic et du pronostic médical. Confronté à un problème de reconnaissance de formes, évoluant au cours du temps, l'intérêt de la méthode du Bootstrap pour mesurer la stabilité de ces formes, caractérisées par des données scalaires ou vectorielles, a été démontré. Certains aspects portant sur la validation théorique de l'utilisation de la méthode du Bootstrap doivent être étudiés. Une implémentation de cette méthode dans les systèmes d'analyses d'images en routine clinique sera alors possible, en particulier pour l'analyse de l'ADN.

La deuxième approche a aboutit à l'élaboration d'un modèle ou plutôt d'une plate-forme de simulation de l'émergence et de la croissance d'une tumeur dans un tissu différencié sain. Outre l'aspect relativement nouveau de la modélisation biologique proposée, l'architecture

informatique de ce modèle permet d'appréhender la plupart des problèmes liés à l'analyse statistique et à l'interprétation des échantillons tumoraux.

Les travaux réalisés pendant cette thèse participent aux nombreux efforts réalisés ces dernières années pour une meilleure compréhension des caractéristiques biologiques des tumeurs humaines, à travers l'analyse cytométrique d'échantillons tumoraux.

Perspectives

Qualité des analyses cytométriques

Au niveau de l'acquisition d'histogrames d'ADN en temps réel, nous allons développé un algorithme plus orienté vers l'aide au diagnostic(Figure 1). En conservant l'étude de la stabilité des variances Bootstrap des primitives des histogrammes comme principe de base, nous pensons l'intégrer dans une structure plus complète. Une telle approche, qui fait appel aux outils et aux principes de l'Intelligence Artificielle Distribuée, permettra d'intégrer des informations de natures différentes et pertinentes pour la mesure de la qualité de l'analyse. Ces informations, telles que le type de diagnostic désiré, la nature du tissu (les préparations provenant de lavages (vessie) ou de ponctions par aiguilles fines (sein) ne présentent pas les mêmes caractéristiques), les informations cliniques sur le patient ou la qualité de la préparation doivent être intégrées dans la définition et le contrôle de la qualité de l'analyse et donc du diagnostic. D'autres applications que l'analyse de l'ADN pourront bien sûr bénéficier de cette approche.

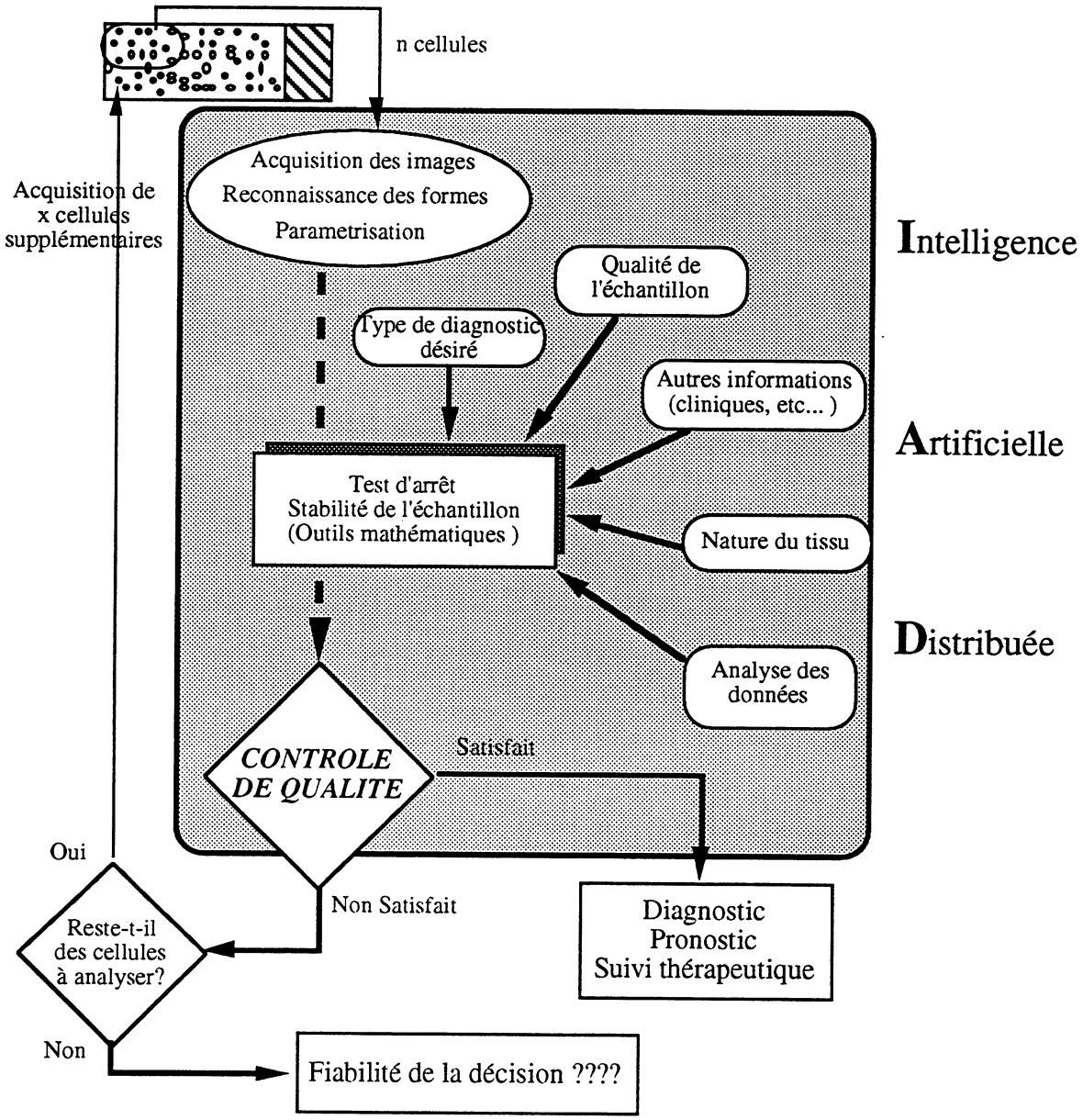


Figure 1 : Utilisation de l'**I**ntelligence **A**rtificielle **D**istribuée pour la définition de la qualité et de la fiabilité de l'analyse cytométrique d'un échantillon biologique.

Plate-forme de modélisation

Au niveau du modèle de simulation de l'évolution d'une tumeur, les perspectives du travail effectué durant ma thèse ont déjà pris forme dans le cadre du projet MOSAIC (

MOdélisation et **S**imulation de l'**A**dhérence et des **I**nteraction **C**ellulaires). Ce projet intègre et réunis les travaux réalisés dans notre équipe sur la modélisation biologique, la sociologie cellulaire et les systèmes multi-agents. L'objectif prncipal concerne la conception et l'implantation d'une plate-forme informatique d'étude et de simulation des mécanismes de prolifération cellulaire et d'homéostasie tissulaire. Nous proposons un outil de simulation de la "vie cellulaire" basé sur une modélisation simplifiée de la cellule et de ses interactions avec son environnement afin de comparer dans un premier temps, les résultats de la simulation avec ceux de la biologie expérimentale. Il s'agit ensuite d'étudier dans quelle mesure cette plate-forme s'avère un moyen intéressant d'investigation et de compréhension des mécanismes mis en jeu dans le développement d'un cancer. L'Intelligence Artificielle Distribuée nous paraît offrir un cadre technologique apte à répondre à ces exigences, en permettant une expression ascendante, distribuée et locale des comportements, fondée sur la notion d'agent autonome de type "réactif". Cette approche ascendante signifie que seules les règles de comportement de chaque composant (cellules ..) ainsi que le système de communication sont spécifiés. Le comportement global et "social" du tissu, comportement à la fois fonctionnel et structural émerge des interactions entre les différents composants du tissu.

Les premiers résultats illustrent d'une part l'émergence de l'homéostasie d'un tissu différencié et d'autre part l'émergence d'une tumeur au sein de ce tissu.

La première phase de conception de cette nouvelle plate-forme a fait l'objet d'une publication, présentée aux journées **IAD & SMA**, de Toulouse du 7 et 8 avril 1993.

MODELE DE SIMULATION DE L'EMERGENCE ET DE LA CROISSANCE DES TUMEURS CANCEREUSES DANS UN TISSU SAIN

M. Guillaud, R. Marcelpoil, O. Baujard, E. Beaurepaire
Laboratoire TIM3 - ERFMQ
Bâtiment CERMO, BP 53 X
38041 Grenoble Cedex FRANCE
Tel : 33-76-51-48-13 Fax : 33-76-51-49-48
Email : guillaud@imag.fr, baujard@imag.fr, repaire@imag.fr, Marpoil@imag.fr

MOTS-CLES

Vie artificielle. Emergence. Homeostasie. Cancer. Modélisation

RESUME

Cet article décrit la conception et l'implantation d'une plate-forme informatique d'étude et de simulation des mécanismes de prolifération cellulaire et d'homéostasie tissulaire. Nous proposons un outil de simulation de la "vie cellulaire" basé sur une modélisation simplifiée de la cellule et de ses interactions avec son environnement afin de comparer dans un premier temps, les résultats de la simulation avec ceux de la biologie expérimentale. Dans un deuxième temps, nous verrons dans quelle mesure cette plate-forme s'avère un moyen intéressant d'investigation et de compréhension des mécanismes mis en jeu dans le développement d'un cancer. L'Intelligence Artificielle Distribuée nous paraît offrir un cadre technologique apte à répondre à ces exigences, en permettant une expression ascendante, distribuée et locale des comportements, fondée sur la notion d'agent autonome de type "réactif". Cette approche ascendante signifie que seules les règles de comportement de chaque composant (cellules ..) ainsi que le système de communication sont spécifiés. Le comportement global et "social" du tissu, comportement à la fois fonctionnel et structural émerge des interactions entre les différents composants du tissu.

Les premiers résultats illustrent d'une part l'émergence de l'homéostasie d'un tissu différencié et d'autre part l'émergence d'une tumeur au sein de ce tissu.

