7.05 The Use of Subsystems to Encode Biosynthesis of Vitamins and Cofactors

Andrei L. Osterman, Burnham Institute for Medical Research, La Jolla, CA, USA

Ross Overbeek, Fellowship for Interpretation of Genomes, Burr Ridge, IL, USA

Dmitry A. Rodionov, Burnham Institute for Medical Research, La Jolla, CA, USA

7.05.1 The Goal

Our ability to acquire and annotate complete genomes has accelerated rapidly since the first completely sequenced genome became available in 1995.[1] Prior to that point in time, our understanding of the biochemical networks that sustain life were largely limited to a relatively few *model organisms*. This is largely true even for the most universal and the best-studied metabolic pathways, such as biosynthesis of the ubiquitous cofactors and their precursors (vitamins) covered in this volume. Hundreds of complete genomes are currently available for analysis, and the availability of thousands is just a few years away. This wealth of data for the first time opened an opportunity to practically assess the famous statement of Jacques Monod: "What is true for *E. coli* is true for the elephant." The expectation was that the identification of genes associated with pathways of interest (e.g., biosynthesis of cofactors) would allow us to establish the presence or absence and to reconstruct the details of these pathways across multiple diverse species. Many research groups who understood the power of comparative genome analysis for projecting the knowledge of genes and pathways from model organisms to others have asked the question "How can we organize the data from thousands of genomes to support analysis of specific areas of metabolism?" The initial efforts to address this question led to the establishment of metabolic reconstruction technology based on comparative analysis.[2,3] The development of the first genomic integrations connecting genes with formally encoded biochemical reactions and pathways have undoubtedly benefited the research community.[4–10] In this chapter, we briefly describe a technology for implementing one particular style of organization termed *the subsystems-based approach*.[11] The proven utility of this approach is illustrated here by examples from vitamin and cofactor biosynthesis. We discuss the encoding, projection, assessment of variation, and prediction of novel aspects of representative pathways. Although we largely limit the discussion of specific examples to one subsystem, 'Biosynthesis of Riboflavin' (vitamin B_2) and the derived cofactors, flavin mononucleotide (FMN) and dinucleotide (FAD), the same approach is applicable to many other cofactors described in this volume.

One goal of subsystems-based annotation is to offer researchers interested in a single biochemical process a 'summary' of exactly how the process is implemented in each of the sequenced genomes. This summary attempts to clarify the different variations one sees in the basic process, which genes play roles in the process, and which open questions remain. We believe that a well-organized collection of the genomic data becomes a framework to support research into the remaining open questions. A growing collection of subsystems capturing a substantial fraction of the Core Metabolic Machinery projected across hundreds of completely sequenced genomes, as well as the tools for its further expansion and curation is provided in The SEED database.[11] A status of The SEED subsystems (that are also available through the National Medical Pathogen Data Resource, NMPDR,[12]) encoding the biosynthesis of several major cofactors are listed in **Table 1**.

Table 1 Examples of vitamin and cofactor biosynthesis subsystems in The SEED database

#	Vitamin	Cofactors	Subsystem in SEED	Roles[a]	Genomes[b]	Genes[c]	Rare roles[d]	Core roles[e]
1	H	Biotin	Biotin biosynthesis	17	415	2780	6	8
2	B_5	CoA	Coenzyme A biosynthesis	18	720	6701	5	10
3	B_{12}	Cobalamine	Coenzyme B_{12} biosynthesis	61	201	4585	16	22
4		PQQ	Coenzyme PQQ synthesis	6	65	345	0	6
5	B_9/B_{11}	Folates	Folate biosynthesis	28	593	8231	7	13
6		Heme	Heme and siroheme biosynthesis	15	596	6474	0	11
7		Lipoate	Lipoic acid metabolism	3	655	1639	0	3
8	B_3	NAD(P)	NAD and NADP cofactor biosynthesis	32	747	7156	15	9
9	B_6	PLP, PMP	Pyridoxin (vitamin B_6) biosynthesis	9	578	2818	0	5
10	B_2	FMN, FAD	Riboflavin, FMN, and FAD metabolism	17	412	3492	7	8
11	B_1	Thiamin-PP	Thiamin biosynthesis	32	310	2971	11	6
12	K	Ubiquinone	Ubiquinone biosynthesis	15	490	4583	6	8
				253		51 775	29%	43%

[a] The number of distinct functional roles (isofunctional protein families) in a subsystem.
[b] The number of genomes that contain an operational variant of a given subsystem.
[c] The total number of genes associated with a set of functional roles in a subsystem (from all genomes counted in the column 'genomes').
[d] Roles that are present in <10% genomes within a subsystem.
[e] Roles that are present in >50% genomes within a subsystem.

