



User's Guide

PARALIGN version 5.0

31 March 2008

Copyright © 2001-2008 Sencel Bioinformatics AS. All rights reserved.

Written by Torbjørn Rognes, Sten Morten Misund-Asphaug, Per Eystein Sæbø & Ophélie Aussedat.

Published by Sencel Bioinformatics AS, Oslo, Norway.

Website: www.sencel.com

E-mail: sencel AT sencel DOT com

The most recent version of this document is available at the following address:

<http://www.sencel.com/images/usersguide.pdf>

Copying is prohibited without written permission from the publisher.

Preface

PARALIGN has been and is developed in a close collaboration between the company Sencel Bioinformatics AS, and a bioinformatics research group at the University of Oslo and Rikshospitalet Medical Centre, Oslo, Norway.

The following people have been involved in the development of PARALIGN: Ophélie Aussedat, Espen Hannisdal, Jon K Lærdahl, Sten Morten Misund-Asphaug, Jon R. Myrseth, Torbjørn Rognes, Torgeir A. Ruden & Per Eystein Sæbø.

We are very grateful to all technical support from the Centre for Information Technology Services at the University of Oslo.

We are also most grateful to the financial support from the Research Council of Norway.

Oslo, 31 March 2008

The PARALIGN development team

Table of Contents

Preface.....	3
1 Introduction.....	5
2 Software installation.....	6
2.1 System requirements.....	6
2.2 Registering for access.....	6
2.3 Downloading the software.....	7
2.4 Installing the files.....	7
2.5 License management.....	7
3 Database preparation.....	9
3.1 Simple databases.....	9
3.2 Database alias files.....	9
4 Searching.....	12
4.1 Selecting search algorithm.....	12
4.2 Selecting sequence types, translation and strands to search.....	12
4.3 Specifying the sequence database.....	13
4.4 Specifying a subset of the database to search.....	13
4.5 Specifying the query sequence.....	14
4.6 Specifying alignment scoring parameters.....	14
4.7 Specifying which hits and alignments to show.....	15
4.8 Selecting the display format of the results.....	15
4.9 Using multiprocessor systems.....	16
4.10 Summary of command line options.....	17
5 The search results.....	20
5.1 Results in standard format.....	20
5.2 Results in HTML format.....	22
5.3 Results in tab-separated text format.....	22
5.4 Results in XML format.....	23
5.5 Calculation of scores and expect values.....	23
6 References.....	25
7 Legal information.....	26
7.1 Patents.....	26
7.2 Copyright.....	26
7.3 Trademarks.....	26
7.4 General license terms and general conditions of Sencel.....	27
7.5 OpenSSL license.....	33
7.6 SSLeay license.....	34
7.7 MPICH license.....	35

1 Introduction

PARALIGN is a program for rapid and sensitive similarity searches in genetic sequence databases, based on local sequence alignments.

The program takes advantage of various forms of parallelism to achieve a high speed with the most accurate algorithms for sequence alignment. Most importantly, it uses multimedia technology, also known as SIMD technology, to compare up to 16 symbols of the query with a symbol from the database sequence, in parallel. The code has been carefully tuned to various microprocessors for optimal performance. The actual microprocessor in the computer will be detected and the code optimized for that processor will be used. PARALIGN also features support for computers with multiple microprocessors (SMP), which speeds up the search process substantially. In addition, PARALIGN support computer clusters using MPI (message passing interface) communication.

PARALIGN can be used to search either a nucleotide or amino acid sequence database for sequences similar to (and possibly homologous to) a given nucleotide or amino acid query sequence. The comparison can be performed using either the rigorous Smith-Waterman optimal local sequence alignment algorithm or with the heuristic, though almost as sensitive, sequence alignment algorithm known as ParAlign.

PARALIGN can perform comparisons of either two nucleotide sequences or two amino acid sequences. The program can also translate a nucleotide query and/or nucleotide database sequence into an amino acid sequence on the fly while searching before the comparison is made. This allows one to search EST sequence data, for instance, using an amino acid query.

The alignment of the sequences is performed using a selectable symbol substitution score matrix and selectable gap open and extension penalty parameters. Alignments with an associated expect value (E-value) within a specified range will be reported.

PARALIGN handles multiple queries. Submitting several queries simultaneously will reduce overhead; as a consequence, results can be expected to appear faster than when submitting queries one at a time.

PARALIGN can produce HTML output (with sequence links) that can be viewed using a web browser, facilitating its use as a part of an intranet web search service. The program can also output the results in a tab-separated text file format or in XML format that can easily be parsed by another program (e.g. a Perl or shell script).

Please note that PARALIGN is mainly designed to find subtle similarities between remotely related sequences. If you are only looking for highly similar sequences there are other programs that will be faster.

The algorithms and implementations used in PARALIGN are described in detail in the scientific papers by Rognes and Seeberg (2000) and Rognes (2001). The web server based on PARALIGN is described by Sæbø *et al.* (2005). Please cite these publications when referring to PARALIGN in scientific publications.

For additional information about PARALIGN visit Sencel's website at <http://www.sencel.com/>

Online searches in the major public sequence databases with PARALIGN running on a powerful computer system is available at <http://www.paralign.org/>

Sencel Bioinformatics is grateful to NCBI for permission to redistribute the formatdb tool, the OpenSSL project and Eric Young for their cryptographic software, and the Argonne National Laboratory and Mississippi State University for the MPICH software.

2 Software installation

This chapter will explain how to get and install the PARALIGN software. If the software has already been installed, you can skip to the next chapter.

2.1 System requirements

This software requires one of the combinations of microprocessor and operating system shown in the table below.

Microprocessor	Operating system
Intel Pentium MMX, Pentium II, Pentium III, Pentium 4, Celeron, Xeon, Xeon MP, and similar processors AMD K6, K6-2, K6-III, Duron, Athlon, Athlon XP, Athlon MP, Athlon 4, and similar processors AMD Athlon 64, Opteron, and similar processors	Linux kernel 2.4 or later
Intel Itanium 2	Linux kernel 2.4 or later
Intel Itanium 2	HP-UX 11i version 2 or later
Motorola PowerPC G4 IBM PowerPC 970 (G5)	Mac OS X version 10.2 or later
HP Alpha EV6 or later	HP Tru64 Unix version 5.1 or later

The files included with PARALIGN does not consume much disk space, but disk space is needed to store the databases files to search. It is recommended that you also have enough RAM in in the computer to hold the entire sequence database in memory. The amount needed is roughly the size of the “.nsq” or “.psq” files of a NCBI formatted database.

2.2 Registering for access

Before the software can be downloaded and used, the following is needed:

- a username and password that grants access to the download area of Sencel's website
- a valid license file in order to run the software

To get a license file and a username and password, you need to register. To register, complete and submit the form on the following page:

<http://www.sencel.com/products/evaluate.html>

You should receive an email, normally within one business day, with the license file and the username and password.

If you are going to use PARALIGN for commercial purposes, in a for-profit company, or other non-academic organisations, you need a commercial license. By registering, commercial users may acquire an evaluation license valid for one month. To order PARALIGN, go to the following page:

<http://www.sencel.com/products/order.html>

2.3 Downloading the software

To download PARALIGN, follow these steps:

1. Go to the following page in your browser:
<http://www.sencel.com/products/download.html>
2. Select the appropriate software package for your computer system and click on the appropriate link.
3. Read the “General license terms and general conditions of Sencel” and indicate your acceptance by clicking on the appropriate link at the bottom of the page.
4. Enter, as requested, the username and password that was sent to you by email after registration.
5. The download will start and you will be prompted to select where to store the installation file.
6. Save the installation package in a directory of your choice (e.g. in your home directory).
7. Change the current directory to the directory where you saved the package.
8. Extract the files in the installation package by issuing a command similar to the following (depending on the name of the installation package file name):

```
gzip -dc paralign-4.0.0-linux-i386.tar.gz | tar xf -
```
9. The files will be extracted and put into a directory called `paralign-4.0.0-linux-i386` (or similar, depending on the package selected) in the current directory.

2.4 Installing the files

When the installation package has been extracted, the files below will be available in the directory.

