

# A systematic analysis of lineage-specific evolution in metabolic pathways

Himanshu Ardawatia, David A. Liberles\*

*Computational Biology Unit, BCCS, University of Bergen, 5020 Bergen, Norway  
Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA*

Received 13 April 2006; received in revised form 30 July 2006; accepted 10 August 2006

Available online 24 August 2006

Received by T. Gojobori

## Abstract

In a search for the lineage-specific evolution of pathways between human, chimpanzee, mouse, and rat, orthologous gene families were generated from genome sequences. For each family, a model-based ratio of nonsynonymous to synonymous nucleotide substitution rates was calculated. Where the free-ratio model of individual ratios on each branch was supported, these families were mapped to two databases of metabolic pathways (KEGG and BioCyc) and the lineage-specific evolution of pathways was evaluated. The most similar pathway evolution was seen between mouse and rat, while the evolutionary pattern between human and chimpanzee was less correlated. Individual pathways in the human lineage were observed to evolve in a faster, lineage-specific manner, including the pathway involving arachidonic acid metabolism (identified through the KEGG analysis) and pyrimidine metabolism (identified through both analyses).

© 2006 Elsevier B.V. All rights reserved.

**Keywords:** Positive selection; Systems biology; Mammals

## 1. Introduction

Understanding the comparative genomics of closely related species can provide mechanistic details on the evolution of biochemical diversity and novelty and how it relates to the phenotypic divergence of species (Crawford, 1989; Liberles, 2005). Complete genome sequencing coupled to individual gene sequencing has expanded the capacity to examine large numbers of genes from a comparative perspective between closely related species. This initially enabled the case by case search for positive selection in individual gene families (Benner et al., 1998; Yang and Bielawski, 2000). Subsequently, a large scale comparative analysis generated lists of genes undergoing positive selection

along all lineages for which data was available (Endo et al., 1996; Benner et al., 2000; Liberles et al., 2001; Clark et al., 2003; Roth et al., 2005; Roth and Liberles, 2006). However, these lists of genes have been treated independently, lacking a view of how proteins that interact biochemically or genetically may be co-evolving, leading to a pathway-level appreciation of positive selection and the generation of evolutionary novelty.

There are several layers of complexity to this type of analysis. The first layer (the focus of this paper) involves the differential evolution of genes that are conserved across species. Ultimately, gene duplication and loss complicate this perspective, potentially magnifying the evolutionary effects (Teichmann and Babu, 2004; Roth et al., in press).

Large scale experimental methods have generated interaction maps of proteomes from species with complete genomes (Jeong et al., 2000) and traditional metabolic pathways that are well characterized can also provide a highly reliable starting point for defining which proteins interact biochemically in the same pathway (Kanehisa et al., 2004). A phylogenetic perspective has previously been used to examine the evolution of metabolic enzymes from Prokaryotes (Dandekar et al., 1999; Forst and Schulten, 1999, 2001). These approaches suffered from the

*Abbreviations:* BioCyc, The BioCyc Collection of Pathway/Genome Databases; CDS, coding sequence; Ka, the nonsynonymous nucleotide substitution rate; Ka/Ks, the ratio of nonsynonymous to synonymous nucleotide substitution rates; KEGG, The Kyoto Encyclopedia of Genes and Genomes; Ks, the synonymous nucleotide substitution rate; RBBH, reciprocal best BLAST hit.

\* Corresponding author. Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA. Tel.: +1 307 766 5206; fax: +1 307 766 5098.

*E-mail address:* [liberles@uwyo.edu](mailto:liberles@uwyo.edu) (D.A. Liberles).

problem of treating a pathway as a unit and attempting to define a distance between entire pathways. The complexity of this approach precludes doing this for multicellular Eukaryotes because of the large number of gene duplication and loss events.

Here we present a study of phylogenetic analysis of human, chimp, mouse and rat proteins where the nonsynonymous to synonymous nucleotide substitution rate ratio (Ka/Ks) is calculated in a model-based approach to test for lineage-specific rates (Yang, 1998). All gene families, where duplication has not occurred or has occurred recently in a lineage-specific manner and where the lineage-specific model was supported were then mapped onto pathways in the KEGG database (Kanehisa et al., 2004) and the BioCyc database (Karp et al., 2005) to look for lineage-specific differences in the evolution of specific pathways.

## 2. Materials and methods

All, known and novel (Ensembl definitions), CDS (coding sequences) for the four species *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, and *Rattus norvegicus* were retrieved from FTP downloads of the Ensembl database (Birney et al., 2006). The sequences were then translated in all 6 reading frames and the longest peptide sequences without any stop codons among all frames were selected for further analysis.

Inter- and intra-species all-against-all BLAST was performed with this set of proteins for all four species using MPI-BLAST (<http://mpiblast.lanl.gov>; Qi and Lin, 2005). Reciprocal best hits (all-versus-all) were taken from parsed BLAST results (exp 1.0e-20; Identity cutoff=40%, Overlap cutoff=100) and all sequences sharing reciprocal best hit criteria with each other were given membership of specific sequence families, including the intra-species best hit as a control. In every protein family, a redundancy check was done to filter out protein sequences belonging to the same gene (either due to being in the same chromosomal location or due to chance mis-annotation in

Table 2

The phylogenetic structure of the gene families obtained is shown

No duplications	Lineage-specific duplications	Duplications deep in the tree
2394	99	2255

Only those families with no duplications or recent lineage-specific duplications were retained for the analysis.

