CISC 436/636 Computational Biology &Bioinformatics (Fall 2016)

Lecture 1

Course Overview

Li Liao

Computer and Information Sciences

University of Delaware

Administrative stuff

- Webpage: **http://www.cis.udel.edu/~lliao/cis636f16**
- Syllabus and tentative schedule (check frequently for update)
- Office hours: 12:45PM-1:45PM Tuesdays and Thursdays.
  - Appointments
- Collect info (name, email, major, programming language)
- Introduce textbook and other resources
  - URLs, PDF/PS files, or hardcopy handout
  - A reading list
- Workload
  - 4 homework assignments (hands-on to learn the nuts and bolts)
    - Language issue: Perl is strongly recommended (A tutorial is provided)
  - Mid-term and final exams
- Late policy: 10% off per class and no later than one week.

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
- 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
- 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) ~ 2500 members
- Severe shortage for qualified bioinformatians

Conferences

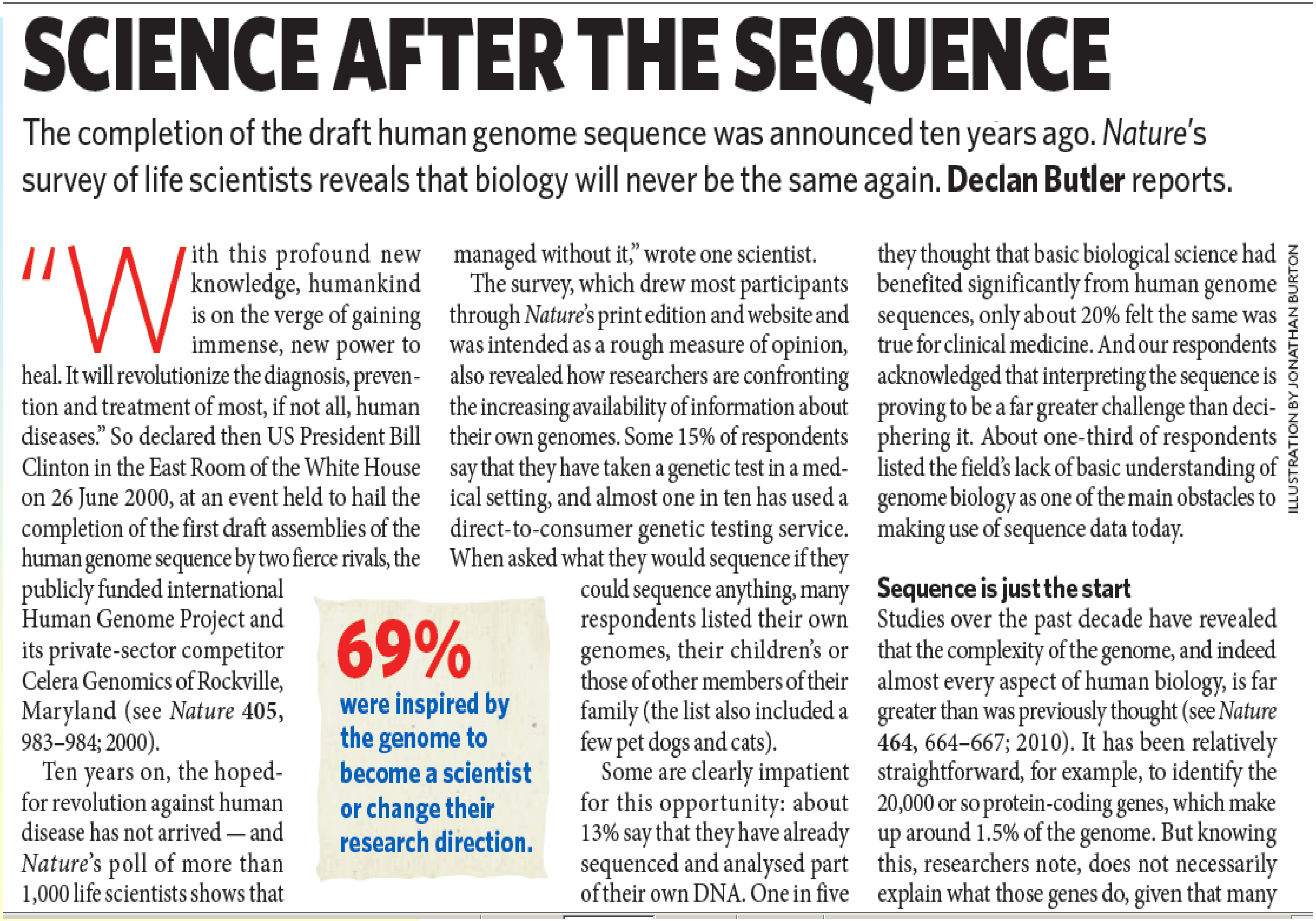
- ISMB (Intelligent Systems for Molecular Biology) started in 1992
- RECOMB (International Conference on Computational Molecular Biology) started in 1997
- PSB (Pacific Symposium on Biocomputing) started 1996
- TIGR Computational genomic, started in 1997
- …

Journals

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

A short history: 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).

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).

69%

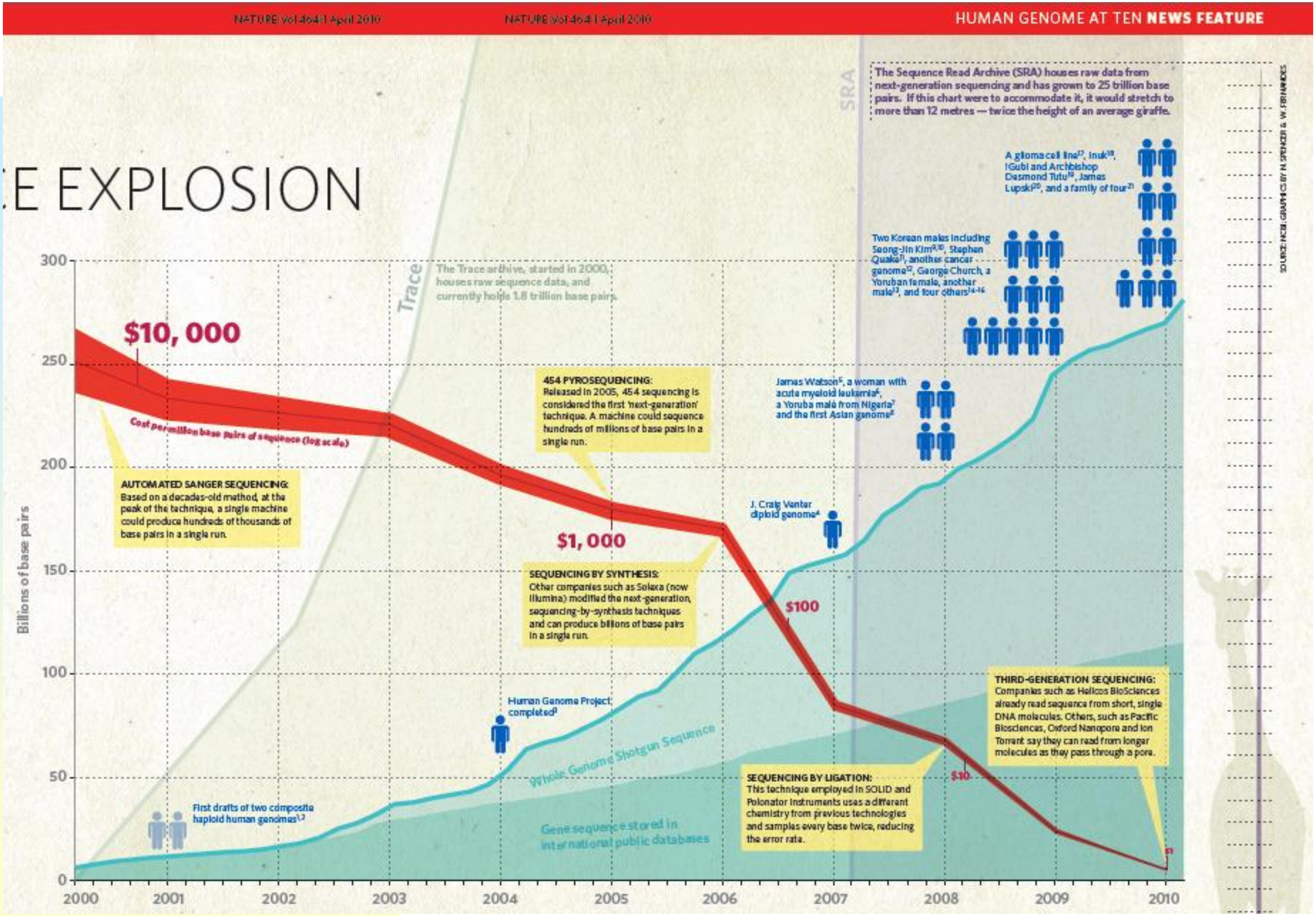
were inspired by the genome to become a scientist or change their research direction.

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



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

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.

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.

