Universal system of gene-enzyme nomenclature in the genome era applied to aromatic biosynthesis

Roy A Jensen[1*], Carol A Bonner[2], and Jian Song[3]
Emerson Hall, University of Florida, PO Box 14425, Gainesville, Florida, 32604

Corresponding author
Emerson Hall, University of Florida, PO Box 14425, Gainesville, Florida, 32604 USA. E-mail: rjensen@ufl.edu, Telephone: 352-475-3019

Abstract (< 250 words)

Background:

The accurate annotation of functional roles for newly sequenced genes of genomes is not a simple matter. Function is, of course, related to amino-acid sequence and to domain structure – but not always in straightforward ways. Even where given functional roles have been identified experimentally, the application of an uneven and erratic nomenclature has generated confusion on the part of annotators and has produced errors that tend to become progressively compounded in database repositories.

Results:

The pathway that is deployed in nature for aromatic biosynthesis exemplifies an accumulation of chaotic nomenclature and a variety of annotation dilemmas. We view this pathway as one that is sufficiently complex to pose most of the common problems, and yet is one that at the same time is of a manageable size. A set of guidelines has been developed for naming genes of aromatic-pathway biosynthesis and the corresponding gene products, and these can be generalized for application to other metabolic subsystems.

Conclusion:

It is urgent that a universal system of gene-enzyme nomenclature be implemented. A system of nomenclature for aromatic biosynthesis is presented that is logical, consistent, and evolutionarily informative.

Background

Dilemma

The prime objective of genome annotation is to relate gene sequences to functional roles. The functional roles are attached to the acronyms that comprise the annotation nomenclature. Unfortunately, a logical and universal convention for acronym assignment has never been adopted, and nomenclature ambiguities are well embedded in the contemporary literature. Examples include cases where, as a matter of idiosyncratic choice, the same genes from different organisms have been named differently. Alternatively, sometimes entirely different genes carry the same name. In addition, fused genes have often been assigned a single acronym prior to the realization that single genes in other genomes encode the separate domain components of the original multi-domain fusion. Not surprisingly, this has happened often with fused genes in *Escherichia coli*, an organism studied intensively prior to availability of massive comparative genomic information. The gene-fusion dilemma has generated a particularly large number of annotation mistakes, e.g., where single stand-alone genes have been annotated the same as the wrong fusion member of a two-domain gene. So what to name the newly recognized unfused genes? Clearly, as an initial step of logical naming, some backtracking is in order so that acronym names can be applied at the level of the unfused genes encoding their uni-domain products. (Even with an ideal future of universal nomenclature in place, occasional adjustments might be needed since genes not previously known to be separable may from time to time be found to have divided counterparts in previously unstudied organisms.)

Although increasingly awkward, the various inconsistent and anomalous patterns of acronym usage have been tolerable, because the pool of available model organisms has been relatively small. In the recent past, most of those engaged with the use of particular gene names have, in fact, been the experts and scholars who developed most of the experimental knowledge base. In contrast, the majority of those who access and assign gene names now are annotators and non-experimentalists who work to assemble and interpret new genomes at global levels. In this era of genomics the proliferation of pet acronyms is an increasingly less affordable luxury, and, in fact, has become a nightmare that simply begs for resolution. One might ask whether contemporary repositories of genome annotations, as they exist, are already too formidable for backtracking and re-annotation to be feasible. These already-huge

repositories are nevertheless only a fraction of what can be anticipated in the near future, and therefore, it seems inevitable, indeed urgent, that a universal nomenclature be implemented.

There is at least a general level of awareness among an increasingly sophisticated core of database users that they must navigate around the aforementioned problems. But an even more challenging dimension of the nomenclature dilemma stems from complicated (albeit intriguing) relationships between function and protein structure (see Brown *et al*. [1] and Orengo and Thornton [2] for a current perspective.) Intuitively, there is a natural tendency to suppose that identical functions of proteins must coincide with very similar structures. However, at one extreme, a single functional role may have been captured by unrelated proteins (analogs), i.e., the functional capability evolved independently more than once (convergence). At the other extreme, a single homology group of proteins may be populated with members that possess varied functional roles, e.g., different substrate specificities. Recent reports [3, 4] teach that substrate specificity may vary, sometimes with surprising ease, across a given protein family, but against a backdrop where the basic reaction chemistry deployed within the family is inevitably conserved very strongly. Thus, nomenclature assignments can be complicated when guidance for annotation is sought from sequence similarity because completely different structures can evolve to support the same functional roles, yet very similar structures can support different functions.

Our group has been preoccupied with comparative aspects of aromatic metabolism for almost 40 years. Throughout the last decade we have given considerable thought to issues of nomenclature in relationship to what is a relatively large and complex metabolic system. We have implemented a continuum of suggestions [5-12], some of which have been further revised or even abandoned in favor of perceived improvements that are presented here. This has at least been a starting point of experience for an issue that is not trivial. Hopefully, our contribution here (which is implemented by a clickable visualization on our website at http://aropath.lanl.gov/Visualizations/AroPath/AroPath.htm) might provide a blueprint, or at least a sounding board, for the resolution of nomenclature issues at a broader and more global level.

Results and Discussion

Biochemical diversity

Figure 1 provides a visualization for the primary biochemical pathway of aromatic biosynthesis. This figure includes the known pathway deviations and indicates analog variations for enzyme steps that are functionally identical. Other representations of this pathway that are available for general use are neither up-to-date nor completely correct. (For example, the KEGG pathway (http://www.genome.ad.jp/kegg/pathway/map/map00400.html), although very widely used and blessed with some innovative features, contains over a dozen errors and omissions. It is only in fairly recent times that the realization has materialized that the pathway of aromatic amino acid biosynthesis is not universal in terms of the metabolite flow route deployed. Two major variations are currently known. **(i)** Occasionally aromatic biosynthesis begins with different steps (AroA' and AroB' instead of AroA and AroB). **(ii)** Alternative intermediates are formed and utilized between prephenate and *L*-phenylalanine, on the one hand, and between prephenate and *L*-tyrosine, on the other hand. Such alternative intermediates exemplify a phenomenon whereby a reversed order of reaction steps are embedded within multi-step pathways that begin and end with the same compounds and where the overall chemistry deployed is identical.

Even where the pathway flow of intermediary metabolites is managed by exactly the same reactions, a further aspect of diversity is that these reactions are not necessarily carried out in different organisms by homolog enzymes. Indeed, analog representatives of AroA, AroC, AroE, AroH, PheA, and TrpC are known to exist. In addition, distinct homolog divergence sometimes yields clearly separable sub-homolog groupings. Since appreciation of both the diversity of pathway flow routes and of the enzyme catalysts has been increasingly unmasked by the exponential increase in genome sampling, elucidation of additional diversity can certainly be anticipated.

p-Aminobenzoate (PABA) synthesis

Although our analysis does not yet extend to the vitamin-like branches, we include PABA synthase as an exception in Fig. 1 because it is so closely related to anthranilate synthase (the first step of *L*-tryptophan biosynthesis). PABA is important as a component of folate and is produced from chorismate. PABA synthase and anthranilate synthase both use identical

substrates (chorismate and glutamine) and both release pyruvate and glutamate as reaction products. The *p*-aminobenzoate and anthranilate (*o*-aminobenzoate) products differ only in the placement of the amino group on the ring. It is thus not surprising that the amino-acid sequences of the two subunits of anthranilate synthase (TrpAa and TrpAb) exhibit high identities to the PabAa and PabAb homologs, respectively. However, PABA synthase deploys a third subunit that has no counterpart in anthranilate synthase. The 2-amino-2-deoxy-isochorismate (ADIC) intermediate created in the anthranilate synthase reaction is evidently so unstable that the ADIC lyase reaction, which releases pyruvate, occurs spontaneously without enzymatic assistance. On the other hand, the 4-amino-4-deoxy-chorismate intermediate of the PABA synthase reaction is much more stable and requires ADC lyase (PabAc) to catalyze the final aromatization and release of pyruvate. The sequence of PabAc has no counterpart in the sequence of TrpAa. It is interesting that PabAc requires pyridoxal 5'-phosphate (PLP) and is a homolog of PLP-dependent branched-chain aminotransferases.

The three proteins of PABA biosynthesis are included in a clickable visualization at AroPath (http://www.aropath.lanl.gov/Visualizations/AroPath/AroPath.htm), a website having links to a detailed table that supplies representative query gi numbers which are hyperlinked to NCBI.

Functional roles viewed within a context of phylogenetic relationships

The correct inference of functional roles is frustrated, not only by the aforementioned uneven character of the relationships that tie functional roles to enzyme structures, but also by uncertainty about what boundaries of amino-acid identity will guarantee a common functional role [13, 14]. Indeed, at one extreme, enzymes that exhibit 98% amino-acid sequence identity catalyze different reactions [15]. In spite of these complications, we assert that functional roles can usually be superimposed to hierarchical evolutionary positions within a homologous protein series. But it may be challenging. Consider the case of $AroA_{I\alpha}$ and $AroA_{I\beta}$, which have the same function (DAHP synthase) and share phylogenetic space in the $AroA_I$ Superfamily, with both being qualitatively distinct from members of yet another group of DAHP synthases: the analog $AroA_{II}$ Superfamily (Fig. 2). At the hierarchical level of I_β proteins, members of $AroA_{I\beta}$ are joined by KdsA. (If KdsA, a single cohesive grouping,

were being discussed within a context of interest in its phylogenetic relationship with other proteins, it would be more precisely be called $KdsA_{I\beta}$ within our scheme.) KdsA proteins have a functional role (KDOP synthase) that differs from that of $AroA_{I\beta}$. In this case, KdsA and $AroA_{I\beta}$ (different function) share more overall sequence similarity than $AroA_{I\beta}$ and $AroA_{I\alpha}$ (same function). This illustrates a collection of different protein groups that could be susceptible to chaotic annotation, as indeed has been the case, However, once the hierarchical relationships have been discerned following the appropriate exploitation of experimental work, each group can then be appreciated to possess a functional role that is cohesively associated with a particular phylogenetic slice. In other words, the phylogenetic thread can usually be traced with respect to functional roles, even though the thread may have a meandering character. New sequence queries that are addressed to a correct background of annotation will generate the correct functional annotation [9].

