

Analyse statistique des séquences biologiques

*modélisation markovienne,
alignements et motifs*

Grégory Nuel
Bernard Prum

Table des matières

Introduction	21
PREMIÈRE PARTIE. MODÉLISATION DE SÉQUENCES	37
Chapitre 1. Chaînes de Markov	39
1.1. Un exemple	39
1.2. Modèles d'indépendance	41
1.2.1. Modèle shuffle	44
1.3. Rappels sur les chaînes de Markov	44
1.3.1. Introduction	44
1.3.2. Loi de X_i , loi stationnaire	47
1.3.3. Chaînes de Markov d'ordre supérieur à un	50
1.3.4. Générateur infinitésimal	51
1.4. Application aux séquences biologiques	52
1.4.1. Estimation du modèle, loi à portée d	52
1.4.2. Loi stationnaire	54
1.5. Choix d'un modèle	55
1.6. Les chaînes de Markov phasées	58
1.7. Diverses chaînes de Markov parcimonieuses	59
1.7.1. Les VLMC	61
1.7.2. Les PMM	62
1.7.3. Les MTD	64
1.8. Les chaînes de Markov dérivantes	65
1.8.1. Dérive linéaire	66
1.8.2. Dérive générale	68
1.9. Notes bibliographiques	68

Chapitre 2. Chaînes de Markov cachées	69
2.1. Motivation	69
2.2. Les modèles HMM	70
2.2.1. Modèle M1M1	70
2.2.2. Modèle M1M m	71
2.2.3. Longueur des plages, ergodicité	71
2.3. Inférence	72
2.3.1. La vraisemblance	72
2.3.2. L'algorithme EM	73
2.3.2.1. La phase M :	74
2.3.2.2. La phase E :	75
2.3.2.3. L'algorithme récursif EM	77
2.3.2.4. EM et gel final	77
2.3.3. Qualité des estimations	79
2.3.3.1. Rappels	79
2.3.3.2. Vraisemblance conditionnelle	80
2.3.3.3. Loi des estimateurs	81
2.3.3.4. Conclusion	82
2.3.4. L'algorithme SEM	82
2.3.5. L'algorithme de Viterbi	83
2.4. Les SHMM	85
2.5. Exemples d'applications	86
2.5.1. Modélisation des gènes	86
2.5.2. Profils HMM	88
2.5.2.1. Les RBS	89
2.5.2.2. Les sites donneurs et accepteurs	89
2.6. Chaînes de Markov cachées et score local	90
2.6.1. Algorithme de Viterbi	93
2.6.2. Algorithme EM	93
2.6.3. Variance de l'estimateur	95
2.6.3.1. Information de Fisher	95
2.6.3.2. Variances	96
2.6.3.3. Variance pour les scores	97
2.6.3.4. Simulations	97
2.6.4. Score local avec m segments	99
2.6.4.1. Vraisemblance	99
2.6.4.2. Algorithme de Viterbi	100
2.6.4.3. Algorithme EM	100
2.6.4.4. Information de Fisher	101
2.7. Logiciels et notes bibliographiques	101
2.7.1. Quelques logiciels	101
2.7.2. Notes bibliographiques	101

DEUXIÈME PARTIE. MOTIFS	103
Chapitre 3. Compter les motifs	105
3.1. Définitions	105
3.1.1. Alphabet	105
3.1.2. Séquence	106
3.1.3. Mot	106
3.1.3.1. Palindromes	106
3.1.4. Motif	107
3.1.5. Que compter ?	107
3.1.5.1. Pour un mot	107
3.1.5.2. Pour un motif	108
3.2. Automates	109
3.2.1. Langages	109
3.2.2. Automates Finis Déterministes	110
3.2.3. Comptages	112
3.3. Algorithmes	114
3.3.1. Construction d'automates	114
3.3.1.1. Automate non déterministe	115
3.3.1.2. Déterminisation	118
3.3.1.3. Minimisation	120
3.3.1.4. Heuristique	120
3.3.2. Arbres de suffixes	124
3.3.3. Arbres de préfixes	129
3.4. Notes bibliographiques	132
Chapitre 4. Statistiques de motifs	135
4.1. Cyrano de Bergerac	135
4.2. Statistique de motifs	139
4.3. Pattern Markov Chain	140
4.3.1. Modèle M_0	141
4.3.2. Un cas simple	142
4.3.3. Modèle M_m	144
4.4. Calculs exacts	150
4.4.1. Finite Markov Chain Imbedding	151
4.4.2. Algorithmes	154
4.4.2.1. Développements asymptotiques	156
4.4.3. Temps d'attente	157
4.4.3.1. Etudier la répartition des motifs	158
4.4.3.2. Simuler la répartition des motifs	160
4.4.3.3. Considérations numériques	161
4.4.4. Moments	163
4.4.4.1. Espérance	163
4.4.4.2. Variance	165

