PRECISE SIMILARITY OF MANY HUMAN PROTEINS TO PROTEINS OF PROKARYA

Roy Britten
California Institute of Technology 101 Dahlia Corona del Mar Ca 92625

Abstract

Proteins originated in early forms of life and have long survived, because they have always been required. Some recognizably similar proteins are found in all sequence comparisons between species, no matter how distant, including prokaryotes and eukaryotes. Reported here are observations on the relationships of human proteins to the proteins of 458 prokaryotes for which protein libraries are available. Each of these libraries includes a protein that matches a human protein with a BLAST score of 573 or more, indicating excellent conservation of certain amino acid sequences. A majority of these proteins also match a yeast protein and other eukaryote proteins with comparable accuracy, indicating that protein conservation is responsible in most cases rather than the horizontal transfer (HGT) between eukaryotes and prokaryotes. Rare examples of HGT are apparently also seen.

Very many significant matches are seen as the criterion is opened, including 20,596 human proteins that match at least one prokaryote protein with expectation of 10^{-3} or less. Individual prokaryote proteins accurately match parts of many modern human proteins that have a wide range of functions showing directly that many proteins of different functions have evolved from an ancestral protein by duplication, rearrangement and divergence of function. The implication is that most or all modern proteins derive from the proteins of the last common ancestor with prokaryotes through many such events.

Introduction.

Past evidence for early origin of proteins comes from many studies, for example from the sharing of structural folds of proteins among the three super-kingdoms examined by Yang et al (1) . They defined and studied 1244 protein fold superfamilies (FSF). Of these 654 occurred in at least one species of each of archaea, bacteria or eukarya among the 174 species they utilized. They even found 49 FSF present in all three super-kingdoms among their examples. This evidence clearly indicates sharing of some functional regions among the proteins of all living forms. It is not yet possible to draw the tree of relationships back to the last common ancestor of eukarya and prokarya (LCAEP), presumably because of extensive horizontal transfer (HGT) among the prokarya (2). Since HGT has occured in recent times between primitive eukarya and prokarya (3) it is likely that HGT occurred between the early eukarya and prokarya. Thus the effective last common ancestor between eukarya and prokarya genes was probably later than the occurrence of the first eukarya in the fossil record and the date is unknown.

In every species tested, including human (4), the proteins are almost all related to other proteins of the same organism, showing the extent of past duplication. The

percentage of the set of proteins that match others in the same set is always large, ranging from about a third of all proteins in some prokaryotes with few proteins to almost all proteins in most species examined.

The approach here is to compare human proteins (representing eukaryotes) with the proteins of many prokaryotes, using BLASTp (5). Comparisons are made to human proteins because the human library is nearly complete and well studied. Comparisons are reported at varied criteria of precision to yield a more full description. A number of human proteins make almost full length precise matches to prokarya proteins.

The date of the branching between prokaryotes and eukaryotes has been variously estimated. The earliest known eukaryotic fossil is about 1,500 MYA (6,7). There are traces of eukaryotic steranes in 2700 million year old Australian shales (8) that may imply the existence of eukaryotes at that time. Based on protein comparisons an estimate of 2 billion years was made (9) later extended to 2.5 billion years(10). Thus 2.5 billion is a useful round number adding up to 5 billion years summing the time in both lineages, though no strong argument could be used against smaller estimates depending on how long a period of massive HGT persisted between prokarya and the eukarya as they advanced in form and complexity. The increasing need for complex 5' control regions and many transcription factors in eukaryotes probably reduced the significance of HGT from prokaryotes, because in order to be useful a newly arrived gene would need to develop or share such control features.

A particularly significant observation is that individual prokaryote proteins match well with many human proteins which have a variety of different functions. These specific cases demonstrate that many proteins have evolved by duplication and divergence of function in the eukaryote lineage since the LCSE (last common sharing event). Tests with a large library of prokaryote protein sequences containing the protein libraries for 458 prokaryote species indicates that more than half of human proteins still retain significant though weak sequence relationship to prokaryote proteins. The overall process has been mutualism of protein evolution and organism evolution, each totally dependant on the other but with independent time courses.

RESULTS

Best human protein matches to many prokaryote species proteins

A library of the proteins is available for each of a set of 458 prokaryote species including 28 archaea. There are about 1,436,050 proteins in this collection. These protein sequences were compared with the human protein library build 36, using BLASTp (5). In the first step in the analysis the maximum score was listed for a protein from each prokaryote species matching any human protein. BLASTp score values for these maxima ranged from 595 to 1373, all extremely good scores. These matches have better than 50% amino acid sequence match and typically include most of the length of the human protein and the prokaryote protein. The average maximum score per species for archaea was 817.2 and for all others was 1027.4. Table 1 lists the best of them, including only 22

human proteins since all but 5 of them make the best match with the proteins of many prokarya. One human protein, NP_000245, `5-methyltetrahydrofolate-homocysteine methyltransferase` is the best match for 124 of the species.

