Latent Dirichlet Allocation

Alberto Bietti

Trop d'information

Topic modeling

- Découvrir la structure thématique cachée dans chaque document d'une archive
- Permet:
  - Organisation automatique des documents par thème
  - Compréhension d'un document
  - Recherche
  - Résumé

Découvrir les thèmes d'un corpus de documents

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Annotation d'images

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE

SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

Exploration de thèmes

Mais aussi…

- Trouver des relations entre les thèmes
- Hiérarchies de thèmes
- Evolution des thèmes au cours du temps
- Prédiction de liens entre articles
- *Collaborative filtering* (ex: recommandation de films)

Latent Dirichlet Allocation (LDA)

- Modèle probabiliste génératif d'un corpus de documents
- Modèle Bayésien hiérarchique à 3 niveaux
- Pour chaque document:
  - On choisit un mélange de thèmes (*topic mixture*)
  - Pour chaque mot, on choisit un thème dans ce mélange
  - Et on choisit un mot depuis ce thème
- Apprentissage de ces caractéristiques par inférence

Modèle

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

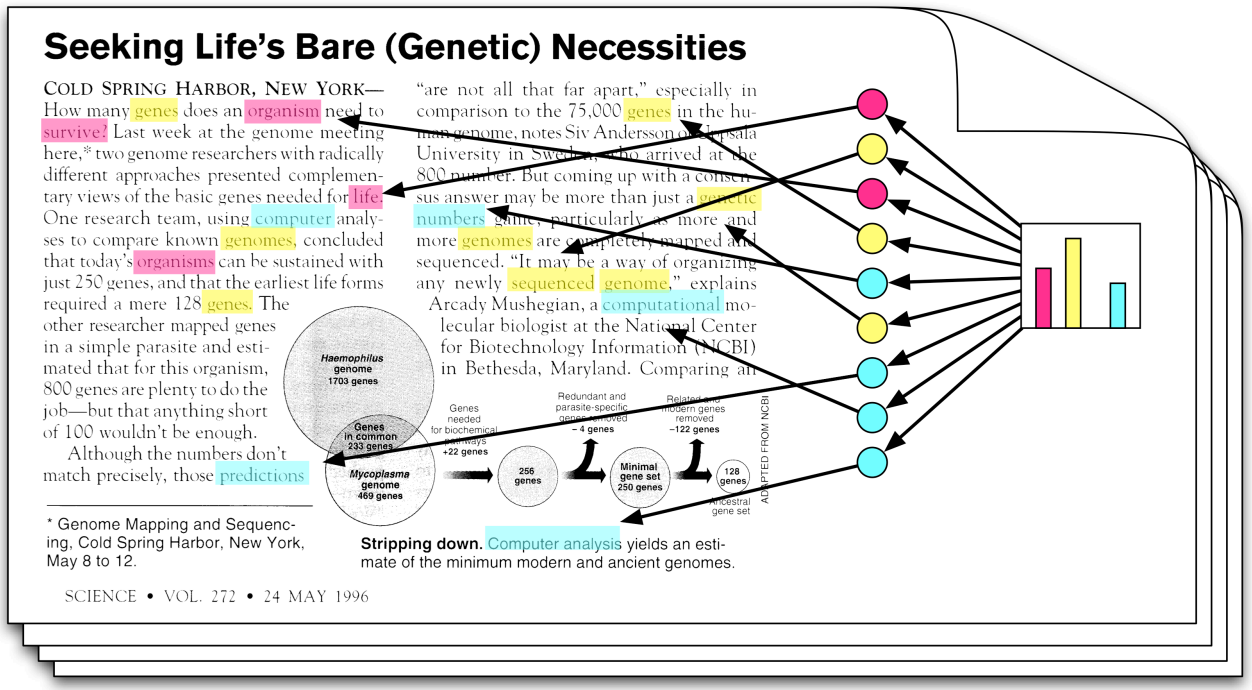
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

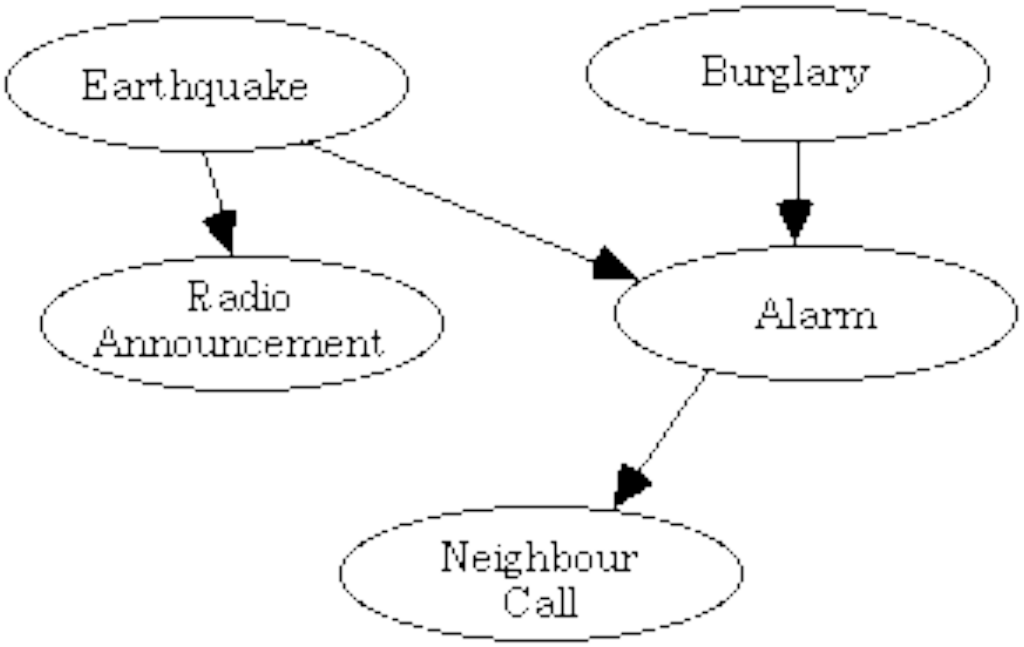
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

