Phylogenetic differences in content and intensity of periodic proteins

Derek Gatherer[1] and Neil R. McEwan[2]

[1]*MRC Virology Unit, Institute of Virology, University of Glasgow, Church St., Glasgow G11 5JR*
[2]*Rowett Research Institute, Greenburn Road, Bucksburn, Aberdeen AB21 9SB*

Correspondence to:

Derek Gatherer
MRC Virology Unit,
Institute of Virology,
University of Glasgow,
Church St.,
Glasgow
G11 5JR
UK

Tel: +44 141 330 6268
Fax: +44 141 337 2236
Email: d.gatherer@vir.gla.ac.uk

Short title: Protein periodicity: a comparative study

Key words: Proteins, periodicity, bioinformatics, Perl, archaea, evolution, genomics, proteome

Abstract (247 words): Many proteins exhibit sequence periodicity, often correlated with a visible structural periodicity. The statistical significance of such periodicity can be assessed by means of a chi-square-based test, with significance thresholds being calculated from shuffled sequences. Comparison of the complete proteomes of 45 species reveals striking differences in the proportion of periodic proteins and the intensity of the most significant periodicities. Eukaryotes tend to have a higher proportion of periodic proteins than eubacteria, which in turn tend to have more than archaea. The intensity of periodicity in the most periodic proteins is also greatest in eukaryotes. By contrast, the relatively small group of periodic proteins in archaea also tend to be weakly periodic compared to those of eukaryotes and eubacteria. Exceptions to this general rule are found in those prokaryotes with multicellular life-cycle phases, e.g. *Methanosarcina sps*. or *Anabaena sps.*, which have more periodicities than prokaryotes in general, and in unicellular eukaryotes, which have fewer than multicellular eukaryotes. The distribution of significantly periodic proteins in eukaryotes is over a wide range of period lengths, whereas prokaryotic proteins typically have a more limited set of period lengths. This is further investigated by repeating the analysis on the NRL-3D database of proteins of solved structure. Some short range periodicities are explicable in terms of basic secondary structure, e.g. alpha helices, while middle range periodicities are frequently found to consist of known short Pfam domains, e.g. leucine-rich repeats, tetratricopeptides or armadillo domains. However, not all can be explained in this way.

Introduction:

Almost as soon as protein sequences began to be determined, it was observed that many proteins have a tendency to periodicity in their sequences (Eck and Dayhoff 1966; Zimmerman et al. 1968). The most extreme examples of periodicity are proteins with invariant, or near invariant, tandem repeats, e.g. polyubiquitin genes in eukaryotes, encoding direct repeats of 76 amino acids. Most periodicities, however, are subtler and frequently not visible to the naked eye, and are termed "cryptic periodicities" (Gatherer and McEwan 2003), or "latent periodicities" (Korotkova et al. 1999; Laskin et al. 2003). Over the years, debate has occurred concerning the extent of such periodicity, its origins and functional significance. With the increasing availability of solved protein structures, it has become apparent that many sequence periodicities reflect structural periodicities. These range from repeats of large domains down to the common short-range periodicity at n= 7 due to alpha helices (Gruber and Lupas 2003; McLachlan and Stewart 1976) and which may simply indicate a protein rich in coiled-coil regions.

Controversy soon arose concerning the origins of such periodicity. Some (Barker and Dayhoff 1977; Ivanov and Ivanov 1980; Ohno 1984; Ohno 1988; Ycas 1976) viewed periodicity as a coincidental phenomenon, deriving from the hypothesized origins of proteins from more literally repetitive concatenations of sequences encoding small oligopeptides. Such concatenations were proposed to have occurred at the dawn of cellular life, in what White & Jacobs (White and Jacobs 1993) referred to as the "protein synthetic big bang". An examination of the distribution of protein sequence lengths by Trifonov (Trifonov 1985) led to a similar conclusion. Others emphasized functional explanations in terms of secondary structure (Eisenberg et al. 1984; Zhurkin 1981). These may be characterised as broadly neutralist and selectionist theories, respectively, or as referred to by White & Jacobs (White and Jacobs 1993), the "starter set hypothesis" and "random origins hypothesis". Under the starter set hypothesis, following the "protein synthetic big bang" literally repetitive proteins in early cellular organisms accumulated

mutations, thus weakening periodicity over evolutionary time. It was therefore predicted that residual weak "cryptic" or "latent" periodicity should be commonly found in proteins, and that this periodicity need not have any explanation in terms of natural selection on the current or past function of the protein. Pursuing the "protein synthetic big bang" metaphor, we refer to this as the "periodicity background radiation", or the "aftersound" (Laskin et al. 2003). By contrast, the random origins hypothesis claimed that periodicity arose out of non-repetitive sequence where consistent selection pressure for a repetitive structure was applied. In such a case, periodicity in those proteins would strengthen over evolutionary time, it would be much rarer, and it would almost always have some functional explanation. There would therefore be no periodicity background radiation effect.

The original debate was conducted in the context of a model of genome evolution that essentially involved only point mutation and simple insertion/deletion events. This model has shifted considerably in the last two decades, with the discovery that genomic DNA is subject to many mechanisms of rearrangement and amplification, referred to by Dover (Dover 2002) as "mechanisms of DNA turnover", and the realisation that such events may have consequences for protein function and ultimately even human molecular disease (Ashley and Warren 1995). Horizontal transmission has also become known as an important factor in the acquisition of novel genetic function, particularly in prokaryotes. It is thus not necessary to require only a single "big bang" followed by some billions of years of simple mutation and selection, but also possible to envisage several "mini-bangs" during the intervening period caused by amplification events, replication slippage, retrotransposition, mobile element-mediated duplication etc. However, if there really were an original "big bang", the periodicity background radiation effect would still be predicted, regardless of any subsequent "mini-bangs". More recent events would then presumably be seen as stronger periodicities, overlaid on a general background of cryptic periodicity.

An early survey of periodicity in 38 protein sequences from a wide variety of organisms was by Ivanov & Ivanov (Ivanov and Ivanov 1980), who concluded that it was widespread within their data set. Specifically in bacteria, Vaara (Vaara 1992) identified eight proteins with periodicity at $n = 6$. By contrast, White & Jacobs (White and Jacobs 1993) used a Run Test to identify proteins with composition deviating from randomness, and were able to find evidence of non-random distribution in no more than 10% of their set of 1789 unrelated sequences. Since periodic sequences are non-random in composition, a Run Test might be expected to identify them (as well as identifying other types of non-random sequence), and this may therefore be taken as some evidence against widespread periodicity. Recently, with the availability of much larger sets of predicted protein sequences, a variety of approaches have been used, notably informational entropy (Korotkova et al. 1999), and oligopeptide word frequencies (Katti et al. 2000). All of these have detected a range of periodicities in a wide variety of data sets. Korotkova *et al.* (Korotkova et al. 1999) estimated that 10% of proteins in SwissProt were periodic, matching the figure of White & Jacobs (White and Jacobs 1993).

The first examination of periodicity in a complete predicted protein set ("proteome") of a single species was by Gatherer & McEwan (Gatherer and McEwan 2003). The *E. coli* proteome has no periodicity background radiation, contrary to the prediction of the neutralist starter set hypothesis. The number of proteins judged periodic at the 5% and 1% significance levels is similar in both the real *E. coli* proteome and a randomly generated proteome with identical amino acid composition. Nevertheless, at the 0.1% significance level, the *E. coli* proteome has several fold more periodic proteins than random sequence, thus revealing the existence of a small core of periodic proteins in a largely non-periodic proteome. This does not necessarily refute the neutralist theory, since it may be that the passage of evolutionary time has been long enough to erase any traces of ancestral tandem repeat origin in the majority of proteins, i.e. the background radiation has simply faded away. However, it does imply that the issue may be undecidable simply by reference to modern proteins, and that on balance the selectionist theory is more likely.

In order to determine if these conclusions previously obtained from *E. coli* are generally applicable, we examined 45 complete proteomes, including an updated version of the *E. coli* proteome. Species were chosen to include representatives of all three superkingdoms ("superkingdom" is used as the most fundamental division of the cellular tree of life, following the NCBI Taxonomy nomenclature - http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root). The extent of periodicity, the strength of periodicity in the most periodic proteins, and the distribution of significant periodicities over different period lengths, were all examined.

Methods:

45 complete proteomes were downloaded on the 19th December 2003, from the EBI proteomes page (http://www.ebi.ac.uk/proteome). Here, the word 'proteome' is used as shorthand for the complete predicted protein set of each genome. Where available, the non-redundant TrEMBL and Swiss-Prot sets were obtained. These were converted into the 6-letter Dayhoff alphabet (Stanfel 1996), using the OddCodes.pm module from BioPerl (http://www.bioperl.org). The Dayhoff alphabet clusters the amino acids as follows into 6 groups: C; AGPST; DENQ; HKR; ILMV; FWY.

A chi-square test compares the position-specific occurrence of each of the six Dayhoff residue categories in the protein with its expected occurrence under a null hypothesis of equal distribution. Yates correction is applied wherever the expected values are less than 5 and the difference between observed and expected is greater than 0.5. Significance is assessed by automated reference to chi-square tables. For degrees of freedom greater than 30, the Fisher-Yates approximation is used to generate a z-score. Periodicity was measured from n=2 to n=100. Fuller details are given in our previous paper (Gatherer and McEwan 2003).

