Google Matrix Analysis of DNA Sequences

Vivek Kandiah and Dima Shepelyansky

Laboratoire de Physique Théorique, IRSAMC, UMR 5152 du CNRS
Université Paul Sabatier, Toulouse

Supported by EC FET open project NADINE

21 june 2014

Introduction : motivation

- Large and accurate genomic dataset available for several species[1].
- Interest in detection of specific/rare patterns in a given sequence.
- New viewpoint of directed network.

Google matrix : $G_{ij} = \alpha S_{ij} + (1-\alpha)/N$

with $S_{i,j} = T_{i,j} / \sum_j T_{i,j}$ where T describes the transitions between nearby words.

[1] http://www.ensembl.org/

Introduction : from DNA sequence to network

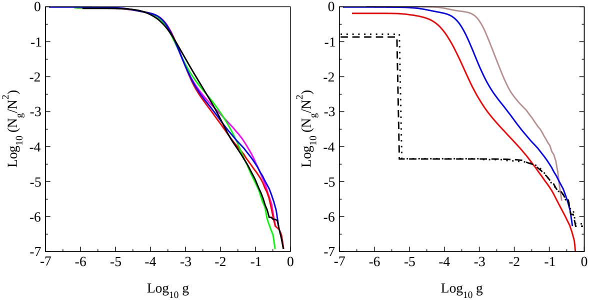
- Bos Taurus (Bull,$L \approx 2.9 \cdot 10^9 bp$); Canis Familiaris (Dog,$L \approx 2.5 \cdot 10^9 bp$); Loxondonta Africana (Elephant,$L \approx 3.1 \cdot 10^9 bp$); Homo Sapiens (Human,$L \approx 1.5 \cdot 10^{10} bp$) and Danio Rerio (Zebrafish,$L \approx 1.4 \cdot 10^9 bp$).

$$...TCG\,\underbrace{ATAT}_{W_{k-1}}\,\underbrace{CTGG}_{W_k}\,\underbrace{TAAC}_{W_{k+1}}\,CTA...$$

$$\rightarrow W_{k-1} \rightarrow W_k \rightarrow W_{k+1} \rightarrow$$

- $T_{ij} \rightarrow T_{ij} + 1$ whenever word j points to word i.
- Full matrix limit, $L/mN^2 \approx$ 10 to 100 transitions per elements at $m = 6$.
- Webpages $\approx$ 10 links per node on average with $N \approx 2 \cdot 10^5$.

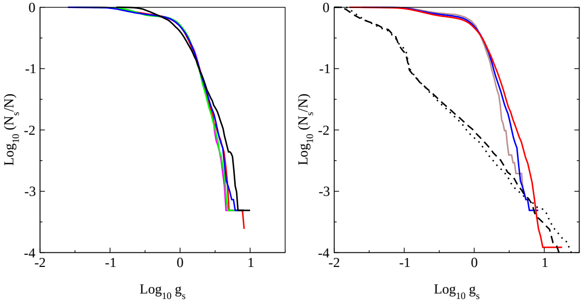
Statistics of Google matrix elements



Integrated fraction N_g/N^2 of Google matrix elements with $G_{ij} > g$ as a function of g. *Left panel :* Various species with 6-letters word length: bull BT (magenta), dog CF (red), elephant LA (green), Homo sapiens HS (blue) and zebrafish DR(black). *Right panel :* Data for HS sequence with words of length $m = 5$ (brown), 6 (blue), 7 (red). For comparison black dashed and dotted curves show the same distribution for the WWW networks of Universities of Cambridge and Oxford in 2006 respectively.

- Oscillations but universal decay law $N_g \propto 1/g^{\nu-1}$ with $\nu \approx 2.5$ (range $-5.5 < \log_{10} g < -0.5$).
- Distribution of outgoing links in WWW networks decay with $\tilde{\nu} \approx 2.7$.

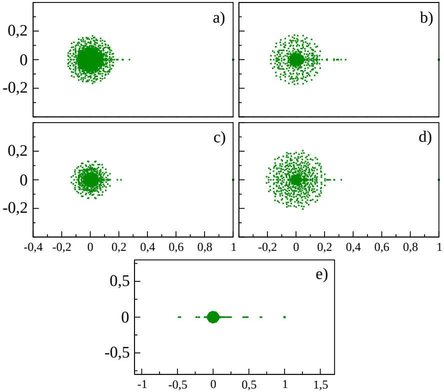
Statistics of Google matrix elements



Integrated fraction N_S/N of sum of ingoing matrix elements with $\sum_{j=1}^{N} G_{i,j} \geq g_S$. Left and right panels show the same cases as above in same colors. The dashed and dotted curves are shifted in x-axis by one unit left to fit the figure scale.

