Computers Chem. Vol. 17, No. 2, pp. 203–207, 1993
Printed in Great Britain.

0097-8485/93 $6.00 + 0.00

BLAZE™: AN IMPLEMENTATION OF THE SMITH–WATERMAN SEQUENCE COMPARISON ALGORITHM ON A MASSIVELY PARALLEL COMPUTER*

DOUGLAS L. BRUTLAG,[1]† JEAN-PIERRE DAUTRICOURT,[2] RON DIAZ,[2] JEFF FIER,[3] BRUCE MOXON[3]‡ and RICHARD STAMM[3]

[1]Department of Biochemistry, Stanford Medical School, Stanford, CA 94305, [2]IntelliGenetics Inc., 700 East El Camino Real, Mountain View, CA 94040 and [3]MasPar Inc., 749 North Mary Avenue, Sunnyvale, CA 94086, U.S.A.

(Received 24 November 1992; in revised form 17 February 1993)

Abstract—We have implemented the Smith and Waterman dynamic programming algorithm on the massively parallel MP1104 computer from MasPar and compared its ability to detect remote protein sequence homologies with that of other commonly used database search algorithms. Dynamic programming algorithms are normally too computer intensive to permit full databases search, however on the MP1104 a search of the Swiss-Prot database takes about 15 s. This nearly interactive speed of database searching permits one to optimize the parameters for each query. Most of the common database search methods (FASTA, FASTDB and BLAST) gain their speed by using approximations such as word matching or eliminating gaps from the alignments which prevents them from detecting remote homologies. By using queries from protein super families containing a large number of family members of diverse similarities, we have measured the ability of each of these algorithms to detect the remotest members of each super family. Using these super families, we have found that the algorithms, in order of decreasing sensitivity are BLAZE, FASTDB, FASTA and BLAST. Hence the massively parallel computers allow one to have maximal sensitivity and search speed simultaneously.

INTRODUCTION

Most word-based or index-based sequence search algorithms such as FASTA (Pearson & Lipman, 1988), FASTDB (Brutlag *et al.*, 1990), or BLAST (Altschul *et al.*, 1990) use approximate methods initially to find regions of local homology. These approximate methods are not as sensitive as the original full dynamic programming algorithms of Needleman & Wunsch (1970) or Smith & Waterman (1981). Even when indexing methods use a word size of one (FASTA and FASTDB) or allow inexact word matching (FASTDB and BLAST), they do not consider all possible alignments and can miss biologically important sequence similarities (Brutlag *et al.*, 1990).

The Needleman–Wunsch and Smith–Waterman methods gain their sensitivity by comparing each base or amino acid in the query with each base or amino acid in the database. In addition, they can utilize a table of amino acid similarity scores (or log-likelihood replacement scores or PAM matrices) to increase the significance of similarities of proteins that are related to each other in evolution (Schwartz & Dayhoff, 1979). One can also vary the penalty for the introduction of an insertion-deletion gap based on its length. Given a table of amino acid similarity scores and specific gap penalties, the Needleman–Wunsch and Smith–Waterman algorithms result in scores which are optimal for either the global or local alignment of two sequences respectively.

The approximate methods were developed to be able to align a query with every sequence in a database in a reasonable amount of computer time and memory with serial processors (Wilbur & Lipman, 1983). Comparing a sequence with current databases using the Smith–Waterman or Needleman–Wunsch algorithms is extremely computer intensive requiring either a massively parallel computer or a supercomputer. They also require memory proportional to the length of the query times the length of the database (Barsalou & Brutlag, 1991).

In this work, we have implemented the full dynamic programming algorithm of Smith & Waterman (1981) as modified by Gotoh (1982) on a massively parallel MasPar MP1104 computer (4096 4-bit processors with a total of 256 megabytes of memory). This implementation, named BLAZE, affords us the complete sensitivity of the Smith and Waterman algorithm using PAM matrices and gap penalties, while simultaneously maintaining interactive performance. Searches of the Swiss-Prot 21 protein database (7,866,594 residues) with a 100 amino acid query take 14.4 s with a single penalty for the

* The preliminary version of this work was presented during the *Second International Workshop of Open Problems in Computational Molecular Biology*, Telluride Summer Research Center, Telluride, Colo., 19 July–2 August 1992.

