Probabilistic Topic Modeling for Genomic Data Interpretation

Xin Chen1, Xiaohua Hu[1], Xiajiong Shen[3], and Gail Rosen[2]

[1]College of Information Science and Technology, Drexel University, Philadelphia, PA, USA, [2]Department of Electrical and Computer Engineering at Drexel University in Philadelphia, PA, USA, [3]College of Computer and Information Engineering, Henan University, Henan, China

Outlines

What's the major research questions of this study?

- We use our data mining framework to investigate two problems:
  1) How do strains within species share the species core gene and how the non-core genes distribute among each strain?
  2) Do genes with similar codon patterns share similar functional role?

Our research objective:

- We aim to develop a new method that is able to analyze the genome-level composition of DNA sequences, in order to characterize a set of common genomic features shared by the same species and tell their functional roles

Outlines

The research task and approach:

1. Access to the gene region information of reference sequences from the NCBI database.
2. Apply a composition-based approach to break down reference sequences into sub-reads called the ‘N-mer’ and represent the sequences by N-mer features.
3. Study the genome-level statistic patterns (such as latent topics) of the ‘N-mer’ features via topic modeling. Study the mutual information between latent topics and gene regions.

Related topics in this presentation:

- Core and distributed genes
- Structure annotation V.S. functional annotation
- Homology-based V.S. composition-based approaches
- Topic Models

…

Outlines

The research task and approach:

1. Access to the gene region information of reference sequences from the NCBI database.
2. Apply a composition-based approach to break down reference sequences into sub-reads called the ‘N-mer’ and represent the sequences by N-mer features.
3. Study the genome-level statistic patterns (such as latent topics) of the ‘N-mer’ features via topic modeling. Study the mutual information between latent topics and gene regions.

Related topics in this presentation:

- Core and distributed genes
- Structure annotation V.S. functional annotation
- Homology-based V.S. composition-based approaches
- Topic Models

…

Core and distributed genes

- In 2005, Medini et al. described the concept of the pan-genome, which is the “entire genome” of an entire species, instead of each thinking about each strain’s genome individually.
  - For example, strains of E. Coli are hypothesized to only share 1,560 core genes (approximately 1/3 of the genes of any given strain), which means many more are “dispensable”

- The distributed genome hypothesis assume that genetic elements can be classified as two types in each strain – genes shared by all the strains and genes that are “dispensable” (only contained within a subset of the strains).

- For clarity, we refer to those essential to each strain as “*core genes”* and those that can vary from strain to strain as “*distributed genes”*.

The distributed genome hypothesis (Ehrlich, 2008)

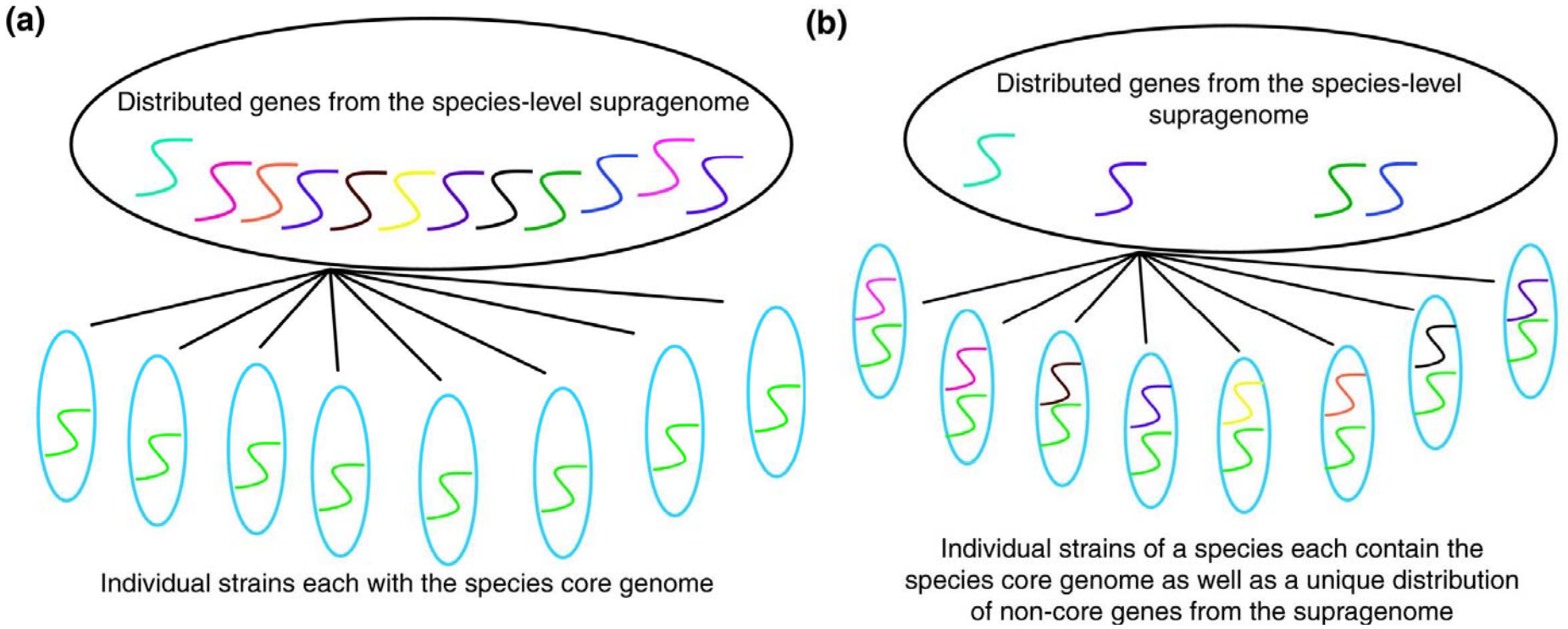


Illustration of the distributed genome hypothesis (Ehrlich, 2008)

