IDIOSYNCRATIC EVOLUTION OF CONSERVED EUKARYOTE PROTEINS THAT ARE SIMILAR IN SEQUENCE TO ARCHAEAL OR BACTERIAL PROTEINS

Roy J Britten

California Institute of Technology and University of California Irvine.
101 Dahlia Ave Corona del Mar Ca 92625

Abstract

Sequence comparisons have been made between the proteins of 571 prokaryote species including 46 archaea and 525 bacteria and the set of human proteins. Highly conserved eukaryotic proteins are often strikingly similar in sequence to archaeal and bacterial proteins. Yet in many cases similarity to archaeal proteins is not correlated to the similarity to bacterial proteins. In these comparisons there are hundreds of eukaryote proteins that match well archeal proteins, but do not match recognizably to bacterial proteins, while thousands of proteins match well to bacterial proteins but not recognizably to archeal proteins. Forty percent of the 21,440 human proteins that significantly match prokaryote proteins are in this extreme idiosyncratic category. These relationships have been preserved over billions of years since the last common ancestor or sharing of protein genes between prokaryotes and eukaryotes. For each of the 21,440 members of this set of human proteins (that make significant matches to any of the 1.8 million proteins in this set of prokaryote species protein libraries) it is certain that each protein has important functions both in prokaryotes and eukaryotes and the precursor proteins have been important in the precursor species of both. That is the only explanation for the preservation of amino acid sequence similarity for the billions of years since the last common ancestor or period of sharing of proteins. Comparisons were made between the proteins of Arabidopsis thaliana and Saccharomyces cerevisiae to the proteins of the 571 prokaryote species. The results agreed with the human comparisons indicating that the conclusions apply to eukaryotes generally.

Introduction

The clue that initiated this work was the observation using BLASTp sequence comparison that human eukaryotic translation initiation factor (NP_001406) makes high scoring matches with all the 46 archaea in this collection with bit score ranging from 397 to 263 while bit scores with bacteria are much lower ranging from 125 to 36.3, limited by the criterion. The matching region for the archaea is almost the same for all examples from positions 39 to 459 of the human amino acid sequence. Pfam search gave three domains extending altogether from 39 to 459. They include two regions of the elongation factor Tu and the initiation factor eIF2 gamma, C terminal. In most cases the archaeal protein is identified as translation factor IF2 gamma but in some cases identified as protein synthesis factor GTP binding. This is an example of idiosyncratic history of a protein family. At some times in its history this protein sequence evolved more rapidly in bacteria than in archaea, even though it was well conserved in all archaea and moderately

well in some bacteria. In other bacteria its function must have been replaced by a protein with a different sequence. Due to the observation of this idiosyncracy a search was made for such cases among the relationships between eukaryotic and prokaryote protein sequences and many examples were found including many much more extreme cases.

It is well known that histones are present in archaea and absent from bacteria. Many eukaryotic mitochondrial ribosomal proteins are present in bacteria and fewer are recognized in archaea. On the other hand eukaryotic nuclear ribosomal proteins are well represented in archaea, but not in bacteria. These observations are currently being examined in order to clarify the origin of eukaryotes (1-5). They form good examples of extreme idiosyncratic behavior in the evolution of proteins and it can be said to be a known phenomenon, even if not so named. The evolution of proteins can be very different depending on the class of organisms in which they function. As far as my search has gone it has not been recognized that the concept idiosyncratic applies to a large fraction of proteins as indicated by the presence of many protein sequences similar to eukaryotic proteins in archaea and not bacteria and many others recognized in bacteria but not in the 46 archaea studied here.

