Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Spoilt for choice: protein target selection in a time of plenty

James Raftery[a]* and John R. Helliwell[a,b]

[a]Department of Chemistry, University of Manchester, Manchester M13 9PL, England, and [b]CLRC, Daresbury Laboratory, Warrington WA4 4AD, Cheshire, England

Correspondence e-mail: jrtest@spec.ch.man.ac.uk

Experiences in the application of Boolean logic to the clusters of orthologous groups of proteins (COGs) database for target selection in the *Mycobacterium tuberculosis* genome are described.

Received 2 November 2001
Accepted 1 March 2002

1. Introduction

Before the structure the crystal, before the crystal the protein and, increasingly, before the protein the gene. Once, the choice of structure was conditioned by what was available, but with the advent of whole genome sequencing and high-throughput crystallography, it is becoming a problem of filtering the choices to a reasonable number. Fortunately, the bioinformatics community has provided tools that can assist in this task. We describe below the path we followed in selecting our targets from the *Mycobacterium tuberculosis* genome (Cole *et al.*, 1998) for the *M. tuberculosis* Structural Genomics Consortium (www.doe-mbi.ucla.edu/TB/). Since we do not have access to gene-knockout results to analyse the *M. tuberculosis* genome into non-essential and pathogenicity/viability genes, we decided to exploit a natural experiment, the *M. leprae* genome (Cole *et al.*, 2001), which to a first approximation is the *M. tuberculosis* genome after massive gene knockout (from 3927 proteins to 1605).

1.1. Methods

The NIH has constructed a site (http://www.ncbi.nlm.nih.gov/COG/) which has organized the microbial genomes in an excellent, if not yet ideal, way. The proteins are related across genomes into clusters of orthologous groups (COGs; Tatusov, 1997, 2001). Each COG, by linking individual proteins/groups of paralogs from at least three of the more than 30 lineages, represents an ancient conserved domain. The core COG genomes are from unicellular eukarya, bacteria and archaea, but these have been supplemented by proteins from two multicellular eukaryotes, the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*. The steps we used were

1=M. leprae .AND. M. tuberculosis: common set.

Then, because we wanted to look at genes required by *M. tuberculosis* and organisms following similar strategies and not just those that every organism must have, we used the group with the smallest genome (Fraser *et al.*, 1995), the Mycoplasmas, to remove the minimum set of genes for independent existence,

2=1 .AND. .NOT. Mycoplasma group.

To simplify the situation for any therapeutic consequences of our structures-to-be, we removed, as proxies for *Homo sapiens*, COGs with *C. elegans*/*D. melanogaster*/yeast components,

3=2 .AND. .NOT. (worm or fly)
4=3 .AND. .NOT. (yeast).

Finally, since part of the survival strategy of *M. tuberculosis* is the ability to survive as an intracellular inhabitant of macrophage and this is shared with Chlamydia,

5=4 .AND. (Chlamydia).

The number of 'surviving' potential targets for each of the sets is

$$1 = 877$$
$$2 = 586$$
$$3 = 321$$
$$4 = 227$$
$$5 = 65.$$

The 65 finally selected COGs are shown in Table 1. These expand into 94 unique proteins.

2. Discussion and concluding remarks

As only 2585 out of 3927 *M. tuberculosis* proteins have at present been placed in COGs, the process starts with a certain loss, although as time progresses the number of COG entries can be expected to increase. A desirable addition to the present COG scheme is to expand it to incorporate *H. sapiens*. The structure of COGs makes it quite reasonable that closely related genomes be assigned just one component of the COG. For example, *M. tuberculosis* and *M. leprae* occupy just one component of the COG, as do *Ureaplasma*

Table 1
COGs.