Ensembl). The longer transcript with higher similarity with other sequences in the same family was chosen in such cases.

Multiple sequence alignments of the filtered translated protein sequences in each protein family were calculated using MUSCLE (Edgar, 2004) with default parameters. Corresponding multiple DNA sequence alignments were built using protein alignments (obtained in a previous step) as templates. The binary rooted distance phylogenetic trees were calculated for all CDS sequences in each protein family using the distance algorithm encoded in the Darwin library (Gonnet et al., 2000). This step served to confirm that the sequences obtained by pairwise best BLAST hit were orthologs with the same species hit as an outgroup.

To check whether the gene trees obtained from the above procedure were coherent with the species tree of the four species under investigation, the Forester algorithm (Zmasek and Eddy, 2001) was used to infer speciation and duplication events on the gene trees by comparison to the common species tree for the four species. Finally categories with ‘no duplications’, ‘duplications within lineages’ and ‘duplications outside lineages somewhere deep in the tree’ were defined for filtering of families and corresponding trees for further analysis.

The DNA multiple sequence alignments and phylogenetic trees obtained in the previous step for each protein family were used as input for calculating maximum likelihood estimates of Ka/Ks ratio for each branch of the phylogenetic tree under a codon-based substitution model. Two models were used. The null model used a single Ka/Ks ratio for all branches of the tree (selective pressure assumed to be the same under all lineages) and the free-ratio model used different Ka/Ks ratio parameters for each branch of the phylogenetic tree (selective pressure

Table 1

Following the all-against-all BLAST searches, the initial composition of families is shown

	Family composition (H = human, M = mouse, C = chimpanzee, R = rat)	Number of families
1	HHMMP (6)	55
2	HHMMRPP (7)	193
3	HHMMRRP (7)	67
4	HHMMRRPP (8)	435
5	HHMMP (5)	66
6	HHMMRPP (6)	210
7	HHMRRP (6)	31
8	HHMRRPP (7)	116
9	HMMMP (5)	342
10	HMMRPP (6)	30
11	HMMRRP (6)	343
12	HMMRRPP (7)	58
13	HMRP (4)	2488
14	HMRPP (5)	116
15	HMRRP (5)	241
16	HMRRPP (6)	31
Total	All	4822

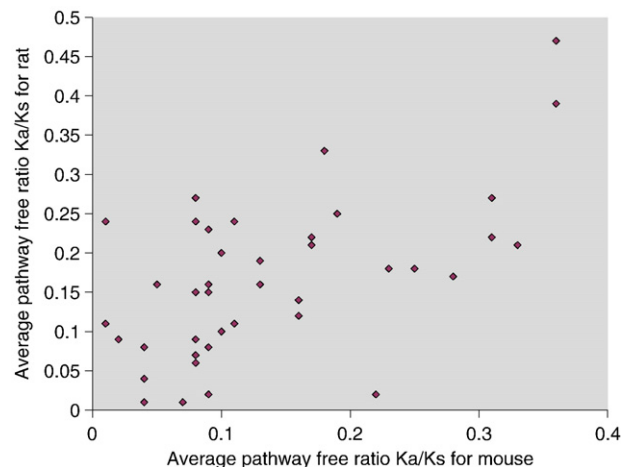


Fig. 1. For all pathways where the free-ratio model had significant datapoints for both mouse and rat, the average lineage-specific Ka/Ks value is plotted for the two lineages.

assumed to vary among lineages). The null model with few parameters is “nested” in the free-ratio model (with more parameters). Codeml from the PAML package (Yang, 1997) was used for these Ka/Ks calculations.

Likelihood ratio tests were performed to determine whether the free-ratio model provided significantly better fit to the data than the null model. For this purpose the likelihood ratio test statistic was calculated as below,

$$\text{LRT} = 2 * (\ln L1 - \ln L2)$$

where,  $\ln L1$  = LogLikelihood under free-ratio model and  $\ln L2$  = LogLikelihood under the null model. The likelihood ratio test statistic (LRT) was compared to the critical values of the chi-square distribution (with degrees of freedom = number of parameters in free-ratio model – number of parameters in null model), evaluated at a 95% significance level.

Pathway data and corresponding sequences for all four species were obtained from the comprehensive KEGG database (Kanehisa et al., 2004). Human, mouse, and rat protein sequences for proteins present in various KEGG metabolic and signaling pathways were downloaded from the FTP site of the KEGG database. Chimp sequences have not been systematically mapped in KEGG, so chimp sequences were extrapolated to human pathways. Intra-species MPI-BLAST was performed between species-specific sequences in different protein families and corresponding KEGG sequences in different pathways and reciprocal best hits were selected (BLAST result parsing with  $\text{exp } 1.0e-20$ ; Identity cutoff=40%; Overlap cutoff=100). The memberships of the corresponding reciprocal best hits obtained above were then investigated with respect to protein families and KEGG pathways. An identical procedure was followed with the BioCyc (specifically, HumanCyc) Pathway database (Karp et al., 2005), with extrapolation from human pathways.

