Bioinformatics
Explained

Bioinformatics explained: BLAST versus Smith-Waterman

July 4, 2007

CLC bio
Gustav Wieds Vej 10 · 8000 Aarhus C · Denmark
Telephone: +45 70 22 55 09 · Fax: +45 70 22 55 19
www.clcbio.com · info@clcbio.com

CLC bio
a result of science

Bioinformatics explained: BLAST versus Smith-Waterman

Database similarity searches are fundamental in bioinformatics as they are some of the best ways of gaining information about unknown genes and proteins and thus being able to characterize and predict functionality based on homology of new sequencing results from the laboratory.

So, database searching and identification of homology is fundamental in bioinformatics today and different methods are in different ways relevant and important to research. Both time and accuracy is of great significance as handling of large-scale projects is required in a suitable time as well as precision should of course not be compromised. The time aspect is of considerable importance as ever growing amounts of data are entering the research laboratories. The growing amounts of data provide us with more and more information and it also becomes more and more important to get the best quality results - we cannot risk overlooking any information hidden in our data material.

The requirements for database similarity searches should of course be reflected by the algorithms used for performing these identifications of homology. This article will therefore look into some advantages and disadvantages of two of the most widely used algorithms within this field, the Smith-Waterman and the BLAST algorithms.

Commonly used algorithms

Two of the most commonly used algorithms for database similarity searching are BLAST [Altschul et al., 1990] and Smith-Waterman [Smith and Waterman, 1981].

The Smith-Waterman algorithm is a method of database similarity searching which considers the best local alignment between a query sequence and sequences in the database being searched. The Smith-Waterman algorithm allows consideration of indels (insertions/deletions) and compares fragments of arbitrary lengths between two sequences and this way the optimal local alignments are identified [Smith and Waterman, 1981].

See *Bioinformatics explained: Smith-Waterman* on `http://www.clcbio.com/be/` to learn more about the Smith-Waterman algorithm.

BLAST (Basic Local Alignment Search Tool) also identifies homologous sequences by database searching [Altschul et al., 1990]. BLAST identifies the local alignments between sequences by finding short matches and from these initial matches (local) alignments are created. The BLAST algorithm is a development of the Smith-Waterman algorithm suggesting a time-optimized model contrary to the more accurate but timeconsuming calculations of the Smith-Waterman algorithm [Altschul et al., 1990].

See *Bioinformatics explained: BLAST* on `http://www.clcbio.com/be/` to learn more about BLAST.

Comparison of Smith-Waterman to BLAST

Comparisons between Smith-Waterman and BLAST show notable differences in the two methods according to accuracy and speed of the searches.

An example from Shpaer et al. [Shpaer et al., 1996] compares accuracy between the two search methods by use of a large objective test data set. Accuracy is compared as a means of selectivity and sensitivity of the search and relates to both the number of results obtained and to the quality of these. For the comparison, BLASTP version 1.4 distributed by NCBI [Altschul et al., 1990] and the publicly available software implementation of Smith-Waterman SSEARCH from the FASTA distribution were running the algorithms using default parameter settings. Test data were generated from the PIR release 40 database where homologous sequences are indicated by the classification of protein super families (SF). Query sequences were selected to represent all nontrivial super families in the PIR release 40 database and 502 protein sequences were chosen to act as queries against the test database consisting of all entries with a SF number assigned in the database [Shpaer et al., 1996]. The method of "missed@equivalence" point [Pearson, 1995, Pearson, 1991] was used to determine accuracy of the search algorithms. Results of this comparison of Smith-Waterman with BLAST on protein sequence similarity shows that the number of false positives as well as false negatives is significantly lower for Smith-Waterman than it is for BLAST and the risk that sequence similarity readily detected by Smith-Waterman will be missed using BLAST is considerable. In short, the conclusion of the comparison between the two sequence similarity search methods is that the Smith-Waterman algorithm performs significantly better than BLAST on accuracy [Shpaer et al., 1996]. Similar results have previously been obtained by Pearson [Pearson, 1995].

The algorithm used is not the only parameter in the measure of accuracy. Also, scoring matrices and penalties can affect this, and for instance optimization of scoring matrices can improve accuracy of the search significantly. This is the case for both the Smith-Waterman and the BLAST methods.

Data example of results obtained by Smith-Waterman and BLAST searches

Smith-Waterman has more than one time been compared to BLAST and has been proved more accurate, returning more qualified hits. An example is illustrated below.

