3986–3990 *Nucleic Acids Research, 1998, Vol. 26, No. 17*

Protein sequence similarity searches using patterns as seeds

Zheng Zhang, Alejandro A. Schäffer[1], Webb Miller, Thomas L. Madden[2], David J. Lipman[2], Eugene V. Koonin[2] and Stephen F. Altschul[2,*]

Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA, [1]Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD 21224, USA and [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received May 7, 1998; Revised and Accepted July 8, 1998

ABSTRACT

Protein families often are characterized by conserved sequence patterns or motifs. A researcher frequently wishes to evaluate the significance of a specific pattern within a protein, or to exploit knowledge of known motifs to aid the recognition of greatly diverged but homologous family members. To assist in these efforts, the pattern-hit initiated BLAST (PHI-BLAST) program described here takes as input both a protein sequence and a pattern of interest that it contains. PHI-BLAST searches a protein database for other instances of the input pattern, and uses those found as seeds for the construction of local alignments to the query sequence. The random distribution of PHI-BLAST alignment scores is studied analytically and empirically. In many instances, the program is able to detect statistically significant similarity between homologous proteins that are not recognizably related using traditional single-pass database search methods. PHI-BLAST is applied to the analysis of CED4-like cell death regulators, HS90-type ATPase domains, archaeal tRNA nucleotidyltransferases and archaeal homologs of DnaG-type DNA primases.

INTRODUCTION

In the analysis of a protein or DNA sequence, particular interest often focuses upon a small region, domain or sequence pattern. A natural question is whether there are other related sequences that share the same pattern. The most widely used tools for sequence similarity search allow matching between arbitrary regions of the query and database sequences (1–5). In contrast, many motif-based search methods seek database sequences that match a pre-specified pattern (6–12). If this pattern is too weak, or not specified with sufficient precision, the number of matches may be very large, most being of no biological relevance. On the other hand, an overly-specific pattern may exclude many sequences of interest.

We describe here the pattern-hit initiated BLAST (PHI-BLAST) program, whose hybrid strategy addresses a type of question frequently asked by researchers: namely, is a particular pattern seen in a protein of interest likely to be functionally relevant, or does it occur simply by chance? To address this question, we combine a pattern search with a search for statistically significant sequence similarity. These two approaches were combined previously in a program that explored the output of a BLAST search for conserved patterns (10). PHI-BLAST implements a reverse strategy which is computationally more efficient, and which we believe will be of greater utility. Specifically, the similarity search is restricted to a subset of the sequence database comprised of the sequences that contain the given pattern.

The input to PHI-BLAST consists of a protein or DNA sequence, along with a specific pattern occurring at least once within the sequence. The pattern is currently required to be a sequence of residues or sets of residues, with ‘wild cards’ and variable spacing allowed; all PROSITE patterns (12), for example, have this form. For each match between an instance of the pattern in the query sequence and an instance in a database sequence, PHI-BLAST constructs a high-scoring local alignment that includes the match. All resulting alignments are sorted by score and evaluated statistically.

This approach has greatest utility when it is suspected that a few residues comprising a small motif may be crucial for the biological function of interest. Showing that this pattern occurs within an extended and statistically significant alignment of the query sequence with one or more database sequences greatly reduces the likelihood that the pattern is spurious. Conversely, insisting on the presence of the pattern and hence searching a reduced sequence space may aid the detection of subtle similarities that blend into the background noise in a regular BLAST search.

THE PHI-BLAST ALGORITHM

To search for matches to a given pattern, we adapted a method of Baeza-Yates and Gonnet (13) and Wu and Manber (14). This

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: altschul@ncbi.nlm.nih.gov

method permits simple patterns to be represented in a single computer word and matches to be found very efficiently. When the pattern is relatively complex, for example consisting of many rigid parts and/or having wide ranges of spacer lengths, our program first searches for the rigid part that is least likely to match by chance alone, and then performs local searches for the remaining pattern elements.

For each instance of the input pattern in a database sequence, paired with an instance in the query, PHI-BLAST attempts to find the optimal local alignment (1,15) containing the aligned patterns. This can be done rigorously by applying dynamic programming (16,17) to the parts of the two sequences preceding and the parts following the pattern. The alignment returned is required to begin at the corner of the path graph, but is permitted to end anywhere within the graph. The difficulty with this approach is that, to guarantee optimality, a very large portion of the path graph needs to be searched, and this requires inordinate time in a database search (18). Accordingly, we have used the gapped extension heuristic described in Altschul *et al.* (5) and Zhang *et al.* (18). Basically, path graph cells are considered only if the score of the best alignment leading into them falls no more than X below the best score yet found. For sufficiently large values of the X parameter, this approach almost always returns the optimal local alignment.

