research articles Journal of proteome research

AUDENS: A Tool for Automated Peptide de Novo Sequencing

Jonas Grossmann,[†,||] Franz F. Roos,[‡,||] Mark Cieliebak,[‡] Zsuzsanna Lipták,[§] Lucas K. Mathis,[‡] Matthias Müller,[‡] Wilhelm Gruissem,[†] and Sacha Baginsky*[,†]

Institute of Plant Science, Swiss Federal Institute of Technology, ETHZ, 8092 Zürich, Institute of Theoretical Computer Science, Swiss Federal Institute of Technology, ETHZ, 8092 Zürich, and Universität Bielefeld, Technische Fakultät, AG Genominformatik, 33594 Bielefeld, Germany

Received March 21, 2005

We present AUDENS, a new platform-independent open source tool for automated de novo sequencing of peptides from MS/MS data. We implemented a dynamic programming algorithm and combined it with a flexible preprocessing module which is designed to distinguish between signal and other peaks. By applying a user-defined set of heuristics, AUDENS screens through the spectrum and assigns high relevance values to putative signal peaks. The algorithm constructs a sequence path through the MS/MS spectrum using the peak relevances to score each suggested sequence path, i.e., the corresponding amino acid sequence. At present, we consider AUDENS a prototype that unfolds its biggest potential if used in parallel with other de novo sequencing tools. AUDENS is available open source and can be downloaded with further documentation at http://www.ti.inf.ethz.ch/pw/software/audens/.

Keywords: mass spectrometry • de novo sequencing • spectra preprocessing

1. Introduction

The database-dependent identification of proteins by mass spectrometry is well established and several software packages are available that allow the efficient identification of proteins from tandem mass spectra.[1] SEQUEST[2] and Mascot[3] are widely used software tools for the assignment of peptide sequences to tandem mass spectra. Further statistical evaluation of the assignment reliability may be required to decrease false positive identification rates.[4] Although the database search approach is quite efficient, it has significant shortcomings when database information about the proteins under investigation is limited. This is particularly problematic in cases where no genome information is available. In addition to missing database entries, databases could be erroneous, alternative splicing could occur, and peptides could carry a post-translational modification (PTM). In all of the aforementioned cases, peptide identification by mass spectrometry is more difficult or even impossible using the database-dependent approach.

One possibility to circumvent this limitation is de novo sequencing, i.e., deriving an amino acid sequence exclusively from the information contained in the MS/MS spectrum, without a sequence database. Since manual de novo sequencing is a very time-consuming process, several software tools were developed that extract an amino acid sequence from an MS/MS spectrum. Among the commercially available tools are Sherenga (SpectrumMill), DeNovoX (ThermoFinnigan) and PEAKS.[5] Academic publications include Lutefisk,[6,7] a Hidden Markov Model approach,[8] and PepNovo.[9] The performance of all de novo sequencing software tools inevitably suffers from inherent limitations of MS/MS spectra analysis making reliable automated de novo sequencing difficult. The mass accuracy is a critical issue, and it was shown that a high mass precision allows for other approaches such as composition-based de novo sequencing.[10] Other problems include incomplete b- and y-ion series (gaps) and additional peaks that are independent from the peptide sequence (i.e., chemical noise and instrumental noise). Apart from the b- and y-ion series, which are usually the most prominent signal peaks in ion trap spectra, there are other peptide-derived peaks such as a-ions, neutral losses of water or ammonia, internal fragments and peaks originating from chemical rearrangements.

Although these issues significantly complicate the analysis of MS/MS data, we believe that the information contained in an MS/MS spectrum can be better exploited. In this paper, we describe AUDENS, a tool for automated de novo sequencing which uses heuristic signal peak recognition to improve the sequencing performance.[11] AUDENS utilizes some of the characteristics of peptide fragmentation in CID experiments to identify b- and y-ions while using the additional information that is contributed by other peaks before the de novo sequencing algorithm is applied. In a preprocessing step, these filters are applied to all measured peaks occurring in a spectrum, and a relevance value is assigned to each peak. Peaks are never removed. If their relevance is not increased in any preprocess-

* To whom correspondence should be addressed: Sacha Baginsky, Institute of Plant Sciences, Swiss Federal Institute of Technology, ETH Zentrum, LFW E18, CH-8092 Zürich, Switzerland. Tel: +41 1 632 38 66. Fax: +41 1 632 10 44. E-mail: sacha.baginsky@ipw.biol.ethz.ch.

† Institute of Plant Science, Swiss Federal Institute of Technology.

‡ Institute of Theoretical Computer Science, Swiss Federal Institute of Technology.

§ Universität Bielefeld, Technische Fakultät, AG Genominformatik.

|| These authors contributed equally to this work.

10.1021/pr050070a CCC: $30.25
Published on Web 08/09/2005

ing step, they keep their initial relevance value of 1. After relevances have been assigned to peaks on the basis of the preprocessing filters, the spectrum is subjected to a Weighted-Chen-et-al-Algorithm for noisy data, which is a slight variation of the dynamic programming algorithm for noisy data by Chen et al.[12] The algorithm is described in more detail in Baginsky et al.[11] We systematically analyzed the de novo sequencing performance of AUDENS on different datasets and compared it to the performance of two other freely available tools, Lutefisk and PepNovo.

2. Materials and Methods

Software. AUDENS is a freely available open source tool published under the GNU general public license 2.0 or higher. Since it is implemented in JAVA, it is platform-independent. It features a graphical user interface and a batch sequencing option. AUDENS can be downloaded at the following URL: http://www.ti.inf.ethz.ch/pw/software/audens/.