Annotation by homologous transfer was used to extrapolate the pathway membership of a given protein to other proteins of other species in the same protein family. For example, a pathway map for a mouse protein in a family was extrapolated to other species sequences in the family and similarly in case of other species.

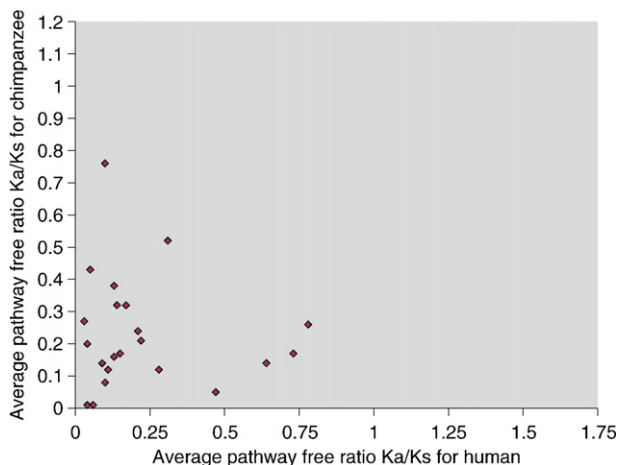


Fig. 2. For all pathways where the free-ratio model had significant datapoints for both human and chimpanzee, the average lineage-specific Ka/Ks value is plotted for the two lineages.

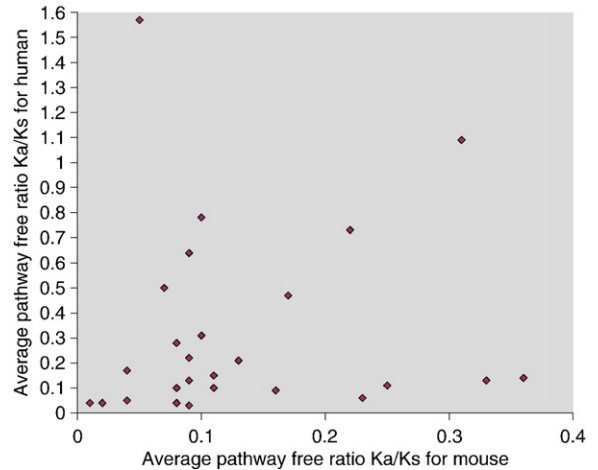


Fig. 3. For all pathways where the free-ratio model had significant datapoints for both human and mouse, the average lineage-specific Ka/Ks value is plotted for the two lineages.

The Ka/Ks values for all individual proteins (where the free-ratio model was supported and where it was not) were mapped to the appropriate KEGG pathway and the mean and standard error of the mean were calculated for each lineage and pathway. This was done for both the families with lineage-specific duplicates and those without. For families with lineage-specific duplicates, the Ka/Ks value of the branch from the ancestral duplication node to the speciation node was used. Spearman's rank correlation studies were done for finding correlations between pathway Ka/Ks averages among all species and a heatmap with hierarchical clustering was generated for the data. All statistical analysis including the hierarchical clustering of pathway Ka/Ks average per species and heatmap generation was done in R statistical programming language (<http://www.r-project.org>) or with MayDay (Dietzsch et al., 2006).

The significance of the KEGG pathway evolution was further assessed by changing the pathway boundaries systematically by both removing 1–2 enzymes from the boundaries of each pathway and adding 1–2 enzymes at the boundary from

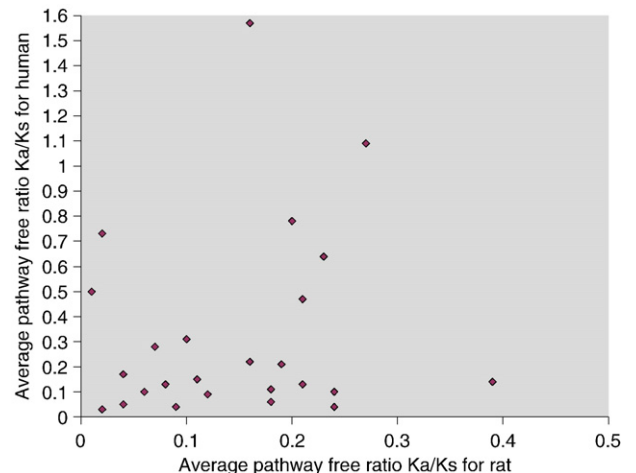


Fig. 4. For all pathways where the free-ratio model had significant datapoints for both human and rat, the average lineage-specific Ka/Ks value is plotted for the two lineages.

neighboring pathways. These perturbed pathways were then analyzed for significantly different average Ka/Ks ratios for all species and for the control, as assessed by the standard error of the mean.

