UNIVERSITÀ DI PARMA

UNIVERSITÀ DEGLI STUDI DI PARMA

Department of Chemistry, Life sciences and Environmental Sustainability
PhD program in Biotechnology and Life Sciences
XXXIII cycle

Filling the holes in metabolic pathways through Big Data analysis and experimental validation

Coordinator: Professor **Marco Ventura**
Tutors: Professor **Riccardo Percudani**
Professor **Ditlev E. Brodersen**

PhD student: **Marco Malatesta**

2017/2018-2019/2020

Summary

Outline

Aim of research

The aim of the research presented in this thesis is the development of bioinformatic methods for the identification of unknown genes coding already characterized enzymes, also known as *"pathway holes"*.

As case studies, we have considered two different pathway holes, corresponding to enzymatic activities characterized experimentally several years ago, but whose genes have not yet been identified. The first pathway hole that we tried to identify is represented by a missing gene in the biosynthetic pathway of carnitine coding for an enzyme named hydroxytrimethyllysine aldolase (E.C. 4.1.2.-).

The second one, is represented by a missing gene coding for cysteine lyase (E.C. 4.4.1.10), an enzyme discovered in the yolk sac of fertilized chicken eggs, that could be involved in a possible shortcut of the biosynthetic pathway of taurine. The first study is still in progress, while the second one has been published during my PhD [1].

Pathway holes

Metabolism is defined as the set of chemical reactions in organisms. These reactions, as it has been known for a long time, are mostly catalyzed by enzymes, which make them occur much faster, even millions fold. Most of the reactions in metabolism (~75%) are well characterized [2] since we know the enzyme-coding gene, and in some cases the enzyme kinetic parameters, its inhibitors and other characteristics (as its 3D structure, catalytic mechanism, etc.). Even in organisms such as *Homo sapiens* or *Escherichia coli*, where there is much more information available, we can find genes that code for enzymes with unknown function, or vice versa, experimentally determined enzymatic reactions whose gene has not yet been identified.

In the first case, the research focuses on the mapping of a protein sequence within a set of enzymatic reactions. This makes the search quite complicated because the enzyme could be involved in a reaction that is not yet known, so that the search space is unlimited.

In the second case, the research focuses on the mapping of an enzymatic reaction within a proteome. This research is usually simpler because the search space is limited and can be further reduced based on experimental data, such as tissue and/or cellular localization, specific activity, molecular weight, cofactors involved. All these data can narrow the search space to a limited number of candidates.

An enzymatic reaction without an identified gene is defined in Metacyc as a "*pathway hole*" [3]. The name recalls the lack of a known gene within a pathway, but it is also applied to reactions that are not necessarily contextualized within a metabolic pathway. These pathway holes include both orphan enzymes i.e., reactions for which the presence of an enzyme has been demonstrated, and ill-defined reactions, for which there is very little information and not even the certainty of their enzymatic dependency.

Several methods have been devised to select the most promising candidates for filling pathway holes. For example, one can choose some candidate genes using Bayesian methods which can predict the genes coding a pathway hole, based on a homology analysis [3]. In addition, if the reaction considered is similar to other reactions involving known genes, one can search for homologous (more specifically, paralogous) genes.

If the other enzymes involved in the same pathway have been already characterized, a co-occurrence analysis can be performed to select the most co-occurrent gene. Candidate genes can also be selected based on gene neighborhood (particularly in prokaryotes), and similarity in the expression profile or in protein localization with other proteins of the same pathway [4].

In some cases, the main candidate gene for a pathway hole corresponds to a protein with unknown function. On the contrary, in many other cases, we can find as the main candidate, an enzyme with an already characterized function. In this latter case the protein could have a

secondary activity corresponding to our pathway hole, due to the promiscuity observed in many enzymes [5].

PLP-dependent enzymes

Both enzymes studied in this work require pyridoxal-5-phosphate (PLP), the biologically active form of vitamin B6, as cofactor [6], [7]. Thanks to this valuable information, we could carry out the analysis not on the entire proteome, but on the limited set of PLP-dependent enzymes of an organism, referred to as the PLPome. Thanks to this biochemical data, we were able to exploit and use the B6 database (http://bioinformatics.unipr.it/B6db), a specific database for enzymes dependent on vitamin B6, which contains all the PLP-dependent proteins of various organisms classified by family, activity and structure [8].

PLP-dependent enzymes constitute a ubiquitous family in living beings and are mostly involved in the metabolism of amino acids. Although these proteins may have different structures and catalyze various reactions, such transamination, decarboxylation and racemization [9], they originate from a few ancestors; this allows easy the identification of the PLPome of an organism through bioinformatics. Pyridoxal-5'-phosphate is covalently linked by enzymes thanks to a conserved lysine residue which forms a C=N double bond between the aldehyde group of the cofactor and the ε-nitrogen of the amino acid forming a Schiff base (imine). This reaction leads to the formation of the "internal aldimine" [10] (**Figure 1**, left). In the presence of the substrate the PLP detaches from the lysine to attach itself to a primary amine, which is often that of an amino acid, forming an "external aldimine" through a transaldimination reaction (**Figure 1**, right).

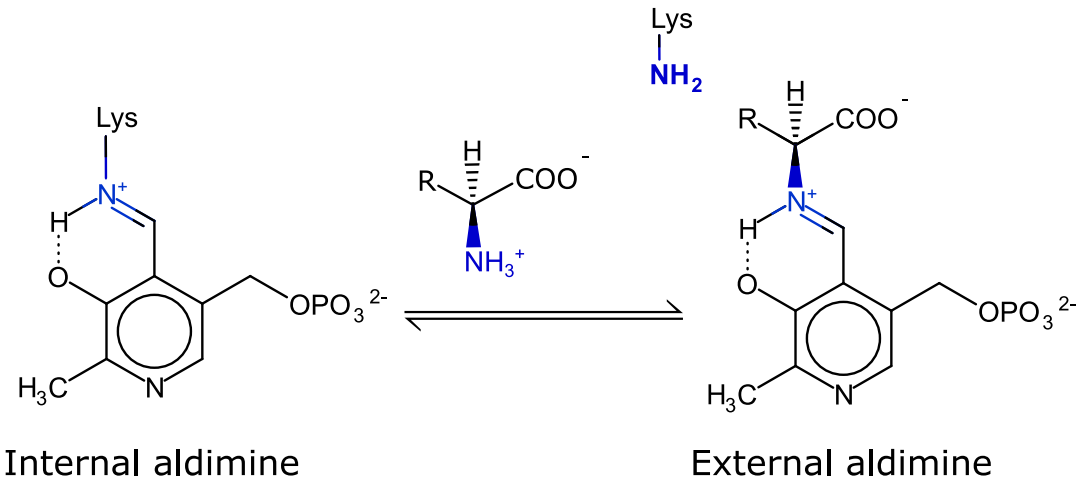


Figure 1. *Transaldimination of PLP-dependent enzymes.*

Vertebrate PLPomes

PLPomes in vertebrates are fairly conserved and only a few exceptions represented by species specific genes. If we consider for example the comparison between the *Homo sapiens* and *Mus musculus* PLPomes, we find 55 common (orthologous) genes with only 2 different enzymes for each species. One of the murine specific genes is Tha1 which encodes for a putative threonine aldolase, which in *Homo sapiens* and other mammals, including some other primates, is a pseudogene [11].

Beyond mammals, with a comparison, for example, between *Homo sapiens* and *Gallus gallus*, we find only 4 enzymes present only in *Gallus gallus* and 8 only in *Homo sapiens*. However, most of the enzymes specifically present in *Homo sapiens* are actually non-neofunctionalized gene duplications, but genes coding for enzymes with identical activities, carried out in a different cellular or tissue context. Consider for instance the three human genes encoding glycogen phosphorylases: PYGL, PYGM, and PYGB which act respectively in the liver, muscle, and brain; we find 2 genes in *Gallus* defined as PYGM-like and therefore a third gene is missing.

Another example is serine hydroxymethyltransferase, which is encoded by 2 genes in *Homo sapiens*, SHMT1 and SHMT2, representing the cytosolic and mitochondrial form, respectively. Only the cytosolic form is present in *Gallus*.

These little differences found in vertebrate's PLPomes can be exploited as a starting point for the identification of pathway holes, in cases in which the activity has been found in a particular organism or group of organisms, as illustrated by the identification of cysteine lyase that will be presented in **Chapter 2**.

Bibliography

1 Malatesta, M. *et al.* (2020) Birth of a pathway for sulfur metabolism in early amniote evolution. *Nat. Ecol. Evol.* 4, 1239–1246

2 Caspi, R. *et al.* (2020) The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* 48, D445–D453

3 Green, M.L. and Karp, P.D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5, 76

4 Nagy, L.G. *et al.* (2020) Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. *Nucleic acids research* DOI: 10.1093/nar/gkz1241

5 Peracchi, A. (2018) The limits of enzyme specificity and the evolution of metabolism. *Trends Biochem. Sci.* 43, 984–996

6 Dunn, W.A. *et al.* (1982) The effects of 1-amino-D-proline on the production of carnitine from exogenous protein-bound trimethyllysine by the perfused rat liver. *J. Biol. Chem.* 257, 7948–7951

7 Tolosa, E.A. *et al.* (1969) Reactions catalysed by cysteine lyase from the yolk sac of chicken embryo. *Biochim. Biophys. Acta* 171, 369–371

8 Percudani, R. and Peracchi, A. (2009) The B6 database: a tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families. *BMC Bioinformatics* 10, 273

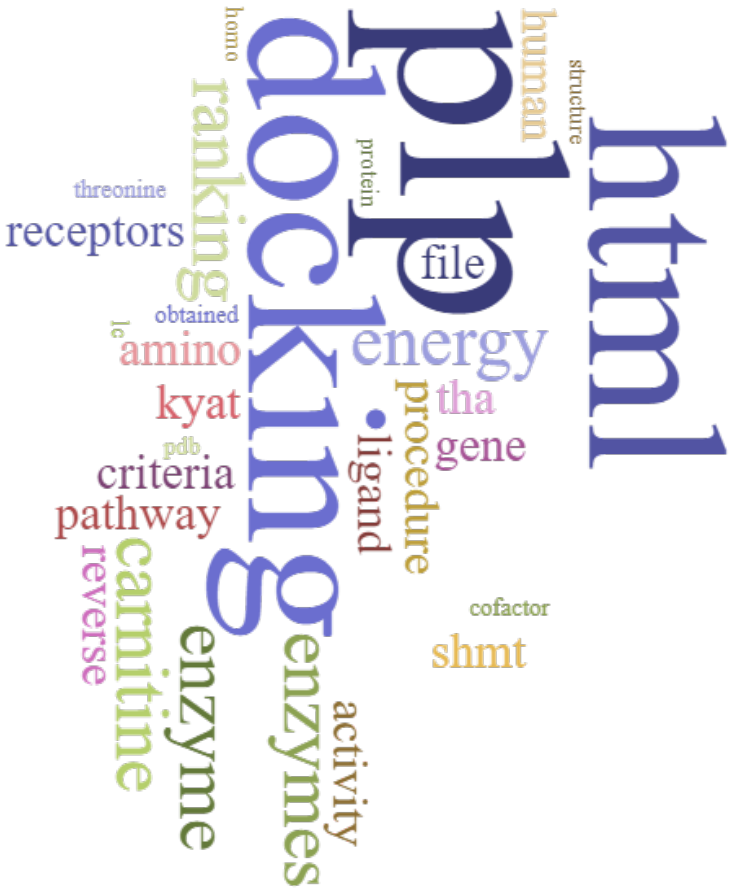
9 Catazaro, J. *et al.* (2014) Functional evolution of PLP-dependent enzymes based on active-site structural similarities. *Proteins* 82, 2597–2608

10 Oliveira, E.F. *et al.* (2011) Mechanism of formation of the internal aldimine in pyridoxal 5'-phosphate-dependent enzymes. *J. Am. Chem. Soc.* 133, 15496–15505

11 Edgar, A.J. (2005) Mice have a transcribed L-threonine aldolase/GLY1 gene, but the human GLY1 gene is a non-processed pseudogene. *BMC Genomics* 6, 32

Chapter 1

Development of a reverse docking procedure for identification of a *pathway hole* in carnitine biosynthesis



Introduction

Carnitine biosynthesis

Carnitine (L-3-hydroxy-4-trimethylaminobutyrate) is an amino compound essential for energy metabolism, in fact its main function is the transport of fatty acyl-CoAs from the cytosol to the mitochondria where they enter the biochemical process of β-oxidation [1].
Carnitine was found the first time in bovine homogenates, hence the name from Latin word *carnem*, coming from meat. It was initially thought to be a vitamin (vitamin B_T), for its indispensability for the growth of *Tenebrio molitor,* the darkling beetle mealworm.
Carnitine is present in most animal species as fungi, protists, and to a lesser extent also in plants [2],[3] and for some of these organisms, which are unable to synthesize it, carnitine could have an effective role of vitamin. Many other organisms instead, including most mammals, are able to synthesize carnitine through four enzymatic reactions that make up the biosynthetic pathway. The presence of the enzymatic pathway has been demonstrated in simpler organisms such as some unicellular fungi (e.g., *Candida albicans*) [4]. In many eukaryotic species, however, are present orthologs of the known genes encoding the enzymes of the pathway, suggesting a more extensive presence of this biosynthetic pathway even in the absence of complete experimental evidence.

Human carnitine biosynthesis

In *Homo sapiens,* the carnitine biosynthesis begins with the degradation of N^6-trimethyllysine (TML) [5], a post-translationally modified amino acid released from the degradation of proteins such as histones, calmodulin, cytochrome c, and myosin [6]. The transformation occurs initially at the submitochondrial level with an oxidation reaction [7], only to then move into the cytosol with the other three steps. (**Figure 1**)

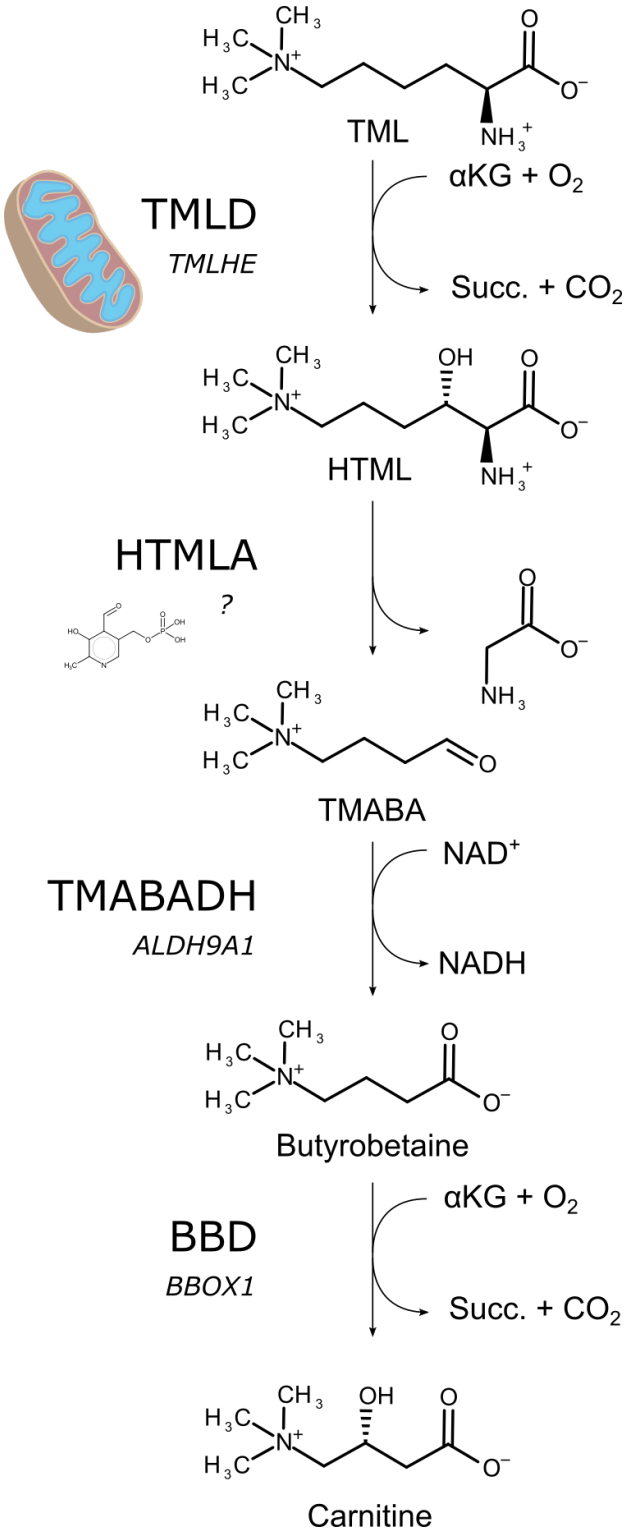


Figure1. *Trimethylated lysine is converted into carnitine through four enzymatic reactions. Only the first reaction occurs in the mitochondrion, the name of the enzymes is in bold and the genes that encode them in italics. The second enzyme does not yet have a characterized gene, but it requires PLP as cofactor. TML: N6-trimethyllysine, TMLD: TML dioxygenase, HTML: 3-hydroxy-TML, HTMLA: HTML aldolase, TMABA: 4-N-trimethyl-aminobutyraldehyde, TMABADH: TMABA dehydrogenase, BBD: butyrobetaine dioxygenase, αKG: α-ketoglutarate, Succ.: succinate.*

TMLD

The gene encoding the first enzyme of the pathway is TMLHE, located on the long arm of the X chromosome (Xq28). It is composed of 11 exons and it is expressed to a greater extent in the liver. The encoded protein has a mass of ~87 kDa with homodimer association and presents a mitochondrial localization sequence at its N-terminus. The enzyme TML dioxygenase (TMLD) catalyzes the addition of a hydroxyl group on the beta carbon of TML, using 2-oxoglutarate, Fe^{2+} and molecular oxygen as cofactors [7], producing 3-hydroxy-N^6-trimethyllysine (HTML).

HTMLA

The second reaction of the pathway involves a retro-aldol cleavage of HTML with the formation of 4-*N*-trimethyl-aminobutyraldehyde (TMABA) and glycine by HTML aldolase enzyme (HTMLA). The gene of this enzyme in humans is still unknown, although the enzymatic activity was partially characterized:

- The enzyme uses pyridoxal-5'-phosphate (PLP) as cofactor. This has been demonstrated in the rat liver with the use of amino-D-proline, a PLP antagonist, capable of blocking the reaction pathway by accumulating the HTML substrate in cells treated with this compound [8];
- The protein is localized at the cellular level in the cytosol and at the tissue level in liver, kidney and heart, in a decreasing order of enzymatic activity [9].
- The only organism in which the gene of this enzyme has been identified is *Candida albicans* [4]. In this saccharomycetales the reaction is catalyzed by the gene product of *GLY2,* a paralogue gene of *GLY1* which encodes for a threonine aldolase that similarly requires PLP as cofactor. In fact, threonine is an amino acid with a hydroxyl group on the β carbon, as well as HTML.

TMABA-DH

The gene ALDH9A1 encodes the third enzyme of the pathway, the TMABA dehydrogenase (TMABA-DH), and it is located on chromosome 1 (1q24.1). TMABA-DH is a highly conserved enzyme that uses nicotinamide adenine dinucleotide (NADH) as cofactor and can catalyze other reactions besides its main one, the conversion of TMABA in γ-butyrobetaine (BB) [10].

BBD

The last enzyme of the pathway is encoded by the BBOX1 gene which is located on the chromosome 11 (11p14.2) and it is a paralogue of TMLHE. The γ-butyrobetaine dioxygenase (BBD) catalyzes the hydroxylation of BB to yield L-carnitine, the final product [11]. Unlike its

paralogue, the BBD does not have a sequence for mitochondrial localization and carries out its activity in the cytoplasm. Both enzymes, TMLD and BBD, have two domains the hydroxylase one (TauD; aa 108-366) and the accessory one (DUF971; aa 9-92) without unknown function yet.

Hypotheses for the HTMLA pathway hole

The first hypothesis about the unidentified gene of the second enzyme HTMLA concerns the possible existence of a homolog of the *GLY2* gene of *Candida albicans*. However, *Homo sapiens* does not have orthologs of the *GLY1* or *GLY2* genes of *C. albicans*. A candidate mammalian ortholog, named Tha1, was in fact lost in some primates becoming a pseudogene due to a nonsense mutation. Most other mammals, such as *Mus musculus*, instead possess Tha1 which codes for an enzyme without a characterized activity [12]. This gene could encode the HTMLA enzyme in the organisms in which it is present. However, the possibility that human HTMLA is an ortholog of *Candida albicans* threonine aldolase is excluded. Therefore, there must be another PLP-dependent enzyme that catalyzes the aldolase reaction in *Homo sapiens*.

For the identification of HTMLA gene we decided to investigate the human and murine PLPomes (the full set of PLP-dependent enzymes of an organism) with the aim to discriminate between two possible hypotheses:

- In mammals, the orthogroup represented by the putative threonine aldolase Tha1 in *Mus musculus* encodes the HTMLA enzyme of the carnitine pathway, while in some primates, including *Homo sapiens*, there is another PLP-dependent gene capable of carrying out the activity of HTMLA, consequently to the pseudogenization of their threonine aldolase ortholog.
- The HTMLA coding gene is different from Tha1 and is conserved in all mammals including primates such as *Homo sapiens*.

In *Homo sapiens* there are two main candidates for this reaction, given their ability to act as aldolases towards beta hydroxylated substrates such as threonine or *allo*-threonine. They are the two cytosolic and mitochondrial isozymes with serine-hydroxymethyltransferase activity (SHMT1 and SHMT2).

The physiological reaction of these enzymes is the interconversion between glycine and serine, through the transfer of a hydroxymethyl (CH3-OH) by tetrahydrofolate (THF), the active form of vitamin B9. PLP binds the amino acid amino group, while THF acts on the side chain. In these enzymes an aldolase activity was also reported, which does not require THF as a cofactor, but only PLP [13].

An activity towards HTML has been observed using the SHMT1 enzyme in crystalline form [7], and also the soluble forms of SHMT1 and SHMT2 [14]. These studies report a low activity of the enzymes towards HTML, although a calculation of the catalytic parameters was not performed.

Through phylogenetic and co-occurrence analysis between the known enzymes of the pathway and the mammalian PLPomes, we were unable to find a valid candidate based on our bioinformatic analysis. In fact, we did not identify any PLP-dependent gene co-occurring with the other three genes of the pathway (**Figure 2**). For this reason, we have decided to select candidates through a structure-based analysis using molecular docking, and in particular a technique known as reverse docking.

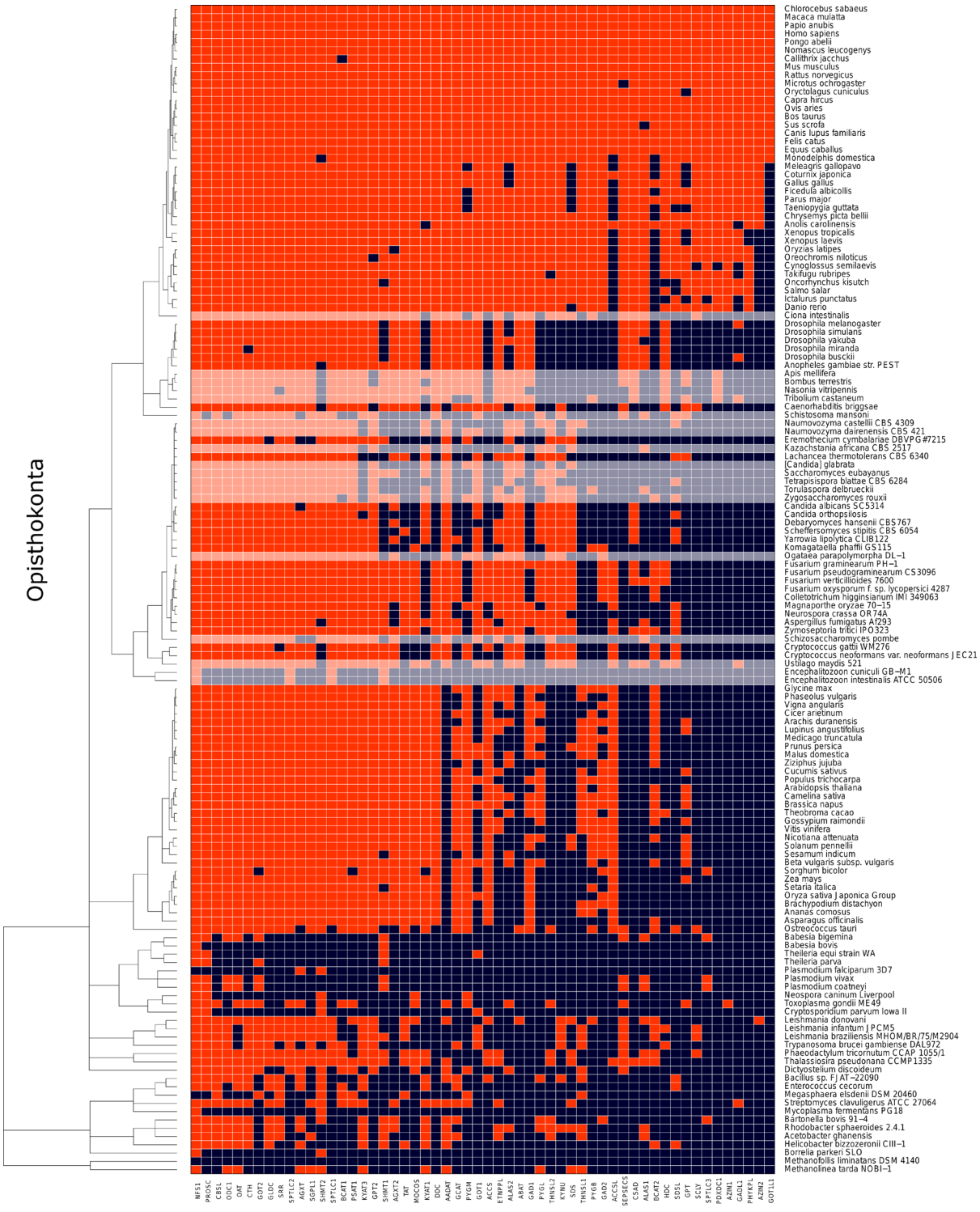


Figure 2. *Occurrence analysis of the human PLPome in different organisms. Faded rows indicate organisms in opisthokonta which do not have both the two paralogues TMLHE and BBOX1.*

Enzyme-substrate docking

Molecular docking is a widespread technique principally used for rational drug design. This technique could be applied even as reverse docking, used so far for drug repositioning, when we want to discover new targets of already known drugs [15].
The receptors used in a docking screening are usually targets of molecules, which bind to a binding site of the protein without being transformed. In a reverse docking, used for example for drug repositioning, the protein-ligand (P-L) predicted complex could be an enzyme with its inhibitor, or a receptor with its ligand. The conformation with the lowest free binding energy will be selected as the best result. Instead, if one wants to use reverse docking for enzyme identification using its substrate as ligand, must consider the enzyme-substrate (ES) complex. Ideally, the best molecule that should be used for reverse docking in the case of enzymes is the transition state of the reaction. Some attempts have been made to simulate the transition state, including the use of the high-energy intermediates to predict substrates of the enzymes [16]. Even if the use of high-energy intermediates provides better results in comparison to the use of substrate ground state, there is no guarantee that the predicted ES complex could be catalytically competent, due to the intrinsic limits of this technique [16].

Molecular docking with Autodock

To perform the reverse docking, we used Autodock4.2 [17] which is composed by two separate programs, Autogrid and Autodock, that require respectively a grid_parameter_file.gpf (GPF) and a docking_parameter_file.dpf (DPF) as input files, as well as the receptor and the ligand in pdbqt format.
Autogrid is used for the construction of non-binding energy maps for each type of ligand atom within a predetermined grid of *NNN* dimensions centered in *xyz* coordinates, corresponding to the site or the region in which one wants to dock the ligand. During the docking procedure, each ligand position will be energetically evaluated by adding the values obtained by interpolating the points of the map around each atom of the ligand.
The GPF contains, as input elements, the dimensions and the coordinates of the grid, the name under which the various calculated maps will be saved, and the grid spacing, which corresponds by default to 0.375 Å (¼ of the C-C bond).
Autodock uses as input a DPF that contains the parameters for the calculation of the binding mode. These include the docking algorithm, the number of runs, the maximum number of energy evaluations, the population size, and all energy maps calculated in advance with Autogrid. The most used algorithm, which appears to be the most effective, is the Lamarckian Genetic Algorithm (LGA) [18]. The LGA is able to choose the best conformation of the ligand

through random generations of an initial random population of individuals (different conformations of the ligand), which are selected, mated, mutated, subjected to a selection of the best conformation for each generation (natural selection). This ultimately leads to the generation of an offspring with subsequent structural optimization (through a minimization of energy), inherited by the children conformations; hence the "Lamarckian" term that takes up the theory of acquired characters.

Receptors availability

The greatest difficulty we encounter when we approach to molecular docking is probably the availability of a resolved protein structure which is not always available. In PDB (https://www.rcsb.org/stats) there are more than 170 thousand structures with around 133 thousand of these representing protein structures. However, the genes represented by these structures are only ~52,000, given the redundancy of many structures per gene e.g., with different substrates, different mutants, or simply different authors.

For example, in the PDB there are more than 49,000 structures representing only ~9,000 genes of *Homo sapiens*, this means that just under half of the human proteome has a resolved structure, despite being one of the most studied organisms.

Hence the need to obtain a structural model of a protein, even when it is not solved experimentally, that somehow fills the holes in the PDB database, with an ongoing development of techniques such as homology modelling, fold recognition and AI structure predictors. In the SWISS-MODEL website (https://swissmodel.expasy.org/), there is a repository for the reference organisms such as *Homo sapiens*, *Mus musculus* and *Escherichia coli* that contains all available structures, including PDB data, and pre-built models obtained by homology [19]. In the human archive 18,000 sequences of 20,000 proteins of the *Homo sapiens* proteome are modeled or solved experimentally. This is of great help if you want to use a structure not present in the PDB but bearing in mind the possible limitations of using these models in docking, especially in protein-ligand docking where the orientation of a single residue side chain can greatly affect the result. This last aspect also occurs using PDB structures with low resolutions higher than 2.5-3 Å, which could be penalized, perhaps not energetically, but certainly by the concreteness of the results.

