Cleavage Fragment Statistics for Peptide Mass Fingerprinting with Weight-Accumulating Markov Models

Sven Rahmann

Algorithms and Statistics for Systems Biology Group
Genome Informatics, Faculty of Technology, Bielefeld University, Germany

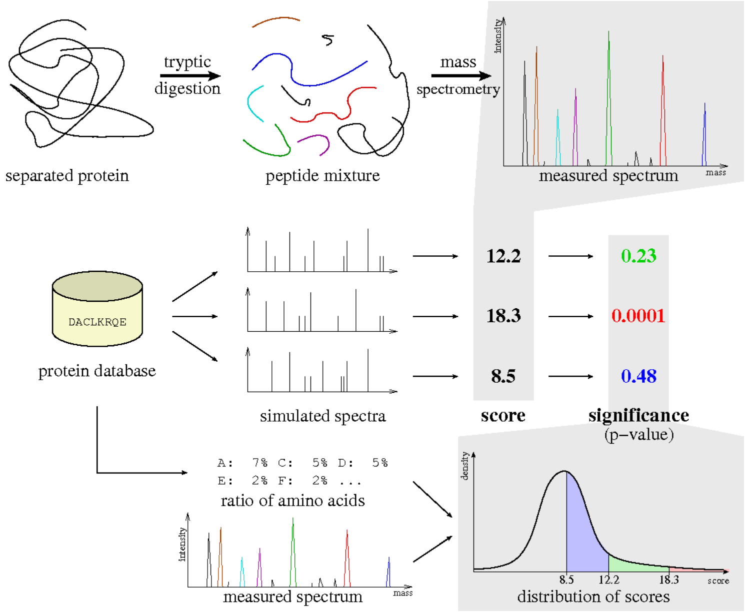
DMV-Jahrestagung, Bonn, 21.09.2006

Peptide Mass Fingerprinting

Protein Identification

- Isolate all copies of one protein from a cell
- Digest these proteins deterministically into fragments (peptides)
- Measure fragment masses by mass spectrometry
- Compare peptide mass fingerprint (PMF) to predicted PMF of database proteins
- Return database protein that “fits best”
- Compute significance of “best fit”

Peptide Mass Fingerprinting



Protein Space – just strings

Definition (Protein sequence)

A protein is a word of some length $\ell \geq 1$ over amino acid alphabet Σ.

Protein Space – just strings

Definition (Protein sequence)

A protein is a word of some length $\ell \geq 1$ over amino acid alphabet Σ.

Definition (Random protein model)

For a given length $\ell \geq 1$ and amino acid frequencies $f = (f(a))_{a \in \Sigma}$, assign a probability to every protein sequence $s = (s_1, \ldots, s_\ell)$:

$$\mathbb{P}_\ell(S = s) = \prod_{i=1}^{\ell} f(s_i).$$

Protein Space – just strings

Definition (Protein sequence)

A protein is a word of some length $\ell \geq 1$ over amino acid alphabet Σ.

Definition (Random protein model)

For a given length $\ell \geq 1$ and amino acid frequencies $f = (f(a))_{a \in \Sigma}$, assign a probability to every protein sequence $s = (s_1, \ldots, s_\ell)$:

$$\mathbb{P}_\ell(S = s) = \prod_{i=1}^{\ell} f(s_i).$$

No masses so far

Protein masses – weighted strings

Definition (Amino acid mass)

Every amino acid a has a mass distribution $\mathcal{L}_a$, derived from

- isotopic distributions of its component atoms,
- modification probabilities,
- mass distributions of modifying groups.

Protein masses – weighted strings

Definition (Amino acid mass)

Every amino acid a has a mass distribution $\mathcal{L}_a$, derived from

- isotopic distributions of its component atoms,
- modification probabilities,
- mass distributions of modifying groups.

Definition (Protein mass)

Every amino acid s_i of protein $s \in \Sigma^\ell$ has a random mass μ_{s_i} drawn from its distribution $\mathcal{L}_{s_i}$.

$$\mu_s = \mu_{s_1} + \mu_{s_2} + \cdots + \mu_{s_\ell} \quad \text{and} \quad \mathcal{L}_s = \mathcal{L}_{s_1} \star \mathcal{L}_{s_2} \star \cdots \star \mathcal{L}_{s_\ell}.$$

Protein Cleavage – getting the PMF

Definition (Standard cleavage scheme)

A standard cleavage scheme (Γ, Π) is specified by

- a set Γ of cleavage characters
- a set Π of prohibition characters

Semantics: cut after aa from Γ unless followed by aa from Π.