† Author for correspondence.
‡ Present address: Teknekron Inc., 1080 Marsh Road, Menlo Park, CA 94025, U.S.A.

insertion or extension of a gap and 35 s when using a full affine gap penalty. This represents about 55 million residue comparisons per second or 22 million comparisons per second respectively. These rapid rates of sequence comparison are possible due to the (1) the large number of processors employed, (2) the efficiency of the implementation of the Smith–Waterman algorithm and (3) the ability to hold the entire database in memory at all times. These rapid search rates also permit full database inversion. One can compare all sequences in a database with all of the others. Results of such a database inversion can be used to classify sequences in superfamilies, discover evolutionary relationships, or discover conserved coding segments.

Here, we compare the accuracy and the sensitivity of three rapid sequence database search algorithms FASTA, FASTDB and BLAST, with the Smith–Waterman algorithm as implemented in BLAZE.

ACCURACY OF BLAST, FASTA AND FASTDB SCORES COMPARED WITH BLAZE

Figure 1 shows the correlations between the scores returned from database searches using BLAST, FASTA and FASTDB, each compared with the scores returned by BLAZE. In each search, the α-chain of human hemoglobin was used as a query to search the Swiss-Prot 15 database. The top 1000 scores were saved from each search and then the

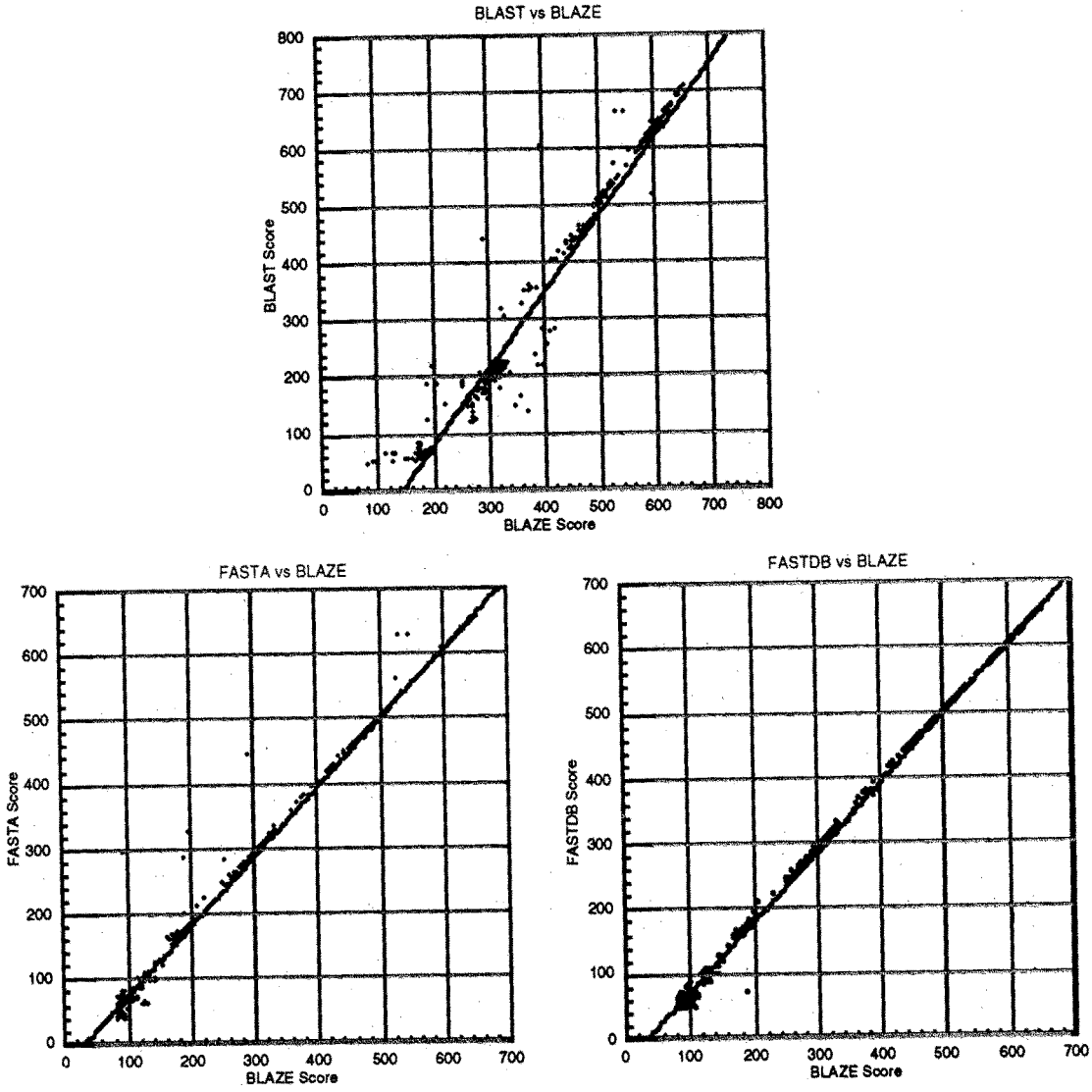

Fig. 1. The human hemoglobin α-chain sequence (HBA_HUMAN in Swiss-Prot) was used as a query to search the Swiss-Prot 15 database using four different algorithms, BLAST (PAM 250), FASTA (k-tuple = 1), FASTDB (k-tuple = 1, PAM 250, gap penalty 3.0 and gap-size-penalty 0.05) and BLAZE (PAM 250, gap-penalty 3). The top 1000 scores from each search were recorded and the scores common to each pair of algorithms listed above were plotted and the correlation coefficient calculated. The correlation coefficients for each of the searches was: BLAST vs BLAZE, $r = 0.9606$; FASTA vs BLAZE, $r = 0.9894$; FASTDB vs BLAZE, $r = 0.9983$.

scores for each database sequence in common between each search and the corresponding BLAZE search were plotted on a correlation graph. An identical PAM 250 was used for each search.