File	Type	Contents
<code>paralign</code>	executable	The standard version of the PARALIGN program
<code>mparalign</code>	executable	A special version of PARALIGN for multi-processor systems
<code>formatdb</code>	executable	A program for database preparation
<code>LICENSE</code>	text file	The software license for PARALIGN
<code>README</code>	text file	The latest information about PARALIGN
<code>formatdb.html</code> / <code>formatdb.txt</code>	HTML document / text file	Information about how to use the NCBI formatdb program
<code>usersguide.pdf</code>	PDF document	This user's guide

You may copy the program files into a directory for executable binaries on your system, like the “`/usr/local/bin`” directory or your private “`~/bin`” directory. You may also want to include these directories in your `PATH` environment variable, if they are not already.

The `formatdb` software and documentation is provided by the National Center for Biotechnology Information (NCBI) in USA.

2.5 License management

After registering you should receive a license file from Sencel by email. The license file is a small text file that is usually called `sencel.lic` and looks like this:

```

-----BEGIN SENCEL LICENSE-----
PRODUCT PARALIGN
VERSION 1
SYSTEM linux-ia64
ID 12345
USER City University, Somewhere.
NOTICE For non-commercial use only.
COUNT 1
START 2005-01-01
EXPIRE 2005-12-31
SIGN CrHp5DD6MkMLabcdehijelm/fQpIPK30Gfh8765IPK30GfhHeVKt3ZGC3Co/e
SIGN cNrixx7hXqtdv99zqiNzMg==
-----END SENCEL LICENSE-----

```

The license file contains information about the product, the licensee, the number of licenses issued, and the license period. PARALIGN requires a valid license file to run. You must store a copy of the license file in a directory that is accessible when running the program. It is illegal to distribute the license file to others or to use the software on more computers than granted by the license, or in any other way not granted by the license. The full license text can be found in this document in the chapter Legal information.

When PARALIGN starts, it will look for the license file in the following four locations:

1. In the file specified by the “-g” option when starting the program. The full path and file name must be specified, for instance by a command like the one below. If there is no “-g” option specified, this step will be skipped.

```
paralign -g /opt/licenses/sencel.lic database query
```

2. In the file specified by the environment variable SENCEL_LICENSE_FILE. The full path and filename must be specified. If the environment variable SENCEL_LICENSE_FILE is not defined, this step will be skipped.
3. In the file “sencel.lic” in the users home directory (~/.sencel.lic), as specified by the environment variable HOME.
4. In the file “sencel.lic” in the current directory (./sencel.lic).

PARALIGN will use the first license file it finds that fits the format. If that file is invalid for any reason, the software will display an error message.

The recommended method of specifying the location of the license file is by setting the environment variable SENCEL_LICENSE_FILE. This variable should be set in an initialisation file that is automatically executed for all users at login (e.g. “/etc/profile”, “/etc/csh.login” or a similar file). In a shell like sh, bash, ksh or zsh, the environment variable may be set by a command like the one below:

```
export SENCEL_LICENSE_FILE=/usr/local/paralign/sencel.lic
```

In a shell like csh or tcsh, it may be set like this:

```
setenv SENCEL_LICENSE_FILE /usr/local/paralign/sencel.lic
```

If there is only one user of the software, the license file can simply be stored in the user's home directory.

3 Database preparation

PARALIGN can search databases prepared by the NCBI formatdb tool, which is also used to create databases for BLAST.

3.1 Simple databases

Sequences to be included in the database must first be formatted in the FASTA file format and then processed by the formatdb tool. The formatdb tool is provided by the National Center for Biotechnology Information (NCBI), USA. An executable version of formatdb and the accompanying file “formatdb.txt” or “formatdb.html” (depending on platform) is included with the PARALIGN software.

The file “formatdb.txt” or “formatdb.html” contains detailed information about the formatdb tool. However, creating a searchable protein database from the FASTA formatted file “ecoli” is as easy as issuing a command like this:

```
formatdb -i ecoli
```

The following command will give a brief overview of the formatdb options:

```
formatdb -
```

The formatdb tool will create a database in the NCBI database format. Each database consists of at least 3 files. These three files have file name extensions as shown in the table below. Only the base name, not including the dot and the three last letters should be specified on the command line when starting PARALIGN.

Sequence type	File contents	File name extension
Nucleotide	Indices	.nin
Nucleotide	Headers	.nhr
Nucleotide	Sequences	.nsq
Amino acid	Indices	.pin
Amino acid	Headers	.phr
Amino acid	Sequences	.psq

PARALIGN will work both with the database format (version 4) created by the current version of formatdb and the database format (database version 3) created by previous versions of formatdb.

In some cases, ParAlign may operate slowly when searching databases that consists of very long sequences (on the order of 100 000 symbols or more). It is recommended that such sequences are split into smaller, overlapping sequences.

Public sequence databases may be obtained from the following sources:

NCBI: <http://www.ncbi.nih.gov/>

EBI: <http://www.ebi.ac.uk/>

3.2 Database alias files

PARALIGN supports database alias files (ending in .nal or .pal), which allows for virtual databases consisting of multiple database files or subsets of other databases. The formatdb tool may be used to generate such files. PARALIGN accepts alias files similar to what NCBI BLAST accepts.

Alias files representing virtual databases comprised of several other databases

It is possible to query several databases at the same time by creating an alias file. The file name of an alias file containing protein sequences should have the .pal extension, while nucleotide databases should have the .nal extension. A simple alias-file may look like example 1 below.

```
# Example 1
# Alias file example, multiple databases
# filename: est.nal
#
TITLE Complete EST database
DBLIST /db/ncbi/est.00 /db/ncbi/est.01 /db/ncbi/est.02 /db/ncbi/est.03
```

An alias file may include references to other alias files on the DBLIST line.

Alias files representing virtual databases comprised of a subset of other databases

It is also possible to create virtual databases that comprise subsets of other databases. Such subsets may be represented by (1) a text file of gi numbers, (2) a binary file of gi numbers, or (3) a mask file. Alias files and binary gi files may be created using the formatdb tool.

The subset of sequences that should be included in the database may be specified with a text file containing a list of gi numbers, as in the alias file given in example 2 below.

```
# Example 2
# Alias file example, subset with gi list as text file
# filename: example2.pal
#
TITLE      Example2
DBLIST     nr
GILIST     gi.txt
LENGTH    748973
NSEQ      4893
```

An alias file similar to the one above may be created using this command:

```
formatdb -t "Example2" -i nr -L example2 -F gi.txt
```

In this case the alias file refers to a plain text file (gi.txt) containing a list of gi numbers for the sequences that should be included in the subset. The file should simply contain one gi number per line.

Alternatively, the gi numbers may be present in a binary file. Using binary gi files increases the searching speed. Binary gi files can be converted from text gi files using formatdb like this:

```
formatdb -F gi.txt -B gi.bin
```

This will create a binary gi file called “gi.bin” based on the list of gi numbers in the text file “gi.txt”.

```
# Example 3
# Alias file example, subset with binary gi list
# filename: example3.pal
#
TITLE      Example3
DBLIST     nr
GILIST     gi.bin
LENGTH    748973
NSEQ      4893
```

To create an alias file as above that refers to a binary gi file, use a command like this:

```
formatdb -t "Example3" -i nr -L example3 -F gi.bin
```

A third possibility is to use a mask file. Example 3 below demonstrates the use of a mask file. The alias file must refer to a mask file (e.g. swissprot.00.msk) that contains a binary mask of which sequences belong to the subset. Mask files are created by the NCBI genmask tool.

```
# Alias file generated by genmask
# Date created: Wed Feb  8 22:57:05 2006
#
TITLE      Non-redundant SwissProt sequences
DBLIST     nr.00
OIDLIST    swissprot.00.msk
LENGTH     71129662
NSEQ       191138
MAXOID     2056815
MEMB BIT 1
# end of the file
```

4 Searching

The PARALIGN program is normally started by typing “paralign” on the command line, followed by the base name of the sequence database to be searched and the name of the file containing the query sequence. Additional options to the program may be specified anywhere on the command line. The general syntax is:

```
paralign [options] database query
```

All options must start with a dash (-) followed by a letter specifying the behavior to change. Some options must be followed by a space and an argument that should be either an integer (for options with the letters “bcdpqrsuv”), a floating-point number (for “kl”) or a string (for “fgim”).