1.INTRODUCTION

Cet article décrit les travaux de conception et d'implantation d'une plate-forme informatique pour l'étude et la simulation des mécanismes de prolifération cellulaire et d'homéostasie tissulaire. L'ambition profonde est de progresser dans la compréhension du

développement des cancers et de la progression des tumeurs. Plus modestement, nous proposons un outil informatique de simulation de la " vie cellulaire" basé sur une modélisation simplifiée de la cellule et de ses interactions avec son environnement afin de comparer, dans un premier temps, les résultats de la simulation aux résultats biologiques et médicaux. Dans un deuxième temps seulement, nous verrons dans quelle mesure cet outil de simulation s'avère être un moyen intéressant d'investigation et de compréhension des mécanismes mis en jeu dans le développement d'un cancer. Des modèles explicites des comportements, des échanges cellulaires et de l'environnement doivent être proposés. Une première implantation de cette plate-forme, basée sur les techniques de l'Intelligence Artificielle Distribuée a été développé à partir de travaux déjà existants (Baujard, 92) (Guillaud, 93) (Marcelpoil, 92a). Cet article présente les premiers résultats, après un rapide état de l'art.

2. PROBLEMATIQUE

Le travail présenté ici se situe au carrefour de la biologie du cancer, de la sociologie cellulaire, et de la vie artificielle. L'idée selon laquelle les structures naturelles (comme beaucoup de structures artificielles) représentent un compromis entre l'ordre et le chaos est de plus en plus intégré dans divers domaines scientifique comme la biologie, la cristallographie et la physique (Weaire, 84). En biologie, cette attention accrue est dirigée par la compréhension des mécanismes qui dirigent les structures tissulaires dans le but de déterminer l'organisation (Kayser,91) et les types de relations que les cellules sont capables d'établir entre elles. La notion de voisinage d'un point est au même titre cruciale pour l'analyse des formes (Ahuja,82)(Zahn, 71). O'Callaghan, en 1975, montra que la définition standard basée sur les k plus proches voisins ne reflétait pas l'association naturelle de certains groupes de points. Depuis, un nombre important d'études converge vers l'utilisation de graphes dérivé du graphe de Voronoï qui possède des propriétés très intéressante sur les voisinages de points (Dussert, 86)(Lorz, 90) (Venema, 91) (Marcelpoil,92a).

Concernant les applications cliniques, des données récentes montrent que certains cancers humains expriment des taux de récepteurs cellulaires anormaux pour les facteurs de croissance, dont l'importance est associée à une prolifération cellulaire rapide et donc à un mauvais pronostique. L'influence du niveau en récepteur aux facteurs de croissance a déjà été étudié par plusieurs auteurs (Pollack, 91). Dans ce contexte, la modélisation d'une tumeur s'avère une approche très intéressante pour comparer les résultats expérimentaux obtenus sur des cultures cellulaires *in vitro* ou *in vivo* (Foulds, 75).

Notre étude se place naturellement dans le domaine de la vie artificielle (V.A.), dont un concept clé est l'émergence de comportements (Langton, 88). La méthodologie la plus courante en V.A. est la suivante: on définit des règles de comportement pour les parties d'un système, puis on les autorise à interagir les unes avec les autres. Avec des règles correctes et le bon agencement des parties entre elles, le phénomène recherché va émerger spontanément des

interactions, comme dans le système naturel étudié. Un comportement observé non immédiatement déductible des loi locales est dit émergent.

Le paradigme multi-agents -et plus particulièrement l'approche fondée sur des agents "réactifs"- est particulièrement bien adapté au domaine de la V.A.. Dans les systèmes réactifs, chaque agent est une entité autonome possédant un comportement très simple et placée dans un environnement donné, sans représentation explicite de cet environnement et sans historique de ses actions. Ces approches ont été utilisées en robotique (Steels, 90) (Brooks, 86) et en éthologie (Drogoul, 92). Dans ces systèmes, l'environnement joue un rôle très important : il représente le monde dans lequel un agent existe (Maruichi, 90). Les agents communiquent à travers lui, et dans ces architectures en essaim (Deneubourg, 92) les interactions agents/environement suffisent pour coordonner le fonctionnement du groupe. En ce qui concerne l'émergence de comportement, qui est un point central des mondes réactifs, des travaux essaient de symboliser les phénomènes émergents pour construire sur eux des comportements de plus haut niveau (Wavish, 92) ou étudient la régulation d'un comportement émergent en agissant sur des paramètres qui peuvent l'influencer (Kephart, 89).

3. MODELES

3.1 Modèle biologique

Nous nous proposons de nous appuyer sur un modèle qui a la particularité de poser les variations de sensibilité des cellules à deux hormones (hormones de croissance et hormones de différenciation) comme étant responsables du passage d'un fonctionnement stable (normal) d'un tissu à un fonctionnement "anormal" où les cellules prolifèrent anarchiquement ; ceci étant la conséquence d'un dérèglement tant au niveau de la fonction cellulaire qu'au niveau des interactions cellulaires. Notre modèle est basé sur un certain nombre d'hypothèses biologiques décrite en détail dans (Guillaud, 93).

Le modèle proposé simule un tissu normal ayant pour parties intégrantes trois types cellulaires fonctionnellement différents. Ces trois types cellulaires à partir desquels une stabilité structurelle et fonctionnelle peut être appréhendée dans un sens biologique sont respectivement :

- Les cellules en cycle, ou proliférantes, i.e. en train de se diviser et qui donneront deux nouvelles cellules indifférenciées,

- les cellules différenciées (ou cellules spécialisées comme les cellules cardiaques), qui libèrent des hormones de différenciation et qui ont une durée de vie limitée,

- les cellules dites "latentes", (ou indifférenciées), dont le devenir (retour en cycle ou différenciation) dépend à la fois de leurs propriétés et potentialités intrinsèques (taux de récepteurs hormonaux, sensibilité aux agents mutagènes, etc..), de la concentration en hormones de croissance et en hormones de différenciation dans le milieu environnant et des contraintes de voisinages.

Le temps passé dans chacun de ces compartiments est exprimé en heures et les proportions respectives sont paramétrisables. Pour simuler la croissance cellulaire, chaque cellule progresse séquentiellement par cinq étapes : G1a->G1b->S->G2->M ou quitte le cycle entre G1a et G1b pour l'état G0.

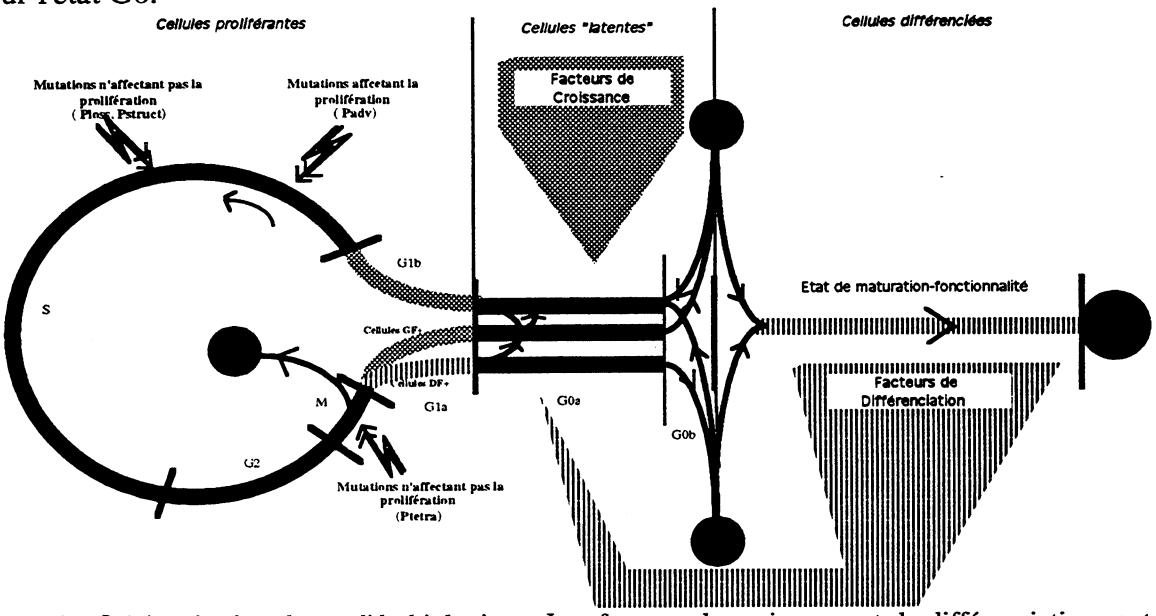


Figure 1 : Schématisation du modèle biologique. Les facteurs de croissance et de différenciation sont en compétition pour la maturation des cellules "latentes". Les cercles noirs représentent les cellules mortes.

Récepteurs et facteurs de différenciation et de croissance.

Le modèle repose sur la disponibilité de deux facteurs extracellulaires envers les cellules dites "latentes" (en G0) : les facteurs de croissances (GF) et les facteurs de différenciation (DF) qui réagissent et se couplent sur les mêmes récepteurs. Le nombre de GF disponible dans le milieu reste constant à chaque étape de la simulation et est égal au nombre de récepteurs à occuper pour qu'une cellule deviennent GF-compétente. Au contraire, les DF sont synthétisés par les cellules différenciées elles-mêmes. Durant sa vie, toute cellule différenciée synthétise et libère autant de DF qu'elle a eu besoin pour se différencier. Deux coefficient d'affinité facteurs-récepteurs x et y peuvent être ajustés de telle façon que les deux cellules soeurs n'aient pas la même affinité pour GF ou DF. Les récepteurs d'une cellule sont x fois plus affine au GF qu'au DF (cellule GF+), tandis que les récepteurs de sa soeur sont y fois plus affine au DF qu'au GF (cellule DF+). Il faut noter que l'efficacité du couplage facteur-récepteur et donc de la transition G0 dépend donc de (i) la concentration en facteurs, (ii) du nombre de récepteurs par cellules, (iii) de la différence d'affinité pour les GF et les DF. Comme le nombre de récepteurs et la concentration en GF et en DF peuvent être modifiés par des mutations génétiques, l'efficacité de couplage facteur-récepteur peut varier d'une cellule à l'autre.

Comportement des cellules dites latentes.

