How representative are the known structures of the proteins encoded by [AU:OK?] a complete genome? A comprehensive structural census

Mark Gerstein

Background: Determining how representative the known structures are of the proteins encoded by [AU:OK?] a complete genome is important for assessing to what extent our current picture of protein stability and folding is overly influenced by biases in the structure databank (PDB). It is also important for improving database-based methods of structure prediction and genome annotation.

Results: The known structures are compared to the proteins encoded by eight complete microbial genomes in terms of simple statistics such as sequence length, composition and secondary structure. The known structures are represented by a collection of nonhomologous domains from the PDB and a smaller list of 'biophysical proteins' on which folding experiments have concentrated. The proteins encoded by the genomes are considered as a whole and divided into various regions, such as known-structure homologue, low complexity (nonglobular), transmembrane or linker. Various tests are performed to assess the significance of the reported differences, in both a practical and a statistical sense.

Conclusions: The proteins encoded by the genomes are significantly different from those in the PDB. Their sequence lengths, which follow an extreme value distribution, are longer than the PDB proteins and much longer than the biophysical proteins. Their composition differs from the PDB proteins in having more Lys, Ile, Asn and Gln and less Cys and Trp. This is true overall and especially for the regions corresponding to soluble proteins of as yet unknown fold. Secondary-structure prediction on these uncharacterized regions indicates that they contain on average more helical structure than the PDB; differences about this mean are small, with yeast having slightly more β structure and *Haemophilus influenzae* and *Helicobacter pylori* more α structure. Further information is available through the GeneCensus system at http://bioinfo.mbb.yale.edu/genome.

Address: Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA.
E-mail: Mark.Gerstein@yale.edu

Key words: biophysics, protein folding, protein structure, structure databank

Received: 17 July 1998
Revisions requested: 18 August 1998
Revisions received: 11 September 1998
Accepted: 28 October 1998

Published: xx xxx 1998
http://biomednet.com/elecref/13590278003xxxxx

Folding & Design xx xxx 1998, 3:000–000

[AU: this paper is somewhat longer than our normal 12-page limit. Please consider parts that could be shortened.]

Introduction

The advent of complete genome sequences allows us to reassess our understanding of proteins and, in particular, protein structure. Most accounts of what proteins 'look like' have underlying them an implicit statistical picture of the 'average protein' — its composition, length, and so forth. This statistical picture is based to a great extent on the properties of the known structures in the Protein Databank (PDB), however, and the selection of proteins in the PDB is highly biased by the preferences of individual investigators and by the physical constraints on what will crystallize (or can be studied by NMR spectroscopy). The selection of proteins encoded by a complete genome, by contrast, is in a sense unbiased, representing the total complement of proteins necessary for an organism to live.

The objective of this paper is to understand how biased the collection of known structures is by comparing them in a statistical fashion to the proteins in a number of recently completed genomes. The comparison focuses on simple measures such as the distribution of sequence lengths, amino acid and secondary structure composition, and the occurrence of transmembrane segments and low-complexity regions. Beyond simply illuminating the biases in the structure databank, this work has important implications for database-based structure prediction and modeling algorithms. These algorithms all essentially try to extrapolate what has already been seen in the database to a new uncharacterized protein. For instance, secondary-structure prediction consists of observing the patterns of amino acids that are associated (albeit often weakly) with helices and strands in the databank, and then assigning secondary structure to an unknown protein based on the occurrence of these patterns in its sequence [1–8]. If the database is so biased that the amino acid and secondary structure composition in the known structures is not rep-

Table 1

Composition of the PDB.

	Soluble PDB (PS)	All-β (PB)	All-α (PA)	Mixed (PAB)
Number of sequences	1135	266	207	662
Number of amino acids	192313	42845	29908	119560
Average length	169	161	144	181
Residue:				
A	8.40%	6.8%	9.2%	8.7%
C	1.72%	1.6%	1.4%	1.8%
D	5.91%	5.9%	5.8%	6.1%
E	6.29%	5.2%	7.3%	6.3%
F	3.94%	4.2%	4.2%	3.9%
G	7.79%	8.4%	6.4%	7.9%
H	2.19%	2.1%	2.2%	2.2%
I	5.54%	5.4%	5.1%	5.8%
K	6.02%	5.6%	6.5%	5.9%
L	8.37%	7.3%	9.6%	8.4%
M	2.15%	1.7%	2.4%	2.2%
N	4.57%	5.3%	4.4%	4.5%
P	4.70%	5.1%	4.4%	4.6%
Q	3.73%	3.5%	4.2%	3.7%
R	4.78%	4.2%	5.4%	4.8%
S	5.97%	7.2%	5.7%	5.6%
T	5.87%	7.2%	5.2%	5.5%
V	6.96%	7.6%	5.7%	7.1%
W	1.46%	1.7%	1.5%	1.3%
Y	3.64%	3.8%	3.5%	3.7%

Statistics relating to the length and composition of some representative subsets of the PDB. PS is 1135 domains derived from applying multiple-linkage clustering to the soluble proteins in the PDB. PA, PB and PAB are proper subsets of PS corresponding to all-α, all-β, and mixed proteins.

resentative of that in the unknown protein, however, one would not expect prediction (by any algorithm) to be very meaningful.

This work follows up on much recent analysis of genomes (or partial genomes). Automated methods have been developed for annotating whole genomes [9–11], for example, and the number of membrane proteins encoded has been surveyed [12–16]. Genomes have also been compared on the basis of the frequencies of oligonucleotide and oligopeptide words [17–20], and censuses have been done on the occurrence of various fold families in genomes [21–24]. This paper also follows up on recent work looking at how the effects of compositional and length biases in the PDB affect various prediction and sequence comparison methods [25–27].

Results

A representative selection of structures from the PDB

The obvious first step in this analysis is deciding exactly what structures one should take as representing the known structures. One could, for instance, take all the structures in the PDB, of which there are currently about 5500 (5493 identifiers and 10781 domains, see the Materials and methods section). This would be clearly biased, though, by the fact that for some, but not all, of the proteins in the PDB there are many mutant or highly homologous structures or many structures of the same protein in different conformations or liganded states. (For instance, there are 154 structures for immunoglobulin variable domains and 222 structures for T4 lysozyme, but only a single structure for the equally important tyrosine kinase and topisomerase II proteins.) Brenner *et al.* [28], in fact, report that 9 out of 10 of the new structures deposited in the PDB are just minor variants of what is already in the database. Consequently, for analyses such as the one here, the PDB is usually clustered into a representative set of unique chains or domains, and this is done here based on amino acid sequence by a new algorithm (described in the Materials and methods section). It gives 1135 representative domains of soluble proteins — the amino acid composition of these domains is shown in Table 1 and their length distribution in Figure 1. The average length of a PDB domain is ~170 residues. (A few very long domains skew this average, so that the most common length, i.e. the mode, is around 120 residues.) The length and composition statistics for these 1135 domains will serve as a standard for comparison against the genome proteins. When, in the following discussion, reference is made to the "average length of the known structures" or the "composition of the PDB", one is directed to the statistics in Table 1 and Figure 1.

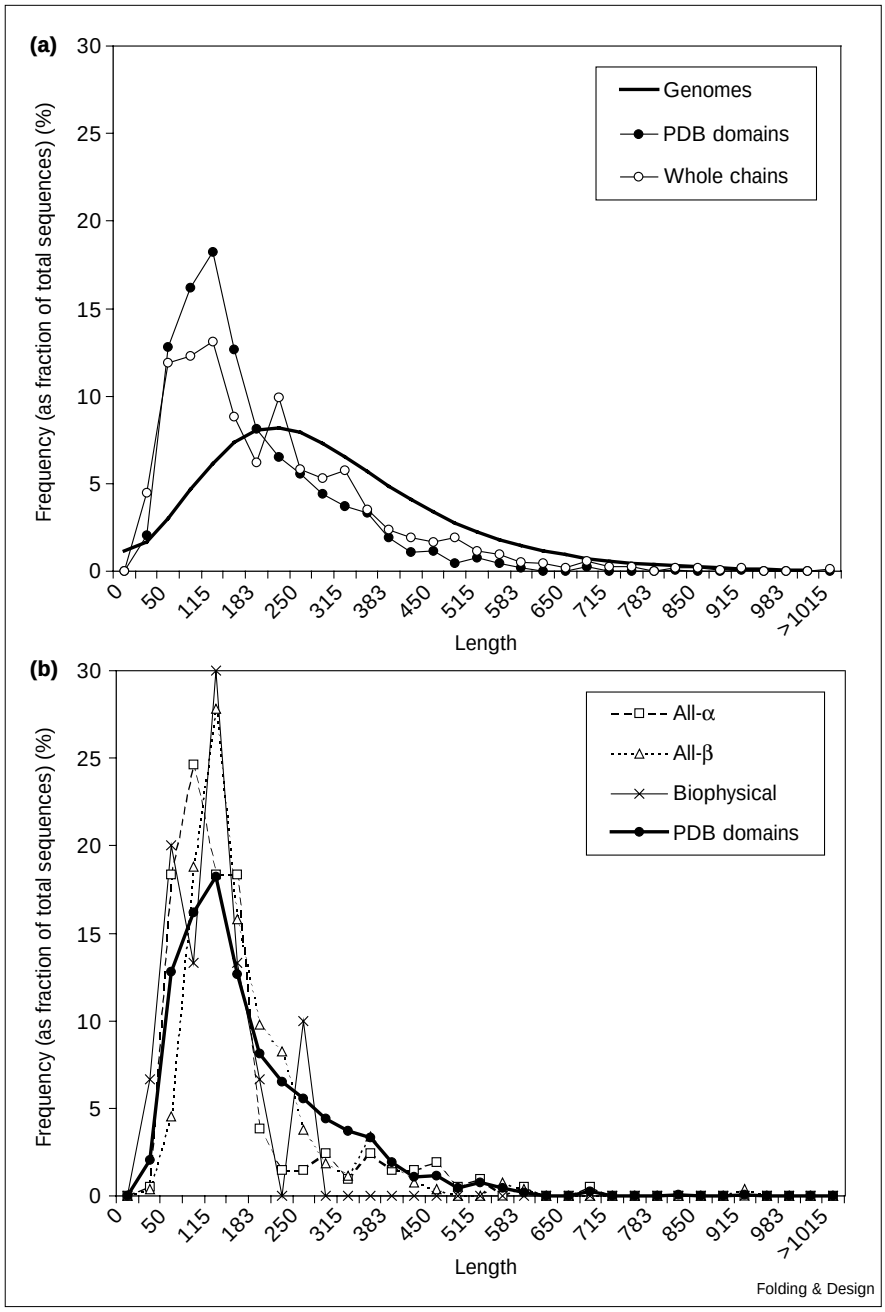
Some basic definitions for comparing composition

In what follows, the composition of a reference is compared with the compositions of other data sets. Usually, this will be the PDB versus whole genomes, and for concreteness this terminology will be used in this section, but other situations will also arise, such as genome versus genome or PDB versus biophysical proteins. It is worthwhile defining a few basic terms relevant to these comparisons. The absolute difference in composition of a particular amino acid i in genome g versus the reference is $D_{abs}(i,g) = C(i,g) - C(i,PS)$, where $C(i,g)$ is the fraction of amino acid i in genome g and $C(i,PS)$ is the analogous fraction in the reference, that is, in PS, the soluble part of the PDB. The relative difference can also be computed: $D_{rel}(i,g) = D_{abs}/C(i,PS)$. Note that both the absolute and relative differences are unitless percentages. This can lead to ambiguity — for example, how does one know whether the statement "the *Escherichia coli* genome has 10% more Ala than the 8% in the PDB" means that the *E. coli* genome has 8.8% or 18% Ala. To avoid this problem, absolute differences will be followed throughout the discussion by 'abs' and relative ones by 'rel'.

