Species and strain-specific patterns of low-complexity proteins in *Escherichia* and *Mycobacteria*

Tannistha Nandi[†], Krishnamoorthy Kannan[#] and Srinivasan Ramachandran[†,*]

[†]G N Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Mall Road, Delhi 110 007, India
[#]School of Biotechnology, GGS Indraprastha University, Kashmere Gate, Delhi 110 006, India

The patterns of evolution of proteins of low sequence complexity (LC proteins) are beginning to receive attention. In this study we have carried out comparative analysis of LC proteins from closely related strains of *Escherichia coli* and *Mycobacteria*. Species-specific differences were observed in all functional super-classes between *E. coli* and *Mycobacteria*. Inter-species comparisons of overall distribution of LC proteins within the genus *Mycobacteria* revealed differences in cellular processes (CP), transport and membrane-associated and characteristic (CH) super-classes. Strain-specific differences in *E. coli* were most apparent in the CH super-class, whereas in *M. tuberculosis* differences were observed in both CP and the CH super-classes. Further, differences among the *E. coli* strains with respect to the CH super-class could be directly correlated with their pathogenic vs non-pathogenic features, whereas with respect to the same super-class, two strains of *Mycobacterium tuberculosis* differed mainly in the number of proteins with a role in virulence.

THE availability of complete genome sequences of over 50 bacteria offers us new opportunities to investigate several issues of fundamental interest to biology using comparative genomics. One such aspect is the evolution of proteins with differing sequence complexities.

Majority of the proteins identified so far are of high sequence complexity. In general, these proteins tend to have a globular or near globular shape[1–3]. Many proteins such as enzymes belong to this category. Recently, large-scale sequencing has revealed a large number of proteins with low sequence complexity (LC proteins; containing significant proportion of simple sequences) in different organisms[2,4]. While proteins of high sequence complexity have been the focus of analysis with respect to structure–function relationships over the last few decades, LC proteins have received less attention. Earlier reports described LC proteins as having either structural or transmembrane characteristics[3]. The vast number of LC proteins identified by genome sequencing in various organisms calls for analysing the patterns of evolution of these proteins and identifying species-specific and strain-specific differences with respect to function.

Here, we report the comparative analysis of LC proteins from the bacteria *Escherichia coli* and *Mycobacteria*, and the identification of species and strain-specific LC proteins. We also describe a general computational procedure for rapid identification of such species and strain-specific LC protein-coding genes for further experimental investigations.

The complete genome files[5–9] of protein sequences in FASTA format were downloaded from the NCBI database. LC proteins were identified using the computer program ScanCom, according to the criteria described previously[1,2]. The ScanCom algorithm identifies the extent of reiterations of dipeptides in a given protein sequence, compares it with the maximum possible dipeptides for the same amino acid composition and classifies proteins into either the high complexity or low complexity category[1,2].

Sequence comparisons were carried out using programs available from The Wisconsin Package Ver. 10.1 and from the NCBI website. Sequence analysis of LC proteins was re-verified using genome BLAST with default settings. Differences in representation arising due to annotation differences or to partial sequences possibly inferred from internal start sites were ignored. To identify species and strain-specific differences, LC proteins with either no homologues or greater than 50% sequence diversity in either species or strains were noted. Identification of homologous proteins to a given sequence was carried out according to the standard procedure involving masking the low-complexity sequences to reveal true orthologous relationships. After the identification of homologues, sequence differences among them were assessed using the entire sequence.

In order to identify species and strain patterns in LC proteins, we have carried out systematic grouping of the LC proteins into different super-classes based on their functional annotation. The procedure was simplified by combining the information class and the metabolic class into a super-class called cellular processes (CP). The transport and membrane-associated classes were combined into transport and membrane-associated (TM) super-class. After the LC proteins were placed into the CP and TM super-classes, the remaining consisted of several proteins whose functional roles correlate with the characteristic biology of a given species or strain. Therefore, these proteins were placed into characteristic (CH) super-class. This procedure, a slight modification of Riley's method[10], enables a rapid comparative analysis of the genomic information on LC proteins[11].

Let us consider species-specific patterns. A summary of the LC proteins from two strains of *E. coli* and *Mycobacterium tuberculosis* and *M. leprae* is provided in Table 1. It is apparent that the distribution pattern in the *Mycobacteria* contrasts that of *E. coli*. The normalized

*For correspondence. (e-mail: ramu@cbt.res.in)

number of LC proteins in the *Mycobacteria* is higher (~3–6-fold) in the CP super-class. The opposite trend was observed in the TM super-class of the *M. tuberculosis* strains. The normalized number was lower by about two fold. The distribution in the CH super-class shows striking differences (~3–10-fold) among different species and strains.

When the number of LC proteins is compared within the strains of the same species, it is apparent that the two *E. coli* strains differ with respect to the CH super-class. In contrast, the two *M. tuberculosis* strains have similar proportion of LC proteins in this super-class.