In summary, Fig. 2 gives an example of four distinct groupings that sort out at different hierarchical levels of evolutionary relationship (indicated by the horizontal arrows). The three protein groupings designated at the bottom-left of Fig. 2 are homologs that gather under the Superfamily-I umbrella. Superfamily II, shown at the right, is probably of independent origin. The evolutionary scenario indicated is one that follows the recruitment hypothesis [16] whereby an ancestral 3-deoxy-ald-2-ulosonate phosphate synthase of broad specificity preceded the acquisition (via gene duplication and then differential substrate specialization) of the narrow-specificity enzymes seen in contemporary organisms. The DAHP synthase domain of *B. subtilis* has greater sequence similarity to the KDOP synthase of *E. coli* than to the DAHP synthase of *E. coli*. A **single** functional role, DAHP synthase (AroA), is represented at **multiple** (three) phylogenetic-node positions. But once the $AroA_{I\alpha}$, $AroA_{I\beta}$, and $AroA_{II}$ groups are pinpointed as discrete groupings within their hierarchical boundaries, the functional roles can be correctly fitted to a phylogenetic perspective. Given such insight, new query sequences can be expected to return results that place them within the phylogenetic space defined by the hierarchical boundaries.

Examples are given at the bottom of Fig. 2 of some sequences that carry relatively uncomplicated acronyms, i.e., Ctep $AroA_{1\alpha}$, Tmar $AroA_{I\beta}$, and Hpyl $AroA_{II}$. The additional examples chosen for inclusion on the far bottom line of Fig. 2 include some sequences having particular variant features that are captured by the acronym. Thus, *E. coli* $AroA_{I\alpha\_w}$ (see **rule viii** in following section) is a tryptophan-inhibited species of DAHP synthase, one of three

differentially regulated $AroA_{I\alpha}$ paralogs in that organism. The *B. subtilis* $AroA_{I\beta}$ domain is linked via an N-terminal fusion with a chorismate mutase of the Superfamily-I homology type (Table 5) to yield $AroH_I$•$AroA_{I\beta}$ (**rule i**). DAHP synthase proteins that belong to the Superfamily-II homology type exist in both bacteria and in higher-plant plastids. The higher-plant protein is designated *$AroA_{II}$ because it possesses a cleavable transit peptide (see **rule xi**).

Proposed nomenclature guidelines

(i) Gene products will have names that parallel the genes encoding them, e.g., *aroA* encodes AroA. (This does not prevent other descriptions of the gene product, e.g., AroA being the universal acronym proper for the enzyme known as DAHP synthase). Standard 4-letter gene and gene product acronyms are referred to as the '**acronym proper**'. Note that additional symbols and conventions that convey important information, as described herein, are appended to the acronym proper.

(ii) Genes in a pathway or pathway segment are named in the order of the reactions catalyzed by the gene products, e.g., AroA, AroB and AroC, which catalyze the first three reactions of aromatic biosynthesis, are encoded by *aroA, aroB, and aroC*.

(iii) The smallest unit of naming is at the level of discrete catalytic or allosteric domains. Multi-domain fusions are designated with intervening bullets, e.g., *tyrA•aroF* encodes fused catalytic domains that abound elsewhere as stand-alone *tyrA* and *aroF* domains. Note that a catalytic domain such as TyrA is actually a "supradomain" consisting of an N-terminal cofactor domain and a C-terminal catalytic domain, and potentially these could be named separately. In fact, the cofactor domain [17] is widely distributed in combination with domains other than the TyrA catalytic domain. However, the separate two domains of TyrA are thus far not known to possess any functional roles as independent entities. TyrA catalytic domains always coexist with the cofactor domain to form an intimate functional unit, and the functional site may very well be created between the two domains [18]. In this case, the TyrA name is currently applied to the supradomain, so named as the smallest functional unit at the present time. Hence, in this context of function, we sometimes apply "supradomain" as an equivalent of "domain" in those cases where the smallest functional unit appears to be the supradomain.

Identical functional roles will be associated with identical acronym-proper labels, regardless of whether the gene products are homologs or analogs. On the other hand, note that it is occasionally possible for enzymes catalyzing different reactions to carry the same acronym proper if they are embedded in the same overall metabolic conversion (see **rule x**).

(iv) If an enzyme consists of subunits, the corresponding gene and gene-product names are designated with additional lower-case letters, e.g., the anthranilate synthase complex consists of the large aminase subunit TrpAa and the small amidotransferase subunit TrpAb, these being encoded by *trpAa* and *trpAb*, respectively.

(v) Distinct allosteric domains are designated with 3 capital letters. One example is *pheA•ACT* encoding PheA•ACT.

(vi) Different homology classes (analogs) that have independently acquired the same function are designated with Roman-numeral subscripts, e.g., $aroA_{I}$ and $aroA_{II}$ encode analogs that catalyze the same reaction. The latter exemplifies a case where, on structural grounds, the apparent analogs could possibly be distant homologs that diverged sufficiently to mask definitive recognition of the homology (given the limitation of current resources). However, we do not infer homology if it cannot be proven.

If a homology class consists of distinct, well-separated subgroups, additional lower-case Greek subscripts can be appended to designate them, as illustrated by $AroA_{I\alpha}$ and $AroA_{I\beta}$ (Fig. 2). If there were no known analogs, then any well-separated sub-homolog groups would be designated without Roman-numeral subscripts, as is exemplified by $TyrA_{\alpha}$ and $TyrA_{\beta}$.

(vii) If different enzyme reactions converge upon a common intermediate as is the case in early aromatic biosynthesis, genes within one of the convergent branches are designated with a 'prime'. Usually this would apply to the least widely distributed branch. Thus, AroA and AroB, on the one hand, and AroA' and AroB', on the other hand, describe different initial routes that converge to provide exactly the same product (dehydroquinate) to AroC for use as its substrate [19]. Thus, AroA and AroA' each catalyze the first committed step of aromatic biosynthesis in different organisms, but the particular reactions catalyzed are not the same. And the same is true of AroB and AroB'.

Our subsystem coverage does not yet include the large number of connecting links that will be added. Of these, the metabolic linkage to NAD biosynthesis via quinolinate comes to

mind because the alternative tryptophan-to-quinolinate and aspartate-to quinolinate pathways [20] will exemplify another instance of pathway convergence to a common intermediate.

(viii) Paralogs, which originated from <u>recent</u> gene duplications and which have no obvious differential functional specializations (one-function paralog family), are distinguished from one another with underscore numbers. Recent gene duplicates (e.g., *trpD*_1, *trpD*_2, and *trpD*_3) might have selective value via manifestation of a gene-dose effect, or they might include pseudogenes destined for elimination (apparently a common phenomenon). If one of the multiple paralogs seems to be uniquely suited to carry out the function corresponding to that of a well-characterized single-gene ortholog in organismal relatives, the preference would be to label it *trpD*_1. For example, such a scenario would apply in the situation (see [11]) where *trpD*_1 occupies a perfect and complete tryptophan operon in some cyanobacteria, whereas the extra-operonic *trpD*_2 and *trpD*_3 paralogs exhibit especially long branches on a protein tree and lack one or more amino-acid residues known to be important for catalysis (thus being likely pseudogenes). In comparisons of the same genes in a collection of organisms where some of the organisms support multiple paralogs and others do not, the single genes cannot properly be labeled the same as a particular paralog member present in a multi-paralog organism. Thus, for example, organisms with a single *trpD* gene would simply be denoted *trpD* since the latter has equally orthologous relationships with each of the recent paralogs present in sister organisms [21].

(ix) Same-function ancient paralogs that vary in some specialized feature carry appropriate underscore notations. Ancient paralogs arose from gene duplications that preceded speciation [21]. Ancient paralogs are usually differentially specialized, and those with different catalytic functions will carry names that reflect different pathway roles. (The ancient AroA and KdsA paralogs of Fig. 2 would be examples). However, occasional ancient paralogs have retained the same enzymatic function and hence share the same acronym proper, but they are differentially specialized in some other way. For example, the trio of paralogous DAHP synthases in enteric bacteria are $AroA_{I\alpha}$ proteins that are subject to differential regulation by feedback inhibition: $AroA_{I\alpha\_W}$ by tryptophan, $AroA_{I\alpha\_F}$ by phenylalanine, and $AroA_{I\alpha\_Y}$ by tyrosine.

Occasionally some member species of two-paralog lineages possess a single remnant paralog, which, in addition to its usual function, has acquired the function of the lost sister paralog, thus being bifunctional. In such cases **the name of the surviving paralog (identified by**

homology or by operon context) is given first and in bold fonts, and separated from the name of the missing paralog by a double 'slash'. Examples of such relatively rare bifunctional proteins covered in this article are **PabAb** // TrpAb and **AroJ$_{I\beta}$ //** HisG in a small clade of *Bacillus* species, as well as **HisD** // TrpC$_{II}$ in Actinomycete bacteria.

(x) Different substrate specificities of homologs typically support different functional roles in **different** pathways e.g., the aforementioned DAHP synthase/KDOP synthase dichotomy (Fig. 2). An exception is represented by phenomena where the order of enzymatic reaction steps in a multi-step pathway is variable. **If homologs having different substrate specificities are embedded within the flow route of the same pathway such that they perform equivalent functional roles at the overall pathway level, they will share the same acronym proper, with the differing specificities indicated with subscript identifiers.** This is exemplified by the alternative flow routes between prephenate and *L*-tyrosine in Fig. 1. The tyrosine-pathway dehydrogenases catalyze different reactions, being specific for prephenate, for *L*-arogenate, or able to utilize both. However, at the broader pathway level, the functional role of each of the three is identical. Namely, in each case the cyclohexadienyl substrate is aromatized via an oxidative reaction that is driven by elimination of the ring-attached carbon dioxide. The three variant specificities are indicated with lower-case, rightward subscripts: TyrA$_p$, TyrA$_a$, and TyrA$_c$, respectively.

If substrate ambiguity of such one-pathway homologs extends to a second substrate, specificity for a second substrate can be designated with leftward subscripts**,** e.g., $_{NAD}$TyrA$_p$ is a tyrosine-pathway dehydrogenase specific for the NAD^+/prephenate couple, whereas $_{NADP}$TyrA$_a$ refers to specificity for the $NADP^+$/arogenate couple.

(xi) Genes that encode cleavable signal (or transit) peptides are denoted by leading-asterisk superscripts. Thus, *aroH$_{I\alpha}$* encodes cytoplasmic chorismate mutase, whereas **aroH$_{I\alpha}$* encodes periplasmic (or secreted) chorismate mutase.

