PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology

Per Eystein Sæbø[1], Sten Morten Andersen[2], Jon Myrseth[1], Jon K. Laerdahl[1] and Torbjørn Rognes[1,2,3,*]

[1]Centre for Molecular Biology and Neuroscience, Institute of Medical Microbiology, University of Oslo and Rikshospitalet-Radiumhospitalet HF, NO-0027 Oslo, Norway, [2]Sencel Bioinformatics AS, Motzfeldts gate 16, NO-0187 Oslo, Norway and [3]Department of Informatics, University of Oslo, PO Box 1080, NO-0316, Oslo, Norway

Received February 15, 2005; Revised and Accepted March 23, 2005

ABSTRACT

PARALIGN is a rapid and sensitive similarity search tool for the identification of distantly related sequences in both nucleotide and amino acid sequence databases. Two algorithms are implemented, accelerated Smith–Waterman and ParAlign. The ParAlign algorithm is similar to Smith–Waterman in sensitivity, while as quick as BLAST for protein searches. A form of parallel computing technology known as multimedia technology that is available in modern processors, but rarely used by other bioinformatics software, has been exploited to achieve the high speed. The software is also designed to run efficiently on computer clusters using the message-passing interface standard. A public search service powered by a large computer cluster has been set-up and is freely available at www.paralign.org, where the major public databases can be searched. The software can also be downloaded free of charge for academic use.

INTRODUCTION

Similarity searching is an essential part of sequence analysis and searching public sequence databases to gain more information about a sequence is one of the tasks carried out most frequently in bioinformatics.

Web services for database homology searches are provided at www.ncbi.nih.gov/BLAST by the National Center for Biotechnology Information (NCBI) in the United States using the NCBI BLAST program (1), and at www.ebi.ac.uk/Tools/similarity.html by the European Bioinformatics Institute (EBI) by using the NCBI BLAST program as well as the FASTA (2) and the WU-BLAST (Gish 1996–2004; http://blast.wustl.edu) programs. Many other institutions provide similar services.

PARALIGN is an alternative to the services provided by the NCBI, EBI and others. The PARALIGN service employs two different very sensitive algorithms that enable the identification of similarities that may pass undetected by the algorithms used by the other services. As the service is running on a powerful computer cluster, searches are carried out at a high speed challenging that of other services.

