22

# Emergence of the Diversified Short ORFeome by Mass Spectrometry-Based Proteomics

Hiroko Ao-Kondo, Hiroko Kozuka-Hata and Masaaki Oyama
*Medical Proteomics Laboratory, Institute of Medical Science, University of Tokyo*
*Japan*

## 1. Introduction

In proteomics analyses, protein identification by mass spectrometry (MS) is usually performed using protein sequence databases such as RefSeq (NCBI; http://www.ncbi.nlm.nih.gov/RefSeq/), UniProt (http://www.uniprot.org/) or IPI (http://www.ebi.ac.uk/IPI/IPIhelp.html). Because these databases usually target the longest (main) open reading frame (ORF) in the corresponding mRNA sequence, whether shorter ORFs on the same mRNA are actually translated still shrouds in mystery. In the first place, it had been considered that almost all eukaryotic mRNAs contains only one ORF and functions as monocistronic mRNAs. It is now known, however, that some eukaryotic mRNAs had multiple ORFs, which are recognized as polycistronic mRNAs. One of the well-known extra ORFs is an upstream ORF (uORF) and it functions as regulators of mRNA translation (Diba et al., 2001; Geballe & Morris, 1994; Morris & Geballe, 2000; Vilela & McCarthy, 2003; Zhang & Dietrich, 2005). For getting clues to the mystery of diversified short ORFs, full-length mRNA sequence databases with complete 5'-untranslated regions (5'-UTRs) were essentially needed (Morris & Geballe, 2000; Suzuki et al., 2001).

The oligo-capping method was developed to construct full-length cDNA libraries (Maruyama & Sugano, 1994) and the corresponding sequence were stored into the database called DBTSS (DataBase of Transcriptional Start Site; http://dbtss.hgc.jp/) (Suzuki et al., 1997, 2002, 2004; Tsuchihara et al., 2009; Wakaguri et al., 2008; Yamashita et al., 2006). Comparing the dataset of DBTSS with the corresponding RefSeq entries, it was found that about 50 % of the RefSeq entries had at least one upstream ATG (uATG) except the functional ATG initiator codon (Yamashita et al., 2003). Although it had been suggested that upstream AUGs (uAUGs) and uORFs play important roles for translation of the main ORF, none of the proteins from these uORFs was detected in biological experiments in vivo. Our previous proteomics analysis focused on small proteins revealed the first evidence of the existence of four novel small proteins translated from uORFs in vivo using highly sensitive nanoflow liquid chromatography (LC) coupled with the electrospray ionization-tandem mass spectrometry (ESI-MS/MS) system (Oyama et al., 2004). Large-scale analysis based on in-depth separation by two-dimensional LC also led to the identification of additional eight novel small proteins not only from uORFs but also from downstream ORFs and one of them was found to be translated from a non-AUG initiator codon (Oyama et al., 2007). Finding of these novel small proteins indicate the possibility of diverse control mechanisms of translation initiation.

In this chapter, we first introduce widely-recognized mechanism of translation initiation and functional roles of uORF in translational regulation. We then review how we identified novel small proteins with MS and lastly discuss the progress of bioinformatical analyses for elucidating the diversification of short coding regions defined by the transcriptome.

## 2. Translational regulation by short ORFs

It is well known that 5′-UTRs of some mRNAs contain functional elements for translational regulation defined by uAUG and uORF. In this section, we show how uAUG and uORF have biological consequences for protein synthesis on eukaryotic mRNAs.

### 2.1 Outline of translation initiation

Initiation of translation on eukaryotic mRNAs occurs roughly as follows (Fig. 1) (Kozak, 1989, 1991, 1999).

1. A small (40S) ribosomal subunit binds near the 5′-end of mRNA, i.e. the cap structure.
2. The 40S subunit migrates linearly downstream of the 5′-UTR until it encounters the optimum AUG initiator codon.
3. A large (60S) ribosomal subunit joins the paused 40S subunit.
4. The complete ribosomal complex (40S + 60S) starts protein synthesis.

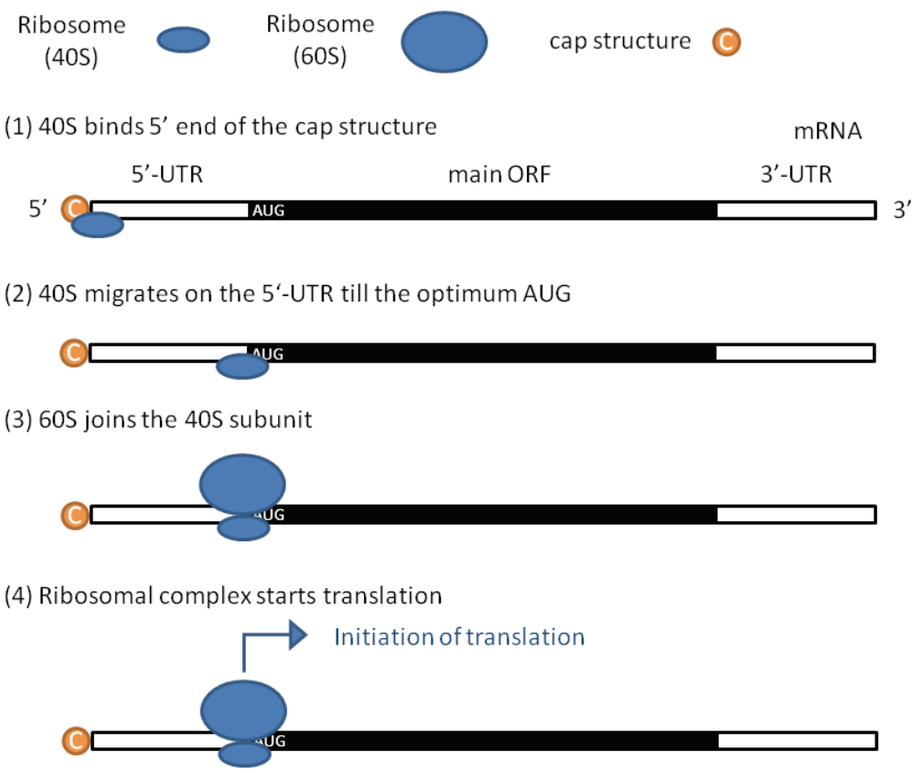



Fig. 1. The proposed procedure for initiation of translation in eukaryotes. The black region indicates the main ORF of the mRNA.

In addition to the above mechanism, initiation of translation without the step of ribosome scanning is also known. It is called "internal initiation", which depends on some particular structure on an mRNA termed internal ribosome entry site (IRES).

