UNIVERSITÀ DI PARMA

Università degli Studi di Parma

Department of Chemistry, Life Sciences
and Environmental Sustainability
Ph.D. program in Biotechnology and Biosciences
XXXV cycle

Identification of missing links in eukaryotic metabolism through co-evolutionary analysis

Coordinator:
Prof. Marco Ventura

Supervisor:
Prof. Riccardo Percudani

Ph.D student:
Elena Dembech

2019/2020 – 2021/2022

Summary

Aim of the research

This dissertation describes the experimental validations of a new bioinformatics tool that enables the identification of coevolving eukaryotic genes by comparing phylogenetic profiles, i.e. vectors that describe the presence/absence of genes in a set of genomes; the establishment of coevolutionary associations between genes could help elucidate their function, in particular in identifying component of structural complex or metabolic proteins.

Two eukaryotic gene families have been investigated computationally and experimentally, leading to the association of functions to genes with unknown or uncharacterized functions. They were found to participate in two metabolic pathways, the purine degradation pathway in Metazoa and the tryptophan catabolic pathway in vertebrates, respectively.

This dissertation is divided into three chapters. Chapter 1 provides an overview of the phylogenetic profiling approach, with the aim of introducing the reader to an understanding of the main issues of this evolution-based computational method applicable to the discovery of new molecular functions.

In Chapter 2, we validated our computational method through the identification of the *malate synthase-like* gene found in Metazoa as a missing ureidoglycolate lyase involved in the last step of purine degradation. In addition, we demonstrated the maintenance of malate synthase activity in some marine invertebrates that possess two copies of the *malate synthase* gene.

As ureidoglycolate lyase catalyzes the conversion of (*S*)-ureidoglycolate to glyoxylate and urea, we revealed a connection between the glyoxylate cycle and purine catabolism. We identified the missing ureidoglycolate lyase using phylogenetic profiles constructed using orthogroups, i.e. groups of orthologous genes, collected in Orthodb v.10. These two experimental validations allowed us to further investigate the evolution of malate synthase genes.

The results described in this chapter were obtained in collaboration with other researchers of the University of Parma and submitted to the peer-reviewed journal "Proceedings of the National Academy of Sciences" (PNAS)[1]; a proposal for a new name (ureidoglycolate lyase) and symbol (*ugl*) for the zebrafish gene has been submitted to the Zebrafish Information Network (ZFIN) database.

In Chapter 3, we demonstrated the importance of the correct orthogroups identification for the construction of phylogenetic profiles and for the prediction of the functional relationships between genes.

The unpublished results presented in this chapter concern the identification of the missing human 2-aminomuconate reductase, an enzyme involved in the tryptophan degradation pathway that catalyzes the reduction and deamination of 2-aminomuconate. In particular, we

have discovered that this enzymatic activity is carried out by the *ASPDH* gene, whose bacterial orthologs have been described to take part in the *de novo* NAD biosynthesis.

To detect significant correlations with tryptophan catabolism orthogroups that have helped the discovery of this unknown gene, it was necessary to build orthogroups *de novo* using the Diamond algorithm and then to create new phylogenetic profiles. The first method presented (Chapter 2) was indeed unable to identify a candidate gene for this pathway hole but we overcame this obstacle by considering novel human-centered orthogroups, resulting in the detection of *ASPDH a*s the missing gene.

Chapter 1

Overview of phylogenetic profiles

Homology and coevolution of protein-coding genes

Protein interaction is one of the most relevant biological networks and it is important to elucidate how the protein-protein interaction processes have diversified in living organisms over the course of evolution[2].

The mutual evolution of interacting biological entities can be observed at different levels, ranging from individual amino acid sites[3] to genome[4]. In the former case, if two residues are structurally in close contact, a single amino acid substitution requires mutation of the other amino acid in order not to perturb the functionality of the entire protein, leading to the coevolution of the two residues[5].

Additionally, genes encoding proteins that interact in a cellular complex, in signaling pathways, and in metabolic routes have been demonstrated to coevolve and we can have a trace of this event considering whole genomes. For example, Respiratory Complex I subunits illustrate the coevolution of genes involved in a cellular complex; in fact, it has been observed that genes responsible for NADH oxidation and for the proton pump tend to coevolve following a modular evolution[6]. A similar evolution has been demonstrated for CatSper genes, which have coevolved to preserve the structural conformation and control CatSper pH sensitivity among mammals[7]. Moreover, the coevolution of a common signaling pathway with intracellular endosymbioses[8] has been demonstrated in plants.

The more significant relationship between genes belonging to different species is homology: orthologous and paralogs genes can be used to build gene families based on the idea that orthologs are the same genes in different species and have evolved from a common ancestor, while paralogs have originated by duplication within the genome.

A gene family is a group of genes that have a genome sequence similarity and are derived from a common ancestral gene; phylogenetic trees of a gene family can be obtained through a variety of bioinformatics methods. By considering multiple complete genomes, it is possible to delineate highly conserved protein families in one or more domains of life[9] and identify coevolved pairs of organisms or biomolecules to understand living systems.

Gene-level coevolution occurs when the existence of a gene in a genome is correlated with the presence of other genes. Phylogenetic trees have always been used as a method to detect coevolution for many protein families using whole protein sequences[10]. The comparison of evolutionary histories of coevolved genes, i.e. the construction of phylogenetic trees, results in a similar topology of the corresponding trees[11].

Indeed, it is necessary to consider that phylogenetic inheritance could be a confounding factor in the analysis of eukaryotic gene coevolution, particularly when a gene takes part in several metabolic routes or have multiple copies in some species; additionally, the similarity between profiles due to a common vertical inheritance may give rise to misunderstandings in associating genes on the basis of their functional relation[12].

The construction of phylogenetic profiles, i.e. a method used to look for the co-occurrence of protein families among organisms, is essential to achieve the goal of studying protein evolution.

The inability to annotate all enzymes by homology-based methods leaves members of metabolic pathways uncovered, and related metabolic activities remain "orphans" and not assigned to any genomic sequence. There are also cases of "local orphans" for which the coding sequences have not been identified in an organism of interest, even though the gene responsible for the reaction has been identified in other organisms[13].

The computational method of phylogenetic profiling thus aims to reconstruct metabolic networks and fill in metabolic gaps.

Phylogenetic profiles

Phylogenetic profiling (PP) is an established bioinformatics method used to predict functional interactions between proteins. Proteins that act together tend to have coordinated evolution through the tree of life[14]. The phylogenetic profiling technique is used to identify protein-coding gene functions and test whether they are functionally related[15]. Predictions based on *in silico* approaches regarding functionally linked proteins also aim to assign a molecular function to uncharacterized components that have been found to be associated with known components[16]. In addition, PPs and the associated co-occurrence network generated could find application in identifying disease-protein associations[17,18].

Phylogenetic profiling was first applied using binary vectors, which indicate the presence or absence of homologous proteins in different organisms[19]. This innovative method was first applied for the analysis of microorganisms and later extended to all living kingdoms.

The application of this initial approach to PP has made it possible to establish relationships between proteins of different organisms, such as in *Saccharomyces cerevisiae* and other yeasts[20], in the model organism *Drosophila melanogaster* to study eukaryotic cilia and flagella[21], and to identify novel components involved in the mammalian meiotic methylation process in *Caenorhabditis elegans* and *Homo sapiens*[16].

The phylogenetic profiling method has been implemented to predict protein interactions and to understand the molecular mechanisms that determine their physical and functional linkage in living cells[22], as well as to trace their evolutionary history. About 80% of all existing eukaryotic genome projects involve genomic sequencing[23] of Fungi and Metazoa, with very little coverage of eukaryotes[24].

However, phylogenetic profiling has allowed making functional predictions for about 10-15% of the human genome; it has otherwise been demonstrated that human gene pairs having the highest coevolution scores are implicated in metabolism or have structural roles[25].

Reference genomes, orthogroup construction, and phylogenetic profiling

Over the past century, several efforts have been undertaken to understand whether there is a possibility of finding a single reference genome, but it has emerged that no single genome can represent diversity within a species. For this reason, the scientific community has paid attention to the identification of the pan-genome, which consists of a universally shared central genome added with a variable part and could be used to understand the complexity of living being genomes. Preliminary studies have described that the mammalian pan-genome is more closed, while the pan-genomes of other groups of organisms are broader[26]. Otherwise, the increase in sequencing capacity with NGS has greatly increased the understanding of the genome and has enabled the completion of genome sequencing of thousands of organisms[27]. Although the human genome was sequenced more than two decades ago, the biochemical and cellular function of many human genes remains unknown, and the availability of several genomes is necessary for the application of phylogenetic profiles. In fact, the expansion of the number of reference genomes should improve the accuracy of this method[22].

Binary phylogenetic profiles are usually generated by searching for homologous protein-coding genes in reference genomes. If a homolog exceeds the threshold, this information is inserted in a vector pair containing information about the presence and the absence of the gene in the genome. When the organism presents a sequence, the entry for the organism/genome is equal to one; if no homologs are found in the genome, the entry is zero. The presence or absence of a homolog in each reference organism is encoded in a string using binary values 1 and 0 and resulting in the construction of a binary vector for each orthogroup (Fig.1).

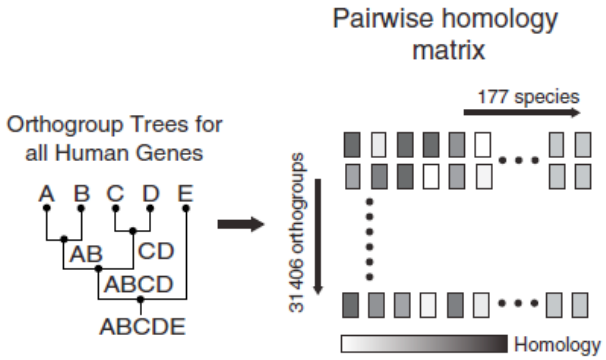


Figure 1. Orthogroup identification and pairwise matrix construction (adaptation from Dey et al., 2015[28]). An example of binary vector construction[28]; the procedure starts with the identification of orthogroups and is followed by the construction of vectors for each gene in all the organisms involved in the analysis.

Orthologous sequences must be correctly linked[29] in each species to create orthogroups, i.e. groups of orthologous genes in selected organisms; a pairwise BBH (bidirectional best hit, or BRH) approach can outperform sophisticated tree-based algorithms in the presence of large genomes and several organisms[30].
In addition to a de novo construction of orthogroups, datasets of previously identified orthologous genes collected in databases are freely available, such as OrthoDB[49].

An effective approach to constructing orthogroups was introduced by Dey and colleagues[28], who described a method to identify the closest homologous gene identified using the longest human protein-coding gene as the query for a homology search. They also confirmed the best result for human genes in other organisms by performing the BRH test with BLAST[28]. After having established the BRH for each organism, paralogs have been added to orthogroups.
In fact, the exclusion of duplicated genes from orthogroups could provide the most unbiased functional predictions[28] since two events of genome-wide duplications have led paralogs to diverge functionally during evolution[31].

In eukaryotes, conservation between two proteins can be quantified on a continuous scale from 0 to 100, and the choice of minimum threshold to identify orthologous and paralogous genes can result in the construction of different PP. The PP construction and the estimation of similarity between two gene families are affected by the requirements of strict sequence similarity and are indeed altered depending on the threshold[18,32,33].

The threshold and the relativity score can generate biases; several PPS methods have tried to reduce these biases using different approaches, such as using a continuous conservation scale instead of a cutoff[14].

Phylogenetic profiles and transitions between the presence and absence of genes in reference genomes also depend strongly on the number of species considered in the analysis and their order in the vector pair; larger datasets also allow linking related evolution events to background organisms and ecological variables[12].

Several approaches have been explored with the goal of extracting information from phylogenetic data, and PP has been widely used to predict protein function; methods that account for shared inheritance are suitable for Big Data analysis.

After identifying orthogroups using a proper algorithm, independent losses in multiple lineages can be used to model gene gains and losses and to compare PP. Matrix construction could be a useful method to represent the phylogenetic profile by considering the distances between orthologs and evaluating statistical parameters for describing gene coevolution. Parameter choice has been otherwise demonstrated to affect the detection of coevolution and the optimization of parameters[14].

Parsimony and maximum likelihood (ML) methods were explored to establish the correct phylogenesis of organisms; using the ML algorithm, Barker and colleagues built a phylogenetic tree of species and PPs based on an ordered vector following that phylogenesis[34]. They then calculated the number of negative and positive tests by comparing their dataset with a Yeast Database[35] and calculating specificity and sensitivity based on the number of true positives and false positives and negatives. This method appears to have greater accuracy and sensitivity and appropriately weights gene acquisitions and losses.

Phylogenetic profile comparison

Several methods have been improved to compare coevolving PP, such as Hamming distance[16,19], Pearson correlation coefficient[13,36], Jaccard similarity[37], Fisher's exact test[12], and others. Other normalizations included singular value decomposition (SVD)[38] and normalized phylogenetic profile (NPP)[33]. Here I will report a quick overview of the most robust methods used to validate the construction of PP.

The similarity between PP vector pairs has been frequently measured through Hamming distance (HD)[16,39], which corresponds to the number of positions that presents different entries

for two binary vectors, and it is used to score the similarity between profile pairs; the effectiveness of this method was validated with datasets of known human protein complexes[16] resulting in a robust approach for coevolution evaluation.

Quantifying the relationship between the distance of genes in a network and the average gene coevolution was also tested by Chen and colleagues[13], who calculated a distance metric and derived a linear Pearson correlation for all genes after establishing a background correlation between all non-metabolic and metabolic genes. They then calculated a heuristic score to optimize parameters by minimizing the log sum of the ranks for all metabolic enzymes. The performance of this algorithm actually depends on the completeness of the network in the dataset, and the authors reported that the algorithm requires the combination of multiple sequences- and context-based associations to have a more accurate assessment of significant pairwise associations[13].

Proposing a new metric based on distance correlation, another research group[14] demonstrated that different strategies can overcome Pearson's correlation. In a recent paper, PP vectors were constructed using BLASTP bit scores, given a set bit score threshold and normalizing bit scores on the value of the query against itself. In these vectors, low or missing bit scores indicate the lack of significant homologous genes, and the resulting matrix considers minimal changes during protein evolution. Standard scaling of the values was performed after being transformed using $\log_2$, and the resulting NPP matrix was subjected to Pearson correlation to identify pairwise coevolution. Comparison with other methods confirms that the choice of parameters for the description of coevolution can influence the performance of the method itself.

The Jaccard index, widely used to solve ecological problems, has been applied to phylogenetic profiling and several applications have been described in the literature. However, it has been found that the Jaccard is strongly influenced by track size and is consistently low for small datasets; furthermore, it has been proposed that this index can be used in the quantification of genomic co-occurrence in conjunction with other multiple alternative measures to avoid the typical limitations[40] of the method.

An alternative method based on the enumeration of shared "runs" or "transitions" has been investigated with the aim of scoring independent loss events[25]. These types of scoring metrics do not require the full pattern of gene presence and absence and could be applied to thousands of genes across phylogeny. Specifically, Dey and colleagues defined a weighted co-occurrence score based on the number of shared transitions with the goal of distinguishing between lower

and higher confidence[28]; they also took into account the mismatch penalty for each pairwise comparison, thus presenting an "unbiased genome-wide strategy for a species-centered phylogenetic loss analysis"[28].

The ROC (receiver operator characteristic) curve and the associated AUC (area under the curve) has been used to compare the performance of different methods[39,41]; ROC analysis allows the identification of optimal thresholds based on a given cost function and shows how sensitivity changes depending on specificity. Calculating the graphical AUC under the line segments (with the false positive rate on the x-axis and the true positive rate on the y-axis) helps to understand the predictive variables. PPP (PrePhyloPro) was found to be the most suitable method for predicting true positives (Fig.2A). This result was supported by calculating the precision and recall for each method (Fig.2B), which showed that some methods are more appropriate for finding true functional and structural linkages.

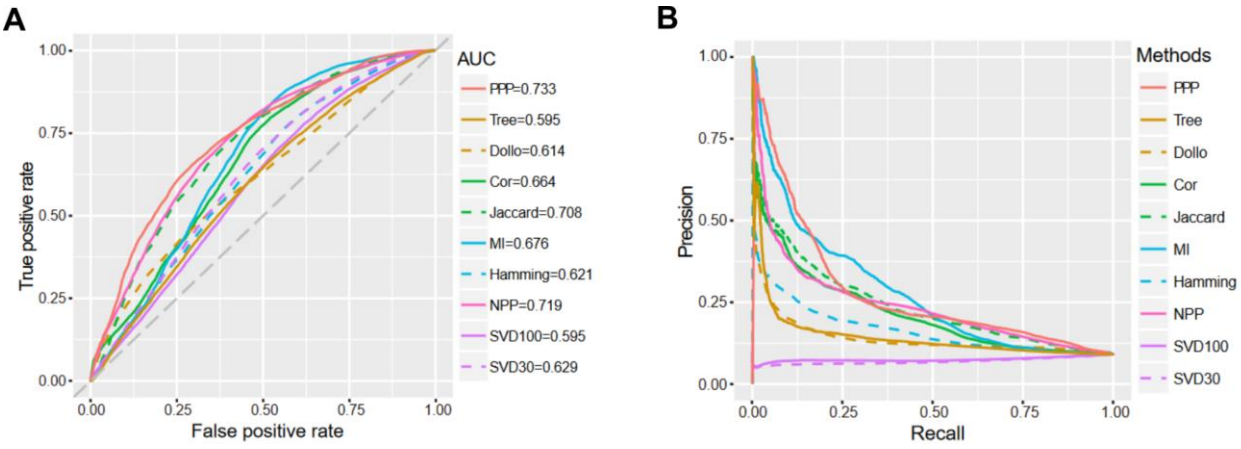


Figure 2. Comparison of method performance (image adaptation from Niu *et al.*, 2017)[39]. ROC curves (A) and PR curves (B) of PPP compared with other methods for performance evaluation.

Quality assessment was investigated by comparing SwissProt keywords[19], subcellular compartmentalization, or shared membership in protein complexes, and data from the analysis were often matched with Databases, such as KEGG[42], to validate and support the identification of functional relationships.

A supervised machine-learning approach for constructing phylogenetic profiles was introduced in 2021[43]; the authors of this innovative procedure utilized "clade-wise" coevolution of functionally related genes and focused their attention on the functional annotation of less-studied genes, considering from kingdom to species level. They computed a species-by-gene

matrix similar to other previously described methods; on the other hand, they introduced for the first time a binary classification model trained and supported by Reactome pathways and they compared real clades to randomized clades. In fact, this provides a practical alternative to full inference on phylogenetic trees. On the other hand, this method seems to have low performance in "young" genes and requires an implementation to focus on either small or large pathways.

Otherwise, a comparison of this novel method based on machine learning with previous ones demonstrated its robustness (Fig.3A) and its ability to associate genes involved in the same metabolic pathway (Fig.3B).

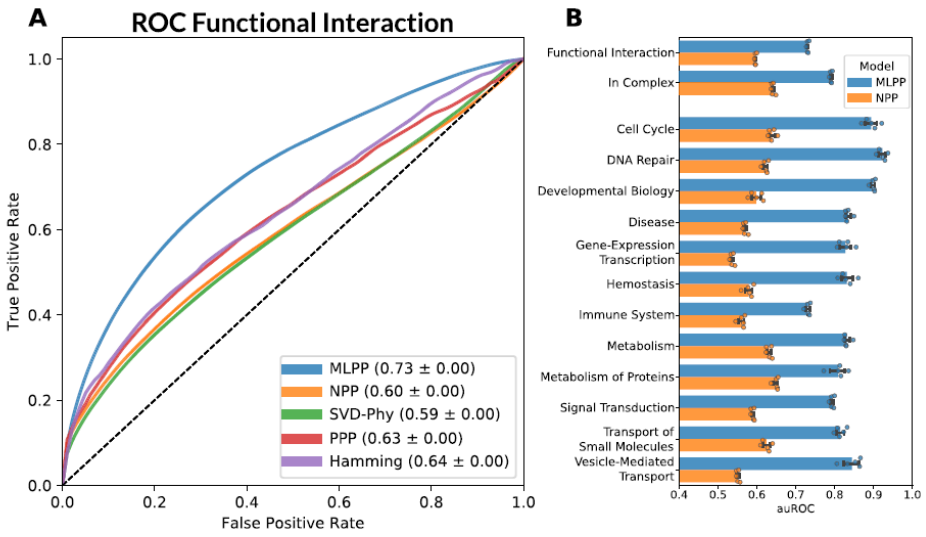


Figure 3. Method comparison for prediction of functional interactions (image adaptation from Stupp *et al.*, 2021[43]). (A) MLPP method compared against other phylogenetic profiling approaches: ROC curve and AUC area. (B) MLPP interaction context prediction from Reactome.

Despite numerous examples of phylogenetic profile construction and validation, it has yet to be determined which method is more effective and least dependent on the bias.

Module identification

Many PP methods have been implemented with the presence of IDs, functional annotation, and curated PubMed citations. Since biological networks are widely regarded to be intrinsically modular[36,44], different approaches could lead to the identification of distinct molecular clusters. It has been observed that, when comparing modules obtained by various methods, part of these modules converges and overlaps. Interestingly, the most reliable module associations are enriched for genes involved in metabolic pathways and functional complexes[25], and all the tested methods can detect closer links between members of known evolutionary networks. Some clusters are enriched for genes assigned to molecular pathways; other genes in the same cluster may function in related pathways[33].

One of the most effective approaches to identify modules containing related genes has proven to be ranking the most related genes using Pearson's correlation coefficient[33], followed by assessing the significance of modules among protein datasets, such as Molecular Signatures Database (MSigDB)[45] or Human Phenotype Ontology (HPO)[46,47].

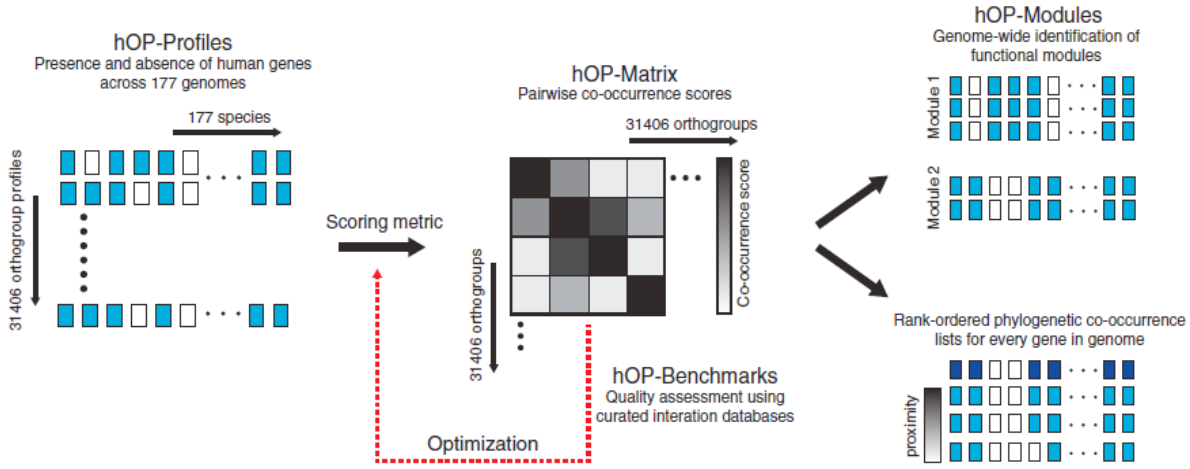


Figure 4. Strategy for phylogenetic profiling centered on human genes[28]. Profile construction and binary phylogenetic vectors based on human-centered orthogroup; establishment and optimization of co-occurrence scores and quality assessment with curated databases; identification of functional modules.

In addition, an unbiased genome-wide approach to identify modules has been described in the work carried out by Dey and colleagues. Indeed, they identified coevolutionary modules (hOP modules) composed of more than 2 profiles (Fig.4); they started by clustering pairs from significant comparisons and then added one-by-one the orthogroups that reached a threshold

and had a weighted average phylogenetic co-occurrence score relative to the previously added members. This approach allowed them to explore the modular architecture of specific cellular processes and to identify undiscovered members of certain pathways or complexes.

As an alternative to the triangular linkage, the Markov Cluster algorithm (MCL)[48] can be useful for assembling groups of proteins to form functional clusters. The MCL approach is based on probability and graph flow theory; it allows simultaneous classification of global relationships and calculation of the probabilities of transitions between nodes.

In this dissertation, we propose a novel phylogenetic profiling metric to score and assess the significance of correlated presence/absence transitions between tree-ordered genomes, based both on a large dataset of eukaryotic orthologous genes from OrthoDB, grouped in modules on the basis of their co-occurrence, or applied to a *de novo* dataset of human-centered orthogroups built using Diamond software.

Bibliography

1. Dembech Elena, Malatesta Marco, De Rito Carlo, Mori Giulia, Cavazzini Davide, Secchi Andrea, Morandin Francesco, PercudanI Riccardo. Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions. (under review, 2023).
2. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* 14, 249–261 (2013).
3. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci.* 110, 15674–15679 (2013).
4. Sloan, D. B. *et al.* Cytonuclear integration and co-evolution. *Nat. Rev. Genet.* 19, 635–648 (2018).
5. Burger, L. & van Nimwegen, E. Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. *PLoS Comput. Biol.* 6, e1000633 (2010).
6. Gabaldón, T., Rainey, D. & Huynen, M. A. Tracing the Evolution of a Large Protein Complex in the Eukaryotes, NADH:Ubiquinone Oxidoreductase (Complex I). *J. Mol. Biol.* 348, 857–870 (2005).
7. Hwang, J. Y. *et al.* Dual Sensing of Physiologic pH and Calcium by EFCAB9 Regulates Sperm Motility. *Cell* 177, 1480-1494.e19 (2019).
8. Radhakrishnan, G. V. *et al.* An ancestral signalling pathway is conserved in intracellular symbioses-forming plant lineages. *Nat. Plants* 6, 280–289 (2020).
9. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A Genomic Perspective on Protein Families. *Science* 278, 631–637 (1997).
10. Goh, C.-S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. Co-evolution of proteins with their interaction partners 1 1Edited by B. Honig. *J. Mol. Biol.* 299, 283–293 (2000).
11. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* 95, 6073–6078 (1998).
12. Barker, D. & Pagel, M. Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes. *PLoS Comput. Biol.* 1, e3 (2005).
13. Chen, L. & Vitkup, D. Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol.* 7, R17 (2006).
14. Bloch, I. *et al.* Optimization of co-evolution analysis through phylogenetic profiling reveals pathway-specific signals. *Bioinformatics* 36, 4116–4125 (2020).
15. Pellegrini, M. Using Phylogenetic Profiles to Predict Functional Relationships. in *Bacterial Molecular Networks* (eds. van Helden, J., Toussaint, A. & Thieffry, D.) vol. 804 167–177 (Springer New York, 2012).
16. Cheng, Y. & Perocchi, F. ProtPhylo: identification of protein–phenotype and protein–protein functional associations via phylogenetic profiling. *Nucleic Acids Res.* 43, W160–W168 (2015).
17. Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M. & Homouz, D. Analyzing a co-occurrence gene-interaction network to identify disease-gene association. *BMC Bioinformatics* 20, 70

(2019).
18. Tabach, Y. *et al.* Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. *Mol. Syst. Biol.* 9, 692 (2013).
19. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. S. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96 8, 4285–8 (1999).
20. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83–86 (1999).
21. Avidor-Reiss, T. *et al.* Decoding Cilia Function. *Cell* 117, 527–539 (2004).
22. Sun, J., Li, Y. & Zhao, Z. Phylogenetic profiles for the prediction of protein–protein interactions: How to select reference organisms? *Biochem. Biophys. Res. Commun.* 353, 985–991 (2007).
23. Richards, S. It's more than stamp collecting how genome sequencing can unify biological research. *Trends Genet.* 31, 411–421 (2015).
24. Dawson, S. C. & Fritz-Laylin, L. K. Sequencing free-living protists: the case for metagenomics. *Environ. Microbiol.* 11, 1627–1631 (2009).
25. Dey, G. & Meyer, T. Phylogenetic Profiling for Probing the Modular Architecture of the Human Genome. *Cell Syst.* 1, 106–115 (2015).
26. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* 21, 243–254 (2020).
27. McCombie, W. R., McPherson, J. D. & Mardis, E. R. Next-Generation Sequencing Technologies. *Cold Spring Harb. Perspect. Med.* 9, a036798 (2019).
28. Dey, G., Jaimovich, A., Collins, S. R., Seki, A. & Meyer, T. Systematic Discovery of Human Gene Function and Principles of Modular Organization through Phylogenetic Profiling. *Cell Rep.* 10, 993–1006 (2015).
29. Dalquen, D. A. & Dessimoz, C. Bidirectional Best Hits Miss Many Orthologs in Duplication-Rich Clades such as Plants and Animals. *Genome Biol. Evol.* 5, 1800–1806 (2013).
30. Kristensen, D. M., Wolf, Y. I., Mushegian, A. R. & Koonin, E. V. Computational methods for Gene Orthology inference. *Brief. Bioinform.* 12, 379–391 (2011).
31. Blomme, T. *et al.* The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* 7, R43 (2006).
32. Enault, F., Suhre, K., Poirot, O., Abergel, C. & Claverie, J.-M. Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res.* 32, W336–W339 (2004).
33. Tabach, Y. *et al.* Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* 493, 694–698 (2013).
34. Barker, D., Meade, A. & Pagel, M. Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* 23, 14–20 (2007).
35. Guldener, U. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.* 33, D364–D368 (2004).
36. Glazko, G. V. & Mushegian, A. R. Detection of evolutionarily stable fragments of cellular

pathways by hierarchical clustering of phyletic patterns. *Genome Biol.* 5, R32 (2004).
37. Brilli, M., Fani, R. & Lio, P. Current trends in the bioinformatic sequence analysis of metabolic pathways in prokaryotes. *Brief. Bioinform.* 9, 34–45 (2007).
38. Franceschini, A., Lin, J., von Mering, C. & Jensen, L. J. SVD-phy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics* 32, 1085–1087 (2016).
39. Niu, Y., Liu, C., Moghimyfiroozabad, S., Yang, Y. & Alavian, K. N. PrePhyloPro: phylogenetic profile-based prediction of whole proteome linkages. *PeerJ* 5, e3712 (2017).
40. Salvatore, S. *et al.* Beware the Jaccard: the choice of similarity measure is important and non-trivial in genomic colocalisation analysis. *Brief. Bioinform.* 21, 1523–1530 (2020).
41. Muschelli, J. ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. *J. Classif.* 37, 696–708 (2020).
42. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30 (2000).
43. Stupp, D. *et al.* Co-evolution based machine-learning for predicting functional interactions between human genes. *Nat. Commun.* 12, 6454 (2021).
44. Luo, F. *et al.* Modular organization of protein interaction networks. *Bioinformatics* 23, 207–214 (2007).
45. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* 1, 417–425 (2015).
46. Altenhoff, A. M. *et al.* OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* 49, D373–D379 (2021).
47. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* 49, D1207–D1217 (2021).
48. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13, 2178–2189 (2003).
49. Zdobnov, E. M. *et al.* OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 49, D389–D393 (2021).

Chapter 2

Identification of a hidden evolutionary association among purine degradation pathway and glyoxylate cycle

Introduction

Purine catabolism

Nucleotide metabolism, and in particular degradations of purines to urate, is an ancient and indispensable set of metabolic reactions which operates in all living organisms; it is characterized by extreme variability across species and its final products vary from species to species depending on their need to recycle carbon, and especially nitrogen[1] (Fig.1).

Purines are a structural component of several coenzymes and are required for the cells' growth, proliferation, and survival.

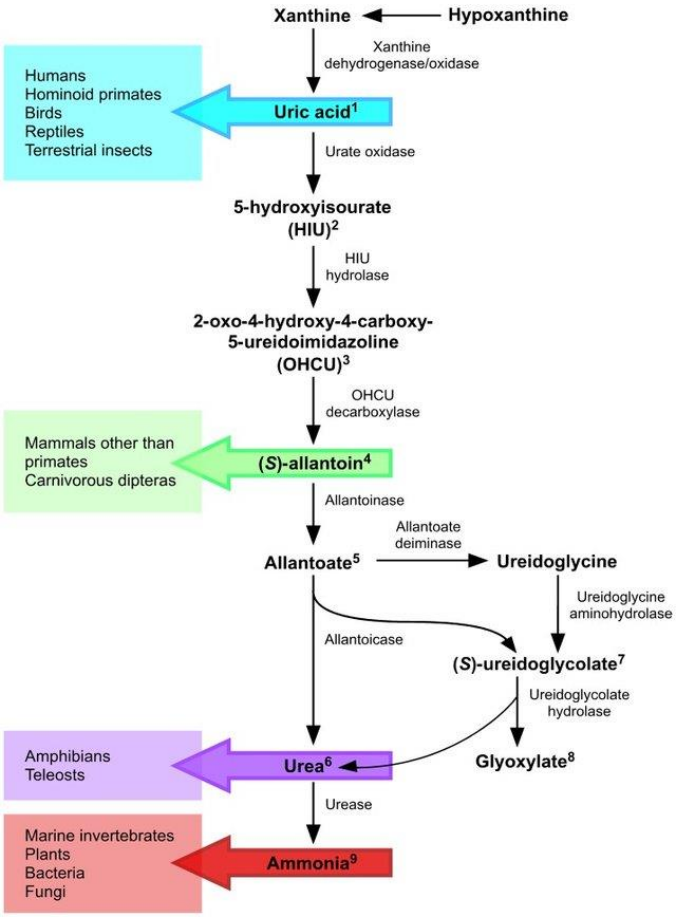


Figure 1: Purine degradation pathway[1]. Reactions of purine catabolism with emphasis on the end products (colored arrows) in different groups of organisms (colored boxes).

Purine catabolism universally starts with adenine/guanine conversion to xanthine, followed by xanthine oxidation to form uric acid or superoxide radicals, the latter with the purpose of

protecting cells against bacterial infection[2]. The metabolic reactions that follow uric acid production are diversified in the different species. In fact, some of the enzymes involved in the purine catabolism pathway have been lost, leading to a change in the end product over the course of evolution, depending on the role of nitrogen in organisms' physiology.

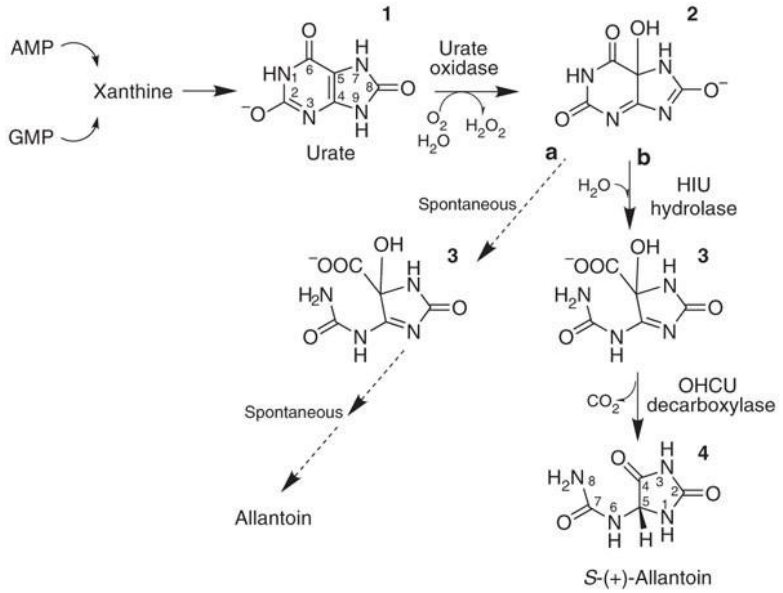


Figure 2: Purine degradation pathway to allantoin[3]. Alternative fates of urate: dashed arrows indicate spontaneous reactions that lead to racemic allantoin formation; solid line arrows indicate consecutive enzymatic reactions that lead to (*S*)-allantoin synthesis.

The complete pathway route followed these enzymatic steps: urate is converted in (*S*)-allantoin through subsequent reactions of oxidation (Uox), hydrolysis (Urah), and decarboxylation (Urad)[3], or could spontaneously be turned into (*S/R*)-allantoin after urate oxidation (Fig.2).
(*S*)-allantoin is then degraded to allantoate by allantoinase (Aln) and allantoate is stereospecifically hydrolyzed to (*S*)-ureidoglycolate by allantoicase (Allc) (Fig.3). In plants and bacteria, allantoate could be converted into (*S*)-ureidoglycolate in a two-step reaction: a first ammonia molecule is released from allantoate, producing (*S*)-ureidoglycine (Fig.3, molecule 4), which is then converted into (*S*)-ureidoglycolate with the release of a second ammonia molecule[4] or could be differently metabolized (Fig.3, reactions b and f).
This two-step reaction is an alternative to the release of a urea molecule from allantoate, therefore it does not require the enzymatic conversion of urea into ammonia and CO_2 by urease. (*S*)-ureidoglycolate is then metabolized to glyoxylate by the ureidoglycolate lyase (UGL) releasing urea[5], which is further converted into ammonia and CO_2 by urease. An alternative way

to release glyoxylate from ureidoglycolate could derive from the enzymatic activity of ureidoglycolate amidohydrolase (UGH) in plants and bacteria[6].

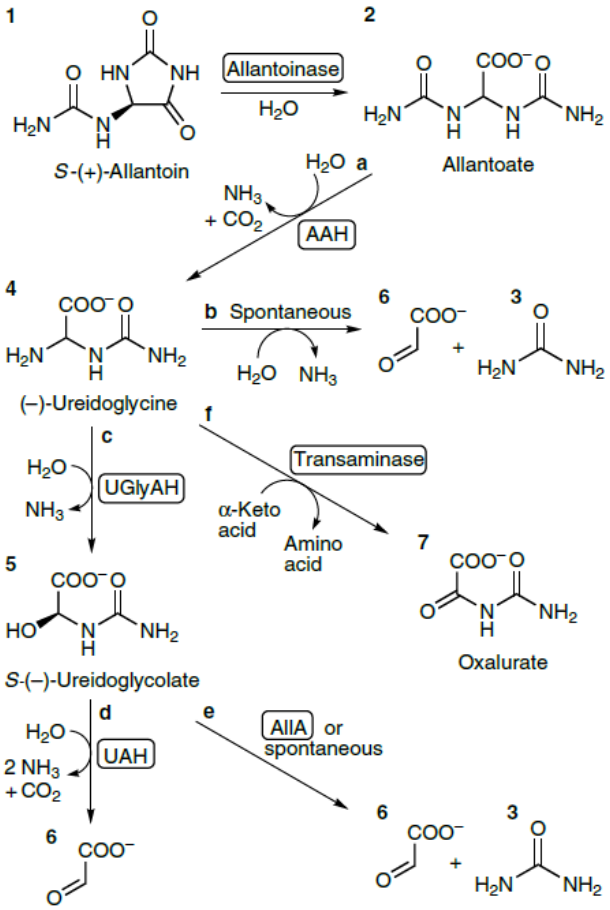


Figure 3: (*S*)-allantoin degradation pathways[4]. Alternative routes for (*S*)-allantoin degradation through spontaneous or enzymatic reactions that lead to the release of different metabolites.

Differences in purine catabolism across phylogenesis

As just mentioned, the purine nucleotides are oxidatively degraded via uric acid and allantoin to CO_2 and NH_3 in most plants; in these organisms, nitrogen plays a key role in growth control and in reproduction[7]. Plants can recycle purines to glyoxylate in order to recover all nitrogen molecules from the purine ring and to stock nitrogen in the form of allantoate or allantoin in their tissues. For example, soybeans can fix nitrogen and use allantoin for their primary nitrogen supply, and its breakdown results in the complete catabolism of the purine ring system in the endoplasmic reticulum[8] without releasing urea as an intermediate during reactions[9].

Most microorganisms, such as bacteria or yeasts, produce degrading enzymes involved in the complete breakdown of urate to ammonia[10]. In bacteria, metabolic reactions partially overlap with those of the Metazoa pathway, but additional reactions exist to make purine oxidation efficient for gaining energy. For example, *Bacillus* spp. possess a transaminase (UGXT) which recognizes (*S*)-ureidoglycine and glyoxylate as substrates and uses them to recover glycine and amino acids in general, and to survive under nitrogen and carbon starvation[11].
The capability of bacteria to metabolize urate and allantoin may contribute to their fitness and has been causally linked to human diseases, such as gout: for example, *Klebsiella spp*. may contribute to purine and uric acid catabolism in the gut microbiome, and their decreased abundance in patients that suffer from gout can contribute to exacerbate the pathology[12]. Moreover, *Escherichia coli* and other bacteria own a metabolic branch point enzyme able to convert (*S*)-ureidoglycolate in oxalurate in a NAD-dependent manner[13].

Allantoin is used as a nitrogen source also by fungi species: *Saccharomyces cerevisiae* and other yeasts are allowed to use allantoin as a nitrogen source thanks to the well-known DAL operon[14] that includes six contiguous genes involved in its breakdown (Fig.4). Other fungal species possess DAL homologous genes scattered around their genomes[15].

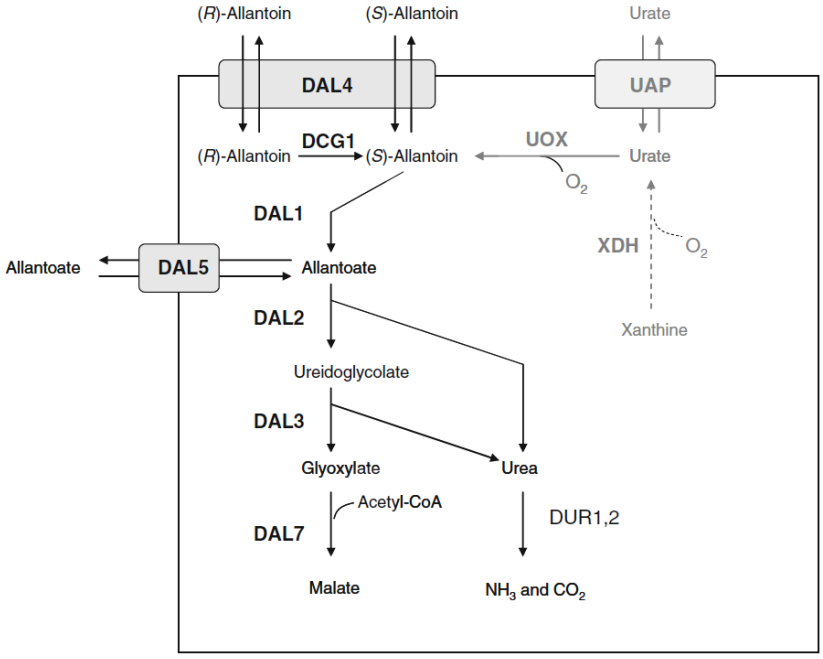


Figure 4: Allantoin metabolism in *S. cerevisiae*[14]. Metabolism of allantoin and catalytic reactions is required to obtain malate, NH_3, and CO_2 as final products. *DAL* genes and their products are involved in the catalysis of these reactions.

Animals instead partially degrade purines to urate, allantoin, allantoate, urea, or ammonia. Considering vertebrates, it is generally accepted that in reptiles, birds, apes, and humans the final product is urate as a consequence of the loss or the pseudogenization of genes responsible for the last steps of the pathway[3]. In placental mammals, the pathway ends with allantoin production; teleosts, amphibians, monotremes, and marsupials present a complete metabolic route where the final product is urea[16].

In apes and humans, the end product is urate because of the inactivation of the uricolytic pathway; when urate removal through urine is not sufficiently efficient, urate accumulates in tissues and can cause pathological conditions[17]. In fact, urate is poorly soluble at physiological pH and deposits as crystals in joints and in the kidney, causing gout, renal stones, and renal failure because up to 90% of filtered urate is reabsorbed mainly through specific transporters[12]. On the other hand, urate is capable of reacting with biologically relevant oxidants to form allantoin, confirming its fundamental role in free radical scavenging[18] ; urate is indeed capable of chelating and inactivating free metals[19,20] and could be involved in the cognitive process[21]. The loss of the final steps of the pathway could be an adaptation to terrestrial life since urate has low solubility in water and can be excreted with very little water.

Coevolution of purine degradation genes

The genes involved in the purine degradation pathway are an example of the evolution of a metabolic genes cluster in eukaryotes, and the variability of the enzymes reflects the different purposes of the metabolic pathway in different species. Through the analysis of gene co-occurrence, subtle functional divergences have been identified between homologous proteins involved in reactions of the same metabolic pathway.

Uox, Urah and Urad genes play a role in urate detoxification and are tightly conserved in organisms capable of its degradation[3,22,23]. Urate oxidases (Uox) are present in both prokaryotes and eukaryotes and are absent in birds, reptiles, and humans which release urate as the last purine metabolite. In these organisms, the two other genes (Urah and Urad) of the urate degradation show a comparable trend of gene loss and preservation depending on Uox presence with a relaxation of the purifying selection after the split of hominoids: in fact, Uox pseudogenization has resulted in the subsequent pseudogenization of Urah and Urad because of their uselessness in the absence of the previous reaction[23].

In this and other cases, genes involved in consecutive steps of a metabolic pathway exhibit a genomic association and their occurrence in the genome depends on cellular metabolic needs. Allantoinases are divided into two groups, one similar to dihydroorotase and metal-dependent hydantoinase (es. DAL1 in *Saccharomyces spp.*)[15] and a second similar to polysaccharide deacetylase which does not require any metal as a cofactor (es. puuE in *Pseudomonas spp*.)[113]. DAL1 homologous genes have been found in plants, fungi, and Metazoa[15] (insects, fishes, and amphibia) while *puuE* genes are found mostly in bacteria[113]. Both enzymes release allantoate, which is indeed hydrolyzed by allantoicase or allantoate amidohydrolases as previously described. All the enzymes involved in (*S*)-allantoin and allantoate catabolism are always co-occurrent in organisms that possess the complete urate degradation pathway.

Few exceptions occur in some organisms; the presence of allantoicase is very interesting in *Homo sapiens* since hominoids have a truncated purine degradation to urate and lack all the genes involved in the subsequent reactions. Probably this enzyme is not involved in the same metabolic pathway and has changed its function.

Ureidoglycolate metabolism and ureidoglycolate lyase features

Ureidoglycolate is an unstable compound that spontaneously decays to glyoxylate and urea[24] but there exist three enzymes that act on ureidoglycolate in the last reaction of purine breakdown.

A first one is ureidoglycolate dehydrogenase, which catalyzes the oxidation of ureidoglycolate to oxalurate in presence of NADH[13].

A second enzyme that processes ureidoglycolate is ureidoglycolate amidohydrolase[25], a manganese-dependent enzyme, localized only in the endoplasmic reticulum of plant cells, that converts ureidoglycolate to glyoxylate and releases two molecules of ammonium and one molecule of CO_2 (Fig.5, reaction on the left).

The third enzyme is ureidoglycolate (urea) lyase, which is mainly present in bacteria, fungi and in some plants and releases one molecule of urea and one molecule of glyoxylate from ureidoglycolate[25] (Fig.5, reaction on the right). The latter enzymatic activity has been detected mostly in bacteria; interestingly, it has been identified within the liver peroxisome of fish, which possess a complete urate degradation pathway. Glyoxylate, one of the final products of the reaction, is a stable compound present in a hydrated form in aqueous solution at neutral pH[9].

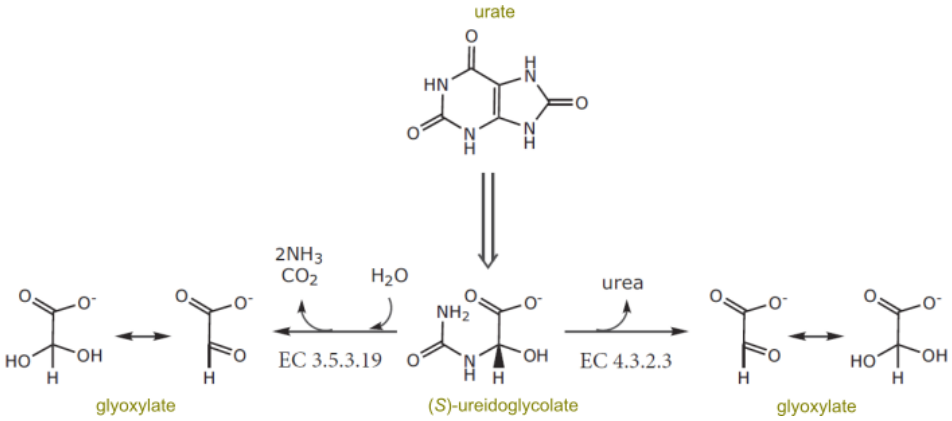


Figure 5: Alternative routes of nitrogen release from ureidoglycolate (image modified from Percudani *et al*., 2013)[25]. Starting from (*S*)-ureidoglycolate, the reaction on the left is catalyzed by ureidoglycolate amidohydrolase, and the reaction on the right is catalyzed by ureidoglycolate lyase.

Fish ureidoglycolate lyase consists of two 64 kDa identical subunits and exhibits a dependency on metals similar to bacterial enzymes[26]. However, no homologous genes of bacterial UGL have been found in the genomes of fishes and the gene responsible for this enzymatic activity

has not been identified yet. In this dissertation, I will describe the identification of the gene encoding the ureidoglycolate lyase gene in vertebrates and other Metazoa.

Glyoxylate cycle

The glyoxylate cycle is a metabolic pathway that is found typically in plants, bacteria, fungi, and microalgae[27]. Glyoxylate shunt reactions result in the production of succinate enabling the production of carbohydrates from fatty acids (Fig.6). In particular, the glyoxylate shunt avoids Krebs cycle decarboxylative steps and takes advantage of acetyl-Coenzyme A produced by beta-oxidation[28].

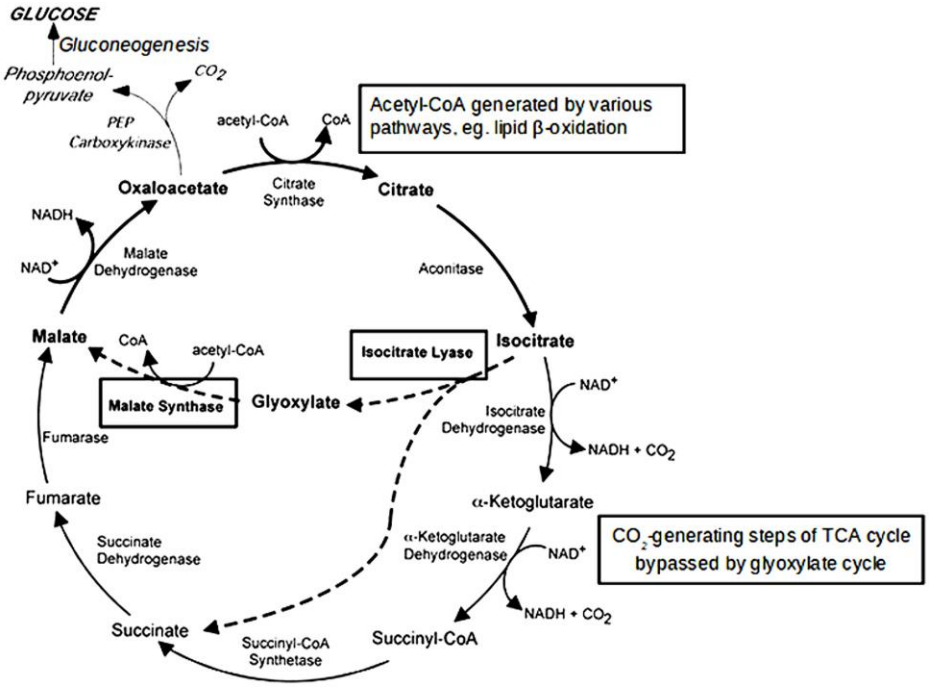


Figure 6: TCA and glyoxylate cycles[29]**.** TCA cycle (black arrows) and glyoxylate cycle (dashed arrows) reactions.

Acetyl-Coenzyme A reacts with the oxaloacetate forming citrate through the citrate synthase[30]; citrate is then converted by aconitase[31] into isocitrate, which is indeed split into glyoxylate and succinate. Succinate is delivered into mitochondria to participate in the Krebs cycle and to be converted into malate. Malate is exported and then converted by cytosolic malate dehydrogenase[32] to oxaloacetate. Glyoxylate released in peroxisomes from isocitrate is then condensed with acetyl-Coenzyme A to produce malate by malate synthase (MS).

In plants, growth is dependent on the glyoxylate cycle because it allows to obtain carbohydrates as energy sources from the degradation of the storage lipids before photosynthesis. Glyoxylate cycle enzymes in plants are localized in peroxisomes but could act also in the cytosol[33,34].

In *Arabidopsis thaliana*, isocitrate lyase (ICL) is indispensable for optimal growth and for oxaloacetate and succinate production through gluconeogenesis contrary to MS whose absence does not interfere with the metabolic process[36]; indeed, unlike isocitrate lyase mutants, malate synthase mutants are able to employ an alternative gluconeogenic mechanism thus demonstrating the non-essentiality of the canonical pathway. In plants, this is instead in the process of infection of some pathogenic microorganisms. In pathogens, the glyoxylate cycle is crucial for host organism survival and, for example, is indirectly involved in the biogenesis of mycotoxins, cell wall components, and in the production of precursors for melanization (i.e. the synthesis of melanin to encapsulate pathogens) of infection structures[36]. Studies on *Candida glabrata* have shown that the absence of ICL1 causes a block in the growth of the pathogen if citrate is present as the sole carbon source[36]. In addition to this, it has been demonstrated that fungal pathogens are unable to use carbon sources if ICL is inactivated[37].

The glyoxylate cycle has been demonstrated to have a pivotal role in the interaction of bacterial species with both plant and mammalian hosts; *Pseudomonas spp*. upregulates replicative genes, including MS and ICL resulting in its own replication and persistence inside the host[38]. In *M. tubercolosis*, glyoxylate cycle enzymes and in particular MS are involved in either enzymatic activity or in the pathogen adherence enhancement[39]. For these and other well-investigated features, the glyoxylate cycle has a relevant role in fungal and bacterial virulence in plants and animals[38].

Saccharomyces cerevisiae, together with *Caenorhabditis elegans,* which have been demonstrated to have acquired from bacteria a bifunctional protein with both ICL and MS functions[40], are anhydrobiotic organisms capable of suspending their life by losing all the water from their body in periods of severe drought[41]. In particular, to survive in hard conditions, they alter the energy metabolism by favoring the synthesis of sugars. In these organisms, the glyoxylate cycle allows the storage of oxaloacetate generating fewer ATP and NADH than the Krebs cycle[42]. Otherwise, the activity of the bifunctional protein in *C. elegans* and in other nematodes is correlated with lipid depletion and carbohydrate synthesis during embryogenesis[43].

