A workflow to increase the deection rate of proteins from unsequenced organisms in high-throughput proteomics experiments

ARTICLE *in* PROTEOMICS · DECEMBER 2007

Impact Factor: 3.81 · DOI: 10.1002/pmic.200700474 · Source: PubMed

7 AUTHORS, INCLUDING:

Jonas Grossmann
ETH Zurich
66 PUBLICATIONS **1,688** CITATIONS
SEE PROFILE

Bernd Fischer
German Cancer Research Center
48 PUBLICATIONS **1,726** CITATIONS
SEE PROFILE

Katja Baerenfaller
ETH Zurich
27 PUBLICATIONS **1,029** CITATIONS
SEE PROFILE

Joachim M Buhmann
ETH Zurich
299 PUBLICATIONS **8,931** CITATIONS
SEE PROFILE

RESEARCH ARTICLE

A workflow to increase the detection rate of proteins from unsequenced organisms in high-throughput proteomics experiments

Jonas Grossmann[1], Bernd Fischer[2]*, Katja Baerenfaller[1], Judith Owiti[1], Joachim M. Buhmann[2], Wilhelm Gruissem[1] and Sacha Baginsky[1]*

[1] Institute of Plant Sciences, ETH Zurich, Zurich, Switzerland
[2] Institute of Computational Science, ETH Zurich, Zurich, Switzerland

Received: May 16, 2007
Revised: August 28, 2007
Accepted: August 29, 2007

We present and evaluate a strategy for the mass spectrometric identification of proteins from organisms for which no genome sequence information is available that incorporates cross-species information from sequenced organisms. The presented method combines spectrum quality scoring, *de novo* sequencing and error tolerant BLAST searches and is designed to decrease input data complexity. Spectral quality scoring reduces the number of investigated mass spectra without a loss of information. Stringent quality-based selection and the combination of different de novo sequencing methods substantially increase the catalog of significant peptide alignments. The *de novo* sequences passing a reliability filter are subsequently submitted to error tolerant BLAST searches and MS-BLAST hits are validated by a sampling technique. With the described workflow, we identified up to 20% more groups of homologous proteins in proteome analyses with organisms whose genome is not sequenced than by state-of-the-art database searches in an *Arabidopsis thaliana* database. We consider the novel data analysis workflow an excellent screening method to identify those proteins that evade detection in proteomics experiments as a result of database constraints.

Keywords:
De novo peptide sequencing / Error tolerant BLAST searches / Plant proteomics / Spectrum quality scoring

1 Introduction

The systematic analysis, identification, and quantification of proteins and their interactions and PTMs are increasingly important fields in life sciences because they provide information about metabolic pathway abundance and protein-network structures. This type of information plays an important role for biotechnology because it enables scientists to make qualitative and quantitative predictions about the impact of a genetic modification on the cellular system. The application of proteomics and related technologies for the analysis of plants is severely hampered by the lack of publicly available sequence information for most economically relevant crop plants, since only a few plant species are sequenced to date including *Arabidopsis thaliana* [1], rice (*Oryza sativa L.*) [2], and poplar (*Populus trichocarpa*) [3]. In order to circumvent this limitation, different strategies and tools were developed to make unsequenced organisms amenable to high-throughput proteomics [4–11]. However, an evaluation of their performance in an integrated proteomics strategy using high-throughput shotgun MS data is currently missing.

Correspondence: Dr. Sacha Baginsky, Institute of Plant Sciences, ETH Zurich, Universitätstrasse 2, CH-8092 Zurich, Switzerland
E-mail: sbaginsky@ethz.ch
Fax: +41-1-632-1079

Abbreviation: PAM, percent accepted mutation

* Both authors contributed equally to this work.

In principle, two different approaches can lead to an increase in protein identifications from unsequenced organisms. In the first approach, MS/MS data are searched against a protein database of an evolutionarily closely related organism. Those spectra that are left unassigned despite originating from a true peptide must be identified and subjected to a more comprehensive protein database search in order to increase the rate of protein identifications. However, as a matter of principle of database-dependent searches, only proteins can be identified that contain at least one peptide with exactly the same sequence as the peptide from a protein in the database. With increasing evolutionary distance this will be an increasingly severe restriction.

In the second approach, the amino acid sequence of a peptide is extracted from the MS/MS spectrum *de novo*, *i.e.*, in a fully database-independent manner using exclusively the information contained in the MS/MS spectrum. Several software tools for peptide *de novo* sequencing are now available and some of them provide sufficiently good results when applied to high-quality spectra [12–18]. Pevtsov *et al.* [19] found in a systematic comparison of different *de novo* peptide sequencing methods that the best performing tools are NovoHMM [16] and PepNovo [15], which are both publicly available. PepNovo uses a probabilistic network whose structure reflects the physicochemical characteristics of peptide fragmentation in CID. NovoHMM is a generative probabilistic model for tandem mass spectra that implicitly separates the prefix and suffix peaks in the spectra.

The approaches modified database search and *de novo* sequencing can also be combined in order to increase the comprehensiveness of protein identifications from unsequenced organisms. One recently described strategy uses a combination of peptide *de novo* sequencing and specialized BLAST searches to identify peptides on the basis of their homology to peptides in the database [4–6]. This strategy involves error tolerant database searches that permit very large search spaces. BLAST searches are tolerant against point mutations and can also cope with PTMs, even if this is not modeled explicitly. In a simulation study, this method demonstrated its potential for cross-species peptide identification provided that the evolutionary distance between the organisms is not too large [6]. To date this strategy was successfully employed for the analysis of low-throughput data, *e.g.*, for the identification of proteins after 2-DE [4], and its performance in a high-throughput context has not been systematically documented yet.

In this paper, we evaluate an alternative workflow that uses a combination of error tolerant BLAST searches and peptide *de novo* sequencing on high-throughput data. Several sequential filtering steps are employed after *de novo* sequencing, which significantly decrease the search space of the BLAST search and its induced complexity. This allows us to lower the acceptance thresholds for the alignments, such leading to a higher protein identification rate. On the basis of our results with three different unsequenced plant species, we discuss the suitability of our proposed workflow to increase the detection rate of proteins in high-throughput proteomics experiments.

