LES CAHIERS DE L'ANALYSE DES DONNÉES

A. M. ALKAYAR

Structure des introns au voisinage d'un point limité entre exon et intron sur une séquence d'ADN

Les cahiers de l'analyse des données, tome 21, n° 2 (1996), p. 149-164

<http://www.numdam.org/item?id=CAD_1996__21_2_149_0>

Les Cahiers de l'Analyse des Données
Vol. XXI - 1996 - n°2 - pp.149-164.

STRUCTURE DES INTRONS AU VOISINAGE D'UN POINT LIMITE ENTRE EXON ET INTRON SUR UNE SÉQUENCE D'ADN

[STRUCTURE INTRON]

A. M. ALKAYAR *

Le présent article fait suite aux analyses de F. MURTAGH, dans [EXON-INTRON]; travail auquel nous renvoyons pour une description du problème avec le format des données.

Dans [EXON-INTRON], comme il est de règle en analyse discriminante, l'espace, où les individus sont affectés aux classes par la recherche de plus proches voisins, est construit en prenant explicitement en compte la répartition en classes des individus (du moins, de ceux d'un échantillon de base): le tableau principal de l'analyse est le rectangle, croisant l'ensemble des 48 modalités descriptives, avec celui des 3 modalités {EI, IE, N} de la variable, cls, à estimer.

Sans prétendre d'abord contribuer à la discrimination des types de jonction, mais seulement afin d'en poursuivre la description, nous avons analysé le tableau de BURT, 48×48, des modalités descriptives. Or il est apparu que la répartition des séquences en trois classes, certes inhérente à notre corpus, sort clairement de l'analyse du tableau carré; et domine même les deux premiers facteurs, sans être aucunement corrélée aux suivants. En sorte qu'on a pu reprendre, dans le plan (1, 2), le problème de discrimination; et obtenir des taux de discrimination exacte peu inférieurs à ceux de [EXON-INTRON].

Quant à la structure des limites entre intron et exon, l'analyse confirme, ce qui était déjà apparu: que les traits caractéristiques en sont presque tous du côté de l'intron; soit au début, dans un jonction EI; soit à la fin, dans une IE. Tandis qu'il va sans dire que rien ne caractérise les séquences sans jonction comprises dans la modalité négative N.

Voyant la forte structure des données, nous avons cru que celles-ci méritaient une analyse ultérieure, fondée sur un codage par triplets des

(*) Docteur de l'Université Pierre et Marie CURIE.

Les cahiers de l'analyse des données 0339 3097/96/02/$ 4.00/

séquences prises dans toute leur longueur; et non plus en se restreignant aux 12 caractères (bases: T, A, C, G) médians, codés un par un. Cette analyse complémentaire, comme les précédentes, éclaire la structure, en général; et peut servir à la discrimination.

1 Étude des séquences fondée sur l'analyse du tableau carré croisant avec elles-mêmes les 48 modalités descriptives des douze bases centrales

1.1 Analyse factorielle du tableau 48 × 48: comparaison avec l'analyse du tableau 48 × 3

De façon précise, le tableau de BURT est celui construit au §3 de [EXON-INTRON], d'après un corpus de base de 2164 séquences.

```
en pr : 48 modal x 3 cl
trace : 2.318e-1
rang  :    1     2
lambda: 1373   945 e-4
taux  : 5924  4076 e-4
cumul : 5924 10000 e-4
```

SIGJ	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR
ci-dessous éléments principaux									
EI	1000	240	423	591	857	612	-242	143	148
IE	1000	241	312	89	26	14	541	974	746
N	1000	519	265	-315	837	375	-139	163	106

Les résultats de l'analyse du rectangle 48 × 3 étant rappelés ci-dessus, considérons ceux de l'analyse du carré 48 × 48.

analyse du tableau de BURT 48 × 48, calculé d'après 2164 séquences de base

trace :	2.880e-1									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	391	308	144	133	121	116	107	101	95	89 e-4
taux :	1357	1070	499	464	419	403	370	350	330	309 e-4
cumul :	1357	2427	2926	3389	3809	4212	4582	4932	5262	5571 e-4

SIG	Qlt10	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR	F 3	CO2	CTR	F 4	CO2	CTR
ci-dessous éléments supplémentaires															
EI	978	20	28	605	900	188	-142	49	13	9	0	0	87	19	11
IE	958	20	21	15	1	0	532	940	184	11	0	0	-38	5	2
N	973	43	18	-287	686	91	-181	273	46	-9	1	0	-23	4	2

Ce n'est pas sans quelque surprise que nous avons trouvé chacune des 3 modalités supplémentaires, {EI, IE, N}, corrélée à plus de 94% avec le plan (1, 2); (les corrélations avec les axes suivants ne pouvant, quant à elles, être que négligeables); alors que la contribution générale des facteurs 1 et 2 à l'inertie globale n'est que de ≈ 24% (cf: cumul).

Dans le plan (1, 2), les deux modalités de jonction effective, {EI, IE}, sortent de l'enveloppe convexe du nuage des 48 modalités descriptives principales; et la modalité négative, N, est à la périphérie de ce nuage. Sur le plan (1, 2) du §2 de [EXON-INTRON], l'ensemble des modalités figure trois fois: comme colonnes principales jx; comme lignes supplémentaires ix; et par les points milieux hx. Il apparaît que l'ensemble des jx est, relativement aux modalités decriptives, moins excentrique que dans la présente analyse; les hx et, *a fortiori*, les ix sortant, certes, plus nettement.

Afin de préciser cette comparaison, on a, sur l'ensemble des 48 modalités descriptives, calculé des corrélations entre facteurs issus des deux analyses, 48 × 3 et 48 × 48. En bref, les deux premiers facteurs issus du tableau carré, reproduisent ceux issus de l'analyse du rectangle [à une faible rotation près; d'ailleurs visible sur les graphiques plans: le point IE étant ici presque exactement sur le demi-axe (F2>0); alors qu'il s'en écarte visiblement dans l'analyse du rectangle]. Au-delà, il n'y a pas de corrélation notable; d'autant moins qu'il convient de considérer les coefficients de corrélation au carré, lesquels expriment les contributions des facteurs d'un système à la reconstitution de l'inertie de ceux de l'autre.

```
axe ≈ analyse du BURT 48x48  ;  fac ≈ analyse du rectangle 48x3
corrélations calculées sur 48 modalités
corr(axe1,fac1) = .963  corr(axe2,fac1) = .173  corr(axe3,fac1) = .0225
corr(axe1,fac2) =-.164  corr(axe2,fac2) = .954  corr(axe3,fac2) = .0162
```

