The symbolic package Regexpcount

Pierre Nicodème

Laboratory of statistics

CNRS - Genopole - Evry

and

Algorithms Project

INRIA - Rocquencourt

Some examples

INPUT

– $\Sigma = \{a, b\}$ alphabet

– aba pattern

– p_a, p_b letter probabilities

OUTPUT

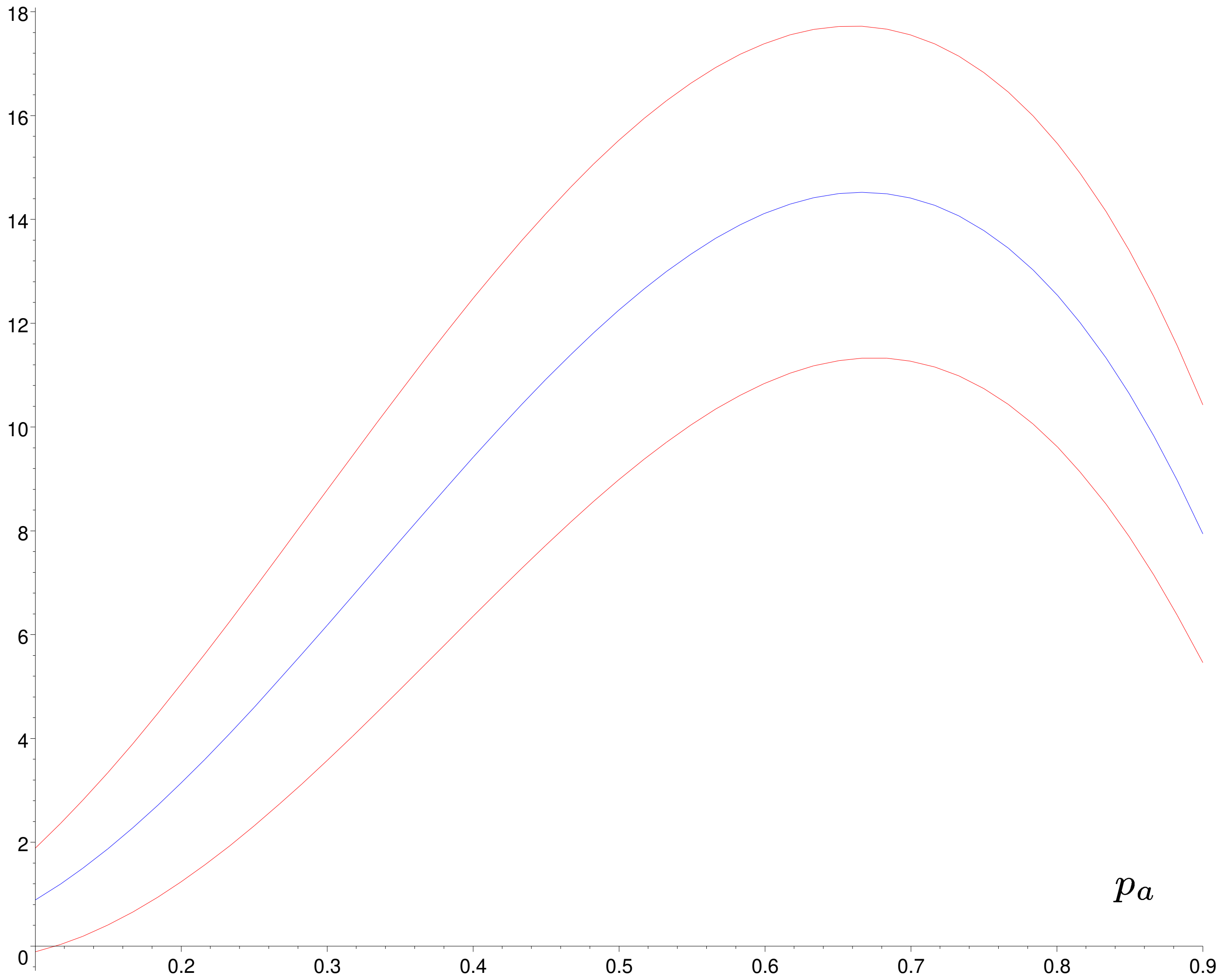
- number of matches

$$\frac{1 + p_a p_b (1-u) z^2}{1 - z + p_a p_b (1-u) z^2 - p_a p_b^2 (1-u) z^3} = \sum \pi_{n,k} u^k z^n$$

