LES CAHIERS DE L'ANALYSE DES DONNÉES

CH. MULLON
F. COLONNA

Correspondance entre structures primaire et secondaire dans les protéines

Les cahiers de l'analyse des données, tome 5, n° 1 (1980), p. 75-85

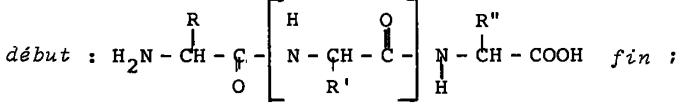
<http://www.numdam.org/item?id=CAD_1980__5_1_75_0>

Les Cahiers de l'Analyse des Données
Vol. V - 1980 - n° 1 - p. 75-85

CORRESPONDANCE ENTRE STRUCTURES PRIMAIRE ET SECONDAIRE DANS LES PROTÉINES [STRUC. PROT.]

par Ch. Mullon (1)
et F. Colonna (2)

1 *Le problème biochimique* : Une protéine est un polymère résultant de l'enchaînement d'acides aminés rentrant chacun dans la formule générale : $H_2N - \underset{R}{CH} - COOH$; selon le groupement d'atomes (radical) que représente la lettre R, on a divers acides aminés dont 20 sont à la base de toutes les structures biologiques (cf *Cahiers* Vol IV n°2 [Code Gén] § 1 ; 1979).L'enchaînement de ces molécules se fait par liaison peptidique, suivant la formule :



début : $H_2N - \underset{R}{CH} - \underset{O}{C} - \left[N - \underset{R'}{CH} - \underset{O}{C} - \right] N - \underset{H}{CH} - COOH$ *fin* ;

sur cette formule semi-développée on a indiqué suivant la convention usuelle le *début* et la *fin* de la chaîne ; le motif entre crochets (au milieu de la formule) représente une suite d'acides aminés généralement différents. Si la chaîne comprend e.g. une dizaine d'acides aminés on parle d'oligopeptide ; mais les protéines proprement dites sont composées e.g. de 50 à 300 acides aminés.

On peut écrire la formule *séquentielle* d'une protéine comme une suite de sigles à trois lettres (Gly pour glycine, Ala pour alanine etc.) représentant chacun un amino-acide (cf [Code Gén.]).Une notation plus condensée (mais moins transparente) consiste à affecter à un amino-acide une lettre unique (G pour glycine, A pour Alanine ... ; *loc. cit.* p. 214 ; tableau). Ainsi il apparaît bien que d'un point de vue linguistique, une protéine est un texte formé sur un vocabulaire de 20 lettres. C'est ce qu'on appelle la *structure primaire*.

Du point de vue biologique, on ne peut toutefois se borner à considérer la structure primaire : même si au niveau des ribosomes la protéine est synthétisée séquentiellement (par *traduction* d'un arn messager : *loc. cit.* p. 213) comme un fil souple, les interactions entre les différentes zones de ce fil produisent un repliement, aboutissant à une forme très compacte qu'on a comparée à celle d'une *pelote* . Il s'en faut toutefois de beaucoup que cette pelote soit un amas désordonné : le repliement aboutit à une forme déterminée, dont la surface présente des accidents (aspérités, concavités...) auxquelles la protéine doit sa spécificité enzymatique. C'est pourquoi après la structure primaire, on doit considérer les structures secondaire et tertiaire. De façon précise, par *structure secondaire* , on entend actuellement une configuration particulière prise par une portion de la séquence, longue de

(1) *Docteur 3° cycle. Université P. et M. Curie*

(2) *Laboratoire de Biophysique Moléculaire. Université de Nancy I.*

Le présent travail est extrait de la thèse de Ch. Mullon (Paris 3° cycle ; 1980) et porte sur des recherches effectuées en commun par F.C.&Ch.M.

quelques acides aminés ; par *structure tertiaire*, on entend la configuration spatiale d'ensemble proprement dite.

En principe, la structure globale d'une protéine, est déterminée à partir de la structure primaire par la condition de minimiser l'énergie libre. Le chimiste sait qu'outre les liaisons covalentes (liaisons peptidiques d'abord entre amino-acides consécutifs ; et aussi ponts disulfures entre deux cystéines situés à quelque distance dans la chaîne) mettant en jeu chacune plus de 50kcal/mole,il existe d'autres liaisons plus faibles (moins de 7 kcal/mole) de types divers : forces de Van der Wals, liaisons ioniques, liaisons hydrogènes. Mais on ne sait pas, pour l'instant, déterminer par le calcul la structure globale résultant de l'ensemble de ces forces qui concourent à replier un fil de structure primaire donnée. Si la structure tertiaire a pu être décrite de façon précise pour un certain nombre de protéines, c'est seulement en appliquant à des échantillons purifiés et cristallisés les méthodes usuelles de diffraction des rayons X : exploit qui valut le prix Nobel à Perutz et Kendrew !

