A Word Recurrence Based Algorithm To Extract Genomic Dictionaries

Vincenzo Bonnici, Giuditta Franco, Vincenzo Manca

Department of Computer Science,

University of Verona, Italy

vincenzo.bonnici@univr.it

UNIVERSITÀ di VERONA
Dipartimento di INFORMATICA

The Thirteenth International Conference on Bioinformatics,
Biocomputational Systems and Biotechnologies
BIOTECHNO 2021
May 30, 2021 to June 03, 2021 - Valencia, Spain

Vincenzo Bonnici

Department of Computer Science, University of Verona, Italy
vincenzo.bonnici@univr.it

Vincenzo Bonnici is Temporary Assistant Professor at the Department of Computer Science of the University of Verona (Italy). He received the bachelor's degree and the master's degree in Computer Science from the University of Catania (Italy) in 2008 and 2011, respectively. He received the Ph.D. in Computer Science from the University of Verona in 2015. In 2013-2014 he ha been Visiting Researcher Scholar at the Institute for Genomics and Bioinformatics (IGB), University of California, Irvine, and in 2017 he has done an internship at the Fondazione per la Ricerca e la Cura dei Linfomi del Ticino, Istituto Oncologico della Svizzera Italiana, Bellinzona. From 2015 to 2019 he has been research associate at the University of Verona .

Bioinformatics, Computational Biology, Algorithms and data structures, Graph Theory and Parallel computing are the main research interests of Vincenzo Bonnici. He has worked on the search of substructures within biomedical graphs, since his master's degree period. He has extended such topic to the creation, integration, and creation of biological networks. His research topics also include computational genomics by developing innovative methods based on Information Theory for the analysis of genomic sequences. Phylogenomic and pangenomics are also among his research interests in which he studies computational methods for computing homology among biological sequences. He has also worked in the System Biology field by studying the physical and functional relations of non-coding RNAs. Parallel computing is a cross-cutting theme for his research studies, and he has developed parallel solutions on top of different architectures, from SMPs to GP-GPUs. As a result of his research activities, he has published several papers in relevant scientific journals (Bioinformatics, BMC, IEEE/ACM TCBB, Nature – Scientific Reports), he won the international contest on graph matching within the 2014 ICPR conference and a best poster award at the Jacob T. Schwartz International School for Scientific Research, and he has been speaker at 12 international scientific conferences.

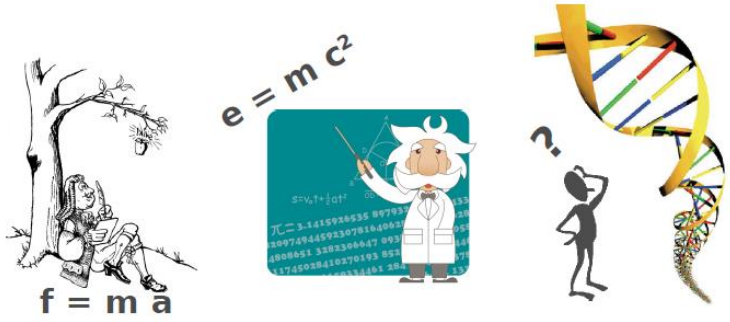
InfoGenomics

The **InfoGenomics** project aims at providing a systematic approach from an informational point of view by making use of informational analysis and well-characterized genomic features: **indices, distributions, representations (and visualizations).**

The **ENCODE** project created an encyclopedia of DNA elements by annotating the human genome in terms of bio-chemical function. It provided evidence that 80% of the human genome, first considered junk regions, is covered by functional elements.

This scenario has an **informational basis linked to DNA fragments** related to such functional elements.
An integration between biochemical and informational analysis could provide new possibilities for interpreting data and to discover **principles of genome organization and functions**.

Maths in action



Manca, V.: **Infobiotics: information in biotic systems**. Springer (2013)
Castellini, A., Franco, G., Manca, V.: **A dictionary based informational genome analysis.** BMC Genomics, 13, 485 (2012)
Bonnici, V., Manca, V. (2016). Informational laws of genome structures. *Scientific reports*, *6*(1), 1-10.
The Encode Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** Nature 489, 57–72 (2012)

Extraction of Genomic Dictionaries

Goal: to extract a set (dictionary) of **factors** from a genomic sequence **G** that are evaluated according to the following aspects:

- **length** of words
- (sequence and positional) **coverage** over G

Mini-max **criteria** are the methods for the evaluation:

i.e. extract a set of words minimizing their lengths and maximizing their sequence coverage.

