doi:10.1016/j.jmb.2005.04.027

J. Mol. Biol. (2005) **349**, 745–763

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The Complement of Enzymatic Sets in Different Species

Shiri Freilich[1]*, Ruth V. Spriggs[1,2], Richard A. George[1,2]
Bissan Al-Lazikani[2], Mark Swindells[2] and Janet M. Thornton[1]

[1] *EMBL-EBI, Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD, UK*

[2] *Inpharmatica Ltd, 60 Charlotte Street, London W1T 2NU UK*

**Corresponding author*

We present here a comprehensive analysis of the complement of enzymes in a large variety of species. As enzymes are a relatively conserved group there are several classification systems available that are common to all species and link a protein sequence to an enzymatic function. Enzymes are therefore an ideal functional group to study the relationship between sequence expansion, functional divergence and phenotypic changes. By using information retrieved from the well annotated SWISS-PROT database together with sequence information from a variety of fully sequenced genomes and information from the EC functional scheme we have aimed here to estimate the fraction of enzymes in genomes, to determine the extent of their functional redundancy in different domains of life and to identify functional innovations and lineage specific expansions in the metazoa lineage. We found that prokaryote and eukaryote species differ both in the fraction of enzymes in their genomes and in the pattern of expansion of their enzymatic sets. We observe an increase in functional redundancy accompanying an increase in species complexity. A quantitative assessment was performed in order to determine the degree of functional redundancy in different species. Finally, we report a massive expansion in the number of mammalian enzymes involved in signalling and degradation.

Keywords: bioinformatics; enzyme evolution; functional redundancy; genome design; mammalian enzymes

Introduction

The availability of a growing number of sequenced genomes facilitates comparative studies between species to reveal conservation and diversification patterns between phylogenetic lineages relative to their last common ancestor. Such studies identify the differences between the three domains of life as well as the differences between closely related species. Recent studies used the gene content information to achieve a more general perspective of the principles of genome design and complexity that are beyond a direct phylogenetic relationship.[1–4] Such studies had focused on elucidating the expansion pattern of various functional groups of proteins within archaea, bacteria and eukaryota. Metabolic genes, for example, were shown to have a roughly constant fraction of the gene content of a genome in each domain of life, i.e. a constant and not-necessarily common fraction for bacteria, archaea and eukaryota.[3] Proteins involved in "small molecule metabolism" were found to be clearly over-represented in small bacterial genomes.[2]

Gaining an overview of how the total protein content, devoted to a specific biological process, varies in different organisms is useful before looking in depth at particular processes/pathways. Linking information on the size of a protein group in a species with information on its functional diversity provides an insight into the ways in which genome expansion affects the functional repertoire in different species and in different domains of life. A comparative study of the functional repertoire can further relate a functional innovation to a process that is either unique to a species or to a group of species. Ultimately, we wish to understand how the phenotype evolved in response to genome evolution.

Here we consider in detail the enzyme complement. Several factors make the set of enzymes a natural candidate for such a study: enzymes are a

Abbreviations used: EC, Enzyme Commission; KEGG, Kyoto Encyclopedia of Genes and Genomes.

E-mail address of the corresponding author: shirigo@ebi.ac.uk

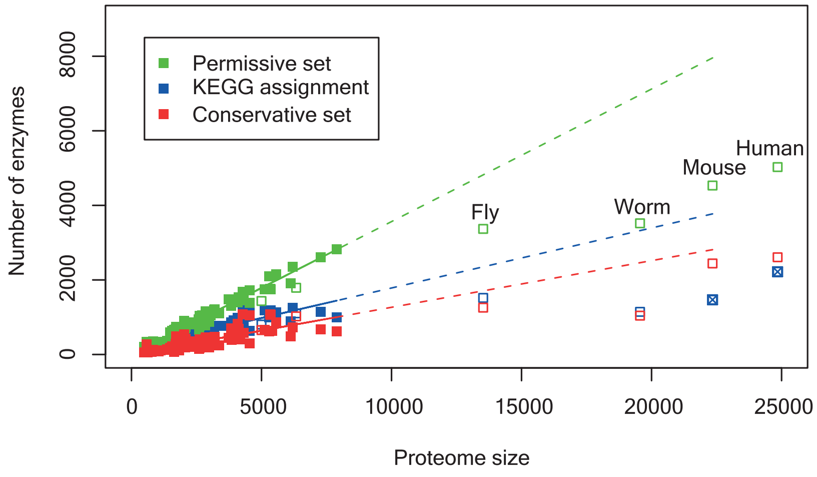

Figure 1. Number of enzymes per species *versus* proteome size. Filled squares, prokaryote species; open squares; eukaryote species; crossed squares, incomplete proteome for human and mouse in KEGG compared to these proteomes in the Biopendium™ (proteome size here is the one retrieved from the Biopendium™). The straight lines represent the regression line calculated for each enzyme set of the prokaryote species. The dotted lines are the extensions of the lines calculated for prokaryote species.

relatively conserved group[5,6] that has been studied extensively, therefore, there are several unique classification systems available. The Enzyme Commission (EC) scheme is a universal reaction classification system common to all species.[7] Such a system is a fundamental requirement for studying function divergence. Metabolic pathway databases, such as KEGG,[8] WIT[9] or EcoCyc,[10] relate an enzymatic reaction to a higher level cellular process, enabling us to make the link between genotypic and phenotypic changes.

Here, we first examine whether the fraction of the enzymes in the genome is constant within and across the three domains of life. A constant fraction of metabolic genes within each domain was previously reported.[3] The analysis here concerns all enzymes in a species, i.e. all proteins classified under the EC scheme (including proteins involved in micro and macro molecule metabolism, signalling and degradation). Our second goal is to characterise the pattern of expansion of enzymatic sets in different species and different domains. In particular, we examine the extent to which two different modes contribute to expansion: the broadening of the reaction repertoire of an organism (more enzymatic reactions) or an increase in its functional redundancy (more proteins performing the same function). Finally, we have characterised differences between the composition of enzymatic sets in different species, especially functional innovations and lineage-specific expansions in the metazoa lineage.

Results

Constructing the full complement of enzymes in different species

To identify the fraction of proteins that are enzymes we started from the complete list of highly curated enzymes in SWISS-PROT[11] and performed a PSI-BLAST[12] search against 85 fully sequenced genomes: 63 bacteria, 16 archaea and six eukaryota species. That is, we infer enzyme function if the sequence is sufficiently similar to one of the validated enzymes in the query list. To explore the consequences of this assumption, for every species in the analysis we define three sets of proteins:

(i) The conservative set: all proteins in a species matching an enzyme from the query list with more than 40% identity. Previous studies have shown that enzymes exhibiting more than 40% sequence identity share in most cases the same function.[13] Therefore, using a 40% identity cut-off enables us to transfer the functional annotation from the query protein to its hits with reasonable confidence.

(ii) The permissive set: all proteins recognising an enzyme from the query list after three PSI-BLAST iterations with an E-value cut-off of 10^{-3}. The use of PSI-BLAST identifies more distantly related homologues, often with low sequence identity ($<20\%$). Such distant relatives have

Table 1. Regression and correlation coefficients of different enzyme sets (Figure 1)

	Prokaryote species		Eukaryote species	
	Correlation coefficient (R^2)	Regression coefficient[a] ± std. error	Correlation coefficient (R^2)	Regression coefficient[a] ± std. error
Permissive set	0.98	0.36 ± 0.007	0.98	0.17 ± 0.017
KEGG assignments	0.90	0.16 ± 0.009	0.77	0.05 ± 0.019
Conservative set	0.79	0.13 ± 0.011	0.84	0.08 ± 0.026

[a] For a linear regression line ($y = ax + b$) the regression coefficient is the constant a; it is the slope of the regression line.

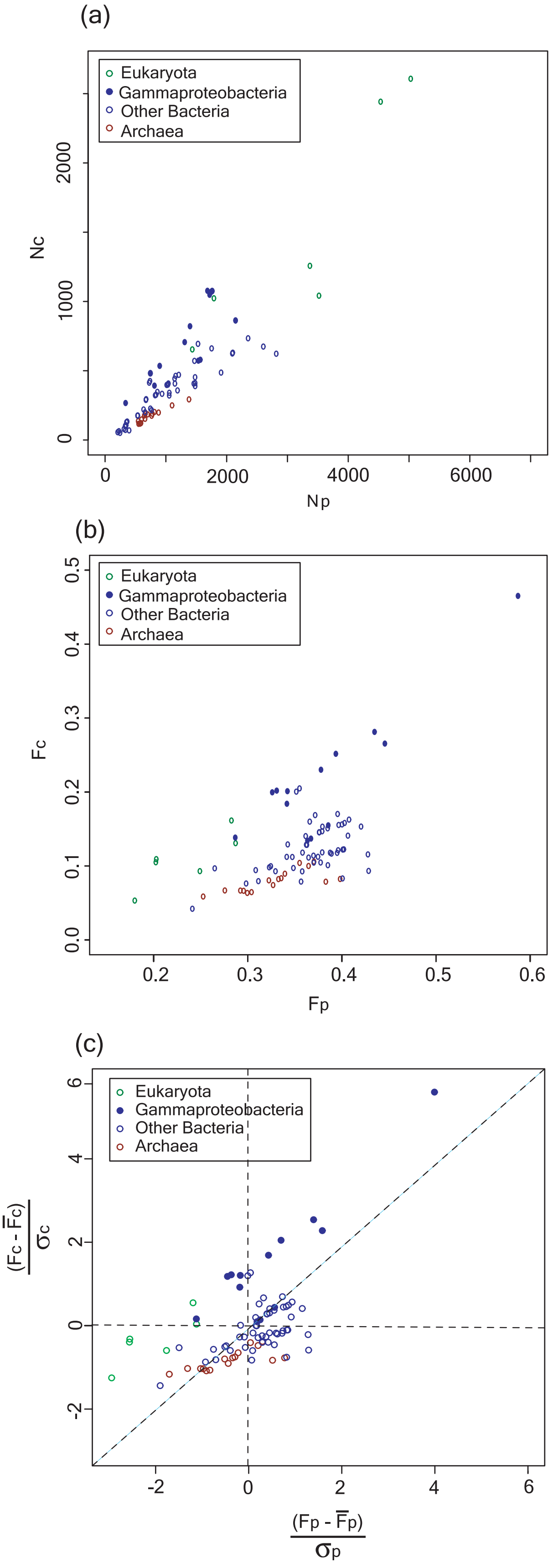

Figure 2. (a) The number of enzymes in a species calculated using different cut-offs. (b) The fraction of enzymes in a species calculated using different cut-offs. (c) The standard deviation from the mean value for the genomic fraction of enzymes in all 85 species examined (*z*-score). *Np*, *Nc*, the number of enzymes in the permissive and the conservative set, respectively; *Fp*, *Fc*, the fraction of the permissive and the conservative set, respectively, in each genome; $\bar{F}p, \bar{F}c$, the mean fraction of the permissive and the conservative set, respectively, in all 85 species; σ*p*, σ*c*, standard deviation of the permissive and the conservative set, respectively.

often evolved new functions;[13] these sets will include enzymes and non-enzymes.

