14, 8605–8613
57 Hayashi, H. *et al.* (1997) *J. Bacteriol.* 179, 4246–4253
58 Cornet, P. *et al.* (1983) *Biotechnology* 1, 589–594
59 Lemaire, M. and Béguin, P. (1993) *J. Bacteriol.* 175, 3353–3360
60 Zverlov, V.V. *et al.* (1994) *Biotechnol. Lett.* 16, 29–34
61 Kirby, J. *et al.* (1997) *FEMS Microbiol. Lett.* 149, 213–219
62 Ohmiya, K. *et al.* (1997) *Biotechnol. Genet. Eng. Rev.* 14, 365–414
63 Karita, S., Sakka, K. and Ohmiya, K. (1997) in *Rumen Microbes and Digestive Physiology in Ruminants* (Onodera, R. *et al.*, eds), pp. 47–57, Japan Sci. Soc. Press
64 Tamaru, Y. *et al.* (1997) *J. Ferment. Bioeng.* 83, 201–205
65 Fanutti, C. *et al.* (1995) *J. Biol. Chem.* 270, 29314–29322
66 Li, X., Chen, H. and Ljungdahl, L. (1997) *Appl. Environ. Microbiol.* 63, 4721–4728

Variation and evolution of the citric-acid cycle: a genomic perspective

Martijn A. Huynen, Thomas Dandekar and Peer Bork

The presence of genes encoding enzymes involved in the citric-acid cycle has been studied in 19 completely sequenced genomes. In the majority of species, the cycle appears to be incomplete or absent. Several distinct, incomplete cycles reflect adaptations to different environments. Their distribution over the phylogenetic tree hints at precursors in the evolution of the citric-acid cycle.

M.A. Huynen, T. Dandekar and P. Bork are in the European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, and in the Max-Delbrück-Centre for Molecular Medicine, 13122 Berlin-Buch, Germany;*
M.A. Huynen is also in the Bioinformatics Group, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.
**tel: +49 6221 387372, fax: +49 6221 387517, e-mail: huynen@embl-heidelberg.de*

Completely sequenced genomes have provided a new way of analysing the biochemical pathways in a species: using the presence of genes encoding the enzymes that catalyse its reactions[1,2]. By studying the variation in metabolic pathways and the way that they are encoded in a rapidly growing set of sequenced genomes, we can elucidate their evolution. Here, we present an investigation of the presence and absence of genes, in prokaryotes and yeast, that code for the enzymes involved in the citric-acid cycle (CAC), including variations such as the reductive CAC and the branched citric-acid pathway, the glyoxylate shunt, and in the reactions connecting the CAC to pyruvate and phosphoenolpyruvate.

Our analysis has combined a thorough examination of sequence data, which included improving the annotation of genes in the GenBank genome database, with an analysis of the biochemical data on the compared species. We examined the genomes of unicellular organisms published to date, including those of four Archaea, 14 Bacteria and one Eukaryote. For an overview of the published genomes, including references, see `http://www.tigr.org/tdb/mdb/mdb.html`.

Variability of the pathway

The genes involved in the CAC and its connections to pyruvate and phosphoenolpyruvate in the various genomes are indicated in Table 1 and a graphical display of the reaction steps for which genes can be found in the selected genomes is given in Fig. 1. The first striking feature in most of the genomes is the incompleteness of the CAC. Only the four largest genomes, those of *Escherichia coli*, *Bacillus subtilis*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae*, and the small genome of *Rickettsia prowazekii*, encode the genes for a complete CAC. In the other genomes, the cycle has gaps or is completely absent. In these incomplete cycles, the genes that are present generally code for reactions that are connected to each other, suggesting there are functional connections between the genes. In incomplete cycles, the last part of the oxidative cycle (steps 6–8 in Fig. 1a), leading from succinate to oxaloactetate, is the most highly conserved, whereas the initial steps (steps 1–3), from acetyl CoA to 2-ketoglutarate, show the least conservation.

When interpreting the role of incomplete CACs, it is important to realize that, as well as the oxidation of acetyl CoA, the CAC also plays a role in the generation of intermediates for anabolic pathways. Specifically, 2-ketoglutarate (between steps 3 and 4), oxaloacetate (between steps 8 and 1) and succinyl CoA (between steps 5 and 6) are starting points for the synthesis of glutamate, aspartate and porphyrin, respectively. The autotrophic species that are missing a small part of the CAC are still able to generate 2-ketoglutarate, oxaloacetate and succinyl CoA from

0966-842X/99/$ - see front matter PII: S0966-842X(99)01539-5

Table 1. Citric-acid cycle enzymes and genes[a,b]

Enzymes[c]	Genes	Species						
		Escherichia coli	*Haemophilus influenzae*	*Helicobacter pylori*	*Rickettsia prowazekii*	*Bacillus subtilis*	*Mycobacterium genitalium*	*Mycobacterium tuberculosis*
Citrate synthase (1) EC 4.1.3.7	*gltA*	EC0720	–	HP0026	RP844	*citZ* *citA*	–	TB0896 TB0889
Aconitase (2) EC 4.2.1.3	*acnA*	EC1276 EC0771	–	–	RP799	*citB*	–	TB1475
	acnB	EC0118	–	HP0779	–	–	–	–
Isocitrate dehydrogenase (3) EC 1.1.1.42	*icdA*	EC1136	–	HP0027	RP265	*citC*	–	TB3339 TB0066
2-ketoglutarate dehydrogenase (4) EC 1.2.4.2	*sucA*	EC0726	HI1662	–	RP180	*odhA*	–	TB1248
EC 2.3.1.61	*sucB*	EC0727	HI1661	–	RP179	*odhB*	–	TB1248 TB2215
2-ketoglutarate ferredoxin oxidoreductase (4) EC 1.2.7.3	*korA*	–	–	HP0589	–	–	–	TB2455[d]
	korB	–	–	HP0590	–	–	–	TB2454[d]
	korC	–	–	HP0591	–	–	–	–
	korD	–	–	HP0588	–	–	–	TB2455
Succinyl-CoA synthetase (5) EC 6.2.1.5	*sucC*	EC0728	HI1196	–	RP433	*sucC*	–	TB0951
	sucD	EC0729	HI1197	–	RP432	*sucD*	–	TB0952
Succinyl-CoA–acetoacetate-CoA-transferase EC 2.8.3.5	*scoA*	–	–	HP0691	–	*yxjD*	–	TB2504
	scoB	–	–	HP0692	–	*yxjE*	–	TB2503
Fumarate reductase[f] (6) EC 1.3.99.1	*frdA*	EC4154	HI0835	HP0192	–	*sdhA*	–	TB1552 TB0248
	frdB	EC4153	HI0834	HP0191	–	*sdhB*	–	TB1553 TB0247
	frdC	EC4152	HI0833	HP0193	–	*sdhC*	–	TB1554
	frdD	EC4151	HI0832	–	–	–	–	TB1555
Succinate dehydrogenase (6) EC 1.3.99.1	sdhA	EC0723	–	–	RP128	–	–	TB3318
	sdhB	EC0724	–	–	RP044	–	–	TB3319
	sdhC	EC0721	–	–	RP126	–	–	TB3316
	sdhD	EC0722	–	–	RP127	–	–	TB3317
Fumarase (7) EC 4.2.1.2	*fumA, B* (Class I)	EC1612 EC4122	–	–	–	–	–	–
	fumC (Class II)	EC1611	HI1398	HP1325	RP665	*citG*	–	TB1098
Malate dehydrogenase (8) EC 1.1.1.37	*mdh*	EC3236	HI1210	–	RP376	*citH*	MG460[d]	TB1240
	Archaeal type	EC0517 EC0801 EC3575	HI1031	–	–	*yjmC*	–	–

