Quantile distributions of amino acid usage in protein classes

Samuel Karlin, B.Edwin Blaisdell and Philipp Bucher[1]

Department of Mathematics, Stanford University, Stanford, CA 94305, USA and [1]Bioinformatique, Institut Suisse de Recherches, Expérimentales sur le Cancer, Ch. des Boveresses 155, CH-1066 Epalinges s, Lausanne, Switzerland

A comparative study of the compositional properties of various protein sets from both cellular and viral organisms is presented. Invariants and contrasts of amino acid usages have been discerned for different protein function classes and for different species using robust statistical methods based on quantile distributions and stochastic ordering relationships. In addition, a quantitative criterion to assess amino acid compositional extremes relative to a reference protein set is proposed and applied. Invariants of amino acid usage relate mainly to the central range of quantile distributions, whereas contrasts occur mainly in the tails of the distributions, especially contrasts between eukaryote and prokaryote species. Influences from genomic constraint are evident, for example, in the arginine:lysine ratios and the usage frequencies of residues encoded by G + C-rich versus A + T-rich codon types. The structurally similar amino acids, glutamate versus aspartate and phenylalanine versus tyrosine, show stochastic dominance relationships for most species protein sets favoring glutamate and phenylalanine respectively. The quantile distribution of hydrophobic amino acid usages in prokaryote data dominates the corresponding quantile distribution in human data. In contrast, glutamate, cysteine, proline and serine usages in human proteins dominate the corresponding quantile distributions in *Escherichia coli*. *E.coli* dominates human in the use of basic residues, but no dominance ordering applies to acidic residues. The discussion centers on commonalities and anomalies of the amino acid compositional spectrum in relation to species, function, cellular localization, biochemical and steric attributes, complexity of the amino acid biosynthetic pathway, amino acid relative abundances and founder effects.

Key words: amino acid usages/quantile distributions/weak and strong amino acid codon types

Introduction

Detailed knowledge of amino acid (aa) usage within and among protein sets may assist in appraising a particular sequence. For example, if a certain protein is reported to be rich (or poor) in a given aa type, one would like to know how significant this circumstance is among a broad collection of proteins from a similar source. From this perspective, invariants and contrasts with respect to aa usage are identified and interpreted for protein sequence collections of several species, including human, *Drosophila*, yeast, *Escherichia coli* and *Bacillus subtilis*; for open reading frames (ORFs) in three large human virus genomes, human cytomegalovirus (CMV), Epstein–Barr virus (EBV) and vaccinia; for various human protein subclasses (e.g. nuclear, glycoprotein and enzyme); and for *E.coli* enzyme.

Our motivation for these analyses derives from an interest in the following biological and evolutionary issues: (i) How do aa usages compare and contrast across species, say *E.coli* versus human, *E.coli* versus *B.subtilis*?; (ii) What is the nature of aa usage per protein in relation to function, cellular localization, evolutionary history and other biological criteria?; (iii) How do aa usages of similar biochemical, charge or steric attributes relate? For example, how do the quantile distributions compare for Lys and Arg (both positively charged), for Asp and Glu (both negatively charged), for Gly and Ala (both of small size), for Ser and Thr (having similar post-translational modification potential), for the amide side chains residues Gln and Asn, among strongly hydrophobic amino acids (Leu, Ile, Val, Phe, Met) and for relationships related to evolutionary substitutability?; (iv) other perspectives on aa compositional preferences relate to the complexity of the biosynthetic pathways for the different aa, to aa relative abundances, to aa distributions along the sequences, to intra and extracellular pH, to codon biases, and to founder effects.

Residue usage across protein subsets has been the subject of a number of comparative studies. Sueoka (1960) noticed a general correlation between deoxynucleotide and aa composition for a variety of organisms. King and Jukes (1969) determined the aa composition of 53 vertebrate polypeptides (total 5492 residues) and claimed, excepting arginine, concordance of observed frequencies with expectations derived from random codon choices. Nakashima *et al.* (1986) investigated the influence of folding types on residue usage. Doolittle (1986, pp. 55–59) compared the aa composition of *E.coli* and human protein sequences and observed the reduced use of cysteine in *E.coli* (putatively all prokaryotes versus higher eukaryotes). McCaldon and Argos (1988) organized peptides ranging from 2 to 11 residues and projected certain preferences in protein sequences. Ikemura *et al.* (1990) and D'Onofrio *et al.* (1991) analyzed the aa composition of individual mammalian proteins under the isochore hypothesis (see also Aissani *et al.*, 1991). All these comparative studies have centered on average residue usages of different protein collections.

Our results are based on more robust quantile distributions and stochastic ordering concepts applied to different amino acid classifications. For a given residue type (e.g. individual aa, cationic, anionic, aggregate hydrophobics) and a given protein collection *C* (e.g. all protein sequences of a particular species or function class), a histogram of use for the residue type was generated. Concretely, for each protein sequence of *C*, the frequency of the residue type in the sequence was determined and the totality of all these frequencies was described by a histogram of the given residue usage with respect to *C*. The quantile distribution is the cumulative representation of this histogram. Thus, the quantile distribution $Q(x)$ of a given residue type for a given set of proteins indicates the fraction of proteins in which that residue type occurs with a frequency $\leq x\%$. The medians (the 0.50 quantile point) and 80% quantile range (corresponding to the 0.10–0.90 quantile levels) are major statistical measurements. The 0.01, 0.05, 0.95 and 0.99 quantile points of aa usage provide standards by which to assess extremes of aa usage for any particular protein or protein family. Quantile distributions for the different protein sets were determined for each individual aa (Tables I and II), for the aa groups of positively and negatively charged residues and for total and net charge values (Table III), for the aggregate of the major hydrophobic aa (Table IV) and for strong and weak aa codon types (Table V) (see Materials and methods).

Materials and methods

Data

Protein sets were compiled from SWISS-PROT release 17 (Bairoch and Boeckmann, 1991). Duplicate and highly similar sequences were culled to remove redundancies with the aid of the program PROSET (Brendel, 1992). The fruit fly and yeast sets respectively contain proteins from *D.melanogaster* and

Table I. Quantile distributions of amino acid usages in different species

Human

Quantile	Min.	.01	.05	10	.25	.50	75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	8	38	76	188	376	564	676	714	744	751		
E	0.3	2.0	3.4	4.1	5.0	6.3	7.8	9.4	11.2	16.8	20.2	6.7	2.5
D	0.1	1.3	2.4	3.0	4.0	5.0	5.8	6.7	7.4	9.7	15.3	5.0	1.6
K	0.0	0.8	2.4	3.0	4.2	5.6	7.1	8.8	9.9	14.1	28.8	5.8	2.6
R	0.5	1.5	2.5	3.1	4.1	5.0	6.2	7.5	8.6	12.0	19.1	5.3	2.0
H	0.0	0.3	0.9	1.1	1.6	2.3	2.9	3.7	4.1	5.6	12.6	2.4	1.1
L	0.8	3.3	5.5	6.4	7.8	9.4	11.0	12.7	14.0	17.1	22.8	9.5	2.7
I	0.0	0.8	1.5	2.1	3.2	4.5	5.6	6.9	7.5	9.5	11.0	4.5	1.8
V	0.6	2.2	3.7	4.2	5.2	6.4	7.6	8.7	9.3	10.5	12.6	6.4	1.7
M	0.0	0.4	0.8	1.0	1.5	2.1	2.7	3.3	3.8	4.7	6.5	2.2	0.9
F	0.0	0.5	1.5	2.1	2.9	3.8	4.8	5.7	6.2	7.5	9.5	3.8	1.5
Y	0.0	0.2	1.0	1.5	2.2	3.0	3.7	4.6	5.1	6.1	7.3	3.0	1.2
W	0.0	0.0	0.2	0.4	0.8	1.2	1.9	2.5	2.9	3.7	5.9	1.4	0.9
P	0.8	1.5	2.6	3.4	4.5	5.6	7.3	9.5	11.0	17.3	36.6	6.2	3.1
G	1.3	2.9	4.0	4.6	5.7	6.9	8.4	10.0	11.2	24.8	28.8	7.4	3.2
A	1.2	3.0	4.1	4.5	5.5	6.8	8.5	10.4	11.6	14.2	26.1	7.2	2.5
S	0.8	3.1	4.2	4.7	5.7	7.1	8.5	10.1	10.9	14.4	17.5	7.3	2.2
T	0.0	1.7	3.0	3.5	4.3	5.3	6.3	7.4	8.3	11.0	13.5	5.4	1.7
Q	0.5	1.4	2.3	2.7	3.4	4.2	5.1	6.3	7.3	10.9	25.6	4.5	1.9
N	0.0	0.6	1.6	2.1	2.9	3.8	4.7	5.7	6.4	7.6	9.2	3.9	1.4
C	0.0	0.0	0.4	0.7	1.2	1.9	2.9	4.9	6.1	7.4	11.2	2.3	1.7

E. coli

Quantile	Min.	.01	.05	10	.25	.50	.75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	8	36	71	178	355	533	639	675	703	710		
E	0.7	1.3	2.5	3.3	4.8	6.3	7.5	8.4	9.1	11.0	15.9	6.1	2.0
D	0.3	1.3	2.4	2.9	4.2	5.4	6.3	7.1	7.7	8.8	10.3	5.3	1.6
K	0.3	1.1	1.9	2.5	3.4	4.5	5.7	7.1	8.1	9.8	15.7	4.7	1.9
R	1.1	1.7	2.7	3.1	4.4	5.7	7.0	8.0	8.8	10.7	12.9	5.7	1.9
H	0.0	0.2	0.6	1.0	1.5	2.3	3.0	3.8	4.3	5.2	6.5	2.3	1.1
L	3.5	5.2	6.5	7.2	8.4	10.1	11.9	13.4	15.1	17.3	21.2	10.3	2.6
I	1.7	2.5	3.4	3.8	4.6	5.7	6.8	8.3	9.2	10.6	13.7	5.9	1.8
V	1.9	3.7	4.6	5.2	6.1	7.1	8.4	9.6	10.4	11.9	13.9	7.3	1.8
M	0.3	0.7	1.1	1.5	1.9	2.6	3.3	4.1	4.7	5.7	6.3	2.7	1.1
F	0.5	1.0	1.7	2.2	2.8	3.5	4.5	5.8	6.6	8.2	13.4	3.8	1.5
Y	0.0	0.5	1.1	1.4	1.9	2.7	3.5	4.4	5.0	7.2	8.7	2.8	1.3
W	0.0	0.0	0.0	0.3	0.6	1.2	1.9	2.6	3.1	4.3	6.7	1.3	1.0
P	1.1	1.6	2.3	2.8	3.5	4.3	5.2	6.1	6.9	8.3	16.4	4.4	1.4
G	0.8	2.9	4.2	5.0	6.3	7.7	9.0	10.2	11.0	12.3	14.2	7.6	2.0
A	2.0	4.0	5.6	6.6	8.0	9.4	11.0	12.6	13.5	15.0	30.9	9.6	2.5
S	1.5	2.4	3.2	3.8	4.5	5.5	6.5	7.6	8.4	10.0	12.4	5.6	1.6
T	1.3	2.5	3.3	3.8	4.4	5.2	6.0	7.0	7.8	9.1	13.1	5.3	1.4
Q	0.0	1.2	1.9	2.3	3.2	4.2	5.2	6.4	7.3	9.0	15.0	4.3	1.6
N	0.6	1.3	1.8	2.2	2.9	3.8	4.6	5.6	6.3	8.0	9.7	3.9	1.4
C	0.0	0.0	0.0	0.2	0.5	1.1	1.5	2.1	2.6	3.8	7.9	1.1	0.9