(iii) The Kyoto Encyclopaedia of Genes and Genomes (KEGG) database predicted species' enzymatic set. The KEGG database aims to contain the complete information available for functionally annotated enzymes in fully sequenced genomes.[8,14] The annotations are retrieved from well-established sources including SWISS-PROT, GenBank and the original genome projects. KEGG also introduces its own predictions based on orthologous relationships.

In Figure 1, the sizes of these different sets were plotted against proteome size. For prokaryote species all three estimates indicate that the expansion of enzymes correlates with proteome expansion (Table 1). A linear model accurately describes the expansion of the permissive sets as indicated by a correlation coefficient of $R^2 = 0.98$. The correlation coefficients for the KEGG sets and the conservative sets are lower, but still significant at 0.90 and 0.79, respectively. The actual protein composition of the conservative and the KEGG sets in *Escherichia coli* almost completely overlaps. In addition, the reaction composition is more than 80% identical with 74 out of 80 prokaryote species and more than 75% identical with the remaining six species. We find that the KEGG prediction for the number of enzymes is usually an intermediate between the permissive prediction and the conservative prediction, suggesting that the cut-off used for the conservative set is stricter than the one used for the KEGG assignments.

For the four multicellular species, worm, fly, mouse and human, the trend is less obvious and the low number of analysed species prohibits definitive conclusions. If the regression line observed from prokaryote species is extended to eukaryote species (Figure 1), the number of enzymes in the multicellular metazoa is lower than predicted. This suggests a slower rate of expansion for the enzymatic sets relative to the increase in proteome size in multicellular species. The decrease is especially obvious for the permissive set.

The expansion of the eukaryotic permissive and conservative sets appears to correlate with genome size, with the exception of the worm genome that has almost the same number of enzymes as the smaller fly genome. The reasons for this are unclear. A decrease in the number of enzymes in worm

Table 2. Size and fraction of enzyme sets in species

Species		Proteome size	Permissive set			Conservative set			Representation in experimental studies[a]
			Number of enzymatic proteins	Fp[b]	$(Fp-\bar{F}p)/\sigma p$[c]	Number of enzymatic proteins	Fc[b]	$(Fp-\bar{F}p)/\sigma c$[c]	
E	*H. sapiens*	24,847	5027	0.20	−2.56	2608	0.10	−0.39	1.14
	M. musculus	22,345	4532	0.20	−2.55	2442	0.11	−0.32	1.11
	D. melanogaster	13,525	3369	0.25	−1.77	1257	0.09	−0.59	−0.19
	C. elegans	19,556	3519	0.18	−2.94	1042	0.05	−1.25	−0.52
	S. cerevisiae	6333	1790	0.28	−1.20	1023	0.16	0.55	2.49
	S. pombe	5000	1435	0.29	−1.12	654	0.13	0.04	0.50
B	*E. coli*	4279	1684	0.39	0.69	1077	0.25	2.05	4.55
	E. coli_O157	5324	1761	0.33	−0.38	1075	0.20	1.22	3.44
	E. coli_O157J	5361	1748	0.33	−0.46	1070	0.20	1.18	3.40
	S. typhimurium	4553	1720	0.38	0.43	1048	0.23	1.69	3.68
	Y. pestis	4083	1397	0.34	−0.18	821	0.20	1.21	1.36
	Buchnera	574	337	0.59	3.99	267	0.47	5.60	0.54
	H. influenzae	1714	745	0.43	1.39	482	0.28	2.54	0.57
	P. multocida	2015	898	0.45	1.58	535	0.27	2.28	0.17
	X. fastidiosa	2832	812	0.29	−1.13	392	0.14	0.16	−0.61
	X. campestris	4181	1535	0.37	0.24	573	0.14	0.14	−0.49
	X. axonopodis	4312	1567	0.36	0.18	579	0.13	0.09	−0.51
	V. cholerae	3835	1310	0.34	−0.19	706	0.18	0.92	−0.17
	P. aeruginosa	5567	2146	0.39	0.56	863	0.16	0.44	0.12
	N. meningitidis	2079	738	0.35	0.04	426	0.20	1.27	−0.12
	N. meningitides_A	2065	726	0.35	−0.02	414	0.20	1.20	−0.15
	R. solanacearum	5116	1752	0.34	−0.18	661	0.13	0.01	−0.56
	H. pylori	1576	538	0.34	−0.19	177	0.11	−0.27	−0.31
	H. pylori_J99	1491	534	0.36	0.09	176	0.12	−0.18	−0.29
	C. jejuni	1634	657	0.40	0.84	200	0.12	−0.10	−0.32
	R. prowazekii	835	351	0.42	1.15	128	0.15	0.41	−0.22
	R. conorii	1374	364	0.26	−1.50	133	0.10	−0.53	−0.44
	M. loti	7275	2603	0.36	0.08	674	0.09	−0.60	−0.53
	S .meliloti	6205	2352	0.38	0.45	734	0.12	−0.17	−0.18
	A. tumefaciens	5402	2101	0.39	0.62	630	0.12	−0.20	−0.37
	A. tumefaciens_C	5299	2093	0.39	0.72	624	0.12	−0.18	−0.38
	B. melitensis	3198	1212	0.38	0.45	470	0.15	0.31	−0.35
	C. crescentus	3737	1480	0.40	0.74	453	0.12	−0.12	−0.58
	B. subtilis	4112	1528	0.37	0.32	694	0.17	0.67	1.76
	B. halodurans	4066	1470	0.36	0.15	571	0.14	0.20	−0.40
	S. aureus_N315	2625	1041	0.40	0.75	408	0.16	0.45	0.04
	S. aureus_Mu50	2748	1033	0.38	0.39	400	0.15	0.28	−0.05
	S. aureus_MW2	2632	1013	0.38	0.55	397	0.15	0.37	−0.07
	L. monocytogenes	2846	1160	0.41	0.93	463	0.16	0.57	−0.43
	L. innocua	3043	1144	0.38	0.39	443	0.15	0.28	−0.48
	L. lactis	2267	860	0.38	0.45	347	0.15	0.41	0.23
	S. pyogenes	1697	671	0.40	0.73	289	0.17	0.69	−0.07
	S. pyogenes_M18	1845	675	0.37	0.22	295	0.16	0.52	−0.17
	S. pneumoniae	2094	838	0.40	0.81	327	0.16	0.46	0.02
	S. pneumoniae_R6	2043	823	0.40	0.85	323	0.16	0.49	0.02
	C. acetobutylicum	3848	1482	0.39	0.55	389	0.10	−0.46	−0.46
	C. perfringens	2723	1055	0.39	0.59	321	0.12	−0.18	−0.51
	T. tengcongensis	2588	937	0.36	0.16	333	0.13	0.00	−0.66
	M. genitalium	484	207	0.43	1.28	56	0.12	−0.22	−0.36
	M. pneumoniae	689	227	0.33	−0.40	64	0.09	−0.59	−0.49
	M. pulmonis	782	335	0.43	1.29	73	0.09	−0.59	−0.60
	U. urealyticum	614	246	0.40	0.81	51	0.08	−0.76	−0.54
	M. tuberculosis	3927	1482	0.38	0.42	411	0.10	−0.40	−0.19
	M. tuberculosis_CDC1551	4187	1459	0.35	−0.07	407	0.10	−0.52	−0.20
	C. glutamicum	3040	1056	0.35	−0.09	341	0.11	−0.27	−0.08
	S. coelicolor	7897	2816	0.36	0.06	623	0.08	−0.83	-0.43
	F. nucleatum	2067	765	0.37	0.29	216	0.10	−0.40	−0.63
	C. trachomatis	895	331	0.37	0.29	102	0.11	−0.24	−0.25
	C. muridarum	916	334	0.36	0.20	102	0.11	−0.29	−0.29
	C. pneumoniae	1054	342	0.32	−0.48	105	0.10	−0.48	−0.53
	C. pneumoniae_AR39	1112	343	0.31	−0.76	105	0.09	−0.57	−0.54
	C. pneumoniae_J138	1069	345	0.32	−0.51	105	0.10	−0.51	−0.54
	B. burgdorferi	1638	395	0.24	−1.90	69	0.04	−1.44	−0.41
	T. pallidum	1036	309	0.30	−0.93	79	0.08	−0.87	−0.53
	Synechocystis	3167	1149	0.36	0.17	407	0.13	−0.00	−0.12
	Anabaena	6129	1908	0.31	−0.71	487	0.08	−0.82	−0.39

Table 2 (*continued*)

Species		Proteome size	Permissive set			Conservative set			Representation in experimental studies[a]
			Number of enzymatic proteins	Fp^{b}	$(Fp-\bar{F}p)/\sigma p^{c}$	Number of enzymatic proteins	Fc^{b}	$(Fp-\bar{F}p)/\sigma c^{c}$	
	D. radiodurans	3182	1192	0.37	0.37	358	0.11	−0.27	−0.61
	A. aeolicus	1560	634	0.41	0.91	220	0.14	0.21	−0.58
	T. maritime	1858	745	0.40	0.82	227	0.12	−0.11	−0.05
A	*M. jannaschii*	1785	661	0.37	0.30	189	0.11	−0.38	−0.01
	M. acetivorans	4540	1380	0.30	−0.83	293	0.06	−1.07	−0.43
	M. mazei	3371	1102	0.33	−0.44	250	0.07	−0.91	−0.39
	M. thermoautotrophicum	1873	665	0.36	0.04	195	0.10	−0.41	0.02
	M. kandleri	1687	544	0.32	−0.52	136	0.08	−0.80	−0.52
	A. fulgidus	2420	812	0.34	−0.29	202	0.08	−0.75	−0.60
	Halobacterium	2622	766	0.29	−1.03	175	0.07	−1.03	−0.52
	T. acidophilum	1482	590	0.40	0.77	122	0.08	−0.77	−0.42
	T. volcanium	1499	574	0.38	0.51	118	0.08	−0.83	−0.56
	P. horikoshii	1801	599	0.33	−0.34	148	0.08	−0.77	−0.26
	P. abyssi	1769	645	0.36	0.20	177	0.10	−0.48	−0.21
	P. furiosus	2065	701	0.34	−0.23	185	0.09	−0.65	−0.15
	A. pernix	1840	551	0.30	−0.91	117	0.06	−1.08	−0.60
	S. solfataricus	2977	880	0.30	−0.97	198	0.07	−1.03	−0.28
	S. tokodaii	2826	779	0.28	−1.31	189	0.07	−1.03	−0.55
	P. aerophilum	2605	658	0.25	−1.71	153	0.06	−1.16	−0.66

E, eukaryota; B, bacteria; A, archaea.

[a] Standard deviation distance from mean value for level of representation (see Materials and Methods).

[b] Number of enzymes/proteome size (fraction of enzymes).

[c] Standard deviation distance from mean (mean fraction of enzymes in all 85 species). $\bar{F}p=0.35$, $\bar{F}c=0.13$, $\sigma p=0.058$, $\sigma c=0.060$.

compared to fly can also be observed in the KEGG sets. The low number of mouse and human enzymes in KEGG might be due in part to the partial representation of the mouse and human proteomes in the KEGG database. The total number of mouse and human proteins in KEGG (downloaded in February 2004) is approximately 2/3 of the current ENSEMBL estimate (August 2004), so the apparent low values are misleading.

