Université d’Évry Val d’Essonne

RÉPARTITION DE MOTIFS DANS DES SÉQUENCES D’ADN

Mémoire présenté par

Stéphane Robin

en vue de l’obtention de l’

Habilitation à Diriger des Recherches

soutenu devant un jury composé de

Pr Jean-Marc Azais	Université Paul Sabatier, Toulouse	rapporteur
Pr Camille Duby	Institut National Agronomique Paris-Grignon	
Pr Warren Ewens	University of Pennsylvania	rapporteur
Pr Bernard Prum	Université d’Évry - Val d’Essonne	rapporteur
Dr François Rodolphe	Institut National de la Recherche Agronomique	rapporteur
Dr Jean-Loup Risler	Université d’Évry - Val d’Essonne	
Pr Martin Vingron	Max Planck Institut für Moleculare Genetik	

le

1er juillet 2002

version du 27 août 2002

À Théo et Victor, mes grands

À Christel, mon épouse

Remerciements

Depuis une dizaine d'années, j'appartiens département OMIP (!) de l'INA-PG. Parmi les membres de ce département, je tiens à remercier tout spécialement Camille Duby pour son accueil lors de mon arrivée à l'agro et pour ce qu'elle m'a appris depuis. Je tiens également à remercier Jean-Jacques Daudin pour le plaisir que j'ai eu à travailler avec lui sur plusieurs articles présentés dans ce mémoire et sur d'autres à venir j'espère.

La majeure partie des travaux présentés ici ont été menés dans le cadre du groupe Statistique des Séquences Biologiques. La diversité des sujets abordés et la liberté des discussions au sein de ce groupe ont été essentielles à ma compréhension de nombreux problèmes aussi bien biologiques que statistiques. J'en remercie tous les membres et particulièrement Sophie Schbath qui l'anime.

L'unité Mathématique, Informatique et Génome m'a accueilli pendant deux ans. Cette expérience m'a convaincu de l'intérêt de faire cohabiter des chercheurs de disciplines différentes dans les mêmes murs. Je remercie François Rodolphe et ses Migales pour le plaisir que j'ai eu à travailler parmi eux.

Ce mémoire a été relu et corrigé par Christel, mon épouse favorite, qui accepté de plonger dans cette prose barbare. Pour cela, comme pour tant d'autres choses, je lui voue une reconnaissance éternelle.

Je remercie grandement Bernard Prum de ses conseils et de son soutien tout au long de la préparation de cette habilitation.

Enfin et surtout, je tiens à remercier mes parents qui m'ont donné les moyens de faire aujourd'hui un métier que j'aime.

Table des matières

Table des figures

Liste des tableaux

Notations et abréviations

Les notations données ici sont celles des chapitres I.2 et I.3 de la partie introductive de ce mémoire. Elles constituent également un « consensus » des notations utilisées dans les chapitres des parties II et III qui peuvent cependant varier d'un chapitre à l'autre. Le lecteur est invité à en vérifier la définition dans chaque chapitre.

$\mathcal{A}$	alphabet des quatre nucléotides $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$
bps	paires de bases
$\mathcal{E}(\lambda)$	loi exponentielle de Poisson de paramètre λ
$\mathbb{E}(Z)$	espérance de la variable aléatoire Z
$\phi_Z(t)$	fonction génératrice de la variable aléatoire Z
$\mathcal{G}(a)$	loi géométrique de paramètre a
i.i.d.	indépendantes et identiquement distribuées
$\mathbb{I}\{A\}$	indicatrice de l'événement A
M00	modèle de Bernoulli avec probabilités égales
M0	modèle de Bernoulli avec probabilités inégales
Mm	modèle de chaîne de Markov d'ordre m
$\boldsymbol{\mu}$	distribution stationnaire d'une chaîne de Markov
N	nombre d'occurrences du motif
$\mathcal{P}(\lambda)$	loi de Poisson de paramètre λ
PC	modèle de Poisson composé
$\mathcal{PA}$	loi de Pólya-Aeppli
$\mathcal{PC}$	loi de Poisson composée
$\pi(\cdot;\cdot)$	probabilité de transition d'un modèle Mm
$\mathbf{\Pi}$	matrice de transition d'une chaîne de Markov
$\mathbb{V}(Z)$	variance de la variable aléatoire Z
v.a.	variable aléatoire
$\mathbf{w}$	mot simple
$\mathcal{W}$	motif (ou famille de mots) : $\mathcal{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_I)$
X	position de la première occurrence du motif
X_n	position de la n-ème occurrence du motif
X'_n	position du n-ème train d'occurrences du motif
Y	distance entre deux occurrences successives du motif
Y^r	distance cumulée d'ordre r

Première partie

Introduction

Chapitre 1

Présentation générale

Sommaire

Ce mémoire présente une revue des travaux de recherche que j'ai menés depuis une dizaine d'années. Ce chapitre liminaire présente un survol plus large des différents sujets que j'ai pu aborder. La partie la plus importante est consacrée à l'étude de la répartition de motifs dans des séquences d'ADN, sujet auquel sont consacrés les chapitres 2 et 3 et les parties II et III. La dernière partie présente des reproductions d'une sélection d'articles.

II : Articles sur les occurrences de mots ou de motifs dans des chaînes de Markov.

III : Articles sur la répartition des occurrences dans des modèles de Poisson composés.

IV : Articles sur d'autres thèmes, sans rapport avec l'analyse des séquences.

1.1 Historique

A l'exception d'un récent détachement de deux ans au sein de l'unité Mathématique, Informatique et Génome (MIG) de l'INRA, j'ai effectué toute ma carrière à l'Institut National Agronomique Paris-Grignon (INA-PG), au sein de l'unité de Mathématiques où j'ai été recruté en 1991, après une année de thèse. C'est donc en tant qu'enseignant-chercheur que j'ai effectué la quasi totalité de mes travaux.

Cette première section passe en revue, dans un ordre à peu près chronologique, la plupart des sujets que j'ai abordés jusqu'à présent. Par nature, le statut d'enseignant-chercheur oblige à travailler dans des domaines variés. J'ai choisi ici trois exemples portant sur des sujets assez différents ; chacun d'eux a donné lieu à une publication reproduite dans la partie IV.

La section suivante sera consacrée aux applications en génomique qui constituent aujourd'hui l'essentiel de mon travail de recherche. Deux thèmes y sont abordés :

- la répartition de motifs dans les séquence d'ADN (sujet de ce mémoire) ;
- l'analyse des données issues de biopuces, sujet vers lequel je m'oriente à présent.

1.1.1 Durées de vie

Mon premier sujet de recherche a été celui de ma thèse effectuée sous la direction du Pr C. Huber à l'université Paris V. Il portait sur les durées de vie et plus particulièrement sur des tests non-paramétriques – les tests du *logrank* – de comparaison de survies entre des patients ayant reçu des traitements différents. L'objectif était de d'évaluer la sensibilité de ces tests à un déséquilibre dans l'attribution des traitements. Ce travail a donné lieu à une publication (Robin (1996)), cf chapitre 10).

J'ai ensuite mis de côté ce sujet mais le co-encadrement avec J.-J. Daudin (INA-PG) et S. Moez (École Nationale Vétérinaire d'Alfort) de la thèse de T. Lorino sur des *Modèles statistiques pour des données de survie corrélées* me donne l'occasion d'y retravailler.

1.1.2 Matière grasse laitière

Un contrat de recherche avec l'Association pour la Recherche dans l'Industrie Laitière (ARILAIT) a été l'occasion de mon premier travail dans le domaine agronomique, ou plus exactement agro-alimentaire. Ce contrat visait à étudier les *Propriétés rhéologiques de la matière grasse laitière* (le beurre).

L'analyse thermique différentielle est une technique utilisée pour caractériser la rhéologie (la dureté) de la matière grasse. Elle fournit des courbes (*thermogrammes*) indiquant la quantité d'énergie absorbée ou libérée par la matière grasse au cours de sa fusion ou de sa cristallisation. Pour comprendre l'influence de la composition de la matière grasse sur sa rhéologie, nous avons tenté de modéliser ces courbes composées d'un ensemble de pics en tentant de séparer ces différents pics. Cette séparation fait appel à des techniques de régression non linéaire.

Ce travail a nécessité le développement d'un logiciel permettant une analyse graphique et manuelle des thermogrammes. Cette analyse préliminaire fournit des valeurs initiales pertinentes pour ajuster le modèle non-linéaire (Robin (1999), cf chapitre 11). Ce travail a fourni les sujets de deux stages du DEA de Statistique (université d'Orsay) et d'un stage de troisième année informatique de l'INA-PG. Un autre article portant sur l'analyse d'un ensemble de thermogrammes obtenus dans le cadre d'un plan d'expériences factoriel est en cours de rédaction.

1.1.3 Distinction variétale

En dehors des travaux tournés vers la génomique, mon sujet de recherche le plus important ces dernières années porte sur la distinction variétale qui est une des missions du Groupe d'Étude et de contrôle des Variétés et des Semences (GEVES). Le GEVES est en charge de l'homologation des variétés végétales nouvelles. Cette homologation ne peut être obtenue pour une nouvelle variété que si cette variété est suffisamment distincte (d'un point de vue phénotypique) des variétés déjà référencées. Cette étape de distinction nécessite des dispositifs expérimentaux de plus en plus lourds du fait de l'accroissement de la collection de référence.

Sur la proposition de C. Baril (responsable scientifique du GEVES), nous avons tenté de prendre en compte une information issue du typage moléculaire des variétés afin de réduire le nombre de comparaisons entre variétés. Pour cela, nous avons choisi de prédire la valeur des différentes variables phénotypiques entrant dans les critères de distinction à partir des données moléculaires, puis de calculer des distances entre ces phénotypes prédits. Ces distances prédites permettent ensuite de déterminer quelles variétés devront être effectivement comparées phénotypiquement, c'est à dire au moyen d'une implantation au champ.

Ces travaux ont donné lieu à deux stages d'étudiants du DEA de Statistique dont le premier, effectué par G. Nuel constitue, à mon sens, un exemple de réussite. Ce stage a abouti à deux publications (Nuel *et al.* 2001a et 2001b) dont la première est présentée au chapitre 12. Ces travaux ont également permis le développement, au GEVES, d'un programme de *Prédiction de distances phénotypiques* (PREDIP) par S. Paillard. Ils se poursuivent actuellement par la thèse de F. Mortier que je co-encadre avec A. Bar-Hen (Université de Marseille) et C. Baril. L'objectif de cette thèse est de généraliser la prédiction de distances au cas où les variables phénotypiques sont à la fois quantitatives et qualitatives.

1.2 Statistique et génome

Enfin, depuis environ cinq ans, l'essentiel de mon travail de recherche se concentre autour des applications en génomique. Cette recherche se fait au sein du groupe *Statistique des Séquences Biologiques* (SSB) dont on trouve une présentation complète (composition, publications, logiciels) sur le site maintenu par S. Schbath à l'adresse suivante :

`www-mig.versailles.inra.fr/ssb/`.

Ce groupe réunit aujourd'hui une vingtaine de chercheurs, majoritairement statisticiens, issus de quatre unités :
- Laboratoire *Statistique et Génome* de l'université d'Évry,
- Unité *Mathématique, Informatique et Génome* de l'INRA de Jouy-en-Josas,
- *Atelier Biologie Informatique Statistiques Sociolinguistique* de l'université de Rouen,
- Unité mixte de recherche INA-PG / INRA de Biométrie.

Les activités de ce groupe ont accompagné l'explosion de la bioinformatique en travaillant sur différents problèmes comme les alignements de séquences, les scores associés à de tels alignements, la détection de gènes, les statistiques de motifs, l'analyse des données de biopuces, *etc.* Les deux derniers sujets sont ceux auxquels je me suis essentiellement consacré ; l'article de Bacro *et al.* (2002) porte toutefois sur les scores de séquences qui y sont associés.

1.2.1 Statistique des motifs

Les statistiques des motifs ont constitué le sujet pionnier des recherches du groupe SSB. Le problème initial étant de détecter des motifs (c'est à dire des mots des quelques lettres dans l'alphabet $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$) significativement sur- ou sous-représentés. Pour ce type d'analyse, les chaînes de Markov constituent le modèle de référence. C'est dans ce cadre que J.-J. Daudin et moi-même avons travaillé sur les occurrences de mots dans des séquences aléatoires de lettres (Robin et Daudin (1999) et (2001), cf chapitres 4 et 5) et notamment sur leur répartition le long de ces séquences. Ces travaux ont rejoint ceux portant sur les fréquences (Robin et Schbath (2001), cf chapitre 6) et se poursuivent encore dans le cadre du modèle de Poisson (Robin (2002), cf chapitres 8 et 9).

Ce thème constitue le sujet central de ce mémoire et une présentation complète en est donnée aux chapitres 2 et 3.

1.2.2 Données issues de biopuces

Analyse statistique du transcriptome

Parmi les avancées technologiques récentes en biologie moléculaire, les « biopuces » ou « puces à ADN » constituent sans doute l'une des plus prometteuses. Ces puces peuvent contenir plusieurs dizaines de milliers d'ADN complémentaires et permettent notamment de mesurer le niveau d'expression de la quasi totalité des gènes d'un organisme. Cette mesure est fondée sur la quantité d'ARN transcrits, d'où le nom de *transcriptome*.

Cependant, l'abondance de ces mesures, leur sensibilité aux conditions expérimentales, leur variabilité intrinsèque en rendent l'interprétation très délicate (Sekowska *et al.* (2001)) et le recours à des outils statistiques apparaît aujourd'hui nécessaire même aux biologistes les plus rétifs.

Une grande variété de problèmes

L'analyse du transcriptome est aujourd'hui un des piliers de la *génomique fonctionnelle* qui vise à déterminer la fonction des différents gènes d'un organisme. Une expérience typique d'analyse de transcriptome consiste à mesurer le niveau d'expression des gènes d'un organisme dans un ensemble de conditions, ou à des instants successifs, ou dans un ensemble de tissus. L'analyse de ces données mobilise un très grand nombre de méthodes statistiques.

Normalisation des données : Comment corriger un bruit de fond hétérogène sur le support ? Faut-il retrancher ce bruit de fond des mesures ?

Classification : Peut-on définir des groupes de gènes ayant des profils d'expression similaires dans une série de conditions ? La typologie obtenue est-elle « répétable » ?

Analyse différentielle : Quels sont les gènes dont le niveau d'expression varie significativement entre deux conditions ? Quels sont les gènes qui ont des variations comparables ?

Discrimination : Peut-on distinguer, par exemple, un tissu sain d'un tissu cancéreux à partir des niveaux d'expression d'un ensemble de gènes ? Comment peut-on déterminer cet ensemble ?

Réseau de régulations : Peut-on déduire de mesures dynamiques les relations entre les expressions des différents gènes ? Peut-on notamment déterminer si l'expression de certains gènes favorise ou réprime l'expression de certains autres et accéder ainsi aux « réseaux de régulation » ?

Un grand nombre de méthodes statistiques classiques peuvent être utilisées pour tenter de répondre à ces questions. Le diffusion de ces méthodes est une de mes contributions à la recherche dans ce domaine depuis environ deux ans. Cette diffusion se fait par des séminaires individuels ou par l'organisation de manifestations de plus grande ampleur (voir paragraphe 1.B.3). À ce titre, mon apport à ce domaine provient, pour l'instant, tout autant des mes activités d'enseignement que de mes activités de recherche proprement dites.

Des adaptations et des développements

Cependant, pour chacune des méthodes énumérées plus haut, le nombre de gènes à considérer simultanément induit une difficulté supplémentaire. Ainsi des problèmes statistiques anciens sont reposés par ces données et des développements méthodologiques seront certainement nécessaires pour les résoudre de façon satisfaisante. Citons ici deux exemples.

Comparaison multiples : En analyse différentielle, le problème des comparaisons multiples se pose presque systématiquement : il s'agit alors de savoir comment contrôler le niveau global de l'ensemble des comparaisons deux à deux de 10 000 ou 50 000 gènes.

Test de Kruskal-Wallis : Les propriétés de ce test utilisé en analyse de la variance non-paramétrique sont connus asymptotiquement. Le cadre asymptotique usuel suppose que le nombre de répétitions tend vers l'infini, ce qui n'est pas le cas dans les expériences de transcriptome. Par contre, le grand nombre de gènes suggère un autre cadre dans lequel c'est le nombre de niveau du facteur « gène » qui tend vers l'infini. Ce changement de cadre a notamment des conséquences sur la variance de la statistique de test.

D'autre part, l'apparition de ces données coïncide avec la diffusion de méthodes dont les propriétés statistiques ne sont connues que depuis peu. C'est le cas des *Support Vector Machines* (SVM) qui sont des algorithmes de discrimination extrêmement efficaces déjà utilisés pour distinguer des tissus de types différents (sain / cancéreux, par exemple). Comme souvent dans ce type de problème, il convient d'éviter les problèmes de sur-ajustement aux données d'apprentissage, ce qui peut être fait grâce à des techniques de choix de modèle.

Bibliographie

Bacro, J.-N., Daudin, J.-J., Mercier, S. et Robin, S. (2002). Back to the local score in the logarithmic case: a direct and simple proof. *Ann. Inst. Statist. Math.* (*to appear*).

Nuel, G., Robin, S. et Baril, C. (2001a). Predicting phenotypic distances using a linear model: the case of varietal distinctness. *J. Appl. Statist.* **28 (5)** 607–621.

Nuel, G., Baril, C. et Robin, S. (2001b). Varietal distinctness assisted by molecular markers: A methodological approach. *Acta Horticulturae.* **546** 65–71.

Robin, S. (1996). Sensitivity analysis of non-parametric tests for censored data. *Commun. Statist. - Theory Meth.* **25 (8)** 1693–1715.

Robin, S. (1999). A model for thermograms. *Biometrics.* **55 (1)** 37–43.

Robin, S. (2002). A compound Poisson model for words occurrences in DNA sequences. *J. R. Statist. Soc. C.* (*to appear*).

Robin, S. et Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36** 179–193.

Robin, S. et Daudin, J.-J. (2001). Exact distribution of the distances between any occurrences of a set of words. *Ann. Inst. Statist. Math.* **36 (4)** 895–905.

Robin, S. et Schbath, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.* **8 (4)** 349–359.

Sahli, A., Trécourt, P. et Robin, S. (1997). One sided acceptance sampling by variables: the case of the Laplace distribution. *Commun. Statist. - Theory Meth.* **26 (11)** 2817–2834.

Sekowska, A., Robin, S., Daudin, J.-J., Hénaut, A. et Danchin, A. (2001). Extracting biological information from dna arrays: an unexpected link between arginine and methionine metabolism in *bacillus subtilis*. *Genome Biology*. **2 (6)** `http://genomebiology.com/2001/2/6/research/0019.1`.

Annexe

1.A Encadrement d'étudiants

1.A.1 Encadrements de stages d'étudiants

- G. Bourguignon (3-ème année INA-PG, 92) : stage au laboratoire de Biométrie de l'INA-PG sur le thème *Mise au point d'un logiciel d'analyse de thermogrammes*. Ce stage a permis la mise au point du logiciel ANATHERM.
- E. Gautherat (DEA de Statistique de Paris Sud, 93) : stage au laboratoire de Biométrie de l'INA-PG sur le thème *Déconvolution de thermogrammes*.
- J. Labarre (3-ème année INA-PG, 94) : stage à l'Association Française du Froid (CEMAGREF) sur la mise en place et l'analyse des résultats d'une *Enquête sur les fluides frigorigènes utilisés par l'industrie agro-alimentaire*.
- D. Belhaj (DEA de Statistique de Paris Sud, 95) : stage au CEMAGREF sur le thème *Authentification automatique de jus de fruits à partir de chromatogrammes*.
- G. Nuel (DEA de Statistique de Paris Sud, 98) : stage conjoint à l'INA-PG (laboratoire de biométrie) et au GEVES sur le thème *Prédictions de distances phénotypiques à partir de données moléculaire pour la distinction variétale*. Ce stage a donné lieu à deux publications et à plusieurs conférences.
- N. Legall (DEA de Statistique de Paris 6, 99) : stage conjoint à l'INA-PG (laboratoire de biométrie) et au GEVES sur le thème *Prédiction de distance phénotypique fondées sur des variables discrètes à partir de données moléculaires*
- C. Jubin (DESS de bioinformatique de Versailles, 99) : stage au laboratoire de Biométrie de l'INA-PG *Insertion au serveur Statistiques des séquences Biologiques d'un programme calculant la distribution exacte des occurrences d'un mot dans une séquence*. Ce stage a permis la mise au point du logiciel DEMOS, bientôt disponible sur le web.
- M. Lecomte (2-ème année ENSAI, 01) : stage au laboratoire INRA-MIG sur le thème *Classification de mutants d'Arabidopsis à partir de spectres infrarouges*.
- S. Ledent (DEA de Modélisation stochastique de Paris 7, 01) : stage au laboratoire INRA-MIG sur le thème *Répartition de motifs dans des séquences d'ADN hétérogènes*.

1.A.2 Encadrements et comités de thèses

Co-encadrements de thèses

- T. Lorino (École Nationale Vétérinaire d'Alfort) *Modèles statistiques pour des données de survies corrélées*, codirection ($3 \times 1/3$) avec J.J. Daudin (INA-PG) et M. Sanaa (ENVA), débutée en septembre 98 soutenue le 21 Mai 2002
- F. Mortier (Univ. de Lille) *Généralisation de la distance de Mahalanobis au mélange de données continues et discrètes: Application à la distinction variétale*, codirection ($3 \times 1/3$) avec A. Bar-Hen (Univ. Marseille-Luminy) et C. Baril (GEVES), débutée en novembre 99

- P. Delmar (Ecole Centrale) *Analyse statistique des données de transcriptome*, codirection ($3 \times 1/3$) avec C. Saguez (école Centrale) et J.J. Daudin (INA-PG)

Comité de thèses

S. Degrelle, *Croissance et différenciation de l'embryon de mammifères en début de gestation : étude par génomique fonctionnelle de l'embryon bovin normal et cloné*, INRA Jouy-en-Josas

1.B Publications et animation scientifique

1.B.1 Publications

Publications dans des revues à comité de lecture

Voir bibliographie page 18.

Thèse de doctorat

Thèse de Doctorat en Statistique de l'université Paris V intitulée *Analyse de sensibilité de tests non-paramétriques adaptés aux données censurées* sous la direction de Mme C. Huber soutenue en décembre 1993 devant un jury composé de

- Mme C. Huber, professeur à Paris V, présidente,
- Mme C. Duby, professeur à l'INA-PG, rapporteur,
- M T. Moreau, directeur de recherche à l'INSERM, rapporteur,
- M J. Bretagnolle, professeur à Paris XI,
- M B. Bru, professeur à Paris V,
- M J.-J. Daudin, professeur à l'INA-PG
- M C. Gouriéroux, professeur au Centre de Recherche en Économie et Statistique Théoriques (CREST).

Mention très honorable. (La période de la thèse a été interrompue par un an de service national.)

Publications d'enseignement

- Daudin J.J., Robin S., Vuillet C. (1999), *Inférence Statistique : Idées, démarches, exemples*, Presses Universitaires de Rennes / Société Française de Statistique
- Polycopiés de l'INA-PG
  - *Analyse de composantes principales* (50 pages) ;
  - *Introduction aux chaînes de Markov* (70 pages) ;
  - *Simulating random variables* (35 pages) ;
  - notes de cours.
- Polycopiés du Cours Supérieur d'Amélioration Génétique des Animaux Domestiques (Rappels de probabilités, Estimation, Tests, Modèle linéaire : 200 pages)

1.B.2 Congrès et séminaires

Congrès

- *Analyse de sensibilité de tests non-paramétriques adaptés aux données censurées* : Journée de Statistique de l'Association pour la Statistique et ses Utilisateurs (Vannes, 93).
- *Modélisation de thermogrammes* : Journées Agro-Industries et Méthodes Statistiques de l'ASU (Dijon, 95).
- *Exact distribution of words occurrences in DNA sequences* :
  - Mathematical Analysis of Biological Sequences (Rouen, 97),
  - Workshop on Mathematical and Statistical Aspects of Molecular Biology (Cardiff, 98),
  - Journées de Statistique, Société Française de Statistique (Grenoble 99),
  - Workshop in bioinformatics and Statistical Genetics (Göteborg, 00),
  - Garchy Seminar on Mathematical Statistics and Applications, (Garchy, France, 00),
  - Journées MAS, Phénomènes et Méthodes Stochastiques (Rennes, 00),
  - Journées Traitement et Analyse des Séquences (Évry, 00),
  - Journées de Génomique de l'Institut Elie Cartan (Nancy, 01).
- *Predicting phenotypic distances through mocular data* : Electrophoresis (conférence invitée, Rouen, 99).
- *Comparaison de différentes approximations de la loi du comptage d'un mot*, poster, Journées Ouvertes en Biologie Moléculaires, (Montpellier, 00)
- *Statistical results in « in silico » analyses of sequences*, Computational Molecular Biology, Eurandom (Eindhoven, 01)
- *Méthodes de classifications*, Journées Analyse Statistique du Transcriptome (Paris, 01)

Séminaires

- *Analyse de sensibilité de tests non-paramétriques adaptés aux données censurées* :
  - Séminaire du Laboratoire de Statistique de Paris V (93),
  - Séminaire de Statistique du CREST (94).
- *Occurrences de motifs dans une séquence aléatoire de lettres* :
  - Séminaire « jeunes chercheurs » Paris VI-ISUP (97),
  - Séminaire du laboratoire de Probabilité et Statistique de l'université de Lille (99),
  - Séminaire du laboratoire de Probabilité de l'université de Nancy (00),
  - Séminaire du laboratoire de Biométrie de l'INRA-Jouy en Josas (01).
  - Séminaire « Statistique et Génome » de l'Ecole Nationale Supérieure (02).
- *Quelques exemples d'analyse d'expérience de transcriptome*, Séminaire du Centre de génétique moléculaire et cellulaire, CNRS / Université Claude Bernard Lyon I (02).

1.B.3 Animation scientifique

La communauté « bioinformatique » française est en grande partie fédérée par une action intitulée *Informatique, Mathématique, Physique pour la Génomique* (IMPG). Cette action est structurée en groupes qui ont pour vocation de faciliter les échanges entre les chercheurs des différentes disciplines concernées. J'ai participé à l'organisation de la réunion du groupe *Traitement et Analyse des séquences* (novembre 1999, environ 200 personnes) et j'anime le groupe *Analyse Statistique du Transcriptome* dont la première réunion s'est tenue en juin 2001 (150 personnes), la prochaine aura lieu en avril 2002.

D'autre part, je participe à la mise en place d'un atelier de formation INSERM sur les méthodes statistiques pour l'analyse du transcriptome pour 2003.

Chapitre 2

Occurrences de motifs dans les séquences d'ADN

Sommaire

2.1 Intérêt biologique

Les *statistiques sur les motifs* font partie des outils élémentaires de l'analyse des génomes. Elles consistent à étudier la fréquence ou la répartition de motifs dans une ou plusieurs séquences pour en déduire soit des propriétés fonctionnelles spécifiques de ces motifs, soit des caractéristiques plus large de la séquence étudiée.

2.1.1 Vocabulaire

Formellement, une séquence d'ADN sera considérée ici comme une suite de lettres issues de l'*alphabet* de quatre nucléotides $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. L'ADN ayant deux brins (complémentaires), chacune de ces lettres correspond en fait à un couple de nucléotides (ou bases). C'est pourquoi les positions des mots ou les distances les séparant sont exprimées en *paires de bases* (bps).

On s'intéresse ici aux occurrences de *mots* ou de *motifs* dans une suite aléatoire de lettres. Les termes « mot » et « motif » n'ont pas de définition universellement admise dans l'analyse des génomes. Les définitions données ici sont donc purement locales.

Mot : Un mot $\mathbf{w}$ est une suite *parfaitement* définie de lettres de l'alphabet $\mathcal{A}$. Par exemple, `gctggtgg` ou `gatc` sont des mots.

Motif : Un motif $\mathcal{W}$ est un ensemble de mots : $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots\}$. En pratique il s'agira toujours de mots semblables entre eux à quelques différences près. Par exemple `gNtggtgg` (où `N` est n'importe quelle lettre de $\mathcal{A}$) est un motif : il correspond à l'ensemble de mots

$$\mathcal{W} = \{\mathtt{gatggtgg}, \mathtt{gctggtgg}, \mathtt{ggtggtgg}, \mathtt{gttggtgg}\}.$$

2.1.2 Trois exemples

On présente ici trois exemples de mots ou de motifs dont l'analyse fait appel à des méthodes statistiques présentées dans ce mémoire. Dans le premier exemple on cherche à connaître un motif inconnu *a priori* ; dans les deux suivants on s'intéresse à la répartition des occurrences d'un motif connu *a priori*.

Motifs promoteurs

Les motifs promoteurs sont des motifs situés en amont des gènes. C'est sur ces motifs que vient se greffer l'enzyme (la polymérase) responsable de la transcription du gène situé en aval. Les motifs promoteurs jouent une rôle central dans la régulation de l'expression des gènes et sont donc activement recherchés.

Les connaissances sur la structure physique de la polymérase peuvent permettre de dresser le portrait-robot d'un motif promoteur. Ainsi, chez *B. subtilis*, ces motifs doivent être composés de deux mots de 6 lettres, séparés d'une distance comprise entre 16 et 18 lettres. Au sein d'une même espèce, un grand nombre de gènes sont précédés de motifs semblables. On peut ainsi espérer reconnaître ces motifs en recherchant dans les régions situées en amont des gènes (les *régions promotrices*) des motifs correspondant à ce portrait-robot. Un motif présent *significativement* souvent dans ces régions sera un bon candidat promoteur.

Motifs CHI

Les motifs CHI (pour *Crossover Hotspot Instigator*) existent chez plusieurs espèces bactériennes. Chez *E. coli* il s'agit du mot `gctggtgg`, chez *H. influenzae* c'est le motif $\mathcal{W} = \texttt{gNtggtgg}$ déjà vu plus haut. Comme son nom l'indique, ce motif est impliqué dans le processus de recombinaison, mais il possède également une fonction de protection contre la dégration du génome par une enzyme (RecBCD chez *E. coli*). Cet enzyme progresse le long du double brin d'ADN en le dégradant : ce mécanisme s'arrête lorsque l'enzyme rencontre une occurrence du CHI.

Compte tenu de cette fonction de protection, on s'attend à ce que ce motif soit particulièrement fréquent dans le génome. C'est ce que montre en effet la recherche de mots de fréquence exceptionnelle brièvement présentée au paragraphe 2.3. Au delà de cette sur-représentation globale, il est intéressant d'étudier sa répartition le long du génome en faisant l'hypothèse suivante : si certaines régions contiennent une densité particulièrement élevée d'occurrences du CHI, c'est sans doute que ces régions doivent être particulièrement protégées, donc qu'elles ont une fonction importante pour l'organisme. Dans ce cas, on d'intéressera naturellement à détecter des régions *significativement* riches en occurrences de ce motif.

Palindromes

Les palindromes constituent un exemple symétrique du CHI. Un palindrome est un mot *auto-complémentaire*, c'est à dire que lu de droite à gauche en utilisant les règles d'appariement de Watson-Crick ($\texttt{a} - \texttt{t}$, $\texttt{g} - \texttt{c}$), on retrouve le même mot. `gttaac` est un palindrome comme le montre le schéma suivant :

$$\begin{array}{cccccc} \texttt{g} & \texttt{t} & \texttt{t} & \texttt{a} & \texttt{a} & \texttt{c} \\ | & | & | & | & | & | \\ \texttt{c} & \texttt{a} & \texttt{a} & \texttt{t} & \texttt{t} & \texttt{g} \end{array}.$$

Ainsi, un palindrome se retrouve de façon symétrique sur les deux brins complémentaires de la molécule d'ADN.

Chez *E. Coli*, les palindromes de longueur 6 sont connus pour être des sites de restriction, c'est à dire des sites susceptibles de donner lieu à une cassure du génome. Cette propriété est sans doute impliquée dans un mécanisme de défense du génome. Contrairement au CHI, ces palindromes sont peu fréquents dans le génome de *E. coli*, ce qui n'est pas surprenant puisqu'ils constituent des points de fragilité. En faisant un raisonnement analogue à celui fait pour le CHI, on sera donc tenté de rechercher des zones *significativement* pauvres en palindromes.

Analyse statistique

Dans chacun de ces exemples, le mot *significativement* apparaît pour marquer le recours nécessaire à l'outil statistique. En effet, parmi tous les motifs candidats à être des promoteurs, il y en existe nécessairement un plus fréquent que les autres. De même il existe forcément une région plus riche que les autres en occurrences du CHI. La question est de savoir si ce motif le plus fréquent ou cette région la plus riche recèle vraiment une information.

Pour chacun de ces exemples, l'outil statistique permet de déterminer si l'observation s'écarte vraiment de ce à quoi on peut s'attendre, ou si elle peut être le fruit du hasard. Dans le premier cas, on conclura que le phénomène observé est digne d'intérêt, dans le second on l'oubliera. Bien entendu, pour avoir une idée de ce à quoi on peut s'attendre, il faut disposer d'un modèle permettant de décrire l'ensemble des possibles.

2.2 Les chaînes de Markov homogènes et leur interprétation

2.2.1 Chaînes de Markov homogènes

La modélisation des séquences d'ADN par des chaînes de Markov vient assez naturellement à l'esprit du statisticien : c'est une modélisation assez simple qui permet de décrire un enchaînement d'états (ici les quatre lettres `a`, `c`, `g`, `t`) et dont les propriétés probabilistes sont bien connues.

Une telle modélisation suppose que la séquence d'ADN est constituée lettre après lettre, chaque lettre étant tirée au sort dans l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. On note

$$\mathbf{S} = (S_1 S_2 \dots S_\ell)$$

une séquence (aléatoire) de ℓ lettres. Un modèle de chaîne de Markov homogène suppose que la probabilité que la lettre S_x soit un a (pour $a \in \mathcal{A}$) dépend des m lettres précédentes $(S_{x-m}, S_{x-m+1}, \dots, S_{x-1})$ et, conditionnellement à ces m lettres, est indépendante des lettres antérieure à S_{x-m-1}. En d'autres termes, toute l'information sur S_x apportée par le passé est contenue dans les m dernières lettres. Le nombre entier m est l'*ordre* de la chaîne de Markov, il mesure la mémoire du processus.

La dépendance d'une lettre vis à vis des m précédentes est caractérisée par la *probabilité de transition* (ou *transition* tout court)

$$\pi(a_1, \dots, a_m; b) = \Pr\{S_x = b | S_{x-m} = a_1, \dots S_{x-1} = a_m\}.$$

L'hypothèse d'homogénéité de la chaîne requiert que cette probabilité de transition ne varie pas le long de la séquence (*i.e.* ne dépende pas de la position x).

Notation pour différentes chaînes de Markov

Différentes versions de cette modélisation peuvent être envisagées. On en définit ici les notations.

M00 est le modèle le plus élémentaire : il suppose que les lettres sont tirées indépendamment des précédentes et qu'elles sont équiprobables ($\pi(a) \equiv 1/4$). Ce modèle est parfois appelé modèle de Bernoulli « équilibré » ou « uniforme ».

M0 suppose toujours l'indépendance entre les lettres mais prévoit que les probabilités $\pi(a)$ puissent être différentes pour les différentes lettres a. On parle alors de modèle de Bernoulli.

M1 est le modèle de chaîne de Markov d'ordre 1, archétype des chaînes de Markov. Il est caractérisé par les probabilités de transition $\pi(a, b)$ (pour $a, b \in \mathcal{A}$).

Mm désigne de façon générale le modèle de chaîne de Markov d'ordre m.

2.2.2 Chaîne de Markov d'ordre 1

Le modèle M1 est, sans conteste, le modèle de chaîne de Markov le plus utilisé (dans ce mémoire comme ailleurs). Cette popularité est évidemment due à sa simplicité. Elle vient aussi du fait qu'on peut, en théorie du moins, généraliser les propriétés obtenues sous M1 au modèle Mm au moyen de la technique du *changement d'alphabet.*

On peut en effet définir une chaîne de Markov d'ordre 2 comme une chaîne de Markov d'ordre 1 sur l'alphabet $\mathcal{A}^2$ constitué de toutes les « lettres » [`aa`], [`ac`], [`ag`], *etc.* (les crochets sont là pour rappeler que [`ac`] constitue une seule lettre de $\mathcal{A}^2$). Les transitions π^* du modèle M1 sur l'alphabet $\mathcal{A}^2$ sont définies à partir des transitions π du modèle M2 sur $\mathcal{A}$ de la façon suivante :

$$\begin{array}{rcll} \pi^*([a_1 a_2]; [b_1 b_2]) & = & \pi(a_1, a_2; b_2) & \text{si } a_2 = b_1, \\ & = & 0 & \text{sinon.} \end{array}$$

De même, le modèle Mm peut se définir comme un modèle M1 sur $\mathcal{A}^m$:

$$\begin{array}{rcll} \pi^*([a_1 \dots a_m]; [b_1 \dots b_m]) & = & \pi(a_1, \dots, a_m; b_m) & \text{si } a_2 = b_1, \dots, a_m = b_{m-1}, \\ & = & 0 & \text{sinon.} \end{array}$$

Cette technique est décrite en détail dans la thèse de Nuel (2001). Il faut cependant noter que les transitions π^* ainsi définies sont très majoritairement nulles, ce qui induit des problèmes numériques qui limitent parfois la mise en pratique de cette généralisation.

Distribution stationnaire

Une propriété importante du modèle M1 est l'existence d'une distribution stationnaire $\boldsymbol{\mu}$ vers laquelle la chaîne de Markov converge. Cette propriété est vraie sous les hypothèses d'irréductibilité et d'apériodicité, en pratique toujours vérifiées par les séquences d'ADN. La distribution $\boldsymbol{\mu} = [\mu(\mathtt{a}), \mu(\mathtt{c}), \mu(\mathtt{g}), \mu(\mathtt{t})]$ est caractérisée par la propriété (vraie pour toute lettre b de $\mathcal{A}$):

$$\mu(b) = \sum_{a \in \mathcal{A}} \mu(a)\pi(a; b).$$

Ainsi, si la lettre S_x est distribuée selon $\boldsymbol{\mu}$, la lettre S_{x+1} et toutes les suivantes le sont également.

La convergence vers cette distribution signifie que, quelle que soit la première lettre de la séquence, la distribution de la lettre S_x se rapproche de $\boldsymbol{\mu}$ quand x croît :

$$\lim_{x \to \infty} \Pr\{S_x = a\} = \mu(a).$$

Cette convergence se fait à vitesse exponentielle. Comme, de plus, on étudie des séquences de grande longueur, on s'autorise en général à considérer que la chaîne de Markov est stationnaire, c'est à dire que, en toute position x, on a $\Pr\{S_x = a\} = \mu(a)$. Ainsi, le mot $\mathbf{w} = (w_1 \dots w_k)$ a la probabilité

$$\mu(\mathbf{w}) = \mu(w_1) \prod_{u=2}^{k} \pi(w_{u-1}; w_u) \tag{2.1}$$

d'apparaître en une position quelconque de la séquence.

2.2.3 Interprétation des chaînes de Markov

La modélisation des séquences biologiques par des chaînes de Markov peut sembler contestable et est effectivement contestée par bon nombre de biologistes. L'idée qu'une séquence d'ADN puisse être comparée à une séquence de lettres produite par un « singe dactylographe » ayant plus ou moins de mémoire est en effet assez difficile à admettre.

En fait, cette modélisation se justifie plutôt en termes d'information contenue dans le modèle. Comme on l'a vu plus haut, le modèle Mm est caractérisé par les transitions $\pi(a_1, \dots, a_m; b)$ dont l'estimation naturelle (qui est également celle du maximum de vraisemblance) est

$$\hat{\pi}(a_1, \dots, a_m; b) = \frac{N(a_1 \dots a_m b)}{N(a_1 \dots a_m)}$$

où $N(a_1 \dots a_m b)$ (resp. $N(a_1 \dots a_m)$) est le nombre d'occurrences du mot $(a_1 \dots a_m b)$ (resp. $(a_1 \dots a_m)$) dans la séquence. Ainsi, l'estimation des transitions du modèle Mm est essentiellement fondée sur le comptage de tous les mots de longueur $m+1$ (ou « $(m+1)$-*mots* »).

On peut donc voir le modèle Mm comme le modèle qui prend en compte les fréquences de tous les $(m+1)$-mots. L'ordre m mesure ainsi la quantité d'information contenue *a priori* dans la modélisation. Cette interprétation est certainement beaucoup plus satisfaisante que celle du singe dactylographe et on a ici l'exemple d'une modélisation probabiliste justifiée par son interprétation statistique.

Modèle de permutation

Mm n'est pas le seul modèle possible prenant en compte les fréquences de tous les $(m+1)$-mots. Un autre modèle naturel suppose que la séquence d'ADN observée a été tirée au hasard (uniforme) dans l'ensemble des séquences de même longueur obtenues par permutations des lettres mais en respectant tous les comptages $N(a_1 \dots a_{m+1})$. Les propriétés théoriques de ce modèle sont plus mal connues que celles des chaînes de Markov, essentiellement à cause de la difficulté combinatoire. Ce modèle peut cependant être utilisé pour des simulations mais, là encore, l'algorithme permettant de simuler des séquences respectant les comptages en $(m+1)$-mots est assez complexe.

Le modèle de permutation et les chaînes de Markov sont deux modèles équivalents en termes d'interprétation biologiques : aucun des deux n'est vraiment plus satisfaisant que l'autre. L'utilisation plus fréquente des chaînes de Markov provient simplement du fait qu'elles sont mieux maîtrisées d'un point de vue théorique.

Il est important de noter que le modèle Mm et le *modèle de permutation* sur les $(m+1)$-mots sont égaux « en espérance » mais pas « en variance ». En effet, même si toutes les transitions sont positives, il se peut (même si c'est très improbable) que, sur une longueur finie, une chaîne Mm n'utilise pas un des $(m+1)$-mots. A l'opposé, le modèle de permutation fixe de façon certaine le nombre de passages par chacun d'entre eux. Ainsi le modèle Mm rejoint le modèle de permutation sur l'espérance du nombre d'occurrences de $(m+1)$-mots, mais autorise une variabilité interdite par le modèle de permutation.

Cette différence est à l'origine de l'écart observé au chapitre 6 entre les approximations de type « KB » et « PRT » de la loi du comptage N. L'approche KB (Kleffe et Borodovsky 1992) se situe dans le cadre des chaînes de Markov, alors que l'approche PRT (Prum *et al.* 1995), présentée dans le même cadre, est conditionnelle aux comptages des $(m+1)$-mots et se situe donc implicitement dans un modèle de permutation. La variance PRT des comptages est d'ailleurs inférieure à la variance KB.

2.3 Occurrences de motifs et chaînes de Markov

L'étude de la distribution du nombre $N(\mathbf{w})$ d'occurrences d'un mot $\mathbf{w}$ dans une chaîne de Markov a constitué le sujet pionnier des travaux du groupe *Statistiques des Séquences Biologiques*. L'article de Prum *et al.* (1995) en propose une approximation gaussienne, la thèse de Schbath (1995b) reprend ces travaux et propose une approximation par une loi de Poisson composée (Schbath 1995a) et le logiciel *R'MES* (Bouvier *et al.* 1999) met ces différentes méthodes à disposition des biologistes. Plus récemment, la thèse de Nuel (2001) propose une approche par grandes déviations ainsi qu'un logiciel opérationnel disponible à l'adresse suivante

`http : //stat.genopole.cnrs.fr/GDon`.

À la suite de ces travaux, J.-J. Daudin et moi avons travaillé sur la répartition des occurrences d'un mot (et plus seulement sur leur comptage) dans une chaîne de Markov en nous concentrant sur l'obtention de résultats exacts. Les deux publications principales sur ce sujet sont reproduites aux chapitres 4 et 5. La première traite du cas d'un seul mot alors que la seconde étudie un ensemble de mots en se concentrant sur les distances entre occurrences.

2.3.1 Occurrences d'un mot

On s'intéresse aux occurrences d'un mot $\mathbf{w} = (w_1 \dots w_k)$ dans une séquence aléatoire de lettres $\mathbf{S} = (S_1 S_2 \dots)$ générée selon une chaîne de Markov d'ordre 1 de matrice de transition $\mathbf{\Pi}$ sur l'alphabet $\mathcal{A}$.

Positions et distances entre occurrences

L'étude de la répartition des occurrences de $\mathbf{w}$ repose sur l'étude

- des *positions* $(X_1, X_2, \dots)$ de ses occurrences successives

• et des *distances* Y qui les séparent.
La figure 2.1 illustre ces définitions. Par convention, la position d'une occurrence est définie par la position de la dernière lettre dans la séquence. Cette convention est commode, mais arbitraire et absolument pas générale.

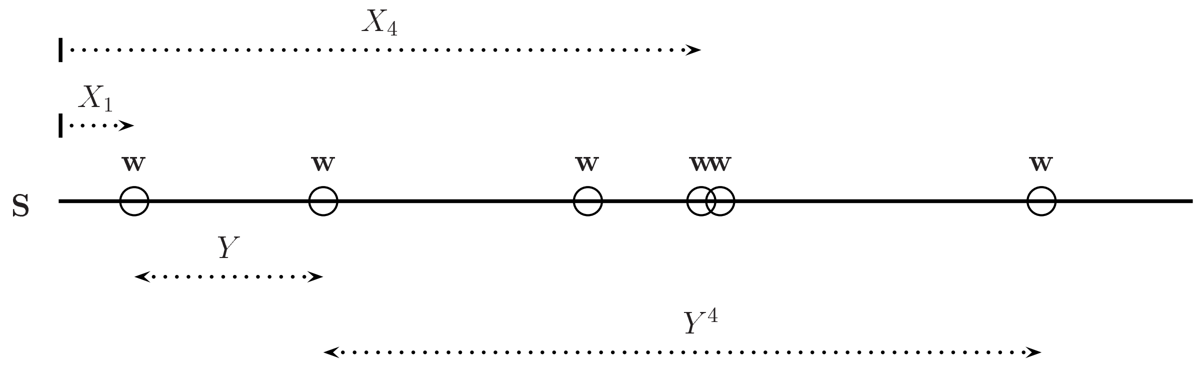


FIG. 2.1 – *Occurrences d'un mot* $\mathbf{w}$ *:* X_1 *et* X_4 *sont les positions de ses première et quatrième occurrences ;* Y *est la distance entre deux occurrences successives et* Y^4 *la distance cumulée d'ordre 4.*

Notre objectif est de déterminer les *distributions exactes* des variables aléatoires (v.a.) X, X_n, Y et Y^r. L'intérêt des distances cumulées apparaîtra dans le chapitre 3.

Méthode d'obtention de la distribution

La distribution de la position $X = X_1$ de la première occurrence est obtenue à partir de sa *fonction génératrice* (de probabilité). En notant $p(x) = \Pr\{X = x\}$, la fonction génératrice ϕ_X de X est définie par

$$\phi_X(t) = \sum_{x \geq 1} p(x) t^x.$$

Cette fonction génératrice est obtenue par un raisonnement en deux étapes.
(i) On établit une récurrence sur les probabilités $p(x)$:

$$p(x) = f[p(1), \ldots, p(x-1)]$$

(théorème 1, paragraphe 4.2.1).
(ii) On déduit la fonction génératrice $\phi_X(t)$ en sommant cette récurrence sur $x \geq 1$ et en multipliant par t^x (théorème 2, paragraphe 4.2.2).
On obtient la fonction génératrice ϕ_Y de la distance Y selon le même principe. Cette distance a un sens dans le modèle M1 car, dans ce modèle, les distances séparant les occurrences successives sont indépendantes et identiquement distribuées (i.i.d.).

Les fonctions génératrices des positions ultérieures X_n et des distances cumulées Y^r s'obtiennent ensuite directement grâce à l'indépendance des distances :

$$\phi_{X_n}(t) = \phi_X(t)[\phi_Y(t)]^{n-1}, \qquad \phi_{Y^r}(t) = [\phi_Y(t)]^r. \tag{2.2}$$

Mise en oeuvre

Le calcul des probabilités $p(x)$ ne requiert pas la connaissance de la fonction génératrice ϕ_X : la récurrence obtenue à l'étape (i) du raisonnement permet de calculer tous les $p(x)$ (pour x compris entre 1 et ℓ). Cependant, l'ordre de cette récurrence croît linéairement avec x et le temps de calcul des $p(x)$ est donc quadratique en ℓ, ce qui peut être rédhibitoire pour les longues séquences.

On peut également obtenir les probabilités $p(x)$ en opérant un développement en série de la fonction ϕ_X (c'est d'ailleurs de cette propriété que les fonctions *génératrices de probabilités* tirent leur nom). Or il apparaît que les fonctions ϕ_X et ϕ_Y sont des fractions rationnelles. Le degré de leur dénominateur est égal à la longueur k du mot. Les fonctions ϕ_{X_n} et ϕ_{Y^r} sont également des fractions rationnelles par voie de conséquence.

Ce résultat est important car les coefficients du développement en série d'une fraction rationnelle satisfont une relation de récurrence dont l'ordre est donné par le degré du polynôme situé au dénominateur. Cette propriété permet de calculer les probabilité $p(x)$ au moyen d'une récurrence d'ordre fixe k et d'obtenir ainsi des temps de calculs linéaires en ℓ. On peut ainsi ajouter une troisième étape au raisonnement :

(iii) On calcule les $p(x)$ par une récurrence (d'ordre k) en développant $\phi_X(t)$ en série.

(De ce fait, les astuces de calcul proposées en appendice du chapitre 4 perdent un peu de leur intérêt.)

Cependant, le calcul de la fonction génératrice nécessite une diagonalisation de matrice et donc, pour des x petits, la récurrence établie au point (i) peut être préférable.

Caractéristique de la distribution des distances

La connaissance des fonctions génératrices ϕ_X et ϕ_Y permet de calculer toute la distribution des variables X et Y (voir paragraphe suivant) donc, en particulier, leurs moments. On dégage ainsi deux grandes caractéristiques concernant les distributions des distances entre occurrences.

1. L'espérance de la distance entre occurrences dépend seulement de la fréquence du mot :
$$\mathbb{E}(Y) = 1/\mu(\mathbf{w}).$$
La fréquence $\mu(\mathbf{w})$ dépend de la longueur du mot et de sa composition en `a`, `c`, `g` et `t` (cf équation (2.1)).
2. La variance de la distance dépend de la fréquence du mot et de sa *structure de recouvrement*. La formule explicite de $\mathbb{V}(Y)$ n'est pas très informative et requiert des notations définies au chapitre 4.

La figure 2.2 reprend la figure 4.1 du chapitre 4 qui est obtenue sous M00 mais rend compte de phénomènes existant sous M1. Dans M00, les deux mots (`gatc`) et (`aaaa`) ont la même fréquence $1/\mu(\mathbf{w}) = (1/4)^4 = 1/256$. Les espérances des distances sont donc égales :
$$\mathbb{E}[Y(\texttt{gatc})] = \mathbb{E}[Y(\texttt{aaaa})] = 256,$$
mais les variances sont très différentes :
$$\mathbb{V}[Y(\texttt{gatc})] = (256.2)^2 \quad < \quad \mathbb{V}[Y(\texttt{aaaa})] = (326.7)^2.$$

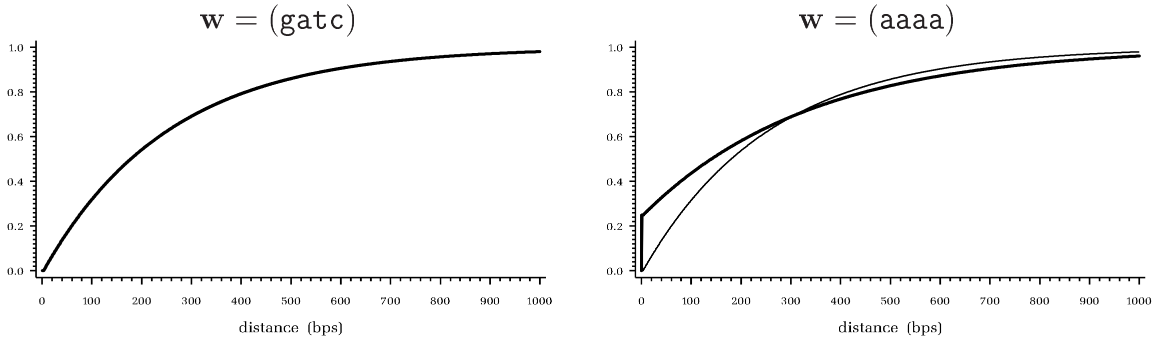


FIG. 2.2 – *Fonction de répartition de la distance Y dans le modèle M00 pour les mots* $\mathbf{w} = (\texttt{gatc})$ *et* $\mathbf{w} = (\texttt{aaaa})$. *Comparaison avec la loi géométrique de paramètre* $1/\mu(\mathbf{w})$ *(trait fin, masqué pour* $\mathbf{w} = (\texttt{gatc})$*).*

La différence de variance manifeste une plus grande régularité dans la répartition des occurrences de (gatc).

Structure de recouvrement d'un mot

L'importance de la structure de recouvrement a été pour la première fois décrite par Guibas et Odlyzko (1981). L'irrégularité de la répartition de (aaaa) vient de sa capacité d'*auto-recouvrement*: ce mot peut se reconstituer à partir des trois dernières lettres d'une de ses occurrences au prix d'un seul (a) supplémentaire. On dit que (aaaa) est fortement recouvrant alors que (gatc) est « non-recouvrant ». On notera dans la suite $a(\mathbf{w})$ la probabilité de cet événement appelée *probabilité de recouvrement.* $a(\texttt{aaaa})$ vaut 1/4 dans le modèle M00 et $\pi(\texttt{a};\texttt{a})$ dans M1.

Le cas du mot (aaaa) est évidemment caricatural, mais de nombreux mots possèdent cette propriété d'auto-recouvrement. Ainsi une des versions du CHI de *H. influenzae* (ggtggtgg) peut se recouvrir sur 1, 2 et 5 lettres comme le montre la figure 2.3.

1 lettre :								*g*	g	t	g	g	t	g	g
w	g	g	t	g	g	t	g	g							
5 lettres :				*g*	*g*	*t*	*g*	*g*	t	g	g				
2 lettres :							*g*	*g*	t	g	g	t	g	g	

FIG. 2.3 – *Recouvrements du mot* ggtggtgg.

La probabilité $a(\mathbf{w})$ est définie par

$$a(\mathbf{w}) = \Pr\{Y(\mathbf{w}) < k\} \tag{2.3}$$

et peut se calculer à partir de ϕ_Y.

Schbath (1995b) en propose une autre formulation à partir des *périodes* du mot $\mathbf{w}$. La notion de période se comprend sur l'exemple de la figure 2.3 en y observant qu'un recouvrement sur 5 lettres (qui laisse 3 lettres non recouvertes) aura toujours lieu avant un recouvrement sur 2 lettres (qui laissent donc $6 = 2 \times 3$ lettres non recouvertes). Les recouvrements sur 2 lettres sont issus de la même période que ceux sur 5 lettres et ne seront jamais comptabilisés en tant que tels mais comme des successions de deux recouvrements sur 5 lettres. Dans cet exemple, la probabilité de recouvrement vaut donc

$$\begin{aligned} a &= \pi(\mathtt{g};\mathtt{g})\pi(\mathtt{g};\mathtt{t})\pi(\mathtt{t};\mathtt{g})\pi(\mathtt{g};\mathtt{g})\pi(\mathtt{g};\mathtt{t})\pi(\mathtt{t};\mathtt{g})\pi(\mathtt{g};\mathtt{g}) && \text{(pour 1 lettre)} \\ &\quad +\pi(\mathtt{g};\mathtt{t})\pi(\mathtt{t};\mathtt{g})\pi(\mathtt{g};\mathtt{g}) && \text{(pour 5 lettres).} \end{aligned}$$

Cette probabilité joue un rôle central dans les modèles de Poisson composés.

2.3.2 Occurrences de plusieurs mots

Les résultats obtenus sur les occurrences d'un seul mot $\mathbf{w}$ peuvent se généraliser aux occurrences d'un motif (famille de mots)

$$\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots\}.$$

On se concentre sur les distances entre occurrences qui sont les plus utiles dans les applications biologiques, notamment pour l'étude leur répartition.

Distances entre occurrences

On note Y_{ij} la distance séparant une occurrence du mot $\mathbf{w}_i$ de l'occurrence suivante du mot $\mathbf{w}_j$ *à condition qu'aucun autre mot de $\mathcal{W}$ n'apparaisse entre temps*; si un mot de $\mathcal{W}$ autre que $\mathbf{w}_j$ (mais y compris $\mathbf{w}_i$) apparaît avant $\mathbf{w}_j$, Y_{ij} est infinie. La figure 2.4 illustre ces notations pour une famille de trois mots. On cherche maintenant à déterminer les distributions de variables Y_{ij}.

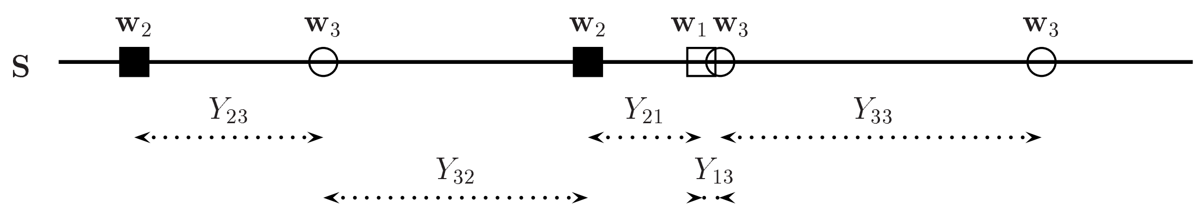


FIG. 2.4 – *Distances séparant les occurrences de trois mots $\mathbf{w}_1$ (□), $\mathbf{w}_2$ (■) et $\mathbf{w}_3$ (○).*

Méthode d'obtention de la distribution

La distribution des Y_{ij} est obtenue selon le même principe que celles de X et Y dans le cas d'un seul mot. La différence principale réside dans le fait que toutes les distributions s'obtiennent simultanément. On note $q_{ij}(y) = \Pr\{Y_{ij} = y\}$ et $\phi_{ij}(t)$ la fonction génératrice de Y_{ij} : $\phi_{ij}(t) = \sum_{y \geq 1} q_{ij}(y) t^y$. On rappelle que, comme la distance Y_{ij} peut être infinie, on a $\sum_y q_{ij}(y) < 1$.

On obtient la distribution des Y_{ij} en trois étapes.

(i) On établit une récurrence sur les $q_{ij}(y)$:

$$q_{ij}(y) = f\left[q_{kl}(z) : 1 \leq z \leq y-1, 1 \leq k,l \leq I\right]$$

en notant I le nombre de mots dans la famille $\mathcal{W}$ (lemmes 1 et 2, paragraphe 5.2.2).

(ii) On en déduit que la matrice $\mathbf{\Phi}(t)$ de terme général $\phi_{ij}(t)$ vérifie

$$\mathbf{F}(t) = \mathbf{\Phi}(t)[\mathbf{I} + \mathbf{F}(t)]$$

où $\mathbf{I}$ est la matrice identité et $\mathbf{F}(t)$ une autre matrice dont les éléments sont des fonctions génératrices calculables directement (lemme 3 et théorème, paragraphe 5.2.2).

(iii) Si $[\mathbf{I} + \mathbf{F}(t)]$ est inversible, on obtient les fonctions génératrices $\phi_{ij}(t)$ en calculant

$$\mathbf{\Phi}(t) = \mathbf{F}(t)[\mathbf{I} + \mathbf{F}(t)]^{-1}.$$

Mise en oeuvre

L'étape (iii) est évidemment la plus délicate. La condition d'inversibilité de $[\mathbf{I} + \mathbf{F}(t)]$ n'est pas un obstacle. La difficulté réside dans l'inversion effective de cette matrice dont chaque élément est une fraction rationnelle.

Cette inversion pose des problèmes numériques importants et qui doivent être résolus avec soin. En effet, les coefficients de chaque fraction rationnelle $\phi_{ij}(t)$ interviennent dans son développement en série qui donne les probabilités $q_{ij}(y)$. Ils doivent donc être déterminés avec une grande précision.

Pour l'instant, on inverse formellement la matrice $[\mathbf{I} + \mathbf{F}(t)]$ au moyen d'un logiciel de calcul symbolique (de type *Maple*). Cette solution limite cependant beaucoup l'application de ces résultats. Ainsi on a pu faire ce calcul pour le motif CHI de *H. influenzae* (4 mots de 8 lettres) mais pas pour les 64 palindromes de 6 lettres, ni même pour les 16 palindromes de 4 lettres.

Caractéristiques de distances

Concernant la distribution des distances Y_{ij}, on fait les mêmes observations que pour un seul mot. La distribution de la distance Y_{ij} dépend principalement

- de la fréquence $\mu(\mathbf{w}_j)$ du mot $\mathbf{w}_j$
- et de la probabilité $a(i,j)$ de recouvrement du mot $\mathbf{w}_j$ sur le mot $\mathbf{w}_i$.

Là encore, ces caractéristiques des distances entre occurrences fonderont les modèles de Poisson composés.

2.4 Application à la recherche de sites promoteurs

Ces résultats sur la répartition de motifs dans le modèle M1 ont trouvé une application dans la recherche de sites promoteurs qui a abouti à l'article (soumis) reproduit au chapitre 7.

2.4.1 Motifs promoteurs

Les connaissances biologiques, portant sur la structure de la polymérase (cf paragraphe 2.1.2), induisent à penser que les motifs promoteurs sont des *motifs structurés* composés de deux mots $\mathbf{v}$ et $\mathbf{w}$ séparés par une distance (légèrement) variable. Un tel motif, noté $\mathbf{m}$, est défini par la notation

$$\mathbf{m} = \mathbf{v}(d_1 : d_2)\mathbf{w} \tag{2.4}$$

où d_1 et d_2 sont les valeurs extrêmes admises pour la distance séparant $\mathbf{v}$ et $\mathbf{w}$.

Définition de régions promotrices

L'objectif est de déterminer, pour une espèce donnée, le motifs structurés qui sont effectivement des promoteurs. On commence donc par rechercher des motifs structurés de la forme (2.4) dans un grand nombre de régions promotrices situées en amont de gènes connus. L'ensemble d'apprentissage de cette méthode est donc constitué d'un ensemble de régions (quelques centaines) de longueurs variables (autour de 100 bps) chacune susceptible de contenir (au moins) un site promoteur.

Reconnaissance de motifs candidats

Parallèlement, on se donne les longueurs des deux mots $\mathbf{v}$ et $\mathbf{w}$ et les distances d_1 et d_2. La détection efficace (en temps de calcul) d'un grand nombre de motifs structurés dans un grand nombre de régions est en soit un problème informatique. Il est résolu par un algorithme de *pattern matching* (cf. paragraphe 7.2.2). La rapidité de cette étape est d'autant plus cruciale que de tels motifs sont fréquemment connus sous forme de *consensus*, c'est à dire avec une certaine imprécision. Ainsi, dans l'exemple traité au chapitre 7, on recherche des motifs de la forme $\mathbf{m} = \mathbf{v}(d_1 : d_2)\mathbf{w}$ mais en autorisant *une erreur au plus* sur l'ensemble des deux mots $\mathbf{v}$ et $\mathbf{w}$.