2 Materials and methods

2.1 Preparation of plastids and protein extracts from spinach, bell pepper, and Cassava

In order to isolate plastids from spinach and bell pepper, we performed Percoll density gradient centrifugation to separate plastids from other cell organelles and cellular debris as described [20]. In the case of spinach chloroplasts, soluble proteins were released from the organelles by osmotic disruption of the envelope membranes and further fractionated by ion exchange chromatography and ammonium sulfate precipitation, before they were analyzed by LCQ-XP IT MS [20, 21]. The bell pepper chromoplast proteins were extracted from isolated organelles by a serial extraction strategy that uses extraction buffers with increase in solubilization capacity, essentially as described [22–25]. All soluble proteins as well as peripheral and integral membrane proteins were used for the subsequent mass spectrometric analyses. Protein extract of soluble proteins from Cassava was prepared as previously described [26] from three different root samples and one leaf sample.

2.2 Mass spectrometric analysis

Mass spectrometric analyses were performed as described previously [16, 23]. In brief, protein samples were run on a 1-D SDS gel, the gel was cut into 8–12 slices (depending on the complexity of the sample) and the peptide mixtures were generated by an in-gel tryptic digest. Peptide mixtures were further fractionated by RP chromatography on C18-material with a column coupled online to an LCQ XP IT mass spectrometer (LC-ESI-MS). For the spinach and bell pepper data acquisition, during a 2–3 h chromatography, the mobile phase was adjusted to increasing ACN concentrations from 5% (start conditions) to 80% (wash step) to elute peptides from the C18-column on the basis of their hydrophobicity. Eluting peptides were analyzed by one full MS scan and three consecutive MS/MS scans of the three most intense parent ions (37% normalized collision energy). For the Cassava data acquisition, the parameters were slightly different. In brief, during 90 min chromatography, the gradient was developed from 5 to 80% ACN in 70 min, while one full MS scan and four consecutive MS/MS scans of the four most intense parent ions (37% normalized collision energy) were acquired.

2.3 Data analysis and interpretation

Database searches were performed with the SEQUEST software (Thermo Finnigan) that was used to search the *A. thaliana* database (TAIR 6, Dec. 2006, incl. common MS

contaminants keratin and trypsin). The peak lists were created by the SEQUEST software for every MS/MS scan with a total ion count of at least 5×10^4, minimal peak count of 35, and a precursor ion mass in the range of 300–2000 *m/z*. Data were searched against the database restricted to tryptic peptides without modifications except for cysteine carbamidomethyl (Δ average mass: 57.0513) and methionine oxidation (Δ average mass: 15.9994) tolerating a parent mass error of ± 2 Da and a daughter ion error of ± 0.8 Da. All SEQUEST data were further analyzed with the Trans-Proteomic Pipeline (TPP V. 2.23.05), which includes PeptideProphet and ProteinProphet [27, 28]. Peptide and protein identifications with minimum probabilities of 0.9 were regarded as being significant.

For *de novo* sequencing, we used PepNovo or NovoHMM as described. PepNovo was integrated in our pipeline with shell scripts that execute the windows binaries provided for PepNovo. The parameter set was used in default mode for IT data with the tryptic model. For each spectrum, the highest scoring sequence and the corresponding score were parsed and passed to the next step. NovoHMM was applied with default parameters for IT data and the highest scoring sequence and the corresponding posterior probability were parsed and passed to the next step. From a subset of the spinach data, we used those spectra that were confidently identified in a database search for a comparison with PeaksStudio 4.5. We used the recently released version of PeaksStudio 4.5 and combined the Peaks auto *de novo* sequencing option (IT settings, max. 2 missed cleavages, carbamidomethylation at cysteines, max. 2 modifications *per* peptide, one reported sequence *per* spectrum) with a Spider search (same settings as for the *de novo* option) using the same sequence database for all tools (TAIR6 including contaminants) for the standard test and the *in silico* altered *Arabidopsis* database.

3 Results and discussion

3.1 A modified analysis pipeline for proteome analyses of unsequenced organisms

The new strategy we present here integrates a set of methods to increase sensitivity for proteomics data that were generated in a high-throughput setting. It was shown by Shevchenko *et al.* [4] that one can receive reliable protein identifications in low-complexity samples by a combination of *de novo* sequencing with an error tolerant MS-based BLAST search (MS-BLAST). Error tolerant search strategies were so far almost exclusively used for low-complexity samples because of the high false discovery rate in complex samples [6]. We demonstrate here that the modified workflow can yield acceptable performance in a high-throughput context. The basis of the modified workflow is the combination of several filter steps that reduce the input data complexity and, therefore, limit the size of the search space, which lowers the significance threshold for peptide alignment scores in MS-BLAST searches (summarized in Fig. 1). First, a spectrum quality filter is applied to the dataset to identify high-quality spectra that were not identified in the first round database search. Only these are subsequently submitted to *de novo* sequencing (Fig. 1). The *de novo* sequencing results were obtained by employing PepNovo and NovoHMM with the proposed standard setups for IT data and were only accepted if they exceeded a reliability-score threshold. These sequential filtering steps considerably decrease the number of spectra that are subsequently analyzed with an error tolerant MS-BLAST search resulting in higher numbers of significantly identified homologous peptides/proteins.

