Biochemical and Biophysical Research Communications 397 (2010) 120–126

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

tsORFdb: Theoretical Small Open Reading Frames (ORFs) database and massProphet: Peptide Mass Fingerprinting (PMF) tool for unknown small functional ORFs

Hyoung-Sam Heo [a,b], Sanghyuk Lee [c], Ji Min Kim [b], Yeon Ja Choi [b], Hae Young Chung [b,*,1], S. June Oh [d,**,1]

[a] Interdisciplinary Research Program of Bioinformatics, College of Natural Science, Pusan National University, Gumjung-gu, Busan 609-735, Republic of Korea
[b] Department of Pharmacy, College of Pharmacy and Molecular Inflammation Research Center for Aging Intervention, Pusan National University, Gumjung-gu, Busan 609-735, Republic of Korea
[c] Division of Life and Pharmaceutical Sciences, Ewha Womans University, 11-1 Daehyun-dong, Seodaemun-gu, Seoul 120-750, Republic of Korea
[d] Department of Pharmacology, College of Medicine and UHRC, Inje University, Gaegum2-dong, Busanjin-gu, Busan 614-735, Republic of Korea

ARTICLE INFO

Article history:
Received 10 May 2010
Available online 23 May 2010

Keywords:
Bioinformatics
Database
ORFeomics
Peptide Mass Fingerprinting
Proteomics
sORFs (Small Open Reading Frames)

ABSTRACT

Peptide mass fingerprinting (PMF) has become one of the most widely used methods for rapid identification of proteins in proteomics research. Many peaks, however, remain unassigned after PMF analysis, partly because of post-translational modification and the limited scope of protein sequences. Almost all PMF tools employ only known or predicted protein sequences and do not include open reading frames (ORFs) in the genome, which eliminates the chance of finding novel functional peptides. Unlike most tools that search protein sequences from known coding sequences, the tool we developed uses a database for theoretical small ORFs (tsORFs) and a PMF application using a tsORFs database (tsORFdb). The tsORFdb is a database for ORFeome that encompasses all potential tsORFs derived from whole genome sequences as well as the predicted ones. The massProphet system tries to extend the search scope to include the ORFeome using the tsORFdb. The tsORFdb and massProphet should be useful for proteomics research to give information about unknown small ORFs as well as predicted and registered proteins.

1. Introduction

Matrix-assisted laser desorption/ionization–time-of-flight–mass spectrometry (MALDI–TOF–MS) is routinely used for efficient protein identification by peptide mass fingerprinting (PMF) [1–5]. PMF compares the observed peptide mass peaks with theoretical masses from the protein sequence databases. PMF tools available on the web include Mascot (http://www.matrixscience.com/search_form_select.html), Aldente (http://www.expasy.org/tools/aldente/), ProFound (http://prowl.rockefeller.edu/) [6], PepMAPPER (http://www.bioinf.manchester.ac.uk/mapper/), and ProteinProspector (http://prospector.ucsf.edu/). Even with these elaborate tools, however, many peaks remain unassigned after PMF analysis partly because of post-translational modifications and the limited scope of protein sequences.

UniProtKB/Swiss-Prot (http://www.expasy.org/sprot/) [7], the most widely used database of protein sequences for PMF, only includes experimentally verified proteins. This limited scope can be alleviated by including translated proteins from predicted genes and/or expressed sequence tags (EST's). According to human genome annotation, non-coding genomic regions account for 98–99% of the human genome [8].

Computational prediction of eukaryotic gene structure is crucial for the assessment of genome contents. There are two main categories in the current gene prediction programs: *de novo* predictors and expression-based predictors [9]. However, these are not sufficient for complete genome annotations. The number of protein-coding genes encoded in the eukaryotic genomes is still unknown. GENECODE, which seeks to identify all protein-coding genes, is an important sub-project of ENCODE (the Encyclopedia of DNA Elements) (http://www.genome.gov/10005107) [10].

Recently, peptides encoded by small (or short) functional ORFs (sORFs) were shown to control differentiation and development, and were defined as a new eukaryotic gene family [11–13]. One problem for current databases of protein sequences for PMF is the fact that initial annotation of the genome included only experimen-

* Corresponding author. Address: Department of Pharmacy, College of Pharmacy and Molecular Inflammation Research Center for Aging Intervention, Pusan National University, 30 Changjeon-dong, Kumjung-gu, Busan 609-735, Republic of Korea. Fax: +82 51 518 2821.
** Corresponding author. Address: Department of Pharmacology, Inje University College of Medicine, Gaegum2-dong, Busanjin-gu, Busan 614-735 Republic of Korea. Fax: +82 1540 3463 3669.
E-mail addresses: hyjung@pusan.ac.kr (H.Y. Chung), o@biophilos.org (S. June Oh).
[1] These authors contributed equally to this work.

0006-291X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2010.05.093

tally verified information, and therefore sORFs encoding functional proteins were largely missed [14,15] and were only considered following detection of expression [16–20]. Like the phenotypic consequences of gene disruption, large numbers of sORFs have not been assessed, and sORFs are underrepresented in genomic and proteomic databases, libraries, and other collections [11].