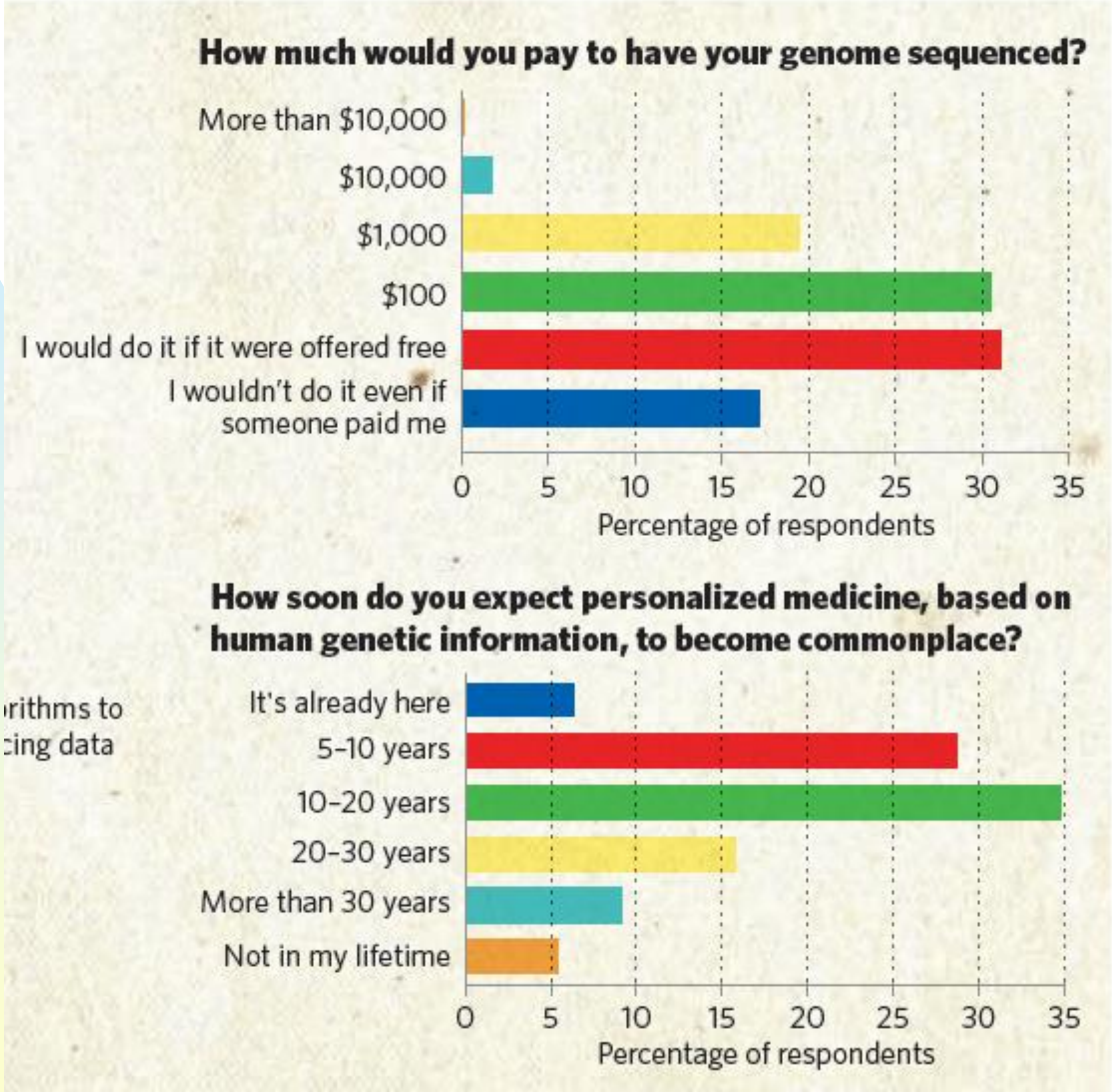
454 PYROSEQUENCING: Released in 2005, 454 sequencing is considered the first 'next-generation' technique. A machine could sequence hundreds of millions of base pairs in a single run.

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.

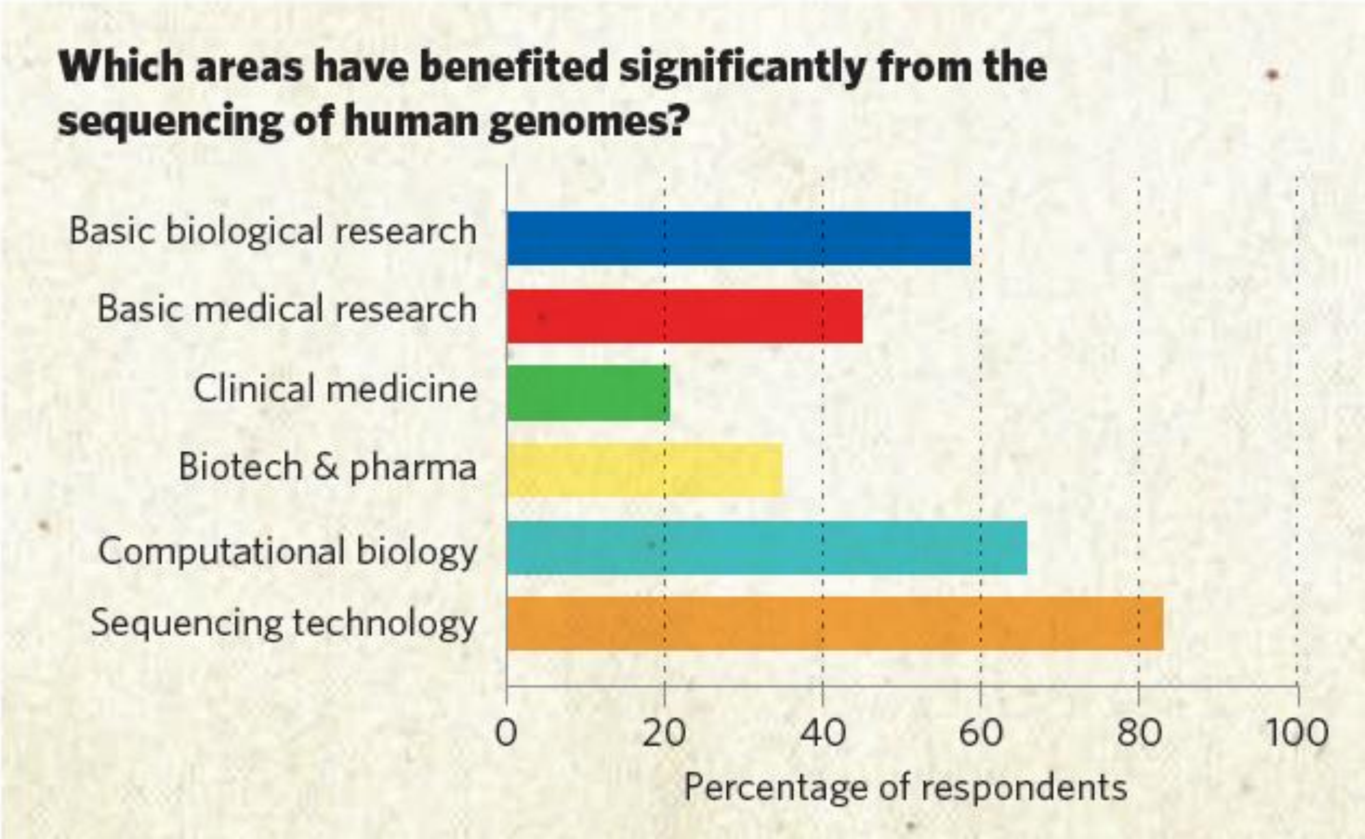
Human Genome at Ten:



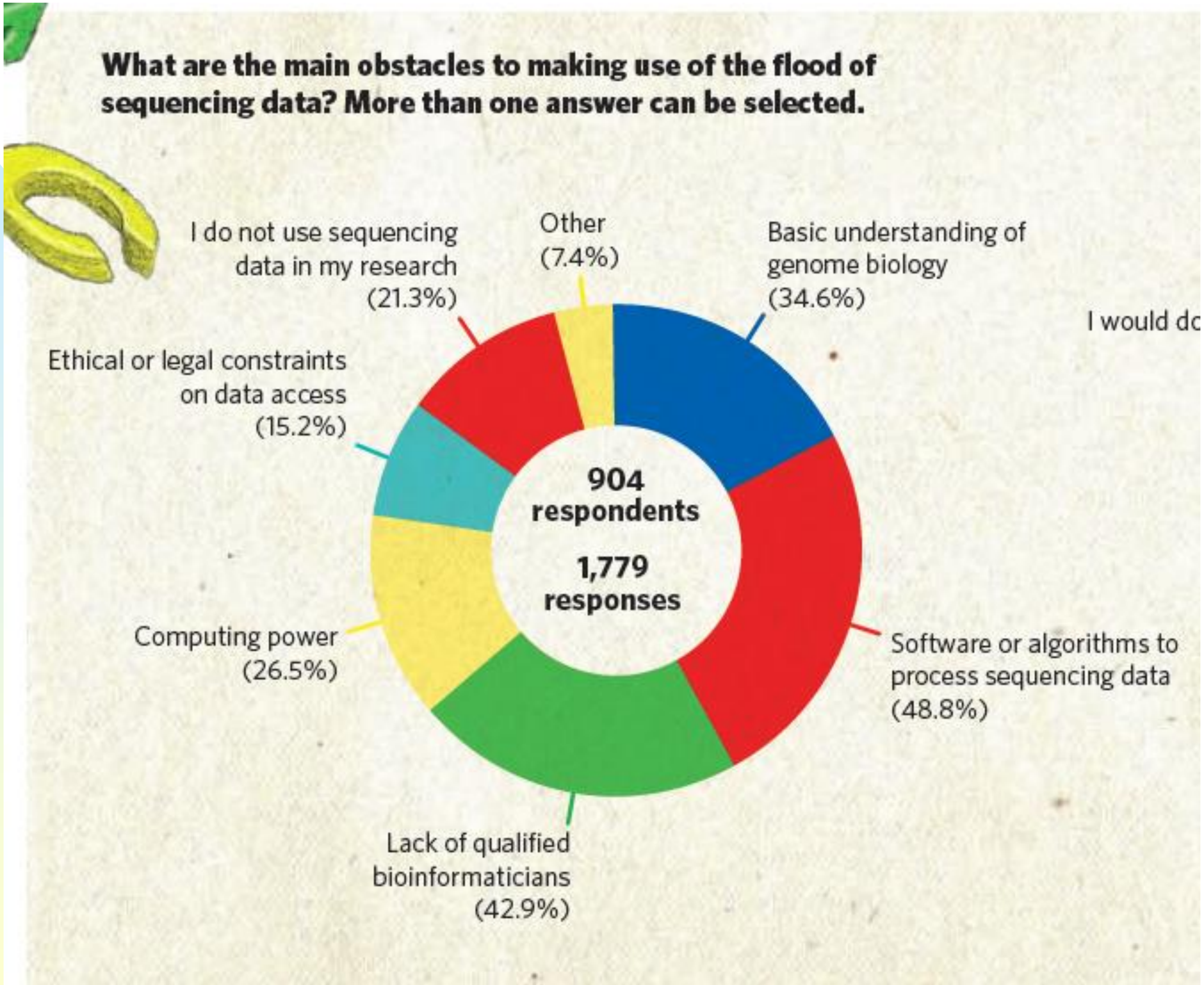
How much would you pay to have your genome sequenced?

How soon do you expect personalized medicine, based on human genetic information, to become commonplace?

Human Genome at Ten:



Human Genome at Ten:



What are the main obstacles to making use of the flood of sequencing data? More than one answer can be selected.

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 to
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

New Sites

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

Gunar
directo
bioinfo

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.

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

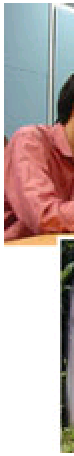
ROBERT TRIENDL

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



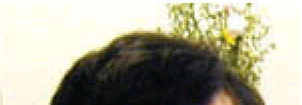
Resea
Biolog
joining
bioinfo
toward

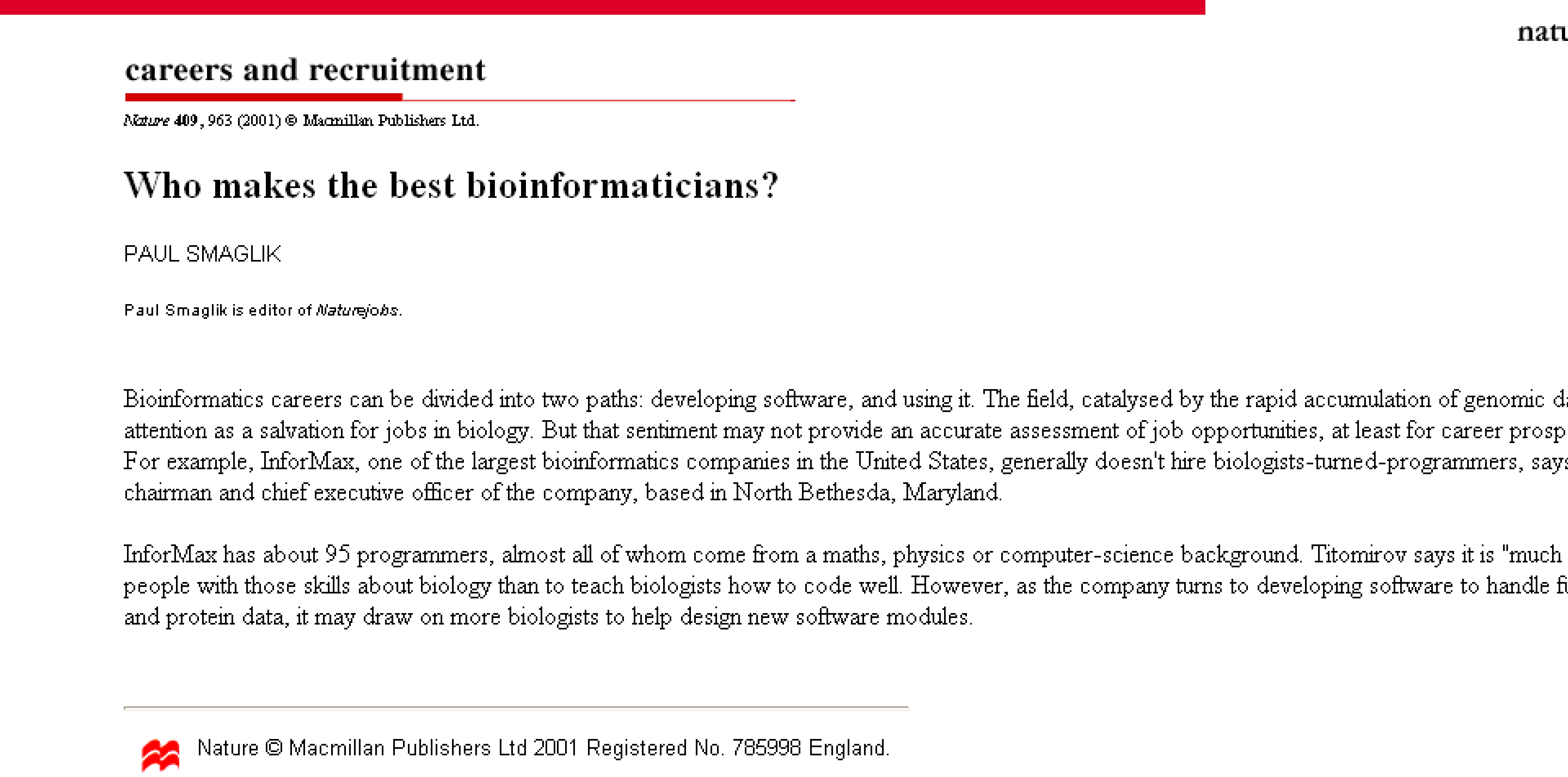
This year, Japan's Ministry of Education, Science, Sports and Culture started to upgrade bioinformatics education at national universities by creating additional staff positions and funding both undergraduate courses and graduate-level informatics training.

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.





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.