Les cellules filles provenant de la mitose passe dans l'étape G0a durant laquelle elles synthétisent un nombre de récepteurs qui sera égal à celui de leur mère pour ensuite passer dans

le compartiment G0. Les cellules G0 ont deux états possible : G0a et G0b. Les cellules G0a augmentent leur nombre de récepteurs jusqu'à doubler leur nombre initial. Si la concentration en GF ou DF n'est pas suffisante, les cellules G0a entre en G0b. Si cette situation demeure durant toute la durée de vie de la cellule, les cellules G0b meurt. Par contre, si la concentration en GF ou DF devient suffisante avant leur mort, les cellules G0b repartent respectivement soit en cycle soit vont se différencier. Une cellule G0 a donc trois "devenirs" possibles :

- rester en G0a puis en G0b jusqu'à la mort,
- quitter l'état G0a ou G0b pour retourner en cycle
- quitter l'état G0a ou G0b pour se différencier et être fonctionnelle jusqu'à sa mort.

Engagement cellulaire vers la croissance ou la différenciation.

Une cellule dite "latente" est GF-compétente quand la proportion g de ses récepteurs initiaux est occupé par des GF est suffisante pour repartir en cycle et donc entrer en phase G1b après un temps égal au temps passé en G0a. Parallèlement, une cellule dite latente est DF-compétente quand la proportion d de ses récepteurs initiaux occupés par des DF est suffisante pour devenir différencier après un temps nécessaire pour achever la phase G0a. Les proportions g et d sont fixées à l'initialisation.

Modélisation des mutations (Figure 1)

Le modèle gère deux sortes de mutations, agissant durant le cycle cellulaire c.à.d. ne touchant que les cellules en train de se diviser.

Le premier type d'anormalités qui intervient avec une probabilité *Padv* durant la phase S, simule des mutations de gènes impliqués dans les mécanismes de prolifération :

- Une cellule subissant une première mutation double son nombre de récepteurs mais a toujours besoin du même nombre de GF ou de DF pour quitter la phase G0. Une cellule simplement muté a donc plus de chance de devenir compétente qu'une cellule normale.

- Une cellule subissant une deuxième mutation synthétise et libère, pendant la phase S, autant de GF dont elle a eu besoin pour quitter G0. Cette cellule devient donc GF-autocrine (auto productrice de GF).

Comme ces mutations sont héréditaires, toute la descendance de la cellule mutée conserve les mêmes caractéristiques (cette descandance forme un clone cellulaire).

Le second type de mutations, survenant durant la phase S ou M, modifie seulement le génome de la cellule : il peut s'agir de tétraploïdisation (*Ptetra*), de perte de chromosomes (*Ploss*) ou d'anomalies structurales (*Pstruct*).

3.2 Modèle informatique

Modélisation d'un agent Cellule

Nous proposons donc de baser notre modèle sur le modèle d'agent présenté sur la Figure 2. Un agent cellule peut être décomposé en 4 parties: perception, connaissance, action et comportement. L'agent est placé dans un environnement, modélisé de façon explicite. La partie *perception* permet à l'agent de percevoir son environnement (perception des ressources disponibles dans le milieu, perception indirecte des cellules voisines). La partie *connaissance* décrit les ressources internes (nombres de chromosomes, nombre de récepteurs etc..) et l'état de la cellule (position dans le cycle). La partie *action* définit les actions que la cellule peut faire (primitives). Ces actions peuvent être internes (changement d'état) ou externe (division cellulaire, libération d'hormones). La partie *comportement* modélise finalement le cycle cellulaire. Selon les ressources externes et internes, la cellule changera son état et donc modifiera ses ressources propres comme celles de l'environnement.

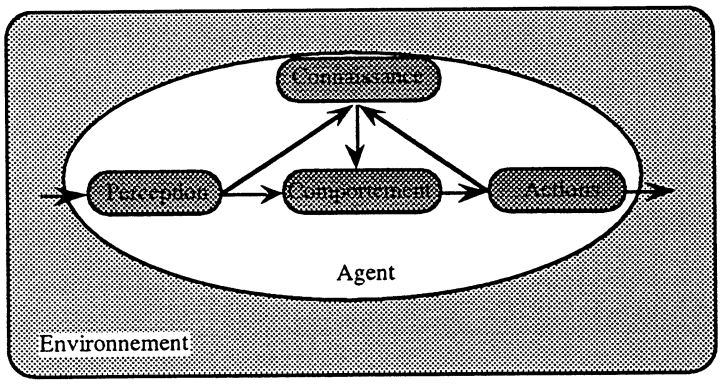


Fig. 2: Modèle d'un agent cellule

Modélisation de l'environnement

Dans notre implémentation, l'environnement topographique est formalisé par le diagramme de Voronoï (fig.4). Chaque germe du diagramme de Voronoï correspond à une cellule, et son polygone associé représente l'environnement de la cellule. Les caractéristiques de cet environnement (hormones, substances chimiques, milieu nutritif,...) peuvent être représentées au moyen de structures attachées à chaque polygone. Une cellule perçoit son propre environnement et peut également agir sur cet environnement. Elle peut par exemple se scinder en deux nouvelles cellules ou libérer des hormones.

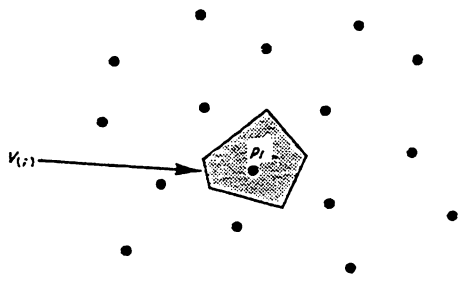


Fig. 4 : Etant donné un ensemble (S) de points, V(i) est le polygone de Voronoï of pi (un point de S). Le polygone associé à pi est l'ensemble des points (x,y) plus proche de pi que n'importe quel point de S.

L'activité de l'environnement consiste à mettre à jour régulièrement les ressources disponibles (augmentation de la quantité d'hormones, etc...) ainsi que la topographie du tissu en tenant compte des cellules qui meurent et des cellules qui naissent.

3.3. Implantation

La plate-forme est implantée en C/C++ avec une interface graphique X Window et fonctionne actuellement sur un réseau de stations de travail (Figure 3). Elle se compose essentiellement de deux parties : l'environnement et les groupes cellulaires. L'environnement est fourni avec sa propre interface graphique qui permet d'afficher le diagramme de Voronoï associé à une population cellulaire, de modifier des ressources locales à chaque cellule et d'activer les outils topographiques. Il fonctionne actuellement sur une station Silicon Graphics. Les cellules sont des instances concurrentes de classes C++, implantant le modèle d'agent. L'interface graphique de la plate-forme permet d'effectuer l'initialisation des différents paramètres du modèle, d'intervenir en cours de simulation (arrêt ou pause) et d'accéder aux outils statistiques qui permettent de suivre l'évolution de certains paramètres, numériques ou graphiques.

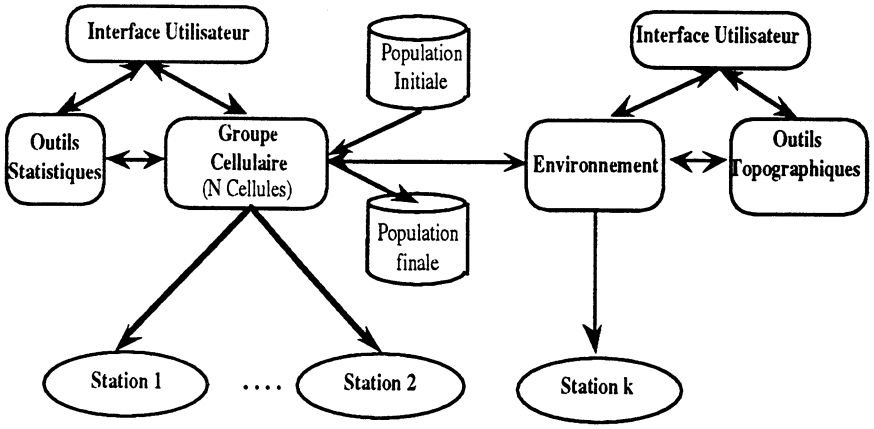


Fig. 3 La plate-forme : Vue de l'architecture globale.

3.4. Outils statistiques

L'établissement d'un diagnostic ou d'un pronostique tumoral nécessite souvent l'analyse d'un échantillon tumoral, pour effectuer par exemple l'analyse de l'histogramme d'ADN de la population. Le problème de l'échantillonnage est ici crucial. L'intérêt du modèle proposé est qu'il permet à l'utilisateur de simuler et d'étudier la variabilité intra-tumorale, tout au long de l'évolution de la tumeur. Tous les outils statistiques intégrés à notre plate-forme permettent de suivre, de représenter certaines caractéristiques de la population (degré de ploidïe, histogramme des âges, évolution du taux d'hormones). L'approche locale de notre plate-forme permet entre autre de suivre des sous-populations de cellules et de les comparer entre elles. *Une émergence fonctionnelle peut ainsi être appréhendée et quantifiée.*

3.5. Outils topographiques

Nous cherchons à explorer les relations entre les fonctions cellulaires et les positions spatiales. Ceci nous amène à résoudre deux problèmes. Le premier est de déterminer la topographie globale de la population, c.à.d. d'approcher la notion de tissu et non plus seulement de groupe de cellules. Le second problème est de déterminer l'environnement local d'une cellule en fonction de toute la population. Nous avons donc développé d'une part un modèle de paramétrisation et de quantification de la topographie de la population cellulaire (Marcelpoil,92a) et d'autre part un modèle de paramétrisation et de quantification des relations cellulaires locales. Ces approches sont basées sur la partition de l'espace de Voronoï construit à partir de l'ensemble des points représentant les positions des cellules.