On peut ici comparer utilement les valeurs propres issues des deux analyses. Dans l'analyse 48 × 48, les deux premières valeurs propres sont nettement séparées des suivantes; mais elles n'apportent ensemble qu'un quart de l'inertie du nuage. Dans l'analyse 48 × 3, toute l'inertie est sur deux axes et les valeurs propres sont environ trois fois plus fortes que celles, de même rang, issues du tableau 48 × 48. Dans cette dernière analyse, relativement à l'autre, les distances absolues sont donc divisées par $\approx\sqrt{(3)}$. Au contraire, si pour placer les modalités {EI, IE, N} - pricipales dans l'analyse 48 × 3, supplémentaires dans l'analyse 48 × 48 - on applique la formule de transition, la différence d'échelle entre les facteurs sur les 48 modalités descriptives, réduits ici relativement à ce qu'ils sont là, est exactement compensée par le changement du facteur $1/\sqrt{(\lambda)}$, plus fort ici que là: d'où, à la rotation près des axes 1 et 2, à peu près les mêmes faxcteurs sur {EI, IE, N}; ce qu'attestent les extraits de listage.

corrélations calculées sur 1061 individus (échantillon d'épreuve)

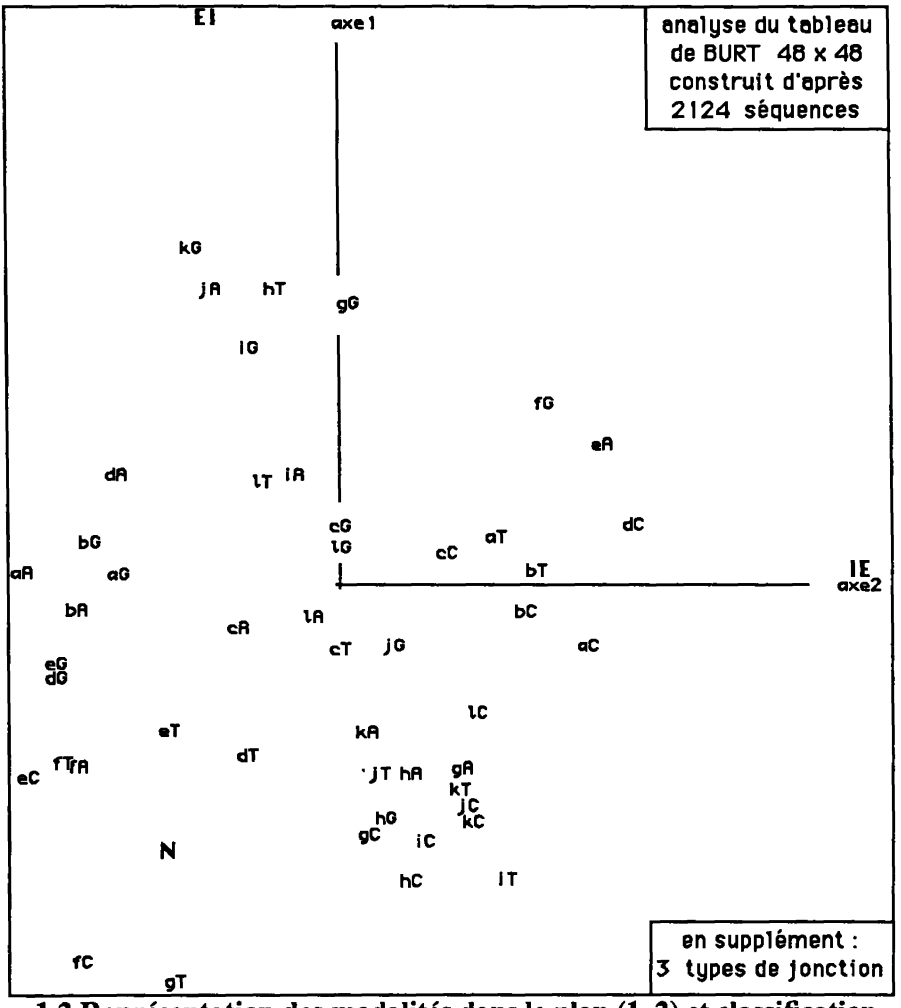
```
corr(axe1,fac1)  =  .976     corr(axe1,fac2)  = -.218
corr(axe2,fac1)  =  .133     corr(axe2,fac2)  =  .967
```

Les mêmes considérations valent pour les séquences individuelles. Les formules de régression linéaire citées ici, attestent que les facteurs sont équivalents; le décalage du zéro - d'ailleurs faible: de l'ordre du centième - s'explique parce que les corrélations concernent non l'échantillon de base sur lequel a été calculé le tableau de BURT, mais l'échantillon d'épreuve; lequel nous intéresse particulièrement pour l'analyse discriminante.

```
corr(axe1,fac1) = .976                    corr(axe2,fac2) = .967
fac1 - .012 ≈ .964 * (axe1 - .012)        fac2 + .006 ≈ .967 * (axe2 + .005)
axe1 - .012 ≈ .988 * (fac1 - .012)        axe2 + .005 ≈ .967 * (fac2 + .006)
```

corrélations et régressions calculées sur 1061 individus (échantillon d'épreuve)

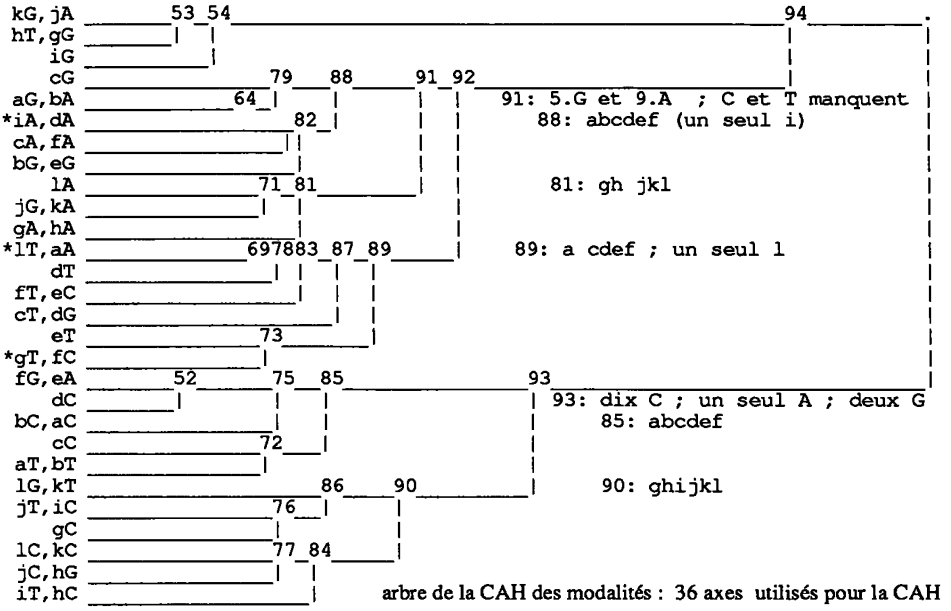
C'est pourquoi, dans le plan (1, 2) issu de la présente anlyse, l'image des 1061 séquences de l'échantillon d'épreuve est conforme à ce qui a été vu dans [EXON-INTRON]; (cf. *infra*, §1.3).