Your coffee habit may be gene

By Jacqueline Howard, CNN

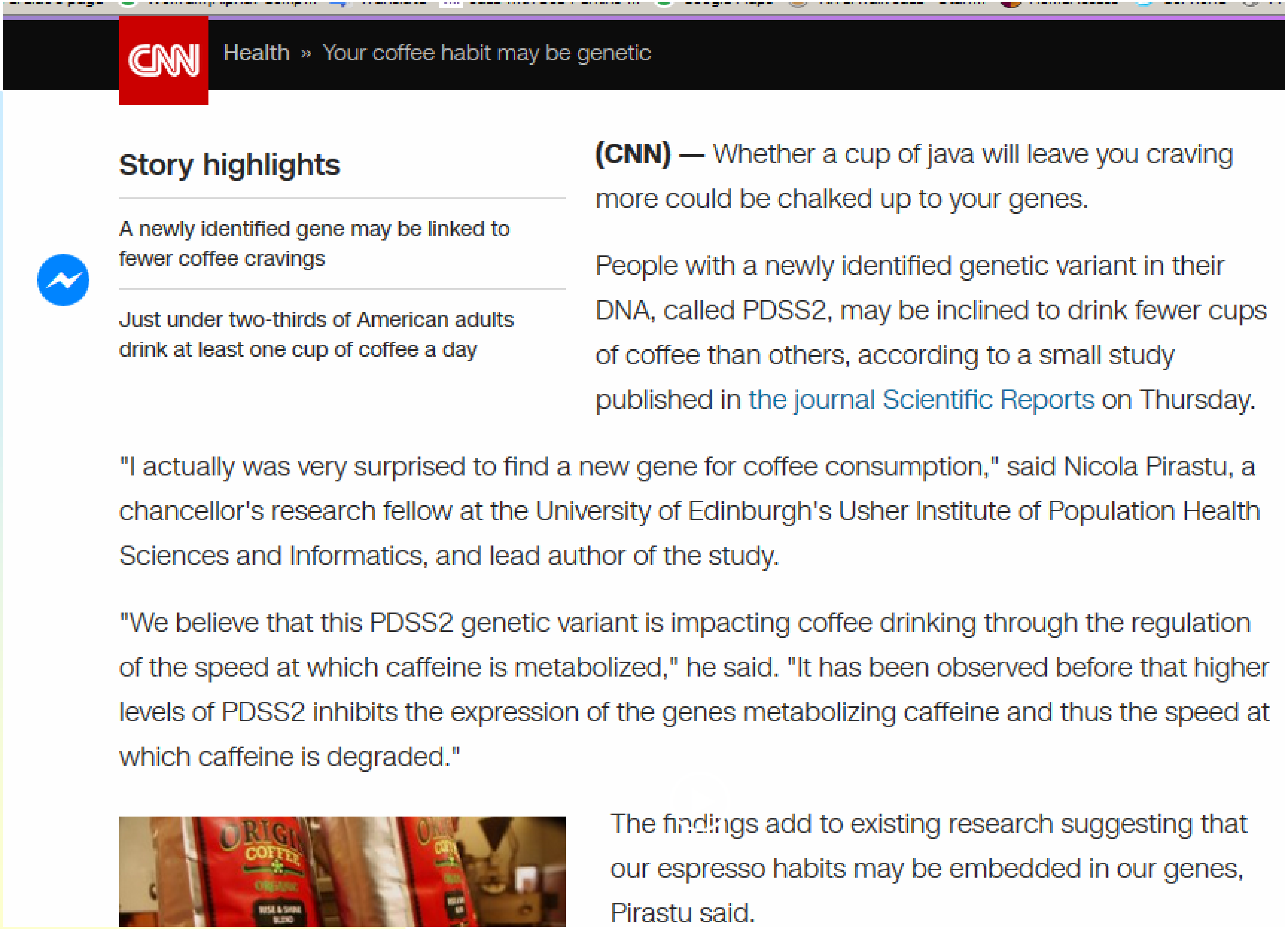
Updated 6:37 AM ET, Fri August 26, 2016

KEVORK DJANSEZIAN/GETTY IMAGES

Top

Photos: Coffee's health history

2015 headline: Coffee is practically a health food – How about coffee's effects on your overall risk of



Story highlights

A newly identified gene may be linked to fewer coffee cravings

Just under two-thirds of American adults drink at least one cup of coffee a day

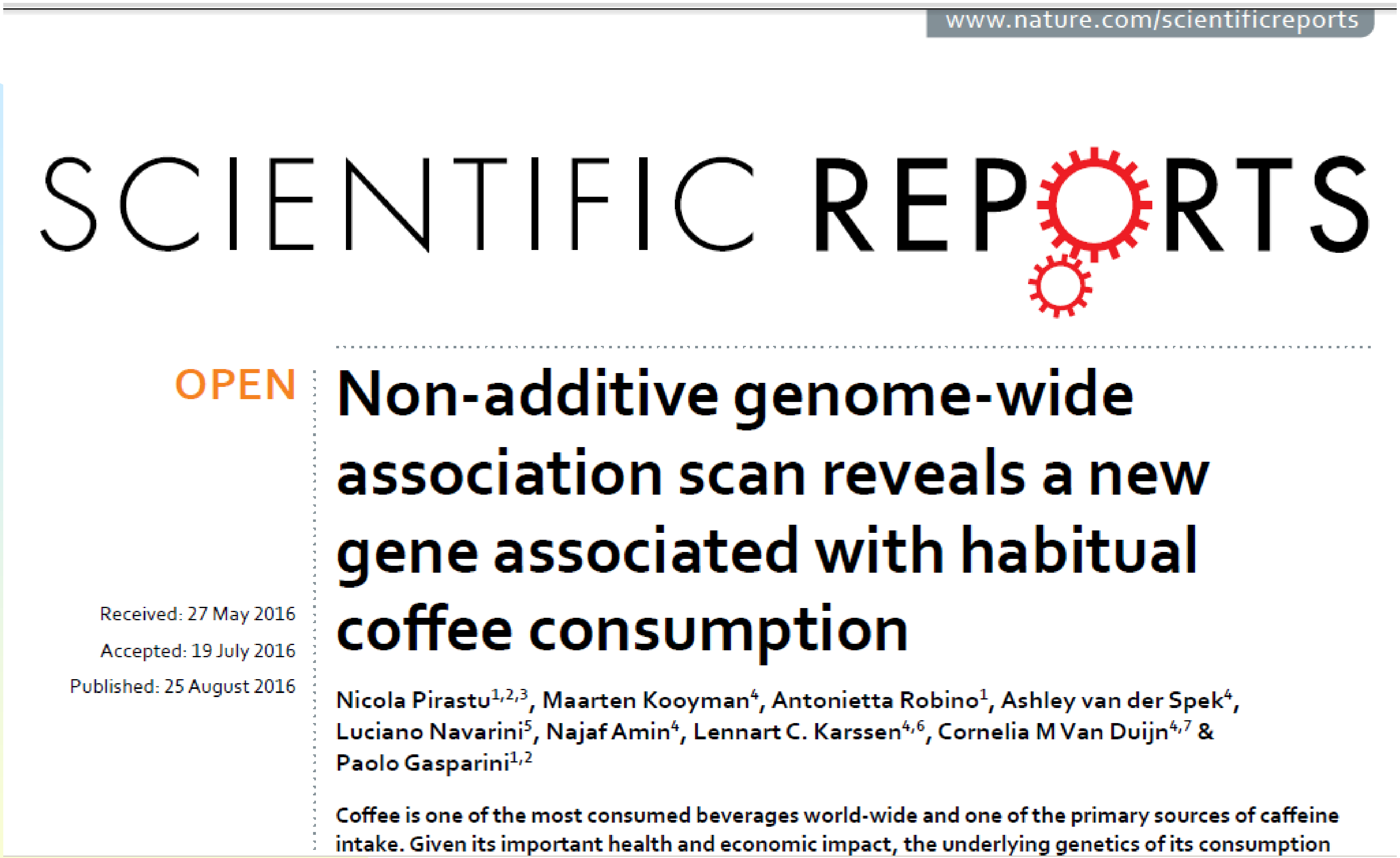
(CNN) — Whether a cup of java will leave you craving more could be chalked up to your genes.

People with a newly identified genetic variant in their DNA, called PDSS2, may be inclined to drink fewer cups of coffee than others, according to a small study published in the journal Scientific Reports on Thursday.

"I actually was very surprised to find a new gene for coffee consumption," said Nicola Pirastu, a chancellor's research fellow at the University of Edinburgh's Usher Institute of Population Health Sciences and Informatics, and lead author of the study.

"We believe that this PDSS2 genetic variant is impacting coffee drinking through the regulation of the speed at which caffeine is metabolized," he said. "It has been observed before that higher levels of PDSS2 inhibits the expression of the genes metabolizing caffeine and thus the speed at which caffeine is degraded."

The findings add to existing research suggesting that our espresso habits may be embedded in our genes, Pirastu said.



www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Non-additive genome-wide association scan reveals a new gene associated with habitual coffee consumption