Because PHI-BLAST performs a gapped extension whenever an instance of the input pattern is encountered in the database, reasonable execution times depend upon such instances being relatively rare. Therefore, we allow only patterns that are expected to occur less frequently than once per 5000 database residues. Any pattern that contains four completely specified residues, or three specified residues whose average background frequency is ≤5.8%, passes this test. Of course, the more specific the input pattern, the faster PHI-BLAST will run. The frequency with which a pattern will occur within the database can be estimated easily (19) from background amino acid frequencies (20).

STATISTICAL ANALYSIS

An alignment A produced by PHI-BLAST may be divided into three parts: the region A_0 spanned by the input pattern, and the local alignments A_1 and A_2 produced to either side of A_0 by the gapped extension routine. Either or both of A_1 and A_2 may be empty. Correspondingly, the score S of the alignment may be divided into the scores S_0, S_1 and S_2. For the purpose of statistical analysis, it is easiest to assume that all alignment regions A_0 that satisfy the input pattern are of equal biological plausibility, and therefore to ignore their scores. Accordingly, each alignment produced by PHI-BLAST is ranked by its reduced score $S' = S_1 + S_2$. For a given value x, we wish to estimate how many alignments are expected to have a reduced score $S' \geq x$ purely by chance.

In general, the input pattern is chosen because it is known to correspond to some feature of biological interest. Therefore, we make no statistical inference from the number of times the pattern is observed to occur within the query sequence (n_q) and the database as a whole (n_d). We simply record $N = n_q n_d$, the number of distinct pattern pairs that may seed a PHI-BLAST local alignment.

The simplest model of protein sequences is as random strings of amino acids, chosen independently with specific background probabilities for the various possible residues. To estimate the random distribution of S', we start by considering the distribution of the scores S_1 and S_2 of which it is the sum. Each of these scores can be thought of as the result of the gapped extension routine applied to a pair of random sequences. In the limit of large values for the X-dropoff parameter (5,18), S_1 is the score of the optimal local alignment required to start at a particular point P. The much studied Smith–Waterman alignment score (1) is just this constrained local alignment score, maximized over all path graph points P. The distribution of Smith–Waterman scores has been established empirically to follow an extreme value distribution, whose scale or decay parameter λ does not change with increasing search space sizes (4,21–24). This implies (25) that the distribution of S_1 should have an exponential tail, with decay parameter λ equal to that of the extreme value distribution for Smith–Waterman scores. Some simple calculus then yields that for sufficiently large scores x, the distribution of $S' = S_1 + S_2$ has the form $Prob(S' \geq x) \approx C(\lambda x + 1)e^{-\lambda x}$ for some constant C. The scores of optimal local alignments constrained to contain distinct pattern pairs may be correlated, but the expected number of alignments attaining a given score is independent of such correlation. Therefore, the expected number of chance alignments produced by PHI-BLAST with reduced score at least x is

$$E(S' \geq x) \approx CN(\lambda x + 1)e^{-\lambda x} \qquad \mathbf{1}$$

Tables of λ for a variety of amino acid substitution matrices and gap costs have been reported (4), and their validity tested on a large number of protein families (26). The values for λ employed here differ slightly from those published previously (4), because we have re-estimated λ using larger and therefore more accurate simulations. The parameter C of equation **1** is new and requires its own estimation. Random simulation (data not shown) using the background amino acid frequencies of Robinson and Robinson (20) yields $C \approx 0.6$ for the BLOSUM-62 matrix (27) in conjunction with the complete range of affine gap costs useful for standard protein sequence comparison (4). We will consider the validity of equation **1** after discussing several biological examples.

IMPLEMENTATION AND EXAMPLES

To enhance the utility and functionality of a WWW-based version of PHI-BLAST, we have nested it between two other programs. While one may define a pattern based upon specific knowledge concerning the query sequence, a researcher often wishes to search a pattern-database for any well-characterized motifs the query may contain. To streamline this latter approach, we have implemented a program that first searches the PROSITE database (12) with the query; any patterns found may then be used to launch a PHI-BLAST database search. To facilitate more detailed analysis of PHI-BLAST output, we allow it automatically to serve as the basis for constructing a position-specific score matrix for further database searching via the position-specific iterated BLAST (PSI-BLAST) program (5). Like other BLAST family programs, PHI-BLAST incorporates a pre-filter for protein regions of biased amino acid composition (low complexity) that often corrupt database searches (28,29).