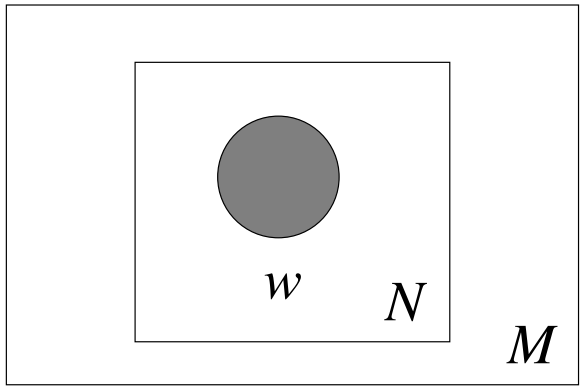
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

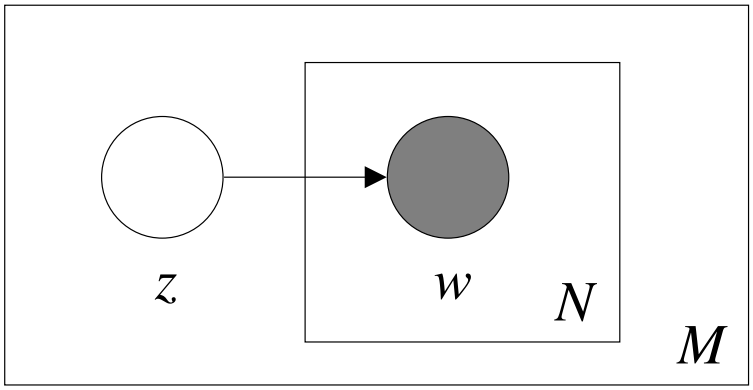
Modèles graphiques



Modèles graphiques

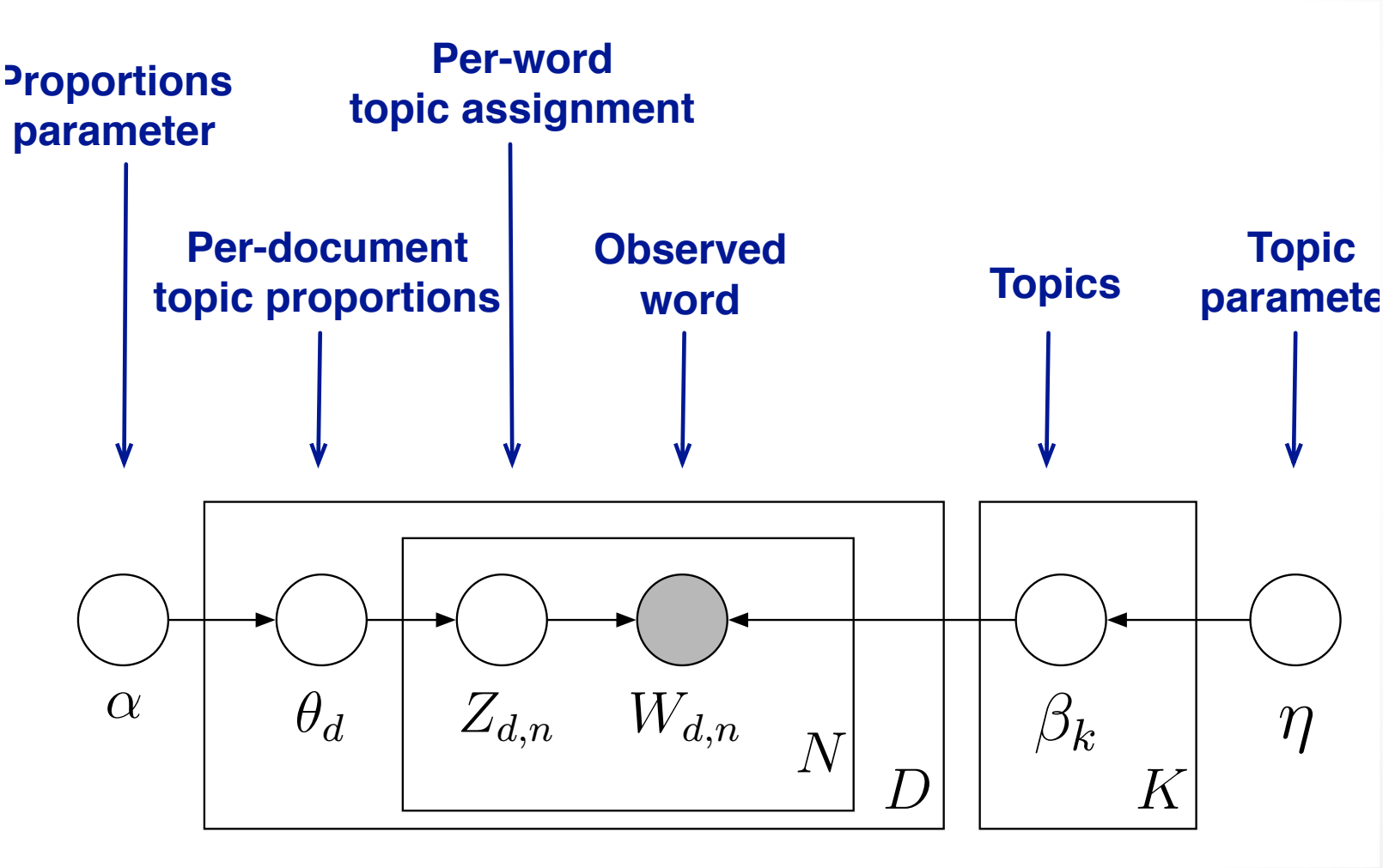


(a) unigram