The z-score thresholds for 5%, 1% and 0.1% significance were determined empirically for each proteome, by shuffling each protein and repeating the chi-square test. This differs from the procedure previously adopted (Gatherer and McEwan 2003), where a randomly generated proteome, with the same compositional content as the *E. coli* proteome, was used as a negative control. The shuffling process provides a better negative control, since it removes any potential artefacts due to differing protein length distribution, or compositional content, of each proteome (which can occur, data not shown). As an additional negative control, a randomly generated proteome was scanned for periodicities, then shuffled and re-examined. The proportion of periodicities in both shuffled and non-shuffled random sequences is essentially identical, thus confirming that shuffling does not introduce any additional artefacts (data not shown). Two further differences between the method reported here and our original method (Gatherer and McEwan 2003), are that periodicity is here measured from n=2 to n=100, rather than only up to n=50, and also that only a single alphabet is used instead of nine alternative alphabets. The Dayhoff alphabet is chosen from the original nine because it reflects likely substitution patterns in evolution. This is relevant given the original "big bang" hypothesis of descent of periodicities from ancestral tandem repeats.

In order to analyse the potential relationship between periodic sequence and periodic structure, NRL-3D (Pattabiraman et al. 1990), a database comprising the sequences of proteins available in the Protein Data Bank (PDB) of solved 3-D structures (Berman et al. 2000), was also analysed in the same way. Pfam Hidden Markov Model-defined domains (Bateman et al. 2004) were determined by reference to the Pfam website (http://www.sanger.ac.uk/Pfam).

The overall distribution of periodicities in each species, the "periodicity profile", was determined by subtracting the number of periodicities at the 1% significance level, for each value of n, from the corresponding number in the shuffled proteome. This corrects for the slight tendency of shuffled sequence to have more periodicities at low values of n (data not shown).

The full list of periodic proteins for each of the 45 species and Perl scripts for periodicity analysis are freely available from the authors (mailto:d.gatherer@vir.gla.ac.uk).

Results and Discussion

1. Extent of periodicity in 45 proteomes

The summary of the periodic proportions of each proteome at the 5%, 1% and 0.1% significance levels are shown in Figures 1, 2 and 3, respectively. [Place here Figure 1, Figure 2 Figure 3]. It should be remembered that there is no qualitative gap between periodic and non-periodic sequences. Inevitably some false positives will arise by random chance with probability equal to the significance level, and likewise some proteins in which a periodicity is functionally genuine, but statistically weak, will be missed.

The overall pattern for most species is indicative of a core of periodic proteins within a non-periodic majority, as previously seen in *E. coli* (Gatherer and McEwan 2003). Nevertheless, it is clear that this periodic core is not the same size in all species, and that some species seem to lack it altogether. Eukaryotes (coloured black in Figures 1-3) tend to be more periodicity-rich than the two prokaryote superkingdoms, and archaea (coloured white in Figures 1-3) tend to be relatively periodicity-poor. This effect becomes more pronounced as the significance level decreases, but a Mann-Whitney U-test shows that it is statistically significant ($p<0.05$) at all three periodicity thresholds displayed. However, even in the most periodicity-rich proteome it should be noted that in excess of 91% of proteins are not significantly periodic, even at the 5% level. It is therefore clear that the most radical prediction of the starter set theory of protein periodicity, that periodicity should be found in a majority of proteins, thus constituting a periodicity background radiation, is not supported by evidence from any of the 45 species examined. This corroborates the evidence of White & Jacobs (White and Jacobs 1993) that only 10% of a set of nearly two thousand unrelated protein sequences from various species were distinguishable from random sequence in a Run Test, which would also be expected to detect periodic proteins. A slightly higher estimate of periodicity, at 30% for

human genes in the EMBL database, was previously obtained by Korotkov *et al*. (Korotkov et al. 1997). However, that analysis was done on DNA rather than protein sequences, and using a different algorithm.

The human proteome has the highest percentage of periodic proteins at all three significance levels (Figures 1-3). The mouse proteome takes second place, except at the 5% level where it is the third most periodicity-rich (Figure 1). At the opposite extreme, an archaeal species is the most periodicity-poor at all three significance levels (Figures 1-3). At the 0.1% level, 7 of the 8 most periodicity-poor proteomes are archaeal (Figure 3). However, there are exceptions to this general tendency. *Methanosarcina acetivorans*, despite being archaeal, is the third most periodicity-rich proteome at the 1% and 0.1% significance levels, and the fifth richest at the 5% level (Figures 1-3). Furthermore, a small number of eukaryotes tend to be periodicity-poor. For instance, at both the 5% and 1% significance levels (Figures 1 & 2) the eukaryotic species *Guillardia theta*, a cryptomonad, and *Encephalitozoon cuniculi*, a microsporidium, are found near the bottom of the table. It is notable that those eukaryotes that are relatively periodicity-poor are unicellular. Conversely, *Methanosarcina mazei* and *M. acetivorans* have proportions of periodic proteins comparable to those of multicellular eukaryotes. These two species are part of the family Methanosarcinaceae, the only known archaeal family to form multicellular structures (Galagan et al. 2002). It is also notable that the eubacterial species with the highest proportion of significantly periodic proteins at the 1% and 0.1% levels is *Anabaena*, being in both cases the 5th most periodic species out of 45 (Figs 2 & 3), and which may also adopt a multicellular form under certain circumstances (Golden and Yoon 2003)

There are also similarities and dissimilarities in the periodicity content of related species. For instance, the human and mouse proteomes are close together at all three significance levels (Figures 1-3). By contrast, three members of the archaeal genus *Pyrococcus* exhibit some divergence. *P. furiosus* and *P. horikoshii* are among the most periodicity-poor species at the 5% and 1% levels (Figures 1 & 2), whereas *P. abyssi* is relatively

periodicity-rich. Similarly, *Thermoplasma volcanicum* has far more periodicities than *Thermoplasma acidophilum* at the 0.1% level (Figure 3). In fact, the latter has no convincing periodicities at all at the 0.1% level.

2. Intensity of maximum periodicities

In shuffled proteomes, periodicity scores rarely exceed z=4. A real periodic protein may have a z-score approaching 100. However, such intense periodicities are not found in all proteomes. Figure 4 shows the maximum z-score found in each of the 45 species. [Place here Figure 4] The tendency for eukaryotes, and especially multicellular eukaryotes, to have proportionally more periodic proteins (Figures 1-3), is seen to be mirrored by their tendency to have stronger z-scores (Figure 4). The Mann-Whitney U-test confirms that this is statistically significant (p<0.01). The top 6 species in terms of their highest z-score are all eukaryote, 5 of which are multicellular. Conversely, 6 of the bottom 9 are archaeal.

On further examination, such very high z-scores are seen to be always the products of long (i.e. high values of n), near perfect repeats. Examples occur in all three superkingdoms, although in archaea they are virtually confined to the genus *Methanosarcina*. However, in multicellular eukaryotes, such long periodicities tend to be longer and more perfectly repetitive than in eubacteria or archaea. For example, the human apolipoprotein A precursor (APOA_HUMAN) has 32 near-identical copies of a 114-mer Kringle domain (Pfam PF00051), producing a z-score of 94.9. Nothing matches this in any prokaryote. The nearest equivalent is the hypothetical protein Q9I2M3 in *Pseudomonas aeruginosa*, which has 18 slightly variable copies of an 82-mer Pfam-B_29 element, giving a z-score of 38.1. The most periodic archaeal protein is found in *Methanosarcina acetivorans* hypothetical protein MA3293 (Q8TKV1), which has 12 variable copies of a 48-mer containing a 34-mer tetratricopeptide (Pfam PF00515), giving a z-score of 31.9.

By contrast, in the cryptomonad unicellular eukaryote *Guillardia theta*, the maximum periodicity is barely above that found in shuffled sequence. This is also true for the four archaeal species, *Thermoplasma acidophlium*, *Thermosplama volcanicum*, *Methanopyrus kandleri* and *Pyrococcus furiosus*. This also serves to illustrate how the strength of the maximum periodicity and the number of significant periodicities are not necessarily related. *Thermoplasma volcanicum* has 0.6% of its proteins significant at the 0.1% level (Figure 3), indicating a reasonable core of significantly periodic proteins, and placing it 16th out of 45 species at the 0.1% level. However, this core of periodic proteins is mostly weak in its z-scores, and at low values of *n*. Overall, those species with proportionally fewer periodic proteins also tend to have weaker periodicities in those proteins.

If the maximally periodic proteins in each species are examined, many of the periodic elements may be seen to be known sequence domains listed in the Pfam database (Bateman et al. 2004). For instance, the 114-mer periodicity in human apolipoprotein A precursor, mentioned above, is caused by a series of Kringle domains (Pfam PF00051). Those cases where the maximally periodic protein has an identifiable repetitive Pfam domain, or can be explained by some other know repetitive element, are listed in Table 1. In total, in the most periodic proteins of 23 of the 45 species, the periodic element can be identified as a repeated Pfam domain. Additionally, in a further 3 species, the periodicity in the most periodic protein can be identified as being the result of alpha helices. The tetratricopeptide, or TPR, domain (PF00515) is responsible for the strongest periodicity in a total of 6 species: *Aquifex aeolicus*, (hypothetical protein AQ_854; O67021), *Methanobacterium thermautotrophicum*, (O-linked GlcNAc transferase; O26176), *Methanococcus janaschii*, (hypothetical protein MJ1345; YD45_METJA), *Methanosarcina acetivorans*, (hypothetical protein MA3293; Q8TKV1) *Methanosarcina mazei* (conserved hypothetical protein; Q8Q0F8) and *Pasteurella multocida* (hypothetical protein PM2006; Q9CJJ9), of which it is notable that 4 are archaea. Periodicity at n=17 and multiples is frequently encountered in all three superkingdoms, especially in *Methanosarcina acetivorans*, *M. mazei*, *Methanococcus thermautotrophicum*, *Anabaena* sp. and *Schizosaccharomyces pombe*. TPR is thus an important periodic domain across a

wide phylogenetic spectrum. In fact, of the 71 proteins significantly periodic at the 0.1% level in *M. acetivorans*, 19 are periodic as a consequence of the presence of multiple TPR domains. The periodic length associated with the TPR domain need not necessarily be *n*=34. Within the *M. acetivorans* hypothetical protein MA3293 (Q8TKV1) the periodicity is actually *n*=48, a 34-mer TPR domain alternating regularly with 14 residues of spacer. In *Pasteurella multocida* hypothetical protein PM2006 (Q9CJJ9), 2 residues of spacer are found on the end of each TPR, thus giving a periodicity of *n*=36.