Protein Cleavage – getting the PMF

Definition (Standard cleavage scheme)

A standard cleavage scheme (Γ, Π) is specified by

- a set Γ of cleavage characters
- a set Π of prohibition characters

Semantics: cut after aa from Γ unless followed by aa from Π.

Example (Trypsin)

$\Gamma = \{\mathtt{K}, \mathtt{R}\}$, $\Pi = \{\mathtt{P}\}$; cuts after lys or arg unless followed by pro.
SwissProt frequencies: $f(\mathtt{K}) + f(\mathtt{R}) = 11.25\%$, $f(\mathtt{P}) = 4.83\%$.

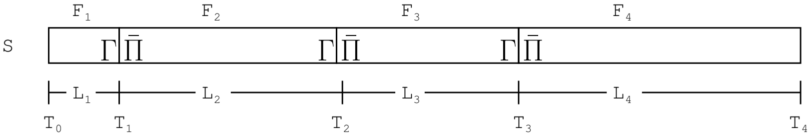
Protein Cleavage – getting the PMF

Definition (Standard cleavage scheme)

A standard cleavage scheme (Γ, Π) is specified by

- a set Γ of cleavage characters
- a set Π of prohibition characters

Semantics: cut after aa from Γ unless followed by aa from Π.



Example (Trypsin)

$\Gamma = \{\mathtt{K}, \mathtt{R}\}$, $\Pi = \{\mathtt{P}\}$; cuts after lys or arg unless followed by pro.
SwissProt frequencies: $f(\mathtt{K}) + f(\mathtt{R}) = 11.25\%$, $f(\mathtt{P}) = 4.83\%$.

Computational problems

For given

- random protein model and sequence length,
- amino acid mass distribution, and
- cleavage rules,

Computational problems

For given

- random protein model and sequence length,
- amino acid mass distribution, and
- cleavage rules,

determine efficiently

- the distribution of the number of fragments,

Computational problems

For given

- random protein model and sequence length,
- amino acid mass distribution, and
- cleavage rules,

determine efficiently

- the distribution of the number of fragments,
- the distribution of the fragment lengths,

Computational problems

For given

- random protein model and sequence length,
- amino acid mass distribution, and
- cleavage rules,

determine efficiently

- the distribution of the number of fragments,
- the distribution of the fragment lengths,
- the joint length-mass distribution, and

Computational problems

For given

- random protein model and sequence length,
- amino acid mass distribution, and
- cleavage rules,

determine efficiently

- the distribution of the number of fragments,
- the distribution of the fragment lengths,
- the joint length-mass distribution, and
- mass occurrence probabilities: probability that there exists at least one fragment with mass in a given range

Possible approaches

- Enumeration of all 20^ℓ protein sequences
  - Exact, but infeasible for $\ell \geq 10$

Possible approaches

- Enumeration of all 20^ℓ protein sequences
  - Exact, but infeasible for $\ell \geq 10$
- Sampling of random proteins
  - Not exact (rare events)

Possible approaches

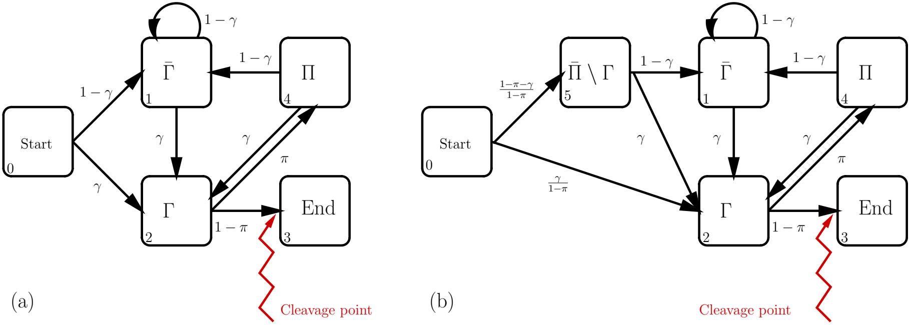
- Enumeration of all 20^ℓ protein sequences
  - Exact, but infeasible for $\ell \geq 10$
- Sampling of random proteins
  - Not exact (rare events)
- Estimation from database
  - Large database required for high precision
  - Also not exact (rare events)