1.2 Représentation des modalités dans le plan (1, 2) et classification

Comme au §2.3 de [EXON-INTRON], l'ensemble des 48 modalités sera considéré simultanément dans le plan (1, 2) et sur l'arbre de la CAH: relativement à l'exposé précédent, celui-ci est plus libre, vis-à-vis de la discrimination: la structure y sortant directement du corpus de séquences.

Nous avons effectué la CAH, d'une part d'après les deux premiers facteurs, d'autre part, dans l'espace de dimension 36 (i.e.: 48-12: nombre des modalités - nombre des variables) issu de l'analyse du tableau carré de BURT. Les deux structures étant similaires on ne présentera que la seconde, afin de compléter, sans répétition inutile, l'article précédent.



arbre de la CAH des modalités : 36 axes utilisés pour la CAH

Il est frappant que la hiérarchie s'interprète bien comme un système de classes formées de modalités qui renvoient, quasi exclusivement, soit aux 6 positions {abcdef} antérieures à la jonction éventuelle; soit à {ghijkl}, qui viennent après. Dans le plan (1, 2), on voit, en bref, que celles de ces classes qui, écartées de l'origine, contribuent à la structure globale s'opposent entre elles suivant l'axe 1, si elles concernent les places {ghijkl}; et suivant l'axe 2 si elles concernent {abcdef}. Ainsi apparaît la prédominance de l'intron, annoncée dans l'introduction. Suivant l'axe 2, où se détache IE, il s'agit d'un intron précédant la jonction, donc occupant les places {abcdef}: en particulier, du côté (F2>0) on voit que C y est fréquent en {a, b, c, d} et T, en {a, b}; au contraire, vers (F2<0), on voit que, dans ce même intron précédant la jonction IE, A est rare en {a, b, c, d} et G en {a, b, d, e}. Vers (F1>0) se détache EI: jonction avec un intron qui est sur {ghijkl}: s'y associe la clase 54 des modalités {kG, jA, hT, gG, iG}: la présence universelle de G en g (après la jonction) est bien connue; et le reste confirme ce qui est noté au §1.2 de [EXON-INTRON]; à l'opposé, vers (F1<0) mais incliné dans la direction (F2>0) pour s'opposer exactement à EI, prédomine l'ensemble des modalités afférentes aux places ghijkl; notamment {gC, hC, iC, jC, kC, lC} et {iT, jT, kT} qui sont dans la classe 90 de modalités.

D'autre part, à un niveau hiérarchique supérieur à celui des classes, en abcdef ou ghijkl, que nous avons considérées, la CAH met, dans la branche 93, (10/12) des C; et dans 92, (9/12) des A; et on trouve, dans le plan (1, 2), que les modalités xC peuvent être renfermées dans un demi-plan, limité à une droite passant par l'origine et contenant le quadrant (F1<0; F2>0).

1.3 Analyse discriminante dans l'espace issu de l'analyse du tableau carré de BURT

	EI	IE	N
36EI	244	17	30
36IE	11	236	21
36N	0	2	479

	EI	IE	N
2EI	240	18	37
2IE	15	229	44
2N	0	8	470

	2EI	2IE	2N
36EI	280	9	2
36IE	7	271	11
36N	8	8	465

affectation de 1061 séquences au centre le plus proche : analyse 48 × 48

36 : affectation dans l'espace de dim 36 ; 2 : affectation dans le plan (1, 2)

affectations exactes par 36 : 959/1061 = 90,385% ; par 2 : 939/1061 = 88,5%

1.3.1 Affectation à un ensemble de trois centres

Si les centres {EI, IE, N}, modalités supplémentaires, étaient exactement dans le plan des axes (1, 2), l'affectation serait la même dans ce plan ou dans l'espace de dimension 36: en effet, il est équivalent de comparer les distances d'un point M de l'espace ambiant à plusieurs points d'un plan Π (ou plus généralement d'un sous-espace linéaire quelconque) et de conparer les distances à ces points de la projection orthogonale, pr(M), de M sur Π. Mais, dans le cas présent, les centres {EI, IE, N} s'écartent quelque peu du plan (1, 2): et il se trouve que le taux d'affectation exacte de l'échantillon d'épreuve augmente de ≈2% quand on prend en compte tous les facteurs. On notera que ce taux dépasse ceux obtenus au §3.2.1 de [EXON-INTRON] dans l'affectation aux centres j (éléments principaux) ou même i (éléments supplémentaires): mais il est inférieur à celui obtenu avec les centres h.

1.3.2 Affectation à la classe de l'individu le plus proche

	EI	IE	N
rxEI	226	13	10
rxIE	16	219	21
rxN	13	23	520

	EI	IE	N
xEI	230	16	19
xIE	13	214	21
xN	12	25	511

	xEI	xIE	xN
rxEI	236	7	6
rxIE	15	225	16
rxN	14	16	526

affectation de 1061 séquences à la classe de l'individu de base le plus proche

+ renvoie à l'analyse du rectangle 48 × 3 ; x , à l'analyse du carré 48 × 48

l'initiale r indique que l'échantillon de base est réduit aux 1840 (436+401+1003) cas bien affectés

