ELSEVIER

Cell Biology International 28 (2004) 171–178

Cell Biology International

www.elsevier.com/locate/cellbi

Evolution of multicellularity in Metazoa: comparative analysis of the subcellular localization of proteins in *Saccharomyces*, *Drosophila* and *Caenorhabditis*

Einat Hazkani-Covo[a], Erez Y. Levanon[b], Galit Rotman[b], Dan Graur[c], Amit Novik[b]*

[a] *Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel*
[b] *Compugen Ltd, 72 Pinchas Rosen St, Tel Aviv 69512, Israel*
[c] *Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5001, USA*

Received 15 July 2003; revised 21 October 2003; accepted 20 November 2003

Abstract

A comparison of the subcellular assignments of proteins between the unicellular *Saccharomyces cerevisiae* and the multicellular *Drosophila melanogaster* and *Caenorhabditis elegans* was performed using a computational tool for the prediction of subcellular localization. Nine subcellular compartments were studied: (1) extracellular domain, (2) cell membrane, (3) cytoplasm, (4) endoplasmic reticulum, (5) Golgi apparatus, (6) lysosome, (7) peroxisome, (8) mitochondria, and (9) nucleus. The transition to multicellularity was found to be characterized by an increase in the total number of proteins encoded by the genome. Interestingly, this increase is distributed unevenly among the subcellular compartments. That is, a disproportionate increase in the number of proteins in the extracellular domain, the cell membrane, and the cytoplasm is observed in multicellular organisms, while no such increase is seen in other subcellular compartments.

A possible explanation involves signal transduction. In terms of protein numbers, signal transduction pathways may be roughly described as a pyramid with an expansive base in the extracellular domain (the numerous extracellular signal proteins), progressively narrowing at the cell membrane and cytoplasmic levels, and ending in a narrow tip consisting of only a handful of transcription modulators in the nucleus. Our observations suggest that extracellular signaling interactions among metazoan cells account for the uneven increase in the numbers of proteins among subcellular compartments during the transition to multicellularity.

1. Introduction

After their synthesis in the cytosol, the proteins in a eukaryotic cell are sorted to one or more subcellular locations (for reviews, see McNew and Goodman, 1996; Munro, 1998; Omura, 1998; Schlenstedt, 1996). The signals for the subcellular sorting of a protein are mostly encoded within its amino-acid sequence. For example, a protein may contain an amino-terminal signal sequence that directs its entry into a target subcellular compartment. Such signals are found, for instance, in proteins destined to enter the endoplasmic reticulum and mitochondria (Rusch and Kendall, 1995). In many cases, these signals are proteolytically removed during or after entry into the target site.

Other destination-determining sequences are also known. For example, the signal anchor for retention in the membrane is an amino-terminal sequence that is not cleaved (Nilsson et al., 1994), the KDEL motif for retention in the endoplasmic reticulum appears in the carboxy-terminus of the protein (Pelham, 1990), and the nuclear localization signal for entering the nucleus may appear anywhere on the protein (Boulikas, 1993). Given that the subcellular destination of a protein is dictated by “amino-acid-sequence signatures”, or at least characteristic amino-acid compositions, computational tools may be used to predict the subcellular localization of proteins for which relevant empirical data is lacking. In this study, we use ProLoc, a computational tool that predicts the subcellular localization of proteins

* Corresponding author: Tel.: +972-3-7658585; fax: +972-3-7658555
E-mail address: amitn@compugen.co.il (A. Novik).

1065-6995/04/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cellbi.2003.11.016

based solely on their amino-acid sequences. ProLoc has recently been used for large-scale annotation of human proteins through the Gene Ontology Consortium (Xie et al., 2002).

Complete genomic sequences of several eukaryotes are currently available (Adams et al., 2000; C. elegans Sequencing Consortium, 1998; Goffeau et al., 1996; Lander et al., 2001), enabling the comparison of their respective proteomes. Previous studies of complete sets of putative proteins of eukaryotes compared protein-domain numbers (Aravind and Subramanian, 1999; Chervitz et al., 1998; Copley et al., 1999; Lander et al., 2001; Rubin et al., 2000). However, to the best of our knowledge, an evolutionary comparison of the subcellular distribution of whole proteins and whole proteomes has not yet been attempted.

