Modeling *Science*

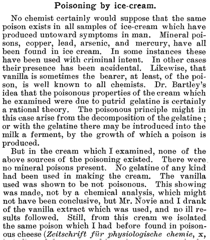
David M. Blei

Department of Computer Science
Princeton University

October 3, 2007

Joint work with John Lafferty (CMU)

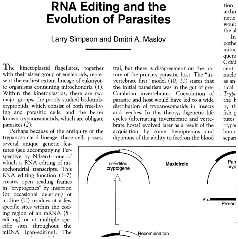
Modeling *Science*



Poisoning by ice-cream.

No chemist certainly would suppose that the same poison exists in all samples of ice-cream which have produced untoward symptoms in man. Mineral poisons, copper, lead, arsenic, and mercury, have all been found in ice cream. In some instances these have been used with criminal intent. In other cases their presence has been accidental. Likewise, that vanilla is sometimes the bearer, at least, of the poison, is well known to all chemists. Dr. Bartley's idea that the poisonous properties of the cream which he examined were due to putrid gelatine is certainly a rational theory. The poisonous principle might in this case arise from the decomposition of the gelatine ; or with the gelatine there may be introduced into the milk a ferment, by the growth of which a poison is produced.

But in the cream which I examined, none of the above sources of the poisoning existed. There were no mineral poisons present. No gelatine of any kind had been used in making the cream. The vanilla used was shown to be not poisonous. This showing was made, not by a chemical analysis, which might not have been conclusive, but Mr. Novie and I drank of the vanilla extract which was used, and no ill results followed. Still, from this cream we isolated the same poison which I had before found in poisonous cheese (*Zeitschrift für physiologische chemie*, x,



RNA Editing and the Evolution of Parasites

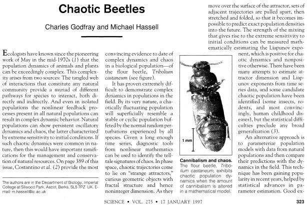
Larry Simpson and Dmitri A. Maslov

The kinetoplastid flagellates, together with their sister group of euglenoids, represent the earliest extant lineage of eukaryotic organisms containing mitochondria (1). Within the kinetoplastids, there are two major groups, the poorly studied bodonids-cryptobiids, which consist of both free-living and parasitic cells, and the better known trypanosomatids, which are obligate parasites (2).

Perhaps because of the antiquity of the trypanosomatid lineage, these cells possess several unique genetic features (see accompanying Perspective by Nilsen)—one of which is RNA editing of mitochondrial transcripts. This RNA editing function (3–7) creates open reading frames in "cryptogenes" by insertion (or occasional deletion) of uridine (U) residues at a few specific sites within the coding region of an mRNA (5'-editing) or at multiple specific sites throughout the mRNA (pan-editing). The

tral, but there is disagreement on the nature of the primary parasitic host. The "invertebrate first" model (10, 11) states that the initial parasitism was in the gut of pre-Cambrian invertebrates. Coevolution of parasite and host would have led to a wide distribution of trypanosomatids in insects and leeches. In this theory, digenetic life cycles (alternating invertebrate and vertebrate hosts) evolved later as a result of the acquisition by some hemipterans and dipterans of the ability to feed on the blood

tion
arthr
netic
woul
the al
In
pothe
mitoc
quen
Crithi
cent
nucle
as an
tical
Trypo
the b
by th
fish p
tutes
trypa
branc
separ



Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (1) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al*. (2) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (3).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

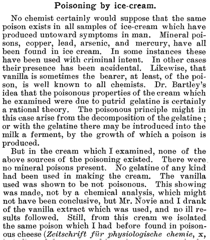
The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

SCIENCE • VOL. 275 • 17 JANUARY 1997

323

- Our data are *Science* from 1880-2002, courtesy of JSTOR.
- JSTOR is an on-line archive that scans the original volumes and performs optical character recognition on the scans.
- This process results in 130K documents, 76M words.

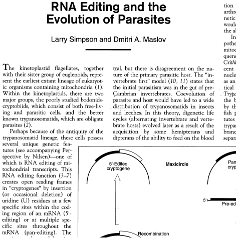
Modeling *Science*



Poisoning by ice-cream.

No chemist certainly would suppose that the same poison exists in all samples of ice-cream which have produced untoward symptoms in man. Mineral poisons, copper, lead, arsenic, and mercury, have all been found in ice cream. In some instances these have been used with criminal intent. In other cases their presence has been accidental. Likewise, that vanilla is sometimes the bearer, at least, of the poison, is well known to all chemists. Dr. Bartley's idea that the poisonous properties of the cream which he examined were due to putrid gelatine is certainly a rational theory. The poisonous principle might in this case arise from the decomposition of the gelatine ; or with the gelatine there may be introduced into the milk a ferment, by the growth of which a poison is produced.

But in the cream which I examined, none of the above sources of the poisoning existed. There were no mineral poisons present. No gelatine of any kind had been used in making the cream. The vanilla used was shown to be not poisonous. This showing was made, not by a chemical analysis, which might not have been conclusive, but Mr. Novie and I drank of the vanilla extract which was used, and no ill results followed. Still, from this cream we isolated the same poison which I had before found in poisonous cheese (*Zeitschrift für physiologische chemie*, x,



RNA Editing and the Evolution of Parasites

Larry Simpson and Dmitri A. Maslov

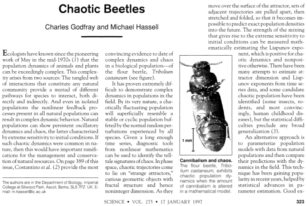
The kinetoplastid flagellates, together with their sister group of euglenoids, represent the earliest extant lineage of eukaryotic organisms containing mitochondria (*1*). Within the kinetoplastids, there are two major groups, the poorly studied bodonids-cryptobiids, which consist of both free-living and parasitic cells, and the better known trypanosomatids, which are obligate parasites (*2*).

Perhaps because of the antiquity of the trypanosomatid lineage, these cells possess several unique genetic features (see accompanying Perspective by Nilsen)—one of which is RNA editing of mitochondrial transcripts. This RNA editing function (*3–7*) creates open reading frames in "cryptogenes" by insertion (or occasional deletion) of uridine (U) residues at a few specific sites within the coding region of an mRNA (5'-editing) or at multiple specific sites throughout the mRNA (pan-editing). The

tral, but there is disagreement on the nature of the primary parasitic host. The "invertebrate first" model (*10, 11*) states that the initial parasitism was in the gut of pre-Cambrian invertebrates. Coevolution of parasite and host would have led to a wide distribution of trypanosomatids in insects and leeches. In this theory, digenetic life cycles (alternating invertebrate and vertebrate hosts) evolved later as a result of the acquisition by some hemipterans and dipterans of the ability to feed on the blood

tion
arthr
netic
woul
the al
In
pothe
mitoc
quen
Crithi
cent
nucle
as an
tical
Trypa
the b
by th
fish p
tutes
trypa
branc
separ

5'-Edited cryptogene
Maxicircle
Pan cryp
5'
Pre-ed
Recombination



Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al*. (*2*) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (*3*).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

SCIENCE • VOL. 275 • 17 JANUARY 1997 323

- Discover the hidden thematic structure with hierarchical probabilistic models called *topic models*.
- Use this structure for browsing, search, and similarity assessment.

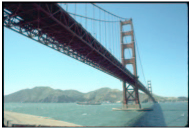
Discover topics from a corpus

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

Annotate unlabeled images

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE



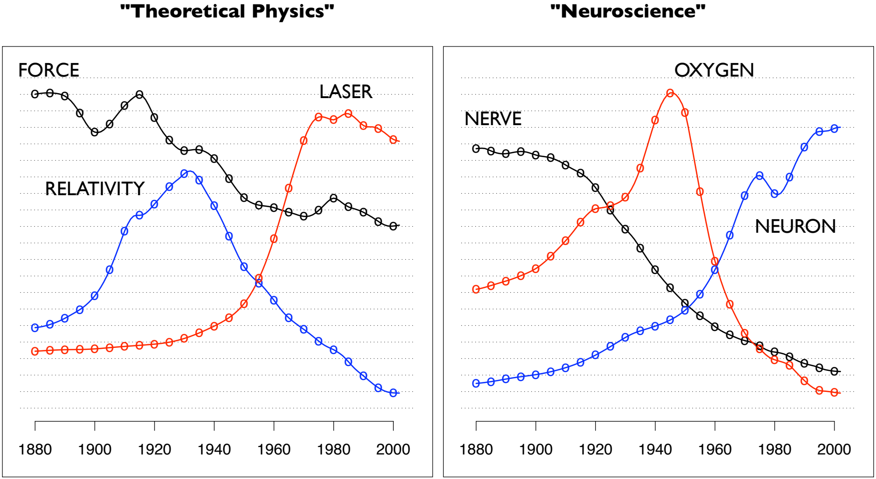
SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

