LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI
F. MURTAGH

Discrimination des jonctions entre exon et intron dans les séquences d'acide désoxyribonucléique

Les cahiers de l'analyse des données, tome 21, n° 2 (1996), p. 133-148

<http://www.numdam.org/item?id=CAD_1996__21_2_133_0>

Les Cahiers de l'Analyse des Données
Vol. XXI - 1996 - n°2 - pp.133-148.

DISCRIMINATION DES JONCTIONS ENTRE EXON ET INTRON DANS LES SÉQUENCES D'ACIDE DÉSOXYRIBONUCLÉIQUE

[EXON-INTRON]

J.-P. BENZÉCRI
*F. MURTAGH**

La présente étude, fait suite à [DONNÉES RÉSEAUX]: comme dans ce dernier article, on applique les méthodes usuelles d'analyse multidimensionnelle à des données prises dans une collection constituée afin de mettre à l'épreuve des méthodes de discrimination rentrant dans le domaine étendu des réseaux de neurones.

1 Position du problème et examen des données

1.1 Structure formelle des séquences d'ADN

Il ne nous appartient pas de reproduire un des plus beaux chapitres de la biologie moléculaire. Nous voulons seulement faire voir, dans la vie, l'origine du problème de linguistique formelle considéré ici.

Il est établi, depuis quelque 40 ans, que les acides nucléiques renfermés dans les chromosomes, donnent, suivant un code universel, la formule des protéines produites dans l'organisme. En bref, le chromosome consiste en des molécules rangées en double hélice; la partie signifiante de la chaîne étant une suite de bases qu'on peut assimiler à un texte écrit dans un alphabet dont les quatre lettres sont:

{(T): Thymine; (A): Adénine; (C): Cytosine; (G): Guanine}

Ce texte s'analyse, sans ambiguïté, comme une suite orientée de mots de trois lettres, dits triplets, ou codons; dont chacun, exception faite de {TAA, TGA, TAG}, qui sont dépourvus de signification, définit un acide aminé: e.g., l'alanine, peut être codée par l'un des 4 triplets {GCT, GCA, GCC, GCG}, commençant par GC. De sorte qu'une suite de triplets peut se traduire, par un processus physicochimique complexe, en une suite d'acide aminés, c'est-à-dire, en une protéine. Plus précisément, de la longue séquence d'un

(*) Informaticien à l'observatoire de l'ESO;
Karl-Schwarzschild-Straße ; D-85748 Garching bei München.

chromosome, seuls certains des segments, copiés d'ADN en ARN messager, régissent ensuite, dans le cytoplasme, au niveau des ribosomes, l'assemblage d'une protéine; les autres étant éliminés lors de l'épissage de l'ARN. Et il y a, sur une séquence d'ADN, entre certains couples de triplets consécutifs, la marque d'une séparation potentielle, délimitant un exon, servant de code pour une protéine; d'un intron, qui n'a pas cette valeur. [Par exemple, un gène fonctionnel peut consister en 3 exons, séparés par deux introns; mais il existe aussi des gènes, à structure continue, définis par un segment unique d'ADN.]

```
EI   CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG
IE   TCTCATACCTTTTCTCTGGGGTCATTCCAGGTATGAGACAGAGTTGAACCTGCGCATGAG
N,   ACGGAGCGAGTCTGGAACCTGATCAGATACATCTATAACCAAGAGGAGTACGCGCGCTAC
```

Les données considérées ici consistent en 3190 séquences de 60 caractères de l'alphabet {T, A, C, G}. De ces séquences, issues de chromosomes de divers primates, chacune comprend 20 triplets; et l'on sait si la rencontre du 10-ème et du 11-ème triplet marque une jonction Exon-Intron, EI; une jonction Intron-Exon, IE; ou ne marque pas de limite de l'une ou de l'autre sorte, N; la répartition approximative de l'ensemble étant: EI, 25%; IE, 25%; N, 50%. L'analyse mathématique doit faire la preuve de son efficacité en discriminant entre les séquences rentrant dans les trois classes {EI, IE, N}.

Les données sont réelles; la structure du corpus, artificielle. Mais peut-être le poids de ≈ 50%, attribué à la modalité négative, N, a-t-il été bien choisi en vue d'applications à des données ultérieures. Nous voulons croire que le problème, tel qu'il est posé, offre un véritable intérêt pour la biologie.

1.2 Bilan statistique du voisinage des points de jonction

Avant toute analyse multidimensionnelle, il vaut la peine de fixer son attention sur les deux seules lettres de rang 30 et 31, entre lesquelles peut se trouver une jonction.

en ligne: carac 30, 31
toutes les séquences

3	EI	IE	N
GG	625	365	146
GC		110	105
GA		217	100
GT		74	80
CG	23		43
CC			118
CA		1	124
CT			133
AG	60		122
AC	1		97
AA			99
AT		1	94
TG	57		151
TC			83
TA			54
TT			106

Il apparaît que le point final, comme le point initial, d'un intron est toujours marqué par un G (guanine); la règle ne souffrant que 3 exceptions. En effet, dans la colonne EI du tableau, les nombres non nuls (un chiffre 1, sur la ligne AC, excepté) sont dans les lignes {GG, CG, AG, TG} (ce qu'on notera, en bref: GG et xG): le 31-ème caractère, début de l'intron, est un G. De même, dans la colonne IE, exception faite de deux chiffres 1, dans les lignes CA et AT, les nombres non nuls sont dans les quatre lignes {GG, GC, GA, GT} (on dira: GG et Gx). Seule la colonne N s'étend sur l'ensemble des 16 lignes du tableau.

