The Minimal-Gene-Set

-Kapil Rajaraman(rajaramn@uiuc.edu)

PHY498BIO, HW 3

The number of genes in organisms varies from around 480 (for parasitic bacterium *Mycoplasma genitalium*) to the order of 100,000 for eukaryotes. The question is: What is the minimum number of genes needed for a cell to sustain itself? These genes would be operating under maximum favorable conditions (all essential nutrients would be available, and there would be no external stress like temperature/pressure jumps). The motivation behind the study was to see if comparative genomics could be combined with biochemical and genetic data to determine this minimum, the knowledge of which would be useful in determining the essential genes in almost all species. In this essay, we discuss Eugene Koonin's work in the field [1].

Initial studies

It is clear that the upper bound on the number of genes is 480. The lower bound would be given by the number of genes necessary for the basic operations of the cell like transcription, replication, repair, as well as membrane properties and transport systems.

The complete sequencing of parasitic bacteria *Haemophilus influenzae* and *M. genitalium* was the first step to a comparative genomic approach to the minimal-gene-set issue. The usefulness of these two species is because (1) unlike multicellular organisms, unicellular organisms do not normally take up proteins from their environment, and the essential properties must be encoded in the genome, and (2) these two species belong to phylogenetically different groups of parasitic bacteria, and so the genes shared by the two genomes are likely to be essential.

The study showed 240 direct counterparts or orthologs. However, these 240 genes were not sufficient in themselves to account for all the metabolic processes. It was, therefore, necessary to include the possibility of the same function being performed by unrelated or distantly related proteins due to nonorthologous gene displacement (NOD). For these two bacteria, the NOD cases comprised around 5% of the 256 gene-set.

Clusters of orthologous genes (COG) approach

This approach is based on the notion that any group of at least three proteins from distant genomes that are more similar to each other than to any other proteins from the same genomes most probably belong to a family of orthologs. This approach entails finding all the triangles of best hits from the complete matrix of pairwise comparisons between proteins encoded in the analysed set of genomes and then merging the triangles with a common side to form the complete orthologous families. Additional searches are performed using PSI-BLAST to add weakly conserved proteins that have been missed by the first step.

The current COG collections have the following features:

(a) 55%-83% of the proteins encoded in the bacterial and archaeal genomes belong to COGs (Table 1). This implies that a good number of genes are conserved for these species through evolution.
(b) The COGs are, for the most part, non-ubiquitous, i.e, most of the COGs only cover a few clades. Among the 2112 COGs in the collection of patterns (of proteins represented by COGs) at the time this paper was written, as many as 1234 unique patterns were seen. This is due to clade-specific gene loss and horizontal gene transfer.

Due to this non-ubiquitousness, the role of NOD is more important than assessed in the initial study. The original minimal gene-set members' status after the COG assessment is shown in Table 2.

Table 2 also shows a group of minimal-set members that are conserved in bacteria. These genes code essential functions, but archaea and eukaryotes have evolved unrelated implementations of these functions. Namely, this is a manifestation of NOD on a larger scale.

It is also seen that the different categories of proteins have different phylogenetic patterns. This is true of the minimal gene set as well as the full COG collection. Components of the translation machinery and RNA polymerase subunits are ubiquitous. The replication-repair mechanisms are conserved among the bacteria, whereas the metabolic mechanisms are phyletically scattered, showing the presence of more NOD.

On the whole, the minimal gene set provides appreciably similar results to the COG approach. More than one third of the proteins from the original gene set show more phyletic scattering in the full COG collection. About a half of these are missing in only a couple of bacterial clades, and are, therefore, highly conserved genes, and are expected to code essential functions. The other genes may be NOD cases or non-essential genes, and can only be distinguished by examining the specific biological function of the proteins.

