

Utilisation of a novel structured network database in efficient storage and management of genomics data

Ali Reza Zamli, Zayed Albertyn, Poh Yang Ming, Ching Soo Meng and Robert G. Hercus

SynaBASE™ is a novel structured network database, which efficiently stores genomics sequence data based on subsequence pattern constructs and their inter-relationships. Storage efficiency is maximised using appropriate build parameters without the loss of 'meaningful' data. Effectiveness of such a novel approach is shown using the example of 100 Streptomyces pneumoniae R6 genomes that have each been randomly mutated at 1% based on the original sequence. Examples of curated genes of varying length from this bacterial strain have been shown to match exactly with that of all mutated genomes. As more divergent genomic sequences get stored in SynaBASE, new unique patterns are generated leading to an increase in the amount of space required.

Introduction

Genomic data from multiple individuals is the basis of advancing from the traditional “one fits all” blockbuster drug, towards “personalised” medicine. The increasing complexity and exponential growth of genomics data makes mining and analysis a progressively more daunting task. The need for proficient data storage and management along with efficient analysis tools has never been so critical.

Until recently, these types of applications required mainframes or supercomputers with custom data management applications, which were both expensive to develop and difficult to maintain. Current solutions that are available for biological data are limited to flat file systems, suffix trees or relational databases such as Oracle® [1]. SynaBASE™, a novel structured network database, has demonstrated unique capabilities in overcoming storage efficiency and data redundancy. SynaBASE achieves this by needing to only store each

unique sequence pattern once to represent disparate data sets. As new sequence data is added, only data that extends the existing subsequence pattern is stored [2]. Ultimately, increases in database size will be directly proportional to differences between new and existing data.

Storage efficiency is further maximised through SynaBASE's ability to organise subsequences into a network structure. The commonalities and differentials between biological data are efficiently managed as structured subsequences for high level scientific analysis on a single platform, such as exhaustive genome-scale comparisons [3].

In this paper we demonstrate, using the Streptomyces pneumoniae R6 genome, that data storage in SynaBASE can scale extremely efficiently without any loss of meaningful data.

Methods

The genomic sequence of *Streptococcus pneumoniae* R6 strain (gi|15902044 |ref|NC_003098.1, 2038615bp) was used in this study. The sequence was obtained from NCBI [4] and a simulation program was used to randomly mutate the sequence 100 times at a 1% mutation rate to yield 100 independent genomic sequences with 1% variation. These variant genomes were used to build a SynaBASE using constant and varying join limit build parameter settings. Join limit is defined as the number of times a particular sequence pattern must occur before it is extended to another sequence pattern. A set of characterised gene sequences from *Streptococcus pneumoniae* R6 of known positions and varying lengths were used to analyse the databases using SynaSearch, SynaSuite's basic alignment search tool.

The *Streptococcus pneumoniae* R6 genomic sequence was also mutated once each at 2%, 5%, 10%, 15% and 20% and each of the iterations was copied 20-fold for building SynaBASE with a constant join limit of 12. This was used to simulate sequence divergence. The amount of patterns of subsequences generated and stored in SynaBASE was subsequently analysed. All build processes and analyses were conducted on a single CPU 1.3 GHz Intel Itanium HP server running Linux Red hat.

Results and Discussions

Using a build process with a join limit increment of 2 units for every increase of 1 additional genome stored, the database size starts to plateau after 20 genomes as seen from Figure 1. This reflects the nature of SynaBASE whereby as more highly similar genomic sequences are stored, the amount of storage space required decreases. This in turn correlates to the number of subsequences generated. This is also a more efficient means of handling biological sequence data as compared to using a suffix tree data structure generated by applications such as Mummer [5], which are built using linear space.

