

Nucleotide Searches: SynaSearch-Bulk demonstrates a 50 to 500 fold performance improvement and increased sensitivity over NCBI-BLAST

Poh Yang Ming, Wagied Davids, Zayed Albertyn, Ali Reza Zamli and Robert G. Hercus

SynaSearch-Bulk™ is a multiple query database search and alignment program developed to interrogate nucleic acid patterns stored in SynaBASE™. To investigate performance enhancements, benchmark analysis was undertaken to compare the results for SynaSearch-Bulk to NCBI BLAST in conducting nucleotide-nucleotide searches. Using equivalent hardware, SynaSearch-Bulk was 50 to 500 fold faster than NCBI BLASTN using a data set of 687,929 bacterial ORFs. Accuracy analysis indicated that SynaSearch-Bulk achieved greater sensitivity to that of BLASTN. Thus sequence similarity searches utilising the SynaBASE network architecture showed significant improvements in search times whilst demonstrating equivalent or better accuracy and sensitivity.

Introduction

The need for proficient data storage and management along with efficient analysis tools is becoming increasingly critical due to the exponential increase in the amount and complexity of biological data.

Applications to search nucleotide and protein sequences that are currently available include the widely-used Basic Local Alignment and Search Tool (BLAST) [1], BLAT [2], WU-BLAST (<http://blast.wustl.edu>), Gapped BLAST and PSI-BLAST [3].

For large-scale sequence similarity searches, an important question is the choice of suitable data structures for storing volumes of data for further processing. Flat files, relational databases and suffix trees, are a few examples of data structures used by search algorithms. These tools and applications use intricate algorithms whose performance deteriorates due to poor scalability with extremely large datasets. Thus, there exists a pressing need for bioinformatics applications or tools that can

perform ultra-fast and accurate sequence searching without the cost and manpower implications of managing large supercomputers or multi-node server clusters.

SynaBASE™ is a proprietary structured database system designed to manage sequence data by storing unique subsequences for more rapid and sensitive sequence analysis [4, 5]. During the SynaBASE build process, exhaustive overlapping k-mers covering the entire sequence are produced and indexed for searching. Due to subsequence redundancy being minimised in SynaBASE, a fraction of the space is required for efficient data storage. Alignment to these subsequences reduces the time cost of obtaining the best matches. SynaSearch™ is an application designed to search nucleotide and protein sequences stored in SynaBASE. A command line utility for SynaSearch, SynaSearch-Bulk™ was developed based on the same core principles of searching sequence databases with “multiple” query sequences.

Sequence similarity searches NCBI Bacterial DNA were conducted to benchmark the performance and overall accuracy of SynaSearch-Bulk and BLASTN. Results indicate that using SynaSearch-Bulk to search against data, which is intelligently and efficiently structured within SynaBASE, leads to a performance level capable of meeting the demands of contemporary genomics applications.

Materials and Methods

Computing resources

All analysis were conducted on a single HP Itanium 2 Rx5670 4 IA-64 1.3GHz CPU, 32GB RAM, running RedHat AS2.1 with Java(TM) 2 SDK Standard Edition 1.4.2 for Linux IA64 Platform.

Nucleotide-Nucleotide searches

687,929 ORFs corresponding to bacterial protein-coding sequences were downloaded from NCBI (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>).

A single SynaBASE (Bacteria Nucleotide-SynaBASE) was built and used as a database for subsequent searches. Searches were performed with 100, 1000, 5000 and 10000 sequences selected at random from this database and used as queries against these ORFs in SynaBASE. A BLAST database of NCBI bacteria data was constructed using the formatdb command for BLASTN searches for the above queries.

The parameter settings used for the nucleotide-nucleotide with SynaSearch-Bulk and BLAST is shown in Table 1A and 1B. Note that BLASTN was run using default settings.

SynaSearch-Bulk algorithm and parameters

SynaSearch works by seeding alignments with multiple overlapping or non-overlapping word matches to the database over a given window length of the query sequence, whereas BLAST is restricted to single word matches, which are subsequently extended. Therefore in cases where percent ID is quite low (<20 % ID) BLAST finds more smaller matches based on a single hit. However, at more intermediate and higher percent identities (35% to 80%), SynaSearch finds more significant result because it requires multiple matches of variable length subsequences within a query window. SynaBASE alignment seeds are allowed to overlap with each other where they are subsequently merged.