2.4.2 Analyse statistique

Ce premier crible permet de dégager des motifs structurés susceptibles d'être des promoteurs. Ces candidats sont sélectionnés en fonction de leurs fréquences d'apparition : les vrais motifs doivent être présents dans un grand nombre de régions promotrices différentes. La question statistique est alors la suivante, parmi les motifs candidats quels sont ceux qui sont présents *significativement souvent* dans les régions promotrices. En termes d'occurrences de motifs, cette question est liée à la position $X(\mathbf{m})$ de la première partie du motif $\mathbf{m}$ occurrence puisque la probabilité d'apparition dans une séquence de longueur ℓ vaut $\Pr\{X(\mathbf{m}) \leq \ell\}$.

Probabilités d'apparition en un point

On a présenté au paragraphe 2.3.2 des résultats sur les lois des distances séparant les occurrences. Ces résultats s'adaptent assez facilement au calcul de la probabilité d'occurrence $\mu(\mathbf{m})$ d'un motif $\mathbf{m}$ en un point quelconque. En supposant que la séquence est une chaîne de Markov stationnaire,

1. on calcule $\mu(\mathbf{v})$,
2. on calcule la probabilité que $\mathbf{w}$ apparaisse entre d_1 et d_2 bps plus loin,
3. $\mu(\mathbf{m})$ est le produit de ces deux probabilités.

Dans ce cas (les distances d_1 et d_2 étant typiquement inférieures à 20 ou 25 bps), le calcul de la loi de la distance peut se faire au moyen de la récurrence établie à l'étape (i) de la méthode présentée au paragraphe 2.3.2. On évite ainsi l'étape (iii) qui serait impraticable puisque, pour des mots $\mathbf{v}$ et $\mathbf{w}$ de longueur 6 en autorisant une erreur, il faudrait traiter deux familles $\mathcal{V}$ et $\mathcal{W}$ contenant chacune 19 mots.

Probabilités d'apparition dans une séquence

La connaissance de la probabilité d'apparition $\mu(\mathbf{m})$ en un point quelconque ne suffit cependant pas à déterminer la loi de la position $X(\mathbf{m})$ de la première occurrence. On a vu dans les paragraphes précédents que cette loi dépendait aussi fortement de la probabilité d'autorecouvrement du motif.

Or il apparaît très vite que la structure de recouvrement du motif $\mathbf{m}$ est impossible à appréhender à cause de l'intervalle $(d_1 : d_2)$. En effet, cet intervalle vierge autorise toutes les combinaisons possibles et il est hors de question (en termes combinatoires) d'explorer toutes les façons dont une séquence d'une vingtaine de lettres –quelconques– peut se recouvrir elle-même. Ainsi, il n'est pas possible (au moyen des résultats présentés ici du moins) de déterminer exactement la probabilité qu'un motif structuré apparaisse dans une séquence de longueur donnée.

Approximation d'ordre k

L'idée la plus simple pour approcher la probabilité $\Pr\{X(\mathbf{m}) \leq \ell\}$ est d'utiliser une approximation géométrique de paramètre $\mu(\mathbf{m})$:

$$\Pr\{X(\mathbf{m}) > \ell\} = [1 - \mu(\mathbf{m})]^{\ell - |\mathbf{m}|+1} \simeq \exp[-\mu(\mathbf{m})(\ell - |\mathbf{m}| + 1)]$$

en notant $|\mathbf{m}|$ la longueur du motif $\mathbf{m}$. Mais la confrontation avec des simulations montre que cette approximation –qui néglige complètement les recouvrements du motif– donne de très mauvais résultats, précisément pour les motifs $\mathbf{m}$ pour lesquels le premier mot $\mathbf{v}$ est très recouvrant.

On améliore beaucoup l'approximation en considérant la probabilité d'occurrence du motif en une position x conditionnellement à sa non-occurrence en $x-1, x-2, \ldots, x-k$. On obtient ainsi une approximation *d'ordre* k de $\Pr\{X(\mathbf{m}) \leq \ell\}$ (cf paragraphe 7.3.1). En pratique, des simulations montrent que l'approximation d'ordre 1 fournit d'excellents résultats (voir figure 7.1). Cette méthode permet de détecter une série de motifs en accord avec le promoteur consensus connu chez *B. subtilis* (voir figure 7.2).

2.5 Des chaînes de Markov aux processus de Poisson composés

2.5.1 Loi du nombre d'occurrences

Résultats exacts

Les résultats sur les positions des occurrences donnent accès à la loi du comptage du nombre d'occurrences. En effet, la probabilité d'observer n occurrences ou plus est égale à la probabilité que la n-ème occurrence survienne avant la fin de la séquence. En notant $N_\ell(\mathbf{w})$ le nombre d'occurrences du mot $\mathbf{w}$ dans une séquence de longueur ℓ, on a donc

$$\Pr\{N_\ell(\mathbf{w}) \geq n\} = \Pr\{X_n(\mathbf{w}) \leq \ell\}.$$

On connaît donc, sans effort supplémentaire, la loi exacte du comptage dans une chaîne de Markov. Cette approche n'est cependant pas la seule à fournir cette loi exacte (voir Régnier (2000)).

Résultats asymptotiques

Les approximations de cette loi sont également assez nombreuses (Kleffe et Borodovsky 1992, Prum *et al.* 1995, Schbath 1995a) et la connaissance de la loi exacte permet d'en comparer les précisions. L'article reproduit au chapitre 6 présente ces comparaisons à partir de simulations et propose des règles pour choisir telle ou telle approximation en fonction de la longueur de la séquence, de la fréquence du motif et du degré de précision souhaitée (voir tableau 6.1). Les principales conclusions ressortant de cette étude sont les suivantes.

- L'approximation de type Poisson composée est la meilleure dans la majorité des cas (sauf pour des motifs extrêmement fréquents). Les bornes de Chen-Stein sont très pessimistes par rapport à la précision effectivement observée.
- Les distribution du comptage sous le modèle de chaîne de Markov et le modèle de permutation (cf paragraphe 2.2.3) sont très différentes.
- L'imprécision due à l'estimation des transitions de la chaîne de Markov induit de plus grandes erreurs que l'usage d'une loi approchée.

Estimation des probabilités de transitions

Aussi inquiétant soit-il, le dernier point du paragraphe précédent mérite d'être discuté. En effet, parler d'estimation des transitions suppose que la séquence observée est en fait gouvernée par une chaîne de Markov dont les « vrais » paramètres sont inconnus. D'un point de vue biologique, cela revient à dire qu'il existe une « vraie matrice » de transition du génome de *E. coli* (ou de *B. subtilis*).

La variabilité entre les génomes d'individus d'une même espèce existe assurément, mais elle n'a probablement pas grand chose à voir avec la variabilité observée entre les trajectoires d'une chaîne de Markov. Ainsi il est sans doute inutile de s'inquiéter de la valeur des « vraies » transitions, tout simplement parce qu'elle n'existent pas. Il semble plus sage de se contenter des transitions estimées sur la séquence observée, sans leur accorder une valeur universelle, mais en considérant qu'elles permettent de prendre en compte *a priori* un certain type d'information.

2.5.2 Loi de Poisson composée

Comme le montre l'article reproduit au chapitre 6, la loi de Poisson composée fournit une très bonne approximation de la loi du comptage $N(\mathbf{w})$ dans une chaîne de Markov. Il apparaît, de plus, qu'elle est fondée sur moins de paramètres que les chaînes de Markov. Il est donc tentant d'utiliser des modèles de Poisson pour décrire les occurrences d'un ou plusieurs mots.

Définition

La loi de Poisson composée peut être définie de deux façons :

1. Soient $(Z_1, Z_2, \dots)$ des variables de Poisson indépendantes de paramètres respectifs $(\lambda_1, \lambda_2, \dots)$:
$$Z_k \sim \mathcal{P}(\lambda_k),$$
la somme
$$N = \sum_{k \geq 1} Z_k$$
suit une loi de Poisson composée.
2. Soient C une variable de Poisson de paramètre λ
$$C \sim \mathcal{P}(\lambda)$$
et $\{K_1, K_2, \dots\}$ une suite de v.a. entières i.i.d. de loi g :
$$\Pr\{K_c = k\} = g(k) \qquad \text{pour } c \geq 1 \text{ et } k \geq 0,$$
la somme
$$N = \sum_{c=1}^{C} K_c$$
suit une loi de Poisson composée.

On peut montrer que ces deux définitions sont équivalentes en prenant
$$\lambda_k = \lambda g(k) \qquad \text{pour } k \geq 1.$$
La notation de cette loi n'est pas complètement fixée. Suivant la première définition, on trouve parfois $\mathcal{PC}(\lambda_1, \lambda_2, \dots)$. On utilise ici une notation plus cohérente avec la seconde définition :
$$N \sim \mathcal{PC}(\lambda; g).$$
Selon cette définition, la loi de Poisson composée est la loi du comptage d'un événement (l'occurrence du motif) survenant par paquets ou par « trains » (les suites d'occurrences se recouvrant), ces trains étant en nombre poissonnien.

Loi de Pólya-Aeppli

La loi de Pólya-Aeppli correspond au cas particulier où (selon la première définition) les variables $\{K_1, K_2, \dots\}$ suivent une loi géométrique $\mathcal{G}(1-a)$:

$$g(k) = a^{k-1}(1-a).$$

Cette loi est gouvernée par deux paramètres (λ et a) et est notée $\mathcal{PA}(\lambda; a)$. Pour une variable N suivant cette loi, on a

$$\begin{aligned} \Pr\{N=n\} &= e^{-\lambda} a^n \sum_{c=1..n} \frac{1}{c!} \binom{n-1}{c-1} \left[\frac{\lambda(1-a)}{a}\right]^c && \text{pour } n > 0, \\ &= e^{-\lambda} && \text{pour } n = 0 \end{aligned}$$

et

$$\mathbb{E}(N) = \lambda/(1-a), \qquad \mathbb{V}(N) = \lambda(1+a)/(1-a)^2.$$

Pour un mot non recouvrant, le paramètre a est nul et on retrouve une loi de Poisson simple.

Approximation de la loi du comptage

Dans le cadre d'un modèle Mm, Schbath (1995a) montre que la loi du comptage $N(\mathbf{w})$ peut être approchée par une loi de Poisson composée qui est en fait une loi $\mathcal{PA}(\lambda; a)$ dont le paramètre a est la probabilité de recouvrement du mot $a(\mathbf{w})$ et dont le paramètre $\lambda(\mathbf{w})$ vaut

$$\lambda(\mathbf{w}) = \mu(\mathbf{w})[1 - a(\mathbf{w})].$$

Dans le cas du modèle M1, les paramètres $\mu(\mathbf{w})$ et $a(\mathbf{w})$ sont exactement ceux donnés aux équations (2.1) et (2.3) de ce chapitre.

L'interprétation du paramètre $\lambda(\mathbf{w})$ est la suivante : $\lambda(\mathbf{w})$ représente la fréquence des trains et chaque train commence par une occurrence de $\mathbf{w}$ (de fréquence $\mu(\mathbf{w})$) *non recouverte à gauche* (d'où le terme $[1 - a(\mathbf{w})]$).

Autre distribution pour la taille des trains

On vient de voir qu'une distribution géométrique de la taille K des trains permet d'obtenir une formulation explicite de la loi du comptage N. Si la loi g n'est pas géométrique, on peut néanmoins calculer la probabilité $\Pr\{N = n\}$ au moyen d'un algorithme en $O(n^2)$ proposé par Barbour *et al.* (1992) (cf proposition 1, chapitre 8). La loi géométrique n'est donc pas une hypothèse nécessaire pour calculer efficacement la loi du comptage. On verra plus loin qu'elle induit des propriétés importantes sur les distances entre occurrences.

2.5.3 Processus de Poisson composé pour un mot

Cette approximation par une loi de Poisson composée suggère une autre modélisation du processus des occurrences d'un mot ou d'une famille de mots. Cette modélisation semble particulièrement efficace pour étudier l'homogénéité de la répartition des occurrences d'un mot ou d'un motif. C'est l'objet de l'article reproduit au chapitre 8. La principale différence avec la modélisation par chaîne de Markov est que la séquence est vue comme une *ligne continue* et non plus comme un enchaînement discret de lettres.

Modèle pour un seul mot

Le modèle de Poisson composé (PC) est fondé

- d'une part sur le processus ponctuel $\{C(x)\}_{x\geq 0}$ des occurrences des trains,
- d'autre part sur la loi commune g des tailles $(K_1, K_2, \dots)$ de ces trains.

Le processus $\{C(x)\}$ est supposé poissonnien (d'intensité λ), ce qui constitue une *hypothèse nulle* dans l'étude de la régularité de la répartition des occurrences. Les tailles des trains suivent une loi géométrique décrivant une succession de recouvrements de probabilité a. On définit ainsi la loi du processus de comptage $\{N(x)\}_{x\geq 0}$:

$$N(x) = \sum_{c=1}^{C(x)} K_c \sim \mathcal{PA}(\lambda x; a).$$

En se souvenant que le paramètre λ est égal au produit $(1-a)\mu$, et en reprenant l'espérance de la loi de Pólya-Aeppli, on retrouve $\mathbb{E}[N(\ell)] = \mu\ell$: l'espérance du comptage est égale à la fréquence du mot multipliée par la longueur de la séquence.

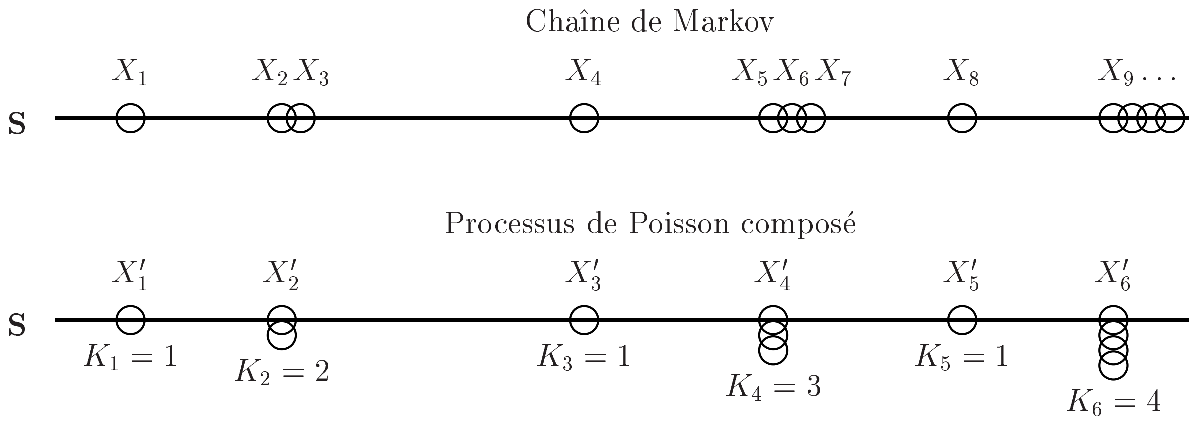


FIG. 2.5 – *Représentations des occurrences d'un mot selon une chaîne de Markov et un processus de Poisson composé.*

Dans cette partie, on note $(X'_1, X'_2, \dots)$ les positions de trains pour les distinguer des positions $(X_1, X_2, \dots)$ des occurrences. Dans le modèle PC, les occurrences qui se recouvrent sont supposées avoir la même position. Ainsi, dans la seconde partie de la figure 2.5, les positions X_2 et X_3 sont égales (à X'_2), de même que X_5, X_6 et X_7 ($= X'_4$) et X_0, X_{10}, X_{11} et X_{12} ($= X'_6$).

On retrouve ici la limite de l'approximation poissonnienne (réservée aux mots rares) : pour que cette modélisation soit acceptable, il faut que la longueur du mot ou du motif soit négligeable par rapport aux distances qui séparent les trains.

Estimations de paramètres

On a vu dans les paragraphes précédents que le modèle Mm pouvait fournir des estimations des paramètres λ et a à partir des probabilités de transitions. Dans une séquence de longueur ℓ, ayant observé le nombre $C(\ell)$ de trains et le nombre total $N(\ell)$ d'occurrences, on peut calculer les estimateurs du maximum de vraisemblance (EMV) de ces paramètres :

$$\hat{\lambda} = \frac{C(\ell)}{\ell}, \qquad \hat{a} = \frac{N(\ell) - C(\ell)}{N(\ell)}$$

qui s'interprètent respectivement comme la fréquence des trains le long de la séquence et la proportion d'occurrences recouvertes parmi les $N(\ell)$ observées.

Il est important de noter que les EMV de λ et a prennent en compte la fréquence observée du motif dans la séquence et sont donc, par nature, ajustés à cette fréquence. Ce n'est, par exemple, pas le cas des estimateurs donnés par le modèle M1 pour des mots d'une longueur supérieure à 2. Comme on l'a vu à la section 2.2, le modèle de Markov ajusté à la fréquence d'un mot de longueur k est le modèle M$(k-1)$. Or ce modèle requiert l'estimation de 4^k paramètres (en fait $3 \times 4^{k-1}$ paramètres indépendants), ce qui est beaucoup si on s'intéresse aux occurrences d'un seul mot. Pour l'étude de la répartition d'un seul mot, le modèle PC (à 2 paramètres) est donc nettement plus parcimonieux que le modèle M$(k-1)$.

Distance entre occurrences

Comme on l'a dit plus haut, le modèle PC sera principalement utilisé pour l'analyse de la répartition des occurrences le long de la séquence. Ces analyses sont, entre autres, fondées sur l'étude des distances entre occurrences.

Dans le modèle PC pour un mot, une distance entre deux occurrences successives peut être

- soit nulle si les occurrences se recouvrent (probabilité a),
- soit distribuée selon une loi exponentielle λ si elles appartiennent à deux trains successifs.

Ainsi la distribution de la distance Y est un mélange d'une masse a en zéro ($\delta\{0\}$) et d'une loi exponentielle de paramètre λ :

$$Y \sim a\delta\{0\} + (1-a)\mathcal{E}(\lambda)$$

et sa fonction de répartition vaut

$$\Pr\{Y \leq y\} = 1 - (1-a)e^{-\lambda y} \qquad \text{pour } y \geq 0. \tag{2.5}$$

De plus ces distances sont indépendantes (la probabilité de recouvrement ne dépend pas des recouvrement passés) et donc, la distribution des distances cumulées d'ordre r s'obtient par convolution r fois de la fonction de répartition de Y :

$$\Pr\{Y^r \leq y\} = \sum_{j=0..r} \binom{r}{j} a^j (1-a)^{r-j} G(y; \lambda, r-j) \tag{2.6}$$

en notant $G(y; \lambda, r)$ la fonction de répartition de la loi $\gamma(\lambda, r)$ (cf section 8.3).

2.6 Modèle de Poisson composé pour plusieurs mots

Le modèle PC peut être également utilisé pour modéliser les occurrences d'une famille de mots $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots\}$. Nous allons voir que si, formellement, cette généralisation ne pose pas de problème, sa mise en oeuvre en pose quelques uns.

Modèle pour plusieurs mots

Comme le modèle PC pour un mot, le modèle PC pour plusieurs mots suppose que les occurrences des mots de la famille surviennent par trains dont les occurrences sont distribuées selon un processus de Poisson (d'intensité λ). La différence avec le modèle pour un seul mot vient essentiellement des probabilités de recouvrement. On rappelle qu'on note

$$a(i, j) = \Pr\{\mathbf{w}_j \text{ recouvre } \mathbf{w}_i\}.$$

La connaissance de ces probabilités fournit la loi de la taille des trains. En effet, la composition d'un train (*i.e.* la succession des mots qui le compose) est décrite par une chaîne de Markov sur les mots $\mathcal{W}$ (plus un état supplémentaire représentant la fin du train). Les probabilités de transition de cette chaîne sont les $a(i, j)$. Conditionnellement à son premier mot, la taille d'un train est donc distribuée comme un temps de séjour parmi les mots de $\mathcal{W}$ avant d'atteindre l'état « fin du train ».

Distribution des distances

De même que la taille des trains est connue conditionnellement au premier mot du train, la distance entre deux occurrences dépend du premier mot. Les distances séparant les occurrences ne sont donc pas identiquement distribuées. On montre qu'elles ne sont pas indépendantes non plus (un exemple illustrant ces deux caractéristiques est donné au paragraphe 3.2.4).

Ces deux remarques ont évidemment de nombreuses conséquences défavorables. L'une d'entre elles est qu'on ne peut pas donner la loi générale de la distance cumulée. On verra au chapitre suivant que, dans ce cas, on ne peut pas utiliser les distances cumulées pour étudier l'homogénéité de la répartition des occurrences (cf paragraphe 3.2.4).

Il existe cependant certaines familles de mots pour lesquelles les distances entre occurrences sont i.i.d. Ainsi, le CHI d'*H. influenzae* (cf paragraphe 2.1.2) est un motif constitué de 4 mots :

$$\mathcal{W} = \{\texttt{gatggtgg}, \texttt{gctggtgg}, \texttt{ggtggtgg}, \texttt{gttggtgg}\}$$

qui peuvent tous se recouvrir sur 1 lettre (g) et peuvent tous être recouverts par le troisième d'entre eux sur 2 (gg) et 5 lettres (ggtgg). Pour ce motif donc les probabilités de recouvrement $a(i, j)$ ne dépendent pas du premier mot $\mathbf{w}_i$ et les distances entre occurrences sont i.i.d. Cette heureuse exception sauve l'application présentée aux chapitres 5 et 8.

Estimation des paramètres

L'estimation des probabilités de recouvrement $a(i, j)$ ne pose pas de problème en théorie. Elles peuvent être calculées dans un modèle Mm ou estimées par maximum de vraisemblance. L'EMV de $a(i,j)$ fondé sur l'observation d'une série d'occurrences des mots de $\mathcal{W}$ vaut simplement

$$\hat{a}(i,j) = \frac{\text{nombre de recouvrements observés de } \mathbf{w}_j \text{ sur } \mathbf{w}_i}{\text{nombre d'occurrences observées de } \mathbf{w}_i}.$$

Cependant, pour de grandes familles de mots, le nombre de paramètres $a(i,j)$ constitue en lui même une limite. Si on s'intéresse, par exemple, aux palindromes de longueur 6 qui sont au nombre de 64, il y a potentiellement $64 \times 64 = 4096$ probabilités de recouvrement à estimer. En fait, ce nombre se réduit à peu près au quart en déduisant les probabilités nulles du fait de la structure de recouvrement des mots.

Le modèle PC pour plusieurs mots utilise donc beaucoup de paramètres, ce qui constitue une limite à son application. On peut contourner ce problème en retournant aux chaînes de Markov pour modéliser les recouvrements des mots. Les probabilités de recouvrement peuvent, par exemple, être estimées à partir des 12 paramètres indépendants d'un modèle M1. Cette solution fournit de bons résultats (y compris avec le modèle M0) dans l'exemple présenté à la section 8.5. On aboutit ainsi à une combinaison dans laquelle la modélisation globale est dévolue au processus de Poisson et la modélisation locale aux chaînes de Markov.

Bibliographie

Barbour, A. D., Chen, L. H. Y. et Loh, W.-L. (1992). Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Prob.* **20** 1843–1866.

Bouvier, A., Gélis, F. et Schbath, S. (1999). *R'MES: Recherche de Mots Exceptionnels dans les Séquences d'ADN – Version 2.* Guide de l'utilisateur. INRA, Biométrie, F78352 Jouy-en-Josas. (`http://www.inra.fr/bia/J/AB/genome/RMES/welcome.html`).

Guibas, L. J. et Odlyzko, A. M. (1981). Periods in strings. *J. Combinatorial Theory A.* **30** 19–42.

Kleffe, J. et Borodovsky, M. (1992). First and second moment of counts of words in random texts generated by Markov chains. *Comp. Applic. Biosci.* **8** 433–441.

Nuel, G. (2001). *Grandes déviations et chaînes de Markov pour l'étude des occurrences de mots dans les séquences biologiques.* PhD thesis, Université d'Évry Val d'Essonne.

Prum, B., Rodolphe, F. et de Turckheim, E. (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* **57** 205–220.

Régnier, M. (2000). A unified approach to word occurrence probabilities. *Discrete Applied Mathematics.* **104** 259–280.

Schbath, S. (1995a). Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics.* **1** 1–16. (`http://www.emath.fr/ps/`).

Schbath, S. (1995b). *Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN.* PhD thesis, Université René Descartes, Paris V.

Chapitre 3

Étude de l'homogénéité de la répartition

Sommaire

3.1 Introduction

La plupart des résultats présentés dans le chapitre précédent peuvent être utilisés pour détecter une hétérogénéité dans la répartition d'un motif. L'objectif de ces analyses est la détection de régions « significativement riches » ou « significativement pauvres » en occurrences d'un motif donné comme le motif CHI ou les palindromes (cf 2.1.2). Pour la clarté de l'exposé, on se limitera ici à la recherche de régions riches, la recherche de régions pauvres étant (à peu près) le problème symétrique.

3.1.1 Modélisation des séquences

Comme dans le chapitre précédent, il est nécessaire de recourir à une modélisation des séquences pour pouvoir déterminer ce qui est significatif et ce qui ne l'est pas. L'article reproduit au chapitre 5 utilise un modèle de chaîne de Markov M1 ; l'article et le projet d'article présentés dans la partie III se placent dans le cadre de processus de Poisson composés (PC).

Les deux modèles M1 et PC sont des modèles homogènes le long de la séquence. Cette homogénéité constitue en quelque sorte l'hypothèse nulle dans la détection de régions riches. On peut cependant remarquer que, dans le modèle PC, cette homogénéité porte intrinsèquement sur les occurrences du motif $\mathcal{W}$ lui-même (puisque les paramètres de ce modèle ne décrivent que ces occurrences-là). Dans le modèle M1 par contre, l'homogénéité porte sur l'ensemble du comportement de la séquence. La détection d'hétérogénéité dans les modèles de Markov prend donc l'homogénéité comme hypothèse nulle, mais spécifie l'alternative en ne s'intéressant qu'à des hétérogénéités affectant la répartition du motif $\mathcal{W}$.

3.1.2 Détection d'hétérogénéités

Deux méthodes pour détecter des hétérogénéités dans la répartition d'un motif sont présentées dans ce chapitre : la première est fondée sur les distances cumulées et la seconde sur une fenêtre mobile. La figure 3.1 les illustre.

Distances cumulées

L'utilisation des distances cumulées d'ordre r (ou « r-scans ») a été proposée par Dembo et Karlin (1992) et appliquée à des données génétiques par Karlin et Macken (1991). L'idée première est de fonder l'analyse de la densité des occurrences d'un motif sur les distances qui les séparent. Les distances « simples » Y séparant des occurrences immédiatement successives ne fournissent pas un outil très puissant. On gagne sensiblement à considérer des distances cumulées Y^r (les définitions seront précisées à la section 3.2).

Les occurrences du motif étant en nombre fini, les distances qui les séparent le sont également. Cette approche est donc intrinsèquement discrète, que la modélisation de la séquence elle-même le soit aussi (modèle Mk) ou non (modèle PC). On verra dans la suite que cette approche est la mieux contrôlée d'un point de vue théorique mais qu'elle repose sur des hypothèses pas toujours acceptables.

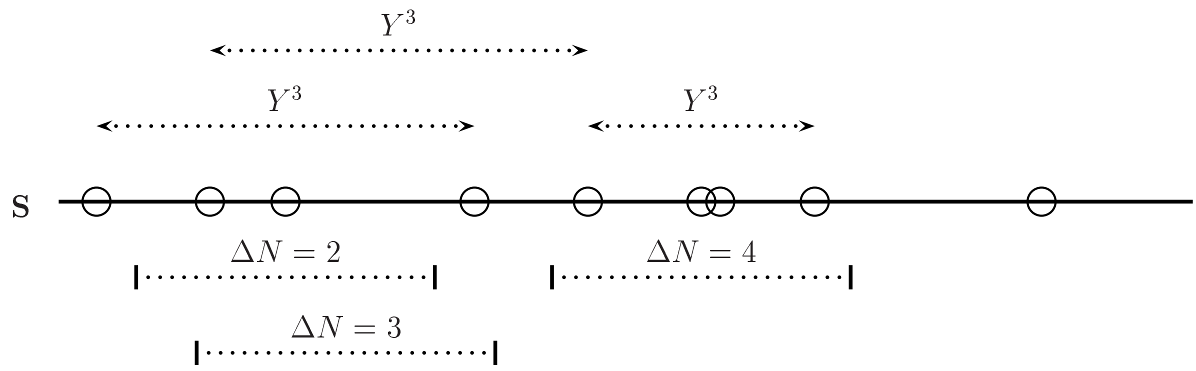


FIG. 3.1 – *Analyse de l'homogénéité de la répartition par les distances cumulées Y^r et par les comptages ΔN dans une fenêtre mobile de largeur fixe.*

Fenêtre mobile

La seconde approche est sans doute plus intuitive. Elle consiste à faire glisser un intervalle (ou *fenêtre*) de longueur fixe le long de la séquence et à compter le nombre ΔN d'occurrences contenues dans cette fenêtre. Cette approche n'est utilisée dans ce mémoire que dans le cadre du processus de Poisson composé, donc d'une modélisation continue de la séquence. On verra que cette méthode est moins exigeante, en termes d'hypothèses, que celle fondée sur les distances cumulées, mais aussi que son assise théorique est moins bonne.

Interprétations comparées

L'indicateur naturel de la fréquence d'un motif dans une région est le rapport

$$\frac{\text{nombre d'occurrences dans la région}}{\text{longueur de la région}}.$$

L'approche par distances cumulées propose de fixer le nombre d'occurrences (r) et d'étudier la variabilité de la longueur Y^r nécessaire pour couvrir ces r occurrences ; on considère alors le rapport r/Y^r. L'approche par fenêtre mobile au contraire fixe la longueur de la région (Δx) et étudie la variabilité du nombre d'occurrence ΔN ; on considère dans ce cas le rapport $\Delta N/\Delta x$. Chacune de ces deux méthodes fixe donc un des termes du rapport (le numérateur ΔN pour les distances cumulées et le dénominateur Y^r pour la fenêtre mobile) et laisse l'autre aléatoire.

3.2 Distances cumulées

3.2.1 Définitions et notations

On s'intéresse ici aux occurrences d'un motif $\mathcal{W}$. On note $\{X_1, X_2, \dots\}$ les positions de ses occurrences successives et $\{Y_1, Y_2, \dots\}$ les distances (simples) qui les séparent :

$$Y_i = X_{i+1} - X_i.$$

Les distances cumulées d'ordre r $\{Y_1^r, Y_2^r, \dots\}$ sont définies par

$$Y_i^r = \sum_{j=1}^{i+r-1} Y_j = X_{i+r} - X_i.$$

Si la séquence contient n occurrences, il y a $n-1$ distances et $n-r$ distances cumulées d'ordre r.

Hypothèse centrale

Dans cette approche, on fait l'hypothèse que *les distances* $\{Y_i\}_{i\geq 1}$ *sont i.i.d.* de fonction de répartition F :

$$F(y) = \Pr\{Y \leq y\}.$$

En fait, l'hypothèse d'équidistribution n'est pas absolument requise (cf paragraphe 3.2.2) mais l'indépendance l'est.

Loi des distances cumulées

Grâce aux hypothèses d'indépendance et d'équidistribution, la fonction de répartition F_r des distances cumulées $\{Y^r\}_{i\geq 1}$ s'obtient en convoluant r fois F. On peut ainsi calculer la probabilité pour qu'une distance cumulée soit inférieure à une longueur y :

$$\Pr\{Y^r \leq y\} = F_r(y)$$

et, réciproquement, pour un niveau α, déterminer la longueur y telle que

$$F_r(y) = \alpha.$$

Si F est connue (par exemple grâce aux résultats du chapitre 2), on peut déclarer qu'une distance cumulée est significativement courte, et donc que la région couverte par cette distance est significativement riche.

Loi des distances cumulées extrêmes

En pratique cependant, on calcule toutes les distances cumulées et on s'intéresse spontanément à la plus courte puisqu'elle révèle la région la plus riche. On s'intéresse donc à la distance

$$Y_{\min}^r = \min\{Y_1^r, Y_2^r, \dots, Y_{n-r}^r\}$$

dont on a besoin de connaître la loi pour savoir si la région la plus riche l'est significativement. La détermination de cette loi, ou plutôt son approximation, est l'objet du paragraphe suivant.

3.2.2 Méthode de Chen-Stein

Cas indépendant

Notons W_i la variable valant 1 si la distance cumulée Y_i^r est inférieure à y et 0 sinon :

$$W_i = \mathbb{I}\{Y_i^r \leq y\}.$$

Les W_i sont des variables de Bernoulli dont les espérances p_i sont toutes égales à $F_r(y)$. Notons de plus $R(y)$ le nombre de distances cumulées inférieures à une distance y donnée :

$$R(y) = \sum_{i=1}^{n-r} W_i. \tag{3.1}$$

Si les distances cumulées $\{Y_1^r, Y_2^r, \ldots, Y_{n-r}^r\}$ était indépendantes, $R(y)$ suivrait une loi binomiale :

$$R(y) \sim \mathcal{B}[n - r, F_r(y)].$$

Pour une longueur de séquence tendant vers l'infini ($\ell \to \infty$) –pour un nombre d'occurrences tendant également vers l'infini : $n \to \infty$– et une distance y petite, cette loi binomiale tendrait vers une loi de Poisson

$$R(y) \approx \mathcal{P}[(n - r)F_r(y)].$$

La méthode de Chen-Stein propose d'approcher la loi de $R(y)$ par cette loi de Poisson dans le cas dépendant (alors qu'elle n'est valable que dans le cas indépendant) et fournit une majoration de l'erreur ainsi commise.

Approximation poissonnienne de type Chen-Stein

Arratia *et al.* (1989) font une présentation très claire et intuitive des approximations de type Chen-Stein en se concentrant sur le cas d'un processus à temps discret, comme le processus des indicatrices $\{W_i\}_{i\geq 1}$. Le résultat central porte sur la distance en variation totale (notée d_{TV}) entre $R(y)$ et une v.a. $Z(y)$ suivant une loi de Poisson de paramètre $[(n - r)F_r(y)])$:

$$d_{\text{TV}}[R(y), Z(y)] = \sup_A |\Pr\{R(y) \in A\} - \Pr\{Z(y) \in A\}|$$

où le sup est pris sur tous les boréliens A. Cette distance donne la plus grande erreur qu'on puisse commettre, en termes de probabilités, en utilisant la loi de Poisson à la place de la vraie loi. Cette distance en variation totale est majorée par une somme de trois termes :

$$d_{\text{TV}}[R(y), Z(y)] \leq b_1 + b_2 + b_3.$$

b_1, b_2 et b_3 sont définis à partir d'une série de voisinages B_i définis pour chaque W_i :

$$\begin{aligned}
b_1 &= \sum_i \sum_{j \in B_i} p_i p_j, \\
b_2 &= \sum_i \sum_{j \in B_i, j \neq i} \mathbb{E}(W_i W_j), \\
b_3 &= \sum_i \mathbb{E}\left[\,(|\mathbb{E}(W_i) - p_i|)|\; \{W_j\}_{j \notin B_i}\right]
\end{aligned}$$

b_1 et b_2 traitent de la dépendance entre W_i et les W_j de son voisinage, b_3 traite de la dépendance vis-à-vis de l'extérieur de ce voisinage.

Application aux distances cumulées

Dans le cas des distances cumulées, la définition du voisinage est assez naturelle. Il est clair (cf figure 3.1) que chaque Y_i^r est dépendant des $r-1$ précédents et des $r-1$ suivants. *Si les distances simples sont indépendantes*, la distance cumulée Y_i^r est indépendante des Y_j^r tels que $|i-j| \geq r$. Donc, en utilisant le voisinage $B_i = \{j : |i-j| < r\}$, le terme b_3 est nul et on obtient la majoration suivante (cf paragraphe 5.3.2) :

$$d_{\mathrm{TV}}[R(y), Z(y)] \leq \left(1 - e^{-[n-r]F_r(y)}\right) \left[(2r-1)F_r(y) + 2 \sum_{k=1..r-1} F_k(y)\right].$$

Cette approximation de la loi de $R(y)$ fournit une approximation de celle de $Y^r_{\min}$. En effet, on a

$$\Pr\{Y^r_{\min} > y\} = \Pr\{R(y) = 0\} \simeq \Pr\{Z(y) = 0\}$$

et donc

$$\Pr\{Y^r_{\min} \leq y\} \simeq \exp\left[-(n-r)F_r(y)\right].$$

3.2.3 Détection de régions riches

Détermination du seuil de significativité

On peut à présent rechercher la distance y^* en dessous de laquelle la plus petite distance cumulée est significative au niveau α :

$$\exp[-(n-r)F_r(y^*)] = \alpha \tag{3.2}$$

soit $y^* = F_r^{-1}[-\log\alpha/(n-r)]$.

La distance seuil y^* n'est pas exacte puisqu'elle est obtenue grâce à une approximation. On peut en fournir un encadrement en calculant les distances y^*_{inf} et y^*_{sup} qui vérifient l'équation (3.2) à la distance $d_{\mathrm{TV}}[R(y), Z(y))]$ près :

$$\begin{aligned}
\alpha &= \exp[-(n-r)F_r(y^*_{\mathrm{inf}})] + d_{\mathrm{TV}}[R(y^*_{\mathrm{inf}}), Z(y^*_{\mathrm{inf}})], \\
\alpha &= \exp[-(n-r)F_r(y^*_{\mathrm{sup}})] - d_{\mathrm{TV}}[R(y^*_{\mathrm{sup}}), Z(y^*_{\mathrm{sup}})].
\end{aligned}$$

La détermination de la distance seuil y^* (et de son encadrement $[y^*_{\mathrm{inf}}; y^*_{\mathrm{inf}}]$) repose uniquement sur la loi des distances cumulées. Elle peut donc se faire dans le cadre du modèle M1 ou du modèle PC en utilisant respectivement les résultats (2.2) et (2.6) du chapitre 2.

Application au CHI d'*H. influenzae*

La détection de régions riches en occurrences du motif CHI dans le génome d'*H. influenzae* est étudiée dans le modèle M1 dans l'article du chapitre 5 et dans le modèle PC dans celui du chapitre 8. La figure 3.2 reprend la figure 8.1 de ce dernier chapitre. La fréquence des occurrences du motif y est représentée par le rapport r/Y^r qui est une estimation locale de l'intensité du processus des occurrences (on observe r occurrences en une distance Y^r). Cette figure permet de comparer les résultats obtenus avec les deux modèles. On y observe que le modèle M1 détecte 4 pics significatifs (au dessus de $r/y^*_{\min}$) et un pic douteux (entre $r/y^*_{\max}$ et $r/y^*_{\min}$) alors que le modèle PC n'en détecte aucun.

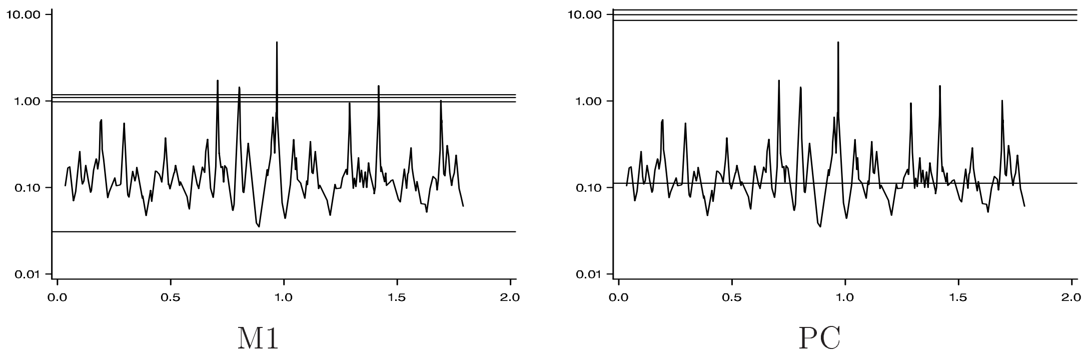


FIG. 3.2 – *Comparaison des seuils y^* pour le motif CHI de H. influenzae. L'abscisse indique la position le long du génome et l'ordonnée le rapport r/Y^r (en échelle logarithmique). Les 4 références horizontales donnent, respectivement, le rapport $r/\mathbb{E}(Y^r)$ sous le modèle considéré, $r/y^*_{\max}$, r/y^* et $r/y^*_{\min}$ pour $\alpha = 0.05$.*

On observe également dans la figure 3.2 que le modèle M1 est mal ajusté en moyenne puisque tout le graphe se situe au dessus de $r/\mathbb{E}(Y^r)$. Ce biais provient du fait que le motif CHI est sur-représenté par rapport au modèle M1 dans le génome de *H. influenzae* (comme dans la plupart des génomes où il intervient). Le modèle PC est ajusté par construction à la fréquence observée et ne présente pas ce biais. Le modèle de chaîne de Markov permettant de corriger ce biais serait le modèle M7 qui utilise $3 \times 4^7 = 19\ 152$ paramètres. On retrouve ici l'intérêt du modèle PC pour détecter une hétérogénéité pour un seul mot : il n'a recours qu'à deux paramètres (λ et a) spécifiquement liés au motif d'intérêt.

3.2.4 Limites de l'approche par distances cumulées

Hypothèses sur les distances simples

On a vu que l'approximation de type Chen-Stein appliquée aux distances cumulées utilise les hypothèses d'indépendance et d'équi-distribution des distances simples $\{Y_1, Y_2, \dots\}$.

De façon plus générale, ce type d'approximation porte sur des sommes de v.a. de Bernoulli $\{\xi_1, \xi_2, \dots\}$ dépendantes. Si ces variables ont des espérances $\{p_1, p_2, \dots\}$ différentes, on obtient une approximation de la loi de leur somme par une loi de Poisson $\mathcal{P}(\sum_i p_i)$. Ainsi, l'hypothèse d'équi-distribution n'est pas absolument requise (même si elle simplifie grandement les calculs).

Par contre, l'indépendance des distances simples $\{Y_1, Y_2, \dots\}$ est nécessaire pour limiter la portée de la dépendance entre les distances cumulées $\{Y_1^r, Y_2^r, \dots\}$ et annuler le terme b_3 de la borne de Chen-Stein.

Répartition des occurrences d'un mot

On a vu au chapitre 2 que, dans le modèle M1 comme dans le modèle PC, les distances séparant les occurrences d'un mot $\mathbf{w}$ sont i.i.d.. Ainsi, l'approche par distances cumulées peut toujours être utilisée pour étudier l'homogénéité de la répartition des occurrences d'un mot.

Répartition des occurrences d'une famille de mots

Dans le cas général d'une famille de mots $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, \dots\}$ les distances séparant les occurrences des différents mots de la famille ne sont pas i.i.d.. On propose ici l'exemple d'une famille de trois mots :

$$\mathcal{W} = \{\mathbf{w}_1 = \texttt{tagc}, \mathbf{w}_2 = \texttt{agca}, \mathbf{w}_3 = \texttt{aagg}\}.$$

Dans cet exemple, $\mathbf{w}_2$ a une forte probabilité de recouvrir $\mathbf{w}_1$ et a une probabilité plus faible d'être recouvert par $\mathbf{w}_3$ qui ne peut être recouvert par aucun mot de $\mathcal{W}$. Dans un modèle M00, la matrice $\mathbf{A}$ de terme général $a(i, j)$ vaut

$$\mathbf{A} = \begin{pmatrix} 0 & 1/4 & 0 \\ 0 & 1/64 & 1/64 \\ 0 & 0 & 0 \end{pmatrix}.$$

Dans cet exemple (comme dans le cas général), la distance Y_i dépend du mot de $\mathcal{W}$ qui apparaît en X_i. Si ce mot est $\mathbf{w}_3$, il ne pourra pas y avoir de recouvrement et la distance Y_i ne pourra pas être nulle dans le modèle PC ou inférieure à 4 dans le modèle M1.

D'autre part, si on observe un recouvrement ($Y_i = 0$ dans PC et $Y_i < 4$ dans M1), il s'agit plus probablement d'un recouvrement de $\mathbf{w}_2$ sur $\mathbf{w}_1$ que de $\mathbf{w}_3$ sur $\mathbf{w}_2$. Ainsi, la distance suivante ne peut pas partir de $\mathbf{w}_1$ (qui ne recouvre personne) et elle part plus probablement de $\mathbf{w}_2$ que de $\mathbf{w}_3$. Cette distance n'est donc pas indépendante de la précédente.

Le cas particulier du CHI d'*H. influenzae* est donc un exemple heureux. Comme on l'a vu à la section 2.6, la structure de recouvrement des quatre mots composant ce motif permet de faire l'hypothèse d'indépendance et d'équi-distribution des distances simples. Nous nous intéresserons dans la section suivante aux occurrences de palindromes de longueur 6 pour lesquels cette hypothèse ne tient plus.

3.3 Fenêtre mobile

L'analyse de la répartition par fenêtre mobile est présentée au chapitre 8 et reprise au chapitre 9. Elle est définie dans le cadre du modèle de Poisson composé et on se limite à ce modèle dans toute ce paragraphe.

3.3.1 Principe et notations

L'approche par fenêtre mobile se fonde sur le comptage du nombre d'occurrence dans une fenêtre de largeur fixe Δx mobile le long de la séquence. On rappelle que le processus de comptage des occurrences est noté $\{N(x)\}_{x \geq \Delta x}$ et on définit

$$\Delta N(x) = N(x) - N(x - \Delta x)$$

le comptage dans la fenêtre de largeur Δx finissant en x.

Loi du comptage

Dans le modèle PC, quelle que soit la distribution g de la taille K des trains (géométrique ou non), on connaît la loi du comptage $\Delta N(x)$ en une position x quelconque de la séquence (cf paragraphe 2.5.2). En tout point x, $\Delta N(x)$ est distribué comme le comptage dans une séquence de longueur Δx :

$$\Delta N(x) \sim \mathcal{PC}(\lambda \Delta x; g).$$

On peut ainsi calculer la probabilité qu'une fenêtre contienne plus de n occurrences, ou calculer un seuil de significativité n (au niveau α) défini par

$$n = \min\{m : \Pr[N(\Delta x) \geq m] \leq \alpha\}.$$

Cependant, et comme dans l'approche par distances cumulées, ce résultat n'est pas suffisant puisqu'on s'intéresse à la loi du comptage *dans la fenêtre la plus riche*, c'est à dire à la loi de la v.a.

$$\Delta N_{\text{sup}} = \sup_{\Delta x \leq x \leq \ell} \{\Delta N(x)\}.$$

3.3.2 Significativité de la fenêtre la plus riche

L'approximation de la loi de ΔN_{sup} demeure (à ma connaissance du moins) un problème ouvert. La difficulté provient essentiellement du fait que le processus $\{\Delta N(x)\}$ est un processus à temps continu.

Heuristique

Le livre de Aldous (1989) est sans doute l'un des plus riches sur les approximations poissonniennes. Appliquée à la loi de ΔN_{sup}, l'heuristique proposée par Aldous amène à approcher la loi du nombre $M(n)$ de positions x auxquelles le processus $\{\Delta N(x)\}$ dépasse un seuil n (élevé) par une loi de Poisson (éventuellement composée). On peut ainsi approcher, entre autre, la probabilité que ce nombre de dépassements soit nul :

$$\Pr\{M(n) = 0\} \simeq \exp[-(\ell - \Delta x) \Pr\{N(\Delta x) > n\}]$$

et donc approcher la loi de ΔN_{sup} :

$$\begin{aligned} \Pr\{\Delta N_{\text{sup}} \leq n\} &= \Pr\{M(n) = 0\} \\ &\simeq \exp[-(\ell - \Delta x) \Pr\{N(\Delta x) > n\}]. \end{aligned} \tag{3.3}$$

Le recours à la loi de Poisson *composée* vient de la structure de dépendance du processus $\Delta N(x)$ qui fait que les dépassements auront tendance à survenir par paquets.

Le livre de Aldous est cependant très particulier dans la mesure où il présente la plupart des résultats de façon très intuitive et sans démonstration. Il a cependant inspiré une grande partie de la littérature sur les approximations poissonniennes et est, par exemple, constamment cité dans les articles de Barbour *et al.* (voir la bibliographie du chapitre 8).

Approximation de type Chen-Stein

La méthode de Chen-Stein est applicable à des processus à temps continu comme le processus $\{\Delta N(x)\}_{x \geq 0}$ (voir notamment Barbour et Brown, 1992b). La qualité de cette approximation est mesurée par une distance entre la vraie loi du nombre de dépassements et la loi de Poisson. La distance en variation totale n'est pas la seule possible ; des résultats existent également concernant la distance de Wasserstein et la distance de Hellinger. Les propriétés et les relations entre ces différentes distances sont présentées à la fin de Barbour *et al.* (1992).

À des fins pratiques, la distance en variation totale semble la plus utile dans la mesure où elle s'exprime directement en termes de probabilité. Malheureusement, il semble très difficile d'obtenir des résultats sur cette distance pour le processus $\{\Delta N(x)\}$ (cf paragraphe 8.4.2). Le seul résultat applicable est fourni par le théorème 2.6 de Barbour et Brown (1992b) : il montre que l'approximation (3.3) de la loi de ΔN_{sup} est optimale en termes de distance en variation totale, mais ne permet pas de calculer explicitement le majorant de cette distance. On dispose donc d'une approximation optimale sans en connaître la précision.

3.4 Prise en compte d'une hétérogénéité

Les modèles Mk et PC sont fondés sur l'hypothèse d'homogénéité de la loi des occurrences le long de la séquence. Or les séquences génomiques sont hétérogènes ou *segmentées*. On cite ici deux exemples de segmentation et les conséquences qu'elles peuvent avoir sur la répartition de certains motifs.

1. Le rapport des fréquences entre les nucléotides `a` et `t` d'une part et `g` et `c` d'autre part n'est pas constant tout au long des génomes. Il existe des régions « riches en `g-c` » et d'autres « pauvres en `g-c` ». Cette différences de composition induit mécaniquement des différences de fréquences pour les motifs eux-mêmes riches en `a-t` ou en `g-c`.
2. Certaines régions des génomes (les « gènes ») codent pour des protéines et d'autres non. Une altération dans une région codante peut avoir des conséquences létales pour l'organisme. On peut alors penser que la fréquence de motifs protecteurs (comme le CHI) doit être systématiquement plus élevée dans ces régions.

Dans certains cas, ces segmentation sont connues *a priori*

- soit de façon déterministe (les limites des régions sont strictement définies) comme dans le cas de séquences annotées,
- soit de façon plus nuancée (sous forme de probabilité d'appartenance de chaque position à l'une ou l'autre des catégories) si la segmentation de la séquence provient par exemple

d'un modèle de chaîne de Markov cachée.

L'objectif du projet d'article présenté au chapitre 9 est d'étudier la répartition des motifs en corrigeant l'effet de cette hétérogénéité connue *a priori*.

3.4.1 Modélisation de l'hétérogénéité

Le modèle utilisé au chapitre 9 est un modèle de Poisson composé semblable à celui du chapitre 8 (présenté à la section 2.5) mais hétérogène. Le problème principal réside dans la modélisation de cette hétérogénéité. On note S le nombre de catégories de régions. On a par exemple

$S = 2$ si on distingue les régions riches en `g-c` des régions pauvres en `g-c` ;

$S = 3$ si on distingue les régions inter-géniques (non-codantes), géniques codant dans le sens direct et géniques codant dans le sens indirect.

On suppose que l'information sur l'hétérogénéité est connue sous la forme de probabilités d'appartenance à chaque catégorie s en chaque position x

$$\tau_s(x) = \Pr\{\text{la position } x \text{ de la séquence appartient à la catégorie } s\}.$$

Si la segmentation est déterministe, les probabilités $\tau_s(x)$ valent toutes 0 ou 1.

Occurrences de trains

Le processus d'occurrences des trains $\{C(x)\}_{x\geq 0}$ est modélisé par un processus de Poisson hétérogène d'intensité

$$\lambda(x) = \sum_{s=1}^{S} \tau_s(x)\lambda_s.$$

$(\lambda_1, \ldots, \lambda_S)$ sont les intensités qui caractérisent la fréquence des trains dans les différents types de régions. L'*intensité cumulée* $\Lambda(x, y)$ entre deux positions x et y définie par $\Lambda(x, y) = \int_x^y \lambda(z)\mathrm{d}z$ se décompose en

$$\Lambda(x, y) = \int_x^y \sum_{s=1}^{S} \tau_s(z)\lambda_s \mathrm{d}z = \sum_{s=1}^{S} \tau_s(z)\Pi_s(x, y) \tag{3.4}$$

en notant $\Pi_s(x, y)$ la *probabilité cumulée* entre x et y dans la catégorie s : $\Pi_s(x, y) = \int_x^y \tau_s(z)\mathrm{d}z$.

Taille des trains

De façon générale, la loi g de la taille des trains est modélisée par un mélange de distributions $(g_1, \ldots, g_S)$ correspondant à chacune des catégories :

$$g(k \mid x) = \Pr\{K = k \mid X = x\} = \sum_s \tau_s(x)g_s(x).$$

Le cas où les lois sont géométriques est, comme au chapitre 8, particulièrement favorable. Encore plus favorable est le cas où les loi g_s sont géométriques *et égales entre elles* : $g_1 = \cdots = g_S$.

3.4.2 Correction de l'hétérogénéité

Homogénéisation de la séquence

L'influence de l'hétérogénéité sur les occurrence des trains est décrite par l'équation (3.4). Si on sait estimer les paramètres $(\lambda_1, \ldots, \lambda_S)$ (et connaissant les $\{\tau_s(x)\}$), cette équation permet de rendre homogène le processus $\{C(x)\}$ des occurrences des trains grâce à la transformation

$$\tilde{X}'_i = \Lambda(0, X'_i) \tag{3.5}$$

où X'_i désigne la position du i-ème train.

Le processus $\{\tilde{C}(\tilde{x})\}$ sautant de 1 à chaque $\tilde{X}'_i$ est alors un processus de Poisson homogène (d'intensité 1). Si l'hétérogénéité n'affecte pas la taille des trains, on est ainsi ramené au modèle de Poisson composé homogène déjà étudié. On peut ainsi appliquer les techniques fondées sur les distances cumulées ou sur une fenêtre mobile dans une séquence « homogénéisée » de longueur $\tilde{\ell} = \Lambda(0, \ell)$.

Distances cumulées

Le cas du processus de Poisson hétérogène est étudié dans l'article de Dembo et Karlin (1992) présentant les distances cumulées. Les auteur y calculent un seuil de significativité y^* pour la distance cumulée la plus courte $Y^r_{\min}$. Ce seuil fait intervenir l'intensité cumulée sur l'ensemble de la séquence $\Lambda(0, \ell)$ et est le même en toute position. Ainsi, si on s'intéresse à un motif $\mathcal{W}$ riche en `g-c` et qu'on prend en compte une hétérogénéité dans la fréquence en `g-c` de la séquence, on s'expose à détecter, malgré tout, les régions riches en un $\mathcal{W}$ précisément dans des régions riches en `g-c`.

En utilisant la transformation (3.5), on peut d'abord définir un seuil $\tilde{y}^*$ pour la plus petite distance cumulée homogénéisée $\tilde{Y}^r_{\min}$ (en notant $\tilde{Y}_i = \tilde{X}_{i+1} - \tilde{X}_i = \Lambda(X_i, X_{i+1})$). Ce seuil est constant le long de la séquence homogénéisée mais devient variable le long de la séquence réelle quand on applique la transformation inverse de (3.5) pour obtenir $y^*(x)$.

Ainsi, en reprenant l'exemple d'un motif riche en `g-c` et d'une segmentation *a priori* entre régions riches et pauvres en `g-c`, le seuil $y^*(x)$ sera plus bas dans les régions riches en `g-c` que dans les régions pauvres : on exigera une densité d'occurrences plus grande dans les régions favorables à l'apparition du motif pour déclarer la zone significativement riche.

Cas général

Si l'hétérogénéité affecte également la taille des trains, il n'existe pas de transformation simple analogue à (3.5) permettant de rendre la distribution g homogène le long de la séquence. Dans ce cas donc, la loi des distances cumulées Y^r_i ou la loi des comptages $\Delta N(x)$ dans la fenêtre mobile dépend de la position dans la séquence. Les méthodes de détection d'hétérogénéités présentées dans le cas du processus de Poisson homogène sont encore applicables mais avec un coût calculatoire bien plus grand.

Bibliographie

ALDOUS, D. (1989). *Poisson Approximations via the Poisson Clumping Heuristic.* Springer-Verlag: New-York.

ARRATIA, R., GOLDSTEIN, L. et GORDON, L. (1989). Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Prob.* **17** 9–25.

BARBOUR, A. D. et BROWN, T. (1992a). The Stein-Chen method, point process and compensators. *Ann. Prob.* **20** 1504–1527.

BARBOUR, A. D. et BROWN, T. (1992b). Stein's method and point process approximation. *Stoch. Proc. Appl.* **43** 9–31.

BARBOUR, A. D., HOLST, L. et JANSON, S. (1992). *Poisson Approximation.* Oxford-University Press.

DEMBO, A. et KARLIN, S. (1992). Poisson approximations for r-scan processes. *Ann. Appl. Prob.* **2 (2)** 329–357.

KARLIN, S. et MACKEN, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Amer. Statist. Assoc.* **86** 27–35.

Deuxième partie

Occurrences de motifs et chaînes de Markov

Chapitre 4

Exact distribution of word occurrences in a random sequence of letters

Robin S. & Daudin J.J. (1999)

Journal of Applied Probability, **36** (1), 179–193

Abstract. The study of the distribution of the distance between words in a random sequence of letters is interesting in view of application in genome sequence analysis. In this paper we give the exact distribution probability and cumulative distribution function of the distances between two successive occurrences of a given word and between the n-th and the $(n+m)$-th occurences under three models of generation of the letters : iid with the same probability for each letter, iid with different probabilities and Markov process. The generating function and the first two moments are also given. The point of studying the distances in yplace of the counting process is that we get some knowledge not only about the frequency of a word but also about its longitudinal distribution in the sequence.

Keywords. distance between occurrences, genome sequence analysis, Markov chain, patterns, waiting time.

Contents

4.1 Introduction

Statistical methods are useful to analyze genome DNA sequences. Much work have been done to find words with unexpected frequencies using Gaussian asymptotic methods (Prum *et al.* (1995), Cowan (1991)) or Poisson or compound Poisson approximation (Godbole (1991), Geske *et al.* (1995), Schbath (1995)). The study of the distances between occurrences, giving more insight about the inhomogeneities of the DNA sequence, have been studied from an empirical point of view (Henaut *et al.* (1996)) or from an asymptotic one (Karlin and Macken (1991)).

This problem is also closely related to the waiting times of success runs and to the distributions of order k. The classical framework for the run problem has been proposed by Feller (1968). The distributions of success runs of fixed length k have been termed as distribution of order k (Philippou *et al.* (1983), Johnson *et al.* (1992)). Extensions to the Markov case have been proposed by Schwager (1983), Aki and Hirano (1993) and Uchida and Aki (1995). A systematic study for the waiting time distribution of the r-th appearance of a success run has been made by Koutras (1997) using a Markov chain imbedding technique. Mori (1989) deals with repetitions of a single letter in an alphabet of more than two letters and gives the limit behaviour of the number of different words observed before a given word appears.

In this paper we deal with a more general problem: we consider a random sequence of letters taking their values in an alphabet of more than two letters and we study the waiting times before the n-th occurrences of a general pattern (or word) of k letters. Our work may be applied to the particular case of the runs and our results are in agreement with these papers.

Exact results on the distribution of the distances between words are not so intricate as it would appear at a first glance. Blom and Thorburn (1982) have given a powerful and simple way to obtain the generating function in the iid case. These results have been extended to the Markov by Chryssaphinou and Papastavridis (1990) and semi-Markov cases under quite general assumptions by Rudander (1996). The former did not give an explicit generating function except in the case of a two letters alphabet. In this paper we use Blom & Thorburn's way to obtain a recurrence relation between the probabilities in the iid and Markov cases. We give a modified recurrence relation which is more efficient and allows us to obtain the probabilities and the cumulative probability function even for long sequences. Therefore, the exact method we propose, which is naturally useful for short sequences ($< 10^4$ bases), may also be used to validate asymptotic methods.

We give new results on the distance between the n-th and the $(n+m)$-th occurrences and obtain associate generating functions and the two first moments.

Using our results and assuming that a Markovian model is true, it is possible to give an answer to the following questions :

1. Does a given word have an unexpected frequency?

2. Is its longitudinal distribution along the sequence homogeneous? The alternatives are a too regular or a clustered distribution.

3. Does a given short fragment of the sequence contain too many occurrences of the word (or too few)?

4. Is the distance between two different words unexpectedly short or long?

In the first part of the paper, we deal with the general mathematical problem without any reference to the genome. The second part is devoted to some brief application to DNA sequences.

4.1.1 General notations

Let $\mathbf{S} = (S_1 S_2 ...)$ be a random sequence of letters taking their values in an alphabet $\mathcal{A}$ with cardinality N and $W = (w_1 ... w_k)$ be a given word of length k. In the following, we shall say that W occurs at position x in the sequence $\mathbf{S}$ if the word W ends at position x:

$$\{W \text{ at } x\} = \{(s_{x-k+1} .. s_x) = (w_1 ... w_k)\} .$$

Let us denote by X_n the random position of the n-th occurrence of W in $\mathbf{S}$ and

$$p(x, n) = \Pr\{X_n = x\} = \Pr\{W \text{ at } x \text{ for the } n\text{-th time in } \mathbf{S}\} ,$$

For $n = 1$, we denote: $X = X_1,\ p(x) = p(x, 1)$. In the models we study here all distances between two successive occurrences are independent and identically distributed (iid). So we can define Y_r to be the distance between the n-th and the $(n + r)$-th occurrence for any $n > 0$

$$Y_r = X_{n+r} - X_n$$

and denote $q(\cdot, r)$ its distribution:

$$q(y, r) = \Pr\{Y_r = y\} .$$

where, as with X_n, we denote $Y = Y_1$ and $q(y) = \Pr\{Y = y\} = q(y, 1)$.

4.1.2 Overlapping structure

The distribution of the occurrences along the sequence depends of the word itself. For example, it is obvious that the length of the word has a great influence on how often it occurs. It also depends on its overlapping structure e.g. on the way the word can overlap himself. This structure is given by the overlapping indicator ε (called *leading number* by Conway):

$$\varepsilon(u) = \mathbf{1}\{(w_{k-u+1} .. w_k) = (w_1 ... w_u)\} \qquad \text{for } 1 \le u \le k ,$$

$\varepsilon(u)$ is equal to one if the first u letters of the word are the same (and in the same order) as the last u letters. If $W = (gagag)$, $\varepsilon(u)$ is 1 for $u = 1, 3, 5$ and 0 for $u = 2, 4$.

Obviously, we always have $\varepsilon(k) = 1$. In the following, we shall call *non-overlapping* a word for which all $\varepsilon(u)$ are null except for $u = k$.

In the case of the success runs of length k we have $\varepsilon(u) = 1$ for $u = 1, ...k$. Therefore we obtain the results of Koutras (1997) for the type III distribution as a particular case.

4.1.3 Models

Simplest model (M00) : First of all, we shall consider the case of a random sequence of iid letters taken with uniform probabilities on $\mathcal{A}$. This model has already been studied by Blom and Thorburn (1982) and Blom (1982) but we use it as an introduction and give a unified presentation of the results. This model is denoted M00 in Prum *et al.* (1995) with reference to a Markov chain with no memory (first "0") and uniform distribution on $\mathcal{A}$ (second "0").

In any position x, a letter a has probability N^{-1} to occur and the whole word W has probability N^{-k} (for $x \geq k$).

General iid sequence (M0) : All the results obtained in the M00 model can be generalized to an iid sequence of letters with any distribution μ on $\mathcal{A}$. This model is called M0 in Prum *et al.* (1995) since the sequence has no memory.

