Novel Computational Methods for Large Scale Genome Comparison

Todd J. Treangen[1,2] and Xavier Messeguer[3]

[1]Institut Pasteur, Microbial Evolutionary Genomics, CNRS, URA2171
[2]UPMC Univ Paris 06, Atelier de BioInformatique, Paris, France
[3]Dept. of Software, Universitat Politècnica de Catalunya, Barcelona, Spain

Summary. The current wealth of available genomic data provides an unprecedented opportunity to compare and contrast evolutionary histories of closely and distantly related organisms. The focus of this dissertation is on developing novel algorithms and software for efficient global and local comparison of multiple genomes and the application of these methods for a biologically relevant case study. The thesis research is organized into three successive phases, specifically: (1) multiple genome alignment of closely related species, (2) local multiple alignment of interspersed repeats, and finally, (3) a comparative genomics case study of *Neisseria*. In Phase 1, we first develop an efficient algorithm and data structure for maximal unique match search in multiple genome sequences. We implement these contributions in an interactive multiple genome comparison and alignment tool, M-GCAT, that can efficiently construct multiple genome comparison frameworks in closely related species. In Phase 2, we present a novel computational method for local multiple alignment of interspersed repeats. Our method for local alignment of interspersed repeats features a novel method for gapped extensions of chained seed matches, joining global multiple alignment with a homology test based on a hidden Markov model (HMM). In Phase 3, using the results from the previous two phases we perform a case study of neisserial genomes by tracking the propagation of repeat sequence elements in attempt to understand why the important pathogens of the neisserial group have sexual exchange of DNA by natural transformation. In conclusion, our global contributions in this dissertation have focused on comparing and contrasting evolutionary histories of related organisms via multiple alignment of genomes.

Keywords: Comparative genomics, genome alignment, interspersed repeats, suffix tree, Hidden Markov Model, DNA uptake sequences, homologous recombination.

1 Introduction

Recent advances in sequence data acquisition technology have provided low-cost sequencing and will continue to fuel the rapid growth of molecular sequence databases. According to the Genomes OnLine Database v2.0 [1], as of June 2008 there are a total of 3825 genome sequencing projects, including 827 that are completed and published. This current and future wealth of available genomic data embodies a wide variety of species, spanning all domains of life . This well documented increase in genome sequence data will allow for unprecedented, in depth studies of evolution in closely related species through multiple whole genome comparisons.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 68–73, 2009.
springerlink.com

Sequence alignment has proven to be a versatile tool for comparing closely and distantly related organisms, nucleotide at a time. However, genome sequences can range in size from millions (i.e. bacteria) up to billions (human genome) of nucleotides, requiring extremely efficient computational methods to perform non-trivial sequence comparisons. Traditionally, sequence alignment methods such as local and global alignment have employed dynamic programming for calculating the optimal alignment between a pair of sequences. While progress has been made [2, 3, 4], optimal global multiple alignment under the sum-of-pairs objective function remains intractable[5] with respect to the number of sequences under comparison, and new heuristics are needed to scale with the dramatic increase of available sequenced genomes.