The main purpose of this study is to characterize the evolutionary differences in the subcellular compartmentalization of proteins between unicellular and multicellular eukaryotes (*Saccharomyces cerevisiae* versus *Drosophila melanogaster* and *Caenorhabditis elegans*). Such a characterization is crucial in our attempts to understand the evolutionary transition to multicellularity. Since we focus on the transition to multicellularity, we chose not to deal with the highly derived apomorphic human genome. During evolution, multicellularity arose numerous times in both prokaryotes and eukaryotes (Kaiser, 2001). In the lineage leading to Metazoa, the transition was accompanied by changes in composition and complexity that varied widely among the different subcellular compartments (Gerhart and Kirschner, 1997). For example, a disproportionate increase is observed in the number of cell-adhesion proteins in the extracellular matrix (Hynes, 1999). Multicellularity also requires an increase in the relative number of proteins affecting the transfer of information among cells, e.g. hormones, neurotransmitters, and signal transduction proteins (Downward, 2001). Since the subcellular distribution of these proteins is not uniform, disproportionate changes may occur in the protein constitution of some compartments.

2. Materials and methods

2.1. Methodology

ProLoc is a computational tool for predicting the subcellular localization of eukaryotic proteins. It uses a series of properties derived from the primary amino acid sequence of a protein to assign it to one of nine subcellular compartments: (1) extracellular domain, (2) cell membrane, (3) cytoplasm, (4) endoplasmic reticulum, (5) Golgi apparatus, (6) lysosome, (7) peroxisome, (8) mitochondria, and (9) nucleus. The program also assigns proteins to subdivisions within each of the above compartment; a total of 24 sublocations. As shown below, the amino-acid properties that are used for the subcellular assignments are (1) N-terminal sequences, (2) protein motifs, (3) amino-acid composition, (4) isoelectric point (pI), and (5) protein length.

1. *N-terminal protein sequences.* Signal sequences are commonly used to assign proteins to subcellular compartments (e.g. Emanuelsson et al., 2000). ProLoc uses signal peptides, signal anchors, and mitochondrion targeting signals (Neupert, 1997; Nilsson et al., 1994; von Heijne, 1985).
2. *Protein motifs.* ProLoc uses motifs that have been previously shown to characterize the sorting process of proteins to subcellular compartments. For example, the KDEL, SKL and SV40-like motifs characterize ER, peroxisome and nuclear proteins, respectively (Dingwall and Laskey, 1991; McNew and Goodman, 1996; Pelham, 1990). Another motif used by the program is the transmembrane segment. ProLoc also searches for compartment-specific domains and signatures from the Pfam and PROSITE databases (Bateman et al., 2000; Hofmann et al., 1999). The 181 matrices that are unique to a particular subcellular compartment were collected from Pfam-A version 5.3 (Bateman et al., 2000), and were used by ProLoc.
3. *Amino acid composition.* The distribution of the amino-acid residues of proteins from different subcellular compartments may differ considerably. For example, integral membrane proteins are rich in hydrophobic amino acids, while nuclear proteins are poor in hydrophobic amino acid residues and rich in charged residues.
4. *pI.* Different subcellular compartments have different pH values, so the proteins that function in them often have different isoelectric points (pI). For example, lysosomal proteins have an acidic pI, while nuclear proteins have a more basic pI.
5. *Protein length.* Protein lengths may vary among subcellular locations. For example, mitochondrial proteins tend to be smaller than proteins of the cell membrane.

ProLoc starts with several groups of proteins whose subcellular localization has been unambiguously determined empirically. These groups are then used to “train” the program in the proper assignment of proteins to the different compartments. Simply put, ProLoc builds an experimental profile of properties for proteins in each of the subcellular compartments, and then compares each protein with each of the subcellular profiles. The end product is a list detailing how well a protein fits into each of the subcellular profiles. The best fit is used as a compartmental assignment.

