文字列の頻度分布による共通パタン発見

池田, 大輔
九州大学情報基盤センター

山田, 泰寛
九州大学大学院システム情報科学府

廣川, 佐千男
九州大学情報基盤センター

出版情報：情報処理学会研究報告 : 自然言語処理. 2003 (98), pp.25-32, 2003-09
バージョン：

文字列の頻度分布による共通パタン発見

池田 大輔、山田 泰寛*、廣川佐千男

九州大学情報基盤センター

*九州大学システム情報学研究院

共通パタン発見問題

最長共通部分列問題

NP完全

F172GNL154

自然言語処理でのパタン発見

- 定型表現や未知語の抽出
  - Nagao & Mori (1994)など
  - 棋譜からの定石発見(中村2002)
- 固定長文字列(n-gram)の頻度を利用
  - 頻度の高いところが抽出したいところ

長さをどう決めるかが問題

目的

- 共通パタン発見問題の定式化
- 問題を解くアルゴリズムの構築
  - 理論的に「解く」
  - 高速かつ頑健な実装

- はじめに
- パタン言語
- テンプレート発見問題
- 実験
- まとめ

F172GNL154

はじめに

F172GNL154

共通/頻出パタンの重要性

- データ/テキスト/Webマイニング
  - 隠れた有用な規則を見つける
- 遺伝子情報学
  - 重要な機能を司る部分は共通
- 圧縮、キャッシュ
- 自然言語処理
  - 定型表現、情報抽出

頻度による情報抽出

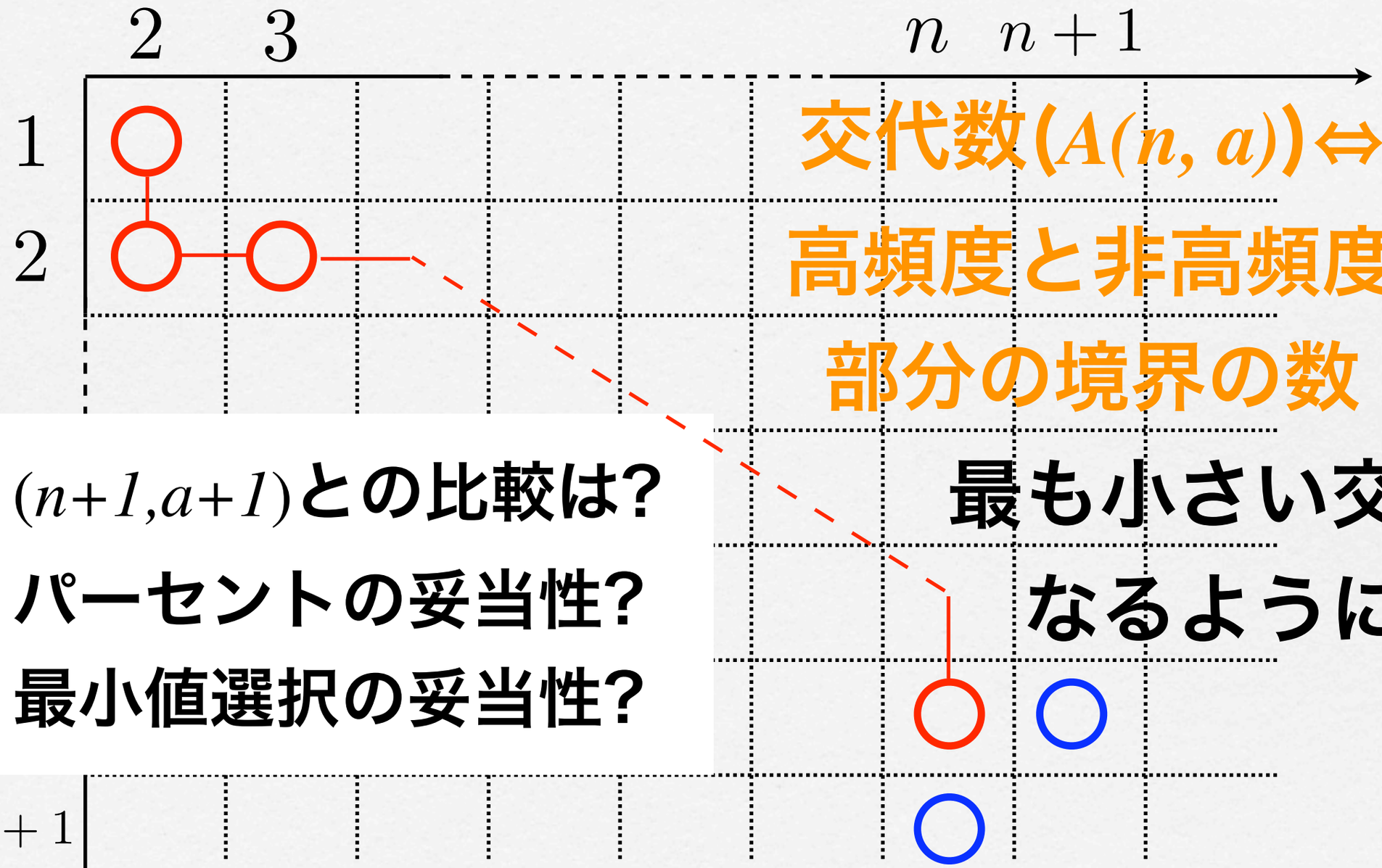
Ikeda *et. al.*, Discovery Science 2001 (LNCS 2226)

- テンプレート部分は高頻度であろう

→高頻度部分を自動で検出する

- 高頻度⇔長さnの部分文字列の上位a%
- 交代数を用いて(n, a)を自動的に決定
- 様々な言語のHTML/XMLファイルに対し
  高精度でテンプレート抽出に成功

