ANALYSE STATISTIQUE DE L'ADN MODELISATION PROBABILISTE PAR LES CHAINES DE MARKOV PUIS SIMULATION ET DETECTION DE BIAIS

Une nouvelle utilisation du logiciel ANAGENE (INRP)

Guy RUMELHARD

Résumé :
L'analyse des séquences d'ADN comprenant plusieurs milliers ou plusieurs centaines de millions de nucléotides, librement accessibles dans les banques de données, ne relève pas, comme on a pu le croire, de la simple « lecture ». Les mathématiques jouent un rôle fondamental pour repérer les « plages » homogènes ou les « mots » ayant un rôle que le biologiste doit ensuite confirmer au laboratoire. Nous proposons ici une approche de la modélisation à l'aide des chaînes de Markov accessible aux élèves de lycées.

L'analyse fonctionnelle indirecte de l'ADN a longtemps commencé par la recherche des protéines synthétisées (enzymes, hormones, constituants des membranes, hémoglobines, globulines,…), et des ARN. On détermine la ou les cellules qui les fabriquent, et on recherche alors l'ARN messager présent dans la cellule au moment de la synthèse. A partir de celui-ci il est possible de « remonter » à l'ADN du chromosome par l'intermédiaire d'un ADNc copié, mais chez les eucaryotes une grande partie de l'ADN n'est pas transcrit ou transcrit mais pas traduit.

L'analyse structurale directe de l'ADN dont le rôle est inconnu est devenue indispensable depuis que l'analyse complète des séquences de l'ADN d'un organisme est possible. La suite des opérations s'est **nécessairement renversée**, sans rendre pour autant l'analyse fonctionnelle obsolète, mais en la complétant et en guidant la recherche des séquences fonctionnelles. En effet le chercheur se trouve devant plusieurs milliers (pour le moindre virus), plusieurs millions (pour les bactéries) et plusieurs centaines de millions de nucléotides (pour les eucaryotes) dont la signification lui échappe. De plus on commence à savoir que si chez les bactéries 95% de cet ADN correspond à des protéines, chez l'homme 95% de cet ADN n'a pas de signification actuellement connue. Par ailleurs les gènes sont constitués d'introns qui ne sont pas transcrits en m ARN, certains m ARN ne sont pas traduits en protéines, et une partie de l'ADN correspond à des virus intégrés. L'espoir de pouvoir « lire » directement l'ADN et de comprendre son rôle s'est évanoui. Cet espoir naïf correspondait à l'attitude naturaliste pour laquelle « comprendre c'est voir ». Pour aborder ces questions les chercheurs font appel, depuis 15 ans environ, à des modèles mathématiques mis au point récemment et qui donnent lieu actuellement à des thèses en mathématiques. Ils ont été mis en œuvre grâce à des ordinateurs puissants qui demandent parfois plusieurs heures de calcul. Cette analyse *in silico* n'est pas un pis aller temporaire. Les mathématiques jouent désormais un *rôle créateur*[1] incontournable dans la recherche, dans l'ADN de « plages homogènes » qui correspondent à des « séquences fonctionnelles », et des « petits mots » de l'ADN ayant une fonction connue.

On peut se demander si ce travail est définitivement inaccessible aux lycéens. Nous allons précisément tenter de décrire une approche de ces méthodes qui soit accessible à des élèves

[1] Rumelhard Guy (2001) Le rôle créateur des mathématiques en sciences de la vie *Biologie Géologie* 4 p. 715-729

de lycée. Il n'existe actuellement qu'un seul livre[2] réellement opératoire mais d'un niveau élevé en mathématiques et quelques articles en français[3] de difficulté variable. Une bibliographie considérable existe en anglais.

Le cheminement du raisonnement est le suivant :

- L'observation et l'analyse statistique de séquences d'ADN permet de déterminer quelques paramètres, telle la fréquence des lettres prises isolément, des successions de deux lettres, ou de trois lettres et plus,
- On propose alors plusieurs modélisations probabilistes de cette séquence,
- Puis on réalise une simulation à partir des divers modèles,
- Enfin la confrontation à la séquence permet de déterminer d'éventuels biais, c'est-à-dire des écarts entre la simulation et la réalité

On peut commencer par l'analyse mathématique de séquences d'ADN assez courtes sur lesquelles ont peut travailler « à la main » sur papier et dont la fonction est connue pour rechercher et comprendre l'intérêt des *modélisations probabilistes* (Chaînes de Markov, loi de Gauss ou de Poisson). Ce travail est accessible en classe au lycée.

On comprendra ensuite comment analyser mathématiquement des séquences inconnues, en recherchant par exemple les chaînes de Markov cachées à l'aide de fenêtres glissantes et d'un algorithme EM, mais on ne peut pas le réaliser en classe sauf à utiliser un programme dont on ne comprend pas l'organisation. On peut par contre comprendre que ce travail mathématique prépare le travail du biologiste qui doit rechercher au laboratoire la fonction des séquences ou des mots distingués et repérés.
Ces modèles mathématiques sont principalement au nombre de trois :

- Chaînes de Markov, (CM)
- Modèles de Markov cachés (HMM, hidden Markov models),
- Algorithme EM (estimation, maximisation), qui se déplace par fenêtres glissantes,

Recherche d'un modèle

Il est devenu habituel de dire que l'ADN se présente comme un texte composé à l'aide de quatre lettres A,C,G,T, qui s'enchaînent *sans interruption* et qui est *orienté* avec un début et une fin. Prenons cette comparaison au sérieux et non pas comme une vague image. Cette analogie avec un texte en langue française nous servira de modèle pour comprendre l'organisation de l'ADN. Notons donc quelques observations et quelques procédés d'analyse

[2] Robin Stéphane, Rodolphe François, Schbath Sophie (2003) *ADN, mots et modèles*. Paris : Belin collection Echelles.
On trouvera des informations plus générales dans : Dardel F., Képès F. (2002) *Bio informatique. Génomique et post-génomique*. Paris : Les éditions de l'école polytechnique

[3] Prum Bernard (2004) Mathématiques et biologie *APMEP* n° 440 p. 337-348 ;
Prum Bernard (2001) La recherche automatique des gènes. *La Recherche* n°346 p. 84-87 ;
Prum Bernard (2000) Une approche statistique de l'analyse des génomes. *La Revue du Palais de la Découverte* 276, 56-65 ;
Prum Bernard (2002) Trouver un gène responsable du cancer. *L'explosion des mathématiques* p. 28-31 ;
Prum Bernard (2000) Les chaînes de Markov dans l'analyse des génomes, *Matapli* 62, 24 ;
Prum Bernard, Muri-Majoube Florence (2001) Une approche statistique de l'analyse des génomes. Gazette n°89
Morange Michel (2005) Avant propos. Dossier N°46. *Pour la Science*. Janvier-Mars ;
Schbath S. (2003) *A la recherche de mots de fréquence exceptionnelle dans les génomes* Images des mathématiques CNRS vol 3

des lettres et des mots d'un texte écrit avec un alphabet comprenant voyelles et consonnes. Il faut noter que certaines langues ne comportent pas de voyelles à l'écrit du moins.

1. ANALYSE D'UN TEXTE COMPOSÉ DE LETTRES (VOYELLES ET CONSONNES)

1.1. Un obstacle à l'analyse d'un texte

La description statistique et la modélisation probabiliste d'un texte écrit dans une langue peut sembler incongrue comme toute mathématisation des œuvres culturelles et des comportements humains, puisque l'essentiel réside dans *la signification* des mots et des phrases ou dans le *plaisir esthétique* que le texte procure. Dans un registre voisin, celui de la musique, et bien qu'il existe des musiques aléatoires au premier rang desquels se présente la campanologie, c'est à dire certaines façons traditionnelles de faire sonner des carillons constitués de nombreuses cloches, il semble totalement incongru de compter les types de notes, leur fréquence, leur enchaînement dans une sonate, une symphonie ou tout autre morceau d'un compositeur célèbre. Sur des textes écrits ce type d'analyse a cependant été réalisé de manière fructueuse, particulièrement depuis l'apparition des théories de la communication et de l'information[4] au milieu du XX ème siècle et initialement dès 1902 par le mathématicien russe Andreï Markov.

1.2. Limites de l'analogie entre texte et ADN

Soit un texte français (ou anglais, allemand), écrit avec les 26 lettres de l'alphabet. Mais en fait les mots sont séparés, et il existe divers signes de ponctuation, ainsi que des lettres accentuées (aigu, grave, circonflexe, tréma, cédilles, apostrophe). Il y a donc plus de 40 « lettres ». L'ADN n'est constitué que de quatre lettres sans séparations ni accentuation. De plus l'ADN est « lu » par groupes de trois lettres avec trois cadres différentes ayant chacun éventuellement une signification.

1.3. Premier modèle pour rendre compte du texte : la fréquence de chaque lettre est constante, mais varie selon les langues

L'occurrence des lettres d'un texte peut sembler *a priori* quelconque[5] et n'obéir à aucune régularité, en particulier pas d'un texte à l'autre, ni même d'un chapitre à l'autre dans un même texte. Il n'en est rien. On peut vérifier empiriquement que la fréquence de chaque lettre tend vers une valeur constante dans un texte suffisamment long et homogène (au moins 1 000 à 10 000 lettres). Cette valeur présente une certaine variance selon la taille du texte. Chaque texte devient un échantillon d'un texte idéal, sauf cas particuliers de textes volontairement écrit en ne faisant pas appel à toutes les lettres. On peut vérifier empiriquement que cette fréquence est différente selon les différentes langues. Les documents de cryptographie fournissent tous les résultats souhaités. On trouve par exemple pour les voyelles en français et en anglais les proportions relatives suivantes (pour 1) :

---

[4] Shannon Claude, Weaver Warren (1949) *The Mathematical Theory of Communication*. Urbana Ill. The University of Illinois Press. ; Mandelbrot Benoït (1954) Structure formelle des textes et communication. Deux études. *Word 10 n°1, 1-27*

[5] Robert-Schwartz Claudine (2003) *L'empereur et la girafe* Diderot, réédité sous le titre *Contes et décomptes de la statistique : Une initiation par l'exemple*. Vuibert p.15-26 textes de Baudelaire, Poincaré, Shakespeare, Neyman, Georges Perec. Etc. p. 168-175.

