Introduction

Data, information and knowledge

Informatics, in particular bioinformatics, is employed routinely in state-of-the-art bioanalytical chemistry and proteomics. In order to optimize the use of informatics for investigating complex biological phenomena, one has to think about the central term “information” for a moment. Why is information so important and meaningful in our world and how do we have to interact with information in order to exploit its value to a full extent?

First, we have to appreciate that information is not just a side effect of the interaction of physical objects but that it can be understood as a fundamental and most basic essence of our universe. In a philosopher’s view our world is not built of things (or physical objects) but of facts. Interested in understanding the mechanisms of life, this view reminds us that we might have to concentrate on information and communication between things, rather than only on the things themselves. Information, in this context, is a form of data that is pointing to relevant facts. But the primary goal in bioanalytical research is not to create information or data but to create knowledge. Knowledge does not arise just by collecting data. Instead, data has to be filtered and processed to create information. To create knowledge from information we first have to understand the structure of information, its properties and functions. Only when putting new information into context with already existing information are we able to create meaning and thus knowledge (Figure 1).

Information can tell us what is the case. For example, if we look at a measured mass value, this information tells us that the structure and composition of the underlying substance might be anything out of a large set of possibilities. Often

Accurate mass as a bioinformatic parameter in data-to-knowledge conversion: Fourier transform ion cyclotron resonance mass spectrometry for peptide de novo sequencing

B. Spengler
University of Giessen, Institute of Inorganic and Analytical Chemistry, Schubertstrasse 60, Building 16, D-35392 Giessen, Germany

With the availability of ultra-precise mass spectrometric biomolecular data, the accurate mass is becoming a physical quantity of high interest for bioinformatics tools and strategies. Fourier transform ion cyclotron resonance mass spectrometry with electrospray ionization or matrix-assisted laser desorption/ionization sources now allows the easy determination of amino acid composition of medium size, unknown peptides when employing combinatorial calculation of hypothetical parent and fragment ion masses. This new method, which in a second step, allows the reliable de-novo sequencing of completely unknown peptides (“composition-based sequencing (CBS)”) appears to open a wide new field of bioanalytical investigation. It has been shown that even unspecifically cleaved proteins can be identified easily and reliably when accurate mass evaluation is combined with protein database search tools. It is quite clear that, while the nominal mass of a peptide has obviously no useful correlation to biomolecular structure, the accurate mass, instead, has a strong and direct correlation to structure that so far has not been exploited considerably by bioinformatic tools. It has already become obvious that accurate mass evaluation is going to become a central goal for bioinformatics strategies in the near future. Strategies for extracting structural, and even functional, information out of accurate mass values of biomolecules still have to be developed. Examples and prospects of accurate mass evaluation in bioinformatics for the field of proteomics are outlined in the following.

Keywords: de novo sequencing, peptides, FT-ICR, accurate mass, composition-based sequencing
much more important, however, is that information can tell us, what is not the case. In the following example (Figure 2), information tells us that the underlying compound cannot be a peptide (because of non-existing elemental compositions of possible peptides corresponding to the mass value of the example).

An important question, thus, is about the best strategy in processing information. From this point of view, information theory appears to become an important issue of bioanalytics in the coming years. Deleting redundant information and extracting irreducible information will be among the basic techniques of information-handling proteomics systems, as well as translation of observations into what is called information space. New types of question will arise, such as “what is the information space of an MHC system?”, or more basic, “what is the meaning of an accurate mass in information space?” Information connects facts. The last question will immediately lead us to ask whether there is a correlation between accurate mass and structure or (more speculative) between accurate mass and biological function.

**Accurate mass and structure**

To investigate a possible correlation between mass and structure, we need to be able to determine mass with the highest accuracy available, typically in the range of 1 ppm or below. Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) routinely provides such accuracy both with electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). The importance of mass accuracy results from the fact that combining the mass-defective accurate mass values of the atoms of large molecules leads to a mass distribution of possible accurate masses of molecules that is not homogeneous, but discrete, meaning that certain mass areas are forbidden for organic molecules and only discrete mass values are allowed. An instrument that determines mass with an uncertainty smaller than the gap between two possible discrete accurate mass values, allows us to achieve direct information on the possible elemental composition of the underlying substance.

This information can be used in different ways (Figure 3). Even without ultra-high mass accuracy, mass signals can be classified into certain substance classes. This process which we call “mass-based classification” (MBC), allows us to differentiate between, for example, peptides and lipids by looking at the accurate mass of the precursor ion. This is possible to some extent (and up to an upper mass limit) because the elemental composition of lipids is rather different from that of peptides, shifting accurate masses to higher values compared to peptides.14

Another application making use of very high mass accuracies is the determination of the composition of molecular ions of a known class. Units (“elements”) or building blocks of such molecule classes could be, for example, atoms, amino acids, monosaccharides or nucleotides. Calculating the number of possible compositions of a molecular ion of a certain mass shows that the uncertainty (variability) is going down in a stepped function of the accuracy of mass determination. For peptides in the mass range around 1000 u, for example, an accuracy of about 0.1 ppm leads to a fully determined system with respect to compositional variability (see Figure 4).