Here, in any position x, a letter a has probability $\mu(a)$. Denoting

$$\tau(u,v) = \prod_{z=u}^{v} \mu(w_z)$$

the whole word W has probability $\mu(W) = \tau(1,k)$. The generating functions are the same as in the M00 case but the polynomial δ (which will be defined in section **2**) is

$$\delta(t) = \sum_{u=1}^{k} \varepsilon(u) \frac{t^u}{\tau(1,u)} \; .$$

The great similarity of the M00 and M0 models causes us to ask the reader to write the results for the M0 case on his own.

First order Markov model (M1) : Another interesting model is the case where the sequence $\mathbf{S}$ is a first order Markov chain with transition probability π: in any position x and any letters a and b of $\mathcal{A}$,

$$\pi(a,b) = \Pr\{S_{x+1} = b \mid S_x = a\} \; .$$

In the following, we shall assume that the sequence is in its steady state i.e. that, in any position x, a letter a has probability $\mu(a)$ to occur where μ is the stationary distribution of π :

$$\mu = \mu.\pi \; .$$

Here we use the notation $\tau(u,v)$ to denote the probability to observe successively $w_u, w_{u+1}, ...w_v$ given the fact that the letter w_{u-1} has occurred just before: for any $2 \leq u \leq v \leq k$

$$\tau(u,v) = \prod_{x=u}^{v} \pi(w_{x-1}, w_x) \; .$$

It is straightforward that, in any position x, the word W is observed with probability $\mu(W) = \mu(w_1)\tau(2,k)$.

4.2 Distribution of the first occurrence

4.2.1 Recurrence

A recurrence on $p(x)$ is given by Blom and Thorburn (1982) (theorem 2.1) in the M00 model. It is obtained by decomposing $\{W \text{ at } x\}$ into $\{W \text{ at } x \text{ for the first time}\}$ (which has probability $p(x)$) and $\{W \text{ at } x \text{ not for the first time}\}$. The last term is itself decomposed in two parts: if the occurrence of W at x is not the first, the first one has occurred at some position $z < x$ and different cases are considered according to the distance between z and x:

$$\Pr\{W \text{ at } x \text{ not for the first time}\} = \sum_{z \leq x-k} p(z) N^{-k} + \sum_{z=x-k+1}^{x-1} p(z)\varepsilon(k-x+z)N^{z-x} \,.$$

We finally get for $x > k$

$$p(x) = N^{-k} - N^{-k}\sum_{z=k}^{x-k} p(z) - \sum_{z=x-k+1}^{x-1} p(z)\varepsilon(k-x+z)N^{z-x} \tag{4.1}$$

and $p(x) = 0$ if $x < k$, $p(k) = N^{-k}$.

Note that the geometric distribution $\Pr\{X = x\} = (1 - N^{-k})^{x-k-1}N^{-k}$ which is sometimes used to deal with such problem (see Hénaut & *al* . Henaut *et al.* (1996)) is not correct since it does not satisfy to (4.1). We shall see in the following that neither the mean nor the variance are the same. The geometric distribution is exact only with words of length 1.

Pekoz (1996) has developed the Stein-Chen's techniques to bound errors for the geometric approximation. Using (4.1) we can compute the distance in total variation d_{TV} between the true distribution and its geometric approximation under the M00 model and compare it to Pekoz's bound. For example, for $W = (\texttt{gatc})$ we compute $d_{TV} = 0.0043$ while Pekoz gives 0.012 and for $W = (\texttt{aaaa})$, $d_{TV} = 0.28$ and Pekoz gives 0.32 . Pekoz's bound seems to be better for overlapping words (such as $W = (\texttt{aaaa})$) than for non-overlapping ones.

Theorem 1 *In a first order Markov sequence, the distribution of the first occurrence X is given by the following recurrence: for $x > k$*

$$\begin{aligned} p(x) = \;& \mu(W) - \textstyle\sum_{z=k}^{x-k} p(z)\pi^{x-z-k+1}(w_k, w_1)\tau(2,k) \\ & - \textstyle\sum_{z=x-k+1}^{x-1} p(z)\varepsilon(k-x+z)\tau(k-x+z+1,k) \end{aligned} \tag{4.2}$$

and $p(x) = 0$ for $x < k$, $p(k) = \mu(W)$ ($\pi^n(a,b)$ is the transition probability from a to b in n steps).

Proof. Since the process in its steady state, $\{W \text{ at } x\}$ occurs with probability $\mu(W)$. $p(x)$ is the probability of $\{W \text{ at } x \text{ for the first time}\}$; the first sums consider all events $\{W \text{ in } z \text{ for the first time and then at } x\}$ for $z < z - k$ and the second for $x - k < z < x$. In the first case, the process has to go from letter w_k to letter w_1 in exactly $(x - z - k + 1)$ steps (which occurs with probability $\pi^{x-z-k+1}(w_k, w_1)$) and one has to observe successively

$w_2, w_3 ..., w_k$ (which occurs with probability $\tau(2,k)$). In the second case, if the overlapping is possible (e.g. if $\varepsilon(k-x+z)=1$), the first $(k-x+z)$ letters are given and one only has to observe the sequence of letters $w_{k-x+z+1}, w_{k-x+z+2}, ...w_k$ (probability $\tau(k-x+z+1,k)$). ■

4.2.2 Generating function

Theorem 2.2 in Blom and Thorburn (1982) gives the probability generating function of $\phi_X(t)=\mathbb{E}(t^X)$ in the M00 case:

$$\phi_X(t)=\left[1+(1-t)\delta\left(\frac{1}{t}\right)\right]^{-1} \tag{4.3}$$

where $\delta(t)=\sum_{u=1}^{u=k}\varepsilon(u)(Nt)^u$. This result is derived from the recurrence (4.1): since $\phi_X(t)=\sum_{x\geq k}p(x)t^x$, one has
$\phi_X(t)=N^{-k}\sum_{x\geq k}t^x-N^{-k}\sum_{x\geq k}t^x\sum_{z=k}^{x-k}p(z)-\sum_{x\geq k}t^x\sum_{z=x-k+1}^{x-1}p(z)\varepsilon(k-x+z)N^{z-x}$
$=N^{-k}\dfrac{t^k}{1-t}-N^{-k}\sum_{z\geq k}p(z)t^z\sum_{x\geq z+k}t^{x-z}-\sum_{z\geq 1}p(z)t^z\sum_{x=z+1}^{z+k-1}\varepsilon(k-x+z)\left(\dfrac{N}{t}\right)^{z-x}$
$=N^{-k}\dfrac{t^k}{1-t}-N^{-k}\phi_X(t)\dfrac{t^k}{1-t}-\phi_X(t)\sum_{v=1}^{k-1}\varepsilon(k-v)\left(\dfrac{N}{t}\right)^{-v}$;
moving the last term on the other side and multiplying by $\left(\frac{N}{t}\right)^k(1-t)$ gives

$$\phi_X(t)\left(\frac{N}{t}\right)^k(1-t)\left[1+\sum_{v=1}^{k-1}\varepsilon(k-v)\left(\frac{N}{t}\right)^{-v}\right]=1-\phi_X(t)$$

which is equivalent to $\phi_X(t)(1-t)\delta(t^{-1})=1-\phi_X(t)$ since $\varepsilon(k)\equiv 1$. The result follows.

Theorem 2 *In the M1 model, the generating function of X is*

$$\phi_X(t)=\left[\gamma_{k,1}(t)+(1-t)\delta\left(\frac{1}{t}\right)\right]^{-1}. \tag{4.4}$$

where $\delta(t)=\sum_{u=1}^{k}\dfrac{\varepsilon(u)t^u}{\mu(w_1)\tau(2,u)}$ *and* $\gamma_{k,1}(t)=\dfrac{1-t}{t\mu(w_1)}\sum_{z\geq 1}\pi^z(w_k,w_1)t^z$.

Proof. The principle of the proof is the same as for (4.3). ■

Note that in a model M0, $\pi^n(a,b)=\mu(b)$ for all $n>0$, so that $\gamma_{k,1}(t)=(1-t)\sum_{z\geq 0}t^z=1$, which is in agreement with (4.3) with the same definition of δ. Rudander (1996) gives another form of this generating function and the associated moments. Chryssaphinou and Papastavridis (1990) give a system of linear equations statisfied by $\phi_X(t)$ and some other generating functions but do not give explicitly $\phi_X(t)$.

4.2.3 Moments

A straightforward application of the general properties of ϕ_X gives the first two moments of X: $\mathbb{E}(X) = \phi'_X(1)$, $\mathbb{V}(X) = \phi''_X(1) + \phi'_X(1)\,[1 - \phi'_X(1)]$. In the M00 case, we get

$$\begin{aligned} \mathbb{E}_{\mathrm{M00}}(X) &= \delta(1)\ , \\ \mathbb{V}_{\mathrm{M00}}(X) &= \delta^2(1) - 2\delta'(1) + \delta(1)\ . \end{aligned}$$

An amazing consequence is that, in a random sequence of fair coin tossing ($N = 2$), the mean position of the first occurrence of the word $W = (01)$ is $\delta_{(01)}(1) = 2^2 = 4$ but for the word $W = (11)$ it is $\delta_{(11)}(1) = 2 + 2^2 = 6$. The variances are also very different since $\mathbb{V}(X)$ is 4 for (01) and 22 for (11)! Blom (1982) discusses the consequences of this result.

It has to be pointed out that $\delta(t) = (Nt)^k$ for a non-overlapping word so that $\mathbb{E}_{\mathrm{M00}}(X)$ is always greater for an overlapping word than for a non-overlapping word with same length. The same remark can be made for the variance.

In the first order Markov model, one needs to evaluate the first and second derivatives of $\gamma_{k,1}(t)$ to give the moments of X using ϕ_X , . This can be done using another form of this function given by the following lemma. Let $\mathbf{\Pi}$ be the transition matrix of the Markov chain: $\mathbf{\Pi}_{a,b} = \pi(a,b)$.

(H) : $\mathbf{\Pi}$ admits a decomposition $\mathbf{\Pi} = \mathbf{A}.\mathbf{\Lambda}.\mathbf{B}$ where $\mathbf{A}.\mathbf{B} = \mathbf{B}.\mathbf{A} = \mathbf{I}$ and $\mathbf{\Lambda}$ is a diagonal matrix where $\lambda_1 = 1$ and $1 > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_N|$.

Lemma 3 *Under assumption* **(H),** *one has*

$$\gamma_{k,1}(t) = 1 + \frac{1-t}{\mu(w_1)} \sum_{i=2}^{N} \frac{\lambda_i}{1 - t\lambda_i} a_{w_k,i} b_{i,w_1} \tag{4.5}$$

where $a_{i,j}$ and $b_{i,j}$ are the general terms of $\mathbf{A}$ *and* $\mathbf{B}$.

Proof. One has $\gamma_{k,1}(t) = \mathbf{\Gamma}_{w_k,w_1}(t)$ where $\mathbf{\Gamma}(t)$ is the matrix defined for any $t > 0$ as

$$\mathbf{\Gamma}(t) = \frac{1-t}{t\mu(w_1)} \left[(\mathbf{I} - t\mathbf{\Pi})^{-1} - \mathbf{I}\right] = \frac{1-t}{t\mu(w_1)} \sum_{z \geq 1} (t\mathbf{\Pi})^z$$

and $\mathbf{I}$ is the $N \times N$ identity matrix.

Using **(H)**, it follows that

$$\begin{aligned} \mathbf{\Gamma}(t) &= \frac{1-t}{t\mu(w_1)} \mathbf{A}.\left[(\mathbf{I} - t\mathbf{\Pi})^{-1} - \mathbf{I}\right].\mathbf{B} = \frac{1-t}{t\mu(w_1)} \sum_{i=1}^{N} \left(\frac{1}{1-t\lambda_i} - 1\right) \mathbf{a}_i.\mathbf{b}_i \\ &= \frac{1-t}{\mu(w_1)} \sum_{i=1}^{N} \frac{\lambda_i}{1-t\lambda_i} \mathbf{a}_i.\mathbf{b}_i \end{aligned}$$

where $\mathbf{a}_i$ denotes the i-th column of $\mathbf{A}$ and $\mathbf{b}_i$ the i-th row of $\mathbf{B}$ so that $(\mathbf{a}_i.\mathbf{b}_j)_{u,v} = a_{u,i} b_{j,v}$. Note that $\mathbf{a}_1 = [1, 1, ...1]'$ and $\mathbf{b}_1 = \mu$ the stationary distribution. Using the fact that $\lambda_1 = 1$, we get

$$\mathbf{\Gamma}(t) = \frac{\mathbf{a}_1 \mathbf{b}_1}{\mu(w_1)} + \frac{1-t}{\mu(w_1)} \sum_{i=2}^{N} \frac{\lambda_i}{1-t\lambda_i} \mathbf{a}_i.\mathbf{b}_i$$

and the lemma is established. ■

$\gamma'_{k,1}$ and $\gamma''_{k,1}$ can be derived from (4.5) and assume a quite simple form for $t = 1$ since, for $m \geq 1$,

$$\gamma_{k,1}^{(m)}(1) = -\frac{m!}{\mu(w_1)} \sum_{i=2}^{N} \frac{\lambda_i^m}{(1-\lambda_i)^m} a_{w_k,i} b_{i,w_1} .$$

for $m \geq 1$. This gives the first two moments of X under M1:

$$\begin{aligned} \mathbb{E}_{\text{M1}}(X) &= \delta(1) - \gamma'_{k,1}(1) , \\ \mathbb{V}_{\text{M1}}(X) &= \left[\delta(1) - \gamma'_{k,1}(1)\right]^2 - 2\delta'(1) - \gamma''_{k,1}(t) + \left[\delta(1) - \gamma'_{k,1}(1)\right] . \end{aligned}$$

For example, with $\mathcal{A} = \{0, 1\}$ and $\mathbf{\Pi} = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, the mean position of the first occurrence of $W = (01)$ is $11/2$ and its variance is $45/4$. For $W = (11)$ the mean is the same and the variance is $81/4$. In this model, successions of identical letters are expected to be frequent but, for the mean, it is exactly balanced by the respective overlapping structures of the words. The variance is still greater for the overlapping word than for the non-overlapping one.

4.3 Distribution of the distance between two successive occurrences

4.3.1 Recurrence

A second point of interest is the distance between an occurrence of W at position x_0 and the next occurrence at, say, $x_0 + Y$.

Theorem 4 *In the uniform iid model (M00), the probabilities $q(y) = \Pr\{Y = y\}$ are given by*

$$\begin{array}{l} \textit{if } 0 < y < k \textit{ then} \\ \qquad q(y) = N^{-y}\varepsilon(k-y) - \sum_{z=1}^{y-1} q(z)\varepsilon(k-y+z)N^{z-y} , \\ \textit{if } y \geq k \textit{ then} \\ \qquad q(y) = N^{-k} - N^{-k}\sum_{z=1}^{y-k} q(z) - \sum_{z=y-k+1}^{y-1} q(z)\varepsilon(k-y+z)N^{z-y} . \end{array} \tag{4.6}$$

Proof. For $y \geq k$, the recurrence is the same as (4.1). For $y < k$, we use the same decomposition but here $\Pr\{W \text{ at } x_0 + y \mid W \text{ at } x_0\}$ is not N^{-k} but $N^{-y}\varepsilon(k-y)$ since the occurrence overlaps necessarily with the preceding; the other term is analogous to the case $y \geq k$ but the first sum is empty. ■

Theorem 5 *In the M1 model, the distribution of the distance Y between two successive occurrences is given by the following recurrence:*

$$
\begin{array}{l}
\textit{if } 0 < y < k \textit{ then} \\
\quad q(y) = \varepsilon(k-y)\tau(k-y+1,k) - \sum_{z=1}^{y-1} q(z)\varepsilon(k-y+z)\tau(k-y+z+1,k) \ , \\
\textit{if } y \geq k \textit{ then} \\
\quad q(y) = \pi^{y-k+1}(w_k,w_1)\tau(2,k) - \sum_{z=1}^{y-k} q(z)\pi^{y-z-k+1}(w_k,w_1)\tau(2,k) \\
\qquad\qquad - \sum_{z=y-k+1}^{y-1} q(z)\varepsilon(k-y+z)\tau(k-y+z+1,k) \ .
\end{array}
\tag{4.7}
$$

Proof. The result is obtained in the same way as (4.6). ■

We remark that, even when $\varepsilon(k-y)=1$, $q(y)$ can be null: some distances corresponding to possible overlapping patterns are impossible. This will be seen again in section **6.1**.

4.3.2 Generating function

The generating function ϕ_Y in M00 is given in Blom and Thorburn (1982) (corollary 2.3) by remarking that, to observe W, one has to observe its greatest overlapping sub-word V and then wait for the completing letters, so that $X_W = X_V + Y$.

This result can also be derived directly from (4.6) by integrating the series $\sum_{y\geq 1} q(y)t^y$. This leads to

$$
\phi_Y(t) = 1 - \frac{(1-t)\left(\frac{N}{t}\right)^k}{1+(1-t)\delta(\frac{1}{t})} \ . \tag{4.8}
$$

Here again, we can get the first two moments of Y :

$$
\begin{array}{rcl}
\mathbb{E}_{\mathrm{M00}}(Y) & = & N^k \ , \\
\mathbb{V}_{\mathrm{M00}}(Y) & = & N^k\left[2\delta(1)-(2k-1)\right] - N^{2k} \ .
\end{array}
$$

The mean distance between two successive occurrences depends only on the length of the word.

This is not true for the variance. $\mathbb{V}_{\mathrm{M00}}(Y)$ can also be written as the sum of $N^k\left[N^k+1-2k\right]$ and $N^k\delta_V(1)$ (where $\delta_V(t)=\delta_W(t)-(Nt)^k$): the first term is common to all words and the second term is null for non-overlapping words. The general notion is that all words with length k are equally frequent but overlapping words are more irregular and occurs in "trains".

Theorem 6 *In the M1 model, the generating function of Y is*

$$
\phi_Y(t) = 1 - \frac{(1-t)\mu(W)^{-1}t^{-k}}{\gamma_{k,1}(t)+(1-t)\delta\left(\frac{1}{t}\right)} \tag{4.9}
$$

where δ and $\gamma_{u,v}$ are defined as in (4.4).

Proof. This generating function is obtained in the same way as the generating function ϕ_X given in (4.4). ■

Note that this result is consistent with the M00 case (4.8).

4.3.3 Moments

The moments of Y are derived from (4.9) using (4.5):

$$\begin{aligned}
\mathbb{E}_{\mathrm{M1}}(Y) &= \frac{1}{\mu(W)}\,, \\
\mathbb{V}_{\mathrm{M1}}(Y) &= \frac{1}{\mu(W)}\left\{2\left[\delta(1)-\gamma'_{k,1}(1)\right]-(2k-1)\right\}-\left(\frac{1}{\mu(W)}\right)^2 .
\end{aligned}$$

4.4 Distribution of the n-th occurrence

4.4.1 Recurrence

Theorem 7 *In the M00 model, the distribution of the position X_n of the n-th occurrence is given by*

$$p(x,n) = N^{-k} - N^{-k}\sum_{z=1}^{x-k} p(z,n) - \sum_{z=x-k+1}^{x-1} p(z,n)\varepsilon(k-x+z)N^{z-x} - \sum_{m=1}^{n-1} p(x,m) \quad (4.10)$$

for $x \geq k$ and $p(x,n)=0$ for $x < k+n-1$, $p(k,1)=N^{-k}$ (with convention $p(\cdot,0)=0$).

Proof. The principle is the same as in (4.1) and (4.6): N^{-k} is the probability that W occurs at position x, $p(x,n)$ is the probability that it is the n-th occurrence, the first two sums scans the cases where the n-th has already occurred in z ($z \leq x-k$ or $x-k < z < x$) and the last term sums all the probabilities that this occurrence is the m-th with $m < n$. ■

Theorem 8 *In the M1 model, the distribution of the position X_n of the n-th occurrence is given by*

$$\begin{aligned}
p(x,n) = \mu(W) - \sum_{z=1}^{x-k} p(z,n)\pi^{x-z-k+1}(w_k,w_1)\tau(2,k) \quad (4.11) \\
-\sum_{z=x-k+1}^{x-1} p(z,n)\varepsilon(k-x+z)\tau(x+z+1,k) - \sum_{m=1}^{n-1} p(x,m)
\end{aligned}$$

for $x \geq k$ and $p(x,n)=0$ for $x < k+n-1$, $p(x,1)=p(x)$ (with convention $p(x,\cdot)=0$ for $x \leq 0$).

Proof. The principle is the same as in (4.10). ■

4.4.2 Generating function

Corollary 9 *The generating function of X_n is*

$$\phi_{X_n}(t) = \phi_X(t) \times \left[\phi_Y(t)\right]^{n-1} .$$

Proof. This comes easily from the fact that $X_n = X + \sum_{m=1}^{n-1} Y_m$ and that $\{X, Y_1, ...Y_{n-1}\}$ are independent in M00 as in M1. ■

As a by-product of this, it follows that, in any model,

$$\mathbb{E}(X_n) = \mathbb{E}(X) + (n-1)\mathbb{E}(Y) , \quad \mathbb{V}(X_n) = \mathbb{V}(X) + (n-1)\mathbb{V}(Y) .$$

4.5 Distribution of the distance between the n-th and the $(n+r)$-th occurrence

4.5.1 Recurrence

Theorem 10 *In the M00 model, the probability $q(y,r) = \Pr\{Y_r = y\}$ is given by the following recurrence:*

$$q(y,r) = N^{-k} - N^{-k} \sum_{z=1}^{y-k} q(z,r) - \sum_{z=y-k+1}^{y-1} q(z,r)\varepsilon(k-y+z)N^{z-y} - \sum_{m=1}^{r-1} q(y,m) \quad (4.12)$$

for $y \geq k$ and $q(y,r) = 0$ for $y < r$, $q(y,1) = q(y)$ (with convention $q(y,\cdot) = 0$ for $y \leq 0$).

Theorem 11 *In the M1 model, the distribution of Y_r is given by the following recurrence:*

$$\begin{aligned} q(y,r) = {} & \pi^{y-k+1}(w_k, w_1)\tau(2,k) - \sum_{z=1}^{y-k} q(z,r)\pi^{y-z-k+1}(w_k,w_1)\tau(2,k) \\ & - \sum_{z=y-k+1}^{y-1} q(z,r)\varepsilon(k-y+z)\tau(k-y+z+1,k) - \sum_{m=1}^{r-1} q(y,m) \end{aligned} \quad (4.13)$$

for $y \geq k$ and $q(y,r) = 0$ for $y < r$, $q(y,1) = q(y)$ (with convention $q(y,\cdot) = 0$ for $y \leq 0$).

4.5.2 Generating function

Corollary 12 *For both M00, M0 and M1 models, the generating function of Y_r is*

$$\phi_{Y_r}(t) = [\phi_Y(t)]^r \ .$$

Proof. The distances Y are iid and Y_r is the sum of r of them. ■

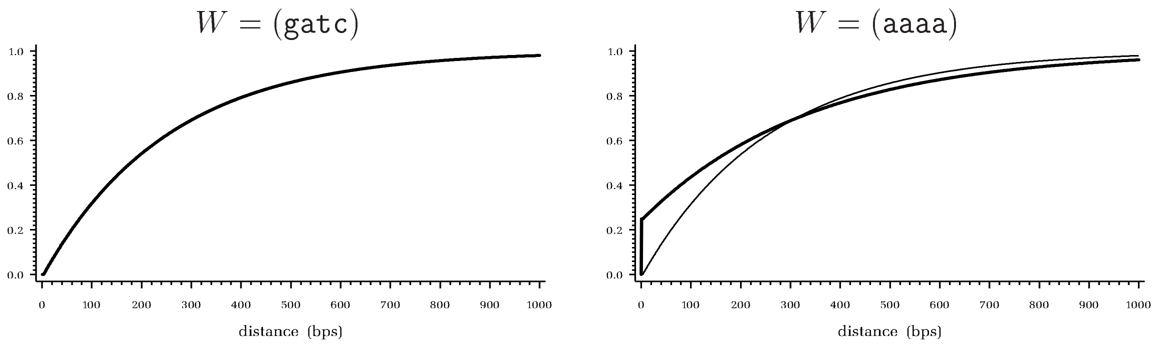
It is straightforward that $\mathbb{E}(Y_r) = r\mathbb{E}(Y)$ and $\mathbb{V}(Y_r) = r\mathbb{V}(Y)$.

4.6 Application to genome analysis

We present here two applications of the above results to genome analysis. Here we are interested in the distribution of words so we use results concerning the distances Y and Y_r. When the sequence **S** in a part of the genome we have $N = 4$ and $\mathcal{A} = \{\mathtt{a}, \mathtt{g}, \mathtt{c}, \mathtt{t}\}$.

4.6.1 Quality of the geometric approximation

The geometric distribution is very widely used to study the distribution of a "word" along the genome (see for example Henaut *et al.* (1996)). As shown above, this distribution is not the exact one but the error made when using it depends greatly on the word. Figure **4.1** shows the difference between the exact and the geometric distributions for a non-overlapping word: $W = (\texttt{gatc})$ (studied in Henaut *et al.* (1996)) and a very overlapping one: $W = (\texttt{aaaa})$ under the M00 model.



(The two curves coincide.)

Figure 4.1: Comparison of the exact (fat line) and geometric (thin line) pdf's under the M00 model for $W = (\texttt{gatc})$ and $W = (\texttt{aaaa})$.

For both words we have $\mu(W) = \frac{1}{256}$ which is the parameter p for the geometric approximation: $\Pr\{Y = y\} \simeq (1-p)^{y-k}p$ for $y \geq k$. As it can be seen from Figure **4.1**, the approximation is quite satisfying for $(\texttt{gatc})$ but very bad for $(\texttt{aaaa})$: the distance in total variation between the exact and the approximate distribution is 65 times greater for $(\texttt{aaaa})$ ($d_{TV} = 0.28$) than for $(\texttt{gatc})$ ($d_{TV} = 0.0043$). These results are general and are due to the fact that the overlapping structure is not taken into account in the geometric distribution.

Note that, despite the fact $W = (\texttt{aaaa})$ can overlap with 1 or 2 letters, we have $\Pr\{Y = y\} = 0$ for $y = 3$ and 2. The reason for this is that, for example, to get a distance $y = 2$, one needs six successive 'a ', so that a distance $y = 1$ is necessarily observed first. This can be related to the periodical structure of ε studied in Schbath (1995).

4.6.2 Comparison with the empirical distribution

We want now to study the distribution of $W = (\texttt{gatc})$ in the sequence " *ecomori*" which is a part of the genome of Escherichia coli. The length of this sequence is $l = 111,402$ base pairs (e.g. letters) and it contains 495 occurrences of W. We use a model M1 (adjusted to the sequence) with transition matrix

$$\mathbf{\Pi} = \begin{pmatrix} 0.2927 & 0.2182 & 0.2219 & 0.2672 \\ 0.2208 & 0.2428 & 0.3223 & 0.2141 \\ 0.2558 & 0.3201 & 0.2228 & 0.2014 \\ 0.1695 & 0.3157 & 0.2309 & 0.2838 \end{pmatrix}.$$

This model gives $\mu(W) = 0.003734, \mathbb{E}(Y) = 267.8, \mathbb{V}(Y) = (264.5)^2$ so the expected number of occurrences under M1 is $l/\mathbb{E}(Y) = 415.9$. Figure **4.2** shows that short distances are more frequent in the *ecomori* sequence than expected. For example, the median observed distance is $y = 135$ while the theoretical median is $y = 188$. Based on this we conclude that (`gatc`) is unexpectedly frequent in this sequence.

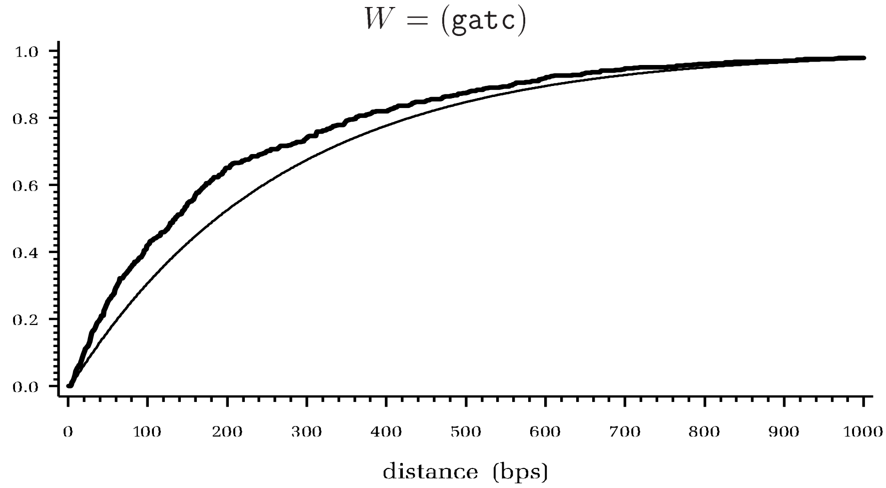


Figure 4.2: Comparison of the empiric pdf in "*ecomori*" (fat line) and theoretical pdf under M1 (thin line) for $W =$ (`gatc`).

4.6.3 Counting statistics

Another way to detect words with unexpected frequencies is to count the occurrences of W. The recurrences (4.10) and (4.11) give the exact distributions of the count $N(l)$ of word W in sequence $\mathbf{S}$ with length l since $\Pr\{N(l) < n\} = \Pr\{X_n > l\}$ respectively in M00 and M1 models. One may use this result to assess the probability of obtaining n words in a sequence of length l by computing $\Pr\{N(l) \leq n\} = \Pr\{X_n \geq l\}$ for a rare word or $\Pr\{N(l) \geq n\} = \Pr\{X_n \leq l\}$ for a frequent word.

For example, we get $n = 48$ occurrences of $W =$ (`gatc`) in the first $l = 10,000$ base pairs of "*ecomori*". Using (4.11), we get $\Pr\{X_{48} \leq 10,000\} = 5.00\%$ for the M1 model. So, in the first $10,000$ base pairs of "*ecomori*", (`gatc`) is unexpectedly frequent.

Using the asymptotic normality of X_n

$$\Pr\{X_n \leq l\} \simeq \Pr\left\{\mathcal{N}(0,1) \leq [l - n\mathbb{E}(Y)]\,[n\mathbb{V}(Y)]^{-1/2}\right\},$$

we get $\Pr\{X_{48} \leq 10,000\} \simeq \Pr\{\mathcal{N}(0,1) \leq -1.559\} = 5.95\%$.

4.7 Distance between two words

The results of section **3** may be easily adapted to the case of the distance Z between two words V and $W = (w_1 ... w_k)$ in this order. We assume that neither V is included in W

nor W in V. Let $W^* = (w_1 ... w_{k^*})$ be the largest overlapping sub-word between V and W: V ends with W^* and W starts with W^*.

As it is proved by Blom and Thorburn (1982) in the M00 case, in expression (4.6) we only have to replace the first $\varepsilon(k-y)$ with $\varepsilon^*(k-y)$ where $\varepsilon^*(u)$ is the overlapping indicator of W upon W^* with u common letters. The other $\varepsilon(u)$'s appearing in (4.6) are the usual $\varepsilon(u)$ describing the self-overlapping structure of W. Under the model M1, we have

$$\delta^*(t) = \sum_{u=1}^{k^*} \varepsilon^*(u) \frac{t^u}{\mu(w_1)\tau(2,u)} , \qquad \delta(t) = \sum_{u=1}^{k} \varepsilon(u) \frac{t^u}{\mu(w_1)\tau(2,u)} .$$

We obtain

$$\phi_Z(t) = \frac{\gamma_{k^*,1}(t) + (1-t)\delta^*(t^{-1})}{\gamma_{k,1}(t) + (1-t)\delta(t^{-1})} ,$$

therefore

$$\begin{aligned}
\mathbb{E}_{\mathrm{M1}}(Z) &= \delta(1) - \delta^*(1) , \\
\mathbb{V}_{\mathrm{M1}}(Z) &= [\mu(W)]^{-1} \left\{ 2\left[\delta(1) - \gamma'_{k,1}(1)\right] - (2k-1) - [\mu(W)]^{-1} \right\} \\
&\quad + [\mu(W^*)]^{-1} \left\{ 2\left[\delta^*(1) - \gamma'_{k^*,1}(1)\right] - (2k'-1) - [\mu(W^*)]^{-1} \right\} .
\end{aligned}$$

Since Z is the position of the first occurrence of W after V, we denote Z_n the position of the n-th occurrence: $Z_n = Z + Y_{n-1}$ (with Y_r defined as in introduction for the word W). We may obtain the pdf, the pgf and the moments of Z_n in the same way.

Bibliography

AKI, S. and HIRANO, K. (1993). Discrete distributions related to succession events in a two state markov chain. In *Statistical Sciences and Data Analysis*, (K. Matusita, M. L. Puri, and T. Hayakawa, ed.), 467–474. VSP Publishers, Amsterdam.

BLOM, G. (1982). On the mean number of random digits until a given sequence occurs. *J. Appl. Prob.* **19** 136–143.

BLOM, G. and THORBURN, D. (1982). How many random digits are required until given sequences are obtained ? *J. Appl. Prob.* **19** 518–531.

CHRYSSAPHINOU, O. and PAPASTAVRIDIS, S. (1990). The occurrence of sequence patterns in repeated dependent experiment. *Theory Prob. Appl.* **35 (1)** 145–152.

COWAN, R. (1991). Expected frequencies of DNA patterns using Whittle's formula. *J. Appl. Prob.* **28** 886–892.

FELLER, W. (1968). *An Introduction to Probability Theory and its Applications.* Wiley, New York.

GESKE, M. X., GODBOLE, A. P., SCHAFFNER, A. A., SKOLNICK, A. M. and WALLSTROM, G. L. (1995). Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Prob.* **32** 877–892.

GODBOLE, A. P. (1991). Poisson approximations for runs and patterns of rare events. *Adv. Appl. Prob.* **23** 851–865.

HENAUT, A., ROUXEL, T., A., G. and MOSZER, I. (1996). Uneven distribution of gatc motifs in the *escherichia coli* chromosome, its plasmids and its phageson number of occurrences of sucess runs of specified length in a two-state Markov chain. *J. Mol. Biol.* **257** 574–585.

JOHNSON, N. L., KOTZ, S. and KEMP, A. W. (1992). *Univariate Discrete Distributions.* Wiley: New-York.

KARLIN, S. and MACKEN, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Amer. Statist. Assoc.* **86** 27–35.

KOUTRAS, M. (1997). Waiting times and number of appearances of events in a sequence of discrete random variables. In *Advances in Combinatorial Methods and Applications to Probability and Statistics*, (N. Balakrishnan, ed.), 253–261, Statistics and Industry and Technology Series. Birkhauser, Boston.

MORI, T. (1989). On the number of different patterns preceding a given one. *Studia. Sci. Math. Hung.* **24** 355–364.

PEKOZ, E. A. (1996). Stein's method for geometric approximation. *J. Appl. Prob.* **33** 707–713.

PHILIPPOU, A. N., GEORGHIOU, C. and PHILIPPOU, G. N. (1983). A generalized geometric distribution and some of its properties. *Statist. Prob. Letters.* **1** 171–175.

PRUM, B., RODOLPHE, F. and DE TURCKHEIM, E. (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* **57** 205–220.

RUDANDER, J. (1996). *On the first occurrence of a given pattern in a semi-Markov process.* PhD thesis, University of Uppsala.

SCHBATH, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics.* **1** 1–16. (http://www.emath.fr/ps/).

SCHWAGER. (1983). Run probabilities in sequences of markov-dependent trials. *J. Amer. Statist. Assoc.* **78** 168–175.

UCHIDA, M. and AKI, S. (1995). Sooner or later waiting time problems in a two-state markov chain. *Ann. Inst. Statist. Math.* **47** 415–433.

Appendix

4.A Computational simplifications

Some recurrences given in the paper involve a number of terms too large to be computed, but all of them can be simplified and reduced. We give here efficient simplifications which give computational times that are linear as a function of x (or y) instead of quadratic, as in Blom and Thorburn (1982). These simplifications are useful for large distances: for small x or y, one may use the exact formulas given in the theorems of the paper.

4.A.1 Position of the first occurrence

M00 model: The recurrence (4.1) uses x terms. Calculating $p(x)-p(x-1)$, we get

$$p(x) = p(x-1) - N^{-k}p(x-k) - \sum_{z=x-k+1}^{x-1} \left[p(z)-p(z-1)\right] \varepsilon(k-x+z)N^{z-x}$$

which uses only k terms.

M1 model: Formula (4.7) leads to very heavy computations and, for example, calculating $p(y)-p(y-1)$ does not reduce it. Yet we can use the convergence of $\pi^n(a,b)$ towards $\mu(b)$ and make an approximation for α large enough which gives for $x>\alpha>k$

$$\begin{aligned} p(x) \;\simeq\; & p(x-1) - p(x-\alpha)\mu(W) \\ & + \sum_{z=x-k+1}^{x-1} \left[p(z-1)-p(z)\right] \varepsilon(k-x+z)\tau(k-x+z+1,k) \\ & + \tau(2,k) \left[\sum_{z=x-\alpha}^{x-k-1} p(z)\pi^{x-z-k}(w_k,w_1) - \sum_{z=x-\alpha+1}^{x-k} p(z)\pi^{x-z-k+1}(w_k,w_1) \right] \end{aligned}$$

with the convention $p(x)=0$ for $y \leq 0$. This reduces greatly the computational time.

The errors made by the approximation $\pi^n(a,b)$ by $\mu(b)$ must be very tiny (e.g. $<10^{-10}$) because they are cumulated in the recurrence. In practice, $\alpha=15$ is sufficient to obtain a precision of 10^{-15} in the genome case, where $N=4$.

Distance between two successive occurrences: Under M00, M0 and M1, the recurrences for $q(y)$ are the same as for $p(x)$ as soon as $y>k$ so the preceding simplifications holds also.

4.A.2 Position of the n-th occurrence

M00 model: Calculating $p(x,n) - p(x,n-1)$ gives a recurrence on n for $n \geq 2$ (using $p(x,1) = p(x)$)

$$
\begin{aligned}
p(x,n) &= N^{-k} \sum_{z=1}^{x-k} [p(z,n-1) - p(z,n)] \\
&\quad + \sum_{z=x-k+1}^{x-1} [p(z,n-1) - p(z,n)]\, \varepsilon(k-x+z) N^{z-x}
\end{aligned}
$$

and calculating $p(x,n) - p(x-1,n)$ gives another recurrence on x

$$
\begin{aligned}
p(x,n) &= p(x-1,n) + N^{-k} [p(x-k,n-1) - p(x-k,n)] \\
&\quad + \sum_{z=x-k+1}^{x-1} \{[p(z,n-1) + p(z-1,n) - p(z,n) - p(z-1,n-1)] \\
&\quad \times \varepsilon(k-x+z) N^{z-x}\}
\end{aligned}
$$

which reduces greatly the number of terms to be computed.

M1 model: $p(x,n) - p(x,n-1)$ gives

$$
\begin{aligned}
p(x,n) &= \tau(2,k) \sum_{z=1}^{x-k} [p(z,n-1) - p(z,n)]\, \pi^{x-z-k+1}(w_k, w_1) \\
&\quad + \sum_{z=x-k+1}^{x-1} [p(z,n-1) - p(z,n)]\, \varepsilon(k-x+z) \tau(k-x+z+1,k)\ .
\end{aligned}
$$

Then, calculating $p(x,n) - p(x-1,n)$ and using the convergence of $\pi^n(a,b)$ towards $\mu(b)$, we get, for α large enough and $y > \alpha$,

$$
\begin{aligned}
p(x,n) &\simeq p(x-1,n) + \tau(2,k)\mu(w_1) [p(x-\alpha,n-1) - p(x-\alpha,n)] \\
&\quad + \sum_{z=x-\alpha+1}^{x-k} \{[p(z,n-1) + p(z-1,n) - p(z,n) - p(z-1,n-1)] \\
&\quad\quad \times \pi^{x-z-k+1}(w_k, w_1) \tau(2,k)\} \\
&\quad + \sum_{z=x-k+1}^{x-1} \{[p(z,n-1) + p(z-1,n) - p(z,n) - p(z-1,n-1)] \\
&\quad\quad \times \varepsilon(k-x+z) \tau(x+z+1,k)\}
\end{aligned}
$$

which only uses α terms.

Distance between the n-th and the $(n+r)$-th occurrences: Here again, the recurrence giving $q(y,r)$ is the same as $p(x,r)$ for y large enough so the preceding formula holds.

Chapitre 5

Exact distribution of the distances between any occurrences of a set of words

S. Robin, J.-J. Daudin (2001)

Annals of the Institute of Mathematical Statistics, **53**, (4), 895–905

Abstract. The distribution of the distance between two (or more) successive occurrences of a specific word in a random sequence of letters is known under different models. In this paper, a more general problem is studied: the distribution of the distance between two (or more) successive occurrences of any word of a given set under a Markov model for the sequence. The generating function and a recurrence for obtaining the probabilities are given. These results are applied to study the distribution of the "CHI" motif in the genome sequence of *Haemophilus influenzae.*

Keywords. distance between occurrences, genome sequence analysis, semi Markov process.

Contents

5.1 Introduction

Let $S = \{S_1 S_2, \ldots, \}$ be a sequence of repeated trials with possible outcomes taken from an alphabet $\mathcal{A}$. Let w be a specific string of letters of $\mathcal{A}$, called a "word" or a "pattern". The exact distribution of the waiting time until w appears has been studied by many authors, often for $\mathcal{A}$ =$\{0,1\}$. Among recent references, Robin and Daudin (1999) have given the generating probability function and a recurrence relation for computing the probabilities under a first order Markovian model for S. A more complete source of references may be obtained in Koutras (1997) who gives the relation between the generating function of the waiting time and the generating function of the number of occurrences of w.

What happens when we consider more than one word?

Let $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ be a set of words of respective lengths $\{k_i\}_{i=1..m}$ not included in each other. In this paper, we study the waiting time until the first word of $\mathcal{W}$ appears. This problems occurs in some applications:

- In genome sequence analysis where the same biological function is fulfilled by different words. For example, a particular pattern of nucleotide called CHI (for *Crossover Hotspot Initiator*) possess a function of protection of the genome and is characteristic of each species. It is unknown for recently sequenced species but is expected to be frequent and regularly spaced. For some organism, the same function is fulfilled by few different words: for example, the CHI of *Haemophilus influenza* is (`gNtggtgg`) where `N` can be any of the four letters of the alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. The knowledge of the exact distribution of the distance between the occurrences of several words would be of great help to study the longitudinal distribution of such a motif.

- A particular case of this problem is *the sooner and later waiting time problem* for success and failure runs where $\mathcal{A}$ =$\{0,1\}$ and $\mathcal{W} = \{1, \ldots, 1,\ 0, \ldots, 0\}$ with k_1 '1' and k_0 '0'. This problem is found in reliability and in psychology.

The general problem has been studied by Breen *et al.* (1985) using renewal theory and only in the non-overlapping and independent case and by Chryssaphinou and Papastavridis (1990) in the overlapping and first order Markovian case. Mori (1991) gives a limit theorem for the waiting time till each of a given set of patterns of same length in a sequence of iid random variables distributed uniformly on any alphabet.

The *sooner and later waiting time problem* has been recently studied by Aki and Hirano (1993) and Uchida and Aki (1995) under a first order Markovian model, by Aki and Balakrishnan (1996) under a markovian model of order two. Koutras and Alexandrou (1997) have generalized the problem with a three letters alphabet under a general non homogeneous model.

The finite Markov chain imbedding technique provides a quite general method to calculate the distribution of the first occurrence of different motifs (Fu and Koutras (94) or Koutras and Alexandrou (1997)) or of the counting of runs (Fu (1996)). However, for complex motifs, especially for overlapping words that are not simple runs, the transition matrix is not straightforward to obtain. Furthermore this matrix has dimensions proportional either to the total length of the motifs or to the length of the sequence. This point seems to make the method untractable for large scaled problems such as DNA analysis.

In this paper we give the exact distribution and the generating function of the waiting time in the case of any set of words with any alphabet and a first order Markov model.

We first present the model and the notations. In the second part, we derive the probability generating function. The third section is devoted to the law of the $r-$scans and the fourth presents an application to the CHI of *Haemophilus influenza.*

Set of words. Let us study the occurrences of a set of words $\mathcal{W} = \{\mathbf{w}_i\}_{i=1..m}$ not included in each other and of respective lengths $\{k_i\}$. $w_{i,u}$ denotes the $u-$th letter of $\mathbf{w}_i : \mathbf{w}_i = (w_{i,1}, , \ldots ,, w_{i,k_i})$. The position of a word in the sequence is the position of its last letter.

Model for the sequence. Let us consider a sequence $\{S_x\}_{x\geq 1}$ of letters taken from an alphabet $\mathcal{A}$. The sequence is assumed to be a homogeneous first order Markov chain (MC1) with transition probability $\mathbf{\Pi}$ with general term $\pi(a, b)$

$$\forall (a, b) \in \mathcal{A} \times \mathcal{A}, \forall x \geq 1, \quad \Pr\{S_{x+1} = b \,|\, S_x = a\} = \pi(a, b).$$

In the following, $\pi^{(n)}(a, b)$ shall denote the general term of $\mathbf{\Pi}^n$ and $\mu = [\mu(a)]_{a\in\mathcal{A}}$ shall denote the stationary distribution of $\mathbf{\Pi}$ which satisfies $\mu.\mathbf{\Pi} = \mu$.

Notations. For any words $\mathbf{w}_i$ and $\mathbf{w}_j$ of $\mathcal{W}$, let us use the following notations.

- $\varepsilon_{ij}(u)$ is the overlapping indicator of $\mathbf{w}_j$ over $\mathbf{w}_i$ with u letters which equals one when the first u letters of $\mathbf{w}_j$ are the same as the last u letters of $\mathbf{w}_i$: $\varepsilon_{ij}(u) = \mathbb{I}\{(w_{i,k_i-u+1}, \ldots , w_{i,k_i}) = (w_{j,1}, \ldots , w_{j,u})\}$, where $\mathbb{I}\{A\}$ equals one if A is true and zero otherwise. The condition of non-inclusion of the words implies that $\varepsilon_{ij}(u) = 0$ for $u = \min(k_i, k_j)$ if $i \neq j$ and for $u > \min(k_i, k_j)$ in any case.

- $\tau_i(u, v)$ is the probability to observe $w_{i,u}$ to $w_{i,v}$ given $w_{i,u-1}$:

$$\tau_i(u, v) = \prod_{x=u}^{v} \pi(w_{i,x-1}, w_{i,x})$$

  with convention $\tau_i(u, 1) = 1$.

- $d_{ij}(t)$ is the overlapping polynomial which does not take into account complete overlap:

$$d_{ij}(t) = \sum_{u=1}^{\min(k_i,k_j)-1} \frac{\varepsilon_{ij}(u)t^u}{\tau_j(2, u)}.$$

5.2 Semi-Markov model for the occurrences of the words

5.2.1 Model and notations

The way the words occur along the sequence is completely described by two processes

- the positions of the occurrences of the different words of $\mathcal{W} : \{X_n\}_{n\geq 1}$;

- the words occurring at each of these positions $\{I_n\}_{n\geq 1}$:

$$\{I_n = i\} \quad \Longleftrightarrow \quad \{\text{the word occuring at } X_n \text{ is } \mathbf{w}_i\}.$$

This describes a semi-Markov process with states in $\mathcal{W}$. This process is assumed to be homogenous along the sequence.

Let $p_{ij}(y)$ denote the probability that the first word of $\mathcal{W}$ occurring after $\mathbf{w}_i$ is $\mathbf{w}_j$ and that it appears y positions after:

$$p_{ij}(y) = \Pr\{(X_{n+1} - X_n = y) \cap (I_{n+1} = j) \mid I_n = i\}.$$

$p_{ij}(y)$ does not depend of n thanks to the Markovian structure of the sequence. Let ϕ_{ij} be the generating function of the $p_{ij}(y)$'s:

$$\phi_{ij}(t) = \sum_{y\geq 1} p_{ij}(y)t^y.$$

Let $q_{ij}(y)$ denote the probability that the word $\mathbf{w}_j$ occurs y positions after $\mathbf{w}_i$:

$$q_{ij}(y) = \sum_{r\geq 1} \Pr\{(X_{n+r} - X_n = y) \cap (I_{n+r} = j) \mid I_n = i\}.$$

The difference between $p_{ij}(y)$ and $q_{ij}(y)$ is that $q_{ij}(y)$ concerns any occurrence of $\mathbf{w}_j$ after any number of renewals of any word of $\mathcal{W}$, while $p_{ij}(y)$ concerns only the first occurrence of $\mathbf{w}_j$ among the set $\mathcal{W}$. Let $f_{ij}(t)$ be the generating function of the $q_{ij}(y)$'s:

$$f_{ij}(t) = \sum_{y\geq 1} q_{ij}(y)t^y.$$

If the first word is distributed according to some distribution ν over $\mathcal{W}$, we shall denote $p_{\nu j}(y) = \sum_i \Pr\{(X_{n+1} - X_n = y) \cap (I_{n+1} = j) \mid I_n = i\}\nu(i)$ where $\nu(i) = \Pr\{I_n = i\}$. We shall denote $\phi_{\nu j}(t)$ the corresponding generating function.

At last, we shall denote $p_{i\bullet}(y)$ the probability that the first occurrence after $\mathbf{w}_i$ of any word of $\mathcal{W}$ occurs y position later : $p_{i\bullet}(y) = \sum_j p_{ij}(y)$, and $\phi_{i\bullet}(t)$ the corresponding probability generating function.

5.2.2 Matrices of generating functions

Lemma 1 *The probabilities $p_{ij}(y)$'s and $q_{ij}(y)$'s satisfy the following relation*

$$q_{ij}(y) = p_{ij}(y) + \sum_{l=1}^{m} \sum_{z=1}^{y-1} q_{il}(z)p_{lj}(y-z). \tag{5.1}$$

Proof. The first term of the right hand side gives the probability that the first word of $\mathcal{W}$ occurring after $\mathbf{w}_i$ is $\mathbf{w}_j$ and that it appears y positions after. The second term calculates the probability of $\mathbf{w}_j$ y position after $\mathbf{w}_i$ is not the first one by conditioning on all possible occurrences of any word of $\mathcal{W}$ between positions 1 and $y-1$. ■

Lemma 2 *The probability $q_{ij}(y)$ is*

$$\begin{aligned} q_{ij}(y) = \; & \varepsilon_{ij}(k_j - y)\tau_j(k_j - y + 1, k_j)\mathbb{I}(y < k_j) \\ & + \pi^{(y-k_j+1)}(w_{i,k_i}, w_{j,1})\tau_j(2, k_j)\mathbb{I}(y \geq k_j). \end{aligned} \tag{5.2}$$

Proof. The first term of the right hand side considers the case where $\mathbf{w}_j$ overlaps $\mathbf{w}_i$ $(y < k_j)$ and the second term considers the case where it does not $(y \geq k_j)$. ∎

Lemma 3 *The generating function $f_{ij}(t) = \sum_{y \geq 1} q_{ij}(y) t^y$ is*

$$f_{ij}(t) = t^{k_j}\tau_j(2, k_j)\left[d_{ij}\left(\frac{1}{t}\right) + \frac{g_{i,j}(t)}{t}\right]$$

where $g_{ij}(t) = \sum\limits_{u \geq 1} \pi^{(u)}(w_{i,k_i}, w_{j,1}) t^u$.

Proof. Multiplying the left side of equation (5.2) by t^y and summing over all positive y's, we get for $f_{ij}(t)$:

$\sum\limits_{y=1}^{k_j-1} \varepsilon_{ij}(k_j - y)\tau_j(k_j - y + 1, k_j)t^y + \sum\limits_{y \geq k_j} \pi^{(y-k_j+1)}(w_{i,k_i}, w_{j,1})\tau_j(2, k_j)t^y$

$= t^{k_j} \sum\limits_{z=1}^{k_j-1} \varepsilon_{ij}(z)\tau_j(z + 1, k_j)t^{-z} + \tau_j(2, k_j)t^{k_j-1} \sum\limits_{u \geq 1} \pi^{(u)}(w_{i,k_i}, w_{j,1})t^u$

$= t^{k_j}\tau_j(2, k_j)\left[\sum\limits_{z=1}^{k_j-1} \frac{\varepsilon_{ij}(z)}{\tau_j(2, z)} t^{-z} + \frac{1}{t}\sum\limits_{u \geq 1} \pi^{(u)}(w_{i,k_i}, w_{j,1})t^u\right]$ and the lemma is proved since $\varepsilon_{ij}(u) = 0$ for $u > \min(k_i, k_j)$. ∎

If we consider the $(m \times m)$ matrices $\mathbf{F}(\mathbf{t})$, $\mathbf{D}(t)$ and $\mathbf{G}(t)$ with respective general terms $f_{ij}(t)$, $d_{ij}\,(t)$ and $g_{ij}(t)$, and the $m-$diagonal matrix $\mathbf{T}(t)$ with general term $\tau_j(2, k_j)t^{k_j}$ $(j = 1..m)$, we obtain the following matrix decomposition :

$$\mathbf{F}(t) = \left[\mathbf{D}\left(\frac{1}{t}\right) + \frac{1}{t}\mathbf{G}(t)\right]\mathbf{T}(t).$$

Considering that $\mathbf{H}(t) = \left[(\mathbf{I} - \mathbf{\Pi}t)^{-1} - \mathbf{I}\right]$ is the generating matrix of the transition probabilities between any couple of letters of $\mathcal{A}$, $\mathbf{G}(t)$ is the $m \times m$ generating matrix of the transition probabilities between the last letter of each word of $\mathcal{W}$ and each first letter: $g_{ij}(t) = [\mathbf{H}(t)]_{w_{i,k_i}, w_{j,1}}$.

Theorem *The generating matrices $\mathbf{\Phi}(t) = [\phi_{ij}(t)]_{i,j=1..m}$ and $\mathbf{F}(t)$ satisfy*

$$\mathbf{F}(t) = \mathbf{\Phi}(t)\left[\mathbf{I} + \mathbf{F}(t)\right]$$

where $\mathbf{I}$ is the $m \times m$ identity matrix.

Proof. Let us multiply each side of equation (5.1) by t^y and sum over all positive y's: the left term gives $f_{ij}(t)$, the right term gives $\phi_{ij}(t)+\sum_{l=1}^{m}\phi_{il}(t)f_{lj}(t)$ and the theorem is proved. ■

A consequence of this theorem is that, if t is such that $[\mathbf{I}+\mathbf{F}(t)]$ is invertible, one has

$$\mathbf{\Phi}(t)=\mathbf{F}(t).\left[\mathbf{I}+\mathbf{F}(t)\right]^{-1}. \tag{5.3}$$

It is hence of a great importance to know if $[\mathbf{I}+\mathbf{F}(t)]$ is invertible. Since $\mathbf{F}(t)$ is a matrix of rational functions, the determinant of $[\mathbf{I}+\mathbf{F}(t)]$ is also a rational function. Therefore it has a finite number of poles and roots and $[\mathbf{I}+\mathbf{F}(t)]$ is invertible for every t except a finite number. Note that $\mathbf{F}(0)=\mathbf{0}$ and $\mathbf{F}(1)$ is not defined.

Since $\mathbf{\Phi}(t)$ is a matrix of rational functions, we can consider that it is characterized by all its coefficients which may be calculated by the aid of (5.3) on any open interval containing no pole and no root. From now on, we shall consider that $\mathbf{\Phi}(t)$ is known.

The following corollary is a direct application.

Corollary 5 *The probability generating function $\phi_{i\bullet}(t)$ of the distance $Y_{i\bullet}$ between an occurrence of $\mathbf{w}_i$ and the next occurrence of any word of $\mathcal{W}$ is $\phi_{i\bullet}(t)=\phi_i(t).\mathbf{1}$ where $\phi_i(t)$ denotes the $i-th$ row of $\mathbf{\Phi}(t)$.*

Proof. $\phi_{i\bullet}(t)$ is defined by $\phi_{i\bullet}(t)=\sum_{y\geq 1}p_{i\bullet}(y)t^y$ where $p_{i\bullet}(y)$ is the probability that the next word of $\mathcal{W}$ occurs y letters after $\mathbf{w}_i$: $p_{i\bullet}(y)=\sum_{j=1}^{m}p_{ij}(y)=\Pr\{X_{n+1}-X_n=y\mid I_n=i\}$. It is clear that $\phi_{i\bullet}(t)=\sum_{j=1}^{m}\sum_{y\geq 1}p_{ij}(y)t^y=\sum_{j=1}^{m}\phi_{ij}(t)=\phi_i(t).\mathbf{1}$. ■

Corollary 6 *The generating function of the $p_{\nu j}(y)$ is $\phi_{\nu j}(t)=\nu.\mathbf{\Phi}_j(t)$ where $\mathbf{\Phi}_j(t)$ is the $j-th$ column of $\mathbf{\Phi}$.*

The generating function of the $p_{\nu\bullet}(y)$ is

$$\phi_{\nu\bullet}(t)=\nu.\mathbf{\Phi}(t).\mathbf{1}.$$

Proof. The first part of the corollary is obvious since $p_{\nu j}(y)=\sum_i \nu(i)p_{ij}(y)$. The second comes from $p_{\nu\bullet}(y)=\sum_j p_{\nu j}(y)$. ■

Sequence with a given beginning. Corollary **6** can be applied to the particular case where the sequence begins with one of the words of $\mathcal{W}$, say $\mathbf{w}_1$. In this case we shall consider the distribution $\nu=[1,0,\ldots,0]$.

If the sequence does not begin with a word of $\mathcal{W}$ but with some other word $\mathbf{w}_0$, we can apply the following result.

Corollary 7 *The generating vector $\phi_0(t)=[\phi_{01}(t),\ldots,\phi_{0m}(t)]$ satisfies*

$$\phi_0(t)=\mathbf{f}_0(t)\left[\mathbf{I}-\mathbf{\Phi}(t)\right].$$

Proof. For the generating functions, equation (5.1) is equivalent to $f_{ij}(t) = \phi_{ij}(t) + \sum_{l=1}^{m} f_{il}(t)\phi_{lj}(t)$. Here we consider $i = 0, \dots, m$ and $j = 1, \dots, m$ since $\mathbf{w}_0$ is only taken into account as a beginning word. Therefore, Theorem 4 can be extended as follows

$$\begin{bmatrix} \mathbf{f}_0(t) \\ \mathbf{F}(t) \end{bmatrix} = \begin{bmatrix} \phi_0(t) \\ \mathbf{\Phi}(t) \end{bmatrix} + \begin{bmatrix} \mathbf{f}_0(t) \\ \mathbf{F}(t) \end{bmatrix} \mathbf{\Phi}(t)$$

and the corollary is proved. ■

5.2.3 Probabilities and moments

Several elementary results can be derived from the preceding theorems and lemmas.

Transition probabilities between words. Since $\sum_{y\geq 1} p_{ij}(y) = \Pr\{I_{n+1} = j \mid I_n = i\}$, $\phi_{ij}(1)$ is the probability that $\mathbf{w}_j$ is the first word of $\mathcal{W}$ to occur after $\mathbf{w}_i$. It is therefore the transition probability from $\mathbf{w}_i$ to $\mathbf{w}_j$:

$$s(i, j) = \Pr\{I_{n+1} = j \mid I_n = i\}$$

and the transition matrix $\mathbf{S} = [s(i, j)]_{i,j=1..m}$ is $\mathbf{S} = \mathbf{\Phi}(1)$.

The stationary distribution of $\mathbf{S}$ can produce a particular distribution ν.

Moments of the distances. It is well known that the expectation and variance of a random variable with generating function ψ are respectively $\psi'(1)$ and $\psi''(1) + \psi'(1)[1 - \psi'(1)]$.

The generating function of the distance between an occurrence of $\mathbf{w}_i$ and the next occurrence of $\mathbf{w}_j$ given that $\mathbf{w}_j$ is the first word of $\mathcal{W}$ to occur after $\mathbf{w}_i$ is $\phi_{ij}(t)/s(i, j)$. These conditional expectation and variance are easy to derive from $\phi'_{ij}(t)$ and $\phi''_{ij}(t)$.

The moments of $Y_{\nu\bullet}$ can also be derived using Corollary **6**.

Computation of the probabilities. The probabilities $p(y)$'s can be computed using the Taylor expansion of $\phi(t)$ which needs the computation of the y-th derivative of $\phi(t)$ at $t = 0$. A more efficient way is to obtain a finite recurrence relation between the $p(y)$'s using the fact that $\phi(t)$ is a rational function. We have used this method to compute $p(y)$ for $y = 1$ to more than 30 000 in a few minutes on a PC.

5.3 $r-$scans

r -scans have been used by Karlin and Macken (1991) in a genome analysis context to study the homogeneity of the distribution of a motif along a sequence. It is defined as the distance between an occurrence and the $r-$th next:

$$Y^r = X_{n+r} - X_n.$$

It is clear that its distribution depends on n only through I_n.

5.3.1 Exact distribution

Let us now consider the r-scan Y_i^r between an occurrence of $\mathbf{w}_i$ and the $r-$th next occurrence of any element of $\mathcal{W}$. Let us denote

$$p_{ij}^r(y) = \Pr\{(X_{n+r} - X_n = y) \cap I_{n+r} = j \mid I_n = i\}$$

and the corresponding generating function $\phi_{ij}^r(t) = \sum_{y\geq 1} p_{ij}^r(y)t^y$.

Theorem 8 *The generating function matrix $\mathbf{\Phi}^r(t)$ of general term $\phi_{ij}^r(t)$ is*

$$\mathbf{\Phi}^r(t) = [\mathbf{\Phi}(t)]^r .$$

Proof. Let us proceed by recurrence and assume that the theorem is true up to $r-1$. One has $p_{ij}^r(y) = \sum_{l=1}^m \sum_{z=1}^{y-1} p_{il}(z)p_{lj}^{r-1}(y-z)$ so $\phi_{ij}^r(t)$ is equal to $\sum_{l=1}^m \left[\sum_{z\geq 1} p_{il}(z)t^z\right] \times \left[\sum_{y\geq z+1} p_{lj}^{r-1}(y-z)t^{y-z}\right] = \sum_{l=1}^m \phi_{il}(t)\phi_{lj}^{r-1}(t)$ so $\mathbf{\Phi}^r(t) = \mathbf{\Phi}(t)\mathbf{\Phi}^{r-1}(t)$. ■

The following two corollaries come straightforward.

Corollary 9 *The probability generating function $\phi_{i\bullet}^r(t)$ of the r-scan Y_i^r is*

$$\phi_{i\bullet}^r(t) = \phi_i^r(t).\mathbf{1}$$

where $\phi_i^r(t)$ is the $i-$th row of $\mathbf{\Phi}^r(t)$.

Corollary 10 *Let ν be a distribution over $\mathcal{W}$. If the first word is distributed according to ν, then the probability generating function of the $p_{\nu\bullet}^r(y)$ is*

$$\phi_{\nu\bullet}^r(t) = \nu.\mathbf{\Phi}^r(t).\mathbf{1}.$$

First word $\mathbf{w}_0$ given. Using the notation of Corollary **7**and the properties of the generating functions of sums of independent random variables, the following corollary is obvious.

Corollary 11 *If the first word is $\mathbf{w}_0$, the generating function of the $p_{0\bullet}^r(y)$ is*

$$\phi_{0\bullet}^r(t) = \phi_0(t).\mathbf{\Phi}^{r-1}(t).\mathbf{1}$$

Computation of the probabilities. The probabilities $p^r(y)$'s could be computed using $\phi^r(t)$ but this appears to lead to numerical instability so it seems better to use the r-th convolution of the $p(y)$.

5.3.2 Chen-Stein approximation for extremal $r-$scans

Dembo and Karlin (1992) have obtained results (Theorem 1) on the distribution of minimum (and maximum) r-scans using a Poisson approximation. Let $N^-(a)$ denote the number of r-scans shorter than (or equal to) a in a sequence of n distances: they approximate the distribution of $N^-(a)$ by a Poisson distribution with parameter $\lambda = (n-r+1)\Pr\{Y_r \leq a\}$.