Drosophila melanogaster

Quantile	Min.	.01	.05	10	.25	.50	.75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	3	12	23	57	114	171	205	216	225	227		
E	0.8	1.0	2.6	3.4	4.6	5.9	7.2	8.2	10.6	14.8	27.6	6.1	2.7
D	1.0	1.4	2.3	3.0	3.9	5.1	6.1	7.3	7.9	9.5	10.2	5.1	1.7
K	2.0	2.3	2.8	3.3	4.1	5.2	6.8	9.1	10.3	13.5	26.6	5.8	2.7
R	0.4	1.9	2.6	3.1	4.1	5.0	6.2	7.7	9.0	11.8	20.8	5.4	2.2
H	0.0	0.0	1.0	1.3	1.9	2.6	3.7	4.5	5.9	7.2	7.9	2.8	1.4
L	2.2	2.6	4.8	5.5	6.6	8.0	9.5	11.0	11.9	14.7	15.3	8.1	2.3
I	1.2	1.3	2.2	2.5	3.5	4.9	6.0	7.3	7.9	9.0	14.2	4.9	1.8
V	1.0	2.0	3.0	3.4	4.5	5.7	6.9	8.1	8.8	10.2	10.9	5.7	1.8
M	0.0	0.2	0.9	1.2	1.6	2.3	3.0	3.9	4.7	6.0	6.7	2.4	1.1
F	0.0	0.8	1.2	1.6	2.4	3.4	4.4	5.3	6.1	7.1	9.1	3.5	1.5
Y	0.0	0.5	1.2	1.5	2.0	2.9	4.0	4.7	5.3	6.7	9.0	3.1	1.3
W	0.0	0.0	0.0	0.2	0.5	0.8	1.2	2.0	2.3	3.3	4.2	1.0	0.7
P	0.0	1.3	2.6	3.2	4.2	5.3	6.7	8.2	9.4	14.9	17.0	5.6	2.4
G	1.9	2.1	2.9	3.7	4.8	6.3	8.0	10.0	11.3	17.6	27.6	6.8	3.1
A	1.7	2.7	4.2	4.7	5.7	7.5	9.1	10.9	12.6	15.7	19.5	7.7	2.7
S	1.6	3.3	4.1	5.3	6.4	7.6	9.9	12.2	13.6	15.5	18.4	8.2	2.8
T	1.1	1.7	3.2	3.8	4.5	5.5	6.4	7.5	8.0	10.6	42.3	5.7	2.9
Q	1.0	1.5	2.4	2.7	3.4	4.5	6.2	9.1	12.1	15.7	18.9	5.3	2.9
N	1.0	1.6	2.6	3.1	3.9	4.9	5.9	7.1	8.4	11.0	15.1	5.0	1.8
C	0.0	0.0	0.2	0.4	0.9	1.6	2.4	3.4	4.8	8.5	9.0	1.9	1.5

Yeast (Saccharomyces cerevisiae)

Quantile	Min.	.01	.05	10	.25	.50	75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	5	22	44	108	216	324	388	410	427	431		
E	0.7	2.4	3.4	4.4	5.4	6.5	7.6	8.9	10.0	12.7	14.4	6.6	2.0
D	1.8	2.4	3.5	4.0	4.9	6.0	6.8	8.0	8.7	12.1	16.3	6.0	1.8
K	1.5	2.9	4.1	4.9	6.1	7.4	8.6	9.8	10.9	13.2	15.0	7.4	2.0
R	0.4	1.3	2.1	2.7	3.5	4.3	5.2	6.3	6.8	8.9	10.7	4.4	1.5
H	0.0	0.4	0.8	1.0	1.4	2.1	2.7	3.2	3.6	4.8	5.7	2.1	0.9
L	0.2	4.7	6.0	6.8	8.0	9.1	10.5	11.7	12.6	14.5	15.7	9.2	2.0
I	0.7	2.8	3.5	4.3	5.2	6.3	7.4	8.3	8.9	10.4	13.1	6.3	1.6
V	0.9	2.3	3.4	3.9	4.7	6.0	7.1	8.1	8.8	10.2	11.0	6.0	1.7
M	0.0	0.3	0.9	1.2	1.6	2.0	2.5	3.2	3.6	4.6	9.1	2.1	0.9
F	0.0	1.2	2.2	2.6	3.6	4.3	5.2	5.9	6.7	8.0	9.5	4.4	1.4
Y	0.5	0.9	1.4	1.7	2.5	3.3	4.1	4.8	5.3	5.9	10.0	3.3	1.2
W	0.0	0.0	0.0	0.2	0.5	0.9	1.4	2.1	2.4	3.0	3.8	1.0	0.7
P	0.0	1.3	2.3	2.8	3.6	4.3	5.1	6.0	6.7	9.0	11.1	4.4	1.4
G	1.2	2.0	2.7	3.3	4.2	5.5	7.1	8.4	9.2	10.8	19.0	5.7	2.1
A	0.5	2.7	3.2	3.8	4.8	6.2	8.0	9.5	10.4	12.6	20.9	6.5	2.3
S	3.8	4.0	4.8	5.6	6.5	7.7	9.2	11.3	12.6	16.0	26.5	8.1	2.6
T	0.5	2.6	3.7	4.2	4.9	5.6	6.5	7.3	8.0	10.6	25.2	5.8	1.7
Q	1.1	1.6	2.0	2.4	3.0	3.6	4.5	5.4	6.7	11.7	26.9	4.0	2.1
N	1.5	2.0	2.8	3.4	4.3	5.3	6.5	8.1	9.1	14.4	23.3	5.7	2.2
C	0.0	0.0	0.0	0.3	0.6	1.1	1.6	2.1	2.5	3.3	5.3	1.2	0.8

Bacillus subtilis

Quantile	Min.	.01	.05	10	.25	.50	75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	2	7	14	34	68	102	122	129	134	135		
E	0.9	1.2	2.5	3.6	6.4	8.1	9.6	10.4	11.1	11.7	14.3	7.7	2.5
D	0.7	0.9	1.9	3.5	4.7	5.6	6.5	7.3	7.9	8.8	9.0	5.5	1.6
K	1.0	2.6	3.8	4.8	5.8	7.2	8.5	10.4	11.5	12.5	13.0	7.3	2.1
R	1.1	1.3	1.9	2.2	3.1	4.3	5.3	6.3	7.1	8.5	9.2	4.2	1.6
H	0.3	0.3	0.5	0.9	1.6	2.0	2.8	3.3	4.1	4.5	4.8	2.2	0.9
L	3.8	3.8	5.8	6.8	7.8	8.8	10.3	12.1	13.9	14.9	17.4	9.1	2.3
I	2.9	3.3	4.2	4.9	5.8	7.2	8.8	10.3	11.7	13.5	16.1	7.4	2.2
V	3.4	3.6	4.6	4.8	5.8	7.0	8.1	9.2	10.0	11.5	12.2	7.0	1.7
M	0.8	0.8	1.3	1.6	2.0	2.7	3.2	3.7	4.3	4.5	4.6	2.6	0.8
F	1.6	1.6	1.9	2.2	2.6	3.7	4.8	6.4	8.5	10.2	10.3	4.1	1.9
Y	0.3	1.0	1.5	1.8	2.5	3.0	3.8	4.5	5.4	7.0	7.0	3.2	1.2
W	0.0	0.0	0.0	0.0	0.3	0.6	1.1	2.1	2.6	3.6	5.2	0.9	0.8
P	0.7	1.1	1.7	2.2	2.9	3.7	4.4	4.9	5.2	5.9	7.5	3.6	1.1
G	2.5	2.5	3.8	4.3	5.2	7.1	8.8	10.1	10.6	12.2	13.0	7.1	2.2
A	1.8	2.5	3.8	4.3	5.2	7.1	9.5	10.9	12.8	13.5	14.1	7.4	2.6
S	2.3	3.3	3.9	4.3	4.9	6.1	7.3	8.2	9.0	11.3	13.4	6.3	1.7
T	2.3	2.8	3.5	3.7	4.3	5.3	6.2	7.1	7.7	11.7	12.1	5.4	1.5
Q	0.8	0.9	1.7	2.1	2.7	3.7	4.5	6.0	6.8	7.9	8.0	3.8	1.5
N	0.4	1.1	1.9	2.4	3.1	4.1	5.5	7.1	7.9	8.9	10.7	4.4	1.8
C	0.0	0.0	0.0	0.0	0.2	0.7	1.2	1.6	2.3	3.3	4.4	0.8	0.7

Human Cytomegalovirus

Quantile	Min.	.01	05	.10	.25	.50	.75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	2	6	12	29	58	87	104	110	114	115		
E	1.3	1.3	1.9	2.3	3.6	5.0	6.2	7.4	8.0	9.6	12.8	5.0	1.9
D	1.3	1.5	1.8	2.3	3.2	4.6	5.6	6.1	7.1	8.0	9.1	4.5	1.6
K	0.0	0.3	0.4	1.3	1.9	2.6	3.4	4.4	5.9	7.2	7.3	2.7	1.4
R	2.6	3.1	3.8	4.5	5.8	7.8	9.8	11.4	14.4	15.8	16.9	8.0	2.9
H	0.6	0.9	1.5	1.7	2.3	3.1	3.8	4.9	5.9	7.0	7.4	3.2	1.3
L	3.2	3.5	5.8	6.9	8.5	10.2	12.3	14.1	16.3	17.1	24.6	10.5	3.2
I	0.0	0.8	1.2	1.4	2.2	3.4	4.8	6.5	6.8	7.9	7.9	3.6	1.8
V	3.5	3.5	4.9	5.3	6.6	7.8	8.7	10.1	10.7	12.0	12.8	7.7	1.8
M	0.0	0.3	0.4	0.8	1.2	1.9	2.7	3.3	4.2	4.8	5.3	2.0	1.0
F	0.9	1.2	1.8	2.3	2.9	3.7	4.7	5.4	6.5	8.2	10.2	3.9	1.5
Y	0.0	0.3	0.5	1.3	2.4	3.2	4.0	5.2	6.0	7.5	9.2	3.3	1.6
W	0.0	0.0	0.2	0.3	0.7	1.4	2.3	2.9	3.6	4.2	5.7	1.6	1.1
P	1.9	1.9	2.8	3.2	4.2	5.6	7.4	9.3	11.8	15.2	16.2	6.1	2.8
G	2.2	2.2	2.8	3.7	4.4	5.8	7.1	9.4	10.2	16.7	19.3	6.2	2.6
A	2.1	3.8	4.2	4.8	6.2	7.5	9.4	10.8	11.8	14.1	14.6	7.8	2.3
S	2.9	3.4	4.0	4.8	6.2	7.4	8.8	10.6	11.7	13.2	13.8	7.6	2.2
T	2.8	3.3	3.6	4.0	5.0	6.5	8.4	9.7	11.4	18.3	19.6	6.9	2.8
Q	1.3	1.3	1.5	1.9	2.5	3.3	4.2	5.1	5.8	7.0	8.3	3.4	1.3
N	0.4	0.6	0.9	1.4	2.1	2.9	4.0	5.6	6.6	9.8	9.9	3.3	1.8
C	0.5	0.7	1.0	1.3	1.7	2.4	3.3	4.2	4.8	7.5	9.4	2.7	1.4

Vaccinia

Quantile	Min.	.01	05	.10	.25	.50	75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	2	6	11	28	55	83	99	105	109	110		
E	2.0	2.5	3.4	3.8	4.3	5.4	6.3	7.6	9.1	9.4	9.4	5.5	1.6
D	3.8	3.8	4.2	4.5	5.6	6.6	7.4	8.5	9.3	11.1	13.0	6.6	1.6
K	2.5	3.8	4.9	5.2	6.3	7.5	8.5	9.4	10.2	10.8	12.8	7.4	1.6
R	0.6	0.8	2.0	2.3	3.2	4.0	4.8	5.7	6.0	7.5	7.5	4.1	1.3
H	0.0	0.4	0.7	1.0	1.5	2.0	2.7	3.2	3.6	5.4	5.4	2.1	1.0
L	3.7	3.7	5.7	5.9	7.5	8.6	9.9	11.3	11.7	13.2	15.5	8.7	2.0
I	4.6	4.7	6.4	6.7	8.0	9.4	10.4	11.4	12.5	13.6	15.6	9.2	1.9
V	2.9	3.1	3.8	4.4	5.4	6.3	7.5	8.2	8.9	9.3	9.9	6.3	1.5
M	0.8	0.8	1.3	1.5	2.0	2.7	3.4	4.1	4.8	5.4	6.6	2.8	1.1
F	1.5	1.6	1.8	2.5	3.5	4.7	5.8	6.7	7.3	8.5	9.1	4.7	1.6
Y	0.5	1.6	2.6	3.2	4.2	5.1	6.1	7.0	7.9	8.5	8.7	5.2	1.6
W	0.0	0.0	0.0	0.0	0.3	0.6	1.0	1.5	1.7	2.3	2.4	0.7	0.5
P	0.8	0.8	1.1	1.9	2.4	3.4	4.1	4.9	5.9	6.6	10.7	3.4	1.4
G	1.3	1.4	2.1	2.2	3.0	3.9	4.8	5.5	6.4	7.8	10.6	4.0	1.4
A	0.9	0.9	1.6	2.0	2.8	3.4	4.4	5.7	6.6	8.4	11.2	3.7	1.6
S	3.4	4.9	5.4	5.8	6.7	7.8	9.1	10.1	12.1	14.8	14.8	8.0	1.9
T	2.4	3.5	4.1	4.4	5.0	6.0	7.1	8.2	9.9	12.5	16.2	6.3	1.9
Q	0.3	0.4	0.8	1.0	1.6	2.1	2.7	3.3	3.9	5.6	9.6	2.2	1.2
N	2.9	2.9	4.3	5.1	5.9	6.8	7.8	8.8	9.3	10.1	11.8	6.9	1.6
C	0.3	0.3	0.5	0.6	1.3	1.9	2.7	3.6	4.1	6.3	6.8	2.1	1.2

