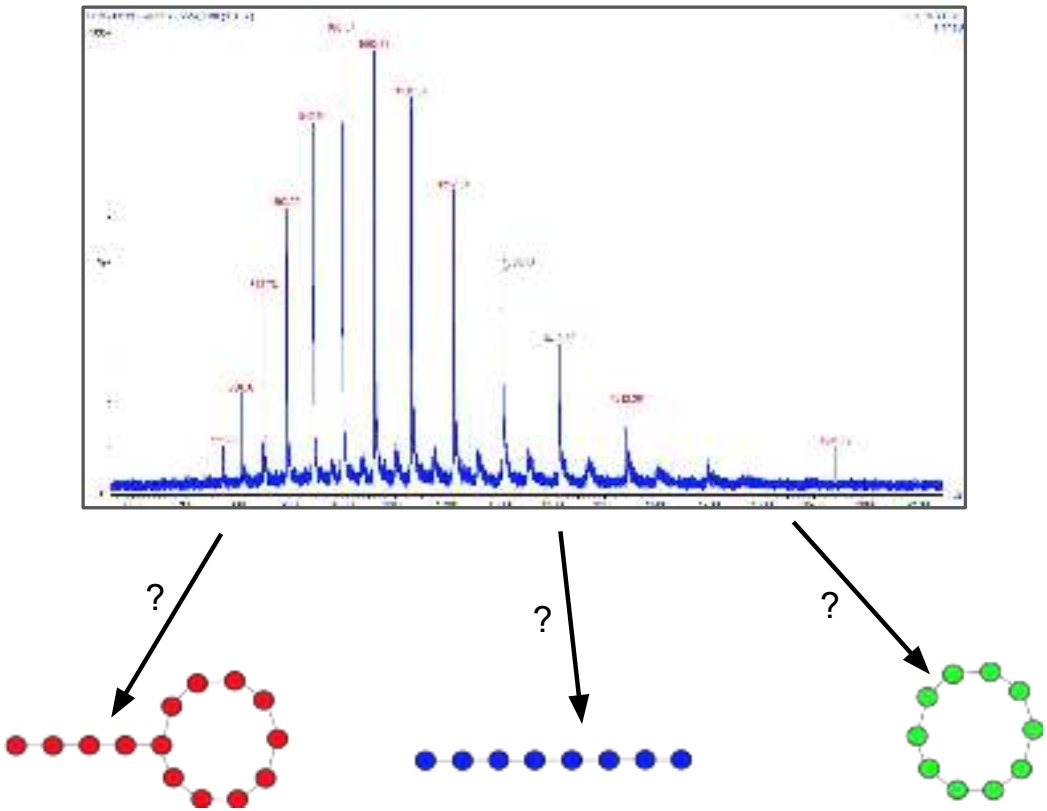
Assessing Significance of Peptide Spectrum Match Scores

Ananstasiia Abramova Anton Korobeynikov

Bioinformatics institute

Saint Petersburg, 2016

Peptide identification



- The goal is to perform peptide identification.
- However we are unaware of any suitable function, that can compute proper *distance*.
- We use function *Score*, that measures peptide-spectrum *similarity*.
- Therefore we have to calculate peptide match score significance.

Significance of Peptide Spectrum Match Scores

We aim to compute:

$$p = \mathbb{P}(Score(Spectrum, peptide) > S^*) = \mathbb{P}(peptide \in A).$$

Here $peptide \in \Omega$ is as follows (Mohimani et al., 2013):

- fixed structure,
- fixed number of initial components,
- fixed sum of initial components (parent-mass of a peptide).

Significance of Peptide Spectrum Match Scores

We aim to compute:

$$p = \mathbb{P}(Score(Spectrum, peptide) > S^*) = \mathbb{P}(peptide \in A).$$

Here $peptide \in \Omega$ is as follows (Mohimani et al., 2013):

- fixed structure,
- fixed number of initial components,
- fixed sum of initial components (parent-mass of a peptide).

In case of linear peptides

For fixed k, M:

$$peptide = (m_1, \ldots, m_k),\ m_i > 0,\ \sum_{i=1}^{k} m_i = M.$$

Monte-Carlo Method

Standard Monte-Carlo approach

$X_1, \ldots, X_N$ — sequence of iid, uniformly distributed peptides from Ω

$$\widehat{p}_N = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{X_i \in A\}}$$

Monte-Carlo Method

Standard Monte-Carlo approach

$X_1, \ldots, X_N$ — sequence of iid, uniformly distributed peptides from Ω

$$\widehat{p}_N = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{X_i \in A\}}$$

- Problem: if p is small, $\mathbb{D}\widehat{p}_N = \frac{p(1-p)}{N}$ is large relative to p (in our case $p \approx 10^{-10}$).
- Standard solution: importance sampling.