Evaluation criteria

Let **W** to be the resultant dictionary output from an elongation procedure
We apply the following measurements:

- **Word Length Distribution** of words in **W**

- **Coverage** of **W** in **G**
  - A position **p** in G is said **covered** if there exists **at least** one word **w** in W such that **G[i,j]** is an **occurrence** of **w** and **i≤p≤j**

  - The **sequence** coverage **cov(G,W)** is defined as ratio between the number of covered positions of G by the words in W, and the total number of positions in G
    - Since real genomes may contains **N**, the total number of positions in G is intended without considering the positions where G[i]=N

  - A position p of G can be covered by several words in W, or multiple times by the same word
    - The **position** coverage **cov(G,W,p)** is defined as the total number of times that p is involved in an occurrence of a word in W
    - We calculate the average (**Avg**) and the standard-deviation (**SD**) of **cov(G,W,p)** for p belonging to the set of covered positions

Extraction of Genomic Dictionaries

State of the art (come examples)

Genome assembly is still an open problem

- Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., ... & Yorke, J. A. (2012). **GAGE: A critical evaluation of genome assemblies and assembly algorithms.** Genome Research, 22(3), 557-567.
- Treangen, T. J., & Salzberg, S. L. (2011). **Repetitive DNA and next-generation sequencing: computational challenges and solutions.**
- Nature Reviews Genetics, 13(1), 36-46.

Special factors and **sequence reconstructability**: the dictionary is composed by the minimal hapaxes (ShUS, shortest unique substring) of the sequence .

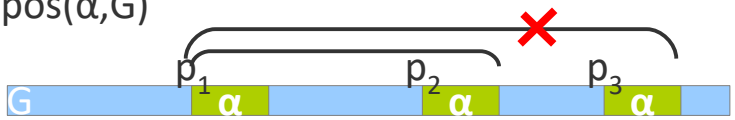
- Carpi, A., & De Luca, A. (2001). **Words and special factors.** Theoretical Computer Science, 259(1), 145-182.
- Mignosi, F., Restivo, A., & Sciortino, M. (2002). **Words and forbidden factors.** Theoretical Computer Science, 273(1), 99-117.
- Mignosi, F., Restivo, A., & Sciortino, M. (2002, January). **Forbidden factors and fragment assembly**. In Developments in Language Theory (pp. 349-358). Springer Berlin Heidelberg.
- Dudík, M., & Schulman, L. J. (2003). **Reconstruction from subsequences.** Journal of Combinatorial Theory, Series A, 103(2), 337-348.
- Fici, G., Mignosi, F., Restivo, A., & Sciortino, M. (2006). **Word assembly through minimal forbidden words**. Theoretical computer science, 359(1), 214-230.
- Piña, C., & Uzcátegui, C. (2008). **Reconstruction of a word from a multiset of its factors.** Theoretical Computer Science, 400(1), 70-83.

Some well-known genomic elements are **clustered** along the genome sequence.

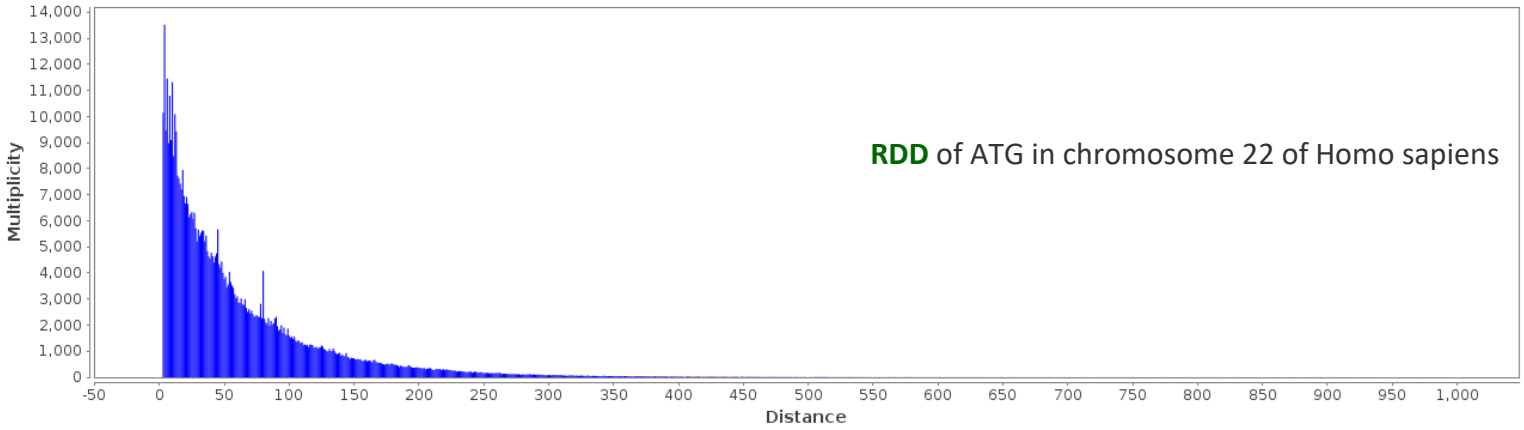
- Extract words by looking at their **clustering coefficient**: based on a comparison between **RDD** and an **empirical** geometric distribution.
- Select **top k-mers**
  - Hackenberg M., Rueda A., Carpena P., Bernaola-Galvan P´., Barturen G., and Oliver J. L.. **Clustering of dna words and biological function: A proof of principle**. Journal of theoretical biology, 297:127–136, 2012.
- **Elongate** words from seeds
  - Carpena P., Bernaola-Galv an P., Hackenberg M., Coronado AV., and Oliver JL. **Level statistics of words: Finding keywords in literary texts and symbolic sequences**. Physical Review E, 79(3):035102, 2009.
  - Carpena P., Bernaola-Galvan P., and Ivanov PC. **New class of level a statistics in correlated disordered chains**. Physical review letters, 93(17):176804, 2004.

