

C. BOURGARIT

## Propriétés combinatoires du code génétique

*Les cahiers de l'analyse des données*, tome 4, n° 2 (1979),  
p. 211-217

[http://www.numdam.org/item?id=CAD\\_1979\\_\\_4\\_2\\_211\\_0](http://www.numdam.org/item?id=CAD_1979__4_2_211_0)

© Les cahiers de l'analyse des données, Dunod, 1979, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

## PROPRIÉTÉS COMBINATOIRES DU CODE GÉNÉTIQUE [CODE GEN]

par C. Bourgarit (1)

Les abondantes données qu'accumule de nos jours la biochimie semblent offrir matière à des analyses statistiques multidimensionnelles. Tandis que certaines études (comme celles de G. Flamenbaum sur la conformation des protéines cf Cahier Vol IV, n° 3; à paraître) doivent tenir compte de la disposition même des atomes et des forces de valence, d'autres ne retiennent des formules démontrées au terme de recherches aussi patientes qu'ingénieuses, qu'un aspect en quelque sorte linguistique. A ces dernières études, la présente note voudrait initier le statisticien (§ 1) tout en présentant quelques remarques dont la forme intéressera peut-être le biologiste (§ 2).

### 1 Les protéines et leur synthèse

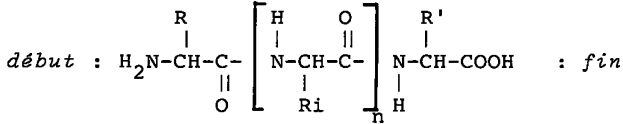
1.1 Structure primaire des protéines : Plus de la moitié de la masse de tout être vivant est constituée par de l'eau ...; certains ions ( $\text{Na}^+$ ,  $\text{K}^+$ ...) jouent un grand rôle; le squelette des vertébrés se distingue par une forte minéralisation. Mais la caractéristique chimique des êtres vivants est la présence de molécules complexes aux longues chaînes d'atomes de carbone : les *molécules organiques*. Parmi celles-ci prédominent les *protéines* aux rôles multiples : enzymes (catalyseurs), hormones (e.g. insuline), et protéines de soutien (fonction mécanique plutôt que chimique).

L'élément constitutif des protéines est l'*acide  $\alpha$ -aminé* dans sa configuration spatiale L (sur la structure spatiale des molécules, cf G. Flamenbaum ; *loc. cit.*) : toute protéine est premièrement une séquence de plusieurs dizaines d'acides aminés (pour une séquence de 10, on emploiera plutôt le terme de *peptide* que celui de protéine) ; si l'on fait abstraction de la disposition spatiale (fort complexe parfois, et toujours essentielle à la fonction) la formule d'une protéine est donc comme un *mot* (fort long) dont les lettres sont prises dans un *alphabet* qui n'est autre que l'ensemble des 20 acides aminés intervenant généralement dans la constitution des êtres vivants : on dit que cette formule décrit la *structure primaire* des protéines, par opposition aux enroulements, replis et entrelacs spatiaux... Outre ces 20 aminoacides dont l'assemblage est régi par le code génétique, plusieurs autres (e.g. l'hydroxyproline) se rencontrent dans des protéines et des peptides, auxquels ils sont incorporés par des réactions particulières.

Chimiquement, les acides aminés sont tous compris dans une formule unique  $\text{R-CHNH}_2\text{-COOH}$  ; formule où la lettre R représente un groupement d'atomes (ou *radical*) qui peut être réduit à un seul atome d'hydrogène H pour l'acide aminé nommé *glycine* ; est un méthyle ( $\text{CH}_3$ ) dans l'*alanine* ; un groupement avec 7 atomes de carbone ( $\text{CH}_2\text{-C}_6\text{H}_5$ ) dans la *phénylalanine* ; etc. Du point de vue linguistique qui est le nôtre, il faut noter que la

(1) Assistante au laboratoire de statistique. Université Pierre et Marie Curie

molécule d'acide aminé a deux pôles : acide COOH et amine (base) NH<sub>2</sub> ; en sorte que chaque maillon de la chaîne protéique est orienté ce qui définit pour chacun de ces *mots* un sens de parcours fixé par convention de la base vers l'acide : le COOH du n-ème acide aminé (la n-ème lettre) se liant au NH<sub>2</sub> du n+1-ème (avec élimination d'une molécule d'eau) pour former une liaison *peptidique* : d'où le schéma :



Vingt acides aminés : cela fait moins qu'il n'y a de lettres : on a donc pu convenir d'un système de symboles à une lettre : G pour glycine ; A pour alanine ; F pour phenylalanine ; ainsi la formule d'une protéine s'écrit véritablement comme un mot. Toutefois des sigles à trois lettres : GLY pour glycine, ALA pour alanine, PHE pour phenylalanine... sont plus faciles à lire : dans une formule écrite avec ces symboles, c'est le sigle à lettres qui constitue véritablement l'élément.

