Bioinformatics

(Globex, Summer 2015)

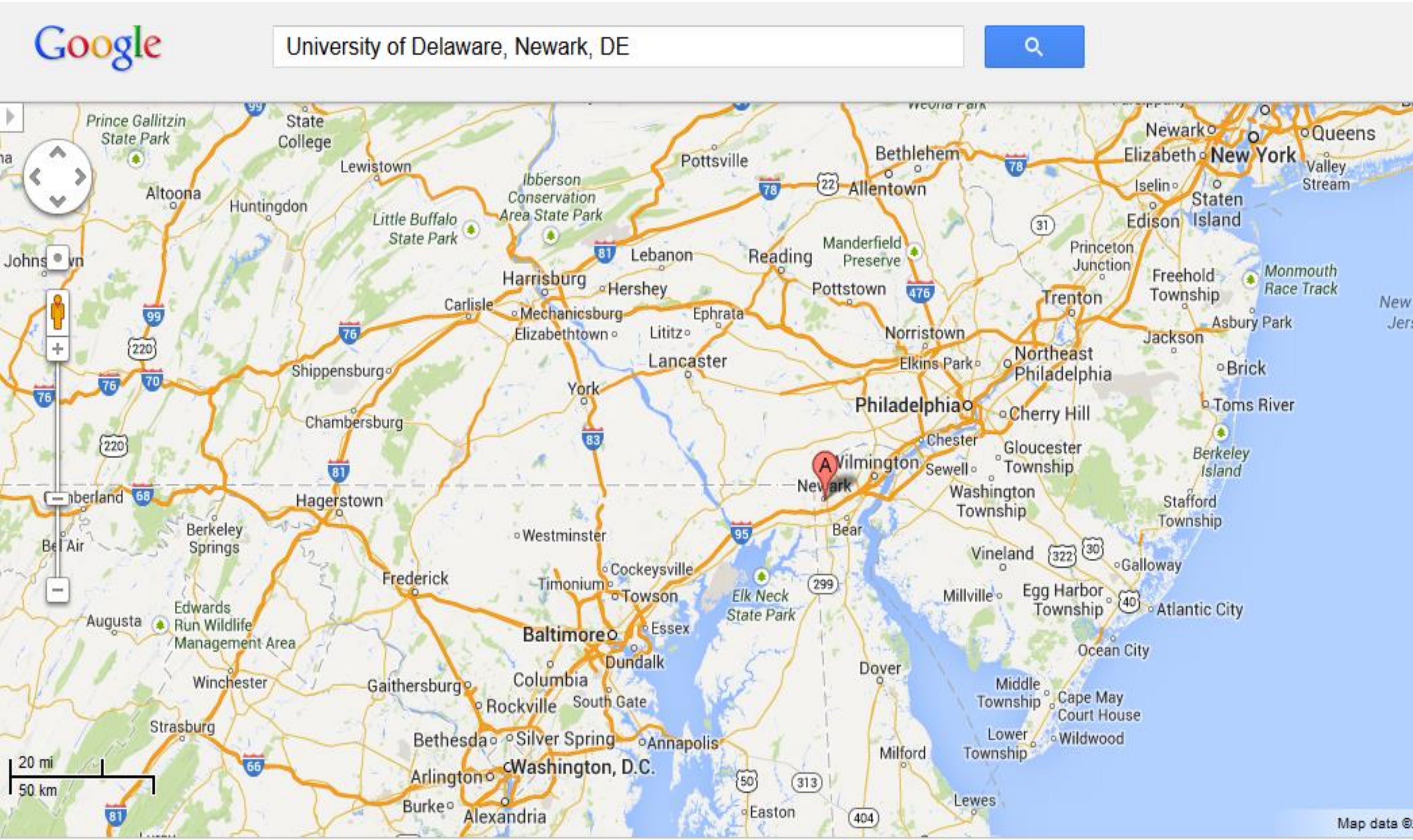
Lecture 1

Course Overview

Li Liao
Computer and Information Sciences
University of Delaware

Dela-where?


Loading...
Calgary
Ontario
Quebec
Winnipeg
Vancouver
Washington
Montana
North Dakota
Minnesota
Montreal
New Brunswick
Nova Scotia
Maine
Wisconsin
Toronto
Oregon
Idaho
Wyoming
South Dakota
Michigan
New York
Vermont
New Hampshire
Iowa
Chicago
Nebraska
Illinois
Pennsylvania
Massachusetts
Ohio
New York
Rhode Island
Nevada
United States
Indiana
San Francisco
Utah
Colorado
Kansas
Missouri
West Virginia
Connecticut
California
Kentucky
Virginia
New Jersey
Oklahoma
Arkansas
Tennessee
North Carolina
Delaware
Los Angeles
Arizona
New Mexico
Maryland
Dallas
Mississippi
South Carolina
District of Columbia
Phoenix
Alabama
San Diego
Texas
Georgia
Louisiana
San Antonio
Houston
Florida
Gulf of California
Monterrey
Gulf of Mexico
Mexico
Havana
Cuba
Guadalajara
Mexico City
Dominican Republic

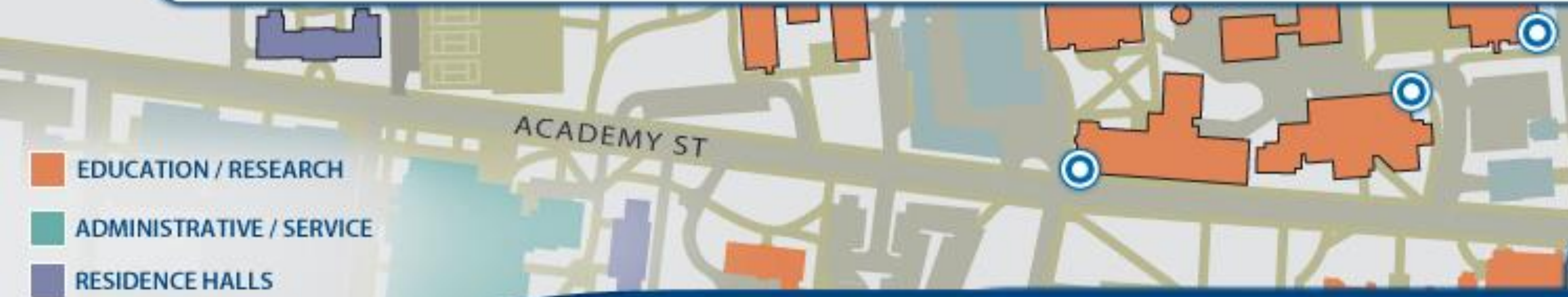


University of Delaware, Newark, DE

UNIVERSITY OF DELAWARE

SMITH HALL

Smith Hall is named in memory of Edward Laurence Smith, dean of Delaware College from 1915-23; it features a campus computer center and is home to the Departments of Computer and Information Science and of Political Science and International Relations.



🟧 EDUCATION / RESEARCH
🟩 ADMINISTRATIVE / SERVICE
🟪 RESIDENCE HALLS
🟫 ATHLETICS

VIDEO TOURS

Administrative stuff

- Syllabus and tentative schedule
- Workload
  - 4 homework assignments (40%)
  - Mid-term exam (20%)
  - Final exam (40%)
- Late policy: Only one day delay is allowed; subject to 15% penalty.

Bioinformatics Books

- **Markketa Zvelebil and Jeremy Baum, Understanding Bioinformatics, Garland Science, 2008.**
- Dan E. Krane & Michael L. Raymer, Fundamental Concepts of Bioinformatics, Benjamin Cummings 2002
- R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1998.
- João Meidanis & João Carlos Setubal. Introduction to Computational Molecular Biology. PWS Publishing Company, Boston, 1996.
- Peter Clote and Rolf Backofen, Computational Molecular Biology: An Introduction, Willey 2000.
- Dan Gusfield. Algorithms on String, Trees, and Sequences. Cambridge University Press, 1997.
- P. Baldi and S. Brunak, Bioinformatics, The Machine Learning Approach, The MIT press, 1998.
- D.W. Mount, Bioinformaics: Sequence and Genome Analysis, CSHLP 2004.

Molecular Biology Books

Free materials:

- Kimball's biology
- Lawrence Hunter: Molecular biology for computer scientists
- DOE’s Molecular Genetics Primer

Books:

- Instant Notes series: Biochemistry, Molecular Biology, and Genetics
- Molecular Biology of The Cell, by Alberts et al

Bioinformatics

- use and develop computing methods to solve biological problems

The field is characterized by

- an explosion of data → Big data!!!
- difficulty in interpreting the data
- a large number of open problems
- until recently, relative lack of sophistication of computational techniques (compared with, say, signal processing, graphics, etc.)

Why is this course good for you?

- Expand your knowledge base
  - Bioinformatics is a computational wing of biotechnology.
  - Computational techniques in bioinformatics are widely useful elsewhere
    - - Cybergenomics: detect, dissect malware
    - Big data!!!
- Job market is strong ☺

Industry is moving in

- IBM:
    - BlueGene, the fastest computer with 1 million CPU
    - Blueprint worldwide collects all the protein information
    - Bioinformatics segment will be $40 billion in 2004 up from $22 billion in 2000
- GlaxoSmithKline
- Celera
- Merck
- AstraZeneca
- …

Computing and IT skills

- Algorithm design and model building
- Working with unix system/Web server
- Programming (in PERL, Java, etc.)
- RDBMS: SQL, Oracle PL/SQL

People

- International Society for Computational Biology (www.iscb.org) ~ 3400 members (as of 2015)
- Severe shortage for qualified bioinformatians / bioinformaticists