Chlamydia trachomatis	*Treponema pallidum*	*Synechocystis*	*Aquifex aeolicus*	*Archaeoglobus fulgidus*	*Methanococcus jannaschii*	*Methanobacterium thermoautotrophicum*	*Pyrococcus horikoshii*	*Saccharomyces cerevisiae*
–	–	*gltA*	AQ0150(c)	AF1340	–	MT0962(c) MT1726(c)	–	YNR001C
–	–	–	– AQ1784	–	–	–	–	YLR304C
–	–	slr0665	–	–	–	–	–	–
–	–	*icd*	AQ1512	AF0647	–	–	–	YOR136W
CT054	–	–	–	–	–	–	–	YIL125W
CT055 CT400	–	–	–	–	–	–	–	YDR148C
–	–	–	–	AF0469	MJ0276	MT1033	PH1662 PH1666	–
–	–	–	–	AF0468	MJ0537	MT1034	PH1661 PH1665	–
–	–	–	–	AF0471	MJ0536	MT1035	PH1660 PH1663	–
–	–	–	–	AF0470	MJ0146	MT1032	PHunnot[e]	–
CT821	–	*sucC*	AQ1620 AQ1306	AF1540 AF2186	MJ0210	MT1036	–	YGR244C
CT822	–	*sucD*	AQ1622 AQ1888	AF1539 AF2185	MJ1246	MT0563	–	YOR142W
–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	–	–	–
CT592	–	–	AQ0594	AF0681	MJ0033	MT1502	–	–
CT591	–	–	AQ0655 AQ0553	AF0682	MJ0092	MT1850	–	–
CT593	–	–	–	AF0684	–	–	–	–
–	–	–	–	AF0683	–	–	–	–
–	–	*frdA*	–	– –	–	–	–	YJL045W YKL148C
–	–	*sdhB* *sdhB*	– –	–	–	–	–	YLL041C
–	–	–	–	–	–	–	–	YMR118C
–	–	–	–	–	–	–	–	–
–	–	–	AQ1780(n) AQ1679(c)	AF1099(n) AF1098(c)	MJ1294(n) MJ0617(c)	MT1735(n) MT0963(n) MT1910(c)	PH1683(n) PH1684(c)	–
CT855	–	*fumC*	–	–	–	–	–	YPL262W
CT376	–	*citH*	AQ1665 AQ1782	AF0855	MJ0490	MT0188	–	YKL085W
–	–	–	–	–	MJ1425	MT1205	PH1277	–

continued

Enzymes[c]	Genes	Species						
		Escherichia coli	*Haemophilus influenzae*	*Helicobacter pylori*	*Rickettsia prowazekii*	*Bacillus subtilis*	*Mycobacterium genitalium*	*Mycobacterium tuberculosis*
Isocitrate lyase (glyoxylate pathway) (9)								
EC 4.1.3.1	*aceA*	EC4015	–	–	–	–	–	TB0467
Malate synthase (glyoxylate pathway) (10)								
EC 4.1.3.2	*aceB*	EC4014	–	–	–	–	–	–
Phosphoenol pyruvate carboxykinase (ATP) (11)	*glcB*	EC2976	–	–	–	–	–	TB1837
EC 4.1.1.49	*pckA*	EC3403	HI0809	–	–	*pckA*	–	–
Phosphoenol pyruvate carboxykinase (GTP) (11)								
EC 4.1.1.32	*pck*	–	–	–	–	–	–	TB0211
Phosphoenol pyruvate carboxylase (11)								
EC 4.1.1.31	*ppc*	EC3956	HI1636	–	–	–	–	–
Malic enzyme (12)								
EC 1.1.1.40	*mez*	EC2463	HI1245	–	RP373	*ytsJ* *yqkJ*	–	–
EC 1.1.1.38	*sfcA*	EC1479	–	–	–	*malS*	–	TB2332
Pyruvate carboxylase/ oxaloacetate decarboxylase[g] (13)								
EC 4.1.1.3	*oadB*	–	–	–	–	–	–	–
EC 6.3.1.4	*pycA/ accC*	– –	–	–	*pycA*	–	TB2967	–
EC 6.4.1.1	*pycB/ oadA*	–	–	–	*pycA*	–	TB2967	– –
Pyruvate dehydrogenase (14)								
EC 1.2.4.1	*aceE* (E1)	EC0114	HI1233	–	–	–	–	TB2241
EC 1.2.4.1	*pdhA* (E1-a)	–	–	–	RP261	*pdhA*	MG274	TB2497
EC 1.2.4.1	*pdhB* (E1-b)	–	–	–	RP262	*pdhB*	MG273	TB2496
EC 2.3.1.12	*aceF*	EC0115	HI1232	–	RP530	*pdhC*	MG272	TB2495 TB2215
EC 1.8.1.4	*lpdA*	EC0116	HI1231	–	RP460 RP805	*pdhD*	MG271	TB0462
Pyruvate ferredoxin oxidoreductase								
EC 1.2.7.1	*porA*	EC1378	–	HP1110	–	–	–	–
	porB	EC1378	–	HP1111	–	–	–	–
	porC	EC1378	–	HP1108	–	–	–	–
	porD	EC1378	–	HP1109	–	–	–	–