RESULTS

BLASTp comparison. BLASTp comparison was made between the human proteins (build 36) and the set of 571 prokaryote species for which protein libraries were available at the start of this work. In this collection there are protein libraries of 46 archaea and 525 bacteria. The matches of each human protein were examined and the best match with a protein of each prokaryote species was determined. The bit scores of the best matches to the proteins of each archaeal species were added and divided by 46 to obtain a measure of archaeal matches. Also the bit scores of the best matches to the proteins of each of the bacterial species were added and the sum divided by 525. A plot (not shown) indicates many cases where either the bacterial average is higher than the Archaeal average or vice versa. There are many examples that have zero scores for archaea or bacteria but significant scores for the other group. A number of tests were made to best bring out this information and the decision was made to describe the maximum values for matches of each human protein in all the archaea or all the bacteria. Fig 1 shows the maximum bit scores with archaea and bacteria for each human protein in build 36.

What is striking in Fig 1 is the number of human proteins that plot on the left edge corresponding to zero bacterial bit score and those along the bottom edge corresponding to zero archaeal score. It is notable that the regions near these line along the axes are empty because the criterion that the expectation be less than 1e-3 eliminates poor matches. Thus the meaning of a zero score is that the match fell below the criterion and cannot be considered to be significant. There are actually 669 human proteins for which the bacterial maximum bit score is zero and the archaeal maximum bit score is greater than 38 (set by the criterion). In addition there are 7906 human proteins for which the archaeal maximum score is zero and the bacterial maximum score is greater than 38. The distribution of the maximum scores for the proteins when the bacterial or archaeal scores

are zero is valuable because it allows us to form an opinion as to which examples have been evidently idiosyncratic in their evolution and this data will be presented after a new scoring method is introduced.

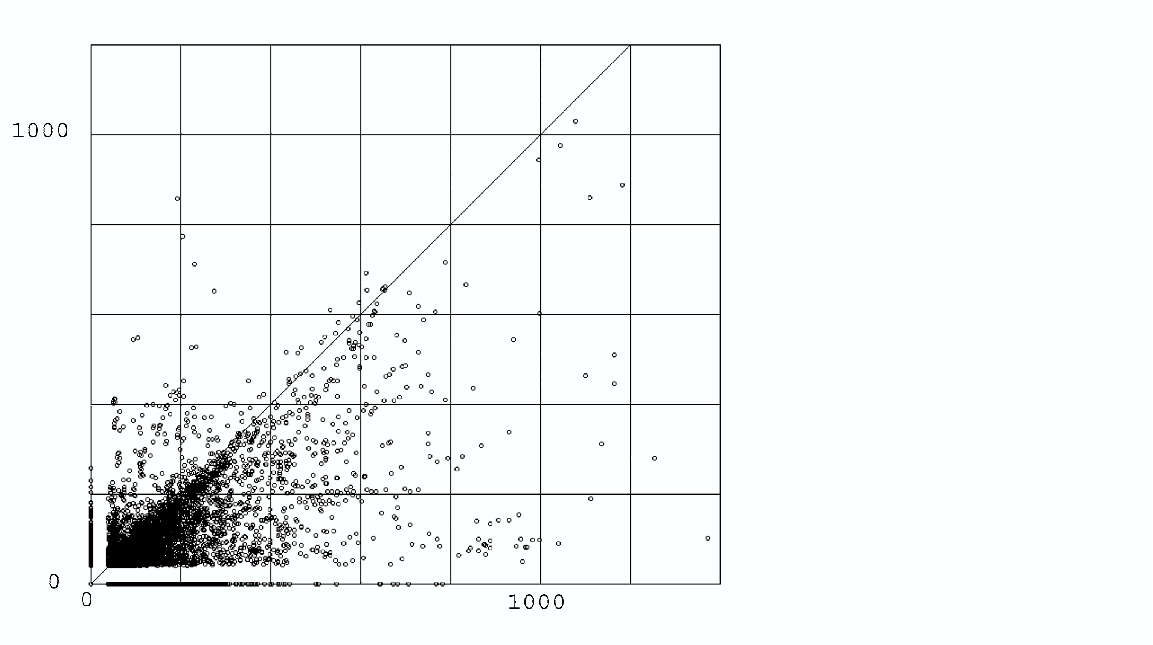

Figure 1 Maximum bit scores for Archaeal vs bacterial proteins matched to human proteins.

