Evolutionary origins of the endocannabinoid system

John M. McPartland a,⁎, Isabel Matias b, Vincenzo Di Marzo b, Michelle Glass c

a GW Pharmaceuticals, 53 Washington Street Ext., Middlebury, VT, 05753, USA
b Endocannabinoid Research Group, Institute of Biomolecular Chemistry, Consiglio Nazionale delle Ricerche, Via Campi Flegrei 34, 80074 Pozzuoli (Napoli), Italy
c Department of Pharmacology, University of Auckland, Private Bag 92019, New Zealand

Received 30 September 2005; received in revised form 4 November 2005; accepted 9 November 2005
Available online 23 January 2006

Abstract

Endocannabinoid system evolution was estimated by searching for functional orthologs in the genomes of twelve phylogenetically diverse organisms: Homo sapiens, Mus musculus, Takifugu rubripes, Ciona intestinalis, Caenorhabditis elegans, Drosophila melanogaster, Saccharomyces cerevisiae, Arabidopsis thaliana, Plasmodium falciparum, Tetrahymena thermophila, Archaeoglobus fulgidus, and Mycobacterium tuberculosis. Sequences similar to human endocannabinoid exon sequences were derived from filtered BLAST searches, and subjected to phylogenetic testing with ClustalX and tree building programs. Monophyletic clades that agreed with broader phylogenetic evidence (i.e., gene trees displaying topographical congruence with species trees) were considered orthologs. The capacity of orthologs to function as endocannabinoid proteins was predicted with pattern profilers (Pfam, Prosite, TMHMM, and pSORT), and by examining queried sequences for amino acid motifs known to serve critical roles in endocannabinoid protein function (obtained from a database of site-directed mutagenesis studies). This novel transfer of functional information onto gene trees enabled us to better predict the functional origins of the endocannabinoid system. Within this limited number of twelve organisms, the endocannabinoid genes exhibited heterogeneous evolutionary trajectories, with functional orthologs limited to mammals (TRPV1 and GPR55), or vertebrates (CB2 and DAGLβ), or chordates (MAGL and COX2), or animals (DAGLα and CB1-like receptors), or opisthokonta (animals and fungi, NAPE-PLD), or eukaryotes (FAAH). Our methods identified fewer orthologs than did automated annotation systems, such as HomoloGene. Phylogenetic profiles, nonorthologous gene displacement, functional convergence, and coevolution are discussed.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Anandamide (AEA); 2-arachidonyl glycerol (2-AG); N-acyl-phosphatidylethanolamine (NAPE); NAPE-selective phospholipase D enzyme (NAPE-PLD); Diacylglycerol lipase α (DAGLα); Diacylglycerol lipase β (DAGLβ); Fatty acid amide hydrolase (FAAH); Monoglyceride lipase (MAGL); Cyclooxygenase 2 (COX2); Cannabinoid receptors (CB1 and CB2); Vanilloid receptor (TRPV1); GPR55

Abbreviations: AEA, anandamide; BLAST, Basic Local Alignment Search Tool; CB1, cannabinoid receptor subtype 1; CB2, cannabinoid receptor subtype 2; COX2, cyclooxygenase subtype 2; DAGLα, diacylglycerol lipase subtype α; DAGLβ, diacylglycerol lipase subtype β; FAAH, fatty acid amide hydrolase; FAS, functional assessment score; GPCR, G-protein coupled receptor; MAGL, monoglyceride lipase; NAPE, N-acyl-phosphatidylethanolamine; NAPE-PLD, NAPE-selective phospholipase D enzyme; PCR, polymerase chain reaction; THC, tetrahydrocannabinol; TM, transmembrane; TRPV1, vanilloid receptor; 2-AG, 2-arachidonyl glycerol; genomes examined: Hs, human, Homo sapiens; Mm, mouse, Mus musculus; Tr, puffer fish, Takifugu rubripes; Ci, sea squirt, Ciona intestinalis; Ce, nematode, Caenorhabditis elegans; Dm, fruit fly, Drosophila melanogaster; Sc, brewer’s yeast, Saccharomyces cerevisiae; At, thule cress, Arabidopsis thaliana; Pf, malaria apicomplexan, Plasmodium falciparum; Tr, ciliate, Tetrahymena thermophila; Af, archaen, Archaeoglobus fulgidus; Mt, bacterium, Mycobacterium tuberculosis; A, alanine; C, cysteine; D, aspartate; E, glutamate; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

⁎ Corresponding author. Tel./fax: +1 802 388 8304.
E-mail address: mcpruitt@verizon.net (J.M. McPartland).

0378-1199/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
1. Introduction

Ten genes that encode proteins involved in endocannabinoid signalling have been identified to date. More awaits discovery. Half a dozen endocannabinoid ligands are recognized, but research has elucidated the proteins that metabolize only anandamide (AEA) and 2-arachidonyl glycerol (2-AG). AEA is biosynthesized from N-acyl-phosphatidylethanolamine (NAPE) by a NAPE-selective phospholipase D enzyme (NAPE-PLD) (Okamoto et al., 2004). 2-AG is biosynthesized by two diacylglycerol lipases, DAGLα and DAGLβ (Bisogno et al., 2003). Fatty acid amide hydrolase (FAAH) primarily catabolizes AEA (Deutsch et al., 2002). 2-AG is catabolized by monoglyceride lipase (MAGL, monoacylglycerol lipase, Dinh et al., 2002) and by cyclooxygenase 2 (COX2, prostaglandin-endoperoxide synthase, Kozak et al., 2003). AEA and 2-AG act as agonists at cannabinoid receptors 1 and 2 (CB1 and CB2), a pair of G-protein-coupled receptors (GCPRs) named after their exogenous ligand, Δ9-tetrahydrocannabinol (THC) (Mechoulam et al., 1998). AEA also gates the vanilloid receptor (TRPV1, transient receptor potential channel vanilloid receptor 1, Zygmun et al., 1999) and GPR55, an orphan GPCR (Brown et al., 2003).