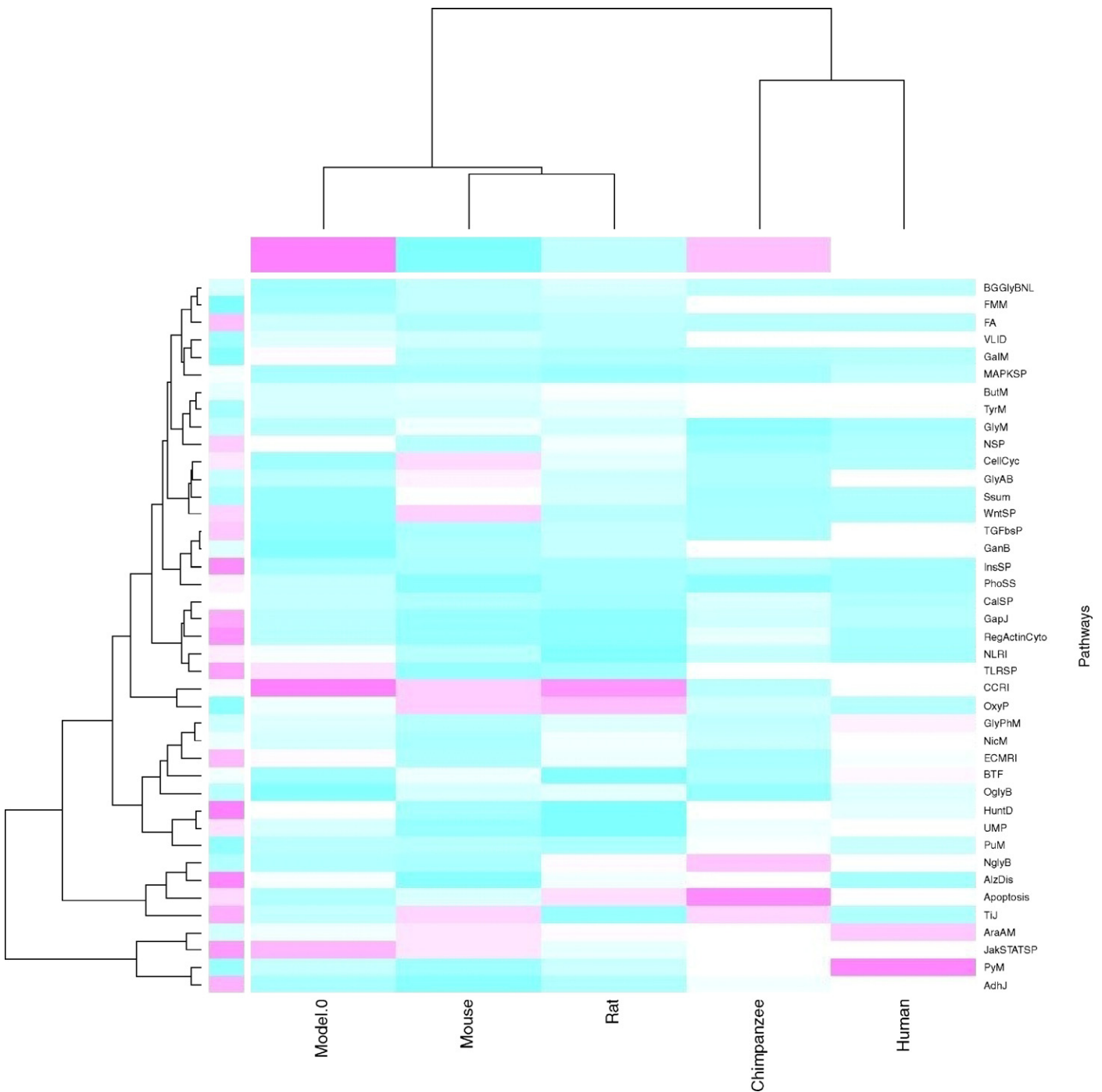
### 3. Results

#### 3.1. Family construction

33,869 human CDS, 39,649 chimpanzee CDS, 31,535 mouse CDS and 35,951 rat CDS were downloaded for analysis.

Using all-versus-all inter- and intra-species reciprocal best BLAST hits (RBBH) among all non-redundant sequences, gene families were defined. Families with sizes varying from 4 sequences to 8 sequences were generated, as shown in Table 1. Multiple sequence alignments and phylogenetic trees were calculated for these families as described.

All gene trees can be reconciled to a species tree by assuming gene duplication or loss events. In general, gene trees need to be mapped to the species tree in order to correctly define the orthologous and paralogous (including in-paralogous) relationships within a given protein family and this was done using the



Forester algorithm (Zmasek and Eddy, 2001). Thus, categories with ‘no duplications’, ‘duplications within lineages’ and ‘duplications outside lineages somewhere deep in the tree’ were defined, as shown in Table 2.

Only families which had either no duplications or which had only lineage-specific duplications were used for pathway mapping (2493 families). Lineage-specific duplications were found in 47 rat protein families, 54 mouse protein families, only 2 human protein families, and no lineage-specific duplication was found in the chimp lineage in these families.

