Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography B, 787 (2003) 91–99
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

Review

Protein database of *Caenorhabditis elegans*

Hiroyuki Kaji, Toshiaki Isobe*

Department of Chemistry, Graduate School of Science, Tokyo Metropolitan University, Minami-osawa 1-1, Hachioji, Tokyo 192-0397, Japan

Abstract

Whole genome sequencing of the free-living nematode *Caenorhabditis elegans* is a prominent achievement in genomics and uncovers the existence of enormous known and unknown gene products. Characterization and linking of all gene products are the next challenging theme of biology. Genome-wide researches are already progressing on *C. elegans* and the fruits of these efforts are accessible through the internet. To link the sequence–function relationship, proteomic research has been applied to provide comprehensive information of the worm proteins. In addition to 2-dimensional gel electrophoresis for visualization of the proteome, recent advances in liquid chromatography (LC)-based technologies have allowed the large-scale analysis of proteins and are at cutting-edge of high-throughput analysis of focused proteome.

Keywords: Reviews; *Caenorhabditis elegans*; Proteomics; Proteins

Contents

*Corresponding author.
E-mail address: isobe-toshiaki@c.metro-u.ac.jp (T. Isobe).

1570-0232/02/$ – see front matter
doi:10.1016/S1570-0232(02)00338-0

1. Introduction

Whole genome sequence of an organism provides extremely important information, in which amino acid sequence of all gene products are documented. The information seems like a complete set of puzzle pieces. A final purpose of life science is to complete a magnificent picture that accounts for the complicated biological system by connecting the puzzle pieces. Proteomics is a field that aims to supply the comprehensive information to elucidate each gene function through the systematic and large-scale analysis of the proteins. The multicellular organism is composed of multiple kinds of cells. The protein complement which constitutes each, i.e. proteome [1], is specific to the cell and has very dynamic characteristics, which changes every moment with various conditions such as an external stimulation, in contrast to the genome, which is essentially identical in nearly all cells. The structure and the quantity of the protein species and their changes under various conditions must reflect the biological activity of the cell. Now, it has become widely accepted that there is little correlation between the amount of mRNA and that of protein, and the activity of the protein is controlled often by various post-translational modifications. In addition, body fluids such as serum, cerebrospinal fluid, and urine, which would be used for diagnostics, do not contain mRNA. These facts become a basis of the idea in which direct analysis of protein is essential for understanding the mechanisms of the biological action.

The fundamental approach of proteomics is that it identifies and characterizes individual protein component by mass spectrometric analysis, after separation of the proteins obtained from biological specimens. The main method selected for the protein separation is two-dimensional gel electrophoresis (2-DE), which has the highest resolution among current separation techniques available. The mass spectrometry (MS) is chosen as a detection–analysis system because of its high sensitivity and the ability to generate strict information on the basis of protein structure. Then the protein is identified by correlation with sequence database using the constraining parameters such as a set of peptide masses or partial amino acid sequence (sequence tag) with its parental peptide mass. This means that it is impossible to identify a protein of which the amino acid sequence is not known or not included in the sequence database. Thus proteomics depends on the information from genomics and vice versa, so the two fields might be complementary. In model organisms presently, information of probable candidate proteins could be obtained by searching nucleotide sequence of the gene or predicted amino acid sequence database. Confirmation by the extensive EST analysis leads to the construction of more accurate amino acid sequence databases.

The nematode *Caenorhabditis elegans* (*C. elegans*) is a good material for proteome analysis. This free-living worm of about 1 mm body length is the first multicellular model animal of which genome sequence was determined [2]. After the publication of its sequence, a partial undecoded region was also determined and the predicted genes had been confirmed by EST information. The current amino acid sequence database contains 20 219 entries (Wormpep66, 27/09/01) [3]. In addition, it has remarkable features in that all cell lineages from a single fertilized egg to adult worm (959 somatic cells (nuclei) for hermaphrodite) and the circuit of 302 nerve cells is completely elucidated [4]. Furthermore, convenient techniques of gene knockout have been developed and many mutants are isolated. RNA mediated interference (RNAi) is also a powerful tool to analyze the effect of expression interference of the gene. Under the circumstances, many genome-wide analyses have already been started on the worm, e.g. EST project, RNAi, Transcriptome (DNA chip) analysis, yeast 2-hybrid assay, in situ hybridization, etc. Proteomic research would also contribute to its part of the overall understanding of the worm. In this review, the protein database essential for proteomic study of the nematode are introduced as summarized in Fig. 1, along the flow of genetic information from DNA to protein and their association. Actual proteome researches are outlined with our current attempts.