Overall rationale in support of the acronym scheme. The above nomenclature scheme is intended to relate the acronym library to the evolutionary thread. This is not absolutely necessary to the extent that the single critical need is to implement a consistent, universal assemblage of acronyms. However, a significant advantage of the system proposed is that a

given acronym is designed to convey a large amount of biochemical and evolutionary knowledge. Information is conveyed, not only by what is **present** in the acronym, but also by what is **absent**. For example, consider the hypothetical Xyz pathway in which one encounters the gene product $XyzC_{II}$•, encoded by *xyzC*$_{II}$•. Even being unfamiliar with the Xyz pathway, one knows (because of the 'C') that reference is being made to the third enzyme in the pathway. The Roman-numeral subscript reveals that this enzyme is one of at least two analog classes, and the bullet informs that there is a C-terminal fusion. $XyzC_{II}$• cannot be a subunit component; otherwise there would be a lowercase letter immediately after the acronym proper. There is no cleavable signal or transit peptide: otherwise there would be a leading asterisk. It is not a member of a one-function paralog family, otherwise we would see underscore notations. It is not a member of a homolog family that separates into distinct subgroups; otherwise an α, β, etc. would follow the Roman-numeral subscript. $XyzC_{II}$• has not expanded its functional repertoire by "borrowing" a second functional role that is exercised elsewhere in the lineage by a paralog relative; otherwise the acronym for the "borrowed" functional role of the lost paralog would be applied (with separation by a 'double slash') after that of the surviving paralog.

Nomenclature that anticipates future changes

Table 1 and Table 2 show the application of these nomenclature guidelines to the biosynthetic pathways for tryptophan and histidine, respectively. Both pathways exhibit some gene fusions in *Esherichia coli*, and the fusion designations are illustrated below the tables. Thus, in *E. coli*, the designation *trpAb*• indicates that the *trpAb* domain has a fusion linkage at the C-terminal end, whereas •*trpB* denotes a *trpB* domain that carries a fusion linkage at its N-terminus. Likewise, whereas *hisH* and *hisF* are free-standing genes in many organisms, they are fused in *E. coli* (hence being denoted there as *hisH*•*hisF*). Both pathways contain enzymes that deploy subunit components: anthranilate synthase (TrpAa and TrpAb), tryptophan synthase (TrpEa and TrpEb), and imidazole glycerol-P synthase (HisEa and HisEb). Each subunit possesses a distinct catalytic function that participates in a more complex overall reaction.

The tryptophan [22] and histidine [23] pathways have long enjoyed status as model pathways in model organisms, and it was generally assumed for many years that these pathways as summarized in Table 1 and Table 2 were universal. However, in the recent genomic era some

unexpected variations have surfaced, and more can be anticipated. Tables 3 and 4 (and the immediately following discussion) illustrate how the new variations can be accommodated to the rules of nomenclature suggested herein.

Tryptophan biosynthesis updated

In Actinomycete bacteria related to *Mycobacterium*, *trpC* (encoding phosphoribosyl-anthranilate isomerase) is absent from the genome. The function of the missing enzyme has in fact been captured by HisD, which also functions as the phosphoribosyl isomerase in the histidine pathway [24]. This breadth of isomerase specificity is unusual because contemporary HisD proteins are generally specific for the isomerase reaction in the histidine pathway. Since broadened isomerase specificity is seen to be achievable, it is tempting to wonder if HisD and TrpC are, in fact, homologs (i.e., paralogs that arose from an ancient common ancestor having broad substrate specificity). If so, this is masked by substantial divergence. With this finding in Actinomycete bacteria, the previously known classical *trpC* (of Table 1) becomes $trpC_I$ in Table 3, and the newly discovered *trpC* found in Actinomycetes is denoted ***hisD*** *//* $trpC_{II}$ (see **rule ix**).

In another example of paralog loss accompanied by the rescue of its function by a surviving sister paralog to yield a bifunctional enzyme, *Bacillus subtilis* and a small clade of close relatives possess a bifunctional ***pabAb*** *// trpAb* gene [25, 26]. Contemporary PabAb and TrpAb proteins are different-function homologs that are difficult to discriminate on the criterion of amino acid sequence because they fail to separate into two distinct and cohesive subclusters on a phylogenetic tree. Fortunately, however, *trpAb* can be presumed to be the lost paralog in the latter case because the surviving paralog gene is located in the *pab* operon [25]. Therefore, we name the surviving paralog gene ***pabAb*** *// trpAb*. (***pabAb*** appears first and in bold fonts because it is the surviving paralog, one that has additionally acquired the closely related function that is shown second.) It is interesting to consider that paralog loss and broadened substrate specificity of the surviving paralog may, in fact, be a reversal of the steps of gene duplication and divergent paralog specialization envisioned to be a general evolutionary scenario [16].

Finally, among prokaryotes we recently analyzed a distinct second subcluster of TrpEb that has a widespread, but erratic, distribution. This relatively rare species has been denoted

TrpEb_2 to distinguish it from the typical form, called TrpEb_1 [12]. According to the bottom section of **rule vi**, TrpEb_1 and TrpEb_2 can better be named $TrpEb_{\alpha}$ and $TrpEb_{\beta}$, respectively (Table 3). $TrpEb_{\beta}$ has been postulated to have two different functional roles. (**i**) In a few cases the variant TrpEb is the sole species of TrpEb present, and by default it appears likely that this variant TrpEb species ($TrpEb_{\beta}$) must indeed have a functional role as tryptophan synthase in these few organisms. It is particularly suggestive that the companion TrpEa subunits in these organisms form a distinct and cohesive phylogenetic cluster with one another on a protein tree, and only in exactly those organisms whose variant $TrpEb_{\beta}$ exists in the absence of the widespread $TrpEb_{\alpha}$. These variant $TrpEa_{\beta}$ and $TrpEb_{\beta}$ subunit pairs likely have evolved unique subunit-subunit contacts with one another [12] that differ from the classical ones established between $TrpEa_{\alpha}$ and $TrpEb_{\alpha}$ [27]. (**ii**) In other cases $TrpEb_{\beta}$ co-exists with $TrpEb_{\alpha}$. Since the latter is undoubtedly the competent functional subunit of tryptophan synthase, the function of the former is a matter of speculation. If the idea is affirmed that these $TrpEb_{\beta}$ proteins have captured the function of a lost homolog (serine deaminase) as proposed by Xie *et al*. [12], they would be renamed **$TrpEb_{\beta}$** // Sda_{II}. The primary serine deaminase elsewhere in the lineage (but missing in the organisms having both $TrpEb_{\alpha}$ and $TrpEb_{\beta}$ would then be named $SdaA_{I}$. If this were to prove correct, it would appear that one way (bioinformatic) to differentiate $TrpEb_{\beta}$ from **$TrpEb_{\beta}$** // Sda_{II} is that the former will coexist uniquely with the distinctive set of $TrpEa_{\beta}$ subunit partners.

Histidine biosynthesis updated

The first enzyme of histidine biosynthesis is a highly regulated phosphotransferase that is unrelated to other classes of phosphotransferases. It has recently been found that feedback inhibition is a accomplished in two very distinct ways. Some organisms have a "short" version of HisA (recently called $HisG_{S}$) that requires another subunit (recently called HisZ), a homolog of class II aminoacyl-tRNA synthetases, for both activity and feedback inhibition [28]. We name these subunits HisA and HAD_{I} (Table 4). HAD denotes 'histidine allosteric domain'. Since HAD_{I} does not appear to have a distinct catalytic role, **rule iv** is not invoked to name it HisAb in combination with HisA (which would then be named HisAa). HAD_{I} is probably required for catalytic activity because it stabilizes HisA.

In other organisms a "long" form of HisA exists (recently called $HisG_{L}$). This is the classical enzyme of such organisms as *E. coli*. $HisG_{S}$ and the N-terminal portion of $HisG_{L}$ are

homologous. Approximately the C-terminal 100 amino-acid residues of $HisG_L$ are unique and constitute a domain that is asserted to be an allosteric domain accomplishing feedback inhibition ([29] and references therein). We propose to name this protein HisA•HAD_{II}. Thus, HAD_I and HAD_{II} can be viewed as analogs that have converged to impart allostery to HisA: HAD_I via complex formation and HAD_{II} via gene fusion.

Prokaryotes that mobilize a HisG homolog for aromatic aminotransferase function usually possess two differentially specialized paralog genes, one (*hisG*) that typically is within a histidine operon and the other *(aroJ$_{I\beta}$)* located elsewhere. hisG *and aroJ$_{I\beta}$* were previously called *hisH_n* and *hisH_b* in order to connote their properties of **n**arrow specificity and **b**road specificity [8]. A few organisms are known where a single *aroJ$_{I\beta}$* gene fulfills a dual function in both histidine and aromatic biosynthesis. Thus, the surviving $AroJ_{I\beta}$ paralog has captured the histidine-pathway function of the lost HisG paralog. In *B. subtilis* ***aroJ$_{I\beta}$*** *// hisG*) is within a mixed-function supraoperon [26, 30] that also contains *tyrA*.

The common-pathway reactions to prephenate

Table 5 shows the nomenclature of gene/protein families with respect to the seven common-pathway steps of chorismate biosynthesis and, in addition, chorismate mutase (the initial step common to PHE and TYR biosynthesis). The far-left column includes broad catalytic roles that may include different analog types. If so, the analog groupings are shown as Superfamilies in the next column to the right. We have been conservative here. For example, on structural grounds MurA and AroF [31] may very well be distant homologs, as is also the case for the $AroH_I$ and $AroH_{III}$ pair [32]. Bold type indicates specific function in aromatic biosynthesis, and the steps are named in the fourth column in the order of reaction sequences in the pathway, i.e., AroA, AroB, etc. Other homolog relatives that possess different functions are also shown in order to illustrate functional roles in a context of broad evolutionary relationships. In addition, it is of interest that a functional role of one grouping sometimes has the potential to be extended to nearby homolog relatives via “suppressor” mutation(s).

AroA. In the first section of Table 5 AroA is seen to separate into two distinct analog groups of 3-deoxy-ald-2-ulosonate phosphate synthases at the level of Superfamily, one of which is further subdivided (recall the previous presentation in Fig. 2). Any enzyme catalyzing the DAHP synthase reaction is denoted AroA. Roman-numeral subscript identifiers indicate placement within homology Superfamily I or II, and if in Superfamily I, whether it belongs to Family α or β. In order to be fully described in its homology context, KdsA (an enzyme of lipopolysaccharide synthesis) would be denoted as $KdsA_{I\beta}$. Most enteric bacteria possess three regulatory paralogs of DAHP synthase, each differentially subject to allosteric inhibition by phenylalanine, tyrosine, or tryptophan. Appending the additional subscript identifiers $_F$, $_Y$, and $_W$ provides the appropriate connotations of allosteric specificity, respectively ($AroA_{I\alpha\_F}$, $AroA_{I\alpha\_Y}$ and $AroA_{I\alpha\_W}$). The $AroA_{II}$ class is subdivided. $AroA_{II}$ includes a variety of microbial species and has a cytoplasmic location. The members of $*AroA_{II}$ are all from higher plants and possess a transit peptide for localization to plastids.

