RESEARCH ARTICLE

Probabilistic Topic Modeling using LDA of Taxonomic Structure of Genomic Data

Dnyati S. Randhave[1], Kiran V. Sonkamble[2]

[1]Department of CS & IT, Dr. BAMU, Aurangabad, Maharashtra, India
[2]Department of CS & IT, Dr. BAMU, Aurangabad, Maharashtra, India

[1] dnyati.randhave@gmail.com; [2] kiransonkamble@gmail.com

Abstract— Many data mining techniques have been proposed for fulfilling various knowledge discover tasks in order to achieve the goal of retrieving useful information for user. Various type of probabilistic topic modeling use LDA model of taxonomic structure of genomic data. It can be generated using the techniques, such as that enable homology based approach for further study the functional core. In the planned method, the recognition of key functionality groups is achieved by generative topic modeling. This paper present the research on concept of developing an effective taxonomy model of data analysis, How effectively to extract useful information from unlabeled data. Most of generative the topic model can be used to the model of the large of data information obtained by homology –based approach and study the microbial core. The model considers each sample as a 'document', which has a mixture of functional groups, while each functional group (also known as a 'latent topic') is a weight mixture of species. Therefore, estimating the generative topic model for taxon large quantity data will uncover the distribution over latent functions (latent topic) in each sample.
Keywords— Data mining, Bioinformatics (genome or protein) databases, Language models, Metagenomics

I. INTRODUCTION

The developed sequencing techniques and meta-genomics has dramatically changed the way of genomics data acquiring and analyzing. Next generation methods (such as Roche/454 Sequencing and Illumina Sequencing) are able to extract very large amount (100 ~ 1000 MB) of DNA fragment sequences from an environmental sample (like the ocean, soil and human body) in only a single run (the acquired data is also known a 'meta-genomic data'). The next generation sequencing methods not only provides efficient ways to generate genomic data, but also enable bioinformatics researchers to exploit genomic data [1]from very large amount of uncultured microbial samples instead of from isolated organism and culture. In analysis meta-genomics data

both, taxonomic categorization and functional explanation are difficult because the obtained large amount of DNA remains are always relatively short and may from mixture of very large amount it's usually difficult and to laborites to assemble and annotate the meta-genomic reads. Microbial genome sequencing started in the late 1990s, and now, one decade later, we researchers have access to hundreds of genome sequences.[2] This advance has revolutionized the way research is conducted, and microbial biology has solidly transitioned into the era of post-genomics. Researchers routinely have access to the full catalog of the genes within a genome, thereby eliminating any misperception that genes function in isolation, and thus facilitating the discovery and characterization of genes as parts of genetic networks. This paradigm shift, inspired by DNA sequencing, has led to the development of other high-throughput genomic techniques such as the DNA microarray. The major challenge of the post-genomic era is to interpret the overwhelming amount of data that is now available to all microbial life scientists. Functional genomics and systems biology seek to address this challenge by utilizing enormous genome-scale datasets to predict functional interactions between genes, both within a genetic network and within a genome. Many of the most advanced techniques and algorithms used to make these predictions require a large variety of genomic datasets that are currently only available to the most well-studied model organisms; for the majority of prokaryotic model organisms, the only available types of genomic data exist in the form of a genome sequence and a DNA microarray.

In this review, we will focus on the homology-based approach of functional and taxonomic structure of genomics by describing how the most common types of genomic data can be utilized to construct an experimental pipeline. We will summarize the different forms of genomics datasets, with emphasis on how these data are used to make functional predictions. We also present a case study for the integration of datasets into an experimental pipeline for the identification and verification of functional interactions, including how these datasets can be applied to bacterial systems biology. Finally, we describe the relationship between functional genomics and the evolution of model organism databases, with emphasis on genome annotation. Functional genomics data can be subdivided into sequenced-based and experiment based sets. Sequence-based datasets apply fundamental genetic principles to entire genomes the genome.

For example, the principle of conserved operons can be used to predict the function and functional interactions of unknown open reading frames (ORFs) based on the clustering of ORFs into putative operons. Experiment-based datasets, on the other hand, are adaptations of established molecular biology protocols scaled-up to become 'high throughput. For example, DNA microarrays represent the adaptation of standard hybridization techniques applied on a genomic scale. While there is a clear delineation between sequenced-based and experiment-based datasets, the primary utilization of these datasets is identical: the large-scale prediction of functional interactions. From a systems biology perspective, the de facto structure of a functional interaction can be reduced to a binary construct between two proteins or genes. Although this construct is a drastic over simplification, it allows for both sequence-based and experiment-based datasets be converting and combining into interaction sets. Stripped of all subtlety that exists in within a genome, every pair of genes either interacts (1) or does not (0).Because this binary characterization of an interaction bears little resemblance to reality, any interaction predicted using a functional genomics dataset must be experimentally verified and characterized with respect to spatial and temporal variables; at present, this process still occurs one interaction at a time. Further complicating this process is a lack of consensus regarding the definition of the term "functional interaction". The problem lies in the literal interpretation of the word "interaction"