$\pi_{n,k}$ proba k occurrences of aba in texts of size n

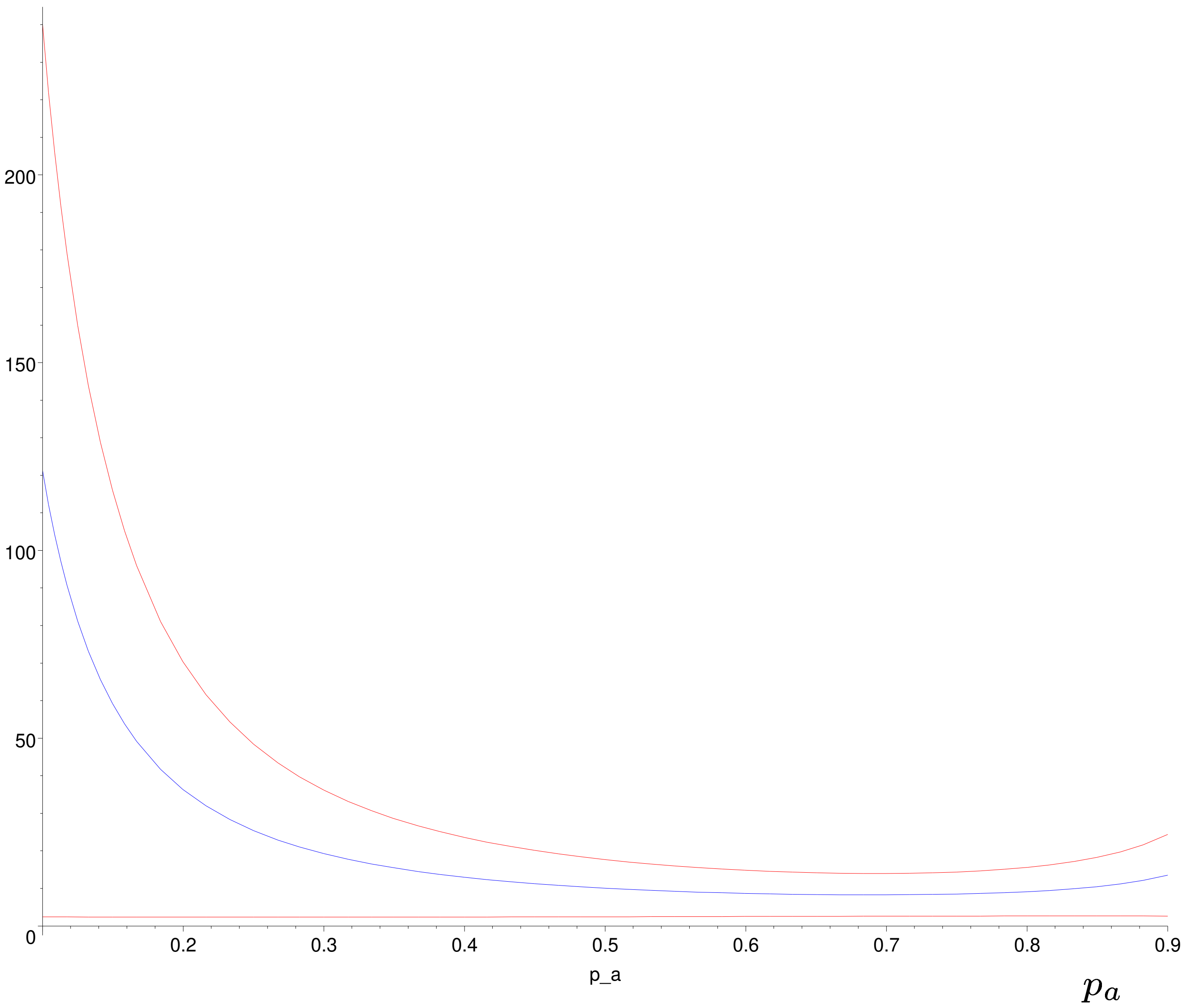
$\rightarrow$ expectation, variance

- waiting time for the motif ..., with error ...



