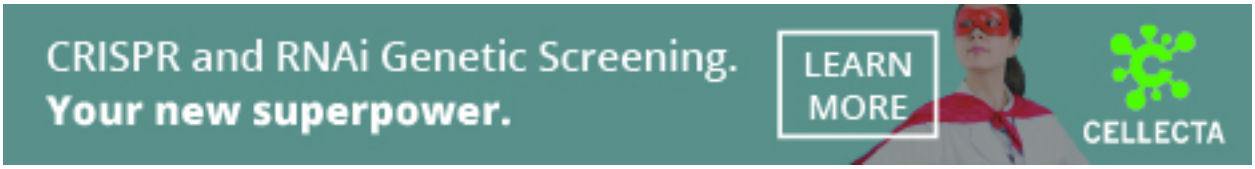
Nonrandom Tripeptide Sequence Distributions at Protein Carboxyl Termini

Gregory J. Gatto, Jr. and Jeremy M. Berg

Genome Res. 2003 13: 617-623
Access the most recent version at doi:10.1101/gr.667603



Letter

Nonrandom Tripeptide Sequence Distributions at Protein Carboxyl Termini

Gregory J. Gatto Jr. and Jeremy M. Berg[1]

Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

The availability of complete genome sequences enables the statistical analysis of sequence features without significant database-imposed bias. The carboxyl termini of proteins often contain regions associated with protein targeting and enhanced translational termination. We analyzed the frequency of occurrence of C-terminal tripeptides in representative archaeal, bacterial, and eukaryotic genomes. The sequence distribution in prokaryotic genomes nearly matches that generated by the randomization of the observed tripeptide set. In contrast, eukaryotic genomes contain large numbers of overrepresented sequences. Some of these correspond to highly repeated sequences from either duplicated endogenous genes or transposon open reading frames. Gratifyingly, others represent previously known targeting signals or sequences associated with an increase in translational termination efficiency. However, a number of overrepresented tripeptides have not been previously noted and may represent novel functional sequences. For example, the sequence XSS may enhance translational termination efficiency in plants, whereas FWC may be a targeting or processing signal for certain amino acid permeases in yeast.

The complete genomic sequence of an organism provides a unique opportunity for insight into the mechanisms of biological processes. The generation of these data has proceeded at a remarkable rate. Initially performed on prokaryotes, complete genome sequencing efforts now have been successfully completed on a diverse set of eukaryotes as well, including the recent landmark achievement of a substantially complete human genome sequence (Lander et al. 2001; Venter et al. 2001). Once an entire genomic DNA sequence has been elucidated, it is possible to predict the location and identity of the entire complement of open reading frames (ORFs) and, hence, the sequences of essentially all the proteins for a given organism. Importantly, this set of ORFs contains little bias as to the nature of the protein product. One application for this information is the determination of patterns and correlations within protein sequences. Because of the large number of ORFs within a given genome, the utilization of statistical methods can be particularly powerful, enabling the distinction between chance occurrences and the results of biological selection. Moreover, one can compare the results of these analyses across a number of species and generate testable hypotheses about the evolutionary history of these sequence distributions.

One example of the potential utility of complete genome analysis which does not appear to have been extensively explored is the study of amino acid sequences located at the ends of the polypeptide chain. These residues often are of functional significance, as they can serve as determinants for interactions with other proteins (Chung et al. 2002). Binding sites located at the termini of proteins have particular advantages over internal sites. For example, they are commonly solvent-exposed and, hence, easily accessible by their appropriate binding partners. Moreover, these regions can evolve readily, as they usually do not require compensatory adjustments at some distant position along the polypeptide chain. In addition, a sequence that must be terminally located adds information content to the signal: The sequence alone is not sufficient, its position relative to either the amino or carboxyl terminus is also critical. From a computational perspective, terminal signal identification within a genomic sequence database is simplified considerably, assuming that the translation initiation and termination codons have been correctly identified.

One prominent example of recognition sites at the termini of proteins includes certain organelle targeting signals. A protein targeted to a particular organelle typically contains a short signal sequence that is sufficient to direct that protein to its appropriate compartment (Blobel and Dobberstein 1975). In some cases, these targeting signals are located at sequence termini. For example, the signal for many proteins to be retained within the lumen of the endoplasmic reticulum (ER) is the C-terminal tetrapeptide -Lys-Asp-Glu-Leu-COO^- (KDEL; Munro and Pelham 1987). Additionally, many proteins destined to reside within the lumen of the peroxisome contain the peroxisomal targeting signal-1 (PTS1), a C-terminal tripeptide with the consensus sequence -Ser-Lys-Leu-COO^- (SKL; Gould et al. 1989). Also, several known targeting signals are found at the N-termini of proteins, including the signal sequence for proteins entering the secretory pathway and the mitochondrial targeting sequence (Schatz and Dobberstein 1996).