AroB. AroB has no known analog relatives that exercise the function of DAHP cyclization, the generally present second step of chorismate biosynthesis. It is interesting that two other carbocyclases exist as distant homologs. 2-Deoxy-*scyllo*-inosose (DOI) synthase from *Bacillus circulans* catalyzes the carbocycle-forming reaction from glucose 6-P as an initial step of antibiotic formation [33]. ValA, the initial enzyme of validamycin antibiotic biosynthesis in *Streptomyces hygroscopicus* utilizes *D*-sedoheptulose 7-phosphate in a comparable carbocycle-forming reaction [34].

AroA' and AroB'. In some *Archaea* and in a few *Bacteria,* aromatic biosynthesis is initiated with completely different substrates than are used in the classical biosynthesis scheme. This is a most striking recent discovery by a single author [19]. Here *L*-aspartate semialdehyde and 6-deoxy-5-ketohexose 1-phosphate are condensed by AroA' to form 2-amino-3,7-dideoxy-*D*-*threo*-hept-6-ulosonate. This is utilized by AroB', which catalyzes both oxidative deamination and carbocyclase reactions to produce dehydroquinate. This is then the point of convergence of the novel pathway with the classical pathway. The discovery of a completely different interface of carbohydrate metabolism with aromatic biosynthesis sets the stage for intriguing new metabolic insights.

AroC. AroC is represented by two broadly distributed analogs: $AroC_I$ and $AroC_{II}$. In the reverse direction, dehydroquinase can function for catabolism of shikimate and/or quinate. Although $AroC_I$ has frequently been referred to as “biosynthetic dehydroquinase” and $AroC_{II}$

as "catabolic dehydroquinase", it would appear that $AroC_{II}$ functions in the biosynthetic direction about as often as $AroC_I$. Although *E. coli* was demonstrated to use $AroC_I$ for biosynthesis at an early time, most enteric bacteria (even ones as close to *E. coli* as *Klebsiella* and *Yersinia*) deploy $AroC_{II}$ for biosynthesis (deduced because $AroC_{II}$ is the sole species of AroC present in these genomes). *Pseudomonas aeruginosa* possesses two paralogs of $AroC_{II}$, one engaged in biosynthesis and the other in catabolism [5], etc. *Bacillus subtilis* employs $AroC_I$ for biosynthesis and $AroC_{II}$ for catabolism.

AroD. As is the case with AroC, the shikimate/quinate dehydrogenases function in the forward direction for biosynthesis and in the reverse direction for catabolism of shikimate and/or quinate. Quinate, in particular, is one of the most abundant sources of carbon and energy on the planet [35]. The Superfamily-I enzymes are NAD(P)-dependent dehydrogenases, whereas the analogs that populate Superfamily-II are membrane-associated dehydrogenases that use pyrrolo-quinoline-quinone as the cofactor instead. The latter is a catabolic quinate dehydrogenase. AroD members (the α Subfamily) do not utilize quinate and are strictly specific for NADPH, as expected for a reductive step in biosynthesis. YdiB, in contrast, is a catabolic enzyme that can use either of the oxidized cofactors, as well as either shikimate or quinate. SDH-L is specific for shikimate and NADP, similar to AroD, but it has been speculated to provide some other unknown function [36].

Reversed order of reactions AroC and AroD in some lineages? Dehydroquinate is converted to shikimate via a dehydratase reaction and a dehydrogenase reaction. It has been pointed out that these reactions could easily occur in either order, as well as that the individual alternative reactions have been demonstrated [35, 37]. As illustrated in Fig. 3, the classical-pathway step order generates dehydroshikimate as a unique intermediate, whereas the alternative order generates quinate as a unique intermediate. There has been relatively little experimental effort to assess the extent to which various apparent dehydroquinate dehydratases might in fact function as quinate dehydratases. Such quinate dehydratases have the potential to couple with enzymes such as the aforementioned substrate-ambiguous YdiB as a quinate dehydrogenase to function as the lower pathway shown in Fig. 3. The state of knowledge here can be compared to a time prior to the finding that arogenate dehydrogenase existed as a potential alternative to prephenate dehydrogenase in nature [38].

AroE. There are two analog types of AroE (shikimate kinase). $AroE_I$ is generally related to NMP kinases, whereas $AroE_{II}$ is generally related to GHMP kinases (Table 5). The

bioinformatic framework of logic and the experimental followup to demonstrate $AroE_{II}$ is a premier example of new-gene discovery in the genomic era [39]. $AroE_{II}$ is thus far restricted to *Archaea* in its distribution. It is interesting that $AroE_{I}$ and $AroE_{II}$ illustrate a case where two analogs share the same convergently evolved substrate specificity, yet each is related to subsets within their own homology groups that exhibit different substrate specificities. This is entirely reminiscent of the situation presented in Figure 2.

AroF. AroF is a highly conserved enolpvruvyl transferase (Superfamily I) that might be distantly related to the MurA type of carboxyvinyl-transferase (Superfamily II) that participates in cell wall biosynthesis [31].

AroG. Chorismate synthase (AroG) exhibits a structure that thus far shows no similarity to any other structurally characterized protein [40]. Chorismate synthase has an absolute requirement for reduced FMN. At the enzymological level, *Neurospora crassa* and *Saccharomyces cerevisiae* have a "built-in" NADPH:FMN oxidase activity that is usually supplied externally by other organisms. A basis for this difference at the level of amino-acid sequence is not yet apparent.

AroH. Chorismate mutases supply the phenylalanine and tyrosine branches with prephenate. Chorismate mutases are distributed throughout an assemblage designated in Table 5 as having cyclohexadienyl mutase as the broad catalytic activity. The cyclohexadienyl mutases are represented by three analog Superfamilies: $AroH_{I}$, $AroH_{II}$ and $AroH_{III}$. Crystal structures have been presented for each: $AroH_{I}$ from *Escherichia* coli [41], $AroH_{II}$ from *Bacillus subtilis* [42], and $AroH_{III}$ from yeast [43]. $AroH_{I}$ and $AroH_{II}$ are small prokaryote sequences, with $AroH_{I}$ having the broadest distribution. Some organisms possess both $AroH_{I}$ and $AroH_{II}$ simultaneously, e.g. *Bacillus subtilis*. $AroH_{III}$ is much larger than $AroH_{I}$ and $AroH_{II}$ due to the presence of a complex allosteric region that interacts with all three aromatic amino acids, tryptophan being an activator, and tyrosine and phenylalanine being inhibitors. Valid support has been presented on structural grounds that $AroH_{I}$ and $AroH_{III}$ could be distant homologs [32], but we take a conservative stance on this point for the present. $AroH_{I}$ subdivides into a number of families, some of which have different functional roles. PchB utilizes isochorismate instead of chorismate, and functions in pyochelin biosynthesis. PapB functions in *p*-amino-phenylalanine biosynthesis, a step needed for antibiotic synthesis [44]. However, PchB is known to have weak activity as chorismate mutase [45]. *$AroH_{I}$ is a novel

chorismate mutase that has a signal peptide which mobilizes it to the periplasm for some unknown function [5].

Note that $AroH_I$ is fully described as $AroH_{I\alpha}$ (Table 5). However, since other families, except $^*AroH_I$, do not exhibit chorismate mutase function, it is unnecessary to use the Family descriptor. (Although $^*AroH_I$, the sole occupant of the β family, functions as chorismate mutase, it is already differentially marked by the asterisk). If new chorismate mutases emerge in other families, an appropriate nomenclature is in place.

Phenylalanine and Tyrosine Branches

The pathway branches that lead from prephenate to *L*-phenylalanine and from prephenate to *L*-tyrosine exhibit a phenomenon in which the overall multi-step chemistry is identical; yet this is accomplished with individual reactions that are different. The biosynthesis of either phenylalanine or tyrosine from prephenate proceeds in two steps. Depending upon different combinations of substrate specificity that prevail in the enzymes of different organisms, the steps of transamination and dehydration (in the case of phenylalanine) or transamination and dehydrogenation (in the case of tyrosine) can occur in opposite order. The alternative order of reactions generates different intermediates that intervene between exactly the same starting substrate (prephenate) and exactly the same ultimate products (phenylalanine or tyrosine). This is detailed in Fig. 1. (A zoom-in view of the appropriate portion of Fig. 1 can be accessed at http://www.aropath.lanl.gov/Visualizations/TyrPath/TyrPath.htm). Thus, when prephenate is transaminated as the penultimate step of tyrosine and/or phenylalanine biosynthesis, *L*-arogenate is a unique intermediate. On the other hand, if transamination occurs as the final step, then phenylpyruvate and 4-hydroxyphenylpyruvate are unique intermediates of phenylalanine and tyrosine biosynthesis, respectively. Thus, *L*-arogenate and phenylpyruvate are alternative intermediates of phenylalanine biosynthesis, depending upon the order of the transaminase and dehydratase reactions. Likewise, *L*-arogenate and 4-hydroxyphenylpyruvate are alternative intermediates of tyrosine biosynthesis, depending upon the order of the transaminase and dehydrogenase reactions. The creation of unique intermediates, while not affecting the overall biochemical transformation, has profound physiological meaning in terms of the placement of the metabolic branchpoint, the pattern of allosteric control, and existence of differentially vulnerable antimetabolite targets. A few

examples follow. In *E. coli* prephenate marks the branchpoint, and the prephenate-utilizing dehydratase and the prephenate-utilizing dehydrogenase are the focal points of allosteric control. At the other extreme in higher plants, *L*-arogenate is at the branchpoint, and the arogenate-utilizing dehydratase and arogenate-utilizing dehydrogenase are the focal points of allosteric control. Illustrative of another pattern, in cyanobacteria prephenate represents the point of pathway divergence, just like *E. coli*. However, in this case the prephenate-utilizing dehydratase and the arogenate-utilizing dehydrogenase are the foci of allosteric control. (The aromatic aminotransferases are never subject to feedback inhibition because they are freely reversible reactions).

Rule x applies to aromatic-pathway dehydrogenases because, regardless of specificity differences they perform the same overall function of decarboxylation and aromatization in *L*-tyrosine biosynthesis. Likewise **rule x** applies to aromatic-pathway dehydratases because, regardless of specificity differences, they perform the same overall function of decarboxylation, dehydration, and aromatization in *L*-phenylalanine biosynthesis. Aromatic aminotransferases generally contribute to both phenylalanine and tyrosine biosynthesis. And again, regardless of the reaction order with respect to the functional linkage with aromatic-pathway dehydratases or dehydrogenases, the overall function is the same, namely to convert the ring-attached pyruvyl moiety to an alanyl moiety.