Table 1 The highest scoring human proteins matching the proteins of The 458 prokarya protein sets

Species[1]	Max score[2]	ID[3]
69	1137	NP_000161
124	1373	NP_000245
1	663	NP_000928
6	851	NP_000929
3	1100	NP_001024
9	612	NP_001677
1	595	NP_001681
57	872	NP_001866
44	961	NP_002188
1	833	NP_002853
3	808	NP_002854
9	692	NP_004125
16	1183	NP_004332
2	874	NP_005600
2	682	NP_009057
2	1254	NP_035194
1	1113	NP_036525
15	927	NP_038203
13	1030	NP_060040
4	634	NP_068746
73	1165	NP_071504
1	1090	NP_497341

1. Number of prokarya species for which this human protein gave the best match
2. Maximum score for this human protein
3. Identifier for this human protein.

The chance of accidental sequence match or convergence is negligible for the typical length and precision for the best matches shown in Table 1. That leaves almost no doubt that they are derived from a last common ancestor. These examples are simply the best matches of very many good matches. Table 2 shows how many different human proteins find matches as a function of match quality as measured by the BLAST score. The last line is for matches better than the minimum significant match (expectation 10^{-3}). This result shows that a majority of human proteins are related weakly but significantly to prokaryote proteins, suggesting that human proteins are primarily derived from very early proteins.

Table 2 - Number of human proteins matching the proteins
of the 458 prokaryote species as a function of BLASTp score

Score limit[1]	Number[2]
1000	12
800	53
660	106
600	153
400	514
200	1964
30	20,596

1/ lower limit of BLASTp scores
2/ The number of different human proteins
(build 36) that match with this score or higher

Possibilities of horizontal transfer

An issue is whether these good matches are the result of effective conservation of the amino acid sequence over the 2 billion year period since the last common ancestor of eukarya and prokarya (LCAEP) in both lineages or has horizontal transfer occurred. One way to approach this question is to examine the distribution of good matches to these proteins among a number of species of eukaryotes. For this study the 106 human proteins have been selected that match prokaryote proteins with scores of 660 or better.

Table 3 Matches to eukaryotic species proteins of
The set of 106 human proteins.

better scores[1]	>100[2]	species
22	58	Saccharomyces cerevisieae
26	51	Arabidopsis Thaliana
68	101	Cenorhabditis elegans
71	100	Drosophila melanogaster
82	103	Gallus gallus
78	102	Dania rerio

1/ The number of cases in which the best match to the eukaryote species protein has a higher score than the best match to a prokaryote protein.

2/The number of cases in which the score was greater than 100 for a match to a protein from the eukaryote species.

Table 3 shows that moderately conserved versions of almost all of the 106 proteins are present in the set of proteins of these 6 species that are a small sample of the eukarya. The bottom 4 species (animalia) include proteins that match very well to two thirds or more of the set of 106 proteins. Almost all the 106 proteins match moderately well to proteins of the animal species (>100 score). More divergence of these proteins has occurred in the evolution of the plant lineages and the fungi leading to the modern yeast.

Only a few of them (22 and 26) have a better score in matching a yeast or plant protein than the prokaryote match. Fewer are recognizable (58 and 51) with a score of 100 or more.

Table 4 shows that there is some variability in the selective history of these proteins. Principally the 4 animalia show high scores with a few exceptions. For example certain of the 53 human proteins matches with individual animal species may show lack of recognition or low scores for best matches to proteins while the others show good matches. However In a majority of cases the yeast protein score is low and usually the plant score is also low for the same human protein. In more than a third of the cases the score for all 6 of the species shown on table 4 is greater than 300 - a very good match. For only two of the proteins, placed at the bottom of the list, there are very poor matches to all of the six species proteins. These two are listed as human hypothetical proteins and may be examples of transfer from a prokaryote to the human lineage, without the transcription regulatory system having developed. These are the only possible examples of HGT to eukarya during evolution of the animalia, on table 4. A table like table 4 of all of the scores of the 106 proteins (with max scores of 660) with the 6 eukaryotic species was assembled (not shown) and these two are the still the only examples of possible recent HGT. Some of the poor matches are significant, and need further exploration.

To further explore significance the high sequence match scores each of the 53 human proteins was matched with all of the proteins of the 458 prokaryote species and the number of these species was counted that included proteins with good matches (expectation less than 1e-100). As shown in the last column of Table 4 good matches are spread widely among the different species of prokarya. Practically all of the 53 match well with proteins of many of the 458 prokarya species. Five even find good matches in a majority of the prokarya species. The table of the 106 species (not shown) gives a very similar pattern . The average over the 106 human proteins was matches with proteins of 170 different prokarya species at this very high criterion (expectation 1e-100). Sequence relationships indicating highly conserved proteins are not rare among the prokarya or eukarya.

Table 4 Best Scores (BLASTp)of 53 human proteins (that score over 800 with Prokarya) vs proteins of 6 eukarya species.