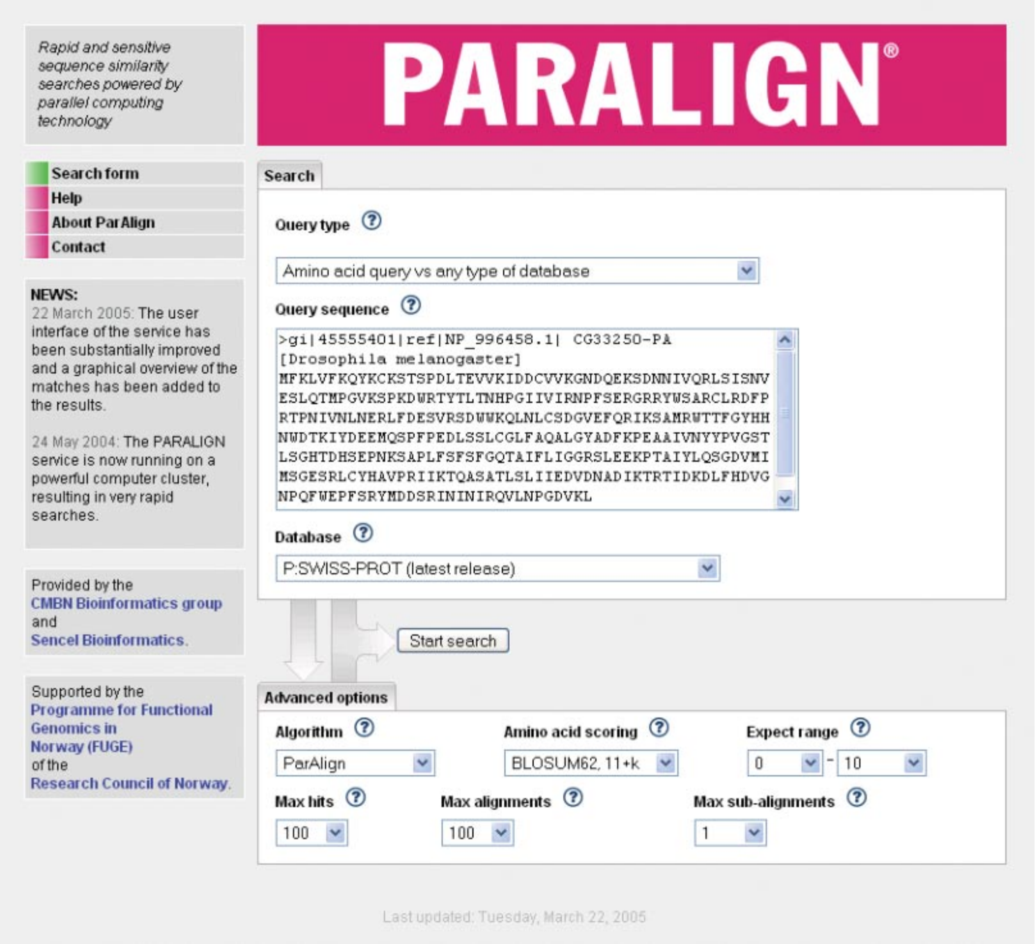
The service is most useful for the identification of both nucleotide and amino acid database sequences that have significant, but possibly limited, similarity to the query sequence. It is especially useful for the detection of remote protein homologs in cases where no other protein family members are known. Sensitive nucleotide sequence comparisons are useful for the detection of unwanted cross-hybridization loci of siRNAs, microarray probes or PCR primers, or when looking for subtle similarities between nucleotide sequences, e.g. non-coding RNA sequences.

The search form is located at www.paralign.org where the query sequence is entered and the database selected (Figure 1). One may choose from three different types of searches: (i) amino acid query sequence against any type of database, (ii) translated nucleotide query sequence against any type of database, or (iii) a direct comparison of a nucleotide query sequence against a nucleotide sequence database. The maximum number of hits, alignments and suboptimal alignments that are to be shown, as well as the search algorithm, the *E*-value range and the scoring parameters (score matrix and gap penalties) may be selected. When the search is completed, information about the query, database and search parameters will be shown together with a graphical overview of the hits, a list of hits (with hypertext links to the alignments below) and alignments (with hypertext links to the corresponding NCBI Entrez entries) (Figure 2).

SEARCH METHODS

The web service employs the PARALIGN software to carry out searches in sequence databases. The PARALIGN software

*To whom correspondence should be addressed. Tel: +47 22844787; Fax: +47 22844782; Email: torbjorn.rognes@medisin.uio.no



Rapid and sensitive sequence similarity searches powered by parallel computing technology

PARALIGN®

NEWS:
22 March 2005: The user interface of the service has been substantially improved and a graphical overview of the matches has been added to the results.

24 May 2004: The PARALIGN service is now running on a powerful computer cluster, resulting in very rapid searches.

Provided by the CMBN Bioinformatics group and Sencel Bioinformatics.

Supported by the Programme for Functional Genomics in Norway (FUGE) of the Research Council of Norway.