En français	A = 0,15	E = 0,42	I = 0,16	O = 0,12	U = 0,15	Y = 0,006 total = 1,00
En anglais	A = 0,21	E = 0,31	I = 0,21	O = 0,16	U = 0,06	Y = 0,04 total = 1,00

On peut tenter de modéliser un texte en faisant seulement appel à ces fréquences, mais il faut simplifier pour pouvoir le faire « à la main ». Dans ce premier modèle, en ne tenant compte que des voyelles et des consonnes, il faudrait supposer qu'un "mot" construit uniquement avec des voyelles (v) et des consonnes (c) est obtenu par un *tirage aléatoire avec remise*, de v ou de c dans une urne dite de Bernoulli dont la composition correspond aux fréquences des v et des c et ne change pas au fur et à mesure des tirages. On dit encore que ce processus est "sans mémoire" car le tirage d'un v ou d'un c ne dépend pas du tirage précédent. Autrement dit la suite des v et des c n'obéit à aucune nécessité de succession autre que le hasard. En fait, en tirant "au hasard" dans une urne les v et les c pour simuler un texte artificiel en fonction de ce premier modèle très simple, il y a "peu de chances" que l'on forme des mots correspondant à ceux qui existent. Ici une estimation numérique de cette affirmation la confirmerait. Ce modèle qui suppose que les lettres sont tirées au hasard dans une urne avec remise, est donc très insuffisant.

Il suffit cependant pour décrypter un texte rendu artificiellement illisible par le procédé de cryptage ci-dessous, et pour séparer éventuellement deux types de langues (français et anglais mélangés) et reconnaître des mots.
Pour se placer dans la même situation inconnue que lorsqu' on "lit" une séquence d'ADN imaginons que chacune des 26 lettres non accentuée, a été remplacée par une autre selon une règle simple homogène pour tout le texte. Par exemple on décale l'alphabet d'une lettre et A est remplacé par un B, B par C et ainsi de suite. De plus toutes les coupures entre les mots, tous les signes de ponctuation et tous les accents ont été supprimées. Ceci a pour effet de rendre le texte incompréhensible.

Ce codage par décalage arbitraire détruit le sens initial et ne produit en général pas de mots nouveaux inattendus. Par contre dans l'ADN qui est lu par groupes de trois lettres le décalage du "cadre de lecture" d'un seul nucléotide suffit pour produire parfois une autre protéine dont la composition est totalement différente dans la mesure précisément où tous les triplets de nucléotides ont une signification. Plusieurs virus, dont le VIH, utilisent partiellement ce procédé consistant à lire la même séquence de nucléotides deux fois de manière décalée pour produire deux protéines différentes. Cette remarque vise à rappeler que l'analogie entre un texte en langue française et des séquences de nucléotides, présente des limites. Une analogie ne contient pas en elle même les limites de son emploi comme modèle. On peut bien évidemment penser *a priori* qu'un texte littéraire ou scientifique est beaucoup plus complexe dans son organisation qu'une séquence de nucléotides. Par ailleurs cette remarque vise à lutter contre une habitude journalistique consistant à dire que le texte de l'ADN est « crypté » c'est à dire volontairement caché (par qui, et dans quel but ?!) et qu'il faudrait le "décrypter". Il n'en est rien !

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z	alphabet
B C D E F G H I J K L M N O P Q R S T U V W X Y Z A	code
LESCOUPURESENTRELESMOTSSONTSUPPRIMEES	phrase sans coupure ni accents
MFTDPVQVSFTFVUSFMFTNPUTTPOUTVQQSJNFFT	phrase codée

Pour décoder un texte codé comme la phrase précédente on observe la fréquence des différentes lettres du texte illisible et on remplace les lettres par celles des fréquences standard de la langue choisie. Si le texte est suffisamment long on aboutit à un résultat qui laisse cependant de nombreuses incertitudes pour les lettres de faible fréquence. Mais si on connaît la langue on reconnaît un mot même sans avoir identifié toutes les lettres. Par ailleurs la fréquence des « w » ou des « u » (entre autre) n'est pas la même en français et en anglais et constitue un indice du mélange des deux langues si le texte est suffisamment long. Évidemment ce travail peut sembler inutile pour notre propos, il prend cependant de l'intérêt si le texte comprend des mots ou des phrases de différentes langues mélangés, car on comprend alors que le but sera de commencer à les séparer.

1.4. Deuxième modèle : la succession (= transition = enchaînement) des voyelles et des consonnes rendent mieux compte du texte

Il semble cependant évident que les successions des voyelles et des consonnes ne sont pas aléatoires. Les doubles consonnes sont nombreuses, les doubles voyelles également. La succession de trois voyelles ou de trois consonnes existe également, mais, en français les successions de quatre ou plus sont rares et correspondent souvent à des mots étrangers. Supposons que dans un texte d'environ 2000 lettres la lettre « e » apparaisse dans 12% des cas du total des lettres (voyelles et consonnes). Si les lettres successives étaient indépendantes les unes des autres, la succession de deux lettres « ee » sans accent devrait apparaître avec une probabilité de p = 0 ,12*0,12 = 0,014 soit environ 28 fois. Il n'en est rien. En fait « e » est souvent suivi d'une consonne, plus spécialement d'un l, d'un n, d'un m, d'un s, d'un t,… ou d'un espace (fin de mot).
Etudier de manière systématique cette deuxième approche, c'est-à-dire toutes les successions de deux, puis trois, puis quatre lettres est fastidieux. On peut le réaliser « à la main » dans un premier temps, à condition de simplifier le travail. On remplace les lettres par la simple désignation de voyelle (v) ou de consonne (c).

- La *fréquence des c et des v* est différente et connue empiriquement. Elle est par exemple :

45,1 % de voyelles et 54,9 % de consonnes en français.

- On pose donc que l'apparition d'un type de lettre (voyelle ou consonne) dans un mot d'une langue donné n'est pas indépendante de la lettre précédente ou des deux ou trois lettres qui l'ont précédé. On propose donc un autre modèle qui améliore la description du texte en prenant en compte des *successions de lettres*. On commencera par deux lettres et on calculera la fréquence de chaque succession : cc, cv, vc, vv. On peut alors se poser la question de la façon suivante. Quelle est la probabilité d'avoir une consonne après une voyelle, autrement dit la succession cv ? Quelle est la probabilité d'avoir une voyelle après une consonne, autrement dit la succession vc. Pour le dire autrement, est-ce que la présence d'une voyelle augmente ou diminue la probabilité de présence d'une voyelle à la lettre suivante. Par exemple en français la lettre q est "souvent" suivie d'un "u", mais les doubles voyelles et les doubles consonnes existent aussi. On peut compliquer la question en se demandant quelle est la probabilité d'avoir une voyelle après deux consonnes, d'avoir une consonne après deux voyelles, etc… Un retour sur la langue française montre aisément que l'alternance régulière de voyelles et de consonnes (comme dans le mot : totalité) est peu fréquente. Par contre la succession de deux voyelles est "souvent" suivie d'une consonne et la succession de deux consonnes est "souvent" suivie d'une voyelle.

- Si on prend en compte des *successions de trois v ou de trois c*, on rendra mieux compte de la langue analysée. Le retour sur la langue française montre que cette fois les successions de trois voyelles sont plus rares, et presque toujours suivies d'une consonne (sauf en fin de mot ou cela dépend du mot suivant), les successions de trois consonnes ne sont pas très fréquentes. Les successions de quatre, ou de cinq c le sont encore moins en français du moins. Autrement dit la nouvelle lettre à écrire dépend de la lettre qui la précède immédiatement, et plus sûrement des deux, trois, quatre lettres précédentes. *Ce processus a une mémoire*. Cette mémoire est cependant limitée. Il est vraisemblablement inutile de remonter très loin. L'apparition d'une lettre est certainement indépendante de la présence d'une voyelle ou d'une consonne 10 ou 20 lettres avant (sauf cas de poème avec des rimes), mais il faudrait tester ces modèles successifs prenant en compte des successions de deux, trois, n lettres.

Bien évidemment ces successions varient d'une langue à l'autre. Il existe une lettre de l'alphabet cyrillique щ que l'on écrit en français avec 5 consonnes : chtch. Les groupes schl, schm sont absents en français sauf pour des mots allemands importés. Etc. On pourra donc distinguer les différentes langues d'après la fréquence de ces successions de deux, trois, quatre lettres, et reconnaître en quelle langue est écrit un texte inconnu.

Ce procédé d'analyse a été théorisé par le mathématicien russe Andreï Markov (1902). Il a appliqué sa théorie aux 20 000 premières lettres du roman de Pouchkine *Eugène Onéguine*. On parle de *chaînes de Markov* ou de *processus markovien*. Comme bien souvent en mathématiques ce travail est très polyvalent et s'applique dans de nombreux domaines : économie, écologie, éthologie, musique, génétique, dynamique des populations, épidémiologie, moteurs de recherche sur Internet, etc.

1.5. Importance de l'écart entre le modèle et l'observation

Selon l'un des sens du mot modèle on dira que l'on cherche ici à *représenter* le mieux possible une langue grâce à un modèle. Mais on peut aussi et même surtout s'intéresser à *l'écart observé entre le modèle ou plutôt ses simulations et le résultat empirique*. Le modèle tente de représenter le résultat *attendu* par exemple si l'on travaille sur un texte crypté, dont on connaît la langue ou le mélange de deux langues. Les séquences qui retiennent l'attention sont surtout *celles qui s'en écartent*. Cet écart est source de questions nouvelles et de recherche. S'agit-il de fragments intercalés d'une autre langue ? Il relance le travail. Voilà le vrai sens de la modélisation dans ce cas.

1.6. Applications à un texte français de six phrases[6]

Analysons un texte constitué par les dix premières phrases extraites du texte de Michael Keane[7]. Nous ne reproduisons pas ici le texte lui-même, mais seulement le résultat des dénombrements.

phrase	c	v	cc	cv	vc	vv
1	126	107	45	81	80	25
2	142	123	49	93	92	30

[6] Ce travail a été réalisé par Michèle ARTIGUE lors des journées de formation interdisciplinaires (Mathématiques, Physique, SVT) à l'IREM de Paris

[7] Keane Michael (2003) *Marches aléatoires renforcées* in *Leçons de mathématiques d'aujourd'hui* (tome 2) Éditions Cassini

3	74	82	21	52	53	26
4	89	79	23	64	65	14
5	120	96	39	80	80	13
6	113	82	44	67	67	13
total	664	569	221	437	437	121

Total c + v = 1233
Total cc + cv + vc + vv = 1216
Fréquence observée de c = 664 / 1233 = 0,54 donc la fréquence de v = 0,46
Dans la langue française (fréquence théorique) c = 0,549 et v = 0,451 soit un écart de 0,9 % car la longueur de ce texte est insuffisante.