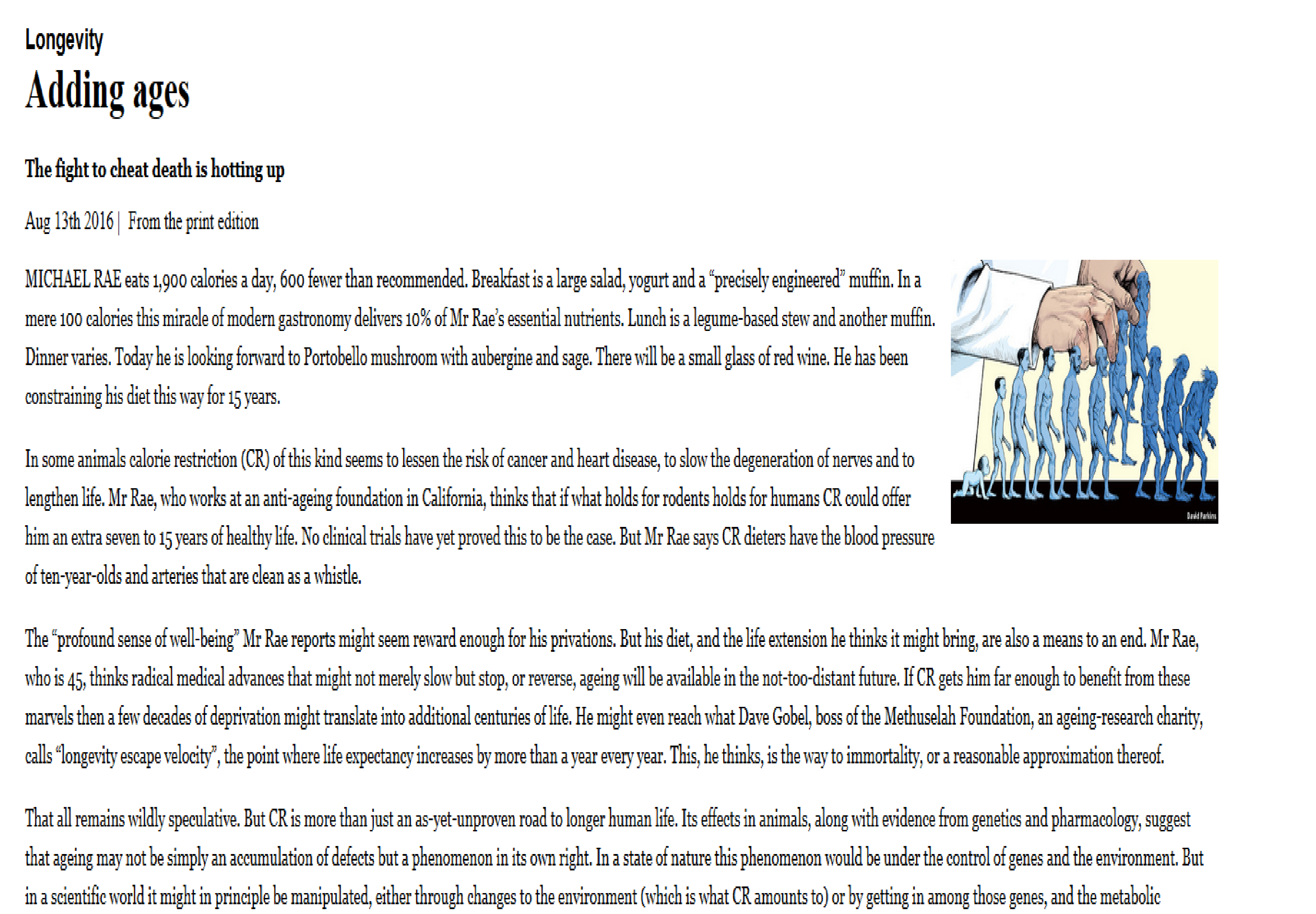
Received: 27 May 2016

Accepted: 19 July 2016

Published: 25 August 2016

Nicola Pirastu[1,2,3], Maarten Kooyman[4], Antonietta Robino[1], Ashley van der Spek[4], Luciano Navarini[5], Najaf Amin[4], Lennart C. Karssen[4,6], Cornelia M Van Duijn[4,7] & Paolo Gasparini[1,2]

Coffee is one of the most consumed beverages world-wide and one of the primary sources of caffeine intake. Given its important health and economic impact, the underlying genetics of its consumption



Longevity

Adding ages

The fight to cheat death is hotting up

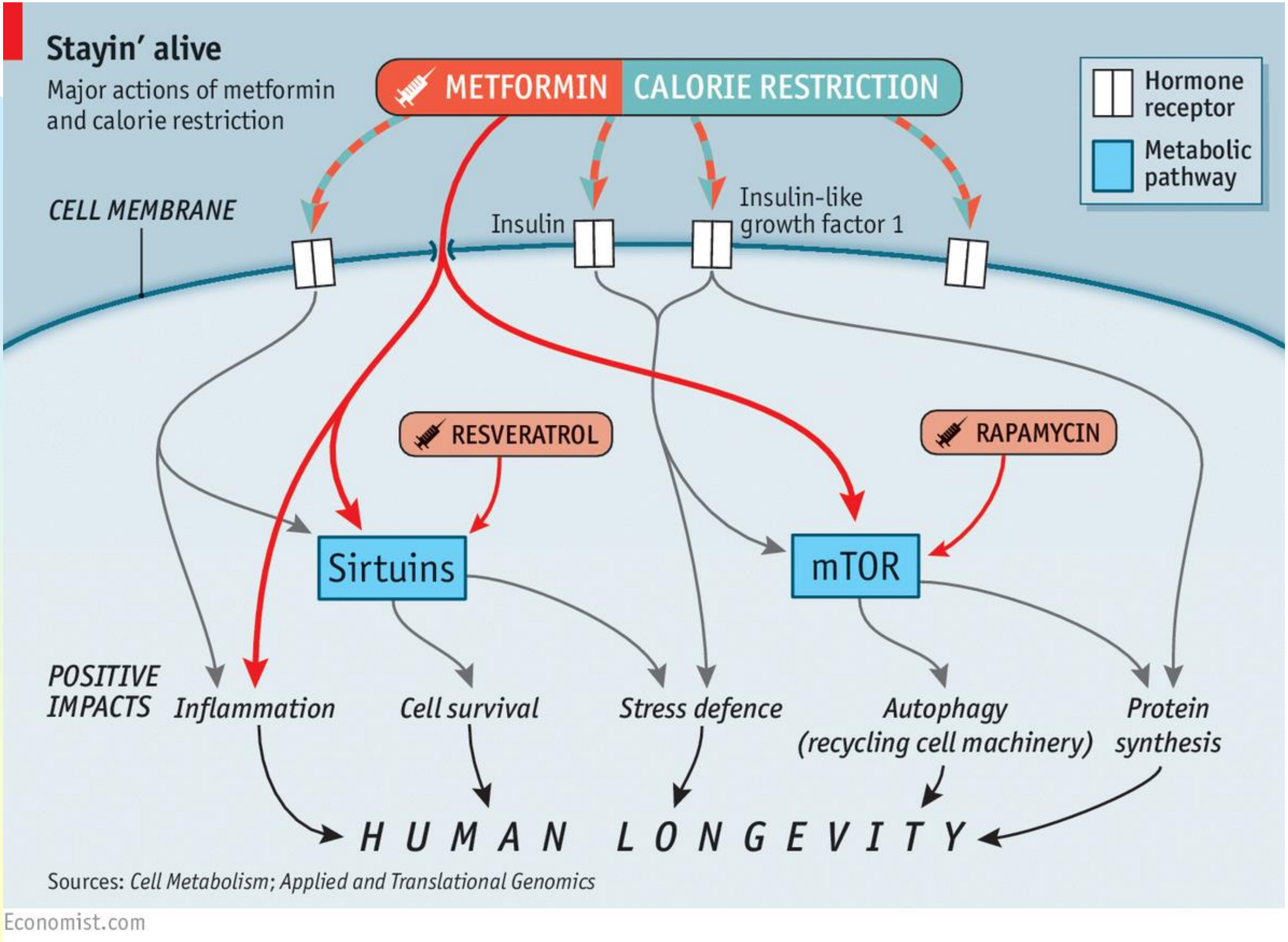
Aug 13th 2016 | From the print edition

MICHAEL RAE eats 1,900 calories a day, 600 fewer than recommended. Breakfast is a large salad, yogurt and a "precisely engineered" muffin. In a mere 100 calories this miracle of modern gastronomy delivers 10% of Mr Rae's essential nutrients. Lunch is a legume-based stew and another muffin. Dinner varies. Today he is looking forward to Portobello mushroom with aubergine and sage. There will be a small glass of red wine. He has been constraining his diet this way for 15 years.

In some animals calorie restriction (CR) of this kind seems to lessen the risk of cancer and heart disease, to slow the degeneration of nerves and to lengthen life. Mr Rae, who works at an anti-ageing foundation in California, thinks that if what holds for rodents holds for humans CR could offer him an extra seven to 15 years of healthy life. No clinical trials have yet proved this to be the case. But Mr Rae says CR dieters have the blood pressure of ten-year-olds and arteries that are clean as a whistle.

The "profound sense of well-being" Mr Rae reports might seem reward enough for his privations. But his diet, and the life extension he thinks it might bring, are also a means to an end. Mr Rae, who is 45, thinks radical medical advances that might not merely slow but stop, or reverse, ageing will be available in the not-too-distant future. If CR gets him far enough to benefit from these marvels then a few decades of deprivation might translate into additional centuries of life. He might even reach what Dave Gobel, boss of the Methuselah Foundation, an ageing-research charity, calls "longevity escape velocity", the point where life expectancy increases by more than a year every year. This, he thinks, is the way to immortality, or a reasonable approximation thereof.