Name	prok	Sc	At	Ce	Dm	Gg	Dr	num
NP_000083	878	56	82	1290	1193	1967	1093	15
NP_001035194	1254	419	139	549	922	315	400	27
NP_001837	818	44	73	1323	1196	1553	1205	14
NP_001024	1100	1084	1133	1203	1241	1157	1445	92
NP_001838	844	45	75	1383	1238	1523	1113	16
NP_071504	1165	1215	382	1530	1625	373	1985	389
NP_000246	940	0	0	1035	0	1317	356	128
NP_542196	863	57	86	831	813	2843	2261	18
NP_001866	1106	1491	766	1415	1420	1103	1486	407
NP_002854	957	762	707	1204	1255	1525	1437	195
NP_000084	920	61	97	857	835	3334	2364	17
NP_004332	1183	2082	734	2435	2776	1805	3430	407
NP_000245	1373	58	82	1632	54	650	63	227
NP_203699	955	66	95	1503	1333	1887	1609	15
NP_203700	954	66	95	1499	1333	1888	1608	15
NP_000079	939	57	93	842	833	1729	2410	21
NP_060040	1035	1172	1417	825	1524	152	897	58
NP_542197	863	57	86	831	813	2938	2478	18
NP_000161	1137	993	1103	1127	1273	1738	1573	198
XP_001129650	991	1120	1369	759	1460	152	835	54
NP_000929	857	1395	1568	1932	2085	2325	1778	33
NP_112730	862	53	85	1209	1085	1404	1163	14
NP_000082	862	53	85	1310	1182	1539	1196	14
NP_038203	967	1199	1170	1272	1349	2130	471	140
NP_001089	930	1048	223	1162	1148	1478	1340	13
NP_112733	862	53	85	1078	967	1236	1163	14
NP_002152	967	1199	1170	1272	1349	2130	471	140
NP_000911	1165	1215	382	1530	1625	373	1985	389
NP_112734	840	53	81	1058	954	1162	1161	14
NP_056534	879	50	84	825	839	2055	1918	17
NP_000085	1047	51	86	1141	1092	1377	1031	17
NP_036525	1113	665	1249	1114	1249	0	1617	139
NP_542412	873	50	89	826	786	2281	2598	17
NP_542411	875	50	89	825	787	2423	2344	17
NP_378667	844	45	75	1383	1238	1523	1113	16
NP_000370	868	0	1190	1149	1366	2006	1837	63
XP_497341	1164	232	113	439	521	840	798	48
NP_542410	873	50	89	825	786	2285	2586	17
NP_002853	982	752	707	1213	1279	1470	1539	195
NP_004125	834	834	794	969	994	1200	1147	440
NP_000081	858	63	91	913	851	1563	1795	21
NP_000384	895	49	86	840	821	2592	2250	21
NP_005600	999	751	712	1233	1286	1441	1489	195
NP_078966	814	279	92	371	1020	262	199	4
NP_001836	878	50	79	1424	1304	2409	1343	16
NP_000080	889	48	82	791	786	2125	2042	20
NP_149162	907	56	85	876	857	1847	2564	21
NP_001835	930	56	85	878	827	1886	2683	21
NP_000486	951	66	95	1504	1335	1894	1612	15
NP_002188	996	231	1116	1148	1236	1592	399	273
NP_001845	863	57	86	831	813	2897	2261	18
XP_947380	851	53	0	83	94	88	97	131
XP_292122	904	46	0	48	45	0	49	31

Columns: 1 human ID; 2 prokarya best score; 3 Yeast score; 4 plant score; 5 C elegans score;
6 D melanogaster score; 7 Chicken score; 8 Fish score.
9 number of prokarya species with a protein matching at expectation less than 1e-100

There has been a suggestion that if a protein is present among vertebrates and absent from invertebrates it might have been the result of horizontal transfer (11). For the 106 cases, not shown, there are no cases in which there is a low score for all of yeast,

plant and Ce the representatives of the non-vertebrates, except for the two cases mentioned above which have low scores with all 6 species. For other cases there is no suggestion of HGT after the branch between the lineages leading to vertebrates and non-vertebrates. A study of HGT among eukaryotes (3) lists many cases of transfer of proteins from prokaryotes to lower eukaryotes such as Diplomonads and fungi and an example of Agrobacterium genes to the plant Nicotiana. In some but not all cases phagotrophy was a likely cause. No cases are described of inter-domain transfer to vertebrates or HGT among vertebrates. Clearly the cases of amino acid sequence similarity we have observed between prokaryotes and human proteins are not likely due to recent HGT, except for the two mentioned above. It is not surprising that these 106 well conserved proteins are typically well preserved in the animal species, as shown for the best 53 on table 4. The low scores or non recognition of a set of them in yeast and plant presumably reflect the differences in the needs for some of the functions of these proteins in their evolutionary processes, but could be due to ancient HGT, as the early animal ancestors evolved .

