

ASSESSMENT OF SMITH-WATERMAN SEQUENCE  
SEARCH TOOLS IMPLEMENTED IN BIOACCELERATOR,  
FDF AND MASPAR

BIOSTANDARDS REPORT

18 FEBRUARY, 1997

## Summary

Analysis of database (SWISS-PROT) search performance, using a set of query protein sequences, by the three implementations of Smith-Waterman search algorithm (namely, Bioccelerator, Fast Data Finder - FDF, and MasPar) indicated that database search speed assumes an asymptotic behaviour as the length of the query sequence increases. Given the configurations available for this study, FDF is faster followed by Bioccelerator and MasPar. An increase in the subject database size brings about a disproportionate increase in MasPar search time while in Bioccelerator and FDF it is nearer to a proportionate increase.

Analysis of the ranking of search results by the three implementations indicated that the hit sequences are appropriately classified into groups of evolutionarily/functionally related members and all three implementations produce the same groups. The range of scores for the closely related group is distinctly higher than that of other groups. In every instance of the test queries (except in one case) the three implementations pick all the related sequence entries from SWISS-PROT database as top hits before picking up any unrelated sequences. Even though the default gap penalties (especially the gap extension) are different in the implementations, the ranking order of the hits are remarkably similar among the implementations (in general a reordering of  $\pm 5$  takes place and such a reordering always occurs within a group). In 9 of 17 instances of the query, the scores of the top group is similar among the implementations while in the remaining 8 cases a difference of 2-10% can be observed.

## Contents

<b>Summary</b>	<b>2</b>
<b>Introduction</b>	<b>3</b>
<b>Aim and Scope of This Study</b>	<b>3</b>
<b>Hardware &amp; Software Implementations</b>	<b>4</b>
<b>I. Search Speed Performance</b>	<b>4</b>
<b>Protocol</b>	<b>4</b>
<b>Enumeration of ‘Raw’ Search Time</b>	<b>5</b>
<b>Results</b>	<b>6</b>
Raw Search Time and Speed	6
Effect of change in database size on raw search time	9
<b>II. Ranking of the Search Results</b>	<b>11</b>
<b>Protocol</b>	<b>11</b>
<b>Results</b>	<b>11</b>
Rank-wise Grouping of Resultant Hits	11
Selectivity	14
Ranking Order And Score Difference Among The Implementations	15
<b>Suggested Further Readings</b>	<b>15</b>

## Introduction

Searching databases for similarity to a newly determined sequence is a vital research tool in computational biology. To tap the potential benefits arising out of such a requirement, a number of hardware and software implementations are being marketed by commercial vendors. Three such implementations are available at EBI. They are Bioccelerator (from Compugen Limited), Fast Data Finder - FDF (from Paracel online systems), and MasPar (from MasPar or DIGITAL) implementation of MPSRCH (from Oxford Molecular Ltd). Bioccelerator is a specifically designed dynamic programming hardware engine to facilitate sequence analysis. Fast Data Finder is a single-purpose reconfigurable engine, containing a very large number of custom processors, primarily designed for string pattern-matching for text searching and subsequently adapted for sequence analysis. MasPar belongs to a family of general purpose parallel computers and MPSRCH is a suite of biological sequence programs designed to run on MasPar.

Smith-Waterman dynamic programming algorithm, which finds the most similar subsequences of two sequences, has been generally recognised as the most sensitive sequence comparison method currently available. All the three implementations search sequences in protein and DNA databases for similarity to the query sequence by using Smith-Waterman algorithm as the core sequence comparison method. In addition Bioccelerator and FDF have implemented generalised profile method for representing conserved regions. Bioccelerator has further implemented multiple sequence alignment.

The objective of this study is to make an assessment of the performances of the implementations. The features assessed are (i) the speed of search, and (ii) the ranking of search results. Results of sequence similarity search in SWISS-PROT protein database, using a set of test query protein sequences, are presented in this report.

## Aim and Scope of This Study

Each of the three implementations uses a front end computer which accepts a query sequence and passes it onto the specific hardware engine for searching the database and in turn gets the matching sequence entries. Ranking the matching sequences and reconstructing the alignments are carried out by the front end computer. For the reasons that (a) a certain amount of interleaving between the search engines and the front end computers is done in Bioccelerator but not in FDF and MasPar, and (b) the front end computers are different in capabilities, it was decided to examine only the 'raw' search time pertaining to the individual hardware implementations irrespective of the front end computers. Such a 'raw' search time is the time required for the hardware implementation to search the complete subject database against a given query sequence and to report back the top scoring sequence identifier. The questions on the ranking of search results that were addressed are (i) whether the hits with top scores in each implementation are same and whether they are ranked in the same order as well as with similar scores, and (ii) whether all the members of the protein family (pertaining to the query sequence) are picked up with similarity scores higher than those of unrelated sequences.

