

COMPUTING ALIGNMENT SEED SENSITIVITY WITH PROBABILISTIC ARITHMETIC AUTOMATA

Inke Herms and Sven Rahmann

Bielefeld University, TU Dortmund

SEEDED ALIGNMENT

- Task: identification of homologous sequences
- heuristics as BLASTn for fast large-scale genomic sequence comparison
 - ① filter candidate sequences that match a short **seed**
 - ② investigate candidates by exact local alignment methods

SEEDED ALIGNMENT

- Task: identification of homologous sequences
- heuristics as BLASTn for fast large-scale genomic sequence comparison
 - ① filter candidate sequences that match a short **seed**
 - ② investigate candidates by exact local alignment methods
- quality crucially depends on the seed:
 - ▶ longer seed \longrightarrow lower sensitivity
 - ▶ shorter seed \longrightarrow lower specificity
- trade-off between search speed and sensitivity

ALIGNMENT REPRESENTATIVE STRING

Representation of random alignments: $\mathcal{A} \in \Sigma^*$

- ① $\Sigma = \{0, 1\}$ with 0 - mismatch and 1 - match

EXAMPLE

Query	G	C	G	A	A	T	G	C	C	T
Target	G	C	C	A	A	C	G	C	T	T
\mathcal{A}	1	1	0	1	1	0	1	1	0	1

ALIGNMENT REPRESENTATIVE STRING

Representation of random alignments: $\mathcal{A} \in \Sigma^*$

- ① $\Sigma = \{0, 1\}$ with 0 - mismatch and 1 - match

EXAMPLE

Query	G	C	G	A	A	T	G	C	C	T
Target	G	C	C	A	A	C	G	C	T	T
\mathcal{A}	1	1	0	1	1	0	1	1	0	1

- ② $\Sigma = \{0, 1, 2, 3\}$ allowing for **indels**

EXAMPLE

Query	G	C	-	A	A	T	G	C	C	T
Target	G	C	C	A	A	C	G	C	-	T
\mathcal{A}	1	1	2	1	1	0	1	1	3	1

ALIGNMENT REPRESENTATIVE STRING

Representation of random alignments: $\mathcal{A} \in \Sigma^*$

- 1 $\Sigma = \{0, 1\}$ with 0 - mismatch and 1 - match

EXAMPLE

Query	G	C	G	A	A	T	G	C	C	T
Target	G	C	C	A	A	C	G	C	T	T
\mathcal{A}	1	1	0	1	1	0	1	1	0	1

- 2 $\Sigma = \{0, 1, 2, 3\}$ allowing for **indels**

EXAMPLE

Query	G	C	-	A	A	T	G	C	C	T
Target	G	C	C	A	A	C	G	C	-	T
\mathcal{A}	1	1	2	1	1	0	1	1	3	1

HOMOLOGY MODEL

- describe random alignments with known degree of similarity
- model **representative string** \mathcal{A} rather than random sequences

HOMOLOGY MODEL

- describe random alignments with known degree of similarity
- model **representative string** \mathcal{A} rather than random sequences
 - ▶ $\Sigma = \{0, 1\}$: set match probability p ,
mismatch probability $q = 1 - p$

HOMOLOGY MODEL

- describe random alignments with known degree of similarity
- model **representative string \mathcal{A}** rather than random sequences
 - ▶ $\Sigma = \{0, 1\}$: set match probability p ,
mismatch probability $q = 1 - p$
 - ▶ general Σ : define Markov chain (Σ, P, p^0) , e.g.

$$P = \begin{array}{c} \begin{array}{cccc} & 0 & 1 & 2 & 3 \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} p_0 & p_1 & p_g & p_g \\ p_0 & p_1 & p_g & p_g \\ p_0^* & p_1^* & p_g & 0 \\ p_0^* & p_1^* & 0 & p_g \end{pmatrix} \end{array} \end{array}$$

according to D.Mak, Y.Gelfand, and G.Benson 2006

SEED MODEL

SEED

A **seed** $\pi = \pi[0]\pi[1] \dots \pi[L - 1]$ is a string over an alphabet of “care”-positions (1) and “don’t care”-positions (e.g. *, ?) with

