Fast Multiple String Matching Using Streaming SIMD Extensions Technology

Simone Faro

faro@dmi.unict.it

University of Catania, Department of Mathematics and Computer Science

M. Oğuzhan Külekci

oguzhan.kulekci@tubitak.gov.tr

TÜBİTAK – National Research Institute of Electronics and Cryptology

Prague Stringology Conference

Prague, Czech Republic, Sep. 1–4, 2013

Multiple Exact String Matching

- Text $T = t_0 t_1 t_2 \ldots t_{n-1}$, $|T| = n$
- Finite alphabet $\Sigma = \{\epsilon_0, \epsilon_1, \ldots, \epsilon_{\sigma-1}\}$, $|\Sigma| = \sigma$
- Pattern set $\mathcal{P} = \{P_0, P_1, P_2, \ldots, P_{r-1}\}$, $|P| = r$
- Length of the patterns $\{m_0, m_1, m_2, \ldots, m_{r-1}\}$, $m = m_0 + m_1 + \ldots + m_{r-1}$
- $P_j = p_j^0 p_j^1 p_j^2 \ldots p_j^{m_j - 1}$

Find all exact occurrences of patterns in $\mathcal{P}$ on text T.

$$\{\forall\ i, \exists\ j : 0 \leq i \leq n - m_j, 0 \leq j < r, t[i \ldots i + m_j - 1] = P_j\}$$

Motivation

Sample Applications

- Computational genomics, e.g., metagenomics
  Given the DNA extracted from a sample (text), determine patterns of interest (pattern set) occurring inside.

- Network intrusion detection and anti-virus software
  Determine if the set of predetermined harmful patterns (malicious code segments, viruses, etc...) exists in a file or in a communication.

- Can also be useful as a filter in approximate string matching.
  When k mismatches are allowed on a m–symbols long pattern, there exists an exact matching partition of length at least $z = \lfloor \frac{m}{k} \rfloor - 1$. Thus, compute all z–symbols long factors of pattern, search them on T ,and perform a verification on each detection.

Previous Studies

- Search patterns one-by-one over T by using a fast exact single pattern matching algorithm, e.g., KMP, BM, or others. $O(m + rn)$ worst case time complexity, where in many practical cases sub-linear performances are achieved on the average.
- Aho-Corasick automaton $O(m + n)$.
- Optimal average complexity is $O(n \log_\sigma(rm')/m')$ [Navarro&Fredriksson'04], and first achieved by Set-Backward-DAWG-Matching (SBDM) that builds an indexing structure for the reverse strings of $\mathcal{P}$. Notice that $m' = min\{m_0, m_1, \ldots, m_{r-1}\}$.

Previous Studies

SBDM implementations

- Bottleneck in SBDM is index construction time and space consumption.
- Partially solved by SBOM (set backward oracle matching) [Allauzen et al. '99].
- Enhanced implementation in ESBOM [Faro&Lecroq'09].

Hash-based solutions

- Use an index table for blocks of q–characters.
- WM algorithm by Wu & Manber'94.

Previous Studies

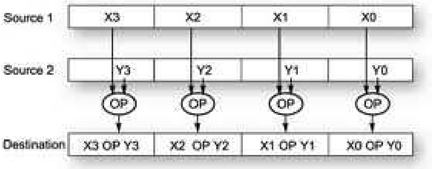
Benefiting from bit-parallelism

- Bit-parallelism [Baeza-Yates&Gonnet'92]: Using computers intrinsic bit operations may cut down the number processes by a factor up to computer word size w.
- Extension of bit-parallel algorithms to multi pattern case.
- Such an efficient solution has been proposed by MBNDM [Rivals et al.'09]: Compute a superimposed pattern among the set, perform filtering via BNDM.

Previous Studies

Benefiting from single-instruction multiple data (SIMD) instructions

- Using recent SIMD technology to speed up operations.



Intel SIMD Technology

- MMX (64-bit registers)
- SSE2,SSE3,SSE3e (128-bit registers)
- SSE4.1, SSE4.2 (special text processing instructions)
- AVX, AVX2 (256-bit registers)

Previous Studies

Originally, SIMD was aimed to help vector operations in image/video processing. However, it helps a lot in practical packed string matching also.

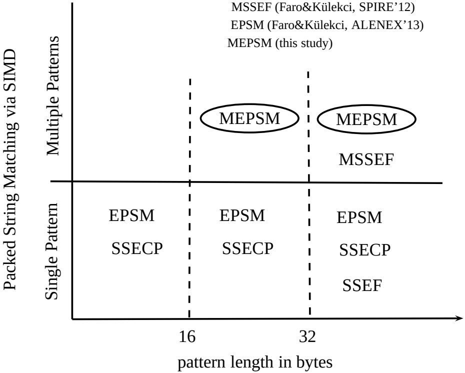
Theoretical work

- Fredriksson, SPIRE'02
  Speeding up pattern matching with super-alphabets
- Bille,CPM'09
  Fast searching in packed strings
- Belazzougui&Raffinot, CIAC'13
  Average Optimal String Matching on Packed Strings
- Ben-Kiki et al.,
  `www.cs.haifa.ac.il/oren/Publications/bpsm.pdf`
  Towards optimal packed string matching

Previous Studies

Packed String Matching Algorithms with Implementations

SSEF (Külekci, PSC'09)
SSECP (Ben-kiki et al., FSTTCS'11)
MSSEF (Faro&Külekci, SPIRE'12)
EPSM (Faro&Külekci, ALENEX'13)
MEPSM (this study)



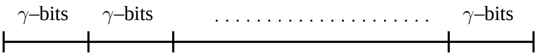
This study presents an efficient solution (MEPSM) on packed multiple-string matching for patterns longer than 16–bytes.

