

Improvement of proteomic data analysis and its application to *Halobacterium salinarum*

Carolina Garcia-Rizo, Friedhelm Pfeiffer, Christian Klein, Frank Siedler, and
Dieter Oesterhelt*

Department of Membrane Biochemistry, Max-Planck Institute of Biochemistry,
Am Klopferspitz 18a, D-82152 Martinsried, Germany

Running head: Improvement of proteomic data analysis

* To whom correspondence should be addressed

Abstract

Motivation

Open reading frame (ORF) prediction is the first step in genome annotation and experimental validation by proteomic analysis is required to show that the predicted ORFs actually encode proteins. This is especially necessary for *Halobacterium salinarum* in which - due to a scarcity of stop codons - more than 86% of the chromosome contains in addition to the ORF for the encoded gene at least one additional spurious ORF of at least 100 codons. Another general problem is the correct prediction of the start codon. In addition, software tools for proteomic identification are fallible, producing both false positive and false negative results. In order to correct these deficiencies, we set out to develop a toolbox for post-processing of proteomic results. It should rapidly select relevant mass peaks that should be subjected to more detailed experimental analysis while the sample is still in the mass spectrometer.

Results

By rapidly selecting relevant mass peaks from a large dataset, the toolbox initiated post source decay (PSD) experiments and thus permitted an

enhancement of the information obtained from the original MALDI-TOF peptide fingerprints. In this way, the following problems were addressed: (a) unambiguous identification of proteins which produced only a low MASCOT score, (b) correction of start codon assignments, (c) validation of the start codon assignment, (d) verification of missed trypsin cleavage sites, and (e) identification of posttranslational modifications.

The toolbox was applied to the proteome of *Halobacterium salinarum* strain R1 (DSM 671) (www.halolex.mpg.de) and should be useful in the study of other proteomes.

Keywords: Peptide mass fingerprints / Protein identification / Proteomic analysis / Posttranslational modification / Bioinformatics / *Halobacterium salinarum*

Introduction

Proteome analysis provides both static and dynamic information about the set of proteins present in an organism. The complete set of possible proteins of a cell can be predicted, once its genome is sequenced, using various gene finder algorithms. In *Halobacterium salinarum*, which has a chromosome size of 2.0 Mbp and total genome size of 2.6 Mbp, the occurrence of a large number of

spurious open reading frames (ORFs) results from the scarcity of stop codons which is due to a distinctive bias at the mononucleotide (high GC content) and dinucleotide level (underrepresentation of the TA sequence). More than 86% of the chromosome and 77% of the genome contains in addition to the ORF for the encoded gene at least one additional spurious ORF of at least 100 codons. This results in 9312 ORFs (93% being longer than 100 codons) of which only 2784 are considered to be protein-coding genes. In addition, the reading frame is open 5' of the start codon for many of the proteins from *H. salinarum*. In order to establish which ORFs really encode proteins, we carried out proteome analysis which in combination with bioinformatic procedures allows a higher degree of confidence in the determined protein inventory for this organism.

Protein identification is achieved by MALDI-TOF peptide mass fingerprint analysis (Pappin *et al.*, 1993). The MASCOT search engine (www.matrixscience.com) used for protein identification returns a “probability-based MOWSE score” (Pappin *et al.*, 1993) for each protein. This score is based on a comparison of the theoretical mass of tryptic peptides with experimental mass peaks, and reflects the unlikelyhood that a given set of mass peaks occurs in a sample by chance. Due to the strong correlation between the number of matching peaks and the MOWSE score, it is mandatory to maximize the number of matching peaks and the correctness of their assignment.

Additional matching peaks may be identified by correcting the start codon assignment. On the other hand matching peaks may be incorrectly assigned due to a casual number correlation (false positives) and will therefore artificially elevate scores. Discrimination between real and misassigned peaks is possible by post source decay experiments (PSD) (Talbo *et al.*, 2001, Rapp *et al.*, 2000) and TOF-TOF MS or ESI/MS/MS methods which generate sequence tags (Mann and Wilm, 1994, Aebersold and Mann, 2003) from peptides. Validation of the correct assignment of matching peaks is especially important when the existence of that peak confirms biological annotation information, e.g. the N-terminal peptide or a posttranslationally modified peptide.

Our approach aims to select from a large data set those individual mass peaks which are most relevant for improvement of the reliability of proteomic identifications and should maximize the information obtained from the sample. Mass peak selection is fast enough to permit continuation of the experiment while the sample is still in the mass spectrometer. Combining genome and proteome analysis, we have developed a bioinformatic toolbox to fulfil this goal. It analyses MASCOT results, identifies single mass peaks of interest suitable for further experimental analysis, e.g. PSD, allowing experimentalists to initiate these analyses more rapidly than previously possible. This approach demonstrates a synergistic effect obtained from integration of experimental and bioinformatic procedures.

Materials and Methods

2D gel electrophoresis and mass spectrometric analysis

Proteins were analysed from *H. salinarum* strain R1 (DSM 671). Details of 2D gel electrophoresis and mass spectrometric analysis procedures are provided elsewhere (Tebbe *et al.*, in preparation, Klein *et al.*, in preparation). Briefly, up to 800 spots from 2D gels are picked using the Bruker Spotpicker Proteineer SP. After tryptic digestion, peptide extracts together with the α -cyano-4-hydroxycinnamic acid matrix are applied automatically to a Bruker SCOUT384 MALDI target. Samples are then analysed by a MALDI-TOF mass spectrometer (Bruker Reflex III). Calibration and peak detection are done by the vendor's software packages.

