TREIZIÈME COLLOQUE GRETSI - JUAN-LES-PINS DU 16 AU 20 SEPTEMBRE 1991

# Performances comparées de méthodes de classification de chromosomes par analyse d'images

D.BARBA X.QIU G.RAMSTEIN

IRESTE/LATI, La Chantrerie, CP 3003

44087 Nantes Cedex 03, FRANCE

RÉSUMÉ

Les travaux effectués dans ce papier s'inscrivent dans la perspective de la conception et la réalisation d'un système de classification entièrement automatique des chromosomes.
Des connaissances de cytogénéticiens ont conduit à un ensemble de méthodes d'extraction des caractéristiques chromosomiques et de reconnaissance des chromosomes. Différentes méthodes de caractérisation du profil ont été testées dont une permettant de caractériser localement les bandes chromosomiques fiables. Une évaluation systématique des performances sur une base de données de chromosomes non triés montre que les nouvelles méthodes sont performantes.

ABSTRACT

The works presented in this paper have been realized into the framework of the design and the realization of an automatic system for chromosomes classification.
The knowledge of the cytogeneticians was used for the development of a set of methods which extract the features of chromosomes and recognize the chromosomes. A local method is presented, which characterize the reliable banding patterns of the chromosomes. A systematic performance valuation shows that our tools favourably stand comparison with those recently designed by the other researchers.

## I - INTRODUCTION

De façon schématique, le problème de la classification automatique des chromosomes se décompose de la manière suivante.

A partir de métaphases de bonne qualité observées sous microscope, il s'agit de segmenter l'image métaphasique, c'est-à-dire d'isoler chacun des chromosomes observés du fond. Enfin il reste à reconnaître chacun des chromosomes et à le classer suivant une classification prédéfinie en 24 classes. L'ensemble des chromosomes classés d'une métaphase constitue un Caryotype qui est ensuite analysé et interprété par le cytogénéticien.

Tout naturellement des méthodes ont été proposées et mises en oeuvre pour réaliser automatiquement ou, beaucoup plus souvent, semi-automatiquement tout ou partie de ces différentes phases [1]. Les nombreuses erreurs introduites à différents niveaux de la procédure globale font que seuls des systèmes semi automatiques sont à ce jour opérationnels dans les laboratoires de cytogénétique [2]. Quoiqu'un gain important ait déjà été obtenu avec des systèmes interactifs, les recherches ont continué pour concevoir et réaliser des systèmes plus automatiques. Ainsi, trois versions d'algorithmes ont été récemment proposées pour extraire l'axe médian d'un chromosome [3].

Ensuite, les méthodes de l'intelligence artificielle peuvent jouer un rôle important en permettant l'utilisation d'outils variés de traitement [4].

Si l'un des axes actuels de recherche concerne la conception des systèmes experts pour la construction des caryotypes, les recherches portant sur la caractérisation efficace et robuste des chromosomes et les méthodes de classification associées sont de grande importance car les performances sur images métaphasiques ne sont pas encore satisfaisantes. Nous avions présenté antérieurement [5] des méthodes efficaces d'amélioration et de segmentations d'images métaphasiques et d'extraction utiles à la description morphologique du corps chromosomique. Dans ce papier nous présentons la suite de cette étude en insistant sur les deux aspects complémentaires de description des bandes chromosomiques et de classification des chromosomes.

## II - CARACTERISATION AUTOMATIQUE DES CHROMOSOMES

### 1) Caractérisation morphologique

A la différence de ce qui était proposé pour des chromosomes homogènes, des procédures assez semblables ont été proposées pour la classification des chromosomes en bande [9].

L'extraction des caractéristiques morphologiques consiste d'abord à localiser le centromère, l'axe médian et les points d'extrémité du chromosome [8].

Les principales difficultés de localisation de l'axe médian proviennent de la variation considérable de la morphologie du chromosome (recourbement du chromosome, le repliement des chromatides et la variation de la longueur et de l'apposition centromérique).

Nous avons développé une méthode [5] pour la localisation des points caractéristiques des chromosomes en bande R qui surmonte les difficultés rencontrées par un automate vers les extrémités du chromosome et évite la recherche des points d'extrémité et du centromère sur le contour entier. Il réduit aussi considérablement leur mauvaise localisation.

**2) Caractérisation des bandes**

Les bandes d'un chromosome se placent parallèlement le long de l'axe médian et sont généralement orthogonales à l'axe médian. La courbe monodimensionnelle obtenue par projection est appelée profil.

