

AAGCTTACTCCGCATATTAAGTTTTGACTTTCCTGTGTTTTATTGGAACCTTTTGTTCCCTGAAGCACTAA  
CATTTCATTTTTTTCAGCAGTTAGCCACCTCCCTAATGGTCAACAGTTCCTTCAGGGCAATCAGG  
TCACTGCTACAAATGCGGTTCAGCCCAAATGATGCAGAATAATTAATCATTGTAAATGTTGGC  
CCTCAAAACAGTGTCTGGAATAAGGGATAACATGACATTTATCATGATCAGTGGAACTGGTTAATCT  
GGTCAGCAGTAAATGAAATTTTAGCTATTTAGCTGACTGGGCTTTTATTTTGTTTATTTGATTTTGAG  
AATTGATAATTAATCTGTAAGAGAGTGCTAGGCTCTGCACCTTTTGTTTTCTATCCAGCT  
ACTGACAAATGATTAATTTAGCTTTTGGGAGAACCAATGTTAAAGAGAAAGGAAGCA  
GGGACAGGCAATGATGATCACTGGGGTAGCCATATATCATATTOGCATGGTATGAT  
ATAATCATAATTAATGATGATCACTTTTGGATTTGGCTGCCCACTTGGTATTATGAG  
ATATTGCAATTAATAATTTAGAAAGCCCTAGTTGCTTGAATCTCTCAATCTAA  
ACTCAGTTTCTGTTGTAATGTGTGATAACAGTTTCAGACTTTTTTTTCTTTTCAG  
CGCTTCACTTAATTAATTTAGTTGTGCTTCCCTAATAAGCTTCCCTCCTTCTCTAAC  
ACAAGTAAAGCCCTGGACCAATCTCTAGGAATGCTGGCAATACAAGTGGAGAGAG  
CAGCTCTGCTGCAATGACAAACGGAGCAGTTGCTGAGCCACGTTTTG  
TGAGGCAATGATTTTATACTGGTGATTAGGATAATTTAGCATTGTTGGGAGGAA  
GGTCCTGTGCTTAAATATGCAGTATTACATTTTATGTTATGCCCTAATGCACA  
GCCCTACCACAAATTCAGGCAATTTTCAGGCTGATGAGGAAAGGATTTTATGATGCCATT  
TAAAGAAAGAAGGCATCAATGCAATCAATTAGAATGAGGTACAAAACAGATGAGCTAATGGCAAAAGA  
CTGATCCAAAAGGTTCCCATCAATCTAACATCACCTTGTAAGCTGAGTTACAAATAGACCTTTTCTT  
TCAGAATTGAGATACGGGAGTGTGATCTTTTCACATTGTCCTATTGGGGTGGTGAGAAATGTTTT  
ATTCAGTTTTTCATGAAACTATTAATAATGATCCATTATATGTTTCAGCACATTGTGTTGCACATAAAGACTT

# BIOKEMIJSKA INFORMATIKA

**prof. dr. Gregor Anderluh**

Kemijski inštitut, Ljubljana

Oddelek za biologijo, Biotehniška fakulteta, UL, Ljubljana

**gregor.anderluh@ki.si**

# **BIOKEMIJSKA INFORMATIKA**

**Uvod; znanost in računalniki, internet**

**Podatkovne zbirke, biološki podatki, bibliografske podatkovne zbirke**

**Določanje zaporedij DNA, genomika, analize nukleotidnih zaporedij**

**Določanje primarne zgradbe proteinov, analize zaporedij proteinov**

**Prileganje zaporedij, iskanje zaporedij v podatkovnih zbirkah, grupiranje zaporedij**

**Drugi biološki podatki**

**Določanje 3D zgradbe proteinov, analize 3D zgradbe**

# VIRI

- <http://web.bf.uni-lj.si/bi/biokemija/bioinfo/bioinfo.htm>

Spletna stran o bioinformatiki (ne preveč osvežena)

- Literatura

Popoln seznam na zgornjih spletnih straneh

**ATTWOOD, TK, PARRY-SMITH DJ (1999) Introduction to Bioinformatics. Prentice Hall, Harlow, England.**

**MOUNT, D. 2004. Bioinformatics. Sequence and Genome Analysis, 2nd edition. Cold Spring Laboratory Press, Cold Spring Harbor, New York.**

- gregor.anderluh@ki.si

# OBVEZNOSTI

## **Predavanja - prof. dr. Gregor Anderluh**

2 uri tedensko;

- pregled področja
- teoretične osnove bioinformatških orodij
- nekateri primeri iz literature
- prosojnice na spletni učilnici (<http://ucilnica.fkkt.uni-lj.si>)

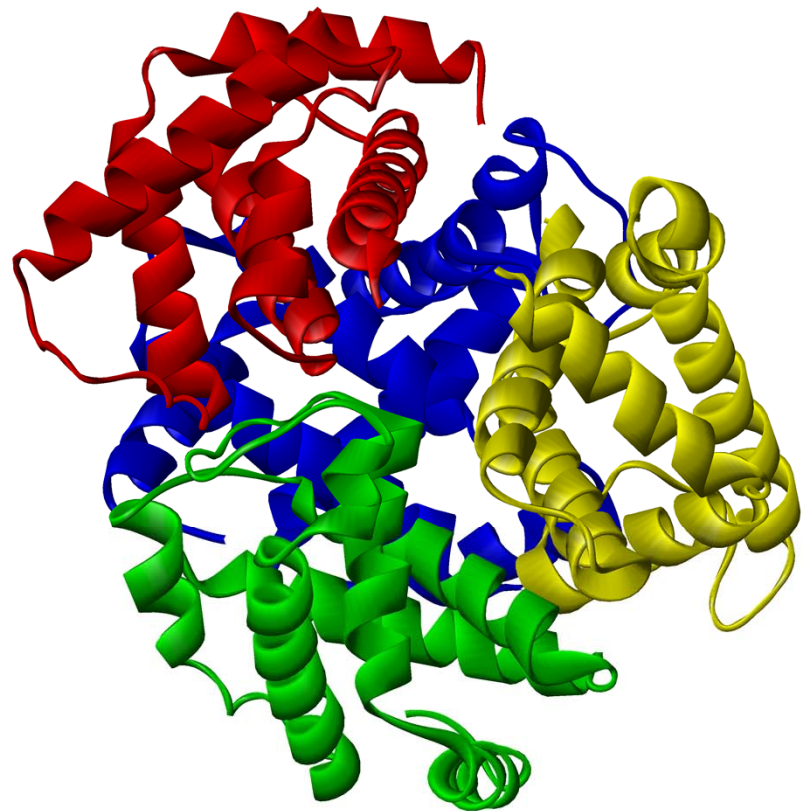
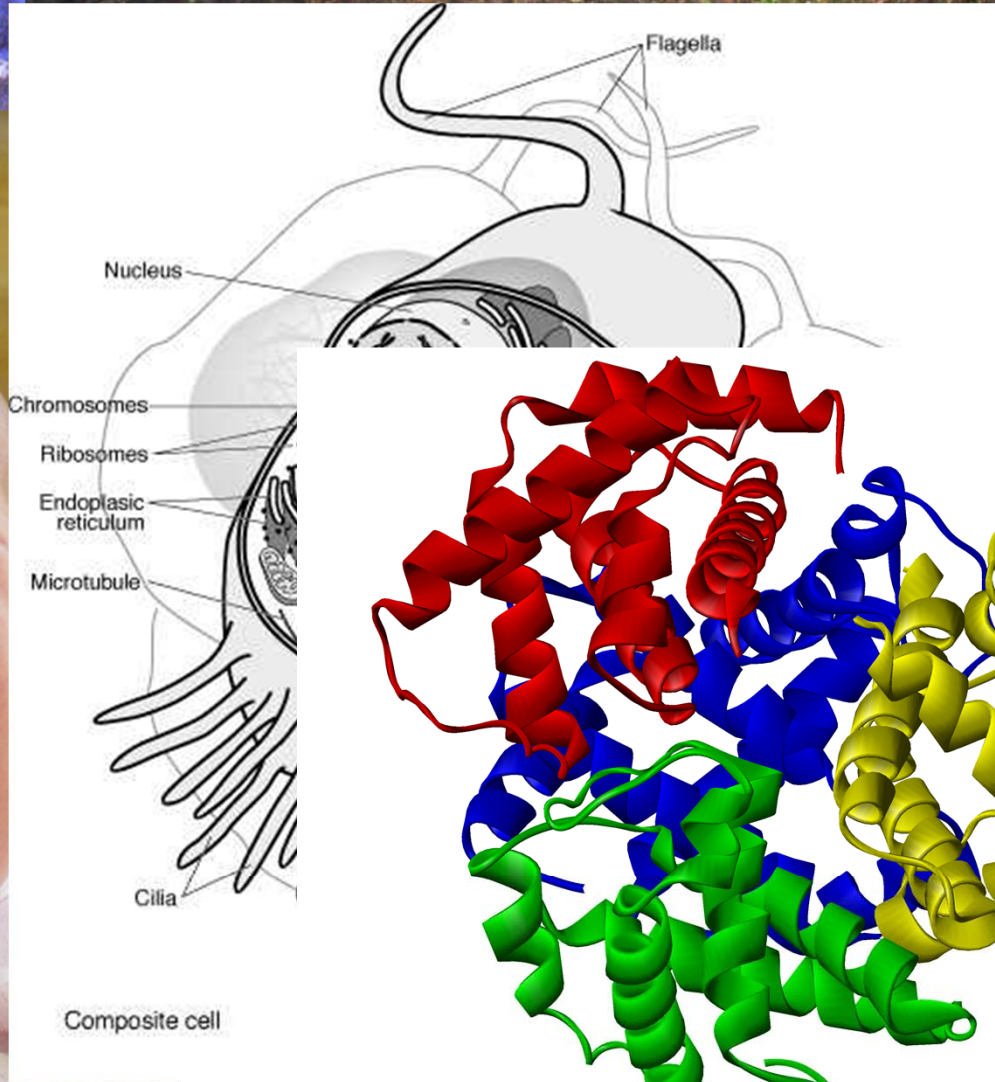
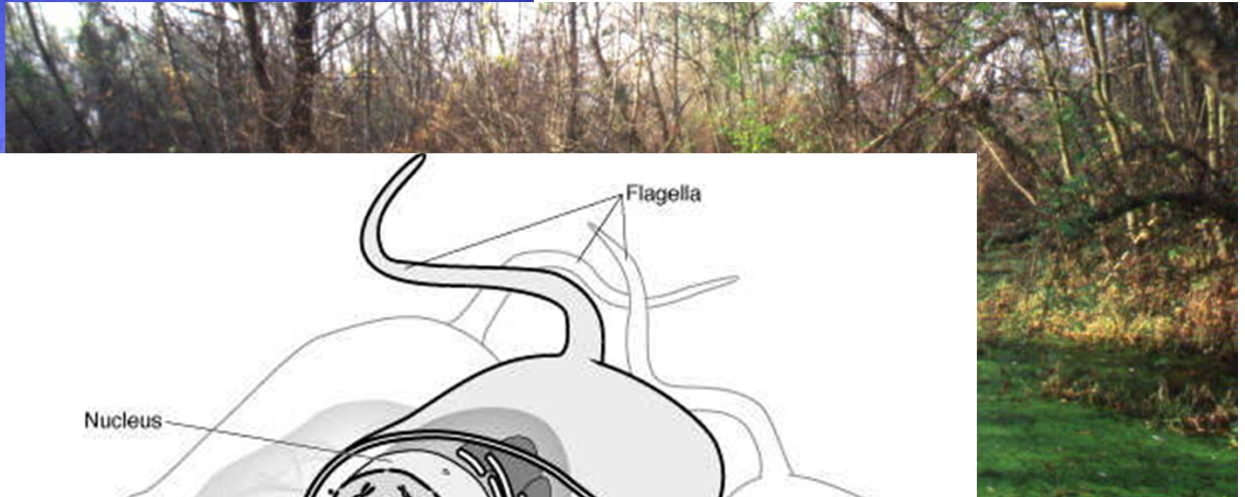
## **Vaje - dr. Miha Pavšič in Aljaž Gaber, univ. dipl. biokem.**

