Gene sequences useful for predicting relatedness of whole genomes in bacteria

Daniel R. Zeigler

Bacillus Genetic Stock Center, Department of Biochemistry, The Ohio State University, Columbus, OH 43210, USA

Correspondence
Daniel R. Zeigler
zeigler.1@osu.edu

Thirty-two protein-encoding genes that are distributed widely among bacterial genomes were tested for the potential usefulness of their DNA sequences in assigning bacterial strains to species. From publicly available data, it was possible to make 49 pairwise comparisons of whole bacterial genomes that were related at the genus or subgenus level. DNA sequence identity scores for eight of the genes correlated strongly with overall sequence identity scores for the genome pairs. Even single-gene alignments could predict overall genome relatedness with a high degree of precision and accuracy. Predictions could be refined further by including two or three genes in the analysis. The proposal that sequence analysis of a small set of protein-encoding genes could reliably assign novel strains or isolates to bacterial species is strongly supported.

INTRODUCTION

The current concept of the bacterial species – a 'genomically coherent' group of strains that share many common traits (Rosselló-Mora & Amann, 2001) – rests squarely on the availability of reliable techniques to quantify the relatedness of bacterial genomes. Although many such techniques exist [for a recent review, see Gürtler & Mayall (2001)], analysis of DNA–DNA hybridization values remains the 'gold standard' for defining bacterial species (Wayne *et al.*, 1987; Stackebrandt *et al.*, 2002). Techniques required to obtain these values, however, tend to be expensive and time-consuming and often require specialized instruments or radioactive labels (Johnson, 1994). Furthermore, as many parameters affect DNA–DNA reassociation, hybridization values may be difficult to reproduce in different laboratories.

An alternative approach to quantification of genome relatedness is to compare selected DNA sequences for a group of bacterial strains. The core technology for this method, DNA sequencing, is relatively rapid and inexpensive, highly reproducible and readily available to virtually any research group through specialized sequencing centres. Databases of gene sequences and computer applications to compare them are, likewise, freely available. For these reasons, DNA sequence analysis has taken an increasingly important role in taxonomic studies in recent years.

The bacterial species concept, however, requires that entire genomes are compared. If sequence analysis is to augment, or even replace, DNA–DNA hybridization in defining species, it is paramount that taxonomists identify genes that can represent whole genomes reliably for the purposes of comparison. Recently, an ad hoc committee for the re-evaluation of the species definition in bacteria issued a call for identification of a set of such genes (Stackebrandt *et al.*, 2002). The committee's consensus was that analysis of at least five genes of diverse chromosomal loci and wide distribution could provide sufficient information to distinguish a bacterial species from related taxa. Once a species was defined in this way, sequence information from a single member of this gene set may be enough to assign additional strains to the species (Stackebrandt *et al.*, 2002).

It is an open question how much information any given gene sequence can provide about the genome that contains it. Sequence differences between related organisms within a given gene are presumably due to slow, continual acquisition of random mutations, which are subject to selection and inherited vertically. Differences seen in whole-genome comparisons, however, are the sum of this vertical inheritance and any number of horizontal transfer events that involve simultaneous acquisition of many genes through transformation, conjugation or bacteriophage infection. Genome reduction through gene inactivation and deletion further complicates the picture. The relative rates of these factors – random mutation, horizontal transfer and genome reduction – are poorly understood.

The goal of the current study is to obtain statistical evidence that individual gene sequences diverge at a rate that reflects the overall rate of genome divergence and to identify genes that could best serve as predictors of genome relatedness. Publicly available bacterial genome sequences were used to

Published online ahead of print on 9 May 2003 as DOI 10.1099/ijs.0.02713-0.

A spreadsheet listing sequence identity scores for all candidates with each genome comparison is available as supplementary data in IJSEM Online.

identify over 30 genes that satisfy the ad hoc committee's criteria (Stackebrandt *et al.*, 2002). When closely related bacteria were compared, the frequency of identical residues in individual gene alignments correlated with the frequency of identical residues in whole-genome alignments with $R^2 \geqslant 0{\cdot}9$ for each of eight genes from this set. The highest-scoring sequence from the set, *recN*, could be used to predict whole-genome relatedness with high accuracy for a test set of 44 bacterial genomes. Combining data from two or three genes could further refine this prediction. It appears that a small number of carefully selected gene sequences can indeed equal, or perhaps even surpass, the precision of DNA–DNA hybridization for quantification of genome relatedness.

