Motif Statistics

Motivation: Molecular biology tries to establish relations between chemical form and biological function.

One important "chemical form": sequence data (DNA, RNA, proteins).

Task: Discern *signal* from *noise*.

Here: *Motifs* (simple regular expressions) representing families of similar (due to common ancestors) sequences;

Example (protein encoding):
$[LIVM](2) - x - D - D - x(2,4) - D - x(4) - R - R - [GH]$

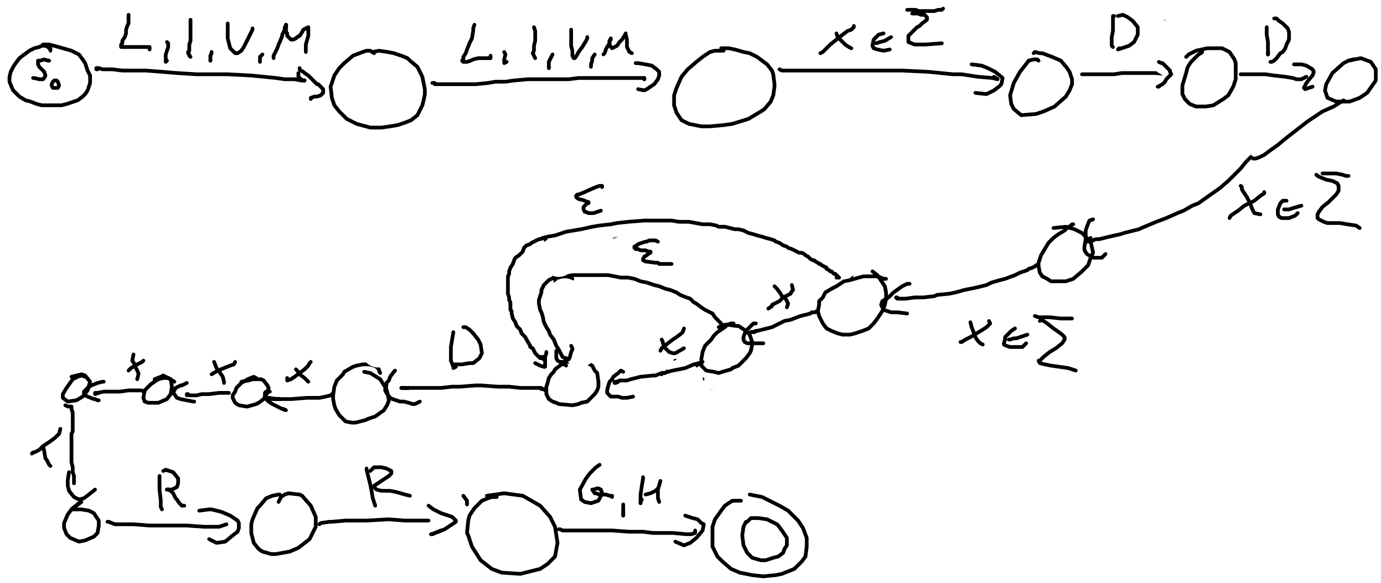
What is the expected number of occurrences of a motif in a random text?
Representation of motifs via finite automata (equivalent to so-called regular expressions):

Definition

A deterministic finite automaton (DFA) A *is given by a tuple* $A = (S, \Sigma, s_0, \delta, F)$ *with*

- S *a finite set of states;*
- Σ *a finite set of symbols (input alphabet);*
- $s_0 \in S$ *the initial state;*
- $\delta : (S \times \Sigma) \mapsto S$ *the transition function;*
- $F \subseteq S$ *set of accepting states.*

Example: Automaton for
$[LIVM](2) - x - D - D - x(2,4) - D - x(4) - R - R - [GH]$



Notation: For DFA A we denote by $\mathcal{L}(A)$ the set of words accepted by A (language).

Remarks:

1. Motifs always describe a finite set of finite strings;
2. the language accepted by a DFA is not necessarily finite (but the accepted strings are);
3. the methods we will consider here apply to every DFA thus can also be used in connection with infinite languages.

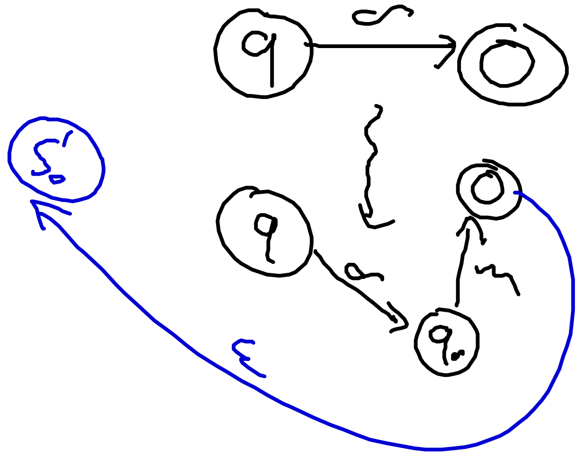
Plan of analysis:

1. Design finite automaton that "reads" all words over Σ but *signals* occurrence of motif;
2. translate automaton into generating function;
3. apply techniques from analytic combinatorics to determine expected number of occurrences (and more).