Saccharomyces cerevisiae only. The viral protein sequences correspond to known ORFs. Several nonexclusive protein subsets were defined relying on the SWISS-PROT keyword index encompassing the human nuclear subset, a human glycoprotein subset and human and *E.coli* enzyme subsets. In every data set sequences shorter than 200 residues were excluded to reduce statistical fluctuations. This length limitation excluded ~25% of all proteins. Only ~10% of enzymes fall below the 200 aa criterion. Small proteins might differ in their aa composition from the bulk sequences. However, proteins of small size would, with >20 aa types, produce a preponderance of outlier observations, thus distorting the statistical analysis.

Quantile distributions and stochastic orderings

For each residue type and a specified protein class C of an organism, the quantity $y = Q(x)$ is the fraction of proteins of C which carry the specified residue type at a frequency at most x. The quantile distributions are displayed for the quantile levels y = min., 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99, max., as representatitive of the whole distribution relative to the appropriate class of proteins and residue types. The distributions are not Gaussian and for most aa the standard deviation is larger than what would be expected from the mean aa frequency and a protein size distribution based on a binomial model. A quantile distribution $\tilde{Q}(\cdot)$ is said to be stochastically larger than the quantile distribution $Q(\cdot)$ if $\tilde{Q}(x) < Q(x)$ for all x. This relation implies that at each y the usage x corresponding to the quantile distribution $\tilde{Q}(\cdot)$ exceeds the usage corresponding to the quantile distribution $Q(\cdot)$ and, more generally, each monotone transformation on levels of usage is similarly ranked; for statistical elaborations see Pečarić *et al.* (1992). Stochastic dominance is

Table II. Quantile distributions of amino acid usages in different subclasses of human and *E.coli* proteins

Nuclear proteins (human)

Quantile	Min.	.01	.05	10	.25	.50	75	90	.95	.99	Max.	Mean	Stdev
Rank	1	1	4	8	19	38	57	69	73	76	76		
E	2.5	2.5	3.4	4.5	5.2	6.8	8.3	9.2	10.7	16.8	16.8	7.1	2.4
D	0.5	0.5	1.7	2.3	3.4	5.0	6.0	7.8	10.0	12.0	12.0	5.0	2.3
K	2.7	2.7	3.3	3.6	4.4	5.3	6.8	8.6	12.4	28.8	28.8	6.3	3.7
R	1.4	1.4	3.0	3.4	4.4	5.3	6.9	8.2	10.1	19.1	19.1	5.9	2.7
H	0.0	0.0	0.4	0.8	1.3	1.9	2.6	3.5	4.6	7.1	7.1	2.1	1.3
L	1.3	1.3	4.0	4.9	6.4	8.4	10.4	11.3	12.4	13.3	13.3	8.3	2.7
I	0.7	0.7	0.9	1.2	2.2	3.2	4.8	5.6	6.5	7.3	7.3	3.5	1.6
V	1.4	1.4	2.3	3.4	4.5	5.2	6.2	7.1	7.7	8.6	8.6	5.2	1.5
M	0.0	0.0	0.9	1.0	1.5	2.3	2.9	3.7	4.4	6.0	6.0	2.3	1.0
F	0.0	0.0	0.8	1.7	2.2	2.9	3.7	4.2	5.3	6.2	6.2	2.9	1.2
Y	0.0	0.0	0.4	1.1	1.8	2.5	3.3	4.8	5.0	5.6	5.6	2.6	1.3
W	0.0	0.0	0.0	0.0	0.3	0.6	0.9	1.5	2.3	3.1	3.1	0.7	0.6
P	2.2	2.2	2.7	4.1	5.3	8.0	10.1	13.0	14.8	20.0	20.0	8.2	3.5
G	2.9	2.9	3.6	4.2	5.2	6.7	8.3	12.0	14.8	24.8	24.8	7.5	3.4
A	3.0	3.0	4.1	4.8	6.3	7.6	9.6	11.1	13.1	26.1	26.1	8.1	3.2
S	2.1	2.1	4.3	5.1	7.3	8.9	10.5	13.3	14.4	15.1	15.1	9.0	2.9
T	2.1	2.1	2.9	3.2	4.0	5.2	5.9	7.8	8.7	11.6	11.6	5.2	1.8
Q	0.5	0.5	2.2	2.9	3.7	4.6	5.7	7.3	9.6	17.7	17.7	5.0	2.4
N	0.6	0.6	1.5	1.8	2.3	3.3	4.6	5.5	5.9	6.6	6.6	3.5	1.4
C	0.0	0.0	0.1	0.3	0.9	1.5	2.4	3.3	4.0	4.3	4.3	1.7	1.1

Glycoproteins (human)

Quantile	Min.	.01	.05	10	.25	.50	75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	4	16	32	80	159	239	[illegible]	303	315	318		
E	1.5	1.9	3.3	3.7	4.7	5.7	6.9	8.2	9.1	12.1	19.0	5.9	1.9
D	0.8	1.7	2.7	3.0	3.9	4.8	5.5	6.3	6.7	8.2	15.3	4.8	1.4
K	0.0	0.5	1.7	2.5	3.7	4.8	6.4	7.4	8.3	9.7	11.1	5.0	1.9
R	1.3	1.7	2.4	2.9	3.9	4.9	5.9	7.4	8.4	11.4	13.3	5.1	1.9
H	0.3	0.4	0.9	1.1	1.7	2.4	2.9	3.6	4.2	5.1	12.6	2.4	1.1
L	3.3	3.6	5.5	6.5	7.8	9.5	11.2	13.3	14.9	17.4	22.8	9.7	2.8
I	0.0	0.8	1.5	2.0	3.2	4.2	5.4	6.6	7.4	8.8	11.0	4.3	1.8
V	0.8	2.9	3.9	4.5	5.5	6.6	7.8	8.8	9.5	10.5	12.2	6.7	1.7
M	0.0	0.5	0.7	1.0	1.4	1.8	2.5	3.2	3.7	4.5	6.5	2.0	0.9
F	0.4	0.5	1.7	2.4	3.0	4.0	5.1	5.9	6.3	7.8	9.5	4.1	1.4
Y	0.0	0.6	1.0	1.6	2.3	3.1	3.9	4.7	5.4	6.2	7.3	3.1	1.3
W	0.0	0.2	0.5	0.7	1.0	1.5	2.1	2.7	3.1	4.0	5.9	1.6	0.9
P	0.8	2.1	3.3	4.0	4.7	5.9	7.4	9.8	11.2	18.9	33.2	6.6	3.2
G	1.3	2.7	4.1	4.7	5.7	7.1	8.2	9.7	11.2	27.6	28.6	7.6	3.7
A	1.6	3.0	3.7	4.3	5.1	6.2	7.5	9.3	10.4	13.4	17.0	6.5	2.1
S	2.3	3.7	4.3	5.4	6.4	7.5	8.6	9.8	10.4	11.9	14.5	7.5	1.8
T	0.4	2.5	3.2	3.8	4.8	5.7	6.6	7.9	9.6	11.3	13.5	5.8	1.8
Q	0.5	1.6	2.4	2.8	3.4	4.2	5.1	6.0	6.6	8.4	13.8	4.4	1.5
N	0.3	1.1	1.8	2.3	3.2	4.1	5.0	6.3	6.9	8.2	9.2	4.2	1.5
C	0.0	0.3	0.6	0.9	1.5	2.4	3.7	6.1	6.9	7.9	8.5	2.9	1.9

Enzymes (human)

Quantile	Min.	.01	.05	10	.25	.50	.75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	3	14	27	67	133	199	239	252	263	265	-	-
E	1.9	2.3	3.9	4.2	5.2	6.3	7.6	8.7	9.4	11.1	11.7	6.4	1.7
D	1.9	2.6	3.4	3.9	4.5	5.2	5.7	6.5	7.1	8.5	9.7	5.2	1.1
K	0.0	1.4	2.4	3.0	4.2	5.7	6.8	8.1	9.2	11.8	17.5	5.7	2.1
R	1.5	2.3	2.6	3.4	4.1	5.1	6.3	7.1	7.6	8.9	13.3	5.2	1.5
H	0.3	0.5	1.2	1.5	1.9	2.5	3.0	3.7	4.0	4.4	4.6	2.5	0.9
L	3.3	5.4	6.1	7.1	8.3	9.5	10.9	12.2	12.8	14.9	16.8	9.6	2.0
I	1.2	1.3	2.7	3.2	4.0	5.0	5.9	7.1	7.5	8.7	11.0	5.0	1.5
V	2.8	3.6	4.5	4.7	5.6	6.6	7.8	8.8	9.4	10.5	11.4	6.7	1.5
M	0.4	0.8	1.0	1.2	1.8	2.3	2.8	3.3	3.7	4.6	5.0	2.3	0.8
F	0.0	0.9	2.4	2.8	3.3	4.2	5.1	6.0	6.6	7.5	9.1	4.2	1.3
Y	0.7	0.9	1.4	1.9	2.6	3.3	4.1	4.9	5.7	6.2	7.2	3.4	1.2
W	0.0	0.0	0.4	0.6	1.0	1.5	2.1	2.6	3.1	3.8	5.9	1.6	0.8
P	2.3	2.6	3.3	3.7	4.6	5.3	6.2	7.4	8.1	9.9	15.0	5.5	1.6
G	2.9	4.1	4.8	5.4	6.2	7.2	8.5	9.7	10.1	11.4	19.0	7.4	1.8
A	2.1	2.7	4.4	4.8	5.6	6.8	8.5	10.6	11.4	13.3	15.0	7.2	2.2
S	2.8	3.3	4.3	4.7	5.5	6.4	7.4	8.5	9.1	10.6	11.5	6.5	1.5
T	2.3	2.6	3.3	3.7	4.4	5.1	5.8	6.6	7.2	8.4	11.3	5.2	1.2
Q	1.4	1.5	2.3	2.7	3.3	4.0	4.7	5.4	6.1	7.6	8.5	4.1	1.2
N	0.6	1.5	2.0	2.4	3.2	3.9	4.6	5.7	6.3	7.6	8.6	4.0	1.3
C	0.0	0.2	0.8	1.0	1.4	1.8	2.7	3.8	4.7	6.5	7.2	2.2	1.2

Enzymes (E.coli)

