analytical chemistry

Article

pubs.acs.org/ac

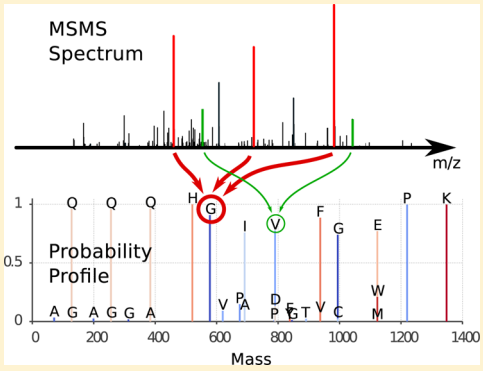
MS/MS Spectra Interpretation as a Statistical–Mechanics Problem

Mauro Faccin† and Pierpaolo Bruscolini*

Departamento de Física Teórica & Instituto de Biocomputacíon y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, c/Mariano Esquillors s/n, 50018 Zaragoza, Spain

Supporting Information

ABSTRACT: We describe a new method for peptide sequencing based on the mapping of the interpretation of tandem mass spectra onto the analysis of the equilibrium distribution of a suitably defined physical model, whose variables describe the positions of the fragmentation sites along a discrete mass index. The model is governed by a potential energy function that, at present, we derive ad hoc from the distribution of peaks in a data set of experimental spectra. The statistical–physics perspective prompts for a consistent and unified approach to de novo and database-search methods, which is a distinctive feature of this approach over alternative ones: the characterization of the ground state of the model allows the de novo identification of the precursor peptide; the study of the thermodynamic variables as a function of the (fictitious) temperature gives insight on the quality of the prediction, while the probability profiles at nonzero temperature reveal, on one hand, which fragments are more reliably predicted. On the other hand, they can be used as a spectrum-adapted, a posteriori score for database search. Results obtained with two different test data sets reveal a performance similar to that of other de novo and database-search methods, which is reasonable, given the lack of an aggressive optimization of the energy function at this stage. An important feature of the method is that it is quite general and can be applied with different choices of the energy function: we discuss its possible improvements and generalizations.



Tandem mass spectrometry is widely used in the field of biochemical analysis of unknown samples of protein, generally embedded in an automated high-throughput pipeline, that yields huge amounts of data, requiring an automated spectrum-interpretation tool. In principle, a tandem mass spectrum contains all the necessary information about the peptide it comes from. In practice, reading out the sequence is always a difficult task, since each spectrum is the statistical outcome of the microscopic rules governing the energy transfer and stochastic fragmentation of the precursor peptide under collisions, in the presence of "noise" sources of different kinds. Ab initio predictions of the spectrum on the grounds of just the physics of molecular collisions are practically impossible, so that the identification of the precursor sequence involves the use of ad hoc score functions to rate the match between the theoretically predicted spectrum and the experimental one. Most often, the search space is limited to the sequences of known proteins ("database search" approach), which is more practical and efficient but also more limited, than de novo methods, inferring the peptide from just the information contained in the spectrum. Remarkably, a central problem is related to assessing the reliability of the prediction, mainly related to the prediction of false positives, i.e., wrong sequences with high scores. This is true also for database-search methods, especially in the case of proteins of low abundance, yielding few good spectra and reducing the possibility of cross-validation of the predictions. Indeed, it is estimated that just 20% of MS/MS spectra is successfully identified by database-matching algorithms.[1] Unfortunately, since the scores are not rooted on a sound modeling of the fragmentation process, the probability distribution of the score values cannot be derived from first principles and involves strong approximations. Thus, several methods have been proposed to score the value of the predictions, in a postprocessing of the sequencing process,[2] either recurring to empirical, database-dependent estimates of error rates or searching a suitably designed decoy database, to estimate the probability that the resulting score could be obtained by a match to a random peptide. Other approaches train a classification algorithm on spectra of known identity,[3] to distinguish correct and incorrect matches. As pointed out by Kim and co-workers,[2] the need for a decoy database is a consequence of the inability to solve the spectrum matching problem: Given a spectrum S and a threshold T for a scoring function, find the probability that a random peptide matches S with score greater than T. This quantity is quite difficult to estimate correctly and is usually replaced by the false discovery rate, which is not a characteristic of the individual spectrum but rather an average property, i.e., the fraction of incorrect guesses among all identifications with score greater than T.

The situation is perhaps even worse for de novo interpretation, where all the score functions are tested against a test bed of spectra, whose interpretation is considered reliable.

Received: September 3, 2012
Accepted: April 12, 2013
Published: April 12, 2013

   dx.doi.org/10.1021/ac4005666 | *Anal. Chem.* 2013, 85, 4884–4892

Thus, while indications of the average precision and recall of the predictions can be given for each method, there is not a robust way to determine how reliable a particular interpretation is. Moreover, the sequence space explored is much huger, and there is no decoy database to learn a null-hypothesis distribution. In general, the best scoring sequence might not be the correct one. A common feature to all sequencing algorithms is that they report an ordered list with the best ranking solutions, that is not granted, though, to include the correct one, for any reasonable length of the list.

Another characteristic of de novo methods is the use of dynamic programming algorithms in the search for the best and suboptimal solutions,[4] that can be seen as a special case of a max-sum algorithm for graphical models. Indeed, the problem can be cast as a graphical model where the graph is either a "spectrum graph"[5] or a mass-array,[6] while the score-function to maximize is simply the sum over the contributions of each N-terminal and C-terminal ("prefix" and "suffix") peptides, with their related fragment ions, to the matches between theoretical and observed peaks. Even if in the following we are not going to use the techniques developed for graphical models, we take advantage of the connection between max-sum and sum-product algorithms on graphical models and the equilibrium of a corresponding statistical mechanics model,[7] to propose a statistical–physics approach to the de novo interpretation problem.

In physical terms, using the negative score as an energy, we can say that de novo algorithms find the minimal energy state ("ground state"), or zero-temperature equilibrium, in the appropriate sequence space, with suboptimal solutions representing the first "excited states" of the energy landscape. In statistical mechanics, a way to explore a system in the vicinity of its ground state consists of raising its temperature and studying the equilibrium solution, as more and more excited states come into play. The equilibrium state will not give us the individual details (in this case, the sequence) of the excited states, but the averages of the state variables can be related to the probability of finding a residue at a certain position, informing us on the regions where the interpretation is more robust and prompting us for the implementation of a score for database search. Moreover, "high" values of the average energy and entropy at low temperature could reveal unreliable interpretations, related either to spectra with a few good matches or to high-entropy energy landscapes, with many alternative sequences at a small gap from the ground state one.Thus, the introduction of an artificial temperature and the study of the resulting thermodynamic equilibrium could also provide some valuable internal indication of the quality of the interpretation, in a sense analogous to the false positive rate mentioned above.