7.05.2   More Precisely, *What Is a Subsystem*?

Subsystem-based annotation seeks to organize the genomic data relating to a small, well-defined set of 'functional roles' that make up a pathway. For metabolic pathways these collections of functional roles include mostly enzymes, sometimes enhanced with transporters and transcriptional regulators. We refer to this set of functional roles as a 'subsystem'. Each functional role is typically associated with a set of homologous genes (members of a single protein family) that implement this role in specific organisms. We create a 'populated subsystem' as a spreadsheet in which the columns represent functional roles, and the rows represent specific genomes. Each cell in the spreadsheet contains the genes that encode proteins that implement the functional role in a specific genome.

A number of factors determines the scope, phylogenetic coverage, accuracy, and level of completion of each subsystem. Of course, the extent of the experimental evidence, the curator's depth of knowledge in the respective area of metabolism, and the stage of the curator's analysis (usually reflected in the notes attached to every subsystem) constitute the main factors. Overall, the small collection of 12 subsystems shown in **Table 1** includes >50 000 annotated genes spanning ∼750 analyzed diverse genomes (among them ∼90% are bacterial, with a relatively small fraction of archaeal and only about a dozen of representative eukaryotic genomes). Briefly looking at **Table 1**, one may notice substantial variations in the number of functional roles (from three in lipoate metabolism up to 61 in coenzyme B_{12} biosynthesis, with a more typical size being ∼15–25 roles), as well as in the number of genomes containing an operational variant of each pathway (from 65 to over 700). Functional roles included in each subsystem may be split into two main categories depending on their occurrence in the genomes. On average ∼40% of the functional roles constitute the core of a subsystem, and they are rather universal (present in >1/2 of the analyzed genomes), whereas ∼30% of the included roles correspond to less ubiquitous and species-specific aspects of the pathway (present in <1/10 of all genomes). A subsystem core often includes the most conserved and universal enzymes, whereas transcriptional regulators, uptake transporters, and rare alternative forms of enzymes frequently constitute a labile periphery of the subsystem.

To illustrate the concept of a subsystem, we will consider the subnetwork of biochemical transformations that convert GTP and ribulose-5-phosphate first into vitamin B_2, and then into FMN and FAD cofactors. This subsystem was chosen due to a combination of reasons: (1) the high level of biochemical understanding (as reflected in Chapter 7.02); (2) the ubiquitous and essential nature of flavin cofactors in the three kingdoms of life; (3) a relatively simple topology, which includes the *de novo* biosynthesis of B_2 (replaced by salvage in some species) followed by its two-step conversion into FMN and FAD cofactors; and (4) broad conservation of most biochemical reactions and enzymes combined with some interesting variations between species that allow us to illustrate the application of comparative genomic techniques.