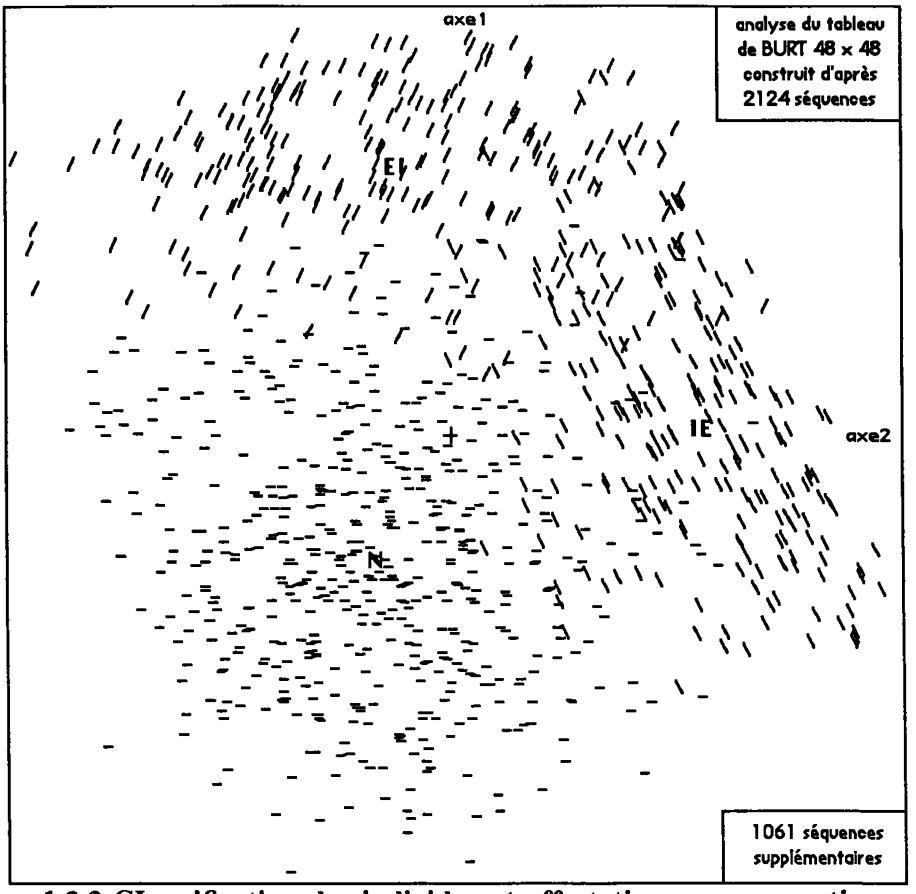
affectations exactes par rx : 965/1061 = 90,95% ; par r+ : 980/1061 = 92,37%

	r+EI	r+IE	r+N
rxEI	236	10	3
rxIE	11	233	12
rxN	17	18	521

N.B. pour la restriction aux séquences bien affectées, cf. [EXON INTR0N], §3.2.3.

Il apparaît qu'après analyse du tableau carré 48 × 48, le taux d'affectation exacte à l'ensemble de base réduit est de ≈91%; peu inférieur à celui obtenu après analyse du rectangle 48 × 3; les affectations des séquences d'épreuve étant généralement les mêmes avec les deux méthodes.

Avec l'analyse du tableau de BURT carré, on tend vers ce que les spécialistes de la reconnaissance des formes appellent parfois: "apprentissage sans maître"; certes l'information relative au type sert pour affecter les séquences d'épreuve; mais l'espace où sont calculées les distances d'affectation est construit sans prendre en compte cette information; du moins explicitement, car la composition de l'échantillon est telle qu'elle impose la structure en trois types de jonction.



1.3.3 CLassification des individus et affectation par apprentissage sans maître

```
4237 _____________4238_____________________________________________.
4233 _____________|                4238: 458EI + 74IE + 81N = 613   |
4130 _____________4240_____________________________4246_____________|
4139 _____________| 4240: 21EI + 343IE + 93N = 457|
4242 ___________________4245_____________________|
4244 ___________________|  4245: 31EI + 94IE + 929N = 1054
```

Dans la partition en 3 classes issue de la CAH des individus de base, on ne reconnaît qu'approximativement les 3 types de séquences; mais, dans une partition en 18 classes, beaucoup de subdivisions sont à peu près pures. En les prenant comme centres d'affectation, on pourrait tenter une discrimination, touchant au domaine de l'apprentissage sans maître; à la réserve près formulée ci-dessus, au §1.3.2, que la structure est fortement imprimée dans le corpus.

2 Analyses fondées sur un codage par triplets des séquences prises dans toute leur longueur

2.1 Correspondance entre codons et places

Avec un alphabet de 4 bases {T, A, C, G}, il y a un ensemble I de 64 triplets possibles; pour lesquels nous conservons d'abord la notation directe: {TTT, TTA, TTC, TTG, TAT, TAA,...}. Dans notre corpus, une occurrence d'un triplet peut être caractérisée par le type, {EI, IE, N}, de la séquence à laquelle il appartient et par sa position (ou rang, de 1 à 20) au sein de cette séquence: soit un ensemble J de $3 \times 20 = 60$ modalités de place; auxquelles on attribue les sigles explicites:

{EI01, EI02, ..., EI20, IE01, IE02, ..., IE20, N01, N02, ..., N20}.

De ce point de vue l'ensemble des occurrences du corpus peut être ventilé suivant un tableau $I \times J$, 64×60:

k(i, j) = nombre des occurrences du triplet *i* occupant une place de type *j*;

par exemple: k(GTG, EI11) = 351, nombre des occurrences de GTG comme onzième triplet dans une séquence ayant une jonction de type EI; (donc au début d'un intron, après un exon).

croisement : 64 triplets × (3 × 20) places (distinguées par type de séquence)

trace :	9.114e-1									
rang :	1	2	3	4	5	6	7	8	9	10
lambda :	3288	2729	1079	564	229	181	129	100	70	61 e-4
taux :	3607	2994	1183	619	252	199	141	110	77	67 e-4
cumul :	3607	6601	7785	8403	8655	8854	8995	9105	9182	9249 e-4

2.1.1 Interprétation des facteurs

L'analyse de ce tableau appelle immédiatement l'attention sur les fonctions des triplets observées au §1.2 de [EXON-INTRON] en faisant une suite de croisements; et elle offre, de la structure du corpus, une vue d'ensemble qui incite à en reprendre la discrimination dans une nouvelle voie.

Les principaux facteurs étant créés par des associations entre un petit nombre d'éléments i et j, l'interprétation est donnée sans graphique, d'après des extraits de listage; et sera corroborée par la CAH

SIGI	QL10	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR
GTA	998	13	243	4140	978	659	-562	18	15
GTG	992	25	106	1944	973	286	-214	12	4
SIGJ	QL10	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR
EI11	1000	12	352	5121	981	957	-707	19	22

L'axe 1 est créé par l'association entre les triplets {GTA, GTG} (surtout GTA) et la place EI11, au début de l'intron dans une jonction EI.