4.4.5. Lois jointes	171
4.4.6. Plusieurs séquences	173
4.4.7. Modèle hétérogène	175
4.5. Approximations gaussiennes	176
4.5.1. Cas Markov	176
4.5.2. Lois jointes	177
4.5.3. Loi de (N_m, N_{m+1})	177
4.5.4. Approche fondée sur les martingales	183
4.5.5. Modèle shuffle	184
4.5.5.1. Espérance	185
4.5.5.2. Variance	187
4.5.5.3. Approximation gaussienne	188
4.6. Approximations binomiales	188
4.6.1. Prise en compte de l'estimation des paramètres	193
4.7. Approximations de Poisson composées	198
4.7.1. Mots recouvrants	198
4.7.1.1. Structure d'un mot périodique	198
4.7.1.2. Occurrences par paquets	199
4.7.1.3. Calcul de θ	200
4.7.1.4. Loi de Poisson géométrique	200
4.7.1.5. Exemple du traitement par AFD	201
4.7.1.6. Cas de motifs	201
4.7.2. Matrice d'autorecouvrement	202
4.7.3. Résultat principal	204
4.7.4. Cas Poisson	205
4.7.5. Cas Poisson géométrique	205
4.7.6. Cas général	211
4.8. Grandes déviations	213
4.8.1. Introduction	213
4.8.2. Niveau 1	215
4.8.2.1. Calculs numériques	217
4.8.3. Niveau 2	218
4.8.3.1. Mise en œuvre pratique	220
4.9. Comparaison des méthodes	220
4.9.1. Complexités	221
4.9.2. Comparaison Markov versus shuffle	225
4.9.3. Grandes déviations précises	225
4.9.4. Cas extrêmes	226
4.9.5. Cas réels	229
4.9.6. Conclusions	232
4.10. Notes bibliographiques	234

Chapitre 5. Motifs biologiques	237
5.1. Chi	237
5.2. Régulation	244
5.3. PROSITE	246
5.4. Scan statistics	247
5.5. Notes bibliographiques	251
5.5.1. Chi	251
5.5.2. Régulation	252
5.5.3. Prosite	252
5.5.4. Scan statistics	252
TROISIÈME PARTIE. ALIGNEMENTS DE SÉQUENCES	253
Chapitre 6. Score local d'une séquence	255
6.1. Définition	255
6.1.1. Segment de score maximal	255
6.1.2. Algorithme linéaire	256
6.1.3. Segments sous-optimaux	258
6.2. Significativité exacte	259
6.2.1. Cas simple	260
6.2.2. Extension aux scores rationnels	264
6.2.3. Extension au cas markovien	265
6.3. Approximations asymptotiques	265
6.3.1. Runs de 1, loi de Gumbel	265
6.3.1.1. Exemple simple	265
6.3.1.2. Application au cas Bernoulli	267
6.3.2. Loi asymptotique du score d'une séquence	267
6.3.3. Validité des approximations	270
6.4. Notes bibliographiques	271
Chapitre 7. Alignement de deux séquences	273
7.1. Introduction	273
7.1.1. L'évolution ponctuelle	273
7.1.2. Matrices d'évolution	275
7.1.2.1. Evolution des séquences nucléotidiques	275
7.1.2.2. Modèle de Jukes et Cantor	277
7.1.2.3. Modèle de Kimura	277
7.1.2.4. Autres modèles	278
7.1.2.5. Evolution des séquences protéiques	278
7.1.2.6. Evolutions non ponctuelles	279
7.2. L'alignement de deux séquences d'ADN	280
7.2.1. Nombre d'alignements possibles	283

7.3. Score global : Needleman et Wunsch	285
7.3.1. L'algorithme de programmation dynamique	285
7.3.2. Recherche de l'alignement : le trace-back	287
7.3.3. Complexité, algorithme SL	290
7.3.3.1. Algorithme SL (Space Linear)	290
7.4. Score local : Smith et Waterman	291
7.4.1. Alignement global de séquences tronquées	291
7.4.2. Programmation dynamique et trace-back	292
7.5. Scores de gap affines	294
7.6. Significativité	295
7.7. Heuristiques, BLAST	295
7.8. Alignement et HMM	297
7.9. Notes bibliographiques	300
Chapitre 8. Alignements multiples	303
8.1. Une heuristique d'alignement multiple	305
8.1.1. L'arbre guide : CLUSTAL	306
8.2. Ouverture vers la phylogénie	307
8.2.1. Phylogénie et distances	307
8.2.2. Phylogénie et parcimonie	308
8.2.2.1. Calcul du coût d'un arbre	308
8.2.2.2. Recherche de l'arbre le plus parcimonieux	309
8.2.3. Phylogénie et vraisemblance	310
8.2.3.1. L'algorithme PhyML	312
8.3. Notes bibliographiques	313
Chapitre 9. Matrices de similarité	315
9.1. Les matrices PAM	316
9.2. Discussion critique	318
9.3. Les matrices BLOSUM	319
9.4. Autres matrices	321
9.4.1. Sensibilité au choix de $\mathbb{S}$	321
9.5. Notes bibliographiques	322
ANNEXES	323
A. L'algorithme EM	325
A.1. La phase M	326
A.2. La phase E	326
A.3. L'algorithme récursif	327
A.4. Variances des estimateurs	327
A.5. Notes bibliographiques	329

B. Arbres, distances et algorithme NJ	331
B.1. Distances d'arbre, distances phylogénétiques	331
B.2. L'algorithme NJ	333
B.3. Notes bibliographiques	335
C. Valeurs propres et vecteurs propres	337
C.1. Analyse spectrale	337
C.1.1. Matrices positives	337
C.1.2. Matrices stochastiques	339
C.1.3. Méthode de la puissance $n^{\text{ième}}$	340
C.2. Algorithme QR	341
C.3. Algorithme d'Arnoldi	345
C.4. Notes bibliographiques	347
Bibliographie	349
Index	359