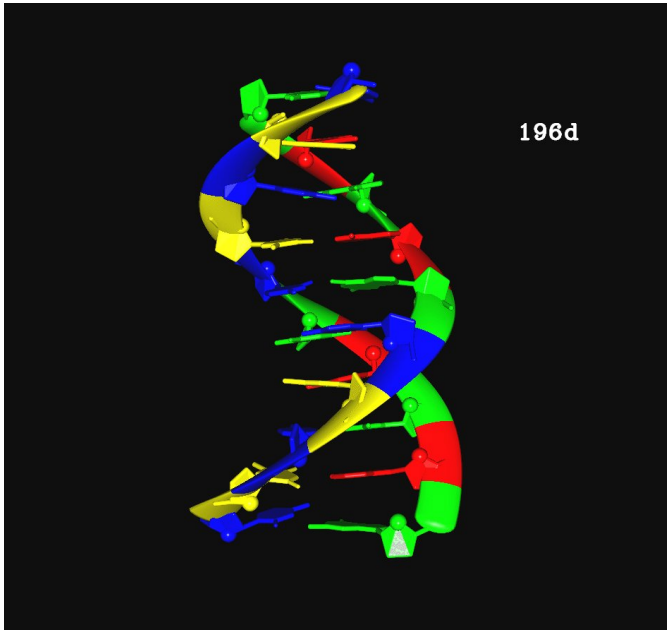
Assessing the significance of Sets of Words

V. Boeva, J. Clément, M. Régnier and M. Vandenbogaert

Moscow, Marne-la-Vallée-CNRS, INRIA, Biozentrum

CPM 2005 – June 22, 2005

Genome analysis



- Structure of the DNA
- Over-(and under) represented DNA motifs
- Regulation sites in genes

Paradigm: biological/random comparison

Paradigm

Comparing mathematical criteria in biological and random sequences, one can extract biological features.

Example

- If a pattern occurs with different frequencies in a real sequence and a random sequence, then it could have a biological meaning.
- When searching for over-represented or under-represented patterns, we must test that such a pattern is not generated by randomness itself.

Paradigm: biological/random comparison

Paradigm

Comparing mathematical criteria in biological and random sequences, one can extract biological features.

Example

- If a pattern occurs with different frequencies in a real sequence and a random sequence, then it could have a biological meaning.
- When searching for over-represented or under-represented patterns, we must test that such a pattern is not generated by randomness itself.

Over-represented patterns

Biological sequence

TTCATTATCTCCATTCGCTGGTGGGCAAGGACTTGAGCTATCGCCCTTTC...
GCATAAAGTTATTCATAAACTGTCAGGGGTTCGGTTGCCGCTGGTGGAAC...
AGGCTGGTGGACGCCTACGTTATTTTGCTGGTGGACTGGAAATCATCTAG...
TCCAACGAAATAGCTGGTGGTCTACACTCATATCGTTATTAACAAACGAA...
AGAAACTAATGGGTGTCACAGCTGGTGGGCTCGTATTTTGTAGGAGGTCA...

Random sequence

ATATATATATTTATCTTGCAACTCGGAGAATTCTATTAATATATGAACGA...
ACGTAGATGACAACAATTAGCATGTGGATTTGTAAGGTAAGTTTCTTGTG...
CGTTGGTTGGTCATCGATGCAATGAATGAGTCGTTTAAAATAAGACTCGA...
TTGTCTCTCAAGTTTTTTTTGCATTACCATTCTAAGCTGGTGGATATAGG...
GTTTACAAGTTTTAACCTTTTGTCACTCGTCACCTTATGTGTGGCTTTAA...

→ *Chi* Motif in E. COLI.

Over-represented patterns

Biological sequence

TTCATTATCTCCATTCGCTGGTGGGCAAGGACTTGAGCTATCGCCCTTTC...
GCATAAAGTTATTCATAAACTGTCAGGGGTTCGGTTGCCGCTGGTGGAAC...
AGGCTGGTGGACGCCTACGTTATTTTGCTGGTGGACTGGAAATCATCTAG...
TCCAACGAAATAGCTGGTGGTCTACACTCATATCGTTATTAACAAACGAA...
AGAAACTAATGGGTGTCACAGCTGGTGGGCTCGTATTTTGTAGGAGGTCA...

Random sequence

ATATATATATTTATCTTGCAACTCGGAGAATTCTATTAATATATGAACGA...
ACGTAGATGACAACAATTAGCATGTGGATTTGTAAGGTAAGTTTCTTGTG...
CGTTGGTTGGTCATCGATGCAATGAATGAGTCGTTTAAAATAAGACTCGA...
TTGTCTCTCAAGTTTTTTTTGCATTACCATTCTAAGCTGGTGGATATAGG...
GTTTACAAGTTTTAACCTTTTGTCACTCGTCACCTTATGTGTGGCTTTAA...