Acquisition of Datasets. We tuned and tested the sequencing performance of AUDENS on three different datasets. First, we acquired an MS/MS dataset from four protein fractions derived from cauliflower. The crude extract was loaded onto a 1D-SDS gel and bands were cut out. Each band was digested with trypsin and analyzed by LC−MS/MS. The samples were measured on an LCQ ion trap (ThermoFinnigan, San Jose, CA). The dynamic exclusion function was enabled to allow two measurements of the same parent-ion during one minute followed by an exclusion of one minute. Each full scan was followed by four data dependent MS/MS scans. The MS/MS data were submitted to a SEQUEST database search (Thermo Finnigan, San Jose, CA) using a protein database of *Arabidopsis* (which is a sequenced relative of cauliflower) and known contaminants. Cysteines were allowed as either unmodified or carboxyamidomethylated forms and methionines were allowed as either unmodified or oxidized forms. 7488 files were generated from a total of 3921 scans. The SEQUEST results were statistically validated by PeptideProphet.[13] Only doubly charged spectra were used since they are more amenable to de novo sequencing than singly or triply charged ones. Twenty-nine doubly charged spectra were assigned with high confidence, i.e., they exceeded a PeptideProphet confidence value of 0.99. This means that the probability of the peptide database assignment being correct is estimated at more than 99% by the Bayesian model on which PeptideProphet is based and, conversely, the risk of an assignment being a false positive is assumed to be below 1%. A second dataset from four *Arabidopsis* protein fractions was acquired on the same machine and under the same conditions as for the cauliflower dataset. 127 doubly charged spectra exceeded a PeptideProphet confidence value of 0.99. To avoid a potential instrument-specific bias, we also used a previously published dataset generated by Keller et al.[14] This set was obtained upon request from the authors and it consisted of 18 known control proteins which were analyzed in 22 runs on an LCQ ion trap machine. From a total of 37 044 SEQUEST assignments, the authors assembled a list of 2784 positive and manually verified hits, of which 1533 spectra labeled as doubly charged were used for the evaluation of the individual preprocessing filters.

Data Preprocessing. AUDENS accepts peak lists as input, which are either directly generated by the mass spectrometer software or can easily be converted from other formats. Preprocessing of MS/MS data was performed to distinguish between b- and y-ions and other peaks. To this end, we applied a set of preprocessing filters to MS/MS data that were implemented in AUDENS, and we evaluated the ability of each filter to distinguish between b- and y-ions and other peaks. The basic preprocessing filters fall into three categories: (1) Intensity filter (2) Complement filter (3) Offset filter (e.g., presence of characteristic isotope patterns, ammonia loss or water loss from amino acids).

Intensity Filter. The peak intensity filter uses peak height as a criterion to distinguish b- and y-ions from other peaks. Two approaches are used for peak height assessment, (i) peak height as a percentage of the base peak (the highest peak) in the spectrum (intensity filter) and (ii) peak height in a mass window of 50−200 Da (window filter). The intensity filter applies a user defined threshold value and assigns a higher relevance to peaks above the threshold (e.g., 10% of the base peak). The window filter slides a window of 50−200 Da through the spectrum and assigns a higher relevance to the highest 1−5 peaks in the window. The heuristic basis for this filter is the observation that peak intensity in ion trap MS/MS spectra is usually highest in the intermediate m/z range, while both the high and the low mass ranges are usually composed of peaks with lower intensities. The minimal window size was chosen since the smallest amino acid has a mass of 57 Dalton, and thus at most two true peaks can occur within this mass interval.

Complement Filter. The complement filter searches for two peaks A and B for which the sum of both masses equals the originally determined parent mass within a certain mass tolerance ($\text{mass}_{\text{b-ion}} + \text{mass}_{\text{y-ion}} = \text{mass}_{\text{parent ion M+2H}}$). The basis for this filter is the usually high symmetry of MS/MS spectra from doubly charged parent ions. Two peaks that sum up to the parent mass are increased in relevance.

Offset Filter. For neutral losses, the following scheme applies: In the case of ammonia, if there is a peak A and a peak B at an offset of −17 Da from peak A, the relevance score of peak A is increased by a user-defined amount. The same applies to water loss with an offset of −18 Da. This is done because neutral losses are characteristic for true peptide peaks. For isotope patterns, the following scheme applies: If a peak A has a neighbor peak B at an offset of +1 Dalton, then the relevance of A is increased. The setting can be changed so that more than one isotope peak is required to increase the relevance of peak A. Missing isotope peaks (e.g., +0 and +2 present, +1 absent) are not tolerated and peak heights in the isotopic patterns are not taken into account.

Parent Mass Adjustment. The precision of the parent mass measurement is crucial for the reliability of de novo sequencing. The sequencing algorithm uses both b-ions and y-ions to derive a sequence, using the complementarity to fill potential gaps in either of the two series using the equation $\text{mass}_{\text{b-ion}} + \text{mass}_{\text{y-ion}} = \text{mass}_{\text{parent ion } M+2H}$. The necessary tolerance window for ion trap data is usually in the order of ± 1−2 Da. We observed that the measurement errors were not unbiased. The distribution indicates that the measured mass usually falls in a region between the theoretical average mass and the theoretical monoisotopic mass of a peptide (data not shown). Because de novo sequencing is much more straightforward on monoisotopic masses than on average masses, we corrected the parent mass measurement by multiplication with a factor of 0.9993. This factor corresponds to the sum of the monoisotopic masses of all 20 amino acids divided by the sum of their average masses and proved to perform well (data not shown). Different frequencies of amino acids in various organisms were not taken into account; all amino acids were weighted equally.

