A Revolutionary Application of a Novel Structured Network Database for Genome to Genome Comparisons

Zayed I. Albertyn, Arif Anwar, Nataraj Dongre, Johan Poole-Johnson, Ching Soo Meng and Robert G. Hercus.

Key biological data management bottlenecks currently restrict the full exploitation of applications such as iterative, poly-genome level haplotyping, genotyping and multi-genome comparisons. In this study we apply a novel structured database, SynaBASE, to demonstrate genome to genome level analysis. Our findings revealed detection of sequence conservation and duplication at the fraction of the time and storage space taken by conventional methods. SynaBASE achieves this through multiple revolutionary approaches. Firstly, storage demands are minimised, as patterns, not flat files, are stored and structured. Secondly, patterns are extended in a non-redundant fashion until they are unique, enabling true scaling in terms of database size, construction speed and analysis, making the analysis of 100s or 1000s of genomes a realistic possibility. Finally, the significance of patterns, rather than frequency, is determined and quantified, hence both aims of high throughput and output are achieved. We foresee multiple applications for SynaBASE in research focused upon Personalised medicine, Ultra-high-throughput genotypic and sequencing and true poly-genome syntenic analysis.

Introduction

Mammalian genomes are an intense focus of modern biotechnological and pharmaceutical research. Genome sequence comparison forms the basis for understanding gene function and evolution within and between organisms [1]. High-throughput comparative genomics projects are computationally intensive with the rate limiting factors being speed and scalability of algorithms and data complexity [2]. The rapid comparison of newly sequenced genomes against well characterized benchmarks is also becoming an important consideration in target analysis and toxicology studies. In addition, new applications such as personalized medicine/ haplotyping, ultra-high-throughput sequencing/ syntenic analysis and genotyping will require a paradigm shift in how data is stored, structured, analysed and accessed.

The identification of evolutionary conserved sequences between genomes is largely dependent on efficient local/global sequence alignment tools accessing sequence data in flat files or relational databases. We have developed an intelligent sequence database system, SynaBASE™, which has the ability to reconstruct genome-genome alignments from a structured network of sequence patterns, rather than flat files. The system identifies patterns and their significance and intelligently stores and structures relationships between patterns, hence minimizing redundancy and maximizing information content. This seminal approach enables sequence manipulation capabilities such as pairwise alignment, database searching and data mining at a fraction of the time taken by conventional methods and at a scale that was previously not possible. Hence making the routine analysis of 100s of genomes possible.

Methods

To demonstrate the application of SynaBASE an internal genome comparison and visualisation tool, SynaCompare™, was utilized to align each chromosome of the NCBI34 draft of the human genome to its NCBI33 [3] counterpart using the latter genome build as the query database. Software performance was assessed as the query dataset was known to show significant sequence identity with the target database. (An advantage of the system is that sequences need not be masked for repeats before alignment).

Results and Discussion

Database construction time was non-linear and progressively scales lower as more data is added (data not shown). Data access time is largely unaffected by database size as it scales at $\log_n$ base 2, where n is the database or query size.

Regions of high similarity between two sequences and results within the context of pattern frequency were found for all matches in the target database (Figure 1). By examining results with reference to pattern frequency deduction of whether pattern matches between two sequences have a low/high frequency in the target database were determined. The implications are that a simple pairwise comparison allows one to infer whether matches between two genomes are within a set of frequently occurring sequence motifs e.g. simple and complex repeats, or less frequent unique sequence patterns such as coding regions. The application therefore uses sequence information from the structured network database to produce accurate and highly informative results.

Self-comparison of chromosome 19 (Figure 1) shows matching regions colour coded according to their frequency in the human genome (NCBI33 Build). The moderately frequent blue areas in the dot plot highlight areas of sequence duplication/repeats in chromosome 19. The same region (around 20Mbp) matches twice on the same chromosome, indicating that this sequence is not unique in the database and is duplicated on the same chromosome at approximately 58Mbp. Such comparisons may be routinely done within a matter of minutes on a single CPU machine. Figure 2 (overleaf) shows that the largest human chromosome (chromosome 1, 246Mbp) self comparison takes 18.5 minutes.

The most popular alignment tools utilize suffix trees and have been reported to do self comparisons of the human genome in 4.5 days on a single CPU machine [4]. Based on our results of alignment time versus chromosome query length (Figure 2), we estimate that SynaCompare would take a maximum of 43.8 hours (1.8 days) to do a full human genome self comparison of all chromosomes with pattern significance values referenced against the target database. Therefore the application could be used to directly infer whether sequence conservation within/between species lie within coding/noncoding regions.