Phylogenetic histories of the four receptors and six enzymes have not been explored very deeply. We use CB1 as an example: the GenBank website (National Center for Biotechnology Information, www.ncbi.nlm.nih.gov) lists CB1 orthologs in 65 mammalian species, but only four non-mammalian vertebrates, and one invertebrate. Elphick and colleagues used phylogenetic tree analysis to identify orthologs of cannabinoid receptors in Takifugu (Fugu) rubripes (Elphick, 2002) and the sea squirt Ciona intestinalis (Elphick et al., 2003). They performed BLAST searches and constructed gene trees with CB1 and CB2 sequences from other species, based on ClustalX alignments and neighbor-joining methods. If gene tree topology was congruent with tree topology, the sequences were deemed orthologous. This approach has widespread appeal, although it does not represent a comprehensive phylogenetic analysis, due to its automated nature and lack of functional analysis (Brinkman and Leipe, 2001). McPartland (2004) explored endocannabinoid gene phylogeny by screening the genomes of twelve phylogenetically diverse organisms (Fig. 1). Unfortunately not all major clades were sampled in this study, because representative organisms await genome sequencing, such as reptiles, lophotrochozoans, cnidarians, and poriferans. Some clades contained one available whole-genome sequence: primates (human, Homo sapiens, Hs), tunicates (sea squirt, Ciona intestinalis, Ci), apicomplexans (Plasmodium falciparum, Pf), and ciliates (Tetrahymena thermophila, Ti). Other clades offered several choices, from which we chose the best-mutagenesis studies) were mapped upon a ClustalX alignment, and the presence or absence (conservation or mutation) of each motif was visually examined.

The purpose of this study is to combine phylogenetic tree analysis (e.g., Elphick et al., 2003) with phylogenomic comparisons (e.g., McPartland, 2004). Both of these methods, however, utilize similarity-based algorithms. Sequence similarity may not overlap with orthology, and certainly does not equate with functional conservation (Zmasek and Eddy, 2002). Thus we mapped functional analysis onto phylogenomically assembled gene trees, to better estimate the functional origins of the endocannabinoid system.

2. Methods

2.1. Genome sampling

To access deep-level phylogenetic signal, we screened the genomes of twelve phylogenetically diverse organisms (Fig. 1). Unfortunately not all major clades were sampled in this study, because representative organisms await genome sequencing, such as reptiles, lophotrochozoans, cnidarians, and poriferans. Some clades contained one available whole-genome sequence: primates (human, Homo sapiens, Hs), tunicates (sea squirt, Ciona intestinalis, Ci), apicomplexans (Plasmodium falciparum, Pf), and ciliates (Tetrahymena thermophila, Ti). Other clades offered several choices, from which we chose the best

---

Fig. 1. Species tree. A phylogenetic “Tree of Life,” based on broad phylogenetic evidence (reviewed by Benton and Ayala, 2003), including the twelve phylogenetically diverse species used in this study (in bold font) and species discussed in the text. Mammals are marked by a bar labeled “M.” Animals are divided into vertebrates (V) and invertebrates (I), and also divided into four physiological groups: deuterostomes (D), lophotrochozoans (L), ecdysozoans (E), cnidarians (C), and poriferans (P). The three domains (supraphyla) are eukaryotes, archaens (A), and bacteria (B).
characterized model organism per branch: for rodents, we chose mouse *Mus musculus* (*Mm*); for fish, puffer fish *Takifugu rubripes* (*Tr*); for insects, fruit fly *Drosophila melanogaster* (*Dm*); for nematodes, *Caenorhabditis elegans*, (*Ce*); for fungi, the brewer's yeast *Saccharomyces cerevisiae* (*Sc*); for plants, thale cress *Arabidopsis thaliana*, (*At*); for prokaryotes, the archaen *Archaeoglobus fulgidus* (*Af*); and the bacterium *Mycobacterium tuberculosis* (*Mt*). The chimp genome has been partially sequenced, but the current draft lacks four endocannabinoid genes (data not shown). Genome databases of *Hs*, *Mm*, *Tr*, *Ci*, *Dm*, *Ce*, *Sc*, *At*, *Pf*, *Af*, and *Mt* were obtained from GenBank (www.ncbi.nlm.nih.gov), with *Tr* and *Ci* cross-referenced with Ensembl (www.ensembl.org/Multi/blastview) and the Joint Genome Institute (JGI, http://genome.jgi-psf.org/cgi-bin/runBlast). The *Ti* genome came from TIGR (http://www.tigr.org/tdb/e2k1/ttg).

### 2.2. Similarity screening and tree building

The 12 genomes were searched with gapped BLAST (blastP and tblastn, Altschul et al., 1997) using ten query sequences: *Hs*BC1 (GenBank:NP_057167), *Hs*BC2 (NP_001832), *Hs*TRPV1 (NP_542437), *Hs*GPR55 (NP_005674), *Hs*FAAH (NP_001432), *Hs*MAGL (NP_009214 not NP_001003794), *Hs*COX2 (NP_009954), *Hs*NAPE-PLD (NP_945341), *Hs*DAGLα (NP_006124 not BAA31634), and *Hs*DAGLβ (NP_61918). Reciprocal best hits were filtered by means of a threshold *E* value (<0.01) and sequence length (75% of query over subject). Paired hits with nearly equal *E* values were analyzed as paralogs and provided useful information for gene trees (e.g., the paralogs CB1 and CB2). Best hits from each genome were aligned with ClustalX (www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html) using default parameters. Alignments were manual edited to minimize indel events before generating tree sequences with a neighbor-joining (NJ) algorithm. We detected several polyphyletic sequences by placing two outgroups per tree: the *Hs* sequence with closest similarity to the query sequence found in the *Hs* genome, and a more distantly related *Hs* outgroup to root the tree. Outgroups characterized by in vitro functional studies were included whenever possible. TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) recursively optimized the alignments and generated graphic outputs, with branch lengths proportional to distances between sequences. Confidence values for NJ trees were generated by bootstrapping, based on 1000 resampling replicates. Bootstrap support for clade stability was given at nodes; bootstrap values ≥ 500 supported monophyly of the clade (Brinkman and Leipe, 2001).