Recurrence Distance Distribution (RDD)

- Let $\alpha \in D_k(G)$
- **Recurrence** $(p_1,p_2) : p_1 , p_2 \in pos(\alpha,G)$
- **Minimal recurrence** $(p_1,p_2) : \nexists\ p' : \ p_1 < p' < p_2 \ \ AND \ \ p' \in pos(\alpha,G)$
- The **recurrence distance** is given by $p_2 - p_1$



- Given a word **α**, the distribution assigns to any **distance n**, going from 1 to some maximum value, the number of times it occurs at distance **n** from its previous occurrence



RDD of ATG in chromosome 22 of Homo sapiens

•Achuth Sankar S Nair and T Mahalakshmi. **Visualization of genomic data using inter-nucleotide distance signals**. Proceedings of IEEE Genomic Signal Processing, 408, **2005**.
•Vera Afreixo, et. al. **Genome analysis with inter-nucleotide distances.** Bioinformatics, 25(23):3064–3070, **2009**.

Recurrence Distance Distribution (RDD)

- If we suppose a sequence as generated by a **Bernoullian process**
- the occurrences of words within the sequences can be considered independent and the occurrence of k of them follows a **Poisson distribution**

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

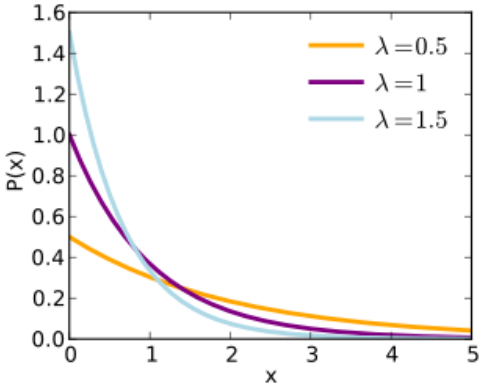
where **e** is the **Euler's number**
and **λ** is the **variance** (the mean value of the distribution)

- It does not describe where such occurrences are, neither the distances between them.

- The **Waiting Time** of a **Poisson Distribution** is the **Exponential Distribution**
  (or its discrete counterpart that is the geometric distribution)

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

- In our case
  - **x** is the **distance** between two occurrences
  - **P(x)** is the **number of times** two occurrences appear at distance **x**



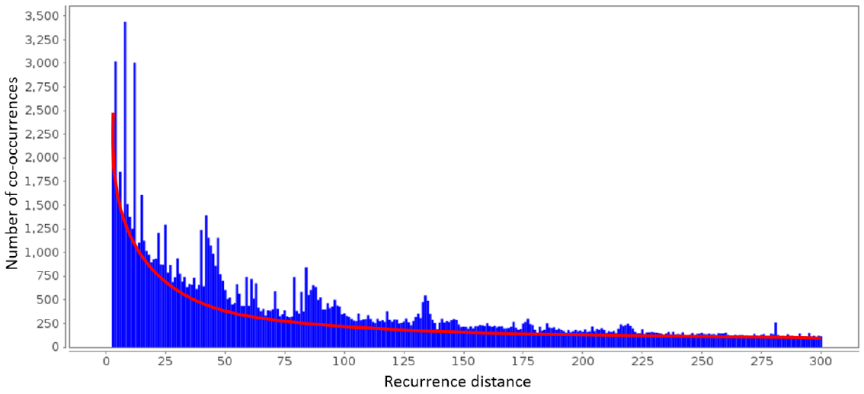
A novel word recurrence based approach

Approach

- Characterize words by the **divergence** of their RDD (Recurrence Distance Distribution) to a theoretical distribution
- The divergence is used as a measure of the **information content of a word**
- **Elongate** low expressive words until they acquire a reasonable level of significance

Random deviation of a word α

1) **Extract** the RDD of α in G, R_α
2) Remove distribution noise (peaks)
3) Force $\mathbf{R_\alpha}$ to be a probability distribution
4) **Estimate** an exponential distribution $\mathbf{E_\alpha}$ from R_α
5) Force E_α to be a probability distribution
6) Calculate the **random deviation** as $\mathbf{r(\alpha)} = \max(KL(R_\alpha,E_\alpha), KL(E_\alpha,R_\alpha))$
7) where **KL** is the **entropic divergence** (namely the Kullback-Leibler divergence)



A novel word recurrence based approach

Ideally, **extract** α **such that:**

$$r(\alpha[1,|\alpha|-1]) < r(\alpha) < r(\alpha x), \qquad \text{for } x \in \{A,C,G,T\}$$

We could scan over $D_k(G)$, for sever values of **k,** but it would be extremely **expensive**
Solution: use an **elongation procedure** in order to elongate **seed words** up to more expressive words.
Discard seeds from the resultant dictionary.