Outlines

The research task and approach:

1. Access to the gene region information of reference sequences from the NCBI database.
2. Apply a composition-based approach to break down reference sequences into sub-reads called the ‘N-mer’ and represent the sequences by N-mer features.
3. Study the genome-level statistic patterns (such as latent topics) of the ‘N-mer’ features via topic modeling. Study the mutual information between latent topics and gene regions.

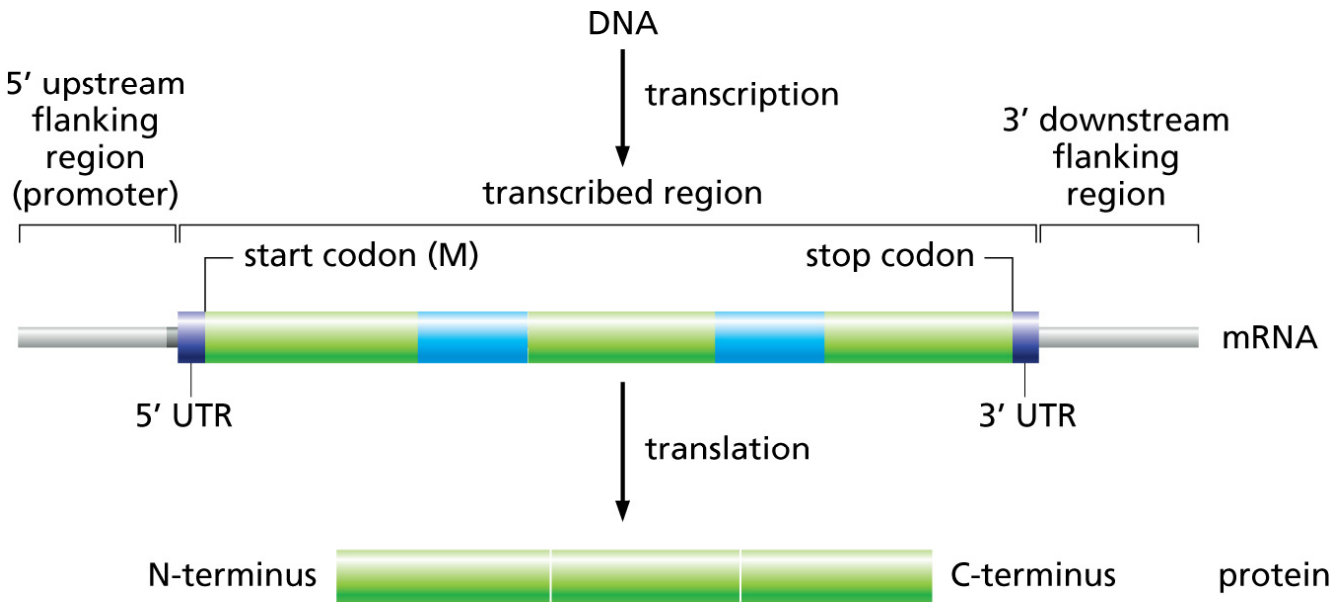
Related topics in this presentation:

- Core and distributed genes
- Structure annotation V.S. functional annotation
- Homology-based V.S. composition-based approaches
- Topic Models

…

Genomic Data Annotation

- Structural annotation
  - Annotating the regions of known open reading frames (ORF's), non-coding genes (rRNA, tRNA, miRNA), Promoters and UTR's in the DNA sequences



Genomic Data Annotation (Continue)

- Functional annotation
  - Uncover the major gene functions related to the genomic sequences
  - Requires explaining the biochemical activity (a.k.a. molecular function) of gene product, identifying the biology process to which the gene or gene product contribute (including information about enzyme, pathway and metabolic capabilities related to the gene).
  - Can be either homology-based or composition-based

Outlines

DREXEL UNIVERSITY

The research task and approach:

1. Access to the gene region information of reference sequences from the NCBI database.
2. Apply a composition-based approach to break down reference sequences into sub-reads called the ‘N-mer’ and represent the sequences by N-mer features.
3. Study the genome-level statistic patterns (such as latent topics) of the ‘N-mer’ features via topic modeling. Study the mutual information between latent topics and gene regions.

