

Genome-wide Analyses of Carboxyl-terminal Sequences*

Jean-Ju Chung‡§, Hongmei Yang‡, and Min Li¶

Sequence motifs at the protein carboxyl termini in linear polypeptides are uniquely positioned and functionally capable of serving as recognition signatures for a variety of cellular and biochemical processes. At the proteome level, it is unknown whether and what carboxyl-terminal sequences might be particularly conserved, which may be directly related to specific biological functions shared among certain groups of proteins. To investigate this question, we analyzed the terminal sequences of reported yeast open reading frames, which presumably constitute the predicted, entire proteome of *Saccharomyces cerevisiae*. The results show that there are both known and novel terminal sequences. They are conserved at a frequency similar to that of functionally important, experimentally confirmed signals such as the HDEL sequence that mediates the endoplasmic reticulum retention and/or retrieval. The findings support the notion that there may be additional carboxyl-terminal signals, and the conserved motifs could be experimentally tested for currently unknown biological functions. Similar analyses were also applied to the limited proteome databases of other organisms with overall consistent findings. Therefore, indexing a proteome according to its carboxyl-terminal sequences may provide a means for functional classification and determination of proteins. *Molecular & Cellular Proteomics* 2:173–181, 2003.

Each protein has one single terminal α -carboxyl group that links directly to the adjacent, last peptide bond. This position, referred to here as a carboxyl terminus, combined with preceding residues often serves as a signature recognition motif capable of conferring a variety of biochemical reactions that are restricted to this position of protein and of essential physiological functions. Some of the known functions include protein trafficking, subcellular anchoring of proteins, targeted protein degradation, and the static and dynamic formation of macromolecular complexes (for a review, see Ref. 1).

In an increasing number of systems, protein carboxyl-terminal sequences are found to be highly conserved among homologues of different species. Structural studies have identified a large number of protein domains that specifically

recognize protein carboxyl termini (2, 3). Using random peptide selection, these protein domains were shown to display high specificity for binding to carboxyl-terminal sequences (4, 5). For example, cystic fibrosis transmembrane conductance regulator (CFTR)¹ proteins, whose malfunction causes cystic fibrosis, have an identical carboxyl-terminal sequence (TRL) in species from *Xenopus* to human (5). This sequence is necessary for appropriate subcellular expression of the CFTR channel in heterologous systems such as polarized Madin-Darby canine kidney cells (6). In addition, the motif is responsible for binding to CAP70 (CFTR-associated protein, 70 kDa), a protein that contains four protein interaction PDZ domains. The multivalent binding of CAP70 to two or more CFTR molecules potentiates the chloride channel activity (7). Similarly, within a given species, conserved carboxyl-terminal motifs have also been found among structurally and/or functionally distinct proteins. These conserved motifs often code specific biological activities, such as HDEL, which is a recognition signal for ER retention and/or retrieval (8), and CAAX, which serves as a substrate site for lipidation (9). Due to an increasing number of genomes that have been sequenced and their corresponding proteome information becoming available, it is now becoming feasible to investigate questions such as whether and to what extent the protein carboxyl-terminal motifs are conserved within a given proteome. If so, do those motifs indeed confer certain conserved functionality that has been determined experimentally? In addition, the conserved sequences with unknown function may be topics of further experimental studies.

To determine whether and what carboxyl sequence motifs are conserved in yeast, we compiled terminal sequences of 6,213 predicted yeast proteins that are longer than 50 amino acid residues. The analyses of frequency of contiguous carboxyl-terminal sequence suggest that conserved motifs are directly correlated to certain shared functions including but not limited to protein targeting. The function of the identified motifs may be investigated experimentally.

MATERIALS AND METHODS

Database for Protein Sequences—The protein sequences of 6,213 proteins from *Saccharomyces cerevisiae* longer than 50 residues were extracted from the National Center for Biotechnology Information (NCBI) database (<ftp.ncbi.nih.gov>). Other genome sequences to run the program were downloaded from NCBI at <ftp.ncbi.nih.gov>, and the reference history can be tracked from www.ncbi.nlm.nih.gov/

From the Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205

Received, January 22, 2003, and in revised form, April 5, 2003

Published, MCP Papers in Press, April 7, 2003, DOI 10.1074/mcp.M300008-MCP200

¹ The abbreviations used are: CFTR, cystic fibrosis transmembrane conductance regulator; ER, endoplasmic reticulum.

