Logical and Probabilistic Reasoning for Genomic Rearrangement Detection

Keith Flanagan, Matthew Pocock, Pete Lee, Anil Wipat
University of Newcastle upon Tyne
Newcastle upon Tyne, United Kingdom
keith.flanagan@ncl.ac.uk

Abstract

Genome sequence and other related bioinformatics databases continue to grow at an ever increasing rate. It is becoming increasingly necessary for computational methods to aid biologists in the systematic derivation of knowledge from this data. To facilitate machine-based techniques for deriving biological knowledge from the analysis of bioinformatics data, it is helpful to represent biological concepts and data using formal knowledge representation techniques, such as ontologies.

This paper briefly presents an ontology and combined logical / probabilistic reasoning system that is capable of systematically identifying multiple types of genomic rearrangement events by analysing pairwise comparison data. The resulting comparisons are annotated with terms from the ontology, thus ascribing biological meaning to these regions. This allows the biologist to focus their efforts upon further analysing the regions of interest, while filtering out less interesting regions.

1 Introduction

Bacteria have a remarkable propensity to dynamically modify their own genomes, either by re-arranging already present genetic material, or by gaining genetic material from other organisms in their environment. For instance, redundant or corrupt genes can be deleted [Mira *et al.*, 2001] to enhance efficiency. Multiple copies of genes can be made, allowing a range of related functionalities to be developed. Genes may be inserted into the genome from another organism by 'piggybacking' them in on mobile genetic material [Hentschel and Hacker, 2001]. As the alterations sometimes confer a competitive advantage, these processes are a major factor in bacterial evolution. By determining how bacterial genomes have evolved, it is possible to develop a deeper understanding of how bacteria function and in the ways that the different "species" are related to one another.

Large-scale and systematic (post-)genomics projects are producing huge amounts of data at an ever increasing rate. For instance, DNA sequencing technology has improved to such a degree that it is now possible to make surveys of the genomes of all micro-organisms in an environment within the time and cost comparable to that required in the pre-genomic era to sequence a single gene [Handelsman, 2004]. Some understanding of the function of newly sequenced genes can be gained by comparing them to genes with known functions. The quantity of data produced by comparing all sequences with each other scales in the order of n^2 on the total number of sequences. This directly challenges the traditional paradigm where a biologist manually examines all of the data. This is compounded by the tendency of new analysis tools to produce more data of more types to analyse. Tools such as Act(Artemis Comparison Tool) [Carver *et al.*, 2005] increase the efficiency with which a biologist can search for biologically interesting features embedded in this data, allowing them to screen more data more efficiently. Ultimately though, as a biologist is still required to identify interesting biological features from the raw data, even the vast efficiency gains provided by these tools will be outstripped. Techniques that can systematically mine this data for the "interesting" information will be required to help biologists home in upon relevant biological knowledge.

2 Automated Analysis

In order to automate the mining of comparative genomics data, the necessary domain knowledge required to perform reasoning must first be represented in a machine-readable and machine-interpretable form, such as an ontology. Once an ontology has been developed, it can be manipulated using reasoners such as RACER [Haarslev and Möller, 2001]. Using the rules of the ontology, inferences can be made about marked-up data. For instance, if we know that a particular region in one genome is highly similar to a particular region in another genome, we may infer that this is a 'matching region'. If we are then told that the region in the other genome appears 'backwards' with respect to the first genome, we can further infer that this region is an instance of the more-specific concept 'inversion'.

Unfortunately, due to the nature of biological data, logical reasoning on its own is generally not sufficient. Biological quantities are often not discrete values and there are often significant uncertainties associated with the data. A Bayesian inference approach is suited to reasoning with biological data, especially given the ability of Bayesian methods to deal with partial or uncertain data.

3 Genomic Rearrangements Ontology

The ontology [Flanagan *et al.*, 2004] used by this project is currently under active development. It is split into several modules for performing inferences under different contexts.

1. Physical Components - provides a basic set of terms for describing physical entities, and their compositions, such as genomes, cells and nucleotides.
2. Single Sequence - allows regions of a genome sequence to be annotated with biological meanings, for example, as a gene or a transcription factor binding site.
3. Pair-wise Comparison - describes the similarities and differences between two genome sequences in terms of edits required to transform one into the other (e.g., insertion and deletion events) and also in terms of the biological consequences of the edits (e.g., conservation, inversion or loss of genetic material).
4. Evolutionary History - provides terms for describing how a set of sequences are related to each other through evolutionary trees and horizontal transfer events. Biological implications, such as orthology, paralogy and xenology can then be attached to these sequences.