A detailed description of ProLoc (including technical minutiae) can be found at http://www.labonweb.com/cgi-bin/proloc/search.cgi.

Table 1
Predictive accuracy of ProLoc. The numbers show percentage of accuracy in predicting subcellular location using the training–testing approach and the jackknifing approach, with and without Pfam matrices. Results of training-testing approach are using proteins in the testing set and (in parentheses) in the training set

Predictive accuracy for subcellular compartments	Training–testing		Jackknifing		Total number of proteins
	No-Pfam	Pfam	No-Pfam	Pfam	
Extracellular domain	70 (69)	78 (76)	65	74	884
Cell membrane	79 (80)	83 (84)	80	84	1850
Cytoplasm	60 (61)	64 (63)	56	59	462
ER	78 (86)	80 (86)	74	79	217
Golgi	50 (61)	67 (73)	37	56	59
Lysosome	74 (88)	80 (88)	64	74	105
Peroxisome	56 (76)	56 (76)	39	44	41
Mitochondria	67 (77)	74 (80)	69	79	288
Nucleus	76 (76)	82 (80)	75	79	1092
Weighted mean	74 (75)	79 (80)	72	78	4998

2.2. Protein datasets for quality control of ProLoc

In order to train ProLoc and to estimate the accuracy of its predictions, we extracted a subset of vertebrate proteins from the SwissProt 39 database (Bairoch and Apweiler, 2000) for which the subcellular localization is clearly annotated. Proteins with ambiguous localizations or whose descriptions included words such as "probable," "potential," and "by similarity" were omitted. We also omitted proteins that did not begin with the amino acid methionine. To minimize bias favoring more well-studied protein families over others, we used Holm and Sander's (1998) near-neighbor redundancy algorithm with a threshold of 90% similarity on the relevant fraction of SwissProt data.

Testing ProLoc performance was accomplished by using two approaches: the training-testing test and the Jackknife test. With the first approach, the 4998 proteins were divided into two parts: 4/5 of the proteins were used as a training set, and 1/5 as a test set. The Jackknife test (Mardia et al., 1979), also known as the "leave-one-out" test, was used for cross-validation of the prediction. During the jackknifing process each protein in the dataset is in turn selected as a tested protein while all the parameters are calculated on the remaining proteins. The localizations predicted by ProLoc were compared to empirically determined localizations. Prediction accuracy was defined as the percentage of correct assignments from among all assignments. The overall prediction accuracy was somewhat lower where Pfam domains were not used as shown in Table 1.

The weighted mean prediction accuracy of ProLoc was 78% (jackknifing) to 79% (training-testing) while using Pfam matrices. If we take into account the two or three best predictions for probable subcellular locations, the jackknifing percentages become 88% and 92%, respectively. Single-compartment prediction accuracy ranged from 44% (jackknifing) or 56% (training-testing) for the peroxisome proteins to 84% (jackknifing) or 83% (training-testing) for the cell membrane proteins. As described by Chou and Zhang (1995), when the number of proteins in a given set is not large enough (e.g. peroxisome and Golgi apparatus), the leave-one-out test may result in a severe loss of information.

2.3. Protein datasets of yeast, nematode and fruitfly

The initial dataset included all known and predicted proteins from *D. melanogaster* (14,080 proteins), *C. elegans* (19,704 proteins), and *S. cerevisiae* (6310 proteins). The yeast proteins were from SGD (*Saccharomyces* Genome Database (Cherry et al., 1997), http://genome-www.stanford.edu/*Saccharomyces*/), the nematode proteins were from Wormpep 32 (Sonnhammer and Durbin, 1997; http://www.sanger.ac.uk/Projects/C_elegans/wormpep/), and the fruitfly proteins were from BDGP (Berkeley *Drosophila* Genome Project (Rubin et al., 2000), http://www.fruitfly.org/sequence/). We omitted proteins that did not begin with methionine as well as proteins shorter than 13 amino acids. The final query datasets included 6308 yeast proteins, 19,677 nematode proteins, and 14,019 fruitfly proteins.

