P4-4
S1-2
(Oral)

Automatic protein clustering as a basis of automatic annotation

Naoki Sato: University of Tokyo, Graduate School of Arts and Sciences

<Summary>
Development of new generation sequencers enabled genome sequencing feasible for every organism in a laboratory. A typical data flow of de novo seuqencing includes (1) assembly of sequence reads, (2) estimation of open reading frames, (3) annotation of proteins, and (4) finding RNA genes. The annotation is normally performed by BLASTP searches against several different databases. However, it is usually hard to find a plausible annotation by just looking at the results of BLASTP searches.
Here I propose a potentially automatic method of annotation that exploits automatic protein clustering using the software GCLUST, which estimates proper similarity threshold for each list of homologs using 'entropy-optimized organism count' method (Sato 2009). The software has been used to construct a homolog database including both prokaryotic and eukaryotic proteins (http://gclust.c.u-tokyo.ac.jp/). For use in genome annotation, we need de novo clustering including many genomes of related organisms as well as genomes of representative organisms. Application of protein clustering in the annotation in *Arthrospira platensis* was the first successful case (Fujisawa et al. 2010). I present here results of protein clustering of total predicted proteins in two draft genomes of cyanobacteria along with total predicted proteins of 41 cyanobacteria available at NCBI. For each of the resultant protein clusters, an alignment and a phylogenetic tree were also prepared for assistance in functional annotation. The quality of alignments was evaluated by counting ill-aligned proteins (missing N- or C-terminus, or insertion/deletion), which was 4-13% of total predicted proteins in most cyanobacterial genomes. Annotation may be automated by extracting significant key words already assigned for member proteins of clusters or by comparison with reference protein clusters.

Nature Precedings : doi:10.1038/npre.2010.5086.1 : Posted 22 Oct 2010

1. Introduction

Current way of genome sequencing

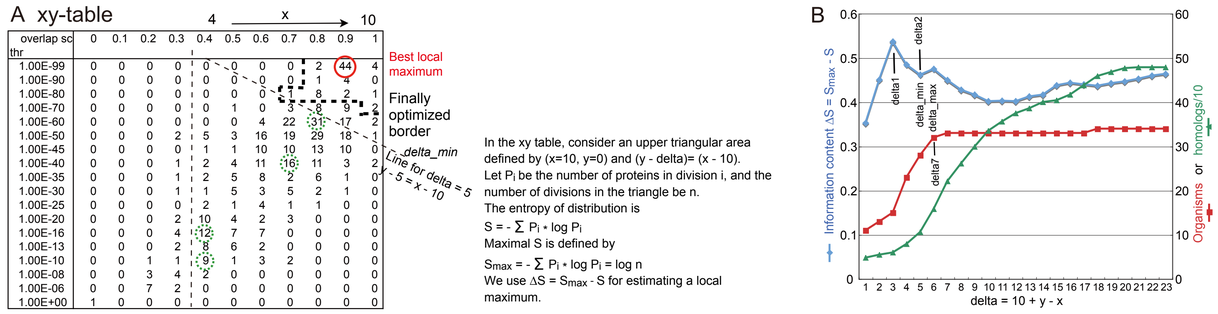
- DNA isolation from bacterial cells
- Library construction
- Sequencing (454 etc)
- Assembly (newbler, MIRA)
- Annotation pipeline (MiGAP：Sugawara & Kurokawa labs.)
  - ORF estimation（MetaGeneAnnotator)
  - RNA genes estimation (tRNAscan-SE, RNAmmer)
  - BLAST (COG, RefSeq, TrEMBL)：to be improved by all-against-all BLASTP and automatic clustering with Gclust → *de novo* ortholog clusters
  - Automatic annotation
  - N-terminal correction

3. Automatic protein clustering with Gclust software

Entropy-optimized organism count (EOOC) method

Problem: All previous clustering of proteins used a simple threshold value such as E = 1.0 x 10^{-6} (criterion may be other parameters), but similarity of proteins is very different in different protein families: *eg*. PsaA: 10^{-150}, PsbO: 10^{-45}, PsbL: 10^{-10}. Use of a single threshold should produce unusually large clusters containing unrelated proteins and divergent paralogs.