Let us now consider strain-specific differences. The great majority of LC proteins in both strains of *E. coli* were assigned the same functions in the genome sequence. In order to identify strain-specific patterns, the comparative distribution of LC proteins in the different functional super-classes and classes for the two *E. coli* strains K12 and O157 were analysed. These are listed in Table 2. With respect to the CP super-class, LC proteins identified in O157 are involved in metabolic roles, whereas in K12 the lone LC protein is involved in replication. In the case of TM super-class, in O157 they belong to the group of ABC transporters, a protein of prophage CP-933, a polyamine transport system protein, sugar transport, and a cyanate transporter. In K12, the transporters belong to citrate-dependent iron uptake, and a membrane protein of unknown function. In the case of CH super-class, K12 has no LC protein that does not have a homologue in the O157 strain. On the other hand, in O157 there are several prophage CP-933 encoded proteins, exoproteins, intimin receptor protein, secreted proteins EspB and SepZ, and the type III secretion apparatus protein that do not have homologues in the K12 strain.

Table 1. Distribution of LC proteins in the functional super-classes CP, TM and CH in *Escherichia* and *Mycobacteria*

Species	Number of LC proteins in different functional classes[a]			
	CP	TM	CH	H[b]
E. coli K12	5.40	12.20	4.46	12.43
E. coli O157	6.59	13.18	12.70	12.45
M. tuberculosis H37Rv	32.44	7.15	43.93	88.89
M. tuberculosis cdc1551	30.79	8.06	42.77	87.73
M. leprae	15.60	24.33	14.35	28.70

[a]Number of LC proteins in each functional super-class was normalized to the total number of proteins scanned using ScanCom[2] and upscaled by a factor of 1000.
[b]H, Group of proteins with no known function and are annotated as hypothetical.

Table 2. Strain-specific differences in distribution of LC proteins in different functional super-classes and their assigned functions in *E. coli* K12 vs O157[a]

Functional super-class	Function assigned to LC protein in strain K12	Function assigned to LC protein in strain O157
CP	Calcium-binding protein required for initiation of chromosome replication	Putative acyl carrier protein
TM	Citrate-dependent iron transport (2) Membrane protein	Transport (2) ABC transport (2) Membrane protein of prophage CP-933X Ribose-specific transport
CH		Proteins of phage CP-933 (26) Putative adhesion Exoproteins (2) Intimin receptor protein Secreted protein EspB SepZ Type III secretion apparatus

[a]The number of proteins if greater than 1 is mentioned in brackets.

As in the case of *E. coli*, majority of LC proteins between the two *Mycobacterial* strains have the same functional roles. Strain-specific differences were observed in the CP and CH super-classes. These observations are summarized in Table 3. In the TM super-class, the two strains had identical functional representation of the LC proteins. A transhydrogenase subunit in H37Rv has no homologous protein in the CDC1551. Likewise, a adenylate cyclase in the CDC strain has a protein in H37Rv that differs in sequence by more than 50%. A DNA-binding protein in the CDC strain has no similar protein in H37Rv. With respect to the CH super-class, the differences are in the number of the proteins belonging to the PGRS, PPE and PE families between the two strains.

Species-specific differences between *M. leprae* and *M. tuberculosis* H37Rv are as follows. With respect to the CP super-class, no difference was observed in the LC proteins of *M. leprae* when compared with *M. tuberculosis*. Two membrane proteins in the TM super-class and one protein (a secreted protein) have no reported homologues in *M. tuberculosis* (Table 4). Although there are sequence differences (an average of 20%) between the homologous LC proteins from *M. leprae* and *M. tuberculosis*, functionally they are highly similar except in the presence of the above-mentioned genes coding for LC proteins in *M. leprae*.

Differences in the LC protein distribution among different strains and species in various functional super-classes arise due to two reasons: first, the presence or absence of the protein (or the encoding gene) itself and second, the differences in the proportion of reiterated motifs (low complexity sequences) within the same protein. Here, we have focused on the presence or absence of an LC protein, and LC proteins exhibiting greater than 50% sequence diversity in inter-species and strain comparisons.

Table 3. Strain-specific differences in distribution of LC proteins in different functional super-classes and their assigned functions in *M. tuberculosis* H37Rv vs CDC1551[a]

Functional super-class	Function assigned to LC protein in strain H37Rv	Function assigned to LC protein in strain CDC1551
CP	Transhydrogenase subunit alpha	Adenylate cyclase, DNA-binding protein CopG family
CH	PE (32), PE_PGRS (61), PE_PGRS(wag22), PPE (60)	PE (31 proteins), PE_PGRS (50 proteins), PPE (56 proteins)

[a]The number of proteins if greater than 1, is mentioned in brackets.