To evaluate the contribution of each step in the workflow for an increased protein detection rate, we performed a

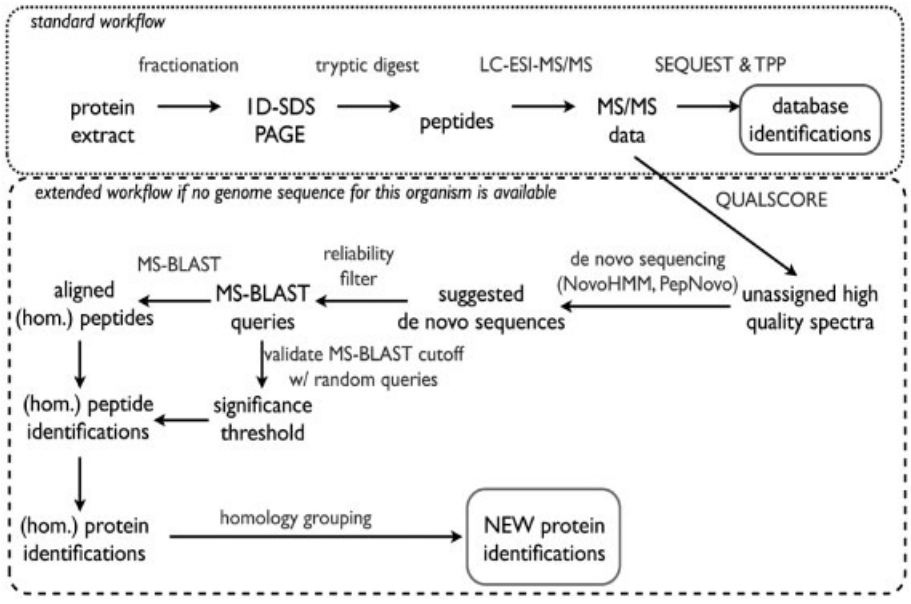

Figure 1. Protein identification workflow. The upper box depicts a standard workflow for protein identification from a complex peptide mixture with database search tools. It comprises a protein fractionation step using 1-D SDS-PAGE followed by in-gel tryptic digest, LC-ESI-MS/MS, and database searches using statistical tools for data validation. The lower box shows the extension of the standard workflow for protein identification in cases where only little sequence information is available for the organism under investigation. In order to select the subset of spectra, which are suitable for the modified analysis pipeline, we employ a number of filtering steps to reduce the complexity of the data.

detailed analysis of the results obtained from 16 LC-MS/MS runs from a spinach chloroplast protein fraction [21]. We compared the proteins, which were identified in our improved workflow with those identified in a normal database search against an *A. thaliana* protein database. Since we are searching for homologous peptide hits in the database, we restrict our comparison to groups of homologous proteins, while considering two proteins as being homologous, if the *p*-value of their BLAST alignment is smaller than 0.01. For the performance evaluation, we divided the identified protein groups into three categories: (i) those that were only identified by MS-BLAST searches, (ii) those that were identified both by MS-BLAST and by standard protein database searches, and (iii) those that were only identified by database searches.

3.2 Assessment of the MS-BLAST performance on highly complex proteomics datasets

The MS-BLAST search is the ultimate step in the modified workflow and the most crucial one. Therefore, we first assessed the threshold values and the reliability of MS-BLAST protein identifications, before evaluating the performance of each of the upstream filters. The previously reported thresholds on MS-BLAST alignment scores [5, 6] are not applicable to our high-throughput scenario, because in our case information about the number of distinct proteins and peptides is missing. This necessitates a scoring strategy that differs from the one proposed previously [6]. Because we are dealing with samples of unknown complexity, we cannot take nonsignificant peptide alignments into account for an additive protein alignment score. We, therefore, decided to consider only those peptide alignments for protein identifications that exceed a certain significance threshold. Such a threshold depends on the complexity of the query and on the size of the database, as random hits are more likely to occur with increasing database size, increasing peptide sequence length, and a higher number of peptide sequences. Since the peptide length and the number of peptides vary highly between experiments, we decided to estimate the threshold individually for each MS-BLAST query.

For this, we sampled 500 query sets of the same size (same number of peptide sequences and same length of single peptide sequences) by employing a first order Markov model, *i.e.*, we considered the probability distribution of the amino acids occurring in the database (*A. thaliana* in our case) and the dependence of an amino acid on its predecessor. The different query sets consisting of 500 sampled queries, each were submitted to MS-BLAST. For each query set, we evaluated the maximum MS-BLAST alignment score over all 500 queries. We determined the MS-BLAST alignment score cutoff such that the maximum alignment score of 5% of the random queries was higher than the cutoff score.

To demonstrate the dependency of the threshold on the database size, we depicted the cumulative probability distribution of the maximum alignment score of a random sample in Fig. 2A using protein databases of different sizes. On the *x*-axis, we have plotted the alignment score threshold and on the *y*-axis we have depicted the percentage of queries that exceeded this threshold (cumulative probability). The alignment score threshold is set where the curve of the corresponding database approaches the 5% level. As can be seen, the threshold for the nonredundant database is much

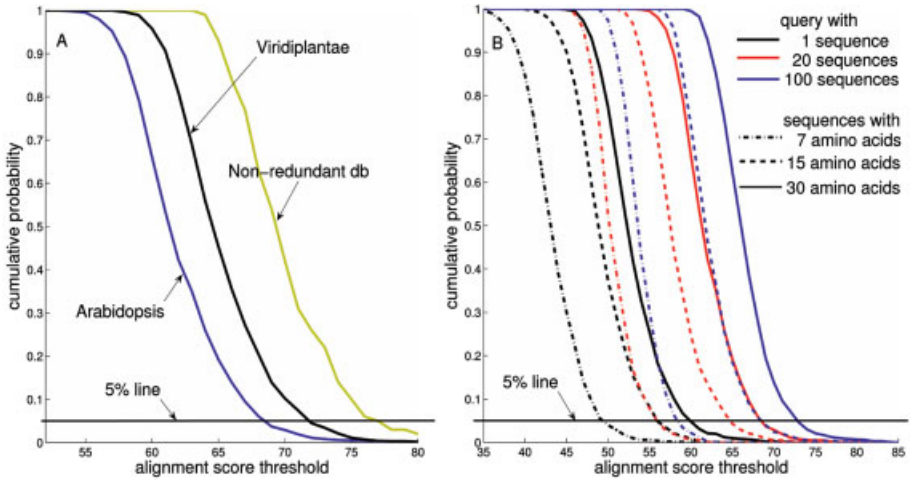

Figure 2. Reliability of peptide alignment depending on database and query size. The plots in A and B show the maximum alignment score dependency on different database sizes (A) and different query sizes assessed on the number of sequences and the number of amino acids *per* sequence (B). Every curve depicts the proportion of random queries that exceed a certain alignment score. The larger the database size (*Arabidopsis*: 29 340 protein entries, 12.6 million symbols; *Viridiplantae*: 382 533 protein entries, 133.2 million symbols; nonredundant db: 800 989 protein entries, 271.4 million symbols) and the larger the query size, the higher is the probability that a significant number of random queries exceed a reliability threshold, such generating false positive peptide alignments.