For analysis of the *D. melanogaster* and *C. elegans* proteins, ProLoc was trained on all proteins from Metazoa (5954) for which the subcellular localization is unambiguously annotated. For analysis of the *S. cerevisiae* proteins, ProLoc was trained on all annotated proteins from Ascomycota (1025).

3. Results

3.1. Predicted subcellular distribution of proteins in yeast, nematode and fruitfly

The numbers of yeast, nematode and fruitfly proteins assigned to each subcellular compartment are summarized in Table 2. Lysosomal proteins could not be

Table 2
Assignment of known and hypothetical proteins encoded by the genomes of yeast, nematode and fruitfly to each of 9 subcellular compartments

	Yeast		Nematode		Fruitfly	
	Number of proteins	Percentage	Number of proteins	Percentage	Number of proteins	Percentage
Extracellular domain	173	2.74	1876	9.53	1556	11.10
Cell Membrane	331	5.25	3829	19.46	1681	11.99
Cytoplasm	831	13.17	6049	30.74	3589	25.60
ER	402	6.37	1465	7.45	587	4.19
Golgi apparatus	64	1.01	66	0.34	93	0.66
Lysosome	[a]	[a]	338	1.72	388	2.77
Peroxisome	368	5.83	862	4.38	544	3.88
Mitochondria	1396	22.13	1809	9.19	1382	9.86
Nucleus	2743	43.48	3383	17.19	4199	29.95
Total	6308	100	19,677	100	14,019	100

[a] Lysosomal proteins could not be predicted in yeast (see text).

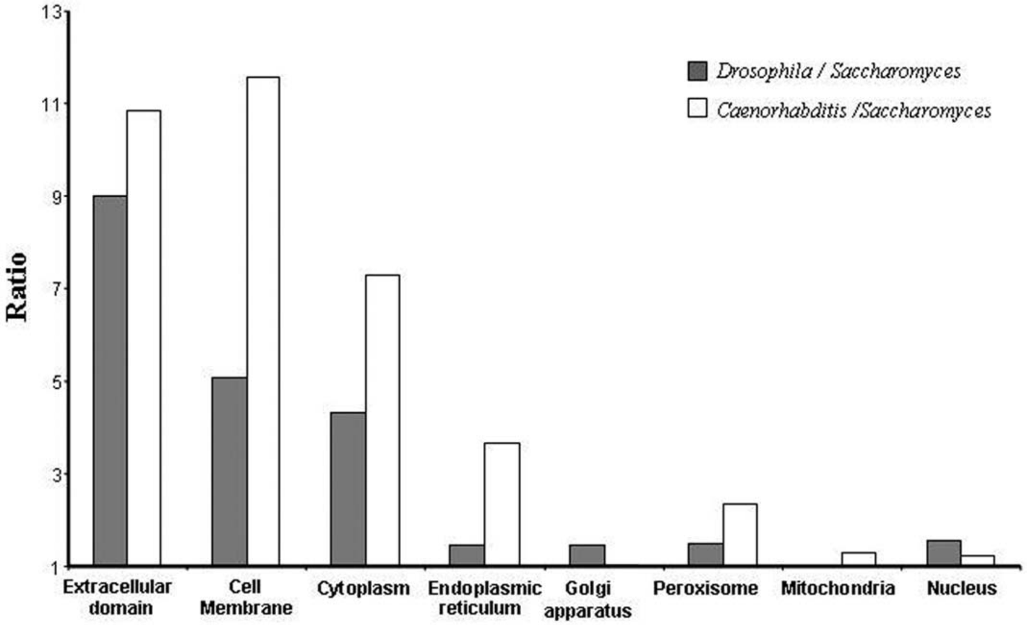

Fig. 1. Ratios of protein numbers between nematode and yeast (white columns), and between fruitfly and yeast (gray columns) in the various subcellular compartments.

predicted in yeast since the number of known lysosomal proteins in this organism is too small for a meaningful statistical estimate.