The terms defined in these modules are used for a variety of purposes, including consistency checking, markup of raw data, automated Java class generation and annotation, labelling the nodes of a Bayesian belief network and associating raw data with Bayesian belief network inputs.

4 Reasoning and Inference

In this work we carry out probabilistic inference to attempt to classify genomic rearrangements by analysing evidence from comparative data sources such as BLAST [Altschul *et al.*, 1990] and IslandPath [Hsiao *et al.*, 2003]. We employ Bayesian belief networks designed to infer which of several hypotheses is probable given evidence about the similarity of a pair of genomic regions. For instance, the analysis of similarity hits from a BLAST report may indicate that a particular region seems to have been inserted or repeated. If IslandPath data indicates that the same region has an abnormal G+C content in comparison to the surrounding sequence, the insertion hypothesis is strengthened [Hentschel and Hacker, 2001], while the repeat hypothesis is weakened.

Bayesian inference networks have been constructed that are able to infer successfully which of several biological interpretations are supported by a combination of BLAST and IslandPath data. The inputs of these networks are associated with bioinformatics concepts, and the outputs with biological concepts. Given bioinformatics data marked up with the bioinformatics concepts, the normal ontological inference rules can be used to apply the networks systematically to all relevant input data, leading to classifications of genomic rearrangements and the genomic features that result.

5 Discussion

Discovery of novel arrangements of features on genome sequences is a thriving and on-going area of research [Nascimento *et al.*, 2004]. It is an essential step in mapping important areas of a genome, in an attempt to determine the function of such regions and/or the processes that formed them.

At the present time, the described inference system can classify regions of interest on a single genome that have a pre-determined definition (for instance, Correia elements [Parkhill *et al.*, 2000]). It is also able to recognise pair-wise rearrangement events between two genomes (for instance, insertions) and in some cases it is able to annotate these rearrangement events biologically (e.g., 'Genomic Island').

All data resulting from the inference networks and algorithms is annotated with both ontology terms and a confidence value. This facilitates further automated or manual analyses using the shared understanding provided by the ontology.

This paper has presented a project that is still very much a "work in progress". While there remain many aspects that still need to be investigated, and further inference networks that need to be developed, the work presented here indicates that automated analyses of this nature may be viable as a method for systematically deriving interesting biological information from bioinformatics data.

References

[Altschul *et al.*, 1990] S. F Altschul *et al.* Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[Carver *et al.*, 2005] Tim J. Carver *et al.* Act: the artemis comparison tool. *Bioinformatics*, page bti553, 2005.

[Flanagan *et al.*, 2004] K. Flanagan *et al.* Ontology for genome comparison and genomic rearrangements. *Comparative and Functional Genomics*, 5, 2004.

[Haarslev and Möller, 2001] Volker Haarslev and Ralf Möller. Racer system description. In R. Gorand A. Leitsch and T. Nipkow, editors, *Proceedings of International Joint Conference on Automated Reasoning, IJCAR'2001*, pages 701–705, June 2001.

[Handelsman, 2004] Jo Handelsman. Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4):669–685, December 2004.

[Hentschel and Hacker, 2001] Ute Hentschel and Jörg Hacker. Pathogenicity islands: the tip of the iceberg. *Microbes and Infection*, 3(7):545–548, June 2001.

[Hsiao *et al.*, 2003] William Hsiao *et al.* Islandpath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, 19(3):418–420, 2003.

[Mira *et al.*, 2001] Alex Mira *et al.* Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, 17(10):589–596, October 2001.

[Nascimento *et al.*, 2004] A.L.T.O. Nascimento *et al.* Genome features of *Leptospira interrogans* serovar copenhageni. *Brazilian Journal of Medical and Biological Research*, 37:459–478, 2004.

[Parkhill *et al.*, 2000] J. Parkhill *et al.* Complete dna sequence of a serogroup a strain of *Neisseria meningitidis* z2491. *Nature*, 404:502–506, March 2000.