higher than for the Arabidopsis database. In a second plot (Fig. 2B), we show the dependency of the threshold on the number of peptides and the length of the peptides. One can observe that the threshold has to be increased, if the number or the length of the peptides in the query set increases. The suggested method takes this into account by estimating an appropriate threshold for each MS-BLAST query based on the query size and the size of the underlying database. This procedure permits a stable estimate of the false positive rate of such a high-throughput approach, optimizing the cutoff value for every query individually. The described individual assessment of MS-BLAST thresholds was done for all subsequent analyses and benchmarking tests.

3.3 Performance assessment of different steps in the workflow: QUALSCORE

All the following steps are evaluated with the above described benchmark dataset consisting of 16 LC-MS/MS runs with spinach chloroplast protein samples. The first tool employed in the workflow is QUALSCORE [25], which allows extracting high-quality spectra most likely derived from peptides that were not assigned in a standard database search against the *A. thaliana* protein database. For an analysis with peptides from an unsequenced organism, the number of such unassigned spectra is expected to be relatively high and only these high-quality unassigned spectra are investigated further. This first filter step is able to reduce the number of spectra, which are to be further processed by *de novo* sequencing and MS-BLAST searches by more than 70%. The effect of the first QUALSCORE filter step on the analysis and the anticipated results is depicted in Fig. 3, where we show the sensitivity of the complete workflow to different spectral quality thresholds. Figure 3A depicts the distribution of quality scores retrieved with the spinach dataset. Figure 3B depicts the dependency of significant protein identifications retrieved with the complete workflow on the quality score threshold. By lowering the spectral quality threshold from 5 to 0, the number of protein groups identified by MS-BLAST increases (Fig. 3B, solid black line). For lower thresholds, the number of subsequent protein group identifications with MS-BLAST remains almost constant (Fig. 3B). From this observation, we conclude that we can restrict the analysis to the set of high-quality spectra above a QUALSCORE threshold of zero. At this threshold, about 70% of the low-quality spectra can be discarded without loss of information (Fig. 3A and B).

3.4 Assessment of the *de novo* sequencing approaches and realiability filtering

The next step is automated *de novo* sequencing. Ideally, a *de novo* sequencing method would suggest one unambiguous amino acid sequence *per* spectrum and the summed amino acid masses should be the mass of the precursor ion. As neither the spectra nor the sequencing tools are perfect, these features are often not met. We, therefore, employed two of the best performing *de novo* sequencing tools. With NovoHMM [16], we estimated a full amino acid sequence with b- and y-ions that sum up to the parent mass, whereas PepNovo [15] produces only the most reliable sequence tags. The *de novo* sequencing results were further processed and filtered by a reliability filter. This filter uses posterior probability estimates or average reliabilities of the assigned *de novo* sequences in order to further decrease data complexity (reliability filter). NovoHMM can assign a posterior prob-

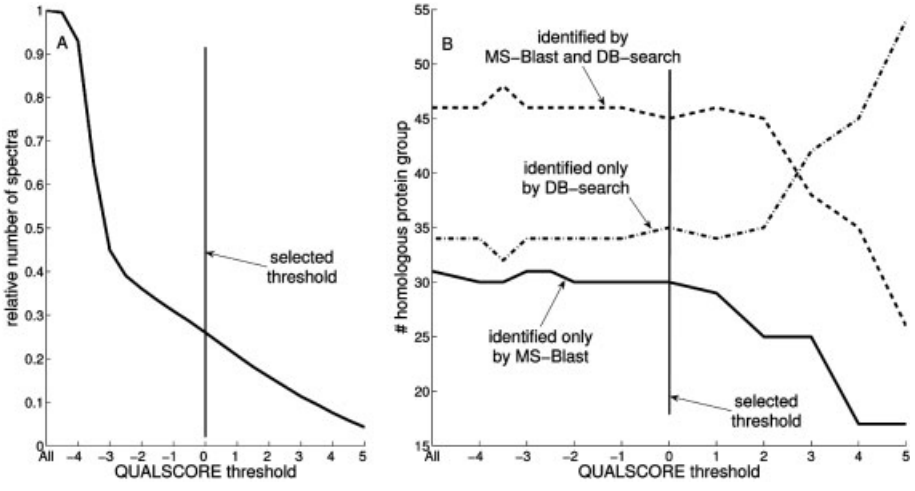

Figure 3. Effect of QUALSCORE stringency on downstream protein identification. (A) The spectrum quality threshold determines how many spectra are used for downstream analysis. (B) The number of identified protein groups as a function of the QUALSCORE threshold. We find that at QUALSCORE thresholds above 0, the rate of significant downstream protein identifications significantly drops (solid and dashed curves). This shows that a significant amount of useful information is discarded if the quality threshold is too stringent (*e.g.*, above 0). For the subsequent experiments, we thus selected a score of 0 that did not seriously affect the rate of significant protein identifications.

ability to every amino acid in a given sequence. *De novo* sequences produced by NovoHMM are accepted if the mean posterior probability of the four amino acids with the largest posterior probability exceeds 0.5. PepNovo was applied with default parameters (tryptic model). The output was parsed and the associated reliabilities provided for each residue were averaged. Based on the identifications from the database, we found that an average reliability cutoff of 0.5 was useful for high-quality spectra (data not shown). In this step, normally about 40% of the total data are discarded.

