Journal of Proteomics & Bioinformatics - Open Access
Research Article JPB/Vol.2/February 2009

Medaka Proteome Study: Use of Cross-species Matching and Expressed Sequence Tag Data for Protein Identification

Mezhoud Karim[1*], Praseuth Danièle[2], Marie Arul[3,4], Puiseux-Dao Simone[1] and Edery Marc[1]

[1]FRE 3206 MNHN-CNRS, Cyanobactéries, cyanotoxines et environnement, Muséum national d'Histoire naturelle, 57 rue Cuvier, F-75231, Paris Cedex 05, France
[2]CNRS, UMR5153; Inserm, U565: USM0503, Muséum national d'Histoire naturelle, 57 rue Cuvier, F-75231, Paris Cedex 05, France
[3]Plateforme de Spectrométrie de Masse et de Protéomique, Département Régulation Développement et Diversité Moléculaire, Muséum national d'Histoire naturelle, 63, rue Buffon, F-75231, Paris Cedex 05, France
[4] Molécules de Communication et Adaptation des Micro-Organismes, FRE 3206 CNRS, Paris, F-75005 France

*Corresponding author: Dr. Karim Mezhoud, FRE 3206 MNHN-CNRS, Cyanobactéries, cyanotoxines et environnement, Muséum national d'Histoire naturelle, 57 rue Cuvier, F-75231, Paris Cedex 05, France, Tel: (33) 1 40 79 32 12; Fax: (33) 1 40 79 35 94; E-mail: kmezhoud@mnhn.fr

Received January 06, 2009; Accepted February 09, 2009; Published February 20, 2009

Citation: Mezhoud K, Praseuth D, Marie A, Puiseux-Dao S, Edery M (2009) Medaka Proteome Study: Use of Cross-species Matching and Expressed Sequence Tag Data for Protein Identification. J Proteomics Bioinform 2: 067-077. doi:10.4172/jpb.1000063

Abstract

Proteomics aims to understand gene function and molecular processes of the living cell through a large-scale study of the expressed proteins. Although proteomics approach is largely applied on experimental models such as rodents, other animal models such as medaka fish (*Oryzias latipes*) also attracts considerable interest in proteomics field. Medaka is one of the most studied animal model in reproductive or developmental biology. Although, its genome has been sequenced, only a few proteins are available in the databases.

We present the procedure used in one experiment as a model of identification of proteins, in this case from liver of medaka treated by a hepatotoxic cyanotoxin, the microcystin.

Because *O. latipes* is not listed in mascot search engine, the identification of proteins (selected because modified by the treatment) was done on closed related species. First, 15 spots were analyzed using peptide mass fingerprinting (PMF). However none could be reliably identified using mascot engine. The reason for low sequence coverage is due to the fact that the outcome of the research depends on the completness of the protein database obtained from NCBInr and the availability of cross-species for protein identification. In order to improve the identification, the PMF search was combined with a search in a medaka nucleotide specific database and confirmed by MS/MS ion searches which revealed successful. This identification procedure has been successfully applied in different experiments with the medaka model.

Keywords: Bioinformatics; Databases; Mass spectrometry; Medaka; Protein identification; Proteomics

Abbreviations

Blast: Basic Local Alignment Search Tool; Blastp: Blast for protein database; EBI: European Bioinformatics Institute; EST: Expressed sequence tag; i.d.: Identify; IDA: Information-dependent acquisition; MS: Mass Spectrometry; MS/MS: tandem mass spectrometry; NCBI: National Centre of Biological Information; NCBInr: NCBI nonredundant protein database; NCBI-EST: NCBI Expressed Sequence Tag database; PAGE: Polyacrylamide gel electrophoresis; PMF: Peptide Mass Fingerprinting; rpsblast: Reverse-position specific Blast (search conserved domains on a protein); tblastn: translate blast nucleotide; 2-D: Two dimensional; 2-DE: Two dimensional electrophoresis

Introduction

Proteomics is aiming at understanding gene function and at characterizing molecular processes of the living cell through a large-scale study of proteins demonstrated as more or less expressed or post-transcriptionally modified in a given biological context. In such studies, protein identification represents a key step. In toxicology proteomics is largely in use but on a limited number of experimental models such as rodents. However other animal models are more and more experimented: the fish medaka (*Oryzias latipes*) is one of the most studied in reproductive or developmental toxicology for example (Wittbrodt et al., 2002; Wester et al., 2004). But if its genome is sequenced (Kasahara et al., 2007), only a few proteins have been characterized in databases which restricts proteomics studies (Henrich et al., 2003).