That all remains wildly speculative. But CR is more than just an as-yet-unproven road to longer human life. Its effects in animals, along with evidence from genetics and pharmacology, suggest that ageing may not be simply an accumulation of defects but a phenomenon in its own right. In a state of nature this phenomenon would be under the control of genes and the environment. But in a scientific world it might in principle be manipulated, either through changes to the environment (which is what CR amounts to) or by getting in among those genes, and the metabolic


Stayin' alive
Major actions of metformin and calorie restriction
METFORMIN
CALORIE RESTRICTION
Hormone receptor
Metabolic pathway
CELL MEMBRANE
Insulin
Insulin-like growth factor 1
RESVERATROL
RAPAMYCIN
Sirtuins
mTOR
POSITIVE IMPACTS
Inflammation
Cell survival
Stress defence
Autophagy (recycling cell machinery)
Protein synthesis
HUMAN LONGEVITY
Sources: Cell Metabolism; Applied and Translational Genomics
Economist.com

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

Cell: the basic unit of life

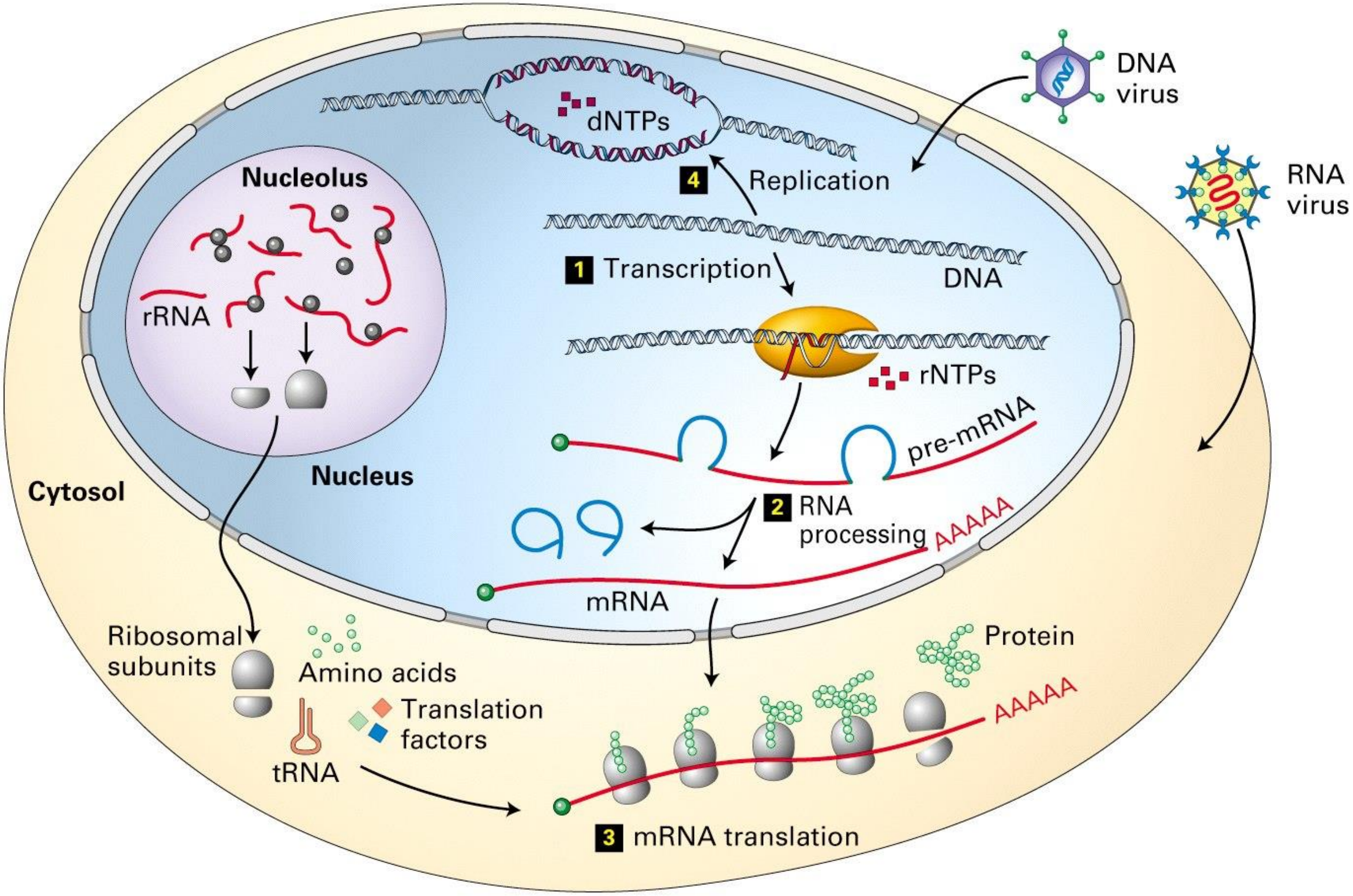
Chromosome (DNA)

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

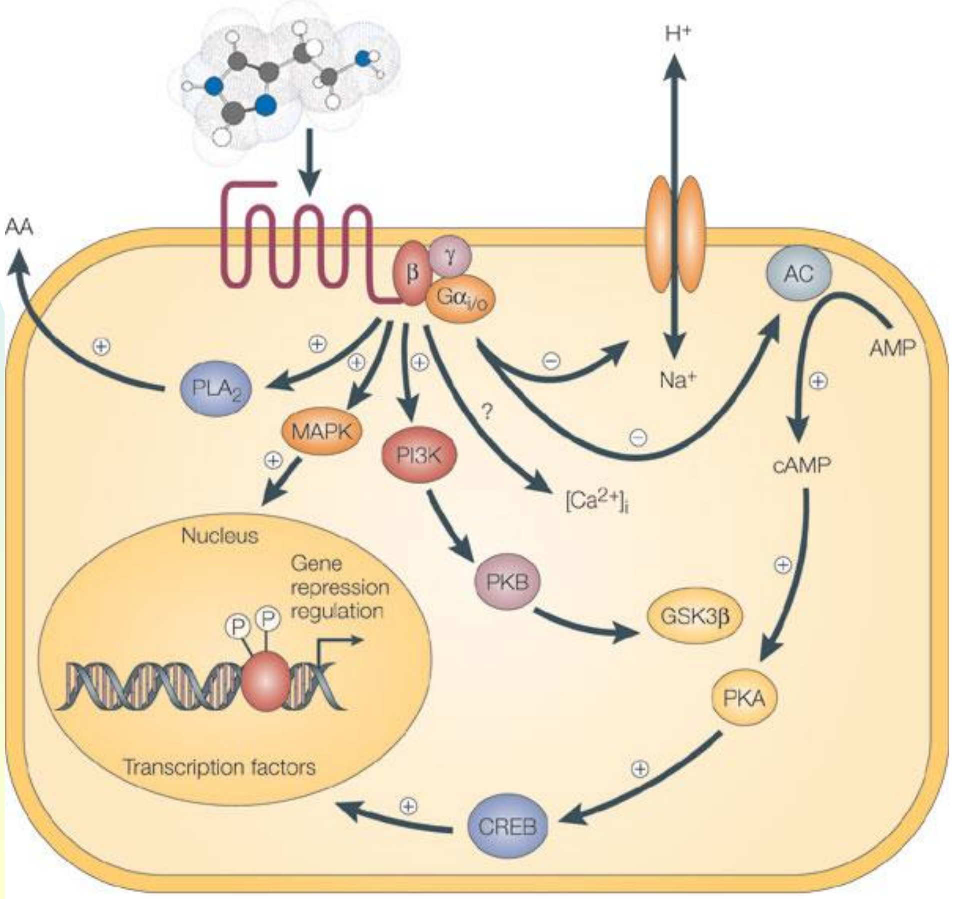
Genes: segments on DNA that contain the instructions for organism's structure and function

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



Protein-Protein Interaction plays essential roles in cellular processes



Nature Reviews | Drug Discovery

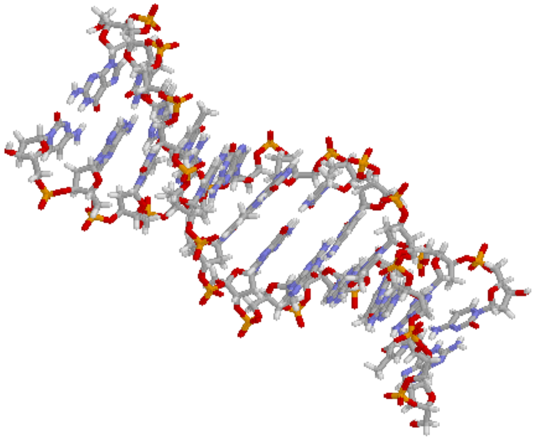
Small molecules:

> 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 hydrogen bonds between complementary bases