Quantile	Min.	.01	.05	10	.25	.50	.75	.90	.95	.99	Max.	Mean	Stdev
Rank	1	4	18	36	90	180	270	324	342	357	360	-	-
E	1.3	2.3	3.8	4.4	5.3	6.5	7.6	8.4	8.6	9.7	10.9	6.4	1.6
D	1.2	2.4	3.1	3.8	4.9	5.7	6.5	7.2	7.6	8.5	9.5	5.6	1.3
K	1.1	1.5	2.4	2.8	3.6	4.6	5.7	6.8	7.5	8.4	10.6	4.7	1.5
R	1.5	1.9	2.8	3.4	4.5	5.5	6.6	7.5	8.0	9.2	10.6	5.5	1.6
H	0.3	0.6	1.0	1.3	1.8	2.4	3.1	3.8	4.2	5.1	5.7	2.5	1.0
L	3.5	4.8	6.4	7.2	8.2	9.6	11.0	12.6	13.4	15.2	17.3	9.7	2.1
I	1.7	2.5	3.5	3.8	4.7	5.7	6.7	7.8	8.5	10.3	12.0	5.8	1.6
V	3.5	4.0	4.8	5.3	6.1	7.0	8.3	9.4	10.1	11.1	13.9	7.2	1.6
M	0.4	0.8	1.3	1.6	2.0	2.6	3.3	3.9	4.3	5.3	5.8	2.7	0.9
F	1.0	1.2	1.9	2.3	2.8	3.5	4.4	5.5	6.2	8.0	8.7	3.7	1.3
Y	0.3	0.7	1.3	1.6	2.1	2.8	3.6	4.3	4.7	7.2	8.2	2.9	1.2
W	0.0	0.0	0.2	0.3	0.6	1.1	1.8	2.6	3.2	4.2	6.7	1.3	1.0
P	1.1	1.9	2.8	3.1	3.7	4.4	5.2	6.0	6.6	7.4	8.8	4.5	1.2
G	1.9	3.1	4.7	5.5	6.6	7.9	9.0	10.1	10.9	11.7	14.2	7.8	1.8
A	2.0	3.9	5.9	7.0	8.3	9.7	11.0	12.6	14.0	15.4	17.8	9.7	2.3
S	2.1	2.3	3.1	3.7	4.4	5.2	6.0	6.9	7.4	8.4	11.9	5.2	1.3
T	1.3	2.7	3.5	4.0	4.5	5.2	6.0	6.8	7.3	9.1	10.1	5.3	1.2
Q	0.0	1.5	1.9	2.4	3.2	4.0	4.9	6.0	6.5	7.4	10.3	4.1	1.4
N	0.9	1.4	2.0	2.5	3.1	3.9	4.5	5.3	5.9	7.7	8.2	3.9	1.2
C	0.0	0.0	0.3	0.4	0.8	1.2	1.6	2.0	2.3	3.8	4.6	1.3	0.7

Table III. Quantile distributions of charge types in different species and viral protein sets

Positively charged amino acids (K+R)

	Min.	.01	.05	10	.25	.50	.75	.90	.95	.99	Max.	Mean	Stdev
Human	3.9	5.4	7.3	8.1	9.3	10.8	12.5	14.3	15.7	23.0	30.2	11.1	2.9
Dros.	4.3	4.8	6.9	7.6	9.2	10.6	12.9	14.5	15.8	24.2	27.0	11.1	3.3
Yeast	3.1	5.1	7.6	8.7	10.1	11.7	13.5	14.9	15.9	19.3	23.4	11.8	2.7
E. coli	3.0	5.0	6.1	7.4	9.0	10.4	11.8	13.1	13.8	16.2	20.5	10.4	2.3
B sub.	4.9	5.0	6.9	7.9	9.8	11.3	13.5	15.0	16.0	17.4	18.7	11.5	2.7
CMV	4.2	5.4	5.8	7.2	9.2	10.8	12.2	14.4	16.5	18.4	19.4	10.7	2.8
EBV	2.6	2.6	5.1	6.7	8.3	9.6	11.1	12.2	13.6	14.9	14.9	9.7	2.3
Vacc	5.4	6.3	8.0	8.7	10.2	11.5	12.9	13.9	15.6	16.3	16.7	11.5	2.2
Hum. nuc.	4.7	4.7	8.1	8.7	9.3	11.6	13.5	15.7	20.3	30.2	30.2	12.1	3.9
Hum. gp	4.8	5.4	6.8	7.5	8.7	9.9	11.0	12.7	14.1	16.0	19.4	10.0	2.1
Hum. enz.	3.9	5.0	8.3	8.8	9.6	10.9	11.9	13.1	14.3	16.2	23.0	10.9	2.0
Ec. enz.	3.0	5.2	7.4	8.1	9.2	10.2	11.4	12.5	13.3	14.3	16.2	10.3	1.8

Negatively charged amino acids (D+E)

	Min.	.01	.05	10	.25	.50	.75	.90	.95	.99	Max.	Mean	Stdev
Human	0.4	4.9	6.8	8.0	9.6	11.4	13.2	15.2	17.7	22.9	26.2	11.6	3.3
Dros.	2.0	4.7	5.8	7.1	9.2	11.0	12.8	15.2	17.7	20.2	34.9	11.2	3.7
Yeast	3.5	5.4	7.7	9.1	10.8	12.4	14.2	16.1	17.5	21.6	27.5	12.6	3.0
E. coli	2.0	3.5	5.1	6.7	10.0	11.9	13.3	14.6	15.4	17.3	20.3	11.4	3.0
B. sub	2.8	3.1	4.1	8.5	12.1	14.0	15.4	16.8	17.8	18.5	21.3	13.2	3.6
CMV	2.7	3.1	4.1	5.3	7.1	9.8	11.4	13.1	15.0	18.5	18.8	9.5	3.1
EBV	3.1	3.1	4.4	5.5	7.6	9.6	11.1	12.8	13.3	17.3	17.3	9.4	2.7
Vacc	6.5	8.0	8.9	9.6	10.9	11.8	13.4	14.6	16.6	18.9	18.9	12.1	2.1
Hum. nuc.	3.6	3.6	5.0	7.5	9.1	11.9	14.2	17.5	19.7	26.2	26.2	12.1	4.3
Hum. gp.	3.9	4.9	6.5	7.4	9.2	10.6	12.1	13.5	14.6	18.8	23.9	10.7	2.6
Hum. enz.	4.9	6.4	8.3	9.0	10.2	11.5	13.0	14.4	14.8	17.5	19.3	11.6	2.2
Ec. enz.	3.5	5.2	7.7	9.4	11.1	12.2	13.5	14.5	15.0	16.4	17.3	12.1	2.2

Total charge (K+R+E+D)

	Min.	.01	.05	10	.25	.50	.75	.90	.95	99	Max.	Mean	Stdev
Human	6.3	11.2	15.3	17.1	19.4	22.1	25.3	29.2	32.1	40.3	50.0	22.7	5.3
Dros.	7.0	11.0	13.1	15.3	18.6	21.9	25.7	29.3	32.2	44.2	56.6	22.4	6.1
Yeast	7.3	12.6	16.5	18.6	21.8	24.3	26.9	29.9	32.4	36.7	42.9	24.4	4.7
E. coli	7.1	9.0	11.5	14.7	19.6	22.4	24.9	26.7	28.2	31.1	33.3	21.8	4.7
B. sub	9.4	9.4	11.2	15.7	22.9	25.4	28.5	30.6	32.8	34.8	35.8	24.7	5.7
CMV	9.6	9.9	11.7	13.5	17.1	20.8	23.4	26.2	27.3	30.1	30.9	20.2	4.7
EBV	9.0	9.0	11.5	13.0	17.3	19.5	21.3	23.3	24.5	28.3	28.3	19.1	3.8
Vacc	13.0	13.8	18.6	19.4	21.5	23.6	25.8	27.9	29.0	30.5	33.5	23.6	3.4
Hum. nuc.	9.7	9.7	14.0	17.3	19.9	22.2	27.6	33.7	40.0	50.0	50.0	24.2	6.9
Hum. gp.	9.3	12.1	15.0	16.3	18.2	20.6	22.9	25.2	26.9	32.2	36.3	20.7	3.8
Hum. enz.	12.1	13.9	16.7	18.4	20.4	22.3	24.7	27.2	28.7	31.4	41.6	22.5	3.6
Ec. enz.	8.3	11.0	16.1	18.1	20.4	22.5	24.8	26.3	27.2	29.2	30.0	22.3	3.6

Net charge (K+R-D-E)

	Min.	.01	.05	10	.25	.50	.75	90	.95	.99	Max.	Mean	Stdev
Human	-14.6	-8.8	-5.1	-3.7	-2.2	-0.7	1.0	2.9	4.9	7.9	26.6	-0.5	3.3
Dros.	-13.2	-8.4	-5.1	-3.6	-1.9	-0.3	1.3	3.2	4.8	12.3	22.3	-0.1	3.5
Yeast	-15.3	-10.6	-6.0	-4.3	-2.3	-0.7	0.9	2.6	3.8	7.0	10.3	-0.8	3.1
E. coli	-10.1	-7.1	-4.2	-3.7	-2.6	-1.5	0.4	2.1	3.1	7.6	12.1	-1.0	2.5
B. sub.	-7.5	-6.7	-5.8	-4.9	-3.7	-2.4	0.2	2.8	3.6	4.7	5.7	-1.7	2.9
CMV	-7.1	-6.7	-3.1	-2.0	-1.2	0.6	3.0	5.2	10.4	13.3	13.8	1.2	3.7
EBV	-10.6	-10.6	-5.0	-3.7	-1.7	0.0	2.2	3.6	6.3	9.7	9.7	0.3	3.3
Vacc.	-9.0	-9.0	-5.0	-3.8	-1.9	-0.3	1.2	2.2	3.0	4.5	4.5	-0.6	2.6
Hum. nuc.	-8.6	-8.6	-6.2	-3.8	-2.3	-0.4	1.7	3.5	6.0	26.6	26.6	0.0	4.3
Hum. gp.	-14.6	-8.9	-5.3	-3.8	-2.2	-0.8	-1.0	2.9	4.7	7.2	8.0	-0.7	2.9
Hum. enz.	-8.0	-7.8	-4.2	-3.0	-2.0	-0.6	0.8	1.5	2.7	4.4	7.8	-0.7	2.1
Ec. enz.	-6.4	-5.5	-4.2	-3.9	-2.9	-2.1	-1.0	-0.6	1.4	4.4	6.9	-1.8	1.8

Table IV. Quantile distributions of major hydrophobic residue usage in different protein sets

Major hydrophobic residues (F+I+L+M+V)												
	Min.	.01	.05	10	.25	.50	.75	.90	.95	.99	Max.	Mean	Stdev
Human	4.2	12.6	18.3	20.7	23.7	26.8	29.5	31.6	33.4	38.4	41.6	26.4	4.8
Dros.	6.8	14.0	16.6	17.8	21.0	25.0	28.2	31.2	32.7	34.2	35.9	24.6	5.0
Yeast	3.7	16.9	20.6	23.0	26.1	28.3	29.9	32.4	34.0	39.1	43.7	27.9	4.1
E. coli	12.4	21.8	24.2	25.5	27.5	29.2	31.1	36.3	40.5	43.8	45.8	29.9	4.4
B. sub.	19.7	20.8	24.0	25.1	27.8	29.6	31.6	34.6	42.5	46.4	47.9	30.3	5.0
HCMV	10.2	15.5	17.4	20.1	24.8	27.8	30.1	36.2	39.8	41.7	43.5	27.8	5.9
EBV	4.2	4.2	14.9	18.1	24.0	27.8	33.1	30.2	36.9	42.6	42.6	26.9	6.6
Vacc	20.3	20.9	23.5	26.7	30.5	32.3	33.7	35.4	36.5	39.9	41.2	31.8	3.7
Hum. nuc.	10.4	10.4	13.5	14.4	20.6	22.7	25.3	27.8	29.5	31.6	31.6	22.2	4.6
Hum. gp.	6.1	12.9	18.9	21.1	23.8	26.6	29.6	33.0	34.6	38.5	41.6	26.7	5.0
Hum. enz.	21.5	23.4	25.2	26.0	27.4	28.9	30.2	32.1	33.8	40.7	45.8	29.2	3.0
Ec. enz.	12.9	19.8	22.7	23.9	26.2	28.2	29.7	31.4	32.7	34.6	36.1	27.9	3.0

Table V. Quantile distributions of strong and weak codon type amino acids in different protein sets