Solution: For each list of protein similarity data (may include two variables), organisms are counted from the top. This is justified by the fact that orthologs are usually found near the top, and then, paralogs, then partially similar proteins. However, there are many different cases. Entropy of distribution is useful in estimating a proper threshold. In the actual implementation in Gclust, the two measures are considered to obtain orthologs and highly related paralogs.



Other merits of the Gclust software:

1. Automatic domain identification based on homology region data of BLAST. This is used for excluding multidomain proteins.
2. Automatic identification of transit peptides based on domain identification.
3. Comparison of both prokaryotic and eukaryotic proteins in a single dataset. This is the reason why Gclust is the only software that can analyze proteins of endosymbiotic origin.
4. A powerful replacement of COG. Gclust clusters are suitable for annotation of data from new generation sequencer.

5. Prospects

For annotation of better quality

- Proposal:

  Annotation of a new genome should be performed in the framework of related genomes
- This process is assisted by automatic clustering of all proteins in related genomes
- N-terminal correction will give better alignment
- Consensus annotation may be easy

Reference
Sato, N. (2009) Gclust: trans-kingdom classification of proteins using automatic individual threshold setting. *Bioinformatics* 25: 599-605

2. Revision of functional categories

Problems of functional categories of COG and a revision

COG is based on clustering of proteins of 61 organisms, which are divided into 14 groups.

Only ortholog groups shared by ≥ 3 groups are assigned COG number

Category 'Energy production and conversion' does not include photosynthesis

Difficulty in assignment of transporter and DNA-binding proteins to a correct functional category

New functional categories of proteins

This classification is implemented in CyanoClust database.
Number of clusters is shown for each sub-category.

Main category	Sub-category	Clusters
gene expression		
	genome structure	11
	replication	10
	repair, recombination, modification and nuclease	23
	transcription	13
	ribosome, translation	111
	RNA metabolism	13
	sorting	11
	processing and degradation	14
	regulation	12
	signal transduction	12
	stress response and chaperon	24
metabolism		
	photosystem	35
	respiration	15
	hydrogenase	5
	inorganic carbon	6
	ATP synthase	10
	central metabolism and sugar	41
	lipid	32
	porphyrin, heme, cytochrome, pigment	38
	phosphorus and sulfur	11
	nitrogen and amino acid	93
	nucleotide	33
	cofactor biosynthesis	60
	other metabolism	24
cellular structure		
	extracellular matrix	28
	cell division	13
	transport and membrane	30
unclassified	unclassified	4
	hypothetical	248
Total		980