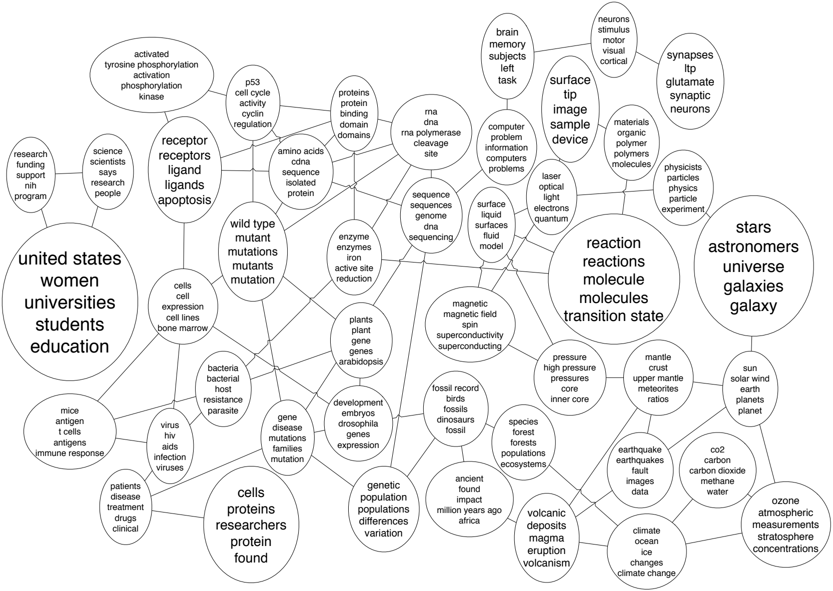
Model the evolution of topics over time



"Theoretical Physics"

"Neuroscience"

Model connections between topics



activated
tyrosine phosphorylation
activation
phosphorylation
kinase

p53
cell cycle
activity
cyclin
regulation

proteins
protein
binding
domain
domains

rna
dna
rna polymerase
cleavage
site

brain
memory
subjects
left
task

neurons
stimulus
motor
visual
cortical

synapses
ltp
glutamate
synaptic
neurons

surface
tip
image
sample
device

materials
organic
polymer
polymers
molecules

computer
problem
information
computers
problems

research
funding
support
nih
program

science
scientists
says
research
people

receptor
receptors
ligand
ligands
apoptosis

amino acids
cdna
sequence
isolated
protein

sequence
sequences
genome
dna
sequencing

surface
liquid
surfaces
fluid
model

laser
optical
light
electrons
quantum

physicists
particles
physics
particle
experiment

united states
women
universities
students
education

wild type
mutant
mutations
mutants
mutation

enzyme
enzymes
iron
active site
reduction

reaction
reactions
molecule
molecules
transition state

stars
astronomers
universe
galaxies
galaxy

cells
cell
expression
cell lines
bone marrow

magnetic
magnetic field
spin
superconductivity
superconducting

plants
plant
gene
genes
arabidopsis

bacteria
bacterial
host
resistance
parasite

pressure
high pressure
pressures
core
inner core

mantle
crust
upper mantle
meteorites
ratios

sun
solar wind
earth
planets
planet

mice
antigen
t cells
antigens
immune response

virus
hiv
aids
infection
viruses

gene
disease
mutations
families
mutation

development
embryos
drosophila
genes
expression

fossil record
birds
fossils
dinosaurs
fossil

species
forest
forests
populations
ecosystems

earthquake
earthquakes
fault
images
data

co2
carbon
carbon dioxide
methane
water

patients
disease
treatment
drugs
clinical

cells
proteins
researchers
protein
found

genetic
population
populations
differences
variation

ancient
found
impact
million years ago
africa

volcanic
deposits
magma
eruption
volcanism

climate
ocean
ice
changes
climate change

ozone
atmospheric
measurements
stratosphere
concentrations

Outline

1. Introduction

2. Latent Dirichlet allocation

3. Dynamic topic models

4. Correlated topic models

Outline

1. Introduction

2. Latent Dirichlet allocation

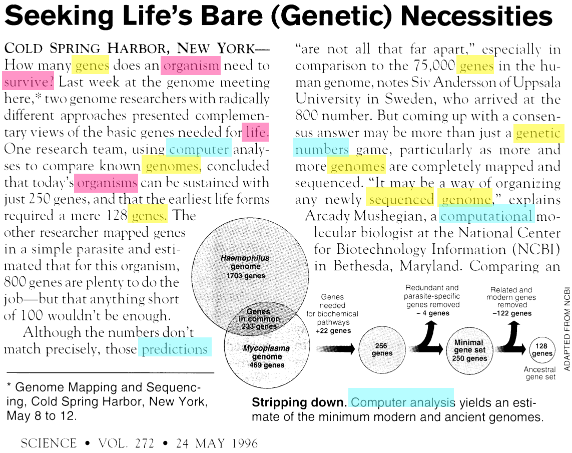
3. Dynamic topic models

4. Correlated topic models

Probabilistic modeling

- Treat data as observations that arise from a generative probabilistic process that includes hidden variables
  - For documents, the hidden variables reflect the thematic structure of the collection.
- Infer the hidden structure using *posterior inference*
  - What are the topics that describe this collection?
- Situate new data into the estimated model.
  - How does this query or new document fit into the estimated topic structure?

Intuition behind LDA



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 – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set

ADAPTED FROM NCBI

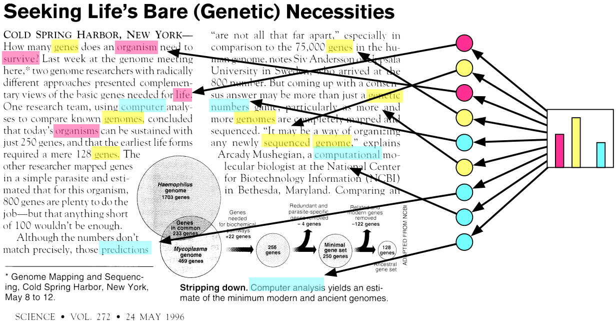
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Simple intuition: Documents exhibit multiple topics.

Generative process



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 – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

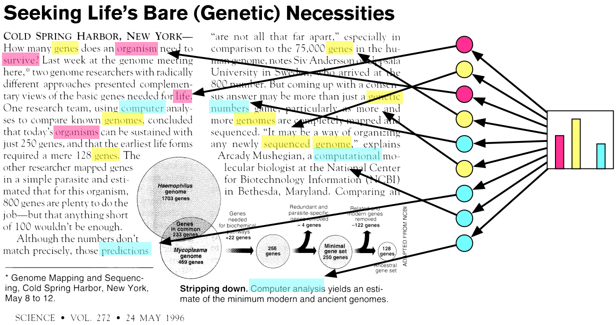
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- Cast these intuitions into a generative probabilistic process
- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

Generative process



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 – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

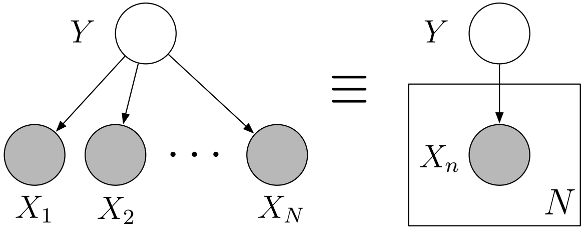
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

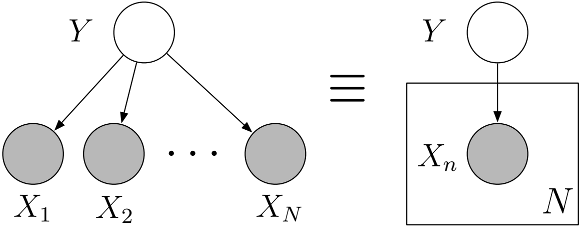
- In reality, we only observe the documents
- Our goal is to infer the underlying topic structure
  - What are the topics?
  - How are the documents divided according to those topics?