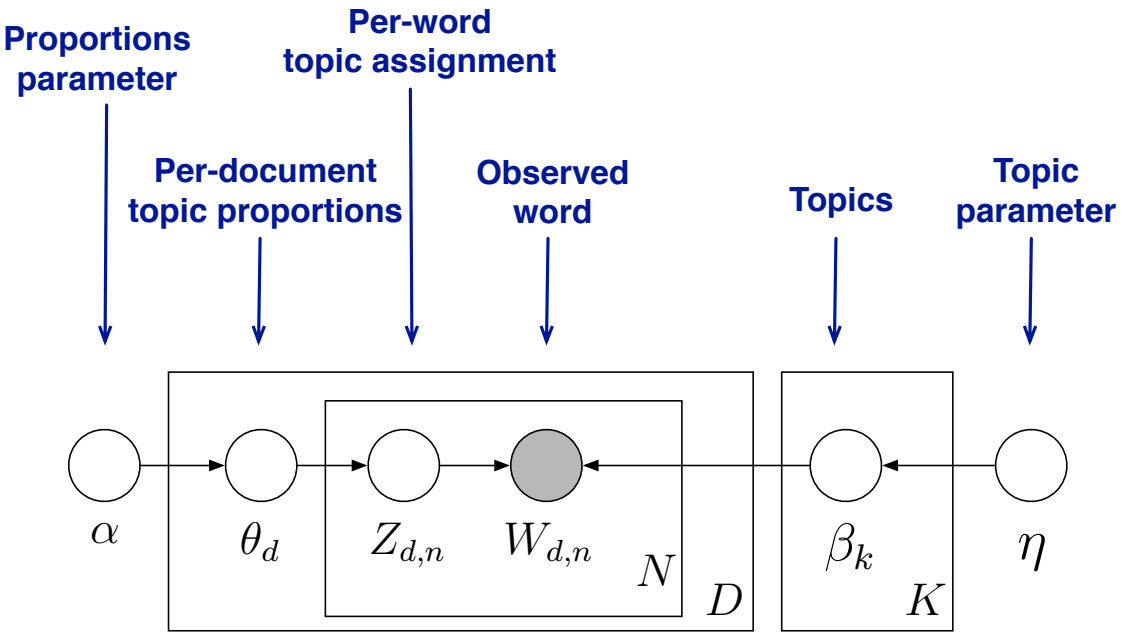


(b) mixture of unigrams

Modèle graphique LDA



Modèle graphique LDA



$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|\beta_{z_n})$$

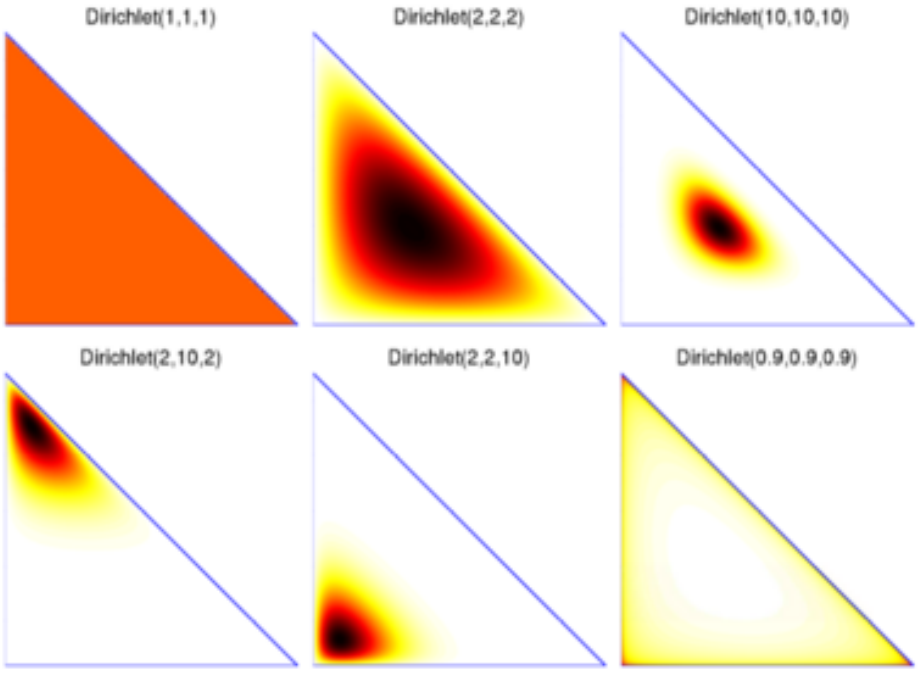
Processus de génération