Strong codon type amino acids (A+G+P)												
	Min.	.01	.05	.10	.25	.50	75	.90	.95	.99	Max.	Mean	Stdev
Human	10.5	12.4	14.2	15.3	16.9	19.6	23.3	27.1	30.6	50.3	63.4	20.8	6.2
Dros.	8.3	11.3	12.4	13.4	16.4	20.0	23.0	26.3	28.8	34.6	46.5	20.1	5.3
Yeast	7.0	8.6	10.8	11.6	13.4	16.5	19.3	21.9	23.2	24.3	31.5	16.6	3.9
E. coli	7.3	12.1	15.5	17.2	19.4	21.7	24.0	25.8	27.3	29.2	38.2	21.6	3.6
B sub.	7.0	9.3	11.3	12.2	14.6	18.6	20.9	24.0	25.8	28.6	28.6	18.1	4.4
HCMV	8.9	11.0	12.7	14.1	16.4	19.1	22.5	26.3	33.7	34.4	46.9	20.1	5.7
EBV	15.8	15.8	16.3	16.9	19.0	21.6	26.6	31.6	37.3	61.6	61.6	24.1	8.1
Vacc.	4.5	5.1	6.8	7.4	8.8	10.9	13.3	14.5	15.6	18.6	20.3	11.1	3.0
Hum nuc.	14.3	14.3	15.0	16.1	18.0	21.9	27.4	32.7	38.3	45.0	45.0	23.7	6.8
Hum. gp.	10.5	12.1	14.1	15.3	17.0	19.5	22.3	26.2	30.6	52.6	55.3	20.7	6.6
Hum. enz.	12.3	13.3	15.2	16.1	17.4	19.7	22.5	25.4	26.5	30.2	32.9	20.2	3.6
Ec. enz.	7.3	12.8	17.3	18.4	19.9	22.3	24.1	25.7	27.0	29.2	34.8	22.1	3.2
Weak codon type amino acids (F+I+K+N+Y)													
	Min.	.01	.05	10	.25	.50	75	.90	95	.99	Max	Mean	Stdev
Human	2.9	8.7	11.8	14.3	17.6	21.7	24.7	26.9	28.5	31.9	36.3	21.0	5.1
Dros.	12.6	13.3	14.8	15.9	18.8	22.0	25.0	28.3	30.0	34.8	41.2	22.2	4.8
Yeast	11.2	17.1	21.4	22.5	24.9	27.0	29.4	31.2	32.1	34.9	42.1	27.0	3.6
E. coli	12.0	12.7	14.8	16.0	18.2	21.0	23.5	26.0	27.6	32.5	40.8	21.0	4.0
B sub.	19.2	19.9	21.0	21.1	23.1	25.7	28.7	32.6	35.9	37.6	45.1	26.4	4.4
HCMV	4.7	5.8	8.1	9.9	13.6	16.4	19.6	24.7	27.7	29.5	30.3	16.9	5.3
EBV	1.4	1.4	9.8	10.9	13.8	16.6	19.9	22.0	23.5	25.6	25.6	16.6	4.8
Vacc	18.5	22.1	27.1	28.4	31.7	33.8	36.3	37.8	38.5	40.8	41.3	33.4	3.9
Hum nuc	10.4	10.4	10.9	12.9	15.0	18.5	21.8	24.4	26.7	32.4	32.4	18.8	4.7
Hum. gp.	2.9	7.8	11.6	13.5	16.8	21.0	24.8	26.9	28.6	30.8	33.2	20.6	5.3
Hum. enz.	9.1	11.3	15.3	16.1	20.0	22.6	24.9	27.3	28.7	32.1	33.1	22.3	4.3
Ec. enz.	12.4	13.8	15.5	16.5	18.6	21.0	23.1	25.1	26.6	34.3	40.8	21.1	3.7
Strong minus weak (G+A+P−F−I−K−N−Y)													
	Min.	.01	.05	.10	.25	.50	75	.90	.95	.99	Max.	Mean	Stdev
Human	−22.8	−17.4	−14.1	−12.1	−9.1	−5.0	0.2	7.3	10.3	19.1	28.7	−3.8	7.6
Dros.	−31.4	−21.3	−15.9	−13.1	−9.3	−5.9	−1.2	3.6	5.1	10.0	16.6	−5.4	6.8
Yeast	−40.5	−23.3	−19.9	−18.2	−15.2	−11.4	−8.4	−6.0	−3.6	2.6	13.0	−11.7	5.3
E. coli	−29.4	−19.5	−11.8	−9.3	−6.5	−3.3	0.3	3.5	5.0	8.6	10.5	−3.2	5.4
B sub.	−32.8	−26.2	−22.6	−19.6	−13.8	−10.8	−7.1	−5.1	−2.4	0.2	1.2	−11.3	5.9
HCMV	−18.8	−17.5	−12.9	−10.5	−3.1	0.9	6.2	9.3	16.8	19.4	23.2	1.0	8.2
EBV	−6.7	−6.7	−5.4	−5.0	−2.4	1.6	8.6	14.0	17.6	37.0	37.0	3.9	8.6
Vacc.	−33.1	−30.4	−29.4	−27.3	−25.4	−21.7	−18.8	−16.5	−13.1	−10.8	1.8	−21.6	5.0
Hum nuc.	−12.9	−12.9	−11.3	−9.4	−6.6	−0.9	5.7	10.9	17.4	19.1	19.1	0.4	8.0
Hum. gp.	−17.8	−16.6	−14.2	−12.0	−9.1	−4.7	0.4	7.6	11.0	20.1	26.7	−3.5	7.7
Hum. enz.	−19.7	−16.8	−14.2	−12.3	−9.2	−6.0	−2.3	2.9	5.8	10.4	12.8	−5.3	5.8
Ec enz.	−29.4	−20.5	−10.2	−8.0	−6.2	−3.0	0.1	2.4	4.3	7.6	8.2	−3.1	4.9

designated by ≻ (see Figures 1 and 4 for examples). A quantile distribution Q is said to be a stochastic dilation (more spread) of $\tilde{Q}$ (designated $Q \gg \tilde{Q}$) if Q and $\tilde{Q}$ have about equal means and the plot of $\tilde{Q}(x)$ crosses the plot of $Q(x)$ once from below to above as x traverses its domain. The notion of stochastic dilation provides information about the relative degree of spread comparing the two distributions. Thus, $Q \gg \tilde{Q}$ entails that the distribution Q has a larger variance than the distribution $\tilde{Q}$ and, more generally, the expectation of any convex function of the variable x is larger for Q than for $\tilde{Q}$.

Compositional anomalies (outliers)

The usage of a given aa type in a protein sequence of length N is considered anomalously high or low if its number of occurrences is three binomial standard deviations above or below the 0.9 and 0.1 quantile points respectively of an appropriate reference set, the binomial standard deviation being defined as $\surd(x^*)(1 - x^*)N$ where x^* is the frequency of the aa type satisfying $Q(x^*) = 0.9$ or $Q(x^*) = 0.1$ respectively.

Table VI. Outlier statistics[a]

Amino acid	No. of proteins of outlier status							
	human		*Drosophila*		yeast		*E.coli*	
	high	low	high	low	high	low	high	low
E	20	–	6	–	3	1	1	1
D	5	1	–	–	6	–	–	–
K	14	3	2	–	2	–	–	–
R	5	2	2	–	1	1	–	–
H	3	–	2	–	1	–	1	–
L	6	8	2	–	–	3	5	–
I	2	2	1	–	1	2	1	–
V	1	3	–	1	–	2	–	1
M	–	–	1	–	1	–	–	–
F	2	2	–	–	–	1	4	–
Y	1	1	1	–	1	1	4	–
W	2	–	–	–	–	–	2	–
P	20	3	6	3	4	2	2	1
G	18	1	–	–	1	–	–	1
A	6	1	6	2	2	1	2	–
S	8	3	2	1	7	–	2	–
T	9	2	1	1	4	1	1	–
Q	11	–	6	–	11	–	2	–
N	7	4	4	–	8	–	1	–
C	12	–	6	–	1	–	3	–
Total no. of proteins	751		227		431		710	
No. of proteins with outliers of any kind	114		45		43		30	
Percentages	15.2		19.8		10.0		4.7	

[a] For definition see Materials and methods.

Results

Tables I–V display the quantile distributions of individual aa usage, cationic, anionic, and aggregate hydrophobic usage and usage of residues classified by codon type. Sample sizes (number of sequences) for each set in Table I exceed 110, reaching 751 sequences for the human collection and 710 sequences for the *E.coli* data; only sequences of at least 200 residues are considered (see Materials and methods).

Some of the quantile distributional differences reflect on the sample sizes and the codon compositional biases extending from yeast (overall genomic G + C% ≃ 41%), *B.subtilis* (≃43%), *E.coli* (≃52%), human (≃53%) and *Drosophila* (≃55%) (Cherry, 1991). The human viral genomes of CMV and vaccinia differ sharply in G + C content, 58% and 38% respectively.

Medians and central 80% quantile range

For all categories of aa usage and organism type, the mean and median values are close, generally showing a slightly greater mean. The 0.1–0.9 quantile ranges for most aa types tend to be of similar length across species. The medians of negatively charged residues over most species are about 11.4–12.2%, but the human high extreme levels have substantially higher usage frequencies, 0.99 quantile = 23.0% compared with 16.2% in *E.coli*.

Amino acids of most and least frequent usage for various species

The most frequently used aa (in terms of mean and median values) in almost all species is Leu, although in *E.coli* Ala is a virtual tie. The least frequently used aa is, generally, Trp in the eukaryotic species and in the viruses, and Cys in the prokaryotes, *E.coli* and *B.subtilis*. Cys, generally, is used ~1% in the

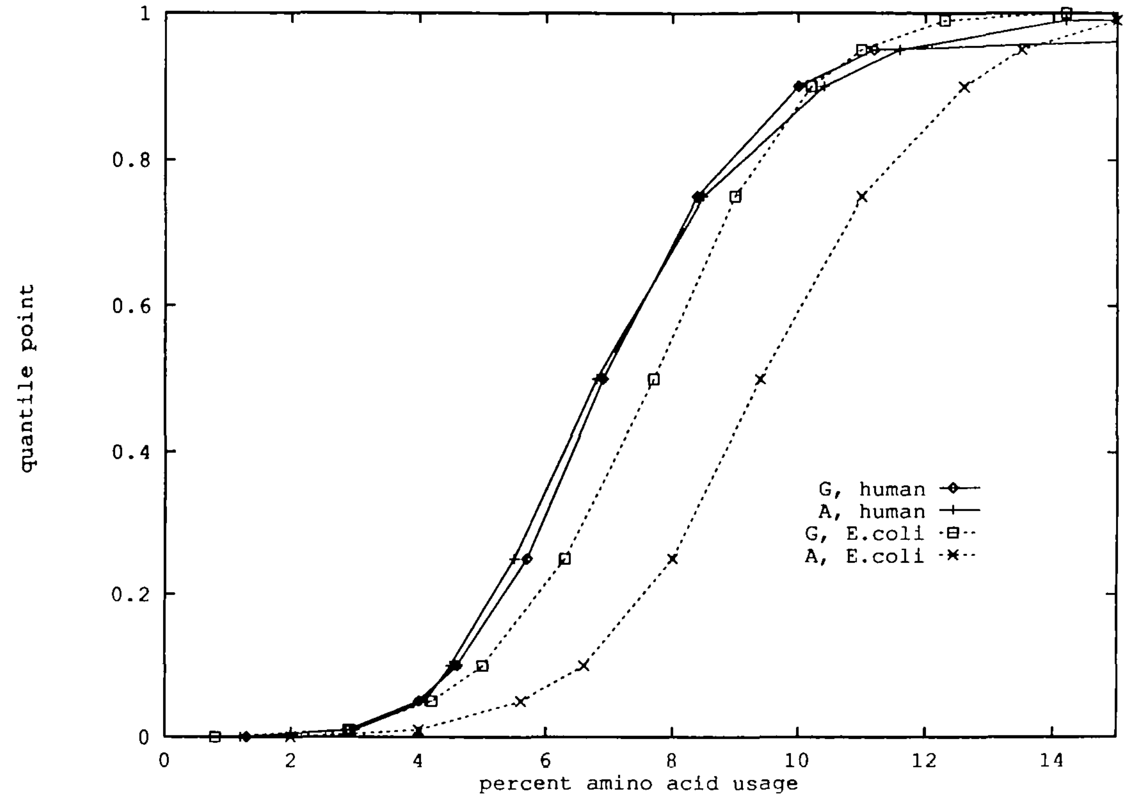

Fig. 1. Glycine and alanine quantile distributions in human and *E.coli*. Note that stochastic ordering implies nonintersecting distribution functions with the stochastically bigger distribution entirely on the right side.

unicellular species compared with >2% in the higher eukaryotes. Cys usage entails quantile distributions that markedly deviate between human and *E.coli*. Nearly 10% of the *E.coli* proteins, compared with ~5% of the human proteins, are devoid of Cys residues. At the high extreme, *E.coli* lacks Cys-rich proteins (99% quantile = 3.8%), whereas the 99% quantile = 7.4% in the human protein collection.