In this paper, we present the use of complete genome information to study patterns of C-terminal tripeptide sequences within the entire ORF complement of a set of representative organisms. Genomes of the following species were studied: *Methanococcus jannaschii* (an archaeon), *Escherichia coli* (a bacterium), *Saccharomyces cerevisiae* (a yeast), *Arabidopsis thaliana* (a plant), *Caenorhabditis elegans* (a nematode), and *Homo sapiens* (humans). These genomes were analyzed by de-

[1]Corresponding author.
E-MAIL jberg@jhmi.edu; FAX (410) 502-6910.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.667603.

termining, for a given genome, how many occurrences of a tripeptide sequence can be defined as overrepresented beyond that expected by chance alone. To achieve this goal, the collection of C-terminal tripeptides from a given genome was randomized multiple times, then compared to the original set.

As might have been anticipated, analysis of the archaeal and bacterial genomes revealed little deviation from the randomized results. Many of the overrepresented sequences within the *E. coli* genome can be explained by transposon ORFs, repeated sequences, or possible effects of certain codons on translation termination efficiency. However, analysis of the eukaryotic genomes yielded a collection of sequences that appear distinctly overrepresented, even after ruling out homologous ORFs. Some of these sequences can be attributed to known targeting signals or binding sites, but others remain less well characterized. Possibilities for selected sequences will be discussed.

RESULTS

Sequence distributions were analyzed using two approaches. The first analysis method determines sequence significance by estimating how many times a sequence must occur to be considered unusually abundant. To accomplish this task, a tally was generated of the number of occurrences for each of the 8000 possible C-terminal tripeptides within a given genome. Then, a "dummy" tripeptide database was created in which the sets of amino acids used at positions −1, −2,

Table 1. Position-Specific Amino Acid Frequencies (Expressed as a Percentage) at the Three C-Terminal Positions for Each of the Genomes Studied

	M. jannaschii				*E. coli*				*S. cerevisiae*			
AA	−1	−2	−3	ORF	−1	−2	−3	ORF	−1	−2	−3	ORF
A	2.4	1.9	2.7	5.4	10.2	9.4	7.6	9.5	5.2	4.2	4.8	5.5
C	0.8	0.5	1.0	1.3	1.2	1.2	1.1	1.2	1.7	1.5	1.4	1.3
D	4.0	3.2	4.3	5.5	4.1	3.7	4.1	5.1	5.1	5.2	5.4	5.8
E	13.3	8.6	9.0	8.7	8.2	6.4	8.3	5.7	6.5	5.8	6.2	6.5
F	4.5	5.4	3.9	4.3	3.6	2.8	3.4	3.9	5.3	5.8	5.5	4.5
G	4.9	6.7	4.6	6.3	6.7	6.5	5.7	7.4	2.6	4.5	4.6	5.0
H	1.4	1.1	1.2	1.4	4.3	2.6	2.7	2.3	2.7	2.3	2.3	2.2
I	10.0	9.6	11.2	10.5	3.7	4.0	5.2	6.0	7.1	6.3	6.6	6.6
K	17.2	17.6	14.8	10.4	10.9	8.7	7.0	4.4	11.5	10.9	8.8	7.3
L	11.7	11.2	12.3	9.5	7.9	9.6	10.1	10.6	10.5	9.3	10.3	9.6
M	0.6	2.6	2.3	2.3	1.3	2.1	1.7	2.8	2.2	1.9	2.3	2.1
N	4.2	6.1	5.4	5.3	4.0	4.4	4.2	4.0	7.1	5.5	5.8	6.1
P	1.5	1.4	2.0	3.4	2.9	3.2	4.7	4.4	2.3	2.9	3.6	4.3
Q	3.7	1.9	1.3	1.4	6.6	5.4	4.7	4.4	4.4	4.0	4.1	3.9
R	5.6	6.5	6.3	3.8	8.7	7.5	7.9	5.5	4.7	6.5	4.9	4.5
S	3.8	4.6	5.2	4.5	6.2	7.0	5.9	5.8	6.7	8.7	8.3	9.0
T	1.9	3.2	2.7	4.0	1.1	5.3	4.7	5.4	3.8	5.3	4.6	5.9
V	2.9	4.2	5.5	6.8	4.8	6.3	6.5	7.1	5.3	4.6	5.7	5.6
W	1.5	0.8	0.9	0.7	1.6	1.1	1.5	1.5	1.7	1.0	1.3	1.0
Y	4.1	2.8	3.6	4.4	2.2	2.8	3.0	2.9	3.5	3.8	3.7	3.4
	C. elegans				*A. thaliana*				*H. sapiens*			
AA	−1	−2	−3	ORF	−1	−2	−3	ORF	−1	−2	−3	ORF
A	5.2	4.1	4.8	6.3	6.2	4.9	5.5	6.3	5.3	6.0	6.4	7.0
C	2.7	2.2	2.0	2.1	2.6	2.2	2.0	1.8	3.2	2.7	2.4	2.2
D	4.3	4.7	4.6	5.3	4.5	4.8	4.9	5.5	4.5	5.0	4.4	4.9
E	5.7	5.5	5.4	6.5	5.1	5.8	5.1	6.8	5.1	6.9	6.8	7.1
F	8.5	5.7	5.7	4.9	5.7	5.0	5.0	4.3	4.6	3.4	3.8	3.7
G	3.0	4.4	4.2	5.3	3.7	5.5	4.9	6.4	3.7	6.2	6.1	6.8
H	3.4	2.6	2.2	2.3	2.6	2.5	2.4	2.3	3.5	2.8	2.7	2.5
I	6.4	6.1	6.3	6.2	5.4	5.1	5.1	5.3	4.5	3.7	3.6	4.4
K	8.8	10.1	8.7	6.5	6.2	7.7	7.1	6.4	7.5	7.9	7.2	5.7
L	9.3	8.1	8.7	8.7	10.2	8.6	9.9	9.5	11.1	8.5	9.2	9.9
M	1.9	2.2	2.1	2.6	2.1	2.0	2.1	2.4	2.1	1.9	2.1	2.2
N	7.4	6.6	5.1	4.9	5.1	4.5	4.2	4.4	4.0	3.8	3.3	3.7
P	2.6	3.8	4.6	4.9	4.2	4.3	5.0	4.8	5.6	6.1	6.2	6.2
Q	5.0	4.0	3.9	4.1	3.2	3.5	3.9	3.5	4.8	4.8	4.6	4.7
R	4.5	6.7	5.9	5.2	6.6	7.2	6.3	5.4	5.2	6.4	5.6	5.6
S	6.7	8.6	9.0	8.0	10.0	10.8	10.7	9.0	9.5	9.4	10.0	8.0
T	3.2	4.8	6.5	5.8	4.6	5.6	5.4	5.1	4.9	5.2	6.4	5.3
V	6.1	5.3	5.8	6.2	6.9	5.6	6.0	6.7	6.5	4.9	4.9	6.1
W	1.1	1.4	1.1	1.1	1.5	1.5	1.4	1.3	1.4	1.5	1.6	1.2
Y	4.3	3.1	3.4	3.2	3.7	3.1	3.2	2.9	3.1	2.8	2.6	2.7