Pentapeptide repeats are responsible for the most periodic protein in a further 3 species. In the eubacteria *Yersinia pestis* and *Rickettsia prowazekii*, the pentapeptide repeats are from Pfam family PF00805, and in the eukaryote *Caenorhabditis elegans*, they are GETHR repeats from family PF05671. In a further 4 species, *Shigella flexneri*, *Pyrococcus abysii*, *Pyrococcus horikoshii* and *Archaeoglobus fulgidus*, a single methyl-accepting chemotaxis protein (MCP) signalling domain (Pfam PF00015) is found in the most periodic protein. Although this is not a repeated domain, it is probable that its high content of alpha helices is the cause of the periodicity at *n*=7.

As well as relatively well-characterised Pfam-A domains, some periodicities appear to be the result of repetitions of the Prodom-derived Pfam-B domains. In *E. coli*, the hypothetical protein ECs0371 (YAHH_ECOLI) is the most strongly periodic, scoring z=8.2 at *n*=31, caused by 3 near perfect repeats of a Pfam-B_31546 domain. This surpasses the periodicity of z=7.5 at *n*=7, found in side tail fiber protein homolog from lambdoid prophage (STFR_ECOLI), the "winning protein" in an earlier draft of the *E. coli* proteome, identified in our previous paper (Gatherer and McEwan 2003). A single copy of the same 31-mer Pfam-B_31546 element is also found in the hypothetical protein ecs0371 (Q8X367).

However, periodicity is not always easy to correlate with Pfam domains. In *Agrobacterium tumifaciens*, the most periodic protein is ice-nucleation protein homolog (Q8U8W4), with z=35.7 at *n*=8 and multiples, caused by repetitions of a few apparently

unrelated 64-mers, each 64-mer being represented a handful of times. This protein is a complex mixture of Pfam-B domains, of which the boundaries do not reflect those of the 64-mers. In this particular case, it is suggested that the periodicity analysis perhaps reveals more about the protein's internal structure than Pfam-B. In some cases, there are no Pfam domains of any kind to be seen. For instance, the most periodic protein in *Pyrobaculum aerophilum* is PaREP7 (Q8ZY83), with a z-score of 10.3 at n=25, caused by a set of 8 slightly variable 25-mers in the middle third of the protein. There is no Pfam domain in this protein, except for a short stretch of Pfam-B_47108 in the extreme N-terminus, which is outside the periodic region. There is also no protein in the PDB that is significantly periodic at n=25, so there is no model anywhere for such a periodicity. In *Clostridium tetani*, putative sialidase EC 3.2.1.18 (Q898J4) has a periodicity of z=10.4 at n=45, caused by 6 slightly variable copies of a 45-mer covering the entire protein except for the first 22 residues, but has no Pfam domains.

4. Structural correlates of periodicities

Where a protein is periodic at n, it will necessarily be periodic at multiples of n, although the z-score will be lower (except in cases where the repeat is perfect). We refer to this as the "multiple effect". This effect is not peculiar to our algorithm, but can be seen in other algorithms for calculating periodicity, (e.g. Fig. 2 of Coward & Drablos (Coward and Drablos 1998)). Examination of the distribution of repeats over all values of n tested, which we refer to as the "periodicity profile", permits the identification of possible families of periodic proteins. The word 'family' must be used with caution here, as it must not be taken to imply homology by descent, but merely a group of proteins with the same significant periodic length. In 17 out of the 45 species, n=7 is the major such group. This is not likely to represent a gene family but rather to reflect the 7-mer periodicity known to exist in alpha-helices. In the PDB, this may be seen in structures such as 1C1G (pig tropomyosin - Figure 5A), having periodicity of z=9.9 at n=7. [Place here Figure 5] Similarly, periodicity at n=11 may also involve alpha helices, e.g. in 1AV1 (apolipoprotein A - Figure 5B), where z=8.8. It is important to note, however, that not all

periodicities at multiples of 7 or 11 are necessarily "multiple effects". For instance, in the PDB, 1BK6 (yeast karyopherin A chain - Figure 5C) has periodicity of z=5.4 at *n*=42, and the periodicity corresponds to an armadillo/beta-catenin-like repeat domain (Pfam PF00514). Although the armadillo domain is constructed largely of alpha helix, and is therefore also periodic at *n*=7, the periodicity at *n*=42 is not a "multiple effect", since it is stronger than the periodicity at *n*=7 in that protein. 1MEY chain F (a designed protein representing a consensus of zinc finger proteins - Figure 5D) has periodicity of z=4.7 at *n*=28, corresponding to a C2H2 type zinc finger (Pfam PF00096) which has both alpha helix and beta sheet elements. 1D0B (*Listeria monocytogenes* internalin - Figure 5E) with periodicity of 7.7 at *n*=22 represents a leucine-rich repeat (LRR - Pfam PF00560), also with both alpha and beta structure, and is not a multiple effect of an alpha helical periodicity at *n*=11. Even when a structure is not available in PDB, it is sometimes possible to speculate about structure by reference to Pfam. For instance, the most periodic protein found in *Porphyromonas gingivalis* is a leucine-rich protein (Q7MTS7), at *n*=22 with a z-score of 12.1. In the absence of any other evidence, the only speculation that could be made would be that this was a multiple effect of alpha helical periodicity at *n*=11. However, the periodic region also corresponds to a leucine-rich repeat Pfam domain. This suggests that the structure of Q7MTS7 may be similar to that of 1D0B. It is notable that the z-score of Q7MTS7 is much greater than that of 1D0B. This is because 1D0B only has 8 leucine-rich repeats, as opposed to 17 in Q7MTS7. The LRR domains in Q7MTS7 are also identical, whereas those in 1D0B are considerably divergent. This suggests that an application of knowledge of periodicity may be found in homology modelling.

However there are also several 'families' of periodicities that cannot be compared to any structures in PDB, since no significant periodicities at those values of *n* are found in the PDB. These include the groups periodic at *n*=23 and *n*=42 in *Streptomyces coelicolor*, at *n*=31 in *E. coli*, at *n*=39 in *Schizosaccharomyces pombe*, at *n*=33 in *Bacillus cereus*, and at *n*=23 in *Drosophila melanogaster* and *Methanosarcina acetivorans*. However, Pfam analysis reveals that in no case are all of the members of the above 'families' in any one

species periodic due to a single kind of repeat or Pfam domain. Rather, there are often a handful of different kinds of periodic elements that can generate periodicity at any particular value of *n* (results not shown). It has been suggested (Del Carpio-Munoz and Carbajal 2002) that periodicity may be useful in detection of remote homologies. The above observations imply that it would be wise to combine this method with a Pfam analysis, in order to avoid being misled by unrelated periodic elements that share the same value of *n*.

5. Comparison of periodicity in related species

It is interesting to observe the similarity between the periodicity profile at the 1% level, in human and mouse (Figure 6). [Place here Figure 6] Although the human proteome has more periodicities than the mouse, the pattern of peaks is virtually identical. The only discrepancy is a slightly elevated number at n=35 in the human proteome. By contrast, two members of the genus *Bacillus*, *B. cereus* and *B. anthracis* (Figure 7 A,B) have very different profiles. [Place here Figure 7] The most periodic protein in *B. cereus*, collagen adhesion protein (Q81GX1), does not appear to be present in *B. anthracis*, as the latter has no protein significantly periodic at n=93. This was confirmed by BLAST searches; BLASTP found no homologous *B. anthracis* protein, and TBLASTN found no region of the *B. anthracis* genome capable of coding for such a homologue.

This situation in *Bacillus* is mirrored in the genus *Methanosarcina*, by *M. acetivorans* and *M. mazei* (Figure 7 C,D), the only archaea with substantial numbers of periodic proteins. There are also visible differences between the three members of the genus *Pyrococcus* studied here, *P. furiosus*, *P. horikoshii* and *P. abyssi* (Figure 8), the two members of the genus *Sulfolobus*, *S. tokodaii* and *S. solfataricus* (Figure 9 A,B), and the two members of the genus *Thermoplasma*, *T. volcanicum* and *T. acidophilum* (Figure 9 C,D). [Place here Figure8, Figure 9] However in *Pyrococcus* and *Sulfolobus*, there are few significant periodicities. In *Pyrococcus furiosus* the maximum periodicity is z=4.0 in acetyl/acyl transferase related protein (Q8U2R4). This is caused by the presence of the bacterial

transferase hexapeptide (Pfam PF00132). The consensus sequence for this short repetitive element is [LIV]-G-X(4). *P. horikoshii*, by contrast, is one of several species which have a methyl-accepting chemotaxis protein (in this case O59504) as the most periodic protein. This is alpha helix-rich and therefore scores z=7.8 at *n*=7. The top four periodic proteins in *P. horikoshii* (O59504, O58181, O58196, O58227) are all periodic at *n*=7. These constitute a family of methyl-accepting chemotaxis proteins found in *P. horikoshii* but not in *P. furiosus*. BLAST searches confirm that no sequence capable of coding for a homologue of these proteins is found in the *P. furiosus* genome. In *P. abyssi*, the situation appears most similar to *P. horikoshii*, with a family of proteins periodic at *n*=7 and its multiples occupying the top 5 spaces in its periodicity table (Q9UYB8, Q9UYF0, Q9V1K4, Q9UYF8, Q9UYE0). These are all methyl-accepting chemotaxis proteins, and homologues of the corresponding top periodic proteins in *P. horikoshii*. *T. acidophilum* acetolactate synthase large chain related protein (Q9HKB0) gives the highest z-score in this species at z=3.1, less than the maximum z-score in the shuffled version of this proteome.