TABLE I
Carboxyl-terminal sequence distribution frequency of yeast proteins

X ₂ -COOH		X ₃ -COOH		X ₄ -COOH		X ₅ -COOH		X ₆ -COOH		X ₇ -COOH		X ₈ -COOH	
KK	152	SKK	21	EVGE	16	REVGE	16	SREVGE	16	SSREVGE	16	DSSREVGE	16
SK	83	KKK	20	KWIH	16	NKWIH	16	TNKWIH	16	LTNKWIH	16	LLTNKWIH	16
KL	70	AKK	17	TIAN	14	YTIAN	14	IYTIAN	14	GIYTIAN	14	DGIYTIAN	13
LL	68	VGE	17	HDEL	11	RPETY	5	GFGLFD	5	IYTAIPK	5	LHLRPETY	5
LK	59	WIH	16	KSKK	7	GVYGG	5	LRPETY	5	MGFGLFD	5	DMGFGLFD	5
RK	54	IAN	15	ALLL	6	FGLFD	5	YTAIPK	5	HLRPETY	5	LHLRPETY	5
SS	52	DEL	15	EEVD	6	TAIPK	5	LRPGTY	5	HLRPETY	5	GIYTAIPK	5
SL	50	SKL	13	PGTY	5	RPETY	5	AAAMLL	4	KRLALPA	4	RVLGVVYC	4
KI	50	EKK	12	VYGG	5	AAMLL	5	LGVVYC	4	PTVEEVD	4	FKRLALPA	4
EL	47	LLL	12	KKNN	5	GVVYC	4	TVEEVD	4	GRIYISE	4	GPTVEEVD	4
KN	46	LSK	12	GLFD	5	LALPA	4	SGVYGG	4	VLGVVYC	4	VFSGVYGG	4
KE	45	LKK	12	AMLL	5	VEEVD	4	RIYISE	4	FSGVYGG	4	EGRIYISE	4
EK	44	LLK	11	AIPK	5	IYISE	4	RLALPA	4	PLAKKKE	3	SCSEESLA	3
AN	44	KKE	11	PETY	5	NALYS	3	VVDTSK	3	AAAAMLL	3	AGAAALLL	3
RR	43	GKK	10	LLKQ	4	FNNNT	3	MYGCHT	3	CSEESLA	3	RYVVDTSK	3
LS	42	DSK	9	ALPA	4	KRLHN	3	EKRLHN	3	HMYGCHT	3	FHMYGCHT	3
EE	41	FWC	9	ALVA	4	GKFK	3	LLLAI	3	GAAALLL	3	ASLLLLAI	3
RL	41	QKI	9	KKEK	4	SEIGW	3	FGKFK	3	AATNAKQ	3	RFHGDGNKL	3
KQ	41	LSI	9	KKFK	4	VCCPS	3	SEESLA	3	RLSEIGW	3	EPLAKKKE	3
IL	40	MLL	9	DYFL	4	AALLL	3	LKYFGR	3	SLLLLAI	3	TAATNAKQ	3
NK	39	DEE	9	VVYC	4	SLENL	3	ATNAKQ	3	NDTLKQ	3	SYEKRLHN	3
LI	39	RRK	9	STFY	4	KSKK	3	LAKKKE	3	YVVDTSK	3	RAINALYS	3
TL	38	SSS	9	SSKL	4	AKKKE	3	HDGNKL	3	YEKRLHN	3	YNDTLKQ	3
TK	38	KKN	8	YISE	4	LLAI	3	INALYS	3	FHDGNKL	3	SDLFSEVE	2
SI	37	EVD	8	KLLK	4	YGCHT	3	DTLLKQ	3	AINALYS	3	HMYSSSL	2
GL	37	LDL	8	KVSK	3	LHDEL	3	AAALLL	3	ELKYFGR	3	EVNIGIKQ	2
ST	36	TKK	8	LSKK	3	VKKEK	3	LSEIGW	3	TKAMSSR	2	IVDGKVLK	2
SE	36	LLV	8	YFGR	3	FNFTK	3	PFSTFY	2	VEHVAKA	2	QITSSITS	2
FL	36	KEK	8	EIGW	3	TNAKQ	3	ARRNAD	2	LFTHAPV	2	DFYDAFYN	2
DL	35	KKD	8	GIKQ	3	EESLA	3	ARSEDK	2	VLKNYSK	2	RTVHRSLD	2

locuslink/refseq.html. Information regarding yeast protein expression, localization, and functional properties were taken from the following websites unless specifically cited: genome-www.stanford.edu/Saccharomyces/, genome-www4.stanford.edu/cgi-bin/SGD, www.proteome.com/, and www.ncbi.nlm.nih.gov/locuslink/refseq.html.

Statistical Analysis—The programs used in database downloading, parsing, and subsequent statistical analysis were written in Perl5.6 and run on a PC Pentium 700 computer. The data were output in Microsoft Excel format. Protein sequence alignment was performed using McAlign of DNASTar™.

