Molecular BioSystems

RSC Publishing

REVIEW

View Article Online
View Journal | View Issue

Searching for a needle in a stack of needles: challenges in metaproteomics data analysis

Cite this: *Mol. BioSyst.*, 2013, **9**, 578

Received 5th October 2012,
Accepted 3rd December 2012

DOI: 10.1039/c2mb25415h

www.rsc.org/molecularbiosystems

Thilo Muth,[a] Dirk Benndorf,[b] Udo Reichl,[ab] Erdmann Rapp[a] and Lennart Martens*[cd]

In the past years the integral study of microbial communities of varying complexity has gained increasing research interest. Mass spectrometry-driven metaproteomics enables the analysis of such communities on the functional level, but this fledgling field still faces various technical and semantic challenges regarding experimental data analysis and interpretation. In the present review, we outline the hurdles involved and attempt to cover the most valuable methods and software implementations available to researchers in the field today. Beyond merely focusing on protein identification, we provide an overview on different data pre- and post-processing steps, such as metabolic pathway analysis, that can be useful in a typical metaproteomics workflow. Finally, we briefly discuss directions for future work.

Introduction

Advances in high throughput DNA sequencing have led to new perspectives on the molecular interactions in environmental microbial communities by retrieving comprehensive sequence information. Metagenomics[1] nowadays provides insight into the phylogenetic structure and functional potential of microbial populations and shows its variety and distribution in natural habitats.[2]

Shifting from the genome to the proteome level, metaproteomics[3] or whole community proteomics[4] aims to additionally investigate the functional profile of microbial communities, *e.g.* the immediate catalytic potential of a microbial community. The strategy pursued to achieve this insight is taken directly from more traditional proteomics approaches. Proteins are isolated from the sample, and subsequently digested with a protease such as trypsin to obtain peptides, and these are then analyzed by tandem mass spectrometry (MS/MS) to obtain fragmentation spectra.[5] These experimental spectra are then either compared to theoretical spectra obtained after *in silico* digest of a protein sequence database, or a sequence is read from them *de novo* to identify the original peptides.[6] These peptides are then in turn used to infer the proteins that were originally derived from the sample.[7]

Compared to the traditional single-organism proteomics approaches however, metaproteomics research presents several unique challenges, most notably the high complexity and heterogeneity of the samples. Indeed, microbial communities may contain hundreds to thousands of different species, with estimations for complete metagenomes predicting a complexity of more than 100 times that of the human genome.[8] Furthermore, a second common hurdle concerns the restricted availability of (predicted) protein databases as metagenomic sequence information is often not available. The resulting lack of protein sequences to match experimental spectra against them leads to a low amount of successful peptide and protein identifications, *e.g.* in samples from wastewater treatment,[9] biogas plants[10] or human gut.[11] Apart from these two primary challenges, microbial samples further carry the intrinsic biological problems of containing many homologous proteins, while performing horizontal gene transfer and exhibiting strain variety as well. A final issue relates to the low reliability of protein extraction from complex biological matrices.[12,13]

In the present review, we outline the most common bioinformatics and data analysis techniques used to tackle these problems with processing metaproteomics data. Besides providing an overview on algorithms and software used, we also mention as-yet unresolved challenges and drawbacks in metaproteomics data analysis and provide an outlook for urgently needed development in the future.

[a] *Max Planck Institute for Dynamics of Complex Technical Systems, Bioprocess Engineering, Magdeburg, Germany*
[b] *Otto-von-Guericke University, Bioprocess Engineering, Magdeburg, Germany*
[c] *Department of Medical Protein Research, VIB, Ghent, Belgium*
[d] *Department of Biochemistry, Ghent University, A. Baertsoenkaai 3, B-9000 Ghent, Belgium. E-mail: lennart.martens@vib-ugent.be; Fax: +32 92649484; Tel: +32 92649358*

Metaproteomics workflow steps and relevant software

An overview of the three steps in a typical metaproteomics workflow, along with a summary of the relevant data processing challenges and key technology platform for each step, is provided in Fig. 1.

Gene prediction and high-throughput sequencing

An important cornerstone of proteomics data processing consists of matching experimental spectra against theoretical spectra obtained from sequence databases. Ideally, the database would fully contain the sample under study, and it would therefore be advantageous to possess the full coding potential of a sample in the form of its complete metagenome. Using next-generation sequencing technologies such as pyrosequencing[14] short reads (100–500 base pairs) can be produced to cover the coding potential of the sample.[15] More detailed information on next-generation DNA sequencing can be found in more specialized comprehensive reviews.[16,17] However, in metagenomics the prediction of intact genes from these short sequence reads is challenging as the single genome assembly traditionally employed in DNA sequencing for whole genomes cannot be applied. Several metagenomic gene prediction programs have therefore been developed, including MetaGeneAnnotator,[18] Orphelia[19] and GeneMark.[20] A benchmark comparison of the algorithms and an approach to combine the different gene prediction methods can be found in the paper of Yok and Rosen.[21] As one example for metagenomic analysis and annotation, the Metagenomics RAST[22] online platform can be used. Furthermore, comprehensive reviews on the specialist and complex topic of metagenomics can be found in Wooley *et al.*[23] and Thomas *et al.*[24]

With a suitable search space defined, either through metagenomics or the use of existing sequence databases, the actual metaproteomics data processing workflow can begin, as detailed in the next sections.