Besides nematodes, the glyoxylate cycle is thought to be absent in Metazoa but MS genes have been identified in arthropods, echinoderms and vertebrates. An exception is the placental mammal gene that contains stop codons and does not encode for a functional protein due to its pseudogenization[44].

MS genes in several vertebrates do not contain base-substitution mutations or insertion/deletions that modify their transcription and translation, and their CDS are homologues, with high sequence similarity, to the well-studied bacterial and yeast malate synthases; these lines of evidence suggest possible maintenance of malate synthase proteins even though they could have changed their enzymatic functions, and proofs of MS activity and glyoxylate cycle existence in vertebrate tissues are lacking. The evidence of possible maintenance of MS in vertebrates was derived from a phylogenetic reconstruction of MS sequences from all living kingdoms[44]. To support the loss of the glyoxylate cycle in most vertebrates, ICL genes are absent in all Metazoa except echinoderms and nematodes.

The presence or absence of an operative glyoxylate cycle in vertebrate tissues still remains a matter of debate.

Isocitrate Lyase and Malate Synthase

Of the five enzymes involved in this metabolic shunt, citrate synthase, aconitase, and malate dehydrogenase are present in the Krebs cycle[28], while only isocitrate lyase and malate synthase are specific to the glyoxylate cycle (Fig.7). Attempts to reproduce an inverse metabolic pathway to the glyoxylate cycle in *Escherichia coli* have led to be aware that the reaction of isocitrate lyase is reversible while malate synthase is not able to catalyze the release of glyoxylate from malate[45].

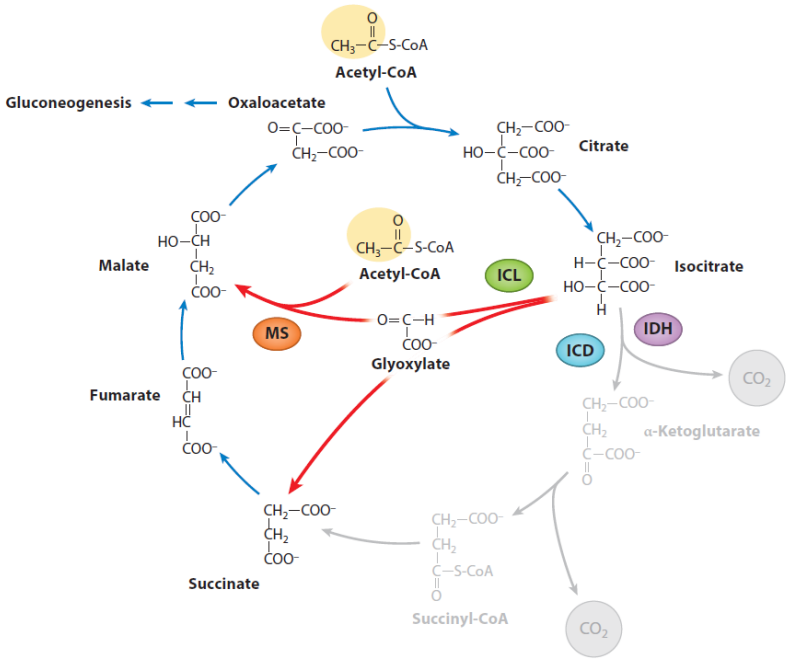


Figure 7: Overview of glyoxylate shunt and Krebs cycle[28]. Glyoxylate shunt (red arrows) and its interaction with the TCA cycle. Reactions colored in grey are bypassed by ICL and MS activities.

Isocitrate lyase catalyzes the reversible aldol cleavage of isocitrate to glyoxylate and succinate[28]. The molecular mass of these proteins ranges from 60 to 65 kDa whereas bacteria could own a smaller protein[38]. For the activity of isocitrate lyases, a tetrameric assembly is necessary; each monomer presents a Triose Isomerase central barrel with several helices which converge in the formation of a unique catalytic center. To catalyze the reaction, isocitrate lyase can bind the substrate only if it first coordinates a metal ion which increases the affinity for isocitrate[114]. Different ICL proteins in a unique organism could be explained by their involvement in the metabolism of different development stages, particularly in plants[46].

Malate synthase is an enzyme that belongs to the class of transferases and catalyzes the following reaction in the glyoxylate cycle: acetyl-CoA + H_2O+ glyoxylate ⇄ (*S*)-malate + CoA[28]. Main plant and fungal malate synthases have C-terminal PTS motifs to target the protein to peroxisomes and the molecular mass is similar to isocitrate lyases[47]; no homology has been found with citrate synthases which catalyzes a similar reaction. All malate synthases share structural similarity and possess a characteristic TIM barrel domain[48] which accommodates the glyoxylate in the active site near a metal center and the acetylated cofactor in a proper binding cavity.

The main difference between the malate synthases from different clades is the quaternary and oligomeric structure: plants possess an octameric protein[49] which could be assembled differently depending on the tissue expression; bacterial enzymes are always monomeric[50] while malate synthases are tetramer or dimer in yeasts[51,52]. The presence of Mg^{2+} is fundamental for the catalytic activity of the enzyme: the metal ion interacts directly with two oxygen atoms of the glyoxylate, two water molecules, and with protein residues possessing acid/base catalytic properties[53].

Two isoforms of malate synthase have been discovered: the so-called isoform A (MSA)[50] is present in archaea, plants, and other organisms, while isoform G (MSG)[54] is found only in bacteria. The structural difference between these two isoforms lies in an α/β domain, which is present in MSG and absent in MSA, and whose functionality is still unclear. Amino acids involved in the binding of either the substrate, the metal (Fig.8), or the cofactor have been identified, enabling the understanding of the catalytic mechanism of malate synthases A[50,55] and malate synthases G[56].

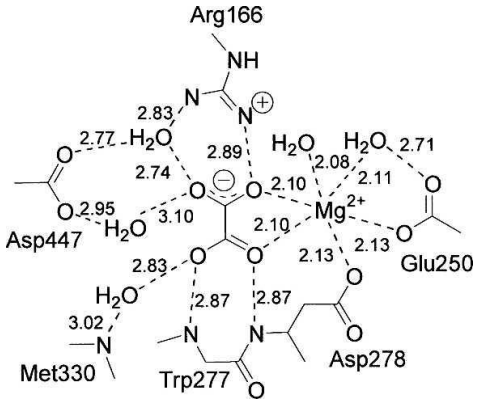


Figure 8: Active site of MSA[50]. Schematic diagram of the *Escherichia coli* MSA active site.

As described before, purine degradation and glyoxylate cycle are united by the fact that they are both located in the peroxisome and that glyoxylate is at the same time a product and a metabolite for both processes.

A functional connection is also suggested by the presence of *MS* genes included in bacterial purine degradation operons, such as the *glcB* gene of *Mycobacterium tuberculosis*, and by the fact that malate synthases and purine catabolism are co-regulated in fungi[57]. In fact, the Degradation of Allantoin Locus (DAL) of *Saccharomyces cerevisiae*, a gene cluster for purine catabolism[14], includes the DAL7 malate synthase gene and all the genes involved in allantoin degradation, suggesting a strict and ancient connection between the two pathways.

Here, we will discuss the highly significant association between the final step of purine degradation and the glyoxylate cycle in early Metazoa, supported by our innovative phylogenetic analysis (Dembech *et al.*, under review).

Results and Discussion

Pipeline and metrics for coevolutionary analysis of eukaryotic genes

An optimal method and metrics for comparing phylogenetic profiles and subsequent analysis have not yet been established, especially when considering datasets from eukaryotes. We have recently proposed and described a new computational procedure suitable for large-scale analysis, with the aim to reveal a novel coevolutionary relationship between molecular components (Dembech *et al*., under review); moreover, we have implemented a statistical significance assessment for correlated presence/absence transitions between gene pairs.

To build the coevolutionary matrix (Fig.9) we have considered eukaryotic orthogroups which are included in the OrthoDB collection[58], and we have filtered the dataset to include only genes present in more than 1% of the 1264 organisms considered for the analysis.

Considering the Orthodb dataset, these filters selected 60,912 of the 406,503 orthogroups and nearly 159M genes (>90% of the dataset); genes included in our analysis include putative orthologs among protist (148), metazoan (448), plant (117), and fungal (549) species.

The matrix of binary profiles (Fig.10A) was then built considering 60,912 orthogroups (matrix columns) and 1264 organisms (matrix rows); the presence of one or more copies of the gene in each genome was considered as “1”, while the absence was considered as “0”.

A taxonomy-constrained phylogenetic tree was used to order genomes; in particular, the transposed profile matrix was used to solve unresolved relationships of the NCBI tree that we have taken as the reference. The same tree was orientated in three different ways based on ladderization (RL right-ladderized; LL left-ladderized, NL non-ladderized) and then used to generate and order profiles.

The score and significance of coevolutionary transitions were considered when comparing each tree-ordered profile and we decided not to consider the similarity between profiles. In fact, the measure of similarity between vectors frequently appears to considerably affect the identification of putative functional gene modules and could end up in missing information[59].

We decided to focus on shared state transitions (1->0 or 0->1) that emerged from phylogenetic profiling[60], and this enumeration depends on the number of concomitant events of gene loss or acquisition considering at least two orthogroups. This approach allows us to establish whether two genes have been inherited similarly i.e. have coevolved.

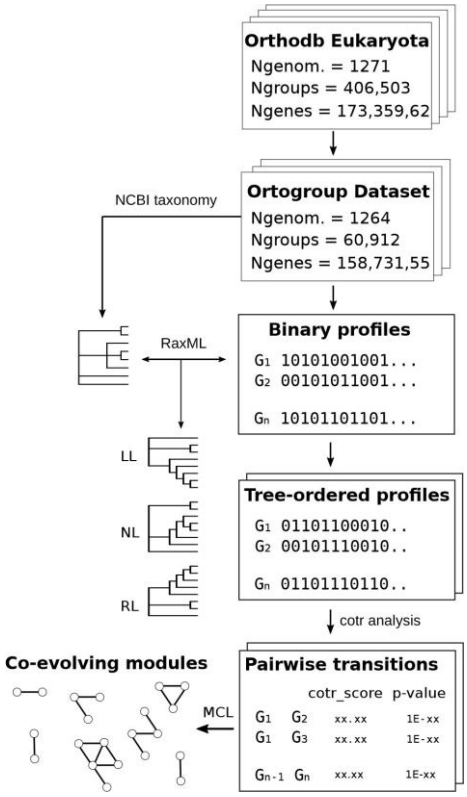


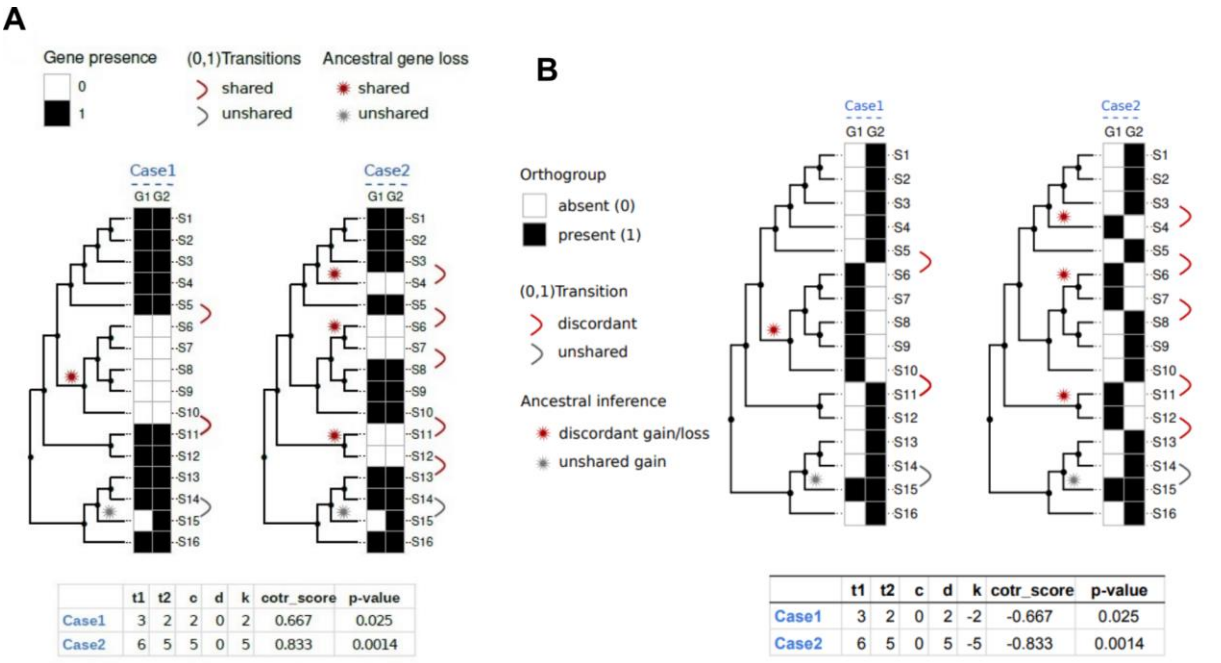
Figure 9: Pipeline for coevolutionary analysis. Scheme of the workflow used for coevolutionary analysis. Orthogroups were downloaded from OrthoDB and used to build binary profiles and tree-ordered profiles. *Cotr_scores* and *p-values* were evaluated for each pairwise comparison and the co-evolving modules were established with MCL (Markov Cluster Algorithm) clustering method.

The co-transition score (*cotr_score*) for each pairwise comparison was calculated as a function of the number of total transitions for each orthogroup/vector (t1 and t2, with no relevance of the order), and the number of concordant (c) and discordant (d) state transitions.

The *cotr_score* can quantify the evolutionary transition similarity for each orthogroup pair. If the *cotr_score* has a positive value between 0 and 1, phylogenetic profiles are correlated (Fig.10A) proportionally to the absolute value; on the contrary, if the *cotr_score* is negative, the phylogenetic profiles are assumed to be anti-correlated (Fig.10B).

Profiles with a higher number of concordant transitions have *cotr_scores* with a higher value with respect to profiles with a lower number. It is possible however to obtain high *cotr_scores* also if profiles have a low number of concordant transitions, but it is necessary to consider the possibility of a random co-occurrence. Moreover, incorrect *cotr_score* evaluations could depend on biological exceptions or gene pseudogenization, together with sequencing errors. For example, errors in bird genome sequencing can be attributed to artifacts that present an extreme percentage of GC bases in chromosomes[61].

We also estimated the *p-value* from the matrix, which calculates the probability to obtain a particular *cotr_score* by chance with Fisher's exact test, and an *adjusted p-value* calculated as the *p-value* correction for multiple comparisons considering all datasets. Orthogroup pairs reaching the predetermined 10^{-3} level of significance in all tree orientations were clustered using the Markov clustering method (MCL)[62] to identify coevolving modules considering two or more orthogroups.



	t1	t2	c	d	k	cotr_score	p-value
Case1	3	2	2	0	2	0.667	0.025
Case2	6	5	5	0	5	0.833	0.0014

	t1	t2	c	d	k	cotr_score	p-value
Case1	3	2	0	2	-2	-0.667	0.025
Case2	6	5	0	5	-5	-0.833	0.0014

Figure 10: Schematic description for coevolutionary metrics. (A) Measurement of *cotr_score* and *p-value* for two evolutionary scenarios (Case 1 and 2). 16 species (S1-S16) and two orthogroups (G1 and G2) were considered; they both have 15 matches and 1 mismatch of presence (black square) and absence (white square). The *cotr_score and p-value* are more significant in Case 2, consistent with the higher occurrence of shared gene losses, and the higher confidence of functional association. (B) Similar analysis considering anti-correlated genes.

Coevolutionary analysis results and module identification

We have therefore obtained more than 4.7 M significant pairwise comparisons considering *p-values* < 10^{-3} and ordering eukaryotic species according to an RL tree (Fig.11A); moreover, 57,716 pairs with significant *adjusted p-values* <10^{-3} emerged from our analysis. Only 530 of these had a negative score but we did not consider them for our analysis.

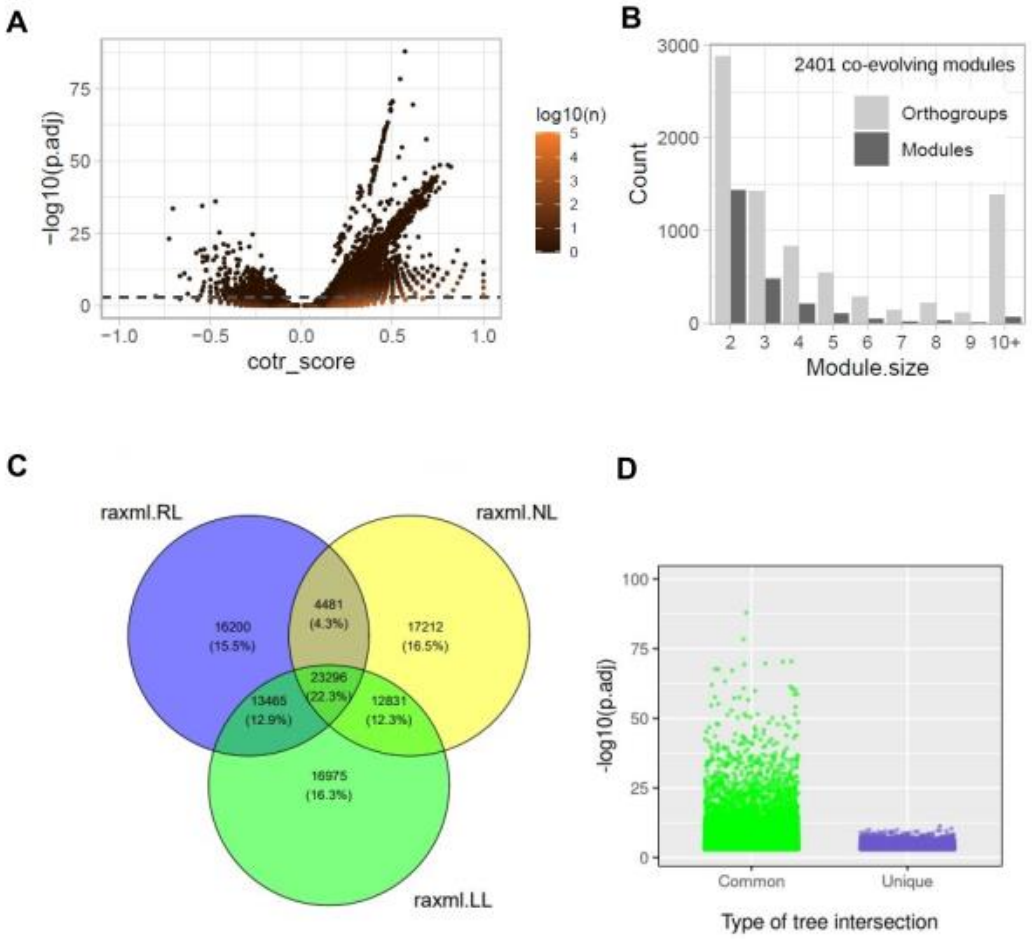


Figure 11: General results of the coevolutionary analysis and different tree orientation comparison. (A) Volcano plot of the relation between *cotr_score* and significance (*adjusted p-values*; p.adj) of 4,727,281 orthogroup pairs with *unadjusted p-value* <10^{-3} (dashed grey line); the numerosity of individual dots is indicated by the color (see legend). Transitions were calculated using an RL tree. (B) Size distribution of 2401 co-evolving modules obtained with MCL algorithm to individual orthogroup pairs; 23,296 pairs with positive *cotr_score* and significant *p.adj value* in all fully resolved tree orientations were considered. (C) Venn diagram showing the overlap among significant (p.adj<10^{-3}) orthogroup pairs obtained with right-ladderized (LR), left-ladderized (LL), and non ladderized (NL) raxml trees. (B) *p-values* of individual orthogroup pairs common to the three datasets (green dots, N=23,296) or uniquely present in the LR dataset (blue dots, N=16,200).

LL and NL trees allowed us to obtain similar quantitative results with respect to the RL tree, but the three sets extracted from the analysis overlapped for only 15-16% of clusters found and shared in all tree orientations (Fig.11C). Pairs with the most significant *p-values* were found in the subset shared by different tree orientations (Fig.11D).

We considered for our analysis the subset shared by all the orientations, which included 23,296 orthogroup pairs and we obtained 2401 coevolving modules connecting a total of 7825 orthogroups (Fig.11B). These modules are mostly composed of more than two orthogroups (63%). 18% of the orthogroups are contained in modules with more than 10 members.

The module distribution across the genomes highlighted the fact that only a small fraction contains orthogroup whose genes are spread in all eukaryotes and most of the modules are characteristics of certain species (Fig.12).

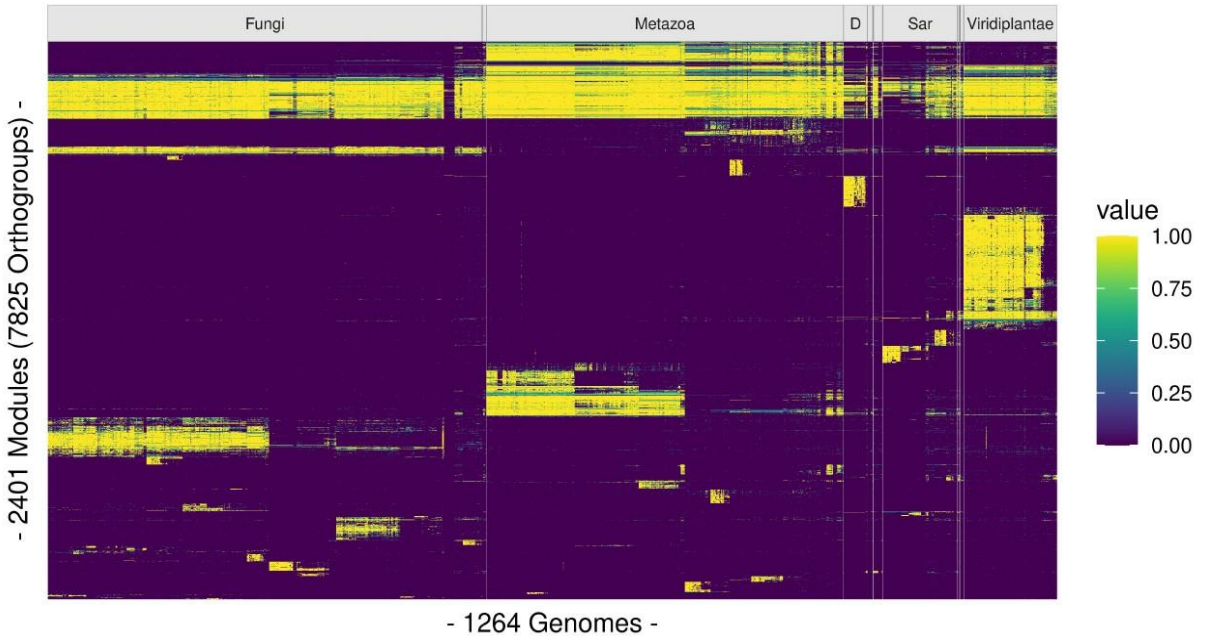


Figure 12: Organism distribution of co-evolving modules. Orthogroups belonging to the same module present in individual genomes are indicated by colors (see scale bar). Vertical lines indicate the taxonomic group boundaries at the kingdom level. D stands for Discoba; groups with <20 members are unlabelled. Species are ordered according to an RL tree polarized with Viridiplantae as the starting node.

Considering the kingdom level, 37 % of the modules represent exclusively Viridiplantae genes, 21% Metazoa, and 14% Discoba (Fig.13A). The majority of coevolutionary signals, i.e. the concordant transitions in coevolving modules, occur at the intersection of taxonomic groups (Fig.13B) and their occurrence reflects the phylogeny representation. Such signals are not found in Opisthokonts (including Fungi and Metazoa) but in the remaining branches of

eukaryote phylogeny which are less represented by complete genomes than this eukaryotic clade. At lower taxonomic levels, groups with large fractions of associated modules include Oomycota, Mollusca, and Chordata among phyla, Mammals, Agaricomycetes, and Sordariomycetes among classes, Lepidoptera and Culicidae among orders and families, respectively (Fig.13B).

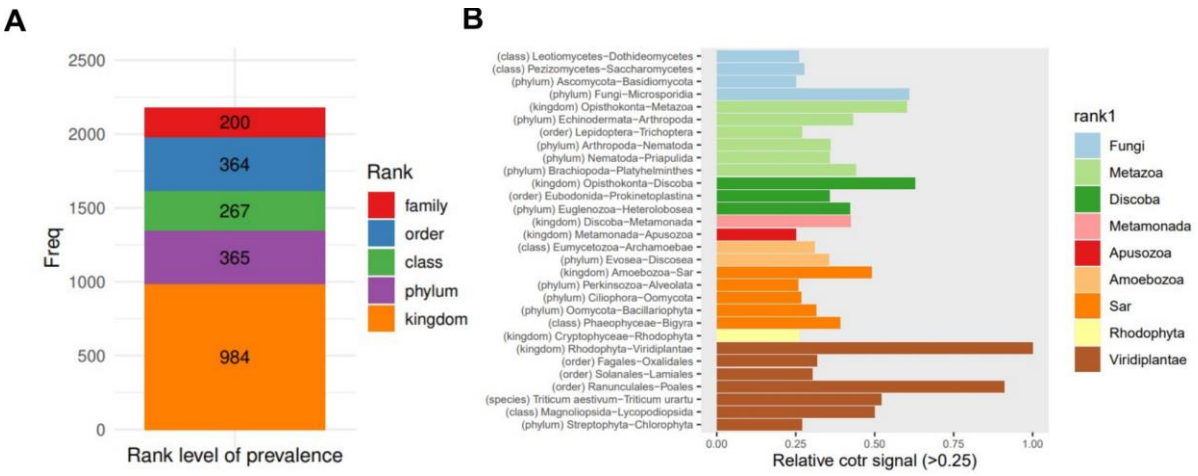


Figure 13: Distribution of coevolutionary signals across genomes. (A) Modules were assigned to the most basal taxonomic rank in which >50% of module orthogroups are found in >50% of taxa belonging to a group of the rank level (rank level of prevalence). The number of modules reaching prevalence in the different taxonomic ranks is shown in the stacked plot. (B) Relative co-transition signals (cotr) were calculated as the relative number of modules with presence/absence transitions per genome. The corresponding taxonomic intersections are reported, with indications of the most basal group-level difference (group rank in parenthesis) with respect to the antecedent point in the profile vector.

Performance evaluation

The construction of the ROC curves and the consequent calculation of the AUC area generated were carried out to estimate the performance of our phylogenetic analysis. The ROC analysis could be applied to binary classification problems and could lead to determining whether a *cotr_score is a* true positive, true negative, false positive, or false negative[63].

It was possible to compare our dataset with data from yeast interactome[64] considering information about physical pair interactions between proteins. Interacting pairs could not have a coevolutionary signal in binary profiles especially if they are spread in all the organisms, as well as coevolving proteins could not have an experimentally detected interaction.

However, the use of these or other similar reference sets for comparing a coevolutionary analysis is justified by the absence of independent evidence of gene coevolution, contrary to the availability of independent evidence for sequence homology[65].

Considering different tree orders for both a fully resolved tree ('tree') and a partially resolved NCBI taxonomy tree ('ncbi', version September 2022), we observed no appreciable differences in the performance of our method, even though AUC values were better for all tree if compared with random orientations (Fig.14B). Otherwise, we obtained an improvement of AUC if consecutive state transitions were penalized (Fig.14A).

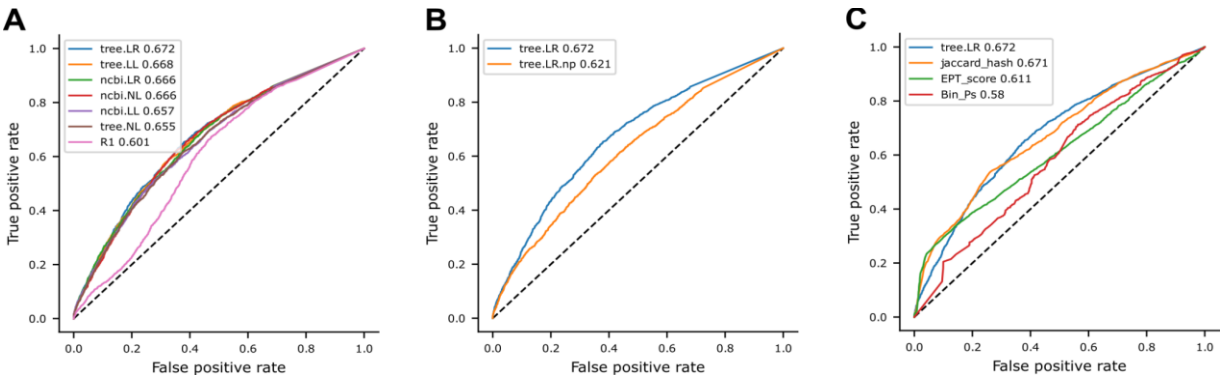


Figure 14: Performance ROC curves. s. (A) Performances obtained with the species vector ordered according to fully resolved ('tree') and partially resolved ncbi ('ncbi') trees, each in the three different orientations, plus a random tree ('R1'). (B) Performances obtained with penalized (blue) and unpenalized (orange) consecutive state transitions ranked by *p-values*. (C) Performance obtained with a chosen tree orientation ('tree.LR') with other methods of phylogenetic profile analysis: Pearson correlation (Bin_Ps), Enhanced Phylogenetic Tree (EPT)[66], and Jaccard Min-Hash[67] (Jaccard_Hash).

We also compared the performance of our method applied to a tree orientation ('tree.LR') with other methods and we obtained AUC values similar to a recently published phylogeny-aware method[67] and higher than those obtained with more conventional methods (Fig.14C). Differences observed in this method comparison, however, can depend on differences both in the scoring system and in the construction of orthologous groups[68,69].

Our method can suffer from bias which depends on errors in genome sequencing, gene calling, and gene collection in orthogroups; it could indeed result in the detection of false negative or positive results. Mostly negative scores could originate from erroneous orthogroup determination; in fact, two or more orthogroups presenting homology among their genes could derive from the splitting of a unique gene family. This may be the case of two complementary orthogroups defined similarly as "L-aspartate dehydrogenase" (1182611at2759) and "putative L-aspartate dehydrogenase" (901738at2759), but we will discuss this interesting case in the last chapter of this thesis.

We have also identified two large modules (Fig.15) with high numbers of transitions in plants and in various eukaryotic species; the former was found to be the collection of chloroplast genes (module #4), while the latter contains genes encoded by the mitochondrial genome (module #35). The high number of transitions that have been detected from the pairwise comparison between orthogroups in the two modules depends on their inclusion in the genomes; organisms whose organellar genome are not included in the database appear not to possess these genes, which have certainly coevolved and are absent due to sequencing errors.

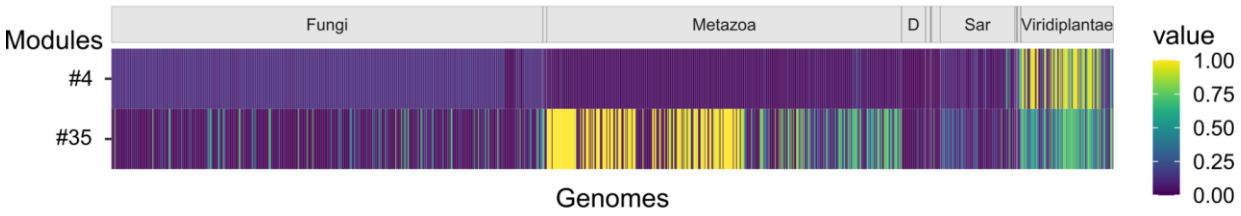


Figure 15: Evidence that Module #4 and #35 genes are encoded by organellar DNA. Modules #4 and #35 were analyzed for the presence of organellar DNA.

Assessing the functional relationships of coevolving orthogroups

The statistical assessment of the functional relationships between orthogroups which appear to coevolve from our analysis was carried out with the aid of pathway annotations found in databases (Fig.16).

Orthogroup pairs were ranked according to the significance of their evolutionary comparison; interestingly, we observed that the position of each pair depends mainly on the degree of overlap in Gene Ontology (GO) experimental annotations for the Cellular Component (CC) and Biological Process (BP) aspects, and less for Molecular Function (MF). This highlights the fact that our procedure is sensitive for the detection of gene products from orthogroup pairs which are located together in cellular compartments, take part in a molecular complex, or participate together in biological pathways; in contrast, they are less likely to share the molecular activities.

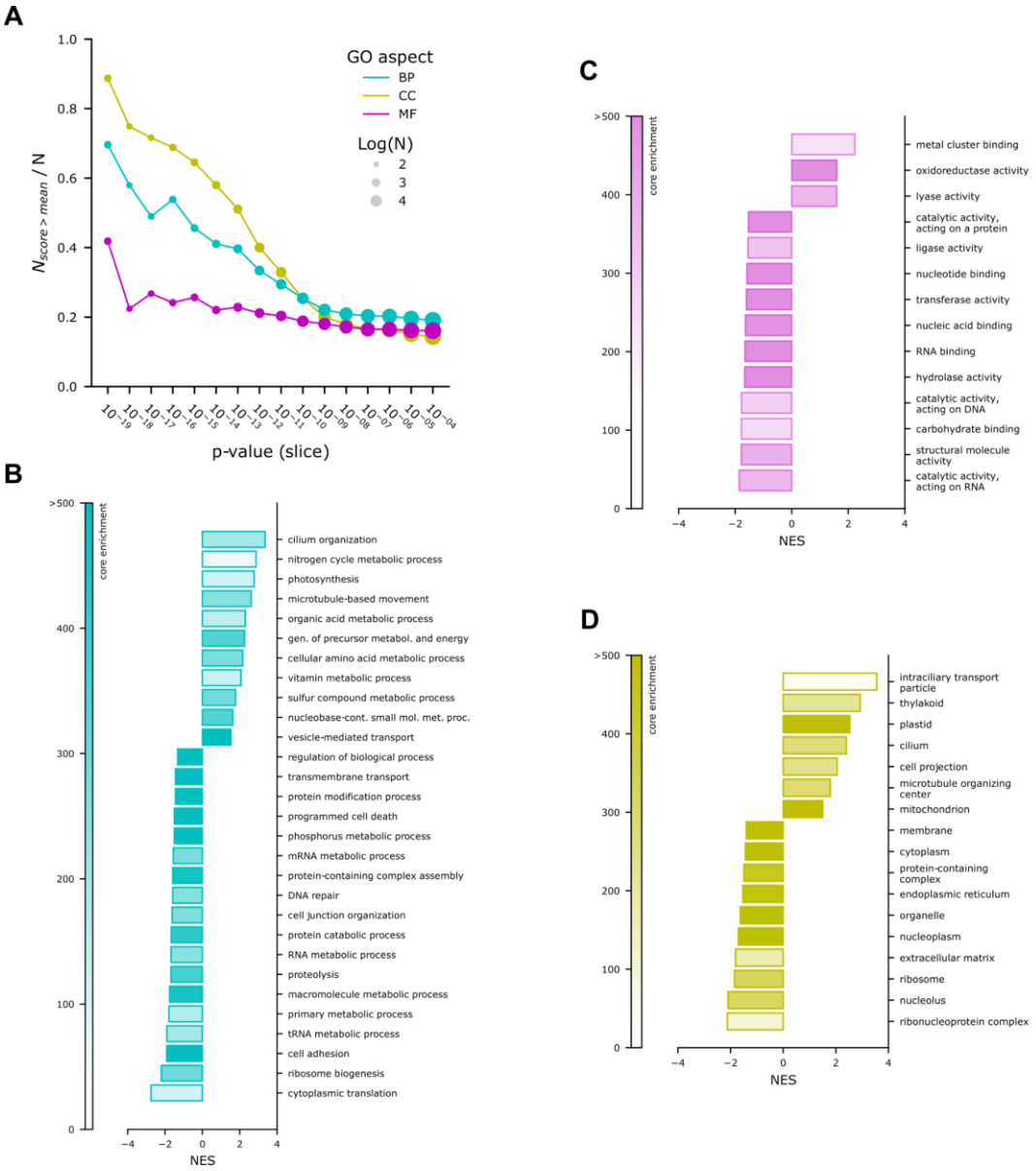


Figure 16: Functional relationships of co-evolving orthogroups. (A) Overlap of Gene Ontology (GO) experimental annotation in orthogroup pairs binned by *unadjusted p-values*. The fraction with an annotation overlap score above the mean is reported for three GO aspects: Cellular Components (CC), Biological Process (BP), and Molecular Function (MF). (B) Enrichment barplot of GO BP using an enrichment p-value cut-off of 0.01; NES = normalized enrichment score. (C) Enrichment bar plot of GO MF. (D) Enrichment bar plot of GO CC.

By performing an enrichment analysis (Fig.16A), we identified enriched BP terms of the GO-slim subset (Fig.16B) such as cilium organization, photosynthesis, microtubule-based movement, amino acid metabolism, and vesicular transport. As expected, universally conserved biological processes lack coevolving gene pairs together with GO terms elements related to the metabolism of proteins (protein catabolism) and nucleic acids (DNA, mRNA, and

tRNA metabolic processes). Ciliary structure and DNA-containing organelles were enriched also in the CC category (Fig.16D); MF terms from GO (Fig.16C) were found only for generic enzymatic activities, such as “oxidoreductase”.

A module-level analysis showed that about 55% of the modules lack any experimental annotations, and about 20% have all annotated orthogroups. The remaining 25% emerges to have only one or more annotated orthogroups in the module and provides the possibility to predict the function of the unannotated ones through the associations with known genes. In a substantial number of modules, there is no partial consensus, so they can represent false positive associations or still unknown connections.

Identification of gene candidates for pathway holes through co-transition analysis

With the aim to identify and experimentally validate novel functional associations predicted by our method, we focused on the 263/2401 modules mapping in KEGG metabolism.

With our method, we have identified several significant associations between genes and modules containing either known genes or genes not assigned to metabolic pathways. The detection of clusters converging and overlapping with modules obtained with other PP methods support the performance of our method, particularly if they contain members of well-studied molecular complexes or metabolic pathways.

As proof of principle, we have identified a module (#60) containing all the components involved in the assembly of the CatSper channel (Tab.1); this provides a close match with very recent predictions on the formation of a physical complex[70] necessary for sperm mobility. In addition, we have detected other modules containing metabolic partners involved, for example, in the FANC pathway (Tab.2), a metabolic route implicated in DNA replication and damage response and related to Fanconi Anemia[71], or in the assembly of physical complexes involved in cellular signaling such as KICSTOR complex[72] (Tab.3) and FAD-PAD[73] binary complex (Tab.4).

By predicting the enzymatic activities to be studied *in vitro*, this approach allows us to take advantage of the annotated orthogroups present in the modules, but it is necessary to have knowledge about the reactions for which no genes have been identified, the so-called “pathway holes”[74]. In that case, the unassigned genes identified by our coevolution analysis may be considered candidates for the unassigned pathway reactions if they share significant *p-values* with one or more components of the pathway.

1269955at2759	Cation channel sperm associated 3
1270192at2759	Cation channel sperm associated 4
213907at2759	cation channel sperm-associated protein subunit delta
846266at2759	Cation channel sperm-associated protein subunit beta
110808at2759	Cation channel sperm-associated protein, subunit gamma
580290at2759	cation channel sperm-associated protein 2
1306120at2759	EF-hand domain
920162at2759	Cation channel sperm associated 1
1380438at2759	Protein of unknown function DUF4579

Table 1: Module #60. Orthogroup cluster containing components of CatSper channel.

1128880at2759	Integrin-alpha FG-GAP repeat-containing protein 2
365605at2759	KICSTOR complex protein C12orf66-like
347778at2759	protein SZT2
692881at2759	kaptin
413007at2759	DEP domain

Table 2: Module #175. Orthogroup cluster containing components of KicStor complex.

169407at2759	FANCI solenoid 2 domain
979208at2759	Fanconi anaemia protein FANCD2
1002812at2759	Fanconi anemia complex, subunit FancL, WD-repeat containing domain
1431031at2759	Ubiquitin-conjugating enzyme E2

Table 3: Module #288. Orthogroup cluster containing components of FANC complex.

1363017at2759	Flavin prenyltransferase PAD1, mitochondrial
588541at2759	Ferulic acid decarboxylase 1

Table 4: Module #957. Orthogroup cluster containing PAD1 and FAD1 components.

Identification of a novel coevolutionary connection between MS and ALLC gene families

Ureidoglycolate lyase, i.e. the enzyme responsible for ureidoglycolate conversion into glyoxylate and urea, acts in the last step of the purine degradation pathway in most eukaryotes. The gene encoding for ureidoglycolate lyase has been identified in some bacteria[5], plants[75], and fungi[76] but has remained unidentified in Metazoa despite the demonstration of its activity in vertebrate tissues[26].

We examined our data on coevolutionary relationships between eukaryotic orthogroups with the aim of discovering significant associations between gene families annotated in the KEGG pathway and attributing functions to unknown genes.

From our analysis, the purine degradation pathway was found to be one of the most enriched pathways in terms of coevolving genes; since genes involved in purine metabolism have been described previously to coevolve across phylogenesis[3], we were not surprised by this result.

We observed that the orthogroup annotated in Orthodb v.10.1 as "Malate synthase" (MS, 358540at2759) has significant associations with the "Uricase" orthogroup (Uox, 906540at2759), which includes genes responsible for the first step in urate peroxisomal degradation, and with "Allantoicase" (Allc, 563639at2759), which catalyzes the hydrolysis of allantoate to ureidoglycolate before the ureidoglycolate lyase reaction in the pathway.

Orthogroup ID	OG ID - Orthodb v.10	OG
Malate synthase	358540at2759	MS
Isocitrate lyase	905115at2759	ICL
Uricase	906540at2759	UOX
5-hydroxyisourate hydrolase	1453185at2759	URAH
Oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase	1547390at2759	URAD
Allantoicase	563639at2759	ALLC

Table 5: Orthogroups from glyoxylate cycle and purine catabolism. Schematic table of orthogroups from Orthodb v.10 implicated in our analysis. Rows containing orthogroups from the glyoxylate cycle are colored in pink and from the purine degradation pathway in light blue.

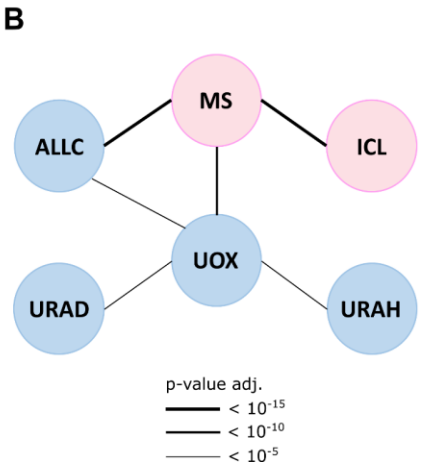
"Malate synthase" is also associated with "Isocitrate lyase" (905115at2759), which is considered its partner due to its involvement in the glyoxylate cycle in bacteria, plants, and fungi together with MS. As support for the hypothesis of a possible novel relation between "Malate synthase" and "Allantoicase", we noticed that they both take part in the same module (#976) in our collection.

At the same time, MS emerged to be significantly associated with other genes clustered in module #53 and involved in purine degradation reactions (Tab.5; Fig.17A); in addition to this, it has been previously hypothesized that malate synthase genes in animals had acquired a new function still to be characterized[44].

Genes involved in urate metabolism such as Uox, Urah, and Urad are significantly interconnected (Fig.17A); their co-occurrence revealed by *cotr_scores* reflects Uox gene inactivation and the consequent dismissal of the following genes of the pathway, since in Uox absence they are unable to perform their catalytic function.

A

OG1	OG2	cotr_score	p-value	p-value adj.
ICL	MS	0.3140	1.12E-25	1.97E-18
ALLC	MS	0.2721	6.66E-25	1.17E-17
URAH	UOX	0.2323	2.19E-19	3.86E-12
URAD	UOX	0.2079	5.20E-17	9.15E-10
UOX	MS	0.2047	3.87E-15	6.80E-08
ALLC	UOX	0.1899	7.85E-14	1.38E-06
URAH	ALLC	0.1657	1.09E-10	1.90E-03
URAH	MS	0.1503	1.09E-09	1.90E-02
URAD	ALLC	0.1463	2.51E-08	4.33E-01
URAD	MS	0.1299	7.40E-08	1.00E+00
ALLC	ICL	0.1128	6.17E-07	1.00E+00
URAD	URAH	0.1280	2.94E-06	1.00E+00
ICL	UOX	0.0940	7.31E-05	1.00E+00



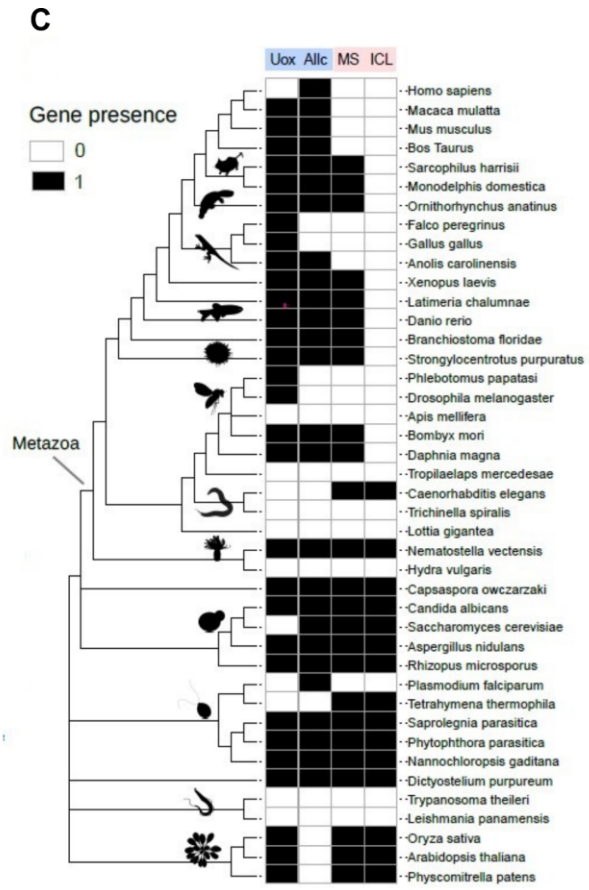


Figure 17: Identification of a relation between MS and purine catabolism genes through concordant transition analysis. (A) *Cotr_score*, *p-value*, and *p-value adj.* for pairwise comparison between glyoxylate cycle and purine degradation orthogroups. (B) Schematic network showing significant cotr associations of orthogroups considered in the analysis; different line thickness represents *p-values adj.* (C) Phylogenetic profile similarity of coevolving genes from purine catabolism (Allc and Uox) and glyoxylate cycle (MS and ICL).

Based on the *p-value* calculated for pairwise comparisons, MS is more closely related to Allc and ICL than Uox to its enzymatic partners Urah and Urad (Fig.17B); that is because MS genes are always present in Metazoa that possess Allc and they are both absent in organisms whose purine metabolism is truncated at acid uric formation (Fig.17C), while Uox is maintained in some organisms which have independently lost Urah and/or Urad. Moreover, MS is coincident with ICL in SAR, plants, and fungi where the glyoxylate cycle occurs (Fig.17C).

This novel coevolutionary linkage between MS and Allc has been successfully identified with the introduction of our novel metric since the MS profile overlaps with purine metabolism genes in a limited group of organisms (Fig.18); in fact, only co-transition rankings allowed us to detect significant associations with Allc and Uox, while the expected association with ICL has been retrieved also with other methods. Organisms holding all these four genes probably have maintained both the entire purine metabolism pathway and the glyoxylate shunt.

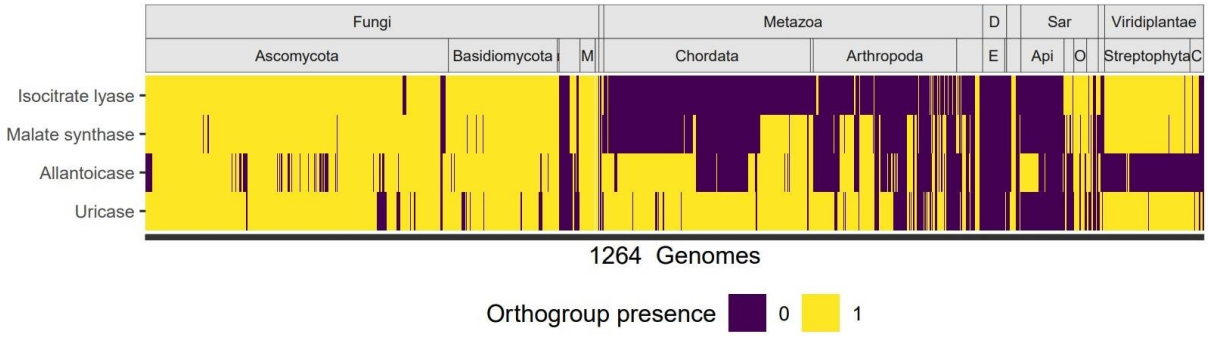


Figure 18: Phylogenetic profile (PP) similarity of co-evolving genes of the glyoxylate cycle and purine degradation. Scheme of absence (violet) and presence (yellow) of ICL, MS, ALLC, and UOX orthogroups across 1264 species ordered according to phylogeny.

MS protein is suitable to be the candidate for ureidoglycolate lyase activity

The involvement of glyoxylate both in the purine degradation pathway and in the glyoxylate shunt could be a possible explanation for the novel association detected by the signal of coevolution of Allc and MS genes in eukaryotes. Malate synthase has been described as a critical enzyme of the glyoxylate shunt and it is involved in the Claisen-like glyoxylate and acetyl-CoA condensation into (*S*)-malate with the concomitant release of CoA (Fig.19B), which is an intermediate of the Krebs cycle[28]; otherwise, allantoicase is involved in allantoate conversion to ureidoglycolate[3] (Fig.19A), the glyoxylate precursor.

The coevolutionary association is clearly noticeable in Metazoa, organisms in which the presence of the glyoxylate cycle has never been demonstrated and ICL enzymatic activities are undetectable in animals except from nematodes[40,77]; on the other hand, malate synthase protein was detected with cytochemistry experiments suggesting the conservation of a homologous protein in vertebrate tissues[78].

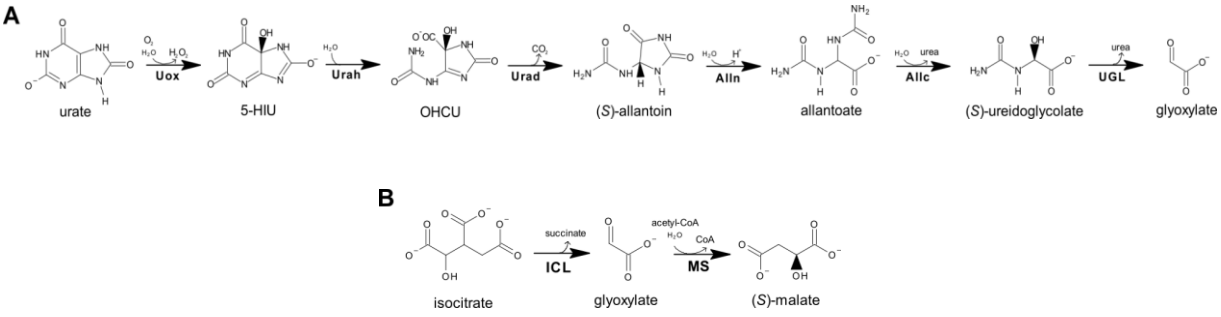


Figure 19. Purine degradation pathway and glyoxylate shunt reactions. (A) Schematic reactions of purine catabolism, starting from urate to the last metabolite, glyoxylate: Allc is involved in urea release from allantoate. (B) Key reactions of glyoxylate pathway: MS acts by converting glyoxylate into (*S*)-malate.

This evidence could be suggestive of a possible gene adaptation and a consequent loss of the malate synthase activity in order to acquire a new molecular function.

We performed the frequency analysis of the PTS1 considering the C-terminus of “malate synthase” proteins collected in the OrthoDB database (which are defined as “malate synthase” and “malate synthase-like” proteins) and we confirmed a prevalent peroxisomal localization consistent with the common localization[16] of the last three steps of purine catabolism (Fig.20A) in vertebrates.

On the basis of the following computational and experimental proofs, the MS gene in Metazoa could be proposed as responsible for ureidoglycolate lyase activity, the last unassigned step of the purine degradation pathway.

*Dr*MSL is the best candidate for ureidoglycolate lyase

Ureidoglycolate lyase activity in Metazoa was described first by Takada and colleagues[26]; the author focused on the presence of this catalytic activity in sardine and mackerel purified liver tissues, but they did not succeed in the identification of the corresponding gene. However, they have established the molecular mass of the UGL protein (64 kDa) with sucrose density gradient separation experiments and its peroxisomal localization.

To test if members of this orthogroup could take part in the purine degradation pathway, we verify if *Danio rerio* "malate synthase-like" protein (*Dr*MSL) could be the missing UGL protein since it is the only gene that has not yet been identified in this pathway.

We applied the expected molecular weight (with a tolerance of 10%) and the PTS1 pattern [SAC][KRHSN][LM] at the protein level, and we excluded zebrafish proteins with homologs in *H.sapiens* (Fig.20B); this latter filter has been utilized because ureidoglycolate lyase has never been detected in human tissues, according to the truncation of the purine degradation pathway. Commands are reported in Supplementary Information.

We found only two candidates that suited the UGL features: "protein brambleberry precursor" (NP_001257483.1, see S.I.) and "malate synthase-like" (NP_001188337.1).

Besides the reported computational evidence, our hypothesis of the reassignment of the MS gene to the last step of purine degradation is supported by the evidence of compatibility between the predicted molecular mass (62.5 kDa) of *Dr*MSL protein and the experimentally determined molecular mass of the fish protein with ureidoglycolate lyase activity and by the presence of a PTS1 signal at the C terminus of *Dr*MSL, which is necessary to target the protein to peroxisomes. Furthermore, it has been recently proposed that *Dr*MSL may convert the glyoxylate from purine degradation to malate[79].

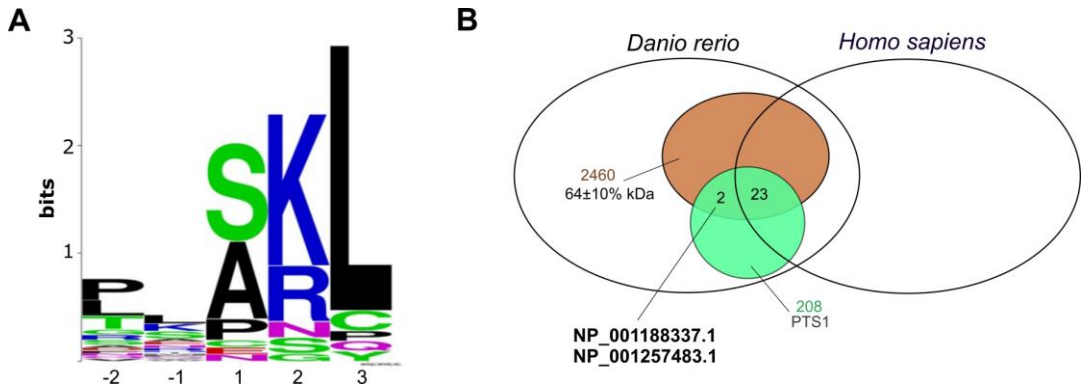


Figure 20: Identification of a *Danio rerio* gene candidate with UGL features. (A) Presence of a C-terminal PTS1 motif in MS and MSL sequences of selected eukaryotic organisms depicted with a sequence logo. (B) Venn diagram of *Homo sapiens* and *Danio rerio* proteomes intersected and filtered with three expected UGL features (MW: 64 kDa ± 10%, PTS1 signal, present only in *D. rerio*). "Malate synthase-like" and "protein brambleberry precursor" are indicated by accession numbers written in bold characters.