The fourth column for each organism (ORF) shows the amino acid frequencies across the entire ORF array.

and −3 across all ORFs were independently jumbled, then rejoined to form a new set of tripeptides. Typically, 1000 iterations of this jumbling procedure were performed. The sequence tallies from these randomized "dummy" sets were then averaged and compared with the observed tally. The importance of performing this analysis in a position-dependent fashion for each organism is highlighted in Table 1. There is considerable variation in the amino acid usage profile not only from organism to organism, but also from the terminal residues to the rest of the protein sequence within a given genome. The scrambling of an existing set of amino acids in a position-dependent manner thus takes into account this unequal distribution of amino acid frequencies. Moreover, the repetitive nature of this analysis allows for the calculation of the distributions for the expected number of sequences that occur a particular number of times.

A sampling of the results from the genome jumbling method is shown in graphic form in Figure 1, in which the number of sequences is plotted as a function of the number of occurrences for each sequence. For a prokaryote such as the archaeon *Methanococcus jannaschii* (Fig. 1A), the observed number of sequence occurrences generally falls within one standard deviation of the randomized sets. These data contrast sharply with the results from the eukaryote *Saccharomyces cerevisiae* (Fig. 1B). For this organism, there is a collection of sequences that appear in the genome more times than predicted by chance alone. Table 2 lists the nonrandomly distributed sequences observed in all the genomes studied. The expected and observed number of sequences with a given number of occurrences is shown. Where possible, a likely reason why that sequence might be highly repeated is indicated (see below).

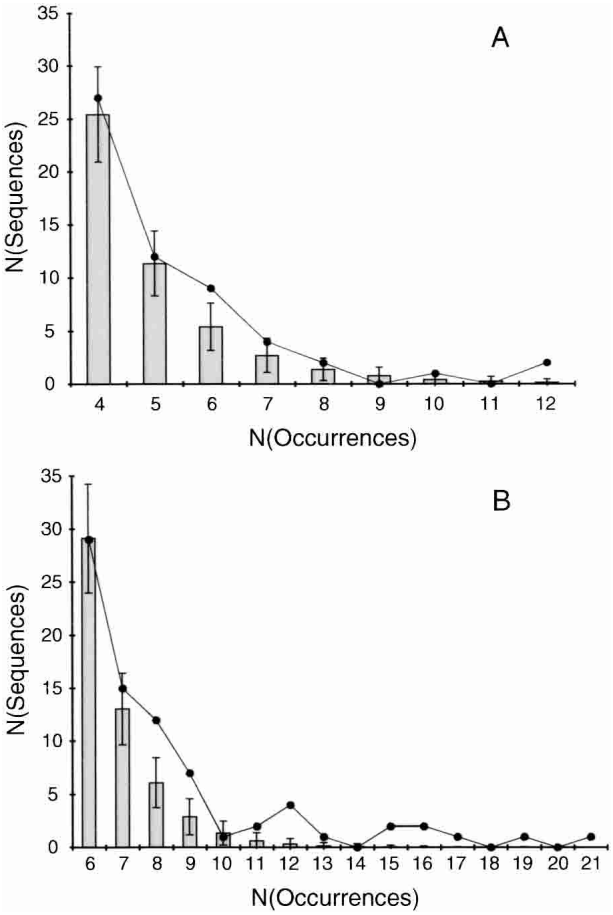


Figure 1 Comparison of observed vs. expected sequence occurrences in *Methanococcus jannaschii* (*A*) and *Saccharomyces cerevisiae* (*B*). The number of tripeptide sequences is plotted as a function of the number of times that that sequence occurs in the genome. The line indicates the genomic data, and the bars show the results from 1000 iterations of the jumbling procedure. Error bars are drawn at one standard deviation.