Software

BLAST and Smith-Waterman output is compared by BLASTP implementation [Altschul et al., 1990] in CLC Combined Workbench 3.0 and CLC Bioinformatics Cell 1.04 accelerating Smith-Waterman searches and integrating the algorithm with CLC Combined Workbench. Both algorithms run as local database searches and are run with the BLOSUM62 matrix and an expect value of 10.

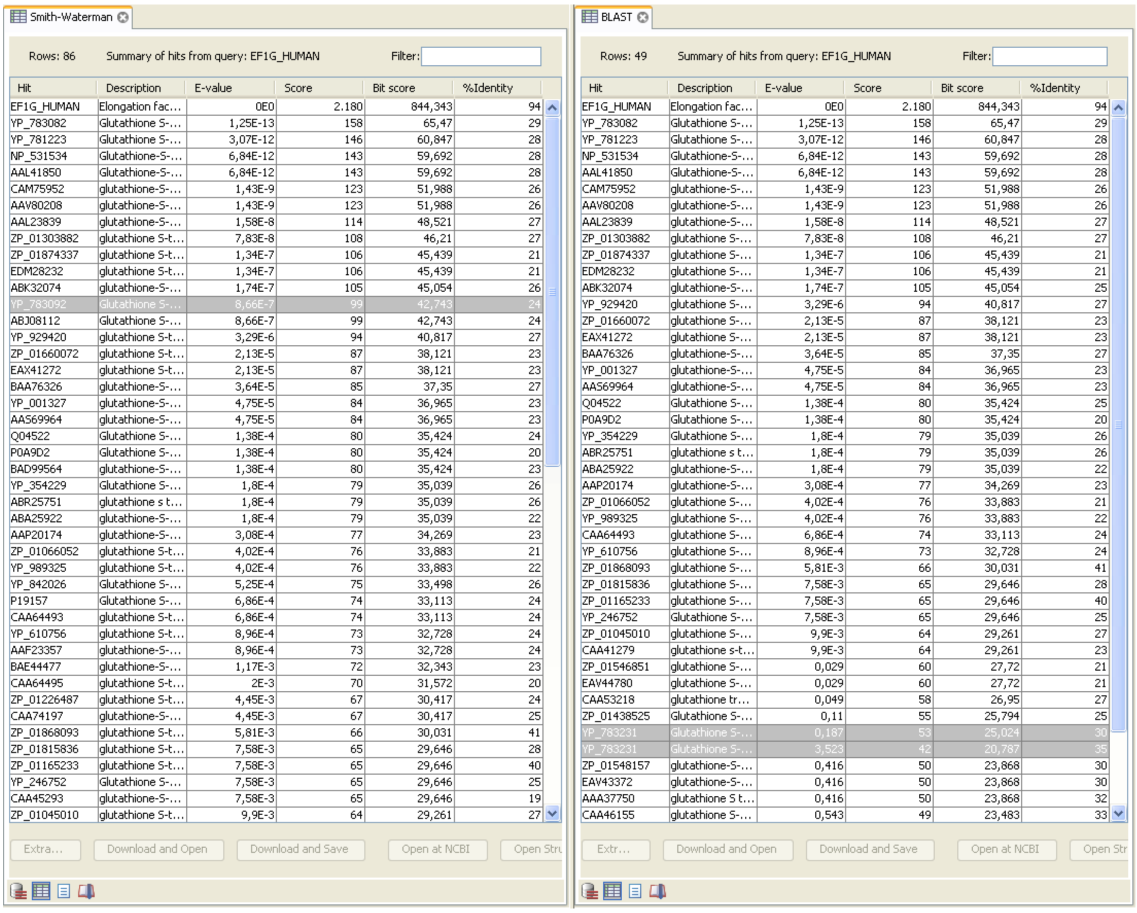
Test data

The family of Glutathione S-transferases (GST) is a diverse enzyme family having members within the species of plants, animals, and prokaryotes. Elongation factors have shown homology with the GST family and Elongation factor 1γ (EF1γ) has been shown to contain GST-related domains [Koonin et al., 1994] as well as GST activity has been proved for EF1γ expressed in *Escherichia coli* [Kobayashi et al., 2001]. For comparison of output from BLAST and from Smith-Waterman searches respectively, the human EF1γ (NCBI accession P26641) acts as query sequence in a match against a database containing 100 randomly chosen GST sequences identified by a simple search for this protein family at NCBI.