Possible approaches

- Enumeration of all 20^ℓ protein sequences
  - Exact, but infeasible for $\ell \geq 10$
- Sampling of random proteins
  - Not exact (rare events)
- Estimation from database
  - Large database required for high precision
  - Also not exact (rare events)

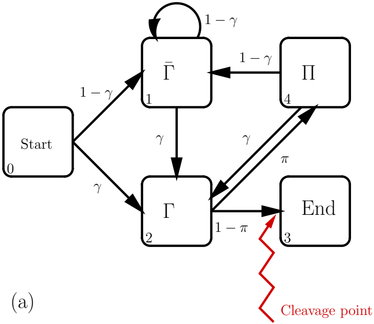
Is there an exact and efficient method?

“Weight-accumulating Markov Chains (WAMMs)”



WAMM: generative probabilistic cleavage model
Left: Initial fragment. Right: Following fragments.

“Weight-accumulating Markov Chains (WAMMs)”

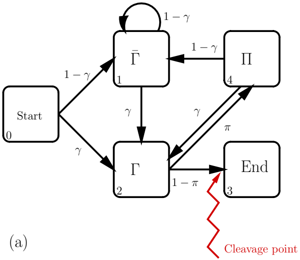


WAMM: generative probabilistic cleavage model

Left: Initial fragment. Right: Following fragments.

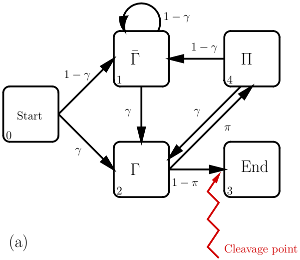
A WAMM can be derived from a standard cleavage scheme (Γ, Π), or from more complicated cleavage rules.

Using WAMMs for Probability Computations



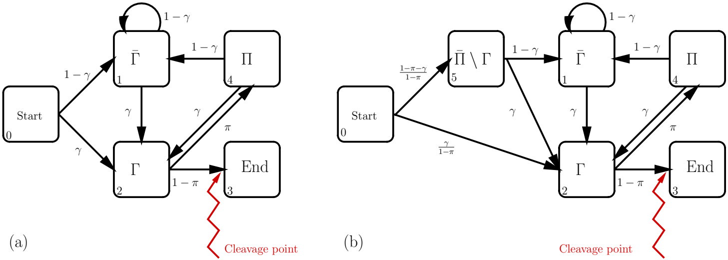
- $h_i^l[m] := \mathbb{P}(\text{in state } i \text{ after } l \text{ steps, accumulated mass } m)$,
- $g_i[m] := \mathbb{P}(\text{mass} = m \mid \text{State} = i)$,

Using WAMMs for Probability Computations



- $h_i^l[m] := \mathbb{P}$(in state i after l steps, accumulated mass m),
- $g_i[m] := \mathbb{P}(\text{mass} = m \mid \text{State} = i)$,
- Matrix $P :=$ Transition matrix of the WAMM.

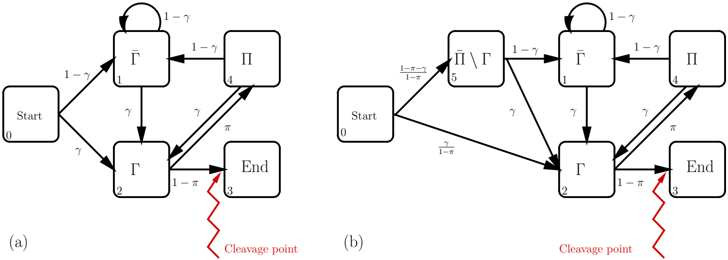
Using WAMMs for Probability Computations



- $h_i^l[m] := \mathbb{P}(\text{in state } i \text{ after } l \text{ steps, accumulated mass } m)$,
- $g_i[m] := \mathbb{P}(\text{mass} = m \mid \text{State} = i)$,
- Matrix $P :=$ Transition matrix of the WAMM.