Related topics in this presentation:

- Core and distributed genes
- Structure annotation V.S. functional annotation
- Homology-based V.S. composition-based approaches
- Topic Models

…

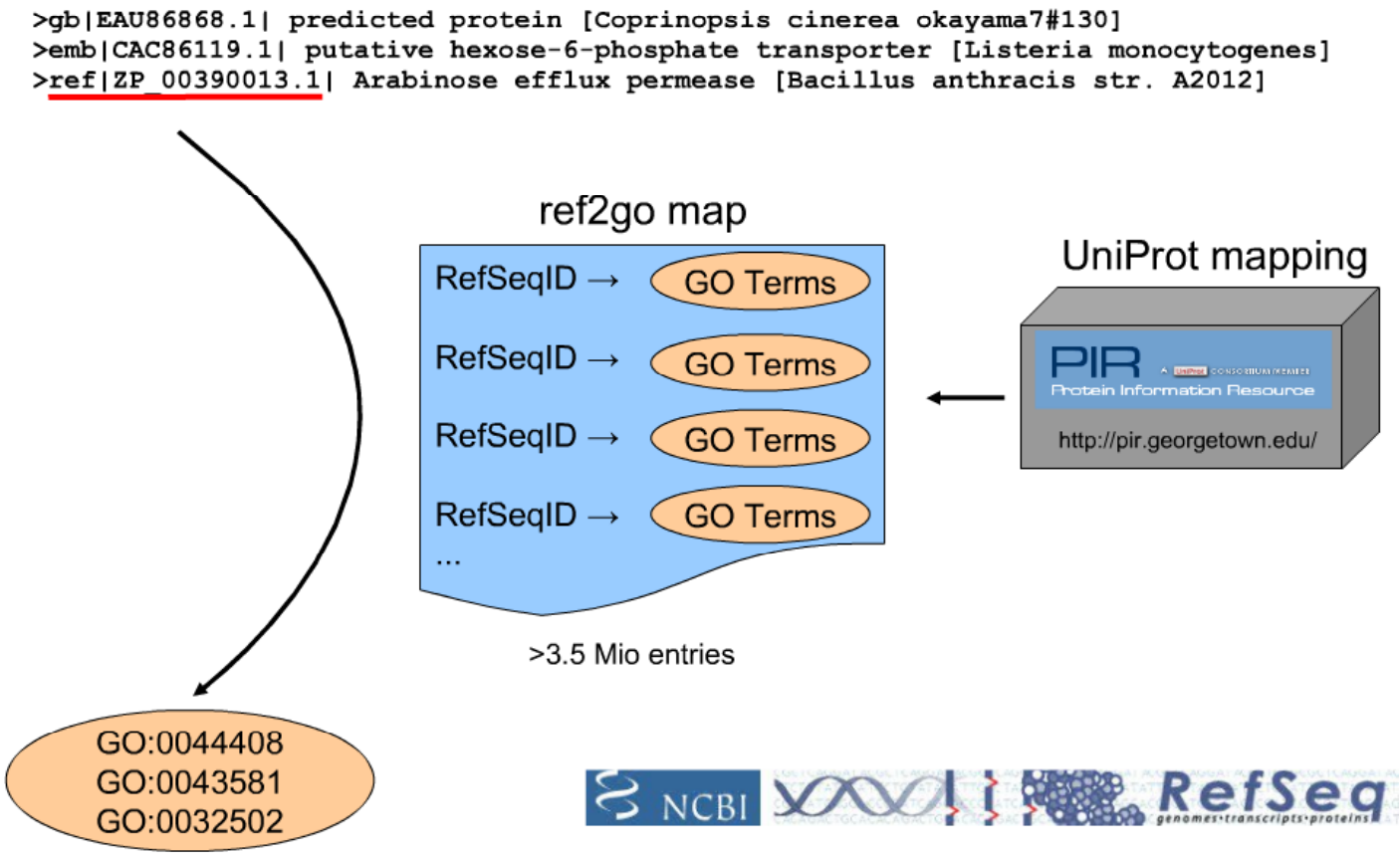
Homology-based functional annotation (Richter and Huson, 2009)

- Homology-based approach has been recently introduced to achieve functional annotation for metagenomic reads (Richter and Huson, 2009).

- The framework begins with a homology based BLASTX algorithm to match the metagenomic fragments against the reference sequences in NCBI database.

- The BLASTX hits will associate fragments with related protein ID and gene names. After that, with the help of the Gene Ontology (GO) database to refer associated gene names to corresponding GO terms, thus provides an overview of gene function and products for metagenomic fragments.

Homology-based functional annotation (Richter and Huson, 2009)



GO terms obtained from database identifier mapping (Richter and Huson, 2009)

The Problems with Homology-based Functional Annotation Methods

1. Homology-based approaches very much reply on the result of local sequence alignment (such as BLAST and BLASTX) to the known open reading frames (ORF).
   - The BLAST-like local alignment may either return hundreds of hits, or return no hits, depending on the threshold of E-value used. In the latter case, the current methods are unable to provide any functional annotation. In the former case, it usually lacks of a proper tie-breaker to further reduce the hits, which makes the functional annotation some how ambiguous (with hundreds of probable explanation)
2. The homology-based functional annotation methods did not provide any insight about the “major” functional capabilities of genomes (like which gene functions are more commonly shared by strains from the same species), as there is no priority for the annotated GO terms.

Composition-based functional annotation

- Recently, the composition-based approaches, which break down the DNA sequences into sub-reads named as ‘N-mers’, have achieved good results in many genomic data analysis tasks (such as genome classification).