If one wants to get a sense of how the composition of a genome differs in an overall sense from the reference, one can compute an average difference in the RMS sense:

Figure 1

The distribution of sequence lengths of structures in the PDB. (a) Length distribution of domains and chains of PDB structures compared with those of genome proteins. The genome distribution is a best-fit, extreme value distribution (see Figure 2). The structure distribution is derived from clustering the PDB as described in the Materials and methods section. One can assess whether the distribution of lengths of genome sequences is significantly different from that of PDB chains in a standard fashion via computation of a χ^2 statistic [96]. Doing this indicates that the differences are statistically significant in a literal sense (i.e. there are enough counts), with the chance of the distributions being identical being less than $1e^{-100}$. (The χ^2 value is 1127 for 31 histogram bins.) (b) The distribution of PDB domain lengths in more detail, breaking down the domains into three subcategories: all-α, all-β, and biophysical.



$$R_X(g) = \sqrt{\left\langle D_X^2(i,g)\right\rangle_i} \qquad (1)$$

In the preceding expression, if X is 'abs', one is evaluating an absolute root-mean square difference R_{abs}, denoted by 'rms abs'. in the text. Likewise, if it is 'rel', one has a relative rms difference, denoted R_{rel} and 'rms rel' in the text. (The averaging here is over all 20 amino acids. It can, of course, instead be performed over secondary structure or genomes to give quantities such as R_{rel}(Ala), but the essential point of distinguishing between relative and absolute compositional differences remains the same.)

Subdivision of the PDB: structural classes and 'biophysical proteins'

Using the original definitions of Levitt and Chothia [29], on a very rough level the domains of known structure can be subdivided into those with all-α, all-β, or mixed struc-

Table 2

Biophysical proteins.

PDB	Select	Length	Class	Name
(a) A list of the proteins used for this study				
1sty	–	137	β	Staph nuclease
1cgp	a:9–137	129	β	CAP
1bgh	–	85	β	Gene V protein
1pht	–	83	β	SH3 domain
1tpf	a:	250	α/β	TIM
1wsy	a:	248	α/β	Trp synthase
8dfr	–	186	α/β	DHFR
2rn2	–	155	α/β	Ribonuclease H
1brs	d:	87	α/β	Barstar
1gbs	–	185	α+β	Hen lyzozyme
119l	–	162	α+β	T4 lysozyme
193l	–	129	α+β	α-Lactabumin
7rsa	–	124	α+β	RNAse A
1brn	l:	108	α+β	Barnase
1fkd	–	107	α+β	FK506
9rnt	–	104	α+β	RNAse T1
1sha	a:	103	α+β	SH2 domain
1ubi	–	76	α+β	Ubiquitin
1cse	i:	63	α+β	CI-2 inhibitor
1igd	–	61	α+β	B1 domain
1mbd	–	153	α	Globin
1hrc	–	105	α	Cytochrome c
2wrp	r:	104	α	Trp repressor
1lli	a:	89	α	Cro repressor
1cop	d:	66	α	Lambda repressor
1rpo	–	61	α	ROP
1myk	a:	47	α	Arc repressor
2zta	a:	31	α	GCN4 zipper
1btl	–	263	M	β-Lactamase
1bpi	–	58	S	BPTI
Average		116		

(b) Composition of the biophysical proteins.

Residue	Hydrophobic/ polar (H/P)	Soluble PDB (PS)	Biophysical proteins (BP)	Relative difference (BP/PS –1)
P	H	4.7%	3.7%	–21%
F	H	4.0%	3.2%	–19%
M	H	2.1%	1.8%	–16%
D	P	6.0%	5.1%	–16%
V	H	7.0%	6.2%	–12%
C	H	1.7%	1.5%	–9%
S	P	6.0%	5.7%	–5%
G	–	7.8%	7.7%	–1%
I	H	5.6%	5.5%	–1%
N	P	4.6%	4.6%	0%
W	H	1.4%	1.5%	1%
T	P	5.8%	6.0%	2%
L	H	8.4%	8.7%	5%
A	–	8.4%	8.8%	6%
Y	–	3.7%	3.9%	6%
H	P	2.2%	2.4%	6%
Q	P	3.7%	4.0%	6%
R	P	4.8%	5.2%	9%
E	P	6.2%	7.0%	13%
K	P	5.9%	7.7%	30%

(a) The 30 'biophysical proteins', which have an average length of 116 amino acids. [AU: please explain the entries in the 'Select' column] (b) The composition of these proteins in comparison to the soluble part of the PDB. Note that the average PDB protein is larger, with an average domain size of 169 residues, and is expected to contain proportionally more hydrophobic residues relative to polar ones than the smaller biophysical proteins. This is indicated in the last column, which shows the change in composition (relative using $D_{rel}(g,i)$ as defined in the text) and the hydrophobic/polar labeling of the amino acids. The net absolute change in hydrophobic residue composition is the total of $D_{abs}(BP,i) = C(BP,i) - C(PS,i)$ summed over all the residues labeled 'H': +2.7%. The net change in hydrophilic residue composition is the same quantity summed over all residues marked with 'P': –2.3%. Note that A, G and Y are left out of these sums. These have a net change of +0.4%, so there is no overall change. [AU: move part (a) to Supplementary material? Include No. of sequences, No. of amino acids and Average length, as for Table 1, instead?]

ture (which includes both α/β and α+β). As shown in Table 1, each group of domains has a different amino acid composition, as expected, differing by 1.3% rms abs (23% rms rel). Making up 18% of the total, the all-α domains tend to be shorter than the overall average (144 residues). The all-β domains are of average length (23% of the total; 161 residues), and the mixed domains, making up the remainder of the total, are (necessarily) slightly longer than the average (181 residues).

In addition, a list of structures corresponding to 'biophysical proteins' was assembled (Table 2). The 30 proteins on this list are supposed to represent the small group of proteins – a subset of those with known structure – on which folding experiments have been done, that is, the proteins that underlie our picture of the folding process. As described in the Materials and methods section, these were chosen in a somewhat subjective fashion, based on literature searches and discussions with colleagues. The biophysical proteins are almost all single-domain and, with a mean length of ~120 residues, are significantly smaller than the average domain in the PDB (Figure 1). As shown in Table 2b, they have mostly moderate or small differences in amino acid composition compared to the average PDB domain (differing by ~0.6% rms abs or 13% rms rel).

One can interpret these differences in terms of the biophysical proteins being considerably smaller than the average PDB domain. This implies that they have a larger surface area relative to buried core and hence more polar and charged residues on the surface relative to hydrophobic ones in the core. As shown in Table 2b, this is largely what is observed. Comparing the composition of the biophysical proteins to those in the PDB, one finds in total that the hydrophobic residues decrease by 2.7% abs and the hydrophilic residues increase by 2.3%. Moreover, five of the six largest decreases are hydrophobes and, likewise, the five largest increases are hydrophiles. The considerably fewer prolines in the biophysical proteins may also be due to the fact that Pro can potentially complicate the folding pathway and so is disfavored by investigators studying folding.

Genomes used and their overall size

The genomes considered in this analysis, listed in Table 3, are the first eight genomes to be completely sequenced. They represent a diverse comparison, being drawn from the three kingdoms of life (Eukarya, Eubacteria and Archaea) and from wildly different external environments (from room temperature and pressure to 85°, 200 atmospheres and from normal pH to highly acidic). They also represent microbes with a wide range of genome sizes and modes of life, from parasite to autotroph. As shown in Figure 2, the distribution of sequence lengths (L) is similar in all eight genomes. It is unimodal with a long tail

Table 3

Genomes and abbreviations used.

Genome	Abbreviation	Size (Mb)	Reference	Website
Haemophilus influenzae	HI	1.83	[87]	http://www.tigr.org/tdb/mdb/hidb/hidb.html
Mycoplasma genitalium	MG	0.58	[88]	http://www.tigr.org/tdb/mdb/mgdb/mgdb.html
Methanococcus jannaschii	MJ	1.66	[89]	http://www.tigr.org/tdb/mdb/mjdb/mjdb.html
Synechocystis sp.	SS	3.57	[90]	http://www.kazusa.or.jp/cyano/cyano.html
Mycoplasma pneumoniae	MP	0.81	[91]	http://www.zmbh.uni-eidelberg.de/M_pneumoniae/MP_Home.html
Saccharomyces cerevisiae	SC	13	[92]	http://genome-www.stanford.edu/Saccharomyces
Helicobacter pylori	HP	1.66	[37]	http://www.tigr.org/tdb/mdb/hpdb/hpdb.html
Escherichia coli	EC	4.60	[93]	http://www.genetics.wisc.edu

and approximately follows an extreme value distribution: $F(L) = 1.25 \exp(-(L-210)/140 - \exp(-(L-210)/140))$. There is no periodicity observed (e.g. for multiples of 125, as suggested by Berman *et al.* (1994) **[AU: which ref is this?]**). The observed fall-off can be rationalized in terms of physical arguments [30].

The average length of a genome-encoded protein sequence is 340 amino acids, appreciably larger than that of the average protein domain, ~170 amino acids, and also larger than an average PDB chain, ~205 amino acids (Figure 1 and Table 4). This average is greatly inflated, however, because of a few extremely long sequences. The most common length for a genome sequence is roughly the size of a single domain (i.e. the mode, ~190). As has been remarked on before [31], yeast has a preponderance of very long protein sequences compared with the bacterial genomes. In particular, about 13% of yeast sequences have a length of more than 833 residues (five PDB domains) compared with the average of 5% in all eight genomes. (This leads to the average yeast sequence being ~470 residues, significantly greater than the genome average of 340.) Interestingly, the mycoplasms (MP and MG) have a relatively high proportion of rather long sequences in their small genomes, indicating that some of these long sequences may be essential. Overall, MJ appears to have the shortest sequences. For yeast there is a distinct spike in the length distribution around 100 residues; this is almost undoubtedly an artifact and reflects the still-not-finalized state of the genome data (see Materials and methods section and [32]).

Overall genome amino acid composition

As shown in Table 5, the genome proteins have some significant differences from the PDB proteins in terms of their overall amino acid composition. The greatest differences are in the amino acids Lys, Ile, Gln and Asn, which are more common in the genome proteins than in the

Figure 2

The distribution of lengths of sequences in eight microbial genomes. An extreme value distribution fit to the observed distribution is shown by the bold line. Note there are some sequences longer than 983 amino acids that are not indicated in the graph. The two-letter abbreviations are defined in Table 3.