Using the Chen-Stein method, assuming that the distances Y are i.i.d., they give an upper bound $b(a)$ for distance in total variation between the distribution of $N^-(a)$ and a Poisson distribution $\mathcal{P}[\lambda(a)]$:

$$b(a) = \left(1 - e^{-\lambda(a)}\right)\left[2(r-1)\Pr\{Y_r \leq a\} + 2\sum_{i=1}^{r-1}\Pr\{Y_i \leq a\}\right].$$

Using the exact distribution of $Y_{\nu\bullet}$ given above, we can calculate $\Pr\{Y_r \leq a\}$ for any r and any a, and then calculate the exact values of $\lambda(a)$. r-scan can be used to detect region with a high (or low) density of a family of words. Usually, one computes the threshold a^* such that $\exp[-\lambda(a^*)] = \alpha$ (where α is a given risk). If the smallest r-scan is smaller than a^*, it is said significantly small.

Case of a renewal process. The i.i.d. hypothesis required for the validity of the Chen-Stein bound does not hold in the general case. This hypothesis is true if the semi-Markovian process is a renewal process, e.g. when the distribution of Y_{ij} does not depend on i, as in the example proposed in section **5.4.2**. In this case, we may calculate the Chen-Stein bound $b(a)$ and derive two more thresholds:

$$a^*_{\text{inf}} : \exp[-\lambda(a)] + b(a) = \alpha, \qquad a^*_{\text{sup}} : \exp[-\lambda(a)] - b(a) = \alpha.$$

It is only possible to arrive at a conclusion when the smallest r-scan is either smaller than a^*_{inf} or larger than a^*_{sup}. When $b(a) > \exp[-\lambda(a)]$ for some $a \geq a^*$, it may be impossible to obtain a^*_{sup} which is not defined in this case. According to our experience, this happens fairly often, especially when the ratio between the total number of occurrences and r is not sufficiently high.

5.4 Applications

5.4.1 Sooner waiting time problem

The sooner waiting problem has been extensively studied. For example Koutras and Alexandrou (1997) have generating function in a sequence of trinary trials. This problem considers 'pure words' i.e. words made of repetitions of the same letter. The most often studied case is the success and failure runs where $\mathcal{A}$ =$\{0,1\}$ and $\mathcal{W} = \{(1,\ldots,1),\ (0,\ldots,0)\}$ with $|\mathbf{w}_1| = k_1$ and $|\mathbf{w}_0| = k_0$. Our results can be applied to deal with any size of alphabet. To calculate the generating function we only need the transition matrix $\mathbf{\Pi}$, the overlapping polynomial matrix $\mathbf{D}(t)$ and the diagonal matrix $\mathbf{T}(t)$.

For example, in the simplest case of success and failure runs, we get

$$\mathbf{D}(t) = \begin{bmatrix} \pi(0,0)\dfrac{1-[\pi(0,0)t]^{-k_0}}{1-[\pi(0,0)t]^{-1}} & 0 \\ 0 & \pi(1,1)\dfrac{1-[\pi(1,1)t]^{-k_1}}{1-[\pi(1,1)t]^{-1}} \end{bmatrix}$$

and

$$\mathbf{T}(t) = \begin{bmatrix} t^{k_0}\pi(0,0)^{k_0-1} & 0 \\ 0 & t^{k_1}\pi(1,1)^{k_1-1} \end{bmatrix}.$$

5.4.2 Genome analysis: CHI of *Haemophilus influenza*

We study here the Crossover Hotspot Initiator of *Haemophilus influenza* which is known to protect the genome against restriction enzymes and is therefore expected to be particularly frequent and regularly spaced.

There are $m = 4$ versions of this pattern which are all $k_i \equiv 8$ letters long:

$$\mathcal{W} = \{\texttt{gatggtgg}, \texttt{gctggtgg}, \texttt{ggtggtgg}, \texttt{gttggtgg}\}.$$

It can be noted that $\mathbf{w}_3$ overlaps itself and the other words with 1, 2 and 5 letters. The other words overlap each other with only 1 letter. The overlapping structure does not depend on the first word, the overlapping polynomials $d_{ij}(t)$ and the generating functions $\phi_{ij}(t)$ do not depend on i so $\phi_{i\bullet}(t) = \phi_{\nu\bullet}(t)$ and the probability distribution function (pdf) of the distance and of the r-scans do not depend on the first word. Hence the distances are i.i.d. and we are in the renewal case described in section **5.3.2**. For the same reason, the generating matrix $\mathbf{\Phi}(t)$ and the transition matrix $\mathbf{S}$ have all their rows equal; any row of $\mathbf{S}$ gives its steady distribution ν.

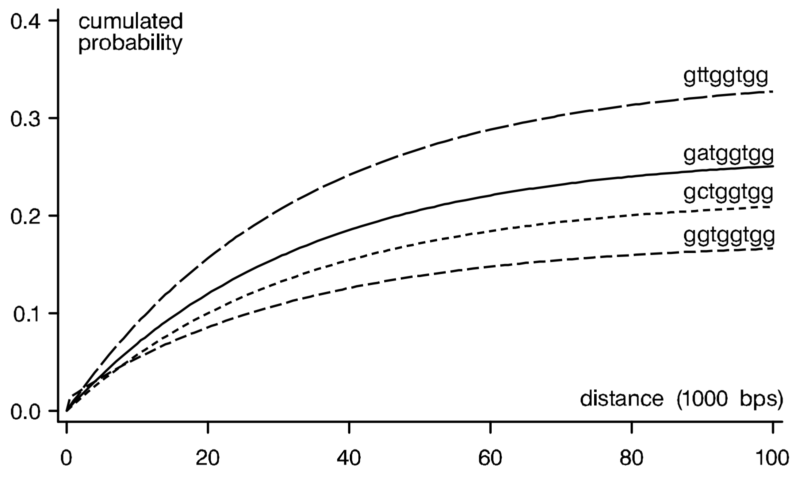


Figure 5.1: Distribution of the distances Y_{ij}'s between the different versions of the CHI motif of *H. influenza*.

The model. Assuming that the sequence is generated according to a first order Markov model (adjusted to the sequence) with transition matrix

$$\mathbf{\Pi} = \begin{pmatrix} 0.3827 & 0.1547 & 0.1639 & 0.2988 \\ 0.3429 & 0.1874 & 0.2156 & 0.2541 \\ 0.2693 & 0.2641 & 0.1974 & 0.2692 \\ 0.2302 & 0.1595 & 0.2202 & 0.3901 \end{pmatrix}$$

the expected distance between two occurrences is $\mathbb{E}(Y_{\nu\bullet}) = 32\,535.8$ and the standard deviation is $32\,916.6$. Under this model, the generating functions $\phi_{ij}(t)$ are all ratios of two polynomials with degree 10 and the transition probabilities between words $s(i,j)$ are simply given by the steady distribution ν :

$$\nu = \begin{pmatrix} 0.2632 & 0.2195 & 0.1738 & 0.3435 \end{pmatrix}.$$

The cumulative probabilities $\sum_{z=1..y} p_{ij}(z)$ are given in figure **5.1**. Remember that they are not cumulative distribution functions since they do not reach 1: the asymptotic values of this distributions are equal to $s(i,j)$ which equal in this case to $\nu(j)$. We see that the $p_{i3}(y)$'s are superior to all other $p_{ij}(y)$'s for very small y: this is due to the fact that $\mathbf{w}_3$ has more chance to overlap other word.

The data. The complete genome of *H. influenza* is 1,830,022 base pairs (bps) long. The number of occurrences of $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{w}_4$ are respectively 28, 56, 76 and 63 so there are 223 occurrences and $n = 222$ distances between them. The mean distance is $8,087.4$ bps, the standard deviation $9\,544.1$ bps, the smallest distance 3 bps (observed 8 times) and the greatest $70\,722$ bps. The distance 3 is necessarily obtained by an overlap of $\mathbf{w}_3$ on another word.

Use of r-scans The heterogeneity of the CHI motif in the genome can be checked using 3-scans (see figure **5.2**): the 5% threshold of the simple Poisson approximation is $a^* = 2744$ ($r/a^* = 1.09\,10^{-3}$) but, taking the error bound $b(a)$ into account we get $a^*_{\text{inf}} = 2537$ ($r/a^*_{\text{inf}} = 1.18\,10^{-3}$) and $a^*_{\text{sup}} = 3071$ ($r/a^*_{\text{sup}} = 0.98\,10^{-3}$). We observe 4 peaks exceeding the upper bound for the density r/a^*_{inf} and one candidate for which we can not decide.

Two interpretations of this result can be given:

- The peaks reveal real rich regions for the CHI motif.
- The general frequency of the CHI motif is much higher than expected under the M1 model. In this case, the peaks may be simple consequences of the departure from the model.

In this case, the second interpretation seems more valid since the density is everywhere higher than expected. The mean and variance of a 222-scan under the M1 model are respectively $\mu = 7\,223\,007$ and $\sigma = 490\,445$; the distance between the first and the last occurrence of the motif is $y^{222} = 1\,795\,407$. Using the central limit theorem for Y^{222}, we get a gaussian score of -11.1, which is highly significant.

The upper bound a^*_{sup} can not be computed for the 4-scan because $e^{-\lambda(a)} - b(a)$ turns out to be negative.

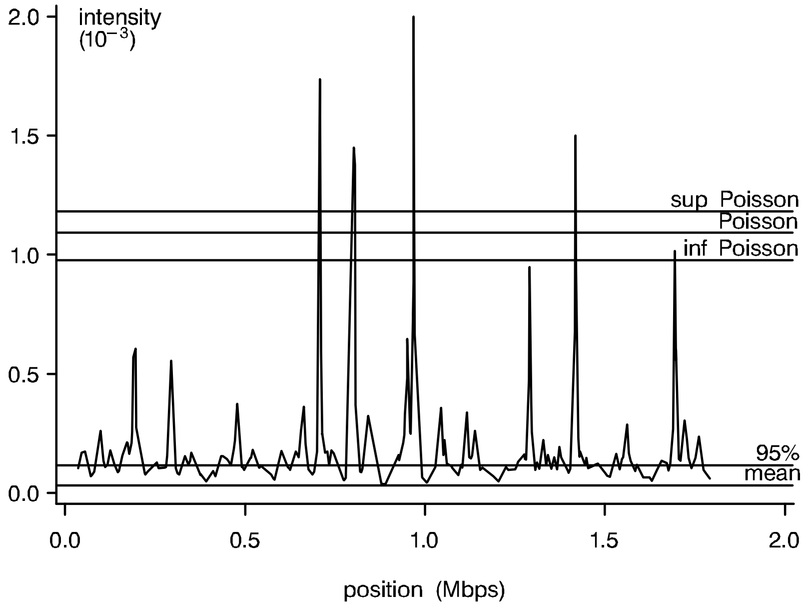


Figure 5.2: Density of the CHI motif using 3-scans in the genome of *H. influenza*: mean $= 1/\mathbb{E}Y$, 95% $= r/y^r_{95\%}$ where $\Pr\{Y^r < y^r_{95\%}\} = 95\%$, Poisson $= r/a^*$, inf Poisson $= r/a^*_{\text{sup}}$, sup Poisson $= r/a^*_{\text{inf}}$ (the highest peak at position 0.97 Mbps is truncated, its real height is 4.8 10^{-3}).

Acknowledgment. We thank an anonymous referee for his helpful technical remarks.

Bibliography

Aki, S. and Balakrishnan, S., N.and Mohanty. (1996). Sooner and later waiting times problems for success and failure runs in higher order markov dependent trials. *Ann. Inst. Statist. Math.* **48 (4)** 773–787.

Aki, S. and Hirano, K. (1993). Discrete distributions related to succession events in a two state markov chain. In *Statistical Sciences and Data Analysis*, (K. Matusita, M. L. Puri, and T. Hayakawa, ed.), 467–474. VSP Publishers, Amsterdam.

Breen, S., Waterman, M. S. and Zhang, N. (1985). Renewal theory for several patterns. *J. Appl. Prob.* **22** 228–234.

Chryssaphinou, O. and Papastavridis, S. (1990). The occurrence of sequence patterns in repeated dependent experiment. *Theory Prob. Appl.* **35 (1)** 145–152.

Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes. *Ann. Appl. Prob.* **2 (2)** 329–357.

Fu, C. J. (1996). Distribution of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica.* **6** 957–974.

Fu, C. J. and Koutras, M. V. (94). Distribution theory of runs: A markov chain approach. *J. Amer. Statist. Assoc.* **427** 1050–1058.

Karlin, S. and Macken, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Amer. Statist. Assoc.* **86** 27–35.

Koutras, M. (1997). Waiting times and number of appearances of events in a sequence of discrete random variables. In *Advances in Combinatorial Methods and Applications to Probability and Statistics*, (N. Balakrishnan, ed.), 253–261, Statistics and Industry and Technology Series. Birkhauser, Boston.

Koutras, M. and Alexandrou, V. (1997). Sooner waiting time problems in a sequence of trinary trials. *J. Appl. Prob.* **34** 593–609.

Mori, T. (1991). On the waiting time til each of some given patterns occurs as a run. *Prob. Theory Rel. Fields.* **67** 313–323.

Robin, S. and Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36** 179–193.

Uchida, M. and Aki, S. (1995). Sooner or later waiting time problems in a two-state markov chain. *Ann. Inst. Statist. Math.* **47** 415–433.

Chapitre 6

Numerical comparison of several approximations of the word count distribution in random sequences

Robin S., Schbath S. (2001)

Journal of Computational Biology, **8** (4), 349–359

Abstract. The exact distribution of word counts in random sequences and several approximations have been proposed in the past few years. The exact distribution has no theoretical limit but may require prohibitive computation time. On the other hand, approximate distributions can be rapidly calculated but, in practice, are only accurate under specific conditions. After making a survey of these distributions, we compare them according to both their accuracy and computational cost. Rules are suggested for choosing between Gaussian approximations, compound Poisson approximation and exact distribution. This work is illustrated with the detection of exceptional words in the phage *Lambda* genome.

Keywords. approximate distributions, exact distribution, exceptional words, simulations, word count.

Contents

6.1 Introduction

Counts of words or motifs are part of the elementary statistics used in biological sequence analysis, for example to find significantly under- or overrepresented words. Knowledge of the statistical distribution of these counts is necessary to assess the significance of the observed results.

The exact distribution of the count of a word is known under the hypothesis that the letters are independent (Bernoulli) or under a Markov model: see Régnier (1999) or Robin and Daudin (1999). It is theoretically possible to calculate this distribution for any word in any sequence. In practice, it is impossible to compute it in a reasonable time for long sequences or for very frequent words. On the other hand, two kinds of approximations exist: Gaussian approximations (Kleffe and Borodovsky (1992), Prum *et al.* (1995)) and Poisson approximations (Schbath (1995), Geske *et al.* (1995)). The asymptotic framework in which these approximations are valid are different but they both require that the length of the sequence tends to infinity. Their main advantage is that, in most cases, they require much shorter computation times.

This work consists then in comparing these different methods considering both the accuracy of the approximations and the computation time. We aim to answer the question: *Which method should I use to study the frequency of this word in a sequence of this length?*. We distinguish here the case where the parameters of the model are known (probabilistic approach) from the case where they need to be estimated on the analyzed sequence (statistical approach). In the second case, the estimated expected count is not independent from the observed count.

We recall in Section 6.2 the different results existing on the distribution of the count of a given word (more technical details are given in the appendix). We describe in Section 6.3 the simulations design (factors and comparison criteria). Section 6.4 presents the results of the probabilistic approach; those of the statistical approach form Section 6.5. As a concrete situation, we finally study in Section 6.6 the influence of the different existing methods on the detection of under- or overrepresented 3-, 6- and 9-nucleotides in the complete genome of the phage *Lambda*.

6.2 Exact distribution and approximations

We are concerned with the occurrences of a k-letter word $w = w_1 \cdots w_k$ in a random sequence S of ℓ letters from a finite alphabet $\mathcal{A}$. The exact distribution of the word count and the asymptotic distributions presented here are available under a Markov model for the sequence (see Schbath (2000) or Reinert *et al.* (2000) for an overview). Using Markov models is a way to take into account the observed frequencies of short words: nucleotides for Markov model of order zero (Bernoulli model), dinucleotides for first order Markov model, etc. However, we restrict ourselves to the Bernoulli model because the distributions, especially the exact distribution, are more easily computable. Increasing the order of the Markov model does not affect the quality of the approximation but increases the computation time. We then assume that the letters $\{S_i\}_{i=1,\dots,\ell}$ are independent and identically distributed according to a given distribution μ on $\mathcal{A}$. Under this model, the probability that the word w occurs at a given position in S is

$$\pi(w) = \mu(w_1) \times \cdots \times \mu(w_k).$$

Let $N = N_\ell$ denote the number of overlapping occurrences of w in the sequence S. For instance $w =$ `aaa` occurs four times in `aaatgaaaaac`, starting at positions 1, 6, 7 and 8. We will focus on the distribution of N defined by

$$p(n) = \Pr\{N = n\}, \quad \forall n \geq 0,$$

or by its cumulative distribution function $P(n) = \Pr\{N \leq n\}$.

6.2.1 Exact distribution

The exact distribution of the count N is given in Régnier (1999) through its probability generating function, which is a rational function. The probabilities $p(n)$ are the coefficients of the Taylor expansion of this generating function; in the case of a rational function, this expansion can be obtained with a finite recurrence.

The exact distribution of the count N_ℓ can also be obtained via the distribution of the position X_n of the n-th occurrence of w in S, thanks to the equality $\Pr\{X_n \leq \ell - k + 1\} = \Pr\{N_\ell \geq n\}$, for all n. Because the generating function of X_n given in Robin and Daudin (1999) is explicit and the recurrence is easy to program (see Appendix 6.A.1), we shall use it to calculate the $p(n)$'s. In the most favorable cases, the computation time is in $O(kn \log \ell)$ but, for the sake of numerical stability, an elementary algorithm in $O(kn\ell)$ is often more satisfying.

6.2.2 Gaussian approximations (KB and PRT).

If the distribution μ of the letters is known, the mean $\mathbb{E}_\mu(N)$ and variance $\mathbb{V}_\mu(N)$ of the count N can be calculated (see Kleffe and Borodovsky (1992) or Appendix 6.A.2 for the formulas). Using a central limit theorem, the distribution of N converges to a Gaussian distribution as the length of the sequence grows to infinity.

For the case where the distribution μ of the letters is unknown but estimated with the vector $\widehat{\mu}$ of the observed frequencies of the letters in a given sequence, Prum *et al.* (1995) propose another Gaussian approximation. This approximation takes the dependency between the observed count and the expected count $\mathbb{E}_{\widehat{\mu}}(N)$ into account: in this, the normalization factor is different from the standard deviation of the count (see Appendix 6.A.3).

Both of these approximations are based on the fact that the expected count is a linear function of the length of the sequence, which is valid only for frequent words. Let p_{KB} and p_{PRT} denote the corresponding Gaussian distributions.

6.2.3 Compound Poisson approximation (CP).

For rare words, Schbath (1995) and Geske *et al.* (1995) consider a sequence of length ℓ growing to infinity and a word of length k_ℓ growing as $\log \ell$. In this asymptotic framework, the approximate distribution of the count, denoted p_{CP}, is a compound Poisson distribution, the parameters of which are given explicitly in Appendix 6.A.4. For nonoverlapping words (a word is said to be nonoverlapping if and only if two of its occurrences can never overlap in a sequence), the compound Poisson distribution is an ordinary Poisson distribution.

A bound of the approximation error is known from the Chen-Stein method (see Schbath (1995) or Reinert and Schbath (1998)): it is in $\ell k \pi(w)^2 \simeq \mathbb{E}_\mu(N) \times k\pi(w)$ and tends to zero when ℓ goes to infinity (see also Erhardsson (1999, 2000)). This convergence still holds when using $\widehat{\mu}$ instead of μ.

6.3 Design

In the remainder we shall consider sequences of letters taken in the alphabet of the nucleotides $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

The parameters of the p_{PRT} and p_{CP} distributions are calculated using C++ subroutines of the *R'MES* software (Bouvier *et al.* (1999)). The exact distribution and its mean and variance used in the p_{KB} distribution are calculated using a C subroutine of the EXCEP software (Klaerr-Blanchard *et al.* (2000)).

Factors. To compare the quality of the approximations in a sufficiently wide range of situations, we considered the following factors:

- *The length ℓ of the sequence* (10 levels): $\ell = 10^i$ for $i = 1.5, 2.0, 2.5, \ldots, 6.0$.

- *The distribution μ of the nucleotides* $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ (2 levels): uniform $\mu = (.25, .25, .25, .25)$ and HI (adjusted to the genome of *H. Influenzae* in the replication direction, known to be particularly unbalanced) $\mu \simeq (.31, .18, .20, .31)$.

- *The word w*: this factor gathers eight levels corresponding to all the possible combinations of three binary subfactors: the length of the word ($k = 3$ or 9 nucleotides), the frequency of the nucleotides present in the word (rare or frequent with respect to the HI distribution) and its overlapping structure (overlapping or nonoverlapping). The eight selected words are: `ttt`, `att`, `ccc`, `ccg`, `atatatata`, `ttatatata`, `cgcgcgcgc` and `cgcgcgcgg`.

Quality criteria. For each situation corresponding to a particular combination (ℓ, μ, w), we calculate several criteria measuring the quality of each approximation. We present here only the main ones.

- *The total variation distance* $d_{\mathrm{TV}}(p, \hat{p})$ between the exact distribution p and its approximation $\hat{p}$:

$$d_{\mathrm{TV}}(p, \hat{p}) = \sup_{A \subset \mathbb{N}} |p(A) - \hat{p}(A)| = \frac{1}{2} \sum_{n \geq 0} |p(n) - \hat{p}(n)|$$

which measures the greatest error one can make, in terms of probability, when using $\hat{p}$ instead of p (Barbour *et al.* (1992)). This distance will be our main criterion to evaluate the global quality of an approximation.

- *The Kolmogorov distance* $d_K(p, \hat{p})$:

$$d_{\mathrm{K}}(p, \hat{p}) = \max_{n \geq 0} |P(n) - \widehat{P}(n)|$$

This distance is less severe than the total variation distance since it considers only events of the form $A = (-\infty; n]$. Yet these events are usually the ones we are interested in, especially when checking over- or underrepresentation of words.

- *The maximal difference between point probabilities*: $\max_{n \geq 0} |\hat{p}(n) - p(n)|$.

- *The effective levels $\tilde{\alpha}$ of six quantiles of the distribution $\hat{p}$.* For the six levels $\alpha =$ 0.05%, 0.5%, 5%, 95%, 99.5% and 99.95%, we calculate first the quantile $\hat{u}_\alpha$ of $\hat{p}$ such that $\widehat{P}(\hat{u}_\alpha) = \alpha$ and then the exact level $\tilde{\alpha}$ such that $P(\hat{u}_\alpha) = \tilde{\alpha}$. These levels give an indication of the quality of the approximation for the tail of the distribution.

6.4 Probabilistic approach

Goal. We consider here the case where the parameters of the model (i.e., the distribution μ of the nucleotides) is known. In this case, the exact distribution $p = p^\mu$ of the count is known and the only two relevant approximate distributions are p_{CP} and p_{KB} as the vector μ of the frequencies does not need to be estimated. The exact distribution and the two approximate distributions are then calculated in a deterministic way for every of the 160 possible combinations of the factors (ℓ, μ, w).

Quality of the approximation. For every couple (μ, w) and for both approximations, the maximal difference between point probabilities (see Section 6.3) tends to zero when the length of the sequence raises to infinity (not shown here). However, this does not guarantee good behavior of the approximation of the whole distribution for which the total variation distance is a much more relevant criterion. Figure 6.1 shows the evolution of this distance as a function of the expected count $\mathbb{E}(N)$, for $\mu = \mathrm{HI}$.

The Gaussian approximation KB gives satisfying results ($d_{\mathrm{TV}} \leq 1\%$, an acceptable threshold for tests of level 5% or 10%) when the expected count exceeds 500. The good quality of this approximation for small expected counts ($\mathbb{E}(N) \leq 10^{-2}$) is due to the fact that the distributions, exact and approximate, are concentrated on the event $\{N = 0\}$. The bad quality of that for intermediate expected counts is especially due to the nonsymmetrical shape of the exact distribution.

The results regarding long and short words do not cover the same range of value for the expected count $\mathbb{E}(N)$, but the continuity of the results issued from these two groups leads us to conclude the expected count is the main parameter that determines the quality of the KB approximation.

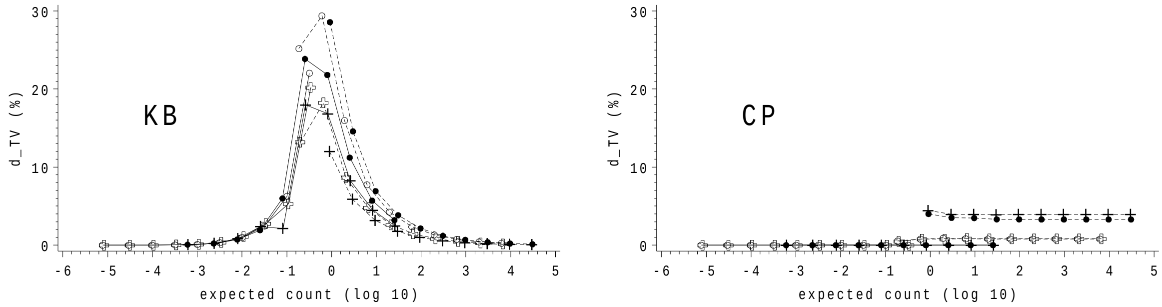


Figure 6.1: Total variation distance between p and, successively, p_{KB} and p_{CP} as a function of the expected count for the HI distribution of the nucleotides. The 8 curves correspond to the 8 considered words: solid line for long words, dashed for short ones; filled symbols for frequent words, empty for rare ones; circles for overlapping words, crosses for nonoverlapping ones.

The CP approximation gives good results for the long words ($k = 9$) and the short rare words (`ccc` and `ccg`). For the short but frequent words (`ttt` and `att`) this approximation is less satisfactory. It appears here the expected count is not the only parameter that determines the quality of the approximation: the probability $\pi(w)$ has also an influence on this quality. This observation agrees with the Chen-Stein bound (see Section 6.2.3). Moreover, the total variation distance does not tend to zero when the length of the sequence tends to infinity. To observe the announced convergence, the length of the word would need to increase with the length of the sequence. It has to be pointed out that the Chen-Stein bound gives a pessimistic measure of the quality of the approximation: it exceeds 100% for $w = \texttt{ttt}$ and $\ell \geq 100$, and is equal to 14% for $w = \texttt{ccg}$ and $\ell = 100,000$, whereas the real distance in total variation is only 0.8%.

The above remarks are still valid looking at the effective levels or at the Kolmogorov distance that varies here between 50% and 70% of the total variation distance (not shown here).

The results obtained with the uniform distribution μ are coherent with the ones observed with HI, except for the distinction between rare and frequent words which has no meaning in this case.

Computation time. Because the numerical computations of the three distributions p, p_{KB} and p_{CP} do not required the same amount of time, we have listed in Table 6.1 these computation times for different word lengths and different sequence lengths. We have used the uniform distribution for the letters, so no distinction is made between rare and frequent words. For a given word length, 3 or 9, the computation time is averaged on the two overlapping words defined in Section 6.3. Instead of considering the 2×10 possible couples for (k, ℓ), we have considered five cases with respect to the order of magnitude of the expected count. Finally, by computation time for the distribution p (or p_{KB} or p_{CP}), we mean the time needed to calculate all the point probabilities $p(n)$ for n between 0 and $N_{\max}$. The quantity $N_{\max}$ depends on the word, on ℓ, and on μ and is chosen in such a way that the exact tail probability is lower than 10^{-6}; it is identical for all the distributions, exact or approximates.

$\mathbb{E}(N)$	word length	ℓ	KB d_{TV}	KB time (ms)	CP d_{TV}	CP time (ms)	Exact time (ms)
$\simeq 0.01$	3	< 1 bp	*out of the design*				
	9	3.2 kb	1.12	0.07	$< 10^{-3}$	**0.08**	**5.65**
$\simeq 1$	3	0.1 kb	20.7	$< 10^{-3}$	2.24	0.01	**0.02**
	9	316 kb	**15.7**	**0.40**	**0.002**	**0.25**	**1334**
$\simeq 50$	3	3.2 kb	2.70	0.04	1.81	7.69	53.7
	9	> 1 Mb	*out of the design*				
$\simeq 500$	3	32 kb	**0.87**	**0.10**	**1.81**	**302.7**	**3434**
	9	> 1 Mb	*out of the design*				

Table 6.1: Total variation distance (%) and computation time (ms) for the p_{KB}, p_{CP} and exact distributions for four magnitudes of the expected count. For a given word length, the total variation distance and the computation time are averaged on the two corresponding overlapping words. Computation times have been obtained with a SUN Entreprise 450 (Ultra SPARC II 300MHz), 1.5 Go RAM.

We clearly see from Table 6.1 that the approximate distributions can be calculated more rapidly than the exact distribution. As soon as one wants to process a large number of words, the exact distribution becomes quite unreachable in long sequences or for frequent words.

It is important to note that the computation times given in Table 6.1 are the ones for overlapping words. For nonoverlapping words, it makes quite no difference except for the CP approximation; indeed the compound Poisson distribution reduces to a Poisson distribution that is much faster to calculate. Moreover, if one is interested only in the tail distribution of the count $\Pr\{N \geq n\}$, there is no need to calculate all the point probabilities of the Gaussian and Poisson distributions.

Suggested rule. To choose which distribution to use, we suggest balancing the quality of the approximation and the computation time required to compute the different distributions (see Table 6.1). The rule seems to be as follows:

- to use the exact distribution when $\mathbb{E}(N) \leq 0.01$ (however, CP is 70 times faster for an excellent quality of approximation and could be preferred for many iterations);

- to use the exact distribution for short words ($k = 3$) but p_{CP} for long words ($k = 9$) when $\mathbb{E}(N) \simeq 1$;

- to use the exact distribution when $\mathbb{E}(N) \simeq 50$ or p_{KB} for many iterations; and

- to use p_{KB} when $\mathbb{E}(N) \geq 500$: the CP approximation is both slower and less accurate, and the exact distribution is too time consuming.

6.5 Statistical approach

Simulations. We consider now the case where the distribution μ of the nucleotides is estimated with the observed frequencies in a DNA sequence, as is commonly done in practice. For each letter distribution μ (uniform and HI) and each sequence length, we simulate 1,000 sequences from which we successively

- estimate the empirical distribution $\widehat{\mu}$;
- compute the four approximate distributions given $\widehat{\mu}$: $p^{\widehat{\mu}}$, $p^{\widehat{\mu}}_{\text{KB}}$, $p^{\widehat{\mu}}_{\text{PRT}}$ and $p^{\widehat{\mu}}_{\text{CP}}$; and
- calculate the different criteria described in Section 6.3 using the exact distribution p^{μ}.

Of course, in practice, the exact distribution p^{μ} is not available since μ is unknown and we can only use the four approximate distributions $p^{\widehat{\mu}}$, $p^{\widehat{\mu}}_{\text{KB}}$, $p^{\widehat{\mu}}_{\text{PRT}}$ and $p^{\widehat{\mu}}_{\text{CP}}$. However, these simulations allow us to get the exact distribution and to compare it to each of its approximations. Moreover, we are able to determine the influence of the estimation of the parameters by comparing $p^{\widehat{\mu}}$ to p^{μ}. From here on, the approximate distribution $p^{\widehat{\mu}}$ is denoted by "exact" distribution.

Because $p^{\widehat{\mu}}$ is long to compute, especially for long sequences (cf. Table 6.1), we reduced the number of simulations from 1000 to 100 for sequences longer than 10^5. The results presented in Figure 6.2 are then obtained from $65,000$ simulations.

Total variation distance. From Figure 6.2, one observes the following main trends:

1. Both Gaussian approximations KB and PRT have similar behaviors; the compound Poisson distribution p_{CP} looks like the "exact" one $p^{\widehat{\mu}}$.
2. All four approximations are of good quality when the expected count is small ($\mathbb{E}(N) \leq 1$) simply because the distributions are all concentrated on the event $\{N = 0\}$. This has been already observed in the probabilistic approach when no estimation is done (cf. Section 6.4).
3. Surprisingly, when the expected count is high ($\mathbb{E}(N) \gg 1$), all four approximations provide the same very bad quality. The total variation distances remain quite constant as $\mathbb{E}(N)$ grows to infinity, contrarily to the maxima of the point probabilities that converge to zero (not shown here). The bad behavior of the "exact" distribution $p^{\widehat{\mu}}$ compared to p^{μ} shows that the penalization due to the estimation of the parameters is very important. By analogy with the triangular inequality

$$d_{\text{TV}}(\widehat{p}^{\widehat{\mu}}, p^{\mu}) \leq d_{\text{TV}}(\widehat{p}^{\widehat{\mu}}, p^{\widehat{\mu}}) + d_{\text{TV}}(p^{\widehat{\mu}}, p^{\mu}),$$

   the approximation cost $d_{\text{TV}}(\widehat{p}^{\widehat{\mu}}, p^{\widehat{\mu}})$ seems negligible compared to the estimation cost $d_{\text{TV}}(p^{\widehat{\mu}}, p^{\mu})$. In the case of the CP approximation, the three distances of the above inequality are, respectively illustrated by Figure 6.2 (CP), Figure 6.1 (CP) and Figure 6.2 ("exact").

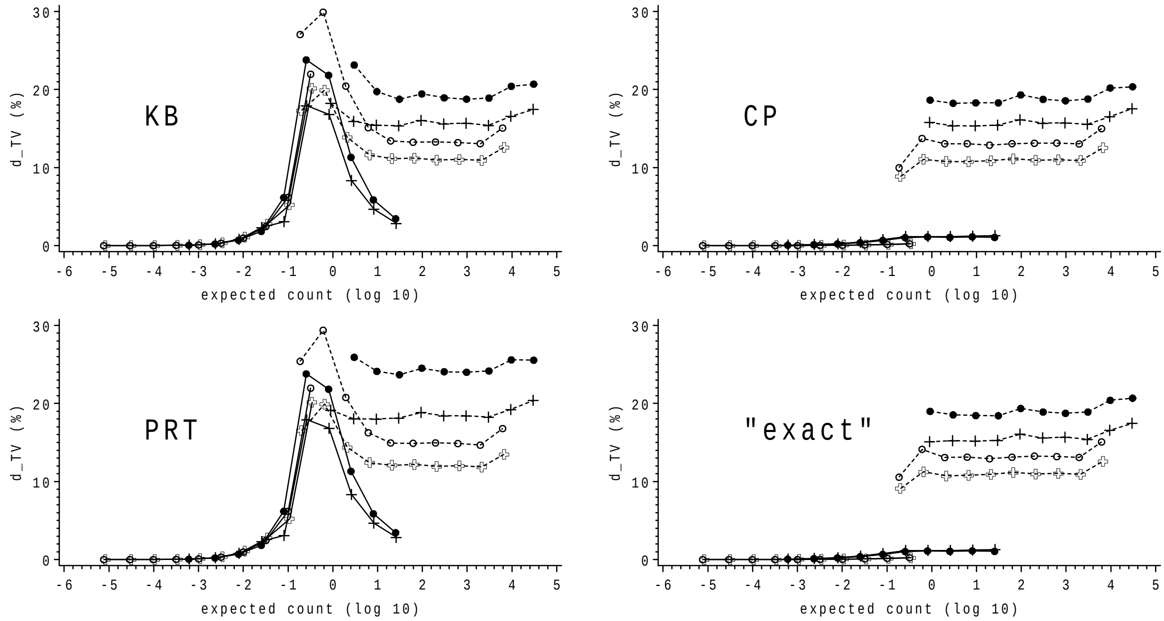


Figure 6.2: Total variation distance between p^{μ} and, successively, $p^{\widehat{\mu}}_{\text{KB}}$, $p^{\widehat{\mu}}_{\text{PRT}}$, $p^{\widehat{\mu}}_{\text{CP}}$ and $p^{\widehat{\mu}}$ (`"exact"`) as a function of the expected count for the HI distribution of the nucleotides (same legend as Figure 6.1). These values are mean calculated over 1,000 simulations, except the last two of each curve that are calculated on 100 simulations.

4. For a fixed high expected count, the bigger the word probability $\pi(w)$ the worse the quality of the approximations. This quality is also worse for overlapping words.

The Kolmogorov distance here is very close to the total variation distance. Hence, the four previous trends are not artifacts of the total variation distance.

Effective levels. In most sequence analyses, the main interest concerns the tail distribution of the count, so the global criterion of the total variation distance may be too strong; the same applies for the Kolmogorov distance. To study the quality of the four approximations in the tail distribution, we then look at the effective levels described in Section 6.3.
We get the result that all four approximations overestimate the effective levels associated with 95%, 99.5% and 99.95% but, as the expected count grows, they get closer to the announced levels even if no convergence is guaranteed. As an example, Table 6.2 gives, for an expected count close to a thousand, the effective levels associated with the announced levels of 95% and 99.95%. For this criterion, the compound Poisson approximation CP gives the best results.

The effective levels associated with 5%, 0.5% or 0.05% are all higher than those announced and do not seem to get closer to them as the expected count increases.

Computation time. Times required to compute the distributions $p^{\widehat{\mu}}$, $p^{\widehat{\mu}}_{\text{KB}}$ and $p^{\widehat{\mu}}_{\text{CP}}$ are the same using $\widehat{\mu}$ instead of μ and are given in Table 6.1. Since the computation times of

Announced levels (%)	Effective levels (%) $p^{\widehat{\mu}}_{\mathrm{KB}}$	$p^{\widehat{\mu}}_{\mathrm{PRT}}$	$p^{\widehat{\mu}}_{\mathrm{CP}}$	$p^{\widehat{\mu}}$
95.00	93.03	90.04	*93.38*	93.02
99.50	98.93	97.81	*99.05*	98.35
99.95	99.83	99.46	*99.86*	99.84

Table 6.2: Effective levels of the quantiles of each of the approximate distributions associated with the announced levels 95%, 99.5% and 99.95%.

the variances of $p^{\widehat{\mu}}_{\mathrm{KB}}$ and $p^{\widehat{\mu}}_{\mathrm{PRT}}$ are equivalent in the Bernoulli model, we can also refer to Table 6.1 for the Gaussian distribution $p^{\widehat{\mu}}_{\mathrm{PRT}}$.

Conclusion. The suggested rule to choose between the approximations is the one stated in Section 6.4 putting together the Gaussian approximations. The crucial point is that the estimation of the parameters has a disastrous effect on the quality of the four approximations.

6.6 Application to unexpected oligonucleotides in DNA sequences

We focus now on a practical problem in DNA sequence analysis: the detection of oligonucleotides that occur with an unexpected frequency in a given sequence. The problem consists in calculating for each word w (let us say of a given length) the critical probability – the so called p-value – $\Pr\{N(w) \geq n(w)\}$, where $n(w)$ is the observed count of w in the DNA sequence. A high p-value corresponds to an underrepresented word, whereas a small p-value corresponds to an overrepresented word. However, with the typical case where the model parameters are not known, one can only use the distribution $\widehat{\mu}$ estimated according to the DNA sequence. One then has four methods to approximate the p-value and the challenging question is "which one to use ?".

For this purpose, we analyzed the phage *Lambda* (48,502bps), and we calculated successively for each oligonucleotide of length 3, 6 and 9 the four approximate p-values using the "exact" distribution $p^{\widehat{\mu}}$, the Gaussian distributions $p^{\widehat{\mu}}_{\mathrm{KB}}$ and $p^{\widehat{\mu}}_{\mathrm{PRT}}$, and the compound Poisson distribution $p^{\widehat{\mu}}_{\mathrm{CP}}$.

Figure 6.3 displays the significance of the counts of the words in the genome of *Lambda*. Each point corresponds to a word: its x-coordinate is the (normalized) p-value of w under p^{μ} and the y-coordinate is its (normalized) approximate p-value using one of the distributions $p^{\widehat{\mu}}_{\mathrm{KB}}$, $p^{\widehat{\mu}}_{\mathrm{PRT}}$ or $p^{\widehat{\mu}}_{\mathrm{CP}}$. For convenience, we use the probit normalization: denoting ϕ the cumulative distribution function of the standard Gaussian distribution, the normalized p-values given on the graph are $\phi^{-1}(p\text{-value})$.

All the graphs show a very good coherence between the approximations. Despite the nonlinear shape observed for the Gaussian approximations KB and PRT, the ordering of the words is very well conserved using any of the four methods. The evolution of the shape of the curves for KB and PRT, as the length of words increases, confirms that the

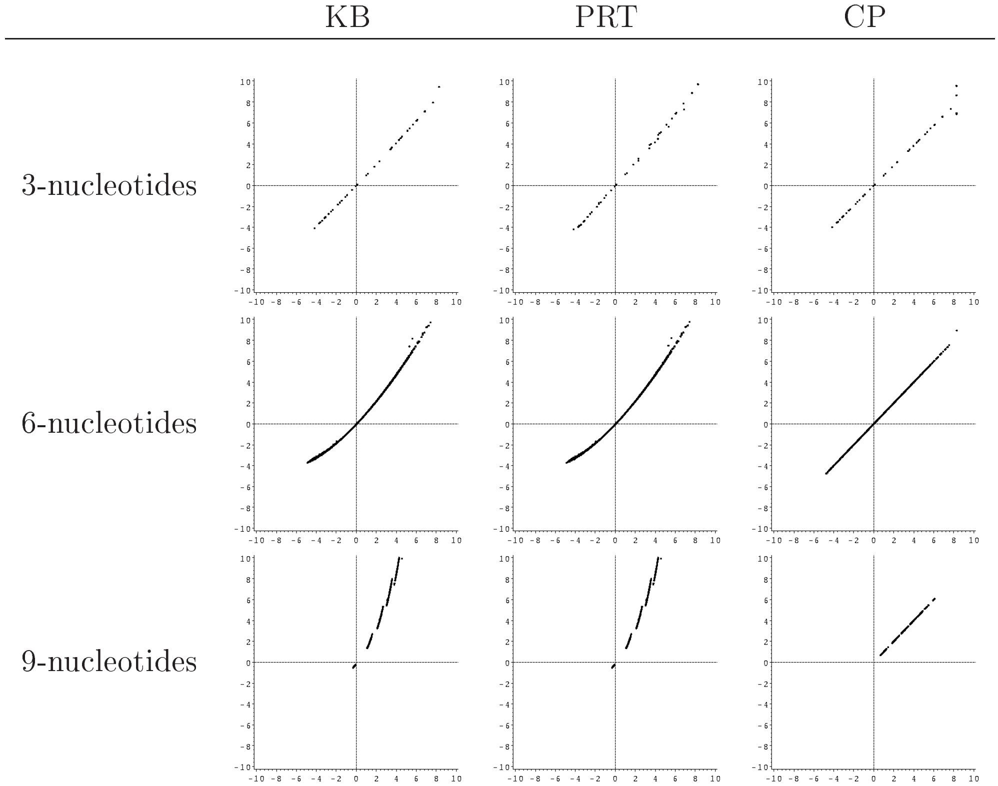


Figure 6.3: The p-values of the counts of all the oligonucleotides of lengths 3, 6 and 9 in the genome of the phage *Lambda* (48,502bps). Comparison of the KB, PRT and CP approximations (y-axis) with the "exact" significancy (x-axis). All the p-values are normalized according to the probit transformation.

rarer the word the worse the Gaussian approximations (see Section 6.2). Surprisingly, the CP approximation gives p-values very similar to the "exact" ones, even for short words. Kendall's correlation coefficient, based on the number of inversions between the orders, is greater than 0.98 for all the approximations for the 3-nucleotides and greater than 0.998 for the oligonucleotides of lengths 6 and 9.

Note that the calculation of the "exact" p-values of the 262,144 9-nucleotides took 6 hours while it only required around 20 seconds using the Gaussian or the compound Poisson approximation.

Therefore, we strongly recommend first using one of the KB, PRT or CP approximation to detect the correct set of most under- and most overrepresented oligonucleotides. This set of exceptional words can then be analyzed more accurately using the "exact" distribution.

Conclusion

This work has been performed under the Bernoulli model (Markov model of order zero). Increasing the order of the Markov model does not affect the quality of the approximations but increases the computation time. In the first order Markov model, times are slightly greater than in the Bernoulli model. However, considering higher orders implies increasing the alphabet and then manipulating matrices with big dimensions. The computations of the exact distribution and the Gaussian distribution p_{KB} are then difficult since they require one to diagonalize such matrices.

As expected, the Gaussian approximations are the most efficient for frequent words. The present work shows that words with expected count exceeding hundreds can be studied with these approximations. On the contrary, for rarer words, they give very bad results. For these words, the compound Poisson approximation is, unsurprisingly, adapted; surprisingly, it also performs well for frequent words.

The catastrophic influence of the estimation of the parameters really was not expected at the beginning of this work. The comparison of p^{μ} and $p^{\widehat{\mu}}$ is the most demonstrative example. This raises the problem of the existence of a true distribution μ different from the empirical one $\widehat{\mu}$, when analyzing biological sequences. Biologically, it could seem reasonable to consider that $\widehat{\mu}$ is in fact the true distribution and then to work conditionally to it. But in this case, the induced model is not a Markovian model anymore. We can think of a sampling model without replacement for which the only existing method is the conditional approach of Prum *et al.* (1995): the PRT Gaussian approximation of this paper. At the present time, there is no literature either on the exact distribution of the count or on a Poisson approximation for rare words.

When no specific words are a priori determined according to some biological information, we need to exhaustively analyze a large set of words and we suggest the following strategy. Using any of the four distributions ($p^{\widehat{\mu}}$ $p^{\widehat{\mu}}_{\mathrm{KB}}$, $p^{\widehat{\mu}}_{\mathrm{PRT}}$ and $p^{\widehat{\mu}}_{\mathrm{CP}}$) give the same ordering of the words with respect to the exceptionality of their frequency. It is then natural to use the fastest one to calculate. This preliminary analysis reveals a subset of exeptionnal words that has to be reanalyzed more accurately with the exact distribution.

Acknowledgment. The authors want to thank E. Coward for lending his programs calculating the exact distribution of the count.

Bibliography

Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation*. Oxford-University Press.

Bouvier, A., Gélis, F. and Schbath, S. (1999). *R'MES : Recherche de Mots Exceptionnels dans les Séquences d'ADN – Version 2*. Guide de l'utilisateur. INRA, Biométrie, F78352 Jouy-en-Josas.
(`http://www.inra.fr/bia/J/AB/genome/RMES/welcome.html`).

Erhardsson, T. (1999). Compound Poisson approximation for Markov chains using Stein's method. *Ann. Prob.* **27** 565–596.

Erhardsson, T. (2000). Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains. Preprint.

Geske, M. X., Godbole, A. P., Schaffner, A. A., Skolnick, A. M. and Wallstrom, G. L. (1995). Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Prob.* **32** 877–892.

Guibas, L. J. and Odlyzko, A. M. (1981). Periods in strings. *J. Combinatorial Theory A.* **30** 19–42.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions.* Wiley: New-York.

Klaerr-Blanchard, M., Chiapello, H. and Coward, E. (2000). Detecting localized repeats in genomic sequences: A new strategy and its application to *B. subtilis* and *A. thaliana* sequences. *Computer Chem.* **24 (1)** 57–70.

Kleffe, J. and Borodovsky, M. (1992). First and second moment of counts of words in random texts generated by Markov chains. *Comp. Applic. Biosci.* **8** 433–441.

Prum, B., Rodolphe, F. and de Turckheim, E. (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* **57** 205–220.

Reinert, G., Schbath, S. and Waterman, M. (2000). Probabilistic and statistical properties of words. *J. Comp. Biol.* **7** 1–46.

Reinert, G. and Schbath, S. (1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol.* **5** 223–253.

Régnier, M. (1999). A unified approach to word occurrence probabilities. Discrete Applied Mathematics, Special Issue on Computational Biology, (*to appear, preliminary version at RECOMB'98*).

Robin, S. and Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36** 179–193.

Schbath, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics.* **1** 1–16. (`http://www.emath.fr/ps/`).

Schbath, S. (2000). An overview on the distribution of word counts in Markov chains. *J. Comp. Biol.* **7** 193–201.

Appendix

6.A Calculating the distributions

The exact distribution and the parameters of the limiting distributions depend on the overlapping structure of the word w, in particular on the periods of w (Guibas and Odlyzko (1981)).

Definition A period of the word $w = w_1 \cdots w_k$ is an integer $p \in \{1, \ldots, k-1\}$ such that $w_i = w_{i+p}$, $\forall i = 1, \ldots, k-p$. Let $\mathcal{P}(w)$ be the set of periods of w.

Let $\mathcal{P}'(w)$ be the set of the *principal* periods of w, namely, the periods that are not strictly multiple of the minimal period of w, and define the quantities

$$\begin{aligned} A_i(w) &= \mu(w_1) \times \cdots \times \mu(w_i) \qquad \text{for } i = 1, \ldots k, \\ A(w) &= \sum_{p \in \mathcal{P}'(w)} A_p(w). \end{aligned}$$

6.A.1 Exact distribution

The recurrences given here can be found with different notations in Robin and Daudin (1999): they give a simple way to calculate the probability

$$q(x, n) = \Pr\{\text{the } n\text{-th occurrence of } w \text{ starts exactly at position } x\}.$$

We use the convention that $q(x, n)$ is null as soon as x or n is negative. The probability at the first position $q(1, n)$ is equal to $\pi(w)$ if $n = 1$ and null otherwise. The following $q(x, n)$'s can be calculated according to the recurrence:

$$\begin{aligned} q(x,n) =& \\ & q(x-1,n) - \mu(w)\big[q(x-k,n) - q(x-k,n-1)\big] \\ & - \sum_{p \in \mathcal{P}(w)} A_p(w)\big[q(x-p,n) - q(x-p-1,n) - q(x-p,n-1) + q(x-p-1,n-1)\big]. \end{aligned}$$

In a sequence of length ℓ, the cumulative distribution function of the count $P(n)$ is equal to the probability that the $(n+1)$-th occurrence of w starts after position $(\ell-k+1)$:

$$P(n) = \sum_{x=\ell-k+2}^{\infty} q(x, n+1) = 1 - \sum_{x=1}^{\ell-k+1} q(x, n+1).$$

6.A.2 Gaussian approximation (KB)

When the letter distribution μ is known, the expectation of N is simply given by $\mathbb{E}_\mu(N) = (\ell - k + 1)\pi(w)$. The variance given by Kleffe and Borodovsky (1992) reduces in the

Bernoulli model to

$$\begin{aligned}\mathbb{V}_\mu(N) &= \mathbb{E}_\mu(N) + 2\sum_{p\in\mathcal{P}(w)}(\ell-k-p+1)\pi(w_1\cdots w_p w_1\cdots w_k)\\ &\quad + \pi^2(w)\Big[\ell(1-2k)+3k^2-4k+1\Big].\end{aligned}$$

The distribution of N is then approximated by the Gaussian distribution with mean $\mathbb{E}_\mu(N)$ and variance $\mathbb{V}_\mu(N)$.

6.A.3 Gaussian approximation (PRT)

When the letter distribution μ is unknown but estimated by $\widehat{\mu}$, the distribution of N is then approximated by the Gaussian distribution with mean $\mathbb{E}_{\widehat{\mu}}(N) = (\ell-k+1)\widehat{\pi}(w) = (\ell-k+1)\widehat{\mu}(w_1)\times\cdots\times\widehat{\mu}(w_k)$ and variance $\widehat{v}$. The variance given by Prum *et al.* (1995) is $\mathbb{E}_{\widehat{\mu}}[N-\mathbb{E}_{\widehat{\mu}}(N)]^2$ and reduces in the Bernoulli model to

$$\begin{aligned}\widehat{v} &= \mathbb{E}_{\widehat{\mu}}(N) + 2\sum_{p\in\mathcal{P}(w)}(\ell-k-p+1)\widehat{\pi}(w_1\cdots w_p w_1\cdots w_k)\\ &\quad + (\ell-k+1)\widehat{\pi}^2(w)\Big[(k-1)^2 - \sum_{a\in\mathcal{A}}\frac{\nu^2(a)}{\widehat{\mu}(a)}\Big],\end{aligned}$$

where $\nu(a)$ is the number of a in the word w. Note that the variance $\hat{v}$ is asymptotically smaller than $\mathbb{V}_\mu(N)$: $\lim_{\ell\to\infty}\frac{1}{\ell}\big[\widehat{v}-\mathbb{V}_\mu(N)\big] = \pi^2(w)\Big[k^2-\sum_{a\in\mathcal{A}}\frac{\nu^2(a)}{\mu(a)}\Big] \le 0.$

6.A.4 Compound Poisson approximation (CP)

The compound Poisson distribution $\mathrm{CP}(\widetilde{\lambda}_1,\widetilde{\lambda}_2,\ldots)$ that approximates the distribution of the number of occurrences N of w in the Bernoulli sequence $\{S_i\}_{i=1\ldots,\ell}$ (Schbath (1995)) is the distribution of $\sum_{j\ge1} j\,Z_j$ where the Z_j's are independent random Poisson variables with parameters $\widetilde{\lambda}_j$ given by

$$\widetilde{\lambda}_j = (\ell-k+1)\pi(w)\big[1-A(w)\big]^2 A^{j-1}(w).$$

Note that if w is not periodic ($\mathcal{P}(w)=\emptyset$) then $A(w)=0$, $\widetilde{\lambda}_j=0$ for $j\ge2$, and the compound Poisson variable reduces to a single Poisson variable Z_1.

Since the expression of $\widetilde{\lambda}_j$ is geometric, it turns out the compound Poisson distribution $\mathrm{CP}(\widetilde{\lambda}_1,\widetilde{\lambda}_2,\ldots)$ is exactly the Pólya-Aeppli distribution with parameters ($\widetilde{\lambda} := \sum_j\widetilde{\lambda}_j, A(w)$); the Pólya-Aeppli distribution with parameters (λ, v) is defined like the distribution of $\sum_{c=1}^{C} K_c$ where C is a random Poisson variable with mean λ and the random variables K_c's are independent and identically distributed according to the geometric distribution with parameter v. With $\widetilde{\lambda} = (\ell-k+1)\pi(w)[1-A(w)]$, we then have (see

Johnson *et al.* (1992), p378):

$$\begin{cases} p_{\mathrm{CP}}(0) = \exp(-\widetilde{\lambda}), \\ p_{\mathrm{CP}}(n) = \exp(-\widetilde{\lambda}) A^n(w) \displaystyle\sum_{c=1}^{n} \frac{1}{c!} \begin{pmatrix} n-1 \\ c-1 \end{pmatrix} \left[\frac{1 - A(w)}{A(w)} \widetilde{\lambda} \right]^c, \ n \geq 1, \end{cases}$$

where $\binom{a}{b}$ denotes the binomial coefficient $a!/(b!(a-b)!)$. The mean and the variance of $\mathrm{CP}(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots)$ are respectively $(\ell - k + 1)\pi(w) = \mathbb{E}_\mu(N)$ and $\mathbb{E}_\mu(N)[1 + A(w)]/[1 - A(w)]$.

Chapitre 7

Occurrence probability of structured motifs in random sequences

Robin S., Daudin J.-J., Richard H., Sagot M.-F., Schbath S.
(2002)

Submitted to *Journal of Computational Biology*

Abstract. The problem of extracting from a set of nucleic acid sequences motifs which may have biological function is more and more important. In this paper, we are interested in particular motifs that may be implicated in the transcription process. These motifs, called structured motifs, are composed of two ordered parts separated by a variable distance and allowing for substitutions. In order to assess their statistical significance, we propose approximations of the probability of occurrence of such a structured motif in a given sequence. An application of our method to evaluate candidate promoters in *E. coli* and *B. subtilis* is presented. Simulations show the goodness of the approximations.

Key-words. Markov models, motif occurrences, promoters, structured motifs.

Contents

7.1 Introduction

The problem of extracting from a set of nucleic acid sequences motifs which may have a biological function (typically, such motifs represent DNA binding sites) has been concerning biologists, statisticians and computer scientists for many years. The motifs that may be considered in terms of algorithms are becoming increasingly sophisticated. In particular, various motifs may be extracted simultaneously. This enables to start addressing a possible cooperative effect between binding sites implicated in a same biological process such as transcription for instance. However, it is in the statistical evaluation of such motifs that we have being lagging behind. Yet it is important to be able to evaluate with enough accuracy how unexpected are such motifs given a model for the sequences. Indeed, exhaustively extracting motifs that satisfy certain constraints (e.g., maximum error rate allowed, minimum frequency of sequence occurrence) is not enough in most cases. Especially when trying to identify weakly conserved motifs, flexibility is required and this may lead to many motifs which satisfy the constraints. Most of these motifs can be "explained" by the sequence composition and only a few stand a chance of being related to a "true" biological object, such as a binding site.

The complex motifs that we are becoming increasingly more efficient in extracting from a set of sequences correspond to what has been called "structured motifs" (see Marsan and Sagot (2000a), (2000b)). They are motifs that may be composed of two or more ordered parts, called "boxes". Each box is separated from the next one by a distance which may take any value inside an interval. The intervals may be different for two pairs of consecutive boxes in a structured motif. Each box may also exhibit a different degree of conservation, i.e., the maximum number of errors allowed (in general, only substitutions) against the corresponding part in the motif may vary from one box to the other. Finally, the structured motif must appear in a minimum number of the input sequences.

Various ways of calculating the probability for a motif composed of just one box to occur in a sequence, possibly allowing substitutions, have been known for some time (Blom and Thorburn (1982), Fu (1996), Robin and Daudin (2001)). Recently, the problem of calculating the probability of exact occurrence of a motif composed of two boxes separated by a fixed distance has also been addressed by van Helden *et al.* (2000). However, the problem of calculating the probability of occurrence of a motif composed of two or more boxes separated by variable distances and allowing for substitutions has not yet been faced. This paper proposes a first method for addressing this problem in the case of structured motifs composed of two boxes separated by a variable distance. It is well known that the probability of occurrence strongly depends on the overlapping structure of the motif, i.e. on the possibility for a motif to overlap itself. Since this overlapping structure cannot be completely described in the case of structured motifs, the method proposed in this paper is not exact but uses generalized geometric approximations in order to calculate the required probability with a good precision.

The paper is organized as follows. Section 7.2 defines a structured motif and recalls a combinatorial algorithm for extracting structured motifs given certain constraints they must satisfy. Section 7.3 presents a way of approximating the probability of occurrence of structured motifs. Finally, Section 7.4 presents an application of the method to obtain the statistical significance of the occurrences of some particular structured motifs in sequences coming just upstream from genes of *B. subtillis* (first dataset) and *E. coli* (second dataset). The particular structured motifs are candidate promoters. The results are compared, both in terms of performance and accuracy, to statistics obtained by using simulation.

7.2 Structured motifs

7.2.1 Basic definitions

Let Σ be the alphabet of nucleotides, Σ^+ the set of all non-empty words on the Σ alphabet. We denote by

$$\mathbf{m} = [\mathcal{V}(d_1 : d_2)\mathcal{W}]$$

a structured motif composed of two boxes (sets of words) $\mathcal{V} \in \Sigma^+$ and $\mathcal{W} \in \Sigma^+$ separated from one another by a distance between d_1 and d_2, for d_1 and d_2 two non negative integers that are fixed. In the particular case where $\mathcal{V}$ and $\mathcal{W}$ are reduced to unique words $\mathbf{v}$ and $\mathbf{w}$, the structured motif is simply denoted by

$$\mathbf{m} = [\mathbf{v}(d_1 : d_2)\mathbf{w}].$$

We now define what it means for a structured motif $\mathbf{m}$ to *occur* in a sequence $\mathbf{S}$ of length ℓ. We must first define what it means for a box $\mathcal{V}$ of a motif $\mathbf{m}$ to occur in $\mathbf{S}$. $\mathcal{V}$ is said to occur in $\mathbf{S}$ if $\mathbf{S}$ contains at least one occurrence of a word $\mathbf{x}$ of $\mathcal{V}$, that is, if there exists $\mathbf{S}', \mathbf{S}'' \in \Sigma^+ \cup \emptyset$ such that $\mathbf{S} = \mathbf{S}'\mathbf{x}\mathbf{S}''$. In our case, we are interested in $\mathcal{V} = \mathcal{N}_e(\mathbf{v})$ and $\mathcal{W} = \mathcal{N}_e(\mathbf{w})$ where $\mathbf{v}$ and $\mathbf{w}$ are fixed words of Σ^+ and $\mathcal{N}_e$ denotes the e-neighborhood. The e-neighborhood of $\mathbf{v}$ is the set of all words that may be obtained from $\mathbf{v}$ by making at most e substitutions; $\mathbf{v}$ and all its e-neighbors have the same length denoted by $|\mathbf{v}|$. The set $\mathcal{N}_e(\mathbf{v})$ contains therefore all the words of $\Sigma^{|\mathbf{v}|}$ that are at a Hamming distance at most e from $\mathbf{v}$. A structured motif $\mathbf{m}$ is now said to *occur* in a sequence $\mathbf{S}$ if there is at least a pair of occurrences $(\mathbf{x}, \mathbf{x}')$ of $\mathcal{V} \times \mathcal{W}$, such that the distance between the end position of $\mathbf{x}$ and the start position of $\mathbf{x}'$ is between d_1 and d_2.

Finally, a structured motif $\mathbf{m}$ is said to be *valid* in relation to a set of sequences $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \cdots\}$ if, given a positive integer $q \leq |\mathcal{S}|$, $\mathbf{m}$ occurs in at least q sequences of $\mathcal{S}$ (it may occur more than once in a sequence). In the remainder, q is called the *quorum*.

7.2.2 Extracting structured motifs

The basic idea of the algorithm consists in efficiently spelling all valid structured motifs, that is, motifs satisfying the constraints (maximum error rate, distance between the two boxes and quorum). This is obtained by doing dynamic programming between a lexicographic tree $\mathcal{M}$ of the potential motifs with an index of the sequences. The index is given under the form of a generalized suffix tree $\mathcal{T}$ (Bieganski *et al.* (1994), Crochemore and Rytter (1994), Gusfield (1997)) and is built in a preprocessing step. The tree $\mathcal{M}$ of potential motifs is virtual: it is never built. Its traversal is done on the fly at the same time as a traversal of $\mathcal{T}$. The dynamic programming recurrence formulas are easy to obtain when substitutions only are allowed. They indicate how to find the occurrences of a motif $\mathbf{m}\alpha$, $\alpha \in \Sigma$, from the occurrences of $\mathbf{m}$. For ease of exposition, we assume that the structured motifs **m** sought are composed of two boxes, each of same length k, separated by a fixed distance d. Occurrences of **m** are nodes, possibly virtual (Marsan and Sagot (2000a), (2000b)) in $\mathcal{T}$. They are identified by a pair (v, e_v) where v is the node and e_v is the number of substitutions between the motif **m** and the string labeling the path from the root of $\mathcal{T}$ to v (such path is noted s_v). Then:

Algorithm 7.2.1 *Let e be the maximum number of substitutions allowed, and let us note $ancestor_\lambda(v)$ the ancestor at distance λ of node v in $\mathcal{T}$ (when λ is 1, such ancestor is the father of v – it is denoted by $father(v)$). The pair (v, e_v) is an occurrence of motif $\mathbf{m}\alpha$ if:*

case $|s_v| \leq k$ **or** $|s_v| \geq (k+d)$**:** *the pair $(ancestor_\lambda(v), e_{ancestor_\lambda(v)})$ is an occurrence of* **m** *where λ is d if $|s_v| = (k+d)$, 1 otherwise and:*

- *there exists an arc leaving $father(v)$ with a label starting with the letter α (in this case e_v will be equal to $e_{father(v)}$);*
- *such an arc does not exist, but $e_{father(v)} < e$ (in this case, e_v will be equal to $e_{father(v)} + 1$);*

case $|s_v| = (k+1)$**:** *the pair $(father(v), e_{father(v)})$ is an occurrence of* **m**.