As an example, consider the following command:

```
paralign -n nt query.fsa
```

This will start a search of the nucleotide database called “nt” with a nucleotide query sequence called “query.fsa”. A direct nucleotide versus nucleotide sequence comparison will be carried out as indicated by the “-n” option.

If the program detects a problem with the options specified, an error message will be displayed.

To see a list of options to the program with a brief explanation, specify the “-h” option.

4.1 Selecting search algorithm

By default, the program uses the heuristic ParAlign search algorithm. The Smith-Waterman algorithm may be selected using the “-y” option. Searches using the ParAlign algorithm are much faster than with the Smith-Waterman algorithm, but the results cannot be guaranteed to be identical, although they usually are.

4.2 Selecting sequence types, translation and strands to search

By default, PARALIGN expects an amino acid query sequence and an amino acid sequence database, and will make direct comparisons of the sequences.

Using the “-n”, “-t”, “-x” and “-s” options this behaviour may be changed so that PARALIGN will use a nucleotide query sequence and/or a nucleotide sequence database, either of which will be automatically translated into amino acid sequences using the standard genetic code prior to comparison.

By specifying the “-n” option, PARALIGN will perform expect both a nucleotide query sequence and a nucleotide sequence database, and will perform a direct comparison of these nucleotide sequences, plus a comparison of the query sequence with the complementary database sequence.

By specifying the “-t” option, PARALIGN will expect a nucleotide sequence database that will be translated into all six possible reading frames before comparison with the query sequence. By specifying the “-x” option, PARALIGN will expect a nucleotide query sequence that will be translated into all six possible reading frames before comparison with the sequences in the database. The “-t” and “-x” options may also be combined.

When the query is a nucleotide sequence, the search may be performed using only on the upper (forward) strand of the query sequence by specifying the “-s 1” option or only on the lower (reverse complement) strand by the “-s 2” option. By default both strands are used (as specified by option “-s 3”). The table below shows all combinations of these options.

In addition, PARALIGN supports the highly experimental “-o” option, that is designed to find RNA sequences that will hybridize. When this option is specified, PARALIGN will assume that

both the query and the database sequences are RNA sequences and compare them using a special RNA-RNA hybridization score matrix. Otherwise, the “-o” option is similar to the “-n” option.

Options	Query	Database	Comparison	Contexts	Comment
(none)	amino acid	amino acid	amino acid	1	
-t	amino acid	nucleotide	amino acid	6	
-n	nucleotide	nucleotide	nucleotide	2	Both strands of query
-n -s 1	nucleotide	nucleotide	nucleotide	1	Only upper strand of query
-n -s 2	nucleotide	nucleotide	nucleotide	1	Only lower strand of query
-x	nucleotide	amino acid	amino acid	6	Both strands of query
-x -s 1	nucleotide	amino acid	amino acid	3	Only upper strand of query
-x -s 2	nucleotide	amino acid	amino acid	3	Only lower strand of query
-t -x	nucleotide	nucleotide	amino acid	36	Both strands of query
-t -x -s 1	nucleotide	nucleotide	amino acid	18	Only upper strand of query
-t -x -s 2	nucleotide	nucleotide	amino acid	18	Only lower strand of query

4.3 Specifying the sequence database

The first argument on the command line (apart from the options) must be the full path and base name of the sequence database to search. By default, the program expects an amino acid sequence database, but if the “-n” or the “-t” option is specified, a nucleotide sequence database is expected. The database must be pre-formatted in the NCBI format using the `formatdb` tool, as described in chapter 3.

Databases representing multiple databases or subsets of other databases may be searched using alias files as described in section 3.2. When PARALIGN is started, the base name of the alias file should be used as the database name, like this:

```
paralign protein-alias-file query
```

or, for a .nal-file

```
paralign -n nucleotide-alias-file query
```

4.4 Specifying a subset of the database to search

By default, PARALIGN will compare the query sequence with all sequences in the database. However, it is possible to limit the search to a given range of sequences in the database using the “-u” and “-v” options. The number of the first sequence to search is indicated with the “-u” option, while the index of the last sequence to search is indicated with the “-v” option. For instance, to search only the first 1000 sequences in a database called “ecoli”, use a command like this:

```
paralign -u 1 -v 1000 ecoli query.fsa
```

The search may also be limited to a subset of sequences in the database by specifying a list of taxid's or gi numbers. Using the “-G” option, a plain text file with a list of gi numbers (one number per line) or a binary gi file may be specified. See section 3.2 for how to create binary gi files. Only sequences with gi numbers in this list will then be included in the search. Example:

```
paralign -G gpcr.gi nr query.fsa
```

In this example, the file “gpcr.gi” could contain gi numbers for all the G-protein Coupled Receptor (GPCR) proteins.

Alternatively, only sequences from certain organisms may be searched. Using the “-T” option, a plain text file with a list of taxid numbers (taxonomy identifiers) (one number per line) may be specified. Only sequences from the organisms with taxid numbers on this list will then be included in the search. The taxid numbers specified must match exactly to the taxid numbers specified in the database. Not all databases contain the taxid information. One cannot use taxid numbers that represent several organisms, e.g. all primates (taxid 9443), in stead you'll have to list all the individual taxid numbers for organisms that belong to the group. For more information on taxid numbers, see www.ncbi.nlm.nih.gov/Taxonomy. Example:

```
paralign -T primates.tax nr query.fsa
```

Here, the file “primates.tax” could contain all taxid numbers for primate organisms.

4.5 Specifying the query sequence

The second argument on the command line (apart from the options) must be the full pathname and filename of a file containing the query sequence or sequences. By default, the program expects an amino acid query sequence, but if the “-n” or the “-x” option is specified, a nucleotide sequence query will be expected. Illegal symbols in the sequence will be ignored.

The query sequence file must be in one of the following file formats:

- Plain text format
- FASTA format
- EMBL/SWISSPROT flat file format
- GenBank/GenPept flat file format

Several sequences may be included in the query file. These sequences may be in either of these formats: FASTA, EMBL/SWISSPROT, GenBank/GenPept.

4.6 Specifying alignment scoring parameters

Score matrices and gap penalties may be specified using the “-m”, “-q” and “-r” options.

By default, the program performs alignment score computations using the BLOSUM62 amino acid substitution score matrix. In addition, a gap open penalty of 11 and a gap extension penalty of 1 are used, resulting in gap penalties of $11+k$, where k is the length (in amino acids) of the gap.

If nucleotide comparison has been selected using the “-n” option, the NT nucleotide substitution score matrix will be used by default. The default gap open penalty with this matrix is 5 and the default gap extension penalty is 2, resulting in gap penalties of $5+2k$, where k is the length (in nucleotides) of the gap.

Because the calculation of statistical parameters, like the expect value, is based on pre-computed constants, only certain combinations of matrices and gap penalty parameters are allowed. The table below shows the allowed combinations.

Matrix	Allowed gap penalties
BLOSUM45	14+2k
BLOSUM62	4+5k, 5+3k, 5+4k, 5+6k, 6+2k, 6+3k, 6+4k, 6+5k, 6+6k, 6+7k, 7+2k, 7+3k, 7+4k, 7+5k, 7+6k, 7+7k, 7+8k 8+k, 8+2k, 8+3k, 8+4k, 8+5k, 9+k, 9+2k, 9+3k, 9+4k, 10+k, 10+2k, 10+3k, 11+k , 11+2k, 11+3k, 12+k
BLOSUM80	6+2k, 7+k, 7+2k, 8+k, 8+2k, 9+k, 9+2k, 10+k, 10+2k, 11+k, 11+2k, 13+2k
PAM30	9+k
PAM70	10+k
NT	1+k, 2+k, 3+k, 5+2k
NT2	10+10k, 12+4k
RNA	5+2k, 8+2k

4.7 Specifying which hits and alignments to show

Normally, a list of up to 100 of the best matching database sequences having an expect value in the default range of 0 to 10 will be shown. The alignment with the query sequence of up to 100 of these sequences will also be shown. If there are several hits within the same database sequence, only the single best alignment within the same database sequence will be shown by default.