In order to implement such scheme, we need, first, a proper way to map the problem of spectra interpretation on that of finding the thermodynamic equilibrium of a suitable physical system and, second, an efficient way to perform calculations. In the following, we will deal with both issues: we will introduce a discrete unidimensional system, whose states encode all possible amino-acid sequences of appropriate mass. We will state the general form of the energy function of the model in terms of just on-site and next-neighbors interaction. We will calculate exactly the partition function of the model as well as some thermodynamic observations, resorting to a transfer-matrix technique, and we will discuss how the equilibrium results can be mapped back to MS/MS spectra interpretation and how to assign a significance value to the resulting sequence prediction.

METHODS

We look at an experimental MS/MS spectrum Σ as the result of stochastic fragmentation of an ensemble of identical parent peptides P^*, generating "true" peaks, overlapped with the extra peaks produced by R, a noise source. Σ, R, and P^*, should be considered as random variables, with different probability distributions.

To estimate the probability that a proposed sequence P is indeed the true precursor ion P^*, we formulate the problem of spectra interpretation as a Bayesian inference problem. As detailed in the Supporting Information, we can write the probability $p(P|\Sigma,R)$ of a parent peptide P given a spectrum Σ and the noise source R as:

$$p(P|\Sigma,R) = p(\Sigma|P,R)\frac{p(P)}{p(\Sigma)} \tag{1}$$

where we have assumed that R is independent from Σ and P, so that its a priori probability $p(R)$ simplifies out. Interpreting the spectrum will be equivalent to finding the peptide sequence P' that maximizes the probability 1:

$$P' = \arg\max_{P}[p(\Sigma|P,R)p(P)] \tag{2}$$

(We neglect the denominator, independent from P.) P' is the best prediction we can give of the true precursor ion P^*. In the above expressions, the a priori probability of sequence P, $p(P)$, is unknown, but some reasonable assumptions can be made on it (see Supporting Information). However, the key points here are as follows: first, how to define a physical model that encodes all the sequence space compatible with the given parent mass; second, how to estimate the probability $p(\Sigma|P,R)$ to observe the spectrum Σ given a sequence P and a source of statistical noise R and how to encode such information in the model, and third, how to explore the configuration space efficiently to find P'. Let us start with the first issue. The position of a MS/MS peak informs on the m/z ratio of the corresponding fragment but does not depend on its sequence. We can use this fact to avoid the book-keeping of a combinatorial number of possible sequences and to define a physical system whose variables carry information on the accumulated mass and charge at each site, as well as some other quantities that we will need to completely characterize the parent peptide. As in ref 6, we define a mass array of $M + 1$ sites from 0 to M, the discretized monoisotopic mass of the parent peptide (see Supporting Information for details). Any sequence of total mass M will map onto a set of "fragmentation sites" ν in the lattice, corresponding to the masses of the prefix fragments ending at each peptide bond. We introduce a list $\mathcal{A}$ of all the residues that can appear in a peptide sequence (possibly enlarged to include post-translational modifications of the residues, if needed), together with their *characteristic numbers* $\omega(a)$, $(a \in \mathcal{A})$, that specify their discretized mass, maximal charge, and list of neutral losses they can undergo. The pattern of characteristic numbers is specific to each residue (apart from the degeneracy Ile-Leu; see Table S1 in the Supporting Information).

We map the possible amino-acid sequences to model configurations by introducing, for each site, a variable $r \in [0, r_{max}]$, where r_{max} is the biggest mass in $\mathcal{A}$. We enforce (as an

dx.doi.org/10.1021/ac4005666 | Anal. Chem. 2013, 85, 4884−4892

energy constraint, see eq 4) that the only values allowed at ν are $r_\nu = r_{\nu-1} + 1$ or $r_\nu = 0$, the latter just holding when $r_{\nu-1} = m(a) - 1$, for some residue $a \in \mathcal{A}$; see Figure 1. The above rule,

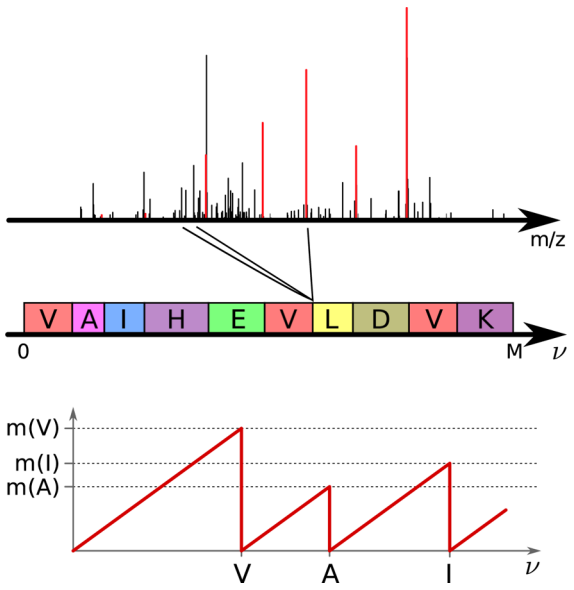


Figure 1. Schematic representation of mapping onto the physical model: a proposed sequence (middle) is laid down on the discretized mass axis; its peptide bonds correspond to the possible fragmentations sites, from which several N-terminal and C-terminal ions are produced. These ions can match peaks of the true parent (in red), noise peaks (in black), or no peak at all of the experimental spectrum (top): accordingly, an energy is associated to the configuration. Bottom: The actual model configuration corresponding to the first residues in the sequence: the r_ν-variables, in red, show a sawtooth profile, increasing up to the residue mass and then dropping to zero. For simplicity, the other dynamic variables, accounting for charge, neutral losses, etc., are not reported. Notice that the state σ_ν at each site does not inform on the sequence of the prefix (or suffix) peptides; however, the best sequence is recovered by the calculation of the average probabilities, eq 10.

together with the boundary conditions $r_0 = r_M = 0$, generates all possible sequences of total mass M, with $r_\nu = 0$ at each fragmentation site ν. To locate the corresponding theoretical peaks, we need to know, at each ν, the maximal charge of a N-terminal ion (q_ν^N) and a C-terminal one (q_ν^C), according to the charge of the precursor peptide and the number of K, R, and H residues, as well as the number and kind of neutral losses they can present. Moreover, a binary variable $\pi_\nu = 0, 1$ will be used to implement enzyme-specific cleavage rules: in the following, we will specify to the most common case of trypsin.