Based on these assumptions, the zebrafish gene annotated in databases as "malate synthase-like" (*Dr*MSL) shares key features with ureidoglycolate lyases described in fish liver extracts, suggesting the involvement of malate synthase-like genes in purine degradation pathway.

Sequence and structure analysis of MS and MSL protein

To verify the loss of the ancestral malate-synthase activity and the acquisition of the new ureidoglycolate lyase activity, we compared MS and MSL sequences with the sequence of the extensively studied *Escherichia coli* Malate synthase A (*Ec*MSA). Lohman and colleagues[50] identified the amino acids responsible for acetyl-CoA binding and glyoxylate coordination in the presence of an Mg^{2+} ion in *Ec*MSA. In particular, Lys101 and Tyr154 are involved in the 3'-phosphate binding while His368 in 5' phosphate binding; Asn105, Met102, and Pro369 are responsible for adenine ring packing; Ala367 and Thr95 establish hydrogen bonds with the N6 position of the adenine ring (Fig.21A).

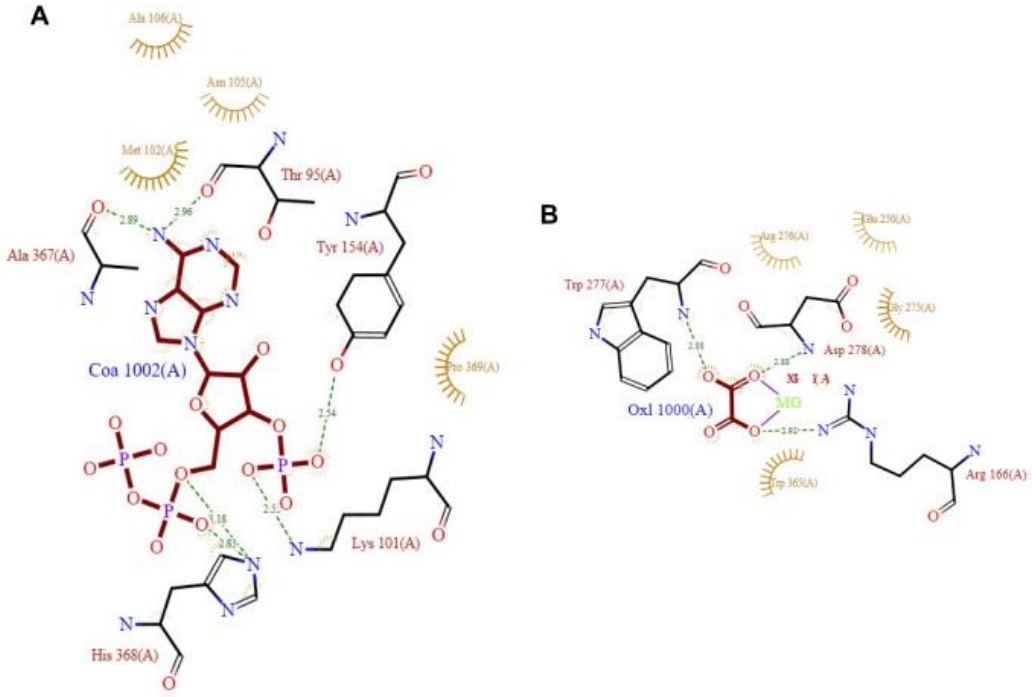


Figure 21. *Ec*MSA ligand interactions described for the 3CV2 model[50]. (A) Schematic diagram of the *Ec*MSA residues involved in acetyl-Coenzyme A binding obtained from PDBsum. Atoms are depicted color-by-element and atomic distances are reported in green. **(B)** Schematic diagram of the *Ec*MSA residues involved in metal coordination and substrate binding obtained from PDBsum. Glyoxylate is substituted with oxalate (in dark red).

The authors have also identified the residues, Trp277, Asp278, and Arg166, and three water molecules involved in hydrogen bonds with the substrate (Fig.21). Arg166 is also involved in coordinating the metal ion. Other amino acids assigned to the active site (Glu250, Met330, Asp447) are involved in the coordination of the metal ion or in binding water molecules, allowing the correct assembly of the catalytic site.

To gain insights into the function of MSL proteins, a multiple alignment of selected MS (group 1) and MSL (group 2) sequences together with *Ec*MSA was performed and visualized, and the conservation of residues with a functional role in malate synthase catalysis was investigated (Fig.22). Considering MSL sequences in Amoebozoa and Metazoa, the totality of the residues involved in acetyl-CoA coordination is lost and has been mutated into amino acids with different biochemical characteristics (Fig.22, orange triangles). This observation suggests a possible change in function as a result of the loss of the acetyl-CoA binding capability.

Additional evidence supporting a different role of the two protein families is the existence of other differently preserved residues (Fig.22, cyan stars) between the two sequence groups. These amino acids have been identified considering 13 Å surrounding the metal ion and matching this analysis with the PDBsum database[80].

In association with these point mutations, we have noticed a remarkable insertion of two amino acids (Fig.22, purple column) together with a close substitution of a conserved glycine in proteins that belong to group 2. We may hypothesize that these insertions, together with the other point mutations, could modify the binding specificity toward molecular substrates and cofactors and could interfere with the possibility to make room for the substrate.

By contrast, residues involved in glyoxylate binding and metal ion coordination (Fig.22, blue triangles) have been conserved during evolution in both MS and MSL sequences. The maintenance of these critical residues, together with surrounding point mutations, suggest the possibility that MSL proteins bind a substrate with similar biochemical characteristics with respect to glyoxylate or with a similar functional group distribution, as could be ureidoglycolate.

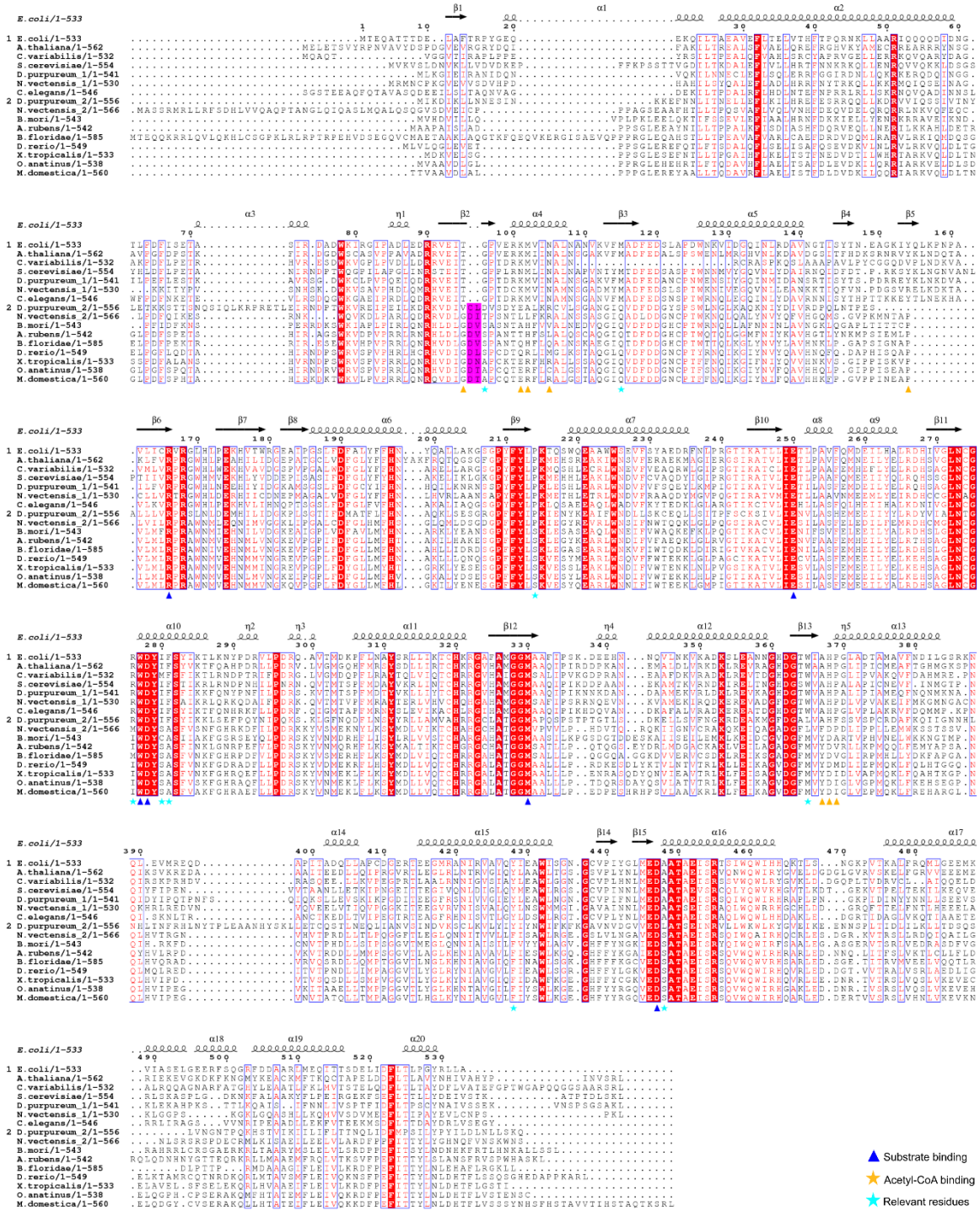


Figure 22: Multiple alignment of MS and MSL sequences. Multiple alignment of MS sequence (group 1) from bacteria, fungi, SAR, and plants and MSL sequences (group 2) from metazoa and SAR. Amino acids conserved in all sequences are shaded red. Residues involved in the metal ion and substrate binding are pointed with blue triangles, and residues involved in acetyl-Coenzyme A binding with orange triangles. Cyan stars indicate residues differently conserved in group 1 in respect of group 2. The inserted dyad in MSL proteins is placed near acetyl-Coenzyme A sites and is shaded purple.

To examine the surrounding of the active site and to confirm the evidence emerging from the multiple alignment analysis, a 3D model of *Dr*MSL was built, which was taken as the reference for the analysis of all vertebrates' MSL protein structures.

The *Dr*MSL (NP_001188337.1) 3D structure was modeled using *Escherichia coli* Malate synthase A (PDB ID: 3CUZ) as a template and then superimposed on the *Ec*MSA monomer; the QMEAN of the model is 0.75, probably due to the good level of identity between the two sequences (36.56%).

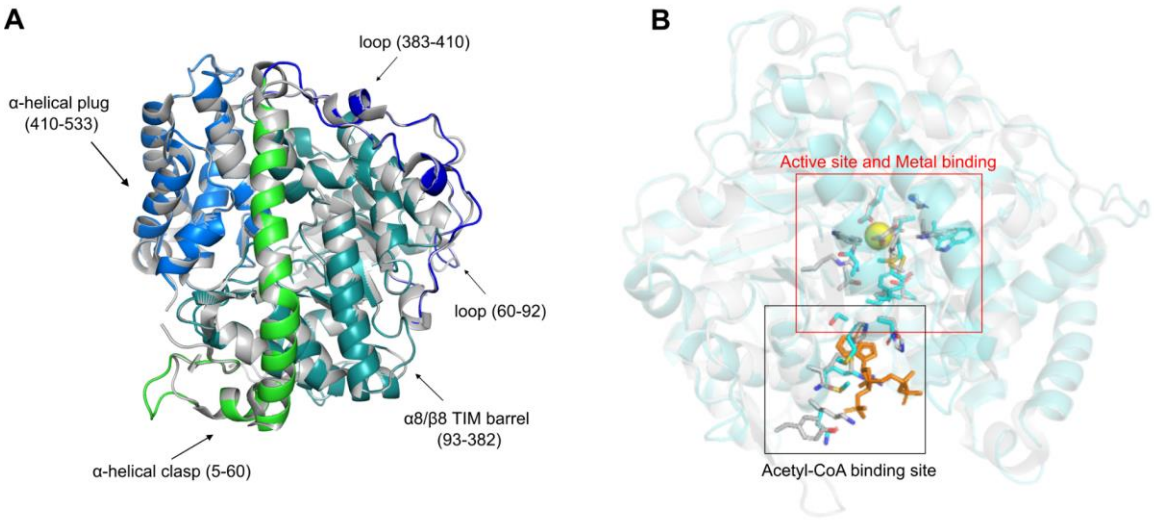


Figure 23: Homology modeling of *Dr*MSL and superimposition with 3CUZ template. (A) Structural domains of *Dr*MSL (colored cartoon) model superimposed with *Ec*MSA structure (grey cartoon). Each blue and green shades correspond to a different *Dr*MSL domain, each of which is pointed by arrows with additional information. (B) Comparison between active site (red box) and Acetyl-CoA binding pocket (black box) for *Dr*MSL 3D homology model (light blue cartoon) superimposed on the experimental structure of *Ec*MSA (3CV2, grey cartoon). Amino acids involved in MS catalysis are drawn in sticks. *Ec*MSA is crystallized together with acetyl-CoA (orange sticks), substrate analog (oxalate, red sticks), and a metal ion (Mg^{2+}, yellow sphere).

The superimposition reveals a striking similarity between the two protein structures and a similar domain organization; in Fig.23, *Dr*MSL is colored with different green and blue shades to highlight structural domains shared with the grey structure of *Ec*MSA.

As described for *Ec*MSA[50], *Dr*MSL is characterized by the presence of an N-terminal clasp (Fig.23A, light green) and a C-terminal plug (Fig.23A, light blue) both formed by a helical structure and connected with two loops (Fig.23A, blue) to the main core of the typical MS domain[81,55], the α8/β8 TIM barrel (Fig.23A, petrol green).

The superimposition of the two models allowed us to compare both the active site and the acetyl-CoA binding pocket (Fig.23B) described for *Ec*MSA[50] in order to investigate a possible difference in *Dr*MSL function, which could reallocate the MSL proteins to the purine degradation pathway as proposed.

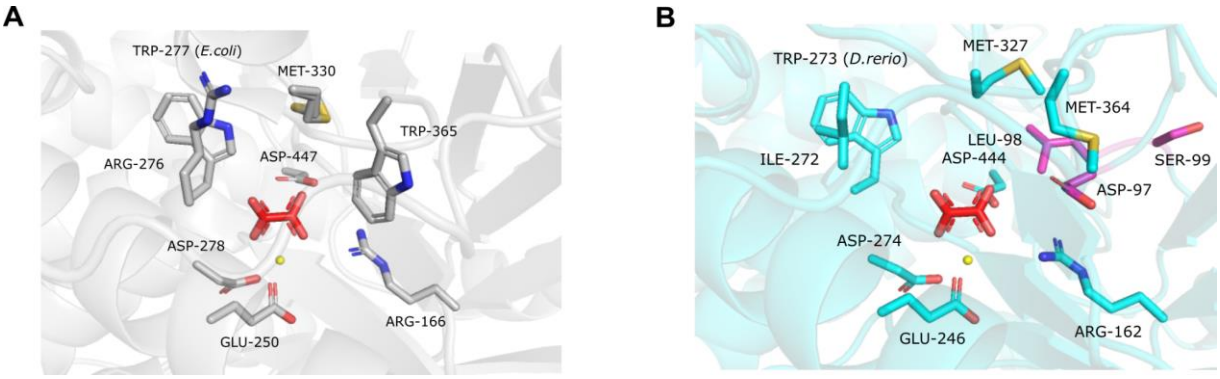


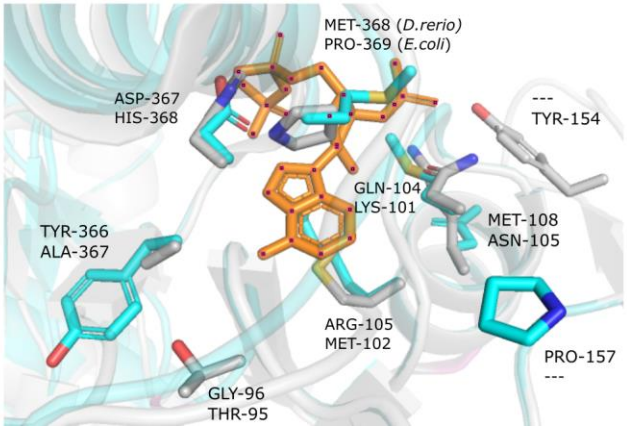
Figure 24: Comparison between *Ec*MSA and *Dr*MSL active sites. (A) Metal binding environment and substrate site of *Ec*MSA (grey model). The key residues are represented by grey sticks, oxalate is depicted in red sticks, and Mg^{2+} is represented as a yellow sphere. (B) Parallel representation of the metal environment and putative active site for the *Dr*MSL model (light blue cartoon). The key residues are depicted as light blue sticks and the inserted dyad as purple sticks as in Fig.22.

The amino acids involved in substrate binding and in metal coordination are spatially conserved. In fact, the comparison between the two 3D structures demonstrates the complete spatial superimposition of Asp274, Arg162, and Glu246 of *Dr*MSL with the corresponding residues of the bacterial protein (Fig.24). The surrounding of the *Dr*MSL putative active site does not overlap entirely with that of *Ec*MSA and presents some point mutations that distinguish MSL from MS proteins, such as Met364, Ile272 (Fig.22, cyan star; Fig.24s).

In addition, the insertion of the newly identified dyad DL (Fig.24B - purple stick) is clearly placed inside the active site and is interposed between amino acids (Met327, Asp444, Arg 162) which take part in metal and water molecules coordination.

We can speculate that mutated and inserted residues are probably participating in a partial rearrangement of the active site and in its size alteration although they did not participate directly in the catalysis, and we can hypothesize a partial modification of the overall active site with the possibility to bind a novel substrate that shares a similar charge distribution with glyoxylate.

A



B

*Dr*MSL	*Ec* MSA
GLY-96	THR-95
GLN-104	LYS-101
ARG-105	MET-102
MET-108	ASN-105
TYR-366	PRO-157
TYR-366	ALA-367
ASP-367	HIS-368
MET-368	PRO-369

Figure 25: Comparison between *Ec*MSA acetyl-CoA binding site and *Dr*MSL model. (A) The close-up comparison of the *Ec*MSA acetyl-CoA binding site shows the mutation of all the superimposed residues in the *Dr*MSL structure. (B) Table of *Ec*MSA residues involved in acetyl-CoA binding and the corresponding *Dr*MSL amino acids detected from both sequence and structural alignments.

Residues involved in the definition of acetyl-Coenzyme A binding pocket have different biochemical characteristics between the two proteins (Fig.25); pronounced conformational changes within the binding sites and a charge distribution no longer suitable for acetyl-CoA coordination are observed in *Dr*MSL structure. To be noted are the loss of the side chain ring of *Ec*MSA Pro369, which ensures the stacking with the adenine ring of acetyl-CoA, and the basic/acid or hydrophobic/hydrophilic changes in the side chains of the involved residues.

To make the point, the consideration of the overall conservation of these critical residues for substrate recognition was combined with the evidence of a probable loss of the malate synthase activity in Metazoa and SAR, consequent to the inability of acetyl-Coenzyme A binding.

Our analysis provides evidence of the probable *Dr*MSL loss of the ancestral activity, consistent with the absence of glyoxylate cycle in vertebrates, and the acquisition of a new molecular function. However, because there is no computational evidence of its ability to catalyze the reaction, we decided to study *in vitro* the catalytic properties of the *Dr*MSL protein.

Enzymatic and chemical strategy to synthesize ureidoglycolate

To demonstrate ureidoglycolate lyase activity acquisition for *Dr*MSL, we explored both a chemical and an enzymatic procedure to obtain ureidoglycolate, the putative substrate of the proposed reaction and the product of the previous reaction in purine catabolism.

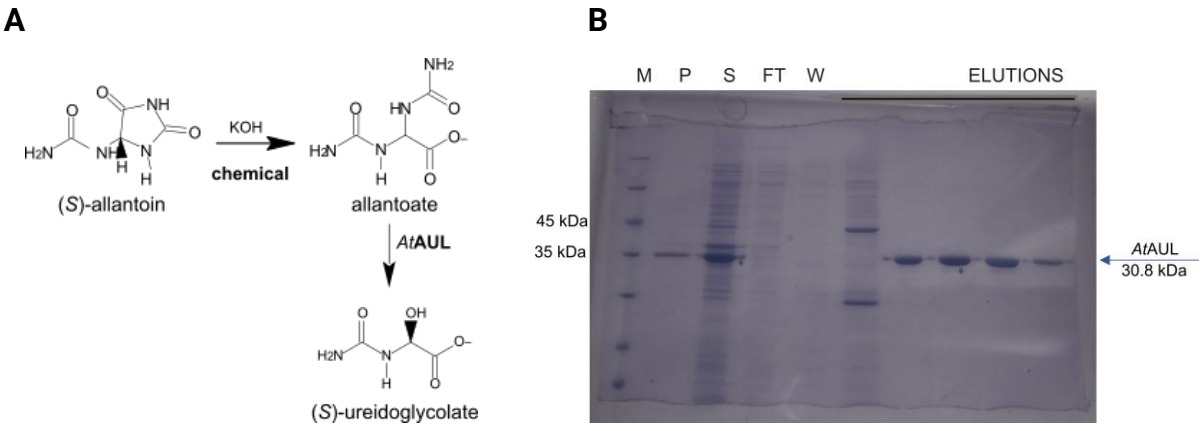


Figure 26: Enzymatic strategy to obtain (*S*)-ureidoglycolate. (A) Scheme of the reactions required to obtain (*S*)-ureidoglycolate. (B) SDS PAGE gel (12%) of *At*AUL protein fractions obtained after sonication and purification with FPLC: M - marker; P - non-soluble fractions; S - soluble fractions; FT, flow through; W - washing; ELUTIONS, eluted fractions.

We first tried to obtain (*S*)-ureidoglycolate with an enzymatic approach: we synthesized allantoic acid through basic hydrolysis of (*S*)-allantoin (Fig.26A, upper reaction) and we used it as substrate to obtain ureidoglycolate (Fig.26A, lower reaction).

We expressed in a recombinant form the *UGLYAH2* gene (Fig.26B) from *Agrobacterium tumefaciens (At*AUL*)*[6], which encodes for an enzyme able to convert allantoate to (*S*)-ureidoglycolate releasing urea. We monitored the reaction and we confirmed the consumption of allantoate and the formation of (*S*)-ureidoglycolate with ^{1}H NMR but we did not clearly observe the completion of the enzymatic reaction and the accumulation of our desired compound, which however partially decays into glyoxylate (Fig.27B).

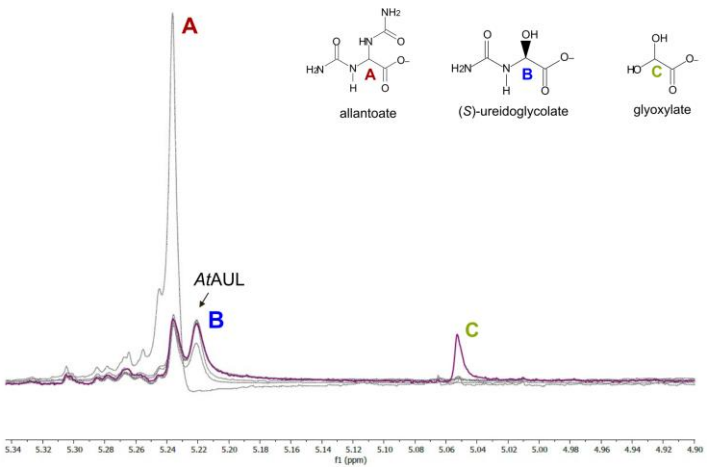


Figure 27: Enzymatic conversion of allantoate to form (*S*)-ureidoglycolate catalyzed by *At*_AUL. Superposition of time-resolved ^{1}H NMR spectra in an enlarged region corresponding to the proton signals of allantoate (5.24 ppm), ureidoglycolate (5.22 ppm), and glyoxylate (5.05 ppm) which appears after hours from the *At*AUL addition.

The positive aspect of this enzymatic approach is obtaining only the enantiomer recognized as substrate by the enzyme; on the other hand, it requires a large amount of (*S*)-allantoin and a significant amount of enzyme to obtain enough substrate to perform our planned investigations. In addition, due to the instability of ureidoglycolate and its tendency to spontaneously hydrolyze in solution, the substrate should be synthesized in conjunction with the experiments and should require an active and stable enzyme, which should be quickly removed to avoid side reactions.

To bypass these limitations, we then synthesized the racemic mixture of (*S/R*)-ureidoglycolate through the urea and glyoxylate chemical condensation according to a previously described procedure[9,82] with some enhancements (see Materials and Methods).

^{1}H and ^{13}C NMR analyses were performed on the white powder obtained (Fig.28A) validating the success of the chemical reaction. Ureidoglycolate proton bound to the chiral carbon was detected at 5.225 ppm in ^{1}H NMR spectrum (Fig.28B), along with a lower signal at 5.055 ppm related to the glyoxylate proton; on top of this, ^{13}C NMR spectrum confirmed the synthesis of ureidoglycolate since its carbons were detected at 76.69 ppm (C2), 163.31 ppm (C4) and 178.49 ppm (C1) consistently with its structure[9], and together with the two signals from glyoxylate (Fig.28C).

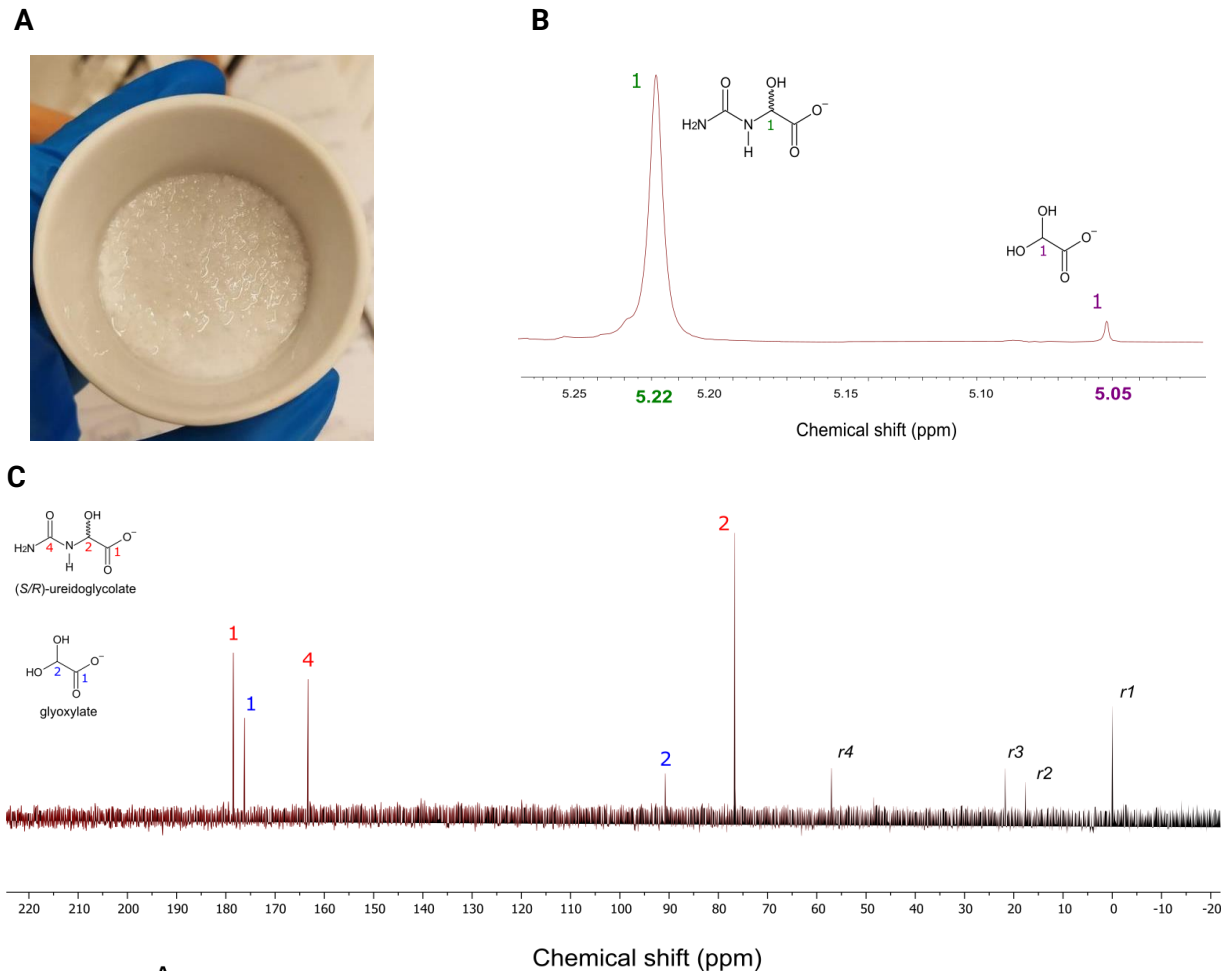


Figure 28: Chemical synthesis of (*S/R*)-ureidoglycolate. (A) Ureidoglycolate white powder was obtained with crystallization and precipitation after 5 hours at 30°C. (B) (*S/R*)-ureidoglycolate (15 mM) ^{1}H NMR spectrum in 50 mM phosphate buffer and 95% d_2O, at room temperature. We observe the proton signal of UG at 5.22 ppm and the proton signal of glyoxylate residues at 5.05 ppm. (C) Ureidoglycolate ^{13}C NMR spectrum (75 mM) resuspended in potassium phosphate buffer 50 mM and 95% d_2O, at room temperature. Carbon signals of UG give peak signals at 76.69 ppm (C2), 163.31 ppm (C4), and 178.49 ppm (C1).

*Dr*MSL production and characterization

The recombinant *Dr*MSL protein was overexpressed in the bacterial host *E. coli* and induced with the addition of lactose and glucose into the medium to avoid basal expression and to reduce the toxicity of protein before induction. Attempts to purify *Dr*MSL after addition of IPTG lead to protein accumulation in inclusion bodies and in the consequent failure to find it in the soluble fraction (data not shown).

Soluble fractions deriving from lactose-induced cells were purified by performing liquid affinity chromatography (FPLC) and *Dr*MSL was extracted from the total protein pool of host cells. The soluble cell fraction was applied to a cobalt resin and the recombinant protein was coordinated with the N-terminal His-Tag to the metal ions of the column; the total process had an average yield of about 3 mg per liter because of the low solubility of the protein. SDS-PAGE confirmed the expected molecular weight for the recombinant protein and its purification (Fig.29A).

*Dr*MSL ureidoglycolate lyase activity and glyoxylate release

We first confirmed the acquisition of the proposed enzymatic activity, i.e. ureidoglycolate lyase, for *Dr*MSL protein by checking the lysis of ureidoglycolate with ^{1}H nuclear magnetic resonance. The rapid decay of the ureidoglycolate proton signal at 5.225 ppm was monitored after the addition of *Dr*MSL freshly purified, together with the simultaneous increase of the novel peak of the diolic proton of glyoxylate at 5.055 ppm (Fig.29B).

We converted peak intensities into chemical species concentrations from time-resolved ^{1}H NMR spectra in order to compare the velocity of the reaction with the spontaneous decay of ureidoglycolate at room temperature (Fig.29C): in fact, ureidoglycolate nonenzymatically hydrolyzes to urea and glyoxylate with a half-life of a few hours in water solution and in presence of bivalent ions[24].

We fitted the experimental data points corresponding to ureidoglycolate spontaneous hydrolysis with the first-order equation $S_{[t]} \sim S_0 * e^{(-k * t)}$ (Fig.29C, dotted black line), and the values of the initial concentration S_0 and the slope k calculated were respectively 51.15 ± 0.42 mM and 8.09 ± 2.29 s^{-1}. From the comparison between ^{1}H NMR time-evolution of ureidoglycolate and glyoxylate proton signals, we observed a significant difference between these two conditions, and it was possible to ascribe changes in spectra over time observed after *Dr*MSL addition to the ureidoglycolate lyase activity of the protein.

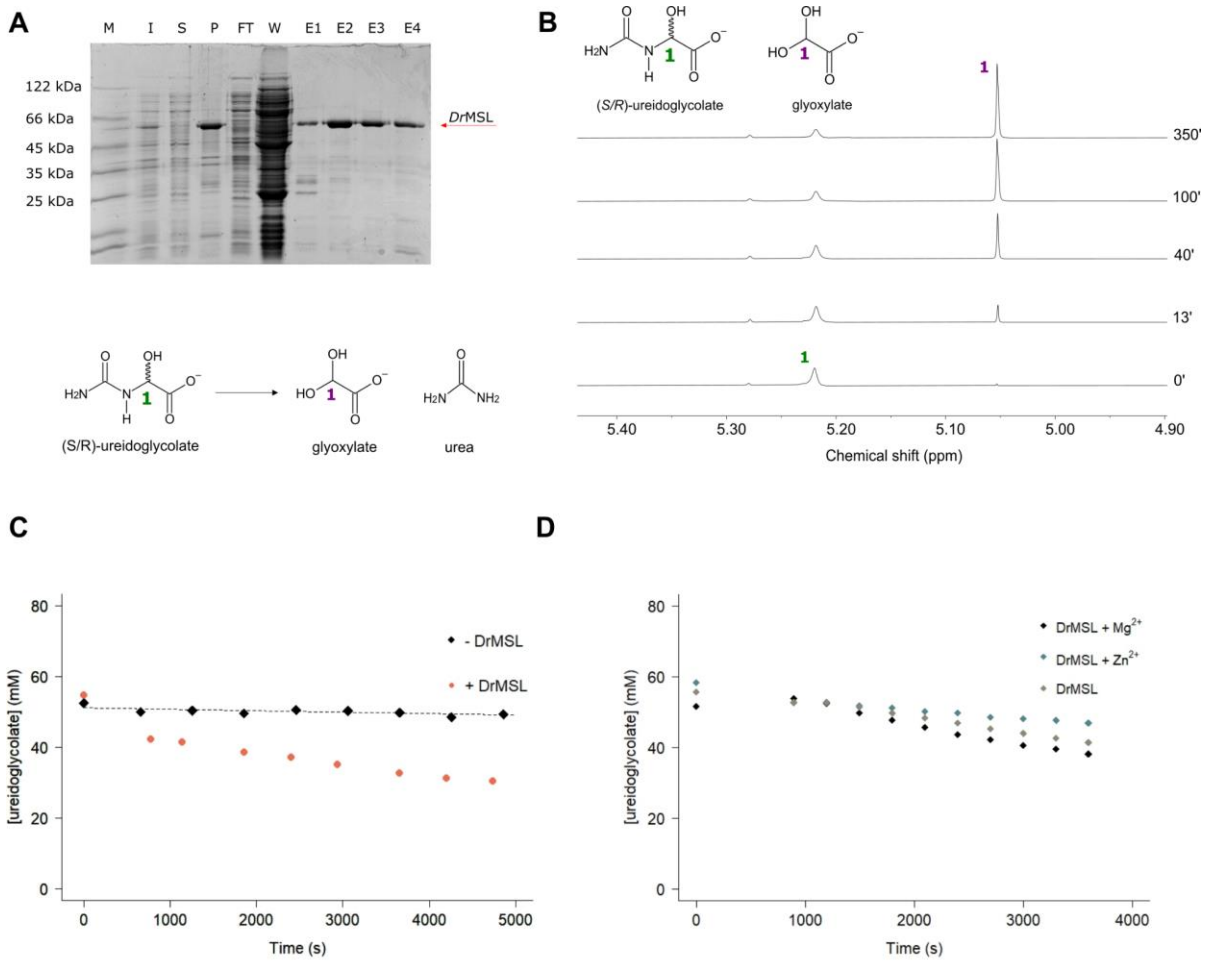


Figure 29: *Dr*MSL recombinant expression, induction, and characterization. (A) SDS-PAGE analysis (12%) of *Dr*MSL protein expression, solubility, and purification with FPLC. Lane M, marker; Lane I, auto-induced expression; Lane S, soluble fraction; Lane P, insoluble fraction; Lane FT, flow-through; Lane W, washing; Lanes E, eluted fractions. (B) Stacked plots of 52.5 mM ureidoglycolate ^{1}H NMR spectrum were recorded at different time points in presence of 2 µM *Dr*MSL and with $MgCl_2$. (C) Ureidoglycolate kinetics followed for spontaneous decay (black diamonds) and for enzymatic activity (red circles) by monitoring proton signals at 5.225 ppm. (D) Ureidoglycolate kinetics were monitored in the presence of the EDTA metal chelator (grey diamonds) and upon incubation with $MgCl_2$ (black diamonds) or $ZnCl_2$ (light blue diamonds).

*Dr*MSL was treated with EDTA for 2 hours to chelate the metals that have bound to the protein during expression and purification steps, and then incubated with divalent ions ($MgCl_2$ and $ZnCl_2$). Time-resolved ^{1}H NMR spectra of *Dr*MSL reactions in the absence or in the presence of metal ions showed no significant differences (Fig.29D), even though *Dr*MSL appeared to hydrolyze ureidoglycolate faster when bound to Mg^{2+}.

However, this approach suffers some limitations: NMR technique has a low sensitivity and requires concentrated samples to detect proton signals[83]. In our experimental conditions, probably the enzyme is saturated and the excess of ureidoglycolate could inhibit its activity,

resulting in underestimation of the reaction rate and in a slower conversion of the substrate to products. Moreover, the addition of paramagnetic metals could interfere with the correct acquisition of NMR spectra[84]; for these reasons, no *Dr*MSL activity in presence of Mn^{2+} has been reported. These limitations were overcome by assaying *Dr*MSL enzymatic activity with a spectrophotometric approach.

*Dr*MSL stereospecificity and selectivity for (*S*)-ureidoglycolate

The persistence of one-half peak of ureidoglycolate after the end of the reaction in ^{1}H NMR spectra can be explained considering the stereospecific activity of most ureidoglycolases which recognize only the (*S*)-ureidoglycolate[6,8].

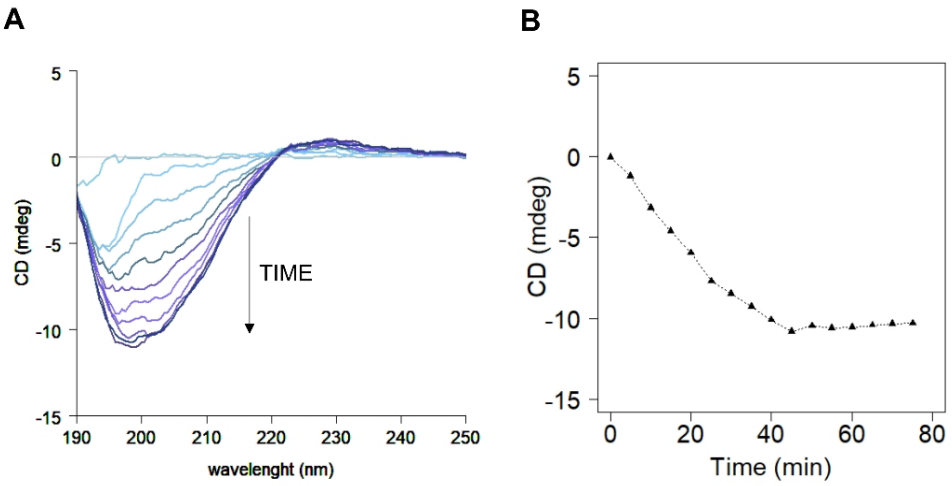


Figure 30: Stereospecificity of *Dr*MSL for (S)-ureidoglycolate. (A) Superimposed time-resolved CD spectra of (*S*)-ureidoglycolate consumption after the addition of 1 µM *Dr*MSL to (*S/R*)-ureidoglycolate racemic mixture, showing accumulation of the (*R*)-ureidoglycolate after the enzymatic reaction. (B) Decrease of CD signal at 200 nm during time after the addition of 1 µM *Dr*MSL.

We confirmed this stereospecificity following the UGL reaction through circular dichroism spectroscopy; specifically, ureidoglycolate lyase activity was followed by recording CD spectra at 5 minutes intervals.

In the presence of the chemically prepared racemic mixture, the accumulation of an optically active compound was observed in the CD spectra, which displayed a negative peak at 200 nm over time (Fig.30A). Peak formation occurs after the enzymatic consumption of (*S*)-ureidoglycolate and is consistent with the persistence of (*R*)-ureidoglycolate in solution. Glyoxylate released during the reaction does not emit CD signals.

Kinetics at 200 nm measured by circular dichroism clearly showed the increment of the *R/S* enantiomer ratio and the negative rise of the CD peak that corresponds to (*R*)-ureidoglycolate (Fig.30A), which is related to (*S*)-ureidoglycolate consumption during the enzymatic activity. After 40 minutes, (*S*)-ureidoglycolate was totally consumed and the spontaneous hydrolysis of the (*R*)-ureidoglycolate into the non-optically active glyoxylate can be detected by monitoring the slow reset of the CD signal (Fig.30B).

Urea release from ureidoglycolate

Ureidoglycolate lyase (EC 4.3.2.3) releases urea and glyoxylate, while ureidoglycolate amidohydrolase (EC 3.5.1.116) found in plants and bacteria releases glyoxylate and ammonia[4,9]. To distinguish between these two enzymatic activities, the release of urea from the ureidoglycolate substrate was confirmed following the enzymatic release of ammonia from urea in presence of urease[25], an enzyme involved in the hydrolysis of urea into one molecule of CO_2 and two molecules of NH_3.

The formation of NH_3 was confirmed by the detection of its reversible condensation with α-ketoglutarate into glutamate with the concomitant oxidation of NAD(P)H to $NAD(P)^+$ or vice versa[86], in presence of glutamate dehydrogenase (GDH). NADH has a characteristic absorbance peak at 340 nm, while NAD^+ does not absorb at this wavelength: variations in the absorbance at this wavelength are indicative of a transition between the oxidized and reduced species.

After the addition of *Dr*MSL to the reaction mixture containing GDH and urease, urea released from ureidoglycolate by *Dr*MSL is enzymatically converted into NH_3, causing a decrease in absorbance at 340 nm (Fig.31A).

The exclusion of urease from the reaction mixture (Fig.31B) and its addition after *Dr*MSL catalysis confirmed the release of NH_3 only from urea and not directly from (*S*)-ureidoglycolate. This evidence confirms the lyase activity for *Dr*MSL as described for bacteria and fungi but only hypothesized for the missing enzyme in fishes; indeed, the activity assay described in the literature[87] monitored only glyoxylate release and did not allow to distinguish between the release of ammonia and the release of urea.

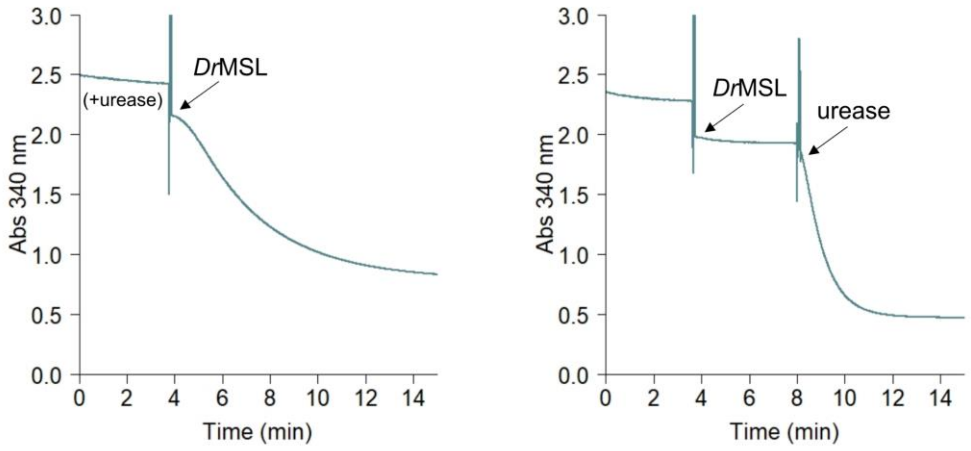


Figure 31: Enzymatic urea release from (*S*)-ureidoglycolate. (A) *Dr*MSL enzymatic activity and urea release from ureidoglycolate were monitored with a continuous coupled assay with GDH at 340 nm in presence of urease in the reaction mixture. (B) *Dr*MSL enzymatic activity and release of ammonia from urea were detected at 340 nm before and after the addition of urease.

*Dr*MSL characterization with LDH coupled assay

We monitored UGL activity spectrophotometrically with a continuous coupled assay with lactate dehydrogenase[87] by following the NADH oxidation at 340 nm consequent to glyoxylate release and reduction to glycolate (Fig.32A). LDH coupled assay has been optimized to not interfere with *Dr*MSL activity: the enzyme was used at the maximum velocity and was added before *Dr*MSL in order to consume the residual glyoxylate to glycolate and not to affect the initial rate evaluation of UGL reaction.

We evaluated the dependence of the initial rate of the hydrolysis by using increasing (*S*)-ureidoglycolate concentrations. Since *Dr*MSL exhibited Michaelis-Menten behavior[88], we estimated kinetics parameters by fitting our experimental data with the Michaelis-Menten equation (Fig.32B). We calculated a turnover number (K_{cat}) of 0.448 ± 0.025 s^{-1}, a Michaelis constant (K_M) of 16.44 ± 4.11 µM for (*S*)-ureidoglycolate and a consequent catalytic efficiency (K_{cat}/K_M) of $2.7*10^4$ s^{-1} M^{-1}.

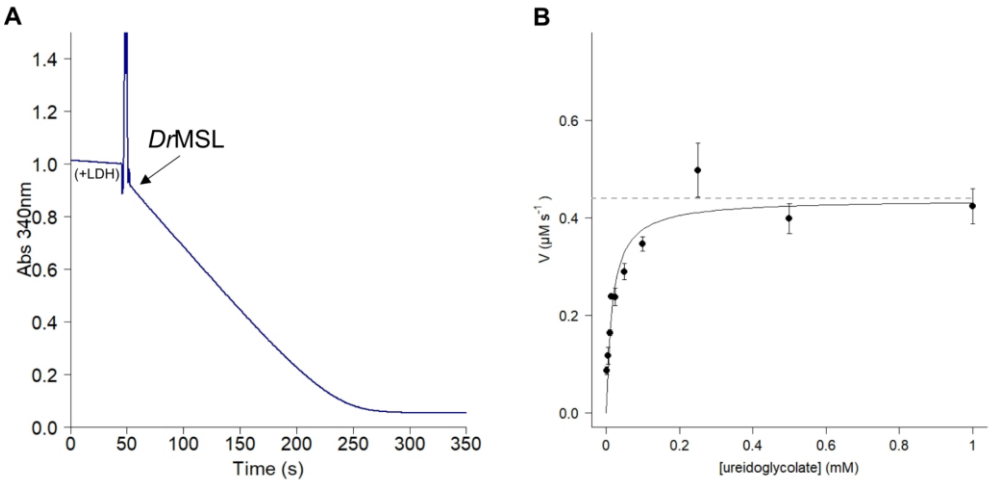


Figure 32: Kinetics characterization of *Dr*MSL (A) Time course of NADH oxidation at 340 nm in the determination of ureidoglycolate lyase activity for *Dr*MSL in presence of LDH. (B) *Dr*MSL activity dependence on the concentration of (*S*)-ureidoglycolate concentrations, measured by the coupled assay with LDH in KP buffer (pH 7.6) at 25 °C. The reaction mixture contained ~0.25 mM NADH, 14 µM LDH, 1 µM *Dr*MSL, and ureidoglycolate. Error bars represent the standard deviation of three independent replicates. Experimental data points were fitted with the Michaelis-Menten equation.

*Dr*MSL has a prevalent monomeric structure

As discussed previously, *Dr*MSL has overall maintained the ancestral tridimensional structure with respect to *E. coli* MSA; to verify the native oligomeric state, we set up a size exclusion chromatography in NaP 20 mM/NaCl 60 mM. Gel filtration revealed the presence of an equilibrium between two different oligomeric states, which probably correspond to the *Dr*MSL monomer and homodimer as detected in fish extracts[26].

The presence of two forms is indicated by the presence of two absorption peaks at 280 nm (Fig.33A); we tested their enzymatic activity by performing LDH coupled assay with 1.94 µg *Dr*MSL and the monomer appears to be more active if compared to the homodimer (Fig.33B).

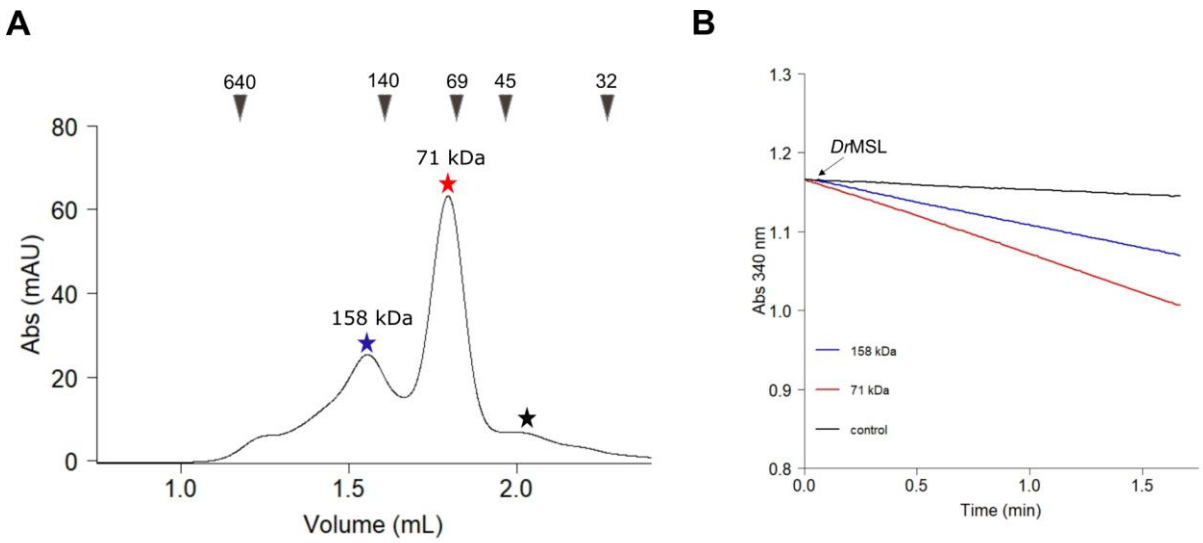


Figure 33: *Dr*MSL structural conformation and relative UGL activity. (A) Size-exclusion chromatography of 1.33 µg/µL DrMSL purified with FPLC. The monomeric form of the protein is pointed by a red star, and the dimeric form by a blue star. The black star indicates the fraction chosen as a negative control to test the enzymatic activity of the eluted fractions. (B) Time-course of NADH oxidation in the determination of ureidoglycolate lyase activity was monitored with a coupled assay at 340 nm after the addition of eluted samples.

We treated the enzyme preparation with a chelating agent (EDTA), and we added different divalent ions (Mg^{2+}, Mn^{2+}, Ca^{2+}, Zn^{2+}) to test the effective requirement of a metal for the correct binding of the substrate and thereby for the UGL reaction catalysis.

In contrast to the metal dependency of homologous eukaryotic malate synthases, which are indeed dependent on the presence of a Mg^{2+} ion in the active site[51,89], *Dr*MSL catalyzes the UGL reaction either in the absence or in the presence of divalent ions and displays only minor differences in the enzymatic activity after treatment with different metals (Fig.34). For example, we observed a modest increase in the activity in the presence of $MnCl_2$ and $MgCl_2$, and a slight decrease in the activity after treatment with $CaCl_2$.

This result is not in agreement with experiments reported in the literature for vertebrate ureidoglycolate lyase[26]; at difference with our procedure, the authors precipitated the protein with ammonium persulfate after its purification from liver tissues. This treatment could have had a strong impact on metal removal if compared with an EDTA treatment, and the addition of the metal ion may have not been enough for protein refolding and for catalysis recovery.

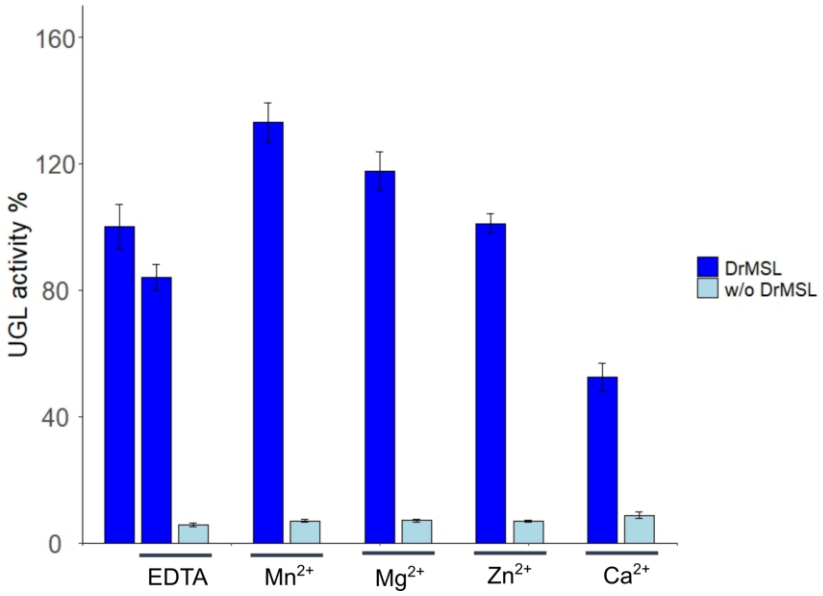


Figure 34: Loss of metal dependency in *Dr*MSL protein. Effects of 2 h treatment with 1 mM EDTA and 1 mM bivalent metal ions incubation on *Dr*MSL UGL activity. Data are expressed as the average ± standard deviation of three independent replicates.

On the other hand, EDTA treatment is an established method for studying the metal dependence of the enzyme and this approach would not fail unless EDTA cannot reach the metal site for structural reasons. In addition to this, our protein was purified mostly in a monomeric form, in contrast to the dimeric form obtained by the authors[26], who did not show the activity of the monomer. Assays performed on different protein complexes have probably led to different experimental results.