Algorithm 1: ExtractFrom(G,D_0)

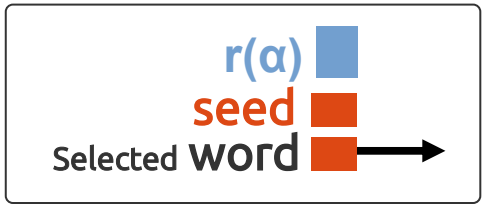
```
W ← {∅}
foreach α ∈ D_0 do
    Elongate(α, W)
end
W ← W \ D_0
return W
```

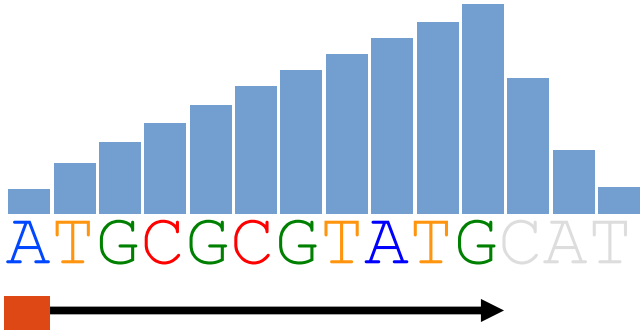
Algorithm 2: Elongate(α,W)

```
if r(αx) ≤ r(α), ∀x ∈ Γ then
    W ← W ∪ {α}
end
else
    foreach x ∈ Γ do
        if r(αx) > r(α) then
            Elongate(αx, W)
        end
    end
end
```

Elongation procedure

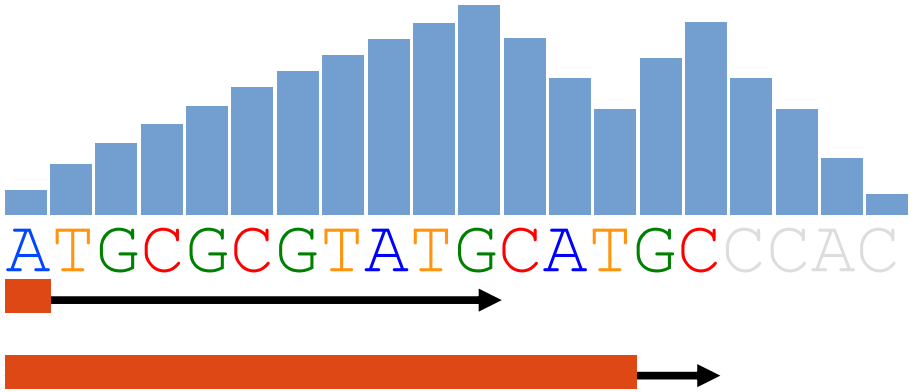
Words, factors and roots





Basic elongation strategy

Elongate a seed until r(α) increases

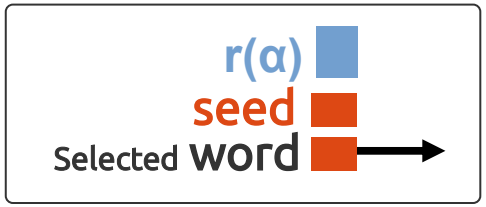


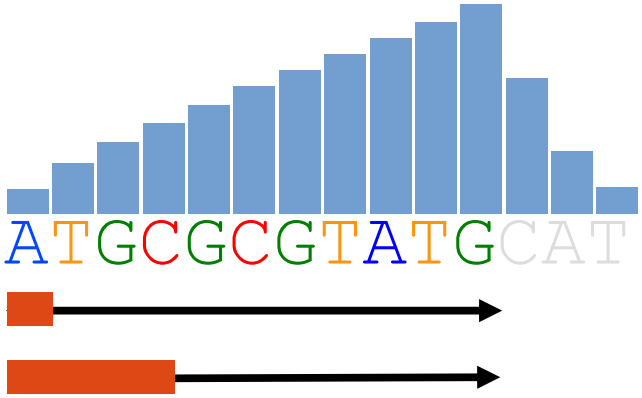
Subword jumping

Use long seeds in order to discover words that contain smaller words

Elongation procedure

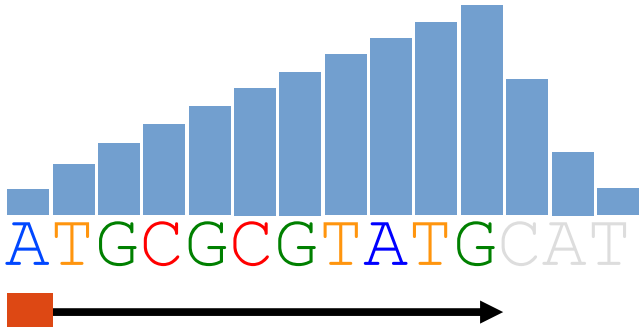
Words, factors and roots





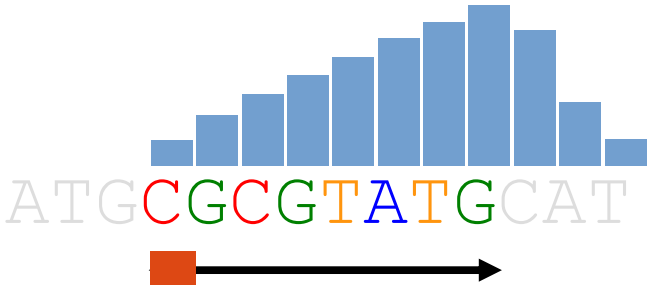
Elongation strategy is inversely inclusive w.r.t. seeds