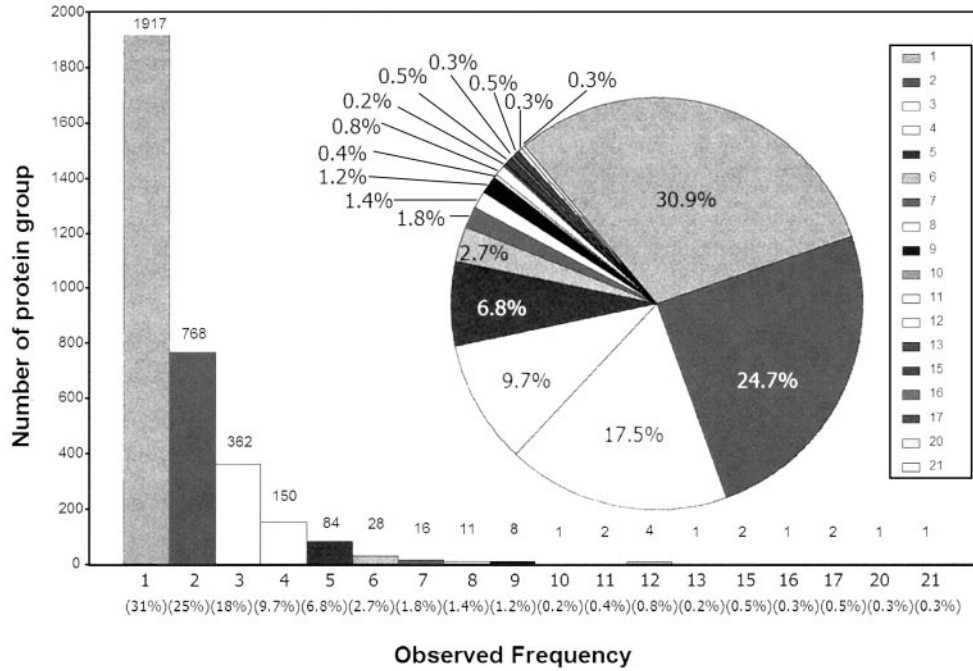
RESULTS

Certain carboxyl tri- or tetra-amino acid sequences, such as SKL for peroxisome targeting (10) and HDEL (or KDEL, depending on the species) for ER retention and/or retrieval (8), are recognition sequences for trafficking proteins to appropriate subcellular compartments or microdomains. To investigate specific common sequence motifs that are shared by a group of otherwise different proteins, we have compiled an abundance of carboxyl peptide sequences from several species including *S. cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. Because the amino acid usage bias is found primarily at the last 8 residues (11, 12), our calculation was performed for dimers and trimers through octamers. The yeast genome has been sequenced

with very limited numbers of gene duplication, which are more suitable for this analysis. The top 30 hits of each are shown in Table I. The complete lists of hits are available from our website (www.molecularinteraction.org/listofpublication.htm).

Of 6,213 yeast proteins that are longer than 50 amino acids, we analyzed the frequency of tripeptide and tetrapeptide sequences (Fig. 1). Fig. 1A shows that 31% of (or 1,917) proteins have unique tripeptide terminal sequences; 25% of (or 1,536 (768 × 2)) proteins have terminal sequences that are found only twice (Fig. 1A). When the same analysis was applied to the tetrapeptidic sequences, about 83% of (or 5,151) proteins have distinct carboxyl tetrapeptide terminal sequences, and ~12% of (or 766 (383 × 2)) proteins have their terminal sequences found only in one other protein (Fig. 1B). Of the remaining 297 proteins (~5%) their terminal sequences are identical to three or more proteins. Analyses of the highest 30 hits identify several conserved motifs. The most abundant three hits of tetrapeptides are EVGE (16 hits), KWIH (16 hits), and TIAN (14 hits). The EVGE is found in YRF1-like helicases in multiple alleles. KWIH is a motif found in an open reading frame of Ty transposons, which are presumably present in multiple locations of yeast genome. TIAN is the terminus shared by multiple, nearly identical seripauperin proteins.