Database Search

The identification of proteins is performed using the MASCOT search engine (which is locally installed) with the “probability-based MOWSE score algorithm”, and a database of 9312 ORFs (2784 protein-coding genes plus 6528 spurious ORFs) from the genome sequence of *H. salinarum* strain R1 (www.halolex.mpg.de). Experimental peptide fingerprint data are compared with those of a theoretical digest of all ORFs from this database.

For a database of 9312 ORFs, a score of 52 or higher is considered significant at the 5% level according to MASCOT. This would mean that only 5% of the ORFs with a score >52 are false positives. This prediction turned out to greatly underestimate the proportion of false positives under our experimental conditions. We use this value of 52 as a “significance cutoff” in our evaluation, as detailed below.

To minimize false positive identifications, a MOWSE score higher than 72 (20 above significance cutoff) is considered as identification and a MOWSE score of higher than 92 (40 above significance cutoff) is considered as reliable identification. We identified 6 false positive identifications with scores of 72, 74, 75, 84, 84, and 96. At least 9 (13%) of the 66 proteins with a score of 62-71 and at least 83 (49%) of the 168 proteins with a score of 52-61 are false positives. Therefore, the actual proportion of false positives occurring above a score of 52 is much higher than predicted by MASCOT (only 5% false positives are expected at a score of 52, 0.5% at 62, 0.05% at 72, and 0.0005% at 92). The MASCOT formula does not take into account peptides originating from self-digestion of trypsin and other frequently occurring contaminating peptides.

Due to the high rate of false positives, proteins with a MOWSE score below 72 are not considered to be identified in our automatic procedure, but may be promoted to the category of “manually identified” when several criteria are

fulfilled (for details see Tebbe *et al.*, in preparation). Currently, 56 proteins are in the category “manually identified”.

PSD spectra are submitted to the MASCOT search engine in sequence query mode using the same database of 9312 ORF’s from *H. salinarum*.

Algorithm

The toolbox reads the MASCOT result files and the underlying peaklists. It analyses the mass spectrum of the protein identified by MASCOT with the knowledge of its amino acid and nucleic acid sequence from the genome.

Theoretical tryptic peptides from the protein are assigned as matched or unmatched peptides, and mass peaks as matched or unmatched peaks.

Alternative peptide sequences (e.g. by reassigning the start codon) and alternative peptide masses (e.g. by considering post-translational modifications) are computed and compared with the masses of the unmatched peaks. If a new match is found, further analysis by PSD of this peak is requested if both, the intensity and the distance to the neighbour peaks, are suitable for PSD.

Results and discussion

In predicting the protein inventory of *H. salinarum* R1 (DSM 671) from the complete genome sequence for this strain (www.halolex.mpg.de), we encountered a severe ORF overprediction problem, evident from the fact that different gene finder programs gave results which are inconsistent with respect to ORF selection and start codon assignment. The genome after manual reannotation is considered to encode 2784 proteins covering 88 % of the genome (90% of the chromosome). However, it contains a total of 9312 ORFs, of which 8649 are longer than 100 codons. The longest of these 6528 spurious ORFs,

Fig1

OE3324A1F, has 1341 codons (Fig.1). In order to determine the protein inventory of this organism, proteomics methods were employed. Proteins were separated by 2D gels and proteins in individual spots were identified by MALDI-TOF peptide fingerprint analysis followed by identification using the MASCOT search engine. The MASCOT search is against a sequence database containing all predicted ORFs from the *H. salinarum* R1 genome (9312 entries). Start codon assignment is complicated by the fact that frames are open 5' of the start codon, one extreme example being the *nuoI* gene which codes for a protein of 153 amino acids. The start codon overlaps with the stop codon of the preceding *nuoH* gene (ATGA) in a manner typical for transcription units and thus can be considered as correctly assigned. Nevertheless, the reading frame of the *nuoI* gene is open for an additional 653 codons in front of its start codon and this extension completely overlaps with the *nuoH* gene and even partly with the

fused nuoCD gene, both of which have been reliably identified by MALDI-TOF peptide fingerprint analysis.

To ensure the reliability of identification, we developed a toolbox for analysis of MASCOT results. The reliability of identification can be increased by (a) increasing the number of matching peptides detected and (b) extraction of a sequence tag from an individual mass peak by PSD analysis. The toolbox identifies individual mass peaks relevant in this context. It proved to be useful for the following applications:

- Validation of low-score identifications
- Handling of problems that may be encountered when a spot contains more than one protein
- Detection of incomplete tryptic digestion
- Detection of posttranslational modifications
- Correction of start codon misassignments by gene finders
- Validation of the start codon assignment

The basic principle of the toolbox for detection of additional matching peptides is to compute the mass of alternate or modified tryptic peptides and to evaluate whether they correspond to those of as yet unmatched peaks in the spectrum. The toolbox highlights individual mass peaks relevant and suitable for further analysis and thus permits rational design of PSD experiments while the sample is still in the mass spectrometer.

The methods described below by specific examples are based on a data set of 745 identified proteins, of which 576 were reliably identified.

Validation of insecure identifications

We use the term “reliable identification” under very stringent conditions (MOWSE score higher than 92) in order to virtually exclude false positive identifications (see Materials and Methods).