Cependant comme la méthode cristallographique est ardue et coûteuse, certains biochimistes songent à déterminer statistiquement des relations approchées entre structure primaire, et structure secondaire. Même si la protéine forme un tout dont la structure définitive (tertiaire) ne s'explique que par l'application, l'ajustement de segments éloignés de la chaîne ; on espère qu'existent au niveau local (c'est-à-dire par exemple sur un segment formé de 5 amino-acides) des déterminismes de structure assez précis que l'analyse statistique observerait, alors que les calculs de chimie quantique ne suffisent pas à les prédire.

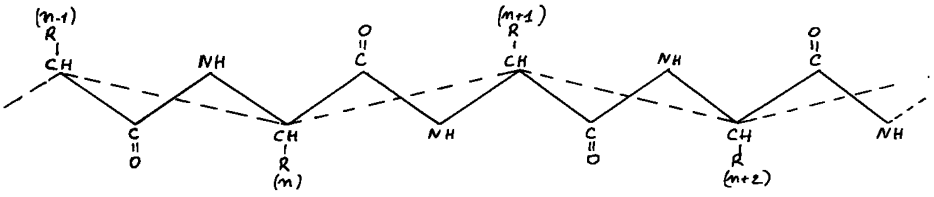
A une telle analyse est consacré le présent travail.

A la base, comme données structurelles exactes, on a choisi 19 protéines complètement décrites par la méthode cristallographique :

(Soit:Chymotrypsine, carboxypeptidase, BTPI, lysosyme, myoglobine, cytochrome C, oxydized HP, calcium binding parvalbumine, ribonuclease S, Bence Jone Immunoglobuline, carbonic anhydrase B et C, flavodoxine, ferredoxine, lambda immunoglobuline FAB...). Après avoir extrait (par la classification automatique, § 2) un schéma simplifié de la structure secondaire, on met en rapport ce schéma avec la structure primaire (§ 3) ; les correspondances observées nous paraissent s'ordonner suivant un ensemble cohérent dont certains éléments sont nouveaux ; elles ne suffisent toutefois pas à prédire une structure (§ 4).

2 *Schéma de la structure secondaire*

2.1 La séquence des angles de torsion peptidiques : Un article antérieur (cf Flamenbaum et col. ; [CONFORMATION], *in Cahiers* Vol IV n°3 ; § 2 pp 341 sqq ; 1979) explique avec figures à l'appui, la description des molécules. Sans reprendre ici cette description à laquelle nous renvoyons le lecteur, nous voulons concentrer l'analyse sur un seul point : la séquence des angles de torsion peptidiques. Redonnons la formule d'une protéine moins schématiquement qu'au § 1 :



les liens marqués de deux traits sont les liaisons peptidiques, unissant deux amino-acides successifs , entre segments de la ligne brisée NH - CHR - CO - NH - CHR - CO - NH .... (ligne en dents de scie), les angles (angles de valence) sont à peu près de 110°. La figure ci-dessus est plane ; la molécule réelle ne l'est pas : en particulier suivant chaque liaison peptidique CO - NH, les deux plans définis respectivement par le segment précédent et le segment suivant (plans CHR- CO -NH et plan CO - NH - CHR) forment un angle (noté ω dans la figure 2 de [CONFORMATION] p. 343).

Ici on s'intéressera seulement à la ligne brisée (non plate) formée de groupements CHR successifs : cette ligne n'est pas une ligne *réelle* formée de liaisons covalentes, mais une ligne *virtuelle* , en tireté sur la figure, définie par des atomes non directement liés entre eux ; l'angle de torsion correspondant à une telle ligne (i. e. l'angle entre les plans $\{CHR_{n-1}, CHR_n, CHR_{n+1}\}$ et $\{CHR_n, CHR_{n+1}, CHR_{n+2}\}$, plans qui se coupent suivant la droite CHR_n, CHR_{n+1}), sera appelé par son angle de torsion peptidique virtuel ; c'est l'angle α défini par Flory. Une convention d'orientation précise permet d'assigner à cet angle une détermination unique entre 0° et 360°. La figure plane en dents de scie correspond à la valeur 180° des angles α de torsion peptidiques (ainsi que des angles de torsion, φ et ψ affectant les autres liaisons) ; au contraire des angles de torsion de 0° (ou ce qui est égal 360°) décrivent une chaîne enroulée s'inscrivant dans un cercle.

Dès le début des études sur la structure des protéines, est apparue l'importance de deux valeurs typiques de l'angle α : 200° et 50°. Quand tous les angles α sont fixés à 200° (valeur proche des 180° de la figure) on parle de forme *feuillet* β. Tandis que la valeur 50°, caractérise l'*hélice* α : la valeur 0° ne peut être gardée sur toute la chaîne : car elle implique une superposition des atomes sur un cercle; mais avec un décalage minime, on a un empilement de spires, une hélice. Cependant entre les deux formes typiques du feuillet et de l'hélice (d'ailleurs approximativement réalisées, ainsi qu'il apparaît sur l'histogramme des valeurs de α ; cf *infra*) il y a des formes de transition : les *tournants* (en anglais *turns*) que l'analyse exposée ici tente de caractériser.