METHODS

Genomic sequence data. All genomic sequences analysed in this study were obtained from publicly available databases. GenBank accession numbers for the complete but unannotated *Neisseria gonorrhoeae* and *Bacillus anthracis* genomes are AE004969 (available at ftp://ftp.genome.ou.edu/pub/gono/) and AAAC01000001, respectively. Sequences for *Bordetella bronchiseptica* RB50, *Bordetella parapertussis* 12822, *Bordetella pertussis* Tahoma 1, *Chlamydophila abortus*, *Corynebacterium diphtheriae* NCTC 13129, *Mycobacterium bovis* AF2122/97, *Neisseria meningitidis* FAM18 and *Yersinia enterocolitica* 8081 were obtained from the Sanger Institute microbial genomes site (http://www.sanger.ac.uk/Projects/Microbes/). All other sequences were downloaded from the NCBI microbial genomes site (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html). Individual gene sequences were obtained from the genomic sequences, following confirmation through BLAST analysis of their uniqueness within the genome and their high sequence similarity ($P<10^{-6}$) to their *Escherichia coli* K-12 and *Bacillus subtilis* 168 orthologues.

Individual gene and whole-genome alignments. For calculating DNA sequence identity for individual genes, sequences obtained from related organisms were aligned with CLUSTAL W and a distance matrix was computed (Thompson *et al.*, 1994). Pairs of whole genomes were aligned by using the NUCMER application (Delcher *et al.*, 2002) with the following parameters: breakLen = 500, minCluster = 40, diagFactor = 0·15, maxGap = 250 and minMatch = 12. To ensure that the algorithm found all possible alignments, each pair was analysed twice with the reference and query files swapped. The resulting output files, giving the coordinates of regions of sequence similarity between the two genomes, were combined and duplicate regions were removed from the list. When two neighbouring regions shared overlapping end-points, the common segment was divided equally between them. Two similarity estimates were calculated from each genomic sequence comparison. DNA sequence identity for conserved regions was calculated as the mean sequence identity of the homologous regions, weighted by each region's length in nucleotides. DNA sequence identity for a pair of whole genomes was calculated by multiplying the sequence identity of their conserved regions by the ratio of the net length of the conserved regions to the mean length of the two genomes.

Statistical analysis. Univariate linear regression models were used to assess the predictive ability of sequence identities for each individual gene with respect to sequence identities for whole-genome alignments. Step-wise linear regression procedures were used to find the subset of genes that best predicted the whole-genome alignment. Prediction intervals were calculated and the upper/lower limits of these intervals were used to determine the cut-point for the desired 70 % alignment.

RESULTS AND DISCUSSION

DNA sequence identity between related bacterial genomes

Among publicly available bacterial genome sequences are several groups of organisms that are related at the genus or species level. At the beginning of this study (in July 2002), 44 genomes that could be grouped into 16 genera were identified (Table 1). It was known that the genera *Escherichia*, *Salmonella* and *Shigella* were highly related (Brenner *et al.*, 1969, 1972; Crosa *et al.*, 1973), so they were treated as if they belonged to a single genus. Genome sequences within each of the 16 groups were aligned in pairwise fashion by the NUCMER application (Delcher *et al.*, 2002) as described in Methods. These alignments identified two sets of sequences, those conserved between both genomes and those unique to each genome. The first set would primarily include elements of the ancestral genome that have been inherited vertically by both species. DNA sequence identity of these conserved regions can be quantified (Table 1), providing an index of the degree of genome divergence that can be explained by random mutation and selection. DNA sequence identity for whole-genome sequences can also be quantified to measure divergence due to all processes, including vertical and horizontal transfer and genome reduction (Table 1).

It is encouraging that the whole-genome sequence identity figures calculated in this study correlate well with available genome similarity measurements obtained by DNA–DNA hybridization. For example, Brenner *et al.* (1972) and Crosa *et al.* (1973) obtained similarity estimates for *Salmonella typhimurium* with *Salmonella typhi* (88 %), *E. coli* (45–46 %) and *Shigella flexneri* (38–39 %) and for *E. coli* with *Shigella flexneri* (84 %). Among the *Chlamydiaceae*, similarity estimates for *Chlamydia muridarum* with *Chlamydia trachomatis* (65 %) (Weiss *et al.*, 1970) and among various strains of *Chlamydophila pneumoniae* (94–100 %) (Fukushi & Hirai, 1989) were also comparable with the estimates in this study. Finally, Somerville & Jones (1972) measured no detectable competition between *Bacillus subtilis* and *Bacillus anthracis* genomic DNA under conditions in which *Bacillus anthracis* competed with other members of the *Bacillus cereus* group at levels from 59 to 100 %. Whilst these hybridization studies were conducted by differing protocols, the fact that they yielded similar results to those found in Table 1 suggests that DNA–DNA hybridization studies and whole-genome alignments are measuring the same quantity, i.e. sequence identity, by compatible methods.