### 2.2 The relationship between uORF and main ORF

In case that an mRNA contains a uORF, two models for the initiation of translation are suggested (Fig. 2) (Hatzigeorgiou, 2002). One is called "leaky scanning" and the other is "reinitiation". If the first AUG codon is in an unfavorable sequence context defined by Kozak (see the section 3.2), a small ribosomal subunit (40S) ignores the first AUG and initiates translation from a more favorable AUG codon downstream located. This phenomenon is known as "leaky scanning" (Fig. 2-(A)). In case that a complete ribosomal complex translates a main ORF after termination of translation of the uORF on the same mRNA, it is termed "reinitiation" (Fig. 2-(B)).

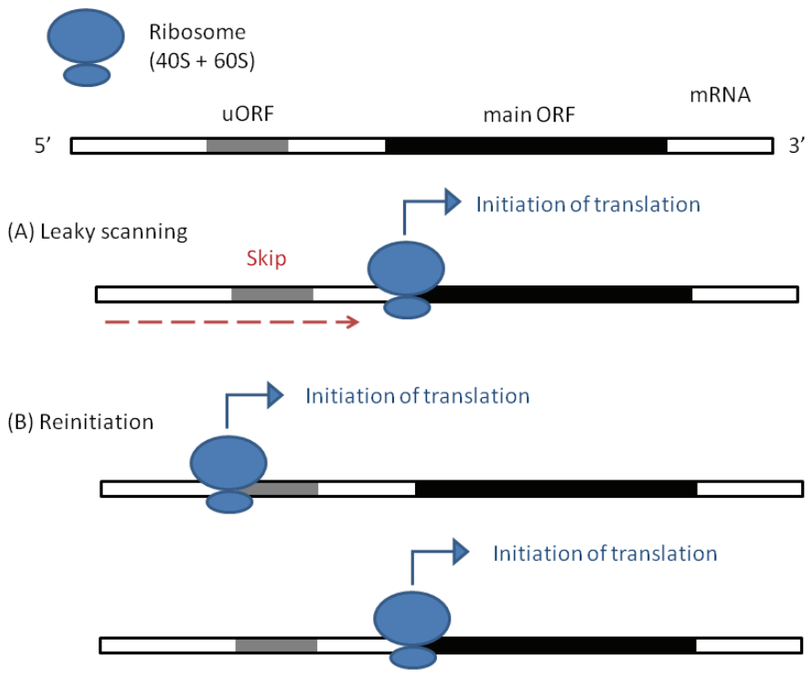



Fig. 2. The irregular models of ribosome scanning on eukaryotic mRNAs.
(A) Leaky scanning and (B) Reinitiation. Gray regions indicate uORFs on the mRNA, whereas black ones represent the main ORFs.

The relations between two ORFs are classified into three types as follows; (1) A distant type; in-frame/out-of-frame, (2) A contiguous type; in-frame and (3) An overlapped type; in-frame/out-of-frame (Fig. 3). *In-frame* means that a uORF and the main ORF are on the same frame of the mRNA sequence, whereas *out-of-frame* means that they are on the different frame. According to the previous analysis of the accumulated 5'-end sequence data, the average size of uORF was estimated at 31 amino acids and 20 % of ORFs were categorized into Type (3) (Yamashita et al., 2003).

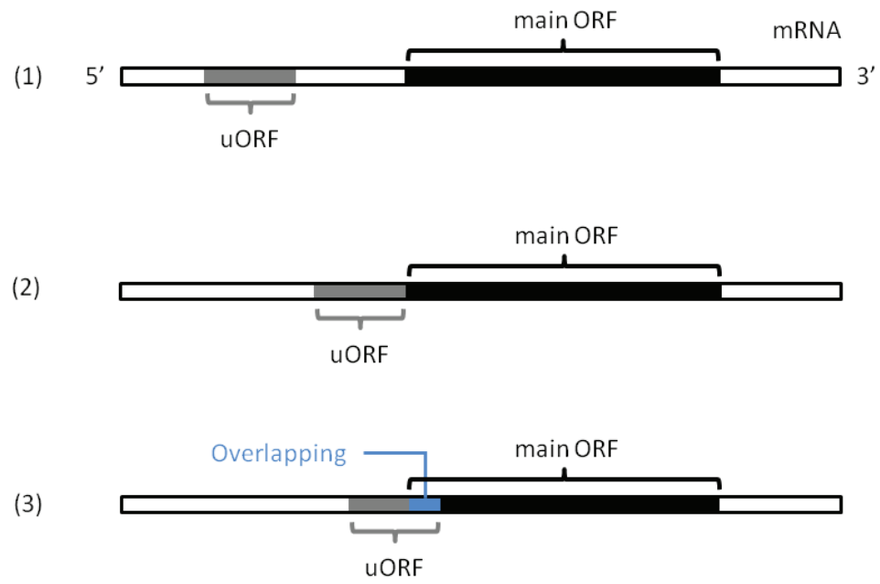



Fig. 3. The location of a uORF and the main ORF on the mRNA.
(1) A distant type, (2) A contiguous type and (3) An overlapped type. Types (1) and (3) have two subtypes based on the frames of two ORFs. One is defined by the same reading frame (in-frame) and the other is by the different one (out-of-frame). Gray and black regions indicate uORFs and the main ORFs on mRNAs, respectively, whereas a blue one represents an overlap.

These different relations might bring about different events in initiating translation. In eukaryotes, it has a tendency to increase an efficiency of reinitiation if the distance between a uORF and the main ORF is long (Kozak, 1991; Meijer & Thomas, 2002; Morris & Geballe, 2000). Therefore, the ORFs classified as Types (2) and (3) would be difficult to be regulated by reinitiation. It is also said that reinitiation occurs only when the length of uORF is short (Kozak, 1991), whereas the sequence context of an inter-ORF's region, that of upstream of uORF, uORF itself and even the main ORF can also affect reinitiation (Morris & Geballe, 2000). On the contrary, the ORFs of Type (3) might easily cause leaky scanning (Geballe & Morris, 1994; Yamashita et al., 2003). As a special case, when a termination codon of the uORF is near the AUG initiator codon of the downstream ORF, within about 50 nucleotides, ribosomes could scan backwards and reinitiate translation from the AUG codon of the downstream ORF (Peabody et al., 1986).