A major challenge of biological science is elucidating the interactions of cellular networks. The underlying complexity arises from intertwined non-linear and dynamic interactions among large numbers of biological molecules, such as genes, proteins, and metabolites, either known or unknown. The reductionism has successfully identified many components and many interactions. However, many questions remain unanswered. How many components have not yet been characterized? Recently, peptides encoded by sORFs were shown to play key roles in various biological processes. Recent genome annotation included only those regions consisting of at least 100 contiguous codons, and therefore sORFs encoding functional peptides were largely missed and were only considered following detection of expression.

Deciphering the biological big picture is one of the main goals in the life sciences. In proteomic research, the current big picture includes many blanks, because current proteomic research focuses on the "usual suspects," such as well-known proteins. Perhaps, some of these blanks are unknown sORFs. In order to fill in the blanks, in proteome research, we propose here a new conceptual approach and method. A database for all potential tsORFs from the whole genome sequence and a PMF application using this database were designed and constructed. For the PMF application, an optimized scoring solution was implemented. With this new concept and method, we sought a verification method using theoreti-

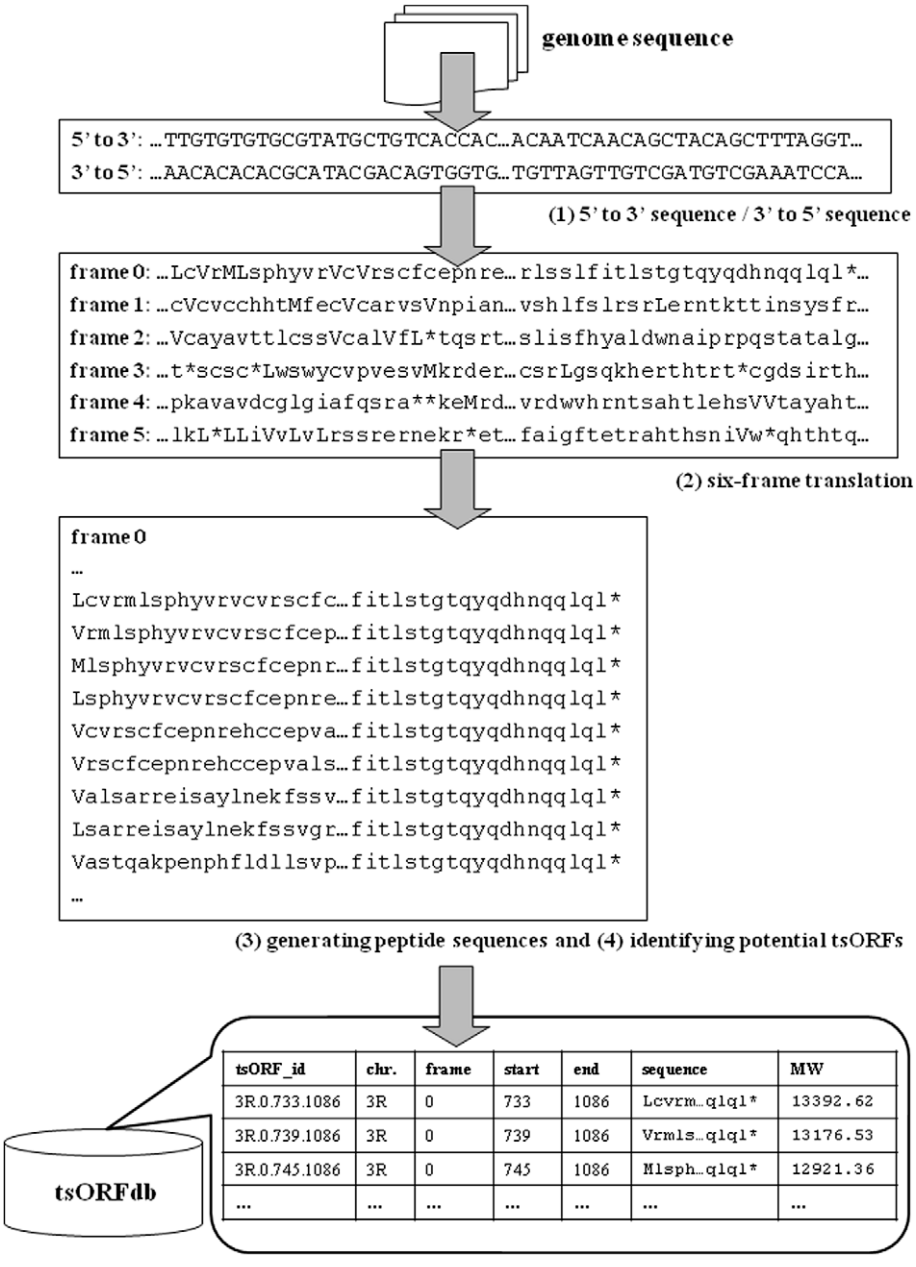

Fig. 1. tsORFdb construction. Each chromosome nucleotide sequence file was downloaded from the UCSC genome bioinformatics site. Processing tsORFs involves the following steps: (1) six reading frame sequence generation from 5′ to 3′ direction and from 3′ to 5′ direction, (2) six-frame translation of each chromosomal nucleotide sequence regardless of gene structure, (3) generating peptide sequences using termination codon, and (4) identifying initiation sites and alternative initiation sites from each peptide sequence to derive potential tsORFs.

cal identification of known sORFs. This research extended the search to include all potential tsORFs from the genome map, the tsORFdb, and developed a web-based PMF application using a database for all potential tsORFs, the massProphet system. The results helped identifying previously unknown sORFs in proteomic research.