### 3.2. Measuring selection with the Ka/Ks ratio

Two models were used for calculating Ka/Ks, the null model where a single Ka/Ks ratio was estimated for all

branches of the tree (selective pressure was assumed to be the same under all lineages) and the free-ratio model, where different Ka/Ks ratios were estimated for each branch of a phylogenetic tree (selective pressure was assumed to vary among lineages) (Yang, 1998). The null model with few parameters is “nested” in the free-ratio model (with more parameters). Likelihood ratio tests were performed to test if the free-ratio model provided significantly better fit to the data than the null model at the 95% significance level, given the extra parameters. At the 95% significance level, for 1848 out of 2493 protein families, the free-ratio model was not found to fit the data significantly better than the null model. In the other 605 protein families, the free-ratio model was found to fit the data significantly better than the null model.

Fig. 5. A heatmap shows the relative values of the average Ka/Ks for each pathway as derived from the KEGG database (Kanehisa et al., 2004), where free-ratio values were significant. The average one ratio values in each pathway are also shown to provide a baseline for the families where that model was supported (Model.0). Hierarchical clustering shows the relationships of the average Ka/Ks value in each row (pathway) and column (species) to each other. The annotation legend for the pathways is shown below.

Pathway ID (KEGG heatmap)	Pathway name (KEGG)
BGGlyBNL	Blood group glycolipid biosynthesis — neo-lactoseries
FMM	Fructose and mannose metabolism
FA	Focal adhesion
VLID	Valine, leucine and isoleucine degradation
GalM	Galactose metabolism
MAPKSP	MAPK signaling pathway
ButM	Butanoate metabolism
TyrM	Tyrosine metabolism
GlyM	Glycerolipid metabolism
NSP	Notch signaling pathway
CellCyc	Cell cycle
GlyAB	Glycosylphosphatidylinositol (GPI) — anchor biosynthesis
SsuM	Starch and sucrose metabolism
WntSP	Wnt signaling pathway
TGFbSP	TGF-beta signaling pathway
GanB	Ganglioside biosynthesis
InsSP	Insulin signaling pathway
PhoSS	Phosphatidylinositol signaling system
CalSP	Calcium signaling pathway
GapJ	Gap junction
RegActinCyto	Regulation of actin cytoskeleton
NLRI	Neuroactive ligand–receptor interaction
TLRSP	Toll-like receptor signaling pathway
CCRI	Cytokine–cytokine receptor interaction
OxyP	Oxidative phosphorylation
GlyPhM	Glycerophospholipid metabolism
NicM	Nicotinate and nicotinamide metabolism
ECMRI	ECM–receptor interaction
BTF	Basal transcription factors
OglyB	O-Glycan biosynthesis
HuntD	Huntington’s disease
UMP	Ubiquitin mediated proteolysis
PuM	Purine metabolism
NglyB	N-Glycan biosynthesis
AlzDis	Alzheimer’s disease
Apoptosis	Apoptosis
TiJ	Tight junction
AraAM	Arachidonic acid metabolism
JakSTATSP	Jak-STAT signaling pathway
PyM	Pyrimidine metabolism
AdhJ	Adherens junction

### 3.3. Mapping Ka/Ks to pathways and pathway evolution rate metric

From all the four lineages, proteins found to have significant free-ratio Ka/Ks values were mapped to 40 metabolic pathways in the KEGG database. The datasets including the branch leading to lineage-specific duplicates and the datasets excluding them were correlated at  $r = .93$  for mouse and  $r = .99$  for human

