

Inference of Multi-Genome Sequence Conservation from Structured Pattern Information

Zayed I. Albertyn, Tay Liang Chung, Ching Soo Meng, and Robert G. Hercus.

The post genomics era offers new and exciting opportunities for elucidation of the complex processes that govern gene function, genome organization and evolution. With the availability of genomes from different organisms, a key challenge is to present this information in a concise view showing evolutionary conserved regions. SynaBASE™ is the world's first structured pattern database system that stores sequence patterns and their relationships for fast and efficient querying. A multi-species SynaBASE of six syntenic vertebrate species chromosomes was built and queried. By applying a unique attribute of SynaBASE called SIGNIFICANCE, it was possible to deduce evolutionary conserved regions using human chromosome 7 as the reference sequence. Results were compared to phylogenetic conservation scores from the UCSC human genome browser. SynaBASE pattern SIGNIFICANCE matched phylogenetic Hidden Markov Model conservation predictions based on BLASTz genome alignments. By deducing the presence of evolutionary conserved regions from raw sequence data in a single step with SIGNIFICANCE, multi genome SynaBASEs have the capability to be utilised as a powerful platform for novel applications in comparative genomics and molecular evolution.

Introduction

Comparative genomics addresses the problem of inferring an evolutionary relationship between the genomes of two organisms based on sequence alignment of protein-coding sequences, cDNA information (where available) as well as non-coding genomic DNA [1]. A logical extension of this principle would thus be to combine related genome data across multiple species to produce a concise view of sites that have been conserved in these organisms over the course of evolution [2].

A myriad of algorithms such as BLASTz, Pipmaker and Mummer, align entire full-length chromosome sequences to complement views of homology between two organisms [3-5]. Once a series of these alignments across multiple species become available, the consolidated information can produce a multiple sequence alignment. The final result is usually presented in the context of a reference genome, e.g. the human genome, and shows areas of sequence identity, the location of potentially homologous genes or sequence conservation.

An application capable of highlighting sequence conservation between the genomes of 6 different mammalian species was addressed in a single step utilising pattern SIGNIFICANCE (a unique attribute of SynaBASE™). SIGNIFICANCE is defined as the probability of predicting the precursor and

successor characters in a pattern [6]. With the ability to automatically store and learn patterns and their relationships within genome sequence data, SynaBASE is capable of reconstructing a concise view of sequence conservation within the context of any reference sequence.

Methods

Mouse chromosome 6, rat chromosome 4, dog chromosome 14, chimpanzee chromosome 6, chicken chromosome 2 and human chromosome 7 were used as the source sequence data for building a multi-species SynaBASE (MSS)[7-11]. These chromosomes were defined as syntenic by Ensembl v23.34e.1 to human chromosome 7 [12]. A SynaBASE of these six genome sequences was subsequently built and queried for pattern SIGNIFICANCE with reference to human chromosome 7. The same chromosome sequence was also queried against the human, mouse and dog genome versions of SynaBASE.

Results & Discussion

Indexing and constructing all the patterns in a SynaBASE of 6 evolutionary related chromosomes took 101 minutes on a HP® server using a single 1.3GHz Itanium™ Intel CPU.

In the context of related chromosome sequences, the implications are that SynaBASE stores computed information describing these shared patterns [13]. Therefore patterns that are highly probable in a reference query, e.g. Hs chr 7 (with respect to the MSS) represent sequences that are shared or conserved with other patterns in that sequence or patterns from other syntenic regions. SIGNIFICANCE peaks on human chromosome 7 correlate strongly with the UCSC conservation track (available at www.genome.ucsc.edu) in Figure 1. These results show a cross corroboration of genome sequence conservation predicted from raw pattern information in the SynaBASE MSS.

In addition to analysing human chromosome 7 against the MSS, the same 500kb region was mined against SynaBASE versions of the human, mouse and dog genomes (See Figure 2 on the next page). Patterns from the human genome seemed to contribute to the high peaks at 65kb and 115kb that were absent from the SIGNIFICANCE plot in mouse and dog. Analysis of these translated patterns against the SWISS-PROT version of SynaBASE identified patterns with keywords relating to perlecan - a heparin sulphate proteoglycan. These genes are conserved across metazoan life forms [17].