De nombreux outils peuvent être construits sur le diagramme de Voronoï, en particulier pour quantifier la topographie et pour étudier l'ordre et le désordre en de nombreux points, par une simple lecture de l'espace paramétrique. *Une émergence structurale peut ainsi être appréhendée et quantifiée.*

Pour mesurer et analyser l'émergence des environnements locaux cellulaires, nous avons développé une méthode basée sur l'arbre d'Ulam (Marcelpoil, 92b). Un arbre d'Ulam est un graphe dont la forme est caractéristique de l'environnement local cellulaire, et nous mesurons l'évolution de cet environnement en calculant les différences entre les formes des arbres associés à chaque cellule. Les arbres d'Ulam (fig.5) sont représentés sur le diagramme de Voronoï. Cette méthode permet aussi d'étudier des sous-ensembles cellulaires particuliers correspondant à des configurations spécifiques. Une corrélation simple peut être établie entre la fonction d'une cellule et ses paramètres locaux. *Le lien entre émergence fonctionnelle et émergence structurale est également appréhendé dans ce modèle.*

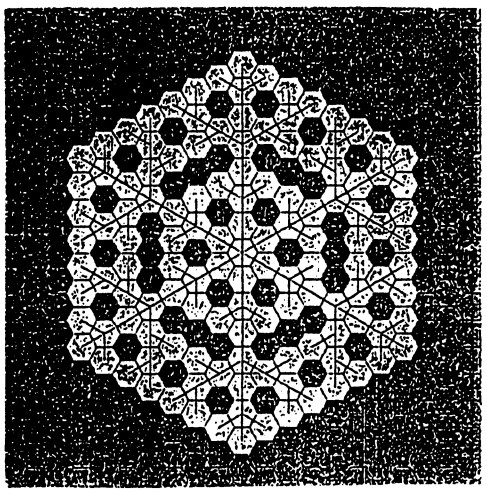


Fig. 5 : Etant donné un ensemble (S) de cellules situées aux noeuds d'une matrice parfaitement triangulaire, l'arbre d'Ulam construit à partir de n'importe quel cellule de la population est un flocon de neige parfait. La forme de l'arbre est caractéristique de l'environnement local de la cellule.

En résumé, la possibilité de comparer les résultats statistiques et topographiques est d'un grand intérêt pour étudier les liens entre l'état pathologique d'une cellule donnée (ou d'un groupe) et son environnement tissulaire.

4. RESULATS

4.1. Emergence d'une population cellulaire stable (homéostasie)

La première série d'expériences a pour but de tester et d'étudier la faculté de notre modèle à simuler une population cellulaire à l'état stable. Pour cela, nous avons initialisé une population de 1000 cellules comprenant 90% de cellules différenciée, 5% de cellules latentes et 5% de cellules proliférantes. L'évolution de la population, simulée pendant 30 000 heures (3.5 années), est illustrée par la Figure 6. La Figure 6a montre que, durant une première période d'ajustement, le nombre total de cellules augmente et diminue de façon importante. Cette phase d'ajustement, (environ 1000 heures), résulte d'une part de la progression des cellules naissantes vers l'état "latent" et d'autre part de la transition de l'état "latent" vers l'état différencié, alors que dans le même temps le nombre de cellules proliférantes diminue. Le modèle auto-régule la proportion de cellules dans les trois compartiments selon les durées de vie maximum fixées à l'initialisation. Si la durée de vie des cellules différenciées était fixée à 2000 heures, (contre 1000 heures pour la figure 6a), la phase d'ajustement durerait 2000 heures.

Après cette phase d'ajustement les proportions de cellules dans les trois compartiments (proliférantes, latentes, différenciées) se stabilisent, ainsi que le nombre total de cellules. Les proportions de cellules proliférantes se stabilisent à 1% pour les cellules proliférantes, 14% pour les cellules latentes et 85 % pour les cellules différenciées. Notons que si les paramètres initiaux sont trop loin de cet équilibre, le modèle devient incapable de s'auto-réguler et la population cellulaire devient indifférenciée et croît exponentiellement.

Les proportions relatives des facteurs GF et DF disponibles dans le milieu sont montrées dans la Figure 6b. La proportion de GF est légèrement supérieure à celle de DF tout au long de la simulation. Néanmoins, certaines variations brusques apparaissent de temps en temps (au maximum 15% d'amplitude), inhérentes à la régulation du système.

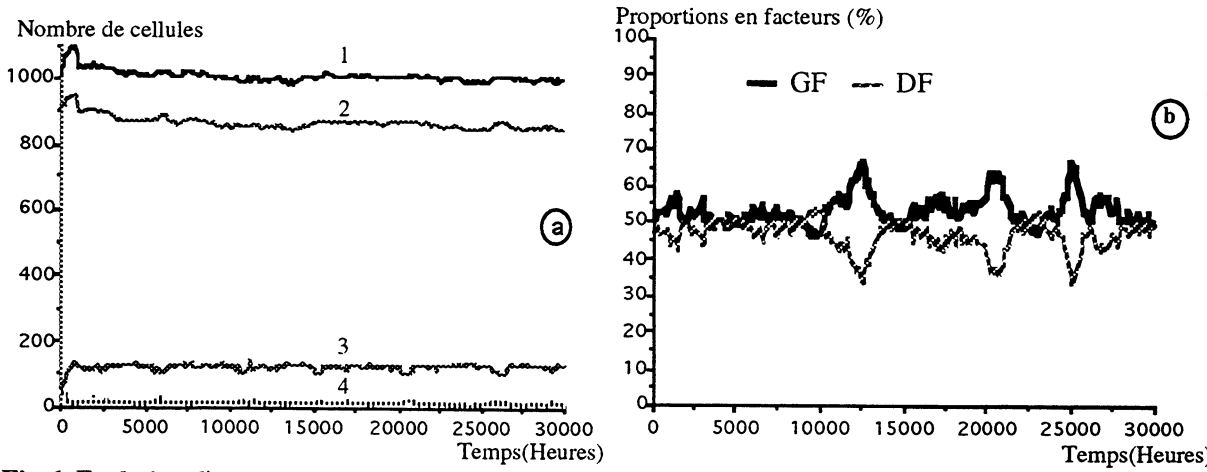


Fig 6. Evolution d'une population normale en fonction du temps. **a)** Evolution du nombre total de cellules (1), du nombre de cellules différenciées (2), du nombre de cellules latentes (3) et du nombre de cellules proliférantes (4). **b)** Evolution de la proportion de GF et de DF en fonction du temps.

Concernant l'émergence structurale, quelles que soient les conditions topographiques initiales, après un certain temps, une population normale de cellules convergera vers une population bien ordonnée topographiquement (Fig.8, gauche).

4.2 Emergence d'une tumeur.

La Figure 7 montre l'évolution d'une population cellulaire tumorale. Une fois l'état stable atteint (flèche), une probabilité de mutation *Padv* (0.01) affecte toutes les cellules passant en phase S. La population augmente alors de façon exponentielle. La concentration en GF et DF devient de plus en plus chaotique. Juste après 1000 heures, la concentration en GF devient trop forte par rapport à celle en DF; toutes les cellules, y compris certaines cellules DF+, partent alors en cycle, et le nombre de cellules augmente considérablement. La vitesse de cette phase néoplasique dépend de la valeur de Padv : plus grande est cette probabilité, plus rapide est l'émergence d'un néoplasme.

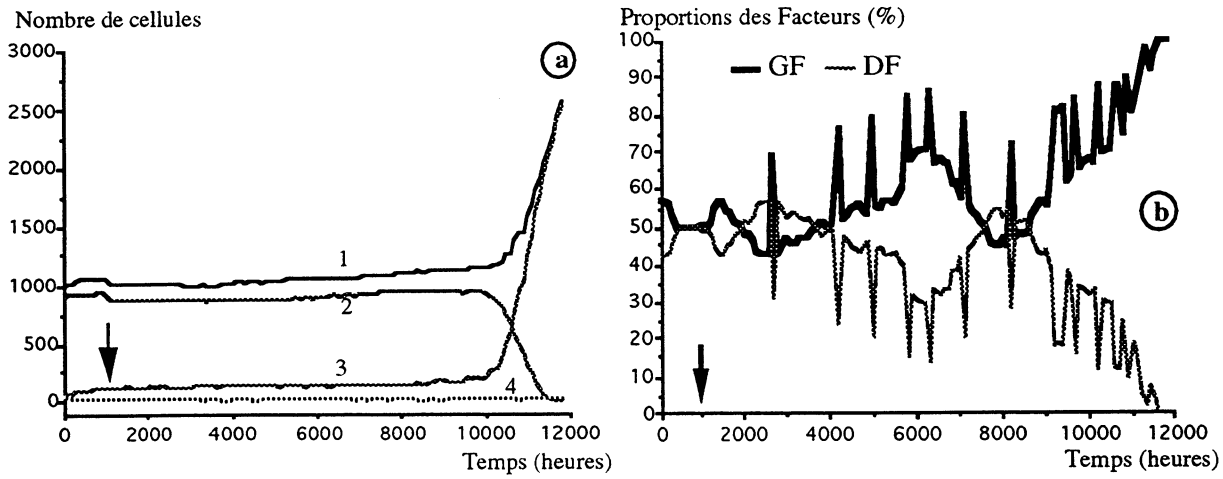


Fig. 7. Évolution du nombre de cellules **(a)** de toute la population (1), du nombre de cellules différenciées (2), du nombre de cellules latentes (3) et du nombre de cellules proliférantes (4) et des proportions en DF et GF **(b)** en fonction du temps. Les flèches indique le début des mutations, avec Padv = 0.01.

Concernant les aspects topographiques, la forme finale du clone tumoral dépend de plusieurs facteurs comme le temps de génération, les ressources de la cellule, l'adhérence des cellules voisines, etc.... La tumeur va s'accroître (ou va dégénérer) pour aboutir à l'émergence d'une structure possédant une topographie totalement différente (Fig. 8, droite).

5. CONCLUSIONS

La majorité des modées de prolifération cellulaire (Calderon, 91) abordent ce problème de façon globale, par l'établissement d'équations différentielles, perdant ainsi l'aspect de

localité et d'individualité, qui est primordial dans notre application. En effet, certaines hypothèses ou observations cliniques ont mis en effet en évidence la possibilité d'une origine monoclonale des cancers, c.a.d. le développement de tumeurs à partir d'une seule cellule ou de plusieurs cellules isolées à différents endroits de la tumeur. L'étude de l'apparition et du développement ou non d'une tumeur dans un tissu sain impose donc une approche individuelle et locale des composants du tissu. De plus, nous savons qu'une cellule cancéreuse n'induit pas forcément l'émergence d'une tumeur. C'est la capacité de son environnement (inertie structurelle et fonctionnelle) à supporter et à réagir à certaines modifications locales (hormonales, mécaniques, nutritionnelles, etc...) qui dictera la réponse, i.e. prolifération ou dégénérescence des cellules cancéreuses. La modélisation par des agents autonomes entre eux mais reliés par leur environnement s'avère particulièrement adéquate pour tenir compte à la fois de l'aspect local et de l'aspect social de l'homéostasie tissulaire.