Such studies require the combination of two-dimensional gel electrophoresis (2-DE) and mass spectrometry (MS) technology, in association with protein or nucleotide database search. Two MS identification strategies are possible. Peptide mass-fingerprinting (PMF) has become a high-throughput method that is easy to handle and rapid; however, the results largely depend on the range of proteins from various species listed in the database searched. A more sophisticated technology is based on the use of sequence tags for peptide-fragment information and subsequent matching to the nucleotide database using Mascot engine (Perkins et al., 1999). If protein sequences of the specific organism under study are not available, resorting to cross-species protein identification is suitable (Cordwell et al., 1995). Depending on the organisms studied, the possibility and reliability of cross-species spot identification through PMF-matching vary greatly. Considering peptide masses, the results of cross-species identification yield low scores, with limited sequence coverage. Consequently, applying tandem mass spectrometry (MS/MS) becomes useful and permits identification of the amino acid sequences; in this case, matching the data with that in the nucleotide database (NCBI-EST) generally produces more results than the theoretical peptide-mass database (NCBInr). Moreover, modern sequencing technologies and projects generate a large quantity of DNA sequences at high speeds. Medaka is one of the interesting organisms whose genome sequence has been published (Kasahara et al., 2007); nevertheless, its proteome has not yet been characterised. Large-scale *in situ* hybridisation screenings have been carried out, and the expression data sorted in the databases are accessible through web interfaces. The medaka fish home page has grouped the major medaka genome-sequencing projects (http://biol1.bio.nagoya-u.ac.jp:8000/). About 527,715 nucleotides and 1,933 proteins have been characterised in the European Bioinformatics Institute (http://www.ebi.ac.uk/) and the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) projects (as of December 2008). However, compared to a few other species, the nucleotide and protein databases of *Oryzias latipes* are still incomplete in public databases. Therefore, the finishing step and annotation process is time-consuming. It has been proved in this article that medaka protein identification using PMF and orthologous proteins can yield valuable information, but the procedure needs to be completed using tandem mass technology. These results could be used as guidelines for determining the feasibility of the shotgun proteomics approach using medaka fish as an experimental model.

Methods

2-dimensional polyacrylamide gel electrophoresis and staining

Freshly isolated fish livers were fractionated into cytosolic, membranous/organellar, nuclear and cytoskeletal fractions using the Subcellular proteome extraction kit from Calbiochem, according to the manufacturers instructions. For sodium laurylsulfate PAGE (SDS-PAGE), approximately 20 µg proteins of the different subcellular fractions were electrophoresed in 12% acrylamide gels according to Laemmli (1970). For two dimension PAGE (2D-PAGE), the samples were further treated with trichloroacetic acid precipitation so as to remove salts. The electrofocusing

was then performed using commercial IPG strips (Immobiline dry strips, 7 cm, pH 4-7 non linear, GE Healthcare, England). The strips were passively rehydrated overnight with 200 ìg proteins in solution in 125 µ L 2 M thiourea, 6 M urea and 2 % CHAPS (w/v). The isoelectrofocusing was then performed at 16°C with the following potential ramp program: 0-50 V in 2 min, 50 V for 30 min, 50-200 V in 15 min, 200 V for 30 min, 200-2000 V in 30 min, 2000 V for 3 hours (maximum current per strip: 25 mA). Prior to the second dimension, the IPG strips were equilibrated twice for 20 min with gentle shaking in 8 mL SDS-containing equilibration buffer [50 mM Tris-HCl, pH 8.8, 6 M urea, 30 % glycerol (v/v), 2 % SDS (w/v) and traces of bromophenol blue]. DTT (1 % final, w/v) was added at the first equilibration step. Iodoacetamide (2.5 % final w/v) was added to the equilibration buffer at the second step. The strips were loaded on top of a 1 mm-thick 11 % acrylamide SDS-PAGE gel, along with molecular weight markers (Molecular Probes). Each 2-D gel was stained with SYPRO Ruby stain (Molecular Probes) and was imaged on Typhoon 9410 (GE Healthcare, England) apparatus.