Since *Dr*MSL is able to catalyze the reaction even in the absence of divalent ions (Fig.34), the acquisition of a new molecular function could have result in the loss of the central role of the metal ion, which probably has been replaced by polar residues in the substrate binding. Considering the results obtained with SEC analysis and metal dependency investigations, and the evidence from literature[26], we can assume that *Dr*MSL is present in solution as a mixture of monomer and dimer, and that the former is enzymatically active, while the latter is inactive.

*Dr*MSL inability to catalyze the synthesis of ureidoglycolate

Time-resolved ^{1}H NMR spectra confirmed the *Dr*MSL inability to catalyze the inverse reaction, i.e. the synthesis of ureidoglycolate, in the presence of 15 mM glyoxylate and 20 mM urea, since no signal corresponding to ureidoglycolate proton was formed within 3 hours after the enzyme addition (Fig.35A). This evidence accentuated the preference of *Dr*MSL for ureidoglycolate rather than for glyoxylate as supposed from the observation of partial modification of the amino acids surrounding the active site and modification of the substrate binding pocket.

Loss of malate synthase activity in vertebrates

Given the structural similarity to the bacterial malate synthase, the MS catalytic activity was assayed for *Dr*MSL.

We performed the spectrophotometric malate synthase assay measuring absorbance variations at 232 nm suggestive of the Coenzyme A release from acetyl-Coenzyme A with the concomitant malate formation[52] (Fig.35B, green panel); malate synthase activity was not verified with NMR due to the large quantities of acetyl-CoA required to have discrete signals at ^{1}H NMR.

No decrease in absorbance has been noticed after *Dr*MSL addition to reaction mixtures (Fig.35C) containing 0.5 mM glyoxylate and 0.25 mM acetyl-CoA, thus confirming the loss of malate synthase activity. The enzymatic activity was assayed in the presence of 3 mM $MgCl_2$ because of the evidence of a central role of the metal in acid-base catalysis of eukaryotic malate synthases[51].

We also performed a malate synthase assay using ureidoglycolate in place of glyoxylate to exclude that *Dr*MSL could act as a bifunctional enzyme and convert (*S*)-ureidoglycolate to (*S*)-malate in a two steps reaction (Fig.35B, blue panel). No variations in absorbance were observed during time after the addition of ureidoglycolate to the reaction mixture containing *Dr*MSL (Fig.35D) except for the substrate and phosphate buffer contribution so we demonstrated the loss of malate synthase activity for *Dr*MSL as a probable consequence of the breakdown of acetyl-Coenzyme A binding capability.

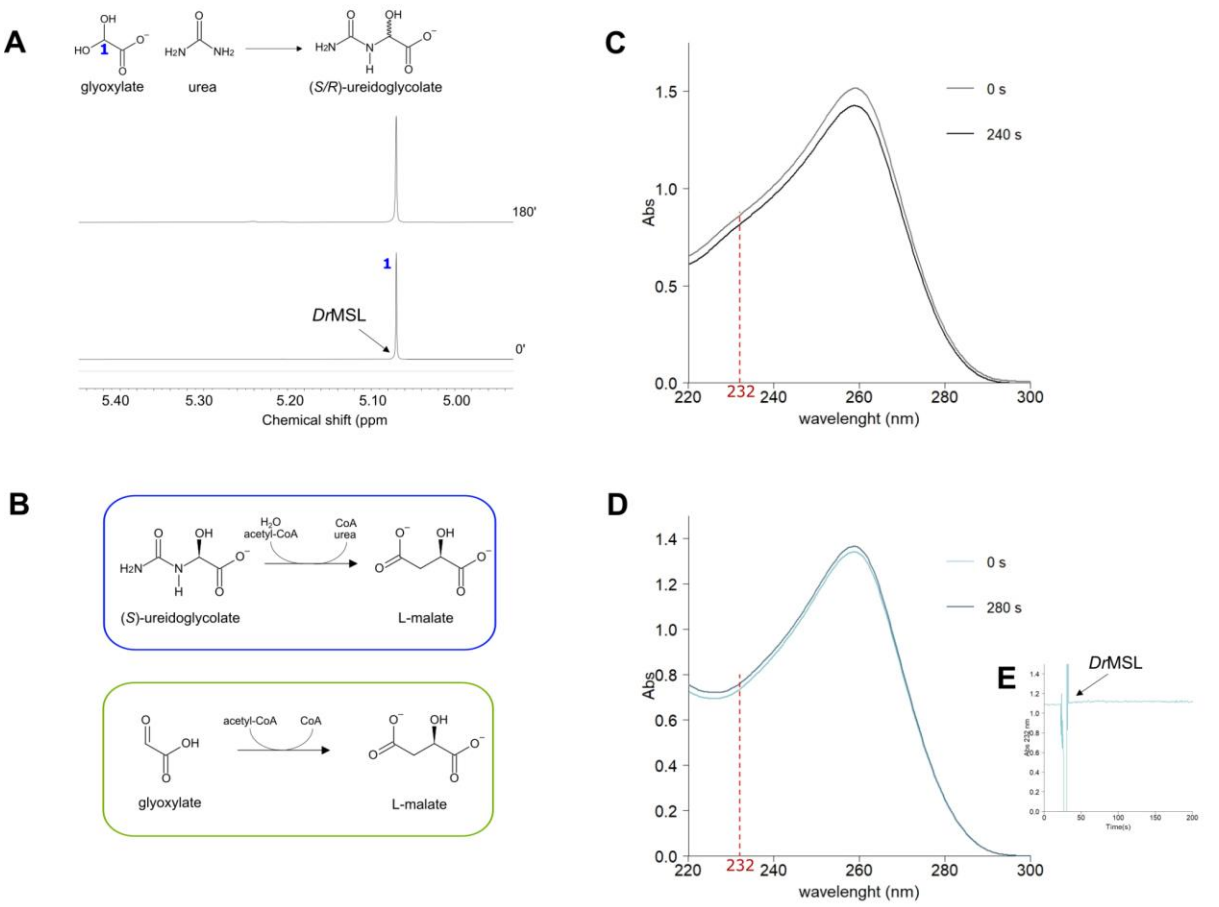


Figure 35: *Dr*MSL failure in UGL reverse reaction and in malate synthase catalysis. (A) Stacked plot of ^{1}H-NMR spectra of 15 mM glyoxylate and 20 mM urea in 95% D_2O recorded before and after the addition of 2 µM *Dr*MSL (3 mM $MgCl_2$). (B) Bifunctional reaction hypotesized for *Dr*MSL, i.e. the synthesis of malate from ureidoglycolate (upper blue panel) and malate synthase reaction (green lower panel). (C) Superimposed spectra of malate synthase assay before (grey line) and after (black line) the addition of 0.5 mM glyoxylate in presence of 2.5 µM *Dr*MSL (220 nm - 300 nm range); no activity has been detected in these conditions. (D) Superimposed spectra of malate synthase assay before (light blue) and after (dark blue) the addition of 0.5 mM ureidoglycolate in presence of 3 µM *Dr*MSL in the reaction mixture (220 nm - 300 nm range); no activity has been detected in these conditions. (E) Time course of CoA release from acetyl-CoA in the determination of *Dr*MSL malate synthase activity monitored at 232 nm in presence of 0.5 mM ureidoglycolate instead of glyoxylate.

Evolutionary and functional divergence of MS and UGL

We then investigated the distribution of MS and MSL genes in eukaryotes and prokaryotes across a phylogenetic reconstruction.

We estimated an unrooted phylogenetic tree of MS and MSL sequences with a maximum likelihood clustering method (Fig.36) and we observed that sequences from bacteria, fungi, plants, green algae and few metazoan known to have MS activity form a separate branch from metazoan sequences clustering with *Dr*MSL.

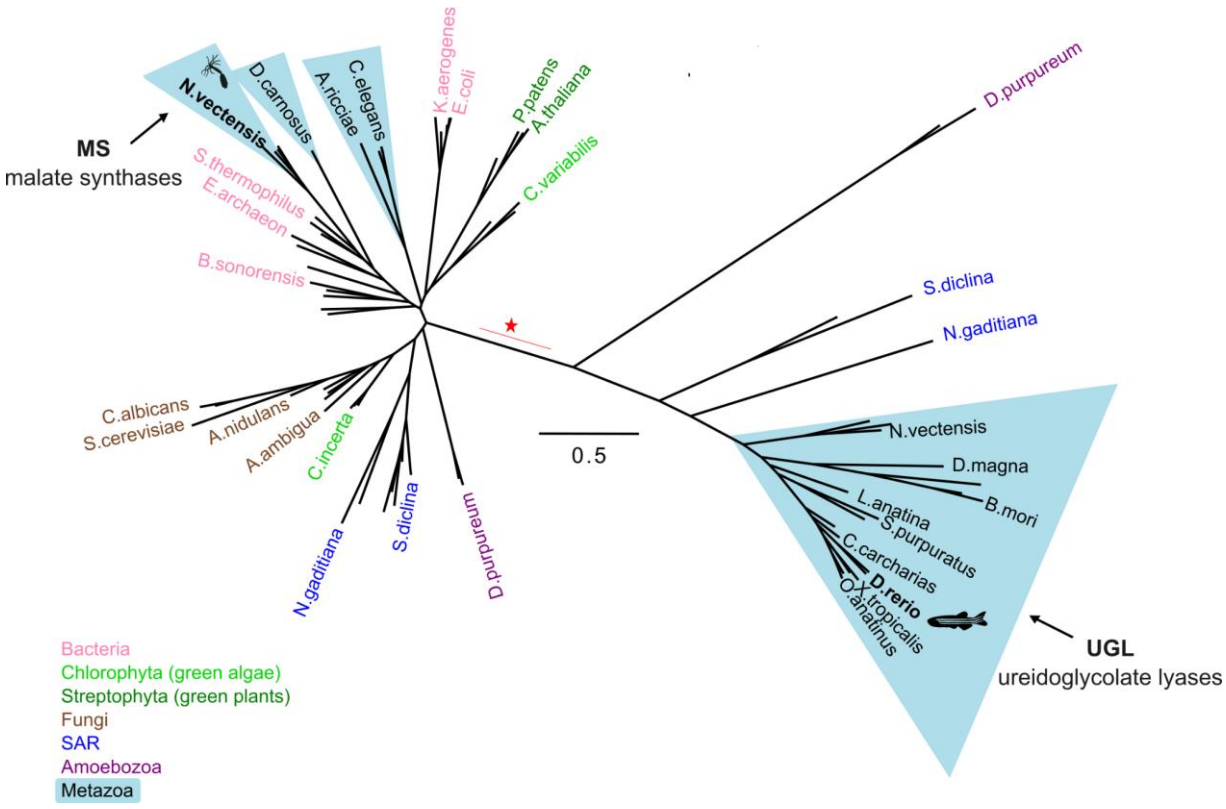


Figure 36: Phylogeny of MS and UGL sequences. Unrooted tree built with Maximum Likelihood estimation using the LG method; the scale bar corresponds to the number of calculated substitutions per site (0.5). Selected terminal organisms are labeled with the abbreviated species name and are colored according to taxonomy as indicated in the legend. Malate synthase (MS) and ureidoglycolate lyase (UGL) in Metazoa are indicated by arrows and included in light blue triangles. Sequences described in this thesis are decorated with animal silhouettes. The branch of inferred genetic duplication between MS and UGL is marked with a red star; the red lines indicate the uncertain node position.

The presence of two distinct groups suggests that MSL originated from duplication of MS followed by neofunctionalization. The sequences of metazoans that cluster together with *Dr*MSL (hereafter "UGL") are found in Amoebozoa, SAR, and Metazoa; therefore, the UGL gene

has been lost in some Insecta groups, in Mollusca, and in Amniota (Mammalia and Sauropsida).

We observed the coexistence of an ancestral MS gene and a duplicate gene in SAR species (*Phytophthora* spp., *Saprolegnia* spp., *Dictyostelium* spp.) and in cnidarians (*Orbicella* spp., *Nematostella* spp., etc). The sequences corresponding to the duplicate gene clusters basal to metazoan UGL sequences.

Sequences that are not clustered either with MS or UGL sequences, i.e. in Amoebozoa, probably share intermediate characteristics with both proteins. This hypothesis is also supported by the presence of intermediate characteristics in protein sequences (Fig.22). *D. purpureum* and other SAR possess two sequences of the MS gene: the first one shares features and is clustered with malate synthases; on the contrary, the other gene copy has lost some residues involved in acetyl-Coenzyme A and has acquired the additional amino acid dyad placed in the active site characteristic of the UGL proteins.

Structural and functional characterization of malate synthase in Metazoa

As evidenced by the phylogenetic analysis, the MS gene is present in duplicate in some organisms, such as marine invertebrates, Amoeboza and Oomycotes probably as a consequence of a duplication event. In particular, *N. vectensis* presents two copies of the gene which are clustered separately in the phylogenetic tree (Fig.37); the sequence that is grouped with UGL proteins probably is the result of the duplication event followed by the acquisition of the new molecular function.

N. vectensis has probably lost the ancestral MS copy of the gene but has acquired an MS copy from bacteria through a horizontal gene transfer (HGT) event. If the MS sequence was the ancestral one, then it would be clustered together with those from SAR and Amoebozoa; furthermore, it has been ruled out that the sequence is the result of bacterial contamination and sequencing errors by analyzing the nearby genes on the chromosome and confirming the presence of introns. In addition, the MS copy emerges from computational analysis to have a peroxisomal localization as expected while the MSL protein is predicted to be cytoplasmatic.

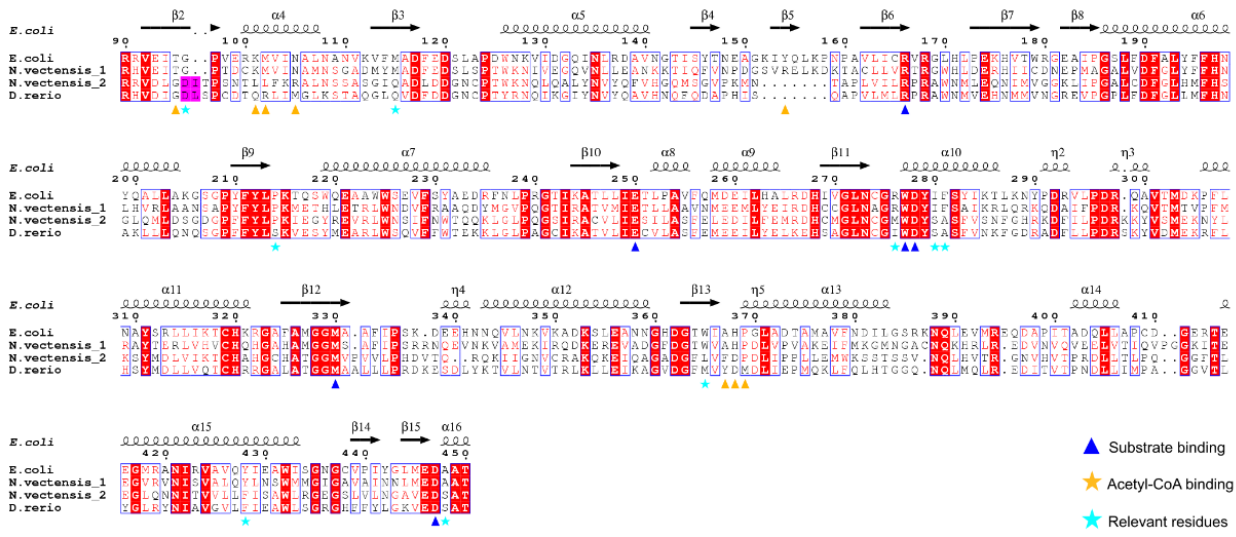


Figure 37: Multiple alignment of selected MS and MSL sequences. Multiple alignment of *Ec*MSA with the *Dr*MSL protein experimentally validated as an ureidoglycolate lyase and the two paralogs from *N.vectensis*. Amino acids conserved in all sequences are shaded red. Residues involved in the metal ion and substrate binding are pointed with blue triangles, and residues involved in acetyl-Coenzyme A binding with orange triangles. Cyan stars indicate residues differently conserved in MS in respect of MSL/UGL. The inserted dyad is shaded purple.

When analyzed for residue conservation in a multiple alignment, the metazoan sequence (Fig.37, N.vectensis_2) grouped with UGL proteins in the ML tree showed loss of acetyl-CoA binding sites if compared with *Ec*MSA. In contrast and as expected, the sequence (N.vectensis_1) clustered with bacterial and other metazoans MS (i.e. *C. elegans*) shares all the relevant residues involved in malate synthase catalysis (Fig.37). Both sequences present strict conservation of the glyoxylate binding site.

To confirm the evidence emerging from the multiple alignment analysis, a 3D model was built for the sea anemone protein (*Nv*MS; XP_001639526.1) possessing MS features. The *Nv*MS 3D structure was modeled using *Ec*MSA (PDB ID: 3CUZ, 3CV2) as template, similarly as done for *Dr*MSL; the QMEAN of the model is 0.83, with a good level of identity between the two sequences (43.35%).

The superimposition reveals a clear similarity between the two protein structures; *Nv*MS (Fig.38, light green cartoon) shares all structural domains with *Ec*MSA (Fig.38, grey cartoon), with little differences in the reciprocal arrangement of the helices due to the synonymous substitution of residues in side chains. Unlike *Dr*MSL, *Nv*MS has maintained almost all the amino acids which bind acetyl-CoA. Therefore, the comparison between the 3D structures

confirms the total maintenance of both the active site and the cofactor binding pocket despite the sequence divergence.

According to the structural and sequence analyses, the starlet sea anemone should present both enzymatic activities.

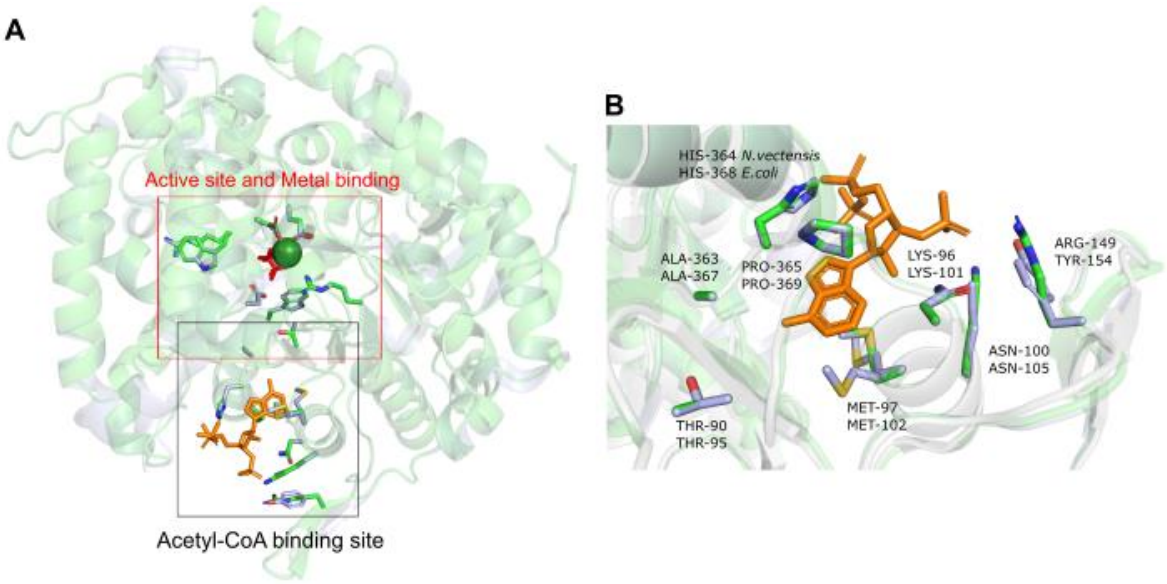


Figure 38: *Nv*MS has maintained the ability of acetyl-CoA binding. (A) Comparison between active site (red square) and Acetyl-CoA binding pocket (black square) for *Nv*MS 3D homology model (light green cartoon) superimposed on the experimental structure of *Ec*MSA (3CV2, grey cartoon). Amino acids involved in MS catalysis are drawn in sticks. (B) Close-up comparison between acetyl-CoA binding sites shows the overall maintenance of the *Nv*MS residues except for the peripheral *Ec*MSA tyrosine substituted with Arg149.

Malate synthase activity in sea anemones

To confirm the maintenance of malate synthase activity in *Nv*MS and in other organisms which possess a gene with MS features, *Nv*MS was overexpressed in *E. coli*, induced with 0.5 mM IPTG and purified using the same procedure described for *Dr*MSL. The purification yielded approximately 14 mg/L and *Nv*MS appeared to be more soluble compared to *Dr*MSL (Fig.39A). We performed the malate synthase assay in presence of 0.5 mM glyoxylate and 0.25 mM acetyl-CoA, and we confirmed the preservation of the malate synthase enzymatic activity, which is evident from the absorbance decrease at 232 nm (Fig.39B, blue line). As expected, *Nv*MS exhibits a close dependence on Mg^{2+} availability in the active site and no activity has been monitored in the absence of the metal ion (Fig.39B, light blue line). The superimposition of spectra acquired in the wavelength range from 220 nm to 300 nm shows a change in the

acetyl-CoA spectrum at wavelengths below 260 nm after the addition of *Nv*MS, confirming the acetyl-CoA conversion into CoA (Fig.39C).

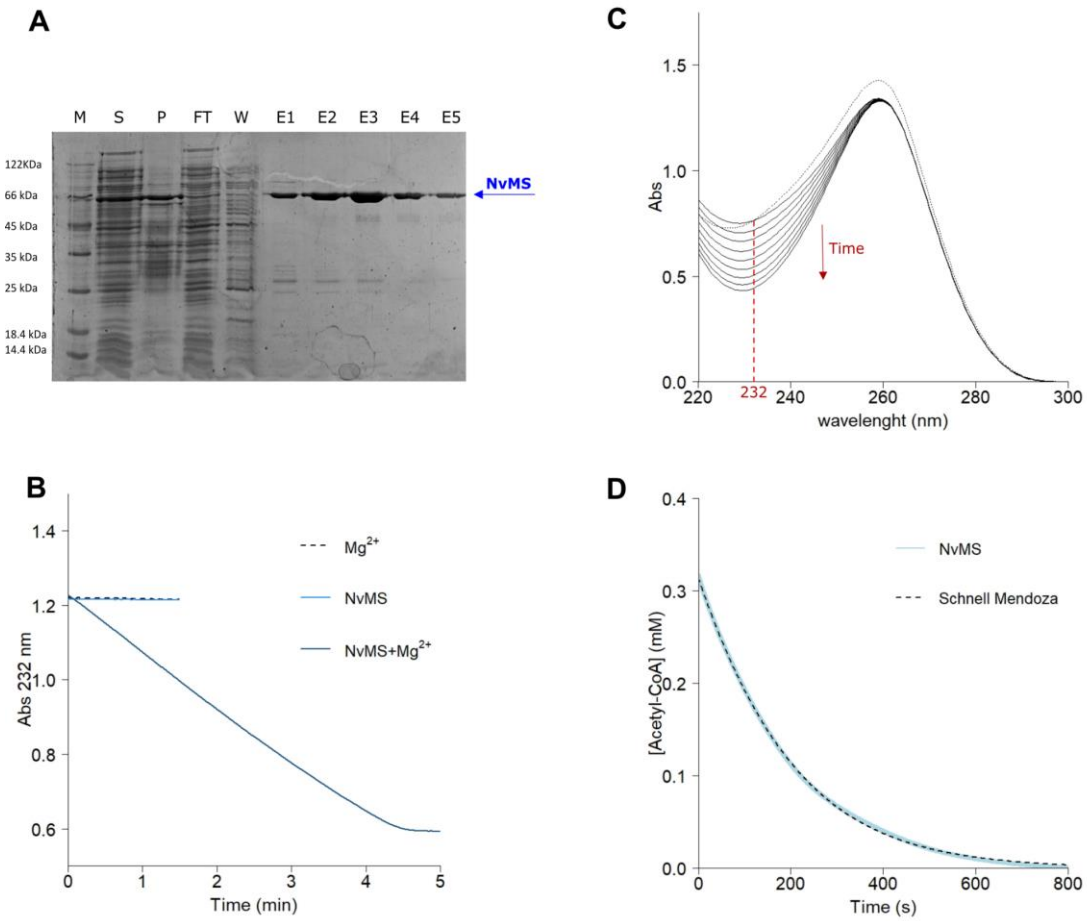


Figure 39: Recombinant expression and characterization of *Nv*MS. A) SDS-PAGE (12%) of *Nv*MS expression and purification through FPLC: M, marker; S, supernatant; P, pellet; FT, flow-through; W, washing; E, elutions. **(B)** Time course of acetyl-CoA condensation monitored at 232 nm. The assay was carried out in presence of 0.25 mM acetyl-CoA and 0.50 mM glyoxylate, in the presence of 1 mM $MgCl_2$ (dashed black line), 0.125 µM *Nv*MS (light blue line) or both $MgCl_2$ and *Nv*MS (dark blue line). (C) Superimposed spectra of malate synthase assay before (dashed black line) and after (solid black lines) the addition of 0.5 mM glyoxylate in presence of 2.5 µM *Dr*MSL (220 nm - 300 nm range) and acquired at 30-seconds intervals. **(D)** Kinetics of acetyl-CoA conversion into CoA (light blue line) at 232 nm fitted with the Schnell Mendoza equation (dashed black line) in presence of 0.5 mM glyoxylate, 0.3 mM acetyl-CoA and 0.125 µM *Nv*MS.

We fitted acetyl-Coenzyme A concentration derived from a single time-course kinetics recorded at 232 nm with the Schnell-Mendoza equation and we evaluated Michaelis-Menten parameters (Fig.39D): a Michaelis constant (K_M) of 1.27 ± 0.03 mM and a turnover number

(K_{cat}) of 59.27 ± 1.35 s^{-1} for acetyl-Coenzyme A and a consequent catalytic efficiency (K_{cat}/K_M) of 4.67*10^4 s^{-1} M^{-1}.

The capability of performing malate synthase activity emerged to be strictly dependent on the presence of a metal ion (Mg^{2+}) from the measurement of the enzymatic activity at 232 nm; *Nv*MS appears to have loosen MS activity in its absence and its specific activity was measured solely if the metal ion is present in the reaction mixture, together with the substrates.

*Dr*MSL and *Nv*MS activities comparison

Malate synthase activity has been observed for *Nv*MS (Fig.40A, black dots), which showed a dependency on the presence of a metal ion for MS catalysis (Fig.40B). Additionally, it does not catalyze the ureidoglycolate lyase reaction (Fig.40C).

*Dr*MSL have maintained only UGL reaction, as detected with LDH coupled assays at 340 nm, and shows no significant variations in activity in the presence of metals (Fig.34), also in the presence of preferred metal (Fig.40C) described for yeasts and bacteria[51,52]; additionally, as predicted from our computational analysis, *Dr*MSL loss malate synthase activity (Fig.40B).

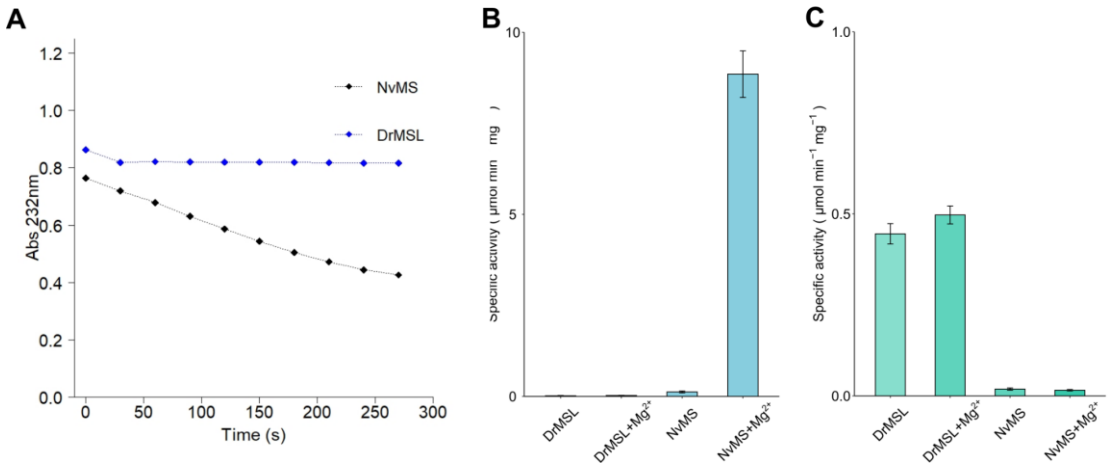


Figure 40: *Nv*MS and *Dr*MSL enzymatic activities. (A) Kinetics of malate synthase assay for *Nv*MS (black diamonds) and *Dr*MSL (blue diamonds) recorded at 30-second intervals at the fixed wavelength of 232 nm, in presence of 0.25 mM acetyl-Coenzyme A and 0.50 mM glyoxylate. (B) Malate synthase specific activities of 2.5 µM *Dr*MSL and 0.125 µM *Nv*MS measured in the presence or absence of $MgCl_2$ at 232 nm. (C) Ureidoglycolate lyase specific activities of 1 µM *Dr*MSL and *Nv*MS 5 µM in the presence and in the absence of $MgCl_2$ measured at 340 nm.

Subcellular compartmentalization of UGL and MS proteins

The final reactions of the purine degradation pathway occur in animal and plant peroxisomes. Uricase is localized in peroxisomes together with catalase which is involved in the decomposition of the H_2O_2 from enzymatic (hypo)xanthine and urate oxidations[90].

Urate oxidase is located in the peroxisomes of fish, amphibians, and mammals, but was lost in some vertebrates due to pseudogenization events[2] during evolution. Other enzymes involved in the pathway could be found in the cytoplasm or mainly in peroxisomes, as for allantoinase and allantoicase[91]. The subcellular localization is diverse in different species and enzymes of purine catabolism have been detected in the cytoplasm, mitochondria, and peroxisome.

Glyoxylate shunt has been described to be localized in glyoxysomes and in peroxisomes for several Plantae clades[92], along with MS and ICL enzymes in yeasts.

With the aim to provide evidence of UGL peroxisomal compartmentalization, we predicted the presence of PTS1 and PTS2 motifs in UGL and MS sequences from the MS gene family and we performed a multiple alignment (Fig.41). Although the N- and C-terminal of the sequences are not fully conserved, both PTS2 and PTS1 are present in many organisms.

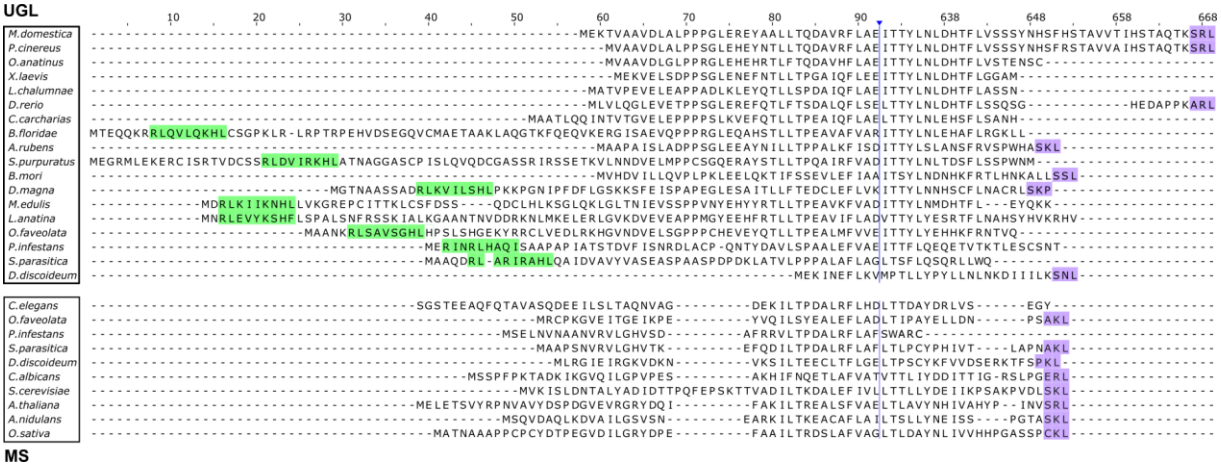


Figure 41: PTS signals in MS and UGL sequences. Multiple alignment of N-terminal and C-terminal UGL (upper box) and MS (lower box) sequences, displayed and modified with Jalview. PTS1 and PTS2 are highlighted in purple and light green; columns 95-630 of the alignment were excluded from the image.

Most eukaryotic proteins appear to be localized in peroxisomes; however, some vertebrate sequences lack PTS signals and could be localized in the cytoplasm, probably due to the differentiation of the ureide degradation pathway during evolution.

To describe subcellular localization of glyoxylate cycle and purine catabolism key components, we compared the presence of PTS signals of UGL and MS sequences with their enzymatic partners, Allc and ICL (Fig.42A).

Considering MS and UGL, signals appear not to be conserved among all the species included in our analysis. PTS1 is spread both in MS and UGL sequences but the amino acid frequencies vary depending on the gene (Fig.42B). MS shows a strict conservation of the tripeptide, in particular the lysine and the leucine at the extreme C-terminus while residues at the first position vary; conversely, the C-terminus of UGL is variable and do not entirely adhere to the original consensus SKL.

PTS2 is exclusively characteristic of UGL sequences and is conserved mostly in basal organisms (SAR) and in marine invertebrates (cnidarians, mytilids, and brachiopods), and rarely in vertebrates, and shows the conservation of the arginine in position 1 and the histidine in the position 8 of the nine-amino acids consensus signal for all the sequences (Fig.42B).

Yeast UGL genes are not represented in this chronogram because *S. cerevisiae* proteins presenting ureidoglycolate lyase activity is encoded by a non-homologous gene that belongs to the DAL operon[93]. These proteins are localized into peroxisomes as other proteins of the pathway. In addition to this, yeast owns two MS gene copies, which are part of the same cluster, but their localization in the cytoplasm or in peroxisomes depends on the growth medium[33]. Although PTS is absent in yeast ICL sequence, proteins are imported into the peroxisome in a Pex5p-dependent manner due to the presence of a non-canonical PTS signal[33] or are directed to the cytoplasm.

MS and ICL are the key enzymes of the glyoxylate cycle and, as in yeast, they need to cooperate in peroxisome in order to catalyze the subsequent isocitrate and acetyl-CoA conversion into succinate and malate[94]. This is evident from their coexistence in the genome of these organisms (Fig.42A).

On the other hand, UGL and Allc catalyze two subsequent reactions to convert allantoate into glyoxylate with the release of two urea molecules, but their subcellular localization may differ. UGL is co-occurrent with Allc due to their involvement in the same pathway but the localization of Allc remains unclear as well as its function, in particular in organisms that do not have the UGL copy and the other genes of the pathway.

In some organisms, only UGL has PTS motifs that determine a peroxisomal localization while Allc lacks any PTSs and are cytoplasmic. Allc shows only PTS1 motifs differently from UGL which has both PTS1 and PTS2 depending on the group of organisms considered.

The different occurrence of PTS signals, which destine soluble proteins to peroxisomes, confirm the heterogeneous subcellular localization of these enzymes despite the strict coevolutionary connection between the purine degradation pathway and the glyoxylate cycle.

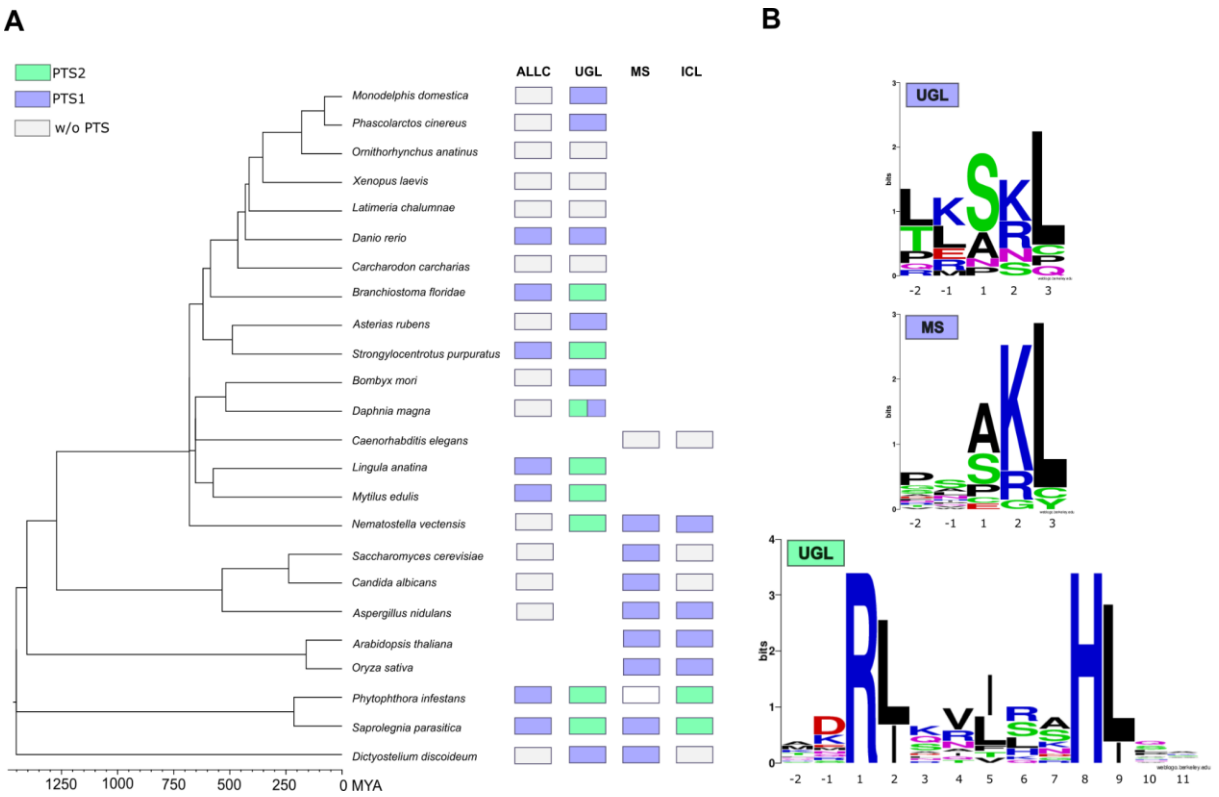


Figure 42: Distribution of PTS in proteins of purine degradation pathway (ALLC and UGL) and glyoxylate cycle (MS and ICL). (A) Eukaryotic phylogeny chronogram considering UGL, MS, ALLC, and ICL genes and the relative presence of PTS1 (purple boxes) and PTS2 (light green boxes) in their protein sequences. Divergence times were derived from TimeTree; *Lingula adamsi, Leptasterias auletica, Strongylocentrotus pallidus* were replaced with *Lingula anatina, Asterias rubens, and Strongylocentrotus purpuratus* to date correctly the species. For the analysis, only organisms that present either the MS or UGL copy were considered. (B) Sequence logos of PTS2 (N-terminal, residues 1-9) and PTS1 (C-terminal, residues 1-3) considering UGL and MS from eukaryotes.

Conclusions

Considering the novel relation between glyoxylate cycle and purine catabolism that emerged from our analysis, our computational method proved to be an effective approach for building phylogenetic profiles and identifying associations between orthogroups with the support of statistical validation. It allows the identification of local correlations in phylogenetic profiles based on a statistical evaluation of the coevolutionary gene transitions significance, without penalizing mismatched profile regions. In addition, it has been shown that the order for eukaryotic species on the phylogenetic tree in the matrix affects either the assessment of gene transition across each orthogroup or the number of concordant transitions calculated for each pair, resulting in different co-transition scores and significance.

A few months ago, the last version of OrthoDB v.11 has been released; it could be interesting to apply our analysis to this extended dataset, which has rebuilt orthogroups and includes about 2000 sequenced genomes. The evaluation of coevolution could also be extended to different datasets comprising orthogroups built with different methods.

Our approach turned out to be robust to predict novel interactions between molecular components involved in related metabolic activities; among them, our results suggested the existence of an evolutionary link between two distinct metabolic pathways in eukaryotes.

As a proof of principle, we validated our computational tool by experimentally demonstrating the functional relationship hypothesized between “malate synthase” orthogroup with “allantoicase” and “isocitrate lyase” orthogroups based on significant *cotr_score* and *p-values*. Supported by structural and sequence analysis of MSL eukaryotic proteins, and in particular on zebrafish MSL, our working hypothesis was that the MS coding gene in Metazoa could have shifted its function to the enzymatic activity of ureidoglycolate lyase.

Preliminary studies on *Dr*MSL have revealed the acquisition of the ureidoglycolate lyase activity and the loss of malate synthase activity as a consequence of the loss of the acetyl-Coenzyme A binding capability and of the active site rearrangement; both the enzymatic activities have been tested by performing specific enzymatic assays.

UGL activity has been monitored with either ^{1}H NMR, or circular dichroism, or a continuous spectrophotometric coupled assay at 340 nm; MS activity has been measured by monitoring the release of Coenzyme A from acetyl-Coenzyme A.

We performed similar structural and functional analysis on *Nv*MS, a protein from *Nematostella vectensis* which interestingly presents typical features of malate synthases, and we

demonstrated the conservation of the ancestral MS enzymatic activity and the consequent maintenance of one of the main activities related to the glyoxylate cycle. In fact, we have experimentally confirmed the different substrate preferences for *Nv*MS and *Dr*MSL under our reaction conditions, and we evaluated the different impact of the metal ion in the active site for both proteins.

The phylogenetic analysis performed considering MS and MSL sequences allowed us to clarify the evolution of the malate synthase gene family.

To explain the occurrence of two gene copies in some organisms and the existence of two subgroups in "Malate synthase" orthogroup, we assumed a possible scenario that implies a gene duplication event preceding the separation between Amoebozoa and Metazoa more than 1.4 Gya[95]. It remains unclear whether the ancestral copy of the gene possessed a malate synthase function or encoded a bifunctional protein with both MS and UGL activities. In the first scenario of neofunctionalization, the gene duplication event could have been followed by the loss of the acetyl-CoA binding domain necessary for MS catalysis and the neofunctionalization into UGL.

In the alternative scenario of subfunctionalization, the two distinct gene copies could have originated from a bifunctional ancestral gene in early eukaryotes and then could have maintained only one of the two activities. The hypothesis of gaining malate synthase function starting from an ancestral UGL gene is less plausible due to the several sequence mutations needed to acquire the ability to bind the cofactor.

The maintenance of the MSL gene and the acquisition of a MS gene through a HGT event from bacteria has led to the possibility for these organisms to maintain the glyoxylate cycle and, at the same time, to end the purine catabolism with the release of glyoxylate.

Moreover, the loss of the ICL gene in all Metazoa except for nematodes and cnidarians was confirmed by our co-occurrence analysis and represents clear evidence of the maintenance of the glyoxylate cycle in organisms which have acquired the real MS.

This metabolic connection gives these animals the advantage to use the purine ring of nucleic acids in presence of acetyl-CoA as an energy source for their metabolism since the glyoxylate cycle allows the recycle of lipid metabolites to produce sugars for gluconeogenesis.

According to this scenario, ureidoglycolate and glyoxylate seem to have had a crucial role in connecting Krebs cycle metabolisms and purine degradation in early eukaryotes (Fig.43), serving as a source of intermediates for ATP production through the electron transport chain.

From our result, interesting conclusions can be drawn:

- The analysis of concordant transitions between orthogroups pairs could be a successful approach in the detection of novel interactions between molecular components.
- *Malate synthase* genes have been maintained in organisms lacking the glyoxylate cycle and have changed their function into ureidoglycolate lyase.
- Evolutionary relatedness between MS and UGL genes suggests that the original function was the malate synthase activity and that the ureidoglycolate lyase activity occurred after a gene duplication event in early eukaryotes. After this event, UGL activity has been lost in Mammalia and Sarcopterygii as a consequence of the earlier truncation of the purine degradation pathway during evolution.
- Early eukaryotes present a functional MS gene copy which, together with ICL, participates in the glyoxylate cycle.

Further studies will focus on the application of our phylogenetic profile analysis and on solving functional relations not yet clarified.

Regarding the specific case of MS and UGL, further experiments with site-directed mutagenesis will clarify the role of critical amino acids that have determined the loss of the ancestral malate-synthase activity and the acquisition of the new ureidoglycolate lyase activity.

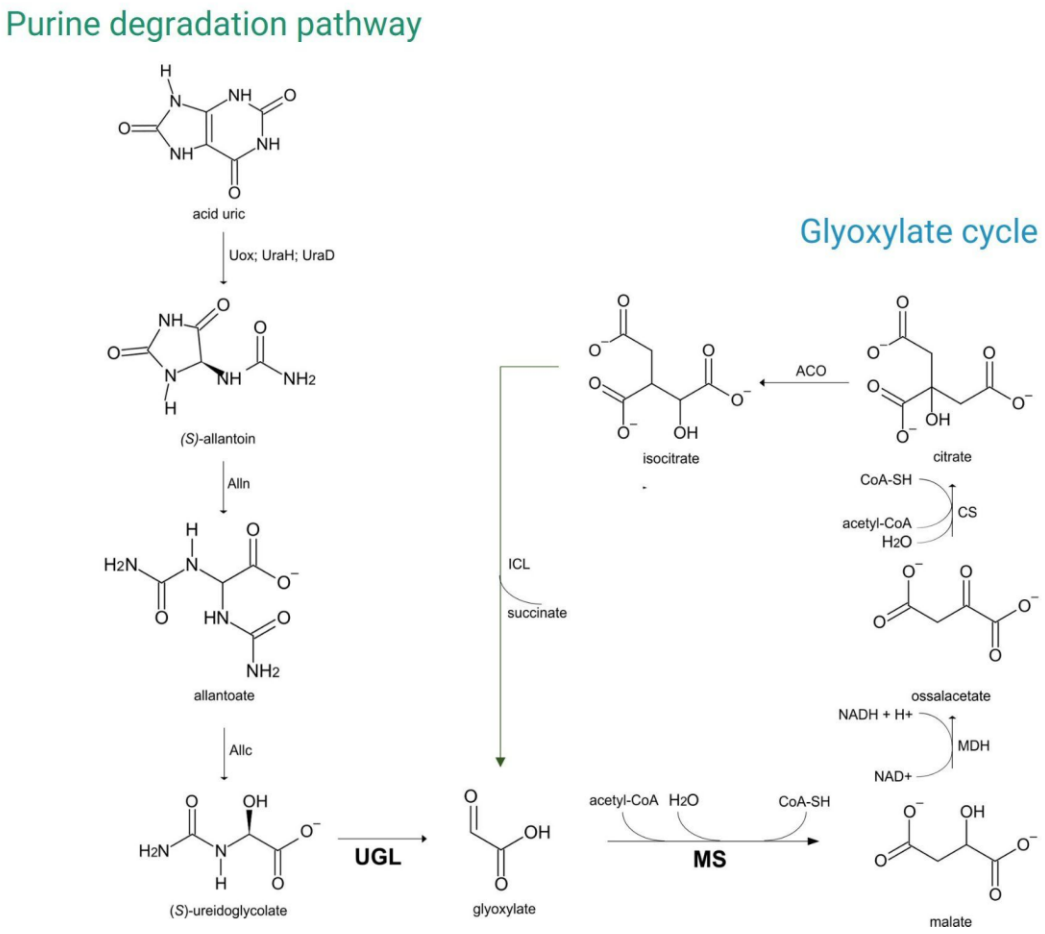


Figure 43. Purine degradation pathway and glyoxylate cycle. The metabolic connection between the Krebs cycle and purine degradation was hypothesized for early eukaryotes.

Materials and Methods

Phylogenetic profile construction

Orthogroups were downloaded from the OrthoDB database (v. 10.1)[58] and parsed with R scripts to generate gene tables of presence and absence. Data were filtered to retain only orthogroups at the Eukaryota level (according to the OrthoDB hierarchy) and present in at least 1% of genomes.

Genome columns were ordered according to the NCBI tree (684 internal nodes) or to a fully resolved tree (1263 internal nodes) obtained with RAxML (v.8.2.12)[96] using the BINCAT model on the transposed profile matrix and the NCBI tree as a constraint.

The 'ladderize' function of the ape package[97] was used to build different tree orders. Viridiplantae was chosen as the starting node in eukaryote phylogeny to polarize unrooted trees and determine leaf orders in case of uncertain roots because casual polarization of the eukaryotic tree can produce incoherent splitting as observed in Opisthokonta.

Co-transition analysis

Cotransitions enumeration was used to implement a memory-efficient algorithm in Python (https://github.com/lab83bio/Cotransitions), considering '1' for the presence and '0' for the absence of a gene in orthogroups (rows) in a list of species (columns) ordered according to phylogeny. Sets of column positions with present->absent ('-1') and absent->present ('1') transitions were determined for each orthogroup.

The intersection of the two sets of vectors was used to obtain the number of concordant (same sign) and discordant (different sign) transitions for each pairwise comparison.

The *cotr_score* (co-transition score) was calculated with an R script as

$$cotr\_score = k / (t_1+t_2-abs(k)),$$

t1 and t2 are the total number of transitions for orthogroups 1 and 2, and k is the difference between concordant and discordant transitions for the pairwise comparison. Correlated transitions range from 0 to 1, and anticorrelated transitions from 0 to -1.

The significance *p-value,* i.e., the probability to obtain by chance an observed *cotr_score*, was calculated through the one-tailed Fisher's exact test.

The statistical analysis was applied to a 2x2 contingency table, considering n as the total number of positions (or genomes) in the transition vector:

$abs(k)$	$t1 - abs(k)$	$t1$
$t2 - abs(k)$	$n - t1 - t2 + abs(k)$	$n - t1$
$t2$	$n - t2$	n

Fisher's exact test can also be applied to the 2x2 table of presence/absence transitions similarly to 2x2 table of presence/absence states of two genes.

P-values for individual tests were adjusted for multiple tests using the Holm correction, obtaining *adjusted p-values*.

To obtain co-evolving modules, orthogroup pairs with adjusted p-values $<10^{-3}$ were clustered with MCL using the $-\log_{10}(p\_adj)$ as a similarity measure, and an inflation ('-I') parameter of 2.5.

Pathway analysis and mapping with GO and KEGG databases

The dataset of orthogroups selected using our filters were annotated with Uniprot-mapped GO terms if experimental annotation (evidence codes: 'IDA', 'IMP', 'IPI', 'IEP', 'IGI') was available for at least one orthogroup member, as well as with KEGG maps and modules of the general and metabolism section of the KEGG pathway database.

Uniprot accessions were obtained for each orthogroup from Orthodb xrefs files (odb10v1_gene_xrefs.tab.gz) and the entire GO annotation dataset (goa_uniprot_all.gaf) was downloaded from Uniprot. Each orthogroup was then associated to GO annotations, considering at least one experimental GO for the orthogroup. Among the 4705005 pairs, there were 46951 unique orthogroups; of these 13324 were associated with 950787 GO Uniprot: 494250 Biological Process, 240736 Cellular Component, and 215801 Molecular Function. The selected GOs were mapped to their SLIM version with the subset 'goslim_generic' or 'goslim_pir' subsets from OWLTOOLS software (https://github.com/owlcollab/owltools).

Orthogroup KEGG maps and modules were retrieved through the OrthoDB API using the orthogroup (og) id as a query ("https://v101.orthodb.org/group?id={og}"). Maps and modules of the metabolism section (kid="00xxxx") were considered separately from the general pathway section.

Orthogroup Set Enrichment Analysis (OSEA)

GSEA function (https://rdrr.io/pkg/clusterProfiler/man/GSEA.html) of the "clusterProfiler" R library (http://yulab-smu.top/biomedical-knowledge-mining-book/universal-api.html)[98] was used to perform the enrichment analysis because it was designed to accept user-defined annotation.

Orthogroup and *p-value* dataset was ordered on the base of the lowest significance value considering the entire set of pairs for each orthogroup. The *p-values* -log10(x) were transformed with the reciprocal square root function of the scipy.stats.boxcox library and normalized in the range -6 to +6.

Python package pygosemsim (https://github.com/mojaie/pygosemsim) was used to calculate the GO semantic similarity between gene pairs in order to determine the biological similarity between orthogroups.

Sequence and structure analysis of "malate synthase" and "malate synthase-like" proteins

Sequences of MS and MS-like proteins for phylogenetic and functional analysis were downloaded either from OrthoDb v.10.1[58] or by performing a homology search in BlastP[99] (NCBI) to derive the complete sequence of proteins present in partial form or annotated erroneously in *"malate synthase-like"* orthogroup (354580at2759).

Multiple alignments of amino acid sequences were performed using ClustalX 2.0[100], modified with Jalview 2.11.2.[101], and displayed with graphical enhancements using ESPript 3.0[102].

The resolved structure of a bacterial "malate synthase A"[50] (PDB ID: 3CUZ, 3CVZ) was taken as a reference to identify residues with a relevant role in the function and structure of proteins included in the analysis. 3CUZ was also used as the template to perform *Dr*MSL and *Nv*MS protein modeling in SWISS-MODEL[103]; both models were structurally aligned and visualized with PyMol 2.5[104].

The phylogenetic tree was inferred with the maximum likelihood estimation using the "Le Gascuel" (LG) model selected by the automatic procedure[105] in PhyML 3.0[106]; it has been visualized with a radial layout and unrooted with FigTree 1.4. Eukaryotic phylogeny and chronograms were supported by the evolutionary timescale derived from TimeTree[95].

Protein sequences were scanned with PSort I (plant sequences) and PSort II (animal and yeast sequences) servers to make predictions about protein cellular localization; peroxisomal compartmentalization was confirmed by performing homology research in PHIBlast with PTSs

patterns[107]. The predicted N-terminal and C-terminal signal peptides were used to generate frequency plots using WebLogo 2.8.

Sequence and structural analysis images were decorated with animal silhouettes downloaded from PhyloPic and modified using Inkscape 1.1[108].

*Dr*MSL identification as the candidate for UGL activity

Danio rerio putative peroxisomal malate synthase (*Dr*MSL) was identified as a possible candidate for UGL activity with a brief set of commands and the analysis returned two results: "malate synthase-like" protein and "protein brambleberry precursor" (see Supplementary Information).

*At*AUL preparation and allantoic acid chemical synthesis

Allantoate was chemically synthesized by mixing 2.8 g of allantoin in 20 mL of 1 M KOH at room temperature for 30 min. The reaction mixture was then placed in ice and precipitated overnight with 9 volumes of EtOH 96%.

To obtain the *At*AUL protein in a recombinant form, we followed the procedure described by Puggioni and colleagues[6] and we confirmed the correct purification of the enzyme loading eluted samples on a 12% SDS PAGE gel. *At*AUL enzymatic activity was assayed following the allantoate consumption and the (*S*)-ureidoglycolate formation with ^{1}H NMR spectroscopy.

Potassium ureidoglycolate synthesis protocol

To test ureidoglycolate lyase activity, (*S/R*)-ureidoglycolate was synthesized from urea and glyoxylate condensation following an adapted procedure from literature[9].