→ *Chi* Motif in E. coli.

Significance of a pattern?

We need to characterize the "probabilistic behaviour" of a pattern.

Problem

There exist measures expressed by expressions and recurrences which can be cumbersome to handle (+ numerical instability)

Our contribution

- A rewriting of exact matricial formula to get tractable formula for the probability of first occurrence of a motif and first co-occurrence of a pair of motifs (here a motif can be a set of words)
- Exhibit a few combinatorial parameters for sets of words
- We consider a positional pattern ($\approx$ affinity matrices) for which efficient computation of these parameters is possible

Significance of a pattern?

We need to characterize the "probabilistic behaviour" of a pattern.

Problem

There exist measures expressed by expressions and recurrences which can be cumbersome to handle (+ numerical instability)

Our contribution

- A rewriting of exact matricial formula to get tractable formula for the probability of first occurrence of a motif and first co-occurrence of a pair of motifs (here a motif can be a set of words)
- Exhibit a few combinatorial parameters for sets of words
- We consider a positional pattern ($\approx$ affinity matrices) for which efficient computation of these parameters is possible

Evaluation of the significance of a pattern H

Let

- $O_n(H)$ = Random variable counting the number of occurrences of the pattern H in a random text of length n.
- $\mathrm{Obs}(H)$ = the number of occurrences of the pattern H in the biological sequence.

How to estimate the significance?

- z-score:

$$Z(H) = \frac{\mathrm{E}[O_n(H)] - \mathrm{Obs}(H)}{\sqrt{\mathrm{Var}\, O_n(H)}}$$

  [Meaningful for a normal distribution, not too far from the mean]
- p-values: $p(H) = \Pr\{O_n(H) \geq \mathrm{Obs}(H)\}$ [Large deviations techniques]
- Probability of first occurrence $\mathcal{F}_n = \Pr\{O_n(H) > 0\}$ [related to waiting time]

Evaluation of the significance of a pattern H

Let

- $O_n(H)$ = Random variable counting the number of occurrences of the pattern H in a random text of length n.
- $\mathrm{Obs}(H)$ = the number of occurrences of the pattern H in the biological sequence.

How to estimate the significance?

- z-score:

$$Z(H) = \frac{\mathrm{E}[O_n(H)] - \mathrm{Obs}(H)}{\sqrt{\mathrm{Var}\, O_n(H)}}$$

  [Meaningful for a normal distribution, not too far from the mean]

- p-values: $p(H) = \Pr\{O_n(H) \geq \mathrm{Obs}(H)\}$ [Large deviations techniques]
- Probability of first occurrence $\mathcal{F}_n = \Pr\{O_n(H) > 0\}$ [related to waiting time]

Evaluation of the significance of a pattern H

Let

- $O_n(H)$ = Random variable counting the number of occurrences of the pattern H in a random text of length n.
- $\mathrm{Obs}(H)$ = the number of occurrences of the pattern H in the biological sequence.

How to estimate the significance?

- z-score:

$$Z(H) = \frac{\mathrm{E}[O_n(H)] - \mathrm{Obs}(H)}{\sqrt{\mathrm{Var}\, O_n(H)}}$$

  [Meaningful for a normal distribution, not too far from the mean]
- p-values: $p(H) = \Pr\{O_n(H) \geq \mathrm{Obs}(H)\}$ [Large deviations techniques]
- Probability of first occurrence $\mathcal{F}_n = \Pr\{O_n(H) > 0\}$ [related to waiting time]

Evaluation of the significance of a pattern H

Let

- $O_n(H)$ = Random variable counting the number of occurrences of the pattern H in a random text of length n.
- $\mathrm{Obs}(H)$ = the number of occurrences of the pattern H in the biological sequence.

How to estimate the significance?

- z-score:

$$Z(H) = \frac{\mathrm{E}[O_n(H)] - \mathrm{Obs}(H)}{\sqrt{\mathrm{Var}\, O_n(H)}}$$

  [Meaningful for a normal distribution, not too far from the mean]
- p-values: $p(H) = \Pr\{O_n(H) \geq \mathrm{Obs}(H)\}$ [Large deviations techniques]
- Probability of first occurrence $\mathcal{F}_n = \Pr\{O_n(H) > 0\}$ [related to waiting time]

Probabilistic models

These criteria suppose an underlying probabilistic model

- Shuffling (exact) model: fix a parameter k and keep the same distribution of factors of length k as in a reference sequence [hard to study!]
- Bernoulli model: $(p_i)_{i\in\Sigma}$ [memoryless]
- Markov model: $\mathbb{P} = (p_{i|j})_{i,j\in\Sigma}, (\pi_i)_{i\in\Sigma}$ [finite context]

Our work concerns Bernoulli and Markov model.