Il existe trois grandes familles de méthodes de caractérisation du profil :

- La **Comparaison Directe** d'un profil inconnu avec chacune des courbes de référence d'une classe donnée.
- La **Caractérisation Globale** (Fourier, Legendre, etc..).
- La **Caractérisation Locale** consistant à coder la structure de pics et de vallées du profil.

Pour les méthodes de caractérisation globale, nous avons utilisé et comparé les bases de FOURIER (F1,...F18), de Legendre (L1,...,L12) et Walsh-Hadamard (W1,...,W12) ainsi que les caractéristiques WDD (i) de Piper et Granum [13]. De plus nous avons introduit quatre nouvelles caractéristiques globales : les trois premiers moments (My1, My2, My3) et le nombre de passage du profil par sa valeur moyenne (ZCR).

Pour la troisième famille de méthodes par caractérisation locale, outre la décomposition gaussienne du profil ou la représentation en séquences de code de transition (BT-séquences [12]), nous avons conçu également une approche totalement différente de ces deux méthodes [11]. Elle étudie les caractéristiques locales du profil sans le décomposer. Par contre, nous déterminons les statistiques des bandes importantes utilisées par les cytogénéticiens pour reconnaître les chromosomes. Ces statistiques indiquent les positions et les propriétés des segments intéressants des profils de chaque classe chromosomique.

Cette méthode possède les avantages suivants. Elle utilise moins de temps de calcul que les autres méthodes de localisation locale du profil et évite le problème de non-stabilité de la décomposition non-linéaire. Elle localise au maximum la dégradation locale du profil, tandis que pour les autres méthodes, cette dégradation locale influence plus ou moins les caractéristiques extraites par la décomposition du profil.

## III - CLASSIFICATION DES CHROMOSOMES

**1) Base de données**

La base de données que nous avons utilisée est composée de chromosomes provenant de treize métaphases de bonne qualité. (peu de chevauchements). Cette base de données est tout à fait réaliste des conditions réelles opératoires dans un laboratoire.

Les classifications effectuées ont été respectivement menées à partir d'un jeu d'apprentissage et d'un jeu d'essai. Nous avons utilisé la méthode "N-1" présentée par GROEN [15] par laquelle les treize métaphases sont équivalentes à vingt-cinq métaphases regroupées en jeu d'apprentissage comprenant douze métaphases et en jeu d'essai comprenant treize métaphases.

**2) Classifieur Bayesien**

Si un chromosome inconnu est représenté par un vecteur de caractéristiques $X = [x_1,...,x_{Nb}]$, Nb étant le nombre des caractéristiques utilisées, alors dans la phase d'apprentissage, on peut établir pour chaque phase chromosomique $W_i$ un centre représenté par un vecteur $M_i$ et une matrice de covariance $R_i$ de taille NbxNb : pour i = 0,..., Nbcls-1 (W0 et W23 représentent les classes chromosomiques X et Y).

On peut supposer que les caractéristiques utilisées dans le vecteur X sont indépendantes. Cette hypothèse est raisonnable et utilisée souvent par les auteurs [14].

Deux autres hypothèses souvent utilisées sont que les caractéristiques sont distribuées suivant une loi normale pour chacune des classes chromosomiques et que les probabilités a priori qu'un vecteur $W \in W_i$ sont constantes pour i = 0,..., Nbcls-1. Ainsi, la probabilité a priori $P_i(X)$ d'un vecteur x inconnu est :

$$P_i(X) = \frac{\exp(-1/2\, D_i(X))}{(2\Pi)^{Nb/2}\, |\Sigma_i|^{1/2}}$$

où $D_i(X)$ est la distance de MAHALANOBIS entre X et $M_i$ :
$D_i(X) = (X-M_i)\ {}^{T}R_i^{-1}\ (X-M_i)$
Soit $L_i(X) = \mathrm{Log}\ (\Pi_i\ (X))$
Si $L_i(X)$ est le minimum des $P_k(X)$, le vecteur inconnu X est classé selon la classe $X_i$.