Un problème de discrimination ternaire, entre EI, IE et N, se pose seulement sur la ligne GG, pour les séquences ayant un G, à la fois, en position 30 et 31. Dans les cas Gx = {GC, GA, GT}, il n'y a que discrimination binaire, entre IE et N; de même, dans les cas xG ={CG, AG, TG}, discrimination binaire entre EI et N; les 9 autres cas, xx = {CC, CA, CT, AC, AA, AT, TC, TA, TT}, des paires sans G, ne peuvent, sauf exceptions rarissimes, marquer une jonction proprement dite. Le problème n'est pourtant pas résolu, ne fût-ce qu'approximativement, par ces remarques; car plus du tiers des cas sont sur la ligne GG.

Sans prétendre achever la discrimination par une recherche arborescente de conditions, nous étendrons quelque peu l'examen du voisinage du point, 30-31, de l'éventuelle jonction. En nous bornant aux séquences 30=31=G, nous considérons séparément la distribution des triplets centraux, de rang 10 et 11.

en ligne: carac 32,33
séquences de centre GG

3	EI	IE	N
GG		11	7
GC	2	21	14
GA		32	12
GT		15	9
CG	2	7	5
CC		20	12
CA	2	17	2
CT		41	9
AG	1	25	16
AC		15	14
AA		30	13
AT		23	8
TG	294	50	10
TC	16	25	5
TA	294	14	6
TT	14	19	4

Sur un premier tableau, dont les lignes sont étiquetées par les caractères 32 et 33, on note qu'à une jonction GG, l'intron commence, le plus souvent, par GTG ou GTA: en effet, dans la colonne EI, prédominent les valeurs 294, portées sur chacune des deux lignes TG et TA; la part de EI et N étant particulièrement faible dans la ligne TA (i.e.: 32=T; 33=A; 11-ème triplet=GTA).

Sur un deuxième tableau, dont les lignes sont étiquetées par les caractères 28 et 29 de la séquence, se signalent, par leur poids élevé, les lignes CA, AA.

en ligne: carac 28,29
séquences de centre GG

3	EI	IE	N
GG	19		14
GC	14		4
GA	67		13
GT	11		9
CG	16		7
CC	24		6
CA	175	278	13
CT	30		16
AG	32		11
AC	11		3
AA	136	9	14
AT	27		9
TG	22		10
TC	11		5
TA	11	78	7
TT	19		5

Sur la ligne AA, figure quasi exclusivement la modalité EI: donc un 10-ème triplet CCG, l'un des quatre codons, {CCT, CCA, CCC, CCG}, pour la proline, suivi d'un G suggère fortement une jontion EI.

Sur la ligne CA, il s'agit de séquences où le 10-ème triplet, (suivi d'un G), est CAG: soit l'un des deux triplets, {CAA, CAG}, codant pour la glutamine. Les jonctions correpondantes sont quasi exclusivement de type EI ou IE, la modalité N étant très rare.

```
en ligne: carac 32,33
centre GG, précédé de CA
 3    EI   IE    N
GG          9
GC     1   16    3
GA         27
GT         12    2
CG     2    6
CC         15
CA         11    1
CT         37
AG         21    1
AC         11    2
AA         14    3
AT         16
TG    71   38
TC    11   21
TA    83   10    1
TT     8   14
```

Cherchant une règle pour distinguer entre ces deux cas, on considérera un nouveau bilan, restreint aux séquences dont le 10-ème triplet est CAG, suivi de G; les lignes du tableau étant étiquetées par les caractères 32 et 33.

On retrouve ce qui a été vu sur le bilan des caratères 32, 33 pour les séquences dont le centre est GG: il n'y a de jonction EI que si le 11-ème triplet commence par GT; mais alors, entre les deux modalités de jonction, EI et IE, la discrimination reste à faire.

2 Analyse, sous forme dijonctive complète, de 3185 séquences restreintes aux 12 bases des 4 triplets médians

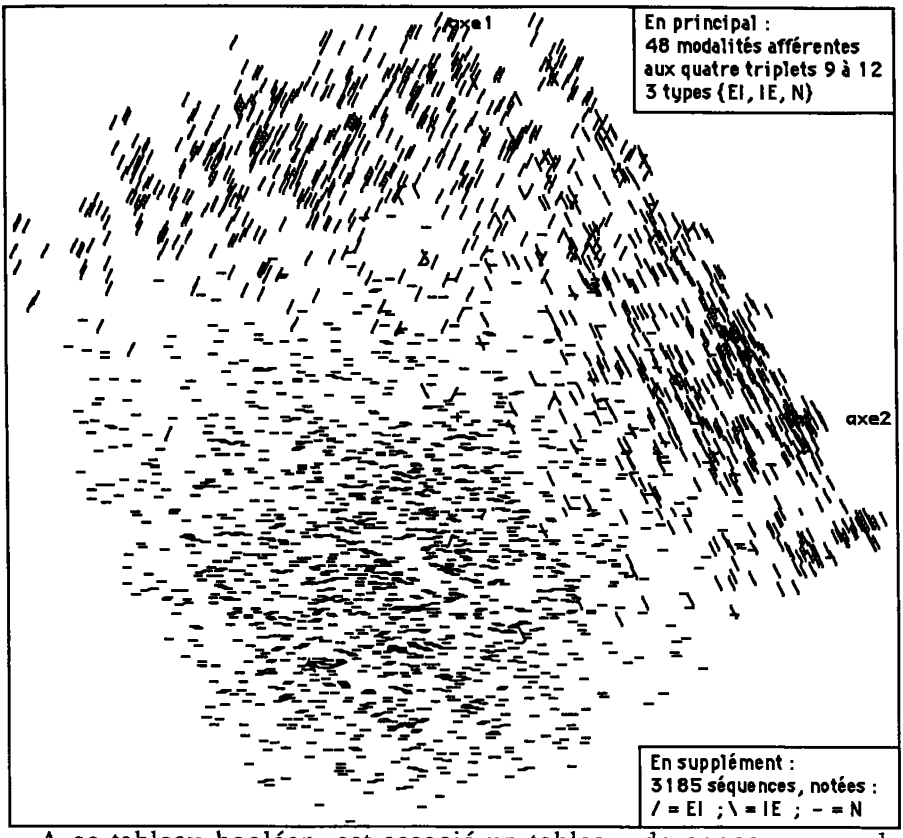
Au §1.2, les caractères des séquences ont été considérés en conjonction ou en disjonction: ici, un ensemble de caractères sera considéré simultanément en formant des combinaisons linéaires continues de variables logiques.