It is reasonable to correlate mass accuracy with parameters known from information theory such as “entropy”

\[
m/z = 1004.203 \text{ u ([M + H]^+), monoisotopic signal}\\
\text{What is the case? } \rightarrow \text{ Substance has one out of thousands of different elemental compositions}\\
\text{What is NOT the case? } \rightarrow \text{ Substance is not a peptide!}\\
\]

Figure 2. Information from data can be an inclusion or exclusion list.
and “missing information”. In Boltzmann’s theory, absolute entropy of a system can be defined as a function of the number of possibilities a system has (Figure 4). In this view, entropy can be interpreted as being equal to the amount of missing information, which would be necessary to describe the system completely (for example information on the positions and velocities of all individual molecules in a gas volume). In our case, the possibilities (or “microstates”) are the possible compositions that lead to the same macrostate, i.e. the same mass within a given uncertainty. Going to a higher mass accuracy thus means reducing the absolute entropy of the investigated system considerably. It is obvious, on the other hand, that entropy can be reduced only to a certain limit for peptide composition analysis, which is given by the fact that different amino acid compositions can have the same elemental composition. The number of amino acid combinations having the identical accurate mass is exponentially growing with mass. An ideal mass spectrometer with unlimited mass accuracy would thus not necessarily improve the ability to determine the amino acid composition of a larger peptide unequivocally from the precursor ion mass alone. It is thus necessary to acquire more information from other sources to reduce entropy further and to determine the stronger system. This information is available from tandem mass spectrometry (MS/MS) data, since fragment ions can be evaluated in terms of possible composition in the same way as precursor ions. The composition information of fragment ion signals can be combined with composition information of the precursor ion, resulting in intersectional composition information (Figure 5). In most cases, a fully determined situation is already achieved by interpreting three to five fragment ion signals, resulting in an unequivocal amino acid composition of the peptide.

In terms of information content, knowing the composition of a peptide is already very close to knowing the sequence. The final step in a CBS of unknown peptides is very simple compared to common de novo sequencing algorithms. A simple similarity test of the remaining possible sequences of the determined amino acids is usually sufficient to identify, uniquely and reliably, the real sequence. Theoretical MS/MS spectra are compared with the observed MS/MS spectrum and overlaps and deletions are evaluated into a matching score.

Results and discussion

An example for high-speed database-independent de novo sequencing of peptides is described in the following. A synthetic peptide unknown to the operator was investigated by static nanospray ionization on a Thermo Fisher "Finnigan LTQ FT" instrument. The doubly-charged peptide ion signal at \([M + 2H]^+ = 555.7523 \text{ u}\) was selected for MS/MS analysis, corresponding to a singly-charged monoisotopic mass of \([M + H]^+ = 1110.4973 \text{ u}\). Table 1 summarizes the determination of the peptide composition. Calculation of all possible amino acid combinations (AAC) having a monoisotopic mass of 1110.4973 \text{ u} +/- 1.0 ppm resulted in 2519 possibilities. This number was reduced by a factor of almost nine after evaluating the first fragment ion. The signal at 272.1717 \text{ u} was found to be a y-type ion with a PrR1 composition and no other possibility of reasonable fragments. The remaining possibilities of peptide composition, taking into account the presence of PR, were 294. Fragment ion 373.0794 \text{ u} was found to have two possibilities, a b-type ion, EpY, and a y-type ion, UpY. Presence of a pY reduced the number of possible peptide compositions to 53 and, at the same time, proved that the unknown peptide is phosphorylated. Evaluation of fragment ion 400.2665 \text{ u} testified to the presence of lysine, while 468.1528 \text{ u} showed the presence of...
(2) Enzymatic digestion is not restricted to using a well-defined single enzyme but can be performed randomly, with an unknown set of enzymes. Autoproteolysed samples can be used to validate results of the first method (PMF).

(3) The specificity and quality of separation steps can be more relaxed, since mixtures of proteins or complex multi-source peptide samples can be analyzed with sufficient reliability.

(4) Knowledge is created through the combination of new information (CBS of unknown peptides) with existing information (protein sequence databases).

Table 1. Course of determination of the amino acid composition of the unknown peptide observed at [M + H]* = 1110.49730 u.