2. Protein database of *C. elegans*

2.1. Proteome predicted from genome sequence

2.1.1. Protein database: Wormpep

The nematode chromosome is composed of 5 autosomes with 1 sex chromosome, and their overall

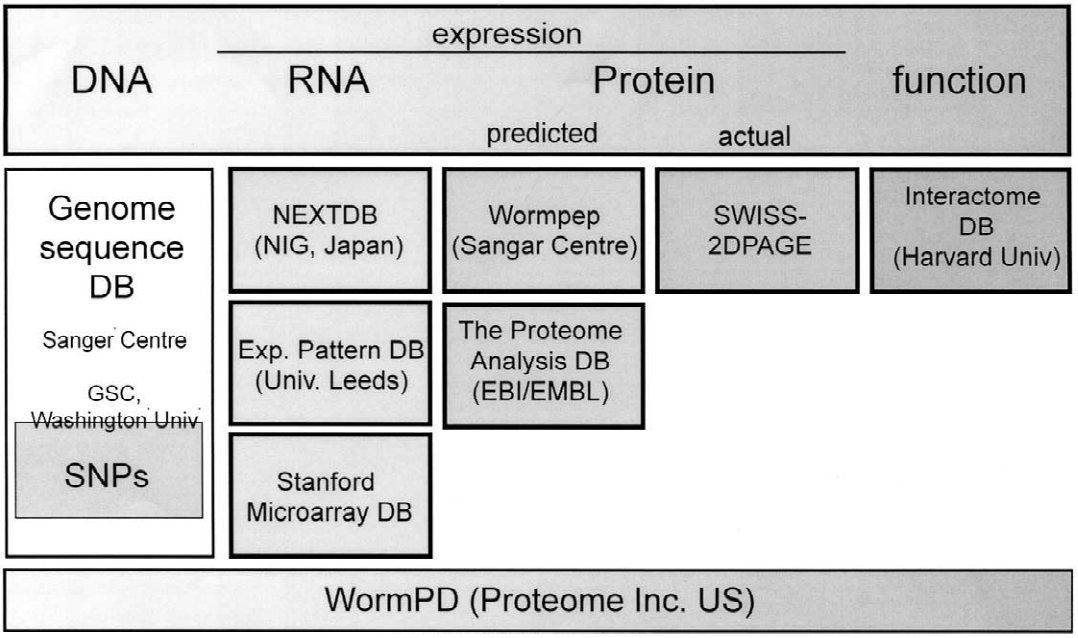

Fig. 1. Databases of *C. elegans*. Useful databses introduced in this review are presented along the flow of genetic information from DNA to protein and their association.

length is about 100 mega base pairs. The sequencing of the genome was achieved by the collaborative research of the Sanger centre (UK) and the Genome Sequencing Center at Washington University (USA). From the base sequence, an individual gene is predicted and each deduced amino acid sequence is accumulated in the protein database ''Wormpep'' [3]. The database has been updated at frequent intervals, where the current (27/09/2001) version wormpep66 contains 20 219 entries. Among them, 675 splice variants were included. Each data contains the CDS (coding sequence) identifier; the Wormpep accession number; a locus name (gene name); a keyword with the brief annotation (protein name); a TREMBL or Swiss-Prot accession number; and a protein_id followed by amino acid sequence. Using these parameters alone or in combination, the database could be searched. In addition to the most current content, previous databases and their history could also be downloaded. Fig. 2 shows a 2-D display of all predicted proteins contained in Wormpep66, where molecular mass (M_r; vertical axis) and isoelectric point (p*I*, horizontal) of each is calculated from their amino acid sequences using an in-house program based on the program of Skoog and Wichman [5] using the fundamental Henderson–Hasselbach equation. The vertical axis is in the logarithmic scale to imagine a 2-DE profile. Post-translational modification was not considered for the calculation. Average number of amino acid residues of all protein precursors is 440 with a corresponding molecular mass of 50 kDa. In the 2-D display, the predicted worm proteins show two clusters with a clear boundary around pH7 as in those of bacteria such as *E. coli* [6] and *H. influenzae* [7]. This may relate to the solubility or stability of proteins in an intracellular environment where proteins have been evolved, while actual reason is unclear.