These data allow us to estimate how well one factor involved in genome divergence – random mutation within vertically transferred genetic material – correlates with total divergence between pairs of genomes. Univariate regression analysis of whole-genome sequence identity with respect to conserved-region sequence identity (Fig. 1a) showed an excellent fit to the linear model ($P<0{\cdot}001$, $R^2=0{\cdot}871$). The

Table 1. DNA sequence comparisons for related bacterial genomes

'Grouping' indicates the genera (or in the case of *Escherichia*, *Shigella* and *Salmonella*, the closely related genera) used to group the sequences for alignment studies. Genomes within each group were aligned in pairwise fashion by using the NUCMER application, as described in Methods. 'Conserved-region' and 'whole-genome' DNA sequence identity scores are described in the text.

Grouping and genome comparison		**DNA sequence identity**	
		Conserved regions	**Whole genomes**
Bacillus			
Bacillus anthracis A2012	*Bacillus halodurans* C-125	0·742	0·039
Bacillus anthracis A2012	*Bacillus subtilis* 168	0·747	0·057
Bacillus halodurans C-125	*Bacillus subtilis* 168	0·769	0·052
Buchnera			
Buchnera aphidicola Sg^{T}	*Buchnera* sp. APS	0·780	0·705
Chlamydia			
Chlamydia muridarum MoPn^{T}	*Chlamydia trachomatis* D/UW-3/CX	0·809	0·736
Chlamydophila			
Chlamydophila pneumoniae AR39	*Chlamydophila pneumoniae* CWL029	0·999	0·999
Chlamydophila pneumoniae AR39	*Chlamydophila pneumoniae* J138	0·999	0·999
Chlamydophila pneumoniae CWL029	*Chlamydophila pneumoniae* J138	0·999	0·999
Clostridium			
Clostridium acetobutylicum ATCC 824^{T}	*Clostridium perfringens* 13	0·754	0·086
***Escherichia*-like**			
Escherichia coli K-12 MG1655	*Escherichia coli* O157 : H7 EDL933	0·977	0·812
Escherichia coli K-12 MG1655	*Salmonella typhi* CT18	0·809	0·534
Escherichia coli K-12 MG1655	*Salmonella typhimurium* LT2	0·809	0·537
Escherichia coli K-12 MG1655	*Shigella flexneri* 2a 301	0·978	0·859
Escherichia coli O157 : H7 EDL933	*Salmonella typhi* CT18	0·808	0·490
Escherichia coli O157 : H7 EDL933	*Salmonella typhimurium* LT2	0·807	0·500
Escherichia coli O157 : H7 EDL933	*Shigella flexneri* 2a 301	0·975	0·795
Salmonella typhi CT18	*Salmonella typhimurium* LT2	0·982	0·883
Salmonella typhi CT18	*Shigella flexneri* 2a 301	0·813	0·536
Salmonella typhimurium LT2	*Shigella flexneri* 2a 301	0·809	0·520
Helicobacter			
Helicobacter pylori 26695	*Helicobacter pylori* J99	0·930	0·876
Listeria			
Listeria innocua CLIP 11262	*Listeria monocytogenes* EGD-e	0·882	0·779
Mycobacterium			
Mycobacterium leprae	*Mycobacterium tuberculosis* CDC1551	0·777	0·360
Mycobacterium leprae	*Mycobacterium tuberculosis* H37Rv	0·777	0·362
Mycobacterium tuberculosis CDC1551	*Mycobacterium tuberculosis* H37Rv	0·998	0·993
Mycoplasma			
Mycoplasma genitalium G-37^{T}	*Mycoplasma pneumoniae* ATCC 29342	0·749	0·253
Mycoplasma genitalium G-37^{T}	*Mycoplasma pulmonis* UAB CT	0·725	0·018
Mycoplasma pneumoniae ATCC 29342	*Mycoplasma pulmonis* UAB CT	0·743	0·012
Neisseria			
Neisseria meningitidis MC58	*Neisseria meningitidis* Z2491	0·957	0·908
Rickettsia			
Rickettsia conorii NIAID Malish 7^{T}	*Rickettsia prowazekii* Madrid E	0·847	0·654
Staphylococcus			
Staphylococcus aureus Mu50	*Staphylococcus aureus* MW2	0·984	0·922
Staphylococcus aureus Mu50	*Staphylococcus aureus* N315	0·999	0·981
Staphylococcus aureus MW2	*Staphylococcus aureus* N315	0·985	0·933
Streptococcus			
Streptococcus agalactiae 2603V/R	*Streptococcus pneumoniae* R6	0·787	0·164
Streptococcus agalactiae 2603V/R	*Streptococcus mutans* UA159	0·768	0·239
Streptococcus agalactiae 2603V/R	*Streptococcus pyogenes* M1 GAS SF370	0·800	0·296

Table 1. cont.