### 2.3. Functional assessment

Whereas BLAST evaluated sequence similarity, a number of algorithms evaluated the patterns and motifs within sequences, and classified the sequences into function-based protein families. We used two protein prediction programs. Pfam (www.sanger.ac.uk/Software/Pfam) implemented a hidden Markov model (HMM) upon gapped multiple sequence alignments. Prosite (http://au.expasy.org/prosite) used a position-specific scoring matrix (PSSM) upon ungapped multiple sequence alignments. Pfam and Prosite are primary profilers and combined automated and human curation, unlike other databases (e.g., CDD, COG, SMART). We also queried sequences with two algorithms that predicted secondary structures. TMHMM (www.cbs.dtu.dk/services/TMHMM) was chosen over Tmpred, SVMtm, and PSORT after it outperformed the other transmembrane region prediction programs in an accuracy trial run upon the ten human endocannabinoid proteins (data not shown). PSORT (http://psort.nibb.ac.jp) was chosen as a subcellular localization predictor, after it outperformed ESLpred, LOC-target, PAIence, SubLoc, and TargetP (data not known). When the prediction for a potential ortholog matched the prediction for an endocannabinoid protein, the potential ortholog scored 1. If the algorithm predicted a different structure or function, the potential ortholog scored 0. The positions of critical AA motifs that served vital roles in endocannabinoid protein function (e.g., ligand-binding and catalysis) were mapped upon ClustalX alignments, and the presence or absence of each AA motif was visually examined, and scored with either 1 (indicating presence of a motif) or 0 (indicating substitution or deletion of a motif).

### Table 1

<table>
<thead>
<tr>
<th>Protein</th>
<th>Amino acid residue motifs</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB1 + CB2</td>
<td>F3.25, K3.28, V3.32, W5.43, L5.50, C175 in EL-2, L6.33-A6.34</td>
<td>McPartland and Glass, 2003&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>CB2-specific motifs</td>
<td>S3.31, T3.35, S4.53A, S4.57A, F5.46V, LDV not MDI in IC-3, C313M, C2.59Y</td>
<td>McPartland and Glass, 2003&lt;sup&gt;a&lt;/sup&gt;; Zhang et al., 2005&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>TRPV1</td>
<td>R114, R491, Y511, S512, T550, E761, C-terminus motif</td>
<td>McPartland, 2004&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>GPR55</td>
<td>K2.60, FV3.28, (pxpx) (6.43-6), K7.36</td>
<td>Unpublished data</td>
</tr>
<tr>
<td>FAAH</td>
<td>P129, PPLP (310-3) part of SH3, I491</td>
<td>Matias et al., 2005&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>MAGL</td>
<td>GxSxG(120-4)-D239-H269 C242</td>
<td>McPartland, 2004&lt;sup&gt;b&lt;/sup&gt;; Saario et al., 2005&lt;sup&gt;*&lt;/sup&gt;</td>
</tr>
<tr>
<td>COX1 + 2 COX2-specific motifs</td>
<td>R120, V349, Y355, G526, S530, L531 T383H, R513H, V523L, L505F</td>
<td>Matias et al., 2005&lt;sup&gt;b&lt;/sup&gt;; Schneider et al., 2004&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>NAPE-PLD</td>
<td>D147, HxHxDH(185-90), H253, D284, H331</td>
<td>Bambai et al., 2004&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>DAGL&lt;sub&gt;α&lt;/sub&gt;</td>
<td>GxSxG(441-5), D495, H429</td>
<td>Okamoto et al., 2004&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> Motifs evaluated in point-mutation studies.

<sup>b</sup> Review article that cites primary literature.

<sup>c</sup> Motifs evaluated in computer-modeling studies.
Conservative AA substitutions were allowed, based on the BLOSUM62 substitution matrix, using a cut-off value of $+1$. The list of critical AA motifs is presented in Table 1. Each queried sequence was given a functional assessment score (FAS), a sum of the four profilers (Pfam, Prosite, PSORT, and TMHMM) and a variable number of functional mapping motifs (between three and eight, see Table 1). FAS scores finalized the transition from sequence similarity (a quantitative measure) to phylogenetic homology (a statement of common ancestral origins) to functional equivalence (a qualitative characterization of protein utility).

3. Results

All twelve genomes expressed one or more sequences that shared similarities with the queried human endocannabinoid sequences. Databases for the $Tr$ and $Ci$ genomes did not cross-reference with consistency, as noted below. Pfam and Prosite tended to classify sequences too broadly, in large functional families, and not in specific homologous series. For example, all the sequences BLASTed from queries $CB1$, $CB2$, and GPR55 fell into the same profile, “GPCR.” This lack of specificity inflated FAS scores. Sequences that claded together in phylogenetic trees shared similar FAS scores.

3.1. $CB1$ and $CB2$

BLAST found no sequences that met threshold in the genomes of $Sc$, $At$, $Pf$, $Tt$, $Af$, or $Mt$. In the combined $CB1$–$CB2$ gene tree (Fig. 2), $HsCB1$ claded with $MmCB1$ and a pair of $TrCB1$ paralogs, all with high FAS scores. $HsCB2$ claded with $MmCB2$ and $TrCB2$, all with high CB2-specific FAS scores. The $Ci$ and $Ce$ orthologs sistered basal to the $CB1$ and $CB2$ clades. Together, this monophyletic clade separated with good bootstrap support from outgroups $HsEDG1$ and $HsEDG2$, the sequences with greatest similarity to $CB1$ and $CB2$ in the $Hs$ genome ($E=2e-28$ and $E=3e-29$, respectively, with $HsCB1$). The $Dm$ sequence passed the BLAST threshold ($E=3.9e-39$), but lacked synapomorphy with cannabinoid receptors, it placed between the EDG clade and the distal outgroup $Hs$ADRA1A (an alpha adrenergic GPCR).