- When considering the genome file as a document, the ‘N-mers’ can to some extend be considered as a kind of ‘code words’ that compose a genome fragment (we may consider the A,T,C,G nucleotides as letters, so N-mers bear an analogy with N-letter words)

The analogy between text documents and genome sequences

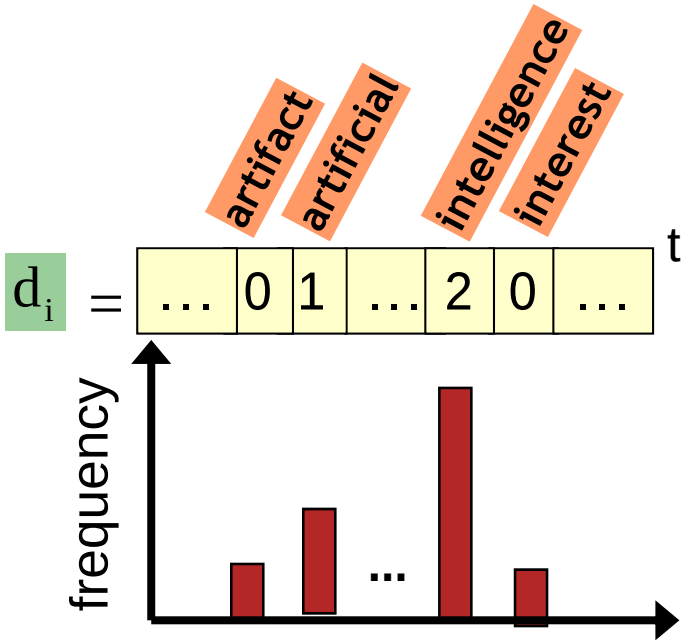
D = Document collection

W = Lexicon/Vocabulary

intelligence w_j

Texas Instruments said it has developed the first 32-bit computer chip designed specifically for artificial intelligence applications [...]

d_i

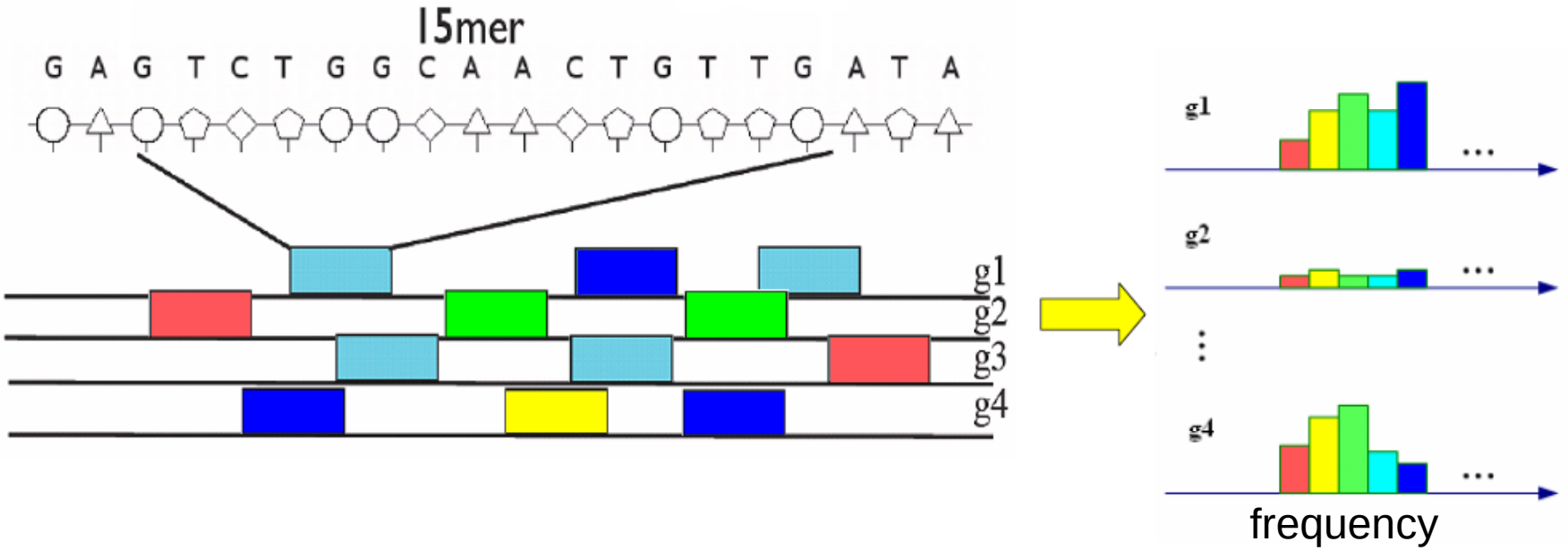


Document-Term Matrix

D	w_1	...	w_j	...	w_J
d_1					
...			...		
d_i		...	$n(d_i, w_j)$	...	
...			...		
d_I					

The analogy between text documents and genome sequences (continued)

- Let’s consider genome sequences as documents, the A,T,C,G nucleotides as letters, while N-mers as N-letter words:

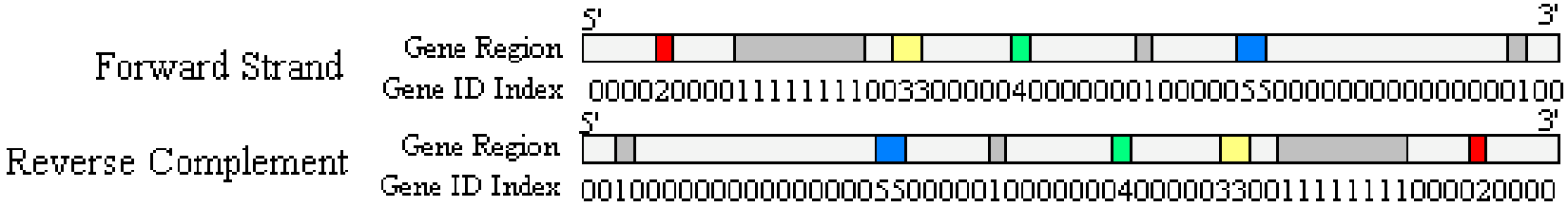


Gene Region Tagging for N-mer Sequence

- We use the “BioJava” package to acquire gene region information from the NCBI database using GenBank accession number s.
- By matching each N-mer location against the gene regions, we tag each N-mer feature with corresponding gene ID.