Searches performed	Word Size	Window length	Indels	Score	E-Value cut-off
Nucleotide	11	300	5	14	10

Table 1A: SynaSearch parameter settings for nucleotide-nucleotide and protein-protein searches.

BLAST program	Word size	Matrix	Gaps (open, extension)	E-value cut-off
BLASTN	11	+1 match, -3 mismatch	5, 2	10

Table 1B: Default BLAST parameters used for searching.

SynaSearch methodology is also different to BLAST in that it uses a 'greedy' approach to finding matches in SynaBASE e.g. if a match is not found at a word size equal to 11 then the algorithm will look for SynaBASE subsequences matching at word lengths of 10. If matches at word length 10 are not found then the process of stepping back one character is iterated.

Results and Discussion

Nucleotide-nucleotide searches

A data set consisting of 687,929 NCBI Bacterial nucleotide sequences was used to illustrate the efficiency of SynaBASE in storing and searching of biological sequence data.

Number of nucleotide sequences	BLASTN		SynaSearch-Bulk v1.1.1		BLASTN: SynaSearch-Bulk
	Total time (s)	Average time (s)	Total time (s)	Average (s)	
100	25879	258.8	444	4.4	58.3
1000	324366	324.4	2287	2.3	141.8
5000	1922172	384.4	4327	0.9	444.2
10000	4580247	458.0	8100	0.8	565.5

Table 2: Variation of search time with increasing number of queries against a SynaBASE of 687929 NCBI Bacteria All nucleotide sequences.

