A Statistical Model of Proteolytic Digestion

I-Jeng Wang[1], Christopher P. Diehl[1] and Fernando J. Pineda[2]
[1]*Research & Technology Development Center, Johns Hopkins University Applied Physics Lab*
[2]*Dept. of Molecular Microbiology & Immunology, Johns Hopkins Bloomberg School of Public Health*
{i-jeng.wang; chris.diehl}@jhuapl.edu, fernando.pineda@jhu.edu

Abstract

We present a stochastic model of proteolytic digestion of a proteome, assuming the distribution of parent protein lengths in the proteome, the relative abundances of the 20 amino acids in the proteome, and the digestion "rules" of the enzyme used in the digestion. We derived a closed form expression for the fragment mass distribution for a large class of enzymes including the widely used Trypsin. The expression uses the distribution of lengths in a mixture of proteins taken from a proteome, as well as the relative abundances of the 20 amino acids in the proteome. The agreement between theory and the in silico digest is excellent.

1. Introduction

We present a rigorous stochastic model of proteolytic digestion of a proteome, assuming (1) the distribution of parent protein lengths in the proteome, (2) the relative abundances of the 20 amino acids in the proteome and (3) the digestion "rules" of the enzyme used in the digestion. The model can be used for hypothesis testing in protein identification software. Current protein identification software (e.g. MOWSE [1], MassSearch [2], and Mascot [3]), is based on histograms of digested proteomes, or on models derived from unrealistic assumptions (e.g. a uniform distribution of proteolytic fragments). The proposed model accounts analytically for the mixture of proteins that constitutes a microorganism's proteome.

It is useful to organize *digestion models* into a taxonomy according to (1) the class of digestion rules obeyed by the enzyme and (2) the order of the Markov model used to model the amino acid sequence. We denote by C(*m*) the class of enzymes whose digestion rules depend on *m* adjacent residues. Table 1 shows a selection of enzymes and their corresponding cleavage rules. In this paper, we assume that the cleavage rules are deterministic. We classify amino acid sequence models according to the order of the Markov chain used to model the sequence. For example, M(0) denotes the simplest sequence models assuming that the adjacent amino acid residues are independent; M(1) denotes the next most complex models assuming that the next residue depends on the previous amino acid residue; and so forth.

Table 1. Cleavage rules for selected reagents

Digesting Reagent	**Class C(1) Cleavage Rules**
Armillaria	Before {C | K}
AspN	Before D
Thermolysin	Before {F | I | L | M | V | W}
Clostripain	After R
LysC	After K
Pancreatic Elastase	After {A| G | S | V}
Digesting Reagent	**Class C(2) Cleavage Rules**
CNBr_Cys	{Before C | After M}
Chymotrypsin	After {F | L | M | W | Y} if not followed by P
Trypsin	After {K | R} if not followed by P
V8 Ammonium Acetate	After E if not followed by P
V8 Phosphate Buffer	After {D | E} if not followed by P
Hydroxylamine	After N if followed by G
Mild Acid Hydrolysis	After D if followed by P

Our methodology views the enzymatic digestion problem as a regenerative process to which we can apply Wald's equation and its generalizations to simplify the computation of the fragment mass distribution. In particular, when overlaid with a cleavage process (either C(1) or C(2)), a protein sequence (with models M(0) or M(1)) can be equivalently modeled as a regenerative process with the cleavage sites as the regeneration points. Hence the regenerative cycles (the fragments) are i.i.d.. Using Wald's first lemma and its extensions [4], the fragment mass distribution can be conveniently partitioned into two terms: the length distribution and the conditional mass probability density. Rigorous closed form expressions for these terms have been derived for the digestion models {C(1), M(0)} and {C(2), M(0)}in [5]. In this paper, we present results for {C(1), M(0)} and a subset of {C(2), M(0)} with the class of enzymes $C_P(2)$ that exhibit Proline blocking. The class $C_P(2)$ includes the widely used enzyme, Trypsin, which cleaves after Lysine or Arginine unless followed by Proline. The agreement between theory and the *in silico* digest is excellent as illustrated by the examples given in Section 4.

2. Models and decomposition

2.1. Sequence and fragment models

Throughout the paper, we assume that the protein sequence is i.i.d.. That is, we only consider class M(0) models. We use $S = X_1 X_2 X_3 \ldots$ to denote an infinite random protein sequence, where the X_i's are random variables taking value from a finite alphabet $\mathcal{A}$ (the set of 20 amino acids). We use $p_a, a \in \mathcal{A}$, to denote $\mathrm{P}\{X_n = a\}$, and $p_{\mathcal{C}}$, $\mathcal{C} \subset \mathcal{A}$, to denote $\mathrm{P}\{X_n \in \mathcal{C}\}$.