Gene Region

Gene ID	Gene Name	from..to	Length	Reverse Complement
2	aroH	573..701	129	Y
1	N/A	1185..2177	993	N
3	sppA	2376..2621	246	N
4	nblB	3306..3452	147	Y
1	N/A	4269..4379	111	N
5	typA	5037..5255	219	N
1	N/A	7113..7265	153	N



Outlines

The research task and approach:

1. Access to the gene region information of reference sequences from the NCBI database.
2. Apply a composition-based approach to break down reference sequences into sub-reads called the ‘N-mer’ and represent the sequences by N-mer features.
3. Study the genome-level statistic patterns (such as latent topics) of the ‘N-mer’ features via topic modeling. Study the mutual information between latent topics and gene regions.

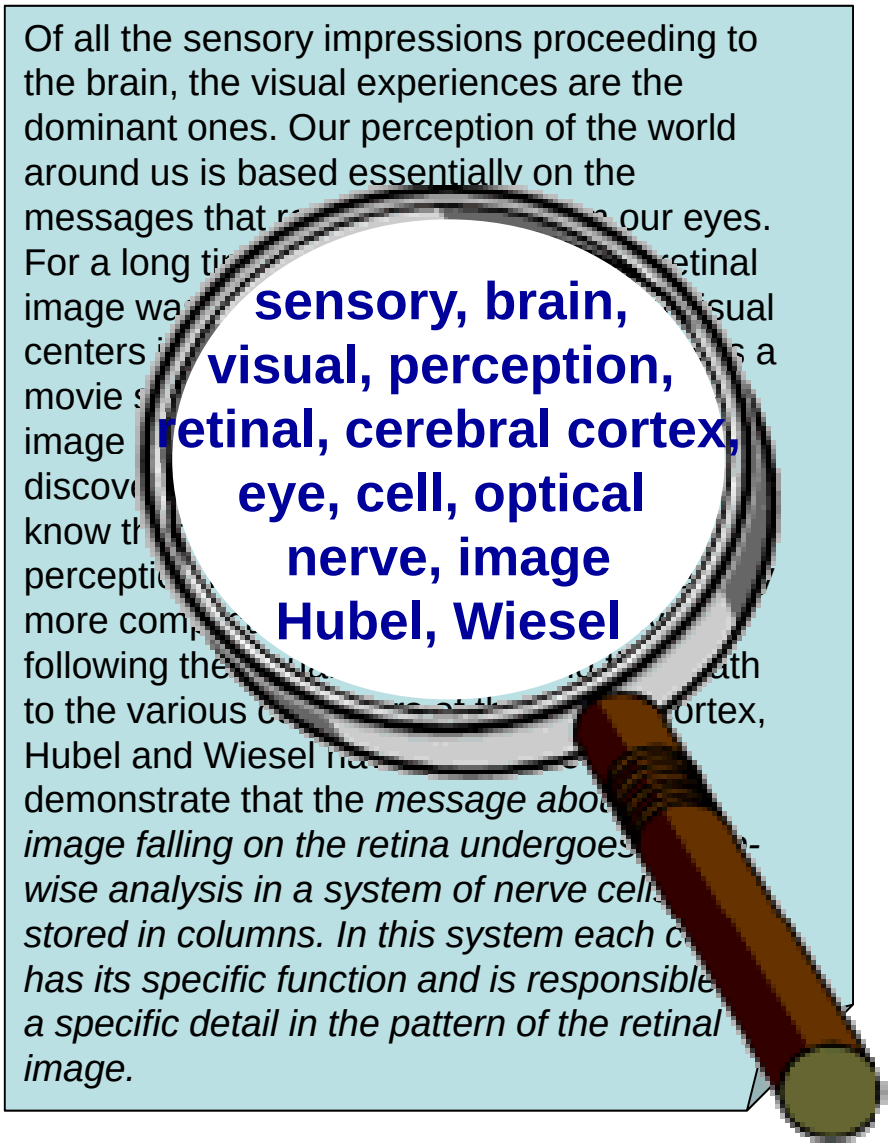
Related topics in this presentation:

- Core and distributed genes
- Structure annotation V.S. functional annotation
- Homology-based V.S. composition-based approaches
- Topic Models

…

Topic Modeling - Intuitive

- Intuitive
  - Assume the data we see is generated by some parameterized random process.
  - Learn the parameters that best explain the data.
  - Use the model to predict (infer) new data, based on data seen so far.