1. Choisir $\theta \sim Dirichlet(\alpha)$.
2. Pour chaque mot w_n:
   - Choisir un topic $z_n \sim Multinomial(\theta)$
   - Choisir un mot $w_n \sim Multinomial(\beta_k)$, avec $k = z_n$.

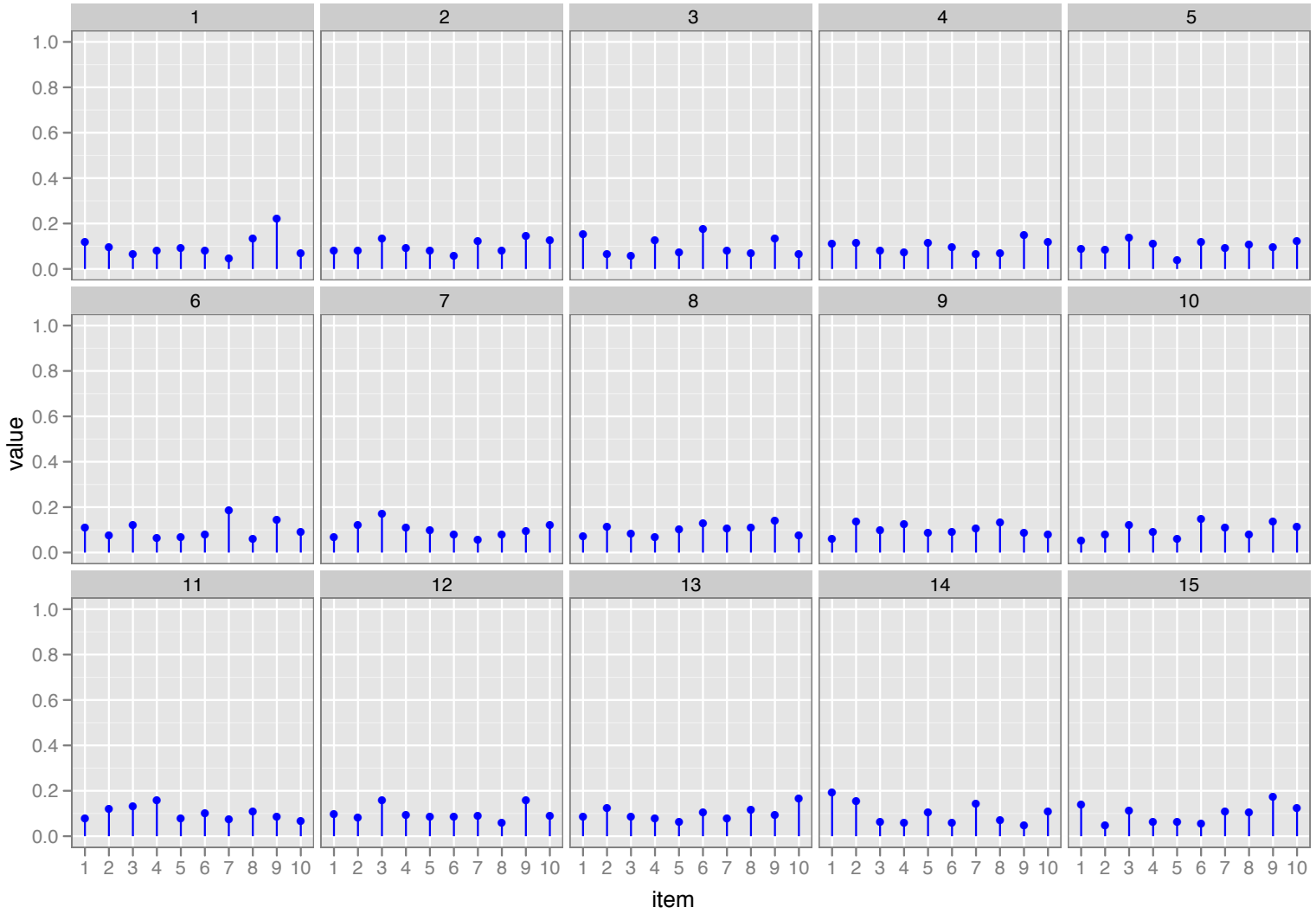
Dirichlet?