SEED MODEL

SEED

A **seed** $\pi = \pi[0]\pi[1] \dots \pi[L - 1]$ is a string over an alphabet of “care”-positions (1) and “don’t care”-positions (e.g. *, ?) with

- restriction $\pi[0] = \pi[L - 1] = 1$,
- **length** $L = |\pi|$,
- **weight** $\omega = \#$ of “care”-positions,

SEED MODEL

SEED

A **seed** $\pi = \pi[0]\pi[1] \dots \pi[L - 1]$ is a string over an alphabet of “care”-positions (1) and “don’t care”-positions (e.g. *, ?) with

- restriction $\pi[0] = \pi[L - 1] = 1$,
- **length** $L = |\pi|$,
- **weight** $\omega = \#$ of “care”-positions,
- generalized string $G(\pi) = A_0 \dots A_{L-1}$, where A_i : set of potential characters specified by $\pi[i]$, e.g. $G(1 * 1 * 1) = [1][01][1][01][1]$
- **pattern set** $\mathcal{PS}(\pi)$: all words matching $G(\pi)$, e.g.
 $\mathcal{PS}(\pi) = \{10101, 10111, 11101, 11111\}$

SEED MODEL

- consecutive seed (W. Pearson and D. Lipman 1988, S.F. Altschul et al. 1990)

SEED MODEL

- consecutive seed (W. Pearson and D. Lipman 1988, S.F. Altschul et al. 1990)
- spaced seed (B. Ma et al. 2002, J. Buhler et al. 2003, B. Brejová et al. 2004, K.P. Choi et al. 2004)

SPACED SEED

$\pi \in \{1, *\}^L$, where $* \mapsto [0, 1]$

Example: $\pi = 1 * 1 * 1$

$\leftrightarrow \mathcal{PS}(\pi) = \{10101, 10111, 11101, 11111\}$

SEED MODEL

- consecutive seed (W. Pearson and D. Lipman 1988, S.F. Altschul et al. 1990)
- spaced seed (B. Ma et al. 2002, J. Buhler et al. 2003, B. Brejová et al. 2004, K.P. Choi et al. 2004)
- indel seed (D. Mak et al. 2006)

INDEL SEED

$\Sigma = \{0, 1, 2, 3\}$, $\pi \in \{1, *, ?\}^L$ where $* \mapsto [0, 1]$ and $? \mapsto [\epsilon, 0, 1, 2, 3]$

Example: $\pi = 1 * 1 ? 1$

$\leftrightarrow \mathcal{PS}(\pi) = \{1011, 1111, 10101, 10111, 10121, 10131,$
 $11101, 11111, 11121, 11131\}$

SEED MODEL

- consecutive seed (W. Pearson and D. Lipman 1988, S.F. Altschul et al. 1990)
- spaced seed (B. Ma et al. 2002, J. Buhler et al. 2003, B. Brejová et al. 2004, K.P. Choi et al. 2004)
- indel seed (D. Mak et al. 2006)
- multiple spaced seed $\Pi = \{\pi_1, \dots, \pi_m\}$ (M. Li et al. 2004, G. Kucherov et al. 2005, Y. Sun et al. 2005)

SEED MATCHING

spaced seed

1010111110
1*11*1

indel seed

1112011031
11?*11

HIT POSITION

π hits \mathcal{A} at position n if $\exists M \in \mathcal{PS}(\pi)$ s.t. $\mathcal{A}[n - |M| + 1, n] = M$.

A multiple seed Π matches \mathcal{A} if at least one component does.

QUESTIONS

GIVEN

- homology model
- target length t of alignments
- seed π
- maximal match number K

DEFINE

$V_t(\pi) = \#$ of matches of π

QUESTIONS

GIVEN

- homology model
- target length t of alignments
- seed π
- maximal match number K

