Research Article

Journal of Molecular Microbiology and Biotechnology

J Mol Microbiol Biotechnol 2017;27:199–212
DOI: 10.1159/000478907

Received: May 3, 2017
Accepted: June 22, 2017
Published online: August 30, 2017

Influence of NanoLC Column and Gradient Length as well as MS/MS Frequency and Sample Complexity on Shotgun Protein Identification of Marine Bacteria

Lars Wöhlbrand[a] Ralf Rabus[a] Bernd Blasius[b] Christoph Feenders[b]

[a]General and Molecular Microbiology and [b]Mathematical Modeling, Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

Keywords

Shotgun proteomics · Bacteria · Nano-liquid chromatography · Protein identification · Ion trap · Mass spectrometry

Abstract

Protein identification by shotgun proteomics, i.e., nano-liquid chromatography (nanoLC) peptide separation online coupled to electrospray ionization (ESI) mass spectrometry (MS)/MS, is the most widely used gel-free approach in proteome research. While the mass spectrometer accounts for mass accuracy and MS/MS frequency, the nanoLC setup and gradient time influence the number of peptides available for MS analysis, which ultimately determine the number of proteins identifiable. Here, we report on the influence of (i) analytical column length (15, 25, or 50 cm) coupled to (ii) the applied gradient length (120, 240, 360, 480, or 600 min), as well as (iii) MS/MS frequency on peptide/protein identification by shotgun proteomics of (iv) 2 marine bacteria. Longer gradients increased the number of peptides/proteins identified as well as the reproducibility of identification. Furthermore, longer analytical columns strictly enlarge the covered proteome complement. Notably, the proteome complement identified with a short column and applying a long gradient is also covered when using longer columns with shorter gradients. Coverage of the proteome complement further increases with higher MS/MS frequency. Compilation of peptide lists of replicate analyses (same gradient length) improves protein identification, while compilation of analyses with different gradient lengths yields a similar or even higher number of proteins using comparable or even less total analysis time.

Introduction

Mass spectrometry (MS)-based protein identification is nowadays an indispensable tool for proteomic research and life sciences in general. Besides gel-based protein separation methods, like 2D gel electrophoresis, gel-free protein identification by means of nano-high-performance liquid chromatography (HPLC)-driven peptide decomplexation coupled to MS-based detection has become a widely used approach. This method enables identification of >1,000 proteins in a single nano-liquid chromatography (nanoLC)-MS experiment – depending on the organism and type of sample [Zhang et al., 2013]. The most commonly used approach to cover a broad range of cellular proteins, so-called shotgun proteomics, is based on the in-solution proteolytic digest of a whole cell lysate (typically

E-Mail karger@karger.com
www.karger.com/mmb

Lars Wöhlbrand
ICBM, Carl von Ossietzky University of Oldenburg
Carl-von-Ossietzky-Strasse 9–11
DE–26111 Oldenburg (Germany)
E-Mail lars.woehlbrand@uni-oldenburg.de

using trypsin) [Gilmore and Washburn, 2010; Nogueira and Domont, 2014]. The generated peptides are subsequently separated by reversed-phase nanoLC, and the eluent is continuously analyzed by an online-coupled mass spectrometer, applying MS/MS. Commonly, linear gradients of increasing amounts of organic solvent in the eluent are applied with typical gradient times of 0.5–2 h [Nogueira and Domont, 2014; Xie et al., 2012]. The analysis of complex samples requires a high degree of decomplexation, i.e., separation efficiency, to achieve maximal proteome coverage. Although very high proteome coverage may be achieved by two-dimensional LC-MS/MS experiments [Yates et al., 2009], i.e., initial strong cation exchange chromatography followed by reversed-phase nanoLC, this technique requires rather large sample amounts, and its operation is demanding as well as time consuming, for which reason one-dimensional nanoLC-MS/MS is mostly preferred.

In nanoLC-MS experiments in general, the mass spectrometer accounts for mass accuracy and time constraints for successive MS and MS/MS cycles. On the other hand, the nanoLC setup, and especially the applied gradient, determines the number of peptides available for ionization and mass-spectrometric analyses per time interval. Here, the number and diversity of coeluting peptides additionally impact the efficiency of the electrospray ionization (ESI) due to competition for space and charge leading to suppressive effects [for overview see Cech and Enke, 2001]. All these factors directly affect the amount of detectable peptides and, hence, the number of identifiable proteins. While the characteristics of the mass spectrometer are fixed (i.e., instrument inherent) some parameters of the nanoLC can be varied to optimize separation. State-of-the-art nanoLC systems allow for very (ultra)high pressure (>1,000 bar), making use of small inner diameter capillaries and columns with small packing material [MacNair et al., 1997; Xie et al., 2012; Zhang et al., 2013]. Although the resolution power of beads <2 μm in combination with very long analytical columns (e.g., up to 2 m) has been demonstrated [MacNair et al., 1997; Shen et al., 2005; Zhou et al., 2012], the vast majority of shotgun analyses applies columns with 2 to 3 μm beads and short to intermediary column lengths (commonly 15–25 cm) [Xie et al., 2012]. Independent of the column/hardware LC setup, longer gradient times generally allow for higher peak capacities [Guiochon, 2006; Köcher et al., 2011; Liu et al., 2007]. This effect is, however, accompanied by peak broadening, and the long gradient times reduce the sample throughput of the instrument – a crucial factor in daily routine.

Such advanced separation methods and analytical setups were developed with samples from human cell lines, lower eukaryotes (e.g., *Caenorhabditis elegans*), and yeast, all of which comprise a high sample complexity (e.g., ~25,000 and ~7,000 potential gene products in case of human plasma and yeast, respectively). In contrast, bacterial samples are characterized by a lower complexity, since their genomes usually harbor <5,000 protein coding sequences (e.g., 4,288 for *Escherichia coli* K12) [Blattner et al., 1997]. The vast majority of proteomic studies on bacteria applied samples of pathogens, including *E. coli* strains. Environmental bacteria, however, apparently differ from the latter with respect to the number of different proteins formed as well as their abundance, most likely due to the fundamentally different habitats, i.e., nutrient-rich host versus oligotrophic environments [Rabus, 2014].

In the present study, two different, marine, environmental bacteria were investigated: *Phaeobacter inhibens* DSM 17395 and *Desulfobacula toluolica* Tol2. The heterotrophic aerobe *P. inhibens* belongs to the globally distributed and abundant alphaproteobacterial *Roseobacter* group [Buddruhs et al., 2013; Ruiz-Ponte et al., 1998], members of which are conceived to be important players in the mineralization of organic matter in the oceans [for overview see Buchan et al., 2005]. The strictly anaerobic, sulfate-reducing *D. toluolica* is a metabolically versatile member of the deltaproteobacterial family *Desulfobacteraceae* [Rabus et al., 1993] that play a pivotal role in the terminal oxidation (to CO_2) of organic carbon in (anoxic) marine sediments [for overview see Rabus et al., 2015]. In contrast to well-studied "standard" bacteria like *E. coli*, environmental bacteria have to cope with nutrient limitation in their habitats, which is reflected in their overall physiology. This difference becomes evident from pronouncedly slower growth rates, for example: while doubling times of <30 min were observed for *E. coli* in glucose-containing media, *P. inhibens* maximally achieves 6.9 h [Zech et al., 2013] and *D. toluolica* only 27 h (utilizing toluene) [Rabus et al., 1993].

In this study, in-solution digests of whole cell lysates of these 2 environmental bacteria were used to study the influence of (i) analytical column length coupled to (ii) the applied gradient length on peptide/protein identification by shotgun proteomics using (iii) 2 different ion trap mass spectrometers. Samples were subjected to separation using analytical columns of 15, 25, or 50 cm length in combination with linear gradients of 120, 240, 360, 480, or 600 min, respectively. Eluting peptides were analyzed by two 3D ion trap mass spectrometers, differing mainly