SynaSuite™ stores biological sequence patterns and their relationships. Therefore accessing precomputed similarity information for alignments significantly increases speed. The results show that it's possible to derive the relative frequency of patterns in an aligned region with reference to the entire human genome database.

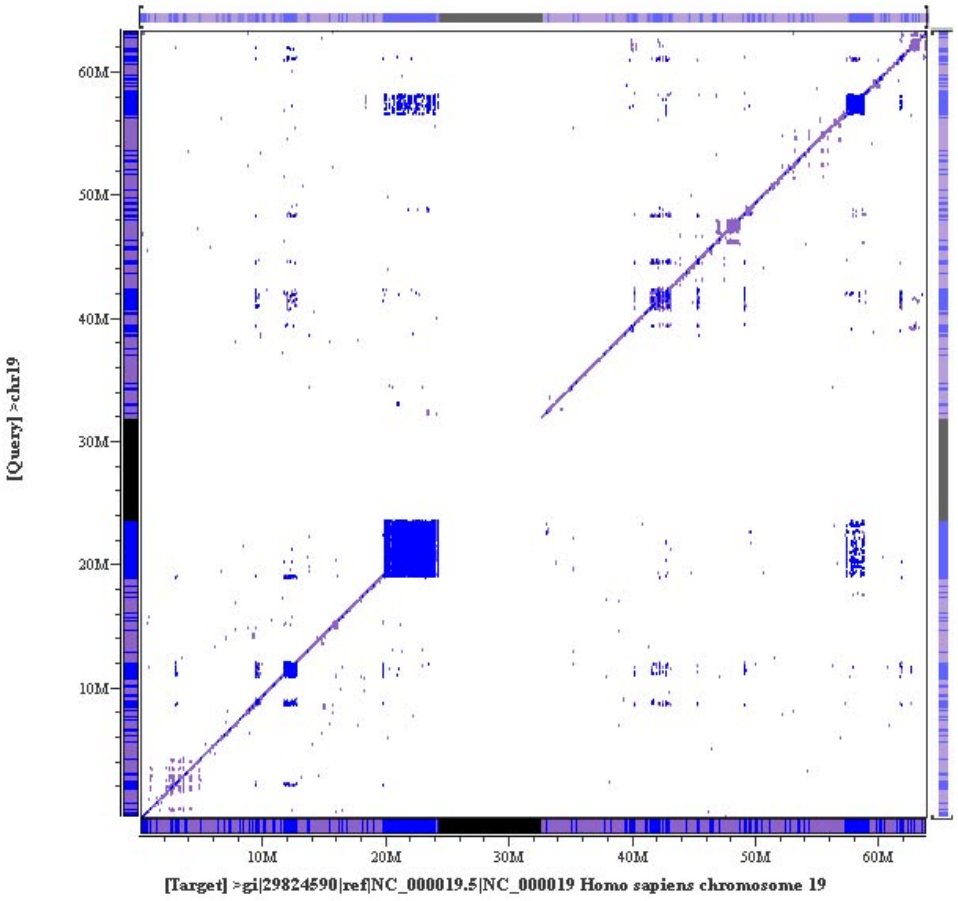

Figure 1. Dot plot of chromosome 19 self comparison between NCBI33 (y-axis) and NCBI34 (x-axis). Pattern frequencies in the human genome are colour coded as medium (blue), low pattern (violet) and no pattern (black). The NCBI34 chromosome 19 query sequence was uploaded to the server to align against the NCBI33 chromosome 19 target using the SynaCompare™ application. Conserved regions are clearly seen in the match diagonal of the plot. Blue patches indicate frequently occurring sequence blocks that are duplicated along the length of the sequence, indicating repeated or duplicated regions.

Conclusions

By comparisons of two versions of the Human Genome in SynaBASE, a pseudo multi-individual comparison was conducted. Significant improvements in speed, database size and scale were demonstrated.