Fréquence théorique calculée dans le cas du modèle de l'urne de Bernoulli dit M0, appliqué au cas observé pour des transitions lettre à lettre.

cc = 0,54 * 0,54 = 0,29 cv = vc = 0,54 * 0,46 = 0,25 vv = 0,46 * 0,46 = 0,21

216 * 0,25 = 304 ; 1216 * 0,29 = 353 ; 1216 * 0,21 = 255

	cc	cv	vv
théorique	353	304	255
observé	221	437	121

Il y a donc un "gros" écart, ce que peut confirmer un calcul du Chi carré avec deux degrés de liberté. Donc l'urne de Bernoulli n'est pas un bon modèle de la situation.
On peut alors refaire le même travail sur un texte en langue anglaise pour rechercher s'il y a des différences et si elles sont significatives par rapport à un texte français. Pour l'ADN il en sera de même dans la mesure où il est constitué d'un mélange de « plages homogènes », de longueur variée, et ayant des « styles » différents. On cherchera à distinguer ces plages.

2. PRESENTATION DES CHAINES DE MARKOV[8]

Ce calcul permet de rendre compte des enchaînements "verticaux" d'une génération à l'autre, ou "horizontaux" le long d'un texte ou d'une séquence d'ADN par exemple.

2.1. Processus généalogiques[9]

Analysons la *transmission "verticale"* des gènes c'est à dire d'une génération à l'autre entre parents n'ayant aucune relation de parenté entre eux et enfants. Il faut et il suffit de tenir compte de la génération qui précède soit (n - 1) pour prévoir les allèles transmis à la génération n. Les autres générations n'apportent aucune information supplémentaire. Si la *chaîne de Markov* ne tient compte que de la génération immédiatement précédente on la

[8] Rittaud Benoît (2004) Les chaînes de Markov, outil universel. *La Recherche* n°378 p. 80-81 ; on trouvera sur Internet de nombreux cours de mathématique sur les chaînes de Markov, les chaînes de Markov cachées et les algorithmes E M.
Girault Maurice (1964) *Première initiation aux processus de Markov*. Revue de Statistique Appliquée. Vol XII n°3 p. 5-14
[9] Engel Arthur (1970) *L'enseignement des probabilités et des statistiques* Ed Cedic 2 tomes. § 6 et 7

nomme : M1. Par contre si la transmission se fait par accouplement entre frères et sœurs ou entre parents et enfants (ce qui est courant chez les animaux) le modèle se complique.

2.2. Succession horizontale des lettres d'un texte

Le jeu de pile ou face rend bien compte d'une succession d'événements. C'est un processus idéalement sans mémoire. Chaque nouveau lancé est *totalement indépendant* des précédents. Par généralisation on peut le considérer comme un cas particulier de chaîne de Markov et l'écrire M0. Ce *concept d'indépendance* est au cœur des modélisations mendéliennes. Dans les jeux de hasard on espère toujours que les chances ne sont pas tout à fait égales, et qu'il existe *une certaine mémoire*, autrement dit qu'un essai nouveau dépend des essais précédents, d'où la recherche d'hypothétiques "martingales". Il faudrait tenir compte du coup précédent, mais aussi des deux, trois, n coups précédents. A la limite le jeu a une mémoire sinon infinie, du moins *périodique*. Le principe des chaînes de Markov consiste à modéliser cette mémoire, c'est-à-dire la dépendance d'un événement par rapport aux événements précédents, autrement dit des successions.
Les générateurs supposés aléatoires des calculatrices scolaires, s'appuient sur un processus périodique mais qui se répète avec un délai suffisamment grand pour que les calculs ordinaires n'en tiennent pas compte. Ils simulent le "hasard". On parle de générateurs pseudo - aléatoire équidistribués. Le modèle des tirages dans une urne dite de Bernoulli avec remise immédiate, est dit sans mémoire. Nous évoquerons ici les tirages sans remise ou avec remise différée d'un ou plusieurs coups qui tentent de simuler diverses formes de dépendance, sans les analyser.

3. ANALYSE DE L' ADN

Pour comprendre le travail des chercheurs sur les séquences d'ADN inconnues on peut commencer par effectuer *deux types de recherche* sur des séquences connues :

- l'utilisation des modèles de Markov permet, par l'analyse des fréquences de lettres prises isolément ou la fréquence de certains *enchaînements* de deux, trois, n lettres, de déceler dans la séquence totale différentes *plages homogènes* ayant des « styles » différents, exons et introns par exemple, c'est-à-dire des proportions des différentes lettres A, C, G,T spécifiques et des probabilités de transition de deux lettres également spécifiques, ce qui doit permettre de les distinguer dans une séquence complète.

- la recherche de la *fréquence et de la répartition de petits mots* de deux, trois, … huit, dix, douze lettres ou plus, dont la fonction est connue ce qui doit aussi permettre d'aider à distinguer des séquences ayant des fonctions différentes selon que les mots sont trop fréquents ou trop rares ou même absents, trop espacés ou trop groupés non pas dans l'absolu, mais en fonction du modèle de Markov retenu du type M0, M1, M2, etc.

Dans les banques d'ADN seule la séquence 5' 3' c'est-à-dire le brin dit « non codant » est écrit. Il est aisé de retrouver la séquence complémentaire grâce à la correspondance A/T et C/G. Cette séquence est nommée « brin codant ». Ce brin est transcrit au moins partiellement

en ARN messager. Dans certains cas cette séquence dite « non codante » peut être cependant partiellement lue et donner naissance à d'autres protéines car elle n'est pas le palindrome de la précédente, sauf pour certains groupes de 4 à 13 lettres qui sont souvent des sites de fixation des enzymes de restriction.

Le problème semble plus compliqué que la succession des v et c puisque nous devons prendre en compte quatre lettres et non pas deux. Par ailleurs il peut sembler incongru de décrire à l'aide de statistiques et modéliser à l'aide de modèles probabilistes la succession des quatre lettres si l'on pense que la question principale est celle de leur *signification fonctionnelle* et non pas de leur *fréquence statistique*, ou de leur *répartition selon un modèle aléatoire* ou non. On espère que la réponse à la seconde question permettra de guider la recherche de la signification fonctionnelle.

3.1. ADN, processus aléatoires et pression de sélection

On suppose que la succession des quatre lettres de l'ADN résulte initialement ou lors des mutations plus récentes, de *processus aléatoires. L'alea est au cœur* des modèles que l'on va proposer. Cette hypothèse est raisonnable car d'une part plusieurs phénomènes se présentent comme une simple combinatoire des quatre lettres, et d'autre part lors de la transmission "verticale" (d'une génération à l'autre par méiose et fécondation) ou "horizontale" (d'une bactérie à l'autre par exemple par échange de plasmides, de chromosome lors de conjugaisons, ou par l'intermédiaire d'un bactériophage) plusieurs mécanismes génétiques actuels ou historiques, sont *modélisés par un processus aléatoires* : séparation des paires de chromosomes au cours de la méiose, mutations, crossing over et cassures diverses, etc. La *pression de sélection* a dû modifier ce caractère aléatoire, mais il doit en rester *des traces* dans les parties du génome non soumises à cette pression de sélection telles que les introns ou les séquences intergéniques.

Précisons en quoi *l'aléa* est *au cœur* de certaines modélisations de mécanismes génétiques. Parmi les *mutations* de l'hémoglobine[10] on peut noter par exemple que l'acide glutamique (Glu) qui est codé par les deux codons GAA et GAG est remplacé, selon les mutations, par la valine (Val), ou la lysine (Lys), ou la glutamine (Gln) ou la glycine (Gly). Ces mutations créent des anémies. Ces mutations résultent de la modification d'un seul nucléotide à la fois, en première, seconde ou troisième position du triplet. Pour savoir si tous les cas possibles de mutation se sont réalisés on notera que TAA et TAG qui peuvent résulter d'une mutation de g en t en première position correspondent à des codons stop et conduisent à une protéine non fonctionnelle donc à un individu qui n'existe pas car il ne peut pas vivre, tandis que les deux mutations restantes n'ont simplement pas encore été trouvées ou bien correspondent à des situations également non fonctionnelles, c'est-à-dire létales. Tous les cas possibles (sauf deux) existent donc réellement ou potentiellement, même s'ils ne sont pas équiprobables. Le degré de viabilité induit les fréquences observées.
Cet effet de sélection va introduire *des biais, ou des écarts par rapport au modèle* de Markov proposé qui ne rend compte que des séquences « banales », c'est-à-dire précisément celles qui ne nous intéressent pas.

Si les mutations qui se traduisent phénotypiquement par des maladies, ou par un caractère létal sont peu nombreuses pour une protéine donnée, le polymorphisme dû à des mutations "neutres" est parfois considérable. Dans le cas de la mucoviscidose, on connaît actuellement plus de huit cents mutations. Une seule mutation en position 508 rend malade.

[10] Rumelhard Guy (1998) Le jeu des possibles et la réfutation. *Biologie Géologie* (APBG) n°1 p. 113-119

En présentant le tableau du code génétique de manière différente de la façon habituelle, on peut faire apparaître une propriété des triplets de bases qui codent les acides aminés.

Acides aminés				**Codons**
Alanine	Ala	A		GC**A** GC**C** GC**G** GC**T**
Glycine	Gly	G		GG**A** GG**C** GG**G** GG**T**
Proline	Pro	P		CC**A** CC**C** CC**G** CC**T**
Thréonine	Thr	T		AC**A** AC**C** AC**G** AC**T**
Valine	Val	V		GT**A** GT**C** GT**G** GT**T**
Isoleucine	Iso	I		ATA ATC - ATT
Methionine	Met	M	START	ATG
Arginine	Arg	A		CG**A** CG**C** CG**G** CG**T** AGA AGG
Leucine	Leu	L		CT**A** CT**C** CT**G** CT**T** TTA TTG
Sérine	Ser	S		TC**A** TC**C** TC**G** TC**T** AGT AGC
Asparagine	Asn	N		AAC AAT
Ac. Aspartique	Asp	D		GAC GAT
Cystéine	Cys	C		TGC TGT
Glutamine	Gln	Q		CAA CAG
Ac. Glutamique	Glu	E		GAA GAG
Histidine	His	H		CAC CAT
Lysine	Lys	K		AAA AAG
Phénylalanine	Phe	F		TTC TTT
Tyrosine	Tyr	Y		TAC TAT
Tryptophane	Trp	W		TGG
STOP				TAA TAG TGA

Tableau du code génétique.

