Re-pairアルゴリズムを用いた効率よいVF符号

吉田諭史，喜田拓也*

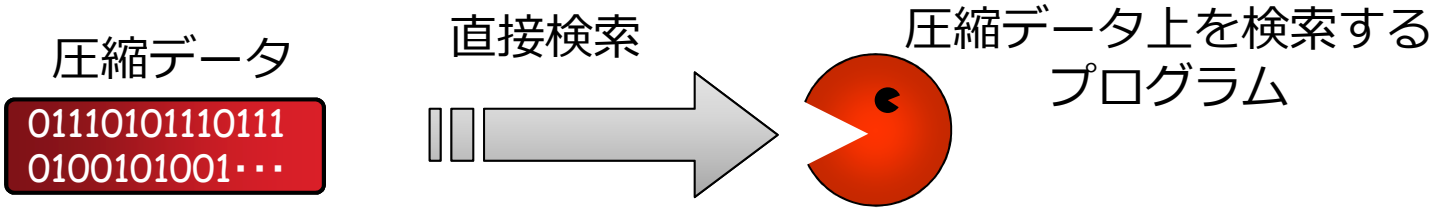
*北海道大学大学院情報科学研究科コンピュータサイエンス専攻

背景：VF符号

入力データを可変長の文字列に分割して，それぞれ固定長の符号語に変換する符号化手法.

特長：圧縮データが取り扱いやすい

➪情報検索，データマイニングが高速に

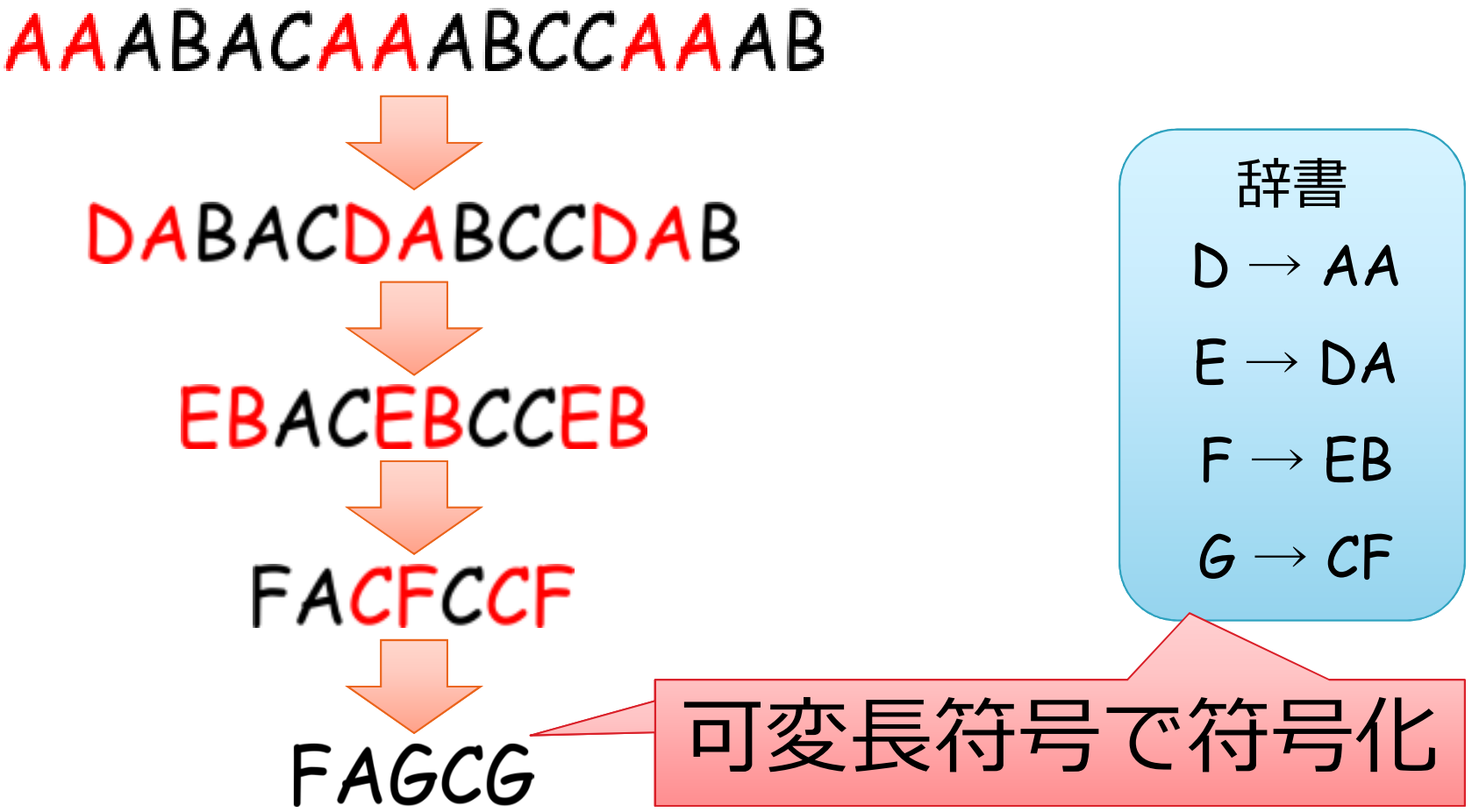


		入力テキスト（情報源記号）	
		固定長 (F)	可変長 (V)
圧縮テキスト（符号語）	固定長 (F)	**FF符号** 等長符号	**VF符号** Tunstall符号
	可変長 (V)	**FV符号** Huffman符号	**VV符号** Re-pair

高い圧縮率を達成するVF符号の開発が必要

Re-pairアルゴリズム

文法変換に基づく圧縮の一種で，すべてのbigramがuniqueになるまで最頻出のbigramをひとつの記号に置き換える.



本研究の主結果

- Re-pairアルゴリズムを用いたVF符号を提案した.
- 提案アルゴリズムを実験的に評価した.
  - 圧縮率は，自然言語について，gzipよりも良い結果が得られた.
  - 圧縮速度は，オリジナルのRe-pairとほとんど変わらない結果となった.
  - 伸長速度は，bzip2よりも良い結果が得られた.