consequence of a lack or excess of DMPK protein. Resolution of the alternative mechanisms underlying the development of DM requires the production of transgenic mouse models that include an expanded CTG repeat. This must be the next goal of DM research.

Acknowledgements

We thank Gill Bates and Sita Reddy for helpful discussions. Work in our laboratory is funded by the Muscular Dystrophy Association.

References

1 Harper, P. (1989) *Myotonic Dystrophy. Major problems in Neurology 21*. Saunders
2 Brook, J.D. *et al.* (1992) *Cell* 68, 799–808
3 Carango, P., Noble, J.E., Marks, H.G. and Funanage, V.L. (1993) *Genomics* 18, 340–348
4 Sabouri, L.A. *et al.* (1993) *Nat. Genet.* 4, 233–238
5 Fu, Y.H. *et al.* (1993) *Science* 260, 235–238
6 Hoffmann-Radvayi, H. *et al.* (1993) *Hum. Mol. Genet.* 2, 1263–1266
7 Krahe, R. *et al.* (1995) *Genomics* 28, 1–14
8 Wang, J.Z. *et al.* (1995) *Hum. Mol. Genet.* 4, 599–606
9 Jansen, G. *et al.* (1996) *Nat. Genet.* 13, 316–323
10 Reddy, S. *et al.* (1996) *Nat. Genet.* 13, 324–335
11 Justice, R.W. *et al.* (1995) *Genes Dev.* 9, 534–546
12 Wang, Y.H. *et al.* (1994) *Science* 265, 669–671
13 Pearson, C.E. and Sinden, R.R. (1996) *Biochemistry* 35, 5041–5053
14 Timchenko, L.T., Timchenko, N.A., Caskey, C.T. and Roberts, R. (1996) *Hum. Mol. Genet.* 5, 115–121
15 Boucher, C.A. *et al.* (1995) *Hum. Mol. Genet.* 4, 1919–1925
16 Otten, A.D. and Tapscott, S.J. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 12, 5465–5469
17 Shaw, D.J. *et al.* (1993) *J. Med. Genet.* 30, 189–192
18 Jansen, G. *et al.* (1995) *Hum. Mol. Genet.* 4, 843–852
19 Shaw, D.J. *et al.* (1993) *Genomics* 18, 673–679
20 St Johnston, D. (1995) *Cell* 81, 161–170.
21 Rastinejad, F., Conboy, M.J., Rando, T.A. and Blau, H.M. (1993) *Cell* 75, 1107–1117
22 Rastinejad, F. and Blau, H.M. (1993) *Cell* 72, 903–917
23 Taneja, K.L. *et al.* (1995) *J. Cell Biol.* 128, 995–1002
24 Willems, P. (1994) *Nat. Genet.* 8, 213–215
25 Zeitlin, S. *et al.* (1995) *Nat. Genet.* 11,155–163
26 Duyao, M.P. *et al.* (1995) *Science* 269, 407–410
27 Nasir, J. *et al.* (1995) *Cell* 81, 811–823
28 Burright, E.N. *et al.* (1995) *Cell* 82, 937–948
29 Ikeda, H. *et al.* (1996) *Nat. Genet.* 13,196–202
30 Bingham, P.M. *et al.* (1995) *Nat. Genet.* 9, 191–196
31 The Dutch–Belgian Fragile X Consortium (1994) *Cell* 78, 23–33

LETTER

Non-orthologous gene displacement

Unrelated proteins with the same biochemical activity can evolve independently[1]. Without complete genome information, however, it is not possible to ascertain whether a particular function is encoded by homologous or non-homologous genes in two different species. The chance always remains that the counterpart to a given gene in one species has not been sequenced in the other. In an attempt to prove and to assess quantitatively the postulate that distinct proteins in different species can be responsible for the same function, we compared the first two genomes to be sequenced completely, those of the parasitic bacteria *Haemophilus influenzae*[2] and *Mycoplasma genitalium*[3]. For this problem to be addressed, the function of the gene in question has to be essential in both organisms. Despite this limitation, we found a considerable number of cases where apparently indispensable cell functions are encoded by non-orthologous genes.