Grouping and genome comparison		DNA sequence identity	
		Conserved regions	Whole genomes
Streptococcus agalactiae 2603V/R	*Streptococcus pyogenes* MGAS315	0·794	0·278
Streptococcus agalactiae 2603V/R	*Streptococcus pyogenes* MGAS8232	0·794	0·279
Streptococcus pneumoniae R6	*Streptococcus mutans* UA159	0·769	0·188
Streptococcus pneumoniae R6	*Streptococcus pyogenes* M1 GAS SF370	0·780	0·186
Streptococcus pneumoniae R6	*Streptococcus pyogenes* MGAS315	0·779	0·185
Streptococcus pneumoniae R6	*Streptococcus pyogenes* MGAS8232	0·779	0·189
Streptococcus mutans UA159	*Streptococcus pyogenes* M1 GAS SF370	0·765	0·232
Streptococcus mutans UA159	*Streptococcus pyogenes* MGAS315	0·761	0·222
Streptococcus mutans UA159	*Streptococcus pyogenes* MGAS8232	0·766	0·226
Streptococcus pyogenes M1 GAS SF370	*Streptococcus pyogenes* MGAS315	0·984	0·904
Streptococcus pyogenes M1 GAS SF370	*Streptococcus pyogenes* MGAS8232	0·982	0·914
Streptococcus pyogenes MGAS315	*Streptococcus pyogenes* MGAS8232	0·982	0·925
Xanthomonas			
Xanthomonas axonopodis 306	*Xanthomonas campestris* ATCC 33913^T	0·844	0·523
Yersinia			
Yersinia pestis CO92	*Yersinia pestis* KIM	0·999	0·997

simplest interpretation of this result is that during bacterial speciation, the various forces that change genome sequence content act at discrete rates in a time-dependent manner. As a result, the ratio between sequence differences in conserved regions and overall genome sequence differences remains fairly constant, at least while the bacteria are related as closely as those analysed here. This interpretation, if true, means that the proposal of Stackebrandt *et al.* (2002) is quite reasonable: a rational definition of bacterial species could be based on sequence analysis of a set of conserved genes. If the relatedness of whole genomes can be measured by examining the subset of genes they share, then a small but representative set of shared genes should successfully predict genome relatedness.

DNA sequence identity between individual gene orthologues

Candidate genes for a 'species prediction set' were identified from these bacterial genomes by applying four criteria that were consistent with the overall goal of this study: to develop a practical tool for bacterial taxonomy. First, the genes must be widely distributed among genomes, with orthologous sequences appearing in most, if not all, free-living bacteria. Secondly, because the occurrence of gene families could make sequencing and alignment technically difficult, each of the 'prediction set' genes must be unique within a given genome, without close paralogues that could confuse analysis. Thirdly, individual gene sequences must

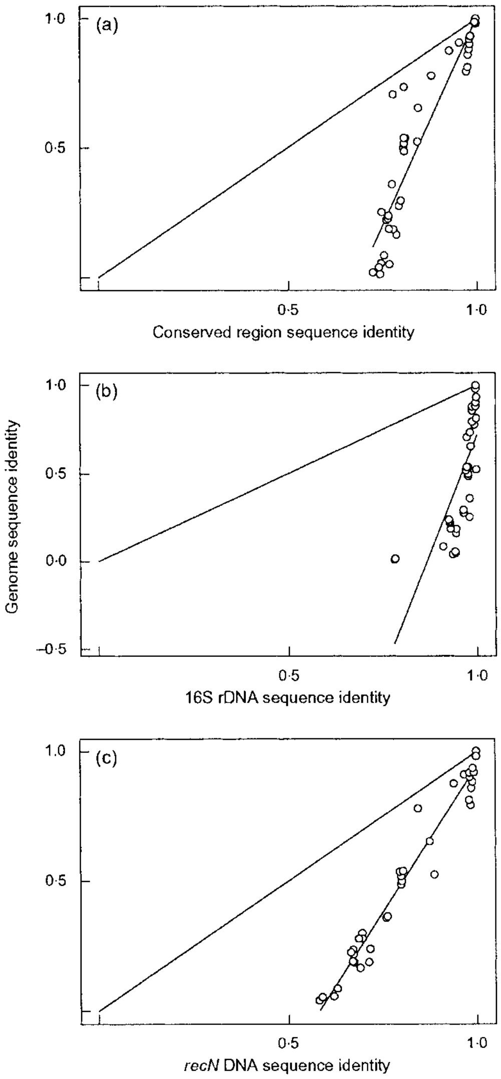

Fig. 1. Linear regression analysis of whole-genome sequence identity versus sequence identities of (a) conserved regions; (b) 16S rDNA sequences; and (c) *recN* sequences. Lines indicate the calculated best fit for the data and the hypothetical fit if the slope of the regression line was equal to 1.

be long enough to contain phylogenetically useful information but short enough to be sequenced economically with a small set of primers. Finally, the sequences must predict whole-genome relationships with acceptable precision and accuracy. The first criterion was satisfied by preliminary genome comparisons among the bacteria listed in Table 1, which yielded a pool of over 100 candidate genes (not shown). BLAST searches eliminated candidates that had close paralogues within one or more of the genomes. Sequences that encoded merely hypothetical proteins were also struck from the list, as were those that were deemed to be shorter (<900 bp) or longer (>2250 bp) than optimum. The 32 candidate genes that remained after application of these criteria are listed in Table 2. The 16S rRNA gene, although it does not encode a protein, was added to the list of candidates because of its wide usage in taxonomic studies.