Limits of molecular docking

One of the main weaknesses of molecular docking could be the static nature of the receptor. There have been many efforts to overcome this problem, for example using software that allow flexible docking, or with the use of molecular dynamics simulations once obtained a predicted binding mode by a docking run [20]. This limit is especially relevant if we consider the induced

fit that exists to a different extent for any proteins. For example, if the structure of an enzyme has been solved in the presence of its substrate or substrate's analog, the conformation of the active site will already be fitted for the molecule binding, and therefore the docking will be decidedly more consistent than that obtained starting from a structure solved in the absence of its substrate. When we apply reverse docking to a receptor set, we encounter additional difficulties compared to docking to a single receptor. The main one is probably the heterogeneity of the receptor set, due to different quality and conditions of the structural models. In fact, in a reverse docking screening we would like our dataset to be as homogeneous as possible, to avoid incurring biases that can mislead our analysis.

Active site aromatic boxes

In the PDB structure of BBD, the last enzyme of the pathway, there are some aromatic residues in the active site that interact with the quaternary ammonium, also known as quat, of the butyrobetaine substrate with a cation-π interaction between the positive charge of the quat, and the partial negative charge of the π orbitals of aromatic rings (**Figure 3a**). Thanks to a structural alignment with the structure model of the paralogue TMLD, the first enzyme of carnitine biosynthesis, we see how the aromatic residues are perfectly conserved (**Figure 3a**). We also find this feature in the active site of TMABADH, the third enzyme of the pathway, by docking its substrate TMABA (**Figure 3b**). These active sites have a characteristic composition called *aromatic box*, spread in proteins that bind N-trimethylated substrates, such as cholinergic receptors or histone tail methylase [21]. For this reason, it is conceivable that also the second enzyme of the carnitine pathway, HTMLA, could also have this feature since the quat is present in all intermediates of this pathway (see **Figure 1**).

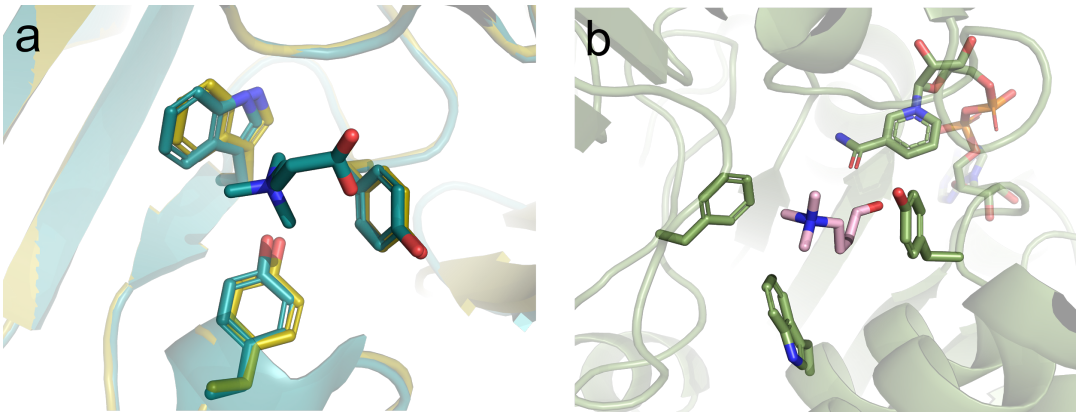


Figure 3. *(**a**) Structural alignment between 3O2G structure in PDB representing BBD, and the model from human SWISS-MODEL repository of TMLD. Butyrobetaine is crystallized together with BBD in blue, and there is a well conserved aromatic box in both paralogues. (**b**) 6VR6 structure in PDB*

representing TMABADH, the third enzyme of the pathway, where the substrate TMABA (pink) was docked using Autodock4.2.

Results

Tuning the reverse docking procedure

We developed a procedure that allows us to perform a reverse docking screening, starting by a receptor set of pdb file, a coordinates file with the center of the grid for each structure, and a manually prepared ligand. Our procedure is written in python3, and through some precompiled script of Autodock4.2 and Autodocktools, it can carry out a reverse docking screening by compiling all the input files necessary for Autogrid and Autodock, providing the final output in a csv file summarizing the energies reported in the different dlg files of the screening.

Retrievement of the receptor sets from the SWISS-MODEL repository

We used human and mouse PLPomes (see **Table S1, S2**) as our receptor sets, retrieving the protein accessions of the corresponding genes in the B6DB using the option "genomic search". Through a dedicated python script, we downloaded all the available structural models of the PLP-dependent enzymes of *Homo sapiens* (202) and *Mus musculus* (117) from the SWISS-MODEL repository. In our dataset of 319 receptors, there are 133 experimentally solved structures in the human set and 22 in the mouse one. These experimentally solved structures are with or without active site ligands and at different resolutions. This makes the dataset very heterogeneous with the risk of penalizing genes that are less represented or represented by models with a bad *Qualitative Model Energy ANalysis* (QMEAN). It should be noted, however, that most of the models in our dataset have acceptable or good QMEAN values (**Figure 4**).
Although the representation of each gene in our set is very heterogeneous, we decided to use all the available structures in order not to miss potentially interesting docking results.

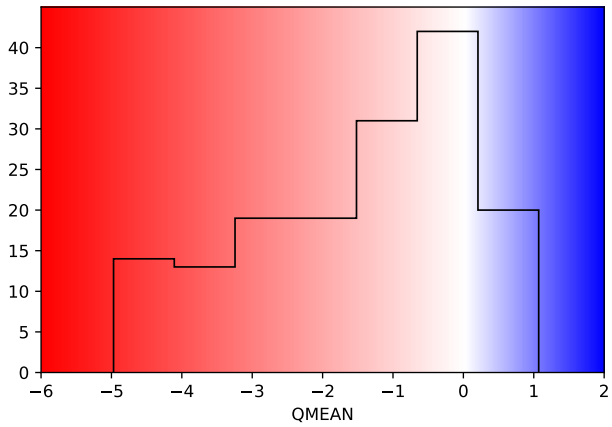


Figure 4. *QMEAN distribution of the 164 models in both the receptors sets. Red-white-blue color gradient is used according to SWISS-MODEL. Most of the models in our dataset have acceptable or good QMEAN values (>-4).*

Grid positioning on the catalytic lysine

To optimize and improve the docking results, it is necessary to limit the sampling space (grid box) to the active site, in order to explore only the space of interest; this strategy permits to shorten the calculation times and obtain statistically more significant results in terms of population of the clusters.

To automate this process, we used the Uniprot data of the catalytic lysine of each sequence in our PLPomes, classified in PTM section as *"N6-(pyridoxal phosphate)lysine"*, and we retrieved the coordinates of the relative structures through the script *get_models_coord.py* obtaining a table containing the names of the receptors and the corresponding coordinates used as center of the grid. As dimension we used a cube of 55 points (20.625 Å) per side which corresponds to 8773 Å^3, sufficient to cover the active site of all the enzymes considered.

Four of the 57 entries considered for *Homo sapiens* (see **Table S1**), and five of the 57 entries considered for *Mus musculus* (see **Table S2**), miss the catalytic lysine in the PTM features of the corresponding Uniprot record, and for AZIN1, Azin1, Ldc1, we retrieved manually the correct residue number in the B6DB, while for the enzymes SPTLC1, Sptlc1, AZIN2, Azin2, PDXDC1, Pdxdc1 and PDXDC2 we were not able to identify a catalytic lysine.

Since many structures are homo-multimers with multiple active sites, and sometimes these sites are not perfectly equal, we decided to analyze all the active sites of all the available structures of the entire dataset. For this reason, we have considered 216 sites for the *Mus musculus* receptor set and 415 for the *Homo sapiens* one.

Reverse docking procedure validation

In addition to our ligand of interest HTML and to the PLP cofactor, used as an energy tare, seven amino acids, that are known substrates for 15 human and 15 mouse enzymes of our receptor sets (**Table 1**), were chosen as ligands in our docking procedure. This allowed us to validate our procedure and select the ranking criterion that in the screening of each ligand identified the corresponding enzymes in the top positions with respect to the entire dataset.
In fact, in the Autodock output (dlg file), the conformations obtained in each run are clustered based on the RMSD with a default threshold of 2 Å so that each cluster represents a different binding mode (**Figure 5**). As ranking criterion one can consider the cluster with the best (i.e., lowest) energy (BC), that represents the binding mode energetically favorite, or the best energy of the largest cluster (LC), that is instead the statistically favorite one. In an optimal result we expect these two clusters coincide (**Figure 5,** left panel), but this is not always the case. Often BC has few conformations which achieved better energy than most docking runs (**Figure 5,** right panel).

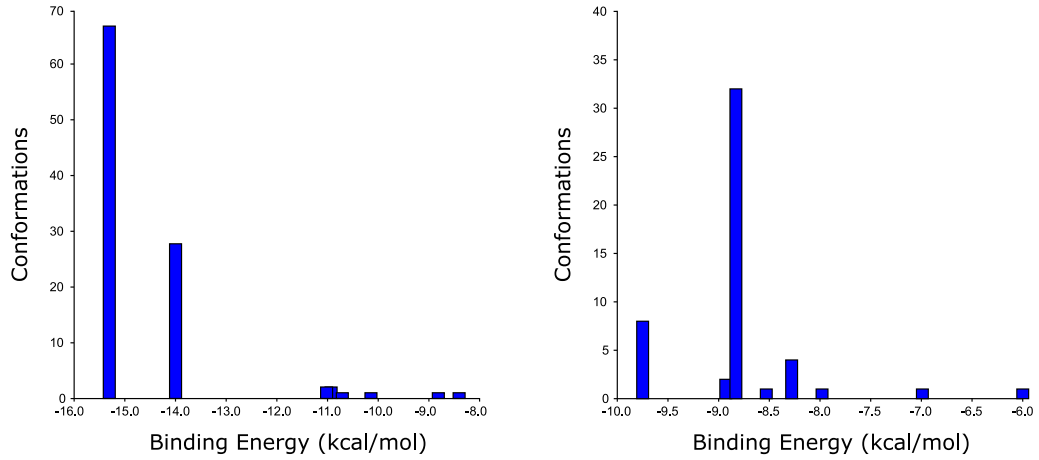


Figure 5. *The largest cluster (LC) is the one with most conformations, while the best cluster (BC) is the one with the lowest energy. In the left panel LC and BC coincide, and there is a preference for the binding mode either for an energetic and a statistical point of view. In the right panel, LC is different from BC, and here one must choose whether to consider one or the other.*

Given the catalytic mechanism of PLP-dependent enzymes, we used as ligands the amino acids bound covalently to the cofactor as external aldimine (**Figure 6**) that is the first step of the reaction mechanism. The use of a molecule with additional anchor points for the docking could be advantageous to guide the substrate in the right direction in the active site, while and at the same time, providing a more solid RMSD clustering.

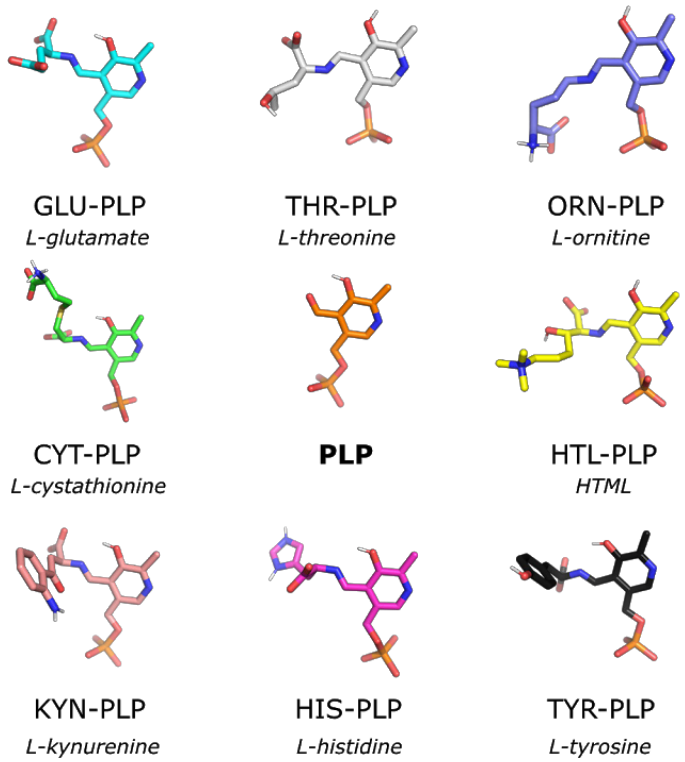


Figure 6. *External aldimine ligands used for the screening procedure validation. Each ligand is named with the amino acids' common names in italic and with its PDB three-letter codes plus the "PLP" suffix.*

Substrates	Enzymes
Kynurenine (KYN-PLP)	KYAT1, KYAT3, KYNU, AADAT, Kyat1, Kyat3, Kynu, Aadat
Tyrosine (TYR-PLP)	TAT, DDC, Tat, Ddc
Histidine (HIS-PLP)	HDC, Hdc
Glutamate (GLU-PLP)	GAD1, GAD2, GOT1, GOT2, Gad1, Gad2, Got1, Got2
Cystathionine (CYT-PLP)	CTH, Cth
Threonine (THR-PLP)	SDS, SDSL, Sds, Sdsl
Ornithine (ORN-PLP)	OAT, Oat

Table 1. *Substrates used as ligand for the entire screening and the relative enzymes used as positive controls as receptors. The corresponding human (uppercase) and mouse (lowercase) enzymes are indicated according to the Uniprot records.*

The QMEAN values and the number of structures per enzyme was not significantly correlated with the energies obtained in the screening in each ligand, as demonstrated by the Pearson index obtained always less than |0.38| (**Table S2, S3**).

LCaaM is the best ranking criteria

We initially considered four ranking criteria (**Figure 7,** four leftmost panels) that consider the best cluster and the largest cluster with their best (BCB, LCB) and with their mean values (BCM, LCM). All the Autodock output dlg files have been parsed with *get_atom_energy.py* (see **Methods**) that creates a table with all the energy values considered. This table was then sorted with the different ranking criteria and it was kept only the best result for each enzyme. The comparison between the rankings obtained by the screening with the different ligands, highlighted a bias due to the presence of the PLP cofactor in all ligands used. The energy contribution given by PLP, drags the different receptors of the datasets into the same positions, which inevitably maintain similar positions in the ranking even with very different substrates (**Figure 8a**). The mean Pearson correlation between the ranking of every receptor obtained with PLP-docking and the ranking obtained with all the other substrates was > 0.71 in all considered criteria (**Table 2**).

To overcome this problem, we have tried to eliminate the energy contribution of the PLP by considering only the non-binding energies of the atoms corresponding to the amino acid in each ligand, represented by electrostatic and van der Waals interactions calculated in each run for every atom.

We have thus considered four additional ranking criteria (**Figure 7,** four rightmost panels) analogous to the previous ones, but with the exclusion of the contribution of PLP cofactor to the binding energies (BCaaB, BCaaM, LCaaB, LCaaM).

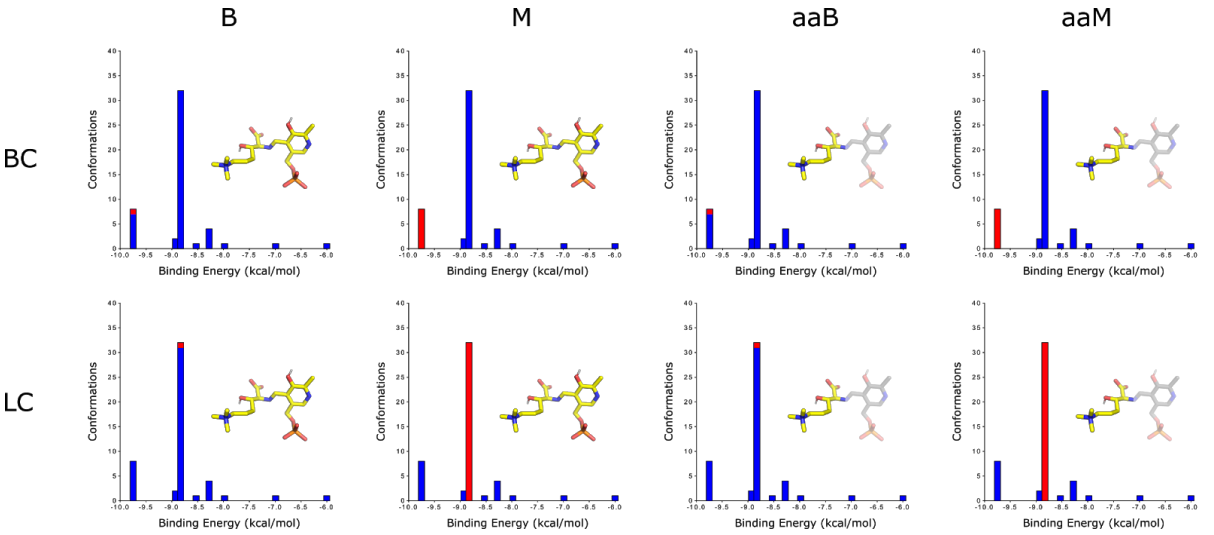


Figure 7. *The different ranking criteria for best cluster (BC) and largest cluster (LC) considered in the validation procedure. The best energy of each cluster (B, aaB), and the average of all the energies of the cluster (M, aaM), including or not the energy contribution of the PLP. Red color highlights the conformation considered for the calculations: all the cluster conformations (M), or only that with the lowest energy (B). PLP are grey faded in the structure figure of the criteria *aa*, indicating the exclusion of the cofactor energies.*

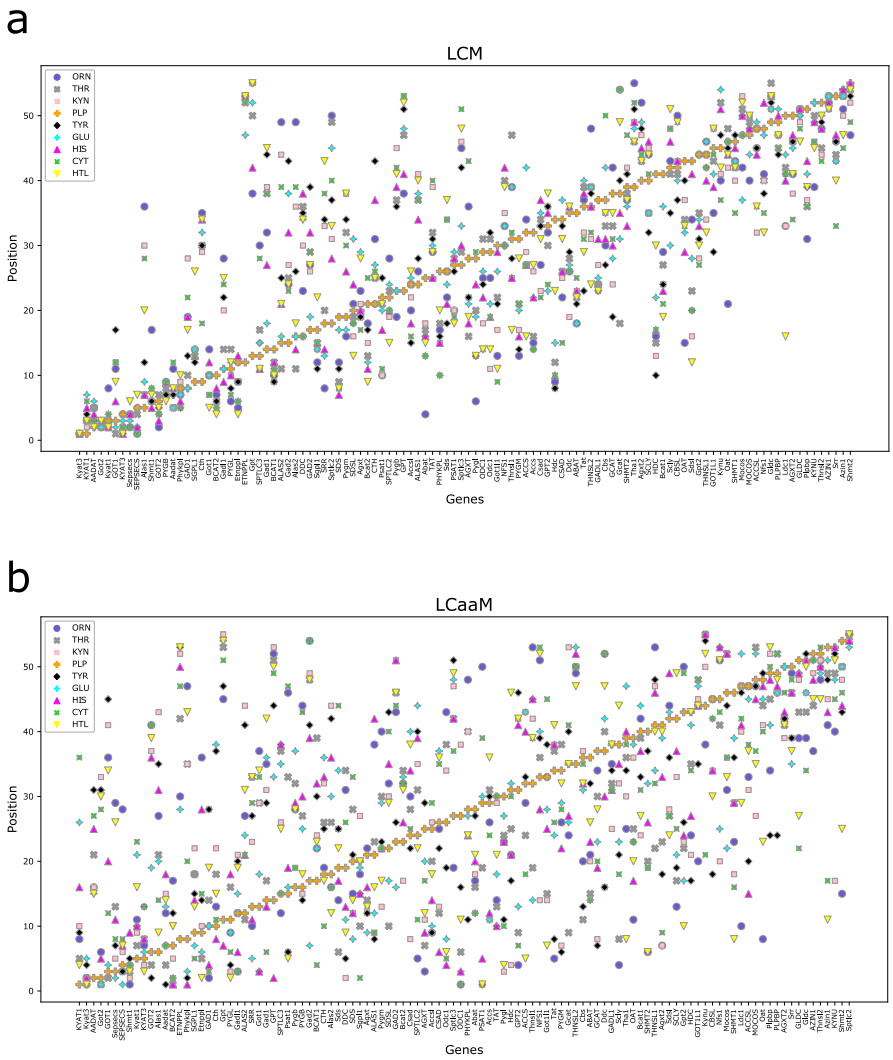


Figure 8. *Scatter plots of the ranking of the different receptors in the dataset (y axis) using the nine different ligands, ordered on the x axis according to the ranking obtained by reverse docking with the PLP ligand (orange triangle). Using the LCM criterion (**a**), data points obtained with the other ligands tend to cluster around the diagonal. By contrast, when the energies of the amino acid alone (LCaaM, **b**) are considered, this clustering decreases. It is also noted that clustering is stronger in the two extreme corners of the plots.*

Looking at the correlation between the PLP and the other ligands in all ranking criteria (**Table 2, Figure 8b**), all the criteria that exclude the PLP (*aa*) have a lower correlation than the criteria that consider the whole ligand.

This lower correlation makes the results more solid, decreasing the bias in the energy ranking. Certainly, the presence of PLP in the ligand somehow influences the binding mode of the amino acid as well, but by excluding its energies later, we were able to minimize the bias due to the PLP binding energies.

	PLP BCB	PLP LCB	PLP BCaaB	PLP LCaaB	PLP BCM	PLP LCM	PLP BCaaM	PLP LCaaM
ORN	0.650652	0.646924	0.323785	0.326486	0.637377	0.656929	0.248393	0.314732
THR	0.861159	0.786717	0.511697	0.550116	0.850698	0.775153	0.562593	0.607821
KYN	0.716878	0.721062	0.450645	0.427935	0.598235	0.663091	0.431017	0.372133
TYR	0.736658	0.724790	0.493590	0.539998	0.709422	0.696641	0.446955	0.465746
GLU	0.856099	0.795390	0.693027	0.677774	0.849481	0.806040	0.702233	0.686751
HIS	0.806877	0.826657	0.580395	0.664232	0.792232	0.813154	0.627182	0.654608
CYT	0.670204	0.681464	0.466164	0.498383	0.641066	0.651375	0.479440	0.469207
HTL	0.742858	0.710449	0.461790	0.468180	0.709917	0.668987	0.463197	0.400319
Mean	0.755173	0.736682	0.497637	0.519138	0.723553	0.716421	0.495126	0.496415

Table 2. *Correlation between PLP ranking and the other ligands in every ranking criterion colored by a gradient from white (min value) to dark blue (max value) of the dataframe. The last row shows the mean value for each column. The ranking criteria that consider only the amino acid (*aa*) are lighter than the others that include PLP cofactor, showing a lower mean correlation.*

From the "letter-value plot" (**Figure 9a**) it is possible to appreciate how all the ranking criteria considered are better than a random ranking used as a control. The one that obtained the highest median and third quartile ranking values was the criterion that considers the mean of the energies of the amino acid alone in the largest cluster (LCaaM). In theory, by using this ranking criterion, we have a 50% chance of finding the right enzyme in the first 10 positions (as shown by dotted line in **Figure 9a**), and a 75% chance of finding it in the first 18 ones.

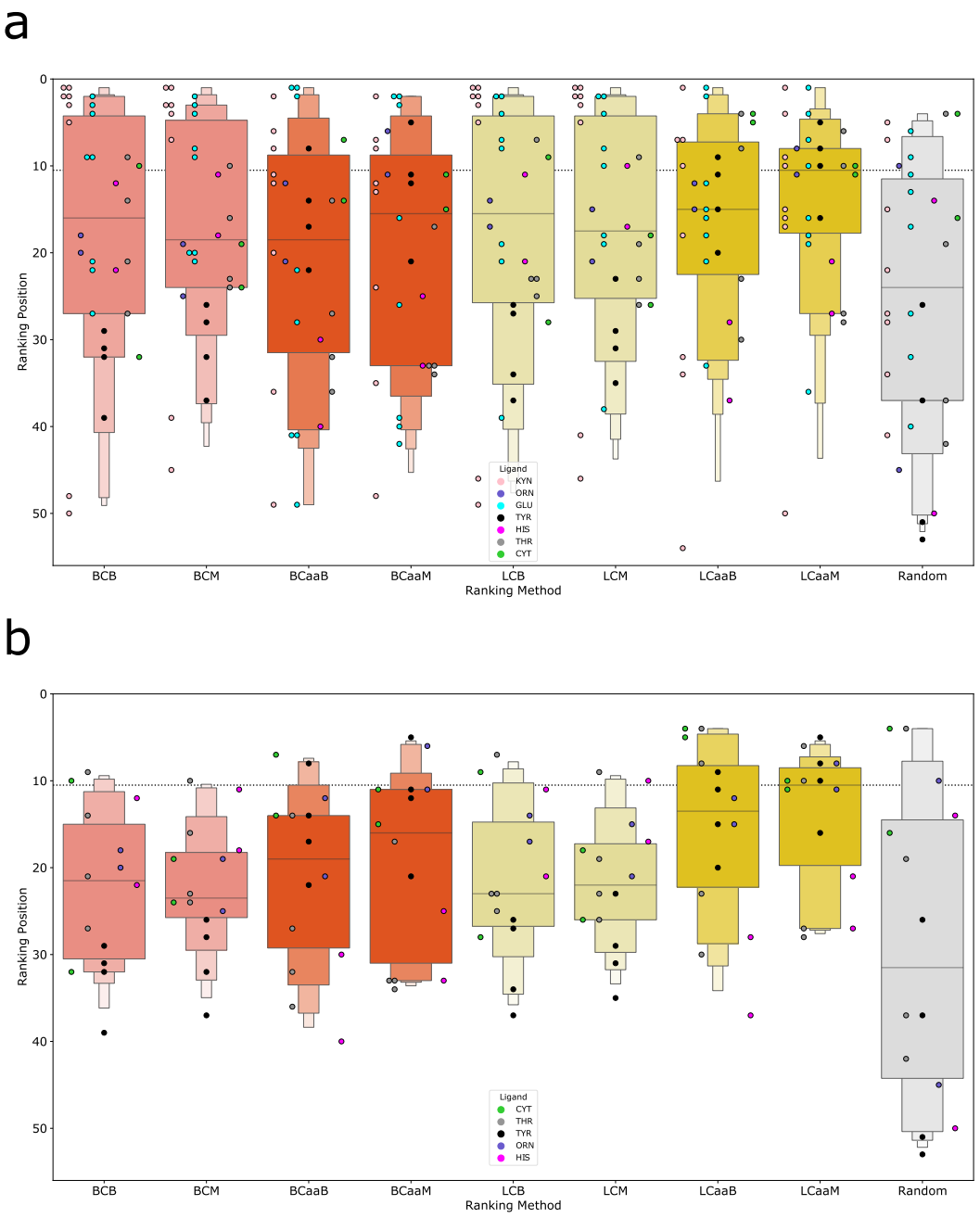


Figure 9. *Letter-value plots of different ranking criteria. (**a**) Colored dots indicate the ranking of the receptors known to bind the different ligands (y axis), obtained by applying different ranking criteria (x axis). BC ranking criteria are colored in orange, while LC criteria are colored in yellow. The darker shades refer to the *aa* rankings. The grey plot shows a random sorting. The dotted line shows the lower median value for all criteria distribution, that correspond always to LCaaM criterion. Different ligands used in the validation are marked in different colors as shown in legend. (**b**) The same ranking criteria excluding KYN and GLU ligands.*

Almost all the ranking criteria that consider the largest cluster (LC*) have obtained better results than that ones that refer to the best cluster (BC*) (**Figure 9**). Considering LC, we see how there is a progressive improvement of the positions, firstly by evaluating its mean value (LCM) and then considering only the energy contribution of the amino acid (LCaaB) and the respective mean value (LCaaM) (**Figure 9**).

The first quartile of the distributions of all the criteria that consider the energies of the entire ligand (not aa) highlighted with lighter shades, is very squashed upwards. This is due to the presence of some enzymes in our dataset (i.e., KYAT1,3, GOT1) that in the validation procedure are highly rewarded by the PLP binding energies and occupy the first positions of the ranking irrespectively of the bound aa, as evidenced by the LC scatter plot (see **Figure 8a, b**).

In fact, by eliminating from the distribution the ligands of the receptors, that are most influenced by the bias (KYN and GLU, indicated by pink and cyan dots), the ranking distribution worsen in the case of the non *aa* criteria, while it remains almost unchanged with the *aa* criteria (**Figure 9b**).

Reverse docking with HTML

HTML_PLP screening reveals candidates for HTMLA in *Homo sapiens*

We used the LCaaM ranking criterion, which achieved the best results in the screening with ligands with known receptors, for a screening with the HTML substrate. The aim of this screening was to find some valid candidates in human and mouse PLPomes to be experimentally validated as the possible HTML aldolase.

In the case of the human PLPome, we notice several interesting hits (**Table 3**). In the 1st position we find the enzyme phosphoserine aminotransferase (PSAT1), which was disregarded due to the very few conformations in the LC (only 8%), suggesting that there is no statistically favorite binding mode of HTML-PLP in this receptor, despite the lowest energy value.

The second enzyme of the ranking is the liver form of glycogen phosphorylase (PYGL), a unique case of a PLP-dependent enzyme that does not form the external aldimine with the substrate, but uses the cofactor phosphate group in catalysis [22]. For this reason, it was regarded as an unlikely candidate, even if a role of PYGL as HTML aldolase would represent an intriguing connection between carbohydrate and fat catabolism.

In the 3rd and 4th positions, on the other hand, we find two interesting isozymes with kynurenine aminotransferase activity (KYAT1, KYAT3). These enzymes also have a secondary lyase activity towards the cysteine-S-conjugated compounds, which gives them the alternative name of CCBL1 and CCBL2.