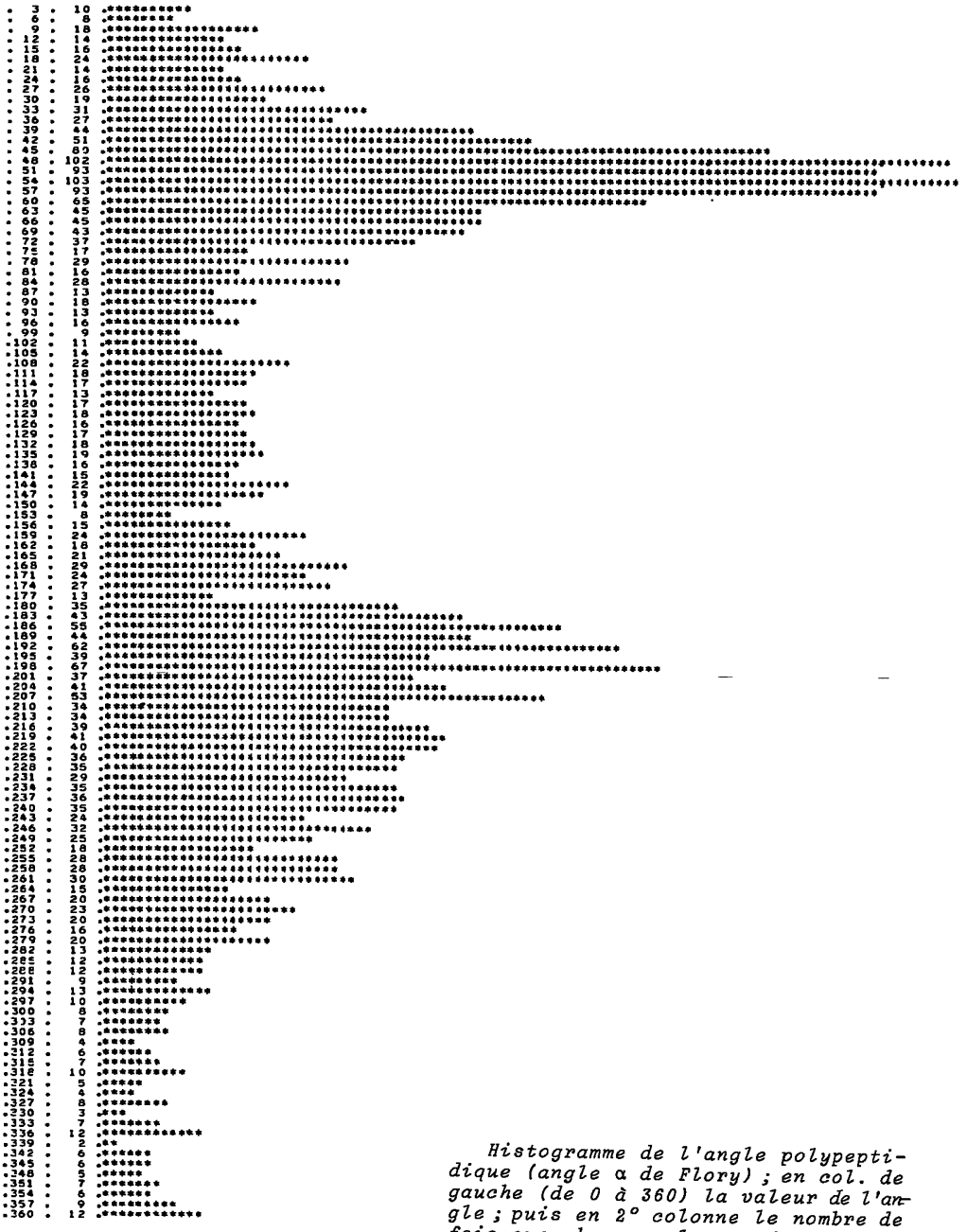
2.2 *Quatre angles pour un pentapeptide* : La distribution des valeurs individuelles de l'angle α est donnée par un histogramme : on distingue deux pics, centrés approximativement sur les valeurs typiques de 50° et 200° (ou un peu plus) ; reste à préciser comment ces valeurs se succèdent dans la chaîne. Pour cela on a considéré les sous-séquences de 4 angles α : en bref si on note 1-2-3-4-5-6-7-8.... -n la suite des n amino-acides dont est formée une protéine, on a (n - 4) pentapeptides successifs : 1-2-3-4-5 ; 2-3-4-5-6 ; 3-4-5-6-7 ; etc. dont chacun offre une suite de quatre liaisons peptidiques : donc de 4 angles α. Les 19 protéines fournissent ainsi un échantillon de 2908 quadruplets. Pour des raisons de temps-calcul la classification automatique a porté sur un sous-échantillon de 1178 quadruplets .