[a]If a gene numbering system is available, the genes are indicated with their gene numbers, and the initials of the species (TB for *M. tuberculosis*); otherwise, gene names are used. Multiple copies of a gene that are a result of recent gene duplications or that do not have separate orthologs in multiple species are in one column. Genes that align only with part of a larger protein are indicated with a (c) for a carboxy-terminal alignment and an (n) for an amino-terminal alignment.
[b]Paralogous gene displacements in the citric acid cycle are shown by a lightly shaded background. Non-homologous gene displacements are indicated with a darker background.
[c]The numbers after the enzyme names correspond to those in Fig. 1.
[d]Indicates tentative identifications.

Chlamydia trachomatis	*Treponema pallidum*	*Synechocystis*	*Aquifex aeolicus*	*Archaeoglobus fulgidus*	*Methanococcus jannaschii*	*Methanobacterium thermoautotrophicum*	*Pyrococcus horikoshii*	*Saccharomyces cerevisiae*
–	–	–	–	–	–	–	–	YER065C
–	–	–	–	–	–	–	–	YNL117W
–	–	–	–	–	–	–	–	
–	–	–	–	–	–	–	–	YKR097W
CT710	TP0122	–	–	–	–	–	PH0312	–
–	–	*ppc*	–	AF1486	–	MT0943	PH0016	–
–	–	*me*	–	AF1727	–	–	PH1275	YKL029C
–	–	–	–	–	–	–	–	–
–	TP0057	–	–	AF2084	–	–	PH1283	–
–	–	AQ1470	AF0220 AQ1517 AQ1664	MJ1229	MT1917	–	YBR218C	YGL062W
TP0056	–	AQ1614	AF1252(c)	MJ1231	MT1107	PH0834		
–	–	AQ1520	AF2085(n)	–	–	–	YGL062W	
–	–	–	–	–	–	–	–	–
CT245	–	slr1934	–	–	–	–	–	YER178W
CT246	–	*pdhB*	–	–	–	–	–	YBR221C
CT247	–	*odhB*	–	–	–	–	–	YNL071W
CT557	–	*pdhD*	AQ0736	–	MJ0636	MT1648	–	YFL018C
–	TP093	*nifJ*	AQ1167 AQ1195	AF1701	MJ0267	MT1739	– PH0684	–
–	TP093	*nifJ*	AQ1168 AQ1196	AF1702	MJ0266	MT1738	– PH0685	–
–	TP093	*nifJ*	AQ1169 AQ1200	AF1699	MJ0269	MT1740	PH0678	–
–	TP093	*nifJ*	AQ1171a AQ1192a	AF1700	MJ0268	MT1740	– PH0682	–

[e] A new gene that is orthologous to the *korD* gene of the other Archaea is present in position 1469328 to 1469148 of the *P. horikoshii* genome (complementary strand).

[f] Genes that are significantly more similar to the *E. coli* succinate dehydrogenase than to the *E. coli* fumarate reductases are only found in *R. prowazekii*, *M. tuberculosis*, *S. cerevisiae* and *Synechocystis*. In the other species the genes are equidistant to either enzyme, and are in the fumarate reductase row.

[g] The genes *oadB* and *accC* provide information about the direction of the pyruvate carboxylase and oxaloacetate decarboxylase (*pycB* and *oadA*) reactions in the cell. The presence of *accC* and *oadB* is only indicated if *pycB* or *oadA* is present.

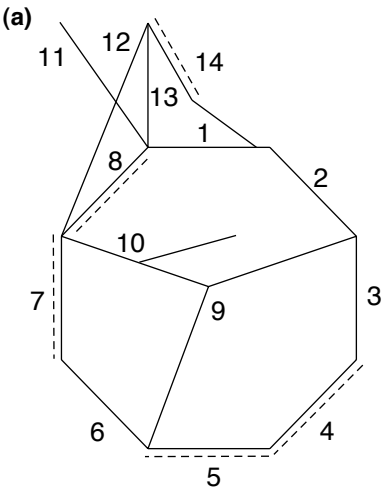
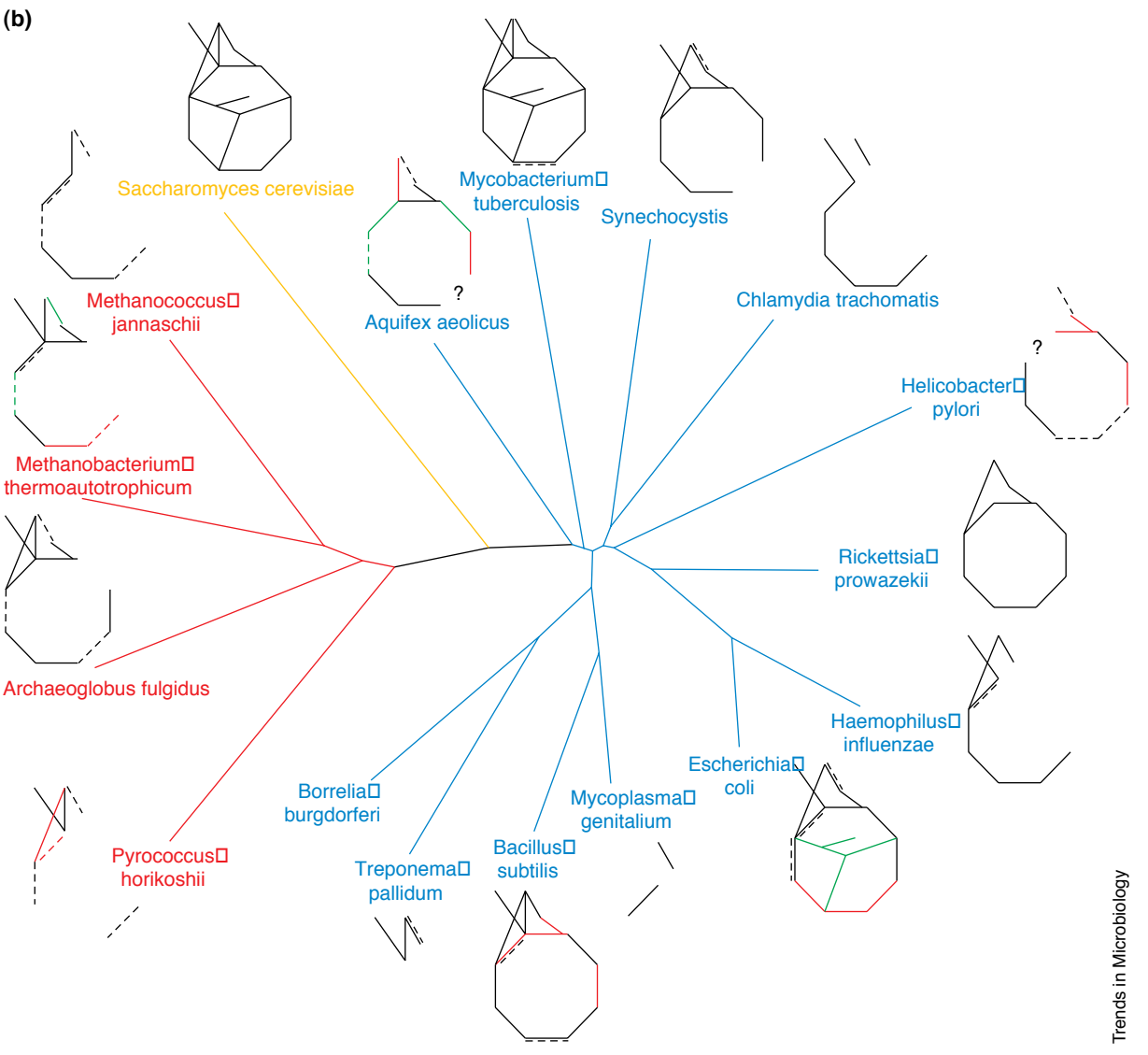