This method identifies a set of tripeptides that can be considered overrepresented based on the number of times these sequences appear within a given genome. To further highlight the importance of a particular sequence, the expected number of occurrences for each individual tripeptide was determined based on positional amino acid frequencies. If $f(a,b)$ is the frequency of amino acid a at position b, then the expected number of occurrences of the C-terminal tripeptide XYZ, N_{Exp}(XYZ), can be expressed as the product:

$$N_{Exp}\,(\mathrm{XYZ}) = f\,(\mathrm{X},-3)\cdot f\,(\mathrm{Y},-2)\cdot f\,(\mathrm{Z},-1)\cdot (N_{ORF})$$

where N_{ORF} is the total number of ORFs within the genome of the organism. Comparison of N_{Exp}(XYZ) with N_{Obs}(XYZ), the observed number of occurrences of XYZ, can give an indication of the significance of that sequence. Although simple to calculate, the results from this frequency method must be interpreted carefully. Sequences that have a relatively low number of occurrences but contain the less frequent amino acids (such as cysteine and tryptophan) tend to show inflated N_{Obs}(XYZ)/N_{Exp}(XYZ) ratios. The results from the frequency method for each sequence, expressed as the N_{Obs}(XYZ)/N_{Exp}(XYZ) ratio, are also presented in Table 2. The application of these ratios to the sequences identified in the jumbling analysis serves to emphasize particular sequences over others.

DISCUSSION

Reasons for High-Frequency Tripeptides

Although a major goal of this study was to identify previously unrecognized binding sites and signals, there are additional expected causes for overrepresented C-terminal tripeptides. First, many organisms possess highly conserved proteins, the genes for which have been duplicated many times within their genomes. For example, the human genome contains multiple copies of each of the histone-encoding genes. Thus, although the sequence KGK is observed 17 times, 11 of them can be attributed to isoforms of the histone protein H2A. In other cases, the actual protein product itself is less well understood. In *A. thaliana*, there exists a large novel protein family named AtPCMP (Aubourg et al. 2000). The members of this family appear to be unique to plants, and their function is not known. Of this group, 57 ORFs contain the "H motif" at their C-terminus. Consequently, they end with the same tripeptide, DYW.

A related reason for finding highly repeated C-terminal sequences is the result of multiple copies of transposon proteins within the genome. In *S. cerevisiae*, the sequence WIH is found 16 times. This tripeptide corresponds to the final three amino acids of the transposon Ty1 gag-pol protein product. All 16 ORFs from this genome ending in WIH encode such