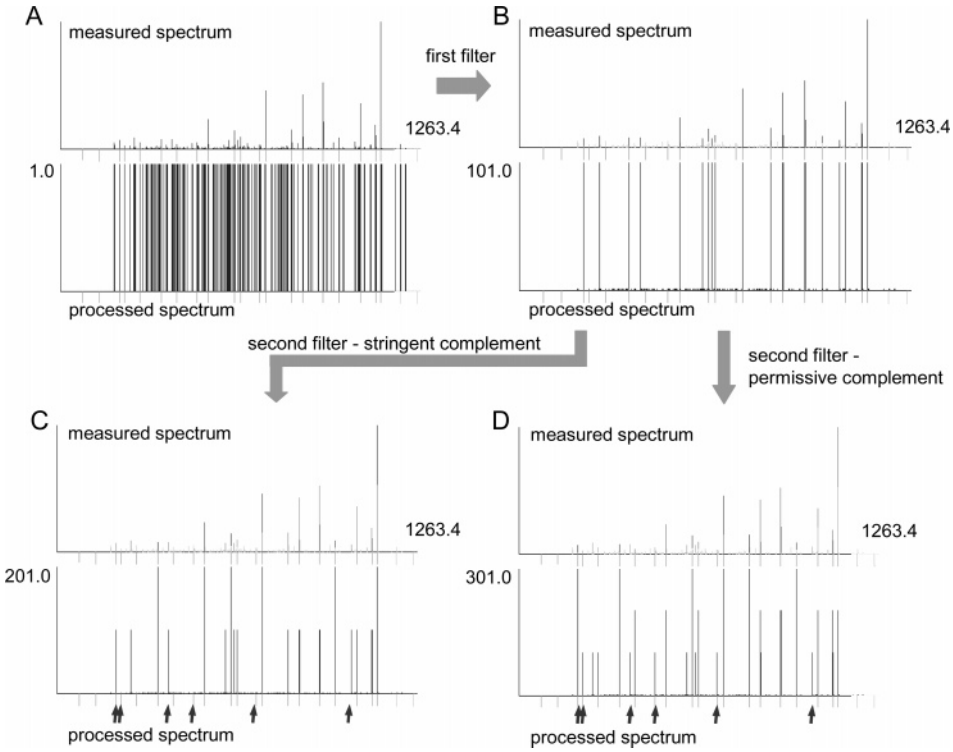

Figure 1. Illustration of spectra preprocessing. In A−D, the measured spectrum is presented in the upper part and the processed spectrum after application of a filter in the lower part. Theoretical b- and y-ions are indicated with small ticks at the *m*/*z* axis. Peaks that differ significantly between the measured and the processed spectrum are indicated with arrowheads. The applied filters were as follows (A) No filter was applied prior to preprocessing. The relevance of each peak is set to one. (B) The percentage filter was applied to increase the relevance of all peaks that are above a threshold of 0.1% of the base peak. An increment of 100 units was added to the initial relevance of 1. (C) Concatenation of the percentage filter with the complement filter in the stringent mode, 100 units are added too. (D) Concatenation of the percentage filter with the complement filter in the permissive mode, 100 units are added as well.

Concatenation of Filters with the Complement Filter. The concatenation of two filters should increase the specificity, i.e., a peak that fulfills several filter criteria (as described above) is more likely to be a true peak. A concatenation does not mean that filters are applied to the spectrum serially but rather that peaks must fulfill two requirements at once if their relevance is to be increased. The complement filter can be combined with other filters since it does not by definition exclude other criteria. The concatenation was used as follows: 1. Permissive concatenation: In the permissive version, only one of the two complementing peaks has to fulfill the additional criterion of the other filter, but both peaks are given a higher relevance. 2. Stringent concatenation: In the stringent version, both complementing peaks have to comply with all applied criteria to gain a higher relevance. The concept of the two concatenation strategies is illustrated in Figure 1.

Evaluation of Filter Performance: Peak Recognition. Peak recognition was defined as follows: From the known amino acid sequence of each peptide, we calculated the theoretical masses of all b- and y-ions. This was done to evaluate the filter performance, whereby the in silico generated spectrum was compared to the measured and preprocessed spectrum. Peaks in the measured spectrum which fell into a tolerance window of 0.5 Da around b- and y-ions in the theoretical spectrum were labeled as signal peaks and the remaining ones as "other peaks". In cases where a signal peak was increased in relevance, the peak was considered "recognized".

Optimization of Filter Relevances. When a preprocessing filter is applied and a certain peak meets a criterion, the relevance value for this peak is increased. The amount of the increase is defined individually for each filter and is called filter relevance in the configuration file. The optimal amounts were unknown at the outset and the goal of the optimization procedure was thus to adjust filter relevances such as to maximize the overall sequencing performance.

We optimized the filter relevance for each filter, i.e., the relative weighting of the filters with respect to each other. A complete screening of all possible parameter permutations would not be computationally feasible for all parameters at once. On the other hand, an isolated optimization is problematic since the optimum of one parameter may strongly depend on other parameters, whose values may still be far from their own optimum. To circumvent those problems, we used a screening design approach first introduced by Plackett and Burman,[15] which is based on specifically designed multiparameter optimization matrices. We used three inequivalent 20 × 20 Hadamard matrices, one of Williamson type, one of first Paley type and one of Tonchev IV type.[16]

In the first step, we set all filter relevances to an arbitrarily chosen starting value of 100. We either decreased or increased parameters in a variety of combinations according to the screening design matrices. We chose binary Hadamard matrices, i.e., in each optimization step two different relevance values were tested per filter. In the case of a positive matrix entry, the corresponding initial value was multiplied by a factor of 1.41 (square root of 2), and for a negative matrix entry, the initial value was divided by 1.41. Each column in the matrix accounts for one configuration setting, while each row corresponds to one filter parameter (filter relevance). The three matrices combined allow for a screening of 20 parameters

within 60 experiments, where each experiment corresponds to one configuration setting applied to the spectrum training dataset. We optimized 14 parameters (filter relevances) using this approach. After each step, we evaluated whether the higher or the lower value of each parameter was more likely to produce better results. The better value was then used as the starting value for the next iteration.