A [one enzyme**:**one reaction] concept is convenient, but not always correct. Substrate ambiguities are probably more widespread than generally realized. At one extreme, a terpene synthase is known to produce 52 different sesquiterpenes from a single substrate [46]. The tyrosine/phenylalanine segments illustrate many examples of substrate ambiguity that involve aminotransferase, dehydratase, or dehydrogenase steps. Prephenate aminotransferase, phenylpyruvate aminotransferase, and 4-hydroxyphenylpyruvate aminotransferase can be a suite of reactions catalyzed by a single enzyme or they can be accomplished by closely related paralogs of overlapping function. Cyclohexadienyl dehydrogenase functions as either prephenate dehydrogenase or arogenate dehydrogenase (depending upon relative substrate availability). Likewise, cyclohexadienyl dehydratase functions as either prephenate dehydratase or arogenate dehydratase (depending upon relative substrate availability). In many organisms it is quite likely that a mixture of both flow routes to phenylalanine and/or tyrosine is ongoing at the same time.

Cyclohexadienyl dehydrogenases are very closely related to prephenate-specific dehydrogenases and to arogenate-specific dehydrogenases, and sequence motifs that predict the various profiles of substrate specificity are not yet worked out [47]. We suspect that changes in specificity have occurred independently so many times that it will be difficult to unravel the evolutionary thread without analysis of many very closely related organisms.. The comparable situation appears to apply to the set of phenylalanine-pathway dehydratases.

Hence, regardless of whether the dehydratase or dehydrogenase reactions are the first or second steps proceeding from prephenate, the acronym proper applied is PheA or TyrA, respectively. In each case, the companion step is a broad-specificity aminotransferase belonging to one of at least three homology subdivisions. The application of the AroJ acronymn for aromatic aminotransferases reflects the sense that they can be viewed as an extension of the common-pathway assemblage that is sidetracked only by the dehydratase /dehydrogenase divergence. The aromatic aminotransferases belong equally to phenylalanine and tyrosine biosynthesis in that they have an affinity for a cyclohexadienyl or aromatic ring possessing a pyruvyl sidechain, which is converted to an alanyl sidechain. Phenylalanine and tyrosine biosynthesis deploy in common all of the enzyme reactions represented by AroA through AroH, as well as AroJ. Only the PheA/TyrA diversion accounts for the ultimate production of two different amino acids, differing only in the presence of a 4-hydroxy substituent at the 4-position of the aromatic ring.

Aromatic aminotransferases. Aromatic aminotransferases belong to a large and ancient lineage that has been well studied [48]. Tracking back to the deepest level is the “Alpha division of PLP-dependent proteins. Splitting from this is the “Aminotransferase superfamily”. One of four divergences yields the “Family I aminotransferases”. Different suites of signature amino-acid motifs have been discerned at each of the latter hierarchical levels. In addition, Jensen and Gu [8] have sorted out seven subfamilies within aminotransferase Family I. This is based upon motifs that exhibit differing patterns of conservation distributed around the 11 invariant anchor residues of Family I. Specificity for transamination of aromatic amino acids is thus far known to be represented by members within three of the seven Subfamilies (See Table 6). Subfamily Iα generally contains aspartate aminotransferases of varied breadth of substrate specificity. *E. coli* TyrB is a broad-specificity variant of AspC that is specialized for function as aromatic aminotransferase, largely because it has been captured by the *tyrR* regulon [49]. Subfamily Iβ consists of two

distinctly diverged clusters of imidazoleacetol-P aminotransferase. One (HisG) has narrow specificity and is frequently encoded by a gene present in the histidine operon. The other ($AroJ_{I\beta}$) has a broad-substrate capability to function as an aromatic aminotransferase. Finally, Subfamily Iγ contains aminotransferases of varied function, some of which can function as aromatic aminotransferases [50].

Given the states of broad substrate specificity and free reversibility of reaction that typify aminotransferases, the *in vivo* functional roles are heavily orchestrated by particular modes of regulation. Thus, PhhC from *Pseudomonas aeruginosa* exhibits *in vitro* properties that are very similar to the *E. coli* paralogs AspC and TyrB. Whereas AspC functions as a fundamental aspartate aminotransferase, TyrB and PhhC are effectively specialized by the regulation schemes in place. Thus, *tyrB* is a member of the large *tyrR* regulon that regulates aromatic biosynthesis [49], and *phhC* is a member of the phenylalanine hydroxylase operon regulated by *phhR* [51]. It is instructive that a mutational deficiency of *aspC* in *E. coli* can be suppressed by *tyrB*, provided that the regulatory constraints normally imposed upon tyrB are removed [52]. Aminotransferases typically are freely reversible reactions. Aromatic aminotransferases can be expected to work in a biosynthetic direction if phenylalanine or act as a repressor, and to work in the catabolic direction and if phenylalanine or tyrosine acts as an inducer.

Tyrosine biosynthesis. Considerable bioinformatic analysis of tyrosine biosynthesis has been published recently [47]. All biosynthetic dehydrogenases performing the critical step of aromatization and decarboxylation belong to the single TyrA Superfamily. Within this Superfamily there is a great deal of diversity with respect to the specificity for both the cyclohexadienyl substrate and for the pyridine nucleotide substrate. Acronyms used to describe TyrA specificity patterns are given in Table 7. Thus far, amino-acid sequence motifs that correlate reliably with sequence specificities have not been determined, and it may be that multiple combinations of active-site combinations can lead to similar substrate-specificity profiles. A comparison of X-ray crystal structures from proteins representing differing substrate specificities is badly needed. Good choices would be that from *Bacillus subtilis* [53] (absolutely specific for prephenate) and that from *Synechocystis* [54] (absolutely specific for arogenate). A recent crystal structure has been completed [18] for the TyrA protein from *Aquifex aeolicus*, which is an NAD^+-dependent cyclohexadienyl dehydrogenase that exhibits a great preference for prephenate over *L*- arogenate.

A recent analysis has identified a large number of "cohesion groups", whose members are congruent with 16S rRNA expectations, except for occasional "intruders" (probable LGT events implied). The phylogenetic orientation of cohesion groups with one another on the protein tree is uncertain because of insufficient bootstrap support. However, these cohesion groups (which can be viewed at http://www.aropath.lanl.gov/TyrPath/TyrCG_index.html) are expected to merge as new sequences become available to fill phylogenetic gaps. TyrA cohesion groups fall into two distinct protein families, termed $TyrA_{\alpha}$ and $TyrA_{\beta}$. The larger $TyrA_{\alpha}$ family includes the majority of bacterial TyrA cohesion groups, whereas the $TyrA_{\beta}$ family includes the *Archaea* and some *Bacteria*. Strikingly, the assemblage of lower-gamma Proteobacteria belong to the $TyrA_{\beta}$ family, whereas the upper-gamma Proteobacteria and all other Proteobacteria belong to the $TyrA_{\alpha}$ Family. We speculate that the $TyrA_{\beta}$ family is distinguished by indel alterations that are associated with functional interactions of the TyrA domain with another domain, e.g., with the N-terminally fused $AroH_I$ domain in *E. coli*.

Phenylalanine biosynthesis. Phenylalanine-pathway dehydratases fall into two classes (Table 8). The widely distributed $PheA_I$•ACT has a cytoplasmic location in prokaryotes (and a plastid location in higher plants). The narrowly distributed $*PheA_{IIc}$ has a cleavable signal peptide and is translocated to a periplasmic location in Gram-negative bacteria or is apparently secreted in Gram-positive bacteria. It seems likely that $*PheA_{IIc}$ (as well as $*AroH_I$ and $*AroJ_{I\beta}$) perform some sort of role in secondary metabolism, rather than exerting a role in primary biosynthesis.

Superfamily-I dehydratases, denoted as $PheA_{Ip}$•ACT, are typically specific for prephenate and possess a C-terminal ACT domain that is responsible for complex patterns of allosteric inhibition and activation [55]. The probable loss of allostery following point mutations in the *Buchnera* sp. $PheA_{Ip}$•ACT is an interesting case of endosymbiont innovation to provide the host with an unregulated supply of phenylalanine [56]. Relatively few organisms known to utilize an arogenate-specific dehydratase have genome sequences available, but among the complete genomes, *Gluconobacter oxidans* possesses $pheA_{Ia}$•*ACT*, and higher plants possess $*pheA_{Ia}$•*ACT* (whose product is translocated to chloroplasts). Thus far, no clues are available from sequence comparisons to distinguish prephenate-specific dehydratases from arogenate-specific dehydratases. As speculated for the foregoing aromatic-pathway dehydrogenases, multiple combinations of substrate-binding residues capable of yielding the same specificity profiles may have evolved independently in different lineages. Regardless of substrate

specificity, a C-terminal Act domain is inevitably present. One exception would appear to be the dehydratase-encoding gene of *Magnetospirillum magnetotactum* (gi 23010242), which appears to lack the allosteric domain entirely.

Periplasmic cyclohexadienyl dehydratases, which are encoded by *$pheA_{IIc}$, share a crowded phylogenetic space with the abundantly distributed the periplasmic LAO binding proteins. Four amino-acid residues are known to be needed for catalytic function in *P. aeruginosa* *$PheA_{II}c$. These residues (unpublished X-ray crystal results of J. Liang and J. Clardy) are Y_{47}, W_{85}, R_{110}, and D_{195} (numbered according to the *Pseudomonas aeruginosa* sequence). However, these residues are also highly conserved in LAO binding proteins, although perhaps less so. In contrast to the clear homology of the N-terminal regions of *$PheA_{IIc}$ and LAO binding proteins, the C-terminal portion of *$PheA_{IIc}$ proteins and of LAO binding proteins are quite divergent from one another, and this allows the separation of *$PheA_{IIc}$ proteins from LAO binding proteins via sequence comparison. It is interesting that two small subgroups diverge from the main *$PheA_{IIc}$ cluster because this might prove to reflect narrowed substrate specificities for prephenate (in which case the acronym would be *$PheA_{IIp}$) or for *L*-arogenate (in which case the acronym would be *$PheA_{IIa}$).

Different-function homolog sequences that crowd a common phylogenetic space

A difficult dilemma of acronyn usage occurs when sequences conferring different function crowd the same phylogenetic space. This is because without a clean correlation of functional divergence and sequence divergence, the sorting of functional roles can be quite challenging. For example, the two subunits of anthranilate synthase (TrpAa and TrpAb) are inter-mixed with the corresponding subunits of PABA synthase (PabAa and PabAb) when protein trees of TrpAa/PabAa or TrpAb/PabAb are constructed. Fortuitously, most of these can be assigned the correct function because of the high frequency of both *trp* and *pab* operons. So far no structural motifs or conserved pattern of residues are known that will generally discriminate between the two subunit pairs.