Thus, the worm has quite a wide variety of proteins even in their size and p*I*, however, once a partial amino acid sequence of *C. elegans* protein is obtained by any way, and then probable candidate could be obtained by searching wormpep database using BLAST. Since there are a few researches on *C. elegans* at protein level, this predicted protein database is a pivotal site for sequence-based research.

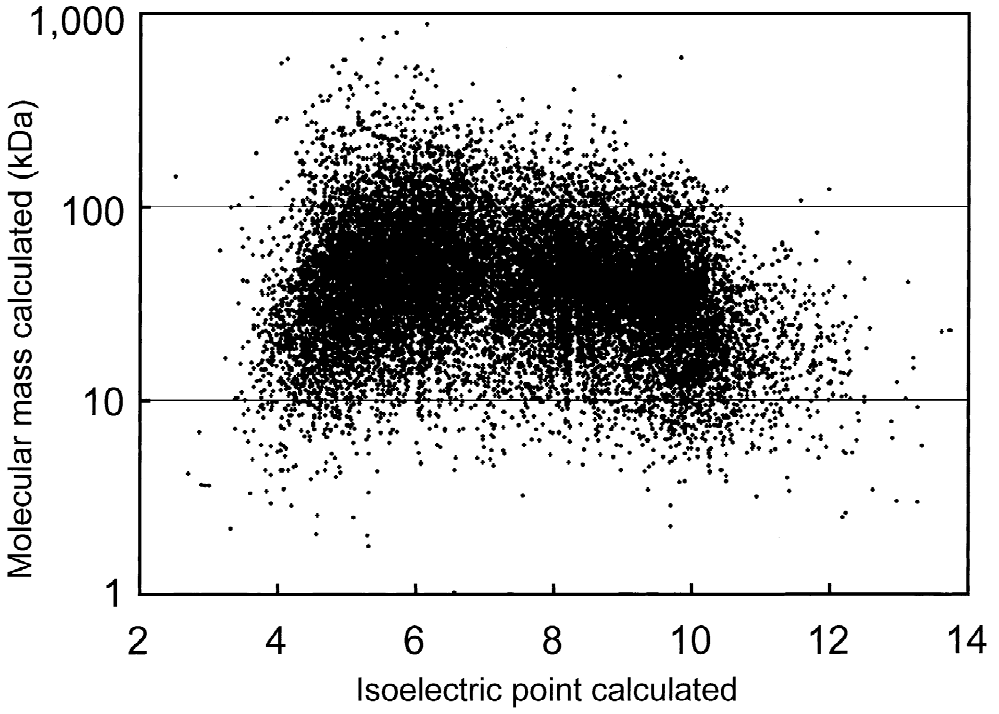

Fig. 2. 2D-display of the predicted worm proteome. Isoelectric point and molecular mass of all *C.elegans* proteins were calculated from their amino acid sequences obtained from protein database Wormpep66 (20 219 entries) constructed by The Sanger Centre. The *y*-axis is presented as a logarithmic scale.

2.1.2. Proteome Analysis database

The Proteome Analysis database managed by EBI (European Bioinformatics Institute: the EMBL Outstation) provides comprehensive statistical analyses of the proteome predicted from genome-sequenced organisms [8]. The analyzed data derived from 7 eukaryotes, 10 archaea, and 39 bacteria are enumerated. In the database on *C. elegans* proteins, a non-redundant proteome set of 17 946 entries is included. The contents are statistic data compiled by InterPro and CluSTr [9], and the comparative results with those of other eukaryotes such as human, *Drosophila melanogaster*, *Saccharomyces celevisiae* and *Arabidopsis thaliana*. InterPro database provides information of functional sites, domains, and families of proteins. On *C. elegans*, 11 842 proteins (66%) are matched with InterPro entries and the remaining 34% have no portion homologous to that of known proteins. The worm proteome contains a major protein family that is common to the eukaryotes, such as protein kinase and C2H2 type zinc-finger. Rhodopsin-like GPCR (G-protein coupled receptor) superfamily, C4-type steroid receptor zinc finger, ligand-binding domain of nuclear hormone receptor and C-type lectin domain are the major complements in the worm, as is characteristic in multicellular organisms. The most remarkable feature of the worm protein is a presence of worm specific protein families or domains designated as ''nematode 7-TM chemoreceptor'' and ''nematode cuticle collagen N-terminal domain''. From studies of statistical analyses of integral membrane proteins, *Caenorhabditis elegans* and *Homo sapiens* seemed to have a preference for proteins with seven transmembrane segments. This information is also interesting, in view of evolutionary events such as gene-duplication and diversification. The worm is equipped with many protein complements for intercellular signaling; therefore, it is suitable for proteomic studies of various biological events characteristic of multicellular organisms, such as development and nervous system function.