Sequence similarity searches were conducted using SynaSearch-Bulk and BLASTN on the respective data sets with queries of 100, 1000, 5000, and 10000 random sequences. SynaSearch-Bulk was over 50 to 500 times faster than BLASTN (see Table 1A and 1B above). The gain in search time is due to the efficient structuring of data whereby all subsequence patterns and their inter-relationships are efficiently accessed from SynaBASE. As more data is stored in SynaBASE, data coverage increases and the system becomes more efficient in detecting sequence patterns and similarities. Table 2 is the data for the searches. Sample output of SynaSearch-Bulk is given in Figure 1 overleaf.

```

Query= emb|AL646053.1|:1215605-1217656
      (2052 letters)

Database: all_bacteria
        687,929 sequences; 640,296,292 total letters

Sequences producing significant alignments:

emb|AL646053.1|:1215605-1217656          4056  0E0
emb|BX571965.1|:2849770-2851533        578  2E-162
gb|CP000010.1|:c655374-653623          570  4E-160
gb|AE004091.1|:c2096356-2094329        273  7E-71

>emb|BX571965.1|:2849770-2851533
      Length=1764
      Score = 578 bits (292), Expect = 2E-162
      Identities = 1246/1564 (79%), Gaps = 0/1564 (0%)
      Strand=Plus/Plus

Qry: 340  GACTGGCGCGTCAACGCGCAACGCGCAACAGGGCTACAGCCTGGGCGGCTGATCCTGAACGTGGCCGGCA 409
          |||
Tgt: 103  GACTGGCGCGTCAACGCGCAACGCGCAACAGGGCTATTCGCTCGGCGGCTGATCCTGAACGTGTCGGCA 172

Qry: 410  AGGTCACCGCCAACTACTGGCTCTGCACGTCTACAGCCCCGAGGCCGGGCACGCGCATCGCGAGGGCGA 479
          |||
Tgt: 173  AGGTGATCGCGAACTACTGGCTGAGCCAGTCTACCCGAGCGCGATCGGGCAAGCGCACCGCAATGCGGA 242

Qry: 480  CCTGCACATCCACGATCTCGACATGCTGTCGGGCTACTGCGCGGGATGGTCCCTGCGCCAGCTGCTGACC 549
          |||
Tgt: 243  TCTGCACATCCACGATCTCGACGTGCTGTCGGGCTACTGCGCGGGCTGGTTCGTCACGCTGCTCAAC 312

Qry: 550  GAGGGCTTCAACGCGCGTCCCGGGCAAGGTGGAGGCCACGCCCGCGCCATATGTCGGCGGGCCATCGGCC 619
          |||
Tgt: 313  GAAGGGTTGAACGCGCGTCCCGGGCAAGGTGAGTCTGGGGCCGCCGAAGCACATGTCGAGCGCGTCCGGCC 382

Qry: 620  AGATCGTCAACTTCTCGGCACGCTGCAGAACGAGTGGGCCGGCGCAGGCGTTTCAGCTCGTTCGACAC 689
          |||
Tgt: 383  AGATCGTGAACCTTCTCGGCACGCTGCAGAACGAATGGGCGGGCGCAGGCGTTTCAGCTCGTTCGACAC 452

Qry: 690  CTACATGGCCCCGTTTCGTGCGGCGGACGCCATGTGCTACGCGCGGTCAAGCAGGCCATGCAGGAGCTG 759
          |||
Tgt: 453  GTACATGGCGCGTTCGTGCGCCGCGACGCGCTCACCTACGCCGAAGTGCAGTCCGTCAGGAACTG 522

Qry: 760  ATCTACAACCTGAACGTGCCACGCCGCTGGGGCACGCAGACGCCCTTACCAACCTGACATTCGACTGGA 829
          |||
Tgt: 523  ATCTACAACCTGAACGTGCCGTCACGCTGGGGCACGCAGACGCCGTTTCAGAACCTGACGTTTCGACTGGA 592

Qry: 830  CCTGCCCGGCGACCTGCGCGAGCAGATCCCCCTACCTTGGCGGGGAGGAGATGCCGTTACCTACGGCGA 899
          |||
Tgt: 593  TCTGCCCCGAGGATCTGCGCGAGCAAGTGCCTGATCGCCGGCGAAGAGATGCCGTTACCTACGGCGA 662

Qry: 900  CCTCCAGCCCGAGATGGACATGATCAACCGCGCTACATCGAGGTGATGATGGCCGGCGATGCCAAGGGC 969
          |||
Tgt: 663  TCTGCAGCCCGAAATGGACATGATCAACAGCGTACATCGAGGTGATGACAGCGGGCGACCGCGGGC 732

Qry: 970  CGCGCCTTCACTTCCCCATCCCGACGTACAACATCAGCCCGATTTTCGACTGGGACCACCCCAACACCA 1039
          |||
Tgt: 733  CGCGTGTTCAGTTCCTCCGATTCGACCTACAACATCAGCCCGATTTTCGACTGGCACAGCCCGAACGCGC 802

```

Figure 1: Sample output of SynaSearch-Bulk for nucleotide searches showing the alignment of the 2nd ranked with a 79% identity to the query.

Accuracy and Sensitivity Analysis

For accuracy analysis, a subset of sequences was selected for the nucleotide-nucleotide searches described above. Both SynaSearch-Bulk and NCBI BLASTN were able to find the corresponding sequences from their respective searches. Thus SynaSearch did not lose any search accuracy with improved search speeds. Overall, a good agreement was observed between BLASTN and SynaSearch-Bulk sequence similarity search results (Table 4). Figure 2, on the next page, shows a cumulative number of hits over % identity of SynaSearch and BLASTN.

SynaSearch-Bulk found more alignments than BLASTN at intermediate percent identities i.e. between 35% and 80% due to some notable differences in the two algorithms. In cases where no match is found at the specified word size, the SynaSearch-Bulk algorithm looks for overlapping matches at shorter word size until it finds matches. Therefore this approach coupled with overlapping matches detects more regions of local pairwise sequence identity during a database search.

SynaSearch	Score	NCBI	Score
embBX119912.1_190095-191966	1872	embBX119912.1_190095-191966	1872
emb BX119912.1 :c1619042-1618917	126	emb BX119912.1 :c1619042-1618917	126
emb BX119912.1 :189235-190170	76	embBX119912.1_189235-190170	76

Table 4: An example validation of SynaSearch results using as query embBX119912.1_190095-191966 against the NCBI Bacteria All database.