A



B

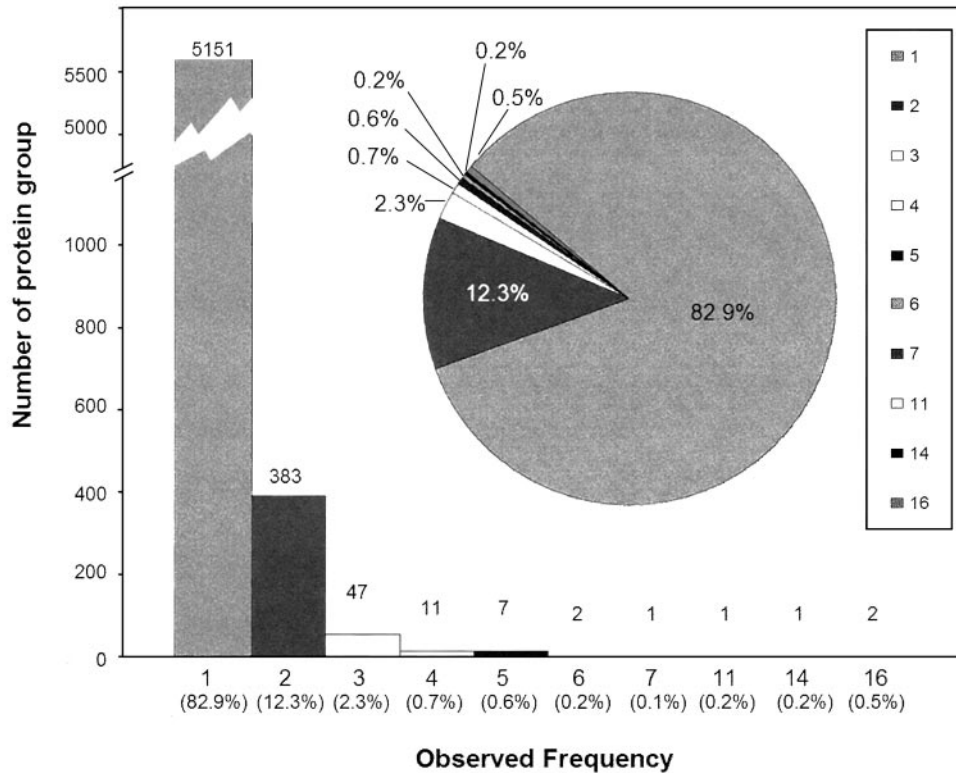


FIG. 1. Frequency distribution of *S. cerevisiae* proteins according to the sequence of the last 3 (A) and 4 residues (B). In the bar graphs, the y axis indicates the number of protein groups that have unique carboxyl-terminal ends, and the x axis represents how many proteins belong to each group and the corresponding percentage in the genome. Therefore, $\Sigma xy = 6,213$. Numbers inside the pie chart represent the percentage of total proteins comprising each group. The percentage value of each group is also identified in parentheses below each group.

TABLE II
Summary of yeast proteins ending with HDEL

Genes	Localization	Microenvironment	Cellular role
<i>PDI1</i>	Endoplasmic reticulum/microsomal fraction	Soluble	Protein disulfide isomerase: roles in protein folding, protein modification, protein degradation
<i>SED4</i>	Endoplasmic reticulum	Integral membrane	Vesicular transport
<i>YDR057W</i>	Endoplasmic reticulum	Unknown	Unknown
<i>CYP5</i>	Endoplasmic reticulum	Soluble	Isomerase/chaperones: roles in protein folding and vesicular transport
<i>SEC20</i>	Other vesicles of the secretory/endocytic pathways	Integral membrane	Vesicular transport
<i>EUG1</i>	Endoplasmic reticulum	Soluble	Isomerase/chaperones: roles in protein folding, protein modification, protein degradation, cell stress
<i>KAR2</i>	Endoplasmic reticulum	Soluble	Heat shock protein/hydrolase/ATPase, chaperones: roles in protein folding/protein translocation
<i>LHS1</i>	Endoplasmic reticulum	Soluble	Chaperones: protein translocation
<i>MPD2</i>	Endoplasmic reticulum	Unknown	Oxidoreductase, isomerase: roles in protein folding, protein modification, protein degradation
<i>MPD1</i>	Endoplasmic reticulum	Soluble	Isomerase, chaperones: roles in protein folding, protein modification, protein degradation
<i>KRE5</i>	Endoplasmic reticulum	Soluble	Transferase: roles in protein modification, cell wall maintenance

These motifs are the termini of the same sets of nearly identical proteins that appear to have been duplicated in the genome.² Consistent with this notion, almost identical numbers of hits of these three sequences are also found as the top hits of carboxyl-terminal 5-mers, 6-mers, 7-mers, and 8-mers (Table I). Other motifs include HDEL (11 hits) and A(M/L)LL (11 hits) tetrapeptide motifs, SKL (13 hits) and LSK (12 hits) tripeptide motifs, and two different dilysine motifs (XKK (152 hits) and KKXX (112 hits)) (see Table I).² Because these hits come from proteins of unrelated sequences, it is consistent with the notion that the conservation of these motifs reflects certain shared properties. It is also quite apparent that the most abundant tri- or tetrapeptide motifs are limited to less than 0.2% of total proteins in the proteome. A similar percentage but with different sequences has also been found in other proteomes such as *D. melanogaster*, *C. elegans*, and *H. sapiens* (see Table V).

The carboxyl HDEL motif is a well known recognition signal that directs proteins to ER retention and/or retrieval. Interestingly 10 of the 11 hits have been reported to be ER-specific proteins despite the fact that they are quite different in terms of overall polypeptide length, protein domain organization, biochemical function, and membrane topology (Table II). The only one within the group that is not strictly for the ER localization is Sec20, an integral membrane protein that is involved in vesicular trafficking in ER and Golgi apparatus (14, 15). While the localization information is incomplete, six of the eight HDEL sequence entries from the *Drosophila* sequence database are ER-resident proteins (Table III and the data from www.molecularinteraction.org/listofpublication.htm). Thus,

the HDEL-ended proteins, which represent ~0.2% of total proteins in the yeast proteome, could have been used to predict their shared properties with high confidence, which raises questions concerning other conserved but functionally unknown motifs.