PHI-BLAST may detect subtle relationships that escape standard database similarity searches, but this potential depends upon the specification of an amino acid pattern likely to be conserved within the protein family of interest. We discuss four

examples involving protein families whose original description depended critically upon detecting relatively weak sequence similarities. In each case, PHI-BLAST reports a subtle but structurally and functionally relevant relationship. The alignments suggesting these relationships are not all statistically significant but, in each database search output ranked by *E*-value, they appear immediately after the alignments involving clear family members, thereby prompting further analysis. In contrast, any of these similarities reported by gapped BLAST (5) are preceded by a number of alignments with smaller *E*-values involving unrelated sequences. The four examples discussed below are summarized in Table 1. All searches were performed on the non-redundant (NR) protein sequence database maintained by the NCBI (30).

CED4-like cell death regulators

The *Caenorhabditis elegans* protein CED4 is a regulator of programmed cell death (apoptosis). CED4 contains the classical P-loop motif involved in phosphate binding and found in a great variety of ATPases and GTPases. ATP binding by CED4, and the role of ATP in its function, have been demonstrated (31,32). In a gapped BLAST search of the NR database, CED4 shows statistically significant sequence similarity to only one protein, the human apoptosis regulator Apaf-1, in which the P-loop is conserved (33,34). However when PHI-BLAST is used, requiring conservation of the P-loop (Table 1), the best hit after Apaf-1, with *E*-value 0.038, is to a plant disease resistance protein, *Arabidopsis thaliana* T7N9.18 (35). Further sequence comparison shows that animal apoptosis regulators and putative plant ATPases involved in disease resistance share several conserved motifs, suggesting that they have a common origin and may have similar roles in programmed cell death (L.Aravind, V.M.Dixit and E.V.Koonin, unpublished observations). Before the Apaf-1 sequence became available, this conclusion had been reached through a laborious comparison of CED4 to a large number of different ATPases (32). Because the Apaf-1 sequence is highly similar to homologous plant proteins, the connection between CED4 and the plant proteins can be easily demonstrated by iterative database search (5). Even without Apaf-1, however, PHI-BLAST is able immediately to establish this link.

HS90-type ATPase domains

We used PHI-BLAST to investigate the subtle but structurally validated relationship between the ATPase domains in the MutL DNA repair proteins, type II topoisomerases, histidine kinases and HS90 family proteins (36,37). The output identified a new family of eukaryotic proteins that contain the same type of predicted ATPase domain, but that in standard database searches do not show significant similarity to any known member of the superfamily. A PHI-BLAST search with the *Escherichia coli* MutL protein (38) as query showed moderate similarity (*E*-value 0.017) to the *C.elegans* protein ZC155.3 (39) that was originally described as having 'weak similarity to Bovine synaptocanalin I'. Subsequent database searches with this worm protein sequence as query revealed homologs in humans (KIAA0136) (40) and plants (41,42), whereas a PHI-BLAST search also showed convincing similarity to MutL family members (best *E*-value 6×10^{-5}). Elucidation of the function of this new family of eukaryotic ATP-utilizing enzymes will be of considerable interest; the synaptocanalin domain apparently was fused to the worm protein by exon misassembly.

Archaeal tRNA nucleotidyltransferases

The archaeal tRNA nucleotidyltransferases (Cca) are a distinct family of nucleic acid polymerases (43) that in standard database

Table 1. Detection of subtle protein sequence relationships using PHI-BLAST

Conserved domain or motif under investigation	Pattern[a]	GenBank (30) accession no. of query	Top non-trivial relevant hit found by PHI-BLAST		Top non-trivial relevant hit found by BLAST	
			Accession no.	*E*-value	Accession no.	*E*-value
A. P-loop ATPase domain in apoptosis regulators and plant stress response proteins	[GA]xxxxGK[ST]	231729	2213598	0.038	2961373	4.7
B. ATPase domain in mismatch repair protein MutL, type II topoisomerases, histidine kinases, and HS90 molecular chaperones	hxhxDxGxG	127552	488200	0.017	2495364	1.8
C. Nucleotidyltransferase domain in archaeal tRNA nucleotidyltransferases	DhDhhh	2826366	2650333	0.061	2650333	8.6
D. Motif VI of superfamily II helicases in archaeal homologs of bacterial DNA primases	QxxGRx[GA]R	2128723	2499099	0.54		