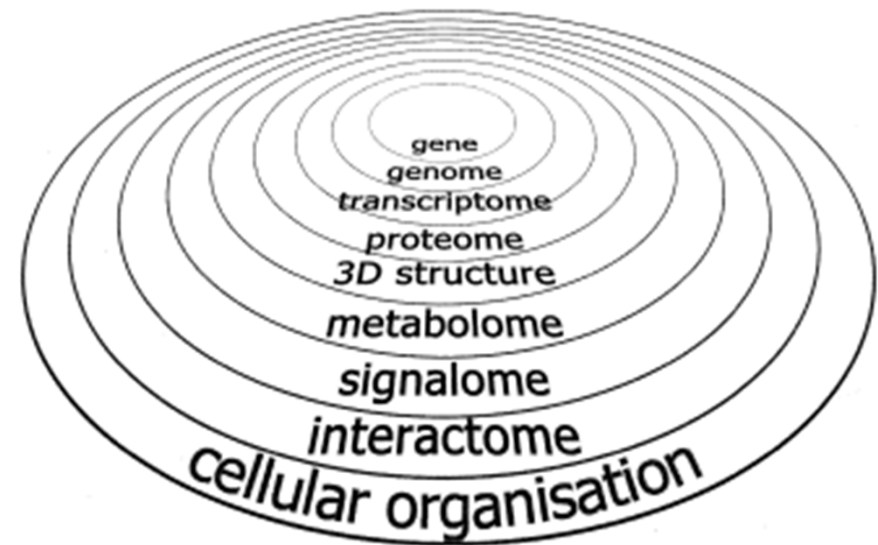
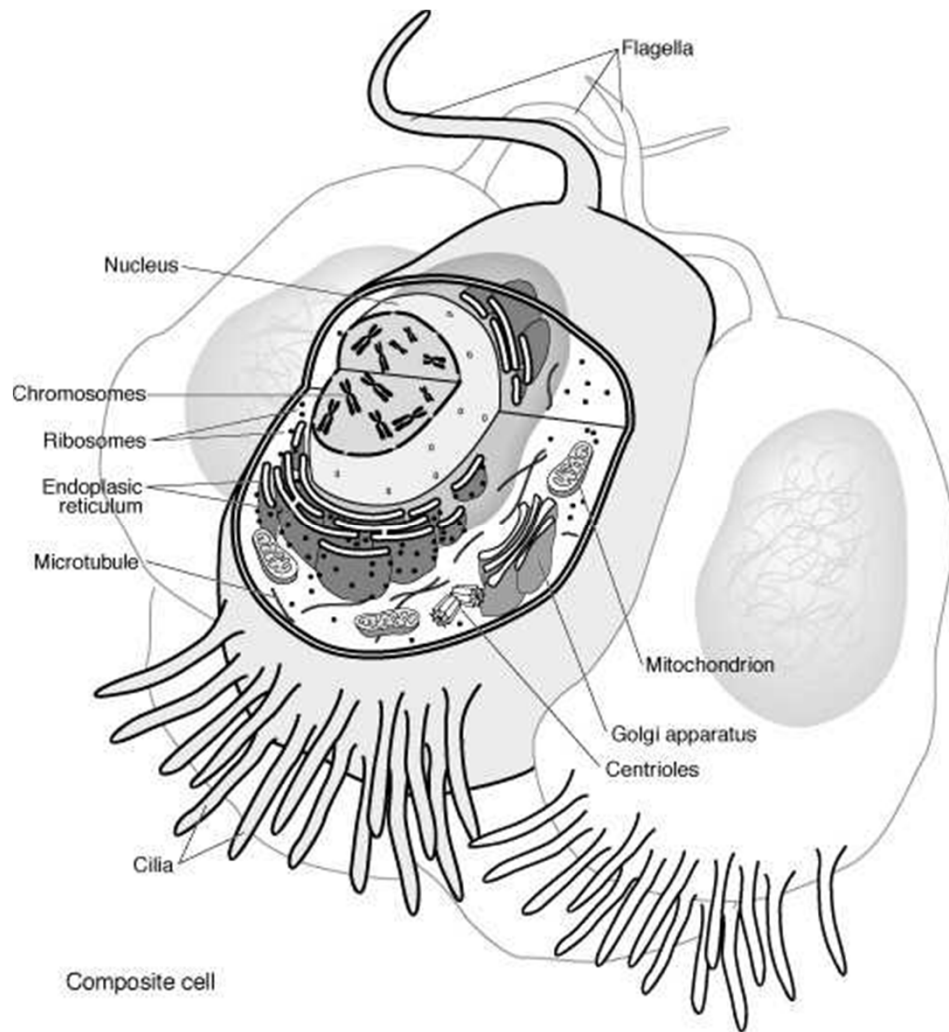
3 ure tedensko;

- praktična uporaba bioinformatških orodij
- material bo sproti objavljen na spletni učilnici

## **Obveznosti**

- kolokvij iz vaj (praktično – v računalniški učilnici)
- pisni izpit (pogoj za pristop k izpitu je opravljen kolokvij)





Vihinen M (2001) *Biomol. Eng.* 18:241-248

# BIOKEMIJA, MOLEKULARNA BIOLOGIJA

biokemijski procesi, encimi

## OSNOVNI CELIČNI GRADNIKI

DNA, beljakovine, zgradba beljakovin, zgradba DNA,

razvoj molekularne biologije (**PRE-GENOMIKA**)

“high-throughput” biologija, določanje nukleotidnega zaporedja DNA genoma (**GENOMIKA**)