In Table 1, we show the results of a detailed comparison of different *de novo* sequencing methods (PepNovo and NovoHMM). The number of significant peptide hits after applying reliability filtering is only slightly larger for PepNovo, when both doubly and triply charged peptides are submitted to MS-BLAST, compared to a submission of only doubly charged peptides either to PepNovo or NovoHMM. From this we conclude that the estimation of triply charged peptides with PepNovo is not very reliable and that NovoHMM has almost the same performance as PepNovo (Table 1). Furthermore, the combination of both *de novo* sequencing tools performs best as it achieves the highest number of significant peptide identifications for this dataset. This suggests that PepNovo and NovoHMM analyze the data in a complementary manner. Our data indicate that merging the results of the two *de novo* sequencing methods has a beneficial effect on significant peptide identifications even though the threshold for acceptance increases with the larger query size (65.5 for the combination of both methods, 63.5 for NovoHMM, and 63.7 for PepNovo). The maximum number of identified peptides is 630 for PepNovo, 620 for NovoHMM, and 870 for the combination of both. The results clearly show that the filtering has a beneficial effect, because the MS-BLAST thresholds decrease concomitantly with the query size and the lower thresholds lead to an increase in the number of significant peptide hits. This analysis demonstrates that a combination of the suggested *de novo* sequencing methods together with the reliability filter has a positive effect on the number of identified peptides. The reliability filter is tested for its usefulness by submitting either all *de novo* sequenced peptide strings to MS-BLAST or only those above a reliability threshold (Table 1 – reliability filter (Y/N)) After applying the realibility filter, the number of identified protein groups by MS_BLAST only are 67 for PepNovo, 58 for NovoHMM, and 71 for the combination, resulting in a total of 100 for PepNovo, 95 for NovoHMM, and 115 for the combination.

3.5 Submission to MS-BLAST

The sequences originating from the same LC-MS/MS run that passed the foregoing filter steps are subsequently submitted to MS-BLAST together in one batch query. The idea behind treating each LC-MS/MS run as analytical unit for the alignment is that peptides from one LC-MS/MS run are derived from one specific set of proteins. Peptides from different LC-MS/MS runs on the other hand are assumed to be independent and, therefore, submitted separately. The *de novo* sequencing results having passed all filter steps were made compatible with MS-BLAST searches by adding the amino acid letter B to the N-terminus of the peptide (B in MS-BLAST can be either a lysine or an arginine). Since many *de novo* sequencing methods do not provide guesses for the amino acids at the beginning and the end of the sequence, we filled these gaps with X, XX, XXX, or XXXX (X in MS-BLAST can be any amino acid) according to the assumed gap length. In those cases in which the gap length was difficult to define, the sequences were written in all possible combina-

Table 1. The effect of *de novo* sequencing

No of spectra	*De novo* sequencing			Reliability filter applied (Yes/No)	MS-BLAST		Protein group identifications		
	Tool	Included charge states	No. of peptide strings		Threshold ($p<0.05$)	Peptide hits	MS-BLAST only	MS-BLAST and database	Database only
8153	PepNovo	2 and 3	8153	N	66.5	543	34	60	143
8153	PepNovo	2 and 3	3647	Y	63.7	630	33	67	136
4076	PepNovo	2	4076	N	63.8	625	34	67	136
4076	PepNovo	2	3346	Y	63.4	624	33	67	136
4076	NovoHMM	2	4076	N	63.8	591	41	58	145
4076	NovoHMM	2	3456	Y	63.5	620	37	58	145
8153	PepNovo and NovoHMM	2 and 3, 2	12 229	N	66.6	825	47	67	136
8153	PepNovo and NovoHMM	2 and 3, 2	7103	Y	65.5	870	44	71	132

Results obtained with two different *de novo* sequencing methods (PepNovo and NovoHMM) and a combination of both. With PepNovo, we performed sequencing for doubly and triply charged peptides, whereas for NovoHMM we only sequenced doubly charged spectra. Furthermore, a reliability filter was applied to the *de novo* generated amino acid sequences as described in the text.

tions. All peptide sequences that passed the above *de novo* sequencing filters are submitted to MS-BLAST and the MS-BLAST alignment score thresholds are estimated as described above. A common error source in *de novo* sequencing is the replacement of one amino acid with two other amino acids with the same total mass (N = GG, Q = GA, W = AD, SV, EG) and *vice versa*. Although it may be possible to further improve the alignment results, we refrained from adjusting our data accordingly, in order to keep the search space smaller. Furthermore, one can add the peptides that are already identified by database search in the MS-BLAST search. This can increase the evidence for peptides matching an already identified protein, but it has no influence on peptides that match a new protein. The results of our approach on all datasets are provided in Table 2.

To measure the precision that can be achieved with the modified workflow, we extracted those spectra that were already identified in the database search, and pushed them through our MS-BLAST analysis pipeline. This way, we created a training set that was restricted to known peptide sequences from a selected set of MS/MS spectra. Altogether, 1435 spectra that gave rise to the identification of 120 proteins in the database search were assembled in this training set. Using the workflow described here, the values for our training set are 23.9% recall and 93.7% precision at the spectrum level and 57.7% recall and 85.7% precision at the protein level. The precision is the ratio of accepted identifications that is correct. The recall is the ratio of correct identifications among all possible correct identifications.

The differences between the precision/recall values at the spectrum and at the protein level illustrate the protein inference problem, which results from the loss of connectivity between peptide and protein [29]. Additional filter steps may be required to increase the reliability of protein assignments on the basis of MS-BLAST alignments (see also below for an example). The most reliable way to minimize the protein inference problem issue is to accept only those protein identifications that are based on two or more distinct and significant peptide alignments. Filtering the data by this criterion, we could increase the precision value to 100% at a recall of 26.3% at the protein level suggesting that reliable protein identifications are retrieved when two distinct peptides are required for protein alignment.