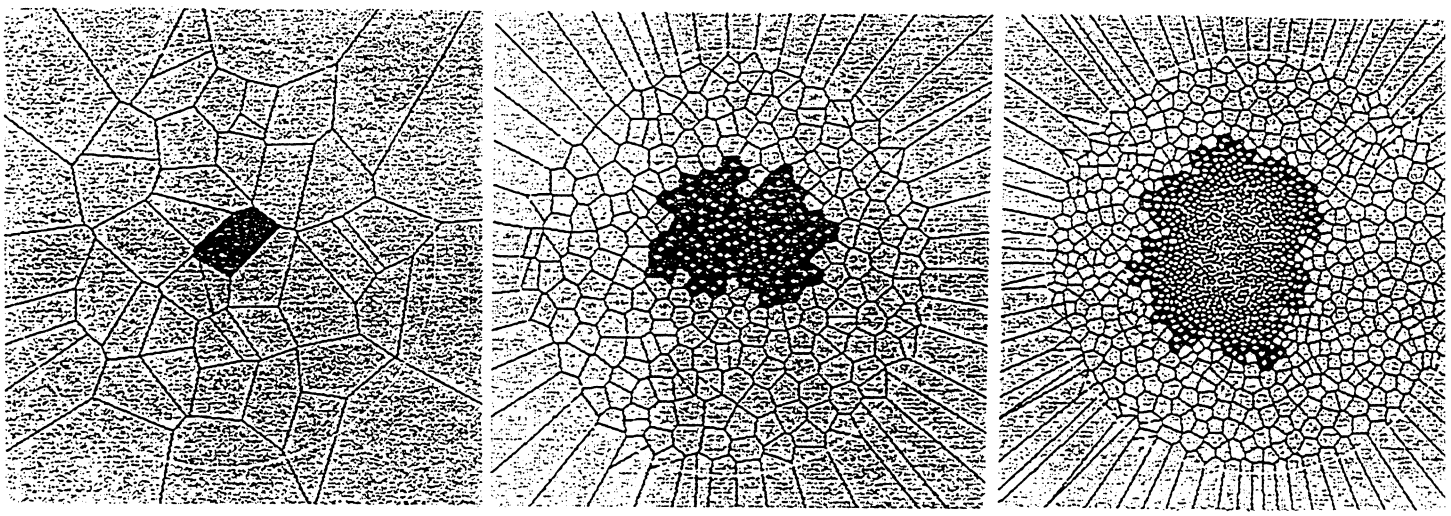


Fig. 8. Evolution d'une population (polygone gris) qui contient une cellule mutée (polygone noir). De la gauche ver la droite, le temps réel a été augmenté de plusieurs temps de génération de cellules normales. Le temps de cycle des cellules cancéreuses étant plus faible que celui des cellules normales, le clone tumoral croît plus vite que les clones normaux.

6. BIBLIOGRAPHIE

Baujard O., Pesty S. and Garbay C.
A Programming Environment for Distributed Applications Design in Artificial Intelligence. *Applications of Artificial Intelligence X : Knowledge-Based Systems,* pp 110-116. SPIE The International Society for Optical Engineering, Orlando 92,1992.

Marcelpoil R. and Usson Y.
Methods for the study of cellular sociology: Voronoï diagrams and parametrization of the spatial relationships. *J. theor. Biol.* 154:359-369. ,1992a

Guillaud M. and Brugal G.
Computer Model for the Emergence of Neoplasia in Growing Cell Populations. Part I : Simulation of Growth Parameters (soumis), 1993.

Weaire D, Rivier N
Soap, Cells, and Statistics - Random patterns in Two Dimensions.

Contemp. Phys. 25:59-99, 1984

Kayser K, Sandau K, Böhm G, Dietmar Kunze K, Paul J
Analysis of soft tissue tumors by an attributed minimum spanning tree.
AQCH 13:329-334,1991

Ahuja N
Dot Pattern Processing Using Voronoi Neighborhoods.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 4:336-343,1992

Zahn CT
Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters.
EEE Trans Comput 20:68-86,1971

Dussert C., Rasigni M., Palmari J., Rasigni G., Llebaria A. & Marty A.
Minimal spanning tree analysis of biological structures.
J Theor Biol 125:317-323, 1987.

Lorz U : Cell-area Distributions of Planar Sections of spatial Voronoï Mosaics.
Materials Characterization. 25:297-309,1990

Venema HW
Determination of nearest neighbors in muscle fiber patterns using a generalized version of the Dirichlet tesselation.
Pattern Recognition Letters. 12:445-449,1991

Pollak M., Boyarsky P. and Gora P.
A mathematical model describing consequences of abnormality high levels of epidermal growth factor receptor on the proliferation of neoplastic cells.
Cancer Investigation, 9,5, 513-520,1991

Foulds L.
Neoplastic Development, Vol. 2, New York: Academic press. ,1975

Langton C.G.
Artificial Life.
Artificial Life, SFI Studies in the Sciences of Complexity, Ed. C. Langton, Addison-Wesley Publishing Company,1988

Steels L.
Cooperation between distributed agents through self-organisation.
Decentralized A.I. pp 175-196. Y. Demazeau & J.P. Müller Eds. Elsevier Science Publishers,1990

Brooks R.A.
A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation.* RA-2, April,14 23,1986

Drogoul A., Corbara B. & Fresneau D.
Applying EthoModelling to social organization in ants.
Biology and Evolution of Social Insects, pp. 375-383, Billen Ed., Leuven University Press, Begium,1992

Maruichi T., Ichikawa M. & Tokoro M..
Modeling Autonomous Agents and their Groups.
Decentralized A.I. pp 215-249. Demazeau & Müller Eds. Elsevier Science Publishers. ,1990

Wavish P.
Exploiting Emergent Behaviour in Multi-Agent Systems.
Decentralized A.I. 3. Werner & Demazeau Eds. Elsevier Science Publishers. ,in press.

Kephart JO., Hogg T. & Huberman B.A.
Dynamics of Computational Ecosystems : Implications for DAI.
Distributed Artificial Intelligence. Vol 2, pp 79-95. Morgan Kaufmann Publishers,1989.

Zheng Z., Polakowska R., Johnson A. and Golsmith L.A.
Transcriptional control of epidermal growth factor receptor by retinoic acid.
Cell growth and differentiation, Vol 3, 225-232, April 1992

Strickland S., Smith K. and Marotti K.
Hormonal inductionn of differentiation in teratocarcinoma stem cells : generation of parietal endoderm by retinoic acid and dibutyryl cAMP.
Cell. 21 : 347 ;1980

Marcelpoil R, Bertin E. and Usson Y
Methods for the study of cellular sociology: Ulam trees on Voronoï diagrams and parametrization of local relationships.
J. theor. Biol. (soumis), 1992b

Calderon Calixto P.
Modeling tumor growth
Mathematical Biosciences 103: 97-114, 1991.

REFERENCES BIBLIOGRAPHIQUES

Al nachawati I.
Processus de classification séquentiels non arborescents pour l'aide au diagnostic.
Thèse d'université, Université de Grenoble 1, 1985.

Altman D.G and Andersen PK.
Statistics in medicine ; 8 : 771-783, 1989.

Atkin NB., Richards BM.
Deoxyribonucleic acid in human tumours as measured by spectrophotometry of Feulgen stain : a comparison of tumours arising at different sites.
Br. J. Cancer ; 10 : 669-687, 1956

Atking NB.
Prognosis significance of modal DNA value and other factors in malignant tumours, nased on 1465 cases
Br. J. Cancer ; 40 : 210-215, 1979

Atkin NB.
Cytophotmetric DNA determination correlatedto kariotype, particulary in cancer.
Analyt Quant Cytol Histol ; 9 : 96-104, 1987

Atking NB.
The clinical usefulness of determining ploidy patterns in human tumors as measured by slide-based feulgen microspectrophotometry
Analyt Quant Cytol Histol ; 13 : 75-79, 1991.

Auer G.U., Caspersson T.O. and Wallgren A.S.
DNA Content and Survival in Mammary Carcinoma.
Analyt Quant Cytol Histol ; 2 : 161-165, 1980.

Auer G.U., Eriksson E. and Azavedo E, et al.
Prognostic significance of nuclearDNA content in mammary adenocarcinomas in humans.
Cancer Res.; 44 : 393 -390, 1984

Aufferman W., Böcking A.
Early detection of precancerous lesions in dysplasias of the lung by rapid DNA image cytometry
Analyt Quant Cytol Histol ; 7 : 218-226, 1987

Ault KA
Detection of small numbers of monoclonal B lymphocytes in the blood of patients with lymphoma.
N. Engl. J. Med ; 301 : 924-928, 1979

Baak JPA.
Manual of Quantitative Pathology in cancer diagnosis and prognosis
Springer-Verlag Berlin Heidelberg, 1991

Banks D.L.
Smoothing the bayesian Bootstrap.
Technical Report 415, Departement of statistics, Carnegie Mellon University, Piitsburgh, Pennsylvannia, 1988 (a).

Banks D.L.
Exact comparison of bootstraps methods.
Technical Report 427, Departement of Statistics, Carnegie Mellon University, Piitsburgh, Pennsylvannia, 1988 (b).

Bartels P.H.
Numerical evaluation of cytological data. I. Description of profil
Analyt Quant Cytol Histol ; 1: 20-28, 1979 (a)

Bartels P.H.
Numerical evaluation of cytological data. II. Comparison of profiles
Analyt Quant Cytol Histol ; 1: 77-83, 1979 (b)

Bartels P.H, Weber J.E and Bibbo M.
Ploidy Pattern Analysis, Statistical Considerations.
Analyt Quant Cytol Histol ;7, 126-130, 1985.

Bartels P.H., Thompson D. and Weber J.E.
Expert systems in histopathology. IV. The management of Uncertainty
Analyt Quant Cytol Histol ; 14 : 1-13, 1992 (a)

Bartels P.H., Thompson D. and Weber J.E.
Expert systems in histopathology. V. DS theory, certainty factors and Possibility theory
Analyt Quant Cytol Histol ; 14 : 165-174, 1992 (b)

Bartels PH., Thompson D., Bibbo M. and Weber JE.
Bayesian belief network in quantitative Pathology.
Analyt Quant Cytol Histol ; 14: 459-473, 1992 (c)

Beerman H., Smith V.T.H.B.M, Kluin PM., Bonsing B.A., Hermans J. and Cornelisses C.J.
Flow cytometric analysis of DNA stemline heterogeneity in primary and metastatic breast cancer.
Cytometry ; 12: 147-154, 1991.