最適な(n, a)の決定



交代数($A(n, a)$)⇔

高頻度と非高頻度

部分の境界の数

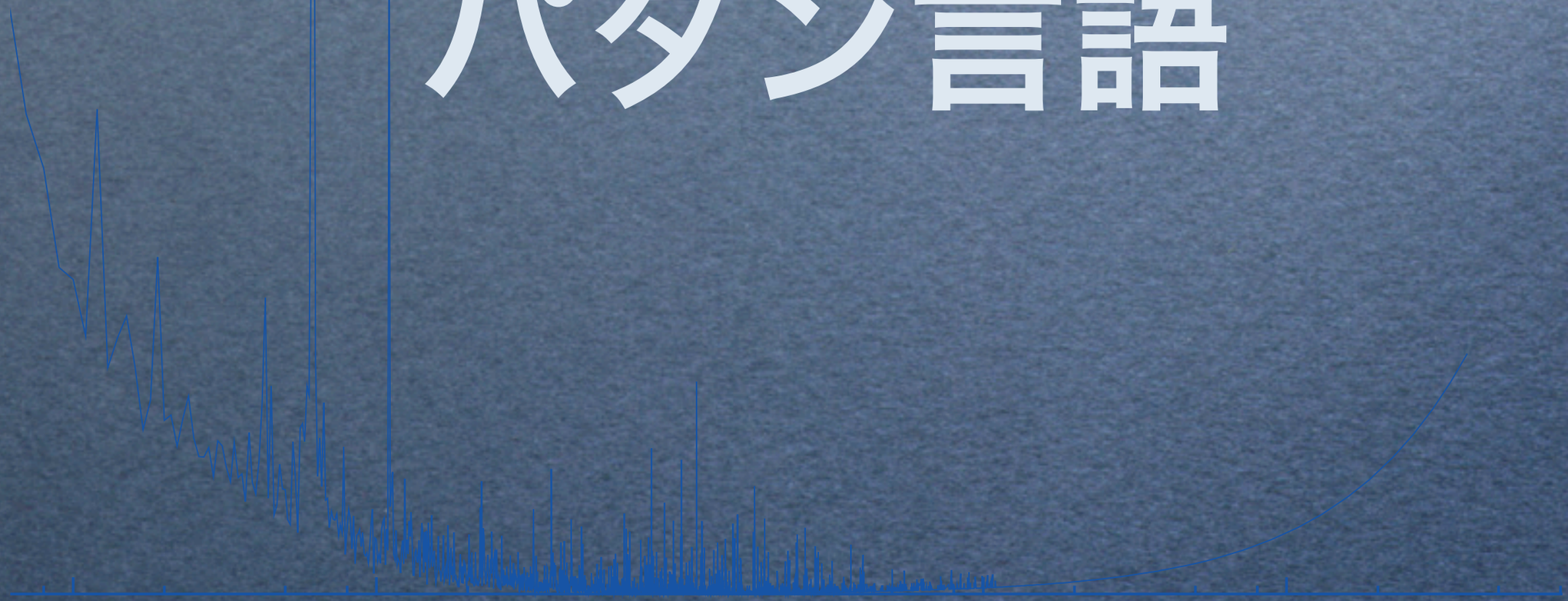
最も小さい交代数に

なるように遷移

- $(n+1,a+1)$との比較は?
- パーセントの妥当性?
- 最小値選択の妥当性?

F172GNL154

パタン言語



F172GNL154

パタン言語

- Σ：定数、V：変数
- パタン$\Leftrightarrow \Sigma \cup V$上の文字列
  - 例：$p = axbyax(\Sigma = \{a, b\}, V = \{x, y\})$
- 代入$\Leftrightarrow\ \theta : V \to \Sigma^*$
- パタンpの言語 $L(p) = \{w \in \Sigma^* \mid \exists\theta; p\theta = w\}$

パタン言語の学習

- パタン言語の学習
  - 入力：(有限)文字列の集合S
  - 問題：以下を満すパタンpを見つけよ

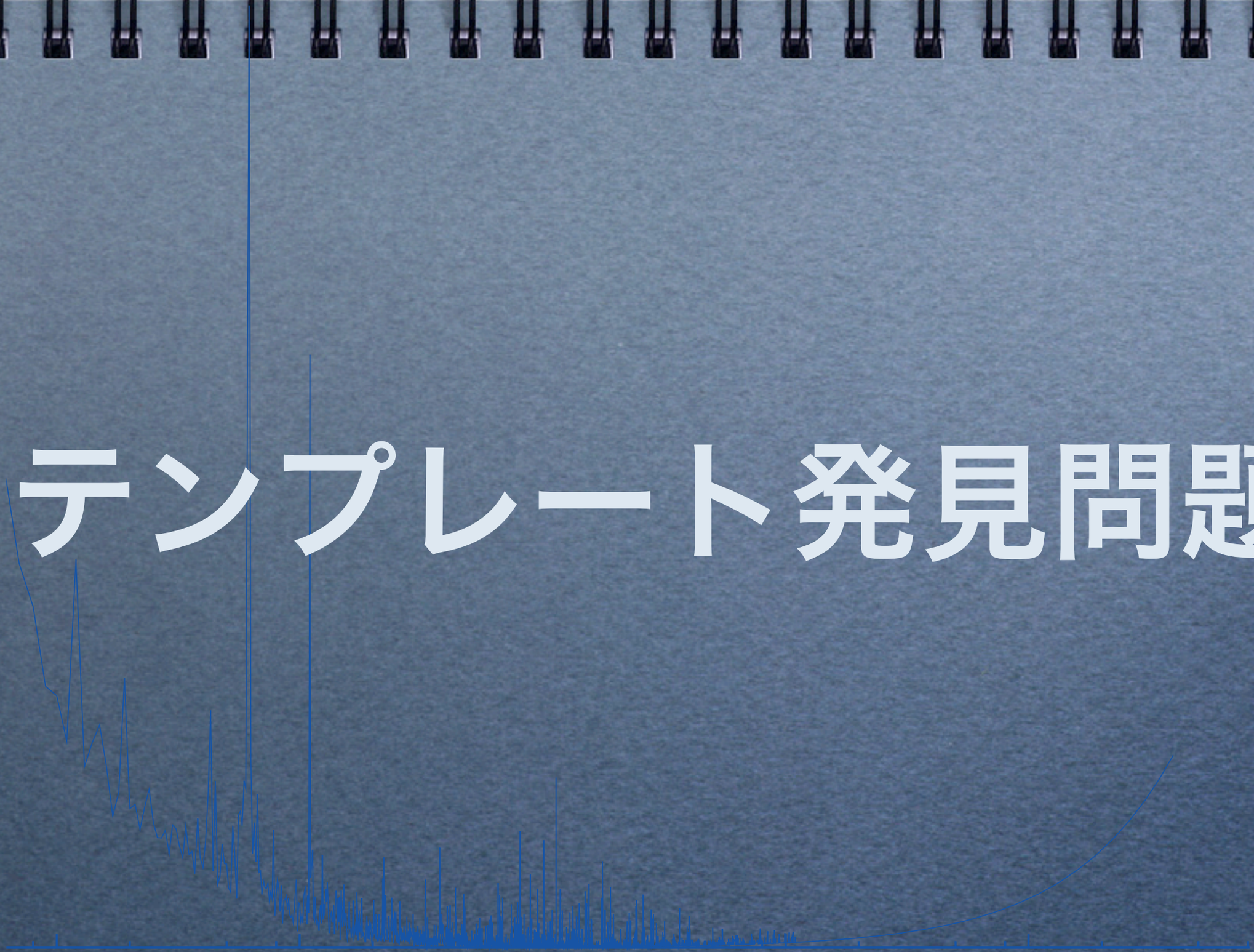
$$S \subseteq L(p) \quad \forall S; p = x \rightarrow S \subseteq L(p) = \Sigma^*$$

$$\nexists q; L(q) \subseteq L(p)$$

パタン言語学習の難しさ

- 1変数パタン(Angluin)
- Regular Pattern (Shinohara)
  - 各変数は高々1回しか現われない
- k-variableパタン(Kearn & Pit)
  - ここまで制限すると多項式時間学習可能