2.1 Tableau des données et tableaux analysés

Dans quelques cas, les données publiées comportent un doute: au lieu d'une base déterminée, on indique seulement une alternative entre deux bases. Pour les quatre triplets médians, de telles lacunes ne se trouvent que dans 5 des 3190 séquences: nous ne considérerons désormais que les 3185 séquences restantes.

Pour la discrimination des types de jonction, on procédera comme dans [DONNÉES RÉSEAUX], §2, pour l'appréciation de l'activité thyroïdienne. [Quant aux lignes incomplètes, ou pourrait procéder, comme on l'a fait dans le même article, au §1, pour la classification des tumeurs mammaires: ici, il faudrait, en cas de doute, partager la masse 1 entre les modalités concurrentes.]

On part du tableau 3185×13, donnant, sous forme numérique, pour chaque séquence, sa classe, cl, puis les 12 bases des 4 triplets médians: {a, b, c, d, e, f, g, h, i, j, k, l}: e.g. j renvoie à la 1-ère base du 12-ème triplet (34–ème caractère dans la séquence complète; et 10-ème, dans la partie centrale, seule considérée ici). Après codage sous forme disjonctive complète, on a un tableau booléen (en 0, 1), 3185×51. Pour les 3 modalités du type de jonction, on conserve les sigles {EI, IE, N}; les modalités des bases reçoivent des sigles de deux lettres; la première (en bas de casse) désignant la position; et la deuxième (en capitale) la base proprement dite: e.g. jA pour: Alanine en position j.



A ce tableau booléen, est associé un tableau, de cooccurrence des modalités, ou tableau de BURT, 51 × 51. Ce tableau est analysé avec, en principal, les trois colonnes des modalités de jonction; et les 48 lignes des modalités des 12 bases. Restent donc en lignes supplémentaires les 3 modalités de jonction. Avec trois colonnes principales seulement, l'analyse ne fournit que deux facteurs non triviaux: l'ensemble des résultats se voit dans un plan. Les modalités {EI, IE, N} y figurent deux fois: comme colonnes principales et comme lignes supplémentaitaires; celles-ci étant beaucoup plus excentriques que celles-là; et un point hx est marqué au milieu du segment joignant deux modalités, ligne et colonne, de même nom.

2.2 Analyse de correspondance: le nuage des 3185 séquences

À l'analyse du sous-rectangle de BURT ainsi défini, on adjoint en colonnes supplémentaires le tableau disjonctif complet décrivant les 3185 séquences. Afin de montrer le potentiel de discrimination de l'analyse, chaque séquence est marquée, dans le plan (1, 2), par un sigle attribué à sa classe; soit: / = EI, \ = IE, - = N.

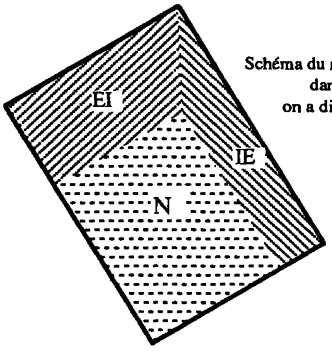


Schéma du nuage des 3185 séqences dans le plan (1, 2): on a distingué trois zones

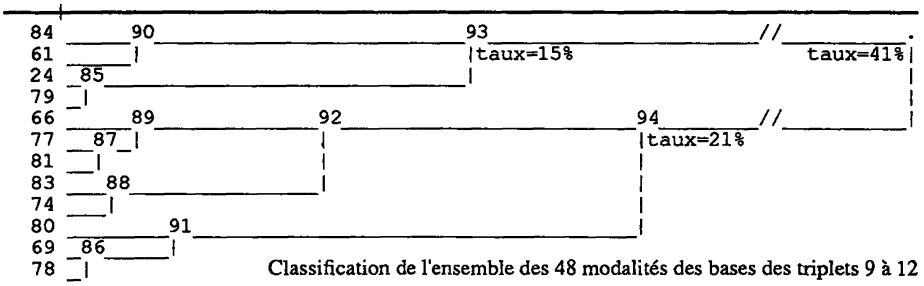
On distingue nettement 3 zones afférentes aux 3 classes de jonction; il y a quelque empiètement au niveau des frontières; mais rares sont les séquences qu'on voit isolées au sein d'une zone autre que celle de leur classe.

Schématiquement, on assimilera le nuage des 3185 séquences à un rectangle, diposé obliquement; au coin inférieur gauche, la zone N occupe un sous-rectangle; la bande restante se partage, au niveau de l'angle supérieur droit, en deux trapèzes afférents à EI et IE. Cette disposition se prête bien à l'analyse discriminante (cf. *infra*, §3).