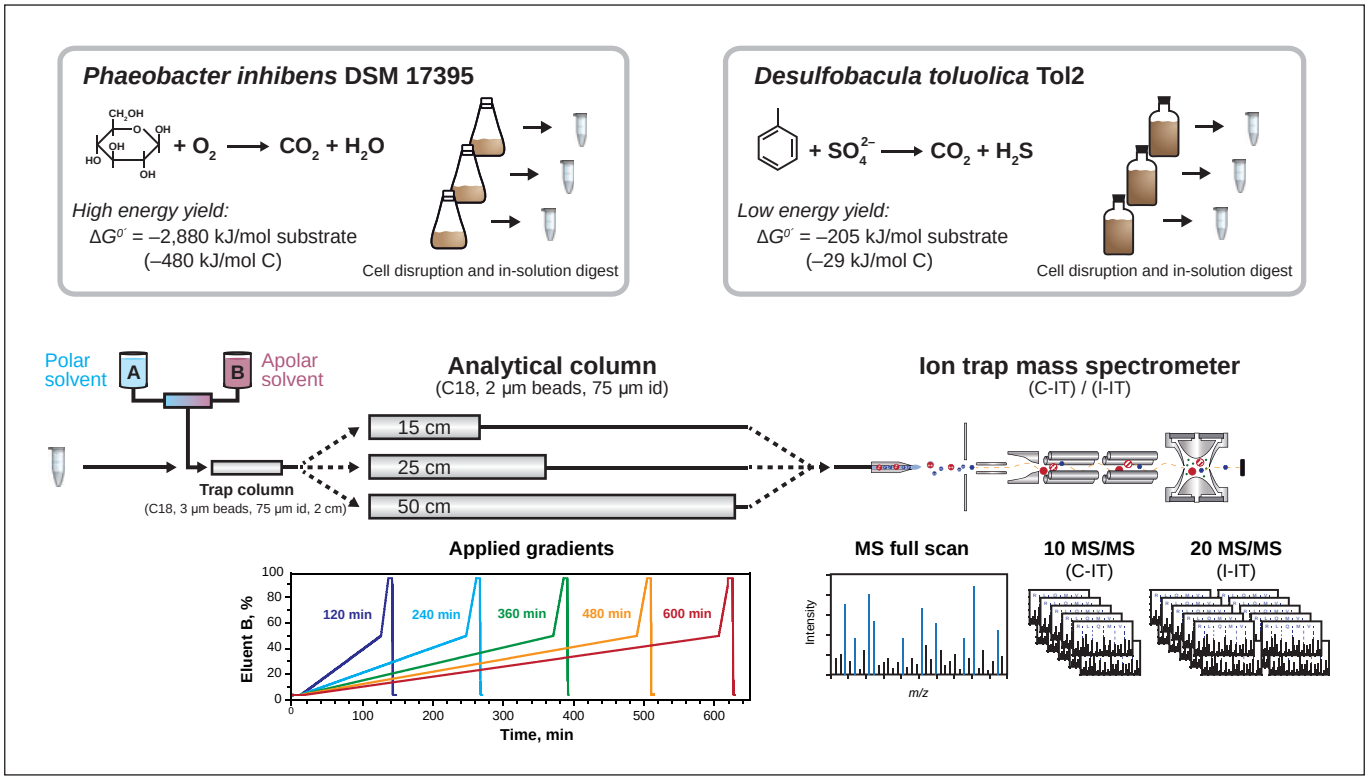


Fig. 1. Schematic representation of the experimental setup. Three biological replicate cultures of 2 marine bacteria, the obligate aerobic *Phaeobacter inhibens* DSM 17395 and the obligate anaerobic (sulfate-reducing) *Desulfobacula toluolica* Tol2, were used for whole-cell shotgun proteomic analyses (reaction equation not stoichiometric). Generated peptide mixtures were separated by nanoLC applying 3 different analytical columns of 15, 25, or 50 cm length, as well as 5 different linear gradients of 120, 240, 360, 480, or 600 min. Peptide masses were analyzed by 2 different, online-coupled ion trap mass spectrometers with acquisition of 10 MS/MS (classic ion trap, C-IT) and 20 MS/MS (improved ion trap, I-IT) per full-scan MS, respectively. Protein identification was performed with ProteinScape on a Mascot server using genome-derived, organism-specific databases. id, inner diameter.

in the MS/MS duty cycle (10 vs. 20 MS/MS per full-scan MS). Three biological replicate samples were analyzed per bacterium and column/gradient setup with both ion traps (yielding a total of 180 shotgun analyses).

Results and Discussion

Experimental Design

An overview of the experimental design of this shotgun proteomic study is provided in Figure 1. Two physiologically different marine bacteria were studied: water-column-inhabiting, aerobic (energy-rich) *P. inhibens* and sediment-inhabiting, anaerobic (energy-limited) *D. toluolica*. To account for biological variation, 3 independent cultures (biological replicates) were generated per organism and subjected to whole-cell, in-solution digest. Aliquots (1 µg) of the same peptide preparation were analyzed by nanoLC-MS/MS applying linear gradients of 120, 240, 360, 480, or 600 min (Table 1) in combination with an analytical column of 15, 25, or 50 cm length. Peptide mixtures were analyzed with 2 different 3D ion trap mass spectrometers. While both instruments feature a similar mass accuracy (±0.15 u), their main difference is the acquisition of 10 MS/MS scans per full-scan MS (classic ion trap, C-IT) in comparison to 20 MS/MS per full scan (improved ion trap, I-IT). The same search criteria were applied to identify proteins based on the respective genome sequence. In the following, a single analysis with a given column, gradient, and MS-setup (e.g., a preparation of *P. inhibens* analyzed with a 15 cm column applying a 240 min gradient and mass detection by C-IT) will be referred to as an analysis set.

Table 1. NanoLC gradient programs for peptide separation

Solvent B[a], %	Gradient time, min					Trap column
	120	240	360	480	600	
4	0	0	0	0	0	load[b]
4	12	12	12	12	12	nanoflow[c]
50	120	255	375	495	615	
95	130	270	390	510	630	
95	135	275	395	515	635	
4	136	276	396	516	636	
4	143	288	402	522	642	load[b]

[a] Composition: 80% (v/v) acetonitrile, 0.1% (v/v) formic acid.
[b] Trap column in loading pump flow (6 µl/min).
[c] Trap column in nanoflow (300 nl/min).

Dataset

In case of *P. inhibens*, application of the different gradient and column combinations yielded (on average) identification of 272–983 (C-IT) and 474–1,352 (I-IT) different proteins per analysis set based on the detection of 815–4,400 (C-IT) and 1,419–7,388 (I-IT) peptides, respectively (Fig. 2). For *D. toluolica* samples, 185–558 (C-IT) and 317–688 (I-IT) different proteins were identified on the basis of 1,074–3,510 (C-IT) and 1,726–4,632 (I-IT) detected peptides, respectively. Including all analysis sets, a total of 1,434 (C-IT) and 1,860 (I-IT) different proteins of *P. inhibens* as well as 831 (C-IT) and 1,092 (I-IT) proteins of *D. toluolica* were detected, which corresponds to a maximum coverage of genome-encoded proteins of 48.0 and 24.9% for *P. inhibens* and *D. toluolica*, respectively (45 analyses per organism and MS setup).

Data Variability

To assess reproducibility of replicate analyses, coefficients of variation (CVs) were determined for peptide retention times considering peptides detected in all 3 replicates per analysis set (413–4,880 peptides). A very low CV with a median of 1.1% (online supplementary Fig. S1; see www.karger.com/doi/10.1159/000478907 for all online suppl. material) was observed for peptide retention times (irrespective of the mass spectrometer used), which demonstrates highly reproducible chromatographic conditions per analysis set. However, few analysis sets apparently comprise higher variation. To some extent, the increased variation may be attributed to subtle hardware changes between replicate analyses (e.g., shortened silica capillaries), causing an offset in peptide retention time for a single replicate. In addition, some column-gradient-sample combinations may give rise to unfavorable coelution of a large number of peptides. This may result in suppressive effects (e.g., charge competition during ionization) that lead to selection for MS/MS at different times.

Overall, compared to C-IT detection, variability applying I-IT detection was slightly higher (median CV 2.0%), most pronounced in case of *D. toluolica* samples. For both instruments, the applied active exclusion settings promote acquisition of identifying spectra of (highly) abundant peptides at their peak maxima. However, these settings may be suboptimal for lowly abundant peptides, rendering off-maximum acquisition more likely. Since the higher MS/MS frequency of the I-IT covers a larger fraction of the coeluting peptides, also a larger share of lowly abundant peptides is covered. Thus, in replicate analyses, identifying off-maximum spectra can occur at different retention times, increasing their respective variation.

Mascot score variation of proteins identified in all analysis sets per MS method (i.e., 135–274 and 228–532 for C-IT and I-IT analyses, respectively) revealed CVs ranging from 8.0–26.0% (C-IT) and 9.0–24.0% (I-IT), respectively (online suppl. Fig. S2, S3). Notably, CVs were always highest with the short 120 min gradient and rapidly decreased with prolonged gradient length per column. With long gradients (480 or 600 min), the CVs of *P. inhibens* samples reached ~12%, while the CVs of *D. toluolica* samples were even lower (9–10%).

Peak Capacity

Peak capacity is a measure for chromatographic performance of the applied gradient and column combination of an LC setup. It is defined as the maximum number of peaks that can be separated in a chromatographic analysis [Neue, 2005] based on the determined peak width. Since the type of mass spectrometer did not influence nanoLC separation performance, only the I-IT data are discussed here (for C-IT data see online suppl. Fig. S4). Overall, chromatographic characteristics were almost similar for both tested bacteria, and the observed differences may be due to the different numbers of peptides present in the analyzed *P. inhibens* and *D. toluolica* samples, respectively (for details see below).