The scores given by BLAST correlated well with the BLAZE algorithm for all the database sequences whose scores were >50% of the maximal score (the query vs itself). This includes all of the α-chain hemoglobins and many of the β-chain hemoglobins. The correlation for scores <50% of the maximal value was significantly poorer.

FASTA gave much better correlation of scores with BLAZE down to 20% of the maximal score. Below this value the scores correlated poorly. There were also a number of outliers which gave unusually high FASTA scores compared with their corresponding BLAZE scores. The nature and reason for these outliers was not examined further.

FASTDB gave the best correlation of scores with BLAZE extending down to as low as 10% of the maximum possible score. A single sequence (out of 469 hemoglobins plotted) gave an unusually low score with the FASTDB algorithm compared with its BLAZE score. The separation of the hemoglobins into various groups (α-chain family, β-chain family, myoglobins and other globins) was also obvious from the clustering of groups of proteins along the correlation curve. Examination of the list of ranked sequences from the BLAZE searches show that all but two of the hemoglobin α-chain were ranked higher than all of the hemoglobin β-chains and no complete β-chain ranked lower than any myoglobin. In general, the scores of the hemoglobin sequences in common between pairs of methods correlated well (correlation coefficients between $r = 0.9606$ and $r = 0.9983$, see Fig. 1).

We also performed another test using a query sequence that is much less highly conserved than the globin family. We choose the MetR protein from *Salmonella typhimurium* (METR_SALTY in Swiss-Prot) as our query for this test. This protein is a positive regulator of the methionine operon and is a member of the LysR family of bacterial activator proteins originally described by Henikoff *et al.* (1988). These proteins all share a helix-turn-helix DNA binding motif but only limited sequence homology (between 19 and 31% of the maximal matching score). Using the sequence of METR_SALTY as a query, the four algorithms only had 29 members of the LysR family in common (Table 2). When the scores for these 29 proteins were correlated with the BLAZE score, the correlation coefficients fell in the range 0.4–0.9 (BLAST vs BLAZE, $r = 0.3938$; FASTA vs BLAZE, $r = 0.8528$; FASTDB vs BLAZE, $r = 0.8955$). Again this reflects the increased ability of FASTA and FASTDB to find the best local alignments when compared with the dynamic programming method of BLAZE.