3.2. Comparison between unicellular and multicellular organisms

Multicellular organisms exhibit a dramatic increase in the number of proteins in the extracellular domain (173 in yeast vs. 1876 in nematode and 1556 in fruitfly, an 11- and 9-fold increase, respectively), the cell membrane (331 vs. 3829 and 1681, a 12- and 5-fold increase, respectively), and the cytoplasm (831 vs. 6049 and 3589, a 7- and 4-fold increase, respectively) (Fig. 1). In contrast, the endoplasmic reticulum exhibits a more modest increase in *C. elegans*, and almost no increase in *Drosophila* (402 vs. 1465 and 587, a 3.5- and 1.5-fold increase, respectively). We see almost no increase in protein numbers in the other compartments.

3.3. Comparison between nematode and fruitfly

The difference in the total number of predicted proteins between nematode (19,677) and fruitfly (14,019) is confined primarily to the cytoplasm (2460 proteins, 43% of the difference), the cell membrane (2148 proteins, 38%), and the endoplasmic reticulum (878 proteins, 16%). A smaller difference is seen in the mitochondria (427 proteins, 8%), the extracellular domain (320 proteins, 6%), and the peroxisome (318 proteins, 6%). A similar number of lysosomal proteins is observed in the two metazoans. The nucleus is the only compartment that shows a larger number of proteins in the fruitfly than in the nematode. A 9 by 2 contingency test of the number of proteins in the different subcellular compartments indicates that the internal pattern of subcellular assignment is significantly different between *Drosophila* and *Caenorhabditis* (χ^2=1,178; *df*=8; $P \ll 0.001$).

4. Discussion

Several methods have been used to predict the subcellular localization of proteins. Three such examples are SignalP (Nielsen and Krogh, 1998), PSORT (Nakai and Horton, 1999), and NNPSL (Reinhardt and Hubbard, 1998). As described by Nakai (2000), the prediction systems for the localization of protein are roughly divided into methods based on amino acid composition alone (e.g. Cedano et al., 1997; Chou and Elrod, 1999; Reinhardt and Hubbard, 1998; Zhou and Doctor, 2003), methods based on sorting signals alone (Emanuelsson et al., 2000; Nielsen and Krogh, 1998), and methods that combine both amino acid composition and sorting signals (e.g. Nakai and Horton, 1999). Meanwhile, methods that incorporate the quasi-sequence-order effects of an entire protein chain (e.g. Chou, 2001) and methods that take functional domain information into account (e.g. Chou and Cai, 2002) have been developed. In the current study we used ProLoc, which can be classified as a method combining both amino-acid composition and sorting signals.

Let us now look at the wrong predictions of ProLoc. In each organelle, we determined the most "popular" false subcellular assignments. We found that ProLoc inaccuracies are somehow related to the cellular process of protein localization. All nuclear-encoded mRNAs are translated on cytosolic ribosomes. Ribosomes synthesizing nascent proteins in the secretory pathway are directed to the rough ER by a signal sequence. After translation is completed in the ER, these proteins move via transport vesicles to the Golgi apparatus, from whence they are further sorted to several destinations. Hence, the organelles involved in this pathway are the ER, Golgi apparatus, lysosome, cell membrane and extracellular domain.

Synthesis of all other nuclear-encoded proteins is completed on "free" cytosolic ribosomes, and the completed proteins are released into the cytosol. These proteins remain in the cytosol unless they contain a specific signal sequence that directs them into the mitochondrion, the peroxisome or the nucleus. We found that in those cases where ProLoc predicts the wrong subcellular location, chances are that it will predict a location within the correct sorting pathway. This phenomenon is unsurprising for proteins sorted via the ER pathway, since these proteins possess particular signal peptides (or variant signal anchors). As far as ER proteins are concerned, 7% of them are wrongly assigned to the cell membrane and 5% to the lysosome. As far as Golgi proteins are concerned, 15% of them are wrongly assigned to the ER. As far as lysosome proteins are concerned, 11% of them are wrongly assigned to the cell membrane. Similarly, from among the cell-membrane proteins, 5% are wrongly assigned to the ER, and from among the extracellular-domain proteins, 6% are wrongly assigned to the lysosome and 5% to the cell membrane.