As a consequence, many of our identifications remain insecure (169 proteins). In addition, false negatives can be promoted to the category of “manually identified” (56 proteins). To validate insecure or manual identifications, one of the matching peptides is subjected to further experimental analysis, for example

Fig 2 PSD analysis. Fig.2 shows the PSD spectrum of a mass peak of 1149.55 Da $[M+H]^+$, which unambiguously identifies the decapeptide QFAVDEDAVR from protein OE4459R (ribosomal protein L31.eR). Peaks represent the complete y series and part of the b series. The same peptide from this protein was identified by a MASCOT search in sequence query mode against all ORFs from *H. salinarum*.

Spots with more than one protein

One gel spot may contain more than one protein and MALDI-TOF allows the identification of all the constituent proteins. However, one mass peak may correspond to tryptic peptides from more than one protein. In 224 samples, more than one protein was identified. In 59 of these, at least one of the peptides could be assigned to more than one protein present in the spot. An example is given in

Fig3

Fig. 3. The mass peak of 1985.13 Da $[M+H]^+$ is ambiguous as it is consistent with the sequence LPIQDVYTISGIGTVPVGR from protein OE4721R (score 131, translation elongation factor aEF-1 alpha chain) and with the sequence ELQEVAIEAQDNIDEIR from OE3718F (score 90, cell division protein ftsZ3). PSD analysis assigned this peak to protein OE4721R (Fig. 3). There are 12 peaks representing b- and y-fragments. The same peptide from protein OE4721R was identified by a MASCOT search in sequence query mode against all ORFs from *H. salinarum*. No reasonable assignments were possible for the peptide from OE3718F.

Upon reanalysis of the peptide fingerprint data for protein OE3718F, omitting the peak at 1985.13 Da, the MOWSE score was lowered from 90 to 73.

Incomplete tryptic digestion

Missed trypsin cleavage sites are quite common and thus, the MASCOT search engine provides the user with the opportunity to set the maximum number of

missed cleavage sites. Increasing the number of missed cleavage sites results in a general decrease of MOWSE scores. Thus, we decided that missed cleavage sites should not be included in MASCOT calculations but should be dealt with by our toolbox. Unmatched peaks are examined to see whether their mass corresponds to that of two consecutive peptides. Potential missed cleavage sites were found in 72% of all mass spectra with an average of 2 matching

Fig4 incompletely cleaved peptides per spectrum. Fig. 4 shows the PSD spectrum of a peak of 1929.96 Da $[M+H]^+$ that demonstrates incomplete cleavage after Lys in the sequence DKYTYPDEFEPSSLGR in the ORF OE4330F (a probable phosphoesterase). There are 11 peaks representing b- and y-fragments. The same peptide from this protein was identified by a MASCOT search in sequence query mode against all ORFs from *H. salinarum*.

Correction of start codon assignments and validation of the N-terminus

Start codon assignment proved to be highly unreliable for the genome of *H. salinarum*. Therefore, the reliable identification of the N-terminus is an important part of the annotation. We routinely attempt to identify incorrect start codon assignments by integrating genomic and proteomic data. Alternative sequence versions are computed using all possible start codons and examined for the occurrence of new tryptic peptides which correspond to an unmatched

Fig5 peak. Fig.5 demonstrates correction of the start codon assignment for protein OE3710R (trkA domain protein trkA6). The previously unmatched peak at 1030.56 Da $[M+H]^+$ becomes an additional matching peak when the codon for Met-35 is assigned to be the start codon instead of the codon for Met-1 as initially predicted by two different gene finders (Glimmer, Delcher *et al.*, 1999, and Orpheus, Frishman *et al.*, 1998). The corresponding PSD spectrum assigned this peak to the sequence MDIVIVGAGR, which is the new N-terminal peptide of protein OE3710R (Fig.5). There are 12 peaks representing b- and y-fragments. The same peptide from this protein was identified by a MASCOT search in sequence query mode against all ORFs from *H. salinarum*.

It should be noted that this procedure can be generally applied to validate the N-terminus of identified proteins. In 20% of all spectra, a mass peak corresponding to the N-terminal peptide is visible with the N-terminal Met removed in half of the proteins.

Posttranslational modifications

We do not permit variable post-translational modifications in the MASCOT search as this results in a general decrease of the MOWSE score. Instead, all tryptic peptides of the candidate protein are computationally modified by a series of posttranslational modifications and compared to the unmatched peaks.

As a more complex procedure the search for a new terminus and posttranslational modification is combined.

No mass peak for the N-terminal peptide (MDPDLADDYISHPDQLQEAAEEALR) of OE4052F (DNA helicase mcm) was found. The toolbox identified a potential new N-terminus 48 codons upstream to the originally predicted start codon. There are two matching peptides in this 48-aa extension, already proving that the start codon has to be reassigned (data not shown). However, there was no match to the N-terminal peptide itself with or without the N-terminal methionine. The toolbox then computes the masses of posttranslationally modified, e.g. acetylated peptides and compares these to the masses of the unmatched peaks of the spectrum. The mass peak at 2099.96 Da $[M+H]^+$ would match under the assumption that the start codon is reassigned, the methionine is cleaved off, and the N-terminal alanine is acetylated (acetyl-AQAANQELVDQFEEFYR). This peak was highlighted by the program and suggested for validation by PSD.

Conclusions

Our toolbox successfully prompted experiments that led to the improved accuracy of protein identification by MALDI-TOF analysis. The additional information extractable from mass spectrometric experiments increased the number and the reliability of identifications but also reduced false positive peak

assignments. Additional biological information can be obtained (e.g. detection of post-translational modification), corrected (e. g. reassignment of start codon) or validated. Although this toolbox was developed for and applied to *H. salinarum*, it can be used to improve the analysis of proteome data from any organism.