6. Short range periodicities apparently not correlated with structure

The above summary has concentrated on cases where the periodicity can be explained by either a demonstrated repetitive structure found in the PDB, as a Pfam domain, or as a known periodic secondary structural feature such as alpha helix. However, there are many cases where significant "short range" periodicities are seen, here defined as a periodicity of *n*=9 or less. Many of these are immediately visible on inspection as simple repetitive elements. For instance, in *Treponema pallidum* the most periodic protein, hypothetical protein TP0470 (O83483), has three TPR Pfam domains, but also a periodicity of z=10.3 at *n*=8 and multiples, caused by the presence of the octapeptide RKEAEEAR in 17 exact copies and one slight variant, in the C-terminal half of the protein. Another strong example is the periodicity at *n*=6, z=9.1, in hypothetical protein ma1459 (Q8TQT1) in *Methanosarcina acetivorans.* Such regions are often identified in Pfam as "low complexity". A few of these are common enough to be classifiable as Pfam

domains in their own right, for instance the pentapeptide repeats mentioned above, or the bacterial transferase hexapeptide (PF00132) which is present in the most periodic protein of *Pyrococcus furiosus*, acetyl/acyl transferase related protein (Q8U2R4). As well as short-range low complexity periodicities, there are occurrences of longer periodicities within Pfam low complexity regions. For instance, in *Pasteurella multocida*, electron transport complex protein rnfC (RNFC_PASMU) has a moderately long periodic length of n=24, z=5.2, and the region is classified as low complexity. In eukaryotes, *Schizosaccharomyces pombe* hypothetical serine/threonine repeat containing protein (Q9HDY9) has a periodicity of z=17.8 at n=36 within a Pfam low complexity region.

A third class of short periodicities exists, however, which are not low complexity, nor are they correlated with any obvious repetitive structural element. As an illustration of this, consider 1A8P from PDB (Figure 5F), described as NADPH\:ferredoxin oxidoreductase from *Azotobacter vinelandii*. This structure does not appear periodic on visible inspection, but has a z-score of 3.4 at period n=5, and is thus one of the 20 proteins in the PDB that are significantly periodic at the 0.1% level. If converted to the Dayhoff alphabet, the position-specific occurrence of each residue is tabulated in Table 2, with the expected occurrence under the null hypothesis of equiprobability (exp) and the chi squares for the rows and columns (chi pos, chi res) calculated using the CHITEST function in Microsoft Excel.

Table 2 shows that positions 2, 3 and 4 all deviate from the null hypothesis at the 5% significance level. Position 5 by contrast has a distribution of Dayhoff residues very close to the expected. Also, the distribution of the third (DENQ) and sixth (FWY) Dayhoff clusters deviates from equiprobability at the 5% level, and the fifth (ILMV) cluster at the 1% level. Therefore the overall significant periodicity of protein 1A8P is caused principally by the excess of the fifth Dayhoff cluster (ILMV) in positions 3 and 4 of a 5-mer period length. This is illustrated in Figure 10. It can be seen that these residues are distributed along the length of the protein. [Place here figure 10]

This class of repeat provides the most typical example of what was predicted by the neutral model of protein periodicity, i.e. a steady background of subtle sequence periodicity in a protein that has evolved a non-periodic structure. Such proteins are, however, relatively rare as they constitute a minority of periodic proteins, which themselves are only a small part of a largely non-periodic proteome in all species.

6. Conclusions and potential Evolutionary significance

In summary, our previous analysis of periodicity in *E. coli* (Gatherer and McEwan 2003) demonstrated that this species has a core of significantly periodic proteins within a mainly non-periodic proteome. The present reanalysis of the updated *E. coli* proteome, in comparison with 44 other species from all three superkingdoms of life, demonstrates that the pattern of periodicity in *E. coli* is typical of many eubacteria (Figures 1-3). However, this comparison also reveals that many eukaryotes and archaea differ noticeably from the initial *E. coli* example. In particular, there are proportionally more periodic proteins in eukaryotes, especially multicellular eukaryotes, and most of the species with a low proportion of periodic proteins are archaea. Secondly, in those species with plentiful periodicities, the periodicities tend to be stronger. Thirdly, the distribution of periodicities over *n*, is also variable, in some cases reflecting phylogenetic relationships, e.g. in the similarity of human and mouse, and in other cases failing to do so, e.g. in the genus *Bacillus*. Fourthly, there are some periodicities that recur in numerous species, e.g. at n=7, 11 or 34. Finally, it is possible relate many periodicities to known structures in the PDB or known domains in Pfam, but others are more difficult to explain.

The central conclusion of the present study is that there is no ubiquitous “periodicity background radiation” to be found in any of the 45 proteomes studied. This is evidence against the “protein synthetic big bang” (terminology of White & Jacobs (White and Jacobs 1993). However, it remains possible, or even probable given what is now known about mechanisms of DNA turnover (Dover 2002), that “mini-bang” events may have

taken place at different times. The presence of strongly periodic elements in a minority of proteins is difficult to explain unless some mechanism of internal duplication is involved. Since many of these periodic elements are literally repetitive, or exhibit a high degree of conservation, they must either be due to recent duplication events, upon which mutational drift has not had time to become apparent, or they must be relics of older events that have been strongly preserved by natural selection. Where the periodic elements are Pfam domains or other typical structures, natural selection would seem to be the most plausible explanation. In addition to strongly periodic elements, there are also some proteins exhibiting periodicity, and particularly short range periodicity (low n), not correlated with structure, for instance 1A8P described above, or the heat shock proteins discussed by Ohno (Ohno 1988). This is more difficult to explain by natural selection, and may represent a decaying relic of much older duplications. However, it is confined to a minority of significantly periodic proteins, which are themselves a small minority of proteins as a whole, so it is impossible to infer any "protein synthetic big bang" with confidence. It is also possible that the "protein synthetic big bang" simply occurred too long ago to have left any traces in modern proteins. The opposing "random origins hypothesis" (White and Jacobs 1993) is more consonant with the overall findings of the present study, but it would be very difficult to defend the idea that it is the exclusive mechanism, simply because many periodicities are so strong, and because it is known that DNA is not just subject to point mutation. Evolving a literal repetitivity out of a random background sequence simply by accumulation of mutations would seem to be impossible except under the most ferocious and consistent natural selection.

The correlation coefficient between number of proteins in each species and the maximum z-score in that species, is 0.732, with a t-test demonstrating that the correlation is statistically significant ($p<0.001$). Likewise, the correlation coefficient between the total number of proteins in each species and its percentage of periodic proteins at the 5%, 1% and 0.1% significance levels is 0.507, 0.683 and 0.758 respectively, with t-tests demonstrating that all these correlations are statistically significant ($p<0.001$ in all three cases). These correlations are explained by the fact that eukaryotes, and especially

multicellular eukaryotes, tend to have larger proteomes than prokaryotes, and that archaea tend to have smaller proteomes than eubacteria. *M. acetivorans*, which exhibits a eukaryote-like content of periodic proteins, also has the largest known archaeal proteome. Conversely, unicellular eukaryotes with little protein periodicity, such as *Guillardia theta* or *Encephalitozoon cuniculi*, also have the smallest eukaryote proteomes. This suggests that multicellularity requires not only a larger protein set than unicellularity, but that the cellular mechanisms leading to periodicity are under some kind of selection. Either periodicity is positively selected in multicellular species, or selected against in unicellular species. The idea that unequal crossover is a mechanism for the generation of repeat sequences (Smith 1976) helps to explain why meiotic eukaryotes have more periodicity than the asexual prokaryotes, but it does not explain the differences between eubacteria and archaebacteria.

Within the 45 species studied here, species with larger proteomes also tend to have longer average protein sizes (r = 0.669, p<0.001). This may be due in part to their larger complement of periodic proteins. If periodicity were produced by internal duplication within proteins, then one would expect periodic proteins to be longer than non-periodic proteins. Correlation coefficients between average protein length in a species and its percentage of periodic proteins at the 5%, 1% and 0.1 % significance levels are 0.326, 0.390 and 0.492 respectively, with t-tests demonstrating that all these correlations are statistically significant at p<0.05, p<0.01 and p<0.001 respectively. These correlations are slightly weaker than the others described above, but this is to be expected since duplication at short values of n may not appreciably lengthen a protein, or as a result of duplication events where some of the periodicity has diminished with time. Although the general conclusion of this paper is that periodicity is under natural selection, the presence of apparently non-functional residual periodicity in some proteins, such as 1A8P mentioned above, and the tendency of average protein length in a species to correlate with the content of periodic proteins does suggest that the neutralist theory may be correct in it proposed mechanism, i.e. internal duplication of short n-mers, if not in its “big bang” scenario. Splitting the “big bang” into a series of small and relatively rare ongoing “little

bangs", superimposed on a more general random origins hypothesis with active natural selection on the resulting periodicities, seems to be most consonant with the existing data.