Protein chemistry and sample preparation

2-D PAGE protein spots of interest were excised and washed thrice in 25 mM NH4HCO3 for 10 min and dehydrated in 100 % acetonitrile (ACN) for 5 min. Acetonitrile was evaporated under vacuum and the gel pieces were rehydrated with 15 µL of a 12.5 ng/µL trypsin solution (25 mM NH4HCO3, 5 mM CaCl2; trypsin from Promega) on ice for 45 min. The excess liquid was removed and the gel pieces were incubated overnight at 30°C in 20 µL 25 mM NH4HCO3. The resulting peptide mixture was extracted from the gel by sonication. Desalting the peptides was performed either using Zip-Tips (Millipore). Elution of the desalted peptides was with 80 % ACN and the peptide mixture was vacuum-dried. Peptides were redissolved in 2–5 µL 1 % formic acid.

MS techniques

Matrix-assisted laser desorption/ionisation time of flight MS

Peptidic mixtures (2 µL) were mixed with an equal volume of a saturated solution of α-cyano-4-hydroxy-cinnamic acid (Sigma) in 50% acetonitrile (ACN), 0.05% trifluoroacetic acid. The new mixtures were allowed to co-crystallise on the matrix-assisted laser desorption/ionisation (MALDI) target plate (dried-droplet method). Mass spectra were recorded in the positive ion mode on a MALDI time of flight (MALDI-TOF) mass spectrometer (Voyager-DE Pro, Applied Biosystems). Mass/charge (m/z) ratios were measured in the reflection/delayed extraction mode, with an accelerating voltage 20 kV, grid voltage 150 V, 64.5% guide wire voltage and low mass gate 600. Protein identification using PMF was conducted using the MASCOT search engine (http://www.matrix-science.com). The search parameters were defined as follows: NCBInr database, all taxa, trypsin digest (allowed up to 1 missed cleavage), cystein optionally modified by carbamidomethylation, methionine optionally modified by oxidation, monoisotopic, positive mode, maximal mass tolerance of 0.1–0.5 Da.

Electrospray-ionisation quadrupole -TOF MS

Peptide mixture was mixed with 90% ACN, 10% methanol. Nano-electrospray-ionisation (Nano-ESI) experiments were carried out using a hybrid quadrupole time-of-flight (Q-TOF) mass spectrometer (Pulsar i, Applied Biosystems), equipped with a nano-ESI source. About 2 µL of the digest were loaded into the nanocapillary; MS and MS/MS data were acquired using an information-dependent acquisition (IDA) method with the following settings: charge-state selection from $[M+2H]^{2+}$ to $[M+4H]^{4+}$; intensity threshold: 5 counts; the collision energy setting was automatically determined by the IDA method based on the charge state of each precursor ion. Following the IDA-based gas-phase fragmentation data acquisition, precursor ions were excluded for 60 s. The acquired data were formatted into Mascot-compatible data using the mascot.dll script in Analyst QS v1.1 software shipped with the mass spectrometer. A no redundant protein database was initially searched (NCBInr). If the results were inconclusive, a nucleic acid database was searched by choosing expressed sequence tag (EST)-others from the master result page. The search was queried using the acquired mass-spectral data with the following settings: taxonomy: all taxa; trypsin digest (allowed up to 1 missed cleavage); cysteinyl residues optionally modified by carbamidomethylation; methionyl residues optionally modified by oxidation; maximal mass tolerance for precursor and fragment ions: 100 ppm.

Sequence search and medaka nucleotide databases

Database search was carried out by feeding the MS data to the Mascot search engine (http://www

matrixscience.com/). The search was conducted for all Taxa. The returned results (generally in orthologous species) were matched to specific medaka EST database. Two databases were used: the medaka gene-expression pattern database (Henrich et al., 2003) and the medaka EST database (http://medaka.lab.nig.ac.jp/) (see results for more details).