For all column lengths, peak width increased with increasing gradient length (online suppl. Fig. S4A, B). The smallest average peak widths were observed for the 50 cm column for all gradients (minimum <0.38). While the 15 cm column yielded largest peak widths (minimum <0.51),

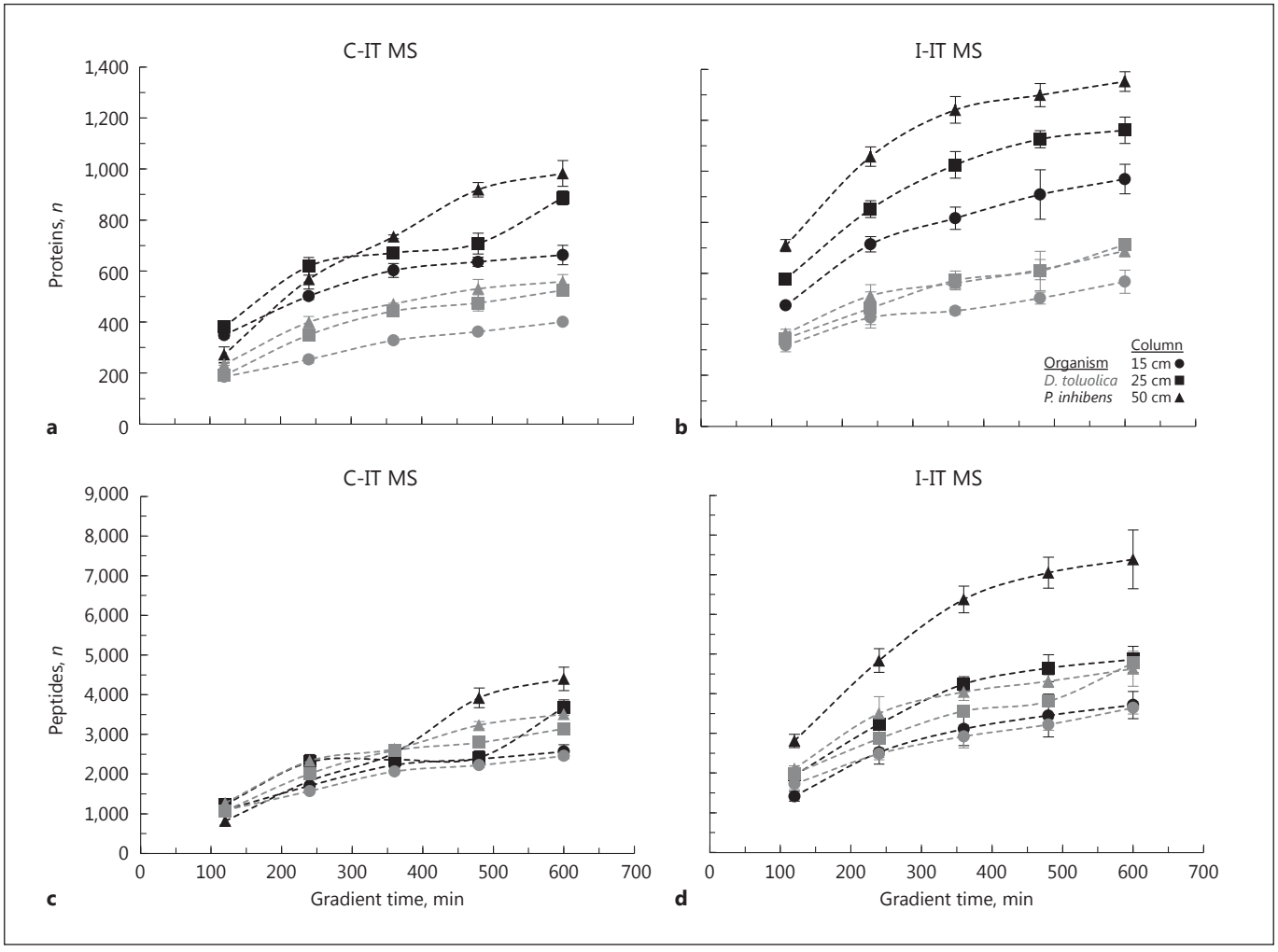


Fig. 2. Number of identified proteins (**a**, **b**) and peptides (**c**, **d**) for *Phaeobacter inhibens* as well as *Desulfobacula toluolica* samples and both MS setups. The different column lengths are indicated. C-IT, classic ion trap; I-IT, improved ion trap.

intermediary ones were observed for the 25 cm column (minimum <0.43). For all columns, peak width increased approximately linearly with gradient time (steeper slopes for shorter column lengths), in agreement with a study of HeLa cell lysates [Köcher et al., 2011].

Corresponding calculated peak capacities were highest for the 50 cm column, ranging from 180 to 590 (Fig. 3a). Lowest peak capacity was determined for the 15 cm column (150–430) and intermediate capacities in case of the 25 cm column (170–500). These results, also with respect to absolute values, are consistent with observations for eukaryotic samples [Hsieh et al., 2013; Köcher et al., 2011]. For all columns, the peak capacity was linearly correlated with gradient time (average $R^2 = 0.985$).

Peak capacity and the number of identified peptides of all analyses also revealed a linear relation (assuming an intercept of zero, $R^2 = 0.800$) (Fig. 3b). For each individual column-organism combination, the linear relation was even stronger (R^2 ranging from 0.919 to 0.993), with the *P. inhibens* analysis sets comprising the 50 cm column and longer gradients (≥360 min) deviating from the respective analysis sets with shorter columns. Overall, the observed relations agree with the reported linear correlation in case of eukaryotic samples [Fairchild et al., 2010; Hsieh et al., 2013; Köcher et al., 2011]. Hence, basic chromatographic parameters of complex peptide mixture separation by nanoLC are independent of sample origin and its complexity, resulting in similar performance for both pro- and eukaryotic samples.

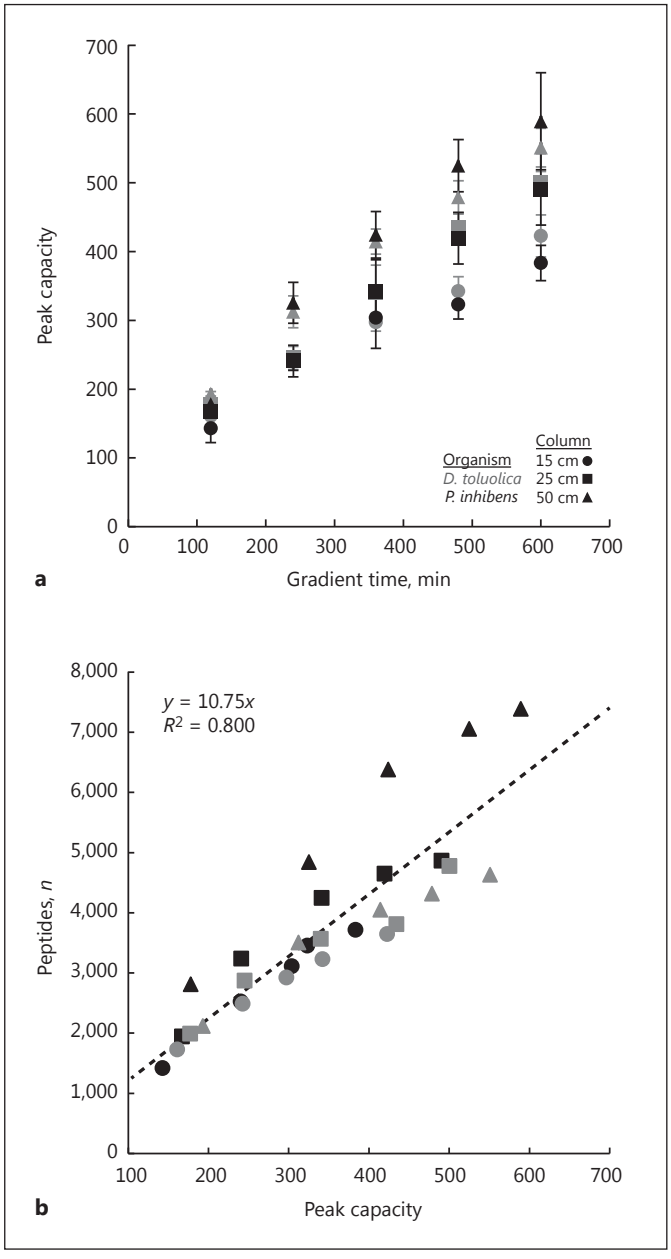


Fig. 3. Peak capacity versus gradient time (**a**) for *Phaeobacter inhibens* and *Desulfobacula toluolica* samples (improved ion trap detection). Linear relation between the number of detected peptides and the peak capacity (**b**) is indicated by a dashed line ($R^2 = 0.800$). The different column lengths are indicated.