Bengtsson E.
The measuring of cell features.
Analyt Quant Cytol Histol ; 9 : 212-217, 1987

Beran R. and Srivastava M.S.
Bootstrap tests and confidence regions for functions of a Covariance matrix.
The Annals of Statistics ; 13 : 95-115, 1985.

Berthier P et Bouroche J.M.
Analyse des données multi-dimensionnelles.
Système décision, PUF, 164 - 174, 1975.

Besse P.
Stabilité de l'Analyse en Composantes Principales par ré-échantillonage, Approximation par la théorie des pertubations

Publication du laboratoire de statistique et probabilité. ISSN 0248-3289, Université Paul Sabatier. Toulouse, 1989

Bhattacharya P.K, Bartels P.H, Bahr G.F and Wied G.L.
A test statistic for detecting the presence of abnormal cells in a sample.
Acta Cytol.; 15 : 533-544, 1971 .

Bhattacharya P.K, Bartels P.H, Taylor J and Wied G.L.
A decision procedure for automated cytology : Test statistic for detecting sample abnormality, and inadequacy.
Acta Cytol. ; 17: 538-548, 1973.

Bibbo M., Bartels P.H, Dytch H.E and Wied G.L.
Ploidy measurements by High-Resolution Cytometry.
Analyt Quant Cytol Histol ; 7 : 81-88, 1985

Bièche I., Lidereau R.
Analyse moléculaire des tumeurs du sein : développements récents
Bull Cancer ; 79 : 1115-1133, 1992

Böcking A., Adler CP., Common HH., Hilgart M., Granzen B., Auffermann C.
Algorithm for a DNA cytophotometric diagnosis and grading of Malignancy
Analyt Quant Cytol Histol ; 6 : 1-8, 1984

Böcking A., Chatelain R., Homge M., Gilissen, Wohltmann D.
Representativity and reproductibility of DNA malignancy Grading in different carcinomas
Analyt Quant Cytol Histol ; 11, 81-86, 1989 (a)

Böcking A., Chatelain R., Bisterfield S., Noll E., Biesterfelf D., Wohltmann C., Goecke C.
DNA grading of malignancy in breast cancer. Prognostic validity, reproductibility and comparison with other classifications.
Analyt Quant Cytol Histol ; 11: 73-80, 1989 (b)

Brugal G.
Image analysis of microscopic preparations.
In: Jasmin G., Proschek L. (eds) Methods and achievements in experimental pathology. Karger, Montreal, 1-13, 1984

Brugal G., Quirion C., Vassilakos P.
Detection of bladder cancers using a SAMBA 20 cell image processor
Analyt Quant Cytol Histol ; 8 : 187-194, 1986

Brugal G.
Traitement et analyse d'images en biologie fondamentale et clinique.
Spectra biologie ; 5 : 27-35, 1987.

Calderon C.P. and Kwembe T.A.
Modeling tumor growth
Mathematical Biosciences ; 103 : 97-114, 1991.

Carey FA., Lamb D., Bird CC.
Intratumoral heterogenity of DNA content in Lung Cancer
Cancer ; 65 : 2226-2269, 1990

Caspersson T., Lomakka G., Casspersson O.
Quantitative cytochemical methods for the study of tumor cell populations
Biochem Pharmacol ; 4 : 113-127, 1960.

Charpin C, Martin P.M, Lissitzky J.C, Jacquemier J, Kopp F, Pourreau-Schneider N, Lavaut M.N et Toga M.
Estrogen receptors immunocytochemical assay (ER-ICA) and laminin detection in 130 breast carcinomas and computarized (SAMBA200) multiparametric quantitative analysis on tissue sections.
Bull. Cancer.; 73 : 651-664, 1986

Chassery JM, Montanvert A.
Géometrie discrète en Analyse d'images
Dans Traité des nouvelles technologies. Hermes, Paris, 1991

Chatelein R., Willms A., Biesterfeld S., Auffremann W., Böcking A.
Automated Feulgen staining with a temperature-controled staining machine.
Analyt Quant Cytol Histol.; 11: 211-217, 1989

Dantas ME, Brown JP., Thomas MR.,et al.
Detection of melanoma cells in bone marrow using monoclonal antibodies.
Cancer ; 52 : 627-632 , 1983

Davidson A.C, Hinkley D.V and Schechtman E.
Efficient Bootstrap simulation.
Biometrika ; 73 : 555 - 566, 1986.

Dawson AE., Austin RE. and Weinberg DS.
Nuclear grading of breast carcinoma by image analysis
American Journal of Clinical Pathology ; 95 : 4- 11, 1991

Diaconis P. and Efron B.
Computer intensive methods in statistic
Scientific American ; 248, 116-130, 1983.

Diday E.
Une nouvelle méthode en classification automatique et reconaissance des formes : La méthode des nuées dynamiques.
Revue de Statistiques Appliquées ; 19 :19-32,1971.

Diday E, Lemaire J, Pouget J et Testu F.
Eléments d'analyse de données.
Dunod. Bordas, Paris, 167-293, 1982.

Do K-A and Hall P.
On importance resampling for the Bootstrap
Biometrika ; 78, 161-167, 1991

Dubuisson B.
Diagnostic et reconnaissance des formes.
Hermes, Paris, 79-111, 1990.

Duijndam WAL, Smeulders AWM., Van Duijn P., Verweij AC.
Optical errors in scanning stage absorbance cytophotometry : I: Procedures for correcting apparent integrated absorbance values for distributional, glare and diffraction errors.
J. Histochem. Histochem ; 28 : 388-394, 1980

Dytch HE., Bartels PH., Bibbo M. Pishotta F. and Wied GL.
The rejection of noncellular artifcats in Papanicolaou-stained slide specimens by an automated high-resolution system : identification of important cytometric featues.
Analyt Quant Cytol Histol.; 5 : 241-249, 1983.

Dytch H.E. and Wied G.L.
Aryificial Neural networks and their use in quantitative pathology

Analyt Quant Cytol Histol ; 22 : 379-393, 1990

Efron B.
Computers and the theory of Statistics ; Thinkink the unthinkable.
Siam Review ; 21 : 460-480, 1979 (a).

Efron B.
Bootstrap methods : Another look at the jacknife.
The annals of statistics ; 7 : 1-26, 1979 (b).

Efron B.
The Jacknife, the Bootstrap and other Resampling Plans.
S.I.A.M. Philadelphia, Pennsylvania, 1982.

Efron B and Tibshirani R.
The Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy
Statistical Science ; 1 : 54-75, 1985.

Fages R.
La notion de dispersion en classification automatique.
Communication aux Journées de statistiques. Nice. 22-26 Mai 1978.

Falkmer UG.
Methodological Aspects on DNA cytometry
Image and flow cytometry on malignant tumors of the breast, the endotrium, the brain, and the salivary glands.
T-tryck, Stocholm, 1989

Falkmer UG.
Methodoligic sources of errors in image and flow cytometric assessments of the malignacy potential of prostatic carcinoma
Hum.Pathol ; 23 : 360-367, 1992.

Fallienus A.
DNA content and prongosis in Breast cancer.
T-tryck, Stocholm, 1986

Ferno M., Baldetorp B., Ewers S-B., Idvall I., Olsson H., Sigurdsson H. and Killander D.
One or multiple samplings for flow cytometric DNA analyses in Breast cancer- prognostic Implications.
Cytometry ; 13 : 241-249, 1992

Feulgen R., Rossenbeck H.
Mikroskopic-chemister Nachweiss einer Nucleinsäure wom yupus Thymus-nucleinsäure und die darauf beruhende selektive Färbung von Zekleinen in mikroskopischen Präparaten.
Z. Physiol Cem. ; 135 : 203-248, 1924

Foulds L.
Tumor progression
Cancer Research ; 17, 355-356, 1957

Galloway MM.
Texture analysis using grey level runlenghts.
Comput Graph. Im. Proc ; 4 : 172-174, 1975

Giaretti W.A, GAIS P., Jutting U., Rodenacker K. and Doermer P.
Correspondances between chromatin morphology as derived by digital image analysis and autoradiographic labelling pattern.

Analyt.Quant.Cytol.Histol.; 5 : 79 - 89, 1983.

Giroud F.
Cell nucleus pattern analysis : geometric and densitometric featuring, automatic cell phase identification.
Biol. Cell. ;44 : 177 - 188, 1982.

Giroud F., Gauvain C., Seigneurin D. and Von Hagen V.
Chromatin Texture changes related to proliferation and maturation in erythrocytes
Cytometry ; 9 : 339-348, 1988

Giroud F., Guillaud M., Chancel G., Montmasson MP.
How instrumentation and fixation can affect the reproductibility and the accuracy of densitometric mesurements ?
Anal.Cell Pathol ; 4 : 160, 1992

Goldstein DJ.
Errors in microdensitometry
Histochem. J.; 13 : 251-267, 1981

Goss G., Petras RE., Perkins A., Miller M.
Effects of refixation and reprocessingon the quality of slides prepared from paraffin embedded tissues.
The journal of Histotechnology; 15 : 43-47, 1992.

Guilbault G.G.
Practical fluoreccence: Theory, Methods and Techniques
Marcel Deller, Inc.New-York , 1973

Harden DG., Taylor AMR.
Chromosomes and neoplasia.
In : Harris H., Kirshhhorn K., eds. Advances in human genetics. London: Plenum Press ; 1-69, 1979

Haroske G., Bergander St, Konig R., Meyer W.
Application of malignancy-associated changes of the cervical epithelium in a hierarchic classification concept.
Anal.Cell Pathol. ; 2 : 189-198, 1990.

Heathfiel H., Bose D., Kirkham N.
A computer-Based decision support system for diagnostic histopathology of the breast
Path. Res. Pract. ; 188 : 418-424, 1992.