The scores are the BLASTp calculated bit scores. Vertically are the maximum archaeal scores and horizontally are the maximum bacterial scores for matches to each human protein. The gap near the axes is due to the elimination of poor matches by the criterion of expectation less than 1e-3. The average of the bacterial maximal scores (121bits) is greater than the average of the maximal archaeal scores (88bits) and the ratio is 1.37, for the cases where neither is zero. Note the large number of points along the axes. These points include 40% of all the 21,440 points plotted.

Scoring by percent match times fraction of length matched. The bit score reported by BLASTp puts weight on longer proteins, but the simple score of percent match times fraction of length reported in the match (FP) avoids this problem and opens up an important new set of observations. In this work the fraction of the length of the prokaryote protein is used in the calculation. The resulting graph, Fig 2, is similar to that with bit scores but shows the presence of many relationships spread out more as this method emphasizes cases where the bit score is intermediate or low. The FP score gives a better image of the scores of archaea when no matches are found with bacteria and vice versa, but the number is the same as for the Fig 1 bit scores. Fig 3 shows the distribution of these scores.

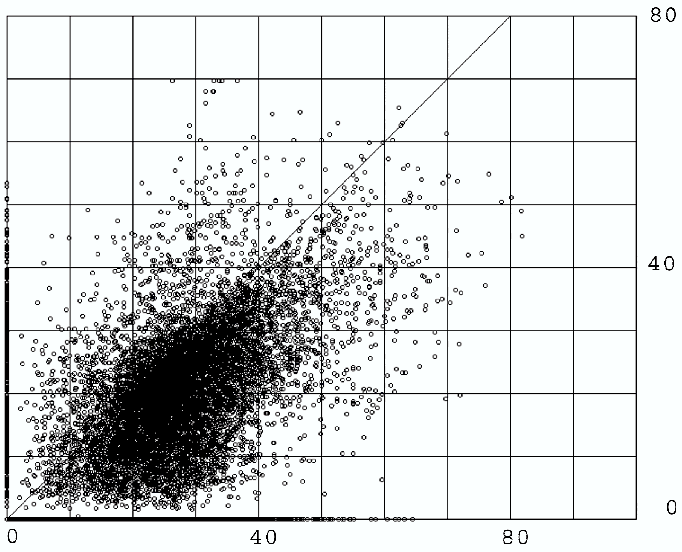

Figure 2 Maximum FP scores for Archaeal vs bacterial proteins aligned with human proteins.

The scores are FP which is the product of the percent amino acid sequence match multiplied by the fraction of the length of the protein in the match Vertically are the maximum archaeal scores and horizontally are the maximum bacterial scores for matches to each human protein. The gap near the axes is due to the elimination of poor matches by the criterion of expectation less than 1e-3. The average of the bacterial maximal scores is greater than the average of the maximal archaeal scores and the ratio is 1.28. As with the Fig1 plot of bit scores there are a number of points on the axes. These points include 40% of all the 21,440 points plotted.

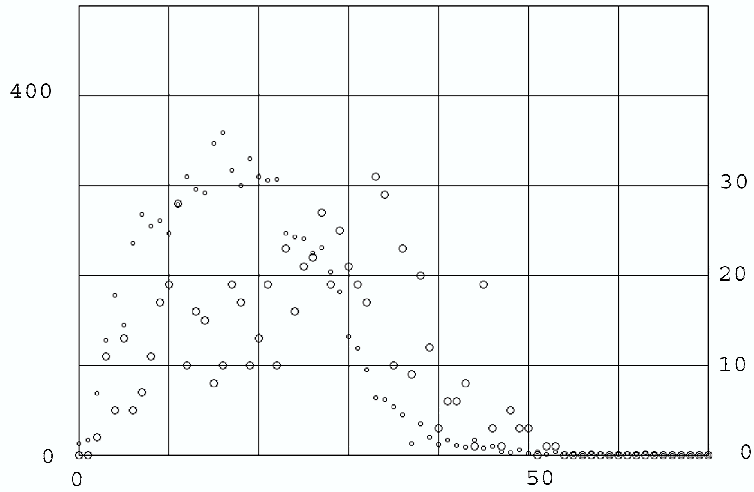

Figure 3. Number of human proteins matching Archaea or bacteria when the other is zero VS. FP score.