3.2. TRPV1

BLAST found no sequences that met threshold in the genomes of $Sc$, $At$, $Pf$, $Tt$, $Af$, or $Mt$. Paralogs from $Tr$ did not cross-reference consistently between databases: the best hit ($E=7.9e-282$) was identified as SINFRUP0000155081 by Ensembl but SINFRUP00000066444 by JGI. Its nearly equal paralog ($E=1.1e-262$) was identified as SINFRUP00000162427 by Ensembl but FRUP00000162428 by JGI (with indels), and SINFRUP00000085541 by GenBank. In the phylogenetic tree (Fig. 3), $Tr$162427 sistered with $Hs$TRPV1 and $Mm$TRPV1 (all with high FAS scores), whereas $Tr$155081 (with a low FAS score) claded with outgroup $Hs$TRPV4, the sequence with greatest similarity to TRPV1 in the $Hs$ genome. The $Ci$, $Ce$, and $Dm$ sequence passed the BLAST threshold but lacked synapomorphy with TRPV1, they placed between the TRPV4 clade and the distal outgroup $Hs$TRPA1, with low FAS scores.

3.3. GPR55

BLAST found no sequences that met threshold in the genomes of $Sc$, $At$, $Pf$, $Tt$, $Af$, or $Mt$. Functional mapping of GPR55 posed a problem, no point mutation studies have been done. GPR55 shares ligands with $CB1$ but little sequence similarity ($E>1$ with $HsCB1$); according to Brown et al. (2005), GPR55’s affinity for AEA represents a case of convergent evolution. Within the $Hs$ genome, the GPR55 sequence most closely resembles that of GPR23 ($E=3e-35$). The ligand of GPR23 is lysophosphatidic acid (LPA), which GPR23 shares with EDG2 (Noguchi et al., 2003). EDG2 shares little sequence similarity with GPR23 ($E=1$), but does resemble $CB1$ ($E=3e-29$). Arachidonoyl-(20:4)-LPA shares structural characteristics with AEA (although LPA is charged), so the ligand binding sites of GPR55 and GPR23 may overlap. We aligned GPR55, GPR23, $CB1$, and EDG2, and examined the point-mutation studies done on EDG receptors (Fujinawa et al., 2005). Several residues aligned as potential ligand-binding sites for GPR55 (Table 1). In the gene tree (Fig. 4), $Hs$GPR55 and $Mm$GPR55 claded together, with perfect FAS scores. $Tr$141960 claded with outgroup $Hs$GPR23. The $Ci$, $Ce$, and $Dm$ sequences placed between the GPR23 clade and the distal outgroup $HsCB1$, with lower FAS scores.
3.4. FAAH

BLAST identified sequences with similarity to *Hs*FAAH in all genomes. Tr sequences did not cross-reference between databases: the best hit, Ensembl SINFRUP00000146100, deleted 31 residues from its c-terminal that were present in JGI FRUP000000146100 and GenBank FRUP00000055162, but the JGI and GenBank sequences exhibited different, inconsistent deletions. The FAAH tree (Fig. 5) bifurcated into two major clades: *Hs*FAAH claded with sequences from *Mm*, *Tr*, *Ci*, *Ce*, *Sc*, and *Tt*, with variable FAS scores. *Hs*Amidase, the sequence with greatest similarity to FAAH in the *Hs* genome, claded with sequences from *Tr* and *Dm*, with low FAS scores.

3.5. MAGL

We tested *Hs*MAGL isoform A (NP_009214) and not isoform B (NP_001003794), although the latter produced identical lists of putative orthologs (data not shown). BLAST identified sequences with similarity to *Hs*MAGL in all genomes. The MAGL tree (Fig. 6) bifurcated into two major clades: *Hs*MAGL claded with sequences from *Mm*, *Tr*, and *Ci*, with good FAS scores. A second clade bearing no congruence with the species tree (Fig. 1) contained sequences from *Tt*, *Pf*, *At*, *Sc*, and *Dm*, sister to the prokaryote sequences *Af* and *Mb*, with variable FAS scores.

3.6. COX2

BLAST found no sequences that met threshold in the genomes of *Sc*, *Pf*, *Tr*, *Af*, or *Mt*. Four sequences with nearly equal E scores were BLASTed from the *Ci* genome. The best hit, Ensembl ENSCINP00000012732, was identical to JGI ci0100149048 and nearly identical to the second hit (ENSCINP00000012733) and the fourth hit (ENSCINP00000012734), except for inconsistent indels in the latter two sequences. The third hit, ENSCINP00000013352, was identical to JGI ci0100139817, except for a 24 residue n-terminal insertion. In the COX2 tree (Fig. 7), *Hs*COX2 claded with sequences from *Mm* and *Tr*, with high FAS scores. This clade separated with good bootstrap support from outgroup *Hs*COX1, the sequence with greatest similarity to COX2 in the *Hs* genome. Basal to the COX2 and COX1 clades, *Ci* paralogs sistered with good bootstrap and high FAS scores. These were followed by a series of sequences that
showed no congruence with the species tree (Fig. 1), and low FAS scores.

3.7. NAPE-PLD

BLAST found no sequences that met threshold in the Ci, Dm, At, or Tr genomes. In the tree (Fig. 8) HsNAPE-PLD claded with sequences from Mm and Tr, with very high FAS scores.

Below this group, the gene tree did not display topographical congruence with the species tree; a clade of prokaryotes sistered with metazoans. The FAS scores were problematic; Pfam and Prosite had difficulty profiling NAPE-PLD, and functional mapping was based upon motifs identified by computer modeling rather than robust experimental point-mutation studies (Table 1). The sequence with greatest similarity to NAPE-PLD in the Hs genome, the interleukin 20 receptor (the outgroup), bears little resemblance to it. This may have created long branch length artifacts in the gene tree.

Fig. 6. Gene tree of COX2 orthologs and outgroups. Sequence names are followed by FAS scores, tallying the presence of COX2 enzyme motifs in the following sequences: HsCOX2 (NP_000954), MmCOX2 (NP_053528), Tr139832 (SINFRUP0000139832), C12732 (ENSCHINP000012732), C113352 (ENSENNP0000113352), CeC46A5.4 (NP_501272), DmCG10211-PA (NP_609883), and outgroup sequences HsCOX1 (NP_000953), HsThyPeroxA (NP_000538) and HsThyPeroxB (NP_783650).