Conferences

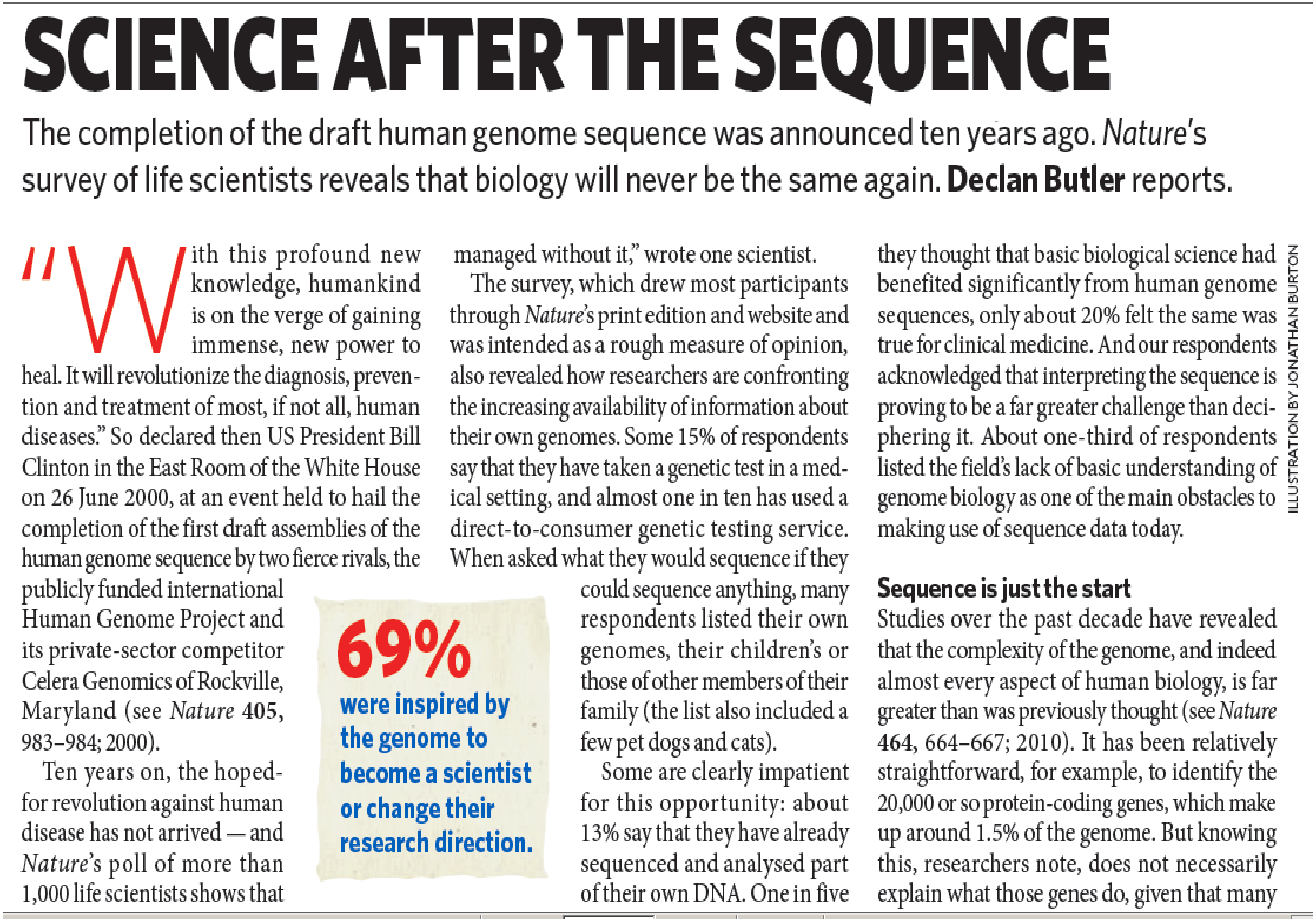
- ISMB (Intelligent Systems for Molecular Biology) started in 1992
- RECOMB (International Conference on Computational Molecular Biology) started in 1997
- PSB (Pacific Symposium on Biocomputing) started 1996
- ACM-BCB
- BIBM …
- …

Journals

- Bioinformatics
- BMC Bioinformatics
- PLOS Computational Biology
- Journal of Computational Biology
- Genome Biology
- Genomics
- Genome Research
- Nucleic Acids Research
- ...

A short history after human genome: 2000 -- 2010

Human Genome at Ten:



SCIENCE AFTER THE SEQUENCE

The completion of the draft human genome sequence was announced ten years ago. *Nature*'s survey of life scientists reveals that biology will never be the same again. **Declan Butler** reports.

"With this profound new knowledge, humankind is on the verge of gaining immense, new power to heal. It will revolutionize the diagnosis, prevention and treatment of most, if not all, human diseases." So declared then US President Bill Clinton in the East Room of the White House on 26 June 2000, at an event held to hail the completion of the first draft assemblies of the human genome sequence by two fierce rivals, the publicly funded international Human Genome Project and its private-sector competitor Celera Genomics of Rockville, Maryland (see *Nature* **405**, 983–984; 2000).

69% were inspired by the genome to become a scientist or change their research direction.

Ten years on, the hoped-for revolution against human disease has not arrived — and *Nature*'s poll of more than 1,000 life scientists shows that

managed without it," wrote one scientist.

The survey, which drew most participants through *Nature*'s print edition and website and was intended as a rough measure of opinion, also revealed how researchers are confronting the increasing availability of information about their own genomes. Some 15% of respondents say that they have taken a genetic test in a medical setting, and almost one in ten has used a direct-to-consumer genetic testing service. When asked what they would sequence if they could sequence anything, many respondents listed their own genomes, their children's or those of other members of their family (the list also included a few pet dogs and cats).

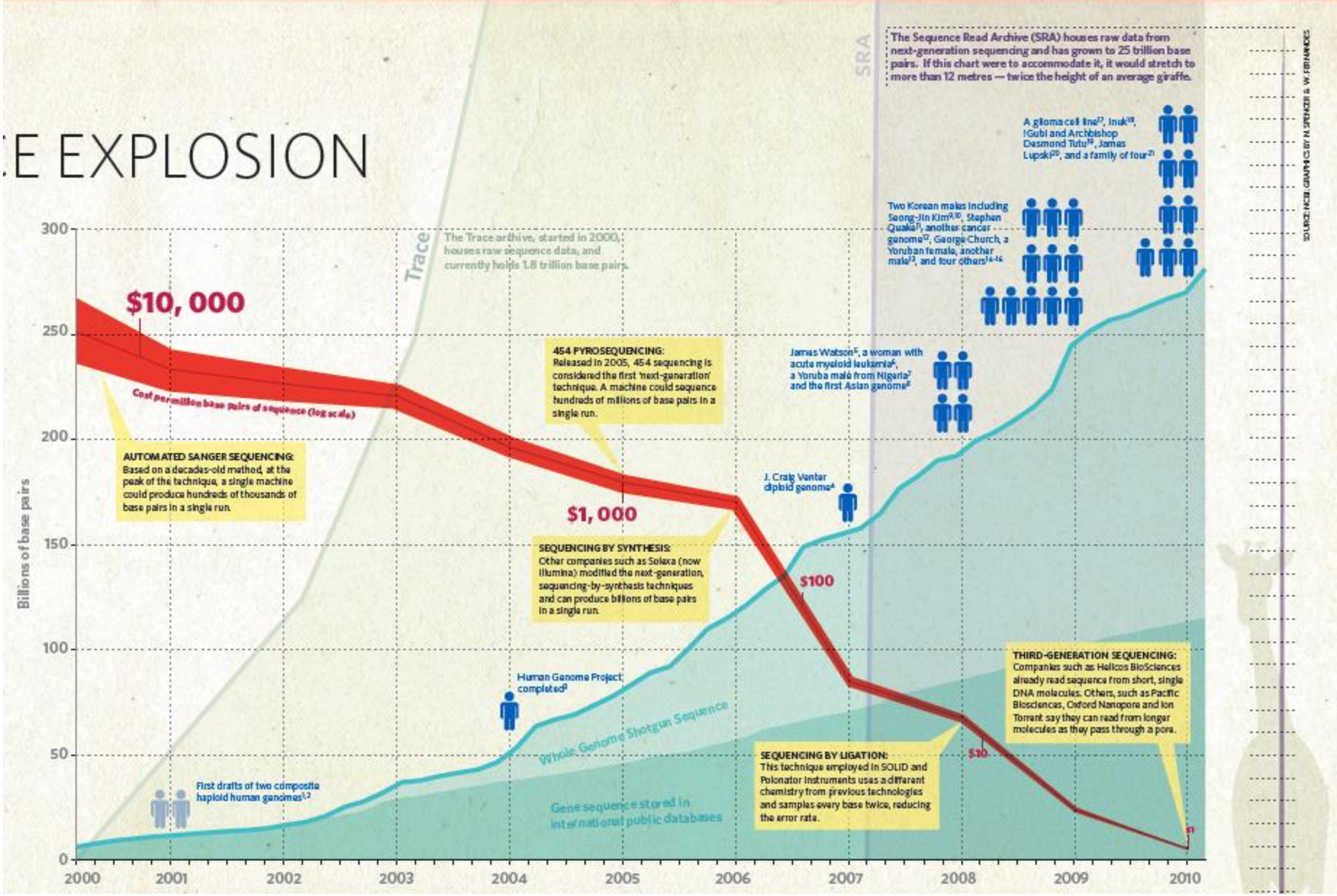
Some are clearly impatient for this opportunity: about 13% say that they have already sequenced and analysed part of their own DNA. One in five they thought that basic biological science had benefited significantly from human genome sequences, only about 20% felt the same was true for clinical medicine. And our respondents acknowledged that interpreting the sequence is proving to be a far greater challenge than deciphering it. About one-third of respondents listed the field's lack of basic understanding of genome biology as one of the main obstacles to making use of sequence data today.

Sequence is just the start

Studies over the past decade have revealed that the complexity of the genome, and indeed almost every aspect of human biology, is far greater than was previously thought (see *Nature* **464**, 664–667; 2010). It has been relatively straightforward, for example, to identify the 20,000 or so protein-coding genes, which make up around 1.5% of the genome. But knowing this, researchers note, does not necessarily explain what those genes do, given that many

ILLUSTRATION BY JONATHAN BURTON

E EXPLOSION



The Sequence Read Archive (SRA) houses raw data from next-generation sequencing and has grown to 25 trillion base pairs. If this chart were to accommodate it, it would stretch to more than 12 metres — twice the height of an average giraffe.

The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.

AUTOMATED SANGER SEQUENCING: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

454 PYROSEQUENCING: Released in 2005, 454 sequencing is considered the first 'next-generation' technique. A machine could sequence hundreds of millions of base pairs in a single run.

SEQUENCING BY SYNTHESIS: Other companies such as Solexa (now Illumina) modified the next-generation, sequencing-by-synthesis techniques and can produce billions of base pairs in a single run.

SEQUENCING BY LIGATION: This technique employed in SOLID and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

THIRD-GENERATION SEQUENCING: Companies such as Helicos BioSciences already read sequence from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.