<table>
<thead>
<tr>
<th>Mass M + H obs.</th>
<th>Type of fragment ion</th>
<th>Number of AAC at ±1.0 ppm</th>
<th>Delta [ppm]</th>
<th>AA compositions</th>
<th># Parent AAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1110.49730</td>
<td>M + H</td>
<td></td>
<td></td>
<td></td>
<td>2519</td>
</tr>
<tr>
<td>272.17166</td>
<td>y</td>
<td>1</td>
<td>-0.21</td>
<td>(P,R)_1</td>
<td>→ 294</td>
</tr>
<tr>
<td>373.07940</td>
<td>b</td>
<td>2</td>
<td>-0.34</td>
<td>(E_pY)</td>
<td>→ 53</td>
</tr>
<tr>
<td></td>
<td>y</td>
<td>2</td>
<td>-0.34</td>
<td>(U_pY)</td>
<td></td>
</tr>
<tr>
<td>400.26652</td>
<td>y</td>
<td>1</td>
<td>-0.40</td>
<td>(P,K,R)_1</td>
<td>→ 12</td>
</tr>
<tr>
<td>468.15277</td>
<td>b</td>
<td>1</td>
<td>-0.55</td>
<td>(U_L_pY)</td>
<td>→ 4</td>
</tr>
<tr>
<td>514.30945</td>
<td>y</td>
<td>2</td>
<td>-0.30</td>
<td>(G,P,K,R)_1</td>
<td>→ 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-0.59</td>
<td>[(G_2/N_1)E,K,L,P,U,R,pY]</td>
<td></td>
</tr>
<tr>
<td>1110.49795</td>
<td>M + H</td>
<td>2 under cons. of P&gt;0 K&gt;0 R&gt;0 pY&gt;0</td>
<td>-0.59</td>
<td>[(G_2/N_1)E,K,L,P,U,R,pY]</td>
<td></td>
</tr>
</tbody>
</table>

UL. Finally, with the evaluation of 514.3095 u only two possible peptide compositions remained, differing in iso- lemental G_2 and N_1 respectively. The total computing time for determining the composition was less than 100 ms. The final composition string was

$$\left[ \frac{G_2}{N_1} \right] E.K.L.P.U.R.pY_1$$

with $$(G_2/N_1)$$ indicating the two possibilities and [...] indicating that the actual sequence can be any permutation of all amino acids or groups within the brackets. Sequence determination, as a second step, was performed by comparing theoretical fragmentation spectra and observed MS/MS data. A total number of 20,160 possible sequences for the above composition was reduced before comparison by taking into account group information from fragment ion analysis (such as PR and ULpY forming groups). The resulting sequence was determined correctly as

U_LpYENKPR.

Total computing time was less than 80 ms for the second step, thus less than 200 ms were needed for the complete de novo sequencing procedure.

Advantages of composition-based sequencing

One of the attractive applications of composition-based approaches is to improve the quality of common database search strategies in proteomics and to reduce the preparational requirements of analytical procedures. Combining database search methods^{2,3,4} such as peptide mass fingerprinting (PMF) with CBS of peptides allows the improvement of proteome analysis in several ways.^{2}

(1) CBS is a second and independent analytical method that can be used to validate results of the first method (PMF).

(2) Enzymatic digestion is not restricted to using a well-defined single enzyme but can be performed randomly, with an unknown set of enzymes. Autoproteolysed samples can thus be analyzed reliably as well.

(3) The specificity and quality of separation steps can be more relaxed, since mixtures of proteins or complex multi-source peptide samples can be analyzed with sufficient reliability.

(4) Knowledge is created through the combination of new information (CBS of unknown peptides) with existing information (protein sequence databases).

Making use of increased accuracy of mass determination is a rather obvious way to go, almost naturally leading to higher quality results, even when using the same already established data analysis techniques. One has to understand, however, that when passing a certain level of accuracy the informational quality of mass spectrometric data is changing completely, leading us to new ways of data interpretation that can be extremely useful and efficient. The discrete nature of accurate masses has, so far, not been exploited much but is now accessible routinely due to the availability of reliable and easy to use FT-ICR mass spectrometers in many proteomics laboratories. It is necessary to understand the goal and the prospects of this kind of data interpretation to produce new highly efficient bioinformatic software tools.

Acknowledgment

Financial support by the European Union (STREP LSHG-CT-2005-518194) is gratefully acknowledged.

References


FT-ICR-MS for Peptide Sequencing


“The world is all that is the case.”; “The world is the totality of facts, not of things.” in: Ludwig Wittgenstein, Tractatus Logico-Philosophicus, Cambridge, UK (1921).

Received: 28 November 2006
Revised: 26 February 2007
Accepted: 28 February 2007
Publication: 20 March 2007