The recipe for the synthesis is described below:

- 5 g of glyoxylic acid monohydrate (Sigma-Aldrich)
- 3 g of solid KOH (Alfa Aesar)
- 8 g of urea (VWR Life Science)
- 2 mL of cold ddH_2O
- KOH 1 N solution (VWR BDH Chemicals)
- ethanol (EtOH) 96% (VWR BDH Chemicals)

Glyoxylic acid and KOH tablets were mixed with cold water and additional KOH was added to adjust the pH to 7.0. Once the powders have been dissolved, urea was added to the solution;

the reaction mixture was heated to 30 °C and vigorously mixed for approximately five hours. The product of the condensation was precipitated overnight at -20 °C using 225 mL of EtOH 96%, which corresponds to 9 volumes with respect to the reaction mixture.

The precipitate was collected on a filter paper using a vacuum pump and a funnel, washed with EtOH 96%, and re-precipitated in 50 mL of water and 450 mL of EtOH 96% overnight. The white powder obtained after two precipitations was then re-filtered and crystallized in 1 part of ddH_2O and 9 parts of EtOH 96% increasing and lowering the temperature.

We obtained approximately 6.5 g of white powder using this protocol; we confirmed the successful synthesis of the racemic ureidoglycolate through ^{13}C and ^{1}H NMR spectroscopy.

(*S/R*)-ureidoglycolate was freshly resuspended in phosphate buffer and maintained in ice just prior the usage.

Vector construction and protein expression and purification

Danio rerio "malate synthase-like" CDS sequence (NM_001201408.1) and *Nematostella vectensis* "malate synthase" (XM_001639476.3) were cloned into pET-28a(+)-TEV plasmid from GenScript Biotech; these expression vectors carry an additional 6xHisTag and a TEV protease recognition site at the N-terminal of the recombinant proteins.

Chemical and physical protein parameters, useful to optimize purification protocol, were calculated using tools provided by ExPasy[109]. *Dr*MSL and *Nv*MS isoelectric point (pI), extinction coefficients (in units of M^{-1} cm^{-1}, at 280 nm), and molecular mass were calculated with ProtParam considering the additional sequences of the protein expressed in recombinant form.

Both the constructs were electroporated into the bacterial host *Escherichia coli* BL21-CodonPlus DE3 strain (Novagen) and grown on LB agar medium added with antibiotics (50 µg/mL kanamycin for plasmid resistance, 34 µg/mL chloramphenicol for host resistance).

Single positive clones were added to LB broth (1% NaCl, 1% tryptone, 0.5% yeast extract) and gene expression was auto induced for 16 h at 20 °C in presence of 0.05% lactose and 0.2% glucose. Alternatively, induction was carried out using 0.5 mM IPTG for *Nv*MS clones. Cells were collected by centrifugation (8000 g, 15 min, 4 °C), and stored at -20 °C.

Pellets were resuspended in a proper lysis buffer (50 mM NaH_2PO_4, 150 mM NaCl, pH 8.0, 10 % glycerol), sonicated (35-40 W, 1 s on - 1 s off for 30 min), and then harvested for 45 minutes, 14000 rpm at 4 °C. Overexpressed proteins were separated from the soluble fractions and

purified on an FPLC system for affinity chromatography (Akta Pure 25 M, GE Healthcare) using a cobalt-charged column (HisTrap HP 5 mL) and taking advance of the 6xHisTag.

After the removal of contaminants with a washing buffer (50 mM NaH_2PO_4, 150 mM NaCl, 20 mM imidazole, pH = 8.0), proteins were eluted from the column with elution buffer (same as washing, plus 500 mM imidazole). A VivaSpin™ protein concentrator (Cytiva) with a suitable cutoff (50 kDa) was used to concentrate proteins (50 mM KH_2PO_4, 150 mM NaCl, pH 7.8) and remove imidazole and other contaminants. Protein fractions were loaded on 12% SDS-PAGE gel to confirm the correct protein size. The amount of protein purified was quantified spectrophotometrically at 280 nm with Lambert-Beer law. Proteins were stored at -80 °C after the addition of 10% glycerol.

Activity assay with NMR spectroscopy assays

Nuclear magnetic resonance was used to assay ureidoglycolate spontaneous decay in water solution, ureidoglycolate lyase activity, and the reverse synthesis of ureidoglycolate starting from glyoxylate and urea.

Enzymatic and non-enzymatic reactions were followed with ^1H NMR spectra, which were acquired every 3 or 5 minutes in kinetic mode at 25 °C with a Jeol ECZ600R spectrometer in non-spinning mode (spectral width as 14450, relaxation delay as 5 s, size as 65536 data points).

Reaction mixtures were prepared in 5 mm diameter tubes with 50 mM KH_2PO_4 in 90% d_2O and 10% H_2O; the DANTE presat sequence was applied to reduce water signal during spectra acquisition.

About 55 mM (*S/R*)-ureidoglycolate was used to follow its spontaneous decay during the time and to assay for UGL activity; 10 mM glyoxylate and 15 mM urea were used for texting the reverse reaction. Except for the ureidoglycolate spontaneous hydrolysis, the reactions were started with the addition of 1 µM or 2 µM *Dr*MSL previously incubated with 1 mM $MgCl_2$.

Changes in proton peak areas in ^1H NMR spectra were consistent with the release of glyoxylate (increase in 5.055 ppm peak) from ureidoglycolate (decrease in 5.225 ppm peak).

Spectra were revised using MestReNova v.14.2[110].

Proton peak areas were integrated and have been converted into concentration data to fit experimental data of ureidoglycolate spontaneous decay with the first-order equation:

$$[S] = [S]_0 * \exp(-k*t)$$

where [S] is the time-dependent concentration of substrate, $[S]_0$ is the initial concentration of the substrate in the reaction mixture, t is the time since the beginning of the reaction, and k is the pseudo-first-order rate constant.

Activity assays with circular dichroism spectrophotometry

The stereospecificity of the UGL reaction was confirmed using circular dichroism, following the persistence of (*R*)-ureidoglycolate in the reaction mixture after the consumption of the (*S*)-enantiomer in presence of the enzyme.

The reaction mixture was prepared in a 1 mm-path-length transparent quartz cuvette and contained 1 µM *Dr*MSL in 10 mM potassium phosphate buffer (KH_2PO_4/K_2HPO_4). UGL reaction was started by adding 2.5 mM (*S/R*)-ureidoglycolate and spectra were acquired at 5 minutes intervals. (*S/R*)-ureidoglycolate spectrum was obtained in absence of the enzyme in the mixture to compare the overall CD signal before and after the enzymatic reaction.

CD spectra were recorded in 190-250 nm far-UV region using a Jasco J-1500 Circular Dichroism Spectrophotometer and measurements were acquired at 5-minute intervals. The temperature was fixed at 25°C with a Peltier thermostatic cell.

UGL activity assay with UV-visible spectrophotometry

Ureidoglycolate lyase activity was estimated using a stoichiometric and continuous coupled assay with LDH (from rabbit muscle, Sigma) as described previously[87].

The reaction started with the addition of 1 µM *Dr*MSL to the reaction mixture containing different concentrations of the racemic solution of ureidoglycolate, and LDH together with its substrate NADH in 50 mM KH_2PO_4, pH 7.6. Once the reaction was started, (*S*)-ureidoglycolate was hydrolyzed to glyoxylate, which was then reduced to glycolate by LDH with the simultaneous NADH oxidation. Conversion of NADH to NAD^+ was detected by an absorbance decrease at the fixed wavelength of 340 nm.

The enzymatic activity was recorded using a JASCO V-750 UV-Visible Spectrophotometer equipped with a thermostat, placing the reaction mixture in a black quartz cuvette, and maintaining the temperature at 25 °C.

Kinetics parameters of ureidoglycolate lyase reaction were determined at 25 °C in the presence of increasing (*S*)-ureidoglycolate concentrations; the corresponding initial velocities (V_0) for 1 µM *Dr*MSL were fitted to evaluate enzymatic kinetics constant for UGL activity with the Michaelis Menten equation (DRC package in R). Intervals of 30 seconds were considered for

measuring the slope in absorbance signal at 340 nm and to estimate V_0 using NADH extinction coefficient (6220 M^{-1} cm^{-1}).

UGL metal dependency analysis

For the analysis of metal dependency, *Dr*MSL aliquots were treated for two hours with 1 mM EDTA (ethylenediaminetetraacetic acid) to chelate any metal contaminants. The solution was then ultra-filtered with a 50 kDa Vivaspin™ concentrator to retain enzymes and to remove metals bound to the chelator; enzymes fractions were then incubated with 1 mM of four different divalent metals ($MnCl_2$, $MgCl_2$, $ZnCl_2$, $CaCl_2$).

Enzymatic activity dependency on metal ions was determined using 2.5 mM (*S/R*)-ureidoglycolate and 0.25 mM NADH as described previously.

Urea release assay with UV-visible spectrophotometry

The *Dr*MSL-dependent release of urea from (*S*)-ureidoglycolate was confirmed at 340 nm with a continuous coupled assay with glutamate dehydrogenase (GDH from the bovine liver; Sigma-Aldrich) and urease (from jack bean, Fisher Chemical).

1 mM α-ketoglutarate, 0.200 mM (*R/S*)-ureidoglycolate, 0.500 mM NADH, 5.5 U GDH were mixed in 20 mM potassium phosphate, pH 7.6, and the release of urea was monitored after or before the addition of 4 U urease respect of 1 µM *Dr*MSL.

MS activity assay with UV-visible spectrophotometry

Malate synthase activity assay was performed for both *Dr*MSL and *Nv*MS by following the decrease in absorbance from 220 nm to 260 nm as a consequence of Coenzyme A release from acetyl-CoA during glyoxylate condensation to form malate.

The reaction mixtures contained 0.5 mM glyoxylate and 0.25 mM acetyl-CoA in 50 mM KH_2PO_4 pH 7.6; different enzyme concentrations were assayed for MS activity in the presence or absence of 3 mM $MgCl_2$. Enzymatic reactions and acetyl-CoA/CoA spectra were acquired in Spectra Measurement and Time course measurement modes.

Single kinetics at 232 nm was recorded following the absorbance change between acetyl-CoA and CoA; the molar extinction difference of approximately 2000 M^{-1} cm^{-1} was used to evaluate the initial rate of the reaction as described previously[111]. Catalytic parameters (Michaelis constant K_M and maximum velocity V_{MAX}) were calculated in presence of 0.125 µM *Nv*MS fitting

single kinetics with Schnell-Mendoza equation[112] (lamW package; RStudio) and K_{cat} was calculated in consequence:

$[S] = K_M * lambertW_0([S]_0/K_M * exp(([S]_0 - V_{MAX} * t)/K_M))$

$K_{cat} = V_{MAX} / [E]$

where [S] is the time-dependent concentration of substrate, $[S]_0$ is the initial concentration of the substrate in the reaction mixture, t is the time since the beginning of the reaction, K_M is the constant which describes the affinity of the enzyme to its substrate, V_{MAX} is the maximum velocity raised by the enzyme, K_{cat} is the turnover constant at saturating substrate concentrations, and [E] is the enzyme concentration.

*Nv*MS metal dependence was evaluated by comparing time-evolution spectra at the fixed wavelength of 232 nm in the presence and absence of magnesium.

MS assay was performed also using 0.5 mM ureidoglycolate instead of glyoxylate as the substrate to test a *Dr*MSL putative bifunctional activity.

Oligomeric state analysis with size exclusion chromatography

Size exclusion chromatography (SEC) was performed in triplicate with a Superdex 200 column on Akta Pure 25 M, loading 75 µg (1.30 mg/mL) of *Dr*MSL/*Nv*MS in a proper buffer (20 mM NaH_2PO_4 pH 7.0, 50 mM NaCl). The elution was monitored at 280 nm and 1.94 µg from fractions corresponding to absorbance peaks were assayed and compared spectrophotometrically with LDH coupled assay.

The molecular mass of proteins eluted was deduced by setting a calibration curve constructed loading commercial analytical standards: thyroglobulin, bovine serum albumin, ovalbumin, trypsin, lysozyme, and conalbumin.

Data analysis

Experimental data were analyzed, processed, and graphically represented using packages in R version 4.1.2 and RStudio and aesthetically modified with Inkscape 1.1. Chemical and enzymatic reactions were drawn using ChemSketch.

Bibliography

1. Lee, I. R. *et al.* Characterization of the Complete Uric Acid Degradation Pathway in the Fungal Pathogen Cryptococcus neoformans. *PLoS ONE* 8, e64292 (2013).
2. Enzymes involved in purine metabolism - A review of histochemical localization and functional implications. *Histol. Histopathol.* 1321–1340 (1999) doi:10.14670/HH-14.1321.
3. Ramazzina, I., Folli, C., Secchi, A., Berni, R. & Percudani, R. Completing the uric acid degradation pathway through phylogenetic comparison of whole genomes. *Nat. Chem. Biol.* 2, 144–148 (2006).
4. Werner, A. K., Romeis, T. & Witte, C.-P. Ureide catabolism in Arabidopsis thaliana and Escherichia coli. *Nat. Chem. Biol.* 6, 19–21 (2010).
5. McInnch, J. K., McInnch, J. D. & May, S. W. Catalysis, stereochemistry, and inhibition of ureidoglycolate lyase. *J. Biol. Chem.* 278, 50091–50100 (2003).
6. Puggioni, V. *et al.* Gene context analysis reveals functional divergence between hypothetically equivalent enzymes of the purine-ureide pathway. *Biochemistry* 53, 735–745 (2014).
7. Zrenner, R., Stitt, M., Sonnewald, U. & Boldt, R. PYRIMIDINE AND PURINE BIOSYNTHESIS AND DEGRADATION IN PLANTS. *Annu. Rev. Plant Biol.* 57, 805–836 (2006).
8. Todd, C. D. *et al.* Update on ureide degradation in legumes. *J. Exp. Bot.* 57, 5–12 (2006).
9. Winkler, R. G., Blevins, D. G. & Randall, D. D. Ureide Catabolism in Soybeans: III. Ureidoglycolate Amidohydrolase and Allantoate Amidohydrolase Are Activities of an Allantoate Degrading Enzyme Complex. *Plant Physiol.* 86, 1084–1088 (1988).
10. Hafez, R. M., Abdel-Rahman, T. M. & Naguib, R. M. Uric acid in plants and microorganisms: Biological applications and genetics - A review. *J. Adv. Res.* 8, 475–486 (2017).
11. Ramazzina, I. *et al.* An aminotransferase branch point connects purine catabolism to amino acid recycling. *Nat. Chem. Biol.* 6, 801–806 (2010).
12. Méndez-Salazar, E. O. & Martínez-Nava, G. A. Uric acid extrarenal excretion: the gut microbiome as an evident yet understated factor in gout development. *Rheumatol. Int.* 42, 403–412 (2022).
13. Kim, M.-I., Shin, I., Cho, S., Lee, J. & Rhee, S. Structural and Functional Insights into (S)-Ureidoglycolate Dehydrogenase, a Metabolic Branch Point Enzyme in Nitrogen Utilization. *PLoS ONE* 7, e52066 (2012).
14. Wong, S. & Wolfe, K. H. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat. Genet.* 37, 777–782 (2005).
15. Osbourn, A. E. & Field, B. Operons. *Cell. Mol. Life Sci.* 66, 3755–3775 (2009).
16. Hayashi, S., Fujiwara, S. & Noguchi, T. Evolution of Urate-Degrading Enzymes in Animal Peroxisomes. *Cell Biochem. Biophys.* 32, 123–129 (2000).
17. Park, J. H., Jo, Y.-I. & Lee, J.-H. Renal effects of uric acid: hyperuricemia and hypouricemia. *Korean J. Intern. Med.* 35, 1291–1304 (2020).
18. Kand'ár, R., Žáková, P. & Mužáková, V. Monitoring of antioxidant properties of uric acid in humans for a consideration measuring of levels of allantoin in plasma by liquid chromatography. *Clin. Chim. Acta* 365, 249–256 (2006).

19. Allen, R. N., Shukla, M. K., Burda, J. V. & Leszczynski, J. Theoretical Study of Interaction of Urate with Li^{+}, Na^{+}, K^{+}, Be^{2+}, Mg^{2+}, and Ca^{2+} Metal Cations. *J. Phys. Chem. A* 110, 6139–6144 (2006).
20. Becker, B. F. Towards the physiological function of uric acid. *Free Radic. Biol. Med.* 14, 615–631 (1993).
21. Tana, C., Ticinesi, A., Prati, B., Nouvenne, A. & Meschi, T. Uric Acid and Cognitive Function in Older Individuals. *Nutrients* 10, 975 (2018).
22. Mandal, A. K. & Mount, D. B. The Molecular Physiology of Uric Acid Homeostasis. *Annu. Rev. Physiol.* 77, 323–345 (2015).
23. Kahn, K., Serfozo, P. & Tipton, P. A. Identification of the True Product of the Urate Oxidase Reaction. *J. Am. Chem. Soc.* 119, 5435–5442 (1997).
24. 's-Gravenmade, E. J., Vogels, G. D. & Van der Drift, C. Hydrolysis, racemization and absolute configuration of ureidoglycolate, a substrate of allantoicase. *Biochim. Biophys. Acta BBA - Enzymol.* 198, 569–582 (1970).
25. Percudani, R., Carnevali, D. & Puggioni, V. Ureidoglycolate hydrolase, amidohydrolase, lyase: how errors in biological databases are incorporated in scientific papers and vice versa. *Database* 2013, bat071–bat071 (2013).
26. Takada, Y. & Noguchi, T. Ureidoglycollate lyase, a new metalloenzyme of peroxisomal urate degradation in marine fish liver. *Biochem. J.* 235, 391—397 (1986).
27. Kong, F., Romero, I. T., Warakanont, J. & Li-Beisson, Y. Lipid catabolism in microalgae. *New Phytol.* 218, 1340–1348 (2018).
28. Dolan, S. K. & Welch, M. The Glyoxylate Shunt, 60 Years On. *Annu. Rev. Microbiol.* 72, 309–330 (2018).
29. Cheah, H.-L., Lim, V. & Sandai, D. Inhibitors of the Glyoxylate Cycle Enzyme ICL1 in Candida albicans for Potential Use as Antifungal Agents. *PLoS ONE* 9, e95951 (2014).
30. Lewin, A. S., Hines, V. & Small, G. M. Citrate synthase encoded by the CIT2 gene of Saccharomyces cerevisiae is peroxisomal. *Mol. Cell. Biol.* 10, 1399–1405 (1990).
31. Beinert, H. & Kennedy, M. C. Aconitase, a two-faced protein: enzyme and iron regulatory factor [1 2]. *FASEB J.* 7, 1442–1449 (1993).
32. Minárik, P., Tomásková, N., Kollárová, M. & Antalík, M. Malate dehydrogenases--structure and function. *Gen. Physiol. Biophys.* 21, 257–265 (2002).
33. Kunze, M., Kragler, F., Binder, M., Hartig, A. & Gurvitz, A. Targeting of malate synthase 1 to the peroxisomes of *Saccharomyces cerevisiae* cells depends on growth on oleic acid medium: Subcellular localization of yeast Mls1p. *Eur. J. Biochem.* 269, 915–922 (2002).
34. Piekarska, K. *et al.* The activity of the glyoxylate cycle in peroxisomes of Candida albicans depends on a functional β-oxidation pathway: evidence for reduced metabolite transport across the peroxisomal membrane. *Microbiology* 154, 3061–3072 (2008).
35. Cornah, J. E., Germain, V., Ward, J. L., Beale, M. H. & Smith, S. M. Lipid Utilization, Gluconeogenesis, and Seedling Growth in Arabidopsis Mutants Lacking the Glyoxylate Cycle Enzyme Malate Synthase. *J. Biol. Chem.* 279, 42916–42923 (2004).
36. Chew, S. Y. *et al.* Glyoxylate cycle gene ICL1 is essential for the metabolic flexibility and virulence of Candida glabrata. *Sci. Rep.* 9, 2843 (2019).

37. Idnurm, A. & Howlett, B. J. Isocitrate Lyase Is Essential for Pathogenicity of the Fungus *Leptosphaeria maculans* to Canola (*Brassica napus*). *Eukaryot. Cell* 1, 719–724 (2002).
38. Dunn, M. F., Ramírez-Trujillo, J. A. & Hernández-Lucas, I. Major roles of isocitrate lyase and malate synthase in bacterial and fungal pathogenesis. *Microbiology* 155, 3166–3175 (2009).
39. Kinhikar, A. G. *et al.* Mycobacterium tuberculosis malate synthase is a laminin-binding adhesin. *Mol. Microbiol.* 60, 999–1013 (2006).
40. Liu, F., Thatcher, J. D., Barral, J. M. & Epstein, H. F. Bifunctional Glyoxylate Cycle Protein of Caenorhabditis elegans: A Developmentally Regulated Protein of Intestine and Muscle. *Dev. Biol.* 169, 399–414 (1995).
41. Erkut, C., Gade, V. R., Laxman, S. & Kurzchalia, T. V. The glyoxylate shunt is essential for desiccation tolerance in C. elegans and budding yeast. *eLife* 5, e13614 (2016).
42. Erkut, C. & Kurzchalia, T. V. The C. elegans dauer larva as a paradigm to study metabolic suppression and desiccation tolerance. *Planta* 242, 389–396 (2015).
43. Khan, F. R. & McFadden, B. A. Embryogenesis and the glyoxylate cycle. *FEBS Lett.* 115, 312–314 (1980).
44. Kondrashov, F. A., Koonin, E. V., Morgunov, I. G., Finogenova, T. V. & Kondrashova, M. N. Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol. Direct* 1, 31 (2006).
45. Mainguet, S. E., Gronenberg, L. S., Wong, S. S. & Liao, J. C. A reverse glyoxylate shunt to build a non-native route from C4 to C2 in Escherichia coli. *Metab. Eng.* 19, 116–127 (2013).
46. Zhang, J. Z., Gomez-Pedrozo, M., Baden, C. S. & Harada, J. J. Two classes of isocitrate lyase genes are expressed during late embryogeny and postgermination in Brassica napus L. *Mol. Gen. Genet. MGG* 238–238, 177–184 (1993).
47. Sonia Beeckmans. *Glyoxylate cycle*.
48. Wierenga, R. K. The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* 492, 193–198 (2001).
49. Kruse, C. & Kindl, H. Malate synthase: Aggregation, deaggregation, and binding of phospholipids. *Arch. Biochem. Biophys.* 223, 618–628 (1983).
50. Lohman, J. R., Olson, A. C. & Remington, S. J. Atomic resolution structures of Escherichia coli and Bacillus anthracis malate synthase A: comparison with isoform G and implications for structure-based drug discovery. *Protein Sci. Publ. Protein Soc.* 17, 1935–1945 (2008).
51. Schmid, G., Durchschlag, H., Biedermann, G., Eggerer, H. & Jaenicke, R. Molecular structure of malate synthase and structural changes upon ligand binding to the enzyme. *Biochem. Biophys. Res. Commun.* 58, 419–426 (1974).
52. Durchschlag, H., Biedermann, G. & Eggerer, H. Large-Scale Purification and Some Properties of Malate Synthase from Baker's Yeast. *Eur. J. Biochem.* 114, 255–262 (1981).
53. Smith, C. V. *et al.* Biochemical and Structural Studies of Malate Synthase from Mycobacterium tuberculosis. *J. Biol. Chem.* 278, 1735–1743 (2003).
54. McVey, A. C. *et al.* Structural and Functional Characterization of Malate Synthase G from Opportunistic Pathogen *Pseudomonas aeruginosa*. *Biochemistry* 56, 5539–5549 (2017).

55. Anstrom, D. M., Kallio, K. & Remington, S. J. Structure of the *Escherichia coli* malate synthase G:pyruvate:acetyl-coenzyme A abortive ternary complex at 1.95 Å resolution. *Protein Sci.* 12, 1822–1832 (2003).
56. Howard, B. R., Endrizzi, J. A. & Remington, S. J. Crystal Structure of *Escherichia coli* Malate Synthase G Complexed with Magnesium and Glyoxylate at 2.0 Å Resolution: Mechanistic Implications ‡. *Biochemistry* 39, 3156–3168 (2000).
57. Zambuzzi-Carvalho, P. F. *et al.* The malate synthase of *Paracoccidioides brasiliensis Pb* 01 is required in the glyoxylate cycle and in the allantoin degradation pathway. *Med. Mycol.* 47, 734–744 (2009).
58. Kriventseva, E. V. *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 47, D807–D811 (2019).
59. Barido-Sottani, J., Chapman, S. D., Kosman, E. & Mushegian, A. R. Measuring similarity between gene interaction profiles. *BMC Bioinformatics* 20, 435 (2019).
60. Dey, G., Jaimovich, A., Collins, S. R., Seki, A. & Meyer, T. Systematic Discovery of Human Gene Function and Principles of Modular Organization through Phylogenetic Profiling. *Cell Rep.* 10, 993–1006 (2015).
61. Huttener, R. *et al.* Sequencing refractory regions in bird genomes are hotspots for accelerated protein evolution. *BMC Ecol. Evol.* 21, 176 (2021).
62. Enright, A. J. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584 (2002).
63. Sonego, P., Kocsor, A. & Pongor, S. ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief. Bioinform.* 9, 198–209 (2008).
64. Deutekom, E. S., van Dam, T. J. P. & Snel, B. Phylogenetic profiling in eukaryotes: The effect of species, orthologous group, and interactome selection on protein interaction prediction. *PLOS ONE* 17, e0251833 (2022).
65. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* 95, 6073–6078 (1998).
66. Hsieh, S.-Y. *et al.* An Enhanced Algorithm for Reconstructing a Phylogenetic Tree Based on the Tree Rearrangement and Maximum Likelihood Method. in *Intelligent Computing Theories and Methodologies* (eds. Huang, D.-S., Jo, K.-H. & Hussain, A.) vol. 9226 530–541 (Springer International Publishing, 2015).
67. Moi, D., Kilchoer, L., Aguilar, P. S. & Dessimoz, C. Scalable phylogenetic profiling using MinHash uncovers likely eukaryotic sexual reproduction genes. *PLOS Comput. Biol.* 16, e1007553 (2020).
68. Zdobnov, E. M. *et al.* OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 49, D389–D393 (2021).
69. Altenhoff, A. M. *et al.* OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* 49, D373–D379 (2021).
70. Rahban, R. & Nef, S. CatSper: The complex main gate of calcium entry in mammalian spermatozoa. *Mol. Cell. Endocrinol.* 518, 110951 (2020).

71. Nepal, M., Che, R., Zhang, J., Ma, C. & Fei, P. Fanconi Anemia Signaling and Cancer. *Trends Cancer* 3, 840–856 (2017).
72. Wolfson, R. L. *et al.* KICSTOR recruits GATOR1 to the lysosome and is necessary for nutrients to regulate mTORC1. *Nature* 543, 438–442 (2017).
73. Bailey, S. S. *et al.* The role of conserved residues in Fdc decarboxylase in prenylated flavin mononucleotide oxidative maturation, cofactor isomerization, and catalysis. *J. Biol. Chem.* 293, 2272–2287 (2018).
74. Green, M. L. & Karp, P. D. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5, 76 (2004).
75. Muñoz, A., Piedras, P., Aguilar, M. & Pineda, M. Urea Is a Product of Ureidoglycolate Degradation in Chickpea. Purification and Characterization of the Ureidoglycolate Urea-Lyase. *Plant Physiol.* 125, 828–834 (2001).
76. Takada, Y. & Tsukiji, N. Peroxisomal localization and activation by bivalent metal ions of ureidoglycolate lyase, the enzyme involved in urate degradation in Candida tropicalis. *J. Bacteriol.* 169, 2284–2286 (1987).
77. Nakazawa, M. *et al.* Characterization of a Bifunctional Glyoxylate Cycle Enzyme, Malate Synthase/Isocitrate Lyase, of Euglena gracilis: A BIFUNCTIONAL GLYOXYLATE CYCLE ENZYME. *J. Eukaryot. Microbiol.* 58, 128–133 (2011).
78. Davis, W. L., Jones, R. G. & Goodman, D. B. Cytochemical localization of malate synthase in amphibian fat body adipocytes: possible glyoxylate cycle in a vertebrate. *J. Histochem. Cytochem.* 34, 689–692 (1986).
79. Kamoshita, M. *et al.* Insights Into the Peroxisomal Protein Inventory of Zebrafish. *Front. Physiol.* 13, (2022).
80. Laskowski, R. A. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* 29, 221–222 (2001).
81. Bracken, C. D. *et al.* Crystal structures of a halophilic archaeal malate synthase from Haloferax volcanii and comparisons with isoforms A and G. *BMC Struct. Biol.* 11, 23 (2011).
82. Valentine R.C., Wolfe R.S. Glyoxylurea. *Biochemical and biophysical research communications.* 5, 305–308.
83. Emwas, A.-H. M. The Strengths and Weaknesses of NMR Spectroscopy and Mass Spectrometry with Particular Focus on Metabolomics Research. in *Metabonomics* (ed. Bjerrum, J. T.) vol. 1277 161–193 (Springer New York, 2015).
84. Otting, G. Protein NMR Using Paramagnetic Ions. *Annu. Rev. Biophys.* 39, 387–405 (2010).
85. Shin, I., Han, K. & Rhee, S. Structural Insights into the Substrate Specificity of (S)-Ureidoglycolate Amidohydrolase and Its Comparison with Allantoate Amidohydrolase. *J. Mol. Biol.* 426, 3028–3040 (2014).
86. Spanaki, C. & Plaitakis, A. The Role of Glutamate Dehydrogenase in Mammalian Ammonia Metabolism. *Neurotox. Res.* 21, 117–127 (2012).
87. Pineda, M., Piedras, P. & Cárdenas, J. A continuous spectrophotometric assay for ureidoglycolase activity with lactate dehydrogenase or glyoxylate reductase as coupling enzyme.*l. Biochem.* 222, 450–455 (1994).

88. Seibert, E. & Tracy, T. S. Fundamentals of Enzyme Kinetics: Michaelis-Menten and Non-Michaelis–Type (Atypical) Enzyme Kinetics. in *Enzyme Kinetics in Drug Metabolism* (eds. Nagar, S., Argikar, U. A. & Tweedie, D.) vol. 2342 3–27 (Springer US, 2021).
89. Dixon, G. H., Kornberg, H. L. & Lund, P. Purification and properties of malate synthetase. *Biochim. Biophys. Acta* 41, 217–233 (1960).
90. Villalobos-García, D. & Hernández-Muñoz, R. Catalase increases ethanol oxidation through the purine catabolism in rat liver. *Biochem. Pharmacol.* 137, 107–112 (2017).
91. Noguchi, T., Takada, Y. & Fujiwara, S. Degradation of uric acid to urea and glyoxylate in peroxisomes. *J. Biol. Chem.* 254, 5272–5275 (1979).
92. De Bellis, L., Picciarelli, P., Pistelli, L. & Alpi, A. Localization of glyoxylate-cycle marker enzymes in peroxisomes of senescent leaves and green cotyledons. *Planta* 180, 435–439 (1990).
93. Yoo, H. S., Genbauffe, F. S. & Cooper, T. G. Identification of the ureidoglycolate hydrolase gene in the DAL gene cluster of Saccharomyces cerevisiae. *Mol. Cell. Biol.* 5, 2279–2288 (1985).
94. Kunze, M., Pracharoenwattana, I., Smith, S. M. & Hartig, A. A central role for the peroxisomal membrane in glyoxylate cycle function. *Biochim. Biophys. Acta BBA - Mol. Cell Res.* 1763, 1441–1452 (2006).
95. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22, 2971–2972 (2006).
96. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313 (2014).
97. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528 (2019).
98. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* 2, 100141 (2021).
99. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410 (1990).
100. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948 (2007).
101. Procter, J. B. *et al.* Alignment of Biological Sequences with Jalview. in *Multiple Sequence Alignment* (ed. Katoh, K.) vol. 2231 203–224 (Springer US, 2021).
102. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* 42, W320-324 (2014).
103. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303 (2018).
104. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
105. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321 (2010).
106. Lefort, V., Longueville, J.-E. & Gascuel, O. SMS: Smart Model Selection in PhyML. *Mol. Biol. Evol.* 34, 2422–2424 (2017).
107. Kunze, M. The type-2 peroxisomal targeting signal. *Biochim. Biophys. Acta Mol. Cell Res.*

1867, 118609 (2020).
108. Inkscape Project, 2020. Inkscape, Available at: https://inkscape.org.
109. Duvaud, S. *et al.* Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res.* 49, W216–W227 (2021).
110. https://mestrelab.com/.
111. Kratochvil, M. J., Balerud, N. K., Schindler, S. J. & Moxley, M. A. Evidence of a preferred kinetic pathway in the carnitine acetyltransferase reaction. *Arch. Biochem. Biophys.* 691, 108507 (2020).
112. Schnell, S. & Mendoza, C. Time-dependent closed form solutions for fully competitive enzyme reactions. *Bull. Math. Biol.* 62, 321–336 (2000).
113. Serventi F., Ramazzina I., Lamberto I., Puggioni V., Gatti R., Percudani R. Chemical Basis of Nitrogen Recovery through the Ureide Pathway: Formation and Hydrolysis of S-Ureidoglycine in Plants and Bacteria. ACS Chemical Biology. Vol.5, No 2. 203 – 214. (2010).
114. Britton K.L., Langridge S.J., Baker P.J., Weeradechapon K., Sedelnikova S.E., De Lucas J.R., Rice D.W., Turner G. The crystal structure and active site location of isocitrate lyase from the fungus Aspergillus nidulans. Structure. Vol 8 No 4. 349-362. (2000).

Supplementary Information

Danio rerio proteome filtering according to UGL enzyme features as in Y.Takada, T. Noguchi, Biochem. J., 1986, 235, 391–397.

The command-line bash procedure was included in an R markdown document and output in pdf with the knitr library. The candidate *Dr*MSL protein is highlighted in bold.

```
#Download D. rerio and H. sapiens proteomes from Uniprot (one protein sequence per gene)
Url='https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/'
wget -q -O - $Url"Eukaryota/UP000000437/UP000000437_7955.fasta.gz" | gunzip > Dr.fasta
wget -q -O - $Url"Eukaryota/UP000005640/UP000005640_9606.fasta.gz" | gunzip > Hs.fasta
grep -c '>' ?? .fasta
## Dr.fasta:20358
## Hs.fasta:20607
- - - - - - - - - - - - - - -
#Danio proteins with Mw = 64000 +/- 10%
pepstats Dr.fasta -auto -stdout | perl -ne '$name=$1 if (/STATS of (\S+)/);if (/ weight = (\S+)/) {printf "$name\n" if ($1>(64000-64000*0.1) and $1<(64000+64000*0.1) )}' | tee Dr_64kDa.acc | wc -l
## 2043
- - - - - - - - - - - - - - -
#Danio proteins with PTS1**
fuzzpro -pattern "[SAC][KRHSN][LM]>" Dr.fasta -auto -stdout|
perl -lane 'print $F[2] if (/Sequence:/)'|tee Dr_PTS1.acc | wc -l
## 175
- - - - - - - - - - - - - - -
#Mw and PTS1 intersection
grep -f Dr_PTS1.acc Dr_64kDa.acc | tee Dr_PTS1_64kDa.acc | wc -l
## 25
- - - - - - - - - - - - - - -
#Make Blast database
makeblastdb -in Hs.fasta -dbtype prot -parse_seqids -logfile logfile
```

makeblastdb -in Dr.fasta -dbtype prot -parse_seqids -logfile logfile

- - - - - - - - - - - - - - -

#UGL candidates (in *Danio* and not in *Homo*):

blastdbcmd -db Dr.fasta -entry_batch Dr_PTS1_64kDa.acc > Dr_PTS1_64kDa.fasta

blastp -db Hs.fasta -query Dr_PTS1_64kDa.fasta -evalue 1e-3 -num_threads 4|

tee Dr_PTS1_64kDa.blast | grep -B6 "No hits found" |grep "Query="

Query= A0A8M1P3F0 Malate synthase OS=Danio rerio OX=7955 GN=mlsl PE=4 SV=1

Query= I6V1W0 Protein brambleberry OS=Danio rerio OX=7955 GN=bmb PE=2 SV=1*

*Protein involved in karyogamy[1]

1. Abrams E.W., Zhang H., Marlow F.L., Kapp L., Lu S., Mullins M.C. Dynamic Assembly of Brambleberry Mediates Nuclear Envelope Fusion during Early Development. *Cell 150, 521-532, August 3,* (2021).

Chapter 3

Identification of human ASPDH gene as the missing 2-aminomuconate reductase in tryptophan degradation pathway

Introduction

NAD⁺ biosynthesis

Coenzymes deriving from pyridine nucleotides are involved in several enzymatic reactions in living organisms[1]. This class of compounds takes also part in ATP synthesis in mitochondria via the electron transport chain acting as electron carriers[2]; otherwise, they can also act as electron donors to regulate redox status in different reactions. These coenzymes are also used as substrates by NAD/FAD-dependent enzymes[3] in the glycolysis and in the Krebs cycle or act as redox regulators in ion channelling[4].

A benzene molecule with one methine group replaced by a nitrogen atom is the base of pyridine structures (Fig.1); a pyridine with a primary amide group attached in the meta position is part of the NMN (nicotinamide mononucleotide) molecule, which is bound to AMP (adenosine monophosphate). In living cells, these compounds are found either in the reduced or oxidized form and could be (un)phosphorylated[5].

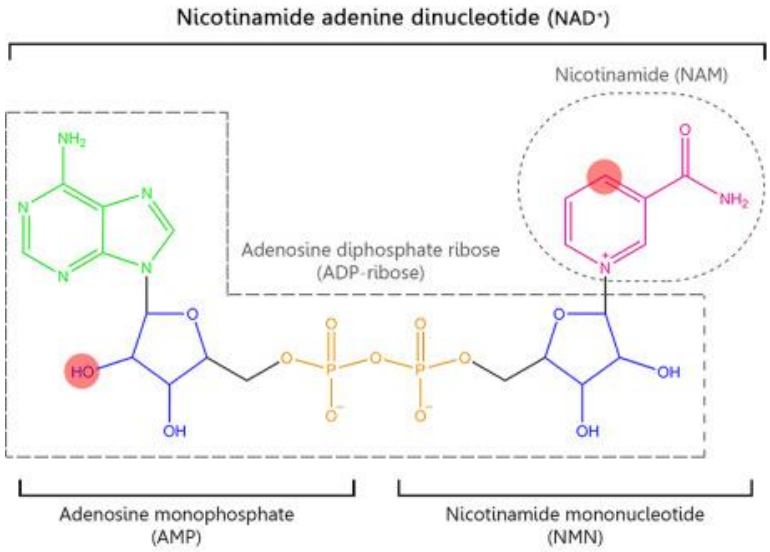


Figure 1: NAD⁺ structure and its derivative compounds (image adapted from Navas *et al.*, 2021) [6]**.** The structure of NAD⁺ with emphasis on AMP (adenosine monophosphate) and NMN (nicotinamide mononucleotide) is included in the grey dotted box, and NAM (nicotinamide) is colored in pink and included in the grey dotted circle. Adenosine is colored green, ribose in blue, and phosphates in orange.

In particular, NAD⁺ (nicotinamide adenine dinucleotide) has a relevant role in redox reaction and it is important for energy metabolism and for hydrogen transfer[7]. NAD⁺ ubiquitous distribution, together with its participation in numerous activities, has effects on many cellular functions

and can affect directly and indirectly many metabolic pathways; moreover, its availability may fluctuate depending on several factors, such as the cellular type and subcellular compartmentalization, the glucose level, and the caloric intake.

NAD^+ is indirectly involved in the response to environmental and nutritional alterations and in DNA damage repair through the activation of sirtuins and mitochondrial metabolism[8]; sirtuins directly regulate PARPs (poly ADP-ribose polymerases) transcriptions and they participate synergically in response to stress and inflammation, and in oocyte maturation and embryo development[9].

In addition to this, the regulation of CD38 and CD157 expression, two glycoproteins that contribute to tumorigenesis[10,11], has been shown to have a strong impact on NAD^+ availability and on control of the expression of proteins related to glycolysis and TCA or active in antioxidant pathways[12]. As consequence, NAD^+ metabolism has a close dependence on the circadian clock and is related to the cellular aging process and chromatin remodeling[13], metabolic disorders, deregulation of the immune system, cellular senescence, and neurodegeneration[14]. Otherwise, NAD^+ depletion leads to degenerative disease its accumulation is related to tumorigenesis[6].

NAD^+ biosynthesis metabolic pathways

NAD^+ levels are controlled by independent metabolic routes (Fig.2) that start from different metabolites and lead to NAD^+ synthesis for the support of all the cellular processes in which it is implicated. In fact, NAD^+ has a central role in cytoplasmatic, mitochondrial, and nuclear metabolism acting as an essential coenzyme, a precursor, and substrate and could be synthesized, consumed, and regenerated with redox reactions catalyzed by the intervention of enzymes and transporters.

Two NAD^+ molecules are reduced to NADH for each glucose molecule during glycolysis. These NADH molecules pass through the mitochondrial membrane and enter thanks to the malate/aspartate shuttle; in this organelle, NAD^+ molecules are then reduced to NADH to support the electron transfer chain. NADH is oxidized by Complex I and its electrons are flown across molecular complexes containing cytochromes, ubiquinones, and others to create a proton gradient used to synthesize ATP[2]. Considering all these reactions, it is clear how NAD^+ has an essential role to ensure the correct homeostasis in living cells.

The *de novo* synthesis, or Kynurenine pathway, starts by capturing dietary-acquired tryptophan through specific transporters. Serum tryptophan can be found in a free state or bound to albumin[15].

As a result of this metabolic route, tryptophan is converted to QUIN (quinolinic acid) through subsequent enzymatic reactions catalyzed by IDO (indoleamine 2,3-dioxygenase), TDO (tryptophan 2,3-dioxygenase), KMO (kynurenine 3-monooxygenase), KYNU (tryptophan 2,3-dioxygenase). The last described step of the pathway is catalyzed by 3HAO (3-hydroxyanthranilic acid oxygenase). It leads to ACMS (α-amino-β-carboxymuconate ε-semialdehyde) formation, which spontaneously cyclizes into QUIN and is then enzymatically converted into NAMN (nicotinamide mononucleotide) by QPRT (quinolinate phosphoribosyltransferase).

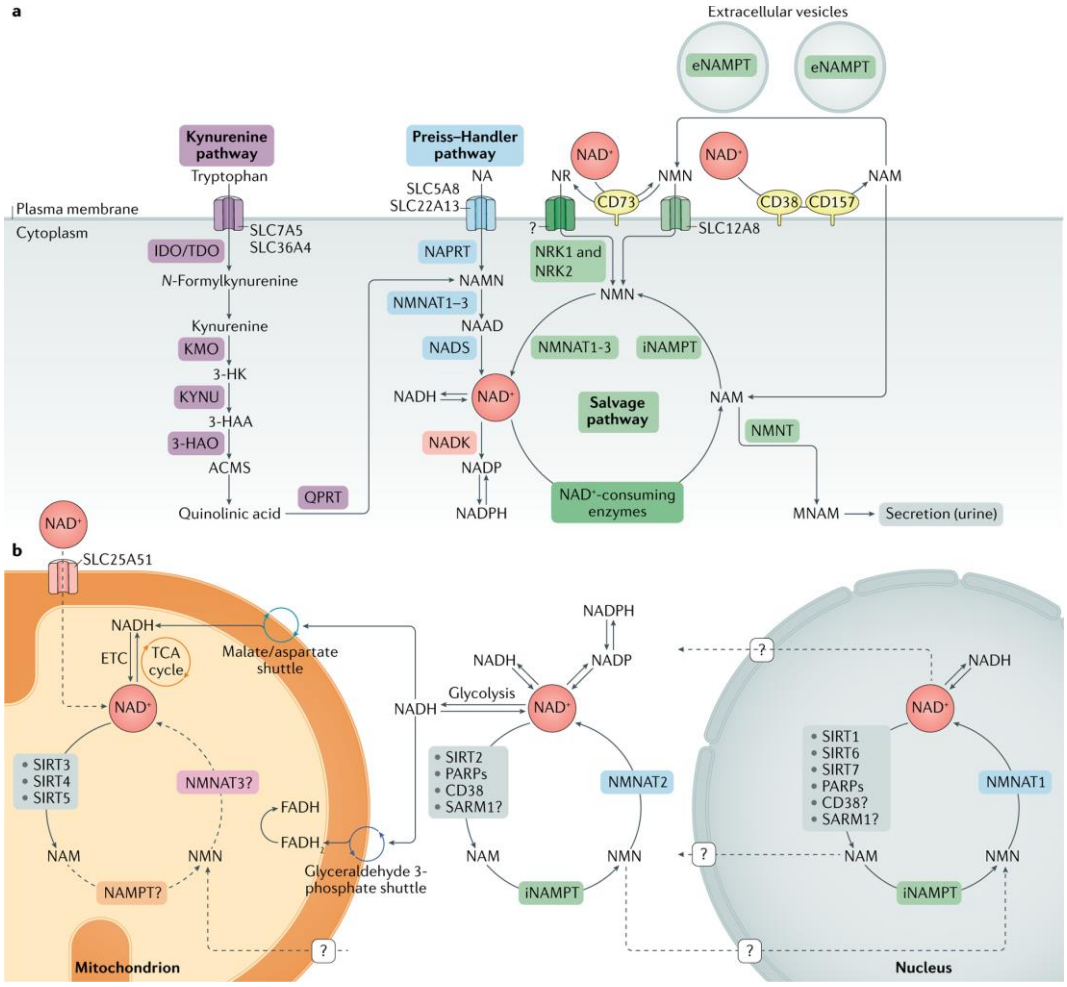


Figure 2: NAD metabolism[14]**.** Schematic reactions involving NAD^+. In the upper panel, reactions that are catalyzed in the cytoplasm are described. In the lower panel, NAD^+ metabolism is in different subcellular compartments (mitochondrion and nucleus).

NAMN is then converged into Preiss–Handler pathway, which starts with NA (nicotinic acid) acquisition from the blood and leads to NAD^+ biosynthesis passing by the two intermediates, NAMN and NAAD (nicotinic acid adenine dinucleotide), by the activity of NAPRT (nicotinic acid phosphoribosyltransferase) and NMNATs (nicotinamide mononucleotide adenylyltransferases). NAAD is finally transformed into NAD^+ by NADS (NAD^+ synthetase).

Sirtuins, PARPs proteins, and the CD38, CD157, and SARM1 are responsible for NAD^+ consumption and conversion into NAM (nicotinamide). The salvage pathway retrieves NAM and converts it into NMN (nicotinamide mononucleotide) by intracellular NAMPT (nicotinamide phosphoribosyltransferase) and then into NAD^+ via the different NMNATs. Prior to being converted into NMN, NAM can be secreted through urine after having been methylated by NNMT (nicotinamide N-methyltransferase).

NMN from the extracellular environment can be imported into the cell via specific transporters while is obtained from intracellular NR by NRKs (nicotinamide riboside kinases).

NMN in cytoplasm originates from NAN NAMPT-dependent conversion and NMNAT2 converts it into NAD^+. NAD^+ is reduced to NADH during glycolysis; the malate/aspartate shuttle and G3P (glyceraldehyde 3-phosphate) shuttle move NADH to the mitochondrial matrix. In particular, $FADH_2$ (flavin adenine dinucleotide) resulting from the G3P shuttle is oxidized by Complex II while NADH by Complex I. SIRT3–SIRT5 in mitochondrion converts NAD^+ in NAM, then in NMN or NAD^+ which is indeed secreted in the cytoplasm by the action of unknown enzymes and transporters. Cytosolic and nuclear NAD^+ equilibrium is maintained by its diffusion through specific nuclear pores and is involved in a nuclear-specific salvage pathway; mitochondrial content is approximately 250 μM[16], nuclear 70 μM[17].

Since NAD^+ precursors (NR, NMN, NAM, and Trp) give a limited contribution to its biosynthesis through Preiss-Handler metabolism and do not sustain high rates of its production, NAD^+ is also produced from NAM in the salvage pathway[12].

NAD^+ biosynthesis in healthy and cancer cells

Tumoral cell metabolism depends on NAD^+ biosynthesis and its depletion could result in the inhibition of cellular growth. Otherwise, epigenetic silencing or NAMPT expression enhancement is associated with the alteration of tumoral cell metabolism[18]. This evidence has provided an interesting therapeutic target to treat malignant cells, such as gliomas, and has led to starting the development of inhibitors direct to NAMPT enzymes. Since NAMPT is directly responsible for modulating immune systems[19], NAMPT has also been chosen to be a

target for altering the immunological state of tumors and its preliminary studies have demonstrated the inhibitor's ability to alter tumor immunologic environment by upregulating surface antigens and limiting the import of NAD^+ precursors[20].

NA acquired and converted into NAD^+ through Preiss–Handler pathway has been used to treat dyslipidemia because of its ability to lower cholesterol[21] by reducing LDL and increasing HDL and enhancing sirtuin activities; this mechanism is probably mediated by cellular receptors. NAM has failed to have the same effects probably because it inhibits sirtuins[22].

The *de novo* biosynthesis pathway and its regulation

The *de novo* biosynthesis in humans and other mammals is one of the main routes of tryptophan degradation and leads to the release of several metabolites, including the NAD^+ precursor QUIN. QUIN is produced via the non-enzymatically cyclization of ACMS[23] which undergoes a variety of utilization, including NAD^+ production.

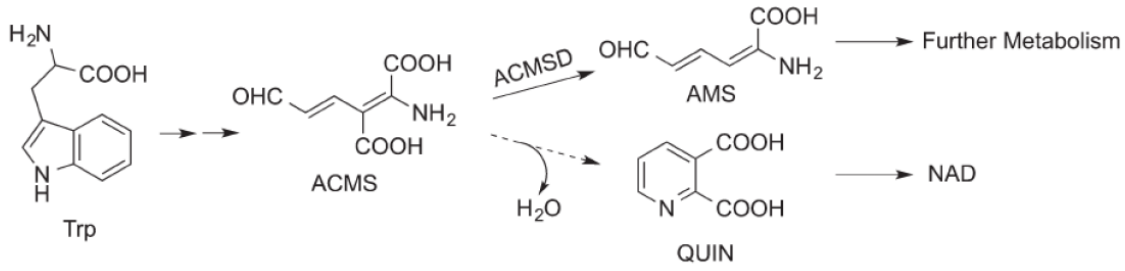


Figure 3. Alternative routes for ACMS metabolism[24]. ACMS could be spontaneously converted into quinolinate (QUIN) with the release of a water molecule. In the alternative, ACMS could be enzymatically turned into AMS by ACMSD activity.

It has been demonstrated that ACMS can alternatively be enzymatically converted into AMS (amino-muconate semialdehyde) by ACMSD and then AMS follows an alternative route that ends in acetyl-Coenzyme A production (Fig.3). ACMSD, which has been mainly expressed in liver, kidney and partially in brain[25] similarly to other genes of the pathway, can be modulated in order to control mitochondrial energetic metabolism and NAD^+ levels since it avoids ACMS enclosure and control QUIN levels[24].

Biochemical and structural analysis of human ACMSD has shown that the conservation of its structural rearrangement allows it to maintain a potentially sophisticated regulatory mechanism[24] in a similar way to the homologous bacterial enzyme[26,27].

ACMS accumulation in cells is regulated by the activity of ACMSD and high levels of activity prevent Trp conversion to NAD^+; on the contrary, the inhibition of ACMSD catalysis leads to QUIN formation. *ACMSD* mutation and protein deficiency have been demonstrated to be associated with cortical myoclonus, epilepsy, and Parkinson's disease due to its accumulation and its inability to cross the blood-brain barrier[28] because of its polar nature.

Recently, the gene involved in the conversion of ACMSD product has been identified; it has indeed been demonstrated that ALDH8A1, a protein previously described as a retinal dehydrogenase able to preferentially oxidize 9-cis-retinal[29], catalyzes the oxidation of 2-AMS (2-aminomuconic semialdehyde) to 2-AM (2-aminomuconate) through a NAD-dependent mechanism[30]. 2-AMS is an unstable compound that tends to spontaneously cyclize to picolinate. Picolinate analogs have been shown to alter serotonin, dopamine, and norepinephrine metabolism in the brain[31], suggesting a poisoning effect when present at higher levels within cells.

The product of 2-AM, an alfa-amino acid with conjugated double bonds, is known to be an intermediate of the biodegradation of nitrobenzene in bacteria and is hydrolyzed to 4-oxalocrotonate with the release of ammonia by 2-aminomuconate deaminase[32].

In mammals, a reaction involving the consumption of 2-aminomuconate has been described[33]: in particular, a 2-aminomuconic acid reductase is proposed to catalyze the stoichiometric conversion of 2-aminomuconate to alpha-ketoadipate (or 2-oxoadipate) and NH_3 in the presence of NAD(P)H (Fig.4, orange box). This reaction was detected following the decrease in absorbance in the surrounding of 325 nm, due to the concomitant 2-aminomuconate reduction and NAD(P)H oxidation after the addition of the enzyme purified from a cat liver tissue.

It has been proposed that the product of 2-AM reduction and deamination is used to produce glutaryl-coenzyme A for glycolysis[34] and for acetyl-Coenzyme A synthesis but the gene responsible for this enzymatic activity has not been identified yet.