Comparison of the permissive and conservative enzyme sets

The expansion patterns of the conservative set and the KEGG set as measured by the slope of the regression line are similar (Table 1) and probably reflect the similarity between the two annotation procedures. In contrast, the expansion of the permissive set is distinguishable from the former two and can be better described by a linear model (higher correlation coefficient). The main difference between the sets is that whilst the first two sets only include close relatives to proteins annotated in SWISS-PROT, the permissive enzyme sets also include more distant relatives (whose function may have diverged). We wanted to examine how the composition of the original query set influences the final hit list in different species under different cut-offs. For each species we plotted *Nc* (number of proteins in the conservative set) against *Np* (number of proteins in the permissive set, Figure 2(a)) and *Fc* (fraction of the conservative set) against *Fp* (fraction of the permissive set, Figure 2(b)). As in all species *Fp* is bigger than *Fc* ($\bar{F}p=0.35$, $\bar{F}c=0.13$) we normalised the relative fraction of each set by comparing it to the mean fraction in all species. For each species we calculated the *z*-score (the standard deviation, σ, distance from the mean): $(Fp-\bar{F}p)/\sigma p$ ($\sigma p=0.058$) and $(Fc-\bar{F}c)/\sigma c$ ($\sigma c=0.060$) (Figure 2(c), Table 2). The diagonal line represents an equal relative fraction of enzymes in the conservative and the permissive sets. In *Chlamydophila pneumoniae*, for example, the conservative set covers about 10% and the permissive set covers about 30% of the genome, which corresponds to similar *z*-score ~0.5 for the two sets (Table 2).

Species plotted above the diagonal line are those in which the relative fraction of enzymes was higher in the conservative set than in the permissive set. Most species that lie well above the diagonal are eukaryota and gammaproteobacteria (Figure 2(c)). Eukaryota and gammaproteobacteria are the most widely studied species as indicated by the composition of our query list (Table 3, see Materials and Methods for further explanation about determination of groups' relative representation in the query list). They are over-represented in the query list. Archaea species on the other hand almost always lie under the diagonal line, possibly due to their low representation in the experimental data.

Table 3. Level of representation of different species groups in SWISS–PROT

	Species group (no. of species)		Total number of genes in the group[a]	Total number of query-list highly related enzymes[b] in the group	Fraction	Standard deviation distance from mean[c]
Within domains	Eukaryota (6)		91,606	5376	0.06	0.95
	Bacteria (63)		184,396	7214	0.04	0.10
	Archaea (16)		37,162	471	0.01	−1.05
Within bacterial subfamilies	Proteobacteria	Gammaproteobacteria (12)	48,630	4745	0.10	2.77
		Betaproteobacteria (3)	9260	128	0.01	−0.36
		Epsilonproteobacteri (3)	4701	78	0.02	−0.26
		Alphaproteobacteria (7)	33,325	429	0.01	−0.40
	Firmicutes	Bacillales (7)	22,072	786	0.04	0.45
		Lactobacillales (5)	9946	309	0.03	0.28
		Clostridia (3)	9159	62	0.01	−0.62
		Mollicutes (4)	2569	20	0.01	−0.59
	Actinobacteridae (4)		19,051	363	0.02	−0.16
	Fusobacterales (1)		2067	5	0.00	−0.79
	Chlamydiales (5)		5046	54	0.01	−0.48
	Spirochaetales (2)		2674	27	0.01	−0.50
	Cyanobacteria (1)		9296	699	0.08	1.93
	Deinococci (1)		3182	11	0.00	−0.75
	Aquificales (1)		1560	7	0.00	−0.71
	Thermotogales (1)		1858	52	0.03	0.17

[a] Relates only to species from the 85 species participating in the analysis.
[b] Query-list highly related enzymes: enzymes with more than 80% identity to a protein in the query list.
[c] Mean: mean value of the enzymatic fraction in the three domains or in the bacterial subfamilies. Within domains: mean = 0.037, standard deviation = 0.023. Within bacterial subfamilies: mean = 0.023, standard deviation = 0.027.

We suggest that the permissive set is less sensitive to the biases in our experimental knowledge, which may explain the improved linear correlation for this set (Figure 1). We use the permissive set to calculate enzymatic fraction, but confine functional analysis to the conservative set.

The enzyme fraction of the genome in different species and domains

Since the permissive enzymatic sets are less biased towards the common model organisms, we have used these sets to calculate the enzymatic

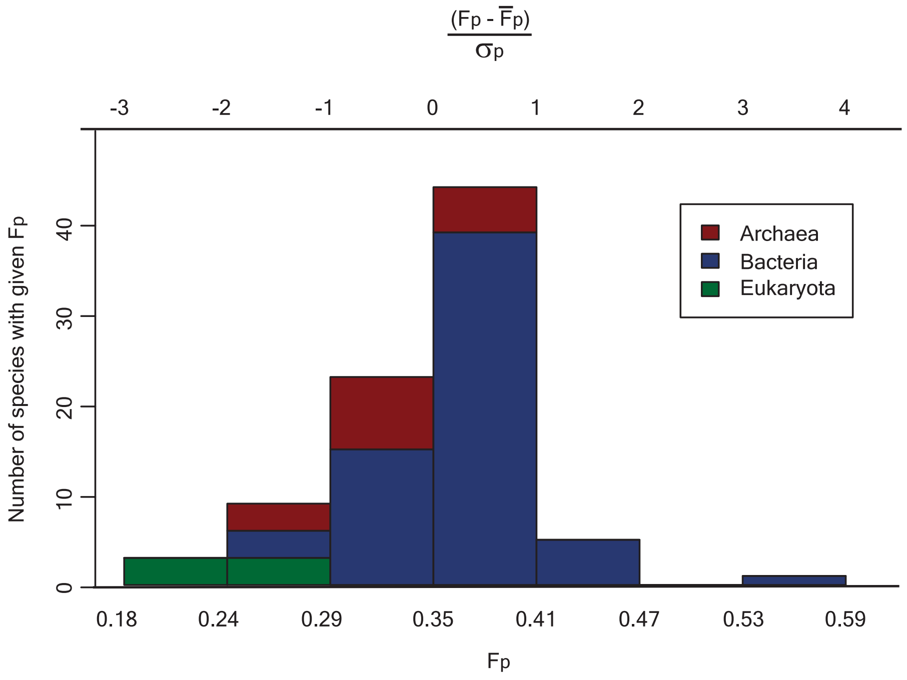

Figure 3. Distribution of the standard deviation from the mean of the fraction of enzymes in 85 species. The values calculated are derived from the permissive enzymatic sets. Eukaryote species: $\bar{F} = 0.23$, $\sigma = 0.05$; Bacteria species: $\bar{F} = 0.37$, $\sigma = 0.05$; Archaea species: $\bar{F} = 0.33$, $\sigma = 0.04$.

fraction (Fe) in different species. The distribution of Fp for species from the three domains of life is shown in Figure 3. $\bar{F}p$ is the average Fp of all 85 species in the analysis. In most archaeal and bacterial species (67 out of 79 species) enzymes and enzyme-related proteins occupy approximately 30–40% of the genome and lie within one standard deviation of the mean value.

All six species in which the fraction of enzymes is significantly higher than the mean (more than one standard deviation) are bacteria. These bacteria are phylogenetically diverse and include one alphaproteobacterium (*Rickettsia prowazekii*), three gammaproteobacteria (*Haemophilus influenzae*, *Pasteurella multocida*, *Buchnera*) and two mollicutes (*Mycoplasma genitalium*, *Mycoplasma pulmonis*). Five are pathogenic bacteria and one is a symbiont (*Buchnera*). Four (*R. prowazekii*, *Buchnera*, *M. genitalium*, *M. pulmonis*) are intracellular, obligatory pathogens or symbionts with a small genome. Intracellular pathogens and endosymbiont bacteria were previously shown to have a relatively high fraction of enzymes.[2] The high metabolic fraction was suggested to be the result of a trend, which occurred in many species independently, in favour of a massive loss of regulatory proteins in intracellular species functioning in a relatively stable environment.[2] Phylogenetic diversity can also be observed in the six prokaryote species with a small fraction of enzymes (more then one standard deviation below the mean). These include three bacteria (alphaproteobacteria, gammaproteobacteria and a spirochaetales) and three archaea (one euryarchaeota and two thermoprotei species). The phylogenetic diversity of these extreme groups supports the notion that the metabolism of an organism better reflects its specific adaptation rather than its phylogenetic history.[2,15]

All eukaryota species in this sample have a relatively low fraction of enzymes (Figure 3) occupying 18–29% of their genome (Table 2). The two yeast species and the fly are between one to two standard deviations from the mean. The lowest fraction of enzymes was recorded for the three metazoan species: worm, mouse and human, where enzymes constitute only 18–20% of the genome. This observation is compatible with previous studies showing that the fraction of proteins involved in metabolism decreases when species complexity grows.[16,17]

We observe here that the increase in the number of enzymes and enzyme-related proteins is correlated with proteome expansion in most prokaryote species (Figure 1). The fraction of enzymes and enzyme-related proteins is approximately constant and ranges between 30 and 40%. The trend is observed in both archaea and bacteria species, where extreme values seem to reflect species-specific adaptations rather than a phylogenetic trend. For example, a relatively high fraction of enzymes was found in species that had a massive loss of regulatory proteins. In contrast, a relatively low fraction of enzymes was found in eukaryote species, a lineage whose phylogenesis involved a massive recruitment of regulatory proteins.[16,18,19] We therefore suggest that the rate of enzyme expansion in a species is approximately constant, and that differences in the fraction of enzymes mainly reflect dramatic changes in the size of other functional categories.

Enzymes recruitment and functional diversification

There are two explanations for the expansion of the number of enzymes as proteome size increases: either a larger variety of reactions has evolved or there are more proteins catalysing the same reactions (isoenzymes), with other differences, such as the control of expression, driving their evolution and retention. Here, we wanted to estimate the level of functional diversification accompanying gene recruitment, and we used the EC scheme† in which each reaction has been assigned a unique four digit identifier. Enzymatic reactions are divided into six classes represented by the first digit in the EC number: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The second digit refers to the subclass, which generally contains information about the type of compound or group involved in the reaction. This subclass definition differs between classes but in most cases it describes the donor group. The third digit further specifies the type of reaction involved, in most cases describing the type of acceptor involved. The fourth digit is a serial number that is usually used to identify the substrate for individual enzymes within a sub-subclass†.[7]

One of the drawbacks in using the EC scheme for determining functional divergence is that some EC numbers describe generic reactions where the compounds are not fully specified. For example, all protein tyrosine-kinases are classified as EC 2.7.1.112 whilst all proteases lie in EC 3.4 but have multiple third and fourth digit depending on their catalytic mechanism and target.[7,20] However, the focus of the EC scheme on reaction type rather than mechanism[7] is in most cases advantageous for this kind of study.

Here, in order to estimate the level of functional diversity accompanying gene recruitment we plotted the number of enzymes in a species against the number of reactions (number of distinct EC numbers assigned in a species). The reaction ratio, R, equals the total number of reactions divided by the number of enzymes in the organism. If R equals 1 then the increase in the enzyme number is entirely due to having more reactions. A smaller ratio implies an increase in the number of enzymes performing the same reaction.