	Entry	Gene names	EC number	LCaaM (kcal/mol)	Num in LC
1	Q9Y617	PSAT1 PSA	2.6.1.52	-7.52	8
2	P06737	PYGL	2.4.1.1	-6.82	27
3	Q6YP21	KYAT3 CCBL2 KAT3	2.6.1.7; 4.4.1.13; 2.6.1.63	-6.80	70
4	Q16773	KYAT1 CCBL1	2.6.1.7; 4.4.1.13; 2.6.1.64	-6.32	65
5	Q96QU6	ACCS PHACS	-	-6.28	9
6	P34897	SHMT2	2.1.2.1	-6.03	34
7	Q9HD40	SEPSECS TRNP48	2.9.1.2	-5.92	43
8	P34896	SHMT1	2.1.2.1	-5.87	19
9	P20711	DDC AADC	4.1.1.28	-5.86	12
10	O15382	BCAT2 BCATM BCT2 ECA40	2.6.1.42	-5.82	71
11	P11926	ODC1	4.1.1.17	-5.73	23
12	P21549	AGXT AGT1 SPAT	2.6.1.51; 2.6.1.44	-5.68	70
13	O95470	SGPL1 KIAA1252	4.1.2.27	-5.63	25
14	P20132	SDS SDH	4.3.1.17; 4.3.1.19	-5.57	75
15	Q8N5Z0	AADAT KAT2	2.6.1.39; 2.6.1.7	-5.56	76
16	P13196	ALAS1 ALAS3 ALASH OK/SW-cl.121	2.3.1.37	-5.54	16
17	P32929	CTH	4.4.1.1	-5.51	31
18	P11216	PYGB	2.4.1.1	-5.50	35
19	Q99259	GAD1 GAD GAD67	4.1.1.15	-5.47	17
20	P04181	OAT	2.6.1.13	-5.44	27
21	P17735	TAT	2.6.1.5	-5.44	14
22	O15270	SPTLC2 KIAA0526 LCB2	2.3.1.50	-5.43	78
23	P54687	BCAT1 BCT1 ECA39	2.6.1.42	-5.42	83
24	Q8IUZ5	PHYKPL AGXT2L2 PP9286	4.2.3.134	-5.37	17
25	Q9NUV7	SPTLC3 C20orf38 SPTLC2L	2.3.1.50	-5.35	82
26	P80404	ABAT GABAT	2.6.1.19; 2.6.1.22	-5.33	71
27	P23378	GLDC GCSP	1.4.4.2	-5.29	34
28	O75600	GCAT KBL	2.3.1.29	-5.28	27
29	Q96GA7	SDSL	4.3.1.19; 4.3.1.17	-5.27	31
30	P0DN79	CBSL	4.2.1.22	-5.26	16
31	P22557	ALAS2 ALASE ASB	2.3.1.37	-5.19	82
32	P11217	PYGM	2.4.1.1	-5.14	20
33	Q9GZT4	SRR	5.1.1.18; 4.3.1.18; 4.3.1.17	-5.05	15
34	P17174	GOT1	2.6.1.1; 2.6.1.3	-4.98	37
35	Q96EN8	MOCOS	2.8.1.9	-4.72	31
36	Q9Y600	CSAD CSD	4.1.1.29; 4.1.1.11	-4.70	28
37	Q4AC99	ACCSL	-	-4.67	9
38	Q6ZQY3	GADL1	4.1.1.11; 4.1.1.29	-4.56	27
39	P00505	GOT2	2.6.1.1; 2.6.1.7	-4.48	32
40	Q8IYQ7	THNSL1	-	-4.45	47
41	Q9BYV1	AGXT2 AGT2	2.6.1.44; 2.6.1.40	-4.29	17
42	Q8TD30	GPT2 AAT2 ALT2	2.6.1.2	-4.22	12
43	P19113	HDC	4.1.1.22	-4.17	29
44	Q8NHS2	GOT1L1	2.6.1.1	-4.13	16
45	O14977	AZIN1 OAZI OAZIN	-	-4.07	8
46	Q05329	GAD2 GAD65	4.1.1.15	-3.89	21
47	Q86YJ6	THNSL2	4.2.3.-	-3.88	55
48	O94903	PLPBP PROSC	-	-3.85	20
49	Q96I15	SCLY SCL	4.4.1.16	-3.84	21
50	P24298	GPT AAT1 GPT1	2.6.1.2	-3.62	53
51	Q16719	KYNU	3.7.1.3	-3.15	28
52	Q9Y697	NFS1 NIFS HUSSY-08	2.8.1.7	-3.11	8
53	Q8TBG4	ETNPPL AGXT2L1	4.2.3.2	-1.48	54

Table 3. *LCaaM ranking of HTML_PLP in Homo sapiens. The right column shows the number of conformations in the LC, that in our case corresponds to the percentage of the totals run in LC (100 runs).*

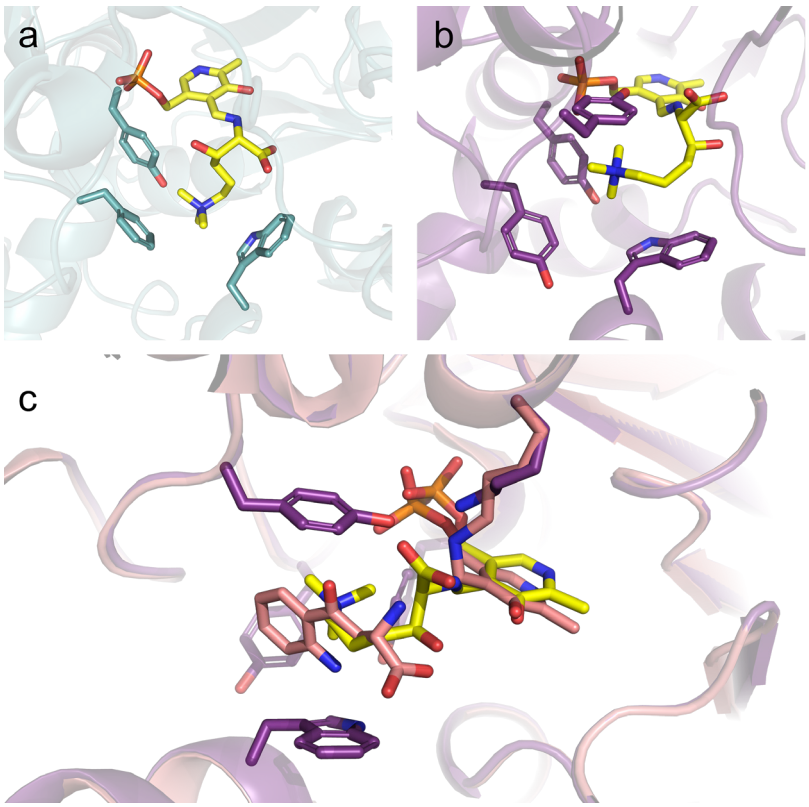


Figure 10. *HTML_PLP docked in human KYATs. (**a**) Lowest energy conformation of LC of HTML-PLP docked in KYAT1, (**b**) Lowest energy conformation of LC of HTML-PLP docked in KYAT3. Aromatic residues that make up the aromatic box are shown as sticks. (**c**) KYAT3 superimposed with structure PDB ID 3E2Z resolved in the presence of kynurenine (pink). HTML-PLP in yellow overlay the PLP cofactor as internal aldimine in the 3E2Z structure.*

By analyzing the active sites of these enzymes that have a high number of conformations in the LC (**Table 3,** "Num in LC" column), the HTML-PLP ligand is very well positioned, with interactions between the cofactor and the protein, mirroring the ones found in resolved structures in the presence of the substrate (**Figure 10c**). In the HTML side chain, there is the typical aromatic box with the cation-π interaction (**Figure 10a, b**), as discussed before (see **Figure 3**).

In 6th and 8th positions we also find the two isozymes, cytosolic and mitochondrial, of serine hydroxymethyltransferase (SHMT1 SHMT2), which are currently the main candidate genes for HTMLA thanks to their aldolase activity found against beta hydroxylated amino acids. In this case we do not find an aromatic box, but the ligand still has other interesting interactions, such as the H-bond between the OH group of the ligand and the nitrogen of the histidine, which is stacked with the ring of the PLP, and the salt bridge between the positive guanidino group of arginine residue and the negative carboxylic group of HTML (**Figure 11a, b**).

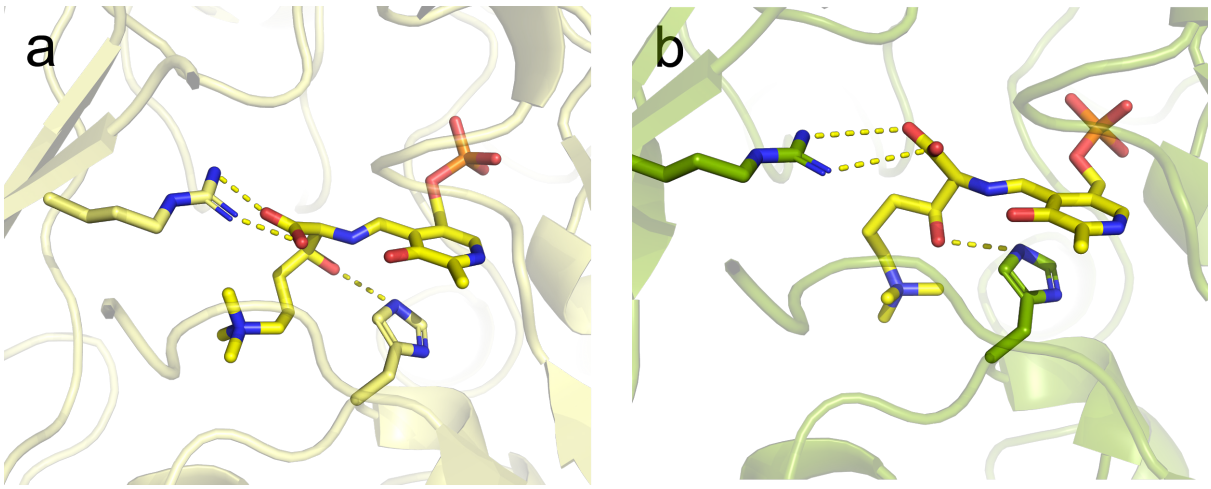


Figure 11. *HTML_PLP docked in human SHMTs. (**a**) Lowest energy conformation of LC of HTML-PLP docked in SHMT1, (**b**) Lowest energy conformation of LC of HTML-PLP docked in SHMT2. Arginine that interacts with the carboxylic group of HTML, and histidine that stacks with PLP, are shown as sticks.*

Finally, in 13th position we find another interesting candidate, sphingosine-1-phosphate lyase 1 (SGPL1). This lyase, which uses phosphorylated sphinganine as substrate, is actually a PLP-dependent aldolase that acts on a substrate that could be considered a structural analog of HTML, with the phosphate group instead of the carboxylic one, and a long aliphatic chain instead of a shorter one (**Figure 12**).
Even in this case it is noted how the histidine stacks with the PLP forms a H-bond with the OH group of HTML, while there are three aromatic residues in the vicinity of trimethylation, which could constitute a possible aromatic box (**Figure 13**).

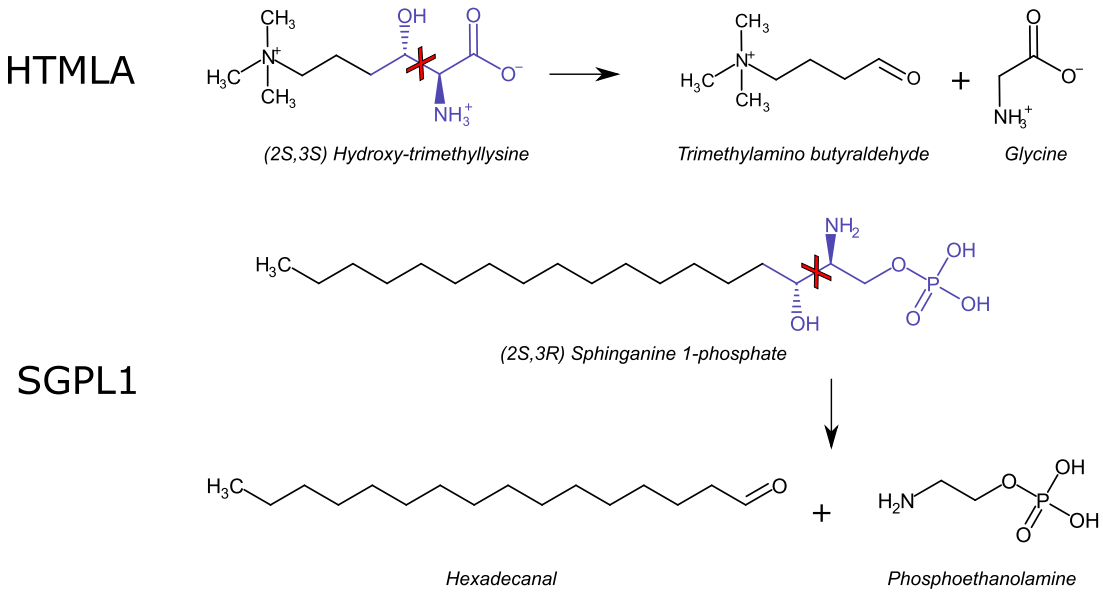


Figure 12. *Structural similarity of HTML aldolase and Sphinganine 1-phosphate lyase reactions. The hydroxyl group is in anti with primary amine bound by PLP. In blue are shaded the structural analogies of the 2 substrates, where the carboxyl group is replaced with phosphate ester in the SGPL1's substrate.*

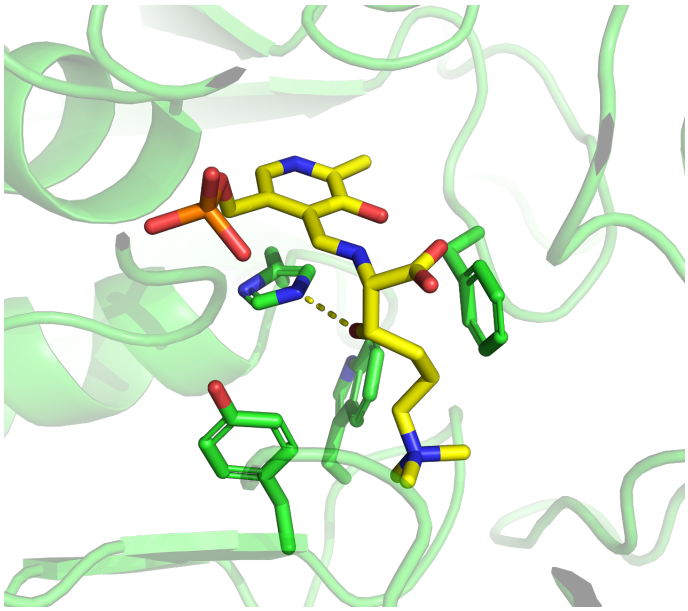


Figure 13. *HTML_PLP docked in human SGPL1. Lowest energy conformation of LC of HTML-PLP docked in SGPL1. Histidine that stacks with PLP and aromatic residues around HTML are shown as sticks.*

HTML-PLP screening in *Mus musculus* PLPome

	Entry	Gene names	EC number	LCaaM (kcal/mol)	Num in LC
1	Q71RI9	Kyat3 Ccbl2 Kat3	2.6.1.7; 4.4.1.13; 2.6.1.63	-7.40	27
2	Q8BWU8	Etnppl Agxt2l1	4.2.3.2	-6.95	23
3	P50431	Shmt1 Shmt	2.1.2.1	-6.62	78
4	Q9ET01	Pygl	2.4.1.1	-6.08	54
5	Q99K85	Psat1 Psa Psat	2.6.1.52	-6.00	8
6	P00860	Odc1 Odc	4.1.1.17	-5.95	30
7	Q7TSV6	Got1l1	2.6.1.1	-5.90	47
8	Q6XPS7	Tha1 Gly1	-	-5.85	40
9	Q8VC19	Alas1	2.3.1.37	-5.67	60
10	Q8BGT5	Gpt2 Aat2	2.6.1.2	-5.61	28
11	O35484	Azin1 Oazi Oazin	-	-5.60	38
12	P08680	Alas2	2.3.1.37	-5.60	67
13	O35423	Agxt Agxt1	2.6.1.51; 2.6.1.44	-5.59	24
14	Q9WVM8	Aadat Kat2	2.6.1.39; 2.6.1.7	-5.58	26
15	Q8VCN5	Cth	4.4.1.1	-5.56	21
16	Q8BTY1	Kyat1 Ccbl1 Kat	2.6.1.7; 4.4.1.13; 2.6.1.64	-5.52	85
17	Q9WUB3	Pygm	2.4.1.1	-5.44	38
18	Q3UX83	Accsl Gm1967	-	-5.43	28
19	Q80WP8	Gadl1	4.1.1.11; 4.1.1.29	-5.40	23
20	Q8R0X7	Sgpl1 Spl	4.1.2.27	-5.40	17
21	P61922	Abat Gabat	2.6.1.19; 2.6.1.22	-5.35	60
22	Q8VBT2	Sds	4.3.1.17; 4.3.1.19	-5.35	25
23	P29758	Oat	2.6.1.13	-5.34	27
24	Q8R238	Sdsl Sds	4.3.1.17; 4.3.1.19	-5.31	30
25	Q9CZN7	Shmt2	2.1.2.1	-5.21	17
26	Q6P6M7	Sepsecs D5Ertd135e	2.9.1.2	-5.20	17
27	P24288	Bcat1 Eca39	2.6.1.42	-5.18	27
28	Q8CI94	Pygb	2.4.1.1	-5.18	44
29	Q3UNZ2	Ldc1 Gm853	-	-5.11	55
30	P05202	Got2 Got-2	2.6.1.1; 2.6.1.7	-5.06	42
31	O35855	Bcat2 Bcatm Eca40	2.6.1.42	-5.03	47
32	Q91WT9	Cbs	4.2.1.22	-4.93	19
33	Q14CH1	Mocos	2.8.1.9	-4.93	31
34	P05201	Got1	2.6.1.1; 2.6.1.3	-4.89	51
35	O88986	Gcat Kbl	2.3.1.29	-4.79	54
36	A2AIG8	Accs	-	-4.79	14
37	Q9QZX7	Srr	5.1.1.18; 4.3.1.18; 4.3.1.17	-4.78	10
38	Q9DBE0	Csad	4.1.1.29; 4.1.1.11	-4.62	23
39	Q8BH55	Thnsl1	-	-4.60	46
40	Q3UEG6	Agxt2	2.6.1.44; 2.6.1.40	-4.59	51
41	Q8QZR1	Tat	2.6.1.5	-4.57	21
42	P48318	Gad1 Gad67	4.1.1.15	-4.56	30
43	Q8R1K4	Phykpl Agxt2l2	4.2.3.134	-4.53	16
44	Q9JLI6	Scly Scl	4.4.1.16	-4.53	13
45	Q80W22	Thnsl2	4.2.3.-	-4.31	11
46	P23738	Hdc	4.1.1.22	-4.26	33
47	O88533	Ddc	4.1.1.28	-4.22	12
48	P48320	Gad2 Gad65	4.1.1.15	-4.00	21
49	Q8BG54	Sptlc3 Sptlc2l	2.3.1.50	-3.88	36
50	Q91W43	Gldc	1.4.4.2	-3.77	15
51	Q9Z1J3	Nfs1 Nifs	2.8.1.7	-3.54	25
52	Q9CXF0	Kynu	3.7.1.3	-3.34	35
53	Q9Z2Y8	Plpbp Prosc	-	-3.18	27
54	Q8QZR5	Gpt Gpt1	2.6.1.2	-3.15	47
55	P97363	Sptlc2 Lcb2	2.3.1.50	-0.74	60

Table 4. LCaaM ranking of HTML_PLP in Mus musculus. *Table shows the position of each name of their best result with LCaaM ranking. The last column shows the number of conformations in the LC.*

Regarding the screening of HTML-PLP towards the *Mus musculus* PLPome, we obtained quite similar results compared to the screening with human genes (**Table 4**). In fact, the enzymes that were in the top positions in the human ranking are also found using the structures of the related murine orthologs. The cases we analyzed that were most interesting are Kyat3 in the first position, Shmt1 in the third and Tha1 in the eighth.

In this case, the paralogue Kyat1 gains the sixteenth position, while Shmt2, the paralogue of Shmt1, the twenty-fifth position. In the screening previously carried out with the PLP alone (see **Figure 8**), Shmt2 is placed in the penultimate position, confirming a condition of the receptor that penalizes docking with the PLP. On the contrary, Shmt1 is instead in the 4th position, despite having almost identical active site residues in terms of primary structure.

Tha1, on the other hand, is the mouse gene that codes for a putative threonine aldolase not experimentally characterized to date. Again, we find an aromatic box around HTML trimethylation (**Figure 14**), making it the best candidate for experimental validation

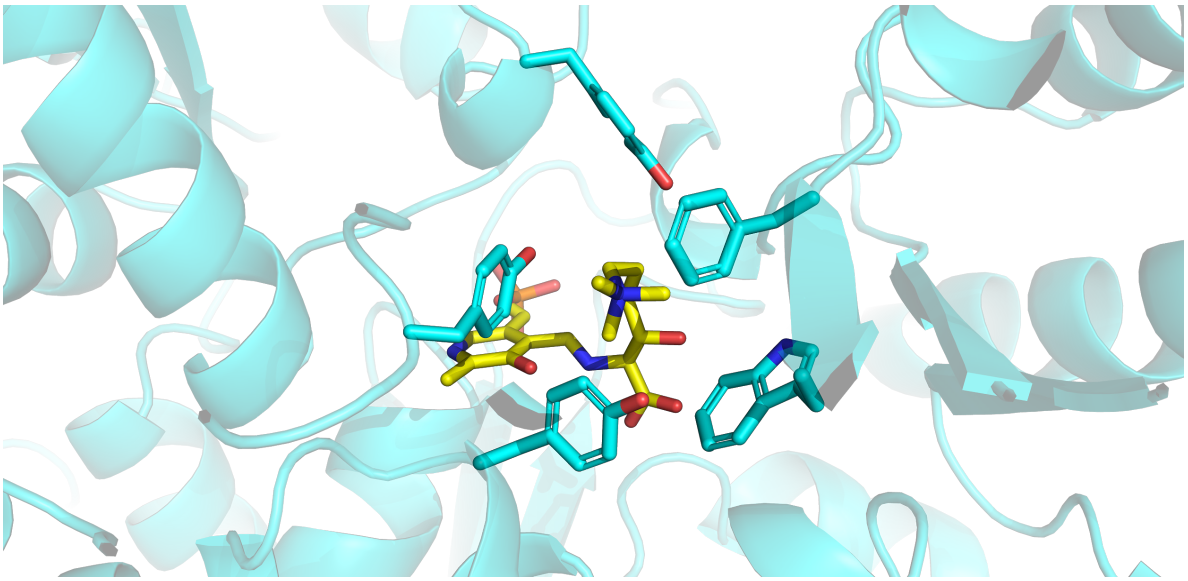


Figure 14. *HTML-PLP docked in mouse Tha1. Lowest energy conformation of LC of HTML-PLP docked in Tha1 model. Aromatic residues that make up the aromatic box are shown as sticks.*

Experimental validation of the mains candidates

We decided to produce the most promising candidate enzymes in recombinant form to tested HTMLA activity. Our main candidates were KYAT1, SHMT1, SHMT2, SGPL1 from *Homo sapiens* and Kyat3 and Tha1 from *Mus musculus*. Expression-ready clones for these enzymes were already present in our laboratory, except for SGPL1 and Tha1 which were purchased from Genscript.
KYAT1, SGPL1, Kyat3, and Tha1 were produced and tested at the University of Parma, while SHMT1 and SHMT2 were produced and tested by our coworkers at the Sapienza University of Rome. Preliminary tests on the HTMLA activity of SHMT1 and SHMT2 were very encouraging, but the studies are still in progress.

Synthesis of HTML

We decided to synthesize enzymatically HTML from TML, since unfortunately it's not commercially available. First, TML was obtained by chemical synthesis from Boc-Lys-OH with a protocol applied by Beatrice Cogliati and Andrea Secchi from University of Parma. About 100 mg of TML was obtained, which purity was verified by ^{1}H and ^{13}C NMR. TML was used as substrate for the second enzymatical step, exploiting the reaction of the first biosynthetic enzyme (TMLD) to produce HTML. We produced the recombinant TMLD enzyme that was able, in the presence of the substrates and cofactors (see **Methods**), to convert efficiently ~90% of TML into HTML. The reaction mixture was finally purified with a cation exchange chromatography, obtaining approximately 60 mg of HTML at the end of the entire procedure, which purity was verified by ^{1}H NMR.

SGPL1

Homo sapiens isoform X2 of SGPL1 (XP_006718116.1), which lacks the N-terminal binding domain to the endoplasmic reticulum (**Figure 15a**), was produced in a recombinant form in *E. coli*. Previously, the activity of SGPL1 was characterized by eliminating the first 58 amino acids of the main isoform NP_003892.2 [23]; on the other hand, in our construct, the first 80 amino acids are missing, which in any case do not constitute the catalytic domain (**Figure 15a**). The protein was obtained with the expected weight of 53 kDa (**Figure 15b**) in a low yield, but enough to be tested through NMR. No aldolase activity was observed against HTML, although we have not tested our protein with its known substrate due to the unavailability of sphinganine 1-phosphate in our laboratory.

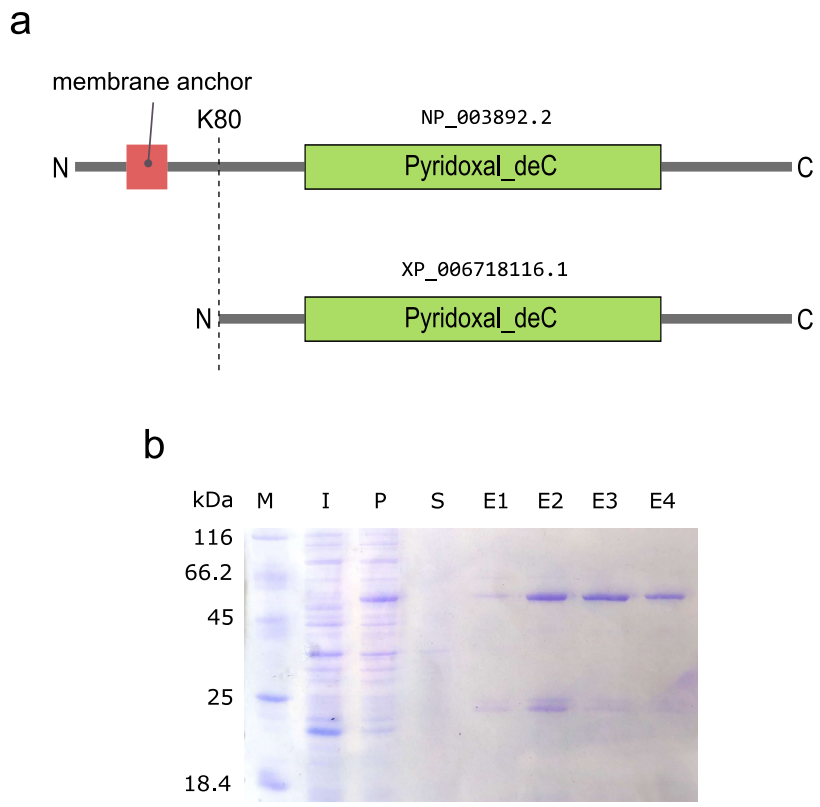


Figure 15. *(**a**) SGPL1 domain composition according to PFAM of the 2 different isoforms (NP_003892.2, **upper**; XP_006718116.1, **lower**). The shorter construct was used for recombinant expression, missing the firsts 80 residues as shown by the dashed line. (**b**) SDS-PAGE of recombinant SGPL1 purification: peqGOLD Marker (M), Induced (I), Pellet (P), Supernatant (S), Elutions (E1, E2, E3, E4).*

KYAT1, Kyat3

The KYAT1 enzyme has a characteristic absorption spectrum of PLP-dependent enzymes, showing a peak at 430 nm (**Figure 16a**), given by the presence of internal aldimine. Through an NMR spectrometry, the absence of activity towards the HTML substrate by both enzymes was observed.

We also wanted to investigate whether KYAT1 could bind HTML without degrading it, looking at the enzymatic activity on DL-kynurenine (**Figure 16c**). In the presence of HTML, KYAT1's activity towards DL-kynurenine decrease (**Figure 16e**) compared to that in the absence of HTML (**Figure 16d**).

We produced even the recombinant form of Kyat3 enzyme of *Mus musculus* and no aldolase activity was observed against HTML assessed by ^{1}H NMR.