Probabilistic models

These criteria suppose an underlying probabilistic model

- Shuffling (exact) model: fix a parameter k and keep the same distribution of factors of length k as in a reference sequence [hard to study!]
- Bernoulli model: $(p_i)_{i\in\Sigma}$ [memoryless]
- Markov model: $\mathbb{P} = (p_{i|j})_{i,j\in\Sigma}, (\pi_i)_{i\in\Sigma}$ [finite context]

Our work concerns Bernoulli and Markov model.

Probabilistic models

These criteria suppose an underlying probabilistic model

- Shuffling (exact) model: fix a parameter k and keep the same distribution of factors of length k as in a reference sequence [hard to study!]
- Bernoulli model: $(p_i)_{i\in\Sigma}$ [memoryless]
- Markov model: $\mathbb{P} = (p_{i|j})_{i,j\in\Sigma}, (\pi_i)_{i\in\Sigma}$ [finite context]

Our work concerns Bernoulli and Markov model.

Probabilistic models

These criteria suppose an underlying probabilistic model

- Shuffling (exact) model: fix a parameter k and keep the same distribution of factors of length k as in a reference sequence [hard to study!]
- Bernoulli model: $(p_i)_{i\in\Sigma}$ [memoryless]
- Markov model: $\mathbb{P} = (p_{i|j})_{i,j\in\Sigma}, (\pi_i)_{i\in\Sigma}$ [finite context]

Our work concerns Bernoulli and Markov model.

Over-(or under-)representation of patterns

Input

- model for the sequence
- n, sequence length
- pattern H (or a set of patterns $\mathcal{H}$)

Question

Find the probabilistic law of the pattern in random sequences of size n (expected values, variances, waiting time, ...)

Two different approaches

- Experimental: A. Denise, M.-F. Sagot, L. Marsan
- Analytical approach

Over-(or under-)representation of patterns

Input

- model for the sequence
- n, sequence length
- pattern H (or a set of patterns $\mathcal{H}$)

Question

Find the probabilistic law of the pattern in random sequences of size n (expected values, variances, waiting time, ...)

Two different approaches

- Experimental: A. Denise, M.-F. Sagot, L. Marsan
- Analytical approach

Analytical approach

Probabilistic methods

[Prum, Rodolphe, de Turkheim 95], [Schbath 97], [Apostolico, Bock, Xuyan 98], [Reinert, Schbath, Waterman 00], ...

Combinatorial methods

Generating functions of probabilities [Régnier, Szpankowski 98], [Nicodème, Salvy, Flajolet 99], ...

Large deviations

[Denise, Régnier 04]

See also Lothaire vol.3 "Applied Combinatorics on Words" to appear soon with a chapter by Reinert, Schbath, Waterman and another by Jacquet, Szpankowski.

Analytical approach

Probabilistic methods

[Prum, Rodolphe, de Turkheim 95], [Schbath 97], [Apostolico, Bock, Xuyan 98], [Reinert, Schbath, Waterman 00], ...

Combinatorial methods

Generating functions of probabilities [Régnier, Szpankowski 98], [Nicodème, Salvy, Flajolet 99], ...

Large deviations

[Denise, Régnier 04]

See also Lothaire vol.3 "Applied Combinatorics on Words" to appear soon with a chapter by Reinert, Schbath, Waterman and another by Jacquet, Szpankowski.

Analytical approach

Probabilistic methods

[Prum, Rodolphe, de Turkheim 95], [Schbath 97], [Apostolico, Bock, Xuyan 98], [Reinert, Schbath, Waterman 00], ...

Combinatorial methods

Generating functions of probabilities [Régnier, Szpankowski 98], [Nicodème, Salvy, Flajolet 99], ...

Large deviations

[Denise, Régnier 04]

See also Lothaire vol.3 "Applied Combinatorics on Words" to appear soon with a chapter by Reinert, Schbath, Waterman and another by Jacquet, Szpankowski.

Analytical approach

Probabilistic methods

[Prum, Rodolphe, de Turkheim 95], [Schbath 97], [Apostolico, Bock, Xuyan 98], [Reinert, Schbath, Waterman 00], ...

Combinatorial methods

Generating functions of probabilities [Régnier, Szpankowski 98], [Nicodème, Salvy, Flajolet 99], ...

Large deviations

[Denise, Régnier 04]

See also Lothaire vol.3 "Applied Combinatorics on Words" to appear soon with a chapter by Reinert, Schbath, Waterman and another by Jacquet, Szpankowski.

