Régression logistiquemultiple et prédiction des facteurs de risque concernant la survie en soins intensifs.

changer le titre

Julia Guerra [1], Maxime Jaunatre [2], Ellie Tideswell [3] | Master 2 BEE Grenoble
Mail [1], Mail [2] Mail [3]| 29 novembre 2019

Todo list

Introduction

L'avancée des techniques de séquençage a permis d'obtenir de grandes quantités d'informations génomiques. Durant ces dernières années, la génétique a embrassé les outils mathématiques de modélisation, parvenant à une caractérisation statistique des données de séquençage. Cette approche a permis de décrire avec précision les motifs observés dans des régions étudiées , et de prédire leurs présences dans des séquences encore inconnues (Wu *et al.* , 2010). Une des méthodes les plus connues dans ce domaine est l'utilisation des chaînes de Markov, introduites premièrement par Churchill (1992) pour l'analyse de séquences génomiques puis par Durbin *et al.* (1998) pour la détection de régions CGI.

Dans le génome des deutérostomiens, la fréquence du dinucléotide C-G est moins importante qu'attendu sous une distribution aléatoire des quatre bases azotées. Ceci est une conséquence des mécanismes de protection contre la mutation spontanée du génome. En effet les bases de Thymine (T) subissent des erreurs de réplication quand elles sont placées sur le même brin et directement après une couple Cytosine-Guanine. L'élimination de ces couples C-G potentiellement dangereux se fait suite à un marquage par méthylation. Cependant, dans certaines régions de l'ADN nommées îlots CpG (ou CGI) ce processus de méthylation est inhibé et donc la fréquence des dinucléotides C-G est donc plus élevée ; par exemple aux alentours de certains promoteurs (Haque *et al.* , 2011, Saxonov *et al.* , 2006, Wu *et al.* , 2010).

La grande variabilité dans la taille, la composition et l'emplacement de ces CGI rend difficile leurs définitions et donc l'établissement d'un algorithme unique permettant leurs détections indubitable (Wu *et al.* , 2010). Ainsi, les modèles de Markov permettent de modéliser les fréquences des nucléotides en fonction de séquences déjà connues ; ces séquences contenant ou non des CGI. En supplément des chaînes de Markov simples, il existe aussi les chaînes de Markov cachées : ces dernières décrivent de nombreux processus réels qui suivent un modèle de Markov, mais qui ne sont que partiellement observables. Une chaîne de Markov cachée permettrait donc l'utilisation d'un seul modèle pour identifier un nucléotide (l'observation) et si ce dernier est à l'intérieur d'un îlot CpG ou non (aussi appelé l'état de la région).

Matériel et méthodes

construire un modèle (markov simple)

jeu de données algo d'apprentissage

$$P_+(Sequence) = log(P_initiale(mot_initial)) + \quad (1)$$

test du meilleur modele

calculer sensi et speci pour une base evaluer tout les sensi et speci tout au long séquence

markov caché et algo viterbi

pourquoi que c'est mieux qu'une fenetre glissante parametrages viterbi - transition entre modèles + et -

smoothing (maxime)

pourquoi comment

Résultats

mus1

table tronquée figure tronquée description du chromosome (nombre d'îlots cpg, taille des cpg, fenetre de smooth)

mus2

description

mus3

description

Discussion

nos résultats sont badass moduler le choix de smoothing améliorations des algos (coder en autre langage), parallelisation du choix de meilleur modele

Ressources

Ce document est disponible en ligne sous format ".Rnw", contenant tout le code néccessaire à la reproduction de l'analyse, réalisée avec un script en langage R (R.Team, 2017), ainsi que le jeu de données de départ. L'ensemble est situé sur Github : `https://github.com/gowachin/BeeMarkov`

Annexes

mus1

table figure

mus2

table figure

mus3

table figure

scripts

Bibliographie

Churchill, Gary A. 1992. Hidden Markov chains and the analysis of genome structure. *Computers and Chemistry*, **16**(2), 107–115.

Durbin, Richard, Eddy, Sean R., Krogh, Anders, & Mitchison, Graeme. 1998. Biological sequence analysis. *Biological sequence analysis*.

Haque, A. N.A., Hossain, M. E., Haque, M. E., Hasan, M. M., Malek, M. A., Rafii, M. Y., & Shamsuzzaman, S. M. 2011. CpG islands and the regulation of transcription. *GENES & DEVELOPMENT*, **25**(1), 1010–1022.

R.Team. 2017. R : A language and environment for statistical computing (Version 3.4. 2)[Computer software]. *Vienna, Austria : R Foundation for Statistical Computing*.

Saxonov, Serge, Berg, Paul, & Brutlag, Douglas L. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(5), 1412–1417.

Wu, Hao, Caffo, Brian, Jaffee, Harris A., Irizarry, Rafael A., & Feinberg, Andrew P. 2010. Redefining CpG islands using hidden Markov models. *Biostatistics*, **11**(3), 499–514.