Using the “-l” option, the upper expect value threshold (default 10) may be adjusted in order to display fewer or more hits.

The “-k” option may be used to turn off display of “obvious” hits and alignments by specifying the lower expect value threshold (default 0).

The maximum number of hits to be shown may be specified by the “-b” option (default 100).

Use the “-a” option to completely disable the display of alignments, or use the “-c” option to specify the maximum number of alignments to show (default 100).

If suboptimal alignments are of interest, the “-d” option can be set to the maximum number of suboptimal alignments wanted, e.g. 10 (default 1). However, only alignments with expect values better (lower) than the limit set by the “-l” option will be shown.

4.8 Selecting the display format of the results

The program's result output is by default displayed in a plain text format suitable for display using a fixed-width font on a terminal. The “-z” option may be used to get an output in HTML format with sequence links into NCBI Entrez. This format is suitable for viewing in a web browser.

The results of the search can also be output in a plain text format that is easy to parse by another program (e.g. a Perl or shell script), by specifying the “-w” option. Alternatively, PARALIGN can

output the results in XML format, if the option “-e” is specified. See section 5.3 for more information about this format.

4.9 Using multiprocessor systems

PARALIGN has been designed for efficient utilization of shared memory symmetric multiprocessor systems (SMP). In such cases, the number of processors to be used can be specified with the “-p” option. The number specified is actually the number of working threads to be used. The default value is 1. For best performance in general, specify a number equal to or lower than the number of microprocessors installed in the computer.

PARALIGN also has support for multiprocessor systems with distributed memory, e.g. clusters, that are communicating using the Message Passing Interface (MPI) standard. The “mpalign” executable should be used on such systems. It is required that the program file, the query sequence file and the database files are available on all nodes and can be accessed using the same pathname. The database will be divided into parts that are processed by each node.

Running the MPI version of PARALIGN requires that an MPI implementation is installed on the computers. The free MPI implementation MPICH may be obtained from the following location:

<http://www-unix.mcs.anl.gov/mpi/mpich/>

On MPI systems, PARALIGN can be started with the following command line (to use 4 processors):

```
mpirun -np 4 mparalign [options] database query
```

The “-p” option may also be used in MPI systems. To run PARALIGN on a cluster with 4 nodes with 2 processors each, for a total of 8 processors, PARALIGN may be started like this:

```
mpirun -np 4 mparalign -p 2 [options] database query
```

4.10 Summary of command line options

- a** **Suppress the computation and display of alignments**
If this option is specified, no alignments will be computed or displayed. The default is to compute and display alignments.
- b #** **Maximum number of hits to show**
Minimum value: 1. Default value: 100
- c #** **Maximum number of alignments to show**
Minimum value: 0. Default value: 100
- d #** **Maximum number of suboptimal alignments to show**
Minimum value: 1. Default value: 1
- e** **Produce XML output**
Using this option, the results of the search will be output in an XML format that is easy to parse by another program (e.g. a Perl, PHP or shell script). See section 5.4 for more information about this format.
- G path** **Specify the name of a file containing gi numbers of sequences to search**
The full path name of a file containing a list of gi numbers may be specified. The file may be a plain text file or a binary file. If this option is used, only sequences with the gi numbers will be searched.
- g path** **Specify license file name**
The full path name of the license file may be specified either by this command line option or by the environment variable `SENCEL_LICENSE_FILE`. If none of these are specified, the user's home directory and the current directory will be searched for a file called "sencel.lic". See chapter 2 for more information about the license file.
- h** **Display help text with options summary**
- k #.#** **Minimum expect value of hits to show**
Hits and alignments with an expect value lower than the specified value will not be shown. The value must be smaller than the value specified by the "-l" option.
Allowed range: 0.0 – infinity. Default value: 0.0
- l #.#** **Maximum expect value of hits to show**
Only hits and alignments with an expect value lower than the specified value will be displayed. The value must be larger than the value specified by the "-k" option.
Argument type: floating point. Allowed range: 0.0 – infinity. Default value: 10.0
- m matrix** **Specify substitution score matrix**
Allowed values for amino acid sequence comparisons: BLOSUM45, BLOSUM62, BLOSUM80, PAM30, PAM70. Allowed values for nucleotide sequence comparisons: NT, NT2. The NT matrix gives a score of +1 for matching nucleotides and -3 for mismatches. The NT2 matrix is a matrix that gives a score of +5 for matching nucleotides and -4 for mismatches.
- n** **Nucleotide sequence comparison**
When this option is present the program will expect a nucleotide query sequence and a nucleotide sequence database. The program will compare the query sequence both with the direct database sequence and with the complementary database sequence. This option may not be used in combination with the "o", "-t" or "-x" options.
- o** **RNA-RNA hybridization**
Warning: this feature is highly experimental and not supported. It is designed to be used to find RNA sequences that will hybridize. When this option is present the program will expect an RNA query sequence and an RNA database sequence, and compare them

using a special score matrix where G:U basepairs are given a positive score in addition to the ordinary basepairs. In this matrix the ordinary A:U and G:C pairs will be scored +5, the special G:U pairs will be scored +2, and mismatches will be scored -3. The program will compare the query sequence both with the direct database sequence and with the complementary database sequence. The query sequence must be reversed compared to a normal query sequence. E-values and statistics are highly unreliable. This option may not be used in combination with the “-n”, “-t” or “-x” options. Otherwise, the “-o” option is similar to the “-n” option.

- p #** **Number of processors to use**
Argument type: integer. Allowed values: 1 to 64. Default value: 1.
- q #** **Gap opening penalty**
Argument type: integer. Allowed values: 0 - 100. Default value: 11 for amino acid sequence comparisons, 5 for nucleotide sequence comparisons.
- r #** **Gap extension penalty**
Argument type: integer. Allowed values: 0 - 100. Default value: 1 for amino acid sequence comparisons, 2 for nucleotide sequence comparisons.
- s #** **Which strands of the query sequence to search**
A value of 1 will force PARALIGN only to search the upper (forward) strand of the query sequence, while a value of 2 will make the program only to search using the lower (reverse complementary) strand. A value of 3 This option is ignored when searching with an amino acid query sequence.
Argument type: integer. Allowed values: 1, 2 or 3. Default value: 3 (both).
- T path** **Specify the name of a file containing taxid numbers of sequences to search**
The full path name of a plain text file containing a list of taxid numbers may be specified. If this option is used, only sequences from the specified organisms will be searched.
- t** **Translate database sequence**
When this option is present, the program will expect a nucleotide sequence database and will translate each database sequence into the six possible amino acid sequences that will subsequently be aligned with the query sequence(s). This option may be used in combination with the “-x” option, but cannot be used in combination with the “-n” or “-o” options.
- u #** **First database sequence to search**
Argument type: integer. Allowed argument range: 1 to the number of database sequences. The number must be equal to or smaller than the value specified with the “-v” option. This option can not be used in combination with -G or -T. Default: 1.
- v #** **Last database sequence to search**
Argument type: integer. Allowed argument range: 1 to the number of database sequences. The number must be equal to or larger than the value specified with the “-u” option. This option can not be used in combination with -G or -T. By default equal to the number of database sequences.
- w** **Produce tab-separated text output**
Using this option, the results of the search will be output in a plain text format that is easy to parse by another program (e.g. a Perl or shell script). See section 5.3 for more information about this format.
- x** **Translate query sequence**
When this option is present the program will expect a nucleotide query sequence and will translate the query sequence into the six possible amino acid sequences that will subsequently be aligned with the database sequence(s). This option may be used in

combination with the “-t” option, but cannot be used in combination with the “-n” or “-o” options.

-y Use the Smith-Waterman algorithm

When this option is present the Smith-Waterman algorithm will be used instead of the default ParAlign algorithm for computing the scores of the sequence alignments while searching. This will increase the time taken for the search by one order of magnitude, but may in a few cases increase sensitivity.

-z Produce HTML output

When this option is present the output will be HTML formatted text suitable for display by a web browser. Links will be generated to NCBI Entrez entries for the sequences.