1.2 Les acides nucléiques : D'autres molécules, moins importantes en masse que les protéines mais essentielles à la vie, sont comme les protéines décrites par des formules séquentielles suivant un alphabet élémentaire, dont la structure est encore plus simple car il ne compte que quatre symboles.

Les *acides nucléiques* sont des chaînes alternées d'éléments : sucre-phosphate-sucres-phosphate-sucres-phosphate... le sucre portant une molécule distinctive - une *base* qui est proprement ce que nous appelons la lettre. Le terme de sucre peut égarer le lecteur qui songe au glucose au saccharose et au fructose... : il s'agit ici de deux molécules particulières (rentrant dans la famille des sucres) qui peuvent être soit le *ribose* soit le *désoxyribose* : dans certains acides nucléiques le sucre est uniquement le *ribose* on parle alors d'acide ribo-nucléique : en abrégé a.r.n. ; dans les autres le sucre est uniquement le *désoxyribose* : ce sont les acides desoxyribonucléiques : en abrégé a.d.n. . Le terme *nucléique* s'explique parce que les acides ont été découverts dans les noyaux des cellules ; plus précisément l'a.d.n. prédomine dans les chromosomes, que l'on considère comme une sorte de formule et de programme reçu des parents (patrimoine héréditaire) : nous devons en reparler (§ 1.3).

Les quatre bases (ou petites molécules organiques porteuses d'un groupement *amino*) sont

pour l'a.r.n. : uracyle ; adénine ; cytosine ; guanine : U.A.C.G. ;

pour l'a.d.n. : thymine ; adénine ; cytosine ; guanine : T.A.C.G. .

Il est reconnu que dans les cellules l'a.d.n. peut servir de modèle - de formule - pour la constitution d'un a.r.n. : en particulier l'a.r.n. messager, qui lui-même sert de formule pour la synthèse *in vivo* (chez les vivants) des protéines (cf *infra*), est une copie de segments de chromosomes. Les processus de copie et de codage prédominent dans la biochimie cellulaire telle que l'ont découverte les recherches faites après 1950 : toutefois les règles de copie ne sont pas toujours simples : un a.r.n. messager peut résulter de la copie de segments non consécutifs de chromosomes, éventuellement répétés : il y a dans les chromosomes comme un dictionnaire, une réserve de *paradigmes* dont l'a.r.n.m. résulte par une *syntaxe* encore inconnue.

Pour les diverses opérations de copie de codage, ou de traduction biunivoque, effectuées à partir des acides nucléiques, on utilise les termes suivants :

*réplication* : décodage d'un a.d.n. pour produire un a.d.n. complémentaire, suivant la règle suivante :

A → T ; T → A ; G → C ; C → G

*transcription* : décodage d'un a.d.n. pour produire un a.r.n. suivant la règle :

A → U ; T → A ; G → G ; C → G

*traduction* : décodage d'un a.r.n.m. pour produire une protéine : les règles de cette traduction constituant le code génétique (cf *infra* § 1.3; § 2).

1.3 La synthèse des protéines chez le vivant : Des techniques très diverses : cristallographie aux rayons X ; microanalyse chimique (chromatographie) ; marquage radioactif ; microscopie électronique... ont permis à la cytologie et à la biochimie de démontrer le schéma suivant qui ne semble pas être contesté par les spécialistes, même s'il pose de nombreuses questions (cf *supra* : rapports entre a.d.n. et a.r.n.m.).

En bref nous distinguerons : la *matière*, son *vecteur*, la *formule*, l'*instrument*.

La *matière* : c'est tout simplement les acides aminés eux-mêmes présents dans la cellule vivante ; et destinés à être assemblés par elle en longues chaînes ou protéines.

Le *vecteur* : l'imprimeur de jadis avait devant lui des lettres de plomb bien rangées dans une casse, et prenait ces lettres à la pince : dans le cytoplasme de la cellule les acides aminés sont dispersés ; mais chacun peut s'attacher à une molécule spécifique rentrant dans la classe définie au § 1.3 : l'a.r.n. de transfert, a.r.n.t. . Cette molécule joue le rôle d'une *pince*, et permet aussi par sa *forme* différente pour chaque acide aminé, d'*étiqueter* en quelque sorte ceux-ci.