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

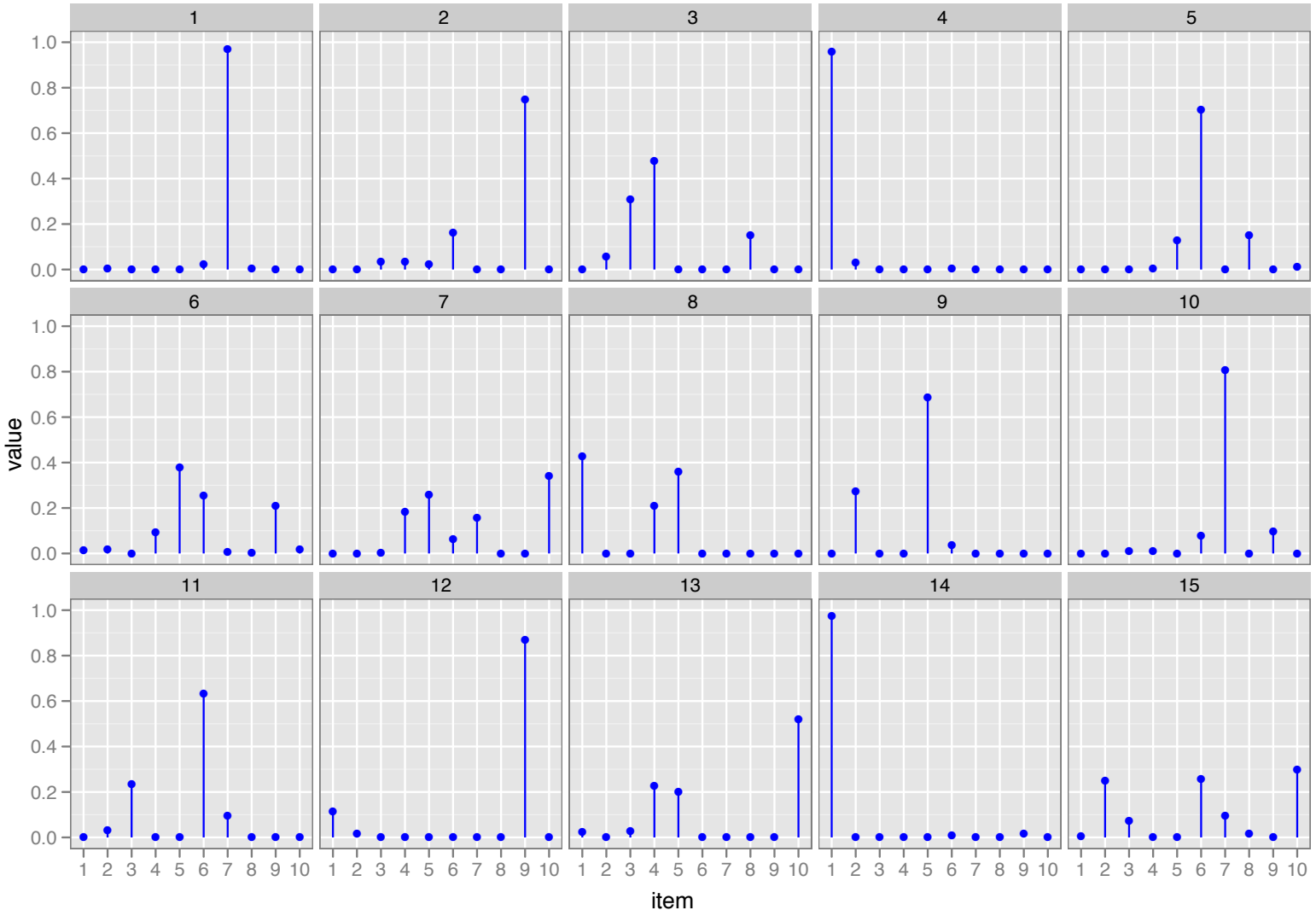
$$\forall i, \theta_i \geq 0 \text{ et } \sum_{i=1}^{k} \theta_i = 1$$