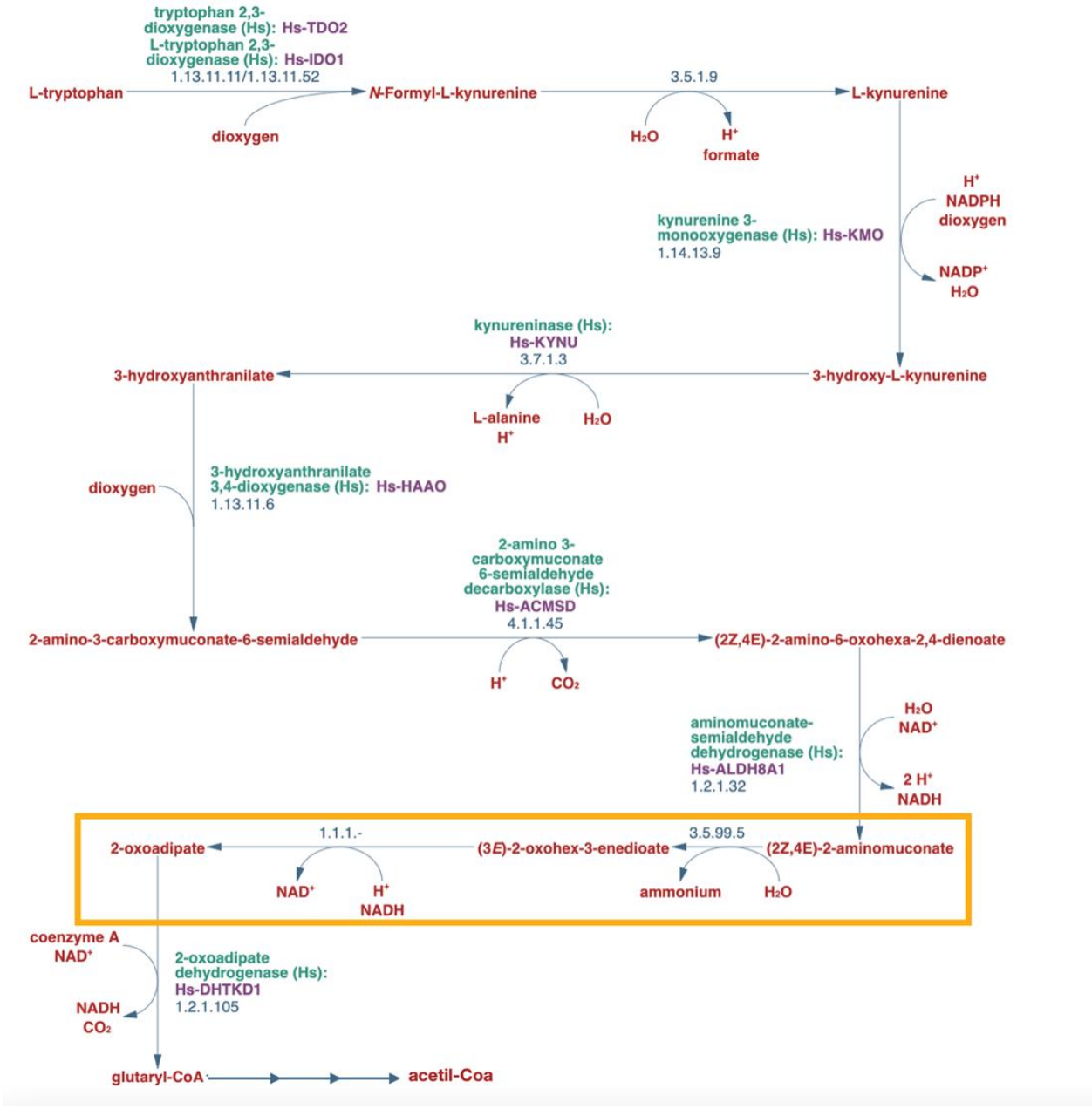


Figure 4: Metabolic pathway of L-Tryptophan degradation in Eukaryotes represented in MetaCyc database. Experimentally validated reactions are reported for degradation of L-tryptophan to acetyl-CoA in *Homo sapiens*, with E.C. number and human genes related to the enzyme responsible for the catalysis. Two-step enzymatic conversion of 2-aminomuconate in 2-oxoadipate is enclosed in the orange box.

The kynurenine pathway in the brain

The kynurenine pathway is carried out in the kidney and liver in humans and other mammals (Fig.5), where amino acids catabolism occurs. Otherwise, brain function may be affected by kynurenine intermediates and Trp availability has an important role to support the kynurenine pathway.

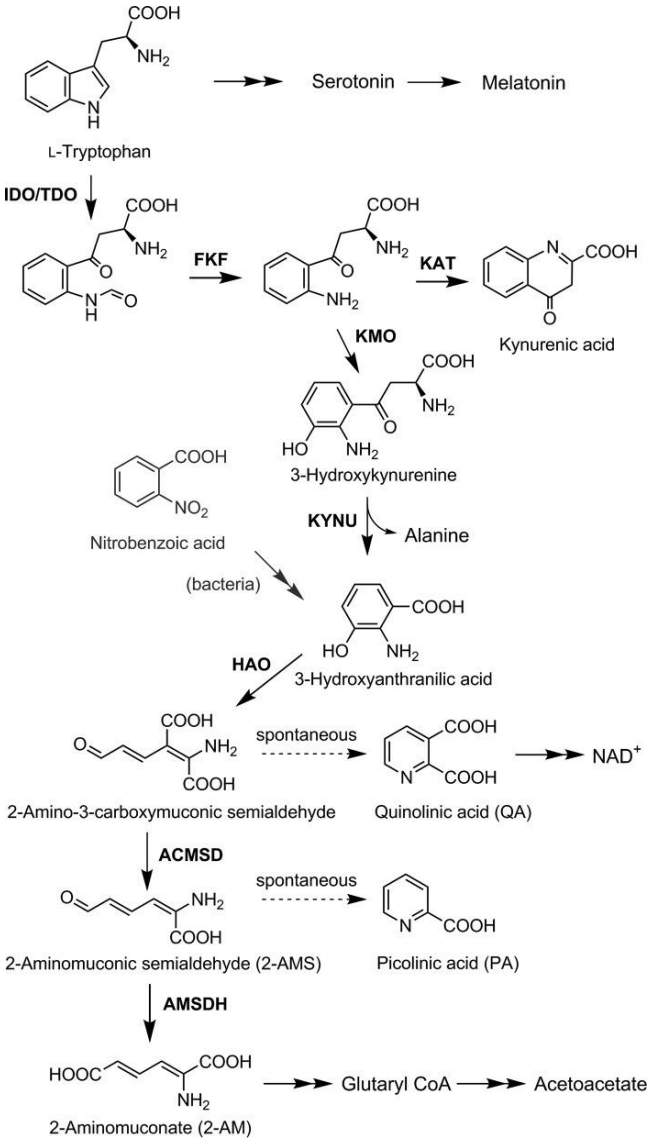


Figure 5. Metabolic route of kynurenine pathway[30]**.** Enzymatic (solid arrows) and spontaneous (dashed arrows) reactions are involved in tryptophan metabolism in mammals.

KYNA (kynurenic acid) has been demonstrated to act as an antagonist of glutamate receptors and has a greater affinity in binding NMDA and other receptors[35]; in addition, KYNA has been discovered to have neuroprotective features[36].

3-HK (3-hydroxykynurenine), 3-HAA (3-hydroxyanthranilate), and anthranilic acid are probably involved in pro- and anti-oxidative processes in the brain but do not have side effects on brain activity[37].

QUIN enters neuronal cells showing a preference for NMDA receptors and stimulates lipid peroxidation in an iron-dependent manner[38]. Otherwise, it has been shown that it might play an active role in mechanisms that undergo neurodegenerative diseases, such as Huntington's disease[39]. An abnormal increase in the ratio between QUIN and KYNA could be at the base of neuronal lesions[40] and may influence depressive disorders[39].

Potential cancer therapies[41,42] or neuroprotective agents[43,44] could be designed or ameliorated to inhibit enzymes that determine negative effects by releasing intermediates in neuronal cells; to complete the knowledge of the role of the kynurenine pathway in the brain, further elucidation of their mechanisms is needed.

Kynurenine pathway and serotonin biosynthesis

Serotonin is directly produced from tryptophan metabolism in the brain and this metabolite is conserved across most living phyla due to its important role in neuroimmune communication[45]; serotonin precursor 5-HTP (5-hydroxytryptophan) is enzymatically produced by TPH (tryptophan hydroxylase) activity and then decarboxylated by AADC (aromatic amino acid decarboxylase). In fact, the synthesis of serotonin is an alternative cerebral route with respect to NAD^+ biosynthesis from tryptophan (Fig.6). Expression and activity of these enzymes, in particular TPH, are regulated by external signals.

Serotonin released from the brain and released into the bloodstream is involved in several functions, including neuronal control of motility[46]. It is otherwise responsible for controlling digestive aspects including insulin release[47]. Serotonin is transported via specific transporters over enterocytes[48] and is then degraded by MAO (monoamine oxidase).

IDO and TDO enzymes, previously described as components of the kynurenine pathway, play central roles in the balance of the two alternative metabolisms. Activation of IDO during inflammation reduces the rate of serotonin synthesis[49] and, in synergy with ACMSD activation, leads to produce picolinate, which has a protective activity against neurotoxicity[50].

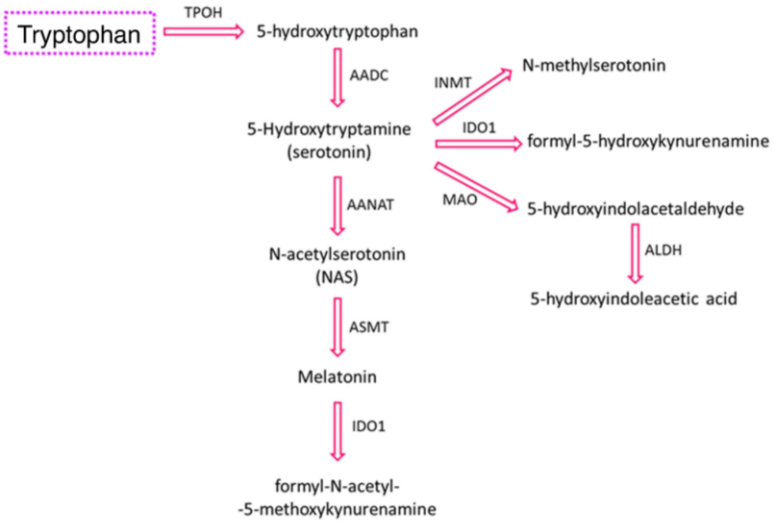


Figure 6: Tryptophan and serotonin catabolism (image adapted from Mondanelli *et al.*, 2021)[45]. Alternative routes for tryptophan usage in cells.

A general dysregulation of serotonin and other neurotransmitters has been detected in patients that suffer from depression or schizoprenia[51,52], suggesting a connection between tryptophan metabolism and other common neuropsychiatric diseases that have an impact on mental health. With regard to depression, cytokine induction activates IDO protein in the brain resulting in tryptophan consumption through kynurenine pathway and shifting the balance from serotonin production[53].

A low-tryptophan diet results in the reduction of serotonin production and circulation, leading to accentuate this disorders[54]; in addition, atherosclerosis, diabetes, and obesity are influenced by several factors, including metabolic dysregulations that derive partially from alterations in Kynurenine and serotonin pathways[55].

This pathway is also related to the persistence of gut microbiota, whose physiology depends on the intermediate of serotonin metabolism[56].

Discoveries of unknown components that regulate this complex metabolic pathway, mainly tryptophan, serotonin, and NAD^+, could help understand pathologies and disorders not yet clarified.

NAD *de novo* biosynthesis in plants, bacteria, and fungi

In animals and in some bacteria species, NAD^+ is mainly obtained from tryptophan by quinolinate production through the Kynurenine pathway as described previously and is then processed through different metabolic routes placed in different organelles.

In plants, only two pathways are involved in the biosynthesis of NAD^+, the *de novo* pathway and the salvage pathway which are partially overlapped.

The *de novo* pathway starts from L-asp (L-aspartate) which is condensed with DHAP (dihydroxyacetone phosphate) into QUIN by LASPO (L-aspartate oxidase) and QS (quinolinate synthase)[57]. The FAD-dependent LASPO is able to convert L-asp to IA (iminoaspartate) preventing the degradation of the unstable product[58]; QS, which is a plastidic enzyme, uses IA as the substrate to form QUIN. To rapidly capture IA, QS and LASPO probably form a physical complex[59]. QUIN is transformed into NMN with the addition of a PRPP (phosphoribosyl phosphonate) moiety by the enzymatic activity of QPT (quinolinate phosphoribosyltransferase).

Enzymatic reactions which are shared by both pathways include NMN conversion into NAD^+ by NMNAT (mononucleotide nicotinate nicotinamide mononucleotide adenyltransferase) and the reaction catalyzed by NADS (NAD synthase). NADK After that, (NAD kinase) is then involved in the ATP-dependent conversion of NAD(H) to NADP(H)[60]. NAD^+ is also catabolized by several enzymes (such as ADPT cyclases, histone deacetylases sirtuins, or PARPs) to support cellular signaling pathways and energy release.

Similarly to mammals, NAD(P)H maintains the cellular redox status and is used inmajor metabolic pathways by plants[61] (Fig.7) In addition, NAD^+ controls the assimilation of carbon and nitrogen and is used in carbon metabolism processes, such as photosynthesis, or for lipid synthesis, it has a central role in the regulation of seed germination, root development, in response to environmental stress, and in immunity[62].

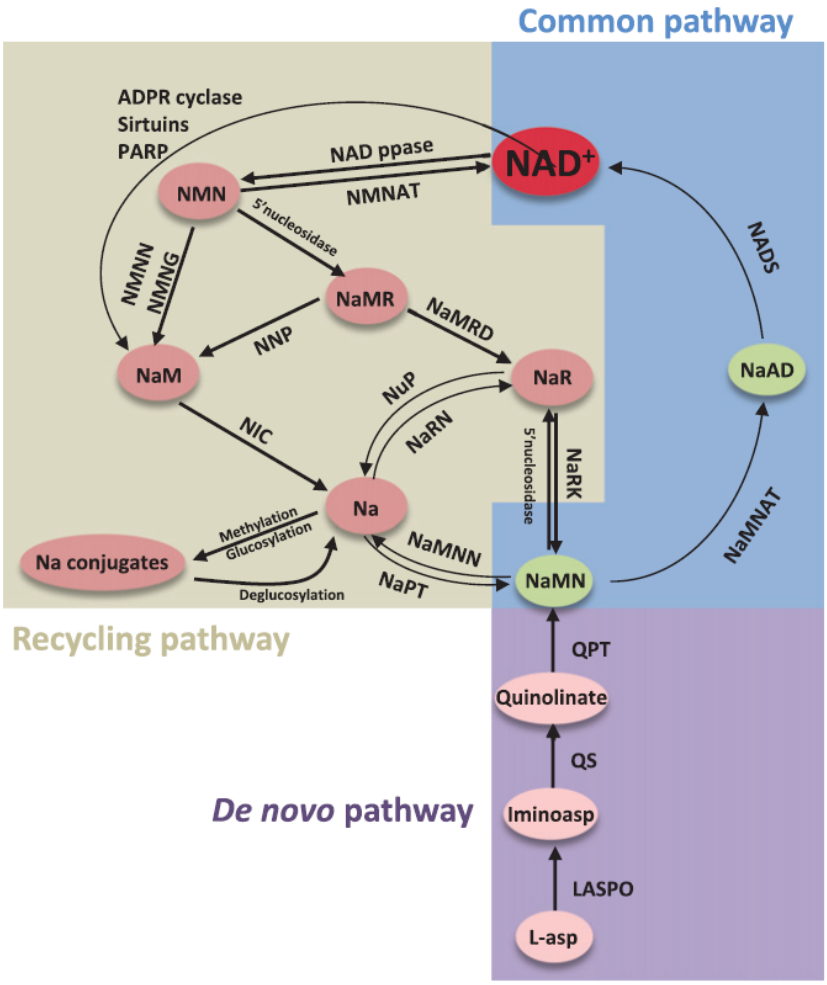


Figure 7: Alternative NAD⁺ biosynthetic pathways in plants[62]**.** Common pathway is highlighted in light blue, recycling pathway in brownish yellow, *de novo* pathway in violet.

In bacteria, the *de novo* synthesis of NAD⁺ involves three steps starting from L-asp and followed by two additional reactions that convert NMN to NAD, similar to what occurs in plants (Fig.8A). The genes involved in these reactions (*nadA, nadB, nadC*) are broadly conserved in bacterial species and are clustered in a specific operon[63], together with other molecular components that play a role in NAD biosynthesis[64]. Conversely , *the nadB* gene lacks in some organisms, as two different types of oxidoreductase have evolved to catalyze the first reaction of the pathway; in fact, in some thermophilic bacteria and in archaebacteria, the aspartate dehydrogenase (L-Aspdh) enzyme catalyzes the NAD(P)-dependent dehydrogenation of L-aspartate to form iminoaspartate, substituting the nadB gene in the *de novo* pathway[65] .

In yeast, NMN is synthesized using tryptophan as a starting compound (Fig.8B) similar to animals. Genes and ORFs responsible for these activities have been identified since they are

homologs to mammalian genes and have been named BNAs (biosynthesis of nicotinic acid); all these genes, except for BNA3[66], are subjected to up-regulation by transcription factor Sum1p[67]. Moreover, the sirtuin Hst1 tightly controls NAD^+ abundance by activating the de novo pathway and BNA genes in case of NAD^+ depletion[68]. Additionally, BNA1, BNA2, and BNA4 require molecular oxygen to allow tryptophan degradation to quinolinate[69].

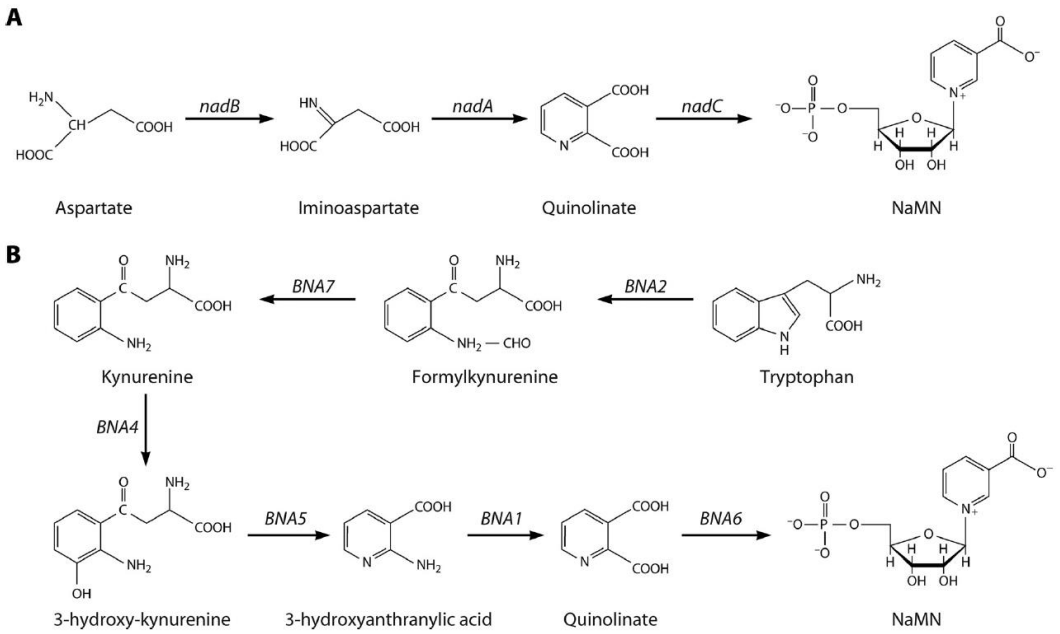


Figure 8. Reaction forming NMN for NAD de novo biosynthesis[70]. (A) NMN synthesis starts from L-Asp in bacteria. (B) NMN synthesis starts from Trp in yeasts.

L-aspartate dehydrogenase enzyme features

Reversible oxidative deamination of L-aspartate to iminoaspartate using either NAD^+ or $NADP^+$, followed by its spontaneous decay into OAA (oxaloacetate), was first observed in a hyperthermophilic bacterium, *Thermotoga maritima*[58]. *T. maritima* L-aspDH (TM1643) showed dehydrogenase activity toward L-aspartate, with a K_M for L-aspartate 20 times lower in the presence of NAD than in the presence of NADP. During this reaction, oxygen is not used as an electron acceptor, whereas LAO (L-aspartate oxygenase, encoded by *nadB* gene) oxidizes L-Asp to iminoaspartate using oxygen or fumarate[71].

It has been also demonstrated TM1643 capability to catalyze the reductive amination of oxaloacetate; for this reaction, NADPH and NADH were equally efficient as electron donors[58].

It has been proposed that L-AspDH enzymatic activities probably have replaced the LAO activity in anaerobic organisms[65] during evolution and are also useful to provide energy and

nitrogen in some bacterial species[72]. In fact, the *nadB* gene is absent in the genome of thermophilic bacteria and in archaea.

TM1643 did not share homology or structural similarity with other proteins involved in amino acid dehydrogenation[58]. After that, a second example of aspartate dehydrogenase protein has been identified in the archaea *Archaeoglobus fulgidus*[65].

A. fulgidus and *T. maritima* L-aspDH are NadB-type enzymes and are included in operons with other genes involved in the *de novo* NAD^+ biosynthesis[73].

Both the two protein structures have been resolved and confirmed that L-AspDH proteins in their native form are homodimeric[58,65] differently from other heterodimeric amino acid dehydrogenases. As expected for archaea enzymes*, the T. maritima* protein is active at higher temperatures.

In organisms such as *P. aeruginosa PAO1*, both L-AspDH and LAO are present, rather suggesting the involvement of L-aspDH in the TCA cycle (formation of OAA) than in NAD^+ biosynthesis. It exhibits similar K_M values for either NAD^+ or $NADP^+$ and substrate specificity for L-Asp and OAA[72]; *Pae*AspDH K_M values calculated were 2.212 mM for OOA , 0.045mM for NADH and 10.1 mM for ammonia. This evidence confirms the reversibility of L-aspartate dehydrogenase activity.

There is no explanation of the presence of a gene homologous to bacterial L-aspDH in most vertebrates and in some invertebrate clades. Since NAD^+ in these organisms is produced starting from tryptophan, there is no need for L-aspartate dehydrogenase activity. On the other hand, no L-aspartate dehydrogenase activity has been reported in eukaryotes[71].

A recent paper has reported that the *L-aspDH* gene in mammals, expressed mainly in the kidney and liver, encodes for a soluble protein able to bind NAADP or NADP performing ITC experiments[74] with higher affinity for NADP. Despite this progress in the study of L-aspDH protein in vertebrates, no hypothesis of its activity has been proposed.
Here, we propose a role in tryptophan degradation for eukaryotic L-aspDH and we demonstrate the NADH-dependent enzymatic activity for this unknown human gene.

Results and Discussion

Human-centered orthogroup identification with DIAMOND software

Orthogroups deriving from OrthoDB have not helped us to identify missing genes belonging to pathways such as the tryptophan catabolism. As an alternative to the construction of phylogenetic profiles, we considered orthogroups that contain at least one or more gene copies in *Homo sapiens*.

The pipeline for phylogenetic profiling and for pairwise comparisons corresponds to the one described in Chapter 1, except for the orthogroup identification.

We performed the analysis of concordant transition by constructing profiles from orthologous gene sets identified through homology searches, considering the same 1264 eukaryotic genomes contained in OrthoDB. Specifically, we carried out a homology search with the fast DIAMOND algorithm[75] using human genes as the queries and then we identified genes from other organisms that were human best hits; protein-coding genes having the most significant E values and being best hit (BH) to human query were included in the orthology group together with the human gene.

This approach, if compared with the previous one, allows to exclude genes that are absent in *Homo sapiens* from the analysis and resolves gene misidentifications or exclusions that occurred using BLAST instead of DIAMOND.

Phylogenetic profiling

Using the novel human-centered orthogroups, we built a matrix of binary profiles considering 20479 orthogroups (matrix columns) and 1258 organisms (matrix rows); the presence of one or more copies of the gene in each genome was considered as “1”, while set as “0” in case of gene absence. We then evaluated the *cotr_score* from the enumeration of concordant and discordant transitions for each pairwise comparison, and we also estimated the *p-value* and the *p-value adjusted* (see Material and Methods in Chapter 2).

We therefore obtained more than 1.3 M significant pairwise comparisons considering *p-values* $< 10^{-3}$ and ordering eukaryotic species according to the NCBI tree.

Identification of a candidate for the missing human 2-aminomuconate reductase

As proof of principle, our co-occurrence analysis allowed us to identify a gene candidate for the missing human 2-aminomuconate reductase, an orphan enzyme involved in the last steps of vertebrate tryptophan catabolism.

In particular, we observed that the orthogroup annotated as "L-aspartate dehydrogenase-like" (ASPDH_HUMAN) had significant associations with orthogroups containing members of the tryptophan degradation pathway (Tab.1), such as T23O_HUMAN, 3HAO_HUMAN, ACMSD_HUMAN and ALDH8A1_HUMAN (Fig.9A, 9C). The *p-values* (Fig.9A) point out significative associations between tryptophan degradation orthogroups and ASPDH_HUMAN.

OG NAME	HUMAN GENE DESCRIPTION
ASPDH_HUMAN	Aspartate dehydrogenase domain-containing protein
3HAO_HUMAN	3-hydroxyanthranilate 3,4-dioxygenase
T23O_HUMAN	Tryptophan 2,3-dioxygenase
KMO_HUMAN	Kynurenine 3-monooxygenase
KYNU_HUMAN	Kynureninase
AL8A1_HUMAN	2-aminomuconic semialdehyde dehydrogenase
ACMSD_HUMAN	2-amino-3-carboxymuconate-6-semialdehyde decarboxylase
KFO_HUMAN	Kynurenine formamidase

Table 1: Orthogroup names and Uniprot description for *ASPDH* and the human genes of tryptophan degradation.

Since *ASPDH-like* genes are mainly distributed in vertebrates and in some invertebrate species but lack in several organisms with the tryptophan degradation pathway, the enumeration of the concordant transitions and the resulting pairwise comparisons are affected by these confounding signals.

Considering known components of tryptophan catabolism, the *p-values* and *cotr_score* are similar to those calculated for ASPDH_HUMAN (Fig.9B) thus supporting the hypothesis of its involvement in the same pathway. As emerged from our analysis (Fig.9A, Fig.9B), ASPDH_HUMAN is in the second position of the most significant profile ranking on the basis of the *cotr_score* and the *p-value*, only after HAAO_HUMAN.

For each eight orthogroups considered, a representative set of 30 organisms were used to build a graphical matrix (Fig.9D) of presence and absence of genes in genomes and the retrieved sequences were examined for phylogenetic distribution.

From this analysis emerged that *ASPDH* are present in Discoba (*N. fowleri*) and in a wide range of animals, but mainly in vertebrates; in addition, they are present also in invertebrate organism

groups, such as in Cnidaria (*H. vulgaris*; *N. vectensis*), in Nematoda (*C. elegans*), and in Arthropoda (*D. magna* - Crustacea; *V. jacobsoni* - Acari).

A

OG1	OG2	cotr_score	p-value
T23O_HUMAN	ASPDH_HUMAN	0.1765	3.61E-09
3HAO_HUMAN	ASPDH_HUMAN	0.1364	6.03E-08
ACMSD_HUMAN	ASPDH_HUMAN	0.1263	2.19E-06
AL8A1_HUMAN	ASPDH_HUMAN	0.0930	1.19E-06

B

OG1	OG2	cotr_score	p-value
3HAO_HUMAN	KYNU_HUMAN	0.0000	9.62E-15
3HAO_HUMAN	ACMSD_HUMAN	0.1465	9.36e-09
T23O_HUMAN	3HAO_HUMAN	0.1316	3.32E-08
3HAO_HUMAN	AL8A1_HUMAN	0.1205	1.90E-07
T23O_HUMAN	ACMSD_HUMAN	0.1071	1.34e-06
KMO_HUMAN	T23O_HUMAN	0.0860	3.51E-05
KYNU_HUMAN	ACMSD_HUMAN	0.1052	2.71E-05
AL8A1_HUMAN	T23O_HUMAN	0.0545	7.57E-04

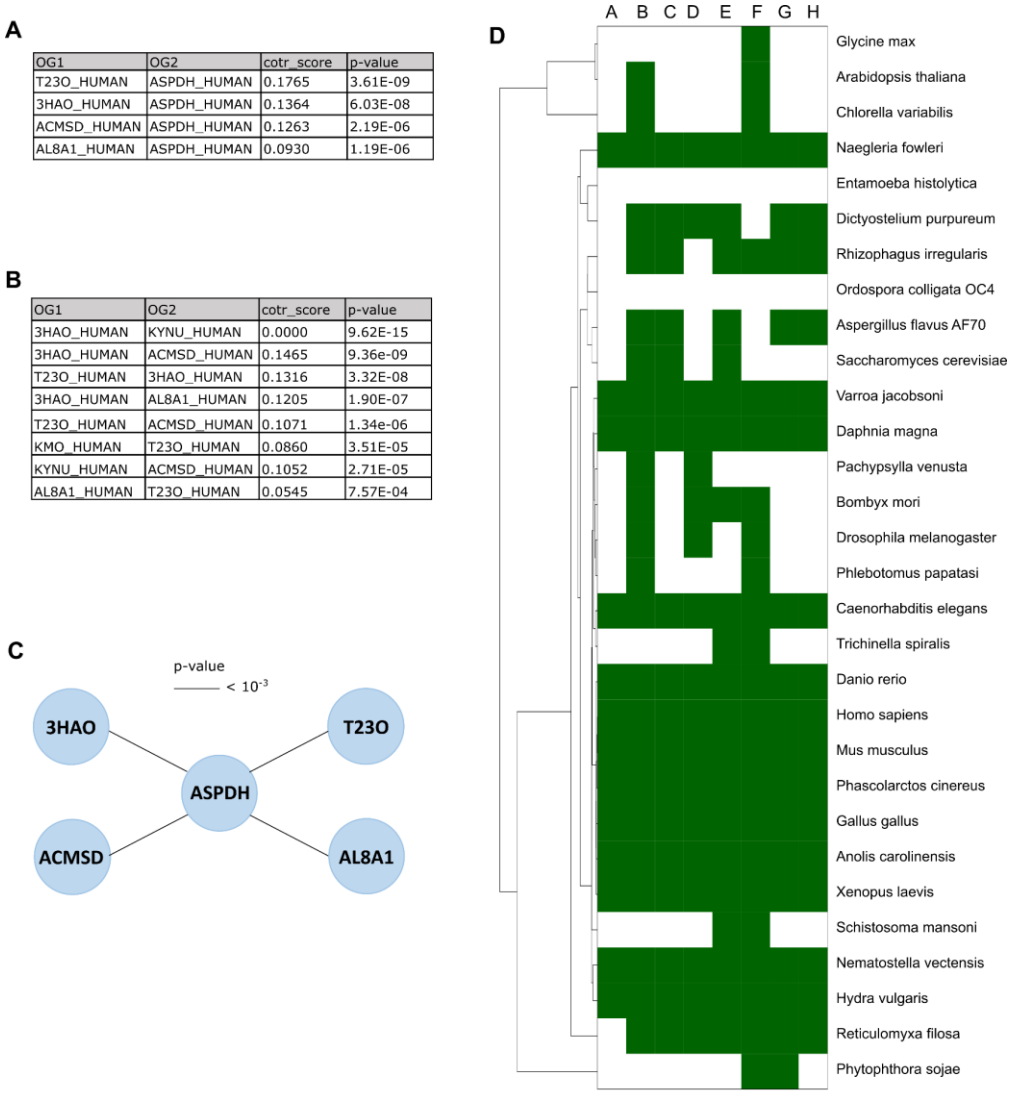


Figure 9: Identification of a relation between ASPDH_HUMAN and tryptophan catabolism genes through concordant transition analysis. (A) *Cotr_score* and *p-value* for pairwise comparison between tryptophan degradation orthogroups and ASPDH_HUMAN. (B) *Cotr_score* and *p-value* for pairwise comparison among tryptophan degradation orthogroups. (C) Schematic network showing significant cotr associations of orthogroups with ASPDH_HUMAN; line thickness represents *p-values* < 10^{-3}. (D) Phylogenetic profile similarity of ASPDH_HUMAN and coevolving genes from tryptophan catabolism. Each row represents an eukaryotic organism, each column represents an orthogroups identified using Diamond: A - ASPDH_HUMAN: B - KMO_HUMAN; C - 3HAO_HUMAN; D - T23O_HUMAN; E - KYNU_HUMAN; F - KFA_HUMAN; G - ACMSD_HUMAN; H - AL8A1_HUMAN.

Organisms having a copy of the *ASPDH* gene present all the other genes included in the analysis; in addition, there are no organisms that present *ASPDH* gene if the other genes are missing.

This strong co-occurrence supports the hypothesis of a close correlation between these genes and *ASPDH*, whose molecular function possibly depends on the presence of the other molecular components.

ASPDH genes are absent in plants (Fig.9D) along with all tryptophan catabolism genes except for KMO and KFA which participate in steroid biosynthesis; this evidence supports their strict functional and evolutionary relationship.

A similar pattern of absence and presence has been found in Fungi and in Insecta (Fig.9D); Insecta probably do not utilize tryptophan as NAD precursor, while Fungi own the BNA operon (see Introduction) containing specific genes for tryptophan metabolism that do not all share homology with vertebrates' genes.

DIAMOND orthogroups and comparison with OrthoDB database

The introduction of human-centered orthogroup has allowed the identification of the described hidden association not yet detected with other phylogenetic profiling methods.

Considering phylogenetic profiles built starting from Orthodb v.10, "L-aspartate dehydrogenase containing domain" profile has been erroneously predicted: in fact, human *ASPDH* homologous genes are contained in two distinct orthogroups called "L-aspartate dehydrogenase" (1182611at2759) and "putative L-aspartate dehydrogenase" (901738at2759) respectively. Genes from vertebrates, particularly from mammals, were randomly split in the two orthogroups and this determined an artificial alternation of present and absent signals along the phylogeny. This led to the erroneous enumeration of the number of concordant transitions in the pairwise comparisons between phylogenetic profiles.

The absence of a signal for *ASPDH-like* gene in some avian species made it difficult to correctly compare "L-aspartate dehydrogenase" orthogroups with the ones involved in tryptophan catabolism with our first computational approach. These genes have not been lost during evolution but they haven't been found in their genomes for technical reasons, for their high GC content and/or for their altered codon usage[76].

Detection of mRNA in avians that apparently lack the protein signal is the proof that *ASPDH* genes are probably hidden by a GC-rich context. In fact, we detected RNA sequences related to missing ASPDH proteins in birds (Tab.2) included in our analysis by performing a homology

search in tBlastN using *Gallus gallus* “putative L-aspartate dehydrogenase” (XP_040512953.1) as query.

AVES SPECIES	ACCESSION NUMBER	IDENTITY	E VALUE
Falco spp.	GGEE01081480.1	78%	7.00E-138
	GFNU01009937.1	72%	8.00E-122
Sturnus vulgaris	GFDQ01000154.1	72%	1.00E-121
	HAMX01123073.1	88%	1.00E-59
Tyto alba	HAMX01154392.1, partial	75%	2.00E-37
Zonotrichia albichollis	GFDQ01018815.1	60%	4.00E-49

Table 2: Evidence of *L-ASPDH-like* genes in aves. Table containing: (I) aves organisms lacking any associated protein in “L-aspartate dehydrogenase” orthogroups in Orthodb v.10; (II) computationally assembled mRNA sequences from primary data such as EST and raw sequence reads (unpublished direct GenBank submission); (III and IV) identity % and E-value respect to *G.gallus* ASPDH-like protein.

In the novel Orthodb v.11[77] release a higher number (1952) of completely sequenced genomes are included.

In the case of *ASPDH* genes, Orthodb v.11 solves the wrong construction of “aspartate dehydrogenase” orthogroups and identifies a single orthogroup containing all the orthologous genes. It also includes 28 bird proteins, e.g. *Tyto alba* protein, with respect to the 6 bird proteins found in Orthodb v.10; conversely, *Gallus gallus* and *Lonchura striata domestica* proteins, present in the previous version, are not included.

ASPDH gene expression and coexpression with tryptophan catabolism genes

The human *ASPDH* gene (GeneID: 554235), which is present in single copy is located on the chromosome 19 (19q13.33) and is formed by eight exons (Fig.10A). Alternative splicing contributes to produce two different isoforms for this gene; the main isoform (NM_001114598.2, NP_001108070.1), composed by seven exons, is 283 amino acid long while a shorter isoform (NM_001024656.3, NP_001019827.2) includes five exons with an alternative 5' transcriptional start site and is 179 amino acids long.

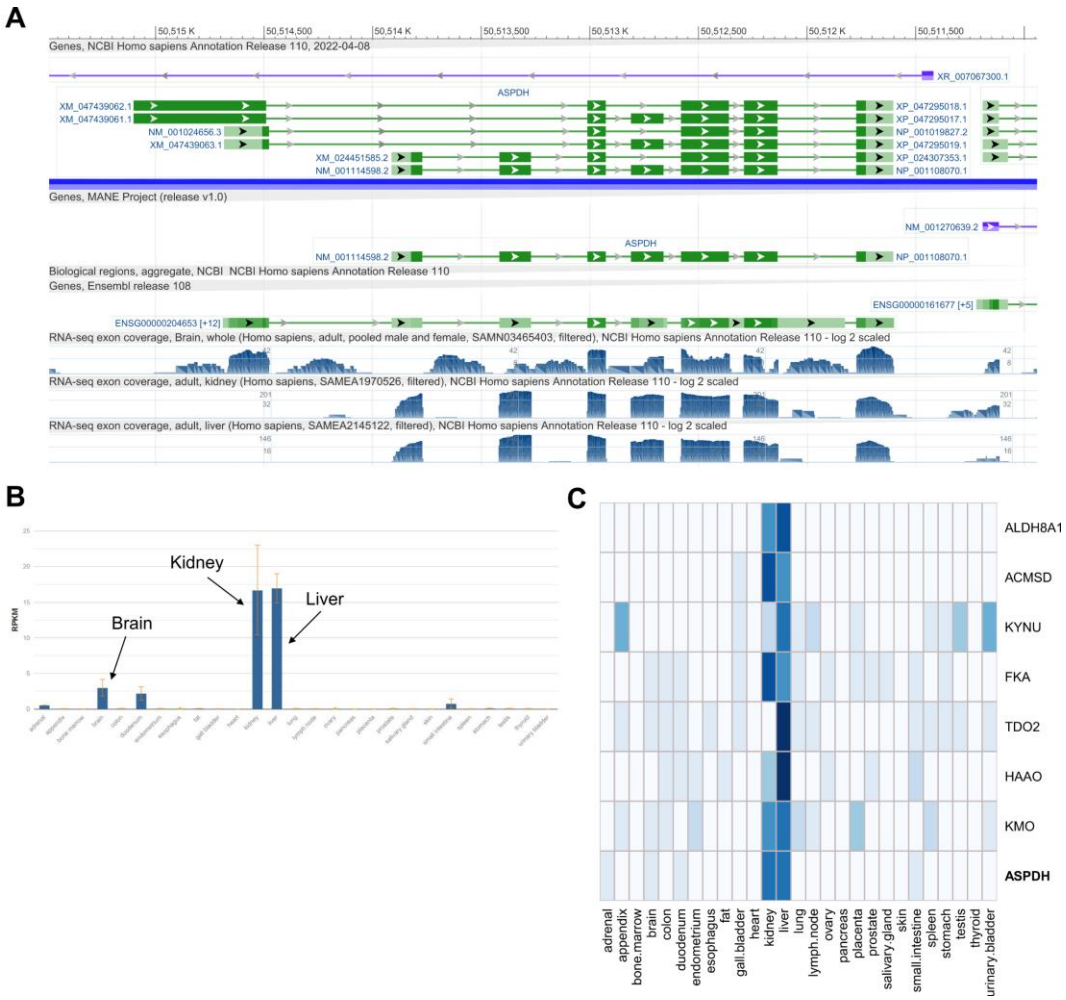


Figure 10: Human *L-aspartate dehydrogenase domain containing* gene. (A) NCBI sequence-viewer representation of the genomic region containing the *ASPDH* gene (GeneID: 554235) on *Homo sapiens* chromosome 19 (ann. release 110). Gene exon structure is represented by green segments while gene introns by green line. Blue bars represent RNA-seq exon coverage (log2 scaled) for human brain (SAMN03465403), kidney (SAMEA1970526), and liver (SAMEA2145122) datasets. (B) *ASPDH* gene expression from HPA RNA-seq normal tissue samples of 95 human individuals. (C) Heatmap of human *ASPDH* gene expression compared with tryptophan catabolism gene expression.

The *ASPDH* gene is highly tissue-specific and is mostly expressed in the kidney and liver, and less abundantly in the brain (Fig.10B). *ASPDH* is expressed similarly to other genes of tryptophan catabolism; in particular, mRNA of all the involved genes have been detected and are most abundant in kidney and liver (Fig.10C).

By performing a sequence analysis with PSort, ASPDH protein is predicted to be cytoplasmic (43.5 %: cytoplasmic, 30.4 %: mitochondrial, 17.4 %: nuclear) similarly to other component of the *de novo* NAD biosynthesis.

Human ASPDH protein (*Hs*_ASPDH) is predicted to have the "homoserine dehydrogenase, NAD binding domain" (CL0139, 9.5e-19, residues 13-124), a structural domain that adopts a Rossmann fold to bind NAD, and the "aspartate dehydrogenase, C-terminal" (CL0063, 4.6e-23, residues 169-261) with unknown function, as emerged from Pfam protein families collection.

The C-terminal domain of homoserine dehydrogenase contributes a single helix to this structural domain (residues 261-273), which is not included in the Pfam model.

NAD binding domain maintenance and active site changes in *Hs*_ASPDH main isoform

To investigate the loss or maintenance of the ancestral enzymatic activity in metazoan ASPDH-like sequences and to examine protein functional features, we compared human ASPDH isoform 1 (*Hs*_ASPDH) with the extensively studied tridimensional model of *Archaeoglobus fulgidus* L-aspartate dehydrogenase (2DC1)[65].

The *Hs*_ASPDH 3D structure (NP_001108070.1) was modeled using *A. fulgidus* 2DC1 as a template and then superimposed on the monomer; the QMEAN of the model is 0.56 probably due to the low level of identity between the two sequences (23.04%). Similar values are found by using *Thermotoga maritima* structural model 1J5P.1 as template[65] for human protein and, since the 3D structural comparison on both the two models was very similar, we considered only the comparison between *Hs*_ASPDH and 2DC1 in this dissertation; in fact, solely the thermophilic protein was crystallized in the presence of both the cofactors and an analog of the substrate, and these additional components are necessary to compare the functional sites of the protein.

The tridimensional structural model appears to be globally overlapping (Fig.11A) and the domain organization was similar to that of *A. fulgidus* and *T. maritima* L-ASPDH. Both proteins present a Rossmann fold motif at the N terminus, which is involved in binding nucleotide cofactors, and a shorter C terminal domain composed of alpha-helices and beta-strands that helps to mediate the dimerization of the protein[58]; the maintenance of both the structural

domains in human protein is supported by the sequence analysis performed with Phyre2 (Fig.11B).

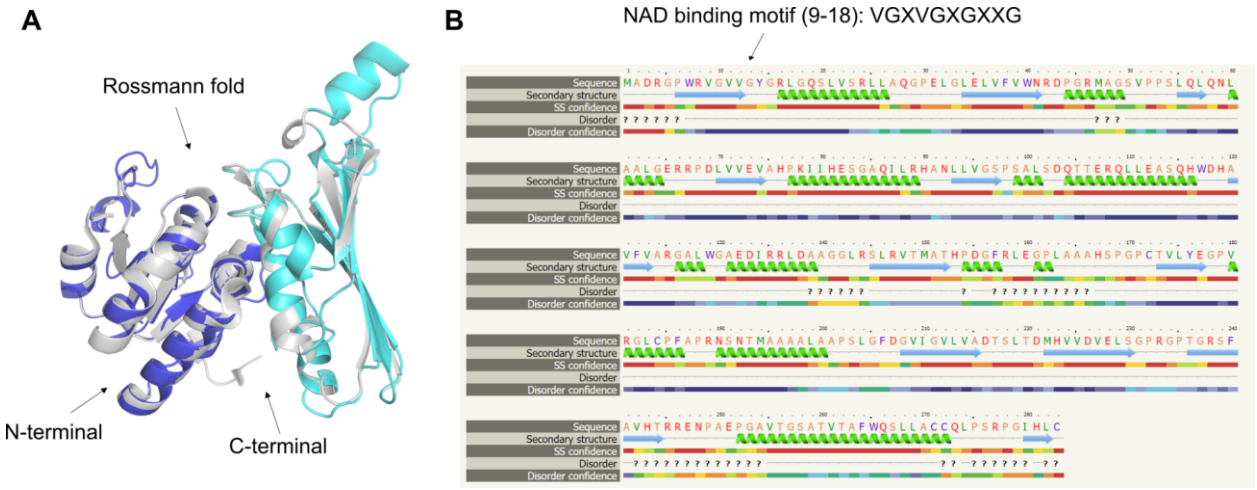


Figure 11: *Hs*_ASPDH structure and sequence analysis. (A) Structural domains of *Hs*_ASPDH model (colored cartoon) superimposed with *Af*_ASPDH 2DC1 structure (grey cartoon). Different structural domains[65] are colored: Rossman fold (dark blue), "aspartate dehydrogenase" (light blue). (B) The secondary structural assignments provided with Phyre2 for the *Hs*_ASPDH structure are shown; the glycine-rich motif is pointed at the N-terminal.

The Rossmann fold has been shown to be involved in the NAD(P) or SAM binding according to the protein function[78], i.e. oxidoreductase or transferases; *Hs*_ASPDH shows the characteristic NAD(P) binding motif VGXGXXG at the N-terminus (Fig.11B), suggesting the maintenance in coordinating the cofactor and its involvement in an oxidoreductase enzymatic function. Despite the low sequence identity between different protein families, this structural domain is strictly conserved in proteins owning an oxidoreductase and dehydrogenase activity, such as aspartate dehydrogenases, glyceraldehyde-3-phosphate dehydrogenases, biliverdin reductase, udp-glucose dehydrogenase, dihydrodipicolinate reductase; in contrast, it has been reported that L-aspDH proteins do not share structural similarity with the superfamilies that include L-amino acid dehydrogenases (L-leucine, L-valine, L-glutamate and L-phenylalanine dehydrogenases)[58].

Yoneda and colleagues[65] identified the amino acids responsible for NAD binding and substrate coordination in the hyperthermophilic protein *Af*_ASPDH used to model *Hs*_ASPDH tridimensional structure. With respect to *Af*_ASPDH NAD binding pocket (Fig.12A), the already mentioned glycine-rich motif (residues 7-12) lies close to adenine ribose, and side chains of Ser59, Tyr66, and Asp65 form hydrogen bonds via a water molecule with N7, N1 and N6 of the

adenine base. 2' and 3' hydroxyl groups of adenine ribose of NAD interact with the side chain of Asp31 by hydrogen bonds while Ala10, Ile11, Arg33, Ala57, and two water molecules are involved in the binding of the adenine phosphate group. Asn212 and Ala57 interact, together with a water molecule, with nicotinamide phosphate; Asn162, Ser81, and backbone oxygen Ala58 with the hydroxyl groups of nicotinamide ribose. Carboxyamide moiety nicotinamide ring establishes interactions with Thr166, and backbone NH and oxygen of Leu80 and Ala108 respectively.

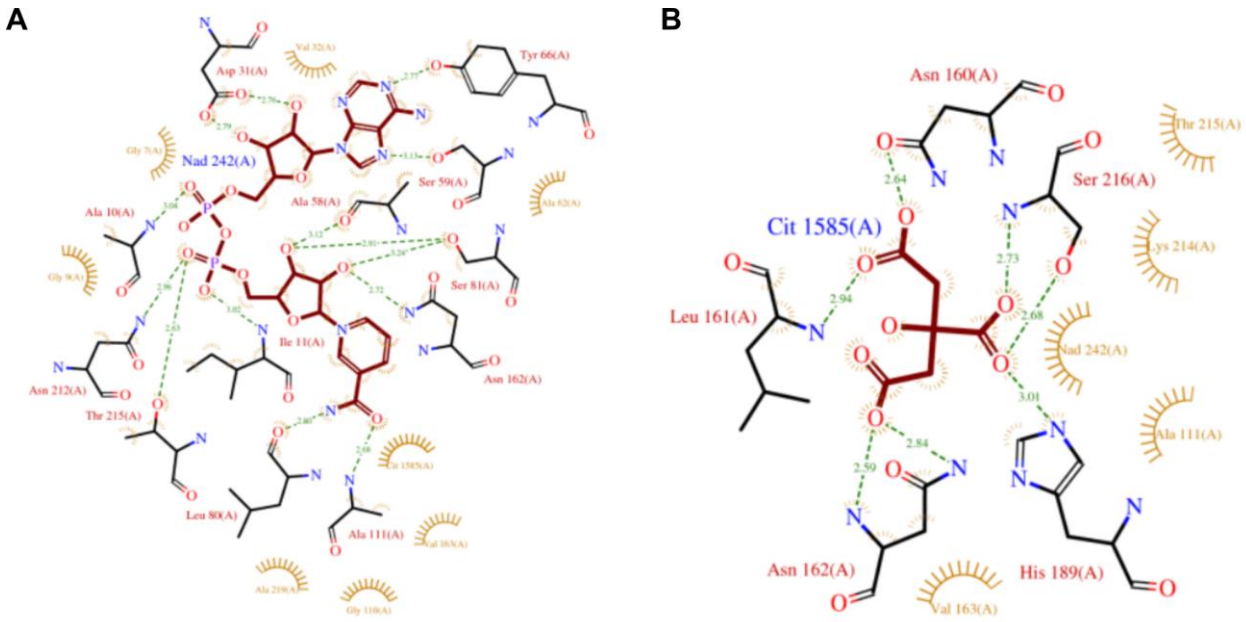


Figure 12: 2DC1 ligand interactions reported in the PDBsum database.[65] (A) LIGPLOT of interactions involving NAD cofactor. (B) LIGPLOT of interactions involving ligand citrate, the substrate analog.

Considering citrate as the substrate analog, the authors have identified six main amino acids involved in its coordination (Fig.12B), even though it is necessary to consider that aspartate possesses an additional NH_2 group and different charge distributions with respect to citrate. Citrate C1 carboxyl group interacts with the Lys134 side chain and with the Leu161 main chain amide; the C5 carboxyl group with the Arg162 side chain and Arg162 and Val163 backbone amides. Ser216, Asn187, and His189 form hydrogen bonds and ionic bonds with C6 carboxylate. C3 hydroxyl interacts with Lys134. Taken together, all these strong and weak interactions cooperate in holding the substrate with NAD.

To gain insight into *Hs*_ASPDH, we used *Af*_ASPDH as the reference to identify if NAD binding pocket is maintained in human protein; together with the structural evidence of Rossmann fold and glycine-rich motif conservation, we observed the conservation of most amino acids involved in the NAD coordination with some exceptions (Fig.13A): in fact, we can observe the switch of Asp65, Tyr66, and Thr166 into amino acids owning a similar side chain for either a sterical hindrance or charge distribution.

Non-conservative substitutions may still favor the ability to coordinate NAD because the residues are not directly involved in the cofactor binding but form hydrogen bonds with the NH_2 or carboxylate of the main chain or interact via water molecules (i.e. Ser59, Ala57); changes in the side chain size have been probably accepted by evolution to accommodate global structural rearrangements.

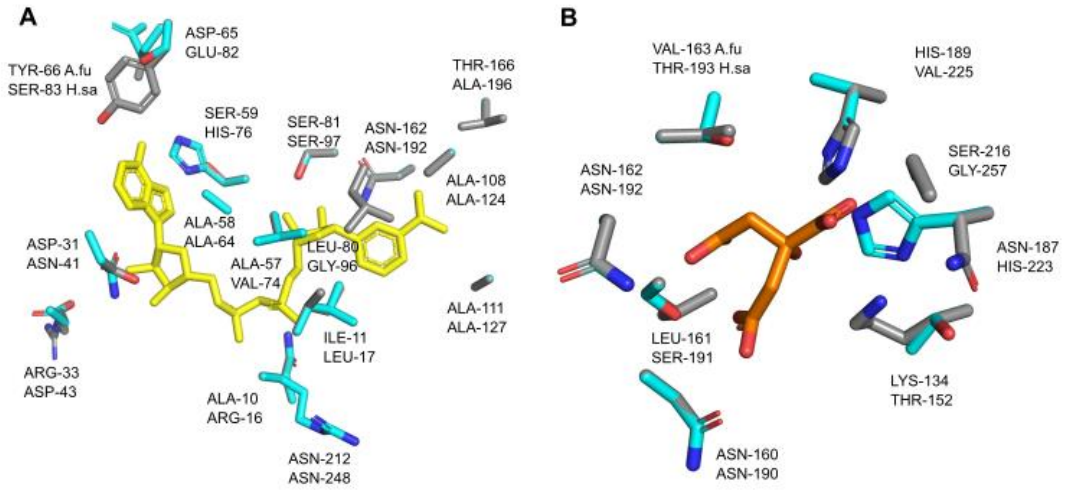


Figure 13: Comparison between NAD⁺ and substrate binding sites. (A) Close-up comparison between NAD binding site of *A. fulgidus* L-ASPDH (2DC1; grey carbons) and the predicted 3D model of *Hs*_ASPDH (light blue carbons) showing the conservative substitutions of residues involved in this interaction. Top labels refer to *A. fulgidus* L-ASPDH and down labels to *Hs*_ASPDH. Nicotinamide adenine dinucleotide is represented as yellow sticks, while residues relevant for its binding in 2DC1 and equivalent residues in *Hs*_ASPDH protein are represented as CPK-colored sticks. (B) Close-up comparison between substrate binding sites showing the conservative substitution of some residues together with the non-conservative substitution of other amino acids (Lys134-Thr152; Asn187-His223; His189-Val225; Leu161-Ser191; Ser216-Gly257). The substrate binding site is represented similarly to NAD binding site.

As for the active site, half of the residues involved in substrate coordination are lost and have been mutated into amino acids with different biochemical characteristics (Fig.13B), suggesting a global rearrangement of the active site. *Af*_ASPDH His189 is substituted by valine in human protein, but histidine replaces in its turn the adjacent Asn187, suggesting the importance of its

imidazole side chain in coordinating the substrate. The acquisition or loss of charged side chains in human protein, i.e. for Leu161-Ser191 and Ser216-Gly257, is in accordance with the hypothesis of a different chemical compound. The conservation of some residues otherwise suggests the possibility to bind a substrate that shares similar charge distribution and possesses some similar functional groups as the archaebacterial protein.

Both the NAD cofactor and the substrate fit in human tridimensional model as emerged by the structural comparison with *A. fulgidus* protein (Fig.14A); the charge distribution of the surface residues in the NAD binding cavity is similar to that of *Af*_ASPDH (Fig.14B-14C) and is suitable for NAD coordination.

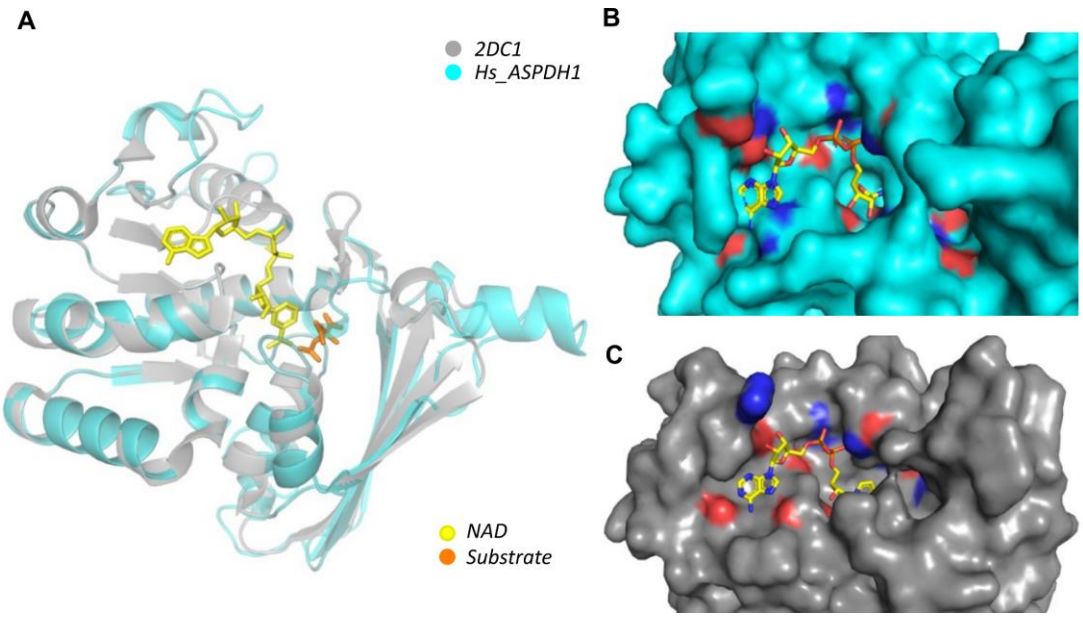


Figure 14: Homology modeling of *Hs*_ASPDH and superimposition with 2DC1 template. (A) Structural model of *Hs*_ASPDH (light blue cartoon) model superimposed with *Af*_ASPDH structure (gray cartoon). NAD and the substrate homolog (citrate) are drawn in yellow and orange sticks, respectively. (B) Surface-accessible NAD binding cavity in *Hs*_ASPDH 3D model with emphasis on positive (blue) and negative (red) charge distribution. (C) Surface-accessible NAD binding cavity in 2DC1 with emphasis on positive (blue) and negative (red) charge distribution.