Table 2. Most Frequent Tripeptide Sequences Observed Within the Genomes Studied

Organism	N(Occ)	N(Seq) (expected)	N(Seq) (observed)	Sequences observed
M. jannaschii	12	0.1 ± 0.3	2	KEE (4.0), LKK (1.8)
(1773 ORFs)	10	0.4 ± 0.6	1	KKL (1.9)
	8	1.4 ± 1.1	2	KKK (1.0), LKE (1.6)
	7	2.7 ± 1.6	4	IIK (2.1), KKE (1.1), LNK (3.0), *RLL* (4.8)
	6	5.4 ± 2.2	9	EKE (1.6), EKL (1.8), IKK (1.0), KIE (1.8), KKD (3.3), KKI (1.3), RKK (1.8), VKE (2.6), VKK (2.0)
E. coli	11	0.0 ± 0.2	1	AKK (3.5)
(4290 ORFs)	10	0.1 ± 0.3	3	KKK (3.5), *RSH* (9.7), *RSR* (4.8)
	9	0.4 ± 0.6	2	EAK (2.5), *RLK* (2.5)
	8	1.2 ± 1.1	7	AAQ (3.9), EEA (3.5), EVK (3.3), GLL (4.3), *LEI* (8.0), LLG (2.9), RRG (4.7)
	7	3.6 ± 1.8	8	*DGE* (7.4), EEV (6.5), GGK (4.1), KLA (2.4), LAS (2.8), *NLA* (4.0), RRR (3.1), SEE (5.4)
S. cerevisiae	21	0.0 ± 0.0	1	SKK (3.3)
(6215 ORFs)	19	0.0 ± 0.0	1	KKK (2.8)
	17	0.0 ± 0.1	1	*VGE* (16.5)
	16	0.0 ± 0.1	2	AKK (4.3), *WIH* (120.2)
	15	0.0 ± 0.2	2	**DEL** (7.3), *IAN* (12.0)
	13	0.1 ± 0.3	1	**SKL** (2.2)
	12	0.3 ± 0.5	4	EKK (1.5), LKK (1.5), LLL (1.9), LSK (1.9)
	11	0.6 ± 0.7	2	KKE (2.9), LLK (1.6)
	10	1.3 ± 1.1	1	GKK (2.8)
	9	2.9 ± 1.7	7	DEE (7.1), DSK (2.7), FWC (158.6), LSI (2.3), MLL (6.5), QKI (4.6), SSS (3.0)
	8	6.1 ± 2.4	12	DDE (7.0), **EVD** (8.8), *IPK* (5.8), KEK (2.2), KKD (2.6), KKN (1.9), LDL (2.3), LLV (2.6), RRK (3.5), SLA (3.2), SSL (1.7), TKK (2.2)
A. thaliana	99	0.0 ± 0.0	1	SSS (3.4)
(25561 ORFs)	54	0.0 ± 0.0	1	*DYW* (95.3)
	43	0.0 ± 0.1	1	SSL (1.4)
	40	0.0 ± 0.2	1	ASS (2.6)
	39	0.1 ± 0.2	2	**DEL** (5.2), TSS (2.6)
	38	0.1 ± 0.3	1	**SKL** (1.8)
	36	0.2 ± 0.4	1	LKL (1.8)
	34	0.2 ± 0.5	1	LLS (1.6)
	32	0.3 ± 0.6	1	EEE (8.2)
	31	0.4 ± 0.6	1	SST (2.3)
	30	0.5 ± 0.7	2	LSS (1.1), STS (2.0)
	29	0.6 ± 0.8	4	KKK (3.4), LLL (1.3), PSS (2.1), RRR (3.8)
	28	0.8 ± 0.9	2	SSI (1.8), VSS (1.7)
	26	1.2 ± 1.1	3	DSD (4.3), GSS (1.9), LVF (3.2)
	25	1.6 ± 1.3	2	DEE (6.7), KKR (2.7)
	24	1.9 ± 1.3	4	FLL (2.1), FSS (1.7), LSL (0.9), RRS (2.1)
	23	2.5 ± 1.5	8	DDE (7.5), EED (6.8), SFL (1.7), SLL (1.0), SSR (1.2), SSV (1.1), VSA (2.3), VTL (2.7)
C. elegans	70	0.0 ± 0.0	1	KKK (4.6)
(19833 ORFs)	45	0.0 ± 0.0	1	*LCE* (20.4)
	38	0.0 ± 0.0	2	**SKL** (2.3), *YNP* (33.7)
	36	0.0 ± 0.0	1	*PGY* (20.8)
	32	0.0 ± 0.0	2	GKK (4.3), *TKY* (5.8)
	30	0.0 ± 0.0	1	*SSK* (2.2)
	28	0.0 ± 0.1	3	DDE (11.5), KKN (2.2), SKK (1.8)
	26	0.1 ± 0.2	2	DSD (7.7), RRK (3.8)
	24	0.2 ± 0.4	5	AKK (2.8), DEE (8.4), KKL (1.5), KRK (2.4), LKK (1.6)
	23	0.3 ± 0.5	5	**AKL** (2.6), **DEL** (4.9), GRK (4.6), KKE (2.3), KKI (2.1)
	22	0.4 ± 0.6	3	EKK (2.3), SKN (1.6), *TNS* (3.9)
	21	0.7 ± 0.8	1	*TRR* (5.4)
	20	1.1 ± 1.1	4	*ERA* (5.3), KKQ (2.3), RKL (1.8), RRR (5.7)
	19	1.8 ± 1.3	7	*DKE* (3.6), FGK (4.3), *INY* (5.4), LGL (2.8), NKK (3.1), SSF (1.5), VSS (2.9)
	18	2.6 ± 1.5	9	EKL (1.8), *FGG* (12.2), KSE (2.1), LFN (2.5), LKI (1.6), RIC (9.5), SRR (3.3), SSS (1.8), VKK (1.8)
H. sapiens	32	0.0 ± 0.0	1	**DEL** (6.3)
(14760 ORFs)	31	0.0 ± 0.0	1	*EKK* (5.3)
	28	0.0 ± 0.0	1	KKK (4.5)
	25	0.0 ± 0.1	1	*LKF* (5.1)
	22	0.1 ± 0.3	1	EEE (6.3)
	21	0.2 ± 0.4	2	LLL (1.6), *SDQ* (6.0)
	20	0.3 ± 0.5	2	LAL (2.2), *SSK* (1.9)
	19	0.4 ± 0.6	3	**EEL** (2.5), LLK (2.2), *WNK* (28.0)

(continued)

