

Performance Benchmarking of a Novel Application in Comparative Genomics Using a Structured Network Database of Genome Patterns

Tan Ka Ju, Ching Soo Meng, and Zayed Albertyn

*As the number of genomes that are sequenced increases, so to does the need for biologists to find more efficient and scalable alignment methods to enable high-throughput sequence analysis. Comparative genomics exemplifies this need, as large, 100-200 mega base pair input sequences need to be accurately aligned for elucidation of biologically significant and evolutionary conserved regions. A novel application, [SynaCompare](#)TM, has been optimised to compare genome patterns stored in the unique [SynaBASE](#)TM architecture. Published benchmarks from Patternhunter, BLASTN and MegaBLAST were used to evaluate the performance of SynaCompare. SynaCompare aligned two *Arabidopsis thaliana* chromosomes to each other in 139 seconds, approximately 3.5 times faster than the best recorded time for Patternhunter. In all instances, SynaCompare found alignments faster than Patternhunter on optimised settings. Another example was the comparison of human chromosomes 21 to 22 in fewer than 5 minutes by SynaCompare versus 1 hour and 27 minutes by Patternhunter (18 times faster). The advantages of SynaCompare's performance over other tools are attributed to patterns from one or many genomes being efficiently structured in a single SynaBASE for fast and scalable querying. SynaBASE is a next generation database technology that uses a structured network pattern approach to enable scalable and efficient comparative genomics analysis for researchers faced with the demands of working with data from multiple genomes.*

Introduction

New challenges in sequence analysis have been posed with the advent of genome data from a multitude of model organisms. Following on from this explosion in biological data, numerous tools have been developed to align genomic sequences from large eukaryotic chromosomes or bacterial genomes [1]. The BLAST family of programs and FASTA are widely used for local alignments with modifications for larger input sequences [2-3]. These algorithms are generally quite sensitive but require expensive cluster hardware for optimal performance in high throughput comparative genomics.

PASH, Patternhunter, LAGAN and Mummer are more recent examples of programs that efficiently align mega base order sequences to each other with moderate CPU requirements, e.g. small servers or desktop workstations [4-7]. The majority of these tools concatenate smaller pattern matches between a query and target database index into much larger and complex alignments. PatternHunter identifies all the homologies between large DNA sequences and it is reported as being much faster and higher in quality

than BLAST [5]. Mummer aligns large sequences to each other using suffix trees generated from query and database sequences [7].

There is a significant time overhead to initially generating database indices or suffix tree structures for subsequent comparison steps in an alignment algorithm. As more data is added, speed performance gradually decreases due to more time spent on sequence pre-processing. The objective of this study was to show that a database that stores sequences as patterns in an intelligent, novel system called SynaBASETM can be applied for more rapid and scalable genome comparisons. SynaBASE automatically calculates and stores associations between sequences based on shared patterns [8]. As more useful data is added to SynaBASE, it becomes more efficient in homology searches. Whole genomes or multiples thereof can thus be stored in SynaBASE and queried in a fraction of the time taken by applications built on conventional databases.

SynaCompare™ is an application that performs pairwise comparisons using SynaBASE patterns. In order to justify the claim that SynaCompare (based on SynaBASE technology) is a novel application in sequence alignment, the performance of this tool was compared to published data on Patternhunter with respect to speed and data scalability. An assessment of alignment accuracy will be discussed in a later study.

Methods

Speed benchmarks from the PatternHunter publication were duplicated in SynaBASE to determine the elapsed time for each sequence comparison (see Tables 1 & 2 for a complete list) [5]. For pairwise comparisons, at least one of the genome/chromosome sequences was built in SynaBASE for querying. Table 1 summarises the genome build information for SynaBASE. All comparisons were run on the Linux HP® Itanium™ architecture using a single 1.3 GHz CPU with 64 GB RAM. Patternhunter, BLAST and MegaBLAST were run on a 700 MHz PC with 1GB RAM as described in [5].

Results & Discussion

All data used in SynaBASE was built from source genomes listed in Table 1 below.

to trigger an alignment with SynaBASE patterns consistently produces alignments in less processing time. In the case of the *Arabidopsis thaliana* comparison, default settings for SynaCompare produced results 3.5 times faster than Patternhunter's optimised double hit model i.e. 139 seconds for SynaCompare versus 498 seconds for Patternhunter (see Table 2 on the next page).