Horizontally is the FP score in % representing the maximal archaeal or maximal bacterial score. Vertically is the number of proteins in bins of 1% in FP. The small circles are counts for proteins that do not match archaeal proteins but match bacterial proteins, using the left scale. The large circles are counts for proteins that do not match bacterial proteins but match archaeal proteins using the right scale

Of the 669 archaeal proteins that match human proteins (where these human proteins do not match any bacterial proteins) 256 are ribosomal proteins. These are similarities to human nuclear ribosomal proteins and it is well known these do not occur in bacteria.(6) Table 1 shows the set of archaeal proteins with highest FP score that do not match any bacterial proteins, after removal of the ribosomal proteins. An interesting group is made of human H4 histone proteins including family members A,B,C,D,E,F,G,H,I,J,K,M. Studies have shown that archaeal species use the histone fold region while bacterial species do not have any histones. There are a large number of what are termed histone like bacterial proteins that bind to DNA. The list of bacterial proteins that do not match archaeal proteins includes ribosomal proteins that match human mitochondrial ribosomal proteins (not shown). Interestingly there are 14 cases in which human mitochondrial ribosomal proteins make good matches to both archaeal and bacterial ribosomal proteins (7).

Many of the proteins on Table 1 are expected, for example, the histones and some proteasomal associated proteins. However the DNA directed RNA polymerases etc. must be examples where the required bacterial function was replaced by a protein of different origin or large evolutionary change in the sequence occurred.