It remained to be determined which of the candidate genes best satisfied the final criterion by predicting whole-genome relationships. For each gene listed in Table 2, orthologous sequences from each of the genomes were divided into genus groups (listed in Table 1). Sequences

Table 2. Single-variable regression analysis for whole-genome sequence identity versus individual gene sequence identity

Gene names used for orthologous sequences are consistently those used in the *E. coli* genome annotation. Gene nomenclature may vary among bacterial species. Orthologous sequences are found in each of the genomes analysed in this study, except as listed under 'Exceptions'. Abbreviations: BAC-ANTH, *Bacillus anthracis*; BUC, *Buchnera*; CHA, *Chlamydia*; CHO, *Chlamydophila*; MYB, *Mycobacterium*; MYP, *Mycoplasma*; MYP-PULM, *Mycoplasma pulmonis*; NEI, *Neisseria*; RIC, *Rickettsia*; STA, *Staphylococcus*; STR, *Streptococcus*; STR-8232, *Streptococcus pyogenes* MGAS8232.

Gene	Product	Exceptions	Slope	SE	R^2
recN	Recombination/repair protein	BUC, CHA, CHO, MYP	2·2460	0·0682	0·965
lig	DNA ligase	BUC, CLO, MYP-PULM	2·2416	0·0790	0·949
dnaX	DNA polymerase III subunits γ, τ		2·2350	0·0801	0·943
glyA	Serine hydroxymethyltransferase	MYB	2·5261	0·1162	0·915
cysS	Cysteine tRNA synthetase		2·4188	0·1095	0·912
thdF	GTP-binding, thiophene oxidation		2·4473	0·1108	0·912
uvrC	Excinuclease ABC, subunit C	CHA	2·2934	0·1063	0·910
ruvB	Holliday junction helicase subunit A	BUC	2·5633	0·1234	0·904
metG	Methionine tRNA synthetase		2·5697	0·1300	0·893
dnaB	Replicative DNA helicase		2·3480	0·1228	0·886
dnaJ	Chaperone with DnaK	MYB	2·3990	0·1305	0·884
lepA	GTP-binding elongation factor	STR-8232	2·9402	0·1646	0·884
argS	Arginine tRNA synthetase		2·3695	0·1296	0·877
ffh	GTP-binding export factor		2·7375	0·1519	0·874
–	All similar sequences in genome		3·2042	0·1797	0·871
serS	Serine tRNA synthetase		2·4724	0·1385	0·871
nrdE	NDP reductase 2, α-subunit	BAC-ANTH	2·5769	0·1534	0·862
ftsZ	Tubulin-like division protein	CHA, CHO	2·3674	0·1451	0·861
metK	Methionine adenosyltransferase	CHA, CHO, RIC	2·6877	0·1670	0·860
trpS	Tryptophan tRNA synthetase		2·4925	0·1518	0·852
atpA	ATP synthase, F1, α-subunit	CHA, CHO	2·7117	0·1729	0·851
dxs	Deoxyxylulose-phosphate synthase	MYP, RIC, STA, STR	2·6133	0·2231	0·846
uvrB	Excision nuclease subunit B	CHA	2·5188	0·1584	0·846
atpD	ATP synthase, F1, β-subunit		3·1999	0·2103	0·843
aspS	Aspartate tRNA synthetase		2·1103	0·1349	0·839
pgi	Glucose phosphate isomerase	NEI, RIC	2·5394	0·1703	0·832
tig	Trigger factor		2·3144	0·1667	0·804
rho	Transcription termination factor	BAC-ANTH, MYP, STR	2·8494	0·2930	0·778
proS	Proline tRNA synthetase		1·9796	0·1575	0·771
recA	DNA strand exchange and renaturation	BAC-ANTH, BUC	2·2265	0·1936	0·750
rpoA	RNA polymerase, alpha subunit		2·4117	0·2245	0·711
eno	Enolase	RIC	2·7169	0·2890	0·658
trpS	Tryptophan tRNA synthetase		1·9515	0·2336	0·598
pgk	Phosphoglycerate kinase	RIC	2·2234	0·2780	0·582
rrsH	16S rRNA		5·4228	0·7354	0·536

within each group were analysed following multiple alignment to compute a distance matrix that quantified sequence identity. Therefore, for each genome comparison, there was a sequence identity score for the whole genomes as well as 33 identity scores for individual genes. A spreadsheet listing sequence identity scores for all candidates with each genome comparison is available as supplementary data in IJSEM Online. These sequence identity scores were analysed by linear regression to determine which gene comparison could best predict genome relatedness.