These data lead to the conclusion that the protein sequence similarities are due to shared ancestry of the proteins. However the effective time of existence of that ancestor is not necessarily the date of separation of lineages leading to present day prokaryotes and eukaryotes (LCAEP). Fossils and chemical traces may establish the earliest known eukaryote as about 2 billion years ago (6,7). However the Andersson (3) study indicates many events of inter-domain HGT to lower eukaryotes. It is easy to postulate that the rampant interspecies HGT among prokarya was equally significant in prokarya evolution in the distant past. Also in the early days of eukarya evolution gene transfer may have crossed the domain boundary to eukaryotes as the boundary was forming. How long after that it continued to be rampant to and among eukaryotes is purely a matter for speculation. It is likely that the rate of HGT was retarded long before the vertebrates originated, due to specialization of tissues and richer requirements for gene expression control. In early eukarya newly acquired genes were not useful until control regions and effective trans regulatory factors were developed.

A rating method that recognizes good matches among shorter proteins. The BLAST score puts emphasis on the length of the match and thus selects longer proteins. The rating method used in this section called FP uses the product of the fraction of the length of the prokarya protein included in the match times the percent match. This FP rating favors the accuracy of the match and the coverage of the prokarya protein by the region of similarity with the human protein. There are 280 different human proteins that match prokarya proteins with a rating of 55 or more. The matching pair of human and prokarya with the largest FP among the proteins of each species were selected. Due to a requirement that the FP rating be greater than 55 there are 452 species protein libraries in the list, out of the 458. However there were only 37 different human proteins in these matches, as shown on table 5. The small number is due to the fact that the prokarya share very many similar proteins among different species, as shown also in Table 1, using the BLASTp scoring method. The average FP was 67.9 for these high scoring matches. That would typically be the result of about 91% of the protein length in the match and 74% of amino acids matching. Two of these proteins also occur on table 1, thus the FP rating

and BLASTp score methods have a large degree of independence in the matches they select.

Interestingly the human proteins were almost always slightly larger than the prokarya proteins in these best matches. In only 14 out of the 452 the prokarya protein was slightly larger and the average ratio of the protein lengths was 1.07 favoring the human proteins. In only two cases the human protein was very much longer. Thus there is a class of human proteins of a few hundred amino acid length with great similarity to prokarya proteins in length and amino acid sequence. Table 5 lists the thirty seven different human proteins and their functional descriptions. Surprisingly there are 7 examples of human proteins described as hypothetical or predicted among the 37. To further examine the relationships the 37 proteins listed on table 5 have been compared with the yeast proteins and 21 have good ratings. Specifically the two hypothetical genes have ratings of 72.4 and 73.2. However the 5 predicted proteins near the bottom of Table 5 include 3 that have poor ratings (19,21,50) and 2 that are absent from the yeast protein matches. These 2 are also missing from the fish D rerio protein matches and could be considered as examples of HGT to the human lineage, after the branch from the fish lineage, that have not had time to become a fully functioning part of the human proteome. All of the other proteins on table 5 have high ratings to D rerio proteins except for 2 of the predicted proteins (ratings 31 and 27) [XP_001132969 and XP_947380]. One of these was previously identified on table 4. Otherwise there is no evidence of HGT for this set of proteins that have matches with maximum ratings to the prokarya proteins.

Table 5 The functions of the 37 human proteins that have the best ratings with prokarya species libraries.
Column 1: abbreviated name (as NP_000150 for example)
Column 2: FP rating.

150	70.6	glutaryl-Coenzyme A dehydrogenase isoform a precursor [Homo sapiens]
246	64.8	methylmalonyl Coenzyme A mutase precursor [Homo sapiens]
523	67.8	propionyl Coenzyme A carboxylase, beta polypeptide [Homo sapiens]
662	66.6	class III alcohol dehydrogenase 5 chi subunit [Homo sapiens]
678	67.5	S-adenosylhomocysteine hydrolase [Homo sapiens]
1025	66.1	ribonucleotide reductase M2 polypeptide [Homo sapiens]
1491	60.8	GDP-mannose 4,6-dehydratase [Homo sapiens]
1677	80.9	hypothetical protein LOC399827 [Homo sapiens]
1684	59.6	hypothetical protein LOC400576 [Homo sapiens]
2037	65.8	glyceraldehyde-3-phosphate dehydrogenase [Homo sapiens]
2159	71.1	isocitrate dehydrogenase 2 (NADP+), mitochondrial precursor [Hs]
2487	76.0	NADH dehydrogenase (ubiquinone) Fe-S protein 8, 23kDa (NADH-coenzyme Q reductase) [Homo sapiens]
2504	60.5	nucleoside-diphosphate kinase 3 [Homo sapiens]
2622	66.1	phosphogluconate dehydrogenase [Homo sapiens]
2970	60.2	sterol carrier protein 2 isoform 1 proprotein [Homo sapiens]
3840	59.2	succinate-CoA ligase, GDP-forming, alpha subunit [Homo sapiens]
4037	68.1	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit precursor [Homo sapiens]
4125	58.7	tumor suppressor candidate 1 [Homo sapiens]
4484	72.5	olfactory receptor, family 13, subfamily D, member 1 [Homo sapiens]
4526	58.1	myelin transcription factor 1 [Homo sapiens]
5462	58.5	glucosamine-6-phosphate deaminase 1 [Homo sapiens]
5608	63.4	ribosomal protein S14 [Homo sapiens]
5800	60.7	peroxiredoxin 2 isoform a [Homo sapiens]
5887	65.8	isocitrate dehydrogenase 1 (NADP+), soluble [Homo sapiens]
5902	70.7	methionine adenosyltransferase II, alpha [Homo sapiens]
9034	64.9	NADH dehydrogenase (ubiquinone) flavoprotein 1, 51kDa [Homo sapiens]
55116	72.6	iron-sulfur cluster assembly enzyme isoform ISCU1 [Homo sapiens]
66953	66.0	peptidylprolyl isomerase A isoform 1 [Homo sapiens]
71415	68.1	methylcrotonoyl-Coenzyme A carboxylase 2 (beta) [Homo sapiens]
77718	75.8	NADH-ubiquinone oxidoreductase Fe-S protein 7 [Homo sapiens]
130141	64.5	PREDICTED: hypothetical protein [Homo sapiens]
132969	69.6	PREDICTED: hypothetical protein [Homo sapiens]
292035	62.3	PREDICTED: similar to olfactory specific medium-chain acyl CoA synthetase [Homo sapiens]
892022	72.2	nicotinamide nucleotide transhydrogenase [Homo sapiens]
932047	62.7	PREDICTED: similar to Phosphoglycerate mutase 1 (Phosphoglycerate mutase isozyme B) (PGAM-B)
947380	81.3	PREDICTED: similar to CG6723-PA [Homo sapiens]
998760	79.5	iron-sulfur cluster assembly enzyme isoform ISCU2 precursor [Hs]