Graphical models (Aside)



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

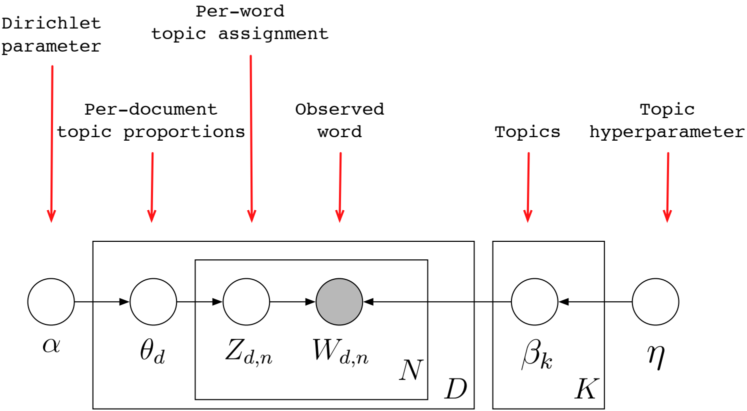
Graphical models (Aside)



- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n \mid y)$$

Latent Dirichlet allocation



Dirichlet parameter

Per-word topic assignment

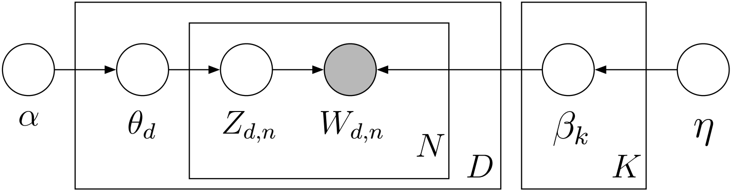
Per-document topic proportions

Observed word

Topics

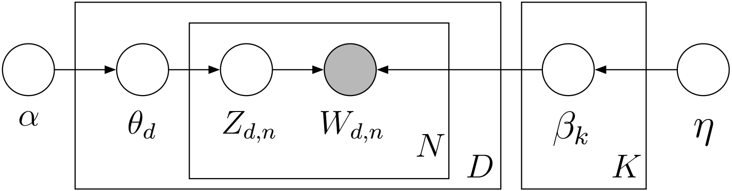
Topic hyperparameter

Latent Dirichlet allocation



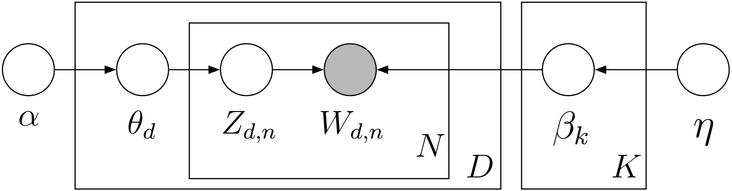
1. Draw each topic $\beta_i \sim \mathrm{Dir}(\eta)$, for $i \in \{1, \ldots, K\}$.
2. For each document:
    1. Draw topic proportions $\theta_d \sim \mathrm{Dir}(\alpha)$.
    2. For each word:
        1. Draw $Z_{d,n} \sim \mathrm{Mult}(\theta_d)$.
        2. Draw $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$.

Latent Dirichlet allocation



- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, etc.

Latent Dirichlet allocation

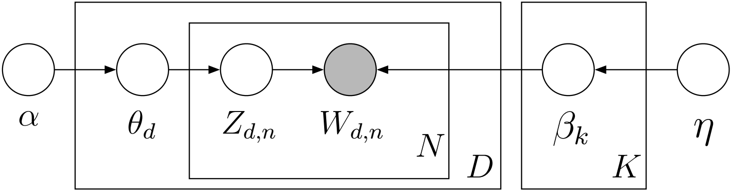


- Computing the posterior is intractable:

$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_{\theta} p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

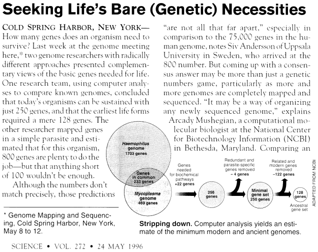
- Several approximation techniques have been developed.

Latent Dirichlet allocation



- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)

Example inference



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 – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set

ADAPTED FROM NCBI

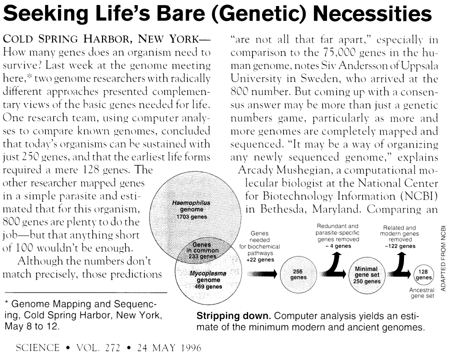
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- **Data**: The OCR'ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.

Example inference



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

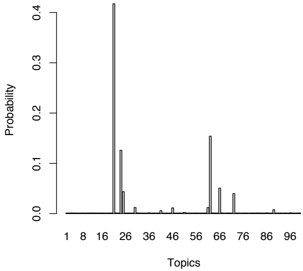
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

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

SCIENCE • VOL. 272 • 24 MAY 1996

→



Example topics

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

LDA discussion

- LDA is a powerful model for
  - Visualizing the hidden thematic structure in large corpora
  - Generalizing new data to fit into that structure
- LDA is a mixed membership model (Erosheva, 2004) that builds on the work of Deerwester et al. (1990) and Hofmann (1999).
  - For document collections and other grouped data, this might be more appropriate than a simple finite mixture
  - See Blei et al., 2003 for a quantitative comparison.
- *Modular*: It can be embedded in more complicated models.
- *General*: The data generating distribution can be changed.
- Variational inference is fast; allows us to analyze large data sets.
- Code to play with LDA is freely available on my web-site, http://www.cs.princeton.edu/∼blei.

Outline

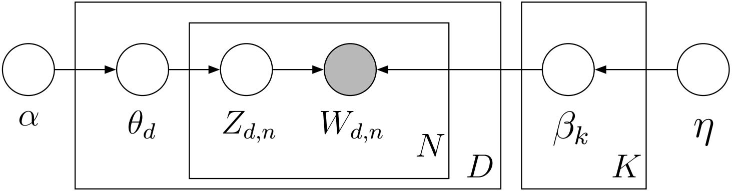
1 Introduction

2 Latent Dirichlet allocation

3 Dynamic topic models

4 Correlated topic models

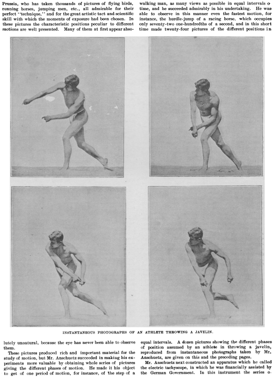
LDA and exchangeability



- LDA assumes that documents are exchangeable.
- I.e., their joint probability is invariant to permutation.
- This is too restrictive.

Documents are not exchangeable

"Instantaneous Photography" (1890)



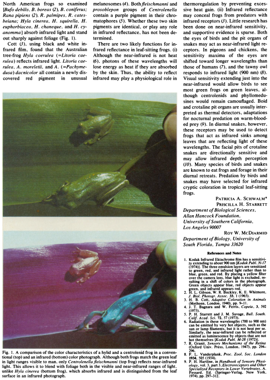
Prussia, who has taken thousands of pictures of flying birds, running horses, jumping men, etc., all admirable for their perfect ''technique,'' and for the great artistic tact and scientific skill with which the moments of exposure had been chosen. In these pictures the characteristic positions peculiar to different motions are well presented. Many of them at first appear absolutely unnatural, because the eye has never been able to observe them.

These pictures produced rich and important material for the study of motion, but Mr. Anschuetz succeeded in making his experiments more valuable by obtaining whole series of pictures giving the different phases of motion. He made it his object to get of one period of motion, for instance, of the step of a walking man, as many views as possible in equal intervals o time, and he succeeded admirably in his undertaking. He was able to observe in this manner even the fastest motion, for instance, the hurdle-jump of a racing horse, which occupies only seventy-two one-hundredths of a second, and in this short time made twenty-four pictures of the different positions in equal intervals. A dozen pictures showing the different phases of position assumed by an athlete in throwing a javelin, reproduced from instantaneous photographs taken by Mr. Anschuetz, are given on this and the preceding pages.

Mr. Anschuetz next constructed an apparatus which he called the electric tachyscope, in which he was financially assisted by the German Government. In this instrument the series o-

INSTANTANEOUS PHOTOGRAPHS OF AN ATHLETE THROWING A JAVELIN.

"Infrared Reflectance in Leaf-Sitting Neotropical Frogs" (1977)



North American frogs so examined [*Bufo debilis*, *B. boreas* (2), *B. coniferus; Rana pipiens* (2), *R. palmipes*, *R. catesbeiana; Hyla cinerea*, *H. squirella*, *H. euphorbiacea*, *H. chaneque*, and *H. cyanomma*] absorb infrared light and stand out sharply against foliage (Fig. 1).