Smith (Smith 1976) simulated a 500bp DNA sequence evolving over 200 generations with varying degrees of mutation and unequal crossover permitted. Fig. 2 of Smith (Smith 1976) shows a starting random sequence and a typical product after 200 cycles of simulation. These data were reanalysed using our algorithm, and the random sequence was found to have a maximum periodicity of z=1.4 at n=13, whereas after 200 cycles it has developed a maximum periodicity of z=44.5 at n=5. It is not possible to establish a significance threshold here since only a single control is compared against a single result, but it is clear that an increase of over 30-fold in maximum z-score is due to a remarkable increase in periodicity, which is also visible to the naked eye. The implication of this is that periodicity can evolve from a random starter set providing that there is a mechanism for generation of short duplications, in this case unequal crossover. It would be interesting to recode the simulation (Smith 1976) and vary the parameters of mutation and unequal crossover to confirm the circumstances under which the random origins hypothesis could result in the spontaneous appearance of periodicity. Unequal crossover as a mechanism for the generation of repeats has a positive feedback component, since the creation of a repeat region also increases the probability of pairing at the next round of sister chromatic exchange, and thereby the chances of another unequal exchange and further expansion of the repeat region. It is tempting to speculate that such a process is more likely in sexually reproducing eukaryotes where meiosis gives a greater opportunity for crossover events. Prokaryotic recombination does not occur directly between genomes but only between genomes and plasmids, phages or other elements horizontally transmitting genetic material between their hosts. Those unicellular eukaryotes that are relatively periodicity-poor, such as the yeasts, *Encephalitozoon* or *Guillardia* have life cycles where meiosis does not occur in every generation (Canning 1988; Cushion 2004).

The most intriguing question is perhaps not why many species have periodic proteins, but why many archaea seem to be virtually devoid of them, or at least have it at a greatly

impoverished level by comparison with eubacteria and especially eukaryotes. The kinds of repetitive structures seen in Figure 5 cannot occur in many archaeal genomes, since if they did they would be manifest as significant periodicities. For instance in *Archaeoglobus fulgidus*, the strongest periodicity is z=6.3 at n=7. This is methyl-accepting chemotaxis protein (O29228) which is rich in alpha helix. The second most periodic protein in this species is also a methyl-accepting chemotaxis protein (O29217) with z=4.4 at n=7. The third is Hypothetical protein AF2031 (O28248) with z=4.3 at n=7 due to a coiled coil-rich region. Aside from O29228, there is no protein in *Archaeoglobus fulgidus* as periodic as any of the PDB structures in Figure 5, the least periodic of which has z=4.7. The only protein significantly periodic at the 1% level in *A. fulgidus* that has a periodicity of any substantial length of n, is hypothetical protein af1881 (YI81_ARCFU), where a Pfam-B_76962 domain gives a periodicity of z=3.4 at n=25. Four archaeal species have even lower maximum z-scores than *Archaeoglobus fulgidus*: *Thermoplasma volcanicum* (max. z=3.2), *Thermoplasma acidophilum* (max. z=3.1), *Pyrococcus furiosus* (max. z=4.0) and *Methanopyrus kandleri* (max. z=3.7). All of these are caused by short-range periodicities rather than periodicity of whole domains or even medium length repeats. It is therefore possible to say that, at least in the 4 or 5 most periodicity-poor archaeal genomes, there is a severe restriction on the permitted tertiary structures that proteins can assume.

Taken together, these results suggest that there are different mechanisms at work, in different species, regulating the production and tolerance of repeated or periodic elements within proteins. It is possible that increased functional redundancy, perhaps as a result of greater internal homeostatic complexity, in larger proteomes, means that the selective pressures on fresh internal sequence duplications are different in ‘advanced’ or ‘higher’ eukaryotes as compared to unicellular organisms. The word ‘advanced’ in this context would also extend to prokaryotes with multicellular tendencies, such as *Methanosarcina* and *Anabaena*. It is also possible that there are simply more processes at work within larger eukaryote genomes that tend to promote, or at least tolerate, internal protein sequence duplication. Either or both of these mechanisms will mean that genome

evolution may be a very different process in different phylogenetic groups. Alternatively, the internal cellular genetic mechanisms may be the same in all species, but external selective pressures may be driving the proteomes in very different directions. The extreme conditions inhabited by many of the archaeal genera (e.g. *Pyrococcus*, *Thermoplasma*, *Methanobacterium*) may impose strong selective constraints against internal amplification of proteins.

The present study only reveals periodicities where the periodic element is between n=2 and n=100, and where it is of equal length. Where periodic elements are separated by spacer regions of variable length, they will not be detected. Equally periodic elements where insertion events have taken place into one of the elements, or where an internal deletion has resulted in loss of part of the element, will also be missed. Further studies could extend to longer periodic lengths simply by increasing the values of n examined. However, coping with 'ragged' periodicities is beyond the scope of the algorithm presented here, as its very definition of periodicity is based on positional asymmetry which requires equal repeat length. Also, since periodicity is here measured across a protein as a whole, shorter regions of periodicity within a longer, mostly non-periodic, protein may be missed. This could be addressed by the use of a sliding window. However, it should be remembered that significance thresholds would need to be recalculated afresh on shuffled sequences, as a sliding window would necessarily involve several tests on a single protein for any particular value of n. This would however increase the likelihood of a false positive result (Bonferroni 1936) and in turn would be likely to be less informative than the algorithm presented here.

The most periodic proteins in each species

Species	Protein	Periodicity at *n*	Pfam Domain	Number of domains
Aeropyrum pernix	Hypothetical tropomyosin (Q9YCN7)	7	N/A (known periodicity at *n*=7 in tropomyosins)	N/A
Anabaena sp.	Hypothetical protein Alr1903 (Q8YVS1)	31	HEAT_PBS (PF03130)	21
Aquifex aeolicus	Hypothetical protein AQ_854 (O67021)	34	TPR (PF00515)	13
Arabidopsis thaliana	Extensin-like protein (Q9STM7)	25	Extensin-like region (PF04554)	4
Bacillus anthracis	Conserved repeat domain protein (Q81Y32)	132	DUF11 (PF01345)	15
Bacillus cereus	Collagen adhesion protein (Q81GX1)	93	Cna protein B-type domain (PF05738)	20
Borrelia pertussis	Hypothetical protein BBI16 (O50870)	18	NUMOD3 motif (PF07460)	12
Caenorhabditis elegans	Hypothetical protein (Q9N5E5)	5	GETHR pentapeptide repeat (PF05671)	47 (each with 5 repeats)
Encephalitozoon cuniculi	Hypothetical protein ECU11_0430 (Q8SU70)	36	SEL1 (a SMART motif)	N/A
Helicobacter pylori	Putative beta-lactamase hcpD (HCPD_HELPY)	36	SEL1 (a SMART motif)	N/A
human	Apolipoprotein(a) precursor (APOA_HUMAN)	114	Kringle (PF00051)	38
Methanobacterium thermautotrophicum	O-linked GlcNAc transferase (O26176)	34	TPR Domain (PF00515)	11
Methanococcus janaschii	Hypothetical protein MJ1345 (YD45_METJA)	34	TPR Domain (PF00515)	8
Methanosarcina acetivorans	Hypothetical protein MA3293 (Q8TKV1)	48 (includes spacer sequence of 14 residues between domains)	TPR Domain (PF00515)	12
Methanosarcina mazei	Conserved hypothetical protein (Q8Q0F8)	34	TPR Domain (PF00515)	46
mouse	Polyubiquitin C (Q9ET23)	76	Ubiquitin (PF00240)	13

Pasteurella multocida	Hypothetical protein PM2006 (Q9CJJ9)	36 (includes spacer sequence of 2 residues between domains)	TPR Domain (PF00515)	12
Porphyromonas gingivalis	Leucine-rich protein (Q7MTS7)	22	Leucine-Rich Repeat (PF00560)	17
Pyrococcus furiosus	Acetyl / acyl transferase related protein (Q8U2R4)	2 (the hexapeptide is more periodic at 2 than 6)	Bacterial transferase hexapeptide (PF00132)	15
Rickettsia prowazekii	Hypothetical protein RP563 (Q9ZCY8)	5	Pentapeptide repeats (PF00805)	9 (each with 8 repeats)
Saccharomyces cerevisiae	Flocculation protein FLO1 precursor (FLO1_YEAST)	45	Flocculin repeat (PF00624)	18
Schizosaccharomyces pombe	Hypothetical protein (Q96WV6)	36	Domain of unknown function DUF963 (PF06131)	147
Streptomyces coelicolor	Putative sensory histidine kinase (O86808)	92	HAMP domain (PF00672)	11
Sulfolobus solfataricus	Microtubule binding protein, putative (Q9UXN4)	7	N/A (known periodicity at n=7 in alpha helices)	N/A
Sulfolobus tokodaii	Hypothetical protein ST1088 (Q972P4)	7	N/A (known periodicity at n=7 in alpha helices)	N/A
Yersinia pestis	Hypothetical protein YPO0510 (Q8ZII8)	5	Pentapeptide repeats (PF00805)	4 (each with 8 repeats)

Table I: The most periodic proteins in each species, where the periodicity is explicable by a repeated Pfam domain or a periodic secondary structural element.