$\alpha = 10$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

$\alpha = 0.1$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

Inférence

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

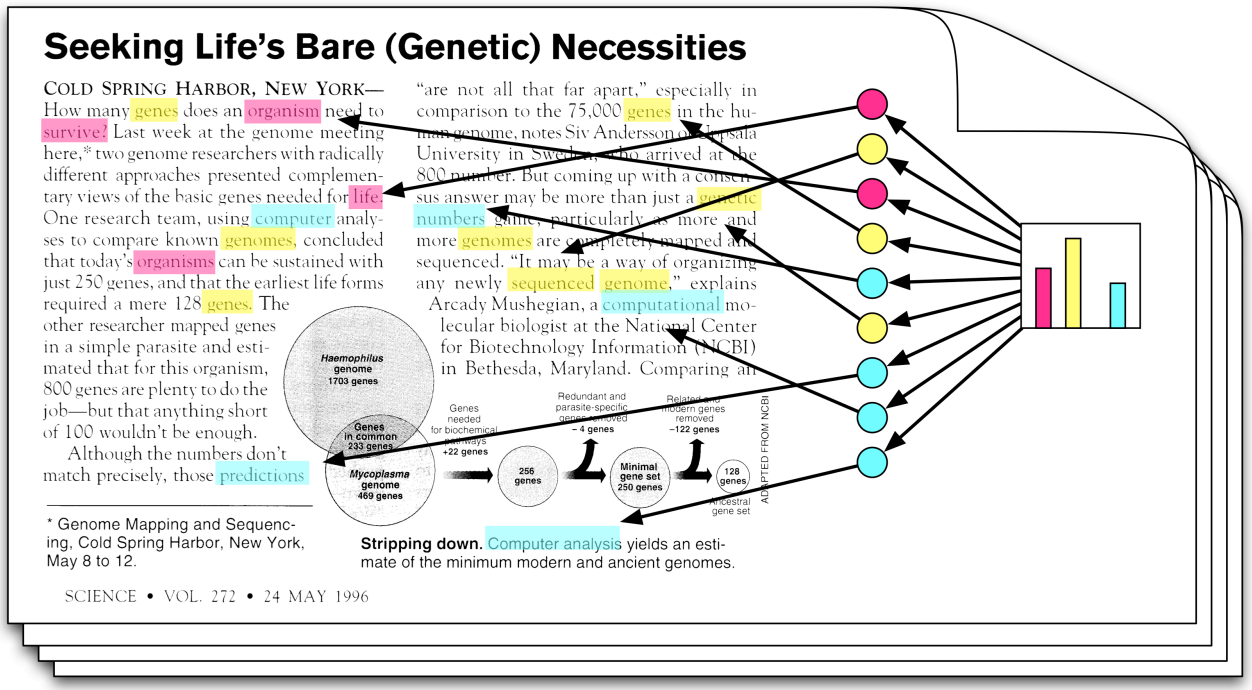
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

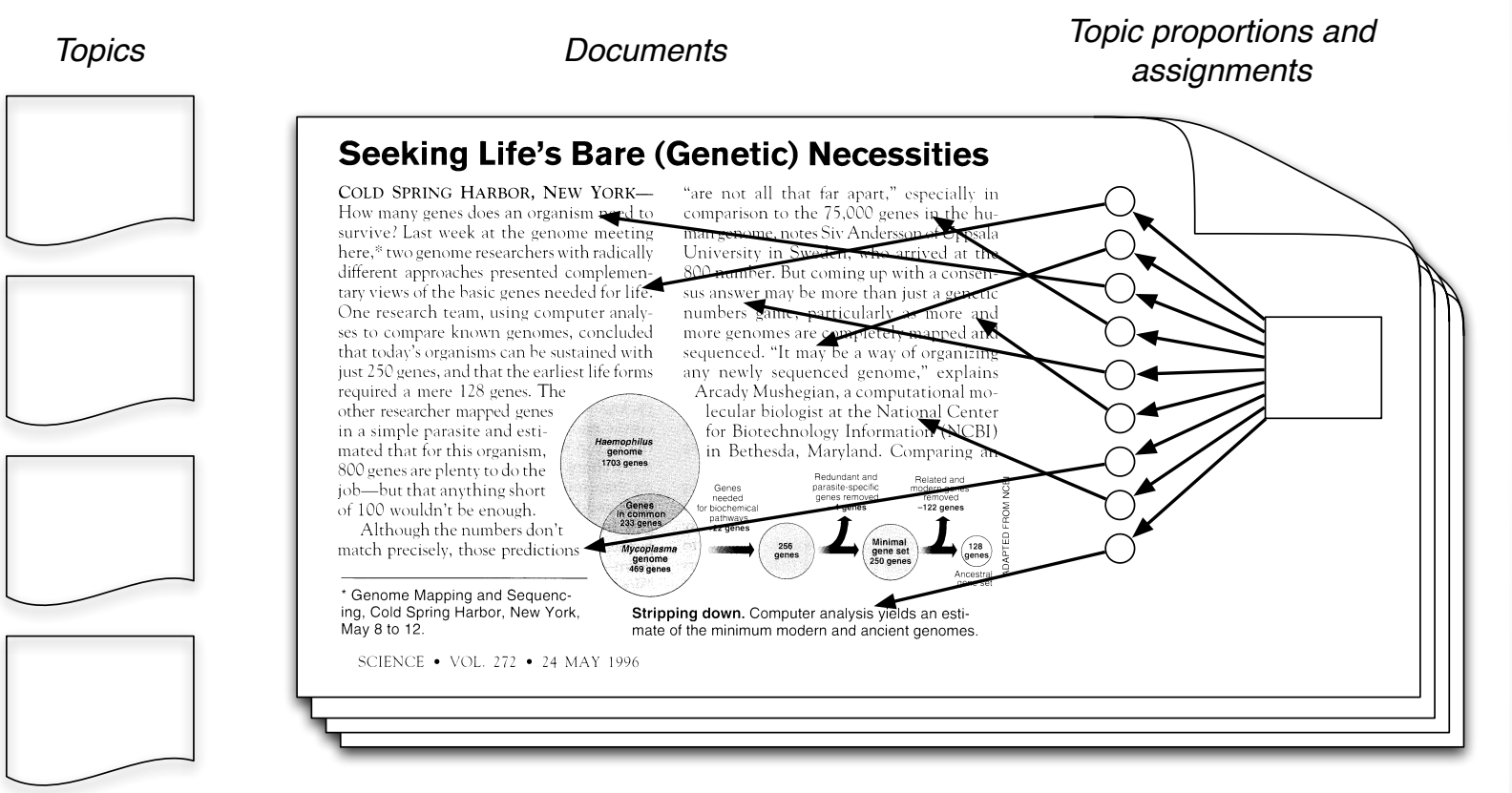
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