Le code génétique est dit *dégénéré* car il suffit de 20 codons pour 20 acides aminés auxquels il faut ajouter les codons STOP et START. Mais cette redondance n'est pas quelconque. Huit triplets présentant au moins quatre possibilités correspondent aux quatre cas possibles de substitution en troisième position. Val = GTT, GTC, GTA, GTG, Gly = GGT, GGC, GGA, GGG, Arg = CGT, CGC, CGA, CGG, Ala = GCT, GCC, GCA, GCG, Thr = ACT, ACC, ACA, ACG, Pro = CCT, CCC, CCA, CCG, Ser = TCT, TCC,TCA, TCG, Leu = TTT, TTC, TTA, TTG. Ici encore cette combinatoire de cas possibles ne signifie pas qu'ils sont équiprobables dans les séquences observées. La sélection peut être intervenue, créant ainsi *des biais par rapport au modèle* de Markov proposé.

Bien évidemment même si au départ, lors de la formation des séquences, ou au moment des mutations, le *processus est aléatoire*, il y a ensuite une *pression de sélection* qui élimine ou privilégie certaines séquences. Donc le modèle aléatoire utilisé *ne cherche pas à représenter toutes les séquences observées* mais seulement celles qui sont "banales" et ne présentent pas d'intérêt en l'état actuel des connaissances. *Détecter des écarts* par rapport à cette constitution théorique aléatoire supposée devra être interprété en termes biologiques c'est à dire en termes de fonction physiologique et d'avantage sélectif éventuel.
Ce n'est pas le seul cas où l'on s'intéresse à *l'écart par rapport au modèle*. Par exemple les chromosomes sexuels X et Y sont équidistribués dans les deux types de spermatozoïdes, et au moment de la fécondation les rencontres sont supposées aléatoires. Si un écart significatif se présente (plus de garçons ou de mâles que de filles ou de femelles ou l'inverse), il invite à chercher la signification physiologique (action de certaines hormones, réactions immunitaires, mutations,…) ou écologique (action de la température) de cet écart. Mais on ne cherche pas à rendre compte mathématiquement de cet écart, c'est-à-dire à l'inclure dans les paramètres définissant le modèle.

3.2. Premier modèle : analyse des fréquences de A, C, G, T

Comme pour le texte français examinons si on peut observer des distinctions entre les différentes séquences de nucléotides en s'appuyant uniquement sur les fréquences des quatre lettres. On peut calculer la fréquence de chacune des quatre lettres, mais il n'y a pas cette fois de fréquence idéale à laquelle se référer puisqu'une séquence d'ADN donnée n'est pas un échantillon d'un texte idéal. Chaque séquence d'ADN d'un organisme donné analysée en totalité risque d'être unique. Les fréquences des quatre lettres risquent d'être spécifiques à chaque séquence. On espère même, par hypothèse, que séquences inter géniques, exons, introns, et ADN viral intégré sont écrites dans des "styles de langues" différentes ce qui devrait permettre de pouvoir les identifier et les séparer le long d'une séquence inconnue. On admet par contre que la séquence d'un exon, ou d'un intron est homogène, c'est à dire que, si l'on admet un modèle probabiliste, les fréquences ne varient que de manière statistique le long de la séquence. Par exemple pour la bactérie *Haemophilus influenzae* dont le chromosome comprend 1.830.140 nucléotides on peut calculer la *fréquence globale* de chaque lettre en supposant que ce chromosome est totalement *homogène,* ce qui n'est certainement pas le cas. Il doit exister plusieurs plages différentes correspondant à plusieurs gènes et qui ont des probabilités de transition différentes. Homogène signifie que la probabilité de transition d'un doublet CG par exemple ne dépend pas de sa position le long de la séquence.
Les banques de données proposent pour *Haemophilus influenzae* :

A = 0,31 C = 0,18 G = 0,20 T = 0,31 total = A + G + C +T = 1

On a vu précédemment que pour un texte de la langue française, l'utilisation directe des fréquences conduit à un modèle ayant très peu d'intérêt pour distinguer des séquences (= des plages) différentes. Nous ferons le calcul ci-dessous pour le virus du SIDA pour le comparer au deuxième modèle.

3.3. Deuxième modèle : analyse des successions (transition, enchaînement) de deux nucléotides

Comme pour le texte en langue française pour lequel la prise en compte des successions de deux lettres rend mieux compte de l'organisation des mots, on peut, dans un premier temps,

rechercher des successions de deux lettres. On se demande quelle est la probabilité que A, C, G, ou T succède à un "A", puis que A, C, G, ou T, succède à un "C", et ainsi de suite. Il y a 16 combinaisons de quatre lettres prises deux à deux. On définit *la matrice de transition* comme la matrice contenant toutes les probabilités de transition (de succession). On forme une matrice carrée : quatre, quatre. On comprend immédiatement que la prise en compte de successions de trois, quatre, cinq lettres va conduire à construire des matrices de 64, 256, 1024 éléments :

$$\Pi = \begin{pmatrix} \mu(a,a) & \mu(a,c) & \mu(a,g) & \mu(a,t) \\ \mu(c,a) & \mu(c,c) & \mu(c,g) & \mu(c,t) \\ \mu(g,a) & \mu(g,c) & \mu(g,g) & \mu(g,t) \\ \mu(t,a) & \mu(t,c) & \mu(t,g) & \mu(t,t) \end{pmatrix}$$

Connaissant les fréquences de A, C, G, T, dans une séquence donnée, on peut calculer les fréquences des successions des 16 combinaisons de deux nucléotides selon le modèle M0 c'est-à-dire comme si les lettres étaient tirées successivement d'une urne de Bernoulli, et les comparer aux probabilités de transition des dinucléotides observées réellement sur cette même séquence.

On trouve dans les banques de données sur Internet le cas du virus du SIDA qui comprend 9718 lettres. La lettre A apparaît 3.410 fois, dont 1.112 fois suivie de A (33%), 561 fois suivie de C (16%), 1 024 fois suivie de G (30%) et 713 fois suivie de T (21%). Un décompte similaire pour les autres lettres C, G, T et leurs successions, conduit au tableau suivant qui donne le dénombrement des mots de deux lettres (la première lettre étant en tête de ligne, la seconde étant en tête de colonne, par exemple AC = 561, CA = 795, etc) dans le génome de HIV.

	A	C	G	T	Total par lettre	
A	1 112	561	1 024	713	3410	35,0%
C	795	413	95	470	1773	18,2%
G	820	457	661	432	2370	24,4%
T	684	342	590	548	2164	22,2%
					9718	100%

En remplaçant chaque nombre par le pourcentage correspondant, on obtient la *matrice de transition* de la chaîne de Markov M1 associée au virus du SIDA :

	A	C	G	T	
A	33%	16%	30%	21%	100%
C	45%	23%	5%	27%	100%
G	35%	19%	28%	18%	100%
T	32%	16%	27%	25%	100%

On remarquera que la matrice M1 n'est pas symétrique, c'est-à-dire que la probabilité que A soit suivi d'un C est de 16%, alors que la probabilité que C soit suivi d'un A est de 45%. Ce

calcul fait apparaître que le mot CG n'apparaît que 95 fois (soit 5%) tandis que GC apparaît 457 fois soit 19%. CG est donc très peu fréquent et ceci est mis en évidence par le modèle M1. Le biologiste propose une explication liée à la méthylation des cytosines que nous expliquerons plus loin.

La matrice ci-dessous calculée pour la chaîne de Markov M0 est symétrique et très différente. La succession A/C se calcule en multipliant le pourcentage de A par le pourcentage de C. Donc celle de C/A sera le même. CG et GC ont la même fréquence.

	A	C	G	T	
A	12%	6 %	9 %	8,5 %	
C	6 %	3 %	4 %	4 %	
G	9 %	4 %	6 %	5,4%	
T	8,5%	4 %	5,4%	5 %	
total	35,5%	17%	24,4%	22,9%	100%

Il faut le rappeler encore une fois, nous ne disons pas que « pour constituer une chaîne de nucléotides, la Nature tire la lettre suivante de l'ADN en fonction de celle qui précède immédiatement », mais nous recherchons dans quelle mesure le mot de deux lettres (qui peut être un sous mot d'un mot de trois lettres ou plus) *s'écarte* du modèle de Markov proposé. Par contre on peut simuler une séquence artificielle d'ADN qu'un biologiste admettra comme « vraisemblable » par ce type de tirage « au hasard ». De même on peut rendre compte des choix des consommateurs pour tel ou tel marque de produit alimentaire par un modèle probabiliste, ce qui ne signifie pas que, dans le magasin, le consommateur choisit « au hasard » en tirant par exemple à pile ou face, ou avec un dé.

En supposant que différentes « plages » de la séquence du virus HIV diffèrent par la fréquence de CG et GC le modèle M0 *ne permettra pas de les distinguer* puisqu'il propose la *même fréquence*. Mais par contre il sert de *révélateur par rapport au modèle M1*. De même le modèle M1 permettrait de *révéler de différences* sur les triplets que l'on modéliserait par un modèle M2. Il peut donc suffire pour nous laisser espérer observer une différence entre exons et introns par exemple dans l'ADN de Mammifères et de l'homme à propos des mots CG et GC, ou d'autres mots de deux lettres. Réciproquement, dans une séquence inconnue, on peut espérer détecter des plages correspondant vraisemblablement à des exons ou des introns. Il n'est pas ici indispensable de considérer des mots de trois lettres. Nous y reviendrons.

3.4. Recherche de triplets ayant une signification

Au lieu de rechercher *toutes les successions de trois lettres* correspondant à un modèle M2, ce qui conduit à examiner les fréquences de 64 triplets, ce qui n'est plus faisable « à la main », on peut se limiter à examiner la *position* et la *fréquence* de quelques triplets caractéristiques dont la fonction est *a priori* connue tels les codons STOP ou START. Il faut ici tenir compte des chevauchements éventuels, et des décalages des trois « cadres de lecture » (= phases de lecture) possibles. A priori un codon STOP ne peut se trouver au milieu d'un exon à moins qu'il ne soit pas reconnu par le cadre de lecture, ou bien que cet exon transcrit ne soit pas

traduit. Sur une séquence inconnue, l'absence de codon stop dans l'un au moins des trois cadres de lecture laisse supposer une fonction dans la synthèse d'une protéine.