## Hardware & Software Implementations

Specifications of the hardware implementations available for this study are as follows:

**Bioccelerator** : Bioccelerator-1, the entry level machine with a single board encompassing 16 processing elements, along with the software Bioccelerator (Version 3). The front end computer is Silicon Graphics Challenge with the operating system IRIX Release 6.2.

**FDF** : FDF-3, with 5 boards encompassing 3360 processors, along with the software Biology Tool Kit (Release Sep 1996). The front end computer is Sun SPARC station 2 with the operating system Sun OS (Release 4.1.3).

**MasPar** : MP-1 Data Parallel Unit, encompassing 4096 processors, along with the software MPSRCH (Release 3.0). The front end computer is DECstation 5000/200 with the operating system Ultrix V4.3 (Rev MP-3.22).

## I. Search Speed Performance

### Protocol

*Query sequences* : A set of 17 protein sequences (Table 1), mostly from SWISS-PROT, of lengths in the range of 16 to 2000 residues.

**Table 1. Test Query Sequences**

No.	query	Query Length	SWISS-PROT ID	Protein	Organism
1.	prion	16	From Prosite database	Signature for prion protein	
2.	ggt	43	See footnote <sup>®</sup>	Signature for gamma-glutamyl transpeptidase	<i>Rattus norvegicus</i>
3.	plasto	105	PLAS_ANAVA	Plastocyanin	<i>Anabaena variabilis</i>
4.	calmod	148	CAL6_ARATH	Calmodulin-6	<i>Arabidopsis thaliana</i>
5.	histone	194	H1_SALTR	Histone H1	<i>Salmo trutta</i>
6.	riboS3	232	RS3_ECOLI	30S ribosomal protein	<i>Escherichia coli</i>
7.	vmat	300	VMAT_MEASY	Matrix Protein	<i>Measles virus</i>
8.	coat	390	COAT_MNSV	Coat protein	Melon necrotic spot virus
9.	amid	505	AMID_PSECL	Amidase	<i>Pseudomonas chlororaphis</i>
10.	dnak	605	DNAK_BACME	dnak heat shock protein	<i>Bacillus megatericum</i>
11.	efg	703	EFG_ECOLI	Elongation factor G	<i>Escherichia coli</i>
12.	ski	750	SKI_CHICK	ski oncogene	<i>Gallus gallus</i>
13.	amdm	810	AMDM_YEAST	AMP deaminase	<i>Saccharomyces cerevisiae</i>
14.	phsg	901	PHSG_YEAST	Glycogen phosphorylase	<i>Saccharomyces cerevisiae</i>
15.	pas8	1030	PAS8_YEAST	Peroxisome biosynthesis protein	<i>Saccharomyces cerevisiae</i>
16.	abl	1520	ABL_DROME	Tyrosine-protein kinase	<i>Drosophila melanogaster</i>

17.	cin2	2005	CIN2_RAT	Sodium channel protein brain II alpha subunit	<i>Rattus norvegicus</i>
-----	------	------	----------	---	--------------------------

---

@**Taken from (Pietrokovski S, 1996 Nucleic Acids Research 24, 3836-3845).**

*Subject database* : SWISS-PROT Release 34 (59,021 entries, 21,210,389 amino acids)

*Search method* : Smith-Waterman method

*Comparison matrix* : Blosum62

*Gap open penalty* : Default values as optimised by the vendors.

Bioccelerator : 10

FDF : 9

Mpsrch\_ppa : 10

*Gap extension penalty* : Default values as optimised by the vendors.

Bioccelerator : 0.5

FDF : 3

MPSrch\_ppa : 2

*Gap penalty* : MPSrch\_pp : 10

It is to be noted that MPSRCH has two programs, namely a faster MPSrch\_pp and a slower MPSrch\_ppa. The difference between the two is that the MPSrch\_ppa uses both gap open and gap extension penalties while MPSrch\_pp uses only a single gap penalty.

### **Enumeration of 'Raw' Search Time**

**BIOCCELERATOR** : The benchmarks as reported by Bioccelerator are (a) raw Bioccelerator search time, (b) search time (which includes the Bioccelerator time as well as the host time in opening the query files and sending the job to Bioccelerator), and (c) Bioccelerator speed (in terms of million cell updates per second which is calculated using the Bioccelerator time). In this study we consider the reported Bioccelerator time and speed.

**FDF** : FDF reports the results of the unix 'time' command which lists the real, system, and user times. The raw FDF search time is the approximated real time (after subtracting the user and system times) obtained with the following parameters : S=255 (to return to the front end computer only those hits with the highest score), V=1 (to list only the top hit) and B=0 (no alignments to be reconstructed). FDF speed was calculated using the formula [(length of the query sequence x length of the database) divided by raw FDF time] and expressed in units of million cell updates/sec.

MasPar : The benchmarks as reported by MPSRCH are (a) raw MasPar time, (b) total search time (which includes raw MasPar time as well as the time on the front end computer), and (c) MasPar speed. The reported raw MasPar time and speed are used in the study.