3.6 Robustness of the method and comparison with other error tolerant search strategies

The robustness of our method, *i.e.*, the recall that can be achieved depends on the evolutionary distance between the organism under investigation and the organism from which the protein database originated. In order to assess the performance of our method for the detection of proteins from organisms with increasing phylogenetic distance, we generated *in silico* variants of the *A. thaliana* protein database. For every amino acid in the original protein database, we sampled a new amino acid according to the transition probabilities in the percent accepted mutations (PAM) matrices. We sampled different artificial databases by applying a series of PAM matrices (PAM2, PAM5, PAM10, . . ., PAM110) that generate an increasing number of point mutations. Based on our validation we expect a precision of 95%, independently of the recall. In Fig. 4A, we plotted the precision *versus* the recall for the spectra in the training set mentioned above. For all recall values, the (experimentally measured) precision is larger than 0.9.

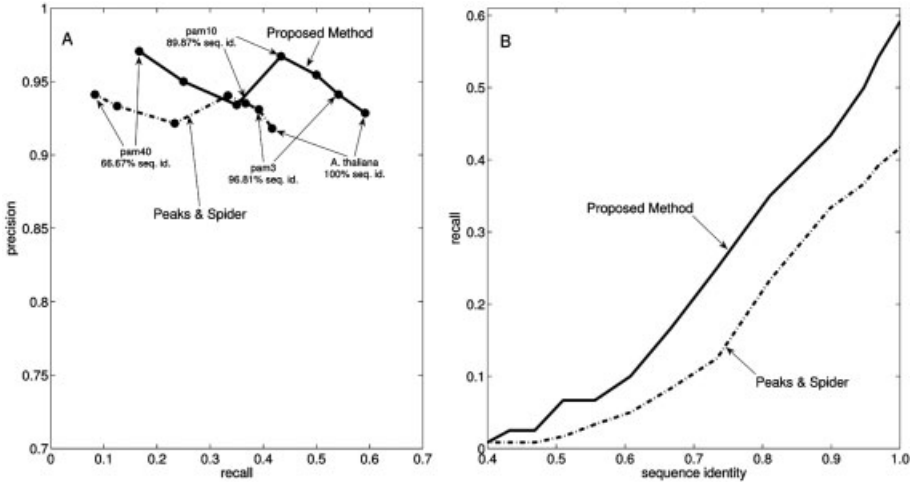

Figure 4. (A) Precision plotted as a function of recall for the proposed method and the Peaks/Spider combination. The precision is similar for the different recall values. (B) The recall as a function of the evolutionary distance. We searched against variations of the *A. thaliana* protein database where the amino acids are mutated (*in silico*) with the PAM matrices. Permutation with the PAM3 matrix yields a protein database with 97% sequence identity to the original *A. thaliana* database, whereas PAM100 yields a protein database with low sequence identity.

The red line in Fig. 4A and B shows a comparison of the workflow proposed here with Spider [7], a commercially available error tolerant analysis tool. We have chosen a cutoff of 45 on the Spider score, which leads to similar precision values above 0.9. Figure 4B reveals that the recall is as expected decreasing with decreasing sequence identity between the organism under investigation and the organism that was used to assemble the protein database. It can be seen that our method outperforms the Spider software on the whole range of evolutionary distances. We tuned Spider such that it produces 20 *de novo* sequences and used the PeptideView results. The presented data were the optimal results that we could retrieve with the Spider software.

3.7 Performance of the modified data analysis workflow in proteomics experiments with three unsequenced plant species

We employed the newly developed and refined workflow also on three large datasets from high-throughput proteomics experiments with protein extracts from spinach chloroplasts, bell pepper chromoplasts, and soluble protein extract from Cassava, three higher plant species for which no complete protein database is available (for extract preparation see Section 2). The standard database search strategy enabled us to identify 5290 peptides from spinach chloroplasts, 4793 peptides from bell pepper chromoplasts, and 1227 peptides from the Cassava leave and root proteome with a confidence score higher than 90% (upper box of Fig. 1). This resulted in 315 identified protein groups in the bell pepper, 204 protein groups in the spinach, and 253 in the Cassava dataset. Most of the protein groups identified with the improved workflow have already been identified in the standard database search.

A number of peptides that gave rise to protein identifications exclusively with the MS-BLAST alignments were found to be enriched for glycine (Tables 1–3 of Supporting Information). Glycine-rich peptides presumably result from keratin, suggesting that their alignments to plant proteins are to be considered false positives. We have, therefore, filtered out peptide alignments based on four consecutive glycines in the sequence (provided in Tables 1–3 of Supporting Information). We filtered out 7, 19, and 3 protein groups from the spinach, bell pepper, and the Cassava datasets, respectively (Fig. 5). After filtering, the MS-BLAST approach yielded 45 new protein groups in bell pepper, 44 in spinach, and 31 in Cassava, which would have been missed with a standard database search (Fig. 5, Table 2).

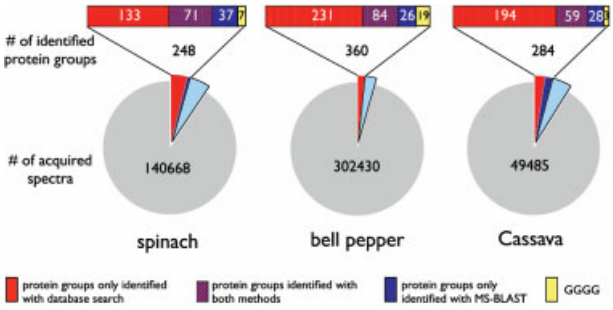

Figure 5. Depiction of protein identification rates in three unsequenced plant species. The gray area in the pie charts represents the fraction of spectra that were left unassigned and were of low quality according to QUALSCORE. The red areas indicate the fractions of spectra that were directly assigned in a database search, and the dark blue sections indicate the fraction of spectra that were identified with the extended analysis pipeline. The light blue pie section represents those spectra that are of high quality but were left unassigned even in the improved workflow. Red parts in the bar diagram represent protein groups that were identified in the standard database search only, protein groups that were confirmed with the extended workflow are depicted in violet, and the protein groups that are exclusively found with the extended analysis pipeline are indicated in dark blue. Because our sampling technique for the assessment of MS-BLAST thresholds does not take overrepresentation of some proteins in a protein fraction into account, strong contamination of the sample with, for example, keratin may cause problems with the alignment of glycine-rich proteins. We have, therefore, applied a glycine filter and excluded those peptides that contain four consecutive glycines from the analysis (yellow piece, labeled "GGGG").