テンプレート発見問題



F172GNL154

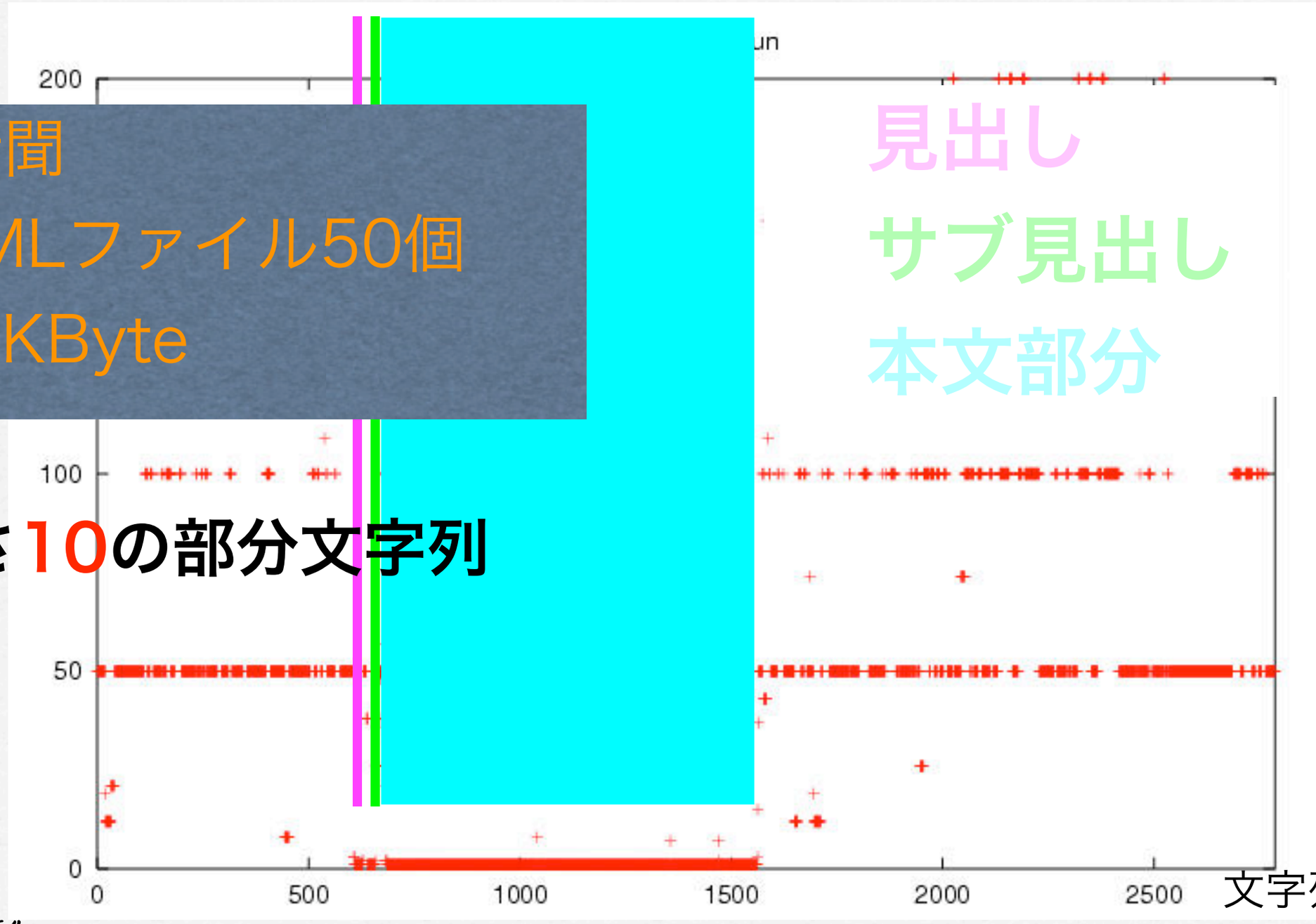
テンプレート

- $p = w_1 x_1 \cdots w_m x_m$ ：パタン
- テンプレート$\Leftrightarrow$ $t = (w_1, \ldots, w_m)$
- tの幅$\Leftrightarrow$ $\min\{|w_i| \mid w_i \in t\}$

共通パタン発見問題

定数部分

位置ごとの頻度

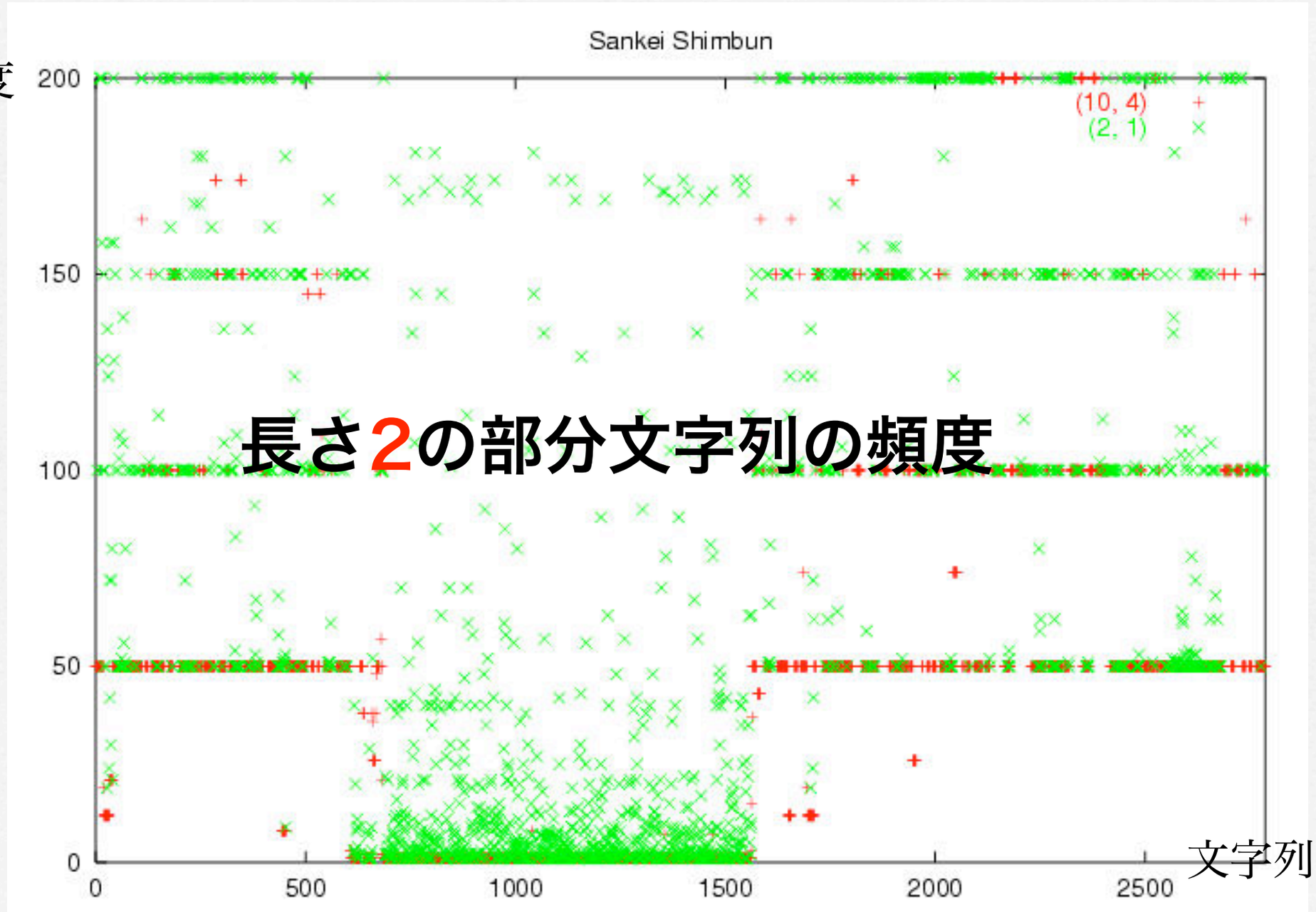


産経新聞
ーHTMLファイル50個
ー328KByte

長さ10の部分文字列

F172GNL154

位置ごとの頻度(Cont.)