2 Multiple Genome Alignment of Closely Related Species

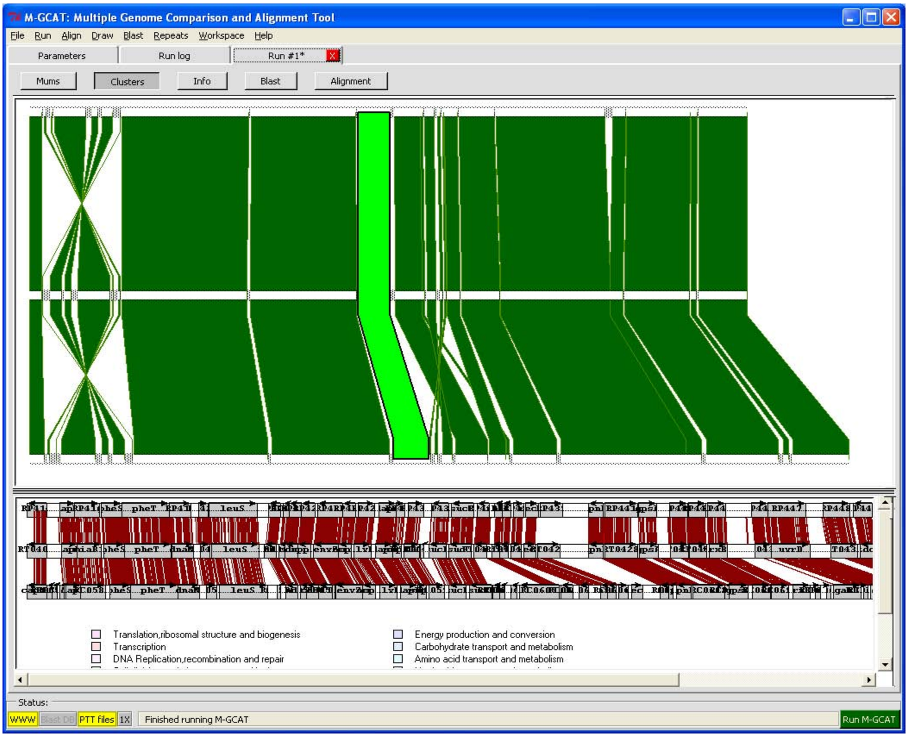
Our methodology for multiple genome alignment follows the idea that the comparison of whole genomes can be efficiently based on detecting unique genomic regions, maximal unique matches (MUMs), which appear only once in each genome. We introduce a Suffix Graph data structure that is a modified version of a directed acyclic graph. The most important feature of our Suffix Graph is that it obtains the most compact representation of a suffix tree so far, while maintaining the topology of the structure. We also present an optimized algorithm for finding unique matches (UMs) and maximal unique matches (MUMs) between multiple DNA sequences using a the Suffix Graph data structure. Our approach for searching for unique substrings($\mathcal{UM}$s) yields significant improvements, in both time and space, over existing methods when dealing with multiple DNA sequences. Specifically, given $S_1, \ldots, S_m$ DNA sequences (where S_1 is the smallest genome), we are able to find the $\mathcal{UM}$s among all of the sequence in linear time $O(|S_1| + \cdots + |S_m|)$ and linear space $O(|S_1|)$ Also, the proposed method only requires 8 bytes/character for sequences $S_2, \ldots, S_m$, to retrieve the positions of the $\mathcal{UM}$s in all sequences [6]. Thus this algorithm has linear space complexity with respect to the smallest sequence and linear time complexity with respect to the sum of the length of all sequences. Using these algorithms & data structures, we have designed and implemented an integrated environment for multiple whole genome comparisons based on our MUM search algorithm. The result is **M**ultiple **G**enome **C**omparison and **A**lignment **T**ool, or **M-GCAT** [7]. M-GCAT is able to compare and identify highly conserved regions in up to 20 closely related bacterial species in minutes on a standard computer, and up to 100 in an hour. M-GCAT also incorporates a novel comparative genomics data visualization interface allowing the user to globally and locally examine and inspect the conserved regions and gene annotations (see Figure 1).

3 Local Multiple Alignment of Interspersed Repeats

During the second phase, we shift the focus to *local* multiple alignment. In this phase, we have developed novel computational methods and software for efficient local multiple sequence alignment of interspersed repeats [8, 9]. Recent advances in sequence data acquisition technology provide low-cost sequencing and will continue to fuel the growth of molecular sequence databases. To cope with advances in data volume, corresponding advances in computational methods are necessary; thus we present an efficient method for local multiple alignment of DNA sequence. Our method for computing local multiple alignments utilizes the MUSCLE multiple alignment algorithm to compute gapped alignments of ungapped multi-match seeds. The method assumes a priori that a fixed number of nucleotides surrounding a seed match are likely to be homologous and, as a result, computes a global multiple alignment on the surrounding region. However, this a priori assumption often proves to be erroneous and results in an alignment of unrelated sequences. In the context of *local* multiple alignment, the fundamental problem with such an approach is that current methods for progressive alignment with iterative refinement compute *global* alignments, i.e. they implicitly assume that input sequences are homologous over their entire length. To resolve the problem, we employ a hidden Markov model able to detect unrelated regions embedded in the global multiple alignment. Unrelated regions are then removed from the alignment and the local-multiple alignment is trimmed to reflect the updated boundaries of homology. We have implemented our method for local multiple alignment of DNA sequence in the `procrastAligner` command-line alignment tool. Experimental results demonstrate that the described method offers a level of alignment accuracy exceeding that of previous methods. Accurately predicting homology boundaries has important implications; for example, tools to build repeat family databases can directly use the alignments without the manual curation required in current approaches and also is likely to aid in the evolutionary analysis of transposon proliferation.

4 Comparative Genomics Case Study of DUS in *Neisseria*

Finally, in this last phase, we employ a comparative genomics approach to study the proliferation of repeat sequence elements in neisserial genomes. Our goal is to understand why the important pathogens of the neisserial group have sexual exchange of DNA by natural transformation. Efficient natural transformation in Neisseria requires the presence of short DNA uptake sequences (DUS), which are highly abundant in their genomes. DUS allow the discrimination between DNA from closely related strains or species and foreign/unrelated DNA. DUS of *Neisseria* spp. is a short signal extending 10 nt, 5'-GCCGTCTGAA-3' [11]. It is present in approximately 2000 copies occupying 1% of the sequenced neisserial genomes. DNA uptake sequences could have proliferated either by selection for transformation or by selfish molecular drive, and through their study we hope to enlighten their evolutionary role. We took advantage of the opportunity provided by the availability of six complete neisserial genomes to globally align the core genome and to define the sets of genes that are ubiquitous and those that were