## BIOINFORMATIKA

razvoj računalništva, pomoč pri delu biokemika,

podatkovne zbirke (proteini, DNA, 3D strukture), programi

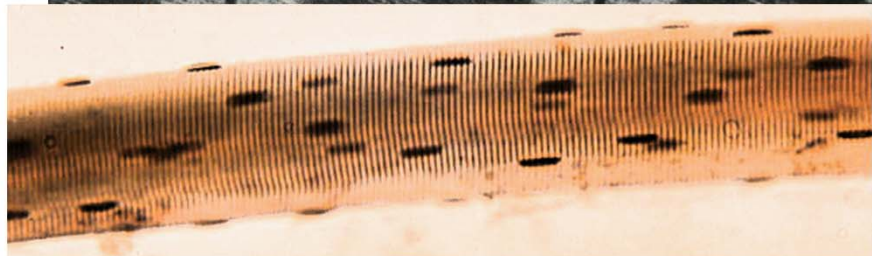
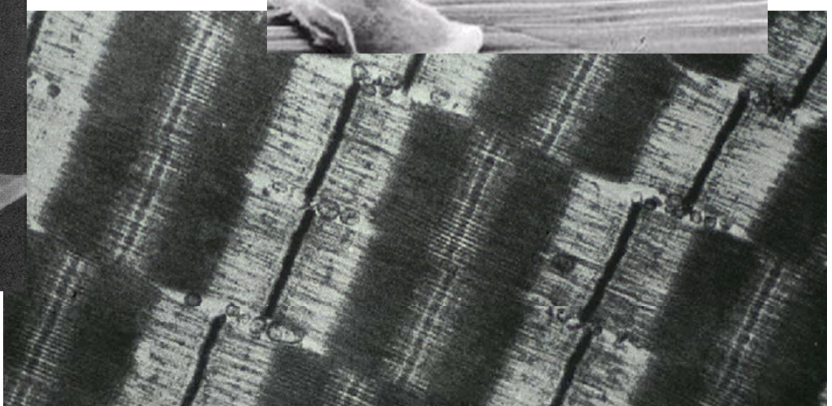
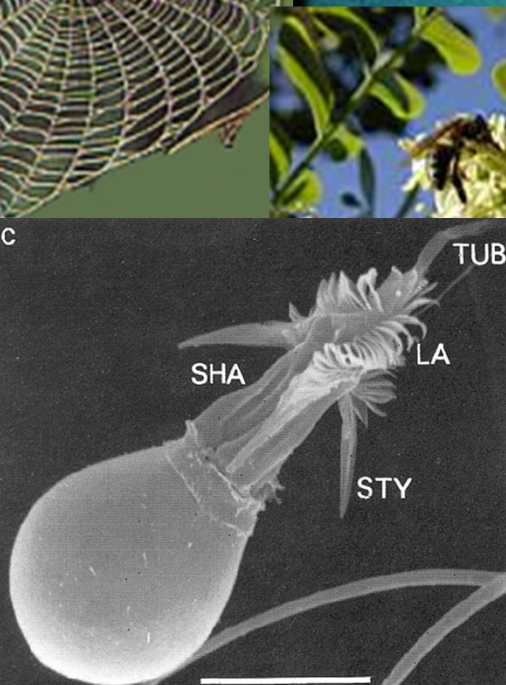
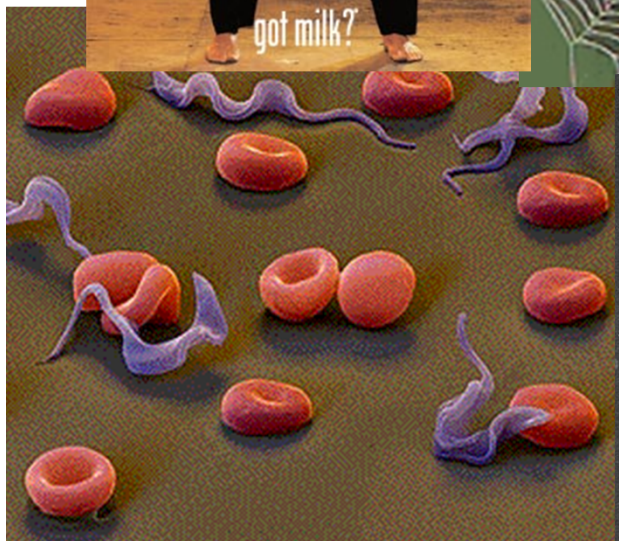
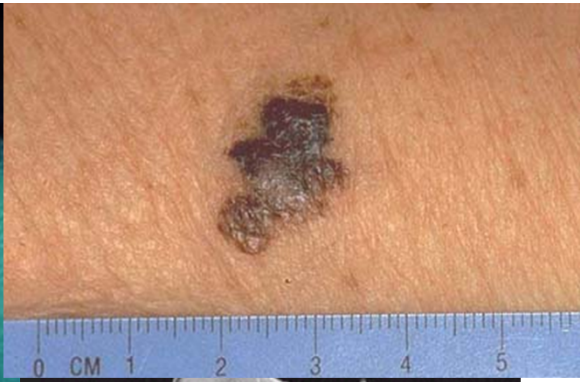
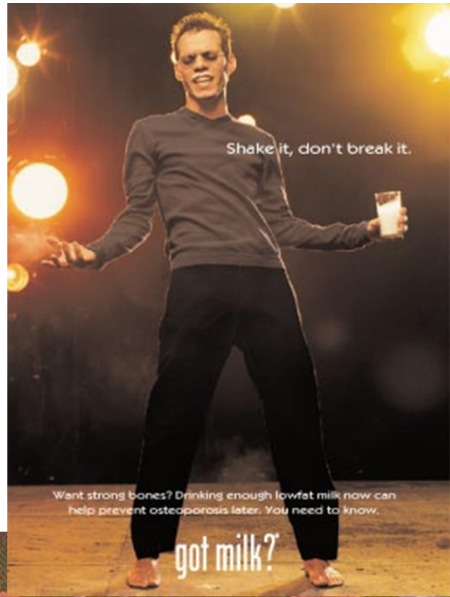
## **POST-GENOMIKA**

strukturna genomika, funkcionalna genomika, metagenomika,

okoljska genomika, transkriptom, proteom, interaktom,

metabolom, integrirane podatkovne zbirke, specializirani

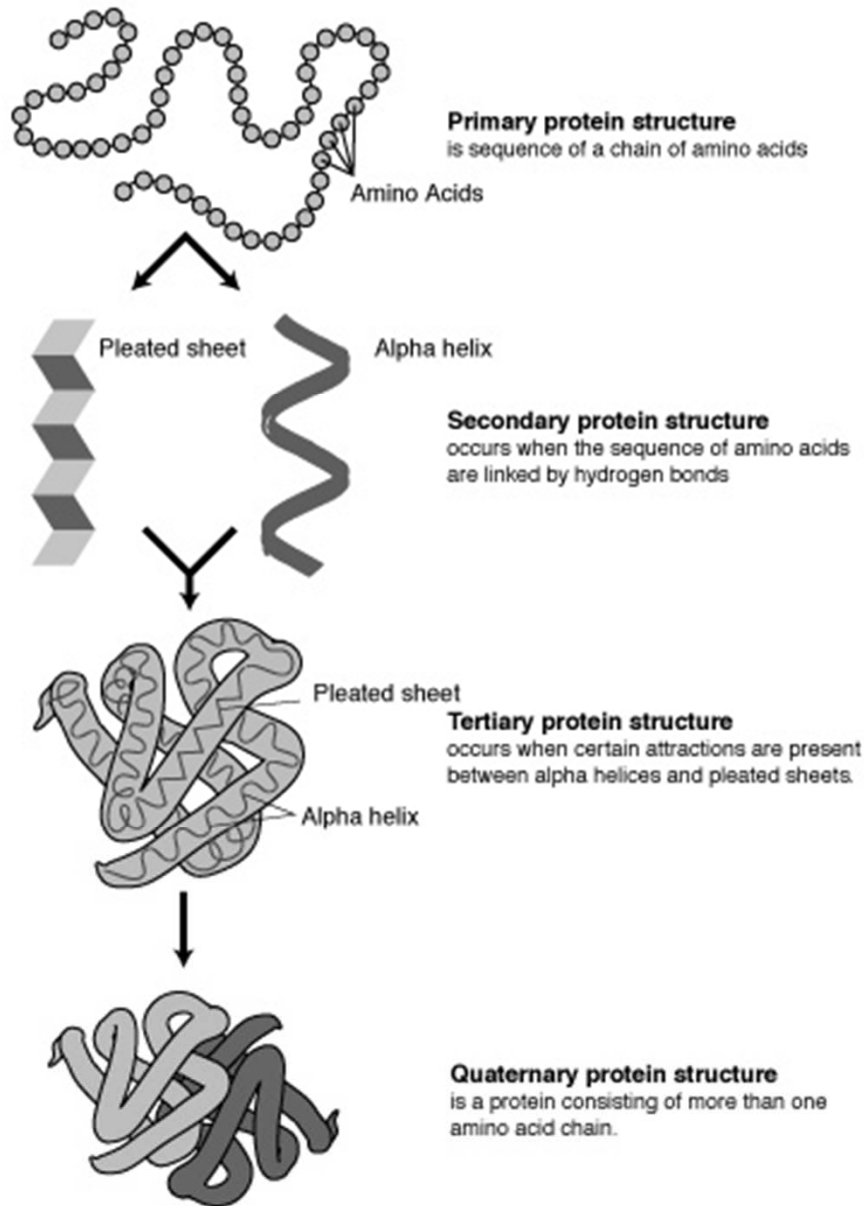
strežniki



The proteins are among the most complicated and enigmatic substances in Nature and appear to be particularly closely related to all that we call Life.