The second group of proteins, i.e. those located in the mitochondria, the nucleus, the cytoplasm and the peroxisome, do not share a common transport pathway. However, a similar phenomenon, albeit of lesser magnitude, is observed. We note, however, that some of the false subcellular assignments in this case are assignments to locations belonging to the ER pathway. For example, 8% of all mitochondrial proteins are assigned to the extracellular domain and 4% to the cytoplasm. Fifteen percent of the nuclear proteins are assigned to the cytoplasm; 13% and 12% of the cytoplasm proteins are assigned to the mitochondria and the nucleus, respectively, and 17% and 14% of the peroxisome proteins are assigned to the ER and cytoplasm, respectively.

Given the erroneous assignments above, it is important to assess whether they affect the conclusions in a significant manner. The answer is most probably negative because of three reasons. First, the false assignments constitute a small percentage of all cases. Second, our estimates of wrong assignments are most certainly inflated because proteins may appear in more than one subcellular compartment. Proteins that shuttle between the cytoplasm and the nucleus constitute one such case in point (Guiochon-Mantel et al., 1994). And finally, because assignment errors in subcellular localization is rather slight, the differences in localization between yeast, nematode and fruitfly are most certainly real rather than artifactual.

Comparing the methods for assigning protein to various subcellular localizations would certainly be an important task, but is unfortunately one that is beyond the aims of this evolutionary study. That said, we note that comparing results pertaining to sets of proteins predicted from complete genomes is far from being a trivial task. There are two main reasons for the difficulties: (1) the numbers and types of predicted subcellular locations by each method are incomparable, and (2) the training sets are different. All predictions concerning subcellular localization should be treated with caution. First, we note that all studies to date have been based on genomic rather than proteomic data. One source of bias may be that currently most genes are assigned a single transcript, which may be a gross underestimation of the number of proteins encoded by a gene. Moreover, ignoring alternative splicing may not only reduce the number of proteins, but may also deprive us of pertinent information, since the splicing pattern of an mRNA transcript may determine the subcellular location of the encoded protein (Black, 2000). The second problem is the existence of proteins that appear in more than one compartment, for example, proteins that shuttle between cytoplasm and nucleus.

Unlike protozoans, metazoans have a number of specialized cell types. In order to coordinate their functions, the cells of a multicellular organism must maintain a constant flow of communication. The developmental complexity of multicellular eukaryotes is based on a system of proteins engaged in extracellular, intercellular, and intracellular signaling (Gerhart and Kirschner, 1997). Signaling requires the transfer of signals, such as hormones and neurotransmitters, from the extracellular domain through the cell membrane to the cytoplasm, and from the cytoplasm to the nucleus (Downward, 2001). Thus, the evolution of multicellularity should be accompanied by an increase in the numbers of proteins in the extracellular domain, the cell membrane, the cytoplasm and the nucleus. Our observations are in full agreement with this line of reasoning. Interestingly, a disproportionate increase in proteins involved in signaling also seems to have occurred in two independently evolved prokaryotic multicellular groups, i.e. Nostocales and Myxobacteria, as well as in a eukaryotic lineage (Volvocales), in which multicellularity evolved independently of that in Metazoa (see Kaiser, 2001).

Multicellular organisms also require cellular adhesion mechanisms, i.e. proteins involved in cell–cell adhesion and cell–matrix interactions, especially for the creation of complex structures during embryogenesis (Hynes, 1994, 1999). Consequently, the transition to multicellularity should be accompanied by an additional increase in the number of cell membrane and extracellular proteins. Since neither signaling nor cellular adhesion are associated with proteins located in the lysosome, peroxisome, mitochondrion, endoplasmic reticulum or Golgi apparatus, these subcellular compartments should not have changed much during the transition to multicellularity.