However, one needs to be careful in suggesting that a highly conserved gene is also essential to the functioning of the cell. This has been shown by knockout mutagenesis, which is used to find nonessential genes by introducing disruptive insertions in selected loci. These studies showed that among the 38 viable knockouts in the minimal gene set, 16 are into genes with a scattered phyletic distribution, but 7 are into universal genes. The first group may be explained as being nonessential to the minimal genome, but the second set perplexed researchers, until the explanation was put forward that the knockout mutagenesis only proves the nonessentiality of the gene under laboratory conditions and in the absence of competition, thereby not reflecting its real life importance.

Nonorthologous Gene Displacement

From the COG approach, NOD has been shown to be more important than originally perceived. (Only around 30% of the genes belong to ubiquitous families). Including the mutagenesis data, it is seen that 20-30 members of the minimal gene set are nonessential.

Whenever a protein is found in bacteria, but not in archaea and eukaryotes (or bacteria and eukaryotes, but not in archaea), NOD is likely (Table 4).

NOD seems to affect all classes of genes. The translation machinery is the most uniform, but NOD is seen in this case too, as seen in Table 4. Thus, one or more genes may displace an existing gene to take over its function. For example, in the DNA replication system, the bacterial components are not orthologous, and in some cases appear to be unrelated to archaeal and eukaryotic units. NOD can also be interpreted in a broader sense, with entire systems and pathways displacing others for a particular role. For example, glycolysis is nearly universal throughout species, but is replaced in *Ricketsiaa prowazakii* by the tricarboxylic-acid cycle.

Owing to the prevalence of NOD, one needs to reassess the concept of the minimal gene set, replacing it instead by a minimal set of functional niches, which can be filled by more than one distinct family of orthologs. With this development, the importance of finding the minimal-gene-set reduces greatly, but construction and analysis of minimal gene sets for different environmental conditions may be useful to predicting subsets of genes required in those conditions. Here again, though, one has to be careful about NOD. Table 5 shows the conserved portion of minimal sets for different lifestyles.

Current Status

Phylogenetic patterns, transposon knockout data, and biochemical reasoning suggest that, in principle, a cell could be supported by a smaller number of proteins than the originally proposed 250. One can also remove the repair systems, and some of the metabolic pathways – leaving us with the translation-transcription-replication mechanism, glycolysis for metabolism, a primitive transport system, and no cell wall. A detailed description can be found in the supplemental material to [1] at the website [4]. Also, there is a web-based program at [2], which shows minimalist description of cells and the genes required. It is doubtful that such a cell could survive under realistic conditions. Moreover, it should be noted that though many genes are individually dispensable, the effect of simultaneous deletion is not known well enough. It is likely that a too many conserved genes may reduce the fitness of cell.

Experimental approaches may identify NOD and help to remove the uncertainties faced by comparative genomics studies. Constructing and manipulating a minimal genome is still a challenge faced by the genomic engineering community. Since one has to perform a number of experiments to identify the effects of simultaneous deletion, it looks like we have to wait a few years before we can construct the minimal genome.

Lastly, there has been some discussion regarding the relevance of the minimal-gene-set concept to the reconstruction of ancestral genomes. The COG approach is not designed for studies of the course of evolution, and has been declared as evolutionarily irrelevant . However, Koonin feels that the universal genes may be a likely heritage of the “last universal common ancestor”. Also, NOD cases can be identified with specific stages in life’s evolution.

Discussion

From Koonin's article, it looks like the construction of the minimal genome is a problem that is solvable (even if it may take some time). The experimental techiques available nowadays are more than enough to carry out knockout mutagenesis studies. One has to be careful in computational studies, because some COGs may be passed over because the orthologous proteins may be too short in some of the species. Sometimes, a COG may be overlooked because the genes may seem to be species-specific. Identifying NOD cases is not an easy task, but some phylogenetic and biochemical reasoning should suffice to locate and explain the mechanisms and how they evolved. Gene sets are yet to be constructed for various environmental lifestyles, and this information would greatly help the researchers working with new types of bacteria and archaea. Once the minimal genes are known, these researchers can then locate NODs and map the evolutionary processes involved. Thus, the minimal gene set (or rather, the minimal functional niche set) is a problem worth pursuing.