Structural analysis and investigation of the arrangement of the amino acids in space are necessary to study *Hs*_ASPDH functional features because spatial superimposition does not entirely coincide with sequence comparison in a multiple alignment performed by selecting bacterial, archaeal, and metazoan sequences (Fig.15) for both NAD coordination (Fig.15, blue triangles) and substrate (Fig.15 yellow stars). For example, His189 involved in substrate binding is maintained in all the selected sequences, suggesting also the maintenance of its

functional role; otherwise, we have already noticed that histidine has been conserved within the active site, but its structural shift is in accordance with an active site alteration.

In this interesting case, the discussion of the probable protein function and role is clearly supported by structure-based investigations together with sequence-based inferences.

The maintenance of critical residues involved in NAD coordination, together with point mutations in residues related to L-aspartate binding, could suggest the possibility to accomodate a different substrate with similar biochemical characteristics or with a similar functional group distribution, as the 2-aminomuconate.

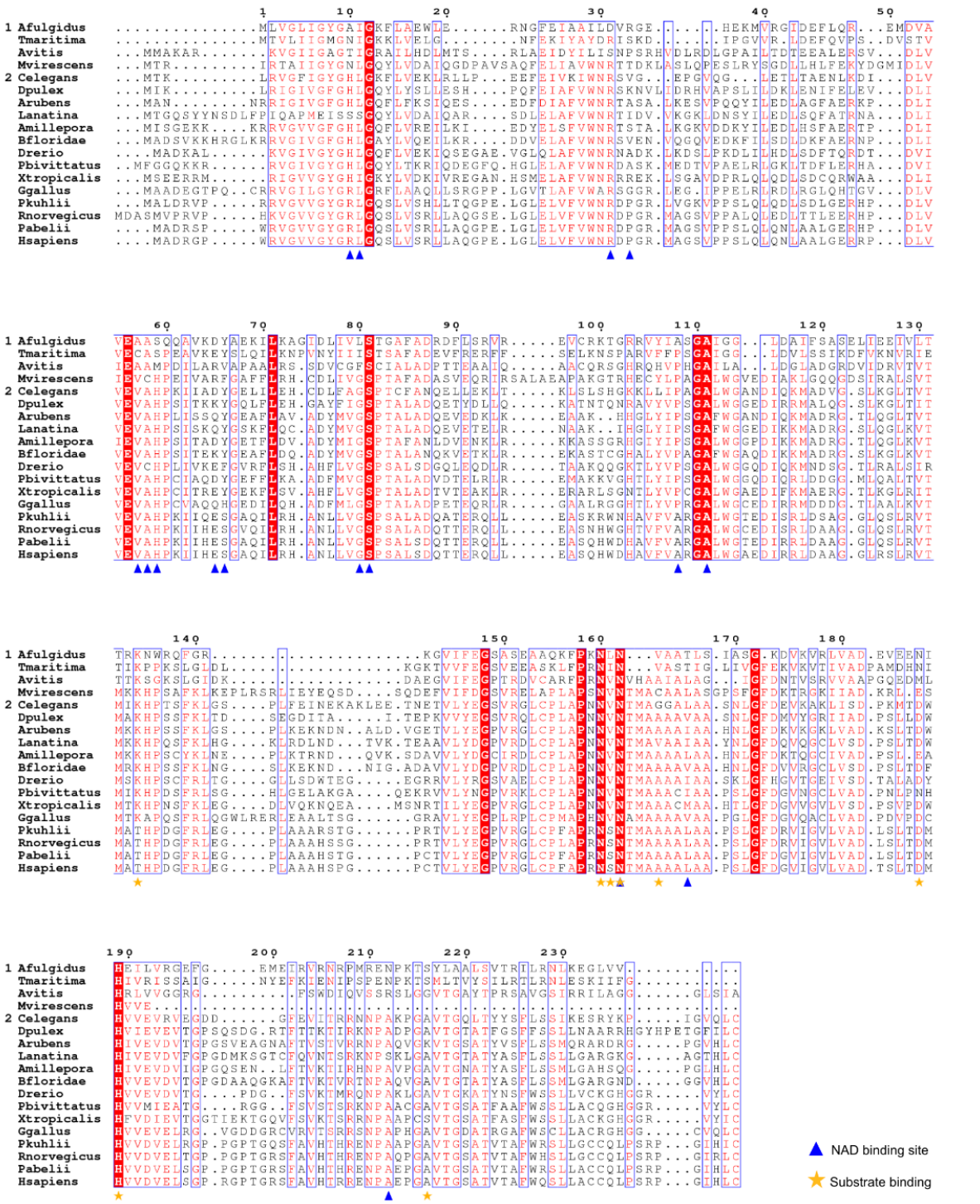


Figure 15: Multiple alignment of ASPDH and ASPDH-like sequences. Multiple alignment of ASPDH proteins (group 1) from bacteria and ASPDH-like sequences (group 2) from Metazoa. Amino acids conserved in all sequences are shaded red. Residues involved in the substrate binding are pointed with yellow stars, and residues involved in NAD binding are pointed with blue triangles.

*Hs*_ASPDH putative substrate and reaction

ASPDH and two other genes involved in the NAD *de novo* biosynthesis, i.e. *nadA* and *nadC*, are included in a three-gene operon in all the microorganisms presenting the pathway. The function of bacterial ASPDH proteins is likely the synthesis of iminoaspartate from L-aspartate; the iminoaspartate product is unstable in an aqueous solution and its hydrolysis leads to oxaloacetate and ammonia (Fig.16A). No activity was detected with D-aspartate, L-glutamate, or L-asparagine. *A. fulgidus*, *T. maritima*, and *P. aeruginosa* L-aspartate dehydrogenase enzyme activities require oxidation of L-aspartate in the presence of NAD or NADP as the electron acceptor.

Both *T. maritima* ASPDH (*Tm*_ASPDH) and *Af*_ASPDH utilize NAD and NADP with the same affinity; the K_M values for L-aspartate, NAD or NADP of *A. fulgidus* ASPDH are quite similar to those of the *T. maritima.* Enzyme activities were unaffected by metal and EDTA treatment and appeared to be independent of their presence[58,65,73].

Since the NAD *de novo* synthesis in humans follows a different metabolic route, and the active site of *Hs*_ASPDH underwent rearrangement, we may hypothesize that the *Hs*_ASPDH gene has modified its activity to perform a different enzymatic reaction. The maintenance of the Rossmann fold and NAD binding cavity is in accordance with the conservation of oxidoreductase activity.

Most human and mammal genes involved in the tryptophan degradation pathway have been identified and the corresponding enzymes have been characterized; however, the last reactions of the pathway need to be clarified. In particular, the gene encoding a 2-aminomuconate reductase, the enzyme involved in the non-reversible reduction of 2-AM in the presence of NAD(P)H with the release of ammonia, has not been identified yet.

The co-occurrence signal emerging from our computational analysis for *Hs*_ASPDH and genes of tryptophan catabolism suggests the involvement of *Hs*_ASPDH in this metabolic pathway. The direction of the missing reaction is opposite to the one catalyzed by *Af*_ASPDH, since human protein requires the reduced form of NAD(P), the NAD(P)H, to ensure the correct reductive deamination.

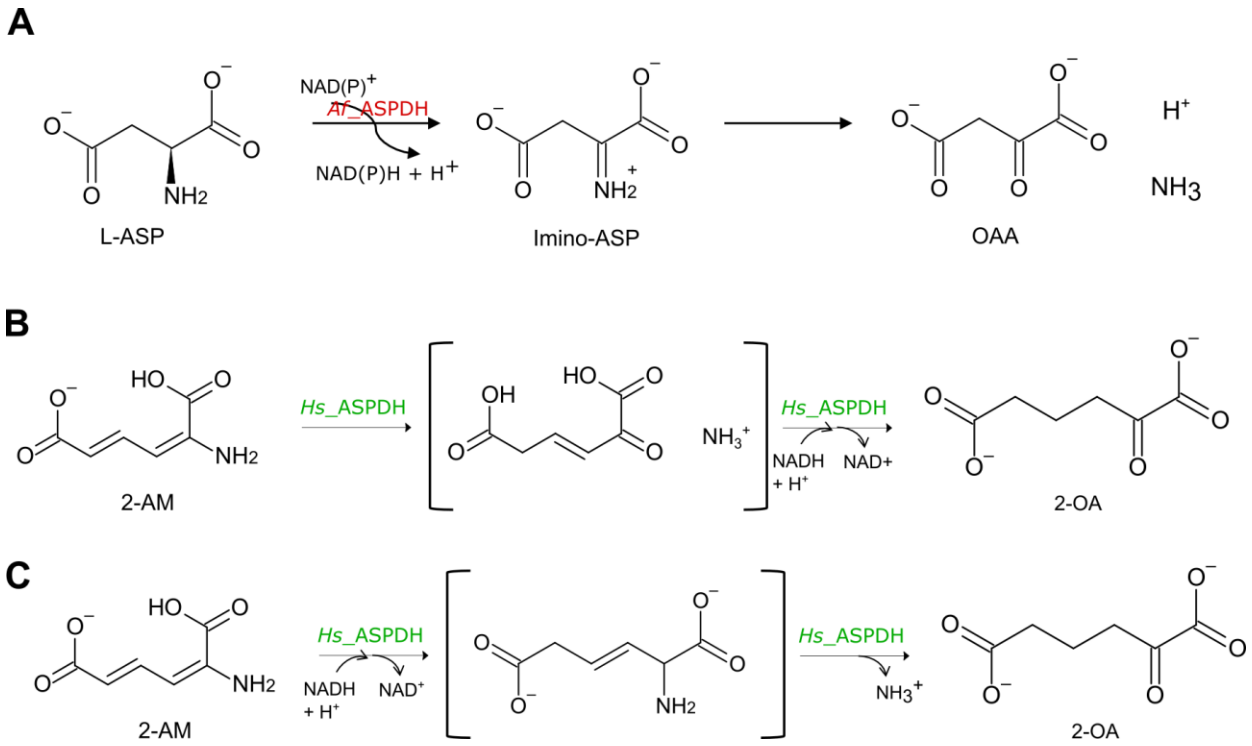


Figure 16: Comparison between the *Af*_ASPDH reaction and the putative *Hs*_ASPDH reactions. (A) L-aspartate oxidation to imino-aspartate in the presence of $NAD(P)^+$ catalyzed by *Af*_ASPDH, followed by IA spontaneous decay into oxaloacetate with the release of ammonium. (B) Deamination of 2-aminomuconate followed by the NADH-dependent reduction of the intermediate to form 2-oxoadipate (C) Reduction of 2-aminomuconate in the presence of NADH followed by the deamination of the intermediate to form 2-oxoadipate.

It is necessary to consider that the conversion of 2-aminomuconate into 2-oxoadipate could take place following two alternative reactions: in the first case, the 2-aminomuconate is deaminated by *Hs*_ASPDH and then reduced to 2-oxoadipate with the concomitant NADH oxidation to NAD^+ (Fig.16B); alternatively, *Hs*_ASPDH could reduce 2-aminomuconate prior to release the ammonia group (16C).

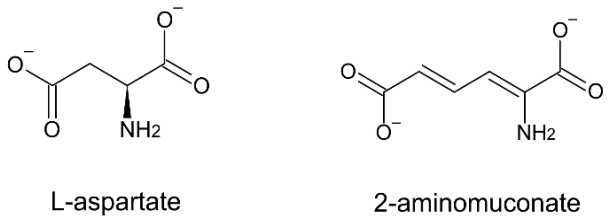


Figure 17: L-aspartate and 2-aminomuconate chemical structures.

One of the most relevant differences between the *Af*_ASPDH substrate, the L-aspartate, and the substrate for the putative reaction proposed, the 2-aminomuconate, is the different lengths of the main chain of the molecules (Fig.17): in fact, L-asp and 2-AM possess 4 and 6 carbons respectively. The different size of 2-AM with respect to L-asp can explain *Hs*_ASPDH need to rearrange the active site to accommodate a larger substrate.

In addition, no carbon-carbon double bonds are present in L-asp in contrast with the two double bonds in 2-AM; otherwise, both the chemical compounds own two terminus carboxylic groups and a closer amine group.

All carbons of the 2-AM show sp^2 hybridization within a planar geometry that allows a global molecular distortion; conversely, L-aspartate possesses 2 of the 4 carbons in a sp^3 hybridization state, resulting in greater torsional freedom.

Taken together, all this evidence and hypothesis have led us to assay both the ancestral and the putative enzymatic activities for *Hs*_ASPDH, with the purpose to clarify the strict evolutionary connection between this scarcely studied human enzyme with the well-known component of the kynurenine pathway.

Enzymatic preparation of 2-aminomuconate

Since 2-aminomuconate is not commercially available, we decided to synthesize it with an enzymatic approach. To obtain the desired compound, it was necessary to use 3-HAA as the initial substrate for three subsequent enzymatic reactions since it is the unique stable compound in the tryptophan degradation pathway prior to 2-AM. In fact, α-amino-β-carboxymuconate-ε-semialdehyde (ACMS) decays into quinolinate, the NAD^+ precursor, while 2-aminomuconic semialdehyde (2-AMS) is unstable and tends to rapidly cyclize into picolinate.

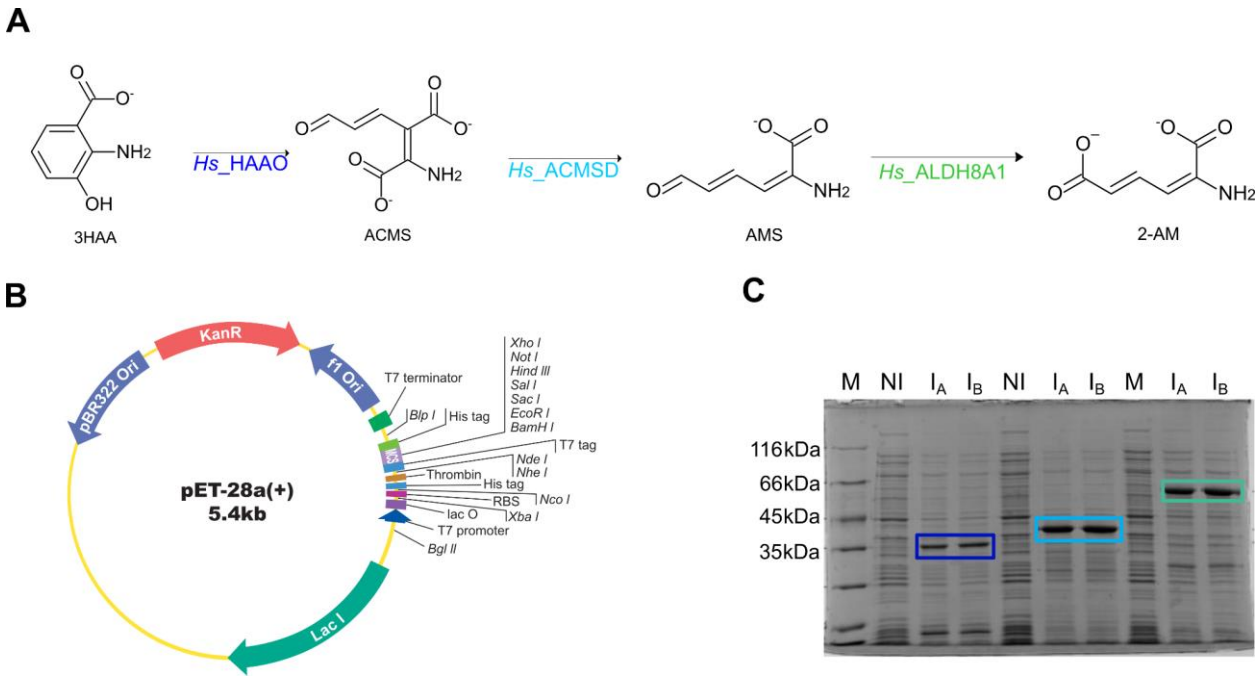


Figure 18: Experimental strategy to obtain the substrate of the putative substrate of *Hs*_ASPDH, the 2-aminomuconate. (A) Metabolic steps to obtain 2-aminomuconate (2-AM) from 3-hydroxyanthranilate (3HAA) with the consecutive enzymatic reactions catalyzed by *Hs*_HAAO, *Hs*_ACMSD, *Hs*_ALDH8A1. (B) *Hs*_HAAO, *Hs*_ACMSD, *Hs*_ALDH8A1 genes were cloned in pET-28a(+) by NdeI/XhoI. (C) SDS PAGE gel (12%) of *Hs*_HAAO (blue box), *Hs*_ACMSD (light blue box), *Hs*_ALDH8A1 (light green box) protein inductions in duplicate: M - marker; NI - not induced; I - induced.

Using the enzymatic approach (Fig.18A), 3-HAA was resuspended in an appropriate buffer to be enzymatically converted into ACMS by the recombinant 3-hydroxyanthranilate 3,4-dioxygenase (*Hs*_HAAO); in the presence of the recombinant amino-carboxymuconate semialdehyde decarboxylase (*Hs*_ACMSD), ACMS is converted into 2-AMS instead of the spontaneous cyclization. 2-AMS is then oxidized in a NAD-dependent manner by the 2-aminomuconic semialdehyde (*Hs*_ALDH8A1).

To succeed in our goal, we ordered the required human gene sequences in pET-28a vectors (Fig.18B). We transformed *E. coli* BL21-codon plus cells with the recombinant constructs and we induced gene expressions with IPTG for 20 °C o/n (Fig.18C). Proteins were purified using FPLC chromatography and assayed to confirm their enzymatic activity.

We report here our improved enzymatic method for the preparation of the 2-aminomuconate as the final product of the enzymatic conversion of 3-hydroxyanthranilate in the presence of the recombinant form of *Hs*_HAAO, *Hs*_ACMSD, *Hs*_ALDH8A1 proteins.

*Hs*_HAAO protein purification and activity assay

3-hydroxyanthranilic acid dioxygenase catalyzes the cleavage of the benzene ring of 3-HAA; in the absence of the specific decarboxylase, the product of the reaction slowly cyclizes in quinolinate (Fig.19A). Its activity is strictly dependent on the presence of a Fe^{2+} ion within the active site, which is involved in the activation of an oxygen molecule and in the progress of the enzymatic reaction[79]; in addition to this, the HAAO/3HAO protein family is characterized by the presence of two conserved histidines and a glutamate residue involved in the binding of the substrates and in the stabilization of the overall protein structure[80].

*Hs*_HAAO was overexpressed in the bacterial system and soluble fractions deriving from previously IPTG-induced cells were purified by performing liquid affinity chromatography (FPLC) on a nickel column; the total process had an average yield of about 20 mg per liter as a result of a good solubility of the protein (Fig.19B).

After rejoining the eluted fractions containing the protein of interest, we spectrophotometrically monitored the reaction following the disappearance of the 3-HAA peak at 315 nm and the rapid formation of ACMS at 360 nm (Fig.19C) in the presence of 1-10 µM *Hs*_HAAO. We confirmed the previously described dependence of the non-heme ferrous iron on the enzymatic activity by adding $(NH_4)Fe(SO_4)_2$ to the enzyme aliquot, since the freshly prepared protein appears to be less active in the absence of added metal ions.

After the total substrate conversion, spontaneous cyclization has been detected by the decrease in absorbance at 360 nm during the time (Fig.19D). Considering an ε_{360nm} of 47,500 M^{-1} cm^{-1} [24], we fitted the experimental data points corresponding to ACMS spontaneous enclosure to quinolinate with the first-order equation S[t] ~ S0*e(-k * t) (Fig.19E): the values of the initial concentration S_0 and the slope k calculated were respectively 39.510 ± 0.007 µM and 0.351 ± 0.008 s^{-1}.

After testing the enzymatic activity, eluted fractions were added with 1 mM of reduced glutathione and 1 mM ammonium iron sulfate[79], frozen in liquid nitrogen, and stored at -80 °C. In the absence of them, protein aliquots rapidly lose their enzymatic activity after thawing.

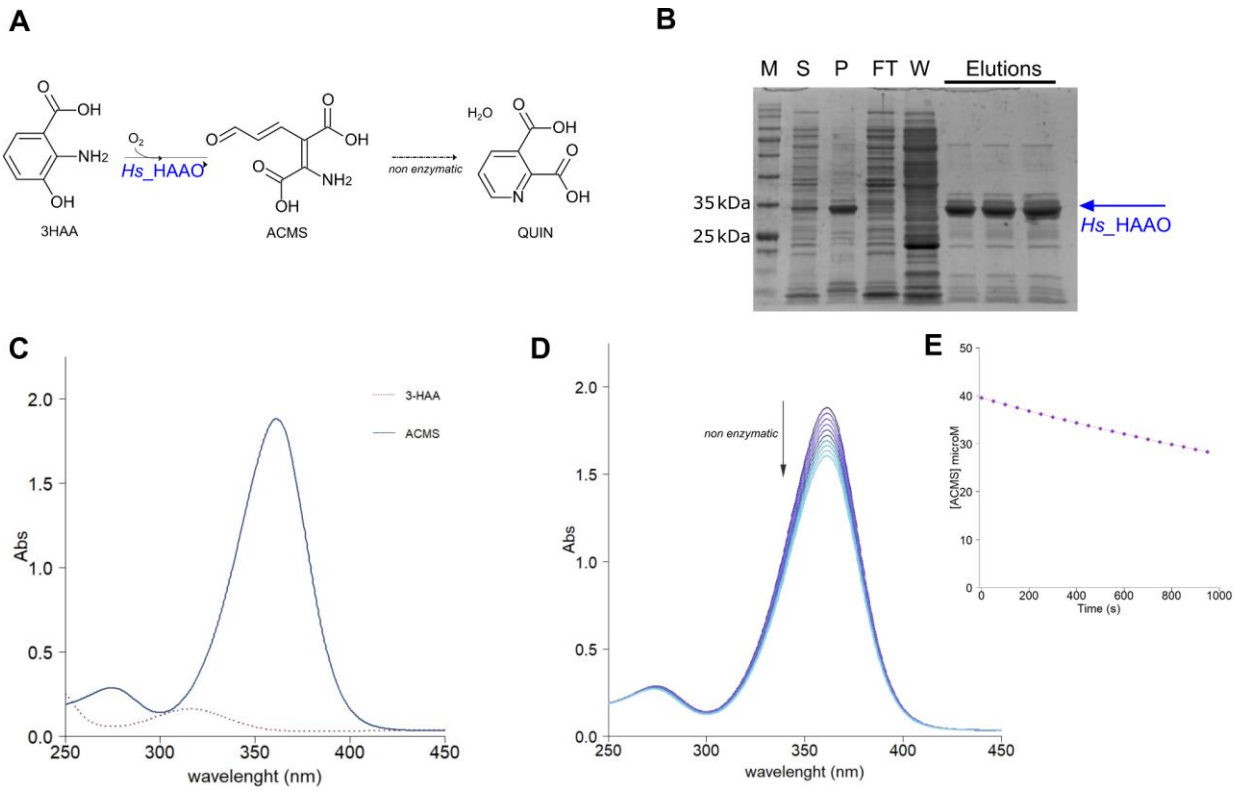


Figure 19: *Hs*_HAAO protein expression, induction, purification, and activity assay. (A) *Hs*_HAAO reaction scheme of 3-HAA oxidative conversion to ACMS, which is followed by its spontaneous cyclization into quinolinate (QUIN). (B) SDS-PAGE 12% of *Hs*_HAAO expression and purification through FPLC: M, marker; S, supernatant; P, pellet; FT, flow-through; W, washing; E, elutions. (C) Enzymatic conversion of 3-HAA (red dotted line) into ACMS (blue solid line) in presence of 10 µM *Hs*_HAAO freshly purified. (D) Spontaneous decay of ACMS into QUIN over time. (E) Kinetics point of ACMS conversion into quinolinate overtime at the fixed wavelength of 360 nm.

*Hs*_AMCSD protein purification and activity assay

ACMSD protein plays a key role in the regulation of NAD synthesis in response to cellular stimuli because it competes with the spontaneous QUIN-forming reaction; additionally, ACMSD activity depends on the presence of divalent metal ions (zinc, cobalt, iron, and manganese)[24]. After having assayed *Hs*_HAAO activity, we expressed and purified the following protein of the pathway, *Hs*_ACMSD, which is required for the 3-HAA conversion to ACMS, (Fig.20A), similarly to *Hs*_HAAO. As described in literature, this protein appears to be poorly soluble (Fig.20B) if expressed without the coexpression of molecular chaperones that improve solubility[24]. Although we did not use chaperon expression, we had an average yield of 4 mg/L which ensured us enough protein to perform enzymatic activity assays.

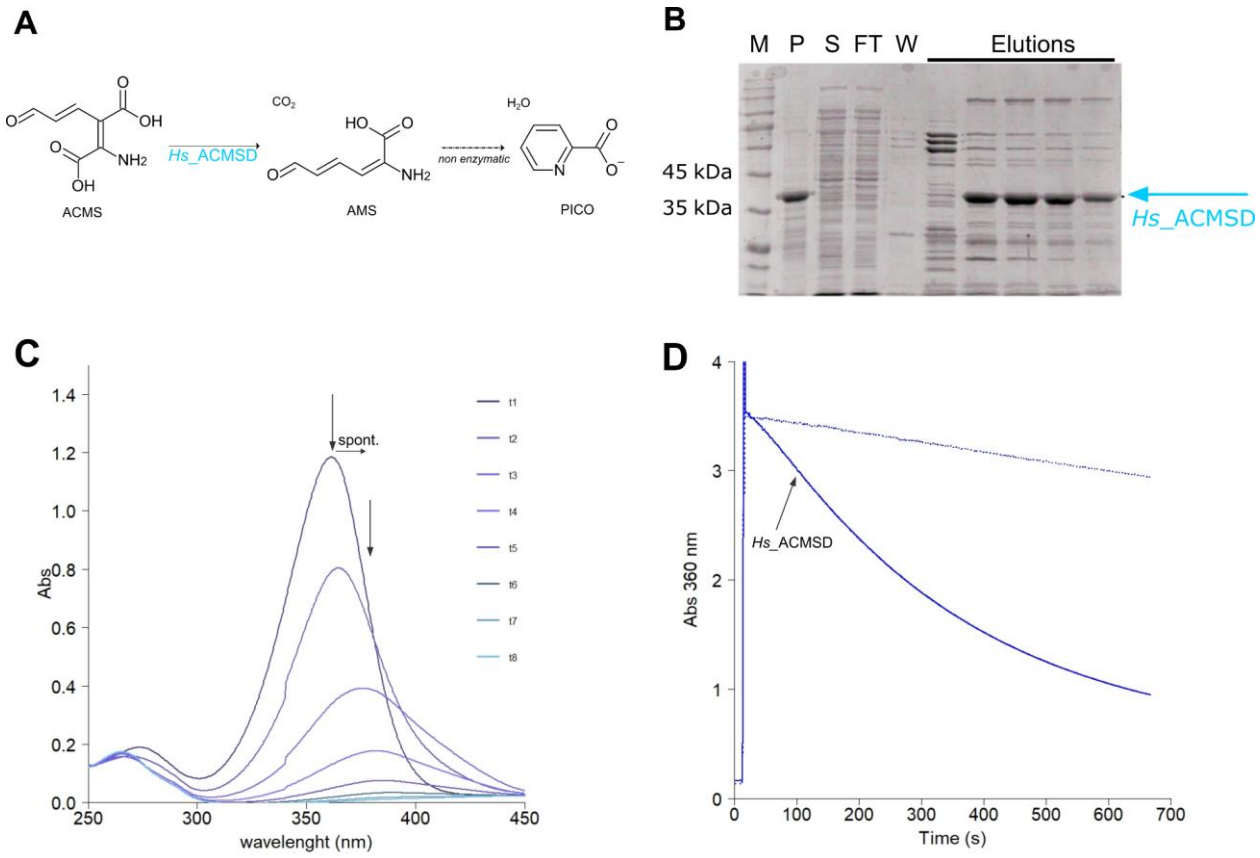


Figure 20: *Hs*_ACMSD protein expression, induction, purification, and activity assay. (A) *Hs*_ACMSD reaction scheme of ACMS decarboxylation into AMS, which is followed by the spontaneous cyclization into picolinate (PICO). (B) SDS-PAGE 12% of *Hs*_ACMSD expression and purification through FPLC: M, marker; P, pellet; S, supernatant; FT, flow-through; W, washing; E, elutions. (C) Enzymatic conversion of ACMS into AMS in presence of 1 µM freshly purified Hs_ACMSD. Spectra were acquired at intervals of 30 s. (D) Time-course comparison between spontaneous decay of ACMS into QUIN (blue dotted line) and enzymatic conversion of ACMS into AMS (solid blue line) at the fixed wavelength of 360 nm.

We detected the conversion of ACMS to 2-AMS with the release of a carbon dioxide molecule after the addition of *Hs*_ACMSD to the reaction mixture containing *Hs*_HAAO and its reaction product obtained from 3-HAA. The occurrence of the ACMSD-catalyzed decarboxylation was observed by monitoring the decrease in absorbance at 360 nm together with a red-shifted shift in the absorbance peak to 375 nm (Fig.20C) as previously demonstrated[81]. In the absence of the following enzyme of the pathway, the 2-AMS rapidly decays to picolinate causing the breakdown in the absorbance spectrum over time (Fig.20C).

*Hs*_ACMSD clearly competes with the nonenzymatic decay of its substrate to QUIN and presents a higher initial velocity V_0 in the presence of the same initial substrate concentration (Fig.20D). The evaluation of the *Hs*_ACMSD chemical parameters for ACMS has several limitations since the substrate saturates the absorbance signal at low concentrations.

It has otherwise been demonstrated that *Hs*_ACMSD is more active in the dimeric form and bound to a divalent zinc ion[24,26].

To optimize protein production, it could be necessary to improve protein expression and stability by adding zinc in the growth medium and/or to perform induction by substituting IPTG with lactose in the medium. In addition to this, gene coexpression with chaperones could cooperate with other optimization to produce a more soluble recombinant protein.

*Hs*_ALDH8A1 protein purification and activity assay

We then expressed, induced, and purified the third protein required, the recently identified dehydrogenase *Hs*_ALDH8A1. This protein has been recently reassigned[30] to the missing 2-AMS dehydrogenase and performs the NAD-dependent oxidation of 2-AMS to 2-AM (Fig.21A). 2-AM is although unstable and tends to spontaneously deaminate to 2-O-3-E in solution (Fig.21A).

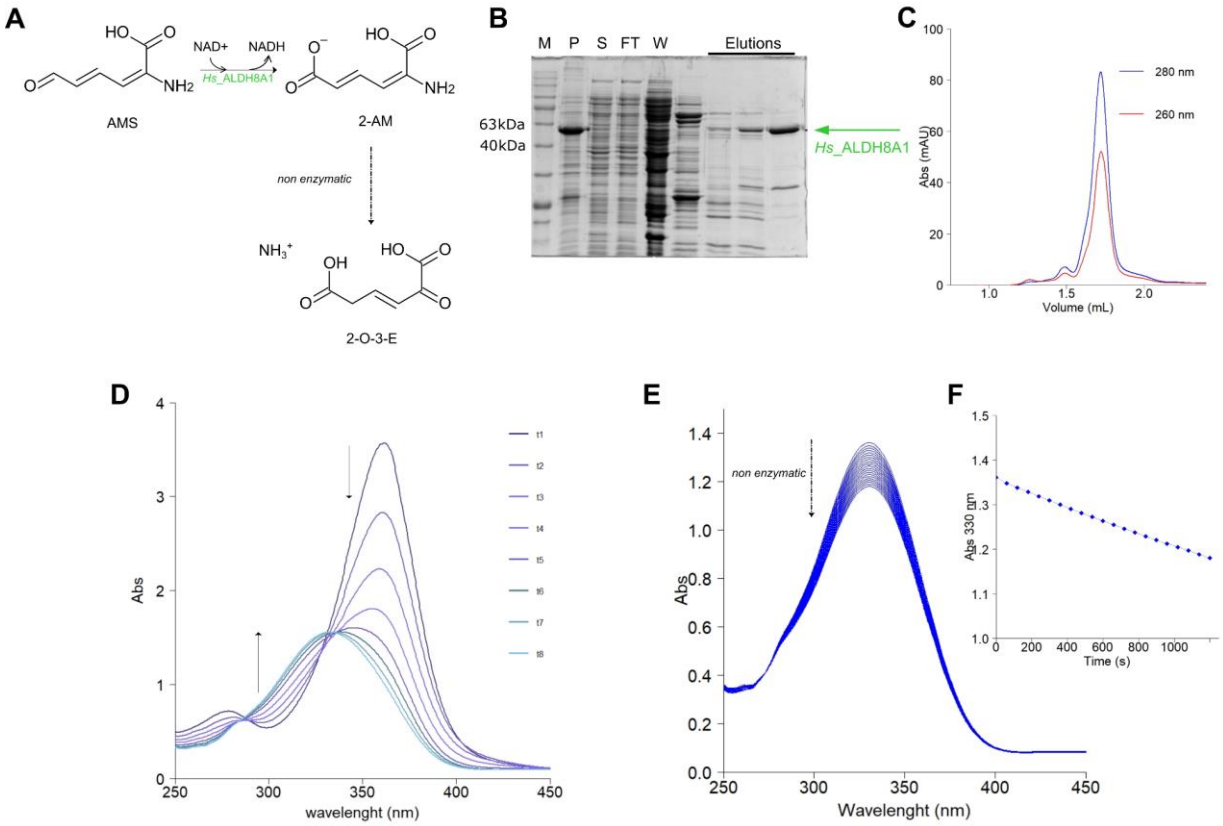


Figure 21: *Hs*_ALDH8A1 protein expression, induction, purification, and activity assay. (A) *Hs*_ALDH8A1 reaction scheme of AMS conversion into 2-AM, which is then followed by the spontaneous deamination into 2-O-3-E (2-oxo-3-hexenedioate.). (B) SDS-PAGE 12% of *Hs*_ALDH8A1 expression and purification through FPLC: M, marker; P, pellet; S, supernatant; FT, flow-through; W, washing; E, elutions. (C) Size-exclusion chromatography of 1.00 µg/µL *Hs*_ALDH8A1 purified with FPLC. The protein appears to be present in a monomeric form (solid blue line). Absorbance spectrum monitored at 260 nm is in agreement with a pure protein sample (solid red line) and with the absence of NAD^+ bound to the enzyme. (D) Enzymatic conversion of AMS into 2-AM in presence of 1 µM freshly purified *Hs*_ALDH8A1. Spectra were acquired at intervals of 30 s. (E) Superimposed spectra of spontaneous 2-AM deamination into 2-O-3-E. Spectra were acquired at intervals of 60 s. (F) Kinetics point of 2-AM deamination overtime at the fixed wavelength of 330 nm.

*Hs*_ALDH8A1 appears to be poorly soluble (Fig.21B) if expressed in a bacterial system and induced with the addition of 0.5 mM IPTG, and the chromatographic purification on a nickel column yielded an average amount of 2 mg/L of protein. As previously demonstrated[30], *Hs*_ALDH8A1 is present in solution in a single monomeric state (Fig.21C) and we observed that the recombinant protein is eluted in the absence of bound NAD molecules (Fig.21C) as deduced by the 260/280 absorbance ratio measured in the correspondence of the protein elution, i.e. 0.62 (peak at 1.71 mL: 0.083 for 280 nm, 0.052 for 260 nm).

We then detected the enzymatic conversion of 2-AMS to 2-aminomuconate after the addition of the eluted fractions to the reaction mixture containing the two previous enzymes, their reaction product, and NAD; we observed a decrease in the absorbance at 360 nm (Fig.21D), but not the absorbance shift to 375 nm because the enzyme prevents the picolinate formation. In addition to this, the characteristic peak of 2-AM is clearly distinguishable at 330 nm at the end of the reaction (Fig.21D). It is necessary to take into consideration that an absorption contribution to the 330 nm peak is given by the enzymatic NAD^+ reduction to NADH, which presents a characteristic absorbance peak at 340 nm and, consequently, absorbs also in the surrounding wavelengths.

The enzymatically produced 2-AM tends to deaminate over time as clearly distinguishable by the decrease in absorbance at 330 nm (Fig.21E, 21F).

After testing the enzymatic activity, eluted fractions were added with 10%, frozen in liquid nitrogen, and stored at -80 °C. In the absence of glycerol, protein aliquots fully lose their enzymatic activity in a few days after thawing.

*Hs*_ASPDH protein expression, induction, and purification

The protein encoded by the human *ASPDH* gene was proposed in this dissertation as the best candidate to act downstream of *Hs*_ALDH8A1 and carry out the enzymatic reduction of 2-aminomuconate to 2-oxoadipate (2-OA) with the release of an ammonium molecule (see Fig.16B-16C) in the presence of NADH.

To confirm this hypothesis, we transformed the recombinant pET-28a(+) vector containing the *Homo sapiens* isoform 1 of the "aspartate dehydrogenase domain-containing protein" sequence (NM_001024656.3) plus an N-terminal His-Tag in the bacterial strain *E.coli* BL21c+. Several tests were performed on cell cultures by combining different inductor concentrations, growth temperatures (37, 30, and 20 °C) and durations (1, 3, and 16 hours) in order to understand which conditions were most suitable to obtain the protein expression and induction (Fig.22A). Best protein induction was achieved by growing transformed cells at 30 °C for 3 hours, after the addition of 1 mM IPTG and 0.5 mg/L L-arabinose as confirmed by SDS PAGE gel electrophoresis (Fig.22A).

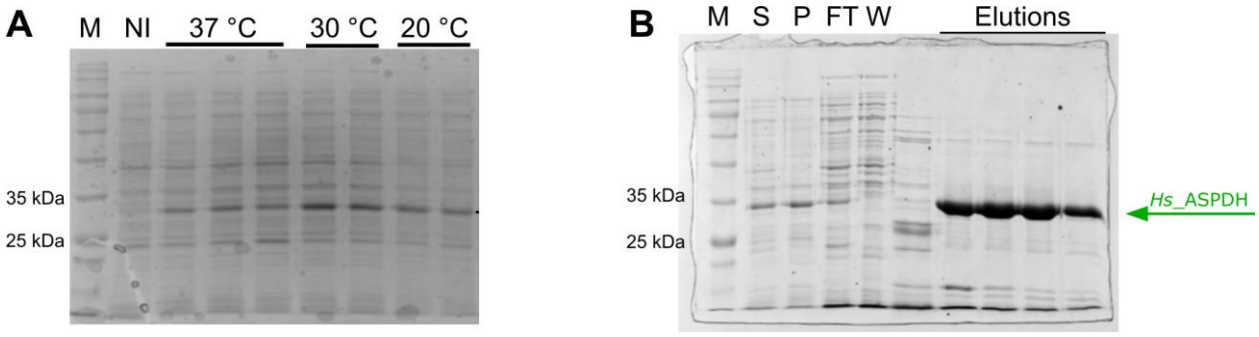


Figure 22: *Hs*_ASPDH protein expression, induction, and purification. (A) SDS-PAGE 12% *Hs*_ASPDH induction at 37 °C, 30 °C and 20 °C after the addition of 0.5 mM IPTG and 0.5 mg/mL L-arabinose. (B) SDS-PAGE 12% of *Hs*_ASPDH expression and purification through FPLC: M, marker; S, supernatant; P, pellet; FT, flow-through; W, washing; E, elutions.

These experimental conditions were chosen to express, induce and purify *Hs*_ASPDH protein using FPLC from 1 liter of transformed cell culture; protein appears to be distributed in equal quantities in both the soluble and insoluble fractions (Fig.22B) and presented the expected molecular weight of ∼31 kDa.

*Hs*_ASPDH acquisition of a novel enzymatic activity in tryptophan catabolism

We started to investigate whether *Hs*_ASPDH could catalyze the enzymatic reduction of 2-aminomuconate to 2-oxoadipate with the release of a molecule of ammonium (see Fig.16B).

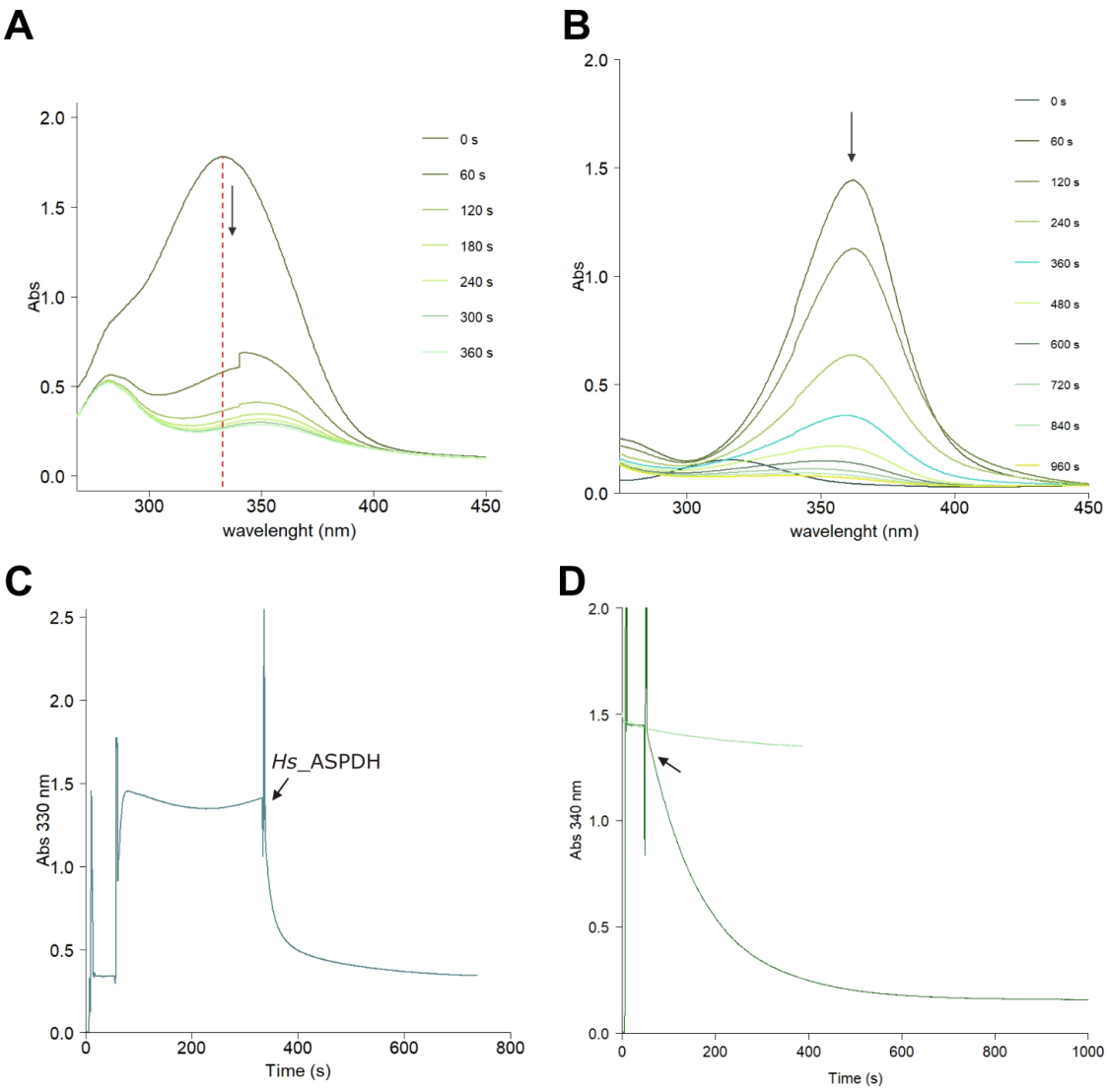


Figure 23: Identification of a novel enzymatic activity for *Hs*_ASPDH. (A) Superimposed spectra of time resolved 2-AM consumption (330 nm peak) after the addition of 1 µM *Hs*_ASPDH to the reaction mixture, at pH =7.6 and 25 °C, recorded at 60 second intervals. (B) Four-steps conversion of 3-HAA to 2-OA in the presence of NAD^+ and the four enzymes required (*Hs*_HAAO, *Hs*_ACMSD, *Hs*_ALDH8A1, *Hs*_ASPDH) recorded at 60 second intervals. (C) Time course of 3-HAA conversion monitored at 330 nm in the presence of *Hs*_HAAO, *Hs*_ACSMD and *Hs*_ALDH8A1 followed by 2-AM consumption after the addition of *Hs*_ASPDH, at pH =7.6 and 25 °C. (D) Enzymatic release of ammonia from 2-AM was monitored with a continuous coupled assay with 4 U GDH at 340 nm in presence of 0.25 mM NADH and 1 mM ᾳ-KG, at pH =7.6, in the reaction mixture once *Hs*_ASPDH reaction was over (dark green line) or in the same conditions but in absence of *Hs*_ASPDH (light green line).

In the presence of the enzymatically synthesized 2-AM, we observed a rapid absorbance decrease at 330 nm (Fig.23A, 23C) and a shift to 350 nm after the addition of *Hs*_ASPDH consistent with the consumption of 2-AM, at 25 °C and pH 7.6.

We followed the four consecutive enzymatic reactions including all the enzymes in the reaction mixture in the presence of NAD^+ (Fig.23B); since we did not observe the appearance of the characteristic peak of NADH at 340 nm after *Hs*_ALDH8A1, we hypothesize that NADH produced by *Hs*_ALDH8A1 is immediately reoxidized to NAD^+ during the *Hs*_ASPDH-dependent 2-aminomuconate reduction. Additionally, the absence of the 2-AM characteristic absorbance peak at 330 nm supports that, when it is released from *Hs*_ALDH8A1, it is rapidly converted into 2-OA by *Hs*_ASPDH.

The initial phase of the *Hs*_ASPDH reaction can be appreciated by following the 2-AM consumption at the fixed wavelength of 330 nm; after the addition of the enzyme, the signal is rapidly abolished due to the reduction of 2-AM to 2-OA and the concomitant NADH oxidation.

We finally tested the *Hs*_ASPDH ability to release an ammonia molecule from 2-AM. In the presence of the products of *Hs*_ASPDH and the substrates of the glutamate dehydrogenases (GDH), we followed the oxidation of NADH at 340 nm recorded after the addition of GDH (Fig.23D). The decrease in absorbance due to NADH oxidation is consistent with the synthesis and reduction of L-glutamate in the presence of alpha-ketoglutarate and the ammonia released during the *Hs*_ASPDH activity. Despite the evidence of the presence of ammonia in the reaction mixture containing *Hs*_ASPDH, this enzymatic assay is not able to discriminate whether ammonia is released before or after the redox reaction.

However, since (I) the signal at 330 nm after *Hs*_ASPDH activity is the result of the absorption of both NADH and 2-AM, (II) the final reaction mixture contains a low percentage of side products, and (III) the commercial GDH contains residual ammonia, it is necessary to optimize and to quantify all the involved compounds to correctly evaluate initial rates of the *Hs*_ASPDH reaction and its kinetics parameters.

*Hs*_ASPDH does not catalyze the aspartate dehydrogenase reaction

Since *Hs*_ASPDH has maintained the NAD^+ binding pocket and the putative active site is not dramatically different from the ancestral one, we investigated whether it is able to catalyze the L-aspartate oxidation to imino-aspartate with the concomitant NAD^+ reduction to NADH (Fig.24A).

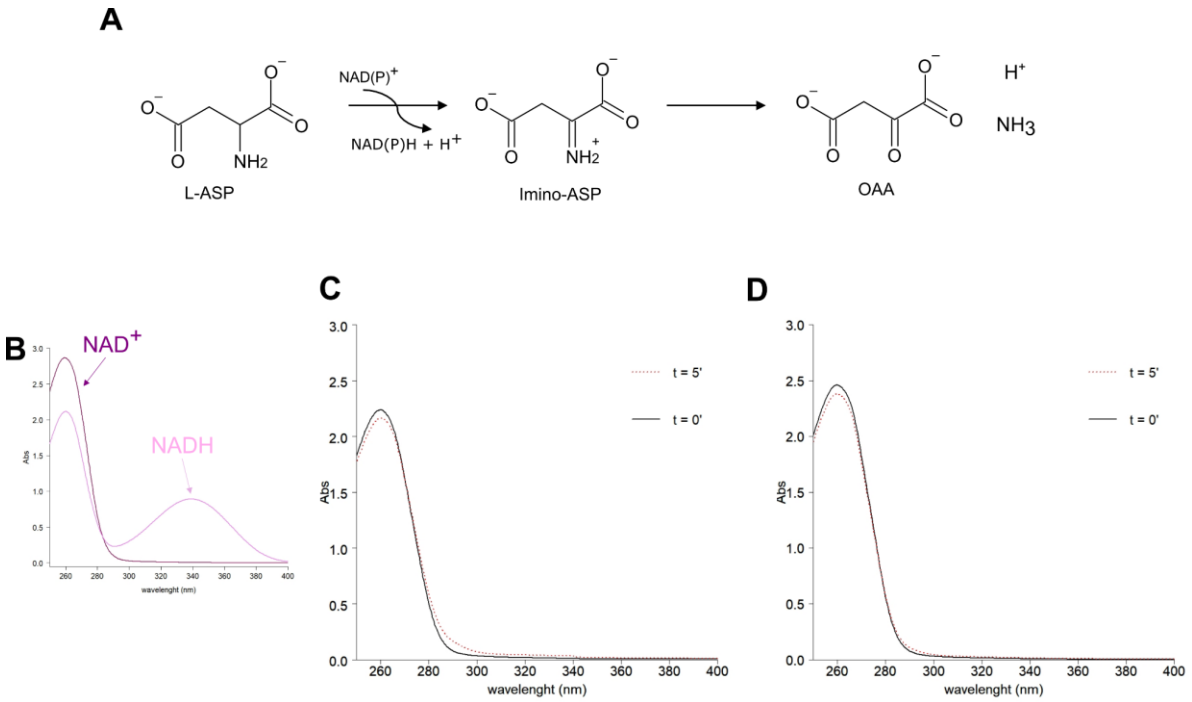


Figure 24: *Hs*_ASPDH has lost the ancestral enzymatic activity. (A) Aspartate dehydrogenases L-Asp oxidation to imino-Asp, followed by its conversion into OAA after deamination. (B) Comparison between NAD^+ and NADH spectra in the wavelength range between 250 nm and 400 nm. (C) L-aspartate dehydrogenase assay in presence of 0.8 mM L-Asp and 0.25 mM $NADP^+$ before (solid black line) and after (dotted red line) the addition of *Hs*_ASPDH, at pH =7.6. (D) L-aspartate dehydrogenase assay in presence of 0.8 mM L-Asp and 0.25 mM NAD^+ before (solid black line) and after (dotted red line) the addition of *Hs*_ASPDH, at pH =7.6.

Since *K.pneumoniae* L-aspartate dehydrogenase is highly specific for $NADP^+$ and, on the contrary, *T. maritima* and *A. fulgidus* can utilize NAD^+ as well as $NADP^+$, we assayed *Hs*_ASPDH L-asp dehydrogenase activity in the presence of either NAD^+ or $NADP^+$ as cofactor.

NAD and NADH absorbance spectra differ because of the presence of the characteristic peak at 340 nm for NADH (Fig.24B). We followed the L-aspartate activity in the presence of the oxidized cofactors and L-asp monitoring variations in the absorbance spectrum in the 250-400 nm range: we did not observe any variations in five minutes after the addition of *Hs*_ASPDH in the reaction mixture, both with NAD^+ and $NADP^+$ (Fig.24C, 24D). This lack of enzymatic activity for *Hs*_ASPDH is consistent with the loss of L-aspartate dehydrogenase activity and of the NAD^+ synthesis starting from L-asp in *Homo sapiens*.

Evolutionary and functional divergence of ASPDH and ASPDH-like proteins

With the aim to clarify the evolutionary history of the observed changes in *ASPDH-like* gene functions with respect to aspartate dehydrogenases, we investigated the distribution of the correspondent genes in eukaryotes and prokaryotes across a phylogenetic reconstruction.

We estimated an unrooted phylogenetic tree of ASPDH-like and ASPDH sequences with the maximum likelihood clustering method (Fig.25) and observed that sequences known to have ASPDH activity in bacteria grouped separately from metazoan sequences. *ASPDH* genes have been found mainly in some Bacteroidetes and Firmicutes, and in Proteobacteria. Other bacterial groups own a different gene that substitute *ASPDH* genes or have a similar function in the pathway, i.e. the aspartate oxidase in *E. coli*; plants and fungi do not possess *ASPDH* gene copies.

We may hypothesize an ancestral horizontal transfer from bacteria to nematodes, followed by the maintenance of the new gene copy in metazoan. Invertebrate sequences (Fig.25, light blue) form a separate cluster respect of the vertebrate sequences (Fig.25, blue). This evidence could be suggestive (I) of the acquisition of a new function for *ASPDH* genes in metazoan species and (II) of the repositioning of the latter in metabolism.