Periodicity in 1A8P

residue	*1*	*2*	*3*	*4*	*5*	*exp.*	*chi res.*
C	0	0	2	0	0	0.4	0.092
AGPST	16	15	14	9	20	14.8	0.374
DENQ	15	19	6	8	7	11.0	**0.019**
HKR	6	10	8	4	6	6.8	0.548
ILMV	7	7	20	21	13	13.6	**<u>0.009</u>**
FWY	8	1	1	9	5	4.8	**0.019**
chi pos	0.194	**0.016**	**0.011**	**0.030**	0.577		

Table II:
Chi-square table showing the deviation from equiprobability at positions 2, 3 & 4, and for residue groups DENQ, ILMV & FWY, in protein 1A8P, NADPH\:ferredoxin oxidoreductase from *Azotobacter vinelandii*

Exp: the expected position-specific occurrence under the null hypothesis of equiprobable distribution. Chi pos: the p-value for the chi-square calculation for each position. Chi res: the p-value for the chi-square calculation for each residue. Bold: significant at the 5% level. Bold underline: significant at the 1% level.

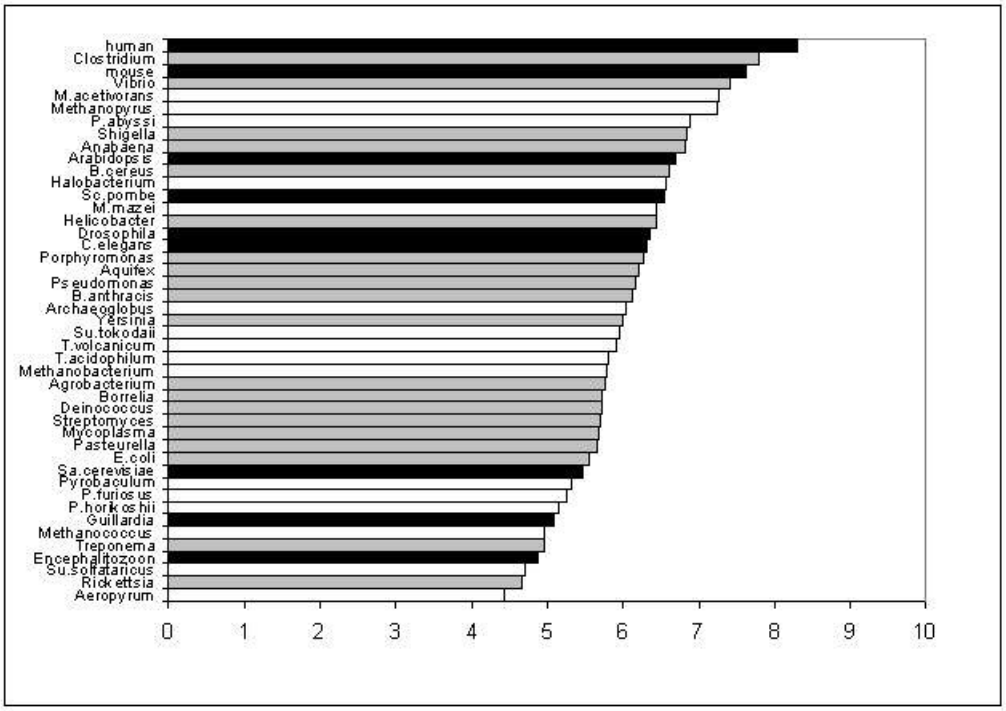

Figure 1: The percentage of proteins periodic at the 5% significance level, in 45 species. Black: eukaryotes. Grey: Eubacteria. White: archaea.

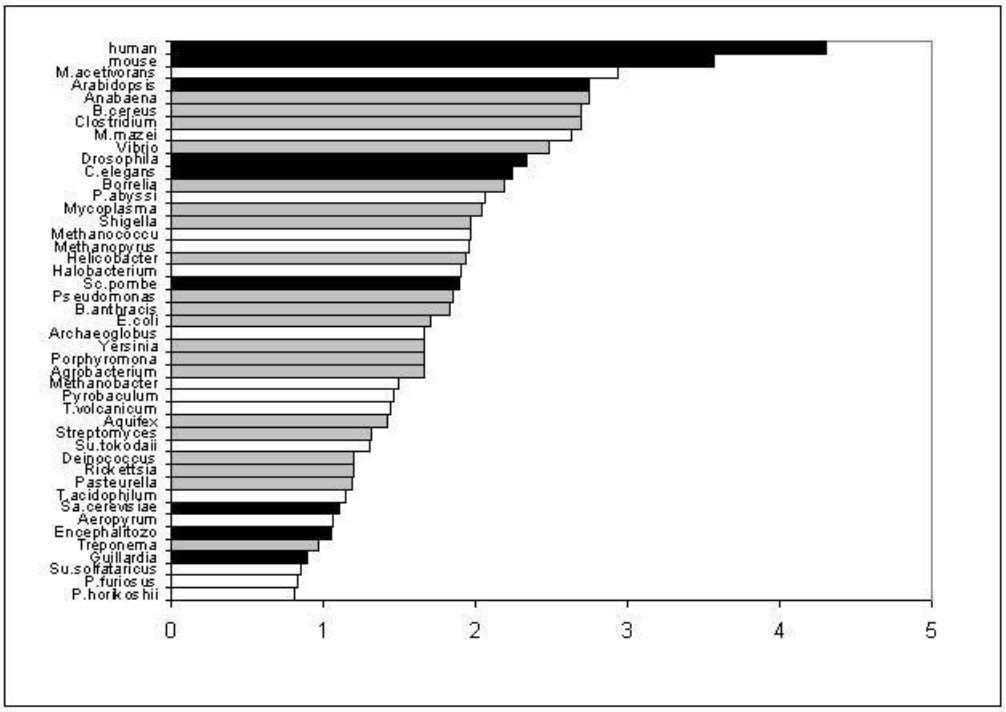

Figure 2: The percentage of proteins periodic at the 1% significance level, in 45 species. Black: eukaryotes. Grey: Eubacteria. White: archaea.

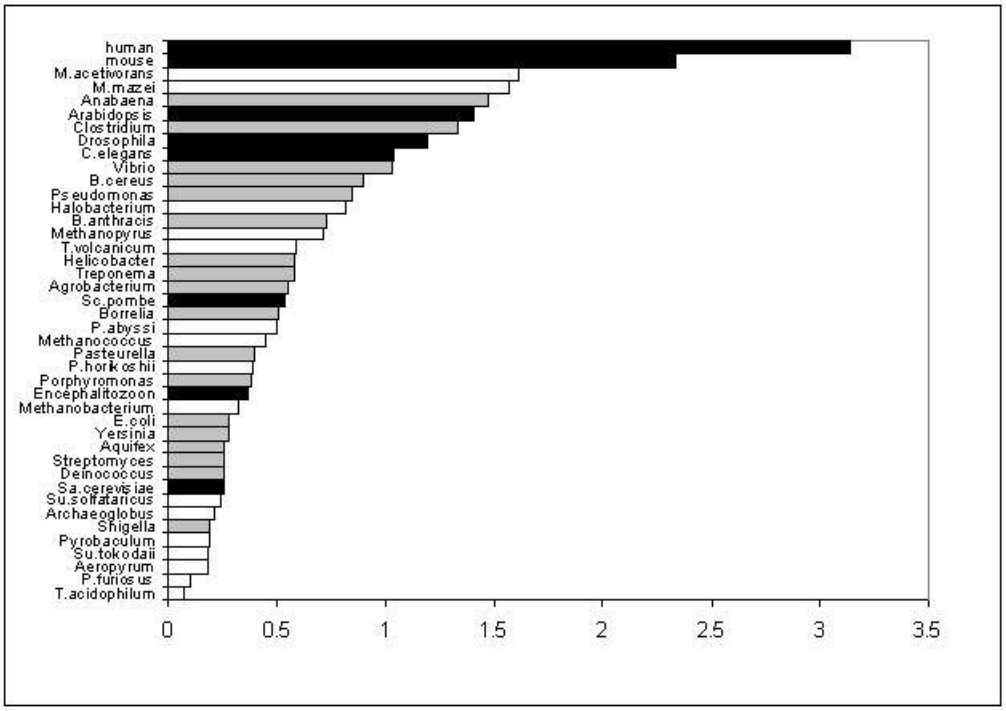

Figure 3: The percentage of proteins periodic at the 0.1% significance level, in 42 species. Black: eukaryotes. Grey: Eubacteria. White: archaea. 3 species included in Figures 1 & 2 are omitted here as their total number of proteins is too small to establish 0.1% significance.