Table 2. *Continued*

Organism	N(Occ)	N(Seq) (expected)	N(Seq) (observed)	Sequences observed
	18	0.7 ± 0.8	3	ASS (2.1), TRL (2.7), TSL (1.8)
	17	1.0 ± 1.0	6	*KGK* (3.4), KRK (3.3), LGL (1.6), LLS (1.6), RKK (3.5), SLL (1.2)
	16	1.6 ± 1.3	5	EDD (7.1), RRR (5.8), SES (1.7), **SKL** (1.2), TEL (2.2)
	15	2.7 ± 1.6	9	GSS (1.9), KRR (4.2), *NKI* (8.5), PSS (1.8), RRK (3.8), SSL (1.0), SSS (1.2), TKL (1.8), TVV (5.0)
	14	4.2 ± 2.0	9	APL (2.2), EKP (3.2), *ERA* (4.1), GKK (2.6), KSS (1.5), LVS (2.2), PGP (4.4), SCC (11.1), TEV (3.3)
	13	6.5 ± 2.4	13	**AKL** (1.6), *CGF* (12.8), DSD (4.7), *DTM* (18.3), EDL (2.3), KKN (3.9), LEA (2.6), PPQ (4.8), **SHL** (2.8), SSP (1.7), SVS (1.9), TSI (3.3), VSS (2.0)
	12	10.1 ± 3.0	20	AAS (2.2), EED (3.8), EKL (1.4), **EVD** (5.4), *FGG* (9.4), KAK (2.5), LKL (1.0), LPQ (3.0), LSL (0.9), LSS (1.0), PAS (2.3), QGL (2.6), RPY (7.6), SEI (2.6), SLS (1.0), SLT (2.0), SSV (1.4), TAL (1.9), TTV (3.8), VLL (1.7)

For each organism, the number of ORFs used for the analysis is indicated. N(Occ) indicates the number of occurrences for a particular sequence in the genome; N(Seq) indicates the number of sequences that appear N(Occ) number of times. The expected value of N(Seq) is derived from the genome jumbling method (with uncertainties shown at one standard deviation). Values in parentheses accompanying each sequence refer to the ratio of the number of times that sequence is observed to the number of times that sequence is expected based on positional amino acid frequencies. Sequences in boldface are known recognition motifs; italicized sequences belong to, entirely or in part, highly repeated sequences (e.g., homologous proteins or transposon ORFs), and underlined sequences take the form XKK (XSS in *A. thaliana*).

proteins. Similarly, all 10 occurrences of RSH and seven of 10 occurrences of RSR from *E. coli* can be explained by transposon ORFs (IS5 and IS1, respectively).

An additional cause of overrepresentation in C-terminal tripeptides may be a consequence of the synthesis of the polypeptide chains themselves. In both *E. coli* (Mottagui-Tabar et al. 1994; Björnsson et al. 1996; Mottagui-Tabar and Isaksson 1997) and *S. cerevisiae* (Mottagui-Tabar et al. 1998), it has been shown that certain amino acids at the last two positions of the polypeptide chain can affect the efficiency of translational termination. Using the analysis presented here, an overrepresentation of ORF sequences that end in XKK is found in all species except *A. thaliana*. For example, in the *C. elegans* genome, the tripeptide KKK is observed 70 times, GKK 32 times, SKK 28 times, AKK and LKK 24 times each, and EKK 22 times. In contrast, the expected number of occurrences of these sequences calculated using position-specific amino acid frequencies are: KKK, 15 times; GKK, 7 times; SKK, 15 times; AKK, 9 times; LKK, 15 times; and EKK, 10 times. These data are consistent with the reported results that in *E. coli*, lysine codons at the −1 (Björnsson et al. 1996) and −2 (Mottagui-Tabar et al. 1994; Mottagui-Tabar and Isaksson 1997) positions can enhance the efficiency of translational termination. Moreover, Arkov et al. (1995) noted that the 5′ contexts of stop codons are similar in both *E. coli* and humans, suggesting that effects on termination efficiency may be related. However, it should be noted that the observation of XKK in *S. cerevisiae* in this study is not consistent with previously observed translational termination effects in this species, which apparently do not depend on lysine (Mottagui-Tabar et al. 1998).

Since the genome jumbling analysis already takes into account the amino acid frequencies at each position, our data not only reflect the previous observation that lysine is found preferentially at the −1 position of many proteins (Berezovsky et al. 1997, 1999), but suggest that the correlated appearance of the terminal dipeptide KK may also be important. Interestingly, the XKK pattern is not found in the *A. thaliana* genome. Instead, an overabundance of XSS sequences is observed. We speculate that in this plant species, the serine residue or its corresponding codons may have a similar effect on translation efficiency as lysine in the other species studied.

Identification of Known and Potential Interaction Motifs

Not surprisingly, in addition to the reasons listed above, this study identified a set of tripeptide sequences that are known to serve as targeting signals or other recognition motifs in eukaryotes. The last three amino acids of the tetrapeptide signal for retention of proteins within the ER (Munro and Pelham 1987), DEL, were identified as being overrepresented in all eukaryotic genomes. In addition, the known variant EEL (Mazzarella et al. 1990) was found in humans. Similarly, the PTS1 (Gould et al. 1989), SKL, was also found in all eukaryotic genomes, as well as the variant AKL in humans and *C. elegans*. Moreover, an additional known C-terminal binding site was found. The sequence EVD occurs eight times in *S. cerevisiae* and 12 times in humans. This sequence corresponds to the last three amino acids of the tetrapeptide recognition motif, EEVD, found in the hsp70/hsp90 protein family, recognized by such proteins as Hop, FKBP52, and PP5 (Young et al. 1998; Buchner 1999; Scheufler et al. 2000).