The reported results are from searches of the NCBI (30) non-redundant protein sequence database (April 9, 1998; 298 842 sequences; 90 087 406 residues). The PHI-BLAST and BLAST algorithms used the BLOSUM-62 substitution matrix (27), in conjunction with penalties of 11+ *k* for gaps of length *k*. BLAST *E*-values were calculated using the statistical parameters $\lambda = 0.270$ and $K = 0.047$, and applying an edge-effect correction (4). PHI-BLAST *E*-values were calculated from equation **1**, using the statistical parameters $\lambda = 0.270$ and $C = 0.6$.

[a]Patterns are described using the one-letter amino acid code. Brackets represent a choice among any of the enclosed amino acids. 'x' represents any amino acid. 'h' represents [ILVMF], a hydrophobic amino acid.

searches do not have detectable similarity to any proteins other than orthologs from other archaeal species. However, they do contain a conserved motif, with two aspartate residues, that resembles the catalytic sites of many other polymerases (44). When this pattern (Table 1) is specified in a PHI-BLAST search with *Methanococcus jannaschii* Cca (45) as query, the top hit outside the archaeal Cca family itself, with *E*-value 0.061, is to hypothetical protein AF0299 from *Archaeoglobus fulgidus* (46), which belongs to a previously described archaeal family of predicted nucleotidyltransferases (47); the third hit (*E*-value 0.13) is to an experimentally characterized streptomycin 3″-adenylyltransferase from *Enterococcus faecalis* (48).

Table 2. Accuracy of PHI-BLAST statistics

Example	Shuffled database		Reversed database	
	Low *E*-val.	Seqs with *E*-val. ≤ 10	Low *E*-val.	Seqs with *E*-val. ≤ 10
A	3.0	4	1.8	2
B	0.64	9	1.1	10
C	0.12	23	1.2	10
D	0.55	12	2.7	2

PHI-BLAST searches were performed on shuffled and reversed versions of the NR database, using the query sequences and associated patterns of Table 1, as well as the same alignment scoring system and statistical parameters λ and *C*. A, CED4-like cell death regulators; B, HS90-type ATPase domains; C, archaeal tRNA nucleotidyltransferases; D, archaeal homologs of DnaG-type DNA primases.

Archaeal homologs of DnaG-type DNA primases

Archaeal homologs of bacterial DNA primases, e.g. *M.jannaschii* protein MJ1206 (45), contain a motif typical of helicases (47), but do not show significant similarity to these proteins in standard BLAST searches. Using *M.jannaschii* MJ1206 and the helicase motif as query, the first non-trivial PHI-BLAST hit, with *E*-value 0.54, is to the well known helicase *Neisseria gonorrhoeae* UvrB (49). The relevance of the helicase motif in the archaeal primase homologs is supported by an extended alignment with the UvrB helicase (L.Aravind, D.D.Leipe and E.V.Koonin, unpublished observations). The similarities uncovered in this example are undetectable with standard database search techniques.

PERFORMANCE EVALUATION

To test the accuracy of the PHI-BLAST statistics given by equation **1**, we used each of the examples above to search 'random databases' constructed from NR by shuffling or reversing each sequence. For each query, the lowest recorded *E*-value, and the number of alignments found with *E*-value ≤ 10, are given in Table 2. For the shuffled database, the geometric mean of the observed numbers of sequences with *E*-value ≤ 10 is 10.0, and no single case diverges from this value by more than a factor of 2.5. This might be expected, as the values of λ and *C* used in equation **1** were calculated employing a random protein model in which all amino acids occur independently. Perhaps surprisingly, Table 2 suggests that under an alternative random protein model, based upon reversed real sequences, these statistics are slightly conservative.

To compare the speed of PHI-BLAST to that of a standard gapped BLAST program (5) we timed both for searches of each of the four examples above against the NR database. Analysis of the results (Table 3) suggests that on the computer system used, ~8 s of each PHI-BLAST run were required to scan the database for pattern hits and for system overhead; the remainder was spent on constructing gapped extensions for all pattern hits found. Clearly, the number of hits generated by the input pattern is a key determinant of PHI-BLAST's speed. For relatively informative patterns PHI-BLAST is very fast, requiring not much more time than that needed to search for pattern hits. For relatively weak patterns, PHI-BLAST expends most of its effort extending hits, and can require time comparable to that for gapped BLAST.