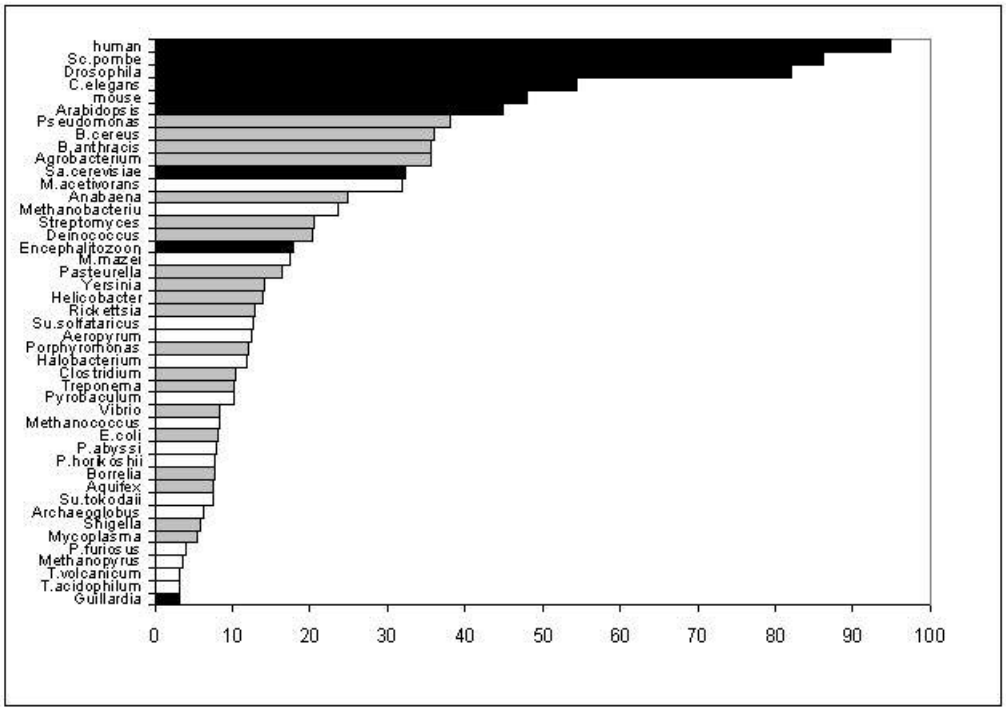

Figure 4: The maximum z-score, representing the degree of periodicity in the most periodic protein, in 45 species. Black: eukaryotes. Grey: Eubacteria. White: archaea.

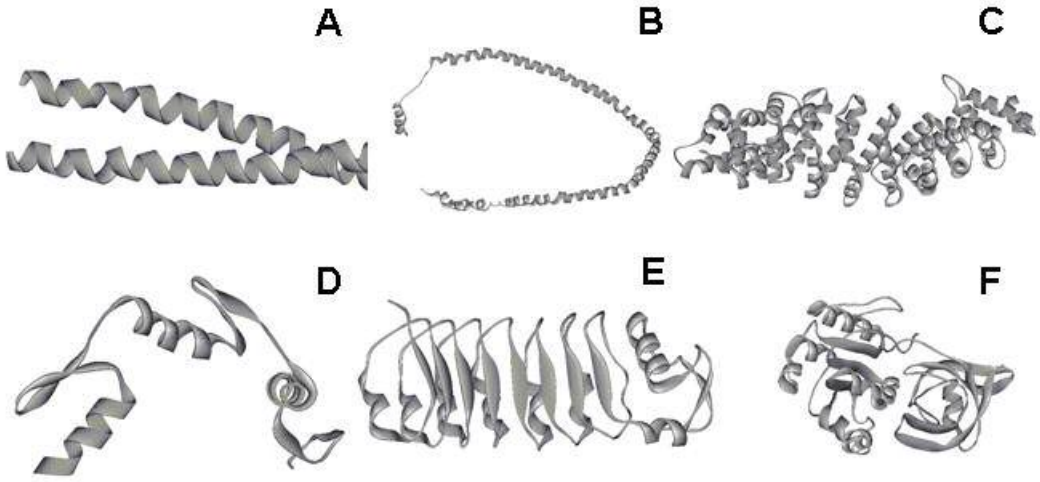

Figure 5. Ribbon diagrams of the tertiary structure of periodic proteins referred to in the text. A: 1C1G, pig tropomyosin, periodicity of z=9.9 at *n*=7. B: 1AV1, human apolipoprotein A, periodicity of z=8.8 at *n*=11. C: 1BK6, yeast karyopherin, periodicity of z=5.4 at *n*=42. D: 1MEY, synthetic zinc finger protein, periodicity of z=4.7 at *n*=28. E: 1D0B, *Listeria* internalin, periodicity of z=7.7 at *n*=22. F: 1A8P, *Azotobacter* NADPH\:ferredoxin oxidoreductase, periodicity of z=3.4 at *n*=5.

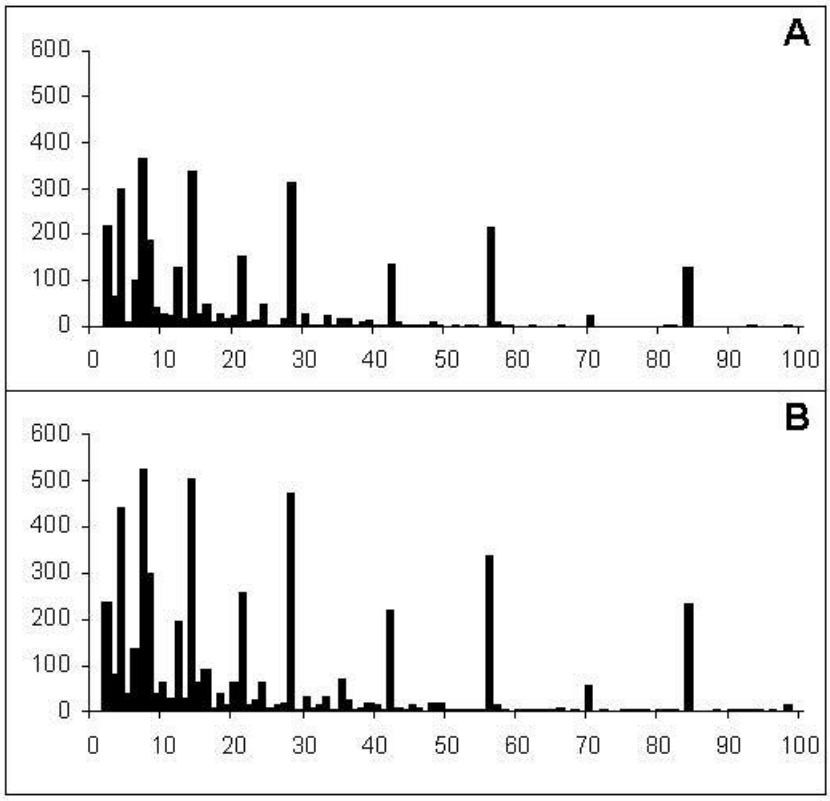

Figure 6. Periodicity profiles for mouse (A) and human (B) proteomes. This is the number of proteins significantly periodic at the 1% level, plotted against the value of n at which they are most significantly periodic. Note the similarity of the profiles, and the peak at *n*=35 found in human but not mouse.

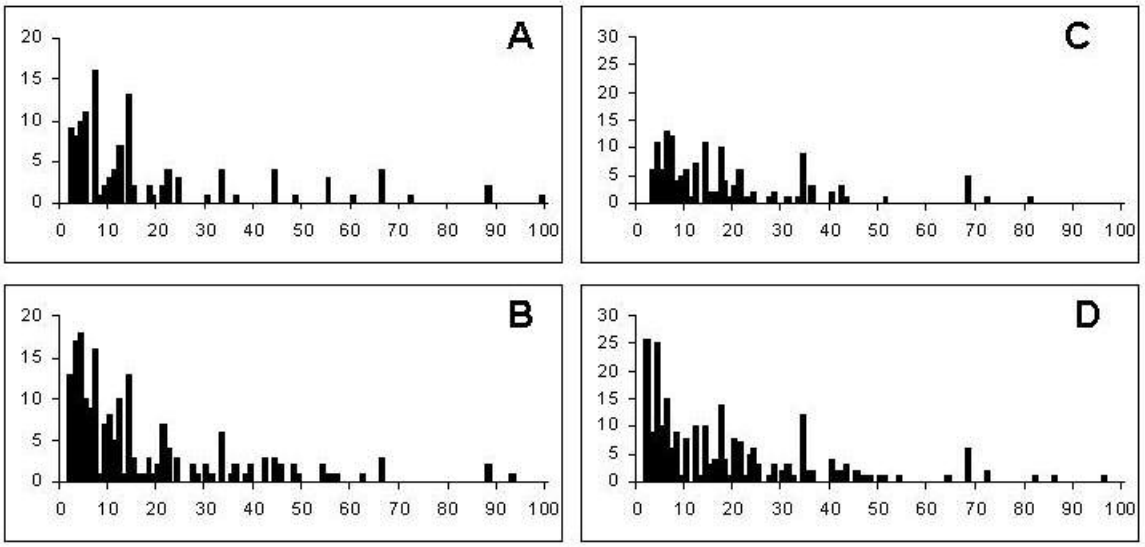

Figure 7. Periodicity profiles for *Bacillus cereus* (A), *Bacillus anthracis* (B), *Methanosarcina acetivorans* (C) & *Methanosarcina mazei* (D). This is the number of proteins significantly periodic at the 1% level, plotted against the value of n at which they are most significantly periodic.

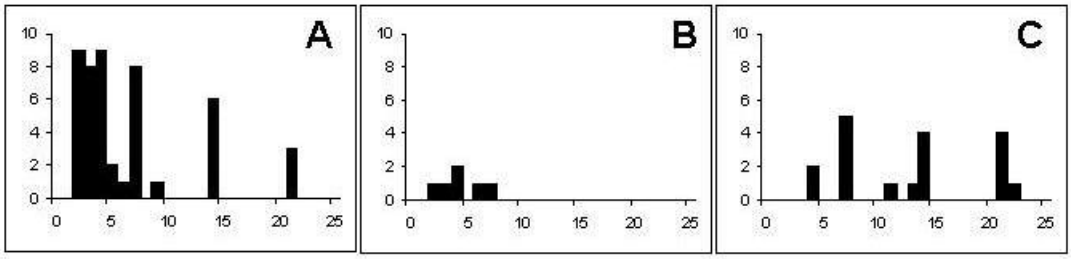

Figure 8. Periodicity profiles for *Pyrococcus furiosus* (A), *Pyrococcus horikoshii* (B) & *Pyrococcus abyssi* (C). This is the number of proteins significantly periodic at the 1% level, plotted against the value of n at which they are most significantly periodic.