Fig. 7. Gene tree of COX2 orthologs and outgroups. All sequence names are followed by two FAS scores. The first FAS tallies the number of COX motifs in the sequence, out of eleven scored motifs. The second FAS tallies the number of COX2-specific motifs, presented as a ratio of COX2/COX1/neither. Sequences include HsCOX2 (NP_000954), MmCOX2 (NP_053528), Tr139832 (SINFRUP0000139832), C12732 (ENSCHINP000012732), C113352 (ENSENNP0000113352), CeC46A5.4 (NP_501272), DmCG10211-PA (NP_609883), and outgroup sequences HsCOX1 (NP_000953), HsThyPeroxA (NP_000538) and HsThyPeroxB (NP_783650).

Fig. 8. Gene tree of NAPE-PLD orthologs and outgroups. Sequence names are followed by FAS scores, tallying the presence of NAPE-PLD enzyme motifs in the following sequences: HsNAPE-PLD (NP_945341), MmNAPE-PLD (NP_848843), Tr142348 (SINFRUP0000142348), CeY37E11AR.4 (NP_500408), ScFmp30p (NP_015222), TrPF11_0452 (NP_701308), AfAF1265 (NP_070093), MmMT0929 (NP_335362), and outgroup sequence HsIL20RA (NP_055247).

Fig. 9. Gene tree of DAGL orthologs and outgroups. Sequence names are followed by FAS scores, tallying the presence of DAGLα enzyme motifs in the following sequences: HsDAGLα (NP_006124), HsDAGLβ (NP_631918), MmDAGLα (NP_932782), MmDAGLβ (NP_659164), Tr167103 (SINFRUP0000167103), Tr148896 (FRUP0000148896), C146113 (c10100146113), CeF42G9.6a (NP_741084), DmCG33174-PD (NP_788900), ScYjr107wp (NP_012641), At1g05790 (NP_172070), T165.m00064 (8254448), and the outgroup HsMAGL1 (NP_009214).
3.8. DAGLα and DAGLβ

BLASTing with either *Hs*DAGLα or *Hs*DAGLβ found no sequences with significant identity in the *Pf*, *Af*, or *Mf* genomes, and hit upon single sequences in the *Ci*, *Ce*, *Dm*, *Sc*, *At*, and *Tt* genomes. BLASTing the *Tr* Ensembl database with *Hs*DAGLα identified SINFRRUP00000167103, whereas *Hs*DAGLβ scored no hits that met threshold. BLASTing the *Tr* JGI database with *Hs*DAGLα and *Hs*DAGLβ hit on the same sequence, FRUP00000148896. The DAGLα clade (Fig. 9) had poor bootstrap support below the sistered vertebrate sequences. Depending upon the degree of manual editing, the *Dm* and *Ci* sequences claded with DAGLα or DAGLβ, or sistered basal to both clades. The final tree was optimized, with all gaps in the alignment removed. The chordate sequences produced high FAS scores, with lower scores in the other sequences.

4. Discussion

BLAST (threshold \( E < 0.01 \)) sensitively identified sequences that shared similarity to queried human sequences. We confidently predict this method did not commit type 2 errors (false negatives). Conversely, we faced the challenge of sorting homologous sequences from homoplastic sequences (sequences sharing similarity because of convergent evolution, not common descent).

4.1. CB1 and CB2

The *Hs* and *Mm* orthologs identified herein (Fig. 2) have been well-characterized in functional studies. The *Tr* sequences have not been functionally tested and would benefit from the attention. *Tr*CB1A and *Tr*CB1B have been described as lineage-specific expansions of CB1, not CB2 (Elphick, 2002). Yet *Tr*CB1A and *Tr*CB1B expressed CB2-specific motifs (S4.53 and S4.57 in Table 1) whilst *Tr*CB2 expressed three CB1-specific motifs (G3.31, MDI in IC-3, and A[conserved]S3.35). The *Tr*CB2 sequence identified herein was 83 amino acid residues longer than *Tr*CB2 described by Elphick (2002).

The CICBR sequence shared 29% identity with *Hs*CB1 and 24% identity with *Hs*CB2. These divergences are greater than the divergence between *Hs*CB1 and *Hs*CB2 (47% identity), suggesting the ancestor of CICBR evolved prior to the CB1–CB2 duplication event, as previously hypothesized by Elphick et al. (2003). The phylogenetic tree upheld this hypothesis, albeit with weak bootstrap support (Fig. 2). The ancestor of CICBR may have functioned like present-day CB1 rather than CB2, judging from CICBR’s CB1-specific FAS score (3/0/5, Fig. 2). CICBR expressed substitutions at two motifs, F3.36 and W5.43, required for mammalian cannabinoid receptors to bind AEA, CP55,940, WIN55212-2, and SR141716A. Transfected receptors with mutations at these sites lost affinity for WIN55212-2, and SR141716A, but retained affinity for AEA and CP55,940 (McAllister et al., 2003). Similarly, *Ci* tissues demonstrated high-affinity binding with \([^{3}H]CP55,940\) (McPartland et al., accepted for publication), but less specific binding with \([^{3}H]SR141716A\) (Matias et al., 2005). CICBR expressed an aromatic residue located i-4 from W.6.48, indicating it may lack the constitutive activity seen in vertebrate receptors (Singh et al., 2002).