Elongation from large seeds include what smaller seed elongate



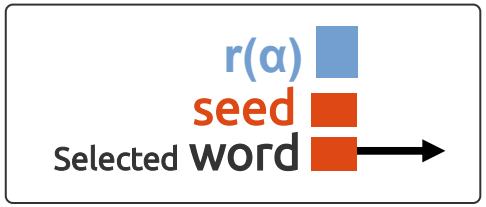
Suffix selection bias

Ideally unwanted, however they may correspond to some roots



Elongation procedure

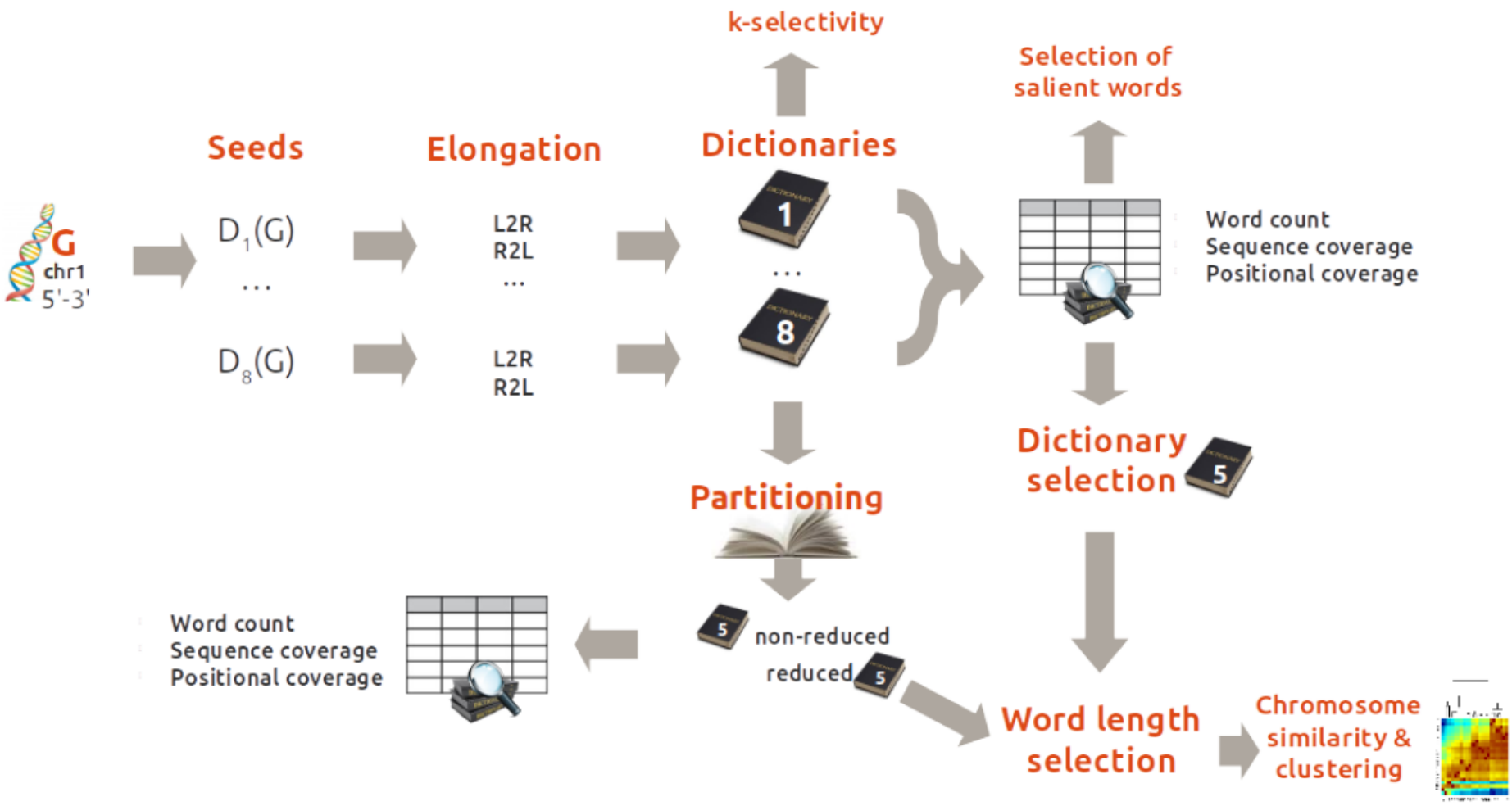
Words, factors and roots



Proto-word selection

Ideally unwanted.
The proper prefix and the proper suffix are selected instead of the real word