5 The search results

The results of a search can be presented in different formats, suitable for viewing either on a terminal or in a browser window. The results can also be output in a format that can easily be parsed by other programs for further processing.

5.1 Results in standard format

The standard output format is divided into several sections, as shown in the example. First, some information about the program will be displayed, followed by information about the query, the database and the search options in effect. Then information about the search will be displayed. Once the search is finished, a list of the resulting hits will be displayed. Finally the alignments will be displayed when they have been computed. The output will contain the following elements:

HEADER	The program name, version, copyright notice, license information and the type of microprocessor detected.
QUERY	The query file name, the sequence length, the sequence, and a short description.
DATABASE	The database file name, a short description, database date, the total length of all sequences, the number of entries, and the length of the longest sequence.
OPTIONS	The substitution matrix name, gap penalties, the number of threads used, the algorithm used, and the expect range.
SEARCH	The search start time, finish time, elapsed time, speed, and number of Smith-Waterman alignments computed.
RESULTS	First, the total number of hits, the number of obvious hits (with a too low e-value), the number of shown hits, and the number of shown aligned hits is listed. Then, for each hit the following is shown: the identifiers of the sequence (gi, accession number etc), a short description of the sequence, strand of query sequence (only when the “-x” option is specified), frame of query sequence (only when the “-x” option is specified), strand of database sequence (only when the “-n”, “-o” or “-t” option is specified), frame of the database sequence (only when the “-t” option is specified), raw alignment score, and the expect value.
ALIGNMENT	For each alignment all of the information in the RESULTS section will be repeated, plus the number and percentage of identical nucleotides (only with “-n” or “-o” option), the number and percentage of identical amino acids (not with “-n” or “-o” options), the number and percentage of similar amino acids (not with “-n” or “-o” options), number of gaps and indels, the length of the alignment, and the actual alignment itself, including starting and ending positions in both sequences (numbers refer to original sequence before translation, if any), and the actual sequence of the aligned parts of the sequences (possibly translated).

An example is shown on the following pages of how the results may look in the standard format by starting PARALIGN with the following command (output has been edited for brevity):

```
paralign /db/swissprot/sp40 end3_ecoli.f -c 2
```

If multiple queries are submitted, the header will be output before the first query. All other queries will start with the word “QUERY”.

PARALIGN 4.0.0 Parallel Sequence Alignment and Database Searching
Software Copyright (C) 1998-2005 Sencel Bioinformatics AS. All rights reserved.
Contains patent pending technology.
Intel Pentium III microprocessor (MMX,SSE) detected.
Licensed to University of Oslo, Norway.
For non-commercial use only. License expires January 1, 2010.

QUERY

File name: end3_ecoli.f
Sequence type: Amino acid
Description: sp|P20625|END3_ECOLI ENDONUCLEASE III (EC 4.2.99.18) (DNA-(A
PURINIC OR APYRIMIDINIC SITE) LYASE).
Length: 211
Sequence: MNKAKRLEILTRLRENNPHPTTELNFSPPFELLIIVLLSAQATDVSVNKATAKLYPVANT
PAAMLELGVGKTYIKTIGLYNSKAENI IKTCRILLEQHNGEVPEDRAALEALPGVGRK
TANVVLNTAFGWPTIAVDTHIFRVCNRTQFAPGKNVEQVEEKLKVVPAEFKVDCHHWLI
LHGRYTCIARKPRCGSCIIEDLCEYKEKVDI

DATABASE

File name: /db/swissprot/sp40
Sequence type: Amino acid
Description: SWISS-PROT Rel 40
Version: 4
Date: Jan 27, 2005 9:22 PM
Residues: 37315215
Sequences: 101602
Longest: 6669

OPTIONS

Algorithm: ParAlign
Database range: 1 to 101602
Score matrix: BLOSUM62
Gap penalties: 11 (open), 1 (extension)
Ungapped L,K,H: 0.32, 0.136, 0.404
Gapped L,K,H: 0.27, 0.047, 0.23
Expect range: 0 to 10
Display limits: 100 hits, 2 alignments and 1 subalignments
Threads: 1

SEARCH

Started: Fri Oct 22 14:23:05 2004
Searching... (please wait)
Completed: Fri Oct 22 14:23:10 2004
Elapsed time: 4.93s
Speed: 1597.1 million cells per second
SW alignments: 636 (0.63%)

RESULTS

Total hits: 33
Obvious hits: 0
Shown hits: 33
Alignments: 2

SEQUENCE	SCORE	E
sp P20625 END3_ECOLI Endonuclease III (EC 4.2.99.18) (DNA-(apurinic	1101	0
sp P44319 END3_HAEIN Endonuclease III (EC 4.2.99.18) (DNA-(apurinic	930	0
sp P57219 END3_BUCAI Endonuclease III (EC 4.2.99.18) (DNA-(apurinic	658	1e-69
sp O05956 END3_RICPR Endonuclease III (EC 4.2.99.18) (DNA-(apurinic	549	6e-57
sp O83754 END3_TREPA Endonuclease III (EC 4.2.99.18) (DNA-(apurinic	545	2e-56
sp P39788 END3_BACSU Probable endonuclease III (EC 4.2.99.18) (DNA-	431	5e-43
sp P46303 UVEN_MICLU Ultraviolet N-glycosylase/AP lyase (UV endonuc	349	2e-33
sp Q9CB92 END3_MYCLE Endonuclease III (EC 4.2.99.18) (DNA-(apurinic	340	2e-32
sp P73715 END3_SYNY3 Endonuclease III (EC 4.2.99.18) (DNA-(apurinic	339	3e-32
sp O69642 END3_MYCTU Endonuclease III (EC 4.2.99.18) (DNA-(apurinic	337	5e-32
sp Q9WYK0 END3_THEMA Endonuclease III (EC 4.2.99.18) (DNA-(apurinic	332	2e-31
sp Q58030 Y613_METJA Putative endonuclease MJ0613.	310	1e-28
sp P54137 END3_CAEEL Probable endonuclease III homolog (EC 4.2.99.1	244	5e-21
sp P29588 GTMR_METTF Possible G-T mismatches repair enzyme (EC 3.2.	201	4e-16
sp Q09907 END3_SCHPO Endonuclease III homolog (EC 4.2.99.18) (DNA-(192	8e-15

Information given for all hits:

- query sequence description header
- database sequence description header
- database sequence length
- strand (+ or -) for the query sequence [only included if the “-n” option is in effect]
- strand (+ or -) and frame for the query sequence [only included if the “-x” option is in effect]
- strand (+ or -) and frame for the database sequence [only included if the “-t” option is in effect]
- alignment score
- expect value

Information given only for aligned hits:

- length of alignment
- number of identical symbols in alignment
- number of similar symbols in alignment (corresponding to a positive value in the score matrix) [not included if the “-n” option is in effect]
- number of gaps
- number of indels
- alignment starting position in query sequence (1-based)
- alignment ending position in query sequence (1-based)
- alignment starting position in database sequence (1-based)
- alignment ending position in database sequence (1-based)

5.4 Results in XML format

The XML format that is specified by the “-e” option is similar to the standard format, but the text is wrapped in XML tags and some links are added. This format is suitable for parsing by other programs, e.g. shell, Perl scripts, or PHP scripts. Note that the PARALIGN XML format is still experimental, and may change in the near future.

5.5 Calculation of scores and expect values

PARALIGN calculates and shows score values and expect values for each of the hits found. The final score values are calculated according to the Smith-Waterman local sequence alignment algorithm (Smith and Waterman 1980).

The degree of similarity between the query and database sequence is indicated by the score of the alignment of the sequences. However, the expect value (E-value) is generally a better indicator of the significance of match than the raw score. The E-value indicates the number of matches expected to be found just by chance having an alignment score at least as good the one at hand. If a match has an E-value of 100 it is usually statistically insignificant, as one would expect 100 at least equally good matches to appear by chance. An E-value of 0.001, however, is usually significant, as such a match would appear by chance only once in a thousand searches. Anyway, the biological significance of a match is often quite different from the statistical significance. In addition, the accuracy of the statistical significance presented by the E-value may vary.