CeC02H7.2 was previously rejected as an ortholog because of its low similarity to *Hs*CB1 (Elphick and Egertova, 2001) and because of it low FAS score (McPartland and Glass, 2001). Nevertheless, CeC02H7.2 claded with cannabinoid receptors, with good bootstrap support. These results show how a combination of BLAST and phylogenetic tree analysis can improve the sensitivity and specificity of ortholog discovery. CeC02H7.2 may be an ortholog, but it might not function as a cannabinoid receptor, its FAS score indicated substitutions at half the motifs required for mammalian CB1 to function. Nevertheless, a radioligand study of nematode neural tissues demonstrated high-affinity binding of \([^{3}H]CP55,940\) (McPartland et al., accepted for publication). Within the *Ce* genome, CeC02H7.2 closely resembled sequences that coded for olfactory receptors. Perhaps the ancestral cannabinoid receptor diverged from a nematode olfactory receptor. Moderate to high densities of CB1 are retained in the human olfactory cortex (Glass et al., 1997) as well as limbic structures, which share primate connections with the olfactory system.

The proximal outgroups, *Hs*EDG1 and *Hs*EDG2, shared the closest similarity to CB1 and CB2 in the *Hs* genome (\( E = 2e-28 \) and \( E = 3e-29 \), respectively, with *Hs*CB1). These sequences have been selected as outgroups in previous cannabinoid studies (Elphick, 2002; Elphick et al., 2003). The *Dm* sequence shared greater similarity with *Hs*CB1 (\( E = 3.9e-39 \)) than *Hs*EDG1 and *Hs*EDG2, but the CLUSTAL NJ algorithm placed the *Dm* sequence distal to *Hs*EDG1 and *Hs*EDG2 in the phylogenetic tree (Fig. 2). *Dm*CG9753-PA was previously rejected as a cannabinoid receptor (McPartland et al., 2001), despite its annotation as a CB1 ortholog by GenBank. In concurrence, radioligand binding studies of *Dm* found no high-affinity binding of \([^{3}H]CP55,940\) or \([^{3}H]SR141716A\) (McPartland et al., 2001). Radioligand binding studies have shown high-affinity binding in *Hydra vulgaris* (a cnidarian, De Petrocellis et al., 1999) and in earthworm (*Lumbricus terrestris*, a lophotrochozoan, McPartland et al., accepted for publication). Unfortunately, no cnidarian or lophotrochozoan genomes have been sequenced yet. The FAS scores of outgroups EDG1 and EDG2 were unexpectedly high, due in part to the lack of specificity by Pfam and Prosite, and due to a cluster of substituted-but-conserved residues (F3.25W, K3.28R, and V3.32M).

Several web-based, automated annotation systems now post lists of putative orthologs of human genes. These systems conflicted with our results and lacked fidelity with each other, due to their reliance upon incorrect or imprecise annotations present in sequence databases. HomoloGene (www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search and DB=homologene) identified “putative homologs” of CB1 in the genomes of *Hm* and *Mm*, and not in *Ce*, *Dm*, *Sc*, *Pf*, or *At* (HomoloGene does not examine the *Tr*, *Ci*, *Tt*, or *Af* genomes). UniGene (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) an automated system that partitions GenBank sequences into sets of gene-oriented clusters, identified “protein similarities” between
HsCB1 and sequences in Hs and Mm, as well as Ce (NP_508760, a sequence removed from GenBank in January 2005) and Dm (NP_477007, a dopamine receptor). Unigene identified similarities between HsCB2 and sequences in Hs and Mm, as well as Ce (T24659, an opioid receptor) and Dm (S68780, a dopamine receptor).

4.2. TRPV1

HsTRPV1 and MmTRPV1 have been well-characterized in functional studies. The Tr paralogs beg functional analysis; the gene tree (Fig. 3) suggested the Tr paralogs descended from a duplication event that gave rise to the TRPV1 and TRPV4 lineages, reciprocal best hits in the Hs genome. TRPV1 and TRPV4 are polyomidal receptors, they both respond to heat, TRPV1 responds to tissue acidity and TRPV4 responds to osmolarity. TRPV1 is gated by capsaicin and AEA; TRPV4 is gated not by AEA but by its FAAH-mediated metabolite, arachidonic acid, and its cytochrome-450-mediated metabolite 5′,6′-epoxyeicosatrienoic acid (Watanabe et al., 2003). TRPV1′s affinity for AEA may have evolved recently, in rodents and primates. The TRPV1 ortholog in chickens has little functional affinity for AEA or capsaicin (Jordt and Julius, 2002). T550 may be a key residue for ligand affinity, substitutions at this site were expressed in four capsaicin-insensitive sequences: chicken TRPV1 (Jordt and Julius, 2002), the Ce sequence identified here (CeOsm9, Tobin et al., 2002), rabbit TRPV2 (Gavva et al., 2005), and HsTRPV2, HsTRPV4, and frog TRPV1 (data not shown). The Tr paralogs also substituted T550 for other residues.

Polyominal branching in the TRPV1 tree (Fig. 3) suggested the Ci–Ce–Dm clade may have expressed greater dissimilarity than the outgroup chosen to root the tree, HsTRPA1. The latter sequence (also known as ANKTM1, a member of the TRP channel family), has some affinity for THC (Jordt et al., 2004). The highly divergent Ci, Ce, and Dm sequences may have grouped together as a consequence of long branch attractions and are not truly related. HomoloGene agreed with our results; it identified putative homologs of TRPV1 in Hm and Mm, and not in Ce, Dm, Sc, Pf, or At (HomoloGene does not include the Tr, Ci, Ti, Mt, and Af genomes). Unigene did not agree entirely with our results; it identified protein similarities between HsTRPV1 and sequences in Hs and Mm, but also in Ce (NP_500372, another Osm-9 receptor).

4.3. GPR55

Functional mapping based upon our estimation of critical motifs (Table 1) proved effective, although FAS scores were inflated by Pfam, Prosite, and TMHMM classifying all the sequences identically as GPRCs. The GPR55 gene tree (Fig. 4) demonstrated the utility of placing two outgroups per tree to detect polyomylphetic sequences. Removing the HsGPR23 sequence collapsed the Hs, Mm, Tr, and Ci sequences into one spurious clade. Instead, placement of HsGPR23 in the alignment demonstrated that Tr141960 shared similarity with GPR55 (thus its detection by BLAST) but was not orthologous with GPR55. HomoloGene identified putative homologs of GPR55 in Hm and Mm, and not in Ce, Dm, Sc, Pf, or At (HomoloGene does not include the Tr, Ci, Ti, Mt, and Af genomes). Unigene identified protein similarities between HsGPR55 and sequences in Hs and Mm, but also in Ce and Dm.