2.1.3. Proteomic research of C. elegans

In contrast to extensive genetic information, the knowledge of the worm protein is quite limited. However, systematic analysis of naturally occurring proteins should be an important approach to confirm the existence of the predicted protein, to know their relative quantities, and particularly to obtain structural information including post-translational modifications. These results could provide insight on the mechanisms for controlling each biological activity. Thus far, several 2-DE separations of *C. elegans* protein were reported, but with earlier applications, it was difficult to identify separated proteins on the 2-D gels due to low sensitivity of available technique [10,11]. Employing immobilized pH gradient (IPG) gel in the first dimension and advancements in detection system, i.e. Edman-based microsequencing and mass spectrometry, made protein identification possible and the number of identified proteins had increased; 12 spots by microsequencing [12] and 69 spots by in-gel digestion followed by MALDI–MS [13], as well as 286 spots on multiple 2D-gels by the same manners [14]. Several 2-DE profiles of the worm protein are located on the database of WORLD-2DPAGE (http://www.expasy.ch/ch2d/2d-index.html). As expected, major proteins identified (152 proteins in Ref. [14]) include many metabolic enzymes for energy generation (49 proteins), proteins for synthesis, folding, and degradation of protein (21), and cell structure proteins (7), which are called ‘‘House-keeping proteins’’. Cellular roles of the identified proteins are compared with those of over-all predicted gene products. These roles are defined in the database WormPD (that will be introduced in Section 2.1.4.1) (Fig. 3). Although the worm genome contains many genes for signal transduction, which occupy about 30% of functionally categorized proteins, no such proteins are detected in major proteins identified. To detect the minor proteins by current technologies, prefractionation of crude extract is necessary. Here, it seems important to note that the major proteins contain considerable