Database size vs number of genomes

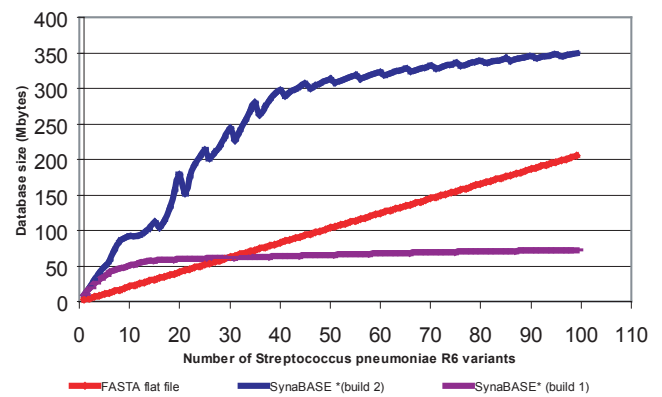


Figure 1: Comparison of FASTA flat file database size and SynaBASE size of various build parameters against the number of *S pneumoniae* mutant genomes built in SynaBASE. *Each genome was randomly mutated at 1% to give variation to simulate intra-species sequence variation. A flat file database 100 times the size of the ~2MBp *S. pneumoniae* genome is needed to store the data whereas SynaBASE requires less storage due to unique patterns being stored only once.

Experimental setup shown by the legend is described in Table 1 below:

	SynaBASE Build Parameters	
	Initial join limit	Increment
Build 1	12	2 per genome
Build 2	12	2 per 5 genomes
FASTA	N/A	N/A

Table 1: Build parameters used for the *Streptococcus pneumoniae* R6 mutation study.

Analysis of the databases that were built was conducted using a set of annotated genes from the *Streptococcus pneumoniae* R6 genome obtained from NCBI's [3] protein translation table output. SynaSearch™ alignments with these queries gave a relatively accurate map to the gene positions as shown in Table 2 below. The positions matched exactly in all the 100 variants of the mutated genomes (data not shown).

Gene name	Length (bp)	Gene ID	NCBI published positions			SynaSearch Highest Scoring Alignments		
			Position		Strand	Position		Strand
			start	stop		start	stop	
ctpC	2073	934119	1892049	1889977	-	1892049	1889977	R-
polA	2634	933214	31306	33939	+	31306	33939	F+
transposase_B	339	934108	1020448	1020110	-	1020448	1020110	R-
gpdA	1017	934131	1879718	1880734	+	1879718	1880734	F+
hk13	1341	933263	470513	471853	+	470513	471853	F+
pstA	816	934117	1876225	1877040	+	1876225	1877040	F+
prtA	6435	933322	562418	568852	+	562418	568852	F+

Table 2: Selected genes of *Streptococcus pneumoniae* R6 (from NCBI's ptt file) used in SynaSearch alignments. Target database used was the entire 100 variants of the *S. pneumoniae* genomes (203, 861, 501 residues).

In the simulation whereby the *Streptococcus pneumoniae* R6 was mutated each at 2%, 5%, 10%, 15% and 20%, there is an increasing number of new patterns generated as the genome sequences stored within SynaBASE become more diverged (Figure 2B). These staggered mutations result in slight increases in storage space proportional to the additional unique patterns generated (Figure 2).

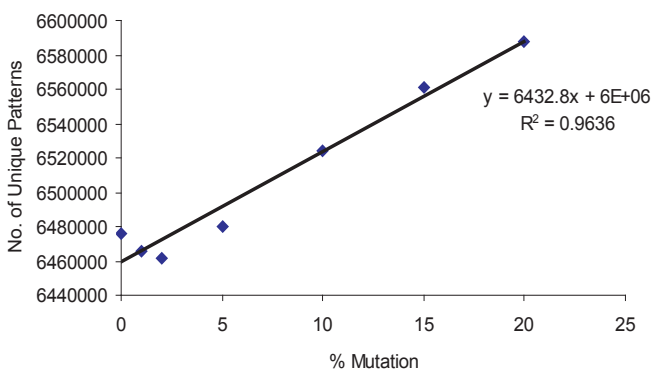
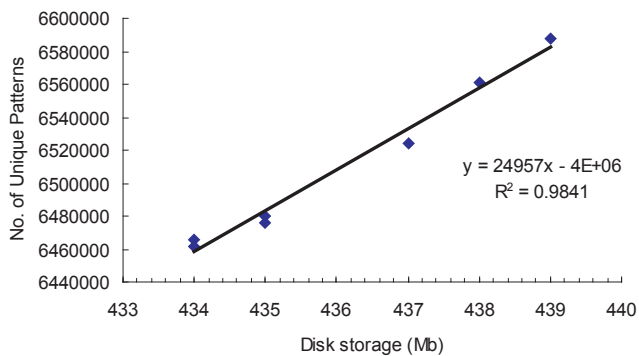


Figure 2: Storage requirements of the unique patterns generated and retained in SynaBASE at various random genome mutations.