SIGI	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR
TAG	989	7	54	-489	36	5	-2452	905	165
CAG	999	35	233	-480	38	25	-2409	960	749
SIGJ	**QLT**	**PDS**	**INR**	**F 1**	**CO2**	**CTR**	**F 2**	**CO2**	**CTR**
EI10	993	12	39	-288	28	3	-1095	405	53
IE10	1000	12	277	-820	32	25	-4483	959	887

L'axe 2 est créé par l'association entre {CAG, TAG} (CAG plus fréquent que TAG) et la place IE10, à la fin de l'intron dans une jonction IE. On notera qu'un élément peut être nettement écarté de l'origine sur un axe sans être corrélé à celui-ci (TAG et CAG sur l'axe 1; GTA et GTG sur l'axe 2); ou sans lui apporter de contribution notable (EI sur l'axe 2).

SIGI	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR	F 3	CO2	CTR
AGT	997	14	82	-144	4	1	211	8	2	2166	884	616
AGA	901	19	8	-92	20	0	128	40	1	547	722	52
AGC	909	16	8	-95	21	0	152	55	1	573	771	50
AGG	925	22	12	-105	23	1	96	20	1	626	831	81
SIGJ	**QLT**	**PDS**	**INR**	**F 1**	**CO2**	**CTR**	**F 2**	**CO2**	**CTR**	**F 3**	**CO2**	**CTR**
EI12	1000	12	108	-189	4	1	261	8	3	2718	902	822

L'axe 3 exprime l'association avec EI12, deuxième place de l'intron dans une jonction EI; seul AGT se signale en cette place avec un fort contraste; les autres triplets AGx ont un profil moins écarté de la moyenne, mais sont bien corrélés à l'axe 3. D'autre part, avec des corrélations modérées (≈200‰) les modalités IE01 à IE09 s'opposent à EI12 sur l'axe3.

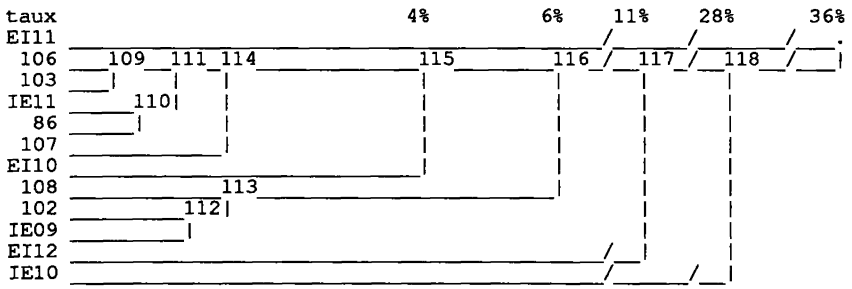
L'axe 4 ne peut être expliqué par un petit nombre de contributions; mais on a sur (F4<0) les huit codons formés des bases T et C {TTT, TTC, TCT, TCC, CTT, CTC, CCT, CCC}, associés, notamment, aux places {IE06...09, EI12}; i.e. à des places dans l'intron distinctes de l'extrémité.

SIG	QLT	PDS	INR	F1	CO2	CTR	F2	CO2	CTR	F3	CO2	CTR	F4	CO2	CTR	F5	CO2	CTR
AAG	954	21	18	-137	24	1	-208	56	3	66	6	1	457	269	80	-611	480	350
EI10	993	12	39	-288	28	3	-1095	405	53	86	2	1	565	108	68	-1099	407	632

L'axe 5 est dominé par l'association entre le triplet AAG et la place EI10, fin de l'exon avant un intron.

En général : EI10 EI11 EI12 et IE06...IE10 sont les seules places à apporter à l'inertie du nuage des INR ≥14‰; parmi celles-ci, il n'y a aucune place, Nx, des séquences rentrant dans la modalité négative; et une seule, EI10, appartient à l'exon, mais au contact immédiat de l'intron.

c	Partition en 12 classes : Sigles des places de la classe c
11	EI11
106	IE15 IE12 EI04 IE18 IE17 EI03 IE13 EI01 EI05 EI09 EI02 EI08
103	IE14 IE20 IE16 EI07 IE19 EI06
31	IE11
86	N17 N09 N03 N12 N11 N20 N13 N07 N04 N02 N08 N16 N15 N01
	N14 N10 N18 N06 N19 N05
107	EI17 EI19 EI15 EI13 EI14 EI16 EI18 EI20
10	EI10
108	IE01 IE02 IE03 IE04 IE05
102	IE06 IE08 IE07
29	IE09
12	EI12
30	IE10



2.1.2 Classification des 60 places et classification des 64 codons

La branche 109, de la CAH des places, comprend les places dans l'exon, E: IE{12...20} EI{01...09}; autres que celles, IE11 et EI10, qui sont sur la jonction; cette branche est proche du centre de gravité (≈CdG).

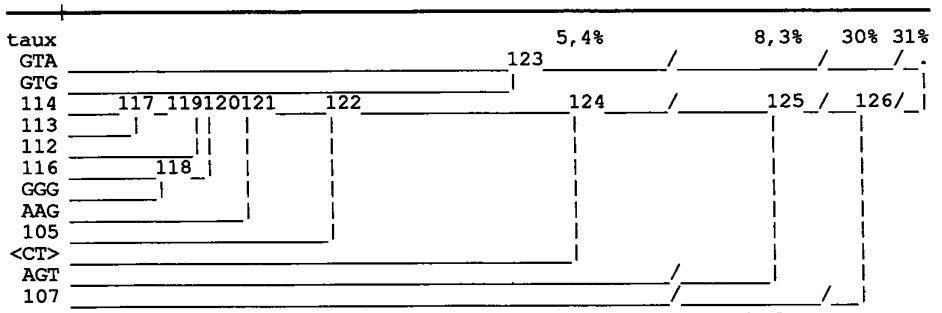
La classe 86, ≈CdG, comprend toutes les places négatives N{01...20}.

La classe 107 comprend les places dans l'intron: EI{13...20}.

La branche 113 comprend les 9 places dans l'intron: IE{01...09}; subdivisées en {IE01...05}, {IE06...08}, {IE09}.