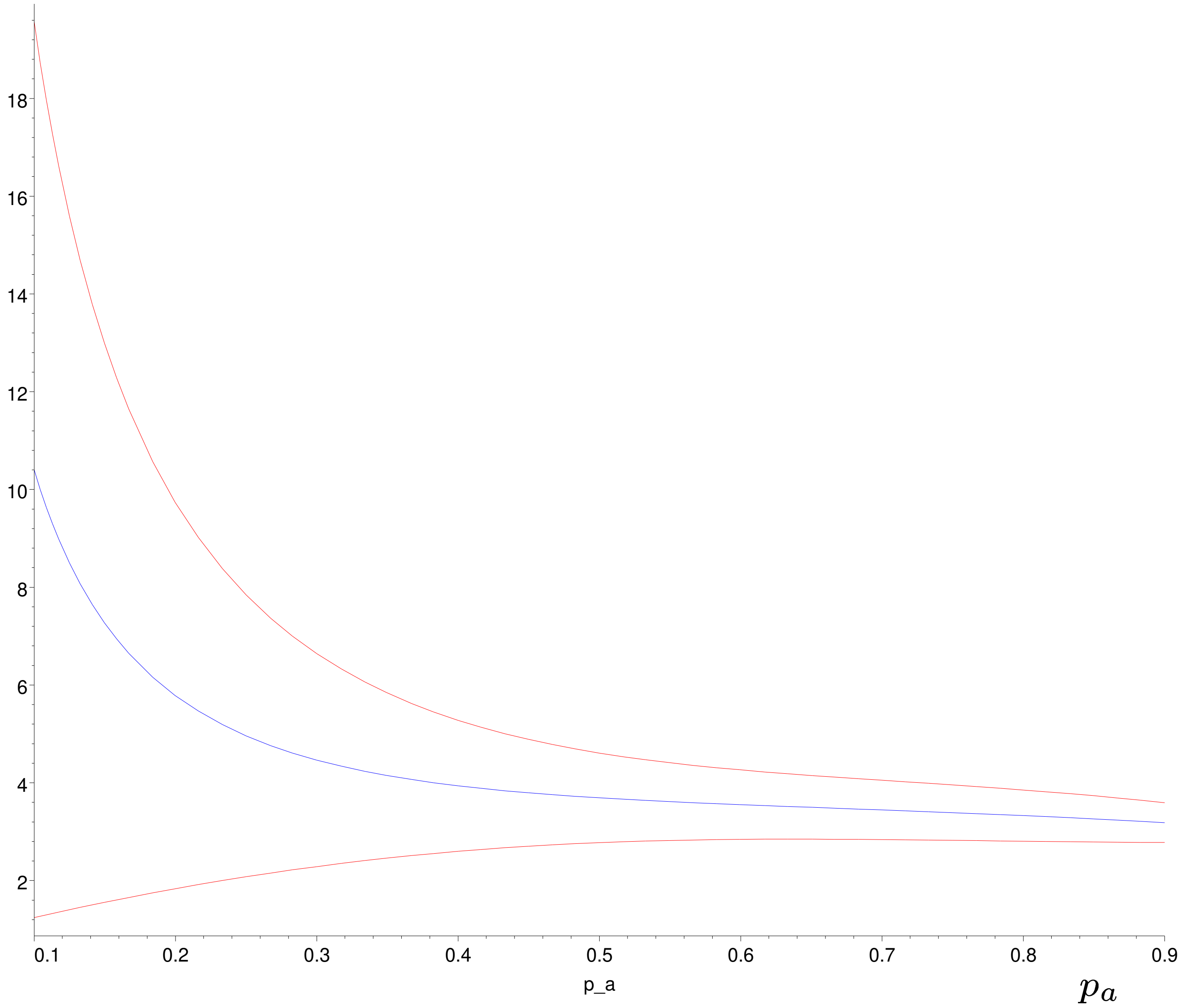
$[\mu, \mu+\sigma, \mu-\sigma]$ (μ expectation, σ standard dev.)

Number of matches of aba in Bernoulli texts



$[\mu, \mu + \sigma, \mu - \sigma]$ (μ expectation, σ standard dev.)

Waiting time for aba in Bernoulli texts



$[\mu, \mu + \sigma, \mu - \sigma]$ (μ expectation, σ standard dev.)

Waiting time for $S_1(aba)$ in Bernoulli texts

$S_1(aba)$: matches with aba with ≤ 1 substitution

Applications and extensions

Problems

- number of matches in random texts
  * of 1 motif (calibration of Prosite motifs in proteoms)
  * of several motifs (constrained occurrences of words in DNA)
- (constrained) waiting time (r-scans of motifs in DNA

Models

- uniform and non-uniform Bernoulli, Markov
- matches with or without errors (insertions, substitutions, deletions)