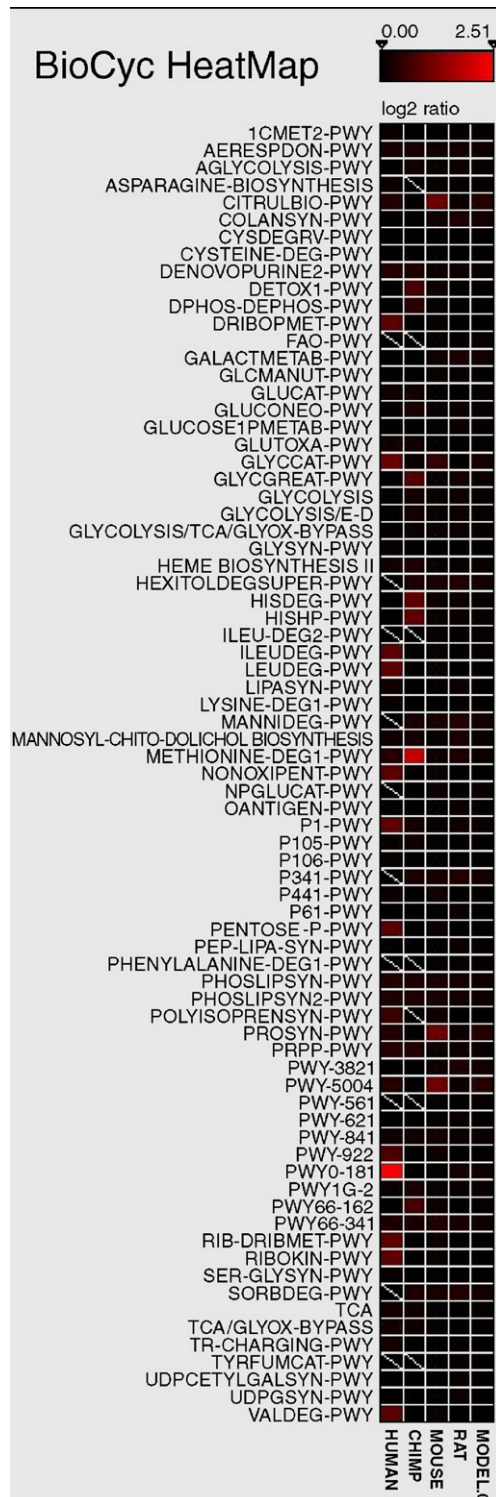


Fig. 6. A heatmap shows the relative values of the average Ka/Ks for each pathway as derived from the BioCyc database (Karp et al., 2005), where free-ratio values were significant. The average one ratio values in each pathway are also shown to provide a baseline for the families where that model was supported (Model.0). The annotation legend for the pathways is shown below the graph. The figure was generated using MayDay (Dietzsch et al., 2006) and '/' indicates boxes for which there was no data.

BioCyc pathway ID	BioCyc pathway name
1CMET2-PWY	Formyl THF biosynthesis I
AERESPON-PWY	Aerobic respiration–electron donors reaction list
AGLYCOLYSIS-PWY	Glycolysis III
ASPARAGINE-BIOSYNTHESIS	Asparagine biosynthesis I
CITRULBIO-PWY	Citrulline biosynthesis
COLANSYN-PWY	Colanic acid building blocks biosynthesis
CYSDEG RV-PWY	L-cysteine degradation VI
CYSTEINE-DEG-PWY	L-cysteine degradation I
DENOVPURINE2-PWY	Purine nucleotides <i>de novo</i> biosynthesis I
DETOX1-PWY	Removal of superoxide radicals
DPHOS-DEPHOS-PWY	NAD phosphorylation and dephosphorylation
DRIBOPMET-PWY	(deoxy)ribose phosphate degradation
FAO-PWY	Fatty acid oxidation pathway I
GALACTMETAB-PWY	Galactose degradation I
GLCMANUT-PWY	<i>N</i> -acetylglucosamine, <i>N</i> -acetylmannosamine and <i>N</i> -acetylneuraminic acid
GLUCAT-PWY	Dissimilation
GLUCONEO-PWY	Glutamate degradation IV
GLUCOSE1PMETAB-PWY	Glucuronogenesis
GLUTOXA-PWY	Glucose and glucose-1-phosphate degradation
GLYCCAT-PWY	Glutamate degradation VII
GLYCGREAT-PWY	Glycine degradation I
GLYCOLYSIS	Glycine degradation II
GLYCOLYSIS/E-D	Glycolysis I
GLYCOLYSIS/TCA/GLYOX-BYPASS	Superpathway of glycolysis+ Entner–Doudoroff
GLYSYN-PWY	Superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass
HEME BIOSYNTHESIS II	Glycine biosynthesis I
HEXITOLDEGSUPER-PWY	Heme biosynthesis II
HISDEG-PWY	Hexitol degradation superpathway
HISHP-PWY	Histidine degradation I
ILEU-DEG2-PWY	Histidine degradation VI
ILEUDEG-PWY	Isoleucine degradation III
LEUDEG-PWY	Isoleucine degradation I
LIPASYN-PWY	Leucine degradation I
LYSINE-DEG1-PWY	Phospholipases
MANNIDEG-PWY	Lysine degradation II
MANNOSYL-CHITO-DOLICHOL BIOSYNTHESIS	Mannitol degradation I
METHIONINE-DEG1-PWY	Dolichyl-diphosphooligosaccharide biosynthesis
NONOXIPENT-PWY	Methionine degradation I
NPGLUCAT-PWY	Non-oxidative branch of the pentose phosphate pathway
OANTIGEN-PWY	Entner–Doudoroff pathway II (non-phosphorylative)
P1-PWY	O-antigen biosynthesis
P105-PWY	Salvage pathways of purine and pyrimidine nucleotides
P106-PWY	TCA cycle variation VIII
	Serine-isocitrate lyase pathway

P341-PWY	Glycolysis II
P441-PWY	<i>N</i> -acetyl neuraminat degradation
P61-PWY	UDP-glucose conversion
PENTOSE-P-PWY	Superpathway of oxidative and non-oxidative branches of pentose phosphate pathway
PEP-LIPA-SYN-PWY	Peptidoglycan and lipid A precursor biosynthesis
PHENYLALANINE-DEG1-PWY	Phenylalanine degradation I
PHOSLIPSYN-PWY	Phospholipid biosynthesis I
PHOSLIPSYN2-PWY	Phospholipid biosynthesis II
POLYISOPRENSYN-PWY	Polysisoprenoid biosynthesis
PROSYN-PWY	Proline biosynthesis I
PRPP-PWY	Superpathway of histidine, purine and pyrimidine biosynthesis
PWY-3821	Galactose degradation III
PWY-5004	Superpathway of citrulline metabolism
PWY-561	Glyoxylate cycle II
PWY-621	Sucrose degradation III
PWY-841	Purine nucleotides <i>de novo</i> biosynthesis II
PWY-922	Mevalonate pathway
PWY0-181	Salvage pathways of pyrimidine deoxyribonucleotides
PWY1G-2	Superpathway of glycolysis and TCA variant VIII
PWY66-162	Oxidative ethanol degradation III
PWY66-341	Cholesterol biosynthesis
RIB-DRIBMET-PWY	Ribose and deoxyribose phosphate metabolism
RIBOKIN-PWY	Ribose degradation
SER-GLYSYN-PWY	Serine and glycine biosynthesis
SORBDEG-PWY	Sorbitol degradation
TCA	TCA cycle
TCA/GLYOX-BYPASS	Superpathway of glyoxylate bypass+TCA
TR-CHARGING-PWY	tRNA charging pathway
TYRFUMCAT-PWY	Tyrosine degradation
UDPNACETYLGALSYN-PWY	UDP- <i>N</i> -acetylgalactosamine biosynthesis
UDPNAGSYN-PWY	UDP- <i>N</i> -acetyl-D-glucosamine biosynthesis
VALDEG-PWY	Valine degradation I