Hecquet B.
Intérêt des modèles mathématiques pour la description de la croissance et le traitement des tumeurs mammaires
VII ièmes Journées de la Société francaise de Sénologie et de Pathologie Mammaire.
Le Touquet, 3-5 Octobre 1985

Hedley DW.
Flow cytometry using parrafin-embedded tissue : five years on.
Cytometry ; 10 : 229-241, 1989.

Hinkley D.V.
Bootstrap Methods
J.R. Statist. Soc. B ; 50 : 321-337, 1988.

Hiraoka Y., Sedat JW., Agard D.A.
The use of a charge-coupled device for quantitative optical microscopy of biological structures
Sciences ; 238 : 36-41, 1987

Holmes-Junca S.
Outils informatiques pour l'évaluation de la pertinence d'un résultat en analyse des données.
Thèse de 3ème Cycle en Mathématique. Université des sciences et techniques du Languedoc.
1985, 16 - 21.

Hosking J.R.M.
The theory of probability weighted moments.
Research report RC12210. IBM Research, Yorktown Heights, 1986

Hosking J.R.M.
Some theoretical results concerning L-moments.
Research report RC14492. IBM Research, Yorktown Heights, 1989

Hosking J.R.M.
L-moments : Analysis and estimation of distributions using Linear combinations of order statistics.
J.R. Statist. Soc. B; 52 : 105-124, 1990.

Hubert P.
Monotone invariant ckustering procedures.
Psychometrika ; 38 : 1-15, 1973.

Humbert C. and Usson Y.
Eucaryotic DNA replication is a topographically ordered process
Cytometry; 13 : 603-614, 1992.

Jambu M.
Classification automatique pour l'analyse des données : 1.méthodes et algorithmes.
Dunod. Bordas, Paris, 1978.

Johns MV.
Importance resampling for Bootstrap confidence intervals.
J. Am. Statist. Assoc.; 83 : 709-714, 1988.

Kiss R., Gasperin P., Verhest A. and Pasteels J.-L.
Modification of tumor ploidy level via the choice of tissue taken as diploid reference in the digital image analysis of feulgen-staine nuclei.
Modern Pathology ; 5 : 655-660,1992.

Laerum O.D., Farsun T.
Clinical application of flow cytometry : A review.
Cytometry; 2 : 1-13, 1981.

Langley FA., Bucley CH., Tasker M.
The use of ROC curves in histocytopatologic decision making
Analyt Quant Cytol Histol ; 7 : 167-173, 1985

Landeweerd GH., Gelsema E.S.
The use of nuclear texture parameters in the automatic analysis of leucocytes
Pattern Recognition ; 10 : 57-61, 1981

Lansing Taylor D., Waggoner A.S, Murphy R.F., Lanni F. and Birge R.R.
Application of fluorescence in the biomedical Science.
Liss, New York, 1986

Lavia LA.
Video Microscopy : System calibration for densitometric analysis
BioTechniques ; 7 , 1018-1025, 1989

Lebart L, Morineau A et Fénelon JP

Traitement des données statistiques:méthodes et programmes.
Dunod, Bordas, Paris, 273 - 353, 1979.

Léger I., Giroud F., Brugal G.
Quantitative analysis of cytosketal proteines through the cell cycle of a MRC-5 fibroblastic cell lines
Analyt Quant Cytol Histol ; 12 : 321-326, 1990

Leuchtenberger C., Leuchtenberger R, Davis AM.
A microspectrophotometric study of the desoxyribose nucleic acid (DNA) content in cells of normal and malignat human tissues
Am. J. Pathol. ; 30 : 65-85, 1954.

Leuchtenb-, erger C., Leuchtenberger R.
Quantitative cytochemical studies on the relation of deoxyribonucleic acid of cells to various pathological conditions.
Biochem Pharm ; 4 : 128-163, 1960

Lo A.Y.
A large sample study of the Bayesian Bootstrap.
The Annals of Statistics.; 15 : 360 - 375, 1987.

Lo A.Y.
A Bayesian Bootstrap for a finite population.
The Annals of Statistics ; 16 : 1684 - 1695, 1988.

MacAulay C., Palcic B .
A comparison of some quick and simple image threshold selection methods for stained cells.
Analyt Quant Cytol Histol ; 10 : 134-138, 1988.

MacAulay C., Palcic B.
An edge Relocation segmentation algorithm
Analyt Quant Cytol Histol ; 3 : 165-171, 1990 (a)

MacAulay C., Palcic B.
Fractal texture features based on optical density surface area : Use in image analysis of cervical cells.
Analyt Quant Cytol Histol ; 12, 394-398, 1990 (b)

MacEachron DL., Gallistel C.R., Tretiak O.J.
Issues in quantitative imaging.
In :Three-dimensionnal neuro-imaging. A.W. Toga ed. Raven Press, Ltd., New York, p.39-71, 1990.

Mahalanobis PC.
Historical note on the D^3-statistic
Sankhya ; 9 : 237-248, 1948

Manly B.F.J.
Multivariate Statistical Methods : A Primer.
Chapman and Hall, New York, 1989

Mathies A., Stryer L.
Single-Molecule Fluorescence Detection : A feasibility study using Phycoerythrin
In: Applications of Fluorescence in Biomedical Science. Eds Taylor DL., Waggoner AS., Murphy RF., Lanni F and Birge RR., Alan R. Liss Inc., pp 129-140

Mellin W.
Cytophotometry in Tumor Pathology : a critical review of methods and applications, and some results of DNA analysis

Path Res Pract. ; 186 : 37-62, 1990

Mesker W.E., Eysackers M.J., Ouwerkerk-van Velzen M.C.M., van Driel-Kulker A.M.J. and Ploem J.S.
Discrepancies in ploidy determination due to specimen sampling errors
Analyt Quant Cytol Histol ; 1: 87-95, 1989

Metezeau P.H., Ronot X., Le Noan-Merdrignac G., Ratineau M.H.
La cytométrie en flux.
Volume 1. MEDSIMc Graw-Hill, 1988.

Meyer J.S., Wittliff JL.
Regional heterogenity in breast carcinoma : thymidine labelling index, steroid hormone receptors, DNA ploidy
Int. J. Cancer ; 47 : 213-220, 1991

Miller R.G.
The Jacknife : a Review.
Biometrika.; 61 : 1-15, 1974.

Mitelman F., Levan G., Nilsson PG. and Brandt L.
Non-random karyotypic evolution in chronic myeloid leukemia.
Int. J. Cancer; 18 : 24-30, 1976

Muller C. et Radoui M.
Reseaux neuromimétiques et analyse des données.
XXIIèmes Journées de Statistiques (ASU), Tours, 1990

Nafe R. and Choriz H.
Introduction of a neuromal network as a tool for diagnostic analysis and classification based on experimental pathologic data
Exp. Toxic. Pathol.; 44 : 17-24, 1992

Nowell P.C.
Mechanisms of tumor Progression
Cancer Research ; 46 : 2203-2207, 1986

Ojima Y., Inui N., Makino S.
Cytochemical studies on tumor cells. V. Measurement of desoxyribonucleic acid (DNA) by feulgen-microspectrophotometry in some human uterine tumors.
GANN ; 51 : 371-376, 1960

Oliver LH., Populsen RS., Toussaint GT.
Estimating false positive and false negative error rates in cervical cell classification
J. Histochem. Cytochem ; 25 : 696-701, 1977.

Ono J., Auer G.
The significance of DNA measurements for the early detection of bronchial cell atypia.
Cytometry ; 3 : 340-344, 1983

Opfermann M., Brugal G., Vassilakos P.
Cytometry of breast carcinoma : Significance of Ploidie Balance and Proliferation index.
Cytometry ; 8 : 217-224, 1987.

Pennes DR., Naylor B., Rebner M.
Fine needle aspiration biopsy of the breast. Influence of the number of passes and the sample size on the diagnostic yield.
Acta Oncologica ; 34 : 673-676, 1990

Piller H.

Microscope photometry.
Springer-Verlag Berlin, Heidelberg, 1987.

Ploem JS., Driel-Kulker AMF.,van Goyarts-Veldtra L. et al.
Image analysis combined with quantitative cytometry results and instrumental developments for cancer diagnosis.
Histochemistry ; 84 : 549-555, 1986.

Pollak M., Boyarsky A. and Gora P.
A mathematical model describing consequences of abnormally high levels of epidermial receptor on the proliferation of neoplastic cells
Cancer investigation; 9, 513-520, 1991.

Pressman NJ.
Markovian analysis of cervical cell image.
J. Histochem. Cytochem ; 24 : 138-144, 1976

Preston K.Jr and Bartels P.H.
Automated Image Prosessing for Cells and Tissue. Progress in Medical Imaging,
Vernon L. Newhouse, Springer-Verlag New York, Inc. 1988.

Rosenfeld A., Davis LS.
Iterative histogram modification.
IEEE Trans Syst Man Cybern ; SMC-8 : 300-302, 1978.

Rosenthal D.L., Suffin SC., Missirlian N., McLatchie C., Castleman K.R.
Techniques in the preparation of a monolayer of gynecologic cells for automated cytology : an overview
Analyt Quant Cytol Histol ; 9 : 55-59, 1987.

Rubin DB.
The bayesian Bootstrap
Annals of Statistics; 9 : 130-134, 1981.

Salmon I., Coibion M., Larsimont D., Badr-El-Din A., Verhest A., Pasteels, J-L, Kiss R.
Comparison of fine needle aspirates to breast cancers to imprints smears by means of digital cell image analysis.
Analyt Quant Cytol Histol ; 13 : 193-200, 1991.

Sanchez L., Regh M., Chatelain R., Böcking A.
Performance of a TV image analysis system as a microdensitometer .
Anal. Quant. Cytol. histol ; 12, 279-284, 1989.

Sandritter W., Carl M., Ritter W.
Cytophotometric measurements of the DNA content of human malignant tumors by means of the Feulgen reaction
Acta Cytol ; 10 : 26-30, 1966.

Santisteban M.-S, Montmasson M.-P, Giroud F., Ronot X. and Brugal G.
Fluoresecence image cytometry of nuclear DNA content versus chromatine pattern : A comparative study of ten fluorochromes.
The Journal Of Histochem and Cytochem ; 40 : 1789-1797, 1992

Sasaki K., Hamano K., Kinjo M. and Hara S.
Intratumoral heterogeneity in DNA ploidy of bladder carcinomas
Oncology ; 49: 219-222,1992

Schaberg ES., Jordan WH., Kuyatt BL
Artificial Intelligence in automated classification of rat vaginal smear cells.