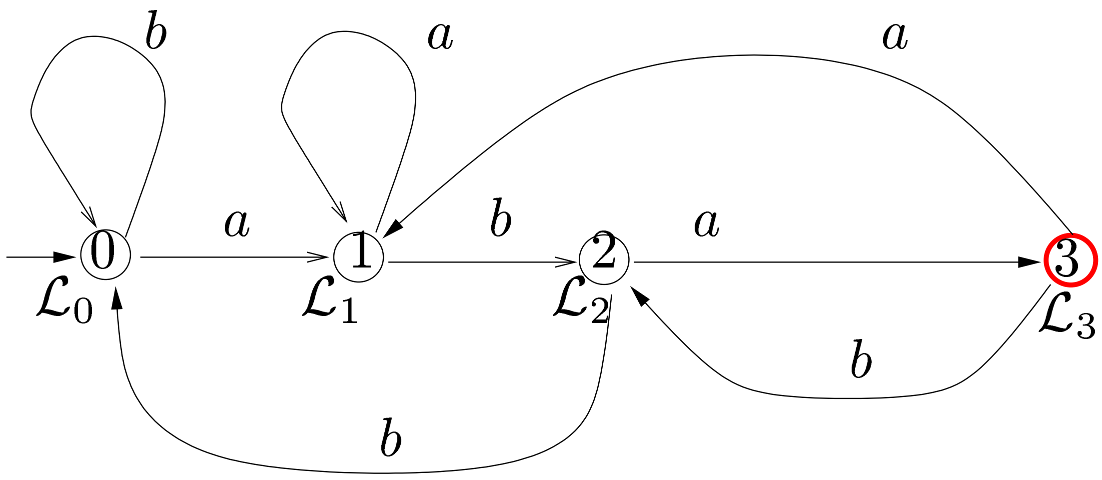
Tools

- automata
- generating functions
- singularity analysis

Number of matches with 1 motif

$$\Sigma = \{a, b\} \quad R = aba, \quad E = \Sigma^* R = \Sigma^* aba$$

$$aabb\,aba\bullet bbabb\,aba\bullet aa\,aba\bullet bbbb$$



Chomsky-Schützenberger

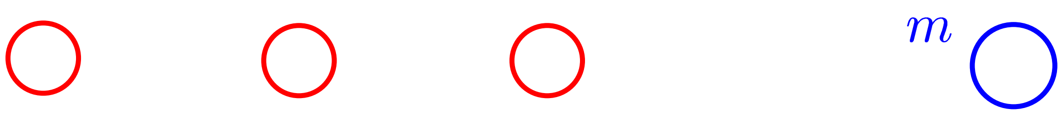
$$\mathcal{L}_0 = a\mathcal{L}_1 + b\mathcal{L}_0, \qquad \mathcal{L}_1 = b\mathcal{L}_2 + a\mathcal{L}_1,$$

$$\mathcal{L}_2 = a\mathcal{L}_3 + b\mathcal{L}_0, \qquad \mathcal{L}_3 = a\mathcal{L}_1 + b\mathcal{L}_2 + \epsilon,$$