Since MasPar can handle upto four jobs at a time, the measurements in this study were done by stopping other job queues on MasPar. Both Bioccelerator and FDF execute one job at a time. For all the three implementations, mean raw search time was calculated over three searches using the same query sequence as well as using sequences of similar lengths.

## Results

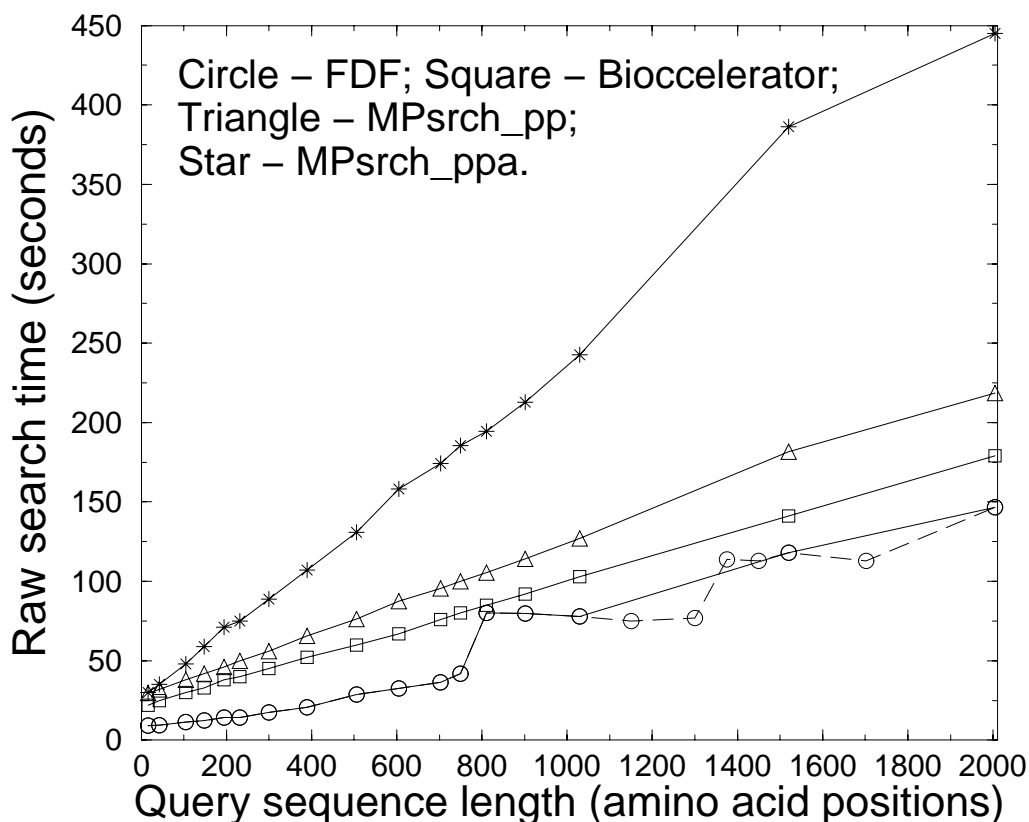
### Raw Search Time and Speed

The raw search time (in seconds) and speed (in terms of million cell updates/sec) as observed for each test query are given in Table 2 and shown in Figures 1-2.

**Table 2. Raw Search Time and Speed for Bioccelerator, FDF, and MPSRCH**

No.	Query	Raw Search Time				Speed (Million cell updates/sec)			
		Biocc	FDF	MPsrch_ pp	MPsrch_ ppa	Biocc	FDF	MPsrch_ pp	MPsrch_ ppa
1.	prion	22.0	9.2	29.62	30.02	15.43	37.9	11.46	11.30
2.	ggt	25.0	9.3	32.01	35.24	36.48	100.7	28.41	25.90
3.	plasto	30.0	11.3	37.88	47.86	74.24	202.3	58.79	46.53
4.	calmod	33.0	12.3	42.03	59.02	95.13	262.0	74.69	53.19
5.	histone	38.0	14.3	46.17	71.20	108.28	294.4	89.12	57.79
6.	riboS3	40.0	14.3	49.89	75.06	123.02	353.2	98.63	65.56
7.	vmat	45.0	17.4	56.10	88.72	141.40	375.4	113.42	71.72
8.	coat	52.0	20.6	65.61	107.17	159.08	412.2	126.08	77.19
9.	amid	60.0	28.6	76.40	131.02	178.52	384.4	140.20	81.75
10.	dnak	67.0	32.7	87.38	158.32	191.53	402.8	146.86	81.05
11.	efg	76.0	36.3	95.40	174.16	196.20	421.6	156.30	85.62
12.	ski	80.0	42.0	99.95	185.44	198.85	388.8	159.16	85.78
13.	amdm	85.0	80.0	105.59	194.33	202.12	220.4	162.71	88.41
14.	phsg	92.0	79.9	114.03	212.71	207.72	248.3	167.59	89.84
15.	pas8	103.0	78.0	127.04	242.77	212.10	287.5	171.98	89.99
16.	abl	141.0	118.0	181.77	386.38	228.65	280.5	177.37	83.44
17.	cin2	179.0	146.6	218.59	444.95	237.58	297.8	194.55	95.58

Fig 1. Raw search times on Bioccelerator, FDF, and MasPar.