The constraints on the charge, neutral losses, and π_ν can also be written in terms of the variables at sites $(\nu - 1)$ and ν (see Supporting Information for details). Basically, they imply that, if $r_\nu > 0$, the other state variables retain the same value of position $(\nu - 1)$: the system *remembers* the state it had at the previous site. If $r_\nu = 0$, a new residue is "started" at ν, and the other state variables are updated according to the pattern of characteristic numbers $\omega(a)$ of the residue a terminating at ν. The above variables inform us on where, along the mass lattice, the fragmentation can occur and what kind of fragments can emerge, while we introduce a variable $\xi_\nu^{s_i} = 1, 0$ according to whether the ion of type s_i was produced at ν or not. Collecting all the state variables introduced above in a global one, σ_ν, the set of all σ_ν, together with their constraints, will describe all the possible parent peptides and specify the relevant information to produce the corresponding fragmentation and ionization patterns.

The behavior of the model will be ruled by the Boltzmann probability associated to each configuration $\phi = \{\sigma_\nu, \nu = 0, ..., M\}$

$$p(\phi|\Sigma,T) = Z^{-1}e^{-\beta H(\phi,\Sigma)} \tag{3}$$

where $\beta = 1/T$ and $H(\phi,\Sigma)$ is the energy function of the model. The partition function $Z = \Sigma_\phi e^{-\beta H(\phi,\Sigma)}$ involves a sum over all the states of the model and thus over *all* the sequences with total mass and charge as the parent peptide. All the relevant observables can be derived from Z, so that its evaluation represents the major challenge. We write the energy as: $H = \Sigma_{\nu=1}^{M} H_{\nu-1,\nu}$, where

$$H_{\nu-1,\nu} = H_\nu^1(\sigma_\nu; \Sigma) + H_{\nu-1,\nu}^2(\sigma_{\nu-1}, \sigma_\nu) \tag{4}$$

Here, the two-sites term of the energy $H_{\nu-1,\nu}^2$ implements the constraints representing the allowed transition from a site to the following one: it is zero if the values of $\sigma_{\nu-1}$ and σ_ν are compatible and infinity otherwise, thus strictly prohibiting forbidden configurations (see Supporting Information). The connection to the experimental spectrum is contained in the single-site term H^1, that represents the energy of a fragmentation in ν and depends only on the fragmentation pattern allowed by the state variable σ_ν at site ν and the quality of the match of the resulting ions to the experimental spectrum Σ. The simple form of eq 4, with just on-site and next-neighbors interactions, rules out the possibility of accounting for the effects that distant residues may have on the fragmentation at ν or of excluding that different ions match the same peak. On the other hand, it allows the exact calculation of the sequence that best matches the spectrum, which should coincide with the true parent if H is sufficiently good. Thus, eq 4 represents a trade-off between accuracy of the score-function and accuracy of the exploration of the configuration space.

In order to find an explicit expression for H_ν^1 of eq 4, we have to deal with the second key issue: how to estimate the probability $p(\Sigma|P,R)$ to observe the spectrum Σ given a sequence P and a source of statistical noise R. If we had a good physical characterization of the fragmentation process from first principles, we could associate a probability to any fragmentation and compare the predicted distribution of intensities with the experimental spectrum. The lack of such a suitable ab initio characterization prompts us for a different approach, where information, collected from a database of spectra, is used to derive an ad hoc energy potential; the experimental spectrum then acts as an "external field" driving the system toward the most-likely parent sequence. We will use the latter approach here, which is common to all de novo methods: for any given sequence P, we will calculate its fragment ions and will match the experimental peaks to such theoretical ions, rewarding the associations according to an ad hoc defined energy (depending on peak intensities and positions) derived from the analysis of a database of reliably interpreted spectra (the "learning data set" LSET, in the following; see Supporting Information for its definition and for calculation details). We finally obtain:

$$p(P|\Sigma,R) \propto \prod_{\nu\in\mathcal{F}(P)} \prod_{s_i\in\mathcal{T}_\nu(P)} \left(\sum_{\xi_\nu^{s_i}=0,1} e^{-H_\nu(s_i,\xi_\nu^{s_i})} \right) \tag{5}$$

dx.doi.org/10.1021/ac4005666 | Anal. Chem. 2013, 85, 4884−4892

where $\nu \in \mathcal{F}(P)$ indicates the possible fragmentation sites of peptide P (i.e., the peptide bonds, in our framework) and $s_i \in \mathcal{T}_\nu(P)$ are the different kinds of ion species ($s_i = y$, b, etc.) that theoretically can be generated at ν.[8] We have introduced:

$$H_\nu(s_i, \xi_\nu^{s_i}) = \mu + \delta_{\xi_\nu^{s_i},1}(\delta_{I(s_i)} h_\nu^M(s_i, I(s_i)) + (1 - \delta_{I(s_i)}) h_\nu^{\bar{M}}(s_i)) \quad (6)$$

where μ is an energy cost, associated to the a priori probability to produce a fragment ion of any kind, that will be chosen to reproduce, on average, precursor peptides of the correct length (see Supporting Information); $\delta_{I(s_i)} = 1$ if there is an experimental peak sufficiently close to the position of the theoretical fragment s_i, and $\delta_{I(s_i)} = 0$ otherwise; $h_\nu^M(s_i, I(s_i))$ and $h_\nu^{\bar{M}}(s_i)$ are energies associated to the matching of the ion s_i, produced at ν, to its "image peak" $I(s_i)$ in the spectrum or to the lack of matching peaks for s_i, respectively. They are written as:

$$h_\nu^M(s_i, I(s_i)) = -\log \frac{p_\nu(s_i: I(s_i))}{p(R: I(s_i))} \quad (7)$$

$$h_\nu^{\bar{M}}(s_i) = -\log p_\nu(s_i: \varnothing) \quad (8)$$

where $p_\nu(X : I(s_i))$ is the probability that the peak $I(s_i)$ was indeed produced by $X = s_i$ or by noise $X = R$, and $p_\nu(s_i : \varnothing)$ is the probability that the ion s_i produced at ν did not yield a peak in the spectrum. Such probabilities are obtained from the LSET (see Supporting Information) by considering the distribution, in the (mass,intensity)-plane, of the peaks corresponding to each type of ion products and to noise. Notice that the expressions of the energies eqs 7 and 8 correspond to the negative of the score usually adopted in de novo methods (e.g., refs 5 and 6), so that our approach can be applied also with different definitions of the probabilities. Comparing eq 3 with eq 5, it can be seen that the former reproduces the latter with $T = 1$ and the identification

$$H_\nu^1(\sigma_\nu, \Sigma) = \sum_{s_i \in \mathcal{T}(\sigma_\nu)} H_\nu(s_i, \xi_\nu^{s_i}) \quad (9)$$

which completes the mapping between the interpretation problem and the physical model.