Indeed, the predicted protein distribution in the three organisms in our study agrees reasonably well with the above expectation. The number of proteins in the extracellular domain, the cell membrane and the cytoplasm increased dramatically in fruitfly and nematode compared to yeast (9–11-fold in the extracellular domain, 5–12-fold in the cell membrane, and 4–7-fold in the cytoplasm), while only minor changes were observed in protein numbers of mitochondria, peroxisome and Golgi apparatus.

Several core biological processes that involve maintenance and expression of genetic material, as well as its duplication and division during the cell cycle, take place in the nucleus. The signaling pathways in multicellular organisms are known to lead to changes in gene expression and to be carried out by transcriptional regulators (Downward, 2001). Therefore, one may expect the evolution of multicellular organisms to be accompanied by changes in the number of such regulators. However, a comparative analysis of zinc-binding transcription factors shows that the total number of such proteins does not vary much between yeast and nematode (Clarke and Berg, 1998), and core biological functions are carried out by a similar number of proteins in both organisms (Chervitz et al., 1998). Taken together, these observations suggest that the number of nuclear proteins has not changed dramatically during the evolution from unicellular to multicellular organisms. Our findings of only a small increase in the number of nuclear proteins in nematode and fruitfly strengthen the conclusions of Clarke and Berg (1998) and Chervitz et al. (1998).

Our data indicate that the highest increase in protein numbers during the evolutionary transition to multicellularity was confined to the extracellular domain and the cell membrane; a considerable, albeit smaller, increase occurred in the cytoplasm, and only a small increase is noted in the nucleus. Signal transduction pathways may be numerically described as an approximate inverted pyramid with an expansive base in the extracellular domain (the numerous extracellular signal proteins), a somewhat smaller number of membrane receptors, still smaller numbers of cytoplasmic proteins, and ending at the narrow tip of the pyramid in a handful of transcription modulators. Our observations indicate that, to a great extent, signal transduction mechanisms may account for the uneven increase in numbers of proteins among subcellular compartments during the transition to multicellularity.

In a study by the International Human Genome Sequencing Consortium (Lander et al., 2001), humans appear to have more proteins involved in cytoskeleton construction, transcription and translation, as well as defense and immunity, than the nematode and fruitfly. There appear to be only modest differences in the number of protein domains between vertebrates and invertebrates, but the former have more distinct protein-domain architectures, defined as the linear arrangement of domains within a polypeptide. This difference is most prominent in the recent evolution of novel extracellular and transmembrane architectures in the human lineage, suggesting that extracellular and transmembrane proteins were important during vertebrate evolution. A similar conclusion regarding the evolution from unicellular to multicellular metazoa can be reached from our findings of a dramatic increase in the number of proteins in these cellular compartments.

The nematode has 5658 more predicted genes than the fruitfly. Our results suggest that this difference is confined mainly to the cytoplasm (43%), the cell membrane (38%) and the endoplasmic reticulum (16%).

Cytoskeletal proteins are believed to be encoded by 5% of the nematode genes, but by only 2% of the fruitfly genes (Rubin et al., 2000). This difference alone may account for about 28% of the difference in the number of cytoplasmic proteins between these organisms. In addition, 5% of all nematode genes encode G-protein-coupled receptors, which are membrane proteins. About 100 of these have clear similarity to receptors identified in other animals, while around 1000 are nematode-specific and are thought to encode chemoreceptors (Bargmann, 1998). These species-specific proteins may provide an explanation for the difference in membrane proteins between nematode and fruitfly.

Acknowledgements

We wish to thank Sergey Gribov, Nili Guttmann-Beck, Sergey Nemzer, Michal Preminger, Hershel Safer and Artur Shnayder for their help. We gratefully acknowledge support from the Magnet Da'at Consortium of the Israel Ministry of Industry and Trade. DG was supported by the Research Foundation of Tel Aviv University.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG et al. The genome sequence of *Drosophila melanogaster*. Science 2000;287:2185–95.

Aravind L, Subramanian G. Origin of multicellular eukaryotes-insights from proteome comparisons. Curr Opin Genet Dev 1999; 9:688–94.

Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000; 28:45–8.