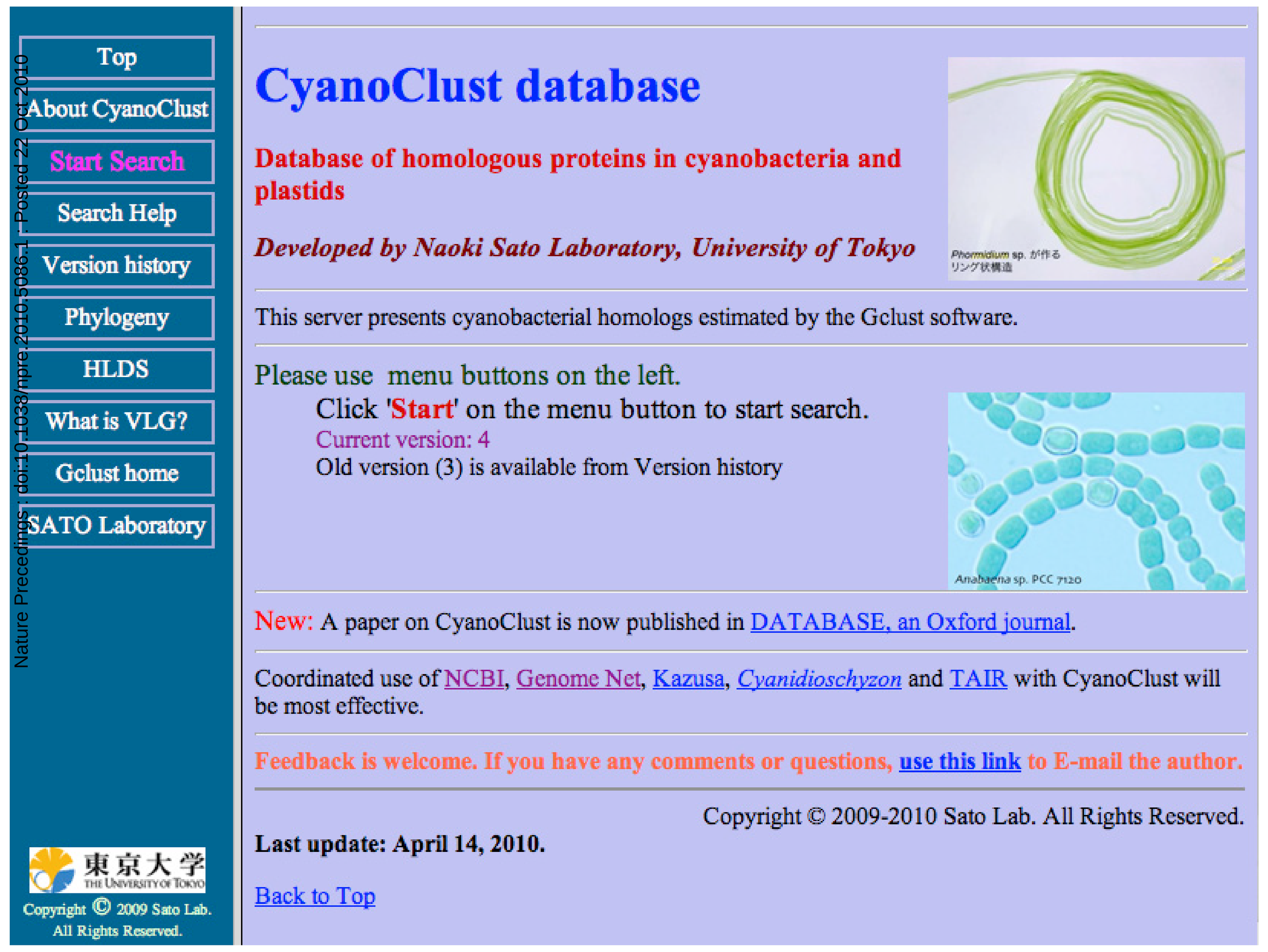
Annotation of universally (at least within a phylum) conserved proteins

Each of the 980 conserved protein clusters in 41 cyanobacteria have been given a biologically correct annotation (see above).

This annotation can be transferred to a new cluster constructed for (new + 41 species).