We consider C(1) and $C_P(2)$ enzyme classes. For C(1) enzymes, cleavage occurs at sites adjacent to amino acids in the set $\mathcal{C}$. For $C_P(2)$ enzymes, cleavage occurs after any amino acid in $\mathcal{C}$ if the following amino acid is not Proline. To facilitate our discussion, we will use $\{F_n\}$ to denote a generic random fragment sequence resulting from the application of a cleavage rule to a random protein sequence S. Here, we provide an informal discussion on why $\{F_n\}$ is i.i.d. except for the first fragment F_1 (please refer to [5] for a rigorous proof).

Amino acids in M(0) proteins are i.i.d. and thus our distributions do not depend on which end of the protein we choose as the N-terminus. Therefore, without loss of generality, we can assume that for class C(1) enzymes, cleavage sites always occur *after* an amino acid in $\mathcal{C}$. For models in {C(1), M(0)}, it is clear that the cleavage event $\{X_n \in \mathcal{C}\}$ is a regenerative point for the protein sequence S. Hence the process between cleavage (that is, the fragments $\{F_n\}$) is i.i.d.. For models in {C(2), M(0)}, a similar conclusion can be drawn by considering an equivalent augmented model for the protein sequence [5]. In the context of this augmented model, an i.i.d. fragment process $\{F_n\}$ can be defined from the regenerative cycles.

Given that $\{F_n\}$ is i.i.d., we will use F to denote a generic random fragment in the sequel.

2.2. Decomposition of mass distribution

For hypothesis testing, it is useful to analytically compute the expected number of fragments with mass m, denoted by $H(m)$, given any specific mass m of interest. Based on the i.i.d. property of fragments, we can establish a decomposition that enable us to treat the fragment length distribution and mass probability density separately in the derivation of $H(m)$. This is accomplished by applying Wald's first lemma and its extensions [4]. Here we state the decomposition without proof (see [5] for a proof):

$$H(m) = \sum_k \underbrace{\left[\sum_{N=k} n(N) L_{N,k}\right]}_{\text{Length Distribution}} \underbrace{\mathrm{P}\{M(F) = m \mid L(F) = k\}}_{\text{Mass Probability Density}}, \quad (1)$$

where $n(N)$ is the number of proteins of length N in the database of interest; $L_{N,K}$ is the expected number of fragments of length k from a protein sequence of length N; $M(F)$ is the mass of F; and $L(F)$ is the length of F. This decomposition described by (1) significantly simplifies the computation of $H(m)$.

3. Fragment mass distributions

For models in both {C(1), M(0)} and {C_P(2), M(0)} we obtain the closed form expression for the fragment mass distribution $H(m)$ by deriving expressions for $L_{N,K}$ and $\mathrm{P}\{M(F) = m \mid L(F) = k\}$.

3.1. Digestion model {C(1), M(0)}

For models in {C(1), M(0)},

$$L_{N,k} = \left[(N-k)p_{\mathcal{C}} + 1\right](1-p_{\mathcal{C}})^{k-1} p_{\mathcal{C}}. \quad (2)$$

The closed form for the mass probability density is given by

$$\mathrm{P}\{M(F) = m \mid L(F) = k\} = \sum_{|\vec{v}|=k} \sum_{|\vec{s}|=1} \left[(k-1)! \prod_{r \notin \mathcal{C}} \frac{\hat{p}_r^{v_r}}{v_r!}\right] \left(\prod_{r \in \mathcal{C}} \hat{p}_r^{s_r}\right) \rho(m - m_{\vec{v},\vec{s}}, \sigma_{\vec{v},\vec{s}}), \quad (3)$$

where the vectors $\{\vec{v}, \vec{s}\}$ are 20-dimensional vectors representing the amino acid composition of a peptide and its c-terminus, respectively, and $\hat{p}_r$ is defined by

$$\hat{p}_r \triangleq \frac{p_r}{z_r},$$

$$z_r = \begin{cases} \sum_{a \in \mathcal{C}} p_a, & r \in \mathcal{C}, \\ \sum_{a \in \mathcal{A} \setminus \mathcal{C}} p_a, & r \notin \mathcal{C}. \end{cases}$$

The term $\rho(m - m_{\vec{v},\vec{s}}, \sigma_{\vec{v},\vec{s}})$ denotes either the isotopic mass distribution for a peptide with composition $\{\vec{v}, \vec{s}\}$, or a phenomenological peak width.