By definition, orthologs are genes that are related by vertical descent from a common ancestor and encode proteins with the same function in different species. By contrast, paralogs are homologous genes that have evolved by duplication and code for proteins with similar, but not identical, functions[4]. Comparison of the 1703 putative proteins encoded by the *H. influenzae* genome[2,5] with the 469 *M. genitalium* protein sequences, revealed that 233 showed sufficiently high similarity for them to be considered orthologs (the criteria for distinguishing orthologs from paralogs are discussed in Ref. 5, in which the genome of *H. influenzae* was compared with the available 75% of the *Escherichia coli* genome). In addition, by comparing the remaining *M. genitalium* and *H. influenzae* proteins with those in the non-redundant sequence database, we identified 12 clear-cut cases where essential cellular functions in the two bacteria were encoded by non-orthologous (i.e. unrelated or paralogous) genes (Table 1). We call this phenomenon non-orthologous gene displacement, although its origins can be quite diverse.

Nucleoside diphosphate kinase (NDK) appears to be a case where an essential function has been taken over by an unrelated protein. This example shows the impact of the concept of non-orthologous gene displacement on the analysis of complete genome sequences. Knowing that a function, in this case the formation of nucleoside triphosphates from diphosphates catalyzed by NDK, is indispensable but that the given genome does not encode homologs of the respective proteins from other organisms, one should screen functionally unassigned gene products for likely candidates. In *M. genitalium*, we found two plausible candidates for a novel NDK, both of which contain sequence motifs typical of nucleoside and nucleotide kinases.

The case of the repair DNA polymerase shows the complexity of the scenarios leading to non-orthologous gene displacement. It appears that a duplication of the DNA polymerase III (Pol III) gene in the common ancestor of Gram-positive and Gram-negative bacteria (which also encoded the

 0168-9525/96/$15.00
PII: S0168-9525(96)20010-1

TABLE 1. Non-orthologous genes coding for the same function in *Mycoplasma genitalium* and *Haemophilus influenzae*

Enzyme	*M. genitalium* Gene[a]	*M. genitalium* Orthologs[b]	*H. influenzae* Gene[a]	*H. influenzae* Orthologs[b]	Comment[c]
No sequence similarity between *M. genitalium* and *H. influenzae* proteins					
Phosphoglycerate mutase	MG430 (*yibO*)	PMGI_BACSU PMGI_ECOLI PMGI_MAIZE	HI0757 (*gpmA*)	PMG1_ECOLI PMGM_HUMAN not in G(+)	*Escherichia coli* encodes both types of enzymes
L-lactate dehydrogenase	MG460	LDH_BACSU LDHM_HUMAN	HI1739B (*lctD* or *lldD*)	LLDD_ECOLI G(+)	The HI enzyme is distantly related to eukaryotic cytochrome B2
Lipoate-protein ligase	MG270	LPLA_ECOLI SCYJL046W_1	HI0027 (*lipB*)	LIPB_ECOLI S51458 (yeast)	*E. coli* and yeast encode both types of enzymes
Nucleoside diphosphate kinase	MG264?[d] MG268?[d]	None	HI0876 (*ndk*)	NDK_ECOLI NDKB_HUMAN	The two predicted kinases in MG are candidates for this indispensable activity
DNA polymerase, repair	MG261 (*dnaE*)	DP3A_HAEIN DP3A_ECOLI	HI0856 (*polA*)	DPO1_ECOLI DPO1_MYCTU	MG encodes two homologs of DNA polymerase III. MG261 is the likely repair polymerase as it belongs to a putative repair operon[6]
RNase H	MG262?[d]	DPO1_BACCA DPO1_HAEIN	HI0138 (*rnhA*); HI1059 (*rnhB*)	RNH_ECOLI RNH1_YEAST RNH2_ECOLI MC326_1 (*M. capricol.*) SC23CDS_13 (yeast)	MG262 is homologous to the 5′–3′ exonuclease domain of DNA polymerase I. It is predicted to replace the two unrelated RNases H of HI in primer removal during DNA replication
Glycyl-tRNA synthetase	MG251	SYG_HUMAN	HI0927 (*glyQ*) HI0924 (*glyS*)	SYGA_ECOLI SYGB_ECOLI *CTU20547_1* (Chlamydia) G(−)	The MG enzyme contains one subunit, the HI counterpart two
Paralogs in *M. genitalium* and *H. influenzae*					
Prolyl-tRNA synthetase	MG283	YHI0_YEAST	HI0729 (*proS*)	SYP_ECOLI YER7_YEAST	Yeast encodes both types of enzymes
Cytidine deaminase	MG052	CDD_BACSU CDD_HUMAN	HI1350 (*cdd*)	CDD_ECOLI	The MG cytidine deaminase is more closely related to eukaryotic enzymes than to those from G(+) bacteria
Pyruvate dehydrogenase E1 component α and β subunits	MG273 MG274	ODPA_BACSU ODPA_HUMAN ODPB_BACSU ODPB_HUMAN	HI1233 (*aceE*)	ODP1_ECOLI G(−)	Two subunits in G(+) bacteria (including MG) and eukaryotes, one subunit in G(−) bacteria