References

[1] Eugene Koonin, How Many Genes Can Make a Cell: The Minimal-Gene-Set Concept, Annu. Rev. Genomics Hum. Genet. , 1, 2000, pg 99 and references therein

[2] www.e-cell.org

[3] http://www.ncbi.nlm.nih.gov/CBBresearch/Koonin/

[4] http://genom.annualreviews.org/cgi/content/full/1/1/99/DC1

Table 1. Coverage of completely-sequenced genomes by conserved families of orthologs

	Number of genes	
Species[a]	**Total**	**In COGs (% of total)**
Bacteria		
Aquifex aeolicus	1526	1265 (83%)
Thermotoga maritima	1852	1437 (78%)
Rickettsia prowazekii	834	632 (76%)
Mycoplasma genitalium	480	366 (76%)
Haemophilus influenzae	1694	1246 (74%)
Chlamydia trachomatis	895	612 (68%)
Treponema pallidum	1033	677 (66%)
Escherichia coli	4292	2752 (64%)
Bacillus subtilis	4100	2600 (63%)
Helicobacter pylori	1577	996 (63%)
Mycoplasma pneumoniae	678	408 (60%)
Chlamydia pneumoniae	1053	629 (60%)
Synechocystis sp.	3168	1883 (59%)
Borrelia burgdorferi	1256	656 (52%)
Archaea		
Archaeoglobus fulgidus	2411	1703 (71%)
Methanobacterium	1871	1319 (70%)
Methanococcus jannaschii	1747	1227 (70%)
Pyrococcus horikoshii	2072	1276 (62%)
Eukaryotes		
Saccharomyces cerevisiae	5932	2052 (35%)

[a]Within bacteria and archaea, the species are ordered by the percentage of genes included in clusters of orthologous groups of proteins (COGs).

TABLE 2 Phylogenetic patterns in the full COG[a] collection and in the minimal gene set classified by functional classes of proteins

Functional class of proteins	Phylogenetic pattern (number and %)					
	Full COG collection			Minimal gene set		
	Universal	Conserved in bacteria	Scattered[b]	Universal	Conserved in bacteria	Scattered[b]
Translation, ribosome structure, and biogenesis	53 (30%)	33 (18%)	93 (52%)	53 (57%)	33 (36%)	6 (7%)
Transcription	4 (5%)	2 (2%)	78 (93%)	4 (50%)	2 (25%)	2 (25%)
Replication, recombination, repair	5 (4%)	15 (12%)	108 (84%)	5 (17%)	15 (54%)	8 (29%)
Metabolism	9 (1%)	7 (1%)	686 (98%)	9 (12%)	7 (9%)	62 (79%)
Cellular processes: chaperone functions, secretion, cell division, cell wall biogenesis	9 (2%)	12 (3%)	374 (95%)	9 (28%)	12 (38%)	10 (34%)
Miscellaneous	1 (0%)	6 (1%)	637 (99%)	1 (6%)	6 (38%)	9 (56%)
Total	81 (4%)	75 (4%)	1953 (92%)	81 (32%)	75 (30%)	97 (38%)

[a]COG, cluster of orthologous groups of proteins.

[b]Defined in this case as missing at least one bacterial species; thus, within the full COG collection, this category includes proteins that are universally conserved in archaea and eukaryotes.

TABLE 3 Members of the original minimal gene set that show scattered phylogenetic patterns and are predicted to be dispensable (examples)