Then
$$h_i^l[m] = \sum_{m'} \left(\sum_k h_k^{l-1}[m - m'] \cdot P_{ki} \right) \cdot g_i[m']$$

Using WAMMs for Probability Computations

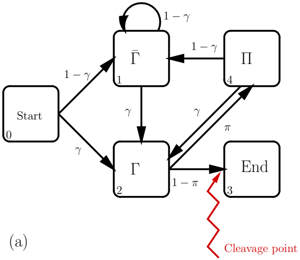


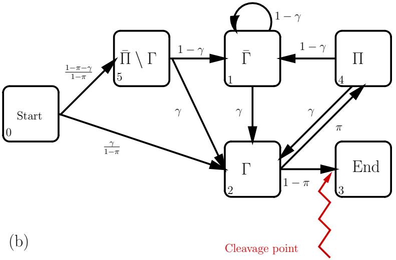
- $h_i^l[m] := \mathbb{P}(\text{in state } i \text{ after } l \text{ steps, accumulated mass } m)$,
- $g_i[m] := \mathbb{P}(\text{mass} = m \mid \text{State} = i)$,
- Matrix $P :=$ Transition matrix of the WAMM.

Then $$h_i^l[m] = \sum_{m'} \left(\sum_k h_k^{l-1}[m - m'] \cdot P_{ki} \right) \cdot g_i[m']$$

$$\mathbb{P}(\text{fragment has length } l \text{ and mass } m) = h_{\text{,,End''}}^{l+1}[m]$$

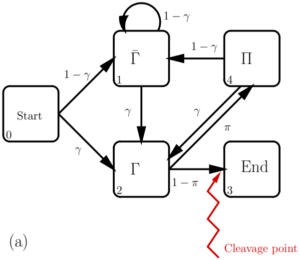
Using WAMMs for Probability Computations





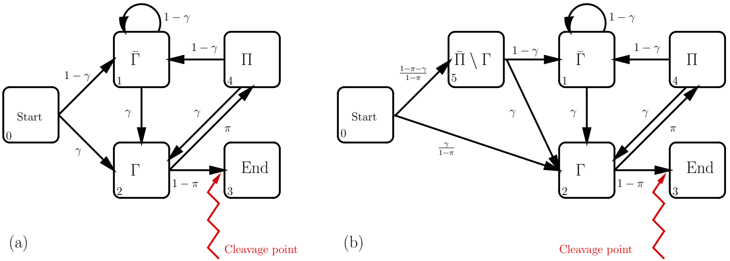
- Matrix $H^{(l)} := (h_i^l[m])_{m \in \text{masses},\ i \in \text{states}}$
  (contains the joint mass-state distribution after l steps),

Using WAMMs for Probability Computations



- Matrix $H^{(l)} := (h_i^l[m])_{m \in \text{masses},\ i \in \text{states}}$ (contains the joint mass-state distribution after l steps),
- Matrix $G := (g_i[m])_{m \in \text{masses},\ i \in \text{states}}$,
- Matrix $P :=$ Transition matrix of the WAMM.

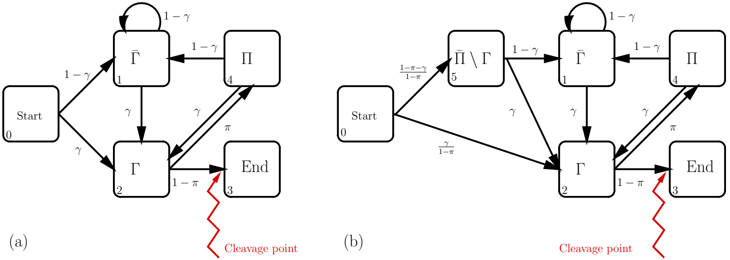
Using WAMMs for Probability Computations



- Matrix $H^{(l)} := (h_i^l[m])_{m \in \text{masses},\ i \in \text{states}}$ (contains the joint mass-state distribution after l steps),
- Matrix $G := (g_i[m])_{m \in \text{masses},\ i \in \text{states}}$,
- Matrix $P :=$ Transition matrix of the WAMM.