Peptide Detection and Protein Identification

In general, both the number of detected proteins and peptides increased with prolonged gradients, though with decreasing slope (Fig. 2). Furthermore, the number of detected peptides and proteins increased with column length for all gradients, resembling observations reported for eukaryotic samples, e.g., *C. elegans* [Hsieh et al., 2013]. The only exception is the observed low number of peptides and proteins detected in case of the *P. inhibens* sample using the 50 cm column with short (120 and 240 min) gradients in combination with the C-IT mass spectrometer (Fig. 2a, c). Here, the analytic capacity of the long column appears to be impaired by the short gradient applied. Since this effect is restricted to the C-IT mass spectrometer, it is most likely due to coelution of a large number of peptides, which can be compensated by the higher MS/MS frequency of the I-IT. Notably, the number of detected peptides was rather similar for all columns and both samples when the 120 min gradient was applied (i.e., ~1,200 C-IT and ~2,000 I-IT) – the only exception being the 50 cm column with the I-IT (2,800) (Fig. 2c, d).

The number of identified proteins was on average 1.7-fold (±0.2, C-IT) and 2.0-fold (±0.2, I-IT) higher for *P. inhibens* than *D. toluolica* samples (Fig. 2a, b). Interestingly, the number of detected peptides was rather similar: 1.0-fold (±0.2) in case of C-IT and 1.3-fold (±0.2) for I-IT detection (Fig. 2c, d). In case of the I-IT, however, 1.5-fold more peptides were detected for *P. inhibens* using the 50 cm column (only 1.2-fold for the 15 and 25 cm column, respectively). Although the numbers of identified proteins pronouncedly differ between the organisms, the difference in the numbers of detected peptides is rather small. Accordingly, the number of identifying peptides per protein is pronouncedly higher in case of *D. toluolica* than *P. inhibens*, irrespective of the mass spectrometer used (Fig. 4). While only 18–35% of all proteins were identified by 5 or more peptides in case of *P. inhibens*, the respective share accounted for 58–65% for *D. toluolica*, indicating that the latter forms fewer different proteins (i.e., lower proteome complexity) than *P. inhibens*. This is in agreement with the disparate thermodynamics of their physiologies. Complete oxidation (to CO_2) of glucose under oxic conditions allows *P. inhibens* to generate –2,880 kJ/mol glucose (i.e., –480 kJ/mol C), whereas complete toluene oxidation coupled to sulfate reduction in *D. toluolica* yields only –205 kJ/mol toluene (i.e., –29 kJ/mol C). Therefore, the energetically limited *D. toluolica* apparently only forms the proteome complement essential under the given conditions, thereby avoiding energy expenditure for proteins not required [Wöhlbrand et al., 2013]. In contrast, *P. inhibens* forms a wide range of different proteins during growth with glucose [Zech et al., 2013]. In consequence, when analyzing similar amounts of total protein (i.e., 1 µg), the lower protein diversity in case of *D. toluolica* samples results in larger amounts of

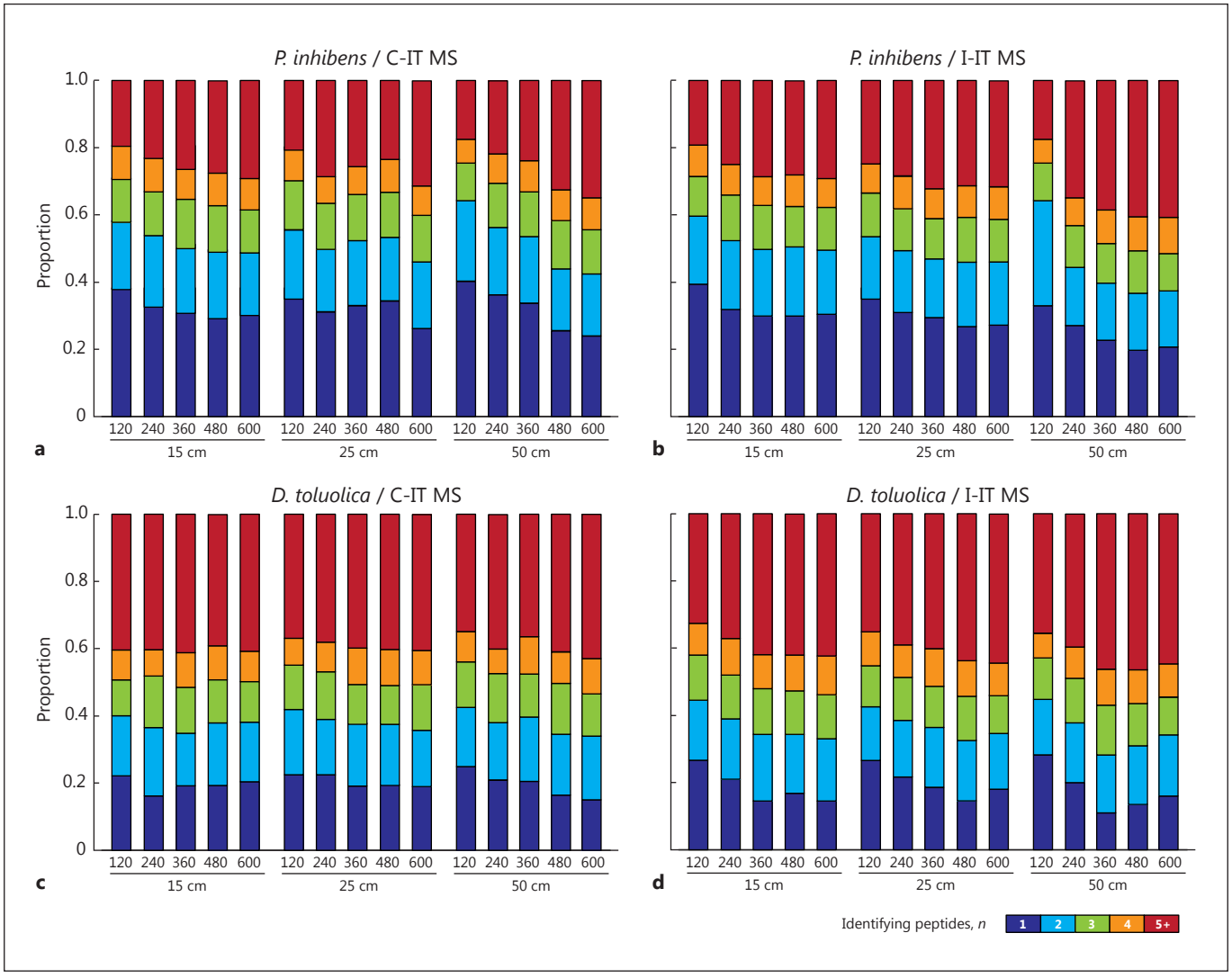


Fig. 4. Proportion of the number of protein-identifying peptides for *Phaeobacter inhibens* (**a**, **b**) and *Desulfobacula toluolica* (**c**, **d**) samples. Gradient times and column lengths are indicated on the horizontal axis. C-IT, classic ion trap; I-IT, improved ion trap.

detectable peptides per protein than for *P. inhibens*, which explains the higher share of 5+ peptide identifications.

The relative distribution of identifying peptides per protein changed with increasing gradient length. Generally, while the share of proteins identified by 2, 3, and 4 peptides remained rather stable, the share of single peptide identifications decreased (e.g., from 40 to 25%, *P. inhibens,* 50 cm column, C-IT) as the 5+ peptide identifications increased (e.g., from 15 to 30%, *P. inhibens,* 50 cm column, C-IT; Fig. 4) with 2 exceptions: (i) *P. inhibens* samples with 25 cm column coupled to C-IT detection (Fig. 4a) (here, the number of single peptide identifications successively increased while 5+ identifications decreased for the 240, 360, and 480 min gradients [note the stable number of detected peptides and proteins for these setups; Fig. 2a]) and (ii) *D. toluolica* samples analyzed with the 15 cm column and C-IT detection, where the relative distribution is similar for all gradients (Fig. 4c). This stability may be attributed to the lower protein diversity of the *D. toluolica* proteome coupled to a higher dynamic range of protein abundance as compared to *P. inhibens*. Apparently, *D. toluolica* forms proteins essential for growth under given conditions in large amounts, while less important proteins remain lowly abundant [Wöhlbrand et al., 2013, 2016]. In

contrast, *P. inhibens* forms a large number of different proteins with rather homogenous abundance [Zech et al., 2013]. In case of *D. toluolica*, the most intense peptide ions automatically selected for MS/MS belong to highly abundant proteins, which are, therefore, identified with a large number of identifying peptides. Elongation of the gradient applied will select more peptides of these abundant proteins, but also single peptides of lower abundant proteins that are newly identified, such that both single and 5+ peptide detection increases likewise. For longer columns, as well as I-IT detection, a larger fraction of the *D. toluolica* proteome is covered, including a large share of lowly abundant proteins (see section Comparison of Analysis Sets). In case of the *P. inhibens* samples, the higher protein diversity and lower abundance range render more proteins detectable though with a lower number of identifying peptides. Due to the more homogeneous abundance, elongated gradients mainly increase the number of identifying peptides per protein and to a lower extend yield new, single peptide identifications.