First drafts of two composite haploid human genomes[1,2]

Human Genome Project completed[3]

J. Craig Venter diploid genome[4]

James Watson[5], a woman with acute myeloid leukemia[6], a Yoruba male from Nigeria[7] and the first Asian genome[8]

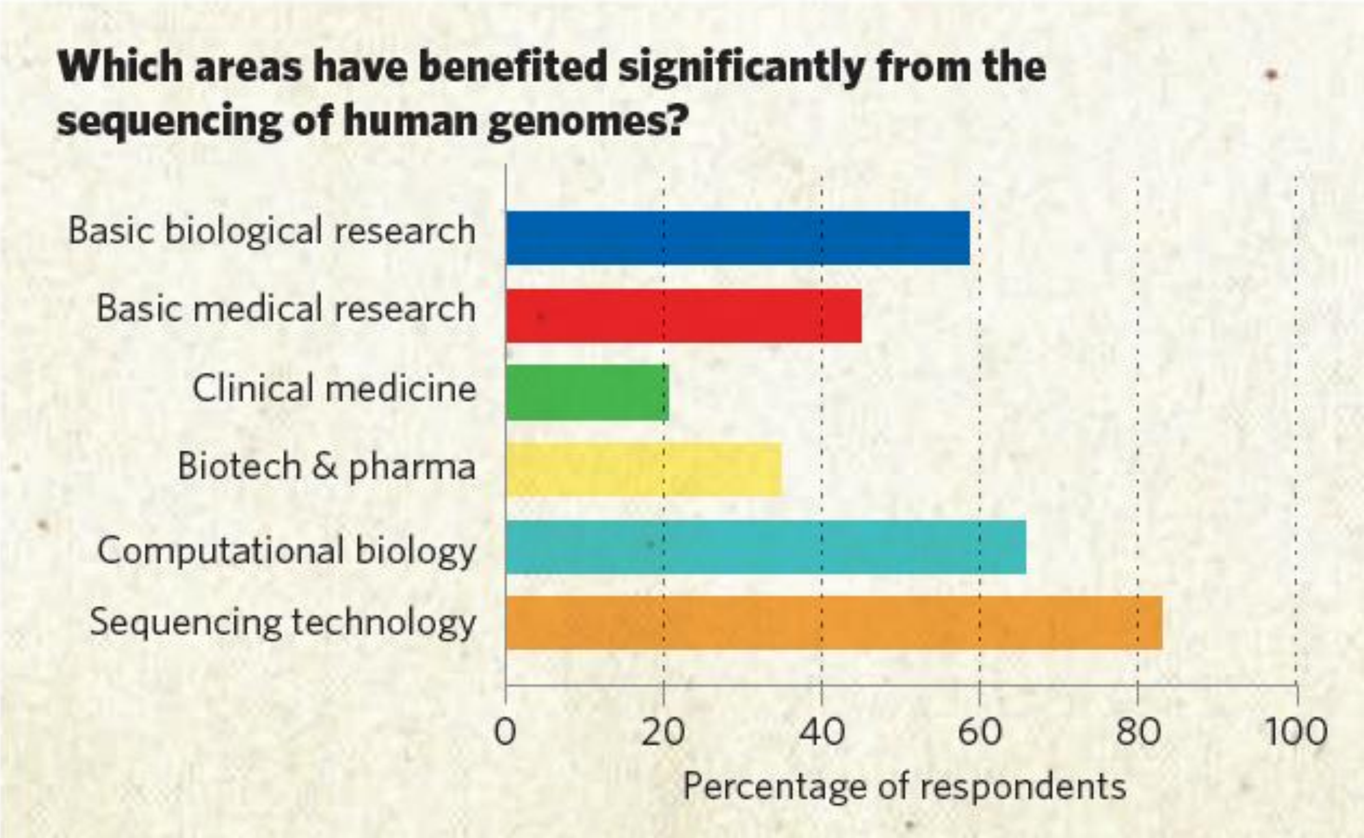
Two Korean males including Seong-Jin Kim[9,10], Stephen Quake[11], another cancer genome[12], George Church, a Yoruban female, another male[13], and four others[14-16]

A glioma cell line[17], Inuk[18], !Gubi and Archbishop Desmond Tutu[19], James Lupski[20], and a family of four[21]

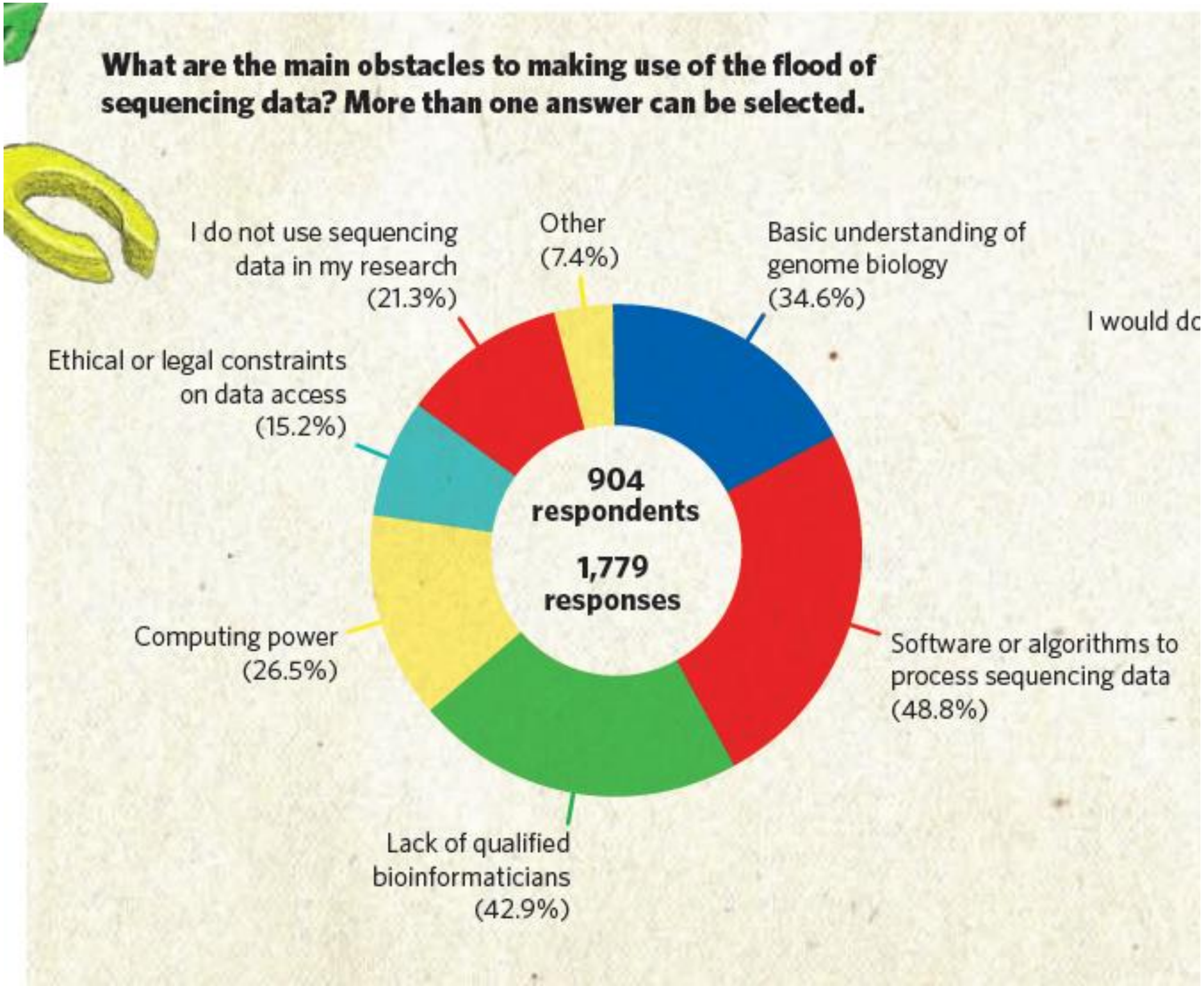
SOURCE: NCBI; GRAPHICS BY N. SPENCER & W. FERNANDES

Human Genome at Ten:

Human Genome at Ten:



Human Genome at Ten:



careers and recruitment

Nature **389**, 420 - 422 (1997) © Macmillan Publishers Ltd.

Running to catch up in Europe

HELEN GAVAGHAN

Helen Gavaghan is a science and technology writer based in Hebden Bridge, Yorkshire, UK.

Across Europe, the story is the same. Demand for those skilled in bioinformatics exceeds supply. Like biochemistry and biophysics befo
the barriers between traditional academic fields, and demanding flexibility and a new way of thinking from its adherents.

Computational biology has meant different things to different people. Not too long ago, says Hans Prydz of the University of Oslo's Biote
handling NMR data or analysing Doppler echograms. Now renamed bioinformatics, it means looking for patterns in DNA and RNA, pr
modelling proteins and mining massive databases that continue to grow. When the DNA database run by the European Bioinformatics In
contained 700,000 nucleotides: now there are more than a billion.

Driven by the scientific and commercial importance of bioinformatics in genomics and drug discovery and development, governments, un
responding with varying degrees of vigour and success to the skills shortage and are seeking ways to cross the boundaries between disci
physics, mathematics, computer science, statistics, protein chemistry, genetics and molecular biology.

At European level, the EBI, based near Cambridge (United Kingdom), is funded to the sum of about DM 9 million ($5 million) by memb
Israel via their contributions to the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. Contributions from the ph
industries roughly double the institute's income. The EBI, an offshoot of EMBL, develops tools for bioinformatics, seeks innovative ways
training courses for academics and industrialists. Initiatives with industry include the Industry Affiliates Initiative, which helps small and me
and apply new techniques; the BioTitan Project, running nodes to enable faster access to databases; and the Biostandards project, funde
European Union for promoting and developing standards.

National initiatives also exist, particularly in the United Kingdom and Germany. Says Andrew Lyall, responsible for bioinformatics at Gla
in pretty good shape." There are two government-financed initiatives in the United Kingdom, both of which received a second lease of lif

One of these schemes, supported by the Biotechnology and Biological Sciences Research Council (BBSRC), coordinates the UK bioinf
the scheme has concentrated on developing software that would enable biologists without information technology (IT) skills to use some
their trade that are found on the World Wide Web. At a meeting earlier this month, the steering committee of the scheme decided to cha
Brass, who runs a masters' degree course in bioinformatics at the University of Manchester and is a member of the committee, says, "We

Nature **404**, 686 - 687 (2000) © Macmillan Publishers Ltd.

Training: United States gives priority to skills shortage

POTTER WICKWARE

Bioinformatics marries together a wide range of scientific disciplines, but with a global shortage of skilled researchers, trai

[WASHINGTON] Industry is draining bioinformatics talent from universities faster than it can be replenished. This is good news for the peo
news for the institutions that are scrambling to provide it, says Francis Ouellette, at the University of British Columbia's Center for Molec
Ouellette and Christoph Sensen at Canadian Bioinformatics Resource, in Halifax, Nova Scotia, run a four-part survey series (one week
genomics, proteomics and tools development), which introduces people to the field. Ouellette worries that the series is only a temporary

Sensen stresses the difficulties academic groups have in finding and retaining talent. "In two years of looking I haven't found a person will
environment. PhDs either go to a company or to a nice warm place in the United States where they also get more money. But there is an
academia because that's where much of the real science is done."