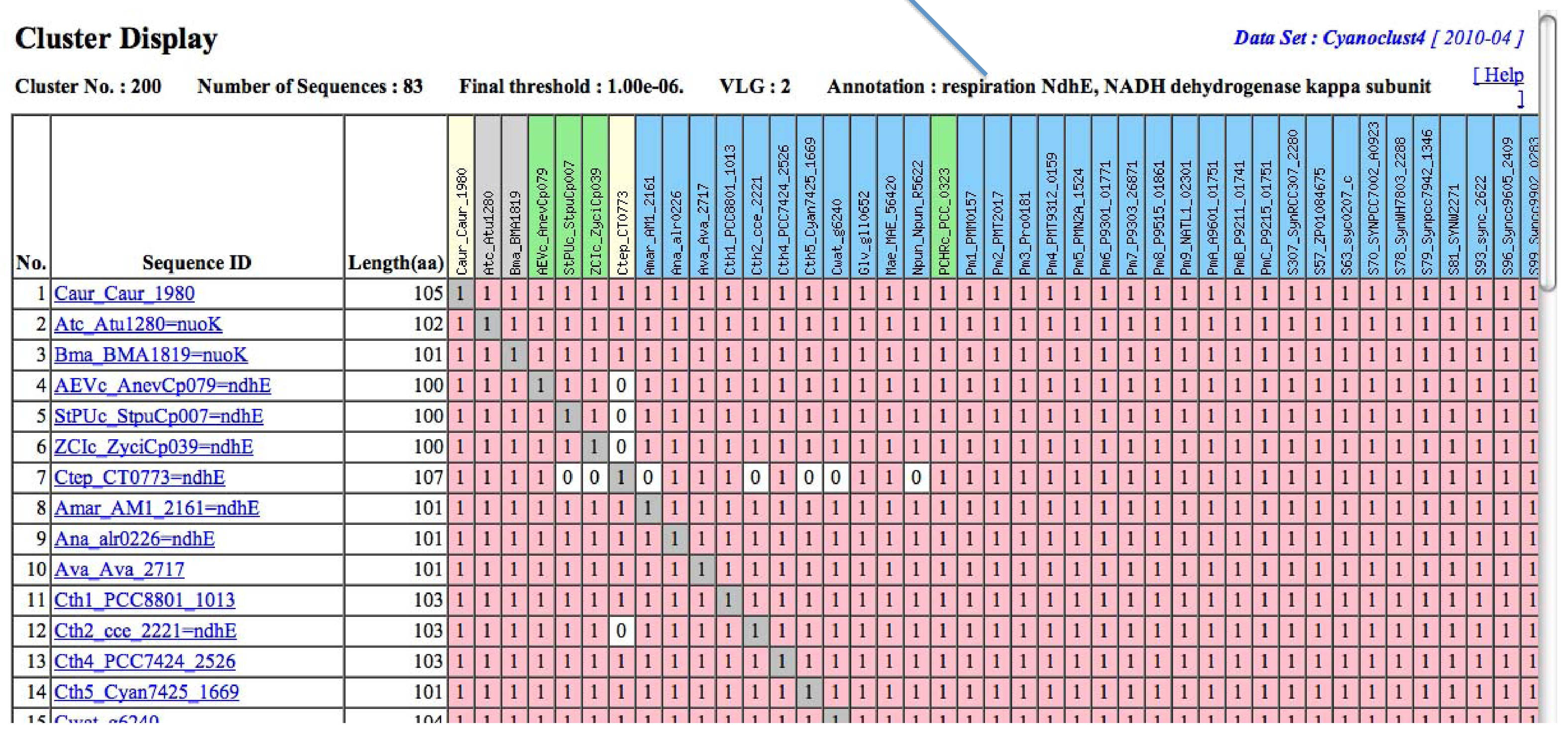
4(B) Improvement of annotation

CyanoClust is a database of ortholog groups of cyanobacteria



Annotation for universally conserved proteins is implemented in CyanoClust

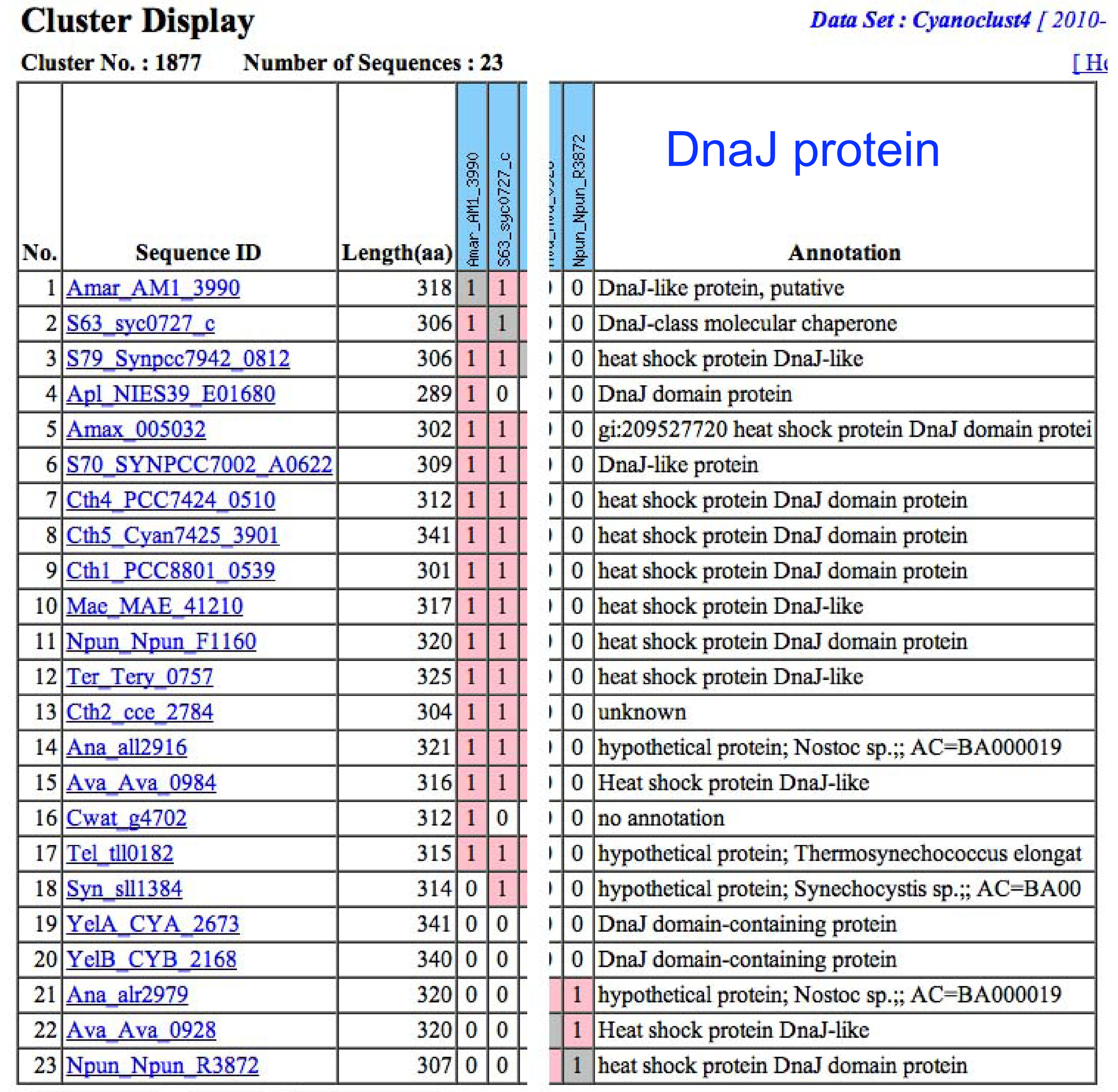
Annotation : respiration NdhE, NADH dehydrogenase kappa subunit



Cluster-based annotation is useful in avoiding 'Inherited strange annotation'

- Some annotations are inherited from those of other genomes based on unreliable homology or improper biological knowledge
- Cluster-based annotation is not susceptible for such inappropriate inheritance of annotation, even though individual annotations (given for original databases) may be variable or sometimes unreliable