**3) Classification d'un chromosome avec les méthodes de comparaison directe des profils**

Deux jeux de profils de référence CR-M et CR-S sont possibles. Chacune des courbes CR-M est obtenue par la moyenne des profils de différents chromosomes d'une même classe. Les courbes CR-S sont les profils d'un jeu de prototypes de chromosomes. La longueur des courbes de référence a été normalisée.

a) Une méthode de **PréclassificationCorrélation** : nous effectuons d'abord une préclassification Bayesienne avec le vecteur des caractéristiques [Sr, Lg, IC, My1, 2, 3, ZCR]. La classification plus fine est basée sur la mesure de similarité entre le profil inconnu et chacune des courbes de référence par corrélation.

b) Une méthode de **PréclassificationComparaisonDynamique** avec une même préclassification que précédemment, mais la classification plus fine est effectuée en fonction de la distance entre le profil inconnu et chacune des courbes de référence mesurée par la méthode de Comparaison Dynamique.

c) Une méthode de **PréclassificationEQM** où la préclassification est toujours la même. Pour obtenir la classification plus fine, les erreurs quadratiques moyennes entre le profil inconnu et les courbes de référence sont calculées.

d) Une méthode **SansPréclassificationEQM**. On calcule d'abord l'erreur quadratique moyenne $E_{pfl}$ entre le profil inconnu et chacune des courbes de référence, et ensuite l'erreur quadratique moyenne $E_{morph}$ entre les paramètres

morphologiques (Sr, Lg, IC) du chromosome inconnu et leur valeur standard mesurée est utilisée comme discriminant pour la classification du chromosome, k étant un poids variable.

### 3-1 Préclassification

Nous avons ajouté aux paramètres morphologiques (surface Sr, Longueur Lg et indice centromérique IC) des informations extraites des bandes (les paramètres My1, My2, My3 et ZCR) pour améliorer les résultats de la préclassification. Ces paramètres sont aisés à mesurer. De plus, ils sont robustes et efficaces pour cette classification.

Deux étapes forment la préclassification. Les chromosomes sont d'abord affectés à une classe par le classifieur Bayesien. Ensuite, les chromosomes classés sont regroupés en Groupes de DENVER [7] selon les relations Groupes-Classes. Nous définissons ici deux niveaux de robustesse pour les résultats de cette préclassification. Si les deux premiers minimums des $L_k(x)$ affectent le chromosome dans les classes du même groupe, où le premier minimum des $L_k(x)$ est inférieur à 75 % du second, le résultat de la préclassification est invalidé : le chromosome sera comparé avec les 24 classes.

Les résultats obtenus avec le vecteur de paramètres VP2 = [Sr, Lg, IC, My1, My2, My3, ZCR] sur le jeu d'essai et aussi sur le jeu d'apprentissage sont donnés dans le tableau 1. On observe que les résultats obtenus avec le vecteur VP2 sont meilleurs que ceux obtenus avec le vecteur VP1 d'un facteur 1.5 en pratique.

| jeu | taux d'erreur dans les classes | taux d'erreur dans les groupes |
|---|---|---|
| essai | 36.2 % | 6.9 % |
| apprentissage | 29.0 % | 5.7 % |

Tableau 1 - Résultats de taux d'erreur de préclassification

### 3-2 Classification fine

Après avoir procédé à la préclassification d'un chromosome dans un des Groupes de DENVER $G_d$ (d = 1,...,7), le groupe $G_d$ comportant Nd classes chromosomiques $Wd_j$ (j = 1,...,Nd), le profil P(i) du chromosome inconnu est ensuite comparé avec chacune des courbes de référence $Rd_j(i)$. La normalisation du profil P(i) (à Nbp points) en longueur et en dynamique engendre le profil normalisé P'(i). Il en est de même pour chacune des courbes de référence $Rd_j(i)$. En essayant séparément deux jeux de profils de référence RC-M et RC-S, les quatre méthodes suivantes ont été testées pour la classification du chromosome.

**a)** La similarité $C_j$ entre P(i) et $Rd_j(i)$ mesurée par corrélation.
Si $C_j$ est le maximum des C1 pour l = 1,..,Nd, le chromosome inconnu est classé à la classe $W^d_j$.

**b)** Nous avons utilisé la procédure suivante de programmation dynamique pour mesurer la distance dynamique moyenne $Dcd_j$ entre les deux courbes P(i) et $R^d_j(i)$. Une matrice $T_j$ est définie, dont l'élément courant $T_j(l,k)$ est en effet une distance cumulée calculée entre P(l-1) et $R^d_j(k-1)$ mesurée par comparaison dynamique. La distance t(k,l) entre ceux-ci est définie comme $t(k,l) = |P(l-1) - R^d_j(k-1)|$. On calcule $T_j(l,k)$ et $Dcd_j$ par les relations suivantes :

$$T_j(l,k) = \text{Min} \begin{matrix} T(l,k-1) + t(l,k) \\ T(l-1, k-1) + 2t(l,k) \\ T(l-1,k) + t(l,k) \end{matrix}$$

$$Dcd_j = T(Nb1, Nb2)/Nbp$$

Le chromosome est classé dans la classe $W^d_j$, si $Dcd_j$ est le minimum des $Dcd_l$ pour l = 1,...,Nd.