### 2.3 The role of short ORFs in translation regulation

The 5'-UTR elements such as uAUGs and uORFs are well known as important regulators for translation initiation. In case of some genes that have multiple uORFs, considerably different effects can be generated on the translation of the main ORF depending on which combination of uORFs is translated. Some uORFs seem to promote reinitiation of the main ORFs and the others seem to inhibit it. It is supposed that these effects are caused by the nucleotide sequences of the 3' ends of the uORFs, that of uORFs or protein products encoded by uORFs. Such differential enhancement of translation are considered to be one of the responses of adaptation to the environment (Altmann & Trachsel, 1993; Diba et al., 2001;

Geballe & Morris, 1994; Hatzigeorgiou, 2002; Iacono et al., 2005; Meijer & Thomas, 2002; Morris & Geballe, 2000; Vilela & McCarthy, 2003; Wang & Rothnagel, 2004; Zhang & Dietrich, 2005). In addition to that, various factors or events are known to influence on the translational inhibition of the main ORF; the presence of arginine, a stalling of a ribosomal complex at the termination or an interaction between a ribosomal complex and the peptide encoded by the uORF, which indicates that down-regulated controls by uORFs are general (Diba et al., 2001; Geballe & Morris, 1994; Iacono et al., 2005; Meijer & Thomas, 2002; Morris & Geballe, 2000; Vilela & McCarthy, 2003; Zhang & Dietrich, 2005).
As for downstream ORFs, there is also a report that a peptide encoded in the 3′-UTR may be expressed (Rastinejad & Blau, 1993). However, whether and how the peptides control the translation initiation of the main ORF is still unknown.

## 3. Variability of translation start sites

How a ribosomal complex (40S + 60S) recognizes an initiator codon on the mRNA is a matter of vital importance for defining the proteome. Here we present a part of already proposed elements for regulation of translation initiation.

### 3.1 The first-AUG rule

Traditionally, the first-AUG rule is widely recognized for initiation of translation (Kozak, 1987, 1989, 1991). It states that ribosomes start translation from the first-AUG on the corresponding mRNA. Although this rule is not absolute, 90-95 % of vertebrate ORFs was established by the first AUG codon on the mRNA (Kozak, 1987, 1989, 1991). Our previous proteomics analysis of small proteins also indicated that about 84 % of proteins in RefSeq were translated from the first AUG of the corresponding mRNAs (Oyama et al., 2004). On the other hand, there are also many negative reports concerning the rule; 29 % of cDNA contained at least one ATG codon in their 5′-UTR (Suzuki et al., 2000); 41 % of transcripts had more than one uAUG and 24 % of genes had more than two uAUGs (Peri & Pandey, 2001); about 50 % of the RefSeq entries had at least one uAUG (Yamashita et al., 2003); about 44 % of 5′-UTRs had uAUGs and uORFs (Iacono et al., 2005). There are also some reports that the first AUG is skipped if it is too close to the cap structure, within 12 (Kozak, 1991) to 14 (Sedman et al., 1990) nucleotides (see the section 3.3). In this chapter, we cited a variety of statistical data on the UTRs. Because they are based on different versions or generations of sequence databases, the data vary widely (Meijer & Thomas, 2002), which is the point to be properly considered.

### 3.2 Kozak's consensus sequence

The strongest bias for initiation of translation in vertebrates is the sequence context called "Kozak's sequence", known as GCCA/GCCATGG (Kozak, 1987). The nucleotides in positions -3 (A or G) and +4 (G) are highly conserved and greatly effective for a ribosomal complex to start translation (Kozak, 1987, 2002; Matsui et al., 2007; Suzuki et al., 2001; Wang & Rothnagel, 2004). The context of an AUG codon in position -3 is the most highly conserved and functionally the most important; it is regarded as strong or optimal only when this position matches A or G, and that in position +4 is also highly conserved (Kozak, 2002). Some reports mentioned that only 0.86 % (Kozak, 1987) to 6 % (Iacono et al., 2005) of functional initiator codons lacked Kozak's sequence in positions -3 and +4, whereas 37 %

(Suzuki et al., 2000) to 46 % (Kozak, 1987) of uATGs would be skipped because of unfavorable Kozak's sequence in both of the positions. On the contrary, another report mentioned that most initiator codons were not in close agreement with Kozak's consensus sequence (Peri & Pandey, 2001).

### 3.3 The length of the 5'-UTR

The length of 5'-UTR is also effective when translation occurs from an AUG codon near the 5′ end of the mRNA (Kozak, 1991; Sedman et al., 1990). About half of ribosomes skip an AUG codon even in an optimal context if the length of 5′-UTR is less than 12 nucleotides (mentioned in the section 3.1) and this type of leaky scanning can be reduced if the length of 5′-UTR is more than or equal to 20 nucleotides (Kozak, 1991). In the traditional analysis based on incomplete 5′-UTR sequences, the distance from the 5' end to the AUG initiator codon in vertebrate mRNAs was generally from 20 and 100 nucleotides (Kozak, 1987). The previous analysis using RefSeq human mRNA sequences indicated that 85 % of 5′-UTR sequences less than 100 nucleotides contain no uAUGs (Peri & Pandey, 2001). The evidence convinced us that the first-AUG rule was widely supported in eukaryotes. In the recent analysis based on full-length 5′-UTR sequences, it is 125 nucleotides long on average (Suzuki et al., 2000) and transcriptional start sites (TSSs) vary widely (Carninci et al., 2006; Kimura et al., 2006; Suzuki et al., 2001). The average scattered length of 5'-UTR was more than 61.7 nucleotides, with a standard deviation of 19.5 nucleotides (Suzuki et al., 2001) and 52 % of the human RefSeq genes contained 3.1 TSS clusters on average (Kimura et al., 2006), which has an over 500 nucleotides interval (Fig. 4). In protein-coding genes, differentially regulated alternative TSSs are common (Carninci et al., 2006). Because the diversity of transcription initiation greatly affects the length of the 5'-UTR, there remain some doubts whether the length of the 5'-UTR contributes to the efficiency of translation initiation. There is also a report that the degree of leaky scanning is not affected by the length of 5′-UTR (Wang & Rothnagel, 2004).

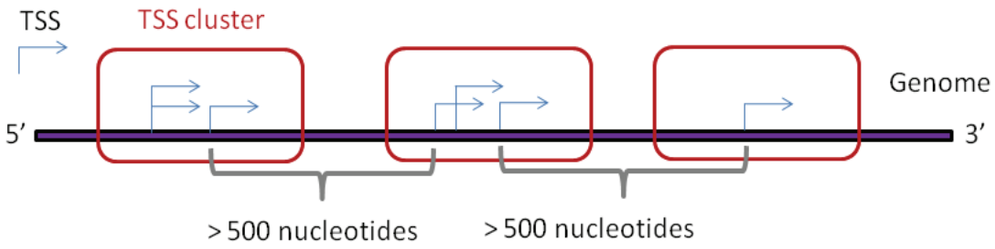



Fig. 4. The schematic representation of the 5′ ends of the TSSs.
Each TSS cluster consists of at least one TSS and has an over 500 nucleotides interval.