Query type: Amino acid query vs any type of database

Query sequence:

```
>gi|45555401|ref|NP_996458.1| CG33250-PA
[Drosophila melanogaster]
MFKLVFKQYKCKSTSPDLTEVVKIDDCVVKGNDQEKSDNNIVQRLSISNV
ESLQTMPGVKSPKDWRTYTLTNHPGIIVIRNPFSERGRRYWSARCLRDFP
RTPNIVNLNERLFDESVRSDWWKQLNLCSDGVEFQRIKSAMRWTTFGYHH
NWDTKIYDEEMQSPFPEDLSSLCGLFAQALGYADFKPEAAIVNYYPVGST
LSGHTDHSEPNKSAPLFSFSFGQTAIFLIGGRSLEEKPTAIYLQSGDVMI
MSGESRLCYHAVPRIIKTQASATLSLIIEDVDNADIKTRTIDKDLFHDVG
NPQFWEPFSRYMDDSRININIRQVLNPGDVKL
```

Database: P:SWISS-PROT (latest release)

Algorithm: ParAlign

Amino acid scoring: BLOSUM62, 11+k

Expect range: 0 - 10

Max hits: 100

Max alignments: 100

Max sub-alignments: 1

Last updated: Tuesday, March 22, 2005

Figure 1. The PARALIGN home page at www.paralign.org contains the search form where the query sequence is entered and the database and the search parameters are selected. Clicking on a question mark opens a window with detailed help for each field.

implements two search methods: (i) the accelerated Smith–Waterman method (3) and (ii) the heuristic ParAlign method (4). To achieve speed, both methods exploit parallel computing technology, as described in more detail below.

The accelerated Smith–Waterman method performs searches using the Smith–Waterman algorithm (5) accelerated with the parallel computing technology to achieve a speed more than six times higher than the implementation in SSEARCH (6) or similar software. The alignment scores calculated are identical to those of other Smith–Waterman implementations.

The ParAlign method is based on a search algorithm that is designed specifically with the parallel computing technology in mind. Despite being a heuristic algorithm similar to NCBI BLAST in speed (1), its sensitivity is similar to that of searches based on the Smith–Waterman algorithm. For ungapped alignments, the alignment scores calculated by the ParAlign algorithm are identical to those of the Smith–Waterman algorithm.