The identification score of proteins detected in all analysis sets of a distinct column setup mostly increased with prolonged gradient time until reaching a plateau of stable scores (online suppl. Fig. S2, S3), similar to the number of protein-identifying peptides. The share of proteins with increasing score steadily decreased, while that of proteins with unchanged score increased correspondingly (up to 69%). Only a low number of proteins revealed decreased identification scores. Interestingly, in case of *P. inhibens* samples applying the 25 cm column and C-IT detection, a rather high share of proteins (41%) revealed a decreased score when comparing the 240 and 360 min analyses (online suppl. Fig. S2C), while scores are similar for the 360 and 480 min analyses. Hence, it seems likely that detection conditions are unfavorable at this gradient (e.g., due to large amounts of coeluting peptides promoting ionization suppression and a higher requirement for MS/MS than possible). This is in agreement with the rather constant number of detected proteins and peptides of these analysis sets (Fig. 2a, c). Overall, with increasing gradient time, more spectra of detectable peptides per protein are acquired until reaching maximal coverage and, therefore, a stable identification score.

Influence of MS/MS Frequency

To assess the influence of the mass-spectrometric detection on protein and peptide identification, the number of identified proteins was related to the corresponding number of peptides (online suppl. Fig. S5A, B), and the number of proteins or peptides detected by I-IT were compared to those of C-IT detection (online suppl. Fig. S5C–F). Irrespective of the column and gradient used, application of I-IT detection allowed for 50 and 30% more identified proteins in case of *P. inhibens* and *D. toluolica* samples, respectively (online suppl. Fig. S5C, D). This improvement was even more pronounced considering peptide detection, revealing a 70% increase for the more complex *P. inhibens* samples (40% for *D. toluolica* samples, online suppl. Fig. S5E, F), though not fully reflecting the twice-as-high MS/MS frequency. An increased MS/MS frequency does not necessarily increase peptide detections by the same factor, since (i) not all acquired peptide MS/MS spectra have a sufficient quality to be considered for subsequent identification and (ii) not only peptides but also other compounds of the eluent may be selected for MS/MS. Overall, the benefit of a higher MS/MS frequency is larger for samples of high complexity.

Reproducibility of Analysis Sets

Comparison of the triple measurements of each analysis set (biological replicates) revealed that at least 49% of all proteins were detected in all 3 replicates (see diagonal in Fig. 5 and online suppl. Fig. S6). This share pronouncedly increased with longer gradients reaching up to 72% (on average 63% covering all analyses). On average, 20% of all proteins were detected in a single replicate only, with highest share at short gradients (up to 30%) that decreases towards longer gradients (as low as 15%) (data not shown). The observed differences with respect to triple/double detections may be attributed to the statistical (i.e., automatic) selection of potential peptides for MS/MS. Although peptide retention time is very reproducible (see above), such that the eluent at a given time contains a similar peptide mixture, respective ion intensities may vary as a result of slight time shifts of the MS scan due to different time spans of previous MS/MS cycles. In consequence, (partly) different precursor ions are selected for MS/MS based on the chosen criteria. This effect becomes less pronounced if the eluent is less complex, which can be achieved by applying longer gradient separation. Notably, this finding is consistent with a previous study of repeated yeast proteome analysis by LC/LC-MS/MS in which 24% of the detected proteins were identified in a single analysis only [Liu et al., 2004].

Comparing 120 and 600 min gradient triple detections, an improvement of 10 percentage points was observed for all columns in case of I-IT detection (Fig. 5). Applying C-IT detection, the improvement increases with longer columns from 3 to maximally 18 percentage

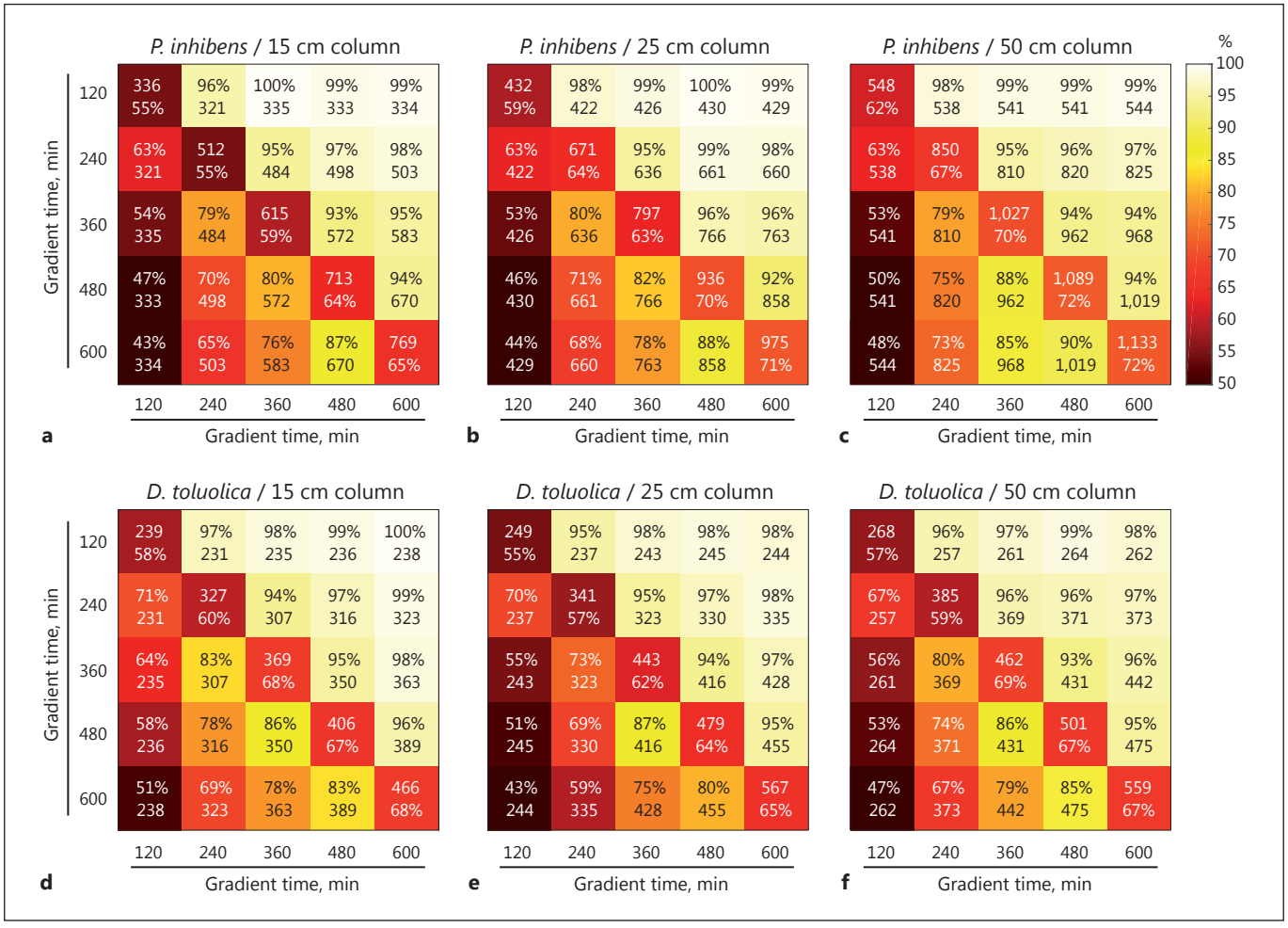


Fig. 5. Similarity matrix of detected proteins per gradient length for each column length for *Phaeobacter inhibens* (**a–c**) and *Desulfobacula toluolica* (**d–f**) samples and improved ion trap detection (for corresponding classic ion trap data see online suppl. Fig. S6). The diagonal contains the number of triple-detected proteins and their share (rounded) of all detected proteins with gradient length. Remaining fields in each row contain the share and number of proteins that are also triple detected with shorter/longer gradient length (comparisons within columns invalid).

points (online suppl. Fig. S6), which may be attributed to the lower MS/MS frequency: the smaller sets of ions selected for MS/MS applying C-IT detection are more likely to contain different ions from the eluting peptides, thereby increasing variation in peptide detection and ultimately protein identification. This effect becomes less pronounced when using longer gradients, since higher proteome coverage is achieved.

Notably, nearly all proteins detected by a short gradient were also detected by longer gradients independent of the mass spectrometer and sample (see rows in Fig. 5 and online suppl. Fig. S6) though the effect is more pronounced with I-IT detection (83–99% C-IT vs. 92–100% I-IT). Hence, with the application of a longer gradient, the set of detected proteins is strictly growing.