In the last case, a jump is made in the tree directly to all the k^{th} descendants of $father(v)$. Such nodes represent a potential starting point for a second box of the motif being spelled. The jump may be performed in two different ways. It is not essential for the purposes of this paper to describe them. Details may be found in Marsan and Sagot (2000a) or (2000b).

It is important to observe that the algorithm is exact. It will exhaustively enumerate all the structured motifs that are valid given the parameters. The algorithm is also efficient in practice, as expected considering its theoretical complexity given in next section.

7.2.3 Complexity of the algorithm

Depending on how the jump is done in the tree $\mathcal{T}$ once a first box for a structured motif **m** has been identified, the time complexity of the algorithm may be $O(|\mathbf{S}|\nu_{2k+d_2}n^2(e,k))$ (tree $\mathcal{T}$ remains unchanged) or $O(|\mathbf{S}|\nu_k n^2(e,k) + |\mathbf{S}|\nu_{2k+d_2}n(e,k))$ (tree $\mathcal{T}$ is temporarily and locally modified) where $|\mathbf{S}|$ is the total number of sequences, ν_λ is the number of (possibly virtual) nodes at level λ in $\mathcal{T}$, and $n(e,k)$ is the number of e-neighbors of a k-letter word. Note that $\nu_\lambda \leq \nu_{\lambda+1} < \ell + |\mathbf{S}|$ where ℓ is the average length of the $|\mathbf{S}|$ sequences. The asymptotic space complexity is the same in both cases, $O(|\mathbf{S}|^2\ell)$, although the second approach has a higher constant.

7.3 Occurrence probability of a structured motif

In this section, we deal with the problem of calculating the probability for a structured motif $\mathbf{m} = [\mathcal{V}(d_1 : d_2)\mathcal{W}]$ to occur in a sequence **S** of length ℓ. This problem is equivalent to the waiting time problem raised in the original paper of Blom and Thorburn (1982). The exact distribution of the waiting time before a given word occurs in a random sequence is given in the Markovian case in Robin and Daudin (1999) and (2001), together with the distribution of the distance between successive occurrences. All these distributions are characterized by their generating functions.

From now on we shall assume that the sequence $\mathbf{S} = (S_1, S_2, \ldots, S_\ell)$ is a stationary Markov chain with transition probabilities $\pi(a,b) = \Pr\{S_{x+1} = b \mid S_x = a\}$, where $a, b \in \Sigma$. Each letter S_x of $\mathbf{S}$ is distributed according to the stationary distribution μ of the Markov chain; μ is defined by $\mu(b) = \sum_{a\in\Sigma} \mu(a)\pi(a,b)$ for all $b \in \Sigma$.

Overlapping structure. One of the most important results given in the papers cited above is that the distribution of the waiting time of a word $\mathbf{v}$ depends on the overlapping structure of the word, i.e. on the possibility for $\mathbf{v}$ to overlap itself. In the case of complex motifs like those studied here, this overlapping structure cannot be completely described so the exact distribution of the waiting time can not be provided. However, we are able to take into account the overlapping structures of the words $\mathbf{v}$ and $\mathbf{w}$.

7.3.1 Generalized geometric approximations

We want to calculate the probability $\gamma(\ell) = \Pr\{\mathbf{m} \in \mathbf{S}\}$. We use the convention that $\{\mathbf{m} \text{ at } x\}$ means that a word of $\mathcal{V}$ starts at x and a word of $\mathcal{W}$ starts y positions later with $|\mathbf{v}| + d_1 \leq y \leq |\mathbf{v}| + d_2$, and $\{\mathbf{m} \text{ not at } x\}$ means either that no word of $\mathcal{V}$ starts at x or that a word of $\mathcal{V}$ starts at x but no word of $\mathcal{W}$ starts y positions later with $|\mathbf{v}| + d_1 \leq y \leq |\mathbf{v}| + d_2$. This probability $\gamma(\ell)$ can be decomposed as follows:

$$\begin{aligned}
1 - \gamma(\ell) &= \Pr\{\mathbf{m} \text{ not at } 1, \mathbf{m} \text{ not at } 2, \ldots, \mathbf{m} \text{ not at } \ell - |\mathbf{m}| + 1\} \\
&= \Pr\{\mathbf{m} \text{ not at } 1\} \times \Pr\{\mathbf{m} \text{ not at } 2 \mid \mathbf{m} \text{ not at } 1\} \times \ldots \\
&\quad \times \Pr\{\mathbf{m} \text{ not at } \ell - |\mathbf{m}| + 1 \mid \mathbf{m} \text{ not at } 1, \ldots, \mathbf{m} \text{ not at } \ell - |\mathbf{m}|\}
\end{aligned} \tag{7.1}$$

where $|\mathbf{m}|$ is the maximal length of the motif (i.e. $|\mathbf{v}| + |\mathbf{w}| + d_2$). We propose to approximate $\Pr\{\mathbf{m} \in \mathbf{S}\}$ by considering the past only up to a fixed order.

Order 0. Approximating $\Pr\{\mathbf{m} \text{ not at } x \mid \mathbf{m} \text{ not at } 1, \ldots, \mathbf{m} \text{ not at } x-1\}$ by $\Pr\{\mathbf{m} \text{ not at } x\} = \Pr\{\mathbf{m} \text{ not at } 1\}$, we get the usual geometric approximation

$$\gamma(\ell) \simeq \gamma_0(\ell) = 1 - [1 - \mu(\mathbf{m})]^{\ell - |\mathbf{m}| + 1} \simeq 1 - \exp[-(\ell - |\mathbf{m}| + 1)\mu(\mathbf{m})] \tag{7.2}$$

where $\mu(\mathbf{m}) = \Pr\{\mathbf{m} \text{ at } x\}$. The last approximation is valid for a small $\mu(\mathbf{m})$.

Order 1. Approximating $\Pr\{\mathbf{m} \text{ not at } x \mid \mathbf{m} \text{ not at } 1, \ldots, \mathbf{m} \text{ not at } x-1\}$ by $\Pr\{\mathbf{m} \text{ not at } x \mid \mathbf{m} \text{ not at } x-1\}$ and denoting $\mu_1(\mathbf{m}) = \Pr\{\mathbf{m} \text{ at } x \mid \mathbf{m} \text{ not at } x-1\}$, we get

$$\begin{aligned}
\gamma(\ell) \simeq \gamma_1(\ell) &= 1 - [1 - \mu(\mathbf{m})][1 - \mu_1(\mathbf{m})]^{\ell - |\mathbf{m}|} \\
&= 1 - [1 - \mu(\mathbf{m})] \exp[-(\ell - |\mathbf{m}|)\mu_1(\mathbf{m})].
\end{aligned} \tag{7.3}$$

The last approximation is valid for a small $\mu_1(\mathbf{m})$.

Order k. Denoting $\mu_k(\mathbf{m}) = \Pr\{\mathbf{m} \text{ at } x \mid \mathbf{m} \text{ not at } x-1, \ldots, \mathbf{m} \text{ not at } x-k\}$ and $\mu_0(\mathbf{m}) = \mu(\mathbf{m})$, we define the approximation of order k:

$$\gamma(\ell) \simeq \gamma_k(\ell) = 1 - [1 - \mu_k(\mathbf{m})]^{(\ell - |\mathbf{m}| - k + 1)} \prod_{i=0}^{k-1} [1 - \mu_i(\mathbf{m})]. \tag{7.4}$$

We use the results on the exact distribution of the distance between words to calculate the probabilities $\mu_k(\mathbf{m})$. For the ease of the reader, we first show how to calculate the probability $\mu_k(\mathbf{m})$ when the boxes $\mathcal{V}$ and $\mathcal{W}$ are reduced to single words $\mathbf{v}$ and $\mathbf{w}$, i.e. when the number e of allowed errors is null (see Section 7.2.1). We then consider the case of multiple words, i.e. when $e > 0$.

In practice, we limit ourselves to the case $k \leq 1$ since this order seems good enough in our applications.

7.3.2 Positions and distances

We first have to define some random variables (r.v.'s). Let us consider the sets of words $\mathcal{V} = \{\mathbf{v}_i\}_{1 \leq i \leq I}$ and $\mathcal{W} = \{\mathbf{w}_j\}_{1 \leq j \leq J}$. According to Section 7.2.1, $\mathcal{V}$ (respectively $\mathcal{W}$) is the e-neighborhood of a word $\mathbf{v}$ (respectively $\mathbf{w}$), so all its elements have the same length $|\mathbf{v}|$ (respectively $|\mathbf{w}|$). Let us define the following r.v.'s:

- $X_{\mathcal{V}}$: the position of the first occurrence of any word of $\mathcal{V}$. This position is given by the position of the first letter of the word: $X_{\mathcal{V}} = \inf\{x : (S_x, \ldots S_{x+|\mathbf{v}|-1}) \in \mathcal{V}\}$.

- $X_{\mathbf{v}_i}(\mathcal{V})$: the position of the first occurrence of $\mathbf{v}_i$ if no other element of $\mathcal{V}$ has occurred before. Formally

$$X_{\mathbf{v}_i}(\mathcal{V}) = \begin{cases} X_{\mathcal{V}} & \text{if } \mathbf{v}_i \text{ is the first word of } \mathcal{V} \text{ to occur,} \\ \infty & \text{otherwise.} \end{cases}$$

  The distribution of $X_{\mathbf{v}_i}(\mathcal{V})$ is denoted $p_{\mathbf{v}_i}(x) = \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = x\}$.

- $Y_{\mathbf{v}_i,\mathcal{W}}$: the distance between the start of an occurrence of $\mathbf{v}_i$ and the first following occurrence of an element of $\mathcal{W}$. Let x be the position of (the start of) an occurrence of $\mathbf{v}_i$; we have $Y_{\mathbf{v}_i,\mathcal{W}} = \inf\{y \geq 0 : (S_{x+y}, \ldots S_{x+y+|\mathbf{w}|-1}) \in \mathcal{W}\}$.

- $Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W})$: the distance between an occurrence of $\mathbf{v}_i$ and the first following occurrence of $\mathbf{w}_j$ if no other element of $\mathcal{W}$ has occurred before. Formally

$$Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W}) = \begin{cases} Y_{\mathbf{v}_i,\mathcal{W}} & \text{if } \mathbf{w}_j \text{ is the first word of } \mathcal{W} \text{ to occur after } \mathbf{v}_i, \\ \infty & \text{otherwise.} \end{cases}$$

  This distance can be null, for example if $\mathcal{V}$ is reduced to one word of one letter (say a) and if $\mathcal{W}$ is reduced to one word w beginning with a. The distribution of $Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W})$ is denoted $q_{\mathbf{v}_i,\mathbf{w}_j}(y) = \Pr\{Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W}) = y\}$.

- $Y_{A,\mathbf{w}_j}(\mathcal{W})$: the distance between the random letter A and the first following occurrence of $\mathbf{w}_j$ if no other element of $\mathcal{W}$ has occurred before. The cumulative probability function of $Y_{A,\mathbf{w}_j}(\mathcal{W})$ depends on the distribution of A:

$$\Pr\{Y_{A,\mathbf{w}_j} \leq y\} = \sum_{a\in\Sigma} \Pr\{Y_{a,\mathbf{w}_j} \leq y\} \Pr\{A = a\}.$$

The distributions of all these r.v.'s are given in Robin and Daudin (2001). They can be calculated thanks to the recursive formulas given in appendix 7.A.1.

7.3.3 Calculation of μ and μ_1 for single words

We consider here the case where the sets $\mathcal{V}$ and $\mathcal{W}$ are respectively reduced to $\mathbf{v}$ and $\mathbf{w}$.

Calculation of $\mu(\mathbf{m})$. To calculate the probability $\mu(\mathbf{m})$, we have to consider only the occurrences of $\mathbf{w}$ starting between positions $|\mathbf{v}| + d_1$ and $|\mathbf{v}| + d_2$ after the start of $\mathbf{v}$. We decompose this probability according to all the possible values of the random letter $A(\mathbf{v})$ that occurs in position $|\mathbf{v}| + d_1 + 1$ after the start of $\mathbf{v}$:

$$\begin{aligned} \mu(\mathbf{m}) &= \mu(\mathbf{v}) \Pr\{Y_{A(\mathbf{v}),\mathbf{w}} \leq d_2 - d_1\} \\ &= \mu(\mathbf{v}) \sum_{a\in\Sigma} \pi^{d_1+1}(v_{|\mathbf{v}|}, a) \sum_{y\leq d_2-d_1} q_{a,\mathbf{w}}(y). \end{aligned} \tag{7.5}$$

where $\mu(\mathbf{v})$ is the probability that the word $\mathbf{v}$ occurs at a given position, namely, $\mu(\mathbf{v}) = \mu(v_1) \prod_{u=2}^{|\mathbf{v}|} \pi(v_{u-1}, v_u)$ where v_u is u-th letter of $\mathbf{v}$. The distribution of the letter $A(\mathbf{v})$ is given by $\Pr\{A(\mathbf{v}) = a\} = \pi^{d_1+1}(v_{|\mathbf{v}|}, a)$ where $\pi^d(a, b)$ is the transition probability from the letter a to the letter b in d steps, i.e. the element (a, b) of the power d of the transition matrix π.

In the case of a fixed distance d_1 $(= d_2)$ between $\mathbf{v}$ and $\mathbf{w}$, the equation (7.5) simply is:

$$\mu(\mathbf{m}) = \mu([\mathbf{v}(d_1)\mathbf{w}]) = \mu(\mathbf{v})\pi^{d_1+1}(v_{|\mathbf{v}|}, w_1)\frac{\mu(\mathbf{w})}{\mu(w_1)}.$$

In the general case, the probability in the right hand side of (7.5) can be calculated using the formulas given in Appendix 7.A.1.

Calculation of $\mu_1(\mathbf{m})$. To calculate $\mu_1(\mathbf{m}) = \Pr\{\mathbf{m} \text{ at } x \mid \mathbf{m} \text{ not at } x - 1\}$, we decompose the event $\{(\mathbf{m} \text{ at } x) \cap (\mathbf{m} \text{ not at } x - 1)\} = \{(\mathbf{m} \text{ at } 2) \cap (\mathbf{m} \text{ not at } 1)\}$ as

$$\big\{(X_\mathbf{v} = 2) \cup [(X_\mathbf{v} = 1) \cap (Y_{A(\mathbf{v}),\mathbf{w}} > d_2 - d_1)]\big\} \cap \{\mathbf{m} \text{ at } 2\}.$$

We then get

$$\begin{aligned} &\Pr\{(\mathbf{m} \text{ at } 2) \cap (\mathbf{m} \text{ not at } 1)\} \\ &= \Pr\{X_\mathbf{v} = 2\} \Pr\{Y_{A(\mathbf{v}),\mathbf{w}} \leq d_2 - d_1\} + \Pr\{X_\mathbf{v} = 1\} \Pr\{Y_{\mathbf{v},\mathbf{v}} = 1\} \Pr\{Y_{A(\mathbf{v}),\mathbf{w}} = d_2 - d_1\} \\ &= p_\mathbf{v}(2) \sum_{a\in\Sigma} \pi^{d_1+1}(v_{|\mathbf{v}|}, a) \sum_{y\leq d_2-d_1} q_{a,\mathbf{w}}(y) + p_\mathbf{v}(1) q_{\mathbf{v},\mathbf{v}}(1) \sum_{a\in\Sigma} \pi^{d_1+1}(v_{|\mathbf{v}|}, a) q_{a,\mathbf{w}}(d_2 - d_1) \end{aligned}$$

and therefore, finally

$$\mu_1(\mathbf{m}) = \Pr\{(\mathbf{m} \text{ at } 2) \cap (\mathbf{m} \text{ not at } 1)\}/[1 - \mu(\mathbf{m})].$$

7.3.4 Calculation of μ and μ_1 for multiple words

We now consider the general case of two boxes $\mathcal{V} = \{\mathbf{v}_i\}_{1 \le i \le I}$ and $\mathcal{W} = \{\mathbf{w}_j\}_{1 \le j \le J}$. We remind that all the elements of $\mathcal{V}$ (respectively $\mathcal{W}$) have the same length $|\mathbf{v}|$ (respectively $|\mathbf{w}|$). We hence consider motifs of the form

$$\mathbf{m} = [\mathcal{V}(d_1 : d_2)\mathcal{W}].$$

Typically, the sets $\mathcal{V}$ and $\mathcal{W}$ will be respectively of the form $\mathcal{N}_e(\mathbf{v})$ and $\mathcal{N}_e(\mathbf{w})$.

Calculation of $\mu(\mathbf{m})$. $\mu(\mathbf{m}) = \Pr\{\mathbf{m} \text{ at } 1\}$ can be calculated decomposing the event $\{\mathbf{m} \text{ at } 1\}$ as $\bigcup_i \bigcup_j \big\{(X_{\mathbf{v}_i} = 1) \cap (Y_{A(\mathbf{v}_i),\mathbf{w}_j}(\mathcal{W}) \le d_2 - d_1)\big\}$, therefore we get

$$\begin{aligned} \mu(\mathbf{m}) &= \sum_i \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = 1\} \sum_j \Pr\{Y_{A(\mathbf{v}_i),\mathbf{w}_j}(\mathcal{W}) \le d_2 - d_1\} \\ &= \sum_i p_{\mathbf{v}_i}(1) \sum_{a \in \Sigma} \pi^{d_1+1}(v_{i,|\mathbf{v}_i|}, a) \sum_j \sum_{y \le d_2 - d_1} q_{a,\mathbf{w}_j}(y) \end{aligned} \tag{7.6}$$

where $v_{i,u}$ is the u-th letter of $\mathbf{v}_i$.

Calculation of $\mu_1(\mathbf{m})$. As in the single word case, we decompose the event $\{(\mathbf{m} \text{ not at } 1) \cap (\mathbf{m} \text{ at } 2)\}$ and we get

$$\begin{aligned} &\Pr\{(\mathbf{m} \text{ not at } 1) \cap (\mathbf{m} \text{ at } 2)\} \\ &= \sum_i \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = 2\} \sum_j \Pr\{Y_{A(\mathbf{v}_i),\mathbf{w}_j}(\mathcal{W}) \le d_2 - d_1\} \\ &\quad + \sum_i \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = 1\} \sum_k \Pr\{Y_{\mathbf{v}_i,\mathbf{v}_k} = 1\} \sum_j \Pr\{Y_{A(\mathbf{v}_k),\mathbf{w}_j}(\mathcal{W}) = d_2 - d_1\} \\ &= \sum_i p_{\mathbf{v}_i}(2) \sum_{a \in \Sigma} \pi^{d_1+1}(v_{i,|\mathbf{v}_i|}, a) \sum_j \sum_{y \le d_2 - d_1} q_{a,\mathbf{w}_j}(y) \\ &\quad + \sum_i p_{\mathbf{v}_i}(1) \sum_k q_{\mathbf{v}_i,\mathbf{v}_k}(1) \sum_{a \in \Sigma} \pi^{d_1+1}(v_{k,|\mathbf{v}_k|}, a) \sum_j q_{a,\mathbf{w}_j}(d_2 - d_1). \end{aligned} \tag{7.7}$$

Special motifs. In the case where the number of admitted errors e is a global one, meaning that it is simultaneously defined on the words $\mathbf{v}$ and $\mathbf{w}$, the elements of $\mathcal{W}$ that are valid to complete the motif $\mathbf{m}$ depend on the element of $\mathcal{V}$ that occurred in the first box. Hence, the calculation of $\mu(\mathbf{m})$ and $\mu_1(\mathbf{m})$ are modified. For example, if we globally accept at most one error ($e = 1$), two cases have to be considered:

1. if $\mathbf{v}$ occurs exactly in the first box then any element of $\mathcal{W} = \mathcal{N}_1(\mathbf{w})$ is valid to complete $\mathbf{m}$;

2. if $\mathbf{v}$ occurs with one error in the first box then only $\mathbf{w}$ is valid to complete $\mathbf{m}$.

For this case, the formulas of $\mu(\mathbf{m})$ and $\mu_1(\mathbf{m})$ are given in appendix 7.A.2.

For an arbitrary global number of admitted errors e, $\mathcal{W} = \mathcal{N}_1(\mathbf{w})$ has to be replaced by $\mathcal{W} = \mathcal{N}_e(\mathbf{w})$ in case 1; Case 2 becomes: if $\mathbf{v}$ occurs with exactly $e' \leq e$ errors then only the elements of $\mathcal{N}_{e-e'}(\mathbf{w})$ are valid to complete $\mathbf{m}$.

7.3.5 Number of occurrences in a set of sequences.

In the application, we consider a set of sequences $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \cdots\}$ and we are interested in the number N of sequences of $\mathbf{S}$ that contain the motif $\mathbf{m}$. Since $\gamma(\ell_s)$ is the probability for the motif $\mathbf{m}$ to occur in a sequence of length ℓ_s, N is distributed like a sum of $|\mathcal{S}|$ independent Bernoulli variables of respective parameters $\gamma(\ell_s)$:

$$N \sim \sum_{s=1}^{|\mathcal{S}|} \mathcal{B}[1, \gamma(\ell_s)]. \tag{7.8}$$

The approximation of order k gives $N \approx \sum_{s=1}^{|\mathcal{S}|} \mathcal{B}[1, \gamma_k(\ell_s)]$ where $\gamma_k(\ell_s)$ is given by equation (7.4).

If all the sequences have the same length ℓ, then N has a binomial distribution: $\mathcal{B}[|\mathcal{S}|, \gamma(\ell)]$.

7.4 Application

7.4.1 Data sets

The purpose of the application presented now is not to exhibit new results but to evaluate the statistical approach introduced in this paper. This is done on data that is clean in the sense that we know the motifs it should contain.

The data consist in two sets of non coding sequences coming just upstream from genes in two well-studied bacterial organisms, *Escherichia coli* and *Bacillus subtilis*. The sets were obtained from Ozoline *et al.* (1998) and Helmann (1995). The transcriptional starting point of the genes have been experimentally determined, or, more rarely, the promoter itself. The sequences in both sets are therefore aligned (on the start of transcription) which enables an easy judgment "by eye" of the biological pertinence of the motifs found. The *E. coli* dataset contains 441 sequences between 60 and 80 bps; the *B. subtilis* dataset contains 131 of length 100bps (except 1 of 99 bps).

The algorithm described in section 7.2 provides a list of candidate promoters for each datasets. For a quorum $q = 4\%$ and a total maximal number of error $e = 1$ on the two boxes, 40 motifs are selected in the *E. coli* dataset and 564 in the *B. subtilis* dataset. We then have to determine their statistical significance, that is the probability these structured motifs occur in at least 4% of the sequences.

7.4.2 Calculations and simulations

For each candidate motif, we calculate according to (7.8) the expected number of sequences in the set containing the motif

$$\mathbb{E}_k(N) = \sum_{s=1}^{|\mathcal{S}|} \gamma_k(\ell_s)$$

for $k = 0, 1$. The corresponding p-value is calculated in the same way to select relevant motifs. For example, considering the motif $\mathbf{m} = \texttt{TTGACAC}(16 : 18)\texttt{TATAATA}$, we get $\gamma_1(100) = 1.4885\ 10^{-4}$. The expected number of sequences containing $\mathbf{m}$ is $131 \times \gamma_1(100) = 1.95\ 10^{-2}$. This motif appears actually in 6 sequences, so the corresponding p-value is

$$\Pr\{\mathcal{B}[131, \gamma_1(100)] \geq 6\} = 7.67\ 10^{-14}.$$

On the other hand, the expected numbers can be estimated using simulations: 2000 simulations give a satisfying precision (see the confidence intervals in Figure 7.1). However, the estimation of the p-values requires a huge number of simulations for an acceptable precision.

The simulations are performed using a first order Markov model. The parameters $\pi(a, b)$ are estimated on the whole set of sequences, and the $|\mathcal{S}|$ sequences are then simulated with respective lengths $(\ell_1, \ldots \ell_{|\mathcal{S}|})$.

7.4.3 Comparison

Figure 7.1 shows that the approximation of order 0 is biased for some motifs. For example, in the *E. coli* dataset, the motif $\texttt{TTTTTT}(5 : 7)\texttt{TTAATT}$ is expected to occur 2.75 times among the 441 sequences according to the order 0 approximation, while simulations give a value of 1.94. The same observation can be made in relation to the *B. subtilis* dataset. A careful analysis shows that all these motifs have a first box $\mathcal{V}$ with high probability of self-overlaps.

The approximation of order 1 corrects this bias and fits very well to the simulated values (see Figure 7.1). 7.8% of the motifs are outside the 95% confidence interval based on the simulated values.

Computation times. The computations were carried out on a Celeron 1Ghz, with an average time of 5.10^{-3} sec. per motif. Hence, the computation time is about 30 sec. for 5 000 motifs using the approximation of order 1 while simulations require 90 mn for the same problem.

7.4.4 Results

Table 7.1 gives the motifs with a p-value smaller than 10^{-16} using the approximation of order 1. This table is a practical result of our method. Applying the same threshold on the p-value with intervals (16 : 18), (17 : 19), (18 : 20) and (19 : 21), we get a list of motifs given in Table 7.2 that seems to agree with the known consensus for the site at -35 and

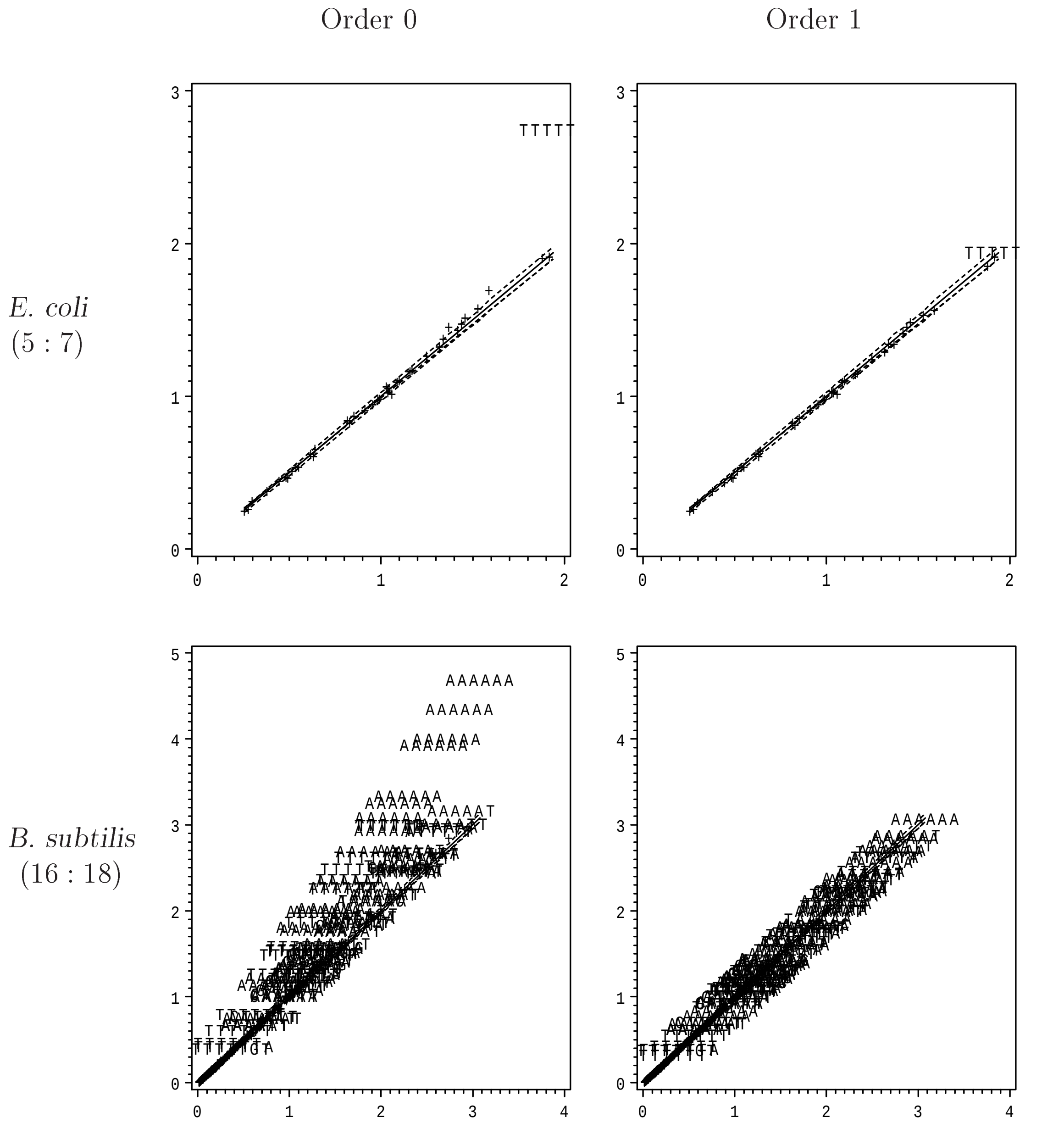


Figure 7.1: Comparison of approximated (y-axis) and simulated (x-axis) expected numbers of occurrences $\mathbb{E}(N)$ for two sets of structured motifs and approximation orders, 0 and 1. Motifs for which the approximation is far from the simulation for order 0 are marked with their first box $\mathbf{v}$.

the `TATA`-box (see Record *et al.* (1996)). It has to be pointed out that motifs obtained with different intervals may be redundant.

v	$(d_1 : d_2)$	**w**	observed N	$\mathbb{E}_1(N)$
`GTTGACA`	$(16:18)$	`ATATAAT`	7	$2.43\ 10^{-2}$
`GTTGACA`	$(16:18)$	`TATAATA`	8	$2.23\ 10^{-2}$
`TGTTGAC`	$(16:18)$	`TATAATA`	10	$2.12\ 10^{-2}$
`TTGACA`	$(16:18)$	`TTATAAT`	12	$1.53\ 10^{-1}$
`TTGACAA`	$(16:18)$	`TACAAT`	9	$9.82\ 10^{-2}$
`TTGACAA`	$(16:18)$	`TATAATA`	10	$5.07\ 10^{-2}$
`TTGACAG`	$(16:18)$	`TATAAT`	9	$7.12\ 10^{-2}$
`TTGACG`	$(16:18)$	`TATAAT`	11	$2.01\ 10^{-2}$

Table 7.1: Highly significant motifs (p-value $< 10^{-16}$) for the *B. subtilis* dataset when the length of the intervals is $(16 : 18)$.

Software availability. These formulas were implemented in a C program (for unix systems), which is available upon request at the following address :

`richard@genopole.cnrs.fr`.

Given parameters (*i.e.* transition matrix and lengths of sequences) and a list of motifs (with their respective number of occurences N) this program computes the expectation and the p-value of order 1 for each of them.

Bibliography

Bieganski, P., Riedl, J., Carlis, J. V. and Retzel, E. (1994). Generalized suffix trees for biological sequence data: applications and implementations. In *Proc. of the 27th Hawai Int. Conf. on Systems Sci.*, 35–44. IEEE Computer Society Press.

Blom, G. and Thorburn, D. (1982). How many random digits are required until given sequences are obtained ? *J. Appl. Prob.* **19** 518–531.

Crochemore, M. and Rytter, W. (1994). *Text Algorithms.* Oxford University Press.

Fu, C. J. (1996). Distribution of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica.* **6** 957–974.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press.

van Helden, J., Rios, A. F. and Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucl. Acids Res.* **28** 1808–1818.

v	$(d_1 : d_2)$	**w**	observed N	$\mathbb{E}_1(N)$
GTTGACA	$(16:18)$	ATATAAT	7	$2.43\ 10^{-2}$
GTTGACA	$(16:18)$	TATAATA	8	$2.23\ 10^{-2}$
TGTTGAC	$(16:18)$	TATAATA	10	$2.12\ 10^{-2}$
TTGACAA	$(16:18)$	TACAAT	9	$9.82\ 10^{-2}$
TTGACAA	$(16:18)$	TATAATA	10	$5.07\ 10^{-2}$
TTGACAG	$(16:18)$	TATAAT	9	$7.12\ 10^{-2}$
TTGACAA	$(17:19)$	ATAATAA	9	$6.97\ 10^{-2}$
TTGTTGA	$(17:19)$	TATAATA	8	$5.17\ 10^{-2}$
GTTGACA	$(17:19)$	ATAATAA	8	$3.09\ 10^{-2}$
GTTGACA	$(17:19)$	TATAATA	8	$2.19\ 10^{-2}$
CTTGACA	$(17:19)$	TATAAT	8	$6.04\ 10^{-2}$
TGTTGAC	$(17:19)$	TATAATA	12	$2.09\ 10^{-2}$
TGTTGAC	$(17:19)$	ATATAAT	7	$2.29\ 10^{-2}$
TTGTTGA	$(18:20)$	TATAATA	8	$5.09\ 10^{-2}$
GTTGACA	$(18:20)$	ATAATGA	7	$1.79\ 10^{-2}$
GTTGTTG	$(18:20)$	TATAATA	7	$2.53\ 10^{-2}$
TGTTGAC	$(18:20)$	ATAATAA	10	$2.90\ 10^{-2}$
TGTTGAC	$(18:20)$	ATACTA	7	$2.77\ 10^{-2}$
TGTTGAC	$(19:21)$	ATAATAA	10	$2.86\ 10^{-2}$
TGTTGAC	$(19:21)$	ATACTA	7	$2.73\ 10^{-2}$
TGTTGAC	$(19:21)$	TATAAT	10	$6.53\ 10^{-2}$
GTTGACT	$(19:21)$	ATAATA	8	$6.25\ 10^{-2}$

Table 7.2: Putative alignment of highly significant motifs (p-value $< 10^{-16}$) for the *B. subtilis* dataset when the length of the intervals is successively $(16 : 18)$, $(17 : 19)$, $(18 : 20)$ and $(19 : 21)$.

Helmann, J. D. (1995). Compilation and analysis of *Bacillus subtilis* α-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucl. Acids Res.* **23** 2351–2360.

Marsan, L. and Sagot, M.-F. (2000a). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comp. Biol.* **7** 345–362.

Marsan, L. and Sagot, M.-F. (2000b). Extracting structured motifs using a suffix tree – algorithms and application to promoter consensus identification. In *RECOMB'00. Proceedings of Fourth Annual International Conference on Computational Molecular Biology*, (R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman, ed.), 210–219. ACM Press.

Ozoline, O. N., Deev, A. A. and Arkhipova, M. V. (1998). Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase. *Nucl. Acids Res.* **25** 4703–4709.

Record, M. T., Reznikoff, W. S., Craig, M. L., McQuade, K. L. and Schlax, P. J. (1996). *Escherichia coli RNA polymerase σ^{70} promoters, and the kinetics of the steps of transcription initiation.* (F. C. Neidhardt, ed.), volume 1. ASM Press.

Robin, S. and Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36** 179–193.

Robin, S. and Daudin, J.-J. (2001). Exact distribution of the distances between any occurrences of a set of words. *Ann. Inst. Statist. Math.* **36 (4)** 895–905.

Appendix

7.A Formulas for probability calculation

7.A.1 Recursive formulas

The distribution of the positions $X_{\mathbf{v}_i}(\mathcal{V})$ and the distances $Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W})$ can be deduced from the results of Robin and Daudin (2001). These distributions are characterized by the matrix containing the generating functions of all the r.v.'s $X_{\mathbf{v}_i}$ and $Y_{\mathbf{v}_i,\mathbf{w}_j}$ for all $\mathbf{v}_i$ and $\mathbf{w}_j$. Unfortunately, the calculation of this matrix requires intensive computation in the case of numerous or long words. We propose here an alternative method based on recurrence formulas. We recall two notations and define a third one:

$$\begin{array}{rcl} p_{\mathbf{v}_i}(x) &=& \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = x\}, \\ q_{\mathbf{v}_i,\mathbf{w}_j}(y) &=& \Pr\{Y_{\mathbf{v}_i,\mathbf{w}_j}(\mathcal{W}) = y\}, \end{array} \qquad r_{\mathbf{v}_i,\mathbf{w}_j}(y) = \Pr\left\{ \begin{array}{c} \mathbf{w}_j \text{ occurs } y \\ \text{positions after } \mathbf{v}_i \end{array} \right\}.$$

The difference between $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$ and $r_{\mathbf{v}_i,\mathbf{w}_j}(y)$ is that, in $r_{\mathbf{v}_i,\mathbf{w}_j}(y)$, the occurrence of $\mathbf{w}_j$ at y is not necessarily the first one after $\mathbf{v}_i$. Remember that the position of a word is given by the position of its first letter.

The probabilities $r_{\mathbf{v},\mathbf{w}}(y)$'s are given for any words $(\mathbf{v}, \mathbf{w})$ by Lemma 2 of Robin and Daudin (2001):

$$\begin{aligned} r_{\mathbf{v},\mathbf{w}}(y) = \varepsilon_{\mathbf{v},\mathbf{w}}(|\mathbf{v}| - y) &\prod_{u=|\mathbf{v}|-y+1}^{|\mathbf{w}|} \pi(w_{u-1}, w_u)\mathbb{I}\{y < |\mathbf{v}|\} \\ &+ \pi^{(y-|\mathbf{v}|+1)}(v_{|\mathbf{v}|}, w_1)\frac{\mu(\mathbf{w})}{\mu(w_1)}\mathbb{I}\{y \geq |\mathbf{v}|\} \end{aligned} \tag{7.9}$$

where the indicator function $\mathbb{I}\{C\}$ equals 1 if C is true and 0 otherwise, and $\varepsilon_{\mathbf{v},\mathbf{w}}(u)$ is the indicator that the first u letters of $\mathbf{w}$ are identical to the last u letters of $\mathbf{v}$: $\varepsilon_{\mathbf{v},\mathbf{w}}(u) = \mathbb{I}\{(w_1 \dots w_u) = (v_{|\mathbf{v}|-u+1} \dots v_{|\mathbf{v}|})\}$. We use the convention $\prod_a^b = 1$ if $a > b$.

Recursive formulas for the positions. The probabilities $p_{\mathbf{v}_i}(x)$'s can be calculated according to Theorem 1 of Robin and Daudin (1999) in the single word case. The generating functions corresponding to the multiple word case are given in Robin and Daudin (2001), but not the recursive formula. We give it here:

$$p_{\mathbf{v}_i}(x) = \mu(\mathbf{v}_i) - \sum_{k=1}^{I}\sum_{z=1}^{x-1} p_{\mathbf{v}_k}(z) r_{\mathbf{v}_k,\mathbf{v}_i}(x - z). \tag{7.10}$$

The probabilities $p_{\mathbf{v}_i}(x)$'s can therefore be calculated thanks to the probabilities $r_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s using (7.9). Note that, in any case, $p_{\mathbf{v}_i}(1) = \mu(\mathbf{v}_i)$.

Recursive formulas for the distances. The probabilities $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s are given by Lemma 1 of Robin and Daudin (2001) only in the case where $\mathcal{V} = \mathcal{W}$. We generalize here this result to the case of different lists:

$$q_{\mathbf{v}_i,\mathbf{w}_j}(y) = r_{\mathbf{v}_i,\mathbf{w}_j}(y) - \sum_{k=1}^{J}\sum_{z=1}^{y-1} q_{\mathbf{v}_i,\mathbf{w}_k}(z) r_{\mathbf{w}_k,\mathbf{w}_j}(y-z). \tag{7.11}$$

The probabilities $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s can be calculated according to the following procedure:

1. calculate the $r_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s and $r_{\mathbf{w}_j,\mathbf{w}_k}(y)$'s using (7.9),

2. calculate recursively the $q_{\mathbf{w}_j,\mathbf{w}_k}(y)$'s according to (7.11),

3. calculate recursively the $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$'s according to (7.11).

The order of the recurrence formulas involved in this procedure is proportional to the distance y. Thus the computation time of $q_{\mathbf{v}_i,\mathbf{w}_j}(y)$ is in $O(y^2)$. In the problem we study here, the distances y are always very short ($y \leq 30$) so the computation time remains very reasonable.

7.A.2 Formulas for the probabilities μ and μ_1 for $e = 1$

We consider the case described at the end of section 7.3.4 where at most one error on both $\mathbf{v}$ and $\mathbf{w}$ is accepted. In this case $\mathcal{V} = \mathcal{N}_1(\mathbf{v})$ and $\mathcal{W} = \mathcal{N}_1(\mathbf{w})$. Both sets $\mathcal{V}$ and $\mathcal{W}$ are reordered in such a way that $\mathbf{v}_1 = \mathbf{v}$ and $\mathbf{w}_1 = \mathbf{w}$, so $\mathcal{V} \setminus \mathbf{v} = \{\mathbf{v}_2, \dots \mathbf{v}_I\}$ and $\mathcal{W} \setminus \mathbf{w} = \{\mathbf{w}_2, \dots \mathbf{v}_J\}$.

In the following, the term $(\mathcal{V})$ will be dropped for the event $X_{\mathbf{v}_i}(\mathcal{V}) = 1$ because $\Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = 1\}$ simply equals $\mu(\mathbf{v}_i)$ that does not depend on $\mathcal{V}$.

Calculation of $\mu(\mathbf{m})$. The element of $\mathcal{W}$ valid to complete $\mathbf{m}$ depends on the element of $\mathcal{V}$ the motif $\mathbf{m}$ starts with. Equation (7.6) has to be decomposed according to the element of $\mathcal{V}$ that occurs in the first box and we get

$$\begin{aligned}
\mu(\mathbf{m}) &= \Pr\left\{(X_{\mathbf{v}_1} = 1) \cap \left[\bigcup_{j\geq 1} \{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W}) \leq d_2 - d_1\}\right]\right\} \\
&\quad + \Pr\left\{\bigcup_{i>1} (X_{\mathbf{v}_i} = 1) \cap (Y_{A(\mathbf{v}_i),\mathbf{w}_1} \leq d_2 - d_1)\right\} \\
&= \Pr\{X_{\mathbf{v}_1} = 1\} \sum_{j\geq 1} \Pr\{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W}) \leq d_2 - d_1\} \\
&\quad + \sum_{i>1} \Pr\{X_{\mathbf{v}_i} = 1\} \Pr\{(Y_{A(\mathbf{v}_i),\mathbf{w}_1} \leq d_2 - d_1)\}.
\end{aligned}$$

Calculation of $\mu_1(\mathbf{m})$. As for the calculation of $\mu(\mathbf{m})$, each term of (7.7) has to be decomposed according to the element of $\mathcal{V}$:

$$
\begin{aligned}
&\Pr\{(\mathbf{m} \text{ not at } 1) \cap (\mathbf{m} \text{ at } 2)\} \\
&= \Pr\{X_{\mathbf{v}_1} = 1\} \Pr\{Y_{\mathbf{v}_1,\mathbf{v}_1} = 1\} \sum_j \Pr\{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W}) = d_2 - d_1\} \\
&\quad + \Pr\{X_{\mathbf{v}_1} = 1\} \sum_{i>1} \Pr\{Y_{\mathbf{v}_1,\mathbf{v}_i} = 1\} \Pr\{Y_{A(\mathbf{v}_i),\mathbf{w}_1}(\mathcal{W}) = d_2 - d_1\} \\
&\quad + \sum_{i>1} \Pr\{X_{\mathbf{v}_i} = 1\} \sum_{j>1} \Pr\{Y_{\mathbf{v}_i,\mathbf{v}_j} = 1\} \Pr\{Y_{A(\mathbf{v}_j),\mathbf{w}_1} = d_2 - d_1 + 1\} \\
&\quad + \sum_{i>1} \Pr\{X_{\mathbf{v}_i} = 1\} \Pr\{Y_{\mathbf{v}_i,\mathbf{v}_1} = 1\} \\
&\qquad \times \left[\Pr\{Y_{A^-(\mathbf{v}_1),\mathbf{w}_1}(\mathcal{W}) = d_2 - d_1 + 1\} + \sum_{j>1} \Pr\{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W}) \leq d_2 - d_1\} \right] \\
&\quad + \Pr\{X_{\mathbf{v}_1}(\mathcal{V}) = 2\} \sum_j \Pr\{Y_{A(\mathbf{v}_1),\mathbf{w}_j}(\mathcal{W}) \leq d_2 - d_1\} \\
&\quad + \sum_{i>1} \Pr\{X_{\mathbf{v}_i}(\mathcal{V}) = 2\} \Pr\{Y_{A(\mathbf{v}_i),\mathbf{w}_1} \leq d_2 - d_1\},
\end{aligned}
$$

where $A^-(\mathbf{v})$ is the (random) letter coming d_1 positions after the end of $\mathbf{v}$, i.e. the letter that just precedes $A(\mathbf{v})$.

Ackowledgement. The authors thank L. Marsan for providing his programs.

Troisième partie

Répartition de motifs et processus de Poisson

Chapitre 8

A compound Poisson model for word occurrences in DNA sequences

Robin S. (2002)

Journal of the Royal Statistical Society, C Series
Applied Statistics (to appear)

Abstract. We present a compound Poisson model describing the occurrences process of a set of words in a random sequence of letters. This model takes into account the frequency of the words and their overlapping structure. This model is compared to the Markov chain model in terms of fit and parsimony. Special attention is given to the detection of poor or rich regions. Several applications of this model are presented and a combination of the Markov and compound Poisson models is proposed in the end.

Keywords. compound Poisson process, DNA sequences, homogeneity checking, Markov chain, words occurrences

Contents

8.1 Introduction

Many DNA sequence analyses are based on the distribution of the occurrences of patterns having some special biological function. The most popular model in this domain is the Markov chain (MC) model that gives a description of the local behaviour of the sequence. One of the basic result obtained under this model is that the distribution of the occurrences of a word strongly depends on its overlapping structure (see the pioneer paper of Guibas and Odlyzko 1981). Exact results on the count (Régnier 2000) or the distances between occurrences (Robin and Daudin 1999, 2001) exist but are very demanding in terms of computation time, especially for high order MC or large sets of words.

On the other hand, the compound Poisson process gives a very natural description of the occurrences of words in a sequence and takes into account the possibility of overlap between occurrences. This model has been proposed several times to approximate the distribution of the number of occurrences in the MC framework (Godbole and Schaffner 1993, Schbath 1995, Geske *et al.* 1995, Reinert and Schbath 1998). Recently, Robin and Schbath (2001) have shown that, in many cases, this approximation is better than a Gaussian one.

We propose the use of a compound Poisson (CP) model to describe the whole occurrence process of a set of words (and not only their count). Such a model will be used to detect rich or poor regions (in terms of number of occurrences) in a genome. Hence, our purpose is different from the one of the papers cited in the previous paragraph: we are not looking for words with an unexpected frequency in a given sequence but for words with unexpected distribution along this sequence. Moreover, we shall work conditionally upon the observed frequency to avoid a systematic confusion between unexpected frequency and unexpected distribution. General tools for checking homogeneity have been proposed by Dembo and Karlin (1992) in a quite general framework and applied by Karlin and Macken (1991) with a simple Poisson model that does not take the overlapping structure into account, while a CP model does.

To illustrate our method we introduce two biological questions in connection with the distribution of some patterns along a whole genome. The first one concerns the CHI sites of the bacteria *H. influenzae* that are known to protect this genome by stopping its degradation performed by a particular enzyme. The degradation stops as soon as the enzyme meets a CHI site. Equivalent mechanisms exist in other species such as *E. coli*. Because of its favourable action, this pattern is expected to be very frequent in the whole genome, and actually is. Moreover, it is expected to be especially frequent in the regions in which alterations may be dramatically damaging for the organism. We will be interested in discovering such rich regions.

The second one concerns the bacteria *E. coli* that produces restriction enzymes corresponding to almost all palindromes made of 6 nucleotides (i.e. letters `a`, `c`, `g` or `t`). A restriction enzyme is able to cut the DNA molecule at specific positions called restriction sites. In our case, the restriction sites are the occurrences of all self-complementary words of length 6 (so-called 6-palindromes). These restriction enzymes and sites seem to be involved in a defense mechanism: *E. coli* produces restriction enzymes and avoids the corresponding restriction sites in its own genome. Phages (parasites) in which 6-palindromes are normally frequent are destroyed by these enzymes. Following the same idea as in the first example, we are intuitively interested here in poor regions in terms of 6-palindromes.

The model is presented in Section 8.2. Section 8.3 is devoted to distances between occurrences and to comparison with MC models. Two methods for checking homogeneities of the distribution of the occurrence are presented in Section 8.4. Section 8.5 presents the results for the two biological applications introduced above.

8.2 Compound poisson model

We are interested in the counting process $\{N(x)\}_{x\geq 0}$ of the occurrences of a set of m words $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ along a sequence $\mathbf{S}$ of length ℓ. To avoid any ambiguous definition of word occurrence, we assume that the elements of $\mathcal{W}$ are not included in each other. To take into account the possibility for the occurrences of these words to overlap, we assume that they may occur in clumps. We therefore propose the following model and notation.

8.2.1 Model and notation

The model is based on the two following assumptions:

A. Let $\{C(x)\}_{x\geq 0}$ denote the counting process of the clumps. $\{C(x)\}$ is a homogeneous Poisson process with constant intensity λ so $C(x) \sim \mathcal{P}(\lambda x)$:

$$\Pr\{C(x) = c\} = e^{-\lambda x}(\lambda x)^c/c!.$$

B. Let $(K_1, K_2, \ldots)$ denote the successive clump sizes. These sizes are independent from $\{C(x)\}$ and i.i.d. with common distribution

$$p(k) := \Pr\{K = k\}, \qquad \text{for } k \geq 1.$$

Under assumptions **A** and **B**, the counting process $\{N(x)\}_{x\geq 0}$ is a compound Poisson process defined by

$$N(x) = \sum_{c=1,\ldots,C(x)} K_c. \tag{8.1}$$

In the following we shall denote the distribution of $N(x)$ by $\mathcal{CP}(\lambda x; p)$. In the literature (Barbour *et al.* 1992b), the same distribution is sometimes denoted $\mathcal{CP}(\lambda_1, \lambda_2, \ldots)$ where $\lambda_k = \lambda x p(k)$.

In some special cases we will also need to describe the content of each clump. Let I_c be the random vector (of random length K_c) that contains the indices of the successive words occurring in the clump c: $I_c = (I_{c,1}, \ldots I_{c,K_c})$ where $I_{c,u} = i$ iff $\mathbf{w}_i$ occurs in position u in clump c. For example, if the clump c begins with $\mathbf{w}_2$, followed by $\mathbf{w}_4$ and ends with $\mathbf{w}_1$ then $I_c = (2, 4, 1)$. Consistently with assumption **B**, the contents are assumed to be independent from $\{C(x)\}$.

We assume here that the length of the words can be neglected with respect to the distance between non-overlapping occurrences so that overlapping occurrences are assumed to stand at the same position in the sequence: this assumption is only realistic for sufficiently rare words.

Probability generating function (pgf). We shall denote ϕ_X the pgf of a random variable X defined by $\phi_X(t) = \mathbb{E}(t^X)$. For example, it is well known that $\phi_{C(x)}(t) = \exp[\lambda x(t-1)]$. A classical property of the pgf is that

$$\phi'_X(1) = \mathbb{E}(X), \qquad \phi''_X(1) = \mathbb{E}(X^2) - \mathbb{E}(X). \tag{8.2}$$

8.2.2 Distribution of the clumps' size

Although the counting process can be described for a general distribution p of the clump size, we emphasise two particular distributions that will be useful for applications. The first one especially holds for one single word ($\mathcal{W} = \{\mathbf{w}\}$) and the second one is its generalisation to the multiple word case.

Geometric sizes. A simple way to describe a single word $\mathbf{w}$ clump occurrence is to assume that $\mathbf{w}$ has a probability a to overlap and that overlaps are independent. In this case, the clump size has a (shifted) geometric distribution $\mathcal{G}(1-a)$:

$$p_{\mathrm{g}}(k) := a^{k-1}(1-a).$$

Note that the $\mathcal{CP}(\lambda x, p_{\mathrm{g}})$ distribution only involves two parameters: the intensity λ of the occurrences of the word and the overlapping probability a. In the following, this version of the CP model is denoted CPg.

When applied to the multiple word case, this model assumes that all the words $\mathbf{w}_i$ have the same overlapping structure so that all of them have the same probability a to be overlapped. Surprisingly, this strong assumption holds for sets $\mathcal{W}$ with real biological interest such as the CHI site of *H. influenzae* (see Sections 8.1 and 8.5.1).

'Multiple' geometric sizes. When studying the occurrences of a set of words, one may generalise CPg model by considering the following probabilities (defined for $1 \le i, j \le m$):

$$\begin{aligned} a(i,j) &= \Pr\{\mathbf{w}_j \text{ overlaps } \mathbf{w}_i\}, \\ a(i,0) &= \Pr\{\text{no element of } \mathcal{W} \text{ overlaps } \mathbf{w}_i\} = 1 - \sum_{j=1..m} a(i,j). \end{aligned}$$

The sequence of the words within a clump is a Markov chain over the state space $\{0, 1, \ldots, m\}$ with transition matrix

$$\mathbf{A} = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{A}_0 & \mathbf{A}^* \end{pmatrix},$$

where $\mathbf{0}$ is a row-vector with m null coordinates, $\mathbf{A}_0$ the column-vector $[a(1,0)\ldots a(m,0)]'$ and $\mathbf{A}^*$ the $(m \times m)$ matrix with general term $a(i,j)$ $(1 \le i, j \le m)$. The absorbing state $\{0\}$ denotes the end of the clump.

In this model, a content $I = (I_1, \ldots, I_K)$ is a sample path of this Markov chain that stops when the state $\{0\}$ is reached. Denoting by $\theta(i)$ the probability that a clump starts with the word $\mathbf{w}_i$ and $\mathbf{\Theta} = [\theta(1) \ldots \theta(m)]$, we get

$$\Pr\{I = (i_1, \ldots, i_k)\} = \theta(i_1) \left[\prod_{u=2..k} a(i_{u-1}, i_u) \right] a(i_k, 0). \tag{8.3}$$

The (random) size K of the clump is the Markov chain sojourn time in the subspace $\{1, \ldots, m\}$ before it reaches $\{0\}$ so, denoting p_{m} the distribution of K (where 'm' stands for 'multiple'),

$$p_{\mathrm{m}}(k) := \mathbf{\Theta}(\mathbf{A}^*)^{k-1}\mathbf{A}_0$$

and the pgf of K is

$$\phi_K(t) = t\mathbf{\Theta}(\mathbf{I} - t\mathbf{A}^*)^{-1}\mathbf{A}_0, \tag{8.4}$$

where $\mathbf{I}$ denotes the $(m \times m)$ identity matrix. The mean and variance of K can be derived using (8.2). Denoting by $\mathbf{D} = (\mathbf{I} - \mathbf{A}^*)^{-1}$,

$$\mathbb{E}(K) = \mathbf{\Theta}\mathbf{D}^2\mathbf{A}_0, \quad \mathbb{V}(K) = \mathbf{\Theta}(\mathbf{D}^2 + 2\mathbf{A}^*\mathbf{D}^3 - \mathbf{D}^2\mathbf{A}_0\mathbf{\Theta}\mathbf{D}^2)\mathbf{A}_0.$$

In the following, this version of the CP model is denoted CPm.

Note that the geometric distribution is a particular case of the multiple geometric distribution where the $a(i,j)$'s are constant functions of i ($a(i,j) \equiv a(\bullet, j)$) so we may set $a = \sum_j a(\bullet, j)$.

8.2.3 Distribution of the count

According to the CP model (8.1), the distribution of the count $N(x)$ is given by

$$\Pr\{N(x) = n\} = \sum_{c=1..n} \Pr\{C(x) = c\} \sum_{k_1 + \cdots + k_c = n} \prod_{i=1..c} p(k_i). \tag{8.5}$$

for $n \geq 1$ and $\Pr\{N(x) = 0\} = \Pr\{C(x) = 0\} = e^{-\lambda x}$. Although this formula involves the sum over all combinations $(k_1, \ldots, k_c)$ such as $\sum_i k_i = n$, these probabilities can be very easily calculated for any distribution p of the clumps' size using the following proposition.

Proposition 1 *The distribution of the count $N(x)$ is given by*

$$\Pr\{N(x) = n\} = \exp(-\lambda x) \sum_{k=1..n} k\lambda x p(k) c_{n,k}$$

where, for $n \geq 1$, $c_{n,n} = 1/n$ and, for $k \geq 1$,

$$c_{n+k,n} = \sum_{j=1..k} \frac{j\lambda x p(j)}{n+k} c_{n+k-j,n}.$$

In addition, $\Pr\{N(x) = 0\} = \exp(-\lambda x)$.

Proof. This is a direct application of Theorem 3 and Lemma 7 of Barbour *et al.* (1992a) with $\lambda_k = \lambda x p(k)$. ■

Generating function. For a general distribution of the clump size, the pgf of the count $N(x)$ is given by

$$\phi_{N(x)}(t) = \phi_{C(x)}[\phi_K(t)] = \exp\{\lambda x[\phi_K(t) - 1]\} \tag{8.6}$$

Applying (8.2), we get $\mathbb{E}[N(x)] = \lambda x \mathbb{E}(K)$ and $\mathbb{V}[N(x)] = \lambda x[\mathbb{V}(K)+\mathbb{E}(K)^2] = \lambda x \mathbb{E}(K^2)$.

Geometric sizes. The $\mathcal{CP}(\lambda x, p_{\mathrm{g}})$ distribution is known has the Pólya-Aeppli or 'geometric Poisson' distribution (Johnson *et al.* 1992, p.378) with parameter λ for the Poisson part and $(1-a)$ for the geometric part. In this case, formula (8.5) reduces, for $n > 0$, to

$$\Pr\{N(x) = n\} = e^{-\lambda x} a^n \sum_{c=1..n} \frac{1}{c!} \binom{n-1}{c-1} \left[\frac{\lambda x(1-a)}{a}\right]^c,$$

where $\binom{a}{b}$ is the binomial coefficient $a!/[b!(a-b)!]$ and $\Pr\{N(x) = 0\} = \exp(-\lambda x)$. The first two moments of $N(x)$ are $\mathbb{E}[N(x)] = \lambda x/(1-a)$ and $\mathbb{V}[N(x)] = \lambda x(1+a)/(1-a)^2$.

'Multiple' geometric sizes. According to (8.6) and (8.4), the pgf of $N(x)$ in the CPm model is

$$\phi_{N(x)}(t) = \exp\{\lambda x[t\mathbf{\Theta}(\mathbf{I} - t\mathbf{A}^*)^{-1}\mathbf{A}_0 - 1]\}. \tag{8.7}$$

Still denoting $\mathbf{D} = (\mathbf{I} - \mathbf{A}^*)^{-1}$, the mean and variance of $N(x)$ are

$$\mathbb{E}[N(x)] = \lambda x \mathbb{E}(K) = \lambda x \mathbf{\Theta}\mathbf{D}^{-2}\mathbf{A}_0, \qquad \mathbb{V}[N(x)] = \lambda x \mathbf{\Theta}(\mathbf{I} + 2\mathbf{A}^*\mathbf{D})\mathbf{D}^2\mathbf{A}_0. \tag{8.8}$$

8.3 Distance between occurrences

Let us denote by $(X_1, X_2, \dots)$ the positions of the successive occurrences of the words in the sequence (with $X_{i+1} = X_i$ in case of overlap) and $(Y_1, Y_2, \dots)$ the distances between these occurrences ($Y_i = X_{i+1} - X_i$). These distances provide essential information about the homogeneity of the occurrences along the sequence. They can be generalised with the r-scans (Dembo and Karlin 1992):

$$Y_i^r = X_{i+r} - X_i = \sum_{i \le j \le i+r-1} Y_j.$$

8.3.1 Distribution of the distances

In the CP model (8.1), a distance Y can be either null, in case of overlap or non-negative when it separates two clumps. Because the clumps occur according to a Poisson process, the distance between two clumps has an exponential distribution with mean $1/\lambda$ denoted $\mathcal{E}(\lambda)$. Thus, Y has a mixture distribution:

$$Y \sim \pi\delta\{0\} + (1-\pi)\mathcal{E}(\lambda), \tag{8.9}$$

where $\delta\{0\}$ denotes the Dirac mass at 0.

In the general case, the probability π can depend either on the position in the clump or on the word that occurs. Thus, in the general case, the Y_i's are neither independent nor identically distributed.

Geometric case. In the CPg model, the probability to observe an overlap is constant and equal to a. Consequently, the distances are i.i.d. with common distribution (8.9), replacing π by a. The cumulative distribution function (cdf) of Y is

$$F(y) = 1 - (1-a)\exp(-\lambda y); \tag{8.10}$$

and its moments are $\mathbb{E}(Y) = (1-a)/\lambda$ and $\mathbb{V}(Y) = (1-a^2)/\lambda^2$.

The distribution of the r-scans Y^r can be obtained by simple convolution:

$$F_r(y) = \sum_{j=0..r} \binom{r}{j} a^j (1-a)^{r-j} G(y; \lambda, r-j) \tag{8.11}$$

where $G(y;\lambda,r)$ is the cdf of the $\gamma(\lambda,r)$ distribution: $G(y;\lambda,r) = \Gamma(y;\lambda,r)/\Gamma(\infty;\lambda,r)$, denoting $\Gamma(y;\lambda,r) = \int_0^y e^{-\lambda u} u^{r-1} \mathrm{d}u$ the incomplete Gamma function that can be easily calculated with many standard software packages. The mean and variance of the r-scan follow: $\mathbb{E}(Y^r) = r\mathbb{E}(Y)$, $\mathbb{V}(Y^r) = r\mathbb{V}(Y)$.

Multiple geometric case. When the clump size has the multiple geometric distribution, the distribution of the distance Y_i depends on the word $\mathbf{w}_i$ occurring at position X_i. In this case, the distribution of Y_i is the same as (8.9), replacing π by $\sum_{j=1..m} a(i,j) = 1 - a(1,0)$. In this case, the distances Y_i's are not i.i.d.

8.3.2 Parameters estimates

Likelihood of the process. Because the contents of the clumps $\{I_c\}$ are independent of their counting process $\{C(x)\}$, the complete log-likelihood of the occurrences of the words of $\mathcal{W}$ in a sequence of length ℓ is

$$\log \mathcal{L}[\{N(x)\}_{0\le x\le \ell}; \{I_c\}_{c=1..C(\ell)}] = \log \mathcal{L}[\{C(x)\}_{0\le x\le \ell}] + \sum_{c=1..C(\ell)} \log \mathcal{L}(I_c). \tag{8.12}$$

Because $\{C(x)\}$ is a Poisson process, $\log \mathcal{L}[\{C(x)\}_{0\le x\le \ell}] = C(\ell)\log(\lambda) - \lambda\ell$ and the maximum likelihood estimates (MLE) of λ (that only depends on $\log \mathcal{L}[\{C(x)\}])$ is $\hat{\lambda} = C(\ell)/\ell$.

The last term of (8.12) can be conditioned by the sizes of the clumps K_c: $\log \mathcal{L}(I_c) = \log \mathcal{L}(K_c) + \log \mathcal{L}(I_c \mid K_c)$ where $\log \mathcal{L}(K_c) = \log p(K_c)$. The log-likelihood of the process $\{N(x)\}$ only depends on the process $\{C(x)\}$ and on the clump sizes $\{K_c\}$ so

$$\log \mathcal{L}[\{N(x)\}_{0\le x\le \ell}] = \log \mathcal{L}[\{C(x)\}_{0\le x\le \ell}] + \sum_{c=1..C(\ell)} \log p(K_c). \tag{8.13}$$

In the case of one single word, $\log \mathcal{L}(I_c \mid K_c)$ is null for every c since the contents of each clump is entirely described by its length so (8.12) and (8.13) are equal.