SENSITIVITY OF FASTA, FASTDB AND BLAST COMPARED WITH BLAZE

In order to measure the sensitivity of the database search algorithms we compared their abilities to discover distantly related homologous sequences. We decided to examine the number of the members of the globin family that can be found using a mammalian α-hemoglobin chain as a query. Since there were 622 globins in the Swiss-Prot Release 22 database (479 intact hemoglobins, 5 hemoglobin fragments, 71 myoglobins, 60 bacterial and other globins and 13 plant leghemoglobins, see Table 1), we asked how many of each of these members of globin family were among the top 650 scores in the database search. We varied the PAM matrix and gap-penalties for BLAZE, FASTDB and the PAM matrix for BLAST until we obtained maximal numbers of these globins.

The number of each of these globin families in the top 650 scores are tabulated in Table 1. It is clear that a PAM value between 200 and 250 are required to detect maximal numbers of the plant leghemoglobins using the mammalian hemoglobin query. It is also clear from the data in Table 1 that there is an optimal value for the gap and gap size penalties. This optimum results from the fact that there are gaps in the best alignment of the mammalian hemoglobin with the plant leghemoglobins. Too high a gap or gap size penalty results in a marked lowering of the score with the leghemoglobins. Lowering the gap and gap size penalty too far increases the score of less highly related sequences which eventually compete for a position in the top 650. Hence, one must be able to vary the gap and gap size penalties independently and in a query dependent fashion in order to obtain the maximal sensitivity of the search.

Another test of the sensitivity of the search algorithms was to utilize the METR_SALTY protein as a query and to measure how many of the known members of the LysR family fall among the top 40 sequences (Table 2). There are 39 members of the LysR family in the Swiss-Prot 21 database. Among these are twelve more distantly related NodD proteins which regulate expression of nodulation genes among the *Rhizobium* and other plant bacterial symbionts. Both BLAZE and FASTDB rank all 39 members of the LysR family before any non-family members. With the default parameter settings, BLAST and FASTA detect only 29 members of the LysR family among the top 40 ranked scores, missing in particular, ten of the distantly related NodD regulatory proteins. By increasing the PAM matrix to 250, BLAST can detect the remaining members of the NodD family.

Yet another criteria can be applied to these searches to judge their sensitivity. The BLAZE, FASTDB and BLAST programs all calculate the likelihood or expectation of finding the score observed between the query and the top scores. Hence, one can judge the sensitivity of the search by the

Table 1. The query HBA_HUMAN was used in a sequence similarity search of the Swiss-Prot 22 database using BLAZE, BLAST, FASTA, FASTDB using parameters shown or the defaults

Program	PAM MATRIX	Gap pen.	Gap size pen.	Hemoglobins	Myoglobins	Other globins	Leghemoglobins	Total found
BLAZE	PAM 250	1	0.0	373	0	0	0	373
BLAZE	PAM 250	1	0.1	472	71	15	0	558
BLAZE	PAM 250	1	0.2	473	71	37	0	581
BLAZE	PAM 250	1	0.5	475	71	55	1	602
BLAZE	PAM 250	1	1	475	71	56	1	603
BLAZE	PAM 250	2	0.0	472	1	6	0	479
BLAZE	PAM 250	2	0.01	472	71	11	0	554
BLAZE	PAM 250	2	0.02	472	71	16	0	559
BLAZE	PAM 250	2	0.05	475	71	43	0	589
BLAZE	PAM 250	2	0.1	475	71	53	1	600
BLAZE	PAM 250	2	0.2	479	71	58	9	617
BLAZE	PAM 250	2	0.4	479	71	60	9	619
BLAZE	PAM 250	2	0.5	478	71	58	9	616
BLAZE	PAM 250	2	1	477	71	57	8	613
BLAZE	PAM 250	2	2	478	71	49	6	604
BLAZE	PAM 250	2	4	478	71	49	6	604
BLAZE	PAM 250	2	8	478	71	49	6	604
BLAZE	PAM 250	4	4	477	71	35	1	584
BLAZE	PAM 250	8	8	476	71	28	0	575
BLAZE	PAM 0	2	0.2	316	3	2	0	321
BLAZE	PAM 50	2	0.2	473	44	11	0	528
BLAZE	PAM 100	2	0.2	474	71	36	1	582
BLAZE	PAM 150	2	0.2	475	71	56	7	609
BLAZE	PAM 200	2	0.2	476	71	59	9	615
BLAZE	PAM 250	2	0.2	479	71	58	9	617
BLAST	PAM 250	∞	∞	484	71	22	0	577
FASTA	PAM 250	1-tup		476	71	27	1	575
FASTDB	PAM 250	3	0.5	479	71	56	7	613
Total in SPT 22				479 + 5 frag.	71	60	13	622