La *formule* : elle est, nous l'avons dit, matérialisée par une molécule d'a.r.n. : l'a.r.n.m. : de façon précise, pour passer de l'a.r.n.m. à la protéine, les bases sont à lire trois par trois, un triplet de bases définissant un acide aminé : par exemple le triplet de bases GCU signifie ALA (ou A = alanine) ; le triplet UUU signifie PHE ( ou F = phénylalanine) ; GCUUUU... signifie donc alanine + phénylalanine + etc. Cette correspondance entre triplets de bases et acides aminés est ce qu'on appelle le *code génétique*. Il est particulièrement remarquable que ce code soit commun à tous les êtres vivants : animaux ou végétaux ; unicellulaires (e.g. bactéries) ou pluricellulaires. En revanche la formule des a.r.n.t. dépend de l'espèce considérée.

L'*instrument* : l'assemblage des acides aminés s'effectue par de petits organites cellulaires (visibles seulement au microscope électronique) les ribosomes, (eux aussi constitués en majeure partie d'a.r.n.); les ribosomes souvent groupés en séquence, suivent de proche en proche la chaîne de l'a.r.n.m., en agrafant des acides aminés, présentés par leur vecteur spécifique l'a.r.n.t., et reconnus conformes aux triplets successifs de l'a.r.n.m. précisément grâce à leur étiquetage par l'a.r.n.t. .

Répétons-le ce merveilleux mécanisme de synthèse devra être éclairé par de nombreuses recherches : par exemple de cinétique chimique... ; mais ce qui est déjà connu, par sa structure formelle intéresse grandement le statisticien. Plus particulièrement, nous considérerons ici le *code génétique*.

2 Etude du code g n tique

2.1 Le tableau de correspondance entre prot ines et bases : Puisqu'il y a quatre bases {U,C,A,G}, il y a  $4 \times 4 \times 4 = 64$  triplets. Ces triplets ont  t  r partis en 21 significations : les 20 acides amin s constitutifs des prot ines usuelles des cellules vivantes ; et un marquant d'extr mit  de s quence comparable   un signe de ponctuation, et not  TAU. Le tableau ci-joint a 21 lignes et 12 colonnes : les lignes sont les 20 acides amin s et le TAU, pour lesquels on a donn  en marge un nom complet, suivi des abr viations en une et trois lettres ; les colonnes sont les quatre bases U,C,A,G qu'on a r p t es trois fois suivant les trois positions 1, 2, 3 qu'elles peuvent occuper dans un triplet. Reste   expliquer les chiffres inscrits dans le tableau :   l'intersection de la ligne LEU et de la colonne U2 on lit : 6 ; cela veut dire qu'il existe 6 triplets de bases dont la signification est leucine, et comportant la base uracyle en position 2 ;   l'intersection de la ligne SER et de la colonne C3 on lit 2 : il y a deux triplets signifiant s rine et dont la troisi me base est la cytosine ; etc.

*A priori* il n'est pas certain que ce tableau (qu'on peut appeler tableau *cumulatif* du code) permette de d terminer sans ambigu t  le code g n tique nous verrons pourtant que tel est le cas. En fait, le tableau a  t  construit parce qu'il se pr te   l'analyse de correspondance ; et les r sultats de l'analyse (expos s ailleurs cf [Corr. Code]) et particuli rement l'introduction   ce probl me ont attir  notre attention sur les particularit s du code expos es au   2.2.

	U <sub>1</sub>	C <sub>1</sub>	A <sub>1</sub>	G <sub>1</sub>	U <sub>2</sub>	C <sub>2</sub>	A <sub>2</sub>	G <sub>2</sub>	U <sub>3</sub>	C <sub>3</sub>	A <sub>3</sub>	G <sub>3</sub>
Tryptophane : W : TRP	1							1				1
Isoleucine : I : ILE			3		3				1	1	1	
Tyrosine : Y : TYR	2						2		1	1		
Phenylalanine : F : PHE	2				2				1	1		
Proline : P : PRO		4				4			1	1	1	1
Leucine : L : LEU	2	4			6				1	1	2	2
Valine : V : VAL				4	4				1	1	1	1
Methionine : M : MET			1		1							1
Lysine : K : LYS			2				2				1	1
Cysteine : C : CYS	2						2		1	1		
Alanine : A : ALA				4	4				1	1	1	1
Arginine : R : ARG		4	2				6		1	1	2	2
Threonine : T : THR			4			4			1	1	1	1
Serine : S : SER	4		2		4		2		2	2	1	1
Glycine : G : GLY				4			4		1	1	1	1
Histidine : H : HIS		2					2		1	1		
Aspartic acid : D : ASP				2			2		1	1		
Asparagine : N : ASN			2				2		1	1		
Glutamic acid : E : GLU				2			2				1	1
Glutamine : Q : GLN		2					2				1	1
Terminaison : Z : TAU	3						2	1			2	1