As function has been experimentally determined for only a small number of proteins,[21] calculating the number of reactions per species requires inferring the function from sequence. Several recent

† http://www.chem.qmul.ac.uk/iubmb/enzyme/rules.html

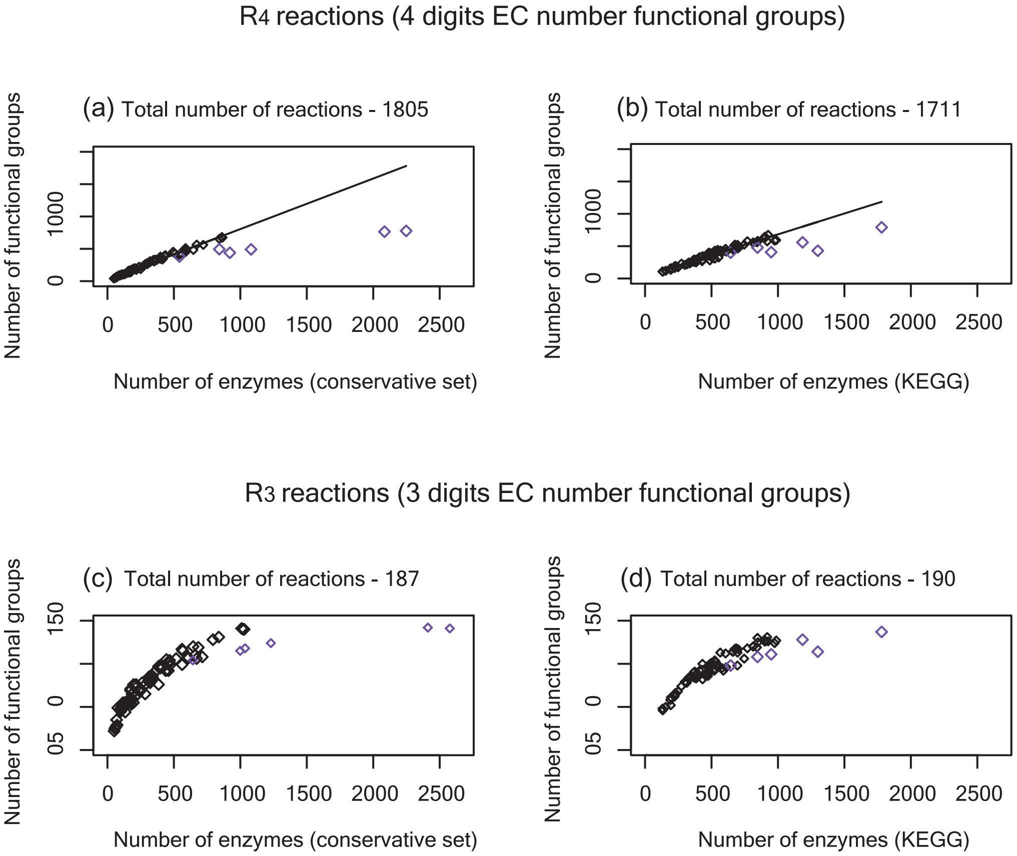

Figure 4. Number of reactions *versus* number of enzymes. Reactions are described either by four-digit assigned EC numbers (a and b) or the three digit assigned EC numbers (c and d). (a) and (c) Number of enzymes per species is the number of proteins with a full EC assignment (a) or a three-digit EC assignment (c). The EC assignments are inferred for proteins sharing 40% sequence identity with a highly curated enzyme from the query list. (b) and (d) Number of enzymes is the number of proteins assigned by KEGG with a full EC number. Black squares, prokaryote species; purple squares, eukaryote species. The straight lines represent the regression line calculated for each set of the prokaryote species. Total number of reactions refers to the number of different reactions in all 85 species examined.

analyses have used EC numbers to study how function changes as homologues diverge.[13,22,23] For single and multi domain proteins, variation in the EC number was found to be rare above 40% sequence identity.[13] Here, based on this observation, we have estimated the number of reactions in the conservative sets by transferring the EC number from a protein in the query list to all hits sharing more than 40% sequence identity. For each species we counted the number of distinct EC reactions. As using functional assignments inferred from sequence can at best only provide an estimation of the number of reactions, we also examined the KEGG assignments obtained by considering orthologous relationships.[8,14]

The reaction ratio for each organism was calculated, counting both the number of distinct EC reactions down to the fourth level (R_4) and distinct EC reactions down to the third level (R_3).

Eukaryotes are more functionally redundant than prokaryote species

The two functionally annotated data sources (KEGG and the conservative set) provide similar observations regarding the diversification of enzyme function (Figure 4(a) and (b)), showing that eukaryote and prokaryote species differ in their expansion pattern. For prokaryotes, plotting the number of reactions against the number of enzymes shows that the two sets are essentially identical, both with a correlation coefficient ~0.99 (Table 4). The regression line obtained (for prokaryotic organisms only, conservative set, Figure 4(a)) is given by the following equation:

$$\text{Number of reactions}$$

$$= 0.78\ \text{Number of Enzymes} + 24$$

Therefore the number of reactions is approximately 4/5 the number of enzymes, i.e. most enzymes perform a single reaction and most reactions are performed by a single enzyme. From the equation, equality between the number of enzymes and the number of reactions is achieved for approximately 110 enzymes. A hypothetical species with 110 enzymes is expected to be able to catalyse 110 reactions without having a single

Table 4. Regression and correlation coefficients of the distribution of the number of different reactions against the number of enzymes

	Prokaryote species		Eukaryote species	
	Correlation coefficient (R^2)	Regression coefficient ± std. error	Correlation coefficient (R^2)	Regression coefficient ± std. error
KEGG assignments (4 digit)[a]	0.98	0.64 ± 0.016	0.85	0.32 ± 0.100
Conservative set (4 digit)[b]	0.99	0.78 ± 0.011	0.99	0.24 ± 0.019
Oxidoreductases (1)[c]	0.98	0.72 ± 0.016	0.99	0.26 ± 0.013
Transferases (2)[c]	0.99	0.76 ± 0.014	0.95	0.17 ± 0.028
Hydrolases (3)[c]	0.97	0.76 ± 0.020	0.99	0.30 ± 0.020
Lyases (4)[c]	0.98	0.74 ± 0.015	0.58	0.24 ± 0.166
Isomerases (5)[c]	0.98	0.67 ± 0.014	0.96	0.12 ± 0.018
Ligases (6)[c]	0.99	0.85 ± 0.017	0.83	0.22 ± 0.075

[a] Figure 4(b).
[b] Figure 4(a).
[c] Figure 5.

isoenzyme. Below 110 enzymes the number of reactions is predicted to exceed the number of enzymes due to multi-functional enzymes. In *Chlamydia muridarum* for example, the number of enzymes in the conservative set is 88, the number of reactions predicted from the equation is 92, and the actual number of reactions is 89.

The linear dependency between the number of enzymes and reactions found for prokaryotic species does not apply for eukaryotic genomes, as the number of reactions in eukaryotic species is lower than that predicted by the prokaryotic regression line, indicating a higher functional redundancy. This observation is compatible with previous studies reporting a general trend, where higher organisms have larger sizes for many of their protein families.[24–27] While both data sets agree, the KEGG assignments suggest a slightly lower increase in reaction ratio in eukaryotic species, probably because of the requirement for an orthologous relationship for functional inference. Thus in prokaryotic species the expansion in the number of enzymes mainly reflects the broadening of the reaction repertoire, whilst in eukaryotes the expansion is to a greater extent the result of increased reaction redundancy.

When studying the distribution of R_3 reactions (reactions detailed down to the third level), a plateau is observed after reaching an enzymatic set size of around 500 proteins (Figure 4(c) and (d)). The number of functional groups in prokaryotic species is similar to the number in similar size eukaryotic species (size relates to estimated number of enzymes). The plateau is the result of the low functional diversity between species with regard to the three-digit EC number reactions. Currently (August 2004), the EC scheme contains 4327 known R_4 reactions, which map to only 236 R_3 and 63 R_2 reactions. The conservative data set includes 1805 R_4 reactions that are mapped to 187 R_3 reactions. As none of the species examined has more than 800 R_4 reactions, there is still diversity in the reaction composition between species, while such diversity is unlikely for the R_3 reactions, where species only have up to 142 reactions. Human (777 R_4 reactions, 141 R_3 reactions) and *E. coli* (682 R_4 reactions, 140 R_3 reactions) for example, share only about 40% (on average) of their R_4 reactions *versus* 80% of their R_3 reactions. Even species with a low number of enzymes share a very similar set of reactions (not shown). Beyond approximately 500 enzymes, species contain almost all R_3 reactions and functional diversification is achieved by diversification in the R_4 EC number reaction composition. It is important to remember however that the inference of function herein is conservative; there is no doubt that in reality most species are likely to have more reactions than predicted here.

Functional diversity of different reaction classes

We have examined whether the increased functional redundancy observed in eukaryote species merely reflects the massive recruitment of enzymes in a limited number of broadly defined reaction types (e.g. protein tyrosine-kinases) or whether it reflects a more general trend. We first divided the enzymes into the six reaction classes of the EC scheme and plotted the number of enzymes and reactions (R_4) in each class against the size of the proteome (Figure 5). Since the results are essentially the same for the KEGG and conservative data sets, only the conservative set is presented and the data in Figure 5 are derived from the data in Figure 4(a). The plot shows that the enzymes are unequally distributed between the six classes and their rates of expansion are radically different. A massive expansion of the transferases and hydrolases functional classes can be observed in metazoa (human, mouse, worm and fly).

Next, we plotted the number of reactions in each class against the number of enzymes in that class (i.e. the number of oxidoreductase reactions in a species against the number of oxidoreductase enzymes in that species) (Figure 6). In prokaryotes, oxidoreductases, transferases and hydrolases seem to be more abundant than lyases, isomerases and ligases. For all functional classes a clear linear