### 3.4 non-AUG initiator codon

In the general translation model, a non-AUG codon is considered to be ignored by ribosomes unless a downstream AUG codon is in a relatively weak context (Geballe & Morris, 1994; Kozak, 1999). In case that an upstream non-AUG codon, such as ACG, CUG or GUG, satisfies Kozak's consensus sequence, it possibly functions as an initiator of translation in addition to the first AUG initiator codon (Kozak, 1999, 2002). Besides Kozak's consensus sequence, downstream stem-and-loop and highly structured GC-rich context in the 5′-UTR could enhance translation initiation from a non-AUG codon (Kozak, 1991, 2002).

## 4. Protein identification by MS

The recent progress of proteomic methodologies based on highly sensitive liquid chromatography-tandem mass spectrometry (LC-MS/MS) technology have enabled us to identify hundreds or thousands of proteins in a single analysis.

We succeeded in the discovery of novel small proteins translated from short ORFs using direct nanoflow LC-MS/MS system (Oyama et al., 2004, 2007). Among 54 proteins less than 100 amino acids that were identified by retrieving several sequence databases with a representative search engine, Mascot (Matrix Science; http://www.matrixscience.com/), four ones were turned out to be encoded in 5′-UTRs (Oyama et al., 2004). This showed the first direct evidence of peptide products from the uORFs actually translated in human cells. In the subsequent analysis using more sophisticated two-dimensional LC system, we also discovered eight novel small proteins (Oyama et al., 2007), five of which were encoded in the 5′-UTR and three were encoded in the 3′-UTR of the corresponding mRNA. Even based on the accumulated DBTSS data, two ORFs had no putative AUG codon, which indicated the possibility that they were translated from non-AUG initiator codon. In the article above, 197 proteins less than 20 kDa were identified by Mascot. The procedure for identifying novel proteins by MS is described as follows.

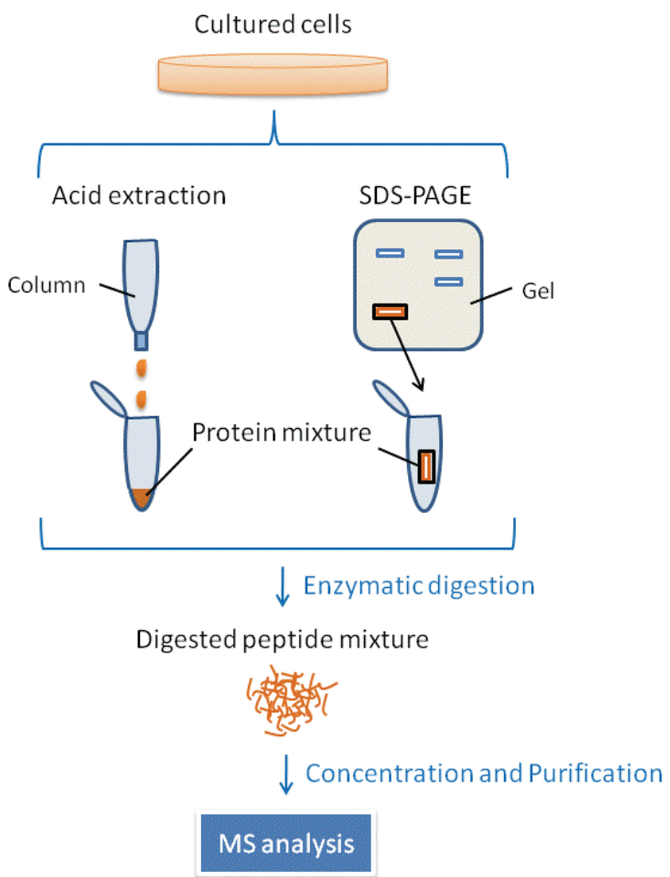


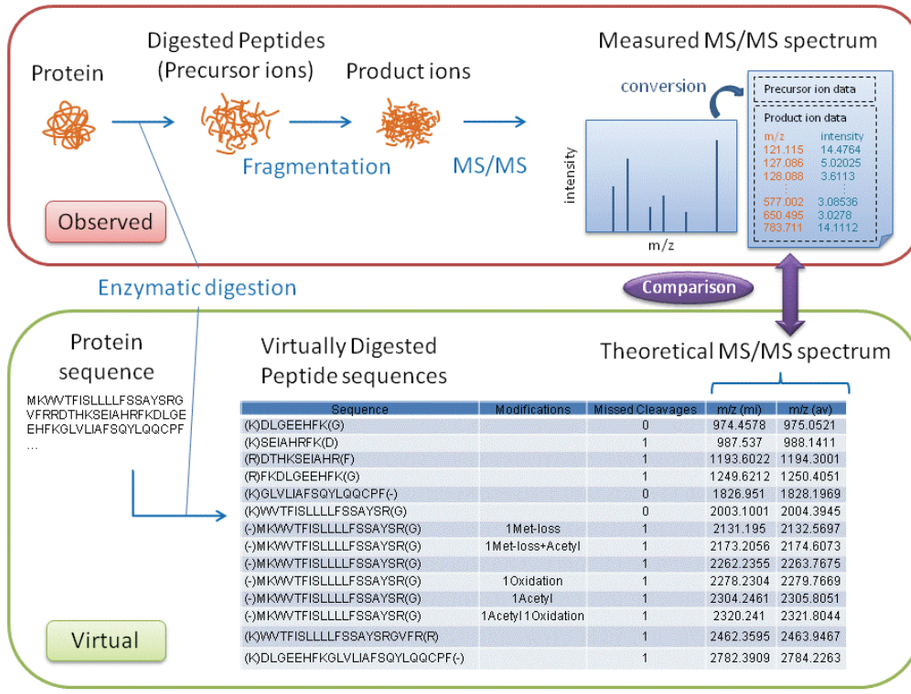

Fig. 5. The procedure for preparing samples for proteomic analyses of small proteins.

### 4.1 Materials and methods

The proteins included in cultured cell lysates were first separated according to their size. Small protein-enriched fraction through acid extraction and SDS-PAGE were treated with enzymes. In case of SDS-PAGE, the digested peptides were extracted from the gel. The samples were desalted and concentrated to introduce into the MS system. The schematic procedure is shown in Fig. 5.