References

- Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics, *Nature*, **422**, 198-207.
- Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER, *Nucleic Acids Res.*, **27**, 4636-4641.
- Frishman, D., Mironov, A., Mewes, H.W. and Gelfand, M. (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes, *Nucleic Acids Res.* **26**, 2941-2947.
- Klein, C., Garcia-Rizo, C., Zischka, H., Siedler, F., Pfeiffer, F. and Oesterhelt, D. (2003) The membrane proteome of *Halobacterium salinarum*, in preparation.

Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags, *Anal. Chem.*, **66**, 4390-4399.

Pappin, D.J.C., Hojrup, P. and Bleasby, A.J., (1993) Rapid identification of proteins by peptide-mass fingerprinting, *Curr. Biol.*, **3**, 327-332.

Perkins, D.N., Pappin, D.J.C., Creasy, D.M. and Cottrell, J.S., (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, **20**, 3551-3567.

Rapp, U., Resemann, A. and Suckau, D. (2000) Detection limits of MALDI-TOF PSD for peptide sequencing from protein digests, *Bruker Daltonik GmbH, Bremen, Germany*. Application note #MT-52.

Roepstorff, P. and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides, *Biomed. Mass Spectrometry*, **11**, 601-601.

Talbo, G.H., Suckau, D., Malkoski, M. and Reynolds, E.C. (2001) MALDI-PSD-MS analysis of the phosphorylation sites of caseinomacropeptide, *Peptides*, **22**, 1093-1098.

Tebbe, A., Klein, C., Bisle, B., Siedler, F., Scheffer, B., Garcia-Rizo, C., Wolfertz, J., Hickmann, V., Pfeiffer, F. and Oesterhelt, D. (2003) Analysis of the cytosolic proteome of *Halobacterium salinarum*, in preparation.