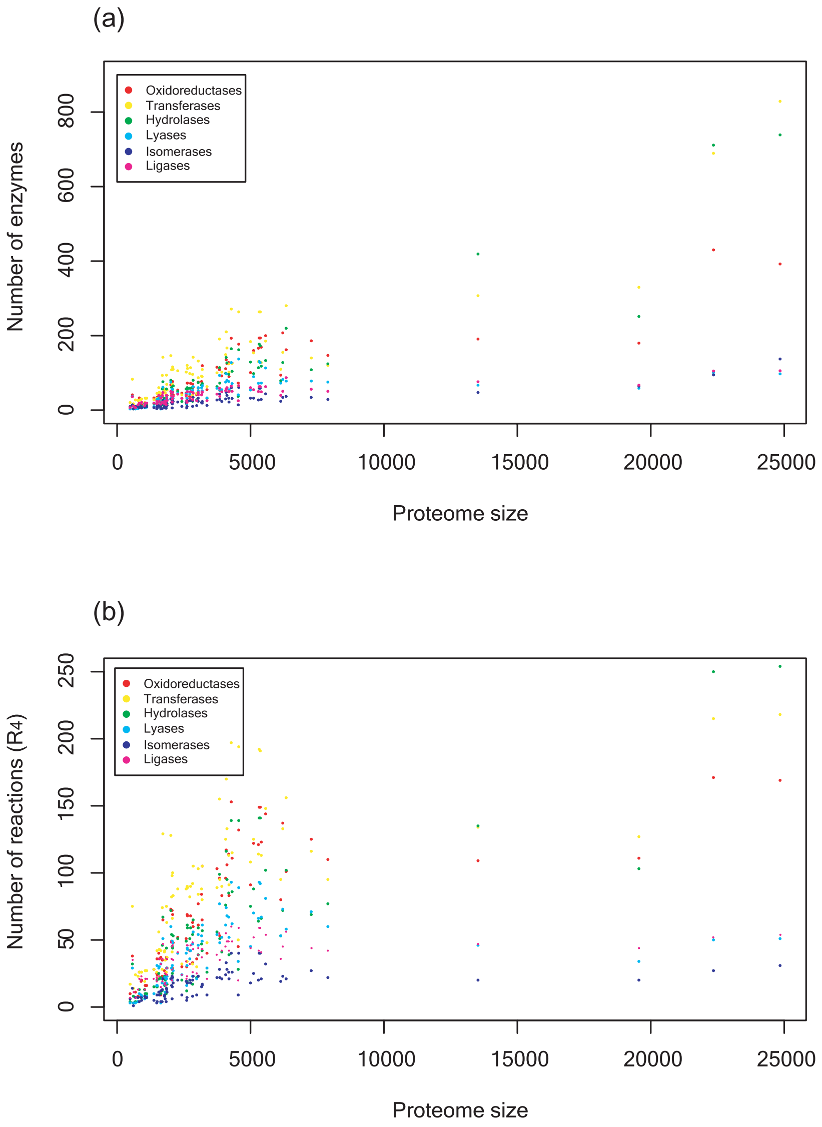

Figure 5. Number of enzymes (a) and number of R_4 reactions (b) against proteome size. The number of reactions and number of enzymes are derived from the conservative set.

dependency exists between the number of enzymes and the number of reactions with a correlation coefficient ranging from 0.97 to 0.98 (Table 4). The slope of the linear regression line is very similar in all plots and ranges between 0.72 and 0.76 in oxidoreductases, transferases, hydrolases and lyases. The slowest rate of functional diversification (regression coefficient of 0.67) was observed for isomerases and the highest rate (regression coefficient 0.85) was observed for ligases.

In eukaryotic species the correlation coefficient is lower in most reaction classes, compared to prokaryote species (Table 4). Unlike prokaryote species, the distribution pattern is not constant and differs between the reaction classes. The rate of expansion of reactions varies and ranges from 0.12 (isomerases) to 0.30 (hydrolases). In one of the reaction classes (lyases) there is no clear linear dependency between the number of reactions and the number of enzymes (correlation coefficient = 0.58). In three reaction classes (oxidoreductases, transferases and hydrolases) there is a massive recruitment of enzymes in mammals although there is no significant increase in the number of reactions. A more moderate increase in the number of mammalian enzymes is observed for the isomerase and ligase reaction classes where the number of reactions is again smaller than the one predicted by the regression line describing the distribution in prokaryote species. Lyases are the only reaction class with a smaller number of enzymes in mammals than in prokaryote species. Other eukaryote species also exhibit an increase in functional redundancy compared to prokaryote species, although more moderate than in mammals. The yeast *Schizosaccharomyces pombe* is the least functionally redundant eukaryote species.

The extent of reaction redundancy: how many reactions in a species are redundant?

In order to quantify how general the functional redundancy is (e.g. how many reactions per species have multiple enzymes), we studied the

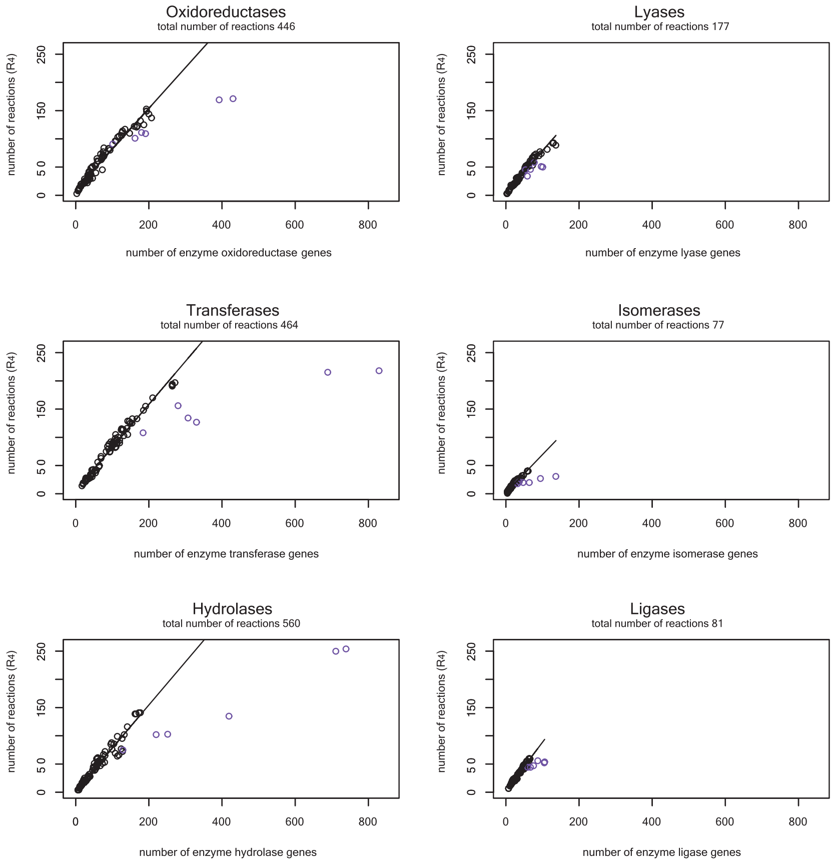

Figure 6. Number of R_4 reactions *versus* the number of enzymes assigned to the reaction class. The number of reactions and number of enzymes are derived from the conservative set. The straight lines represent the regression line calculated for each set of the prokaryote species. Total number of reactions refers to the number of different reactions in all 85 species examined.

distribution of the number of enzymes per reaction in five species representing different levels of complexity (Figure 7). The species examined, ordered from the most to least complex, are the multicellular metazoa human and fly, *Saccharomyces cerevisiae* (unicellular eukaryote), and two prokaryote species: the free living *Escherichia coli* and the intracellular obligate symbiont *Buchnera*. In all functional classes, the fraction of reactions with only a single enzyme assigned increases when species' complexity decreases. Human has the biggest fraction of reactions that are redundant (58%), i.e. the same reaction is performed by more than a single enzyme, followed by fly (44%), yeast (37%), *E. coli* (25%) and *Buchnera* (10%).

Buchnera, the small genome intracellular obligate symbiont, exhibits a lower functional redundancy compared to the free living *E. coli*. Both *E. coli* and *Buchnera* belong to the Enterobacteriaceae subdivision of the gammaproteobacteria. The evolution of the *Buchnera* genome, as of many other obligate parasites and symbionts, involved a massive reduction in the size of the genome.[28] The number of genes in *Buchnera* (~ 500) is approximately 1/10 of the number of genes in the closely related *E. coli* genome. Almost every *Buchnera* gene has a clear orthologue in *E. coli*, indicating that *Buchnera* provides a good approximation of a minimal *E. coli* genome (in the context of an intracellular environment[29]). As almost all of its reactions seem to be assigned to a single gene product it seems that *Buchnera* has lost even the limited functional redundancy observed in *E. coli*. The lack of functional redundancy might be related to the fact that *Buchnera* is an intracellular species functioning in a relatively stable environment.

Increased functional redundancy is observed in species with increased complexity. In eukaryote species, and especially in metazoa where there are many different cell types and environments, many reactions are catalysed by more than a single enzyme. The increase in functional redundancy might also reflect the lack of physical constraint on the size of the eukaryotic genome where the selective pressure to lose redundant genes is less strong than in bacterial genomes.[30]

The most redundant reaction class in human is the hydrolase reaction class where 65% of the reactions are catalysed by more than a single enzyme, followed by the transferase and oxidoreductase functional classes where 58% of the reactions are redundant (Figure 7). These correspond to the classes that are expanded massively (Figure 6). In the isomerase and ligase reaction classes, where one observes a moderate increase in the number of enzymes, 52% of the reactions are redundant. In the lyase reaction class, where the proteins/enzymes ratio in eukaryotes is very similar to the ratio in eukaryotes (Figure 6), only 39% of the reactions are redundant.

We have listed those reactions that are assigned more than 20 enzymes per species (Table 5). Only

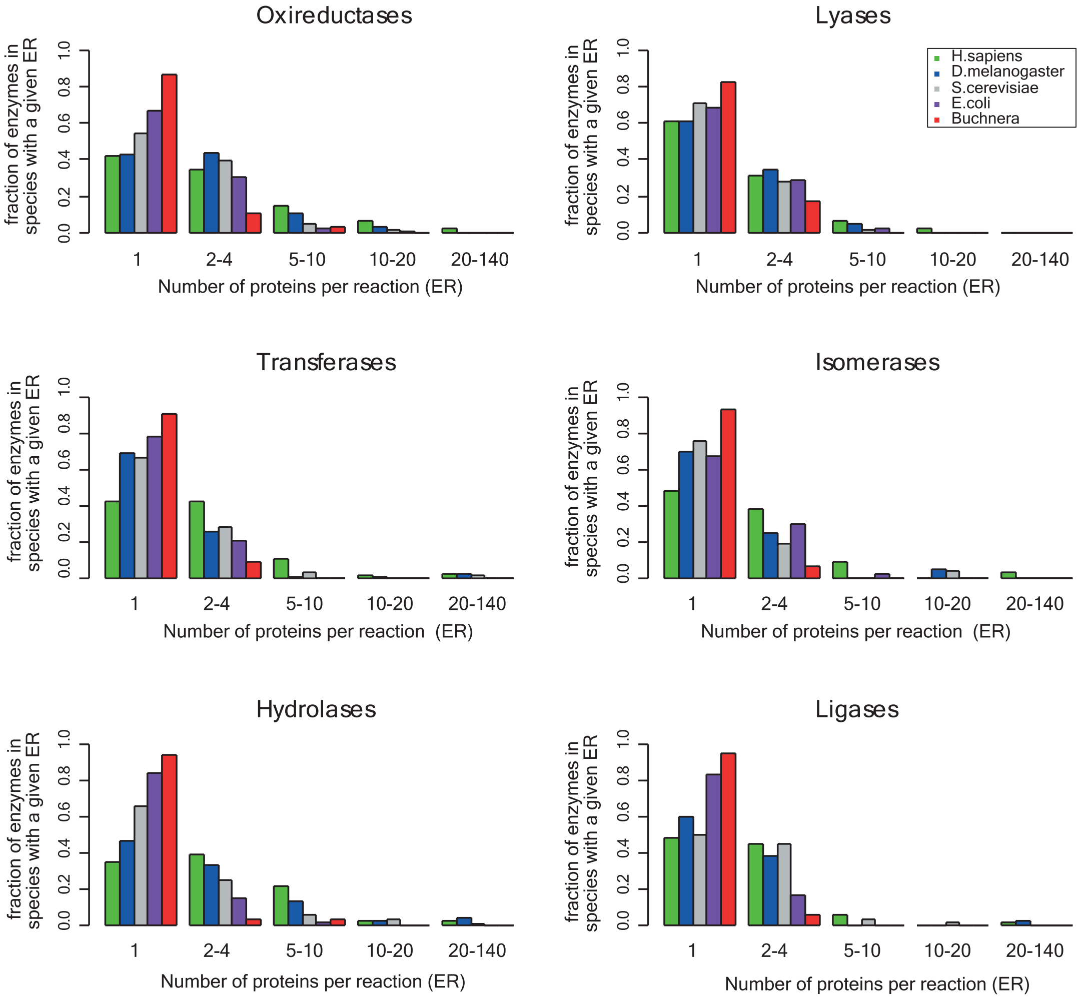

Figure 7. For each class, the fraction of all enzymes in the species with a given "number of enzymes per reaction" (ER). The total number of enzymes (conservative set) is as follows: *H. sapiens*, 2608; *D. melanogaster*, 1257; *S. cerevisiae*, 1023; *E. coli*, 1077; *Buchnera*, 267.