The number of hemoglobins, myoglobins, plant leghemoglobins or other globins found in the top 650 ranked sequences are reported in this table. The last line specifies the total number of each type of globin present in Swiss-Prot found by keyword search (using FINDSEQ).

statistical significance of the alignments of known homologs. A sensitive search will have the highest significance (lowest likelihood or expectation) for known homologs. In data not shown, the order of sensitivity of the three programs which calculate significance are again BLAZE, FASTDB and BLAST. This was true for several different protein super families, some of which required gaps in the alignments and others which did not.

Table 2. The query sequence METR_SALTY was used in a database search of Swiss-Prot 21 either using BALZE, BLAST, FASTA, or FASTDB algorithms using the parameters listed

Method	PAM	Gap penalty	LysR found	LysR missed	Non LysR
BLAZE	50	7	17	22	23
BLAZE	50	14	18	21	22
BLAZE	50	21	18	21	22
BLAZE	50	28	18	21	22
BLAZE	100	6	26	13	14
BLAZE	100	12	33	6	7
BLAZE	100	18	33	6	7
BLAZE	100	24	33	6	7
BLAZE	150	5	29	10	11
BLAZE	150	10	39	0	1
BLAZE	150	15	39	0	1
BLAZE	200	4	29	10	11
BLAZE	200	8	38	1	2
BLAZE	200	12	39	0	1
BLAZE	200	16	39	0	1
BLAZE	250	4	32	7	8
BLAZE	250	8	39	0	1
BLAZE	250	12	39	0	1
BLAZE	250	16	39	0	1
BLAZE	250	20	39	0	1
BLAST	120		29	10	11
BLAST	250		39	0	1
FASTA	120		29	10	11
FASTDB	250	3	39	0	1
Total in	SWP 21		39		

The number of the members of the LysR family of bacterial regulatory proteins found among the top 40 sequences are reported.

DISCUSSION

Word based sequence similarity search algorithms were developed to overcome the computational complexity of the database search on serial computers. By requiring exact word matching in the initial step of homology detection, the search algorithms had to give up the sensitivity afforded dynamic programming algorithms coupled with amino acid replacement matrices. PAM matrices, while allowing one to detect more remotely related sequences, are time consuming even on today's most powerful workstations. We have calculated that the searches performed here by BLAZE in <15 s would take over 75 h on a 25 MHz UNIX workstation using the same algorithm.

The ability to utilize a wide variety of PAM matrices and gap penalties coupled with the interactive speeds of BLAZE make it possible to perform many database searches in a few minutes. BLAZE permits one to identify optimal parameters for detecting any subclass of sequences related to the query interactively. This tool can insure that any statistically significant sequence similarity, no matter how remote, can be detected.

In addition to searching for related groups of proteins, the rapid speed of BLAZE permits comparison of entire databases. For example, comparisons of all the coding regions or exons in GenBank with each other (Dorit *et al.*, 1990) could be accomplished in a few days. We have performed one such database comparison using BLAZE. We compared the GENESEQ™ database of patented protein sequences (17,273 proteins with 2,606,023 residues) vs Swiss-Prot 21, (23,742 proteins with 7,866,594 residues). This database comparison ran in 96 h and showed that 8073 sequences were in common and 9200 proteins were unique to GENESEQ™.