M-GCAT: Multiple Genome Comparison and Alignment Tool
File Run Align Draw Blast Repeats Workspace Help
Parameters
Run log
Run #1*
Mums
Clusters
Info
Blast
Alignment
□ Translation,ribosomal structure and biogenesis
□ Transcription
□ DNA Replication,recombination and repair
□ Energy production and conversion
□ Carbohydrate transport and metabolism
□ Amino acid transport and metabolism
Status:
WWW
Blast DB
PTT files
1X
Run M-GCAT

Finished running M-GCAT

Fig. 1. Visual representation of a multiple genome comparison of 3 *Rickettsia* genomes inside of the M-GCAT interactive genome comparison viewer. There are 22 highly conserved collinear regions displayed, covering approximately 91% of the total genomic sequence. The region highlighted in green and indicated with the black arrow is one of the 22 regions found to be highly conserved among the 3 closely related species. Directly below the global genome comparison viewer, all annotated genes are displayed horizontally as rectangles. The red vertical lines represent the multiple maximal unique matches common to all 3 genomes.

recently acquired or recently lost in each group. Multiple genome alignments provided a new approach in solving the puzzle of the origin and fate of DUS in these genomes, which helped to elucidate the association between these signals and recombination events [10]. A strong correlation between the average distance between DUS and the length of conversion fragments was found, indicating that the process of transformation is tightly linked to and even shaped by a history of recombination.

5 Conclusion

Due to the great amount of DNA sequences currently available, tools which can efficiently and accurately align and compare multiple genomes are essential for identifying evolutionary patterns. We have proposed novel data structures, algorithms, and software for active areas of research in comparative genomics. Additionally, we have performed an comparative genomics analysis to study the evolutionary role of DNA uptake sequences in six strains of *Neisseria*. The novel algorithmic ideas presented in this doctoral thesis for multiple genome comparison based on global and local alignment allow for multiple genome comparison of organisms at varying evolutionary distances. Our global contributions in this dissertation have focused on comparing and contrasting evolutionary histories of related organisms via their genomes.

Acknowledgments

T.J. Treangen would like to especially thank Eduardo PC Rocha of the Institut Pasteur for his guidance and support throughout the thesis. We would also like to thank Aaron E. Darling, Ole Herman Ambur, Mark A. Ragan, Nicole T. Perna, and Tone Tonjum for their collaboration throughout various phases of the thesis. This work has been supported by the LogicTools project (TIN2004-03382) funded by the Spanish Ministry of Science and Technology and the EU program FEDER and AGAUR Training Grant FI-IQUC-2005.

References

1. Liolos, K., Tavernarakis, N., Hugenholtz, P., Kyrpides, N.: The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Research 34, 332–334 (2006)
2. Edgar, R.: MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32 (2004)
3. Thompson, J.D., Higgins, D.G., Gibson, T.: Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680 (1994)
4. Notredame, C., Higgins, D.G., Heringa, J.: T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of Molecular Biology 302, 205–217 (2000)
5. Wang, L., Jiang, T.: On the complexity of multiple sequence alignment. J. Comput. Biol. 1, 337–348 (1994)
6. Treangen, T.J., Roset, R., Messeguer, X.: Optimized search for common unique substrings, on both forward and reverse strands, in multiple DNA sequences. In: Poster proceedings of the 1st internation conference on Bioinformatics Research and Development BIRD (2007)
7. Treangen, T.J., Messeguer, X.: M-GCAT: Interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. BMC Bioinformatics 7, 433 (2006)
8. Darling, A.E., Treangen, T.J., Zhang, L., Kuiken, C., Messeguer, X., Perna, N.T.: Procrastination leads to efficient filtration for local multiple alignment. In: Bücher, P., Moret, B.M.E. (eds.) WABI 2006. LNCS (LNBI), vol. 4175. Springer, Heidelberg (2006)

9. Treangen, T.J., Darling, A.E., Ragan, M.A., Messeguer, X.: Gapped Extension for Local Multiple Alignment of Interspersed DNA Repeats. In: LNBI proceedings of the International Symposium on Bioinformatics Research and Applications ISBRA (2008)
10. Treangen, T.J., Ambur, O.H., Tonjum, T., Rocha, E.P.C.: The impact of the neisserial DNA uptake sequences on genome evolution and stability. Genome Biology 9(3), R60 (2008)
11. Goodman, S.D., Scocca, J.J.: Factors influencing the specific interaction of Neisseria gonorrhoeae with transforming DNA. J. Bacteriol. 173, 5921–5923 (1991)