Pour acquérir une vue d'ensemble de ces 1178 quadruplets, on l'a soumis à l'algorithme de classification ascendante hiérarchique (avec agrégation suivant le moment d'ordre 2), après avoir défini une distance entre deux angles α et α' on prend :

$$D^2(\alpha,\alpha') = (\cos\alpha - \cos\alpha')^2 + (\sin\alpha - \sin\alpha')^2 = 2(1 - \cos(\alpha - \alpha')) ;$$

d'où pour la distance entre deux quadruplets $q = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ et $q' = (\alpha'_1, \alpha'_2, \alpha'_3, \alpha'_4)$:

$$D^2(q,q') = \Sigma\{D^2(\alpha_i, \alpha'_i) \mid i = 1, \ldots, 4\}.$$



Histogramme de l'angle polypeptidique (angle α de Flory) ; en col. de gauche (de 0 à 360) la valeur de l'angle ; puis en 2° colonne le nombre de fois que chaque valeur a été rencontrée sur les 19 protéines étudiées ici.

Numéro	Effectif	Angle 1	Angle 2	Angle 3	Angle 4
1	42	219(39)	56(14)	54(14)	55(25)
2	25	268(31)	64(46)	174(55)	96(36)
3	42	220(31)	139(34)	248(36)	209(45)
4	23	253(32)	147(32)	1(25)	201(146)
5	31	206(33)	35(30)	175(52)	209(35)
6	24	231(36)	60(33)	65(33)	240(43)
7	82	33(37)	89(58)	210(55)	212(50)
8	54	51(25)	41(40)	65(41)	211(38)
9	45	39(43)	195(43)	57(30)	67(41)
10	57	207(36)	238(41)	51(29)	78(46)
11	45	162(64)	239(54)	64(34)	217(55)
12	75	207(40)	215(33)	226(48)	45(39)
13	38	154(40)	7(49)	228(61)	52(42)
14	16	67(28)	180(44)	245(38)	77(39)
15	24	50(24)	60(39)	184(21)	55(20)
16	101	91(60)	228(40)	196(43)	227(35)
17	132	213(38)	214(30)	215(29)	201(29)
18	30	219(30)	211(30)	186(30)	286(25)
19	292	54(18)	52(14)	51(18)	52(20)

TABLEAU 1 : Les 19 classes d'une partition issue de la classification automatique des quadruplets d'angles ; après la moyenne de chaque angle, on a donné entre parenthèses son écart-type.

Classe	Effectif	Angle 1	Angle 2	Angle 3	Angle 4
1	545	51(15)	51(15)	51(16)	51(18)
2	114	40(31)	43(33)	52(34)	194(48)
3	66	49(30)	49(31)	187(47)	48(32)
4	140	42(30)	51(40)	201(52)	205(48)
5	70	36(35)	198(49)	44(34)	48(39)
6	70	45(38)	197(55)	45(38)	202(56)
7	65	46(40)	190(57)	208(55)	44(36)
8	237	44(38)	200(52)	200(43)	215(39)
9	120	206(47)	47(31)	45(30)	50(28)
10	93	220(46)	45(36)	50(41)	199(54)
11	74	190(51)	33(40)	206(55)	42(39)
12	159	210(49)	39(36)	196(54)	199(45)
13	139	209(44)	219(45)	48(34)	47(34)
14	164	210(44)	207(48)	33(37)	198(54)
15	235	201(44)	215(38)	213(45)	39(36)
16	617	202(41)	203(37)	209(35)	208(38)

TABLEAU 2 : Les quadruplets d'angles répartis en 16 classes centrées au voisinage des valeurs typiques de 50° et 200° ; comme au tableau 1, on a précisé moyennes et écarts-types.

2.3 *Résultats de la classification et interprétation schématique* : En se bornant à la partie supérieure de l'arbre issu de la C.A.H., on obtient une partition en 19 classes que nous avons caractérisées chacune par la moyenne et l'écart-type des 4 angles α. Il est satisfaisant de constater (cf tableau 1) que les moyennes sont proches des valeurs typiques de A = 50° et B = 200° et les écarts-types généralement faibles. C'est pourquoi on a décidé de considérer désormais une partition de l'ensemble de tous les quadruplets disponibles en 16 classes ; obtenue en rattachant chaque quadruplet au quadruplet de valeurs typiques A ou B dont il est le plus proche. Ainsi la classe 1 = AAAA comprend les 545 quadruplets rattachés au système (50°, 50°, 50°, 50°) qui caractérise l'hélice α ; la classe 16 = BBBB comprend 617 quadruplets rattachés au système (200°, 200°, 200°, 200°) qui caractérise le feuillet β ; la classe 9 = BAAA, comprend 120 quadruplets voisins de (200°, 50°, 50°, 50°), ce qu'on interprétera comme des débuts d'hélices α ; tandis que 8 = ABBB est la classe des débuts de feuillets β ; les tournants se signalent par des successions BA et particulièrement par des alternances : ABAB, BABA ; etc.