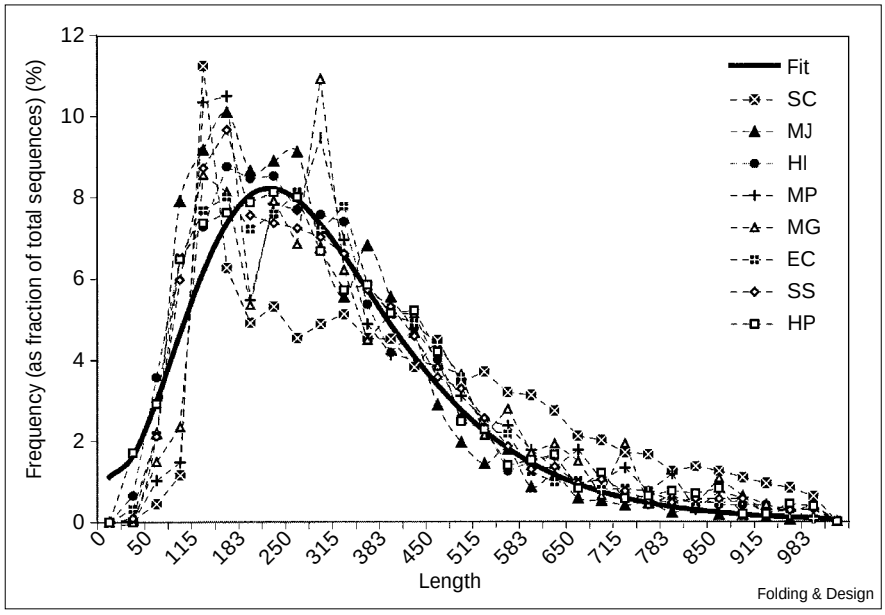


Table 4

Statistics for the lengths of genome-encoded protein sequences.

	PS	Average*	EC	HI	HP	MG	MJ	MP	SC	SS
Average size of a sequence	169	340	317	301	317	364	287	351	466	326
Most common sequence size in histogram (mode)	116	186	249	149	216	283	149	149	116	149
Fraction of sequences > 333 aa (~2 SCOP domains)	9.8%	40%	38%	34%	38%	42%	32%	39%	56%	38%
Fraction of sequences > 833 aa (~5 SCOP domains)	0.1%	4.6%	2.8%	2.4%	3.2%	6.0%	2.1%	5.0%	12.6%	3.9%
Average size of a sequence as a multiple of SCOP domains	2.0	1.9	1.8	1.9	2.1	1.7	2.1	2.8	1.9	
Total number	1135	4290	1680	1577	468	1735	677	6218	3168	
Number of SCOP-sized domains	8017	2982	2955	1006	2939	1404	17119	6099		

*The average over the eight genomes for a row. Note that these simple average values are slightly different from those obtained from integrating the extreme value distribution fit, shown in Figures 1 and 2.

PDB proteins, and Cys and Trp, which are less common. The latter difference may reflect the abundance of Cys and Trp in active sites and binding surfaces, and the prevalence of enzymes and 'binding' proteins in the PDB (e.g. see [33]). Also, the PDB has a clear over-representation of extracellular proteins, which can have disulfides, as opposed to intracellular ones, which cannot. The amino acids that differ most in occurrence between the genome proteins and the PDB also tend to vary the most between genomes. This is especially true for Lys. The amino acids with the greatest similarity in composition to the PDB are Asp, Glu, Thr, Tyr and Val. It is interesting that Asp and Glu are so similar in composition while there are great differences between Gln and Asn.

Table 5

Composition of the genome proteins as compared to the PDB.

	RMS	K	I	C	Q	W	N	F	L	G	A	P	S	R	H	M	E	D	T	Y	V
(a) Absolute composition																					
EC		4.4	6.0	1.2	4.4	1.5	4.0	3.9	10.6	7.4	9.5	4.4	5.8	5.5	2.3	2.8	5.7	5.1	5.4	2.9	7.1
HI		6.3	7.1	1.0	4.6	1.1	4.9	4.5	10.5	6.6	8.2	3.7	5.8	4.5	2.1	2.4	6.5	5.0	5.2	3.1	6.7
SS		4.2	6.3	1.0	5.6	1.6	4.0	4.0	11.4	7.4	8.5	5.1	5.8	5.1	1.9	2.0	6.0	5.0	5.5	2.9	6.7
SC		7.3	6.6	1.3	3.9	1.0	6.1	4.5	9.6	5.0	5.5	4.3	9.0	4.5	2.2	2.1	6.5	5.8	5.9	3.4	5.6
HP		8.9	7.2	1.1	3.7	0.7	5.9	5.4	11.2	5.8	6.8	3.3	6.8	3.5	2.1	2.2	6.9	4.8	4.4	3.7	5.6
MP		8.6	6.6	0.8	5.4	1.2	6.2	5.6	10.3	5.5	6.7	3.5	6.5	3.5	1.8	1.6	5.7	5.0	6.0	3.2	6.5
MG		9.5	8.2	0.8	4.7	1.0	7.5	6.1	10.7	4.6	5.6	3.0	6.6	3.1	1.6	1.5	5.7	4.9	5.4	3.2	6.1
MJ		10.4	10.5	1.3	1.5	0.7	5.3	4.2	9.5	6.3	5.5	3.4	4.5	3.8	1.4	2.2	8.7	5.5	4.0	4.4	6.9
Average		7.5	7.3	1.1	4.2	1.1	5.5	4.8	10.5	6.1	7.0	3.8	6.4	4.2	1.9	2.1	6.5	5.1	5.2	3.3	6.4
SD		2.3	1.4	0.2	1.3	0.3	1.2	0.8	0.7	1.0	1.5	0.7	1.3	0.9	0.3	0.4	1.0	0.3	0.7	0.5	0.6
(b) Difference in composition versus the PDB																					
EC	16	–25	8	–29	19	7	–15	–2	28	–6	13	–5	–3	16	3	28	–7	–14	–7	–22	1
H	17	8	27	–38	24	–21	6	12	26	–15	–2	–20	–2	–6	–7	10	5	–17	–11	–14	–4
SS	20	–29	13	–39	49	9	–13	1	37	–6	1	11	–3	6	–15	–8	–2	–16	–6	–20	–4
SC	21	24	18	–21	5	–27	31	14	15	–36	–34	–7	51	–7	–2	–4	5	–4	0	–8	–20
HP	27	52	29	–34	0	–51	27	36	34	–26	–18	–29	14	–28	–4	2	11	–20	–25	1	–20
MP	28	45	18	–55	44	–17	35	41	24	–29	–20	–25	8	–27	–18	–28	–8	–17	2	–11	–7
MG	36	61	48	–50	27	–32	62	53	28	–41	–33	–36	11	–35	–28	–30	–8	–18	–8	–11	–12
MJ	38	77	88	–23	–61	–49	14	6	14	–19	–35	–28	–25	–20	–35	1	40	–8	–31	20	–2
Average		26	31	–36	13	–23	19	20	26	–22	–16	–17	6	–13	–13	–4	4	–14	–11	–8	–9
RMS		45	39	38	35	31	30	28	27	25	24	23	21	21	18	18	16	15	15	15	11

The table shows (a) the amino acid composition of each genome and then (b) the difference in composition versus the PDB. This latter number is expressed as a relative change $D_{rel}(g,i)$ as defined in the text: $D_{rel}(g, i) = [C(g,i)–C(PS,i)]/C(PS,i)$, where $C(g,i)$ is the genome composition of amino acid i and $C(PS,i)$ is the composition of the corresponding amino acid in the PDB from the PS column in Table 1.

The bottom rows beneath each block give the average, standard deviation (SD), and RMS average of the column above. The column headed 'RMS' gives the RMS average of the amino acid differences in a row, i.e. R_{rel}, as defined in the text. The rows and columns of the table are sorted so that genomes and amino acids with the greatest differences relative to the PDB are in the bottom left-hand corner.

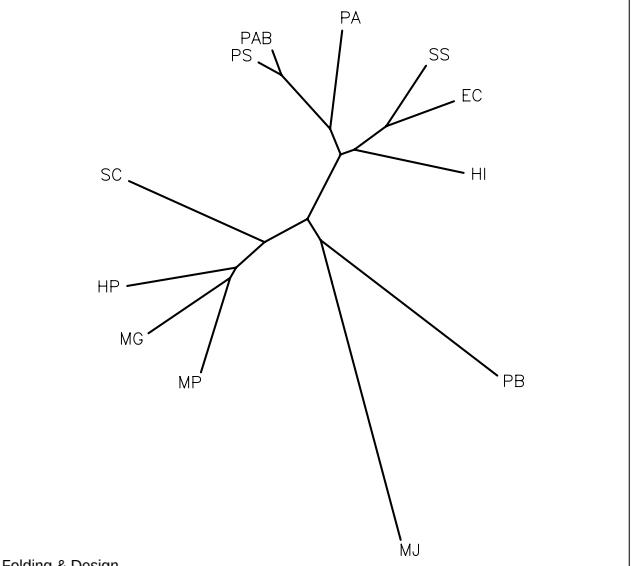
Overall, the genomes with the greatest similarity in composition to the PDB are EC, HI, SS and SC, with the mycoplasms and MJ having the greatest differences. This is perhaps understandable in terms of the great number of *E. coli* and yeast structures in the PDB. Figure 3 shows a cluster tree grouping the genomes based on overall amino acid composition. It is similar in topology to the conventional tree based on 16S ribosomal sequences [34–36]. That is, it groups together the gram-positive bacteria (MP and MG) and the gram-negative bacteria (HI and EC) and positions these two bacterial lineages with the cyanobacteria SS, a roughly equal distance from the eukaryote SC and the archeon MJ. The only problematical organism is HP, which is closer to the mycoplasms in the composition tree. HP is a gram-negative bacterium and should be grouped with EC and HI; however, it has previously been found to be rather problematic in terms of evolutionary classification [37,38]. The overall similarity of the tree in Figure 3 with the conventional tree is notable given the very different properties of the genomes used for defining distance for these trees, that is, overall amino-acid composition versus the specific nucleotide sequence of a single gene.

Each amino acid has a different propensity to confer secondary structure, whether it be α helix, transmembrane helix, or β strand (Table 6 **[AU: tables renumbered from here, please check all table citations carefully.]**). Consequently, the observed differences in amino-acid composition might be expected to give rise to more of one type of secondary structure, for example more helices. This can be tested to some degree through prediction of secondary structure, as discussed below.

An overview of sequence masking

One problem with comparing the overall amino acid composition of the genome proteins with that of the PDB is that this lumps together many distinctly different groups of proteins — membrane proteins, proteins with PDB homologues, completely uncharacterized proteins, and so forth. In this section, an attempt is made to disentangle these different groups. One can think of this process as sequentially applying a number of 'masks' to the genome sequences – first, covering the regions that match a domain in the PDB, then covering the low-complexity regions and transmembrane helices, and finally, short segments between already masked regions are annotated as linkers, connecting domain structures with loops. The part of the genome sequence that remains after all of this consists of structurally uncharacterized domains of soluble proteins. Comparison of these uncharacterized regions with the PDB is the ultimate goal of this analysis, as it most directly addresses the issue of how representative the known structures are of the new and unusual proteins encoded in the genomes.