Several other efforts to parallelize database search have been undertaken. Some of these involve implementation of a general sequence alignment algorithm on other massively parallel computers (Collins & Coulson, 1984). Others involve the implementations on more complex parallel architectures (Lander & Mesirov, 1988; Sittig *et al.*, 1993; Galper & Brutlag, 1990). Other efforts involve implementing the Smith–Waterman algorithm directly in hardware in the form of a customized chip (ASICs) (Hunkapiller *et al.*, 1990). Finally there are efforts to implement some of the approximate methods in parallel (Deshpande *et al.*, 1991) or in hardware as well (White *et al.*, 1991).

The results presented here indicate that general purpose massively parallel computers can give high sensitivity and accuracy of database search at interactive speeds. Processing elements which handle only a few bits at a time are ideally suited for sequence similarity search. A large amount of local memory per processor is necessary to hold the similarity matrices and a subset of the database. With their large numbers of processors, the aggregate memory and bandwidth can exceed that of traditional supercomputers. The MasPar architecture is ideal in that the entire databse can be held in memory overcoming the I/O bottleneck that plagues many other efforts to parallelize sequence comparisons. The MP1104 with 4,096 processors has sufficient memory to hold the entire GenBank and Swiss-Prot database simultaneously Larger computers from MasPar are able to hold the entire human genome in memory at one time. We feel that massively parallel methods hold the key to the searching and sorting genome amounts of sequence data.

Availability—BLAZE hardware and software is available from MasPar Inc. [(408)736-3300) and IntelliGenetics Inc. ((415)962-7300] at the addresses given above.

Acknowledgements—BLAZE™ is a registered trademark of IntelliGenetics Inc. and MasPar Computer Corp. GENESEQ™ and FINDSEQ™ are trademarks of IntelliGenetics Inc.

REFERENCES

Altschul S. F., Gish W., Miller W., Myers E. W. & Lipman D. J. (1990) J. *Mol. Biol.* **215,** 403.

Barsalou T. & Brutlag D. L. (1991) *Md Comput.* **8,** 144.

Brutlag D. L., Dautricourt J.-P., Maulik S. & Relph J. (1990) *CABIOS* **6,** 237.

Collins J. F. & Coulson A. F. (1984) *Nucleic Acids Res.* **12,** 181.

Deshpande A. S., Richards D. S. & Pearson W. R. (1991) *Comput. Appl. Biosci.* **7,** 237.

Dorit R. L., Schoenbach L. & Gilbert W. (1990) *Science* **250,** 1377.

Galper A. R. & Brutlag D. L. (1990) *Parallel Similarity Search and Alignment with the Dynamic Programming Method* (KSL Report 90-74). Stanford University, Calif.

Gotoh O. (1982) J. *Mol. Biol.* **162,** 705.

Henikoff S., Haughn G. W., Calvo J. M. & Wallace J. C. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85,** 6602.

Hunkapiller T., Waterman M., Jones R., Effert J., Chow E., Peterson J. & Hood L. (1990) *Human Genome: 1989–90 Program Report*, p. 101. U.S. Department of Energy, Washington, D.C.

Lander E. & Mesirov J. P. (1988) 257.

Needleman S. B. & Wunsch C. D. (1970) J. *Mol. Biol.* **48,** 443.

Pearson W. R. & Lipman D. J. (1988) *Proc. Natl Acad. Sci. U.S.A.* **85,** 2444.

Schwartz R. M. & Dayhoff M. O. (1979) *Atlas Protein Struct.* **5**(Suppl. 3), 353.

Sittig D. F., Foulser D., Carriero N., McCorkle G. & Miller P. L. (1993) *A Parallel Computing Approach to Genetic Sequence Comparison: The Master–Worker Paradigm.* George Washington University, Washington, D.C. In press.

Smith T. F. & Waterman M. (1981) J. *Mol. Biol.* **147,** 195.

White C. T., Singh R. K., Reintjes P. B., Lampe J., Erickson B. W., Dettloff W. D., Chi V. L. & Altschul S. F. (1991) *BioSCAN: A VLSI-Based System for Biosequence Analysis*, p. 504. IEEE Computer Society Press, Los Alamitos, Calif.

Wilbur W. J. & Lipman D. J. (1983) *Proc. Natl Acad. Sci. U.S.A.* **80,** 726.