Restent isolées, comme de type excentrique: {EI11, IE10, EI12, EI10, IE11}: soit les deux places, 10 et 11, encadrant une jonction IE ou EI; et immédiatement après le début de l'intron: EI12; tandis qu'on a déjà noté, dans 113, la place IE09 qui précède la fin de l'intron.

c	Partition en 12 classes : Sigles des triplets de la classe c
50	GTA
52	GTG
114	GTC GTT ATT TAT TGT ATC GCT GCC ACC TGA ACT TGC CAC
113	GAA GAT GAC CGA CTA GCA ACA CAA TAC AAC AAA ATA AAT CAT
112	TAA TTA TCA CCA CTG TTG CCG
116	GGC GGA CGC ATG GCG TGG CGG GAG ACG TCG
64	GGG
24	AAG
105	GGT CGT AGA AGC AGG
115	CCC CCT CTC CTT TCT TCC TTC TTT
29	AGT
107	CAG TAG



Sur le tableau de croisement ci-dessous, on lit, e.g. que 1250 occurrences des triplets (formés de C et de T) de la classe i115, notée <CT>, sont dans les places {IE06...08}, formant la classe j102. On a, d'après les listages VACOR, marqué "+" les valeurs qui contribuent le plus aux associations entre profils de classes sur I et J (et "x", celles moins importantes). Les corrélations déjà vues d'après l'analyse factorielle sont ainsi précisées et complétées.

croisement des partitions en 12 classes issues du tableau : 64 triplets × (3 × 20) places

12	EI11	j106	j103	IE11	j86	j107	EI10	j108	j102	IE09	EI12	IE10
GTA	+375	54	36	14	262	29	0	20	12	1	0	0
GTG	+351	253	85	50	645	125	11	42	14	3	1	0
i114	29	1811	883	201	6598	1097	20	925	500	46	56	0
i113	3	1689	840	178	6464	810	60	545	135	70	37	2
i112	0	973	592	34	3607	643	86	523	221	x269	15	1
i116	4	1488	764	121	4708	1054	175	294	101	46	62	5
GGG	0	159	79	11	674	+390	19	52	23	1	23	0
AAG	0	242	116	9	706	85	+134	48	1	0	6	17
i105	0	715	409	61	2506	469	40	166	32	4	+312	0
<CT>	0	1399	591	74	5293	1137	30	x1108	+1250	320	12	0
AGT	0	93	55	10	418	58	3	26	2	0	+234	0
i107	0	283	134	2	1079	199	x184	76	4	5	4	+740

2.2 Correspondance entre séquences et modalités des triplets, groupés par positions et par codons

2.2.1 Le tableau de description des séquences sur 72 modalités

Compte tenu des CAH du §2.1.2, les 20 places de triplets sont rangées en 6 classes, notées {A, B, C, D, E, F}, pour les rangs {1-8,9,10,11,12,13-20}. Les lettres A et F renvoient, respectivement, aux huit premiers des 20 triplets et aux 8 derniers: en effet, dans chacun des trois types de séquences, les profils d'occupation des places, par les 64 codons, diffèrent peu au sein de l'une ou l'autre de ces classes (ils sont dans une même classe de la CAH du §2.1.2). Au contraire, on distingue les places médianes {B, C, D, E} = {9,10,11,12}, dont chacune a un profil d'occupation caractéristique, pour l'un ou l'autre des types de jonction EI ou IE.

Pour les codons, on retient la partition en 12 classes, présentée au §2.1.2; ces classes, qu'elles comprennent un seul codon ou plusieurs, sont simplement notées en chiffres, de 01 à 12, dans l'ordre où elles se présentent sur le tableau du contenu des classes: ainsi, 07 désigne l'unique codon GGG; et 10, l'ensemble, <CT> = i115, des 8 codons formés des caractères C et T à l'exclusion de tout autre.

Ceci posé, on considère un ensemble M de 72 modalités ayant chacune un sigle de 3 caractères; dont le premier est une capitale, de A à F; et les deux derniers désignent un nombre, de 01 à 12: e.g., F07, signifiera: présence du codon GGG dans l'un des rangs 13 à 20 sur une séquence; B10, présence de l'un des codons de <CT> au rang 9. L'ensemble M est rangé dans l'ordre lexicographique usuel:

$$M = \{A01\ A02 \ldots\ A12\ B01 \ldots\ B12 \ldots\ E12\ F01\ F02 \ldots\ F12\}$$

Ainsi, à toute séquence d'ADN, *s*, peut être associé un vecteur de description $\{k(s, m) \mid m \in M\}$, où k(s, m) désigne le nombre des triplets de la séquence *s* rentrant dans la modalité (composée) *m*.

```
ACA GCT CCA GAA GTT GAA AAT GCA ATT AGA GTA CCA GGA AAC AGG AGT TTC TTT TCC CTC

0 0 2 1 2 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
      0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 0 2 1 0
```

Pour plus de clarté, on présente une séquence *s* dont les 20 triplets sont séparés par des blancs; et, en-dessous de celle-ci, sur deux lignes, la suite des k(s, m): les deux premières composantes, k(s, A01) et k(s, A02) sont nulles parce que les codons GTA et GTG ne figurent pas parmi les 8 premiers triplets de *s*; k(s, A03) vaut 2, parce que deux des triplets 1 à 8, le 2-nd, GCT et le 5-ème, GTT, rentrent dans la 3-ème classe de la partition des codons donnée au §2.1.2; etc...; k(s, F12) est nul parce que, des deux codons {CAG, TAG}, constituant la classe 12, aucun n'apparaît dans *s* du rang 13 au rang 20.

2.2.2 Enchaînement des analyses et discrimination des types

Les séqences présentant des données manquantes étant écartées, il reste un tableau 3175 × 72, décrivant suivant l'ensemble M des 72 modalités, un ensemble S de 3175 séquences, dont le bilan par types est (762.EI + 765.IE + 1648.N). Les séquences sont réparties en deux sous-ensembles, S1 et S2, suivant la parité de leur rang: S1, (381.EI + 383.IE + 824.N), sert d'échantillon de base; S2, (381.EI + 382.IE + 824.N), d'échantillon d'épreuve. D'après le tableau S1 × M, on calcule un tableau £a × M, formé des trois lignes de cumul par type des séquences de S1.

Dans une première analyse, £a × M est en principal; S2 × M, en supplémentaire: il n'y a que 2 facteurs; dans le plan (1, 2), on affecte chaque séquence de S2 au centre, ligne de £a, dont elle est le plus proche: le taux d'affectation exacte dépasse 94%. À titre complémentaire, l'analyse est reprise en mettant en supplément les 12 modalités Axx et les 12 modalités Fxx afférentes aux ailes initiale et finale de la séquence, l'information relative aux 4 triplets centraux étant seule gardée en principal: d'où un tableau à 48 colonnes. Le taux d'affectation exacte baisse, mais reste proche de 94%.