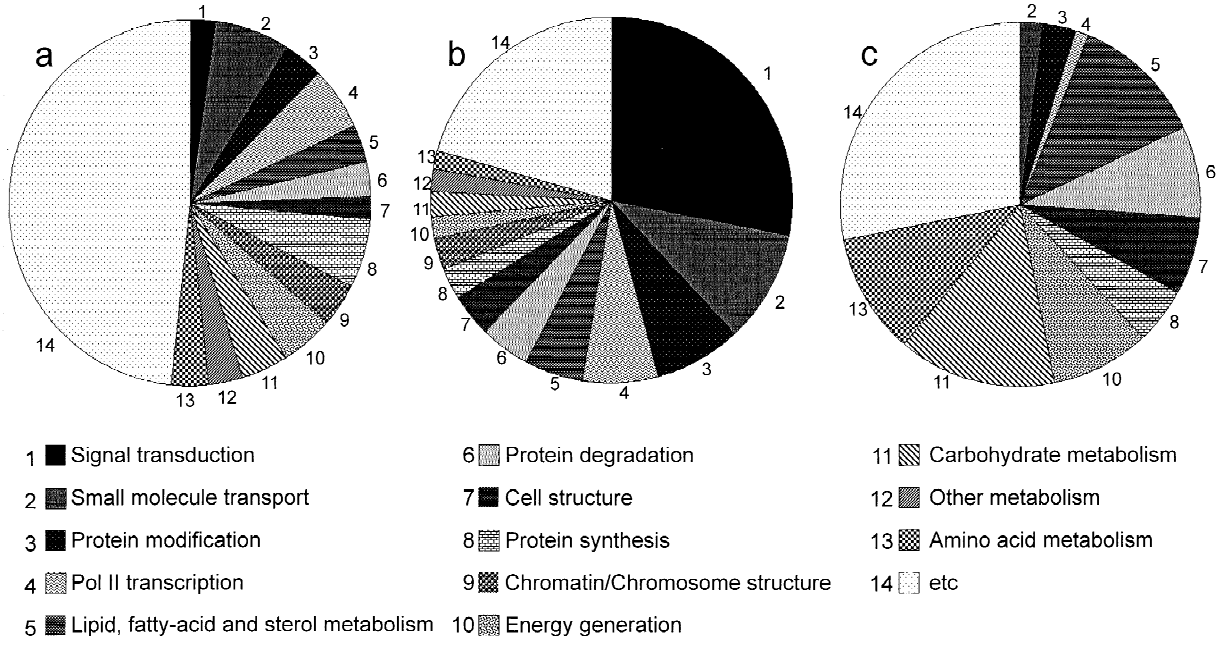

Fig. 3. Distribution of the cellular roles of categorized proteins of yeast *S. cerevisiae* (a: 6026 proteins), *C. elegans* (b: 5793 proteins) and the worm proteins identified through 2D-PAGE (c: 152 proteins) [14]. Each role was obtained from the databases YPD (a) and WormPD (b and c), respectively, maintained by Proteome, USA [18,19]. The worm has many proteins for signal transduction and protein modification compared with those of yeast, suggesting that those proteins are equipped for intercellular communication. However, no proteins for signal transduction were detected in the identified proteins through 2-DE.

hypothetical proteins. This is one of the significance of actual proteome analysis in an open-system. In addition, many structural incompatibilities have been observed between detected spot positions and their calculated values. Molecular mass (M_r) of actual protein [13] is apt to be smaller than that of the calculated one (Fig. 4A). It is known that many proteins after synthesis are cleaved for localization and activation purposes. The removal of signal peptide or pro-region leads to a decrease in M_r. Conversely, attachment of carbohydrate increases the M_r of protein and particularly on SDS–PAGE the effect may emerge remarkably. Isoelectric points (p*I*) of many spots are also shifted toward the acidic side (Fig. 4B). These tendencies suggest that various post-translational modifications, e.g. phosphorylation of side chain hydroxyl group(s), acylation of N-terminal amino group, deamidation, and removal of signal peptide including basic residue(s), have occurred frequently, and p*I* shift toward the basic side by, for example, esterification of carboxyl group may also happen. These modifications are recognized to be very important to control the biological activity of proteins. In fact, proteins identified by MS are predicted to have some fragments, which have modification such as acetylation (+42 u), oxidation (+16), deamidation (+1), dehydroxylation (−18), methylation (+14), and rarely phosphorylation (+80). As described above, the knowledge of the worm protein is limited; therefore, these modifications should be confirmed by other methods. Thus, systematic and large-scale analysis of proteome is one of the key approaches to provide information on structural aspects of each protein.

Current 2-DE based proteomics have several technical limitations, e.g. the identification of those proteins requires time-consuming steps to digest the proteins within the gel, extract individual protein ‘‘spots’’ from the gel, and prepare the extracted samples for MS analysis. Furthermore, some cellular proteins are rarely isolated in 2D-PAGE studies, such as membrane-associated proteins and those with extremes in either p*I* or molecular mass as well as minor proteins. We had developed multi-dimensional liquid chromatography (LC) to separate complex protein mixture automatically with high resolution