22 such reactions were identified for the five model species. For many of the reactions listed the high number of proteins assigned is due to a broad specificity of the EC reaction or to having several proteins working in concert as subunits of a protein complex. Therefore the proteins in many of these reactions are not true isoenzymes, and these big reaction clusters do not usually represent a true functional redundancy. Yet, a comparison of the species distribution of the reactions is of interest, as in many cases it reflects unique species-specific or lineage-specific adaptations. For example, Microsomal P450 has 34 proteins assigned in human. Multicellularity was suggested to be a driving force for P450 duplication as it is a natural choice for making and degrading mammalian signalling molecules like retinoic acid, thromboxane A2, steroids, and ecdysone.[31] Similarly, 22 human proteins are assigned to cyclic-nucleotide phosphodiesterase. Cyclic-nucleotide phosphodiesterase is a regulator of the cAMP signalling pathway, a central pathway in learning and memory.[32] Two reactions that are massively expanded only in the fly, glutathione transferase and cholinesterase, are primarily responsible for metabolic resistance to insecticides.[33] Several transferase and hydrolase reactions were massively expanded in all eukaryotic genomes examined; many of them are reactions involved in signalling and degradation, functional classes that have been massively expanded in metazoa. Only a single reaction was differentially expanded in bacteria; 30 proteins were assigned to the PEP-dependant phosphotransferase enzyme II in *E. coli*, a participant in the chemotactic pathway in motile bacteria.[34]

The origin and composition of the mammalian reaction set

After studying the characteristic patterns of enzymatic expansion we wanted to further analyse the composition of the mammalian reaction set. We divided the 796 reactions identified in mammals (human and mouse) into four phylogenetic groups: universal reactions, eukaryotic-specific reactions, metazoa-specific reactions and mammalian-specific reactions. Mammalian-specific reactions are reactions found only in mammals; metazoa-specific

Table 5. Reactions assigned to more than 20 proteins per species in one of the five model species in Figure 7

Class	Enzyme	Enzyme name	Species	Number of assigned proteins	Description[20]
Oxido-reductases	1.6.5.3	NADH dehydrogenase (ubiquinone)	H.sapiens	46	Mitochondrial protein complex.
	1.6.99.3	NADH dehydrogenase (cytochrome *c* reductase)	*H.sapiens*	45	Protein complex.
	1.9.3.1	Cytochrome-*c* oxidase	*H.sapiens*	21	Mitochondrial complex.
	1.14.14.1	Unspecific mono-oxygenase, microsomal P450	*H.sapiens*	34	A group of heme-thiolate proteins (P-450), acting on a wide range of substrates
Transferases	2.5.1.18	Glutathione transferase	*D.melanogaster*	26	A group of enzymes with broad specificity
	2.7.1.27	Erythritol kinase	*S.cerevisiae*	24	
			D.melanogaster	44	
			H.sapiens	76	
	2.7.1.37	Protein kinase	*S.cerevisiae*	25	A broad specificity group of enzymes that are under review by the NC-IUBMB. Signalling
			D.melanogaster	51	
			H.sapiens	132	
	2.7.1.69	PEP-dependant phospho-transferase enzyme II	*E. coli*	30	Comprises a group of related enzymes
	2.7.1.112	Protein-tyrosine kinase	*D.melanogaster*	28	The reaction includes all enzymes acting as protein-tyrosine kinases. All phosphorylated proteins, regardless of their function and nature, are commonly considered as the substrate in the reaction. Signalling
			H.sapiens	84	
	2.7.7.6	DNA-directed DNA polymerase	*S.cerevisiae*	29	Protein complex
			H.sapiens	22	
	2.7.7.49	RNA-directed DNA polymerase	*H.sapiens*	115	Protein complex
Hydrolases	3.1.1.8	Cholinesterase	*D.melanogaster*	24	Acts on a variety of choline esters and a few other compounds
	3.1.2.15	Ubiquitin thiolesterase	H.sapiens	27	Degradation.
	3.1.3.16	Phosphoprotein phosphatase	*D.melanogaster*	22	A group of enzymes removing the serine- or threonine-bound phosphate group from a wide range of phosphoproteins. Signalling
			H.sapiens	36	
	3.1.3.48	Protein-tyrosine-phosphatase	*H.sapiens*	61	Dephosphorylates *O*-phosphotyrosine groups in phosphoproteins. Signalling
	3.1.4.17	3′,5′-Cyclic-nucleotide phosphodiesterase	*H.sapiens*	22	Regulator of the cAMP signalling pathway[28]
	3.4.21.1	Chymotrypsin	*D.melanogaster*	26	Broad specificity of substrates. Degradation
			H.sapiens	52	
	3.4.21.4	Trypsin	*D.melanogaster*	35	Broad specificity of substrates. Degradation
			H.sapiens	37	
	3.4.25.1	Proteasome endopeptidase complex	*D.melanogaster*	24	Degradation
	3.6.3.14	H+-transporting two-sector ATPase	*S.cerevisiae*	28	Mitochondrial complex.
			D.melanogaster	31	
			H.sapiens	42	
Isomerases	5.2.1.8	Peptidylprolyl isomerase	*H.sapiens*	74	Broad specificity of substrates. Involved in protein folding
Ligases	6.3.2.19	Ubiquitin–protein ligase	*D.melanogaster*	22	All protein-lysine, are commonly considered as the substrate in the reaction. Degradation
			H.sapiens	30	

reactions are those mammalian reactions that are found only in metazoa (but are not mammalian-specific); eukaryote-specific reactions are mammalian reactions that are found only in eukaryote species (but are not metazoan-specific); universal reactions are those mammalian reactions that are also found in prokaryote species.

44% of the mammalian reactions are classified as universal, 13% as eukaryotic-specific, 17% as metazoa-specific, and 26% as mammalian-specific (Figure 8). The high fraction of universal reactions in mammals indicates a high level of conservation of the enzymatic sets, which is compatible with previous studies that had reported the existence of an extensive conserved core of metabolic enzymes common to archaea, bacteria and eukaryota.[5] The distribution of the remaining reaction groups indicates that a more limited set of reactions is eukaryotic or metazoa-specific, suggesting that the phylogenesis of these groups did not involve a massive recruitment of new reactions. Surprisingly, we found that a relatively big group of reactions (26%) is mammalian-specific.

We studied the reaction class distribution of the different reaction groups (Figure 8). The ligase reaction class, one of the most non-redundant reaction classes in human (Figure 7), is the most conserved one: almost all reactions are universal. The hydrolases reaction class, the most redundant reaction class in human, has many enzymes that are metazoa and mammalian-specific. Most of the metazoa and mammalian-specific hydrolases are peptidases (58 and 64%, respectively).

Discussion

We present here a comprehensive analysis of the complement of enzymes in a large variety of species. By using information retrieved from the well annotated SWISS-PROT[11] database, together with sequence information from a variety of fully

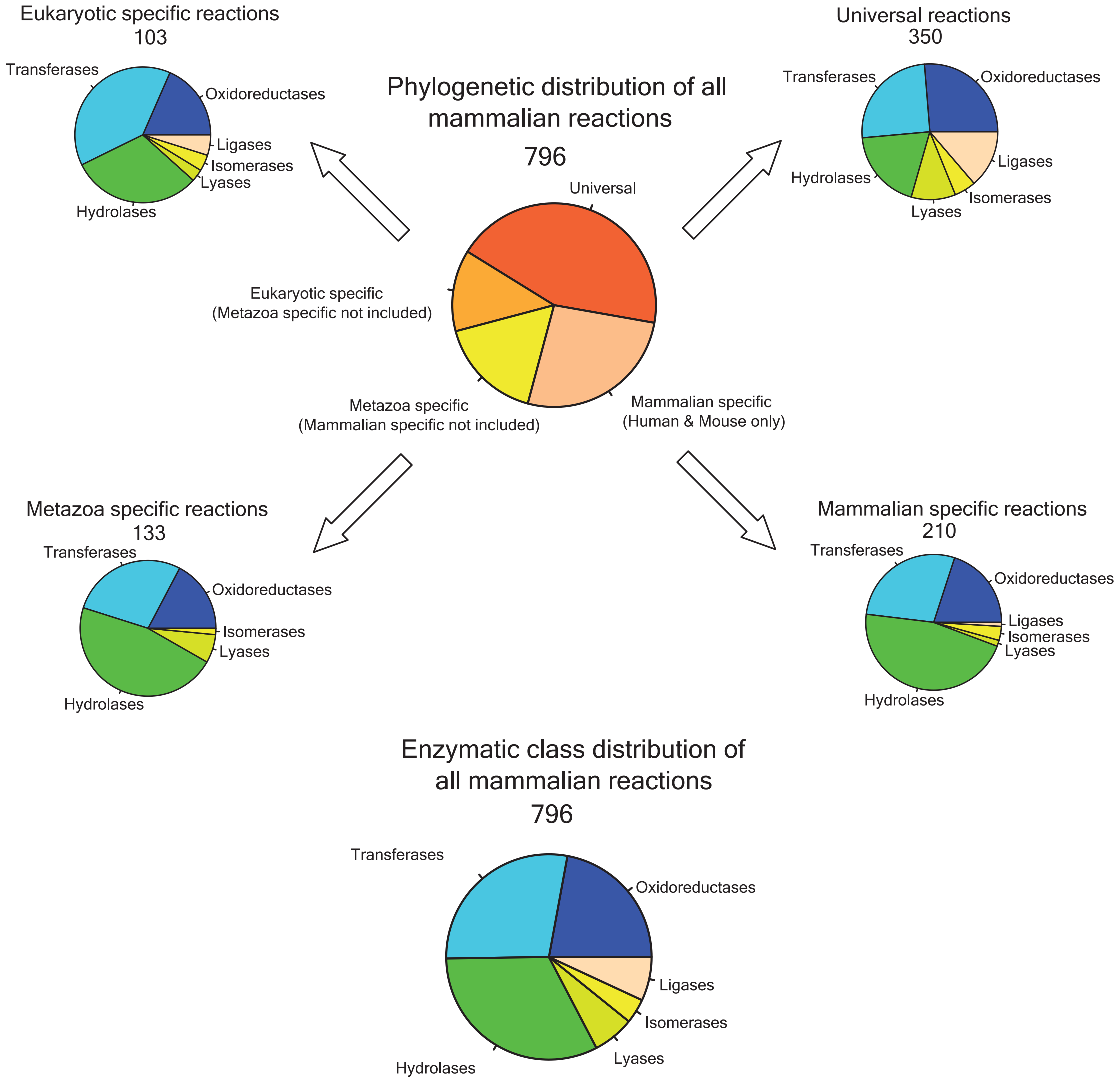

Figure 8. The distribution of mammalian reactions into phylogenetic groups and into reaction classes. The reactions are those identified in the conservative set. Numbers (on top of each circle) represent the number of reactions in each group.

sequenced genomes and information from the EC functional scheme,[7] we have aimed here to estimate the fraction of enzymes in genomes, to determine the extent of their functional redundancy in different domains of life and to identify functional innovations and lineage-specific expansions in the metazoa lineage.

