



NMPDR 

The Project to Annotate 5000 Genomes (up from 1000)


Ross Overbeek
Fellowship for Interpretation of Genomes

April 2007



Topics Covered


- **FIG, the NMPDR and the Project to Annotate 1000 Genomes**
- What do we mean by **Annotations?**
- **Subsystems-Based Annotations**
- **FIGfams**
- **The RAST Server**



www.nmpdr.org

The Fellowship for Interpretation of Genomes

- A framework for continuing a cooperative effort to "**characterize unicellular life**"
- The software project that began with "**the SEED**"
- The **Project to Annotate 1000 Genomes** as conceived of in a bar in Albuquerque (and initiated in late 2003)
- The **NMPDR** as an initial project based on the SEED.



www.nmpdr.org

What Do We Mean by “Good” Annotations?

- Identify genes accurately (at least CDSs and RNAs).
- Supply consistent and accurate gene functions
- Attach gene functions to abstract functional roles that make up subsystems
- Create functional modules from instances of subsystems. These lay the foundation for the development of consistent metabolic reconstructions (and models).



www.nmpdr.org



Strategies for Annotation

1. Annotate a genome at a time
 - a) If done by walking down the genome, genes are treated “out of context”
 - b) Pretty much assures that most annotations will be done by someone who may be a skilled annotator, but has no special expertise in the vast majority of genes annotated
2. Annotate components of the cellular machinery across the entire set of hundreds of genomes
 - a) Allows the annotator to become an expert in known variations of the “subsystem”
 - b) Allows a higher level of consistency



www.nmpdr.org



Metrics for Annotation

1. Consistency: it is desirable that two genes playing identical roles in distinct genomes are assigned the same function
 - How can this be measured?
2. Accuracy: a number of issues are relevant
 - How serious is over-precision?
 - How would one measure “accuracy”?



www.nmpdr.org



Critical Properties of Gene Functions

1. It must be possible to **automatically determine whether or not two gene functions are identical**
2. **Distinct functions have distinct names.** For example, hypothetical proteins are grouped into protein families with unique identifiers.
3. It must be **easy to rename families.**



www.nmpdr.org



What Is a Populated Subsystem?

- A “subsystem” is a set of abstract functional roles
- A “populated subsystem” is a subsystem with an attached spreadsheet connecting specific genes to functional roles in a set of organisms
- The “populated subsystem” should be viewed as the basic unit of annotation



www.nmpdr.org

Subsystem: Chorismate Synthesis

Functional Roles		
Column	Abbrev	Functional Role
1	DAHPS	Phospho-2-dehydro-3-deoxyheptonate aldolase (EC 2.5.1.54)
2	DHQS	3-dehydroquinate synthase (EC 4.2.3.4)
3	DHQQH	3-dehydroquinate dehydratase (EC 4.2.1.10)
4	SDH	Shikimate 5-dehydrogenase (EC 1.1.1.25)
5	SK	Shikimate kinase (EC 2.7.1.71)
6	PSCVT	3-phosphoshikimate 1-carboxyvinyltransferase (EC 2.5.1.19)
7	CS	Chorismate synthase (EC 4.2.3.5)
8	ASI	Alternative step 1 of chorismate biosynthesis

Name functional roles for the chosen Subsystem

Subsets of Roles
Subset Includes These Roles

Define subsets



www.nmpdr.org



FIGfams

- A FIGfam is a protein family such that:
 - All of the members perform the same function
 - All of the members are end-to-end similar (e.g., they share a common domain structure)
- How are they built?
 1. All genes from a single column of a subsystem that are globally similar go into the same FIGfam
 2. Two genes from closely-related genomes that "clearly correspond" go into the same FIGfam
 3. These two sources of "sets" get merged
 4. **Reliable FIGfams:** those that contain members from a subsystem column



www.nmpdr.org



FIGfams: (continued)

- How many exist
 - ~6500 reliable FIGfams
 - ~90,000 far less reliable, often small, FIGfams
- How are they used:
 - To attain consistency and accuracy on central machinery
 - To attain consistency on closely-related genomes
 - A strategy for improvement of overall annotations (projection of subsystems)



www.nmpdr.org



Evaluation of NMPDR Annotations

1. Consistency (by FIG using FIGfams): 99%
2. Completeness (by TIGR): 62% (connections to EC and GO)
3. Accuracy (by TIGR): 98%



www.nmpdr.org



RAST Server: Procedure

1. Register (user)
2. Submit genome (user)
3. Genome annotation process (server)
4. Quality controls (server)
5. Evaluation / Viewing (user)
6. Download results (user)
7. Delete the genome from the server (user)



www.nmpdr.org



Genome annotation process

- Automated process consisting of:
 - Gene calling
 - Initial annotation of function
 - Initial metabolic reconstruction
- Process takes 1-7 hours depending on size and complexity of the genome
- ~20 genomes per day



www.nmpdr.org



RAST Approach

1. Determine phylogenetic neighborhood
2. Use subsystems in closest neighbors to search for active variants in the new genome
3. Then try all subsystems that have not been considered
4. Finally, use standard techniques for remaining genes

Steps 2 and 3 produce "reliable annotations".
29% of the B.meg genes fell into this category.



www.nmpdr.org