Il est clair qu'une description approfondie de la structure secondaire devrait tenir compte non seulement des angles α mais des autres angles de torsion (φ,ψ) ; pour ne rien dire des chaînes latérales...; l'assimilation de tout angle α aux valeurs typiques 50° (A) ou 200° (B) est abusive. Mais dans l'exploration tentée ici, on se bornera désormais à décrire la structure spatiale associée à chaque pentapeptide par une suite de 4 lettres A ou B symbolisant, répétons-le, le quadruplet des angles de torsion peptidiques virtuelles.

3 *Correspondance entre structure secondaire et primaire pour les pentapeptides*

Nous rendons compte ici successivement de deux analyses. Dans la première (§ 3.1), la structure primaire d'un pentapeptide est décrite explicitement par cinq variables, prenant chacune 20 modalités. Dans la deuxième (§ 3.2), le pentapeptide est assimilé à une séquence de 5 chiffres binaires (1 hydrophobe ; ou 0 non-hydrophobe) ; il s'agit, ici encore d'une simplification, destinée à prendre explicitement en compte les effets séquentiels en évitant toutefois la multiplicité des combinaisons.

3.1 *Correspondance entre structure secondaire et acides aminés* : Le tableau analysé comporte 16 colonnes (les 16 configurations typiques de AAAA = hélice à BBBB = feuillet) ; et 100 lignes (les 20 aminés, chacun considéré en 5 positions possibles) ; avec e.g. à l'intersection de la ligne K3 et de la colonne BABA, le nombre de fois (k(K3,BABA) = 2) que l'acide aminé K (la lysine) se trouve en position 3 dans un pentapeptide dont la structure secondaire (spatiale) a été rattachée au schéma BABA (un tournant).

3.1.1 *Calculs de contribution sur le tableau des données* : L'examen direct de ce tableau de contingence peut induire une erreur, car il ne suffit pas de remarquer un nombre élevé ou faible : il faut rapporter ce nombre au produit du poids de sa ligne (i.e. de la fréquence de l'acide aminé considéré) par celui de sa colonne (fréquence de la figure spatiale) et se garder des fluctuations d'échantillonnage (cf *infra* § 3.1.2). En revanche après division du tableau par son total, il est commode de considérer, (comme il est classique en a. des corr.) les quantités $CTR(i,j) = (f_{ij} - f_i f_j)^2 / f_i f_j$ (dont la somme est la trace $\Sigma\lambda_\alpha$) : ces quantités ont été tabulées (cf Ch. Mullon, thèse) en les affectant du signe de la différence $(f_{ij} - f_i f_j)$. Ainsi une

valeur élevée positive de CTR(X_p ,j) signale qu'une position p de l'acide aminé X induit la configuration j ; une valeur nettement négative exprime un rejet ; tandis qu'au voisinage du 0 est l'indifférence.. Tous les résultats lus cas par cas dans la table des CTR(i,j) apparaissent d'ailleurs dans leur ensemble sur les plans issus de l'analyse factorielle (qui montrent encore d'autres résultats) : mais la lecture des CTR(i,j) a le mérite de nous assurer que les proximités apparues sur les graphiques ne sont pas fallacieuses.

Sans reproduire ici les tables, nous énumérons les effets observés ; et publions ensuite les plans 1 × 2 et 3 × 4 issus de l'analyse factorielle. Voici d'abord, par groupes successifs d'acides aminés, ce qu'on lit sur les tables.

Vis-à-vis des structures AAAA (hélice) et BBBB (feuillet) les acides aminés C, D, N, P, G ne manifestent d'affinités en aucune position. En revanche il y a des liens notables entre (D1,BABA) ; (D2,ABAB) ; (N1,BAAB) ; (N2,BBAA) ; (N5,BBBA) ; (N3,BABB) ; (N4,BBAB) ; (G3,BABB) ; (G4,BBAB) ; (G5,AAAB) ; (G5;BBBA).

Les effets les plus significatifs pour ces cinq premiers acides étant l'affinité de la proline P, avec un certain nombre de structures de tournant.

(P2,BAAA) ; (P3,BBAA) ; (P3,ABAA) ; (P4,AABA) ; (P5,AAAB).

Les acides aminés (A,M,L,F,I,V,W,Y), généralement hydrophobes (cf *infra* § 3.2) apparaissent très fréquemment en toute position dans les structures pures AAAA ou BBBB. Plus précisément, certains acides discriminent ces deux structures : l'alanine A, ou la méthionine M sont plus hélicogènes ; tandis que la valine V et l'isoleucine I sont liées au feuillet. Le tryptophane W a un rôle particulier : indifférent à la plupart des structures, c'est un important initiateur d'hélices ; il en est de même de la leucine L.

Aucun effet ne peut être attribué à Q, R, H.