2. Materials and methods

2.1. Data source

Whole genome sequences data were downloaded from the University of California Santa Cruz (UCSC) genome browser database (http://genome.ucsc.edu/) [21]. Human (*Homo sapiens*) whole genome data were released in March 2006 (hg18). Mouse (*Mus musculus*) whole genome data were released in July 2007 (mm9). Fruit fly (*Drosophila melanogaster*) whole genome data were released in April 2006 (dm3). Yeast (*Saccharomyces cerevisiae*) whole genome data were released in October 2003 (sacCer1).

2.2. tsORFdb design and implementation

The existing tools for PMF offer only known protein identifications because these tools use known coding region sequences such as (known or predicted) protein, short cDNA, and EST's. To surpass the limits of existing PMF tools, tsORFdb was designed and constructed to include all tsORFs from whole genome sequences.

All potential tsORFs derived from whole genome sequences downloaded from the UCSC genome browser database. Processing tsORFs included the following steps (Fig. 1) on each chromosome files:

Step 1: six-frame translation of each chromosomal sequence was performed regardless of gene structure;
Step 2: peptide sequences were separated according to the termination site; and
Step 3: initiation sites and alternative initiation sites from separated peptide sequences were identified, and potential tsORFs were derived.

We note that a leucine can be an alternative start codon [22,23]; genetic codes for tsORFdb and non-ATG initiation codon lists are provided in Supplementary Material 1. For example, if a separated peptide sequence was:

'LcVrMlsphyvrvcvrscfcepnrehccepvalsarreisaylnekfssvgrrtap rhthshicrntttyetnvastqakpenphfldllsvprlsslfitlstgtqyqdhnqqlql*'

the potential tsORFs would be:

[1] the first alternative initiation site to the termination site:

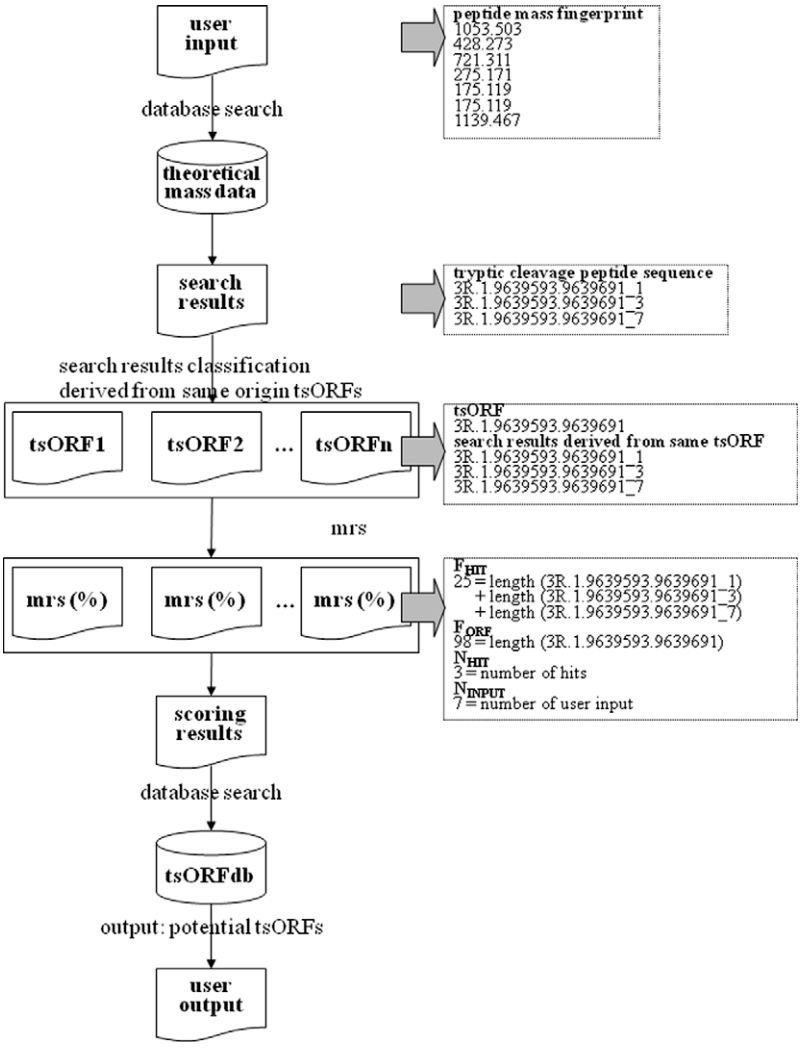

Fig. 2. Schematic process of massProphet scoring. The massProphet scoring algorithm includes the following steps: (1) tracing the origin tsORF for the search result of each peptide mass fingerprint, (2) classifying the search results by tsORF origin, and (3) match ratio score (mrs) calculation for each class.