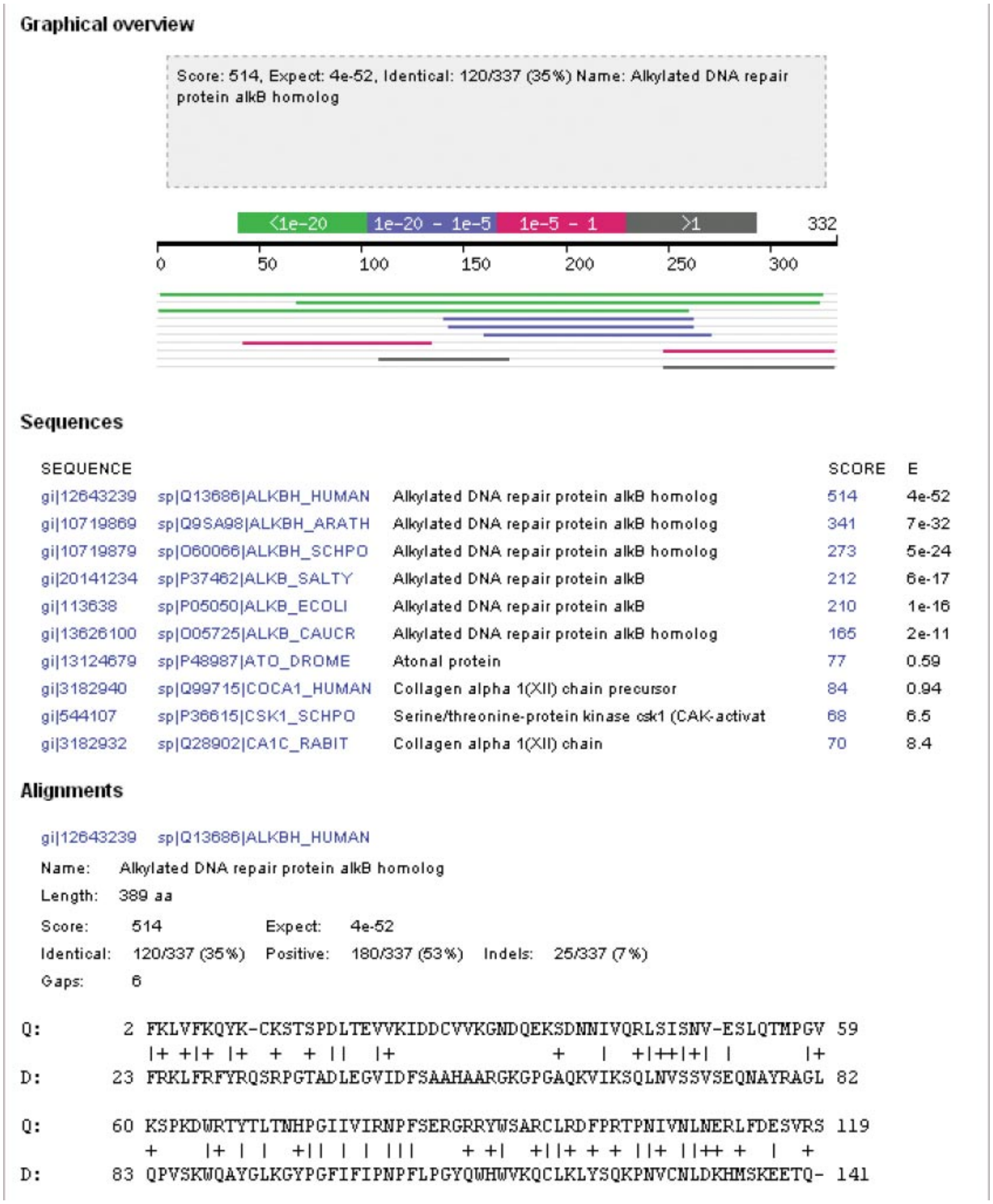
PARALLEL COMPUTING

The software and service makes use of the parallel computing at several levels.

First level: Both methods implemented in PARALIGN employ parallel computing technology known as multimedia technology (MMX/SSE, AltiVec) or single-instruction multiple-data (SIMD) technology to gain high speed. This type of technology was originally designed for rapid processing of images, video and sound, and allows the microprocessor to carry out many operations on small numbers in parallel, instead of one single operation on a large number at a time. Despite being available in most modern processors, the technology is rarely used by other bioinformatics software, except for a recent implementation of the HMMER software (7) using AltiVec (http://hmmer.wustl.edu). In PARALIGN, this technology is used to process the score values of several cells in the alignment matrix in parallel.

Second level: On symmetric multiprocessing (SMP) computers having several microprocessors with shared common memory, multiple threads of the PARALIGN program can run concurrently and search through different parts of the sequence database. These threads are implemented using the standard Posix threads (pthreads).

Third level: Clusters of multiple computing nodes connected together through a network may run one or more instances of PARALIGN on each node that searches through different parts of the sequence database. The communication between the nodes is implemented using the message-passing interface (MPI) standard, as implemented in the MPICH (http://www-unix.mcs.anl.gov/mpi/mpich/) or other packages. The



Score: 514, Expect: 4e-52, Identical: 120/337 (35%) Name: Alkylated DNA repair protein alkB homolog

Sequences

SEQUENCE			SCORE	E
gi|12643239	sp|Q13686|ALKBH_HUMAN	Alkylated DNA repair protein alkB homolog	514	4e-52
gi|10719869	sp|Q9SA98|ALKBH_ARATH	Alkylated DNA repair protein alkB homolog	341	7e-32
gi|10719879	sp|O60066|ALKBH_SCHPO	Alkylated DNA repair protein alkB homolog	273	5e-24
gi|20141234	sp|P37462|ALKB_SALTY	Alkylated DNA repair protein alkB	212	6e-17
gi|113638	sp|P05050|ALKB_ECOLI	Alkylated DNA repair protein alkB	210	1e-16
gi|13626100	sp|O05725|ALKB_CAUCR	Alkylated DNA repair protein alkB homolog	165	2e-11
gi|13124679	sp|P48987|ATO_DROME	Atonal protein	77	0.59
gi|3182940	sp|Q99715|COCA1_HUMAN	Collagen alpha 1(XII) chain precursor	84	0.94
gi|544107	sp|P36615|CSK1_SCHPO	Serine/threonine-protein kinase csk1 (CAK-activat	68	6.5
gi|3182932	sp|Q28902|CA1C_RABIT	Collagen alpha 1(XII) chain	70	8.4