The E-value, as calculated by PARALIGN, is dependent on the raw alignment score (S), the length of the query sequence (M), the length of the database sequence (N), the number of database sequences (C), as well as the scoring system used (substitution matrix and gap penalties). The E-value is calculated according to the following formula:

$$E = C K M^2 N^2 \exp(-\lambda S)$$

Here, M' and N' are the lengths of the query and database sequences subject to a length-correction term, d , and are calculated as follows:

$$M' = \max(1, M - d)$$

$$N' = \max(1, N - d)$$

$$d = \ln(KMN) / H$$

K , λ and H are scoring system dependent constants. The natural logarithm and exponential functions are used.

This method of calculating E-values is similar to how BLAST and FASTA computes E-values, but it does not take sequence composition into account. For more information on calculation of expect-values, see the paper by Altschul and Gish (1996).

As an example, we will calculate the E-value of a protein query sequence of 216 amino acids matching a database protein sequence of 389 amino acids with an alignment score of 169. The database contains 186 234 sequences. Default scoring parameters are used; the substitution matrix is BLOSUM62, the gap open penalty is 10, and the gap extension penalty is 1. For these scoring parameters, the following pre-computed constants are used: $L=0.27$, $K=0.047$ and $H=0.23$. The formulas above give an E-value of $8 \cdot 10^{-12}$, which is shown as 8e-12.

6 References

- Altschul SF and Gish W (1996) *Local alignment statistics*. Methods in enzymology, 266, 460-480.
- Rognes T. (2001) *ParAlign: a parallel sequence alignment algorithm for rapid and sensitive database searches*. Nucleic Acids Research, 29 (7), 1647-1652.
- Rognes T. and Seeberg E. (2000) *Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors*. Bioinformatics, 16 (8), 699-706.
- Smith TF and Waterman MS (1981) *Identification of common molecular sub-sequences*. Journal of Molecular Biology, 147 (1) 195-197.
- Sæbø PE, Andersen SM, Myrseth J, Lærdahl JK, Rognes T (2005) *PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology*. Nucleic Acids Research, 33 (Suppl 2), W535-W538.

7 Legal information

7.1 Patents

The PARALIGN software contains an implementation of the following patent:

- Norwegian patent no 324517:
“Bestemmelse av optimal skåre for lokal sekvenssammenstilling”
(Granted 5 November 2007.)

The software also contains implementations of patent pending sequence alignment methods, as described in the following and related patent applications:

- PCT NO/01/00394 (US patent application no 20040024536):
Determination of optimal local sequence alignment similarity score
- PCT NO/02/00041 (US patent application no 20040098203):
Parallel biological sequence alignments

7.2 Copyright

The PARALIGN software and documentation is copyright © 1998-2008 Sencel Bioinformatics AS. All rights reserved.

7.3 Trademarks

Intel, Pentium, MMX, Celeron, Xeon and Itanium are trademarks or registered trademarks of Intel Corporation, Inc.

AMD, K6, K6-2, K6-III, Athlon, Duron, Opteron and 3DNow! are trademarks or registered trademarks of Advanced Micro Devices, Inc.

AltiVec is a trademark of Motorola, Inc.

HP and Tru64 are trademarks of The Hewlett-Packard Company.

IBM and PowerPC are trademarks of International Business Machines Corporation.

Sun and Sun Microsystems are trademarks or registered trademarks of Sun Microsystems, Inc. UltraSPARC is a trademark of SPARC International, Inc.

Adobe is a trademark of Adobe Systems Incorporated.

UNIX is a registered trademark of The Open Group.

Linux is a registered trademark of Linus Torvalds.

Sencel and Sencel Bioinformatics are trademarks, and PARALIGN is a registered trademark of Sencel Bioinformatics AS.

All other trademarks and registered trademarks are owned by their respective companies.

7.4 General license terms and general conditions of Sencel

1. DEFINITIONS

Install	The installation of the Software into a permanent memory on the System, e.g. hard disk, CD ROM or other storage device.
License Fee	The license fee as specified in the License Agreement.
Licensed Software	Software and/or associated media and materials as specified in the License Agreement.
Sencel	Sencel Bioinformatics AS (Org.no. 983 558 526), Oslo, Norway, a Norwegian limited company.
Software	Software developed or marketed by Sencel.
System	The system as specified in the License Agreement.
The Agreement	The License Agreement and these General Terms and Conditions jointly.
Use	The copying, transmitting or loading of the Software into the temporary memory (RAM) of a computer or the utilization of Software in a CPU.

2. SCOPE OF THE AGREEMENT

Sencel grants Customer a limited, non-exclusive and non-transferable right to Use the Licensed Software on the System for the Customer's internal use in its organisation, subject to the License Agreement concluded between the parties, and these General Terms and Conditions.

3. GRANT OF THE SOFTWARE LICENSE ENTITLES THE LICENSEE TO:

- (a) download one copy of the Licensed Software for Use on the System;
- (b) Install the number of copies of the Licensed Software on the System as is necessary to Use the Licensed Software in accordance with the Agreement;
- (c) Use the Licensed Software on the System on the number of CPUs in concurrent use as specified in the License Agreement;
- (d) download, Install and use the instructional and/or operational manuals as is necessary to Install and Use the Licensed Software;
- (e) support of the Licensed Software at such support charges as specified by Sencel at any time.

4. THE LICENSEE UNDERTAKES:

(a) not to perform any of the acts referred to in this sub-section (a) except to the extent and only to the extent permitted by the applicable law to the Licensee as a lawful user (i.e. a party with an express right to use) of the Licensed Software and only then for the specific limited purposes stated in such applicable law or hereunder. The Licensee undertakes:

- (i) not to copy the Licensed Software other than as specified in Section 3;
- (ii) not to translate, adapt, vary, modify the Licensed Software;
- (iii) not to disassemble, decompile or reverse engineer the Licensed Software;
- (iv) not to remove or circumvent, or attempt removal or circumvention, of any technical device which may have been applied to protect the Licensed Software in any way and/or to prevent unauthorized copying or other unauthorized use, including but not limited to use of

the Licensed Software in excess of the agreed number of CPUs in concurrent use or any license period.

(b) to maintain accurate and up-to-date records of the number and location of all copies of the Licensed Software;

(c) to Install and Use Software only as explicitly permitted, and to supervise and control that such Install and Use is in accordance with the terms of the Agreement;

(d) to replace the current version of the Licensed Software with the upgraded version forthwith upon notification;

(e) not to remove or change any message, label or note on the Licensed Software, nor remove or change any indication concerning Intellectual Property from or in the Licensed Software or materials, including any indications concerning the confidential nature of the Licensed Software.

(f) to reproduce and include the copyright notice of Sencel on all and any copies, whether in whole or in part, in any form, including partial copies or modifications of the Licensed Software made herein;

(g) not to provide or otherwise make available the Licensed Software in whole or in part (including but not limited to program listings, object and source program listings, object code and source code), in any form to any person other than the Licensee's employees without prior written consent from Sencel;

(h) within fourteen (14) days after the date of termination or discontinuance of this Agreement for whatever reason, to return or destroy (as Sencel shall instruct) the Licensed Software and all copies, in whole and in part, in any form including partial copies or modifications of the Licensed Software received or downloaded from Sencel or made in connection with this Agreement and all documentation relating thereto and to furnish Sencel with a certificate, certifying that the same has been done, unless the Licensee has obtained Sencel's prior written authorization to retain one copy for archive purposes only.

5. LICENCE FEE

(a) Sencel shall upon the making available for download of the Licensed Software invoice the Licensee for the License Fee. The Licence Fee shall be due and payable by the Licensee to Sencel within fourteen (14) days after the date of invoice. Sencel may amend the License Fee from time to time by giving to the Licensee not less than thirty (30) days prior notice thereof.

(b) The Licensee may be subject to an additional charge if upgraded versions of the Licensed Software are made available for download. Such additional charge for upgrades shall become payable within thirty (30) days of availability unless the Licensee notifies Sencel that it will not use such upgrades within such period.