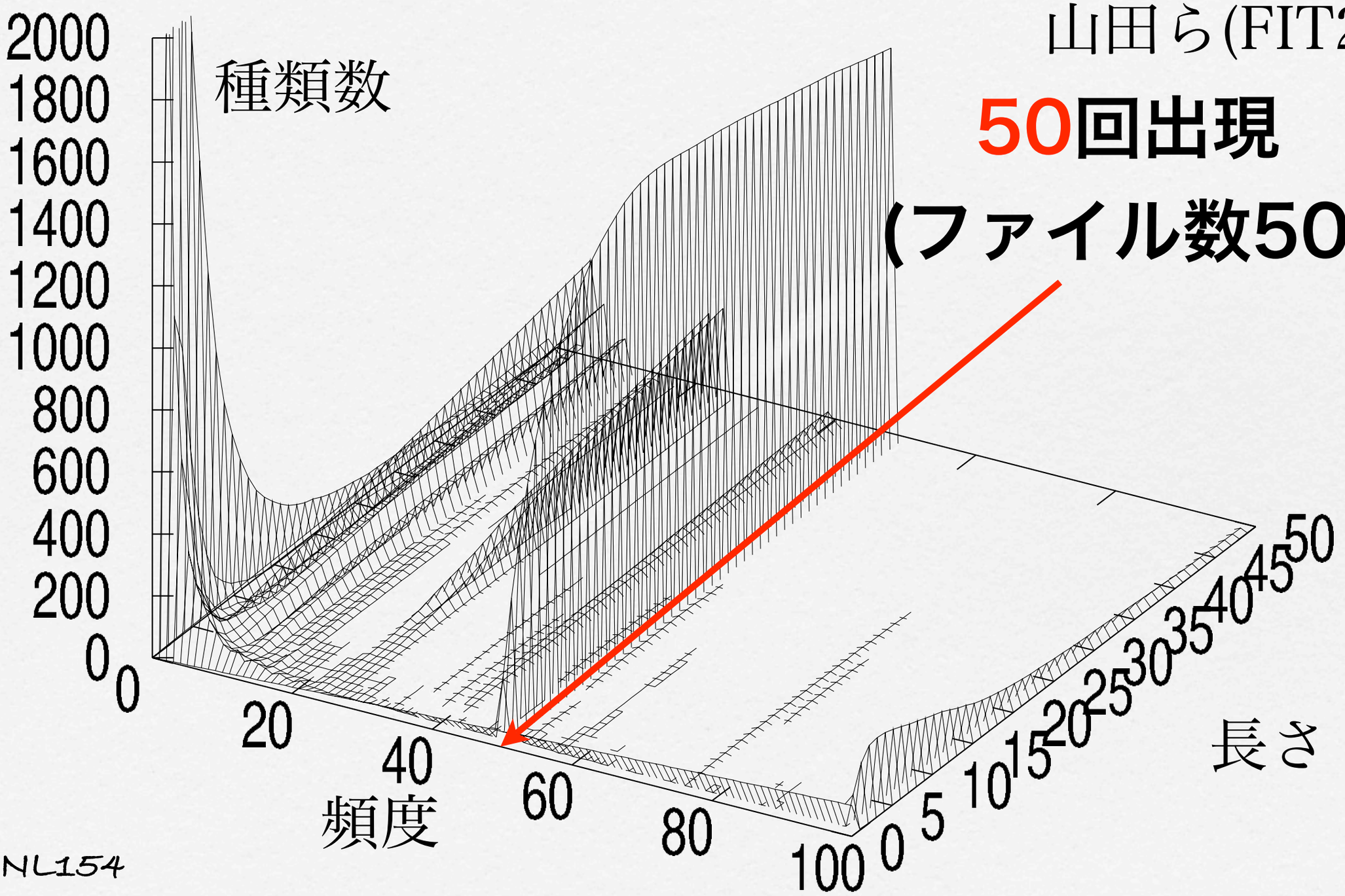
F172GNL154

任意の長さに拡張

山田ら(FIT2003)

50回出現

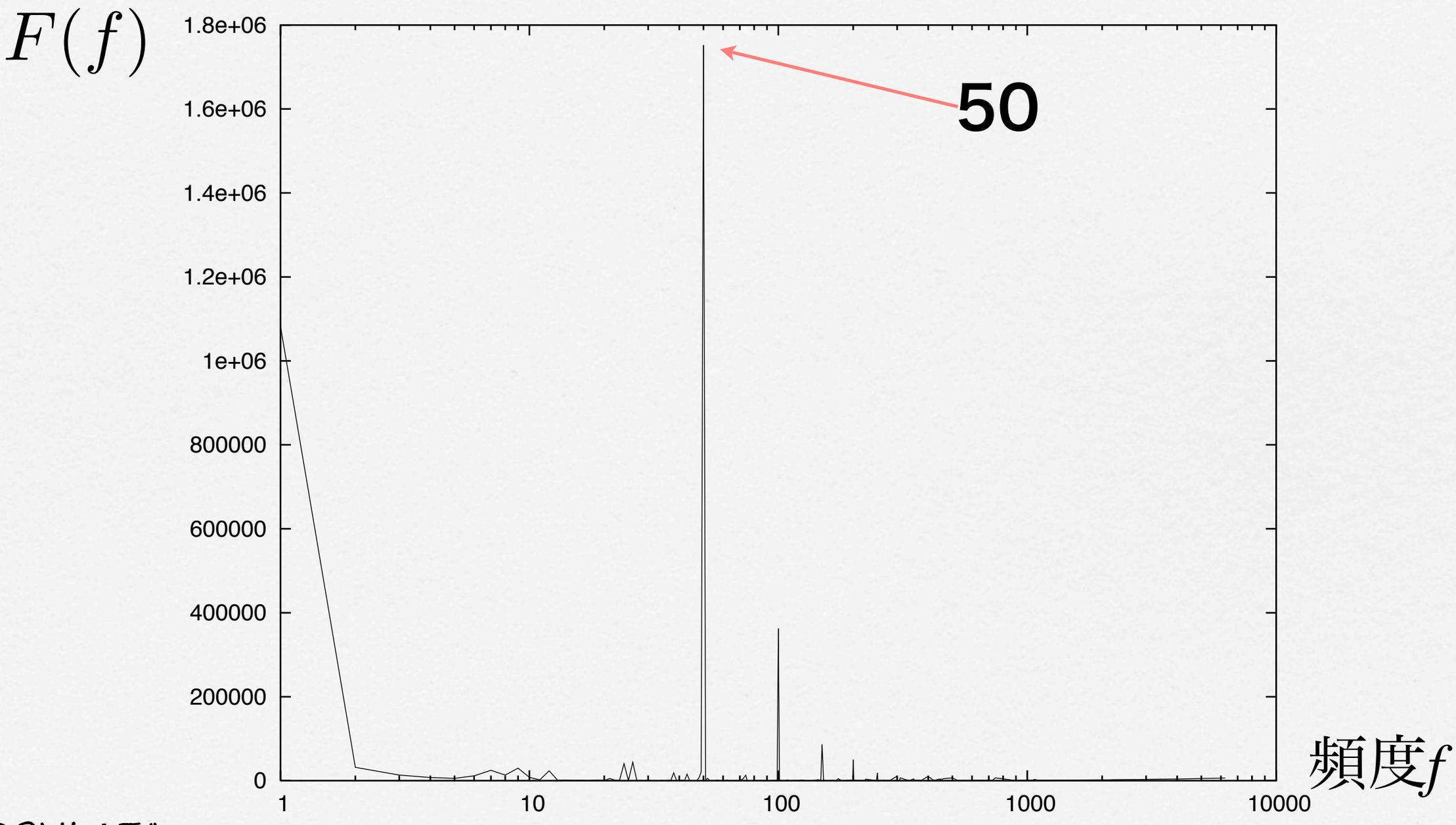
(ファイル数50)