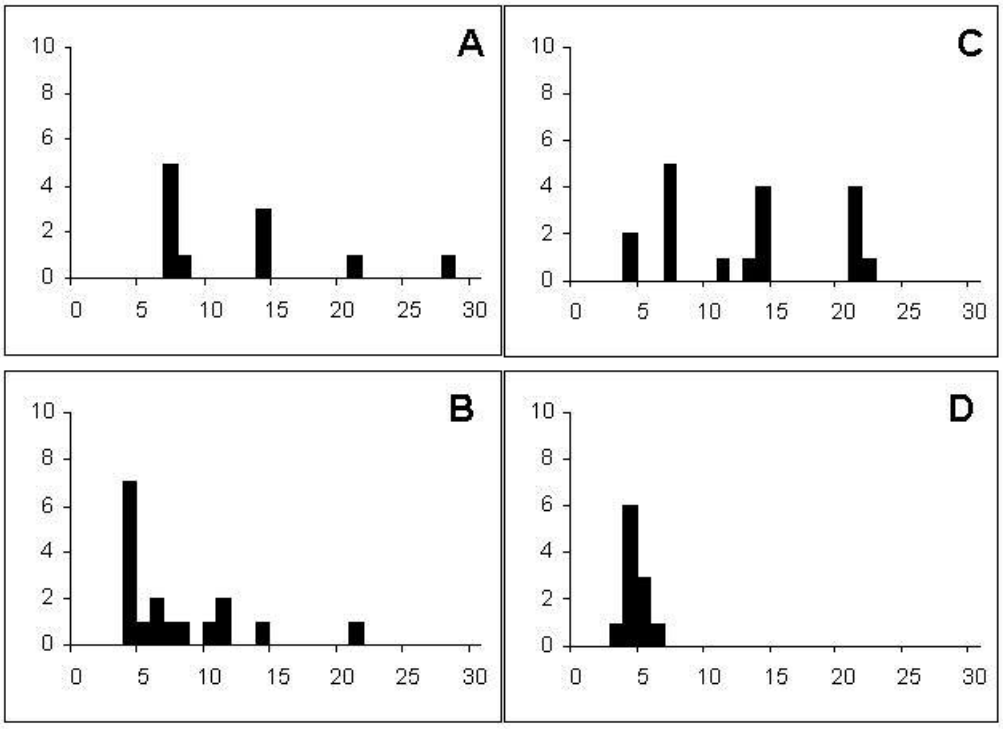

Figure 9 Periodicity profiles for *Sulfolobus tokodaii* (A), *Sulfolobus solfataricus* (B) *Thermoplasma volcanicum* (C) & *Thermoplasma acidophilum* (D). This is the number of proteins significantly periodic at the 1% level, plotted against the value of *n* at which they are most significantly periodic.

>P1;1A8P NADPH\:ferredoxin oxidoreductase - Azotobacter vinelandii
SN**L**NVERV**L**SVHHWNDT**L**FSFKTTRNPS**L**RFENGQFV**MI**GLE**V**DGRP**LM**RAYS**I**ASPNYE
EH**L**EFFS**I**KVQNGPLTSR**L**QHLKEGDE**LM**VSRKPTGT**LV**TSD**LL**PGKH**L**YMLST GTG**L**AP
FMS**L**IQDPEVYERFEKV**VL**IHG**V**RQVNE**L**AYQQFITEH**L**PQSEYFGEA**V**KEK**LI**YYPT**V**T
RESFHNQGRLTD**LM**RSGK**L**FED**I**GLPP**I**NPQDDRAMICGSPS**ML**DESCEVLDGFGLK**I**SP
RMGEPGDY**L**IERAFVEK

Fig 10. Sequence of PDB entry 1A8P. Residues in Dayhoff cluster no. 5 (ILMV) in positions 3 & 4 at *n*=5, are emphasized in underlined bold italic. It can be seen that they are distributed throughout the length of the protein.

References

Ashley C, Warren S (1995) Trinucleotide repeat expansion and human disease. Ann. Rev. Genetics 29:703-728

Barker W, Dayhoff M (1977) Evolution of lipoproteins deduced from protein sequence data. Comp. Biochem. Physiol. B 57:309-315

Bateman A, Coin L, Durbin R, Finn R, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer E, Studholme D, Yeats C, Eddy S (2004) The Pfam Protein Families Database. Nucl. Acids Res. 32:D138-D141

Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P (2000) The Protein Data Bank. Nucl. Acids Res. 28:235-242

Bonferroni C (1936) Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8:3-62

Canning E (1988) Nuclear division and chromosome cycle in microsporidia. Biosystems 21:333-340

Coward E, Drablos F (1998) Detecting periodic patterns in biological sequences. Bioinformatics 14:498-507

Cushion M (2004) Comparative Genomics of Pneumocystis carinii with Other Protists: Implications for Life Style. Journal of Eukaryotic Microbiology 51:30-37

Del Carpio-Munoz C, Carbajal J (2002) Folding pattern recognition in proteins using spectral analysis methods. Genome Informatics 13:163-172

Dover G (2002) Molecular drive. Trends Genet. 18:587-589

Eck R, Dayhoff M (1966) Evolution of the structure of ferridoxin based on living relics of primitive amino acid sequences. Science 152:363-366

Eisenberg D, Weiss R, Terwilliger T (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad. Sci. USA 81:140-144

Galagan J, Nusbaum C, Roy A, Endrizzi M, Macdonald P, FitzHugh W, Calvo S, Engels R, Smirnov S, Atnoor D, Brown A, Allen N, Naylor J, Stange-Thomann N, DeArellano K, Johnson R, Linton L, McEwan P, McKernan K, Talamas J, Tirrell A, Ye W, Zimmer A, Barber R, Cann I, Graham D, Grahame D, Guss A, Hedderich R, Ingram-Smith C, Kuettner H, Krzycki J, Leigh J, Li W, Liu J, Mukhopadhyay B, Reeve J, Smith K, Springer T, Umayam L, White O, White R, Conway de Macario E, Ferry J, Jarrell K, Jing H, Macario A, Paulsen I, Pritchett M, Sowers K, Swanson R, Zinder S, Lander E, Metcalf W, Birren B (2002) The genome of M. acetivorans reveals extensive metabolic and physiological diversity. Genome Res. 12:532-542

Gatherer D, McEwan N (2003) Analysis of sequence periodicity in E. coli proteins: empirical investigation of the 'duplication and divergence' theory of protein evolution. J. mol. Evol. 57:149-158

Golden J, Yoon H (2003) Heterocyst development in Anabaena. Curr. Opin. Microbiol. 6:557-563

Gruber M, Lupas A (2003) Historical review: Another 50th anniversary - new periodicities in coiled coils. Trends Biochem. Sci. 28:679-685

Ivanov O, Ivanov C (1980) Some evidence for the universality of structural periodicity in proteins. J. mol. Evol. 16:47-68

Katti MV, Sami-Subbu R, Ranjekar PK, Gupta VS (2000) Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. Prot. Sci. 9:1203-1209

Korotkov E, Korotkova M, Tulko J (1997) Latent sequence periodicity of some oncogenes and DNA-binding protein genes. Comp. Appl. Biosci. 13:37-44

Korotkova M, Korotkov E, Rundenko V (1999) Latent periodicity in protein sequences. J. mol. Model. 5:103-115

Laskin A, Korotkov E, Chaley M, Kudryashov N (2003) The locally optimal method of cyclic alignment to reveal latent periodicities in genetic texts: the NAD-binding protein sites. Molecular Biology 37:561-570

McLachlan A, Stewart M (1976) The 14-fold periodicity in alpha-tropomyosin and the interaction with actin. J. mol. Biol. 103:271-298

Ohno S (1984) Repeats of base oligomers as the primordial coding sequences of the primeval earth and their vestiges in modern genes. J. mol. Evol. 20:313-321

Ohno S (1988) Codon preference is but an illusion created by the construction principle of coding sequences. Proc. Natl. Acad. Sci. USA 85:4378-4382
Pattabiraman N, Namboodiri K, Lowrey A, Gaber B (1990) NRL_3D: a sequence-structure database derived from the Protein Data Bank (PDB) and searchable within the PIR environment. Protein Sequences & Data Analysis 3:387-405
Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528-535
Stanfel L (1996) A new approach to clustering the amino acids. J. theoret. Biol. 183:195-205
Trifonov E (1985) Segmented structure of protein sequences and early evolution of genome by combinatorial fusion of DNA elements. J. mol. Evol. 40:337-342
Vaara M (1992) Eight bacterial proteins, including UDP-N-acetylglucosamine acyltransferase (LpxA) and three other transferases of Escherichia coli, consist of a six-residue periodicity theme. FEMS Microbiol. Lett. 76:249-254
White S, Jacobs R (1993) The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. J. mol. Evol. 36:79-95
Ycas M (1976) Origin of periodic proteins. Fed. Proc. 35:2139-2140
Zhurkin V (1981) Periodicity in DNA primary structure is defined by secondary structure of the coded protein. Nucl. Acids Res. 9:1963-1971
Zimmerman J, Eliezer N, Simha R (1968) The characterization of amino acid sequences in proteins by statistical methods. J. theoret. Biol. 21:170-201