CONCLUSION

As illustrated by the biological examples discussed above, PHI-BLAST helps both to ascertain the biological relevance of patterns detected within protein sequences, and in some cases to detect subtle similarities that escape a regular BLAST search. We note, however, that PHI-BLAST was specifically designed to combine pattern search with the search for statistically significant sequence similarity, rather than to maximize search sensitivity. Thus in general one should not expect PHI-BLAST, which by its

Table 3. Execution speed of PHI-BLAST

Example	Length of query	No. of instances of pattern in database	PHI-BLAST execution time (seconds)	BLAST execution time (seconds)
A	549	14582	26	77
B	615	2986	12	103
C	449	1890	10	71
D	424	672	9	64

The four examples of Table 1 were used to search the NR database using PHI-BLAST, and BLASTP version 2.0.4. Both programs employed the same substitution and gap costs, and the same *X*-dropoff parameter. This timing experiment was run on one 168 MHz UltraSparc processor of a Sun Ultra Enterprise 4000/5000 server with 768 Mbytes of RAM. This machine runs the operating system Solaris, version 2.6, which is an implementation of UNIX. We used the current Sun C compiler, with the *-O* option for optimization, to compile both programs. The times given are the sum of the user and system times reported by the *time* command, and are for the better of two identical runs. A–D, as in Table 2.

nature is a single-pass search method, to be more sensitive than PSI-BLAST (5). Furthermore, within proteins, residues that are absolutely conserved during evolution constitute a small minority, and even specifying a restricted set of possibilities for a given residue position often excludes many members of a protein family. PHI-BLAST therefore is not the ideal tool for completely delineating a class of related proteins. However, by greatly restricting the size of the search space, PHI-BLAST can allow the similarities of some distant homologs to rise above the background noise that would otherwise obscure them. Such findings can be used subsequently for more extensive family analysis using PSI-BLAST (5) or other tools.

We have developed PHI-BLAST for protein–protein comparison, but plan to extend its applicability. A version that translates a DNA database in all six reading frames for comparison to a protein query would be particularly valuable, and a DNA–DNA comparison version should also find use. We also plan to extend PHI-BLAST so that it may use generalized affine gap costs (50) in place of the traditional affine gap costs (51–54) currently permitted.

Note

Source code for PHI-BLAST is available by anonymous ftp from the machine ncbi.nlm.nih.gov, within the directory 'blast', and the program may be run from NCBI's web site at http://www.ncbi.nlm.nih.gov/

ACKNOWLEDGEMENTS

Z.Z. and W.M. are supported by grant LM05110 from the National Library of Medicine. We thank Dr L. Aravind for helpful discussions.