Cott (*3*), using black and white infrared film, found that the Australian tree-frog *Hyla coerulea* (=*Litoria caerulea*) reflects infrared light. *Litoria caerulea*, *A. moreletii*, and *A.* (=*Pachymedusa*) *dacnicolor* all contain a newly discovered red pigment in unusual melanosomes (*4*). Both *fleischmanni* and *prosoblepon* groups of *Centrolenella* contain a purple pigment in their chromatophores (*5*). Whether these two skin pigments are identical, or play any role in infrared reflectance, has not been determined.

There are two likely functions for infrared reflectance in leaf-sitting frogs. (i) Although the near-infrared is not heat (*6*), photons of these wavelengths will lose energy as heat if they are absorbed by the skin. Thus, the ability to reflect infrared may play a physiological role in thermoregulation by preventing excessive heat gain. (ii) Infrared reflectance may conceal frogs from predators with infrared receptors (*3*). Little research has been done on near-infrared sensitivity, and supportive evidence is sparse. Both the eyes of birds and the pit organs of snakes may act as near-infrared light receptors. In pigeons and chickens, the sensitivity maxima of the eyes are shifted toward longer wavelengths than those of humans (*7*), and the tawny owl responds to infrared light (900 nm) (*8*). Visual sensitivity extending just into the near-infrared would allow birds to see most green frogs on green leaves, although centrolenids and phyllomedusines would remain camouflaged. Boid and crotaline pit organs are usually interpreted as thermal detectors, adaptations for nocturnal predation on warm-blooded prey (*9*). In diurnal snakes, however, these receptors may be used to detect frogs that act as infrared sinks among leaves that are reflecting light of these wavelengths. The facial pits of crotaline snakes are directionally sensitive and may allow infrared depth perception (*10*). Many species of birds and snakes are known to eat frogs and forage in their diurnal retreats. Predation by birds and snakes may have selected for infrared cryptic coloration in tropical leaf-sitting frogs.

PATRICIA A. SCHWALM*
PRISCILLA H. STARRETT
Department of Biological Sciences,
Allan Hancock Foundation,
University of Southern California,
Los Angeles 90007
ROY W. MCDIARMID
Department of Biology, University of
South Florida, Tampa 33620

Fig. 1. A comparison of the color characteristics of a hylid and a centrolenid frog in a conventional (top) and an infrared (bottom) color photograph. Although both frogs match the green leaf in light ranges visible to man, only *Centrolenella fleischmanni* (top frog) reflects near-infrared light. This allows it to blend with foliage both in the visible and near-infrared ranges of light, unlike *Hyla cinerea* (bottom frog), which absorbs infrared and is distinguished from the leaf surface in an infrared photograph.

References and Notes

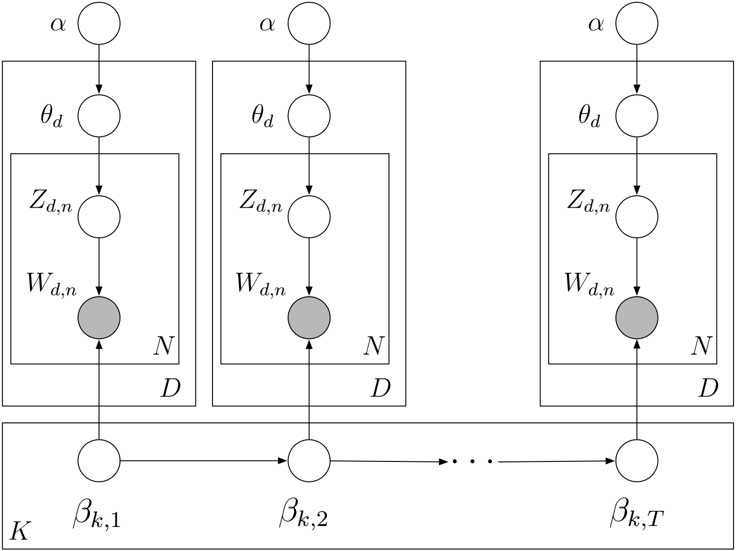
1. Kodak Infrared Ektachrome film has a sensitivity extending to about 900 nm [*Kodak Publ. N-17* (1974)]. The three emulsion layers are sensitized to green, red, and infrared light rather than to blue, green, and red. By placing a yellow filter over the camera lens, blue light is excluded, resulting in a shift of colors in the photograph. Green objects appear blue, red objects appear green, and infrared appears red.
2. H. L. Gibson, W. R. Buckley, K. E. Whitmore, *J. Biol. Photogr. Assoc.* **33**, 1 (1965).
3. H. B. Cott, *Adaptive Coloration in Animals* (Methuen, London, 1940), pp. 9–11.
4. J. T. Bagnara and W. Ferris, *Copeia*, **3**, 592 (1975).
5. P. H. Starrett and J. M. Savage, *Bull. South. Calif. Acad. Sci.* **72**, 57 (1973).
6. Radiation in these wavelengths (700 to 900 nm) can be emitted by very hot objects, such as the sun or lamp filaments, but it is not heat per se. Similarly, the near-infrared can be reflected or emitted as luminescence by objects that are not hot themselves [*Kodak Publ. M-28* (1972)].
7. R. Granit, *Sensory Mechanisms of the Retina* (Oxford Univ. Press, London, 1947), pp. 294–296.
8. F. L. Vanderplank, *Proc. Zool. Soc. London* **1934**, 505 (1934).
9. P. H. Hartline, in *Handbook of Sensory Physiology*, vol. 3, part 3, *Electroreceptors and Other Specialized Receptors in Lower Vertebrates*, A. Fessard, Ed. (Springer-Verlag, New York, 1974), pp. 297–312.

- Documents about the same topic are not exchangeable.
- Topics evolve over time.

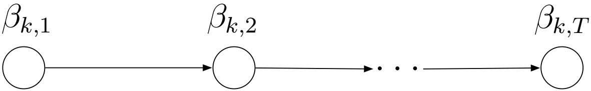
Dynamic topic model

- Divide corpus into sequential slices (e.g., by year).
- Assume each slice's documents exchangeable.
  - Drawn from an LDA model.
- Allow topic distributions evolve from slice to slice.

Dynamic topic models



Modeling evolving topics



- Use a logistic normal distribution to model topics evolving over time (Aitchison, 1980)
- A state-space model on the natural parameter of the topic multinomial (West and Harrison, 1997)

$$\begin{aligned} \beta_{t,k} \,|\, \beta_{t-1,k} &\sim \mathcal{N}(\beta_{t-1,k}, I\sigma^2) \\ p(w \,|\, \beta_{t,k}) &= \exp\left\{\beta_{t,k} - (1 + \textstyle\sum_{v=1}^{V-1} \exp\{\beta_{t,k,v}\})\right\} \end{aligned}$$

Posterior inference

- Our goal is to compute the posterior distribution,

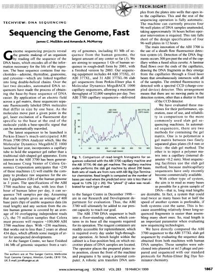
$$p(\beta_{1:T,1:K}, \theta_{1:T,1:D}, \mathbf{z}_{1:T,1:D} \mid \mathbf{w}_{1:T,1:D}).$$

- Exact inference is impossible
  - Per-document mixed-membership model
  - Non-conjugacy between $p(w \mid \beta_{t,k})$ and $p(\beta_{t,k})$
- MCMC is not practical for the amount of data.
- Solution: Variational inference

Variational inference

- Define a family of distributions q on the latent variables indexed by free *variational parameters*.
- Find the member closest in $\mathrm{KL}(q||p)$ to the true posterior.
- Equivalently, maximize the Jensen's bound on the marginal likelihood of the data, within the variational family.
- See Jordan et al. (1999) and Wainwright and Jordan (2003).
- (More details at the end of the talk, if you are interested.)