An example of consensus-based annotation



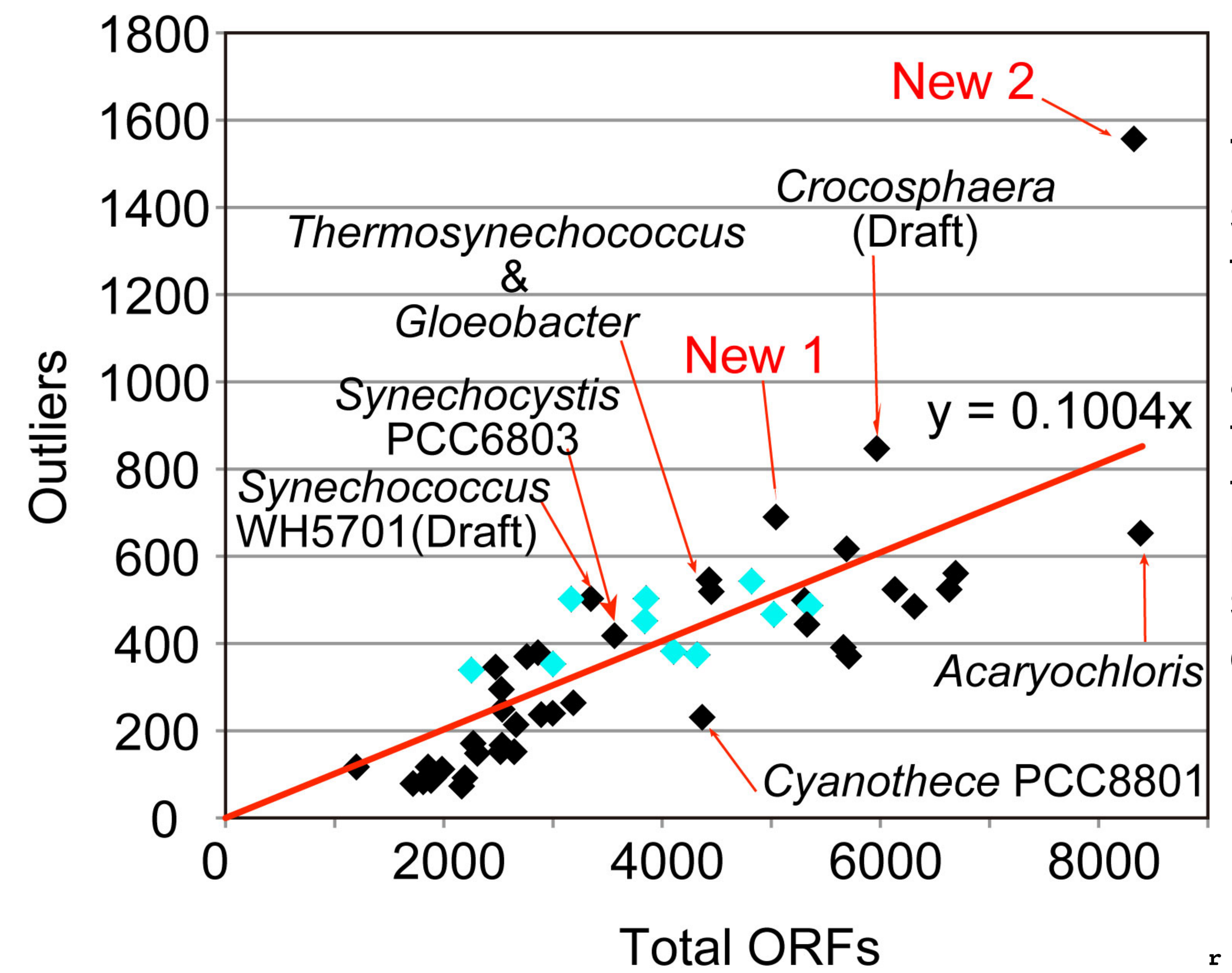
Cluster Display — Data Set : Cyanoclust4 [2010-