Chris Lee, of the Bioinformatics Institute at the University of California, Los Angeles, concurs. Industry has the data, he says. But it lacks
full-service university, as well as the freedom to "sit around talking about problems with people from different backgrounds".

The gap between supply and demand in bioinformatics is receiving official recognition in the United States. The US National Institutes of
bioinformatics mainly through two institutes, the National Human Genome Research Institute and the National Library of Medicine. How
centres outside the NIH must also arise. The NIH approves the concept of developing such "centres of excellence", but has been slow t
infrastructure.

The National Institute of General Medical Sciences has also committed itself to funding training slots, and a fourth branch of the NIH, the
Resources (which is not an institute), has put itself behind shared bio-computational resources at more than a dozen centres nationwide.
Argonne and Oak Ridge laboratories are also huge funders of bioinformatics work, as is, to a somewhat smaller extent, the Department

On the private side, the Howard Hughes Medical Institute (HHMI) has declared that it will appoint investigators in computational biolog
that until now has avoided funding research in what it viewed as engineering disciplines. Now, however, it is becoming clear that biocom
HHMI's biomedical mission, but is one of its most critical elements.

Other support is also issuing from the Alfred P. Sloan Foundation, which has recently called for proposals to fund academic units that cr
in biology. Traditionally, these degrees have not carried the same weight in biology as in engineering or business, where they are terminal

news

Nature **410**, 293 (2001); doi:10.1038/35066710

Singapore invests in bioinformatics

DAVID CYRANOSKI

[TOKYO] Plans for a new institute in Singapore could address one of the most acute skills shortages in science by producing up to 100 tra

The planned Bioinformatics Institute is part of Singapore's US$1 billion-a-year effort to turn the island into a powerhouse of biomedical research. Within five years, the institute should be delivering 100 masters degrees in bioinformatics. This is more than any other institution in the world, says Limsoon Wong, director of the Kent Ridge Digital Bioinformatics Laboratories, and one of the planners behind the institute.

Gunar
directo
bioinfo

According to Wong, training at the institute will go well beyond the curating of data. "We will be training people how to make predictions from the data concerning interaction between proteins, and how to use these data to drive experiments," he says.

The research and teaching institute will be housed temporarily at first, before moving to the planned 'biopolis' science park near the National University of Singapore, when it opens in two years time. The government has yet to announce its funding level, but the park is expected to start with a grant of around S$100 million (US$60 million).

The institute is likely to absorb the existing bioinformatics centre at the National University and to help service the nation's expanding genomics programme. Its own research programme will follow from the interests of the staff who will be recruited internationally.

Gunaretnam Rajagopal, a theoretical physicist at Cambridge University, has been named as the institute's deputy director, and starts wor to build a world-class research organization, with strong encouragement, commitment and active support of the government of Singapore

naturejobs **careers**

Naturejobs **413**, 25 October 2001, 8 (2001); doi:10.1038/35101788

Playing catch-up

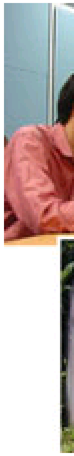
ROBERT TRIENDL

Robert Triendl is a freelance writer based in Tokyo.

Japan's government is belatedly realizing that it needs to increase funding for training in bioinformatics, says Robert Triend field could hinder the country's efforts.

Well-trained bioinformatics specialists in Japan are not just rare — they are virtually non-existent. This is partly because of a lack of formal education in the subject, and the problem is systemic. With little formal recognition of bioinformatics as a field, graduate departments have until recently allocated only a limited number of students to existing bioinformatics teachers. The government recognized the need for more bioinformaticians as it scaled up the country's genomics efforts.

CBRC



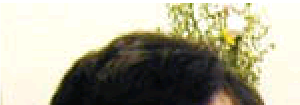
Resea
Biolog
joining
bioinfo
toward

This year, Japan's Ministry of Education, Science, Sports and Culture started to upgrade bioinformatics education at national universities by creating additional staff positions and funding both undergraduate courses and graduate-level informatics training.

Kyoto University's new bioinformatics centre is a product of this new policy. The university is Japan's leading academic centre for bioinformatics research, but until a few months ago all its bioinformatics activities were concentrated in just one laboratory: Minoru Kanehisa's lab at the Institute for Chemical Research.

Tokyo University also plans to increase its bioinformatics education and training activities. And the private Keio University has set up a whole new campus focusing on systems biology and the dynamic modelling of biological systems such as human blood cells.

Part of the ministry's promotion and coordination fund, the programme will provide between US$1 million and US$2 million in additional funding over several years for undergraduate and graduate education in bioinformatics, systems biology, protein functional analysis and software development.



MAR 2001

FULLTEXT

PDF

FIGURES / TABLES

Careers and Recruitment

March 2001 Volume 19 Number 3 pp 285 - 286

University bioinformatics programs on the rise

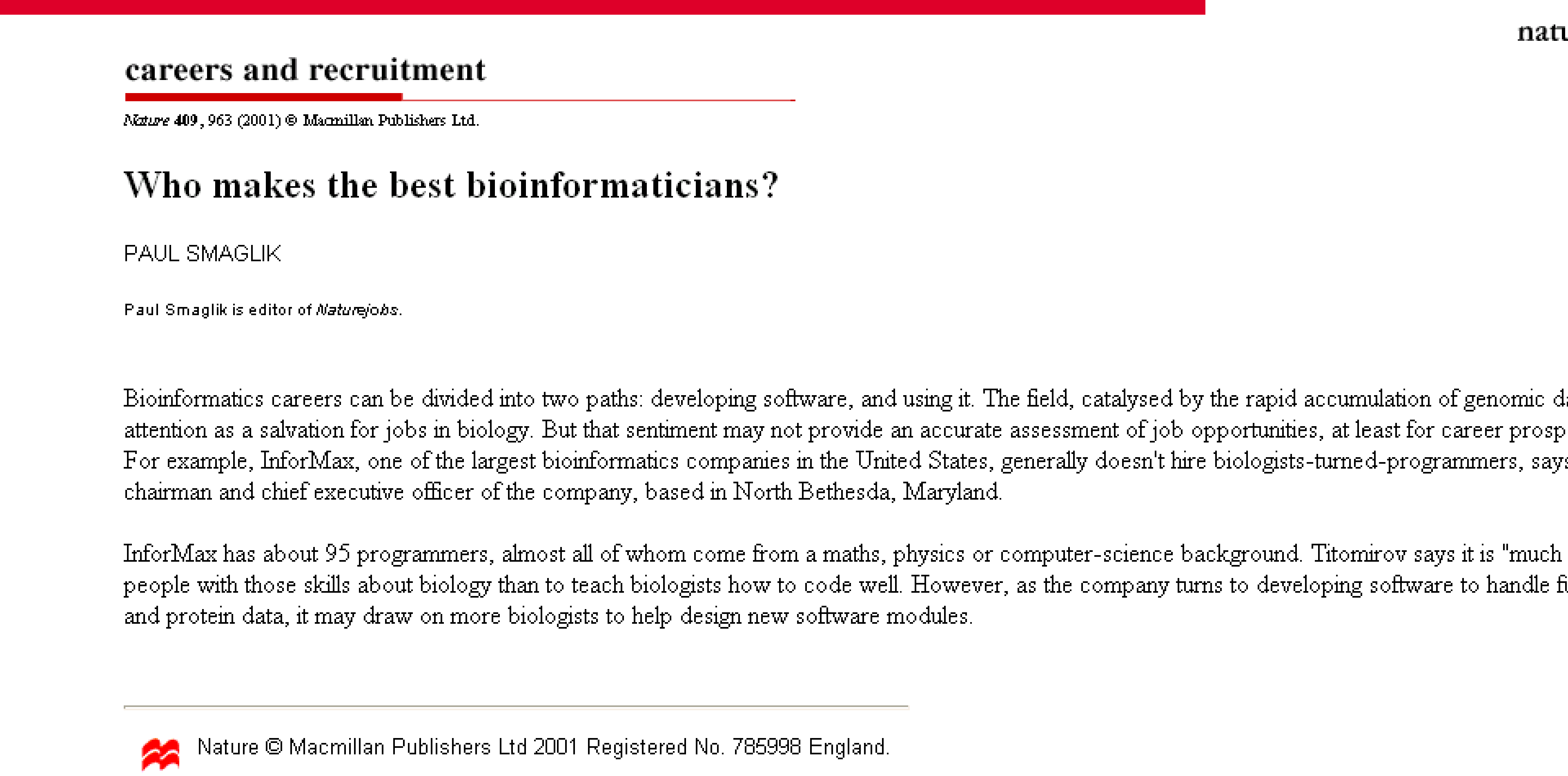
Randy J. Zauhar

Randy J. Zauhar is associate professor of biochemistry and director of the graduate program in bioinformatics, Department of Chemistry and Biochemistry, University of the Sciences in Philadelphia, 600 S. 43rd Street, Philadelphia, PA 19104-4495 (e-mail: r.zauhar@usip.edu).

Fueled by strong demand from students and industry's need for trained bioinformaticists, universities are increasing their offerings in this fast-growing field.

Information is found in the arrangement of things with respect to each other. Whether we observe the arrangement of ink on the page to form words, or the order of nucleotide bases to form a gene sequence, we recognize that information is the key to understanding the world around us. Bioinformatics studies how information is stored, reproduced, and used by living systems. It is not an overstatement to say that bioinformatics is what biology is evolving to become in the 21st century.

The subject matter of bioinformatics has existed as long as there has been molecular biology, but it has emerged as a distinct discipline only in the last two decades. This was prompted by a veritable explosion of information triggered by



natu

careers and recruitment

Nature **409**, 963 (2001) © Macmillan Publishers Ltd.

Who makes the best bioinformaticians?