### 4.2 Protein identification

The samples were analyzed using nanoflow LC-MS/MS system. The purified peptides were eluted with a linear gradient of acetonitrile and sprayed into the high-resolution tandem mass spectrometer. Acquired tandem mass (MS/MS) spectra were then converted to text files and processed against sequence databases using Mascot. Based on the principle that each peptide has a MS/MS spectrum with unique characteristics, the search engine compares measured data on precursor/product ions with those theoretically calculated from protein sequence data (Fig. 6). The MS/MS spectrum file contains mass to charge ratio (m/z) values of precursor and product ions along with their intensity. The measured spectrum lists are searched against sequence databases to identify the corresponding peptide in a statistical manner. The theoretical spectrum lists are totally dependent on the contents of sequence databases themselves.




| Sequence | Modifications | Missed Cleavages | m/z (mi) | m/z (av) |
|---|---|---|---|---|
| (K)DLGEEHFK(G) | | 0 | 974.4578 | 975.0521 |
| (K)SEIAHRFK(D) | | 1 | 987.537 | 988.1411 |
| (R)DTHKSEIAHR(F) | | 1 | 1193.6022 | 1194.3001 |
| (R)FKDLGEEHFK(G) | | 1 | 1249.6212 | 1250.4051 |
| (K)GLVLIAFSQYLQQCPF(-) | | 0 | 1826.951 | 1828.1969 |
| (K)WVTFISLLLLFSSAYSR(G) | | 0 | 2003.1001 | 2004.3945 |
| (-)MKWVTFISLLLLFSSAYSR(G) | 1Met-loss | 1 | 2131.195 | 2132.5697 |
| (-)MKWVTFISLLLLFSSAYSR(G) | 1Met-loss+Acetyl | 1 | 2173.2056 | 2174.6073 |
| (-)MKWVTFISLLLLFSSAYSR(G) | | 1 | 2262.2355 | 2263.7675 |
| (-)MKWVTFISLLLLFSSAYSR(G) | 1Oxidation | 1 | 2278.2304 | 2279.7669 |
| (-)MKWVTFISLLLLFSSAYSR(G) | 1Acetyl | 1 | 2304.2461 | 2305.8051 |
| (-)MKWVTFISLLLLFSSAYSR(G) | 1Acetyl 1Oxidation | 1 | 2320.241 | 2321.8044 |
| (K)WVTFISLLLLFSSAYSRGVFR(R) | | 1 | 2462.3595 | 2463.9467 |
| (K)DLGEEHFKGLVLIAFSQYLQQCPF(-) | | 1 | 2782.3909 | 2784.2263 |

Fig. 6. The principle of protein identification.

A search engine compares measured MS/MS spectrum lists with theoretical ones. The acquired MS/MS spectra are converted into a text file that is composed of precursor ion data and product ion data, in a format defined by the search engine. Product ion data usually consist of multiple pairs of m/z and its intensity. The theoretical m/z values are calculated virtually.

### 4.3 Finding of novel small proteins

For exploring novel small proteins, two types of sequence databases were used; one was an artificial database computationally translated from the cDNA sequences in all the reading frames and the other was an already established protein database. In order to process the comparison of the large-scale protein identification data from the two kinds of databases, several Perl scripts have been developed based on the definition that candidates of novel small proteins were identified only in the cDNA database(s) (Fig. 7). In a result datasheet using RefSeq sequences, each protein was annotated with NM numbers for the cDNA database and with NP numbers for the protein database. The Perl scripts then exchanged NM to NP numbers and evaluated them.

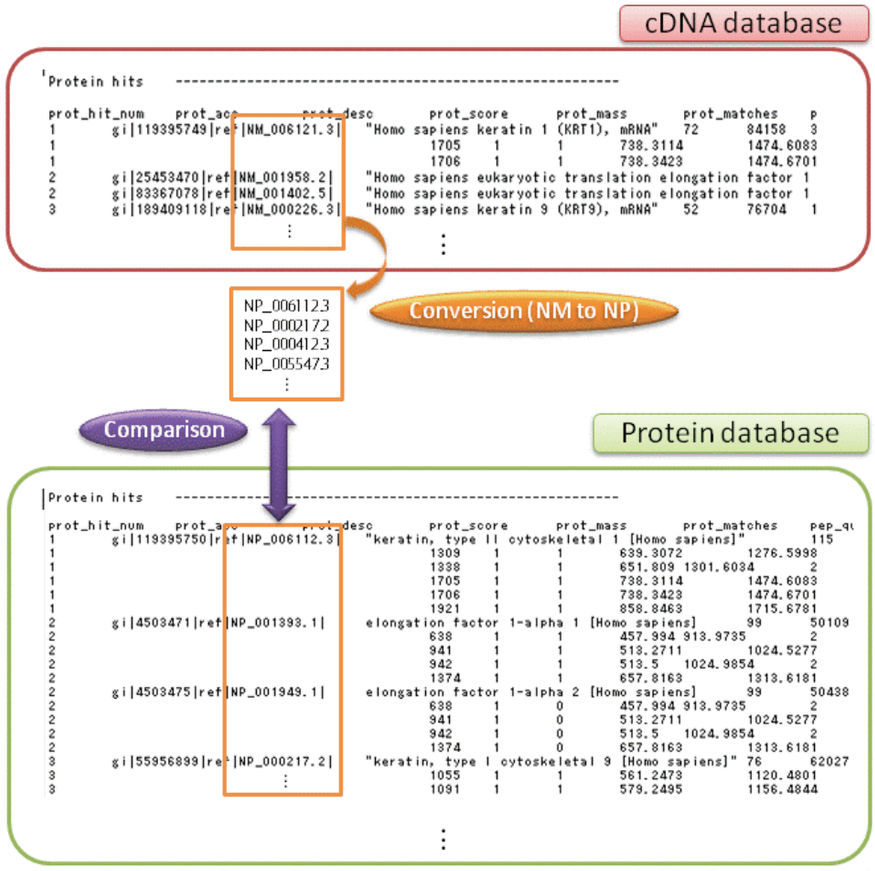



Fig. 7. The algorithm to compare the lists of search results using RefSeq cDNA and protein databases.

The proteins identified from the cDNA database are annotated with NM numbers, whereas, those from the protein database are with NP numbers. To compare these results, it is needed to exchange NM to NP numbers. The NP numbers annotated only from the cDNA database are considered to be candidates of novel proteins.

## 5. Bioinformatics approach

In order to forward MS-based identification of novel coding regions of mRNAs, MS systems, sequence databases and bioinformatics methodologies are required to improve together.

Regarding bioinformatics, two aspects seem to be demanded; one is for retrieving target proteins from an enormous size of database searching results, the other is for constructing platforms to predict novel coding sequences (CDSs).