- Visible differences between species but close to universal decay curve as $N_s \propto 1/g^{\mu-1}$ with $\mu \approx 5$.
- Distribution of ingoing links in WWW networks decay with $\tilde{\mu} \approx 2.1$.

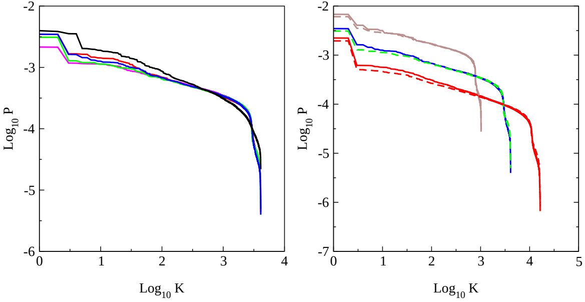
Spectrum and PageRank



Eigenvalue spectrum at $m = 6$ of a) Bos Taurus, b) Canis Familiaris, c) Loxodonta Africana, d) Homo Sapiens and e) Danio Rerio.

- Presence of large gap.
- HS $\sim$ CF and strong differences between mammalian and non mammalian sequences.
- Spectrum of G and G^* are identical.

Spectrum and PageRank



PageRank probability decay of several species at $m = 6$ (left) and Homo Sapiens at $m = 5$, $m = 6$ and $m = 7$ (right).

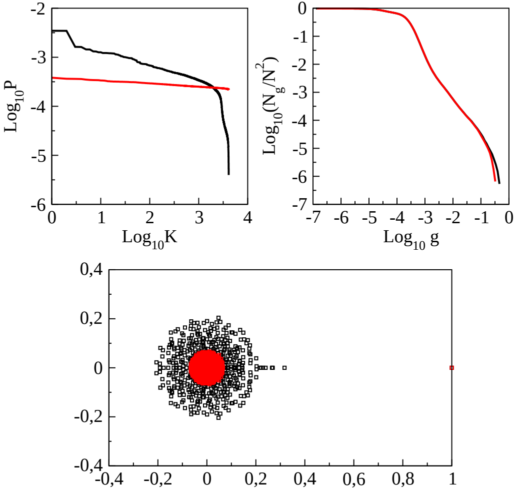
- PageRank $\sim$ frequency of words.
- $P(K) \sim 1/K^{\beta}$ with $\beta = 1/(\mu - 1)$.
- At $m = 6 : \beta = 0.273 \pm 0.005$ (BT), 0.340 ± 0.005 (CF), 0.281 ± 0.005 (LA), 0.308 ± 0.005 (HS), 0.426 ± 0.008 (DR) in the range $1 \leq \log_{10} K \leq 3.3$. Small variation between mammalian species, stable with word length.

Top five (top) and last five (bottom) PageRank entries of DNA sequences.

BT	CF	LA	HS	DR
TTTTTT AAAAAA ATTTTT AAAAAT TTCTTT	TTTTTT AAAAAA AATAAA TTTATT AAATAA	AAAAAA TTTTTT ATTTTT AAAAAT AGAAAA	TTTTTT AAAAAA ATTTTT AAAAAT TATTTT	ATATAT TATATA AAAAAA TTTTTT AATAAA

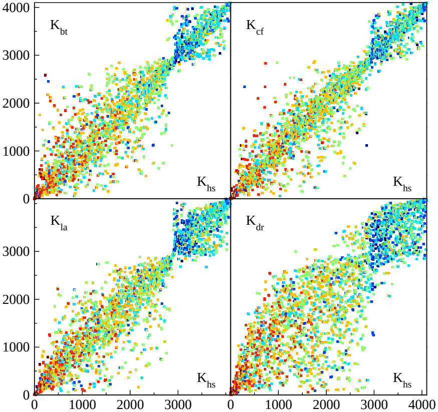
BT	CF	LA	HS	DR
CGCGTA TACGCG CGTACG CGATCG ATCGCG	TACGCG CGCGTA TCGCGA CGTACG CGATCG	CGCGTA TACGCG ATCGCG TCGCGA CGCGAT	TACGCG CGCGTA CGTACG TCGACG CGTCGA	CCGACG CGTCGG CGTCGA TCGACG TCGTCG

Spectrum and PageRank



Random matrix model with distribution of elements corresponding to HS at $m = 6$.