Methodology

Packed string representation

$\gamma = \lceil \log \sigma \rceil$ bits required per symbol

ω is the computer word size



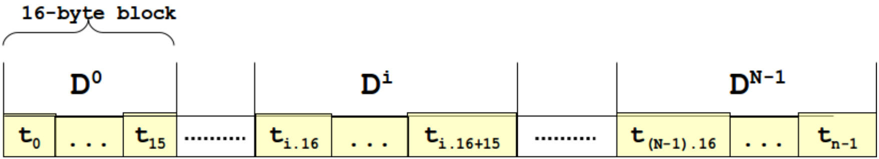
$\alpha = \lfloor \frac{\omega}{\gamma} \rfloor$ symbols per computer word

In practice,

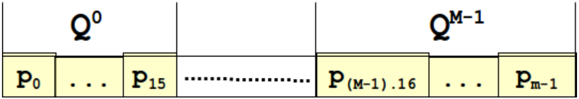
$\gamma = 8$, $w = 128$ (via special SSE registers), and $\alpha = 16$.

Packed Representations

Text



Pattern

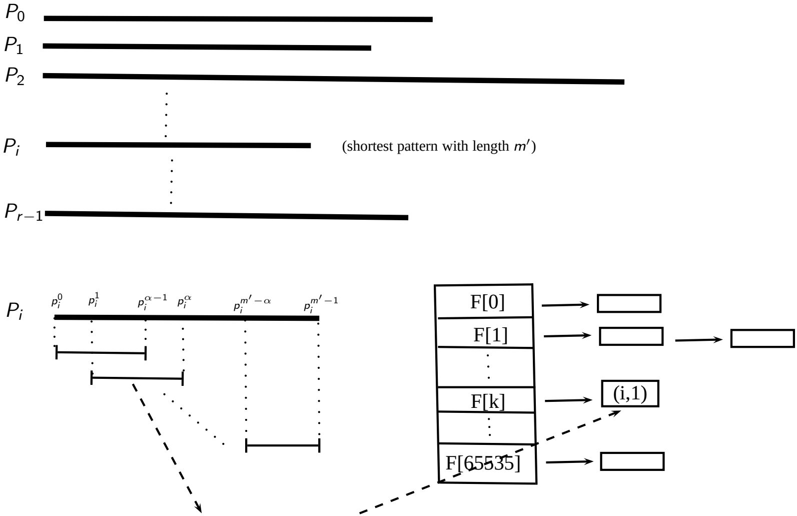


The SSE Instruction Used in The Algorithm

- We use the SSE instruction *packed cyclic redundancy check*.
  - `crc32 = _mm_crc32_u16(acc,B),` B `is 16--bits long`
  - `crc32 = _mm_crc32_u32(acc,B),` B `is 32--bits long`
  - `crc32 = _mm_crc32_u64(acc,B),` B `is 64--bits long`

- The CRC values are all 32-bit, and we use their lower 16–bits as the fingerprint of B input values.

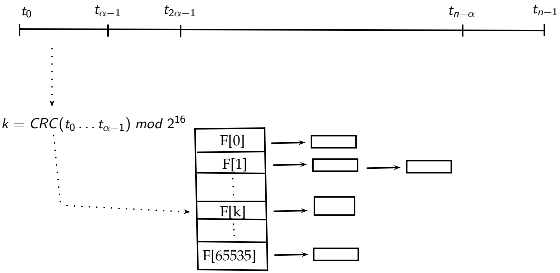
The Preprocessing

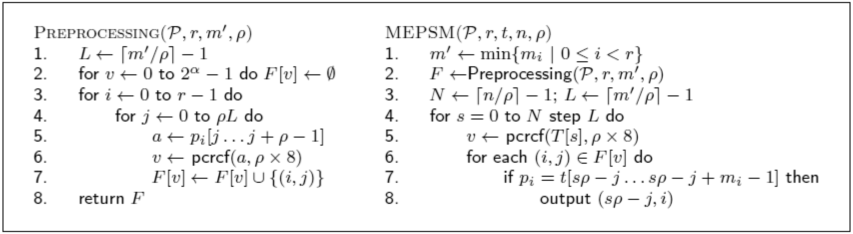


(shortest pattern with length m')