The many functions of human proteins derived from a single prokaryote protein. One *Mycoplasma genitalium* (Mg) protein identified as NP_072883 (HMW2 cytoadherence accessory protein) matches 345 human proteins of the KGMV library, (see methods). Table 6 lists 49 human proteins of the KGMV library with the best matches (expectation better than 1e-22). The variety of human protein types with similarity to one *Mycoplasma genitalium* (Mg) protein gives insight into human protein evolution. Table 6 lists just the best matches from the KGMV library (see methods) out of 345 human proteins that owe their origin, at least in part, to a protein of a last common ancestor shared with *Mycoplasma genitalium*. The fraction of the length of the Mg protein that matches these human proteins at high score is shown in fig 1. Only two match almost full length, while a few match nearly full length. The others match shorter regions all overlapping a central region. In the long history of the eukaryotic lineage leading to apes

since our last common ancestor with prokaryotes parts of this protein sequence have been used for many protein functions.

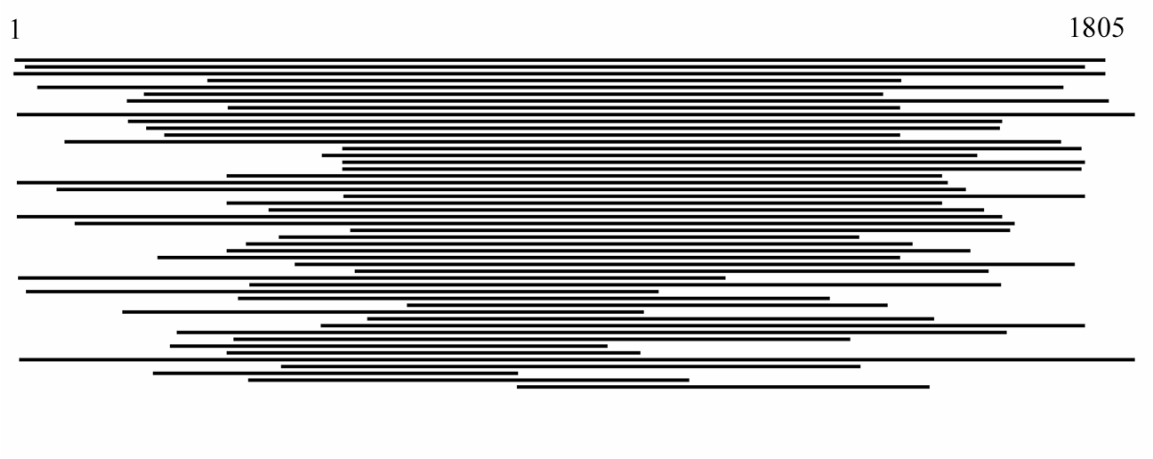

Figure 1 The regions of an Mg protein matching 49 human proteins.
From left to right is the full length (1-1805 amino acids) of the Mg protein NP_072883 (HMW2 cytoadherence accessory protein). The lines show the regions reported by BLASTp for each of the 49 matches which are the same set of human proteins, ordered from top to bottom, by score, as in Table 6 that lists their names. Table 6 also lists the expectation calculated by BLASTp.

