Motifs caches dans les textes

(travail commun avec Philippe Flajolet, Yves Guivarc'h, Wojciech Szpankowski)
Analyse de quicksort optimisé

Le probleme de la recherche d'un motif donne dans un texte est un probleme tres classique et tres bien connu d'un point de vue algorithmique et probabiliste. Ici, le probleme traite est plus general, car nous nous interessons aux occurrences de ce motif lorsque les symboles contigus du motif n'apparaissent plus necessairement de maniere contigue dans le texte. Par exemple, baba n'apparait pas comme motif dans le texte abracadabra, mais il apparait 3 fois comme sous sequence de ce texte: on peut considere que baba est 3 fois "cache" dans le texte. Ce probleme qui se pose naturellement dans des contextes lies a la biologie ou a la securite informatique n'a jamais ete reellement aborde d'un point de vue probabiliste: on se donne un motif, un ensemble de contraintes sur les ecarts acceptes entre les occurrences des symboles du motif, et on se pose la question: que peut-on dire du nombre de fois ou le motif apparait comme sous-sequence admise dans un texte aleatoire. La rponse cette question est essentielle si on veut pouvoir separer les occurrences anormalement frequentes (en plus ou en moins) des occurrences normalement frequentes. Nous repondons ici de maniere tres precise a la question....