Analyt Quant Cytol Histol ; 6 : 446-450, 1992

Schulte EK., Wittekind DH.
Standardization of the Feulgen reaction : the influence of chromatin condensation of the kinetics of acid hydrolysis.
Anal. Cell. Pathol.; 2 , 149-157, 1990

Schwartz G., Schwartz M., Schenk U.
Effect of the spectral properties of monolayer cell preparations for automated cervical cytology on visual evaluation and classification, with an estimation of the number of cells required to be screened.
Analyt Quant Cytol Histol ; 5 : 189-193, 1983

Seigneurin D., Gauvain C. Brugal G.
A quantitative analysis of the human Bone Marrow cragulocytic cell lineage using the SAMBA 200 cell image processor. 1. The normal maturation sequence
Analyt Quant Cytol Histol ; 6 : 168-178, 1984.

Shackney S.E., Smith C.A, Miller B.W., Burholt D.R., Murtha K., Giles H.R., Ketterer D.M. and Pollice A.A.
Model for the genetic evolution of human solid tumors.
Cancer Research ; 49 : 3344-3354, 1989.

Shapiro H.M.
Multistation multiparameter flow cytometry : a critical review and rationale.
Cytometry ; 3 : 227-278, 1983

Silverman B.W and Young G.A.
The bootstrap : To smooth or not to smooth ?
Biometrika.; 74 : 469 - 479, 1987.

Spyratos F., Briffod M., Gentile A., Brunet M., Brault C., Desplaces A.
Flow cytometric study of DNA distribution in cytopunctures of benign and malignant breast lesions
Analyt Quant Cytol Histol ; 9 : 485-494, 1987.

Stenkvist B. and Strande G.
Entropy as an algorithm for the statistical description of DNA cytometric data obtained by image analysis microscopy
Anal.Cell Pathol ; 2 :159-165, 1990

Sugihara H., Hattori T., Fujita S., Hirose K. and Fkuda M.
Regional ploidy variations in signet ring cell carcinomas of the stomach.
Cancer ; 65 :122-129, 1990

Sychra JJ., Bartels PH., Bibbo M., Taylor J., Wied GL.
Computer recognition of binucleation with overlapping in epithelial cells
Acta Cytol ; 22 : 22-28, 1978

Tanaka N., Ikeda H., Ueno T et al.
Field test and experimental use of CYBEST Model 2 for practical gynecologic mass screening.
Analyt Quant Cytol Histol ; 1 : 122-126, 1979

Tanke HJ., Rothbarth PH., vossen JMJ et al
Flow cytometry of reticulocytes applied to clinical hematology
Blood ; 61 : 1091-1097, 1983

Tucker JH., Eason P., Stark. M.
Ellipse test for a reduction of false positive signals in automated cytology.
Acta Cytol.; 22 : 370-376, 1979.

Uckun S.
Model based Reasoning in biomedicine
Critical Reviews in Biomedical Engineering ; 19 : 261-292, 1992.

van Driel-kulker A.M
Application de l'analyse quantitative des images au cyto-diagnostic des cancers du col de l'utérus.
Thèse de l'Université de Grenoble I. 1986.

van Ginneken A.M. and Smeulders A.W.M.
Reasoning in uncertainties. An analysis of five strategies and their suitability in pathology
Analyt Quant Cytol Histol ; 13 : 93-109, 1991

Visser JWM., Marten ACM., Hagenbeek A.
Detection of minimal residual disease in acute leukemia by flow cytometry.
Ann. NY Acad Sci ; 468 : 268-275, 1986

Waggoner A.S..
Fluorescent probesfor analysis of cell structure, function and health by flow imaging cytometry.
In : Applications of Fluorescence in Biomedical Science. Eds Teylor D.L., Waggoner A.S., Murphy R.F., Lanni F. and Birge R.R., Alan R. Miss, pp 3-28. 1986

Weber J.E. and Nielsen T.
Multinomial distribution in ploidy analysis
Analyt Quant Cytol Histol ; 7 : 140-151, 1985 (a)

Weber J.E, Baldessari B.A and Bartels P.H.
Test Statistics for Detecting Aneuploidy and Hyperploidy.
Analyt Quant Cytol Histol ; 7 : 131-139, 1985 (b).

Weber J.F., Bartels P.H., Bartels H.C. and Bibbo M.
Discrimination of DNA ploidy patterns by order statistics.
Analyt Quant Cytol Histol ; 9 : 60-68, 1987.

Wersto R.P., Robert L., Liblit MA., Koss L.G.
Flow cytometry DNA analysis of human solid tumors: A review of the interpretation of DNA histograms
Human Pathology ; 22 : 1085-1098, 1991

Wersto R.P., Liblit R.L., Deitch D., Koss L.G.
Variability in DNA measurements in multiple tumor smples of human colonic carcinoma
Cancer ; 67 : 106-115, 1991

Wheeler N., Suffin SC., Hall TL., Rosenthal DL.
Prediction of cervical neoplasia diagnosis groups: discriminant analysis on digitized cell images.
Analyt Quant Cytol Histol ; 29 : 169-181, 1987.

Wied GL, Bartels PH., Bibbo M., Dytch HE.
Image analysis and quantitative cyto and histopathology.
Techn. Report, N°2, The International academy of cytology, 1988

Wied G.L., Dytch H., Bibbo M., Bartels P.H. and Thompson D.
Artificial intelligence-guided analysis of cytologic data
Analyt Quant Cytol Histol ; 12 : 417-428, 1990

Williams AJ, Santiago S., Lehrman S., Popper R.
Transcutaneous needle aspiration of solitary pulmonary masses: How many passes ?
Am Rev Respir Dis; 136 : 452-454, 1987

Wittekind D.
Standardization of Dye and Stains for automated cell pattern recogniton
Analyt Quant Cytol Histol ; 7 : 6-30, 1985

Zetterberg A., Esposti P.
cytophotometric DNA-analysisof aspirated cells from prostatic carcinoma.
Acta Cytol ; 20 : 46-57, 1976.

MODELISATION ET CONTROLE STATISTIQUE DE L'ANALYSE CYTOMETRIQUE DE LA PLOIDIE EN CANCEROLOGIE

Résumé : L'objectif de cette thèse a été de concevoir deux applications destinées à améliorer la qualité et l'interprétation des analyses cytologiques par cytométrie à balayage. Ces deux applications correspondent à deux approches différentes ; *une approche méthodologique* dont le but était le développement d'un algorithme de reconnaissance de formes dynamiques, évoluant avec la taille de l'échantillon analysé ; *une approche analytique* dont le but était la conception d'un modèle de croissance tumorale en vue d'une meilleure compréhension des effets de l'échantillonnage sur l'interprétation des examens cytologiques tumoraux.
La première application concerne en particulier le contrôle statistique de l'analyse de l'ADN. Le contenu en ADN d'un échantillon tumoral est représenté sous forme d'histogramme. Ces histogrammes d'ADN sont codés par les L-moments, combinaisons linéaires de statistiques d'ordre supérieur. La méthode du bootstrap a été utilisée pour mesurer la variance de ces L-moments pour un échantillon donné. L'évolution des variances Bootstrap des L-moments en fonction de la taille de l'échantillon a permis de détecter à la fois la stabilité des histogrammes d'ADN et l'apparition d'individus nouveaux, en cours d'analyse. Cet algorithme de reconnaissance de formes dynamiques a été aussi appliqué au contrôle statistique de l'acquisition de données multi-dimensionnelles, en suivant la stabilité des valeurs propres d'une analyse en composantes principales effectuée sur des paramètres densitométriques et texturaux.
La deuxième application a abouti à l'élaboration d'une plate-forme de modélisation de l'émergence et de la croissance d'une tumeur dans un tissu sain différencié. Cette modélisation biologique repose sur la compétition entre hormones de croissance et hormones de différenciation sur les cellules souches, compétition assurant la stabilité (homéostasie) ou l'instabilité (cancérisation) du tissu. Contrairement aux modèles déterministes classiques, seuls les comportements et les caractéristiques individuels des cellules sont définis ; les simulations ont montré que le comportement de la population cellulaire était émergent. Cette modélisation cellulaire permet d'étudier la majorité des problèmes rencontrés en cytométrie à balayage, concernant le diagnostic et le pronostic effectués à partir d'échantillons tumoraux: représentativité de l'échantillon, taille de l'échantillon, erreurs instrumentales, interprétation statistique, fiabilité de la décision médicale, etc...

MODELISATION AND CONTROL OF CYTOMETRIC PLOIDY ANALYSIS IN CANCEROLOGY

Abstract : The purpose of this study was to use two applications to improve quality and interpretation of image cytometry analysis. These aplications correspond to two different approaches ; a*methodological approach* to develop a dynamic recognition algorithm of patterns changing with the sample size, and an*analytical approach* to develop a computer model of a growing tumor to study several interpretation problems which arise in the cytometric analysis of human tumors.
The first application concerns statistical control of DNA analysis. The DNA content of tumor sample is usually represented by a histogram. We showed that these DNA histograms can be coded by the L-moments, linear combinations of higher order statistics. The bootstrap method has been used to estimate the variance of the L-moments for a given sample. The evolution of bootstrap variances as function of the sample size permitted an evaluation of the stability of DNA histograms and to detect the advent of new type of cells, as sample size increases. This algorithm has also been applied to the statistical control of multivariate cytometric data acquisition, by assessing the stability of eigen values issu from the Principal Components Analysis of textural and densitometric features of biological samples.
The second application led to the conception of a testbed of a model of the emergence and growth of neoplasm in differentiated normal tissu. This model, based on the major biological mechanisms identified as responsible for tumor initiation and promotion, considers the competition between differentiation and growth factors as responsible for the likelihood of genetic abnormalities which may change the differentiation to proliferation equilibrium. On the opposite of deterministic models, the characteritics and behaviors of individual cells are defined. The simulations have shown that the population equilibrium was a spontaneous development in the model. This computer model allows the study of the majority of problems encountered in image cytometry, for diagnosis and prognosis based on tumor samples : representativeness of the sample, sample size, instrumental errors, statistical interpretation, fiability of the medical decision, etc...