A comparison of the results for the three organisms shows that for Cassava most of the identified proteins were already identified in the standard data analysis workflow (Fig. 5). A potential explanation could be that the evolutionary distance between *Arabidopsis* (the underlying data-

Table 2. Results obtained with the bell pepper, spinach and Cassava datasets

Description		*DeNovo* sequencing	MS-BLAST		Protein group identifications		
Origin of dataset	Spectra in dataset[a)]	No. of peptide strings[b)]	Threshold ($p<0.05$)	Peptide hits	MS-BLAST only	MS-BLAST and database	Database only
Bell pepper	8666	6487	61.6	424	45	84	231
Spinach	8153	7103	65.0	870	44	71	133
Cassava	3234	3100	65.0	783	31	59	194

a) After application of QUALSCORE filtering.
b) After application of the reliability filter.

base organism) and Cassava compared to the other investigated organisms is relatively small. Another possible explanation would be that the Cassava dataset has the highest number of low-quality unassigned spectra, suggesting that this dataset has a lower number of assignable peptide sequences. This could also explain why the additional information obtained with this dataset by applying the modified analysis pipeline was rather small. Nevertheless, it is important to note that for all datasets almost a third of the proteins previously identified in the database search were also confirmed in the MS-BLAST approach, although only those spectra were used for the search that were left unassigned in a standard database search. This finding illustrates that in routine database searches a significant amount of information is currently discarded.

Figure 6 maps proteins that were exclusively identified with the modified workflow and not with the database search onto the RNA and protein synthesis pathway (colored in blue). For most functional categories, the majority of the proteins were already identified in the database search. It is striking, however, that some functional categories are enriched for those proteins that were exclusively identified with the modified workflow (Fig. 6). This is especially true for plastid ribosomal proteins and some proteases, which might be the effect of different evolutionary pressure on proteins with different functions.

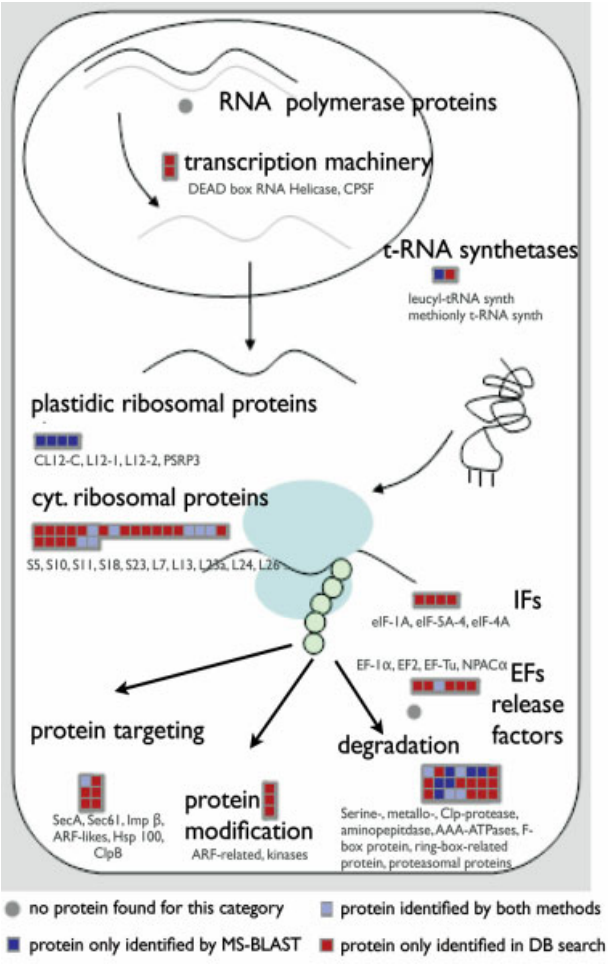

Figure 6. RNA and protein synthesis overview. The figure illustrates cross-species protein identifications using the standard and our improved workflow. Each square box represents one protein homolog from *Arabidopsis* that was identified from Cassava root and leave protein extract in a standard database search (red box), or only after applying the extended workflow (blue box) or the protein was identified in both approaches (light blue box). Gray circles indicate that for this group no homolog could be identified. Note that no plastid ribosomal proteins were identified with the database search approach, whereas four different ribosomal proteins were identified with the improved workflow. Abbreviations are: CPSF, cleavage and polyadenylation specificity factor; NPACα, nascent polypeptide-associated complex alpha; eIF, eukaryotic initiation factor; EF, elongation factor; AAA-ATPase, ATPases associated with diverse cellular activities; ARF-related proteins, ADP ribosylation factor-related proteins. The MapMan tool [30] was used to generate parts of this figure.