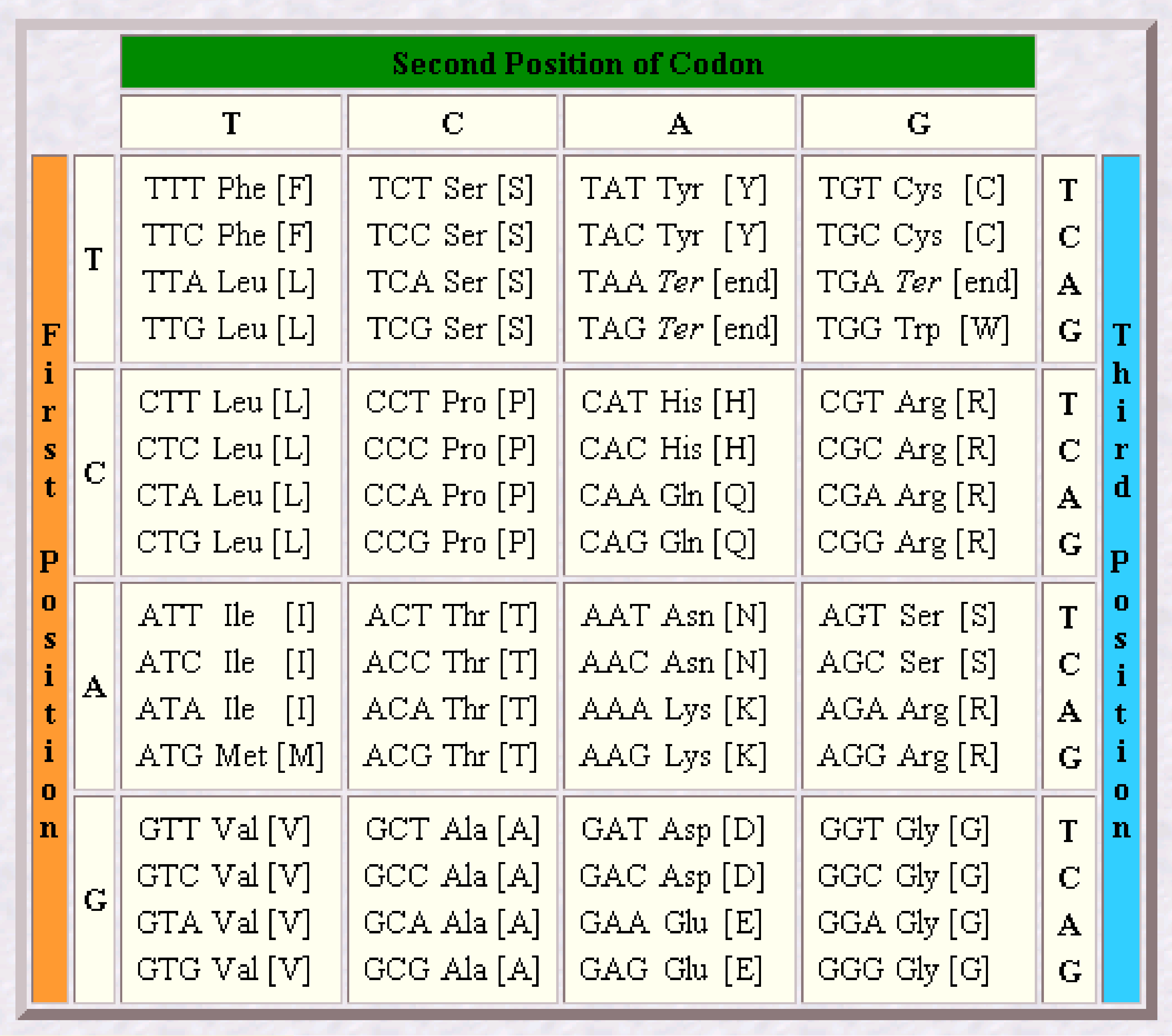
hydrogen bond:

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

Hack the life…

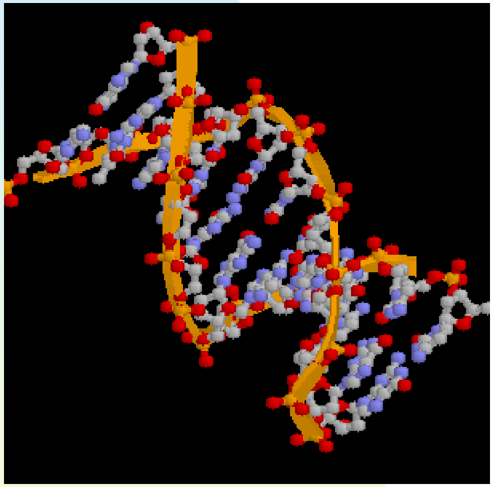
atcgggctatcgatagctatagcgcgatatatcgcgcgtatatgcgcgcatattag
tagctagtgctgattcatctggactgtcgtaatatatacgcgcccggctatcgcgct
atgcgcgatatcgcgcggcgctatataaatattaaaaaataaaatatatatatatgc
tgcgcgatagcgctataggcgcgctatccatatataggcgctcgcccgggcgcga
tgcatcggctacggctagctgtagctagtcggcgattagcggcttatatgcggcga
gcgatgagagtcgcggctataggcttaggctatagcgctagtatatagcggctagc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

Genetic Code: codons

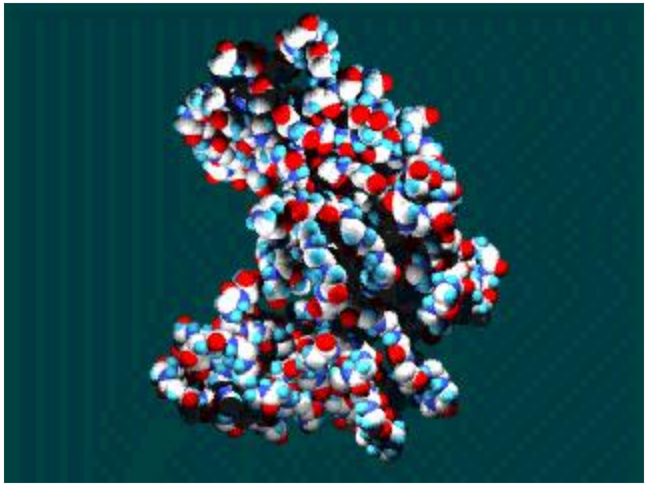


First Position		Second Position of Codon					Third Position
		T	C	A	G		
	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T	
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C	
		TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A	
		TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G	
	C	CTT 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