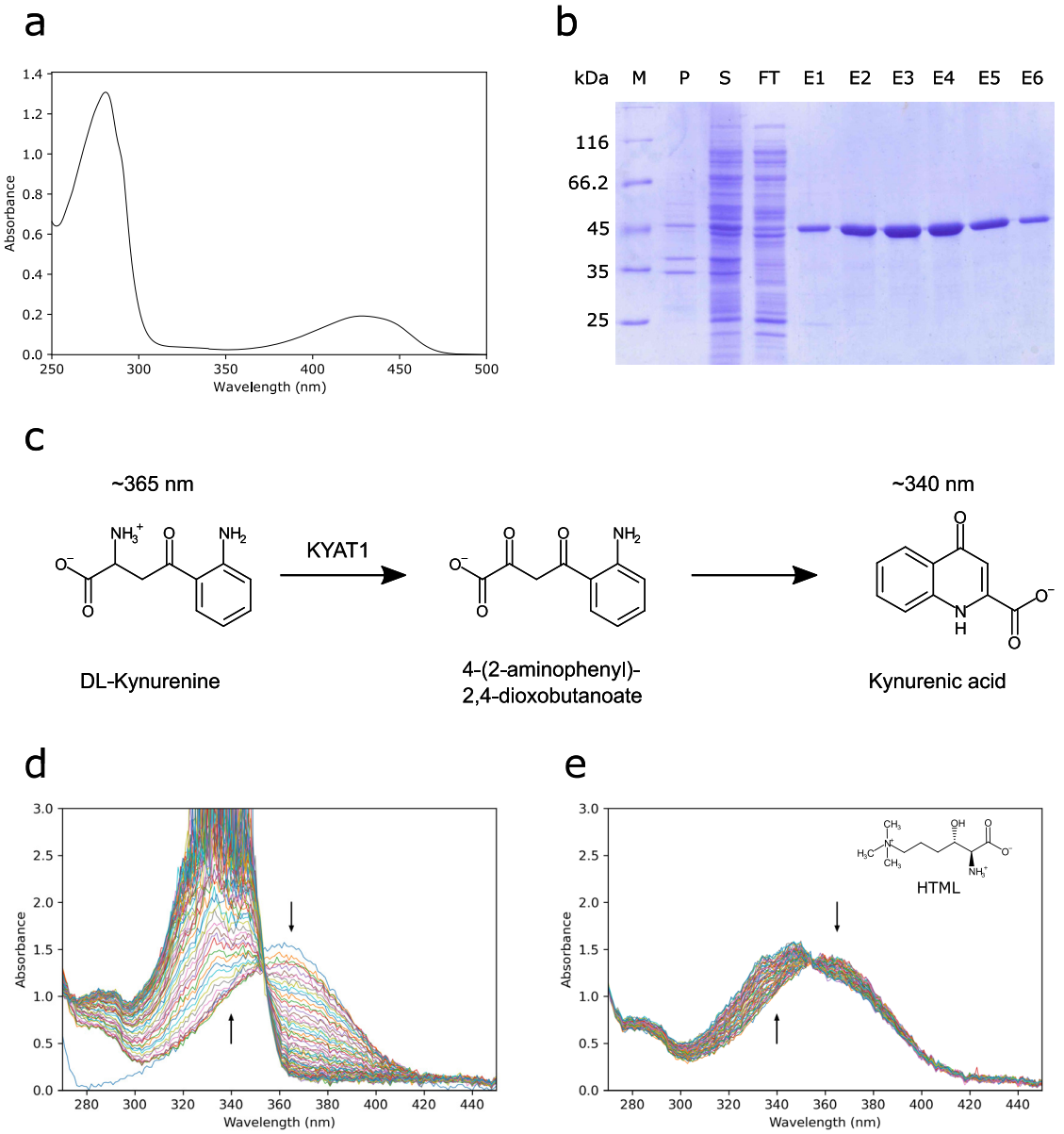


Figure 16. *(**a**) Absorbance spectrum of recombinant KYAT1 (21.3 µM) in NaH_2PO_4 (20 mM), pH 7.0. (**b**) SDS-PAGE of recombinant KYAT1 purification: peqGOLD Marker (M), Pellet (P), Supernatant (S), Flow-through (FT), Elutions (E1, E2, E3, E4, E5, E6). (**c**) Simplified scheme of the kynurenine-glyoxylate aminotransferase reaction of KYAT1 showing formation of kynurenic acid through spontaneous cyclization of an alpha-keto acid intermediate. Spectrophotometric kinetics showing a decrease of the DL-kynurenine (0.3 mM) signal at ~360 nm, and the corresponding increase of the kynurenic acid signal at ~340 nm. Spectra were collected every 30 seconds for 30 minutes, in the absence (**d**) or in the presence (**e**) of 0.5 mM HTML.*

Tha1

The recombinant expression of Tha1 was more complicated, and it was necessary to co-express the construct with GroEL and GroES chaperones (see **Methods**) to obtain a soluble (**Figure 17b**), albeit very unstable, product. In fact, the protein, just eluted from the nickel column, begins to precipitate even at low concentrations. We produced Tha1 protein without the predicted mitochondrial signal, truncating the sequence at the V40 residue (**Figure 17a**), according to the comparison with homolog bacterial genes that lack the N-terminal signal [12].

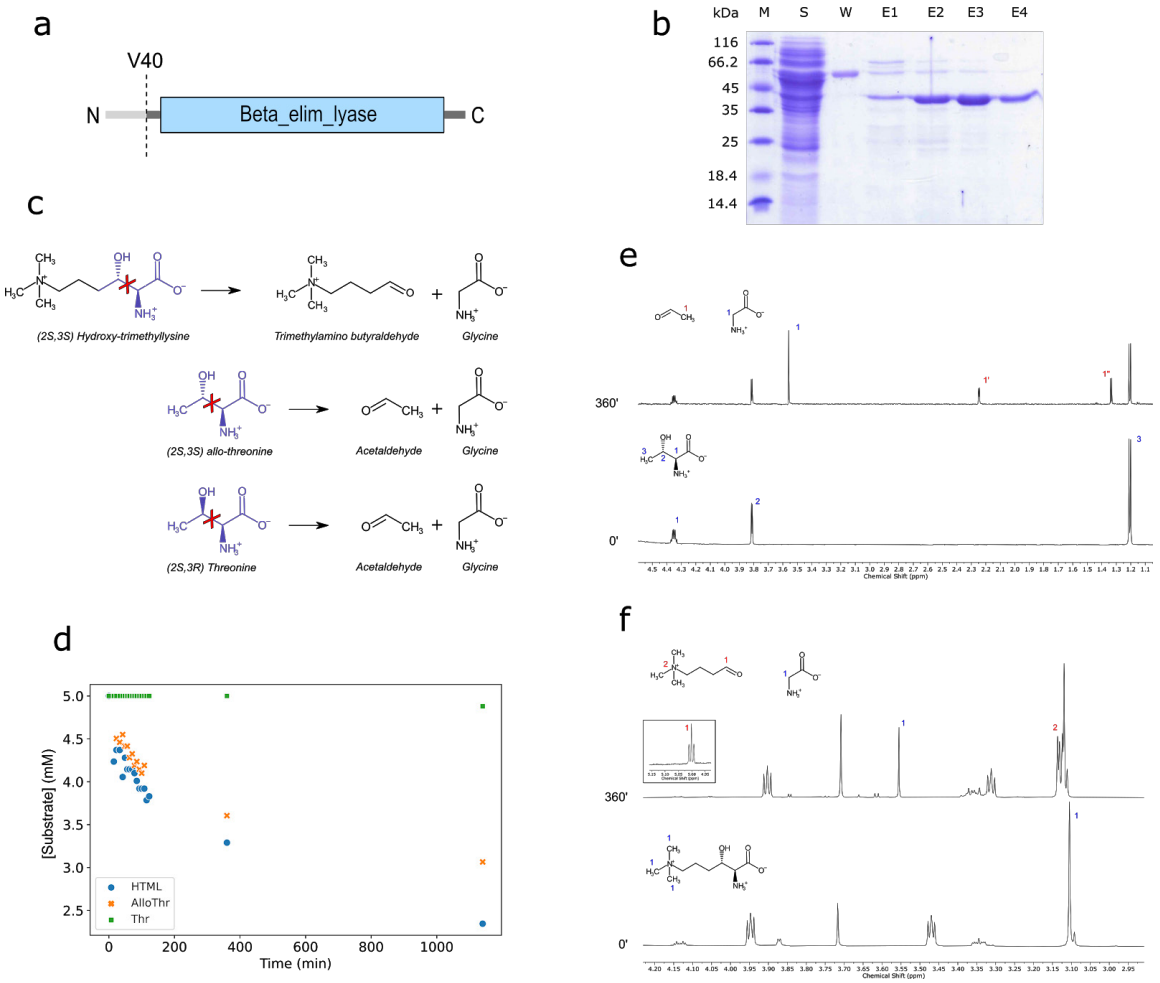


Figure 17. *(**a**) Tha1 domain composition according to PFAM. The dashed line indicates gene truncation for recombinant protein expression. (**b**) SDS-PAGE of recombinant Tha1 purification: peqGOLD Marker (M), Supernatant (S), Washing (W), Elutions (E1, E2, E3, E4). (**c**) Scheme of the aldol cleavage reactions tested to verify the activity of Tha1. In blue are shaded the structures common to the three substrates. (**d**) Decay plot of Tha1 activity towards substrates; each point shows the integral of ^{1}H NMR kinetic of the corresponding substrate. (**e**) ^{1}H NMR spectrum of Tha1 activity in the presence of 5 mM L-threonine at 0' and 360'. (**f**) ^{1}H NMR spectrum of Tha1 activity in the presence of 5 mM HTML at 0' and 360'.*

Nonetheless, we managed to get enough soluble protein to verify its activity. We tested the threonine aldolase enzyme with three different substrates: L-threonine, L-*allo*-threonine and HTML. In fact, *allo*-threonine and HTML share the same absolute configuration (2S,3S) contrary to the L-threonine that has (2S,2R) (**Figure 17c**).

The different diastereomers also demonstrated different behaviors in the Tha1 reaction as assessed by ^{1}H NMR. In the presence of the enzyme Tha1 we observed a negligible consumption of the threonine substrate and a similar consumption of *allo*-threonine and HTML (**Figure 17d**). The aldolase activity was demonstrated, as shown in the time-resolved NMR spectra, with the production of appreciable amounts of glycine (**Figure 17e, f**; signals at 3.6 ppm) and the corresponding aldehydes: acetaldehyde with *allo*-threonine (**Figure 17e**) and TMABA with HTML. (**Figure 17f**).

Allo-threonine and HTML are not completely consumed, and the reaction seems to reach equilibrium at about half the initial concentration of the substrates.

Discussion

To our knowledge, this is the first time that a reverse docking technique has been applied to the molecular identification of an enzyme responsible for an unassigned reaction of a metabolic pathway (pathway hole). Possible limitations of this technique are the availability of 3D structures and the identification of the active site region in the protein structure. Both limitations can be overcome in the case of mammalian PLP-dependent enzymes, as structural models of sufficient quality are available for the human and mouse PLPomes and the active site region can be readily defined by the localization of the catalytic lysine.

We validated our procedure through a screening of different compounds used as positive controls with the PLP cofactor linked to the true substrate (see **Figure 6**). The results, however, were found to be biased by the different binding energies of the enzymes for the cofactor moiety of the external aldimine (see **Figure 7, Table 2**). By considering the energies of the amino acid substrate atoms alone in the docking results, we were able to improve the ranking of the control enzymes. The best ranking criterion, LCaaM, allowed us to find our control enzymes in the top 10 positions with a probability of 50% (see **Figure 9**). By applying the procedure to the dataset of enzymes represented by human and mouse PLPomes, we were able to recover some valid candidates for the HTMLA reaction, which we produced in recombinant form. From our results, two main conclusions can be drawn:

- our procedure allowed us to find in the first 10 positions of both rankings (see **Tables 3,4**), three genes (SHMT1, SHMT2, Tha1) that have been previously implicated in the catalysis of the retro-aldol cleavage step of the carnitine pathway.
- KYAT1, one of the top ten candidates, although unable to catalyze the HTML retro aldol cleavage, was inhibited by HTML (see **Figure 16c-e**), confirming the already established effectiveness of reverse docking in predicting the binding of a molecule in the active site [15].

Preliminary studies of activity on human SHMTs reveal aldolase activity in the presence of HTML (R. Contestabile, personal communication). Further studies will be needed to confirm the involvement of the identified enzymes in the carnitine pathway in mice and humans.

Our working hypothesis is that the gene that codes for HTMLA is Tha1 in *Mus musculus* and the other mammals that possess orthologous genes, since HTML was the preferred substrate among those tested (see **Figure 17**). Humans and other mammals that have lost the Tha1 gene could exploit the secondary activities of SHMT1 and SHMT2 to produce TMABA. Although the evidence of a PLP-dependent enzyme acting as HTMLA has been provided in the rat, it is very likely that it is also in humans, given that all the other aldolases that act on beta hydroxylated amino acids require this cofactor. When exogenous TML is administered to the rat, it is

completely and readily absorbed [24]. In humans, dietary TML is less effective, than γ-butyrobetaine, in increasing the production of carnitine, suggesting a factor limiting in the first part of the biosynthetic pathway. [25]. Both activities of threonine and *allo*-threonine aldolase were observed in rats [26], however, carried out by 2 different enzymes. Based on our results, the enzyme capable of cleaving *allo*-threonine also in rats could be encoded by Tha1. Given the unclear physiological significance of this reaction in mammals, the primary reaction is more likely the one in the carnitine biosynthetic pathway.

Methods

Rev_Docking GitHub repository

All the scripts written and used in this work are collected in the Rev_Docking GitHub repository (https://github.com/Percud/Rev_Docking/), as well as the receptors set, ligand files and other example files.

Establishment of the Human and Mouse dataset receptors

The structures for reverse-docking were downloaded with *get_models_coord.py* by SWISS-MODEL repositories of *Homo sapiens* and *Mus musculus*, using the Uniprot accession number obtained with *convert_ac* function in *revdock.py* module (see below) by the NCBI accessions present in B6DB (http://bioinformatics.unipr.it/B6db/tmp/), corresponding only to PLP-dependent enzymes (PLPomes).
During the procedure, all the structures were deprived of the modified LLP residue corresponding to the internal aldimine where it presents, replacing with LYS name all the atoms relating to the lysine and eliminating the atoms relating to the PLP cofactor.
Any heteroatoms present in the models have been excluded in the procedure by commenting with a hash symbol (#) lines starting with HETATM in the pdb file, in the receptor's preparation by the automatic procedure *rev_docking_p.py*.

Ligands preparation

The ligands used for the validation of the procedure have been selected in order to use amino acids with different properties (negative, positive, hydrophobic, aromatic, etc..). The amino acids were ligated with a covalent bond between their N-alpha groups and the aldehydic group of PLP in the external aldimine state. For ORN-PLP we use the bond with the N-delta group, as external aldimine for ornithine aminotransferase.
All the compounds used in the reverse docking screening were constructed manually with Avogadro, according to their chemical composition, followed by a steepest descent algorithm energy minimization using Ghemical as force field.
The external aldimine molecules were given in input to Chimera to assign charges with Gasteiger using AMBERff14SB as force field and saved in mol2 format.
Finally, the ligands were saved as pdbqt format in Autodocktools, keeping all the rotable torsions.

Coordinates file for the Autodock grid center

The coordinates for grid positioning have been obtained by the *get_models_coord.py* script which i) retrieves the position of the lysine indicated as PTM in the Uniprot database; ii) determines the residue number relative to the lysine in the pdb file through a pairwise alignment of the Uniprot and the PDB sequence converted in FASTA format with *pdb2fasta* (see below), and iii) retrieves the coordinates of the lysine NZ atom, to be used as the center of the grid. The script output is a tab-separated file containing for each pdb file, the coordinates to be used in the GPF.

All the functions called in *get_models_coord.py* were enclosed in the *revdock.py* module:

- *convert_ac:* converts a list of accession numbers, e.g. NCBI protein RefSeq, into accession numbers used as model identifiers, e.g. Uniprot accessions.
- *get_models:* retrieves all the models in the SWISS-MODEL repository based on a list of Uniprot accession numbers, an organism TaxID, and the output directory.
- *pdb2fasta:* writes a FASTA file for each protein chain of a PDB file, inserting an X for the missing residues of the PDB structure.
- *get_features:* given a Uniprot ID, returns a complete dataframe of the records. We exploit it to retrieve the position of the catalytic lysine.
- *match_fasta_position:* determines the residue position of a subject FASTA file based on the number of a query FASTA file through a blastp pairwise alignment. We use it to match residue numbers in the PDB from Uniprot sequences.

Reverse-docking procedure

The reverse docking procedure contained in the *rev_docking_p.py* script requires a ligand in pdbqt format, a coordinates file with a list of xyz coordinates for each structure where the grid will be positioned, and a dataset of receptors in PDB format.

This procedure completes all the steps for the preparation of the receptors, parameter files, grid maps and the docking algorithm, with the help of python scripts provided in the *Utilities24* folder (*MGLTools-1.5.6/MGLToolsPckgs/AutoDockTools/Utilities24/*):

- *prepare_receptor4.py*: adds hydrogens and charge to the protein structure and saves the pdbqt file for all active sites defined in the coordinates file.
- *prepare_gpf4.py:* prepares grid parameters file by adding the coordinates read from the coordinates file.
- *prepare_dpf4.py:* prepares the docking parameter using a reference file with the algorithm parameters.

Once the pdbqt, gpf and dpf files are obtained for each active site of each structure, the procedure runs the Autogrid and Autodock program in multiprocessing mode.
All the fixed parameters for the procedure, such as, for example, the grid size in the gpf or the number of runs and energy evaluations in the dpf, are written in a configuration file (*rd_conf.txt*) where the paths of the input and output files as well as the number of parallel processes are also specified:

```
[DOCKING]
MGL_ROOT = ./MGLTools-1.5.6   # MGL_ROOT environmental variable - path to mgltools
ligand = ./Docking_data/HTML_PLP.pdbqt
receptor_dir = ./Mouse/                 # path to receptor pdb files
coord_file = ./Mouse/Mouse_coord.csv    # table of grid center: receptor chain resnum xyz
outdir = ./Mouse/outdir_HTML_PLP
processes = 200                         # number of processes of multiprocessing

[GPF]
npts = 55,55,55

[DPF]
ga_pop_size = 300
ga_num_evals = 25000000
ga_run = 100
```

For our procedure, we changed run to 100, num_evals to 25,000,000 and pop_size to 300, since we had ligands with more than 10 torsions. The procedure has been carried out in the SkyLake node (4 INTEL XEON E5-6140 2.3GHz, 72 cores, and 384 Gb of RAM) of the HPC facility of the University of Parma. A typical run took about 24 hours for the mouse PLPome and 60 hours for the human one. The running time depends on the degrees of freedom of the ligand and the docking parameters such as run and num evals.

Results analysis

For the analysis of the results, we have written a python script that parses the energy values of the Autodock output file in dlg format. The script *get_atom_energy.py* returns a tab-separated data frame with all the energy considered in ranking criteria, the corresponding run, and information about the number of conformations in the largest and second largest clusters. We have used this script to analyze the different ranking criteria considered for each ligand screening. To obtain the energy for *aa* criteria, we sum the non-binding energy of the only

amino acid, which were previously marked manually with a different three letter code with respect to PLP atoms during the ligand's preparation (see above).
All the graphs reported in this section have been obtained using python codes with the use of the pandas, matplotlib and seaborn modules. The computer codes are contained in the Jupyter notebook called *rev_docking_result.ipynb* provided at the Rev_Docking repository.

HTML synthesis

HTML was synthesized starting from TML through enzymatic conversion to the hydroxylated form. In turn, TML was obtained by chemical synthesis from (2S)-6-amino-2-{[(tert-butoxy)carbonyl]amino}hexanoic acid (purchased from FCH) conducted by Beatrice Cogliati and Andrea Secchi (University of Parma) using a previously described protocol [27]. For the enzymatic HTML synthesis, we produced recombinantly the TMLD enzyme following the protocol of Kazaks et al. [28], who kindly provided us with the corresponding clone. We prepared the reaction mixture according to a modified protocol [29] using triethanolamine instead of the phosphate buffer, which in the presence of Fe^{2+} ion, immediately precipitates. We prepared the following protocol for 30 min in a flask agitated at 37°C in a final volume of 100 mL:

α-keto glutarate	15 mM
ascorbate	5 mM
TML	5 mM
$FeSO_4$	200 μM
TEA	20 mM
DTT	1 mM
TMLD	10 μM

We finally purified the reaction mixture, that contained the enzyme and the other molecules, with cation exchange chromatography, by exploiting the positively charged N-trimethyl group to isolate HTML from the other negatively charged molecules such as ascorbate, 2-oxoglutarate, and succinate. After reaching pH 5.0 with the addition of HCl, the solution was firstly deprived by the enzyme TMLD through a Vivaspin™ centrifugation, then the flow-through was loaded into a 50 mL Superloop of AKTA pure system FPLC and purified using HiTrap 5 mL SP column. We used 0.2 M HCl to elute the molecule with a gradient of 7 CV. We followed the elution on 210 nm and the fractions of the corresponding peak and flow-through of the column were analyzed by ^{1}H NMR spectra using the setting described below.

Plasmid construction

For the construction of *Hs*SGLP1 expression plasmid, the SGPL1 (NCBI GeneID: 8879) CDS sequence (XM_006718053.1) inserted into pET-28b vector was purchased from GenScript (USA Inc.). For the construction of Tha1 expression plasmid, the Tha1 (NCBI GeneID: 71776) CDS sequence (NM_027919.4) without the first 40 amino acids corresponding a predicted mitochondrial signal, inserted into pET-28b expression vector was purchased from GenScript (USA Inc.) The constructs were transformed by electroporation into *E. coli* BL21 with pGRO7 plasmid from Takara™ containing GroEL and GroES chaperonins. The authenticity of all constructs was verified by sequence analysis.
The KYAT1 and Kyat3 clones were kindly given by Alessio Peracchi, of the University of Parma.

Protein expression and purification

We have used the same protocol purification for all the PLP-dependent enzymes produced in this work. Protein expression was performed by inoculating a single colony of every clone in a Liter of autoinducing LB broth obtained by adding 0,5 g/L glucose and 2 g/L lactose to standard LB medium. Cells were grown at 20°C for 16h after a pre-induction phase at 37°C for 8h. Cell pellets were resuspended in 50 mL of Lysis Buffer (NaH_2PO_4 50 mM pH 7.4, NaCl 150 mM, 20 µM PLP), sonicated (1s on/off alternatively at 40 W for 30 min) and centrifuged (14000 rpm for 40 minutes). Supernatant was loaded into a 50 mL Superloop of AKTA pure system FPLC and purified by Affinity Chromatography (AC) using HisTrap 5 mL FF column. Proteins were washed with 4 CV of Washing Buffer (NaH_2PO_4 50 mM pH 7.4, NaCl 150 mM). For Tha1 and SGPL1, we used an additional Washing Buffer (NaH_2PO_4 50 mM pH 7.4, KCl 100 mM, glycerol 10%, sucrose 500 mM, $MgCl_2$ 20 mM, ATP 5 mM, DTT 1 mM) to rid of GroEL which would otherwise be found in the elutions (see lane W in **Figure 17b**). Proteins are eluted with AC Elution Buffer (NaH_2PO_4 20 mM pH 7.4, NaCl 150 mM, 500 mM imidazole). Protein fractions were collected and concentrated by Vivaspin™ centrifugation for dialysis in a Storage Buffer (NaH_2PO_4 50 mM pH 7.4, NaCl 150 mM, DTT 1 mM, glycerol 15%).

NMR spectroscopy

^1H NMR spectra were acquired with a JEOL ECZ600R spectrometer in no spinning mode at 25°C. Samples were loaded in Wilmad ECONOMY NMR tubes, solved in 600 µL of H_2O:D_2O (9:1) with simple DANTE presat sequence for H_2O suppression.

Bibliography

1 Houten, S.M. and Wanders, R.J.A. (2010) A general introduction to the biochemistry of mitochondrial fatty acid β-oxidation. *J. Inherit. Metab. Dis.* 33, 469–477

2 Rebouche, C.J. and Seim, H. (1998) Carnitine metabolism and its regulation in microorganisms and mammals. *Annu. Rev. Nutr.* 18, 39–61

3 Panter, R.A. and Mudd, J.B. (1969) Carnitine levels in some higher plants. *FEBS Lett.* 5, 169–170

4 Strijbis, K. *et al.* (2009) Identification and characterization of a complete carnitine biosynthesis pathway in Candida albicans. *FASEB J.* 23, 2349–2359

5 Tanphaichitr, V. *et al.* (1971) Lysine, a precursor of carnitine in the rat. *J. Biol. Chem.* 246, 6364–6366

6 HIJSZAR, G. Tissue-specific Biosynthesis of z-IV-Monomethyllysine and &-N-Trimethyllysine in Skeletal and Cardiac Muscle Myosin: A Model for the Cell-free Study of Post-translational Amino Acid Modifications in Proteins.

7 Hulse, J.D. *et al.* (1978) Carnitine biosynthesis. beta-Hydroxylation of trimethyllysine by an alpha-ketoglutarate-dependent mitochondrial dioxygenase. *J. Biol. Chem.* 253, 1654–1659

8 Dunn, W.A. *et al.* (1982) The effects of 1-amino-D-proline on the production of carnitine from exogenous protein-bound trimethyllysine by the perfused rat liver. *J. Biol. Chem.* 257, 7948–7951

9 Rebouche, C.J. and Engel, A.G. (1980) Tissue distribution of carnitine biosynthetic enzymes in man. *Biochimica et Biophysica Acta (BBA) - General Subjects* 630, 22–29

10 Hulse, J.D. and Henderson, L.M. (1980) Carnitine biosynthesis. Purification of 4-N'-trimethylaminobutyraldehyde dehydrogenase from beef liver. *J. Biol. Chem.* 255, 1146–1151

11 Englard, S. *et al.* (1985) Gamma-butyrobetaine hydroxylase: stereochemical course of the hydroxylation reaction. *Biochemistry* 24, 1110–1116

12 Edgar, A.J. (2005) Mice have a transcribed L-threonine aldolase/GLY1 gene, but the human GLY1 gene is a non-processed pseudogene. *BMC Genomics* 6, 32

13 Ogawa, H. and Fujioka, M. (1981) Purification and characterization of cytosolic and mitochondrial serine hydroxymethyltransferases from rat liver. *J. Biochem.* 90, 381–390

14 Schirch, L. and Peterson, D. (1980) Purification and properties of mitochondrial serine hydroxymethyltransferase. *J. Biol. Chem.* 255, 7801–7806

15 Kharkar, P.S. *et al.* (2014) Reverse docking: a powerful tool for drug repositioning and drug rescue. *Future Med. Chem.* 6, 333–342

16 Hermann, J.C. *et al.* (2006) Predicting substrates by docking high-energy intermediates to enzyme structures. *J. Am. Chem. Soc.* 128, 15882–15891

17 Morris, G.M. *et al.* (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791

18 Morris, G.M. *et al.* (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* 19, 1639–1662

19 Bienert, S. *et al.* (2017) The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* 45, D313–D319

20 Guterres, H. and Im, W. (2020) Improving Protein-Ligand Docking Results with High-Throughput Molecular Dynamics Simulations. *J. Chem. Inf. Model.* 60, 2189–2198

21 Nagy, G.N. *et al.* (2014) Composite aromatic boxes for enzymatic transformations of quaternary ammonium substrates. *Angew. Chem. Int. Ed. Engl.* 53, 13471–13476

22 Tagaya, M. and Fukui, T. (1984) Catalytic reaction of glycogen phosphorylase reconstituted with a coenzyme-substrate conjugate. *J. Biol. Chem.* 259, 4860–4865

23 Van Veldhoven, P.P. *et al.* (2000) Human sphingosine-1-phosphate lyase: cDNA cloning, functional expression studies and mapping to chromosome 10q22(1). *Biochim. Biophys. Acta* 1487, 128–134

24 Davis, A.T. and Hoppel, C.L. (1986) Effect of starvation on the disposition of free and peptide-linked trimethyllysine in the rat. *J. Nutr.* 116, 760–767

25 Rebouche, C.J. *et al.* (1989) Utilization of dietary precursors for carnitine synthesis in human adults. *J. Nutr.* 119, 1907–1913

26 Karasek, M.A. and Greenberg, D.M. (1957) Studies on the properties of threonine aldolases. *J. Biol. Chem.* 227, 191–205

27 Baba, R. *et al.* (2012) Development of a fluorogenic probe with a transesterification switch for detection of histone deacetylase activity. *J. Am. Chem. Soc.* 134, 14310–14313

28 Kazaks, A. *et al.* (2014) Expression and purification of active, stabilized trimethyllysine hydroxylase. *Protein Expr. Purif.* 104, 1–6

29 Wang, Y. *et al.* (2019) Investigating the active site of human trimethyllysine hydroxylase. *Biochem. J.* 476, 1109–1119

Appendix

	Entry	Gene names	Protein names	EC number	lys
1	Q8N5Z0	AADAT KAT2	Kynurenine/alpha-aminoadipate aminotransferase, mitochondrial	2.6.1.39; 2.6.1.7	263
2	P80404	ABAT GABAT	4-aminobutyrate aminotransferase, mitochondrial	2.6.1.19; 2.6.1.22	357
3	Q96QU6	ACCS PHACS	1-aminocyclopropane-1-carboxylate synthase-like protein 1	-	323
4	Q4AC99	ACCSL	Probable inactive 1-aminocyclopropane-1-carboxylate synthase-like protein 2	-	395
5	P21549	AGXT AGT1 SPAT	Serine--pyruvate aminotransferase	2.6.1.51; 2.6.1.44	209
6	Q9BYV1	AGXT2 AGT2	Alanine--glyoxylate aminotransferase 2, mitochondrial	2.6.1.44; 2.6.1.40	350
7	P13196	ALAS1 ALAS3 ALASH OK/SW-cl.121	5-aminolevulinate synthase, nonspecific, mitochondrial	2.3.1.37	445
8	P22557	ALAS2 ALASE ASB	5-aminolevulinate synthase, erythroid-specific, mitochondrial	2.3.1.37	391
9	O14977	AZIN1 OAZI OAZIN	Antizyme inhibitor 1	-	-
10	Q96A70	AZIN2 ADC KIAA1945 ODCP	Antizyme inhibitor 2	-	-
11	P54687	BCAT1 BCT1 ECA39	Branched-chain-amino-acid aminotransferase, cytosolic	2.6.1.42	222
12	O15382	BCAT2 BCATM BCT2 ECA40	Branched-chain-amino-acid aminotransferase, mitochondrial	2.6.1.42	229
13	P35520	CBS	Cystathionine beta-synthase	4.2.1.22	119
14	P0DN79	CBSL	Cystathionine beta-synthase-like protein	4.2.1.22	119
15	Q9Y600	CSAD CSD	Cysteine sulfinic acid decarboxylase	4.1.1.29; 4.1.1.11	305
16	P32929	CTH	Cystathionine gamma-lyase	4.4.1.1	212
17	P20711	DDC AADC	Aromatic-L-amino-acid decarboxylase	4.1.1.28	303
18	Q8TBG4	ETNPPL AGXT2L1	Ethanolamine-phosphate phospho-lyase	4.2.3.2	278
19	Q99259	GAD1 GAD GAD67	Glutamate decarboxylase 1	4.1.1.15	405
20	Q05329	GAD2 GAD65	Glutamate decarboxylase 2	4.1.1.15	396
21	Q6ZQY3	GADL1	Acidic amino acid decarboxylase GADL1	4.1.1.11; 4.1.1.29	333
22	O75600	GCAT KBL	2-amino-3-ketobutyrate coenzyme A ligase, mitochondrial	2.3.1.29	265
23	P23378	GLDC GCSP	Glycine dehydrogenase	1.4.4.2	754
24	P17174	GOT1	Aspartate aminotransferase, cytoplasmic	2.6.1.1; 2.6.1.3	259
25	Q8NHS2	GOT1L1	Putative aspartate aminotransferase, cytoplasmic 2	2.6.1.1	249
26	P00505	GOT2	Aspartate aminotransferase, mitochondrial	2.6.1.1; 2.6.1.7	279
27	P24298	GPT AAT1 GPT1	Alanine aminotransferase 1	2.6.1.2	314
28	Q8TD30	GPT2 AAT2 ALT2	Alanine aminotransferase 2	2.6.1.2	341
29	P19113	HDC	Histidine decarboxylase	4.1.1.22	305
30	Q16773	KYAT1 CCBL1	Kynurenine--oxoglutarate transaminase 1	2.6.1.7; 4.4.1.13; 2.6.1.64	247
31	Q6YP21	KYAT3 CCBL2 KAT3	Kynurenine--oxoglutarate transaminase 3	2.6.1.7; 4.4.1.13; 2.6.1.63	280
32	Q16719	KYNU	Kynureninase	3.7.1.3	276
33	Q9Y697	NFS1 NIFS HUSSY-08	Cysteine desulfurase, mitochondrial	2.8.1.7	258
34	P04181	OAT	Ornithine aminotransferase, mitochondrial	2.6.1.13	292
35	P11926	ODC1	Ornithine decarboxylase	4.1.1.17	69
36	Q6P996	PDXDC1 KIAA0251	Pyridoxal-dependent decarboxylase domain-containing protein 1	4.1.1.-	-
37	Q8IUZ5	PHYKPL AGXT2L2 PP9286	5-phosphohydroxy-L-lysine phospho-lyase	4.2.3.134	278
38	O94903	PLPBP PROSC	Pyridoxal phosphate homeostasis protein	-	47
39	Q9Y617	PSAT1 PSA	Phosphoserine aminotransferase	2.6.1.52	200
40	P11216	PYGB	Glycogen phosphorylase, brain form	2.4.1.1	681
41	P06737	PYGL	Glycogen phosphorylase, liver form	2.4.1.1	681
42	P11217	PYGM	Glycogen phosphorylase, muscle form	2.4.1.1	681
43	Q96I15	SCLY SCL	Selenocysteine lyase	4.4.1.16	259
44	P20132	SDS SDH	L-serine dehydratase/L-threonine deaminase	4.3.1.17; 4.3.1.19	41
45	Q96GA7	SDSL	Serine dehydratase-like	4.3.1.19; 4.3.1.17	48
46	Q9HD40	SEPSECS TRNP48	O-phosphoseryl-tRNA	2.9.1.2	284
47	O95470	SGPL1 KIAA1252	Sphingosine-1-phosphate lyase 1	4.1.2.27	353
48	P34896	SHMT1	Serine hydroxymethyltransferase, cytosolic	2.1.2.1	257
49	P34897	SHMT2	Serine hydroxymethyltransferase, mitochondrial	2.1.2.1	280
50	O15269	SPTLC1 LCB1	Serine palmitoyltransferase 1	2.3.1.50	-
51	O15270	SPTLC2 KIAA0526 LCB2	Serine palmitoyltransferase 2	2.3.1.50	379
52	Q9NUV7	SPTLC3 C20orf38 SPTLC2L	Serine palmitoyltransferase 3	2.3.1.50	371
53	Q9GZT4	SRR	Serine racemase	5.1.1.18; 4.3.1.18; 4.3.1.17	56
54	P17735	TAT	Tyrosine aminotransferase	2.6.1.5	280
55	Q8IYQ7	THNSL1	Threonine synthase-like 1	-	351
56	Q86YJ6	THNSL2	Threonine synthase-like 2	4.2.3.-	113