4.4. FAAH

The FAAH gene tree (Fig. 5) did not display topographical congruence with the species tree (Fig. 1), necessitating some interpretation, aided by FAS scores. We interpreted the Hs-to-Tr clade as FAAH orthologs. This phylogenetic concept departed from Patricelli and Cravatt (2000), who characterised FAAH as a mammalian enzyme. The Hs-to-Tr clade separated with good bootstrap support from the Hs–Tr–Dm clade of “other amidases.” The literature supports this concept. FAAH-like amidohydrolyases have been extracted from Ci (Matias et al., 2005), the leech Hirudo medicinalis (Matias et al., 2001), Tetrarhymena pyriformis, a species closely related to Tr, (Karava et al., 2001), and even Hydra vulgaris (De Petrocellis et al., 1999). Endocannabininoids have been extracted from Sc (Merkel et al., 2005), consistent with the presence of an FAAH ortholog in the Sc genome. On the other hand, Dm lacks detectable AEA (McPartland et al., 2001), so the lack of an FAAH ortholog in this species should not be surprising. Segregating FAAH from “FAAH-like amidases” may be pedantic, because the At sequence in Fig. 5 has been demonstrated to metabolize AEA with kinetics equal to FAAH (Shrestha et al., 2003). HomoloGene did not agree with our results; it identified putative homologs of FAAH in Hm, Mm, Ce and Dm, and not in Sc, Pf, and At. Unigene on the other hand identified Sc and At sequences sharing similarities with HsFAAH, along with Hs, Mm, and Ce, sequences.

4.5. MAGL

The MAGL tree (Fig. 6) resolved into two major clades. We interpreted the Hs, Mm, Tr, and Ci sequences as MAGL orthologs, with good bootstrap support and good FAS scores. The clade that included Tr, Pf, At, Sc, and Dm had poor bootstrap support and no congruence with the species tree (Fig. 1). Interpreting the Tr-to-Dm clade as MAGL orthologs would suggest that MAGL evolved in organisms with ancient lineages. Indeed a sequence with similarity to HsMAGL was previously reported in the Cowpox virus genome (McPartland, 2004). Functional studies are required to resolve this. HomoloGene identified putative homologs of MAGL in Hs, Mm, and At, and not in Ce, Dm, Sc, or Pf (HomoloGene does not include the Tr, Ci, Ti, Mt, and Af genomes). Unigene identified protein similarities between HsMAGL and the Hs, Mm, Sc, and At sequences evaluated herein.

4.6. COX2

The top branches of the COX tree (Fig. 7) agreed with a COX phylogram by Jarving et al. (2004), except Järving and colleagues additionally BLASTed with HsCOX1, so the COX1
clade included sequences from \textit{Mm} and \textit{Tr}; Järving and colleagues also placed \textit{Ci} paralogs in a clade basal to COX2 and COX1, suggesting the ancestor of the \textit{Ci} sequences evolved prior to the COX2–COX1 duplication event. The ancestral COX gene may have functioned like present-day COX2 rather than COX1, judging from FAS scores. \textit{Ci12732} conserved three COX2-specific motifs (R513, V523, L505) and no COX1-specific motifs; \textit{Ci13352} conserved two COX2-specific motifs (R513, V523) and one COX1-specific motif (F505). The other sequences had low FAS scores and placed in spurious clades that contradicted broader phylogenetic evidence (e.g., monophyly of plant and metazoan sequences). Whether these sequence function like COX2 or COX1 can be ascertained in studies similar to Knight et al. (1999), who inferred \textit{Ci} prostaglandin synthesis with COX2-selective drugs (etodolac) but not with COX1-selective drugs (resveratrol). Homologene identified COX2 homologs in \textit{Hm} and \textit{Mm}, and not in \textit{Ce}, \textit{Dm}, \textit{Sc}, \textit{Pf}, or \textit{At} (\textit{Tr}, \textit{Ci}, \textit{Ti}, \textit{Mt}, and \textit{Af} genomes were not examined). Unigene identified protein similarities between \textit{Hs}COX2 and sequences in \textit{Hs}, \textit{Mm}, and \textit{Ce}.

4.7. \textit{NAPE-PLD}

The \textit{NAPE-PLD} gene tree (Fig. 8) did not display topographical congruence with the species tree (Fig. 1). When gene tree topology lacks congruence with the species tree topology, the sequences in the gene tree may not be orthologous (Brinkmann and Leipe, 2001). The incongruent placement of the \textit{Mt–Af} clade can be explained by the fact that \textit{NAPE-PLD} shares no similarity with other phospholipase D enzymes, rather it resembles \(\beta\)-lactamase enzymes in prokaryotes (Okamoto et al., 2004). Sister to the prokaryotes were \textit{Sc} and \textit{Ce} sequences, which we interpreted as functional \textit{NAPE-PLD} orthologs. When the prokaryotes were removed from analysis, \textit{Sc} and \textit{Ce} claded with the vertebrates, with adequate bootstrap support \((\geq 555 \text{ at both nodes})\), and \textit{Sc} and \textit{Ce} reversed their positions on the gene tree, becoming topographical congruent with the species tree (data not shown). \textit{Sc} and \textit{Ce} FAS scores supported our interpretation, as does the literature. The lipid fraction of \textit{Ce} includes phosphatidyethanolamine and arachidonic acid (Tanaka et al., 1996), which are the feedstock phospholipid and fatty acid necessary for \textit{NAPE-PLD} to produce AEA. The presence of arachidonic acid in \textit{Sc} is debated (Zank et al., 2000; Merkel et al., 2005). The \textit{Sc} ortholog identified by our methods was recently examined by Merkel et al. (2005), who described it as a \textit{NAPE} hydrolyzing phospholipase D. Lack of \textit{NAPE-PLD} in \textit{Ci} was previously reported (Matias et al., 2005) but still surprising, because AEA has been extracted from \textit{Ci} tissues (Matias et al., 2005). Sun et al. (2004) proposed a second mechanism for AEA biosynthesis, where \textit{NAPE} is hydrolyzed by a secretory phospholipase \(A_2\) (PLA\(_2\) group Ib) to NA-lysoPE, which is then cleaved by a lysophospholipase D enzyme to yield AEA. This possibility can be explored after the specific mammalian enzymes have been cloned. Similarly, lack of \textit{NAPE-PLD} in \textit{Ti} was surprising, Siafaka-Kapadai et al. (2005) reported AEA and \textit{NAPE-PLD}-like activity in \textit{Tetrahymena pyriformis}. No \textit{NAPE-PLD} in the \textit{At} genome was also unexpected, because an ortholog of \textit{NAPE-PLD} was reported in tobacco plants (Chapman, 2000). Absence of a \textit{NAPE-PLD} ortholog in \textit{Dm} agreed with ligand extraction studies that found no measurable amounts of AEA in \textit{Dm} tissues (McPartland et al., 2001). Homologene identified putative homologs of \textit{NAPE-PLD} in \textit{Hs}, \textit{Mm}, and \textit{Ce}, and not in \textit{Dm}, \textit{Sc}, or \textit{Pf} (Homologene does not include the \textit{Tr}, \textit{Ci}, \textit{Ti}, \textit{Mt}, and \textit{Af} genomes). Unigene identified protein similarities between \textit{Hs}MAGL and sequences in \textit{Mm} and \textit{Ce}.