Combinatorial analysis : from generating functions to formulas

Methodology

- Equations on languages translate to equations on generating functions. A language $\mathcal{L} \rightarrow L(z) = \sum_{w \in \mathcal{L}} \alpha(w) P(w) z^{|w|}$
- Computing parameters requires extracting coefficients of generating functions.

Extracting coefficient can be tedious (numerical instability, time efficiency), slow and require the use of formal systems like Maple or Mathematica.

Techniques and tools

- Complex analysis
- Algebraic operations on series

Combinatorial analysis : from generating functions to formulas

Methodology

- Equations on languages translate to equations on generating functions. A language $\mathcal{L} \rightarrow L(z) = \sum_{w \in \mathcal{L}} \alpha(w) P(w) z^{|w|}$
- Computing parameters requires extracting coefficients of generating functions.

Extracting coefficient can be tedious (numerical instability, time efficiency), slow and require the use of formal systems like Maple or Mathematica.

Techniques and tools

- Complex analysis
- Algebraic operations on series

Example

Let $F_n(\mathcal{H})$ be the probability that at least one word in the set $\mathcal{H}$ occurs in a random sequence of size n. One has

$$F_{\mathcal{H}}(z) = \sum_{n \geq 0} F_n(\mathcal{H}) z^n = \frac{1}{1-z} H(z) \mathbb{D}(z)^{-1} \, \mathbf{1}_q{}^{\mathrm{t}}$$

Here: if $\mathcal{H} = \{H_1, \ldots, H_q\}$ with the H_i words of length m,
$H(z) = (P(H_1) z^m, \ldots, P(H_q) z^m)$,
$\mathbb{D}(z) = (1-z)(\mathbb{I} + \mathbb{C}(z)) + \mathbb{H}(z)$.

Applying algebraic and combinatorial identities, we get

$$F_{\mathcal{H}}(z) = \frac{1}{1-z} - \frac{1}{1-z+Q_{\mathcal{H}}(z)} \text{with } Q_{\mathcal{H}}(z) = \mathrm{Trace}(\mathbb{H}(z) \mathbb{A}(z)^{-1}).$$

The asymptotics comes from dominant singularity in the complex plane $z \in \mathbb{C}$, obtained by bootstrapping

$$P\{O_n > 0\} \approx 1 - (1 + P(\mathcal{H}) - \widetilde{C}(\mathcal{H}))^{-n}, \text{ when } n\, P(\mathcal{H}) < 1$$

Example

Let $F_n(\mathcal{H})$ be the probability that at least one word in the set $\mathcal{H}$ occurs in a random sequence of size n. One has

$$F_{\mathcal{H}}(z) = \sum_{n \geq 0} F_n(\mathcal{H}) z^n = \frac{1}{1-z} H(z) \mathbb{D}(z)^{-1} \, \mathbf{1}_q{}^{\mathrm{t}}$$

Here: if $\mathcal{H} = \{H_1, \ldots, H_q\}$ with the H_i words of length m,
$H(z) = (P(H_1)z^m, \ldots, P(H_q)z^m)$,
$\mathbb{D}(z) = (1-z)(\mathbb{I} + \mathbb{C}(z)) + \mathbb{H}(z)$.

Applying algebraic and combinatorial identities, we get

$$F_{\mathcal{H}}(z) = \frac{1}{1-z} - \frac{1}{1-z+Q_{\mathcal{H}}(z)} \text{with } Q_{\mathcal{H}}(z) = \mathrm{Trace}(\mathbb{H}(z)\mathbb{A}(z)^{-1}).$$

The asymptotics comes from dominant singularity in the complex plane $z \in \mathbb{C}$, obtained by bootstrapping

$$P\{O_n > 0\} \approx 1 - (1 + P(\mathcal{H}) - \widetilde{C}(\mathcal{H}))^{-n}, \text{ when } n\,P(\mathcal{H}) < 1$$

Example

Let $F_n(\mathcal{H})$ be the probability that at least one word in the set $\mathcal{H}$ occurs in a random sequence of size n. One has

$$F_{\mathcal{H}}(z) = \sum_{n \geq 0} F_n(\mathcal{H}) z^n = \frac{1}{1-z} H(z) \mathbb{D}(z)^{-1} \, \mathbf{1}_q{}^{\mathrm{t}}$$

Here: if $\mathcal{H} = \{H_1, \ldots, H_q\}$ with the H_i words of length m,
$H(z) = (P(H_1)z^m, \ldots, P(H_q)z^m)$,
$\mathbb{D}(z) = (1-z)(\mathbb{I} + \mathbb{C}(z)) + \mathbb{H}(z)$.