**c)** L'erreur quadratique moyenne $E_j$ entre P'(i) et $R'd_j(i)$ :

$$E_j = \sum_{i=1}^{Nbp} (P(i) - R^d_j(i))^2/Nbp$$

$E_j$ est proportionnel à la distance euclidienne entre deux vecteurs P'(I) et $R'd_j(i)$. La distance de MAHALANOBIS $D_j$ peut être également calculée. En supposant que les points des P'(i) et $R'd_j(i)$ sont indépendants, nous avons :

$$D_j = \sum_{i=1}^{Nbp} \frac{(P'(i) - R'^d_j(i))^2}{ri_j^2} \quad / Nbp$$

où $\sigma_{ij}$ est l'écart-type du point i de $R'^d_j(i))$ mesuré pendant les mesures des RC-M.
Si $E_j$ (ou $D_j$) est le minimum des $E_1$ (ou $D_1$) pour l = 1,...,$N_d$, le chromosome est affecté à la classe $W^d_j$

**d)** Dans cette quatrième méthode, l'étape de préclassification n'est pas utilisée. Le profil normalisé P'(i), considéré comme un vecteur de caractéristiques, et les trois caractéristiques morphologiques Sr, Lg et IC sont utilisés dans un classifieur provenant d'une modification du classifieur bayesien pour classer directement les chromosomes selon les 24 classes. Cette modification consistant à ajouter des poids aux caractéristiques morphologiques a pour but d'étudier la relation entre taux d'erreurs de classification et importance des caractéristiques morphologiques par rapport aux informations totales utilisées.

L'évolution des taux d'erreurs de la classification en fonction du poids de k montre que le taux d'erreur est minimum, quand le poids k est égal à 14.

### Analyse des résultats et conclusion

Les résultats Tableau 2 montrent que la méthode de PréclassificationEQM est supérieure à la méthode de PréclassificationCorrélation. Notons que les informations de moyenne et d'écart-type du profil ne sont pas utilisés dans la méthode Préclassification-Corrélation dans la phase de classification plus fine.

Dans la méthode de PréclassificationEQM, la distance de MAHALANOBIS ($D_j$) donne des résultats meilleurs que la distance euclidienne ($E_j$), car les écarts-types de chaque point de la courbe de référence (CR-M) sont pris en compte. Par contre, nous n'avons pas obtenu de résultats aussi bons avec la méthode de PréclassificationComparaisonDynamique. De plus, elle est coûteuse en temps de calcul. Cela peut provenir du choix de la distance t(l,k).
La méthode SansPréclassificationEQM (Tableau 2) engendre des taux d'erreur beaucoup moins élevés que pour les trois autres méthodes. Ils sont comparables avec ceux obtenus par les méthodes de caractérisation globale du profil présentées plus loin. Ceci montre qu'il faut utiliser les caractéristiques morphologiques dans la phase de classification plus fine.

**4) Classification des chromosomes avec des méthodes de caractérisation globale des profils**

Nous avons testé trois groupes de caractéristiques globales extraites des profils pour la classification des chromosomes : les coefficients obtenus par la décomposition en séries de FOURIER des LEGENDRE et de WALSH-HADAMARD.

En utilisant les N premiers coefficients des séries de FOURIER, le meilleur résultat du classifieur bayesien est obtenu quand N est égal à 10 (tableau 2).

**5) Classification des chromosomes avec des méthodes de caractérisation locale des profils**

Les nombres des caractéristiques mesurées sont différents d'une classe à l'autre. Cela pose un problème pour utiliser le classifieur bayesien tel quel, car les vecteurs à l'entrée du classifieur bayesien doivent avoir des dimensions identiques. Nous avons résolu ce problème en modifiant la fonction de discrimination. Au cas où le nombre des caractéristiques n'est pas constant, nous utilisons la contribution moyenne par caractéristique comme discriminant pour classer les chromosomes inconnus.

Les résultats (tableau 2) sont un peu meilleurs que ceux obtenus avec les autres méthodes. Par contre, les temps de calcul sont supérieurs à ceux des méthodes de caractérisation globale du profil.

| méthode | taux d'erreur dans la classe |
|---|---|
| corrélation | 27.6 % |
| EQM | 25.3 % |
| EQM-sans pré-classification | 20.0 % |
| Fourier | 18.8 % |
| Locale | 16.7 % |

Tableau 2 - Résultats de taux d'erreur de classification selon les différentes méthodes utilisées.