Figure 3



A cluster tree based on amino-acid composition. This unrooted tree shows the result of clustering the eight microbial genomes on the basis of differences in amino-acid composition. The distance between two genomes *A* and *B* is defined in terms of amino-acid composition through the following formula for Euclidean distance:

$$D_{abs}(AB) = \sqrt{\sum_{i=1}^{20}\left(C(i,A) - C(i,B)\right)^2} \quad (4)$$

where $C(i,g)$ is the composition of the ith amino acid in genome g. Other measures of distance were also tried, in particular the Hellinger distance [97], which is the same as $D_{abs}(AB)$ except for the replacement:

$$C(i,\cdot) \rightarrow \sqrt{C(i,\cdot)} \quad (5)$$

This treats small differences differently. However, it is found that the resulting tree topology is insensitive to the choice of distance metric – providing a test of the robustness of the results.

From the masking process, one gets two numbers: the fraction of the total amino acids in a genome associated with a particular structural feature, and the number of proteins in the genome (i.e. open reading frames — ORFs) that contain this feature. Most genome analyses have tended to focus on the latter value, characterizing, for example, certain fractions of the proteins in the genome as being membrane proteins. This is somewhat deceptive, however, as a given protein can have many different domains and structural features. For instance, a given ORF can simultaneously match a known domain of a soluble protein and also contain a transmembrane helix.

Sequence similarity to known structures

The first step in 'masking' is just to compare the genome proteins with the structures in the PDB. This was done with standard sequence comparison approaches (see the Materials and methods section). As has been found in

Table 6

Experimentally determined local structure propensities.

	Propensity (kcal/mol)		
	TM helix	α Helix	β Strand
A	–1.6	–1.9	0.0
C	–2.0	–1.1	–0.8
D	+9.2	–1.0	+0.9
E	+8.2	–1.2	–0.2
F	–3.7	–1.0	–1.1
G	–1.0	0.0	+1.2
H	+3.0	–1.1	–0.4
I	–3.1	–1.2	–1.3
K	+8.8	–1.5	–0.4
L	–2.8	–1.6	–0.5
M	–3.4	–1.4	–0.9
N	+4.8	–1.0	–0.5
P	+0.2	+3.0	>3.0
Q	+4.1	–1.3	–0.4
R	+12.3	–1.9	–0.4
S	–0.6	–1.1	–0.9
T	–1.2	–0.6	–1.4
V	–2.6	–0.8	–0.9
W	–1.9	–1.1	–1.0
Y	+0.7	–1.2	–1.6

The transmembrane (TM) helix scale gives the energy in kcal/mol for inserting this amino acid into a membrane [78]. It is used here for the identification of membrane proteins. The α helix and β strand propensity scales are also expressed in kcal/mol. Both scales are derived from protein-unfolding experiments [94,95], but similar scales can be determined from doing statistics on solved crystal structures [82].

numerous previous analyses, about one-eighth of the ORFs in the genomes (13%) were homologous (or identical) to sequences corresponding to known structures, and these structure matches involved 9% of the total amino acids in a genome (Table 7). This number ranges considerably, however. Predictably, yeast has the smallest fraction of its genome matched in terms of total number of residues (6%), and MJ has the smallest fraction matched in terms of fraction of sequences (11%). Conversely, HI has the largest fraction of its amino acids matched to known structures (14%), and MG has the largest fraction of its ORFs that have a structural match (19%).

As is to be expected, the segments matching known structures are more similar in composition to the PDB and to each other (18% and 15% rms rel) than to the genomes overall or to structurally uncharacterized regions (Tables 5 and 8, see later). And the average size of a genome region matching a PDB structure (152 residues) is a bit less than the average size of a domain in the known structures (but this number varies between genomes).

Low-complexity regions, transmembrane helices and linkers

Stretches of low complexity sequence are thought not to fold into globular protein structures [39,40]; they may correspond to fibrous or disordered structures. Consequently, it is doubtful whether they will ever be crystallized. After removing the structure matches (as shown in Table 6), about one-quarter of the remaining residues in the genomes are in low-complexity regions. This number varies considerably between genomes, with MJ having the most and HI the fewest (37% versus 15%). Somewhat surprisingly, MG has a high proportion of its minimal genome devoted to these sequences (28%, more than EC or HI), indicating that they must have some essential role. The

Table 7

Overall statistics for occurrence of different structurally characterized regions.

	Average	SD	EC	HI	HP	MG	MJ	MP	SC	SS
Statistics for amino acids										
Total number	775,998		1,358,465	505,279	500,616	170,400	497,968	237,905	2,900,670	1,033,450
Fraction masked by...										
PDB match	8.7%	3.7%	11.1%	13.7%	8.8%	12.9%	7.1%	9.7%	6.2%	9.0%
Low-complexity region	21.7%	6.9%	16.7%	13.9%	22.2%	28.2%	35.1%	24.7%	23.9%	20.5%
TM helix	4.9%	1.4%	7.3%	6.1%	4.8%	3.8%	2.9%	4.5%	5.2%	5.9%
Linker region	5.1%	0.4%	5.3%	4.8%	4.8%	5.0%	5.0%	5.2%	4.6%	5.1%
Fraction remaining uncharacterized	59.7%	8.9%	59.6%	61.5%	59.4%	50.2%	49.9%	55.8%	60.0%	59.6%
Statistics for ORFs										
Total number	2206	1731	4290	1680	1577	468	1735	677	6218	3168
Fraction containing...										
PDB match	12.6%	4.8%	14.1%	16.8%	12.2%	19.2%	11.0%	14.2%	13.5%	13.2%
Low-complexity region	43.0%	12.6%	34.6%	30.6%	43.2%	51.7%	61.3%	49.3%	56.3%	39.6%
TM helix	28.8%	6.6%	34.6%	27.7%	26.9%	26.7%	19.6%	28.1%	35.6%	36.8%
Linker region	51.0%	9.1%	49.0%	46.1%	50.4%	58.8%	55.0%	56.0%	57.3%	52.8%
Fraction containing...										
Uncharacterized region	76.8%	4.4%	75.2%	73.2%	75.4%	74.8%	68.8%	77.8%	84.0%	79.4%
Characterized region	65.5%	13.7%	64.2%	58.6%	65.2%	74.1%	74.9%	70.9%	79.1%	68.3%

Table 8

Composition of structurally uncharacterized regions.

	RMS	K	Q	N	I	C	W	G	A	F	P	L	R	E	S	M	T	V	Y	H	D
(a) Absolute composition																					
EC		4.8	4.8	4.3	5.6	1.2	1.5	6.9	8.7	3.6	4.5	9.8	6.1	6.5	5.7	2.6	5.3	6.6	3.1	2.6	5.8
HI		6.7	5.1	5.2	6.6	1.1	1.1	6.2	7.5	4.2	3.8	10.0	4.9	7.0	5.7	2.2	5.1	6.4	3.3	2.3	5.5
SS		4.6	5.9	4.2	6.0	1.1	1.5	6.9	7.8	3.9	5.2	10.6	5.5	6.6	5.6	1.9	5.3	6.4	3.2	2.2	5.5
SC		7.5	3.9	6.0	6.7	1.4	1.1	4.8	5.2	4.6	4.3	9.6	4.8	6.7	8.0	2.1	5.6	5.7	3.6	2.3	6.1
HP		9.2	3.8	6.0	7.0	1.2	0.7	5.5	6.5	5.2	3.4	10.5	3.9	7.1	6.5	2.2	4.4	5.4	4.0	2.4	5.2
MP		8.7	5.3	6.7	6.5	0.9	1.2	5.3	6.2	5.6	3.6	9.8	3.8	5.8	5.9	1.6	5.7	6.3	3.6	2.1	5.5
MG		9.6	4.7	7.8	8.0	1.0	1.0	4.7	5.3	5.8	3.1	9.9	3.4	5.8	6.3	1.6	5.4	6.0	3.5	1.9	5.3
MJ		9.9	1.7	5.4	9.4	1.5	0.8	6.3	5.4	4.2	3.7	8.7	4.4	8.5	4.4	2.3	4.2	6.9	4.6	1.8	5.9
Average		7.6	4.4	5.7	7.0	1.2	1.1	5.8	6.6	4.6	3.9	9.8	4.6	6.7	6.0	2.1	5.1	6.2	3.6	2.2	5.6
SD		2.1	1.3	1.2	1.2	0.2	0.3	0.9	1.3	0.8	0.7	0.6	0.9	0.9	1.0	0.3	0.5	0.5	0.5	0.3	0.3
(b) Difference in composition versus the PDB																					
EC	15	–19	30	–8	0	–25	3	–12	4	–10	–3	18	28	5	–5	18	–9	–5	–16	17	–3
H	17	14	36	13	18	–33	–22	–21	–10	5	–18	20	2	14	–5	3	–12	–9	–9	3	–8
SS	19	–22	58	–9	8	–31	8	–12	–7	–3	11	27	16	6	–6	–11	–10	–8	–11	–1	–8
SC	20	28	5	30	20	–14	–21	–38	–38	16	–8	15	0	8	33	–2	–5	–19	0	6	2
HP	26	56	3	30	25	–27	–51	–30	–23	31	–27	25	–19	14	9	0	–26	–23	10	7	–13
MP	27	48	43	44	17	–48	–16	–33	–26	42	–24	17	–21	–7	–1	–29	–3	–10	–1	–5	–9
MG	34	64	25	69	43	–41	–31	–40	–36	46	–33	18	–29	–7	5	–25	–8	–15	–3	–15	–11
MJ	32	68	–55	16	70	–10	–42	–19	–35	5	–20	5	–8	37	–26	5	–28	–2	26	–20	–2
Average		30	18	23	25	–29	–21	–26	–22	17	–15	18	–4	9	0	–5	–13	–11	0	–1	–6
RMS		45	37	34	32	31	29	28	26	26	20	19	18	16	16	16	15	13	12	11	8

This table has an identical format to that of Table 5, but here all the statistics are restricted to the uncharacterized regions of the genomes – i.e. the regions corresponding to soluble protein domains with a definite yet currently unknown fold. There are 160 differences reported in this table; 18 of these are greater than the largest difference between all-α and all-β domains (40% rel, see Table 1): K in MJ, MG, MP and HP; Q in MJ, MP and SS; N in MP and MG; I in MJ and MG; C in MP and MG; W in HP and MJ; G in MG; and F in MG and MP.

low-complexity regions are highly variable in composition and, predictably, very different in composition from the PDB (see Table 9 for specific values).

About 5% of the residues in the genomes are in transmembrane helices (Table 7). This number ranges from a high of 7% in EC to a low of 3% in MJ. The number of sequences with at least a single transmembrane element ranges from a high of ~35% in EC, SC and SS to a low of about 20% in MJ with an average of about 28%.

Segments of sequence already accounted for thus far – i.e. PDB matches, low-complexity regions or transmembrane helices — are considered to be 'characterized' regions. The average length of these regions is ~100 residues, and these segments make up ~35% of the total amino acids in a genome. Short sequences between characterized segments are considered to be linkers, loops or coils connecting known structural elements, whether membrane-spanning helices or known globular domains. Over all the genomes, linker regions are consistently about 11 residues in length and constitute ~5% of the total amino acids (Table 6). Compared to the PDB they are especially enriched in Lys and depleted in Ala and Gly (by –29% and –23% rms rel). This latter result is somewhat contrary to expectation, as one tends to think that the small residues, such as Ala and Gly, occur often in flexible loops connecting domains [41,42].