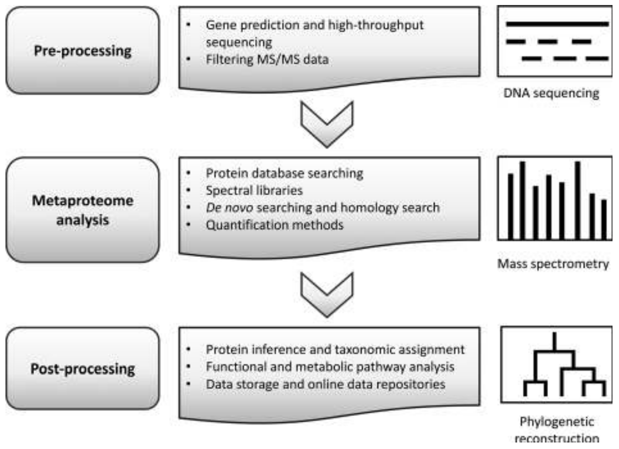
Pre-processing
• Gene prediction and high-throughput sequencing
• Filtering MS/MS data
DNA sequencing
Metaproteome analysis
• Protein database searching
• Spectral libraries
• *De novo* searching and homology search
• Quantification methods
Mass spectrometry
Post-processing
• Protein inference and taxonomic assignment
• Functional and metabolic pathway analysis
• Data storage and online data repositories
Phylogenetic reconstruction

Fig. 1 Overview of a metaproteomics workflow. The three main workflow parts are shown in the boxes on the left, with the middle column showing the relevant data processing steps in each of these three steps. The rightmost column shows the corresponding key technology platform for that step.

Filtering MS/MS data

Prior to attempting peptide and protein identification, it is important to guarantee a minimal degree of quality in the raw data. The massive amount of information produced in metaproteomics experiments renders it necessary to filter out low quality raw spectra *a priori*. While spectrum filters can be applied manually by choosing criteria such as minimum total intensity, minimum peak number or mandatory presence of certain fragment ions, several more advanced and automated methodologies have been described for spectrum quality assessment.[25–27] By applying the quality-based classifier method[26] for 25 000 experimental MS/MS spectra from metaproteomics samples we found that between 13 and 47% of the non-identified spectra could be removed due to bad quality while retaining 98% of the correctly identified spectra (see Table 1). Another promising approach to reduce computational load and simultaneously improve spectral quality is provided by spectral clustering to combine redundant spectra into meta-spectra.[28,29] The clustering method of Flikka *et al.*[28] was applied to the aforementioned dataset and resulted in a reduction of 23 458 non-identified spectra to 16 432 clusters and 2408 meta-spectra (see Table 2).

In contrast, 1542 identified spectra could be combined to 1391 clusters and 108 meta-spectra. These numbers show that the non-identified spectra exhibit a higher redundancy in comparison to the identified spectra.

Protein database searching

In general, the algorithms for protein identification correlate experimental fragment ion spectra with theoretical spectra

Table 1 Spectrum quality classification. The table shows the true-positive (TP) and the true-negative (TN) rate for five different test datasets. The TP rate represents the rate of identified spectra labeled as good spectra. The TN rate is the rate of unidentified spectra classified as bad

Dataset	TP rate	TN rate	No. spectra
Dataset1	0.98881	0.23225	5000
Dataset2	0.98282	0.46995	5000
Dataset3	0.98023	0.18597	5000
Dataset4	0.97826	0.12867	5000
Dataset5	0.97701	0.38384	5000
Sum	0.98119	0.28097	25 000

Table 2 Spectral clustering. The table shows the clustering results for the identified (Id) and unidentified (Non-Id) spectra of the same datasets as used for the spectrum quality classification in Table 1. All spectra were clustered (Id and Non-Id clusters) and clusters with at least two containing spectra were merged into meta-spectra (Id and Non-Id merged)

Dataset	Id spectra	Id clusters	Id merged	Non-Id spectra	Non-Id clusters	Non-Id merged
Dataset1	268	246	18	4732	4110	438
Dataset2	291	278	13	4709	4128	490
Dataset3	354	328	25	4646	4151	439
Dataset4	368	344	22	4632	4172	402
Dataset5	261	195	30	4739	3981	639
Sum	1542	1391	108	23 458	16 432	2408