4 Concluding remarks

Alternative data analysis tools were developed to support peptide identifications in a database-independent manner. The tools described here, particularly when used in combination, have the potential to increase the detection rate of reliable protein identifications in proteome analyses of organisms, whose genomes are not sequenced yet. We devised a workflow, which increases the number of significant identifications by sequential filtering steps with stringent selection criteria. First, a spectrum quality filter is applied to the spectra. Only high-quality unassigned spectra are then submitted to *de novo* sequencing. The *de novo* sequencing results were only accepted if they exceeded a reliability-score threshold. A combination of the *de novo* sequencing algorithms PepNovo and NovoHMM shows the best performance. These sequential filtering steps considerably decrease the number of spectra that are to be analyzed with an error tolerant MS-BLAST search. As it currently stands, the combination of tools as described here is an excellent screening strategy to assess how many peptide spectra have not been identified because of database constraints and to provide a more comprehensive identification of homologous proteins contained in the protein fraction under investigation. Based on our results we strongly recommend to apply stringent filter criteria before accepting MS-BLAST protein identifications. We used, for example, a postprocessing filter that eliminates all sequences with multiple glycines, because they are likely derived from keratin. Furthermore, the user may consider accepting only those alignments that are based on two or more different peptides, in order to minimize the protein inference problem. Either way, validation should also include biochemical experiments to prove the presence of a particular protein in the sample under investigation.

The author(s) would like to acknowledge funds from the ETH Zurich to W. G. and S. B. as well as funding from the Global Challenges Project (Cassava) to J. O. and W. G. This work was also supported by ETH-grant no. TH-5/04-3 to J. B.

5 References

[1] Initiative, T. A. G., Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2000, *408*, 796–815.

[2] Yu, J., Hu, S., Wang, J., Wong, G. K. *et al.*, A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* 2002, *296*, 79–92.

[3] Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J. *et al.*, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006, *313*, 1596–1604.

[4] Shevchenko, A., de Sousa, M. M., Waridel, P., Bittencourt, S. T. *et al.*, Sequence similarity-based proteomics in insects: Characterization of the larvae venom of the Brazilian moth *Cerodirphia speciosa*. *J. Proteome Res.* 2005, *4*, 862–869.

[5] Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A. *et al.*, Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* 2001, *73*, 1917–1926.

[6] Habermann, B., Oegema, J., Sunyaev, S., Shevchenko, A., The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. *Mol. Cell. Proteomics* 2004, *3*, 238–249.

[7] Han, Y., Ma, B., Zhang, K., SPIDER: Software for protein identification from sequence tags with *de novo* sequencing error. *J. Bioinform. Comput. Biol.* 2005, *3*, 697–716.

[8] Waridel, P., Frank, A., Thomas, H., Surendranath, V. *et al.*, Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated *de novo* sequencing. *Proteomics* 2007, *7*, 2318–2329.

[9] Mackey, A. J., Haystead, T. A., Pearson, W. R., Getting more from less: Algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell. Proteomics* 2002, *1*, 139–147.

[10] Bandeira, N., Tang, H., Bafna, V., Pevzner, P., Shotgun protein sequencing by tandem mass spectra assembly. *Anal. Chem.* 2004, *76*, 7221–7233.

[11] Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D. *et al.*, Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* 2004, *76*, 3556–3568.

[12] Chen, T., Kao, M. Y., Tepel, M., Rush, J. *et al.*, A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* 2001, *8*, 325–337.

[13] Johnson, R. S., Taylor, J. A., Searching sequence databases via *de novo* peptide sequencing by tandem mass spectrometry. *Mol. Biotechnol.* 2002, *22*, 301–315.

[14] Ma, B., Zhang, K., Hendrie, C., Liang, C. *et al.*, PEAKS: Powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 2003, *17*, 2337–2342.

[15] Frank, A., Pevzner, P., PepNovo: *De novo* peptide sequencing via probabilistic network modeling. *Anal. Chem.* 2005, *77*, 964–973.

[16] Fischer, B.,Roth, V.,Roos, F.,Grossmann, J. *et al.*, NovoHMM: A hidden Markov model for *de novo* peptide sequencing. *Anal. Chem.* 2005, *77*, 7265–7273.

[17] Grossmann, J., Roos, F. F., Cieliebak, M., Liptak, Z. *et al.*, AUDENS: A tool for automated peptide *de novo* sequencing. *J. Proteome Res.* 2005, *4*, 1768–1774.

[18] Searle, B. C., Dasari, S., Turner, M., Reddy, A. P. *et al.*, High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS *de novo* sequencing results. *Anal. Chem.* 2004, *76*, 2220–2230.

[19] Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C. *et al.*, Performance evaluation of existing *de novo* sequencing algorithms. *J. Proteome Res.* 2006, *5*, 3018–3028.

[20] Baginsky, S., Gruissem, W., Chloroplast mRNA 3′-end nuclease complex. *Methods Enzymol.* 2001, *342*, 408–419.

[21] Baginsky, S., Grossmann, J., Gruissem, W., Proteome Analysis of Chloroplast mRNA Processing and Degradation. *J. Proteome Res.* 2007, *6*, 809–820.

[22] Baginsky, S., Siddique, A., Gruissem, W., Proteome analysis of tobacco bright yellow-2 (BY-2) cell culture plastids as a model for undifferentiated heterotrophic plastids. *J. Proteome Res.* 2004, *3*, 1128–1137.

[23] Kleffmann, T., Russenberger, D., von Zychlinski, A., Christopher, W. *et al.*, The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. *Curr. Biol.* 2004, *14*, 354–362.

[24] Siddique, M. A., Grossmann, J., Gruissem, W., Baginsky, S., Proteome analysis of bell pepper (*Capsicum annum* L.) chromoplasts. *Plant Cell Physiol.* 2006, *47*, 1663–1673.

[25] Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M. *et al.*, Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: Toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics* 2006, *5*, 652–670.

[26] Cabral, G., Carvalho, L. J., Analysis of proteins associated with storage root formation in cassava using two-dimensional gel electrophoresis. *Rev. Bras. Fisiol. Veg.* 2001, *13*, 41–48.

[27] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 2002, *74*, 5383–5392.

[28] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 4646–4658.

[29] Nesvizhskii, A. I., Aebersold, R., Interpretation of shotgun proteomic data: The protein inference problem. *Mol. Cell. Proteomics* 2005, *4*, 1419–1440.

[30] Thimm, O., Blasing, O., Gibon, Y., Nagel, A. *et al.*, MAPMAN: A user-driven tool to display genomics datasets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 2004, *37*, 914–939.