Fig. 1. (a) A graphical representation of the reactions of the citric-acid cycle (CAC), including the connections with pyruvate and phosphoenolpyruvate, and the glyoxylate shunt. When there are two enzymes that are not homologous to each other but that catalyse the same reaction (non-homologous gene displacement), one is marked with a solid line and the other with a dashed line. The oxidative direction is clockwise. The enzymes with their EC numbers are as follows: 1, citrate synthase (4.1.3.7); 2, aconitase (4.2.1.3); 3, isocitrate dehydrogenase (1.1.1.42); 4, 2-ketoglutarate dehydrogenase (solid line; 1.2.4.2 and 2.3.1.61) and 2-ketoglutarate ferredoxin oxidoreductase (dashed line; 1.2.7.3); 5, succinyl-CoA synthetase (solid line; 6.2.1.5) or succinyl-CoA–acetoacetate-CoA transferase (dashed line; 2.8.3.5); 6, succinate dehydrogenase or fumarate reductase (1.3.99.1); 7, fumarase (4.2.1.2) class I (dashed line) and class II (solid line); 8, bacterial-type malate dehydrogenase (solid line) or archaeal-type malate dehydrogenase (dashed line) (1.1.1.37); 9, isocitrate lyase (4.1.3.1); 10, malate synthase (4.1.3.2); 11, phosphoenolpyruvate carboxykinase (4.1.1.49) or phosphoenolpyruvate carboxylase (4.1.1.32); 12, malic enzyme (1.1.1.40 or 1.1.1.38); 13, pyruvate carboxylase or oxaloacetate decarboxylase (6.4.1.1); 14, pyruvate dehydrogenase (solid line; 1.2.4.1 and 2.3.1.12) and pyruvate ferredoxin oxidoreductase (dashed line; 1.2.7.1). **(b)** Individual species might not have a complete CAC. This diagram shows the genes for the CAC for each unicellular species for which a genome sequence has been published, together with the phylogeny of the species. The distance-based phylogeny was constructed using the fraction of genes shared between genomes as a similarity criterion[29]. The major kingdoms of life are indicated in red (Archaea), blue (Bacteria) and yellow (Eukarya). Question marks represent reactions for which there is biochemical evidence in the species itself or in a related species but for which no genes could be found. Genes that lie in a single operon are shown in the same color. Genes were assumed to be located in a single operon when they were transcribed in the same direction and the stretches of non-coding DNA separating them were less than 50 nucleotides in length.