Fig.1

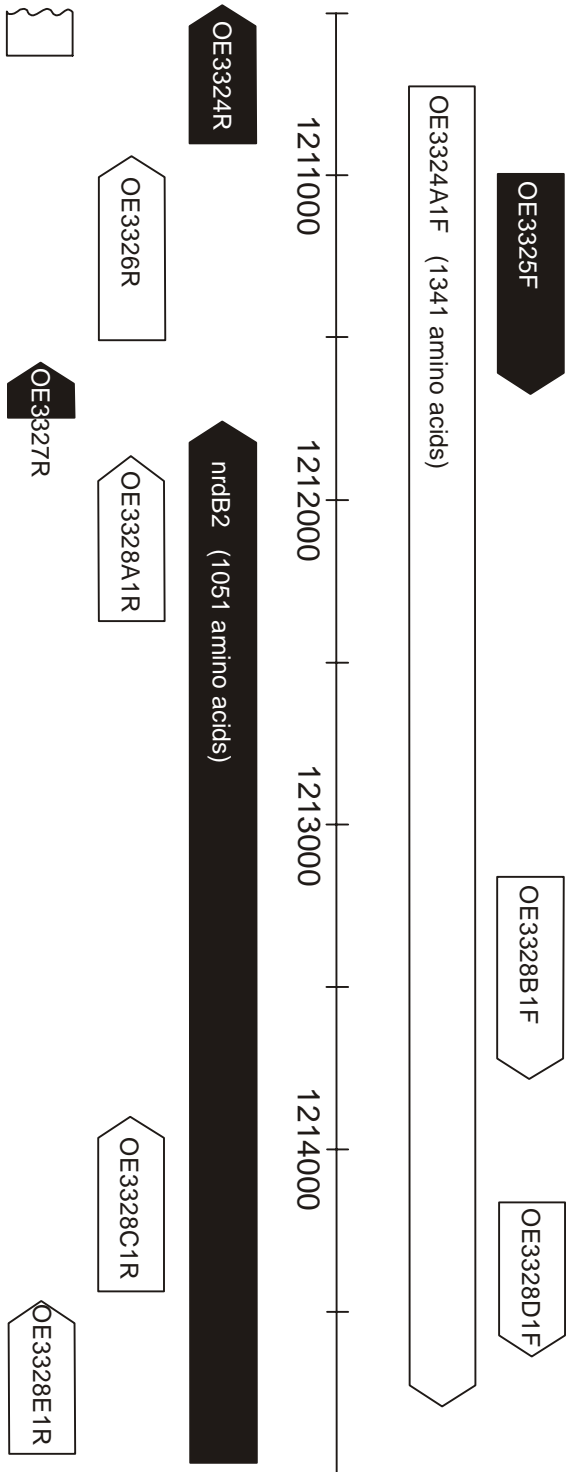


Fig.2

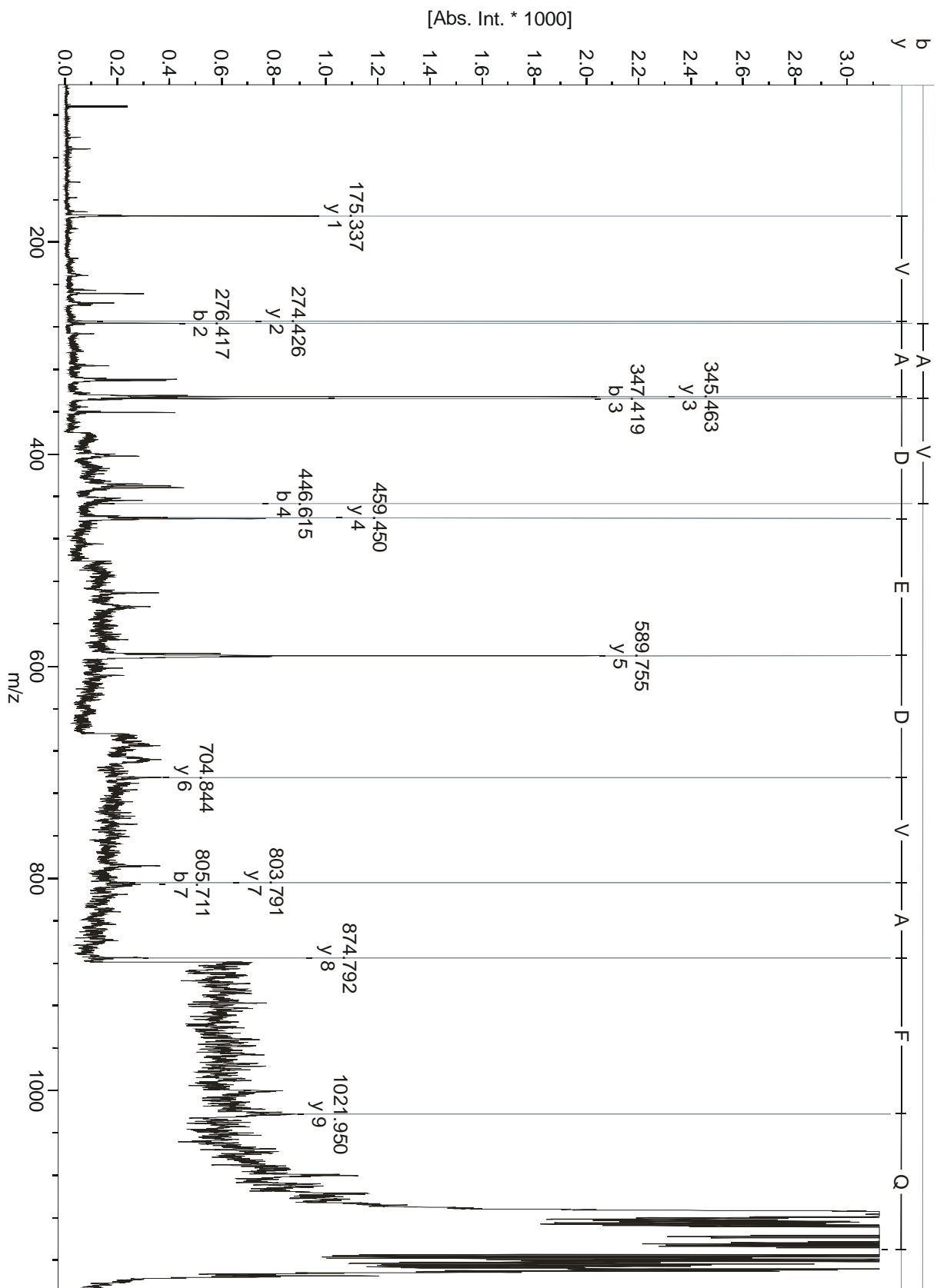


Fig 3a

●**OE4721R Mass:45622 Score:131** translation elongation factor aEF-1 alpha chain

Mr(exp)	Mr(calc)	Delta	Peptide
886.47	886.47	0.00	EHVFLSR
1383.72	1383.80	-0.08	TLGIDELIVAVNK
1521.65	1521.71	-0.06	DLFGQVGFNPDDAK
1523.70	1523.80	-0.10	HQNLAVIGHVDHGK
1846.81	1846.82	-0.01	GGFEFAYVMDNLAEER
1984.12	1984.10	-0.02	LPIQDVYTTISGIGTVPVGR
2590.21	2590.19	0.02	TIEMHHEEVPNAEPGDNVGFNVR
2683.31	2683.32	-0.01	NMITGASQADNAVLVVAADDGVAPQTR
3149.42	3149.45	0.03	GVTIDIAHQEFTTDEYEFTIVDCPGHR

●**OE3718F Mass:42083 Score:90** cell division protein ftsZ3

Mr(exp)	Mr(calc)	Deltas	Peptide
883.45	883.46	-0.01	SFQTFVR
1402.64	1402.69	-0.05	LTLPCEIEGSR
1622.72	1622.77	-0.05	ADLMGLEHIPEENR
1704.77	1704.81	-0.04	EVDNLLVFDNDAWR
1984.12	1983.97	0.15	ELQEVAIEAQDNIDEIR
2081.01	2080.95	0.06	SGESVQGGYDEINEEIVTR
2643.32	2643.31	0.01	GGDYPVSNSDQVASVLLSGVNDVPR