### 5.1 Contribution of sequence databases & bioinformatics to MS-based proteomics

The recent advances in MS-based proteomics technology have enabled us to perform large-scale protein identification with high sensitivity. The accumulation of well-established sequence databases also made a great contribution to efficient identification in proteomics analyses. One of the representative databases is a specialized 5'-end cDNA database like DBTSS and the other is a series of whole genome sequence databases for various species. To investigate the mechanisms in transcriptional control, DBTSS has lately attracted considerable attention because it contains accumulated information on the transcriptional regulation of each gene (Suzuki et al., 2002, 2004; Tsuchihara et al., 2009; Wakaguri et al., 2008; Yamashita et al., 2006). Based on the accumulated data, the diverse distribution of TSSs was clearly indicated (Kimura et al., 2006; Suzuki et al., 2000, 2001). On the other hand, many whole genome sequencing projects are progressing all over the world (GOLD: Genomes Online Database; http://www.genomesonline.org/). Complement and maintenance of sequence databases for various species must help to find more novel proteins across the species. For example, there are several reports that conducted bioinformatical approaches to explore novel functional uORFs by comparing the 5'-UTR regions of orthologs based on multiple sequence alignments (Zhang & Dietrich, 2005), using ORF Finder (http://bioinformatics.org/sms/orf find.html) and a machine learning technique, inductive logic programming (ILP) with biological background knowledge (Selpi et al., 2006), or applying comparative genomics and a heuristic rule-based expert system (Cvijovic et al., 2007). Using advanced sequence databases, new protein CDSs were added as a result of the prediction by various algorithms (e.g. Hatzigeorgiou, 2002; Ota et al., 2004). Based on the well-established cDNA databases, MS could evaluate whether these CDSs are actually translated in a high-throughput manner. Construction of more detailed sequence databases will lead to detection of more novel small proteins in the presumed 5'-UTRs (Oyama et al., 2004). To make good use of those exhaustive sequence databases, bioinformatical techniques, especially data mining tools such as search engines to retrieve target proteins from an enormous size of database search results, are obviously indispensable.

### 5.2 Contribution of MS-based proteomics to sequence databases & bioinformatics

In addition to the technological progress of MS, sequence databases and data mining tools, development of other bioinformatical techniques called prediction tools, are also important. Ad-hoc algorithms for predicting new CDSs, as mentioned above, could be improved by using MS-based novel protein data. Those novel ones can be applied to play a role in a collection of supervised training data for machine learning, pattern recognition or rule-based manual approach. There is an interesting bioinformatical report which hypothesized that a uORF in the transcript down-regulates transcription of the corresponding RNA via RNA decay mechanisms (Matsui et al., 2007). They obtained human and mouse transcripts from RefSeq and UniGene (http://www.ncbi.nlm.nih.gov/unigene) and classified the transcripts into Level 0 (not containing uORF) and Level 1-3 (containing uORF). Then, they prepared the data of expression intensities and half-lives of mRNA transcripts mainly from SymAtlas (now linked to BioGPS; http://biogps.gnf.org/#goto=welcome) and Genome Research website (http://genome.cshlp.org/). Although they suggested that not only the

expression level but also the half-life of transcripts was obviously declined in the latter group, they did not demonstrate any interaction between uORFs and transcripts.
Advanced MS instruments can not only evaluate whether uORFs are actually translated but also quantify time-course changes of their expression levels. Stable isotope labeling with amino acids in cell culture (SILAC) technology enables us to quantify the changes regarding all the proteins in vivo (Oyama et al., 2009). Based on time-course changes of specific peptides, we could also hypothesize some regulatory interactions. In combination with the measurement of the dynamics of the corresponding mRNAs using microarray or reverse transcription-polymerase chain reaction (RT-PCR), transcriptional regulation by short ORFs will be analyzed at the system level.

## 6. Conclusion

Although the roles of 5'-UTR elements, especially uORFs, had been well discussed as translational regulators for the main ORFs in the biological context, whether the proteins encoded by the uORFs were translated had not been approached for a long time. We first unraveled the mystery by demonstrating the existence of novel protein products defined by these ORFs using advanced proteomics technology. Thanks to the progress of nanoLC-MS/MS-based shotgun proteomics strategies, thousands of proteins can now be identified from protein mixtures such as cell lysates. Some of the presumed UTRs are no longer "untranslated", and other noncoding transcripts are no longer "noncoding". One of the novel small proteins revealed in our analysis was indeed defined by a short transcript variant generated by utilization of the downstream alternative promoters (Oyama et al., 2007). Alternative uses of diverse transcription initiation, splicing and translation start sites could increase the complexity of short protein-coding regions and MS -based annotation of these novel small proteins will enable us to perform a more detailed analysis of the real outline of the proteome, along with the translational regulation by the diversified short ORFeome systematically.

## 7. References

Altmann, M. & Trachsel, H. (1993). Regulation of translation initiation and modulation of cellular physiology. *Trends in Biochemical Sciences*, Vol. 18, No. 11, pp. 429-432, Online ISSN 0167-7640; 0376-5067, Print ISSN 0968-0004.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engstrom, P.G., Frith, M.C., Forrest, A.R.R., Alkema, W.B., Tan, S.L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S.M., Wells, C.A., Orlando, V., Wahlestedt, C., Liu, E.T., Harbers, M., Kawai, J., Bajic, V.B., Hume, D.A. & Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, Vol. 38, No. 6, pp. 626-635, Online ISSN 1546-1718, Print ISSN 1061-4036.

Cvijovic, M., Dalevi, D., Bilsland, E., Kemp, G.J.L. & Sunnerhagen, P. (2007). Identification of putative regulatory upstream ORFs in the yeast genome using heuristics and evolutionary conservation. *BMC Bioinformatics*, Vol. 8, Article No. 295, Online ISSN 1471-2105.

Diba, F., Watson, C.S. & Gametchu, B. (2001). 5′UTR Sequences of the Glucocorticoid Receptor 1A Transcript Encode a Peptide Associated With Translational Regulation of the Glucocorticoid Receptor. *Journal of Cellular Biochemistry*, Vol. 81, No. 1, pp. 149-161, Online ISSN 1097-4644, Print ISSN 0730-2312.

Geballe, A.P. & Morris, D.R. (1994). Initiation codons within 5′-leaders of mRNAs as regulators of translation. *Trends in Biochemical Sciences*, Vol. 19, No. 4, pp. 159-164, Online ISSN 0167-7640; 0376-5067, Print ISSN 0968-0004.

Hatzigeorgiou, A.G. (2002). Translation initiation start prediction in human cDNAs with high accuracy. *Bioinformatics*, Vol. 18, No. 2, pp. 343-350, Online ISSN 1460-2059, Print ISSN 1367-4803.