'Lcvrmlsphyvrvcvrscfcepnrehccepvalsarreisaylnekfssvgrr-taprhth
shicrntttyetnvastqakpenphfldllsvprlsslfitlstgtqyqdhnqqlql*'

[2] the second initiation site to the termination site:
'Vrmlsphyvrvcvrscfcepnrehccepvalsarreisaylnekfssvgrrtaprhth-shicrntttyetnvastqakpenphfldllsv prlsslfitlstgtqyqdhnqqlql*'

[3] the third initiation site to the termination site:
'Mlsphyvrvcvrscfcepnrehccepvalsarreisaylnekfssvgrrtaprhth-shicrntttyetnvastqakpenphfldllsvprl sslfitlstgtqyqdhnqqlql*'

[4] the fourth initiation site to the termination site:
'Lsphyvrvcvrscfcepnrehccepvalsarreisaylnekfssvgrrtaprhth-shicrntttyetnvastqakpenphfldllsvprlsslfitlst gtqyqdhnqqlql*'

[5] the fifth initiation site to the termination site:
'Vcvrscfcepnrehccepvalsarreisaylnekfssvgrrtaprhthsh icrntttyetnvastqakpenphfldllsvprlsslfitlstgtqyqdhnqqlql*'

[6] the sixth initiation site to the termination site:
'Vrscfcepnrehccepvalsarreisaylnekfssvgrrtaprhthshicrntttyetn-vastqakpenphfldll svprlsslfitlstgtqyqdhnqqlql*'

[7] the seventh initiation site to the termination site:
'Valsarreisaylnekfssvgrrtaprhthshic
rntttyetnvastqakpenphfldllsvprlsslfitlstgtqyqdhnqqlql*'

[8] the eighth initiation site to the termination site:
'Lsarreisaylnekfssvgrrtaprhthshicrntttyetnvastqak-penphfldllsvprl sslfitlstgtqyqdhnqqlql*'

[9] the ninth initiation site to the termination site:
'Vastqakpenphfldllsvprlsslfitlstgtqyqdhnqqlql*'

[10] the last initiation site to the termination site:'Llsvprlsslfitl stgtqyqdhnqqlql*'

(Capital letters indicate initiation and alternative initiation sites. '*' indicates a termination site.).

The current version of the tsORFdb includes human, mouse, fruit fly, and yeast. The minimum and maximum tsORF lengths of each species are provided in Supplementary Material 2.

2.3. massProphet system design and implementation

The massProphet system consists of two components: [1] theoretical mass fingerprint derived from tsORFdb, and [1] scoring algorithm for assessing the significance of hits.

For each tsORF theoretical peptide masses were calculated from all possible peptide fragments based on the trypsin cleavage rules. The current version of massProphet includes tryptic digested theoretical mass fingerprints, and post-translational modification was not incorporated into the calculation of molecular weights.

Instead of the MOWSE scoring algorithm [24], we implemented a two-stage solution for scoring solution optimized for the massProphet system. The scoring algorithm for massProphet was based on the mapping of tsORFs deduced from the whole genome sequence of an organism to the peptide fragments of the tsORFs. The first stage was based on peptide mass fingerprint search results derived from same-origin tsORFs. This stage included the following steps:

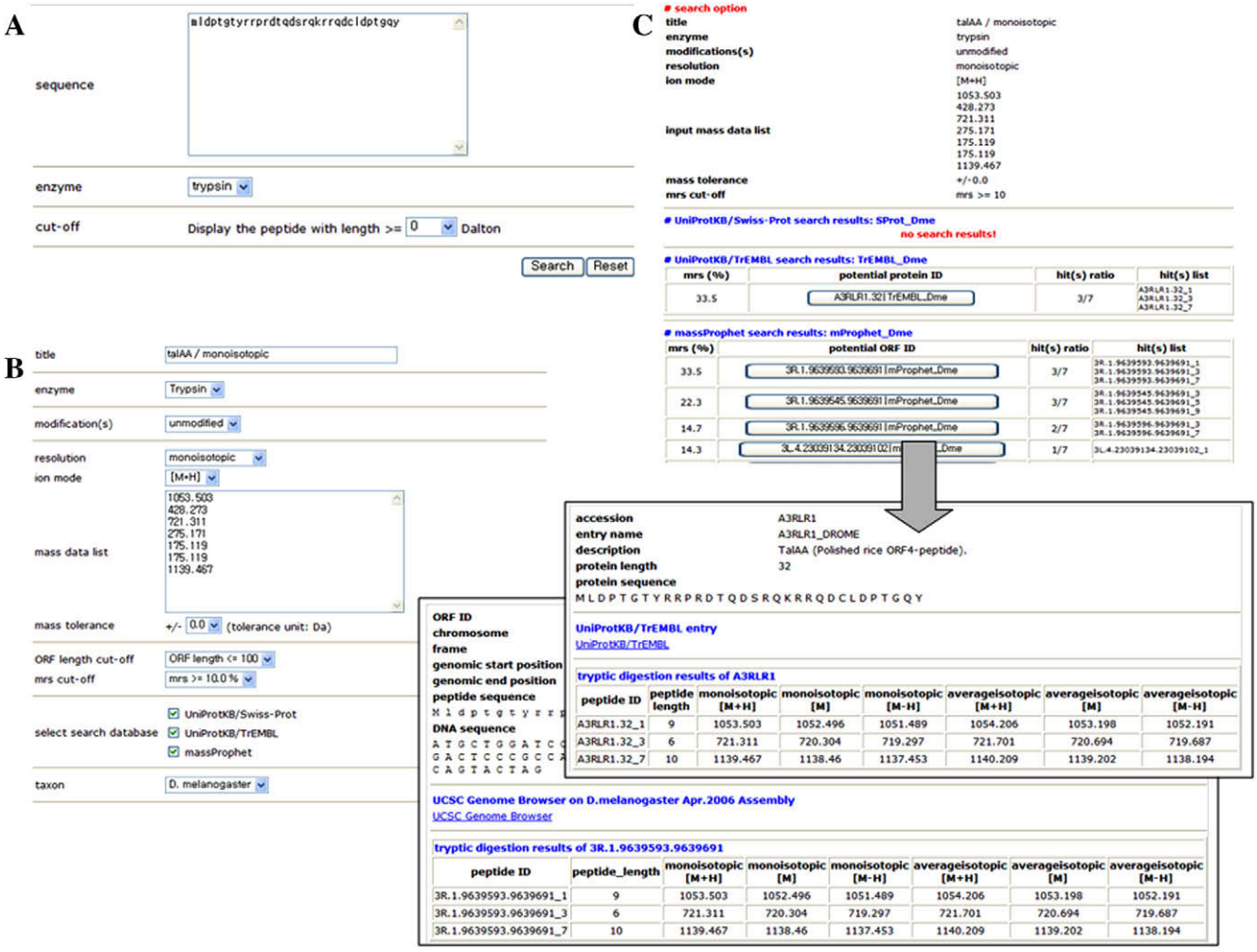

Fig. 3. Querying massProphet. (A) massCalculator for the calculation of theoretical peptide mass fingerprint from user-input peptide sequence. (B) Querying interface of the massProphet system. (C) Comparing search results of the massProphet to UniProtKB/Swiss-Prot and UniProtKB/TrEMBL database search in the massProphet search mode. The current massProphet search results show tsORF information and linkouts to UCSC genome browser.

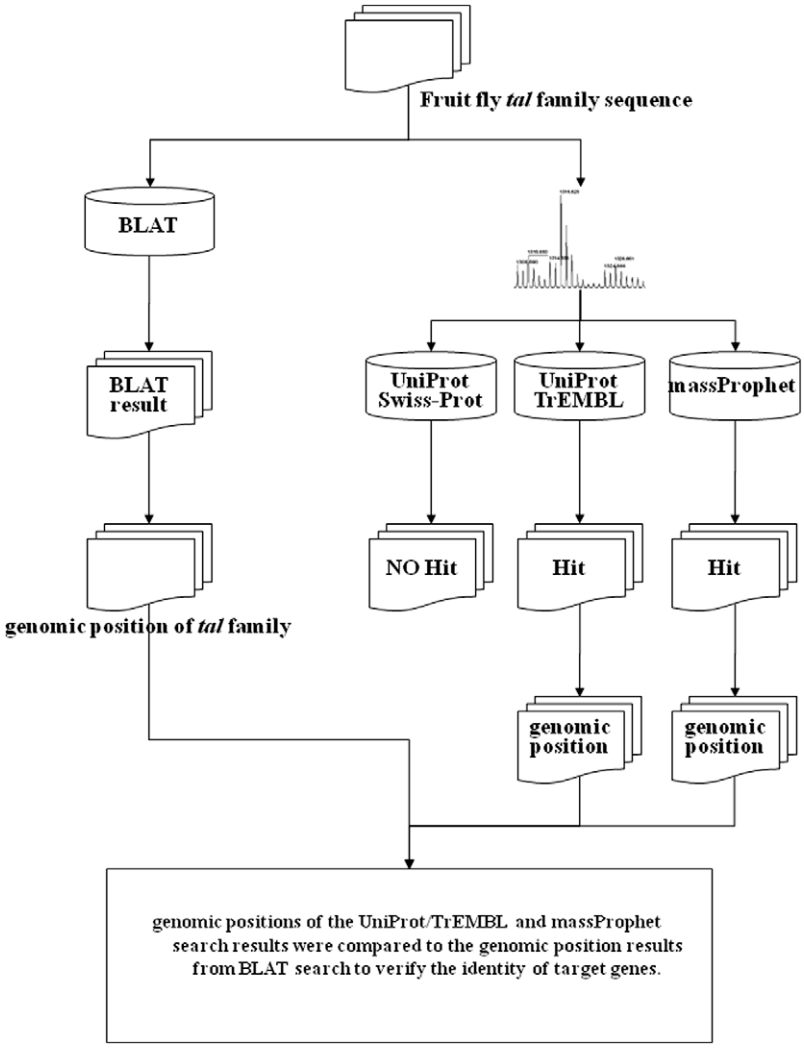

Fig. 4. Verification flow. The verification flow includes the following steps: (1) confirming the genomic position of *tal* gene family using BLAT search, (2) analyzing the theoretical peptide mass fingerprint of *tal* gene family by the massProphet search to yield mrs, and (3) comparing the genomic positions of the massProphet search results to the genomic position results from BLAT search to verify the identity of target genes.