calculated for each peptide derived after *in silico* digest of a protein sequence database. Several commercial and non-commercial software packages exist for this purpose, and these can be applied in metaproteomics data analysis as well. The most prominent commercial search algorithms are SEQUEST[30] and MASCOT.[31] Various high performing free and open source alternatives such as X!Tandem,[32] OMSSA,[33] Myrimatch,[34] Crux[35] and InsPect[36] are available as well. However, each of these algorithms suffers from the issue of false positive identifications, and controlling the overall false discovery rate (FDR) is therefore essential.[37] Several computational techniques have been developed to estimate the FDR at both the peptide and the protein level, including statistical modeling approaches[38,39] and the highly popular target-decoy[40] approach that has already been included in several commercial search engines. The generation of a decoy database is usually done by reversing or shuffling the input database sequences.[41] A major issue comes with the question how to reasonably generate a decoy database. Software packages such as QVality[42] and Percolator[43] make it possible to use the results of any search engine and apply the target-decoy strategy to them. In the case of Percolator, the use of advanced machine learning methods is furthermore employed to increase the sensitivity of identification at constant FDR. Although FDR estimation is useful in the traditional, more limited search space of a single organism's proteome, where traditional search engines provide a sufficiently discriminating score between correct and incorrect identification candidates, this scoring system resolution can deteriorate quickly when the search space increases in complexity, as is the case in metaproteomics.[44] This effect dramatically reduces sensitivity, both in metaproteomics and proteogenomics.[45] One approach to regain some sensitivity by providing a scoring system that better separates correct identifications from incorrect ones, hinges on the combination of multiple database search engines as implemented by iProphet[46] and MSblender.[47] These latter approaches benefit from the partially complementary nature of search engine identifications.[48,49]

Spectral libraries

Database search engines have the disadvantage of having to rely on existing protein sequence information. The alternative approach of spectral library searching is based on recorded and identified high-quality mass spectra as more flexible and accurate references for identification. Here, the experimental mass spectra are compared against reference spectra in the spectral library and their similarity is taken as discriminative measure. Popular spectral library search engines are SpectraST,[50] X!Hunter[51] and the NIST MS search software.[52] A shortcoming of this method is the fact that a reliable reference spectral library has to be available for searching, and this can be difficult to obtain, especially in the case of metaproteomics data. Indeed, spectral libraries are mostly derived from single-organism experiments on model species, most notably human, mouse or yeast. However, the SpecraST tools allow spectral libraries to be home-built,[53] and as a result it will be interesting to devise ways to develop spectral libraries for metaproteomics. Note that spectral identification of the spectra in the library is not necessarily required. One could track the occurrence of non-identified spectra across different experiments and ecosystems, and thus detect significantly represented spectra for more detailed follow-up analysis. Furthermore, interesting perturbations across systems could be detected through marker spectra, even if they have not yet been identified. This approach is quite similar to typical profile-based biomarker discovery strategies,[54] with that difference that MS2 features would be used here instead of MS1 features for profile-based approaches. This reduces the total amount of available data, but also the amount of noise among the data, and may therefore be more successful than the profile-based strategy.

De novo sequencing and homology search

The method of *de novo* sequencing deduces aminoacid sequences directly from fragmentation spectra. It thus relies solely on the information present in the spectra and therefore obviates the need for a protein sequence database. This makes *de novo* searching a very useful tool in metaproteomics research. Furthermore, *de novo* approaches can also identify previously unknown peptide sequences or sequences carrying unexpected post-translational modifications. The major caveat of such methods is that they require very high quality data to function reliably,[55] emphasizing the importance of preprocessing steps such as binning, noise reduction and other filtering techniques.[56] Despite the corresponding low success rate, *de novo* searching often remains the only possibility in the context of metaproteomics experiments where the appropriate protein sequence information is unknown or unavailable. The most prominent *de novo* software packages are currently the freely available PepNovo+[57] and the commercial PEAKS[58] suite, with several other tools available as well.[59] After the generation of *de novo* peptide sequences a BLAST search[60] can be employed in order to identify candidate homology proteins. The MS BLAST homology searching protocol[61] represents a web-based application specifically tailored to MS-based protein similarity searches. However, this approach has to be used with care as the obtained protein matches may result from incorrectly derived peptide sequences due to the error-prone nature of *de novo* sequencing. *De novo* sequencing results therefore require tedious manual inspection, limiting the overall throughput of this method. It is worth noting that BLAST searches do not take into account mass spectral information, and confidently identified aminoacids may well be considered mutated by BLAST without reflecting badly on the downstream alignment score. Cantarel *et al.* used a combination of PepNovo+ and PEAKS for a whole-community proteomics workflow to obtain high confidence consensus sequence tags.[62] The derived *de novo* sequences were then mapped onto predicted protein sequences from metagenomic contigs, establishing the link between assembled metagenomics data and proteomics *de novo* peptide sequence data.

Quantification methods

In order to analyze and compare protein expression levels, methods of protein quantification[63] need to be applied to microbial community samples. Quantification of proteins in 2D-gels worked well for decades, but separation of proteins from environmental samples with 2D-gel electrophoresis is extremely challenging, with time-consuming optimization required for every novel sample type. Furthermore, none of the popular quantification methods such as ICAT[64] or iTRAQ[65] in gel-free proteomics could be successfully applied to measure protein expression in complex environmental samples. To make matters worse, the performance of software for label-based quantitative proteomics is often still lagging behind the latest advances in the techniques themselves.[66] Fortunately, so-called label-free techniques provide a promising alternative for peptide and protein quantification in microbial community proteomics. Label-free quantification has the advantage that it can be applied directly to protein identification data and represents a very simple and straightforward approach for estimating protein abundance.[67] In this approach, the number of identified spectra for a specific protein is counted, and the higher these spectral counts are the higher the assumed protein abundance. An improved approach is provided by the normalized spectral abundance factor (NSAF)[68] that takes into account the fact that proteins with longer sequences usually have more peptide identifications than shorter proteins and thus obtains more consistent results by normalizing for this effect.