We are finally left with the problem of exploring the configuration space to find the one representing the best parent sequence. In the standard approach, a sequence probability is calculated as a product of independent single-node factors, as in eq 5, that score the match between theoretical fragments and the spectrum. In practice, instead of dealing with the product of probabilities, it is handier to use the sum of their logarithms as a score and to find the precursor sequence as the one maximizing such sum. In our approach, the log-probabilities acquire an independent status, as energies of a physical system, and coincide with the logarithm of the probabilities only for $T = 1$. At any T, the equilibrium state of the physical system will be an ensemble of ground state and excited states, encoding the optimal and suboptimal sequences, populated according to the Boltzmann probability eq 3. As T approaches 0, such an ensemble collapses into just the microstate of minimal energy, thus recovering the same solution as the standard approach. At higher temperatures, several model configurations will be populated; the average properties of this ensemble will translate into a sequence profile of the most likely sequences, instead of a list of suboptimal sequences; the latter profile will be useful to explore the possible alternative interpretation and assess the goodness of the prediction. Despite that the number of terms in the sum is exponential in M, a transfer-matrix formalism (see Supporting Information) allows one to calculate the partition function Z exactly, as well as other relevant equilibrium quantities, like the average energy $U = \langle H \rangle$, coinciding at $T = 0$ with the energy of the most likely peptide; the entropy $S = -\Sigma_\phi p(\phi) \ln p(\phi)$, giving a measure of how many sequences are "populated" at a given temperature. The most likely sequence can be found resorting to the quantities

$$p_\nu(a) = \langle \Delta_{\nu-1,\nu}^a \rangle \quad (10)$$

which represent the probability that a residue of type a ends at ν (see Supporting Information): at any temperature, the best sequence P' is recovered by starting on the last site of the lattice $\nu = M$, looking at the most likely residue terminating there and tracking back the position ν of the preceding fragmentation; the process is iterated in backward steps until reaching $\nu = 0$. At $T = 0$, the $p_\nu(a)$ is different from zero just at the fragmentation sites of the lowest energy sequence, while at higher temperatures several configurations, corresponding to suboptimal sequences, will be populated, and $p_\nu(a)$ will be different from zero for an increasing number of positions ν and species a. This will generate a probability profile, allowing the identification of the most reliable fragmentation sites.

An interesting quantity related to the profile is the "sequence entropy"

$$S^s = -\sum_{\nu=0}^{M} p_\nu \log(p_\nu) \quad (11)$$

where $p_\nu = \Sigma_a p_\nu(a)$ is the probability that ν is a fragmentation site.

To test the de novo method, we apply it to the spectra of the test set (TSET1) introduced in ref 5 and later used in ref 9 composed by 280 spectra of double charged peptides of up to 1400 Da, produced with tryptic cleavage. For each spectrum, a reliable interpretation is available.[5] We identify and filter out isotopic peaks by the same procedure applied to the LSET. When greater statistics is needed, we refer to the Extended Learning data set (ELSET, see Supporting Information). We compare, at the level of the fragmentation sites, the predicted best sequence and the predicted probability profile with the provided parent sequence, that we refer to as the "true sequence" P^* in the following. We proceed as follows: for the profiles, we define the total "predicted positive" fraction as the sum of all the probabilities eq 10, at all sites: $PP = \Sigma_\nu \Sigma_a p_\nu(a)$, and analogously, the "true positive" fraction (i.e., the fraction of predicted fragmentation sites that are correct) as $TP = \sum_{\nu \in \mathcal{F}(P^*)} \sum_a p_\nu(a)$, where $\mathcal{F}(P^*)$ is the set of fragmentation sites of the true sequence P^*. The "real positive" value RP is simply the number of residues of P^*. Moreover, we define the corresponding quantities for the predicted best sequence P': PP′ is the length in residues of this sequence and TP′ is the overlap $\mathcal{F}(P') \cap \mathcal{F}(P^*)$ between the sets of fragmentation sites while the real positive number RP′ is the same as before.

To quantify the goodness of the profile, we compute the *precision* $\Pi = TP/PP$ (the fraction of predicted fragmentation sites that are correct), the *recall* $\Gamma = TP/RP$ (the fraction of true fragmentation sites that are correctly predicted), and their harmonic mean, the F-value: $\Phi = 2\Pi\,\Gamma/(\Pi + \Gamma)$ Analogously,

dx.doi.org/10.1021/ac4005666 | Anal. Chem. 2013, 85, 4884−4892

we define the precision Π′, recall Γ′, and F-value Φ′ for the predicted best sequence by the use of TP′, PP′, and RP′.

To perform the database search, we build up the target database by digesting the SwissProt database strictly according to the tryptic rules, allowing at most one missing cleavage. For each peptide x in the database, we calculate:

$$p^{\mathrm{db}}(x) = \prod_{\nu \in \mathcal{F}(x)} p_\nu(a_\nu) \tag{12}$$

where a_ν is the residue of x ending at ν and $p_\nu(a)$ is defined in eq 10 and calculated with our program (in the following: *T-novoMS*). Whence, for the sequence P_1 in the target database that maximizes $p^{\mathrm{db}}(P)$, we calculate the "z-score" for spectrum Σ:

$$z_{\mathrm{T}}(\Sigma) = \frac{(\bar{e} - e(P_1))}{\sigma_e} \tag{13}$$

where $e(P) = -\log(p^{\mathrm{db}}(P))/L(P)$, with $L(P)$ being the length in residues of sequence P and $\bar{e}$ and σ_e being the average and standard deviation of the distribution of e in the database.[10] Then, we generate a decoy database, reversing the protein sequences and applying the same procedure as above,[11] calculating $z_D(\Sigma)$. We apply the method to the same set TSET1 as before and also to a database of 18169 spectra[12,13] (TSET2), calculating the false discovery rate (FDR) and coverage of the predictions. Given a threshold z_0, we define the FDR as FDR = (do + 2db)/(db + tb + to), where db ("decoy better") and tb ("target better") are the number of spectra that get a better z-score in the decoy database or in target database, respectively, and do ("decoy only") and to ("target only") are the number of spectra that get a score above the threshold z_0 only in one of the databases.[14] For TSET1, whose "true" parent sequences are known, we also apply an alternative definition of FDR, based just on the target z-score z_T: for a given z_0, FDR′ = *FP*/PP, where PP is the number of spectra Σ with $z_T(\Sigma) > z_0$ and *FP* is the number of spectra whose true precursor sequence does not coincide with the best in the database, yet $z_T > z_0$. Both for the de novo and database-search approaches, we also check the effects of a prefiltering of the spectra, performed by selecting six peaks in each window of 100 Da and discarding the others as noise.[6]

RESULTS

Low-Temperature Regime: Peptide Identification. The state of the system at $T = 0$ provides the predicted parent sequence but is difficult to study due to computational numerical divergences. We verified that $T = 1$ is sufficiently low to ensure that the minimal energy state is clearly identifiable. We compared the results of the algorithm with other popular de novo sequencing algorithms, such as NovoHMM,[9] Lutefisk,[15,16] Pepnovo,[5] and MS-novo.[6,17] For every spectrum in TSET1, we compare the inferred sequence with the "true" one deposited in the database, computing precision Π′, recall Γ′, and F-value Φ′; then, we average them over all the spectra.[18] The results in Table 1 show that the performance of the model is comparable to that of the common alternative de novo softwares. For instance, the F-value, which combines precision and recall and is 1 only if the model reproduces all and only the correct fragmentation sites, is essentially the same as that of MS-Novo. It is interesting to notice that the correct estimate of the precursor mass is critical for performance and is more relevant than a noise prefiltering of the spectrum, performed as described in Methods. The mass error is due to the discretization of the mass, that involves some truncation of the true residues masses, yielding a sequence-dependent cumulative effect that can cause a unit shift in the parent mass (see Supporting Information).

Table 1. Average Precision Π′, Recall Γ′, and F-value Φ′ of the Match between the Predicted and "True" Parent Sequence, for Different Algorithms[a]

model	Π′	Γ′	Φ′	n
Lutefisk	0.717	0.664	0.688	43
PepNovo	0.691	0.665	0.676	109
NovoHMM	0.786	0.778	0.781	12
MS-Novo	0.767	0.695	0.727	19
T-novoMS	**0.713**	**0.700**	**0.705**	**27**
***T-novoMS*-p**	**0.734**	**0.719**	**0.726**	**15**
T-novoMS-M	0.747	0.732	0.739	0
T-novoMS-p-M	0.755	0.740	0.747	0

[a]The last column shows the number of wrongly estimated precursor masses. In boldface are the results for our method. We report also the results obtained for *T-novoMS* upon prefiltering the spectrum (-p) or forcing the mass of the theoretical sequence as an input (-M), as an ideal case.

Temperature Dependence and Quality Checks. Figure 2 and Figure S2 of the Supporting Information show the probability profile $p_\nu(s_i)$ of eq 10, obtained for a sample spectrum of the precursor sequence ALAEHGIVFGEPK at two different temperatures. These profiles are analogous to those introduced in ref 19 with a different method. The probability of a residue ending at a certain site is 1 or 0 for T approaching 0, and at low T, just a reduced number of states contribute, which

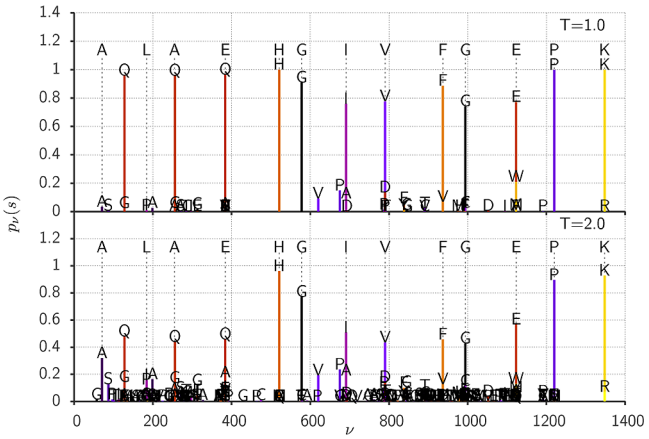


Figure 2. Probability profiles at different temperatures, corresponding to a high quality prediction (Φ = 0.8 at $T = 1$) for a sample spectrum. The height of the bar at ν represents $p_\nu(s)$, and its label is the residue s to which it refers. Bars exceeding 1 represent the true parent sequence, as deposited in the database. Notice that a correct fragmentation is missed (Leu2) at $T = 1$, yielding a wrong peptide length. Also, the first two fragmentations sites are wrongly predicted, causing a wrong interpretation E4Q of the following fragmentation. At lower temperatures, this features is even more evident (see Supporting Information). At $T = 2$, all the fragmentation sites of the true sequence are recovered with nonzero probability, but the ground-state sequence still dominates. Notice that the profile suggests which fragmentations are more reliably predicted. At higher temperatures (not shown), the equilibrium configuration is dominated by entropy and no interesting information can be extracted from the equilibrium state.

dx.doi.org/10.1021/ac4005666 | *Anal. Chem.* 2013, 85, 4884−4892

allows an easy readout of the best sequence. However, the probability profile, for $T > 0$, contains the contribution of every sequence in the conformation space, and upon increasing the temperature, alternative fragmentations appear. Also, the higher energy available allows one to overcome the cost μ in eq 6 of increasing the number of residues, yielding longer precursor peptides on average.