For each gene, the plot of genome sequence identity versus individual gene sequence identity fit a linear model with $P<0·01$. Goodness of fit varied widely among the candidate genes, with R^2 values ranging from 0·536 to 0·965. Eight of the genes had an R^2 value of 0·9 or better, making them outstanding candidates for a 'species prediction' sequence set (Table 2). The individual candidate gene with the poorest ability to predict genome relatedness on a genus or species level was 16S rDNA, the gene that encodes 16S rRNA (Fig. 1b); as others have noted (Fox *et al.*, 1992; Stackebrandt & Goebel, 1994), it is simply too highly conserved to differentiate reliably between closely related taxa. The candidate gene with the greatest potential for predicting genome relatedness at the genus or subgenus level was *recN* (Fig. 1c), a recombination and repair protein-encoding gene that is found in each of the free-living bacterial genomes analysed, as well as in the two *Rickettsia* species. Among genes found in every bacterial genome analysed, the highest-scoring candidate was *dnaX*, a gene that encodes two subunits of DNA polymerase III. Whilst *recN* is slightly superior to *dnaX* in terms of fit to whole-genome data (R^2, 0·965 and 0·943, respectively), the latter sequence may prove to be particularly useful in analysis of taxa characterized by genome reduction.

A parallel study analysed amino acid sequence identities of the predicted products for each of the candidate genes (not shown). In each case, the gene sequences showed a better fit to genome relatedness than the predicted protein sequences. Whilst protein sequences have often been analysed to compare distantly related organisms or ancient gene duplications (Brown *et al.*, 2001; Delcher *et al.*, 2002), DNA sequences may be more useful for distinguishing close phylogenetic relationships.

This 'bacterial species prediction' set differs from sequence sets that were assembled for the purpose of constructing universal phylogenetic trees (Brown *et al.*, 2001). The purpose of those studies is to detect relationships among very distantly related organisms at the domain level, whereas the purpose of the current study is to distinguish between closely related organisms at the species level. Many of the highest scoring sequences in Table 2 have no known orthologues outside bacteria and so cannot be used to construct universal trees. Further, the practical aims of the current study impose selection criteria that would be unnecessary in a universal tree study, such as moderate gene length and absence of close paralogues. Nevertheless, there is some overlap, including certain tRNA synthetases and DNA and RNA polymerase subunits, between the bacterial 'species prediction' set developed in this study and the 'universal tree' set of Brown *et al.* (2001).

Predictive models

Linear regression analysis yielded simple predictive models for high-scoring candidates. Genome relatedness for two related bacterial strains can be predicted by the following formula:

$$SI_{\text{genome}} = -1·30 + 2·25(SI_{recN})$$

where SI_{genome} is the predicted DNA sequence identity shared by the genomes and SI_{recN} is the sequence identity shared by their *recN* orthologues. The actual genome sequence identity is plotted against the *recN* prediction in Fig. 2(a) for each of the genomes listed in Table 1. The

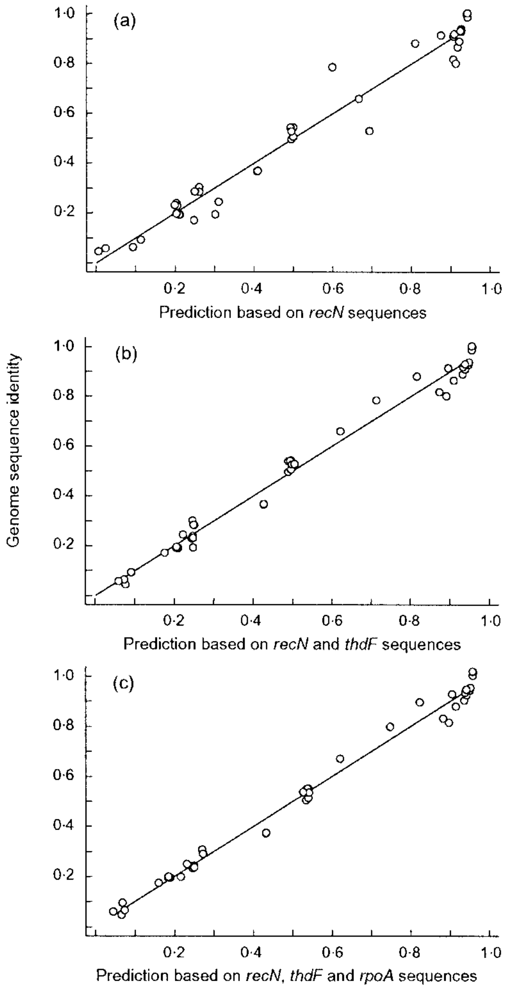

Fig. 2. Linear regression analysis of whole-genome sequence identity scores versus scores predicted based on the models described in the text for (a) *recN*; (b) *recN* and *thdF*; and (c) *recN*, *thdF* and *rpoA*.