Presentation speech, Nobel prize for Chemistry 1958





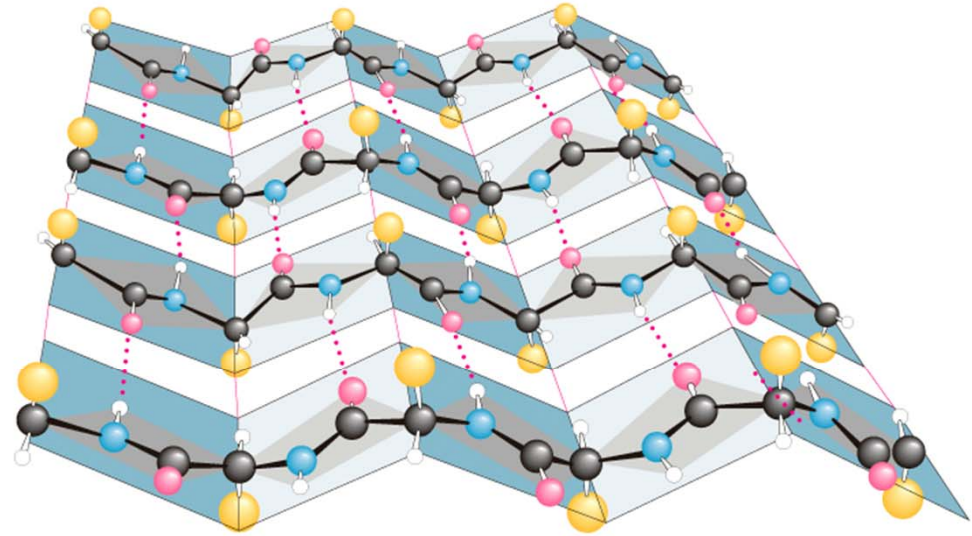
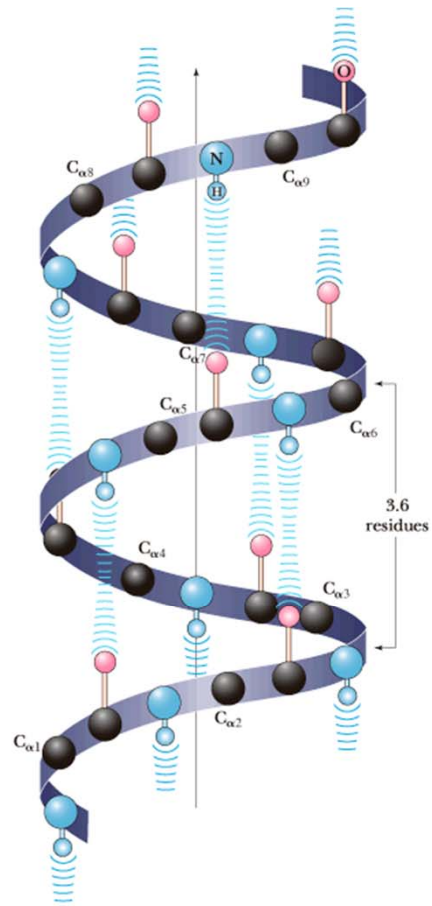
>SLIT\_DROME (P24014):

MAAPSRRTLMPPPFRQLRLLILPILLLLLRHDAVHAEPYSGGFGSSAVS  
 SGGLGSVGIHIPGGGVGVITEARCPRVCSCTGLNVD**C**SHRGLT**S**VPRKI  
**S**ADVERLEL**Q**GNNLT**V**IYETDF**Q**RLTKLRML**Q**LTDN**Q**IHTIERN**S**F**Q**DL  
**V**SLERLDISNNVIT**T**VGRRV**F**KGA**Q**SLRSL**Q**L**D**NN**Q**IT**C**LDEHAF**K**GLV  
**E**LEIL**T**LNNNNL**T**SLPHN**I**F**G**GL**R**LRAL**R**LS**D**NP**F**ACD**C**HL**S**WL**S**R**F**L  
 RSATRLAPYTRCQSP**S**QLKGQ**N**VADLHD**Q**E**F**KCSGLTEHAP**M**ECGAENS  
 CPHPCRCADGIVDCREKSL**T**SVPV**T**L**P**DD**T**TDV**R**LE**Q**N**F**IT**E**L**P**PK**S**F**S**  
**S**FRRL**R**RI**D**LS**N**NNI**S**RIA**H**DAL**S**GL**K**QL**T**TL**V**LYGN**K**IK**D**LP**S**GV**F**K**G**  
**L**GS**L**RL**L**LLNANEI**S**CI**R**KDA**F**RD**L**H**S**LS**L**SL**S**LYD**N**NI**Q**SLAN**G**TF**D**AM  
 KSMKT**V**HLAK**N**PF**I**DC**N**LR**L**W**L**AD**L**Y**L**H**K**NP**I**ET**S**GAR**C**ES**P**K**R**M**H**RR**R**I  
 ESL**R**E**E**K**F**KCS**W**GEL**R**M**K**LS**G**EC**R**MD**S**DC**P**AM**C**H**C**EG**T**TV**D**CT**G**RR**L**KE  
 I**P**RD**I**PL**H**TT**E**LL**L**ND**N**EL**G**R**I**SS**D**GL**F**GR**L**PH**L**V**K**LE**L**K**R**N**Q**L**T**G**I**E**P**  
**N**A**F**E**G**ASH**I****Q**E**L****Q**L**G**EN**K**I**K**E**I****S**N**K**M**F**L**G**L**H**Q**L**K**T**L**N**LY**D**N**Q**I**S**C**V**M**P**G  
**S**F**E**HL**N**SL**T**SL**N**LAS**N**P**F**NC**N**CH**L**AW**F**A**E**CV**R**KK**S**L**N**GG**A**ARC**G**AP**S**K**V**  
 RD**V**Q**I**K**D**LP**H**S**E**FK**S**SEN**S**E**G**CL**G**D**G**Y**C**PP**S**CT**C**T**G**TV**V**AC**S**R**N**Q**L**KE  
 I**P**RG**I**PA**E**T**S**E**L**Y**L**E**S**NE**I**E**Q**I**H**Y**E**R**I**R**H**L**R**SL**T**RL**D**LS**N**N**Q**IT**I**LS**N**Y  
 TFAN**L**TK**L**ST**L**I**I**SY**N**KL**Q**CL**Q**R**H**AL**S**GL**N**N**L**R**V**VS**L**H**G**N**R**I**S**ML**P**E**G**S  
 FED**L**K**S**L**T**H**I**AL**G**SN**P**LY**C**DC**G**L**K**W**F**SD**W**IK**L**D**V**Y**E**P**G**IAR**C**A**E**PE**Q**M**K**  
 DK**L**IL**S**TP**S**SS**F**VC**R**GR**V**R**N**D**I**LAK**C**NAC**F**E**Q**PC**Q**N**Q**A**Q**CV**A**L**P**Q**R**E**Y**Q  
**C**L**C**Q**P**GY**H**G**K**H**C**E**F**MI**D**AC**Y**GN**P**CR**N**NAT**C**TV**L**E**E**GR**F**S**C**Q**C**AP**G**Y**T**G**A**  
**R**C**E**T**N**ID**D**CL**G**E**I**K**C**Q**N**NAT**C**ID**G**V**E**SY**K****C**E**C**Q**P**G**F**S**G**E**F**C**D**TK**I**Q**F**C**S**  
 PEF**N**P**C**ANGAK**C**MD**H**F**T**H**S****C**DC**Q**AG**F**H**G**T**N**CT**D**NID**D**C**Q**N**H**M**C**Q**N**GG**T**  
 C**V**D**G**I**N**D**Y**Q**C**R**C**P**D**D**Y**T**G**K**Y**C**E**GH**N**M**I**S**M**Y**P**Q**T**SP**C**Q**N**H**E**C**K**H**G**V**C**F**Q**  
 P**N**A**Q**GS**D**Y**L**C**R**CH**P**GY**T**G**K**W**C**E**Y**L**T**S**I**S**F**V**H**NN**S**F**V**E**L**E**P**L**R**TR**P**EAN**V**  
 T**I**V**F**SS**A**E**Q**NG**I**L**M**Y**D**G**Q**DA**H**L**A**VEL**F**NG**R**I**R**V**S**Y**D**V**G**N**H**P**V**ST**M**Y**S**F**E**  
 MV**A**D**G**K**Y**H**A**VEL**L**A**I**K**N**F**T**LR**V**DR**G**L**A**R**S**I**I**NE**G**S**N**D**Y**L**K**L**T**TP**M**FL**G**  
 GL**P**V**D**PA**Q**Q**A**Y**K**N**W**Q**I**R**N**L**T**S**F**K**G**C**M**KE**V**W**I**N**H**K**L**V**D**F**G**NA**Q**R**Q**K**I**T**P**  
 G**C**AL**L**E**G**E**Q**Q**E**E**D**E**Q**D**F**MD**E**T**P**H**I**K**E**E**P**V**D**PC**L**EN**K**CR**R**G**S**R**C**V**P**NS  
 N**A**RD**G**Y**Q****C**K**C**K**H**G**Q**R**G**R**Y**C**D**Q**G**E**G**S**T**E**P**P**T**V**T**A**A**ST**C**R**K**E**Q**V**R**E**Y**Y**T**EN  
 DC**R**SR**Q**PL**K**Y**A**K**C**V**G**GC**N**Q**C**CA**K**I**V**RR**R**R**K**V**R**M**V**C**S**NN**R**K**Y**I**K**N**L**D**I**V  
 RK**C**G**C**T**K**K**C**Y

Leucine rich repeat EGF domain



1951



## Struktura $\alpha$ -heliksa in $\beta$ -ploskev

Pauling and Corey (1951) *Proc. Natl. Acad. Sci. USA* 27, 205-211; *Proc. Natl. Acad. Sci. USA* 37, 729-740