While performing such a large-scale analysis, several important limitations must be acknowledged:

(1) The scarcity of sequenced eukaryote species. As only a limited number of eukaryote and metazoa genomes are available it is not yet clear how general the trends we have identified are. Future sequencing of additional unicellular and multicellular eukaryote species will provide a better understanding of the extent to which differences between eukaryote and prokaryote species can be related to the transition from the prokaryote to the eukaryote cell or to the transition from unicellularity to multicellularity.

(2) Inferring function from sequence. As function has been experimentally proven for only a small number of proteins,[21] calculating the number of reactions per species requires the inference of function from sequence. Here, based on previous studies of function-sequence dependency in enzymes,[13] we have transferred annotations from a query protein to hits sharing at least 40% sequence identity, when the query protein has an 80% overlap with the hit. Such annotation transfer has been found to be accurate in the large majority of examined enzyme-pairs, although not in all. As at the moment there are no fully experimentally annotated genomes, we are forced to rely on sequence-based annotations. In a study of the sequence-function dependency in single and multi-domain proteins, Hegyi & Gerstein have shown that in multi-domain proteins in a case of complete coverage along the full length of both proteins, function is conserved in 90% of the protein pairs.[35] We repeated our analysis while conditioning annotation transfer in mutual full coverage (80% of the length) between query and hit. Our observations under the above limitation are consistent with the results obtained by conditioning annotation transfer only by full coverage (80%) of the query protein (results not shown). As shown, our results are also consistent with the results obtained when using the functionally annotated KEGG database.[8,14]

(3) The use of the EC scheme for estimating functional redundancy. Proteins sharing the same EC number can be related in several ways. They may be proteins working in concert (e.g. in protein complexes), duplicated genes that are predicted to have the same function and genes that do not share sequence similarity but still catalyse the same reaction. The two last groups might represent specific adaptations to different regulatory modes. We have performed an analysis in order to verify that the increase in functional redundancy observed in higher species is indeed the result of family expansion rather than the result of having more reactions that describe proteins working in concert (see Materials and Methods).

An additional challenge when using the EC scheme for determining functional divergence is that some EC numbers describe generic reactions where the compounds are not fully specified (e.g. protein-kinases). A few such EC reactions, that have many proteins assigned to them, are listed in Table 5. Here, we have verified that the increased functional redundancy observed in eukaryotic species reflects a general trend rather than lineage-specific expansion of proteins assigned to a limited number of broadly defined reaction types. First, we show that increased functional redundancy is observed for most of the functional classes (i.e. not limited to protein-kinases or protein-peptidases). Next, we show that in most reaction classes extensive recruitment of enzymes is accompanied by a general increase in the number of redundant reactions (i.e. increase in the number of reactions to which more than a single enzyme is assigned). We therefore conclude that, although the EC scheme is not ideal for studying functional redundancy in a few broadly defined reactions such as protein-kinases and peptidases, the functional redundancy reported in the analysis is much more general.

Despite the above caveats, since we are basing our analysis on several sources of information, we believe that the general observations made here are valid.

A few major trends emerge from the analysis:

Enzymatic fraction

The fraction of enzyme-related proteins is approximately constant and ranges between 30 and 40% of the genome in most prokaryote species. The trend observed is common to both archaea and bacteria species where extreme values seem to reflect species-specific adaptations rather than a phylogenetic trend. The relatively low fraction of enzymes found in eukaryote species might be related to a massive recruitment of regulatory proteins.[16,18,19] We therefore suggest that the rate of enzyme expansion in all domains is approximately constant, and that the differences in the fraction of enzymes mainly reflect dramatic changes in the size of other functional categories.

Increased functional redundancy in eukaryote species

In eukaryotic species the enzymes/reactions ratio is higher than in prokaryotic species and therefore the functional redundancy in these species must increase. Whilst enzymatic sets grow when proteome size increases in all species, eukaryotic and prokaryotic species differ in their pattern of expansion. In prokaryotic species the expansion of the enzymatic set mainly reflects the broadening of the reaction repertoire, whilst the expansion of the eukaryotic set, and especially of multicellular eukaryotes, is to a larger extent the result of an increase in reaction redundancy. An increased sequence redundancy accompanying an increase in the number of genes was previously reported.[24–27] The quantitative assessment performed here for the enzymatic set of each species, confirms and quantifies this general trend. Whereas 58% of the enzymatic reactions in human are found to be redundant, less than 10% of the enzymatic reactions in the intracellular symbiont bacteria *Buchnera* are performed by more than a single enzyme.

Two fundamental differences between unicellular prokaryotes and multicellular eukaryotes can contribute to different adaptation strategies regarding the addition of new enzymes. First, genomes of eukaryotic species lack physical constraints on their size and therefore the selective pressure to lose redundant genes is less efficient than that in bacterial genomes.[30] Second, unlike unicellular species, cells of multicellular species function in a relatively stable environment and their metabolic diversity is therefore more likely to represent spatial adaptations.

The origin of the mammalian reaction set

The mammalian set of reactions includes 44% universal reactions. Approximately half of the remaining reactions are mammalian-specific, suggesting that the evolution of eukaryotes and metazoa did not involve a massive recruitment of new reactions.

An increase in the number of enzymes and reactions involved in regulatory processes in the mammalian enzymatic set

Hydrolases are the most prolific mammalian-specific reaction class (46% of mammalian-specific reactions compared to 19% of universal reactions) and most of these reactions are peptidases. The increase in the functional repertoire of hydrolase reactions is accompanied by a massive increase in the number of hydrolase genes (33% of human enzymes compared to 19% of *E. coli* enzymes). A massive increase in the number of genes is also observed for the transferase reactions, where the most massively expanded reactions are protein-kinases. Protein-kinases are involved in signal transduction. Peptidases play an important role in regulating the activity and fate of many proteins and have an essential role in the control of cell behaviour, survival and death.[36]

Here we have discussed the relationship between the expansion of a functional group (of enzymes) and changes in the size and composition of its functional repertoire. Our results suggest that a universal enzymatic core vertically inherited from the eukaryotic ancestral species remains highly conserved in mammals. New enzymes added mostly contribute to an increase in functional redundancy. In the future we intend to focus on the contribution made by new reactions added to mammalian metabolism, and to study the relationship between the increased functional redundancy and mammalian spatial complexity.

Materials and Methods

Construction of the query list

In order to ensure that the enzyme set for each species is as complete and validated as possible, we have listed all the highly curated UniProt[37] enzymes. The list is composed of SWISS-PROT[11] (release 41.0) proteins assigned with an EC number. In order to exclude assignments based on sequence similarity, annotations including "hypothetical protein", "by similarity", "putative" or "probable" were filtered from the list, as well as proteins for which all references are common to more than 20 entries. TrEmbl[11] (release 23) entries with the "experimental" evidence code were included in the list. The final enzyme list known as the "query list" is composed of 23,431 proteins. The distribution of the enzymes between the three domains of life is the following: 14,128 eukaryota (1869 human), 7653 bacteria, 574 archaea, and 1076 viruses. The comprehensiveness of the query list composition aims to cover all possible domain combinations.

Each enzyme in the list was used as a query for a PSI-BLAST search against fully sequenced genomes, including 63 bacteria species, 16 archaea species and six eukaryota species.

Construction of enzymatic sets for each species

Homologues to each of the SWISS-PROT sequences, in the constructed enzymatic query list, were retrieved from the Biopendium™ (version 13, created 21 March 2003)[38] using PSI-BLAST[12] to three iterations. Homologues were only accepted if there was an overlap of 80% between the SWISS-PROT query sequence and the homologue sequence, and if the homologue was found within the pre-defined set of completed genomes. This includes completed genomes from Ensembl human[39] (release 11.31), Ensembl mouse (release 11.3), and the NCBI completed genomes (15 August 2002). It should be remembered that no completed genome is truly complete and the dataset used here is therefore a snapshot of the current knowledge within the Biopendium™.

Since Ensembl fly and worm genomes were not included in Biopendium™ version 13, these genomes were constructed by comparing all fly/worm sequences in the Biopendium™ against the latest available Ensembl genomes (6 August 2003). All sequences from either *C. elegans* or *D. melanogaster* were retrieved from the Biopendium™ and processed. The most complete resulting genomes were 98.8% complete for *C. elegans* (at 90% identity and 80% overlap) and 99.1% complete for *D. melanogaster* (at 95% identity and 80% overlap).

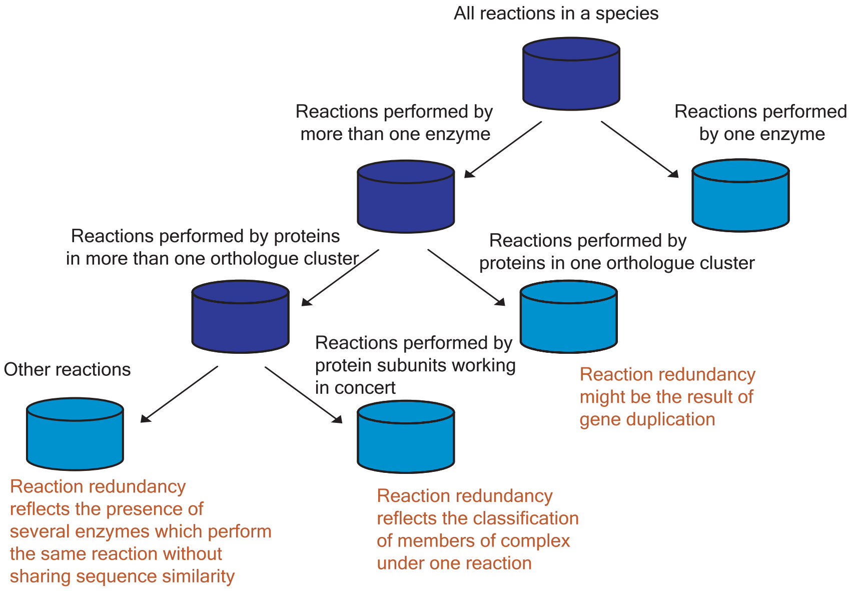
All reactions in a species
Reactions performed by more than one enzyme
Reactions performed by one enzyme
Reactions performed by proteins in more than one orthologue cluster
Reactions performed by proteins in one orthologue cluster
Other reactions
Reactions performed by protein subunits working in concert
Reaction redundancy might be the result of gene duplication
Reaction redundancy reflects the presence of several enzymes which perform the same reaction without sharing sequence similarity
Reaction redundancy reflects the classification of members of complex under one reaction

Figure 9. Description of possible ways in which enzymes that perform the same reaction in a species are related.