Results

Proteins from liver medaka fishes were resolved by 2D gel technology and staining. Spots were then selected for protein identification because modified by the treatment (Figure 1). In the first step, fifteen sample spots were subjected to MALDI-TOF MS analysis and PMF search. Since the protein sequences of medaka fish are not yet characterised and *Oryzias latipes* is a species that is not listed in the database (taxonomy field) using the available search engine, the search was conducted without taxonomy restriction.

Most of the samples could not be identified by PMF due to the poor sequence coverage in spite of a significant number of peptides in the spectrum (Figure 2). Four spots with a modest score level defined by Mascot could be identified by cross-species matching (Table 1, Id. 1, 2, 3, 4).

The matching peptides were then searched by off-line translate-blast nucleotide (tblastn) algorithm, using Expressed Pattern Database (MEPD, Germany, http://ani.embl.de:8080/mepd/) and/or medaka EST database (Japan, http://medaka.lab.nig.ac.jp) (Figure 3). The output of the search was a list of clones from *Oryzias latipes*. The identification of the possible corresponding proteins was considered complete only if the three following conditions were verified for each clone: (1) the clone had a reasonable length; (2) the number of matched peptides was significant; (3) the virtual protein sequence obtained by translation of the clone had a computed mass compatible with the apparent mass observed on the gel. For each clone, an alignment of