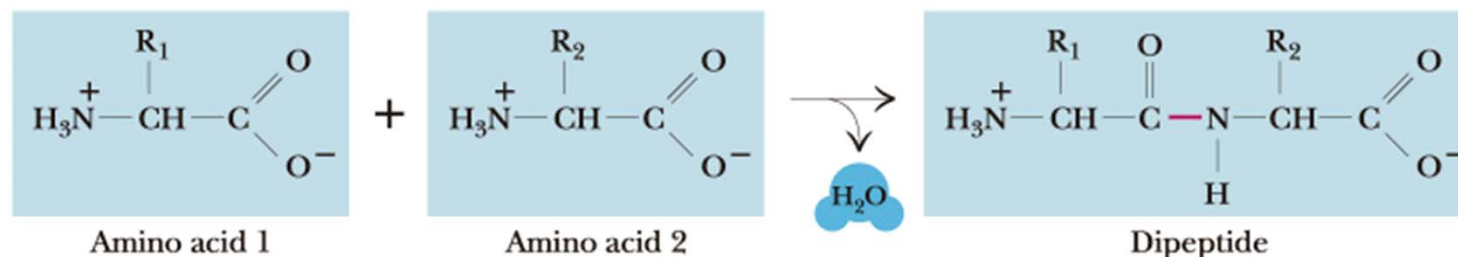
## The Nobel Prize in Chemistry 1954

### Linus Carl Pauling

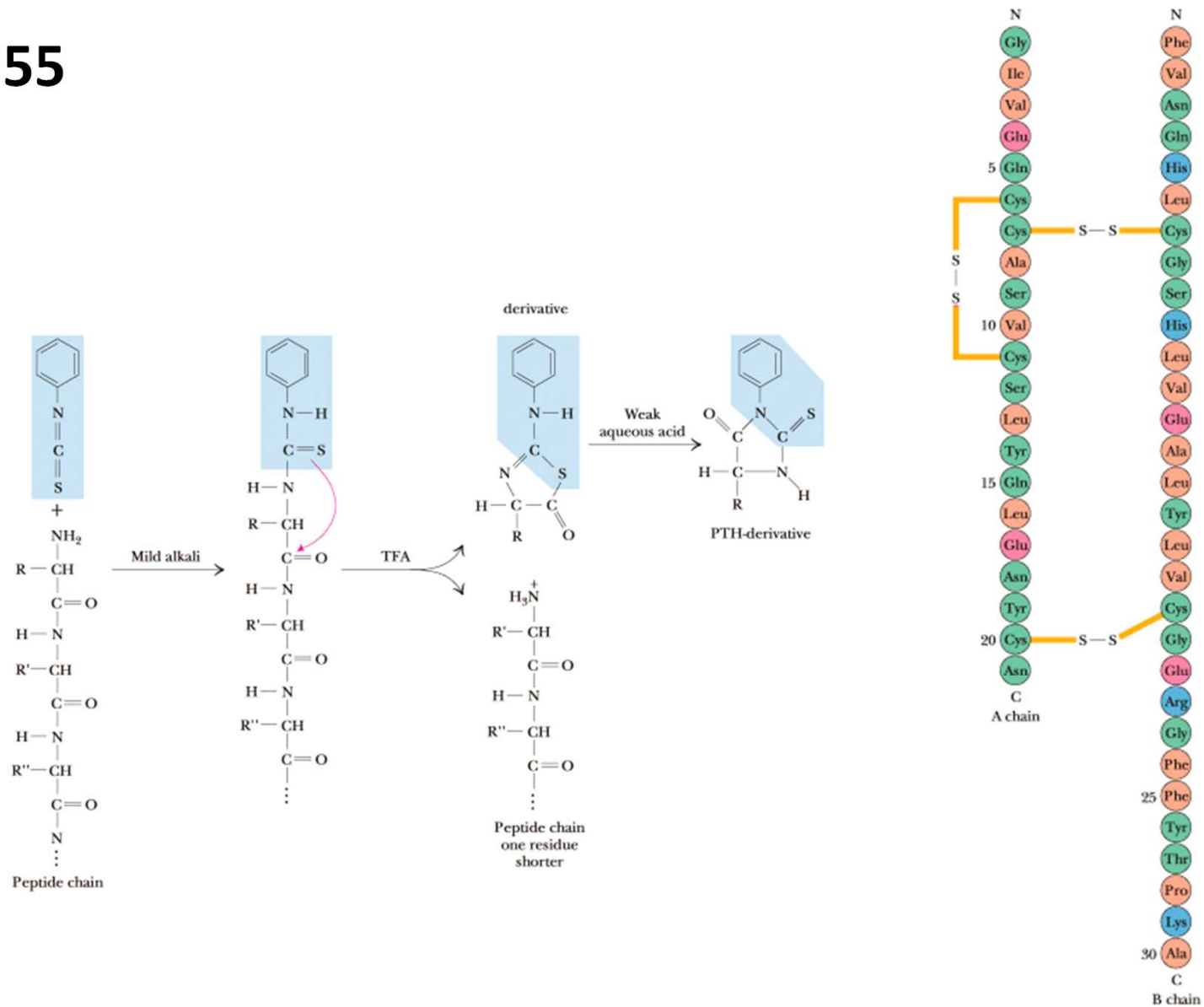
"for his research into the nature of the chemical bond and its application to the elucidation of the structure of complex substances "

Pauling has tried this method in his studies of the structure of proteins with which he has been occupied during recent years. **To make a direct determination of the structure of a protein by X-ray methods is out of the question for the present, owing to the enormous number of atoms in the molecule.** A molecule of the coloured blood constituent hemoglobin, which is a protein, contains for example more than 8,000 atoms.

On this basis Pauling deduced some possible structures of the fundamental units in proteins, and the problem was then to examine whether these could explain the X-ray data obtained. It has thus become apparent that one of these structures, the so-called alpha-helix, probably exists in several proteins.



# 1955



## Prvo aminokislinsko zaporedje proteina (insulin)

Sanger, F., Thompson, E.O.P. and Kitai, R. (1955) *Biochem. J.* **59**, 509



## The Nobel Prize in Chemistry 1958

### Frederick Sanger

"for his work on the structure of proteins, especially that of insulin"

Doctor Frederick Sanger. It sometimes happens that an important scientific discovery is made so to say "overnight" - if the time is ripe and the necessary background is there. Yours is not of that kind. The first successful determination of the structure of a protein is the result of many years of persistent and zealous work, in which the final solution of the problem has been approached step by step. You knew when you began to look into the structure of the insulin molecule 15 years ago that the problem was a formidable one.



## The Nobel Prize in Chemistry 1962

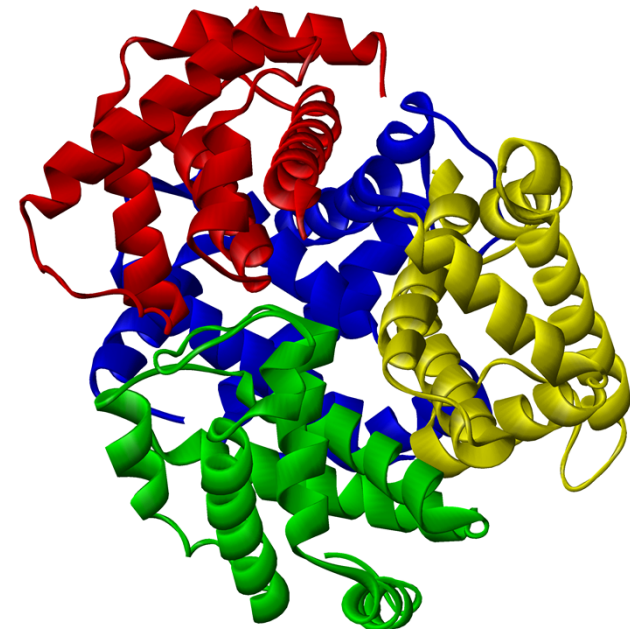
**Max Ferdinand Perutz**

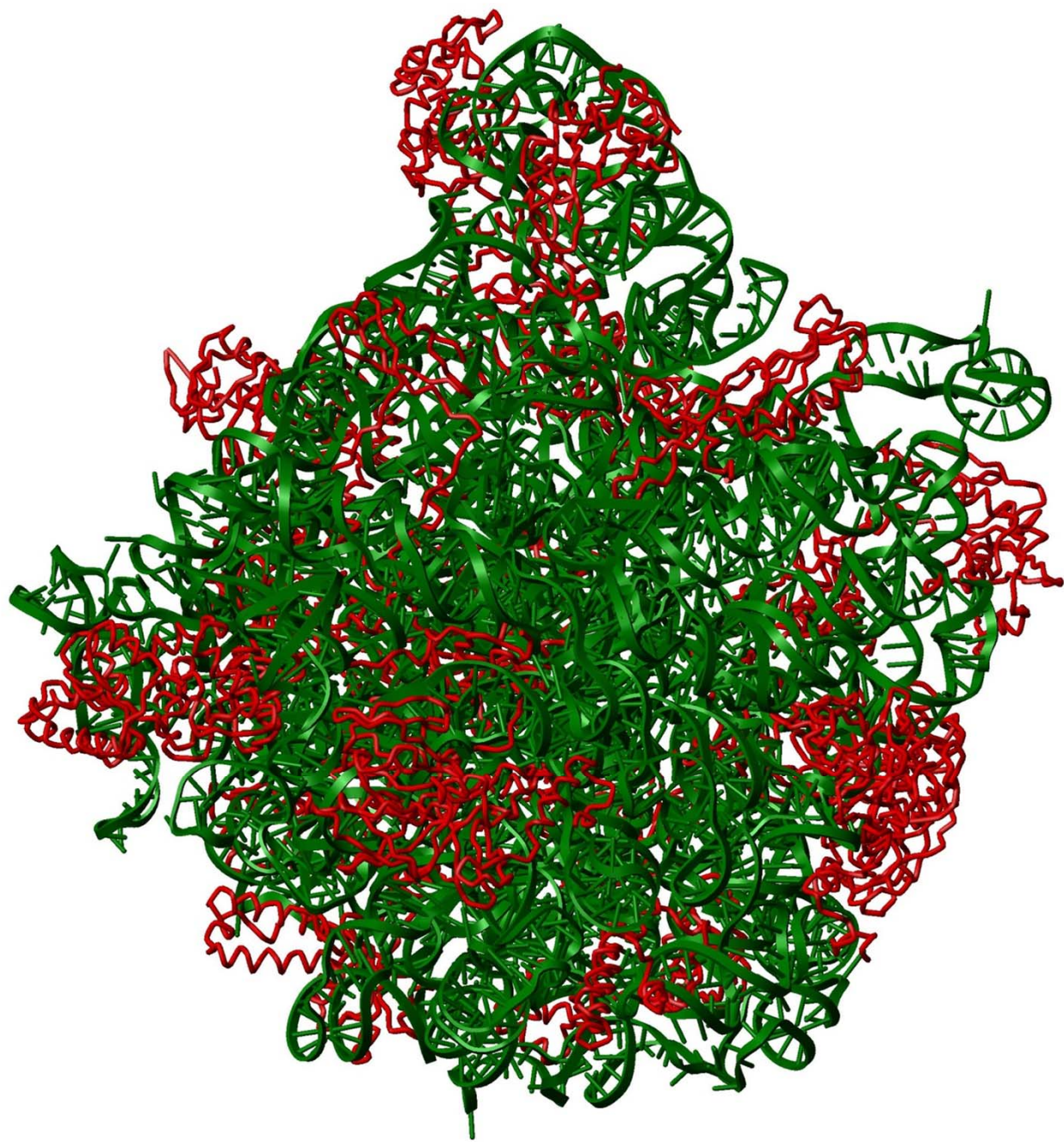
**John Cowdery Kendrew**

"for their studies of the structures of globular proteins"

But even if the path was now open for a direct structure determination of haemoglobin and myoglobin, there was still an enormous amount of data to be processed.