Figure 1 provides a simplified subsystem diagram that schematically shows major intermediary metabolites (depicted by ovals with abbreviations or Roman numerals I through VII) and enzymes (shown as rectangles with abbreviations) known to catalyze the respective reactions (shown by arrows) in at least some of the characterized species. Similar diagrams are broadly used by some of the pathway-oriented databases (such as KEGG[5]) to capture all possible reactions and pathways within a subnetwork, whereas other resources (e.g., MetaCyc[6]) prefer to display organism-specific pathway diagrams. As in KEGG pathway maps, subsystem diagrams in The SEED database provide the ability to highlight the functional roles present in any selected organism. This allows the user to make a preliminary assessment of which of the possible fluxes (shown by thick gray arrows in **Figure 1**) or metabolic scenarios[13] are present in the organism of interest. This depiction of the genes from specific organisms can be used to clearly reveal incomplete functional variants of pathways[14] containing gaps (*missing genes*) and inconsistencies (*out-of-context genes*) that reflect incomplete knowledge or annotation errors. Although some of the revealed problems may be reconciled by similarity-based annotation techniques (e.g., by finding a gene candidate with a lower homology score or by the detailed analysis of gene grouping in a family of paralogues), others may not be effectively addressed without application of additional genome context analysis techniques. Application of these techniques, primarily clustering of functionally related genes on the prokaryotic chromosome,[15] analysis of protein fusion events,[16] co-occurrence profiles[17] and shared regulatory sites,[18] substantially improves the quality and consistency of genomic annotations. Such analysis can lead to accurate prediction of novel gene candidates and other conjectures about pathways that can be further tested by focused experiments (see Osterman and Overbeek[19]
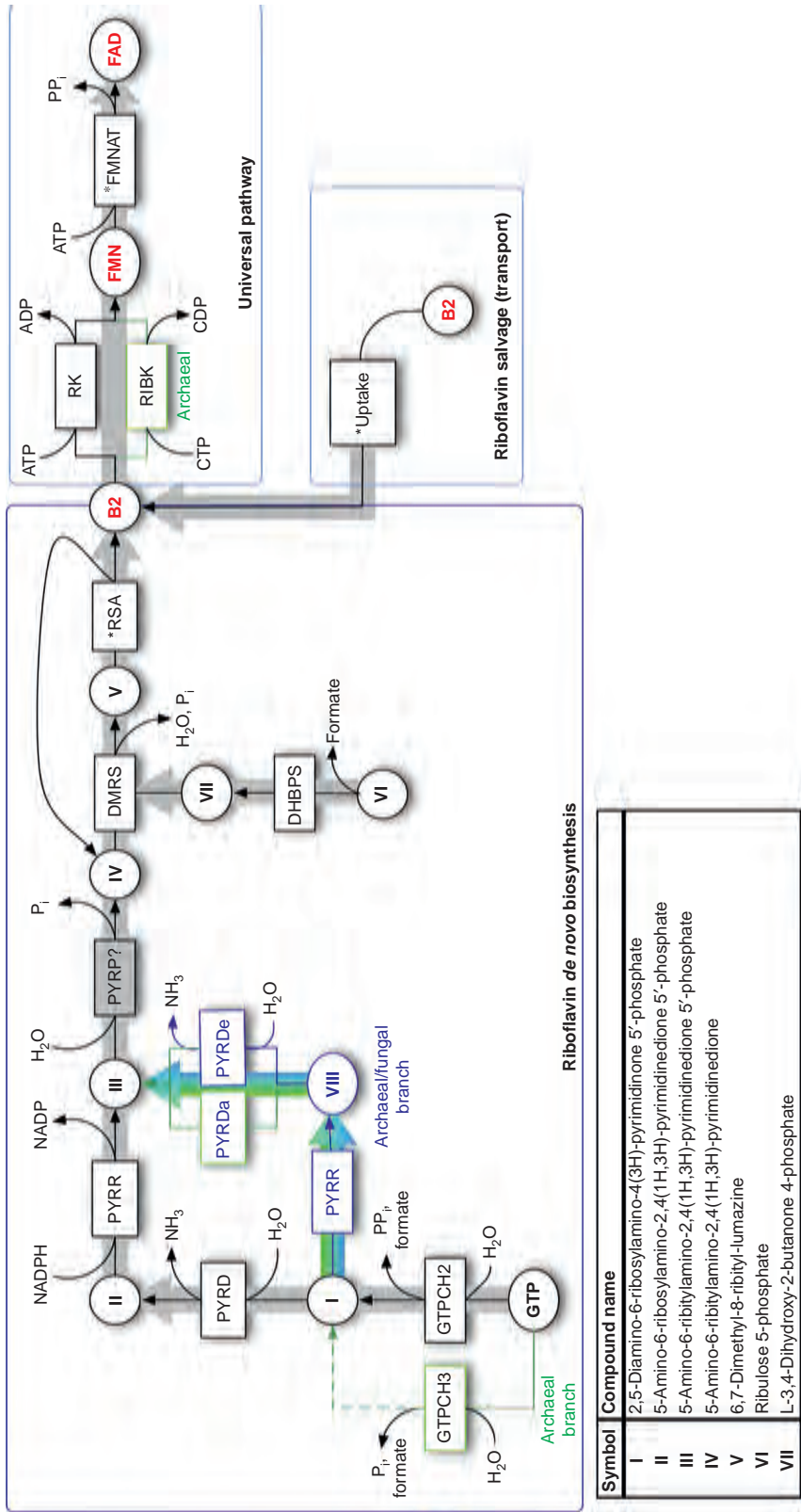
Symbol	Compound name
I	2,5-Diamino-6-ribosylamino-4(3H)-pyrimidinone 5′-phosphate
II	5-Amino-6-ribosylamino-2,4(1H,3H)-pyrimidinedione 5′-phosphate
III	5-Amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione 5′-phosphate
IV	5-Amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione
V	6,7-Dimethyl-8-ribityl-lumazine
VI	Ribulose 5-phosphate
VII	L-3,4-Dihydroxy-2-butanone 4-phosphate

Figure 1 A subsystem diagram of the biosynthesis of riboflavin, FMN, and FAD. Enzymes are indicated by rectangles with abbreviations as in **Table 3**. Pathway intermediates and products are shown in circles by standard abbreviations (GTP, FMN, FAD, and B₂ for riboflavin) or roman numerals enlisted in the inset. Major fluxes are outlined by thick arrows (gray, for bacterial and universal routes and colored, for the archaeal/fungal branch). Subsets of roles marked by '*' combine alternative (nonorthologous) forms of genes (protein families) that play equivalent roles in the subsystem.