Table 6 The human proteins matching at expectation 10^{-22} or better a *Mycoplasma genitalium* protein (NP_072883)

ID[1]	EXP[2]	Description
1813	-59.4	centromere protein E, 312kDa (CENPE)
2078	-56.7	golgi aut, golgin subfamily a, 4 (GOLGA4)
4487	-56.0	golgi aut, golgin subfamily b, macrogolgin
2474	-47.7	myosin, heavy pp 11, smooth muscle (MYH11)
4239	-47.7	thyroid horm recept interactor 11 (TRIP11)
3566	-47.4	early endosome antigen 1, 162kD (EEA1)
7186	-46.7	centrosomal protein 2 (CEP2)
5964	-45.1	myosin, heavy pp 10, non-muscle (MYH10)
16343	-44.4	centromere p F, 350/400ka (mitosin) (CENPF)
181453	-44.0	GRIP and coiled-coil domain cont 2 (GCC2)tv1
201383	-43.0	plectin 1, intermed fil.bind p 500kDa (PLEC1)
2473	-42.7	myosin, heavy pp 9, non-muscle (MYH9)
3292	-42.5	translocated promoter region (TPR)
2470	-42.5	myosin, heavy pp 3, skm, embryonic (MYH3)
3802	-41.0	myosin, heavy pp 13, skm (MYH13)
2471	-39.7	myosin, heavy pp 6, cardiac musc, alpha
257	-39.5	myosin, heavy pp 7, cardiac musc beta (MYH7)
17534	-39.2	myosin, heavy pp 2, skm, adult (MYH2)
4415	-39.0	desmoplakin (DSP)
6185	-38.7	nuclear mitotic apparatus protein 1 (NUMA1)
17533	-38.5	myosin, heavy pp 4, skm (MYH4)
2472	-37.0	myosin, heavy pp 8, skm, perinatal (MYH8)
5963	-36.7	myosin, heavy pp 1, skm, adult (MYH1)
182946	-36.5	ninein (GSK3B interacting p) (NIN), tv 5
147171	-35.7	A kinase (PRKA) anchor p (yotiao) 9 (AKAP9)

```
  24581  -35.0  chromosome 6 open read frame 60 (C6orf60)
   7018  -32.7  centrosomal protein 1 (CEP1)
   2956  -32.7  restin ( intermediate filament-associated)
  18003  -32.3  uveal aut with coiled-coil domain
                and ankyrin repeats(UACA)
  24729  -31.0  myosin, heavy polypeptide 14 (MYH14)
 182926  -30.7  kinectin 1 (kinesin receptor) (KTN1)
 206886  -29.0  sarcoma antigen NY-SAR-41 (NY-SAR-41)
  20242  -28.4  kinesin family member 15 (KIF15)
   2705  -28.0  periplakin (PPL)
   5895  -27.2  golgi aut, golgin subfamily a, 3 (GOLGA3)
  24513  -27.0  FYVE and coiled-coil domain cont 1 (FYCO1)
   4850  -27.0  Rho-assoc, coiled-coil c p kinase 2 (ROCK2)
 178040  -26.2  RAB6 interacting p 2 (RAB6IP2), tv epsilon
  20770  -26.0  cingulin (CGN)
  16195  -25.4  M-phase phosphoprotein 1 (MPHOSPH1)
   6031  -25.0  pericentrin 2 (kendrin) (PCNT2)
   5732  -24.7  RAD50 homolog (S. cerevisiae) (RAD50), tv 1
   2077  -24.5  golgi aut, golgin subfamily a, 1 (GOLGA1)
   5406  -24.3  Rho-assoc, coiled-coil cont p kinase 1(ROCK1)
 183380  -23.4  dystonin (DST), transcript variant 1
   1988  -23.4  envoplakin (EVPL)
  24704  -23.0  chromosome 20 open read frame 23 (C20orf23)
  15687  -22.0  filamin A interacting protein 1 (FILIP1)
   3176  -22.0  synaptonemal complex protein 1 (SYCP1)


Col 1 Identification is NM_ followed by requisite zeros
      and the number; giving coding and protein sequences.
Col 2 log (base 10) of the expectation quoted by Blastp.
Abreviations p: protein; aut: autoantigen; pp:polypeptide;
tv:transcript variant;assoc: associated; cont: containing;
 skm: skeletal muscle
```

More details of the relationships. This Mg protein (NP_072883) was also compared with the build 36 human protein library listing 34,180 proteins and about a hundred of these proteins matched with expectation 10-21 or better. An attempt to associate these sequence similarities with protein domains was made using Pfam search (pfam.janella.org). No domains were recognized in NP_072883 while six domains were recognized in the three top human proteins listed on table 6. They were all distinct from one another and all relatively short. The long conserved regions shown on Fig 1 cannot easily be connected to domains and the conclusion is that some other important aspect of protein sequence or structure has been conserved over billions of years. Further work would be required to identify the significance of so long a region..