3.5. Rechercher des « petits mots » de 4 à 13 nucléotides

Ici encore, au lieu d'examiner toutes les combinaisons de mots de quatre (256), cinq (1024), ou huit (65536) lettres ou plus on peut se limiter aux petits mots qui ont une signification connue telle que les *sites de fixation des enzymes de restriction* qui comprennent de 4 à 13 nucléotides, le site de réparation de huit lettres nommé Chi des bactéries (Crossover Hotspot Instigator), les promoteurs, les enhancer, etc.
Les sites de fixation des enzymes de restriction ont une signification fonctionnelle chez les bactéries et les virus bactériophages, mais n'ont aucune signification fonctionnelle chez les Mammifères ou l'homme. Leur répartition est vraisemblablement aléatoire. Chez les bactéries (Escherichia coli par exemple) et les bactériophages (par exemple lambda, bactériophage de E. Coli) les enzymes de restrictions permettent aux bactéries de découper l'ADN d'un bactériophage parasite. Il ne faut cependant pas qu'elles découpent leur propre ADN. Les sites de fixation des enzymes de restriction qui sont souvent des palindromes et plus largement tous les palindromes seront donc vraisemblablement « sous représentés » c'est-à-dire moins fréquents qu'une répartition aléatoire.

Selon un procédé habituel en biologie moléculaire, les êtres vivants deviennent des outils pour analyser d'autres êtres vivants[11]. On utilise les virus ou les enzymes comme outils pour détecter rapidement d'éventuelles mutations ponctuelles ou pour fabriquer des fragments d'ADN plus petits que l'on peut insérer dans un plasmide par exemple pour l'amplifier puis l'analyser. On peut alors observer le nombre, la position de ces sites et évaluer les distances entre les sites. Leur répartition peut être examinée sous une hypothèse aléatoire telle une loi de Gauss ou une loi de Poisson. On évalue *l'écart dans la répartition par rapport à cette loi*. Les mots sont "trop" fréquents, "trop" rares, ou totalement absent.

4. UTILISATION DU LOGICIEL ANAGENE (INRP)

On peut utiliser la banque de données du logiciel ANAGENE (INRP) et certaines de ses fonctions pour montrer le principe de la recherche :
- des fréquences des lettres isolément,
- des fréquences de successions de deux, trois lettres,
- du nombre de la position et des distances entre des petits mots le long d'une séquence.

Le logiciel réalise de manière automatique certaines de ces opérations. Il sera aisé de l'améliorer en les réalisant toutes. En attendant certaines opérations peuvent se réaliser à la main sans trop de difficulté dans la mesure où les séquences ne sont pas trop longues (1000 à 3500 nucléotides). Dans ce but, le logiciel permet *d'imprimer les séquences sur papier* en mettant la réglette supérieure sous forme de triplets pour repérer le cadre de lecture. On peut comparer dans l'ordre la séquence complète, la séquence d'ADNc (copié), la séquence de la partie strictement codante, c'est-à-dire sans les exons qui ne sont pas transcrits, ni les promoteurs, en alignant les séquences de manière à déterminer les exons et les introns, et *d'imprimer aussi ce résultat sur papier*.

4.1. Premier modèle M0 : fréquences des lettres A, C, G, T

[11] Morange Michel (2005) La place des instruments dans les transformations de la biologie au XXème siècle. *Bull. Hist. Epistém. Sci. Vie, 12* (2) 219-229

La banque de données du logiciel n'offre que trois types de gène en « entier » (pour autant qu'un gène ne soit qu'un fragment d'ADN !) accompagné de l'ADNc (ADN copié) ou du mRNA (ARN messager), et de la séquence strictement codante c'est-à-dire sans les exons qui ne sont pas traduits, ni le promoteur, mais c'est suffisant pour travailler.

Soit la séquence HLAa 0201.adn (gène qui correspond à un marqueur cellulaire du système immunitaire) de 3528 nucléotides. Elle présente un promoteur, 8 exons et 7 introns qui sont connus. Elle présente de nombreux allèles. On peut donc procéder à la détermination des fréquences des quatre lettres A, C, G, T pour vérifier empiriquement si elle est la même dans les exons et les introns. Après avoir sélectionné une séquence un clic sur I (information) donne les fréquences des quatre lettres A, C, G, T dans l'ADN entier, dans l'ADN copié et dans l'ADN strictement codant (tableau ci-dessous).

Commençons par l'observation des fréquences de A, C, G, T dans l'ADN de HLA a 0201 (3528 paires de bases, dont 1098 seulement codent pour des acides aminés). Il faudrait calculer les variances pour chaque pourcentage, mais c'est difficile à cause des trois phases de lecture, et nous ne le ferons pas ici. Les séquences d'ADN entier ne sont vraisemblablement pas homogènes et on peut s'attendre à trouver une différence entre les « plages » correspondant aux exons et aux introns. Nous avons calculé ci-dessous l'exon n°3, l'exon n°8 et l'intron n°3. Il apparaît des différences nettes pour les lettres G et T, mais ces différences ne suffisent pas pour caractériser tous les exons et les introns.
Nous donnons ci-dessous à titre de comparaison quelques exemples de résultats contrastés sur des protéines différentes pour lesquels il n'y a pas de différence de fréquence entre les quatre lettres entre les exons et les introns (FSH), ni même entre les quatre lettres qui sont proches de 25% (Tyralba). FSH est un gène qui correspond à une hormone de l'hypophyse, et Tyralba 1 est le gène de la tyrosinase, enzyme qui intervient dans la synthèse des pigments de la peau. Pour FSH on n'observe pas de différence entre exons et ADN total. Pour Tyralba les proportions sont différentes de HLA et FSH, mais nous ne disposons pas de l'ADNc.

	ADN entier	ADN exons	ADN codant	Exons n°3	Intron n°3	Exon n°8
A	19,8%	20,5%	20,0%	21,1%	20,7%	21,4%
C	27,8%	27,7%	29,4%	28,3%	28,7%	23,8%
G	30,0%	30,4%	33,5%	35,6%	24,3%	24,5%
T	22,4%	21,3%	17,0%	14,9%	26,2%	30,1%
	3528 pb	1684 pb	1098 pb	275pb	599pb	554pb

FSH « entier »	3055 nucléotides	a = 28,8%	c = 19,2%	g = 21,2%	t = 30,8%
FSH copié (exons)	2439 nucléotides	a = 28,8%	c = 20,2%	g = 21,2%	t = 29,8%

Tyralba « entier »	1590 nucléotides	a = 26,0%	c = 24,4%	g = 23,1%	t = 26,5%

4.2. Modèle M 1 : successions de deux lettres

La matrice décrite précédemment donne les 16 groupes possibles de deux lettres. ANAGENE ne permet pas actuellement de faire ce travail de manière automatique. Il faut le faire « à la main » ce qui est assez rapide surtout si on répartit le travail entre les élèves !

Le tableau ci-dessous donne pour HLAa 0201 les valeurs « attendues » sous M0 pour les groupes de deux lettres en utilisant les proportions du tableau précédent. La probabilité « attendue » de CG dans l'ADN entier s'obtient en multipliant 27,8 % par 30,0%. Ce tableau est symétrique (le % de CG = GC).

AA 3,9% AC 5,5% AG 5,9% AT 4,4%	CA 5,5% CC 7,7% CG 8,3% CT 6,2%	TA 4,4% TC 6,2% TG 6,7% TT 5,0%	GA 5,9% GC 8,3% GG 9,0% GT 6,7%
Total = 100%			

Les fréquences de transition *observées* (dénombrées à la main sur le tirage papier) sont très différentes des fréquences calculées :

- Pour CG = 3,4% (120 / 3528) au lieu de 8,3%,

et en distinguant les exons = 3,9% (65 / 1684), et les introns 2,8% (52 / 1844).
Mais pour l'exon n°3 = 8,5% et l'intron n°3 = 2,2%

- Pour TA = 2,0% (au lieu de 4,4%)

La prise en compte des successions de deux lettres permet donc cette fois d'observer des différences à propos de CG ou de TA dont on peut espérer qu'elles suffiront à caractériser les exons et les introns, et donc, réciproquement, de les rechercher dans des séquences inconnues.

Plus précisément voici quelques résultats *observés* (dénombrés à la main) pour les 16 groupes de deux lettres, exprimés en % pour l'exon n° 3 (275 nucléotides) et l'intron n°3 (599 nucléotides).

Exon 3	Intron 3	Exon 3	Intron 3
AA 2,5	3,9	TA 2,5	2,0
AC 7,5	4,9	TC 3,5	9,8
AG 7,8	6,9	TG 6,8	7,9
AT 2,5	4,6	TT 1,8	4,2
CA 7,1	6,8	GA 7,8	8,1
CC 6,8	10,1	GC 9,6	3,7
CG 8,5	2,2	GG 11,4	8,6
CT 5,3	10,1	GT 4,6	5,7

Il est à noter que AA, AT, TT et TA sont très peu fréquents dans cet exon. Ceci pourrait être mis en relation avec le fait que les codons (triplets) TAA, TAG, TGA sont des codons stop donc absents dans au moins un des cadres de lecture.
Les successions CG, ou GC sont fréquentes dans cet exon et très faibles dans l'intron. La fréquence de CG suffirait peut être à distinguer les exons et les introns, ou bien, plus largement d'autres types de « plages ». Nous verrons plus loin la signification de la succession CG.

Pour Tyralba la fréquence de toutes les successions de deux lettres devrait être 0,25*0,25 = 0,0625 = 1 / 16. On peut donc penser que chaque succession de deux lettres peut apparaître statistiquement environ 100 fois (1590 / 16). Voici quelques valeurs observées :

Proche des valeurs attendues : AA = 102, TT = 112, CC = 102,

Très inférieures : CG = 22 et GC = 84, TA = 61,

Supérieures : TG = 124

Rappelons que ce travail n'a d'intérêt que pour justifier le travail réciproque, c'est à dire la recherche de distinctions permettant de caractériser les exons et les introns dans une séquence inconnue, la distinction de toute autre plage.