Table S1. *Human PLPome used in the reverse docking screening according to B6DB. Uniprot accessions, gene names, protein main names, EC number and the number of the catalytic lysine retrieved from Uniprot database.*

	Entry	Gene names	Protein names	EC number	lys
1	Q9WVM8	Aadat Kat2	Kynurenine/alpha-aminoadipate aminotransferase, mitochondrial	2.6.1.39; 2.6.1.7	263
2	P61922	Abat Gabat	4-aminobutyrate aminotransferase, mitochondrial	2.6.1.19; 2.6.1.22	357
3	A2AIG8	Accs	1-aminocyclopropane-1-carboxylate synthase-like protein 1	-	324
4	Q3UX83	Accsl Gm1967	Probable inactive 1-aminocyclopropane-1-carboxylate synthase-like protein 2	-	417
5	O35423	Agxt Agxt1	Serine--pyruvate aminotransferase, mitochondrial	2.6.1.51; 2.6.1.44	231
6	Q3UEG6	Agxt2	Alanine--glyoxylate aminotransferase 2, mitochondrial	2.6.1.44; 2.6.1.40	349
7	Q8VC19	Alas1	5-aminolevulinate synthase, nonspecific, mitochondrial	2.3.1.37	447
8	P08680	Alas2	5-aminolevulinate synthase, erythroid-specific, mitochondrial	2.3.1.37	391
9	O35484	Azin1 Oazi Oazin	Antizyme inhibitor 1	-	-
10	Q8BVM4	Azin2 Adc Odcp	Antizyme inhibitor 2	-	-
11	P24288	Bcat1 Eca39	Branched-chain-amino-acid aminotransferase, cytosolic	2.6.1.42	222
12	Q91WT9	Cbs	Cystathionine beta-synthase	4.2.1.22	116
13	Q9DBE0	Csad	Cysteine sulfinic acid decarboxylase	4.1.1.29; 4.1.1.11	305
14	Q8VCN5	Cth	Cystathionine gamma-lyase	4.4.1.1	211
15	O88533	Ddc	Aromatic-L-amino-acid decarboxylase	4.1.1.28	303
16	Q8BWU8	Etnppl Agxt2l1	Ethanolamine-phosphate phospho-lyase	4.2.3.2	278
17	P48318	Gad1 Gad67	Glutamate decarboxylase 1	4.1.1.15	404
18	P48320	Gad2 Gad65	Glutamate decarboxylase 2	4.1.1.15	396
19	Q80WP8	Gadl1	Acidic amino acid decarboxylase GADL1	4.1.1.11; 4.1.1.29	362
20	O88986	Gcat Kbl	2-amino-3-ketobutyrate coenzyme A ligase, mitochondrial	2.3.1.29	262
21	Q91W43	Gldc	Glycine dehydrogenase	1.4.4.2	759
22	P05201	Got1	Aspartate aminotransferase, cytoplasmic	2.6.1.1; 2.6.1.3	259
23	Q7TSV6	Got1l1	Putative aspartate aminotransferase, cytoplasmic 2	2.6.1.1	249
24	P05202	Got2 Got-2	Aspartate aminotransferase, mitochondrial	2.6.1.1; 2.6.1.7	279
25	Q8QZR5	Gpt Gpt1	Alanine aminotransferase 1	2.6.1.2	314
26	Q8BGT5	Gpt2 Aat2	Alanine aminotransferase 2	2.6.1.2	340
27	P23738	Hdc	Histidine decarboxylase	4.1.1.22	312
28	Q8BTY1	Kyat1 Ccbl1 Kat	Kynurenine--oxoglutarate transaminase 1	2.6.1.7; 4.4.1.13; 2.6.1.64	247
29	Q71RI9	Kyat3 Ccbl2 Kat3	Kynurenine--oxoglutarate transaminase 3	2.6.1.7; 4.4.1.13; 2.6.1.63	281
30	Q9CXF0	Kynu	Kynureninase	3.7.1.3	276
31	Q3UNZ2	Ldc1 Gm853	Gene model 853,	-	-
32	Q14CH1	Mocos	Molybdenum cofactor sulfurase	2.8.1.9	264
33	Q9Z1J3	Nfs1 Nifs	Cysteine desulfurase, mitochondrial	2.8.1.7	260
34	P29758	Oat	Ornithine aminotransferase, mitochondrial	2.6.1.13	292
35	P00860	Odc1 Odc	Ornithine decarboxylase	4.1.1.17	69
36	Q99K01	Pdxdc1	Pyridoxal-dependent decarboxylase domain-containing protein 1	4.1.1.-	-
37	Q8R1K4	Phykpl Agxt2l2	5-phosphohydroxy-L-lysine phospho-lyase	4.2.3.134	278
38	Q9Z2Y8	Plpbp Prosc	Pyridoxal phosphate homeostasis protein	-	47
39	Q99K85	Psat1 Psa Psat	Phosphoserine aminotransferase	2.6.1.52	200
40	Q8CI94	Pygb	Glycogen phosphorylase, brain form	2.4.1.1	681
41	Q9ET01	Pygl	Glycogen phosphorylase, liver form	2.4.1.1	681
42	Q9WUB3	Pygm	Glycogen phosphorylase, muscle form	2.4.1.1	681
43	Q9JLI6	Scly Scl	Selenocysteine lyase	4.4.1.16	247
44	Q8VBT2	Sds	L-serine dehydratase/L-threonine deaminase	4.3.1.17; 4.3.1.19	41
45	Q8R238	Sdsl Sds	Serine dehydratase-like	4.3.1.17; 4.3.1.19	48
46	Q6P6M7	Sepsecs D5Ertd135e	O-phosphoseryl-tRNA	2.9.1.2	284
47	Q8R0X7	Sgpl1 Spl	Sphingosine-1-phosphate lyase 1	4.1.2.27	353
48	P50431	Shmt1 Shmt	Serine hydroxymethyltransferase, cytosolic	2.1.2.1	251
49	Q9CZN7	Shmt2	Serine hydroxymethyltransferase, mitochondrial	2.1.2.1	280
50	O35704	Sptlc1 Lcb1	Serine palmitoyltransferase 1	2.3.1.50	-
51	P97363	Sptlc2 Lcb2	Serine palmitoyltransferase 2	2.3.1.50	377
52	Q8BG54	Sptlc3 Sptlc2l	Serine palmitoyltransferase 3	2.3.1.50	371
53	Q9QZX7	Srr	Serine racemase	5.1.1.18; 4.3.1.18; 4.3.1.17	56
54	Q8QZR1	Tat	Tyrosine aminotransferase	2.6.1.5	280
55	Q6XPS7	Tha1 Gly1	L-threonine aldolase	-	242
56	Q8BH55	Thnsl1	Threonine synthase-like 1	-	351
57	Q80W22	Thnsl2	Threonine synthase-like 2	4.2.3.-	113

Table S2. *Mouse PLPome used in the reverse docking screening according to B6DB. Uniprot accessions, gene names, protein main names, EC number and the number of the catalytic lysine retrieved from Uniprot database.*

	BCB	LCB	BCaaB	LCaaB	BCM	LCM	BCaaM	LCaaM
ORN	-0.331219	-0.302320	-0.220447	-0.195878	-0.301752	-0.298282	-0.277000	-0.236733
THR	-0.324124	-0.323983	-0.302576	-0.267098	-0.322939	-0.341028	-0.333539	-0.313465
KYN	-0.252282	-0.245647	-0.144551	-0.149288	-0.209821	-0.236270	-0.174780	-0.172164
PLP	-0.127195	-0.072796	-0.099475	-0.051006	-0.147660	-0.093012	-0.136703	-0.083014
TYR	-0.250609	-0.207126	-0.234798	-0.152018	-0.216443	-0.200270	-0.178904	-0.138395
GLU	-0.287984	-0.271693	-0.229996	-0.073668	-0.280426	-0.274366	-0.237204	-0.096094
HIS	-0.264122	-0.282422	-0.218892	-0.202621	-0.273740	-0.274106	-0.236556	-0.181910
CYT	-0.234137	-0.235894	-0.235715	-0.181348	-0.173071	-0.172371	-0.268710	-0.195661
HTL	-0.244020	-0.239865	-0.224641	-0.266979	-0.184935	-0.215541	-0.193404	-0.241753
Mean	-0.257299	-0.242416	-0.212343	-0.171100	-0.234532	-0.233916	-0.226311	-0.184354

Table S3. *Correlation between QMEAN value and the ranking position of all ligands in every ranking criterion considered, colored by a gradient from white (min absolute value) to dark blue (max absolute value) of the dataframe. The last row shows the mean value for each column.*

	BCB	LCB	BCaaB	LCaaB	BCM	LCM	BCaaM	LCaaM
ORN	0.380112	0.377390	0.065151	0.006851	0.371783	0.369082	0.083173	0.010143
THR	0.234132	0.225613	0.131790	0.102088	0.255878	0.262067	0.145983	0.066425
KYN	0.242611	0.212138	-0.051482	0.043043	0.236744	0.227733	0.056200	0.109131
PLP	0.053924	0.000339	0.029792	0.005632	0.069786	0.023749	0.038332	0.017396
TYR	0.236816	0.295307	0.122587	0.133384	0.272608	0.303259	0.148092	0.140647
GLU	0.222613	0.200577	0.137006	-0.060485	0.235838	0.233388	0.170556	0.003545
HIS	0.249786	0.223458	0.073856	0.043652	0.274880	0.238778	0.136592	0.088258
CYT	0.237171	0.265995	0.076299	-0.000505	0.206965	0.253709	0.143226	0.070192
HTL	0.240440	0.233199	0.145611	0.076879	0.249211	0.244279	0.184943	0.135625
Mean	0.233067	0.226002	0.081179	0.038949	0.241522	0.239561	0.123011	0.071263

Table S4. *Correlation between the number of structures per gene and the ranking position of all ligands in every ranking criterion considered, colored by a gradient from white (min absolute value) to dark blue (max absolute value) of the dataframe. The last row shows the mean value for each column.*

Chapter 2

Identification of the sauropsidian pathway for taurine biosynthesis

Introduction

The separation of amniotes occurs ~310 million years ago into two lineages: sauropsids including reptiles and birds, and synapsids whose extant representatives are mammals [1,2]. Before this event, amniotes had developed few characters, such as a protective environment for embryo development called the amniotic or cleidoic egg [3]. After separation, sauropsids and synapsids evolved distinct morphological and physiological adaptations [4]. Despite the availability of several complete genomes for amniotes and other vertebrates, adaptation to terrestrial life is less well understood. Whole genome comparisons have revealed several genes that are not shared between the two classes of amniotes - about 2700 in the comparison between *Gallus gallus* and *Homo sapiens [5]*, but the possibility whether these differences correspond to innovative molecular processes or new metabolic pathways is largely unknown.

Cysteine lyase

Cysteine lyase (EC 4.4.1.10) is an enzyme that catalyzes a peculiar reaction of incorporation of sulfur (sulfite ion) into an amino acid (cysteine) with production of cysteic acid (3-sulfo-L-alanine) and release of reduced sulfur as hydrogen sulfide (H_2S) (**Figure 1**). This activity has been identified in the fertilized chicken egg in the yolk sac [6], a membranous structure that plays a crucial role by providing nutrients to the embryo. This enzymatic activity is firstly observed, during the embryo development, in germ layer cells and increases during the differentiation of the yolk-sac endoderm [7]. The presence of cysteine lyase has been confirmed in the yolk sac of other sauropsids, but not of mammal ones although tested in some species [8]. The partially purified activity [9] was found to require pyridoxal 5'-phosphate (PLP), a key cofactor in amino acid metabolism. Early radiotracer experiments with ^{35}S established that sulfite (SO_3^{2-}) or sulfate (SO_4^{2-}) ions are lastly incorporated into taurine in embryonated eggs, providing evidence that cysteine lyase is involved in a pathway of taurine biosynthesis [6,10].

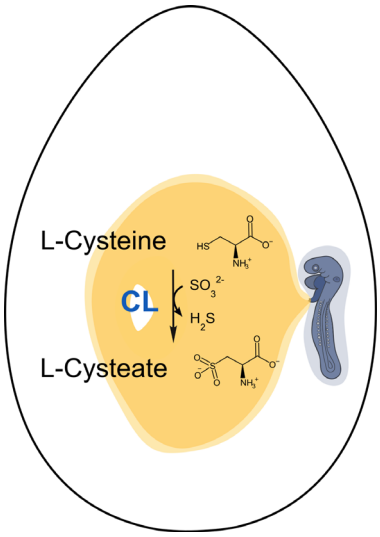


Figure 1. *CL reaction found in fertilized chicken eggs. Cysteine is converted into cysteic acid through a beta-substitution of sulfite ion and thiolic group forming hydrogen sulfide.*

Taurine

Taurine (2-aminoethane sulfonic acid) is an amino acid derivative with important physiological roles in animals [11]. Originally identified in *Bos taurus* bile, taurine is rarely found outside metazoa, but is present in high concentrations (up to 50 mM) in many vertebrate tissues where it is involved in processes such as osmoregulation [12], pH regulation [13], calcium homeostasis [14], cytoprotection [15]. Conjugation of taurine with cholic acid by the liver enzyme bile acid-CoA:amino acid N-acyltransferase (BAAT) produces taurocholate, a major component of vertebrate bile [16]. The low pKa (~1) of taurine sulfonic group ensures solubility of bile salts in the acidic environment of the duodenum.

Biosynthetic pathway

The known pathway of taurine biosynthesis involves formation of hypotaurine from cysteine through cysteine dioxygenase (CDO, EC 1.13.11.20) and cysteine sulfinate decarboxylase (CSAD, EC 4.1.1.29) [17,18], or from cysteamine through cysteamine dioxygenase (ADO, EC 1.13.11.19) [19,20]. The subsequent oxidation of hypotaurine to taurine could occur enzymatically (EC 1.8.1.3) [21] or spontaneously [22,23] (**Figure 2**). In view of the evidence of a cysteine lyase activity, a pathway for taurine synthesis not involving cysteine/cysteamine oxidation and formation of hypotaurine should exist in the developing chicken embryo. However, this pathway is not completely understood due to the lack of knowledge of its genes and proteins.

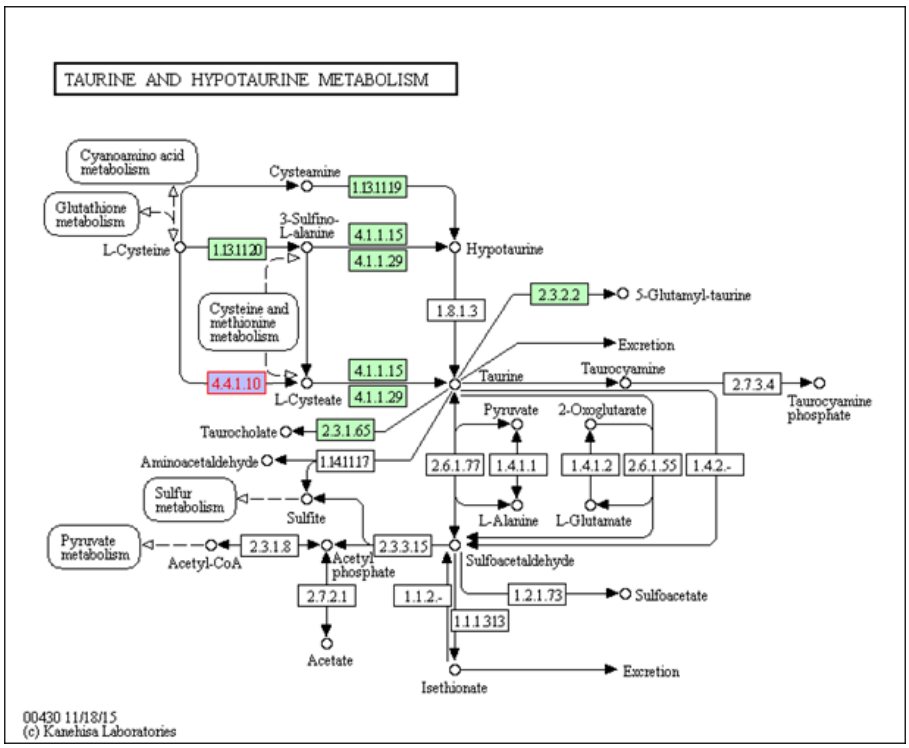


Figure 2. *Kegg scheme of taurine and hypotaurine metabolism. Cysteine lyase is highlighted in red text; in green are shaded the reactions with a characterized gene in Gallus gallus.*

Results

Identification of a candidate cysteine lyase (CL) in *Gallus gallus*

To explain previous observations on the cysteine lyase activity of the chicken embryo [8], we supposed that *Gallus gallus* has a gene encoding a PLP-dependent enzyme, while mammals not. Despite PLP-dependent enzymes catalyze many different reactions, they have few evolutionary origins, enabling to identify an organism's PLPome by bioinformatics. We decided to compare the PLPomes of *Gallus gallus* and *Homo sapiens* (**Figure 3**) with the "whole genome analysis" tool of B6db [24]. This comparison revealed that the genomes of these species encode only a part (~50) of the more than 300 known families deposited in the database, and most of that present in the two species have a one-to-one orthologous relationship (green-shaded entries in **Figure 3**). However, 8 human proteins and 4 chicken proteins (unshaded entries in **Figure 3**) do not have a correspondence in the other species.

Selection

		1. *Gallus gallus*		2. *Homo sapiens*		
Activity	EC (subfamily)	Accession	E-value	Accession	E-value	
Glycine dehydrogenase (decarboxylating).	1:4:4:2	XP_015135548.1	0.0E+000	NP_000161.2	0	
Glycine hydroxymethyltransferase.	2:1:2:1	XP_414824.4	1.6E-287	NP_004160.3	4.2E-295	
				NP_001159828.1	2.4E-277	
Glycine C-acetyltransferase.	2:3:1:29	XP_004950962.1	4.4E-227	NP_001165161.1	3.7E-225	
5-aminolevulinic acid synthase.	2:3:1:37	XP_025010275.1	1.6E-270	NP_000023.2	1.7E-269	
				NP_954635.1	1E-268	
Serine C-palmitoyltransferase.	2:3:1:50 (a)	XP_004949309.1	4.3E-231	NP_006406.1	2.2E-237	
	2:3:1:50 (b)	NP_001006483.1	1.7E-294	NP_004854.1	7.4E-299	
	2:3:1:50 (b)	XP_025005211.1	3.5E-275	NP_001336874.1	4.1E-260	
Phosphorylase.	2:4:1:1	NP_001026205.1	0.0E+000	NP_002853.2	0	
	2:4:1:1	NP_989723.1	0.0E+000	NP_005600.1	0	
				NP_002854.3	0	
Aspartate aminotransferase.	2:6:1:1 (a)	NP_990854.1	1.4E-297	NP_002070.1	5.7E-296	
	2:6:1:1 (a)	NP_990652.1	5.2E-297	NP_002071.2	1.2E-305	
Ornithine--oxo-acid aminotransferase.	2:6:1:13	NP_001006567.1	7.9E-276	NP_001309897.1	3.4E-275	
4-aminobutyrate aminotransferase.	2:6:1:19 (a)	XP_414940.2	0.0E+000	XP_011520702.1	0	
2-aminoadipate aminotransferase.	2:6:1:39 (a)	XP_426286.3	3.6E-236	XP_006714294.1	4.1E-265	
Branched-chain amino acid aminotransferase.	2:6:1:42	XP_416424.1	3.0E-184	NP_005495.2	3.3E-191	
				XP_024307398.1	1.2E-181	
Alanine--glyoxylate aminotransferase.	2:6:1:44	XP_429219.3	0.0E+000	XP_005248394.1	0	
Tyrosine aminotransferase.	2:6:1:5	XP_025010115.1	3.1E-253	NP_000344.1	5.8E-267	
Serine--pyruvate aminotransferase.	2:6:1:51	XP_003641783.2	2.9E-242	NP_000021.1	1.1E-259	
Phosphoserine aminotransferase.	2:6:1:52	XP_424846.3	2.8E-232	NP_478059.1	1E-234	
Kynurenine--oxoglutarate aminotransferase.	2:6:1:7	XP_415485.2	5.8E-262	XP_016870749.1	1.5E-278	
	2:6:1:7	XP_025009029.1	4.8E-268	NP_001336377.1	8.9E-268	
Alanine aminotransferase.	2:6:1:2 (b)			NP_005300.1	6.4E-262	
		XP_015147909.1	9.8E-259	NP_597700.1	4E-267	
Cysteine desulfurase.	2:8:1:7 (a)	NP_001026018.1	3.4E-297	NP_066923.3	1.9E-298	
molybdenum cofactor sulfurtransferase	2:8:1:9	XP_419048.4	1.9E-290	NP_060417.3	0	
O-phospho-L-seryl-tRNA(Sec):L-selenocysteinyl-tRNA synthase	2:9:1:2	NP_001026329.1	0.0E+000	NP_058651.3	0	
Kynureninase.	3:7:1:3	XP_004943069.1	9.1E-240	XP_024308976.1	3.1E-259	
Glutamate decarboxylase.	4:1:1:15 (a)	NP_990244.1	0.0E+000	NP_000808.2	0	
	4:1:1:15 (a)	XP_015137540.1	0.0E+000	NP_001127838.1	0	
Ornithine decarboxylase.	4:1:1:17:1	NP_001161238.1	1.7E-171	NP_002530.1	5.2E-174	
Histidine decarboxylase.	4:1:1:22 (b)	NP_001280217.1	2.1E-283	NP_002103.2	1.3E-285	
Aromatic-L-amino-acid decarboxylase.	4:1:1:28	XP_419032.3	3.7E-275	NP_001076440.1	1.9E-285	
Sulfinoalanine decarboxylase.	4:1:1:29	XP_004939481.1	0.0E+000	NP_997242.2	0	
	4:1:1:29	XP_025001257.1	1.0E-283	NP_057073.4	0	
Sphinganine-1-phosphate aldolase.	4:1:2:27	NP_001007947.1	3.5E-255	NP_003892.2	1.1E-248	
Threonine aldolase.	4:1:2:5	XP_015151050.1	1.0E-164			probable low-specificity L-threonine aldolase 2
Cystathionine beta-synthase.	4:2:1:22	XP_015156364.1	7.4E-231	NP_001171479.1	7.5E-228	
	4:2:1:22	XP_015156382.1	4.8E-214			cystathionine beta-synthase-like
Threonine synthase.	4:2:3:1	XP_025005939.1	6.4E-083	NP_060741.3	2.8E-080	
	4:2:3:1	XP_025004138.1	1.2E-093	NP_079114.3	1.2E-090	
Ethanolamine-phosphate phospho-lyase	4.2.3.2	XP_015132066.2	1.6E-119	NP_112569.2	1.3E-114	
5-phosphonooxy-L-lysine phospho-lyase	4.2.3.134	XP_025010761.1	1.0E-111	NP_699204.1	9.3E-100	
L-serine ammonia-lyase.	4:3:1:17a	NP_001185572.1	5.4E-204	NP_612441.1	1.8E-205	
				NP_006834.2	8.4E-209	
D-serine ammonia-lyase.	4:3:1:18 (b)	XP_003640696.1	7.2E-172			uncharacterized protein CHDSD
Cystathionine gamma-lyase.	4:4:1:1	XP_422542.2	2.2E-280	NP_001893.2	5.1E-291	
1-aminocyclopropane-1-carboxylate synthase.	4:4:1:14	XP_015142682.1	3.2E-082	NP_001027025.2	3.4E-078	
	4:4:1:14	XP_015142683.1	4.6E-081	NP_115981.1	1.4E-081	
Selenocysteine lyase.	4:4:1:16 (b)	NP_001132935.1	0.0E+000	NP_057594.4	0	
Serine racemase.	5:1:1:18	XP_001234891.1	6.6E-180	XP_006721628.1	1.4E-223	
Ornithine decarboxylase paralogue		NP_001280585.1	8.9E-145	NP_443724.1	1E-196	
		NP_001008729.1	4.2E-110	NP_056962.2	2.3E-116	
L-phenylserine dehydratase		XP_004942032.2	2.9E-049			uncharacterized protein SRRL
GOT1L1				NP_689626.2	0	
PROSC		XP_424381.4	2.4E-132	NP_009129.1	5.2E-150	
PDXDC1		XP_004945124.1	9.3E-007	NP_001272379.1	0.00000052	
PDXDC2				XP_016885696.1	0.00000011	

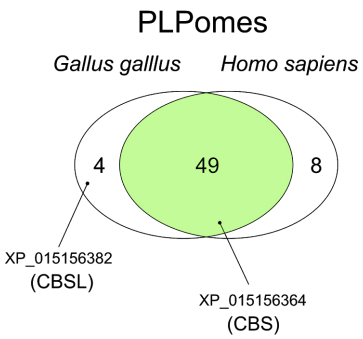


Figure 3. *Comparison between Gallus gallus and Homo sapiens PLPomes as classified by B6db. Orthologous proteins (BRH test) are colored in green. Specific Gallus proteins without corresponding human orthologs are in bold. The Venn diagram summarizes numbers of shared and unique genes in the two organisms, where CBSL is specific of Gallus gallus, in spite of its homolog CBS [25].*

The 4 chicken proteins without a counterpart in humans were examined in detail. The XP_015151050 protein with similarity to threonine aldolase was easily excluded because parallel comparisons including other species revealed that the gene encoding this enzyme is usually present in mammals although it is a pseudogene in humans [26]. According to the B6db classification, the XP_003640696 and XP_004942032 proteins catalyze the cleavage of a carbon-oxygen bond in D-serine [27] and L-phenylserine, respectively, through β-elimination reactions with partial similarity to the cysteine lyase reaction. By contrast, the XP_015156382 protein (gene: LOC418544) was inserted in the same family of cystathionine beta-synthase (CBS), a ubiquitous enzyme in amniotes that catalyzes very similar β-replacement reactions. Indeed, the primary cysteine lyase activity (reaction 1, **Figure 4**) is analogous to the serine "sulfhydrase" activity [28] of CBS (reaction 4, **Figure 4**), while its secondary activity, i.e. formation of lanthionine (reaction 2, **Figure 4**), is analogous to the main CBS activity, i.e. formation of cystathionine in the transsulfuration pathway of cysteine biosynthesis (reaction 3, **Figure 4**).

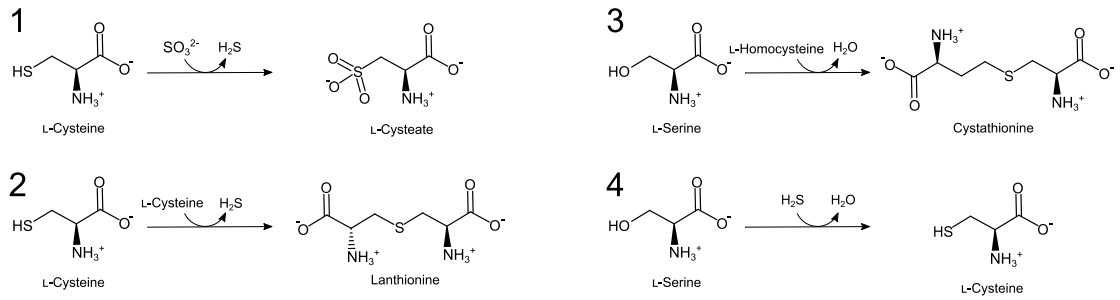


Figure 4. CL reactions compared to CBS reactions. *(**left**) Reactions catalyzed by unidentified cysteine lyase: synthesis of cysteate (**1**) and lanthionine (**2**). (**right**) Reactions catalyzed by cystathionine beta-synthase (CBS): synthesis of cystathionine (**3**) and conversion of serine into cysteine via addition of hydrogen sulfide (**4**). Both enzymes catalyze a beta substitution towards amino acid, with another amino acid molecule (cysteine or homocysteine) or an inorganic one (H_2S or SO_3), with release of sulfide or water [25].*

To analyze the conservation of the active site residues of cystathionine beta-synthase family, we structurally aligned pdb structures of experimentally validated CBS proteins [29] together

with structural models of *Gallus gallus* homologous sequences. The *Gallus gallus* sequence annotated as CBS in RefSeq (XP_015156364) had a perfect conservation of the residues lining active site, while the protein encoded by LOC18544 and annotated as CBS-like (XP_015156382) showed a partial conservation of this region, with a conserved catalytic lysine (K119), and several non-conservative substitutions (**Figure 5**), consistent with the possible acquisition of a divergent function. Gene expression data of chick embryo in the GEISHA database [30] suggested that LOC18544 expression during egg development at an extra-embryonic level is in accordance with the cysteine lyase activity described in literature [7]. Given the evidence of the bioinformatics analysis we decided to characterize the protein encoded by LOC418544 as major candidate for cysteine lyase, and based on the observations described below, we named this gene cysteine lyase (CL).