$$k = CRC(p_i^0 \dots p_i^{\alpha})\ mod\ \ 2^{16}$$

The Searching Phase





```
PREPROCESSING$(\mathcal{P}, r, m', \rho)$
    $L \leftarrow \lceil m'/\rho \rceil - 1$
    for $v \leftarrow 0$ to $2^\alpha - 1$ do $F[v] \leftarrow \emptyset$
    for $i \leftarrow 0$ to $r - 1$ do
        for $j \leftarrow 0$ to $\rho L$ do
            $a \leftarrow p_i[j \dots j + \rho - 1]$
            $v \leftarrow$ pcrcf$(a, \rho \times 8)$
            $F[v] \leftarrow F[v] \cup \{(i, j)\}$
    return $F$
```

```
MEPSM$(\mathcal{P}, r, t, n, \rho)$
    $m' \leftarrow \min\{m_i \mid 0 \leq i < r\}$
    $F \leftarrow$Preprocessing$(\mathcal{P}, r, m', \rho)$
    $N \leftarrow \lceil n/\rho \rceil - 1$; $L \leftarrow \lceil m'/\rho \rceil - 1$
    for $s = 0$ to $N$ step $L$ do
        $v \leftarrow$ pcrcf$(T[s], \rho \times 8)$
        for each $(i, j) \in F[v]$ do
            if $p_i = t[s\rho - j \dots s\rho - j + m_i - 1]$ then
                output $(s\rho - j, i)$
```

Complexity Analysis

Preprocessing

- Space: $O(rm' + 2^{\alpha})$, since rm' alignments inserted into a 2^{α}-row table.
- Time: $O(m)$ (assuming all patterns have same length m').

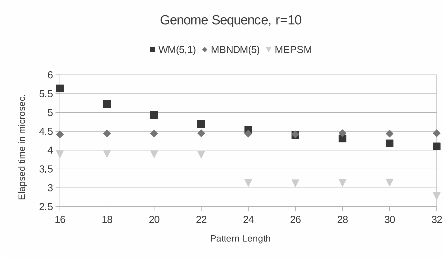
Scanning Phase

- Best case: $O(\frac{n}{\alpha})$
- Worst case: $O(nm)$ (assuming each fingerprint matches with all patterns, and requires verification of whole set)

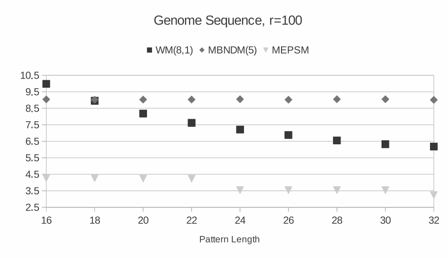
Experimental Results

- SMART platform is used.
- Most competitive multiple pattern matching algorithms are observed to be the WM and MBNDM.
- The q-gram enhanced implementations of both considered (q=3 to 8). For WM also distinct *h* values are considered.
- Experiments conducted on genome, protein sequences, and English text. The averages of 200 runs are reported.
- Pattern set size (r) tested are 10, 100, 1000, and 10000.
- Observed that proposed scheme is most efficient when pattern length varies in between 16 and 32.