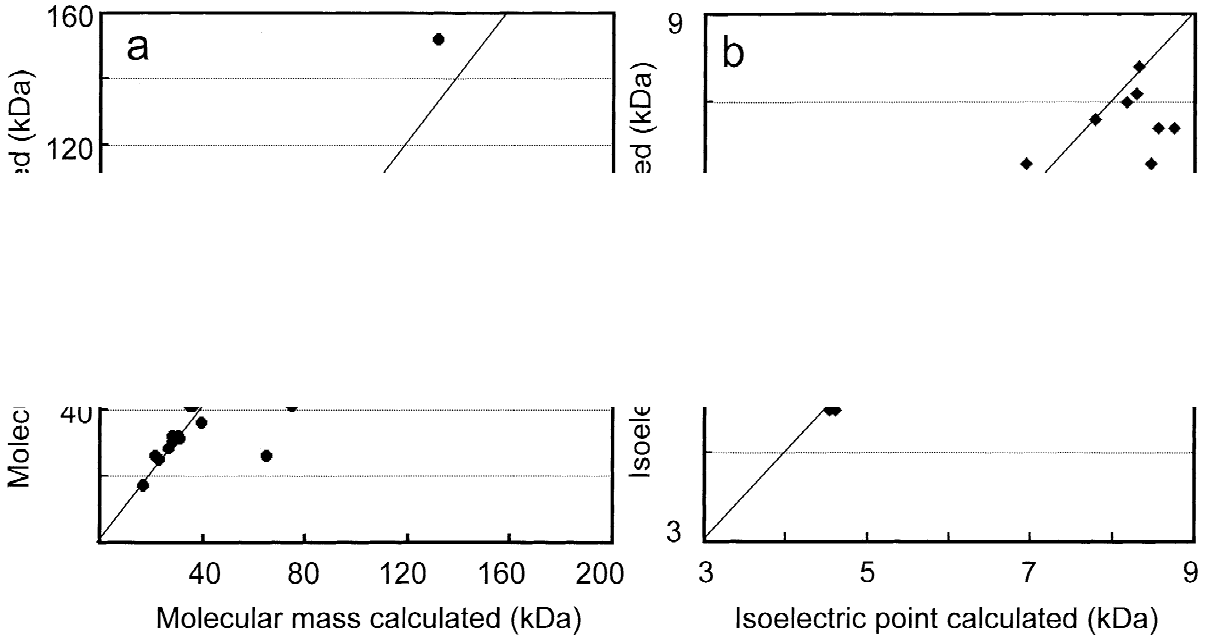

Fig. 4. Differences in molecular mass (M_r: a) and isoelectric point (p*I*: b) between calculated and observed values [13]. The differences in M_r suggest post-translational processing such as proteolytic processing and attachment of carbohydrates and those in p*I* may be due to phosphorylation, acylation, deamidation, and so forth.

and reproducibility and applied the system to separate brain proteins [15]. Recently, an automated micro-scale 2-dimensional LC system (2D-μLC) directly connected to electrospray ionization (ESI)-Q-TOF (time-of-flight) hybrid mass spectrometer for on-line analysis was constructed (Yamauchi, unpublished data). Such a LC–MS–MS has shown to be powerful to identify many hundreds of proteins of a focused proteome, such as multi-protein complexes, functional membrane domains and organelles. As a focused proteome of *C. elegans*, Concanavalin A-bound glycosylated proteins were chosen and their tryptic peptides were applied to the 2D-μLC–MS–MS system. Over 2800 peptides corresponding to independent 333 proteins were identified (Kaji, unpublished data). The identified proteins contained 194 categorized proteins and the rest were hypothetical. Among the 194 proteins, 105 had information on their cellular roles. The number of metabolic enzymes found was 20 (19%). The most remarkable cellular role represented was protein degradation, in which 24 proteases and protease inhibitors were assigned, and an additional 10 were peptidases with unknown role. This result suggests that many proteases and their inhibitors would be secreted or localized in lysosome through endoplasmic reticulum, tagging with carbohydrates. Furthermore, proteins for signal transduction (6 proteins), differentiation (1), and cell cycle control (1) were detected in the glycoproteins. Such a hyphenated system combining LC-based separation and ESI-equipped mass spectrometer that can generate sequence-tag, e.g. Q-Tof hybrid and ion-trap mass spectrometers, must be robust tools for high-throughput identification of focused proteome such as large protein complex, biological machinery, and organelle [16,17].

2.1.4. Database for characterization of C. elegans protein

Once protein candidates of interest are identified, the proteins must be characterized from various perspectives. In this section, useful databases for the characterization of the gene product are introduced.

2.1.4.1. WormPD. BioKnowledge® Library managed by Incyte Genomics contains a *C. elegans* proteome database, WormPD [18,19]. Current version (23/09/2001) includes 19 609 entries; the number of proteins characterized by genomics or biochemistry is 1947 (about 10%) and that of proteins homologous to known proteins is 9313 (47%). Annotation of proteins is progressing day by day, yet proteins with unknown function, 8349 (43%) still remains. This database is searchable by the following parameter alone or combination by Boolean operator; Gene Name/Keyword, Organismal Role, Biochemical Function and Cellular Role, Mutant Phenotype, Subcellular Localization, Post-translational Modifications, Genetic, Calculated Property, and Predicted Motif and by using a partial amino acid sequence. The output contains exhaustive information that starts from the gene name with synonyms and follows the most salient protein properties such as the categories mentioned above, e.g. spatial and temporal patterns of protein expression, partial sequence indices and full sequence of the translated protein, and the calculated values for p*I*, molecular mass, and number of predicted transmembrane segments. Many cross-references to homologues in other organisms, and to other databases and references are also included. From this database, much information could be obtained to suggest a direction on the next information to search for.