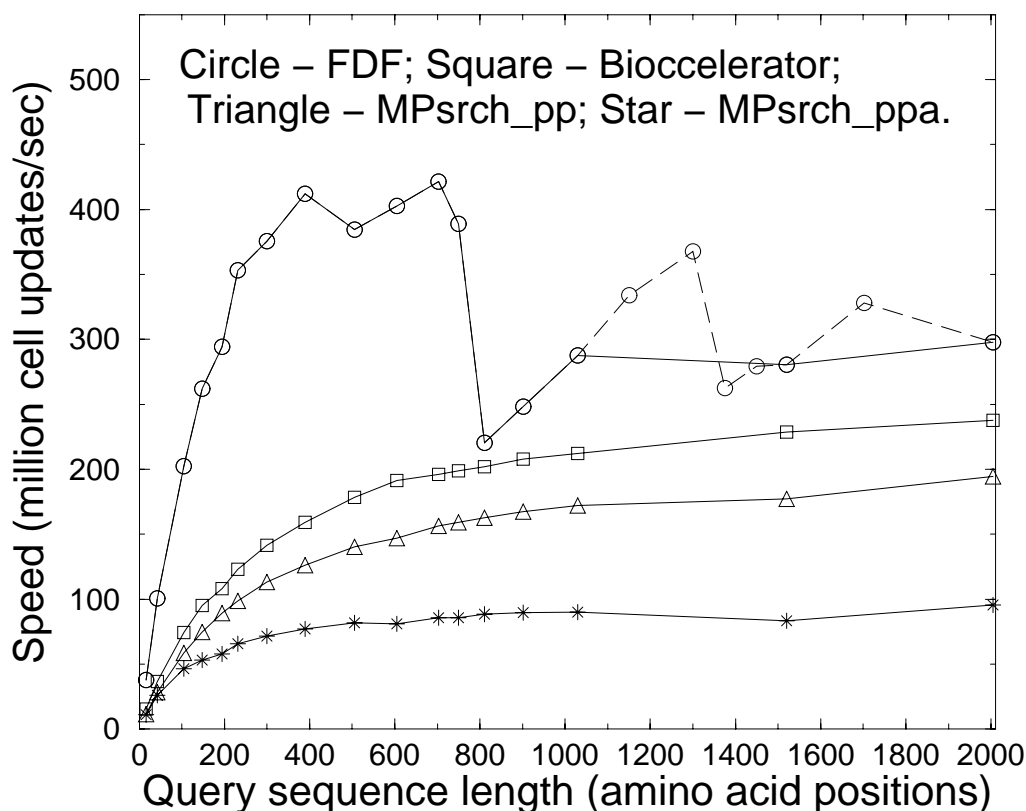
○—○ FDF : Dashed line shows raw search time for the additional queries (see text).

Figures 1-2 show values for 5 additional queries of lengths 1150, 1300, 375, 1450, and 1702. These additional queries were chosen to represent the length range gap between the test query 15 and 17 (Table 1). Except in the case of FDF (shown by dashed line in figures), the data points for these additional queries got plotted onto the line obtained for the test queries.

The data presented in Table 2 and Figures 1-2 points to the following observations.

- (i) The pattern of speed distribution with increase in the length of the query is asymptotic (Fig. 2). In all the implementations, the speed increases rapidly as the query length reaches ~400 residues and shows a saturation for query lengths of 1000 and more.
- (ii) Given the configurations of the hardware available for this study, FDF is the fastest of the three with MPSRCH trailing as the slowest.
- (iii) MPsrch\_ppa is much slower than MPsrch\_pp (for medium to longer queries, there is a 1.5 to 2 fold difference in the search time) for the reason that MPsrch\_pp uses a single gap penalty while MPsrch\_ppa uses two. It is to be noted that the earlier releases of MPSRCH had only the MPsrch\_pp version for considerations of speed.

Fig 2. Search speed of Bioccelerator, FDF, and Maspar



○ —○ FDF : Dashed line shows search speed for the additional queries (see text).