General case. In the absence of a specific model, the estimation of the distribution p only involves $\sum_c \log \mathcal{L}(K_c) = \sum_{k\ge 1} \#\{\text{clumps of size } k\} \log p(k)$ and the estimate distribution $\hat{p}$ is the empirical distribution of the sizes of the clumps.

Geometric case. In the CPg model, the clump size is geometrically distributed so $\sum_c \log \mathcal{L}(K_c) = [N(\ell) - C(\ell)] \log(a) + C(\ell) \log(1-a)$ and the MLE of a is $\hat{a} = [N(\ell) - C(\ell)]/N(\ell)$.

Multiple geometric case. In the CPm model, the estimation of $\mathbf{\Theta}$ and $\mathbf{A}$ require to consider the contents $\{I_c\}$. According to (8.3), the log-likelihood of the contents $I_c = (i_{c,1}, \ldots, i_{c,K_c})$ is $\log \mathcal{L}(I_c) = \log[\theta(i_{c,1})] + \sum_{u=2..K_c} \log[a(i_{c,u-1}, i_{c,u})] + \log[a(i_{c,K_c}, 0)]$. Denoting $N_{ij}(\ell)$ the number of overlaps of $\mathbf{w}_j$ on $\mathbf{w}_i$, $N_i(\ell)$, the total number of occurrences of $\mathbf{w}_i$ and $C_i(\ell)$ the number of clumps beginning with $\mathbf{w}_i$, the MLE of $a(i,j)$ and $\theta(i)$ are

$$\hat{a}(i,j) = N_{ij}(\ell)/N_i(\ell), \qquad \hat{a}(i,0) = 1 - \sum_{1 \le j \le m} \hat{a}(i,j), \qquad \hat{\theta}(i) = C_i(\ell)/C(\ell).$$

Testing CPg versus CPm. In the multiple word case, when considering the geometric model CPg as a restricted version of the multiple geometric CPm, we assume that $a(i,j) \equiv a(\bullet, j)$. In this case, we get for $0 \le j \le m$, $\hat{a}(\bullet, j) = N_{+j}(\ell)/N(\ell)$ where $N_{+j} = \sum_{i=1..m} N_{ij}$. This estimate is consistent with $\hat{a}$ since $\hat{a} = \sum_{j=1..m} \hat{a}(\bullet, j)$.

We may test the hypothesis of homogeneous overlapping structures (i.e. for all $i \in \{1..m\}$, $a(i,j) = a(\bullet, j)$), using the likelihood ratio test. An application of this test is given in Section 8.5.1.

8.3.3 Comparison with markov models

Many statistical analysis of DNA sequences are based on Markov model. In these applications, the sequence is considered as a Markov chain of order d (MCd) on the alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. MC0 model involves 3 independent parameters, MC1 model involves 12 and MCd involves $3 \times 4^{d-1}$. We propose here a brief comparison of the CP and MC models.

Count. The exact distribution of the count $N(\ell)$ is given by Régnier (2000) under any model MCd. In the same framework, Schbath (1995) proposes a compound Poisson approximation of the distribution of $N(\ell)$ that is consistent with CP model presented here. Denoting $\mu(\mathbf{w})$ the probability for the word $\mathbf{w}$ to occur at a given position is the sequence and a the self-overlapping probability of $\mathbf{w}$ (these parameters can be explicitly calculated under any MCd model), the distribution of $C(\ell)$ can be approximated by a Poisson distribution with parameter $\lambda = \mu(\mathbf{w})(1-a)$. The approximate distribution of $N(\ell)$ is then $\mathcal{CP}(\lambda, p_{\mathrm{g}})$ where p_{g} is the geometric distribution $\mathcal{G}(1-a)$.

Other approximations for the distribution of $N(\ell)$ exist (see for example Prum *et al.* 1995 or Kleffe and Borodovsky 1992) but the general comparison made by Robin and Schbath (2001) shows that the compound Poisson approximation works surprisingly well and is the best in almost every situation, except for very frequent words.

Distances. Exact distribution of the occurrence positions and of the distances between occurrences are given by Robin and Daudin (1999) for one word and Robin and Daudin (2001) for a set of words in MC1 (and MC0) model. Even for short words, such as 4-nucleotides, the distribution (8.9) is very close to the exact distribution under the MC1 model (not shown here). However, the quality of the CP model is especially obvious for longer (and rarer) words such as 8-nucleotides.

CP model has a special advantage when the number of word occurrences is far from expected under the MC model. For example, there are 762 CHI sites ($\mathbf{w} = \texttt{gctggtgg}$) in the genome *E. coli* while the expected count under MC1 is 91. The log-likelihood of the occurrences of this word is –8 339 under MC1 and –7 402 under CPg. A natural way to take into account the observed frequency of the word in the sequence is to increase the order of the MC. In the case of the CHI site, one would need a MC7 to take completely its frequency into account but this model would involve around 50 000 parameters while the CP model has only 2.

Note that the number of parameters is a severe criterion for the MC7 model since its 50 000 parameters provide a modelling of the whole sequence of letters. Actually, few of them are really useful to describe the occurrences of one single word: we just need to be able to calculate the stationary probability for the word to occur and its overlapping probabilities.

Anyway, the CP model is a parsimonious way to check the homogeneity of the distribution of the occurrences along the sequence, independently from the total number of occurrences.

Overlapping probabilities. Despite MC models sometimes being inaccurate to describe the entire occurrences process, they can be very useful to describe their local behaviour. In particular, the overlapping probabilities $a(i, j)$ can be calculated very simply using the results of Robin and Daudin (2001) considering that two words overlap if the distance separating their occurrences is smaller than their lengths. Using MLE in the CP model, the m^2 overlapping probabilities have to be estimated one by one; using MCd model, they can be calculated through the $3 \times 4^{d-1}$ transition probabilities. In case of large sets of words (e.g. when $m \geq \sqrt{3 \times 4^{d-1}}$), MC$d$ gives a more parsimonious way to estimate the $a(i, j)$'s. An example of such a combination of CP and MC models is presented in Section 8.5.2.

8.4 Checking homogeneity

In this section, we address the problem of detecting heterogeneities in the distribution of the patterns along the sequence. For this purpose, the CP model described above can be considered as a null hypothesis to be tested. We shall consider two different ways to check homogeneities. The first one is based on the cumulated distances separating word occurrences (r-scans) while the second one considers the number $N(x) - N(x - y)$ of occurrences appearing inside an interval of fixed width y moving along the sequence with x. In both case, we shall be interested in extreme values; the two approaches are related since $\Pr\left\{\min_{i=1..N(\ell)-r}(Y_i^r) < y\right\} = \Pr\left\{\sup_{y \leq x \leq \ell}[N(x) - N(x - y)] > r\right\}$.

8.4.1 r-scans

The simple underlying idea here is that the scans are shorter in region in which the words of $\mathcal{W}$ are more frequent. The choice of r is related to the size of the regions one wants to check. When analysing a whole sequence, we are generally interested in the 'richest' (or the 'poorest') region. We therefore need to know the distribution of the number $R = R(r, y)$ of r-scans shorter than a given threshold y: $R = \sum_{i=1..N(\ell)-r} I\left\{Y_i^r \leq y\right\}$ where $I\{A\}$ is 1 if A is true and 0 otherwise.

This distribution is not known in the general case but Theorem 1 of Dembo and Karlin (1992) gives an approximation of it (conditionally to the observed number of occurrences $n(\ell)$) based on the Chen-Stein method (Barbour *et al.* 1992). If the distances Y_i are i.i.d. with common cdf F (F_r denoting the cdf of the r-scans), the total variation distance between R and the Poisson variable Z with mean $\mathbb{E}(R) = [n(\ell)-r]F_r(y)$ is bounded:

$$d_{\mathrm{TV}}(R, Z) \leq \left(1 - e^{-[n(\ell)-r]F_r(y)}\right)\left[(2r-1)F_r(y) + 2\sum_{k=1..r-1} F_k(y)\right]. \tag{8.14}$$

An equivalent result exists for the number of r-scans greater that y. An application of this is that, for small y, $\Pr\{R = 0\}$ can be approximated by $1 - e^{-[n(\ell)-r]F_r(y)}$. In the case of circular sequences (or genomes), there are $n(\ell)$ r-scans so $\mathbb{E}(R) = n(\ell)F_r(y)$.

The case where the distances $\{Y_i\}$ are i.i.d. exponentially distributed $\mathcal{E}(\lambda)$ – corresponding here to the non-overlapping case – is studied in Karlin and Macken (1991). This result is generalised here to the overlapping case in the Poisson geometric model (thanks to the i.i.d. property): the parameter of the Poisson approximation and the term between brackets in (8.14) can then be calculated using the cdf F_r given in (8.11).

8.4.2 Moving intervals

Another indicator of the homogeneity of word occurrences along the sequence is the count $N(x) - N(x-y)$ in all intervals of fixed width y for $y \leq x \leq \ell$. The probability that a given interval contains more than r occurrences is $\Pr\{N(y) \geq r\}$ and can be calculated using Proposition 1.

When considering the whole sequence, we are interested in the 'richest' region of the sequence. We therefore need to calculate the probability that, in a sequence of length ℓ, at least one interval contains more than r occurrences. We suggest here to use the 'Poisson clumping heuristic' first proposed by Aldous (1989). Let us consider the counting process $M(x)$ that can only jump when clumps occur:

$$\Delta M(x) = \Delta C(x) \times I\{N(x) - N(x-y) \geq r\},$$

where $\Delta M(x) = M(x) - M(x^-)$, and respectively for $\Delta C(x)$. If $M(x)$ is restricted to $y \leq x \leq \ell$, the expected count $\mathbb{E}M(\ell)$ is $(\ell - y)\tau(y, r)$ where $\tau(y, r) = \Pr\{N(y) \geq r\}$. The idea of this heuristic is that, for r 'sufficiently large', the process $\{M(x)\}$ is approximately a Poisson process with intensity $\tau(y, r)$ so that, in particular,

$$\Pr\{M(\ell) > 0\} \simeq 1 - e^{-(\ell-y)\tau(y,r)}. \tag{8.15}$$

In case of circular genomes, $M(x)$ is defined for $0 \leq x \leq \ell$ so $(\ell - y)$ is replaced by ℓ.

Bounds of the Poisson approximation. Although many articles, such as those of Barbour and Brown (1992ab) or Brown and Greig (1996), deal with Poisson approximations for point processes, explicit bounds for the distance between the distribution of $M(\ell)$ and $\mathcal{P}(\tau_{y,r})$ seem very difficult to derive. Explicit bounds only exist for the non-overlapping case, e.g., when the words occur according to a Poisson process (Janson 1984, Barbour *et al.* 1992b) or for other special processes such as those with internal history or with conditionally independent increments .

Theorem 2.6 of Barbour and Brown (1992b) gives a general bound for Poisson approximation of point processes. Denoting $\{Z(x)\}$ the Poisson process with constant intensity $\tau(y,r)$, the total variation distance between $\{M(x)\}$ and $\{Z(x))\}$ is bounded:

$$d_{\mathrm{TV}}(M,Z) \leq \int_y^\ell \mathbb{E}||M-(M_x-\delta\{x\})||\tau(y,r)\mathrm{d}x.$$

where $(M_x-\delta\{x\})$ is the reduced Palm process at point x. There is an additional term to this bound for any other Poisson process so that, in terms of total variation distance, $\{Z(x)\}$ is optimal. Unfortunately, the expected norm $\mathbb{E}||M-(M_x-\delta\{x\})||$ seems rather difficult to derive.

8.5 Case studies

8.5.1 Distribution of the CHI sites of *H. influenzae*

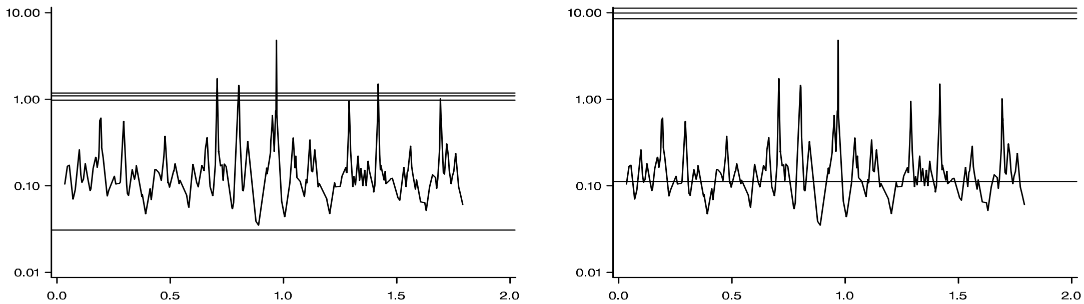
The CHI pattern in the genome of *H. influenzae* is a set of $m=4$ words that only differ by their second letter: $\mathcal{W}=\{\mathtt{gatggtgg}, \mathtt{gctggtgg}, \mathtt{ggtggtgg}, \mathtt{gttggtgg}\}$. There are 223 occurrences of this pattern, in 215 clumps. We propose here to study the homogeneity of its occurrences along the 1 903 356 bps of the whole genome.

Choice of the CP model. All the words of $\mathcal{W}$ can overlap each other using the sub-words (g) and, in addition, the word $\mathbf{w}_3=(\mathtt{ggtggtgg})$ can overlap all the $\mathbf{w}_i$'s (including itself) with the sub-words (gg) and (ggtgg). This peculiar overlapping structure allows to use the p_{g} distribution. To test the CPg model versus the more general CPm, we use the likelihood ratio test (both models use the MLE estimates of all the parameters). Under CPm, the estimate of the transition matrix $\mathbf{A}^*$ is

$$\hat{\mathbf{A}}^* = 10^{-2}\begin{pmatrix} 0 & 0 & 3.57 & 0 \\ 0 & 0 & 5.36 & 0 \\ 0 & 0 & 1.32 & 0 \\ 0 & 0 & 4.76 & 0 \end{pmatrix}.$$

Under CPg, $\mathbf{A}^*$ has a constant row estimated to $10^{-2}(0 \quad 0 \quad 3.59 \quad 0)$ (so $\hat{a}=3.59\%$). The two models gives the same MLE for $\boldsymbol{\Theta}$ and λ ($\hat{\lambda}=1.13\ 10^{-4}$). The log-likelihood of model CPm is $\log\mathcal{L}_{\mathrm{CP}m}=-2\ 490.4$ (for 20 functionally independent parameters) while the one of CPg is $\log\mathcal{L}_{\mathrm{CP}g}=-2\ 491.5$ (for 8 parameters). The likelihood ratio test statistic is $\chi^2=2.16$ with 12 degrees of freedom so we keep model CPg.

Checking homogeneity using r-scans. Under model CPg, the r-scans have common cdf given in (8.11). We use the estimate parameters $\hat{\lambda}$ and $\hat{a}$ given above. To give a comparison with the MC1 results given in Robin and Daudin (2001), we take $r = 3$.



MC1 CPg

Figure 8.1: Comparison of the intensity thresholds for the CHI site of *H. influenzae*. The x-axis gives the positions in Mbps, the y-axis gives the intensity r/Y^r multiplied by 10^3 (in log scale). The four horizontal line give, in ascending order, the theoretical mean intensity, the lower bound of the Chen-Stein approximation, the Chen-Stein threshold and the upper bound.

Figure 8.1 shows the 'intensity' r/Y^r calculated all along the genome. We see that, because the CHI sites are unexpectedly frequent with respect to model MC1, the observed intensity is greater than expected under MC1 all along the sequence. The frequency is intrinsically taken into account by model CP. A direct consequence of this is that the four (or five) significant peaks under MC1 turn out to be non significant under CP.

8.5.2 Homogeneity of the distribution of the 6-palindromes in the genome of *E. coli*

Palindromes are words that are complementary of themselves. Because most of them are restriction sites in *E. coli*, the palindromes of length 6 (and 8) are known to be avoided in its genome ($\ell = 4\ 638\ 868$ bps) . These patterns are sites of frailty of the genome and we are interested in their distribution along the sequence: they occur 54 724 times in 50 941 clumps (the biggest one contains 5 words).

Model selection. We first have to select a model. Table 8.1 describes the fits and log-likelihoods of seven models. The mean clump size is 1.07 and the variance is 0.077. The chi-square test on the distribution of the clump size rejects the p_{g} distribution for K. The likelihood ration test of CPg versus CPm also rejects model CPg. Because there are 64 6-palindromes, the CPm involves 1 272 independent parameters; 1 208 of them are elements of $\mathbf{A}$ corresponding to each possible overlap (there are less than $64 \times 64 = 4\ 096$ possible overlaps because many of them are impossible, such as (`tgatca`) over (`cccggg`)).

Here, we consider five versions of model CPm denoted CPm, CPm$_0$, CPm$_0^*$, CPm$_1$ and CPm$_1^*$. CPm uses the MLE estimates of λ, $\mathbf{A}$ and $\mathbf{\Theta}$, CPm$_0$ (resp. CPm$_1$) model

Model	df	$d_{\mathrm{TV}}(K)$	$\log \mathcal{L}(N)$	AIC(N)
CPe	*6*	–	*-294 503*	*589 017*
CPg	2	13.9 %	-306 660	613 323
CPm	1 272	1.9 %	-294 614	591 772
CPm_0	$m + A - 1 = 67$	1.3 %	-294 658	589 450
CPm_0^*	$A = 4$	0.9 %	-294 606	589 220
CPm_1	$m + A(A-1) = 76$	3.3 %	-294 844	589 840
CPm_1^*	$1 + A(A-1) = 13$	5.8 %	-295 062	590 149

Table 8.1: Comparison of seven models for the occurrences of the 6-palindromes in the genome of *E. coli*. CPe uses the empirical distribution of the clump size; CPm_0, CPm_0^*, CPm_1 and CPm_1^* are modified versions of CPm. df is the number of independent parameters, A is the size of the alphabet, $d_{\mathrm{TV}}(K)$ is the total variation distance between the estimate distribution of the clump size and the empirical one.

uses the MC0 (resp. MC1) estimate of $\mathbf{A}$ and CPm_0^* (resp. CPm_1^*) uses the MC0 (resp. MC1) estimates of both $\mathbf{A}$ and $\mathbf{\Theta}$ (end of Section 8.3.3). In these models, the Markov chain modelling is used locally, only to describe the contents of the clumps. Models CPm, CPm_0, CPm_0^*, CPm_1 and CPm_1^* have comparable fits for the clump size distribution (see $d_{\mathrm{TV}}(K)$ in Table 8.1) and for the global counting process $\{N(x)\}$ itself (see $\log \mathcal{L}(N)$). Model CPm_0^* gives a better log-likelihood than CPm (although CPm only uses MLE) but these two quantities are not directly comparable since the parameters estimated in these two models are not the same. AIC is a criterion that combines goodness of fit and parsimony. Corrected AIC (see Burnham and Anderson 1998) leads to the same conclusions. Results concerning the CPe model – that uses the *empirical* distribution of the clump size – are given as references.

According to AIC, we decide to use model CPm_0^* to study the homogeneity of the distribution of the palindromes. Compared to model MC0, this model only involves one more parameter (λ) that takes into account the global frequency of the clumps. Here, the Poisson modelling deals with the global distribution of the words and the Markov chain (actually Bernoulli) modelling describes their local behaviour.

Checking homogeneity. Because model CPg is rejected, the distance between occurrences can not be considered to be i.i.d. so the r-scan homogeneity checking can not be applied. As an application of the alternative method presented in Section 8.4.2, we consider intervals of length 10 000 bps: the richest one contains 185 palindromes and the poorest 73. The p-value associated to these two quantities are approximated using the Poisson heuristic (8.15) with the estimate distribution of $N(y)$ given by model CPm_0^*; we get

$$\Pr\left\{\min_{y \leq x \leq \ell} N([x-y;x]) \leq 73\right\} > 10\%, \quad \Pr\left\{\max_{y \leq x \leq \ell} N([x-y;x]) \geq 185\right\} < 5\%.$$

So, we observe no significantly poor region but one significantly rich region located between 2 460 567 bps and 2 461 566 bps after the replication origin. The biological interpretation of this last result is not clear. It could result from a recent horizontal transfer of this part of the genome coming from another species, but this hypothesis would need to be confirmed by additional data.

8.6 Conclusion and future works

We have presented a compound Poisson model that describes the occurrences of a possibly overlapping pattern under the hypothesis that it is uniformly distributed along the genome. This model appears to be fitting the real data well and is more parsimonious than Markov chain models. Yet, when the pattern is made of a large set of words, the number of overlapping probabilities to be estimated increases dramatically. In this case, Markov (even Bernoulli) models can be used to describe the overlaps. Hence a combination of Poisson (for the global behaviour) and Markov (for the local one) seems to be a good compromise.

Two methods to detect heterogeneity have been presented. The statistical properties r-scans are well known but this approach requires that all the words have the same probability to overlap, which is a strong assumption. In practice, this implies that the size of the clumps has a geometric distribution. The method based on moving intervals requires weaker assumption but provides only an approximation of the distribution of the associated statistic. Even bounds for this distribution are not available and this seems to be still an open problem.

In this model, the hypothesis of uniform distribution of the pattern along the sequence comes from the Poisson dimension. In many cases, heterogeneity structures are known a priori: coding versus non coding, `g-c` rich versus `a-t` rich regions, etc. An interesting development would be to take into account this prior heterogeneity with a (compound) heterogeneous Poisson process. Such a method would detect poor or rich regions that are not due to the prior heterogeneity.

Acknowledgment. The author thanks especially one of the referees for all his remarks and advices that really helped him to improve this paper.

Bibliography

Aldous, D. (1989). *Poisson Approximations via the Poisson Clumping Heuristic.* Springer-Verlag: New-York.

Barbour, A. D. and Brown, T. (1992a). The Stein-Chen method, point process and compensators. *Ann. Prob.* **20** 1504–1527.

Barbour, A. D. and Brown, T. (1992b). Stein's method and point process approximation. *Stoch. Proc. Appl.* **43** 9–31.

Barbour, A. D., Chen, L. H. Y. and Loh, W.-L. (1992a). Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Prob.* **20** 1843–1866.

Barbour, A. D., Holst, L. and Janson, S. (1992b). *Poisson Approximation.* Oxford-University Press.

Brown, T. C. and Greig, D. (1996). Poisson approximation for point processes via monotone coupling. *Ann. Appl. Prob.* **6** 545–560.

Burnham, K. P. and Anderson, R. A. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach.* Wiley.

Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes. *Ann. Appl. Prob.* **2 (2)** 329–357.

Geske, M. X., Godbole, A. P., Schaffner, A. A., Skolnick, A. M. and Wallstrom, G. L. (1995). Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Prob.* **32** 877–892.

Godbole, A. P. and Schaffner, A. A. (1993). Improved Poisson approximations for word patterns. *Adv. Appl. Prob.* **25** 334–347.

Guibas, L. J. and Odlyzko, A. M. (1981). Periods in strings. *J. Combinatorial Theory A.* **30** 19–42.

Janson, S. (1984). Bounds on the distributios of extremal values of a scanning process. *Stoch. Proc. Appl.* **18** 313–328.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions.* Wiley: New-York.

Karlin, S. and Macken, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Amer. Statist. Assoc.* **86** 27–35.

Kleffe, J. and Borodovsky, M. (1992). First and second moment of counts of words in random texts generated by Markov chains. *Comp. Applic. Biosci.* **8** 433–441.

Prum, B., Rodolphe, F. and de Turckheim, E. (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* **57** 205–220.

Reinert, G. and Schbath, S. (1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol.* **5** 223–253.

Régnier, M. (2000). A unified approach to word occurrence probabilities. *Discrete Applied Mathematics.* **104** 259–280.

Robin, S. and Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36** 179–193.

Robin, S. and Daudin, J.-J. (2001). Exact distribution of the distances between any occurrences of a set of words. *Ann. Inst. Statist. Math.* **36 (4)** 895–905.

Robin, S. and Schbath, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.* **8 (4)** 349–359.

Schbath, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics.* **1** 1–16. (`http://www.emath.fr/ps/`).

Chapitre 9

Checking homogeneity of patterns distribution in heterogeneous sequences

Ledent S., Robin S. (2002)

Contents

9.1 Introduction

9.1.1 Aim and principle

Aim.

- Detect unexpectedly rich or poor regions in terms of occurrences of a particular pattern
- Take into account clumping effects due to the overlapping structure of the pattern
- Take into account an heterogeneity structure of the sequence that is known a priori (genome annotation, `g-c` rich versus `a-t` rich regions, HMM modeling)

Method.

- We use a global modeling that is not based on the sequence of the nucleotides (no Markov chain).
- The sequence is considered as a continuous line and the occurrences as points on this line. Hence, occurrences have to be relatively rare.
- We want thresholds that vary along the sequence with respect to the prior heterogeneity structure.

The case of the heterogeneous Poisson process is already considered in Karlin and Macken (1991). We introduce here the possibility of overlaps thanks to the *compound* Poisson modeling and the way we take the prior heterogeneity into account in different and, on our opinion, preferable.

9.1.2 Definitions and notations

In this paper, we are essentially interested in the counting process $\{N(x)\}_{x\geq 0}$ of the occurrences of a given pattern along a continuous sequence. Denoting $\{X_1, X_2, \dots\}$ the positions of the successive occurrences, this counting process is defined by

$$N(x) = \sum_{n\geq 1} \mathbb{I}\{X_n \leq x\} \tag{9.1}$$

where $\mathbb{I}\{A\}$ equals 1 if A is true and 0 otherwise.

To analyze the distribution of the occurrences along the sequence, we will consider the distances $\{Y_1, Y_2, \dots\}$ between these occurrences:

$$Y_n = X_{n+1} - X_n \qquad \text{for } n \geq 1.$$

We will also consider the cumulated distances of order r $\{Y_1^r, Y_2^r, \dots\}$:

$$Y_n^r = \sum_{i=n}^{n+r-1} Y_i = X_{n+r} - X_n \qquad \text{for } n \geq 1.$$

To take into account the possible overlaps between occurrences of the pattern, we assume that this pattern occurs in clumps. We denote $\{Z_1, Z_2, \dots\}$ the positions of the successive clumps; their counting process $\{C(x)\}_{x\geq 0}$ is defined by

$$C(x) = \sum_{c\geq 1} \mathbb{I}\{Z_c \leq x\}.$$

Denoting $\{K_1, K_2, \dots\}$ the sizes of the clumps, the counting process $\{N(x)\}_{x\geq 0}$ of the occurrences of the pattern is defined by

$$N(x) = \sum_{c=1}^{C(x)} K_c. \tag{9.2}$$

Tied positions. In our modeling, occurrences belonging to the same clump,i.e. overlapping occurrences, have the same position. For example, if the first clump has position Z_1 and size $K_1 = 1$ and the second clump has position Z_2 and size $K_2 = 2$, then the positions of the occurrences will be $X_1 = Z_1$ and $X_2 = X_3 = Z_2$. The positions of the occurrences number 2 and 3 are tied. The distances between these occurrences will be $Y_1 = Z_2 - Z_1$ and $Y_2 = 0$. The next position and distance will be $X_4 = Z_3$ and $Y_3 = Z_3 - Z_2$, etc.

9.2 Heterogeneous compound Poisson model

9.2.1 Heterogeneous Poisson model for the clumps

Poisson process of the clumps' occurrences. We assume that $\{C(x)\}_{x\geq 0}$ is an heterogeneous Poisson process with intensity $\lambda(x)$ that is actually a mixture of S constant intensities $(\lambda_1, \dots, \lambda_S)$:

$$\lambda(x) = \sum_s \lambda_s \pi_s(x) \tag{9.3}$$

where, for all x and all s, $0 \leq \pi_s(x) \leq 1$ and $\sum_s \pi_s(x) = 1$. Index s reveals some kind of underlying 'state'. The proportions (or probabilities) $\pi_s(x)$ of these states may vary along the sequence but not their intensities λ_s's. Intensity λ_s is related to expected frequency of the clumps in state s: in an homogeneous sequence of length ℓ in state s, the expected number of clumps is $\lambda_s \ell$.

Denoting $\Pi_s(x, y)$ the cumulated probability over the interval $[x; y]$: $\Pi_s(x, y) = \int_x^y \pi_s(z)\mathrm{d}z$, the cumulated intensity over the same interval is

$$\Lambda(x, y) = \int_x^y \lambda(z)\mathrm{d}z = \sum_s \lambda_s \Pi_s(x, y). \tag{9.4}$$

We remind here two standard properties of heterogeneous Poisson process (see Kingman (1993)):

A: The number of clumps occurring in the interval $[x; y]$ has a Poisson distribution:

$$C(y) - C(x) \sim \mathcal{P}[\Lambda(x, y)].$$

B: The cumulated distribution function (cdf) of the position of the first clump and of the distance between successive clumps are given by

$$\Pr\{Z_1 > x\} = \exp[-\Lambda(0, x)], \qquad \Pr\{Z_{c+1} - Z_c > y\} = \exp[-\Lambda(Z_c, Z_c + y)]. \quad (9.5)$$

Maximum likelihood estimates of the parameters. The log-likelihood of the heterogeneous Poisson process $\{C(x)\}_{0 \leq x \leq \ell}$ is

$$\begin{aligned} \mathcal{L}(C) = &-\Lambda(0, Z_1) - \sum_{c=2}^{C(\ell)} \Lambda(Z_{c-1}, Z_c) + \Lambda(Z_{C(\ell)}, \ell) + \sum_{c=1}^{C(\ell)} \log[\lambda(Z_c)] \\ = &-\Lambda(0, \ell) + \sum_{c=1}^{C(\ell)} \log[\lambda(Z_c)] \;=\; -\sum_s \lambda_s \Pi_s(0, \ell) + \sum_{c=1}^{C(\ell)} \log[\lambda(Z_c)]. \end{aligned} \quad (9.6)$$

The maximization of $\mathcal{L}(C)$ gives the maximum likelihood estimates (MLE) of the λ_s's; it can be made using the Newton-Raphston algorithm, provided its first and second derivatives given in appendix 9.A.

This log-likelihood is simplified in the case where the probabilities $\pi_s(x)$ are always either 0 or 1. In this case, denoting $C_s(\ell)$ the number of clumps occurring 'in state s': $C_s(\ell) = \sum_{c=1}^{C(\ell)} \mathbb{I}\{\pi_s(Z_c) = 1\}$, $\mathcal{L}(C)$ becomes $\mathcal{L}(C) = -\sum_s [\lambda_s \Pi_s(0, \ell) + C_s(\ell) \log(\lambda_s)]$. In this case, the MLE of λ_s is simply $\hat{\lambda}_s = C_s(\ell)/\Pi_s(0, \ell)$.

This particular result suggests a way to calculate starting values for the Newton-Raphston algorithm in the general case. Replacing $C_s(\ell)$ by the sum of the probabilities $\pi_s(x)$ at each clump position, we propose to start with $\lambda_s^0 = \left[\sum_{c=1}^{C(\ell)} \pi_s(Z_c)\right] / \Pi_s(0, \ell)$. In all the analyzes we made, these starting values were very close (**i.e. within the confidence interval**) to the final estimates.

9.2.2 Size of the clumps

Because we assume that the sequence is heterogeneous a priori, it is natural to assume that the distribution of the clumps' size is also heterogeneous. In the following we denote the distribution of the size K given the position Z of the clump by

$$p(k \mid x) = \Pr\{K = k \mid Z = x\} \qquad \text{for } k \geq 0.$$

If the distribution of the size of the clumps is homogeneous, we simply denote $p(k) = \Pr\{K = k\}$.

General case. Consistently with the modeling of the intensity $\lambda(x)$ introduced in (9.3), we assume that the dependence between the size and the position of a clump has the following form:

$$p(k \mid x) = \sum_s \pi_s(x) p_s(k)$$

where p_s denotes the distribution of the clump size in s.

Mean distribution. In Section 9.2.3 we will be interested in the by the distribution of the count $N(x,y) = N(y) - N(x)$. This distribution depends on the mean distribution of the clump size over the interval $[x;y]$:

$$\overline{p}(k \mid x,y) = \frac{\int_x^y p(k \mid z)\lambda(z)\mathrm{d}z}{\int_x^y \lambda(z)\mathrm{d}z} = \frac{\sum_s \lambda_s \Pi_s(x,y) p_s(k)}{\sum_s \lambda_s \Pi_s(x,y)}.$$

Geometric case. As in Robin (2002), we emphasize the particular case of the geometric distribution. This distribution is especially adapted to the case where the pattern is a single word that has a probability a to overlap itself. If the probability a depends on s, we get:

$$p(k \mid x) = \sum_s \pi_s(x) a_s^{k-1}(1-a_s).$$

If this probability is independent of the state s, we get: $p(k) = a^{k-1}(1-a)$.

Likelihood and parameter estimates. In the general case, we need to consider the conditional likelihood of the clumps' sizes conditional to their position that is

$$\mathcal{L}(\{K_c\} \mid \{Z_c\}) = \sum_c \log p(K_c \mid Z_c) = \sum_c \log \left[\sum_s \pi_s(Z_c) p_s(K_c) \right]. \tag{9.7}$$

The MLE's of the probabilities $p_s(k)$ are obtained by maximizing $\mathcal{L}(\{K_c\} \mid \{Z_c\})$ provided that, for all s, $\sum_{k \geq 0} p_s(k) = 1$. The first and second derivatives of this conditional likelihood are given in appendix 9.A: they are necessary for a Newton-Raphston optimization.

If the distribution of the clump size is homogeneous, $\mathcal{L}(\{K_c\} \mid \{Z_c\})$ reduces to $\sum_c \log p(K_c)$, and the MLE of $p(k)$ is simply

$$\hat{p}(k) = \frac{\#\text{ clumps of size } k}{C(\ell)}.$$

The most convenient case is the homogeneous geometric case that involve one single parameter (a), the MLE of which is $\hat{a} = 1 - C(\ell)/N(\ell)$ (see Robin (2002)).

As for the estimation of the λ_s's, the estimates of the homogeneous case suggest starting values for the maximization algorithm. In the general case, one may set $\hat{p}_s^0(k) = \left[\sum_{c:K_c=k} \pi_s(Z_c)\right] / \left[\sum_c \pi_s(Z_c)\right]$ and in the geometric case $\hat{a}_s^0(k) = 1 - \left[\sum_c \pi_s(Z_c)\right] / \left[\sum_c \pi_s(Z_c) K_c\right]$.

Testing the homogeneity of the clumps size. As we will see in Section 9.3, the hypothesis of homogeneous clump size distribution is very convenient to detect unexpectedly rich or poor regions. The hypothesis "$p_1 = p_2 = \cdots = p_S$" can be tested using the likelihood ratio test. In the case where the probabilities $\pi_s(x)$ are all 0 or 1, the distribution of the clump size can be estimated directly in each state and the chi-square test can be used as well.

9.2.3 Distributions of the count

In view of analyzing the homogeneity of the occurrences distribution along the sequence, we need to know the distribution of the count $N(x, y) = N(y) - N(x)$ in any interval $[x; y]$. The number of clump in $[x; y]$ has a Poisson distribution $\mathcal{P}[\Lambda(x, y)]$ (Property **A**). Each clump occurring in $[x; y]$ has the marginal distribution $\overline{p}(\cdot \mid x, y)$. Hence, equation (9.2) shows that the count $N(x, y)$ has a compound Poisson distribution denoted $\mathcal{CP}[\Lambda(x, y); \overline{p}(\cdot \mid x, y)]$.

The probability $\Pr\{N = n\}$ can be calculated using a $O(n^2)$ algorithm given in Robin (2002) and based on a result of Barbour *et al.* (1992). Practically, because this probability depends on the interval $[x; y]$, it is very time consuming to compute in sequence analysis. Note that, if the distribution of the clump size is homogeneous, this distribution only depends on the cumulated intensity $\Lambda(x, y)$.

Geometric case. In the special (and favorable) case where the mean distribution is geometric with parameter a, the distribution of the count is a Pólya-Aeppli distribution $\mathcal{PA}[\Lambda(x, y); a]$ (see Johnson *et al.* (1992)):

$$
\begin{aligned}
\Pr\{N(x,y) = n\} &= e^{-\Lambda(x,y)} a^n \sum_{c=1..n} \frac{1}{c!} \binom{n-1}{c-1} \left[\frac{\Lambda(x,y)(1-a)}{a}\right]^c && \text{for } n > 0, \\
&= e^{-\Lambda(x,y)} && \text{for } n = 0
\end{aligned}
$$

9.3 Two methods to check spreading homogeneity

We concentrate now on the detection of unexpectedly rich or poor regions in terms of pattern occurrences. To present the method, we will limit ourselves to the detection of rich regions. The natural criterion to evaluate the richness of a region is ratio

$$
\frac{\text{number of pattern occurrences in the region}}{\text{length of the region}}.
$$

We propose two methods based on this kind of ratio. In the first one (Section 9.3.2), the numerator is fixed and the denominator is random; in the second one (Section 9.3.3), the numerator is random and the denominator is fixed. For both methods, we will need to homogenized the clump process (Section 9.3.1).

9.3.1 Correction of the prior heterogeneity

We emphasize that probabilities $\pi_s(x)$ are known a priori. This is a crucial point of our approach. The underlying states refer to biological structure that may be related to the frequency of the pattern. Our purpose is to detect second heterogeneity, i.e. an heterogeneity that is not due to this prior heterogeneity $\lambda(x)$. Hence, this prior structure has to be defined independently of the observed distribution of the pattern along the sequence.

Homogenization of the clump process. Given the probabilities $\pi_s(x)$'s and intensities λ_s's, the heterogeneous clump process $\{C(x)\}$ can be transformed into an homogeneous Poisson process $\{\tilde{C}(\tilde{x})\}$. This homogenization is based on the continuous transform $\mathrm{d}\tilde{x} = \lambda(x)\mathrm{d}x$.

Denoting

$$\tilde{Z}_c = \Lambda(0, Z_c), \tag{9.8}$$

the counting process $\tilde{C}(\tilde{x}) = \sum_{c\geq 1} \mathbb{I}\{\tilde{Z}_c \leq \tilde{x}\}$ is an homogeneous Poisson process with intensity 1. According to Property **B**, the distances $\tilde{Z}_{c+1} - \tilde{Z}_c$ are all independent with common distribution $\mathcal{E}(1)$: for all $c \geq 1$, $\Pr\{\tilde{Z}_{c+1} - \tilde{Z}_c > \tilde{y}\} = \exp(-\tilde{y})$.

In the following $\{\tilde{C}(\tilde{x})\}_{\tilde{x}\geq 0}$ (resp. $\{\tilde{Z}_c\}_{c\geq 1}$) will be called the homogenized process (resp. positions) of the clumps. The clumps are uniformly distributed along the homogenized interval $[0; \widetilde{\ell}]$, with $\widetilde{\ell} = \Lambda(0, \ell)$. Homogenized occurrences positions $\{\tilde{X}\}_{n\geq 1}$, distances $\{\tilde{Y}\}_{n\geq 1}$ and counting process $\{\widetilde{N}(\tilde{x})\}_{\tilde{x}\geq 0}$ are defined in the same way.

9.3.2 Cumulated distance

The first method that we propose to detect heterogeneity in the distribution of the occurrences is based on the r-scan approach introduced in Karlin and Macken (1991) and Dembo and Karlin (1992). This approach is based on the cumulated distances of order r (r-scans) defined in Section 9.1.2. The idea is to detect unexpectedly rich regions by looking at the minimal cumulated distances (poor regions are detected by maximal distances).

Chen-Stein approximation. The key results given in Dembo and Karlin (1992) is the approximate distribution of $\min_{n\geq 1}\{Y_n^r\}$ that is obtained thanks to the Chen-Stein method (see Arratia *et al.* (1989)). Denoting F_r the common cdf of the Y_c^r's, and given the total number of occurrence $N(\ell)$, the number of cumulated distances smaller than a (low) threshold y has approximately a Poisson distribution with parameter $[N(\ell)-r]F_r(y)$. Hence the distribution of the shortest cumulated distance is approximately

$$\Pr\{\min_{n\geq 1}[Y_n^r] \leq y\} \simeq 1 - \exp\{-[N(\ell) - r]F_r(y)\}. \tag{9.9}$$

Furthermore, the difference between this approximate probability and its true value can be bounded. The so-called 'Chen-Stein' bound is given in Equation (14) of Robin (2002). Equation (9.9) provides a threshold y^* for a given level α: $\Pr\{\min_{n\geq 1}[Y_n^r] \leq y^*\} = \alpha$. The Chen-Stein bound provides a bounding of y^*.

This result is based on the assumption that the distances $\{Y_n\}_{n\geq 1}$ are independent, a it is the case in a Poisson process. Thanks to this hypothesis, the dependence between the Y_n^r's is limited and the Chen-Stein bound can be calculated.

Application to the compound Poisson modeling. The Chen-Stein approximation presented above is still valid if the distances have different distribution. The case of the heterogeneous Poisson process is considered in Karlin and Macken (1991). The authors give a global threshold y^* that is the same all along the sequence.

We propose here to have a non constant threshold in order to take the prior heterogeneity $\lambda(x)$ into account in the threshold itself. To do this, we simply apply the Chen-Stein approximation to the homogenized distances $\{\tilde{Y}_n\}_{n\geq 1}$.

As shown in Robin (2002), in the homogeneous compound Poisson modeling framework, the distances $\{\tilde{Y}_n\}_{n\geq 1}$ are independent if

i) the clump size is homogeneous along the sequence: $p(k \mid x) \equiv p(k)$,

ii) their distribution is geometric: $p(k) = a^{k-1}(1-a)$.

Under conditions *i*) and *ii*), we can apply the results obtained for the homogeneous compound Poisson process by Robin (2002). The homogenized distances $\{\tilde{Y}_n\}_{n\geq 1}$ have common cdf:

$$\widetilde{F}(\tilde{y}) = 1 - (1-a)e^{-\tilde{y}}$$

that can be used to determined the threshold $\tilde{y}^*$ thanks to Equation (9.9).

Extremal cumulated distance. $\tilde{y}^*$ is the 'homogenized threshold' (with level α), that is the threshold for the minimal homogenized cumulated distance $\min_n\{\tilde{Y}_n\}$. We get a non constant threshold $y^*(x)$ $(0 \leq x \leq \ell)$ for the minimal cumulated distances $\min_n\{Y_n\}$ by applying the inverse transform of (9.8) to $\tilde{y}^*$.

The threshold $y^*(x)$ will be lower (i.e. more severe) in regions where the pattern is expected to be frequent according to the prior heterogeneity $\lambda(x)$; it will be higher in region where the pattern is expected to be rare.

9.3.3 Local count

Assumptions *i*) and *ii*) of Section 9.3.2 are restrictive and do not hold in the general case. If the pattern is a set of word having different overlapping probabilities, it is clear that the (homogenized) distances are not independent.

We propose now a second approach that is not based on the distances between occurrences but on the count of these occurrences. The idea is simply to look at the number of patterns occurring in an interval that move continuously along the sequence.

Interval with fixed width. Given a fixed width Δx, the distribution of the count $N(x, x+\Delta x)$ has been studied in Section 9.2.3. In the general case, this distribution depends on both x and Δx. We denote this distribution $q(n|x, x+\Delta) = \Pr\{N(x, X+\Delta x) = n\}$ and $Q(n|x, x+\Delta) = \sum_{m\leq n} q(m|x, x+\Delta)$ the corresponding survival function.

Following Aldous (1989) heuristic, we approximate the probability that the count in the richest interval does not exceed a (high) threshold n^* by

$$\Pr\left\{\sup_{0\leq x\leq \ell-\Delta x}[N(x, X+\Delta x)] \geq n^*\right\} \simeq 1 - \exp\left[-\int_0^{\ell-\Delta x} Q(n^*|x, x+\Delta)\mathrm{d}x\right]. \quad (9.10)$$

This approximation can be view as a result of the Chen-Stein method applied to continuous processes (see Barbour and Brown (1992)). As shown in Robin (2002), it is optimal in terms of total variation distance but the corresponding bound is not tractable.

Practically, the right hand side of (9.10) requires to calculate the distribution of all the possible intervals. The computation time is in $O[\ell(n^*)^2]$, which is very demanding, especially in the case of DNA sequences.

Interval with homogenized fixed width. If the distribution of the clump size is homogeneous, the distribution $q(n|x, x+\Delta)$ only depends on the width Δx. We now concentrate on this case: $p(k \mid x) \equiv p(k)$ that does not need to be a geometric distribution. In this case, the problem is much simplified when applying Aldous' heuristic to the homogenized process $\{\widetilde{N}(\tilde{x})\}_{\tilde{x}\geq 0}$. This approach is coherent with the cumulated distance approach of Section 9.3.2.

Considering a fixed homogenized width $\Delta\tilde{x}$, the (homogenized) number of clumps in $[\tilde{x}; \tilde{x}+\Delta\tilde{x}]$ has a Poisson distribution $\mathcal{P}(\Delta\tilde{x})$. Hence, for all $\tilde{x}$, the homogenized count $\widetilde{N}(\tilde{x}, \tilde{x}+\Delta\tilde{x})$ has the same compound Poisson distribution:

$$\widetilde{N}(\cdot, \cdot + \Delta\tilde{x}) \sim \mathcal{CP}(\Delta\tilde{x}, p).$$

Distribution of extremal count. Let $\tilde{Q}$ denote the survival function of the $\mathcal{CP}(\Delta\tilde{x}, p)$ distribution: $\tilde{Q}(n) = \Pr\{\widetilde{N}(\cdot, \cdot + \Delta\tilde{x}) \geq n\}$. In the case where the distribution of the clump size is homogeneous, we can approximate the distribution of the highest count:

$$\Pr\left\{ \sup_{0\leq\tilde{x}\leq\ell-\Delta\tilde{x}} \left[\widetilde{N}(\tilde{x}, \tilde{x}+\Delta\tilde{x})\right] \geq \widetilde{n}^* \right\} \simeq 1 - \exp\left[-(\widetilde{\ell} - \Delta\tilde{x})\tilde{Q}(\widetilde{n}^*)\right].$$

For a given level α, we can calculated the corresponding homogenized threshold $\widetilde{n}^*$ once for all. As in Section 9.3.2, this constant threshold gives a threshold $n^*(x)$ that varies along the sequence thanks to the inverse transform of (9.8).

9.4 Applications

Bibliography

ALDOUS, D. (1989). *Poisson Approximations via the Poisson Clumping Heuristic.* Springer-Verlag: New-York.

ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1989). Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Prob.* **17** 9–25.

BARBOUR, A. D. and BROWN, T. (1992). Stein's method and point process approximation. *Stoch. Proc. Appl.* **43** 9–31.

BARBOUR, A. D., CHEN, L. H. Y. and LOH, W.-L. (1992). Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Prob.* **20** 1843–1866.

DEMBO, A. and KARLIN, S. (1992). Poisson approximations for r-scan processes. *Ann. Appl. Prob.* **2 (2)** 329–357.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions.* Wiley: New-York.

Karlin, S. and Macken, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Amer. Statist. Assoc.* **86** 27–35.

Kingman, J. F. C. (1993). *Poisson Processes.* Oxford University Press.

Robin, S. (2002). A compound Poisson model for words occurrences in DNA sequences. *J. R. Statist. Soc. C.* (*to appear*).

Appendix

9.A Likelihood and derivatives

9.A.1 Clump occurrences

The first and second derivatives of $\mathcal{L}(C)$ defined in (9.6) are

$$\frac{\partial \mathcal{L}}{\partial \lambda_s} = -\Pi_s(0,\ell) + \sum_{c=1}^{C(\ell)} \frac{\pi_s(Z_c)}{\lambda(Z_c)}, \qquad \frac{\partial^2 \mathcal{L}}{\partial \lambda_s \partial \lambda_t} = \sum_{c=1}^{C(\ell)} -\frac{\pi_s(Z_c)\pi_t(Z_c)}{[\lambda(Z_c)]^2}.$$

9.A.2 Clump sizes

General case. We are looking for the maximum of the conditional likelihood $\mathcal{L}(\{K_c\} \mid \{Z_c\})$ defined in (9.7) with respect to the $p_s(k)$. Because $\sum_{k\geq 1} p_s(k) = 1$ for all s, we only consider the derivatives with respect to the $p_s(k)$'s for $k \geq 2$, replacing $p_s(1)$ by $1 - \sum_{k\geq 2} p_s(k)$ in $\mathcal{L}(\{K_c\} \mid \{Z_c\})$. The first derivative with respect to $p_s(k)$ is

$$\frac{\partial \mathcal{L}}{\partial p_s(k)} = \sum_{c:K_c=k} \frac{\pi_s(Z_c)}{\sum_{u=1}^S \pi_u(Z_c)p_u(k)} - \sum_{c:K_c=1} \frac{\pi_s(Z_c)}{\sum_{u=1}^S \pi_u(Z_c)(1-\sum_{k\geq 2} p_u(k))}.$$

For all s and t, the second derivative with respect to $p_s(k)$ and $p_t(m)$ is, for $k \neq m$),

$$\frac{\partial^2 \mathcal{L}}{\partial p_s(k)\partial p_t(m)} = -\sum_{c:K_c=1} \frac{\pi_s(Z_c)\pi_t(Z_c)}{\left\{\sum_u \pi_u(Z_c)[1-\sum_{k\geq 2} p_u(k)]\right\}^2};$$

otherwise,

$$\frac{\partial^2 \mathcal{L}}{\partial p_s(k)\partial p_t(k)} = -\sum_{i:K_c\neq 1} \frac{\pi_s(Z_c)\pi_t(Z_c)}{[\sum_u \pi_u(Z_c)p_u(k)]^2} - \sum_{c:K_c=1} \frac{\pi_s(Z_c)\pi_t(Z_c)}{\left\{\sum_u \pi_u(Z_c)[1-\sum_{k\geq 2} p_u(k)]\right\}^2}.$$

Geometric case. In this case, the log-likelihood (9.7) becomes

$$\mathcal{L}(\{K_c\} \mid \{Z_c\}) = \sum_{i=1}^{C(\ell)} \log\left[\sum_s \pi_s(Z_c) a_s^{K_c-1}(1-a_s)\right].$$

Its first derivative with respect to a_s is

$$\frac{\partial \mathcal{L}}{\partial a_s} = \sum_{i=1}^{C(\ell)} \frac{\pi_s(Z_c)a_s^{K_c-2}(K_c - K_c a_s - 1)}{\sum_u \pi_u(Z_c)a_u^{K_c-1}(1-a_u)}.$$

For $t \neq s$, the second derivative is

$$\frac{\partial^2 \mathcal{L}}{\partial a_s \partial a_t} = -\sum_{c=1}^{C(\ell)} \frac{\left[\pi_s(Z_c)a_s^{K_c-2}(K_c - K_c a_s - 1)\right]\left[\pi_t(Z_c)a_t^{K_c-2}(K_c - K_c a_t - 1)\right]}{\left[\sum_u \pi_u(Z_c)a_u^{K_c-1}(1-a_u)\right]^2};$$

otherwise,

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial a_s^2} &= \sum_{i=1}^{C(\ell)} \frac{\pi_s(Z_c) a_s^{K_c-3}[(K_c-2)(K_c-K_c a_s-1)-a_s K_c]}{\left[\sum_u \pi_u(Z_c) a_u^{K_c-1}(1-a_u)\right]^2} \\
&\quad - \sum_{c=1}^{C(\ell)} \left[\frac{\pi_s(Z_c) a_s^{K_c-2}(K_c-K_c a_s-1)}{\sum_u \pi_u(Z_c) a_u^{K_c-1}(1-a_u)}\right]^2.
\end{aligned}$$

Quatrième partie

Autres publications

Chapitre 10

Sensitivity analysis of non-parametric tests for censored data

ROBIN S. (1996)

Communications in Statistics - Theory and Methods, **25**, (8), 1693–1715

Abstract. The use of non-parametric tests for the two sample problem such as say the weighted logrank tests for censored data is based on the hypothesis that the probability of belonging to a groupe is the same for each individual. Our problem is : what can be said about the significance level of such a test if this hypothesis does not hold ?

Using a new formulation of the weighted logrank tests statistics, it is shown that, in presence of such an heterogeneity, bounds on the exact (and unknown) significance level may be given. This results allows to lead the sensitivity analysis of those test statistics which are shown to be asymptotically normal.

Keywords. heterogeneity ; robustness ; weighted logrank tests.

Contents

10.1 Sensitivity analysis

10.1.1 Problem

In a paper published in 1987, Rosenbaum (1987) raises the problem of testing the effect of a treatment on a response in observational studies. The question is: what can be said about the effect of a treatment when its assignment is not controlled by the statistician ?

Let Y denote the observed response and Z the treatment. Z is a binary variable belonging to $\{0, 1\}$. We want to test the hypothesis

$$\mathbf{H}_0 : \quad \mathcal{L}(Y|Z = 0) = \mathcal{L}(Y|Z = 1),$$

which says that the response is independent of the treatment. We hope to reject $\mathbf{H}_0$ to accept the alternative hypothesis according to which the treatment has a favorable effect on the response.

The problem raised by P. Rosenbaum is that, generally, Y does not only depend on the treatment but also on certain characteristics of the subject. Some of those characteristics, say V, are observed, and some others, say U, are not. We shall denote $X := (U, V)$ the set of covariates. A vector (X, Y, Z) is then associated to each individual. In an experimental study, Z may be randomly assigned to every subject and so, it is possible to test the *marginal independence* of Y and Z. In this case, X may be dropped since it is identically distributed among the individuals receiving the treatment ($Z = 1$) and the controls ($Z = 0$). On the contrary, in an observational study, the assignment may depend on X which may also have an influence on Y. In this case, the distribution of X may be very different among the treated subjects and among the controls and the *apparent effect* of Z on Y may only reflect this difference. Hence X can not be abandoned and we have to test the *conditional independence* of Y and Z:

$$\mathbf{H}'_0 : \quad \forall x,\ \mathcal{L}(Y|Z = 0, X = x) = \mathcal{L}(Y|Z = 1, X = x).$$

If the unobserved part U of X does not exist (*i.e.* if $X = V$), it is possible to match pairs in the population.

Here, the point of interest is the *latent variable* U. We shall therefore drop V, the observed part of X, and assume that $X = U$.

We shall assume that the probability of receiving the treatment depends on the covariate U. If this covariate was observed and if its influence on the treatment assignment was known, it would be possible to give the distribution of the test statistic. Since it is not observed, the true distribution of the statistic remains unknown but we will show that it is possible to give two bounding distributions.

Let's call *probability function* (denoted p), the function giving the probability of being treated given a particular value of U:

$$p(u) = \mathrm{P}\{Z = 1|U = u\}.$$

Here, and this is an important difference with Rosenbaum (1987), we do not need to specify the function p.

The case $p(u) \equiv p$ will be called *equiprobability* ; the hypothesis of equiprobability is the classical one, it implies that all individuals have the same probability of receiving the treatment.

The sensitivity analysis is the study of the evolution of the bounding distributions when the influence of the *heterogeneity factor* U - on the treatment assignment - grows.

In this paper, the response Y to the treatment is an eventually *right censored duration*. More precisely, if Y_i' denotes the survival time of the i-th subject, one observes $Y_i := \min(Y_i', C_i)$, where C_i is a positive random variable censoring Y_i' on its right. If C_i is lower than Y_i', one has $Y_i = C_i$ and the subject is said to be censored. So, for each subject i, one observes the two-dimensional response (Y_i, D_i) where Y_i is the duration and D_i the death indicator taking the value 1 if $Y_i = Y_i'$ and 0 otherwise.

10.1.2 Notations

In the following, we shall assume without loss of generality that the observed durations are given in ascending order: $y_1 \leq y_2 \leq ... \leq y_N$. Let R_i denote the number of individuals '*at risk*' at time Y_i^- *i.e.* neither dead nor censored immediately before Y_i , R_{Ai} (resp. R_{Bi}) the number of treated individuals (resp. controls) at risk at the same time and M_i the number of observed deaths in Y_i : $R_i := \sum_j \mathbf{1}\{Y_j \geq Y_i\}$, $R_{Ai} := \sum_j Z_j \mathbf{1}\{Y_j \geq Y_i\}$, $R_{Bi} := \sum_j (1 - Z_j)\mathbf{1}\{Y_j \geq Y_i\}$, $M_i := \sum_j D_j \mathbf{1}\{Y_j = Y_i\}$.

Let T be the test statistic ; if the rejection region of $\mathbf{H}_0$ has the form $\{T \leq k\}$, one shall call *significance level of the test*, the probability $\mathrm{P}_{\mathrm{H}_0}\{T \leq k\}$.

10.2 Weighted logrank tests

10.2.1 Historical presentation

We are interested here in a family of tests for the two sample problem with censored data : the *weighted logrank test statistics*. All those statistics can be written as

$$T := \sum_{i=1}^{N} D_i W_i \left(Z_i - \frac{R_{Ai}}{R_i} \right),$$

where the weight W_i only depends on deaths and censorings observed before Y_i and does not depend on the treatments Z_i's:

$$W_i = w\left(\{(Y_j, D_j), j : Y_j < Y_i\}\right).$$

Historically, those statistics have appeared separately in the literature. Gehan (1965) proposes an extension of the Wilcoxon's test statistic to censored data ; Gehan's statistic is based on pairs comparison and is presented as a U-statistic. Cox (1972) uses the logrank test which is defined as the partial likelihood score-test of a censored sample. Harrington and Fleming (1982) consider weights depending on the Kaplan-Meyer survival function estimator. The classical weights are : $W_i = R_i$ (Gehan), $W_i \equiv 1$ (logrank), $W_i = \sqrt{R_i}$ (Tarone & Ware), $W_i = \{\widehat{S}(Y_i)\}^\rho$, $\rho \in [0, 1]$ (Harrington & Fleming).

The point process theory gives another expression of the same statistics :

$$T = \int K \left(\frac{\mathrm{d}N_A}{R_A} - \frac{\mathrm{d}N_B}{R_B} \right),$$

where N_A et N_B are the counting processes of deaths in group A (treated) and B (controls): $N_A(t) := \sum_i D_i Z_i \mathbf{1}\{Y_i \leq t\}$, $N_B(t) := \sum_i D_i(1 - Z_i)\mathbf{1}\{Y_i \leq t\}$ and K is a predictable process featuring the W_i's.

Under the equiprobability hypothesis (*i.e.* $p(u) \equiv p$), the asymptotic normality may be proved invoking the conditional independence of the elements of the sum given the past (in the first way) or using results on convergence of martingales (in the second way).

Remark. For all classical weights, it is clear that, since the y_i's are given in ascending order, w_i decreases as the index i grows. It's also easy to show that the ratio (w_i/r_i) increases with i. We shall use those properties later on.

10.2.2 A new formulation of the test statistics

The tests we study here are based on statistics T given the sequence of the (y_i, d_i)'s. So, the only random element is the corresponding sequence of the treatment assignments Z_i's. Let us use capital letters for the random terms and ordinary letters for the given terms :

$$T := \sum_{i=1}^{N} w_i d_i \left(Z_i - \frac{R_{Ai}}{r_i} \right). \tag{10.1}$$

An expression putting random terms apart from the others would be of great interest to study the distribution of T. Such an expression is given here, writing the statistic T as a simple scalar product between the random vector $\mathbf{Z}$ of the treatment assignments and a vector $\mathbf{t}$ of individual scores. The method to calculate the scores is an extension of the Gehan's method ; let t_{ij} be the result of the comparison of subjects i and j :

$$\begin{array}{lll} t_{ij} & := & -w_j/r_j \quad \text{if subject } i \text{ is certainly dead after subject } j, \\ & & w_i/r_i \quad \text{if subject } j \text{ is certainly dead after subject } i, \\ & & 0 \qquad\quad \text{if nothing is certain,} \end{array}$$

which gives formally

$$t_{ij} := \left(\frac{d_i w_i}{r_i} \right) \mathbf{1}\left\{ y_i \leq y_j \right\} - \left(\frac{d_j w_j}{r_j} \right) \mathbf{1}\left\{ y_j \leq y_i \right\}. \tag{10.2}$$

The individual scores t_i are then defined as follows:

$$t_i := \sum_{j=1}^{N} t_{ij}. \tag{10.3}$$

A 'good' subject is a subject with a negative score.

Lemma 2.1 *Let the t_i's be the individual scores defined in* (10.2) *and* (10.3)*, the statistic T defined in* (10.1) *may be written as:*

$$T = \langle \mathbf{t} \cdot \mathbf{Z} \rangle = \sum_{i=1}^{N} t_i Z_i.$$

Proof: $\sum_{i=1}^{N} t_i Z_i = \sum_{i=1}^{N} \left(Z_i \sum_{j=1}^{N} t_{ij} \right) = \sum_{i=1}^{N} \left\{ Z_i \left(\frac{d_i w_i}{r_i} \right) r_i \right\} - \sum_{j=1}^{N} \left\{ \left(\frac{d_j w_j}{r_j} \right) \sum_{i=1}^{N} Z_i \mathbf{1} \left\{ y_j \leq y_i \right\} \right\} = \sum_{i=1}^{N} \left(\frac{d_i w_i}{r_i} \right) (Z_i r_i - R_{Ai}) = \sum_{i=1}^{N} d_i w_i \left(Z_i - \frac{R_{Ai}}{r_i} \right)$.□

Statistics T can be written as *the score sum of treated individuals.* This new expression puts apart random terms Z_i from those conditionally to whom we work.

If the treatment tends to increase the survival, the statistic will take a negative value since the comparisons (10.2) will generally be favorable to treated individuals and therefore their scores (10.3) will be negative.

The contribution of each individual to the statistic is well evaluated thanks to this new expression of the weighted logrank test statistics. For example, it enhances the fact that a censored subject i is always 'good' since all the comparisons t_{ij} he is involved in are non-positive because no other subject is certainly dead after him : censored individuals are potentially immortal.

10.3 Bounds of the significance level

Lemma **2.1** says that all weighted logrank tests statistics can be written as the scalar product of the score vector $\mathbf{t}$ with the treatment assignment vector $\mathbf{Z}$. The presence of an heterogeneity factor U implies that, for each subject, the probability of receiving the treatment depends on the value of U.

Let us assume that

$$\mathbf{A}: \quad \exists \, (p_{\min}, p_{\max}) \in \,]0;1[^2 : \; \forall u, \; p_{\min} \leq p(u) \leq p_{\max}.$$

This hypothesis slightly reduces the interval in which $p(u)$ takes its values: the extreme non-random cases $p(u) = 0$ and $p(u) = 1$ are eliminated. One has to remark that *the probability function p remains unknown.*

The purpose of this part is to give bounds of the - unknown - significance level of the test which obviously depends on the probabilities $p_i = p(u_i)$. Let us denote α_k this significance level :

$$\alpha_k(\mathbf{t}, \mathbf{p}) = \mathrm{P} \left\{ T = \langle \mathbf{t} \cdot \mathbf{Z} \rangle \leq k | \mathbf{p} \right\},$$

where $\mathbf{t} = (t_1, ..., t_N) \in \mathbb{R}^N$, $\mathbf{p} = (p_1, ..., p_N) \in \mathcal{P} := [p_{\min}, p_{\max}]^N$, $\mathbf{Z} = (Z_1, ... Z_N) \in \{0, 1\}^N$. $\mathbf{p}$ is called the probability vector.

In order to obtain such a result, we have to distinguish two different kinds of treatment assignment :

1. treatments Z_i's are independently assigned to subjects,
2. treatments Z_i's are assigned to subjects given a fixed number of subjects to be treated: $\sum_{1\leq i\leq N} Z_i = n$.

It has to be pointed out that $\mathbf{Z}$ is the only random vector and that we work conditionnaly to $\mathbf{t}$ *i.e.* to the sequence of durations and death indicators (y_i, d_i): here, we have to deal with the random aspect of treatment assignments, not with the random aspect of durations.