and for rat. The datasets including the in-paralog families will be used for the remainder of the discussion.

In comparing the close species relationships, rat and mouse had a correlation coefficient of 0.54 as seen in Fig. 1, while human and chimpanzee had a correlation coefficient of just 0.06 as seen in Fig. 2. Interestingly, human had higher correlation coefficients with the rodents, with  $r=0.21$  for mouse (Fig. 3) and  $r=0.15$  (Fig. 4) for rat. A heatmap showing the relative rates of evolution of the different pathways in the different lineages, as compared with the single rate value for the families where the free-ratio model was not supported is shown in Fig. 5.

Because there may be a subjective nature to the pathway boundaries in the KEGG database, the pathway definitions were perturbed by adding or removing 1–2 enzymes at the boundaries of each pathway. No rapidly evolving pathway was found to evolve significantly differently using this perturbation. Along the human lineage, only the galactose metabolism showed a significant perturbatory effect, where perturbations relaxed the negative selective pressure (0.15) significantly. Perturbations of oxidative phosphorylation involving deletions showed a change

in average Ka/Ks for chimpanzee, mouse, and rat, but not for human.

### 3.4. Mapping Ka/Ks to BioCyc and comparison with KEGG

A similar analysis to that performed for the KEGG database (Kanehisa et al., 2004) was also performed with pathways derived from the BioCyc database (Karp et al., 2005). A heatmap showing the data from the BioCyc database is found in Fig. 6. In comparing the data generated by BioCyc with that generated from KEGG (Fig. 7), a global correlation coefficient of 0.29 was obtained. The correlation was high for all comparisons except for those in the mouse lineage. The reason for the aberrant behavior of the mouse dataset was unclear. The mapping between KEGG and BioCyc, in general was not one-to-one, as most KEGG pathways mapped to several pathways in BioCyc. This is because the pathways in BioCyc are defined more narrowly.

## 4. Discussion

Many major signaling and metabolic pathways were not found to have gene families that supported the free-ratio model across lineages. In analyzing KEGG, the highest average Ka/Ks for all four species was observed for oxidative phosphorylation and cytokine–cytokine receptor pathways. Chimpanzee had one of the highest average Ka/Ks values in the tight junction pathway. An average Ka/Ks of 1.09 (potentially indicating pathway-level positive selection) was observed for arachidonic acid metabolism in the human lineage (although the average was not statistically significantly above 1). Another rapidly evolving pathway in the human lineage is that of pyrimidine metabolism. These are candidate pathways for functional adaptation, as indicated in Fig. 5.

The pyrimidine metabolism pathway in humans was also found to be rapidly evolving in the BioCyc analysis. Several other pathways were found to be rapidly evolving in the human lineage, including ribose and deoxyribose phosphate

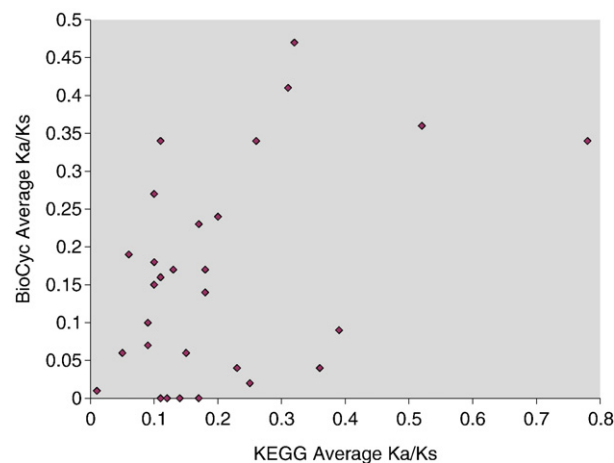


Fig. 7. The correlation between the average pathway Ka/Ks values for each species as measured from the KEGG (Kanehisa et al., 2004) and BioCyc (Karp et al., 2005) databases is shown. The overall correlation coefficient for this comparison is 0.29.

metabolism, the pentose phosphate pathway, several aliphatic amino acid degradation pathways, and mevalonate pathway. Along the chimpanzee lineage, methionine and histidine degradation pathways were identified as rapidly evolving.