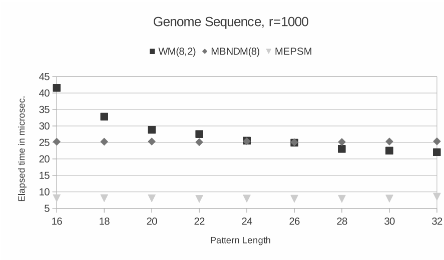
Results on Genome Sequences



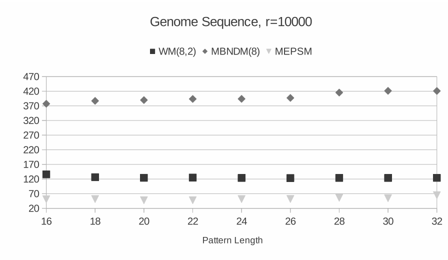
Genome Sequence, r=10



Genome Sequence, r=100

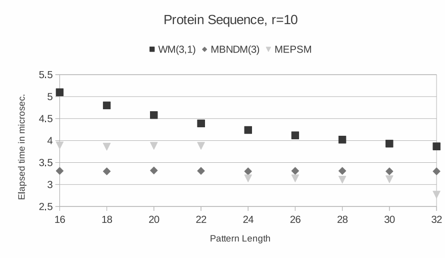


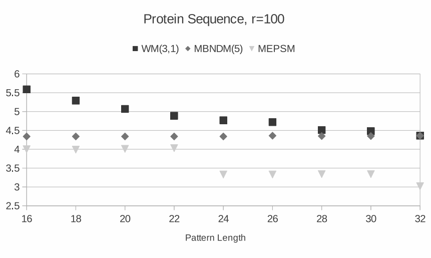
Genome Sequence, r=1000

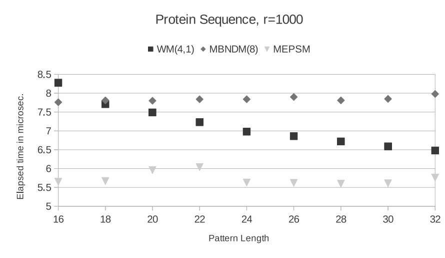


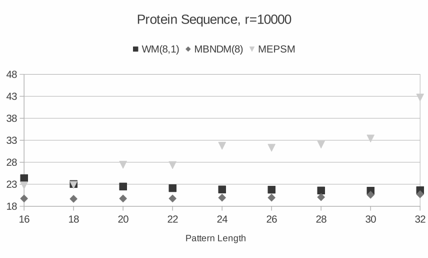
Genome Sequence, r=10000

Results on Protein Sequences

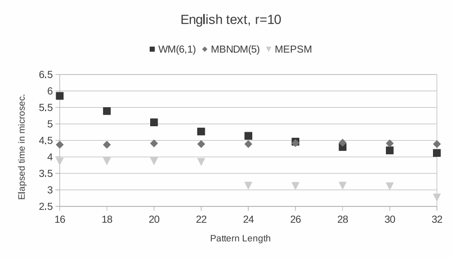


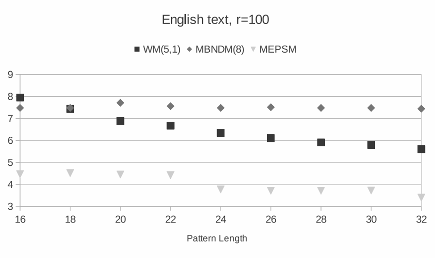


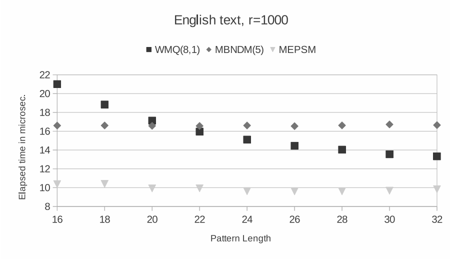


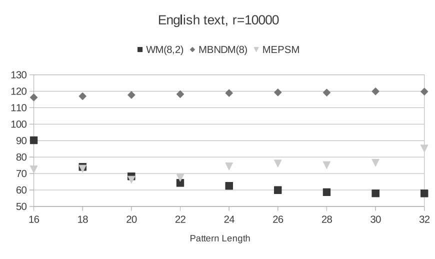


Results on English Text


English text, r=10


English text, r=100


English text, r=1000


English text, r=10000

Overall picture

(A) m	16	20	24	28	32
genome	1.13	1.13	1.41	1.37	1.47
protein	0.85	0.85	1.05	1.06	1.19
nat.lang.	1.13	1.13	1.40	1.37	1.48

(B) m	16	20	24	28	32
genome	2.11	1.92	2.03	1.85	1.90
protein	1.08	1.08	1.30	1.30	1.44
nat.lang.	1.67	1.57	1.66	1.56	1.64

(C) m	16	20	24	28	32
genome	3.12	3.15	3.18	2.91	2.58
protein	1.37	1.37	1.23	1.16	1.07
nat.lang.	1.60	1.67	1.57	1.36	1.24

(D) m	16	20	24	28	32
genome	2.34	2.51	2.33	2.15	1.86
protein	0.86	0.72	0.62	0.62	0.48
nat.lang.	1.24	1.03	0.84	0.78	0.67

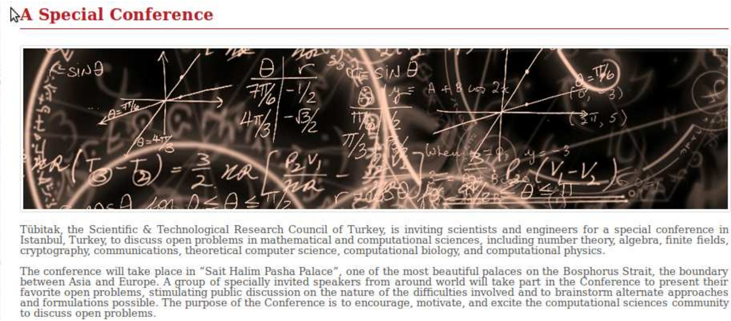
The speed ups obtained via MEPSM compared with the second best results on sets of 10 (A), 100 (B), 1.000 (C) and 10.000 (D) patterns.

Conclusions

- Multiple packed string matching efficient on 16-32 byte long patterns. Work still in progress to cover shorter patterns.
- Scales quite well on increasing size of the pattern set. Alternatives become more advantageous on longer pattern lengths (due to improvement of shift values in that case)
- Performance on new AVX2 instructions set ???
- SIMD instructions make a good job in practice. However, latency/throughput values of instructions should be considered during the design.
- SIMD instructions might be considered in general for algorithm engineering.

Thank you!

- www.opmcs.org
- Free registration!
- Lectures by world renowned 20 invited speakers!
- Spectacular venue at Istanbul Bosphorus.



A Special Conference

Tübitak, the Scientific & Technological Research Council of Turkey, is inviting scientists and engineers for a special conference in Istanbul, Turkey, to discuss open problems in mathematical and computational sciences, including number theory, algebra, finite fields, cryptography, communications, theoretical computer science, computational biology, and computational physics.

The conference will take place in "Sait Halim Pasha Palace", one of the most beautiful palaces on the Bosphorus Strait, the boundary between Asia and Europe. A group of specially invited speakers from around world will take part in the Conference to present their favorite open problems, stimulating public discussion on the nature of the difficulties involved and to brainstorm alternate approaches and formulations possible. The purpose of the Conference is to encourage, motivate, and excite the computational sciences community to discuss open problems.

Let's meet where the continents meet.