Protein inference and taxonomic assignment

The question of how to correctly assign peptides to proteins has been formulated as the protein inference problem, and is eloquently described in ref. 7. Peptides can be shared among protein splice isoforms in eukaryotes, across homologous proteins from different species, or across recurring functional domains, leading to so-called degenerate peptides. These degenerate peptides lead to essentially irresolvable ambiguities in protein identification that can easily lead to incompatible or incorrect data interpretations.[41] In metaproteomics, another protein inference challenge is added at the taxonomy level. Indeed, rather than mapping to multiple proteins, many peptides identified in metaproteomics experiments can map to multiple species, genera or even families. Additionally, database search engines often display only a subset of all possible protein identifications (the so-called best hits) for a limited number of species and this limitation in the provided output has to be taken into consideration for samples of unknown species composition. A promising strategy for taxonomic evaluation is provided by the following workflow: the identified peptide sequences (derived from database or *de novo* searching) are submitted as a preprocessing step to protein BLAST.[60] With the derived results the homology-matching software MEGAN[69] then computes a phylogenetic tree for the dataset by employing the NCBI taxonomy database. Details and instructions on the metagenomic and metaproteomic analysis with MEGAN (free for academic use) can be found in ref. 70. For metaproteomics, Schneider *et al.* built a perl script based workflow (PROPHANE: PROteomics result Pruning and Homology group ANnotation Engine) that fuses protein hits sharing common peptides to a group.[71] Subsequently, the taxonomic affiliation is assigned on the level where the different affiliations of the hits in the tree are converging.[71] Although this approach is promising, it does not take into account that certain aminoacid sequences within a group of hits are conserved and thus should carry less weight in this analysis.

Functional and metabolic pathway analysis

Protein identification lists derived from a database are however not the end result of a proteomics experiment. Typically, a perspective focusing on the meaningful semantic interpretation of the obtained protein data is sought after. This section therefore describes possible strategies and examples for post-processing analysis in the context of metaproteomics research.

Several publicly available databases can be used for the functional annotation of individual identified proteins. The UniProt knowledgebase[72] serves as an excellent starting point for collecting relevant information on specific proteins. The Cluster of Orthologous Groups database (COG,[73,74]) maps both prokaryotic and eukaryotic proteins to COG groups and each group is linked to a COG functional category, *e.g.* aminoacid transport and metabolism. COGs can be used to assign predicted functions to coding sequences, as shown by Schlüter *et al.* for the metagenomic analysis of a biogas-producing microbial community.[75] Furthermore, Kolmeder *et al.* applied the COG classification for the functional analysis of the human intestinal metaproteome.[11] The main issue with COG is that it has not been updated since 2003, so it does not include any novel information obtained since then. The Gene Ontology (GO) project is a collaborative effort that addresses the need for consistent descriptions of gene products across different databases.[76] GO provides three structured vocabularies, so-called *ontologies* that describe biological processes, cellular components and molecular functions independently of associated species. Several tools, *e.g.* the Ontologizer software[77] and the web interface-based DAVID,[78] enable protein analysis *via* ontologies and protein families.

InterPro is a protein domain database providing a wide range of information on protein sequence function and annotation.[79] Its aim is to integrate information from other secondary protein databases on functional sites and domains, such as PROSITE,[80] PRINTS,[81] SMART,[82] Pfam[83] and ProDom.[84] A search against InterPro can quickly reveal functional or known domains in an otherwise uncharacterized protein, and can thus contribute to a functional understanding of the identified proteins.

KEGG (the Kyoto Encyclopedia of Genes and Genomes) is a data resource integrating genomic, biochemical and functional information that focuses on intermediate metabolic and regulatory pathways.[85] The idea is to model expression data and to understand higher-order cellular processes. The data model is based on catalyst activities *via* Enzyme Commission (EC) numbers. This may prove an issue for proteomics data however, as the relation of a protein to its enzymatic function may not be a single link, especially in the case of a polypeptide chain with multiple functions. The KEGG automated annotation server[86] can be used to place proteins into KEGG ontologies. Mapping of identified

COGs onto KEGG pathways was employed by Kolmeder *et al.* and a global metabolic pathway map could thus be retrieved.[11] In addition to taxonomical analysis, the PROPHANE software assigns mass spectra to functional groups using the COG and KEGG database.[71] With PROPHANE, protein annotations are being validated by various complementary approaches, including tools such as ClustalW,[87] BLAST[60] and Bioperl.[88] The Reactome project collects structured information on canonical biological pathways and processes in human,[89] and contains supplementary information with orthologous molecular reactions in mouse, rat, worm and other model organisms. Extending beyond a mere database, Reactome actually resembles a curated journal that is expert-authored and peer-reviewed. However, the main disadvantage of using Reactome for metaproteomics research lies in the fact that this knowledgebase is limited to few species, all of which are higher eukaryotes.