Regions of sequence remaining structurally uncharacterized

After the whole masking process is done, including finding the linkers, one is left with regions of sequence that have not been characterized in a structural sense. These 'uncharacterized regions' presumably fold into soluble, globular protein structures, though some of them could also be part of all-β membrane proteins, such as porins [43]. They provide a suitable comparison for the PDB, which also consists (mostly) of soluble proteins with globular structures.

Uncharacterized regions constitute about ~60% of the amino acids in a genome. Their average size is 186 residues, which is, perhaps not coincidentally, about the size of an average PDB domain. This number is remarkably constant across the genomes with a standard deviation of only ~9% (16) **[AU: is this ref [16]?]**. Interestingly, HI, followed closely by yeast, has the highest fraction of uncharacterized regions (64% and 60%), and MG and MJ have the lowest (50%). The latter value reflects the large number of low-complexity regions in MJ. For yeast there is a large discrepancy in the total number of residues that

Table 9

Difference in composition of various regions versus the PDB.

	Average	SD	EC	HI	HP	MG	MJ	MP	SC	SS
Overall	23%	10%	16%	17%	27%	36%	38%	28%	21%	20%
PDB match	18%	9%	12%	14%	24%	27%	34%	20%	12%	15%
Low-complexity region	36%	13%	32%	33%	39%	50%	52%	40%	42%	35%
TM helix	49%	15%	55%	53%	55%	57%	55%	56%	56%	51%
Linker region	27%	10%	22%	24%	29%	39%	33%	35%	21%	25%
Uncharacterized region	23%	6%	15%	17%	26%	34%	32%	27%	20%	19%

The difference in composition of a specific region of a genome (e.g. linker regions) versus the PDB averaged over all 20 amino acids (in an RMS sense). That is, each value in this table is an $R_{rel}(g)$ value as defined in the text, but now restricted to just comparing the composition of a specific region of the genome. Because of the large number of amino acids involved in all comparisons, all the compositional differences reported here are statistically significant in a literal sense (see text). The number of amino acids compared is listed in Table 7. The smallest number of amino acids compared is for the TM helices in MG: 38% × 170,400 = 6475. Some notes on the compositional differences follow. For the low-complexity regions, the average difference in composition between the genomes is 29% rms rel, with the most variable amino acids being C, H, K, M, Q and W (data not shown). The average difference from the PDB is 36% rms rel, with the genomes being enriched in K, S and I and depleted in C, A, H, Y, W and M. As is necessitated by their definition, the transmembrane regions have a relatively constant composition across the genomes, and relative to the PDB they are depleted in amino acids such as D, E, K, N, Q and R and enriched in A, F, G, I, L, M, V and W.

are uncharacterized versus the few proteins that do not have at least one characterized region (60% versus 29% (= 100% – 71%)). This reflects the large average size of a yeast protein (which can contain multiple structural domains) and highlights the problem, alluded to earlier, of characterizing an entire ORF based on it having a single domain of known structure.

As shown in Table 8, compared to the known structures, the composition of the uncharacterized regions is deficient in Ala, Cys, Gly, Pro and Trp and is enriched in Ile, Lys, Leu, Asn and Gln, with the compositions of Gln and Lys being particularly variable.

Secondary-structure prediction on uncharacterized regions

It is possible to structurally characterize the uncharacterized regions in a rough fashion through prediction of secondary structure. This was done using standard approaches (the GOR program). Surprisingly, despite the differences in amino-acid composition, the overall statistics for secondary structure composition (the number and size of helices and strands) were very similar in all the genomes (Table 10). About 39% of the residues are predicted to be in an α-helical conformation, 17% in a strand conformation and the remainder in a coil conformation. This is markedly more helical than the (predicted) secondary structure composition of the PDB: 31% helical and 21% strand. The difference is consistently observed across all the genomes (standard deviation 2% rms abs), but there are some, relatively small, variations. Yeast has the least helical structure, and HI and HP the most (34% versus 41% and 42%).

How can genomes have such similar secondary structure composition while having such a markedly different amino acid composition (i.e. comparing Tables 8 and 10)? This is analogous to the question: how they can have very different base compositions (AT- or GC-rich) while coding for proteins with similar amino acid composition. To some degree it has do with a 'degeneracy' in the coding of secondary structure propensities and the 'trading off' of residues with equivalent propensities between genomes. This is evident in the similar values calculated for each genome for average helix and strand propensity per residue (i.e. –1.0 and –0.3 kcal/mol, Table 10).

One problem with this analysis of secondary structure composition is that a prediction method is potentially being applied to sequences very different from those it was 'trained' on. That is, the parameters for the prediction programs all derive from the data in the PDB, so the results of running these programs may unduly reflect the biased nature of the PDB and not be indicative of the actual secondary structure in a genome. This problem is particularly acute here in trying to contrast the PDB with the uncharacterized regions in the genomes.

One way of comparing secondary structure without introducing the statistical biases of the PDB is to use the experimentally determined propensities for α and β structure. These propensities were used to identify regions of mostly helical and mostly strand residues – which, hopefully, correspond loosely to all-α and all-β domains — and then statistics were done on the occurrence of these regions. The results (shown in Table 10) indicate many more putative all-α domains in the uncharacterized regions than all-β ones (11% versus 3% of the total residues in the uncharacterized regions). As was observed for the standard secondary structure prediction, yeast has relatively more β structure and HI and HP more α (yeast

Table 10

Predicted secondary structure composition of structurally uncharacterized regions.

	Average	SD	EC	HI	HP	MG	MJ	MP	SC	SS
Total uncharacterized residues	530,488		809,837	310,907	297,265	85,467	248,367	132,692	1,742,937	616,432
Average experimental propensity of these residues										
α propensity	–1.01	0.03	–1.00	–1.02	–1.05	–1.05	–1.01	–1.03	–1.00	–0.96
β propensity	–0.34	0.05	–0.27	–0.33	–0.37	–0.42	–0.36	–0.38	–0.36	–0.26
Fraction of these predicted by GOR to be...										
in coil conformation	45%	2%	44%	43%	42%	44%	43%	45%	49%	46%
in strand conformation	17%	1%	17%	16%	15%	17%	19%	17%	17%	16%
in helical conformation	39%	2%	39%	41%	42%	39%	37%	39%	34%	38%
Fraction of these predicted by experimental propensities to be...										
in all-α domain	11%	3%	11%	12%	16%	11%	8%	10%	9%	8%
in all-β domain	2.7%	2.4%	0.6%	1.0%	2.1%	7.7%	2.4%	3.0%	4.2%	0.6%

The average β-strand and α-helix propensities are derived by computing a weighted average of the propensities in Table 6, using as weighting factors for each residue the fractional composition of it in the uncharacterized regions:

$$\overline{P}(g) = \sum_{i=1}^{20} P(i)C(g,i) \qquad (6)$$

where $P(i)$ is the propensity of amino acid i (from Table 6) and $C(g,i)$ is the composition of amino acid i in genome g (from Table 8).

has 9% all-α and 4% all-β regions as compared to 23% and 1% for HI). This finding of the prevalence of all-β structure in yeast is similar to what was observed in an earlier survey of supersecondary structures in three genomes [22].

Discussion

At this point one has been presented with many statistics, particularly those related to differences in amino acid and secondary structure composition. One is naturally led to ask how significant and meaningful they are. This question can be answered on a number of levels.

Literal statistical significance

First, one can ask whether one has compared enough amino acids for the differences in composition to be significant, compared to the expected random variation. This is statistical significance in a literal sense. It is properly addressed through the calculation of a chi-squared (χ^2) statistic. For purposes of concreteness the following discussion will focus on comparing the amino-acid composition of a particular genome g (e.g. *E. coli*) against the PDB, but the argument is general and can be extended to many of the other statistics presented here. The appropriate null model is that genome g has the same composition as the PDB. One then compares the actual number of counts of each amino acid observed with the number expected, if this were the case, in the calculation of a χ^2 statistic with 19 degrees of freedom:

$$\chi^2 = \sum_i \frac{(O(i)-E(i))^2}{E(i)} = N\sum_{i=1}^{20} \frac{(C(i,g)-C(i,PS))^2}{C(i,PS)} \qquad (2)$$

where the summation is carried out over all 20 amino acids, N is the total number of amino acids in genome g, $O(i) = NC(i,g)$ is the observed number of counts of amino acid i, and $E(i) = NC(i,PS)$ is the expected number of counts, assuming the genome has the same composition as the PDB. Using the definitions given previously, this can be rewritten as:

$$\chi^2 = N\sum_i C(i,PS)D_{rel}(i,g)^2 \qquad (3)$$

Clearly, the χ^2 statistic depends greatly on the number N of amino acids in the genome. For all the compositional differences reported here, N is so large that the χ^2 statistic is highly significant. For instance, in comparing the amino acid composition of the *E. coli* genome to the PDB, the χ^2 statistic is ~30000 (0.022N, for N = 1358465 amino acids). This has a chance (i.e. a p-value) of less than 1e-100 **[AU: please explain the e-value form here]** occurring under the null model. Even if the number of amino acids compared were only 10000, the p-value would still be less than 1e-30. In fact, one only needs compare 2000 amino acids to achieve a significance level of 0.1%. Considering that the composition difference between *E. coli* and the PDB is one of the smallest reported here (only 16% rms rel) and that all the comparisons done here involve many more than 10000 amino acids, one can see that the composition differences reported here are all statistically significant — in the literal sense. This does not mean, though, that they are free of statistical artifacts or systematic biases.