Charged amino acid usages

The 0.1 −0.9 quantile points of negatively and positively charged aa are largely concordant in all the species examined. The extreme values (corresponding to the quantile points 0.01, 0.05, 0.95 and 0.99) show substantial variation between species (especially for unicellular versus multicellular species).

Although the aggregate average positive charge frequency per protein is approximately constant across species, ~11.5%, Lys and Arg individually vary substantially. Thus, Arg is less frequent (actually stochastically smaller; Figure 3) in human compared with *E.coli* proteins. The human nuclear proteins (Tables II and III), on average, contain relatively more positively charged residues than the overall human protein sequences and 2% more than cationic occurrences in human glycoproteins.

The median and mean uses of acidic residues (D + E) are nearly invariant across species, confined to the range 11.2−13.2%, with E on average 6.4% and D on average 5.5%. It seems paradoxical that of the quantile charge tables, CMV proteins use on average the fewest acidic residues (9.5%), although CMV incorporates eight ORF sequences containing hyper acidic charge runs (a hyper charge run is an extremely long run including at least nine contiguous residues of the specified charge type, a rare feature of a protein sequence; see S.Karlin and B.E.Blaisdell, manuscript in preparation).

Independent of species, the net charge on average (and in median terms) is slightly negative (see Table III), with the marked exception of the ORF sequences of the CMV genome. Parenthetically, CMV has >20 substantial (≥400 residues) ORFs with significantly low positive charge usage. The 0.01 −0.99 range of the total charge quantile distributions (Table III) expands (apparently not dependent on the numbers of sequences) with the organism complexity: *E.coli*, 9.0−31%; yeast, 12.6−36.7%; *Drosophila*, 11.0−44.2%; human, 11.2−40.3%. It is intriguing that the total charge, on average, is reduced by >2% in CMV proteins (and all human herpes virus ORFs, data not shown) relative to species proteins.

Hydrophobic residue usages

The aggregate of strong hydrophobic aa (Leu, Ile, Val, Phe, Met) is dominating among the prokaryotic (*E.coli*, *B.subtilis*) proteins at every quantile level compared with all the eukaryotic species examined (Figure 2). The viruses (CMV and vaccinia) at the median level show about the same hydrophobic quantile points as the prokaryotic species, but at the high quantile levels hydrophobic residue usage is lower, similar to the eukaryotic sequences. By contrast, the subclass of the human enzyme quantile distribution dominates stochastically the corresponding quantile distribution of the *E.coli* enzyme set (Table V). The human nuclear protein class has (except at the minimum point) the lowest (subordinating) hydrophobic quantile levels. In particular, the human nuclear proteins are rich in charged and general hydrophilic aa and also proline.

Amino acid usages of strong and weak codon types

The strong codon aa group S = {Gly, Ala, Pro} is translated from codon types SSN (S is the nucleotide C or G, N is any nucleotide) and the weak codon aa group comprises *W* = {Phe, Ile, Lys, Asn, Tyr}. The multicellular eukaryotic protein sets favor greater use of *S* aa compared with the (A + T-rich) unicellular yeast and *B.subtilis* gene sequences. The reverse stochastic ordering holds for *W* aa types (Table IV).

Comparison of the extremes of the quantile distributions

A varied picture is seen from the tails of the distributions. The human protein sequences are partitioned into proteins that are extremely rich or extremely poor in several residue types. Ten different aa are observed to be absent from at least one human protein, including the relatively abundant residue Thr. Only five aa are absent from one or more *E.coli* sequences. At the high usage levels, seven aa in human (Leu, Gly, Ala, Pro, Gln, Glu, Lys) reach quantile frequencies >20%. In *E.coli*, only Leu and Ala surpass this mark.

The percentage of proteins exhibiting quantile distributional outliers, as defined in Materials and methods, varies greatly over species: *Drosophila* (19.8%) > human (15.2%) > yeast (10.0%) > *E.coli* (4.7%). The spectrum of aa giving rise to outliers is strongly species-dependent. Outliers on the high side tend to be hydrophilic and Gly in the eukaryotic species, but hydrophobic in *E.coli*. Thus, the human set includes 20 proteins with very extreme Glu usage, as compared with only one in *E.coli*. Generally, the expanse of the tails of the distributions increases with organismal complexity. This trend is reflected in the number of aa in each species, with minima of 0% and maxima ≥20%: *E.coli*, 5:2; yeast 6:5; *Drosophila*, 7:5; human, 10:7% (Table I).

Comparisons of average (or median) amino acid usages for codon degeneracy classes

Degeneracy-1 group. Pervasively, Trp < Met. For the prokaryotes (*E.coli*, *B.subtilis*), Cys < Trp. Except for the human collection, Cys is the second least frequently used aa.

Degeneracy-2 group. The quantile distributions of the two-degeneracy (2-codon) aa are in the main congruent across species except for Cys and Lys. The following nearly species invariant median use pattern applies for the two-degeneracy aa: Lys, Glu > Asp > Gln, Asn > Phe, Tyr > His > Cys. For the viruses, the genome compositional biases have some influence.

Degeneracy-4 group. A weak general trend indicates Ala > Gly > Val > Thr, Pro. Consistent with the weak base genome bias, usage of Ala and Gly is low in yeast and very low in vaccinia. Pro has low usage in prokaryotes and comparatively high usage in human nuclear proteins. The use of Pro in humans is high compared with that in *E.coli*. This probably reflects the higher average G + C content of human genes and the profusion of collagen-like and other extra-cellular proteins.

Degeneracy-6 group. The general trend entails Leu > Ser > Arg. Deviations from this pattern occur for the nuclear proteins where Ser > Leu.

Stochastic orderings within and between species

Glu ≻ Asp (i.e. Glu usage is stochastically larger than Asp usage, see Materials and methods) for most protein data sets. Lys ≻ Arg in yeast and vaccinia consonant with their weak base-rich genomes. In most data sets there is no consistent stochastic ordering pattern between Lys and Arg.

The stochastic ordering Phe ≻ Tyr (exception vaccinia) holds for most organisms examined despite the fact that Dayhoff *et al.* (1978) rank Phe and Tyr the highest in aa exchange ratio.

For every quantile level the percent use of hydrophobics entails the stochastic ordering (*E.coli*, *B.subtilis*) ≻ yeast ≻ human ≻ *Drosophila* (see Figure 2 and Table IV).

The quantile distributions of acidic versus basic residues (Table

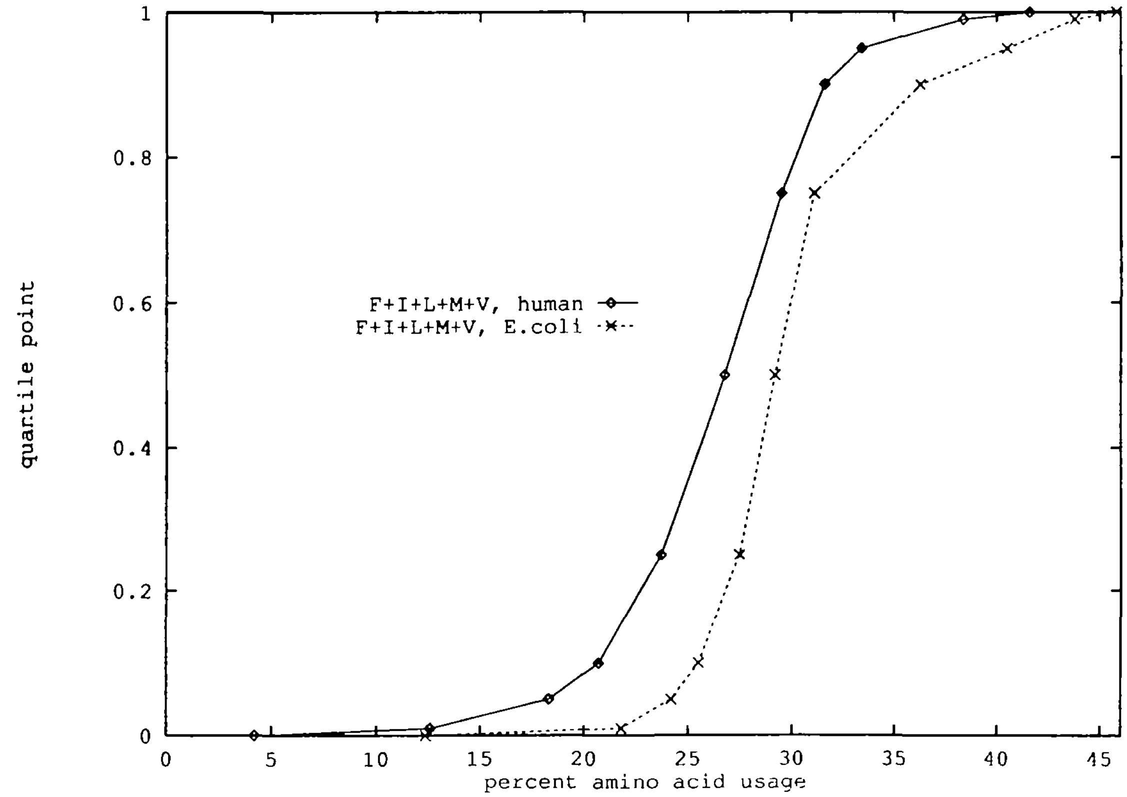

Fig. 2. Quantile distribution plots for *E.coli* and human aggregate hydrophobic aa.

III, Figure 4) for the human nuclear protein set cross at least twice indicating no distinctive preferences in usages, although total charge in nuclear proteins is relatively high compared with glycoproteins.

Discussion

Invariants, contrasts and conundrums

Various questions are raised by the data. How is aa usage affected by protein structure and genomic organization, aa biosynthesis,

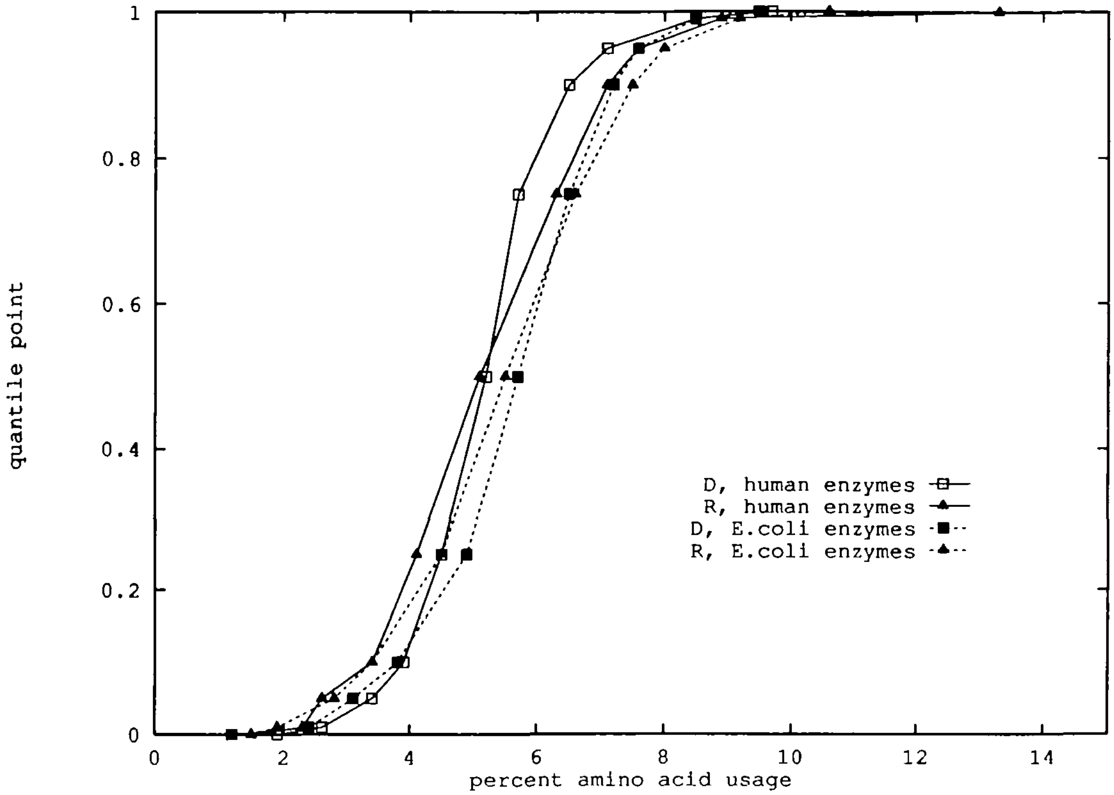

Fig. 3. Arginine and aspartate quantile distribution plots for human and *E.coli* enzyme sequences.