Then $$H^{(l)} = (H^{(l-1)} \cdot P) \star G.$$

Using WAMMs for Probability Computations



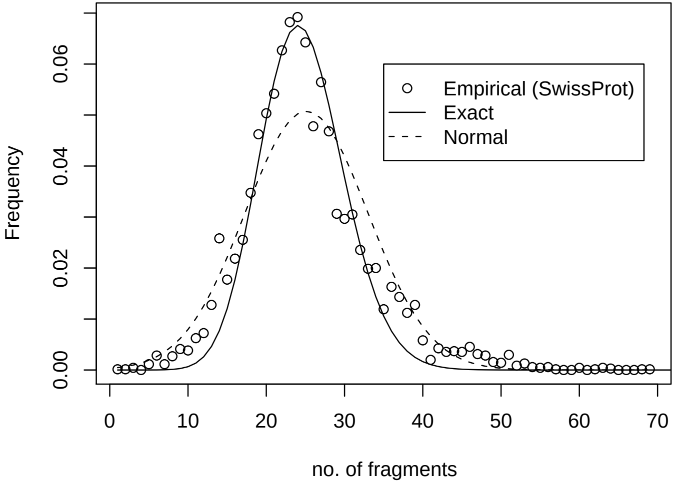
- Matrix $H^{(l)} := (h_i^l[m])_{m \in \text{masses},\, i \in \text{states}}$ (contains the joint mass-state distribution after l steps),
- Matrix $G := (g_i[m])_{m \in \text{masses},\, i \in \text{states}}$,
- Matrix $P :=$ Transition matrix of the WAMM.

$$\text{Then} \qquad H^{(l)} = (H^{(l-1)} \cdot P) \star G.$$

This is an update formula for the mass-state distribution.

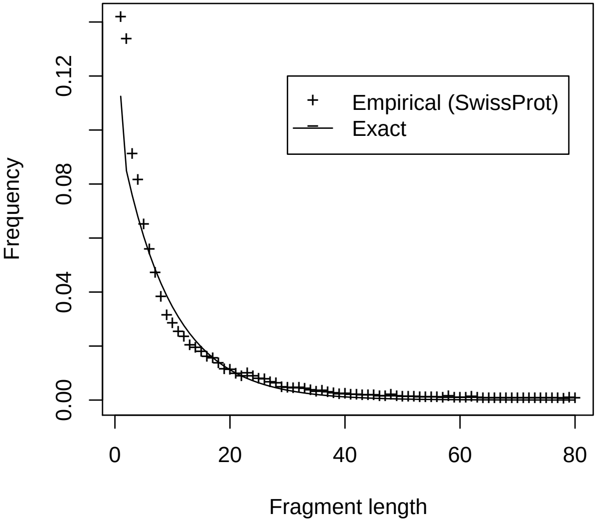
Results: Number of Fragments

Fragment number distribution of proteins of length 207 ± 7.