SynaCompare aligns two sequences with the requirement that either one of them are residing in a SynaBASE. In Table 1, a SynaBASE was built from the query and target sequences except for the mouse genome in the last column. The size of the database affects the speed at which matching patterns are found because the entire database is scanned for alignment seeds. Therefore elapsed time for finding alignments from pattern seeds in SynaBASE will depend on the database used. However, SynaCompare still manages to achieve optimal performance on entire genomes due as access time to SynaBASE patterns scales at log n base 2, where “n” is the database size.

An example of homologous regions reported between the human and mouse X chromosomes by SynaCompare is shown in Figure 1 on the following page. The 152MB by 159MB comparison of these sequences referenced 36,408,557 patterns in a SynaBASE of the mouse genome to report alignments in 16 minutes.

SynaBASE DB Name	Sequences	Source Data Size (bases)	SynaBASE Size (patterns)	Build Time (min)
<i>Arabidopsis thaliana</i> chr2 & chr4 [11]	2	38,290,401	5,473,407	7
<i>E.coli</i> and <i>H.influenza</i> genomes [12]	2	6,469,359	918,214	< 1
<i>Mycoplasma pneumoniae</i> & <i>M. genitalium</i> genomes [12]	2	1,396,468	199,034	< 1
<i>Homo sapiens</i> genome NCBI 33 (chr21 & chr22) [9]	2	96,453,509	9,691,439	13
<i>Mus musculus</i> genome mm2 [10]	22	2,638,213,512	36,408,557	420

Table 1. SynaBASE statistics for selected sequence databases used in SynaCompare.

Results show that SynaCompare performance exceeds that of published data for Patternhunter, BLAST and MegaBlast. Query sizes of less than 5MB that are in the range of small microbial and viral genomes require less processing time and hence less memory. In this range the fastest pairwise alignments from SynaCompare are three times faster as Patternhunter on default settings. Data for the *M. pneumoniae* and *E. coli* genome comparisons in Figure 1 show that using a seed length of 12 bases

The human chromosome 21 versus 22 comparison demonstrates the scalability of using SynaBASE for genome alignments. SynaCompare was able to find alignments between chromosome 21 and chromosome 22 and completed the process approximately 18 times faster than the double-hit model Patternhunter i.e. 293 seconds versus 5250 seconds for SynaCompare and Patternhunter respectively. SynaCompare also compared these same two chromosomes in 1260 seconds using a SynaBASE of the human genome (data not shown). The latter analysis is a one against all comparison, which will enable a further increase in performance. This will be the subject of a later article. Compiling a SynaBASE enables iterative querying for comparative genomics by saving on the scalability overhead of repeatedly building an index for each alignment, as evidenced by comparison to conventional programs.

Seq1	Size	Seq2	Size	PH	PH2	MB28	Blastn	SynaCompare
M. pneumoniae	828K	M. genitalium	589K	10s 65M	4s 48M	1s 88M	47s 45M	3.2s
E.coli	4.7M	H. influenza	1.8M	34s 78M	14s 68M	5s 561M	716s 158M	11.5s
A. thaliana 2	19.6M	A. thaliana chr4	17.5M	5,020s 279M	498s 231M	21,720s 1,087M	∞	139s
H. sapiens chr22	35M	H. sapiens chr21	26.2M	14,512s 419M	5,250s 417M	∞	∞	293s

Table 2. Performance comparison - A Minimum match / seed length of 12 bases was used for all SynaCompare runs. Patternhunter (PH) uses 11 base seed weight [5]. PH2 – Patternhunter with double hit model; MB28- MegaBlast with default seed length = 28. Table entries under PH, PH2, BLASTN and MB28 indicate time and space used; ∞ means out of memory. Figures for Patternhunter, BLASTN and MegaBlast reproduced from [5].