3.2. Digestion model {C_P(2), M(0)}

For models in {C_P(2), M(0)},

$$L_{N,k} = \alpha \left[1 + \left(N - k - 1 + \frac{1}{1-p_{\mathrm{P}}}\right) p'_{\mathcal{C}}\right] \left(1 - p'_{\mathcal{C}}\right)^{k-1} p'_{\mathcal{C}}, \quad (4)$$

where

$$\alpha \triangleq 1 - p_{\mathcal{C}}(1 - p_{\mathrm{P}}),$$

$$p'_{\mathcal{C}} \triangleq p_{\mathcal{C}} \left[\frac{(1-p_{\mathcal{C}})(1-p_{\mathrm{P}})}{1 - p_{\mathcal{C}}(1-p_{\mathrm{P}})}\right],$$

with $p_{\mathrm{P}} = \mathrm{P}\{X_n = \mathrm{P}\}$. Note that when p_{P} is small, (4) can be approximated quite accurately by

$$L_{N,k} \approx \left[(N-k)p'_{\mathcal{C}} + 1\right](1 - p'_{\mathcal{C}})^{k-1} p'_{\mathcal{C}},$$

which is identical to the length distribution for C(1) enzymes given in (2) but with scaled cleavage probability p'_C. The mass probability density can be derived for $\{C_P(2), M(0)\}$ as well. Due to the space limitation, please refer to [5] for the closed form expression.

4. Numerical results

A comparison of analytic and *in-silico* results is shown in Figure 1 and 2. We use an E. coli proteome taken from

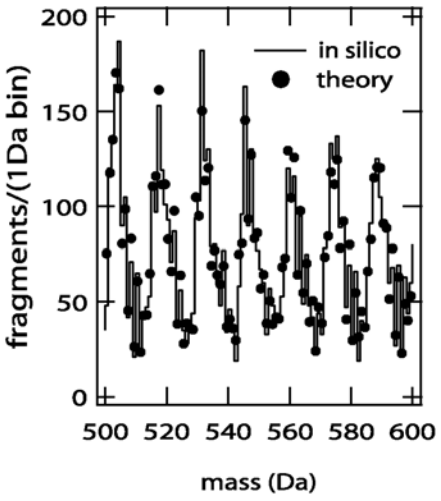

Figure 1. Comparison between analytic expression and in silico digestion for shuffled E. coli proteome over a 100 Da window

the SWISSPROT database. The proteins were shuffled to destroy sequential correlations. The analytic calculation is faster than the corresponding *in silico* digestion.

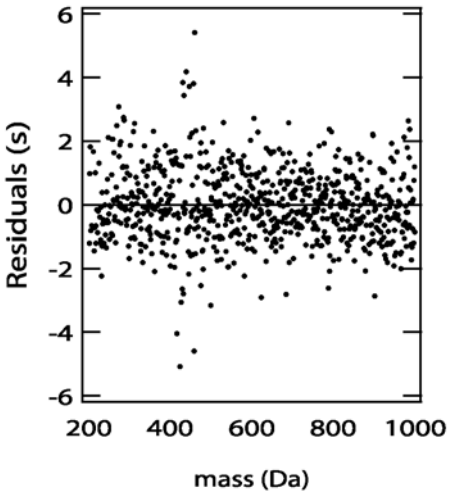

Figure 2. Residues in multiples of Poisson errors (calculated from analytic expression)

Except for unexplained excess dispersion at ~450 Da, the residuals between the *in-silco* and analytic results are consistent with Poisson statistics. It is clear from this figure that the mass distribution can vary by several orders of magnitude over just few Daltons. This result leads us to believe that heuristically derived mass distributions are likely to result in biased p-values in hypothesis tests. This is especially true for small peptides. It becomes a minor effect in heavier peptides. It is typically recommended that protein identification be performed using peptides heavier than ~500 Da due to lack of selectivity of the lighter peptides. Our results suggest that, a more careful treatment of the mass distribution below ~500Da, could make this mass range more informative.

References

[1] D. J. C. Pappin, P. Hojrup, and A. J. Bleasby, "Rapid Identification of Proteins by Peptide-mass Fingerprinting," *Current Biology*, vol. 3, no. 6, 1993, pp. 327–332.

[2] P. James, M. Quadroni, E. Carafoli, and G. Gonnet, "Protein Identification by Mass Profile Fingerprinting," *Biochemical and Biophysical Research Communications*, 195(1), August 1993, pp. 58–64.

[3] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data," *Electrophoresis*, 20, 1999, pp. 3551–3567.

[4] A. Wald, "Sequential Tests of Statistical Hypotheses," *Annals of Mathematical Statistics*, 16, 1945, pp. 117–186.

[5] I-J. Wang, C. P. Diehl, and F. J. Pineda, "A Statistical Model of Proteolytic Digestion," JHU/APL technical report, 2003.