Pattern Set Size r=10

(Genome) m	16	18	20	22	24	26	28	30	32
WM(5,1)	5.64 0.43	5.22 0.42	4.94 0.44	4.70 0.42	4.54 0.43	4.40 0.43	4.31 0.44	4.18 0.43	4.10 0.44
MBNDM(5)	4.42 0.17	4.44 0.17	4.44 0.17	4.45 0.17	4.44 0.17	4.42 0.17	4.45 0.17	4.44 0.17	4.45 0.17
MEPSM	**3.89** 0.01	**3.90** 0.01	**3.89** 0.01	**3.88** 0.01	**3.13** 0.01	**3.12** 0.01	**3.13** 0.01	**3.14** 0.01	**2.78** 0.01

Pattern Set Size r=10

(Genome) *m*	16	18	20	22	24	26	28	30	32
WM(5,1)	5.64 0.43	5.22 0.42	4.94 0.44	4.70 0.42	4.54 0.43	4.40 0.43	4.31 0.44	4.18 0.43	4.10 0.44
MBNDM(5)	4.42 0.17	4.44 0.17	4.44 0.17	4.45 0.17	4.44 0.17	4.42 0.17	4.45 0.17	4.44 0.17	4.45 0.17
MEPSM	**3.89** 0.01	**3.90** 0.01	**3.89** 0.01	**3.88** 0.01	**3.13** 0.01	**3.12** 0.01	**3.13** 0.01	**3.14** 0.01	**2.78** 0.01
(Protein)									
WM(3,1)	5.10 0.44	4.80 0.43	4.58 0.44	4.39 0.43	4.24 0.43	4.12 0.44	4.02 0.43	3.93 0.43	3.87 0.43
WM(6,1)	5.53 0.43	5.10 0.43	4.75 0.44	4.51 0.43	4.33 0.43	4.18 0.43	4.07 0.43	3.93 0.42	3.84 0.42
MBNDM(3)	**3.31** 0.17	**3.30** 0.17	**3.32** 0.17	**3.31** 0.17	3.30 0.17	3.31 0.17	3.31 0.17	3.30 0.17	3.30 0.17
MEPSM	3.89 0.01	3.86 0.01	3.88 0.01	3.88 0.01	**3.14** 0.01	**3.14** 0.01	**3.11** 0.01	**3.12** 0.01	**2.77** 0.01

Pattern Set Size r=10

(Genome) *m*	16	18	20	22	24	26	28	30	32
WM(5,1)	5.64 0.43	5.22 0.42	4.94 0.44	4.70 0.42	4.54 0.43	4.40 0.43	4.31 0.44	4.18 0.43	4.10 0.44
MBNDM(5)	4.42 0.17	4.44 0.17	4.44 0.17	4.45 0.17	4.44 0.17	4.42 0.17	4.45 0.17	4.44 0.17	4.45 0.17
MEPSM	**3.89** 0.01	**3.90** 0.01	**3.89** 0.01	**3.88** 0.01	**3.13** 0.01	**3.12** 0.01	**3.13** 0.01	**3.14** 0.01	**2.78** 0.01
(Protein)									
WM(3,1)	5.10 0.44	4.80 0.43	4.58 0.44	4.39 0.43	4.24 0.43	4.12 0.44	4.02 0.43	3.93 0.43	3.87 0.43
WM(6,1)	5.53 0.43	5.10 0.43	4.75 0.44	4.51 0.43	4.33 0.43	4.18 0.43	4.07 0.43	3.93 0.42	3.84 0.42
MBNDM(3)	**3.31** 0.17	**3.30** 0.17	**3.32** 0.17	**3.31** 0.17	3.30 0.17	3.31 0.17	3.31 0.17	3.30 0.17	3.30 0.17
MEPSM	3.89 0.01	3.86 0.01	3.88 0.01	3.88 0.01	**3.14** 0.01	**3.14** 0.01	**3.11** 0.01	**3.12** 0.01	**2.77** 0.01
(English)									
WM(5,1)	5.91 0.42	5.47 0.43	5.14 0.43	4.88 0.43	4.71 0.43	4.54 0.43	4.40 0.43	4.29 0.43	4.20 0.44
WM(6,1)	5.85 0.43	5.39 0.43	5.05 0.43	4.77 0.42	4.64 0.42	4.46 0.43	4.30 0.43	4.20 0.42	4.12 0.44
MBNDM(5)	4.37 0.17	4.37 0.17	4.41 0.17	4.39 0.17	4.39 0.17	4.42 0.17	4.43 0.17	4.41 0.17	4.39 0.17
MEPSM	**3.86** 0.01	**3.87** 0.01	**3.87** 0.01	**3.85** 0.01	**3.13** 0.01	**3.12** 0.01	**3.13** 0.01	**3.11** 0.01	**2.77** 0.01

Pattern Set Size r=100

Genome *m*	16	18	20	22	24	26	28	30	32
WM(5,1)	9.78 0.44	9.38 0.44	8.96 0.44	8.73 0.44	8.60 0.44	8.42 0.44	8.22 0.44	8.10 0.43	8.07 0.44
WM(8,1)	9.98 0.44	8.96 0.44	8.18 0.44	7.62 0.44	7.21 0.44	6.88 0.45	6.55 0.44	6.32 0.44	6.18 0.45
MBNDM(5)	9.04 0.21	9.02 0.21	9.03 0.22	9.03 0.21	9.05 0.21	9.02 0.21	9.05 0.21	9.05 0.21	9.01 0.21
MEPSM	**4.27** 0.04	**4.26** 0.04	**4.24** 0.04	**4.23** 0.04	**3.54** 0.08	**3.54** 0.08	**3.54** 0.08	**3.54** 0.08	**3.24** 0.12