Bargmann CI. Neurobiology of the *Caenorhabditis elegans* genome. Science 1998;282:2028–33.

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2000; 28:263–6.

Black DL. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell 2000;103:367–70.

Boulikas T. Nuclear localization signals (NLS). Crit Rev Eukaryot Gene Expr 1993;3:193–227.

Cedano J, Aloy P, Perez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997; 266:594–600.

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C et al. Genetic and physical maps of *Saccharomyces cerevisiae*. Nature 1997;387:67–73.

Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS et al. Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science 1998;282:2022–8.

Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001;43:246–55.

Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 2002;277:45765–9.

Chou KC, Elrod DW. Protein subcellular location prediction. Protein Eng 1999;12:107–18.

Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.

Clarke ND, Berg JM. Zinc fingers in *Caenorhabditis elegans*: finding families and probing pathways. Science 1998;282:2018–22.

C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 1998; 282:2012–8.

Copley RR, Schultz J, Ponting CP, Bork P. Protein families in multicellular organisms. Curr Opin Struct Biol 1999;9:408–15.

Dingwall C, Laskey RA. Nuclear targeting sequences—a consensus? Trends Biochem Sci 1991;16:478–81.

Downward J. The ins and outs of signalling. Nature 2001;411:759–62.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000;300:1005–16.

Gerhart J, Kirschner M.. Cells, embryos and evolution: toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability. Malden (MA): Blackwell Science Inc, 1997;243–7.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science 1996;274:546, 563–567.

Guiochon-Mantel A, Delabre K, Lescop P, Milgrom E. Nuclear localization signals also mediate the outward movement of proteins from the nucleus. Proc Natl Acad Sci U S A 1994;91:7179–83.

Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. Nucleic Acids Res 1999;27:215–9.

Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 1998;14:423–9.

Hynes RO. Genetic analyses of cell-matrix interactions in development. Curr Opin Genet Dev 1994;4:569–74.

Hynes RO. Cell adhesion: old and new questions. Trends Cell Biol 1999;9:M33–7.

Kaiser D. Building a multicellular organism. Annu Rev Genet 2001; 35:103–23.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.

Mardia KV, Kent JT, Bibby JM. Multivariate analysis. Academic Press, 1979, p. 322, 381.

McNew JA, Goodman JM. Targeting and assembly of peroxisomal proteins: some old rules do not apply. Trends Biochem Sci 1996; 21:54–8.

Munro S. Localization of proteins to the Golgi apparatus. Trends Cell Biol 1998;8:11–5.

Nakai K. Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 2000;54:277–344.

Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 1999;24:34–6.

Neupert W. Protein import into mitochondria. Annu Rev Biochem 1997;66:863–917.

Nielsen H, Krogh A. Prediction of signal peptides and signal anchors by a hidden Markov model. Proc Int Conf Intell Syst Mol Biol 1998;6:122–30.

Nilsson I, Whitley P, von Heijne G. The COOH-terminal ends of internal signal and signal-anchor sequences are positioned differently in the ER translocase. J Cell Biol 1994;126:1127–32.

Omura T. Mitochondria-targeting sequence, a multi-role sorting sequence recognized at all steps of protein import into mitochondria. J Biochem (Tokyo) 1998;123:1010–6.

Pelham HR. The retention signal for soluble proteins of the endoplasmic reticulum. Trends Biochem Sci 1990;15:483–6.

Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26:2230–6.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK et al. Comparative genomics of the eukaryotes. Science 2000;287:2204–15.

Rusch SL, Kendall DA. Protein transport via amino-terminal targeting sequences: common themes in diverse systems. Mol Membr Biol 1995;12:295–307.

Schlenstedt G. Protein import into the nucleus. FEBS Lett 1996; 389:75–9.

Sonnhammer EL, Durbin R. Analysis of protein domain families in *Caenorhabditis elegans*. Genomics 1997;46:200–16.

von Heijne G. Signal sequences. The limits of variation. J Mol Biol 1985;184:99–105.

Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A et al. Large-scale protein annotation through gene ontology. Genome Res 2002;12:785–94.

Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins 2003;50:44–8.