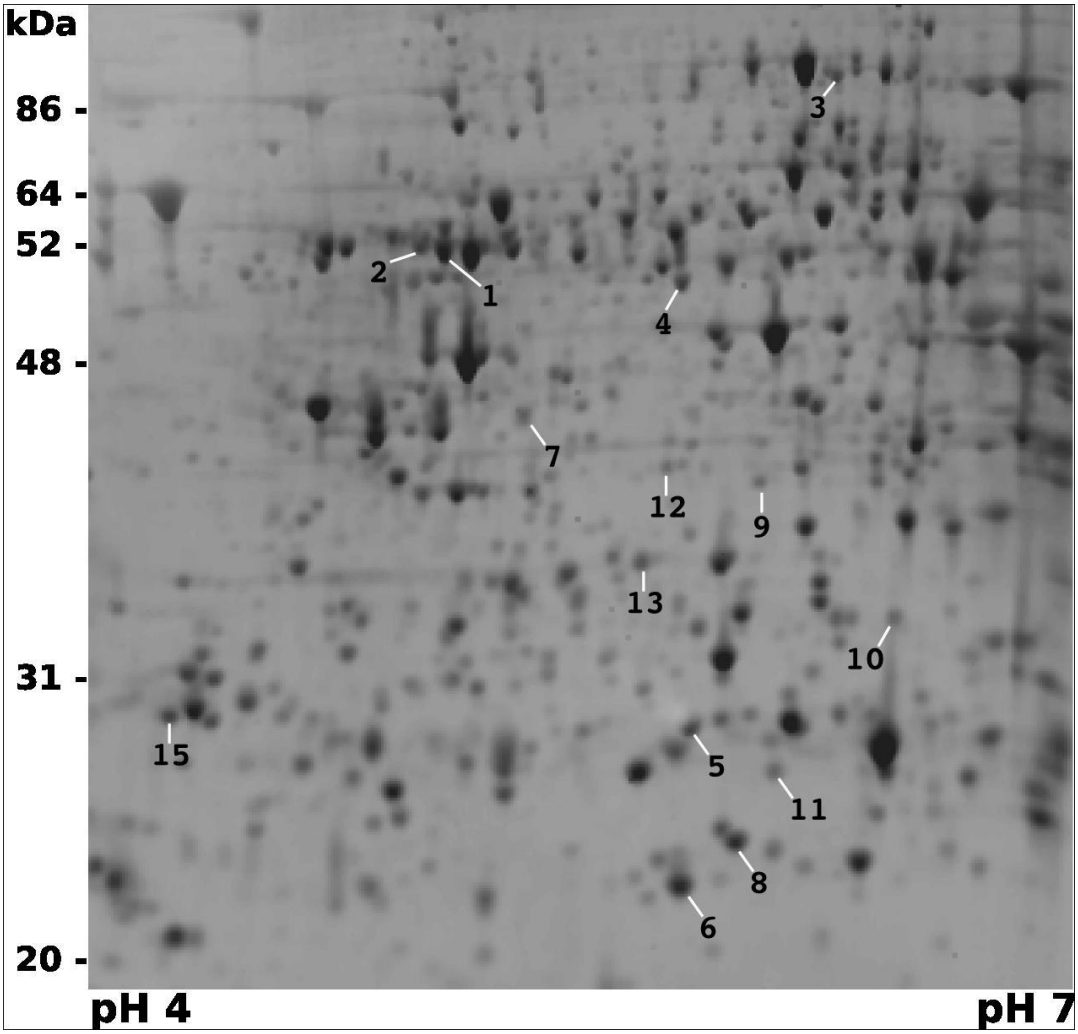

Figure 1: 2-DE gel of medaka hepatocyte cytosolic fraction stained with Sypro Ruby.

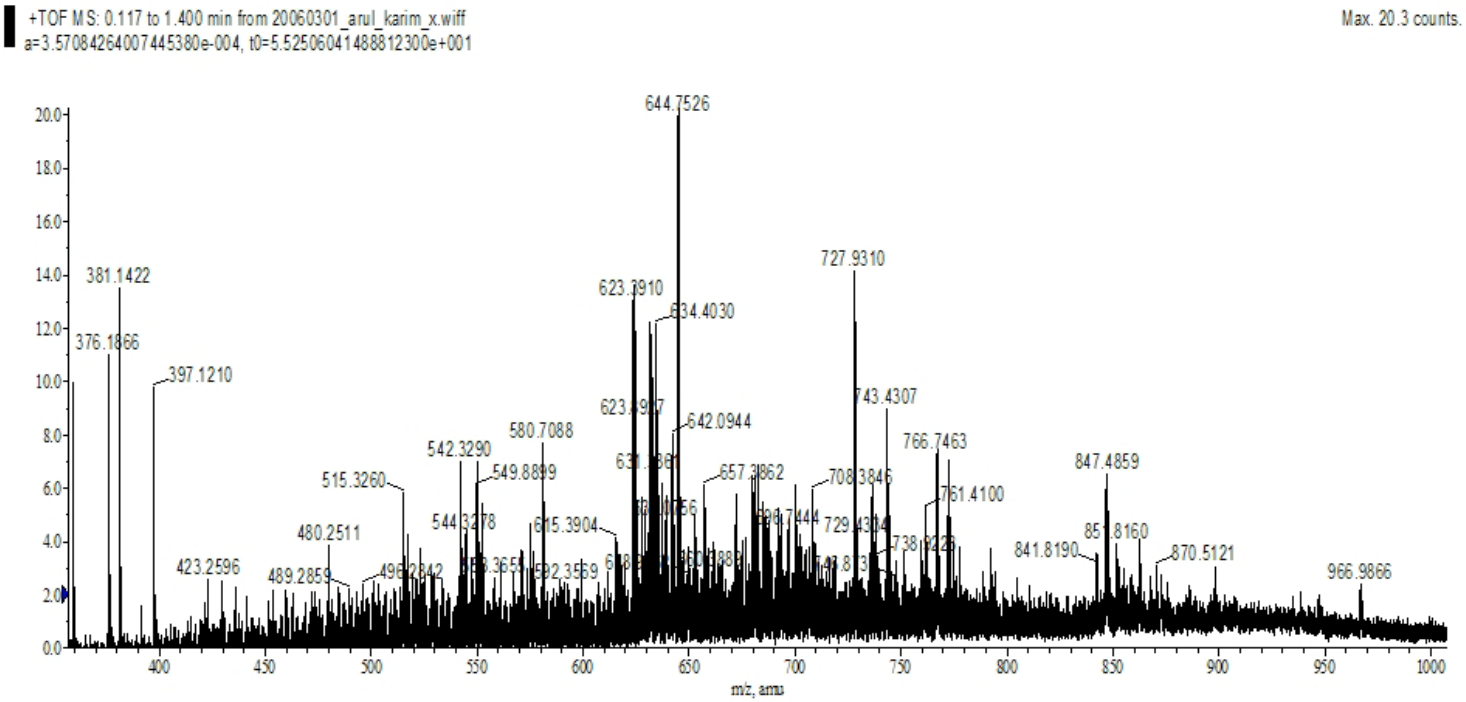

Figure 2: MALDI-TOF MS spectra of mono protonated ions derived from spot 3 (Selenium binding protein 1 or SeBP) in Figure 1.