3.1 Independent assignments

This hypothesis means that the Z_i's are independent Bernoulli random values with respective parameters p_i :

$$\mathbf{B}: \quad \{Z_i, 1 \leq i \leq N\} \text{ independent}, \quad Z_i \sim \mathcal{B}(p_i).$$

Under $\mathbf{B}$, the mean and the variance of T are easily found :

$$\mathrm{E}(T) = \sum_{i=1}^{N} t_i p_i, \quad \mathrm{V}(T) = \sum_{i=1}^{N} t_i^2 p_i(1-p_i).$$

It is also clear that the values of T are spread between the sum of all negative scores and the sum of all positive scores. T reaches those bounds for two particular values of $\mathbf{z}$:

$$\tilde{\mathbf{p}} := \arg\max_{\mathbf{p}\in\mathcal{P}}(P\{\mathbf{Z}=\mathbf{z}|\mathbf{p}\}) \quad \Rightarrow \quad \tilde{p}_i := p_{\max}\mathbf{1}\{t_i < 0\} + p_{\min}\mathbf{1}\{t_i \geq 0\},$$
$$\tilde{\mathbf{p}}' := \arg\max_{\mathbf{p}\in\mathcal{P}}(P\{\mathbf{Z}=\mathbf{z}'|\mathbf{p}\}) \quad \Rightarrow \quad \tilde{p}_i' := p_{\max}\mathbf{1}\{t_i > 0\} + p_{\min}\mathbf{1}\{t_i \leq 0\}.$$

The vectors $\widetilde{\mathbf{z}}$ and $\widetilde{\mathbf{z}}'$ are called *extreme combinations of treatment*. $\widetilde{\mathbf{z}}$ is the most favorable combination since it gives the best value of the statistic (*i.e.* the lowest).

In an intuitive way, if the probability vector $\mathbf{p}$ is such that $\widetilde{\mathbf{z}}$ is a very likely issue of $\mathbf{Z}$, we can think that the real significance level is very different from the significance level calculated under the equiprobability hypothesis. In fact, it must be higher.

Hypothesis $\mathbf{B}$ clearly implies that

$$\mathrm{P}\left\{\mathbf{Z}=\mathbf{z}|\mathbf{p}\right\} = \prod_{i=1}^{N} p_i^{z_i}(1-p_i)^{1-z_i},$$

so, the vector $\widetilde{\mathbf{z}}$ (resp. $\widetilde{\mathbf{z}}'$) has a maximal probability of being observed for the probability vector $\widetilde{\mathbf{p}}$ (resp. $\widetilde{\mathbf{p}}'$):

$$\tilde{\mathbf{z}} := \arg\min_{\mathbf{z}\in\{0,1\}^N}(T = <\mathbf{t}.\mathbf{Z}>) \quad \rightarrow \quad \tilde{z}_i := \mathbf{1}\{t_i < 0\}$$
$$\tilde{\mathbf{z}}' := \arg\min_{\mathbf{z}\in\{0,1\}^N}(T = <\mathbf{t}.\mathbf{Z}>) \quad \rightarrow \quad \tilde{z}_i' := \mathbf{1}\{t_i > 0\}$$

To obtain the bounding of $\alpha_k(\mathbf{t}, \mathbf{p})$, we have to define a pre order on $\mathcal{P}$ which induces a stochastic order on T.

Definition. *Let us denote ' $\leq_{\mathbf{t}}$ ' the pre order on $\mathcal{P}$:*

$$\forall(\mathbf{p},\mathbf{p}')\in\mathcal{P}^2,\quad \mathbf{p}\leq_{\mathbf{t}}\mathbf{p}' \quad\Leftrightarrow\quad \forall i,\ t_i(p_i-p'_i)\leq 0.$$

Lemma 3.1 *Under* **A**, *the vectors* $\widetilde{\mathbf{p}}$ *and* $\widetilde{\mathbf{p}}'$ *are extreme for ' $\leq_{\mathbf{t}}$ ':*

$$\forall\mathbf{p}\in P,\quad \widetilde{\mathbf{p}}\leq_{\mathbf{t}}\mathbf{p}\leq_{\mathbf{t}}\widetilde{\mathbf{p}}'.$$

Proof: The proof is straight forward using **A** :

$$\begin{aligned}\forall\mathbf{p}\in\mathcal{P},\forall i\in\{1,\ldots,N\}: t_i<0 &\Rightarrow \tilde{p}'_i=p_{\min}\leq p_i\leq p_{\max}=\tilde{p}_i,\\ t_i=0 &\Rightarrow t_i\tilde{p}_i=t_ip_i=t_i\tilde{p}'_i=0,\\ t_i>0 &\Rightarrow \tilde{p}_i=p_{\min}\leq p_i\leq p_{\max}=\tilde{p}'_i\square.\end{aligned}$$

We now have to recall that two random values ξ and ζ are said *stochastically ordered* if and only if their respective distribution functions F and G are ordered:

$$\xi\sim F,\zeta\sim G:\quad \xi\leq_{\text{stoch}}\zeta \quad\Leftrightarrow\quad \forall x\in\mathbb{R},\ F(x)\geq G(x)$$

and we give a lemma which will appear to be useful in the sequel.

Lemma 3.2 *For* $(\xi,\ \xi')$ *and* (ζ,ζ'), *two pairs of real independent random values such as* $\xi\leq_{\text{stoch}}\xi'$ *and* $\zeta\leq_{\text{stoch}}\zeta'$, *one has*

$$\xi+\zeta\leq_{\text{stoch}}\xi'+\zeta'.$$

Proof: The direct calculus of $\mathrm{P}\{\xi+\zeta\leq x\}-\mathrm{P}\{\xi'+\zeta'\leq x\}$ shows that this quantity is non-negative.□

Lemma 3.3 *Under* **B**,

$$\forall(\mathbf{p},\mathbf{p}')\in\mathcal{P}^2,\qquad \mathbf{p}\leq_{\mathbf{t}}\mathbf{p}'\ \Rightarrow\ (T|\mathbf{p})\leq_{\text{stoch}}(T|\mathbf{p}')\,.$$

Proof: The lemma is first proved for two vectors $\mathbf{p}$ and $\mathbf{p}'$ of $\mathcal{P}$ having all their coordinates equal except one (say the first) and such that $\mathbf{p}\leq_{\mathbf{t}}\mathbf{p}'$:

$$t_1(p_1-p'_1)\leq 0,\quad \forall i>1: p_i=p'_i.$$

Then we have clearly $(t_1Z_1|p_1)\leq_{\text{stoch}}(t_1Z_1|p'_1)$.

But **B** says that

$$\begin{aligned}(T|\mathbf{p}) &= \left(\sum_{i>1}t_iZ_i|\{p_i,i>1\}\right)+(t_1Z_1|p_1),\\ (T|\mathbf{p}') &= \left(\sum_{i>1}t_iZ_i|\{p'_i,i>1\}\right)+(t_1Z_1|p'_1).\end{aligned}$$

where $\left(\sum_{i>1}t_iZ_i|\{p_i,i>1\}\right)=\left(\sum_{i>1}t_iZ_i|\,\{p'_i,i>1\}\right)$, $\left(\sum_{i>1}t_iZ_i|\{p_i,i>1\}\right)\perp(t_1Z_1|p_1)$ and $\left(\sum_{i>1}t_iZ_i|\,\{p'_i,i>1\}\right)\perp(t_1Z_1|p'_1)$ and then, lemma **3.2** gives, for any $\mathbf{p}$ and $\mathbf{p}'$ equal except their first coordinates,

$$(T|\mathbf{p}) \leq_{\text{stoch}} (T|\mathbf{p}') \,. \tag{10.4}$$

This proof may be extended to any $(\mathbf{p}, \mathbf{p}')$ such as $\mathbf{p} \leq_{\mathbf{t}} \mathbf{p}'$ using a sequence of $(N+1)$ vectors $\mathbf{p}^{(0)}, ..., \mathbf{p}^{(N)}$ such as $\mathbf{p}^{(i)} = [p_1^{(i)}, ..., p_N^{(i)}]$ where $p_j^{(i)} := p_j$ if $j > i$ and p'_j otherwise.

Since $\mathbf{p} \leq_{\mathbf{t}} \mathbf{p}'$, this sequence verifies $\mathbf{p}^{(0)} \leq_{\mathbf{t}} \mathbf{p}^{(1)} \leq_{\mathbf{t}} ... \leq_{\mathbf{t}} \mathbf{p}^{(N)}$, where $\mathbf{p}^{(i)}$ and $\mathbf{p}^{(i+1)}$ are equal except their first coordinates, and the lemma follows using (10.4).□

Lemma **3.2** means that the pre order on $\mathcal{P}$ induces a stochastic order on T: if $\mathbf{p} \leq_{\mathbf{t}} \mathbf{p}'$ then T is stochastically lower given $\mathbf{p}$ than given $\mathbf{p}'$. The bounds of the real significance level can then be given.

Theorem 3.4 *Under hypothesis* **A** *and* **B**,

$$\forall \mathbf{p} \in \mathcal{P}, \ \alpha_k(\mathbf{t}, \widetilde{\mathbf{p}}) \geq \alpha_k(\mathbf{t}, \mathbf{p}) \geq \alpha_k(\mathbf{t}, \widetilde{\mathbf{p}}').$$

Proof: The proof follows immediately from lemmas **3.1** and **3.2** and from the definition of the significance level $\alpha_k(\mathbf{t}, \mathbf{p})$.□

10.3.1 Fixed number of individuals to be treated

Hypothesis **B** is not always realistic: even when the treatment assignment is not totally controlled, the constraint of the number n of subjects to be treated often holds. Therefore, we make here the hypothesis **C** which replaces **B** :

$$\mathbf{C}: \quad \sum_{i=1}^{N} Z_i = n.$$

This new hypothesis implies that $\mathbf{Z}$ does no longer take its values in $\{0,1\}^N$ but in a subset $\mathcal{Z}_n$:

$$\mathcal{Z}_n := \left\{ \mathbf{z} \in \{0;1\}^N : \sum\nolimits_{i=1}^{i=N} z_i = n \right\},$$

and this obviously affects the distribution of $\mathbf{Z}$, *i.e.* the law of the treatment assignments.

The dependence of the Z_i's entirely modifies the bounding method and especially the form of the extreme vectors $\widetilde{\mathbf{p}}$ and $\widetilde{\mathbf{p}}'$. The bounding is obtained using Hollander *et al.* (1977) results about functions decreasing in transposition.

We shall first introduce a notation: let $\mathbf{v}$ be a N -dimensional vector, ${}_{ij}\mathbf{v}$ denotes the N-dimensional vector in which the i-th and the j-th coordinates of $\mathbf{v}$ are inverted:

$$\begin{array}{rll} {}_{ij}v_k := & v_i & \text{if } k = j, \\ & v_j & \text{if } k = i, \\ & v_k & \text{otherwise.} \end{array}$$

Definition 3.5 *A function g on $\mathbb{R}^N \times \mathbb{R}^N$ is said increasing in transposition if and only if, for any $(\mathbf{a}, \mathbf{b})$ of $\mathbb{R}^N \times \mathbb{R}^N$, (i) and (ii) hold:*

$$\begin{array}{ll} (i) & g(\mathbf{a}, \mathbf{b}) = g({}_{ij}\mathbf{a}, {}_{ij}\mathbf{b}), \\ (ii) & (a_i - a_j)(b_i - b_j) \leq 0 \quad \Rightarrow \quad g(\mathbf{a}, \mathbf{b}) \geq g(\mathbf{a}, {}_{ij}\mathbf{b}). \end{array}$$

The increasing in transposition property means that the function g is greater when the coordinates of $\mathbf{a}$ and $\mathbf{b}$ are arranged in different orders than when they are arranged in the same order.

Lemma 3.6 *Under hypothesis* $\mathbf{C}$ *,* $\alpha_k(\mathbf{t}, \mathbf{p})$ *is increasing in transposition.*
Proof: Condition (i) of definition **3.5** for the function $\alpha_k(t,p)$ is easily found by inverting i-th and j -th coordinates of each vector.

To obtain condition (ii), we just have to show that, when $(t_i - t_j)(p_i - p_j)$ is negative, T is stochastically lower given $\mathbf{p}$ than given ${}_{ij}\mathbf{p}$. Therefore, we use lemma **3.2**. Denoting $T_{ij} := t_i Z_i + t_j Z_j$, one has $T = (T - T_{ij}) + T_{ij}$. To obtain independence, we work conditionally to $S_{ij} := Z_i + Z_j \in \{0, 1, 2\}$. Since
$\mathrm{P}_{\mathrm{H}_3}\{T - T_{ij} = \tau_1, T_{ij} = \tau_2 | S_{ij} = s_{ij}\}$
$= \mathrm{P}_{\mathrm{H}_2}\left\{T - T_{ij} = \tau_1 | \sum_{l \neq i,j} Z_l = n - s_{ij}\right\} \times \mathrm{P}_{\mathrm{H}_2}\{T_{ij} = \tau_2 | S_{ij} = s_{ij}\}$,
one has

$$T_{ij} \overset{S_{ij}}{\perp} (T - T_{ij})$$

For any s_{ij}, we then have for $T - T_{ij}$

$$\mathcal{L}\left\{T - T_{ij} | \sum_{l \neq i,j} Z_l = n - s_{ij}, \mathbf{p}\right\} = \mathcal{L}\left\{T - T_{ij} | \sum_{l \neq i,j} Z_l = n - s_{ij},{}_{ij}\,\mathbf{p}\right\},$$

and for T_{ij}, there are three cases

1: $s_{ij} = 0 \quad \Rightarrow \quad T_{ij} \equiv 0 \;\Rightarrow\; \mathcal{L}\{T_{ij} | S_{ij} = 0, \mathbf{p}\} = \mathcal{L}\{T_{ij} | S_{ij} = 0,{}_{ij}\,\mathbf{p}\}$,

2: $s_{ij} = 2 \quad \Rightarrow \quad T_{ij} \equiv t_i + t_j \;\Rightarrow\; \mathcal{L}\{T_{ij} | S_{ij} = 2, \mathbf{p}\} = \mathcal{L}\{T_{ij} | S_{ij} = 2,{}_{ij}\,\mathbf{p}\}$,

3: $s_{ij} = 1 \quad \Rightarrow \quad T_{ij} = t_i$ or $T_{ij} = t_j$.

In the third case, one has, for example, for $t_i \leq t_j$ and $p_i \leq p_j$

$$\begin{aligned} \mathrm{P}\{T_{ij} = t_i | S_{ij} = 1, \mathbf{p}\} &= \mathrm{P}\{Z_i = 1, Z_j = 0 | S_{ij} = 1, \mathbf{p}\} = \tfrac{p_i(1-p_j)}{p_i(1-p_j)+p_j(1-p_i)}, \\ \mathrm{P}\{T_{ij} = t_i | S_{ij} = 1,{}_{ij}\,\mathbf{p}\} &= \mathrm{P}\{Z_i = 1, Z_j = 0 | S_{ij} = 1,{}_{ij}\,\mathbf{p}\} = \tfrac{p_j(1-p_i)}{p_i(1-p_j)+p_j(1-p_i)} \end{aligned}$$

So, it comes,

$$\begin{aligned} \mathrm{P}\{T_{ij} \leq t_i | S_{ij} = 1, \mathbf{p}\} &\geq \mathrm{P}\{T_{ij} \leq t_i | S_{ij} = 1,{}_{ij}\,\mathbf{p}\}, \\ \mathrm{P}\{T_{ij} \leq t_j | S_{ij} = 1, \mathbf{p}\} &= \mathrm{P}\{T_{ij} \leq t_j | S_{ij} = 1,{}_{ij}\,\mathbf{p}\} = 1, \end{aligned}$$

and generally

$$(t_i - t_j)(p_i - p_j) \leq 0 \quad \Rightarrow \quad (T|\mathbf{p}) \leq_{\text{stoch.}} (T|_{ij}\mathbf{p})\,.\square$$

The bounds on $\alpha_k(\mathbf{t}, \mathbf{p})$ come from this lemma.

Theorem 3.7 *Under hypotheses* **A** *and* **C**, *if the individual scores are given in ascending order* ($t_1 \leq t_2 \leq ... \leq t_N$), *there exist two integers m and m' belonging to* $\{1, ..., N-1\}$ *such as*

$$\forall \mathbf{p} \in \mathcal{P}, \quad \alpha_k(\mathbf{t}, \widetilde{\mathbf{p}}) \geq \alpha_k(\mathbf{t}, \mathbf{p}) \geq \alpha_k(\mathbf{t}, \widetilde{\mathbf{p}}')$$

where $\widetilde{p}_i := p_{\max}\mathbf{1}\{i \leq m\} + p_{\min}\mathbf{1}\{i > m\}$ and $\widetilde{p}'_i := p_{\max}\mathbf{1}\{i > N - m'\} + p_{\min}\mathbf{1}\{i \leq N - m'\}$.

Proof: To show that all coordinates of $\mathbf{p}$ and $\widetilde{\mathbf{p}}$ are equal either to $p_{\min}$ or to $p_{\max}$, it suffices to show that the derivative of $\alpha_k(\mathbf{t}, \mathbf{p})$ with respect to any p_i has a constant sign. Let us calculate the derivative of $\alpha_k(\mathbf{t}, \mathbf{p})$:

$$\alpha_k(\mathbf{t}, \mathbf{p}) = \frac{\sum\limits_{\mathbf{z} \in \mathcal{Z}_n} \left(\mathbf{1}\left\{\sum_i t_i z_i \leq k\right\} \times \prod\limits_i p_i^{z_i}(1-p_i)^{1-z_i} \right)}{\sum\limits_{\mathbf{z} \in \mathcal{Z}_n} \left(\prod\limits_i p_i^{z_i}(1-p_i)^{1-z_i} \right)} \tag{10.5}$$

Let $D_i(\mathbf{p})$ denote the denominator of (10.5), one has

$$\begin{aligned}
D_i(\mathbf{p}) &= p_i \sum_{\mathbf{z} \in \mathcal{Z}_n : z_i = 1} \left(\prod_{j \neq i} p_j^{z_j}(1-p_j)^{1-z_i} \right) + (1-p_i) \sum_{\mathbf{z} \in \mathcal{Z}_n : z_i = 0} \left(\prod_j p_j^{z_j}(1-p_j)^{1-z_j} \right) \\
&= p_i \sum_{j \neq i} \left[(1-p_j) \sum_{\substack{\mathbf{z} \in \mathcal{Z}_n : \\ z_i = 1, z_j = 0}} \left(\prod_{k \neq i,j} p_k^{z_k}(1-p_k)^{1-z_k} \right) \right] \\
&\quad + (1-p_i) \sum_{j \neq i} \left[p_j \sum_{\substack{\mathbf{z} \in \mathcal{Z}_n : \\ z_i = 0, z_j = 1}} \left(\prod_{k \neq i,j} p_k^{z_k}(1-p_k)^{1-z_k} \right) \right] \\
&= p_i \sum_{j \neq i} \left[(1-p_j) K_{ij}(\mathbf{p}) \right] + (1-p_i) \sum_{j \neq i} \left[p_j K_{ij}(\mathbf{p}) \right]
\end{aligned}$$

where $K_{ij}(\mathbf{p}) := \sum_{\mathcal{Z}_n^{ij}} \left({}_k p_k^{z_k}(1-p_k)^{1-z_k} \right)$ with $\mathcal{Z}_n^{ij} := \{(z_k, 1 \leq k \leq N,\ k \neq i,j) : \mathbf{z} \in \mathcal{Z}_n, z_i = 0, z_j = 1\}$. This notation makes sense since the structure of $\mathcal{Z}_n$ is such that the sets $\mathcal{Z}_n^{ij}$ and $\{(z_k, 1 \leq k \leq N, k \neq i,j) : \mathbf{z} \in \mathcal{Z}_n, z_i = 1, z_j = 0\}$ are equal. Then, we have

$$D_i(\mathbf{p}) = p_i \sum_{j \neq i} \left[(1-2p_j) K_{ij}(\mathbf{p}) \right] + \sum_{j \neq i} \left[p_j K_{ij}(\mathbf{p}) \right] \tag{10.6}$$

The numerator of (10.5) is

$$\begin{aligned}
&\sum_{\mathbf{z}\in\mathcal{Z}_n}\left(\mathbf{1}\left\{\sum_i t_i z_i \le k\right\}\times\prod_i p_i^{z_i}(1-p_i)^{1-z_i}\right)\\
&= p_i \sum_{\mathbf{z}\in\mathcal{Z}_n : z_i=1}\left(\mathbf{1}\left\{\sum_i t_i z_i \le k\right\}\times\prod_{j\neq i} p_j^{z_j}(1-p_j)^{1-z_j}\right)\\
&\quad +(1-p_i)\sum_{\mathbf{z}\in\mathcal{Z}_n : z_i=0}\left(\mathbf{1}\left\{\sum_i t_i z_i \le k\right\}\times\prod_{i\neq j} p_j^{z_j}(1-p_j)^{1-z_j}\right)\\
&= p_i Q_i(\mathbf{t},\mathbf{p}) + (1-p_i)Q_i'(\mathbf{t},\mathbf{p}).
\end{aligned}$$

With (10.6), it gives

$$\alpha_k(\mathbf{t},\mathbf{p}) = \frac{p_i\left(Q_i(\mathbf{t},\mathbf{p})-Q_i'(\mathbf{t},\mathbf{p})\right)Q_i(\mathbf{t},\mathbf{p})+Q_i'(\mathbf{t},\mathbf{p})}{p_i\sum_{j\neq i}\left[(1-2p_j)K_{ij}(\mathbf{p})\right]+\sum_{j\neq i}\left[p_j K_{ij}(\mathbf{p})\right]}$$

where $K_{ij}(\mathbf{p})$, $Q_i(\mathbf{t},\mathbf{p})$, $Q_i'(\mathbf{t},\mathbf{p})$ *do not depend on* p_i. We now calculate the derivative of $\alpha_k(\mathbf{t},\mathbf{p})$ with respect to p_i :

$$\left(\frac{\partial\alpha_k}{\partial p_i}\right)_{(\mathbf{t},\mathbf{p})} = \frac{Q_i(\mathbf{t},\mathbf{p})\sum_{j\neq i}p_j K_{ij}(\mathbf{p}) - Q_i'(\mathbf{t},\mathbf{p})\sum_{j\neq i}(1-2p_j)K_{ij}(\mathbf{p})}{\left(D_i(\mathbf{u})\right)^2}$$

and we see that the denominator of this quantity (and then its sign) does depend on p_i.

This result means that vectors $\widetilde{\mathbf{p}}$ and $\widetilde{\mathbf{p}}'$ have all their coordinates equal either to $p_{\min}$ or to $p_{\max}$ but does not say how those coordinates are arranged. We can obviously drop the cases $\widetilde{p}_i \equiv p_{\min}$, $\widetilde{p}_i \equiv p_{\max}$, $\widetilde{p}_i' \equiv p_{\min}$, $\widetilde{p}_i' \equiv p_{\max}$ so that $1 < m < N$ and $1 < m' < N$. Finally, lemma **3.6** says that, if the individual scores are given in ascending order ($t_1 \le t_2 \le \ldots \le t_N$), the first coordinates of $\widetilde{\mathbf{p}}$ are equal to $p_{\max}$ and the last are equal to $p_{\min}$ and symmetrically for $\widetilde{\mathbf{p}}'$. □

The integers m and m' have to be determined numerically but the number of candidate values is only $N-1$.

10.4 Asymptotic normality

The purpose of this section is to show that, even in presence of heterogeneity, the weighted logrank test statistics are asymptotically normal. In this section, we shall insist on the case where the assignments are independent. A reference is given to obtain a similar result when the number of individuals to be treated is fixed.

Lemma 4.1 *Under hypothesis* **A** *and* **B** *, for* $T_N := \sum_{i=1}^{i=N} t_{iN} Z_i$*, one has*

$$\frac{\sup_{i=1..N}\{t_{iN}^2\}}{\sum_{i=1}^{i=N} t_{iN}^2} \underset{N\to+\infty}{\longrightarrow} 0 \quad\Rightarrow\quad \frac{T_N - \mathrm{E}(T_N)}{\sqrt{\mathrm{V}(T_N)}} \;{}_{\mathcal{L}}\underset{N\to+\infty}{\longrightarrow} \mathcal{N}(0,1).$$

Proof: We use the Lindeberg condition (Billingsley (1968), theorem **7.2**) for sums of independent variables : let $\{\xi_{iN}, N \geq 1, 1 \leq i \leq N\}$ be a triangular array of independent variables with zero mean and let us denote $\xi_{iN} \sim P_{iN}$, $S_N := \sum_{i=1}^{i=N} \xi_{iN}$ and $\sigma_N^2 :=\mathrm{V}(S_N)$, one has

$$\frac{S_N}{\sigma_N} \underset{N\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,1) \quad \Leftrightarrow \quad \forall\varepsilon>0, \frac{1}{\sigma_N^2}\int_{\{|\xi_{iN}|>\varepsilon\sigma_N\}} \xi_{iN}^2 \mathrm{d}P_{iN} \underset{N\to+\infty}{\longrightarrow} 0.$$

For $\xi_{iN} := t_{iN}(Z_i - p_i)$, we have $\sigma_N^2 = \sum_{i=1}^{i=N} t_{iN}^2 p_i(1-p_i)$. Under **A** and **B** , $\{\xi_{iN}, N \geq 1, 1 \leq i \leq N\}$ is a triangular array of independent variables with zero mean and, denoting $\sigma_{\min}^2 := \inf\{p_{\min}(1-p_{\min}), p_{\max}(1-p_{\max})\}$, we have $\sigma_N^2 \geq \sigma_{\min}^2 \sum_{i=1}^{i=N} t_{iN}^2$ and then

$$|t_{iN}| \leq \varepsilon\sigma_N \quad \Rightarrow \quad \int_{\{|\xi_{iN}|>\varepsilon\sigma_N\}} \xi_{iN}^2 \mathrm{d}P_{iN} = 0$$

since $|\xi_{iN}| \leq |t_{iN}|$. So, the Lindeberg condition is verified as long as every t_{iN}^2 remains negligible with respect to $\sum_{i=1}^{i=N} t_{iN}^2$.$\square$

We now have to show that the condition of lemma **4.1** holds for classical weighted logrank test statistics.

We need to introduce a few notations. Let $\mathcal{T}_N$ be the set of observed death times and K_N its cardinal : $\mathcal{T}_N := \{y_i, 1 \leq i \leq N : m_i > 0\} = \{\tau_1, ..., \tau_{K_N}\}$ where $\tau_1 < ... < \tau_{K_N}$, μ_k the number of deaths at time τ_k : $\mu_k := \sum_{i=1}^{i=N} d_i \mathbf{1}\{y_i = \tau_k\}$ (for any k, $\mu_k > 0$) and ρ_k the number of deaths observed in τ_k or after : $\rho_k := \sum_{i=1}^{i=N} d_i \mathbf{1}\{y_i \geq \tau_k\} = \sum_{h=k}^{h=K_N} \mu_h$. We shall denote further $\rho (= \rho_1)$ the total number of observed deaths : $\rho := \sum_{i=1}^{i=N} d_i = \sum_{k=1}^{k=K} \mu_k$.

Let us introduce two classical hypotheses :

$$\mathbf{D}: \quad \lim_{N\to+\infty} \frac{\rho}{N} = \lambda > 0$$

$$\mathbf{E}: \quad \lim_{N\to+\infty} \frac{\sup_{k=1..K_N} \{\mu_k\}}{\rho} = 0.$$

D means that the total number of observed deaths never becomes negligible with respect to N and **E** means that the number of deaths observed at any time μ_k remains negligible with respect to the total number of deaths. We may remark that **E** implies that K_N grows indefinitely with N.

Theorem 4.2 *Under hypotheses* **A**, **B**, **D** *and* **E** *the Gehan, logrank, Tarone & Ware and Harrington & Fleming test statistics are asymptotically normal.*

Proof: We use the condition of lemma **4.1**. After several developments, the numerator gives

$$\begin{aligned}
\sum_{i=1}^{N} t_{iN}^2 &= \sum_{i=1}^{N} d_i \left(\frac{w_i}{r_i}\right) r_i \left(r_i - m_i\right) \\
&\quad \text{but } r_i \text{ is greater than the corresponding } \rho_k \\
&\geq \sum_{k=1}^{K_N} \mu_k q_k^2 \rho_k \left(\rho_k - \mu_k\right) \\
&\quad \text{where } q_k \text{ denotes } \frac{w_i}{r_i} \text{ for any individual } i \text{ dead in } \tau_k \\
&\geq q_1^2 \sum_{k=1}^{K_N} \mu_k \rho_k \rho_{k+1} \\
&\quad \text{since } \rho_{k+1} = \rho_k - \mu_k \text{ and since } q_k \text{ grows with } k.
\end{aligned}$$

And, using the definition of ρ_k, it comes

$$\sum_{k=1}^{K_N} \mu_k \rho_k \rho_{k+1} = \frac{1}{3}\left(\rho^3 - \sum_{k=1}^{K_N} \mu_k^3\right). \tag{10.7}$$

Equation (10.7) is obtained considering the left term as a stack of parallelepipeds with respective dimensions $(\mu_k \times \rho_k \times \rho_{k+1})$ and the right term as the global volume of the stack. Hypothesis **E** ensures us that this quantity is equivalent to $\rho^3/3$, *i.e.* to $(\lambda N)^3/3$ after **D**.

We now have to study q_1^2 for each classical statistic to obtain an equivalent of a lower bound for the numerator :

logrank	$q_1^2 = \left(\frac{1}{r_1}\right)^2 \geq \frac{1}{N^2}$	$\Rightarrow \sum_{i=1}^{N} t_{iN}^2 \geq \frac{\lambda^3}{3} O(N),$
Gehan	$q_1^2 = 1$	$\Rightarrow \sum_{i=1}^{N} t_{iN}^2 \geq \frac{\lambda^3}{3} O(N^3),$
Tarone & Ware	$q_1^2 = \left(\frac{1}{\sqrt{r_1}}\right)^2 \geq \frac{1}{N}$	$\Rightarrow \sum_{i=1}^{N} t_{iN}^2 \geq \frac{\lambda^3}{3} O(N^2),$
Harrington & Fleming	$q_1^2 = \left(\frac{1}{r_1}\right)^2 \geq \frac{1}{N^2}$	$\Rightarrow \sum_{i=1}^{N} t_{iN}^2 \geq \frac{\lambda^3}{3} O(N).$

Let us now look at the denominator. To obtain the condition of lemma **4.1**, we just have to give an upper bound for the denominator which remains negligible with respect to the numerator.

Since the w_i's are positive and noticing that $t_{iN} = d_i w_i - \sum_j (d_j w_j / r_j) \times \mathbf{1}\{y_j \leq y_i\}$, we have

$$\forall i, |t_{iN}| \leq \max\left\{w_1, \sum_{j=1}^{N} \frac{d_j w_j}{r_j}\right\}.$$

We then have an equivalent of an upper bound for the greatest t_{iN}^2 for each classical statistic : $O\left(\ln^2(\lambda N)\right)$ for the logrank, $O\left(N^2\right)$ for Gehan, $4O\left(N\right)$ for Tarone & Ware and $O\left(\ln^2(\lambda N)\right)$ for Harrington & Fleming. The comparison of those upper bounds' equivalents with the respective lower bounds' equivalents of the denominator insures us that all classical statistics verify the condition of lemma **4.1**.$\square$

To obtain a similar result when the number of individuals to be treated is fixed, we have to consider the statistic $T = \sum_{i=1}^{i=N} t_i Z_i$ as a weighted sum of Bernoulli random values and to study the law of T given the fact that the unweighted sum $\sum_{i=1}^{i=N} Z_i$ is equal to a certain n. The mean and variance of T can be directly calculated but give very heavy formulas. The paper of Barbe (94) gives conditions of asymptotic normality for such statistics.

10.5 Application to Freireich's data

We now present the sensitivity analysis of weighted logrank tests on Freireich's data (*cit.in.* Gehan (1965)). The purpose of this classical study is to test the effect of 6-mercaptopurine against acute leukemia. Among the $N = 42$ subjects, $n = 21$ have been treated with 6-mp ($Z = 1$) and $N - n = 21$ have received a placebo ($Z = 0$). The observed durations are : 6, 6, 6, 6^+, 7, 9^+, 10, 10^+, 11^+, 13, 16, 17^+, 19^+, 20^+, 22, 23, 25^+, 32^+, 32^+, 34^+, 35^+ for treated individuals and : 1, 1, 2, 2, 3, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23 for controls. Those durations are given in week and ($^+$) means that the duration is censored.

10.5.1 Individual scores

Figures 10.1 and 10.2 give the histogram of the individual scores t_i's for the logrank and the Gehan tests. We find here an illustration of a classical property of censored durations : individuals censored at successive time behave as if they were *ex-aequo*. We can see that scores t_i's decrease as durations y_i's grow but this phenomenon holds separately among censored and non-censored subjects, censored subjects always having a negative score (cf. remark at the end of section **2**. We can also notice that the order of the individual scores is not the same for different statistics : for example, we have $t_{(a)} < t_{(b)}$ for logrank and $t_{(b)} < t_{(a)}$ for Gehan.

10.5.2 Sensitivity analysis

Table 10.1 gives the values (denoted k) taken by each classical statistic in Freireich's study and, under **B** and **C**, the null variances and significance levels (computed with the normal approximation).
Under the classical equiprobability hypothesis, none of those tests allows us to accept $\mathbf{H}_0$ at level 5%.

For the sensitivity analysis, we suppose that the heterogeneity factor takes its values in $[0;1]$ and that the probability function has the form

$$p(u) := \frac{e^{\gamma u}}{1 + e^{\gamma}}.$$

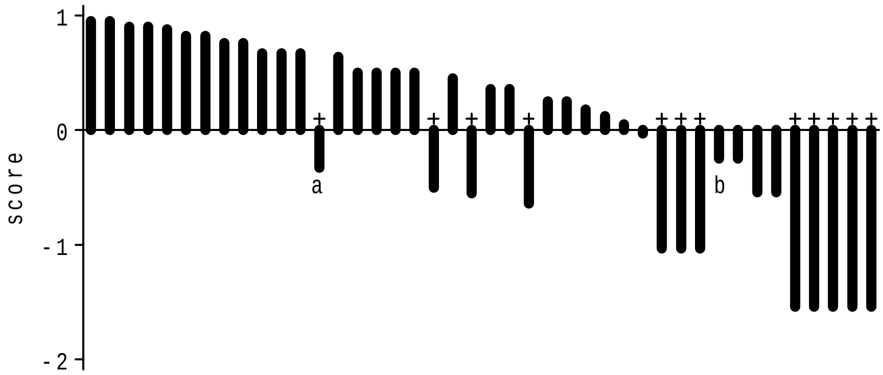


Figure 10.1: Scores of the logrank test on Freireich's dataset('+' = censored).

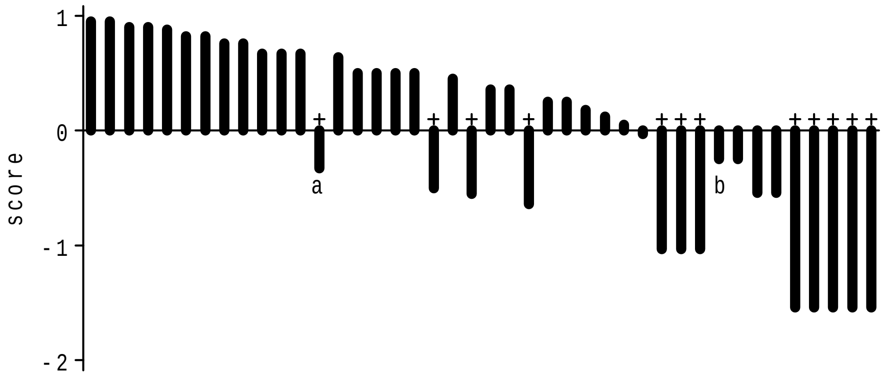


Figure 10.2: Scores of the Gehan test on Freireich's dataset ('+' = censored).

Since $u \in [0;1]$, we have $1/(1+e^{\gamma}) \leq p(u) \leq e^{\gamma}/(1+e^{\gamma})$ and we shall call *sensitivity parameter*, the parameter e^{γ} which measures the influence of U on the treatment assignment : $e^{\gamma}=1$ corresponds to the equiprobable case $p_i \equiv p$ and e^{γ} grows as heterogeneity becomes stronger. Remark that $e^{\gamma}=p_{\max}/p_{\min}$.

The use of a particular probability function is not necessary to lead the analysis : we could compute the significance level for all possible couples ($p_{\min}, p_{\max}$). However, the use of a sensitivity parameter makes the results more interpretable.

In order to analyze the sensitivity of the different tests, we will study how their significance level is modified when e^{γ} grows and, especially, for which value of e^{γ} the significance level will reach 5%. Table 10.2 gives the upper bound $\alpha_k(\mathbf{t},\widetilde{\mathbf{p}})$ of the real (and unknown) significance level $\alpha_k(\mathbf{t},\mathbf{p})$ of the classical weighted logrank tests under **B** and **C** for different values of e^{γ}.

For example, under hypothesis **B**, we read here that if there is an heterogeneity such

	logrank	Tarone & Ware	H. & F. ($\rho = 0.5$)	H. & F. ($\rho = 1$)	Gehan
k	-10.25	-51.16	-8.24	-6.88	-271
$\mathrm{V}_{\mathbf{H}_0,\mathbf{B}}(T)$	6.73	178.25	4.49	3.33	5510
$\mathrm{P}_{\mathbf{H}_0,\mathbf{B}}\{T \leq k\}$	0.00004	0.00006	0.00005	0.00008	0.00013
$\mathrm{V}_{\mathbf{H}_0,\mathbf{C}}(T)$	6.90	182.60	4.60	3.41	5644.39
$\mathrm{P}_{\mathbf{H}_0,\mathbf{C}}\{T \leq k\}$	0.00005	0.00008	0.00006	0.00010	0.00015

Table 10.1: Observed values, variances and significance levels of several statistics on Freireich's dataset

e^γ	logrank	Tarone & Ware	H. & F. ($\rho = 0.5$)	H. & F. ($\rho = 1$)	Gehan
			$\mathrm{P}_{\mathbf{H}_0,\mathbf{B}}\{T \leq k \mid \widetilde{\mathbf{p}}\}$		
1.0	0.00004	0.00006	0.00005	0.00008	0.00013
1.5	0.00188	0.00301	0.00244	0.00364	0.00536
2.0	0.01302	0.02025	0.01681	0.02377	0.03299
2.5	0.04121	0.06219	0.05258	0.07131	0.09474
			$\mathrm{P}_{\mathbf{H}_0,\mathbf{C}}\{T \leq k \mid \widetilde{\mathbf{p}}\}$		
1.0	0.00005	0.00008	0.00006	0.00010	0.00015
1.5	0.00221	0.00360	0.00294	0.00434	0.00631
2.0	0.01466	0.02396	0.01998	0.02796	0.03533
2.5	0.04488	0.07253	0.06161	0.08268	0.10841

Table 10.2: Upper bounds of the significance levels of several statistics on Freireich's dataset

that $p_{\max}/p_{\min} = 2.5$, the real significance level may be 0.09474 (if the individuals' repartition between the two groups is the most favorable to the treatment) and yet the classical p-value is evaluated at 0.00013.

We can see here, and it can be observed on many other examples, that the logrank test is the less sensitive and the Gehan test, the most sensitive. In a general way, it seems that the more contrasted are the w_i's, the more sensitive is the test.

Acknowledgments. I want to acknowledge Professor C. Huber for all her useful advises and suggestions.

Bibliography

Billingsley, P. (1968). *Convergence of probability measures.* Wiley.

Cox, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc. B.* **34** 187–220. (with discussion).

Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika.* **52** 203–23.

Harrington, D. and Fleming, T. (1982). A class of rank test procedures for censored survival data. *Biometrika.* **69** 553–66.

Hollander, M., Proschan, F. and Sethuraman, J. (1977). Functions decreasing in transposition and their applications in ranking problems. *Ann. Statist.* **5** 722–33.

Rosenbaum, P. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika.* **74** 13–26.

Chapitre 11

A model for thermograms

ROBIN S. (1999)

Biometrics, **55** (1), 37–43

Abstract. Thermograms are curves resulting from thermal analysis and are of great interest in the study of various food and biological products physical properties. A method to separate underlying peaks is proposed, and statistical properties of estimates for some characteristic parameters are derived. The total number of peaks can be estimated with a sequential analysis of the residual plots. For each new peak, a statistical criterion is proposed to check if it is significantly different from the noise of the recording. As an example, the method is applied to a summer milk fat fusion thermogram.

Keywords. nonlinear regression, peaks separation, thermal analysis

Contents

11.1 Introduction

Thermal analysis brings informations on the structural and rheological behavior of biological and food products as a function of the temperature. The results of this method are generally presented as a curve called *thermogram* (or Differential Scanning Calorimetry −DSC− recording) where the horizontal axis value is the temperature and the vertical axis value is the enthalpy.

A thermogram is essentially composed of several peaks each representing a component of the considered product. The location of a peak identifies the underlying component and the area gives its proportion. A reliable separation method of the peaks in therefore necessary to derive satisfying quantitative informations on each component from thermal analysis.

Up to now, analysis of thermograms has remained essentially empirical because of the lack of adapted statistical methods and models. The problem of the estimation of the areas is particularly important and very poorly solved with the common empirical methods. For complex thermograms (e.g. with numerous or overlapping peaks) such as milk fat DSC, even the location parameters are very inaccurately estimated with these methods.

Ritter 7 investigates a similar problem for curves obtained by electron spectroscopy: he proposes a Gaussian-Lorentzian model with a *fixed* number of peaks and uses a *Bayesian* estimation method to incorporate some auxiliary informations in the model. In the end, he discusses the consequences of ignoring the prior information on the quality of estimates.

The model we propose here allows to separate numerous peaks and gives very good fitting approximation of the whole thermogram. Figure 11.1 shows an example of its application: such a decomposition is of great help for physicists to infer the composition of the product.

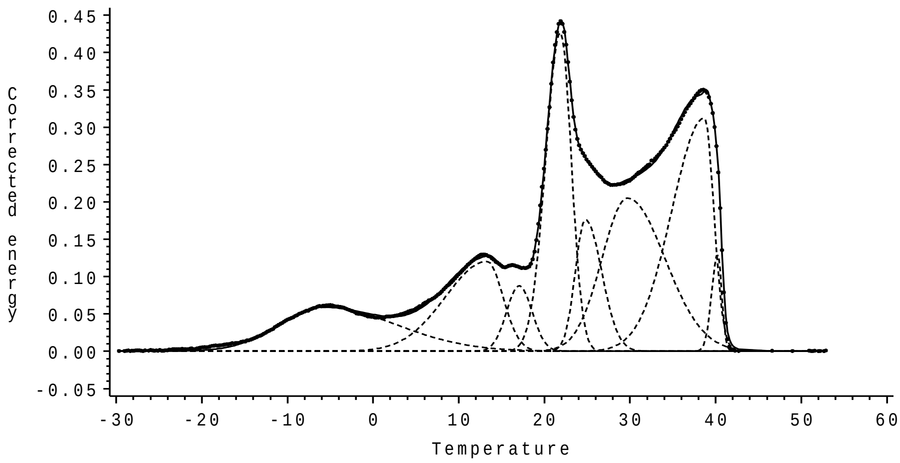


Figure 11.1: 8-peak decomposition plot.

Note that the model we present here is *not a mixture model*. The literature on mixtures is abundant (see for example Roeder 6 or the general literature on classification) but is not suitable for our problem since we do not have a random sample issued from a mixture distribution. The data we deal with are a sequence of regularly spaced abscissas with corresponding random ordinates so we use a *nonlinear regression model.* Despite the model uses 'Gaussian' peaks, it has nothing to do with any density probability function.

In section **2**, several models for a single peak are presented and a choice is made among them, the general model is then constructed when the number of peaks is known. The different steps of the parameters estimation are exposed in section **3** and the statistical properties of the estimates are derived. A sequential method to determine the number of peaks is then proposed in section **4** and section **5** presents an application to a summer milk fat stearic fraction.

11.2 The model

11.2.1 Model for one peak

Peak shape: The physical theory suggests that a peak is made of a linearly ascending part starting from a point called *onset* and an exponentially decreasing part. This description gives the so-called *linear-exponential* model

$$\begin{aligned}\varphi(t;\theta) &= h(t-o)/(s-o) && \text{if } t \le s,\\ &= h\exp\{-v(t-s)\} && \text{otherwise}\end{aligned}$$

where θ is the vector containing the characteristic parameters of the peak. Here $\theta = (o, s, h, v)'$ where h denotes the height (i.e. the maximum) of the peak, s (for *summit*) the temperature corresponding to this maximum, o the temperature of the onset and v the exponential return rate to the baseline. In this model, the area of the peak is $a = h\{(s-o)/2 + v^{-1}\}$.

Before proposing any other model we shall remark that, according to the physical theory, four seems to be the good number of parameters for one peak since the model has to give: the place where the peak stands (i.e. the location parameter), its height, and the increasing and decreasing rates. A model with less parameters would necessarily imply assumptions on the shape of the peak such as symmetry.

Another possible model is the *Gaussian-exponential* model

$$\begin{aligned}\varphi(t;\theta) &= h\exp\{-\lambda^2(t-s)^2/2\} && \text{if } t \le s,\\ &= h\exp\{-v(t-s)\} && \text{otherwise.}\end{aligned}$$

Here $\theta = (\lambda, s, h, v)'$ where λ represents the thickness of the ascending part of the peak: in this model, as observed on real thermograms, there is no rupture of the slope at the onset.

The last model proposed here is the *Gaussian-Gaussian* model

$$\begin{aligned}\varphi(t;\theta) &= h\exp\{-\lambda^2(t-s)^2/2\} && \text{if } t \le s,\\ &= h\exp\{-\gamma^2(t-s)^2/2\} && \text{otherwise}\end{aligned}$$

where $\theta = (\lambda, s, h, \gamma)'$ and γ represents the thickness of the descending part.

The different shapes obtained with these models can be seen on Figure 11.2. As we shall see in the following, the rupture at the top of the peak observed with the first two models is not very realistic.

Comparison of the different shapes: To explain the choice of the single peak function $\varphi(\cdot;\theta)$, we compare the different models in the case of a pure product which is a reference among fatty acids: the lauric acid.

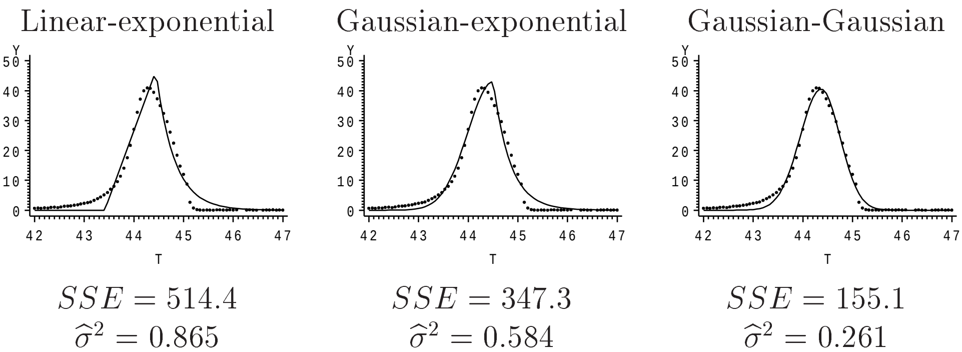


Figure 11.2: Model for the single peak of pure lauric acid. Dots: observed data, solid line: fitted curve.

Figure 11.2 presents the fitted curves obtained with the three models and shows that the hypothesis of a linear increasing part is a bad approximation, likewise the exponential decreasing part. The different models are compared according to their residual sums of squares and estimated variances of errors (SSE and $\widehat{\sigma}^2$: precisely defined in next section). We see that the Gaussian-Gaussian model is more than three times better than the linear-exponential and more than twice better than the Gaussian-exponential. Many other examples give similar results, we shall therefore use the Gaussian-Gaussian model in the following.

11.2.2 General model

We study here a nonlinear regression model for a thermogram with K peaks (K being known)

$$Y(t;\omega) = \Phi(t;\omega) + E(t)$$

where t is the temperature, ω the set of parameters and where the error terms $E(t)$ are supposed to be independent and normally distributed.

Baseline: During the fusion, the properties of a product are modified because its components melt one after each other. This modification implies that the slope of the baseline (which is related to the specific heat $-C_p-$ of the product) changes all along the thermogram. In the common empirical methods, this baseline $b(t)$ is approximated by a broken line but this method gives bad results at the breaking points. We use here an approximated baseline made of two lines and a third-degree spline polynomial joining them

$$\begin{aligned} b(t) &= a_{\min}t + c_{\min} && \text{if } t < t_{\min}, \\ &= b_0 + b_1 t + b_2 t^2 + b_3 t^3 && \text{if } t_{\min} \le t \le t_{\max}, \\ &= a_{\max}t + c_{\max} && \text{if } t > t_{\max}. \end{aligned}$$

The points $t_{\min}$ and $t_{\max}$ are chosen in the linear parts of the thermogram, respectively at the beginning and at the end. Coefficients $(a_{\min}, c_{\min})$ and $(a_{\max}, c_{\max})$ are the slopes and intercept of the respective tangents in $t_{\min}$ and $t_{\max}$ and (b_0, b_1, b_2, b_3) are related to $(a_{\min}, c_{\min}, a_{\max}, c_{\max})$ by the usual relations insuring the continuity of $b(t)$ and $b'(t)$. This expression gives a smoother (and more realistic) shape to the baseline.

The introduction of the baseline in the model will be briefly discussed in the conclusion. In the following, we shall consider that the baseline has been removed from the thermogram, so we study here the corrected enthalpy

$$Y(t) \to Y(t) - b(t).$$

Model for K peaks: The particularity of this model is that the function $\Phi(\cdot;\omega)$ is the sum of K functions $\varphi(\cdot;\theta_k)$ each corresponding to one peak

$$\Phi(t;\omega) = \sum_{k=1}^{K} \varphi(t;\theta_k), \qquad \omega = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}.$$

ω is the set of parameters describing the whole thermogram.

The underlying physical hypothesis is that the thermogram is the simple sum of the K peaks each corresponding to one component. We shall see in the following that, in spite of its simplicity, such a model gives very good results.

Notice that this model is identifiable as are Gaussian mixtures: there is a unique parameter ω giving the curve $y = \Phi(t;\omega)$ for all t.

11.3 Parameter estimation

In the following, for any parameter ψ we shall denote: ψ its true value, $\widetilde{\psi}$ its evaluation with an empirical method, ψ^* the starting value of the nonlinear regression algorithm and $\widehat{\psi}$ its maximum likelihood estimator.

11.3.1 Nonlinear regression: estimation of the θ_k's

Let C be the curve to be modeled and K the number of peaks. C is made of n points: $C = \{(t_i, y_i), i = 1..n\}$. We use the maximum-likelihood estimator which, under the normal hypothesis, is the same as the least-squares estimator: $\widehat{\omega}$ minimizes

$$SSE(\mathbf{u}) = \sum_i (y_i - \Phi(t_i, \mathbf{u}))^2.$$

All computations are made with the *NLIN* procedure of the *SAS* software, generally with the Marquardt algorithm. This algorithm needs the derivatives of $\Phi(t;\omega)$ according to each element of ω for any value of t. These derivatives have rather simple forms, except on the rupture points where the derivatives according to s_k may be infinite, so we have to assume that no point t_i of the observed curve stands at the exact x-value s_k of the rupture of any peak.

Data transformation: To avoid heteroscedasticity for $E(t)$, we shall estimate the regression model

$$Z(t) = \log\{1 + \Phi(t;\omega)\} + E(t)$$

where $Z(t) = \log\{1 + Y(t)\}$ instead of modeling directly $Y(t)$; notice that the constant 1 is chosen arbitrarily. This transformation is coherent with the hypothesis that the variance of the error term $E(t)$ increases with the signal $Y(t)$ (see Huet, Jolivet & Messéan 4 or Seber & Wild 8).

Confidence intervals: The nonlinear least-squares estimator is asymptotically unbiased and normal (see Gallant 3). Denoting $\widehat{\omega}$ the estimator of ω, we have $\sqrt{n}\,(\widehat{\omega} - \omega) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, \mathbf{I}(\omega)^{-1}\sigma^2\right)$ where $\mathbf{I}(\omega)$ is the Fisher's information matrix, and approximately

$$(\widehat{\omega} - \omega) \sim \mathcal{N}\left(\mathbf{0}, \left(\mathbf{F}(\omega)'.\mathbf{F}(\omega)\right)^{-1}\sigma^2\right)$$

where the general term of $\mathbf{F}(\omega)$ is $\mathbf{F}_{ik}(\omega) = \dfrac{\partial\Phi(t_i;\omega)}{\partial\omega_k}$.

This gives the variances of the maximum-likelihood estimators $\widehat{\lambda}_k$, $\widehat{s}_k$, $\widehat{h}_k$ and $\widehat{\gamma}_k$ for any of the K peaks ($k \in 1..K$). These results joined to the classical unbiased estimate $\widehat{\sigma}^2 = (n-4K)^{-1}\times\sum_i (y_i - \Phi(t_i;\widehat{\omega}))^2$ give asymptotic confidence intervals for each parameter.

11.3.2 Initialization

ANATHERM software: The ANATHERM (for '*Thermograms Analysis*') software has been developed in order to exploit the results of the many differential thermic analyses of anhydrous milk fat. Here, this software is used to obtain the starting values ω^* needed by the nonlinear estimation algorithm.

ANATHERM computes characteristics elements of a curve such as: its derivative at any point, the greatest and smallest derivatives on a given interval, the area under the curve between two given points, the x- and y-value of the maximum of the curve in a given interval, *etc* . The derivatives are estimated with a weighted least-squares regression and the weight function is user-defined. For a general presentation, see Robin *et al.* 5.

Initial values: For each peak, the abscissa of the onset $(\widetilde{o})$ and of the maximum $(\widetilde{s})$, the ordinate of the maximum $(\widetilde{h})$ and the greatest ascending and descending slopes $\widetilde{p}_{\mathrm{asc}}$ and $\widetilde{p}_{\mathrm{desc}}$ (with corresponding abscissas $\widetilde{t}_{\mathrm{asc}}$ and $\widetilde{t}_{\mathrm{desc}}$) are evaluated with ANATHERM (the onset is defined as the intersection of the greatest ascending tangent with the baseline).

These coefficients will give the starting values ω^* of the nonlinear regression algorithm but, for some of them, the baseline effect has to be corrected: $o^* = \widetilde{o}$, $s^* = \widetilde{s}$, $h^* = \widetilde{h} - b(\widetilde{s})$. The different models starting values for the other parameters of are obtained using the properties of the functions φ: denoting $p^*_{\mathrm{asc}} = \widetilde{p}_{\mathrm{asc}} - b'(\widetilde{t}_{\mathrm{asc}})$ and $p^*_{\mathrm{desc}} = \widetilde{p}_{\mathrm{desc}} - b'(\widetilde{t}_{\mathrm{desc}})$, one has for the exponential descending part $v^* = -p^*_{\mathrm{desc}}/h^*$, for the Gaussian ascending part $\lambda^* = \sqrt{e}p^*_{\mathrm{asc}}/h^*$ and for the Gaussian descending part $\gamma^* = -\sqrt{e}p^*_{\mathrm{desc}}/h^*$.

11.3.3 Estimation of the onset and area

The estimation of the onset and of the area for any peak is essential to identify the underlying component and to evaluate its proportion in the product. Since the parameters a and o do not explicitly appear in the function $\varphi(\cdot;\theta)$, they have to be estimated using $\widehat{\lambda}$, $\widehat{s}$, $\widehat{h}$, $\widehat{\gamma}$ (index k is omitted here to avoid too heavy notations).

Onset: According to the physical practice, the onset has been defined as the intersection of the greatest ascending tangent with the baseline. It is clear that a Gaussian peak has no real onset but this parameter is very useful to locate the 'significant' beginning of the peak. The greatest ascending tangent is reached when $\partial^2\varphi(t;\theta)/\partial t^2 = 0$ i.e. when $t_{\text{asc}} = s-\lambda^{-1}$; at this point, the slope of the tangent is $\partial\varphi(t;\theta)/\partial t|_{t_{\text{asc}}} = \sqrt{2/e}h\lambda^{-1}$ and this tangent intercepts the baseline at the abscissa $o = s - 2\lambda^{-1}$. The natural estimator of the onset is then

$$\widehat{o} = \widehat{s} - 2\widehat{\lambda}^{-1}.$$

Area: The theoretical area value of a peak is $a = \varphi(t;\theta)\mathrm{d}t = \sqrt{\pi/2}h\left(\lambda^{-1}+\gamma^{-1}\right)$ so that a natural estimator of the area is

$$\widehat{a} = \sqrt{\pi/2}\widehat{h}\left(\widehat{\lambda}^{-1}+\widehat{\gamma}^{-1}\right).$$

Bias and variance approximation of $\widehat{o}$ and $\widehat{a}$: The biases and variances of these estimators may be approximated with the Δ-method using the second order Taylor expansion of $\widehat{o}$ and $\widehat{a}$ about (λ, s, h, γ). The first and second order terms of the expansion give respectively the variance and the bias (see appendix **A**).

We have to note that nonlinear least-squares estimators are only asymptotically unbiased and may be biased for finite samples even for rather simple models as it is shown in Box 1. Therefore, the variances and biases expressions given in appendix are only approximations.

In the example of section **5** as in many others the approximate biases of both $\widehat{o}$ and $\widehat{a}$ remain negligible (see Table 11.1).

Other possible parametrizations: The model could have been parametrized with the respective inverses λ' and γ' of λ and γ to obtain simpler forms for both a and o ($a = h\sqrt{\pi/2}\,(\lambda'+\gamma')$, $o = s-2\lambda'$); then $\widehat{o}$ would be asymptotically unbiased and the variance and bias of $\widehat{a}$ calculated with the Δ-method have much simpler forms. Despite this parametrization seems to be equivalent to the previous one, it leads to numerical instabilities, especially in case of very thin peaks i.e. when λ' or γ' is very small.

Another way of writing the model is to introduce directly the parameters a and o using $\lambda = 2(s-o)^{-1}$ and $\gamma = [a\sqrt{2/\pi} + h(s-o)/2]^{-1}$ but this parametrization also leads to condition number problems and also to very slow convergence of the algorithm.

The first parametrization shall therefore be maintained and both o and a will be estimated with the expressions of $\widehat{o}$ and $\widehat{a}$ given above.

11.4 Determination of the number of peaks

Iterative method: Because of its empirical character, the initialization step presented above is generally insufficient to find all peaks in complex thermograms. An easy way to exhibit new peaks is to study the residual plots. After a K-peak regression model has been fitted to the curve, the residual plot $\widehat{e}_{(K)}(t) = y(t) - \Phi(t, \widehat{\omega}_{(K)})$ can be analyzed as a fictitious thermogram in which new peaks appear. The initial value θ^*_{K+1} for the parameter of the $(K+1)$-th peak may then be evaluated (for example with ANATHERM) and a $(K+1)$-peak model can be fitted with the initial parameter $\omega^*_{(K+1)} = \begin{pmatrix} \widehat{\omega}_{(K)} \\ \theta^*_{K+1} \end{pmatrix}$.

An important practical point is to know if the first step has to be very accurate or may remain quite coarse. Empirically it appears that the results are much more satisfying when as many obvious peaks as possible are detected during the initialization step. On the other hand, even if we are only interested in few main peaks, it seems convenient to establish a complete decomposition to avoid confusion with undetected peaks and get more precise estimates for major ones.

This iterative method needs a stopping rule based on a statistical criterion to see if the new peaks brings up a significant information.

Test for nested models: For each new peak, we have to check if it is significantly different from the noise of the recording. So we need a test for nested model to decide if the $(K+1)$-th peak is a real one rather than a simple artefact.

In order to test a model with K_0 peaks versus a model with K peaks $(K_0 < K)$, we use the likelihood ratio test which leads to a classical Fisher statistics (see Seber & Wild 8, section **5.3**). Formally, if $\omega = \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}$ (ω_1 and ω_2 having respectively $4K_0$ and $4(K - K_0)$ coordinates), we want to test $\mathbf{H}_0 = \{\omega_2 = \mathbf{0}\}$ versus $\mathbf{H}_1 = \{\omega_2 \neq \mathbf{0}\}$. Denoting $\widehat{\omega}_{(K)}$ the estimate under $\mathbf{H}_1$ and $\widehat{\omega}_{(K_0)}$ the estimate under $\mathbf{H}_0$, the likelihood ratio test statistic is $LR = \left(SSE(\widehat{\omega}_{(K_0)})/SSE(\widehat{\omega}_{(K)})\right)^{n/2}$ and is equivalent to

$$F = \frac{SSE\left(\widehat{\omega}_{(K_0)}\right) - SSE\left(\widehat{\omega}_{(K)}\right)}{SSE\left(\widehat{\omega}_{(K)}\right)} \frac{n - 4K}{4\left(K - K_0\right)}$$

which is approximately distributed as $\mathcal{F}_{n-4K, 4(K-K_0)}$.

We have to point out that, since there is an implicit constraint on the parameter h_k which has to be positive, the true value of ω under $\mathbf{H}_0$ is on the boundary of the allowed parameter space. Self & Kung-Yee 9 show that, in this case, the likelihood ratio test statistic is distributed as a mixture of χ^2 and not as a simple χ^2 so F is not exactly distributed as $\mathcal{F}_{n-4K, 4(K-K_0)}$. Nevertheless, the F statistic given above can be directly interpreted as the ratio of the last $(K - K_0)$ peaks mean square to the complete model error variance. Furthermore, one of the referee remarks that the use of the $\mathcal{F}$ distribution generally induces a conservative test so $\mathbf{H}_0$ is not rejected to often.

11.5 Application to a summer milk fat's stearic fraction

We present here an application of this method to the fusion thermogram at 5°C/mn of a summer anhydrous milk fat stearic fraction. The recording contains $n = 599$ points. An ANATHERM computation gives initial parameters for 4 peaks but this decomposition is obviously incomplete: Figure 11.3 gives the fitted 4-peak decomposition ($\widehat{\sigma}^2 = 42.0 \times 10^{-6}$).

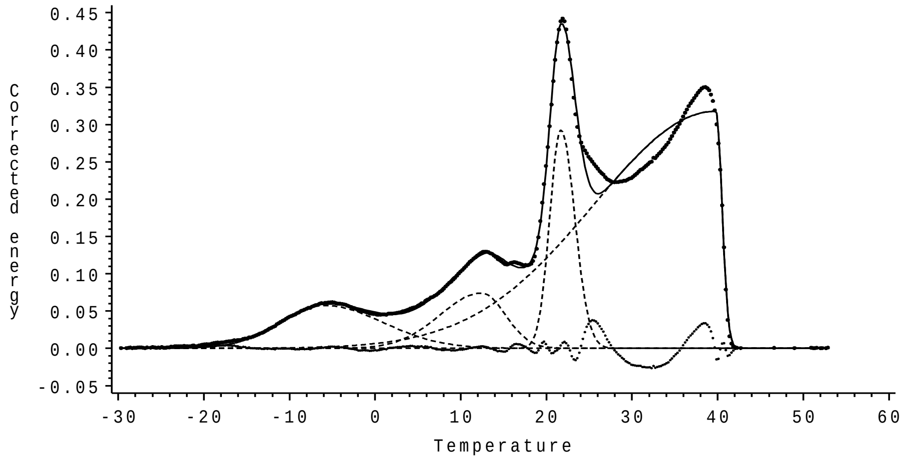


Figure 11.3: Original 4-peak decomposition and residual plots. Dots: observed data, solid line: fitted curve (solid line), dashed lines: peaks, small dots: residuals.