FIT2GNL154

f vs. $F(f)$

$$F(f) = fV(f)$$



F172GNL154

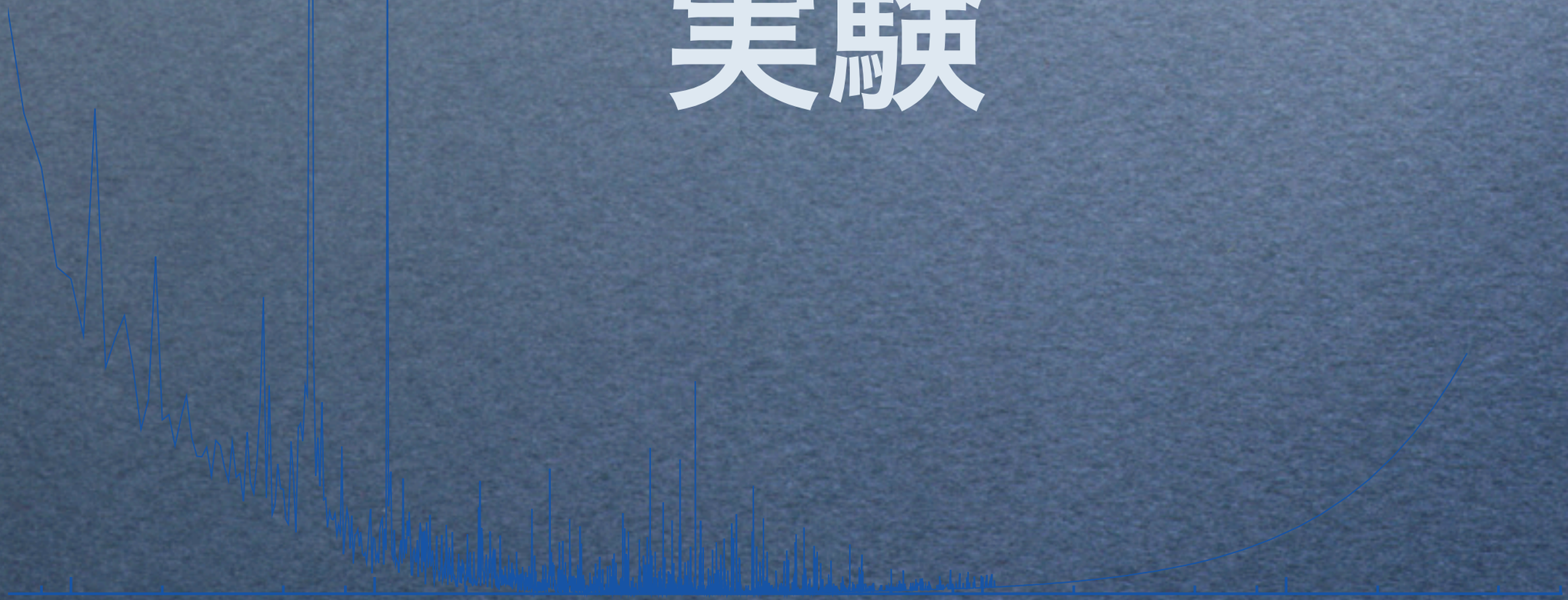
テンプレート発見問題

- 入力：文字列の有限集合S
  - Sは未知の正規パタンpから自然な確率分布に従って生成された
- 問題：以下を満すテンプレートt
  - tは正規パタンpのテンプレートである
  - $S \subseteq L(p)$ かつ $\nexists q; L(q) \subseteq L(p)$

アルゴリズム

- 任意の長さの部分文字列を数える
  - 実際には10〜30程度の長さまで
- すべての頻度fに対し、ちょうどf回出現する文字列の数$V(f)$を計算
- fを横軸(対数)、$F(f)$を縦軸にプロット