Comparison of Analysis Sets

To assess similarities and dissimilarities between all analysis sets, multidimensional scaling (MDS) was applied for visualization (Fig. 6), considering all 1,912 detected proteins for *P. inhibens* and 1,145 for *D. toluolica*, respectively. Close proximity of points in the MDS plot indicates high similarity of corresponding samples, which is evident, for example, for replicates with their high share of triple-/double-detected proteins. Interestingly, the 120 min analyses in combination with C-IT detection cluster

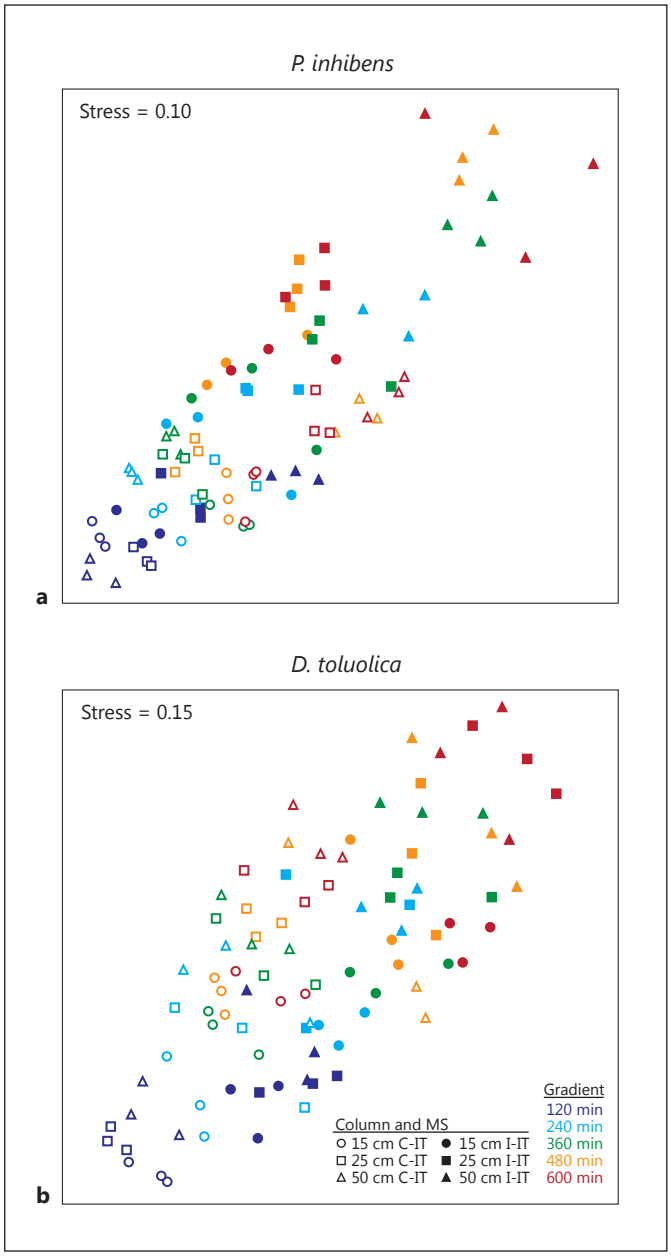


Fig. 6. Multidimensional scaling plot showing analyses of all column/gradient lengths and mass spectrometer combinations for *Phaeobacter inhibens* (**a**) and *Desulfobacula toluolica* (**b**) samples. C-IT, classic ion trap; I-IT, improved ion trap.

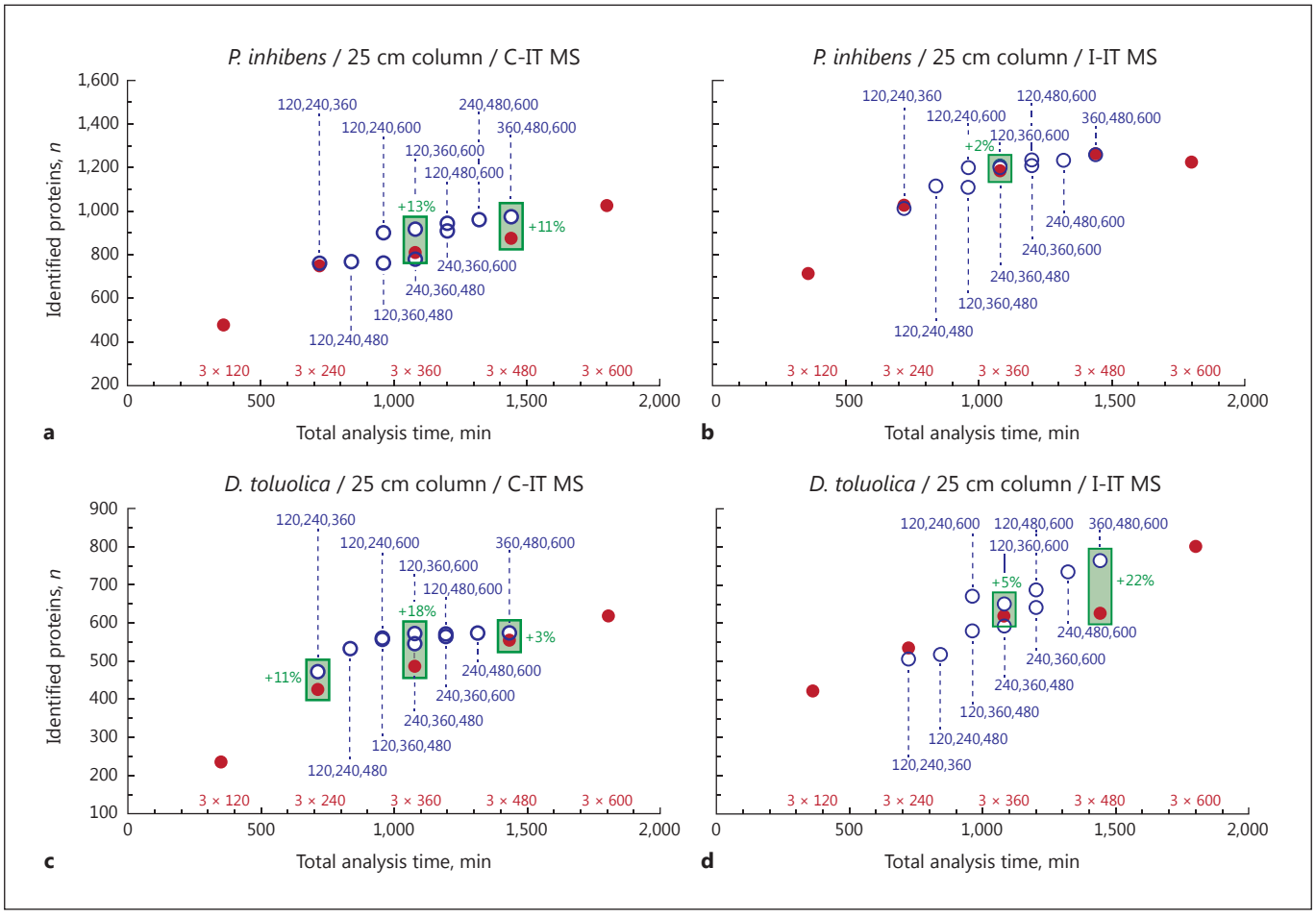
irrespective of column and sample applied. In case of I-IT detection, this pattern is only observed for *D. toluolica*.

For both mass spectrometers, analysis sets of the 15 cm column and long gradients are apparently more similar to analyses applying the 25 cm column with intermediate gradients (e.g., 600 min and 15 cm compared to 240 or 360 min and 25 cm). Similarly, long gradients with the 25 cm column are more similar to 360 or 480 min gradients applying the 50 cm column. Furthermore, short analyses with I-IT detection are more similar to intermediate-length C-IT analyses using the same column, reflecting the higher number of identified proteins applying I-IT detection. Hence, longer LC gradients may compensate for shorter column length as well as lower MS/MS frequency.

Effect of Peptide List Compilation on Protein Identification

Peptide list compilation combines the detected peptides of 2 (or more) analyses to form a new pool of peptides based on which protein identification is performed, thereby improving the latter [Thiele et al., 2010]. In this study, compilations of similar gradient replicates (e.g., 3 × 240 min) yielded higher numbers of identified proteins than single measurements (e.g., 1,024 vs. 851 for 240 min gradient, 25 cm column, I-IT detection, and *P. inhibens* sample). Generally, compilation increased the number of identified proteins by 40–207% (compare Fig. 2 to red circles in Fig. 7 and online suppl. Fig. S7). This increase was most pronounced for compilations of short to intermediate lengths (up to 3 × 480 min) and less for 3 × 600 min. In addition, the benefit of peptide list compilation was larger (i) in case of C-IT detection (83–207%; I-IT detection 66–93%) and (ii) for complex samples (i.e., *P. inhibens*).