Cluster No. : 1877 Number of Sequences : 23

DnaJ protein

No.	Sequence ID	Length(aa)	Amar_AM1_3990	S63_syc0727_c	Npun_Npun_R3872	Annotation
1	Amar_AM1_3990	318	1	1	0	DnaJ-like protein, putative
2	S63_syc0727_c	306	1	1	0	DnaJ-class molecular chaperone
3	S79_Synpcc7942_0812	306	1	1	0	heat shock protein DnaJ-like
4	Apl_NIES39_E01680	289	1	0	0	DnaJ domain protein
5	Amax_005032	302	1	1	0	gi:209527720 heat shock protein DnaJ domain protei
6	S70_SYNPCC7002_A0622	309	1	1	0	DnaJ-like protein
7	Cth4_PCC7424_0510	312	1	1	0	heat shock protein DnaJ domain protein
8	Cth5_Cyan7425_3901	341	1	1	0	heat shock protein DnaJ domain protein
9	Cth1_PCC8801_0539	301	1	1	0	heat shock protein DnaJ domain protein
10	Mae_MAE_41210	317	1	1	0	heat shock protein DnaJ-like
11	Npun_Npun_F1160	320	1	1	0	heat shock protein DnaJ domain protein
12	Ter_Tery_0757	325	1	1	0	heat shock protein DnaJ-like
13	Cth2_cce_2784	304	1	1	0	unknown
14	Ana_all2916	321	1	1	0	hypothetical protein; Nostoc sp.;; AC=BA000019
15	Ava_Ava_0984	316	1	1	0	Heat shock protein DnaJ-like
16	Cwat_g4702	312	1	0	0	no annotation
17	Tel_tll0182	315	1	1	0	hypothetical protein; Thermosynechococcus elongat
18	Syn_sll1384	314	0	1	0	hypothetical protein; Synechocystis sp.;; AC=BA00
19	YelA_CYA_2673	341	0	0	0	DnaJ domain-containing protein
20	YelB_CYB_2168	340	0	0	0	DnaJ domain-containing protein
21	Ana_alr2979	320	0	0	1	hypothetical protein; Nostoc sp.;; AC=BA000019
22	Ava_Ava_0928	320	0	0	1	Heat shock protein DnaJ-like
23	Npun_Npun_R3872	307	0	0	1	heat shock protein DnaJ domain protein