Another highly conserved tetrapeptide motif is A(M/L)LL, which also has 11 hits. Of the 11 proteins, with exceptions of one reported to be in nuclei and one whose localization is currently unknown, the remaining nine proteins are found to have cell wall-related function (Table IV). These nine proteins display difference in length and/or domain organization (Fig. 2). Among them, the mannoprotein genes *DAN1*, *TIR1*, *TIR2*, *TIR3*, *TIR4*, and *TIP1* were shown to be expressed after an anaerobic shift or during cold shock, whereas the *CWP2* gene was down-regulated under the same conditions, suggesting that the seven mannoprotein genes are involved in remodeling of the cell wall (16). The remaining two of the nine proteins, Ylr110c and Ylr194c, are also found to be localized in the cell wall. It is interesting that these nine proteins are all involved in the same pathway and that a subset of these proteins was also identified by other informatic approaches (17). While the precise biochemical role of this motif is currently unknown, the terminal region might be recognized and cleaved prior to the surface expression of these proteins via a glycosylphosphatidylinositol anchor (12).

Conserved motifs with hits less than 10 genes may also be of great significance, but more analyses may be needed. For example, the KSKK motif was found in seven proteins (Rpl34b, Prp21, Cbf5, Ylr302c, Nop12, Ri01, and Svl3). Most of them are associated with RNA-involved structures or functions such as spliceosome and RNA transport.²

Similar analyses also showed that 10 of 13 proteins with the SKL terminal sequence are peroxisomal proteins. This motif is

² J.-J. Chung and M. Li, www.molecularinteraction.org/listofpublication.htm.

TABLE III
Summary of *Drosophila* proteins ending with HDEL

	Gene (GI no.)	Gene name ^a	Protein encoded
1	19921464	CG6453 gene product	α -Glucosidase
2	18859803	CG1837 gene product	Protein disulfide isomerase
3	24645441	CRC gene product	Calreticulin
4	21357739	CG5520 gene product	Glycoprotein 93
5	24586105	SPN4 gene product from transcript CG9453-RB	Serine protease inhibitor 4
6	24586107	SPN4 gene product from transcript CG9453-RA	Serine protease inhibitor 4
7	24657035	FKBP13 gene product from transcript CG9847-RB	Fkbp13
8	17352457	FKBP13 gene product from transcript CG9847-RA	Fkbp13

^a The alternative transcripts of a gene are labeled with the suffixes -RA, -RB, -RC, etc. by FlyBase. These vary due to alternative splicing, variable exons, multiple poly(A) sites, or multiple promoters. Alternative transcripts may or may not encode different protein products.

TABLE IV
Summary of yeast proteins ending with A(M/L)LL

Genes	Localization	Microenvironment	Cellular role
<i>TIP1</i>	Cell wall	Unknown	Serine-rich protein (Tiplp/Tirlp) family, structural protein; cell stress, cell wall maintenance
<i>YBR113W</i>	Unknown	Unknown	Unknown
<i>DAN1</i>	Cell wall	Unknown	Serine-rich protein (Tiplp/Tirlp) family, structural protein; cell stress, cell wall maintenance
<i>YLR110C</i>	Cell wall	Unknown	Structural protein, cell wall maintenance
<i>YLR194C</i>	Cell wall	Peripheral membrane	Unknown
<i>MSH2</i>	nuclear	DNA-associated (direct or indirect)	ATPase, DNA-binding protein; DNA repair, recombination
<i>TIR1</i>	Cell wall	Unknown	Serine-rich protein (Tiplp/Tirlp) family, structural protein; cell stress, cell wall maintenance
<i>YILO11W(TIR3)</i>	Cell wall	Unknown	Serine-rich protein (Tiplp/Tirlp) family; cell wall maintenance
<i>CWP2</i>	Cell wall	Unknown	Serine-rich protein (Tiplp/Tirlp) family, structural protein; cell stress, cell wall maintenance, and others
<i>YOR009W(TIR4)</i>	Cell wall	Unknown	Serine-rich protein (Tiplp/Tirlp) family; cell wall maintenance
<i>TIR2</i>	Cell wall	Unknown	Serine-rich protein (Tiplp/Tirlp) family, structural protein; cell stress, cell wall maintenance

also known as peroxisomal targeting signal type 1 (PTS1). This result is consistent with SKL as a recognition signal for peroxisomal localization (10, 18). This conservation is also found in the higher organisms.² It should be noted, however, that not all conserved sequences give rise to detectable shared features such as subcellular localization. For example, the dilysine motif KKXX is thought to be a retention/retrieval signal in the secretory pathway (19). Of 112 proteins with the KKXX terminus, the shared features were not detectable.³ For processes such as protein ER retention, several other ER localization signals have been reported, and the localization of ER could be a transient step to serve as a check point of proper protein assembly (13). Thus, the presence of conserved sequence motifs may be more suitable for grouping genes with related function or property. But the converse is not valid because it is known that many proteins found in ER do not use the HDEL-mediated localization machinery.