Applying algebraic and combinatorial identities, we get

$$F_{\mathcal{H}}(z) = \frac{1}{1-z} - \frac{1}{1-z+Q_{\mathcal{H}}(z)} \text{with } Q_{\mathcal{H}}(z) = \mathrm{Trace}(\mathbb{H}(z)\mathbb{A}(z)^{-1}).$$

The asymptotics comes from dominant singularity in the complex plane $z \in \mathbb{C}$, obtained by bootstrapping

$$P\{O_n > 0\} \approx 1 - (1 + P(\mathcal{H}) - \widetilde{C}(\mathcal{H}))^{-n}, \text{ when } n\,P(\mathcal{H}) < 1$$

Combinatorial properties of a set of words

Let $F = \texttt{ATAA}$ and $G = \texttt{AACT}$.

```
ATAA
..AACT
...AACT
```

- CT and ACT are right complements.
- The set of right complements of F in G is $C_{F,G}$ the *correlation set*.
- When $F = G$, the *autocorrelation set* is $A_F = C_{F,F} + \varepsilon$ with ε the empty word.
- The set $\widetilde{C}_{F,\mathcal{H}}$ of *minimal right complements* of F in $\mathcal{H}$ is the set of minimal words in $\cup_{G \in \mathcal{H}} C_{F,G}$ for the prefix order.

Combinatorial properties of a set of words

Let $F = \texttt{ATAA}$ and $G = \texttt{AACT}$.

```
ATAA
..AACT
...AACT
```

- CT and ACT are right complements.
- The set of right complements of F in G is $C_{F,G}$ the *correlation set*.
- When $F = G$, the *autocorrelation set* is $A_F = C_{F,F} + \varepsilon$ with ε the empty word.
- The set $\widetilde{C}_{F,\mathcal{H}}$ of *minimal right complements* of F in $\mathcal{H}$ is the set of minimal words in $\cup_{G \in \mathcal{H}} C_{F,G}$ for the prefix order.

Combinatorial parameters

A few combinatorial parameters

$$P(\mathcal{H}) = \sum_{w \in \mathcal{H}} P(w)$$
$$C(\mathcal{H}) = \sum_{F,G \in \mathcal{H}} \sum_{w \in C_{F,G}} P(Fw)$$
$$\widetilde{C}(\mathcal{H}) = \sum_{F \in \mathcal{H}} \sum_{w \in \widetilde{C}_{F,\mathcal{H}}} P(Fw)$$

These suffice to express several quantities

$$\mathrm{E}[O_n(\mathcal{H})] = (n-m+1)\ P(\mathcal{H}),$$
$$\mathrm{Var}[O_n(\mathcal{H})] = (n-m+1)\big(\ P(\mathcal{H}) + (1-2m)\ P(\mathcal{H})^2 + 2\,C(\mathcal{H})\big) + m(m-1)\ P(\mathcal{H})^2 - 2\,\widehat{C}(\mathcal{H})$$
$$\Pr\{O_n > 0\} \approx 1 - (1 + P(\mathcal{H}) - \widetilde{C}(\mathcal{H}))^{-n}, \text{if } n\,P(\mathcal{H}) < 1$$

Better than systems of functional equations only if we know how to compute these quantities efficiently!

Combinatorial parameters

A few combinatorial parameters

$$P(\mathcal{H}) = \sum_{w \in \mathcal{H}} P(w)$$
$$C(\mathcal{H}) = \sum_{F,G \in \mathcal{H}} \sum_{w \in C_{F,G}} P(Fw)$$
$$\widetilde{C}(\mathcal{H}) = \sum_{F \in \mathcal{H}} \sum_{w \in \widetilde{C}_{F,\mathcal{H}}} P(Fw)$$

These suffice to express several quantities

$$\mathrm{E}[O_n(\mathcal{H})] = (n - m + 1)\ P(\mathcal{H}),$$
$$\mathrm{Var}[O_n(\mathcal{H})] = (n - m + 1)\big(\ P(\mathcal{H}) + (1 - 2m)\ P(\mathcal{H})^2 + 2\,C(\mathcal{H})\big) + m(m-1)\ P(\mathcal{H})^2 - 2\,\widehat{C}(\mathcal{H})$$
$$\Pr\{O_n > 0\} \approx 1 - (1 + P(\mathcal{H}) - \widetilde{C}(\mathcal{H}))^{-n}, \text{if } n\,P(\mathcal{H}) < 1$$

Better than systems of functional equations only if we know how to compute these quantities efficiently!

A general method to compute parameters

Right complements are related to borders on words $\Longrightarrow$ They can be computed using a tree-like structure computed thanks to the Aho-Corasick algorithm with complexity $O(\sum_{w \in \mathcal{H}} |w|)$.

When $\mathcal{H}$ is the set of words which are at Hamming distance k from a word H of length m, it remains $O(m^k)$, *i.e.*, exponential with respect to k.