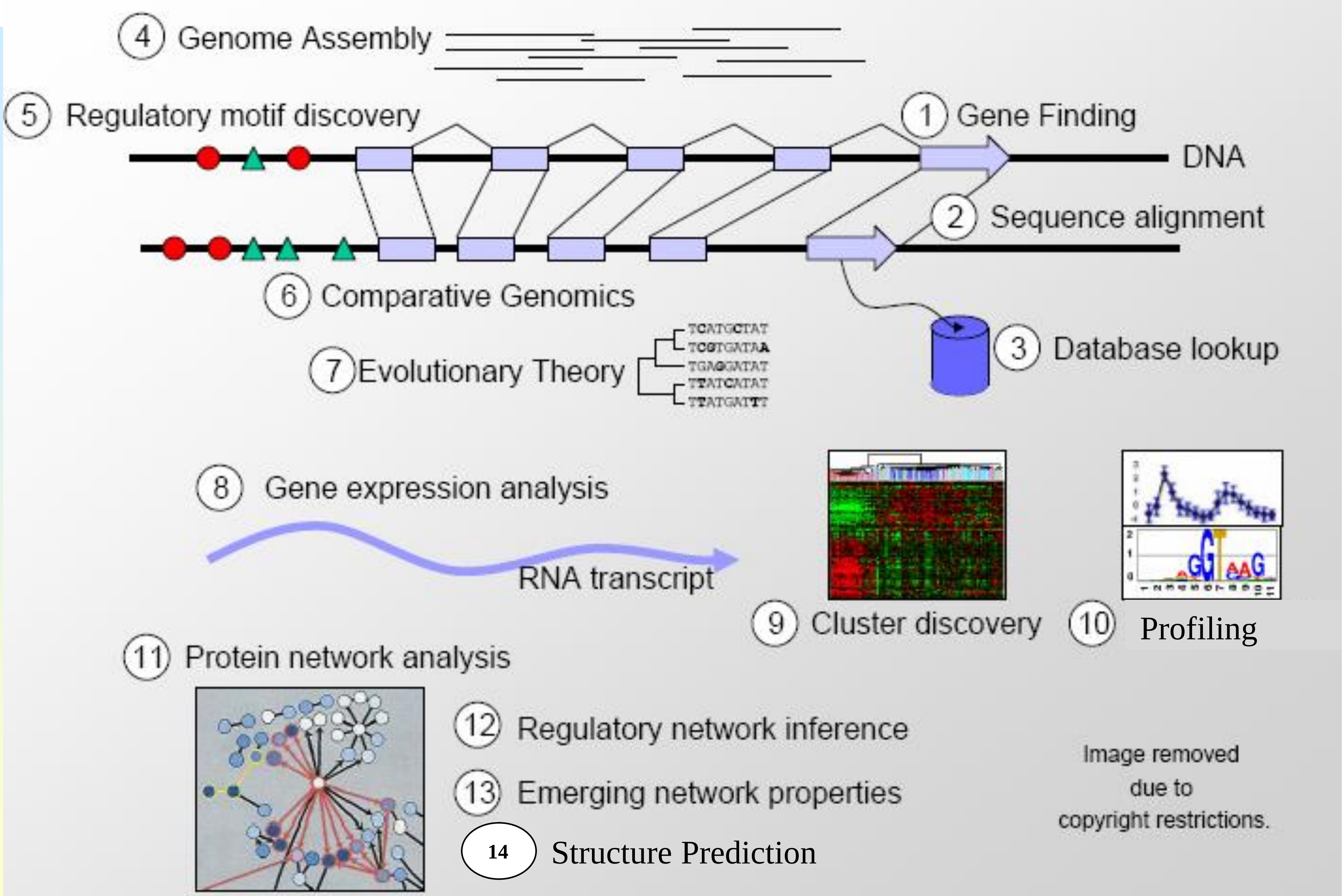


3-D biochemical structure

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

DNA

2 Sequence alignment

6 Comparative Genomics

7 Evolutionary Theory

3 Database lookup

8 Gene expression analysis

RNA transcript

9 Cluster discovery

10 Profiling

11 Protein network analysis

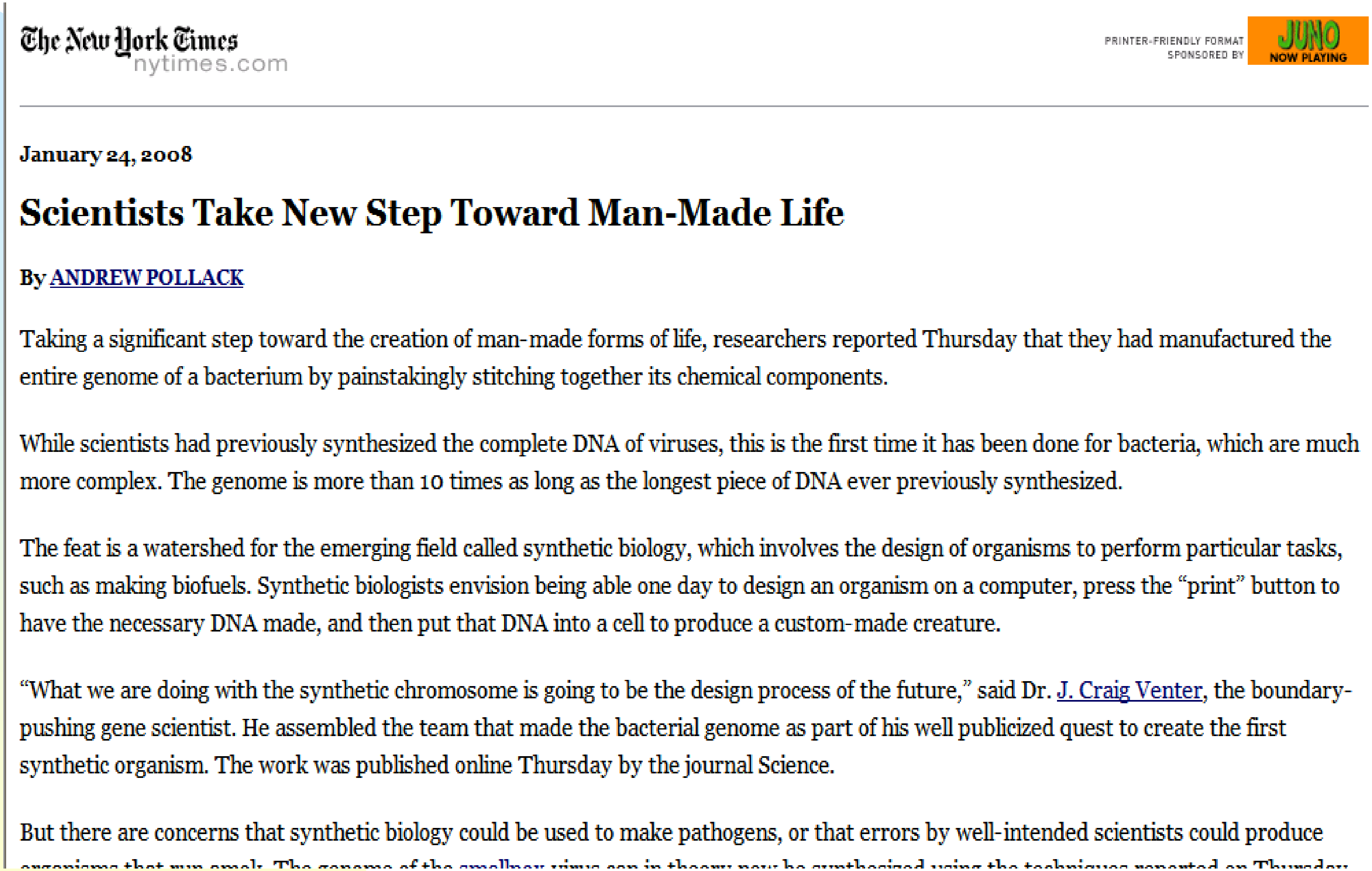
12 Regulatory network inference

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

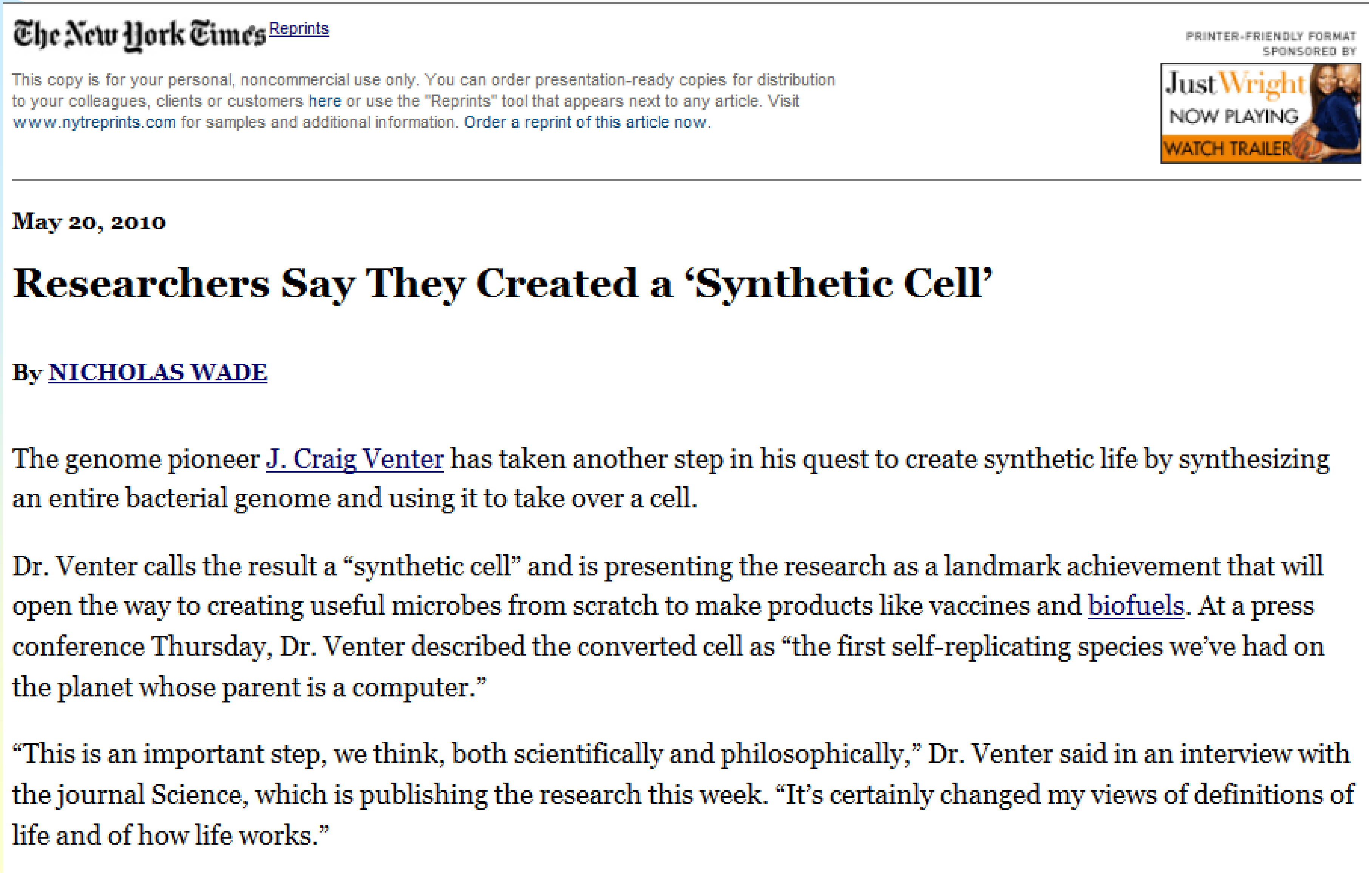
Taking a significant step toward the creation of man-made forms of life, researchers reported Thursday that they had manufactured the entire genome of a bacterium by painstakingly stitching together its chemical components.

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
  - Probabilistic
- Structure prediction
  - RNA Secondary
  - Protein Secondary & tertiary
- Network analysis:
  - Metabolic pathways reconstruction
  - PPI network
  - Regulatory networks (Gene expression)

How should I learn this course?

Come to the class, do homework assignments, reading assignments, and ask questions!

Nuts and Bolts: A lot of facts, new terminologies, models and algorithms

A typical approach to study almost any subject

> what is already known? (what is the state of the art, so you won't reinvent the wheel)

> what is unknown?

o Known unknowns

o unknown unknowns

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
    - ☞ Monte Carlo simulation
- Know the existing resources: Databases, Software, …