Science data



SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 μm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

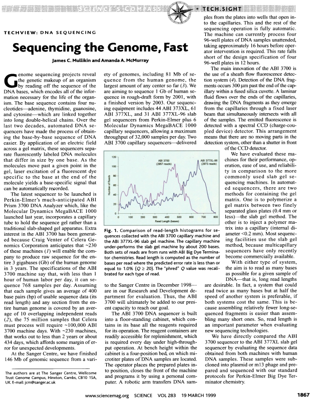
www.sciencemag.org SCIENCE VOL 283 19 MARCH 1999 1867

```
TECHVIEW: DNA S E Q U E  N C I NG


   Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray



Genome sequencing projects reveal
      the genetic makeup of an organism
      by reading off the sequence of the
DNA bases, which encodes all of the infor-
mation necessary for the life of the organ-
ism. The base sequence contains four nu-
cleotides-adenine, thymidine, guanosine,
and cytosine-which are linked together
into long double-helical chains. Over the
last two decades, automated DNA se-
quencers have made the process of obtain-
ing the base-by-base sequence of DNA...
```

- Analyze JSTOR's entire collection from *Science* (1880-2002)
- No reliable punctuation, meta-data, or references
- Restrict to 30K terms that occur more than ten times
- The data are 76M words in 130K documents

Analyzing a document

Original article



SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (*1*) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (*2*), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (*3*). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (*4*). Detection of the DNA fragments occurs 300 µm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

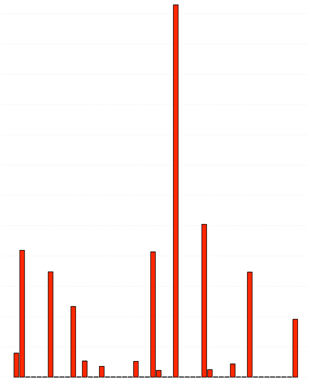
With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

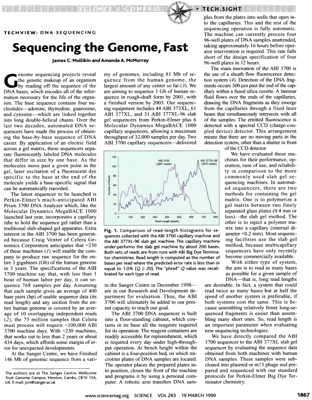
www.sciencemag.org SCIENCE VOL 283 19 MARCH 1999 1867

Topic proportions



Analyzing a document

Original article



SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymidine, guanosine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegaBACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (1) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (2), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a variety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (3). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL, 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegaBACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA samples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (4). Detection of the DNA fragments occurs 300 µm past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fixed laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB10 1SA, UK. E-mail: jcm@sanger.ac.uk

www.sciencemag.org SCIENCE VOL 283 19 MARCH 1999 1867

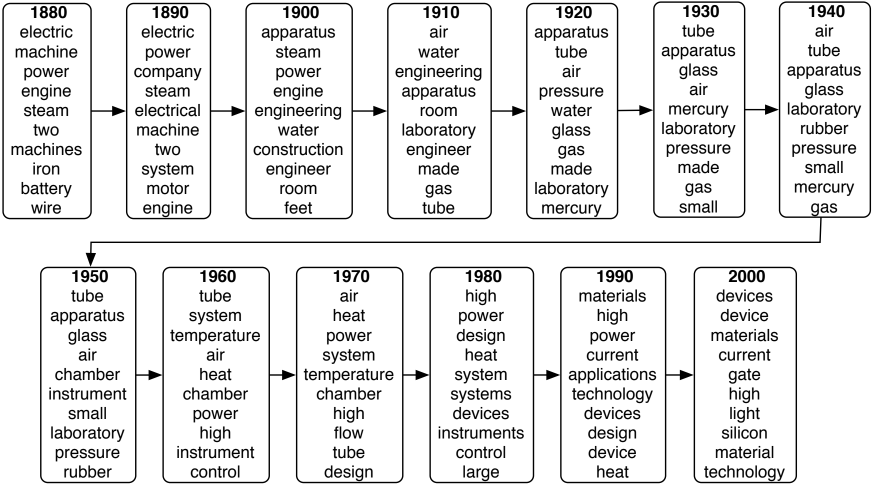
Most likely words from top topics

sequence
genome
genes
sequences
human
gene
dna
sequencing
chromosome
regions
analysis
data
genomic
number

devices
device
materials
current
high
gate
light
silicon
material
technology
electrical
fiber
power
based

data
information
network
web
computer
language
networks
time
software
system
words
algorithm
number
internet

Analyzing a topic



1880
electric
machine
power
engine
steam
two
machines
iron
battery
wire

1890
electric
power
company
steam
electrical
machine
two
system
motor
engine

1900
apparatus
steam
power
engine
engineering
water
construction
engineer
room
feet

1910
air
water
engineering
apparatus
room
laboratory
engineer
made
gas
tube

1920
apparatus
tube
air
pressure
water
glass
gas
made
laboratory
mercury

1930
tube
apparatus
glass
air
mercury
laboratory
pressure
made
gas
small

1940
air
tube
apparatus
glass
laboratory
rubber
pressure
small
mercury
gas

1950
tube
apparatus
glass
air
chamber
instrument
small
laboratory
pressure
rubber

1960
tube
system
temperature
air
heat
chamber
power
high
instrument
control

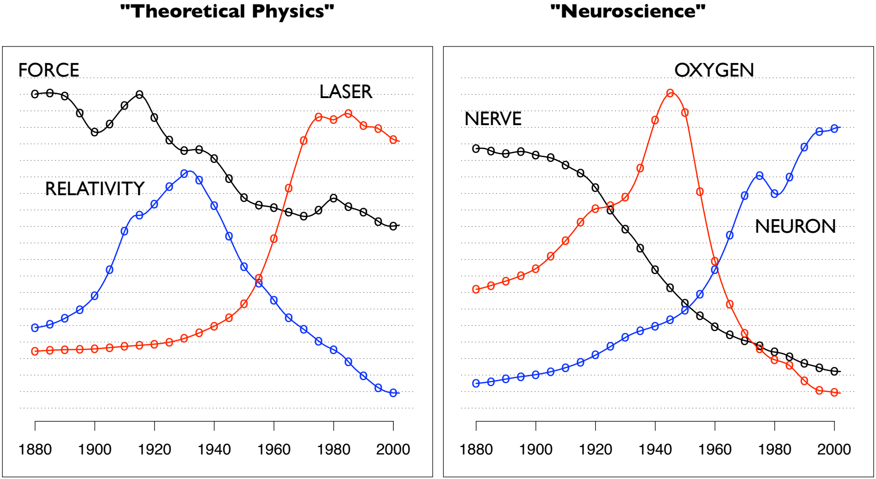
1970
air
heat
power
system
temperature
chamber
high
flow
tube
design

1980
high
power
design
heat
system
systems
devices
instruments
control
large

1990
materials
high
power
current
applications
technology
devices
design
device
heat

2000
devices
device
materials
current
gate
high
light
silicon
material
technology

Visualizing trends within a topic



Time-corrected document similarity

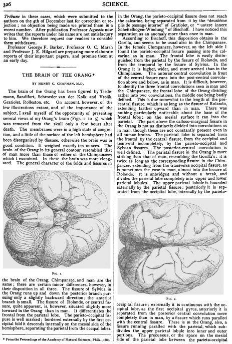
- Consider the expected Hellinger distance between the topic proportions of two documents,

$$d_{ij} = \mathrm{E}\left[\sum_{k=1}^{K}(\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}})^2 \,|\, \mathbf{w}_i, \mathbf{w}_j\right]$$

- Uses the latent structure to define similarity
- Time has been factored out because the topics associated to the components are different from year to year.
- Similarity based only on topic proportions

Time-corrected document similarity

The Brain of the Orang (1880)



326 SCIENCE.

Tribune in these cases, which were submitted to the authors on the 4th of December last for correction or rejection; no objection being made we printed them in a recent number. After publication Professor Agassiz now writes that the reports under his name are not satisfactory to him. We therefore request our readers to consider them withdrawn.

Professor George F. Barker, Professor O. C. Marsh and Professor J. E. Hilgard are preparing more elaborate reports of their important papers, and promise them at an early day.

THE BRAIN OF THE ORANG.*

BY HENRY C. CHAPMAN, M.D.

