# **LUCApedia Search Tutorial**

#### **General search strategy**

LUCApedia uses annotation from the Uniprot database to superimpose multiple datasets related to ancient life. We do not predict annotations ourselves. This has the effect that a given instance of a protein may not be ineligible for certain datasets even if other members of its protein family are. For example, Glutaminyl-tRNA synthetase is probably very ancient and should have hits in multiple datasets within LUCApedia. The *Escherichia coli* version of the protein has hits to six of the LUCApedia datasets. The *Nosema ceranae* version of the protein only has two hits to LUCApedia datasets.

The *N. ceranae* version of the protein is ineligible for mapping to the Harris *et al.* (2003) and Mirkin *et al.* (2003) datasets because these studies are based on the COG database and *N. ceranae* proteins are not included in the COG database. Similarly, it is ineligible for mapping to the Yang *et al.* (2005) and Wang *et al.* (2007) datasets because these studies are based on the SCOP database and the *N. ceranae* version of the protein is not included in the SCOP database.

We recommend searches that use a general protein name that represents a family of proteins, because it is the protein family that is ancient rather than the instance of a protein within an extant organism. Note that we rank the Uniprot entries by the number of hits to datasets within LUCApedia.

					Glutaminyl-tRNA synthetase								
				exact protein name part		protein name Unipro		rot ID					
												Next →	
displaying	first 50 records												
displaying	illot ou records												
protein name	Uniprot ID <sup>1</sup>	Harris et al., 2003 <sup>2</sup> (COG ID <sup>3</sup> )	Mirkin et al., 2003 <sup>4</sup> (COG ID <sup>3</sup> )	Delaye et al., 2005 <sup>5</sup> (Pfam ID <sup>6</sup> )	Yang et al., 2005 <sup>7</sup> (SCOP superfamily ID <sup>8</sup> )	Wang et al., 2007 <sup>9</sup> (SCOP fold ID <sup>8</sup> )	Srinivasan and Morowitz, 2009 <sup>10</sup> (Enzyme commission code <sup>11</sup> )	Ribozyme function (Enzyme commission code <sup>11</sup> )	Nucleotide cofactor usage	Amino acid cofactor usage	Iron sulfur cofactor usage	Zinc cofactor usage	
Glutaminyl- tRNA synthetase	SYQ_ECOLI	COG0008	COG0008	tRNA- synt_1c_C, tRNA- synt_1c	c.26.1	c.26, b.53	-	6.1.1.18	-	-	-	-	
Glutaminyl- tRNA synthetase	SYQ_XYLFA	COG0008	COG0008	tRNA- synt_1c_C, tRNA- synt_1c	-	-	-	6.1.1.18	-	-	-	-	
Glutaminyl- tRNA synthetase	SYQ_YEAST	COG0008	COG0008	tRNA- synt_1c_C, tRNA- synt_1c	-	-	-	6.1.1.18	-	-	-	-	
Glutaminyl- tRNA synthetase	SYQ_NEIMA	-	-	tRNA- synt_1c_C, tRNA- synt_1c	-	-	-	6.1.1.18	-	-	-	-	
Glutaminyl- tRNA synthetase	SYQ_NOSCE	-	-	tRNA- synt_1c_C, tRNA- synt_1c	-	-	-	6.1.1.18	-	-	-	-	
Glutaminyl- tRNA synthetase	E7R1R4_PICAD	-	-	tRNA- synt_1c_C, tRNA- synt_1c	-	-	-	-	-	-	-	-	

Searches for protein names that are specific to an organism are less likely to yield the full extent of results. For example, the locus number of the *N. ceranae* version of Glutaminyl-tRNA synthetase is NCER\_100619. A search for NCER\_100619 will yield only one entry that just so happens to have fewer hits to LUCApedia datasets.

		Ple	ease en	NCER_1006		Uniprot		ed "entry nam Search	e"):			
				exact protein		partial protein name		prot ID				
1 records fou	nd											
protein name	Uniprot ID <sup>1</sup>	Harris et al., 2003 <sup>2</sup> (COG ID <sup>3</sup> )	Mirkin et al., 2003 <sup>4</sup> (COG ID <sup>3</sup> )	Delaye et al., 2005 <sup>5</sup> (Pfam ID <sup>6</sup> )	Yang et al., 2005 <sup>7</sup> (SCOP superfamily ID <sup>8</sup> )	2007 <sup>9</sup> (SCOP	Srinivasan and Morowitz, 2009 <sup>10</sup> (Enzyme commission code <sup>11</sup> )	Ribozyme function (Enzyme commission code <sup>11</sup> )	Nucleotide cofactor usage	Amino acid cofactor usage	Iron sulfur cofactor usage	Zinc cofacto usage
NCER_100619	SYQ_NOSCE	-	-	tRNA- synt_1c_C, tRNA- synt_1c	-	-	-	6.1.1.18	-	-	-	-

## Can't find a protein?

LUCApedia uses all of the names and synonyms provided by the Uniprot database for each of its entries. This list is comprehensive, but not exhaustive. If you are having trouble finding a protein of interest, we recommend the following steps.

- 1) Try using the partial protein name search option rather than the exact protein name search option.
- 2) Try using our browse option to find the protein name and its exact spelling.
- 3) Look for synonyms of your protein name in other databases such as <u>Entrez</u>, <u>KEGG</u>, and <u>BioCyc</u>.

#### Are you an advanced user?

If you are interested in using LUCApedia to do large scale analyses, we recommend that you download the flat files rather than using the web server search function. Flat files and descriptions of their contents can be found on the <u>download page</u>.

## **Still have questions?**

Please contact: Aaron Goldman

106A Guyot Hall

Princeton, NJ 08544-1003 Email: adg@princeton.edu Phone: (609) 258-6724