De Novo Sequencing Results and Performance Evaluation. The output of AUDENS is a ranked list of "multi-sequences". A multi-sequence displays ambiguities in the suggestion and actually corresponds to several sequences, i.e., the multi-sequence ACV(N/GG) yields the sequences ACVN and ACVGG if enumerated. AUDENS uses dynamic programming,[12,17] where the score for each sequence is calculated as the sum of the relevances of all peaks, i.e., of the nodes in the spectrum graph which contribute to the path as either b- or y-ion to construct the suggested sequence. If a permutation of amino acids results in the same relevance score, e.g., from gaps that correspond to dipeptides, then the amino acids that fit into the gap are provided in brackets. If there are several possibilities to match a gap, then the tool produces a "multi-sequence". For the evaluation of de novo sequencing performance we used the following strategy: The *Arabidopsis* dataset and the Keller dataset were analyzed with AUDENS to evaluate its performance with and without preprocessing (see above for a description of the datasets). For each spectrum, the tool produces a list of at most 100 sequence suggestions (multi-sequences) which are sorted by their score. We counted for how many spectra the correct sequence was contained in this list and we calculated the median and the average rank at which the correct amino acid sequence occurred for one whole dataset.

3. Results and Discussion

Sequencing Performance without Preprocessing. We evaluated the sequencing performance of AUDENS without and with preprocessing on the *Arabidopsis* and the Keller datasets described above. The de novo sequencing performance of AUDENS (see Materials and Methods) is shown in Table 1A. It is notable that the sequencing performance differs significantly between the two datasets. For the unpreprocessed *Arabidopsis* dataset, AUDENS generated a high ranking correct sequence for 9.4% of all spectra. The correct sequence appeared at a median rank of 2 and an average rank of 4.3 whereas in the unpreprocessed Keller dataset, AUDENS produced a correct sequence for only 4.6% of the MS/MS spectra and the correct sequence appeared at a median rank of 6 and an average rank of 16.6, respectively. However, the quality of the two datasets differed noticeably. Since the cauliflower and the *Arabidopsis* dataset were assembled using a stringent PeptideProphet cutoff of 0.99 confidence value, this set contained more MS/MS spectra showing a clear fragmentation pattern, i.e., higher quality spectra. De novo sequencing depends on high spectrum quality, and it is therefore not surprising that AUDENS performs better on a dataset that contains more high quality spectra.

De novo sequencing is intuitively more difficult than database searching since the search space is much larger. This weakness must usually be compensated by high spectrum quality, e.g., accurate mass measurements and a favorable signal-to-noise ratio. We could clearly illustrate this by applying the algorithm on 1000 simulated theoretical MS/MS spectra. For spectra with precise masses and no noise, the algorithm produced a high-ranking correct sequence for all of them, with a median rank of 1 and an average rank of 1.7. The performance gradually decreased if Gaussian mass variations and noise were introduced which are inherent to real mass spectrometric measurements (Table 1B).

Table 1. De Novo Sequencing Performance

A. increase of de novo sequencing performance due to preprocessing[a]

dataset:	% sequenced	median rank	average rank
Keller	4.6 (17.1)	6 (3)	16.6 (9.2)
Arabidopsis	9.4 (31.5)	2 (2.5)	4.3 (9.2)

B: de novo sequencing performance without preprocessing on 1000 simulated MS/MS spectra[b]

noise	mass precision [Da standard deviation]	% sequenced (within best 100 ranks)	median rank	average rank
none	0.0	100.0	1	1.7
none	0.1	96.2	1	2.8
100 peaks	0.0	99.1	1	3.2
100 peaks	0.1	92.5	1	5.1
200 peaks	0.0	96.6	2	4.3
200 peaks	0.1	88.8	2	6.4

[a] Shows the performance of de novo sequencing without preprocessing and in brackets with an optimized parameter setting for preprocessing. [b] Shows the de novo sequencing performance without preprocessing on simulated, theoretical MS/MS spectra. In the case of perfect spectra with highest mass precision and no noise, the correct sequence was found at a high rank for all spectra. The introduction of noise and Gaussian mass imprecision somewhat decreased the performance. Uniformly distributed noise peaks were added until the spectrum reached a given number of peaks (100 or 200, respectively) The relevance of the noise peaks was 10% of the theoretical b- and y-ions. A sequencer mass tolerance of 0.3 Da was used throughout the simulation.