The genomes/groupings corresponding to the letters aompkzyqvdrlbcefghsnujxitw of the bitmap are *Archaeoglobus fulgidus*, *Halobacterium* sp. NRC-1, *Methanococcus jannaschii* + *Methanobacterium thermoautotrophicum*, *Thermoplasma acidophilum* + *T. volcanium*, *Pyrococcus horikoshii* + *P. abyssi*, *Aeropyrum pernix, Saccharomyces cerevisiae* + *Candida albicans*, *Aquifex aeolicus*, *Thermotoga maritima*, *Deinococcus radiodurans*, *Mycobacterium tuberculosis* + *M. leprae*, *Lactococcus lactis* + *Streptococcus pyogenes*, *Bacillus subtilis* + *B. halodurans*, *Synechocystis*, *Escherichia coli* K12 + *E. coli* O157 + *Buchnera* sp. APS, *Pseudomonas aeruginosa*, *Vibrio cholerae*, *Haemophilus influenzae* + *Pasteurella multocida*, *Xylella fastidiosa*, *Neisseria meningitidis* MC58 + *N. meningitidis* Z2491, *Helicobacter pylori* + *H. pylori* J99 + *Campylobacter jejuni*, *Mesorhizobium loti* + *Caulobacter crescentus*, *Rickettsia prowazekii*, *Chlamydia trachomatis* + *C. pneumoniae*, *Treponema pallidum* + *Borrelia burgdorferi*, *Ureaplasma urealyticum* + *Mycoplasma pneumoniae* + *M. genitalium*, respectively. The first bitmap place is for fly/worm contributions. The meaning of the functional group letters JKLDOMNPTGCEFHIQRS can be found at http://www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?fun = all.