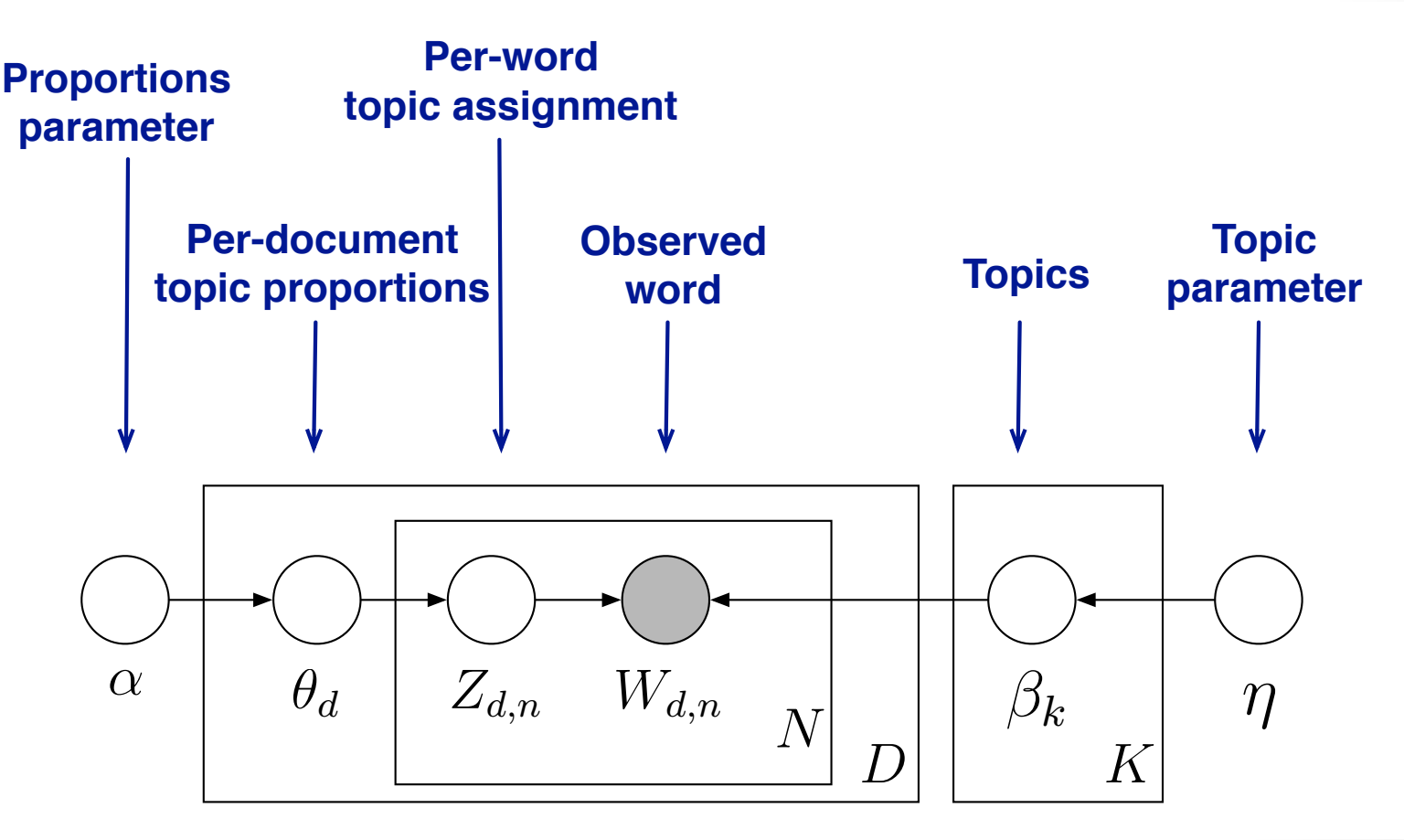
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Inférence