Fig 3b

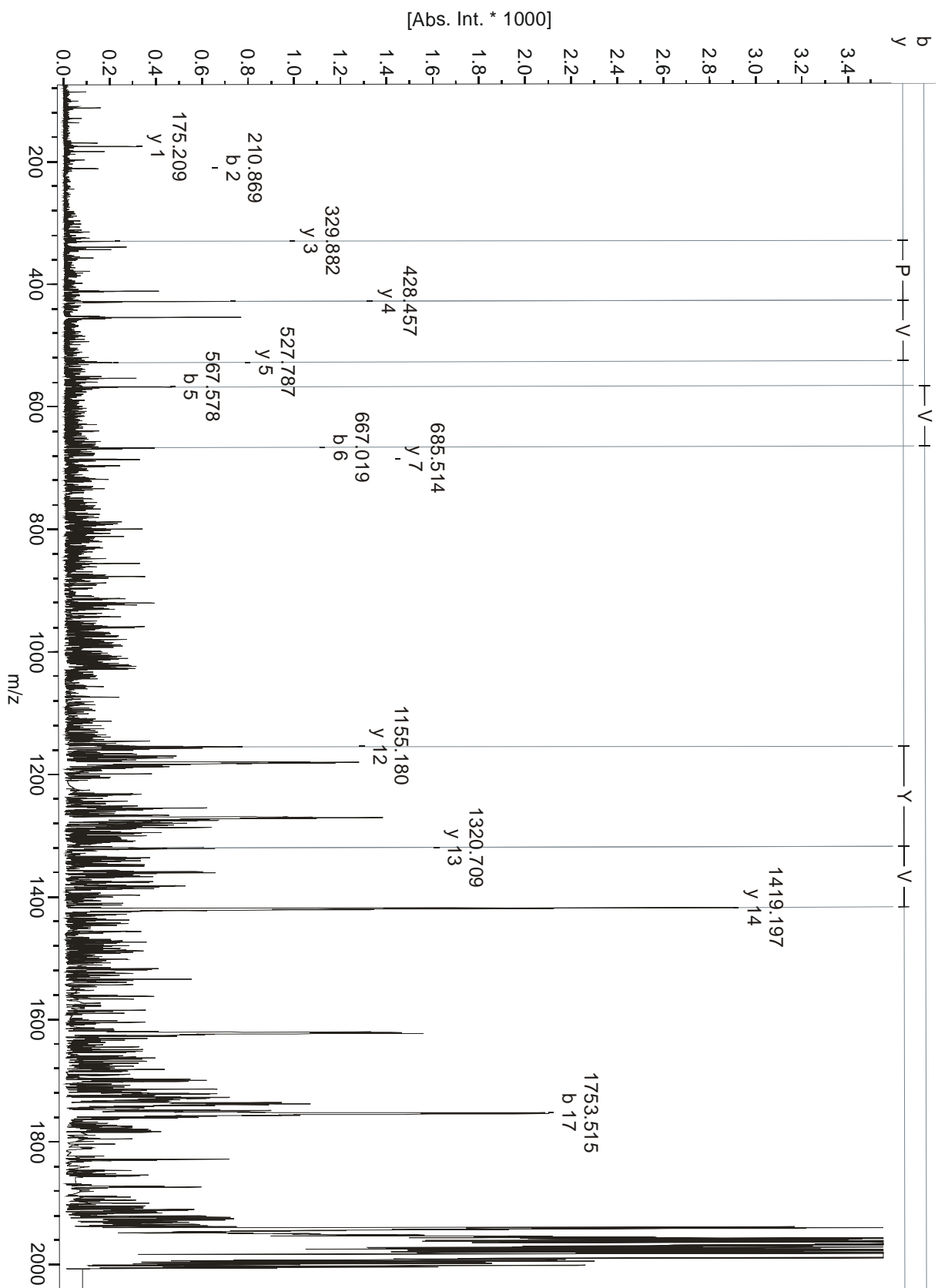


Fig.4

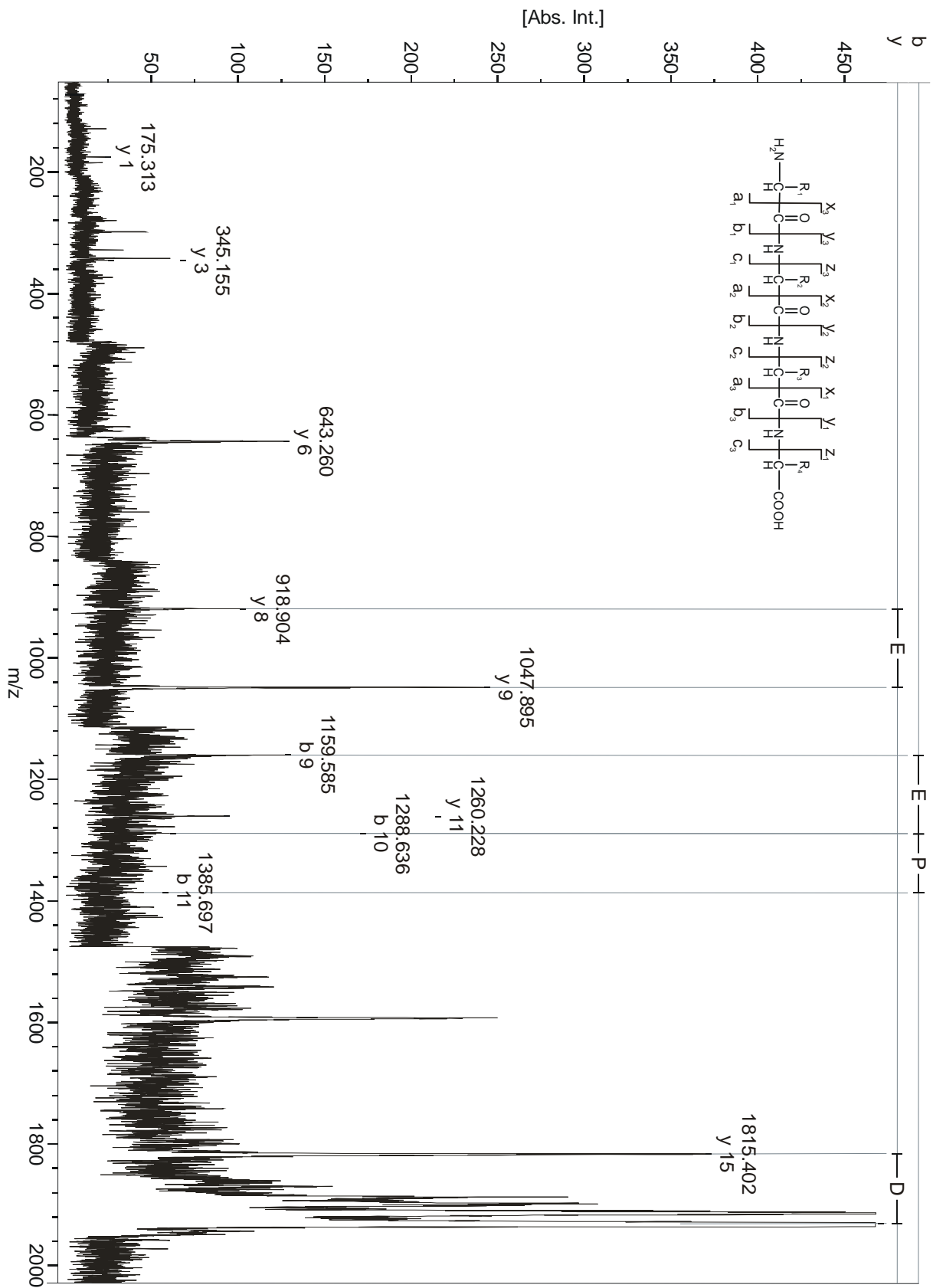


Fig.5a

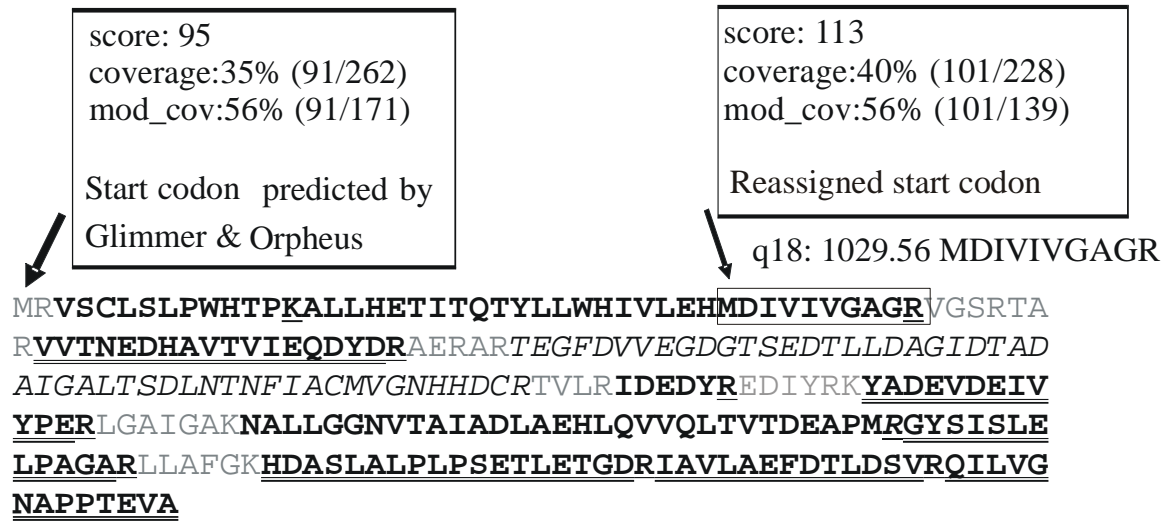


Fig.5b

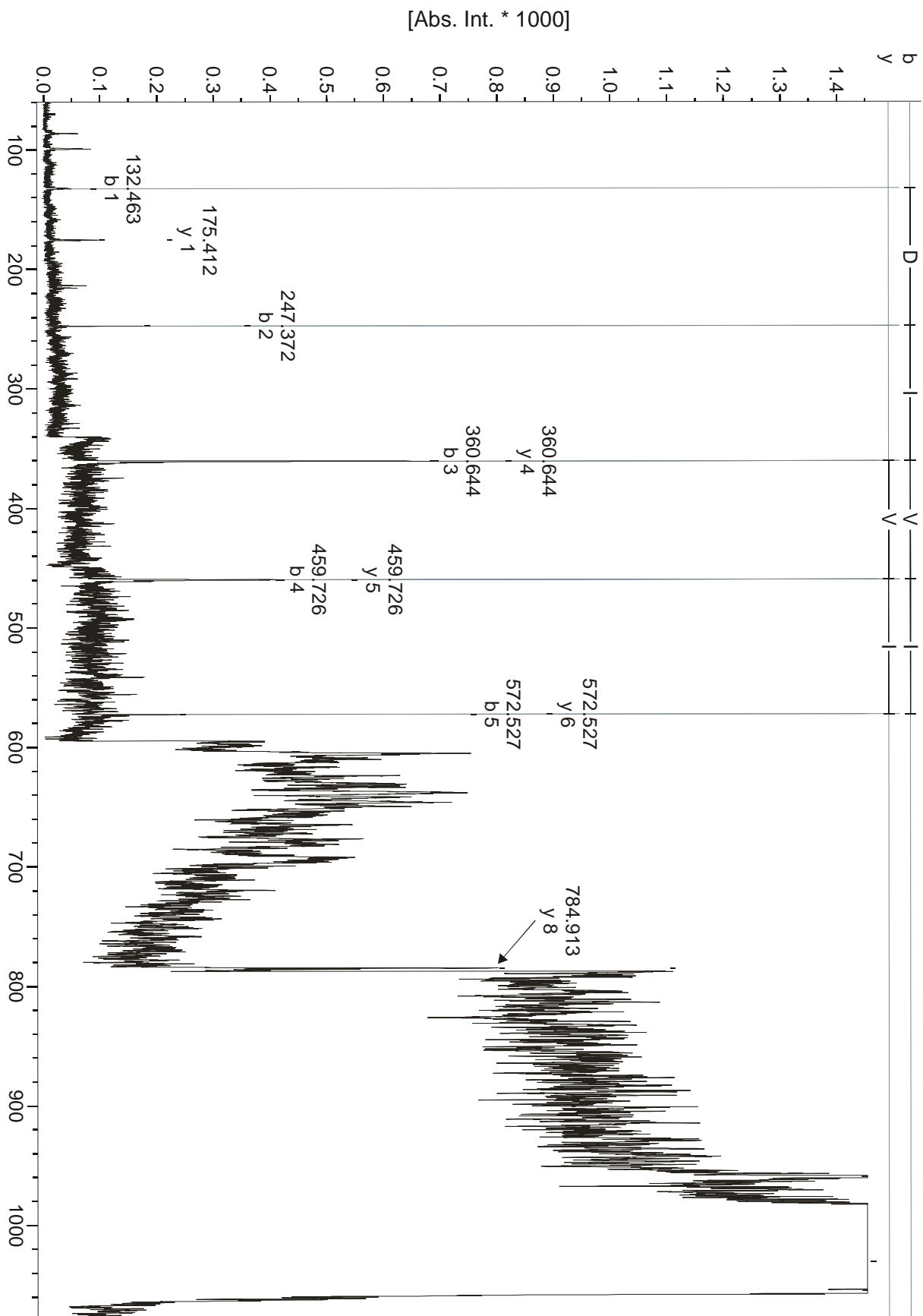


Figure legends

Fig.1 A stretch of DNA sequence from *H. salinarum* strain R1 around position 1,213,000 showing genes encoding real proteins in black and spurious ORFs in white. The longest spurious ORF, OE3324A1F, has 1341 codons. The NrdB2 protein (1051 amino acids) is encoded on the opposite strand and has been reliably identified.

Fig.2 The PSD spectrum of the mass peak of 1149.55 Da $[M+H]^+$ with annotation of the most relevant peaks, i.e. those matching b- and y-fragments with a tolerance of 1 Da. The peak is assigned to the peptide QFAVDEDAVR from protein OE4459R. There are 13 peaks representing b- and y-fragments and an additional 13 peaks representing a-fragments or b- and y-fragments with loss of ammonia (-17 Da). The nomenclature for sequence ions (Roepstorff and Fohlman, 1984) is illustrated in the inset to Fig.4.