The analysis of the residual plot suggests a fifth peak at $s \simeq 26$°C. This peak appears to be highly significant ($F > 500$): Figure 11.4 shows the 5-peak decomposition and residual plots ($\widehat{\sigma}^2 = 8.24 \times 10^{-6}$).

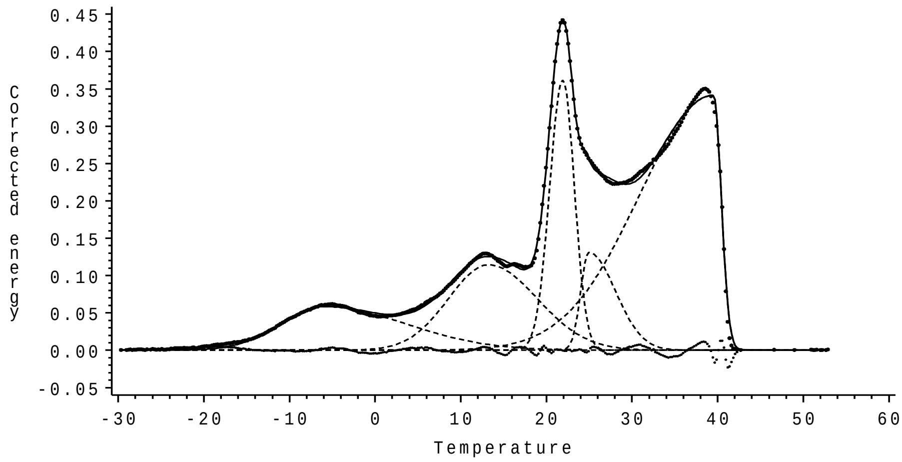


Figure 11.4: 5-peak decomposition and residual plots (same legend as figure 11.3).

New significant peaks can be detected among the successive residual plots at $s \simeq 40°C$, $s \simeq 30°C$ and $s \simeq 17°C$; so the sequential method leads up to a 8-peak decomposition (Figure 11.1: $\widehat{\sigma}^2 = 3.16 \times 10^{-6}$).

Table 11.1 gives the estimates of the main characteristics for the second peak ($s \simeq 13$). We see that the estimates obtained with the classical empirical method (after correction of the baseline effect) are acceptable for the summit's temperature, poor for the height and very inaccurate for both the onset and the area. This shows the interest of a precise modelling of the thermogram. Note that the estimated biases of $\widehat{o}$ and $\widehat{a}$ are very small.

	Summit	Height	Onset	Area
Empirical estimate	$s^* = 13.0$	$h^* = 0.140$	$o^* = 1.88$	$a^* = 1.71$
Least-squares estimate	$\widehat{s} = 13.3$	$\widehat{h} = 0.120$	$\widehat{o} = 3.61$	$\widehat{a} = 1.01$
95 confidence interval	$[13.2; 13.4]$	$[0.119; 0.121]$	$[3.44; 3.73]$	$[0.99; 1.03]$
Estimated bias	–	–	$1.05\,10^{-3}$	$3.80\,10^{-4}$

Table 11.1: Estimates for the 2-nd peak of the 8-peak model.

11.6 Conclusion

The method presented here can give very complete decomposition of complex thermograms. In practice, the number of detected peaks may be high ($K > 10$) but such a modeling has to be physically meaningful. Therefore, the major interest of the method is to give very reliable estimates of various characteristics (onset, area, *etc.*) essentially for a few major peaks. The interest of detecting smaller peaks is that it makes the estimates concerning the major peaks more accurate.

As already said in section **2.2**, a possible extension is to take the baseline into account with a model of the form $Y(t) = \Phi(t, \omega) + b(t, \beta) + E(t)$ where $b(t, \beta)$ is the baseline function. ANATHERM gives starting values of the parameter β for a spline polynomial. This extension can be very useful in case of very flat (and therefore of doubtful physical meaning) peaks and does not entail to many numerical difficulties.

This method can obviously be used with other peak shapes than the Gaussian-Gaussian one but many theoretical peak equations (such as convolution models) lead to intractable expressions and to very heavy computations. The good results obtained here are due to the fact that the model is the simple sum of several elementary peak functions $\varphi(., \theta_k)$.

This method can be applied in many other domains where the problem of peak separation is encountered such as chromatography. In this precise case, the model can be simplified since the symmetry of the peaks is commonly admitted (see Dyson 2) but numerical troubles come from the great number of peaks ($K = 100$ or 500) and require to limit it with a condition such as $h > h_{\min}$ or to use local decompositions on a sequence of intervals.

Support: This work has been supported by the ARILAIT-Recherche research program 'Etude de la cristallisation fractionnée de matière grasse laitiè re anhydre : Composition

et structure en triglycérides'.

Acknowledgment: I would like to acknowledge Pr M. Ollivon and F. Lavigne for all their patient and useful explanations about the physical problem and Pr C. Duby for her helpful advice for the mathematical part. I would also like to thank the anonymous referees for their remarks and their suggestions that helped me a lot to improve my work and presentation.

Bibliography

[1] Box, M.J. (1971). *Bias in nonlinear estimation.* J. Royal Statist. Soc., B, **33**, 171-201.

[2] Dyson, N. (1990). *Chromatographic integration methods.* RSC chromatography monographs.

[3] Gallant, A.R. (1987). *Nonlinear statistical models.* Wiley.

[4] Huet, S., Jolivet, E., Messéan A. (1992). *La ré gression non-linéaire : méthodes et applications en biologie.* INRA.

[5] Robin, S., Bourguignon, G., Lavigne, F., Ollivon, M. (1997). *Analysis of complex DSC recordings: Application to fat-containing foods using ANATHERM software.* submitted to J. Thermal Analysis.

[6] Roeder, K. (1994). *A graphical technique for determining the number of components in a mixture of normals.* J. Amer. Statist. Assoc., **89**, No 426, p487-495.

[7] Ritter, C. (1994). *Statistical analysis of spectra from electron spectroscopy for chemical analysis.* The Statistician, **43** , 111-127.

[8] Seber, G.A.F., Wild, C.J. (1989). *Nonlinear regression.* Wiley.

[9] Self, G.S., Kung-Yee, L. (1987). *Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions.* J. Amer. Statist. Assoc., **82**, 605-610.

Appendix

11.A Bias and variance approximation of $\widehat{o}$ and $\widehat{a}$

The second order expansion of $\widehat{o}$ and $\widehat{a}$ gives

$$\begin{aligned}
\widehat{o} &= \widehat{s} - 2\widehat{\lambda}^{-1} \\
&\approx s - 2\lambda^{-1} + (\widehat{s} - s) + 2\lambda^{-2}(\widehat{\lambda} - \lambda) - 2\lambda^{-3}(\widehat{\lambda} - \lambda)^2, \\
\widehat{a} &= \sqrt{\frac{\pi}{2}}\widehat{h}(\widehat{\lambda}^{-1} + \widehat{\gamma}^{-1}) \\
&\approx \sqrt{\frac{\pi}{2}}[h(\lambda^{-1} + \gamma^{-1}) + (\lambda^{-1} + \gamma^{-1})(\widehat{h} - h) - h\{\lambda^{-2}(\widehat{\lambda} - \lambda) - \gamma^{-2}(\widehat{\gamma} - \gamma)\} \\
&\quad + h\{\lambda^{-3}(\widehat{\lambda} - \lambda)^2 + \gamma^{-3}(\widehat{\gamma} - \gamma)^2\} - (\widehat{h} - h)\{\lambda^{-2}(\widehat{\lambda} - \lambda) - \gamma^{-2}(\widehat{\gamma} - \gamma)\}].
\end{aligned}$$

First order terms give evaluations of the variances (dropping higher order terms):

$$\begin{aligned}
\mathrm{V}(\widehat{o}) &\simeq \mathrm{V}(\widehat{s}) + 4\lambda^{-4}\mathrm{V}(\widehat{\lambda}) + 4\lambda^{-2}\mathrm{Cov}(\widehat{s}, \widehat{\lambda}), \\
\mathrm{V}(\widehat{a}) &\simeq \pi/2\{(\lambda^{-1} + \gamma^{-1})^2\ \mathrm{V}(\widehat{h}) + h^2(\lambda^{-4}\mathrm{V}(\widehat{\lambda}) + \gamma^{-4}\mathrm{V}(\widehat{\gamma})) \\
&\quad - h(\lambda^{-1} + \gamma^{-1})(\lambda^{-2}\mathrm{Cov}(\widehat{h}, \widehat{\lambda}) + \gamma^{-2}\mathrm{Cov}(\widehat{h}, \widehat{\gamma})) + h^2\lambda^{-2}\gamma^{-2}\mathrm{Cov}(\widehat{\lambda}, \widehat{\gamma})\}.
\end{aligned}$$

Second order terms give evaluations of the biases (the mean of the first order terms being null):

$$\begin{aligned}
\mathrm{E}(\widehat{o} - o) &\simeq -2\lambda^{-3}\mathrm{V}(\widehat{\lambda}), \\
\mathrm{E}(\widehat{a} - a) &\simeq \sqrt{\pi/2}\{h\lambda^{-3}\mathrm{V}(\widehat{\lambda}) + h\gamma^{-3}\mathrm{V}(\widehat{\gamma}) - \lambda^{-2}\mathrm{Cov}(\widehat{h}, \widehat{\lambda}) - \gamma^{-2}\mathrm{Cov}(\widehat{h}, \widehat{\gamma})\}.
\end{aligned}$$

The variance and bias of both $\widehat{o}$ and $\widehat{a}$ may be estimated using the estimates of the variances and covariances of $\widehat{\lambda}$, $\widehat{s}$, $\widehat{h}$ and $\widehat{\gamma}$.

Chapitre 12

Predicting distances using a linear model: the case of varietal distinctness

Nuel G., Robin S., Baril C.P. (2001)

Journal of Applied Statistics, **28** (5), 607–621

Abstract. Differences between plant varieties are based on phenotypic observations, which are both space and time consuming. Moreover the phenotypic data result from the combined effects of genotype and environment. On the contrary, molecular data are easier to obtain and give a direct access to the genotype. In order to save experimental trials and to concentrate efforts on the relevant comparisons between varieties, the relationship between phenotypic and genetic distances is studied. It appears that the classical genetic distances based on molecular data are not appropriate for predicting phenotypic distances. In the linear model framework, we define a new pseudo genetic distance which is a prediction of the phenotypic one. The distribution of this distance given the pseudo genetic distance is established. Statistical properties of the predicted distance are derived when the parameters of the model are either given or estimated. We finally apply these results to distinguishing between 144 maize lines. This case study is very satisfactory because the use of anonymous molecular markers (RFLP) leads to saving 29% of the trials with an acceptable error risk. These results need to be confirmed on other varieties and species and would certainly be improved by using genes coding for phenotypic traits.

Keywords. genetic distances, linear model, molecular markers, phenotypic distances, varietal distinctness

Contents

12.1 Introduction

Every year, hundreds of new plant varieties emerge from laboratory and field trials. To supply farms with the latest products of genetic progress in an organized and coherent manner, the entry of a new variety into the official catalogue is a required condition for commercialization. To be registered, a variety must be distinct, uniform and stable (DUS) and show agronomic or technological advantages. GEVES (Groupe d'étude et de contrôle des variétés et des semences) performs technical analyses of new varieties on behalf of the CTPS (Comité Technique Permanent de la Sélection) under the authority of the french Ministry of Agriculture. DUS tests involve observations over one to three growth cycles in two sites. In all cases, morphological and physiological characteristics are examined at several stages of plant growth, and, in some cases, biochemical characteristics as well. All countries that are members of UPOV (Union pour la Protection des Obtentions Végétales) follow the same recommendations and thus assessment methods are harmonized internationally. In this article, we only focus on the distinctness procedure.

At the present time, a variety is accepted as being new if it can be unambiguously distinguished from all other varieties on the basis of at least one characteristic. This differentiation may also be based on a number of small differences. The requirement for a more accurate and objective method to assess the distinctness of two varieties has led the CTPS to encourage feasibility studies on the use of multivariate criteria. Maize is genetically well characterized, and has thus been used as a model system. An investigation of its DNA polymorphism has been undertaken by GEVES in collaboration with maize breeders and INRA (Institut National de la Recherche Agronomique). Every year 350 lines (corresponding to 500 varieties) are tested for DUS and the acceptance rate is about 30%. The distinctness of a new variety is assessed by comparing with standard varieties, called the reference collection, which included 1700 lines (corresponding to 1000 varieties) in 1998 .

Obtaining phenotypic data is both space and time consuming; as a consequence, the number of varieties to be tested every year prevents from numerous repetitions. On the other hand molecular characteristics are comparatively easier to obtain. The aim of this paper is to predict phenotypic distances through molecular data. In the context of varietal distinctness, using these predictions could save many field trials. Efforts could be concentrated on difficult cases using better experimental designs.

Numerous genetic distances and similarity indices already exist; Gower (1985) gives a general review. The simplest measure of genetic distance without any extra information but molecular is the Rogers' distance (Rogers (1972)) calculated as the proportion of markers having different modalities. The relationship between phenotypic (Mahalanobis (1936)) and genetic (Rogers) distances is not linear but may be described as triangular (see figure **1**). This shape, which is observed for all genetic distances, has some biological explanations: (*i*) compensation of elementary traits in complex traits expression, (*ii*) low part of genome involved in phenotypical traits expression, (*iii*) lack of environmental conditions necessary for the expression of all the observed traits, (*iv*) variation of linkage disequilibrium according to the origin of genotypes. As demonstrated in Burstin and A. (1997), this triangular shape would still be observed if one used only molecular markers linked with the phenotypic traits, such as QTL (quantitative traits

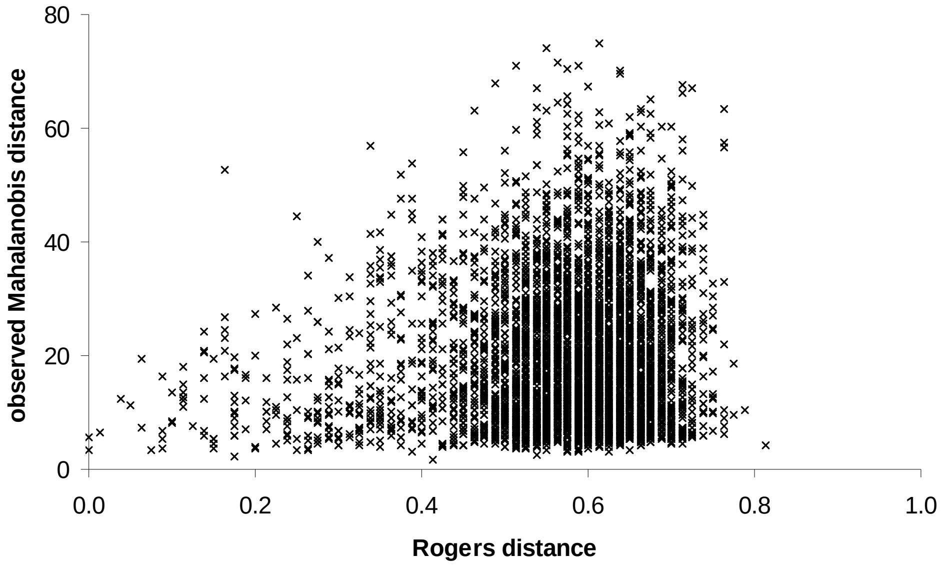


Figure 12.1: Plot of the Mahalanobis distance D versus the Rogers' (genetic) distance.

loci).

Obviously, this kind of relationship prevents one from predicting the phenotypical distance from the molecular one. A new genetic distance has therefore to be defined. We propose to use a linear model for predicting the phenotypic traits from the molecular markers. We are then able to calculate a prediction of the phenotypical distance based only on the molecular data. This predicted distance can therefore be considered as a pseudo genetic distance. Probabilistic and statistical properties of this distance are derived; in particular, its distribution is established. In section **4**, we use this model to distinguish between maize lines.

12.2 Linear model and Distance prediction

Let us consider n individuals each described by p quantitative morphological variables Y^j $(j = 1 \dots p)$ and m molecular markers $(k = 1 \dots m)$ each having s_k modalities corresponding to the alleles observed in the collection. Let $\mathcal{M}$ denote the set of the observed markers $(|\mathcal{M}| = m)$.

12.2.1 The linear model

Model for one variable. Each morphological variable is governed by a set $\mathcal{A}_j$ of markers. We assume here that the relation between Y^j and the markers of $\mathcal{A}_j$ can be modeled by a classical linear model $\mathbf{Y}^j = \mathbf{X}_{\mathcal{A}_j}\theta_{\mathcal{A}_j} + \mathbf{F}^j$ where $\mathbf{X}_{\mathcal{A}_j}$ is the matrix including the genotypes (coded in terms of presence/absence of each modality of each marker) of all individuals for the markers of $\mathcal{A}_j$ and $\theta_{\mathcal{A}_j}$ is the vector of the effects of the different modalities.

Unfortunately we do not observe all the markers of the relevant set $\mathcal{A}_j$ but only the subset $\mathcal{B}_j = \mathcal{M} \cap \mathcal{A}_j$. So we have to consider the restricted model

$$\underset{(n,1)}{\mathbf{Y}^j} = \underset{(n,1+M_j)}{\mathbf{X}_{\mathcal{B}_j}} \underset{(1+M_j,1)}{\theta_{\mathcal{B}_j}} + \underset{(n,1)}{\mathbf{E}^j} \tag{12.1}$$

where $\mathbf{X}_{\mathcal{B}_j}$ contains the genotypes for the markers of $\mathcal{B}_j$ ($M_j = \sum_{k \in \mathcal{B}_j} s_k$, the first element of $\theta_{\mathcal{B}_j}$ is the intercept). The residuals $\{E_a^j\}_{a=1\ldots n}$ are assumed to be independent and normally distributed with the same mean 0 and variance σ_j^2. In this model, the residual term $\mathbf{E}^j$ includes the residual term $\mathbf{F}^j$ of the complete model and the effects of the unobserved relevant markers.

This model appears to give a good representation of the relation between genetic and phenotypic distances. In particular, Burstin and A. (1997) have shown, using simulations, that it reflects the frequently observed triangular relationship (see figure **1**).

¿From now on, to make the notation more convenient, we shall use an alternative but equivalent form of model (12.1):

$$\mathbf{Y}^j = \underset{(n,1+M)}{\mathbf{X}} \underset{(1+M,1)}{\theta^j} + \mathbf{E}^j \tag{12.2}$$

where $\mathbf{X}$ contains the genotypes of all individuals and for all the markers of $\mathcal{M}$ ($M = \sum_{k \in \mathcal{M}} s_k$) and where the elements of θ^j corresponding to the markers of $\mathcal{M} - \mathcal{B}_j$ are set to zero.

In this model, the expected value Y_a^j of j-th variable for the individual a is

$$\mu_a^j = \mathbb{E}\left(Y_a^j\right) = \mathbf{x}_a \theta^j$$

where $\mathbf{x}_a$ denotes the a-th row of the matrix $\mathbf{X}$.

Model for p variables. To consider p variables, we just have to gather the p models (12.2) into a single model for $\mathbf{Y} = [\mathbf{Y}^1 \ldots \mathbf{Y}^p]$:

$$\underset{(n,p)}{\mathbf{Y}} = \underset{(n,M+1)}{\mathbf{X}} \underset{(M+1,p)}{\mathbf{\Theta}} + \underset{(n,p)}{\mathbf{E}} \tag{12.3}$$

where $\mathbf{\Theta} = [\theta^1 \ldots \theta^p]$ and $\mathbf{E} = [\mathbf{E}^1 \ldots \mathbf{E}^p]$.

To take into account the possible dependence of the phenotypic variables, we assume that their common covariance matrix $\mathbf{\Sigma}$ is not necessarily diagonal. The vectors $\mathbf{E}_a = [E_a^1 \ldots E_a^p]$ and $\mathbf{Y}_a = [Y_a^1 \ldots Y_a^p]$ are normally distributed:

$$\mathbf{E}_a \sim \mathcal{N}\left(\mathbf{0}_p; \mathbf{\Sigma}\right), \qquad \mathbf{Y}_a \sim \mathcal{N}\left(\mu_a; \mathbf{\Sigma}\right)$$

where $\mu_a = [\mu_a^1 \ldots \mu_a^p] = \mathbf{x}_a \mathbf{\Theta}$.

In this model, the individuals are assumed to be independent but the variables measured on a same individual are correlated.

12.2.2 Distance Prediction

The phenotypic distance between two individuals is defined as

$$D^2_{\mathbf{M}}(a,b) = \|\mathbf{Y}_a - \mathbf{Y}_b\|^2_{\mathbf{M}} = (\mathbf{Y}_a - \mathbf{Y}_b)\,\mathbf{M}\,(\mathbf{Y}_a - \mathbf{Y}_b)'$$

where $\mathbf{M}$ is a given metric.

The case of Mahalanobis distance. A particularly interesting distance is the Mahalanobis distance D^2_{Maha} for which $\mathbf{M}$ is the inverse of the covariance matrix of $(\mathbf{Y}_a - \mathbf{Y}_b)$, i.e. $\mathbf{M} = [\mathbb{V}(\mathbf{Y}_a - \mathbf{Y}_b)]^{-1} = \frac{1}{2}\mathbf{\Sigma}^{-1}$ so

$$D^2_{\text{Maha}}(a,b) = \frac{1}{2}(\mathbf{Y}_a - \mathbf{Y}_b)\,\mathbf{\Sigma}^{-1}(\mathbf{Y}_a - \mathbf{Y}_b)'.$$

We can define the expected distance $\delta_{\text{Maha}}(a,b)$ as the distance between the expected values of the $\mathbf{Y}_a$'s:

$$\delta^2_{\text{Maha}}(a,b) = \|\mu_a - \mu_b\|^2_{\frac{1}{2}\mathbf{\Sigma}^{-1}},$$

which can be expressed as

$$\begin{aligned}\delta^2_{\text{Maha}}(a,b) &= \frac{1}{2}(\mathbf{x}_a\mathbf{\Theta} - \mathbf{x}_b\mathbf{\Theta})\,\mathbf{\Sigma}^{-1}(\mathbf{x}_a\mathbf{\Theta} - \mathbf{x}_b\mathbf{\Theta})' \\ &= \frac{1}{2}(\mathbf{x}_a - \mathbf{x}_b)\,\mathbf{\Theta}\mathbf{\Sigma}^{-1}\mathbf{\Theta}'(\mathbf{x}_a - \mathbf{x}_b)' = \|\mathbf{x}_a - \mathbf{x}_b\|^2_{\frac{1}{2}\mathbf{\Theta}\mathbf{\Sigma}^{-1}\mathbf{\Theta}'}.\end{aligned}$$

This distance can be considered as a pseudo genetic distance, since it is only calculated through the genotypes $\mathbf{x}_a$ and $\mathbf{x}_b$, with the metric $\frac{1}{2}\mathbf{\Theta}\mathbf{\Sigma}^{-1}\mathbf{\Theta}'$.

Given the genotypes, $D^2_{Maha}(a,b)$ is random while $\delta^2_{Maha}(a,b)$ is not. It is straightforward that this phenotypic square distance has a non-central chi-square distribution with a non-centrality parameter equal to the square predicted distance:

$$D^2_{\text{Maha}}(a,b) \sim \chi^2_p\left[\delta^2_{\text{Maha}}(a,b)\right]. \tag{12.4}$$

This is clear since all the vectors $\mathbf{E}_a$ are independent with mean zero and common covariance matrix $\mathbf{\Sigma}$, so $\mathbb{V}(\mathbf{E}_a - \mathbf{E}_b) = 2\mathbf{\Sigma}$. Hence, for any given level $1-\gamma$ we can give a two-sided prediction interval for $D^2_{\text{Maha}}(a,b)$:

$$\left[\chi^2_{p;\gamma/2}\left\{\delta^2_{\text{Maha}}(a,b)\right\}\ ;\ \chi^2_{p;1-\gamma/2}\left\{\delta^2_{\text{Maha}}(a,b)\right\}\right] \tag{12.5}$$

(where $\chi^2_{p;\beta}(\delta^2)$ is the β-quantile of the non central chi-square distribution) and one-sided interval:

$$\left[\chi^2_{p;\gamma}\left\{\delta^2_{\text{Maha}}(a,b)\right\}\ ;\ \infty\right[.$$

General case. The Mahalanobis distance is not the only distance that can be used, although it is the most convenient. For other distances the following proposition may be established.

Proposition 1 *For a general metric* $\mathbf{M}$*, the distribution of the square distance* D^2 *is a linear combination of non-central chi-squares: using the two diagonal decompositions* $\mathbf{M} = \mathbf{PDP}'$ *and* $\mathbf{D}^{\frac{1}{2}}\mathbf{P}'\mathbf{\Sigma}\mathbf{P}\mathbf{D}^{\frac{1}{2}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$ *where* $\mathbf{P}$ *and* $\mathbf{Q}$ *are orthogonal,* $\mathbf{D}$ *and* $\mathbf{\Lambda}$ *diagonal with* $\mathbf{\Lambda} = \mathbf{diag}\,[\lambda_1, \ldots, \lambda_p]$*, one has*

$$D^2_{\mathbf{M}}(a, b) \stackrel{\mathcal{D}}{=} 2\sum_{j=1}^{p} \lambda_j \xi_j$$

where the ξ_j *are independent,* $\xi_j \sim \chi^2_1\left[(\nu^j)^2/(2\lambda_j)\right]$ *and*

$$\nu = \begin{bmatrix}\nu^1 & \ldots & \nu^p\end{bmatrix} = (\mu_a - \mu_b)\ \mathbf{P}\mathbf{D}^{\frac{1}{2}}\mathbf{Q}.$$

The proof of this proposition and of the following are given in the appendix. According to proposition **1**, a predicted distance can be defined as the linear combination of the non centrality parameters $(\nu^j)^2$:

$$\delta^2_{\mathbf{M}}(a, b) = 2\sum_{j=1}^{p} \lambda_j \frac{(\nu^j)^2}{2\lambda_j} = \nu\nu' = (\mu_a - \mu_b)\mathbf{M}(\mu_a - \mu_b)'.$$

Prediction intervals can thus be constructed even for a general metric $\mathbf{M}$.

12.2.3 The case of an unknown parameter

That the parameter Θ is estimated is not taken into account in the previous results and this may induce some bias and uncertainty in the distance prediction. Furthermore, when managing reference collections, this estimation is made on the elements of the collections, but the results may be applied to other individuals. In this section we give some results about the bias and distribution of the predicted distance. The originality of these results lays in the two following points: i) the difference between the training and validation subsets is taken into account, especially in the expression of the predicted vectors $\widehat{\mathbf{Y}}^j$; ii) the linear model is different for each phenotypic variable, which makes a major difference with the classical multivariate linear model.

Parameter estimates. In the following, for each column vector $(n \times 1)$, we shall consider that the first n_1 coordinates correspond to the training set and the remaining n_0 to the validation set $(n_0 + n_1 = n)$. Decomposing $\mathbb{R}^n$ as $\mathbb{R}^n = \mathbb{R}^{n_1} \oplus \mathbb{R}^{n_0}$, we denote by $\mathbf{\Pi}_1$ the $(n \times n)$ projector matrix that sets the last n_0 coordinates to zero and by $\mathbf{\Pi}_0$ the $(n \times n)$ projector setting the first n_1 coordinates to zero.

Since the rank of $\mathbf{\Pi}_1\mathbf{X}_{\mathcal{B}_j}$ is less than M_j+1, we use side conditions to define the generalized inverse $(\mathbf{X}'_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{X}_{\mathcal{B}_j})^-$ (see Rao and Mitra (1971)). $\widehat{\theta}_{\mathcal{B}_j} = (\mathbf{X}'_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{X}_{\mathcal{B}_j})^-\mathbf{X}'_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{Y}^j$ is an unbiased estimate of the particular value of $\theta_{\mathcal{B}_j}$ satisfying this conditions, i.e. such as $\theta_{\mathcal{B}_j} = (\mathbf{X}'_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{X}_{\mathcal{B}_j})^-\mathbf{X}'_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{X}_{\mathcal{B}_j}\theta_{\mathcal{B}_j}$ (see Scheffé (1959)).

The predicted vector $\widehat{\mathbf{Y}}^j$ is

$$\widehat{\mathbf{Y}}^j = \mathbf{X}_{\mathcal{B}_j}\widehat{\theta}_{\mathcal{B}_j} = \mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{Y}^j \tag{12.6}$$

where $\mathbf{A}_{\mathcal{B}_j} = \mathbf{X}_{\mathcal{B}_j}(\mathbf{X}'_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{X}_{\mathcal{B}_j})^{-}\mathbf{X}'_{\mathcal{B}_j}$. $\widehat{\mathbf{Y}}^j$ is an unbiased estimate of μ^j: $\mathbb{E}(\mathbf{Y}^j) = \mu^j$.

Decomposing $\mathbf{I}_n = \mathbf{\Pi}_1 + \Pi_0$, we get $\widehat{\mathbf{Y}}^j = \mathbf{\Pi}_1\mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{Y}^j + \mathbf{\Pi}_0\mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{Y}^j$. The first term concerns the first n_1 coordinates: $\mathbf{\Pi}_1\mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1$ is the orthogonal projector on the intersection of $\mathcal{B}_j$ and the column space of $\mathbf{\Pi}_1$ and this projection does not depend on the side conditions used to calculate the generalized inverse (see Scheffé (1959)). The second element concerns the last n_0 coordinates but is not exactly equivalent since $\Pi_0\mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1$ is not a projector: it gives the predicted values for the validation set, these values depend on the side conditions.

The predicted matrix $\widehat{\mathbf{Y}}$ is

$$\widehat{\mathbf{Y}} = \left[\ \mathbf{A}_{\mathcal{B}_1}\mathbf{\Pi}_1\mathbf{Y}^1 \quad \ldots \quad \mathbf{A}_{\mathcal{B}_p}\mathbf{\Pi}_1\mathbf{Y}^p\ \right].$$

General case. When $\mathbf{\Theta}$ is unknown, for any pair (a,b), we are able only to calculate an estimate of δ^2 :

$$\widehat{\delta}^2(a,b) = \left\|\widehat{\mathbf{Y}}_a - \widehat{\mathbf{Y}}_b\right\|^2_{\mathbf{M}} = \left(\widehat{\mathbf{Y}}_a - \widehat{\mathbf{Y}}_b\right)\mathbf{M}\left(\widehat{\mathbf{Y}}_a - \widehat{\mathbf{Y}}_b\right)'.$$

Defining the vector $\mathbf{u}_a = [0 \cdots 0\,1\,0 \cdots 0]$ having n coordinates with a one in the a-th column and zero elsewhere, the difference $\widehat{\mathbf{Y}}_a - \widehat{\mathbf{Y}}_b$ is equal to the contrast $(\mathbf{u}_a - \mathbf{u}_b)\,\widehat{\mathbf{Y}}$, so

$$\widehat{\delta}^2(a,b) = (\mathbf{u}_a - \mathbf{u}_b)\,\widehat{\mathbf{Y}}\mathbf{M}\widehat{\mathbf{Y}}'\,(\mathbf{u}_a - \mathbf{u}_b)'.$$

Proposition 2 *The bias of $\widehat{\delta}^2(a,b)$ is positive and equal to*

$$(\mathbf{u}_a - \mathbf{u}_b)\,\mathbf{H}\,(\mathbf{u}_a - \mathbf{u}_b)' = h^a_a + h^b_b - 2h^b_a$$

where the general term h^b_a of the matrix $\mathbf{H}$ is

$$h^b_a = \sum_{j=1}^{p}\sum_{j'=1}^{p}\sigma^{j'}_j m^{j'}_j\left[\mathbf{u}_a\mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{A}_{\mathcal{B}_{j'}}(\mathbf{u}_b)'\right],$$

where σ^j_a and m^j_a are the general terms of $\mathbf{\Sigma}$ and $\mathbf{M}$.

Using Mahalanobis distance with one single model. We shall concentrate now on the case of the Mahalanobis distance ($\mathbf{M} = \frac{1}{2}\mathbf{\Sigma}^{-1}$), in the case where $\mathbf{\Sigma}$ is known. Assuming additionally that the same model is valid for all the phenotypic variables, we give a simpler expression of the bias in corollary **3**.

Corollary 3 *In the particular case of the Mahalanobis distance and if we use the same model for all the $(Y^j)_{j=1\ldots p}$ (i.e. $\mathbf{X}_{\mathcal{B}_j} \equiv \mathbf{X}$), the bias is*

$$\frac{p}{2}(\mathbf{u}_a - \mathbf{u}_b)\,\mathbf{A}\mathbf{\Pi}_1\mathbf{A}\,(\mathbf{u}_a - \mathbf{u}_b)'.$$

This corollary can be simplified if the generalized inverse $(\mathbf{X}'\mathbf{\Pi}_1\mathbf{X})^{-}$ is reflexive (see Rao and Mitra (1971)): in this case $\mathbf{A}\mathbf{\Pi}_1\mathbf{A} = \mathbf{A}$.

Although in practice, both $\mathbf{\Theta}$ and $\mathbf{\Sigma}$ are estimated, the fact that $\mathbf{\Sigma}$ is unknown is not taken into account here. At the present time, we propose to 'plug' in $\widehat{\mathbf{\Sigma}}$ in place of $\mathbf{\Sigma}$ in all the formulae given here.

12.3 Distinctness procedure

12.3.1 Decision rule

The distinctness between two lines is generally determined according to a threshold t on the phenotypic distance:

$$\{\text{lines } a \text{ and } b \text{ are distinct}\} \Leftrightarrow \left\{D^2(a,b) > t\right\}.$$

Formula (12.5) gives a prediction interval for d^2 with level $1-\gamma$. This result can be applied to predict whether lines a and b are distinct or not according to the following decision rule:

- if $\chi^2_{p;\gamma/2}\left[\delta^2(a,b)\right] > t$, the whole prediction interval on D^2 is above the threshold t: lines a and b are predicted to be distinct considering their genotypes;
- if $\chi^2_{p;1-\gamma/2}\left[\delta^2(a,b)\right] < t$, the whole prediction interval on D^2 is beyond the threshold t: lines a and b are predicted to be similar considering their genotypes;
- if $\chi^2_{p;\gamma/2}\left[\delta^2(a,b)\right] < t < \chi^2_{p;1-\gamma/2}\left[\delta^2(a,b)\right]$, the threshold t is included in the prediction interval: no decision can be made on the basis of the simple genotypes so an experiment has to be made to measure the phenotypic traits.

This rule is summarized in figure **2**. The molecular thresholds b_{inf} and b_{sup} are given by

$$t = \chi^2_{p;1-\gamma/2}(b_{\text{inf}}) = \chi^2_{p;\gamma/2}(b_{\text{sup}}).$$

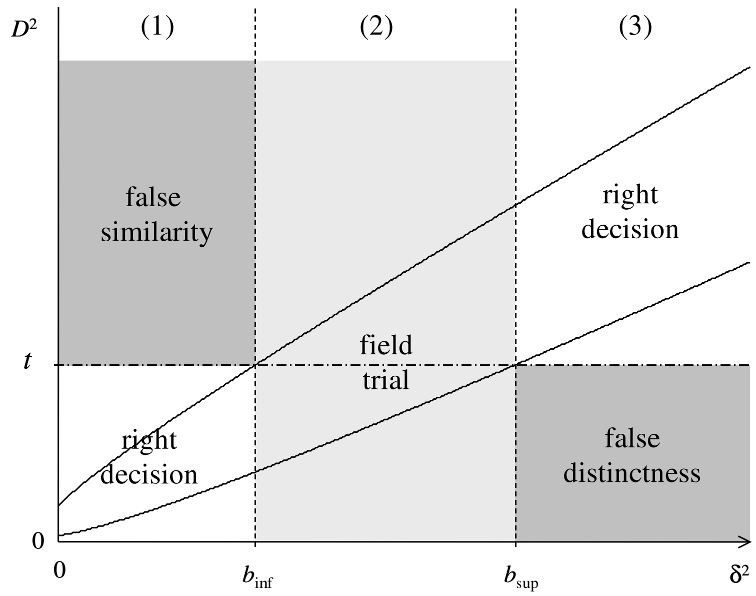


Figure 12.2: Decision rule (solid lines = $\chi^2_{p;\gamma/2}(\delta^2)$ and $\chi^2_{p;1-\gamma/2}(\delta^2)$)

Errors. The decisions made according to molecular data can obviously lead to some errors. Let us consider the cases corresponding to the three vertical zones of figure **2**.

(1) $\delta^2(a,b) < b_{\text{inf}}$ (a and b predicted similar):
if $D^2\,(a,b) < t$ then a and b are effectively similar, the decision is correct;
if $D^2\,(a,b) > t$ then a and b are actually distinct, the decision is wrong.

(2) $b_{\text{inf}} < \delta^2(a,b) < b_{\text{sup}}$: no error can be made since there is no decision; the experiment will show whether a and b are distinct or not.

(3) $\delta^2(a,b) > b_{\text{sup}}$ is the symmetrical case of **(1)**: if $D^2(a,b) > t$ then the decision is correct, else we are misled.

12.3.2 Proposition of new distinctness procedure

Distinctness. The aim of a distinctness procedure is to decide whether candidates lines (set $\mathcal{C}$) are distinct from those of a reference collection (set $\mathcal{R}$). A candidate line c of $\mathcal{C}$ is distinct from the collection $\mathcal{R}$ if it is distinct from all the elements r of $\mathcal{R}$:

$$\{c \text{ distinct from } \mathcal{R}\} \Leftrightarrow \left\{\forall r \in \mathcal{R}:\; D^2(r,c) > t\right\}. \tag{12.7}$$

On the contrary c is said to be similar to $\mathcal{R}$ as soon as it resembles one of the elements of $\mathcal{R}$:

$$\{c \text{ similar to } \mathcal{R}\} \Leftrightarrow \left\{\exists r \in \mathcal{R}:\; D^2(r,c) < t\right\}.$$

Savings. As plant breeders need a phenotypic evaluation of their lines when proposing a new variety for registration, all candidate lines have to be set on trial. Our wish is to avoid a part of the experiments for the reference lines by using the predicted distance (through molecular data) rather than the observed one.

For a reference line r, we can avoid experimentation if a decision can be made for all the molecular comparisons between r and all the candidate lines. Let us denote by $\mathcal{S}$ (for savings) the set of the reference lines which do not need to be set on trial ($\mathcal{S} \subset \mathcal{R}$):

$$\mathcal{S} = \left\{r \in \mathcal{R}:\; \forall c \in \mathcal{C},\, \delta^2\,(r,c) \notin \,]b_{\text{inf}}, b_{\text{sup}}[\right\}. \tag{12.8}$$

False distinctness. A candidate line c can be predicted to be distinct from $\mathcal{R}$ while it is not. This happens when all the following conditions are fulfilled:

D1 c is similar to $\mathcal{R}$: $\exists r \in \mathcal{R},\; D^2(r,c) < t$;

D2 c is not predicted to be similar considering molecular data: $\forall r \in \mathcal{R},\; \delta^2(r,c) > b_{\text{inf}}$;

D3 c is distinct from all the experimented reference lines: $\forall r \in \mathcal{R} - \mathcal{S},\; D^2(r,c) > t$;

D4 none of the errors is corrected by any experiment: $\forall r \in \mathcal{R},\; \delta^2(r,c) > b_{\text{sup}}$ and $D^2(r,c) < t:\; r \in \mathcal{S}$.

The point (**D4**) has some major (and favourable) consequences. It recalls that a candidate line is compared not only to the reference lines from which it is close according to the molecular data, but also to all the reference lines that shall be trialed. If the candidate line c is near the reference line r but these two lines are not predicted as similar, the final error may be avoided if r is predicted to be similar to another candidate c', and therefore, trialed. In this case, multiple candidate lines can reduce savings but avoid some false distinctnesses.

False similarity. A candidate line c can be predicted to be similar to $\mathcal{R}$ while it is actually distinct. This happens when both the following conditions are fulfilled:

S1 c is distinct from $\mathcal{R} : \forall r \in \mathcal{R},\ D^2(r,c) > t$;

S2 at least one decision concerning c is wrong and the concerned reference line is put aside: $\exists r \in \mathcal{R} :\ \delta^2(r,c) < b_{\text{inf}},\ r \in \mathcal{S}$.

In the context of DUS testing, false similarities will never arise because no similarity decision can be made on the basis of single molecular data. One reason is that such an error is much more likely to occur than false distinctness; another reason is that the reference collection has to contain the widest possible biodiversity so it seems preferable to accept some irregular elements (that is to make a conservative choice) rather than to reject original lines.

This paper will therefore focus on the upper bound b_{sup}. In the following, we shall assume that b_{inf} is equal to zero. This is equivalent to applying the decision rule of section 12.3.1 with an unilateral prediction interval $\left[\chi^2_{p;\gamma}(\delta^2)\,;\infty\right]$.

12.4 Application to varietal distinctness

12.4.1 Material

These results have been applied to $n = 144$ maize inbred lines, representative of material released in France, described using both Restriction Fragment Length Polymorphism (RFLP) markers and $p = 10$ discriminant morphological quantitative traits (Bar-Hen and Charcosset (1995)). This genetical material, derived from the maize germplasms corresponding to the different groups cultivated in Europe, is coded to ensure the confidentiality of the results. 100 probes were chosen according to the quality of the hybridization signal and the genome coverage. DNA extracted from 15 young plants was digested with three restriction enzymes, giving rise to 3×100 Enzyme $\times$ Probe Combinations (EPC). Among these 300 EPCs, only 222 provided polymorphic and interpretable profiles corresponding to 95 probes, giving 1098 bands. In order to avoid redundancy, only $m = 80$ monolocus EPCs corresponding to different loci, were selected (see Dillmann *et al.* (1997) for more details).

The ten morphological quantitative traits were observed over four years (from 1989 to 1992) in three sites, with two replications per site following a block design. The experimental design is unbalanced because sites were chosen for each genotype according to its earliness. Most of the inbred lines were evaluated in at least two locations for at least two years, but some of them were evaluated for only one year in two locations. An estimate of each phenotypic effect was given by the residuals of the analysis of variance performed on each trait with main effects: year, location, block and all two-way interaction terms: year×location, year×block. The quantitative traits concerned ear (length, diameter of ear, diameter of cob), number of rows of seeds, plant (height of ear, total plant length), leaf (width of blades) and tassel (length of main axis above lowest side branch, length of main axis above highest side branch, date of male flowering). See Bar-Hen *et al.* (1995) for more details.

For each phenotypic variable, the set $\mathcal{B}_j$ of the relevant markers has been selected using a stepwise procedure using a subset of $n_1 = 100$ training lines of maize (the $n_0 = 44$ remaining lines were used to validate the model). This procedure uses a Fisher test with level α to enter or remove markers from the model. In the following, we shall discuss the choice of the selection level α.

Throughout this application we use Mahalanobis distance in order to take into account the correlations between the phenotypic traits. Furthermore, the estimate $\widehat{\mathbf{\Theta}}$ is calculated using the Moore-Penrose generalized inverse (see Rao and Mitra (1971)) performed with the MatLab software (The Mathworks inc., Version 5.2): this choice has no influence on the predicted value for the 100 training lines but, as said in section **12.2.3**, changes the predictions for the validation set.

12.4.2 Distance prediction

Using the usual $\alpha = 5\%$ level for selecting the relevant markers for each phenotypic variable, we obtain very well-fitted predictions of the phenotypic distances, as shown in figure **3**. It is clear that the triangular shape disappears: the distribution of the observed distance D is concentrated in the prediction interval. Figure **1** plots the Mahalanobis distance versus the Rogers' (molecular) distance for the $n_1(n_1-1)/2 = 4950$ pairs of training lines. The comparison of figures **1** and **3** shows the improvement of the prediction using δ rather than the Rogers' distance: we obtain an almost linear relation instead of a triangular one. Note that all the points have the same y-value on both graphs; only the x-values differ.

Robustness. In practice, this method should be used to compare lines of a reference collection with new lines. In this framework, the reference collection would be the training set and the new lines could be considered as a validation set since they would not be used to estimate $\mathbf{\Theta}$. Hence it is necessary to study the quality of the predictions when considering pairs involving both training and validation lines.

Unfortunately, as shown in table **1**, the quality of the predictions with a selection of the markers at level $\alpha = 5\%$ is very disappointing for such pairs: the observed confidence is 11.9% for a 95% prediction interval. As discussed in Nuel (1998), this selection level is not severe enough and seems to select poorly relevant markers and induces an over-fitting of the model; these selected markers strongly depend on the training set and give very inaccurate predictions for the validation set.

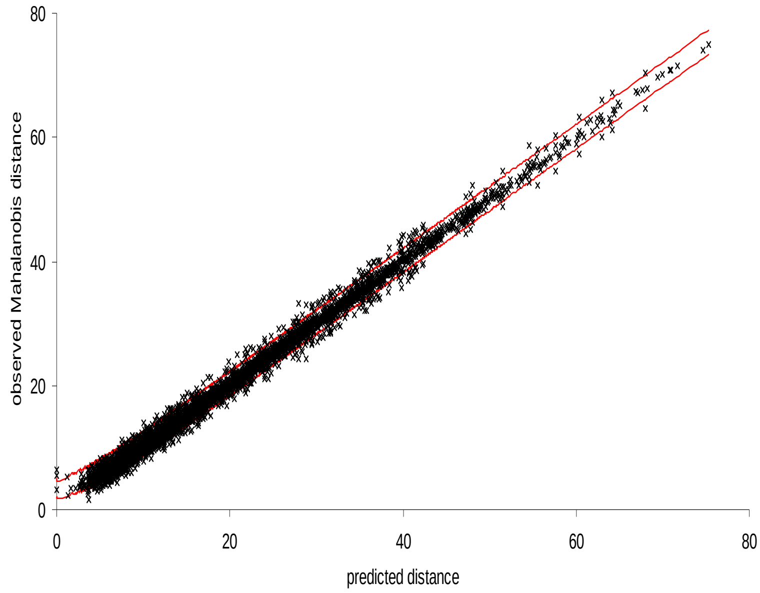


Figure 12.3: Plot of the Mahalanobis distance D versus the (uncorrected) predicted distance with a selection level $\alpha = 5\%$, for couples of training lines. The solid lines give the 95% prediction interval.

selection	predicted distance		corrected predicted distance	
level	training couples	mixed couples	training couples	mixed couples
5%	93.0%	11.9%	93.2%	12.3%
1%	93.9%	76.0%	94.6%	77.7%

Table 12.1: Observed confidence for two selection levels. The expected confidence is 95%. For each level and distance, the first column concerns couples of training lines and the second one concerns mixed couples (one training line and one validation line).

However, a great effort is made by many genetic research teams to detect QTL for traits having an agronomical interest; some substantial progresses can also be expected using expressed sequences tags (EST). Therefore, we can soon expect to get more relevant markers than those used in this example. To try to see what kind of performance could be obtained with such markers, we apply a more severe selection level ($\alpha = 1\%$). In this case, the observed confidence (set table **1**) raises to 76.0% for such pairs (and to 93.9% for pairs of training lines). This results allow us to be optimistic about the performances of the method when used with more accurate predictors such as QTL.

Correction of the bias. As shown in section **12.2.3**, the fact that $\mathbf{\Theta}$ has to be estimated after the data, induces that $\widehat{\delta}^2$ is biased. Figure **4** presents the plots of the observed distance versus the uncorrected and the corrected predicted distance. The bias is calculated according to proposition **2**; when the correction happens to give a negative distance, it is set arbitrarily to zero. This correction shifts the set of points slightly to

the ses of
0.7% both
valic

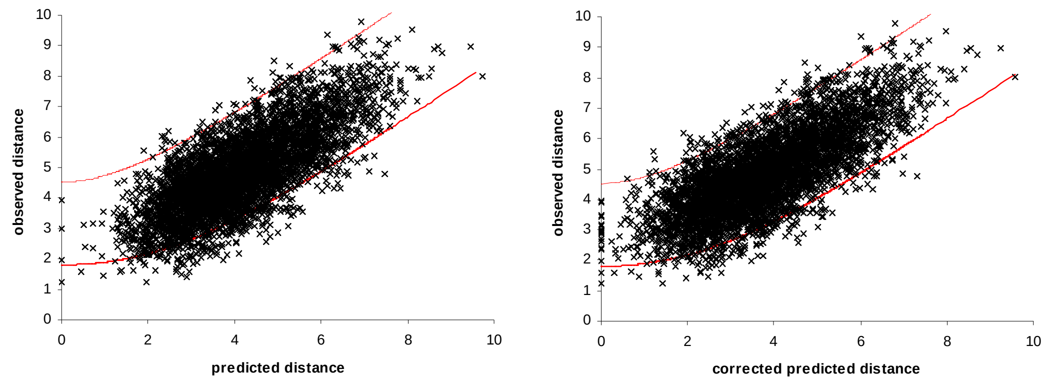


Figure 12.4: Plot of the Mahalanobis distance D versus the uncorrected predicted distance (left hand side) and corrected predicted distance (right hand side) with a selection level $\alpha = 1\%$, for couples of training lines. The solid lines give the 95% prediction interval.

It has to be pointed out that the calculation of the bias is not so easy with ordinary statistical softwares. Although the computation of the predicted distance involves no particular difficulties, the bias requires some special programming. In our case, the correction of the bias does not induce a great improvement so this effort may seem to be rather futile. Nevertheless it is not clear that this is a general result.

12.4.3 DUS

To analyze the performances of the distinctness procedure we are proposing in section **12.3.2**, we simulate several DUS sessions with a varying candidate set $\mathcal{C}$ and a constant reference collection $\mathcal{R}$. As said in section **12.1**, at the present time, the number of candidate lines represent 20% of the size of the reference collection. In order to respect this proportion, we sample 20 lines among the 44, for each simulation (the total number of possible subsets would be $\mathrm{C}_{44}^{20} \simeq 2 \times 10^{12}$).

Phenotypic threshold t. To determine the phenotypic threshold above which two lines are deemed to be phenotypically distinct, we sample 10 000 sets $\mathcal{C}$; for each sample, we calculate t so that the acceptance rate (according to formula (12.7)) is 70%, as indicated in section **12.1**. The final threshold is the mean of these 10 000 values.

False distinctness and savings. The choice of the confidence level γ is a crucial step in the procedure: for a given phenotypic threshold t, the molecular threshold b_{sup}, defined by $t = \chi^2_{p;\gamma}(b_{\text{sup}})$, is a decreasing function of γ. Thus, a small value of γ will provide a high value of b_{sup} and, therefore, a large field trial zone. In this case, we expect very few errors, and savings. Contrariwise, a high confidence level γ will lead to numerous savings and errors.

To optimize the procedure, we consider all the values of γ from 0.5% to 10% by 0.5%. For computational time reasons, we only use 1000 samples to estimate the number of false distinctnesses and savings for each value of γ. These quantities are estimated and not directly calculated because the combinatorial problem seems really intricate.

Results. Figure **5** shows the evolution of the number of false distinctnesses and of the percentage of savings as the confidence level γ grows. We observe that both of these quantities increase step by step; this is due to the limited size of the reference collection ($n_1 = 100$). Since the lower bound b_{inf} is set to zero, the definition 12.8 can be rewritten as

$$\mathcal{S} = \left\{ r \in \mathcal{R} : \inf_{c \in \mathcal{C}} \left[\delta^2 \left(r, c \right) \right] > b_{\text{sup}} \right\}.$$

The set $\mathcal{S}$ changes only when the bound b_{sup} becomes lower than one of 100 different values of $\inf_{c\in\mathcal{C}} \left[\delta^2 \left(r, c\right)\right]$. The same kind of mechanism governs the number of false distinctnesses.

To choose the level γ, we set a maximum acceptable number false distinctnesses. If we are ready only to accept one error among the 20 candidate lines, we see on the left side of figure **5** that the corresponding level is $\gamma = 1.5\%$ which provides 29% of savings (right side of figure **5**). If we accept a slighlty larger error, say 1.2, we get $\gamma = 2.5\%$ and 40%

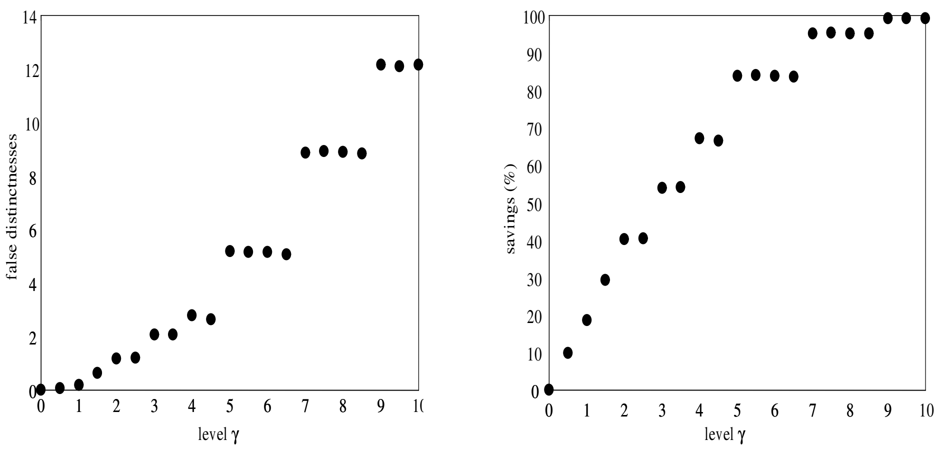


Figure 12.5: Evolution of the number of false distinctnesses (left) and of the percentage of savings (right) as functions of the confidence level γ (%).

12.5 Discussion and Conclusion

One could note that the choice of a linear model, the residuals of which provide an estimation of phenotypic values, is fundamental. It gives the weighting matrix used for the calculation of phenotypic distances. When the interaction terms are significant, it could be further adapted to use as many traits as there are phenotypic traits $\times$ environments combinations. This would increase the number of variables but would probably positively affect the quality of the phenotypic distance estimation.

The distinctness of varieties exclusively based on phenotypical data being both area and time consuming, necessarily leads to compromises between the number of observed varieties and the quality of experimental design, because of the difficulty of mastering all the environmental variation factors. The present model provides a valuable tool for concentrating the field work on the most difficult evaluations. Conservatively, one could avoid 29% of the trials making at most one false distinctness. Furthermore, the use of pseudo genetic distance can be of great help in optimising the experimental design: pairs of lines having a small predicted distance should be trialed in neighbouring plots.

The perspective for this method are as follows:

- Substantial improvements can be expected soon, thanks to important progress made in the localization of genes coding for agronomical traits, providing more relevant markers.

- This approach has to be validated on more data: currently it is being applied to 83 rapeseed varieties described with 324 amplified fragment length polymorphism (AFLP) markers and 10 phenotypic traits (see Lombard *et al.* (2000)).

- Finally, considering the importance of discrete variables noted in the trials, the generalization of this approach to qualitative traits would be of great interest.

Bibliography

BAR-HEN, A. and CHARCOSSET, A. (1995). Relationship between molecular and morphological distances in a maize inbred lines collection. In *Biometrics in Plant Breedings: Application of Molecular Markers*, (J. Van Ooijen and J. Jansen, ed.), 57–66, Proc. of the 9th Meet. Eucarpia, Sect. Biometrics Plant Breed.

BAR-HEN, A., CHARCOSSET, A., BOURGOIN, M. and GUIARD, J. (1995). Relationship between genetic markers and morphological traits in a maize lines collection. *Euphytica.* **84** 145–154.

BURSTIN, J. and A., C. (1997). Relationship phenotypic and markers distances: theoretical and experimental investigations. *Heredity.* **79 (5)** 477–483.

DILLMANN, C., BAR-HEN, A., GUERIN, D., CHARCOSSET, A. and MURIGNEUX, A. (1997). Comparison of rflp and morhological distances between maize *zea mays l.* inbred lines. consequences for germplasm protection purposes. *Theoretical and Applied Genetics.* **96 (5)** 92–102.

GOWER, J. (1985). *Encyclopedia of statistical sciences.* (S. Kotz and N. Johnson, ed.), volume 5, chapter Measures of similarity and distance, 397–405. Wiley: New-York.

LOMBARD, V., BARIL, C. P., DUBREUIL, P. and ZHANG, D. (2000). Genetic relationships and fingerprinting of rapeseed cultivars using AFLP: consequences for varietal registration. *Crop Science, in press.*

MAHALANOBIS, P. C. (1936). On the generalized distances in statistics. *Proc. Natl. Inst. Sci. India.* **2** 49–55.

NUEL, G. (1998). Prédiction de distances phénotypiques à partir de données moléculaires pour la distinction variétale. Master's thesis, INA-PG – Université Paris XI.

RAO, C. R. and MITRA, S. K. (1971). *Generalized inverse of matrices and its applications.* New York: Wiley.

ROGERS, J. (1972). Measures of similarity and genetic distance. In *Studies in Genetics*, volume VII, 145–153. Univ. Texas Publ.

SCHEFFÉ, H. (1959). *The analysis of variance.* Wiley.

Appendix

12.A Proofs

Proposition 1. This can be proved using the same kind of argument as Cochran's theorem (see for example Scheffé (1959)). Since $\mathbf{M}$ is a metric, it is defined and positive, we hence decompose it as $\mathbf{M} = \mathbf{PDP}'$. We obtain

$$D^2_{\mathbf{M}}(a,b) = (\mathbf{Z}_a - \mathbf{Z}_b)(\mathbf{Z}_a - \mathbf{Z}_b)',$$

where $\mathbf{Z}_a = \mathbf{Y}_a\mathbf{PD}^{\frac{1}{2}}$. Since $\mathbf{Y}_a$ is normal with mean μ_a and variance $\mathbf{\Sigma}$, the distribution of $\mathbf{Z}_a$ is

$$\mathbf{Z}_a \sim \mathcal{N}_p\left(\mu_a \mathbf{PD}^{\frac{1}{2}}\ ;\ \mathbf{D}^{\frac{1}{2}}\mathbf{P}'\mathbf{\Sigma}\mathbf{PD}^{\frac{1}{2}}\right).$$

Furthermore, since $\mathbf{\Sigma}$ is positive, one has $\lambda_j \geq 0$ for all $1 \leq j \leq p$. The variance of the transformed variable $\mathbf{T}_a = \mathbf{Z}_a\mathbf{Q} = \mathbf{Y}_a\mathbf{PD}^{\frac{1}{2}}\mathbf{Q}$ is diagonal: $\mathbb{V}(\mathbf{T}_a) = \mathbf{Q}'\mathbf{D}^{\frac{1}{2}}\mathbf{P}'\mathbf{\Sigma}\mathbf{PD}^{\frac{1}{2}}\mathbf{Q} = \mathbf{Q}'\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'\mathbf{Q} = \mathbf{\Lambda}$, so

$$\mathbf{T}_a \sim \mathcal{N}_p\left(\mu_a\mathbf{PD}^{\frac{1}{2}}\mathbf{Q}\ ;\ \mathbf{\Lambda}\right)$$

and $D^2_{\mathbf{M}}(a,b) = (\mathbf{T}_a - \mathbf{T}_b)(\mathbf{T}_a - \mathbf{T}_b)'$.

Hence we have shown that $D^2_{\mathbf{M}}(a,b)$ is the squared norm of a normal vector with mean $\nu = (\mu_a - \mu_b)\mathbf{PD}^{\frac{1}{2}}\mathbf{Q} = [\nu^1\ \dots\ \nu^p]$ and independent coordinates with respective variances $2\lambda_j$, so the proposition is proved. ■

Proposition 2. One has $\mathbb{E}[\widehat{\delta}^2(a,b)] = (\mathbf{u}_a - \mathbf{u}_b)\,\mathbb{E}(\widehat{\mathbf{Y}}\mathbf{M}\widehat{\mathbf{Y}}')\,(\mathbf{u}_a - \mathbf{u}_b)'$ and, according to equation (12.6) and model (12.1),

$$\widehat{\mathbf{Y}}^j = \mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{Y}^j = \mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\left(\mathbf{X}_{\mathcal{B}_j}\theta_{\mathcal{B}_j} + \mathbf{E}^j\right) = \mu^j + \mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{E}^j.$$

So $\widehat{\mathbf{Y}}\mathbf{M}\widehat{\mathbf{Y}}'$ can be decomposed as

$$\widehat{\mathbf{Y}}\mathbf{M}\widehat{\mathbf{Y}}' = \mu\mathbf{M}\mu' + \mathbf{2}\mu\mathbf{MG} + \mathbf{GMG}'$$

where $\mu = [\mu^1\ \dots\ \mu^p]$ and $\mathbf{G} = \left[\mathbf{A}_{\mathcal{B}_1}\mathbf{\Pi}_1\mathbf{E}^1\ \dots\ \mathbf{A}_{\mathcal{B}_p}\mathbf{\Pi}_1\mathbf{E}^p\right]$; its mean is $\mathbb{E}(\widehat{\mathbf{Y}}\mathbf{M}\widehat{\mathbf{Y}}') = \mu\mathbf{M}\mu' + 2\mu\mathbf{M}\mathbb{E}(\mathbf{G}) + \mathbb{E}(\mathbf{GMG}')$. Since each row $\mathbf{E}$ has a normal distribution with mean $\mathbf{0}$, $\mathbb{E}(\mathbf{G})$ is null, so $\mathbb{E}[\widehat{\delta}^2(a,b)]$ is equal to

$$(\mathbf{u}_a - \mathbf{u}_b)\,\mu\mathbf{M}\mu'\,(\mathbf{u}_a - \mathbf{u}_b)' + \mathbb{E}\left[(\mathbf{u}_a - \mathbf{u}_b)\,\mathbf{GMG}'\,(\mathbf{u}_a - \mathbf{u}_b)'\right]$$

where the first term equals $\delta^2(a,b)$ and where the second term is the expectation of a square distance and is hence positive.

The term (a,b) of $\mathbf{GMG}'$ equals $\sum_{j=1}^p \sum_{j'=1}^p m_j^{j'} \times G_a^j G_b^{j'}$. Denoting the general term of $\mathbf{G}$ by G_a^j, one has $G_a^j = \mathbf{u}_a\mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{E}^j$ so

$$\begin{aligned}\mathbb{E}\left(G_a^j G_b^{j'}\right) &= \mathbf{u}_a\mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbb{E}\left[\mathbf{E}^j\left(\mathbf{E}^{j'}\right)'\right]\mathbf{\Pi}_1\mathbf{A}_{\mathcal{B}_{j'}}(\mathbf{u}_b)' \\ &= \sigma_j^{j'}\left[\mathbf{u}_a\mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{A}_{\mathcal{B}_{j'}}(\mathbf{u}_b)'\right].\end{aligned}$$

The expectation of $\mathbf{GMG}'$ gives the result. ■

Corollary 3. Since all the projectors are equal, one has $\mathbf{A}_{\mathcal{B}_j}\mathbf{\Pi}_1\mathbf{A}_{\mathcal{B}_{j'}} = \mathbf{A}\mathbf{\Pi}_1\mathbf{A}$. Applying proposition **2**, we get

$$h_a^b = \left[\mathbf{u}_a\mathbf{A}\mathbf{\Pi}_1\mathbf{A}\left(\mathbf{u}_b\right)'\right]\sum_{j=1}^{p}\sum_{j'=1}^{p}\sigma_j^{j'}m_j^{j'}.$$

Furthermore, in the case of the Mahalanobis distance, the metric $\mathbf{M}$ is $\frac{1}{2}\Sigma^{-1}$, so $\mathbf{\Sigma M} = \frac{1}{2}\mathbf{I}_p$, hence $\sum_{j=1}^{p}\sum_{j'=1}^{p}\sigma_j^{j'}m_j^{j'} = \frac{p}{2}$ and the corollary is proved. ■