- (iv) FDF shows a break in the linearity of the graph (Figures 1 and 2), when the length of the query sequence increases to 750 residues and the difference in the speed with that of Bioccelerator becomes less. This is due to the fact that FDF treats longer queries as multiple fragments of window size 750 residues with an overlap of 150 residues. Such a limitation is due to the following reason : The five boards of FDF are distributed into 1 virtual pipeline board and 4 physical pipeline boards (3360 processors in total). Approximately 40 processors are required for each amino acid position of the query sequence. By virtue of the ability of the virtual pipeline to do data buffering and recycling, the constraint on the length of the query sequence as limited by the number of processors is removed thereby setting the current limit on the length of the query sequence as 750 amino acid positions. Queries of length more than 750 residues are thus too long to fit the virtual pipeline. The search is done by segmenting the query sequence onto overlapping (by 150 residues) multiple fragments of window size 750. The break in the linearity of the graph recurs again at query lengths of ~1400 and at ~2000 (shown by dashed line in Figures 1 - 2). And thus the search time and speed distributions for FDF show a staircase effect.
- (v) The raw search time and speed reported so far were measured with the condition that only top hit be listed. MPsrch\_pp and Bioccelerator always listed the exact SWISS-PROT entry (note that the query sequences are also originally from SWISS-PROT database) while FDF did not return the exact SWISS-PROT entry as the top hit in a

major number of cases of test queries. For this purpose, FDF had to be executed with the options of *rescore* and *sort\_by\_highscore* in which case the exact sequence entry was returned as the top hit. The benchmark times observed with these options are given in Table 3. Since no interleaving is implemented between the search engine and the front end computer, the overall elapsed (real) time becomes much higher in the case of long queries (Table 3). The raw FDF time does not change. Thus there is an overload on the front end computer in the case of FDF. It is to be pointed out that such an overload is dependent on the power of the front end computer.

**Table 3. Benchmark Times on FDF with the Rescore and Sort\_by\_highscore Options.**

No <sup>@</sup>	query	real time	user time	system time	raw FDF time without options
5.	histone	19.9	3.7	0.7	15.5
6.	riboS3	18.3	2.8	0.4	15.1
7.	vmat	21.3	3.1	0.3	17.9
8.	coat	22.9	2.0	0.4	20.5
9.	amid	31.7	2.5	0.4	28.8
10.	dnak	101.1	65.5	0.9	34.7
11.	efg	58.9	20.0	0.5	38.4
12.	ski	46.9	4.6	0.4	41.9
13.	amdm	91.8	8.7	0.6	82.5
14.	phsg	102.4	19.8	0.7	81.9
15.	pas8	133.5	57.1	0.9	75.5
16.	abl	429.5	311.6	1.6	116.3
17.	cin2	257.4	109.0	0.9	147.5

<sup>@</sup>The data for the first 4 queries are not given because the values are not very different from those without the options.

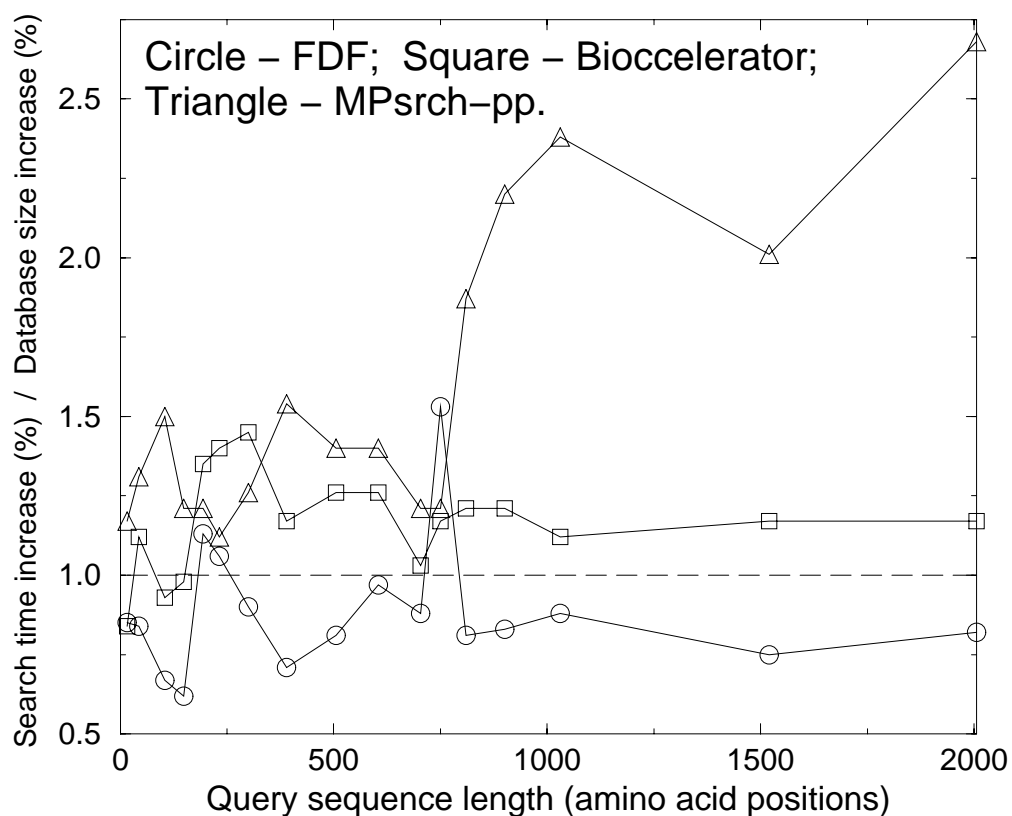
### Effect of change in database size on raw search time