Step 1: for each peptide mass fingerprint search result, the origin tsORF was traced and
Step 2: search results were classified according to the same-origin tsORFs.

The second stage was a match ratio score (mrs) for each class. The mrs was calculated as:

$$\text{mrs} = \left(\sum_{i=1}^{n} F_{HIT} \Big/ \sum_{j=1}^{m} F_{ORF} \right) * (N_{HIT}/N_{INPUT}) \quad (1)$$

where F_{HIT} is the search hit from the same-origin tsORF, F_{ORF} is the trypsin cleavage peptide sequence from the same-origin tsORF, N_{HIT} is the number of hits, and N_{INPUT} is the number of user-inputs. Fig. 2 shows the schematic process of massProphet scoring.

3. Results and discussion

3.1. Querying tsORFdb

The tsORFdb provides a querying interface (Supplementary Material 3A) and the sequence of the tsORFdb in FASTA format (Supplementary Material 3B) [25]. tsORF information includes identification numbers of the tsORF, origin chromosome, origin frame, genomic start position, genomic end position, peptide sequence of the tsORF, and DNA sequence of the tsORF (Supplementary Material 3C).

3.2. Querying massProphet

The massProphet system provides a massCalculator (Fig. 3A) that calculates a theoretical peptide mass fingerprint derived from a user-input peptide sequence and the massProphet peptide mass fingerprint search interface (Fig. 3B). For comparison of the massProphet search results with other databases, our system is able to incorporate UniProtKB/Swiss-Prot and UniProtKB/TrEMBL databases into searches in the massProphet search mode (Fig. 3C). The massProphet search results include identification numbers of the tsORF, origin chromosome, genomic start position, genomic end position, peptide sequence of the tsORF, DNA sequence of the tsORF, and tryptic digestion results of the tsORF. The tsORFdb and massProphet search result provides link to the UCSC genome browser database.

3.3. Verification of the massProphet system

For verification, we identified known functional small ORFs in fruit fly, reported as the tarsal-less (*tal*) gene family [13]. The verification flow (Fig. 4) included the following steps:

Table 1
Genomic information of the fruit fly *tal* gene family.

tal1A	mRNA	EF427619.1:416...451[a]
	Sequence	atggcagcct acttggatcc cactggccag tactaa
	BLAT search result	chr.3R:9639248-9639283(+)
tal2A	mRNA	EF427619.1:529...564[a]
	Sequence	atggccgcct atctggatcc cactggtcag tactga
	BLAT search result	chr.3R: 9639361-9639396(+)
tal3A	mRNA	EF427619.1:636...671[a]
	Sequence	atgtcgcacg atttggaccc cactggcacc tactaa
	BLAT search result	chr.3R: 9639468-9639503(+)
talAA	mRNA	EF427619.1:761...859[a]
	Sequence	atgctggatc ccactggaac ataccggcga ccacgcgaca cgcaggactc ccgccaaaag aggcgacagg actgcctgga tccaaccggg cagtactag
	BLAT search result	chr.3R: 9639593-9639691(+)

[a] http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nuccore&id=126256578.

Step 1: the genomic position of the *tal* gene family was confirmed using a BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) [26] search;
Step 2: the theoretical peptide mass fingerprint of the *tal* gene family was analyzed by the massProphet search to yield mrs; and
Step 3: the genomic positions of the massProphet search results were compared with the genomic position results from BLAT searches to verify the identity of target genes.

For comparison of the massProphet search result with that of other databases, we processed UniProtKB/Swiss-Prot and UniProtKB/TrEMBL [27] sequence data and constructed cognate databases that were used in the massProphet search mode.

Table 1 presents genomic information of the fruit fly gene family (*tal1A*, *tal2A*, *tal3A*, and *talAA*) by BLAT search. We verified the information through comparison of the genomic position of the fruit fly gene family and the massProphet search results using the theoretical mass fingerprint of the fruit fly gene family. Table 2 shows the theoretical verification results of the massProphet system. There is no hit in the UniProtKB/Swiss-Prot search. Using massProphet, we obtained relevant results: total hits for *tal1A*, 485 potential tsORFs; *tal2A*, 485 potential tsORFs; *tal3A*, 15 potential tsORFs; *talAA*, 86 potential tsORFs. Many similar sequences were included in tsORFdb because all potential tsORFs were extracted from the whole genome. The massProphet scoring algorithm calculated the mrs of many search results and offered results in order of probability. In the case of UniProtKB/TrEMBL, we obtained correct search results because the *tal* gene family was published in UniProtKB/TrEMBL.