In addition to compilation of replicate analysis sets of same gradient length, compilations of different gradient lengths using the same column were performed. In most cases, such mixed gradient time compilations increased the number of identified proteins compared to compilations of replicate measurements (2–24% increase) using the same or even less total analysis time (i.e., sum of all compiled gradient times; e.g., 720 min for 3 × 240 min) (Fig. 7; online suppl. Fig. S7). For example, compilation of triplicate 360 min measurements of *P. inhibens* samples using the 25 cm column and C-IT detection yielded 811 different proteins (total analysis time 1,080 min), while compilation of the 120, 360, and 600 min gradients identified even 919 proteins (13% increase) within the same total analysis time (Fig. 7a). Notably, the latter compilation identified even more proteins than the triplicate 480 min analysis compilation (876 proteins, total analysis time 1,440 min), thus saving 6 h of instrument time. In most cases, the best time/identification ratio was observed for combinations of 120, 360, and 600 or 120, 240, and 600 min gradients. Indeed, these combinations join the peptide sets of very different analysis sets, as evident from the MDS plots (Fig. 6), thereby facilitating more comprehensive protein identification.



Color version available online

Fig. 7. Effect of peptide list compilation with equal gradient length analyses (red circles) and different gradient length analyses (blue rings) applying the 25 cm analytical column and either C-IT (**a**, **c**) or I-IT (**b**, **d**) detection on the number of identified proteins for *Phaeobacter inhibens* (**a**, **b**) and *Desulfobacula toluolica* (**c**, **d**) samples. The gradient time of the compiled analysis sets is indicated. The improvement by mixed gradient time compilations is indicated in green. C-IT, classic ion trap; I-IT, improved ion trap.

Due to the higher coverage of the peptide complement applying I-IT detection, the effect of peptide list compilation is less pronounced. Apparently, the benefit for low complex *D. toluolica* samples seems to be higher (up to 24% increase) than for the more complex *P. inhibens* samples (up to 13%). This effect may be due to the detection of single peptides of low-abundance proteins, which by itself are not sufficient for protein identification. Hence, peptide compilation may merge 2 (or more) such peptides in the combined peptide list, thereby elevating the resulting protein identification score above threshold. Therefore, compilation takes more effect in case of *D. toluolica* samples due to the rather large share of proteins with low abundance.

Conclusions

Ideally, all peptides of an analyzed sample should be subject to MS/MS fragmentation to achieve complete coverage of the proteome, but limitations of the mass spectrometer as well as constraints on acquisition time lead to incomplete coverage in practice. Here we demonstrated that (i) the protein diversity of the sample and the protein abundance range influence the number of detectable peptides and proteins, hence, achievable proteome coverage strongly depends on the organism studied and its biological state, e.g., physiological condition, (ii) the chromatographic conditions applied impact on reproducibility of protein identification, this directly affects gel- and label-free quantitative experiments, by enhanc-

ing or impairing the extent and reliability of protein quantification, and (iii) compilation of mixed gradient length analyses facilitates comprehensive identification while saving instrument time. Most likely, the described effects also affect higher resolution mass spectrometers (e.g., Q-TOF or Orbitrap instruments) and may even be more pronounced.

Overall, while MS/MS frequency cannot be changed (since it is inherent to the MS instrument), a smart decision on the nanoLC gradient and column setup may allow for a time-efficient acquisition to achieve the proteome coverage required to answer the scientific questions asked.

Materials and Methods

Bacterial Samples

P. inhibens DSM 17395 was cultivated in minimal mineral medium with glucose (11 mM) as sole source of carbon and energy as described before [Zech et al., 2013]. *D. toluolica* Tol2 was grown in minimal medium supplemented with toluene (2% [v/v] in an inert heptamethylnonane carrier phase) as sole source of carbon and energy under sulfate-reducing conditions, as described before [Rabus et al., 1993]. For each bacterium, 6 parallel cultures (i.e., biological replicates) were performed, 3 of which were subjected to cell harvest, while the others were further incubated to demonstrate "normal" growth. Cells of *P. inhibens* were harvested during early active growth, i.e., 15 h after inoculation at an optical density of 600 nm of 0.15, and *D. toluolica* during linear growth at half maximal optical density (OD_{660} 0.2) as described before [Champion et al., 1999]. Obtained cell pellets were immediately frozen in liquid nitrogen and stored at –80°C until further processing.

In-Solution Digest

For shotgun proteomics, a cell pellet of ~100 mg wet weight was resuspended in 200 µL of lysis buffer (7 M urea, 2 M thiourea, and 30 mM Tris/HCl, pH 8.5), and cells were disrupted by means of the grinding kit (GE Healthcare, Munich, Germany). After incubation for 20 min at 20°C and 1,000 rpm, cell debris was precipitated by centrifugation (10 min, 20,000 *g*, 4°C). Subsequently, the protein content of the soluble fraction was determined according to the method described by Bradford [1976]. A total of 50 µg of protein per sample (1 µg/µL in urea buffer; 0.4 M NH_4HCO_3, 8.0 M urea) was subjected to reduction (55 mM DTT, 30 min at 56°C) and alkylation (100 mM iodoacetamide, 30 min in the dark). The solution was diluted with gradient-grade HPLC water to achieve a final concentration of 2 M urea prior to tryptic digest (1 µg per sample; Trypsin Gold; Promega, Mannheim, Germany) at 37°C for 16 h. Aliquots were frozen in liquid nitrogen and stored at –80°C until nanoLC-MS/MS measurements.

NanoLC-MS/MS

The generated peptides were decomplexed with an Ultimate 3000 nanoRSLC System (ThermoFisher Scientific, Germering, Germany) equipped with 1 of 3 separation columns of 15, 25, or 50 cm length, but with similar characteristics: C18, 2 µm bead size, 75 µm inner diameter (PepMapRSLC; ThermoFisher Scientific). The nanoLC was operated in a trap column setup (PepMap nanoTrap: C18, 3 µm bead size, 75 µm inner diameter, 2 cm length; ThermoFisher Scientific). The nanoLC eluent was online coupled to ESI and an ion trap mass spectrometer applying 2 different setups. (i) ESI was performed with a distal-coated silica tip (10 µm inner diameter; New-Objectives, Woburn, MA, USA) applying a capillary voltage of –1.6 kV with an endplate offset of –0.5 kV and hot nitrogen (150°C) as dry gas. MS analysis was performed with an amaZon classic ETD ion trap mass spectrometer (referred to as classic ion trap, C-IT) (Bruker Daltonik GmbH, Bremen, Germany). Full-scan MS spectra were acquired for *m/z* of 150–2,500 with a maximum of 200,000 trapped ions (max. accumulation time 50.0 ms) and a resolution of 0.3 u (enhanced resolution scan). Per full scan, MS/MS spectra for the 10 most intense masses were acquired with a resolution of 0.5 u (ultra scan), applying active exclusion after a single measurement for 0.25 s. (ii) ESI was performed with a captive spray ion source (Bruker Daltonik GmbH) applying a capillary voltage of –1.3 kV and hot nitrogen (150°C) as dry gas. MS analysis was performed with an amaZon speed ETD ion trap mass spectrometer (referred to as improved ion trap, I-IT) (Bruker Daltonik GmbH), acquiring full-scan MS spectra from *m/z* of 300–1,400 with a maximum of 200,000 trapped ions (max. accumulation time 10.0 ms) and a resolution of 0.3 u (enhanced resolution scan). Per full scan, MS/MS spectra for the 20 most intense masses were acquired with a resolution of 0.5 u (Xtreme scan), applying active exclusion after a single measurement for 0.25 s.

Five different, linear gradients of increasing acetonitrile concentration were applied using 0.1% (v/v) formic acid in gradient-grade HPLC water as solvent A and 80% (v/v) acetonitrile, 0.1% (v/v) formic acid in gradient-grade HPLC water as solvent B. Details on the different gradients are given in Table 1. The analytical pump was operated at a constant flow rate of 300 nL/min. Sample loading onto the trap column was performed with 0.1% (v/v) TFA in gradient-grade HPLC water at a flow rate of 6 µL/min.