Tableau K<sub>1J</sub>

2.2 Particularit s du code

2.2.1 Pr dominance des deux premi res bases : Consid rons d'abord le bloc des quatre premi res colonnes, correspondant   la premi re base du triplet. Dans presque chaque ligne, il y a un seul nombre non-nul, et trois z ros ; font seules exception les lignes LEU, ARG, SER qui n'ont que deux z ros. Prenons l'exemple de la deuxi me ligne : ILE : il y a trois triplets qui signifient isoleucine, mais tous ont pour premi re base A, l'ad nine ; aucun triplet en U1. C1, ou G1 ne peut signifier ILE ; etc.

Le bloc des quatre colonnes suivantes présente une structure analogue : ici encore dans presque chaque ligne, il y a un seul nombre non nul et trois zéros ; les exceptions sont SER et TAU. Mettons donc à part les quatre lignes LEU, ARG, SER et TAU. Il reste 17 lignes (sur 21) présentant la propriété suivante : les triplets de bases associés à l'acide aminé correspondant ont tous même première base et même deuxième base : en d'autres termes l'acide aminé détermine (sauf exceptions) les deux premières bases des triplets qui le signifient ; seule la troisième base est plus ou moins indéterminée.

De façon précise : deux acides aminés sont représentés par un seul triplet :

TRP = UGG ; MET = AUG ;

neuf acides aminés sont représentés par deux triplets, dont les deux premières bases sont déterminées :

TYR = UA{UC} ; PHE = UU{UC} ; LYS = AA{AG} ;

CYS = UG{UC} ; HIS = CA{UC} ; ASP = GA{UC} ;

ASN = AA{UC} ; GLU = GA{AG} ; GLN = CA{AG} ;

un acide aminé est représenté par trois triplets, dont les deux premières bases sont déterminées :

ILE = AU{UCA}

cinq acides aminés sont caractérisés par les deux premières bases , la troisième pouvant être quelconque :

PRO = CCX ; VAL = GUX ; ALA = GCX ; THR = ACX ; GLY = GGX ;

Dans les formules ci-dessus, on a noté X la possibilité d'une quelconque des quatre bases pour la position 3 ; et indiqué entre accolades la possibilité d'alterner entre deux ou trois bases dans cette même position.

On voit que dans cinq cas, le sens est totalement défini par les deux premières bases, la lecture de la troisième permet de choisir entre deux éventualités qui subsistent seules : par exemple après CA si on lit U ou C le sens est HIS et si on lit A ou G le sens est GLN. Le remplacement de U3 par C3 ou C3 par U3 ne change *jamaïs* le sens d'un triplet (on le vérifiera complètement en examinant le reste du code : § 2.2.2) ; le remplacement de A3 par G3 ou G3 par A3 n'importe au sens que dans quatre cas. C'est pourquoi on a dit que les deux premières bases prédominent.

Jusqu'ici, la lecture du tableau de correspondance cumulant les codes, nous a permis de retrouver tous les triplets correspondant à 17 acides aminés ; on verra que les trois autres aminoacides et le TAU sortent aussi sans confusion du même tableau.

**2.2.2 Les synonymes en deux pavés** : Pour 17 sens sur 21, les triplets correspondants forment ce qu'en terme mathématique on peut appeler un *pavé* c'est-à-dire un produit de trois parties finies de l'alphabet des bases UCAG : en fait les deux premières parties (intervalle de la première et de la deuxième coordonnée) sont réduites à un seul élément (les coordonnées 1 et 2 sont déterminées!). Pour les quatre sens exceptionnels on montre qu'il suffit de deux pavés, que le tableau de correspondance nous permet de retrouver :

TAU : trois triplets commençant par U1 : pour les deux dernières bases on a le tableau



dans ce tableau à quatre cases, trois seulement sont occupées (réalisées) ; on sait qu'il y a deux bases dans la colonne A3, deux dans la ligne A2 ; une seule dans la colonne G3 et la ligne G2. Ceci suffit à conclure.

$$\text{TAU} = \{\text{UAA}, \text{UAG}, \text{UGA}\} = \text{UA}\{\text{AG}\} + \text{UGA} ;$$

LEU : six triplets dont la base médiane est U2 ; pour la première et la troisième base on a :