The MetaCyc database is a curated reference database of metabolic pathways from a wide variety of organisms, with a particular emphasis on microorganisms and plants.[90] Being connected to MetaCyc, the BioCyc database offers further species-specific pathway and genome databases.[91] One of the future plans is to integrate the genomes of the Human Microbiome Project.[92] The metaproteomic analysis of the human salivary microbiota by Rudney *et al.* was supplemented by searching the MetaCyc database for prokaryotic pathways.[93] The authors mention the problem that not all sequenced microbial species were included in the database, and that pathway matches were therefore based on data from closely related taxa in some cases.

Additionally, the presence of proteins in common metabolic pathways, protein–protein interaction maps or regulatory networks could also help to interpret the obtained metaproteomics data, possibly even serving to confirm weak protein identifications.[94]

Data storage and online data repositories

Although various data analysis software packages exist, it is often difficult to integrate the output from the various analysis tools. Many laboratories tend to implement their own in-house database systems and scripts, but these local solutions have the disadvantage that the results are mostly not replicable for any other laboratory. It is therefore helpful to organize the data workflow by using a laboratory information management system (LIMS) or data analysis tools that provide a unified relational database system at the backend. Proteus[95] is a commercial data integration and analysis system, capable of storing data from many popular proteomics analysis tools. CPAS,[96] ms_lims,[97] MASPECTRAS,[98] myProMS[99] and OpenMS[100] represent freely available frameworks for mass spectrometry-based data analysis and provide similar storage capabilities. Comparative reviews of these various LIMS systems can be found in ref. 101 and 102.

Local data storage is not the end of the data storage and dissemination chain, however. Indeed, several initiatives have been started over the past years in order to make experimental proteomics data publicly available to the scientific community at large. Among the most popular online databases and repositories for proteomics data are the PRoteomics IDEntifications database (PRIDE),[103] the Global Proteome Machine Database (GPMDB),[104] PeptideAtlas,[105] ProteomeCommons Tranche[106] and NCBI Peptidome.[107] Peptidome has since ceased to exist, but all its data have been safely archived in PRIDE.

In addition to pure storage capacities, these repositories provide possibilities of exchanging data between different laboratories and allow the reanalysis of data, for instance when new search algorithms or functional annotation tools become available. An overview of the functionalities of these online systems can be found in ref. 108–110.

Conclusions

Like all proteomics disciplines, metaproteomics research strongly benefits from the developments of more sensitive instruments and optimized sample preparation. But while both trends increase the amount of acquired data, the corresponding increase in the yield of information does not yet follow suit. The heterogeneity and complexity of the samples raise the question of the validity of the obtained results, and advanced pre- and post-processing steps are needed to bring the performance of the data analysis to the level of the data acquisition. The lack of sequence information remains an important issue for successful peptide and protein identification in metaproteomics, and both *de novo* sequencing approaches as well as alternatives such as spectral library searching have to be improved further to overcome this problem. In addition, software tools required for the reliable functional and taxonomic assignment of primary data are currently lacking yet are urgently needed. The current workflow in processing metaproteomics data depends on collating together pipelines from a variety of different tools from various research fields. It would be of great benefit to the field if a software platform were to be constructed that integrates and improves these various functions in a single application. In opposite of these data analysis and interpretation challenges however, the field of metaproteomics holds significant promise for the future. Driven by the steadily growing amount of available metagenomic sequence information, continued improvements in the analytical and bioinformatics workflows will enable metaproteomics to increasingly contribute to a better understanding of microbial communities in the near future.

Abbreviations

FDR	False discovery rate
MS/MS	Tandem mass spectrometry

Acknowledgements

L.M. acknowledges the financial support of Ghent University (Multidisciplinary Research Partnership ''Bioinformatics: from nucleotides to networks'') and the PRIME-XS project funded by the European Union 7th Framework Program under grant agreement number 262067.

References

1 J. Singh, A. Behal, N. Singla, A. Joshi, N. Birbian, S. Singh, V. Bali and N. Batra, *Biotechnol. J.*, 2009, **4**, 480–494.

2 S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz and E. M. Rubin, *Science*, 2005, **308**, 554–557.

3 P. Wilmes and P. L. Bond, *Environ. Microbiol.*, 2004, **6**, 911–920.

4 J. F. Banfield, N. C. Verberkmoes, R. L. Hettich and M. P. Thelen, *OMICS*, 2005, **9**, 301–333.

5 K. Gevaert, P. Van Damme, B. Ghesquiere, F. Impens, L. Martens, K. Helsens and J. Vandekerckhove, *Proteomics*, 2007, **7**, 2698–2718.

6 M. Vaudel, A. Sickmann and L. Martens, *Expert Rev. Proteomics*, 2012, **9**, 519–532.

7 A. I. Nesvizhskii and R. Aebersold, *Mol. Cell Proteomics*, 2005, **4**, 1419–1440.

8 F. Bäckhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson and J. I. Gordon, *Science*, 2005, **307**, 1915–1920.

9 R. Kuhn, D. Benndorf, E. Rapp, U. Reichl, L. L. Palese and A. Pollice, *Proteomics*, 2011, **11**, 2738–2744.

10 A. Hanreich, R. Heyer, D. Benndorf, E. Rapp, M. Pioch, U. Reichl and M. Klocke, *Can. J. Microbiol.*, 2012, **58**, 917–922.