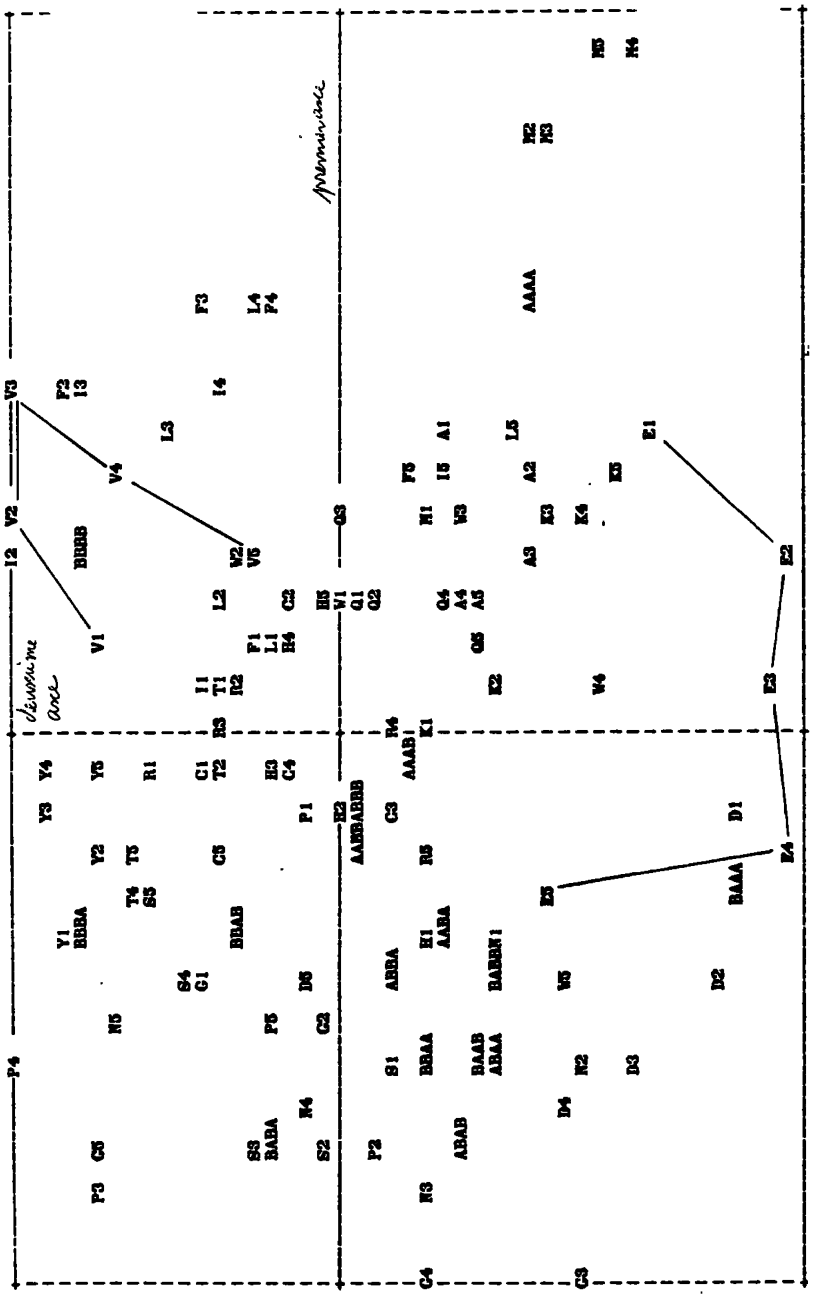
La lysine K , et l'acide glutamique E , sont reliés respectivement aux fin et début d'hélice. On notera en particulier la croissance des coefficients CTR(K_p ,AAAA) de p = 1 à 5 (position initiale à position finale) ; et de même la décroissance des CTR(E_p ,AAAA) de p = 1 à 5.

Enfin l'acide glutamique Q tend à briser la configuration BBBB.

3.1.2 *Résultats de l'analyse factorielle* : Considérons d'abord la suite des valeurs propres et des taux :

rang α	1	2	3	4	5	....	Total
λ_α	0,050	0,028	0,023	0,022	0,017	....	0,217
τ_α	23%	12,7%	10,5%	10,1%	7,9%	....	100%

Les premières valeurs propres sont bien détachées avec des pourcentages d'inertie élevés. Mais les valeurs propres sont faibles. Il est difficile d'appliquer une épreuve de validité : toutefois si l'on considère qu'il s'agit d'un tableau de contingence (bien qu'en fait les faits recensés ne soient pas indépendants entre eux), on se fonde sur ce que le total du tableau de correspondance, multiplié par la trace est double du nombre des cases pour conclure que la moitié de l'inertie est susceptible d'être interprétée (sur ces calculs, liés à



Correspondance entre 20 acides aminés, considérés dans les cinq positions d'un pentapeptide, et les seize structures secondaires-types de ce pentapeptide (cf § 3.1).

l'épreuve du χ^2, cf e.g. [Epr. Val.] TIE B n° 8 § 1). La cohérence de l'interprétation est d'ailleurs assurée pour le plan 1 × 2 ; et le plan 3 × 4 présente des relations qu'on considérera au moins comme des suggestions intéressantes. Mais même si on se satisfait de la validité (que des analyses ultérieures fondées sur un ensemble plus vaste de données pourront confirmer), les faibles valeurs propres, indiquent de faibles différences de profils, c'est-à-dire une liaison peu marquée entre les acides aminés d'une part, et la structure spatiale du pentapeptide où ils s'insèrent d'autre part : ce qui interdit de prédire une structure secondaire à partir de la structure primaire, d'après les tableaux considérés ici (cf *infra* § 4).