Informational analisys pipeline



Elongation procedure

Preferred word lengths in WLD

Hg19, chromosome 22

extracted word length \ seed	1	2	3	4	5	6	7	8
4	2	13	20					
5	31	134	202	272				
6	**63**	**349**	**517**	**995**	**1,261**			
7	57	180	232	350	475	1,343		
8	57	193	277	430	679	3,001	10,668	
9	10	144	241	529	1,073	7,602	29,521	53,314
10	2	**201**	**326**	**794**	**1,391**	**9,126**	**59,951**	**129,872**
11	5	151	233	569	923	4,302	63,089	184,296
12	2	64	91	198	323	973	24,275	97,646
13		21	30	51	81	225	4,592	20,670
14		2	3	10	18	40	875	3,525
15		2	2	5	6	11	190	724
16		4	5	5	5	9	54	165
17		1	1	2	2	3	17	54
18							5	19
19								5
20							1	6
21								3
22								6
23								1

Elongation procedure

Preferred word lengths in WLD

Hg19, chromosome 1

extracted word length \ seed	1	2	3	4	5	6	7	8
4		5	5					
5	17	54	108	179				
6	41	305	666	1,306	1,666			
7	**92**	**337**	**616**	**1,478**	**2,310**	**2,925**		
8	79	178	280	468	593	1,474	4,151	
9	43	142	248	562	811	3,879	14,614	39,347
10	8	**221**	**542**	**1,325**	**2,140**	**9,106**	48,112	144,355
11	**13**	197	479	1,284	2,115	6,986	**50,442**	**224,644**
12		122	297	838	1,363	2,201	24,687	303,163
13	2	53	119	327	579	774	6,403	136,135
14	2	19	36	80	145	194	1,094	20,805
15	2	7	9	21	33	50	291	4,193
16		5	7	12	17	24	99	1,196
17		2	3	5	6	9	27	327
18		1	1	1	1	4	12	128
19							2	43
20							2	15
21								6
23								1
24								1

Elongation procedure

Preferred word lengths and their sequence coverage

Hg19, chromosome 1, sequence coverage

Extracted word length \ seed	1	2	3	4	5	6	7	8
4		0.0291	0.0291					
5	0.0309	0.0790	0.1362	0.1681				
6	0.0269	**0.3149**	**0.5504**	**0.7767**	**0.8426**			
7	**0.0742**	0.2479	0.3878	0.6430	0.7691	**0.8141**		
8	0.0285	0.0616	0.0899	0.1187	0.1384	0.1643	**0.2634**	
9	0.0115	0.0209	0.0303	0.0499	0.0615	0.0714	0.1593	**0.6315**
10	0.0008	0.0054	0.0071	0.0128	0.0206	0.0329	0.0974	0.5388
11	**0.0025**	**0.0077**	**0.0088**	**0.0108**	**0.0127**	0.0174	0.0602	0.3509
12		0.0028	0.0031	0.0081	0.0089	0.0101	0.0342	0.2858
13	0.0000	0.0006	0.0013	0.0054	0.0065	0.0070	0.0155	0.1209
14	0.0035	0.0048	0.0049	0.0056	0.0065	0.0066	0.0101	0.0451
15	0.0026	0.0036	0.0036	0.0050	0.0052	0.0052	0.0065	0.0214
16		0.0016	0.0017	0.0017	0.0017	0.0028	0.0032	0.0090
17		0.0011	0.0011	0.0012	0.0013	0.0013	0.0014	0.0031
18		0.0006	0.0006	0.0006	0.0006	0.0012	0.0012	0.0020
19							0.0000	0.0003
20							0.0000	0.0002
21								0.0001
23								0.0000
24								0.0000

Elongation procedure

Preferred word lengths and their sequence coverage

Hg19, chromosome 1, positional coverage

Extracted word length \ seed	1	2	3	4	5	6	7	8
4		1.0078	1.0078					
5	1.0807	1.1690	1.2411	1.4198				
6	**1.1539**	**1.3022**	**1.6590**	**2.3201**	**2.7715**			
7	1.0934	1.2876	1.4587	1.9817	2.5877	2.9160		
8	1.1569	1.2590	1.3125	1.4228	1.5184	1.5836	1.5572	
9	**1.4480**	**1.5411**	**1.5211**	**1.7039**	**1.8791**	**1.8661**	1.5470	1.7484
10	1.0006	1.1090	1.1033	1.1697	1.1926	1.2632	1.2580	1.5457
11	**4.0810**	**2.1729**	**2.0809**	**1.9100**	**1.7829**	**1.6131**	**1.3009**	1.3658
12		1.0654	1.0624	1.1926	1.1809	1.1716	1.1507	1.3455
13	1.0000	1.0000	1.0000	1.1355	1.3769	1.3530	1.2340	1.3709
14	1.0000	1.0000	1.0000	1.0551	1.2244	1.2235	1.1687	1.3807
15	1.0000	1.1446	1.1445	1.1065	1.1739	1.1725	1.1444	1.2559
16		1.2684	1.2636	1.2588	1.2539	1.1544	1.1447	1.1148
17		1.0000	1.0000	**1.3982**	**1.3957**	**1.3948**	**1.3608**	**1.3440**
18		1.0000	1.0000	1.0000	1.0000	1.0000	1.0015	1.0187
19							1.0000	1.0000
20							1.0000	1.0000
21								1.0000
23								1.0000
24								1.0000

