Statistical Foundations of De Novo Sequencing

Sacha Baginsky[+], Mark Cieliebak[#], Jonas Grossmann[+], Wilhelm Gruissem[+], Torsten Kleffmann[+], Lucas K. Mathis[#]*
[+]*Institute of Plant Sciences* [#]*Institute of Theoretical Computer Science*
ETH Zurich
sacha.baginsky@ipw.biol.ethz.ch, cieliebak@inf.ethz.ch

Automatic interpretation of mass spectrometry data is becoming increasingly important in high-throughput protein identification. Using protein databases of hypothetical and real proteins, tandem mass spectrometry (MS/MS) has been applied successfully to identify proteins in complex mixtures (e.g., using SEQUEST). This approach has shortcomings because it is dependent on available protein sequence databases. Identification of a peptide - even from an excellent MS/MS spectrum - may fail for several reasons: certain amino acids were substituted; a corresponding database entry is erroneous; the protein is a product of an alternative splicing process; or - in the worst case - the protein does not occur in the database. In these cases, de novo sequencing techniques for interpreting tandem mass spectra independent of databases are needed.

In general, de novo sequencing works as follows: first, a set of theoretical peptides that match the given MS/MS spectrum is generated. This set can be very large due to contaminations and measurement errors. In a second step, these theoretical peptides are ranked using heuristics, and those peptides with the highest ranking represent the output. While the first step is computationally rather simple (very efficient (i.e., fast) algorithms have been introduced recently [1,2]), finding suitable heuristics to determine the "correct" peptide in the huge set of matching sequences is a difficult problem. This is especially true for spectra that contain a very large amount of noise.

There are several software packages for de novo sequencing, such as Lutefisk, BioAnalyst or BioWorks. However, they do not allow for efficient and reliable de novo sequencing of arbitrary peptides. It seems obvious that in general one single MS/MS spectrum does not contain sufficient information for reliable de novo sequencing. Therefore, we will implement a toolkit (experimental setup, chemicals and sequencing program) which generates additional data - such as spectra from the acetylated protein - and which will allow for efficient de novo sequencing.

As a first step towards the development of such a toolkit, we investigate the data in MS/MS spectra by statistical means. There are many statistical analyses of MS/MS spectra in the literature (e.g. see [3], and [4] for an overview) but we are not aware of statistics aiming at the specific needs for de novo sequencing.

In our experiments, we use polynucleotide phosphorylase, bovine serum albumine and cytochrome as model proteins whose amino acid sequences are well known. These proteins are digested with Trypsin and tryptic peptides are analyzed by LC-ESI-MS/MS using the ion trap technology for mass determination (LCQ DecaXP, ThermoFinnigan, San Jose, California). The ion trap is set to operate in data-dependent acquisition mode, generating four MS/MS scans for the most intense ions of each MS scan. We distinguish between three types of spectra: *true spectra* that belong to a peptide of the input protein, *contamination spectra* that belong to a peptide from some other protein, and *trash spectra*, for which no peptide can be identified at all. We use Sequest and Lutefisk as a reference to determine the peptide corresponding to a spectrum. For the first two types of spectra, the corresponding peptide sequence is known, and we can distinguish between *true peaks* (those peaks that belong to a peptide ion) and *grass peaks* (peaks due to contamination or measurement errors).

Among others, we are addressing the following questions:

- What is the ratio between true spectra, contamination spectra, and trash spectra?

[*] Authors in alphabetical order

- How many peptides of a tryptically digested protein can be measured by common LC-ESI-MS/MS?
- Is there a threshold for the number of peaks that distinguishes trash and non-trash spectra?
- How much does the parent mass measured during the MS/MS run differ from the theoretical parent mass of the corresponding peptide? What mass tolerance is needed?
- What is the ratio between true and grass peaks in a (true or contamination) spectrum?
- Do true and grass peaks differ significantly in their abundance? Does that allow for a threshold to "mow the grass"?
- For true peaks, what is the average difference between the theoretical ion mass and the measured value?
- If *p* is a true peak, how many isotope peaks (*p+1*, *p+2*, etc.) can be expected? What is the probability for a grass peak *q* that *q+1, q+2,* etc. occur as well?
- Given a parent ion with mass *m* and a true peak *p*, does the complementary peak *m-p+1* occur in the spectrum? What mass tolerance is needed? What is the probability that two grass peaks are complementary?

For example, the following table shows the distribution of complementary peaks in 17 true spectra: Hereby, two peaks are complementary if they sum up to the parent mass of the peptide (up to some constant offset 1). A pair of "true/true" peaks *p* and *q* means that the two peaks are complementary *and* that they both belong to a b- resp. y-ion of the corresponding peptide. Analogously, a "grass/grass" pair represents two complementary peaks that do not belong to b- resp. y-ions.[1]

Total No. of peaks in spectrum	No. of b- and y-ions of the corresponding peptide	True peaks in spectrum	True peaks corresponding to b-Ions	True peaks corresponding to y-Ions	**Complementary Peaks**			
					true/true	true/grass	grass/grass	total
191	28	20	10	10	6	2	6	14
285	26	21	11	10	8	4	26	38
162	24	18	9	9	6	1	12	19
122	26	11	5	6	6	1	7	14
270	30	21	11	10	7	1	25	33
171	32	18	9	9	2	1	11	14
146	36	20	7	13	5	1	6	12
135	24	12	4	8	3	0	7	10
288	18	15	7	8	6	17	32	45
193	26	16	6	10	5	3	14	22
231	28	21	9	12	8	1	12	21
489	22	10	6	4	3	0	55	58
188	26	15	7	8	10	2	17	29
147	18	15	8	7	5	3	13	21
173	24	18	8	10	8	1	8	17
63	22	11	6	5	5	2	4	11
146	20	13	7	6	7	0	3	10

As can be seen from the table, approx. one third of all complementary peaks are pairs of true peaks. Thus the number of "true/true" complementary peaks in a sequence generated by a de novo sequencing program can be used to rank this sequence. Although one such criterion is not sufficient for proper ranking schemes, we are sure that the totality of our statistical material will be of large impact for the design of de novo sequencing strategies.

References

1 T., Chen, M. Kao, M. Tepel, J. Rush, G. M. Church: A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry, Proc. of the 11th SIAM-ACM Symposium on Discrete Algorithms (SODA), p. 389-398, 2000.
2 G. Pandurangan and H. Ramesh: The Restriction Mapping Problem Revisited, Journal of Computer and System Sciences, to appear 2002.
3 M. Wilm, G. Neubauer, M. Mann: Parent Ion Scans of Unseparated Peptide Mixtures. Anal. Chem. 68, p. 527-533, 1996.
4 Peter James: Proteome Research: Mass Spectrometry. Springer, 2001.

[i] The “true/grass” pairs appear due to the mass tolerance of 0.6. The number of “true/true” pairs may exceed the number of true peaks because of two peaks that have the mass of a b- or y-ion of the peptide, up to the mass tolerance.