The slight differences between SynaBASE pattern SIGNIFICANCE and the phastCons track are as a result of (a) not all organisms in the UCSC conservation track are represented in the multi-species SynaBASE and (b) different methods were used to produce the final result. However, it is important to note that all that was required to extract this information from the MSS SynaBASE was the quantification of SIGNIFICANCE. As mentioned before, it took 100 minutes to build the MSS and 3.8 seconds on a single CPU server to obtain SIGNIFICANCE results without the need for multiple alignments of whole genome sequences or construction of a Hidden Markov Model.

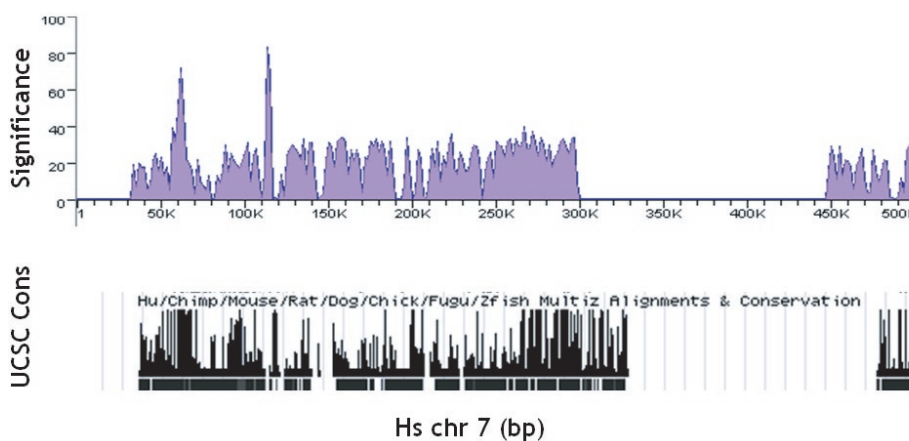


Figure 1. A comparison of SynaBASE multi species genome SIGNIFICANCE and the UCSC genome conservation track plotted for the 1st 500Kb of human chromosome 7. A single query of against SynaBASE produces the top graph in 3.8 seconds whereas the UCSC conservation track is constructed using combination of progressive genome sequence alignments from multiple species and a phylogenetic Hidden Markov Model [14-16].

A query of the 1st 500Kb of human chromosome 7 against the MSS was conducted in 3.8 seconds. The SIGNIFICANCE results correlated almost exactly with the UCSC genome browser view of multiple species conservation computed from human, chimp, mouse, rat, dog, chicken, fugu and zebrafish alignments [16].

The UCSC conservation track is based on the phastCons phylogenetic HMM derived from computed multiple sequence alignments by multiz [14-15]. This track was constructed by “Best-in-genome” pairwise alignments generated for each species using BLASTz. These pairwise alignments were then multiply aligned using multiz, beginning with human-chimp alignments and subsequently adding in mouse, rat, dog, chicken, fugu, and zebrafish. The resulting multiple alignments were then assigned conservation scores by phastCons [15-16].