Alignments

gi|12643239 sp|Q13686|ALKBH_HUMAN

Name: Alkylated DNA repair protein alkB homolog

Length: 389 aa

Score: 514 Expect: 4e-52

Identical: 120/337 (35%) Positive: 180/337 (53%) Indels: 25/337 (7%)

Gaps: 6

```
Q:         2 FKLVFKQYK-CKSTSPDLTEVVKIDDCVVKGNDQEKSDNNIVQRLSISNV-ESLQTMPGV 59
             |+ +|+ |+  +  + ||  |+              +   |  +|++|+| |      |+
D:        23 FRKLFRFYRQSRPGTADLEGVIDFSAAHAARGKGPGAQKVIKSQLNVSSVSEQNAYRAGL 82

Q:        60 KSPKDWRTYTLTNHPGIIVIRNPFSERGRRYWSARCLRDFPRTPNIVNLNERLFDESVRS 119
             +    |+ | |  +|| | | |||    + +|  +||+ + + ||+ ||++ +  |  +
D:        83 QPVSKWQAYGLKGYPGFIFIPNPFLPGYQWHWVKQCLKLYSQKPNVCNLDKHMSKEETQ- 141
```

Figure 2. The search results include a graphical overview of the hits, a list of matches and the sequence alignments. In the graphical overview, the position of the matches relative to the query sequence is indicated with lines coloured according to the *E*-value of the alignment. Hypertext links are provided to further sequence information.

cluster powering the PARALIGN service consists of 33 nodes with 2 Intel Xeon 2.4 GHz microprocessors and 2 GB RAM each. The nodes are running the Red Hat Enterprise Linux 3.0 Workstation operating system, and are connected using Gigabit Ethernet switches.

PROCESSING OF SEARCH REQUESTS

Searches can be carried out in a wide range of sequence databases. The common public databases provided by the NCBI are downloaded weekly and made available for searches. Databases are processed and stored in the same format as used by the NCBI BLAST (1), making it easy to use the same databases for both programs. The databases are distributed to the local hard disks on all nodes in the cluster.

Submissions through the PARALIGN web interface are processed as outlined in Figure 3. The user submits search requests using the search form at www.paralign.org by entering a query sequence and selecting a database and appropriate search parameters. The web server will process the request by running a submission script that checks the input. If the input is ok, a search is submitted to a scheduling system using the