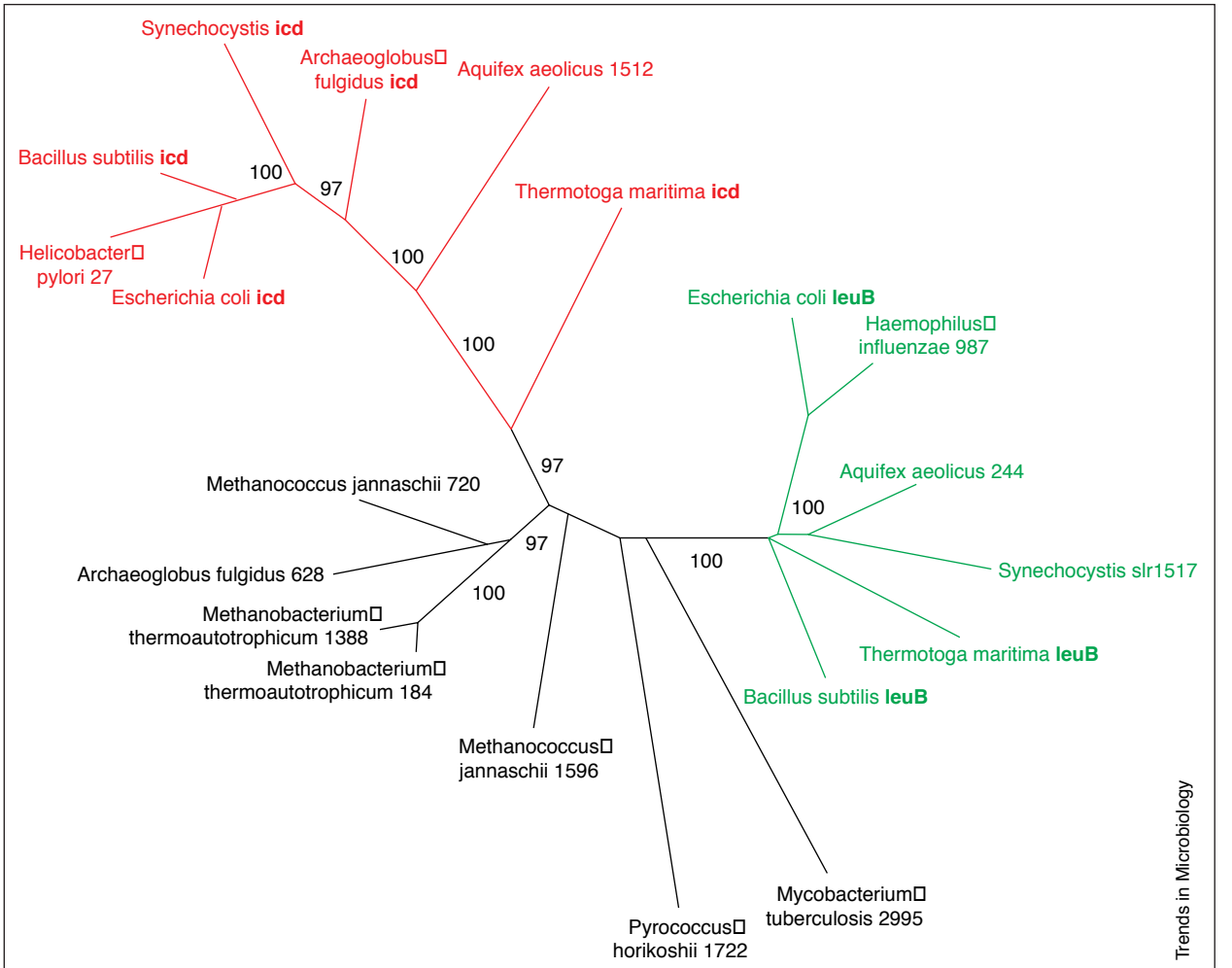


Fig. 2. A neighbour-joining clustering of isopropylmalate-dehydrogenase genes (*leuB*) and isocitrate-dehydrogenase genes (*icd*) in completely sequenced genomes. The genes with experimentally characterized functions are shown in bold, for the others the gene number has been indicated. A set of genes from the Archaea and from *M. tuberculosis* (black) is nearly equidistant from the bacterial isopropylmalate dehydrogenases (green) and the bacterial isocitrate dehydrogenases (red). Two of the archaeal genes (MT1388 and AF0628) are located in leucine operons, identifying them and the other archaeal genes in the lower two-thirds of the tree as isopropylmalate dehydrogenases. Note that the *A. fulgidus icd* gene, a 'true' isocitrate dehydrogenase in the Archaea, was probably horizontally transferred from the Bacteria to *A. fulgidus*[30]. Bootstrap values larger than 90 are indicated.

pyruvate; however, the path from pyruvate to 2-ketoglutarate varies among the species. The autotrophic Bacteria generate 2-ketoglutarate from pyruvate via the right branch of the CAC, activating it in the oxidative direction (clockwise in Fig. 1a), whereas the methanogenic Archaea[3] and *Archaeoglobus fulgidus* generate it via the left branch of the pathway, activating it in the reductive direction (counterclockwise in Fig. 1a).

Precursors of the CAC

The variation in the CAC provides clues to its evolutionary origins, as it shows which incomplete cycles are feasible in which environments. It has been argued that the CAC evolved originally as two separate pathways stemming from pyruvate, with an oxidative branch leading to 2-ketoglutarate and a reductive branch leading to succinyl CoA (Ref. 4). Alternative potential intermediate stages are present in *Synechocystis*, which lacks 2-ketoglutarate dehydrogenase[5], and in the methanogens *Methanococcus jannaschii* and *Methanobacterium thermoautotrophicum*[3]. The pathway in the two Archaea is a particularly interesting candidate for an intermediate stage in CAC evolution, as two of the missing enzymes, isocitrate dehydrogenase and aconitase, have homologues in the leucine-biosynthesis pathway (Fig. 2). Thus, the duplication of two genes from a single pathway could have completed two gaps in the CAC simultaneously.

The methanogenic lifestyle is presumably very old, as reflected by its presence throughout the euryarchaeal phylogeny[6]. The only archaeal isocitrate dehydrogenase found, the one from *A. fulgidus*, probably arose via a horizontal gene transfer from the Bacteria (Fig. 2). This supports the hypothesis that the last common ancestor of the Euryarchaeota did not have aconitase nor isocitrate dehydrogenase, and that its CAC was incomplete.

In the Bacteria, the CAC appears to have been complete by the time the proteobacteria, the low

G+C and the high G+C Gram-positive organisms, diverged from each other, as it is complete in at least one species in each of these taxa. The incomplete cycles in the bacterial pathogens therefore appear to be the result of a secondary loss of genes. Biochemical results from *Aquifex pyrophilus*[7] and other species from the root of the bacterial phylogeny suggest that the complete CAC in the Bacteria was originally operating in the reductive direction as a pathway for carbon assimilation[8].

Operon structure

Operon organization provides information about the functional relatedness of genes and thus suggests pathways. For example, in *M. thermoautotrophicum*, pyruvate ferredoxin oxidoreductase (step 14 in Fig. 1a) and fumarase (step 7) are located in a single operon, indicating that they are part of the same pathway of acetate assimilation that starts with acetyl CoA and operates part of the CAC in the reductive direction. Such a metabolic flux has indeed been observed in methanogens[3]. The operon organization of genes also suggests that *Aquifex aeolicus* has a complete CAC, as there are two operons connecting genes in the first half of the reductive cycle (the 'left' part in Fig. 1) to the second half (the 'right' part in Fig. 1). However, we cannot confirm the presence of a 2-ketoglutarate ferredoxin oxidoreductase (step 4) in *A. aeolicus*.

It has been argued that, as there is little conservation of operon structure in prokaryotic evolution beyond genes of proteins that interact physically[9,10] (Table 1), its role in the (co)regulation of genes is rather limited[11]. The organization of genes of the CAC into operons throughout the phylogenetic tree suggests that the lack of conservation of which particular genes are located in the same operon is rather a result of the high rate of evolution of the (co)regulation of genes.

