4482–4486 Nucleic Acids Research, 1998, Vol. 26, No. 19 © 1998 Oxford University Press

Concordance analysis of microbial genomes

Robert E. Bruccoleri*, Thomas J. Dougherty and Daniel B. Davison

Bristol-Myers Squibb, Pharmaceutical Research Institute, PO Box 4000, Princeton, NJ 08543-4000, USA

Received May 18, 1998; Revised and Accepted August 12, 1998

ABSTRACT

The set of proteins which are conserved across families of microbes contain important targets of new anti-microbial agents. We have developed a simple and efficient computational tool which determines concordances of putative gene products that show sets of proteins conserved across one set of user specified genomes and not present in another set of user specified genomes. The thresholds and the homology scoring criterion are selectable to allow the user to decide the stringency of the homologies. The system uses a relational database to store protein coding regions from different genomes, and to store the results of a complete comparison of all sequences against all sequences using the FASTA program. Using Web technology, the display of all the related proteins for a given sequence and calculation of multiple sequence alignments (using CLUSTALW) can be performed with the click of a button. The current database holds 97 365 sequences from 19 complete or partial genomes and 8 798 905 FASTA comparison results. A example concordance is presented which demonstrates that the target of the quinolone antibiotics could have been identified using this tool.

INTRODUCTION

The emergence of antibiotic resistance in pathogenic bacteria represents a major threat to public health (1). Due to the rapid growth rates of microbial populations, evolution of resistance can occur in relatively brief time frames. Moreover, it is clear that genetic exchange systems can rapidly disperse resistance genes among diverse bacterial species (2). It is therefore essential to develop rapid processes to identify novel antibiotic chemotypes. With the advent of complete bacterial genomic sequences, new insights into bacterial physiology from this information can suggest additional key reactions to serve as targets for antibacterial intervention. The sequence of every potential coding region in many bacterial species are now known (3–15), and large scale comparisons can provide valuable clues as to metabolism and pathogenic mechanisms. However, the great challenge in this analysis is the sheer volume of data. Even relatively simple bacteria such as the mycoplasmas have several hundred genes, and most pathogens have thousands of potential reading frames.

In our practice, we search for conservation across different species of bacteria, and we exclude proteins which are homologous to those found in the yeast genome. Using this tool, we have successfully identified a number of promising protein targets for new antibiotic development conserved among several pathogenic bacteria. The concordance analysis also has utility in retrospectively identifying targets of clinically proven agents, as well as potential conserved resistance elements, such as antibiotic efflux pumps.

One key aspect of our design is to allow the user to specify the criteria that determine if two sequences are related. As the statistical judgments of similarity are still controversial, we felt that the user of the system should be free to select the similarity criterion. The ability to make this selection contrasts with the Clusters of Orthologous Groups (16), where fixed rules were used to relate orthologous gene products.

The implementation of this design requires tabulating all possible sequence homologies within a very low level of similarity, and executing specific queries each time the system is used. Relational database technology provides adequate performance for queries to be answered in a time scale of minutes on a server like the Silicon Graphics Origin 2000. We use the PostgreSQL relational database system (http://www.postgresql.org) (17), because it is efficient, easy to program and freely redistributable. The user interface is the Web, and the processing of user requests is done using Perl (18) and the PostgreSQL/Perl interface written by Edmund Mergl (available as part of PostgreSQL).

To calculate and store all pairs of sequence homologies within very low limits of sequence similarity, we chose the FASTA program. (19). It has the ability to perform gapped alignments, and it also provides several different measures of the quality of the alignment, in particular, *Z*-scores, expectation of randomness, and percent similarity over the homologous regions of sequence.

MATERIALS AND METHODS

Our database currently has three large tables for storing these homologies. The first table contains all the protein sequences using the attributes shown in Table 1. There are two comparison tables, one for each scoring matrix, BLOSUM 62 (20) and PAM 250 (21). Both tables have the identical structure shown in Table 2. All name columns are indexed using a hash table, and the numeric columns are indexed using a B-tree index (22). These indices are necessary for reasonable retrieval performance.