実験

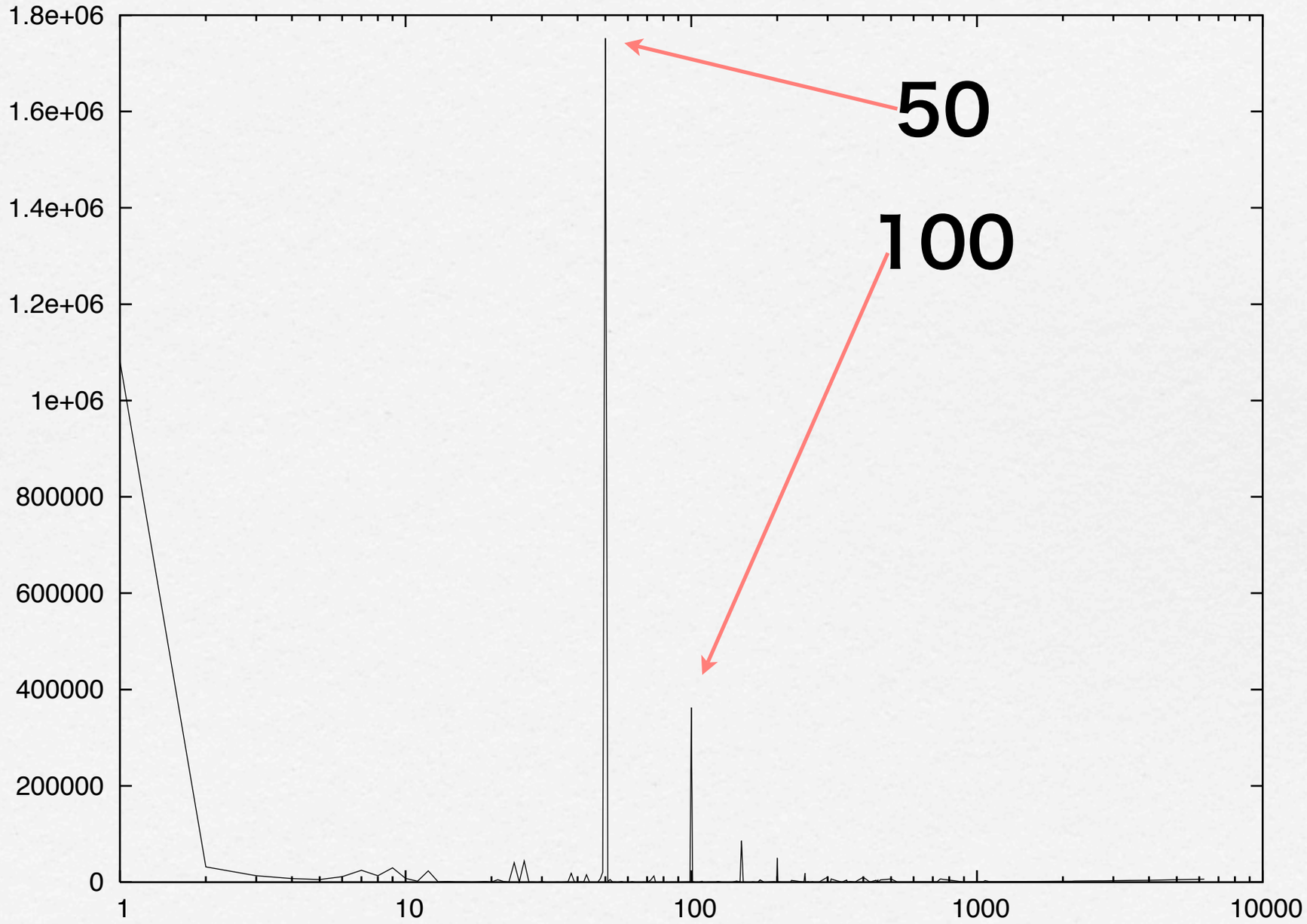


F172GNL154

実験データ

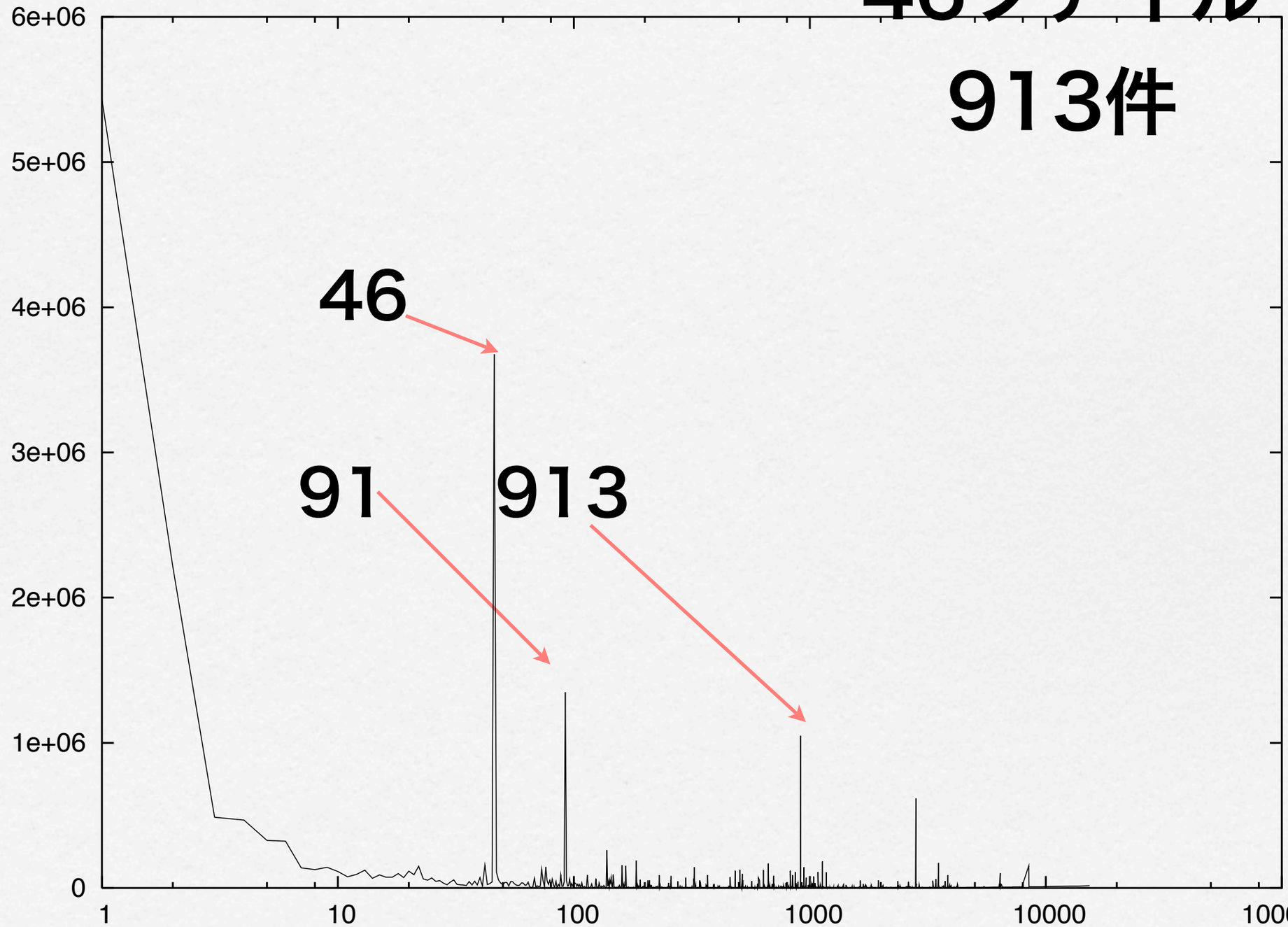
- (複数の)HTMLファイル
  - 前処理は行なわない
- 3種類のデータセット
  - 単一サイトのファイル
  - 単一サイトの周期の異なるテンプレート
  - 複数サイトのファイル