Dans une seconde analyse, le tableau S1 × M est en principal; et S2 × M, en supplément. Dans l'espace des profils, engendré par les 66 axes factoriels, on affecte chaque séquence d'épreuve, s2, à la classe de la séquence de base, s1, la plus proche. Le taux d'affectation exacte n'est que de ≈85%.

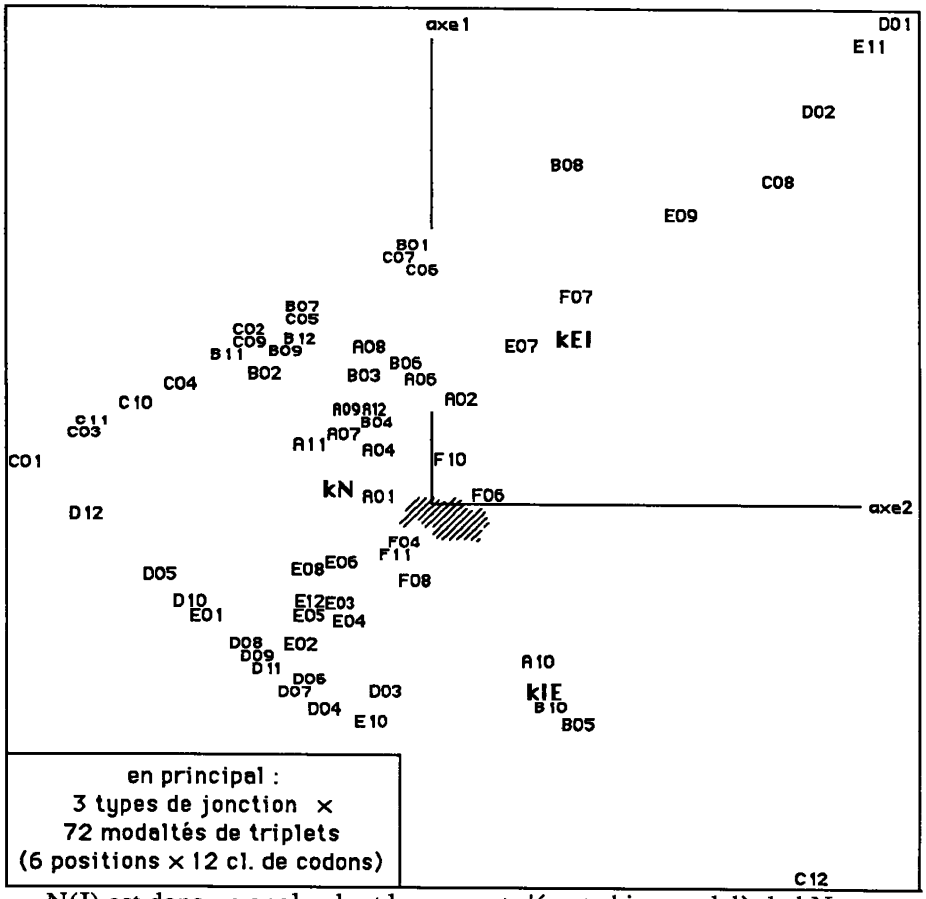
Parce que le succès de la méthode d'affectation aux centres de classes dépend de la courbure des frontières entre les classes, on fait une nouvelle expérience d'affectation aux individus: S1 et S2 étant adjoints en supplément à l'analyse de £a × M, tout s2 est affecté à la classe de la séquence s1 qui en est le plus proche dans le plan (1, 2). Le taux obtenu dépasse 92%; et il approche de 94% si l'on restreint l'échantillon de base au sous-ensemble des séquences de S1 ayant même classe que leur plus proche voisin, au sein de S1 lui-même: cf. *supra*, §1.3.2 et [EXON INTRON], §3.2.3.

2.2.3 Croisement entre 3 types de jonction et 72 modalités des triplets

3 types × ({B,C,D,E} × 12 classes de codons)
trace : 5.707e-1
rang : 1 2
lambda : 3368 2339 e-4
taux : 5902 4098 e-4
cumul : 5902 10000 e-4

SIGJ	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR
qEI	1000	240	446	-1023	987	746	-117	13	14
qIE	1000	241	338	455	259	148	-769	741	610
qN,	1000	519	216	262	288	105	411	712	376

Comme terme de comparaison, nous donnons d'abord des résultats afférents au tableau à 48 colonnes: les profils Axx Fxx, peu contrastés étant écartés, la trace est plus forte; de plus, l'orientation des axes est différente.



N(J) est dans un angle, dont le sommet s'écarte bien au delà de kN:

les modalités, nulles pour la ligne kEI : {C01, D05, D07, D08...12, E01}, sont sur le côté inférieur, opposé à kEI;

les modalités nulles pour la ligne kIE : {B01, B07, B08, B11, (B12 ≈0), C01...03, C05, C07, C09...11}, sont sur le côté supérieur, opposé à IE;

C12, associé à kIE s'oppose absolument à la ligne des autres Cx;

D01 et D02, associés à kEI, s'opposent aux autres Dx;

