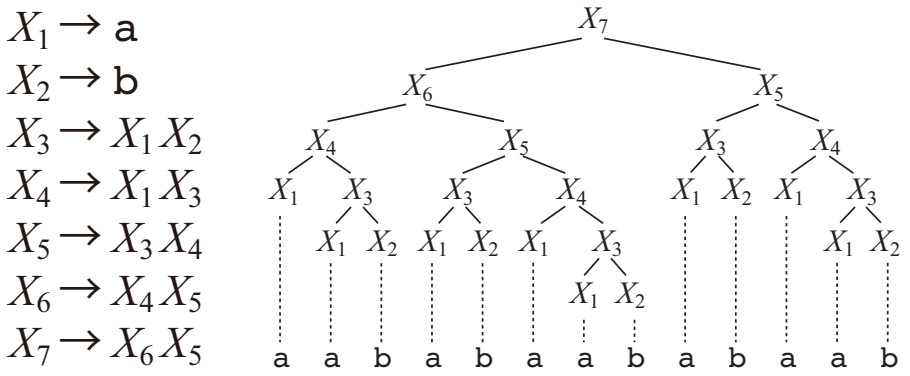
# 圧縮文字列処理のアルゴリズム

九州大学 大学院システム情報科学研究院 情報学部門
坂内 英夫

文字列 (データ) 圧縮とは，与えられた文字列に内在する規則性に基づき，より短い表現に変換する操作である．文字列圧縮は大規模なデータの通信・記憶容量の節約に有効であり，これまでに様々な手法が提案され，広く利用されている．一方で，圧縮された文字列データを実際に処理・解析するためには圧縮表現を展開して元の文字列に戻してから行うのが一般的である．しかし，そのような方法では結局元の大規模データを扱うことになるため，多くの計算資源が要求されることに変わりない．

これに対して「圧縮文字列処理」は，一度得られた圧縮表現を陽に展開せず，直接処理することで処理に必要な記憶容量，更には計算時間をも削減することを目指すアプローチである．圧縮文字列処理アルゴリズムの計算量は圧縮表現のサイズに依存し，表現が短いほど，すなわち，良く圧縮できるデータほど計算に必要な領域と時間は小さくなる．そのため，特に近年増加しているバージョン管理された文書や同一生物種の複数個体のゲノム配列集合など，共通部分が多く圧縮しやすいデータに対して圧縮文字列処理のアプローチによる省領域化・高速化が期待できる．

本講演では，多くの圧縮アルゴリズムの出力をモデル化できる単一の文字列を生成するチョムスキー標準形の文脈自由文法 (Straight Line Program - SLP) を処理対象の圧縮表現とする圧縮文字列処理アルゴリズムについて，最近の研究成果を幾つか紹介する．




文字列 “aababaababaab” を表現する SLP の例 (左) およびその導出木 (右).