(c) Where the Licensee is paying annual Licence Fees, the Licensee may request Sencel to convert to a one-time payment Licence and Sencel shall, upon acceptance of such request, notify the Licensee of the appropriate payment to be made.

(d) An additional fee is payable for each CPU of the System upon which the Licensee wishes to Use the Licensed Software in excess of the number specified in the License Agreement. In the event that the System is inoperable or requires or is under repair, the Licensee shall report the same to Sencel for permission to Use the Licensed Software on a back-up system at no extra charge, until the System is operational.

(e) All charges and fees hereinbefore referred to are exclusive of and net of any taxes, duties or such other additional sums including, but without prejudice to the foregoing generality, value added or

purchase tax, excise tax, tax on sales, property or use, import or other duties, whether levied in respect of the License Agreement, the Licensed Software, its use or otherwise.

6. WARRANTY

(a) The Licensee acknowledges that Licensed Software in general is not error-free and agrees that the existence of such errors shall not constitute a breach of this Agreement.

(b) In the event that the Licensee discovers a material error which substantially affects the Licensee's Use of the same and notifies Sencel of the error within 90 days from the date of this Licence (the "warranty period") Sencel shall at its sole option either refund the Licence Fee or use all reasonable endeavours to correct by patch or new release (at its option) that part of the Licensed Software which does not so comply PROVIDED THAT such non-compliance has not been caused by any modification, variation or addition to the Licensed Software not performed by Sencel or caused by its incorrect use, abuse or corruption of the Licensed Software or by use of the Licensed Software with other Licensed Software or on equipment with which it is incompatible.

(c) To the extent permitted by the applicable law, Sencel disclaims all other warranties with respect to the Licensed Software, either express or implied, including but not limited to any implied warranties of merchantability or fitness for any particular purpose.

(d) Although Sencel does not warrant that the Licensed Software supplied hereunder shall be free from all known viruses it has used commercially reasonable efforts to check for the most commonly known viruses prior to packaging but the Licensee is solely responsible for virus scanning the Licensed Software.

7. SENCEL'S LIABILITY

(a) Sencel shall not be liable to the Licensee for any loss or damage whatsoever or howsoever caused arising directly or indirectly in connection with the License Agreement, the Licensed Software, its use, support or otherwise, except to the extent to which is unlawful to exclude such liability under the applicable law.

(b) Notwithstanding the generality of (a) above, Sencel expressly excludes liability for consequential loss or damage which may arise in respect of the Licensed Software, its use, the System or in respect of other equipment or property, or for loss of profit, business, revenue, goodwill or anticipated savings.

(c) In the event that any exclusion contained in this Agreement shall be held to be invalid for any reason and Sencel becomes liable for loss or damage that may lawfully be limited, such liability shall be limited to the annual Licence Fee or 50 % of the one-time Licence fee.

8. COPYRIGHT, PATENTS, TRADE MARKS AND OTHER INTELLECTUAL PROPERTY RIGHTS

(a) The Licensee acknowledges that the following is to be considered as the intellectual property of Sencel ("Intellectual Property"):

(i) The Software and any product, material or service relating thereto, including but not limited to source-code, object-code, all documentation and manuals, software design or other designs, data registers, databases, interface, algorithms, ideas, concepts, sketches, diagrams and technical data, which Sencel, any of its employees or any subcontractor, has conceived, originated, made, developed or used in connection with the Software, whether or not registered or protected by copyright or patent devised or acquired by Sencel;

(ii) Any trademark(s), trade name(s), logo(s) or product name(s) for which Sencel use and/or is the registered proprietor of, and which are related to the Software, including but not limited to "ParAlign"; and

(iii) "Sencel" and other company names used by Sencel, hereunder, but not limited to all names starting with "Sencel".

(b) The Licensee acknowledges that Sencel has and remains exclusive rights to any and all of the Intellectual Property, and the Licensee shall not during or at any time after the expiry or termination of the License Agreement in any way question or dispute these rights.

(c) The Licensee also acknowledges that it may only use Intellectual Property and other rights belonging to Sencel with the consent of Sencel and during the License Agreement and subject to instructions and guidelines set by Sencel at any time.

(d) The Licensee shall not during or after the expiry or termination of this Agreement, without the prior written consent of Sencel, use, adopt or apply for registration of any name, trade name, trading style, domain name or commercial designation that includes or is similar to or may be mistaken for the whole or any part of any of the Intellectual Property.

(f) The Licensee shall indemnify Sencel fully against all liabilities, costs and expenses that Sencel may incur as a result of work done in accordance with the Licensee's specifications involving infringement of any patent or other proprietary right.

9. CONFIDENTIAL INFORMATION

(a) Any information, data, drawings, specifications, documentation, software listings, source or object code, which Sencel may have imparted and may from time to time impart to the Licensee relating to the Licensed Software or this Agreement, is proprietary and confidential ("Confidential Information").

(b) The Licensee hereby agrees that it shall use Confidential Information solely in accordance with the provisions of the Agreement and that it shall not at any time during or after expiry or termination of the Agreement, disclose the same, whether directly or indirectly to any third party without Sencel's prior written consent.

(c) The Licensee further agrees that it shall not itself or through any subsidiary, agent or third party modify, vary, enhance, copy, sell, lease, license, sub-license or otherwise deal with the Licensed Software or any part or parts or variations, modifications, copies, releases, versions or enhancements thereof or have any Licensed Software or other program written or developed for it based on any Confidential Information.

(d) The Licensee is obligated to inform employees and other persons under the authority, control or supervision of the Licensee, of the Agreement. The Licensee shall ensure that employees, agents, vendors, contractual partners, and any other person under the authority, control or supervision of the Licensee, respects the Agreement. The Licensee shall immediately inform Sencel in writing if non-authorized use is discovered or if such use is suspected.

(e) The foregoing provisions shall not prevent the disclosure or use by the Licensee of any Confidential Information that is or hereafter, through no fault of the Licensee, becomes public knowledge or to the extent permitted by law.

10. INSPECTION

Sencel shall have the right at all reasonable times to inspect the premises of Licensee in order to determine and verify Licensee compliance with the Agreement.

11. FORCE MAJEURE

Sencel shall be under no liability to the Licensee in respect of anything which, apart from this provision, may constitute breach of this Agreement arising by reason of force majeure, namely,

circumstances beyond the control of Sencel which shall include (but shall not be limited to) acts of God, perils of the sea or air, fire, flood, drought, explosion, sabotage, accident, embargo, riot, civil commotion, including acts of local government and parliamentary authority; inability to supply the Licensed Software, materials or support; the breakdown of equipment and labour disputes of whatever nature and for whatever cause arising including (but without prejudice to the generality of the foregoing) work to rule, overtime bars, strikes and lockouts and whether between either of the parties hereto and any or all of its employees and/or any other employer and any or all of its employees and/or between any two or more groups of employees (and whether of either of the parties hereto or any other employer).

12. TERMINATION

(a) In addition to provisions for termination as herein provided, Sencel may by notice in writing to the Licensee terminate this Agreement, in whole or at Sencel's sole option with respect only to the support, if any of the following events shall occur, viz.:

(i) if the Licensee is in breach of any term, condition or provision of this Agreement or required by applicable law and fails to remedy such breach (if capable of remedy) within 30 days of having received written notice of such breach from Sencel;

(ii) if the Licensee, being a body corporate shall present a petition or have a petition presented by a creditor for its winding up, or shall convene a meeting to pass a resolution for voluntary winding up, or shall enter into any liquidation (other than for the purposes of a bona fide reconstruction or amalgamation) shall call a meeting of its creditors, or shall have a receiver of all or any of its undertakings or assets appointed, or shall be deemed by the relevant statutory provisions under the applicable law to be unable to pay its debts.

(b) Upon termination, the Licensee shall pay to Sencel all costs and expenses, including legal and other fees incurred and all arrears of fees, charges or other payments arising in respect of the Licensed Software or this Agreement or otherwise.

(c) Termination, howsoever or whenever occasioned shall be subject to any rights and remedies Sencel may have under this Agreement or in Law.

13. ASSIGNMENT

The Licensee shall not assign or otherwise transfer all or part of the Licensed Software or this Agreement without the prior written consent of Sencel.