*To whom correspondence should be addressed at present address: Center for Advanced Biotechnology and Medicine, Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854, USA. Tel: +1 732 235 5796; Fax: +1 732 235 4850; Email: bruc@acm.org

Table 1. Protein sequence table

Name	Organism prefix followed by gene name. The prefix is a two letter code for the organism which simplifies searches. The gene name is the name used by the group which published the genome.
Purpose	Brief description of the gene product's function. This is taken from the annotation provided by the group which published the genome.
Seq	Amino acid sequence.
Organism	Full name of organism.
Compared	Boolean flag indicating whether this sequence has been compared against the database.

Table 2. Comparison table

Name1	Name of first sequence in a comparison.
Name2	Name of second sequence in a comparison. Collated after Name1.
Z-score	*Z*-score from FASTA.
Expectation	Probability that the sequence alignment is random.
Local overlap ratio	Ratio of identical amino acids to number of amino acids in region of overlap.
Local overlap count	Number of amino acids in region of overlap.
Global overlap ratio	Ratio of identical amino acids to the maximum of the number of amino acids in both sequences.

In addition to these large tables, there are several others which provide essential information. The prefix table provides the relationship between the organism name and the organism abbreviation. The compare_code table lists the command options used to run FASTA with each scoring matrix used.

In order to save time when duplicate queries are made, the results of each concordance query is saved along with all the options used to generate it. If a user requests a concordance which has already been determined, the system will simply return the result. These saved results are deleted whenever new sequences are added to the database.

In order to prepare a concordance of gene products for a given organism, it is necessary to specify the following information. (i) The target genome. (ii) A reference list of species for which homologs to the target genome are to be identified. (iii) The criteria for selecting homologs. (iv) The 'match number' test. The 'match number' specifies the minimum number of reference species where the target must be found in order for a target gene product to be displayed. A setting of '1' will result in the display of every gene in the target genome which has any matches in the reference genomes. A setting of '2' requires at least two matches in the reference genomes, and is functionally equivalent to the criterion used for the Clusters of Orthologous Groups (16) analysis if the target and reference species are all phylogenetically distant. (v) A list of excluded species. If a gene product in the target genome is found within one of these excluded species, then the gene product in the target genome is not displayed. (vi) The criteria for identifying homologs in the excluded species.

The process of generating the concordance begins with the selection of all homologs which include sequences from the target genome and reference species using the criteria specified above. If there is more than one sequence in a particular reference species which matches a gene product in the target genome, then the system will select the best match, where best is defined in terms of the selection criteria. For example, if the selection criteria is *Z*-score, then the sequence with the highest *Z*-score will be used. Next, the 'match number' test is applied, and only those gene products with the requisite number of species with matching genes will be saved. We then select all homologs which include sequences from the target genome and the list of excluded species and which satisfy the second criteria. For each homolog found, we delete the sequence from the target list. Finally, the results are returned to the web browser. The resulting web page also contains a summary of the parameters used in the query and counts of the number of sequences selected and filtered.

Several types of hot links are provided in the returned Web page. It is possible to list the sequence of any gene product in the display. For each target sequence, it is possible to invoke CLUSTALW (23) to calculate a multiple sequence alignment for all the sequences related to the target protein sequence. It is also possible to display a list of every sequence that is homologous to a target gene product. For each match to a target sequence, it is possible to get a sequence alignment of just the matching gene product against the corresponding gene product in the target species.

RESULTS AND DISCUSSION

As of April 1998, our database holds 97 365 sequences from 19 complete or partial genomes and 8 798 905 FASTA comparison results, and we are using this tool to identify new targets for the development of antibiotics. We are focusing on gene products which do not have any known function in the well characterized genomes such as *Escherichia coli* (9) or *Bacillus subtilis* (14). Early results have shown that ~20% of the genes found by our concordance analysis tool are indeed essential in the bacteria we have tested.