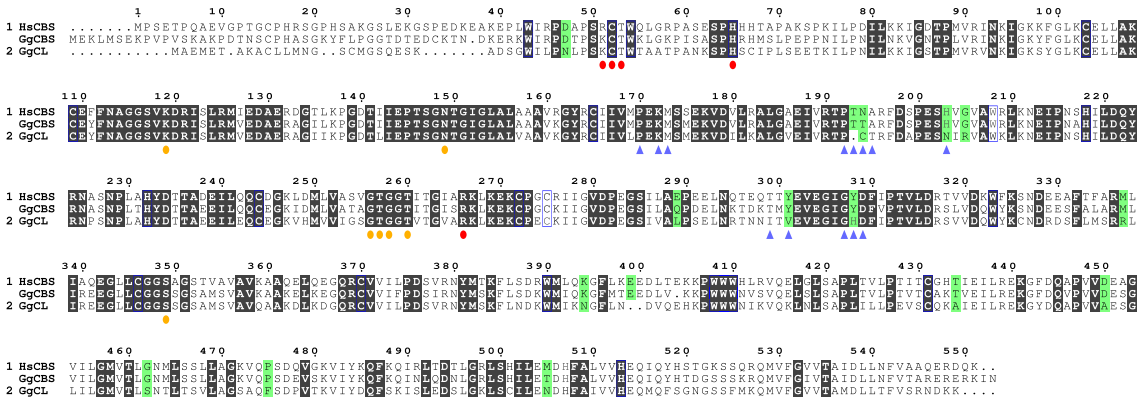


Figure 5. *Multiple sequence alignment of Homo sapiens CBS (HsCBS), Gallus gallus CBS and CL proteins (GgCBS, GgCL). Colored dots indicate residues that recognize heme (red) and PLP (yellow), blue triangles indicate the substrate cavity in the CBS structure (PDB ID 3PC3). Conserved amino acids based on the multiple alignment of 8 putative CL orthologues and 22 CBS sequences from vertebrates are shaded in black. Green shading highlights conserved differences between CBS and CL families [25].*

Recombinant CL is an enzyme with heme and PLP

The *Gallus gallus* CL (*Gg*CL) sequence has about the same similarity with *Gg*CBS (65.1% identity) and *Hs*CBS (64.6% identity), less than the similarity observed between *Hs*CBS and *Gg*CBS (75.3% identity). The multiple alignment shows that CL and CBS sequences are similar over their entire lengths (**Figure 5**). Consistently, the predicted domain for *Gg*CL show a very similar architecture like CBS, with a N-terminal heme-binding motif, a central PLP-binding domain (PALP), and two C-terminal CBS tandem repeats (**Figure 6a**). We can obtain soluble

expression of CL with a truncated form of the protein that lacks the non-catalytic CBS tandem repeats (aa 1-396, **Figure 6a**) as previously reported for the homologous CBS enzyme [31,32],.

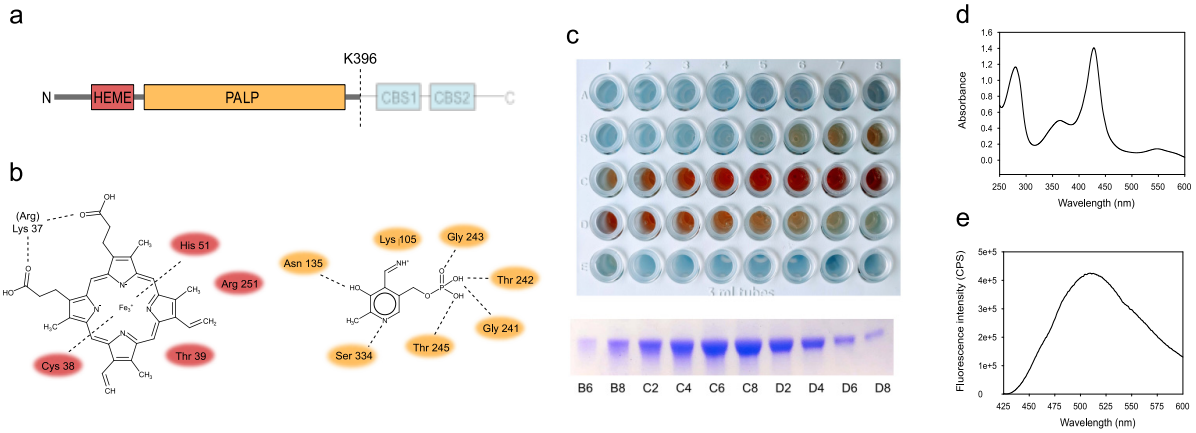


Figure 6. *(**a**) Scheme of GgCL domains. Dashed line indicates the position of gene truncation for recombinant protein expression. Predicted interactions with heme (**b**, left) and PLP (**b**, right) are shown with residues conserved in the alignment of CBS/CL proteins (**Figure**) highlighted in colors. (**c**) GgCL fractions collected after cation-exchange purification, showing a color tone proportional to the protein concentration as determined by SDS-PAGE. (**d**) Absorbance spectrum of recombinant GgCL (16.5 μM) in NaH_2PO_4 (20 mM), pH 7.0. (**e**) Fluorescence emission spectrum (excitation: 412 nm) of recombinant GgCL (22 μM) in NaH_2PO_4, pH 7.0 [25].*

Recombinant His-tagged *Gg*CL was produced with high yield in *E. coli* and purified to homogeneity. *Gg*CL-enriched fractions exhibited a vivid orange color after affinity and ion exchange chromatography (**Figure 6c**). Conservation of the residues for the binding of heme the catalytic cofactor PLP (**Figure 6a, b**) suggests that CL, like CBS, can bind both the cofactors. The absorbance spectrum of recombinant *Gg*CL showed the Soret peak, typical of heme proteins (**Figure 6c**). Usually, PLP-dependent enzymes exhibit an absorption peak at around 410 nm, due to the absorbance of the internal aldimine between PLP and the active site lysine. The presence of PLP cofactor was not apparent in the absorbance spectrum due to the dominant signal of heme in the PLP absorbance range [33]. However, the fluorescence emission spectrum upon excitation at 412 nm showed a peak centered at 510 nm attributable to the ketoenamine tautomer of bound PLP (**Figure 6d**).

CL catalyzes substitution of cysteine thiol with sulfite

The *Gg*CL activity (**Figure 7a-e**) was initially monitored spectrophotometrically by trapping *in situ* generated H_2S with lead acetate to form lead sulfide (PbS), an intensively absorbing

dark compound. H_2S release could be observed in the presence of cysteine alone and was much faster in the presence of sulfite (**Figure 7a**), suggesting that *Gg*CL has the capability to catalyze both cysteine lyase reactions (see **Figure 4**). In the absence of lead acetate, these reactions, despite the small amount of reagents, can be recognizable for their unpleasant odor owing to the high sensitivity of the human nose to H_2S -also known as the rotten egg smell [34].

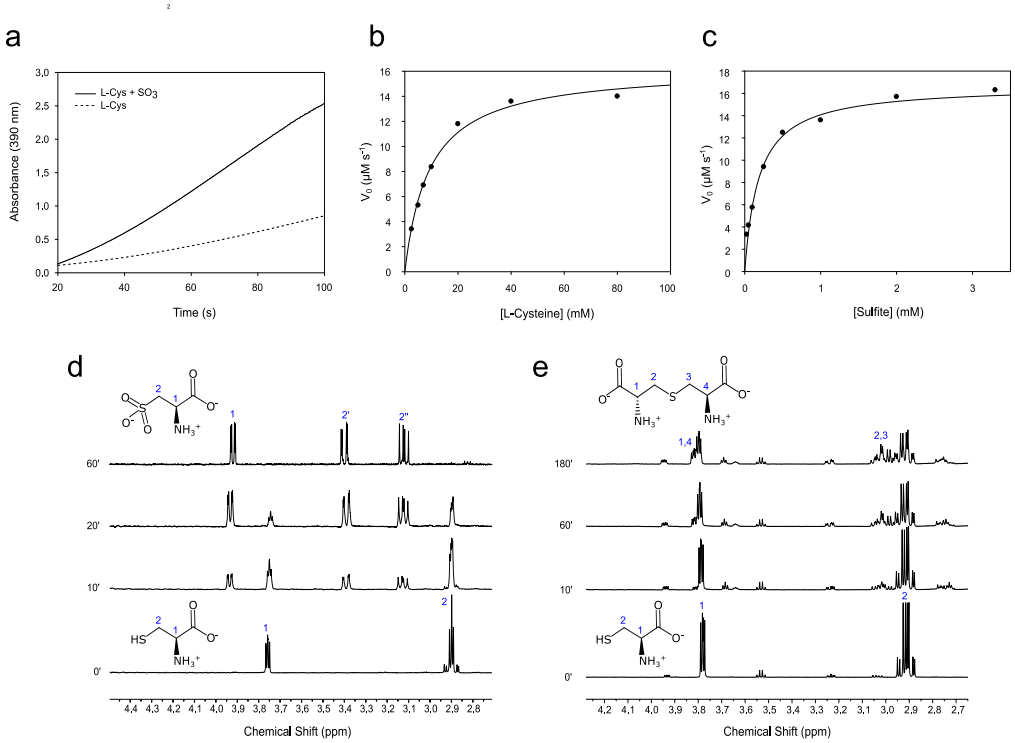


Figure 7. *(**a**) Kinetics of H2S production by the CL main reaction followed spectrophotometrically at 390 nm in 50 mM NaH_2PO_4, pH 7.0 with cysteine (5 mM), lead acetate (0.4 mM), GgCL (1 µM), in the presence (solid line) or in the absence of Na_2SO_3 (5 mM, dashed line). (**b**, **c**) Non-linear fitting to the Michaelis- Menten equation of the dependency on substrate concentrations of the initial reaction velocity of GgCL (1 µM) with fixed (**b**) Na_2SO_3 (5 mM) and (**c**) cysteine (40 mM). (**d**) Time-resolved ^1H NMR spectra of cysteine (5 mM) with the addition of GgCL (1 µM), showing total conversion into cysteic acid. (**e**) Time-resolved ^1H NMR spectra of cysteine (10 mM) with the addition of GgCL (1 µM), showing minor conversion into lanthionine [25].*

The dependence on substrate concentrations of reaction velocity followed Michaelis-Menten (MM) kinetics (**Figure 7b, c**). Fitting to the MM equation gave a k_{cat} value of **16.5 ± 1.72 s^{-1}** and K_m values of 9.2 **± 0.92 mM** (cysteine) and **0.18 ± 0.02 mM** (sulfite).

The formation of the reaction products was directly monitored through time-resolved ^1H NMR spectrometry. Cysteine was completely converted into cysteic acid (CA) in the presence of excess sulfite, (**Figure 7d**), while a minor conversion into lanthionine was observed without

the addition of sulfite (**Figure 7e**). *Gg*CL revealed no activity towards homocysteine and serine, or H_2S and serine, as substrates (**Figure 8a, b**), suggesting that it is not able to catalyze the CBS reactions (see **Figure 4**). Despite its inability to act with a β-replacement on serine, *Gg*CL is nevertheless able to subtract the alpha proton of this substrate (**Figure 8c, d**).

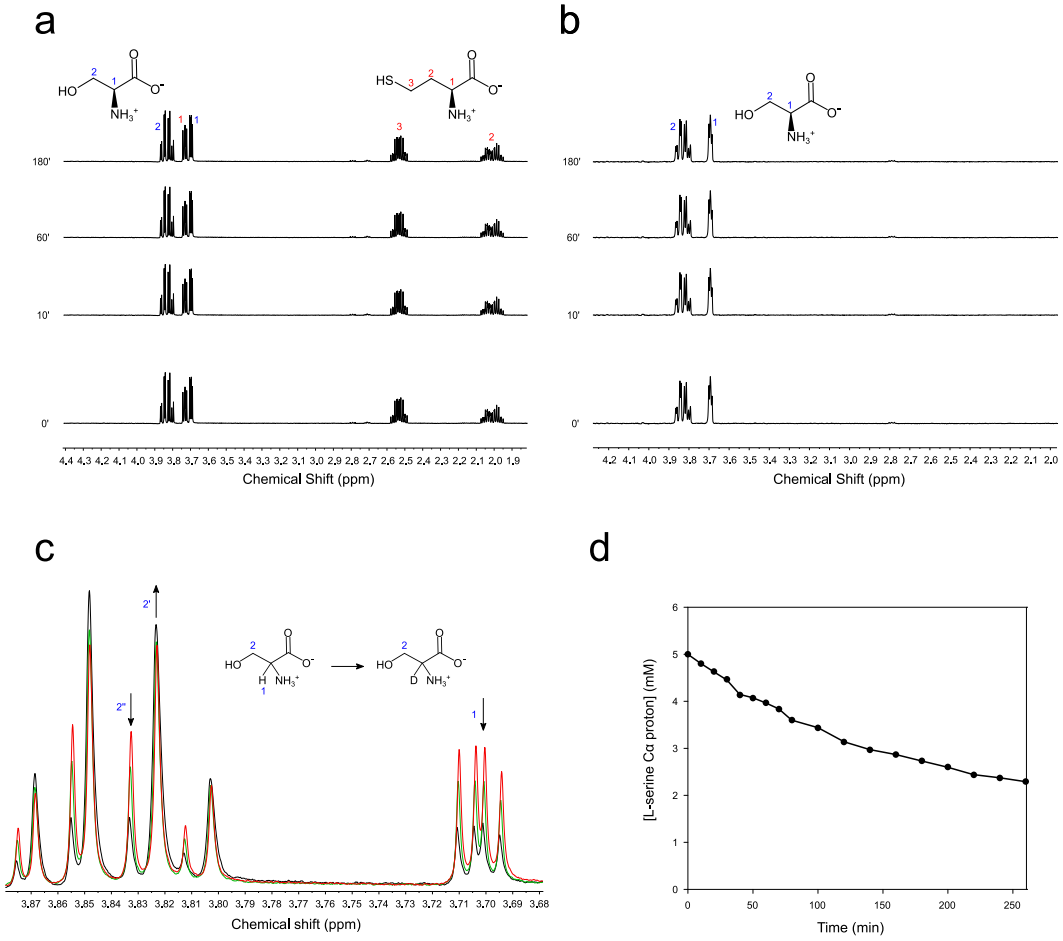


Figure 8. *(**a**) Time-resolved ^{1}H NMR spectra of homocysteine (5 mM) and serine (5 mM) with the addition of GgCL (1 μM). (**b**) Time-resolved ^{1}H NMR spectra of Na_2S (5 mM) and serine (5 mM) with the addition of GgCL (1 μM). (**c**) Exchange of serine alpha proton with deuterium catalyzed by GgCL (1 μM) in 95% D_2O. Spectra were overlayed at time 0' (red), 60' (green), 260' (blue). (**d**) ^{1}H peak integration of serine Cα proton is reported in the plot using the interval 0'-260' [25].*

Gallus gallus CSAD is a specific CA decarboxylase

According to previous observations, the CL product cysteic acid (CA) is finally converted into taurine by the chicken embryo [6,10]. This reaction is presumably catalyzed by a PLP-dependent enzyme because it involves an α-decarboxylation towards an amino acid. Our

previous PLPome comparison did not highlight a hypothetic PLP-dependent decarboxylase specific for *Gallus* (see **Figure 3**). However, *Gallus gallus* has the ortholog of *Homo sapiens* cysteine sulfinic acid decarboxylase (CSAD; EC 4.1.1.29) of the mammalian pathway, an enzyme able to catalyze even the CA decarboxylation, albeit with minor efficiency with respect to its main substrate CSA [35]. The *Gallus* ortholog (XP_025001259) of human CSAD has the PLP binding residues conserved (**Figure 9a, b; Figure 11**) in all vertebrate's orthologues examined. *GgCSAD* was recombinantly expressed in *E. coli* as a PLP protein, with a prevalence for enolimine tautomer of the cofactor (**Figure 9e, f**) which gives the protein the characteristic yellow color of PLP (**Figure 9c**). Decarboxylase activity was tested by ^{1}H NMR (**Figure 10a, b**).

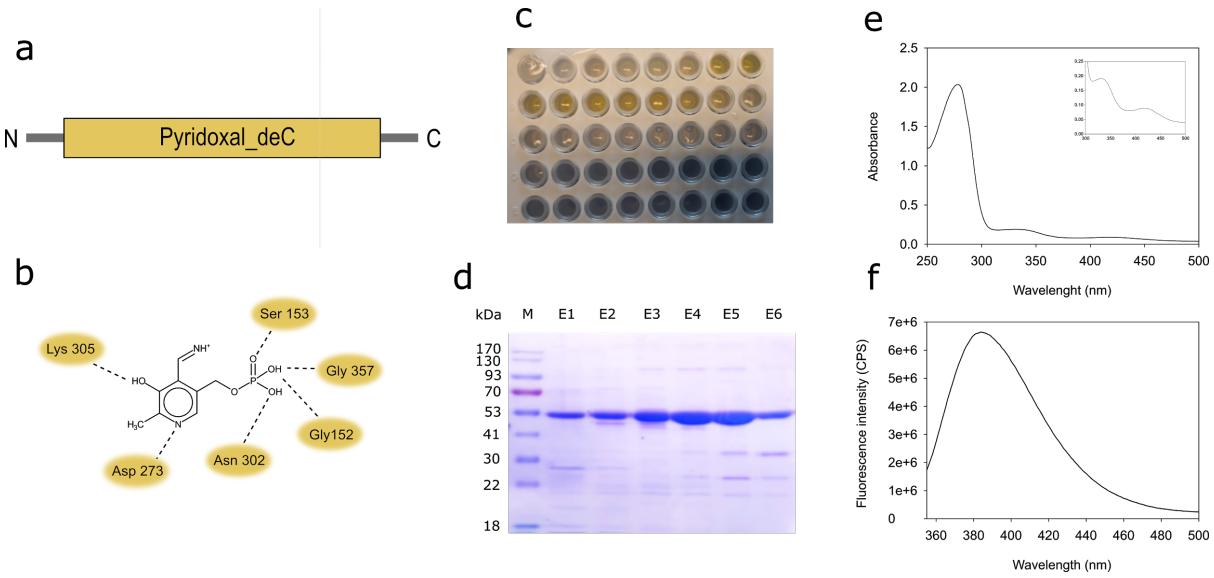


Figure 9. *(**a**) GgCSAD domain composition according to PFAM. Predicted interactions with PLP (**b**) according to conserved residues in the alignment of CSAD ortholog proteins (**Figure 12**). (**c**) GgCSAD fractions collected after cation-exchange purification, showing a color tone proportional to the protein concentration as determined by SDS-PAGE (**d**). (**e**) Absorbance spectrum of purified GgCAD in 20 mM NaH_2PO_4, pH 8.0 and 100 mM NaCl; The wavelength region of PLP tautomers (ketoenamine 415 nm, enolimine 340 nm) is shown in the up-right inset. (**f**) Fluorescence emission spectrum of PLP enolimine tautomer upon excitation at 340 nm [25].*

Unexpectedly, the *Gallus* decarboxylase was able to catalyze CA decarboxylation into taurine (**Figure 10a, c**), with higher efficiency than CSA decarboxylation, serving as a poor substrate (**Figure 10b,c**). Velocity of CA decarboxylation was slower in the presence of CSA (**Figure 11b, c**). This inhibition can be ascribed to CSA itself rather than to its decarboxylation product hypotaurine (**Figure 11a, c**) that is partly formed during the reaction (see **Figure 10b**). For

this reason, we propose to name CA decarboxylase (CAD) the enzyme encoded by the *Gallus gallus* gene annotated automatically as CSAD based on its orthology with the human one.

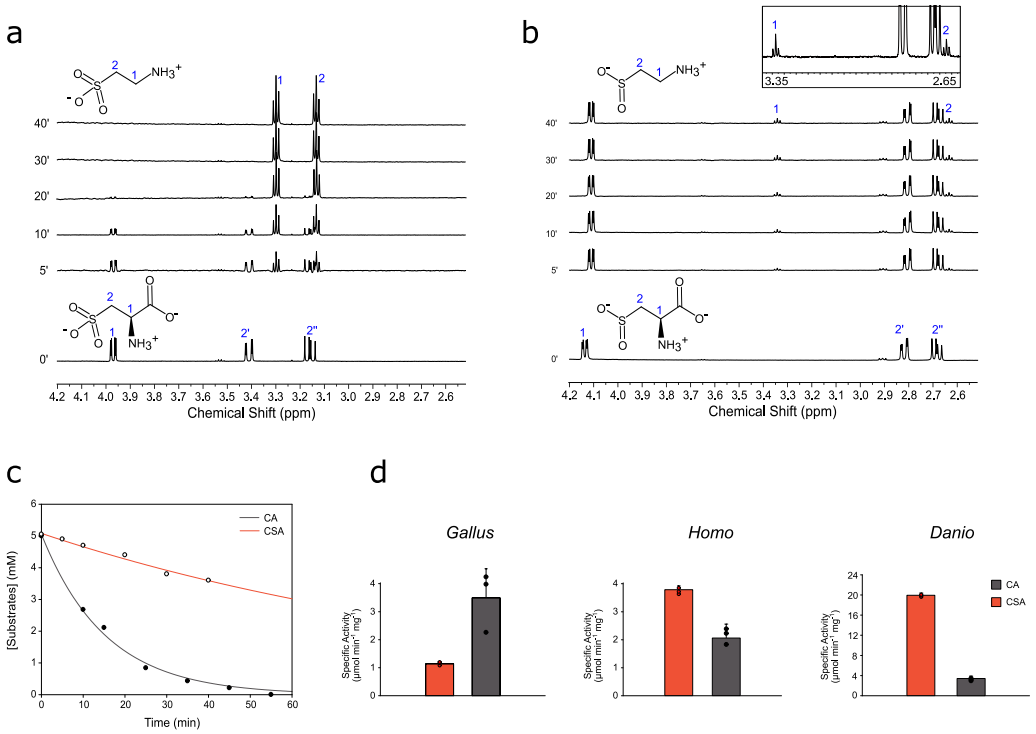


Figure 10. *(**a**) Time-resolved ^{1}H NMR spectra of cysteic acid (5 mM) with the addition of GgCAD (1 μM), showing total conversion into taurine. (**b**) Time-resolved ^{1}H NMR spectra of cysteine sulfinic acid (5 mM) with the addition of GgCAD (1 μM), showing minor production of hypotaurine. (**c**) GgCAD kinetics in the presence of different substrates. The plot shows the time-dependent decrease of cysteic acid (CA) and cysteine sulfinic acid (CSA). The grey curve shows the fitting of the experimental points obtained by NMR peaks integration with the integrated Michaelis-Menten equation (Schnell and Mendoza, 1997) with Km = 6.95 ± 3.23 mM, kcat = 10.54 ± 3.46 s^{-1}; (**d**) Specific activities of Gallus gallus, Homo sapiens, Danio rerio CSAD orthologs with CSA and CA substrates [25].*

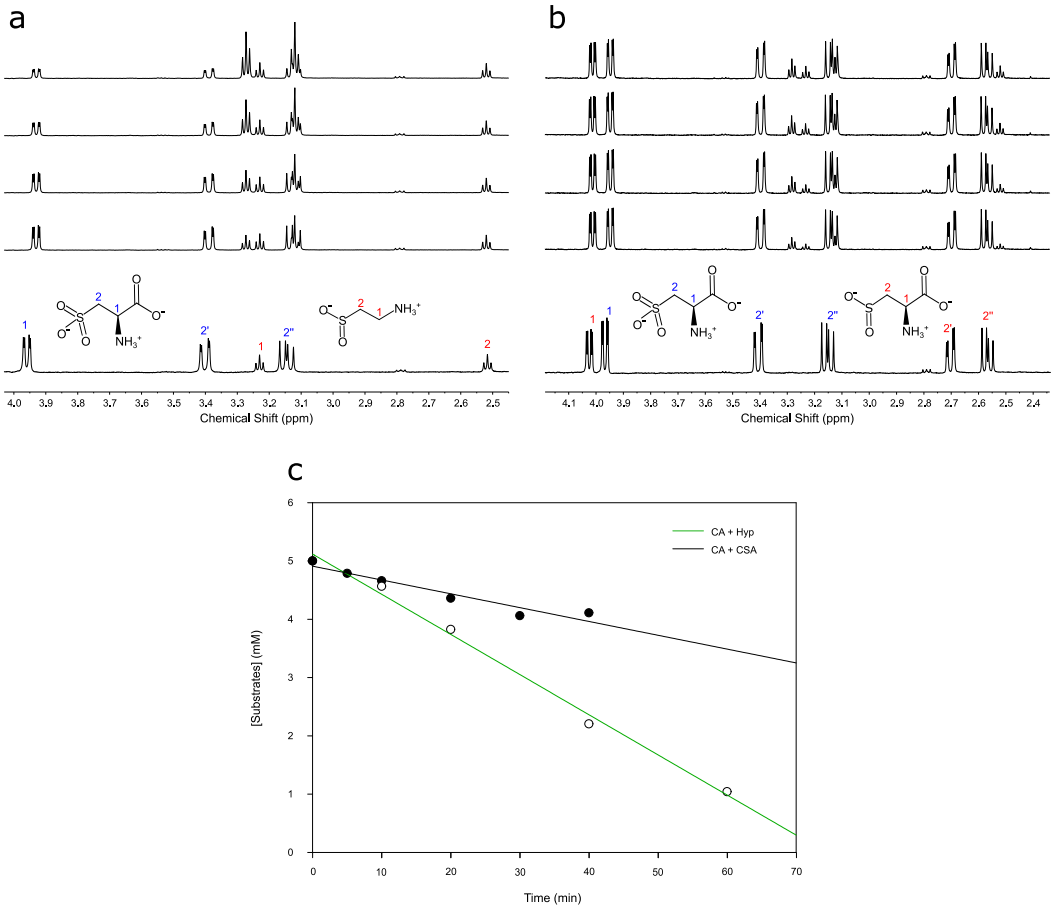


Figure 11. *(**a**) Time-resolved ¹H NMR spectra of hypotaurine (5 mM) and cysteic acid (5 mM with the addition of GgCAD (1 μM), showing slight inhibition of CAD activity. (**b**) Time-resolved ¹H NMR spectra of cysteine sulfinic acid (5 mM) and cysteic acid (5 mM) with the addition of GgCAD (1 μM), showing strong inhibition of CAD activity. (**c**) GgCAD kinetics in the presence of different substrates. The plot shows the time-dependent decrease of cysteic acid in the presence of hypotaurine (CA + Hyp) or cysteine sulfinic acid (CA+ CSA) [25].*

We confirmed the previously reported preference of human CSAD for the cysteine sulfinic acid in our experimental conditions and we observed the same specificity in the CSAD ortholog of a basal vertebrate such as *Danio* rerio (**Figure 10d**). No conserved differences in the active site are found between the comparison of *Gg*CAD and the other sauropsidian orthologs, and CSAD sequences from other non-sauropsidian vertebrates (**Figure 12**). However, we found two conserved mutation (Hydrophobic → Hydrophilic) in the residues located within 10 Å from the active site cavity of the sauropsidian sequences (**Figure 12)**.

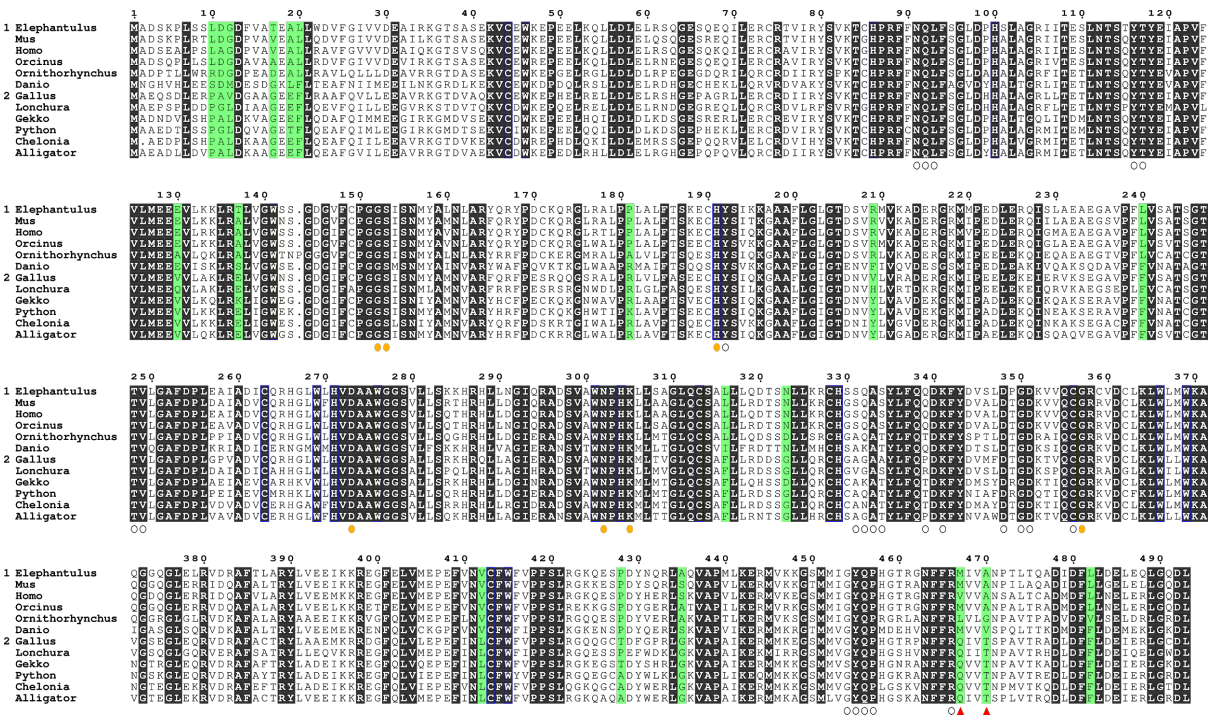


Figure 12. *Multiple alignment of selected Gallus CSAD orthologs. Residues that bind PLP (orange) or surround the active site cavity (white) in the CSAD structure (PDB ID: 2JIS) are marked with filled circles. Conserved residues are shaded in black, while conserved differences between Sauropsids and non-Sauropsids are shaded in green. Residues within 10 Å from the active site that correspond to the green shading, are marked with red triangles, and used as mutation targets (Q467V, T470A)* ***[25]****.*

We analyzed the activity towards the two substrates of single (T470A and Q467V) and double site directed mutants of *Gg*CAD, and we noted showed the contribution of these two substitutions to the preference of *Gallus* protein for CA; in fact, in the double mutant the preference is overturned, obtaining a better activity in the presence of CSA (**Figure 13**).