[a] The *M. genitalium* and *H. influenzae* proteins are designated as in Refs 3 and 2, respectively. The orthologous *E. coli* gene is indicated in parentheses where available.

[b] The SWISS-PROT code is given wherever available. G(−) and G(+) indicate the species range of the orthologs (Gram-negative and Gram-positive, respectively).

[c] Abbreviations: MG, *M. genitalium*; HI, *H. influenzae*.

[d] ? indicates that there are no detectable orthologs, but the functions are evidently essential for the organisms; thus, the detected similarity to other proteins with similar functions is suggestive but needs to be verified experimentally.

repair enzyme Pol I), was followed by a subsequent elimination of one of the Pol III paralogs in the Gram-negative lineage and of Pol I in the *Mycoplasma* lineage[6]. As a result, one of the Pol III paralogs, which is encoded in the same operon as two putative repair enzymes, has probably overtaken the repair polymerase function in the mycoplasmas (Ref. 6 and Table 1). Theoretically, one may imagine that in a minimal organism, a single polymerase would be responsible for replication and repair DNA synthesis[7]. The analysis of the actual *M. genitalium* genes, however, makes a compelling case for non-orthologous displacement.

The situation with RNase H is even more intricate. *H. influenzae*

and *E. coli* each encode two unrelated RNases H. These enzymes are involved in the removal of RNA primers during DNA replication, which is evidently essential for the completion of replication. In *M. genitalium*, there is no homolog of any known RNase H, and the activity is likely to be provided by the MG262 protein that is homologous to the N-terminal, 5′–3′ exonuclease domain of Pol I (Ref. 6 and Table 1). This example shows the limitations of the classical definition of orthology as even though the MG262 protein is more closely related to Pol I than to any other protein in *H. influenzae* or other bacteria, it is obvious that the functions of these proteins only partially overlap.

Finally, as two types of aminoacyl-tRNA synthetases of *M. genitalium* have only eukaryotic orthologs, horizontal gene transfer with subsequent displacement of the bacterial enzymes seems to be a plausible mechanism (Table 1).

In addition to the 11 instances of non-orthologous gene displacement summarized in Table 1, we found more than 10 cases in which it remains unclear whether the *M. genitalium* and *H. influenzae* genes responsible for a particular function are orthologs or paralogs (data not shown). Even though the respective proteins are highly similar in *M. genitalium* and *H. influenzae*, they show approximately the same level of similarity to their eukaryotic or archaeal counterparts, suggesting the possibility of non-orthologous displacement. However, alternative hypotheses, such as an unusual 'burst-like' mode of evolution of these genes, cannot be ruled out.

It is remarkable that as many as 11 non-orthologous genes with the same functions have been found compared with the 233 apparent orthologs in *H. influenzae* and *M. genitalium*. In larger genomes, there are likely to be hundreds of such cases, and they will require special consideration in the reconstruction of metabolic pathways and in comparative genome analysis.

EUGENE V. KOONIN*, ARCADY R. MUSHEGIAN* AND PEER BORK‡
bork@embl-heidelberg.de

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.
‡European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012, Heidelberg and Max-Delbrück-Center for Molecular Medicine, Berlin-Buch, Germany.

References

1 Doolittle, R.F. (1994) *Trends Biochem. Sci.* 19, 15–18
2 Fleischmann, R.D. *et al.* (1995) *Science* 269, 496–512
3 Fraser, C.M. *et al.* (1995) *Science* 270, 397–403
4 Fitch, W.M. (1970) *Syst. Zool.* 19, 99–106
5 Tatusov, R.L. *et al.* (1996) *Curr. Biol* 6, 279–291
6 Koonin, E.V. and Bork, P. (1996) *Trends Biochem. Sci.* 21, 128–129
7 Mushegian, A.R. and Koonin, E.V. *Proc. Natl. Acad. Sci. U. S. A.* (in press)