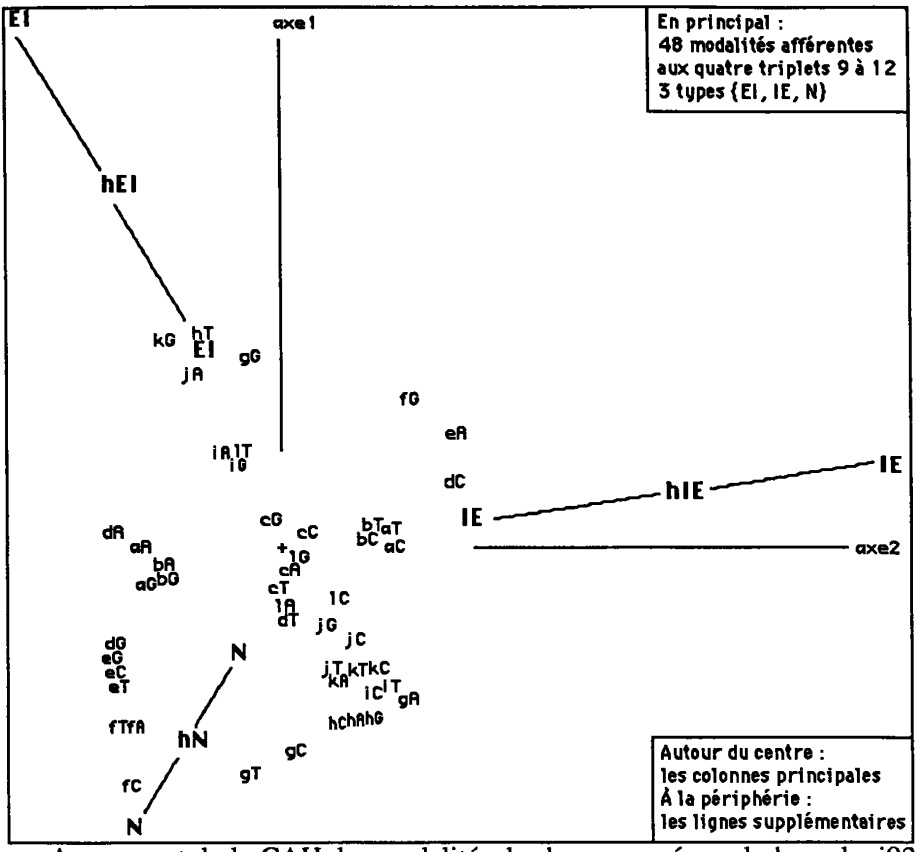
2.3 Nuage des 48 modalités des bases dans les triplets centraux

Pour examiner avec ordre, dans le plan (1, 2), le nuage des modalités des bases, on considérera simultanément la CAH de ces 48 modalités.

c	Partition en 12 classes : Sigles des modalités de la classe c
84	hT kG jA gG
61	iA lT iG
24	fG
79	eA dC
66	aT bT bC aC
77	cG cC lG cA
81	cT lA dT lC jG
83	jC jT kT kA kC iT iC gA hG hA hC
74	gC gT
80	dA aA bA bG aG
69	dG eG eC eT
78	fA fT fC



Classification de l'ensemble des 48 modalités des bases des triplets 9 à 12



Au sommet de la CAH des modalités des bases, se sépare la branche i93, subdivisée en i90 et i85. Les trois modaltés {dC, eA, fG} de i85 sont celles d'un triplet CAG placé au dixième rang d'une séquence. Le bilan statistique du milieu des séquences s'accorde avec l'image du plan (1, 2) pour montrer i85 nettement opposé à N, et associé à IE plutôt qu'à EI.

Les modalités de i90 sont, notamment, celles de triplets GTA et GTG placés au onzième rang. La classe i90 s'écarte de l'origine dans la direction de EI; plus précisément, i90 se subdivise en i60 et i84, celle-ci s'écartant plus que celle-là de l'origine. Il est remarquable que soient dans i84 les modalités jA et kG, qui, bien qu'appartenant au 12-ème triplet, lequel est séparé du milieu par le 11-ème, apportent à l'analyse de fortes contribution (celle de kG n'étant même dépassée que par hT).

Reste la branche i94, divisée en i92 et i91. La classe i91 se subdivise en i80, i69, i78. Cette dernière subdivision s'écarte nettement de O dans la direction de N; et elle comprend les modalités {fA, fT, fC} (autres que fG)

```
en prin: 48 mod de base x 3 classes
trace  :   2.318e-1
rang   :      1     2
lambda :   1373   945   e-4
taux   :   5924  4076   e-4
cumul  :   5924 10000   e-4
```

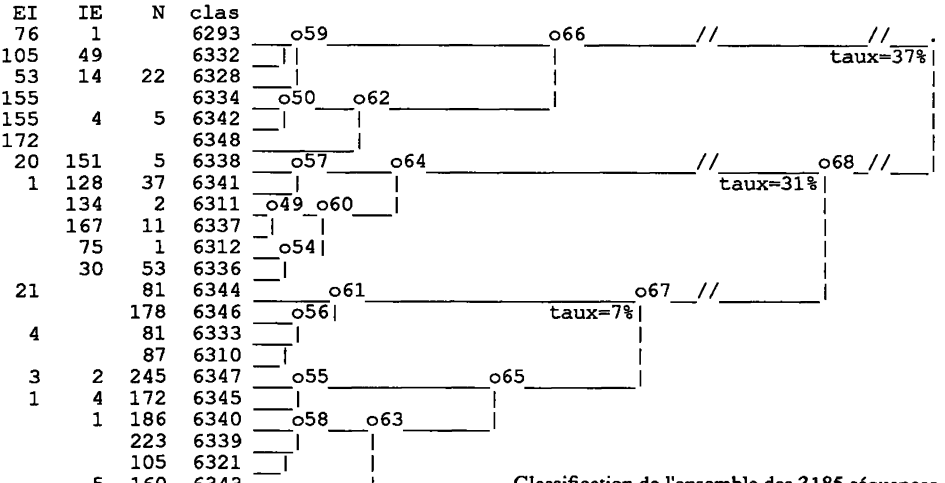
afférentes au 30-ème caractère; modalités fx dont l'association avec N a déjà été vue. Les subdivisions i69 et i80 sont proches de N, mais s'en écartent dans la direction de EI; aucune modalité de i69 ou de i80 n'apporte de forte contribution.