The existence of predictions with very different quality, as measured by the F-values Φ or Φ', proposes the challenge of recognizing a good prediction from a bad one. Table 1 provides information on the average quality of the predictions but not on the quality of an individual one. However, the possibility to tune the temperature to extract information about other low energy states, describing alternative sequences, can provide us with valuable tools to assess the quality of the prediction. Therefore, we look for some thermodynamic quantities correlating with the F-value, that can be used as a predictor of the latter, and analyze ELSET (see Supporting Information), to get a better statistics. We have found that the quantity that best correlates with the F-value is the "sequence entropy" eq 11, that informs us on the number of alternative sequences having an energy close to the minimal one and affects the goodness of the interpretation. Figure 3 shows the distribution

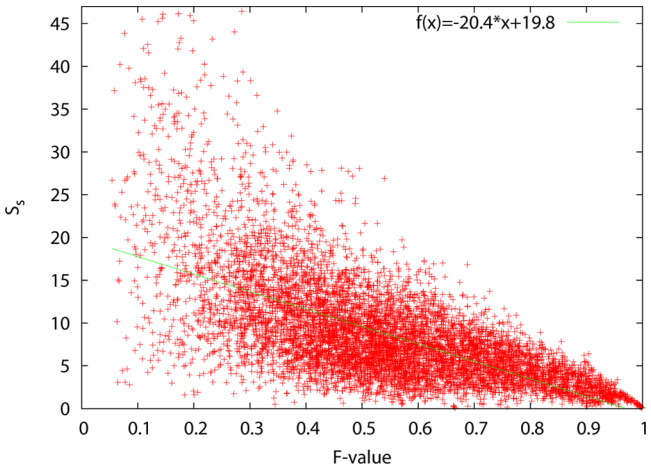


Figure 3. Correlation of the symbol entropy S^s and the quality measure F-value Φ, at temperature $T = 1$; the latter is a sufficiently low temperature to allow one to identify the best precursor sequence but already has a reasonable population in alternative conformations to give information about the low-lying structure of the solution space. The precursor mass is calculated from the true sequence. Data from the 7839 spectra with $Q = 2$ of the ELSET (correlation coefficient $r = -0.648$).

of the experimental spectra according to their entropy and F-value Φ, calculated for all the doubly charged spectra from the learning data set. The value of the correlation coefficient tells us there is a linear trend in the data, even if the distribution is quite broad: at low values of the entropy, this is related to the existence of some spectra for which the best solution is very stable and nevertheless wrong, which can be attributed to a limitation of the design of the energy function. Despite these limitations, important information on the quality of the prediction can be extracted from the data of Figure 3. For instance, we can select $\Phi_0 = 0.8$ as a threshold for "good" predictions and see how the spectra with entropy below (or above) a given threshold are classified according to this criterion.

Table S5 in the Supporting Information reports several indicators of the relationship between sequence entropy and prediction F-value. We see that only at very low entropies $S^s < 1$ we are able to single out good predictions with a high reliability, but the coverage is low (only 10% of the good sequences present such a low entropy). On the other hand, predictions with $S^s > 5$ are of poor quality 89% of the times. Such results are not yet sufficient to provide a definite knowledge of the value of the prediction but are indeed a first step toward the definition of an intrinsic quality indicator of the peptide identification, a feature that is missing in other de novo approaches. Future developments will aim at improving the energy function, to have a less disperse distribution and a shift toward higher values of the F-value.

On the other hand, the information contained in the probability profile allows us to determine which fragmentations are more robustly predicted (i.e., maintain a high probability upon increasing the temperature). Figure 4 reports the

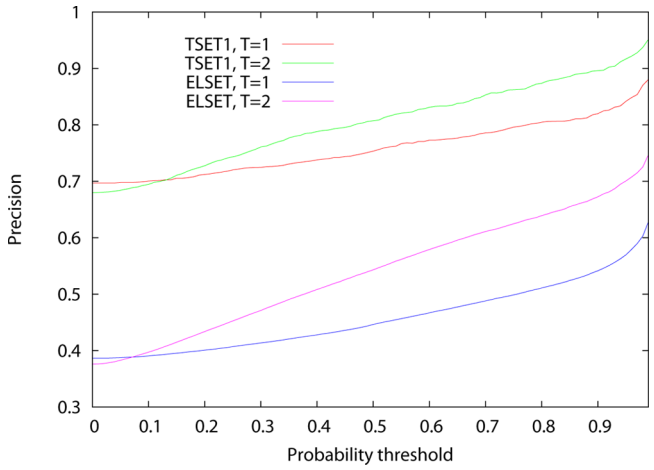


Figure 4. Dependence of the precision Pr = STP/SPP of the predicted fragmentation sites on the probability threshold p_0, calculated at $T = 1$ and $T = 2$ for the data set TSET1 and ELSET, with noise-prefiltering. Here, STP = $\Sigma_{\Sigma \in SET}$TP(Σ) and SPP = $\Sigma_{\Sigma \in SET}$PP(Σ), where SET is either TSET1 or ELSET, PP(Σ) is the number of fragmentation sites ν, in the profile for spectrum Σ, whose fragmentation probability is $p_\nu > p_0$, and TP(Σ) is the number of the correctly predicted fragmentations sites, upon comparison with the true parent of Σ.

dependence of the precision of the prediction on the fragmentation-probability threshold (i.e., the number of correctly predicted fragmentation sites over the total number of predicted sites with probability over the threshold), for two different temperatures, for the ELSET and TSET1. As it may be expected, at both temperatures, sites ν with higher probabilities are more likely to correspond to correct fragmentation sites. However, the precision increases upon raising the temperature, as well as the dependence on the threshold: the possibility to explore alternative regions of the sequence space reveals which fragmentation sites are less reliably predicted. Eventually, every probability profile will loose its significance when the temperature is high enough that the entropy dominates over the energy (see Figure S3 in the Supporting Information). We cannot propose a well-posed recipe for the choice of the optimal temperature, that will be, in general, spectrum dependent: in the following, we will work at $T = 2$ on an empirical basis, leaving a more rigorous analysis for future developments.

dx.doi.org/10.1021/ac4005666 | Anal. Chem. 2013, 85, 4884−4892

***T-novoMS* as a Database Search Tool.** The main reason why de novo sequencing methods perform worse than database-search ones is that in the former case the sequence space explored includes all possible sequences, most of which will not correspond to real protein sequences. We may ask therefore to what extent the efficiency of *T-novoMS* will increase when applied to the reduced sequence space of a database. Different than with other tools, *T-novoMS* provides a probability profile that can be quickly matched against each peptide in the database, associating the score eq 12 to it. Table 2 reveals that *T-novoMS* is able to retrieve correctly 84% of the

Table 2. Frequency of the Assignment the "True" Precursor Peptides to the Best Ranks, According to Our Method *T-novoMS* (at T = 2, with Noise Prefiltering) and MASCOT, among a Peptide Databased Derived from Swiss-Prot[a]

rank	*T-novoMS*	MASCOT
1	236	245
2	10	6
3	2	
>3	4	

[a]In *T-novoMS* database search, sequences differing by I↔L are considered the same. In 14 cases, *T-novoMS* assumes a wrong parent mass, and in the other 14 cases (all corresponding to nontryptic sequences), it assigns a null probability to the "true" sequence. In 29 cases, we were not able to find the position of the "true" sequence in the list reported by MASCOT.

sequences corresponding to the spectra in TSET1 and recover 90% in the first three positions, performing only slightly worse than MASCOT,[20] despite the fact that 49 out of the 280 "true peptides" are nontryptic in some way (according to the strict tryptic rules we impose in the interpretation; see Supporting Information).

