Pattern Matching versus Covid

Philippe Jacquet (IEEE Fellow)
Inria, France
Philippe.jacquet@inria.fr
Definition

• Consider a finite alphabet $A$.
  • Eg $A$={$a,b$}
  • $A^*$ is the set of finite texts written with symbols in $A$.
    • $A^*$ = {$\varepsilon$,”$a$”,”$b$”,”$aa$”,...}

• A text $w$ is factor of a text $X$ if $X = uvw$ with $u,v \in A^*$
  • $w$ = “$aba$”, $X$ = “$abbaabaa$”, $X$ = “$abaabb$”, $X$ = “$ababa$”

• Pattern matching problem:
  • Given $w$ and $X$, determine if $w$ is factor of $X$.

• Application
  • Parsing, compilation, search, edition. compression, etc.
Probabilistic model on sequence

• The simplest: memoryless source, $P(X) =$product of symbol probabilities
  • $|X|$ length of sequence X.
  • Number of occurrences of w in X: $|X|_w$

• $E[|X|_w] = (|X| - |w|)P(w)$.

• When $|w| = o(\sqrt{|X|})$ then $E[|X|_w] \sim |X|P(w)$

• Straightforward extension for Markov sequence generation with finite memory
Word sequence matching problem

• Let \((w_1, w_2, ..., w_k)\) be a sequence of different words and X be a text.
  • We want the \(w_i\) to be factors of X and appear in this order in X.
  • Let \(|X|_{(w_1, ..., w_k)}\) be number of such occurrences in X
    • \(E[|X|_{(w_1, ..., w_k)}] = \left(|X| - \sum_{k} |w_k| \right) \cdot P(w_1) ... P(w_k)\)
      • When \(|w_1| + ... + |w_k| = o(\sqrt{|X|})\), then \(E[|X|_{(w_1, ..., w_k)}] \sim \frac{|X|^k}{k!} \cdot P(w_1) ... P(w_k)\)

• If the order is indifferent
  • \(E[|X|_{\{w_1, ..., w_k\}}] \sim |X|^k P(w_1) ... P(w_k)\)
The suffix/prefix trees

• useful data structures for linear pattern matching.
  • All suffixes (resp prefixes) are stored in a tree structure
  • Longest copy of \( w \) found by exploration of the tree of \( X \)

• Computation of the suffix tree:
  • Naïve algorithm in \( |X| \log |X| \)
  • Rissanen algorithm in \( |X| \)
  • Search in tree is in \( \log |X| \)
Joint Complexity

• Definition: number of common factors between two texts
  • $\varepsilon$, a, e, i, n, s, t, _, a_, an, is, na, s_, ti, _a, _e, _i, ana, is_, nan, _is, anan, nana, _is_, anana
  • $J(X1,X2)=25$
  • Fast to compute: linear via suffix trees superposition.

• Real time classification and trend propagation tracking
  • Agnostic of language
  • Resilient to mispealing

X1=The ape is eating a banana
X2=Ananas is an exotic fruit
Theoretical performance

- When texts X and Y are of same length but from sources with different parameters

  \[ E[J(X, Y)] \sim \frac{|X|^\kappa}{\sqrt{a \log |X| + b}} \text{ with } \kappa < 1 \]

- When texts are of different lengths but from sources with same parameters (e.g., Markov with same transition matrix)

  \[ E[J(X, Y)] \sim \frac{(|X| + |Y|) \log(|X| + |Y|) - |X| \log |X| - |Y| \log |Y|}{h} \]


Our evil guest star:

AACAACCAACCAAACCTTTGGATCCATCTTTGAGATGTGGTTTCTCTCTAAAACGAAAACCTTAAAATCTGTGTGGGGTCTGCTGACTGCGGCTGACATTAGTGGCCTC

ACGATATAAAATCGCACTTCTATTAGAAAAGATGTGACCTTCAAAATTATGGTGATAGTGCAACATTACCTAAAGGCATAATGATGAGTCTCGCAAAATATACTCAACTGTGTCAATATTTAAACACATTAACATTAGCTGTACC

GTAAAAACACAGTTCAATTATTATAAGAAAGTTGATGGTGTTGTCCAACAATTACCTGAAACTTACTTTACTCAGAGTAGAAATTTACAAGAATTTAAACCCAGGAGTCAAATGGAAATTGATTT

CTAACCTTAACTTGCCTGGTTGTGATGGTGGCAGTTTGTATGTAAATAAACATGCATTCCACACACCAGCTTTTGATAAAAGTGCTTTTGTTAATTTAAAACAATTACCATTTTTCTATTACTCT

GCCTCAAAGATTTTGGGACTACCAACTCAAACTGTTGATTCATCACAGGGCTCAGAATATGACTATGTCATATTCACTCAAACCACTGAAACAGCTCACTCTTGTAATGTAAACAGATTTAATGT