Myoglobin, the smaller of the two molecules, contains about 2,600 atoms, and the positions of most of these are now known. But for this purpose, Kendrew had to examine 110 crystals and measure the intensities of about 250,000 X-ray reflections. **The calculations would not have been practicable if he had not had access to a very large electronic computer.** The haemoglobin molecule is four times as large, and its structure is known less thoroughly. In both cases, however, Kendrew and Perutz are currently collecting and processing an even greater number of reflections in order to obtain a more detailed picture.





# 1953

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

- <sup>1</sup> Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
- <sup>2</sup> Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **6**, 285 (1949).
- <sup>3</sup> Von Arx, W. S., *Woods Hole Papers in Phys. Oceanog. Meteor.*, **11** (3) (1950).
- <sup>4</sup> Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

## MOLECULAR STRUCTURE OF NUCLEIC ACIDS

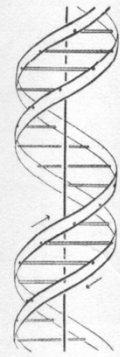
### A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey<sup>1</sup>. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining  $\beta$ -D-deoxyribofuranose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's<sup>2</sup> model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis.

is a residue on each chain every 3.4 Å. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 Å. The distance of a phosphorus atom from the fibre axis is 10 Å. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally<sup>3,4</sup> that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

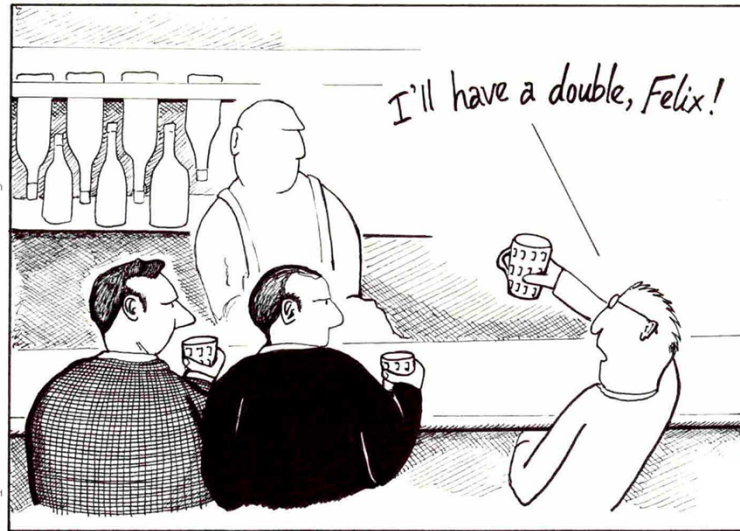
The previously published X-ray data<sup>5,6</sup> on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

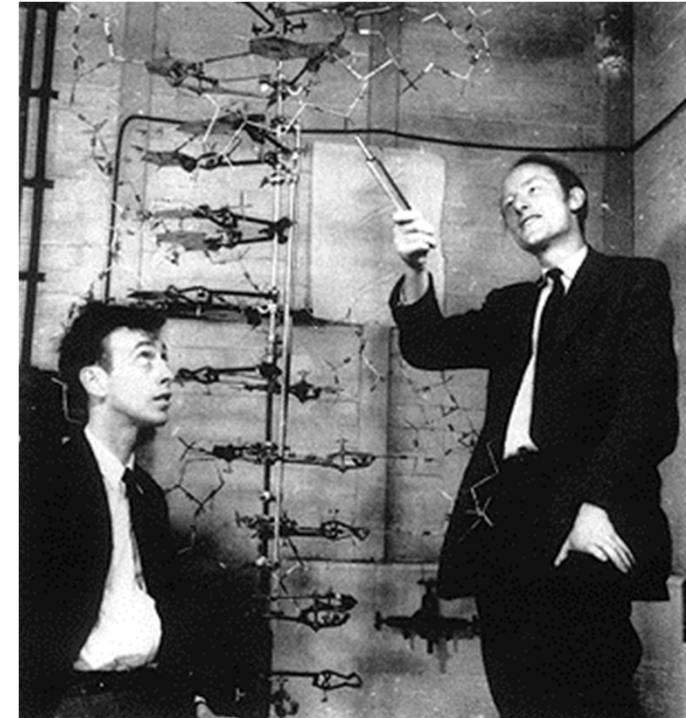
Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

STRANGE MATTER  
by nick d. kim  
strange-matter.com



Cambridge, 1953. Shortly before discovering the structure of DNA, Watson and Crick, depressed by their lack of progress, visit the local pub.







## The Nobel Prize in Physiology or Medicine 1962

**Francis Harry Compton Crick**

**James Dewey Watson**

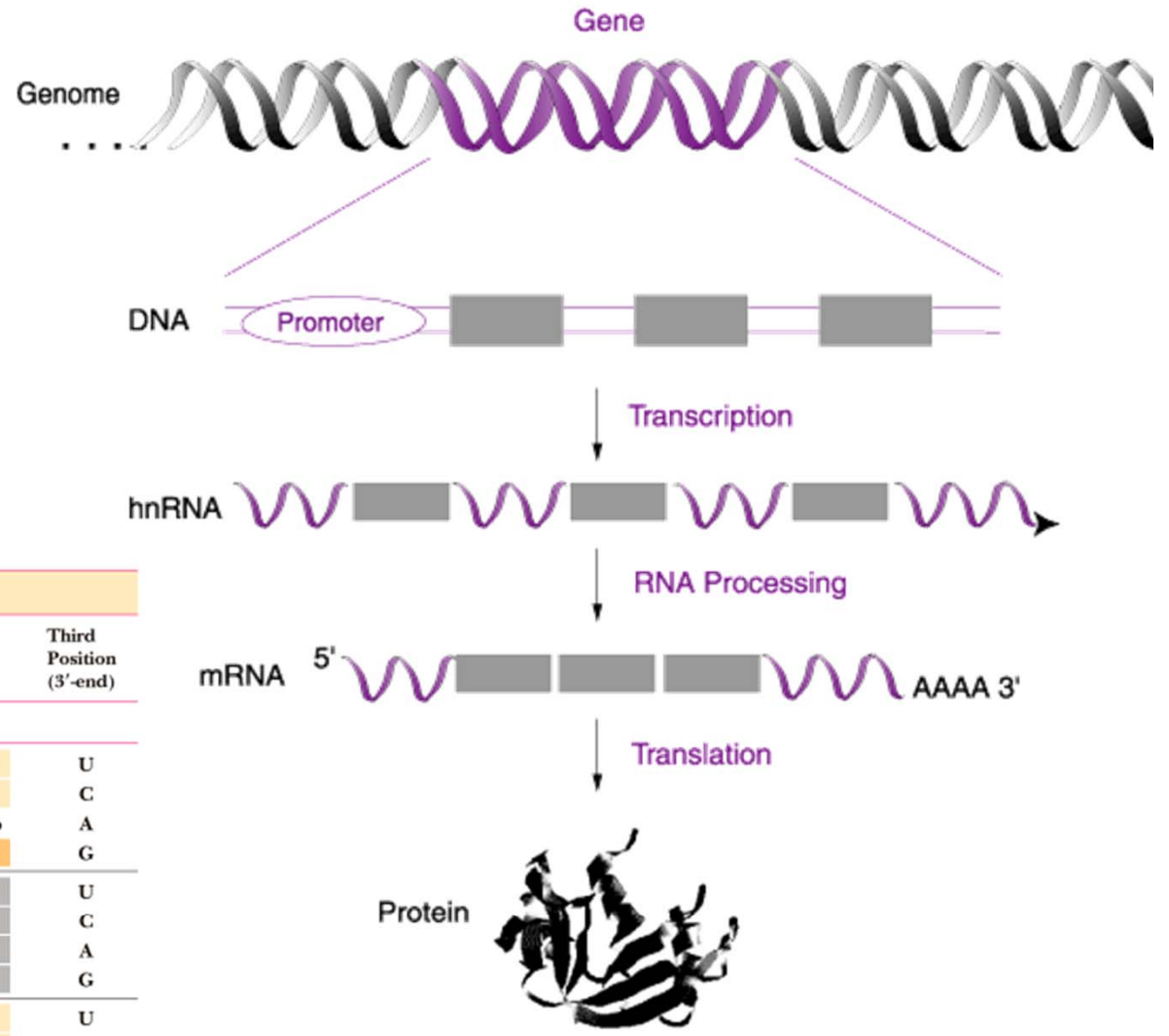
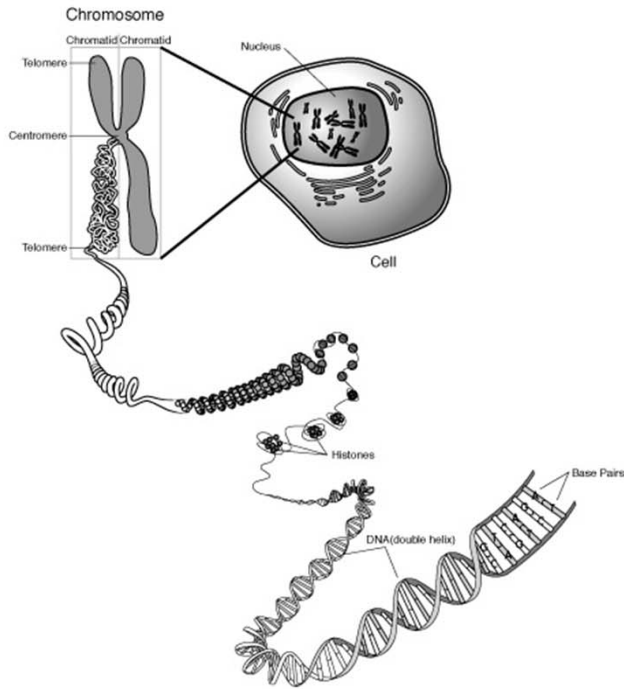
**Maurice Hugh Frederick Wilkins**