Evaluation of Preprocessing Parameters. The goal of the spectra preprocessing step is to better exploit the information contained in an MS/MS spectrum and to improve the reliability of de novo sequencing. This is achieved by assigning relevances to each peak that are calculated from different filters which take into account peptide fragmentation chemistry.[18,19] Peaks that comply with several characteristics of b- and y-ions (e.g., mass dependent isotope distribution, complementary peaks, water loss) are more likely to be true b- or y-ions and can thus be distinguished from other peaks. We used the Keller dataset to calculate the specificity and the sensitivity of each filter on the basis of the peak recognition rules to assess suitable parameter settings for the preprocessing (see Materials and Methods). Specificity was defined as the rate of true negative peak recognition among all other peaks than b- and y-ions. This number provides by inference also information on the false positive identification rate, i.e., how many other peaks were wrongly recognized by the filter. We are aware of the fact that this criterion underestimates the filter performance, since some peaks in a spectrum are considered noise although they are derivatives of b-and y-ions and therefore contain sequence information. Examples for these peaks are the a-ions that arise from a carbon monoxide loss from b-ions. For a comparison of filter performance however, the absolute specificity values are not relevant but rather their dependence on the filter parameters. Sensitivity was defined as the rate of true positive

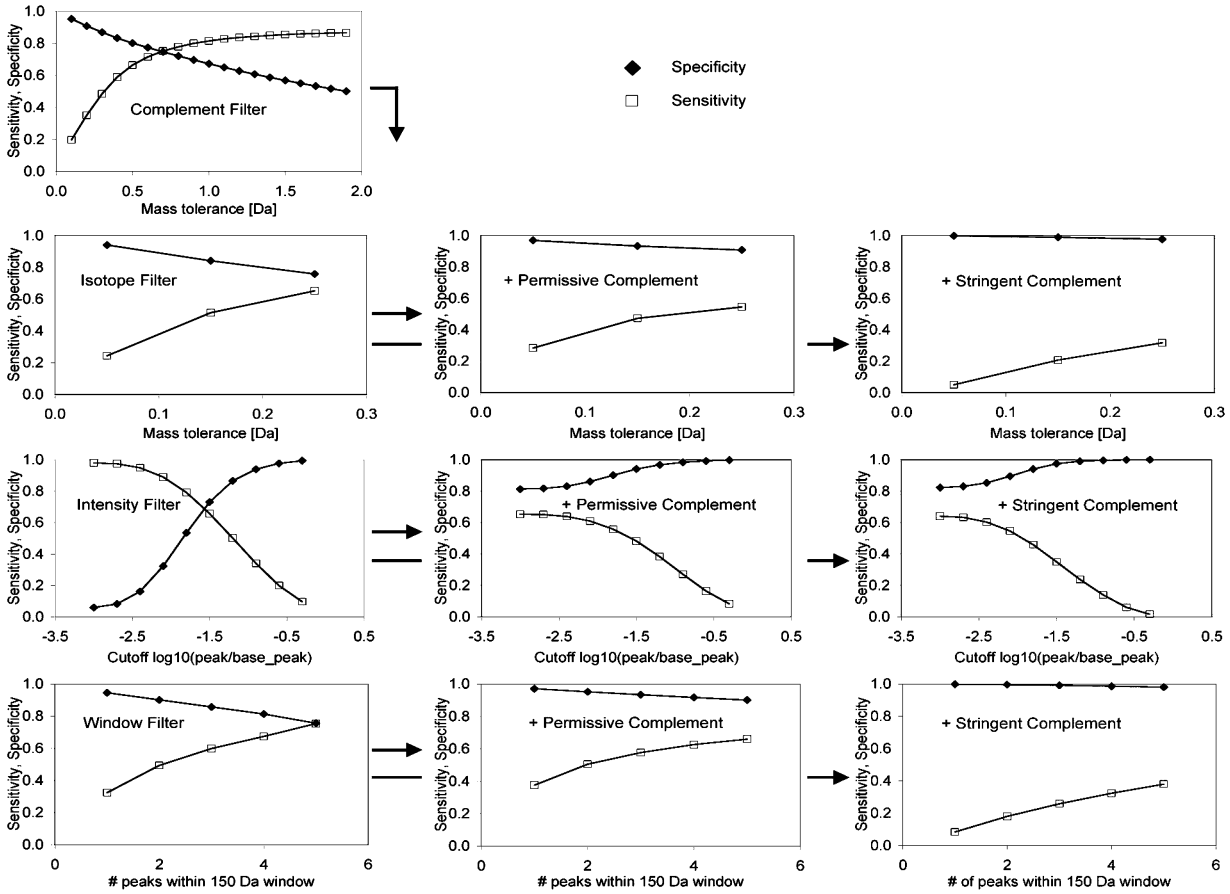

Figure 2. Performance of individual filters: tradeoff between sensitivity and specificity. To compare which filters identify signal peaks best, specificity and sensitivity curves were calculated for each of them (specificity: filled diamonds, sensitivity: empty squares). Each chart shows the result of one filter on the "KELLER dataset". The charts on the left show the performance of the basic filters, and the two charts on the right in each panel show the result of the concatenation of several basic filters. The settings for each filter are indicated on the *x*-axis. For the concatenated complement filters, we used a mass tolerance of 0.9 Da.

peak recognition among all true peaks as defined in the database search results.

The preprocessing filters were applied to all spectra in the dataset using various filter settings (e.g., peak intensity, mass tolerance). Peaks whose relevance was increased by at least one filter were labeled as "recognized" and then compared to the "signal peaks" that came from the comparison of the measured and the in silico generated spectrum. 319 837 measured peaks in 1533 spectra selected from the Keller dataset were taken into account. 37 690 peaks (11.8%) were labeled as signal since they fell in a tolerance window of 0.5 Da around the theoretical b- or y-ions, whereas the other 282 147 peaks (88.2%) were labeled as other peaks. For each peak, the comparison may result in any of the four cases: True positive (labeled as signal peak and subsequently recognized by the filter), false positive (peak other than b- or y-ion whose relevance was falsely increased), true negative (not labeled, not recognized), and false negative (labeled but not recognized) peak recognition. Using these definitions we calculated the sensitivity and specificity values for each filter to investigate their optimal parameter range. The results are shown in Figure 2.