NP_072883 was also compared with the proteins of the fish D rerio and Drosophila melanogaster. The results in both cases were similar to the results of the comparison with human proteins including many hundreds of matches (499 for Drosophila) and a comparable number with expectations less than 10-20. There were a comparable number of matches recognized in rice proteins (529 total) but fewer that

matched with expectation 10-20 or less. There were many matches with yeast (Saccharomyces cerevisiae) but only 13 with expectation 10-20 or less. There is no way that these results could be explained by a horizontal transfer from prokarya to eukarya but a proposal could be made that one of the genes of the eukarya might have been transferred to Mycoplasma genitalium or one of it's relatives. To test this the Mg protein (NP_072883) was compared with all of the proteins of the 458 prokarya. There were one or two close matches and 199 matches with expectation less than 1e-19. The presence of so many similar proteins among many of the prokarya pretty well rules out such a transfer in recent times. It does prevent the identification of the potential precursor protein of the many eukaryote and prokaryote proteins, leaving no doubt that such a protein existed. This data helps to fill in the view that the origin of many eukaryotic proteins occurred in the last common ancestors of bacteria and eukaryotes (LCAEP) or in an early sharing event. Also it suggests that the set of related proteins shown on table 5 and 6 had their origins early in the lineage leading to apes, many as early as the branch leading to insects.

Discussion

Relationships and possible late common protein ancestors

It seems a common opinion that all eukaryote proteins derive from a small number of proteins of early forms (e.g. 12) and this view is likely correct since no serious alternatives are known. Earlier it was shown that almost all human proteins are the result of duplication(4) many of which were ancient events leading to the suggestion that human proteins were the result of duplication, divergence, rearrangement and the evolution of new functions. This work shows that many of the human proteins have amino acid sequence similarity to prokaryotic proteins. The fact (Table 2) that 20,596 human proteins have recognizable similarity (expectation 10-3) to proteins of prokarya supports the view that human proteins are the product of a long process of protein replication and divergence that started with last common protein ancestor shared with prokaryotes. The precise amino acid sequence relationships between the proteins of eukarya and prokarya reported here leave no doubt about the existence of a common ancestral origin.

The time of existence of that ancestor and the occurrence of horizontal gene transfer have been partially examined . Table 4 lists the scores of the 53 best matching human proteins to those of individual species of prokarya. It also lists the best scores for the same human proteins with proteins of 6 eukarya, representing fungi, plants and animalia. With few exceptions the animalia scores are high. In only two cases listed at the bottom are they all low. In these cases, which are hypothetical human proteins, the scores for yeast and plant are also low suggesting possible horizontal transfer after the branch from the human lineage to the lineage leading to modern fish. In one other case described in the section describing a different match quality rating system there is also such a suggestion. These three cases need further examination. Beyond these three cases there is no suggestion of a last protein common ancestor that occurred after the branches to plants and fungi.

In about half of the 53 examples listed on Table 4 the score for yeast or plant proteins or both is high. For the 106 cases examined (not shown) both the yeast and plant scores are over 100 and in 66 cases either plant or yeast proteins score over 100. The examples in which yeast or plant score less than 100 there are two alternatives: the protein diverged or was lost in the lineage of the modern species or it was never present. Further examination of the proteins of many species might resolve this question and establish whether or not HGT occurred. The best that can be said with the current data is that in about half the 106 cases HGT might have occurred after the branch from the vertebrate lineage to yeast and plant lineages but there is no direct evidence that it did occur.

The length of conserved regions and domains

The length of the conserved regions for an example is shown in Table 6 and Fig 1 and in all cases is longer than typical functional modules, which have a mode value of about 100 residues(12, fig 2). A few examples have been examined in which long regions almost the full length of the prokaryote protein have been accurately conserved, often with about half of the amino acids matched. In one example PFAM identifies a single domain 423 residues long but the match between the human and the prokarya (Pm) protein is 962 residues long, with a 46% match. In some other cases (not described) the matches are longer than the domains identified by PFAM. The observations clearly show that regions of proteins extending beyond known domains are matched. Domains are only part of the story of protein function, and evolutionary conservation.

The many human proteins matching a single prokarya protein ancestor. The data of Table 6 show clearly that many human proteins match well to a single prokarya protein and the conclusion is that they derived in part from a common ancestral protein. This is not an isolated case, simply an example chosen because of the number of human related proteins. For example there are 11 Bx (*Burkholderia xenovarans*) proteins that each match 20 human proteins with an expectation of 10^{-20} or less. There are also 6 Bx proteins that each match 47 human proteins, most involving the DEAD box. All of this is what would be expected if the human proteins were the end product of a long period of evolution of proteins with new and old functions that depended on duplication, rearrangement, combination of useful parts, divergence and selection.

Evolution of proteins and species

The replication to form new protein types is a very much slower process than the replication of organisms or the creation of new species. Both proteins and organisms replicate and diverge and undergo selection. Metazoons require tens of thousands of proteins and complex systems of regulation of their expression in many different cell types so that each individual develops and goes through a life cycle meeting procreation, ecological and social requirements. Success and failure adds up to natural selection for individuals and species. In comparison natural selection for proteins depends on their contribution to this complex system. The protein evolution and the biological species evolution are both dependant on each other: an example of mutualism between a set of molecules and a set of biological species.