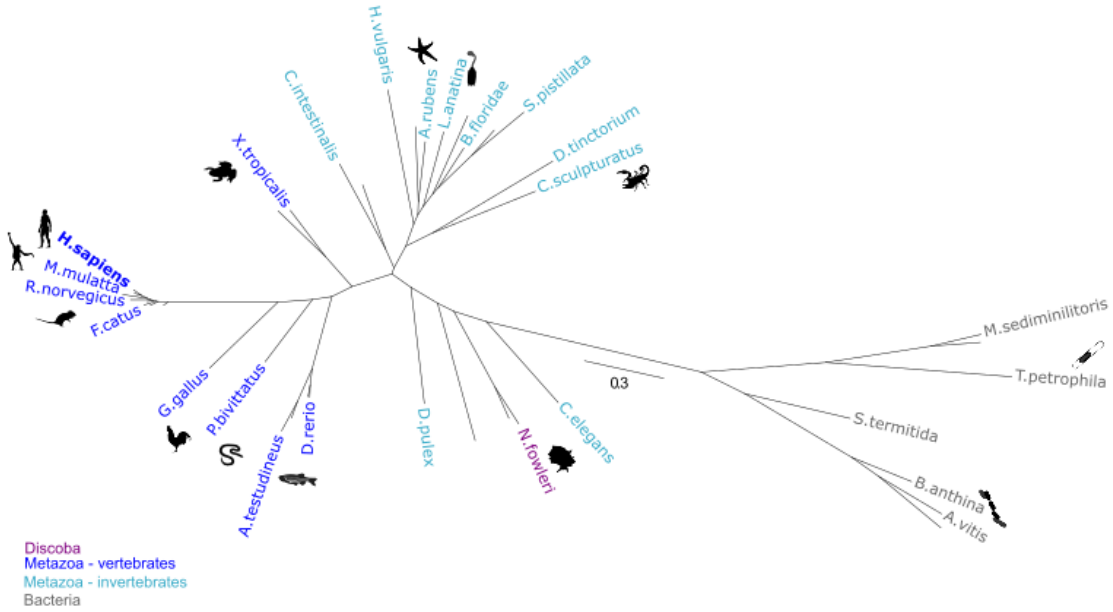


Figure 25: Phylogeny of ASPDH and ASPDH-like sequences. Unrooted tree built with Maximum Likelihood estimation using the LG model; the scale bar corresponds to the number of calculated substitutions per site. Selected organisms are labeled with the abbreviated species name and are colored according to taxonomy as indicated in the legend.

Features of human *ASPDH* isoform 2

As mentioned before, human *ASPDH* gene can produce two different splicing isoforms. The main one has been thoroughly described in this dissertation, but experimental and computational evidence collected in biological databases support the transcription of a shorter isoform.

The isoform 2 contains a distinct 5' UTR and lacks several in-frame portions of the 5' coding region, compared to variant 1, and has a shorter distinct N terminus (Fig.26).

We performed a homology search in tBlastN using EST database and we found multiple translated cDNA sequences that corresponds to human *ASPDH* isoform 2 (Tab.3); in particular, isoform 2 appears to be expressed mainly in neural system (hypothalamus, brain, thymus) and in human retina. No additional information is available on Human Protein Atlas databases since, to confirm the presence of human ASPDH proteins, experiments have been carried out using a rabbit antibody, thus excluding the shorter isoform because these organisms lack it.

ID	Tissue
EL736833.1	human retina
BI602385	hypotalamus
BU729215	human retina
HY149961.1	brain
HY102106.1	thymus
TO8135.1	brain

Table 3: Human cDNAs of *Hs*_ASPDH2 and their tissue expression. Table containing human cDNA IDs corresponding to isoform 2 and relative tissues.

We then confirmed the presence of a shorter isoform in Catarrhini (*Pan* spp., *H. sapiens*, *Pongo* spp., *T. gelada*) and Platyrrhini (*M. mulatta*) by performing a homology search with BlastP and a multiple sequences alignment (Fig.26); conversely, only the 283 amino acid long isoform is transcripted and translated in prosimians.

From the sequence alignment of the isoforms 1 (Fig.26, group 1) and isoform 2 (Fig.26, group 2) deriving from different primates, the absence of two extended portions at the N-terminal of group 2 is noticeable and corresponds to three skipped exons (2nd, 3rd and 5th). In addition, sequences of group 2 present an additional portion that corresponds to the alternative *ASPDH* 5’ UTR and transcription start site.

Furthermore, the amino acid sequences of both ASPDH isoforms are highly conserved in primates (Figure 26, red columns) probably due to their essential functional role. However, the

detection of synonymous substitution, i.e. in position 62, 168, 245 of the multiple alignment, is in agreement with the isoform 2 encoding and traduction into a functional protein.

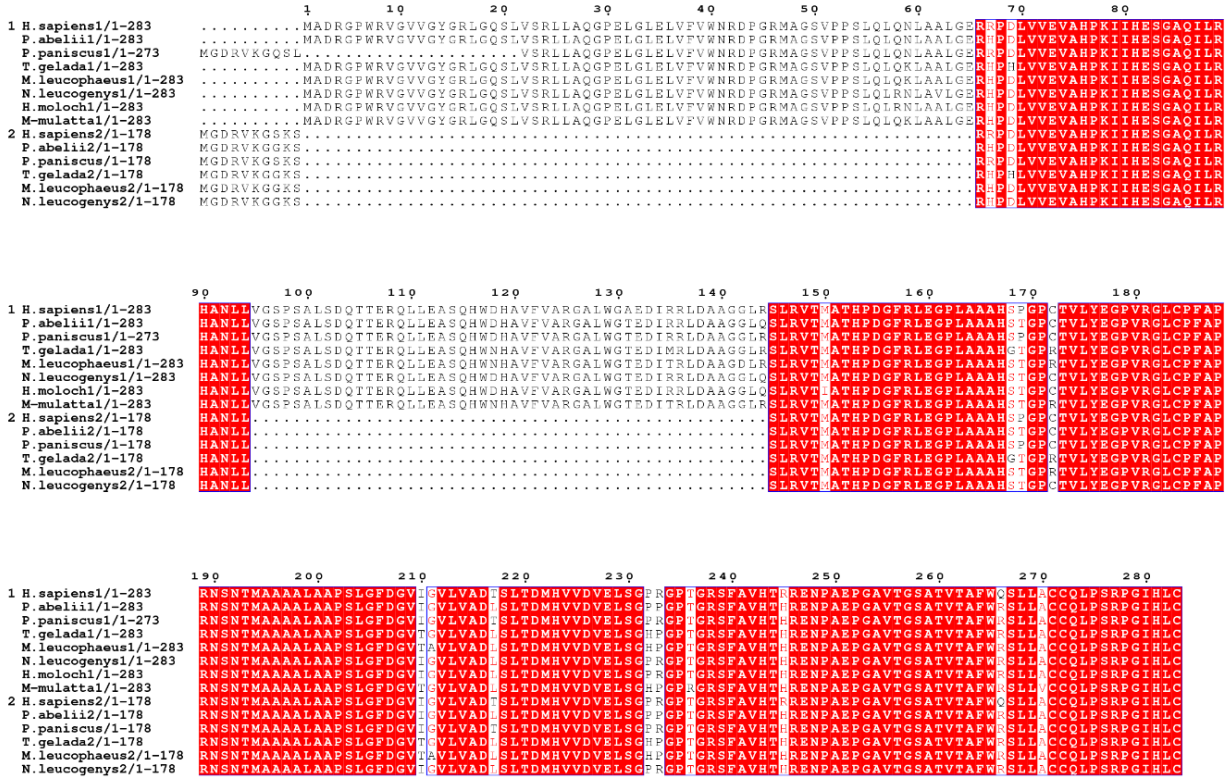


Figure 26: Multiple alignment of ASPDH isoform 1 and 2 in monkeys. Multiple alignment of ASPDH-like isoform 1 (group 1) and ASPDH-like isoform 2 sequences (group 2) from primates.

ASPDH isoform 2 structure and sequence analysis

To gain insights into the function of the shorter ASPDH isoform (*Hs*_ASPDH2), a multiple alignment of the two human isoform with *A. fulgidus* L-aspartate deydrogenase was performed and visualized, and the conservation of residues with a role in the NAD^+ binding was investigated.

The lack of an extended sequence portion at the *Hs*_ASPDH2 N-terminus determines the loss of half of the amino acids involved in the cofactor coordination (Fig.27A, blue triangles) unlike for the main isoform. The loss of the Rossmann fold domain can be corroborated by comparing 2DC1 3D structure and the *Hs*_ASPDH2 structural model built with Swiss Model using 2DC1 as template. We observed the almost complete loss of the series of alternating β-strand and α-helical segments characteristic of the Rossmann fold in *Hs*_ASPDH2, except for an α-helical element which is conserved at the C-terminal and encoded by an exon maintained in both the

isoforms (Fig.27B), and the presence of an additional short alpha-eliches followed by a beta-strand at the N-terminal encoded by the additional exons of *Hs*_ASPDH2.

We may hypothesize that the C-terminal helices is more probably placed next and parallel to other helices with respect to its modeled position (Fig.27B, see violet arrow); this model has been similarly predicted for the partial model of *Hs*_ASPDH2 (AF-M0QZ80-F1) by Alphafold, which have shaped an overall globular structure formed by three propellers resting on the beta sheet.

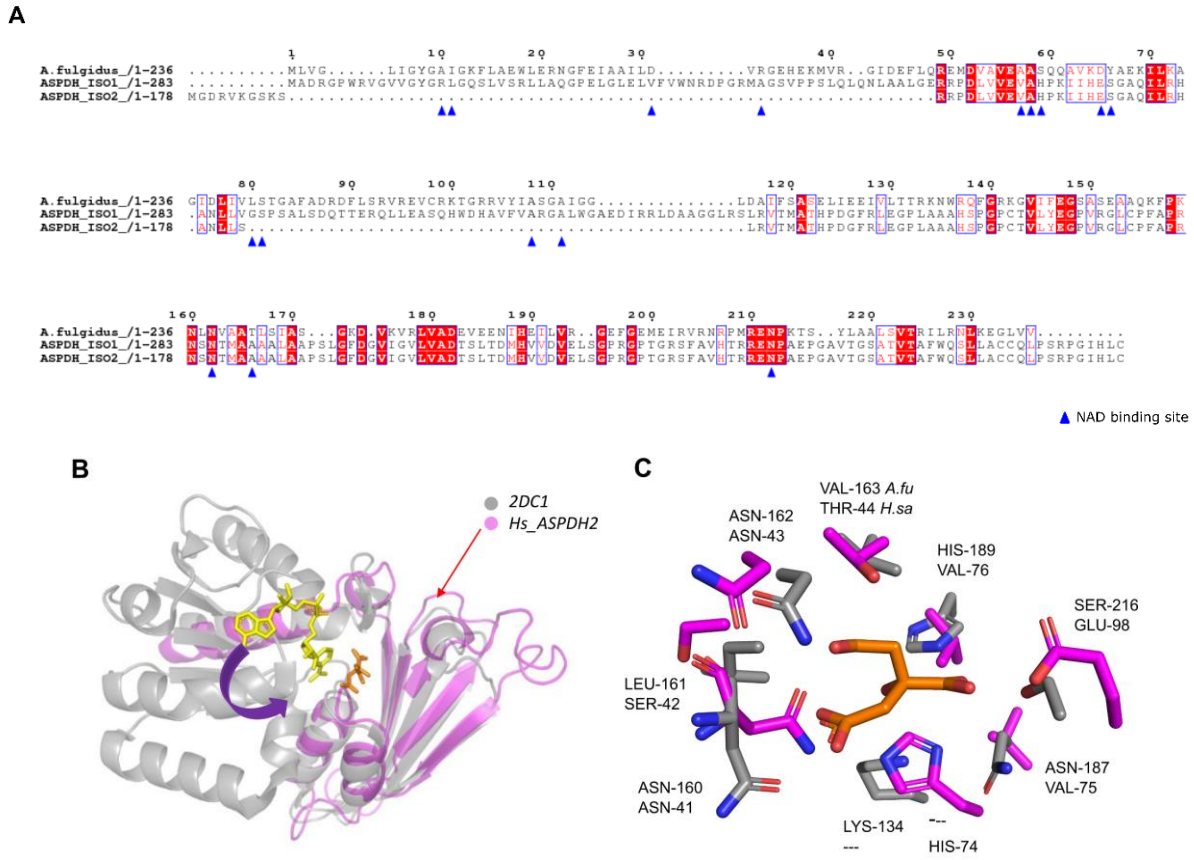


Figure 27: Structure and sequence analysis of human ASPDH shorter isoform. (A) Multiple alignment of isoform 1 and isoform 2 of human *ASPDH* gene with the *A. fulgidus* experimentally validated L-aspartate dehydrogenase. (B) Superimposition of *Hs*_ASPDH2 tridimensional model (purple cartoon) with 2DC1 structure (grey cartoon) in the presence of the substrate and the NAD cofactor, depicted respectively as yellow and orange sticks. (C) Close-up comparison between substrate binding sites showing the substitution of the totality of residues involved. Relevant residues are represented as CPK-colored sticks.

Considering both the structural and sequence comparison with *Af*_ASPDH and *Hs*_ASPDH, we can firmly state that *Hs*_ASPDH2 has lost the Rossmann fold domain and, consequently, its ability to bind the NAD cofactor essential to catalyze oxidoreductase reactions.

Since the substrate binding cavity appears to have been maintained, we compared the substrate surrounding amino acids with the aim to understand if *Hs*_ASPDH2 is able to accommodate L-aspartate or a chemical compound with similar characteristics. Except for two asparagine residues, all the amino acids responsible for the substrate binding described for *Af*_ASPDH protein are mutated into residues with different biochemical features in *Hs*_ASPDH2 protein (Fig.27C); in addition, the substrate binding cavity appear to be shifted, forming a larger empty pocket located in front of the surface and accessible from a tunnel that connect the functional buried cavities with the surface (Fig.27B, red arrow).

*Hs*_ASPDH2 gene expression and induction

We transformed the recombinant pET-28a(+) vector containing the *H. sapiens* isoform 2 of the "aspartate dehydrogenase domain-containing protein" sequence (NM_001024656.3) plus an N-terminal His-Tag in the bacterial strain *E.coli* BL21c+.

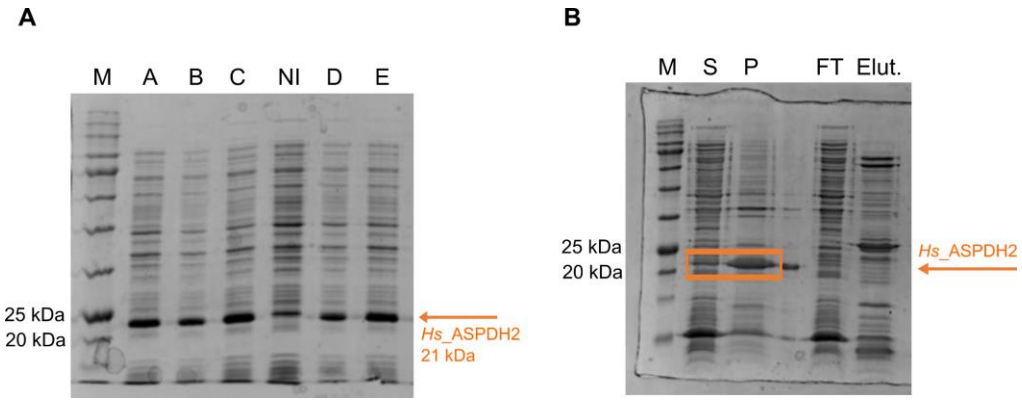


Figure 28: *Hs*_ASPDH2 gene expression and solubility tests. (A) SDS-PAGE (12%) *Hs*_ASPDH induction: M - marker; lane A - 30 °C, 3 h, 1 mM IPTG; lane B - 20 °C, o/n, 0.5 mM IPTG; lane C - 20 °C, o/n, 1 mM IPTG; lane D - 37 °C, 1 h, 1 mM IPTG; lane E - 30 °C, 3 h, 0.5 mM IPTG; N.I.: non induced. (C) SDS-PAGE (15%) of *Hs*_ASPDH expression and purification through FPLC: M, marker; S, supernatant; P, pellet; FT, flow-through; Elut, elution.

Several tests were performed on cell cultures by combining different inductor concentrations, growth temperatures (37, 30, and 20 °C) and durations (1, 3, and 16 hours) to understand what conditions were most suitable to obtain the protein expression and induction (Fig.28). Best induction was achieved by growing transformed cells at 30 °C for 3 hours, after the addition of 0.5 mM IPTG as confirmed by loading protein samples on SDS PAGE electrophoretic gel (Fig.28A). Attempts to purify protein using chromatographic techniques have failed since the recombinant protein expressed using this protocol appears to be insoluble (Fig.28B).

Conclusions

Orthogroup construction has been shown to be a determinant of the results of co-evolutionary analysis, as demonstrated by the identification of the gene that encodes for for the enzyme involved in the conversion of 2-aminomuconate (2-AM) to 2-ketoadipate (2-KA) in eukaryotic tryptophan catabolism. At variance with the method described in Chapter 2, the concordant transitions analysis between phylogenetic profiles were carried out by considering human-centered orthogroups formed after homology searches using DIAMOND software.

Our first phylogenetic profiling method did not allow us to identify a gene candidate for the missing 2-aminomuconate reductase in the tryptophan catabolism since the corresponding orthologous genes were splitted in two complementary OrthoDB orthogroups.

ASPDH_HUMAN, named after the human *ASPDH* gene, emerged as the best candidate to suit 2-aminomuconate reductase features; in Metazoa, it shares indeed similar phylogenetic profiles with other genes involved in the same pathway (AL8A1_HUMAN, ACMSD_HUMAN, 3HAO_HUMAN), with whom shares significant *cotr_scores* and *p-values*. In addition, *ASPDH* human gene shares a comparable tissue expression with the gene of the tryptophan catabolism, in particular with ACMSD and ALDH8A1 that specifically catalyzes the previous reactions in kidney and liver.

ASPDH shares homology with archaeal and bacterial aspartate dehydrogenases that take part in NAD *de novo* biosynthesis; based on *in silico* structural and sequence investigations, we have concluded that *Hs*_ASPDH has maintained the NAD binding pocket with some structural modifications and, probably, its oxidoreductase activity.

We validated our computational prediction through the development of an experimental strategy to obtain the substrate, the 2-aminomuconate, as the product of enzymatic conversions from 3-hydroxyanthranilate, and we expressed and purified in a recombinant form the three *Hs_ASPDH* upstream enzymes (*Hs*_HAAO, *Hs*_ACMSD, *Hs*_ALDH8A1). We spectrophotometrically followed the three consecutive enzymatic reactions in the presence of 3HAA and we confirmed the formation and disappearance of the signals corresponding to the different chemical species involved, some of which cyclize spontaneously over time. We studied the conditions of gene expression and induction of the recombinant *Hs*_ASPDH in the *E. coli* system, and we purified the protein through affinity chromatography.

To confirm the putative enzymatic activity, we monitored the enzymatic production of 2-AM starting from 3-HAA and we observed a rapid decrease in the absorbance at 330 nm after the

addition of *Hs*_ASPDH consistent with the consumption of 2-AM . We also demonstrated the presence of ammonia in solution after *Hs*_ASPDH by performing a coupled assay with GDH, but we did not determine whether deamination occurs before or after substrate reduction.

Taken together, all these lines of evidence confirm the identification of the missing human 2-aminomuconate (deaminating) reductase.

Further studies will focus on *Hs*_ASPDH structural and functional features. Because of the moderate solubility of *Hs*_ASPDH, enhancements of the experimental procedures could avoid protein precipitation in inclusion bodies and could increase the amount of protein purified, enabling its crystallization.

Several experimental investigations will be required to biochemically characterize *Hs*_ASPDH; it would be necessary to evaluate kinetic parameters and to determine Michaelis-Menten constants, and characterize the reaction product with mass spectrometry following similar procedure to what was done for the *Hs*_ALDH8A1 reaction product[30].

To assess biochemical parameters, we need to purify the 2-aminomuconate from the reaction mixture and identify conditions to store it by preventing its spontaneous deamination during time; it could be useful to apply the reaction mixture (containing 2-AM, NADH and the three enzyme) to a positive charged Ion Exchange spin column which can bind the negative-charged 2-AM. Attention should be paid to the working pH since it has been shown that 2-AM tends to deaminate more rapidly in an acidic environment[82].

A similar procedure will be useful to purify and to characterize with mass spectrometry the probable *Hs*_ASPDH reaction product, the 2-oxoadipate, which possesses two negative charged carboxylic groups.

The *ASPDH* human gene also produces a smaller isoform which is expressed only in the brain and has a different length and domain composition. Preliminary studies of the *Hs*_ASPDH2 protein have led us to hypothesize that this shorter isoform lacks the ability to bind the NAD cofactor and, as consequence, has lost is oxidoreductase catalytic activity; additionally, residues involved in the putative substrate binding are mutated with respect to both *Af*_ASPDH and *Hs*_ASPDH.

Our computational transcript analysis confirms that *ASPDH* encodes also a shorter protein in brain tissues; however, attempts to express it in a recombinant form in *E. coli* have failed. If expressed following experimental procedure similar to *Hs*_ASPDH expression, the protein accumulates in the insoluble fraction.

To obtain a stable protein, it could be useful to optimize gene expression and induction by substituting the IPTG inductor with lactose in the growth medium and/or modifying the temperature and the duration of the protein synthesis. An additional approach could be the coexpression of the construct with GroEL and GroES chaperones to assist the folding of the protein. As a further possibility a different eukaryotic host system could be used, such as *Pichia pastoris* or *Saccharomyces cerevisiae*.

The study of this shorter isoform could be relevant, as its expression in the brain could be related to the serotonin pathway, which is produced by tryptophan and has a key role in different psychiatric and neurological disorders.

Materials and Methods

Orthogroups and phylogenetic profile construction

The pipeline for phylogenetic profiling and for pairwise comparisons has been described in Material and Methods of Chapter 2.

Human-centered orthogroups were constructed using 1258 proteomes; homology searches were performed with DIAMOND software[75] using human gene as the query and setting a significant E value 10^{-3}. DIAMOND was used as a high-throughput aligner that is multiple orders of magnitude faster than BLAST. The identified genes from other organisms that were best hits (BH) of the human gene were included in the orthology group. Novel orthogroups were named after Uniprot Entry names corresponding to human genes.

Data were parsed with Perl scripts to generate gene tables of presence and absence. Cotransitions enumerations were used to build phylogenetic profiles for each human-centered orthogroups, considering '1' for the presence and '0' for the absence of a gene in orthogroups (rows) in a list of species (columns) ordered according to phylogeny. Sets of column positions with present->absent ('-1') and absent->present ('1') transitions were determined for each orthogroup using a script in Python.

The *cotr_score* (cotransition score) was calculated with an R script as *cotr_score* = k / (t1+t2-abs(k)), as described in Chapter 2. The significance *p-value* was calculated through the one-tailed Fisher's exact test.

"L-aspartate dehydrogenase like" genes from birds lacking protein signals were identified by performing a homology search in tBlastN using *Gallus gallus* "putative L-aspartate dehydrogenase" (XP_040512953.1), the tsa (Transcriptome Shotgun Assembly) database, an expected threshold = $1e^{-3}$, and limiting the search to birds (taxid:8782).

Sequence and structure analysis of "L-aspartate dehydrogenase" and "L-aspartate dehydrogenase-like" proteins

For sequence and phylogenetic analysis, human "aspartate dehydrogenase domain-containing protein" isoform 1 (NP_001108070.1, *Hs*_ASPDH) and isoform 2 (NP_001019827.2, *Hs*_ASPDH2), and "L-aspartate dehydrogenase-like" proteins were downloaded from OrthoDB v.10.1[83] and used for homology search with BlastP.

Multiple alignments of amino acid sequences were performed using ClustalX 2.0[84], modified with Jalview 2.11.2[85], and displayed with graphical enhancements using ESPript 3.0[86].

The experimental structure of "L-aspartate dehydrogenase" (PDB ID:2DC1)[65] from *Archaeoglobus fulgidus* was taken as a reference to identify residues with a relevant role in the function and structure of proteins included in the analysis. 2DC1 was also used as the template to perform the structural modeling of *Hs*_ASPDH and *Hs*_ASPDH2 using SWISS-MODEL[87]; both models were structurally aligned to 2DC1 and visualized with PyMol 2.5[88].

The phylogenetic tree was inferred with the maximum likelihood estimation using the "Le Gascuel" (LG) model selected by the automatic procedure in PhyML 3.0[89] and visualized with a radial layout and unrooted with FigTree 1.4.

Sequence and structural analysis images were decorated with animal silhouettes from PhyloPic (http://phylopic.org/) and modified using Inkscape 1.1.

Gene expression data for *Hs_ASPDH* and other genes of tryptophan catabolism were downloaded from the publicly available HPA RNA-seq sequencing project at the NCBI, containing RNA-seq data on tissue samples from 95 human individuals in a representative set of 27 major human organs and tissues[90]. The expression data of human genes have been displayed and compared using the *heatmap()* function in R: data have been organized in a color map where each column represents a human tissue and each row represents a gene. Color shades are proportional to the gene expression considering all tissues for each gene.

A tBlastN homology search was performed in the EST database to confirm cDNA expression for *Hs_ASPDH* in human and mammal tissues. Expression data were also inspected in the Human Protein Atlas database (https://www.proteinatlas.org/).

Chemical and physical protein parameters, useful to optimize purification protocol, were calculated using the ProtParam tool provided by ExPasy[91] and structural domain was identified using PFam database.

*Hs*_HAAO, *Hs*_ACMSD, *Hs*_ALDH8A1: vector construction, protein expression and purification

Human "3-hydroxyanthranilate 3,4-dioxygenase" (*Hs*_HAAO) CDS sequence (NM_012205), "amino-carboxymuconate semialdehyde decarboxylase" (*Hs*_ACMSD) CDS sequence (NM_001307983.2), and "aldehyde dehydrogenase 8 family member A1" CDS sequence (NM_022568.4) cloned into pET-28a(+) plasmids were purchased from GenScript Biotech; these expression vectors carry an N-terminal l 6xHisTag followed by a thrombin site to facilitate its removal.

*Hs*_HAAO, *Hs*_ACMSD, and *Hs*_ALDH8A1 isoelectric point (pI), extinction coefficients (in units of M^{-1} cm^{-1}, at 280 nm), and molecular mass were calculated with ProtParam considering the additional sequences of the protein expressed in recombinant form.

The constructs were electroporated into the bacterial host *Escherichia coli* BL21-CodonPlus DE3 strain (Novagen) and grown on LB agar medium added with antibiotics (50 µg/mL kanamycin for plasmid resistance, 34 µg/mL chloramphenicol for host resistance).

Single positive clones were added to LB broth (1% NaCl, 1% tryptone, 0.5% yeast extract) and gene expression was induced for 16 h at 20 °C with the addition of 0.5 mM IPTG (isopropylthio-β-galactoside) or alternatively 0.05% glucose and 0.2% lactose during media preparation. After induction, cells were collected by centrifugation (8000 g, 15 min, 4 °C), and stored at -80 °C.

Pellets were resuspended in a proper lysis buffer containing NaH_2PO_4 50 mM, NaCl 300 mM, glycerol 5% at pH 7.8 in the presence of 1 mM PMSF (serine protease inhibitor) and treated with 1 mg/mL lysozyme for 30 min. After the addition of 5 mM β-merc (beta-mercaptoethanol), the induced cells were sonicated (35-40 W, 1 s on - 1 s off, 15 min) and harvested for 45 minutes, 14000 rpm at 4 °C.

Overexpressed proteins were separated from the soluble fractions and purified on an FPLC system for affinity chromatography (Akta Pure 25 M, GE Healthcare) using a cobalt-charged column (HisTrap HP 5 mL) and taking advance of the 6xHisTag.

After the removal of contaminants with a washing buffer (50 mM Tris-HCl, 300 mM NaCl, 20 mM imidazole, pH = 8.0, 5 mM β-merc), proteins were eluted from the column with elution buffer (same as washing, plus 500 mM imidazole). A VivaSpin™ protein concentrator (Cytiva) with a suitable cutoff (10-30-50 kDa) was used to concentrate proteins (50 mM Tris-HCl, 300 mM NaCl, pH 7.8) and remove imidazole and other contaminants. Protein fractions were

loaded on 12% SDS-PAGE gel to confirm the correct protein size. The amount of the proteins purified were spectrophotometrically quantified at 280 nm with Lambert-Beer law.

Proteins were stored at -80 °C with 10% glycerol; HAAO was also added with 1 mM $FeNH_4SO_4$ and 1 mM glutathione[79] to avoid loss of activity after freezing.

The first-order equation S[t] ~ S0*e(-k * t) was used to evaluate values of the initial concentration S_0 and the slope k for *Hs*_ACMSD.

*Hs*_ASPDH and *Hs*_ASPDH2: vector construction, protein expression and purification

Human "aspartate dehydrogenase domain-containing - isoform 1" CDS sequence (NM_001114598.2) cloned into a modified pET-28a(+) plasmid was purchased from GenScript Biotech (Fig.29); this expression vector carries an N-terminal 6xHisTag followed by a thrombin site to facilitate its removal. The plasmid also contains the araBAD promoter[92] of *E.coli* L-arabinose operon for a thigh control of the protein; this approach allows regulation of carbon sources and is ideal for the expression and the solubility optimization of toxic proteins.

Human "aspartate dehydrogenase domain-containing - isoform 2" CDS sequence (NM_001024656.3) cloned into pET-28a(+)-TEV plasmid was purchased from GenScript Biotech; this expression vector carries a 6xHisTag and a TEV protease recognition site at the N terminus.

Both the constructs were electroporated into the bacterial host *Escherichia coli* BL21-CodonPlus DE3 strain (Novagen) and grown on LB agar medium added with antibiotics (50 µg/mL kanamycin for plasmid resistance, 34 µg/mL chloramphenicol for host resistance); single positive clones were added to LB broth (1% NaCl, 1% tryptone, 0.5% yeast extract).

Gene expression and induction were optimized at 3 h at 30 °C after the addition of 0.5 mM IPTG (isopropylthio-β-galactoside) and 0.5 mg/mL L-arabinose; the best conditions were found after multiple attempts by modifying inductor concentrations, temperature and duration of induction. After the induction, cells were collected by centrifugation (8000 g, 15 min, 4 °C), and stored at -80 °C before usage.

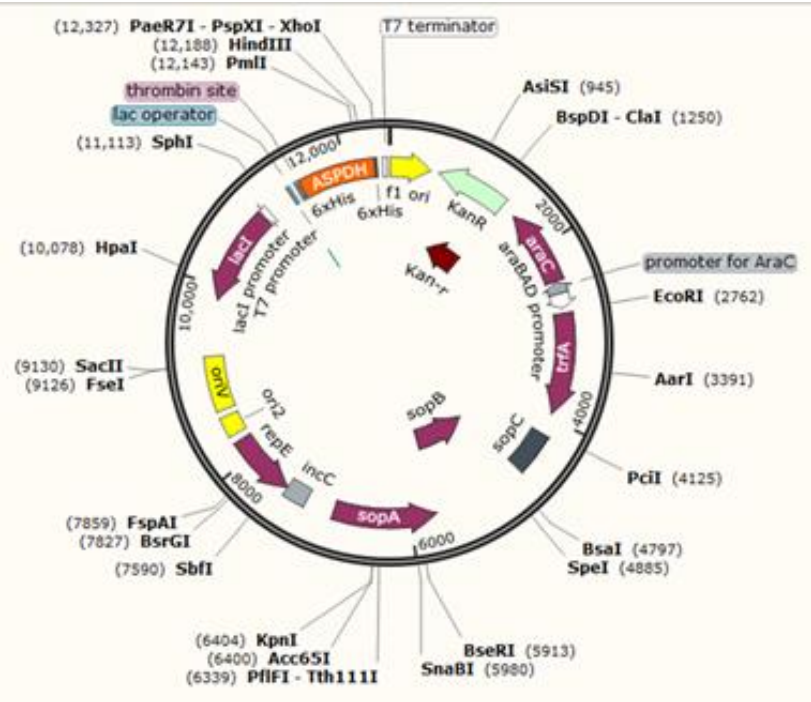


Figure 29. *Hs*_ASPDH isoform 1 expression vector. Modified pET-28(a)+ vector contains: (I) AraC promoter and lac operator suitable for gene induction with IPTG and L-arabinose; (II) a multiple cloning site containing the *ASPDH* CDS sequence cloned in frame with a N-terminal 6xHis Tag and a thrombin site; (III) antibiotic resistance to kanamycin.

For both constructs, pellets were resuspended in a proper volume of lysis buffer containing NaH_2PO_4 50 mM, NaCl 300 mM, glycerol 5% at pH 7.8 in the presence of 1 mM PMSF, and treated with 1 mg/mL lysozyme for 30 min. After the addition of 5 mM β-merc, the induced cells were sonicated (35-40 W, 3 s on - 6 s off, 10 min) and harvested for 45 minutes, 14000 rpm at 4 °C.

Overexpressed proteins were separated from the soluble fractions and purified on an FPLC system for affinity chromatography (Akta Pure 25 M, GE Healthcare) using a cobalt-charged column (HisTrap HP 5 mL), taking advance of the 6xHisTag.

After the removal of contaminants with a washing buffer (50 mM Tris-HCl, 300 mM NaCl, 20 mM imidazole, pH = 8.0, 5 mM β-merc), proteins were eluted from the column with an elution buffer (same as washing, plus 500 mM imidazole). A VivaSpin™ protein concentrator (Cytiva) with a suitable cutoff (10 kDa) was used to concentrate proteins (50 mM Tris-HCl, 300 mM NaCl, pH 7.8) and to remove imidazole and other contaminants. Protein fractions were loaded on 12% or 15% SDS-PAGE gel to confirm the correct protein size. The amount of the purified proteins were spectrophotometrically quantified at 280 nm with Lambert-Beer law.

Proteins were stored at -80°C in the presence or in the absence of 10% glycerol.

Activity assay with UV-visible spectrophotometry

Activity assays for all the enzymes tested were performed by following the evolution of absorbance spectra in the spectral range from 250 nm to 450 nm or at a fixed wavelength, depending on the reaction and on the chemical species involved. All spectra were acquired in Spectra Measurement and Time course measurement modes using a JASCO V-750 UV-Visible Spectrophotometer equipped with a thermostat and placing the reaction mixture in a black quartz cuvette and maintaining the temperature at 25 °C.

The reaction mixtures to assay the freshly purified *Hs*_HAAO activity contained 50 µM 3HAA and *Hs*_HAAO 50 mM in KH_2PO_4/Tris-HCl at pH 7.6.

The reaction mixtures to assay the freshly purified *Hs*_ACMSD activity contained 50 µM 3HAA, *Hs*_HAAO and *Hs*_ACMSD in 50 mM KH_2PO_4/Tris-HCl at pH 7.6.

*Hs*_HAAO and *Hs*_ACMSD activities were also monitored at the fixed wavelength of 360 nm to confirm ACMS formation and to follow ACMS spontaneous decay into QUIN or its enzymatic conversion into AMS.

The reaction mixtures to assay the freshly purified *Hs*_ALDH8A1 activity contained 75 µM 3HAA, 150 µM NAD^+, *Hs*_HAAO, *Hs*_ACMSD, and *Hs*_ALDH8A1 in 50 mM KH_2PO_4/Tris-HCl at pH 7.6. *Hs*_ALDH8A1 activity was also followed at the fixed wavelength of 330 nm to confirm 2-AM formation and its spontaneous deamination.

The reaction mixtures to assay the freshly purified *Hs*_ASPDH activity contained 75 µM 3HAA, 150 µM NAD^+, *Hs*_HAAO, *Hs*_ACMSD, *Hs*_ALDH8A1 and *Hs*_ASPDH in 50 mM KH_2PO_4/Tris-HCl at pH 7.6. *Hs*_ASPDH activity was also followed at the fixed wavelength of 330 nm to confirm 2-AM consumption.

*Hs*_ASPDH was also assayed for L-aspartate dehydrogenase activity by following NAD^+ reduction and monitoring the evolution of absorbance spectra in the spectral range from 250 nm to 400 nm. The reaction mixture contained 2 mM L-aspartate, 0.25 mM NAD^+ or 0.25 mM $NADP^+$, 5 µM *Hs*_ASPDH in 20 mM KH_2PO_4 at pH 7.6.

The *Hs*_ASPDH-dependent release of ammonia from 2-AM was confirmed with a continuous coupled assay at the fixed wavelength of 340 nm in the presence of glutamate dehydrogenase (GDH from the bovine liver; Sigma-Aldrich).

The GDH coupled assay was started at the end of *Hs*_ASPDH assay in the presence of GDH substrates: the 250 µL reaction mixtures contained *Hs*_HAAO, *Hs*_ACMSD, *Hs*_ALDH8A1 and

*Hs*_ASPDH, with their final reaction product obtained starting from 60 µM 3HAA and 150 µM NAD^+. 1 mM ⍺-ketoglutarate, 0.25 µM NADH in 20 mM pH 7.6 potassium phosphate were added to the ended-reaction to enable the GDH-dependent NADH oxidation to NAD^+ as consequence of L-glutamate formation from alpha-ketoglutarate and ammonia; 4 U GDH were added to start the reaction.

Data analysis

Experimental data were analyzed, processed, and graphically represented using packages in R version 4.1.2 and RStudio and aesthetically modified with Inkscape 1.1.
Chemical and enzymatic reactions were drawn using ChemSketch.

Bibliography

1. Berger, F. The new life of a centenarian: signalling functions of NAD(P). *Trends Biochem. Sci.* 29, 111–118 (2004).
2. Luengo, A. *et al.* Increased demand for NAD+ relative to ATP drives aerobic glycolysis. *Mol. Cell* 81, 691-707.e6 (2021).
3. Nakamura, M., Bhatnagar, A. & Sadoshima, J. Overview of *Pyridine Nucleotides* Review Series. *Circ. Res.* 111, 604–610 (2012).
4. Yu, P., Cai, X., Liang, Y., Wang, M. & Yang, W. Roles of NAD+ and Its Metabolites Regulated Calcium Channels in Cancer. *Molecules* 25, 4826 (2020).
5. Zapata-Pérez, R. *et al.* Reduced nicotinamide mononucleotide is a new and potent NAD $^+$ precursor in mammalian cells and mice. *FASEB J.* 35, (2021).
6. Navas, L. E. & Carnero, A. NAD+ metabolism, stemness, the immune response, and cancer. *Signal Transduct. Target. Ther.* 6, 2 (2021).
7. Pollak, N., Dölle, C. & Ziegler, M. The power to reduce: pyridine nucleotides – small molecules with a multitude of functions. *Biochem. J.* 402, 205–218 (2007).
8. Imai, S. & Guarente, L. NAD+ and sirtuins in aging and disease. *Trends Cell Biol.* 24, 464–471 (2014).
9. Pollard, C.-L., Gibb, Z., Swegen, A. & Grupen, C. G. NAD$^+$, Sirtuins and PARPs: enhancing oocyte developmental competence. *J. Reprod. Dev.* 2022–052 (2022) doi:10.1262/jrd.2022-052.
10. Wo, Y. J. *et al.* The Roles of CD38 and CD157 in the Solid Tumor Microenvironment and Cancer Immunotherapy. *Cells* 9, 26 (2019).
11. Quarona, V. *et al.* CD38 and CD157: A long journey from activation markers to multifunctional molecules: CD38 and CD157. *Cytometry B Clin. Cytom.* 84B, 207–217 (2013).
12. Carles Canto´, Keir J. Menzies, and Johan Auwerx. NAD+ Metabolism and the Control of Energy Homeostasis: A Balancing Act between Mitochondria and the Nucleus. 31–53 (2015).
13. Nakahata, Y. & Bessho, Y. The Circadian NAD $^+$ Metabolism: Impact on Chromatin Remodeling and Aging. *BioMed Res. Int.* 2016, 1–7 (2016).
14. Covarrubias, A. J., Perrone, R., Grozio, A. & Verdin, E. NAD+ metabolism and its roles in cellular processes during ageing. *Nat. Rev. Mol. Cell Biol.* 22, 119–141 (2021).
15. Badawy, A. A.-B. Kynurenine Pathway of Tryptophan Metabolism: Regulatory and Functional Aspects. *Int. J. Tryptophan Res.* 10, 117864691769193 (2017).
16. Nakagawa, T., Lomb, D. J., Haigis, M. C. & Guarente, L. SIRT5 Deacetylates Carbamoyl Phosphate Synthetase 1 and Regulates the Urea Cycle. *Cell* 137, 560–570 (2009).
17. Fjeld, C. C., Birdsong, W. T. & Goodman, R. H. Differential binding of NAD $^+$ and NADH allows the transcriptional corepressor carboxyl-terminal binding protein to serve as a metabolic sensor. *Proc. Natl. Acad. Sci.* 100, 9202–9207 (2003).
18. Fons, N. R. *et al.* PPM1D mutations silence NAPRT gene expression and confer NAMPT inhibitor sensitivity in glioma. *Nat. Commun.* 10, 3790 (2019).
19. Cameron, A. M. *et al.* Inflammatory macrophage dependence on NAD+ salvage is a

consequence of reactive oxygen species–mediated DNA damage. *Nat. Immunol.* 20, 420–432 (2019).
20. Li, M. *et al.* Local Targeting of NAD+ Salvage Pathway Alters the Immune Tumor Microenvironment and Enhances Checkpoint Immunotherapy in Glioblastoma. *Cancer Res.* 80, 5024–5034 (2020).
21. Altschul, R., Hoffer, A. & Stephen, J. D. Influence of nicotinic acid on serum cholesterol in man. *Arch. Biochem. Biophys.* 54, 558–559 (1955).
22. Anderson, R. M., Bitterman, K. J., Wood, J. G., Medvedik, O. & Sinclair, D. A. Nicotinamide and PNC1 govern lifespan extension by calorie restriction in Saccharomyces cerevisiae. *Nature* 423, 181–185 (2003).
23. Colabroy, K. L. & Begley, T. P. The Pyridine Ring of NAD Is Formed by a Nonenzymatic Pericyclic Reaction. *J. Am. Chem. Soc.* 127, 840–841 (2005).
24. Huo, L. *et al.* Human α-amino-β-carboxymuconate-ε-semialdehyde decarboxylase (ACMSD): A structural and mechanistic unveiling: Substrate Positioning in Human ACMSD. *Proteins Struct. Funct. Bioinforma.* 83, 178–187 (2015).
25. Pucci, L., Perozzi, S., Cimadamore, F., Orsomando, G. & Raffaelli, N. Tissue expression and biochemical characterization of human 2-amino 3-carboxymuconate 6-semialdehyde decarboxylase, a key enzyme in tryptophan catabolism: Human ACMSD. *FEBS J.* 274, 827–840 (2007).
26. Yang, Y., Davis, I., Matsui, T., Rubalcava, I. & Liu, A. Quaternary structure of α-amino-β-carboxymuconate-ε-semialdehyde decarboxylase (ACMSD) controls its activity. *J. Biol. Chem.* 294, 11609–11621 (2019).
27. Martynowski, D. *et al.* Crystal Structure of α-Amino-β-carboxymuconate-ε-semialdehyde Decarboxylase: Insight into the Active Site and Catalytic Mechanism of a Novel Decarboxylation Reaction . *Biochemistry* 45, 10412–10421 (2006).
28. Martí-Massó, J. F. *et al.* The ACMSD gene, involved in tryptophan metabolism, is mutated in a family with cortical myoclonus, epilepsy, and parkinsonism. *J. Mol. Med.* 91, 1399–1406 (2013).
29. Lin, M. & Napoli, J. L. cDNA Cloning and Expression of a Human Aldehyde Dehydrogenase (ALDH) Active with 9-cis-Retinal and Identification of a Rat Ortholog, ALDH12. *J. Biol. Chem.* 275, 40106–40112 (2000).
30. Davis, I., Yang, Y., Wherritt, D. & Liu, A. Reassignment of the human aldehyde dehydrogenase ALDH8A1 (ALDH12) to the kynurenine pathway in tryptophan catabolism. *J. Biol. Chem.* 293, 9594–9603 (2018).
31. S A Reading. Chromium picolinate. 29–31 (1996).
32. He, Z. & Spain, J. C. A Novel 2-Aminomuconate Deaminase in the Nitrobenzene Degradation Pathway of *Pseudomonas pseudoalcaligenes* JS45. *J. Bacteriol.* 180, 2502–2506 (1998).
33. Nishizuka, Y., Ichiyama, A. & Hayaishi, O. [58] Metabolism of the benzene ring of tryptophan (mammals). in *Methods in Enzymology* vol. 17 463–491 (Elsevier, 1970).
34. Schwarcz, R., Bruno, J. P., Muchowski, P. J. & Wu, H.-Q. Kynurenines in the mammalian brain: when physiology meets pathology. *Nat. Rev. Neurosci.* 13, 465–477 (2012).

35. Parsons, C. G. *et al.* Novel systemically active antagonists of the glycine site of the N-methyl-D-aspartate receptor: electrophysiological, biochemical and behavioral characterization. *J. Pharmacol. Exp. Ther.* 283, 1264–1275 (1997).
36. Foster, A. C., Vezzani, A., French, E. D. & Schwarcz, R. Kynurenic acid blocks neurotoxicity and seizures induced in rats by the related brain metabolite quinolinic acid. *Neurosci. Lett.* 48, 273–278 (1984).
37. Giles, G. I., Collins, C. A., Stone, T. W. & Jacob, C. Electrochemical and in vitro evaluation of the redox-properties of kynurenine species. *Biochem. Biophys. Res. Commun.* 300, 719–724 (2003).
38. Rios, C. & Santamaria, A. Quinolinic acid is a potent lipid peroxidant in rat brain homogenates. *Neurochem. Res.* 16, 1139–1143 (1991).
39. Schwarcz, R., Whetsell, W. O. & Mangano, R. M. Quinolinic Acid: An Endogenous Metabolite That Produces Axon-Sparing Lesions in Rat Brain. *Science* 219, 316–318 (1983).
40. Simon, R. P., Swan, J. H., Griffiths, T. & Meldrum, B. S. Blockade of *N* -Methyl-D-Aspartate Receptors May Protect Against Ischemic Damage in the Brain. *Science* 226, 850–852 (1984).
41. Ala, M. The footprint of kynurenine pathway in every cancer: a new target for chemotherapy. *Eur. J. Pharmacol.* 896, 173921 (2021).
42. Abd El-Fattah, E. E. IDO/kynurenine pathway in cancer: possible therapeutic approaches. *J. Transl. Med.* 20, 347 (2022).
43. Stone, T. W., Forrest, C. M. & Darlington, L. G. Kynurenine pathway inhibition as a therapeutic strategy for neuroprotection: Kynurenines and neuronal viability. *FEBS J.* 279, 1386–1397 (2012).
44. Sas, K., Szabó, E. & Vécsei, L. Mitochondria, Oxidative Stress and the Kynurenine System, with a Focus on Ageing and Neuroprotection. *Molecules* 23, 191 (2018).
45. Mondanelli, G. & Volpi, C. The double life of serotonin metabolites: in the mood for joining neuronal and immune systems. *Curr. Opin. Immunol.* 70, 1–6 (2021).
46. Shajib, Md. S., Baranov, A. & Khan, W. I. Diverse Effects of Gut-Derived Serotonin in Intestinal Inflammation. *ACS Chem. Neurosci.* 8, 920–931 (2017).
47. Walther, D. J. *et al.* Serotonylation of Small GTPases Is a Signal Transduction Pathway that Triggers Platelet α-Granule Release. *Cell* 115, 851–862 (2003).
48. Rudnick, G. & Sandtner, W. Serotonin transport in the 21st century. *J. Gen. Physiol.* 151, 1248–1264 (2019).
49. Oxenkrug, G. F. Metabolic syndrome, age-associated neuroendocrine disorders, and dysregulation of tryptophan-kynurenine metabolism: Metabolic syndrome and tryptophan-kynurenine metabolism. *Ann. N. Y. Acad. Sci.* 1199, 1–14 (2010).
50. Bender, D. A. & McCREANOR, G. M. Kynurenine hydroxylase: a potential rate-limiting enzyme in tryptophan metabolism. *Biochem. Soc. Trans.* 13, 441–443 (1985).
51. Savitz, J. The kynurenine pathway: a finger in every pie. *Mol. Psychiatry* 25, 131–147 (2020).
52. Cervenka, I., Agudelo, L. Z. & Ruas, J. L. Kynurenines: Tryptophan's metabolites in exercise, inflammation, and mental health. *Science* 357, eaaf9794 (2017).
53. Miura, H. *et al.* A link between stress and depression: Shifts in the balance between the

kynurenine and serotonin pathways of tryptophan metabolism and the etiology and pathophysiology of depression. *Stress* 11, 198–209 (2008).
54. Correia, A. S. & Vale, N. Tryptophan Metabolism in Depression: A Narrative Review with a Focus on Serotonin and Kynurenine Pathways. *Int. J. Mol. Sci.* 23, 8493 (2022).
55. Liu, J.-J., Movassat, J. & Portha, B. Emerging role for kynurenines in metabolic pathologies. *Curr. Opin. Clin. Nutr. Metab. Care* 22, 82–90 (2019).
56. Stasi, C., Sadalla, S. & Milani, S. The Relationship Between the Serotonin Metabolism, Gut-Microbiota and the Gut-Brain Axis. *Curr. Drug Metab.* 20, 646–655 (2019).
57. Katoh, A., Uenohara, K., Akita, M. & Hashimoto, T. Early Steps in the Biosynthesis of NAD in Arabidopsis Start with Aspartate and Occur in the Plastid. *Plant Physiol.* 141, 851–857 (2006).
58. Yang, Z. *et al.* Aspartate Dehydrogenase, a Novel Enzyme Identified from Structural and Functional Studies of TM1643. *J. Biol. Chem.* 278, 8804–8808 (2003).
59. Sakuraba, H., Tsuge, H., Yoneda, K., Katunuma, N. & Ohshima, T. Crystal Structure of the NAD Biosynthetic Enzyme Quinolinate Synthase. *J. Biol. Chem.* 280, 26645–26648 (2005).
60. Hunt, L., Lerner, F. & Ziegler, M. NAD – new roles in signalling and gene regulation in plants. *New Phytol.* 163, 31–44 (2004).
61. Smith, E. N., Schwarzländer, M., Ratcliffe, R. G. & Kruger, N. J. Shining a light on NAD- and NADP-based metabolism in plants. *Trends Plant Sci.* 26, 1072–1086 (2021).
62. Gakière, B. *et al.* NAD $^+$ Biosynthesis and Signaling in Plants. *Crit. Rev. Plant Sci.* 37, 259–307 (2018).
63. Marinoni, I. *et al.* Characterization of l-aspartate oxidase and quinolinate synthase from Bacillus subtilis: NadA and NadB from B. subtilis. *FEBS J.* 275, 5090–5107 (2008).
64. Teramoto, H., Suda, M., Inui, M. & Yukawa, H. Regulation of the Expression of Genes Involved in NAD *De Novo* Biosynthesis in *Corynebacterium glutamicum*. *Appl. Environ. Microbiol.* 76, 5488–5495 (2010).
65. Yoneda, K. *et al.* The first archaeal l-aspartate dehydrogenase from the hyperthermophile Archaeoglobus fulgidus: Gene cloning and enzymological characterization. *Biochim. Biophys. Acta BBA - Proteins Proteomics* 1764, 1087–1093 (2006).
66. Wogulis, M., Chew, E. R., Donohoue, P. D. & Wilson, D. K. Identification of Formyl Kynurenine Formamidase and Kynurenine Aminotransferase from *Saccharomyces cerevisiae* Using Crystallographic, Bioinformatic and Biochemical Evidence. *Biochemistry* 47, 1608–1621 (2008).
67. Bedalov, A., Hirao, M., Posakony, J., Nelson, M. & Simon, J. A. NAD $^+$ -Dependent Deacetylase Hst1p Controls Biosynthesis and Cellular NAD $^+$ Levels in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 23, 7044–7054 (2003).
68. Croft, T., Venkatakrishnan, P. & Lin, S.-J. NAD+ Metabolism and Regulation: Lessons From Yeast. *Biomolecules* 10, 330 (2020).
69. Panozzo, C. *et al.* Aerobic and anaerobic NAD $^+$ metabolism in *Saccharomyces cerevisiae*. *FEBS Lett.* 517, 97–102 (2002).
70. Gazzaniga, F., Stebbins, R., Chang, S. Z., McPeek, M. A. & Brenner, C. Microbial NAD Metabolism: Lessons from Comparative Genomics. *Microbiol. Mol. Biol. Rev.* 73, 529–541

(2009).
71. Nasu, S., Wicks, F. D. & Gholson, R. K. The mammalian enzyme which replaces b protein of e. coli quinolinate synthetase is d-aspartate oxidase. *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.* 704, 240–252 (1982).
72. Li, Y. *et al.* A novel l-aspartate dehydrogenase from the mesophilic bacterium Pseudomonas aeruginosa PAO1: molecular characterization and application for l-aspartate production. *Appl. Microbiol. Biotechnol.* 90, 1953–1962 (2011).
73. Li, Y., Ogola, H. J. O. & Sawa, Y. l-Aspartate dehydrogenase: features and applications. *Appl. Microbiol. Biotechnol.* 93, 503–516 (2012).
74. He, X., Kang, Y. & Chen, L. Identification of ASPDH as a novel NAADP-binding protein. *Biochem. Biophys. Res. Commun.* 621, 168–175 (2022).
75. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60 (2015).
76. Huttener, R. *et al.* Sequencing refractory regions in bird genomes are hotspots for accelerated protein evolution. *BMC Ecol. Evol.* 21, 176 (2021).
77. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* 51, D445–D451 (2023).
78. Toledo-Patiño, S., Pascarelli, S., Uechi, G. & Laurino, P. Insertions and deletions mediated functional divergence of Rossmann fold enzymes. *Proc. Natl. Acad. Sci.* 119, e2207965119 (2022).
79. Mitchell Robert, Kang, H. H. & Henderson, L. M. Inactivation during functioning of 3-hydroxyanthranilate oxidase resulting from oxidation of bound ferrous iron. *The Journal of Biological Chemistry* (1963).
80. Brkić, H., Kovačević, B. & Tomić, S. Human 3-hydroxyanthranilate 3,4-dioxygenase (3HAO) dynamics and reaction, a multilevel computational study. *Mol. Biosyst.* 11, 898–907 (2015).
81. Li, T., Ma, J. K., Hosler, J. P., Davidson, V. L. & Liu, A. Detection of Transient Intermediates in the Metal-Dependent Nonoxidative Decarboxylation Catalyzed by α-Amino-β-Carboxymuconate-ε-Semialdehyde Decarboxylase. *J. Am. Chem. Soc.* 129, 9278–9279 (2007).
82. He, Z. & Spain, J. C. Preparation of 2-aminomuconate from 2-aminophenol by coupled enzymatic dioxygenation and dehydrogenation reactions. *J. Ind. Microbiol. Biotechnol.* 23, 138–142 (1999).
83. Kriventseva, E. V. *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 47, D807–D811 (2019).
84. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948 (2007).
85. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinforma. Oxf. Engl.* 25, 1189–1191 (2009).
86. Gouet, P. ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.* 31, 3320–3323 (2003).

87. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303 (2018).
88. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
89. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* 59, 307–321 (2010).
90. Fagerberg, L. *et al.* Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Mol. Cell. Proteomics* 13, 397–406 (2014).
91. Duvaud, S. *et al.* Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res.* 49, W216–W227 (2021).
92. Guzman, L. M., Belin, D., Carson, M. J. & Beckwith, J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* 177, 4121–4130 (1995).