The fruit fly *tal* gene family is located on chromosome 3R. In the case of chromosome 3R, according to the UCSC genome browser database, 5664 mRNAs and 151,581 EST's have been identified. The present study predicted 3737,906 potential tsORFs on chromosome 3R. On the plus strand of chromosome 3R, approximately 0.161% of tsORFs overlapped with EST's on the same chromosome, whereas approximately 0.012% of tsORFs overlapped with EST's on the minus strand of chromosome 3R.

The massProphet system would be useful for proteomics research to give information about unknown short potential proteins or oligopeptides as well as the predicted and registered ones.

Table 2
Verification of the massProphet system. Search results comparison among UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, and massProphet (fruit fly chromosome 3R).

tal1A	UniProtKB/Swiss-Prot	No search result	
	UniProtKB/TrEMBL	Match ratio score	100%
		Accession	A3RLQ8[b]
		Description	*Tal1A (Tal2A)*
	massProphet	Match ratio score	100%
		ORF ID	3R.1.9639248.9639283
tal2A	UniProtKB/Swiss-Prot	No search result	
	UniProtKB/TrEMBL	Match ratio score	100%
		Accession	A3RLQ8[b]
		Description	*Tal1A (Tal2A)*
	massProphet	Match ratio score	100%
		ORF ID	3R.0.9639361.9639396
tal3A	UniProtKB/Swiss-Prot	No search result	
	UniProtKB/TrEMBL	Match ratio score	100%
		Accession	A3RLR0[c]
		Description	*Tal3A*
	massProphet	Match ratio score	100%
		ORF ID	3R.2.9639468.9639503
talAA	UniProtKB/Swiss-Prot	No search result	
	UniProtKB/TrEMBL	Match ratio score	33.5%[a]
		Accession	A3RLR1[d]
		Description	*TalAA*
	massProphet	Match ratio score	33.5%[a]
		ORF ID	3R.1.9639593.9639691

[a] mrs of a correct hit was 33.5% because tryptic peptides from *talAA* ORF sequence included many peptide fragments smaller than 500 Da.
[b] http://www.uniprot.org/uniprot/A3RLQ8.
[c] http://www.uniprot.org/uniprot/A3RLR0.
[d] http://www.uniprot.org/uniprot/A3RLR1.

4. Conclusions

We presented a database for all potential tsORFs from the genome map and a PMF application using the database for all potential tsORFs. The results are stored in a web-enabled database called tsORFdb and a web-based application called massProphet. The purpose of tsORFdb is to list all possible sORFs from whole genome sequence data that can be predicted through bioinformatic approaches. The tsORFdb also may be a useful reference for information on all potential sORFs. Using tsORFdb, massProphet can identify sORFs that are unassigned using UniProtKB/Swiss-Prot database. Although further experimental validation is needed to confirm the proteomic experiments related to sORFs and sORFs identification using tsORFdb and massProphet, the tsORFdb and massProphet are useful databases, as is the PMF application, for proteomics research to give information about unknown small functional open reading frames as well as the predicted and registered proteins. Database schema and statistics for tsORFdb are provided in Supplementary Material 4 and Supplementary Material 5.

tsORFdb and massProphet are available from the following web address: http://bionet.inje.ac.kr/massProphet. The Supplementary Material and on-line help at the tsORFdb and massProphet site describe the application in more detail.

We will update the content of the tsORFdb and massProphet depending on update of genome annotation. The tsORFdb and massProphet will also be included in database of various species.

Acknowledgments

This work was supported by and a National Research Foundation of Korea (NRF) grant (No. 2009-0083538) funded by the Korea government (MEST) and the 2005 Inje University research grant. We are grateful to the 'Aging Tissue Bank' for supplying research resources.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the on-line version, at doi:10.1016/j.bbrc.2010.05.093.