La classe i92 se sudivise en i89 et i88. Il n'y a rien à dire de i89, sinon que ses modalités entourent l'origine. Enfin, i88 est proche de N; particulièrement sa subdivision i74, tandis que i83 s'en écarte vers IE. On a dnas i88 les modalités {gA, gT, gC} (autres que gG) afférentes au 31-ème caractère; modalités, gx, dont l'association avec N a déjà été vue (tandis que gG va avec EI).

2.4 Classification de l'ensemble de 3185 séquences

Avec 3185 individus, la CAH complète a 3184 nœuds, numérotés de 3186 (la paire de séquences agrégée au niveau le plus bas) à 6369 (le sommet de la hiérarchie). Ici, on considère la partition en 22 classes définie par les 21 nœeuds les plus hauts. Sur le graphique arborescent, les nœuds sont marqués par 3 caractères, la lettre 'o' remplaçant les deux chiffres initiaux 63. En marge de l'arbre, le contenu des classes est donné par un bilan, suivant les 3 types de jonction, {EI, IE, N}.

Au sommet de la hiérarchie la branche o66 se sépare du reste, o68.



Classification de l'ensemble des 3185 séquences

```
branche o66: 716.EI +  68.IE +   27.N
branche o64:  21.EI + 685.IE +  109.N
branche o67:  28.EI +  13.IE + 1518.N
total      : 765.EI   766.IE   1654.N
```

bilan des types par branche

Dans chacune des 6 subdivisions retenues de la branche o66, prédominent les jonctions IE: les classes 6334 et 6348 comprennent exclusivement de telles jonctions; le taux de EI ne descend au-dessous de 94% que dans les classes 6332: 105.EI + 49.IE, et 6328: 53.EI + 14.IE + 22.N.

La branche o68 se scinde en o64 et o67; où prédominent respectivement IE et N. Trois des 6 subdivisions de o64, {6311, 6337, 6312}, représentent IE, presque sans mélange. Mais dans 6636, (30.IE+53.N), la modalité négative, N, est plus fréquente que IE.

Dans 3 des 8 subdivisions de la branche o67, on trouve exclusivement N; ailleurs, le taux de N dépasse 95%; fait seule exception la subdivision 6344, (21.IE + 81.N).

D'après le tableau du bilan général, il apparaît qu'en assimilant le type d'une séquence au type prédominant de la branche où celle-ci est classée, on a le type exact dans plus de 91% des cas : (716+685+1518)/3185 = 91,65%. De façon précise, ainsi qu'on l'a noté en décrivant la CAH, la plupart des erreurs sont dans un petit nombre de subdivisions, qui correspondent à l'empiétement des zones dans le plan (1, 2).

3 Discrimination entre séquences d'après les 12 bases des 4 triplets médians

L'analyse factorielle et la CAH montrent que les séquences relevant de types de jonction différents se distinguent quant à la composition des triplets médians. On appliquera donc diverses méthodes d'analyse dicriminante; d'abord, au §3.1, pour considérer l'ensemble des séquences, sans réserver d'échantillon d'épreuve; ensuite, au §3.2, en réservant un tiers des individus, afin d'apprécier, l'efficacité potentielle de la méthode dans son application à des données nouvelles.

3.1 Affectation de la totalité des 3185 séquences aux trois types de jonction

3.1.1 Affectation au centre de classe le plus proche

De façon précise, les tableaux ci-joints donnent les bilans des estimations; la nature véritable (ou donnée pour telle...) de la jonction étant notée en colonne; et les affectations respectives, précédées de i ou de j, sur les lignes.

Il apparaît que le bilan des erreurs varie selon que les cas individuels sont

	EI	IE	N
jEI	726	39	79
jIE	35	717	133
jN	4	10	1442

	EI	IE	N
iEI	653	30	3
iIE	21	549	14
iN	91	187	1637

	iEI	iIE	iN
jEI	685	2	157
jIE	1	582	302
jN	0	0	1456

affectation des séquences d'après les 12 bases médianes

affectés aux jx, profils des colonnes principales; ou aux ix, lignes supplémentaires, qui, sur le plan (1, 2) se projettent beaucoup plus loin de l'origine que les jx. Dans l'affectation aux jx, la modalité négative, jN, n'est presque jamais donnée pour des séquences ayant une jonction EI ou IE; mais des contacts rentrant dans N sont indûment affectés à jEI et jIE. Dans l'affectation aux ix, on a des erreurs de sens opposé. Le croisement entre les deux affectations, aux ix et aux jx, confirme ce qu'on peut appeler: le pouvoir d'attraction de iN (au sein du système des ix).

	EI	IE	N
hEI	709	31	26
hIE	35	700	57
hN	21	35	1571

	hEI	hIE	hN
iEI	679	7	0
iIE	0	549	0
iN	87	201	1627

	hEI	hIE	hiN
jEI	765	8	71
jIE	1	784	100
jN	0	0	1456

affectations exactes: aux hx: 2980/3185 = 93,56% ;
aux ix: 2839/3185 = 89,14% ; aux jx: 2723/3185 = 85,49% .

Ceci suggère de prendre pour centres d'affectaion des points hx, milieux repectifs des segments {ix, jx}. Le taux d'affectations exactes passe au-dessus de 93%. Les croisement entre affectations montre que 288 des séquences affectées à iN vont à hEI ou hIE; corrélativement, aucune des séquences affectées à iEI ou iIE ne va à hN. Au contraire, 171 séquences affectées à jEI ou jIE, vont à hN; tandis que toutes les séquences affectées à jN vont à hN. Ainsi, l'affectation aux hx, corrige les erreurs systématiques notées dans les affectations respectives aux ix et aux jx.