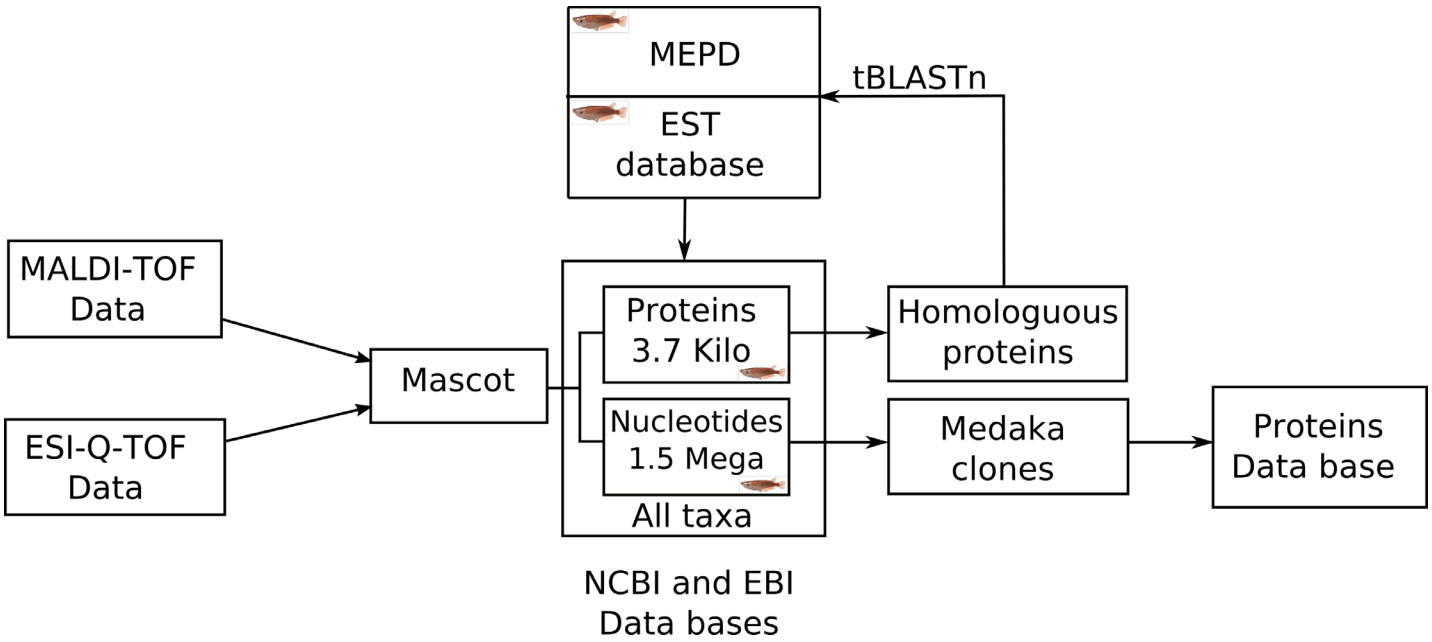

Figure 3: Protein-identification procedure.

Spot Id.	Protein identity	Cov.% Score	Matched peptide sequence
1	Phenylalanine hydroxylase		
	32442452 (nr, *D. rerio*)	25[a]	13/86 peptides
	BJ911084(MEPD, *O. latipes*)	76/79	
2	Phenylalanine hydroxylase		
	32442452 (nr, *D. rerio*)	25[a]	12/86 peptides
	BJ737725 (MEPD, *O. latipes*)	74/79	
3	Selenium binding protein 1		
	XP_707845 (nr, *D. rerio*)	28[a]	14/56 peptides
	BJ007475 (MEPD, *O. latipes*)	28/72	
	AAH5690(nr, *D. rerio*)	5[b]	EEIVYLPCIYR;LILPSLISSR
	BJ007475 (MEPD,*O. latipes*)	43/44	MVEPVEVLWK
	BJ007475 (est, *O. latipes*)	26[b]	STGILKPDYLATVDVDPK
		144/68	EEIVYLPCIYR; IYVIDVGTDPRAPK
4	Keratin 18 type 1		
	CAA74664 (nr, *O. mykiss*)	23[a]	10/67 peptides
	BJ498018 (MEPD, *O. latipes*)	40/54	
	AAC38007 (nr, *C. auratus*)	14[b]	LQDALEEQK; MAMQNLNDR; VMTVTQTLVDGK
	BJ747804 (MEPD, *O. latipes*)	137/54	
5	Unidentified		