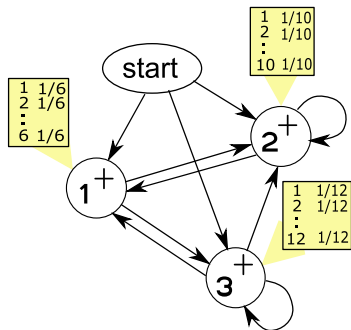
DEFINE

$V_t(\pi) = \#$ of matches of π

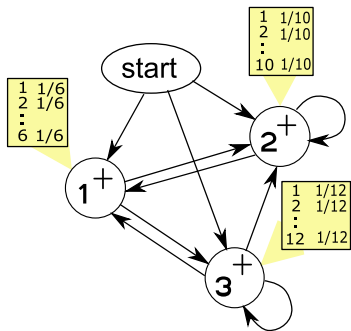
WANTED

- 1 Sensitivity: $\mathbb{P}(V_t(\pi) \geq 1)$
- 2 Hit distribution: $\mathbb{P}(V_t(\pi) = k)$, $k = 0, \dots, K$
(overlapping and non-overlapping hits)

PROBABILISTIC ARITHMETIC AUTOMATA (PAA)



PROBABILISTIC ARITHMETIC AUTOMATA (PAA)

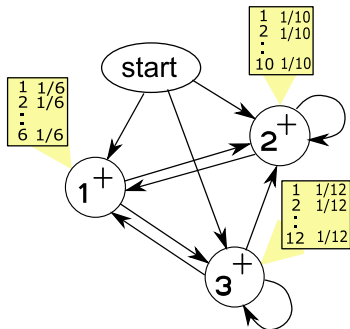


3 2 2 1 3 1

COMPONENTS

- 1 **Markov chain** generates a string over a given alphabet
- 2 in each state a weight is emitted
- 3 **arithmetic operations** on the emissions

PROBABILISTIC ARITHMETIC AUTOMATA (PAA)

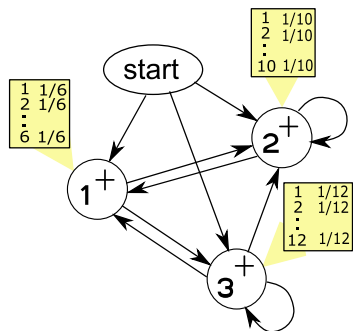


COMPONENTS

- 1 **Markov chain** generates a string over a given alphabet
- 2 in each state a weight is **emitted**
- 3 **arithmetic operations** on the emissions

3	2	2	1	3	1
6	2	9	6	4	1

PROBABILISTIC ARITHMETIC AUTOMATA (PAA)



COMPONENTS

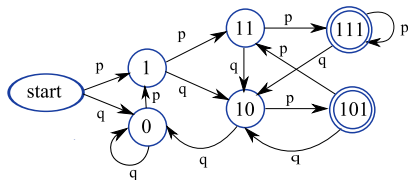
- 1 **Markov chain** generates a string over a given alphabet
- 2 in each state a weight is **emitted**
- 3 **arithmetic operations** on the emissions

3	2	2	1	3	1
6	+ 2	+ 9	+ 6	+ 4	+ 1
6	8	17	23	27	28

PAA LAYOUT

Seed $\pi = 1 * 1$

$$\Sigma = \{0, 1\}, p^0 = (q, p), P = \begin{pmatrix} q & p \\ q & p \end{pmatrix}$$



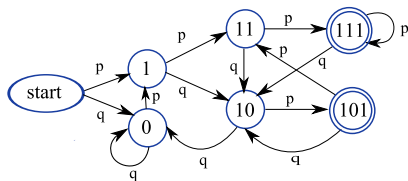
PAA counting overlapping occurrences of $1 * 1$ in a random alignment

- goal: count occurrences of seed patterns in a random alignment
- probabilistic version of Aho-Corasick automaton (T. Marschall, S. Rahmann 2008)
- states $Q = \{start\} \cup \{\sigma \in \Sigma\} \cup \text{prefixes}(\mathcal{PS}(\pi))$
- transition from state u to state v iff v is maximal suffix of $u\sigma$ for a $\sigma \in \Sigma$

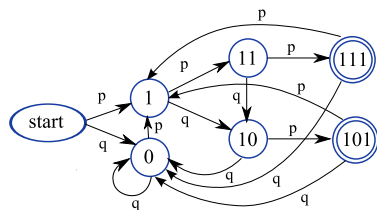
PAA LAYOUT

Seed $\pi = 1 * 1$

$$\Sigma = \{0, 1\}, p^0 = (q, p), P = \begin{pmatrix} q & p \\ q & p \end{pmatrix}$$