11 C. A. Kolmeder, M. de Been, J. Nikkilä, I. Ritamo, J. Mättö, L. Valmu, J. Salojärvi, A. Palva, A. Salonen and M. W. de Vos, *PLoS One*, 2012, **7**, e29913.

12 K. Chourey, J. Jansson, N. VerBerkmoes, M. Shah, K. L. Chavarria, L. M. Tom, E. L. Brodie and R. L. Hettich, *J. Proteome Res.*, 2010, **9**, 6615–6622.

13 L. Giagnoni, F. Magherini, L. Landi, S. Taghavi, A. Modesti, L. Bini, P. Nannipieri, D. Van der lelie and G. Renella, *Eur. J. Soil Sci.*, 2011, **62**, 74–81.

14 M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg, *Nature*, 2005, **437**, 376–380.

15 M. L. Metzker, *Nat. Rev. Genet.*, 2010, **11**, 31–46.

16 J. Shendure and H. Ji, *Nat. Biotechnol.*, 2008, **26**, 1135–1145.

17 D. R. Bentley, *Curr. Opin. Genet. Dev.*, 2006, **16**, 545–552.

18 H. Noguchi, T. Taniguchi and T. Itoh, *DNA Res.*, 2008, **15**, 387–396.

19 K. J. Hoff, T. Lingner, P. Meinicke and M. Tech, *Nucleic Acids Res.*, 2009, **37**, W101–W105.

20 W. Zhu, A. Lomsadze and M. Borodovsky, *Nucleic Acids Res.*, 2010, **38**, e132.

21 N. G. Yok and G. L. Rosen, *BMC Bioinf.* 2011, **12**, 20.

22 F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening and R. Edwards, *BMC Bioinf.* 2008, **9**, 386.

23 J. C. Wooley, A. Godzik and I. Friedberg, *PLoS Comput. Biol.*, 2010, **6**, e1000667.

24 T. Thomas, J. Gilbert and F. Meyer, *Microb. Inf. Exp.*, 2012, **2**, 3.

25 J. W. H. Wong, M. J. Sullivan, H. M. Cartwright and G. Cagney, *BMC Bioinf.* 2007, **8**, 51.

26 K. Flikka, L. Martens, J. Vandekerckhove, K. Gevaert and I. Eidhammer, *Proteomics*, 2006, **6**, 2086–2094.

27 A. I. Nesvizhskii, F. F. Roos, J. Grossmann, M. Vogelzang, J. S. Eddes, W. Gruissem, S. Baginsky and R. Aebersold, *Mol. Cell Proteomics*, 2006, **5**, 652–670.

28 K. Flikka, J. Meukens, K. Helsens, J. Vandekerckhove, I. Eidhammer, K. Gevaert and L. Martens, *Proteomics*, 2007, **7**, 3245–3258.

29 A. M. Frank, N. Bandeira, Z. Shen, S. Tanner, S. P. Briggs, R. D. Smith and P. A. Pevzner, *J. Proteome Res.*, 2008, **7**, 113–122.

30 J. K. Eng, A. L. McCormack and J. R. Yates, *J. Am. Soc. Mass Spectrom.*, 1994, **5**, 976–989.

31 D. N. Perkins, D. J. Pappin, D. M. Creasy and J. S. Cottrell, *Electrophoresis*, 1999, **20**, 3551–3567.

32 R. Craig and R. C. Beavis, *Bioinformatics*, 2004, **20**, 1466–1467.

33 L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi and S. H. Bryant, *J. Proteome Res.*, 2004, **3**, 958–964.

34 D. L. Tabb, C. G. Fernando and M. C. Chambers, *J. Proteome Res.*, 2007, **6**, 654–661.

35 C. Y. Park, A. A. Klammer, L. Käll, M. J. MacCoss and W. S. Noble, *J. Proteome Res.*, 2008, **7**, 3022–3027.

36 S. Tanner, H. Shu, A. Frank, L.-C. Wang, E. Zandi, M. Mumby, P. A. Pevzner and V. Bafna, *Anal. Chem.*, 2005, **77**, 4626–4639.

37 M. Vaudel, J. M. Burkhart, A. Sickmann, L. Martens and R. P. Zahedi, *Proteomics*, 2011, **11**, 2105–2114.

38 A. Keller, A. I. Nesvizhskii, E. Kolker and R. Aebersold, *Anal. Chem.*, 2002, **74**, 5383–5392.

39 B. Y. Renard, W. Timm, M. Kirchner, J. A. Steen, F. A. Hamprecht and H. Steen, *Anal. Chem.*, 2010, **82**, 4314–4318.

40 J. E. Elias and S. P. Gygi, *Nat. Methods*, 2007, **4**, 207–214.

41 L. Martens and H. Hermjakob, *Mol. Biosyst.*, 2007, **3**, 518–522.

42 L. Käll, J. D. Storey and W. S. Noble, *Bioinformatics*, 2008, **24**, i42–i48.

43 L. Käll, J. D. Canterbury, J. Weston, W. S. Noble and M. J. MacCoss, *Nat. Methods*, 2007, **4**, 923–925.

44 N. Colaert, S. Degroeve, K. Helsens and L. Martens, *J. Proteome Res.*, 2011, **10**, 5555–5561.

45 K. Krug, S. Nahnsen and B. Macek, *Mol. Biosyst.*, 2011, **7**, 284–291.

46 D. Shteynberg, E. W. Deutsch, H. Lam, J. K. Eng, Z. Sun, N. Tasman, L. Mendoza, R. L. Moritz, R. Aebersold and A. I. Nesvizhskii, *Mol. Cell Proteomics*, 2011, **10**, M111.007690.