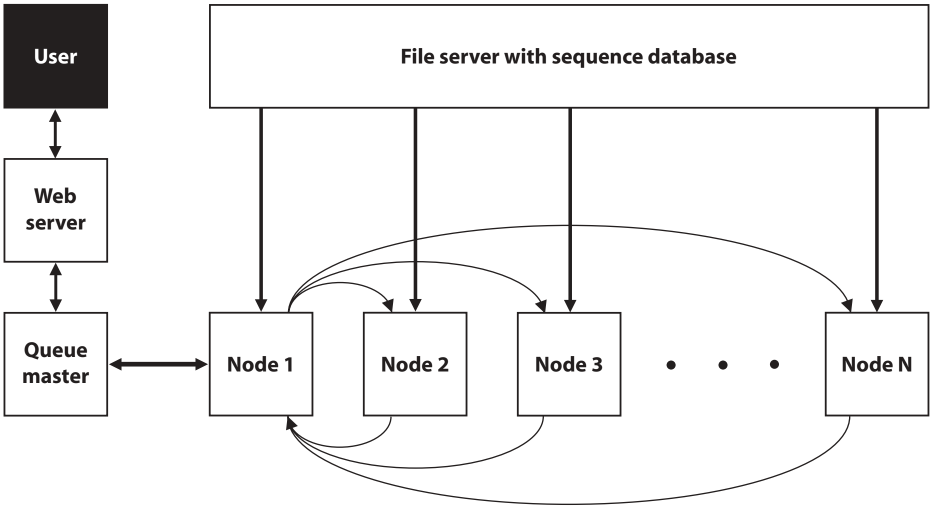


Figure 3. The data flow for distributed searches on the computer cluster is illustrated in this diagram. Database sequences are loaded directly from a file server into memory on each node. The query sequence and the search parameters are transferred from the user, via the web server and queuing system to the nodes. Search results from each node are collected by the first node which then generates the final output that is presented to the user.

TORQUE Resource Manager (www.clusterresources.com/products/torque/), which is based on PBS. Another script then takes over and checks if the results are ready. If not, the user is presented with the information on the status of the search, including the position in the queue, whether the search is running, expected time of finishing, etc. When results are ready, they are presented to the user.

When a search is scheduled to be run on the cluster, the MPI system will start PARALIGN on a selected number of nodes in the cluster. Each node is responsible for searching a predetermined part of the sequence database. The results from each node are collected by the first node, which puts them together to a final result.

The system is designed so that multiple searches in the same database will usually result in the same node searching through exactly the same part of the database. This part of the database will usually be automatically cached in RAM by the operating system the first time it is read from the local hard disk, making subsequent searches go faster because they will require little, if any, disk access.

To get an impression of the total time used by the system from submission to results, we used the *Saccharomyces cerevisiae* Mag protein (UniProt accession no. P22134) of 296 amino acids as a query sequence against different databases using default search parameters. Searching the NCBI non-redundant (nr) protein database containing a total of ~790 million amino acids took 15 s using the ParAlign algorithm and 30 s using the accelerated Smith–Waterman algorithm. A search in the human expressed sequence tag (EST) database (translated) containing ~3200 million nucleotides took 30 s using the ParAlign algorithm and 70 s using the accelerated Smith–Waterman algorithm. In all cases, ~13 s are overhead not used for the actual searches.

AVAILABILITY

The web service is freely available for all. Stand-alone executables are available free of charge for academic use, while commercial use requires a paid license from the company Sencel Bioinformatics AS (www.sencel.com). The stand-alone software is available for five different computing platforms: Linux on i386, Linux on Itanium, HP-UX on Itanium, MacOS on PowerPC and Tru64 Unix on Alpha. The software, as well as additional information about it, is available from the company's website.

If the web service is used in any published work, please cite this paper. Alternatively, the papers describing the methods used may be cited: the accelerated Smith–Waterman algorithm (3) and/or the ParAlign algorithm (4).

Both the stand-alone software and the web service are under development. In particular, we would like to improve the web interface and the presentation of the results. We have also planned a number of improvements to the stand-alone program, including versions for other computer architectures.

ACKNOWLEDGEMENTS

We are grateful for the support from Andreas Botnen and others at the Centre for Information Technology Services (USIT) at the University of Oslo, who set up and are hosting the computers. We also thank the Research Council of Norway, in particular the National Programme for Research in Functional Genomics in Norway (FUGE), as well as the Norwegian Cancer Society for financial support. Funding to pay the Open Access publication charges for this article was provided by the Research Council of Norway.

Conflict of interest statement. Torbjørn Rognes holds stock in the company Sencel Bioinformatics AS, the makers of PARALIGN. The other authors have no conflicts of interest to declare.

REFERENCES

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

2. Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*., **183**, 63–98.
3. Rognes,T. and Seeberg,E. (2000) Six-fold speed-up of Smith–Waterman sequence database searches using parallel processing on common microprocessors. *Bioinformatics*, **16**, 699–706.
4. Rognes,T. (2001) ParAlign: a parallel sequence alignment algorithm for rapid and sensitive database searches. *Nucleic Acids Res*., **29**, 1647–1652.
5. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.
6. Pearson,W. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.
7. Eddy,S. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.