Table 3. Summary of sample concordance

Organism	*E.coli*
Number of sequences	4289
Compared against	*B.subtilis* *H.influenzae* *H.pylori* *M.tuberculosis*
Selection qualifier	Target sequences must match all species
Selection criterion	Expected ≤1.0 × 10^{-24}
Comparison matrix	BLOSUM 62
Number of sequences which match at least one species above	1944
Sequences removed by 'all' requirement	1679
Excluding organism(s)	*S.cerevisiae*
Exclusion criterion	Expected ≤1.0 × 10^{-6}
Sequences removed by exclusion	176
Number of sequences in concordance	89

Table 4. Results of sample concordance

Name	Annotation
accD	acetyl-CoA carboxylase beta subunit
atoA	unknown. N.B. o216; residues 62–206 are 73% identical to amino acids 64–208 from acetyl-CoA transferase ATOA_HAEIN SW: P44874 (223 amino acids)
baeR	transcriptional regulatory protein BaeR
clpP	ATP-dependent clp protease proteolytic subunit
cysE	serine acetyltransferase
dfp	protein dfp
dnaA	chromosomal replication initiator protein DnaA
dnaB	replication protein replicative DNA helicase
dnaE	DNA polymerase III, alpha chain
dnaG	DNA biosynthesis; primase DNA primase
dppB	dipeptide transport system permease protein dppb
dppC	dipeptide transport system permease protein dppc
efp	elongation factor P
era	GTP-binding protein
fabD	malonyl CoA-acyl carrier protein transacylase
fabH	3-oxoacyl-[acyl-carrier-protein] synthase III
fmt	methionyl-tRNA formyltransferase
frr	ribosome recycling factor
ftsK	cell division protein FtsK
ftsZ	cell division protein FtsZ
galU	glucose-1-phosphate uridylyltransferase
gcpE	GcpE protein (protein E)
glmU	unknown. N.B. f456; ?% identical to GLMU_ECOLI SW: P17114; similar to *B.subtilis* tms; correction to previous sequence
glnA	glutamine synthetase
gyrA	DNA gyrase subunit A
hhoA	protease hhoA precursor
htrA	protease DO precursor; heat shock protein HtrA
infA	initiation factor IF-1
infC	initiation factor IF-3
kdpE	unknown. N.B. f225; 92% identical to KDPE_ECOLI SW: P21866
kdtB	unknown. N.B. o159
lig	DNA ligase
mfd	transcription-repair coupling factor (TrcF)
moaA	molybdenum cofactor biosynthesis protein A
moaC	molybdenum cofactor biosynthesis protein C
moeA	molybdopterin biosynthesis MoeA protein
mraY	first step of the lipid cycle reactions in the biosynthesis of the cell wall peptidoglycan phospho-N-acetylmuramoyl-pentapeptide-transferase
mrsA	MrsA protein
murA	udp-n-acetylglucosamine 1-carboxyvinyltransferase
nikC	unknown. N.B. o277
nikD	unknown. N.B. o254
nth	endonuclease III
ompR	positive regulatory gene for ompC and ompF
oppC	oligopeptide transport system permease protein
parC	chromosome partitioning topoisomerase IV subunit
pepA	aminopeptidase A/1
phoP	transcriptional regulatory protein PhoP
pnp	polynucleotide phosphorylase
polA	DNA polymerase I
prlA	preprotein translocase secy subunit
purU	formyltetrahydrofolate deformylase
pyrH	uridine 5′-monophosphate kinase
recA	catalyzes hydrolysis of ATP in the presence of ssDNA, ATP-dependent uptake of ssDNA by duplex DNA and ATP-dependent hybridization of homologous single-stranded DNAs RecA protein