5. **Analyse de quelques exemples**

Il s'agit de rechercher des "traces" d'une constitution aléatoire de l'ADN ainsi que des caractéristiques qui vont créer une "perturbation" par rapport à cet *alea*, perturbations qui ne sont détectables qu'en fonction du modèle utilisé.
On peut télécharger des séquences d'ADN à l'adresse suivante :

http:// www. migale. jouy. inra. fr /banques/
Ou bien par Google : genome jouy inra puis MIGALE
Dans la banque Gen bactériens on ne retiendra que les séquences « fna » entières et sans coupures. La banque MICADO présente les séquences sous une autre forme. On se limitera à des fragments de quelques milliers de nucléotides, car il faut beaucoup de mémoire et un ordinateur rapide.
La banque contient des virus, des bactériophages, des végétaux, des Invertébrés et des Vertébrés (fragments). Si on souhaite imprimer quelques séquences il faut se rappeler que si un virus occupe deux ou trois pages, mais parfois 15, une bactérie en occupe de 4 à 500, et un Vertébré plusieurs centaines de rames !

5.1. Successions de deux lettres

Le bactériophage Lambda, parasite de E. coli, comprend 48502 nucléotides. On trouve dans les banques (www.ebi.ac.uk) le comptage des lettres isolément et des doubles lettres dans les *trois phases*, puis au total ainsi que la séquence complète (14 pages en tirage papier) et les traductions (30 pages). On peut également télécharger cette séquence sur le site de l'académie d'Orléans : http://www.ac-orleans-tours.fr/svt/theme1/telcha2.htm

A	12334	0,25
C	11362	0,23
G	12820	0,26
T	11986	0,25

Le calcul sous M0 donnera des fréquences « attendues » identiques pour toutes les successions de deux lettres puisque les fréquences sont à peu près égales. Il en sera d'ailleurs de même pour le calcul des valeurs « attendues » pour les successions de trois lettres, etc.
Valeurs observées pour les successions de deux lettres :

	total	
AA	3693	0,076
AC	2732	0,056
AG	2574	0,053
AT	3337	0,068
GA	3256	0,067
GC	3179	0,065
GG	3613	0,075
GT	2769	0,057
CA	3214	0,066
CC	3113	0,064
CG	2497	0,051
CT	2536	0,052
TA	2173	0,045
TC	3793	0,078
TG	2676	0,052
TT	3346	0,069
TOTAL	48501	1,000

Le tableau montre des différences nettes pour certaines successions telles CG, TA, qu'il faudrait évaluer à l'aide d'un test de Chi deux. Il est difficile de rechercher « à la main » dans 14 pages les tri et tétranucléotides caractéristiques évoqués précédemment pour lesquels on s'attend à observer des différences. On peut charger la séquence dans Excel par copié / collé et définir une fonction pour l'analyser. Nous y reviendrons.

5. 2. Recherche de triplets ayant une signification connue

Le logiciel ANAGENE ne permet pas actuellement de faire la recherche des triplets de manière automatique. On peut cependant *rechercher à la* main les codons STOP, TAA, TAG, TGA dans les trois "cadres de lecture" successifs à condition de mettre la réglette supérieure par triplets au moment de l'impression sur papier.
Pour HLA a 0201 du logiciel ANAGENE on obtient les résultats suivants :

- Pour l'exon 3 TAG apparaît deux fois mais il est décalé par rapport au cadre de lecture et n'agit donc pas comme codon STOP ce qui est, bien évidemment un critère de reconnaissance,
- Pour l'intron 3 TAA apparaît deux fois, tag apparaît cinq fois et TGA apparaît 16 fois. Ce dernier serait reconnu comme codon STOP dans plusieurs cas, *quel que soit le cadre de lecture*, si l'intron était transcrit et traduit.
- Pour l'exon 8 des codons stop apparaissent quel que soit le cadre de lecture, mais si cet exon est transcrit, il n'est pas traduit.

Il faut noter que les introns et les exons ne sont pas nécessairement coupés à la fin d'un codon, mais éventuellement en plein milieu, autrement dit le cadre de lecture d'un exon ne commence pas nécessairement au premier nucléotide. On peut poursuivre ce travail sur les autres exons et les autres séquences proposées par ANAGENE.

5.3. Fréquences relatives des différents triplets dans les trois phases de lecture

On peut examiner la fréquence relative des différents triplets dans les trois phases de lecture à l'intérieur des gènes de Escherichia coli pour les 10 000 premiers codons (seulement !). On trouve dans les banques de données sur Internet le résultat suivant :

phase 0			
TTT 188	CTT 104	ATT 268	GTT 196
TTC 194	CTC 108	ATC 276	GTC 141
TTA 110	CTA 29	ATA 29	GTA 116
TTG 107	CTG 553	ATG 257	GTG 250
TCT 92	CCT 55	ACT 106	GCT 207
TCC 88	CCC 50	ACC 278	GCC 257
TCA 54	CCA 88	ACA 59	GCA 219
TCG 78	CCG 262	ACG 133	GCG 337
TAT157	CAT 119	AAT 142	GAT 309
TAC 131	CAC 111	AAC 220	GAC 226
TAA 19	CAA 141	AAA 353	GAA 405
TAG 0	CAG 309	AAG 87	GAG 185
TGT 42	CGT 222	AGT 62	GGT 280
TGC 58	CGC 223	AGC 153	GGC 323
TGA 4	CGA 29	AGA 16	GGA 58
TGG 145	CGG 42	AGG 11	GGG 99

Phase 1 TTT 111 TTC 151 TTA 185 TTG 310 TCT 120 TCC 139 TCA 205 TCG 253 TAT 39 TAC 90 TAA 83 TAG 71 TGT 156 TGC 303 TGA 302 TGG 406	CTT 79 CTC 87 CTA 86 CTG 210 CCT 105 CCC 111 CCA 186 CCG 271 CAT 61 CAC 111 CAA 93 CAG 157 CGT 99 CGC 278 CGA 167 CGG 266	ATT 94 ATC 188 ATA 183 ATG 261 ACT 110 ACC 148 ACA 169 ACG 262 AAT 102 AAC 228 AAA 242 AAG 346 AGT 99 AGC 183 AGA 126 AGG 174	GTT 115 GTC 145 GTA 157 GTG 188 GCT 120 GCC 146 GCA 166 GCG 325 GAT 18 GAC 28 GAA 36 GAG 24 GGT 38 GGC 111 GGA 65 GGG 83
2924	2367	2915	1765
Phase 2 TTT 149 TTC 102 TTA 108 TTG 40 TCT 184 TCC 105 TCA 160 TCG 121 TAT 219 TAC 141 TAA 217 TAG 32 TGT 153 TGC 289 TGA 329 TGG 198	CTT 150 CTC 90 CTA 98 CTG 118 CCT 150 CCC 100 CCA 156 CCG 138 CAT 201 CAC 173 CAA 227 CAG 125 CGT 208 CGC 281 CGA 347 CGG 275	ATT 97 ATC 48 ATA 36 ATG 38 ACT 150 ACC 93 ACA 131 ACG 83 AAT 156 AAC 120 AAA 134 AAG 44 AGT 108 AGC 173 AGA 197 AGG 119	GTT 203 GTC 71 GTA 65 GTG 52 GCT 311 GCC 158 GCA 234 GCG 173 GAT 252 GAC 143 GAA 224 GAG 41 GGT 233 GGC 277 GGA 252 GGG 167
2547	2837	1727	2856

On peut observer la fréquence des codons START et STOP, puis comparer les fréquences observées aux fréquences calculées dans le modèle « dee base » M0. On reviendra ensuite sur la fréquence comparée des codons que l'on nomme synonymes, et qui codent le même acide aminé.

TTT 188 1,9 1,2	CTT 104 0,9 1,4	ATT 268 2,8 1,3	GTT 196 1,9 1,5
TTC 194 1,8 1,4	CTC 108 1,0 1,5	ATC 276 2,7 1,4	GTC 141 1,5 1,6
TTA 110 1,1 1,3	CTA 29 0,3 1,4	ATA 29 0,3 1,3	GTA 116 1,1 1,5
TTG 107 1,2 1,5	CTG 553 5,7 1,6	ATG 257 2,5 1,5	GTG 250 2,7 1,8
TCT 92 0,9 1,4	CCT 55 0,6 1,5	ACT 106 0,9 1,4	GCT 207 1,6 1,6
TCC 88 0,9 1,5	CCC 50 0,4 1,6	ACC 278 2,5 1,5	GCC 257 2,5 1,8
TCA 54 0,6 1,4	CCA 88 0,8 1,5	ACA 59 0,5 1,4	GCA 219 2,0 1,7
TCG 78 0,8 1,6	CCG 262 2,6 1,8	ACG 133 1,4 1,7	GCG 337 3,6 2,0
TAT 157 1,5 1,3	CAT 119 1,2 1,4	AAT 142 1,4 1,3	GAT 309 3,2 1,5
TAC 131 1,4 1,4	CAC 111 1,1 1,5	AAC 220 2,4 1,4	GAC 226 2,2 1,7
TAA 19 *	CAA 141 1,3 1,4	AAA 353 3,5 1,3	GAA 405 4,3 1,6
TAG 0 *	CAG 309 3,0 1,7	AAG 87 1,1 1,6	GAG 185 1,8 1,8
TGT 42 0,5 1,5	CGT 222 2,5 1,6	AGT 62 0,7 1,5	GGT 280 2,8 1,8
TGC 58 0,7 1,6	CGC 223 2,4 1,8	AGC 153 1,5 1,7	GGC 323 3,2 2,0
TGA 4 *	CGA 29 0,3 1,7	AGA 16 0,1 1,6	GGA 58 0,6 1,8
TGG 145 1,4 1,8	CGG 42 0,4 2,0	AGG 11 0,1 1,8	GGG 99 1,0 2,2