Peptide/Protein Identification and Data Processing

Acquired mass spectral data were processed using DataAnalysis (version 4.2; Bruker Daltonik GmbH). Subsequent protein identification was performed via the ProteinScape platform (version 3.1; Bruker Daltonik GmbH) on a Mascot server (version 2.3; Matrix Science Ltd., London, UK) against a genomic database of *P. inhibens* and *D. toluolica*, respectively, including a target-decoy strategy. Mascot search parameters were as follows: enzyme trypsin; 1 missed cleavage allowed; carbamidomethylation (C) as fixed, oxidation (M) as variable modification; peptide and MS/MS mass tolerance 0.4 Da; monoisotopic; peptide charge 2+ and 3+; instrument type ESI-TRAP; significance threshold $p < 0.05$. Spectra, the assigned peptides, and peptide scores were imported to ProteinScape by the ProteinExtractor tool [Thiele et al., 2010], and assessment was performed applying an ion score cutoff of 25.0, a minimum peptide length of 5, and peptide decoy with a false discovery rate <1.0%. Protein list compilation was performed via the ProteinExtractor, applying the same parameters. Peak capacity (*P*) was calculated according to Neue [2005]:

$$P = 1 + \frac{t_g}{\frac{1}{n}\sum_{i=1}^{n}\omega_i}, \quad (1)$$

with n being the number of peaks included, t_g the gradient time, and ω_i the peak width of peptide i. Since ω_i is equal to $4\sigma_i$, the peak width was calculated based on the full width at half maximum (FWHM) according to Neue [2005]:

$$\sigma_i = \frac{FWHM_i}{2\sqrt{2\ln 2}}. \quad (2)$$

FWHM values were determined for each analysis using DataAnalysis, and only peaks with a signal-to-noise ratio of ≥10.0 were included.

Multidimensional Scaling

Sample dissimilarity was calculated according to the euclidean distance of protein Mascot scores (nondetection assigned 0) of all assigned different proteins, as described previously [Zech et al., 2011]. The dissimilarity matrix of all pairwise distances was subsequently visualized using MDS [Kruskal and Wish, 1978; Zech et al., 2011]. Positions in the 2D plane were scanned for local minima with the Markov chain Monte Carlo sampling method parallel tempering [Geyer, 1991; Swendsen and Wang, 1986], and results were controlled both visually (Shepard diagrams) and through the Kruskal [1964] stress formula 1.

Disclosure Statement

The authors declare no conflicts of interest.

Acknowledgments

We are grateful to C. Hinrichs (Oldenburg) for technical assistance and S. Liedtke (Dreieich) for interesting discussions. This work was supported by the German Research Foundation (DFG) in the framework of the Collaborative Research Center Roseobacter TRR SFB 51.

References

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: The complete genome sequence of *Escherichia coli* K-12. Science 1997; 277:1453–1462.

Bradford MM: A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 1976;72:248–254.

Buchan A, González JM, Moran MA: Overview of the marine *Roseobacter* lineage. Appl Environ Microbiol 2005;71:5665–5677.

Buddruhs N, Pradella S, Göker M, Pauker O, Pukall R, Spröer C, Schumann P, Petersen J, Brinkhoff T: Molecular and phenotypic analyses reveal the non-identity of the *Phaeobacter gallaeciensis* type strain deposits CIP 105210^T and DSM 17395. Int J Syst Evol Microbiol 2013;63:4340–4349.

Cech NB, Enke CG: Practical implications of some recent studies in electrospray ionization fundamentals. Mass Spectrom Rev 2001;20: 362–387.

Champion KM, Zengler K, Rabus R: Anaerobic degradation of ethylbenzene and toluene in denitrifying strain EbN1 proceeds via independent substrate-induced pathways. J Mol Microbiol Biotechnol 1999;1:157–164.

Fairchild JN, Walworth MJ, Horvath K, Guiochon G: Correlation between peak capacity and protein sequence coverage in proteomics analysis by liquid chromatography-mass spectrometry/mass spectrometry. J Chromatogr A 2010;1217:4779–4783.

Geyer C: Markov chain Monte Carlo maximum likelihood; in Keramidas EM, Kaufman SM (eds): Computing Science and Statistics. Proceeding of the 23rd Symposium Interface, Seattle, Washington, April 21–24, 1991. Fairfax Station, Interface, 1991, pp 156–163.

Gilmore JM, Washburn MP: Advances in shotgun proteomics and the analysis of membrane proteomes. J Proteomics 2010;73:2078–2091.

Guiochon G: The limits of the separation power of unidimensional column liquid chromatography. J Chromatogr A 2006;1126:6–49.

Hsieh EJ, Bereman MS, Durand S, Valaskovic GA, MacCoss MJ: Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. J Am Soc Mass Spectrom 2013;24:148–153.

Köcher T, Swart R, Mechtler K: Ultra-high-pressure RPLC hyphenated to an LTQ-Orbitrap Velos reveals a linear relation between peak capacity and number of identified peptides. Anal Chem 2011;83:2699–2704.

Kruskal J, Wish M: Multidimensional Scaling. Thousand Oaks, Sage, 1978.

Kruskal JB: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29:1–27.

Liu H, Finch JW, Lavallee MJ, Collamati RA, Benevides CC, Gebler JC: Effects of column length, particle size, gradient length and flow rate on peak capacity of nano-scale liquid chromatography for peptide separations. J Chromatogr A 2007;1147:30–36.

Liu H, Sadygov RG, Yates JR 3rd: A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 2004;76:4193–4201.

MacNair JE, Lewis KC, Jorgenson JW: Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns. Anal Chem 1997;69:983–989.

Neue UD: Theory of peak capacity in gradient elution. J Chromatogr A 2005;1079:153–161.

Nogueira FC, Domont GB: Survey of shotgun proteomics. Methods Mol Biol 2014;1156:3–23.

Rabus R: Fifteen years of physiological proteo(geno)mics with (marine) environmental bacteria. Arch Physiol Biochem 2014; 120:173–187.

Rabus R, Nordhaus R, Ludwig W, Widdel F: Complete oxidation of toluene under strictly anoxic conditions by a new sulfate-reducing bacterium. Appl Environ Microbiol 1993;59: 1444–1451.

Rabus R, Venceslau SS, Wöhlbrand L, Voordouw G, Wall JD, Pereira IAC: A post-genomic view of the ecophysiology, catabolism and biotechnological relevance of sulphate-reducing prokaryotes; in Poole R (ed): Advances in Microbial Physiology. Oxford, Academic Press, 2015, vol 66, pp 55–321.

Ruiz-Ponte C, Cilia V, Lambert C, Nicolas JL: *Roseobacter gallaeciensis* sp. nov., a new marine bacterium isolated from rearings and collectors of the scallop *Pecten maximus*. Int J Syst Bacteriol 1998;48(pt 2):537–542.

Shen Y, Zhang R, Moore RJ, Kim J, Metz TO, Hixson KK, Zhao R, Livesay EA, Udseth HR, Smith RD: Automated 20 kpsi RPLC-MS and MS/MS with chromatographic peak capacities of 1000–1500 and capabilities in proteomics and metabolomics. Anal Chem 2005; 77:3090–3100.

Swendsen RH, Wang JS: Replica Monte Carlo simulation of spin glasses. Phys Rev Lett 1986; 57:2607–2609.

Thiele H, Glandorf J, Hufnagel P: Bioinformatics strategies in life sciences: from data processing and data warehousing to biological knowledge extraction. J Integr Bioinform 2010;7:141.

Wöhlbrand L, Jacob JH, Kube M, Mussmann M, Jarling R, Beck A, Amann R, Wilkes H, Reinhardt R, Rabus R: Complete genome, catabolic sub-proteomes and key-metabolites of *Desulfobacula toluolica* Tol2, a marine, aromatic compound-degrading, sulfate-reducing bacterium. Environ Microbiol 2013;15:1334–1355.

Wöhlbrand L, Ruppersberg HS, Feenders C, Blasius B, Braun HP, Rabus R: Analysis of membrane-protein complexes of the marine sulfate reducer *Desulfobacula toluolica* Tol2 by 1D blue native-PAGE complexome profiling and 2D blue native-/SDS-PAGE. Proteomics 2016;16:973–988.

Xie F, Smith RD, Shen Y: Advanced proteomic liquid chromatography. J Chromatogr A 2012;1261:78–90.

Yates JR, Ruse CI, Nakorchevsky A: Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng 2009;11:49–79.

Zech H, Echtermeyer C, Wöhlbrand L, Blasius B, Rabus R: Biological versus technical variability in 2-D DIGE experiments with environmental bacteria. Proteomics 2011;11:3380–3389.

Zech H, Hensler M, Kossmehl S, Drüppel K, Wöhlbrand L, Trautwein K, Hulsch R, Maschmann U, Colby T, Schmidt J, Reinhardt R, Schmidt-Hohagen K, Schomburg D, Rabus R: Adaptation of *Phaeobacter inhibens* DSM 17395 to growth with complex nutrients. Proteomics 2013;13:2851–2868.

Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR 3rd: Protein analysis by shotgun/bottom-up proteomics. Chem Rev 2013;113:2343–2394.

Zhou F, Lu Y, Ficarro SB, Webber JT, Marto JA: Nanoflow low pressure high peak capacity single dimension LC-MS/MS platform for high-throughput, in-depth analysis of mammalian proteomes. Anal Chem 2012;84:5133–5139.