No. of proteins	Phylogenetic bitmap	Functional group	COG	Description
31	-----------r---efghsnujxi--	[N]	COG1560	Lauroyl/myristoyl acyltransferase involved in lipid A biosynthesis
44	-----------rlb-efghsn-jxit-	[M]	COG1686	D-Alanyl-D-alanine carboxypeptidase
17	----------dr---efghsn-jxi--	[K]	COG1678	Putative transcriptional regulator
20	----------drlbcefghs--jxit-	[L]	COG1195	Recombinational DNA-repair ATPase
68	----------drlbcefghsnuj-i--	[M]	COG0791	Cell-wall-associated hydrolases (invasion-associated proteins)
11	---------v-r-b--fg-s----i--	[T]	COG1875	Predicted ATPase related to phosphate starvation-inducible protein PhoH
23	---------vdr--cefghsnujxit-	[L]	COG0817	Holliday junction resolvasome endonuclease subunit
17	---------vdrlb-e-gh-----i--	[E]	COG1438	Arginine repressor
26	---------vdrlb-efghsnujxi--	[L]	COG1570	Exonuclease VII: large subunit
22	---------vdrlbce-gh---j-i--	[G]	COG0448	ADP-glucose pyrophosphorylase
22	---------vdrlbcefghsn-j-i--	[K]	COG1327	Predicted transcriptional regulator; consists of a Zn-ribbon and ATP-cone domains
29	---------vdrlbcefghsnujxit-	[L]	COG0742	N6-adenine-specific methylase
29	---------vdrlbcefghsnujxit-	[LK]	COG1197	Transcription-repair coupling factor: superfamily II helicase
11	--------q-dr---------u--it-	[R]	COG1579	Zn-ribbon protein: possibly nucleic acid binding
56	--------q-dr-bcefghsnuj-it-	[M]	COG0860	*N*-acetylmuramoyl-L-alanine amidase
24	--------q-dr-bcefghsnujxit-	[S]	COG1496	Uncharacterized ACR
25	--------q-drlbcef-h-nujxi--	[I]	COG0623	Enoyl-(acyl-carrier-protein) reductase (NADH)
35	--------q-drlbcefghsnujxi--	[J]	COG1295	tRNA-processing ribonuclease BN
27	--------qv-rlb-efghsnujxit-	[N]	COG1862	Preprotein translocase subunit YajC
27	--------qvdr--cefgh-nujxit-	[M]	COG0815	Apolipoprotein *N*-acyltransferase
28	--------qvdr--cefghsnujxit-	[R]	COG0728	Uncharacterized membrane protein: putative virulence factor
27	--------qvdr-b-efghsnujxit-	[K]	COG1158	Transcription termination factor
25	--------qvdr-bcefghsnuj-it-	[I]	COG0743	1-Deoxy-D-xylulose 5-phosphate reductoisomerase
26	--------qvdr-bcefghsnuj-it-	[M]	COG0821	Essential bacterial protein: involved in density-dependent regulation of peptidoglycan biosynthesis
27	--------qvdr-bcefghsnuj-it-	[IM]	COG0761	Penicillin tolerance protein
28	--------qvdr-bcefghsnujxit-	[J]	COG1825	Ribosomal protein L25 (general stress protein Ctc)
19	--------qvdrlbc------u--it-	[R]	COG1837	Predicted RNA-binding protein (KH domain)
25	--------qvdrlbcefghs--jxit-	[J]	COG1544	Ribosome-associated protein Y (PSrp-1)
29	--------qvdrlbcefghsnuj-it-	[HI]	COG1154	Deoxyxylulose-5-phosphate synthase
25	--------qvdrlbcefghsnujxi--	[TK]	COG0745	Response regulators consisting of a CheY-like receiver domain and a HTH DNA-binding domain
29	--------qvdrlbcefghsnujxit-	[R]	COG0802	Predicted ATPase or kinase
30	--------qvdrlbcefghsnujxit-	[L]	COG1198	Primosomal protein N′ (replication factor Y)–superfamily II helicase
31	--------qvdrlbcefghsnujxit-	[M]	COG0812	UDP-*N*-acetylmuramate dehydrogenase
35	--------qvdrlbcefghsnujxit-	[M]	COG0766	UDP-*N*-acetylglucosamine enolpyruvyl transferase
39	--------qvdrlbcefghsnujxit-	[M]	COG1181	D-Alanine-D-alanine ligase and related ATP-grasp enzymes
40	--------qvdrlbcefghsnujxit-	[M]	COG0773	UDP-*N*-acetylmuramate-alanine ligase
8	--------qvdrlbcefghsnujxit-	[D]	COG0772	Bacterial cell-division membrane protein
63	--------qvdrlbcefghsnujxit-	[M]	COG0768	Cell-division protein FtsI/penicillin-binding protein 2
64	--------qvdrlbcefghsnujxit-	[J]	COG1187	16S rRNA uridine-516 pseudouridylate synthase and related pseudouridylate synthases
10	------z---dr-b--f-----jxit-	[L]	COG2094	3-Methyladenine DNA glycosylase
32	---m----qvdrlbcefghsnujxit-	[M]	COG0771	UDP-*N*-acetylmuramoylalanine-D-glutamate ligase
32	---m----qvdrlbcefghsnujxit-	[M]	COG0770	UDP-*N*-acetylmuramyl pentapeptide synthase
38	---m----qvdrlbcefghsnujxit-	[M]	COG0769	UDP-*N*-acetylmuramyl tripeptide synthase
35	--o--kz-qv-r-bcefghsn-jxi--	[J]	COG1530	Ribonucleases G and E
19	--o-pkz-qv-r--c------u-xit-	[F]	COG1351	Predicted alternative thymidylate synthase
28	--om--z---drlb-efghsnuj-it-	[L]	COG1194	A/G-specific DNA glycosylase
32	--om-k--qvdr-bcefghsnujxit-	[N]	COG0341	Preprotein translocase subunit SecF
35	--om-k--qvdr-bcefghsnujxit-	[N]	COG0342	Preprotein translocase subunit SecD
28	--om-kz-qvdrlbcefghsn-j-i--	[R]	COG2262	GTPases
18	-a--p-z---dr---efg-sn-jxi--	[C]	COG2142	Succinate dehydrogenase hydrophobic anchor subunit
28	-a-m----qv-r-bcefghsnujxi--	[E]	COG0253	Diaminopimelate epimerase
37	-ao-----q-dr-bcefghsnujxi--	[O]	COG1651	Protein disulfide isomerase
10	-ao---z-qv-r--c---------i--	[S]	COG1259	Uncharacterized ACR
9	-aom-------rl-----------i--	[S]	COG1478	Uncharacterized ACR
57	-aom-kz-q-drlb-efg-snujxit-	[K]	COG1475	Predicted transcriptional regulators
46	-aom-kz-qvdrlbcefgh-nujxit-	[P]	COG0803	ABC-type Mn/Zn-transport system: periplasmic Mn/Zn-binding (lipo)protein (surface adhesin A)
49	-aom-kz-qvdrlbcefgh-nujxit-	[P]	COG1108	ABC-type Mn^{2+}/Zn^{2+} transport systems: permease components
27	-aomp-z-q-dr-bcefghsnu--i--	[H]	COG0373	Glutamyl-tRNA reductase
42	-aomp-z-qvdr-bce-g---u--i--	[HR]	COG1060	Thiamine biosynthesis enzyme ThiH and related uncharacterized enzymes
35	-aompkz--vdrlb-e---s--j-it-	[K]	COG1321	Mn-dependent transcriptional regulator
28	-aompkz--vdrlbce---s--j-it-	[R]	COG0396	Iron-regulated ABC transporter ATPase subunit SufC
51	-aompkz--vdrlbce---s--j-it-	[R]	COG0719	Predicted membrane components of an uncharacterized iron-regulated ABC-type transporter SufB
35	-aompkz-q--r-bcef-hsnujxi--	[F]	COG0717	Deoxycytidine deaminase
57	-aompkz-qv-rlbcefghsnujxi--	[EM]	COG0329	Dihydrodipicolinate synthase/*N*-acetylneuraminate lyase
56	-aompkz-qvdr-bcefghsnujxit-	[NO]	COG0616	Periplasmic serine proteases (ClpP class)