$$L_0 = aL_1 + bL_0, \qquad L_1 = bL_2 + aL_1,$$

$$L_2 = aL_3 + bL_0, \qquad L_3 = aL_1 + bL_2 + 1.$$

$$L_0(a, b) = \frac{a^2 b}{1 - a - b}$$



marking: $\Sigma' = \Sigma \cup \{m\}$

$$\bullet \quad \to \quad m$$

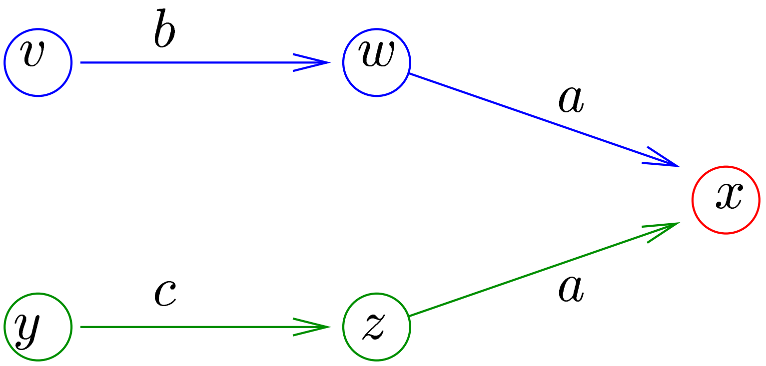
$$aaaabambamaaabamaa$$

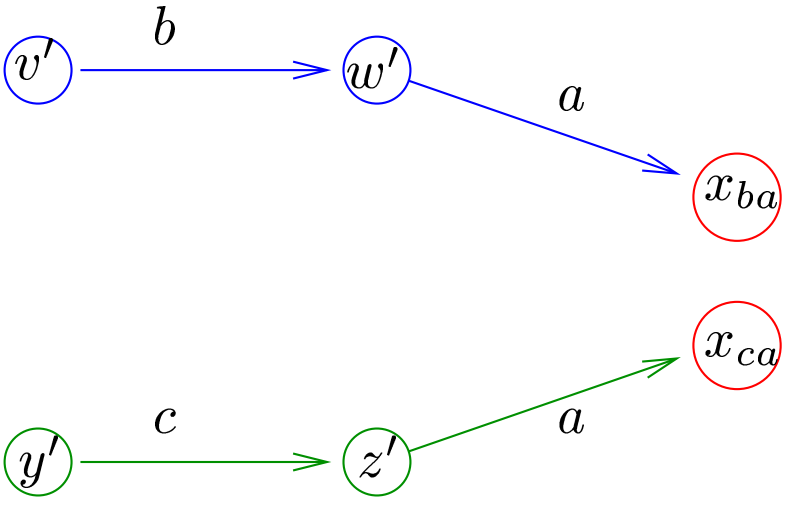
$$L_0 = aL_1 + bL_0 + 1, \qquad L_1 = bL_2 + aL_1 + 1,$$
$$L_2 = amL_3 + bL_0 + 1, \quad L_3 = aL_1 + bL_2 + 1.$$

$$L(a,b,m) = \frac{1 + ab(1-m)}{1 - a - b + ab(1-m) - ab^2(1-m)},$$
$$F(z,u) = L(p_a z, p_b z, u) = \sum \pi_{n,k} u^k z^n$$