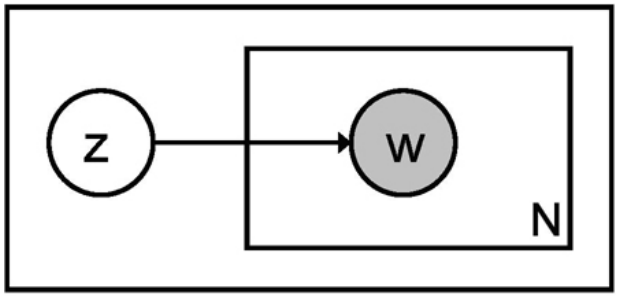


Notations

- Word
  - Basic unit.
  - Item from a vocabulary indexed by {1, . . . ,V}.
- Document
  - Sequence of N words, denoted by w = (w1,w2, . . . ,wN).
- Collection
  - A total of D documents, denoted by C = {w1,w2, . . . ,wD}.
- Topic
  - Denoted by z, the total number is K.
  - Each topic has its unique word distribution p(w|z)

Background & Existing Techniques of Generative Latent Topic Models

- The Naïve Bayesian model



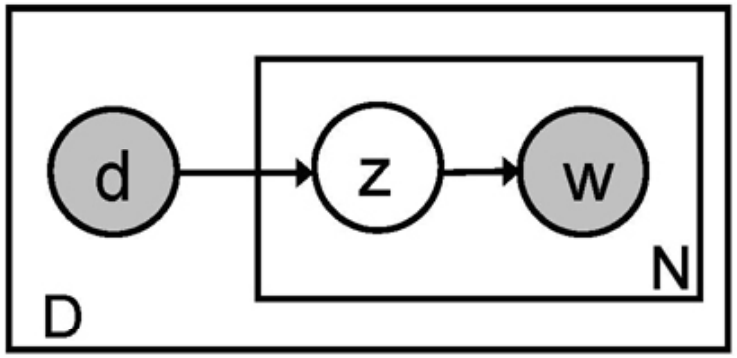
Likelihood of word w given topic z

$$z^* = \arg\max p(z \mid w) \propto p(z)\,p(w \mid z)$$

Word-Topic decision

Prior Probability of Topic z

- The probabilistic latent semantic indexing (PLSI) model



PLSI Model (Hoffman, 2001)

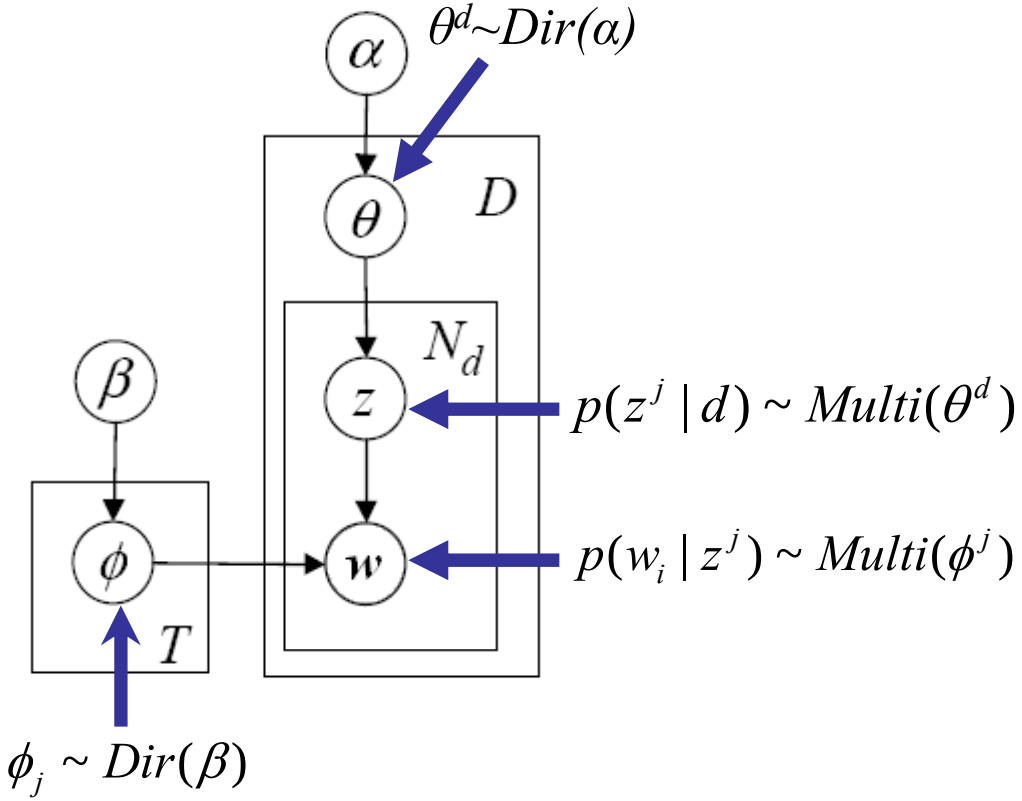
Assumption:

Each document has a mixture of *k* topics.

Fitting the model involves:

Estimating the topic specific word distributions $p(w_i|z_k)$ and document specific topic distributions $p(z_k|d_j)$ from the corpse via maximum likelihood estimation (MLE).

Latent Dirichlet Allocation (LDA) Model (Blei, 2003)



$$p(z_{wi} = j \mid w_i, \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) \propto \frac{\beta + n^{wi}_{-i,j}}{W\beta + n^{\cdot}_{-i,j}} \cdot \frac{\alpha + n^{d}_{-i,j}}{T\alpha + n^{d}_{-i,\cdot}}$$

- In PLSI model, the topic mixture probability $p(z_k|d_j)$ for documents are fixed once the model is estimated. For new coming document, the model needed to be re-estimated. Thus it is not scalable.
- The LDA model treats the probability of latent topics for each document $p(z|d)$ and the conditional probability of words for each latent topic $p(w|z)$ as latent random variables which are subject to change when new document comes.