The influence of database size on raw search time was scrutinised by examining the raw search times on protein databases of different sizes. SP-TREMBL 1.0 (86,039 entries, 25,760,174 amino acids) and Rel33 (52,205 entries, 18,531,384 amino acids) or Rel34 (59,021 entries, 21,210,389 amino acids) of SWISS-PROT were considered. Rel34 of SWISS-PROT showed an increase of 14.4% over Rel33 while the size of SP-TREMBL was higher than that of Rel34 of SWISS-PROT by 21.4%. A parameter indicating the ratio between percentage increase in search time and percentage increase in database size was calculated for every test query in each implementation. A value of 1 for such a ratio indicates a proportionate (linear) increase in search time as the database size increases. A value of more than 1 indicates a disproportionate increase in search time and a value of less than 1 indicates that increase in database size does not affect the speed performance to the extent of change in the database size. The ratios thus calculated are given in Table 4 and they are shown in Figure 3.

**Table 4. Ratio Between Percentage Increase in Search Time and Percentage Increase in Database Size.**

No.	query	query length	Bioccelerator	FDF	MPsrch -pp
1.	prion	16	0.84	0.85	1.17
2.	ggt	43	1.12	0.84	1.31
3.	plasto	105	0.93	0.67	1.50
4.	calmod	148	0.98	0.62	1.21
5.	histone	194	1.35	1.13	1.21
6.	riboS3	232	1.40	1.06	1.12
7.	vmat	300	1.45	0.90	1.26
8.	coat	390	1.17	0.71	1.54
9.	amid	505	1.26	0.81	1.40
10.	dnak	605	1.26	0.97	1.40
11.	efg	703	1.03	0.88	1.21
12.	ski	750	1.17	1.53	1.21
13.	amdm	810	1.21	0.81	1.87
14.	phsg	901	1.21	0.83	2.20
15.	pas8	1030	1.12	0.88	2.38
16.	abl	1520	1.17	0.75	2.01
17.	cin2	2005	1.17	0.82	2.68

Fig 3. Effect of database size change on search time.



The figure indicates that in the case of FDF, the increase in database length does not warrant a proportionate increase in search time; and in the case of Biocelerator and MasPar increase in database length has brought about disproportionate increases in the search time. Such a disproportionate increase in time is very high for MasPar (MPsrch\_pp).

## II. Ranking of the Search Results

As indicated earlier, the test query sequences are from different protein families, of different lengths and are from different taxonomic kingdoms (such as prokaryotes, eukaryotes, and viruses). A comparative study of the search results and the ranking of the resultant hits was carried out. The specific questions that were addressed are (a) whether the closely-related (to the query sequence) hits are well-separated in scores from those that are distantly related; (b) whether the hits are ranked into appropriate groups of differential relation to the query sequence and whether such groupings are common among the implementations; (c) whether the scores and ranking order of the related sequences are the same across the implementations; and (d) how many of the related sequences are picked up by the implementations.

### ***Protocol***

The protocol is same as given earlier. In addition, it is to be mentioned that in the case of MasPar implementation, MPsrch\_ppa was used and the ranking of the resultant hits was by the original scores. FDF was used with the rescore and sort\_by\_highscore options.

### ***Results***

#### Rank-wise Grouping of Resultant Hits

The score-wise ranking of the resultant hits is shown in Table 5. The ranking procedure was found to classify the hit sequences into appropriate groups. Such a grouping was invariant among the implementations. The groupings are of the following three types : (a) according to the major taxonomic kingdoms (prokaryote, chloroplast, mitochondria, archaeobacteria, and eukaryote) - such cases are ggt, plasto, riboS3, amid, dnak, efg, and phsg; (b) according to the taxonomic subclassification within a single kingdom. - such cases are calmod, histone, vmat, coat, and amdm; (c) according to the functional classification - such cases are ski, pas8, abl, and cin2. The range of scores for the groups shown against each test sequence in the table indicated that the group of hits that are closely related to the query sequence were found to be clearly demarked from the rest of the hits. It is also seen that in general there is a 2-3 fold difference in the scores between the closely-related group of hits and the less-closely related group.