14. WAIVER

Failure or neglect by either party to enforce at any time any of the provisions hereof shall not be construed nor shall be deemed to be a waiver of that party's rights hereunder nor in any way affect the validity of the whole or any part of this Agreement nor prejudice that party's rights to take subsequent action.

15. HEADINGS

The headings of the terms and conditions herein contained are inserted for convenience of reference only and are not intended to be part of or to affect the meaning or interpretation of any of these terms and conditions of this Agreement.

16. SEVERABILITY

In the event that any of the terms, conditions or provisions contained in this Agreement shall be determined by any competent authority to be invalid, unlawful or unenforceable to any extent, such

term, condition or provision shall to that extent be severed from the remaining terms, conditions and provisions which shall continue to be valid to the fullest extent permitted by law.

17. NOTICES

Any notice to be given by either party to the other must be in writing and be sent by to the address, mail address, e-mail address or facsimile number of the other party as appearing herein or such other address as such party may from time to time have communicated to the other in writing. Any such notice shall be considered to have been given at the time when actually delivered if delivered by hand, or upon the next working day following sending by facsimile or e-mail or in any other event within 4 days after it was mailed in the manner hereinbefore provided.

18. AMENDMENTS

Any variation to the License Agreement, these General Terms and Conditions, or any other conditions of Licensee, shall not apply unless accepted by Sencel in writing.

19. GOVERNING LAW. DISPUTES.

(a) All contracts made between Sencel and Licensee and these General Terms and Conditions shall be governed by and interpreted in accordance with Norwegian law.

(b) In case of a dispute between the parties which can not be solved by negotiations, the Licensee hereby submit for the benefit of Sencel/Distributor, to the jurisdiction of the Norwegian courts, with the Oslo City Court as due venue. Nevertheless, Sencel has the right to instigate legal proceedings in the country where the Customer has its place of residence or perform business activities.

7.5 OpenSSL license

This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit (<http://www.openssl.org/>). That software is subject to the license below.

```
OpenSSL License
-----

/* =====
 * Copyright (c) 1998-2003 The OpenSSL Project. All rights reserved.
 *
 * Redistribution and use in source and binary forms, with or without
 * modification, are permitted provided that the following conditions
 * are met:
 *
 * 1. Redistributions of source code must retain the above copyright
 * notice, this list of conditions and the following disclaimer.
 *
 * 2. Redistributions in binary form must reproduce the above copyright
 * notice, this list of conditions and the following disclaimer in
 * the documentation and/or other materials provided with the
 * distribution.
 *
 * 3. All advertising materials mentioning features or use of this
 * software must display the following acknowledgment:
 * "This product includes software developed by the OpenSSL Project
 * for use in the OpenSSL Toolkit. (http://www.openssl.org/)"
 *
 * 4. The names "OpenSSL Toolkit" and "OpenSSL Project" must not be used to
 * endorse or promote products derived from this software without
 * prior written permission. For written permission, please contact
 * openssl-core@openssl.org.
 *
 * 5. Products derived from this software may not be called "OpenSSL"
 * nor may "OpenSSL" appear in their names without prior written
 * permission of the OpenSSL Project.
 *
 * 6. Redistributions of any form whatsoever must retain the following
 * acknowledgment:
 * "This product includes software developed by the OpenSSL Project
 * for use in the OpenSSL Toolkit (http://www.openssl.org/)"
 *
 * THIS SOFTWARE IS PROVIDED BY THE OpenSSL PROJECT ``AS IS'' AND ANY
 * EXPRESSED OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
 * IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR
 * PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE OpenSSL PROJECT OR
 * ITS CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,
 * SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT
 * NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES;
 * LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION)
 * HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT,
 * STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE)
 * ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED
 * OF THE POSSIBILITY OF SUCH DAMAGE.
 * =====
 *
 * This product includes cryptographic software written by Eric Young
 * (eay@cryptsoft.com). This product includes software written by Tim
 * Hudson (tjh@cryptsoft.com).
 */
```

7.6 SSLeay license

This product includes cryptographic software written by Eric Young (ey@cryptsoft.com). That software is subject to the license below.

```
/* Copyright (C) 1995-1998 Eric Young (ey@cryptsoft.com)
 * All rights reserved.
 *
 * This package is an SSL implementation written
 * by Eric Young (ey@cryptsoft.com).
 * The implementation was written so as to conform with Netscapes SSL.
 *
 * This library is free for commercial and non-commercial use as long as
 * the following conditions are aheared to. The following conditions
 * apply to all code found in this distribution, be it the RC4, RSA,
 * lhash, DES, etc., code; not just the SSL code. The SSL documentation
 * included with this distribution is covered by the same copyright terms
 * except that the holder is Tim Hudson (tjh@cryptsoft.com).
 *
 * Copyright remains Eric Young's, and as such any Copyright notices in
 * the code are not to be removed.
 * If this package is used in a product, Eric Young should be given attribution
 * as the author of the parts of the library used.
 * This can be in the form of a textual message at program startup or
 * in documentation (online or textual) provided with the package.
 *
 * Redistribution and use in source and binary forms, with or without
 * modification, are permitted provided that the following conditions
 * are met:
 * 1. Redistributions of source code must retain the copyright
 * notice, this list of conditions and the following disclaimer.
 * 2. Redistributions in binary form must reproduce the above copyright
 * notice, this list of conditions and the following disclaimer in the
 * documentation and/or other materials provided with the distribution.
 * 3. All advertising materials mentioning features or use of this software
 * must display the following acknowledgement:
 * "This product includes cryptographic software written by
 * Eric Young (ey@cryptsoft.com)"
 * The word 'cryptographic' can be left out if the rouines from the library
 * being used are not cryptographic related :-).
 * 4. If you include any Windows specific code (or a derivative thereof) from
 * the apps directory (application code) you must include an acknowledgement:
 * "This product includes software written by Tim Hudson (tjh@cryptsoft.com)"
 *
 * THIS SOFTWARE IS PROVIDED BY ERIC YOUNG ``AS IS'' AND
 * ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
 * IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE
 * ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR OR CONTRIBUTORS BE LIABLE
 * FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL
 * DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS
 * OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION)
 * HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT
 * LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY
 * OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF
 * SUCH DAMAGE.
 *
 * The licence and distribution terms for any publically available version or
 * derivative of this code cannot be changed. i.e. this code cannot simply be
 * copied and put under another distribution licence
 * [including the GNU Public Licence.]
 */
```

7.7 MPICH license

This product includes the MPICH implementation of the Message-Passing Interface standard. That software is subject to the license below.

COPYRIGHT

The following is a notice of limited availability of the code, and disclaimer which must be included in the prologue of the code and in all source listings of the code.

Copyright Notice

+ 1993 University of Chicago
+ 1993 Mississippi State University

Permission is hereby granted to use, reproduce, prepare derivative works, and to redistribute to others. This software was authored by:

Argonne National Laboratory Group

W. Gropp: (630) 252-4318; FAX: (630) 252-5986; e-mail: gropp@mcs.anl.gov
E. Lusk: (630) 252-7852; FAX: (630) 252-5986; e-mail: lusk@mcs.anl.gov
Mathematics and Computer Science Division
Argonne National Laboratory, Argonne IL 60439

Mississippi State Group

N. Doss: (601) 325-2565; FAX: (601) 325-7692; e-mail: doss@erc.msstate.edu
A. Skjellum: (601) 325-8435; FAX: (601) 325-8997; e-mail: tony@erc.msstate.edu
Mississippi State University, Computer Science Department &
NSF Engineering Research Center for Computational Field Simulation
P.O. Box 6176, Mississippi State MS 39762

GOVERNMENT LICENSE

Portions of this material resulted from work developed under a U.S. Government Contract and are subject to the following license: the Government is granted for itself and others acting on its behalf a paid-up, nonexclusive, irrevocable worldwide license in this computer software to reproduce, prepare derivative works, and perform publicly and display publicly.

DISCLAIMER

This computer code material was prepared, in part, as an account of work sponsored by an agency of the United States Government. Neither the United States, nor the University of Chicago, nor Mississippi State University, nor any of their employees, makes any warranty express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.