Box 1. Glossary

Gene relationships

Homologous: Genes are homologous if they evolved from a common ancestor.

Orthologous: Genes are orthologous if their divergence reflects a speciation event[31]. For example, α globin in humans is orthologous to α globin in chimpanzees. Orthology is used for function prediction as orthologous genes are, in comparison to paralogous genes, relatively likely to perform the same function. An orthologous relationship between two genes does not guarantee, however, that they have the same function.

Paralogous: Genes are paralogous if their divergence reflects a gene duplication event within a species[31]. For example, α-globin in humans is paralogous to β-globin in humans.

Gene movement

Gene displacement: Occurs when different genes code for proteins that perform the same function.

Non-homologous gene displacement: Occurs if the genes displacing each other are not homologous.

Non-orthologous gene displacement: Includes both paralogous and non-homologous gene displacement[32].

Orthologous gene displacement: Occurs if orthologous genes displace each other by, for example, horizontal gene transfer from one genome to another.

Paralogous gene displacement: Occurs if the genes displacing each other are paralogous.

Gene displacement

The analysis of a biochemical pathway based on its genes assumes that we know at least one example sequence for every gene involved in that pathway, to provide a search image. Non-homologous gene displacement (Box 1) cautions us in this assumption. For example, the fact that *Helicobacter pylori* does not have homologues of the 2-ketoglutarate dehydrogenase of *E. coli* does not mean that the transformation between 2-ketoglutarate dehydrogenase and succinyl CoA (step 4 in Fig. 1a) is absent. *H. pylori* has an alternative enzyme complex for this reaction: 2-ketoglutarate ferredoxin oxidoreductase[12]. Non-homologous gene displacement, apart from being interesting from an evolutionary perspective, is also of interest in finding species-specific drug targets. Indeed, 2-ketoglutarate ferredoxin oxidoreductase is thought to be essential for the viability of *H. pylori*[12].

In the CAC of the published genomes, non-homologous gene displacement can also be observed for malate dehydrogenase, fumarase, pyruvate dehydrogenase and succinyl-CoA synthetase (Table 1). As is the case for 2-ketoglutarate dehydrogenase, the function of pyruvate dehydrogenase is taken over by a ferredoxin oxidoreductase, in this case pyruvate ferredoxin oxidoreductase (POR). Of the two types of malate dehydrogenase, the second was first observed in Archaea, but it does, however, have orthologues in *E. coli*, *Haemophilus influenzae* and *B. subtilis* (Table 1). Two non-homologous classes of fumarase have been distinguished. In the genomes analyzed here, Class I is present in *E. coli* and in the Archaea, whereas Class II are only observed in the Bacteria and *S. cerevisiae*. A succinyl-CoA–acetoacetate-CoA transferase has been shown to be active in *H. pylori*[13], where it takes over the role of succinyl-CoA synthetase in the CAC.

Non-homologous gene displacement of subunits of enzyme complexes also occurs. There are two non-homologous E1 subunits of the pyruvate dehydrogenase complex. One type is found in Gram-positive bacteria and eukaryotes, and the second is present in *E. coli*, *H. influenzae* and *M. tuberculosis*. A frequently occurring type of gene displacement takes place in non-catalytic subunits in, for example, fumarate reductase. Depending on the species, the fumarate reductase protein complex is anchored to the membrane by either two subunits (e.g. in *E. coli*) or one subunit (e.g. in *B. subtilis*), or can even be cytoplasmatic and fused to a heterodisulphide reductase, as in *M. thermoautotrophicum*[14]. Furthermore, membrane-binding subunits are not necessarily orthologous across the species (e.g. compare *E. coli* with *A. fulgidus*). The fusion of orthologous catalytic domains with proteins that are non-homologous to each other points to the embedding of the reactions of the CAC in different cellular contexts.

For example, the genes for pyruvate carboxylase (*pycB*) and oxaloacetate decarboxylase (*oadA*) are orthologues of each other. Their occurrence in the genome in conjunction with either oxaloacetate decarboxylase β chain (encoded by *oadB*) or biotin carboxylase (encoded by *accC*) indicates the flux of the reaction they catalyse. The genomes of the heterotrophic species *Treponema pallidum* and *Pyrococcus horikoshii* contain *oadB* but not *accC*, indicating a catabolic flux (i.e. decarboxylating), whereas the autotrophic species in Table 1 contain *accC*, indicating an anabolic flux. In *A. fulgidus*, which carries both the *oadB* and the *accC* genes, the reaction appears to occur in both directions.

What is considered a gene displacement depends, of course, on the level of functional resolution[15]. For example, is merely replacing a catalyst in a pathway sufficient to qualify as gene displacement or must the cofactors used also be the same? There are several non-homologous enzymes that catalyse the transformation between oxaloacetate and phosphoenolpyruvate but that use different sources of phosphate (ATP, GTP or pyrophosphate; Table 1).

A less drastic form of gene displacement is paralogous displacement (Box 1). There are two types of aconitase synthase, encoded by the genes *acnA* and *acnB*, whose divergence pre-dates that of the radiation in the Bacteria (see Table 1 for more examples). Even though several non-orthologous gene displacements have been identified in the CAC, we cannot claim to know them all. Thus, genome data should always be checked for consistency with experimental data on the pathways in a species, especially when reaction steps are hypothesized to be absent.

Comparing biochemical data with genomic data

The predictions of the presence of biochemical pathways based on the genome analysis shown in Fig. 1 are consistent with the biochemical data on the species analyzed, with a few exceptions, which are discussed below. (An overview of the literature supporting our predictions of the presence and absence of CAC genes is available from `http://dove.EMBL-Heidleberg.DE/Genome/Citric_Acid_Cycle`.) Caution is needed when comparing genomic data from one species with biochemical data from a different species within the same genus, as gene content evolves at a high rate[9]. Even within a species, the gene content varies: a phosphofructokinase gene isolated from *E. coli* is absent from the published *E. coli* genome sequence (T. Dandekar *et al*., submitted).

In *H. pylori*, all of the CAC reactions have been observed[16] except that of succinyl-CoA synthetase. No sequence homologous to a known malate dehydrogenase has, however, been detected in either of the sequenced *H. pylori* genomes. The malate-dehydrogenase activity measured[16] is one order of magnitude lower than the activity of the CAC enzymes for which genes could be found. Also, for the glycolytic enzymes pyruvate kinase and phosphofructokinase low activities have been observed[16] and their genes have not been identified in the *H. pylori* genome[17].