A general method to compute parameters

Right complements are related to borders on words $\Longrightarrow$ They can be computed using a tree-like structure computed thanks to the Aho-Corasick algorithm with complexity $O(\sum_{w \in \mathcal{H}} |w|)$.

When $\mathcal{H}$ is the set of words which are at Hamming distance k from a word H of length m, it remains $O(m^k)$, *i.e.*, exponential with respect to k.

A particular setting: positional pattern with errors

Alphabet $\Sigma(= \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\})$

- Pattern $\mathcal{H} \in \Sigma^m \equiv (\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_m)$ where $\mathcal{H}_i \subseteq \Sigma$
- Neighborhood $\mathcal{N} = (\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_m)$ such that $\mathcal{H}_i \subseteq \mathcal{N}_i \subseteq \Sigma \rightarrow$ states the allowed errors at each position.

Example

$\mathcal{H} = (\{\mathtt{A}, \mathtt{T}\}, \{\mathtt{A}\}, \{\mathtt{G}\}, \{\mathtt{A}\}, \{\mathtt{C}\})$,
$\mathcal{N} = (\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}, \{\mathtt{A}, \mathtt{T}\}, \{\mathtt{G}\}, \{\mathtt{A}, \mathtt{T}\}, \{\mathtt{C}, \mathtt{G}\})$
(Note that in IUPAC code , $\mathcal{H} = \mathtt{WAGAC}$ and $\mathcal{N} = \mathtt{NWGWS}$)
$\mathcal{H}$ = AAGAC + TAGAC
$\mathcal{N}$ = AAGAC + TAGAC + AAGAG+ AAGTC+ AAGTG+ ATGAC+ ATGAG+ ATGTC+ ATGTG+ CAGAC+ CAGAG+ CAGTC+ CAGTG+ CTGAC+ CTGAG+ CTGTC+ CTGTG+ GAGAC+ GAGAG+ GAGTC+ GAGTG+ GTGAC+ GTGAG+ GTGTC+ GTGTG+ TAGAG+ TAGTC+ TAGTG+ TTGAC+ TTGAG+ TTGTC+ TTGTG

A particular setting: positional pattern with errors

Alphabet $\Sigma(= \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\})$

- Pattern $\mathcal{H} \in \Sigma^m \equiv (\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_m)$ where $\mathcal{H}_i \subseteq \Sigma$
- Neighborhood $\mathcal{N} = (\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_m)$ such that $\mathcal{H}_i \subseteq \mathcal{N}_i \subseteq \Sigma \rightarrow$ states the allowed errors at each position.

Example

$\mathcal{H} = (\{\mathtt{A}, \mathtt{T}\}, \{\mathtt{A}\}, \{\mathtt{G}\}, \{\mathtt{A}\}, \{\mathtt{C}\})$,
$\mathcal{N} = (\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}, \{\mathtt{A}, \mathtt{T}\}, \{\mathtt{G}\}, \{\mathtt{A}, \mathtt{T}\}, \{\mathtt{C}, \mathtt{G}\})$
(Note that in IUPAC code , $\mathcal{H} = \mathtt{WAGAC}$ and $\mathcal{N} = \mathtt{NWGWS}$)

$\mathcal{H}$ = AAGAC + TAGAC

$\mathcal{N}$ = AAGAC + TAGAC + AAGAG+ AAGTC+ AAGTG+ ATGAC+
ATGAG+ ATGTC+ ATGTG+ CAGAC+ CAGAG+ CAGTC+ CAGTG+
CTGAC+ CTGAG+ CTGTC+ CTGTG+ GAGAC+ GAGAG+ GAGTC+
GAGTG+ GTGAC+ GTGAG+ GTGTC+ GTGTG+ TAGAG+ TAGTC+
TAGTG+ TTGAC+ TTGAG+ TTGTC+ TTGTG

A particular setting: positional pattern with errors

Alphabet $\Sigma(= \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\})$

- Pattern $\mathcal{H} \in \Sigma^m \equiv (\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_m)$ where $\mathcal{H}_i \subseteq \Sigma$
- Neighborhood $\mathcal{N} = (\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_m)$ such that $\mathcal{H}_i \subseteq \mathcal{N}_i \subseteq \Sigma \rightarrow$ states the allowed errors at each position.