Pattern Set Size r=100

Genome *m*	16	18	20	22	24	26	28	30	32
WM(5,1)	9.78 0.44	9.38 0.44	8.96 0.44	8.73 0.44	8.60 0.44	8.42 0.44	8.22 0.44	8.10 0.43	8.07 0.44
WM(8,1)	9.98 0.44	8.96 0.44	8.18 0.44	7.62 0.44	7.21 0.44	6.88 0.45	6.55 0.44	6.32 0.44	6.18 0.45
MBNDM(5)	9.04 0.21	9.02 0.21	9.03 0.22	9.03 0.21	9.05 0.21	9.02 0.21	9.05 0.21	9.05 0.21	9.01 0.21
MEPSM	**4.27** 0.04	**4.26** 0.04	**4.24** 0.04	**4.23** 0.04	**3.54** 0.08	**3.54** 0.08	**3.54** 0.08	**3.54** 0.08	**3.24** 0.12
Protein									
WM(3,1)	5.59 0.43	5.29 0.43	5.07 0.43	4.89 0.43	4.77 0.45	4.72 0.43	4.51 0.43	4.48 0.44	4.36 0.43
WM(4,1)	6.31 0.43	5.93 0.43	5.66 0.43	5.36 0.44	5.06 0.44	5.09 0.45	4.72 0.44	4.64 0.44	4.50 0.44
MBNDM(5)	4.34 0.26	4.34 0.26	4.34 0.26	4.34 0.25	4.34 0.25	4.36 0.26	4.35 0.25	4.35 0.26	4.36 0.26
MEPSM	**4.00** 0.04	**3.99** 0.04	**4.01** 0.04	**4.03** 0.04	**3.33** 0.08	**3.33** 0.08	**3.34** 0.08	**3.34** 0.08	**3.02** 0.12

Pattern Set Size r=100

Genome *m*	16	18	20	22	24	26	28	30	32
WM(5,1)	9.78 0.44	9.38 0.44	8.96 0.44	8.73 0.44	8.60 0.44	8.42 0.44	8.22 0.44	8.10 0.43	8.07 0.44
WM(8,1)	9.98 0.44	8.96 0.44	8.18 0.44	7.62 0.44	7.21 0.44	6.88 0.45	6.55 0.44	6.32 0.44	6.18 0.45
MBNDM(5)	9.04 0.21	9.02 0.21	9.03 0.22	9.03 0.21	9.05 0.21	9.02 0.21	9.05 0.21	9.05 0.21	9.01 0.21
MEPSM	**4.27** 0.04	**4.26** 0.04	**4.24** 0.04	**4.23** 0.04	**3.54** 0.08	**3.54** 0.08	**3.54** 0.08	**3.54** 0.08	**3.24** 0.12
Protein									
WM(3,1)	5.59 0.43	5.29 0.43	5.07 0.43	4.89 0.43	4.77 0.45	4.72 0.43	4.51 0.43	4.48 0.44	4.36 0.43
WM(4,1)	6.31 0.43	5.93 0.43	5.66 0.43	5.36 0.44	5.06 0.44	5.09 0.45	4.72 0.44	4.64 0.44	4.50 0.44
MBNDM(5)	4.34 0.26	4.34 0.26	4.34 0.26	4.34 0.25	4.34 0.25	4.36 0.26	4.35 0.25	4.35 0.26	4.36 0.26
MEPSM	**4.00** 0.04	**3.99** 0.04	**4.01** 0.04	**4.03** 0.04	**3.33** 0.08	**3.33** 0.08	**3.34** 0.08	**3.34** 0.08	**3.02** 0.12
English									
WM(5,1)	7.95 0.42	7.44 0.44	6.88 0.42	6.67 0.44	6.34 0.43	6.10 0.44	5.91 0.44	5.80 0.44	5.60 0.44
WM(7,1)	8.37 0.39	7.58 0.41	7.02 0.41	6.66 0.40	6.26 0.40	6.00 0.40	5.80 0.41	5.63 0.40	5.47 0.40
MBNDM(5)	8.22 0.24	8.20 0.25	8.08 0.25	8.17 0.25	8.20 0.25	8.19 0.24	8.21 0.25	8.20 0.25	8.21 0.25
MBNDM(8)	7.48 0.28	7.48 0.28	7.71 0.29	7.56 0.28	7.48 0.28	7.51 0.28	7.48 0.27	7.48 0.27	7.44 0.28
MEPSM	**4.46** 0.04	**4.51** 0.04	**4.45** 0.04	**4.42** 0.04	**3.77** 0.08	**3.71** 0.08	**3.71** 0.08	**3.72** 0.08	**3.40** 0.12