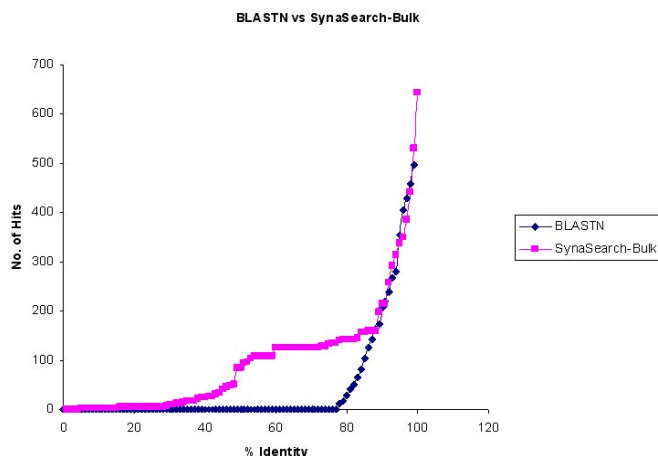


Figure 2: Sensitivity comparison between SynaSearch and BLASTN. The cumulative numbers of alignment hits at percent identity ranges were compared for BLASTN and SynaSearch. Therefore at intermediate pairwise identities, SynaSearch finds more significant alignments. An expectation value cut-off of 10 was used for both applications.

Conclusion

Traditional sequence similarity searches programs such as the BLAST suite of programs are known to be fast and accurate for finding related homologous sequences. However for large-scale nucleotide and protein searches the performance of sequence similarity searches presents a bottleneck in the annotation and discovery process of novel genes.

In this report the speed and accuracy performance of BLAST and SynaSearch-Bulk was benchmarked on nucleotide searches. SynaSearch is a unique sequence similarity search tool built on the structured network database technology of SynaBASE. SynaSearch-Bulk outperformed BLASTN in terms of software performance and demonstrated increased sensitivity compared to that of BLAST. For large-scale nucleotide-nucleotide searches, SynaSearch-Bulk was able to search bacterial ORFs sequences 50 to 500 times faster than BLASTN using identical hardware. More detailed sensitivity analysis indicated that SynaSearch-Bulk achieved comparable accuracy to that of BLAST for nucleotide-nucleotide. SynaSearch-Bulk also returned a higher proportion of significant search results at more intermediate pairwise percent identity ranges (35% to 80%).

In conclusion, SynaSearch displays improved sensitivity to NCBI BLASTN, while SynaSearch-Bulk, its multiple query processing counterpart showed increased performance on large-scale data sets compared with NCBI BLASTN. This demonstrates that SynaSearch-Bulk utilising SynaBASE shows considerable reduction in database search times for enabling more sensitive comparison of genomics data in a high-throughput environment.

References

- [1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403-410.
- [2] Kent, W.J. (2002) BLAT- the BLAST-like alignment tool. *Genome Res.*, 12, 656-64.
- [3] Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.
- [4] Zayed I. Albertyn, Arif Anwar, Nataraj Dongre, Johan Poole-Johnson, Ching Soo Meng and Robert G. Hercus. (2004) A Revolutionary Application of Novel Structured Network Database for Genome to Genome Comparison. Available from www.synamatix.com.
- [5] Zayed I. Albertyn, Tan Ka Ju, Wong Chee San, M. Ramachandran and Johan Iskandar (2004) Reconstructing Alignments Based on Patterns Inherent in Biological Data: A Qualitative Comparison. Available from www.synamatix.com.



Utilising Synamatix technologies to power its online bioinformatics application services.

CLICK HERE FOR YOUR 3-MONTH FREE TRIAL

This Application Note is for distribution to Synamatix members, associate members and mailing list subscribers only. The contents are provided for personal, non-commercial purposes only and are protected by various national and international intellectual property laws, conventions and treaties. All title and intellectual property rights in and to Synamatix, SynaBASE, SynaMine, and SynaSuite and the accompanying printed materials are owned by Synamatix sdn bhd. Other trademarks or names are used only in an editorial fashion and to the benefit of the respective trademark owner with no intention of the infringement of the trademark. All trademarks or service marks are the property of their respective owners.