Algorithm Bernoulli to Markov

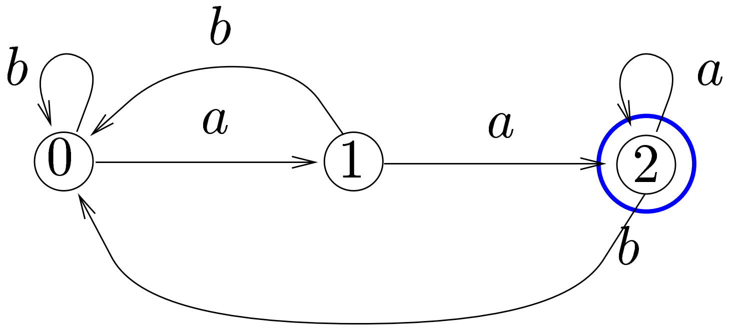
Bernoulli automaton

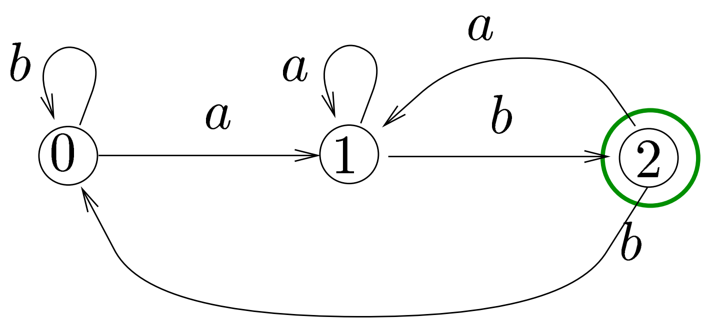


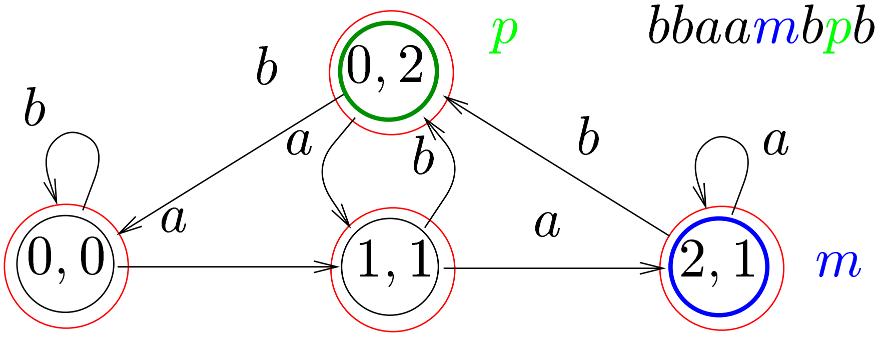


Markov (order 2) automaton

Cartesian product of automata







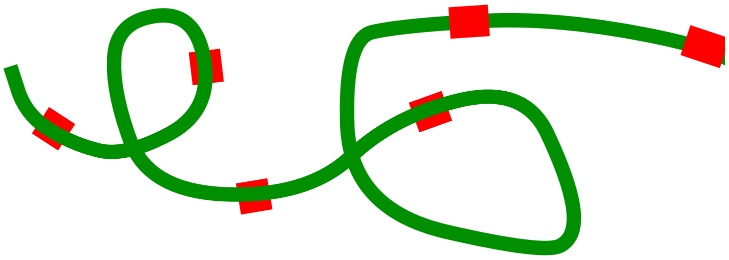
$bbaambpbbbabpaam$

$$F(a,b,m,p) = \frac{1+a-ap}{1-b-ap-abm+abp}$$

$$\mathrm{gf}_{\mathrm{bernoulli}}(z,u,v) = F(z/2,z/2,u,v)$$

$$= \pi_{ijn} u^i v^j z^n = 2\frac{2+z-vz}{4-2z-2vz-uz^2+vz^2}$$

Constrained words in H. Influenzae



gxtggtgg overrepresented protecting motif χ

length of H. Influenzae = 1830140

$\mu(\chi) = 56.26 \qquad \sigma(\chi) = 7.59$

observed(χ)=223

What about w = tggtgggc ?

$\mu(w), \sigma(w)$ knowing that observed(χ)=223 ?

Steps of the computation

Compute

1. a marked automaton B for $\Sigma^*\chi \times \Sigma^* w$
2. a markov automaton M (from B)
3. the probability g.f. $F(z,u,v) = \sum \pi_{i,j,n} u^i v^j z^n$

   ($\pi_{i,j,n}$ = Proba. text size n distrib. like H. Influenzae

   has i occ. χ, j occ. w)
4. the solution β of $\phi(1830140, \alpha) = \text{observed}(\chi)$

$$\text{with} \quad \phi(n,\alpha) = \frac{[z^n] \left. \frac{\partial F(z,\alpha u,1)}{\partial u} \right|_{u=1}}{[z^n] F(z,\alpha,1)}$$

5 · $$\mu(w) = \frac{[z^{1830140}] \left. \frac{\partial F(z,\beta,v)}{\partial v} \right|_{v=1}}{[z^{1830140}] F(z,\beta,1)}$$

6 · $$Moment2(w) = \frac{[z^{1830140}] \left. \frac{\partial}{\partial v} v \frac{\partial F(z,\beta,v)}{\partial v} \right|_{v=1}}{[z^{1830140}] F(z,\beta,1)}$$

steps 5 and 6 are computed by singularity analysis

Demo maple 1

```
> Hgen;
```

$$\{c=Atom,\ x=\mathrm{Union}(a,c,g,t),\ \chi=\mathrm{Prod}(g,x,t,g,g,t,g,g),\ t=Atom,\ g=Atom,\ a=Atom\}$$

```
> word;
```

$$\{\mathrm{W}=\mathrm{Prod}(t,g,g,t,g,g,g,c)\}$$

```
> T:=time():
> auto:=regexptomatchesgram(Hgen union word,CHI,
                            [[chi,m,'overlap'],[W,p,'overlap']]):
  markovauto:=grammarkov(auto,CHI,1,'k');
```

$markovauto := \{w7_g=\mathrm{Union}(\mathrm{Prod}(k_{g,g},w1_g),\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,t},w2_t),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{E}),$

$\mathrm{X}=\mathrm{Union}(\mathrm{Prod}(k_a,w21_a),\mathrm{Prod}(k_c,w21_c),\mathrm{Prod}(k_t,w11_t),\mathrm{Prod}(k_g,w13_g),\mathrm{E}),$

$w6_c=\mathrm{Union}(\mathrm{Prod}(k_{c,c},w21_c),\mathrm{Prod}(k_{c,a},w21_a),\mathrm{Prod}(k_{c,t},w9_t),\mathrm{Prod}(k_{c,g},w13_g),\mathrm{E}),$

$w18_c=\mathrm{Union}(\mathrm{Prod}(k_{c,c},w21_c),\mathrm{Prod}(k_{c,a},w21_a),\mathrm{Prod}(k_{c,t},w9_t),\mathrm{Prod}(k_{c,g},w13_g),\mathrm{E}),\ k_{a,c}=Atom,\ k_{t,t}=Atom,$

$k_{c,c}=Atom,\ k_{c,t}=Atom,\ k_{t,a}=Atom,\ k_{t,c}=Atom,\ k_{g,g}=Atom,\ k_{a,a}=Atom,\ k_{c,a}=Atom,\ k_{a,t}=Atom,\ k_{g,c}=Atom,$

$k_{t,g}=Atom,\ k_{a,g}=Atom,\ k_{c,g}=Atom,\ k_c=Atom,\ k_{g,t}=Atom,\ k_{g,a}=Atom,\ k_t=Atom,\ k_g=Atom,\ k_a=Atom,\ p=\mathrm{E},$