**Table 5. Score-wise Groupings of Hits.**

No.	query	Range of Scores			Groups of hits
		Bioccel	FDF	MPsrch_ppa	
1.	prion	79-79	79-79	79-79	All prions in the database.
2.	ggt	(209-119) (73-69)	(209-103) (73-69)	(209-107) (73-69)	Eukaryotic ggt's. Prokaryotic ggt's.
3.	plasto	(547-530) (277-183)	(547-530) (265-174)	(547-530) (272-181)	Prokaryotic <i>anabaena</i> plastocyanin. Plastocyanins from plastids.
4.	calmod	(756-700) (691-539)	(756-700) (691-539)	(756-700) (691-539)	Calmodulins from plants. Calmodulins from other eukaryotes.
5.	histone	(935-825) (591-223)	(935-814) (542-180)	(935-818) (572-191)	Histone H1's from <i>Salmoniformes</i> . H1's and H5's from other eukaryotes.
6.	riboS3	(1178-1001) (655-609) (590-438) (428-285) (257-195) (183-162) (149-127)	(1178-1003) (656-595) (585-429) (371-264) (249-192) (182-149) (142-118)	(1178-1002) (655-607) (590-437) (418-277) (253-195) (181-154) (142-119)	Prokaryotic ribosomal 30S protein. From prokaryotes. Mix of prokaryotes and plastids. From plastids. From archaeobacteria. From mitochondria. From eukaryotes.
7.	vmat	(1561-964) (451-417) (208-101)	(1561-967) (437-399) (175-86)	(1561-965) (444-410) (185-90)	Matrix protein from morbillivirus of paramyxoviridae. From paramyxovirus of paramyxoviridae. --- same ---
8.	coat	1984 (605-469) (277-223) (149-110)	1984 (530-445) (252-196) (102-73)	1984 (575-445) (267-202) (108-97)	The query coat protein sequence From tombus- and diantho- viridae. From carmoviridae. From necro- and sobemo- viridae.
9.	amid	2662 1169 (383-197)	2662 1154 (385-164)	2662 1168 (384-177)	The query amidase sequence. Prokaryotic amidase. Eukaryotic and putative prokaryotic amidases and related hydrolases.
10.	dnak	3014 (2694-1738) (1723-1646) (1643-1637) (1631-1464)  (1449-1408) (1400-1206)	3014 (2678-1583) (1539-1470) (1467-1447) (1444-1246)  (1244-1198) (1190-1011)	3014 (2690-1707) (1689-1612) (1612-1605) (1591-1401)  (1399-1359) (1355-1162)	The query dnak sequence. From prokaryotes. Mix from mitos, plastids and proks. Mitochondrial stress 70 proteins. Mix of dnak's/HS70's from proks & euks. Eukaryotic 78KD glucose-regulated proteins. Eukaryotic HS7x proteins.

		(1090-241)	(1002-160)	(1076-167)	Mix of all the above.
11.	efg <sup>®</sup>	(3615-3071) (2124-998) (781-710) (632-590)  (563-388)  (385-258) (233-186)	(3615-3067) (2068-523) (377-333) (306-232)  (190-175)  (232-191) (132-111)	(3615-3071) (2098-698) (662-592) (541-422)  (422-345)  (219-210) (157-126)	EF-G's from prokaryotes. EF-G's from other prokaryotes. EF-2's from archaeobacterias. Prokaryotic tetracycline resistance proteins. Eukaryotic EF-2's & prokaryotic RF's. GTP-binding proteins. EF-Tu's of prokaryotes and organelles.
12.	ski	(3887-2318) (1206-888)	(3887-2318) (822-821)	(3887-2318) (1035-874)	<i>ski</i> oncogenes. <i>ski</i> -related oncogenes.
13.	amdM	4302 (1951-1449)  (738-345)	4302 (1797-1423)  (345-314)	4302 (1901-1422)  (583-345)	AMP deaminase from yeast. AMP deaminase from higher eukaryotes. Hypothetical proteins and fragments.
14.	phsg	4734 (2221-1953) (1915-1617)	4734 (2056-1718) (1704-1073)	4734 (2187-1901) (1875-1569)	The query phosphorylase. From eukaryotes. From prokaryotes.
15.	pas8	5316  (1779-1657)  1031  (748-471)	5316  (1552-1501)  872  (665-441)	5316  (1669-1601)  954  (707-463)	The query peroxisome biosynthesis protein. Other peroxisome biosynthesis proteins. Related peroxisome assembly proteins. The other related proteins of AAA family of ATPases.
16.	abl	7935 (2067-1649) (1082-966)	7935 (1919-1496) (943-941)	7935 (1931-1551) (1071-919)	The query tyrosine-protein kinase. <i>abl</i> 's from other eukaryotes. Other tyrosine kinases.
17.	cin2	10379 (8995-6405)  (5620-2000) (1644-633)	10379 (8901-5761)  (4972-1954) (768-327)	10379 (8980-6244)  (5302-1464) (1107-388)	The query sodium channel protein. Other sodium channel proteins from rat. <i>cin</i> 's from other eukaryotes. Related calcium channel proteins.

<sup>®</sup>In the case of FDF, there is a change in the order of the groups of eukaryotic EF-2's and GTP-binding proteins. And also release factors appear in the group of tetracycline resistance proteins.