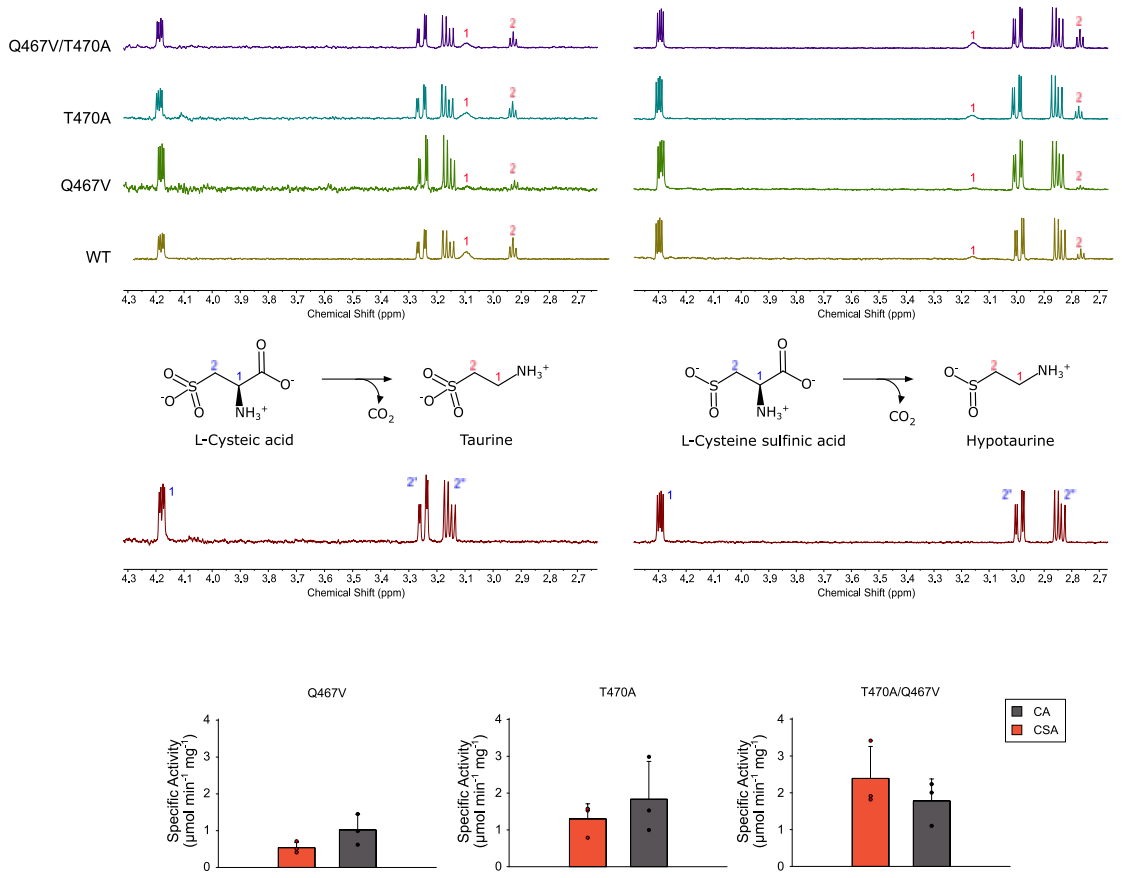


Figure 13. *(**upper**) ^{1}H NMR spectra showing decarboxylase reaction activity of wild-type, single and double mutants (WT, Q467V, T470A, Q467V/T470A, respectively) in presence of cysteic acid (left) or cysteine sulfinic acid (right) after 5' of reaction stopped with 1M of HCl. (**lower**) Specific activities of GgCAD single and double mutants (Q467V, T470A, Q467V/T470A, respectively) with CA and CSA substrates with standard error bars [25].*

CL, CBS, and CSAD are expressed during early stages of embryogenesis

Thanks to the collaboration with GEISHA database from Arizona (http://geisha.arizona.edu), that collects in situ hybridization images of chicken embryos at different stage, we could observe the localization of our three genes. We report here results taken from Malatesta et al. [25]:

*"To determine the expression of CL, CBS and CSAD during early stages of chicken embryogenesis, whole mount in situ hybridization analyses were performed in chicken embryos between 0.5 and 4 days of development (Hamburger-Hamilton [HH] stages 4-24) [36]. At HH stage 4, CL expression was first detected in the extraembryonic endoderm at the boundary of the area pellucida and area opaca (**Figure 14a**). At HH stages 10 and 18, CL mRNAs were broadly detected throughout the extraembryonic endoderm (**Figure 14b, c**). At HH stage 18 and 24, widespread expression was also evident in the embryo proper (**Figure 14c, d**). CBS expression was first detected at HH stage 4 weakly in the epiblast (**Figure 14e**). At HH stage 10, CBS mRNAs were localized to the head region and in the intermediate mesoderm, with strong expression in the primitive blood cells of the extraembryonic blood islands. Broad CBS expression was evident throughout the embryo at HH stages 18 and 24 (**Figure 14g, h**). CSAD expression was first detected at HH stage 4 in the extraembryonic endoderm (**Figure 14i**). Expression in extraembryonic endoderm persisted at HH stages 10 and 18 (**Figure 14j, k**) At HH stage 24, CSAD mRNAs were detected throughout the embryo, with higher levels of expression observed in the liver and mesonephros (arrowhead and arrow, **Figure 14l**).*

*Inspection of the genomic regions containing CL, CBS, and CSAD genes revealed that CL is adjacent to CBS in a head-to-tail orientation on the Gallus chromosome 1 (**Figure 14m**). Analysis of available RNA-seq profiles shows a prevalence of CBS over CL transcripts in the aggregated dataset. Tissue-specific RNA-seq profiles show abundant CBS transcripts in the adult kidney and liver where CL transcripts are barely detected. Conversely, CL transcripts are more abundant than CBS transcripts in the adult duodenum (**Figure 14m**). CSAD is located on Gallus chromosome 33 adjacent to ZNF740 in a head-to-head orientation (**Figure 14n**). The same organization is also observed for the human gene (www.ncbi.nlm.nih.gov/gene/51380), supporting orthology. CSAD transcripts are present in several adult tissues and especially abundant in kidney, liver, and duodenum (**Figure 14n**)."*

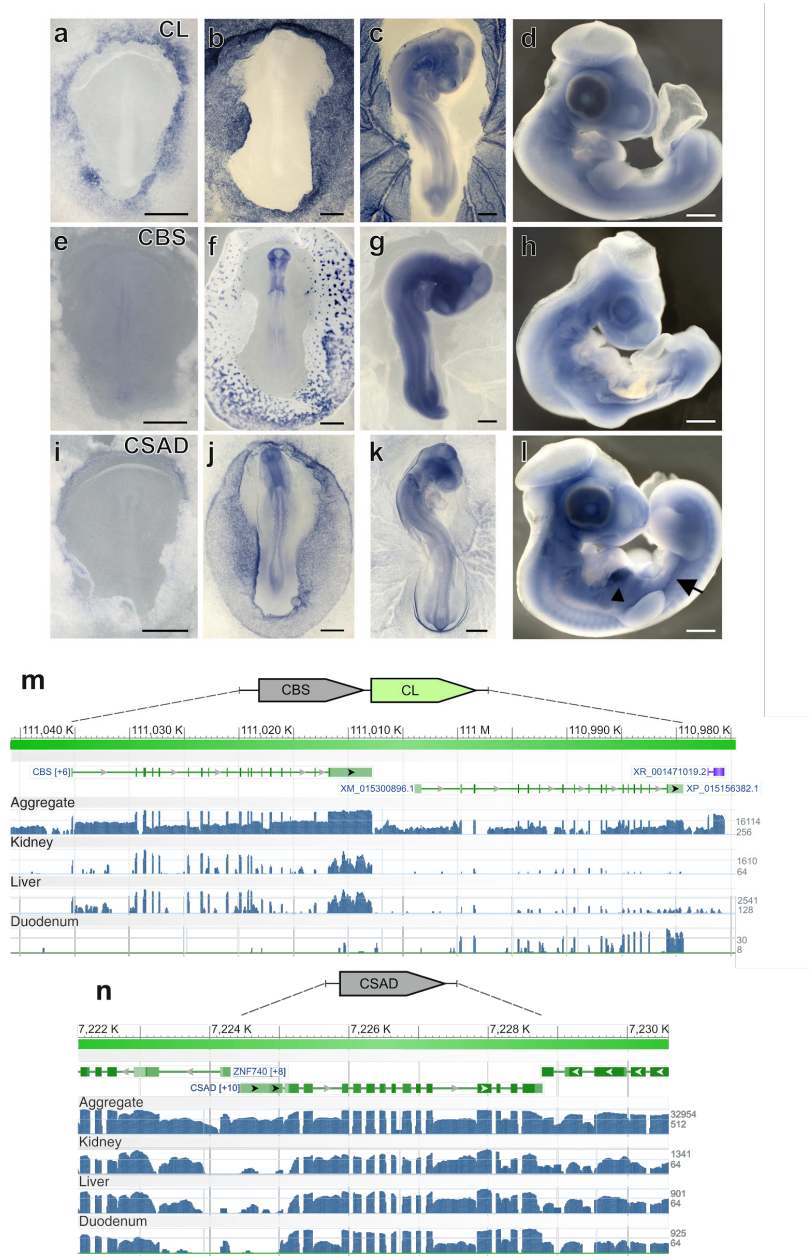


Figure 14. *(**a-d**) In situ hybridization analysis of CL expression in chicken embryos at Hamburger-Hamilton developmental stages 4, 10, 18, and 24, sorted from left to right. (**e-h**) In situ hybridization analysis of CBS expression at HH stages 4, 10, 18, and 24, sorted from left to right. (**i, l**) In situ hybridization analysis of CSAD expression at HH stages 4, 10, 18, and 24, sorted from left to right. (**m**) NCBI Sequence-viewer representation of the genomic region on Gallus gallus chromosome 1 (annotation release 104) encompassing the CBS and CL genes. Gene exon structure is represented by green segments. Blue bars represent RNA-seq exon coverage (log2 scaled) for aggregate, kidney (SAMEA2201372), liver (SAMEA2201470), and duodenum (SAMN03376186) datasets. (**n**) NCBI Sequence-viewer representation of the genomic region on Gallus gallus chromosome 33 (annotation release 104) encompassing the CSAD gene. Tracks are as in panel **m** [25].*

The CL pathway for taurine biosynthesis

We can design the pathway for taurine synthesis in fertilized chicken eggs (**Figure 15**) with the evidence obtained of *Gallus* CL and CAD proteins. The cysteine thiol group replaced with sulfite ion by CL lead to the production of CA, which is efficiently decarboxylated by CAD to obtain taurine, the final product of the pathway (**Figure 15a**, upper branch). The combined use of these enzymes in reaction, cysteine is rapidly and quantitatively converted into taurine in the presence of sulfite with small temporary CA accumulation observed during the reaction (**Figure 15b**). The identification of CL orthologs in all sauropsids (see next paragraph) suggests the possibility of this pathway is conserved in this class of amniotes. Contrary, the synapsids do not have this pathway because the taurine production is due by cysteine oxidation followed by CSA decarboxylation to hypotaurine followed by hypotaurine oxidation [17,18] (**Figure 15a**, lower branch).

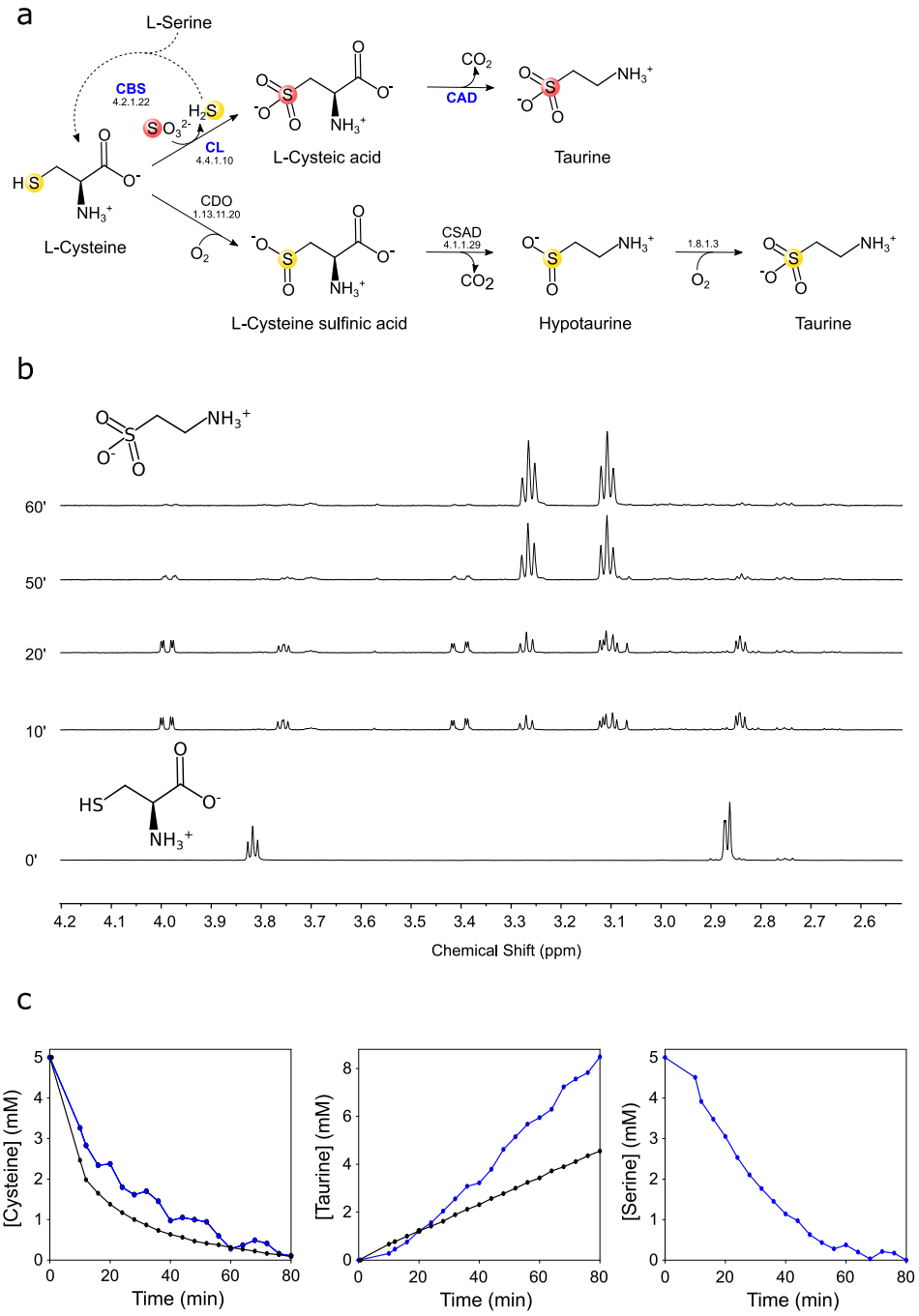
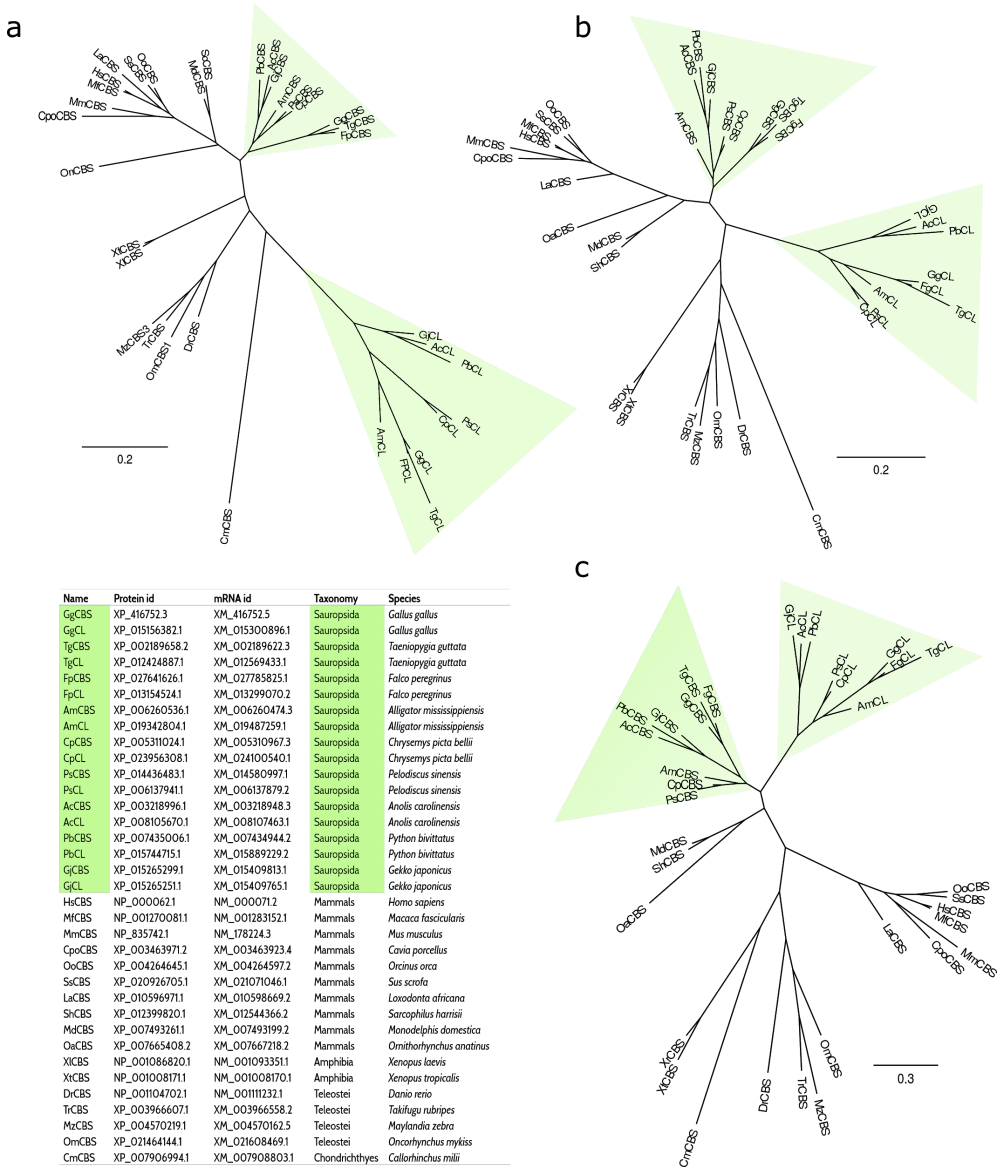


***Figure* 15.** *(**a**) The identified pathway for taurine biosynthesis in sauropsids compared with the known mammalian one. Dashed line shows recycle of hydrogen sulfide into cysteine catalyzed by CBS. (**b**) All-in-one enzymatic taurine production from cysteine. Time-resolved ^{1}H NMR spectra of sulfite (7 mM) and cysteine (5 mM) with the addition of recombinant GgCL (1 µM) and GgCAD (1 µM) proteins. (**c**) ^{1}H peak integration of cysteine (left), taurine (center), and serine (right) NMR corresponding signals using the same conditions of (**b**) with (black dots) or without (blue dots) of GgCBS (4 µM) and serine (5 mM) [25].*

With respect to the mammalian pathway, the sauropsidian pathway for taurine biosynthesis does not involve the oxidation of reduced sulfur and has one less enzyme involved in. The H_2S released in the CL reaction is not wasted, as it can be recycled for cysteine formation using other amino acids as substrates (**Figure 15a**, dashed line). The CBS activity assays revealed its capability to form cysteine from serine and H_2S (see **Figure 4**). Since the coding gene of this enzyme is expressed at early stages in the chicken embryo (see **Figure 14e-h**), CBS should be the enzyme responsible for the serine hydrolase activity detected in the chicken embryo liver [37]. By adding *in vitro* serine and *Gg*CBS to the other reaction components for taurine biosynthesis, the same consumption of cysteine (**Figure 15c**, left) produces twice as much taurine (**Figure 15c**, center) with complete consumption of serine (**Figure 15c**, right).

Origin and conservation of the sauropsidian pathway

CL is very similar to CBS although the enzymatic activity is unmistakably changed due to an accumulation of substitutions in the active site. *Gg*CL protein orthologs are present only in sauropsids, suggesting a birthplace of the protein family in this lineage. Phylogenetic analysis shows that CL sequences form an isolated monophyletic group within the CBS tree of vertebrates (**Figure 16**). In the maximum likelihood (ML) protein tree, the CL clade branches basally to teleostei, while in the ML nucleotide tree, CL branches basally to amniotes (**Figure 16a, b**).

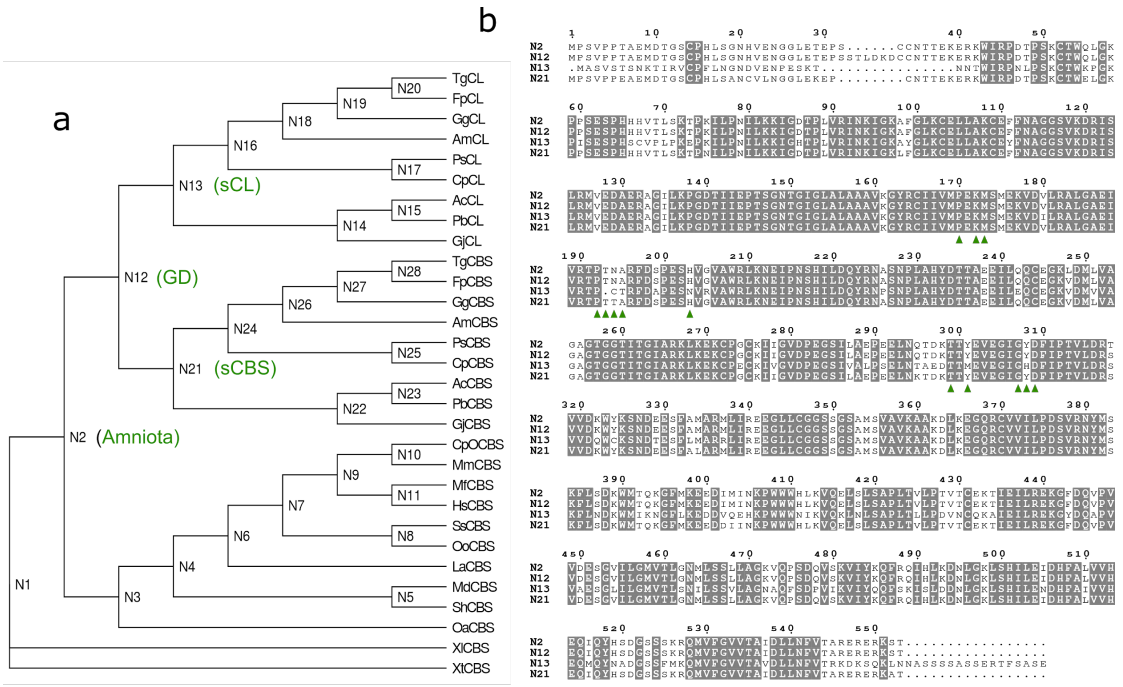


Name	Protein id	mRNA id	Taxonomy	Species
GgCBS	XP_416752.3	XM_416752.5	Sauropsida	*Gallus gallus*
GgCL	XP_015156382.1	XM_015300896.1	Sauropsida	*Gallus gallus*
TgCBS	XP_002189658.2	XM_002189622.3	Sauropsida	*Taeniopygia guttata*
TgCL	XP_012424887.1	XM_012569433.1	Sauropsida	*Taeniopygia guttata*
FpCBS	XP_027641626.1	XM_027785825.1	Sauropsida	*Falco peregrinus*
FpCL	XP_013154524.1	XM_013299070.2	Sauropsida	*Falco peregrinus*
AmCBS	XP_006260536.1	XM_006260474.3	Sauropsida	*Alligator mississippiensis*
AmCL	XP_019342804.1	XM_019487259.1	Sauropsida	*Alligator mississippiensis*
CpCBS	XP_005311024.1	XM_005310967.3	Sauropsida	*Chrysemys picta bellii*
CpCL	XP_023956308.1	XM_024100540.1	Sauropsida	*Chrysemys picta bellii*
PsCBS	XP_014436483.1	XM_014580997.1	Sauropsida	*Pelodiscus sinensis*
PsCL	XP_006137941.1	XM_006137879.2	Sauropsida	*Pelodiscus sinensis*
AcCBS	XP_003218996.1	XM_003218948.3	Sauropsida	*Anolis carolinensis*
AcCL	XP_008105670.1	XM_008107463.1	Sauropsida	*Anolis carolinensis*
PbCBS	XP_007435006.1	XM_007434944.2	Sauropsida	*Python bivittatus*
PbCL	XP_015744715.1	XM_015889229.2	Sauropsida	*Python bivittatus*
GjCBS	XP_015265299.1	XM_015409813.1	Sauropsida	*Gekko japonicus*
GjCL	XP_015265251.1	XM_015409765.1	Sauropsida	*Gekko japonicus*
HsCBS	NP_000062.1	NM_000071.2	Mammals	*Homo sapiens*
MfCBS	NP_001270081.1	NM_001283152.1	Mammals	*Macaca fascicularis*
MmCBS	NP_835742.1	NM_178224.3	Mammals	*Mus musculus*
CpoCBS	XP_003463971.2	XM_003463923.4	Mammals	*Cavia porcellus*
OoCBS	XP_004264645.1	XM_004264597.2	Mammals	*Orcinus orca*
SsCBS	XP_020926705.1	XM_021071046.1	Mammals	*Sus scrofa*
LaCBS	XP_010596971.1	XM_010598669.2	Mammals	*Loxodonta africana*
ShCBS	XP_012399820.1	XM_012544366.2	Mammals	*Sarcophilus harrisii*
MdCBS	XP_007493261.1	XM_007493199.2	Mammals	*Monodelphis domestica*
OaCBS	XP_007665408.2	XM_007667218.2	Mammals	*Ornithorhynchus anatinus*
XlCBS	NP_001086820.1	NM_001093351.1	Amphibia	*Xenopus laevis*
XtCBS	NP_001008171.1	NM_001008170.1	Amphibia	*Xenopus tropicalis*
DrCBS	NP_001104702.1	NM_001111232.1	Teleostei	*Danio rerio*
TrCBS	XP_003966607.1	XM_003966558.2	Teleostei	*Takifugu rubripes*
MzCBS	XP_004570219.1	XM_004570162.5	Teleostei	*Maylandia zebra*
OmCBS	XP_021464144.1	XM_021608469.1	Teleostei	*Oncorhynchus mykiss*
CmCBS	XP_007906994.1	XM_007908803.1	Chondrichthyes	*Callorhinchus milii*

Figure 16. *Unrooted maximum-likelihood (ML) trees obtained from nucleotide and protein multiple alignments of 35 CL and CBS sequences from 26 vertebrate species. Nucleotide and protein accession numbers corresponding to tree tip labels are indicated and the sauropsidian sequences are colored in green. Scale bars indicate the number of calculated substitutions per site. (**a**) Protein ML tree (436 alignment patterns) showing branching of the CL clade basal to teleostei. (**b**) Nucleotide ML tree (1277 alignment patterns) showing branching of the CL clade basal to amniotes. (**c**) Third codon position ML tree (613 alignment patterns) showing branching of the CL clade within Sauropsida [25].*

These phylogenetic reconstructions are muddled by contrasts in evolutionary rates, and possible long-branch attraction (LBA) artifacts causing attraction of the fast-evolving clade (CL) towards the basal clades [38]. The examination of the protein and nucleotide trees proposes that rate differences among CL and CBS are because of amino acid substitutions. The tree got with the first and second (~non-synonymous) codon positions was similar to the

protein tree (not shown), while the tree obtained with the third (~synonymous) codon position showed decreased differences in branch lengths and the expected sister relationship of sauropsidian CBS and CL clades (**Figure 16c.** All the sauropsidian genomes have the CBS-CL locus found in *Gallus gallus* (see **Figure 14m**) in conserved synteny, while it is absent in non-sauropsidian species (**Figure 16**).



	193		194		195		203		301		308	
N2 (Amniota)	T	0.99936	N	0.54534	A	0.99914	H	0.99992	Y	0.99988	Y	0.99992
N12 (GD)	T	0.99934	N	0.59071	A	0.99910	H	0.99990	Y	0.99986	Y	0.99990
N13 (sCL)	-	0.99000	C	0.99937	T	0.98614	N	0.99680	M	0.99757	H	0.99777
N21 (sCBS)	T	0.99998	T	0.99711	A	0.99996	H	0.99998	Y	1.00000	Y	1.00000

Figure 17. *(**a**) Evolutionary dendrogram used in ancestral state reconstructions assuming split of amniote last common ancestor (Amniote; N2) into two lineages before the gene duplication (GD; N12) leading to sauropsidian CL (sCL; N13) and CBS (sCBS; N21). Sequence identifiers are the same as in Figure 16. (**b**) Multiple sequence alignment of reconstructed ancestral sequences corresponding to nodes N2, N12, N13, and N21. Positions with identical residues in the four nodes and human CBS are shaded gray. Active site residues are marked with green triangles. Numeration follows the human CBS sequence. (**c**) Character state probabilities for active site residues substituted in GgCL showing high probability of fixation before the split of extant sauropsids [25].*

This confirms a tandem duplication of CBS gene caused the evolution of CL, occurred after the separation of synapsids and sauropsids lineages in the late Paleozoic, c.a. 300 MYA. The neo duplicated gene diverged by active site mutations in a complete neofunctionalization process, followed by an adaptation of CSAD ortholog for the cysteic acid to the new specificity (co-

option, **Figure 18**). These processes occurred before split of the main sauropsidian clades (**Figure 17**) since the presence of both genes in all sauropsids which are expected to have conserved the alternative pathway for taurine biosynthesis.