REFERENCES

1 Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol.*, **147**, 195–197.
2 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
3 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
4 Altschul,S.F. and Gish,W. (1996) *Methods Enzymol.*, **266**, 460–480.
5 Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
6 Myers,E.W. and Miller,W. (1989) *Bull. Math. Biol.*, **51**, 5–37.
7 Smith,R.F. and Smith,T.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 118–122.
8 Staden,R. (1990) *Methods Enzymol.*, **183**, 193–211.
9 Mehldau,G. and Myers,G. (1993) *Comp. Appl. Biosci.*, **9**, 299–314.
10 Tatusov,R.L. and Koonin,E.V. (1994) *Comp. Appl. Biosci.*, **10**, 457–459.
11 Ogiwara,A., Uchiyama,I., Takagi,T. and Kanehisa,M. (1996) *Protein Sci.*, **5**, 1991–1999.
12 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **25**, 217–221.
13 Baeza-Yates,R. and Gonnet,G. (1992) *Commun. Assoc. Comp. Mach.*, **35**, 74–82.
14 Wu,S. and Manber,U. (1992) *Commun. Assoc. Comp. Mach.*, **35**, 83–91.
15 Sellers,P.H. (1984) *Bull. Math. Biol.*, **46**, 501–514.
16 Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.
17 Sankoff,D. (1972) *Proc. Natl Acad. Sci. USA*, **69**, 4–6.
18 Zhang,Z., Berman,P. and Miller,W. (1998) *J. Comput. Biol.*, **5**, 197–210.
19 Staden,R. (1989) *Comp. Appl. Biosci.*, **5**, 89–96.
20 Robinson,A.B. and Robinson,L.R. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 8880–8884.
21 Smith,T.F., Waterman,M.S. and Burks,C. (1985) *Nucleic Acids Res.*, **13**, 645–656.
22 Collins,J.F., Coulson,A.F.W. and Lyall,A. (1988) *Comp. Appl. Biosci.*, **4**, 67–71.
23 Mott,R. (1992) *Bull. Math. Biol.*, **54**, 59–75.
24 Waterman,M.S. and Vingron,M. (1994) *Stat. Sci.*, **9**, 367–381.
25 Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press, New York, NY.
26 Pearson,W.R. (1998) *J. Mol. Biol.*, **276**, 71–84.
27 Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
28 Wootton,J.C. and Federhen,S. (1993) *Comp. Chem.*, **17**, 149–163.
29 Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) *Nature Genet.*, **6**, 119–129.
30 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) *Nucleic Acids Res.*, **26**, 1–7.
31 Seshagiri,S. and Miller,L.K. (1997) *Curr. Biol.*, **7**, 455–460.
32 Chinnaiyan,A.M., Chaudhary,D., O'Rourke,K., Koonin,E.V. and Dixit,V.M. (1997) *Nature*, **388**, 728–729.
33 Zou,H., Henzel,W.J., Liu,X., Lutschg,A. and Wang,X. (1997) *Cell*, **90**, 405–413.
34 Li,P., Nijhawan,D., Budihardjo,I., Srinivasula,S.M., Ahmad,M., Alnemri,E.S. and Wang,X. (1997) *Cell*, **91**, 479–489.
35 Buehler,E., Dewar,K., Feng,J., Kim,C., Li,Y., Shinn,P., Sun,H., Conway,A., Conway,A., Kurtz,D., *et al.* (1997) GenBank accession no. 2213598.
36 Bergerat,A., de Massy,B., Gadelle,D., Varoutas,P.C., Nicolas,A. and Forterre,P. (1997) *Nature*, **386**, 414–417.
37 Mushegian,A.R., Bassett,D.E., Jr, Boguski,M.S., Bork,P. and Koonin,E.V. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 5831–5836.
38 Tsui,H.T., Mandavilli,B.S. and Winkler,M.E. (1992) *Nucleic Acids Res.*, **20**, 2379.
39 Wilson,R., Ainscough,R., Anderson,K., Baynes,C., Berks,M., Bonfield,J., Burton,J., Connell,M., Copsey,T., Cooper,J., *et al.* (1994) *Nature*, **368**, 32–38.
40 Nagase,T., Seki,N., Tanaka,A., Ishikawa,K. and Nomura,N. (1995) *DNA Res.*, **2**, 167–174.
41 Bevan,M., Hilbert,H., Braun,M., Holzer,E., Brandt,A., Duesterhoeft,A., Hoheisel,J., Jesse,T., Heijnen,L., Vos,P., *et al.* (1998) GenBank accession no. 2961386.
42 Bevan,M., Hilbert,H., Braun,M., Holzer,E., Brandt,A., Duesterhoeft,A., Hoheisel,J., Jesse,T., Heijnen,L., Vos,P., *et al.* (1998) GenBank accession no. 2961387.
43 Yue,D., Maizels,N. and Weiner,A.M. (1996) *RNA*, **2**, 895–908.
44 Dracheva,S., Koonin,E.V. and Crute,J. (1995) *J. Biol. Chem.*, **270**, 14148–14153.
45 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.
46 Klenk,H.P., Clayton,R.A., Tomb,J., 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.
47 Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997) *Mol. Microbiol.*, **25**, 619–637.
48 LeBlanc,D.J., Lee,L.N. and Inamine,J.M. (1991) *Antimicrob. Agents Chemother.*, **35**, 1804–1810.
49 Black,C.G., Fyfe,J.A. and Davies,J.K. (1995) *J. Bacteriol.*, **177**, 1952–1958.
50 Altschul,S.F. (1998) *Proteins*, **32**, 88–96.
51 Gotoh,O. (1982) *J. Mol. Biol.*, **162**, 705–708.
52 Fitch,W.M. and Smith,T.F. (1983) *Proc. Natl Acad. Sci. USA*, **80**, 1382–1386.
53 Altschul,S.F. and Erickson,B.W. (1986) *Bull. Math. Biol.*, **48**, 603–616.
54 Myers,E.W. and Miller,W. (1988) *Comp. Appl. Biosci.*, **4**, 11–17.