3.1.2 Bilan des affectations pour des groupes de séquences définis par le voisinage immédiat du point de jonction

On a vu au §1.2, que la nature de la jonction est partiellement déterminée par les bases placées aux rangs 30 et 31 (f et g selon les notations adoptées pour les analyses factorielles): ceci suggère de reprendre le bilan des affectations en distinguant quatre classes de séquences GG, Gx, xG, xx (x désignant une base quelconque, autre que la Guanine, G).

Il y a peu à dire sur le cas de xx. Les 910 séquences dont le centre est xx (i.e. n'ayant de Guanine ni en f, ni en g; ou encore, ne rentrant ni dans la modalité fG, ni dans gG), rentrent dans la modalité négative N; à 3 exceptions près, une jonction EI et deux jonctions IE. Dans l'analyse générale, ces exceptions sont affectées (par erreur) à hN; de plus, 4 cas rentrant dans N sont indûment affectés à hIE. Il vaut donc mieux affecter à N, sans autre examen, toute séquence de type xx.

	EI	N
hEI	135	11
hIE	0	1
hN	5	304

	EI	N
hEI	135	11
hN	5	305

affectation des séquences dont le centre est xG
à gauche, affectations à 3 centres h ; à droite, à 2 centre

Pour les séquences dont le centre est xG, il n'y a strictement que deux types attestés dans notre corpus: EI ou N. Concurremment avec l'affectation générale de ces séquences aux trois centres {hEI, hIE, hN}, on peut faire une affectation restreinte, à celui des deux centres {hEI, hN} dont la séquence est le plus proche. Le tableau montre qu'en procédant ainsi on évite une affectation erronée (d'un cas de N à IE), le reste du bilan n'étant pas changé.

	IE	N
hEI	6	4
hIE	368	31
hN	27	250

	IE	N
hIE	368	32
hN	33	253

affectation des séquences dont le centre est Gx
à gauche, affectations à 3 centres h ; à droite, à 2 centre

Pour les séquences dont le centre est Gx, notre corpus n'offre que les deux types: IE et N. Comme pour xG, on compare l'affectation générale à celle restreinte à deux centres; ici: {hIE, hN}. On lit sur le tableau que cette dernière méthode permet d'éviter 3 erreurs (affectations de séquences de type N). D'une manière ou d'une autre, la discrimination entre IE et N, pour les séquences en xG, apparaît en butte à plus d'erreurs que la discrimination entre IE et N, pour les séquences en Gx.

	EI	IE	N
πhEI	570	20	8
πhIE	39	334	20
πhN	15	9	118

	EI	IE	N
hEI	574	25	11
hIE	35	332	21
hN	15	6	114

	πhEI	πhIE	πhiN
hEI	591	14	5
hIE	6	370	12
hN	1	9	125

afffectation des séquences dont le centre est en GG
centres πhx issus de l'analyse partielle ; centres hx issus de l'analyse générale

L'affectation des séquences en GG se présente avec la difficulté maxima: les 3 types {EI, IE, N} étant possibles.

Pour ces séquences, nous donnons, d'une part, le bilan des affectations aux centres {hEI, hIE, hN} issus de l'analyse générale des 3185 séquences; et, d'autre part, le bilan des affectations à des centres {πhEI, πhIE, πN}, construits de façon analogue, mais d'après une analyse partielle portant sur l'ensemble des séquences dont le centre est GG. Entre les deux variantes, les différences sont minimes: il faut toutefois noter que nous n'avons pas repondéré les colonnes du sous-rectangle de BURT afférent aux séquences en GG: or dans ce corpus partiel, le poids de N est faible (pour un exemple de repondération des colonnes, cf. [DONNÉES RÉSEAUX], §2).

3.2 Affectation aux classes de jonction avec échantillon d'épreuve

3.2.1 Affectation à trois centres

L'analyse principale du §2 est fondée sur l'ensemble des 3185 séquences pour lesquelles les 12 bases des 4 triplets médians sont connues sans ambiguïté. Le succès de la discrimination fondée sur cette analyse incite à reprendre la même méthode en réservant un échantillon d'épreuve.

On conserve le codage du §2.1; mais le tableau de BURT est construit en recensant les cooccurrences de modalités dans un sous-ensemble S+ formé des 2/3 des séquences. À l'analyse du sous-rectangle de BURT croisant les 48 lignes, afférentes aux modalités des 12 bases, et les 3 colonnes, classes de jonction, on adjoint en supplément le tiers restant, S3, des séquences. Chacune de celles-ci est affectée au centre le plus proche.

De façon précise, comme au §3.1.1, on prend pour centre, dans le plan (1, 2), le milieu, noté %hx, du segment joignant les projections jx et ix de la colonne et de la ligne afférentes à une même classe x. D'autre part, on retient, pour comparaison, les affectations obtenues au §3.1.1 pour les séquences de ce même ensemble S3.

	EI	IE	N		EI	IE	N		%hEI	%hIE	%hN
%hEI	236	13	11	hEI	237	12	12	hEI	259	0	2
%hIE	12	224	16	hIE	12	228	18	hIE	1	252	5
%hN	7	18	524	hN	6	15	521	hN	0	0	542

affectations exactes de l'échantillon d'épreuve aux %hx: 984/1061 = 92,74%

Les tableaux ci-joints donnent les bilans des estimations; la nature véritable (ou donnée pour telle..) de la jonction étant notée en colonne et les affectations respectives, précédées de %h ou de h.

Aucune différence notable n'apparaît entre les bilans des erreurs d'affectation de S3: d'une part, aux %hx, comme échantillon d'épreuve non pris en compte dans le sous-tableau de BURT d'où résultent les centres; d'autre part, aux hx, dont la construction prend en compte toutes les séquences, en particulier celles de S3. Au niveau même des séquences individuelles, un croisement montre que les deux affectations sont les mêmes, à 8 exceptions près.