Fig.3 (A) Two distinct proteins are identified by a MASCOT search in peptide fingerprint mode from one gel spot: OE4721R with score 131 and OE3718F with score 90. The peak of 1985.13 Da $[M+H]^+$ (1984.12 Da $[M]$) matches to one peptide from each of the two proteins. (B) The PSD spectrum of the mass peak of 1985.13 Da $[M+H]^+$ with annotation of the most relevant peaks (see

legend to Fig.2). The ambiguous matching peak of 1985.13 Da $[M+H]^+$ is assigned to the peptide LPIQDVYTISGIGTVPVGR from protein OE4721R. There are 12 peaks representing b- and y-fragments and an additional 7 peaks representing a-fragments or b- and y-fragments with loss of ammonia.

Fig.4 The PSD spectrum of the mass peak of 1929.96 Da $[M+H]^+$ with annotation of the most relevant peaks (see legend to Fig.2). The peak is assigned to the peptide DKYTYPDEFEPSELLGR from protein OE4330F, which has an internal, uncleaved Lys residue. There are 10 peaks representing b- and y-fragments and an additional 5 peaks representing a-fragments or b- and y-fragments with loss of ammonia. The inset illustrates the nomenclature for sequence ions (Roepstorff and Fohlman, 1984).

Fig. 5 Correction of the start codon assignment for the protein OE3710R. (A) Representation of the protein sequence as originally predicted by the two ORF finders Glimmer and Orpheus. Peptides are coded according to identification by mass spectrometry. Bold double underlined: 6 matched peptides [800-4000Da], bold not underlined: 4 unmatched peptides, cursive: 1 peptide >4000 Da, grey: several peptides <800 Da. The Lys and Arg residues of matched and unmatched peptides are singly underlined. The newly assigned N-terminus is boxed and the mass of the previously unmatched peptide of 1030.56 Da $[M+H]^+$ (1029.56 Da

[M]) is indicated. (B) The PSD spectrum of the mass peak of 1030.56 Da $[M+H]^+$ with annotation of the most relevant peaks (see legend to Fig.2). The peak is assigned to the peptide MDIVIVGAGR from protein OE3710R. There are 10 peaks representing b- and y-fragments and an additional 5 peaks representing a-fragments or b- and y-fragments with loss of ammonia.