## Selectivity

The ability of the implementations to pick up all the related sequences as top hits before picking up any unrelated sequence was examined. For every test query sequence, the number of related sequences belonging to the same family were identified by searching the ‘definition’ records from SWISS-PROT through keywords derived from the sequence entry corresponding to the query sequence and also scanning the ‘comments’ records which reports information on similarity. PIR which includes information on protein families was also consulted. The size of the protein family thus arrived is given in brackets under the column ‘Selectivity’ in Table 6. Such a size does not include those SWISS-PROT entries which report sequences of only fragments and not the complete protein.

**Table 6. Selectivity and Ranking Order of Hits by the Three Implementations.**

No.	Query	Selectivity <sup>@</sup>	Ranking <sup>§</sup>	Comments
1.	prion	31 (31)	Not applicable	
2.	ggt	8 (8)	Same	
3.	plasto	35 (35)	± 5	
4.	calmod	45 (47)	Same	See <sup>1</sup> below.
5.	histone	70 (94)	± 5	
6.	riboS3	46 (49)	± 2 (in 2 cases by ± 5)	
7.	vmat	28 (28)	Same	
8.	coat	13 (13)	± 2	
9.	amid	8 (8)	± 4	
10.	dnak	167 (167)	Same in the top 8 hits	
11.	efg	60 (60)	± 2 in the top 17 hits	See <sup>2</sup> below
12.	ski	6 (6)	Same	
13.	amdm	7 (7)	Same	
14.	phsg	20 (20)	± 5 (in one case by 7)	
15.	pas8	82 (86)	± 4 (in one case by 23)	
16.	abl	22 (22)	± 4	See <sup>3</sup> below
17.	cin2	12 (12)	Same	See <sup>4</sup> below

<sup>@</sup>Selectivity - No. of related sequences (of the family size) picked up before other unrelated sequences are picked up. The results are same for all the implementations.

<sup>§</sup>This indicates the differences in the ranking order of hits among the three implementations. The ranking order as produced by FDF was taken as reference. The data given in the table indicates the maximum difference in the ranking order as produced by Bioccelerator and Mpsrch\_ppa against that produced by FDF. Such a data is not applicable in the case of prion since the query is a conserved signature sequence and hence the ranking order of the hits is not relevant.

<sup>1</sup>The remaining two calmodulins (namely, CALM\_YEAST and CALM\_STRPU) are ranked after calcium-binding proteins for the reason that these two proteins possess only 3 and 2 binding sites respectively while calmodulins and calcium-binding proteins possess 4 such sites.

<sup>2</sup>Included are EF-G’s of prokaryotes and organelles, EF-2’s of archaeobacteria and eukaryotes, and the more related prokaryotic tetracycline resistance proteins and GTP-binding proteins. Excluded are EF1’s of eukaryotes and EF-Tu proteins.

<sup>3</sup> In addition to these closely related sequences, there are many protein kinase entries that show high homology (even the 300<sup>th</sup> hit has a percent similarity of 59).

<sup>4</sup> The hits corresponding to sodium channel proteins are followed by 14 calcium channel proteins.

It can be seen from the table that all the members of the family were picked up (except in the cases of calmod, histone, and riboS3) by the three implementations. In the cases of calmod and riboS3 only 2-3 entries were not picked up while in the case of histone, 75% of the family members have been picked up before distantly related members were picked.

### Ranking Order And Score Difference Among The Implementations

The preservation of ranking order of the hits among the implementations was scrutinised. It was observed that there is a maximum shift of +/-5 in the ordering of hits among the different implementations (Table 6). Such a shift always occurred within the individual groups and never crossed the groups. This is significant considering the fact that the default gap open and extend penalties are different among the implementations (see the *protocol*).

As a next step, the scores of the hits were compared among the implementations. The scores obtained for the topmost hit (which is always the same sequence as the query) were the same in all the three implementations, and this meant that the same comparison matrix and ranking function are being used by the implementations. Since the gap penalties (in particular the gap extension) is different among the implementations, the scores for the hits are expected to be different among the implementations. The score ranges of the group of closely-related sequences against each implementation (as given in Table 5) were scrutinised for score differences. In 9 out of 17 cases, the score ranges for the group of closely-related hits are similar among the implementations. Such cases are prion, plasto, calmod, histone, riboS3, vmat, amid, efg, and ski. In the remaining 8 cases, a deviation of 2% to 10% was observed.

### Suggested Further Readings

1. Smith T.F. & Waterman M.S. (1981) Identification of common molecular subsequences. *J Mol Biol* 147, 195-197.
2. Pearson W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 11, 635-650.
3. Pearson W.R. (1995) Comparison of methods for searching protein sequence databases. *Protein Science* 4, 1145-1160.
4. Hughey R. (1996) Parallel hardware for sequence comparison and alignment. *CABIOS* 12, 473-479.