prediction is surprisingly powerful: it fits a linear model with $P<0{\cdot}001$ and $R^2=0{\cdot}965$. The mean absolute residual (i.e. the mean difference between the predicted and actual genome sequence identity) is only 0·044. Similar results (not shown) were obtained by using *dnaX* sequences with the following formula:

$$SI_{\text{genome}}=-1{\cdot}29+2{\cdot}24(SI_{dnaX})$$

A 95 % prediction interval can be obtained from the *recN* prediction formula above. Based on these results, one can make the following conclusions: if the *recN* DNA sequences for two bacterial strains or isolates are <84 % similar, we can be 95 % confident that their genome sequences are <70 % similar and that the bacteria belong to different species. If the *recN* DNA sequences are >96 % similar, then by the same reasoning, we can be 95 % confident that the bacteria belong to the same species. If the *recN* sequences are between 84 and 96 % similar, it is questionable whether the genome sequence identity is greater than or less than 70 %, making the species identity for these strains uncertain.

Step-wise linear regression procedures were used to determine whether inclusion of sequence data from other candidate genes improved the predicting power of *recN*. The results suggest that the best two-gene combination (that is, the combination with the highest R^2 value) was *recN* and *thdF*, which encodes a thiophene oxidation enzyme (Fig. 2b), whereas the best three-gene combination was *recN*, *thdF* and *rpoA*, which encodes the RNA polymerase α-subunit (Fig. 2c). Interestingly, the best sets used a combination of genes that were highly (*rpoA*), moderately (*thdF*) and weakly (*recN*) conserved. Prediction models resulting from this analysis were:

$$SI_{\text{genome}}=-1{\cdot}67+0{\cdot}707(SI_{recN})+1{\cdot}92(SI_{thdF})$$

$$SI_{\text{genome}}=-1{\cdot}88+0{\cdot}52(SI_{recN})+1{\cdot}78(SI_{thdF})+0{\cdot}52(SI_{rpoA})$$

where SI_{genome} is the predicted DNA sequence identity shared by the genomes and SI_{recN}, SI_{thdF} and SI_{rpoA} are the sequence identities shared by the *recN*, *thdF* and *rpoA* orthologues, respectively. For the two- and three-gene models, R^2 values improved to 0·986 and 0·989, respectively, and the mean absolute residuals improved to 0·032 and 0·026, respectively.

Bacterial genome sequences continue to become available in increasing numbers. Following the statistical modelling phase of this study, public databases were re-examined for new whole-genome sequences that either fit within the genus groups analysed before or defined new groups. Several new comparisons were possible, serving as a test of the validity of the predictive models (Table 3). The genomes were analysed as before and for each genome pair, sequence identity scores were computed for whole genomes and for the predictor genes *recN*, *thdF* and *rpoA*. The single-, two- and three-gene sets predicted whole-genome sequence identity with mean residual values of 0·049, 0·032 and 0·040, respectively. This second test set of comparisons yielded results that were completely in harmony with those of the first set of comparisons.

Table 3. Predictions of DNA sequence identity for pairs of genomes

Predictions 1, 2 and 3 are based on the single-, two- and three-gene prediction models that are described in the text.

Genome comparison		DNA sequence identity				Genome prediction		
		recN	***rpoA***	***thdF***	**Genome**	**1**	**2**	**3**
Corynebacterium diphtheriae	*Corynebacterium efficiens*	0·584	0·821	–	0·118	0·014	–	–
Corynebacterium diphtheriae	*Corynebacterium glutamicum*	0·591	0·847	–	0·126	0·030	–	–
Corynebacterium efficiens	*Corynebacterium glutamicum*	0·735	0·872	–	0·336	0·354	–	–
Escherichia coli CFT073	*Escherichia coli* K-12	0·969	1·000	0·950	0·790	0·880	0·839	0·835
Escherichia coli CFT073	*Escherichia coli* O157 : H7	0·972	1·000	0·947	0·765	0·887	0·835	0·831
Escherichia coli CFT073	*Salmonella typhi*	0·802	0·977	0·834	0·509	0·505	0·498	0·530
Escherichia coli CFT073	*Salmonella typhimurium*	0·803	0·976	0·838	0·511	0·507	0·507	0·537
Escherichia coli CFT073	*Salmonella flexneri*	0·969	0·997	0·953	0·752	0·880	0·845	0·839
Mycobacterium bovis	*Mycobacterium tuberculosis* CDC	0·999	1·000	1·000	0·991	0·948	0·956	0·939
Mycobacterium bovis	*Mycobacterium leprae*	0·758	0·883	0·811	0·366	0·406	0·423	0·417
Neisseria meningitidis FAM	*Neisseria gonorrhoeae*	0·954	0·993	0·952	0·804	0·847	0·832	0·827
Neisseria meningitidis FAM	*Neisseria meningitidis* MC58	0·979	1·000	0·974	0·907	0·903	0·892	0·883
Neisseria gonorrhoeae	*Neisseria meningitidis* MC58	0·956	0·993	0·952	0·829	0·851	0·834	0·828
Neisseria gonorrhoeae	*Neisseria meningitidis* Z2491	0·952	0·993	0·961	0·811	0·842	0·848	0·842
Staphylococcus epidermidis	*Staphylococcus aureus* MU50	0·775	0·945	0·805	0·406	0·444	0·424	0·447
Staphylococcus epidermidis	*Staphylococcus aureus* MW2	0·775	0·945	0·807	0·409	0·444	0·427	0·451
Staphylococcus epidermidis	*Staphylococcus aureus* N315	0·775	0·945	0·806	0·412	0·444	0·425	0·449
Yersinia enterocolitica	*Yersinia pestis* KIM	0·822	0·985	0·851	0·523	0·550	0·545	0·574