Statistical proximity



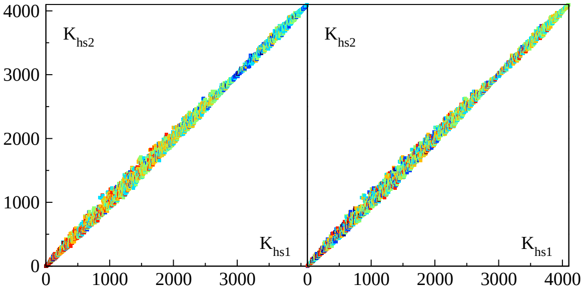
PageRank proximity K — K plane diagrams for different species in comparison with Homo Sapiens.

$$\zeta(s_1, s_2) = \frac{\sqrt{\sum_{i=1}^{N}(K_{s_1}(i) - K_{s_2}(i))^2)/N}}{\sigma_{rnd}}.$$

$$\begin{aligned}
\zeta(HS,CF) &= 0.206, & \zeta(HS,LA) &= 0.238,\\
\zeta(HS,BT) &= 0.246, & \zeta(LA,CF) &= 0.303,\\
\zeta(CF,BT) &= 0.308, & \zeta(LA,BT) &= 0.324,\\
\zeta(DR,HS) &= 0.375, & \zeta(DR,CF) &= 0.414,
\end{aligned}$$

$\zeta(DR,LA) = 0.422$, $\zeta(DR,BT) = 0.425$

Statistical proximity



PageRank proximity $K - K$ plane diagrams between two Homo Sapiens individuals.
$\zeta = 0.031$

Conclusion and Perspectives

- Complex and large gaped spectrum of Google matrix.
- Structural differences and similarities of DNA with WWW through G_{ij}.
- DNA sequence $\mu \approx 5 \rightarrow$ slow PageRank decay $\beta \approx 0.25$ (For WWW $\beta \approx 0.9$).
- PageRank correlations show the statistical similarity between species from a Markov chain point of view.
- Random matrix model reproducing the spectrum.
- Other eigenmodes may highlight a relatively long living relaxation mode and might localize themselves in a paricular set of words.

References

1. Nucleotide sequence bank http://www.ncbi.nlm.nih.gov
2. Academic Web Link Database Project http://cybermetrics.wlv.ac.uk/database/
3. S.Brin and L.Page, *Computer Networks and ISDN Systems* **30** 107 (1998).
4. A.M. Langville and C.D. Meyer C D 2006 *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, 2006.
5. Frahm KM, Shepelyansky DL (2012) *Poincaré recurrences of DNA sequences*, Phys. Rev. E **85**: 016214
6. K.M. Frahm, B. Georgeot and D.L. Shepelyansky, *Universal emergence of PageRank*, J. Phys, A: Math. Theor. **44** (2011) 465101.
7. L.Ermann, K.M.Frahm and D.L.Shepelyansky *Spectral properties of Google matrix of Wikipedia and other networks* submitted to Eur. Phys. J. B 5 Dec 2012
8. Fortunato S *Community detection in graphs* Phys. Rep.486: 75 (2010)
9. Robin S, Rodolphe F, Schbath S *DNA, words and models* Cambridge Univ. Press, Cambridge (2005)
10. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng C-K, Simons M, Stanley HE *Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics* Phys. Rev. E52: 2939 (1995)
11. Halperin D, Chiapello H, Schbath S, Robin S, Hennequet-Antier C, Gruss A, El Karoui M (2007) *Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling*, PLoS Genetics **3(9)**: e153
12. Dai Q, Yang Y, Wang T (2008) *Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison*, Bioinformatics **24(20)**: 2296
13. Reinert G, Chew D, Sun D, Waterman MS (2009) J. Comp. Biology **16(12)**: 1615
14. Burden CJ, Jing J, Wilson SR (2012) *Alignment-free sequence comparison for biologically realistic sequences of moderate length*, Stat. Appl. Gen. Mol. Biology **11(1)** 3
15. Brendel V, Beckmann JS, Trifonov EN (1986) J. Boimolecular Structure Dynamics **4**: 11
16. Popov O, Segal DM, Trifonov EN (1996) Biosystems **38**: 65
17. Frenkel Zakharia M, Frenkel Zeev M, Trifonov EN, Snir S (2009) J. Theor. Biology **260**: 438