Topics
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways 22 genes
256 genes
Redundant and parasite-specific genes removed
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Inférence


Proportions parameter
Per-word topic assignment
Per-document topic proportions
Observed word
Topics
Topic parameter
α
θ_d
$Z_{d,n}$
$W_{d,n}$
N
D
β_k
K
η

Inférence

- Estimer la loi *à posteriori* des variables cachées connaissant les données

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

- Inenvisageable en pratique

Inférence approchée

- **Markov Chain Monte Carlo**
  - Estimer la distribution par échantillonnage
  - *Collapsed Gibbs sampling*
- **Variational inference**
  - Estimer la distribution par une optimisation

Exemple

- 17000 documents du magazine *Science* (1990-2000)
- 11 millions de mots
- 20000 termes uniques

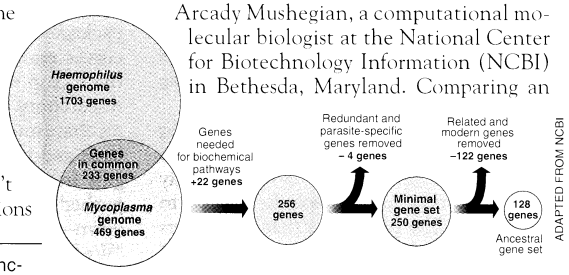
- LDA avec 100 topics (inférence variationnelle)

Exemple

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

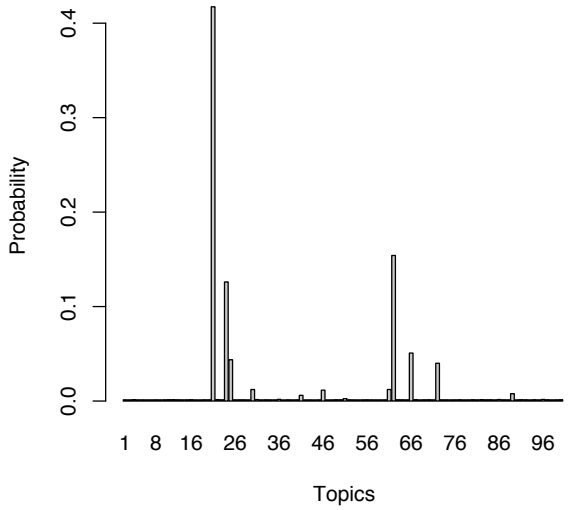
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996



Exemple

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Exemple

problem	model	selection	species
problems	rate	male	forest
mathematical	constant	males	ecology
number	distribution	females	fish
new	time	sex	ecological
mathematics	number	species	conservation
university	size	female	diversity
two	values	evolution	population
first	value	populations	natural
numbers	average	population	ecosystems
work	rates	sexual	populations
time	data	behavior	endangered
mathematicians	density	evolutionary	tropical
chaos	measured	genetic	forests
chaotic	models	reproductive	ecosystem

Exemple


Wikipedia Topics
Relative Presence of Topics in all Documents
{household, population, female}
{film, series, show}
{theory, work, human}
{son, year, death}
{war, force, army}
{system, computer, user}
{album, band, music}
{government, party, election}
{game, team, player}
{god, call, give}
{company, market, business}
{math, number, function}
{film, series, show}
words
related documents
related topics
film
series
show
character
play
make
episode
movie
good
release
feature
television
star
The X-Files
Orson Welles
Stanley Kubrick
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
Married... with Children
History of film
The A-Team
Pulp Fiction (film)
Mad (magazine)
{son, year, death}
{work, book, publish}
{album, band, music}
{woman, child, man}
{law, state, case}
{black, white, people}
{theory, work, human}
{@card@, make, design}
{war, force, army}
{god, call, give}
{game, team, player}
{day, year, event}
{company, market, business}
Stanley Kubrick
related topics
{film, series, show}
{theory, work, human}
{son, year, death}
{black, white, people}
{god, call, give}
{math, energy, light}
Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.
Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature.[1] A recurring theme in his films is man's inhumanity to man. While often viewed as
related documents
Orson Welles
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
The A-Team
Pulp Fiction (film)
Buffy the Vampire Slayer (TV series)
The X-Files
Sunset Boulevard (film)
Jack Benny
{theory, work, human}
words
related documents
related topics
theory
work
human
idea
term
study
view
science
concept
form
world
argue
social
Meme
Intelligent design
Immanuel Kant
Philosophy of mathematics
History of science
Free will
Truth
Psychoanalysis
Charles Peirce
Existentialism
Deconstruction
Social sciences
Idealism
{work, book, publish}
{law, state, case}
{son, year, death}
{woman, child, man}
{god, call, give}
{black, white, people}
{film, series, show}
{war, force, army}
{language, word, form}
{@card@, make, design}
{church, century, christian}
{rate, high, increase}
{company, market, business}

Bibliographie

- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research,* 3:993–1022, 2003.
- David M. Blei. Introduction to probabilistic topic models. In *Communications of the ACM,* à paraître.
- David M. Blei. Probabilistic Topic Models. KDD 2011 tutorial.