This study demonstrates that large genomes can be efficiently compared at speeds and scales that were previously unattainable. This seminal technology utilises structured patterns, relationships between patterns and their significance to deliver increased qualitative (prediction of significance)

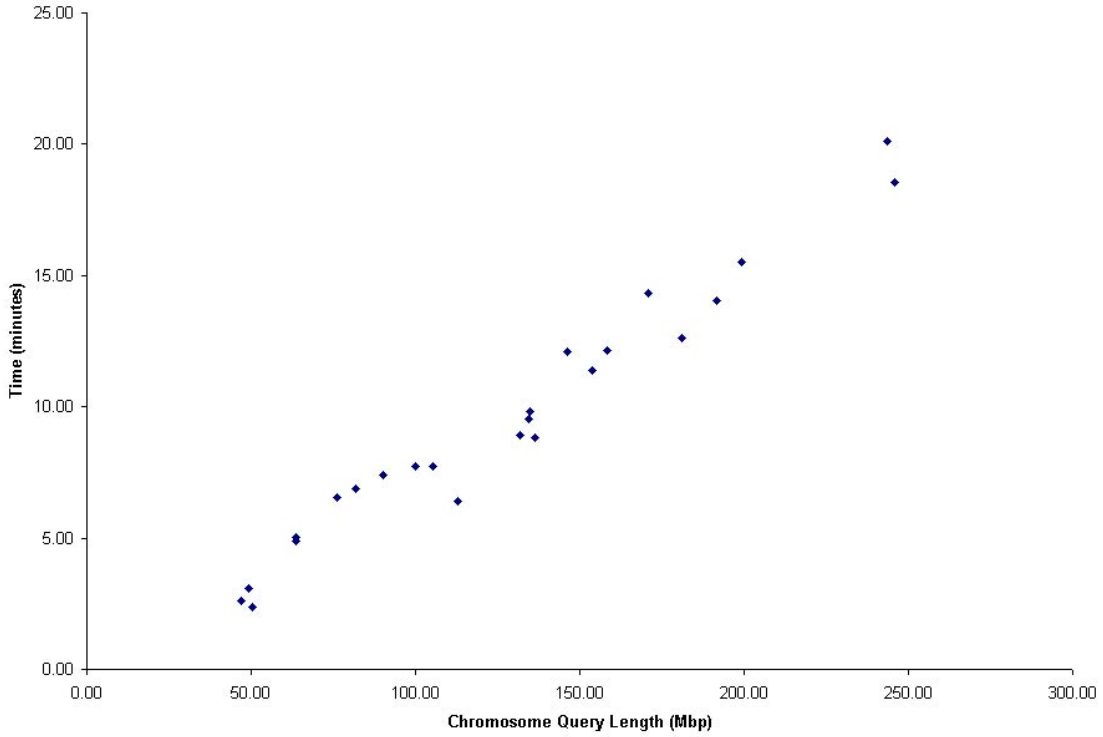

Figure 2. Scatter plot of alignment time versus chromosome query length for the human genome self comparison. Sequences were aligned using SynaCompare on a single 1.3 GHz CPU with 64GB RAM. The average time taken was 9.5 minutes / query.

and quantitative performance (reduction in database size, access time and analysis time).

Unlike conventional schemas, almost limitless scale is achieved as flat files are not stored; instead, redundancy is minimized by storing and structuring patterns and extending them until they become unique. Hence an increase in homogeneous data would not lead to significant increases in database size, as the focus is on the principle of intelligently storing an unlimited amount of sequence data, e.g. multiple copies of the same genome and relationships between them. A natural extension of SynaBASE would be adaptation for high throughput comparisons in the field of personalised medicine, SNP detection and haplotype mapping, and as an overall intelligent archival system capable of keeping pace with high-throughput sequence based applications. All SynaSuite™ applications centre around the concept of finding conserved sequence motifs in the context of the database, e.g. data mining, transcript-to-genome mapping, microarray probe design and multiple sequence alignment.

We are in the process of presenting a detailed study of applications of SynaBASE for multi-genome based haplotype analysis in a subsequent article.

REFERENCES

1. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC. Cross-species sequence comparisons: a review of methods and available resources. Genome Res. 2003 Jan;13(1):1-12. Review.
2. Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin E, Pachter L, Dubchak I. Strategies and tools for whole-genome alignments. Genome Res. 2003 Jan;13(1):73-80.
3. Kent, W.J. and Haussler, D. (2001). Assembly of the Working Draft of the Human Genome with GigAssembler. *Genome Res.* 11:1541-1548
4. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes.Genome Biol. 2004;5(2):R12. Epub 2004 Jan 30.

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