The statistical significance of a given motif may be analyzed in a number of ways. For the yeast genome, the total number of genes (6,213) and the lengths of the corresponding proteins (a combined total length of 2,912,365 amino acids) are known. Therefore it permits a more detailed statistical analysis. Fig. 3 shows the overall frequency of 20 amino acids (*part IV*) and the frequency of each amino at a given carboxyl-terminal position (-1 to -20 with -1 being the last residue at the carboxyl terminus) (*part V*). Most noticeable is the position-specific bias of lysine at the carboxyl terminus. While it is apparent that there is certain bias of amino acid abundance at terminal positions, the abundance-corrected HDEL sequence probability is similar, 7.76×10^{-6} for internal and 8.38×10^{-6} for carboxyl-terminal sequences (Fig. 3, *part VI*). There are 2,893,726 (2,912,365 - 6,213 \times 3) possible (internal and terminal) tetrapeptide sequences but only 6,213 possible carboxyl-terminal tetrapeptide sequences. Thus, the probability of the entire yeast genome to have one protein with the terminal HDEL motif by chance is 5%. As shown, 11 different proteins with terminal HDEL sequence have been identified

³ J.-J. Chung, H. Yang, and M. Li, unpublished results.

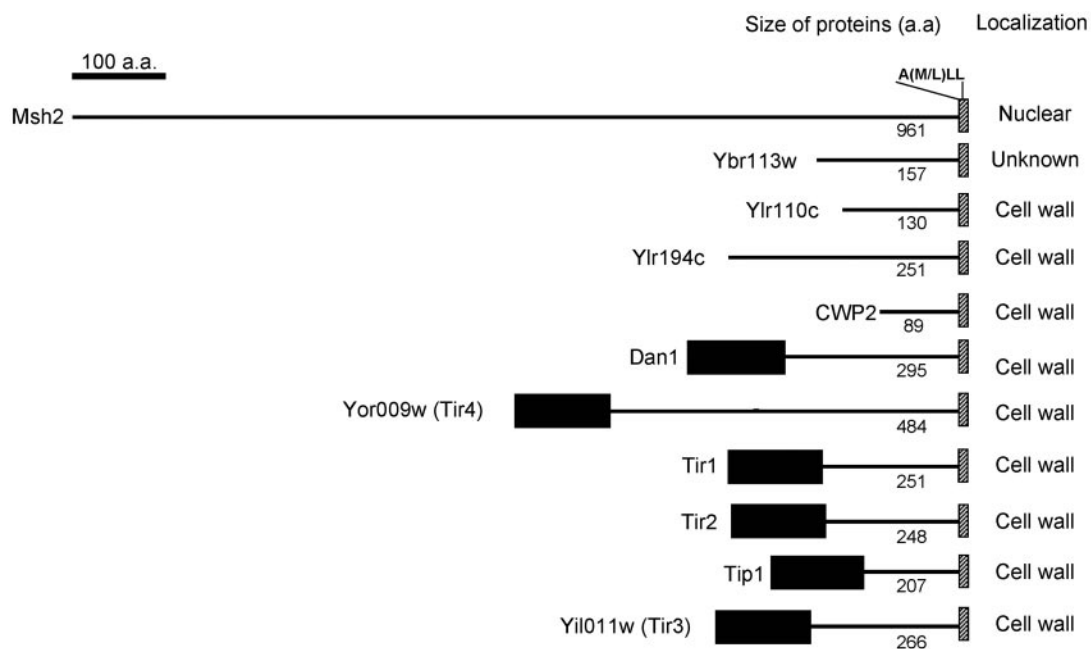


FIG. 2. Schematic diagram of a group of 11 proteins ending with A(M/L)LL found by the last 4 residues as a possible functional index. Filled boxes represent the PAU domain, and hatched boxes represent carboxyl-terminal A(M/L)LL. The length and predicted localization of the proteins are as indicated. The boxes and lines are drawn in scale, and the scale bar is shown on the top of the diagram. a.a., amino acids.

(Table II), representing more than 200-fold above the expected frequency. Similar calculations were performed for A(M/L)LL and SKL, which were 26- and 2-fold above the expected frequency (Fig. 3, part VI). Using this information, one may also calculate the odds of finding the same sequence by chance.

DISCUSSION

Short linear terminal epitopes in proteins as sites for recognition are different from that of internal sequences in a number of ways; most obviously, each protein has only one carboxyl terminus. Most proteins of known structure have their carboxyl terminus exposed, consistent with a role for recognition that could lead to several possible biochemical events including static binding, cleavage (exposed to new terminal sequences), or posttranslational modification (for a review, see Ref. 1). The analyses of yeast proteome (Fig. 3) suggest that it is possible to estimate the number of internal residues needed to confer a similar level of diversity conferred by 5 terminal residues. Generally there are about 500 pentapeptide sequences in a 500-amino acid-long protein but only one carboxyl-terminal pentapeptide. Thus, to confer the same specificity in an internal peptide one needs to make the probability of a random occurrence 500 times less likely; this is accomplished by extending the length by 2 residues ($(1/20) \times (1/20) = 1/400$). In this context, it is intriguing that immunoglobulin or T-cell receptor binding, when recognizing internal peptide, involves 7–9 amino acids in length. In contrast, a PDZ domain binding to carboxyl-terminal peptide recognizes 3–5 residues. Indeed about 500 pentapeptide se-

quences in a 500-residue-long protein is overestimated since each adjacent peptide differs by 1 residue. Interestingly the strongly biased positions in protein carboxyl termini are most pronounced at the last 4 or 5 residues, which according to the calculation here provides sufficient diversity for a typical proteome.