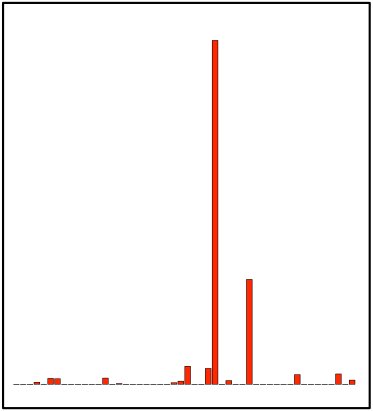
The brain of the Orang has been figured by Tiedemann, Sandifort, Schroeder van der Kolk and Vrolik, Gratiolet, Rolleston, etc. On account, however, of the few illustrations extant, and of the importance of the subject, I avail myself of the opportunity of presenting several views of my Orang's brain (Figs. 1 to 5), which was removed from the skull only a few hours after death. The membranes were in a high state of congestion, and a little of the surface of the left hemisphere had been disorganized by disease, otherwise the brain was in good condition. It weighed exactly ten ounces. The brain of the Orang in its general contour resembled that of man more than those of either of the Chimpanzees which I examined. In these the brain was more elongated. The general character of the folds and fissures in the brain of the Orang, Chimpanzee, and man are the same; there are certain minor differences, however, in their disposition in all three. The fissure of Sylvius in the Orang runs up and down the posterior branch pursuing only a slightly backward direction; the anterior branch is small. The fissure of Rolando, or central fissure, quite apparent, is, however, situated slightly more forward in the Orang than in man. It differentiates the frontal from the parietal lobe. The parieto-occipital fispure is well marked; bordered externally by the first occipital fold it descends internally on the mesial side of the hemisphere, separating the parietal from the occipal lobes. in the Orang, the parieto-occipital fissure does not reach the calcarine, being separated from it by the "deuxième plis de passage interne" of Gratiolet, or "untere innere Scheitelbogen-Windung" of Bischoff. I have noticed this separation as an anomaly more than once in man.

FIG. 1.

According to Bischoff, this disposition obtains in the Gorilla, and seems to be usual also in the Chimpanzee. In the female Chimpanzee, however, on the left side I found the parieto-occipital fissure passing into the calcarine, as in man. The frontal lobe is easily distinguished from the parietal by the fissure of Rolando, and from the temporal by the fissure of Sylvius. In the Orang it is higher, wider, and more arched than in the Chimpanzee. The anterior central convolution in front of the central fissure runs into the post-central convolution above and below, as in man. It is difficult, however, to identify the three frontal convolutions seen in man and the Chimpanzee, the frontal lobe of the Orang dividing rather into two convolutions, the middle one being badly defined. This is due somewhat to the length of the precentral fissure, which is as long as the fissure of Rolando, extending farther upward than in man. There was nothing particularly noticeable about the base of the frontal lobe; on the mesial surface it ran into the parietal. The part above the calloso-marginal fissure in the Orang is not as distinctly divided into convolutions as in man, though these are not constantly present even in all human brains. The parietal lobe is separated from the frontal by the central fissure, from the occipital and temporal incompletely, by the parieto-occipital and Sylvian fissures. The posterior-central convolution is well defined. The parietal fissure in the Orang is more striking than that of man, resembling the Gorilla's; it is twice as long as the corresponding fissure in the Chimpanzee, extending from the transverse occipital fissure, as is sometimes the case in man, almost into the fissure of Rolando. It is unbridged and without a break, and divides the parietal lobe completely into upper and lower parietal lobules. The upper parietal lobule is bounded externally by the parietal fissure; posteriorly it is separated from the occipital lobe, internally by the parieto-occipital fissure; externally it is continuous with the occipital lobe, as the first occipital gyrus, anteriorly it is separated from the posterior central convolution more completely than in man, by a fissure which runs paralled with the central fissure. There is in the Orang, also, a fissure running paralled with the parietal, which subdivides the upper parietal lobule into inner and outer portions. The precuneus, or the space on the mesial side of the parietal lobe between the parieto-occipital

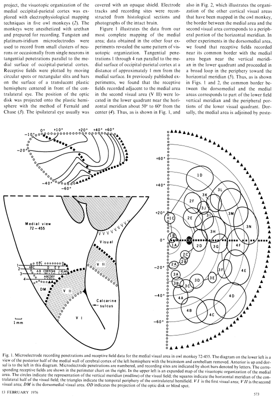
FIG. 2.

* From the Proceedings of the Academy of Natural Sciences, Phila., 1880.



Time-corrected document similarity

Representation of the Visual Field on the Medial Wall of Occipital-Parietal Cortex in the Owl Monkey (1976)



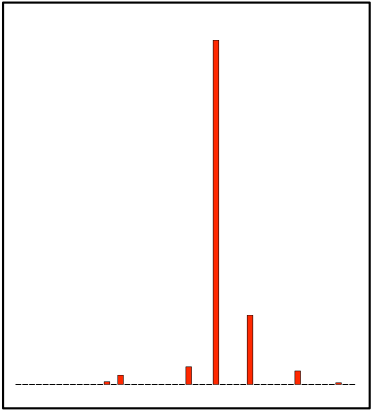
project, the visuotopic organization of the medial occipital-parietal cortex was explored with electrophysiological mapping techniques in five owl monkeys (*2*). The monkeys were anesthetized with urethan and prepared for recording. Tungsten and platinum-iridium microelectrodes were used to record from small clusters of neurons or occasionally from single neurons in tangential penetrations parallel to the medial surface of occipital-parietal cortex. Receptive fields were plotted by moving circular spots or rectangular slits and bars on the surface of a translucent plastic hemisphere centered in front of the contralateral eye. The position of the optic disk was projected onto the plastic hemisphere with the method of Fernald and Chase (*3*). The ipsilateral eye usually was covered with an opaque shield. Electrode tracks and recording sites were reconstructed from histological sections and photographs of the intact brain.

Figure 1 illustrates the data from our most complete mapping of the medial area; data obtained in the other four experiments revealed the same pattern of visuotopic organization. Tangential penetrations 1 through 4 ran parallel to the medial surface of occipital-parietal cortex at a distance of approximately 1 mm from the medial surface. In previously published experiments, we found that the receptive fields recorded adjacent to the medial area in the second visual area (V II) were located in the lower quadrant near the horizontal meridian about 50° to 60° from the center (*4*). Thus, as is shown in Fig. 1, and also in Fig. 2, which illustrates the organization of the other cortical visual areas that have been mapped in the owl monkey, the border between the medial area and the second visual area corresponds to a peripheral portion of the horizontal meridian. In other experiments in the dorsomedial area, we found that receptive fields recorded near its common border with the medial area began near the vertical meridian in the lower quadrant and proceeded in a broad loop in the periphery toward the horizontal meridian (*5*). Thus, as is shown in Figs. 1 and 2, the common border between the dorsomedial and the medial areas corresponds to part of the lower field vertical meridian and the peripheral portions of the lower visual quadrant. Dorsally, the medial area is adjoined by poste-

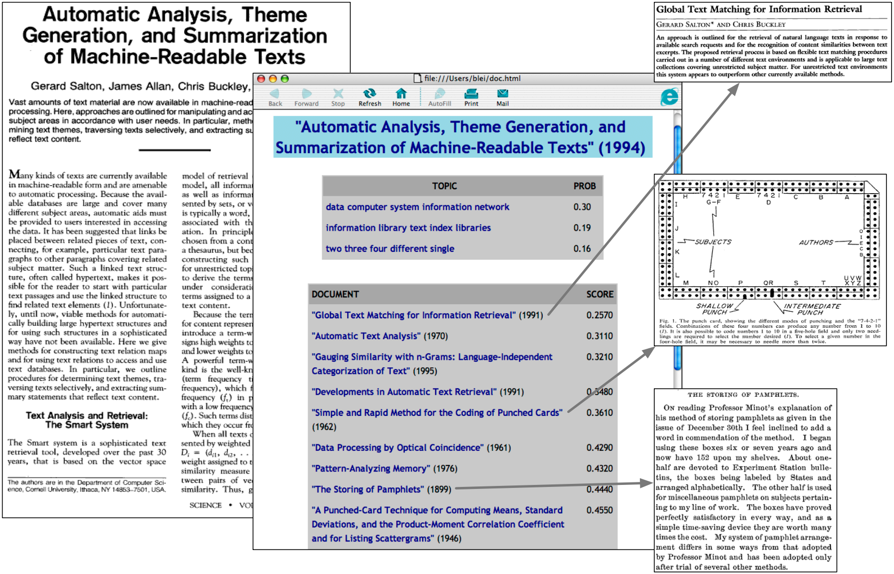
Fig. 1. Microelectrode recording penetrations and receptive field data for the medial visual area in owl monkey 72-455. The diagram on the lower left is a view of the posterior half of the medial wall of cerebral cortex of the left hemisphere with the brainstem and cerebellum removed. Anterior is up and dorsal is to the left in this diagram. Microelectrode penetrations are numbered, and recording sites are indicated by short bars denoted by letters. The corresponding receptive fields are shown in the perimeter chart on the right. In the upper left is an expanded map of the visuotopic organization of the medial area. The circles indicate the representation of the vertical meridian (midline) of the visual field; the squares indicate the horizontal meridian of the contralateral half of the visual field; the triangles indicate the temporal periphery of the contralateral hemifield. *V I* is the first visual area; *V II* is the second visual area; *DM* is the dorsomedial visual area. *OD* indicates the projection of the optic disk or blind spot.