Table 1 The highest FP score set of archaeal proteins for cases where the human protein does not match any bacterial proteins, after removal of the ribosomal proteins. The first column is the maximum FP score in %. Then the brief description of the human protein and on the second line the brief description of the Archaeal protein with the maximum score including the species.

```
  53.338   DNA directed RNA polymerase II polypeptide F [Homo sapiens]
DNA-directed RNA polymerase subunit K [Methanosarcina barkeri str. Fusaro]
  52.776   DNA directed RNA polymerase II polypeptide L [Homo sapiens]
DNA-directed RNA polymerase subunit N [Methanopyrus kandleri AV19]
  47.431   DNA directed RNA polymerase II polypeptide E [Homo sapiens]
DNA-directed RNA polymerase, subunit H (rpoH) [Methanocaldococcus jannaschii DSM 2661]
  44.030   TATA box binding protein like 2 [Homo sapiens]
TATA-box binding [Thermofilum pendens Hrk 5]
  43.358   hypothetical protein LOC115939 [Homo sapiens]
Protein of unknown function DUF367 [Methanosaeta thermophila PT]
  42.861   nucleolar protein family A, member 3 [Homo sapiens]
hypothetical protein STS104 [Sulfolobus tokodaii str. 7]
  42.843   proteasome (prosome, macropain) subunit, alpha type, 8 isoform 1 [Homo]
proteasome subunit alpha [Pyrococcus horikoshii OT3]
  41.643   Sm protein F [Homo sapiens]
Like-Sm ribonucleoprotein, core [Methanosaeta thermophila PT]
  39.689   programmed cell death 5 [Homo sapiens]
hypothetical protein Ta0052 [Thermoplasma acidophilum DSM 1728]
  39.594   U6 snRNA-associated Sm-like protein 5 [Homo sapiens]
Small ribonucleoprotein [Methanococcoides burtonii DSM 6242]
  38.971   proteasome (prosome, macropain) subunit, alpha type, 8 isoform 3 [Homo]
proteasome subunit alpha [Pyrococcus horikoshii OT3]
  38.522   nuclear transcription factor Y, beta [Homo sapiens]
histone HMtA1 [Methanothermobacter thermautotrophicus str. Delta H]
  37.558   small nuclear ribonucleoprotein polypeptide E [Homo sapiens]
hypothetical protein Ta0927 [Thermoplasma acidophilum DSM 1728]
  37.453   small nuclear ribonucleoprotein polypeptide F [Homo sapiens]
Like-Sm ribonucleoprotein, core [Methanosaeta thermophila PT]
  37.012   DNA directed RNA polymerase III polypeptide K [Homo sapiens]
DNA-directed RNA polymerase subunit M [Pyrococcus furiosus DSM 3638]
  36.738   DNA polymerase epsilon subunit 4 [Homo sapiens]
hypothetical protein Mlab_0494 [Methanocorpusculum labreanum Z]
  36.539   Lsm1 protein [Homo sapiens]
small nuclear ribonucleoprotein homolog (Sm-like) [Pyrobaculum aerophilum str. IM2]
  35.712   Sec61 beta subunit [Homo sapiens]
proteasome-activating nucleotidase [Thermococcus kodakarensis KOD1]
  35.685   proteasome alpha 3 subunit isoform 2 [Homo sapiens]
proteasome subunit alpha [Methanopyrus kandleri AV19]
  35.549   proteasome alpha 3 subunit isoform 1 [Homo sapiens]
proteasome subunit alpha [Methanopyrus kandleri AV19]
  35.279   proteasome alpha 6 subunit [Homo sapiens]
20S proteasome, A and B subunits [Staphylothermus marinus F1]
  34.947   proteasome alpha 1 subunit isoform 2 [Homo sapiens]
Proteasome endopeptidase complex [Thermofilum pendens Hrk 5]
  34.947   proteasome alpha 1 subunit isoform 1 [Homo sapiens]
Proteasome endopeptidase complex [Thermofilum pendens Hrk 5]
  34.623   eukaryotic translation termination factor 1 [Homo sapiens]
peptide chain release factor eRF/aRF, subunit 1 [Thermofilum pendens Hrk 5]
  34.546   U6 snRNA-associated Sm-like protein LSm8 [Homo sapiens]
Like-Sm ribonucleoprotein, core [Methanoculleus marisnigri JR1]
  34.464   zinc ribbon domain containing 1 [Homo sapiens]
archaeal transcription factor S [Methanosarcina acetivorans C2A]
  34.410   DNA directed RNA polymerase II polypeptide J-related gene isoform 3 [Homo]
hypothetical DNA-directed RNA polymerase subunit L [Sulfolobus tokodaii str. 7]
  34.005   general transcription factor IIB [Homo sapiens]
transcription initiation factor IIB [uncultured methanogenic archaeon RC-I]
  33.956   Lsm3 protein [Homo sapiens]
Like-Sm ribonucleoprotein, core [Methanosaeta thermophila PT]
  33.920   U6 snRNA-associated Sm-like protein LSm7 [Homo sapiens]
snRNP Sm-like protein [Methanobrevibacter smithii ATCC 35061]
  33.828   histone H4 and 14 variants of H4 all match:
archaeal histone [Methanosphaera stadtmanae DSM 3091]
```

```
  33.752   integrin beta 4 binding protein isoform a [Homo sapiens]
translation initiation factor IF-6 [Natronomonas pharaonis DSM 2160]
```

The histone prokaryotic relatives. It is well known that archaea make use of histones for nucleoid structure while bacteria do not. There are many so called histone-like proteins in bacteria involved in DNA binding. Thus it was worthwhile to look for sequence similarity of eukaryotic histone proteins to the collection of 571 prokaryote protein libraries. For this purpose all 113 human genes that are histones or putative histones were listed and the sequence similarities to the prokaryote proteins assayed. The FP scoring method is more effective since the histones are short proteins. The few matches are shown on Table 2, ordered by bacterial protein maximum FP score. Only in the case of the relatively poor matches to H4 histone family members at the bottom of the table does the match occur with a prokaryote histone. While these are significant matches the expectation scores are barely above the criterion being about 5e-4. The other prokaryote proteins are hypothetical or ribosomal proteins or the odd examples listed in the Table 2 footnote. Except for H4 the histone related protein lineages have followed idiosyncratic evolutionary functional pathways while preserving a significant amino acid sequence similarity.