Statistical relationships of words and topics

Top words from some of the $p(w|z)$

"Arts"	"Budgets"	"Children"	"Education"
new	million	children	school
film	tax	women	students
show	program	people	schools
music	budget	child	education
movie	billion	years	teachers
play	federal	families	high
musical	year	work	public

An example of topic assignment to words

Inference on a held-out document

Topics: "Arts", "Budgets", "Children", "Education".

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants.

Experimental Results and Discussions

- In our experiment, we deal with the problem of uncovering genome-level composition of N-mer latent patterns and explain the functional role of different components. Also, we hope to identify both core genes (genes shared by all the strains) and distributed genes (genes vary from strain to strain).

- We apply the LDA topic model to N-mer sequence data of 635 genomes acquired from the NCBI database. On the convergence of the topic modeling process, each N-mer will be assigned a topic label, from which we will be able to tell the topic-level distribution of N-mer as well as the genome-level distribution of topics

Illustration of top-ranked 9-mers for latent topics learnt by the LDA model

Topic 1	p(Nmer|topic)	Topic 2	p(Nmer|topic)	Topic 3…
GGCGTCGAG	2.19E-04	TCTCGGCAA	2.50E-04	…
GCGATTGCC	1.81E-04	GCCTGCGCG	2.24E-04	…
GGATGCGGC	1.76E-04	GGCACTGGT	1.83E-04	…
CGCCCAGGA	1.67E-04	GGATTATTA	1.69E-04	…
GCGCCGTGG	1.60E-04	GAAGTGGCG	1.63E-04	…
GTTTTATTA	1.57E-04	CGGCTGTTC	1.61E-04	…
TGTCGTGGT	1.49E-04	GCGGCTCAA	1.57E-04	…
CCGGAAGTT	1.49E-04	GGGCGAAGT	1.53E-04	…

Top-ranked latent topics for E. Coli genomes

Genome (Strain) Name	Top-ranked Latent Topics ID
Escherichia_coli_536	79, 69, 61, 68, 92, 87
Escherichia_coli_CFT073	67, 51, 87, 93, 89, 57
Escherichia_coli_HS	93, 27, 94, 58, 95, 67
Escherichia_coli_O157H7	54, 95, 99, 69, 46, 65
Escherichia_coli_UTI89	44, 54, 58, 92, 79, 90
Escherichia_coli_APEC_O1	51, 54, 77, 93, 97, 90
Escherichia_coli_E	62, 97, 43, 94, 75, 92
Escherichia_coli_K12	16, 49, 93, 95, 74, 83
Escherichia_coli_O157H7_EDL933	27, 86, 49, 52, 55, 94
Escherichia_coli_W3110	29, 46, 58, 94, 87, 73

Top-ranked latent topics for P. Marinus genomes

Genome (Strain) Name	Top-ranked Latent Topic ID
Prochlorococcus_marinus_AS9601	3 ,36 ,91 ,58 ,93 ,71
Prochlorococcus_marinus_MED4	3 ,36 ,79 ,91 ,93 ,4
Prochlorococcus_marinus_MIT_9211	12 ,41 ,62 ,74 ,94 ,96
Prochlorococcus_marinus_MIT_9301	3 ,36 ,79 ,91 ,93 ,71
Prochlorococcus_marinus_MIT_9312	3 ,36 ,57 ,79 ,91 ,93
Prochlorococcus_marinus_NATL	3 ,36 ,91 ,93 ,82 ,71
Prochlorococcus_marinus_CCMP1375	82 ,91 ,36 ,81 ,45 ,69
Prochlorococcus_marinus_MIT9313	0 ,89 ,11 ,19 ,68 ,64
Prochlorococcus_marinus_MIT_9215	3 ,36 ,79 ,91 ,93 ,22
Prochlorococcus_marinus_MIT_9303	0 ,89 ,11 ,19 ,68 ,64
Prochlorococcus_marinus_MIT_9515	3 ,36 ,91 ,58 ,93 ,22
Prochlorococcus_marinus_NATL	3 ,36 ,91 ,93 ,71 ,81

Commonly shared top-ranked latent topics for both E. Coli and P. marinus genome sets

Genome Set	Topic Ranking	Topic ID
E. Coli	**Top 40**	1,12,14
	Top 50	1,10,12,14,26,30,63,80,91
P. Marinus	**Top 10**	13,60
	Top 20	8,13,19,33,37,42,47,60,64,68,70,72
	Top 30	8,10,11,13,19,23,24,33,37,42,46,47,60,64,68,70,72,75

- The E. Coli genomes are really diverse, as they rarely share common latent topics among their top-ranked topics. On the contrary, P. Marinus, another genome set we studied, have many common latent topics shared among its different strains. This result suggests that the E. Coli species has more distributed genes than core genes (which may further indicate that E. Coli has experience massive gene loss and gene gain which induce large intra-species genomes variation).

Illustration of gene regions and their most relevant latent topics in P. Marinus genomes

- In order to fully understand the functional roles of uncovered genomic pattern (like core genomes and distributed genomes), it is of great importance to study the relationship between latent topics and gene regions and give it a biological explanation.