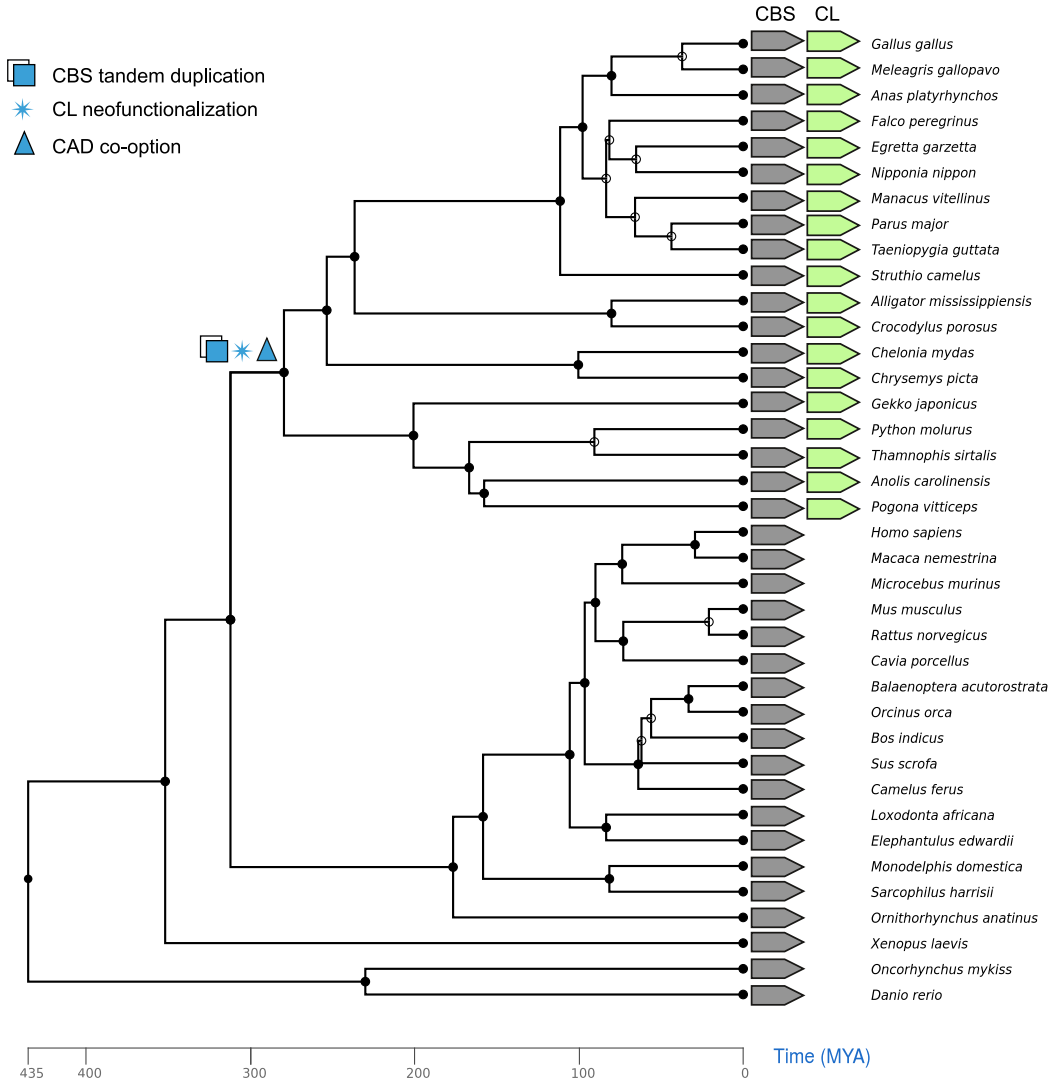


Figure 18. *Key evolutionary events in the origin of the metabolic pathway are mapped on a vertebrate chronogram. Phylogenetic relationships and divergence times are obtained from TimeTree (Kumar et al., 2017); transparent nodes correspond to unresolved relationships in NCBI taxonomy. The organization of the CBS locus in the species is represented at the terminal nodes, showing conservation of CBS-CL synteny in sauropsids and not in synapsids [25].*

Discussion

We have presented evidence for a route to taurine biosynthesis that originated in the common ancestor of birds and reptiles. This finding reveals that vertebrate adaptation to terrestrial habitats - particularly embryonic development in the cleidoic egg - also entailed the evolution of novel enzymes and metabolic pathways. Our investigation started with the aim of filling a gap in knowledge about an enzymatic activity (cysteine lyase) whose existence in the *Gallus* embryo has been suggested by early biochemical evidence, but for which a gene was not identified. The recognition of such common cases of enzymatic functions without assigned genes (pathway holes) is facilitated by the availability of structured information in public digital resources of enzymes (e.g. Enzyme: https://enzyme.expasy.org; Brenda: https://www.brenda-enzymes.info) and metabolic pathways (e.g. Kegg: https://www.genome.jp/kegg; Metacyc: https://metacyc.org).

Knowledge of the requirement of a particular cofactor (PLP) for the enzymatic activity has been the key to the discovery of the CL gene since it allowed restriction in the search to a limited number of proteins expected to use that cofactor. Comparison of the complete catalog of PLP-dependent proteins (PLPome) in two vertebrate species (see **Figure 3**) carried out in the light of the activity distribution among organisms and the reaction mechanism provided enough evidence for the identification of a single candidate gene for the sought function. Such *in silico* PLPome comparisons - aided by the bioinformatics tools of the specialized B6 database described here - could be used to identify genes responsible for other unassigned PLP-dependent activities. However, a similar search strategy can be extended to other cofactors where complete and accurate catalogs of the dependent proteins are available. Pertinently, a field reporting the presence of a cofactor (if any) is provided in enzyme digital records [39].

The experimental characterization of the CL candidate provided clear-cut evidence for its functional assignment and revealed the presence in the protein of a heme prosthetic group in addition to PLP. The presence of heme was unnoticed in studies with the partially purified protein [9], but could be anticipated by the homology of the identified CL sequence to CBS, so far the only PLP-dependent protein known to contain a heme group [32,40]. As in CBS, heme is bound to a separate domain of the protein and not directly involved in PLP-dependent catalysis. Although the presence of heme in CL can be explained by its phylogenetic origin, it is possible that CBS and CL are subjected to a common heme-mediated mechanism of regulation and sensing. By means of the heme, CBS is regulated by the redox state and binding of carbon monoxide to the ferrous atom of the prosthetic group [41,42]. While CBS and CL

catalyze different reactions, they both generate H_2S, a molecule with an emerging role as gaseous messenger in vertebrates [43].

The identification of the cysteic acid decarboxylase (CAD) as the second enzyme of the taurine pathway that can convert the CL reaction product into taurine has been an unexpected outcome of our experimental validation. No such enzyme has been previously described in literature. The *Gallus* ortholog of mammalian cysteine sulfinate decarboxylase (CSAD) was tested given the described ability of CSAD to act as decarboxylase towards CA as secondary substrate. Surprisingly, this enzyme revealed a strong specificity for CA. This evolutionary c shift of preference towards the CL reaction product (see **Figure 10**) gives independent support to the physiological relevance of the alternative sauropsidian pathway for taurine biosynthesis and further suggests that the standard mammalian pathway (see **Figure 15**) is of limited relevance for the production of taurine in sauropsids despite the maintenance of a CDO gene in their genomes that encode for the first enzyme of this pathway.

Evidence in various vertebrate species indicates that taurine has a vital role in adult and fetal life [15,44]. Taurine is essential in mammalian fetal development. In humans, reduced placental taurine transport causes intrauterine growth restriction and developmental deficits [45]. A taurine-restricted diet in pregnant cats, carnivorous animals which are deficient in taurine biosynthesis due to reduced hepatic activity of CDO and CSAD enzymes [46], results in abortion or delivery of growth-retarded kitten with impaired neurological function [44]. Mice disrupted in the gene encoding taurine transporter (taut-/-) show retinal degeneration and reduced offspring [47]. Dependency on mother's supply can extend to early postnatal development, with taurine being a major constituent in the milk of several mammalian species, including humans [48]. Data presented in this work show that in contrast to most mammals, in which taurine is provided during development by maternal transfer, sauropsids have a specific pathway for taurine biosynthesis whose genes are actively expressed during embryo development in the egg (see **Figure 14**).

While some physiological roles of taurine in specific organs such as e.g. retina, brain, muscles, and kidneys have been investigated only in mammals, there is direct evidence of the importance of taurine in the formation of bile salts in sauropsids [49,50]. Only taurine conjugates are found in chicken bile, owing to the inability of the *Gallus* BAAT enzyme to use glycine in substitution of taurine in the conjugation reaction [51]. During chicken embryo development, bile salts aid utilization of lipids, the major source of energy stored in the egg yolk [52]. Fats are absorbed through the yolk sac membrane, extraembryonic structure surrounding the yolk, by specialized endodermal cells containing bile and lipases [53]. Expression of the CL gene in the extraembryonic endoderm during early embryonic development (see **Figure 14a-d**) suggests that conjugation of bile acids for fat digestion is a major use of the taurine produced by the metabolic pathway. However, the more pleiotropic

expression of the CSAD gene, especially during later development (see **Figure 14i-l**), is consistent with a broader physiological role of taurine in different organs as observed in mammals.

Given that synthesis of taurine is required for embryo development in the reptilian egg, a relevant question is why sauropsids evolved a different pathway. The main novelty in the sauropsidian pathway is that oxidized sulfur (SO_3^{2-}) is used to replace cysteine thiol instead of oxidizing it through the CDO reaction. Previous radiotracer experiments have established that also sulfate (SO_4^{2-}) is eventually incorporated into taurine by the chicken egg through enzymatic conversion to sulfite [6]. Sulfate and sulfite are originated by spontaneous oxidation of the reduced sulfur present in the cell, but the animal cells are unable to reduce this sulfur compound to sulfide and incorporate it into other biomolecules. For this reason, in oviparous amniotes all the reduced sulfur required for the embryo development is in advance stored in the egg. The everyday use of unfertilized eggs reveals this sulfur content, for example with the release of H_2S during cooking contributes to the characteristic egg flavor or with the formation of FeS precipitate in the green yolk surface of hard-boiled eggs [54]. The CL pathway represents a more wasteless way to biosynthesize taurine by finding a use for oxidized sulfur that otherwise would be a junk product of cellular metabolism. In addition, functional and phylogenetic links between CBS and CL support the existence of a reduced sulfur cycle in the sauropsidian egg allowing the reuse of H_2S in amino acids (see **Figure 15**). Origin and maintenance of this pathway in the sauropsidian lineage suggest that the need to complete development in a self-contained life-supporting structure imposes a selective pressure on embryo metabolism for efficient use of growth-limiting resources.

For the possibility to use DNA and proteins as "documents of evolutionary history" [55], analysis of genes involved in the metabolic pathway provided insight into the mechanisms by which this pathway originated in sauropsidian ancestors. The initial event has been identified in the segmental duplication of the CBS gene. Following a typical scheme of neofunctionalization, one copy retained the original function while the other developed a novel catalytic ability through accumulation of molecular changes involving one deletion and five substitutions in conserved active site residues (see **Figure 17**). CL neofunctionalization could have immediately offered the possibility to proto reptiles to synthesize taurine via a different pathway by exploiting a secondary activity of an existing enzyme (CSAD). The gene encoding this enzyme has been eventually co-opted without duplication (True, JR 2002) for the new pathway by promoting a secondary activity (CAD) to the main one. Interestingly, this tuning of substrate specificity occurred with substitution of residues located externally to the active site (see **Figures 12, 13**). Origin of a novel pathway for taurine biosynthesis through a single gene duplication and a few point modifications of existing enzymes demonstrates that innovation of metabolic traits in vertebrates can arise from subtle genomic changes that less

specific analyses can easily overlook. Conceivably, other new metabolic pathways could be identified by targeted comparisons of vertebrate genomes.

Methods

All the methods and materials used are the same as those reported in the publication Malatesta et al. [25]

In silico analysis

The *Gallus gallus* and *Homo sapiens* PLPomes side-by-side comparison was performed with the "whole genome analysis" tool available in the B6db (http://bioinformatics.unipr.it/B6db) using the options "exclude isoforms" and "highlight BRH" for an easier identification of the common or unique genes. We perform further comparisons by extending the same procedure to other sauropsids (e.g. *Anolis carolinensis*) or other mammals (e.g. *Mus musculus*) to confirm the conservation of unique genes found in the first one in their respective taxonomic classes.

Substrate-binding cavities of *Gallus* enzymes were obtained by using the experimentally solved structure of human CBS (PDB: 4L3V) and human CSAD (PDB: 2JIS) analyzed through cavity computation by CAVER Analyst 2.0 by setting for CBS and CSAD respectively a Large Probe of 3.00 Å and 4.00 Å, and a Probe of 2.50 Å and 2.80 Å. The residues found in the cavity prediction were highlighted in the multiple sequence alignments (see **Figures 5, 12**) using ESPript 3.0.

Molecular phylogeny

Protein sequences were downloaded from NCBI and aligned with Clustalw 2.1. To obtain coding sequence (CDS) alignments, CDS were extracted from the corresponding mRNA sequences using ORFfinder 0.4.3 with the options "-s 0 -ml 1000 -strand plus -outfmt 1" and aligned based on amino acid alignment with macse v2.03. Phylogenetic trees were built with RAxML v. 7.7.8 using the GTR amino acid substitution matrix with optimization of substitution rates and GAMMA model of rate heterogeneity. The partitioning of codon positions was specified in a partition file to generate separated alignment sets for the first and second codon positions and third codon position using the option '-f s'. Maximum-likelihood reconstruction of ancestral character states including insertions/deletions [56] was obtained with the FastML web server (http://fastml.tau.ac.il/) based on extant CBS and CL sequences and a phylogenetic tree assuming CBS duplication in the sauropsidian ancestor.

Embryo Collection and In Situ Hybridization

Fertile chicken eggs (HyLine, Iowa; not a commercially available source) were incubated in a humidified incubator at 37.5 °C for 0.5 to 5 days. Embryos were collected into chilled chick saline (123 mM NaCl), removed from the vitelline membrane, and cleaned of yolk. Extra-embryonic membranes and large body cavities (brain vesicles, atria, allantois, eye) were opened to minimize trapping of the in-situ reagents. Embryos were fixed overnight at 4°C in freshly prepared 4% paraformaldehyde in PBS, washed twice briefly in PBS plus 0.1% Triton X-100 then dehydrated through a graded MEOH series and stored at -20 °C overnight in 100% MEOH. cDNA templates for generating all antisense RNA probes were obtained by reverse transcriptase-polymerase chain reaction using pooled RNA from embryos between HH stages 4 and 30. Primer sequences were designed using the mRNA sequence in the NCBI database. Embryo processing, antisense RNA probe preparation and whole-mount ISHs were performed as described [57]. A detailed protocol is available for download at http://geisha.arizona.edu.

Vector construction

The *Gg*CL expression vector were constructed with the LOC418544 (NCBI GeneID: 418544) CDS sequence (XM_015300896) cloned into pcDNA3.1+/C-(K)DYK vector and acquire from GenScript (USA Inc.). The coding region of the vector was subsequently amplified using CBSL_Fw, CBSL_Rev primers (for the native form of *Gg*CL) or CBSL_Fw, CBSL_short_Rev (for the truncated form of *Gg*CL) by PCR, with the use of Phusion polymerase, and then subcloned into pET-28b expression vector at NdeI/XhoI sites, generating respectively pET-28b-native*Gg*CL and pET-28b-truncated*Gg*CL. The *Gg*CAD expression vectors (WT, Q467V, T470A, Q467V-T470A) were constructed with *Gg*CAD wild-type sequence (NCBI GeneID: 426184) and acquired from GenScript (USA Inc.) already subcloned into pET-28b expression vector. All the constructs were transformed by electroporation into *E. coli* BL21 Codon Plus strain and the authenticity of all vectors was verified by next-gen sequencing.

Protein expression and purification

All the protein expressions were performed by inoculating a colony of each clone in one Liter of autoinducing LB broth that contains 2 g/L lactose and 0,5 g/L glucose and the other compounds of the standard LB. Cells were incubated for 16h at 30°C (*Gg*CBS, *Gg*CL), or for 16h at 20°C after a pre-induction phase of 8h at 30°C (*Gg*CAD). Cells were harvested and resuspended in 50 mL of Lysis Buffer (NaH_2PO_4 20 mM pH 7.0, NaCl 100 mM, 20 µM PLP), sonicated (1s on/off at 40 W for 0.5h) and centrifuged for 40 minutes at 14,000 rpm. After the

centrifugation, supernatants were purified with Affinity Chromatography (AC) on AKTA Pure system FPLC using the HisTrap 5 mL FF column. Proteins were eluted with a gradient of 7CV of AC Elution Buffer (NaH_2PO_4 20 mM pH 7.0, NaCl 100 mM, 20 μM PLP, 500 mM imidazole). *Gg*CL was further purified with Cation Exchange Chromatography (CIEX) using HiTrap SP FF column; the fractions obtained with AC were collected and diluted in 50 mL of Loading Buffer (MES 20 mM pH 6.5, 20 μM PLP) and eluted in 7CV gradient of CIEX Elution Buffer (MES 20 mM pH 6.5, 1 M NaCl, 20 μM PLP). Protein fractions (**Figure 6c, 9c**) obtained were concentrated with Vivaspin™ centrifugation to proceed with a final purification step using a Size Exclusion Chromatography (SEC) on a Superdex 200 column using SEC Buffer (NaH_2PO_4 20 mM pH 7.0, NaCl 100 mM). *Gg*CAD fractions after the same AC of *Gg*CL were diluted in 50 mL of Loading Buffer (NaH_2PO_4 20 mM pH 8.0, 20 μM PLP) for a further Anion Exchange Chromatography (AIEX) using HiTrap Q FF column, and eluted in 7CV gradient of AIEX Elution Buffer (NaH_2PO_4 20 mM pH 8.0, 1 M NaCl, 20 μM PLP).

UV-Visible and fluorescence spectroscopy

The measurement of the absorption spectra of the purified enzymes and the determination of kinetic parameters of *Gg*CL was performed with JASCO spectrophotometer. The quantification of *Gg*CL was performed by monitoring the absorbance at 428 nm (Soret peak) and using 84900 M^{-1} cm^{-1} as extinction coefficient, previously determined for *Hs*CBS [31]. The quantification of all the orthologous CSAD proteins purified, was performed by monitoring the absorbance at 280 nm and using the corresponding molar extinction coefficients computed with ProtParam (57410 M^{-1} cm^{-1} *Gg*CAD, 72880 M^{-1} cm^{-1} *Dr*CSAD, 61880 M^{-1} cm^{-1} *Hs*CSAD) H_2S release due to CL reactions was followed with time course kinetic at 390 nm using the reported extinction coefficient of PbS 5500 M^{-1} cm^{-1} [58]. Velocities considered for the Michaelis-Menten fitting correspond to the maximum speed of the enzyme, but they did not correspond to the initial velocities of the kinetics, due to a slight delay (see **Figure 7a**). Fitting to the Michaelis-Menten equation of the points were performed with SigmaPlot 14.0.
The presence of PLP cofactor in the enzymes was assessed by fluorescence spectroscopy [59] performed on a FluoroMax-3 spectrofluorometer (HORIBA Jobin Yvon, Kyoto, Japan) set at 20 °C with the emission slits set to 7 nm an integration time of 0.6 seconds. *Gg*CL measurement was obtained with 40 μM of the enzyme in 20 mM NaH_2PO_4 pH 7.0, 100 mM NaCl recorded between 425 nm and 600 nm using an excitation wavelength of 412 nm; *Gg*CAD measurement was obtained with 20 μM of the enzyme in 20 mM NaH_2PO_4 pH 8.0, 100 mM NaCl recorded between 355 nm and 500 nm using an excitation wavelength of 340 nm.

NMR Spectroscopy

^{1}H NMR spectra were acquired with a JEOL ECZ600R spectrometer at 25°C. Sample were loaded in tubes in the presence of 50 mM NaH_2PO_4 pH 7.0 to avoid signals of organic buffers in a final volume of 600 µL (H_2O:D_2O 9:1). For all the spectra measurement was used simple DANTE presat sequence for H_2O suppression.

Bibliography

1 Kumar, S. *et al.* (2017) Timetree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819

2 Reisz, R.R. and Müller, J. (2004) Molecular timescales and the fossil record: a paleontological perspective. *Trends Genet.* 20, 237–241

3 Sander, P.M. (2012) Paleontology. Reproduction in early amniotes. *Science* 337, 806–808

4 Pough, F.H. *et al.* (2013) *Vertebrate life*, (9th edn) Pearson.

5 International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695–716

6 Chapeville, F. and Fromageot, P. (1957) Formation de sulfite, d'acide cystéique et de taurine à partir de sulfate par l'oeuf embryonné. *Biochimica et biophysica acta* 26, 538–558

7 Bennett, N. (1973) Study of yolk-sac endoderm organogenesis in the chick using a specific enzyme (cysteine lyase) as a marker of cell differentiation. *J. Embryol. Exp. Morphol.* 29, 159–174

8 Fisher, J.-L. *et al.* (1982) La localisation de la cystéine lyase et le plan d'organisation du développement embryonnaire chez divers représentants de Vertébrés. *Biol Cell* 46, 291–300

9 Tolosa, E.A. *et al.* (1969) Reactions catalysed by cysteine lyase from the yolk sac of chicken embryo. *Biochim. Biophys. Acta* 171, 369–371

10 Machlin, L.J. *et al.* (1955) The utilization of sulfate sulfur for the synthesis of taurine in the developing chick embryo. *J. Biol. Chem.* 212, 469–475

11 Huxtable, R.J. (1996) Taurine. Past, present, and future. *Adv. Exp. Med. Biol.* 403, 641–650

12 Schaffer, S. *et al.* (2000) Role of osmoregulation in the actions of taurine. *Amino Acids* 19, 527–546

13 Hansen, S.H. *et al.* (2010) A role for taurine in mitochondrial function. *J. Biomed. Sci.* 17 Suppl 1, S23

14 El Idrissi, A. (2008) Taurine increases mitochondrial buffering of calcium: role in neuroprotection. *Amino Acids* 34, 321–328

15 Ripps, H. and Shen, W. (2012) Review: taurine: a "very essential" amino acid. *Mol. Vis.* 18, 2673–2686

16 Falany, C.N. *et al.* (1994) Glycine and taurine conjugation of bile acids by a single

enzyme. Molecular cloning and expression of human liver bile acid CoA:amino acid N-acyltransferase. *J. Biol. Chem.* 269, 19375–19379

17 Ebels, I. *et al.* (1980) Biosynthesis of taurine by rat pineals in vitro. *J. Neural Transm.* 48, 101–117

18 Ye, S. *et al.* (2007) An insight into the mechanism of human cysteine dioxygenase. Key roles of the thioether-bonded tyrosine-cysteine cofactor. *J. Biol. Chem.* 282, 3391–3402

19 Cavallini, D. *et al.* (1976) A new pathway of taurine biosynthesis. *Physiol. Chem. Phys.* 8, 157–160

20 Dominy, J.E. *et al.* (2007) Discovery and characterization of a second mammalian thiol dioxygenase, cysteamine dioxygenase. *J. Biol. Chem.* 282, 25189–25198

21 Veeravalli, S. *et al.* (2020) Flavin-Containing Monooxygenase 1 (FMO1) Catalyzes the Production of Taurine from Hypotaurine. *Drug Metab. Dispos.* DOI: 10.1124/dmd.119.089995

22 Baseggio Conrado, A. *et al.* (2015) Oxidation of Hypotaurine and Cysteine Sulfinic Acid by Peroxidase-generated Reactive Species. *Adv. Exp. Med. Biol.* 803, 41–51

23 Sumizu, K. (1962) Oxidation of hypotaurine in rat liver. *Biochim. Biophys. Acta* 63, 210–212

24 Percudani, R. and Peracchi, A. (2009) The B6 database: a tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families. *BMC Bioinformatics* 10, 273

25 Malatesta, M. *et al.* (2020) Birth of a pathway for sulfur metabolism in early amniote evolution. *Nat. Ecol. Evol.* 4, 1239–1246

26 Edgar, A.J. (2005) Mice have a transcribed L-threonine aldolase/GLY1 gene, but the human GLY1 gene is a non-processed pseudogene. *BMC Genomics* 6, 32

27 Tanaka, H. *et al.* (2011) Crystal structure of a zinc-dependent D-serine dehydratase from chicken kidney. *J. Biol. Chem.* 286, 27548–27558

28 Braunstein, A.E. *et al.* (1971) Specificity and some other properties of liver serine sulphhydrase: Evidence for its identity with cystathionine β-synthase. *Biochim. Biophys. Acta* 242, 247–260

29 Ereno-Orbea, J. *et al.* (2013) Structural basis of regulation and oligomerization of human cystathionine beta-synthase, the central enzyme of transsulfuration. *Proc.Natl.Acad.Sci.USA*

30 Antin, P.B. *et al.* (2014) GEISHA: an evolving gene expression resource for the chicken embryo. *Nucleic Acids Res.* 42, D933-7

31 Kery, V. *et al.* (1998) Trypsin cleavage of human cystathionine beta-synthase into an evolutionarily conserved active core: structural and functional consequences. *Arch. Biochem. Biophys.* 355, 222–232

32 Meier, M. *et al.* (2001) Structure of human cystathionine beta-synthase: a unique pyridoxal 5'-phosphate-dependent heme protein. *EMBO J.* 20, 3910–3916

33 Majtan, T. *et al.* (2014) Domain organization, catalysis and regulation of eukaryotic cystathionine beta-synthases. *PLoS ONE* 9, e105290

34 Leonardos, G. *et al.* (1969) Odor threshold determinations of 53 odorant chemicals. *J. Air Pollut. Control Assoc.* 19, 91–95

35 Agnello, G. *et al.* (2013) Discovery of a substrate selectivity motif in amino acid decarboxylases unveils a taurine biosynthesis pathway in prokaryotes. *ACS Chem. Biol.* 8, 2264–2271

36 Hamburger, V. and Hamilton, H.L. (1992) A series of normal stages in the development of the chick embryo. 1951. *Dev. Dyn.* 195, 231–272

37 Sentenac, A. and Fromageot, P. (1964) La sérinehydrolyase de l'oiseau mise en évidence dans l'embryon et mécanisme d'action. *Biochimica et Biophysica Acta (BBA) - Specialized Section on Enzymological Subjects* 81, 289–300

38 Bergsten, J. (2005) A review of long-branch attraction. *Cladistics* 21, 163–193

39 Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.* 28, 304–305

40 Singh, S. *et al.* (2007) Properties of an unusual heme cofactor in PLP-dependent cystathionine beta-synthase. *Nat. Prod. Rep.* 24, 631–639

41 Shintani, T. *et al.* (2009) Cystathionine beta-synthase as a carbon monoxide-sensitive regulator of bile excretion. *Hepatology* 49, 141–150

42 Taoka, S. *et al.* (1999) Characterization of the heme and pyridoxal phosphate cofactors of human cystathionine beta-synthase reveals nonequivalent active sites. *Biochemistry* 38, 2738–2744

43 Paul, B.D. and Snyder, S.H. (2012) H_2S signalling through protein sulfhydration and beyond. *Nat. Rev. Mol. Cell Biol.* 13, 499–507

44 Sturman, J.A. (1988) Taurine in development. *J. Nutr.* 118, 1169–1176

45 Norberg, S. *et al.* (1998) Intrauterine growth restriction is associated with a reduced activity of placental taurine transporters. *Pediatr. Res.* 44, 233–238

46 de la Rosa, J. and Stipanuk, M.H. (1985) Evidence for a rate-limiting role of cysteinesulfinate decarboxylase activity in taurine biosynthesis in vivo. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry* 81, 565–571

47 Heller-Stilb, B. *et al.* (2002) Disruption of the taurine transporter gene (taut) leads to retinal degeneration in mice. *FASEB J.* 16, 231–233

48 Rassin, D.K. *et al.* (1978) Taurine and other free amino acids in milk of man and other mammals. *Early Hum. Dev.* 2, 1–13

49 Haslewood, G.A. (1971) Bile salts of germ-free domestic fowl and pigs. *Biochem. J.* 123, 15–18

50 Tomaselli, S. *et al.* (2007) NMR-based modeling and binding studies of a ternary complex between chicken liver bile acid binding protein and bile acids. *Proteins* 69, 177–191

51 Center, C. and Francisco, S. Purification and characterization ofcholyl-CoA : taurine N-acetyltransferase from the liver of domestic fowl (Gallus gallus).

52 Noble, R.C. and Cocchi, M. (1990) Lipid metabolism and the neonatal chicken. *Prog. Lipid Res.* 29, 107–140

53 Yadgary, L. *et al.* (2013) Changes in yolk sac membrane absorptive area and fat digestion during chick embryonic development. *Poult. Sci.* 92, 1634–1640

54 Germs, A.C. (1973) Hydrogen sulphide production in eggs and egg products as a result of heating. *J. Sci. Food Agric.* 24, 7–16

55 Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *J. Theor. Biol.* 8, 357–366

56 Ashkenazy, H. *et al.* (2012) FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* 40, W580-4

57 Antin, P.B. *et al.* (2010) Embryonic expression of the chicken Krüppel-like (KLF) transcription factor gene family. *Dev. Dyn.* 239, 1879–1887

58 Chiku, T. *et al.* (2009) H2S biogenesis by human cystathionine gamma-lyase leads to the novel sulfur metabolites lanthionine and homolanthionine and is responsive to the grade of hyperhomocysteinemia. *J. Biol. Chem.* 284, 11601–11612

59 Salsi, E. *et al.* (2011) Exploring O-acetylserine sulfhydrylase-B isoenzyme from Salmonella typhimurium by fluorescence spectroscopy. *Arch. Biochem. Biophys.* 505, 178–185