Possible sampling artifacts: clustering and non-uniform composition of PDB

Figure 4

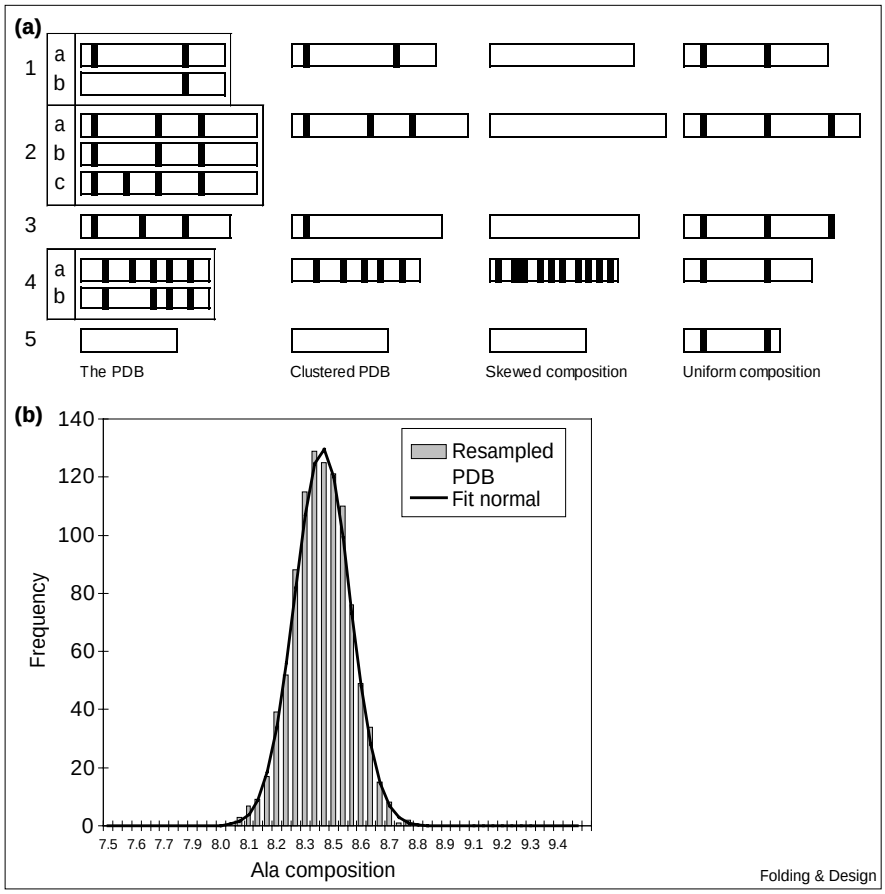


Illustration of possible sampling artifacts affecting the composition of the PDB. The reported composition of the soluble PDB, which acts as a reference here, can be affected by the particularities of the clustering method and by compositional heterogeneity in the PDB. This figure is meant to be read in conjunction with Table 11. (a) A schematic illustrating clustering and sampling bias in the PDB. At the left is a representation of all the sequences in the PDB. The black bands in each sequence are to be read as the occurrence of a particular amino acid. Their frequency (~8.5%) is approximately the same as that of Ala. The transition to the 'clustered PDB' shows how the PDB is clustered into a small number of families. By definition, the difference in amino-acid composition between each member in a cluster is small. Thus, no matter whether one picks representative a, b, or c of the second cluster, the overall composition of the PDB will be nearly the same. This is quantified in the 'SD-CLUST' column [AU: does not correspond to a column in Table 11] of Table 11. The 'skewed composition' and 'uniform composition' columns to the right show two extreme cases of how a given amino acid (e.g. Ala) could be distributed amongst the various cluster representatives. On one extreme, uniform composition, it could be distributed uniformly through each sequence, so that the composition of the PDB would be relatively insensitive to the addition of a new fold (e.g. fold 5). Alternatively, one sequence could be highly compositionally biased, so that it contains much more of a given amino acid than the PDB as a whole. In the extreme case shown here for illustrative purposes, one imagines that all the Ala in the PDB is concentrated in a single sequence (fold 4). Thus, the presence or absence of this sequence would greatly affect the composition of the PDB. One imagines that the actual PDB (clustered PDB) is somewhere between these extremes. The compositional bias of the PDB can be quantified by resampling [44,98,99]. One begins with the 1135 sequences in the soluble PDB (data set PS in Table 11). Then one randomly picks the 1135 sequences with replacement to make a new soluble PDB, called the bootstrap sample PS^*_2. One calculates the composition PS^*_2 and then continues the procedure N times (here $N = 1000$), generating PS^*_2, PS^*_3… PS^*_N and their compositions. The composition of each amino acid i varies between the various bootstrap samples PS^*_j, and one can graph its distribution. (Technically, one should really deal with this in a multivariate sense in terms of multinomial distributions, but it is sufficient here to deal only with projections of these high-dimensional distributions.) This is shown in (b) for Ala. As is evident from the figure, the distribution follows an almost perfectly Gaussian (or binomial) distribution. One can estimate the standard deviation of this distribution and use it to quantify the degree of non-uniformity in composition of the PDB. The larger the width, the greater the bias. For instance, for the uniformly distributed case, the width would be 0. For the case where all the Ala in the PDB is concentrated in one sequence, the width is estimated to be about 8%. (This is derived from making up a specially biased PDB that contains 8.4% Ala but has it all in one sequence.) In both these extreme cases, however, the distribution would no longer be normal. The standard deviations for all 20 amino acids are shown in the SD-RESAMP column of Table 11. The average width is 0.00074 rms abs (1.7% rms rel). Using these standard deviations in conjunction with a 'normal' curve, one can estimate a significance or p-value for each of the observed differences between the genomes and the PDB. Consider the difference between the *E. coli* genome composition and the PDB discussed in the text. This has an average difference of 16% rms rel, about 10 average standard deviations, and so is clearly significant. Finally, using an analogous approach to the one described above, one can also see how compositionally heterogeneous the genomes are and derive compositional distributions for each amino acid in the genome. These result in narrow distributions with even tighter standard deviations than the PDB. Then if one wishes one can use the t-test or a resampling approach to estimate significance for the difference of the means of the two distributions. [AU: could this figure and Table 11 be moved to Supplementary material or this legend significantly reduced?]

One potential statistical artifact could result from the method used to cluster the PDB — that is, different clustering methods could give significantly different PDB compositions (Figure 4). Two statistics that quantify this possible effect are shown in Table 11. For the first statistic, the same basic clustering algorithm used here is

Table 11

Investigation of biases in the composition of the PDB.

	PS	SD-RECLUS	PDB40D	SD-RESAMP
Number of sequences	1135		1217	
Number of amino acids	192313		202415	
Residues:				
A	8.40%	8E-08	8.41%	0.0012
C	1.72%	2E-08	1.63%	0.0006
D	5.91%	4E-08	5.99%	0.0007
E	6.29%	7E-08	6.16%	0.0008
F	3.94%	3E-08	4.02%	0.0006
G	7.79%	5E-08	7.82%	0.0009
H	2.19%	2E-08	2.21%	0.0005
I	5.54%	5E-08	5.55%	0.0007
K	6.02%	1E-07	5.87%	0.0009
L	8.37%	6E-08	8.37%	0.0009
M	2.15%	3E-08	2.21%	0.0004
N	4.57%	6E-08	4.64%	0.0007
P	4.70%	3E-08	4.64%	0.0007
Q	3.73%	5E-08	3.73%	0.0006
R	4.78%	6E-08	4.77%	0.0007
S	5.97%	8E-08	6.00%	0.0009
T	5.87%	7E-08	5.86%	0.0007
V	6.96%	6E-08	6.97%	0.0007
W	1.46%	1E-08	1.47%	0.0004
Y	3.64%	3E-08	3.67%	0.0006

This table gives some indication of how the reported composition of the soluble PDB can be affected by the particularities of the clustering method and by biases in the PDB. The last two columns give measures of how dependent the composition of the soluble PDB (PS) is on the representative structures chosen by the clustering algorithm. As should be evident, the clustering algorithm has little effect on the overall composition calculated for the PDB. The PS-CLUST column (at far right) shows the standard deviation in the composition of PS when different representatives were randomly chosen for each of the 1135 clusters. The average (RMS) deviation is 6E-8 [AU: lowercase "e" used in text, i.e. 6e-8]. The PDB40D column gives the composition of another clustering of the PDB. This is a standard data set available over the web (via http://scop.mrc-lmb.cam.ac.uk [62]). It was prepared in a completely different fashion from the clustering here, so that representative sequences are roughly no more than 40% similar to each other. (There is a length correction via the HSSP equation [74].) Furthermore, unlike PS, PDB40D includes a few membrane proteins. The average (RMS) difference between PS and PDB40 is ~4E-4. The units of PS-SD [AU: what does this mean?] are the same as those of PS composition. Thus, a 8E-8 value for Ala means that with one SD unit, Ala ranges in composition from 0.084 – 0.00000008 to 0.084 + 0.00000008. These very narrow bands imply that the bulk composition does not depend on the choice of cluster representative. The PS-RESAMP column gives a measure of the compositional heterogeneity of the PDB. It gives the standard deviation of the distribution resulting from resampling the PDB, as described in Figure 4. That is, each standard deviation in this table is derived from a Gaussian like that shown for Ala in Figure 4b. [AU: various terms and columns referred to in the footnote (underlined) do not correspond to those in the table, could you please provide consistent terms] [AU: could this table be moved to Supplementary material?]

employed but different cluster representatives are picked at random. This gives essentially no difference to PDB composition (average difference is 8e-8 rms abs). For another statistic, the composition of the PDB was calculated after it had been clustered by a completely different algorithm. This also gave essentially the same PDB composition as reported here, indicating that details of the clustering are not expected to affect the results significantly. (The actual difference in composition between the two clustering methods, 0.03% rms abs and 1.3% rms rel, provides a useful baseline for assessing compositional differences.)

Another potential sampling issue that may affect the quality of the statistics is that the reference PDB data set may be highly heterogeneous in composition. By this, one means that a few sequences with a very biased composition may disproportionately skew the composition of the whole PDB (this situation is illustrated in Figure 4). For instance, imagine if the last two folds added to the PDB were exceedingly Ala-rich. This would imply that the composition of the PDB would be contingent on exactly when the calculation was done (i.e. before or after the new Ala-rich entries were added). The contrasting situation, of course, would be where Ala was distributed uniformly throughout the PDB. Sampling bias is not expected to be as meaningful for the genome compositions, as they are composed of an unbiased and essentially unchanging selection of proteins.

One can measure this lack of uniformity in PDB composition through resampling: making up a 'new PDB' by picking randomly from the original PDB with replacement, calculating a composition of this new PDB and then determining the spread of these compositions. As illustrated in Figure 4, the range of resampled compositions is directly related to how non-uniformly amino acids are spread through the PDB. Figure 4 also shows that the actual resampled compositions are distributed in an approximately Gaussian (or binomial) fashion with a very narrow variance. The narrowness of the spread in resampled compositions is quite small in comparison to the observed differences in composition between the genomes and the PDB (0.00074 rms abs or 1.7% rms rel, which is about one-tenth of the average difference between EC and PDB, 16% rms rel). This implies that the observed differences are statistically significant even when accounting for the variability in the PDB due to its biased and heterogeneous nature.

Meaningful differences and practical significance

Thus far, it has been shown that the reported composition differences are significant in the literal sense of having enough data (counts) and are not unduly affected by the clustering method or the heterogeneity of the PDB. However, this does not really answer the question of whether the observed differences in composition are biologically meaningful or practically useful. The only way to do this is to compare the observed differences to known

differences in composition that have been established to be relevant. One area in which composition differences have been found to be relevant is in the prediction of membrane protein topology. The 'positive-inside' rule based on composition differences of 5–10% abs for Arg + Lys (e.g. 15% inside versus 5% outside) has been found to be quite effective in this regard [45–47]. A number of the composition differences between genomes are comparable to this and involve many more total residues. For instance, the MJ genome has ~4% abs more Arg + Lys than the PDB.

Another helpful yardstick to use in comparing compositions is the difference in composition of all-α and all-β domains. These differ by 23% rms rel on average (1.3% rms abs) and up to 40% rel (24% abs) for particular amino acids. Half of the genomes (HP, MP, MG and MJ) differ more from the PDB than this, both overall and in their uncharacterized regions. Moreover, when comparing the genomes with the PDB, many of the differences in composition for particular amino acids are considerably more than the maximal difference of 40% rel between all-α and all-β domains — such as Lys in MJ, MG, MP and HP (see Table 8 for more examples). Thus, the composition differences between the genomes and the PDB discussed here are comparable to observed variations in composition that are considered significant.