PAUL SMAGLIK

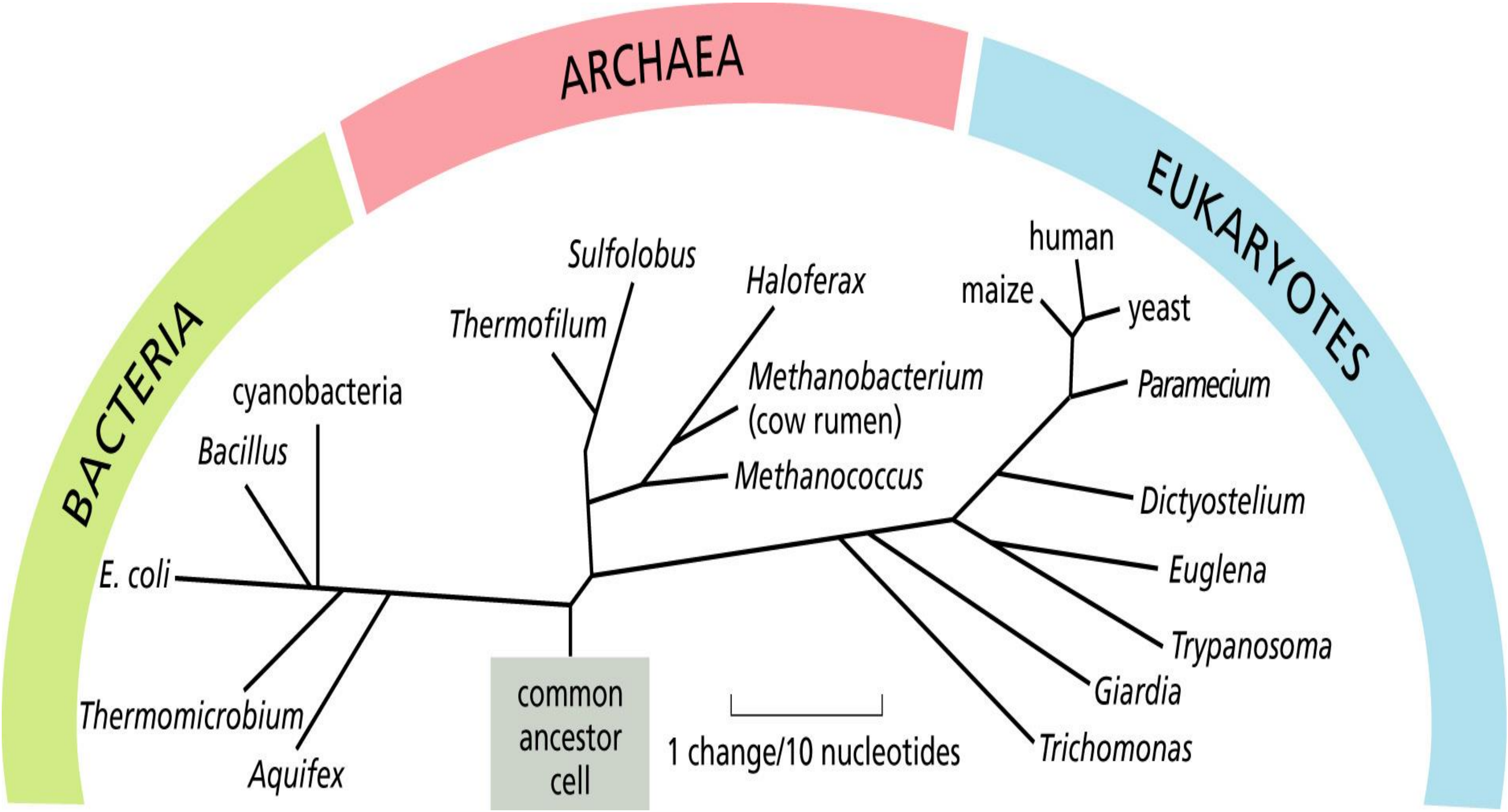
Paul Smaglik is editor of *Naturejobs*.

Bioinformatics careers can be divided into two paths: developing software, and using it. The field, catalysed by the rapid accumulation of genomic d
attention as a salvation for jobs in biology. But that sentiment may not provide an accurate assessment of job opportunities, at least for career prosp
For example, InforMax, one of the largest bioinformatics companies in the United States, generally doesn't hire biologists-turned-programmers, says
chairman and chief executive officer of the company, based in North Bethesda, Maryland.

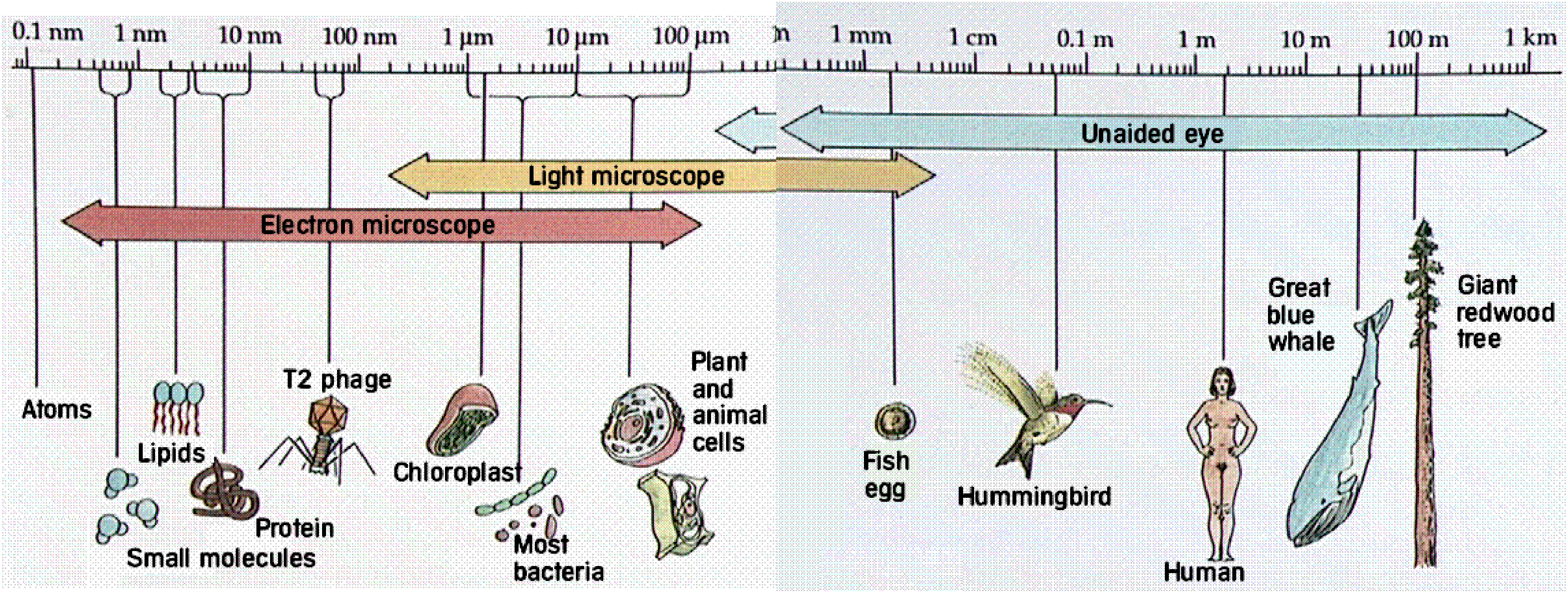
InforMax has about 95 programmers, almost all of whom come from a maths, physics or computer-science background. Titomirov says it is "much
people with those skills about biology than to teach biologists how to code well. However, as the company turns to developing software to handle fi
and protein data, it may draw on more biologists to help design new software modules.

It is "much easier" to teach people with those skills about biology than to teach biologists how to code well.

Tree of Life


ARCHAEA
BACTERIA
EUKARYOTES
Sulfolobus
Thermofilum
Haloferax
Methanobacterium (cow rumen)
Methanococcus
human
maize
yeast
Paramecium
Dictyostelium
Euglena
Trypanosoma
Giardia
Trichomonas
cyanobacteria
Bacillus
E. coli
Thermomicrobium
Aquifex
common ancestor cell
1 change/10 nucleotides

10^{-9} 10^{-6} 10^{-3}


0.1 nm
1 nm
10 nm
100 nm
1 µm
10 µm
100 µm
1 mm
1 cm
0.1 m
1 m
10 m
100 m
1 km
Unaided eye
Light microscope
Electron microscope
Atoms
Lipids
Small molecules
Protein
T2 phage
Chloroplast
Most bacteria
Plant and animal cells
Fish egg
Hummingbird
Human
Great blue whale
Giant redwood tree

Organisms: three kingdoms -- eukaryotes, eubacteria, and archea

Cell: the basic unit of life

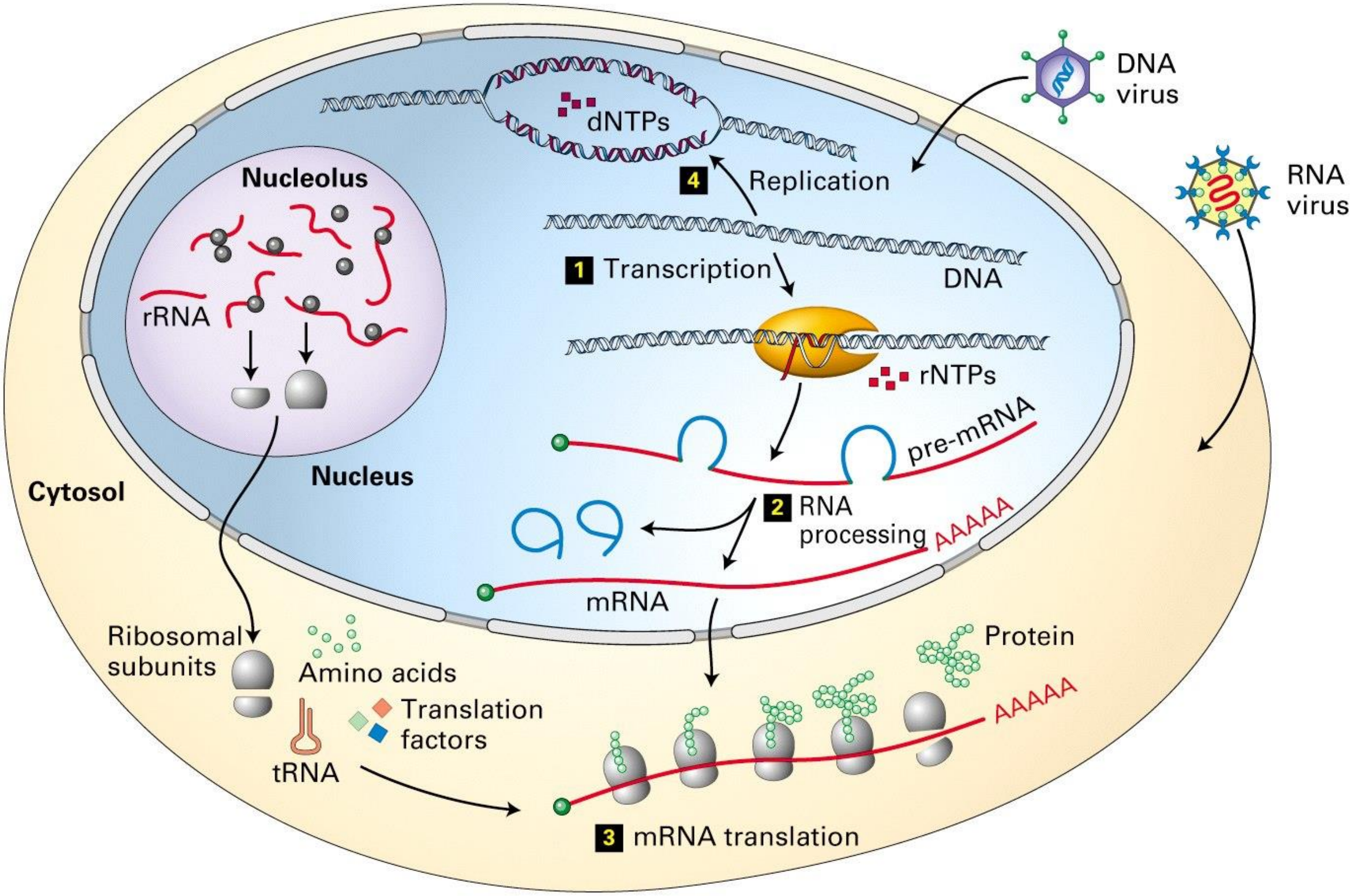
Chromosome (DNA)