3.2.2 Affectation des séquences supplémentaires à la classe de la séquence principale la plus proche

Ainsi qu'on l'a noté au §2.2, le schéma du nuage des séquences suggère une discrimination facile des zones par des lignes droites. Or affecter les séquences au centre le plus proche revient à partager le plan en des zones délimitées par les médiatrices des segments joignant les centres deux à deux.

De ce point de vue, on peut considérer comme fortuit le succès de

l'affectation au triangle des centres hx ou %hx, dont les médiatrices ont l'orientation et la position convenable; tandis qu'avec le triangle des jx ou des ix les médiatrices écornent les zones.

C'est pourquoi, en vue de généralisations ultérieures, nous considérons une autre méthode d'affectation, propre à délimiter des zones de forme quelconque, ne suivant aucun schéma simple: l'affectation au plus proche voisin. De façon précise, pour tout individu de l'échantillon d'épreuve, on détermine l'individu de l'échantillon de base qui en est le plus proche; et on affecte celui-là à la classe dans laquelle rentre celui-ci.

	EI	IE	N
+EI	242	18	10
+IE	10	216	31
+N	3	21	510

	%hEI	%hIE	%hN
+EI	247	12	11
+IE	7	227	23
+N	6	13	515

	hEI	hIE	hN
+EI	249	12	9
+IE	6	231	20
+N	6	15	513

+ désigne l'affectation des 1061 séquences supplémentaires aux 2124 séquences principales
affectations exactes par +x: 968/1061 = 91,23%

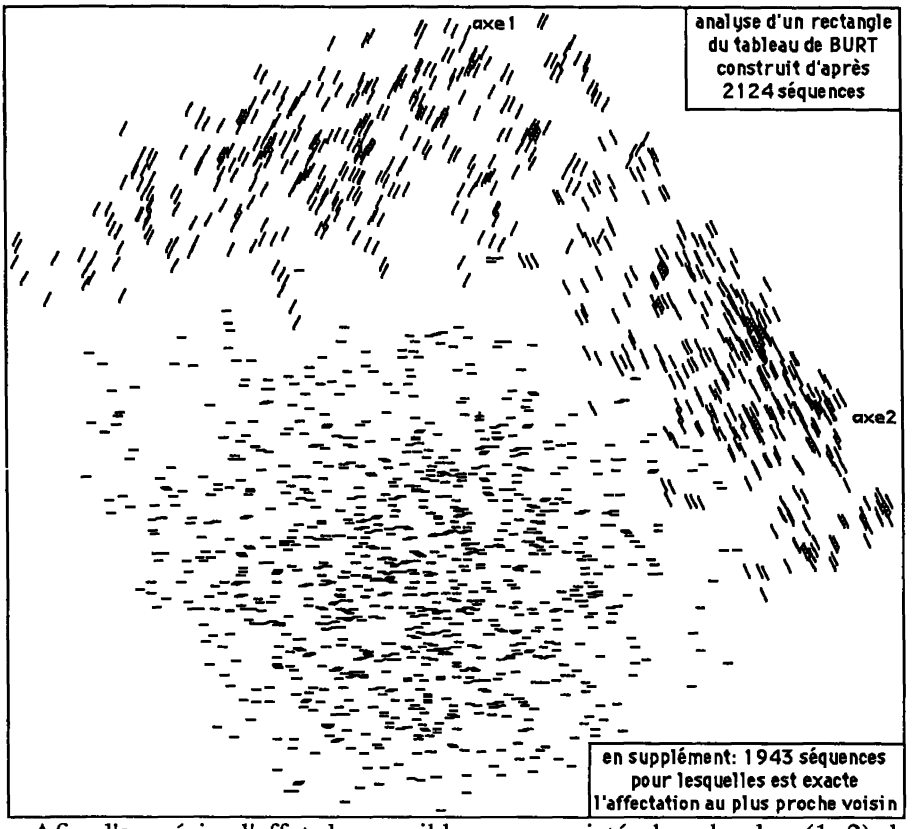
On voit sur les tableaux de bilans que l'affectation d'après le plus proche voisin, pris dans l'ensemble des 2124 séquences principales, diffère peu, dans le cas de nos données des affectations à un système de trois centres, hx ou %hx.

3.2.3 Affectation des séquences supplémentaires à la classe de la séquence la plus proche dans un sous-ensemble choisi

Divers perfectionnements ont été proposés pour la méthode d'affectation d'après le plus proche voisin: par exemple, on peut se fonder, non sur le plus proche voisin seulement, mais sur plusieurs voisins qui offrent une image plus exacte des proportions des divers types dans le voisinage de l'individu supplémentaire à affecter. On citera, à ce propos la méthode de régression avec nombre variable de voisins, appliquée par A. G. HATHOUT à l'analyse de données boursières

Ici, nous partons de la remarque que, parmi les individus de l'échantillon de base, certains, isolés au sein d'une zone où prédomine un type autre que le leur, ne sont pas de bons indicateurs du type de leurs voisins. De tels individus sont donc à éliminer, *a priori*, dans la recherche du plus proche voisin d'un individu supplémentaire.

Voici comment on a procédé pour mettre cette remarque à profit. Pour chacun, ib, des 2124 individus (séquences) de l'échantillon de base, on a déterminé celui, i'b qui en est le plus proche; et ib n'a été conservé comme indicateur du type (des individus de l'échantillon supplémentaire) que si ib et i'b étaient de même type. Ainsi, n'ont été retenu comme indicateurs que 1943 individus sur 2124.