Sur l'ensemble J des 16 configurations, le plan 1 × 2 montre une structure triangulaire : d'une part (F1 > 0) les deux configurations pures AAAA (hélice) et BBBB (feuillet) qui s'opposent sur l'axe 2; ces deux points apportent à eux-seuls plus de 60% de l'inertie du plan 1 × 2 ; et la somme de leurs CO2 - cos2 - avec les axes 1 et 2 est de 921 millièmes pour AAAA et de 806 millièmes pour BBBB) d'autre part (F1 < 0), les autres configurations, d'autant plus écartées du côté F1 < 0, qu'elles sont plus hétérogènes (ABAB ou BABA = tournants) . On notera qu'à l'exception de BBBA (fin de feuillet) et BAAA (début d'hélice) aucune configuration (du côté F1 < 0) ne s'écarte sur l'axe 2 (positif ou négatif).

Quant aux acides aminés on a sur l'axe 1, avec les structures AAAA et BBBB, les acides aminés hydrophobes déjà signalés I,Y,F,M,L (cf § 3.1.1) ; et à l'opposé les petites molécules G,N,D,S,P. Dans le demi-plan F1 > 0, les amino-acides hydrophobes se répartissent sur l'axe 2 suivant leurs affinités avec AAAA ou BBBB. A chacun des amino - acides il correspond dans le plan 1 × 2 un chapelet de cinq points (5 positions) qui apparaît régulier et localisé. La faible amplitude de la trajectoire s'explique par la prédominance des structures AAAA et BBBB qui toutes deux peuvent se prolonger de part et d'autre de chaque position. L'orientation de la trajectoire peut suggérer des effets de position, tel celui déjà signalé (cf § 3.1.1) de l'acide glutamique E pour induire un début d'hélice.

Dans le plan 3 × 4, on note l'affinité de quelques acides aminés (proline P ; asparagine N ; acide aspartique D) pour les structures pliées (tournants) ainsi que l'équivalence conformationnelle, pour ce type de structure, entre la proline en position i + 1 et l'acide aspartique et l'asparagine en position i.

3.2 *Correspondance entre structure secondaire et structure primaire*

résumée : On souhaiterait connaître l'effet sur la structure secondaire des séquences d'acides aminés (et non plus seulement, comme ci-dessus, des acides aminés considérés individuellement). Mais avec 20 amino-acides il y a $(20)^5 = 3{,}2\ 10^6$ pentapeptides possibles! Il est exclu de les considérer tous individuellement. On a donc tenté de résumer par un nombre binaire à 5 chiffres la structure primaire d'un pentapeptide, en notant 1 = hydrophobe ; 0 = non-hydrophobe ; et e.g. 00101 un pentapeptide présentant deux autres hydrophobes en position 3 et 5. En effet la configuration prise pour une protéine dépend des interactions des acides aminés (plus exactement de leur radical R) avec les molécules (polaires) du solvant aqueux. La configuration d'énergie minima, les résidus R polaires à l'extérieur de la protéine , et ceux hydrophobes à l'intérieur : d'où la théorie de Lim (1976) le principal mécanisme du repliement des protéines, repose sur la tendance de celles-ci à constituer des groupes compacts d'acides hydrophobes.

Toutefois la classification des acides aminés reste controversée:

ici on a considéré comme hydrophobes les 8 acides suivants : valine , leucine, isoleucine, méthionine, cystéine, tryptophane, phénylalanine, tyrosine.

On a construit un tableau de contingence 32 × 16 croisant l'ensemble des nombres binaires à 5 chiffres (0,1) (lesquels représentent pour nous des structures primaires résumées) avec l'ensemble des mots de quatre lettres (A, ou B) (lesquels sont une description approchée de la structure spatiale). Aussi on lit e.g. à l'intersection de la ligne 00100 et de la colonne BBBA le nombre k(00100, BBBA) = 12 des pentapeptides de notre échantillon,qui comportent un seul amino-acide hydrophobe, en position 3, et ont une structure de fin de feuillet (trois angles de torsion voisins de 200° ; puis un dernier, voisin de 50°).

Avec une trace de 0,45 et une première valeur propre de 0,15 , cette analyse fournit des résultats significatifs statistiquement parlant. Mais l'interprétation est peut-être plus délicate que pour l'analyse précédente : pour comprendre l'association entre BBBB et 10101 (ou AAAA et 10110), il faut se représenter la disposition spatiale des acides aminés hydrophobes, qui dans l'un et l'autre cas se rapprochent, conformément à la théorie de Lim.

4 *Essais de prédiction de structure spatiale* : Le problème de la prédiction de la structure spatiale à partir de la structure séquentielle (primaire) a donné lieu en biochimie à une abondante littérature (cf e.g. G.E. Schulz et R.H. Schirmer *Principles of protein structure* ; chap. 6 ; Springer 1979) : comme on l'a dit au § 1, le problème est d'une grande importance, même si la solution en semble encore inaccessible.