Table 2 Archaeal and bacterial matches to human histones
Col 1 is maximum FP score to the archaeal proteins. Col 2 is the maximum FP score to the bacterial proteins. The second line is the archaeal protein with maximum FP score.[1]

```
  23.144  63.336   H1 histone family, member 5 [Homo sapiens]
     hypothetical protein Mboo_0185 [Candidatus Methanoregula boonei 6A8]
  19.375  56.427   H1 histone family, member 3 [Homo sapiens]
     Ribosomal protein L32e [Candidatus Methanoregula boonei 6A8]
  28.127  51.741   H1 histone family, member 4 [Homo sapiens]
     hypothetical protein Mboo_0185 [Candidatus Methanoregula boonei 6A8]
  17.660  47.333   H1 histone family, member 0 [Homo sapiens]
     Ribosomal protein L32e [Candidatus Methanoregula boonei 6A8]
  24.423  46.235   H1 histone family, member 2 [Homo sapiens]
     Ribosomal protein L32e [Candidatus Methanoregula boonei 6A8]
  21.378  39.023   H1 histone family, member 1 [Homo sapiens]
     Ribosomal protein L32e [Candidatus Methanoregula boonei 6A8]
   0.000  38.822   histone H1 variant [Homo sapiens]
     -
  31.045  36.143   H2A histone family, member Y isoform 1 [Homo sapiens]
     ADP-ribose binding protein [Methanococcoides burtonii DSM 6242]
   0.000  35.909   H1 histone family, member X [Homo sapiens]
     -
  54.857  34.713   PREDICTED: similar to testis-specific histone 2a [Homo]
     hypothetical protein Tpen_0842 [Thermofilum pendens Hrk 5]
  29.185  34.402   H2A histone family, member Y isoform 3 [Homo sapiens]
     hypothetical protein PAE1111 [Pyrobaculum aerophilum str. IM2]
  15.499  34.132   oocyte-specific histone H1 [Homo sapiens]
     Ribosomal protein L32e [Candidatus Methanoregula boonei 6A8]
  19.016  33.887   H1 histone family, member T, testis-specific [Homo]
     Ribosomal protein L32e [Candidatus Methanoregula boonei 6A8]
  26.537  30.043   core histone macroH2A2.2; H2A histone family, member Y2
     hypothetical protein ST2383 [Sulfolobus tokodaii str. 7]
  33.828   0.000   histone H4 [Homo sapiens]  and 14 variants
     archaeal histone [Methanosphaera stadtmanae DSM 3091]
```

1. The matching bacterial proteins were all hypothetical except for 3 examples of human H2A matching Appr-1 ' ' -p processing enzyme family protein [Leptospira interrogans serovar La1 str 56601] and an H2A to the same protein of [syntrophobacter fumaroxidand MP08].

Other Eukaryote comparisons to prokaryote proteins. To ask whether the human protein comparisons are representative of eukaryote proteins, in general, comparisons were made between yeast *Saccharomyces cerevisiae* as well as *Arabidopsis thaliana* protein libraries and the prokaryote protein libraries. The simplest overall comparison of the relationships with these eukaryotic proteins is the number of examples in which there were matches to bacterial proteins and not archaeal proteins and vice versa. The available set of *Arabidopsis thaliana* proteins was compared using BLASTp with the 571 prokaryote proteins. Similarly the set of 5879 yeast (*Saccharomyces cerevisiae*) proteins was compared with the 571 prokaryote protein libraries with the result that there were 3363 matches. Of these 180 matched archaeal proteins but did not match bacterial proteins. Another 1042 yeast proteins matched bacterial proteins and failed to match archaeal proteins. These patterns are quite similar to that for the human protein comparisons as shown on Table 3 where they are expressed as percent of the whole protein library. It appears quite likely that these numbers will change when a larger set of Archaeal protein libraries is examined. Nevertheless they are quite consistent among the three eukaryotes.