Gene Region	Topic	MI_value	Topic	MI_value	Topic	MI_value
Non-Gene	**3**	0.133928	**36**	0.098731	**26**	0.088407
Unnamed Gene	**0**	0.040136	**8**	0.036109	**70**	0.035463
bioF	**0**	0.294652	**8**	0.286716	**60**	0.21943
proC	**70**	0.265276	**19**	0.235419	**0**	0.232178
...	**...**	**...**	**...**	**...**	**...**	**...**

Illustration of detailed gene function information (enzyme and pathway information, metabolic capabilities) acquired from MetaCyc database

- We exploited the BioCyc (http://biocyc.org/), an openly available, highly accurate metabolite pathway and enzyme database, to provide hierarchical functional annotations for gene regions, which involves enzyme and pathway information as well as the metabolic capabilities.

gene_commonName: gldA
protein_commonName: putative glycerol dehydrogenase
protein_type: Polypeptides
enzrxn_commonName: putative glycerol dehydrogenase
enzrxn_type: Enzymatic-Reactions
reaction_commonName: Glycerol dehydrogenase
reaction_type: EC-1.1.1, Small-Molecule-Reactions
reaction_left: GLYCEROL+NAD
reaction_right: PROTON+DIHYDROXYACETONE+NADH
pathway_commonName: glycerol degradation V
pathway_type: GLYCEROL-DEG
pathway_Comment: Glycerol dissimilation in |FRAME: TAX-83333| is usually initiated by the ATP-dependent |FRAME: GLYCEROL-KIN-CPLX| (encoded by |FRAME: EG10398|), which phosphorylates glycerol to |FRAME: GLYCEROL-3P|. However, upon inactivation of the kinase, it may be replaced by the |FRAME: EG11904| |FRAME: NAD|-linked |FRAME: GLYCDEH-CPLX| |CITS: [6183251]|. This enzyme is cryptic in the wild type, and is only activated by mutation. It exhibits broad substrate specificity (it has a lower Km value for |FRAME: PROPANE-1-2-DIOL| than for |FRAME: GLYCEROL|) and its true physiological role remains uncertain |CITS:[8265357] [6361270]|.

Major gene functions that are most relevant to the commonly shared top-ranked latent topics

Organism	Notes	Topic/Gene	Interesting Similar Keys of Genes Involved
E. Coli	Top 40	Topic 1	Binding, metabolic process, cytoplasm
	Top 50	Topic 26	Binding,cytoplasm, **ATP binding, nucleotide binding, transferase**, metabolic process
		Topic 63	Cytoplasm, **ATP binding, nucleotide binding, ligase activity**, catalytic activity
P. marinus	Top 10	Topic 13	Metal ion binding, oxidation reduction
		Topic 60	Synthesize sugar
	Top 20	Topic 72	Ammonia production/transport
	Paired	hisS	Histidyl-tRNA synthetase and ligase, tRna-charging reactions and pathways
		hisZ	
	Paired	bioF	Reductase
		proC	
	Paired	dfp	Putative enzymes
		ctaC	

- The result provides us an insight into the functional role of the core genome. It also shows that gene pairs relevant to the same latent topics also share some common gene functions, which indicates that the uncovered latent topics are biological informative and useful to the interpretation

Conclusions

- A probabilistic topic modeling method is introduced to interpret the genome-level composition of DNA sequences, in which the concurrence patterns of N-mer features across the whole genome set are modeled as latent topics.

- We show that the proposed probabilistic topic modeling algorithm is capable of characterizing core and distributed genes within a species. It also provides new insights about making genome-level compositions associated with their functional roles.

Questions?

Backup Slides

LDA Model Estimation - Gibbs Sampling Monte Carlo process (Griffiths, 2004)

Probability of a topic being assigned to a word given other observations:

$$p(z_{wi} = j \mid w_i, \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) \propto p(w_i \mid z_{wi} = j, \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) \cdot p(z = j \mid \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}})$$

$$p(w_i \mid z_{wi} = j, \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) = \int p(w_i \mid z = j, \varphi^j, \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) p(\varphi^j \mid \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) d\varphi^j = \frac{\beta + n_{-i,j}^{wi}}{W\beta + n_{-i,j}^{\cdot}}$$

$$p(z = j \mid \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) = \int p(z = j \mid \theta^d) \cdot p(\theta^d \mid \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) d\theta^d = \frac{\alpha + n_{-i,j}^{d}}{T\alpha + n_{-i,\cdot}^{d}}$$

$$p(w_i \mid z = j, \varphi^j, \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) = \varphi^j$$

$$p(\varphi^j \mid \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) \propto p(\mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}} \mid \varphi^j) \cdot p(\varphi^j)$$

in which $p(\mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}} \mid \varphi^j) \sim Multi(\varphi^j)$

and $p(\varphi^j) \sim Dir(\beta)$. **It follows that**

$$p(\varphi^j \mid \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) \sim Dir(\beta + n_{-i,j}^{wi})$$

$$p(\theta^d \mid \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) \propto p(\mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}} \mid \theta^d) \cdot p(\theta^d)$$

Since $p(\mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}} \mid \theta^d) \sim Multi(\theta^d)$

and $p(\theta^d) \sim Dir(\alpha)$

We have $p(\theta^d \mid \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}) \sim Dir(\alpha + n_{-i,j}^{d})$

Mote-Carlo process

- Given the word-topic posterior probability, the Monte Carlo process becomes really straightforward, which is similar to throwing dice (given the probability of each facet to appear) to determine the assignment of topics to each words for the next round.

Given probability for each word:

$$p(z_{wi} = j \mid w_i, \mathbf{w}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{wi}}), j = 1...K$$

New topic assignment for each word.