Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data

G. Nuel, L. Regad, J. Martin and A.-C. Camproux

MAP5 (CNRS 8145), Paris Descartes University
MTi (INSERM 973), Paris Diderot University
IBCP (IFR 128, CNRS 5086), University of Lyon 1

JOBIM
Montpellier, September 7-9, 2010

Search for functional motifs in biological sequences

Motifs facts

- selection pressure ⇒ unusual counts (ex: TFs, CHI, etc.)
- functional motifs are well conserved across sequences
- statistically significant motifs ⇒ good functional candidates

Statistical framework

- $x = x_1 \dots x_\ell$ observed biological sequence
- n observed count of the motif in x
- $X = X_1 \dots X_\ell$ random sequence under a Markov model
- N random count of the motif in X ⇒ p-value $= \mathbb{P}(N \geqslant n)$

Purpose of the talk

How to compute such p-values when considering biological datasets of a large number of short sequences ?

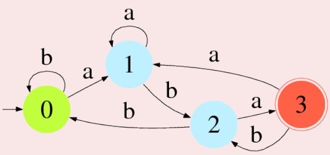
Known methods for a single sequence

Classical approaches

- Monte-Carlo simulations
- approximations (Gaussian, Poisson, Large Deviations)
- exact computations

Minimal Markov chain embedding through DFA

ex. with motif aba over the binary alphabet $\mathcal{A} = \{\text{a}, \text{b}\}$:



$$\mathbf{T} = \begin{pmatrix} \pi_{\text{b,b}} & \pi_{\text{b,a}} & 0 & 0 \\ 0 & \pi_{\text{a,a}} & \pi_{\text{a,b}} & 0 \\ \pi_{\text{b,b}} & 0 & 0 & \pi^{*}_{\text{b,a}} \\ 0 & \pi_{\text{a,a}} & \pi_{\text{a,b}} & 0 \end{pmatrix}$$

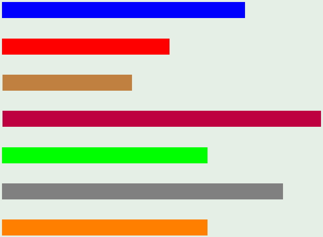
$$G(y) = \sum_{n \geqslant 0} \mathbb{P}(N = n) y^n = \mathbf{u}(\mathbf{P} + y\mathbf{Q})^{\ell}\mathbf{v} \quad \text{with } \mathbf{T} = \mathbf{P} + \mathbf{Q}$$

Dealing with several sequences

Examples of biological datasets with many sequences

- protein databases (ex: 70 000 of length from 10 to 2000)
- upstream regions (ex: 30 000 regions of length 700)
- short reads (ex: 10^6 reads of length 35)

Toy example



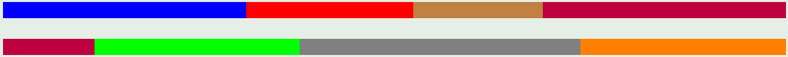
A fragmented dataset

Dealing with several sequences

Examples of biological datasets with many sequences

- protein databases (ex: 70 000 of length from 10 to 2000)
- upstream regions (ex: 30 000 regions of length 700)
- short reads (ex: 10^6 reads of length 35)

Toy example



concatenation ⇒ back to the previous case

$$G(y) = \mathbf{u}(\mathbf{P} + y\mathbf{Q})^{\ell}\mathbf{v} \quad \text{with} \quad \ell = \ell_1 + \ell_2 + \ldots + \ell_7$$

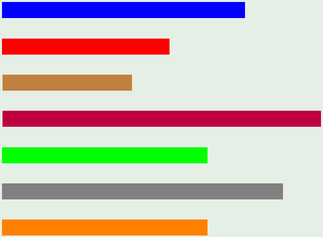
Easy but also ignore all edge effects

Dealing with several sequences

Examples of biological datasets with many sequences

- protein databases (ex: 70 000 of length from 10 to 2000)
- upstream regions (ex: 30 000 regions of length 700)
- short reads (ex: 10^6 reads of length 35)

Toy example



$$G(y) = G_1(y) \times G_2(y) \times G_3(y) \times G_4(y) \times G_5(y) \times G_6(y) \times G_7(y)$$

Two algorithms

Notations

We consider r sequences of lengths $\ell_1 \leqslant \ell_2 \leqslant \ldots \leqslant \ell_r$ and a total of n occurrences of a motif of complexity L (DFA size).

Algorithm 1: compute directly $G(y)$ by recursion

$$G(y) = \mathbf{u}(\mathbf{P} + y\mathbf{Q})^{\ell_1}\mathbf{v} \times \mathbf{u}(\mathbf{P} + y\mathbf{Q})^{\ell_2}\mathbf{v} \times \ldots \times \mathbf{u}(\mathbf{P} + y\mathbf{Q})^{\ell_r}\mathbf{v}$$

$$\Rightarrow \quad O(\ell \times n \times L) \quad \text{with} \quad \ell = \ell_1 + \ldots + \ell_r$$

(also valid with heterogeneous models)