Table 3 idiosyncratic percentages

	Total in library	match%[1]	b+ (zero a)%[2]	a+ (zero b)%[3]
Homo Sapiens	34,180	62.8	23.1	1.95
A thaliana	30,480	60.5	22.1	1.65
S cerevisiae	5,879	57.2	17.7	3.06

1/ Percent of eukaryote library to match prokaryote proteins at criterion of expectation less than 1e-3.
2/ Percent of eukaryote library to match bacterial proteins but not archaeal proteins.
3/ Percent of eukaryote library to match Archaeal proteins but not bacterial proteins.

DISCUSSION

The conclusions of this work are summarized in a few paragraphs. The 21440 human proteins that significantly sequence match archaeal or bacterial proteins are very highly conserved and thus their evolutionary precursors have had a long history of carrying out important functions in both the eukaryotic and prokaryotic lineages, over several billion years.

The history of many of these proteins has been idiosyncratic showing relationship for instance to archaeal proteins but not to bacterial proteins (average 2.2%) or to bacterial proteins but not archaeal proteins (average 21%) from Table 3. This was best shown using the FP score which is the product of the percent amino acid match and the fraction of the prokaryote protein length in the match. The clearest view of the relationships came from selecting the best match of a human protein to the proteins of all the bacterial or archaeal species.

The average of all the matches yields the same numbers in the classes but the presence of many poor matches reduces the scores. The reason that there are fewer Archaeal examples is presumably due to the smaller number of archaeal protein libraries in this study. In fact the ratio of the number of bacterial species libraries (425) to the

number of archaeal species libraries (46) is 9.2 compared to the table 3 columns 4 and 5 averages of 21%/2.2% =9.5.

The pattern of idiosyncracy is about the same whether derived from matches to human, *A thaliana* or yeast protein sequence comparisons. The only possibly significant difference is the 3% of the yeast proteins that make archaeal matches and not bacterial matches.

The majority of the human proteins that show any relationship to prokaryote proteins show somewhat better relationships to bacterial proteins than archeaeal proteins. The bit score ratio is 1.37 for the maximum while the maximal FP score ratio is 1.28. If the average of all FP scores between human proteins and bacterial proteins is compared to the same for archaeal proteins the ratio is 1.18. On Fig 2 clearly many proteins fall above the diagonal but the majority fall below it and this pattern extends from the poorest matches to the best. The very best matches of human proteins to bacterial proteins have an FP of 82% while the very best matches of human proteins to archaeal proteins have an FP score of 70%. The result is similar for the bit scores where the best comparison is almost 1400 fo the bacteria to just over 1000 for the archaea. Thus if we took a majority vote among the 21,440 proteins or let the leaders decide the bacteria come out more closely related to eukaryotes than the archaea, on this basis. This result is different from the relationships derived from ribosomal sequences which places the archaea closer to the eukaryotes. The two sets of results should not be considered in conflict since they are both clearly factual and derive from large bodies of data. They should be considered as giving insight into the complexities of the origins and histories of the prokaryotes.

Not much ancient fossil data is available for prokaryotes but their presence as large numbers of individuals and species seems likely. There are fossils of ancient microbial mats and stromatolites going back 3.4 billion years without species identification except for cyanobacteria and a few other morphotypes (8). The organisms of these early periods were the precursors of archaea or bacteria and there may be many surprises among them. They may include a branch to the eukaryotes, with or without passing through archaeal or bacterial precursor species. The difference in the ribosomal RNA evidence and the protein evidence might be able to be used to construct a precursor that combines these features.

Horizontal transfer of genes (HGT) is common in prokaryotes and occurs between prokaryotes and lower eukaryotes (9,10,11) but is rare between prokaryotes and vertebrates, for example. Figure 2 shows a number of examples where the maximum FP score for human proteins matching bacterial proteins is greater than 70% and up to 82%. An FP score is always lower than the % match since the length of match is almost always less than the protein length. All of these also have high scores to archaeal proteins. There are of course many more cases than the maximum values on Fig 2. There are in fact 484 examples with FP greater than 70% including matches to 17 different humans proteins.