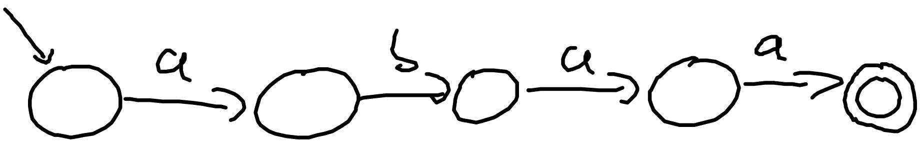
Step 1:

- Given DFA A, modify it to accept $\Sigma^* \cdot \mathcal{L}(A)$; let $A' = (S', \Sigma, s_0', \delta', F')$ be the resulting automaton.
- To mark all matches introduce new symbol m, setting $\Sigma' := \Sigma \cup \{m\}$.
- For all $q \in S'$ and all $\sigma \in \Sigma$ with $\delta(q, \sigma) = f \in F'$ create new state q_σ in S' and set $\delta'(q, \sigma) := q_\sigma$ and $\delta'(q_\sigma, m) := f$.
- For all $f \in F$ set $\delta'(f, \varepsilon) := s_0'$ (restart for next occurrence).

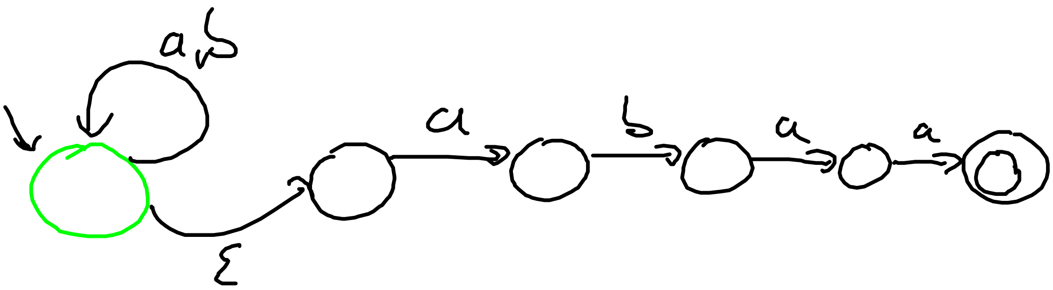
Example:



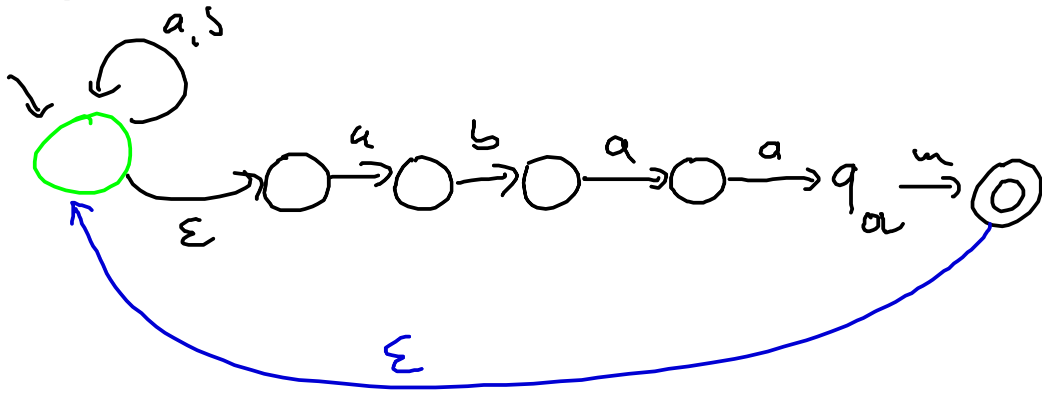
Example: $\Sigma = \{a, b\}$, $\mathcal{L}(A) = \{abaa\}$



$\Sigma^* \cdot \mathcal{L}(A)$:

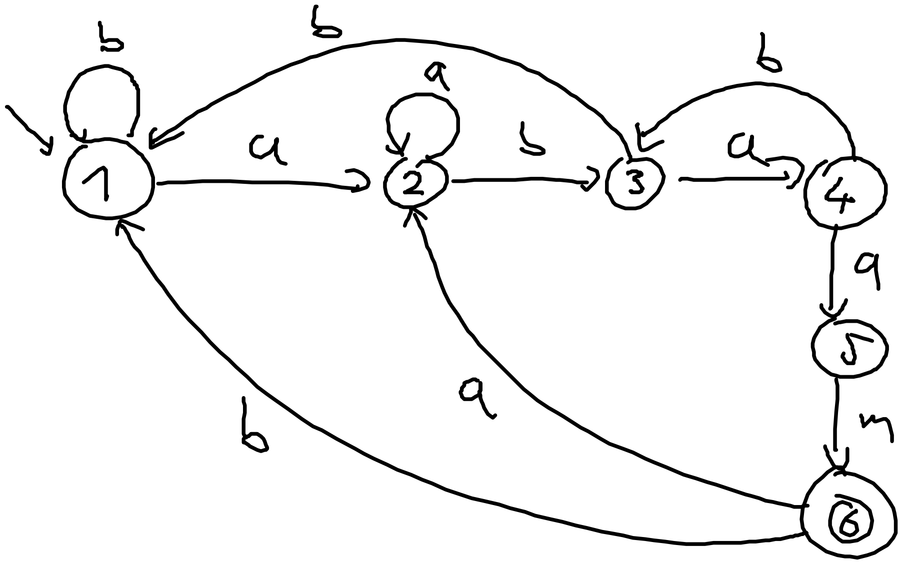


• New state q_a with



• restart

DEA :



Step 2: Here we can resort on Chomsky and Schützenberger: Assuming $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_r = m\}$

- for each state s_i of DFA introduce (ordinary) GF S_i;
- for each state s_i we have $S_i(z_1, \dots, z_r) = (1+) \bigoplus_{\substack{\sigma_j \in \Sigma \\ \delta(s_i, \sigma_j) = s_k}} z_j S_k(z_1, \dots, z_r)$ where term $1+$ iff $s_i \in F$.

$a_0, a_1, a_2, \dots$

$$\downarrow$$

$$\sum_{i \geq 0} a_i z^i$$

$$\Big\downarrow [z^n]$$

a_n

Remark: The resulting GF S_0 is rational.

Example:

$$S_1(\overset{a}{z_1}, \overset{b}{z_2}, \overset{m}{z_3}) = z_1 \cdot S_2(z_1, z_2, z_3) + z_2 \cdot S_1(z_1, z_2, z_3)$$

$$S_2(z_1,z_2,z_3) = z_1 \cdot S_2(z_1,z_2,z_3) + z_2 \cdot S_3(z_1,z_2,z_3)$$

$$S_3(z_1,z_2,z_3) = z_1 \cdot S_4 + z_2 \cdot S_1$$

$$\vdots \qquad S_6 = 1 + z_2 \cdot S_1 + z_1 \cdot S_2$$

$$S_2 = S_1 \left(z_1 + z_2 + z_1^2 (z_2 - 1) z_2 \right) + (1+S_1) z_1^3$$

$$\times z_2 z_3$$

$$S_1 = \frac{z_1^3 z_2 z_3}{1 - z_2 - z_1 (1 + z_1 z_2 (-1 + z_2 + z_1 z_3))}$$

Step 3: Now $P(z,u) := S_0(zp_1, zp_2, \ldots, zp_{r-1}, u)$ is the BGF with

- the coefficient at z^n being associated with all accepted words of length n (symbol m not contributing),
- assuming a BERNOULLI probability model (symbol σ_i shows up with probability p_i), and
- each occurrence of the pattern labeled by variable u.

Example:

Word = abaam $\qquad S_1 \xrightarrow{a} S_2 \xrightarrow{b} S_3 \xrightarrow{a} S_4 \xrightarrow{a} S_5 \xrightarrow{m} S_6$

$$S_1 = z_1 \cdot S_2 = z_1 \cdot z_2 \cdot S_3 = z_1 \cdot z_2 \cdot z_1 \cdot S_4$$

$$= z_1 z_2 z_1 z_1 S_5 = z_1 z_2 z_1 z_1 z_3 S_6$$

$$= z_1 z_2 z_1 z_1 \cdot z_3$$

subst. $\{$ $$z p_1 \, z p_2 \, z p_1 \, z p_1 \, u = z^4 \cdot u \cdot p_1^3 \, p_2$$

Assuming $p_1 = p_2 = \frac{1}{2}$ we get

$$S_1 = \frac{u \cdot z^4}{16 - 16z + 2z^3 + (1+u)z^4}$$

From $P(z, u)$ (for given motif (aka DFA)) we can easily compute

1. the probability of k occurrences in a text of length n;
2. the expected number of occurrences of the motif in a text of length n;
3. the corresponding variance;
4. the limiting distribution.

Example (for 2.):

$$\frac{\partial}{\partial u} P(z,u) \Big|_{u=1} = \underbrace{\frac{z^4(16 - 16z + 2z^3 - z^4)}{4(1-z)^2 (8+z^3)^2}}_{P_1(z)}$$

$$[z^n] \, P_1(z) \sim n \cdot 0.0030864 2\ldots$$