Finally, previously unidentified recognition motifs may emerge from the results of this analysis. For example, the tripeptide FWC was found nine times in the *S. cerevisiae* genome. This particular sequence has several characteristics which make it potentially intriguing as a candidate recognition site: (1) All nine proteins that contain FWC belong to the family of amino acid permeases, and (2) the absolute sequence conservation extends only across the last three residues; upstream residues, although similar, are not identical (Fig. 2). Furthermore, previous work has suggested that amino acid permeases in yeast utilize a unique set of proteins for proper trafficking (Ljungdahl et al. 1992; Kuehn et al. 1996; Gilstring et al. 1999), and that the C-terminal portion of these permeases is important for turnover (Helliwell et al. 2001; Omura et al. 2001). Of the remaining genomes presented in Table 2, we did not identify FWC as an overrepresented tripeptide. These results, along with the data from this study, implicate the C-

```
Agp1p    DEELIKQEDE EYRERLRNGP YWKRVVAFWC 633
Bap2p    DPELMRQEDE ENKEKLRNMS LMRKAYHFWC 609
Bap3p    DPEIMRQEDE ENKERLKNSS IFVRVYKFWC 604
Gap1p    DLDLLKQEIA EEKAIMATKP RWYRIWNFWC 602
Gnp1p    DEELLKQEDE EYKERLRNGP YWKRVLDFWC 663
Hip1p    DLTLRREEMR IERETLAKRS FVTRFLHFWC 603
Sam3p    NLELFKAQKE AEEQLIASKP FYYKIYRFWC 587
Tat1p    LKEYENSESS ENPNSSRSRK FFKRMTNFWC 619
Tat2p    DIEIVKQEIA EKKMYLDSRP WYVRQFHFWC 592
```

Figure 2 Alignment of the amino acid sequences of the nine amino acid permeases from *S. cerevisiae* that end in the tripeptide FWC (highlighted) across the C-terminal 30 residues. Note that there are no other absolutely conserved residues within this region besides the terminal tripeptide.

terminal tripeptide FWC as potentially important for proper localization of the yeast amino acid permeases. Interestingly, in our preliminary analysis of the genome from the yeast *Candida albicans*, we noted that the FWC tripeptide does appear five times: once in the ORF for the amino acid permease Gap1p, and in four additional ORFs with high sequence similarity to Gap1p and other amino acid permeases, as identified through a BLAST search (Altschul et al. 1997). However, we were unable to identify any occurrences of the FWC sequence in the fission yeast *Schizosaccharomyces pombe*, a species more distantly related to *S. cerevisiae* than *C. albicans*, despite the fact that this genome does contain ORFs encoding amino acid permeases with significant similarity to those of *S. cerevisiae*.

Conclusions

The complete sequence of the genome of an organism can serve as a powerful tool for the analysis of sequence patterns and distributions. Since this information comprises a nearly complete array of a large number of data points, statistical methods can be applied to distinguish between chance occurrences and the results of selection. In the work presented here, C-terminal tripeptide sequences from a given genome were analyzed to determine which sequences can be considered overrepresented. Through this study, a collection of nonrandomly distributed tripeptide sequences was identified in eukaryotes. In addition to known targeting signals and binding motifs, several sequences could be explained by ORF homology or effects of certain codons on translational termination efficiency.

For many of these remaining sequences, however, distinguishing features are not readily apparent without more detailed analysis. This was, in part, due to the fact that many of these tripeptides belong to ORFs whose functions have not yet been determined. Hence, it is not yet possible to correlate those C-terminal sequences with a particular function or subcellular localization. Moreover, the difficulty of a comprehensive analysis of these data is enhanced by the large number of overrepresented sequences and the myriad of functions in which their corresponding proteins participate. By making these results available to the scientific community, we hope to enable the identification of additional common themes.

METHODS

Databases of the protein sequences of predicted ORFs from an entire genome were downloaded from the Web sites of the following institutions: The Institute for Genomic Research (*M. jannaschii*, *A. thaliana*; Web site: http://www.tigr.org [Bult et al. 1996; The *Arabidopsis* Genome Initiative 2000]), The *E. coli* Genome Project at the University of Wisconsin-Madison (*E. coli* K-12, Web site: http://www.genome.wisc.edu/k12.htm [Blattner et al. 1997]), The *Saccharomyces* Genome Database at Stanford University (*S. cerevisiae*, Web site: http://genome-www.stanford.edu/Saccharomyces/ [Goffeau et al. 1996]), The Sanger Centre (*C. elegans*, Web site: http://www.sanger.ac.uk/Projects/C_elegans/ [The *C. elegans* Sequencing Consortium 1998]), and the National Center for Biotechnology Information (*H. sapiens*, Web site: http://www.ncbi.nlm.nih.gov/genome/guide/human/ [Lander et al. 2001; Venter et al. 2001]). At the time of this study, the annotated array of predicted ORFs from the human genome was comprised of 14,760 entries.

Programs used for both the frequency method and the genome jumbling method were written in Perl.

ACKNOWLEDGMENTS

G.J.G. was supported in part by a grant through the Medical Scientist Training Program. We thank the NIH for support.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Arkov, A.L., Korolev, S.V., and Kisselev, L.L. 1995. 5′ contexts of *Escherichia coli* and human termination codons are similar. *Nucleic Acids Res.* **23:** 4712–4716.