Results

CLC Combined Workbench returns the results from both search methods in a similar way making it easy to compare the quality of the output. Results are shown as graphic or tabular views (see 1).

The Smith-Waterman search returns a result list of 85 hits from the match between human EF1γ against the 100 members of the GST family. Comparatively, the BLAST search only returns a list of 49 hits. There is obviously a difference in identified homology between sequences for the two search methods as illustrated in figure 1 showing partly the tabular output of the two searches.



Smith-Waterman

Rows: 86 Summary of hits from query: EF1G_HUMAN Filter:

Hit	Description	E-value	Score	Bit score	%Identity
EF1G_HUMAN	Elongation fac...	0E0	2.180	844,343	94
YP_783082	Glutathione S-...	1,25E-13	158	65,47	29
YP_781223	Glutathione S-...	3,07E-12	146	60,847	28
NP_531534	Glutathione-S-...	6,84E-12	143	59,692	28
AAL41850	Glutathione-S-...	6,84E-12	143	59,692	28
CAM75952	glutathione-S-...	1,43E-9	123	51,988	26
AAV80208	glutathione-S-...	1,43E-9	123	51,988	26
AAL23839	glutathione-S-...	1,58E-8	114	48,521	27
ZP_01303882	glutathione S-t...	7,83E-8	108	46,21	27
ZP_01874337	glutathione S-t...	1,34E-7	106	45,439	21
EDM28232	glutathione S-t...	1,34E-7	106	45,439	21
ABK32074	glutathione-S-...	1,74E-7	105	45,054	26
YP_783092	Glutathione S-...	8,66E-7	99	42,743	24
ABJ08112	Glutathione S-...	8,66E-7	99	42,743	24
YP_929420	glutathione S-t...	3,29E-6	94	40,817	27
ZP_01660072	glutathione S-t...	2,13E-5	87	38,121	23
EAX41272	glutathione S-t...	2,13E-5	87	38,121	23
BAA76326	glutathione-S-...	3,64E-5	85	37,35	27
YP_001327	glutathione-S-...	4,75E-5	84	36,965	23
AAS69964	glutathione-S-...	4,75E-5	84	36,965	23
Q04522	Glutathione S-...	1,38E-4	80	35,424	24
P0A9D2	Glutathione S-...	1,38E-4	80	35,424	20
BAD99564	glutathione-S-...	1,38E-4	80	35,424	23
YP_354229	Glutathione S-...	1,8E-4	79	35,039	26
ABR25751	glutathione s t...	1,8E-4	79	35,039	26
ABA25922	glutathione-S-...	1,8E-4	79	35,039	22
AAP20174	glutathione-S-...	3,08E-4	77	34,269	23
ZP_01066052	glutathione S-t...	4,02E-4	76	33,883	21
YP_989325	glutathione S-t...	4,02E-4	76	33,883	22
YP_842026	Glutathione S-...	5,25E-4	75	33,498	26
P19157	Glutathione S-...	6,86E-4	74	33,113	24
CAA64493	glutathione S-t...	6,86E-4	74	33,113	24
YP_610756	glutathione S-t...	8,96E-4	73	32,728	24
AAF23357	glutathione-S-...	8,96E-4	73	32,728	24
BAE44477	glutathione-S-...	1,17E-3	72	32,343	23
CAA64495	glutathione S-t...	2E-3	70	31,572	20
ZP_01226487	glutathione S-t...	4,45E-3	67	30,417	24
CAA74197	glutathione-S-...	4,45E-3	67	30,417	25
ZP_01868093	glutathione S-t...	5,81E-3	66	30,031	41
ZP_01815836	glutathione S-t...	7,58E-3	65	29,646	28
ZP_01165233	glutathione S-t...	7,58E-3	65	29,646	40
YP_246752	Glutathione S-...	7,58E-3	65	29,646	25
CAA45293	glutathione-S-...	7,58E-3	65	29,646	19
ZP_01045010	glutathione S-t...	9,9E-3	64	29,261	27

Extra... Download and Open Download and Save Open at NCBI Open Stru

BLAST

Rows: 49 Summary of hits from query: EF1G_HUMAN Filter:

Hit	Description	E-value	Score	Bit score	%Identity
EF1G_HUMAN	Elongation fac...	0E0	2.180	844,343	94
YP_783082	Glutathione S-...	1,25E-13	158	65,47	29
YP_781223	Glutathione S-...	3,07E-12	146	60,847	28
NP_531534	Glutathione-S-...	6,84E-12	143	59,692	28
AAL41850	Glutathione-S-...	6,84E-12	143	59,692	28
CAM75952	glutathione-S-...	1,43E-9	123	51,988	26
AAV80208	glutathione-S-...	1,43E-9	123	51,988	26
AAL23839	glutathione-S-...	1,58E-8	114	48,521	27
ZP_01303882	glutathione S-...	7,83E-8	108	46,21	27
ZP_01874337	glutathione S-...	1,34E-7	106	45,439	21
EDM28232	glutathione S-...	1,34E-7	106	45,439	21
ABK32074	glutathione-S-...	1,74E-7	105	45,054	25
YP_929420	glutathione S-...	3,29E-6	94	40,817	27
ZP_01660072	glutathione S-...	2,13E-5	87	38,121	23
EAX41272	glutathione S-...	2,13E-5	87	38,121	23
BAA76326	glutathione-S-...	3,64E-5	85	37,35	27
YP_001327	glutathione-S-...	4,75E-5	84	36,965	23
AAS69964	glutathione-S-...	4,75E-5	84	36,965	23
Q04522	Glutathione S-...	1,38E-4	80	35,424	25
P0A9D2	Glutathione S-...	1,38E-4	80	35,424	20
YP_354229	Glutathione S-...	1,8E-4	79	35,039	26
ABR25751	glutathione s t...	1,8E-4	79	35,039	26
ABA25922	glutathione-S-...	1,8E-4	79	35,039	22
AAP20174	glutathione-S-...	3,08E-4	77	34,269	23
ZP_01066052	glutathione S-...	4,02E-4	76	33,883	21
YP_989325	glutathione S-...	4,02E-4	76	33,883	22
CAA64493	glutathione S-...	6,86E-4	74	33,113	24
YP_610756	glutathione S-...	8,96E-4	73	32,728	24
ZP_01868093	glutathione S-...	5,81E-3	66	30,031	41
ZP_01815836	glutathione S-...	7,58E-3	65	29,646	28
ZP_01165233	glutathione S-...	7,58E-3	65	29,646	40
YP_246752	Glutathione S-...	7,58E-3	65	29,646	25
ZP_01045010	glutathione S-...	9,9E-3	64	29,261	27
CAA41279	glutathione s-t...	9,9E-3	64	29,261	23
ZP_01546851	glutathione S-...	0,029	60	27,72	21
EAV44780	glutathione S-...	0,029	60	27,72	21
CAA53218	glutathione tr...	0,049	58	26,95	27
ZP_01438525	Glutathione S-...	0,11	55	25,794	25
YP_783231	Glutathione S-...	0,187	53	25,024	30
YP_783231	Glutathione S-...	3,523	42	20,787	35
ZP_01548157	glutathione-S-...	0,416	50	23,868	30
EAV43372	glutathione-S-...	0,416	50	23,868	30
AAA37750	glutathione S t...	0,416	50	23,868	32
CAA46155	glutathione S-...	0,543	49	23,483	33

Extr... Download and Open Download and Save Open at NCBI Open Str

Figure 1: *Tabular view of BLAST and Smith-Waterman searches of human Elongation factor 1γ matched against a database containing 100 Glutathione S-transferase family members. The results show a significant difference in the number of hits returned by the algorithms. In the left side of the picture is shown the result list returned by the Smith-Waterman search. The row greyed out shows an example of a sequence match identified by Smith-Waterman but not identified by BLAST. The greyed out rows of the BLAST results in the right side of the picture show an example of BLAST returning more hits for a specific comparison of two sequences.*

Searching using the Smith-Waterman algorithm clearly identifies a more detailed pattern concerning similarity between query and database sequences. For instance, the sequence returned in row 13 of the result list from the Smith-Waterman search is not identified by BLAST even though the same database has been searched. From this search for matches between the human EF1γ and the 100 GST family members, we expect to identify some degree of similarity as the elongation factor is distantly related to the protein family represented in the database, and as the database consists of members of the same family, a significant number of homologous are expected.

Furthermore, the results returned by the BLAST search contains in some cases more hits for each sequence, e.g. you get more possible alignments for the same sequence match of query and database sequence. This is explained by the fact that BLAST identifies local alignments by initially finding short matches between two sequences not taking the entire sequence into account whereas Smith-Waterman identifies *the optimal* local alignment for each two sequences compared. This exemplifies how a match identified by BLAST is not necessarily the optimal alignment between query and database sequence whereas the Smith-Waterman algorithm returns one - *the optimal* - local alignment for each pair of compared sequences.