Spot Id.	Protein identity	Cov.% Score	Matched peptide sequence
6	DJ-1 BAD67176 (nr, *O. latipes*) Unnamed protein (RKIP) CAG08164 (nr, *T. nigroviridis*) BJ747456 (MEPD,*O. latipes*) Raf kinase inhibitor protein AU170544 (est, *O. latipes*)	 30[b] 126/50 14[b] 126/50 16[b] 254/61	NVVICPDTSLEEASK GAEEMETVIPVDVMRR QGPYDVVLLPGGMPGAQNLAESPAVK LYEQLAGK; LYTLALTDPDAPSR YGSLEIDELGK; GNDVSSGCVLSDYVGSGPPK; LYTLALTDPDAPSR
7	Acidic ribosomal phosphoprotein P0 AAP20211 (nr, *P. major*) BJ003458 (MEPD, *O. latipes*)	 14[b] 223/49	 IIQLLDDYPK;GHLENNPALEK TSFFQALGITTK DLLLANKVPAAAR
8	Natural killer enhancing factor AAY25400 (nr, *P. olivaceus*)	 27[b]	 SVEETLRL

	BJ714211 (MEPD, *O. latipes*) RahpC-TCA family AV669883 (est, *O. latipes*)	172/47 11[b] 81/65	AVMPDGQFK;QITINDLPVGR GLFIIDDKGVL ; TISTDYGVLKEDEGIAYR IGSLAPDFTAK TISTDYGVLKEDEGIAYR
9	Unidentified		
Spot Id.	Protein identity	Cov.% Score	Matched peptide sequence
10	Proteasome α type 1 AAP20159 (nr, *P. major*) BJ01318 (est, *O. latipes*)	 25[b] 141/49 28[b] 167/62	LVSLIGSK;QGSATVGLK IHQIEYAMEAVK; ALRETLPAEQDLTTK Same 4 peptides + FVFDRPLPTSR
11	Hypoxanthine guanine		

	phosphoribosyl transferase CAA35648 (nr, *C. longicaudatus*) BJ729370(MEPD, *O. latipes*)	 10[b] 62/50	 VIGGDDLSTLTGK SIPMTVDFIR
12	F-actin capping protein A AAR16282 (nr, *T. rubripes*) AM151914 (MEPD, *O. latipes*)	 3[b] 84/51	 ILLNNDNLLR
13	Actin capping protein B subunit AAA52222 (nr, *G. Gallus*) BJ729321 (MEPD, *O. latipes*)	 15[b] 208/47	TGSGTMNLGGSLTR LVEDMENKIR TKDIVNGLR STLNEIYFGK
14	Enolase AAA70080 (nr, *S. pombe*) BJ722994 (MEPD,*O. latipes*)	 15[b]	 IEEELGSR IEEELGDK
15	14-3-3 protein zeta/delta (RKIP-1) P29361 (nr, O. aries) BJ727482 (MEPD, O. latipes)	 5[b] 84/50	 SVTEQGAELSNEER

[a]: MALDI-TOF identification

[b]: LC-ESI-Q-TOF identification

Table 1: Peptide sequences of identified proteins.

its polypeptidic sequence was carried out against the protein sequence initially obtained through cross-species identification to compare both protein sequences. To obtain additional information, tandem MS/MS was carried out. However information on *O. latipes* proteins is poorly available in public databases, and their homologies to proteins from other teleostean species whose genomes are available (*D. rerio, O. mykiss and Tetraodon fugu*) are insufficient. Then, online matching of the data spectrum was carried out with the NCBI-EST database. The remaining proteins were identified through approaches previously described (Altschul et al., 1989). In this case, one or multiple clones coding for the sequenced peptides were selected, and the corresponding polypeptide was reconstructed. If the computed mass of this polypeptide was compatible with the apparent mass observed on the gel, this specific sequence was used for a blastp/rpsblast search in NCBInr (Altschul, 1989) to confirm that the reconstituted sequence and the initial protein were homologous and could have the same putative function.