In *Mycoplasma genitalium* and *Mycoplasma pneumoniae*, by contrast, only malate-dehydrogenase activity has been detected[18]. The gene that could be responsible for this activity, MG460, was classified originally as a lactate dehydrogenase but might have a dual substrate specificity[19].

In *A. pyrophilus*, a close relative of *A. aoelicus*, all enzyme activities of a complete, reductive TCA have been reported[7]. By contrast, in *A. aoelicus*, no suitable candidate for 2-ketoglutarate ferredoxin oxidoreductase has been found. The enzymes that have been proposed as 2-ketoglutarate ferredoxin oxidoreductases[20] are significantly more similar to pyruvate oxidoreductases and ketoisovalerate oxidoreductases, although they do not cluster within any of the known groups of ferredoxin oxidoreductases.

No biochemical data are available on the CAC of *Chlamydia trachomatis*. Interestingly, it has been reported that infection by *Chlamydia psittaci* increases glutamate production in the host cell[21]. Glutamate could be oxidized via 2-ketoglutarate and the incomplete CAC of *C. trachomatis*. The heterotrophic species *P. horikoshii* requires peptides for growth[22]. This could explain its unusual, fragmented cycle, with the isolated ketoglutarate oxidoreductase playing a role in the fermentation of glutamate.

Technical aspects of finding orthologous genes

In determining whether genes coding for the enzymes of specific reactions are present, several techniques can be used that go beyond simple homology and orthology searches. They reveal the extraordinary variety in the evolution of proteins and, consequently, a number of technical pitfalls in the analysis of genome sequences.

Discrimination between orthology and paralogy

Determining whether homologous proteins perform the same function, or merely a similar function is one of the major bottlenecks in genome annotation. In determining functional equivalence, orthology (Box 1) is an important tool. The principal sources of information for identifying orthologous genes are the relative levels of sequence identity. Orthologous genes are expected to have higher levels of sequence identity to each other than to paralogous genes[9]. Using multiple genomes, orthology predictions can be tested for consistency[9,23]. However, patterns of sequence similarity alone are often not sufficient to make a reliable decision.

An example of this is given by the isopropylmalate dehydrogenases of the Archaea (Fig. 2). These genes, none of which has been experimentally characterized, have similar levels of identity to bacterial isopropylmalate dehydrogenases (part of the leucine-synthesis pathway) and bacterial isocitrate dehydratases. Two of the archaeal isopropylmalate dehydrogenase genes, MT1388 and AF0628, are located in an operon together with other leucine-synthesis genes. This identifies them as isopropylmalate dehydrogenases, whose function can subsequently be transferred to a

Questions for future research

- Where does the citric-acid cycle (CAC) originate from? Genome analysis reveals various alternatives for incomplete cycles (Fig. 1b). Have any of these been a precursor to the CAC? More completely sequenced genomes from the root of the bacterial and archaeal phylogenies will help in providing an answer.
- Did the last common ancestor of Bacteria and Archaea have a complete CAC? More complete genome sequences of the Archaea, specifically of the Crenarchaea, for which there is biochemical evidence for a complete CAC, should provide an answer.
- What are the variants of the CAC in eukaryotes, specifically in the a-mitochondrial ones that branch early in the phylogeny?
- Does gene displacement have any functional relevance?
- Are genes that are hypothesized to be part of a single pathway coexpressed? Delineation of regulatory structures and the incorporation of gene-expression data should provide an answer.

larger set of orthologous archaeal genes. Thus, the extra knowledge gained from knowing the context of a gene in the genome can be crucial in the assignment of function[9,15]. At present, there are still genes within this set that are annotated as isocitrate dehydrogenases in the genome and pathway databases. In addition, among the archaeal isopropylmalate dehydratases, some of which were identified originally as aconitases in the genome annotations, correct functional annotation is now possible using the knowledge that two of them lie in a leucine-synthesis operon.

Undiscovered open reading frames

Short genes encoding less than 100 amino acids are not always identified or recognized in annotated genomes. One example is the δ subunit of 2-ketoglutarate ferredoxin oxidoreductase in *P. horikoshii*. The knowledge that a subunit was missing directed a specific search at the DNA level for a gene encoding that subunit, using an orthologous sequence from another archaeon. Thus, the missing subunit was quickly identified in a likely location, adjacent to the other subunits of the oxidoreductase in the genome (Table 1).

The modular organization of proteins

The fusion and splitting of genes is common in the studied genomes (Table 1). In the CAC of Archaea and *A. aeolicus*, the genes are split relatively often. Fused genes are not always characterized as such in the original genome annotations. An example of a fused gene is the 2-ketoglutarate dehydrogenase gene in *M. tuberculosis* (*sucA–sucB*, TB1248), in which the *sucB* gene is fused to the 5′ end of the *sucA* gene. However, the first 130 amino acids of the *sucB* region of the gene, corresponding to the lipoyl domain and a linker region, are missing. An alternative, complete *sucB* gene is present in the TB2215 gene, which has two lipoyl domains. As this type of 2-oxo-acid dehydrogenase forms a multimeric enzyme complex, the subunit with two lipoyl domains might compensate for the one without a lipoyl domain.

Low sequence similarity

A gene for phosphoenolpyruvate carboxylase (PEPC) has not been identified in *M. thermoautotrophicum*, even though the enzyme activity has been observed[24]. Iterative, profile-based homology searches using PSI-Blast[25] and starting with the gene for PEPC from one of the Bacteria, identified genes homologous to PEPC in *M. thermoautotrophicum* (MT0943), *A. fulgidus* (AF1486) and *P. horikoshii* (PH0016), with a significance of $E = 0.001$. PSI-Blast with an *E*-value cutoff of 0.001 has been shown to be a reliable tool for finding homologous relationships[26].

The calculated molecular weight of the protein encoded by MT0943, 59 kDa, is close to that of PEPC in the archaeon *Methanothermus sociabilis* (60 kDa)[27]. The archaeal PEPC candidates are homologous to the carboxy-terminal two-thirds of bacterial PEPCs. This is in accordance with the observations that the PEPC in *M. sociabilis* does not show allosteric regulation and that the amino-terminal region of the *E. coli* PEPC is specifically involved in allosteric regulation[27].