KEGG database

The Kyoto Encyclopedia of Genes and Genomes[8,14] gene catalogue contains its own functional annotations together with SWISS-PROT, GenBank and the original genome project teams' annotations. KEGG's EC assignments for a newly sequenced organism are made by considering orthologous relationships. Orthologue clusters in KEGG are based not only on sequence similarity but also on positional correlation of genes on the chromosome. Their amino acid sequence similarity is determined using the SSEARCH program based on the Smith–Waterman algorithm. The number of enzymes in a species was downloaded at January 2004†.

The number of reactions (fully assigned EC numbers) per species was extracted from the enzyme file in the LIGAND section by automated text parsing (downloaded at February 2004).

Determination of the representation of a species or a group of species in the query list

In order to examine how well a species is represented in the query list, we counted all proteins in a species matching an enzyme from the query list with more than 80% identity. We used the 80% cut-off rather than counting directly the number of enzymes per species in the query list in order to transfer annotations between closely related species. For example, if an enzyme in the query list was identified in one strain of *E. coli* the use of the 80% cut-off will enable us to transfer the annotation to other *E. coli* strains. The average fraction of these highly curated enzymes (as determined by the 80% cut-off) over all species was calculated, as well as the *z*-score (number of standard deviations from the mean in each species, Table 2).

In order to study the level of representation in the query list for each domain of life we repeated the same procedure, done for each species, for each group of species (e.g. bacteria, archaea and eukaryota). The fraction of highly curated enzymes per group was calculated by summing the number of experimentally validated enzymes in all species classified to the group and dividing it by the summed number of genes. As our data set includes many more bacteria species (63) than archaea (16) and eukaryota (6) we have divided the bacterial species into the lower level classification of subfamilies, repeating the same procedure (Table 3). A complete list of entries per species in the query list is available‡.

Classification of mammalian reactions into phylogenetic groups

Mammalian reactions are those reactions identified in the human and mouse conservative set. The classification of the mammalian reactions to a phylogenetic group is derived as follows: universal reactions are mammalian reactions found in more than five prokaryote species and at least a single non-mammalian eukaryote species; eukaryotic-specific reactions are those mammalian reactions found in a yeast species and in no more than five prokaryote species; metazoan-specific reactions are those mammalian reactions found in fly or worm but not in yeast and in no more than five prokaryote species; mammalian-specific reactions are those mammalian reactions found in mammals but not in other eukaryote and in no more than five prokaryote species.

Analysis of the relationships between proteins classified to the same EC reaction

Proteins sharing the same EC number can be related in several ways. They may be protein subunits working in concert (e.g. in protein complexes), proteins that are assigned to the same function due to high sequence similarity or proteins that do not share sequence similarity but are predicted to catalyse the same reaction. We have used the KEGG database in order to estimate the fraction of reactions that are assigned to more than one

† http://www.genome.ad.jp/kegg/docs/upd_genes.html

‡ http://www.ebi.ac.uk/~shirigo/supplemental1.html

Table 6. Distribution of different reaction groups in species

		HAS	MMU	CEL	SCE	SPO	ECO	BSU	PAE	STM
All reactions		816	610	416	491	408	656	531	582	663
Reactions with one enzyme		540	428	270	334	296	513	393	400	525
Reactions with multi enzymes	One orthologue cluster	206 (74%)	131 (72%)	111 (76%)	119 (75%)	79 (70%)	61 (42%)	86 (62%)	114 (62%)	58 (42%)
	Multi orthologue clusters where reactions do not describe proteins working in complex	49 (18%)	35 (19%)	17 (12%)	16 (10%)	10 (9%)	43 (30%)	22 (16%)	23 (13%)	39 (28%)
	Multi orthologue clusters where reactions describe proteins working in complex	21 (8%)	16 (9%)	18 (12%)	22 (14%)	23 (21%)	39 (27%)	30 (22%)	45 (25%)	41 (30%)

Number in brackets represents the % out of all reactions with multi enzymes. Data were extracted from the KEGG database (see Materials and Methods). HAS, *H.sapiens;* MMU, *M.musculus;* CEL, *C.elegans;* SCE, *S.cerevisiae;* SPO, *S.pombe;* ECO, *E. coli;* BSU, *B.subtilis;* PAE, *P.aeruginosa;* STM, *S.typhimurium.*

protein and the fraction of reactions that are assigned to more than a single orthologous cluster (Figure 9). The classification of genes to reactions was extracted from the ligand section in KEGG. The classification of genes to orthologue clusters was extracted from the "ko" section in KEGG. Data were downloaded on August 18th, 2004.†

In 71–76% of the multi-protein reactions in the five eukaryotes examined (*H. sapiens, M. musculus, C. elegans, S. cerevisiae, S. pombe*) all enzymes are classified into a single orthologue cluster, suggesting that the functional redundancy is, in most cases, the result of a gene duplication event (Table 6). A text search was done on those enzymes classified to multi-protein, multi-orthologue-cluster reactions to look for annotations indicating that they are part of a complex (e.g. complex, subunit, component, chain). The high fraction of multi-protein reactions that are classified to one orthologue cluster together with the low fraction of "complex-reactions" in eukaryote species implies that the increase in functional redundancy observed is mainly the result of enhanced gene duplication, rather than an artefact caused by the increase in number of reactions representing proteins working in complex.

In a similar way we have analysed the distribution of various EC groups in the conservative set. The reaction distribution was studied in five species: *H. sapiens, D. melanogaster, S. cerevisiae, E. coli* and *Buchnera*. Clustering of proteins into sequence groups was done according to their similarity to the query protein that identified them, i.e. proteins that were recognised by the same query protein (40% identity, 80% overlap) are clustered together. Similar results to those observed from the KEGG database are obtained using the conservative set. Finally, we have studied the sequence similarity between protein pairs in 156 reactions in human assigned to two proteins. A total of 79% of the protein pairs share more than 40% sequence identity (not shown).

Acknowledgements

We thank Hannes Ponstingl, Kirill Degtyarenko, Juan A.G. Ranea, Sophia Tsoka and Leon Goldovsky for their helpful comments. S.F. is supported by an EMBL fellowship. R.V.S. and R.A.G. are supported by the DTI LINK programme in Applied Genomics funded by the BBSRC.

References

1. Ranea, J. A., Buchan, D. W., Thornton, J. M. & Orengo, C. A. (2004). Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol.* **336**, 871–887.
2. Cases, I., de Lorenzo, V. & Ouzounis, C. A. (2003). Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol.* **11**, 248–253.
3. van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet.* **19**, 479–484.
4. Ranea, J. A., Grant, A., Thornton, M. J. & Orengo, C. A. (2005). Microeconomic principles of optimal genome size in bacteria. *Trends Genet. Sci.* **21**, 21–25.
5. Peregrin-Alvarez, J. M., Tsoka, S. & Ouzounis, C. A. (2003). The phylogenetic extent of metabolic enzymes and pathways. *Genome Res.* **13**, 422–427.
6. Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G. & Little, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science*, **271**, 470–477.
7. Tipton, K. & Boyce, S. (2000). History of the enzyme nomenclature system. *Bioinformatics*, **16**, 34–40.
8. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucl. Acids Res.* **30**, 42–46.
9. Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E., Jr, Kyrpides, N. *et al.* (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucl. Acids Res.* **28**, 123–125.
10. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M. & Pellegrini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucl. Acids Res.* **28**, 56–59.
11. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E. *et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365–370.
12. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped

† ftp://ftp.genome.ad.jp/pub/kegg/

BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

13. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
14. Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **28**, 27–30.
15. Aguilar, D., Aviles, F. X., Querol, E. & Sternberg, M. J. (2004). Analysis of phenetic trees based on metabolic capabilities across the three domains of life. *J. Mol. Biol.* **340**, 491–512.
16. Andrade, M. A., Ouzounis, C., Sander, C., Tamames, J. & Valencia, A. (1999). Functional classes in the three domains of life. *J. Mol. Evol.* **49**, 551–557.
17. Tamames, J., Ouzounis, C., Sander, C. & Valencia, A. (1996). Genomes with distinct function composition. *FEBS Letters*, **389**, 96–101.
18. Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S. *et al*. (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science*, **282**, 2022–2028.
19. Aravind, L. & Subramanian, G. (1999). Origin of multicellular eukaryotes—insights from proteome comparisons. *Curr. Opin. Genet. Dev.* **9**, 688–694.
20. Barrett, A. J. (1994). Classification of peptidases. *Methods Enzymol.* **244**, 1–15.
21. Andrade, M. A. & Sander, C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotechnol.* **8**, 675–683.
22. Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249.
23. Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Struct. Funct. Genet.* **41**, 98–107.
24. Enright, A. J., Kunin, V. & Ouzounis, C. A. (2003). Protein families and TRIBES in genome sequence space. *Nucl. Acids Res.* **31**, 4632–4638.
25. Krakauer, D. C. & Plotkin, J. B. (2002). Redundancy, antiredundancy, and the robustness of genomes. *Proc. Natl Acad. Sci. USA*, **99**, 1405–1409.
26. Muller, A., MacCallum, R. M. & Sternberg, M. J. (2002). Structural characterization of the human proteome. *Genome Res.* **12**, 1625–1641.
27. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, **278**, 631–637.
28. Moran, N. A. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr. Opin. Microbiol.* **6**, 512–518.
29. Moran, N. A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, **108**, 583–586.
30. Lynch, M. & Conery, J. S. (2003). The origins of genome complexity. *Science*, **302**, 1401–1404.
31. Nelson, D. R. (1999). Cytochrome P450 and the individuality of species. *Arch. Biochem. Biophys.* **369**, 1–10.
32. Bolger, G., Michaeli, T., Martins, T., St John, T., Steiner, B., Rodgers, L. *et al*. (1993). A family of human phosphodiesterases homologous to the dunce learning and memory gene product of *Drosophila melanogaster* are potential targets for antidepressant drugs. *Mol. Cell. Biol.* **13**, 6558–6571.
33. Ranson, H., Claudianos, C., Ortelli, F., Abgrall, C., Hemingway, J., Sharakhova, M. V. *et al*. (2002). Evolution of supergene families associated with insecticide resistance. *Science*, **298**, 179–181.
34. Lux, R., Jahreis, K., Bettenbrock, K., Parkinson, J. S. & Lengeler, J. W. (1995). Coupling the phosphotransferase system and the methyl-accepting chemotaxis protein-dependent chemotaxis signaling pathways of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **92**, 11583–11587.
35. Hegyi, H. & Gerstein, M. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* **11**, 1632–1640.
36. Puente, X. S. & Lopez-Otin, C. (2004). A genomic analysis of rat proteases and protease inhibitors. *Genome Res.* **14**, 609–622.
37. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S. *et al*. (2004). UniProt: the Universal Protein knowledgebase. *Nucl. Acids Res.* **32**, D115–D119.
38. Swindells, M., Rae, M., Pearce, M., Moodie, S., Miller, R. & Leach, P. (2002). Application of high-throughput computing in bioinformatics. *Philos. Transact. ser. A Math. Phys. Eng. Sci.* **360**, 1179–1189.
39. Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L. *et al*. (2004). An overview of Ensembl. *Genome Res.* **14**, 925–928.

Edited by M. Levitt

(Received 19 January 2005; received in revised form 10 April 2005; accepted 12 April 2005)