Overall conclusions

This analysis has attempted to determine how representative the known structures are of the proteins encoded in the first eight microbial genomes to be completed, in terms of simple statistics such as sequence length, composition and secondary structure. The sequence lengths of proteins encoded in the genomes, following a long-tailed extreme value distribution, are significantly longer than proteins in the PDB, especially the biophysical proteins (which are quite short). Although the genomes have a roughly similar distribution of transmembrane helices and linker regions, they differ in the relative amount of low-complexity regions and structural homologues they contain. The composition of the genomes, particularly corresponding to the regions of unknown soluble proteins, differs from the PDB in having more Lys, lle, Asn and Gln and less Cys and Trp. Bulk structure prediction applied to the uncharacterized regions of the genomes shows them to have a consistently more helical structure than the PDB, though there are some differences between the genomes, with yeast having more β structure and HI and HP having more α structure.

Furthermore, this analysis has shown that beyond being small, the PDB is also highly biased (with respect to the genomes) in terms of the length and composition of the proteins it contains. Statistical analyses and structure prediction approaches built upon the contents of the PDB need to take these biases into account so that they can be more readily applied to the emerging genome sequence data.

Materials and methods

A relational database of genome sequences and structure assignments

Translated genome sequences were taken from the relevant web sites (Table 3). The genome data are constantly changing and are contingent on the current state of the art in gene finding. The data used in this paper reflect a particular snapshot of this ongoing process. Structures were taken from the PDB via the PDB browser [48,49]. Domain fold and class definitions were taken from SCOP (version 1.35, May 1996) [50–52]. Specific values quoted about the composition of the PDB, such as that it has 5493 total structures and 222 T4 lysozyme structures, refer to the state of the databank when SCOP 1.35 was built. (Since this analysis was performed SCOP 1.37 has been released, which refers to 6497 total structures.) Core structures for each domain were based on refinement of structural alignments [53–56]. The biophysical protein list was constructed in a subjective fashion, based on conversations with colleagues and reading the literature.

Analysis and processing of the data were greatly expedited by the use of a simple relational database, implemented in DBM, Perl5 [57] and mini-SQL (http://Hughes.com.au). This was described in an earlier paper [22] which also contains tables cross-referencing sequence identifiers, structure matches, transmembrane helix positions, and so forth, and cross-tabulation reports giving the occurrence of various patterns. Most of these tables and reports will be made available over the internet (as text tables and via a simple query interface) from the GeneCensus system at http://bioinfo.mbb.yale.edu/genome. The tables are structured in such a way that all the genome features (e.g. location of a transmembrane helix or PDB match) are annotated in a consistent fashion, with thresholds and scoring schemes applied consistently over multiple tools. This attempt at consistency is similar to what has been achieved in other genome annotation systems that aim to integrate multiple tools [58,59].

Sequence comparison

All sequence comparison was done with the FASTA program (version 2.0) with k-tup 1 and an 'e-value' threshold of 0.01 [60,61]. The e-value describes the number of errors per query expected in a single database scan, so a value of 0.01 means that about one out of a hundred cluster linkages will be in error [27,51,62–65]. To extend the sensitivity of the analysis, transitive comparisons were sometimes used [66].

FASTA with k-tup 1 is considered to be one of the most sensitive single-sequence comparison methods, essentially as sensitive as a Smith–Waterman comparison and much faster [62,67]. However, there are a number of other potentially more sensitive methods of comparing sequences to structures that are based on multiple-sequence information – e.g. profiles, hidden Markov models, PSI-BLAST, and motif analysis [68–72]. A number of these were tested and, as expected, they find more homologues for certain folds. The sensitivity improvement is not uniform, however, as these multiple-sequence methods do considerably better matching structures, for which many sequences are known. This creates a subtle bias: one finds more of what one already knows a lot about. This is obviously disadvantageous for a large-scale census where uniform sampling and treatment of the data are more important than sensitivity – i.e. one is more concerned with accuracy in relative rather than absolute numbers. Moreover, cobbling together a census through the use of a disparate collection of tools and patterns creates the problem of devising consistent scores and thresholds. This is particularly acute in the case of manually derived sequence patterns and motifs, as an expert on a particular fold or motif would expect his pattern to find relatively more homologues than a pattern not constructed by an expert. The approach here, apply-

ing the same single-sequence procedure to each fold, circumvents these problems to some degree. Furthermore, it has an added advantage in that it can be performed automatically without manual intervention and, consequently, can easily be scaled up to deal with much larger datasets.

Nevertheless, it is undoubtedly the case that in the future multiple-sequence methods will enable a more complete exploration of the genome. Recent work assigning folds to the MG genome using PSI-BLAST in combination with duplication analysis gives a hint of what is to come [73]. This work shows that by using a very up-to-date database (SCOP 1.37) and a multiple-sequence-based fold-recognition method (PSI-BLAST), a considerably larger percentage of the genome can be assigned to known folds (27% by amino acid and 41% by ORFs in comparison to the values of 13% and 19% in Table 7). These fold assignments were integrated into the sequence-masking procedure done here and this, in turn, shows that 61% of the amino acids in the MG genome can be given a structural annotation (i.e. structure match, transmembrane, low complexity, or linker).

Clustering

The structures in the PDB were clustered into 1135 representative domains. The few membrane protein structures in the PDB were excluded from this clustering so that all the membrane proteins would be identified, in a uniform fashion, by prediction. (This is not expected to be a major factor as, for instance, the yeast genome contains only a single homologue to a known membrane protein structure.) The clustering was similar in spirit to the many previous divisions of the PDB into representative chains (e.g. see [28,62,74,75]). However, a slightly different multiple-linkage algorithm was used [76]. It was designed to be internally consistent with the search method used to identify homologues in the genomes, using the same similarity criteria (a FASTA e-value threshold). The clustering algorithm takes the results of an all-versus-all comparison of the PDB and creates a graph that has one vertex for each sequence and one edge for each similarity score. Each vertex starts out as a cluster of size one. As sequence similarity scores (i.e. e-values) are not commutative, this directed graph is converted to an undirected graph by removing the better scoring edges between pairs. Then each edge is considered in turn, and the two clusters associated by this edge are merged into a single cluster if every member of the first cluster has a good scoring edge between it and every member of the second cluster, and vice versa. The edges are considered in order of decreasing similarity. This has the advantage that close relationships are considered before more distant ones, ensuring that distant relationships are not erroneously used to add a member to a cluster when there exists (for that member) a much closer relationship that would lead to an alternative clustering. Furthermore, this algorithm will produce the same result on the same data set every time; i.e. it is not affected by the order in which the data is traversed. Cluster trees based on distance matrices were built with the Kitsch program, which is part of the Phylip package [77]. Trees were built on the basis of the difference in amino acid composition vectors, as described in the caption to Figure 3. Di-amino acid composition was also used and gave a similar tree.

Transmembrane helix, low-complexity and linker region identification

Transmembrane segments were identified using the GES hydrophobicity scale, shown in Table 6 [78]. The values from the scale for amino acids in a window of size 20 (the typical size of a transmembrane helix) were averaged and then compared against a cutoff of –1 kcal/mol. A value under this cutoff was taken to indicate the existence of a transmembrane helix. Initial hydrophobic stretches corresponding to signal sequences for membrane insertion were excluded. (These have the pattern of a charged residue within the first seven, followed by a stretch of 14 with an average hydrophobicity under the cutoff.) These parameters have been used, tested, and refined on surveys of membrane protein in genomes [12,37,47].

Low-complexity, non-globular sequences were identified with the SEG program [39,40,79] using the standard parameters K(1) = 3.4 and K(2) = 3.75, and a window of length 45. These parameters are the ones used to find 'long' domain-size low-complexity regions. The average size of a low-complexity region found here is ~110 residues. Many of these transmembrane regions are also low-complexity regions (almost half). Taking a conservative approach, it was decided to annotate these doubly identified regions as low complexity, not as transmembrane. This will tend to reduce the total amount of identified transmembrane helices. This is especially true for MJ, which has the largest number of low-complexity regions. SEG is a standard program for the annotation of low-complexity regions and has been integrated into a number of genome analysis systems, in particular the PEDANT system [80]. Characterized regions are considered to be PDB matches, transmembrane helices, or low-complexity regions. Linker regions were considered to be stretches of uncharacterized sequence that connected two characterized regions and were less than 50 residues in length. Linkers also included short sequences at the N or C terminus. Initial Met residues were excluded from the statistics on linker regions.

Secondary structure prediction

Secondary structure prediction was done using the GOR program [1,2,81]. This is a well-established and commonly used method. It is statistically based so that the prediction for a particular residue to be in a given state (say Ala to be in a helix) is directly based on the frequency that this residue occurs in this state in a database of solved structures (taking into account neighbors at ±1, ±2, and so forth). Specifically, version 4 of the GOR program is used here [1]. This bases the prediction for residue *i* on a window from *i*–8 to *i*+8 around *i*, and within this window, the 17 individual residue frequencies (singlets) are combined with the frequencies of all 136 possible di-residue pairs (doublets). The GOR method uses only single-sequence information and because of this achieves lower accuracy (65% versus 71%) than the current 'state-of-the-art' methods that incorporate multiple sequence information [3,82–84]. However, it is not possible to obtain multiple sequence alignments for most of the proteins in each of the genomes. Consequently, bulk predictions of all the proteins in a genome based on multiple-alignment approaches are skewed, in the same sense as discussed above for multiple-sequence-based fold-recognition methods. One gets two distinctly different types of prediction, depending on how many homologues a given protein has. For the bulk prediction done here a simpler single sequence approach was deemed more consistent. This is especially appropriate given that the goal here is to characterize the part of the genome that is least well understood from a structural perspective. Note also that the analysis here is not at all focused on the particular secondary structure prediction for any individual residue. What is of concern is aggregate secondary structure content of whole proteins (and genomes). Prediction of aggregate quantities is expected to be more accurate than the prediction of individual residues [5]. This is evident in the greater success that has been had in predicting overall class of protein fold than just the secondary structure [85,86].

All-α and all-β regions were highlighted in a simple fashion based on the experimental amino-acid propensities (shown in Table 6). A window of length 50 was moved over the sequence, and the average propensity for the amino acids in the window was calculated. If this was less than –0.65 kcal/mol using the β propensity scale, the region was considered all-β. Alternatively, if it was less than –1.25 kcal/mol using the α propensity scale, the region was considered all-α. These thresholds were determined from an analysis of how well they discriminated the known all-alpha and all-beta domains in the PDB.

Acknowledgements

Thanks to Guy Plunket and Mike Cherry for providing information about the genome data; George Weinstock and Steve Norris for providing information on transmembrane folds; Fred Richards, Lynne Regan, and Julie Forman-Kaye for helping with the biophysical protein lists; John Hartigan, Tom Wu,

and Junhyong Kim for help with statistics; Ted Johnson for help with clustering; Hedi Hegyi for carefully reading the manuscript; and Janice Murphy for manuscript preparation.