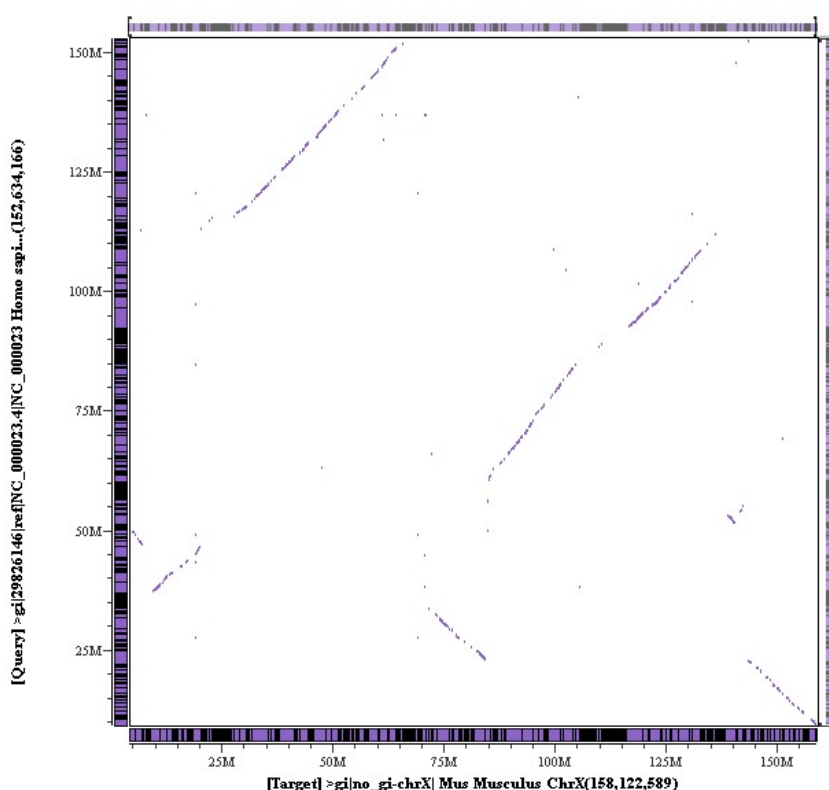


Figure 1 Alignment of the human and mouse X chromosomes - The human X chromosome sequence was used to query a SynaBASE of the mouse genome [10]. Matching regions show areas of possible synteny or conserved non-coding regions. Elapsed time for this query was 16 minutes on a single CPU 1.3 GHz Itanium CPU with default SynaCompare settings.

Conclusion

SynaCompare is a novel application in comparative genomics that aligns large sequences to each other based on the shared patterns between them. It differs from conventional programs such as Patternhunter in that it uses patterns stored in SynaBASE for rapid sequence matching. Based on comparison to published results, SynaCompare found alignments at significantly faster speeds when compared to Patternhunter in every instance of genome alignments. The results demonstrate that SynaBASE is a novel database system for applications in computational biology. This system

provides genome-scale alignments at speeds that exceed current tools. By considering all patterns in a data source sensitivity is maintained. Due to the increased performance and scalability advantages, SynaCompare is well suited to manage and analyse the current explosion in sequence data.

REFERENCES

1. Pollard DA, Bergman CM, Stoye J, Celniker SE and Eisen MB (2004). Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* 5, 6.
2. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* 215,403-10.
3. Pearson WR and Lipman DJ (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci* 85, 2444-8.
4. Kalafus KJ, Jackson AR, Milosavljevic A (2004) Pash: efficient genome-scale sequence anchoring by Positional Hashing. *Genome Res.* 2004 14.672-8.
5. B. Ma, J. Tromp, and M. Li. Super Seeds for Faster and More Sensitive Homology Search. *Bioinformatics* 18: 440-445. 2002.
6. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S and NISC Comparative Sequencing Program (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13,721-31
7. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5,R12.
8. Albertyn, ZI, Wong CS, Tay L and Hercus, G (2004). A new approach to genome-wide annotation based upon calculation of significance from a structured pattern database. Available from www.synamatix.com
9. International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860-921.
10. Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420,520-62.
11. The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, , 796-815.
12. Bacterial Genomes downloaded from NCBI at <ftp.ncbi.nlm.nih.gov>
13. Albertyn ZI, Tay LC, Ching SM and Hercus, RG (2004). Inference of multi-genome sequence conservation from structured pattern information. Available from www.synamatix.com.

To request detailed technical information or feedback please send an email to tech@synamatix.com

This Research Newsletter is for distribution to Synamatix members, associate members and mailing list subscribers only. The contents are provided for personal, non-commercial purposes 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, SynaSearch 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.