Similarly, the substrate specificity of TyrA proteins, whether they be specific for prephenate ($TyrA_p$), specific for *L*-arogenate ($TyrA_a$), or broadly specific ($TyrA_c$), cannot be correlated with homology sub-clusters or discerned by any known motifs at the present time. The PheA

proteins have been less heavily studied, but again the substrate specificities of $PheA_I$ proteins, whether specific for prephenate or specific for *L*-arogenate, cannot be predicted at present from sequence comparisons. Broad-specificity enzymes of the $PheA_I$ homology class (cyclohexadienyl dehydratase) are not currently known, although it seems likely that they exist. Only broad-specificity cyclohexadienyl dehydratase is known to populate the $PheA_{II}$ homology class at present, but the great majority of sequences now available have not been studied experimentally. It would not be surprising if these prove to include prephenate-specific and arogenate-specific enzymes in addition to the cyclohexadienyl dehydratases.

The ease of tracking nodes of function along the evolutionary thread can be expected to vary with the degree of sequence conservation that is typical of a given enzyme reaction. For those reactions which exhibit considerable plasticity, being prone to facile changes in substrate recognition, a more highly refined scale of analysis is necessary. That is, one must zoom in on more sequences from a more compressed phylogenetic slice of organisms. The substrate specificities of TyrA proteins exemplify a scenario where divergence to a given specificity appears to have occurred independently and via different combinations of mutations on many occasions. Given sufficient experimental information with organisms having appropriate phylogenetic spacing, it should be possible to identify multiple nodes of specificity that are reminiscent of those shown in Fig. 2 (albeit more complicated).

Figure Legends

Fig. 1. Biochemical pathway of aromatic amino acid biosynthesis. The segments discussed herein are the "common" pathway linking carbohydrate metabolism to chorismate, the tryptophan branch, the single step from chorismate to prephenate, the phenylalanine branch, and the tyrosine branch. The branch from chorismate to *p*-aminobenzoate is also shown because of the close relationship of PabAa and PabAb to TrpAa and TrpAb. This pathway can also be accessed at http://www.aropath.lanl.gov/Visualizations/AroPath/AroPath.htm.

Clickable acronyms there will return a table from which gi-number queries are hyperlinked to NCBI in order to provide curated representatives of any given functional grouping.

Fig. 2. Placement of DAHP (3-deoxy-**D**-*arabino*-heptulosonate 7-P) synthases (AroA proteins) within their phylogenetic context. The protein tree shown on the left (not drawn to scale) depicts a set of enzymes within Superfamily I that includes KDOP (3-deoxy-**D**-*manno*-octulosonate-8-P) synthase. The Superfamily-II DAHP synthases on the right are analogs of the three co-homolog proteins shown on the left. The evolutionary scenario is that one (or two) ancestral 3-deoxy-ald-2-ulosonate phosphate synthases of broad substrate specificity diverged to yield narrow-specificity DAHP synthases or KDOP synthase. The shaded orange cones depict the emergence of DAHP synthase on three independent occasions, and the horizontal arrows indicate hierarchical positions that mark narrowed specificity to yield DAHP synthases. Single-domain examples of each of the four groupings are given at the bottom (first line), and examples of domains with more complex features are given on the second line. Bacterial abbreviations: *Ctep*, *Chlorobum tepidum*; *Ecol*, *Escherichia coli*; *Tmar*, *Thermotoga maritima*; *Hpyl*, *Helicobacterium pylori*; *Bsub*, *Bacillus subtilis*.

Fig. 3. Plausible reversal of enzymatic steps between dehydroquinate and shikimate. In the classical pathway (upper route), a dehydratase functions first, followed by a dehydrogenase reaction. In the bottom route (as discussed by Jensen *et al*. [37], by Bonner and Jensen [35], and refs. therein) a dehydrogenase functions first, followed by a dehydratase reaction. These two routes produce dehydroshikimate and quinate as unique intermediates of a two-step overall conversion that is otherwise identical.

Tables

Table 1. Key to nomenclature[a] for tryptophan biosynthetic pathway

New gene name	Prior gene name	Protein domain encoded	Reaction order
trpAa	*trpE*	Anthranilate synthase: aminase subunit (α)	1
trpAb	*trpG*	Anthranilate synthase: amidotransferase subunit (β)	
trpB	*trpD*	Anthranilate phosphoribosyl transferase	2
trpC	*trpF*	Phosphoribosyl-anthranilate isomerase	3
trpD	*trpC*	Indoleglycerol phosphate synthase	4
trpEa	*trpA*	Tryptophan synthase, α subunit	5
trpEb	*trpB*	Tryptophan synthase, β subunit	

[a]The nomenclature is at the level of catalytic domain in the order of reaction steps in the pathway. See Xie et *al*. [57] for a detailed rationale supporting the new nomenclature. Overall reactions 1 and 5 consist of two subunits, with α and β subunits assigned the lowercase 'a' and 'b' designations, respectively, according to **rule iv.** The convention of a bullet denotes a fusion, e.g., *trpD•trpC,* as illustrated below.

Consider the *trp* operon of *Escherichia coli*:

Current nomenclature**:** *trpE trpG•D trpC•F trpB trpA* or

trpE trp(G)D trpC(F) trpB trpA

Logical nomenclature**:** *trpAa trpAb•B trpD•C trpEb trpEa*

Table 2. Key to nomenclature for histidine biosynthetic pathway

Step	Enzyme name	Logical nomen-clature	E. coli nomen-clature	E. coli fusion partner
[1]	ATP phosphoribosyl transferase	*hisA*	*hisG*	
[2]	PR-ATP pyrophospholydrolase	*hisB*	*hisI*^{c}	*hisI*^{n}
[3]	PR-AMP cyclohydrolase	*hisC*	*hisI*^{n}	*hisI*^{c}
[4]	Phosphoribosyl-5-amino-1-phosphoribosyl-4-imidazole carboxamide isomerase	*hisD*	*hisA*	
[5]	Imidazole glycerol-P synthase (cyclase subunit) (amidotransferase subunit)	 *hisEa* *hisEb*	 *hisF* *hisH*	
[6]	Imidazole glycerol-P dehydratase	*hisF*	*hisB*$_{d}$	*hisB*$_{px}$
[7]	Imidazole acetol-P aminotransferase	*hisG*	*hisC*	
[8]	Histidinol-P phosphatase	*hisH*	*hisB*$_{px}$	*hisB*$_{d}$
[9]	Histidinol dehydrogenase	*hisI*	*hisD*	

***E.coli his* operon as an example:**

Current nomenclature* *hisG hisD hisC hisB*$_{px}$•*B*$_{d}$ *hisH hisA hisF hisI*^{n}•*I*^{c}

Logical nomenclature *hisA hisI hisG hisH•F hisEb hisD hisE*a *hisC•B*

*Frequently, the fused domains (*hisB*$_{px}$•*hisB*$_{d}$ and *hisI*^{n}•*I*^{c}) have been referred to as '*hisB*' and '*hisI*' (or *hisIE*), respectively.

Table 3. Recent expansions of nomenclature for genes of Trp biosynthesis

Table 1 entry	Updated acronym	Functional reaction(s) encoded[a]	Query Tag (gi number)
trpC	$trpC_I$	PR-anthranilate isomerase	Bsub 143771
	hisD *// $trpC_{II}$*	**PR-5-amino-1-PR-4-imidazole carboxamide isomerase**/(PR-anthranilate isomerase)	Mtub 45593458
trpAb	*trpAb*	Anthranilate synthase: amidotransferase subunit	Ypes 15980206
	pabAb *// trpAb*	PABA synthase: amidotransferase subunit/ **anthranilate synthase: amidotransferase subunit**	Bsub 129521
trpEa	$trpEa_{\alpha}$	Tryptophan synthase, α subunit (ubiquitous)	Ecol 1787514
	$trpEa_{\beta}$	Tryptophan synthase, α subunit (rare)	Ssol 1004320
trpEb	$trpEb_{\alpha}$	Tryptophan synthase, β subunit (ubiquitous)	Ecol 1787515
	$trpEb_{\beta}$	Tryptophan synthase, β subunit (rare)	Ssol 1004319

[a]Abbreviation: PR, phosphoribosyl. In each of the four table blocks shown (separated by bold horizontal lines), the gene entry shown in Table 1 is given in the first column, and is renamed (usually), as shown directly across from it in the second column. Acronyms of newly recognized genes having the same function are shown on the bottom line of each block.

Table 4. Recent nomenclature expansions for His biosynthesis

Table 2 entry	Updated acronym	Functional reactions encoded[a]	Query Tag (gi number)
hisA	*hisA•HAD$_{II}$*	ATP PR-transferase•allosteric domain (HisG$_L$)	Ecol 1788330
	hisA	ATP PR-transferase (HisG$_s$)	Tmar 7387767
	HAD$_I$	Allosteric subunit (HisZ)	Tmar 20978516
hisD	*hisD*	PR-6-amino-1-PR-4-imidazole carboxamide isomerase	Ecol 87082028
	hisD *// trpC$_{II}$*	**PR-6-amino-1-PR-4-imidazole carboxamide isomerase** /PR-anthranilate isomerase	Mtub 45593458
hisG	*hisG*	Imidazoleacetol-P aminotransferase	Ecol 1788332
	aroJ$_{Iβ}$ *// hisG*	**Aromatic aminotransferase**/Imidazoleacetol aminotransferase	Bsub 3123224

[a]Abbreviation: PR, phosphoribosyl. In each of the three table blocks shown, the gene entry shown in Table 2 is given in the first column. The new name, when applicable, is presented directly across from it in the second column. Acronyms of newly recognized genes having the same function are shown on the bottom line of each block.