There is a large number of patterns of approximately 9-14bp in length stored in SynaBASE as shown in Figure 3. The frequency of patterns below 15bp as shown in Figure 4.

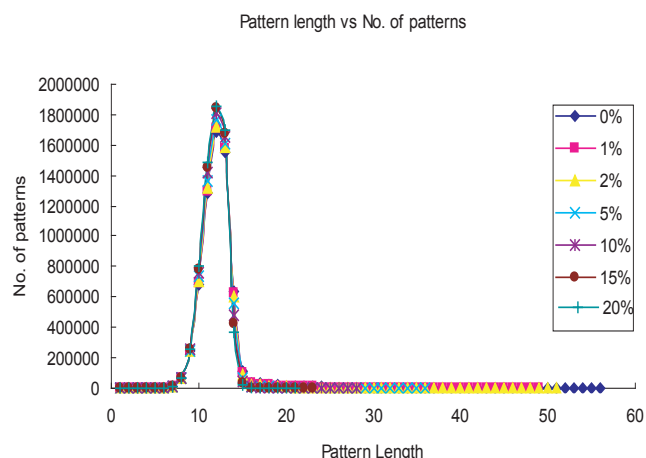


Figure 3: Number of unique patterns (subsequences) generated and stored in SynaBASE based on pattern length. Legend shows the different rates of mutation used.

Pattern length vs Pattern Frequency

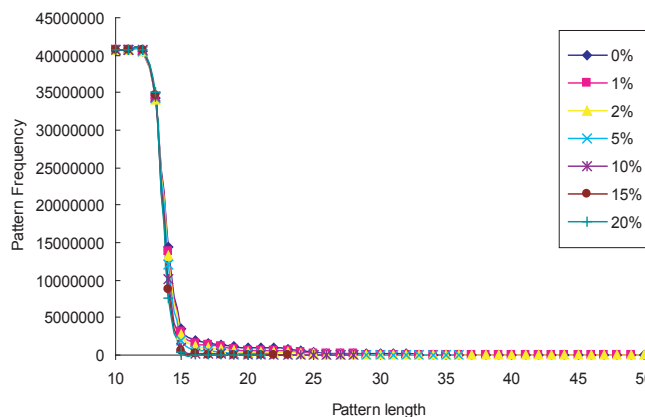


Figure 4: The frequency of patterns (subsequences) occurrence based on pattern length. Legend shows the different rates of mutation used.

Conclusion

It can be inferred that by adjusting the build process when constructing SynaBASE databases; highly efficient scalability can be achieved as multiple copies of highly related genomes are stored as associative subsequences. This paper has demonstrated the enhanced and efficient approach of SynaBASE in managing biological data when compared to conventional databases storing flat files. SynaBASE can be potentially applied in research areas requiring management and analysis of large number of closely related genomes, such as in personalised medicine research, whereby sequences from a population with SNPs can be stored and analysed very quickly.

References

1. Stephens S. M., Chen J. Y., Davidson M. G., Thomas S., and Trute B. M. (2005) Oracle database 10g: a platform for BLAST search and Regular Expression pattern matching in life sciences. *Nucleic Acid Res.* 33, D675-D679.
2. Tan Ka Ju, Ching Soo Meng, Zayed Albertyn (2004) Performance Benchmarking of a Novel Application in Comparative Genomics Using a Structured Network Database of Genome Patterns. Available from www.synamatix.com
3. Zayed I. Albertyn, Arif Anwar, Nataraj Dongre, Johan Poole-Johnson, Ching Soo Meng and Robert G. Hercus (2004) A Revolutionary Application of a Novel Structured Network Database for Genome to Genome Comparisons. Available from www.synamatix.com
4. Wheeler D., Benson D. A., Bryant S., Canese K., Church D. M., Edgar, R., Federhen S., Helmberg W., Kenton D., Khovayko O. et al. (2005) Database resources of the National Center for Biotechnology Information: Update. *Nucleic Acid Res.* 33, D39-D45.
5. Kurtz S., Phillipy A., Delcher, A L., Smoot M., Shumway M., Antonescu C., and Salzberg S L. (2004) Versatile and open software for comparing large genomes. *Genome Biology.* 5(R12), R12.1-R12.9.

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.