As expected, sensitivity and specificity values for each filter are a tradeoff, i.e., increasing the sensitivity of a filter results in a loss of specificity and vice versa (Figure 2). For each of the basic filters we determined a parameter setting (i.e., mass tolerance for isotope filter, complement filter, and offset filter, percentage of base peak for intensity filter) that was a compromise (usually 80−95% specificity) between sensitivity and specificity. With this setting, we determined the performance of filter concatenation between the basic filters (Figure 1) and the complement filter in stringent and permissive concatenation mode. The combination of the basic filters with the stringent complement filter gave higher specificity and a reduced sensitivity compared to concatenation in permissive mode. We assume that this can be attributed to the fact that in ion trap data, b-ions are usually less abundant than y-ions, which reduces the probability that both ions fulfill each filter criterion. Thus, stringent concatenation may not always be adequate since b- and y-ions usually exhibit independent and different fragmentation characteristics in CID experiments.

Optimization of Preprocessing for de Novo Sequencing. We expected the preprocessing to improve the de novo sequencing performance by better distinguishing between signal and other peaks through application of the filters described above. AUDENS features 40 user-adjustable preprocessing and sequencing parameters in total. For all parameters except the 14 filter relevances, we used the settings that were found to be suitable by the sensitivity/specificity measurements of the filters depicted in Figure 2. To optimize the filter relevances we used a screening design approach based on the one described by Plackett et al. and applied it on a test set of 127 spectra, all of high quality (above 0.99 PeptideProphet confidence). Ideally, de novo sequencing should provide sequences at rank 1. More generally, we need an objective function that evaluates the

Table 2. Comparison of de Novo Sequencing Performance of AUDENS to Lutefisk and PepNovo

Correct Sequence	Total # AA	Audens (optimized via screening design)	# AA	Lutefisk (default parameters)	# AA	PepNovo (default, tryptic)	# AA
SKAEAESLYQSK	12	n[215.2]AEAGASLY[128.1][215.2]c	7	[286.17]EAESLYKSK	9	SQAEAESLYGASK	11
AADVGADLVGK	11	nAADVGADLVG[128.1]c	11	WAVGADLVGK	8	SVAVGADLVGK	8
TDALDAAGNTTAAIGK	16	nGSGA[128.1]DAASSA[128.1]GLA[198.1]c	3	[216.08]ALDAA[174.06]AAEALGK	9	SEALPYNGAAKALGK	6
HGVQELEIELQSQLSK	16	nEY[211.1]GGALENGANGTA[128.1]SSc	0	No Sequence found	0	QLGALELKQFK	1
AGEFFASAHR	10	nDAAETVTAG[258.1]c	0	No Sequence found	0	AGEFFAGTHR	8
NIAVGRPDEATRPDALK	17	n[298.1]SPPCTPV[185.1]G[282.1]c	0	[281.11]DGDRVK[174.06]GP[213.12]	1	QVADQSNQELVYYHK	2
AAVIGDTIGDPLK	13	n[241.2]LGDTLGDP[241.2]c	8	[241.10]DDGLVG[326.12]K	2	GSPLGSELGDPLQ	8
QYQALGGGANTVAHGYTK	18	nCMEALLGADTVAH[202.1]CSc	4	[311.12]RHAVTNKGGLAR[246.07]	0	EYKALGGQNTMGLAGPTK	9
QYQALGGGANTVAPGYTK	18	nFGSESGAVTNPLAF[272.1][128.1]c	1	[401.18]YAVTNAGNWGA[144.05]E	0	YQQALGGGANTVAGSCATK	13
SLGAAIIYNK	10	n[200.1]GAALLYN[128.1]c	8	[257.12]AALLYNK	7	EAGAALLYNK	8
LAADTPLLTGQR	12	nLAADTPLLT[341.1]c	9	No Sequence found	0	LAADTPLLTS	9
LVDIGTVTAQQAK	13	n[230.0]PLGTVTANE[212.1]c	6	No Sequence found	0	PLGTVTANFPP	6
IRLENEIQTYR	11	n[171.1]PPAEL[188.1]AGGLLHc	0	[252.09]WGMKLEDA[234.10]	1	LRLQNELQTYR	10
VYVGQGDSGVVYVK	14	nFDF[194.1]FDGEGVDKc	1	[262.10]F[194.08]FDGKGV[245.08]	0	FDVGQGDFPVYVE	8
TLDEQVDQEEFVR	13	n[201.1]AGE[128.1]VHVV[128.1][213.1]ARc	4	[273.12]FEEGWVK[284.20]R	1	GGEQVWGEENMR	6
QISNLQQSISDAEQR	15	n[241.1][200.1]N[128.1]GASLSDAELGNc	7	[284.15]FADSLSKKNP[171.06]R	1	CPNQQSLSDAERK	8
SLGAAIIFNK	10	n[200.1]GAALLF[242.1]c	6	[271.12]GA[226.09][261.11]K	1	GAALLFD	6
NIEQHASDNVNK	12	n[227.15]E[128.1]HASDNV[242.1]c	8	[227.13][246.07]DSAHK[210.13]K	1	QVEQHASDNVNK	10
IGGIGTVPVGR	11	nLGGLGEAD[294.1]c	5	[170.11]GLGTVPRR	7	LGGLGSLPRR	7
TAENFANYTGDQGYPGGR	18	n[172.1]DLCAVYDSS[128.1]GYVNGPc	3	[243.09][142.08]NTFYLYDAC[210.13]	1	PTFYTGDQGYNPR	8
ALEESNYELEGK	12	n[184.2]EESNYELE[185.2]c	8	[332.08]LEYNYYGK	2	SPEESNYELEGK	10
QRTDALDAAGNTTAALGK	18	n[201.2]AA[158.2]G[128.1]SGAADL[185.2]AL	2	No Sequence found	0	QGNALSEMAADPSLALGK	4
WELLQQVDTSTR	12	nHSFGLAGEVDF[298.3]c	2	[315.11]DLKKVNF[168.09]M	4	NAELLQEVDGTTW	6
AAVDSNSQVGSLFQVLR	17	nSSLRAFLSGV[128.1]S[271.3][268.3]c	0	[287.08]VKFLSGVKSNGVLR	3	SNSQVGSLFQVLR	13
YLEAGVSEHAK	11	nYLEPAPGAENGSc	3	[276.11]FFVGAE[259.17]	0	YNEAGVFFMK	6
HFEVDLGEFR	10	nHFEVNHAEEHc	4	H[184.08]ADLDVE[267.14]	1	HFEVDLDAFR	8
DALAEGDKLTLETAK	15	n[242.2]GSLGD[128.1]LY[128.1]AGHSc	4	WLS[174.06]TLKDGE[196.12]R	1	QGGGAEGDQLTLMVAK	10
NLALGGVQEEVTHPSALR	18	n[227.2]ALMHTVEEAGTGGL[278.3]c	3	[227.13]APFHTVEEKRGP[168.09]K	0	QGLLGGVQQEVTHFN	9
QLYSAAWYPLR	11	n[241.3]YSAAWY[210.3]Rc	7	[242.10][142.08]YSTSN[257.12]R	1	LEYSAAGEYPLR	8
	394		124		61		226