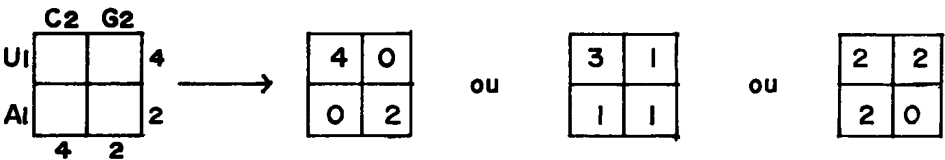
tous les triplets commençant par C1 U2 doivent être possibles ; quant aux deux triplets en U1 U2 leur troisième base ne peut être que A3 ou G3 ; donc :

$$\text{LEU} = \text{CUX} + \text{UU}\{\text{AG}\} ;$$

Le cas de ARG est tout analogue à celui de LEU ; on a :

$$\text{ARG} = \text{CGX} + \text{AG}\{\text{AG}\} ;$$

Le cas de SER est le plus complexe, parce qu'il y a ambiguïté sur chacune des quatre bases des triplets qui le signifient : ici encore on aura recours à un tableau carré symbolisant les combinaisons permises pour les deux premières bases :



à chaque case de ce tableau il peut *a priori* correspondre un nombre quelconque de vingt triplets acceptables : de zéro à 4... ; toutefois on sait que s'il y a quatre triplets en U1C2 il n'y en aura aucun en A1 C2, ni en U1 G2 parce qu'il n'y a que quatre triplets en U1 et quatre triplets en C2 : du point de vue du remplissage des cases du tableau carré, on aboutit donc aux trois possibilités que nous avons figurées. C'est la première qui est à adopter : car dans le reste du code aucun triplet U1 C2 n'a été rencontré : il faut donc qu'ils soient tous ici pour signifier SER. Dès lors on conclut sans peine :

$$\text{SER} = \text{UCX} + \text{AG}\{\text{UC}\} ;$$

ce qui achève la lecture du code d'après le tableau cumulatif.

On notera que chaque fois que plusieurs bases sont possibles en position 1 ou 2 pour les triplets correspondant à un acide aminé donné ou le TAU (et cette ambiguïté n'existe que dans les quatre cas qu'on vient d'examiner en détail) il n'y a toutefois jamais compatibilité entre G1 et une autre base en position 1 ; ni entre U2 et une autre base en position 2. On peut ainsi partager les 21 sens en quatre classes :

VAL : toujours codé par G1 et U2 : VAL = GUX

{ALA, GLY, ASP, GLU} : toujours codés par G1 et par  $\bar{U}2$  (i.e. non U2 : un deuxième symbole qui n'est pas U2)

{ILE, PHE, LEU, MET} : toujours codés par  $\bar{G}1$  et par U2 ;

et les douze autres (11 acides aminés et le TAU), toujours codés par  $\bar{G}1$   $\bar{U}2$  (i.e. sans G en position 1, ni U en position 2).

Cette particularité ressort immédiatement de l'analyse de correspondance du tableau cumulatif des codes (cf [Corr.Code] : on verra dans l'introduction à ce problème, comment s'enchaînent les propriétés du code, les particularités des résultats de l'analyse et les questions mathématiques).

*N.B.* : dans ce raisonnement, on suppose que tous les triplets ont été attribués, chacun une fois et une seule.

2.3 Hasard ou nécessité : Un illustre biologiste s'interrogeait sur le rôle du hasard et de la nécessité dans la genèse de la vie. La structure si particulière du code génétique diffère grandement de ce qu'on attendrait du tirage au hasard de la correspondance entre 64 triplets et 21 significations : il y a là quelque profonde nécessité. D'ailleurs la similitude entre les codes d'acides semblables (tels ASP et GLU ; ALA et GLY) frappe également :

ASP : GA{UC} ; ALA : GCX

GLU : GA{AG} ; GLY : GGX

Le code pourrait être non totalement arbitraire, mais partiellement analogique : on a suggéré de voir dans les triplets comme un moule approximatif des aminoacides qu'ils signifient (indépendamment de l'étiquetage par l'a.r.n. de transfert) (cf J. Reuben, *FEBS Let.* ; 94, 20; 1978; et J.J. Hopfield *PNAS* ; 75, 4334 ; 1978 ; cités et analysés par A. Danchin *la Recherche*, n° 97 ; Fév 1979). A la limite, il n'y aurait qu'un seul code possible.

Peut-être n'y a-t-il qu'un seul monde possible ?

والله أعلم وله التوفيق

(i.e. dans la science qu'en a Dieu tout est certain ; et c'est lui qui fait aboutir le hasard au succès)