References

1. Garnier, J., Gibrat, J.F. & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* 266, 540-553.
2. Gibrat, J., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* 198, 425-443.
3. Rost, B. (1996). PHD: predicting one-dimensional protein secondary structure by profile-based neural networks. *Methods Enzymol.* 266, 525-539.
4. Rost, B. & Sander, C. (1992). Jury returns on structure prediction. *Nature* 360, 540. [AU: one page only?]
5. Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232, 584-599.
6. Benner, S.A., Cohen, M.A. & Gerloff, D. (1992). Correct structure prediction? *Nature* 359, 781. [AU: one page only?]
7. Benner, S.A., Gerloff, D.L. & Jenny, T.F. (1994). Predicting protein crystal structures. *Science* 265, 1642-1644.
8. Benner, S.A. & Gerloff, D.L. (1993). Predicting the conformation of proteins. Man versus machine. *FEBS Lett.* 325, 29-33.
9. Scharf, M., *et al.*, & Sander, C. (1994). GeneQuiz: a workbench for sequence analysis. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology.* pp. 348-353, AAAI Press, Menlo Park, California.
10. Casari, G., *et al.*, & [AU: please provide last author's name] (1995). Challenging times for bioinformatics. *Nature* 376, 647-648.
11. Ouzounis, C., Bork, P., Casari, G. & Sander, C. (1995). New protein functions in yeast chromosome VIII. *Protein Sci.* 4, 2424-2428.
12. Arkin, I., Brunger, A. & Engelman, D. (1997). Are there dominant membrane protein families with a given number of helices? *Proteins* 28, 465-466.
13. Goffeau, A., Slonimski, P., Nakai, K. & Risler, J.L. (1993). How many yeast genes code for membrane-spanning proteins? *Yeast* 9, 691-702.
14. Rost, B., Fariselli, P., Casadio, R. & Sander, C. (1995). Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.* 4, 521-533.
15. Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane segments at 95% accuracy. *Protein Sci.* 7, 1704-1718.
16. Boyd, D., Schierle, C. & Beckwith, J. (1998). How many membrane proteins are there? *Protein Sci.* 7, 201-205.
17. Blaisdell, B.E., Campbell, A.M. & Karlin, S. (1996). Similarities and dissimilarities of phage genomes. *Proc. Natl Acad. Sci. USA* 93, 5854-5859.
18. Karlin, S. & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* 11, 283-290.
19. Karlin, S., Burge, C. & Campbell, A.M. (1992). Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res.* 20, 1363-1370.
20. Karlin, S., Mrazek, J. & Campbell, A.M. (1996). Frequent oligonucleotides and peptides of the haemophilus influenzae genome. *Nucleic Acids Res.* 24, 4263-4272.
21. Gerstein, M. & Levitt, M. (1997). A structural census of the current population of protein sequences. *Proc. Natl Acad. Sci. USA* 94, 11911-11916.
22. Gerstein, M. (1997). A structural census of genomes: comparing eukaryotic, bacterial and archaeal genomes in terms of protein structure. *J. Mol. Biol.* 274, 562-576.
23. Gerstein, M. & Hegyi, H. (1998). Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.*, in press. [AU: vol no? any update?]
24. Gerstein, M. (1998). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins*, in press. [AU: vol no? any update?]
25. Bryant, S.H. & Altschul, S.F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* 5, 236-244.
26. Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. (1994). Issues in searching molecular sequence databases. *Nat. Genet.* 6, 119-129.
27. Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA* 95, 5913-5920.
28. Brenner, S.E., Chothia, C. & Hubbard, T.J. (1997). Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* 7, 369-376.
29. Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature* 261, 552-558.
30. Panchenko, A.R., Luthey-Schulten, Z. & Wolynes, P.G. (1996). Foldons, protein structural modules, and exons. *Proc. Natl Acad. Sci. USA* 93, 2008-2013.
31. Netzer, W.J. & Hartl, F.U. (1997). Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature* 388, 343-349.
32. Das, S., *et al.*, & [AU: please provide last author's name] (1997). Biology's new Rosetta stone. *Nature* 385, 29-30.
33. Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* 2, 101-113.
34. Olsen, G.J., Woese, C.R. & [AU: please provide last author's name] (1994).[AU: please provide title]. *J. Bacteriol.* 176, 1-6.
35. Koonin, E.V., Mushegian, A.R. & Rudd, K.E. (1996). Sequencing and analysis of bacterial genomes. *Curr. Biol.* 6, 404-416.
36. Lansig, P., Lansing, P., Harley, J.P. & Klein, D.A. (1996). *Microbiology 3rd Ed.* [AU: please provide place and name of publisher]
37. Tomb, J.-F., *et al.*, & [AU: please provide last author's name] (1997). The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature* 388, 539-547.
38. Doolittle, R.F. (1997). A bug with excess gastric acidity. *Nature* 388, 515-516.
39. Wootton, J.C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17, 149-163.
40. Wootton, J.C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 266, 554-571.
41. Gerstein, M., Lesk, A.M., Baker, E.N., Anderson, B., Norris, G. & Chothia, C. (1993). Domain closure in lactoferrin: two hinges produce a see-saw motion between alternative close-packed interfaces. *J. Mol. Biol.* 234, 357-372.
42. Gerstein, M. & Krebs, W. (1998). A database of macromolecular movements. *Nucleic Acids Res.* 26, 4280-4290.
43. Weiss, M.S., Abele, U., Weckesser, J., Welte, W., Schiltz, E. & Schulz, G.E. (1991). Molecular architecture and electrostatic properties of a bacterial porin. *Science* 254, 1627-1630.
44. Efron, B. & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science* 253, 390-395.
45. von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* 225, 487-494.
46. von Heijne, G. (1994). Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.* 23, 167-192.
47. Wallin, E. & von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* 7, 1029-1038.
48. Abola, E., Sussman, J., Prilusky, J. & Manning, N. (1997). Protein Data Bank archives of three-dimensional macromolecular structures. *Methods Enzymol.* 277, 556-571.
49. Stampf, D.R., Felder, C.E. & Sussman, J.L. (1995). PDBbrowse – a graphics interface to the Brookhaven Protein Data Bank. *Nature* 374, 572-574.
50. Murzin, A., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
51. Brenner, S., Hubbard, T., Murzin, A. & Chothia, C. (1995). Gene duplication in H. influenzae. *Nature* 378, 140. [AU: one page only?]
52. Hubbard, T.J.P., Murzin, A.G., Brenner, S.E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucleic Acids Res.* 25, 236-239.
53. Altman, R. & Gerstein, M. (1994). Finding an average core structure: application to the globins. In *Proceedings of the Second International Conferene on Intelligent Systems in Molecular Biology.* pp. 19-27, AAAI Press, Menlo Park, CA.
54. Gerstein, M. & Altman, R. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* 251, 161-175.
55. Gerstein, M. & Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures,.In *Proc. Fourth Int. Conf. on Intell. Sys. Mol. Biol.* pp. 59-67, AAAI Press, Menlo Park, CA.
56. Gerstein, M. & Levitt, M. (1998). Comprehensive assessment of

automatic structural alignment against a manual standard, the Scop classification of proteins. *Protein Sci.* 7, 445-456.
57. Wall, L., Christiansen, D. & Schwartz, R. (1996). *Programming Perl*. O'Reilly and Associates, Sebastapol, CA.
58. Gaasterland, T. & Sensen, C.W. (1996). Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie* 78, 302-310.
59. Medigue, C., Moszer, I., Viari, A. & Danchin, A. (1995). Analysis of a *Bacillus subtilis* genome fragment using a co-operative computer system prototype. *Gene* 165, GC37-51. [AU: is 'GC' part of page numbering?]
60. Lipman, D.J. & Pearson, W.R. (1985). Rapid and sensitive protein similarity searches. *Science* 227, 1435-1441.
61. Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA* 85, 2444-2448.
62. Brenner, S., Chothia, C. & Hubbard, T. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA* 95, 6073-6078.
63. Pearson, W.R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* 276, 71-84.
64. Pearson, W.R. (1996). Effective protein sequence comparison. *Methods Enzymol.* 266, 227-259.
65. Pearson, W.R. (1997). Identifying distantly related protein sequences. *Comput. Appl. Biosci.* 13, 325-332.
66. Gerstein, M (1998). Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics*, in press. [AU: vol no? any update?]
67. Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195-197.
68. Bowie, J.U. & Eisenberg, D. (1993). Inverted protein structure prediction. *Curr. Opin. Struct. Biol.* 3, 437-444.
69. Tatusov, R.L., Altschul, S.F. & Koonin, E.V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci. USA* 91, 12091-12095.
70. Eddy, S.R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* 6, 361-365.
71. Dubchak, I., Muchnik, I., Holbrook, S.R. & Kim, S.H. (1995). Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl Acad. Sci. USA* 92, 8700-8704.
72. Altschul, S.F., *et al.*, & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.
73. Teichmann, S., Park, J. & Chothia, C. (1998). Structural assignments to the proteins of Mycoplasma genitalium show that they have been formed by extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA*, in press. [AU: vol no? any update?]
74. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci.* 1, 409-417.
75. Boberg, J., Salakoski, T. & Vihinen, M. (1992). Selection of a representative set of structures from Brookhaven Protein Data Bank. *Proteins* 14, 265-276.
76. Kaufman, L. & Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.
77. Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle.
78. Engelman, D.M., Steitz, T.A. & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophysical Chem.* 15, 321-353.
79. Wootton, J.C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* 18, 269-285.
80. Frishman, D. & Mewes, H.-W. (1997). PEDANTic genome analysis. *Trends Genet.*13, 415-416.
81. Garnier, J., Osguthorpe, D. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120, 97-120.
82. King, R.D. & Sternberg, M.J.E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* 5, 2298-2310.
83. Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.* 5, 1704-1718.
84. Salamov, A. & Solovyev, V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* 247, 11-15.
85. Dubchak, I., Holbrook, S.R. & Kim, S.H. (1993). Prediction of protein folding class from amino acid composition. *Proteins* 16, 79-91.
86. Metfessel, B.A., Saurugger, P.N., Connelly, D.P. & Rich, S.S. (1993). Cross-validation of protein structural class prediction using statistical clustering and neural networks. *Protein Sci.* 2, 1171-1182.
87. Fleischmann, R.D., *et al.*, & [AU: please provide last author's name] (1995). Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science* 269, 496-512.
88. Fraser, C.M., *et al.*, & [AU: please provide last author's name] (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* 270, 397-403.
89. Bult, C.J., *et al.*, & [AU: please provide last author's name] (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273, 1058-1073.
90. Kaneko, T., *et al.*, & [AU: please provide last author's name] (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3, 109-136.
91. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C. & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 24, 4420-4449.
92. Goffeau, A., *et al.*, & [AU: please provide last author's name] (1997). The yeast genome directory. *Nature* 387 (suppl), 5-105.
93. Blattner, F.R., *et al.*, & [AU: please provide last author's name] (1997). The complete genome sequence of *Escherichia coli K-12*. *Science* 277, 1453-1462.
94. Chakrabartty, A., Kortemme, T. & Baldwin, R.L. (1994). Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci.* 3, 843-852.
95. Smith, C.K., Withka, J.M. & Regan, L. (1994). A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry* 33, 5510-5517.
96. Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1992). *Numerical Recipes in C*. Cambridge University Press, Cambridge.
97. Amari, S. (1982). [AU: please provide title]. *Ann. Stat.* 10, 357-387.
98. Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* 1, 54-77.
99. Simon, J. (1993). *Resampling: The New Statistics*. Wadsworth, Boston.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad – for further information, see the explanation on the contents pages.