L'analyse factorielle offre ici un cadre géométrique intéressant. Comme précédemment, bornons-nous aux pentapeptides; la structure géométrique spatiale étant réduite à l'ensemble J des 16 configurations (de AAAA à BBBB). Soit I un ensemble de caractères que peut présenter un pentapeptide du point de vue de sa structure séquentielle (primaire) : par exemple cf § 3.1, avoir tel acide aminé en telle position ; ou encore cf § 3.2, avoir un ensemble d'hydrophobes et de non-hydrophobes disposés d'une façon déterminée (dans les essais de prédiction rapportés ici, on considère un tableau 132 × 16, obtenu en superposant le tableau 100 × 16, du § 3.1 ; et le tableau 32 × 16 du § 3.2 : ce dernier étant affecté du poids 2, compte-tenu de son intérêt particulier; ce qui implique que dans les colonnes supplémentaires p, cf *infra*, les 32 derniers éléments soient multipliés aussi par 2). L'ensemble des molécules étudiées permet de construire un tableau de correspondance k(i,j) = nombre de fois qu'un pentapeptide de forme spatiale j présente le caractère chimique i . En adjoignant à ce tableau des colonnes supplémentaires représentant chacune un pentapeptide p (avec k(i , p) = 1 si p présente le caractère chimique i ; et 0 sinon), on a une représentation spatiale du nuage de quelque 3000 pentapeptides considérés ici, bien adaptée à l'étude des propriétés chimiques (séquentielles, de la structure primaire) les plus liées à la forme. Pour prédire la structure d'un pentapeptide nouveau ps , on voudrait trier ses plus proches voisins parmi les 3000 pentapeptides de forme connue, et lui attribuer pour angles αi (cf § 2.2) les valeurs moyennes des angles de ses voisins (en tenant compte ensuite, cf *infra*, de l'ensemble de la molécule et non plus seulement d'un pentapeptide unique). Mais en l'état actuel des programmes de recherche de voisins (susceptibles toutefois d'être grandement accélérés dans un proche avenir, grâce à l'algorithme d'agrégation en boules de rayon fixe et centres optimisés ; cf *Cahiers* Vol IV n° 3 pp 357 sqq, 1979 et l'algorithme d'agrégation autour de centres variables en boules de rayons bornés cf *Cahiers* Vol IV n° 3 pp 365 sqq) cette méthode aurait demandé

un temps de calcul excessif. On s'est borné à chercher les plus proches voisins de p parmi l'ensemble J des 16 structures spatiales-types : le premier voisin fournissant la structure jugée la plus probable ; et les voisins suivants des hypothèses à examiner après la première. Dès lors la prédiction de la structure d'une protéine est faite en suivant la chaîne des pentapeptides,et en attribuant à chacun d'eux p, une structure j proche de p (dans l'espace issu de l'analyse factorielle) et de plus compatible avec des structures j' , j" proches des points figurant les pentapeptides qui encadrent p : en effet si e.g. on attribue au pentapeptide 8-9-10-11-12 la structure AABB, on ne pourra attribuer à 7-8-9-10-11 que l'une des deux structures BAAB ou AAAB ; (et de même pour 9-10-11-12-13, soit ABBB soit ABBA). Finalement, la structure de la molécule est prédite comme une suite de A et de B, qu'on peut comparer à la structure réelle. Le hasard pur donnerait un taux de concordance de 50% : notre progamme a donné des taux oscillant de 65% à 80% (la valeur typique étant 70%) . Il est intéressant de noter que la plupart des erreurs se groupent en séquences : comme si une erreur sur un pentapeptide, entraînait une fausse orientation dans l'interprétation de tout un segment. Par exemple dans le cas de la carboxypeptidase, qui compte 300 liaisons peptidiques, il y a 195 prédictions exactes (i.e. angles de classe A $\approx 50°$, ou B $\approx 200°$ correctement prédits) soit 65% du tout ; mais les séquences exactes de longueur supérieure ou égale à 4, ont une longueur totale de 150 (soit la moitié du tout), et il y a une séquence exacte de longueur 31 ; le nombre des pentapeptides dont la structure spatiale est correctement prédite est de 100(soit un tiers) ; ce qui est une performance intéressante puisqu'il y a 16 structures possibles (encore qu'inégalement probables : cf § 2.3 ; tableau 2).

Nous conclurons donc que les données analysées ici suffisent à montrer des relations nettes entre structure primaire et structure secondaire des protéines ; mais que la prédiction de celle-ci à partir de celle-là n'est pas atteinte : l'analyse de données plus vastes permettra sans doute d'en approcher ; même si le caractère global du problème de la structure permet d'affirmer qu'une prédiction tout à fait sûre requiert que les suggestions issues de l'analyse statistique soient confrontées à un modèle chimique (calculs d'énergie etc.).