The resultant data in this report have been deposited and are available at www.molecularinteraction.org/listofpublication.htm. The specific abundance of certain amino acids may be related to free energy considerations. For example, the preference for lysine at the carboxyl termini could stem from electrostatic stabilization of helix dipoles. The sequences may be mined in other ways as well. For example, the sequences may be sorted by predetermination of amino acid residues at any given position, such as CAAX or KKXX. Our result suggests that the conserved carboxyl-terminal sequence alone may confer certain important biochemical function, although some of these functions are currently unknown. Thus, these conserved motifs may serve as one criterion for grouping diverse proteins with certain shared properties including biochemical function, membrane association, and/or protein domain organization. It supports the notion that the high information contents at the carboxyl termini encode the signatures for certain fundamental functions including but not limited to subcellular localization (1). In *C. elegans*, HDEL and KDEL were found in seven and four proteins, respectively. Whether the *C. elegans* ER retention machinery is more degenerate in recognizing its anchor site than that of other species remains unanswered.

It is important to note that conserved motifs are likely to be

TABLE V
Top 30 hits of carboxyl-terminal sequences in higher eukaryotes

	<i>C. elegans</i>				<i>D. melanogaster</i>				<i>H. sapiens</i>			
	X_3		X_4		X_3		X_4		X_3		X_4	
1	KKK	73	YLCE	45	KKK	68	EEVD	15	DEL	54	KEKK	28
2	LCE	46	EYNP	37	SKL	33	KKKK	14	LLL	53	EFMA	25
3	YNP	38	APGY	26	DED	31	HGGM	14	SSS	49	TSSK	23
4	SKL	38	ATKY	26	AKL	31	KDEL	13	KKK	49	VLRH	22
5	PGY	36	TTNS	20	DEL	30	SDEV	12	SSL	46	GERA	20
6	GKK	32	GTRR	19	AKK	30	WRPW	12	LLS	45	EEVD	20
7	TKY	32	PINY	18	SSS	27	SAAN	11	EEL	44	SLKF	19
8	SSK	31	GERA	18	LKK	27	LKTG	11	ASS	43	CWNK	19
9	SKK	29	TSSK	18	LLK	24	LLKK	11	LLK	43	GFGG	19
10	DSD	28	KKKK	18	TEL	24	GISY	11	SSK	41	AKGK	18
11	DDE	28	GDKE	17	KSN	24	NRRY	10	SLL	39	NSDK	18
12	KKN	27	GFGG	17	KSK	23	RKCF	10	EKK	38	QKAK	17
13	DEL	27	CRIC	16	SSN	23	SDED	10	KAK	38	TKLG	16
14	DEE	26	AFDH	12	GKK	22	AASQ	10	LGL	37	EEEE	15
15	GRK	26	SDSD	12	SKK	22	ENEF	10	PSS	37	RKCF	15
16	LKK	25	GKKK	11	KLK	22	WAFV	9	PAS	35	SCGF	13
17	RRR	25	SATA	10	RRY	21	EIDN	9	EEE	35	RRRR	13
18	KKL	25	FGRK	10	TDL	20	RAKL	9	RRR	34	SDSD	13
19	RRK	25	RRRR	10	AAN	20	GGDN	9	LTL	34	LVCQ	13
20	AKL	24	MKQH	10	ASK	20	EEEE	9	TSL	33	HDEL	13
21	EKK	23	PKKK	9	KAK	20	VERA	9	GSS	33	KDEL	13
22	KRK	23	YLGP	9	ERA	20	CRTS	9	SVL	33	ASKE	13
23	KKE	23	GCYQ	9	KRK	19	ESKL	9	SSG	32	CGQL	12
24	KKI	22	PPGY	9	DSD	19	NKKK	9	SPS	32	EDTM	12
25	TNS	22	KKKN	9	IFG	19	LRSE	9	LRH	31	GEKP	12
26	AKK	22	GWGK	8	LKL	18	DDED	9	LKF	31	GRRF	12
27	RKL	21	VQFC	8	LQK	18	TASK	9	PGP	31	KEEL	12
28	TRR	21	KSWE	8	TTA	18	MIIE	9	EKL	31	SSSS	12
29	SKN	21	EDDE	8	EKL	18	FQDV	8	PSP	31	CNKI	11
30	FGK	20	PDSP	7	NRR	18	KKYK	8	KGK	31	IGII	11

annotated protein sequences become available, it will be important to mine both types of motifs via informatic tools, which could provide interesting hypotheses that can be tested experimentally.

Acknowledgment—We thank Dr. Yoshiro Hanyu for helpful comments on this manuscript.

* The work in the Li laboratory was supported by grants from the National Institutes of Health (to M. L.) and an established investigator award from the American Heart Association (to M. L.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Both authors contributed equally to this work.

§ A predoctoral fellow of the American Heart Association.

¶ To whom correspondence should be addressed: Dept. of Neuroscience, WBSB 216, Johns Hopkins University School of Medicine, 725 N. Wolfe St., Baltimore, MD 21205. Tel.: 410-614-3692; Fax: 410-614-1001; E-mail: minli@jhmi.edu.