"for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material"



Today no one can really ascertain the consequences of this new exact knowledge of the mechanisms of heredity. We can foresee new possibilities to conquer disease and to gain better knowledge of the interaction of heredity and environment and a greater understanding for the mechanisms of the origin of life. In whatever direction we look we see new vistas. **We can, through the discovery by Crick, Watson and Wilkins, to quote John Kendrew, see «the first glimpses of a new world».**





**The Genetic Code**

First Position (5'-end)	Second Position				Third Position (3'-end)
	U	C	A	G	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met*	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

## The Nobel Prize in Physiology or Medicine 1968

**Robert W. Holley**

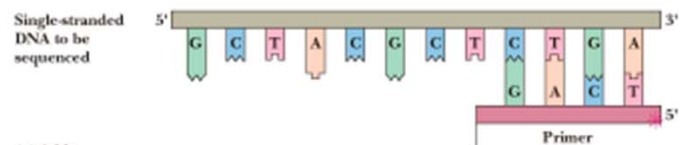
**Har Gobind Khorana**

**Marshall W. Nirenberg**

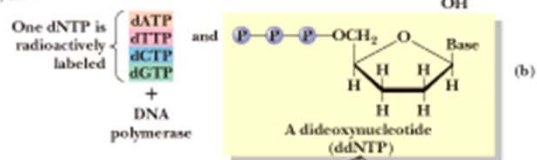
"for their interpretation of the genetic code and its function in protein synthesis"

Dr. Holley, Dr. Khorana, Dr. Nirenberg. At the end of his Nobel lecture, Edward Tatum in 1958 looked into his crystal ball and tried to predict some of the future developments in molecular biology. He suggested among other things that the solution of the genetic code might come during the lifetime of at least some of the members of his audience. This appeared to be a bold prophecy at that time. In reality it took less than three years before the first letters of the code were deciphered and, because of the ingenuity of you three, the nature of the code and much of its function in protein synthesis were known within less than eight years. Together you have written the most exciting chapter in modern biology.

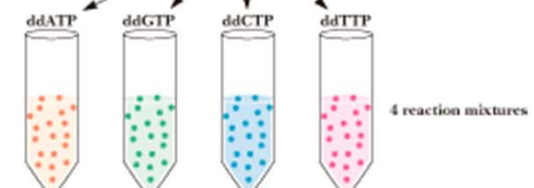




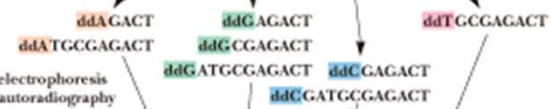
(a) Add:



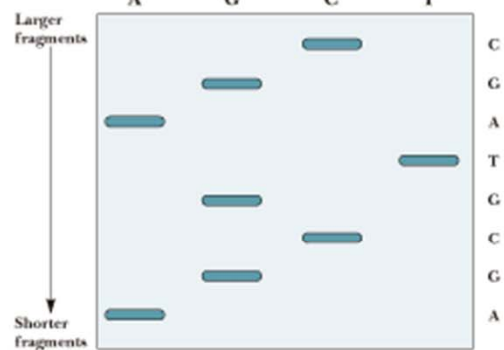
(c)



Reaction products

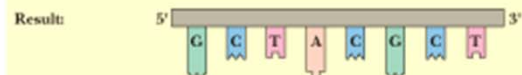


(d)

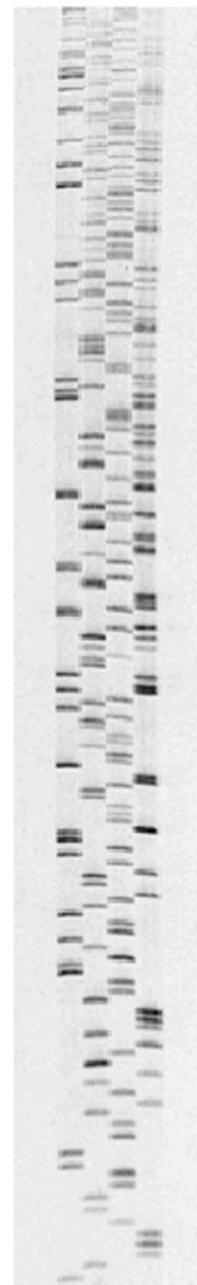


Reading sequence bottom to top: -A-G-C-G-T-A-G-C-

Its complement is the original template strand (3'→5'): -T-C-G-C-A-T-C-G-



ATGC



## The Nobel Prize in Chemistry 1980

**Walter Gilbert**

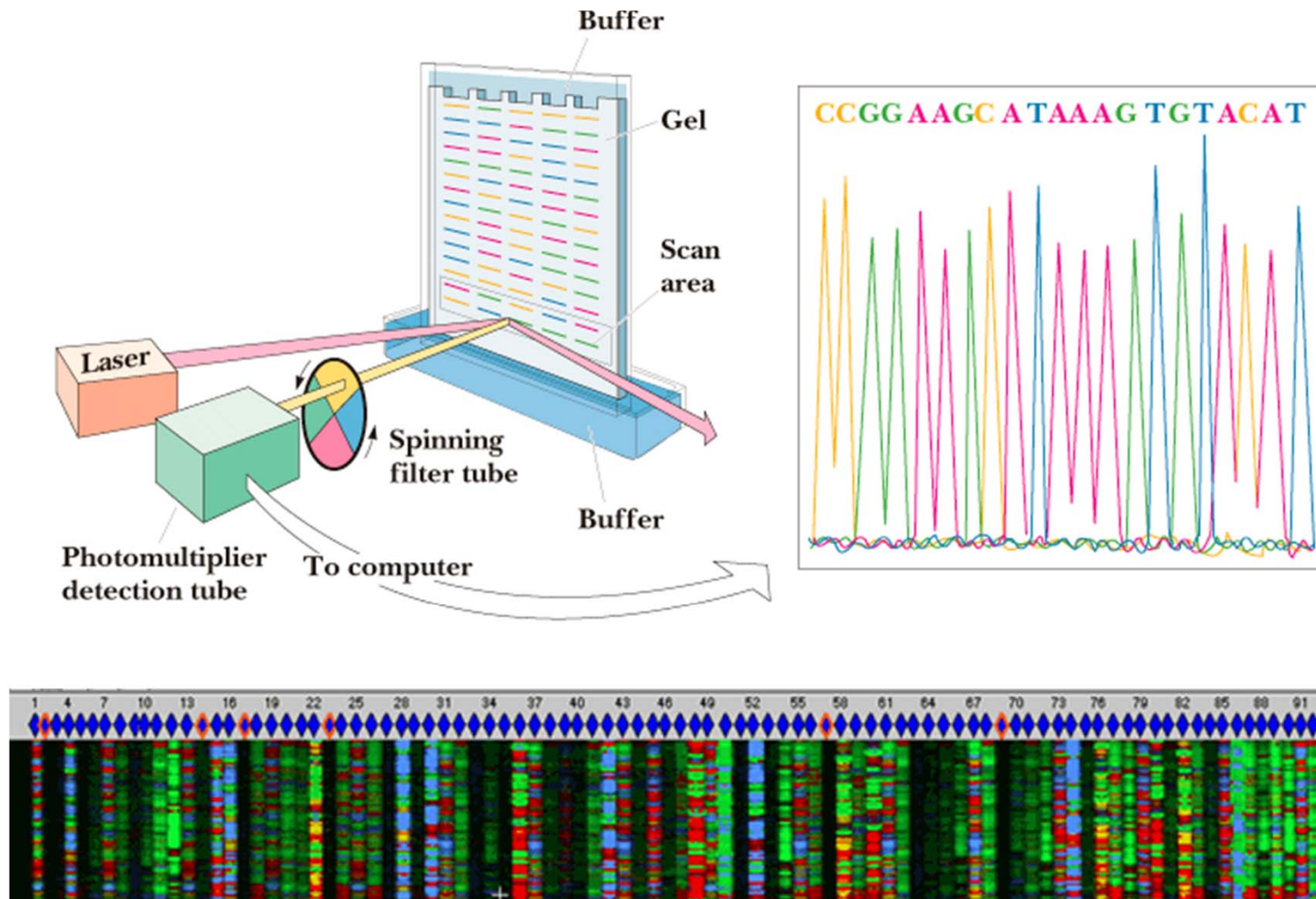
**Frederick Sanger**

"for their contributions concerning the determination of base sequences in nucleic acids"



Sanger is responsible for the first complete determination of the sequence of a DNA molecule. He has established the sequence of the 5375 building blocks in DNA from a bacterial virus called phi-X174.