The highest FP score is between an unidentified human protein and a proline transporter from a Staphyloccus. It is probably not a case of HGT but a TIGR urinary

infection. The nucleotide sequence XM_942287 (from which the postulated protein XP_947380 was translated) can be aligned only with one of 940 sequenced bacterial genomes: Staphylococcus saprophyticus, subsp ATCC 15305 (Ss15305) which is a common cause of human urinary infection. The alignment is 78% match over 1463 nt of the 1539 length. The alignment with the human genome gives a good match (100% over 1539) with a region called chromosome Un (AADB02028357) and an 84% match from 808 to 1539 on chromosome 5. Both of these are almost certainly artifacts. There are no matches to other sequenced mammalian genomes. The inaccuracy of these matches needs explanation. Perhaps the contamination came from an un-sequenced strain of Staphylococcus closely related to Ss15305.

There are 6 human proteins that make matches with bacterial proteins with FP score greater than 75%. The coding sequences of these 6 protein were searched against the prokaryote DNA genomes using NCBI genomic blast program. Four of them failed to make any significant matches and can be ruled out as examples detectable HGT. Two of them however did significantly match a number of bacterial genomic sequences in known gene regions. They are NADH dehydrogenase (ubiquinone) Fe-S protein 8 (NP_002487) and closely related NADH-ubiquinone oxireductase Fe-S protein 7. In human they both include numerous introns and are obviously not the direct product of HGT. However their coding regions match a number of different bacterial protein genes: 6 for the first and 28 for the second, with typically 47% coverage and 80% identical match. Some mammalian genes for these proteins do not have an exon/intron pattern, for example, in the gray mouse-lemur. There probably have been HGT events among bacteria and higher eukaryotes in these two cases in the not too distant past. However HGT does not appear to have been a significant factor for most of the 21,440 proteins described and does not affect the evidence for the idiosyncratic behavior of protein evolution described.

Methods

The BLASTp comparisons were made using the Blastall program on a 4 processor Mac with OS*X operating system. The raw results are available on DVD for those interested in following up the analysis. The protein libraries were downloaded from NCBI. All of the analysis was done with fortran programs (not available). The graphics were done on a windows operating system Dell computer using Photoshop. Many statements are based on the results of searches carried out by NCBI "genomic blast" (HTTP://www.ncbi.nlm. gov/sutils/genom_tabl.cgi?/)

References

1 Sapp J (2005) The prokaryote-eukaryote dichotomy. Microbiol Mol Biol Rev. 69(2):292-305

2 Sandman K Reeve JN (2005) Archaeal chromatin proteins. Curr Opin Microbiol 8(6)656-61

3 Poole AM Penny D (2007) Evaluating hypotheses for the origin of eukaryotes. Bioessays. 29(1) 7 4-84.

4 de Duve C (2007) The origin of eukaryotes Nat Rev Genet 8(5):395-403

5 Gupta RS Golding GB (1996) The origin of the eukaryote cell. Trends Biochem Sci. 21(5):166-71.

6 Cubonova L Sandman K Hallam SJ Delong EF Reeve JN (2005) Histones in crenarchaea. J Bacteriol 187(15):5482-5

7 Lecompte O Ripp R Thierry JC Moras D Poch O (2002) Comparative analysis of ribosomal proteins in complete genomes. Nucleic Acids Res 30 5382-90.

8 Allwood AC Walter MR Kamber BS Marshall CP Burch IW (2006) Stromatolite reef from the Early Archaean era of Australia. Nature 441:714-8

9 Hotopp JC et al (20 authors) (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317:1753-6.

10 Nosenko T Bhattacharya D (2007) Horizontal gene transfer in chromovalveolates. BMC Evol Biol 7:173

11 Rogers MB Patron NJ Keeling PJ (2007) Horizontal gene transfer of a eukaryotic plastid targeted protein gene to cyanobacteria. BMC Biol 5:26.