## IV - DISCUSSION ET CONCLUSION

Dans la première famille de méthodes, nous avons utilisé les informations du profil pour comparer le profil inconnu avec les courbes de référence. Avec les trois premières méthodes, les chromosomes sont d'abord pré-classés dans les groupes de DENVER. Notre méthode comporte deux étapes pour cette préclassification et utilise à la fois les informations morphologiques et de bandes. Avec les seules informations du profil, les chromosomes sont ensuite affectés à une classe. Si les informations morphologiques pouvaient aussi être utilisées dans cette seconde phase, le taux d'erreur serait significativement diminué. La quatrième méthode obtient des résultats meilleurs que les trois premières méthodes, car elle utilise simultanément les informations morphologiques et celles de profil dans un seul classifieur.

Dans la seconde famille de méthodes, nous extrayons les caractéristiques globales du profil qui sont ensuite classés par le classifieur bayesien. Les caractéristiques globales sont les quatre caractéristiques nouvelles auxquelles sont ajoutées les coefficients des séries (FOURIER, LEGENDRE, WALSH). Les résultats obtenus avec les deux premiers jeux de coefficients sont comparables et meilleurs que pour le troisième. Les fonctions de WALSH, riches en hautes fréquences, ne sont pas convenables pour décomposer les profils chromosomiques.

Dans la troisième famille de méthodes, on extrait les caractéristiques locales du profil selon les segments utilisés dans le profil. Cela produit pour chaque classe chromosomique un vecteur dont la longueur est variable. Le classifieur bayesien a été modifié pour effectuer la classification. Les résultats de classification sont meilleurs que ceux obtenus par les autres méthodes. Ce qui montre l'intérêt d'une telle approche qui devra être développée et affirmée.

## BIBLIOGRAPHIE


[1] C.LUNDSTEEN, T.GERDES, J.MAAHR, and J.PHILIP : "Clinical performance of a routine system for semi-automated chromosomes analysis", Am.J;Hum. Genet. 1987, 41, pp 493-502.

[2] D.RUTOVITZ : "introduction", in Automation of cytogenetics, Ed.Lundsteen C. et Piper J.1989, Germany.

[3] R.STEFANELLI : "A comment on an investigation into the skeletonization approach of hilditch", Patt.Recog.1986, Vol 19, pp 13-14.

[4] CJ.TAYLOR, J.GRAHAM, and D.COOPER : "System architectures for interactive knowledge-based image interpretation", Phil.Trans.R., Soc.Lond.1988, A324, pp 457-465.

[5] X.QIU, D.BARBA and S.LELANDAIS : "extraction de paramètres en vue de la reconnaissance automatique des chromosomes", 7ème Congrès, AFCET, Paris, 1989, pp 1549-1564.

[6] JM.ROBERT : "Génétique et cytogénétique cliniques", dirigée par R.DERE et P.RORER, Flammarion, ISBN, 1977.

[7] Denver Conference, "A proposed standard system of nomenclature of human mitotic chromosomes", Lancet, vol2, 1960, pp 1063-1065.

[8] X.QIU : "Vers la classification automatique des chromosomes", Doctorat de Sciences, Université de Nantes, Nov.1990.

[9] J.GRAHAM : "Automation of routine clinical chromosome analysis, I.", Anal. and Quant. Cytol. and Hist., 1987, n°5, pp 383-390.

[10] X.QIU et D.BARBA : "A low level algorithm library for automatic chromosome classification", 1st European Conference on Biomedical Engineering, Nice, Feb.1991, paper n°89.

[11] X.QIU et D.BARBA :"Classification of the chromosomes with density profile", 12ème International Conference IEEE EMBS, PHILADELPHIA, USA, Nov 1-4, 1990.

[12] C.LUNDSTEEN : "Description of chromosome banding patterns by band transition sequences", 1979, 15, pp 418-429.

[13] J.PIPER and E.GRANUM :"On fully automatic feature measurement for banded chromosome classification", Cytometry, 1989.

[14] J.PIPER :"The effect of zero feature correlation assumption on maximum likelihood based classification of chromosomes", Signal Processing, 1987, 12, pp 49-57.

[15] FCA. Groen and M. van der PLOEG : "DNA cytophotometry of human chromosome banded profiles", 1976, pp 547-550.