2.1.4.2. Stanford Microarray Database (SMD). SMD stores raw and normalized data from microarray experiments [20]. For *C. elegans*, a global profile of germ line gene expression is displayed. The data was obtained using 11 917 cDNA fragments corresponding to 63% of predicted *C. elegans* genes to profile gene expression patterns in the germ line [21]. In the profiles, the result of the experiments, ratio and the probability, is searchable using the gene name. The results can be also illustrated graphically [22].

2.1.4.3. Expression pattern databases

2.1.4.3.1. *The Nematode expression pattern database* (*NEXTDB*). This site, which is managed by the Kohara Laboratory of National Institute of Genetics (Japan), provides the information on the expression pattern of the worm cDNA through EST analysis and systematic whole mount in situ hybridization [23]. The images are accessible by search with YK clone

number, cosmid name (e.g. K04H4), gene name, product name, CELK group, homologous protein, motif or domain, chromosome, and/or clone size. If a clone with interest had been investigated, hybridized images from embryo (egg) to adult could be browsed with the worm gene map. The site attaches the databases of EST (by the same Lab) and of RNAi phenotypes ran by Sugimoto's group [24].

2.1.4.3.2. *The Hope laboratory expression pattern database*. This site, maintained by the Hope laboratory of the University of Leeds (UK) [25], provides the expression profiles of 318 genes with images [26]. However, method of approach is different from NEXTDB, i.e. all expression patterns of this site are generated using reporter genes such as beta-galactosidase (*lacZ*) or green fluorescent protein (GFP). The database is searchable not only by gene name (cosmid name) and protein homology (e.g. kinase), but also by tissue, stage, and pattern description of the expression. The data contain colored or luminescence image(s) with detailed condition and description about the expression.

2.1.4.4. Others. Single nucleotide polymorphism (SNPs) data opens in the database of Genome Sequencing Center of Washington University [27,28]. In *C. elegans* genome sequencing, Bristol N2 strain was used as source. Here, SNPs against the strain CB4856 isolated from a Hawaiian island were predicted from 5.4 Mbp of aligned sequences. The number of polymorphism reached up to 6222. This information would be useful even prior to generating mutant worms.

At present, comprehensive analysis leads to the detection of many hypothetical proteins with unknown function; therefore, to reveal their function, it is quite important to know their partners. Interactome database [29] by the Vidal Laboratory of Harvard University includes information of protein–protein interaction based on yeast two-hybrid system [30]. The information is focused on vulval development-related and proteasome-related proteins [31]. This site attaches a link to ''WORFDB'', which is a database of ORF (open reading frame) of the worm [32]. Nearly all genes of the worm are based on prediction from genome sequence and the results have been confirmed using EST data. The database aims to support the prediction with experimental data, i.e. predicted ORFs are amplified by PCR using ORF-specific primers from cDNA library, cloned and sequenced to generate ORF sequence tags (OSTs) [33]. Each data is displayed with primer sequences used for PCR and the image of resulting electrophoregram, and revealed the existence of at least 17 300 genes in the worm. When the identified protein is a hypothetical protein, it is feasible to know whether the predicted gene is actually expressed at mRNA level and to verify the splicing through the site.

2.1.5. Concluding remarks

The existence of various comprehensive databases represents many and parallel progress in genome-wide researches, particularly it is remarkable on *C. elegans*, maybe due to the cooperative attitude of researchers on the worm. As everyone knows, the first achievement of genome sequencing of multicellular animal was on the small round worm, *C. elegans*. However, proteome database based on actual proteomic experiments is limited for public use. As described in the text, it is essential to construct a systematic protocol to prepare sub-proteomes, which could reconstitute the full-proteome. In addition to the attempts for full description of proteome, the comprehensive analysis of protein–protein interaction, or function proteomics, would be important to complement the results of yeast two-hybrid assay and to provide the basis for elucidating the gene function and the cellular networks of *C. elegans*.

Acknowledgements

This work was supported in part by Grants for Integrated Proteomics System Project, Pioneer Research on Genome the Frontier, from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] V.C. Wasinger, S.J. Cordwell, A. Cerpa-Poljak, J.X. Yan, A.A. Gooley, M.R. Wilkins, M.W. Duncan, R. Harris, K.L. Williams, I. Humphery-Smith, Electrophoresis 16 (1995) 1090.