産経新聞の記事ファイル



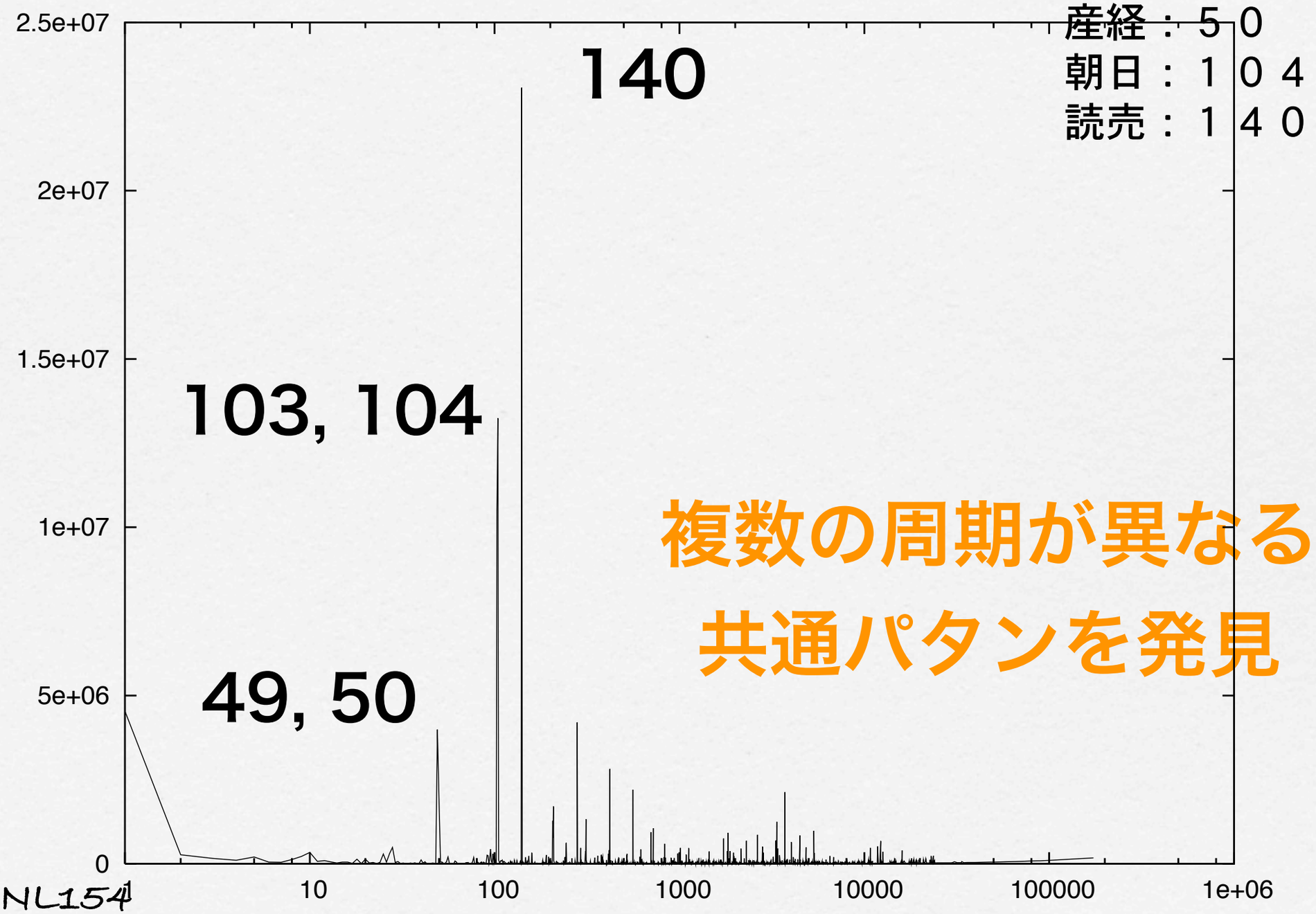
F172GNL154

Yahoo!検索結果



F172GNL154

3サイトのニュース記事



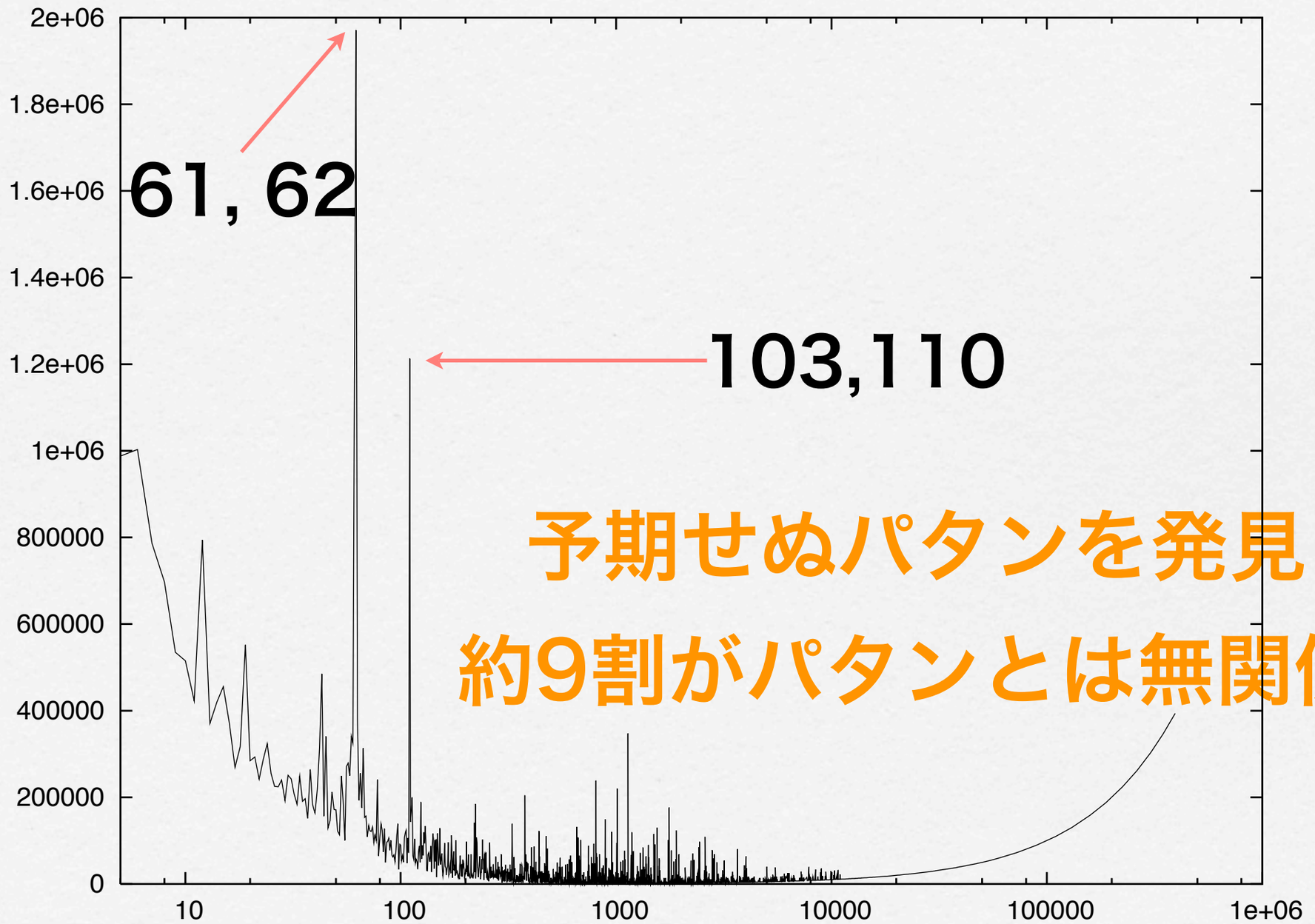
複数の周期が異なる
共通パタンを発見

九大サイト内

九大トップページから深さ3まで
リンクをたどり598ファイルを収集

http://www.kyushu-u.ac.jp/

九大サイト内(Cont.)