M. genitalium gene	Function/activity	Phylogenetic pattern[a]	Transposon insertion knockout reported?
MG012	Ribosomal protein S6 modification Enzyme (glutaminyl transferase)	`am-k---ce--h--gp-----`	No
MG104	Exoribonuclease (RNase B family)	`----yqvceb-hujgp-lin-`	No
MG346	rRNA methylase (SpoU class)	`-------cebrh--gp--in-`	Yes
MG278	Guanosine polyphosphate Pyrophosphohydrolase (SpoT)	`-----qvcebrhujgpo----`	Yes
MG097	Uracil DNA glycosylase	`----y---ebrhujgpo-in-`	No
MG262.1	Formamidopyrimidine-DNA glycosylase	`-------cebrh--gp-----`	No
MG408	Peptide methionine sulfoxide reductase	`--t-yqvcebrhujgp-l---`	Yes
MG448	Conserved domain frequently associated with peptide Methionine sulfoxide reductase	`--t-yqvcebrhujgp-l---`	No
MG049	Purine-nucleoside phosphorylase	`--------eb-huj-------`	Yes
MG052	Cytidine deaminase	`----y-v-ebrh--gpo----`	Yes
MG127	Arsenate reductase	`--------eb-h--gp-----`	No
MG033	Glycerol uptake facilitator	`a-t-y-vceb-h--gpo----`	Yes
MG038	Glycerol kinase	`a---yqvcebrh--gpo----`	No
MG299	Phosphotransacetylase	`------vcebrh--gpol--x`	Yes

[a]Pattern notation: a letter indicates presence of the respective species in the given cluster of orthologous groups of proteins, and a hyphen in the corresponding position indicates its absence. Species abbreviations: a, *Archeoglobus fulgidus*; m, *Methanococcus jannaschii*, t, *Methanobacterium thermoautotrophicum*; k, *Pysococcus horikoshii*; y, *S. cerevisiae*; q, *Aquifex aeolicus*; v, *Thermotoga maritima*; c, *Synechocystis* sp.; e, *E. coli*; b, *B. subtilis*; r, *Mycobacterium tuberculosis*; h, *H. influenzae*; u, *Helicobacter pylori*; j, *H. pylori* J strain; g, *M. genitalium*; p, *M. pneumoniae*; o, *Borrelia burgdorferi*; l, *Treponema pallidum*; I, *Chlamydia trachomatis*; n, *C. pneumoniae*; x, *Rickettsia prowazekii*.

TABLE 4 Nonorthologous gene displacement (NOD) within the minimal gene set (examples)

Function/activity	*M. genitalium* gene	Organisms with NOD (representative)	Phylogenetic pattern[a]	Comment
Lysyl-tRNA synthetase	*MG136*	Archaea, spirochetes, Rickettsia (MJ0359–*M. jannaschii*)	`----yqvcebrhujgp-lin-` `amtk----------ol--x`	Most bacteria an eukaryotes encode class II Lysyl-tRNA synthetase, whereas archaea, spirochetes, and rickettsiae, possess class I enzyme; the two enzymes are unrelated.
Glycyl-tRNA synthetase	*MG251*	Most bacteria (GlyQ, Glys–*E. coli*)	`amtky-----r---gpol---` `-----qvceb-huj----inx`	Mycoplasma, spirochetes, and mycobacteria encode an archaeal/eukaryotic-type, one-subunit glycyl-tRNA synthetase, whereas most bacteria possess a distinct, 2-subunit enzyme. The α-subunit is distantly related to the archaeal/eukaryotic, but they are not orthologs.
Cysteinyl-tRNA synthetase	*MG253*	Archaeal methanogens—*M. Jannaschii*, *M. thermoautotrophicum*	`a--kyqvcebrhujgpolinx` `-mt----------`	In the methanogenic archaea, *M. jannaschaii* and *M. thermoautotrophicum*, the function of cysteinyl-tRNA synthetase is fulfilled by prolyl-tRNA synthetase, which in these organisms is a bifunctional enzyme (34a).
Glutamine activation for translation	*MG098*, *MG099*, *MG100*	γ-Proteobacteria, eukaryotes (GlnS—*E. coli*)	`amtkyqvc-br-ujgpolinx`[b] `----y---e--h---------`	γ-Proteobacteria and eukaryotes possess an aminoacyl-tRNA synthetase for glutamine, whereas most bacteria and archaea use the transamidation mechanism (29).