Table 5. Nomenclature for common-pathway reactions to prephenate

Broad Catalytic Reaction	Super-family	Family	Subfamily	Functional Reaction	Query Tag (gi number)
3-Deoxy-ald-2-ulosonate phosphate synthase	I	α	**$AroA_{I\alpha}$**	**DAHP synthase**	**Ctep 21646940**
		β	**$AroA_{I\beta}$**	**DAHP synthase**	**Tmar 4980844**
			KdsA	KDOP synthase	Ecol 1708631
	II		**$AroA_{II}$**	**DAHP synthase**	**Hpyl 7465174**
			***$AroA_{II}$**	**DAHP synthase**	**Lesc 410486**
DeoC/FbaB aldolase family			**AroA'**	**ADTH synthase**	**Mjan 1591105**
Carbocyclase	I	α	**AroB** ***AroB**	**Dehydroquinate synthase I** **Dehydroquinate synthase I**	**Enid 5822049** **Atha 18425036**
		β	BtrC	DOI synthase	Bcir 11191970
		γ	ValA	Sedoheptulose 7-P cyclase	Shyg 59710122
Deaminating carbocyclase			**AroB'**	**Dehydroquinate synthase II**	**Mjan 1591882**
Dehydroquinase	I		**$AroC_I$** ***$AroC_I$•**	**Dehydroquinase** **Dehydroquinase•**	**Styp 114188** **Atha 15230703**
	II		**$AroC_{II}$**	**Dehydroquinase**	**Mtub 6226839**

Continued overleaf

Shikimate/quinate Dehydrogenase	I	α	**AroD** **•AroD**	**Shikimate Dehydrogenase** **Shikimate dehydrogenase**	**Ecol 16131162** **Atha 15230703**
		β	*YdiB*	Shikimate/quinate dehydrogenase	Ecol 15831653
		γ	SDH-L	Shikimate dehydrogenase	Hinf 42629096
	II		QuiA	Quinate dehydrogenase	Acal 9087181
Small-molecule Kinase	I NMP Kinase	α	Cmk	Cytidylate kinase	Ecol 2506790
			Adk	Adenylate kinase	Ecol 125161
		β	**$AroE_I$** ***$AroE_I$**	**Shikimate kinase** **Shikimate kinase**	**Echr 114199** **Atha 30692396**
	II GHMP Kinase	α	ThrB	Homoserine kinase	Ecol 366419
			GalK	Galactokinase	Ecol 120898
			MvaK	Mevalonate kinase	Spyo 13622041
		β	**$AroE_{II}$**	**Shikimate kinase**	**Mjan 14194467**
Carboxyvinyl-transferase	I		**AroF** ***AroF**	**EPSP synthase** **EPSP synthase**	**Ecol 2506201** **Atha 15225450**
	II		**MurA**	UNAG enolpyruvyl transferase	Ecol 1789580
Chorismate Synthase			**AroG** ***AroG**	**Chorismate synthase** **Chorismate synthase**	**Ecol 114183** **Atha 18402389**

Continued overleaf

Cyclohexadienyl mutase	I	α	**AroH$_I$**	**Chorismate mutase**	**Mjan 2495875**
		β	***AroH$_I$**	**Periplasmic chorismate mutase**	**Paer 11350384**
		γ	PchB	Isochorismate mutase	Paer 12230975
		δ	PapB	ADC mutase	Spri 1575338
	II		**AroH$_{II}$**	**Chorismate mutase**	**Bsub 6234687**
	III		**AroH$_{III}$** ***AroH$_{III}$**	**Allosteric chorismate mutase** ***Allosteric chorismate mutase**	**Scer 6325317** **Atha 18406100**

Homology-group members with specificity directly relevant to aromatic-pathway biosynthesis are shown in bold type. Gi numbers for query sequences that can be used to identify each grouping are shown at the far right. **Abbreviations: ADC**, 4-amino,4-deoxy-chorismate; **ADTH,** 2-amino-3,7-dideoxy-**D**-*threo*-hept-6-ulosonate; **DAHP**, 3-deoxy-**D**-*arabino*-heptulosonate 7-phosphate; **DOI**, 2-deoxy-scyllo-inosose; **EPSP,** enolpyruvylshikimate-3-phosphate; **GHMP**, galactose/homoserine/mevalonate/ phosphomevalonate; **KDOP**, 3-deoxy-**D**-*manno*-octulosonate-8-phosphate; **NMP**, nucleoside monophosphate kinase; UNAG, UDP-*N*-acetylglucosamine; **gi**, gene identification; **Chorismate mutase_P**, periplasmic chorismate mutase; **Ecol,** *Escherichia coli,* ***Tmar,*** *Thermotoga maritima*; **Hpyl**, *Helicobacter pylori*; **Enid**, *Emericella nidulans;* ***Bcir****, Bacillus circulans*; **Styp**, *Salmonella typhi*; **Mtub**, *Mycobacterium tuberculosis*; **Acal**, *Acinetobacter calcoaceticus*; **Pchr**, *Pectobacterium chrysanthemi*; **Spyo**, *Streptococcus pyogenes;* **Mjan**, *Methanococcus jannaschii*; **Paer,** *Pseudomonas aeruginosa*; **Spri**, *Streptomyces pristinaespiralis*; **Scer**, *Saccharomyces cerevisiae*; **Bsub**, *Bacillus subtilis*. Entries having green fonts correspond to proteins of higher plants that possess cleavable transit peptides and become localized in chloroplast organelles.

Table 6. Key to nomenclature of aromatic aminotransferases

Family I subfamily	New gene name	Typical subfamily role	Example	Query Tag (gi number)
α	*aroJ*$_{Iα}$	Aspartate AT (AspC)	[a]Ecol "TyrB"	85676806
β	*aroJ*$_{Iβ}$ **aroJ*$_{Iβ}$	Imidazoleacetol-P AT (HisG) Periplasmic aromatic AT	[b]Hinf "HisH" Paer	1574093 14794424
γ	*aroJ*$_{Iγ}$	Aromatic AT	[c]Llac "AraT"	6318592

[a]*E. coli* and its close relatives have two close paralogs: AspC for general interconversion of aspartate and oxaloacetate, and tyrB for aromatic biosynthesis.

[b]*Haemophilus influenzae* possesses two paralogs of the HisG aminotransferase: one of narrow specificity for imidazoleacetol-P (HisG) and the other of broad specificity that can also function as aromatic aminotransferase. (*Bacillus subtilis* and close relatives lack the typical omnipresent, operon-bounded, narrow-specificity species of HisG, and the remaining broad-specificity **AroJ$_{β}$** // HisG functions for both histidine and aromatic biosynthesis.)

[c]See Rijnen et al. [50].

Table 7. Abbreviations used to designate substrate specificities of tyrA/TyrA homologs

Description of specificity[a]	Abbreviations[b]	
	Gene	Gene Product
Specificity for cyclohexadienyl substrate is unknown	*tyrA*$_x$	TyrA$_x$
Broad-specificity cyclohexadienyl dehydrogenase (CDH)	*tyrA*$_c$	TyrA$_c$
Narrow-specificity prephenate dehydrogenase (PDH)	*tyrA*$_p$	TyrA$_p$
Broad-specificity cyclohexadienyl dehydrogenase having catalytic-core indels in correlation with an extra-core extension	*tyrA*$_{c\_\Delta}$	TyrA$_{c_\Delta}$
Narrow-specificity arogenate dehydrogenase (ADH)	*tyrA*$_a$	TyrA$_a$
TyrA homolog is AGN-specific and NAD^+-specific	$_{NAD}$*tyrA*$_a$	$_{NAD}$TyrA$_a$
TyrA homolog is AGN-specific and $NADP^+$-specific	$_{NAD}$*tyrA*$_a$	$_{NADP}$TyrA$_a$
TyrA homolog is AGN-specific but utilizes either NAD^+ or $NADP^+$	$_{NAD(P)}$*tyrA*$_a$	$_{NAD(P)}$TyrA$_a$
Specificity for both the cyclohexadienyl and pyridine nucleotide substrates is unknown	$_x$*tyrA*$_x$	$_x$TyrA$_x$

[a]The abbreviations CDH, PDH, and ADH (shown in parentheses) have been used frequently in the literature.

[b]Abbreviations in the upper-half table indicate the specificities for the cyclohexadienyl substrate. Abbreviations in the lower-half table indicate specificities for both cyclohexadienyl (right subscripts) and pyridine nucleotide substrates (left subscripts). Combinations not shown can be deduced from the examples given, e.g., a TyrA homolog specific for prephenate and NAD^+ would be designated $_{NAD}$TyrA$_p$.

Table 8. Key to nomenclature for phenylalanine biosynthetic pathway

Super-family	New gene name	Prior gene name	Protein domain encoded	Query Tag (gi numbers)
I	*pheA$_{Ip}$•ACT*	*pheA*	Prephenate dehydratase•allosteric domain	Bsub 130048
	pheA$_{Ia}$•ACT **pheA$_{Ia}$•ACT*		Arogenate dehydratase•allosteric domain *Arogenate dehydratase•allosteric domain	Goxy 58000965 Atha 79317657
II	**pheA$_{IIc}$*	*pheC*	Periplasmic cyclohexadienyl dehydratase	Paer 2997758

Acknowledgements

R. Jensen thanks the National Library of Medicine (Grant G13 LM008297) for its support of scholarly studies.

References

1. Brown SD, Gerlt JA, Seffernick JL, Babbitt PC: **A gold standard set of mechanistically diverse enzyme superfamilies** *Genome Biol* 2006, **7**:R8.
2. Orengo CA, Thornton JM: **Protein families and their evolution-a structural perspective**. *Annu Rev Biochem* 2005, **74**:867-900.
3. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C: **The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli***. *J Mol Biol* 2001, **311**(4):693-708.
4. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective**. *J Mol Biol* 2001, **307**(4):1113-1143.
5. Calhoun DH, Bonner CA, Gu W, Xie G, Jensen RA: **The emerging periplasm-localized subclass of AroQ chorismate mutases, exemplified by those from *Salmonella typhimurium* and *Pseudomonas aeruginosa***. *Genome Biol* 2001, **2**(8):RESEARCH0030.0031-0030.0016.
6. Gosset G, Bonner CA, Jensen RA: **Microbial origin of plant-type 2-keto-3-deoxy-*D-arabino*-heptulosonate 7-phosphate synthases, exemplified by the chorismate- and tryptophan-regulated enzyme from *Xanthomonas campestris***. *J Bacteriol* 2001, **183**:4061-4070.
7. Gu W, Williams DS, Aldrich HC, Xie G, Gabriel DW, Jensen RA: **The *aroQ* and *pheA* domains of the bifunctional P-protein from *Xanthomonas campestris* in a context of genomic comparison**. *Microb Comp Genomics* 1997, **2**(2):141-158.
8. Jensen RA, Gu W: **Evolutionary recruitment of biochemically specialized subdivisions of Family-I within the protein superfamily of aminotransferases**. *J Bacteriol* 1996, **178**:2161-2171.
9. Jensen RA, Xie G, Calhoun DH, Bonner CA: **The correct phylogenetic relationship of KdsA (3-deoxy-d-manno-octulosonate 8-phosphate synthase) with one of two independently evolved classes of AroA (3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthase)**. *J Mol Evol* 2002, **54**(3):416-423.
10. Subramaniam PS, Xie G, Xia T, Jensen RA: **Substrate ambiguity of 3-deoxy-D-*manno*-octulosonate 8-phosphate synthase from Neisseria gonorrhoeae in the context of its membership in a protein family containing a subset of 3-deoxy-D-*arabino*-heptulosonate 7-phosphate synthases**. *J Bacteriol* 1998, **180**:119-127.
11. Xie G, Bonner CA, Brettin T, Gottardo R, Keyhani NO, Jensen RA: **Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in *Xylella* species and in heterocystous cyanobacteria**. *Genome Biol* 2003, **4**(2):R14.
12. Xie G, Forst C, Bonner CA, Jensen RA: **Significance of two distinct types of tryptophan synthase beta chain in Bacteria, Archaea and higher plants**. *Genome Biol* 2002, **3**(1):RESEARCH0004.0001-0004.0013.
13. Tian W, Skolnick J: **How well is enzyme function conserved as a function of pairwise sequence identity?** *J Mol Biol* 2003, **333**:863-882.