Elongation factors seem to share similarities with the Glutathione S-transferase family at sequence level. The query sequence is though only distantly related to the Glutathione S-transferase proteins in the test database and we see how a BLAST search might miss important information and identification of sequence relatedness which is readily detected from the Smith-Waterman search.

Smith-Waterman: Advantages and disadvantages in review

Sequence similarity searches performed using the Smith-Waterman algorithm guarantees you *the* optimal local alignments between query and database sequences. Thus, you are ensured the best performance on accuracy and the most precise results - aspects of significant importance when you cannot afford to miss any information gained from the similarity search as e.g. when searching for remote homology. The Smith-Waterman algorithm being the most sensitive algorithm for detection of sequence similarity has however some costs. Time is a considerable disadvantage and performing a Smith-Waterman search is both time consuming and computer power intensive.

BLAST: Advantages and disadvantages in review

The algorithm behind BLAST increases speed of the database searches compared to the Smith-Waterman algorithm. Similarity between two sequences using BLAST is determined by identifying initial short matches and starting local alignments from these matches. Some matches between query sequences and database sequences may be missed by BLAST, and the method does not *guarantee* identification of the optimal alignment between query and database sequence. This might be a disadvantage but often the results obtained by use of the BLAST algorithm may be sufficient and BLAST has, however, some qualities appealing to research. The method is fast, results are returned in a short time and the tool has become the *defacto* standard within database similarity searching.

Should I use Smith-Waterman or BLAST for my sequence similarity searches?

Both the BLAST and the Smith-Waterman algorithms have qualities for research. BLAST is a very popular algorithm giving you results shortly after starting the search. Smith-Waterman ensures you that you get the optimal local alignment and thereby that you will not miss any important information from your sequence similarity search. The Smith-Waterman algorithm is, however, time consuming and has strong requirements for computer power. So, BLAST can be a good method for initial screening of sequence data when getting an indication of results in a short time is more important than getting the most accurate results. The Smith-Waterman algorithm should, on the other hand, be the choice for your database similarity searches when getting precise results is more important than time.

As demands of improved performance in less time are growing, the similarity search algorithms are still being developed and optimized and there are ways of accelerating the Smith-Waterman algorithm, e.g. based on FPGA chips or by the SIMD technology. By such development the Smith-Waterman algorithm becomes a more reasonable choice for practical database similarity searches as time is reduced significantly.

Other useful resources

Public available Smith-Waterman implementation from the Japanese Institute for Bioinformatics Research and Development http://www-btls.jst.go.jp/cgi-bin/Tools/SSEARCH/index.cgi

Bioinformatics Explained

Public available Smith-Waterman implementation from the FASTA distribution http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=select&pgm=sw

The BLAST web page hosted at NCBI http://www.ncbi.nlm.nih.gov/BLAST

Bioinformatics explained: Smith-Waterman http://www.clcbio.com/be/

Bioinformatics explained: BLAST http://www.clcbio.com/be/

CLC Combined Workbench http://www.clcbio.com/combined

CLC Bioinformatics Cell http://www.clccell.com

Creative Commons License

References

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[Kobayashi et al., 2001] Kobayashi, S., Kidou, S., and Ejiri, S. (2001). Detection and Characterization of Glutathione S-Transferase Activity in Rice EF-1betabetagamma and EF-1gamma Expressed in Escherichia coli. *Biochemical and Biophysical Research Communications*, 288(3):509–514.

[Koonin et al., 1994] Koonin, E., MUSHEGIAN, A., TATUSOV, R., ALTSCHUL, S., BRYANT, S., BORK, P., and VALENCIA, A. (1994). Eukaryotic translation elongation factor 1 {gamma} contains a glutathione transferase domain–Study of a diverse, ancient protein superfamily using motif search and structural modeling. *Protein Science*, 3(11):2045.

[Pearson, 1991] Pearson, W. (1991). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11(3):635–50.

[Pearson, 1995] Pearson, W. (1995). Comparison of methods for searching protein sequence databases. *Protein Science*, 4(6):1145.

[Shpaer et al., 1996] Shpaer, E. G., Robinson, M., Yee, D., Candlin, J. D., Mines, R., and Hunkapiller, T. (1996). Sensitivity and selectivity in protein similarity searches: a comparison of smith-waterman in hardware to blast and fasta. *Genomics*, 38(2):179–191.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.