Aubourg, S., Boudet, N., Kreis, M., and Lecharny, A. 2000. In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. *Plant Mol. Biol.* **42:** 603–613.

Berezovsky, I.N., Kilosanidze, G.T., Tumanyan, V.G., and Kisselev, L. 1997. COOH-terminal decamers in proteins are nonrandom. *FEBS Lett.* **404:** 140–142.

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.

Björnsson, A., Mottagui-Tabar, S., and Isaksson, L.A. 1996. Structure of the C-terminal end of the nascent peptide influences translation termination. *EMBO J.* **15:** 1696–1704.

Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1474.

Blobel, G. and Dobberstein, B. 1975. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J. Cell Biol.* **67:** 835–851.

Buchner, J. 1999. Hsp90 & Co.—a holding for folding. *Trends Biochem. Sci.* **24:** 136–141.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058–1073.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *Caenorhabditis elegans*. A platform for investigating biology. *Science* **282:** 2012–2018.

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.

Gilstring, C.F., Melin-Larsson, M., and Ljungdahl, P.O. 1999. Shr3p mediates specific COPII coatomer-cargo interactions required for the packaging of amino acid permeases into ER-derived transport vesicles. *Mol. Biol. Cell* **10:** 3549–3565.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B.,

Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. 1996. Life with 6000 genes. *Science* **274:** 546, 563–567.
Gould, S.J., Keller, G.A., Hosken, N., Wilkinson, J., and Subramani, S. 1989. A conserved tripeptide sorts proteins to peroxisomes. *J. Cell Biol.* **108:** 1657–1664.
Helliwell, S.B., Losko, S., and Kaiser, C.A. 2001. Components of a ubiquitin ligase complex specify polyubiquitination and intracellular trafficking of the general amino acid permease. *J. Cell Biol.* **153:** 649–662.
Kuehn, M.J., Schekman, R., and Ljungdahl, P.O. 1996. Amino acid permeases require COPII components and the ER resident membrane protein Shr3p for packaging into transport vesicles in vitro. *J. Cell Biol.* **135:** 585–595.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Ljungdahl, P.O., Gimeno, C.J., Styles, C.A., and Fink, G.R. 1992. SHR3: A novel component of the secretory pathway specifically required for localization of amino acid permeases in yeast. *Cell* **71:** 463–478.
Mazzarella, R.A., Srinivasan, M., Haugejorden, S.M., and Green, M. 1990. ERp72, an abundant luminal endoplasmic reticulum protein, contains three copies of the active site sequences of protein disulfide isomerase. *J. Biol. Chem.* **265:** 1094–1101.
Mottagui-Tabar, S. and Isaksson, L.A. 1997. Only the last amino acids in the nascent peptide influence translation termination in *Escherichia coli* genes. *FEBS Lett.* **414:** 165–170.
Mottagui-Tabar, S., Björnsson, A., and Isaksson, L.A. 1994. The second to last amino acid in the nascent peptide as a codon context determinant. *EMBO J.* **13:** 249–257.
Mottagui-Tabar, S., Tuite, M.F., and Isaksson, L.A. 1998. The influence of 5′ codon context on translation termination in *Saccharomyces cerevisiae*. *Eur. J. Biochem.* **257:** 249–254.
Munro, S. and Pelham, H.R. 1987. A C-terminal signal prevents secretion of luminal ER proteins. *Cell* **48:** 899–907.
Omura, F., Kodama, Y., and Ashikari, T. 2001. The basal turnover of yeast branched-chain amino acid permease Bap2p requires its C-terminal tail. *FEMS Microbiol. Lett.* **194:** 207–214.
Schatz, G. and Dobberstein, B. 1996. Common principles of protein translocation across membranes. *Science* **271:** 1519–1526.
Scheufler, C., Brinker, A., Bourenkov, G., Pegoraro, S., Moroder, L., Bartunik, H., Hartl, F.U., and Moarefi, I. 2000. Structure of TPR domain-peptide complexes: Critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. *Cell* **101:** 199–210.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Young, J.C., Obermann, W.M., and Hartl, F.U. 1998. Specific binding of tetratricopeptide repeat proteins to the C-terminal 12-kDa domain of hsp90. *J. Biol. Chem.* **273:** 18007–18010.

WEB SITE REFERENCES

http://www.tigr.org; The Institute for Genomic Research (source of *M. jannaschii* and *A. thaliana* ORF sequences).
http://www.genome.wisc.edu/k12.htm; The *E. coli* Genome Project at the University of Wisconsin-Madison (source of *E. coli* ORF sequences).
http://genome-www.stanford.edu/Saccharomyces/; The *Saccharomyces* Genome Database at Stanford University (source of *S. cerevisiae* ORF sequences).
http://www.sanger.ac.uk/Projects/C_elegans/; The Sanger Centre (source of *C. elegans* ORF sequences).
http://www.ncbi.nlm.nih.gov/genome/guide/human/; The National Center for Biotechnology Information (source of *H. sapiens* ORF sequences).

Received July 29, 2002; accepted in revised form January 28, 2003.