Pattern Set Size r=1000

Genome *m*	16	18	20	22	24	26	28	30	32
WM(8,1)	41.44 0.52	38.26 0.52	37.24 0.54	36.00 0.56	35.08 0.57	34.24 0.59	32.79 0.58	32.42 0.60	32.09 0.60
WM(8,2)	41.55 0.64	32.83 0.69	28.83 0.71	27.56 0.73	25.55 0.73	24.93 0.77	23.02 0.75	22.52 0.78	22.05 0.80
MBNDM(8)	25.22 0.39	25.23 0.39	25.28 0.39	25.09 0.39	25.36 0.39	25.05 0.40	25.14 0.39	25.28 0.40	25.33 0.40
MEPSM	**8.08** 0.40	**8.09** 0.40	**8.05** 0.40	**7.87** 0.40	**7.96** 0.77	**7.92** 0.77	**7.89** 0.77	**7.96** 0.78	**8.53** 1.15

Pattern Set Size r=1000

Genome *m*	16	18	20	22	24	26	28	30	32
WM(8,1)	41.44 0.52	38.26 0.52	37.24 0.54	36.00 0.56	35.08 0.57	34.24 0.59	32.79 0.58	32.42 0.60	32.09 0.60
WM(8,2)	41.55 0.64	32.83 0.69	28.83 0.71	27.56 0.73	25.55 0.73	24.93 0.77	23.02 0.75	22.52 0.78	22.05 0.80
MBNDM(8)	25.22 0.39	25.23 0.39	25.28 0.39	25.09 0.39	25.36 0.39	25.05 0.40	25.14 0.39	25.28 0.40	25.33 0.40
MEPSM	**8.08** 0.40	**8.09** 0.40	**8.05** 0.40	**7.87** 0.40	**7.96** 0.77	**7.92** 0.77	**7.89** 0.77	**7.96** 0.78	**8.53** 1.15
Protein									
WM(4,1)	8.28 0.53	7.71 0.54	7.49 0.56	7.23 0.56	6.98 0.57	6.86 0.57	6.72 0.59	6.59 0.59	6.48 0.60
WM(8,1)	9.87 0.53	8.78 0.54	8.06 0.56	7.54 0.58	7.11 0.58	6.77 0.59	6.52 0.60	6.31 0.62	6.17 0.64
MBNDM(5)	8.47 0.37	8.52 0.39	8.50 0.38	8.51 0.38	8.57 0.38	8.58 0.39	8.51 0.38	8.49 0.38	8.64 0.39
MBNDM(8)	7.76 0.52	7.81 0.52	7.80 0.53	7.84 0.52	7.84 0.53	7.90 0.54	7.81 0.53	7.85 0.53	7.98 0.54
MEPSM	**5.65** 0.40	**5.67** 0.40	**5.96** 0.40	**6.04** 0.41	**5.63** 0.79	**5.62** 0.79	**5.60** 0.79	**5.61** 0.79	**5.76** 1.18

Pattern Set Size r=1000

Genome *m*	16	18	20	22	24	26	28	30	32
WM(8,1)	41.44 0.52	38.26 0.52	37.24 0.54	36.00 0.56	35.08 0.57	34.24 0.59	32.79 0.58	32.42 0.60	32.09 0.60
WM(8,2)	41.55 0.64	32.83 0.69	28.83 0.71	27.56 0.73	25.55 0.73	24.93 0.77	23.02 0.75	22.52 0.78	22.05 0.80
MBNDM(8)	25.22 0.39	25.23 0.39	25.28 0.39	25.09 0.39	25.36 0.39	25.05 0.40	25.14 0.39	25.28 0.40	25.33 0.40
MEPSM	**8.08** 0.40	**8.09** 0.40	**8.05** 0.40	**7.87** 0.40	**7.96** 0.77	**7.92** 0.77	**7.89** 0.77	**7.96** 0.78	**8.53** 1.15
Protein									
WM(4,1)	8.28 0.53	7.71 0.54	7.49 0.56	7.23 0.56	6.98 0.57	6.86 0.57	6.72 0.59	6.59 0.59	6.48 0.60
WM(8,1)	9.87 0.53	8.78 0.54	8.06 0.56	7.54 0.58	7.11 0.58	6.77 0.59	6.52 0.60	6.31 0.62	6.17 0.64
MBNDM(5)	8.47 0.37	8.52 0.39	8.50 0.38	8.51 0.38	8.57 0.38	8.58 0.39	8.51 0.38	8.49 0.38	8.64 0.39
MBNDM(8)	7.76 0.52	7.81 0.52	7.80 0.53	7.84 0.52	7.84 0.53	7.90 0.54	7.81 0.53	7.85 0.53	7.98 0.54
MEPSM	**5.65** 0.40	**5.67** 0.40	**5.96** 0.40	**6.04** 0.41	**5.63** 0.79	**5.62** 0.79	**5.60** 0.79	**5.61** 0.79	**5.76** 1.18
English									
WMQ(8,1)	21.01 0.54	18.84 0.55	17.13 0.55	15.96 0.58	15.11 0.59	14.45 0.60	14.04 0.60	13.55 0.62	13.33 0.62
WM(8,2)	27.18 0.67	20.27 0.70	17.41 0.72	15.89 0.76	14.68 0.77	14.07 0.80	13.16 0.82	12.68 0.86	12.24 0.85
MBNDM(5)	16.60 0.38	16.61 0.38	16.57 0.38	16.56 0.38	16.61 0.38	16.54 0.38	16.62 0.38	16.72 0.38	16.65 0.38
MEPSM	**10.37** 0.40	**10.40** 0.40	**9.92** 0.39	**9.93** 0.40	**9.62** 0.79	**9.61** 0.77	**9.61** 0.78	**9.68** 0.78	**9.83** 1.18