Importance Sampling Approach

- Sample N copies of X independently from a sampling distribution G with corresponding density g.
- Compute the estimate

$$\widehat{p}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{dF}{dG} \mathbb{I}_{\{X_i \in A\}}$$

Importance Sampling Approach

- Sample N copies of X independently from a sampling distribution G with corresponding density g.
- Compute the estimate

$$\widehat{p}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{dF}{dG} \mathbb{I}_{\{X_i \in A\}}$$

- Problem: shape of the set A is complex. How to choose G?
- Solution: if f is a density of F, consider $g(x) \propto w(Score(x))f(x)$.

Resulting Algorithm of p Estimation

1. Estimate weights w using Wang-Landau algorithm.
2. Define transition density γ for Metropolis-Hastings algorithm:
   - Let $x = (\alpha_1, \ldots, \alpha_k)$ be the current state.
   - Sample i from uniform distribution on $\{1, \ldots, k\}$.
   - Sample δ from the uniform distribution on $[-\alpha_i, \alpha_{i+1}]$.
   - Assign $y = (\alpha_1, \ldots, \alpha_i + \delta, \alpha_i - \delta, \ldots, \alpha_k)$
3. Construct Markov chain with equilibrium density $g(x) = w(Score(x))f(x)$
4. Construct estimate:

$$\widehat{p}_N = \frac{\sum_{i=1}^{N} \mathbb{I}_{\{X_i \in A\}} / w(Score(X_i))}{\sum_{i=1}^{N} 1 / w(Score(X_i))}$$

Variance of Estimates

- In order to construct a confidence interval for $\widehat{p}_N$, we need to estimate its variance.
- $X_1, \ldots, X_N$ form a Markov chain (so they are dependent).

There are different approaches (Flegal et al., 2010) for variance estimation in such cases:

- Non-overlapping batch means (Jones et al., 2006)
- Overlapping batch means (Jones et al., 2006)
- Spectral variance methods (Hobert et al., 2002)

Variance of Estimates

- In order to construct a confidence interval for $\widehat{p}_N$, we need to estimate its variance.
- $X_1, \ldots, X_N$ form a Markov chain (so they are dependent).

There are different approaches (Flegal et al., 2010) for variance estimation in such cases:

- Non-overlapping batch means (Jones et al., 2006)
- Overlapping batch means (Jones et al., 2006)
- Spectral variance methods (Hobert et al., 2002)

Stopping Rule for Markov Chain Trajectory

- We need to decide how long to run a MCMC simulation.
- We have to perform p-value calculations *en masse* $\Longrightarrow$ it is impossible to employ fixed time rule or perform Markov Chain diagnostics by hands

Proposal (Flegal et al., 2013)

Denote the posterior variance associated with p by λ_p^2. *Suppose that* $\sqrt{N}(\widehat{p}_N - p) \xrightarrow{d} \mathcal{N}(0, \sigma_p^2)$. *Suppose also that* $\widehat{\lambda}_N \to \lambda_p$ *and* $\widehat{\sigma}_p \to \sigma_p$ *w.p.1. Denote*

- $T_\varepsilon = \inf\{n \geq 0 : 2z_{\delta/2}\widehat{\sigma}_p/\sqrt{N} \leq \varepsilon\widehat{\lambda}_N\}$,
- $C_N = (\widehat{p}_N - z_{\delta/2}\widehat{\sigma}_p/\sqrt{N}; \widehat{p}_N + z_{\delta/2}\widehat{\sigma}_p/\sqrt{N})$.

Then as $N \to \infty$ *or* $\varepsilon \to 0$ *the simulation will stop w.p.1 and*

$$\mathbb{P}(p \in C_{T_\varepsilon}) \to 1 - \delta.$$

Results: surfactin

	$\widehat{p}_N$	95% conf. int.	
MC	$1.3 \cdot 10^{-5}$	$1.19 \cdot 10^{-5}$	$1.4 \cdot 10^{-5}$
MCMC	$1.3 \cdot 10^{-5}$	$1.28 \cdot 10^{-5}$	$1.38 \cdot 10^{-5}$
MS-DPR (Mohimani et al., 2013)	$2.3 \cdot 10^{-5}$	NA	NA