Iacono, M., Mignone, F. & Pesole, G. (2005). uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene*, Vol. 349, pp. 97-105, Online ISSN 1879-0038, Print ISSN 0378-1119.

Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H., Ishii, S., Sugiyama, T., Saito, K., Isono, Y., Irie, R., Kushida, N., Yoneyama, T., Otsuka, R., Kanda, K., Yokoi, T., Kondo, H., Wagatsuma, M., Murakawa, K., Ishida, S., Ishibashi, T., Takahashi-Fujii, A., Tanase, T., Nagai, K., Kikuchi, H., Nakai, K., Isogai, T. & Sugano, S. (2006). Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Research*, Vol. 16, No. 1, pp. 55-65, Online ISSN 1549-5469, Print ISSN 1088-9051.

Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*, Vol. 15, No. 20, pp. 8125-8148, Online ISSN 1362-4962, Print ISSN 0305-1048.

Kozak, M. (1989). The Scanning Model for Translation: An Update. *The Journal of Cell Biology*, Vol. 108, No. 2, pp. 229-241, Online ISSN 1540-8140, Print ISSN 0021-9525.

Kozak, M. (1991). Structural Features in Eukaryotic mRNAs That Modulate the Initiation of Translation. *The Journal of Biological Chemistry*, Vol. 266, No. 30, pp. 19867-19870, Online ISSN 1083-351X, Print ISSN 0021-9258.

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene*, Vol. 234, No. 2, pp. 187-208, Online ISSN 1879-0038, Print ISSN 0378-1119.

Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, Vol. 299, No. 1-2, pp. 1-34, Online ISSN 1879-0038, Print ISSN 0378-1119.

Maruyama, K. & Sugano, S. (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, Vol. 138, No. 1-2, pp. 171-174, Online ISSN 1879-0038, Print ISSN 0378-1119.

Matsui, M., Yachie, N., Okada, Y., Saito, R. & Tomita, M. (2007). Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS Letters*, Vol. 581, No. 22, pp. 4184-4188, Online ISSN 1873-3468, Print ISSN 0014-5793.

Meijer, H.A. & Thomas, A.A.M. (2002). Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. *Biochemical Journal*, Vol. 367, No. 1, pp. 1-11, Online ISSN 1470-8728, Print ISSN 0264-6021.

Morris, D.R. & Geballe, A.P. (2000). Upstream Open Reading Frames as Regulators of mRNA Translation. *Molecular and Cellular Biology*, Vol. 20, No. 23, pp. 8635-8642, Online ISSN 1098-5549, Print ISSN 0270-7306.

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., Kimura, K., Makita, H., Sekine, M., Obayashi, M., Nishi, T., Shibahara, T., Tanaka, T., Ishii, S., Yamamoto, J., Saito, K., Kawai, Y., Isono, Y.,

Nakamura, Y., Nagahari, K., Murakami, K., Yasuda, T., Iwayanagi, T., Wagatsuma, M., Shiratori, A., Sudo, H., Hosoiri, T., Kaku, Y., Kodaira, H., Kondo, H., Sugawara, M., Takahashi, M., Kanda, K., Yokoi, T., Furuya, T., Kikkawa, E., Omura, Y., Abe, K., Kamihara, K., Katsuta, N., Sato, K., Tanikawa, M., Yamazaki, M., Ninomiya, K., Ishibashi, T., Yamashita, H., Murakawa, K., Fujimori, K., Tanai, H., Kimata, M., Watanabe, M., Hiraoka, S., Chiba, Y., Ishida, S., Ono, Y., Takiguchi, S., Watanabe, S., Yosida, M., Hotuta, T., Kusano, J., Kanehori, K., Takahashi-Fujii, A., Hara, H., Tanase, T., Nomura, Y., Togiya, S., Komai, F., Hara, R., Takeuchi, K., Arita, M., Imose, N., Musashino, K., Yuuki, H., Oshima, A., Sasaki, N., Aotsuka, S., Yoshikawa, Y., Matsunawa, H., Ichihara, T., Shiohata, N., Sano, S., Moriya, S., Momiyama,H., Satoh, N., Takami, S., Terashima, Y., Suzuki, O., Nakagawa, S., Senoh, A., Mizoguchi, H., Goto, Y., Shimizu, F., Wakebe, H., Hishigaki, H., Watanabe, T., Sugiyama, A., Takemoto, M., Kawakami, B., Yamazaki, M., Watanabe, K., Kumagai, A., Itakura, S., Fukuzumi, Y., Fujimori, Y., Komiyama, M., Tashiro, H., Tanigami, A., Fujiwara, T., Ono, T., Yamada, K., Fujii, Y., Ozaki, K., Hirao, M., Ohmori, Y., Kawabata, A., Hikiji, T., Kobatake, N., Inagaki, H., Ikema, Y., Okamoto, S., Okitani, R., Kawakami, T., Noguchi, S., Itoh, T., Shigeta, K., Senba, T., Matsumura, K., Nakajima, Y., Mizuno, T., Morinaga, M., Sasaki, M., Togashi, T., Oyama, M., Hata, H., Watanabe, M., Komatsu, T., Mizushima-Sugano, J., Satoh, T., Shirai, Y., Takahashi, Y., Nakagawa, K., Okumura, K., Nagase, T., Nomura, N., Kikuchi, H., Masuho, Y., Yamashita, R., Nakai, K., Yada, T., Nakamura, Y., Ohara, O., Isogai, T. & Sugano, S. (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genetics*, Vol. 36, No. 1, pp. 40-45, Online ISSN 1546-1718, Print ISSN 1061-4036.

Oyama, M., Itagaki, C., Hata, H., Suzuki, Y., Izumi, T., Natsume, T., Isobe, T. & Sugano, S. (2004). Analysis of Small Human Proteins Reveals the Translation of Upstream Open Reading Frames of mRNAs. *Genome Research*, Vol. 14, No. 10B, pp. 2048-2052, Online ISSN 1549-5469, Print ISSN 1088-9051.

Oyama, M., Kozuka-Hata, H., Suzuki, Y., Semba, K., Yamamoto, T. & Sugano, S. (2007). Diversity of Translation Start Sites May Define Increased Complexity of the Human Short ORFeome. *Molecular & Cellular Proteomics*, Vol. 6, No. 6, pp. 1000-1006, Online ISSN 1535-9484, Print ISSN 1535-9476.

Oyama, M., Kozuka-Hata, H., Tasaki, S., Semba, K., Hattori, S., Sugano, S., Inoue, J. & Yamamoto, T. (2009). Temporal Perturbation of Tyrosine Phosphoproteome Dynamics Reveals the System-wide Regulatory Networks. *Molecular & Cellular Proteomics*, Vol. 8, No. 2, pp. 226-231, Online ISSN 1535-9484, Print ISSN 1535-9476.