> circular, also called plasmid when small (for bacteria)
> linear (for eukaryotes)

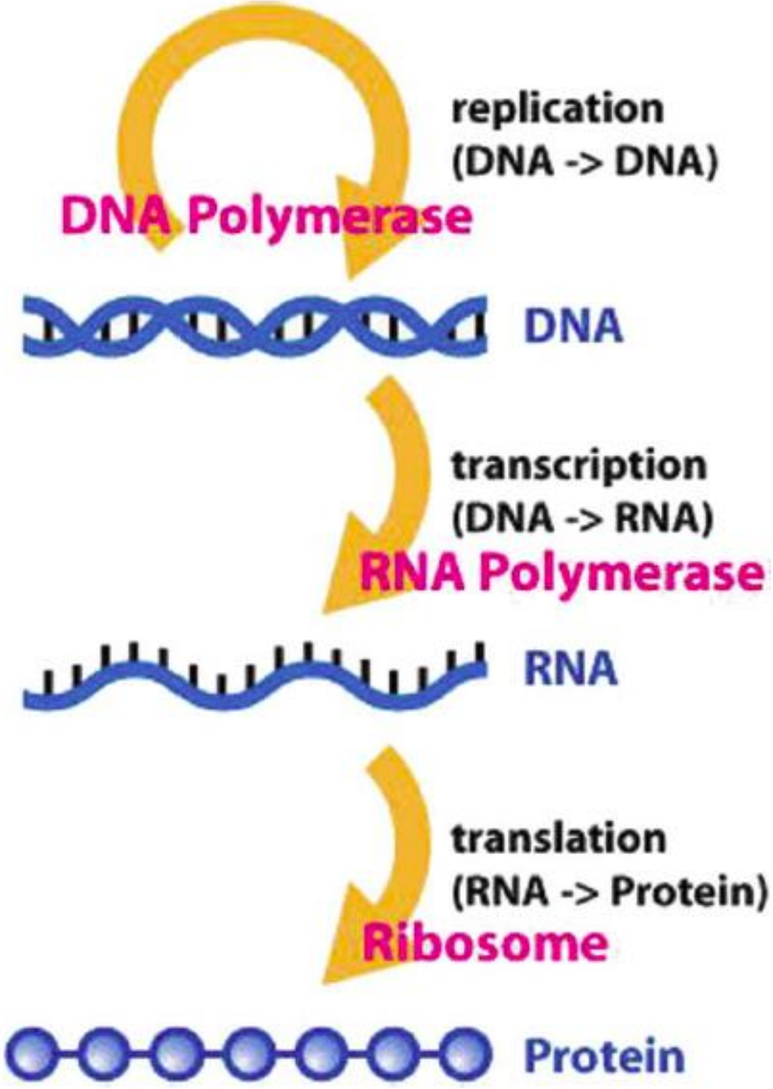
Genes: segments on DNA that contain the instructions for organism's structure and function

Proteins: the workhorse for the cell.
> establishment and maintenance of structure
> transport. e.g., hemoglobin, and integral transmembrane proteins
> protection and defense. e.g., immunoglobin G
> Control and regulation. e.g., receptors, and DNA binding proteins
> Catalysis. e.g., enzymes

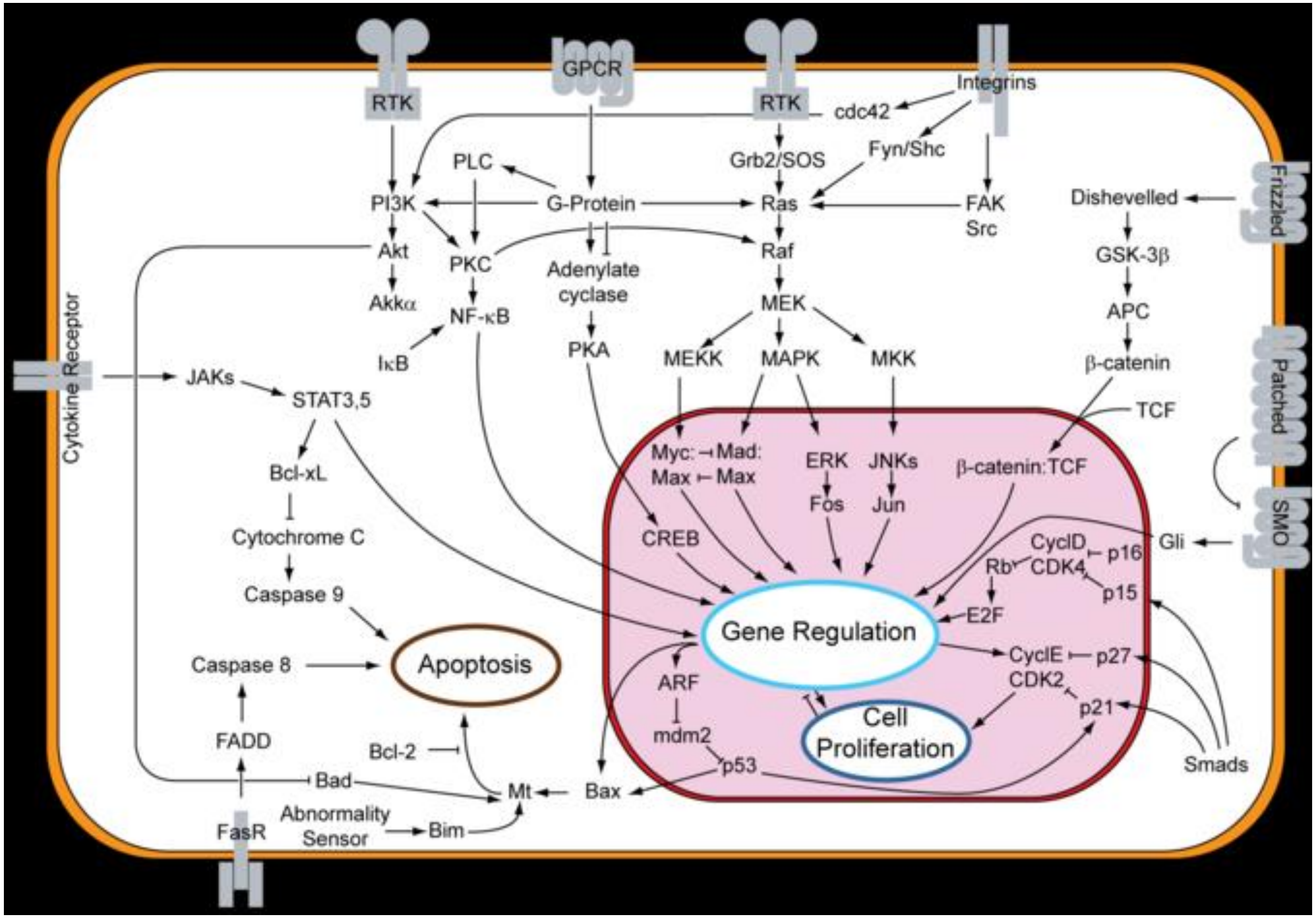
Bioinformatics in a … cell



Central Dogma of Molecular Biology
Genetic Information Flow:



Cellular processes are programmed, regulated, and yet are stochastic



Small molecules:

> sugar: carbohydrate

> fatty acids

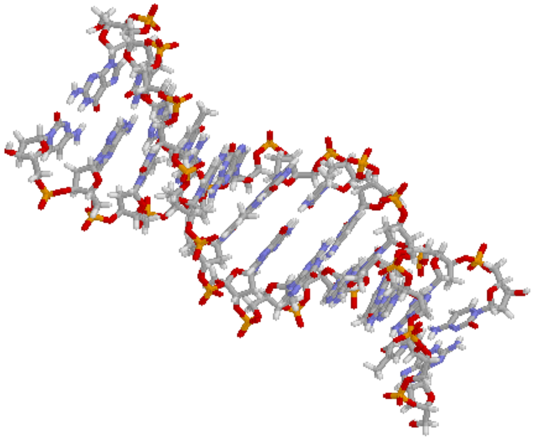
> nucleotides: A, C, G, T --> DNA (double helix, hydrogen bond, complementary bases A-T, G-C)

four bases: adenine, cytosine, guanine, and thymidine (uracil)