Subunit sharing

Lipoamide dehydrogenase (LPD), encoded by the *lpdA* gene, plays a central role in *E. coli* metabolism as it is not only shared between the enzyme complexes for pyruvate dehydrogenase and 2-ketoglutarate dehydrogenase but also with the glycine-cleavage system. Therefore, the presence of *lpdA* in a genome does not necessarily indicate that the complete set of genes for one of the dehydrogenase complexes is present. For example, in the genome of *M. genitalium*, *lpdA* is present only as part of the pyruvate-dehydrogenase enzyme complex, whereas *M. thermoautotrophicum* and *M. jannaschii* have *lpdA* but do not have either of the dehydrogenase complexes.

Comparison with pathway databases

The original published genome annotations are not always correct. For the genes involved in the CAC, there are both over-predictions of occurrence (e.g. for aconitases and isocitrate dehydrogenases in some of the archaeal genomes) and under-predictions. Under-predictions are partly the result of a justifiably conservative approach to function prediction, and partly the result of experimentally analysing the protein's functions after the publication of the genome sequence. For example, a recently sequenced glyceraldehyde-3-phosphate ferredoxin oxidoreductase from *Pyrococcus furiosus*[28] has been discovered to have orthologues in *P. horikoshii* (PH0457) and *M. jannaschii* (MJ1185) that were, until now, annotated as open reading frames.

Errors in genome databases can propagate into pathway databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG; `http://www.tokyo-center.genome.ad.jp/kegg/`) or What is There (WIT; `http://wit.mcs.anl.gov/WIT2/`). It is beyond the scope of this review to quantify the quality of function prediction in genome and pathway databases in detail. The majority of predictions in genome databases, and in pathway databases such as KEGG and WIT, are consistent with the results

shown in Table 1. In a detailed analysis such as that presented here, genome and pathway databases serve as research tools and as a consistency check for predictions based on independent analyses. Inconsistencies between annotations trigger more-in-depth studies of the phylogenetic relations between proteins and circumstantial evidence for function. A complete list of improved annotations, including references to experimental results supporting them, is available from `http://dove.EMBL-Heidleberg.DE/Genome/Citric_Acid_Cycle`.

Outlook

The consistency of the results of sequence analysis with the biochemical data shows that comparative genome analysis can be a reliable tool for predicting pathways in a species. However, processes like gene displacement and sequence divergence imply that complementary biochemical data are necessary, especially as a considerable number of genes in the published genomes still have unknown functions.

One aspect of genome analysis that has not been addressed is the regulation of gene expression. Indeed, the co-occurrence of genes for a pathway in a genome does not guarantee that they are actually expressed together. Regulatory structures in DNA and RNA show very little conservation over the phylogenetic distances of the species compared here[9]. The main tool we have at present for the comparative prediction of co-expression of genes is operon structure. The next big challenge for bioinformatics lies at the level of predicting gene regulation. For this, the integration of gene-expression and proteomics data with genome-sequence data will be fundamental.

Acknowledgements

The research of M.A.H. has been made possible partly by a fellowship of the Royal Netherlands Academy ofArts and Sciences. This work was supported by BMBF. We thank the referees for their comments.

References

1 Gaasterland, T. and Selkov, E. (1995) *Ismb* 3, 127–135
2 Karp, P.D., Ouzounis, C. and Paley, P. (1996) *Ismb* 4, 116–124
3 Fuchs, G. and Stupperich, E. (1978) *Arch. Microbiol.* 118, 121–125
4 Melendez-Hevia, E., Wadell, T.G. and Cascante, M. (1996) *J. Mol. Evol.* 43, 293–303
5 Lucas, C. and Weitzman, P. (1977) *Arch. Microbiol.* 114, 55–60
6 Woese, C.R. (1987) *Microbiol. Rev.* 51, 221–271
7 Beh, M. *et al.* (1993) *Arch. Microbiol.* 160, 306–311
8 Romano, A.H. and Conway, T. (1996) *Res. Microbiol.* 147, 448–455
9 Huynen, M.A. and Bork, P. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 5849–5856
10 Dandekar, T. *et al.* (1998) *Trends Biochem. Sci.* 23, 324–328
11 Lawrence, J.G. and Roth, J.R. (1996) *Genetics* 143, 1843–1860
12 Hughes, N. *et al.* (1998) *J. Bacteriol.* 180, 1119–1128
13 Corthesy-Theulaz, I.E. *et al.* (1997) *J. Biol. Chem.* 41, 25659–25667
14 Heim, S. *et al.* (1998) *Eur. J. Biochem.* 253, 292–299
15 Bork, P. *et al.* (1998) *J. Mol. Biol.* 283, 707–725
16 Hoffman, P.S. *et al.* (1996) *J. Bacteriol.* 178, 4822–4829
17 Tomb, J-F. *et al.* (1997) *Nature* 388, 539–547
18 Manolukas, J.T. *et al.* (1988) *J. Gen. Microbiol.* 134, 791–800
19 Pollack, J.D. (1997) *Trends Microbiol.* 5, 413–419
20 Deckert, G. *et al.* (1998) *Nature* 392, 353–358
21 Ojcius, D.M. *et al.* (1998) *J. Biol. Chem.* 273, 7052–7058
22 Gonzalez, J.M. *et al.* (1998) *Extremophiles* 2, 123–130
23 Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) *Science* 278, 631–637
24 Jansen, K., Stepperich, E. and Fuchs, G. (1982) *Arch. Microbiol.* 132, 355–364
25 Altschul, S.F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
26 Huynen, M.A. *et al.* (1998) *J. Mol. Biol.* 280, 323–326
27 Sako, Y. *et al.* (1996) *FEBS Lett.* 392, 148–152
28 van der Oost, J. *et al.* (1998) *J. Biol. Chem.* 273, 28149–28154
29 Snel, B., Bork, P. and Huynen, M.A. (1999) *Nat. Genet.* 21, 108–110
30 Steen, I.H., Lein, T. and Birkeland, N-K. (1997) *Arch. Microbiol.* 168, 412–420
31 Fitch, W.M. (1970) *Syst. Zool.* 19, 99–110
32 Koonin, E.V., Mushegian, A.R. and Bork, P. (1996) *Trends Genet.* 12, 334–336