and, by extension, the purpose of its modifier, "functional". A functional interaction can be narrowly defined as only those proteins that engage in direct physical contact or it can be broadly defined to include all of the proteins that are involved in a response or pathway. Both of these definitions represent extremes and, as such, contradictory examples abound. This lack of consensus does not represent a failure of the scientific community; it is possible that no rigorous definition exists. The difficulty in constructing a universally accepted definition for 'functional interaction' is similar to the current controversy in finding a definition for a bacterial species or a planetary body Rather than commit to specific pairwise functional interactions, ontologies have been developed to categorize proteins into functional groups including Clusters of Orthologous Groups (COGs) Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG), and Protein Families (Pfam). Although every ontology uses an independent schema for their classifications, the fact that a computer script can convert data from one ontology to another is a strong indication of convergence. Ontological analysis provides insight into the relationship that exists between a gene and its genome by reducing the set of possible interaction partners from the whole genome to a subset. In the following section, we characterize the most common types of sequence-based and experiment-based functional.

II. HOMOLOGY – BASED APPROACHES

The critical first step in homology modeling is the identification of the best structure, if indeed any are available. The simplest method of identification relies on serial pairwise sequence alignments aided by database search techniques such as FASTA and BLAST. [2] More sensitive methods based on multiple sequence alignment – of which PSI-BLAST is the most common example – iteratively update their position-specific scoring matrix to successively identify more distantly related homolog's. This family of methods has been shown to produce a larger number of potential and to identify better for sequences that have only distant relationships to any solved structure. Protein threading, also known as fold recognition or 3D-1D alignment, can also be used as a search technique for identifying templates to be used in traditional homology modeling methods. . When performing a BLAST search, a reliable first approach is to identify hits with a sufficiently low *E*-value, which are considered sufficiently close in evolution to make a reliable homology model. Other factors may tip the balance in marginal cases.

A better approach is to submit the primary sequence to fold-recognition servers or, better still, consensus meta-servers which improve upon individual fold-recognition servers by identifying similarities (consensus) among independent predictions. Often several candidate template structures are identified by these approaches. Some methods can generate hybrid models with better accuracy from multiple. Most methods rely on a single template. Therefore, choosing the best template from among the candidates is a key step, and can affect the final accuracy of the structure significantly. This choice is guided by several factors, such as the similarity of the query and template sequences, of their functions, and of the predicted query and observed template secondary structures. Perhaps most importantly, the coverage of the aligned regions: the fraction of the query sequence structure that can be predicted from the template, and the plausibility of the resulting model. Thus, sometimes several homology models are produced for a single query sequence, with the most likely candidate chosen only in the final step. It is possible to use the sequence alignment generated by the database search technique as the basis for the subsequent model production; however, more sophisticated approaches have also been explored. One proposal generates an ensemble of stochastically defined pair wise alignments between the target sequence and a single identified template as a means of exploring "alignment space" in regions of sequence with low local similarity. "Profile-profile" alignments that first generate a

259

sequence profile of the target and systematically compare it to the sequence profiles of solved structures; the coarse-graining inherent in the profile construction is thought to reduce noise introduced by sequence drift in nonessential regions of the sequence. We present a www server for homology-based gene. The user enters a pair of evolutionary related genomic sequences. The homology-based approaches assign meta-genomic reads into NCBI taxonomy based on the alignment between these reads and standard reference (of known species) in standard databases (such as NCBI NR database). One example of homology-based classification approach is the Metagenome Analyzer (a.k.a. MEGAN) [9]. MEGAN is computer software that achieves taxonomical analysis over large databases. It compares DNA fragments against the database of reference sequence, and extracts taxonomical information from the high score BLAST hits. Based on the taxonomical information, the BLAST hits will be matched to different species and strains of the NCBI taxonomy (the algorithm collect all high score BLAST hits and assign taxon ID to each hit based on NCBI taxonomy, the NCBI Taxonomy contains over 460,000 taxa from different taxonomical ranks such as Kingdom, Phylum, Class, Order,...,). After that, the algorithm will look for the lowest common ancestor (LCA) taxon of all those BLAST hits and then assigned the input sequence fragment with that taxon. In practice, BLASTX algorithm will be used to compare all reads against the standard references database (like the NR (non-redundant) protein database from NCBI). Before loading BLASTX hits of a specific meta-genomic fragment to perform a LCA algorithm, the MEGAN algorithm uses a bit-score threshold to limit the number of BLAST hits. It also discards all the "isolated assignment" by looking into the number of hit with regard to each taxon. Despite these efforts, however, the number of resulting hits may still be up to tens of thousands. The LCA algorithm used in MEGAN is able to visualize the hierarchical structure of the phylogenetic tree. However, it should be pointed out that the resulting taxon assignment may only has a limited resolution, as reads with too much BLAST hits may always be assigned to relative high taxon levels such as genus, family instead of species and strains [8].