13 FEBRUARY 1976

573



Browser of Science



Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts

Gerard Salton, James Allan, Chris Buckley,

Vast amounts of text material are now available in machine-read
processing. Here, approaches are outlined for manipulating and ac
subject areas in accordance with user needs. In particular, meth
mining text themes, traversing texts selectively, and extracting su
reflect text content.

Many kinds of texts are currently available in machine-readable form and are amenable to automatic processing. Because the available databases are large and cover many different subject areas, automatic aids must be provided to users interested in accessing the data. It has been suggested that links be placed between related pieces of text, connecting, for example, particular text paragraphs to other paragraphs covering related subject matter. Such a linked text structure, often called hypertext, makes it possible for the reader to start with particular text passages and use the linked structure to find related text elements (1). Unfortunately, until now, viable methods for automatically building large hypertext structures and for using such structures in a sophisticated way have not been available. Here we give methods for constructing text relation maps and for using text relations to access and use text databases. In particular, we outline procedures for determining text themes, traversing texts selectively, and extracting summary statements that reflect text content.

Text Analysis and Retrieval: The Smart System

The Smart system is a sophisticated text retrieval tool, developed over the past 30 years, that is based on the vector space

The authors are in the Department of Computer Science, Cornell University, Ithaca, NY 14853–7501, USA.

model of retrieval
model, all informa
as well as informa
sented by sets, or v
is typically a word,
associated with th
ation. In principle
chosen from a cont
a thesaurus, but bec
constructing such
for unrestricted top
to derive the terms
under consideratio
terms assigned to a
text content.

Because the term
for content represen
introduce a term-w
signs high weights to
and lower weights to
A powerful term-w
kind is the well-kn
(term frequency t
frequency), which f
frequency (f_t) in p
with a low frequency
(f_c). Such terms dist
which they occur fr

When all texts c
sented by weighted
$D_i = (d_{i1}, d_{i2}, . .$
weight assigned to t
similarity measure
tween pairs of ve
similarity. Thus, g

SCIENCE • VO

file:///Users/blei/doc.html

"Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts" (1994)

TOPIC	PROB
data computer system information network	0.30
information library text index libraries	0.19
two three four different single	0.16

DOCUMENT	SCORE
"Global Text Matching for Information Retrieval" (1991)	0.2570
"Automatic Text Analysis" (1970)	0.3110
"Gauging Similarity with n-Grams: Language-Independent Categorization of Text" (1995)	0.3210
"Developments in Automatic Text Retrieval" (1991)	0.3480
"Simple and Rapid Method for the Coding of Punched Cards" (1962)	0.3610
"Data Processing by Optical Coincidence" (1961)	0.4290
"Pattern-Analyzing Memory" (1976)	0.4320
"The Storing of Pamphlets" (1899)	0.4440
"A Punched-Card Technique for Computing Means, Standard Deviations, and the Product-Moment Correlation Coefficient and for Listing Scattergrams" (1946)	0.4550

Global Text Matching for Information Retrieval

Gerard Salton* and Chris Buckley

An approach is outlined for the retrieval of natural language texts in response to available search requests and for the recognition of content similarities between text excerpts. The proposed retrieval process is based on flexible text matching procedures carried out in a number of different text environments and is applicable to large text collections covering unrestricted subject matter. For unrestricted text environments this system appears to outperform other currently available methods.

Fig. 1. The punch card, showing the different modes of punching and the "7-4-2-1" fields. Combinations of these four numbers can produce any number from 1 to 10 (1). It is also possible to code numbers 1 to 10 in a five-hole field and only two needlings are required to select the number desired (1). To select a given number in the four-hole field, it may be necessary to needle more than twice.

THE STORING OF PAMPHLETS.

On reading Professor Minot's explanation of his method of storing pamphlets as given in the issue of December 30th I feel inclined to add a word in commendation of the method. I began using these boxes six or seven years ago and now have 152 upon my shelves. About one-half are devoted to Experiment Station bulletins, the boxes being labeled by States and arranged alphabetically. The other half is used for miscellaneous pamphlets on subjects pertaining to my line of work. The boxes have proved perfectly satisfactory in every way, and as a simple time-saving device they are worth many times the cost. My system of pamphlet arrangement differs in some ways from that adopted by Professor Minot and has been adopted only after trial of several other methods.

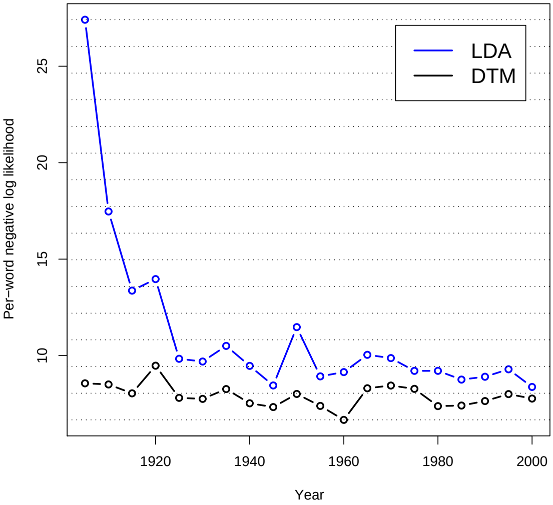
Quantitative comparison

- Compute the probability of each year's documents conditional on all the previous year's documents,

$$p(\mathbf{w}_t \mid \mathbf{w}_1, \ldots, \mathbf{w}_{t-1})$$

- Compare exchangeable and dynamic topic models

Quantitative comparison



Dynamic topic models discussion

- The DTM is a hierarchical model of sequential document collections;
- Exchangeability assumptions should be taken seriously.
- Variational methods allow large scale posterior inference.
- Examining the latent structure yields useful browsing tools
- Some open issues
    - Model selection: choosing the number of topics
    - Variational inference: what are the hidden assumptions?

Outline

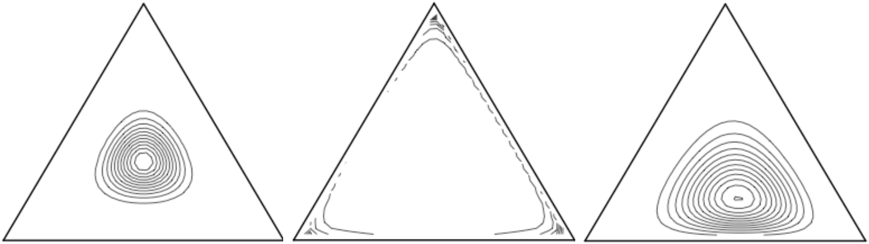
1. Introduction

2. Latent Dirichlet allocation

3. Dynamic topic models

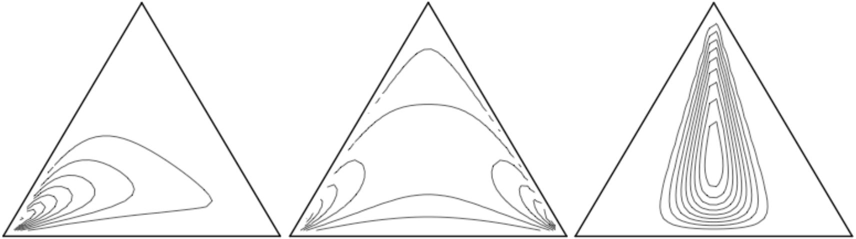
4. Correlated topic models

The hidden assumptions of the Dirichlet distribution



- The Dirichlet is an exponential family distribution on the *simplex*, positive vectors that sum to one.
- However, the near independence of components makes it a poor choice for modeling topic proportions.
- An article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