47 T. Kwon, H. Choi, C. Vogel, A. I. Nesvizhskii and E. M. Marcotte, *J. Proteome Res.*, 2011, **10**, 2949–2958.

48 E. A. Kapp, F. Schutz, L. M. Connolly, J. A. Chakel, J. E. Meza, C. A. Miller, D. Fenyo, J. K. Eng, J. N. Adkins, G. S. Omenn and R. J. Simpson, *Proteomics*, 2005, **5**, 3475–3490.

49 C. Stephan, K. A. Reidegeld, M. Hamacher, A. van Hall, K. Marcus, C. Taylor, P. Jones, M. Muller, R. Apweiler, L. Martens, G. Korting, D. C. Chamrad, H. Thiele, M. Bluggel, D. Parkinson, P. A. Binz, A. Lyall and H. E. Meyer, *Proteomics*, 2006, **6**, 5015–5029.

50 H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein and R. Aebersold, *Proteomics*, 2007, **7**, 655–667.

51 R. Craig, J. C. Cortens, D. Fenyo and R. C. Beavis, *J. Proteome Res.*, 2006, **5**, 1843–1849.

52 S. Stein and D. Scott, *J. Am. Soc. Mass Spectrom.*, 1994, **5**, 859–866.

53 H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, S. E. Stein and R. Aebersold, *Nat. Methods*, 2008, **5**, 873–875.

54 K. R. Coombes, J. S. Morris, J. Hu, S. R. Edmonson and K. A. Baggerly, *Nat. Biotechnol.*, 2005, **23**, 291–292.

55 S. Pevtsov, I. Fedulova, H. Mirzaei, C. Buck and X. Zhang, *J. Proteome Res.*, 2006, **5**, 3018–3028.

56 K. Ning, N. Ye and H. W. Leong, *J. Bioinf. Comput. Biol.*, 2008, **6**, 467–492.

57 A. Frank and P. Pevzner, *Anal. Chem.*, 2005, **77**, 964–973.

58 B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie, *Rapid Commun. Mass Spectrom.*, 2003, **17**, 2337–2342.

59 J. Allmer, *Expert Rev. Proteomics*, 2011, **8**, 645–657.

60 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.

61 A. Shevchenko, S. Sunyaev, A. Loboda, P. Bork, W. Ens and K. G. Standing, *Anal. Chem.*, 2001, **73**, 1917–1926.

62 B. L. Cantarel, A. R. Erickson, N. C. VerBerkmoes, B. K. Erickson, P. A. Carey, C. Pan, M. Shah, E. F. Mongodin, J. K. Jansson, C. M. Fraser-Liggett and R. L. Hettich, *PLoS One*, 2011, **6**, e27173.

63 M. Vaudel, A. Sickmann and L. Martens, *Proteomics*, 2010, **10**, 650–670.

64 S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold, *Nat. Biotechnol.*, 1999, **17**, 994–999.

65 P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson and D. J. Pappin, *Mol. Cell Proteomics*, 2004, **3**, 1154–1169.

66 H. Choi, D. Fermin and A. I. Nesvizhskii, *Mol. Cell Proteomics*, 2008, **7**, 2373–2385.

67 H. Liu, R. G. Sadygov and J. R. Yates, *Anal. Chem.*, 2004, **76**, 4193–4201.

68 B. Zybailov, A. L. Mosley, M. E. Sardiu, M. K. Coleman, L. Florens and M. P. Washburn, *J. Proteome Res.*, 2006, **5**, 2339–2347.

69 D. H. Huson, A. F. Auch, J. Qi and S. C. Schuster, *Genome Res.*, 2007, **17**, 377–386.

70 D. H. Huson and S. Mitra, *Methods Mol. Biol.*, 2012, **856**, 415–429.

71 T. Schneider, K. M. Keiblinger, E. Schmid, K. Sterflinger-Gleixner, G. Ellersdorfer, B. Roschitzki, A. Richter, L. Eberl, S. Zechmeister-Boltenstern and K. Riedel, *ISME J.*, 2012, **6**, 1749–1762.

72 R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi and L.-S. L. Yeh, *Nucleic Acids Res.*, 2004, **32**, D115–D119.

73 R. L. Tatusov, E. V. Koonin and D. J. Lipman, *Science*, 1997, **278**, 631–637.

74 R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D. A. Natale, *BMC Bioinf.* 2003, **4**, 41.

75 A. Schlüter, T. Bekel, N. N. Diaz, M. Dondrup, R. Eichenlaub, K.-H. Gartemann, I. Krahn, L. Krause, H. Krömeke, O. Kruse, J. H. Mussgnug, H. Neuweger, K. Niehaus, A. Pühler, K. J. Runte, R. Szczepanowski, A. Tauch, A. Tilker, P. Viehöver and A. Goesmann, *J. Biotechnol.*, 2008, **136**, 77–90.