4.8. \textit{DAGL\(\alpha\)} and \textit{DAGL\(\beta\)}

The combined \textit{DAGL\(\alpha\)} and \textit{DAGL\(\beta\)} tree proved most challenging. We interpreted the \textit{Hs}-to-\textit{Ce} sequences as \textit{DAGL} orthologs, with a duplication event that led to the eventual divergence of \textit{DAGL\(\alpha\)} and \textit{DAGL\(\beta\)} in vertebrates. Poor bootstrap support, despite optimized editing of the alignment, made the timing of the duplication event difficult to interpret. The tree (Fig. 9) suggested a duplication event ancestral to \textit{Dm}, with subsequent loss of \textit{DAGL\(\beta\)} paralogs in \textit{Dm} and \textit{Ci}. This is not a very parsimonious scenario. Alternatively, Matias et al. (2005) presented a \textit{DAGL} tree constructed without editing, and the \textit{Ci} sequence sistered basal to the \textit{DAGL\(\alpha\)} and \textit{DAGL\(\beta\)} clades. This tree better explains the appearance of paralogs in vertebrates. Additional sequences are needed to better resolve the timing of the \textit{DAGL} duplication event. 2-AG, the product of \textit{DAGL\(\alpha\)} and \textit{DAGL\(\beta\)}, has been extracted from the tissues of vertebrates as well as several invertebrates: sea urchin (\textit{Paracentrotus lividus}), sea slug (\textit{Aplysia sp.}), leeches (\textit{Hirudo} and \textit{Theromyzon tessulatum}), mussel (\textit{Mytilus galloprovincialis}), clam (\textit{Tapes dicussatus}), oyster (\textit{Crassosterea sp.}), insects (\textit{Apis mellifera} and \textit{D. melanogaster}), and even \textit{Hydra vulgaris} (reviewed by McPartland, 2004). Homologene identified \textit{DAGL\(\alpha\)} homologs in \textit{Hm}, \textit{Mm}, and \textit{Dm}, and \textit{DAGL\(\beta\)} homologs in \textit{Hm} and \textit{Mm}, and not in \textit{Ce}, \textit{Sc}, \textit{Pf}, or \textit{At}. Unigene identified proteins similar to \textit{Hs}DAGL\(\alpha\) in \textit{Hs}, \textit{Mm}, and \textit{Ce}, as well as \textit{Sc} and \textit{At}. Unigene identified proteins similar to \textit{Hs}DAGL\(\beta\) in \textit{Hs}, \textit{Ce}, and \textit{At}.

4.9. \textit{Conclusions}

The use of sequence profilers, phylogenetic tree analysis, and functional mapping produced conservative lists of orthologs. We found fewer orthologs than did automated annotation systems. Our results suggested the endocannabinoid system was heterogeneously distributed: functional TRPV1 and GPR55 receptors were limited to mammals; CB2 and \textit{DAGL\(\beta\)} were limited to vertebrates; MAGL and COX2-like enzymes were limited to chordates; CB1-like receptors and \textit{DAGL\(\alpha\)} were limited to bilaterian animals; \textit{NAPE-PLD} was limited to the opisthokonta (animals and fungi), and FAAH was limited to eukaryotes.

Some genes shared phylogenetic profiles; that is, the genes were present/absent in the same species. Phylogenetic profiles indicate functional relationships (Pellegrini et al., 1999). For example, CB1, CB2, TRPV1, GPR55, FAAH, and \textit{NAPE-PLD} were absent in the \textit{Dm} genome; these genes code for proteins
associated with AEA, and Dm tissues lacked detectable levels of AEA (McPartland et al., 2001). Indeed the absence of cannabinoid receptors in insects has been described as a sorting event secondary to the loss of AEA (McPartland et al., accepted for publication). This phylogenetic profile was not shared by DAGLα, which was present in the Dm genome, functionally linked to the presence of 2-AG in Dm tissues. We did not see complementary patterns suggestive of nonorthologous gene displacement, in which unrelated or distantly related proteins are responsible for the same function in different organisms. Although the lack of NAPE-PLD in the Ci genome, despite AEA in its tissues (Matias et al., 2005), suggested an unelucidated AEA biosynthesis pathway had evolved in the chordates. This may also explain the appearance of three “new” AEA-gated receptors in chordates (CB2, GPR55, TRPV1).


Acknowledgements

This work was partially supported by an unrestricted grant from GW Pharmaceuticals, Salisbury, Wiltshire SP4 0JQ, UK. The authors thank anonymous journal reviewers for substantial improvements to this manuscript.

References