Afin d'apprécier l'effet de ce crible, on a projeté, dans le plan (1, 2), le sous-nuage de ces 1943 individus: il n'y a plus de séquences isolées au sein d'une zone autre que celle de leur classe; et, au niveau même des frontières entre zones, l'empiètement est réduit, relativement à celui qu'on a observé au §2.1; dans l'analyse générale.

	EI	IE	N
r+EI	239	12	13
r+IE	12	226	23
r+N	4	17	515

	+EI	+IE	+N
r+EI	256	2	6
r+IE	9	242	10
r+N	5	13	518

	%hEI	%hIE	%hN
r+EI	253	4	7
r+IE	5	242	14
r+N	2	6	528

r+ désigne l'affectation des 1061 séquences supplémentaires
à un ensemble réduit de 1943 séquences principales
affectations exactes par r+x: 980/1061 = 92,37%

On voit sur les tableaux de bilans que l'affectation, r+x, d'après le plus proche voisin, pris dans un ensemble restreint de 1943 séquences principales, améliore quelque peu, dans le cas de nos données, les affectations, +x, d'après

la totalité de l'échantillon de base; la taux d'affectation exacte étant à peine inférieur à celui afférent au système de trois centres, %hx. Succès d'autant plus appréciable que, comme on l'a dit, l'heureuse disposition des centres relativement aux frontières des zones apparaît comme une particularité fortuite de nos données.

Quant aux fluctuations des taux calculés en fonction du partage des données entre échantillon de base et échantillon d'épreuve, nous noterons que les erreurs auxquelles sont en butte les affectations sont de trois types:

certains individus (séquences) sont manifestement au sein d'une zone relevant d'une autre classe que celle que leur attribue le fichier analysé: qu'il s'agisse ou non d'une erreur de saisie, un algorithme ne peut que leur attribuer la classe de la zone où ils tombent;

les domaines afférents aux trois types de jonction sont contigus entre eux: au mieux, l'affectation dans les marges sera le fait d'un hasard non biaisé; et il y aura un certain nombre (d'ailleurs fluctuant) d'erreurs inévitables;

enfin, il se peut que la cloison - plan médiateur etc... - tracée par l'algorithme prenne en travers la frontière optima: d'où des erreurs q'une méthode parfaite doit éviter.

De cepoint de vue, l'examen des graphiques plans vaut mieux que les épreuves de validité usuelles de la statistique.

4 Conclusions et perspectives

Nous soulignerons en conclusion que, si le principe des diverses méthodes d'affectation à des centres est simple et bien connu, le succès d'une application de ces méthodes dépend essentiellement de l'espace euclidien dans lequel on se place; c'est-à-dire de l'analyse multidimensionelle qui, avant toute discrimination, a servi à construire cet espace (cf. T. K. GOPALAN & F. MURTAGH).

Ici, après codage disjonctif complet des variables discrètes, on a soumis à l'analyse de correspondance un sous-rectangle du tableau de BURT croisant, en bref, la variable à expliquer avec les variables explicatives. La méthode d'affectation s'est alors montrée stable; et l'on est en droit de présumer qu'elle pourra servir, ultérieurement, à déterminer la nature d'un contact entre deux triplets, avec un taux d'erreur comparable à celui trouvé dans la présente étude.

Une étude approfondie des mêmes données a montré, d'autre part, qu'il y a quelque latitude, dans le codage et le choix même du tableau analysé; l'atteste l'article de A. M. ELKAYAR, [STRUCTURE INTRON], publié dans ce même cahier.

Références sur les données

Origine générale de la compilation des données:

[PROBEN1] : A set of Neural Network Benchmark Problems and Benchmarking rules; Sept. 1994; par :

Lutz PRECHELT (prechelt@ira.uka.de), Fakultät für Informatik, Universität Karlsruhe; 78128 Karlsruhe, Allemagne.

Les données de [PROBEN1] renvoient à des collections bien connues:

University of California at Irvine machine learning databases archive: (`ics.uci.edu`, directory: `/pub/machine-learning-databases`);

où les données du présent article sont comprises sous le titre:

"Primate splice-junction detection gene sequences (DNA) with associated imperfect domain theory".

Quant à l'utilisation antérieure des données, la référence la plus récente, donc la plus complète, est:

G.G. TOWELL & J.W. SHAVLIK : "Interpretation of Artificial Neural Networks: Mapping Knowledge-based Neural Networks into Rules", in: *Advances in Neural Information Processing Systems*; Vol IV, Morgan KAUFMANN; (1992).

Références sur les méthodes de l'Analyse des Données

T. K. GOPALAN & F. MURTAGH : "The Role of Input Data Coding in Multivariate Data Analysis: The Example of Correspondance Analysis"; à paraître;

HASSAN HAMOUD Anwar : "Diversité des codages permis en analyse discriminante: exemple de données cytologiques"; [CODAGE DISCRI.], in *CAD*, Vol.XXI, n°1, pp.75-82; (1996);

A. G. HATHOUT : "La régression d'après un nombre variable de voisins"; [RÉGR. NVAR. VOIS.], in *CAD*, Vol.VIII, n°1, pp.19-26; (1983);

A. G. HATHOUT : "Régression avec nombre variable de voisins et régression avec stratégie variable"; [RÉG. STRA. VAR.], in *CAD*, Vol.X, n°4, pp.470-476; (1985);

A. G. HATHOUT : "Étude préalable à la constitution d'une cible pour l'identification des valeurs mobilières dans la période comprise entre le 18/10/85 et le 21/11/86"; [VAL. MOB.], in *CAD*, Vol.XII, n°1, pp.91-110; (1987);

F. MURTAGH : "Application de l'analyse factorielle et de l'analyse discriminante à des données colligées pour être soumises à des réseaux de cellules"; [DONNÉES RÉSEAUX], in *CAD*, Vol.XXI, n°1, pp.53-74; (1996).