TGCAGCAGAAACGCTCAAAGCTACTGAGGAGACATTTAAACTGTCTTATGGTATTGCTACTGTACGTGAAGTGCTGTCTGACAGAGAATTACATCTTTCATGGGAAGTTGGTAAACCTAGACCA

CTGCTTCTGGTAATCTATTACTAGATAAACGCACTACGTGCTTTTCAGTAGCTGCACTTACTAACAATGTTGCTTTTCAAACCTGTCAAACCCGGTAATTTTAACAAAGACTTCTATGACTTTGCT

TTGAAATGGGCTAGATTCCCTAAGAGTGATGGAACTGGTACTATCTATACAGAACTGGAACCACCTTGTAGGTTTGTTACAGACACACCTAAAGGTCCTAAAGTGAAGTATTTATACTTTATTAA

GCTACTTTAGACTGACTCTTGGTGTTTATGATTACTTAGTTTCTACACAGGAGTTTAGATATATGAATTCACAGGGACTACTCCCACCCAAGAAAAGCATAGATGCCTTCAAACTCAACATTAAA

ATAGCATCTACAGATACTTGTTTTGCTAACAAACATGCTGATTTTGACACATGGTTTAGCCAGCGTGGTGGTAGTTATACTAATGACAAAGCTTGCCCATTGATTGCTGCAGTCATAACAAGAGA

AGAGCTAATAACACTAAAGGTTCATTGCCTATTAATGTTATAGTTTTTGATGGTAAATCAAAATGTGAAGAATCATCTGCAAAATCAGCGTCTGTTTACTACAGTCAGCTTATGTGTCAACCTAT

TCCCTTGGGATACTATAGCTAATTATGCTAAGCCTTTTCTTAACAAAGTTGTTAGTACAACTACTAACATAGTTACACGGTGTTTAAACCGTGTTTGTACTAATTATATGCCTTATTTCTTTACT

GATTTAAACCAGTTAACTGGTTATAAGAAACCTGCTTCAAGAGAGCTTAAAGTTACATTTTTCCCTGACTTAAATGGTGATGTGGTGGCTATTGATTATAAACACTACACACCCTCTTTTAAGAA

AACCTGAAGAAGAGCAAGAAGAAGATTGGTTAGATGATGATAGTCAACAAACTGTTGGTCAACAAGACGGCAGTGAGGACAATCAGACAACTACTATTCAAACAATTGTTGAGGTTCAACCTCAA

AAGCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATGCAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCAGAC
Coronavirus SARS-2 sequence
Joint Complexity of SARS-2 versus Bat-α Coronavirus (2007)

Cumulated Joint Complexity With SARS-2

Dashed: theoretical with pure random sequences

Base index, parsed from right to left
SARS-2 versus Bat and HIV

Normalized Joint Complexity With SARS-2

Base index parsed from right to left
“Accidental” Insertions of HIV in SARS-2 genome

- Despite not related, there are 19 short common factors of HIV sequence in SARS-2 genome
  - From 19 factors of 20± bases each
  - 20th is the bat-α virus

Maximum joint complexity discrepancy in standard deviation unit
Matching probabilities

• If the factors are fixed, then the probability of such accidental insertions will be smaller than $30,000^{19} 4^{-19 \times 20} \approx 2.10^{-144}$!
  
  • This is the word sequence problem

• Perez-Montagnier conclusion: the SARS-2 Cov virus has been “forged”.
Weak pattern matching

• Let unfix the matching factors

• Assume a corpus of M bases (probably around $10^6$) about HIV sequences (simian, HIV-1, HIV-2 and reversed sequences)
  • Average number of matchings of the corpus over a genome of N bases is $MN \times P($matching$)$.
  • $P(k \text{ matchings}) < \frac{MN \times 4^{-20}}{k} \approx 1.4 \times 10^{-3}$ for $k=19$ and $M = 10^6$, $N = 30,000$ and $P($matching$) = 4^{-20}$.
    • Tolerance for max 3 errors multiplies by $\binom{20}{3}$:
  • $P(19 \text{ matchings}) < 1.6$
Weak matching: The matching between the matchers

• The segment HIV-2-UCI1 (segment 11) versus the other segments

• Similar accidents
  • Is it “forged” (it is from 1993)?
Strong pattern matchings

- RmYN02 “Closely Related to SARS-CoV-2” (Nov 2020)

Strong pattern matchings (continued)

Normalized Joint Complexity With SARS-2

Base index parsed from right to left
SARS-2 Covid closest known ancestor

• RaTG13 Bat Coronavirus
  • Found in 2013 in a cave in Yunnan
  • Six years after Bat-α, low matching.
  • Strong matching with SARS-2, six years later.
insertion in SARS-2 from RaTG13

- with largest common factor, JC can detect the offsets between sequences
  - Offset increment: insertion
  - Decrement: deletion

Offset detected
Via JC between SARS-2
And RaTG13

Genome of SARS-2 per 50bp slice
Thank you

Stay at home
and
Kill Coronavirus!

Safa