Algorithm 2: compute all $G_j(y)$ recursively and combine them

$$G_1(y) = \mathbf{u}(\mathbf{P} + y\mathbf{Q})^{\ell_1}\mathbf{v} \quad G_2(y) = \mathbf{u}(\mathbf{P} + y\mathbf{Q})^{\ell_2}\mathbf{v} \quad \ldots$$

$$\text{then} \quad G(y) = G_1(y) \times G_2(y) \times \ldots \times G_r(y)$$

$$\Rightarrow \quad O(\ell_r \times n \times L) + O(r \times n^2)$$

Examples with Biological datasets

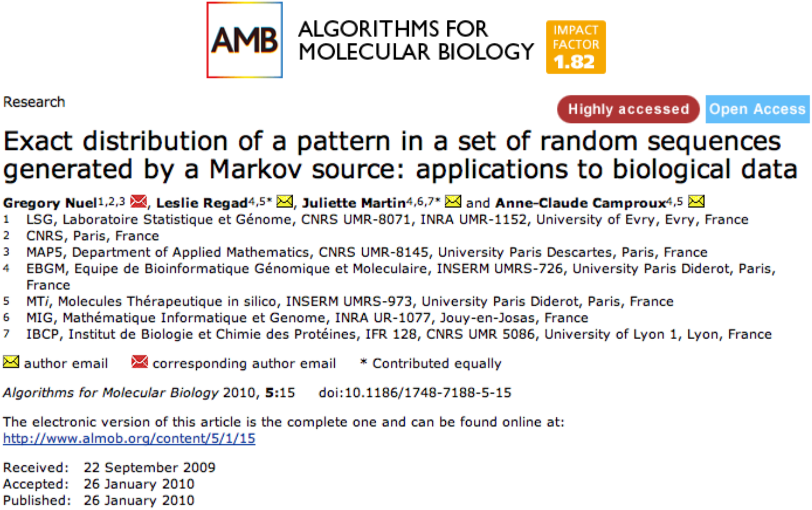
Complete proteome of *E. coli* ($r = 4\,131$, $\ell_1 = 14$, $\ell_r = 2\,358$)

PROSITE signature	L	n	exact
PILI_CHAPERONE	226	10	3.27×10^{-46}
SIGMA54_INTERACT_2	313	12	1.58×10^{-42}
EFACTOR_GTP	320	8	4.43×10^{-20}
ALDEHYDE_DEHYDR_CYS	331	11	5.63×10^{-9}
ADH_ZINC	478	12	8.93×10^{-16}
THIOLASE_1	637	5	5.76×10^{-9}
SUGAR_TRANSPORT_1	796	18	3.75×10^{-8}
FGGY_KINASES_2	2668	5	2.14×10^{-4}
PTS_EIIA_TYPE_2_HIS	2758	8	7.19×10^{-19}
MOLYBDOPTERIN_PROK_3	3907	11	2.59×10^{-35}
SUGAR_TRANSPORT_2	6689	10	1.22×10^{-5}

Examples with Biological datasets

Upstream regions of yeast genes ($r = 1\,371$, $\ell_1 = \ell_r = 800$)

DNA pattern	n	L	homogeneous	heterogeneous
CGCACCC*	28	10	2.95×10^{-3}	3.74×10^{-3}
AAGAAAAA*	427	11	1.31×10^{-99}	1.29×10^{-99}
AACAACAAC	25	10	1.76×10^{-6}	1.38×10^{-6}
TCCGTGGA*	22	11	1.12×10^{-6}	1.55×10^{-6}
GCGCGCGC	18	11	6.52×10^{-10}	1.65×10^{-9}
RTAAAYAA*	391	14	7.70×10^{-12}	1.68×10^{-12}
WWWTTTGCTCR*	15	17	4.15×10^{-1}	4.09×10^{-1}
A{24}	42	27	2.05×10^{-23}	2.14×10^{-22}
TAWWWWTAGM*	212	36	3.08×10^{-9}	3.04×10^{-9}
YCCNYTNRRCCGN*	11	40	3.10×10^{-2}	3.05×10^{-2}
GCGCN{6}GCGC	1	106	8.97×10^{-1}	8.84×10^{-1}
CGGN{8}CGG*	102	183	1.26×10^{-14}	1.73×10^{-13}
GCGCN{10}GCGC	6	464	2.88×10^{-2}	2.84×10^{-2}



AMB ALGORITHMS FOR MOLECULAR BIOLOGY IMPACT FACTOR 1.82

Research

Highly accessed Open Access

Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data

Gregory Nuel[1,2,3] ✉, **Leslie Regad**[4,5*] ✉, **Juliette Martin**[4,6,7*] ✉ and **Anne-Claude Camproux**[4,5] ✉

1 LSG, Laboratoire Statistique et Génome, CNRS UMR-8071, INRA UMR-1152, University of Evry, Evry, France
2 CNRS, Paris, France
3 MAP5, Department of Applied Mathematics, CNRS UMR-8145, University Paris Descartes, Paris, France
4 EBGM, Equipe de Bioinformatique Génomique et Moleculaire, INSERM UMRS-726, University Paris Diderot, Paris, France
5 MT*i*, Molecules Thérapeutique in silico, INSERM UMRS-973, University Paris Diderot, Paris, France
6 MIG, Mathématique Informatique et Genome, INRA UR-1077, Jouy-en-Josas, France
7 IBCP, Institut de Biologie et Chimie des Protéines, IFR 128, CNRS UMR 5086, University of Lyon 1, Lyon, France

✉ author email ✉ corresponding author email * Contributed equally

Algorithms for Molecular Biology 2010, **5**:15 doi:10.1186/1748-7188-5-15

The electronic version of this article is the complete one and can be found online at:
http://www.almob.org/content/5/1/15

Received: 22 September 2009
Accepted: 26 January 2010
Published: 26 January 2010