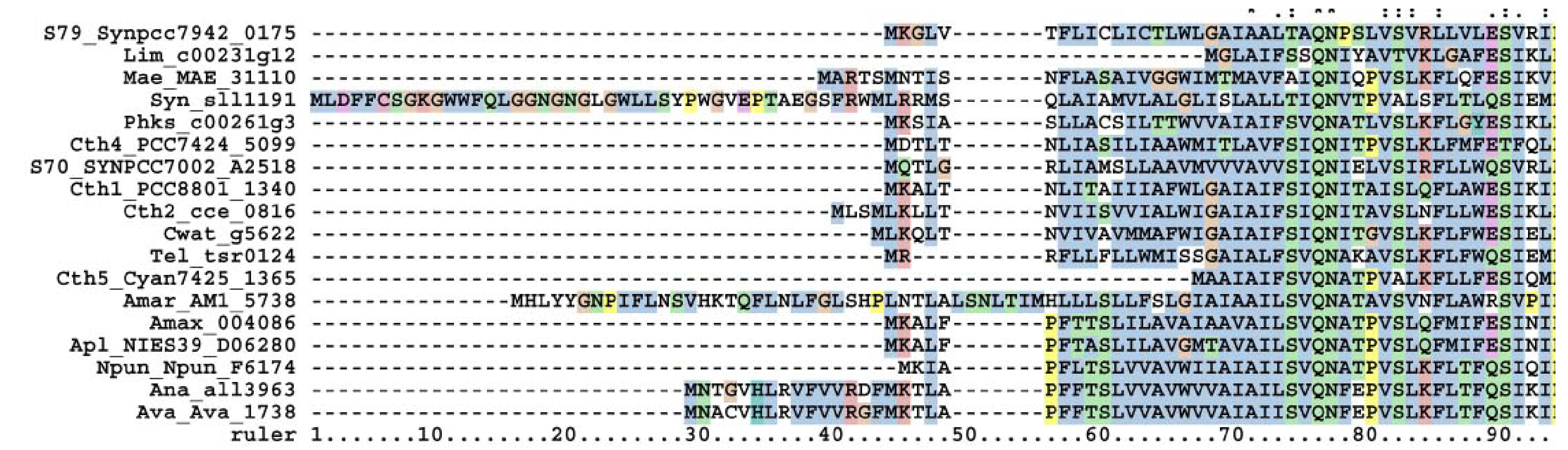
4(A) Improvement of N-termini



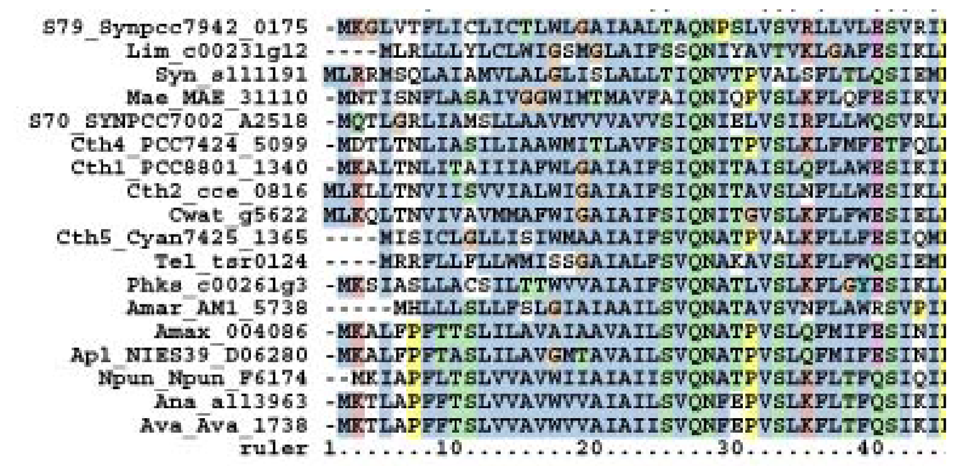
The N-termini can be improved by comparing sequences within a cluster. Outliers as defined as proteins having a protruding or lacking N-terminus are about 10% of total ORF of a genome, but are significantly larger than this average in draft sequences. The N-terminus problem exists also in published genomes. Many of them can be improved by inspecting the N-terminal nucleic acid sequence.

```
      r of sequences: N0 = 18.
 Start position: av0 = 39.500 std0 = 14.927.

 ID              start  N1  av1     std1
Lim_c00231g12      67    2  66.500  0.500

>Lim_c00231g12
-----------------------------------------------------------------MGLAIFSS
QNIYAVTVKLGAFESIKLPLGLVLIFCAGLGAIIMTLFMGFAQKSIQFSFPSIPKFNTSSPRKSFQKSPTKPQKT
SSNQRSANPNNKKEQESRDSFDDDWDDDWG--------------

Hypothetical translation of the upstream sequence
                          (Expected position of init. codon)
----------------------------------------vvvvvvvvv
RFTQRmRQAIIENRFVEEFGHYlTDQKPSTTSANFH*IIKRSSQLSFLK*T*NmlRLLLYLClWIGS
(Possible init. codons are marked as m, v, and l.)

>Lim_c00231g12 revised
-----------------------------------------------------mlRLLLYLClWIGSMGLAIFSS
QNIYAVTVKLGAFESIKLPLGLVLIFCAGLGAIIMTLFMGFAQKSIQFSFPSIPKFNTSSPRKSFQKSPTKPQKT
SSNQRSANPNNKKEQESRDSFDDDWDDDWG--------------
```

Before correction



After correction



Published genomes can also be improved.

Organism	Total ORFs	Outliers	Selected for N-terminal correction	Remaining outliers
Synechocystis sp. PCC 6803	3564	418 (11 %)	199	86
Anabaena sp. PCC 7120	6132	524 (8 %)	189	70
Arthrospira platensis	6630	524 (7 %)	86	38
New cyano 1	5043	690 (13 %)	190	85
New cyano 2	8323	1557 (18 %)	243	112