Example

$\mathcal{H} = (\{\mathtt{A}, \mathtt{T}\}, \{\mathtt{A}\}, \{\mathtt{G}\}, \{\mathtt{A}\}, \{\mathtt{C}\})$,
$\mathcal{N} = (\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}, \{\mathtt{A}, \mathtt{T}\}, \{\mathtt{G}\}, \{\mathtt{A}, \mathtt{T}\}, \{\mathtt{C}, \mathtt{G}\})$
(Note that in IUPAC code , $\mathcal{H} = \mathtt{WAGAC}$ and $\mathcal{N} = \mathtt{NWGWS}$)

$\mathcal{H}$ = AAGAC + TAGAC

$\mathcal{N}$ = AAGAC + TAGAC + AAGAG+ AAGTC+ AAGTG+ ATGAC+
ATGAG+ ATGTC+ ATGTG+ CAGAC+ CAGAG+ CAGTC+ CAGTG+
CTGAC+ CTGAG+ CTGTC+ CTGTG+ GAGAC+ GAGAG+ GAGTC+
GAGTG+ GTGAC+ GTGAG+ GTGTC+ GTGTG+ TAGAG+ TAGTC+
TAGTG+ TTGAC+ TTGAG+ TTGTC+ TTGTG

Computing probabilities

For these patterns, we easily compute

$$\sum_{\substack{w \in \mathcal{N} \\ d(w,\mathcal{H}) \leq k}} P(w), \text{with } d(\cdot,\cdot) \text{ the Hamming distance}$$

- First idea: view sets of words as formal series, and substitute probabilities to symbols (just as affinity matrices)

$$P : \mathtt{TAAGC} \mapsto p_{\mathtt{T}} p_{\mathtt{A}} p_{\mathtt{A}} p_{\mathtt{G}} p_{\mathtt{C}} \qquad \text{(Bernoulli)}$$

$$\mapsto \sum_{i \in \Sigma} \pi_i p_{\mathtt{T}|i} p_{\mathtt{A}|\mathtt{T}} p_{\mathtt{A}|\mathtt{A}} p_{\mathtt{G}|\mathtt{A}} p_{\mathtt{C}|\mathtt{G}} \qquad \text{(Markov)}$$

- Mark the errors: introduce a new variable u to count the number of errors with respect to $\mathcal{H}$.

Mark the errors

The variable u counts the number of errors.
Black: symbols of the patterns marked by $u^0 = 1$, Red: allowed error marked by $u^1 = u$.
Developing

$$(\mathtt{A} + u\mathtt{C} + u\mathtt{G} + \mathtt{T})(\mathtt{A} + u\mathtt{T})\mathtt{G}(\mathtt{A} + u\mathtt{T})(\mathtt{C} + u\mathtt{G}),$$

gives AAGAC + uAAGAG+ uAAGTC+ u^2AAGTG+ uATGAC+ u^2ATGAG+ u^2ATGTC+ u^3ATGTG+ uCAGAC+ uCAGAG+ u^2CAGTC+ u^3CAGTG+ u^2CTGAC+ u^3CTGAG+ u^3CTGTC+ u^4CTGTG+ uGAGAC+ u^2GAGAG+ u^2GAGTC+ u^3GAGTG+ u^2GTGAC+ u^3GTGAG+ u^3GTGTC+ u^4GTGTG+ TAGAC+ uTAGAG+ uTAGTC+ u^2TAGTG+ uTTGAC+ u^2TTGAG+ u^2TTGTC +u^3TTGTG

Mark the errors

The variable u counts the number of errors.
Black: symbols of the patterns marked by $u^0 = 1$, Red: allowed error marked by $u^1 = u$.
Developing

$$(\mathtt{A} + u\mathtt{C} + u\mathtt{G} + \mathtt{T})(\mathtt{A} + u\mathtt{T})\mathtt{G}(\mathtt{A} + u\mathtt{T})(\mathtt{C} + u\mathtt{G}),$$

gives AAGAC + uAAGA$\mathtt{G}$+ uAAG$\mathtt{T}$C+ u^2AAG$\mathtt{T}\mathtt{G}$+ uA$\mathtt{T}$GAC+ u^2A$\mathtt{T}$GA$\mathtt{G}$+ u^2A$\mathtt{T}$G$\mathtt{T}$C+ u^3A$\mathtt{T}$G$\mathtt{T}\mathtt{G}$+ uCAGAC+ uCAGAG+ u^2CAG$\mathtt{T}$C+ u^3CAG$\mathtt{T}\mathtt{G}$+ u^2C$\mathtt{T}$GAC+ u^3C$\mathtt{T}$GA$\mathtt{G}$+ u^3C$\mathtt{T}$G$\mathtt{T}$C+ u^4C$\mathtt{T}$G$\mathtt{T}\mathtt{G}$+ $u\mathtt{G}$AGAC+ $u^2\mathtt{G}$AGA$\mathtt{G}$+ $u^2\mathtt{G}$AG$\mathtt{T}$C+ $u^3\mathtt{G}$AG$\mathtt{T}\mathtt{G}$+ $u^2\mathtt{G}\mathtt{T}$GAC+ $u^3\mathtt{G}\mathtt{T}$GA$\mathtt{G}$+ $u^3\mathtt{G}\mathtt{T}$G$\mathtt{T}$C+ $u^4\mathtt{G}\mathtt{T}\mathtt{G}\mathtt{T}\mathtt{G}$+ TAGAC+ uTAGA$\mathtt{G}$+ uTAG$\mathtt{T}$C+ u^2TAG$\mathtt{T}\mathtt{G}$+ $u\mathtt{T}\mathtt{T}$GAC+ $u^2\mathtt{T}\mathtt{T}$GA$\mathtt{G}$+ $u^2\mathtt{T}\mathtt{T}$G$\mathtt{T}$C +$u^3\mathtt{T}\mathtt{T}\mathtt{G}\mathtt{T}\mathtt{G}$