Elongation procedure

Preferred word lengths and their coverages

$W_k = D_k(W)$

extracted word length k	Word count $\lvert W_k \rvert$	Sequence coverage $cov(G,W_k)$	Positional coverage $avg(cov(G,W_k,p))$
	5	5	5
6	1,666	**0.8426**	**2.7715**
7	**2,310**	0.7691	2.5877
8	593	0.1384	1.5184
9	811	0.0615	**1.8791**
10	**2,140**	0.0206	1.1926
11	2,115	**0.0127**	**1.7829**
12	1,363	0.0089	1.1809
13	579	0.0065	1.3769
14	145	0.0065	1.2244
15	33	0.0052	1.1739
16	17	0.0017	1.2539
17	6	0.0013	**1.3957**
18	1	0.0006	1.0000

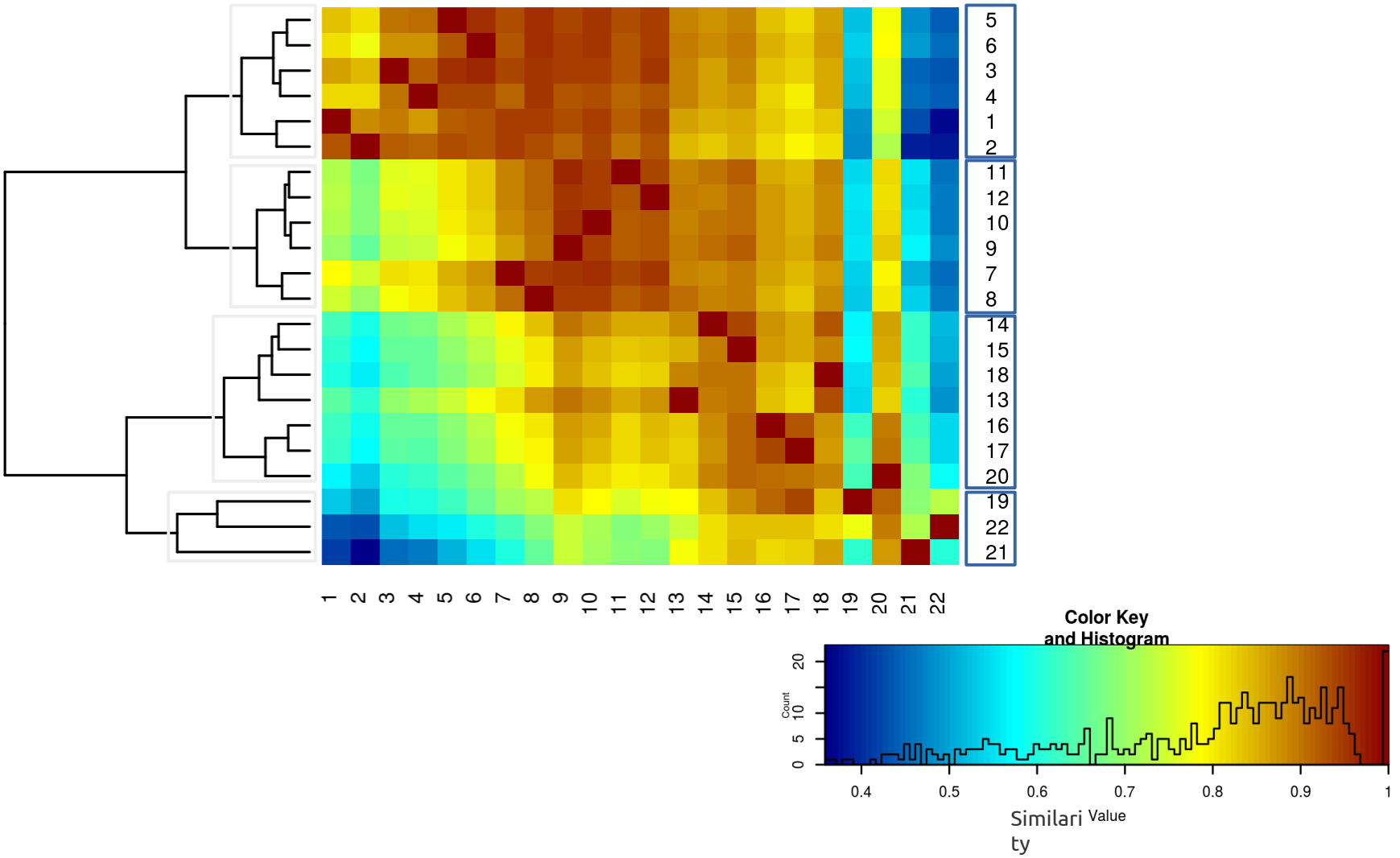
Seeds = $D_5(G)$

Hg19, chromosome 1

Chromosome clustering by dictionary intersection

Similarity = intersection / union

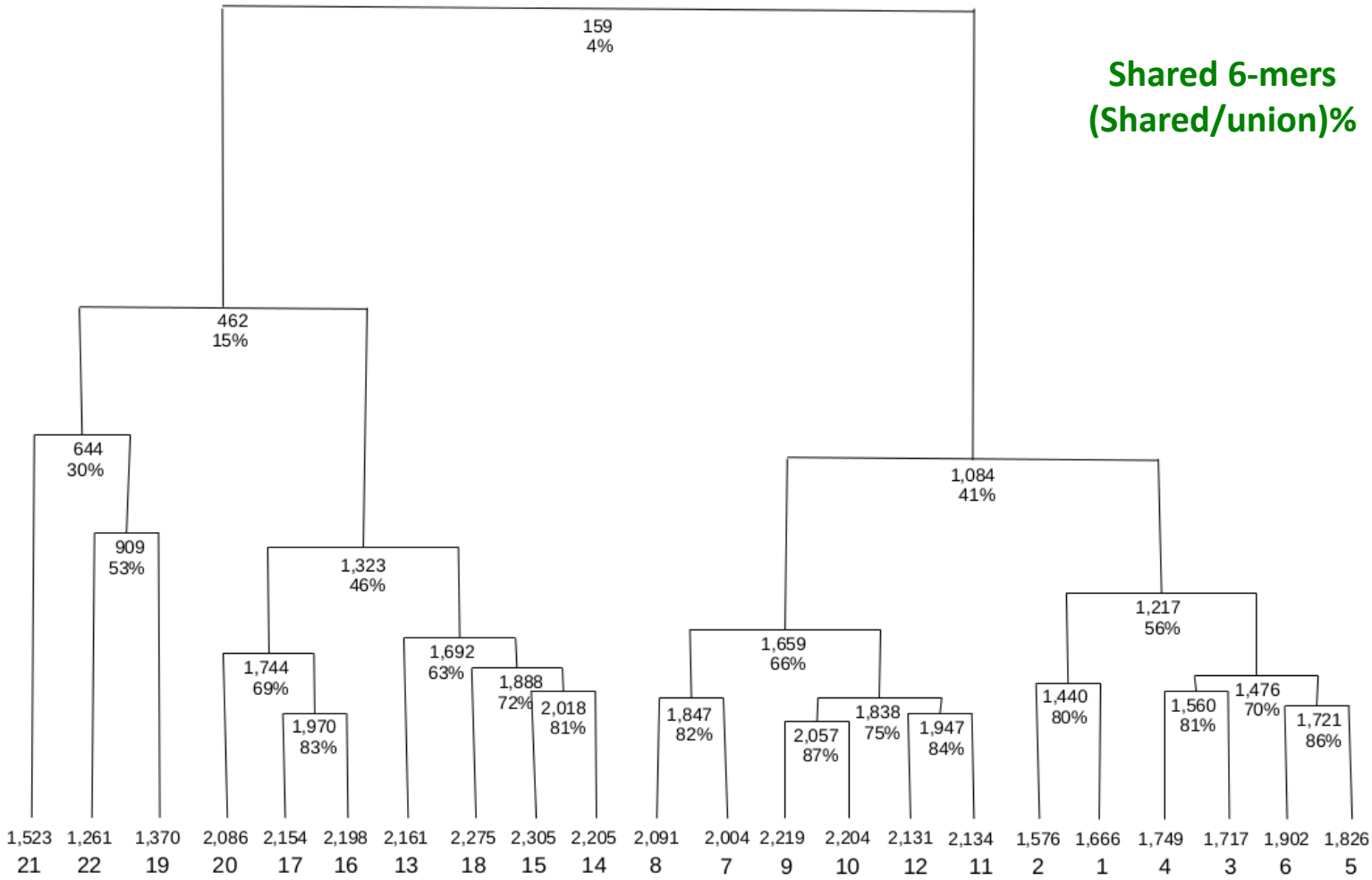
Seed length = 5, word length = 6



Chromosome clustering by dictionary intersection

Similarity = intersection / union

Seed length = 5, word length = 6



Shared 6-mers
(Shared/union)%

Conclusions

- We have described an **information theoretical methodology** to extract relatively small **genomic dictionaries**, which have good properties in terms of genome **coverage** (both sequence and average positional coverage as close as possible to one)
- We have shown that **preferred seed lengths** emerge, from an observation of sequence and positional genome coverage that provide a better coverage.
- Moreover, dictionaries of examers were identified to reveal a clear similarity pattern for **human chromosomes**.
- This preliminary work has shown that **divergence from randomness** can be an interesting measure in deciphering the genome language.