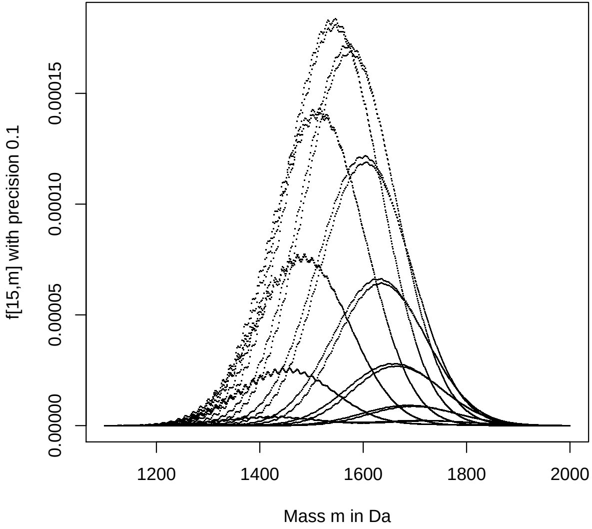
Fragment Lengths

Distribution of fragment lengths of SwissProt proteins



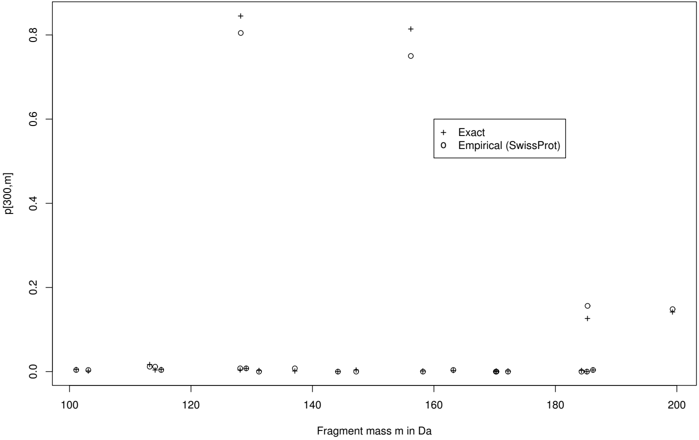
Joint Length-Mass Distribution

Fragment mass distribution; length = 15, High precision = 0.1 Da.



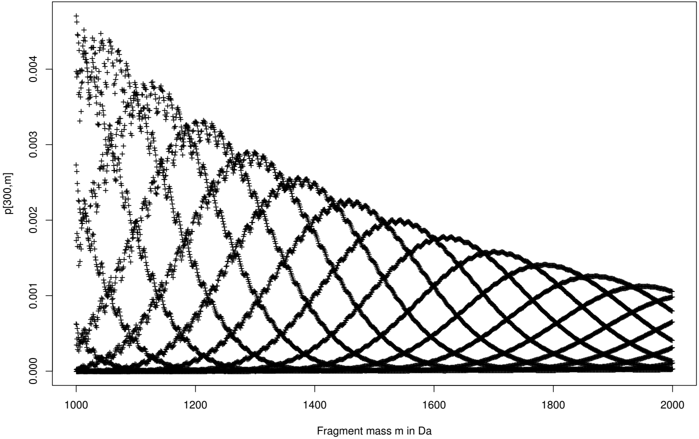
Mass Occurrence Probabilities

Fragment mass occurrence probabilities for proteins of length 300

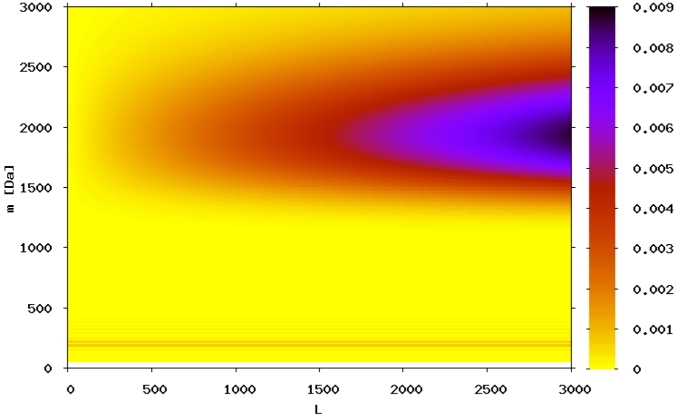


Mass Occurrence Probabilities

Fragment mass occurrence probabilities for proteins of length 300



Mass Occurrence Probabilities



Summary

- New computational framework "WAMM"

Summary

- New computational framework “WAMM”
- Only aa frequencies needed

Summary

- New computational framework "WAMM"
- Only aa frequencies needed
- Elegant formulation and update equation:
  $H^{(l)} = (H^{(l-1)} \cdot P) \star G.$

Summary

- New computational framework "WAMM"
- Only aa frequencies needed
- Elegant formulation and update equation: $H^{(l)} = (H^{(l-1)} \cdot P) \star G$.
- Applicable to probability computations in mass spectrometry, to significance computations for peptide mass fingerprinting, e.g., what's the probability that a random protein contains a fragment with mass in a given range?

Acknowledgments

Joint work with

- Hans-Michael Kaltenbach
- Sebastian Böcker

Acknowledgments

Joint work with

- Hans-Michael Kaltenbach
- Sebastian Böcker

Thanks to

- International NRW Graduate School in Bioinformatics and Genome Research, Bielefeld
- Henner Sudek, Marcel Martin and Tobias Marschall

Universität Bielefeld
Technische Fakultät
AG Genominformatik

International Graduate School
Bioinformatics & Genome Research

Acknowledgments

Joint work with

- Hans-Michael Kaltenbach
- Sebastian Böcker

Thanks to

- International NRW Graduate School in Bioinformatics and Genome Research, Bielefeld
- Henner Sudek, Marcel Martin and Tobias Marschall

Thank you for listening

Questions?