Table 4. *Continued*

relA	GTP pyrophosphokinase
rho	transcription termination factor rho
rplS	50S ribosomal subunit protein L19
rplT	50S ribosomal subunit protein L20
rpoA	RNA polymerase, alpha subunit
rpoD	RNA polymerase, sigma-70 subunit RNA polymerase sigma-70 factor
rpoS	controls a regulon of genes required for protection against external stresses RNA polymerase sigma subunit RpoS (sigma-38)
rpsD	30S ribosomal subunit protein S4
rpsH	30S ribosomal subunit protein S8
ruvB	holliday junction DNA helicase RuvB
sapD	peptide transport system ATP-binding protein SapD
secA	preprotein translocase SecA subunit
sms	Sms protein
spoT	guanosine-3′,5′-bis(diphosphate) 3′-pyrophosphohydrolase; CG site no. 156
tpx	thiol peroxidase
trmD	tRNA(guanine-7)methyltransferase
uvrA	excision nuclease
uvrB	excision nuclease ABC subunit B
xseA	exodeoxyribonuclease large subunit
yabC	hypothetical 34.9 kDa protein in fruR-ftsL intergenic region
yaeM	hypothetical protein in frr 3′ region
ycfH	hypothetical protein in holB-ptsG intergenic region
yejE	hypothetical 38.1 kDa protein in bcr 5′ region
yfgB	hypothetical 43.1 kDa protein in ndk-gcpE intergenic region
yfhI	hypothetical protein in fdx 3′ region
ygaG	hypothetical protein in emrB 3′ region
ygbB	hypothetical 16.9 kDa protein in surE-cysC intergenic region
b0366	unknown. N.B. o255; this 255 amino acid ORF is 41% identical (7 gaps) to 239 residues of an ~280 amino acid protein NRTD_SYNP7 SW: P38046
b0420	unknown. N.B. f620; 73% identical (4 gaps) to 620 residues of YE39_HAEIN SW: P45205 (625 amino acids)
b0571	unknown. N.B. f227; this 227 amino acid ORF is 61% identical (0 gaps) to 225 residues of an ~232 amino acid protein COPR_PSESM SW: Q02540
b0658	unknown. N.B. f292; this 292 amino acid ORF is 23% identical (9 gaps) to 272 residues of an ~440 amino acid protein YTFL_HAEIN SW: P44717
b0661	unknown. N.B. f474; similar to BCHE_RHOCA SW: P26168; similar to LEXA_PSEPU SW: P37453
b0831	unknown. N.B. o306; this 306 amino acid ORF is 46% identical (32 gaps) to 306 residues of an ~344 amino acid protein DPPB_ECOLI SW: P37316
b0832	unknown. N.B. o303; this 303 amino acid ORF is 42% identical (5 gaps) to 278 residues of an ~304 amino acid protein DPPC_ECOLI SW: P37315
b1485	unknown. N.B. f298; this 298 amino acid ORF is 46% identical (7 gaps) to 268 residues of an ~304 amino acid protein DPPC_ECOLI SW: P37315
b2955	unknown. N.B. o378; this 378 amino acid ORF is 67% identical (4 gaps) to 378 residues of an ~384 amino acid protein HEMN_HAEIN SW: P43899

The gene names and annotation are taken from GenBank deposition of the genome. Gene names consisting of the letter ‘b’ followed by four digits are used for those open reading frames which do not have a gene name, and these names correspond to the gene labels in the GenBank entry for this genome, namely version M52. In this context, ‘unknown’ means having no product or function listed in the GenBank entry.

We can also demonstrate the utility of the tool by showing how DNA gyrase, the target of the existing, quinoline class of antibiotics (24), could have been predicted by our tool. Table 3 shows the summary of results of a concordance which sought targets in *E.coli* (9) where *E.coli* targets were matched against *B.subtilis* (14), *Haemophilus influenzae* (3), *Helicobacter pylori* (12) and *Mycobacterium tuberculosis* (personal communication from The Institute of Genomic Research). The protein coding regions for *M.tuberculosis* were predicted from the nearly complete sequence of the CSU#93 strain currently underway at the Institute for Genomic Research. The program, CRITICA (25), was used to identify the protein coding regions in *M.tuberculosis* using the default parameters. Only those matches using the BLOSUM 62 scoring matrix whose probability of being random was estimated to be <1.0 × 10^{-24} were used and which matched in all four genomes. In addition, any sequence which matched sequences in *Saccharomyces cerevisiae* (5) were discarded from the list. The criteria for this exclusionary rule was more encompassing, as we used homologies whose probability of being random was estimated to be <1.0 × 10^{-6}. The resulting gene names and annotations from the *E.coli* genome sequence (9) are shown in Table 4.

The parameters used for generating this concordance were selected in part to produce a short table suitable for publication. One can see a number of interesting characteristics in this list nonetheless. DNA gyrase does appear, and many of the proteins found are those involved with information processing; DNA replication and repair, transcription and translation. Several transporters are found, and 23 proteins have no function attributed to them.