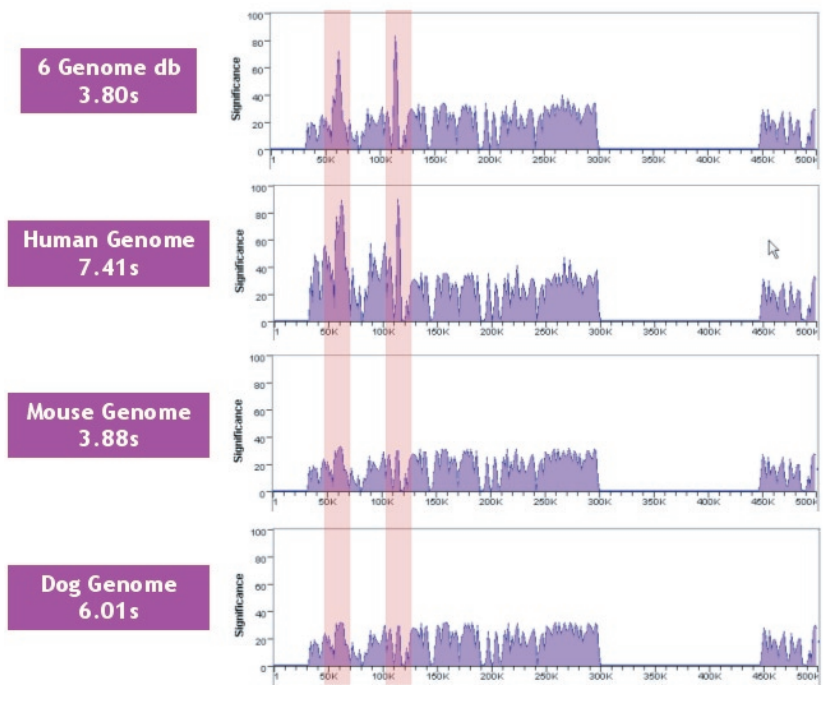


Figure 2. SIGNIFICANCE results for the 1st 500kb of human chromosome 7 queried against the MSS, human, mouse and dog SynaBASE genomes. Regions highlighted in red show the differences in SIGNIFICANCE peaks across the databases.

As more genome data from different organisms become available, it may be more practical to adopt such a system that saves time, computing resources and complements existing methods used in finding evolutionary conserved sequence regions.

Conclusion

Using only SIGNIFICANCE with respect to structured sequence patterns in SynaBASE allows for the inference of sequence conservation across multiple genomes. It provides a quick and easy view of the shared sequence information between organisms without the need for complex multi-step computational pipelines involving days or even weeks on computing clusters.

Conducting such high-level comparative genomics applications on modest hardware would not be possible without the intelligent structuring of pattern data in SynaBASE. As stated in previous work, SynaBASE pattern SIGNIFICANCE may be used in inferring relationships between protein and DNA for a multitude of genomics applications such as protein domain mining, gene prediction, sequence assembly, chromosome-level alignments and microarray probe analysis [6]. Therefore the unique database architecture of SynaBASE coupled with SIGNIFICANCE provides a powerful platform for developing the next generation of sequence analysis tools for applications in biotechnology and biomedical research.

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. Frazer KA, Elnitski L, Church DM, Dubchak I and Hardison RC (2003). Cross-species sequence comparisons: a review of methods and available resources. Genome Res. 13, 1-12. Review.
3. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D and Miller W (2003). Human-mouse alignments with BLASTZ. Genome Res. 13,103-7.
4. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R and Miller W (2000). PipMaker--a web server for aligning two genomic DNA sequences. Genome Res. 10,577-86.
5. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C and Salzberg SL (2004). Versatile and open software for comparing large genomes. Genome Biol. 5, R12.
6. Albertyn ZI, Anwar A, Dongre N, Poole-Johnson, J, Ching SM and Hercus, R.G (2004). A Revolutionary Application of a Novel Structured Network Database for Genome to Genome Comparisons. Available online at www.synamatix.com.
7. Waterston et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-62.
8. Gibbs et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428,493-521.
9. Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL, Pop M, Wang W, Fraser CM and Venter JC (2003). The dog genome: survey sequencing and comparative analysis. Science 301,1898-903.
10. Chimpanzee Genome Sequencing Consortium
11. Kent, W.J. and Haussler, D. (2001). Assembly of the Working Draft of the Human Genome with GigAssembler. Genome Res. 11, 1541-1548.
12. Hubbard et al., 2002. The Ensembl genome database project. Nucleic Acids Research 30, 38-41.
13. Albertyn ZI, Wong CS, Tay LC, Ramachandran M and Hercus RG. A new Approach to Genome-wide Annotation Based upon Calculation of Significance from a Structured Pattern Database. Available online at www.synamatix.com
14. Siepel A and Haussler D (2003). Combining phylogenetic and hidden Markov models in biosequence analysis. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB 2003)*, pp. 277-286.
15. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. Aligning multiple genomic sequences with the threaded blockset aligner (2004). Genome Res. 14,708-15.
16. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. The UCSC Genome Browser Database. Nucleic Acids Res. 31,51-4.
17. Jiang X and Couchman JR (2003). Perlecan and tumor angiogenesis. J Histochem Cytochem. 51,1393-410.

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.