5' end phosphate group

3' end is free

1' position is attached with the base



double strand DNA sequences form a helix via h bonds between complementary bases

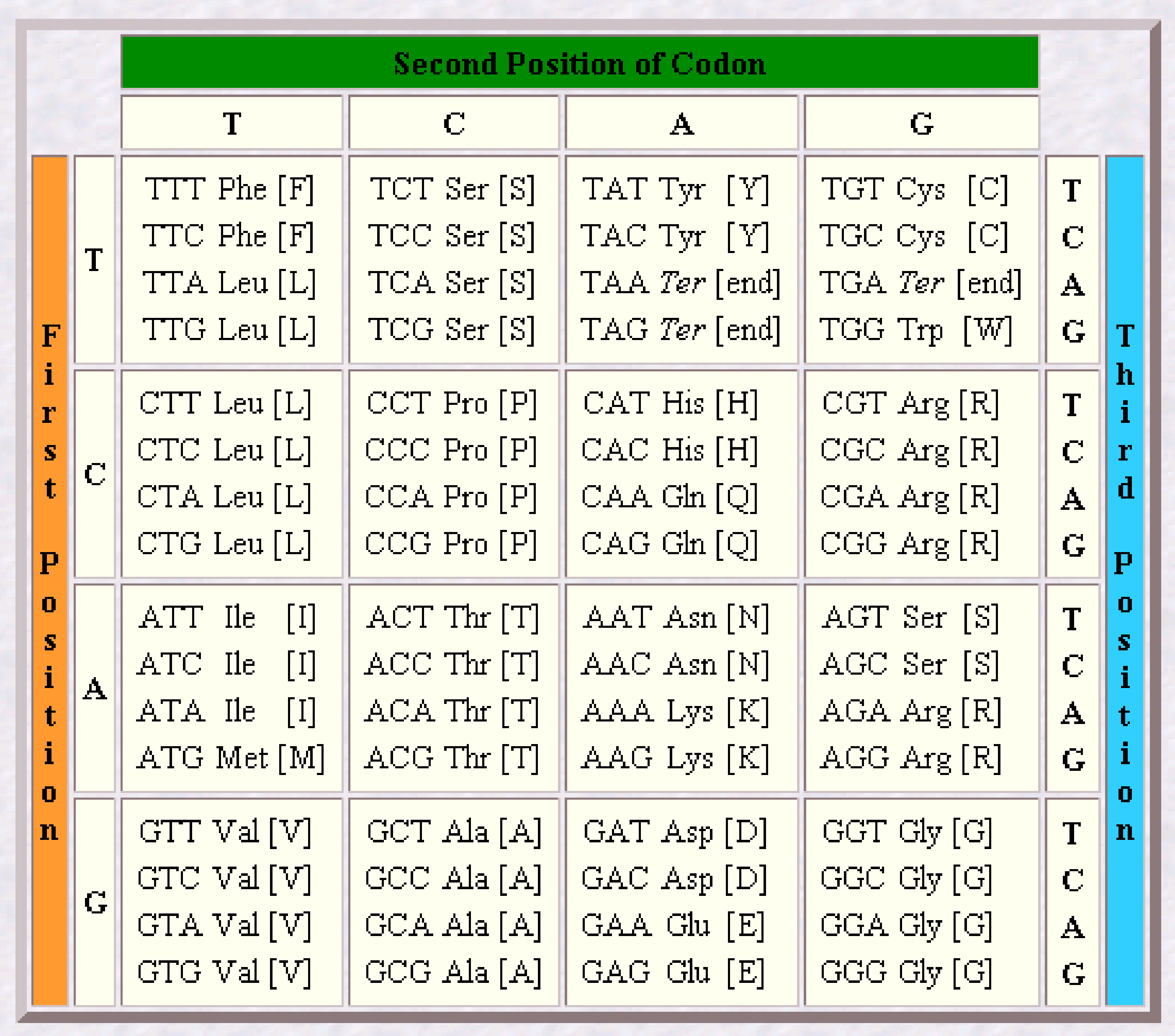
hydrogen bond:

- weak: about 3~5 kJ/mol (A covalent C-C bond has 380 kJ/mol), will break when heated
- saturation:
- specific:

Hack the life…

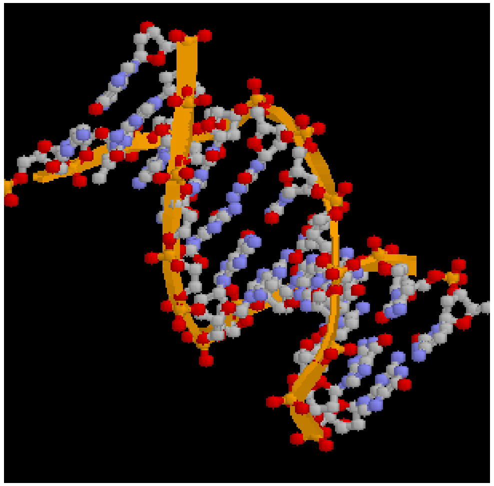
atcgggctatcgatagctatagcgcgatatatcgcgcgtatatgcgcgcatattag
tagctagtgctgattcatctggactgtcgtaatatatacgcgcccggctatcgcgct
atgcgcgatatcgcgcggcgctatataaatattaaaaaataaaatatatatatatgc
tgcgcgatagcgctataggcgcgctatccatatataggcgctcgcccgggcgcga
tgcatcggctacggctagctgtagctagtcggcgattagcggcttatatgcggcga
gcgatgagagtcgcggctataggcttaggctatagcgctagtatatagcggctagc
cgcgtagacgcgatagcgtagctagcggcgcgcgtatatagcgcttaagagcca
aaatgcgtctagcgctataatatgcgctatagctatatgcggctattatatagcgca
gcgctagctagcgtatcaggcgaggagatcgatgctactgatcgatgctagagca
gcgtcatgctagtagtgccatatatatgctgagcgcgcgtagctcgatattacgcta
cctagatgctagcgagctatgatcgtagca..

Genetic Code: codons

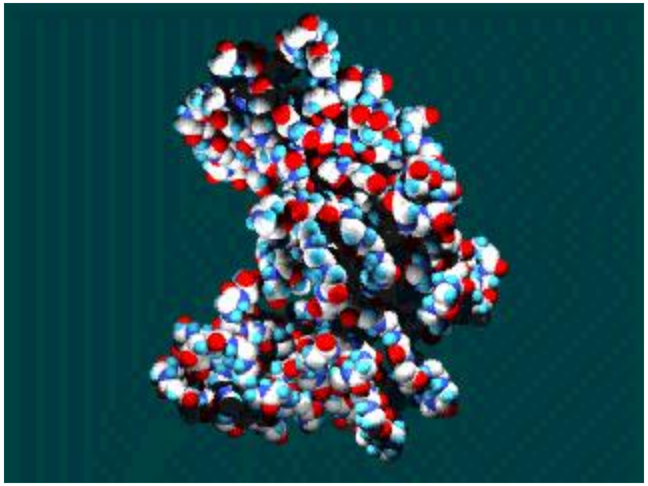


First Position		Second Position of Codon				Third Position
		T	**C**	**A**	**G**	
	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	**T**
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	**C**
		TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	**A**
		TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	**G**
	C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	**T**
		CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	**C**
		CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	**A**
		CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	**G**
	A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	**T**
		ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	**C**
		ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	**A**
		ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	**G**
	G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	**T**
		GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	**C**
		GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	**A**
		GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	**G**

Information Expression



1-D information array

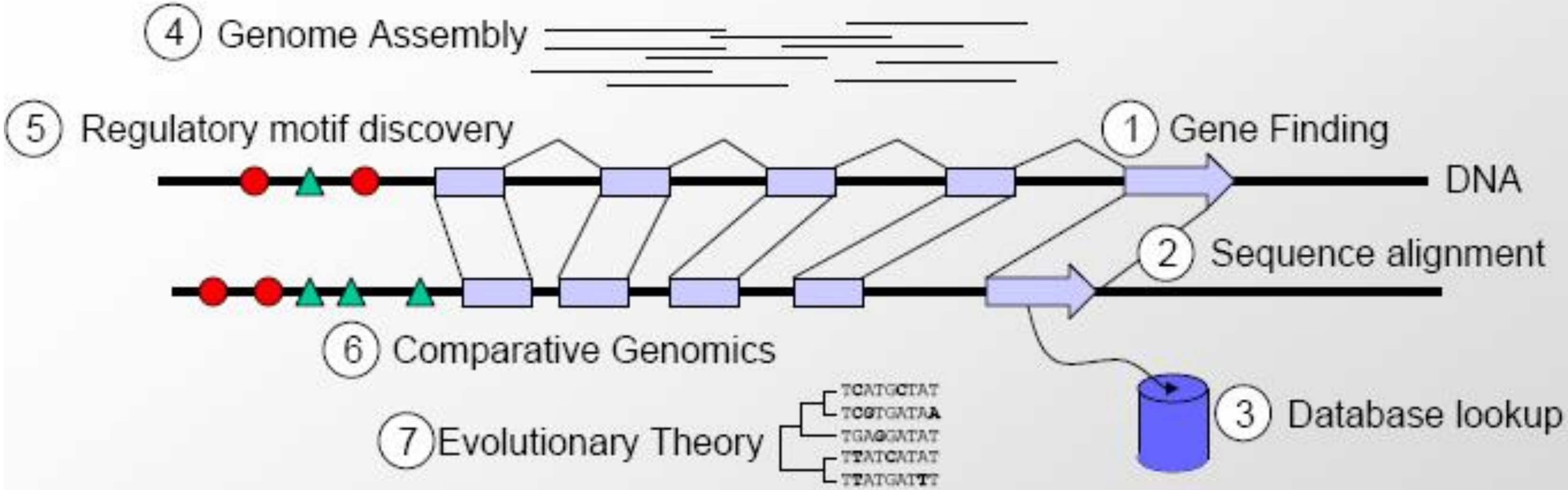


3-D biochemical structure

Challenges in Life Sciences

- Understanding correlation between genotype and phenotype
- Predicting genotype <=> phenotype
- Phenotypes:
  - drug/therapy response
  - drug-drug interactions for expression
  - drug mechanism
  - interacting pathways of metabolism