However, in general, the value of the probability score $p^{db}(x)$, eq 12, is not sufficient to discriminate a good prediction from a bad one: the best ranking sequences in the database for two different spectra could get the same score, but this does not imply that both identifications are correct (or wrong). Thus, it is necessary to associate a quality indicator to the prediction: after considering some candidates, we have seen that the z-score, eq 13, shows the best performance. The z-score measures, for each spectrum, how odd the best sequence is in the database in comparison to all the others, according to the distribution induced by the probability profile. Figure 5 shows the ROC curves obtained with *T-novoMS* upon varying the z-score threshold, for TSET1 and ELSET, together with the corresponding ones from MASCOT, upon varying the e-value threshold. *T-novoMS* yields a better curve than MASCOT for ELSET, which is not really surprising, since ELSET is an extension of the learning set. On TSET1, MASCOT performs better, but the curves are quite close, confirming the results of Table 2.

Table S6 in the Supporting Information reveals that a cutoff z_0 = 6.27 yields a FDR < 1%, with a coverage of 45%: almost half the spectra in TSET1 are identified with less than 1% probability of error. However, we cannot assume that the above value of z_0 is valid in general, for unrelated spectra from different data sets, since the probability score distribution of the sequences in the database will depend on the spectrum. Hence, we cannot extend the above findings, on the relationship between FDR and z_0, to predict the FDR of truly unknown

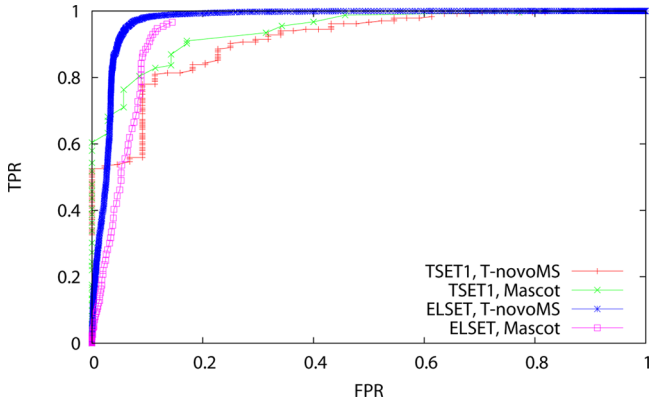


Figure 5. True positive rate (or recall) vs false positive rate for TSET1 and ELSET. The closer the curve to the upper left corner, the better: FPR = 0 and TPR = 1 corresponds to all and only correct identifications.

spectra. In those cases, a valid alternative is to select the threshold z_0 from a comparison with a decoy database, as explained in Methods.

Figure 6 shows the relation between the values of the FDR, obtained upon varying z_0, and the number of spectra that are

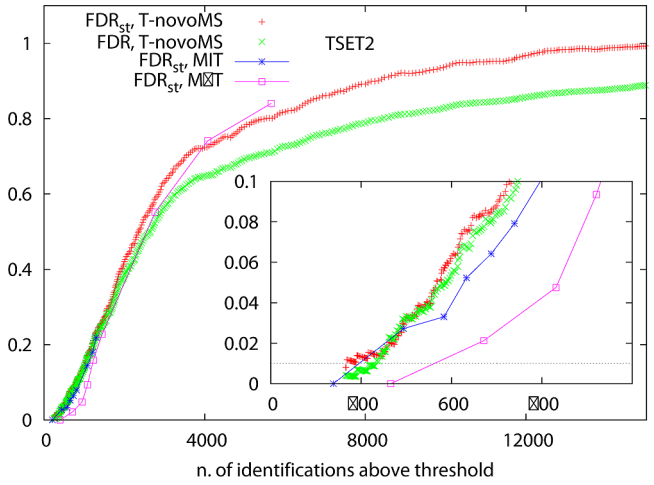


Figure 6. False discovery rate vs number of spectra identified above the threshold, with *T-novoMS* and MASCOT. For *T-novoMS*, FDR refers to the definition given in Methods, and FDR_{st} = ND/NT, where ND and NT are the number of spectra in the decoy and target database with a z-score $z > z_0$; the curves are obtained upon varying z_0. For MASCOT, FDR_{st} = ND/NT, where ND and NT are the number of spectra identified in the decoy and target database with a score exceeding the identity threshold value (MIT) or the homology threshold (MHT). Inset: detail of the most interesting region at FDR < 0.1.

identified with $z > z_0$ in the target database, for TSET2. We plot both the FDR calculated as specified in Methods, according to the definition of ref 14 and that calculated according to the more standard definition, FDR_{st}. Also, Figure 6 reports the corresponding curves obtained with MASCOT, upon varying the significance threshold, which affects the number of identifications above the "identity threshold" and the "homology threshold" in both the target and decoy databases. In this case, we can only calculate FDR_{st}. The inset of the figure gives the detail of the most interesting region FDR < 0.1. We first observe that FDR from ref 14 provides a better curve than

dx.doi.org/10.1021/ac4005666 | *Anal. Chem.* 2013, 85, 4884−4892

FDR_{st}, with smaller values of FDR at a given number of hits. This is also true for TSET1, for which we have found that FDR agrees with FDR′, calculated from the knowledge of the true parent peptide (see Methods and Figure S5 in the Supporting Information). The two curves almost overlap in the most interesting region at FDR < 10%; yet, their difference can still be relevant at FDR < 1%. The inset shows that the curves obtained with *T-novoMS* are very close to that obtained with MASCOT Identity Threshold, which provides a slightly bigger number of hits at a given value of the FDR. The curve obtained with MASCOT homology threshold improves the number of hits, at a given FDR, of a factor of around 2 with respect to *T-novoMS* results.