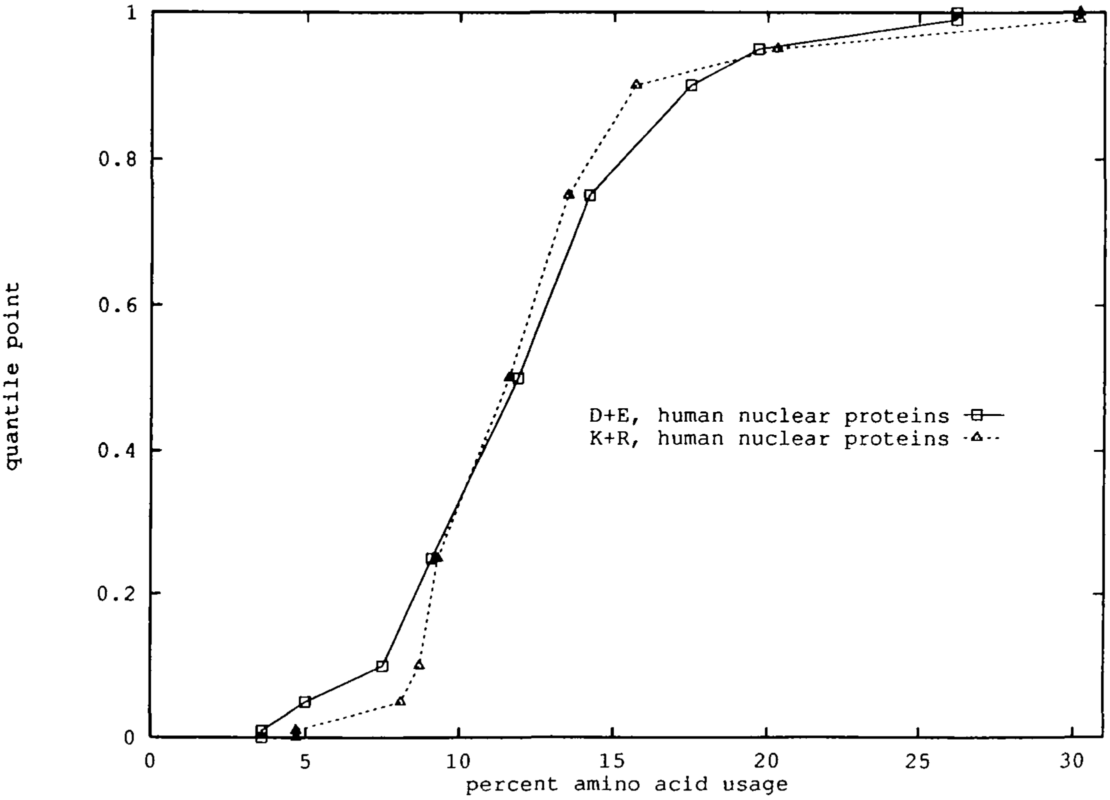

Fig. 4. Acidic versus basic charge quantile distribution plots for human.

relative abundances of free aa, tRNA availabilities and evolutionary founder effects? We highlight and venture interpretations and hypotheses on the principal findings.

Charge compositional biases

For all species sets, the median and mean net charges of a protein are slightly negative (~ −0.5%), whereas the human herpes virus ORFs show on average a slightly positive net charge, ~ +0.3% (Karlin and Brendel, 1992). In this study histidine with positively charged residues was not included because in the normal cellular ambience His is uncharged (Watson *et al.*, 1987; Stryer, 1988). In some previous publications, e.g. Karlin (1990) and Karlin *et al.* (1991), positive charge clusters and runs with and without His were analyzed. The total positive charge (Lys + Arg) per protein is generally constant over species, ~11.5%. Individually, the median Lys and Arg frequencies per protein vary across the different species. For example, in the human set Arg is under-represented, presumably because of CpG suppression, while in *E.coli* Lys is under-represented (see Tables I and II). For the major human herpes viruses, Lys is broadly under-used and Arg broadly over-used as in the prokaryotic data sets (Karlin and Brendel, 1992).

Of all aa Glu frequencies broadly show the greatest departure from proportionality to codon degeneracy. It is curious that evolution did not opt for more acidic aa codons. One might speculate that the code was, in the main, fixed early in evolutionary time, and compensated later by increased availability of acidic tRNA and aminoacyl-tRNA synthetase molecules. In contrast, the average basic residue usage (11.5%) is much closer to the average frequencies under random codon usage. The average level of basic residues drops to ~10.3−10.7% for the enzyme subsets of the human and *E.coli* sequences. This is consistent with the observation that enzymes, in general, rarely feature anomalous charge distributions (Karlin, 1990). In humans the median Lys frequency, 5.7%, is significantly greater than the 3.3% expected from random codon usage. The average of degeneracy-2 arginine (codons AGR) is 2.4% for human but only 0.05% in *E.coli*, both much lower than expected from random codon usage. But the average of degeneracy-4 arginine (codons CGN) is ~3.2% for human and ~5.6% for *E.coli*. It is striking that frequencies of the four charged aa deviate more significantly from proportionality to degeneracy than do any of the 16 uncharged aa.

Why is Glu stochastically larger than Asp, that is, used more at all levels of use? From a structural viewpoint, Asp is recognized as an α-helix breaker, whereas Glu is favorable to α-helix formation. Moreover, the side chain of Glu involves two methylene groups as against a single methylene group in Asp providing greater conformational flexibility. Asp and Glu are encoded by similar codon forms (GAR and GAY respectively), but possibly the juxtaposition of purine−pyrimidine (AY) at codon sites 2 and 3 is sterically unfavorable compared with the purine−purine arrangement (AR). Apropos, polypurine runs for unknown reasons tend to be over-represented in genomic sequences (Bucher and Yagil, 1991).

Why do the majority of species proteins favor a net negative charge (Karlin *et al.*, 1991)? Residues on the surface of proteins presumably need to be highly selective to be able to interact with appropriate structures or avoid interacting with other structures. From this perspective, the general net negative charge may better avoid (mediated by electrostatic repulsion) undesirable interactions with DNA, RNA, membrane surfaces and certain other proteins. The extracellular milieu for metazoans is slightly alkaline, with pH ~7.2−7.4 (Roos, 1981), whereas the intracellular pH is quite variable ranging from 5.0 to 7.2, depending on tissue type and subcellular localizations (Alberts *et al.*, 1983; Stryer, 1988). One might speculate that enzyme activity is 'optimum' at a pH similar to the pH of the host cells, which in mammalian organisms is commonly slightly acidic. Moreover, the protein negative charge tendency can contribute in modulating secretion and intracellular transport, in inducing transcriptional activation and generally in mediating rapid and potent interactions of protein assemblages.

Use of aggregate hydrophobic aa

The major hydrophobic aa tend to be over-represented in the prokaryotes (*E.coli, B.subtilis*) compared with eukaryotic species (human, chicken, *Xenopus*, *Drosophila*, yeast) suggesting that proteins enveloping a substantial hydrophobic core are relatively more common in prokaryotes compared with eukaryotes (Figure 2), but this does not hold for the subclass of enzymes.

Human versus E.coli aa quantile distributions

A natural set of quantile distribution comparisons apply to human versus *E.coli* for each residue type, because sample sizes are about equal (751 and 710 sequences respectively) and both are of broad functional distribution. The following stochastic dominance orderings prevail: *E.coli* ≻ human for residue types Leu, Ile, Val, Ala, Met, Arg, aggregate hydrophobics, basic, emphasizing the major hydrophobics except for the aromatic Phe; human ≻ *E.coli* for residue types Cys, Pro, Ser, Glu, acidic (all with relatively small side chains); no definite stochastic ordering between human and *E.coli* is seen for the remaining residue types. Strikingly, human enzymes ≻ *E.coli* enzymes for the aggregate of hydrophobic aa usages. For both the human and *E.coli* enzyme sets, the quantile distribution of Arg is a stochastic dilation (is more spread out, see Materials and methods) over the quantile distribution of Asp (Figure 3). This property for enzymes is persistent and independent of a bias for or against arginine (e.g. CpG suppression). Such a stochastic dilation is not true for the complete species protein collections.

The large numbers of stochastic orderings attest to the ancient divergence between *E.coli* and human in aa usages. Quantile distributions for the variable of sequence length (for proteins ≥200 residues) show that the median length is ~450 residues for eukaryotic species sequences but only ~370 residues for *E.coli* sequences (data not shown). Consistent to the smaller protein sizes, one might expect more statistical variation and more extremes in aa usages for *E.coli* versus human sequences; however, just the opposite is observed. The abundance of extremes for the human sequences putatively reflects the greater complexity of protein activities in the highly differentiated eukaryotic cells.

Functional and structural determinants

To what extent do protein structural and functional requirements determine aa frequencies? The results in Tables II−V suggest that the human nuclear proteins emphasize hydrophilic residues compared with cellular enzymes and glycoproteins in which hydrophobic residues are foremost. It is generally accepted that charged residues are either exposed to solvent or, if buried, are likely to occur in pairs of opposite charge. From this perspective and the expectation that most proteins would avoid unnecessary assemblages and interactions, it is proposed that surface residues tend to be more acidic than basic, thus reducing undesired nonspecific ionic interactions. This hypothesis is consistent with recent characterizations of aa 'environments' in protein structures (Bowie *et al.*, 1991). In their analysis, for an exposed or partially exposed polar residue environment in many secondary structure contexts, the scores of the aa Glu and Asp are positive, while exposed environment scores of Lys and Arg are negative, signifying that the acidic and basic residues are over-represented and under-represented respectively in this 'surface' environment.

Amino acid biosynthesis and abundance

Do aa which are easier to synthesize and/or to be acquired from external sources tend to be used more in proteins? How is this reflected in intracellular aa abundances? Does the biosynthetic pathway complexity (i.e. the number of enzymatic steps or the nonessential or essential character of the particular aa) reflect on aa usage? In this context, Glu is at the center of the web of aa biosynthetic pathways and Asp is synthesized with but one additional enzymatic step (Stryer, 1988) which may, in part, account for the relatively high residue usages of Glu and Asp (Glu stochastically greater than Asp). Consistent with their significant over-representations, acidic residues often exhibit unusual distributions in protein sequences, including a preponderance of very long acidic runs especially pronounced in connection with multiprotein complexes. Thus, many fundamental nuclear and extra-cellular proteins carry unusually long acidic runs or mixed charge runs favoring acidic residues. This is particularly shown by proteins of the nucleolus and those that are involved in RNA and DNA processing including nucleolin, topoisomerase I, UBF, HMG1, U1snRNP, U2snRNP, myosin light chain kinase, troponins C and T, neurofilament triplet L, lamins A, B and C, CENP-B, calreticulin and others. Long anionic charge runs are also prevalent in many proteins associated with ionic transport including voltage-gated Na^+ channel, nicotinic acetylcholine receptor, AE1 and AE3 anionic exchange proteins, Ca^{2+} transporter and others. In sharp contrast, there are essentially no proteins with very long cationic runs (S.Karlin and B.E.Blaisdell, manuscript in preparation). In fact, the longest cationic run observed among the current human protein collection (751 sequences) is a single nonapeptide (in GC rich DNA binding factor), while there are many proteins carrying anionic runs of >14 residues length (e.g. nucleolin, myc, calreticulin).