In conclusion, the hypothesis that sequences from protein-encoding genes can predict genome relatedness accurately is supported strongly by these studies. Before conducting full-scale studies based on these concepts, however, researchers should keep at least two cautions in mind. First, these results are based on a limited number of bacterial species; as additional genome data become available, the statistical models presented in this paper will doubtless require refinement. Secondly, it is certainly possible that within a given genus, phylogenies constructed by using a small number of genes can be perturbed radically if horizontal transfer has occurred among those genes. Nevertheless, the results in this paper demonstrate that analysis of even a single carefully selected gene, such as *recN* or *dnaX*, could be a powerful tool for discriminating between species represented among a large group of bacterial isolates. Clearly a set of 'species predictor' genes has superior precision when compared to a single predictor gene, but there is a diminishing degree of improvement as each new sequence is added. The five-gene set envisioned by Stackebrandt *et al.* (2002) may be more genes than are actually required to equal or even surpass the power of DNA–DNA reassociation in assigning related bacterial isolates to species.

ACKNOWLEDGEMENTS

The author thanks T. Davidsen and A. Phillippy for helpful advice on using the NUCMER package, S. Shew and M. Abdullah for technical assistance in software implementation and S. Hoshaw-Woodard for performing the statistical analyses reported in this paper.

REFERENCES

Brenner, D. J., Fanning, G. R., Johnson, K. E., Citarella, R. V. & Falkow, S. (1969). Polynucleotide sequence relationships among members of *Enterobacteriaceae*. *J Bacteriol* **98**, 637–650.

Brenner, D. J., Fanning, G. R., Skerman, F. J. & Falkow, S. (1972). Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. *J Bacteriol* **109**, 933–965.

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet* **28**, 281–285.

Crosa, J. H., Brenner, D. J., Ewing, W. H. & Falkow, S. (1973). Molecular relationships among the *Salmonelleae*. *J Bacteriol* **115**, 307–315.

Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478–2483.

Fox, G. E., Wisotzkey, J. D. & Jurtshuk, P., Jr (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* **42**, 166–170.

Fukushi, H. & Hirai, K. (1989). Genetic diversity of avian and mammalian *Chlamydia psittaci* strains and relation to host origin. *J Bacteriol* **171**, 2850–2855.

Gürtler, V. & Mayall, B. C. (2001). Genomic approaches to typing, taxonomy and evolution of bacterial isolates. *Int J Syst Evol Microbiol* **51**, 3–16.

Johnson, J. L. (1994). Similarity analysis of DNAs. In *Methods for General and Molecular Bacteriology*, pp. 655–682. Edited by P. Gerhardt, R. G. E. Murray, W. A. Wood & N. R. Krieg. Washington, DC: American Society for Microbiology.

Rosselló-Mora, R. & Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* **25**, 39–67.

Somerville, H. J. & Jones, M. L. (1972). DNA competition studies within the *Bacillus cereus* group of bacilli. *J Gen Microbiol* **73**, 257–265.

Stackebrandt, E. & Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44**, 846–849.

Stackebrandt, E., Frederiksen, W., Garrity, G. M. & 10 other authors (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* **52**, 1043–1047.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

Wayne, L. G., Brenner, D. J., Colwell, R. R. & 9 other authors (1987). International Committee on Systematic Bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol* **37**, 463–464.

Weiss, E., Schramek, S., Wilson, N. N. & Newman, L. W. (1970). Deoxyribonucleic acid heterogeneity between human and murine strains of *Chlamydia trachomatis*. *Infect Immun* **2**, 24–28.