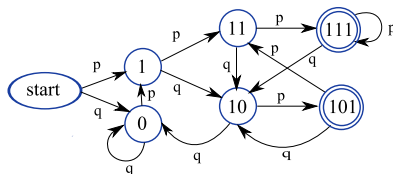


PAA counting overlapping occurrences of $1 * 1$ in a random alignment



PAA counting non-overlapping occurrences of $1 * 1$ in a random alignment

PAA VARIABLES AND PROCESSES



CHARACTERISTICS

- ① state process $(Y_i)_{i \in \mathbb{N}_0}$ with $Y_0 = \text{start}$
- ② emissions: number $C(q)$ of patterns that end in state q
- ③ V_t : accumulated number of seed hits in alignment of length t :

$$V_0 \equiv 0$$

$$V_l = V_{l-1} + C(Y_l)$$

SENSITIVITY AND HIT DISTRIBUTION

THE SENSITIVITY OF π

... is the probability to hit a random alignment of length t at least once, that is

$$S(\pi, t) = \mathbb{P}(V_t \geq 1) = 1 - \mathbb{P}(V_t = 0)$$

SENSITIVITY AND HIT DISTRIBUTION

THE SENSITIVITY OF π

... is the probability to hit a random alignment of length t at least once, that is

$$S(\pi, t) = \mathbb{P}(V_t \geq 1) = 1 - \mathbb{P}(V_t = 0)$$

THE HIT DISTRIBUTION $\mathcal{L}(V_t)$

... is given by

$$\mathbb{P}(V_t = k) = \mathbb{P}(\{\mathcal{A} : |\mathcal{A}| = t, V_t = k\}).$$

SENSITIVITY AND HIT DISTRIBUTION

MARGINALIZATION

$$\mathbb{P}(V_t = k) = \sum_{q \in Q} \mathbb{P}(Y_t = q, V_t = k) = \sum_{q \in Q} h_q^t(k)$$

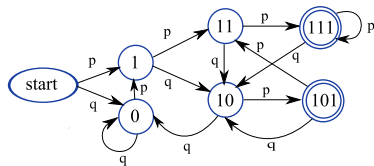
SENSITIVITY AND HIT DISTRIBUTION

MARGINALIZATION

$$\mathbb{P}(V_t = k) = \sum_{q \in Q} \mathbb{P}(Y_t = q, V_t = k) = \sum_{q \in Q} h_q^t(k)$$

RECURRENCE

$$h_q^t(k) = \sum_{q' \in Q} h_{q'}^{t-1}(k - C(q)) T_{q'q}$$



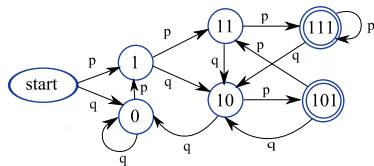
SENSITIVITY AND HIT DISTRIBUTION

MARGINALIZATION

$$\mathbb{P}(V_t = k) = \sum_{q \in Q} \mathbb{P}(Y_t = q, V_t = k) = \sum_{q \in Q} h_q^t(k)$$

RECURRENCE

$$h_q^t(k) = \sum_{q' \in Q} h_{q'}^{t-1}(k - C(q)) T_{q'q}$$



Symbolic representation of recurrences (D. Mak and G. Benson 2007)

RESULTS

Sensitivity: values agree with recent work

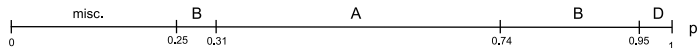
Hit distribution:

compare all seeds with $L = 18$, $\omega = 11$, $\Sigma = \{0, 1\}$, $t = 100$:

sensitivity



probability of at least 2 non-overlapping hits



Seeds A: 111 * 1 * * 1 * 1 * * 11 * 111 (PatternHunter)

B: 111 * * 1 * 11 * * 1 * 1 * 111

C: 11 * * 111 * 1 * * 1 * 111 * 1

D: 111 * 1 * * 11 * 1 * 1 * * 111

CONCLUSION

PAA provides exact, unifying method to compute seed sensitivity

- different homology models
- various seed models
- unifying definitions for gapped and ungapped alignments
- exact sensitivity for multiple seeds
- entire hit distribution
(counting overlapping or non-overlapping seed hits)