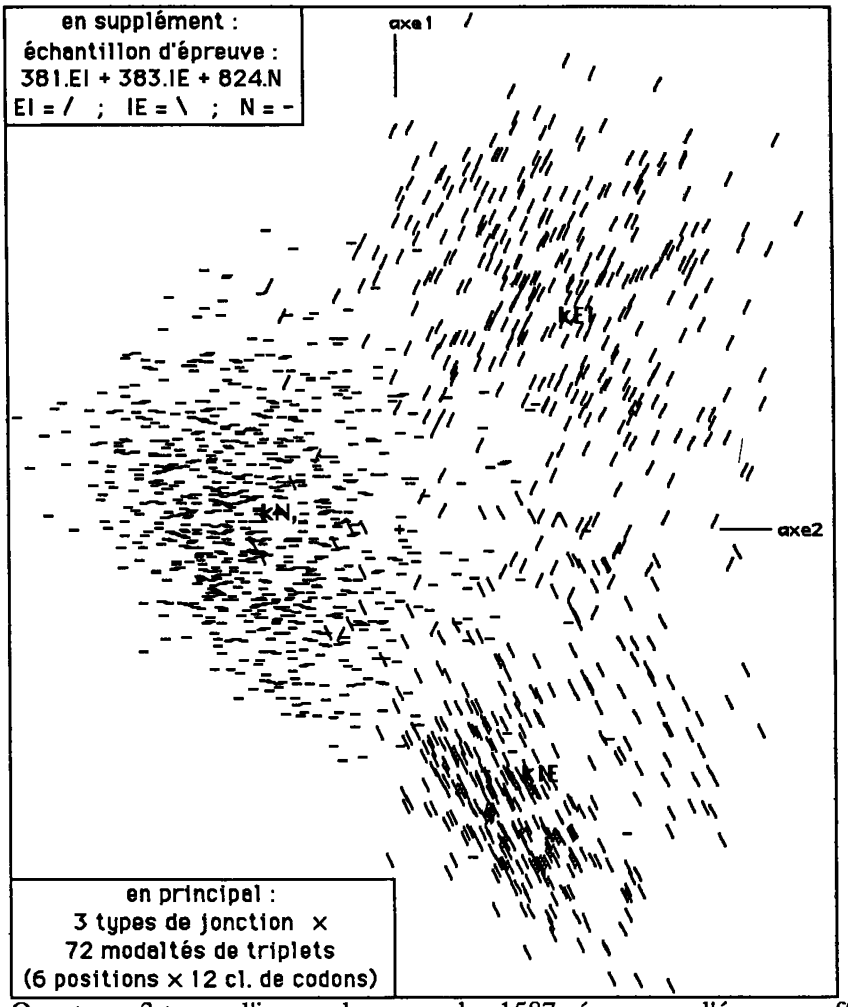
avec k(B12, IE)=1 ≈ 0, le point B12 s'écarte peu du côté supérieur;

les modalités A10, A06 et F07 contribuent fortement à l'inertie.

3 types × ({A,B,C,D,E,F} × 12)

trace	1.532e-1
rang	1 2
lambda	883 649 e-4
taux	5766 4234 e-4
cumul	5766 10000 e-4

SIGJ	QLT	PDS	INR	F 1	CO2	CTR	F 2	CO2	CTR
kEI	1000	240	387	396	636	427	300	364	333
kIE	1000	241	408	-456	802	568	226	198	191
kN	1000	519	204	29	14	5	-244	986	476



Quant aux 3 types, l'image du nuage des 1587 séquences d'épreuve offre, dans le plan (1, 2), une séparation par des zones de faible densité plus nette que celle observée dans aucune autre analyse; d'où résulte un taux maximum d'affectations exactes au centre le plus proche. Avec 48 modalités, le taux est moindre qu'avec 72; ce qu'on expliquera par le rôle des modalités A10, E07...

principal: {A,B,C,D,E,F}

	EI	IE	N
kEI	359	6	19
kIE	12	365	35
kN	10	11	770

Exact: 94,14% (1494/1587)

principal: {B,C,D,E}

	EI	IE	N
qEI	351	13	16
qIE	18	358	30
qN	12	11	778

Exact: 93,7% (1487/1587)

```
S1 x M  :  Analyse avec 1588 séquences en principal; et leurs centres, par type, en supplémentaire
trace   :    3.461e+0
rang    :       1      2      3      4      5...   10...   15...   30      45      60
lambda  :    1377   1089    843    786    724...  658...  613...  518...  442...  354 e-4
taux    :     398    315    244    227    209...  190...  177...  150...  128...  102 e-4
cumul   :     398    712    956   1183   1392... 2378... 3291... 5717... 7769... 9486 e-4

|SIGI| QLT PDS INR|  F 1 CO2 CTR|  F 2 CO2 CTR|  F 3 CO2 CTR|  F 4 CO2 CTR|
+----+-----------+--------------+--------------+--------------+--------------+
| kEI| 898 240  17| -338 461 199|  111  50  27|  292 345 243| -103  42  32|
| kIE| 921 241  18|  421 682 310|  241 224 129|  -11   0   0|  -62  15  12|
| kN,| 846 519   9|  -39  26   6| -164 443 127| -130 281 104|   76  96  38|
```

2.2.4 Affectation à la classe de l'individu de base le plus proche

Le tableau ci-dessus atteste qu'à la différence de ce qu'on a obtenu au §1.1 dans l'espace engendré par les 36 axes, on a ici, en prenant S1 x M pour tableau principal, une médiocre représentation des centres dans l'espace engendré par les premiers des 66 axes factoriels . On expliquera par cela le taux de discrimination médiocre par affectation, dans cet espace, à la classe de l' individu de base le plus proche.

codage des séquences avec, en principal, 72 modalités de M: {A,B,C,D,E,F} x 12
affectation des 1587 individus d'épreuve à un ensemble d'individus de base

	EI	IE	N
pvEI	325	9	64
pvIE	30	342	76
pvN	26	31	684

exact: 1351/1587=85,13%
espace de dimension 36
issu de l'analyse de S1 x M

	EI	IE	N
vqEI	333	7	22
vqIE	12	355	26
vqN	36	20	776

exact: 1464/1587=92,25
plan issu de l'analyse des 3 centres: 3 x 72

	EI	IE	N
vrqEI	349	14	16
vrqIE	5	349	17
vrqN	27	19	791

exact: 1489/1587=93,82%
vq: affectation à S1 ; vrq: affectation à S1 réduit

Au contraire, on a des taux proches du meilleur (obtenu au §2.2.3) si l'affectation est faite au type de jonction du plus proche individu de base; parcticulièrement si cet ensemble est réduit comme on l'explique au §2.2.2.

3 Conclusion: logique discrète et représentation spatiale

Une note trouvée dans la base de données propose, pour la discrimination des types de jonction, un algorithme fondé sur des définitions et implications écrites dans le langage "Prologue". Ainsi, selon l'auteur de la note: "sont correctement identifiés 40% des IE, 4% des EI et 99% des N; 48 exemples sont indûment affectés à IE, mais aucun ne l'est à EI..." (l'affectation du reste des cas tombant dans N...).

L'analyse multidimensionnelle, ainsi que divers algorithmes (cf. TOWELL et coll, cités dans [EXON INTRON]), obtiennent des taux de reconnaissance bien meilleurs. Pourtant, l'observation attentive des séquences, dans [EXON INTRON], puis dans le présent article, découvre une forte structure logique, marquée par des cases nulles dans des tableaux de croisement. Mais pour coordonner en une description globale les multiples facettes de cette structure, le continuum spatial réussit mieux que les systèmes orthogonaux d'axes logiques fixés une fois pour toutes.

Et l'expérience du présent article conduit à prôner l'affectation à un sous-ensemble de l'échantillon de base, réduit par affectation à lui-même.