[2] T.C.e.S. Consortium, Science 282 (1998) 2012–2018.
[3] http://www.sanger.ac.uk/Projects/C_elegans/wormpep/
[4] W.B. Wood (Ed.), The Nematode *Caenorhabditis elegans*, Cold Spring Harbor Laboratory Press, New York, USA, 1988.
[5] B. Skoog, A. Wichman, Trends Anal. Chem. 5 (1986) 82.
[6] R.A. VanBogelen, K.Z. Abshire, B. Moldover, E.R. Olson, F.C. Neidhardt, Electrophoresis 18 (1997) 1243.
[7] A.J. Link, L.G. Hays, E.B. Carmack, J.R. Yates, Electrophoresis 18 (1997) 1314.
[8] http://www.ebi.ac.uk/proteome/
[9] R. Apweiler, M. Biswas, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E.V. Kriventseva, V. Mittard, N. Mulder, I. Phan, E. Zdobnov, Nucleic Acids Res. 29 (2001) 44.
[10] P.H. O'Farrell, J. Biol. Chem. 250 (1975) 4007.
[11] L.A. Meheus, J.J. Van Beeumen, A.V. Coomans, J.R. Vanfleteren, Biochem. J. 245 (1987) 257.
[12] L. Bini, H. Heid, S. Liberatori, G. Geier, V. Pallini, R. Zwilling, Electrophoresis 18 (1997) 557.
[13] H. Kaji, T. Tsuji, K.G. Mawuenyega, A. Wakamiya, M. Taoka, T. Isobe, Electrophoresis 21 (2000) 1755.
[14] S.P. Schrimpf, H. Langen, A.V. Gomes, C. Wahlestedt, Electrophoresis 22 (2001) 1224.
[15] T. Isobe, K. Uchida, M. Taoka, F. Shinkai, T. Manabe, T. Okuyama, J. Chromatogr. 588 (1991) 115.
[16] A.J. Link, J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik, J.R. Yates, Nat. Biotechnol. 17 (1999) 676.
[17] D.K. Han, J. Eng, H. Zhou, R. Aebersold, Nat. Biotechnol. 19 (2001) 946.
[18] http://www.incyte.com/proteome/mainmenu.jsp (registration required; a pay database).
[19] M.C. Costanzo, M.E. Crawford, J.E. Hirschman, J.E. Kranz, P. Olsen, L.S. Robertson, M.S. Skrzypek, B.R. Braun, K.L. Hopkins, P. Kondu, C. Lengieza, J.E. Lew-Smith, M. Tillberg, J.I. Garrels, Nucleic Acids Res. 29 (2001) 75.
[20] http://genome-www4.stanford.edu/MicroArray/SMD/publications.html
[21] V. Reinke, H.E. Smith, J. Nance, J. Wang, C. Van Doren, R. Begley, S.J. Jones, E.B. Davis, S. Scherer, S. Ward, S.K. Kim, Mol. Cell 6 (2000) 605.
[22] http://workhorse.stanford.edu/cgi-bin/genebar/genegraph-.pl
[23] http://nematode.lab.nig.ac.jp/index.html
[24] I. Maeda, Y. Kohara, M. Yamamoto, A. Sugimoto, Curr. Biol. 11 (2001) 171.
[25] http://129.11.204.86:591/default.htm
[26] I.A. Hope, D.G. Albertson, S.D. Martinelli, A.S. Lynch, E. Sonnhammer, R. Durbin, Trends Genet. 12 (1996) 370.
[27] http://genome.wustl.edu/gsc/Projects/C.elegans/SNP/
[28] S.R. Wicks, R.T. Yeh, W.R. Gish, R.H. Waterston, R.H. Plasterk, Nat. Genet. 28 (2001) 160.
[29] http://vidal.dfci.harvard.edu/interactome.htm
[30] A.J. Walhout, S.J. Boulton, M. Vidal, Yeast 17 (2000) 88.
[31] A.J. Walhout, R. Sordella, X. Lu, J.L. Hartley, G.F. Temple, M.A. Brasch, N. Thierry-Mieg, M. Vidal, Science 287 (2000) 116.
[32] http://worfdb.dfci.harvard.edu/
[33] J. Reboul, P. Vaglio, N. Tzellas, N. Thierry-Mieg, T. Moore, C. Jackson, T. Shin-i, Y. Kohara, D. Thierry-Mieg, J. Thierry-Mieg, H. Lee, J. Hitti, L. Doucette-Stamm, J.L. Hartley, G.F. Temple, M.A. Brasch, J. Vandenhaute, P.E. Lamesch, D.E. Hill, M. Vidal, Nat. Genet. 27 (2001) 332.