Codons des 10 000 premiers nucléotides de Escherichia coli.
Colonne 1 : codons ; Colonne 2 : dénombrement sur les 10 000 premiers nucléotides pour la phase de lecture 0 qui est supposée codante d'après la très faible fréquence des codons STOP, par comparaison avec les phases 1 et 2.
Colonne 3 : fréquences observées sur les gènes connus des 10 000 nucléotides (phase 0)
Colonne 4 : fréquence théorique calculée dans le modèle « de base » M0

phénylalanine	leucine	isoleucine	valine
TTT 188 1,9 1,2	CTT 104 0,9 1,4	ATT 268 2,8 1,3	GTT 196 1,9 1,5
TTC 194 1,8 1,4	CTC 108 1,0 1,5	ATC 276 2,7 1,4	GTC 141 1,5 1,6
sérine	CTA 29 0,3 1,4	ATA 29 0,3 1,3	GTA 116 1,1 1,5
TCT 92 0,9 1,4	CTG 553 5,7 1,6	Methionine, start	GTG 250 2,7 1,8
TCC 88 0,9 1,5	TTA 110 1,1 1,3	ATG 257 2,5 1,5	alanine
TCA 54 0,6 1,4	TTG 107 1,2 1,5	thréonine	GCT 207 1,6 1,6
TCG 78 0,8 1,6	proline	ACT 106 0,9 1,4	GCC 257 2,5 1,8
AGT 62 0,7 1,5	CCT 55 0,6 1,5	ACC 278 2,5 1,5	GCA 219 2,0 1,7
AGC 153 1,5 1,7	CCC 50 0,4 1,6	ACA 59 0,5 1,4	GCG 337 3,6 2,0
tyrosine	CCA 88 0,8 1,5	ACG 133 1,4 1,7	Ac.aspartique
TAT 157 1,5 1,3	CCG 262 2,6 1,8	asparagine	GAT 309 3,2 1,5
TAC 131 1,4 1,4	glycine	AAT 142 1,4 1,3	GAC 226 2,2 1,7
stop	CGT 222 2,5 1,6	AAC 220 2,4 1,4	Ac.glutamique
TAA 19 *	CGC 223 2,4 1,8	lysine	GAA 405 4,3 1,6
TAG 0 *	CGA 29 0,3 1,7	AAA 353 3,5 1,3	GAG 185 1,8 1,8
TGA 4 *	CGG 42 0,4 2,0	AAG 87 1,1 1,6	glycine
cystéine	AGA 16 0,1 1,6	histidine	GGT 280 2,8 1,8
TGT 42 0,5 1,5	AGG 11 0,1 1,8	CAT 119 1,2 1,4	GGC 323 3,2 2,0
TGC 58 0,7 1,6	glutamine	CAC 111 1,1 1,5	GGA 58 0,6 1,8
tryptophane	CAA 141 1,3 1,4		GGG 99 1,0 2,2
TGG 145 1,4 1,8	CAG 309 3,0 1,7		

Fréquence des codons synonymes chez Escherichia coli

On peut à cette étape du travail *revenir dans l'autre sens*. Comment détecter dans une séquence inconnue supposée codante laquelle des trois phases est codante ? La fréquence de certains triplets est très variable d'une phase à l'autre. A la limite, pour faire comprendre le travail on pourrait s'intéresser uniquement aux trois codons stop. Ici TAG, TAA et TGA sont très peu fréquents et même TAG est nul en phase un ce qui « signe » donc certainement la phase codante. Il suffit donc de se déplacer le long de la séquence par une « fenêtre glissante » qui se décale de une puis de deux lettres et de recommencer le dénombrement. En fait cette analyse est très sensible et il peut suffire d'analyser 25 codons soit 78 lettres pour obtenir un résultat.

5.4. Nombre et position des sites de coupures des enzymes de restriction et distances entre les sites

En fait un site de quatre nucléotides a des chances d'être présent « au hasard », tandis qu'un site de huit nucléotides et plus, a peu de chances d'apparaître même une seule fois dans des séquences aussi courtes. On peut alors rechercher si la *position* des coupures et donc la *distance entre les coupures* obéissent à une loi de probabilité ou si elle s'en écarte. Si oui il faudra alors rechercher la raison biologique de cet écart. On peut procéder à la recherche du nombre de sites de coupures par les enzymes de restriction de 4 nucléotides présents dans ANAGENE.

Une séquence étant sélectionnée et une enzyme de restriction également il suffit d'un clic sur I (information) pour obtenir le *nombre de coupures*, le *nombre de fragments*, la *position* de ces coupures et la *longueur des fragments* donc *la distance* entre les coupures.

Par exemple on peut se demander combien il y a de sites AGCT correspondant à un palindrome qui permettraient la fixation et la coupure par l'enzyme Alu I.
Si la fréquence des quatre lettres est la même (à peu près 25%) comme dans le cas de Tyralba, le nombre de combinaisons de quatre nucléotides *différents* pris par groupes de quatre serait de 256 et donc leur probabilité de 1/ 256 soit 0,0039. En ne considérant que les sites composés de quatre lettres *différentes*, on pourrait examiner les 24 combinaisons de ces quatre lettres. Mais pour les huit palindromes (ACGT, AGCT, CATG, GATC, GTAC, TCGA, TGCA, CTAG) marqués d'un astérisque dans le tableau ci-dessous sept seulement correspondent à un site de fixation d'une enzyme de restriction actuellement connu. Grâce à l'une des fonctions du logiciel ANAGENE ces enzymes vont nous permettre de repérer rapidement certains sites de quatre lettres différentes.

ACGT* Mae II	CAGT	TGCA* inconnue	TGAC
ACTG	CATG* Nla III	GTCA	GTAC* Rsa I
AGCT* Alu I	GACT	TCGA* Taq I	TCAG
AGTC	GATC* Sau 3a	CTGA	CTAG* Mae I
ATGC	TAGC	CGTA	CGAT
ATCG	TACG	GCTA	GCAT

Pour la séquence complète de HLA comprenant promoteur, exon, intron, soit 3528 nucléotides on s'attendrait à trouver 3528 / 256 = 13 occurrences de AGCT (Alu I).
Pour HLAa 0201 on en observe 12. Leur position est 913, 1714, 1728, 2079, 2094, 2128, 2587, 2669, 2785, 2850, 2973, 3242. Les distances entre les sites (la longueur des fragments)

sont donc 913, 801, 68, 297, 15, 459, 82, 176, 65, 123, 269, 286. La distance moyenne entre les sites est théoriquement 256. On peut réécrire cette suite de distances dans l'ordre croissant : 15, 65, 68, 82, 123, 176, 269, 286, 297, 459, 801. Une loi de Poisson semble décrire ces valeurs mais nous ne le vérifierons pas. Si l'on ne retient que les exons strictement codant le résultat est-il différent ?

D'autres enzymes dont le site de fixation est également de quatre nucléotides permettent d'observer des nombres de coupures potentielles très variables : 9, 27, 11, 16, 4, 4, 7, 20, 4, 12, 3. Plusieurs fréquences de ces sites semblent donc "plus fréquente" ou "plus rare" que celle « attendue » dans le modèle choisi. Resterait à chercher une raison biologique.

Tableau des enzymes de restriction et des sites de fixation sur l'ADN (ANAGENE)

Dans le tableau ci-dessous la lettre « n » représentent n'importe quelle lettre A, C, G ou T ; la lettre « s » la paire C, G ; la lettre « w » la paire A, T ; la lettre « r » les purines A, G ; la lettre « y » les pyrimidines C, T ; et l'astérisque « * » le complément inversé (ex : CCTC, GAGG).

4 nucléotides

Alu I	AGCT
Tha I	CGCG
Hae III	GGCC
Hha I	GCGC
Hpa II	CCGG
Mnl I	CCTC *
Mae I	CTAG
Mae II	ACGT
Mse I	TTAA
Nla III	CATG
Rsa I	GTAC
Sau 3a	GATC
Taq I	TCGA

5 nucléotides

Alw I	GGATC *
Ava II	GGwCC
BbvI	GCwGC
Dde I	CTnAG
Eco RII	CCwGG
Fnu 4H	GCnGC
Fok I	GGATG *
Hga I	GACGC *
Hinf I	GAnTC
Hph I	GGTGA *
Mae III	GTnAC
Mbo II	GAAGA *
Nci I	CCsGG
Ple I	GAGTC *
Sau 96I	GGnCC
Scr FI	CCnGG
Sfa NI	GATGC *

6 nucléotides

Aat II	GACGTC
Acc I	GTmkAC
Afl II	CTTAAG
Afl III	AC ry GT
Aha II	Gr CG y C
Apa I	GGGCCC
Apa LI	GTGCAC
Ase I	ATTAAT
Asu II	TTCGAA
Ava I	Cy CG r G
Avr II	CCTAGG
Bal I	TGGCCA
Bam HI	GGATCC
Ban II	G r GC y C
Bcl I	TGATCA
Bgl II	AGATCT
BsmI	GAATGC *
Bsp HI	TCATGA
Bsp MI	ACCTGC *
Bsp mII	TCCGGA
Bss hII	GCGCGC
Cfr10 I	rCCGGy
Cla I	ATCGAT
Dra I	TTTAAA
Eae I	yGGCCr
Eco 47 III	AGCGCT
Eco RI	GAATTC
Eco RV	GATATC
Fsp I	TGCGCA
Hae I	wGGCCw
Hgi AI	GwGCwC
Hinc II	GTyrAC
Hind III	AAGCTT
Hpa I	GTTAAC
Kpn I	GGTACC
Ksp I	CCGCGG
Mlu I	ACGCGT

Nae I	GCCGGC
Nru I	TCGCGA
Nsi I	ATGCAT
Nar I	GGCGCC
Nco I	CCATGG
Nde I	CATATG
Nhe I	GCTAGC
Nsp I	r CATG y
Pst I	CTGCAG
Pvu I	CGATCG
Sal I	GTCGAC
Sca I	AGTACT
Sdu I	Gd GCh C
Sfu I	TTCGAA
Sma I	CCCGGG
SnaB I	TACGTA
Sna I	GTATAC
Spe I	ACTAGT
Sph I	GCATGC
Ssp I	AATATT
Sst I	GAGCTC
Sst II	CCGCGG
Stu I	AGGCCT
Sty I	CC ww GG
Xba I	TCTAGA
Xho I	CTCGAG
Xho II	r GATC y
Xma III	CGGCCG

7 nucléotides

Bst EII	GGTn ACC
Bsu36 I	CCTn ACC
Dra I I	r GGn CCy
Esp I	GCTn AGC
PpuM I	r GG w CC y
Rsr II	CGGw CCG

8 nucléotides

Not I	GCGGCCGC
Sgr AI	Cr CCGGy G

9 nucléotides

AlwN I	CAGnnnCTG
Dra III	CACnnnGTG
Tht111 I	GACnnnGTC

10 nucléotides

Xmn I	GAAnnnnTTC

11 nucléotides

Bgl I	GCCnnnnnGGC
PflM I	CCAnnnnnTGG

12 nucléotides

Bst XI	CAAnnnnnnTGG

13 nucléotides

Sfi I	GGCCnnnnnGGCC

5.5. Séquences d'ADN synonymes

En se reportant au tableau du code génétique précédent, on constate que si la méthionine est codée par un seul codon ATG, qui est aussi un codon START, certains acides aminés sont codés par plusieurs codons dits synonymes.