14. Rost B: **Enzyme function less conserved than anticipated**. *J Mol Biol* 2002, **318**(2):595-608.
15. Seffernick JL, de Souza ML, Sadowsky MJ, Wackett LP: **Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different**. *J Bacteriol* 2001, **183**(8):2405-2410.
16. Jensen RA: **Enzyme recruitment in evolution of new function.** . *Ann Rev Microbiol* 1976, **30**:409-425.
17. Chothia C, Gough J, Vogel C, Teichmann SA: **Evolution of the protein repertoire**. *Science* 2003, **300**(5626):1701-1703.
18. Sun W, Singh S, Zhang R, Turnbull JL, Christendat D: **Crystal structure of prephenate dehydrogenase from *Aquifex aeolicus:* Insights into the catalytic mechanism** *(in press)* 2006.
19. White RH: ***L*-Aspartate semialdehyde and a 6-deoxy-5-ketohexose 1-phosphate are the precursors to the aromatic amino acids in *Methanocaldococcus jannaschii***. *J Biol Chem* 2004, **43**:7618-7627.
20. Kurnasov O, Goral V, Colabroy K, Gerdes S, Anantha S, Osterman A, Begley TP: **NAD biosynthesis: identification of the tryptophan to quinolinate pathway in bacteria**. *Chem Biol* 2003, **10**(12):1195-1204.
21. Fitch WM: **Homology a personal view on some of the problems**. *Trends Genet* 2000, **16**(5):227-231.
22. Yanofsky C: **Advancing our knowledge in biochemistry, genetics, and microbiology through studies on tryptophan metabolism**. *Annu Rev Biochem* 2001, **70**:1-37.
23. Alifano P, Fani R, Lio P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB: **Histidine biosynthetic pathway and genes: structure, regulation, and evolution**. *Microbiol Rev* 1996, **60**(1):44-69.
24. Barona-Gómez F, Hodgson DA: **Occurence of a putative ancient-like isomerase involved in histidine and tryptophan biosynthesis**. *EMBO Rep* 2003, **4**:296-300.
25. Xie G, Keyhani NO, Bonner CA, Jensen RA: **Ancient origin of the tryptophan operon and the subsequent dynamics of evolutionary change**. *Microbiol Mol Biol* 2003, **67**:303-342.
26. Henner D, Yanofsky C: **Biosynthesis of aromatic amino acids**. In: *Bacillus subtilis and Other Gram-Positive Bacteria: Biochemistry, Physiology, and Molecular Genetics.* Edited by Sonenshein AL, Hoch JA, Losick R. Washington, DC: Am. Soc. Microbiol. ; 1993: 269-280.
27. Weber-Ban E, Hur O, Bagwell C, Banik U, Yang L-H, Miles EW, Dunn MF: **Investigation of allosteric linkages in the regulation of tryptophan synthase: The roles of salt bridges and mono-valent cations probed by site-directed mutation, optical spectroscopy, and kinetics**. *Biochemistry* 2001, **40**:3497-3511.
28. Bovee ML, Champagne KS, Demeler B, Francklyn CS: **The quaternary structure of the HisZ *N*-1-(5' -Phosphoribosyl)-ATPtransferase from *Lactococcus lactis***. *Biochem* 2002, **41**:11838-11846.
29. Champagne KS, Sissler M, Larrabee Y, Doublie S, Francklyn CS: **Activation of the hetero-octameric ATP phosphoribosyl transferase through subunit interface rearrangement by a tRNA synthetase paralog**. *J Biol Chem* 2005, **280**:34096-34104.
30. Xie G, Brettin TS, Bonner CA, Jensen RA: **Mixed-function supraoperons that exhibit overall conservation, albeit shuffled gene organization, across wide intergenomic distances within eubacteria**. *Microb Comp Genomics* 1999, **4**(1):5-28.

31. Eschenburg S, Kabsch W, Healy ML, Schönobrunn E: **A new view of the mechanisms of UDP-*N*-acetylglucosamine enolpyruvyl transferase (Mura) and 5-enolpyruvylshikimate-3-phosphate synthase (AroA) derived from X-ray structures of their tetrahedral reaction intermediate states**. *J Biol chem* 2003, **278**:49215-49222.
32. MacBeath G, Kast P, Hilvert D: **A small, thermostable, and monofunctional chorismate mutase from the archeon *Methanococcus jannaschii***. *Biochemistry* 1998, **37**:10062-10073.
33. Kudo F, Hosomi Y, Tamegai H, Kakinuma K: **Purification and characterization of 2-deoxy-*scyl-lo* inosose synthase derived from *Bacillus circulans*. a crucial carbocyclization einzyme in the biosynthesis of 2-deoxystreptamine-containing aminoglyoside antibiotoics**. *J antibiotics* 1999, **52**:81-88.
34. Yu Y, Bai L, Minagawa K, Jian X, Li L, Li J, Chen S, Cao E, Mahmud T, Floss HG *et al*: **Gene cluster responsible for validamycin biosynthesis in *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008**. *Appl Environ Microbiol* 2005, **71**:5066-5076.
35. Bonner CA, Jensen RA: **Upstream metabolic segments that support lignin biosynthesis**. In: *Lignin and Lignan Biosynthesis.* Edited by Lewis NG, Sarkanan S, vol. 697 (Amer Chem Soc Symp Series). Washington, D.C.: American Chemical Society; 1998: 29-41.
36. Singh S, Korolev S, Koroleva O, Zarembinski T, Collart F, Joachimiak A, Christendat D: **Crystal structure of a novel shikimate dehydrogenase from *Haemophilus influenzae***. *J Biol Chem* 2005, **280**(17):17101-17108.
37. Jensen RA, Morris P, Bonner C, Zamir LO: **Biochemical interface between aromatic amino acid biosynthesis and secondary metabolism**. In: *Plant Cell Wall Polymers: Biogenesis and Biodegradation.* Edited by Lewis NG, Paice MG, vol. 399 (Amer Chem Soc Symp Series). Washington, D.C.: American Chemical Society; 1989: 89-107.
38. Stenmark SL, Pierson DL, Glover FI, Jensen RA: **Blue-green bacteria synthesize L-tyrosine by the pretyrosine pathway**. *Nature* 1974, **247**:290-292.
39. Daugherty M, Vonstein V, Overbeek R, Osterman A: **Archaeal shikimate kinase, a new member of the GHMP-kinase family**. *J Bacteriol* 2001, **183**:292-300.
40. Quevillon-Cheruel S, Leulliot N, Meyer P, Graille M, Bremang M, Blondeu K, Sorel I, Poupon A, Janin J, van Tilbeurgh H: **Crystal structure of the bifunctional chorismate synthase from *Saccharomyces cerevisiae***. *J Biol Chem* 2004, **279**:619-625.
41. Lee AY, Karplus PA, Ganem B, Clardy J: **Atomic structure of the buried catalytic pocket of *Escherichia coli* chorismate mutase**. *J Am Chem Soc* 1995, **117**:3627-3628.
42. Chook YM, Ke H, Lipscomb WN: **Crystal structures of the monofunctional chorismate mutase from *Bacillus subtilis* and its complex with a transition state analog**. *Proc Natl Acad Sci USA* 1993, **90**:8600-8603.
43. Xue Y, Lipscomb WN, Graf R, Schnappauf G, Braus G: **The crystal structure of allosteric chorismate mutase at 2.2 Å resolution.** *Proc Natl Acad Sci USA* 1994, **91**:10814-10818.
44. Blanc V, Gil P, Bamas-Jacques N, Lorenzon S, Zagorec M, Schleuniger J, Bisch D, Blache F, Debussche L, Crouzet J *et al*: **Identification and analysis of genes from *Streptomyces pristinaespiralis* encoding enzymes involved in the biosynthesis of the 4-dimethylamino-*L*-phenylalanine**. *Mol Microbiol* 1997, **23**:191-202.

45. Künzler D, Sasso S, Gamper M, Hilvert D, Kast P: **Mechanistic insights into the isochorismate pyruvate lyase activity of the catalytically promiscuous PchB from combinatorial mutagenesis and selection**. *J Biol Chem* 2005, **280**:32827-32834.
46. Yoshikuni Y, Ferrin TE, Keasling JD: **Designed divergent evolution of enzyme function**. *Nature* 2006, **440**(7087):1078-1082.
47. Song J, Bonner CA, Wolinsky M, Jensen RA: **The TyrA family of aromatic-pathway dehydrogenases in phylogenetic context** *BMC Biology* 2005, **3**:13.
48. Mehta PK, Hale TI, Christen P: **Aminotransferases: demonstration of homolgy and division into evolutionary subgroups**. *Eur J Biochem* 1993, **214**:549-561.
49. Pittard AJ, Camakaris H, Yang J: **The TyrR regulon**. *Mol Microbiol* 2005, **55**:16-26.
50. Rijinen L, Bonneau S, Yvon M: **Genetic characterization of the major lactococcal aromatic aminotransferase and its involvement in conversion of amino acids to aroma compounds**. *App Environ Microbiol* 1999, **65**:4873-4880.
51. Gu W, Song J, Bonner CA, Xie G, Jensen RA: **PhhC is an essential aminotransferase for aromatic amino acid catabolism in *Pseudomonas aeruginosa***. *Microbiology* 1998, **144** 3127-3134.
52. Jensen RA, Calhoun DH: **Intracellular roles of microbial aminotransferases: overlap enzymes across different biochemical pathways**. *Crit Rev Microbiol* 1981, **8**:229-266.
53. Jensen RA, Stenmark SL: **The ancient origin of a second microbial pathway of *L*-tyrosine biosynthesis in prokaryotes**. *J Mol Evol* 1975, **4**:249-259.
54. Bonner CA, Jensen RA, Gander JE, Keyhani NO: **A core catalytic domain of the TyrA protein family: arogenate dehydrogenase from *Synechocystis***. *Biochem J* 2004, **382**:279-291.
55. Riepl RG, Glover GI: **Regulation and state of aggregation of *Bacillus subtilis* prephenate dehydratase in the presence of allosteric effectors**. *J Biol Chem* 1979, **254**(20):10321-10328.
56. Jimenez N, Gonzalez CF, Silva FJ: **Prephenate dehydratase from the aphid endosymbiont (*Buchnera*) displays changes in the regulatory domain that suggest its desensitization to inhibition by phenylalanine**. *J Bacteriol* 2000, **10**:2967-2969.
57. Xie G, Keyhani NO, Bonner CA, Jensen RA: **Ancient origin of the tryptophan operon and the subsequent dynamics of evolutionary change**. *Microbiol Mol Biol* 2003, **67**:303-342.