Other studies examining lineage-specific gene family evolution at the sequence level have found that the functional distribution of rapidly evolving genes is nonrandom (Liberles et al., 2001; Seoighe et al., 2003; Roth and Liberles, 2006). One study characterizing the differences after speciation events and duplication events found significant differences between the two, with interestingly, vertebrate genes with a metabolic/synthetic role showing an excess of rapid evolution following speciation events as compared with gene duplication events (Seoighe et al., 2003).

Hierarchical clustering in Fig. 5 did recover the species tree, indicating support for more similar pathway-level evolution between human and chimpanzee versus mouse and rat, although the low correlation level between human and chimpanzee ( $r=0.06$ ) is curious. At the pathway level, signaling pathways and metabolic pathways did not segregate in general and there was no obvious link between the clustering and the relationships of the pathways. This may indicate the opportunism of evolution in driving lineage-specific evolution through small numbers of changes in a punctuated lineage-specific manner. This mode of evolution is consistent with the expectations from ancestral sequence reconstruction of key metabolic proteins and the mode in which function has changed (Zhu et al., 2005). Further, it has been suggested that different enzymes in the same pathway may be subject to different functional constraints depending upon their roles in pathway function and this may also influence the dataset observed above (Salvador and Savageau, 2006). Future directions include the examination of co-evolution of interacting proteins and the examination of changes in metabolic flux through pathways evolving in a lineage-specific manner through the evolution of specific proteins.

## Acknowledgments

We would like to thank Inge Jonassen and Tim Hughes for helpful discussions, Parallab for providing a computational environment and support, and FUGE, the Norwegian functional genomics research platform for funding this work.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2006.08.013.

## References

- Benner, S.A., Trabesinger, N., Schreiber, D., 1998. Post genomic science: converting primary structure into physiological function. *Adv. Enzyme Res.* 38, 155–180.
- Benner, S.A., Chamberlin, S.G., Liberles, D.A., Govindarajan, S., Knecht, L., 2000. Functional inferences from reconstructed evolutionary biology involving rectified databases — an evolutionarily grounded approach to functional genomics. *Res. Microbiol.* 151, 97–106.
- Birney, E., et al., 2006. Ensembl 2006. *Nucleic Acids Res.* 34, D556–D561.
- Clark, A.G., et al., 2003. Inferring nonneutral evolution from human–chimpanzee orthologous gene trios. *Science* 301, 1960–1963.
- Crawford, I.P., 1989. Evolution of a biosynthetic pathway. *Ann. Rev. Microbiol.* 43, 567–600.
- Dandekar, T., Schuster, S., Snel, B., Huynen, M., Bork, P., 1999. Pathway alignment: application of comparative analysis of glycolytic enzymes. *Biochem. J.* 343, 115–124.
- Dietzsch, J., Gehlenborg, N., Nieselt, K., 2006. MayDay — a microarray data analysis workbench. *Bioinformatics* 22, 1010–1012.
- Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797.
- Endo, T., Ikeo, K., Gojobori, T., 1996. A large scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* 13, 685–690.
- Forst, C.V., Schulten, K., 1999. Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information. *J. Comput. Biol.* 6, 343–360.
- Forst, C.V., Schulten, K., 2001. Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.* 52, 471–489.
- Gonnet, G.H., Hallett, M.T., Korostensky, C., Bernardin, L., 2000. Darwin v. 2.0: an interpreted computer language for the biosciences. *Bioinformatics* 16, 101–103.
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L., 2000. The large scale organization of metabolic networks. *Nature* 407, 651–654.
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M., 2004. The KEGG resources for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280.
- Karp, P.D., et al., 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* 19, 6083–6089.
- Liberles, D.A., 2005. Datasets for evolutionary comparative genomics. *Genome Biol.* 6, 117.
- Liberles, D.A., Schreiber, D.R., Govindarajan, S., Chamberlin, S.G., Benner, S.A., 2001. The Adaptive Evolution Database (TAED). *Genome Biol.* 2 (8), 0028.1–0028.6.
- Qi, Y., Lin, F., 2005. Parallelisation of the BLAST algorithm. *Mol. Cell. Biol. Lett.* 10, 281–285.
- Roth, C., Liberles, D.A., 2006. A systematic search for positive selection in higher plants (embryophytes). *BMC Plant Biol.* 6, 12.
- Roth, C., Betts, M.J., Steffansson, P., Sælensminde, G., Liberles, D.A., 2005. The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. *Nucleic Acids Res.* 33, D495–D497.
- Roth, C., et al., in press. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J. Exp. Zool.*
- Salvador, A., Savageau, M.A., 2006. Evolution of enzymes in a series is driven by dissimilar functional demands. *Proc. Natl. Acad. Sci. U. S. A.* 103, 2226–2231.
- Seoighe, C., Johnston, C.R., Shields, D.C., 2003. Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. *Mol. Biol. Evol.* 20, 484–490.
- Teichmann, S.A., Babu, M.M., 2004. Gene regulatory network growth by duplication. *Nat. Gen.* 36, 492–496.
- Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556.
- Yang, Z., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15, 568–573.
- Yang, Z., Bielawski, J.P., 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503.
- Zhu, G., Golding, G.B., Dean, A.M., 2005. The selective cause of an ancient adaptation. *Science* 307, 1279–1282.
- Zmasek, C.M., Eddy, S.R., 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17, 821–828.