To estimate the independence of protein and organism history it would be worthwhile to estimate their relative rates of replication but that is difficult to do with any accuracy. For the eukaryote species we can make a very rough model, by assuming a steady state even though it has been perturbed by large scale events of extinction. At present there are less than 2 million described biological species and the estimates of the total present number including un-described species range from 10 million to 100 million. Assume that at any one time in the past there were 10 million species with a mean lifetime before extinction of 4 million years (13). Thus on average a couple of new species appears every year and a few go extinct, by this crude calculation. The order of magnitude is probably correct. The calculation suggests that there have been 5 billion species of eukaryotes that appeared and became extinct since the LCAEP. The uncertainty is so great that I usually consider that there have been a billion species, while there might have been 10 billion.

For the proteins we have no idea of the extinction rate. An estimate of the number of proteins present in the LCAEP can be based on the number of proteins that have survived in the prokaryotes, with a maximum presently known of 8702 in Bx (*Burkholderia xenovarans*). The number of types of proteins in the prokaryotes taken together is much larger, but protein evolution has occurred among them , of course, giving rise to new functions. We are left with a crude estimate of 10^4 proteins in the LCAEP as a starting point. The present day number of protein types is about 10^5, suggesting a tenfold growth. This has occurred over about 2 billion years suggesting a duplication every 200 million years on average, not allowing for losses. This is the required minimum rate of protein duplication.

Of course there is another way to count proteins, multiplying the number of biological species by the number of proteins in each and counting polymorphism in populations, leading to a very large number of more than 10^{5+9} but the interest here is in the evolution of types of proteins.

It is hard to say what limits can be placed on losses in the early years of metazoon life but in the last few hundred million years they have not been great. For example a fish D rerio shares about 90% of human proteins at expectation less than 10^{-3}. However proteins are many of them present in families and individual family members could be lost without losing the recognition of fish and human proteins. Little confidence can be placed in these crude estimates, but there seems no doubt that new protein formation is orders of magnitude smaller than that of biological species which has produced a billion species at least.

METHODS

Many comparisons have been made with BLASTp (5) at the most open possible criterion to detect distant relationships. The criterion for a significant match of an expectation of 10-3 or less was chosen because that is the most open criterion at which few if any accidental matches occur as shown by the following test. The proteins of the Archaea, (*Haloquadratum walsbyi*), were compared using BLASTp to a random protein library with the same set of lengths and average composition as the KGMV human protein library. The result was that the 13298 proteins of this library made 4 matches with the random sequences. This was taken as a negligible accidental background level. A limit of 10^{-3} allows a very small background number of accidental matches and is a conservative choice of expectation limit. This open criterion is suitable for recognizing many significant similarities between human proteins and proteins of prokaryotes. Build 35 with 25,193 proteins while build 36 has 34,180. proteins. An Apple G5, a Sun II and a Dell 8200 were used for these studies. To prepare the file of "known protein" genes that include the proteins that have been studied a list of the 25,193 genes with brief identifiers was alphabetized and blocks of were removed, for example those identified as hypothetical or similar to other genes. Then all the members of sets of transcription variants were removed and replaced with the gene that appeared to have the longest variant, with 13,298 remaining, called the KGMV library. For this purpose the length of the transcript was taken from the protein description. For each match the BLASTp program calculates an expectation. An expected frequency of occurrence can be converted to a probability of occurrence using the equation: P = 1 - exp(-E). In the limit as E approaches infinity, P approaches 1. In the limit as E approaches 0, P approaches E.

References

1 Yang S, Doolittle RF, Bourne PE. (2005)*Proc Natl Acad Sci U S A*.;102(2):373-8.
2 Boucher Y, Douady CJ, Papke RT, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF (2003) Annu Rev Genet 37: 283-328
3 Andersson J O (2005)Cell Mol Life Sci 62, 1182-1197
4 Britten RJ.(2006) *Proc Natl Acad Sci U S A*.;103(50):19027-32.
5 Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), *Nucleic Acids Res*. 25:3389-3402.
6 Knoll AH, Javaux EJ, Hewitt D, Cohen P. (2006) Philos *Trans R Soc Lond B Biol Sci*. 361(1470):1023-38.
7 Javaux EJ, Knoll AH, Walter M. (2003) *Orig Life Evol Biosph*. 33(1):75-94
8 Brocks JJ, Logan GA, Buick R, Summons RE.(1999) *Science*. 285(5430):1033-6.
9 Doolittle RF, Feng DF, Tsang S, Cho G, Little E. (1996) *Science*. 271(5248):470-7.
10 Gu X (1997) *Mol Biol Evol*.14(8):861-6
11 Salzberg SL, White O, Peterson J, Eisen J A (2001) Science 292 1903-1906
12 Wheelan SJ, Marchier-Bauer A, Bryant SH (2000) *Bioinformatics* 16:613-18
13 Raup DM (1994) *Proc Natl Acad Sci U S A*.91(15):6758-63.