DNA-dependent DNA polymerase main catalytic subunit	*MG261*[c]	Archaea, eukaryotes (MJ0885, MJ1630—*M. jannaschii*)	`-----qvcebrhujgpolinx`
			`amtky---e------------`
			`amtk----------------`
				The main catalytic subunits of the DNA polymerase in bacteria and in archaea/eukaryotes appear to be unrelated; archaea possess an additional DNA polymerase not seen in eukaryotes or bacteria.

ATPase involved in DNA replication initiation (Dna-A)	*MG469*		`-----qvcebrhujgpolinx`	The ATPases involved in replication initiation in bacteria and in archaea-eukaryotes are distantly related but not orthologous (24). Not

		Archaea, eukaryotes (MTH1412—*M. thermoautotrophicum*)	a-tky----------	the nearly complementary phylogenetic patterns (although NOD is likely for *M. jannaschii*).
Ribonuclease HII (family II)	*MG199*		-----q---b----gp--in-	A complex case of NOD—most bacteria have both a typical RNase HII (family I) and RNase HI. By contrast, mycoplasmas encode only a distinct form of RNase HII [family II; 4], whereas archaea possess only the typical RNase HII.
		Ribonuclease HII, family I: archaea, eukaryotes, most bacteria (RnhB—*E. coli*)	amtkyqvcebrhuj--o-inx	
		Ribonuclease HI: eukaryotes, most bacteria (RnhA—*E. coli*)	----y--cebrhuj---l--x	
Holliday junction resolvase endonuclease subunit	*MG291.1*		-----qvcebrhujgp--inx	Another complex case of NOD. RuvC and its functionally analogous but unrelated archaeal counterpart have been characterized experimentally. The mycoplasmas and *B. subtilis* do not encode orthologs of either of these but, along with most other bacteria, possess a protein that is distantly related to RuvC (L. Aravind and EV Koonin, unpublished data) and could function as an alternative resolvase.
		Archaea (MJ0947 — *M. jannaschii*)	amtk----------	
		Most bacteria (RuvC—*E. coli*)	------vce-rhuj---linx	
Fructose biphosphate aldolase	*MG023*		----yqvcebrhujgpol---	A classic case of NOD in an essential step of glycolysis. The two aldolases are distantly related but not orthologous. Note that the phylogenetic patterns are nearly complementary, but *A. aeolicus* and *E. coli* possess both aldolases.
		Archaea, *Chlamydia*, a few other bacteria (DhnA—*E. coli*)	amtk-q--e---------in-	

[a]The pattern notation and species abbreviations are the same as in Table 3.

[b]B subunit. Other subunits are missing in some species.

[c]MG261 is the ortholog of the replicative polymerase of *H. influenzae*, but in *M. genitalium* this protein is likely to be involved in repair, whereas MG031, a distinct form of DNA polymerase III represented only in gram-positive bacteria, is the likely replicative polymerase (14,18).

Functional class of proteins	Free-living organisms	Autotrophs	Chemoautotrophs	Thermophiles
Translation, ribosome structure and biogenesis	57	64	68	62
Transcription	4	9	11	6
Replication recombination, repair	9	14	18	14
Metabolism	62	152	171	89
Cellular processes: chaperone functions, secretion, cell division, cell wall biogenesis	17	48	50	30
Miscellaneous	13	46	74	44
Total	158	320	379	237

[a]The following species whose genomes have been completely sequenced were included in this analysis: Free-living organisms—*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *P. horikoshii*, *S. cerevisiae*, *A. aeolicus*, *T. maritima*, *Synechocystis* sp., *E. coli*, *B. subtilis*; autotrophs—*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *A. aeolicus*, *Synechocystis* sp.; chemoautotrophs—*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *A. aeolicus*; thermophiles—*A. fulgidus*, *M. jannaschii*, *M. thermoautotrophicum*, *D. horikoshii*, *A. aeolicus*, and *T. maritima*.

Table 5 – Conserved portions of hypothetical minimal gene sets for different lifestyles.