予期せぬパタンを発見

約9割がパタンとは無関係

F172GNL154

まとめ

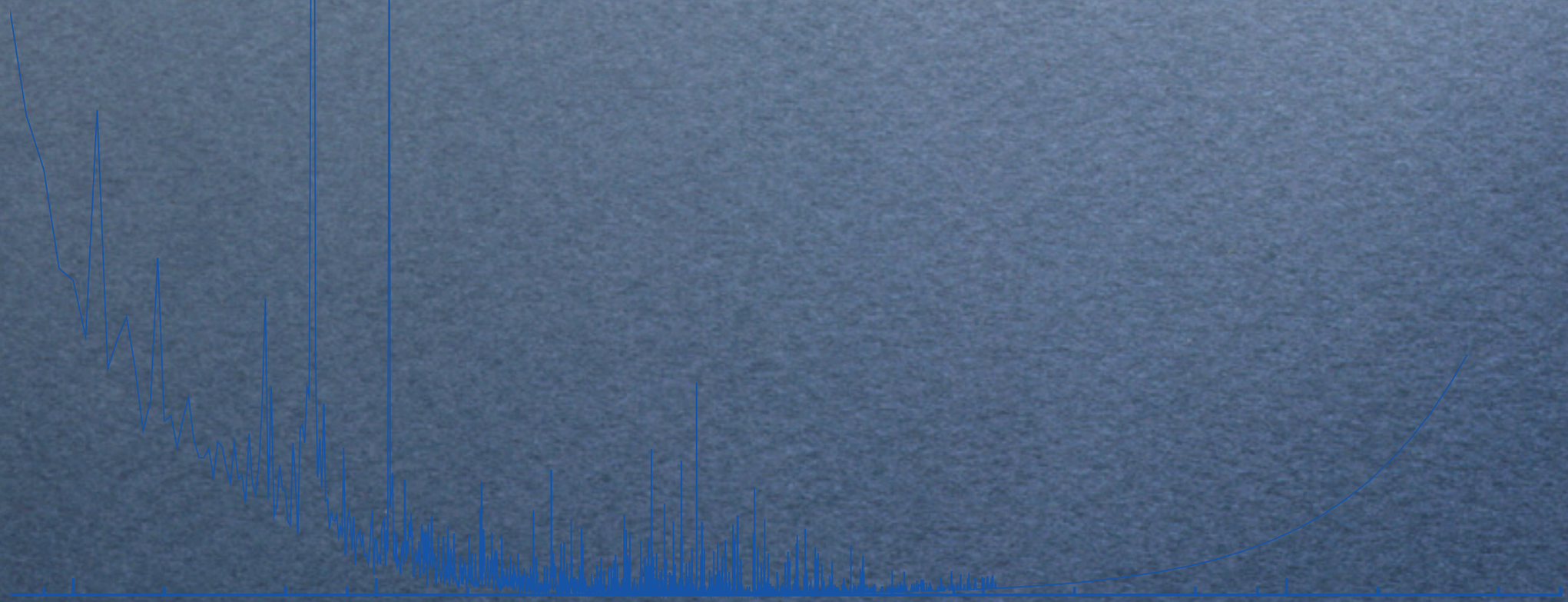
F172GNL154

まとめ

- テンプレート発見問題の定式化
- 共通パタン発見手法の提案
  - 正例のみで動く
  - 入力に対する前提や背景知識が不要
- HTMLファイルによる実験
  - 複数テンプレート
    - 周期の異なるテンプレートの混在も可能
  - ノイズに対し頑健

今後の課題

- パタン抽出の自動化
  - 複数の候補出力や対話的操作
- 「パタン」の拡張
  - 近似文字列や異周期テンプレートを包含
    - 後者は実験的にうまくいくことが示されている
- 精度の理論的な評価



F172GNL154

機械学習

- パタン言語の学習
  - Angluin (1980)
  - 変数の出現を制限する
  - → 共通部分よりも変数の対応
  - 負例(背景)をつかうものもあり

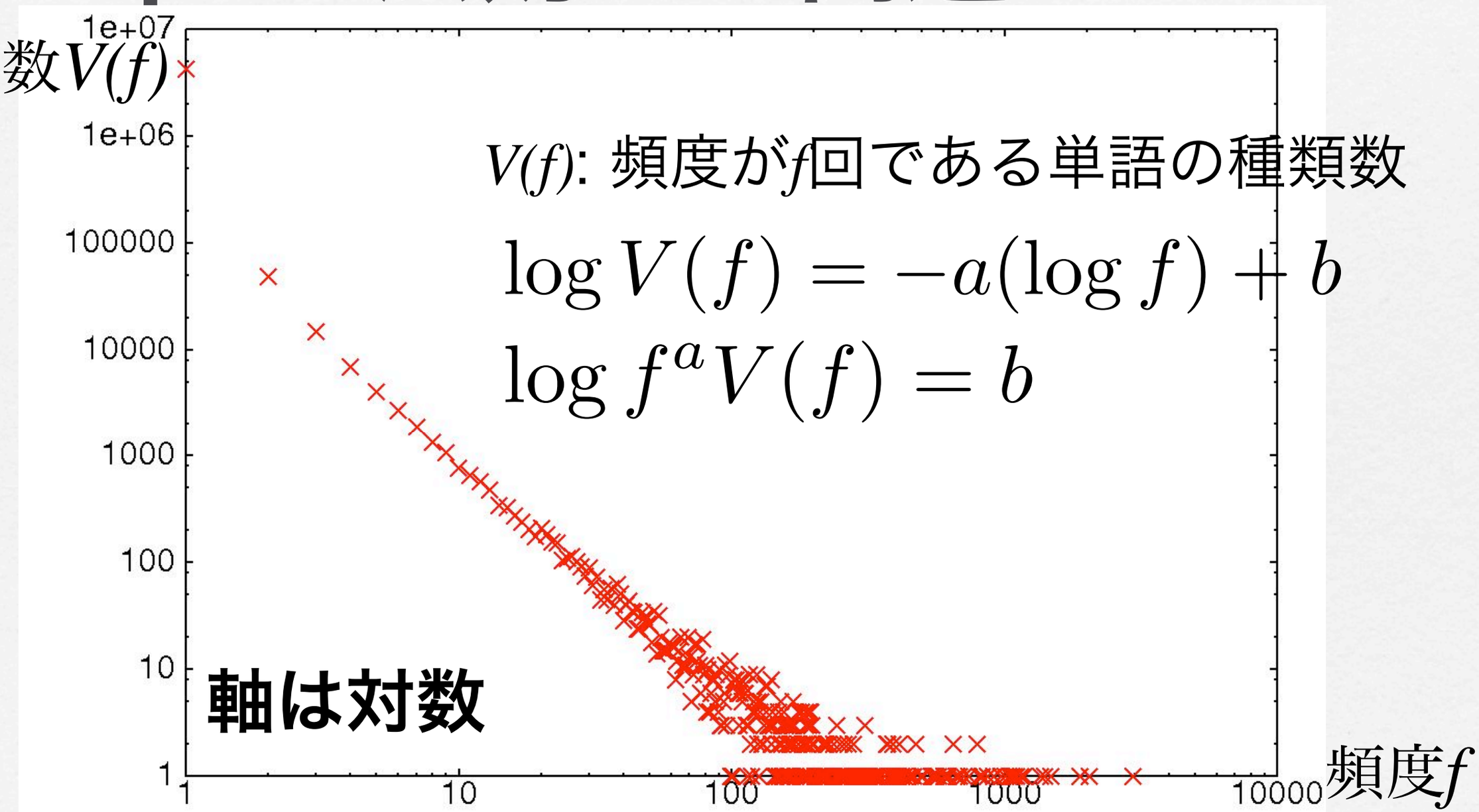
文字列処理

"共通問題"は部分列

- 最長共通部分列問題
  - 入力：(複数の)文字列
  - 出力：入力の共通部分列で最長のもの

NP完全(Maier 1978)

Zipfの法則との関連



九州大学__病院__

九州大学__病院__

Mail to Web master | サイトマップ

九州大学
Kyushu University

キーワードを入れてください 検索

日本語 English

学部・大学院 faculty graduate school

教育・学生生活 campus life

研究 research

社会との連携 society

行事・イベント
入試情報
国際交流・留学
教官情報
図書館
博物館
病院
同窓会・後援会
改革・評価
新キャンパス
広報誌
教職員向け情報
マップ

ホーム > 病院

病　院

トップページに戻る

- 九州大学の病院
  - 医学部附属病院
  - 歯学部附属病院
  - 生体防御医学研究所附属病院

ページのトップへ

トップページに戻る

FI72GNL154