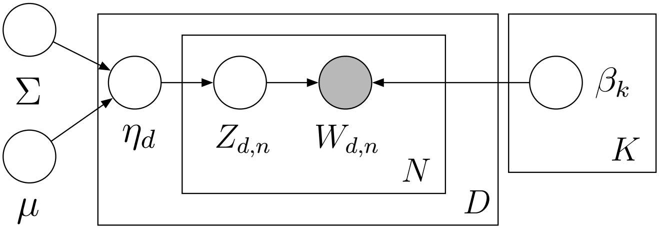
The logistic normal distribution



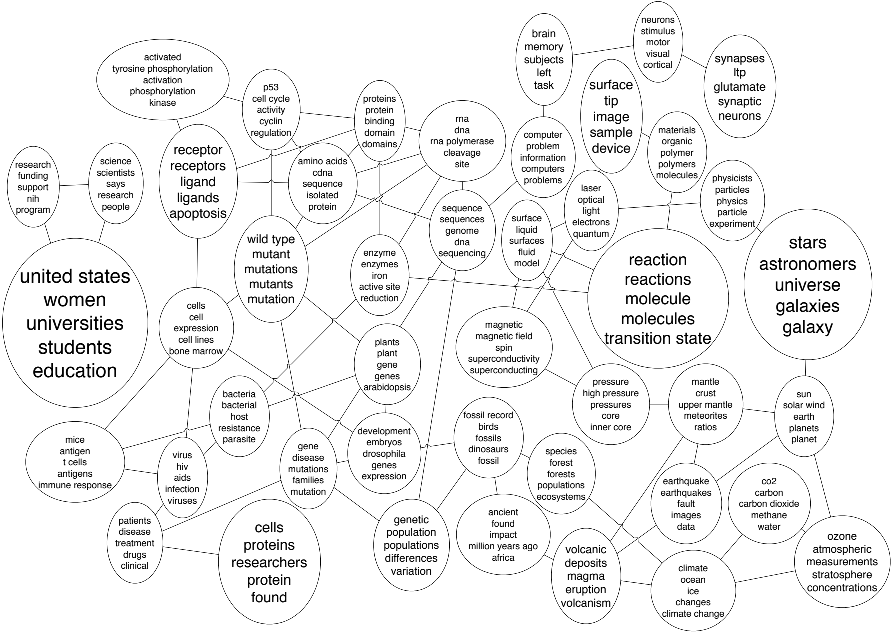
- The logistic normal is a distribution on the simplex that can model dependence between components.
- The natural parameters of the multinomial are drawn from a multivariate Gaussian distribution.

$$
\begin{aligned}
X &\sim \mathcal{N}_{K-1}(\mu, \Sigma) \\
\theta_i &= \exp\{x_i - \log(1 + \textstyle\sum_{j=1}^{K-1} \exp\{x_j\})\}
\end{aligned}
$$

Correlated topic model (CTM)



- Draw topic proportions from a logistic normal, where topic occurrences can exhibit correlation.
- Use for:
    - Providing a “map” of topics and how they are related
    - Better prediction via correlated topics



activated
tyrosine phosphorylation
activation
phosphorylation
kinase

p53
cell cycle
activity
cyclin
regulation

proteins
protein
binding
domain
domains

rna
dna
rna polymerase
cleavage
site

brain
memory
subjects
left
task

neurons
stimulus
motor
visual
cortical

synapses
ltp
glutamate
synaptic
neurons

surface
tip
image
sample
device

computer
problem
information
computers
problems

materials
organic
polymer
polymers
molecules

research
funding
support
nih
program

science
scientists
says
research
people

receptor
receptors
ligand
ligands
apoptosis

amino acids
cdna
sequence
isolated
protein

laser
optical
light
electrons
quantum

physicists
particles
physics
particle
experiment

sequence
sequences
genome
dna
sequencing

surface
liquid
surfaces
fluid
model

wild type
mutant
mutations
mutants
mutation

enzyme
enzymes
iron
active site
reduction

reaction
reactions
molecule
molecules
transition state

stars
astronomers
universe
galaxies
galaxy

united states
women
universities
students
education

cells
cell
expression
cell lines
bone marrow

magnetic
magnetic field
spin
superconductivity
superconducting

plants
plant
gene
genes
arabidopsis

pressure
high pressure
pressures
core
inner core

mantle
crust
upper mantle
meteorites
ratios

sun
solar wind
earth
planets
planet

bacteria
bacterial
host
resistance
parasite

fossil record
birds
fossils
dinosaurs
fossil

mice
antigen
t cells
antigens
immune response

virus
hiv
aids
infection
viruses

gene
disease
mutations
families
mutation

development
embryos
drosophila
genes
expression

species
forest
forests
populations
ecosystems

earthquake
earthquakes
fault
images
data

co2
carbon
carbon dioxide
methane
water

patients
disease
treatment
drugs
clinical

cells
proteins
researchers
protein
found

genetic
population
populations
differences
variation

ancient
found
impact
million years ago
africa

volcanic
deposits
magma
eruption
volcanism

climate
ocean
ice
changes
climate change

ozone
atmospheric
measurements
stratosphere
concentrations

Summary

- Topic models provide useful descriptive statistics for analyzing and understanding the latent structure of large text collections.
- More generally, probabilistic graphical models are a useful way to express assumptions about the hidden structure of complicated data.
- Variational methods allow us to perform posterior inference to automatically infer that structure from large data sets.
- Current research
    - Choosing the number of topics
    - Continuous time dynamic topic models
    - Topic models for prediction
    - Inferring the impact of a document

“We should seek out unfamiliar summaries of observational material, and establish their useful properties... And still more novelty can come from finding, and evading, still deeper lying constraints.” (Tukey, 1962)

Diversion: Variational inference

- Let $x_{1:N}$ be observations and $z_{1:M}$ be latent variables
- Our goal is to compute the posterior distribution

$$p(z_{1:M} \mid x_{1:N}) = \frac{p(z_{1:M}, x_{1:N})}{\int p(z_{1:M}, x_{1:N}) dz_{1:M}}$$

- For many interesting distributions, the marginal likelihood of the observations is difficult to efficiently compute

Variational inference

- Use Jensen's inequality to bound the log prob of the observations:

$$\log p(x_{1:N}) \geq \mathrm{E}_{q_\nu}[\log p(z_{1:M}, x_{1:N})] - \mathrm{E}_{q_\nu}[\log q_\nu(z_{1:M})].$$

- We have introduced a distribution of the latent variables with free *variational parameters* ν.
- We optimize those parameters to tighten this bound.
- This is the same as finding the member of the family q_ν that is closest in KL divergence to $p(z_{1:M} \mid x_{1:N})$.

Mean-field variational inference

- Complexity of optimization is determined by the factorization of q_ν
- In *mean field variational inference* we choose q_ν to be fully factored

$$q_\nu(z_{1:M}) = \prod_{m=1}^{M} q_{\nu_m}(z_m).$$

- The latent variables are independent.
  - Each is governed by its own variational parameter ν_m.
- In the true posterior they can exhibit dependence
  (often, this is what makes exact inference difficult).

MFVI and conditional exponential families

- Suppose the distribution of each latent variable conditional on the observations and other latent variables is in the exponential family:

$$p(z_m \mid \mathbf{z}_{-m}, \mathbf{x}) = h_m(z_m) \exp\{g_m(\mathbf{z}_{-m}, \mathbf{x})^T z_m - a_m(g_i(\mathbf{z}_{-m}, \mathbf{x}))\}$$

- Assume q_ν is fully factorized, and each factor is in the same exponential family:

$$q_{\nu_m}(z_m) = h_m(z_m) \exp\{\nu_m^T z_m - a_m(\nu_m)\}$$

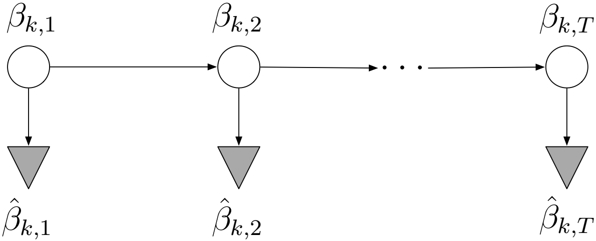
MFVI and conditional exponential families

- Variational inference is the following coordinate ascent algorithm

$$\nu_m = \mathrm{E}_{q_\nu}[g_m(\mathbf{Z}_{-m}, \mathbf{x})]$$

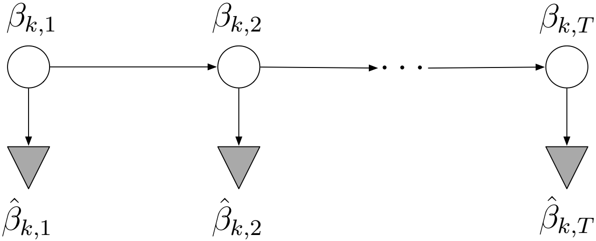
- Notice the relationship to Gibbs sampling

Variational family for the DTM



- Distribution of θ and z is fully-factorized (Blei et al., 2003)
- Distribution of $\{\beta_{1,k}, \ldots, \beta_{T,k}\}$ is a *variational Kalman filter*
- Gaussian state-space model with free *observations* $\hat{\beta}_{k,t}$.
- Fit observations such that the corresponding posterior over the chain is close to the true posterior.

Variational family for the DTM



- Given a document collection, use coordinate ascent on all the variational parameters until the KL converges.
- Yields a distribution close to the true posterior of interest
- Take expectations w/r/t the simpler variational distribution