REFERENCES

- Chung, J. J., Shikano, S., Hanyu, Y., and Li, M. (2002) Functional diversity of protein C-termini: more than zipcoding? *Trends Cell Biol.* **12**, 146–150
- Doyle, D. A., Lee, A., Lewis, J., Kim, E., Sheng, M., and MacKinnon, R. (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* **85**, 1067–1076
- Gatto, G. J., Jr., Geisbrecht, B. V., Gould, S. J., and Berg, J. M. (2000) A proposed model for the PEX5-peroxisomal targeting signal-1 recognition complex. *Proteins* **38**, 241–246
- Stricker, N. L., Christopherson, K. S., Yi, B. A., Schatz, P. J., Raab, R. W., Dawes, G., Bassett, D. E., Jr., Bredt, D. S., and Li, M. (1997) PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences. *Nat. Biotechnol.* **15**, 336–342
- Wang, S., Raab, R. W., Schatz, P. J., Guggino, W. B., and Li, M. (1998) Peptide binding consensus of the NHE-RF-PDZ1 domain matches the C-terminal sequence of cystic fibrosis transmembrane conductance regulator (CFTR). *FEBS Lett.* **427**, 103–108
- Moyer, B. D., Duhaime, M., Shaw, C., Denton, J., Reynolds, D., Karlson, K. H., Pfeiffer, J., Wang, S., Mickle, J. E., Milewski, M., Cutting, G. R., Guggino, W. B., Li, M., and Stanton, B. A. (2000) The PDZ-interacting domain of cystic fibrosis transmembrane conductance regulator is required for functional expression in the apical plasma membrane. *J. Biol. Chem.* **275**, 27069–27074
- Wang, S., Yue, H., Derin, R. B., Guggino, W. B., and Li, M. (2000) Accessory protein facilitated CFTR-CFTR interaction, a molecular mechanism to potentiate the chloride channel activity. *Cell* **103**, 169–179
- Teasdale, R. D., and Jackson, M. R. (1996) Signal-mediated sorting of membrane proteins between the endoplasmic reticulum and the Golgi apparatus. *Annu. Rev. Cell Dev. Biol.* **12**, 27–54
- Zhang, F. L., and Casey, P. J. (1996) Protein prenylation: molecular mechanisms and functional consequences. *Annu. Rev. Biochem.* **65**, 241–269
- Fujiki, Y. (2000) Peroxisome biogenesis and peroxisome biogenesis disorders. *FEBS Lett.* **476**, 42–46
- Berezovsky, I. N., Kilosanidze, G. T., Tumanyan, V. G., and Kisselev, L. L. (1999) Amino acid composition of protein termini are biased in different manners. *Protein Eng.* **12**, 23–30
- Berezovsky, I. N., Kilosanidze, G. T., Tumanyan, V. G., and Kisselev, L. (1997) COOH-terminal decamers in proteins are non-random. *FEBS Lett.*

- 404, 140–142
13. Zerangue, N., Schwappach, B., Jan, Y. N., and Jan, L. Y. (1999) A new ER trafficking signal regulates the subunit stoichiometry of plasma membrane K(ATP) channels. *Neuron* **22**, 537–548
 14. Schleip, I., Heiss, E., and Lehle, L. (2001) The yeast SEC20 gene is required for N- and O-glycosylation in the Golgi. Evidence that impaired glycosylation does not correlate with the secretory defect. *J. Biol. Chem.* **276**, 28751–28758
 15. Cosson, P., Schroder-Kohne, S., Sweet, D. S., Demolliere, C., Hennecke, S., Frigerio, G., and Letourneur, F. (1997) The Sec20/Tip20p complex is involved in ER retrieval of dilysine-tagged proteins. *Eur. J. Cell Biol.* **73**, 93–97
 16. Abramova, N., Sertil, O., Mehta, S., and Lowry, C. V. (2001) Reciprocal regulation of anaerobic and aerobic cell wall mannoprotein gene expression in *Saccharomyces cerevisiae*. *J. Bacteriol.* **183**, 2881–2887
 17. Caro, L. H., Tettelin, H., Vossen, J. H., Ram, A. F., van den Ende, H., and Klis, F. M. (1997) In silico identification of glycosyl-phosphatidylinositol-anchored plasma-membrane and cell wall proteins of *Saccharomyces cerevisiae*. *Yeast* **13**, 1477–1489
 18. Amery, L., Brees, C., Baes, M., Setoyama, C., Miura, R., Mannaerts, G. P., and Van Veldhoven, P. P. (1998) C-terminal tripeptide Ser-Asn-Leu (SNL) of human D-aspartate oxidase is a functional peroxisome-targeting signal. *Biochem. J.* **336**, 367–371
 19. Gaynor, E. C., te Heesen, S., Graham, T. R., Aebi, M., and Emr, S. D. (1994) Signal-mediated retrieval of a membrane protein from the Golgi to the ER in yeast. *J. Cell Biol.* **127**, 653–665