III. GENERATIVE TOPIC MODEL FOR TAXONOMIC DATA ANALYSIS

In this model working on functional groups in microbial communities, the approach is based on taxon large quantity data acquired from homology based approaches. In this paper the generative topic modelling is an unsupervised probabilistic learning method that is able to extract useful information from unlabelled data. Various domain using a text mining has been developed for generative topic models. The bioinformatics or computational biology various domain generative topic model has been previously used to protein- protein relation from MEDLINE abstract of biomedical literature [2] [5]; we use LDA (Latent Dirichlet Allocation) Model [3]

Latent Dirichlet allocation (LDA) is a generative probabilistic model. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words

LDA assumes the following generative process for each document

1. Choose N ~Poisson (ξ)

2. Choose θ~Dir (α).

3. for each of the N words w_n:

    (a) Choose a topic Z_n~Multinomial (θ).

    (b) Choose a word w_n from p($w_n \mid Z_n$,β),   a multinomial probability conditioned o the topic Z_n

Several simplifying assumptions are made in this basic model, some of which we remove in subsequent sections. First, the dimensionality D of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed known and fixed. Second, the word probabilities are parameterized by a k X V matrix β,

Where β_{ij} =p (w^j =1 | Z^j = 1) which for now we treat as a fixed quantity that is to be estimated.

Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that N is independent of all the other data generating variables (q and **z**). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent development.

IV. FUTURE WORK: ONLINE GENERATIVE TOPIC MODELLING

Online processing of text streams is an essential task of many genuine applications. The objective is to identify the underlying structure of evolving themes in the incoming streams online at the time of their arrival. As many topics tend to reappear consistently in text streams, incorporating semantics that were discovered in previous streams would eventually enhance the identification and description of topics in the future. Latent Dirichlet Allocation (LDA) topic model is a probabilistic technique that has been successfully used to automatically extract the topical or semantic content of documents we investigate the role of past semantics in estimating future topics under the framework of LDA topic modeling, based on the online version implemented in Then, this model is incrementally updated according to the information inferred from the new stream of data with no need to access previous data. Since the proposed approach is totally unsupervised and data-driven, we analyze the effect of different factors that are involved in this model, including the window size, history weight, and equal/decaying history contribution. The proposed approach is evaluated using benchmark datasets. Our experiments show that the embedded semantics from the past improved the quality of the document modeling. We also found that the role of history varies according to the domain and nature of text data.

V. CONCLUSIONS

In this paper, a set of novel probabilistic topic models have been proposed to address challenging issues text mining and bioinformatics studies. The contributions are as follows. In this section, we conduct a generative topic modeling experiment for taxonomic analysis. In Section, we apply the LDA topic model to the taxon abundance data of human gut microbial samples. The human gut microbial community taxon abundance data is generated by [4], which is openly accessible via: http://gutmeta.genomics.org.cn/. According to [4], the Illumina GA reads from human gut microbial samples are firstly assembled into longer contigs. After that, the MetaGene program was used to predict open reading frames (ORFs) from those contigs. The predicted ORFs were then aligned to each other and grouped to a non-redundant gene set. The gene taxonomic assignment is achieved by carrying out BLASTP alignment against the NR database. The taxonomical level of each gene is determined by the lowest common ancestor (LCA). As a result of gene taxonomic assignment, the taxon abundance data for each sample can be produced.

REFERENCES

[1]   Xin Chen, Xiaohua Hu, Y. Lim, Xiajiong Shen, E.K. Park and Gail L. Rosen, Member Exploiting the Functional and Taxonomic Structure of Genomic Data by Probabilistic Topic Modeling pp. 980-99 Issue Date: July 2012

[2]   Huson D., Auch A., Qi J., Schuster S., MEGAN analysis of metagenomic data, Genome research 2007

[3]   Blei D., Ng A.. and Jordan M.. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993 1022,2003

[4]   Qin J et al. A human gut microbial gene catalogue established by met genomic sequencing Nature. 2010 Mar 4; 464(7285):59-65

[5]   Zheng B., Mclean D. C., and Lu X., Identifying biological concepts from a protein-related corpus with a probabilistic topic model. BMC   Bioinformatics, 7, 2006

[6]   Medini, D. "The microbial pan-genome," Current Opinion in Genetics and Development. 2005; 15:6,589-594

[7]   Daniel C. Richter and Daniel H. Huson, Functional Metagenome Analysisusing Gene Ontology (MEGAN 4), Talk at the SIG M3 meeting (ISMB 2009), Stockholm

[8]   Huson D., Richter D., Mitra S., Auch A., Schuster S., Methods for comparative metagenomics, BMC Bioinformatics, Vol. 10, No. Suppl 1. (2009), S12

[9]   Huson D., Auch A., Qi J., Schuster S., MEGAN analysis of metagenomic data, Genome resear