However, it is important to recognize that there are limitations in this automated approach for concordance discovery. Distant gene relationships will be missed because the alignment scores will have low statistical significance. For example, Table 4 shows glutamine synthetase being unique to bacteria, but eukaryotes also have a distantly related glutamine synthetase. Another difficulty arises from multi-domain proteins where different functional units are spliced together into one gene product. Common domains, such as nucleotide binding domains, will register highly significant homologies, but these pairings may not reflect functional equivalence for the entire gene. Use of the global overlap ratio field in the comparison table can be helpful for this problem. Nonetheless, the user of this software must be cautious in their interpretation of the results. The use of these simple conservation and exclusion rules provides a powerful mechanism for identifying important genes across any set of genomes. It will be an interesting verification of the power of this concordance method to review the targets suggested here with the targets of new antibiotics discovered and developed in the post-genomic era.

AVAILABILITY

The concordance analysis described in this paper is part of a larger, freely redistributable software package called SEEBUGS (Software for the Examination, Exploration and Broad Understanding of Genome Sequences). In addition, a Web server using this software and public domain sequences is currently available for a limited number of external queries. For more information, please contact Dr Bruccoleri by email at bruc@acm.org or see the URL http://www.cabm.rutgers.edu/~bruc

ACKNOWLEDGEMENTS

We thank Dr Jiri Novotny, Dr Herb Klei and the referees for their insightful remarks on the manuscript, and Drs Wes Cosand, John Barrett and Joseph J. Villafranca for their support in the development of this application. We also thank Drs Jonathan Badger and Gary Olsen for providing us with CRITICA before publication. Permission to publish this analysis of the *M.tuberculosis* partial genome was kindly provided by The Institute for Genomic Research. The URL for accessing the sequence is http://www.tigr.org

REFERENCES

1 Cohen,M.L. (1994) *Trends Microbiol*., **2**, 422–425.

2 Corvalin,P. (1996) *J. Antimicrob. Chemother*., **37**, 855–869.

3 Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.-F., Dougherty,B.A., Merrick,J.M., *et al.* (1995) *Science*, **269**, 496–512.

4 Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M., *et al.* (1995) *Science*, **270**, 397–403.

5 Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., *et al.* (1996) *Science*, **274**, 546–567.

6 Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., *et al.* (1996) *Science*, **273**, 1058–1073.

7 Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S., *et al.* (1996) *DNA Res*., **3**, 109–136.

8 Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and Herrmann,R. (1996) *Nucleic Acids Res*., **24**, 4420–4449.

9 Blattner,F.R., Plunkett,III,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., *et al.* (1997) *Science*, **277**, 1453–1474.

10 Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K., *et al.* (1997) *J. Bacteriol*., **179**, 7135–7155.

11 Klenk,H.-P., Clayton,R.A., Tomb,J.F., White,O., Nelson,K.E., Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D., *et al.* (1997) *Nature*, **390**, 364–370.

12 Tomb,J.F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S., Dougherty,B.A., *et al.* (1997) *Nature*, **388**, 539–547.

13 Fraser,C.M., Casjens,S., Huang,W.M., Sutton,G.G., Clayton,R., Lathigra,R., White,O., Ketchum,K.A., Dodson,R., Hickey,E.K., *et al.* (1997) *Nature*, **390**, 580–586.

14 Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S., *et al.* (1997) *Nature*, **390**, 249–256.

15 Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M., *et al.* (1998) *Nature*, **392**, 353–358.

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

17 Stonebraker,M., Rowe,L.A. and Hirohama,M. (1990) *IEEE Trans. Knowl. Data Eng*., **2**, 125–142.

18 Wall,L., Christiansen,T., Schwartz,R.L. and Potter,S. (1996) *Programming Perl*. O'Reilly & Associates, Inc., 2nd edition.

19 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

20 Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

21 Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure*, Supplement 3, Volume 5. National Biomedical Research Foundation, Washington, DC.

22 Elmasri,R. and Navathe,S.B. (1994) *Fundamentals of Database Systems*. Addison-Wesley, 2nd edition.

23 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res*., **22**, 4673–4680.

24 Fisher,L.M., Lawrence,J.M., Josty,I.C., Hopewell,R., Margerrison,E.E. and Cullen,M.E. (1989) *Am J. Med*., **87**, 2S–8S.

25 Badger,J.H. and Olsen,G.J. (1998) *Mol. Biol. Evol*., in press.