Poster A-19

Logical and probabilistic reasoning for the functional annotation of genomic comparisons.

Authors:
Keith Flanagan (*University of Newcastle*)
Matthew Pocock (*University of Newcastle*)
Phil Lord (*University of Newcastle*)
Robert Stevens (*University of Manchester*)
Anil Wipat (*University of Newcastle*)

Short Abstract: We present an approach for the systematic functional analysis of the variations between bacterial genome sequences. Our approach utilises a combination of logical and probabilistic reasoning techniques, backed by an ontology containing the definitions of domain terms. Ontology terms are also used to form complex queries over analysis result sets.

Long Abstract:

Introduction

Bacterial genomes are in a constant state of flux, balancing the pressure for genomic efficiency and maintenance against the need for acquiring new genes and functionality. Genomic plasticity gives rise to a remarkable number of complex rearrangement events, each of which has functional consequences for the organism.

A number of methods can be used for detecting the consequences of genomic rearrangements from whole genome comparisons. However, these either rely on a large amount of human input, often via visualisations of the sequences being compared, or are aimed at reconstructing and estimating the evolutionary divergence between genomic sequences. Given the rate of high-throughput sequencing there is a real need for computational support in detecting and classifying adaptive changes in genome structure. New approaches for the systematic identification and characterisation of the genomic variation that results from niche-adaptation driven genomic plasticity is vital if we are to maximally exploit genomic sequence data and understand the mechanisms of bacterial evolution and genomic diversity. This study attempts to address these problems by developing a computational means of analysing the results of the comparison of small genomes from whole genome alignment algorithms.

We have developed a system that uses a combination of logical and probabilistic reasoning to detect and classify variation in genomic sequences. In essence, we employ a computational methodology to discover and classify relations between bacterial genomes. Importantly, advanced queries can be performed over the resulting data set, enabling irrelevant information to be filtered and features of interest to be easily located. Data are combined from multiple sources to make inferences regarding the location, function and potential origins of particular regions of genome sequences in the context of a pairwise comparison. Domain terms are defined in an ontology, in order to utilise logical reasoning

techniques during the analysis process. Ontology term definitions and the relations between them are also used to form queries over the results of an analysis.

The resulting tool is able to rapidly process and functionally classify the results of pairwise whole genomic alignments, at a nucleotide sequence granularity, in terms of the sequence properties and the features that lie on them. The user may search for regions with desired functional characteristics and regions of potential genomic rearrangement and horizontal gene transfer are identified. Moreover, the tool attempts to classify complex aggregations of features in terms of their order and function to identify composite features such as transposons and pathogenicity islands.

Approach

Our approach is a multi-layered analysis, starting with a basic set of relations that describe the similarities and differences between two genomes. The results of further analysis stages are then layered on top. Each stage of the process provides another source of information, allowing new classifications to be made, or existing classifications to be refined in light of new findings. At each analysis stage, we employ a combination of logical and probabilistic reasoning.

Logical reasoning allows us to query type membership based on ontology definitions. For instance, we make use of "disjoint" restrictions between terms in order to separate incompatible concepts. Given the nature of biological sequence data, there is often no exact Boolean answer to a given classification decision. Where appropriate, we use Bayesian networks to provide probabilistic classifications. Probabilistic reasoning helps in cases where there may be inexact, missing, or potentially incorrect data. Our implementation semantically annotates the input and output nodes of the Bayesian networks with ontology terms. This allows us to perform secondary logical reasoning over the outputs of the network, if required.

Our approach comprises three stages:

Stage 1: Classification of the rearrangements between two genomes. Similarity data from an alignment tool such as BLAST or MUMmer is acquired and very basic set of relations that describe the similarities and differences between two genomes can then be constructed. These relations include terms such as: ‘Similarity’, ‘Presence-Absence’, ‘Absence-Presence’ and ‘Substitution’’. Probabilistic reasoning provided by a Bayesian network is used to classify the types of sequence 'difference', based on the relative locations of 'Similarity' relations.

Stage 2: Raw sequence difference relations discovered in the first stage are enhanced by incorporating biological annotations and sequence composition information. The more information that is available to support this classification, the higher the probability the classification is made. For instance, we can make inferences about whether a particular region is a potential pathogenicity island. A "strong" candidate for a 'pathogenicity island' classification would be an 'Absence-Presence' relation containing a virulence-related gene, flanked by 'Similarity' relations with a G/C content different from the 'Absence-Presence' region.

On completion of the second stage, it is possible to perform queries such as "find 'Absence-Presence' features containing a transposon", or "find functionality contained in genome A, that is not present in genome B".

Stage 3: During this stage of the analysis we incorporate sequence distance metrics, best blast hits, and sequence composition information in an attempt to determine the evolutionary direction of pairwise relations, and to narrow the search space for the origin of horizontally-transferred regions.

Conclusion

We present a computational methodology to assist in the analysis of bacterial genome comparisons. Our approach allows the functional consequences of genomic rearrangements to be classified. Unlike visualisation based techniques, our method is amenable to a rapid, all-against-all approach.

Basic similarities and differences between genome sequences are used as a foundation on which more complex assertions can be placed. Using a combination of logical and probabilistic reasoning over these assertions, it is possible to make inferences, such as the probable function and type of a mobile genetic element. Once comparison features and relations have been identified and classified, it is possible to perform advanced queries over the result-set. Since each classification is an ontology term, the relations between concepts can be leveraged in these queries providing a concise report consisting only of the concepts of interest.