$m=\mathrm{E},\ w3_g=\mathrm{Union}(\mathrm{Prod}(k_{g,g},w1_g),\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,t},w2_t),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{E}),$

$w13_g=\mathrm{Union}(\mathrm{Prod}(k_{g,t},w12_t),\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{Prod}(k_{g,g},w14_g),\mathrm{E}),$

$w12_t=\mathrm{Union}(\mathrm{Prod}(k_{t,a},w21_a),\mathrm{Prod}(k_{t,c},w21_c),\mathrm{Prod}(k_{t,g},w19_g),\mathrm{Prod}(k_{t,t},w9_t),\mathrm{E}),$

$w2_t=\mathrm{Union}(\mathrm{Prod}(k_{t,a},w21_a),\mathrm{Prod}(k_{t,c},w21_c),\mathrm{Prod}(k_{t,g},w8_g),\mathrm{Prod}(k_{t,t},w9_t),\mathrm{E}),$

$w1_g=\mathrm{Union}(\mathrm{Prod}(k_{g,a},w18_a),\mathrm{Prod}(k_{g,g},w14_g),\mathrm{Prod}(k_{g,c},p,w6_c),\mathrm{Prod}(k_{g,t},w17_t),\mathrm{E}),$

$w11_t=\mathrm{Union}(\mathrm{Prod}(k_{t,a},w21_a),\mathrm{Prod}(k_{t,c},w21_c),\mathrm{Prod}(k_{t,g},w19_g),\mathrm{Prod}(k_{t,t},w11_t),\mathrm{E}),$

$w9_t=\mathrm{Union}(\mathrm{Prod}(k_{t,a},w21_a),\mathrm{Prod}(k_{t,c},w21_c),\mathrm{Prod}(k_{t,t},w11_t),\mathrm{Prod}(k_{t,g},w10_g),\mathrm{E}),$

$w10_g=\mathrm{Union}(\mathrm{Prod}(k_{g,t},w12_t),\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{Prod}(k_{g,g},w4_g),\mathrm{E}),$

$w14_g=\mathrm{Union}(\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{Prod}(k_{g,g},w14_g),\mathrm{Prod}(k_{g,t},w17_t),\mathrm{E}),$

$w15_t=\mathrm{Union}(\mathrm{Prod}(k_{t,a},w21_a),\mathrm{Prod}(k_{t,c},w21_c),\mathrm{Prod}(k_{t,t},w9_t),\mathrm{Prod}(k_{t,g},w5_g),\mathrm{E}),$

$w16_g=\mathrm{Union}(\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{Prod}(k_{g,g},w14_g),\mathrm{Prod}(k_{g,t},w15_t),\mathrm{E}),$

$w5_g=\mathrm{Union}(\mathrm{Prod}(k_{g,t},w12_t),\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{Prod}(k_{g,g},w3_g),\mathrm{E}),$

$w18_a=\mathrm{Union}(\mathrm{Prod}(k_{a,a},w21_a),\mathrm{Prod}(k_{a,c},w21_c),\mathrm{Prod}(k_{a,t},w9_t),\mathrm{Prod}(k_{a,g},w13_g),\mathrm{E}),$

$w4_g=\mathrm{Union}(\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,t},w2_t),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{Prod}(k_{g,g},w14_g),\mathrm{E}),$

$w21_c=\mathrm{Union}(\mathrm{Prod}(k_{c,c},w21_c),\mathrm{Prod}(k_{c,a},w21_a),\mathrm{Prod}(k_{c,g},w13_g),\mathrm{Prod}(k_{c,t},w11_t),\mathrm{E}),$

$w21_a=\mathrm{Union}(\mathrm{Prod}(k_{a,a},w21_a),\mathrm{Prod}(k_{a,c},w21_c),\mathrm{Prod}(k_{a,t},w11_t),\mathrm{Prod}(k_{a,g},w13_g),\mathrm{E}),$

$w19_g=\mathrm{Union}(\mathrm{Prod}(k_{g,t},w12_t),\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{Prod}(k_{g,g},w16_g),\mathrm{E}),$

$w8_g=\mathrm{Union}(\mathrm{Prod}(k_{g,t},w12_t),\mathrm{Prod}(k_{g,c},w18_c),\mathrm{Prod}(k_{g,a},w18_a),\mathrm{Prod}(k_{g,g},m,w7_g),\mathrm{E}),$