Mark the errors

The variable u counts the number of errors.
Black: symbols of the patterns marked by $u^0 = 1$, Red: allowed error marked by $u^1 = u$.
Developing

$$(\mathtt{A} + u\mathtt{C} + u\mathtt{G} + \mathtt{T})(\mathtt{A} + u\mathtt{T})\mathtt{G}(\mathtt{A} + u\mathtt{T})(\mathtt{C} + u\mathtt{G}),$$

gives $\mathtt{AAGAC} + u\mathtt{AAGAG} + u\mathtt{AAGTC} + u^2\mathtt{AAGTG} + u\mathtt{ATGAC} + u^2\mathtt{ATGAG} + u^2\mathtt{ATGTC} + u\mathtt{CAGAC} + u\mathtt{CAGAG} + u^2\mathtt{CAGTC} + u^2\mathtt{CTGAC} + u\mathtt{GAGAC} + u^2\mathtt{GAGAG} + u^2\mathtt{GAGTC} + u^2\mathtt{GTGAC} + \mathtt{TAGAC} + u\mathtt{TAGAG} + u\mathtt{TAGTC} + u^2\mathtt{TAGTG} + u\mathtt{TTGAC} + u^2\mathtt{TTGAG} + u^2\mathtt{TTGTC}$

Mark the errors

The variable u counts the number of errors.
Black: symbols of the patterns marked by $u^0 = 1$, Red: allowed error marked by $u^1 = u$.
Developing

$$(\mathtt{A} + u\mathtt{C} + u\mathtt{G} + \mathtt{T})(\mathtt{A} + u\mathtt{T})\mathtt{G}(\mathtt{A} + u\mathtt{T})(\mathtt{C} + u\mathtt{G}),$$

gives AAGAC + uAAGAG+ uAAGTC+ u^2AAGTG+ uATGAC+ u^2ATGAG+
u^2ATGTC+ uCAGAC+ uCAGAG+ u^2CAGTC+
u^2CTGAC+ uGAGAC+ u^2GAGAG+
u^2GAGTC+ u^2GTGAC+
TAGAC+ uTAGAG+ uTAGTC+ u^2TAGTG+ uTTGAC+ u^2TTGAG+
u^2TTGTC

k errors max $\longrightarrow$ remove words with more than $k = 2$ red letters
$\longrightarrow$ truncate polynomials in u at order k on the degree

Computing the probability (end)

Of course we don't want to develop! We compute iteratively the probabilities by considering successive positions and using truncated polynomials in u.

When the number of errors k is fixed, we have an algorithm in time $O(mk)$ to compute the probability ($m\times$ cost of multiplying a polynomial of degree k by a monomial of degree 1).

The same principles can be extended to a Markov model and the others parameters related to the overlapping structure.

Computing the probability (end)

Of course we don't want to develop! We compute iteratively the probabilities by considering successive positions and using truncated polynomials in u.

When the number of errors k is fixed, we have an algorithm in time $O(mk)$ to compute the probability ($m\times$ cost of multiplying a polynomial of degree k by a monomial of degree 1).

The same principles can be extended to a Markov model and the others parameters related to the overlapping structure.

Computing the probability (end)

Of course we don't want to develop! We compute iteratively the probabilities by considering successive positions and using truncated polynomials in u.

When the number of errors k is fixed, we have an algorithm in time $O(mk)$ to compute the probability ($m\times$ cost of multiplying a polynomial of degree k by a monomial of degree 1).

The same principles can be extended to a Markov model and the others parameters related to the overlapping structure.

Conclusion

We have rewritten exact matricial expressions in a suitable form depending on a few combinatorial parameters which are easy to compute.

Perspectives

- Extension to dyadic – or structured – patterns (M.-F. Sagot and L. Marsan), palindromic patterns, highly repetitive patterns
- Conditional occurrence problem (artifacts)