Pattern Set Size r=10000

Genome *m*	16	18	20	22	24	26	28	30	32
WM(5,2)	119.29 1.72	119.49 1.86	119.68 2.00	122.00 2.18	120.50 2.32	120.28 2.43	120.53 2.57	120.82 2.74	120.94 2.85
WM(8,2)	135.98 1.58	126.30 1.77	124.21 2.00	125.15 2.24	123.64 2.43	123.50 2.61	123.78 2.82	123.62 3.04	123.99 3.22
MBNDM(8)	377.14 1.34	386.98 1.37	389.70 1.37	393.60 1.38	393.82 1.39	397.11 1.40	415.40 1.45	421.28 1.46	420.84 1.46
MEPSM	**50.97** 3.87	**51.55** 3.92	**47.62** 3.97	**47.60** 4.00	**51.52** 7.65	**51.90** 7.66	**55.93** 8.21	**54.60** 8.02	**64.85** 11.78

Pattern Set Size r=10000

Genome *m*	16	18	20	22	24	26	28	30	32
WM(5,2)	119.29 1.72	119.49 1.86	119.68 2.00	122.00 2.18	120.50 2.32	120.28 2.43	120.53 2.57	120.82 2.74	120.94 2.85
WM(8,2)	135.98 1.58	126.30 1.77	124.21 2.00	125.15 2.24	123.64 2.43	123.50 2.61	123.78 2.82	123.62 3.04	123.99 3.22
MBNDM(8)	377.14 1.34	386.98 1.37	389.70 1.37	393.60 1.38	393.82 1.39	397.11 1.40	415.40 1.45	421.28 1.46	420.84 1.46
MEPSM	**50.97** 3.87	**51.55** 3.92	**47.62** 3.97	**47.60** 4.00	**51.52** 7.65	**51.90** 7.66	**55.93** 8.21	**54.60** 8.02	**64.85** 11.78
Protein									
WM(8,1)	24.36 1.54	23.05 1.65	22.48 1.78	22.11 1.91	21.82 2.03	21.76 2.16	21.58 2.25	21.53 2.39	21.63 2.49
MBNDM(8)	**19.75** 1.51	**19.68** 1.51	**19.75** 1.51	**19.76** 1.52	**19.94** 1.55	**19.95** 1.56	**20.06** 1.59	**20.60** 1.60	**20.72** 1.62
MEPSM	22.74 4.05	22.84 4.08	27.43 3.97	27.36 3.95	31.75 7.71	31.29 7.72	32.03 7.74	33.41 7.83	42.73 11.56

Pattern Set Size r=10000

Genome *m*	16	18	20	22	24	26	28	30	32
WM(5,2)	119.29 1.72	119.49 1.86	119.68 2.00	122.00 2.18	120.50 2.32	120.28 2.43	120.53 2.57	120.82 2.74	120.94 2.85
WM(8,2)	135.98 1.58	126.30 1.77	124.21 2.00	125.15 2.24	123.64 2.43	123.50 2.61	123.78 2.82	123.62 3.04	123.99 3.22
MBNDM(8)	377.14 1.34	386.98 1.37	389.70 1.37	393.60 1.38	393.82 1.39	397.11 1.40	415.40 1.45	421.28 1.46	420.84 1.46
MEPSM	**50.97** 3.87	**51.55** 3.92	**47.62** 3.97	**47.60** 4.00	**51.52** 7.65	**51.90** 7.66	**55.93** 8.21	**54.60** 8.02	**64.85** 11.78
Protein									
WM(8,1)	24.36 1.54	23.05 1.65	22.48 1.78	22.11 1.91	21.82 2.03	21.76 2.16	21.58 2.25	21.53 2.39	21.63 2.49
MBNDM(8)	**19.75** 1.51	**19.68** 1.51	**19.75** 1.51	**19.76** 1.52	**19.94** 1.55	**19.95** 1.56	**20.06** 1.59	**20.60** 1.60	**20.72** 1.62
MEPSM	22.74 4.05	22.84 4.08	27.43 3.97	27.36 3.95	31.75 7.71	31.29 7.72	32.03 7.74	33.41 7.83	42.73 11.56
English									
WM(5,2)	91.42 1.91	86.60 2.03	84.85 2.18	82.93 2.35	83.23 2.52	80.96 2.65	80.34 2.81	79.82 2.94	80.08 3.12
WM(8,2)	90.23 1.70	74.00 1.92	68.27 2.16	**64.27** 2.40	**62.50** 2.65	**59.98** 2.86	**58.69** 3.08	**57.97** 3.32	**57.91** 3.54
MBNDM(8)	116.23 1.45	116.98 1.44	117.73 1.47	118.21 1.47	118.91 1.48	119.29 1.51	119.18 1.50	119.94 1.53	119.78 1.53
MEPSM	**72.54** 3.88	**72.94** 3.87	**66.26** 3.92	67.39 3.97	74.36 7.62	76.10 7.63	75.11 7.55	76.54 7.65	85.21 11.35