References

[1] R.M. Caprioli, T.B. Farmer, J. Gile, Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS, Anal. Chem. 69 (1997) 4751–4760.
[2] V. Egelhofer, K. Bussow, C. Luebbert, H. Lehrach, E. Nordhoff, Improvements in protein identification by MALDI-TOF-MS peptide mapping, Anal. Chem. 72 (2000) 2741–2750.
[3] T. Bonk, A. Humeny, MALDI-TOF-MS analysis of protein and DNA, Neuroscientist 7 (2001) 6–12.
[4] V. Egelhofer, J. Gobom, H. Seitz, P. Giavalisco, H. Lehrach, E. Nordhoff, Protein identification by MALDI-TOF-MS peptide mapping: a new strategy, Anal. Chem. 74 (2002) 1760–1771.
[5] W.J. Henzel, C. Watanabe, J.T. Stults, Protein identification: the origins of peptide mass fingerprinting, J. Am. Soc. Mass Spectrom. 14 (2003) 931–942.
[6] W. Zhang, B.T. Chait, ProFound: an expert system for protein identification using mass spectrometric peptide mapping information, Anal. Chem. 72 (2000) 2482–2489.
[7] C.H. Wu, R. Apweiler, A. Bairoch, D.A. Natale, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, B. Suzek, The universal protein resource (UniProt): an expanding universe of protein information, Nucl. Acids Res. 34 (2006) D187–D191.
[8] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D.K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D.S. Gerhard, T.R. Gingeras, Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution, Science 308 (2005) 1149–1154.
[9] M.R. Brent, How does eukaryotic gene prediction work?, Nat Biotechnol. 25 (2007) 883–885.
[10] E. Birney, J.A. Stamatoyannopoulos, A. Dutta, R. Guigo, T.R. Gingeras, E.H. Margulies, Z. Weng, M. Snyder, E.T. Dermitzakis, R.E. Thurman, M.S. Kuehn, C.M. Taylor, S. Neph, C.M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J.A. Greenbaum, R.M. Andrews, P. Flicek, P.J. Boyle, H. Cao, N.P. Carter, G.K. Clelland, S. Davis, N. Day, P. Dhami, S.C. Dillon, M.O. Dorschner, H. Fiegler, P.G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K.D. James, B.E. Johnson, E.M. Johnson, T.T. Frum, E.R. Rosenzweig, N. Karnani, K. Lee, G.C. Lefebvre, P.A. Navas, F. Neri, S.C. Parker, P.J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F.S. Collins, J. Dekker, J.D. Lieb, T.D. Tullius, G.E. Crawford, S. Sunyaev, W.S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I.L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H.A. Hirsch, E.A. Sekinger, J. Lagarde, J.F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J.S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M.C. Dickson, D.J. Thomas, M.T. Weirauch, J. Gilbert, et al., Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, Nature 447 (2007) 799–816.
[11] R. Sopko, B. Andrews, Small open reading frames: not so small anymore, Genome Res. 16 (2006) 314–315.
[12] J.P. Kastenmayer, L. Ni, A. Chu, L.E. Kitchen, W.C. Au, H. Yang, C.D. Carter, D. Wheeler, R.W. Davis, J.D. Boeke, M.A. Snyder, M.A. Basrai, Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*, Genome Res. 16 (2006) 365–373.
[13] M.I. Galindo, J.I. Pueyo, S. Fouix, S.A. Bishop, J.P. Couso, Peptides encoded by short ORFs control development and define a new eukaryotic gene family, PLoS Biol. 5 (2007) e106.
[14] A. Goffeau, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S.G. Oliver, Life with 6000 genes, Science 274 (1996) 546–567.
[15] A. Goffeau, Four years of post-genomic life with 6000 yeast genes, FEBS Lett. 480 (2000) 37–41.
[16] W.M. Olivas, D. Muhlrad, R. Parker, Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs, Nucl. Acids Res. 25 (1997) 4619–4625.
[17] V.E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett, Jr., P. Hieter, B. Vogelstein, K.W. Kinzler, Characterization of the yeast transcriptomes, Cell 88 (1997) 243–251.
[18] A. Kumar, P.M. Harrison, K.H. Cheung, N. Lan, N. Echols, P. Bertone, P. Miller, M.B. Gerstein, M. Snyder, An integrated approach for finding overlooked genes in yeast, Nat. Biotechnol. 20 (2002) 58–63.
[19] G. Oshiro, L.M. Wodicka, M.P. Washburn, J.R. Yates 3rd, D.J. Lockhart, E.A. Winzeler, Parallel identification of new genes in *Saccharomyces cerevisiae*, Genome Res. 12 (2002) 1210–1220.
[20] M.M. Kessler, Q. Zeng, S. Hogan, R. Cook, A.J. Morales, G. Cottarel, Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome, Genome Res. 13 (2003) 264–271.
[21] D. Karolchik, R.M. Kuhn, R. Baertsch, G.P. Barber, H. Clawson, M. Diekhans, B. Giardine, R.A. Harte, A.S. Hinrichs, F. Hsu, K.M. Kober, W. Miller, J.S. Pedersen, A. Pohl, B.J. Raney, B. Rhead, K.R. Rosenbloom, K.E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A.S. Zweig, D. Haussler, W.J. Kent, The UCSC genome browser database 2008 update, Nucl. Acids Res. 36 (2008) (2008) D773–D779.
[22] T. Soldati, B.W. Schafer, J.C. Perriard, Alternative ribosomal initiation gives rise to chicken brain-type creatine kinase isoproteins with heterogeneous amino termini, J. Biol. Chem. 265 (1990) 4498–4506.
[23] S.R. Schwab, J.A. Shugart, T. Horng, S. Malarkannan, N. Shastri, Unanticipated antigens: translation initiation at CUG with leucine, PLoS Biol. 2 (2004) e366.
[24] D.J. Pappin, P. Hojrup, A.J. Bleasby, Rapid identification of proteins by peptide-mass fingerprinting, Curr. Biol. 3 (1993) 327–332.
[25] W.R. Pearson, Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms, Genomics 11 (1991) 635–650.
[26] W.J. Kent, BLAT–the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.
[27] C. O'Donovan, M.J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, R. Apweiler, High-quality protein knowledge resource: SWISS-PROT and TrEMBL, Brief Bioinform 3 (2002) 275–284.