CONCLUSIONS

Mapping the problem of peptide sequencing onto the study of the equilibrium behavior of a physical system allows the construction of a natural combination of a de novo and database-search tool. The quality of the predictions of *T-novoMS* is similar to that obtained with other de novo methods, and it could probably be improved, by further refinements of the learning database and the energy function, as well as by the introduction of immonium fragments, and of a more realistic description of tryptic (or other enzyme) constraints. The introduction of post-translational modifications is straightforward and just involves the extension of the residue "alphabet" (Table S1 in the Supporting Information) to include them.

A crucial feature of the method is the possibility to associate a probability profile to prefix (and suffix) masses in a natural and precise way; the former can then be used to estimate which fragmentations are more reliably predicted. Moreover, such profiles can be used as an accurate, spectrum-specific score, to fish out the correct sequence from a peptide database.

A test with the same data set used for de novo prediction, as well as a much huger one, reveals that the performance of our method is not far from that of MASCOT, one of the most common database-search tools. It is important to stress that our approach not only allows a prediction of the most likely parent sequence (either de novo or by database search) but also prompts for the design of intrinsic quality assessment of the predictions (even if at a probabilistic level), in the spirit of the "spectral dictionaries".[21] We have found that the "sequence entropy" correlates with the quality of the de novo prediction and can give indications on the latter, even if just in a statistical sense. This is a first example of how the exploration of the "thermodynamic properties" of each spectrum, allowed by the statistical−physics perspective, can provide a spectrum-specific insight on the quality of the interpretation, instead of relying on indicators of the average performance. For the database-search prediction, we have found that the quality of the interpretation increases with the z-score, even if at the moment we cannot establish a general quantitative relation between them from first principles, so that we need a comparison with a decoy database to establish the value of the predictions.

At present, *T-novoMS* does not outperform alternative de novo or database-search software, but we think that this is mainly due to the design of the energy function, rather than to the approach itself. Indeed, at this stage, we were more interested in proving the capabilities of the method, rather than in tuning a score function for the best performance. A different characterization of the phenomenological distributions, a softening of the "tryptic" constraints that now strictly rule out combinations of neighboring residues corresponding to missing enzymatic cleavages, and a precursor-mass dependent choice of the μ parameter, are just a few examples of the optimizations that can be attempted in the future refinements of the energy function. Moreover, this statistical−mechanics approach will be valid not only for such refined functions but also for a whole class of different energy functions, of the form given by eqs 4 and 6. Hence, our proposal establishes the basis of an alternative approach to the interpretation of MS/MS spectra, combining de novo and database-search methods in an unified framework, where important insight can be gained from the well established conceptual tools and techniques of statistical physics.

ASSOCIATED CONTENT

Supporting Information

A more detailed presentation of the theory and additional results (while the present version of computer programs are available from the authors upon request). This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pier@unizar.es.

Present Address

[†]Institute for Scientific Interchange (ISI) Foundation, Via Alassio 11/c, 10126 Torino, Italy.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge support from the MICINN (grant FIS2009-13364-C02-01) and the University of Zaragoza (grant UZ2012-CIE-06) and thank B. Fischer for providing them with the test database of reliable spectra. P.B. thanks R. Zecchina, A. Vindigni, S. Franz, and J. Vazquez for fruitful discussion. M.F.'s contract was supported by DGA program B045/2007. The *T-novoMS* algorithm was developed and tested resorting to the computing facilities provided by the BIFI Institute.

REFERENCES

(1) Marcotte, E. M. *Nat. Biotechnol.* **2007**, *25*, 755−757.

(2) Kim, S.; Gupta, N.; Pevzner, P. A. *J. Proteome Res.* **2008**, *7*, 3354−3363.

(3) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.

(4) Hughes, C.; Ma, B.; Lajoie, G. A. *Methods Mol. Biol.* **2010**, *604*, 105−121.

(5) Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77*, 964−973.

(6) Mo, L.; Dutta, D.; Wan, Y.; Chen, T. *Anal. Chem.* **2007**, *79*, 4870−4878.

(7) Yedidia, J. S.; Freeman, W. T.; Weiss, Y. In *Exploring Artificial Intelligence in the New Millenium*; Lakemeyer, G., Nebel, B., Eds.; Morgan Kaufmann: San Diego, CA, 2003; Chapter 8, pp 239−236.

(8) Notice that this expression is general and not restricted to a discrete mass-array for ν, nor to our discrete physical system: given a proposed peptide P, it is a recipe to calculate $p(P|\Sigma,R)$ as a product, over the ions generated at any fragmentation site, of a suitable weight e^{-H_ν}. However, the identification of the fragmentation sites with the discrete lattice sites is introduced in writing eq 9, so we use the same symbol ν throughout the Article to indicate both.

(9) Fischer, B.; Roth, V.; Roos, F.; Grossmann, J.; Baginsky, S.; Widmayer, P.; Gruissem, W.; Buhmann, J. M. *Anal. Chem.* **2005**, *77*, 7265−7273.

(10) With a little abuse, we can think of $e(P)$ as an energy per residue of P, associated to the probability profile induced by the experimental spectrum, so that z is a measure of how "odd" is the best sequence P_1 in the database among all the others.
(11) Wang, G.; Wu, W. W.; Zhang, Z.; Masilamani, S.; Shen, R.-F. *Anal. Chem.* **2009**, *81*, 146−159.
(12) Mathias, R. A.; Wang, B.; Ji, H.; Kapp, E. A.; Moritz, R. L.; Zhu, H.-J.; Simpson, R. J. *J. Proteome Res.* **2009**, *8*, 2827−2837.
(13) Entry PAe003694 in Peptide Atlas Repository; Instrument: LCQ DECA.
(14) Navarro, P.; Vázquez, J. *J. Proteome Res.* **2009**, *8*, 1792−1796.
(15) Taylor, J.; Johnson, R. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.
(16) Taylor, J.; Johnson, R. *Anal. Chem.* **2001**, *73*, 2594−2604.
(17) They were run with the default parameters: NovoHMM was run with the nongrouping option and PepNovo using the CID_IT_TRYP model.
(18) It is fair to notice that MS-novo, Lutefisk, and PepNovo algorithms do not predict the entire sequence if they are not sure, so that their recall will be lower than precision, in general.
(19) Kim, S.; Bandeira, N.; Pevzner, P. A. *Mol. Cell. Proteomics* **2009**, *8*, 1391−1400.
(20) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(21) Kim, S.; Gupta, N.; Bandeira, N.; Pevzner, P. A. *Mol. Cell. Proteomics* **2009**, *8*, 53−69.