Discussion

This study was undertaken to establish a strategy to interrogate the databases in order to characterize proteins from the medaka fish by proteomic approach. The mass of the peptides (during MS scan) and their corresponding fragments (during MS/MS experiments) were used to search the selected database.

The feasibility of cross-species proteome protein identification using PMF approaches has been studied by relying on conservation of some of the genes and associated protein sequences. In general, the data suggest that standard PMF, with a mass accuracy of 0.1 Da results in unsatisfying identification scores. In spite of the relatively high conservation of tryptic cleavage sites, the few mutations that modify internal residues during evolution significantly change the peptide masses, making difficult the identification since many missed peptides thus fall in the "unmatched peptides" category. To facilitate the identification procedure, some additional information for database searching should be used. This includes protein features such as protein mass and isoelectric point, which are both relatively well conserved. However, using these criteria, only four proteins were identified, with poor scores (Table 1). Although medaka genome has been completely sequenced (Kasahara et al., 2007), only 1933 proteins have been characterised in the NCBI protein database (December 2008). Matching PMF data to NCBInr frequently yielded insignificant results through cross species identification in related species.

MS/MS sequencing was systematically performed to complete PMF analyses and EST database was selected since data on *O. latipes* is available. Several EST sequences were returned with high scores and allowed identification in cases where PMF was unsuccessful using the same prepared samples.

This approach provides a framework to identify proteins after 2D electrophoretic separation. It should be fruitful to apply it after use of the shotgun technique, resulting in many isolated proteins being accurately identified.

Acknowledgements

Gildas Mouta-Cardoso is warmly thanked for his technical help with the 2D gels and Lionel Dubost for mass-spectrometry analysis. This work was supported by grants from the ANR 07 SEST CYANOTOX 005, the AFSSET APR EST 2007 10, the Institut National des Sciences de l'Univers/Ecodyn and was further supported by a grant to Dr. Marc Edery from L'Association pour la Recherche sur le Cancer. This study is a contribution to the IRD research group CYROCO UR167.

References

1. Altschul SF (1989) Gap costs for multiple sequence alignment. J Theor Biol 138: 297-309. » CrossRef » Pubmed » Google Scholar
2. Cordwell SJ, Wilkins MR, Cerpa PA, Gooley AA, Duncan M et al. (1995) Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption ionisation/time-of-flight mass spectrometry and amino acid composition. Electrophoresis 16: 438-443. » CrossRef » Pubmed » Google Scholar
3. Henrich T, Ramialison M, Quiring R, Wittbrodt B, Furutani SM et al. (2003) MEPD: a Medaka gene expression pattern database. Nucleic Acids Res 31: 72-74. » CrossRef » Pubmed » Google Scholar
4. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, et al. (2007) The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714-719. » CrossRef » Pubmed » Google Scholar
5. Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227: 680-685. » CrossRef » Pubmed » Google Scholar

6. Mezhoud K, Praseuth D, Puiseux DS, François JC, Bernard C et al. (2008) Global quantitative analysis of protein expression and phosphorylation status in the liver of the medaka fish (*Oryzias latipes*) exposed to microcystin-LR I. Balneation study. Aquat Toxicol 86:166-175.» CrossRef » Pubmed » Google Scholar

7. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551-3567. » CrossRef » Pubmed » Google Scholar

8. Wester PW, Van der Len LT, Vos JG (2004) Comparative toxicological pathology in mammals and fish: some examples with endocrine disrupters. Toxicology 205: 27-32. » CrossRef » Pubmed » Google Scholar

9. Wittbrodt J, Shima A, Schartl M (2002) Medaka — a model organism from the far EAST. Nat Rev Genet 3: 53-64. » CrossRef » Pubmed » Google Scholar