Footnote: The table shows the result of de novo sequencing of 29 example spectra using different de novo sequencing tools such as Lutefisk, AUDENS and PepNovo. With all three tools, only the first ranking sequence was taken into consideration. This is opposed to the Audens optimization step where the objective function included all sequences in the top 100 ranks. The underlined letters were found to be correct, including their position, and their count is indicated for each sequence. AUDENS output: the letters n and c indicate the mass of the additional proton for the n-terminus and the free acid on the c-terminus. Therefore the masses in brackets should be constructed from amino acid residue masses only. The amino acid pairs Isoleucine/Leucine and Lysine/Glutamine (128.1 Da) were considered to be equivalent for the scoring.

performance of the sequencing tool. This function needs to reflect all rank changes, should focus on ranks close to one, and should not require high precision arithmetics to evaluate. One simple way of achieving this is to sum up the inverse of the ranks of the spectra. We called this the IRS (inverse rank sum) and subsequently used it to score the contribution of different relevance settings for successful preprocessing and improvement of de novo sequencing performance.

Once all 127 spectra of the test set had been sequenced with all 60 parameter settings, the performance score IRS was calculated for each setting. We then established the impact of each of the 14 parameters and determined in what direction their value should be adjusted to improve the performance. We compared the IRS between the parameter settings for which a particular parameter had been increased or decreased. Unequal group sizes were taken into account in the calculation. If one group yielded better results than the other, then the parameter value of the better group was used for the next iteration. The performance improved during the first five steps of the iteration whereas in the sixth step, we encountered a slight reduction of the performance. We chose the parameter setting of the fifth iteration for all further evaluations. As a general tendency, the filter relevances of the first filter criteria were reduced with respect to the initial value while the filter relevances of the additional complement criterion were increased. We attribute this to the higher specificity of the concatenated filters. Generally, we found that quite diverse parameter settings yielded similar overall results and that the screening design approach allows us to weed out low performance parameter settings. This is in line with previous experience gathered while manually tuning the tool: out of 25 arbitrary parameter settings, the four best settings together managed to sequence 50% more spectra than the best setting alone. It seems that there is no unique configuration setting that is best for all spectra.

The sequencing performance on the Keller dataset improved from 4.6% of the spectra sequenced at a median rank of 6 to 17.1% of the spectra sequenced at a median rank of 3 (Table 1A). The number of spectra sequenced thus rose by a factor of 3.7. The performance on the *Arabidopsis* dataset improved from 9.4% sequenced at a median rank of 2 to 31.5% sequenced at a median rank of 2.5. The number of spectra that were successfully sequenced thus increased 3.4-fold. In conclusion, the combination of different preprocessing filters is an effective strategy to improve the rate of correctly sequenced spectra and the rank at which the correct sequence occurred. A detailed comparison revealed that different filter combinations might give rise to similar sequencing performance but also that the optimal filter settings for de novo sequencing differ between spectra.

Comparison of AUDENS, Lutefisk and PepNovo. The best performing configuration setting from the screening design was taken as optimal setting for AUDENS. With this setting we sequenced the 29 spectra from the cauliflower dataset which exceed a PeptideProphet threshold of 99% reliability. We purposely chose the cauliflower dataset since it had not previously been used for the tuning. These 29 spectra were also sequenced with Lutefisk and submitted to PepNovo with default parameters. Since the output formats differ, we had to standardize the evaluation. We accepted leucine as isoleucine

(no mass difference) and a mass gap of 128.1 Dalton as lysine or glutamine (mass difference of 0.036 Dalton). Then we checked whether and how many correct amino acids were placed at the correct position in the sequence. AUDENS as well as Lutefisk give a mass gap in brackets if they cannot find an unambiguous sequence fitting in the gap whereas PepNovo produces only one sequence but indicates the confidence for each amino acid. In Table 2 we show and compare the results of the three tools with the correct amino acid sequence as determined by SEQUEST/PeptideProphet.