How are relative intracellular aa concentrations reflected in aa usages? One might expect that cellular aa concentrations correlate negatively with the complexity of the biosynthetic pathway and correlate positively to the nonessential character of the aa. From this perspective, Glu would be a relatively abundant aa in all protein categories (species, function, localization), which it is. In humans the essential aa vary greatly in abundance from least, Trp, to most, Leu. The primary aa of the human biosynthetic pathways, Asp, Glu, Ser, Ala and Gly, from which others derived in part, rank 9, 6, 5, 3 and 2 out of 20 respectively.

Founder effects

Is there a remnant of founder effects relevant to aa usages? Wong and Cedergren (1986) speculate that those aa derived directly from the prebiotic synthesis manifest a higher frequency in today's protein universe. There is the speculation that the earliest peptides were composed of few aa which are the ones most abundant today (Wong, 1988). The chemically and metabolically simplest aa to accumulate are Gly, Asp, Glu and Ala and they are, therefore, considered likely to have been the most abundant in the primitive biosphere. Indeed, the average frequencies of these aa tend to exceed expectations (Tables I and II) compared with random codon usage. In particular, Ala and Gly are much greater than expected in *E.coli* compared with human. Miller (1986) estimated the relative abundances of aa found in meteorites in decreasing rank order to be Gly, Ala, Val, Asp, Glu and Pro, a ranking broadly concordant with aa usages. Evolutionary processes have certainly expanded and diluted the protein repertoire and reduced the amounts of the most over-represented aa. Just as the genetic code is not frozen, as deduced from changing codon assignments (Osawa *et al.*, 1992), aa usages are likely to be in a state of dynamic evolution, with new proteins continuously being formed and others eliminated. Crick *et al.* (1976) speculated that in a methane-rich high temperature environment early translation events favored G + C-rich DNA sequences. This kind of DNA distribution is not consistent with current representations of aa usage (neither for averages nor extremes).

Relative highs and lows in aa usage

The predominance of leucine among protein sequences certainly reflects its important role in hydrophobic core structures, in transmembrane segments, in signal peptides and its prevalence and stability in secondary and tertiary structures. The relatively high alanine frequency in proteins also reflects on its α-helix stability and flexible hydrophobic properties. Interestingly, in human nuclear proteins serine is foremost.

Cysteine exhibits unusual quantile distributions in many species, sharply disparate between human and *E.coli*. Nearly 10% of the *E.coli* proteins and ~5% of the human proteins are devoid of cysteine residues (these include many ribosomal proteins and proteins functioning in mRNA processing). The dearth of cysteine-rich proteins in *E.coli* may reflect the near absence of extracellular proteins, whereas the human collection features many cysteine-rich secreted proteins, e.g. blood-clotting factors, proteins of the complement series and an assortment of glycoproteins. Even *B.subtilis*, in possession of relatively more secreted proteins than *E.coli*, exhibits significantly low cysteine usage. Apropros, no zinc finger proteins have been uncovered to date in prokaryotes (Branden and Tooze, 1991; Luisi, 1992).

The quantile distributions of Gly, Pro and Cys in the human protein sequences exhibit relatively long tails (especially at the high extreme). This may merely reflect the protein sequence sampling bias exemplified by large numbers of collagen types, keratins and excreted proteins—the first two types abundant with Gly and Pro, the latter type enriched with cysteine kringles, EGF-like domains and disulfide bonds. The increasing discovery of zinc finger or other metal ion coordinating nuclear proteins may also be relevant.

The pervasive stochastic dominance ordering Phe ≻ Tyr, valid for all species protein sets, is difficult to explain. The aa occurrences of Phe and Tyr strongly correlate as do all pairs of aromatic aa (Karlin and Bucher, 1992). Phe is a precursor of Tyr in the path to acetyl CoA a precursor of many important biological molecules. It is noteworthy that tyrosine is encoded from the codon TAY where the dinucleotide TpA is pervasively under-represented (Burge *et al.*, 1992) and the least energetically stable among all dinucleotides (Bresslauer *et al.*, 1986; Delcourt and Blake, 1991). Moreover, tyrosine is often an important phosphorylation target site in effecting protein conformational and functional changes which may explain its limited judicious use compared with phenylalanine.

Further data studies

(i) It would be informative to ascertain quantile distributions for the various protein classes in other aa classifications including the chemical, functional, structural and size alphabets (for definitions see Karlin *et al.*, 1991). (ii) The methodology of quantile distributions can also be applied to general DNA sequences in various alphabets (e.g. purine, pyrimidine) and to gene sequences with respect to codons, or silent site frequencies, etc. (iii) Multivariate quantile distributions are also feasible, e.g. examining simultaneously usage frequencies for charge, size and hydrophobicity. (iv) It would be desirable to extend quantile distributions and correlation analyses to other natural function or structure protein classes including sequences of the superimmune family, proteins of cytoskeletal associations, kinases, developmental genes and transcription factors. (v) With respect to evolutionary perspectives, quantile tables of homologous proteins (e.g. the globin family) would be of interest. (vi) Other variables associated with protein classes for which quantile distributions are natural include the length and kDa assessments of a protein sequence, multiplet counts (Karlin and Brendel, 1992) and observations or predictions on secondary structures (number of α-helices, β-strands).

Possible experiments suggested by the data and theory

(i) It is paradoxical that Lys and Arg usage tend to be uncorrelated or negatively correlated but are scored high in the PAM exchange matrix (Dayhoff *et al.*, 1978). Our previous discussion underscored chemical, shape and ionic differences. It would seem interesting to conduct large scale replacements of Lys ↔ Arg in various protein classes, particularly those with extreme frequencies of Lys and Arg, and evaluate consequences on function and structure. (ii) The almost universal stochastic ordering Glu ≻ Asp indicates preference of Glu over Asp at all levels of use. Our previous discussion suggested possible reasons for this. Again, focused studies of replacements of Glu ↔ Asp might help elucidate the relative role of Glu versus Asp. (iii) Similar replacement experiments relevant to the universal stochastic ordering Phe ≻ Tyr could be of interest. (iv) Leu is broadly of abundant use (overall it has the highest frequency in proteins). To what extent and for which protein species can Val or Ile substitute or not substitute for Leu or Ala for Leu? (v) Ser entails the highest frequency in mammalian nuclear proteins. What substitutions preserve functions in these cases?

Quantile implications for sequence comparisons

Contrary to intuition, compositional differences appear to be more pronounced between species than between function classes. This has obvious implications for phylogenetic reconstructions as well as for the statistical evaluation of weak protein sequence similarities. Overestimates of evolutionary distances may result from not taking divergent species-specific compositional constraints into account. The significance of high scoring matches between segments of similarly biased aa composition might be better evaluated on the basis of compositional extremes in the respective species. In this context we might propose sequence comparisons based on asymmetric PAM matrix scores (Dayhoff *et al.*, 1978) as a constituent of a realistic model of protein sequence evolution. The sharp differences of 'strong minus weak' codon types in CMV versus vaccinia quantile distributions may be explained from this perspective.

Many authors, as reviewed in Introduction, have written on variation in the aa composition of proteins generally in terms of averages. This paper shows that composition quantile distributions and the recognition of stochastic dominance relations allow more refined and robust comparative assessments of the aa compositional variation of proteins. These include observations on the universal stochastic dominance of glutamate over aspartate and of phenylalanine over tyrosine, and the sharp contrasts of usage associations between acidic residues versus usage associations between basic residues. Our interpretations and speculations focused on invariants and contrasts of the aa compositional spectrum in relation to species, function, cellular and tissue localization, biochemical and steric attributes, complexity of the different aa biosynthetic pathways, aa relative abundances, tRNA availabilities, translation fidelity and efficiency, early historical events, and evolutionary processes. To what extent some of the results may be artifacts of sample bias in the current collections of sequences of the databases is unclear. The acquisition of more complete genomes (or even chromosomes) over the next decade can help in resolving uncertainties through applications of the concepts and methods of quantile distributions of enhanced power.

Acknowledgements

We are happy to acknowledge the valuable comments on the manuscript from Drs V.Brendel, C.Burge, A.M.Campbell, R.Doolittle, E.Mocarski, R.Sapolsky and M.Zuker. S.K. and B.E.B. were supported in part by NIH grants GMHG00335-03, GM10452-28 and NSF grant DMS86-06244.

References

Aissani,B., D'Onofrio,G., Mouchiroud,D., Gardiner,D., Gautier,C. and Bernardi,G (1991) *J. Mol. Evol.*, **32**, 493–503.

Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (1983) *Molecular Biology of the Cell.* Garland, New York.

Bairoch,A. and Boeckmann,B. (1991) *Nucleic Acids Res.*, **19**, 2247–2249.

Bowie,J.U., Lüthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.

Branden,C. and Tooze,C. (1991) *Introduction to Protein Structure.* Garland, New York.

Brendel,V. (1992) *Math. Comput. Modelling*, **16**, 37–45.

Bresslauer,K.J., Frank,R., Blöcker,H. and Marky,L.A. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.

Bucher,P. and Yagil,G. (1991) *DNA Seq.*, **1**, 159–172.

Burge,C., Campbell,A. and Karlin,S. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.

Cherry,M. (1991) Tables of codon frequencies, calculated for different organisms. Obtained from the EMBL netserver.

Crick,F.H.C., Brenner,S., Klug,A. and Pieczenik,G. (1976) *Origins Life*, **7**, 389–397.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Delcourt,S.C. and Blake,R.D. (1991) *J. Biol. Chem.*, **266**, 15160–15169.

D'Onofrio,G., Mouchiroud,D., Aissani,B., Gautier,C. and Bernardi,G. (1991) *J. Mol. Evol.*, **32**, 504–510.

Doolittle,R.F. (1986) *Of URFS and ORFS.* University Science Press, Mill Valley, CA.

Ikemura,T., Wada,K. and Aota,S. (1990) *Genomics*, **8**, 207–216.

Karlin,S. (1990) In Sarma,R.H. and Sarma,M.H. (eds), *Proceedings of the Sixth Conservation in Biomolecular Stereodynamics.* Adenine Press, Albany, NY, pp. 85–95.

Karlin,S. and Brendel,V. (1992) *Science*, **257**, 39–49.

Karlin,S. and Bucher,P. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 12165–12169.

Karlin,S., Bucher,P., Brendel,V. and Altschul,S. (1991) *Ann. Rev. Biophys. Biophys. Chem.*, **20**, 175–203.

King,J.L. and Jukes,T.H. (1969) *Science*, **164**, 788–798.

Luisi,B. (1992) *Nature*, **356**, 379–382.

McCaldon,P. and Argos,P. (1988) *Proteins: Struct., Funct. Genet.*, **4**, 99–122.

Miller,S.L. (1986) *Chem. Scripta*, **26B**, 5–11.

Nakashima,H., Nishikawa,K. and Ooi,T. (1986) *J. Biochem.*, **99**, 153–162.

Osawa,S., Jukes,T.H., Watanabe,K. and Muto,A. (1992) *Microbiol. Rev.*, **56**, 229–264.

Pečarić,J.E., Proschan,F. and Tong,Y.L. (1992) *Convex Function, Partial Orderings and Statistical Applications.* Academic Press, New York.

Roos,S. (1981) *Physiol. Rev.*, **61**, 296–334.

Stryer,L. (1988) *Biochemistry.* Freeman, New York.

Sueoka,N. (1961) *Proc. Natl Acad. Sci. USA*, **47**, 1141–1149.

Watson,J.D., Hopkins,N.H., Roberts,J.W., Steitz,J.A. and Weiner,A.M. (1987) *Molecular Biology of the Gene*, 4th edition. Benjamin Cummings, Menlo Park, CA.

Wong,J.T. (1988) *Microbiol. Sci.*, **5**, 174–181.

Wong,J.T. and Cedergren,R. (1986) *Eur. J. Biochem.*, **159**, 175–180.

Received on July 1, 1992; revised on September 12, 1992; accepted on September 16, 1992