Challenges in Computational Biology



④ Genome Assembly

⑤ Regulatory motif discovery

① Gene Finding

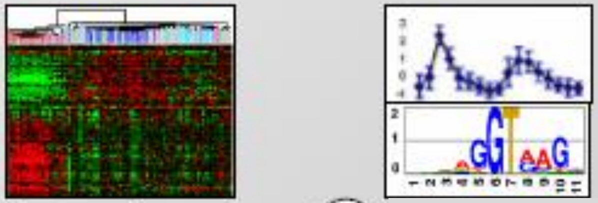
② Sequence alignment

⑥ Comparative Genomics

⑦ Evolutionary Theory

③ Database lookup

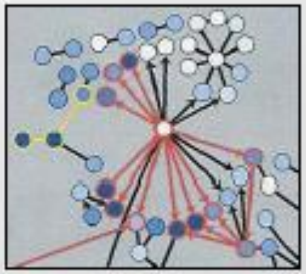
⑧ Gene expression analysis



⑨ Cluster discovery

⑩ Profiling

⑪ Protein network analysis



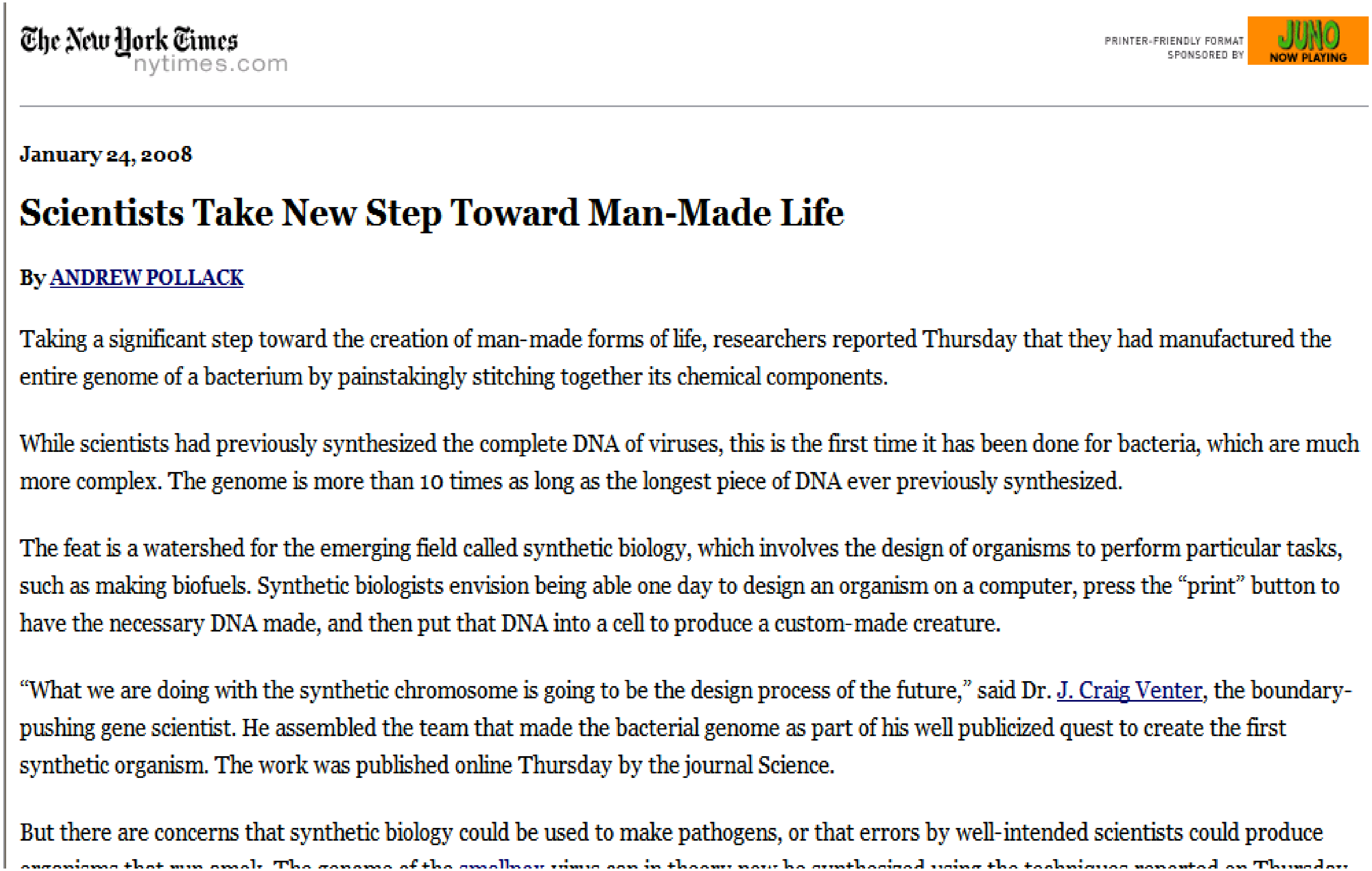
⑫ Regulatory network inference

⑬ Emerging network properties

⑭ Structure Prediction

Image removed
due to
copyright restrictions.

Credit:Kellis & Indyk



The New York Times
nytimes.com

January 24, 2008

Scientists Take New Step Toward Man-Made Life

By ANDREW POLLACK

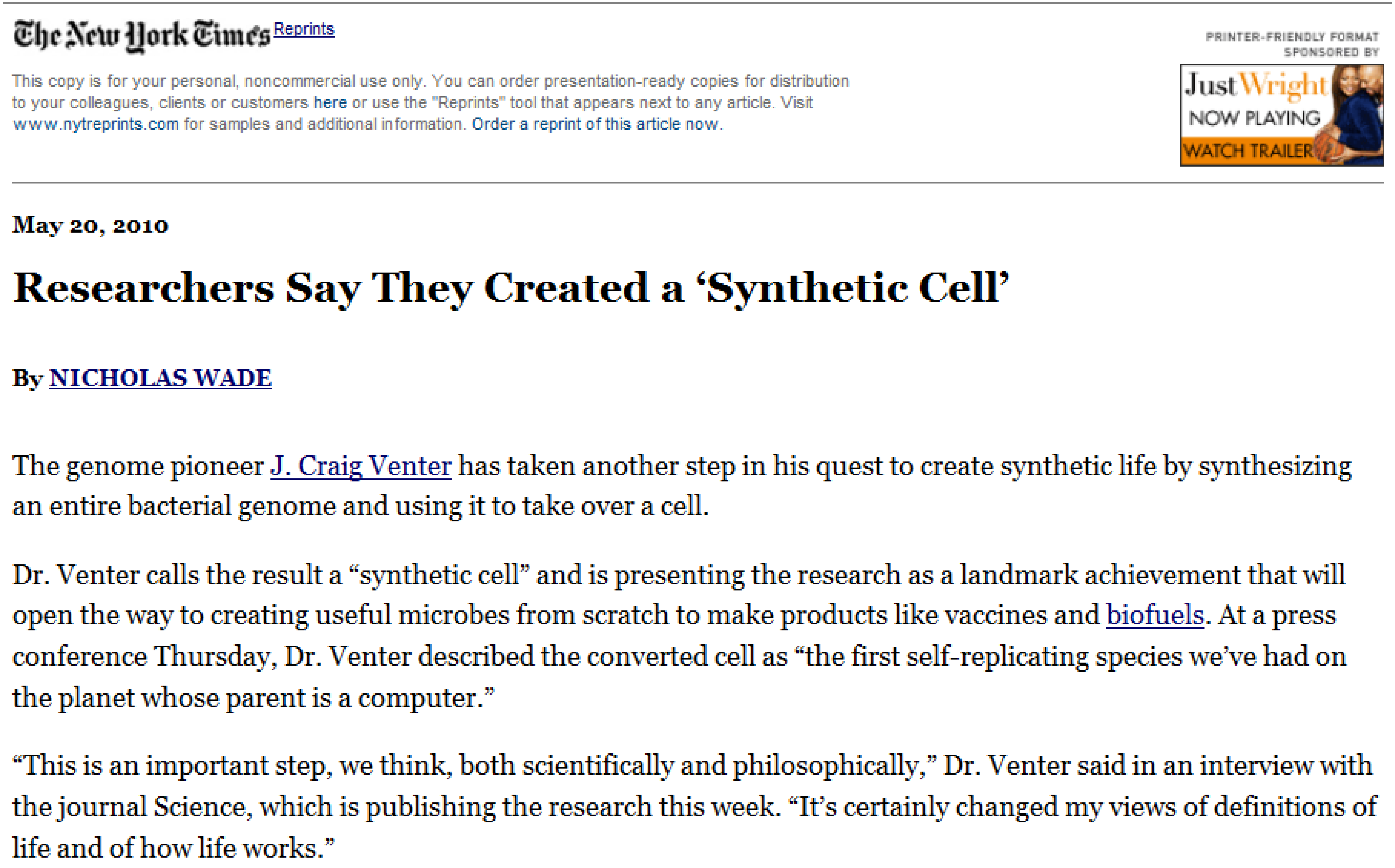
Taking a significant step toward the creation of man-made forms of life, researchers reported Thursday that they had manufactured the entire genome of a bacterium by painstakingly stitching together its chemical components.

While scientists had previously synthesized the complete DNA of viruses, this is the first time it has been done for bacteria, which are much more complex. The genome is more than 10 times as long as the longest piece of DNA ever previously synthesized.

The feat is a watershed for the emerging field called synthetic biology, which involves the design of organisms to perform particular tasks, such as making biofuels. Synthetic biologists envision being able one day to design an organism on a computer, press the “print” button to have the necessary DNA made, and then put that DNA into a cell to produce a custom-made creature.

“What we are doing with the synthetic chromosome is going to be the design process of the future,” said Dr. J. Craig Venter, the boundary-pushing gene scientist. He assembled the team that made the bacterial genome as part of his well publicized quest to create the first synthetic organism. The work was published online Thursday by the journal Science.

But there are concerns that synthetic biology could be used to make pathogens, or that errors by well-intended scientists could produce organisms that run amok. The genome of the smallpox virus can in theory now be synthesized using the techniques reported on Thursday



The New York Times Reprints

May 20, 2010

Researchers Say They Created a ‘Synthetic Cell’

By NICHOLAS WADE

The genome pioneer J. Craig Venter has taken another step in his quest to create synthetic life by synthesizing an entire bacterial genome and using it to take over a cell.

Dr. Venter calls the result a “synthetic cell” and is presenting the research as a landmark achievement that will open the way to creating useful microbes from scratch to make products like vaccines and biofuels. At a press conference Thursday, Dr. Venter described the converted cell as “the first self-replicating species we’ve had on the planet whose parent is a computer.”

“This is an important step, we think, both scientifically and philosophically,” Dr. Venter said in an interview with the journal Science, which is publishing the research this week. “It’s certainly changed my views of definitions of life and of how life works.”

Topics

- Mapping and assembly
- Sequence analysis (Similarity -> Homology):
  - Pairwise alignment (database searching)
  - Multiple sequence alignment, profiling
  - Gene prediction
  - Pattern (Motif) discovery and recognition
- Phylogenetics analysis
  - Character based
  - Distance based
- Structure prediction
  - RNA Secondary
  - Protein Secondary & tertiary
- Network analysis:
  - Gene expression
  - Regulatory networks

How much should I know about biology?

- Apparently, the more the better
  - The least, Pavzner's 3-page "All you need to know about Molecular biology".
  - Chapters 1 & 2 of the text.

- We adopt an "object-oriented" scheme, namely, we will transform biological problems into abstract computing problems and hide unnecessary details.

So another big goal of this course is learn how to do abstraction.

Goals?

At the end of this course, you should be able to

- Describe the main computational challenges in molecular biology.
- Implement and use basic algorithms.
- Describe several advanced algorithms.
  - ☞ Sequence alignment using dynamics programming
  - ☞ Hidden Markov models
  - ☞ Hierarchical clustering
  - ☞ K-means
  - ☞ Gradient descent optimization

- Know the existing resources: Databases, Software, …