For the 29 test spectra, we underlined the correctly positioned amino acids and counted them. For AUDENS, we observe that the correct amino acids often correspond to the mass gap but the algorithm cannot determine the order of the amino acids. We did not accept those ambiguous cases even if the correct sequence appeared. We observe that AUDENS as well as Lutefisk have difficulties to derive a full-length amino acid sequence for most of the test spectra. This is consistent with the fact that MS/MS spectra derived from ion traps do not contain the low molecular mass ions. Therefore, the algorithm is unable to derive the correct order for the first two amino acids. This is indicated by the mass gap, which leads to a multi-sequence. Our comparison revealed that the performance of AUDENS (124 amino acids correctly positioned out of 394) is better than the one of Lutefisk (61 amino acids correctly positioned out of 394) but worse than the performance of PepNovo (226 amino acids correctly positioned out of 394). Spectra preprocessing as used for AUDENS does not improve the performance of Lutefisk or PepNovo however, probably because the parameter settings used were tuned specifically for AUDENS (data not shown).

Nevertheless, we found that in one case, AUDENS produced more accurate results than PepNovo (Table 2). This indicates that the sequencing performance of different tools may vary between spectra, and that using different tools in parallel is likely to increase the overall sequence coverage. Additionally, if several tools agree on the same sequence, this increases the reliability of the result. Thus, AUDENS can be used as a complementary tool to extend the scope of de novo sequencing and to increase coverage of and confidence in the de novo sequencing results. To verify sequence suggestions and to put them in a broader context, they may be searched against protein databases using tools such as the recently published DeNovoID.[20] Unlike the BLAST algorithm, it is specifically designed to use the output of the de novo sequencing tools and accepts partly ambiguous sequences and even mass gaps.

4. Concluding Remarks

AUDENS is a freely available software tool for automated de novo sequencing of peptides from MS/MS data. This tool combines an extensive spectra preprocessing part with the implementation of a dynamic programming algorithm. The way AUDENS is presently designed allows the user to individually set preprocessing parameters. Highly flexible and customized settings are necessary to cope with different MS/MS spectra characteristics. Although spectra preprocessing improves the de novo sequencing performance considerably, the difficulty of interpreting conflicting output remains a limiting step in the reliable automated de novo sequencing of high throughput MS/MS data. The error rate is not yet estimated but we are planning to implement a dedicated postprocessing module in AUDENS that helps the user to assess the probability of sequence correctness from the AUDENS output. Currently, the mass list of AUDENS comprises the twenty standard amino acids but it can be extended to include post-translational modifications as well, at the risk of an increased false positive rate though. At present, we consider AUDENS a prototype that unfolds its biggest potential if used in parallel with other de novo sequencing tools. The complementarity between different tools can be exploited to increase coverage and reliability. AUDENS is fast (approximately 1 s per ion trap spectrum on an Intel Pentium M (Centrino) 1.6 GHz processor) and features a batch sequencing option that is very useful for high-throughput purposes and the automatic processing of large numbers of spectra. Furthermore, it is open source, comprises a graphical user interface, and since it is implemented in JAVA, it can be used on all major operating systems without further adaptation.

Acknowledgment. This project was funded by the ETH Grant SEP TH -41/02-2. The authors would like to thank Riko Jacob and Peter Widmayer for invaluable assistance.

References

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.
(2) Yates, J. R.; Eng, J. K.; Clauser, K. R.; Burlingame, A. L. *J. Am. Soc. Mass Spectrom.* **1996**, *7*, 1089−1098.
(3) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(4) Nesvizhskii, A.; Aebersold, R. *Drug Discov. Today* **2004**, *15*, 173−181.
(5) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337−2342.
(6) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067−1075.
(7) Taylor, J. A.; Johnson, R. S. *Anal. Chem* **2001**, *73*, 2594−2604.
(8) Fischer B.; Roth V.; Buhmann J. M.; Grossmann J.; Baginsky S.; Gruissem W.; Roos F. F.; Peter Widmayer; NIPS Proceedings 2004, Vancouver, 457−464
(9) Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77*, 964−973.
(10) Spengler, B. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 703−714.
(11) Baginsky, S.; Cieliebak, M.; Gruissem, W.; Kleffmann, T.; Lipták, Z.; Müller, M.; Penna, P.; Technical Report no. 383, 2002 ETHZ, Zürich. http://www.ti.inf.ethz.ch/pw/publications/papers02/tr383.pdf.
(12) Chen, T.; Kao, M. Y.; Tepel, M.; Rush, J.; Church, G. M. *J. Comput. Biol.* **2001**, *8*, 325−337.
(13) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.
(14) Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. *Omics* **2002**, *6*, 207−212.
(15) Plackett, R. L.; Burman, J. P. *Biometrika* **1946**, *33*, 305−325.
(16) Colbourn, C. J. *The CRC Handbook of Combinatorial Designs*; CRC Press: Boca Raton, 1996.
(17) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; Stein, C. *Introduction to Algorithms*; MIT Press: Cambridge, MA, 2001.
(18) Papayannopoulos, I. A. *Mass Spectrom. Rev.* **1995**, *14*, 49−73.
(19) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R. *Anal. Chem.* **2003**, *75*, 1155−1163.
(20) Halligan, B. D.; Ruotti, V.; Twigger, S. N.; Greene, A. S. *Nucleic Acids Res.* **2005**, Jul 1, 33 (Web server issue): W376-81.

PR050070A