Thr thréonine ACA, ACC, ACG, ACT, donc A et C sont communs, la troisième base est l'une des quatre,

Gln la glutamine CAA, CAG, donc c et a sont communs, la troisième base est soit G, soit A,

Arg arginine AGA, AGG, CGA, CGC, CGG, CGT, donc seul g en deuxième position est commun à tous les codons,

Ser sérine, AGC, AGT, TCA, TCC, TCG, TCT, donc aucune base n'est commune à tous les codons.

Dans les deux premiers cas au moins on peut supposer que la présence de la troisième base est aléatoire, tirée dans une urne de Bernoulli dont la composition correspond à la fréquence des quatre bases dans la séquence considérée. En effet le même acide aminé étant codé par 2, 4 ou 6 codons les différentes séquences qui présentent l'un ou l'autre codon sont *a priori* équivalentes quant à la traduction en acide aminé.

Si l'on considère maintenant une succession d'acides aminés :

Met / Thr / Gln / Cys / Arg / Pro / Ser / Leu / Phe
A T G / A C. / C A . / T G . / . G . / C C . / . . . / . T . / TT .

Il va apparaître des *blocs* de 1, 2, 5 lettres obligées et des *lacunes* de 1, 2, 5 espaces qui peuvent être remplies par diverses bases.
La logique de remplissage peut se faire selon le modèle M0 pour chaque lacune isolément, indépendamment des bases qui suivent ou précèdent.

Mais on peut également considérer les successions de deux, trois ou 5 bases. La logique de ce remplissage peut dépendre alors de l'apparition ou de l'élimination de petits mots indépendants de la constitution de la protéine elle même. Par exemple la formation ou l'élimination de sites de fixation d'enzymes de restriction. La succession de deux sérines dont les codons ont été rappelés ci-dessus, peut faire apparaître ou non selon les codons utilisés le site de l'enzyme de restriction Alu I, soit A G C T si les codons sont successivement A,G,C et T,C,A ou n'importe lequel des trois autres TCC, TCG, TCT. Ou bien encore TCG puis AGC ou AGT. Les autres combinaisons ne le font pas apparaître. Inversement la succession A G C T peut être produite par les successions Glu Leu (gag / ctg ou gag / ctt), Ala Ala (GCA / GCT), Leu Ala (CTA / GCT), Gly Ala (GGA / GCT), Ser Cys (AGC / TGC), etc.

On peut donc établir le tableau des triplets utilisés de manière préférentielle par un organisme donné et tenter de l’interpréter.

Fréquence des différents triplets synonymes du code génétique chez le colibacille

Phe TTT 19	Ser TCT 10	Tyr TAT 15	Cys TGT 6
Phe TTC 18	Ser TCC 10	Tyr TAC 14	Cys TGC 5
Leu TTA 10	Ser TCA 6	TAA STOP	TGA STOP
Leu TTG 11	Ser TCG 8	TAG STOP	Trp TGG 13
Leu CTT 10	Pro CCT 6	His CAT 11	Arg CGT 25
Leu CTC 10	Pro CCC 6	His CAC 11	Arg CGC 22
Leu CTA 3	Pro CCA 8	Gln CAA 13	Arg CGA 3
Leu CTG 55	Pro CCG 24	Gln CAG 30	Arg CGG 4
Ile ATT 27	Thr ACT 11	Asn AAT 16	Ser AGT 7
Ile ATC 28	Thr ACC 24	Asn AAC 25	Ser AGC 15
Ile ATA 4	Thr ACA 6	Lys AAA 37	Arg AGA 2
Met ATG 27	Thr ACG 12	Lys AAG 12	Arg AGG 1
Val GTT 21	Ala GCT 18	Asp GAT 32	Gly GGT 29
Val GTC 14	Ala GCC 23	Asp GAC 23	Gly GGC 31
Val GTA 12	Ala GCA 20	Glu GAA 44	Gly GGA 7
Val GTG 25	Ala GCG 33	Glu GAG 20	Gly GGG 9

Fréquence des différents triplets synonymes du code génétique chez l'homme

Phe TTT 16	Ser TCT 13	Tyr TAT 13	Cys TGT 10
Phe TTC 23	Ser TCC 18	Tyr TAC 19	Cys TGC 15
Leu TTA 5*	Ser TCA 9	TAA STOP	TGA STOP
Leu TTG 11	Ser TCG 4*	TAG STOP	Trp TGG 14
Leu CTT 11	Pro CCT 16	His CAT 9	Arg CGT 5*
Leu CTC 20	Pro CCC 20	His CAC 14	Arg CGC 11
Leu CTA 6*	Pro CCA 14	Gln CAA 11	Arg CGA 5*
Leu CTG 43	Pro CCG 6*	Gln CAG 34	Arg CGG 4*
Ile ATT 15	Thr ACT 13	Asn AAT 17	Ser AGT 10
Ile ATC 24	Thr ACC 23	Asn AAC 23	Ser AGC 19
Ile ATA 6*	Thr ACA 14	Lys AAA 22	Arg AGA 10
Met ATG 23	Thr ACG 7*	Lys AAG 35	Arg AGG 11
Val GTT 10	Ala GCT 20	Asp GAT 22	Gly GGT 11
Val GTC 16	Ala GCC 29	Asp GAC 29	Gly GGC 25
Val GTA 6*	Ala GCA 14	Glu GAA 27	Gly GGA 17
Val GTG 31	Ala GCG 7*	Glu GAG 41	Gly GGG 17

On note que les codons peu fréquents (marqués de *) sont, chez l'homme, principalement ceux qui contiennent les deux lettres CG *précédées ou suivies* de A, C, G, T. C'est donc la faible fréquence de CG qui détermine la faible fréquence de ces huit codons. La sous représentation du dinucléotide CG dans l'ensemble du génome des Vertébrés semble liée au fait que CG constitue un signal de méthylation de la cytosine. Lorsque cette méthylation porte sur les deux brins, donc le palindrome GC, elle joue un rôle dans l'expression de certains gènes d'origine paternelle ou maternelle que l'on nomme « l'empreinte génétique » (genetic

imprinting). La thymine étant une méthyle uracile, on comprend que la cytosine mute plus aisément en thymine ce qui réduit le nombre de CG en les transformant en TG. On trouve donc une forte fréquence de TG. Toutefois la présence de TG risque de créer les triplets ATG et TGA, ce qui doit en limiter la fréquence.
Autrement dit il faut prendre en compte la fréquence des sous mots de deux lettres. D'une manière plus générale la fréquence d'un mot de 4, 5, 6, 7, 8 lettres doit être examinée en fonction des sous mots de deux, trois, quatre lettres qui le constituent. En considérant les deux lettres CG comme une seule lettre, on peut se ramener alors à un modèle de Markov d'ordre un. L'exemple ci-dessous donne les sous mots d'un mot de cinq lettres.

CGTGG
CGTG GTGG
CGT GTG TGG
CG GT TG GG

6. Retour à l'analyse bio-informatique de séquences non connues : chaînes de Markov cachées (Hidden Markov Modèles : HMM), et algorithme EM par fenêtres glissantes

Ayant réalisé ce travail sur le pouvoir discriminant des modèles M0, M1, sur des séquences connues, ainsi que la recherche directe de petits mots connus, il est maintenant possible de concevoir *le travail de recherche, dans l'autre sens* en aveugle sur des séquences inconnues de plusieurs millions de nucléotides ce qui est le travail des chercheurs en bio-informatique.

6.1. Recherche de plages homogènes.

Pour les séquences d'ADN de Mammifères on peut chercher par exemple à repérer l'existence d'exons et d'introns et supposant que deux chaînes de Markov différentes P1 et P2 ayant des *matrices de transition* différentes en rendent compte. Les limites (début / fin) des exons et introns sont inconnues. On dit encore que les chaînes de Markov sont "cachées". De même pour une bactérie si l'on recherche par exemple un fragment transféré « horizontalement » d'une bactérie dans une autre.
On peut distinguer *deux étapes dans la compréhension* de ce travail. On peut effectuer ce travail de recherche par des découpages arbitraires de la séquence, découpages que l'on déplace le long de la séquence, et que l'on affine progressivement, mais ce n'est pas réalisable à la main ! On considère par exemple des fenêtres de 200 bases se recouvrant successivement de 100 bases. Mais on peut aussi se déplacer d'une seule base à la fois !

Première étape : si on connaît les matrices de transition et qu'il n'y en a que deux P1 et P2 différentes pour les deux types de « plages » que l'on cherche à distinguer on peut calculer la vraisemblance que telle « plage » découpée arbitrairement par la fenêtre ait été fabriquée par l'une ou l'autre des matrices P1 ou P2. Bien évidemment la fenêtre découpée arbitrairement peut se situer à cheval sur deux plages. Puis on recommence en déplaçant la « fenêtre », puis en modifiant sa longueur.

Deuxième étape (inaccessible !) : si on ne connaît pas les matrices de transition, mais que l'on suppose qu'il y en a seulement deux, il est impossible de faire le travail précédent. On propose cependant le travail suivant. Il faut à chaque découpage d'une fenêtre « estimer » la matrice de transition et décider si elle correspond à l'une ou l'autre des séquences de « style » exon ou de « style » intron. Si la fenêtre suivante diffère suffisamment de la précédente on

l’attribue à l’autre « style ». En réitérant ce travail on peut s’approcher progressivement de la matrice de transition réelle. L'algorithme EM permet d'effectuer ce travail par itérations (E pour "estimation", et M pour "maximisation"), mais il n'est pas accessible en lycée, à moins de faire tourner un programme sans comprendre son fonctionnement. Le programme de Recherche R’HOM (département BIA de biométrie et intelligence artificielle de l’INRA) permet le recherche de Régions Homogènes dans les séquences d’ADN.

6.2. La Recherche de Mots Exceptionnels dans une Séquence d’ADN

Elle peut se faire avec le programme R’MES. Dans la réalité on recherche des mots de huit lettres avec des chaînes de Markov d’ordre 6, en supposant l’utilisation de plus d’une dizaine de chaînes de Markov cachées différentes correspondant aux sites promoteurs, aux sites d’initiation de la transcription, aux exons, aux introns, aux séquences intergéniques, aux virus intégrés, aux URL, aux queues poly A, etc…

Nous espérons avoir donné des indications de travail réalisable avec le logiciel ANAGENE ou « à la main », ou à partir de séquences des banques de données. Il est également possible de créer des simulations à l’aide de séquences de 0 et de 1 en utilisant le tableur Excel et son générateur aléatoire ALEA, ce que nous développerons dans un prochain article.