Table 4. Species-specific differences in distribution of LC proteins in different functional super-classes and their assigned functions in *M. leprae* compared with *M. tuberculosis* H37Rv. The corresponding gi numbers are also shown

Functional super-class	Function assigned to LC protein in *M. leprae*
TM	Membrane protein (gi 13093305)
	Integral membrane protein (gi 13093730)
CH	Secreted protein (gi 13092813)

In general, low complexity sequences undergo sequence variations at the DNA level due to slippage during replication and/or unequal crossing-over mechanisms[12]. In pathogenic organisms, sequence variations have been hypothesized to confer survival advantage against the host immune defence mechanisms[13]. In this respect, the genome sequences from different species and strains from a wide phylogenetic spectrum offer an opportunity to investigate the patterns of such sequence variations and their evolution.

E. coli O157 is an enteropathogenic strain, whereas *E. coli* K12 is not. Functional differences in the LC proteins of the two strains of *E. coli*, K12 and O157, are most apparent in the CH super-class. In *E. coli* O157 strain, a substantial number of LC proteins belong to the prophage CP-933, indicating transfer of genes due to phage invasion. Other LC proteins in O157 have functional roles in secretion and adhesion in relation to its pathogenic features. Noteworthy in this regard are the type III secretion apparatus and the *Escherichia*-secreted proteins[14,15]. Thus, it is apparent that in the *Escherichia* strains, differences in LC proteins have arisen due to phage transfer and acquisition of enteropathogenic features.

M. tuberculosis strain CDC1551 is a clinical strain, whereas H37Rv is the commonly used laboratory-virulent strain. Earlier reports described that CDC1551 has large sequence and single nucleotide polymorphisms in various genes compared to H37Rv. Here we observed minor differences in the LC proteins of the two strains with respect to functional diversity. A transhydrogenase in H37Rv has no reported homologue in the CDC1551 strain and likewise, a DNA-binding protein identified in CDC1551 strain does not have a reported homologue in the H37Rv strain. Sequence differences in the adenylate cyclase proteins of the CDC1551 and the H37Rv strains as identified in this work were reported earlier[8]. Differences between the two strains were also observed due to the number of LC proteins belonging to the same family, such as PPE and PE_PGRS family of proteins that have a role in virulence. These observations on LC proteins are in agreement with the current paradigm that the differences in the two strains of *M. tuberculosis* have arisen due to insertion/deletion and nucleotide polymorphisms.

M. leprae has a drastically reduced number of proteins when compared to *M. tuberculosis*[9]. Three LC proteins (two membrane proteins and one secreted protein) in *M. leprae* do not appear to have homologues in *M. tuberculosis*. It would be of interest to characterize the functional roles of the genes encoding these proteins.

It is apparent from this comparative analysis of *E. coli* and *Mycobacteria* that significant functional differences in the LC protein-coding genes in different strains of the same species and different species of the same genus occur in the TM and CH super-classes. In summary, our computational procedure of analysing sequence complexity and subsequent comparative analysis using a modified functional classification scheme can be used for rapid identification of species and strain-specific differences for further experimental characterization.

1. Nandi, T., B-Rao, C. and Ramachandran, S., *J. Biosci.* (*Suppl. 1*), 2002, **27**, 15–25.
2. Nandi, T. *et al.*, *J. Biomol. Struct. Dyn.*, 2003, **20**, 657–668.
3. Wootton, J. C., *Comput. Chem.*, 1994, **18**, 269–285.
4. Romero, P., Obradovic, Z. and Dunker, A. K., *FEBS Lett.*, 1999, **462**, 363–367.
5. Blattner, F. R. *et al. Science*, 1997, **277**, 1453–1474.
6. Perna, N. T. *et al.*, *Nature*, 2001, **409**, 529–533.
7. Cole, S. T. *et al.*, *Nature*, 1998, **393**, 537–544.
8. Fleischmann, R. D. *et al.*, *J. Bacteriol.*, 2002, **184**, 5479–5490.
9. Cole, S. T. *et al.*, *Nature*, 2001, **409**, 1007–1011.
10. Riley, M., *Microbiol. Rev.*, 1993, **57**, 862–952.
11. Nandi, T., Kannan, K. and Ramachandran, S., *In Silico Biol.*, 2003, **3**, 0024.
12. Nishizawa, M. and Nishizawa, K., *Proteins*, 1999, **37**, 284–292.
13. Pizzi, E. and Frontali, C., *Genome Res.*, 2001, **11**, 218–229.
14. Devinney, R., Nisan, I., Ruschkowski, S., Rosenshine, I. and Finlay, B. B., *Infect. Immunol.*, 2001, **69**, 559–563.
15. Umanski, T., Rosenshine, I. and Friedberg, D., *Microbiology*, 2002, **148**, 2735–2744.

ACKNOWLEDGEMENTS. T.N. is a recipient of a fellowship from CSIR. S.R. is a recipient of NMITLI and funds from CSIR. We thank Prof. Samir Brahmachari and Yogendra Singh for discussions.

Received 14 November 2002; revised accepted 10 April 2003