urealyticum, *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. However, as *M. genitalium* is the organism that presently holds the record for the smallest number of genes in a free-living organism (Fraser *et al.*, 1995), it would have been better to use that than

Ureaplasma urealyticum .OR. Mycoplasma
pneumoniae .OR. Mycoplasma genitalium

in 2. Likewise, while the site permits comparison of all genomes, the result cannot then be filtered on-site for those that are part of one component. In fact, after comparing *M. tuberculosis* and *M. leprae*, the rest of the filtering had to be performed using scripts. It should be emphasized that the NIH site is a very considerable achievement and that doubtless these quibbles will be remedied sooner rather than later. Our decision to use the *M. tuberculosis–M. leprae* intersection as an alternative to laboratory gene knockout neatly bypasses one of the problems of that procedure, namely that many knockouts have no phenotype, *i.e.* knockouts can appear to produce no effect but it is possible that the right environment had not been chosen to show its necessity. Our approach has a built-in reality check, *i.e. M. leprae* has survived in the wild as a pathogenic organism. While the relationship between *M. tuberculosis* and *M. leprae* may seem a one-off, similar relationships may turn out to be surprisingly common. For example, *Yersinia pseudotuberculosis* may be considered as ancestral to *Y. pestis*, the causative agent of plague. Both organisms are pathogenic, but by different modes. The recent complete sequencing of *Y. pestis* (Parkhill *et al.*, 2001) shows the genes associated with the *Y. pseudotuberculosis* pathogenicity mode have collapsed to pseudogenes; genes for its own pathogenicity have been obtained from a range of other organisms.

The members of the table have been ordered on the phylogenetic bitmap so identical profiles are clustered together. This phylogenetic profiling has been used to associate proteins that may be functionally linked (Eisenberg *et al.*, 2000). Other criteria such as number of amino acids or methionines may now also be applied.

This is a contribution from the North West Structural Genomics Centre (www.nwsgc.ac.uk).

References

Cole, S. T. *et al.* (1998). *Nature (London)*, **393**, 537–544.

Cole, S. T. *et al.* (2001). *Nature (London)*, **409**, 1007–1011.

Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. (2000). *Nature (London)*, **405**, 823–826.

Fraser, C. M. *et al.* (1995). *Science*, **270**, 397–403.

Parkhill, J. *et al.* (2001). *Nature (London)*, **413** 523–527.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). *Science*, **278**, 631–637.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. & Koonin, E. V. (2001). *Nucleic Acids Res.* **29**, 22–28.