Peabody, D.S., Subramani, S. & Berg, P. (1986). Effect of Upstream Reading Frames on Translation Efficiency in Simian Virus 40 Recombinants. *Molecular and Cellular Biology*, Vol. 6, No. 7, pp. 2704-2711, Online ISSN 1098-5549, Print ISSN 0270-7306.

Peri, S. & Pandey, A. (2001). A reassessment of the translation initiation codon in vertebrates. *Trends in Genetics*, Vol. 17, No. 12, pp. 685-687, Print ISSN 0168-9525.

Rastinejad, F. & Blau, H.M. (1993). Genetic Complementation Reveals a Novel Regulatory Role for 3′ Untranslated Regions in Growth and Differentiation. *Cell*, Vol. 72, No. 6, pp. 903-917, Online ISSN 1097-4172, Print ISSN 0092-8674.

Sedman, S.A., Gelembiuk, G.W. & Mertz, J.E. (1990). Translation Initiation at a Downstream AUG Occurs with Increased Efficiency When the Upstream AUG Is Located Very Close to the 5' Cap. *Journal of Virology*, Vol. 64, No. 1, pp. 453-457, Online ISSN 1098-5514, Print ISSN 0022-538X.

Selpi, Bryant, C.H., Kemp, G.J.L. & Cvijovic, M. (2006). A First Step towards Learning which uORFs Regulate Gene Expression. *Journal of Integrative Bioinformatics*, Vol. 3, No. 2, ID. 31, Online ISSN 1613-4516.

Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. & Sugano, S. (1997). Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene*, Vol. 200, No. 1-2, pp. 149-156, Online ISSN 1879-0038, Print ISSN 0378-1119.

Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H., Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., Isogai, T., Suyama, A. & Sugano, S. (2000). Statistical Analysis of the 5′Untranslated Region of Human mRNA Using "Oligo-Capped" cDNA Libraries. *Genomics,* Vol. 64, No. 3, pp. 286-297, Online ISSN 1089-8646, Print ISSN 0888-7543.

Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S., Okubo, K., Sakaki, Y., Nakamura, Y., Suyama, A. & Sugano, S. (2001). Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO reports*, Vol. 2, No. 5, pp. 388-393, Online ISSN 1469-3178, Print ISSN 1469-221X.

Suzuki, Y., Yamashita, R., Nakai, K. & Sugano, S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Research*, Vol. 30, No. 1, pp. 328-331, Online ISSN 1362-4962, Print ISSN 0305-1048.

Suzuki, Y., Yamashita, R., Sugano, S. & Nakai, K. (2004). DBTSS: DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Research*, Vol. 32 (suppl 1), Database issue D78-D81, Online ISSN 1362-4962, Print ISSN 0305-1048.

Tsuchihara, K., Suzuki, Y., Wakaguri, H., Irie, T., Tanimoto, K., Hashimoto, S., Matsushima, K., Mizushima-Sugano, J., Yamashita, R., Nakai, K., Bentley, D., Esumi, H. & Sugano, S. (2009). Massive transcriptional start site analysis of human genes in hypoxia cells. *Nucleic Acids Research*, Vol. 37, No. 7, pp. 2249-2263, Online ISSN 1362-4962, Print ISSN 0305-1048.

Vilela, C. & McCarthy, J.E.G. (2003). Regulation of fungal gene expression via short open reading frames in the mRNA 5′untranslated region. *Molecular Microbiology,* Vol. 49, No. 4, pp. 859-867, Online ISSN 1365-2958, Print ISSN 0950-382X.

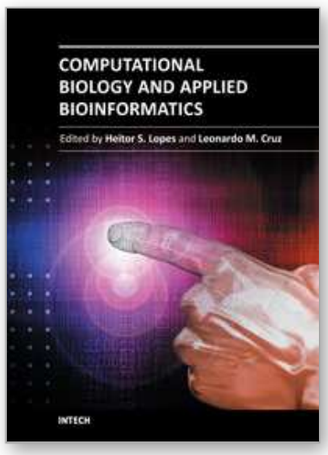
Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S. & Nakai, K. (2008). DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Research*, Vol. 36 (suppl 1), Database issue D97-D101, Online ISSN 1362-4962, Print ISSN 0305-1048.

Wang, X-Q. & Rothnagel, J.A. (2004). 5′-Untranslated regions with multiple upstream AUG codons can support low-level translation via leaky scanning and reinitiation. *Nucleic Acids Research*, Vol. 32, No. 4, pp. 1382-1391, Online ISSN 1362-4962, Print ISSN 0305-1048.

Yamashita, R., Suzuki, Y., Nakai, K. & Sugano, S. (2003). Small open reading frames in 5′ untranslated regions of mRNAs. *Comptes Rendus Biologies*, Vol. 326, No. 10-11, pp. 987-991, Online ISSN 1768-3238, Print ISSN 1631-0691.

Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K. & Sugano, S. (2006). DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. *Nucleic Acids Research*, Vol. 34 (suppl 1), Database issue D86-D89, Online ISSN 1362-4962, Print ISSN 0305-1048.

Zhang, Z. & Dietrich, F.S. (2005). Identification and characterization of upstream open reading frames (uORF) in the 5′ untranslated regions (UTR) of genes in Saccharomyces cerevisiae. *Current Genetics*, Vol. 48, No. 2, pp. 77-87 Online ISSN 1432-0983, Print ISSN 0172-8083.




**Computational Biology and Applied Bioinformatics**

Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4
Hard cover, 442 pages
**Publisher** InTech
**Published online** 02, September, 2011
**Published in print edition** September, 2011

Nowadays it is difficult to imagine an area of knowledge that can continue developing without the use of computers and informatics. It is not different with biology, that has seen an unpredictable growth in recent decades, with the rise of a new discipline, bioinformatics, bringing together molecular biology, biotechnology and information technology. More recently, the development of high throughput techniques, such as microarray, mass spectrometry and DNA sequencing, has increased the need of computational support to collect, store, retrieve, analyze, and correlate huge data sets of complex information. On the other hand, the growth of the computational power for processing and storage has also increased the necessity for deeper knowledge in the field. The development of bioinformatics has allowed now the emergence of systems biology, the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of a living being. This book presents some theoretical issues, reviews, and a variety of bioinformatics applications. For better understanding, the chapters were grouped in two parts. In Part I, the chapters are more oriented towards literature review and theoretical issues. Part II consists of application-oriented chapters that report case studies in which a specific biological problem is treated with bioinformatics tools.



**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen

IntechOpen