$w17_t=\mathrm{Union}(\mathrm{Prod}(k_{t,a},w21_a),\mathrm{Prod}(k_{t,c},w21_c),\mathrm{Prod}(k_{t,t},w9_t),\mathrm{Prod}(k_{t,g},w10_g),\mathrm{E})\}$

```
> tauto:=time()-T; T:=time():
```

$$tauto := .464$$

```
>
```

Demo maple 2

```
> Digits:=30:
> T:=time(): beta:=fsolve(phi,2..5); time()-T;
```

$$\beta := 3.71500260619846885584814600795$$

$$1.308$$

sumofcoeff = [z^n] F(z, beta,1) - Sum of coeff of probability g.f. must sum up to 1

```
> sumofcoeff:=evalcoeff(denphi,beta,1830146,1);
```

$$sumofcoeff := .330376940523924778854588001899\ 10^{69}$$

```
> Digits:=10:
> adnw;
```

$$\left[\left[\frac{186146}{609969}, k_a\right], \left[\frac{336446}{1829907}, k_c\right], \left[\frac{361676}{1829907}, k_g\right], \left[\frac{573347}{1829907}, k_t\right], \left[\frac{213715}{558438}, k_{a,a}\right], \left[\frac{97395}{361676}, k_{g,a}\right], \left[\frac{115361}{336446}, k_{c,a}\right],\right.$$
$$\left[\frac{131957}{573347}, k_{t,a}\right], \left[\frac{45752}{279219}, k_{a,g}\right], \left[\frac{71409}{361676}, k_{g,g}\right], \left[\frac{36261}{168223}, k_{c,g}\right], \left[\frac{126249}{573347}, k_{t,g}\right], \left[\frac{86381}{558438}, k_{a,c}\right], \left[\frac{47765}{180838}, k_{g,c}\right],$$
$$\left.\left[\frac{63061}{336446}, k_{c,c}\right], \left[\frac{91471}{573347}, k_{t,c}\right], \left[\frac{83419}{279219}, k_{a,t}\right], \left[\frac{6953}{25834}, k_{g,t}\right], \left[\frac{42751}{168223}, k_{c,t}\right], \left[\frac{223670}{573347}, k_{t,t}\right], [u, m], [v, p]\right]$$

```
> T:=time():
> neq:=Chomsky_Schutzenberger(markovauto,adnw):
> mvgf:=subs(solve({op(neq)},{seq(op(1,i),i=neq)}),CHI):
> tsolve:=time()-T;T:=time():
```

$$tsolve := 2.451$$

```
> Digits:=30:
> expstd(mvgf,beta,sumofcoeff,1830146);
```

"Conditionned Expectation", 23.5181553330401091613440176359

"Conditionned Standard Deviation", 4.84763947870020039447815146172

```
> tstats:=time()-T;T:=time():
```

$$tstats := .357$$

```
> expstd(mvgf,1,1,1830146);
```

"Expectation", 15.2124052543474542605033150727

"Standard Deviation", 3.90005838002946105422993577166

```
> tauto+tsolve+tstats;
```

$$3.272$$

```
> bneq:=Chomsky_Schutzenberger(auto,berw):
> bern_mvgf:=subs(solve({op(bneq)},{seq(op(1,i),i=bneq)}),CHI);
```

$$bern\_mvgf := -64\,(z^{11}\,u\,v + 64\,z^8\,u\,v - z^{11}\,u - 64\,z^8\,u + 1024\,z^7\,u + 1024\,z^6\,u + 65536\,z^3\,u - z^{11}\,v - 64\,z^8\,v + z^{11} + 64\,z^8 - 1024\,z^7 - 1024\,z^6 - 65536\,z^3 - 4194304) \Big/ (-4194304\,z^4 - 268435456\,z + 4194304\,z^4\,u + 65536\,z^6 + 53248\,z^8\,u - 4096\,z^9\,v + 4194304\,z^3 + 1024\,z^{10}\,v + 4096\,z^9\,u\,v - 4096\,z^9\,u - 1024\,z^{10}\,u\,v - 4096\,z^8\,u\,v - 49152\,z^8 + 4096\,z^9 - 1024\,z^{10} + 16\,z^{13}\,u + 64\,z^{12} - 16\,z^{13} + z^{15} - 64\,z^{12}\,u - z^{15}\,u + 64\,z^{12}\,u\,v + z^{15}\,u\,v - 16\,z^{13}\,u\,v - 64\,z^{12}\,v + 16\,z^{13}\,v - z^{15}\,v - 4194304\,z^3\,u + 268435456 - 65536\,z^6\,u + 1024\,z^{10}\,u)$$

```
>
```

Availability

http://algo.inria.fr/libraries/software.html