76 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nat. Genet.*, 2000, **25**, 25–29.

77 S. Bauer, S. Grossmann, M. Vingron and P. N. Robinson, *Bioinformatics*, 2008, **24**, 1650–1651.

78 D. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane and R. A. Lempicki, *Genome Biol.*, 2007, **8**, R183.

79 R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, E. M. Zdobnov and InterPro Consortium, *Bioinformatics*, 2000, **16**, 1145–1150.

80 C. J. A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch and P. Bucher, *Briefings Bioinf.*, 2002, **3**, 265–274.

81 T. K. Attwood, M. E. Beck, A. J. Bleasby and D. J. Parry-Smith, *Nucleic Acids Res.*, 1994, **22**, 3590–3596.

82 J. Schultz, F. Milpetz, P. Bork and C. P. Ponting, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 5857–5864.

83 A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe and E. L. Sonnhammer, *Nucleic Acids Res.*, 2000, **28**, 263–266.

84 C. Bru, E. Courcelle, S. Carrère, Y. Beausse, S. Dalmar and D. Kahn, *Nucleic Acids Res.*, 2005, **33**, D212–D215.

85 M. Kanehisa and S. Goto, *Nucleic Acids Res.*, 2000, **28**, 27–30.

86 Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa and M. Kanehisa, *Nucleic Acids Res.*, 2007, **35**, W182–W185.

87 J. D. Thompson, T. J. Gibson and D. G. Higgins, *Curr. Protoc. Bioinformatics*, 2002, ch. 2, Unit 2.3.

88 J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson and E. Birney, *Genome Res.*, 2002, **12**, 1611–1618.

89 D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D'Eustachio and L. Stein, *Nucleic Acids Res.*, 2011, **39**, D691–D697.

90 C. J. Krieger, P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee and P. D. Karp, *Nucleic Acids Res.*, 2004, **32**, D438–D442.

91 R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang and P. D. Karp, *Nucleic Acids Res.*, 2008, **36**, D623–D631.

92 J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, C. C. Baker, V. Di Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A. R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson and M. Guyer, *Genome Res.*, 2009, **19**, 2317–2323.

93 J. D. Rudney, H. Xie, N. L. Rhodus, F. G. Ondrey and T. J. Griffin, *Mol. Oral Microbiol.*, 2010, **25**, 38–49.

94 W. W. Goh, Y. H. Lee, M. Chung and L. Wong, *Proteomics*, 2012, **12**, 550–563.

95 Proteus. http://www.genologics.com.

96 A. Rauch, M. Bellew, J. Eng, M. Fitzgibbon, T. Holzman, P. Hussey, M. Igra, B. Maclean, C. W. Lin, A. Detter, R. Fang, V. Faca, P. Gafken, H. Zhang, J. Whiteaker, J. Whitaker, D. States, S. Hanash, A. Paulovich and M. W. McIntosh, *J. Proteome Res.*, 2006, **5**, 112–121.

97 K. Helsens, N. Colaert, H. Barsnes, T. Muth, K. Flikka, A. Staes, E. Timmerman, S. Wortelkamp, A. Sickmann, J. Vandekerckhove, K. Gevaert and L. Martens, *Proteomics*, 2010, **10**, 1261–1264.

98 J. Hartler, G. G. Thallinger, G. Stocker, A. Sturn, T. R. Burkard, E. Körner, R. Rader, A. Schmidt, K. Mechtler and Z. Trajanoski, *BMC Bioinf.* 2007, **8**, 197.

99 P. Poullet, S. Carpentier and E. Barillot, *Proteomics*, 2007, **7**, 2553–2556.

100 M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert and O. Kohlbacher, *BMC Bioinf.* 2008, **9**, 163.

101 C. Piggee, *Anal. Chem.*, 2008, **80**, 4801–4806.

102 C. Stephan, M. Kohl, M. Turewicz, K. Podwojski, H. E. Meyer and M. Eisenacher, *Proteomics*, 2010, **10**, 1230–1249.

103 L. Martens, H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove and R. Apweiler, *Proteomics*, 2005, **5**, 3537–3545.

104 R. Craig, J. P. Cortens and R. C. Beavis, *J. Proteome Res.*, 2004, **3**, 1234–1242.

105 E. W. Deutsch, H. Lam and R. Aebersold, *EMBO Rep.*, 2008, **9**, 429–434.

106 ProteomeCommons.org Tranche. http://www.tranche.proteomecommons.org.

107 L. Ji, T. Barrett, O. Ayanbule, D. B. Troup, D. Rudnev, R. N. Muertter, M. Tomashevsky, A. Soboleva and D. J. Slotta, *Nucleic Acids Res.*, 2010, **38**, D731–D735.

108 J. A. Mead, L. Bianco and C. Bessant, *Proteomics*, 2009, **9**, 861–881.

109 M. Riffle and J. K. Eng, *Proteomics*, 2009, **9**, 4653–4663.

110 J. A. Vizcaino, J. M. Foster and L. Martens, *J. Proteomics*, 2010, **73**, 2136–2146.