Sequence investigations with the methods of Gilbert and Sanger together with the recombinant-DNA technique make excellent tools for continued investigations of the structure and function of the genetic material.



**1986**

Avtomatizacija določanja nukleotidnih zaporedij in uporaba fluorescentnih barvil

Leroy Hood in Mike Hunkapiller



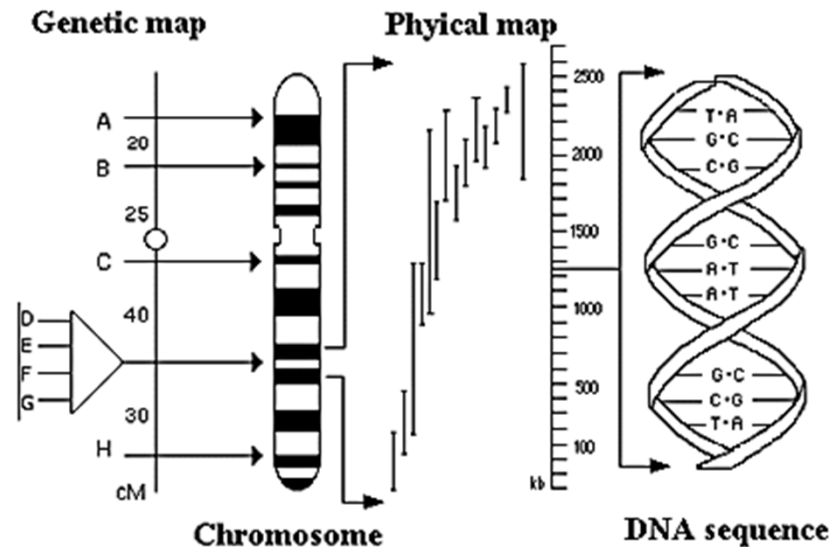
**Genomics** is the study of an organism's genome and the use of the genes. It deals with the systematic use of genome information, associated with other data, to provide answers in biology, medicine, and industry.

Genomics has the potential of offering new therapeutic methods for the treatment of some diseases, as well as new diagnostic methods. Other applications are in the food and agriculture sectors. The major tools and methods related to genomics are bioinformatics, genetic analysis, measurement of gene expression, and determination of gene function.

In biology the **genome** of an organism is the whole hereditary information of an organism that is encoded in the DNA (or, for some viruses, RNA). This includes both the genes and the non-coding sequences. The term was first coined, in 1920, by Hans Winkler, Professor of Botany at the University of Hamburg.

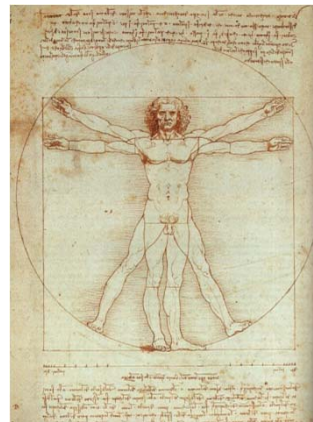
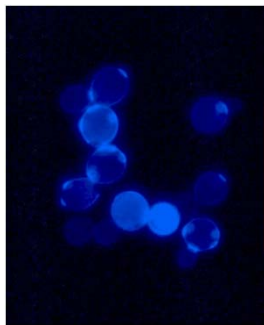
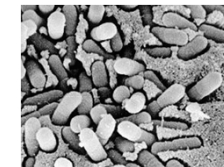
## Wikipedia

1986 revija GENOMICS  
(Roderick Thomas)





	Število baznih parov $\times 10^6$	Število genov	Določen
Bakteriofag $\phi$ X174	0.005	10	1977
<i>Mycoplasma genitalium</i>	0.58	483	1995
<i>Hemophilus influenzae</i>	1.83	1738	1995
<i>M. tuberculosis</i>	4.41	3959	1998
<i>Escherichia coli</i>	4.6	4377	1997
<i>Saccharomyces cerevisiae</i>	12.00	5885	1996
<i>Caenorhabditis elegans</i>	95.50	19.820	1998
<i>Drosophila melanogaster</i>	180.00	13.601	2000
<i>Arabidopsis thaliana</i>	117.00	25.498	2000
Človek	3300.00	$\approx 34.000$	2001



# Hierarchical shotgun sequencing

Genomic DNA



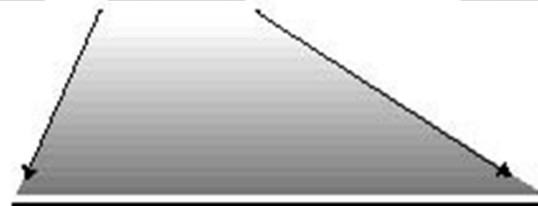
BAC library



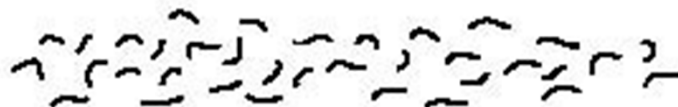
Organized mapped large clone contigs



BAC to be sequenced



Shotgun clones



Shotgun sequence

```
...ACCGTAAATGGGCTGATCATGCTTAAA  
TGATCATGCTTAAACCCTGTGCATCCTACTG...
```

Assembly

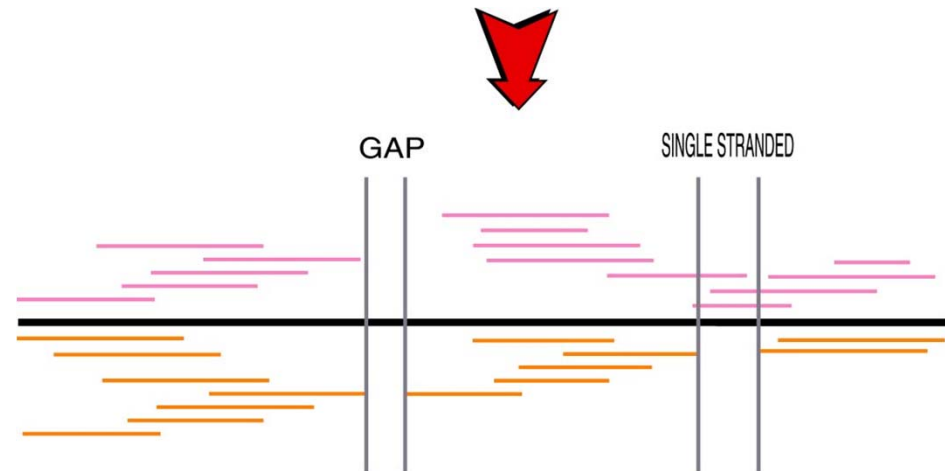
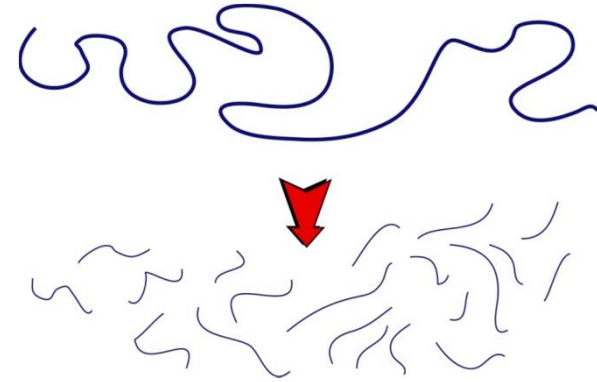
```
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...
```



300 ABI PRISM® 3700 DNA Analyzers, Applied Biosystems



800 povezanih Compaq Alpha-based 64-bit postaj, vsaka sposobna več kot 250 bilijonov primerjav zaporedij na uro.



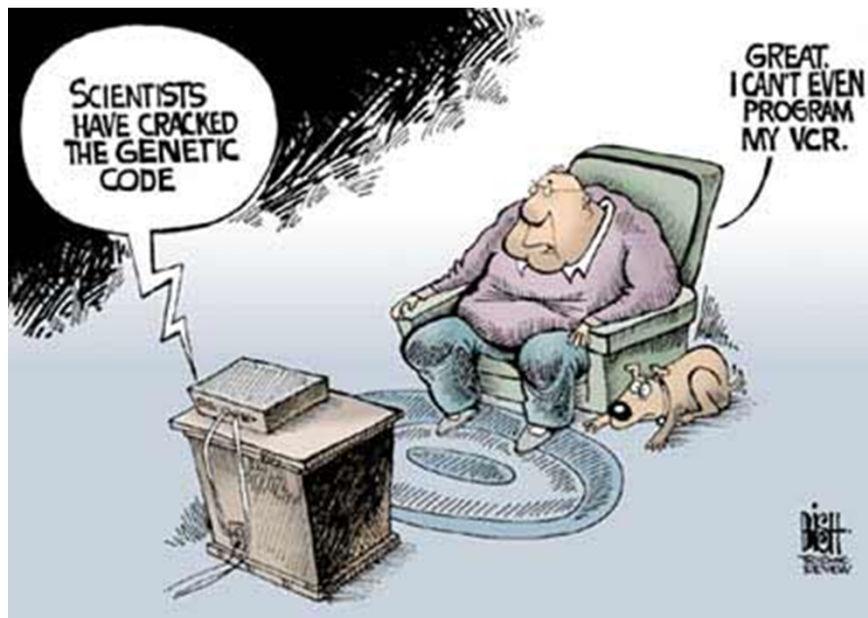
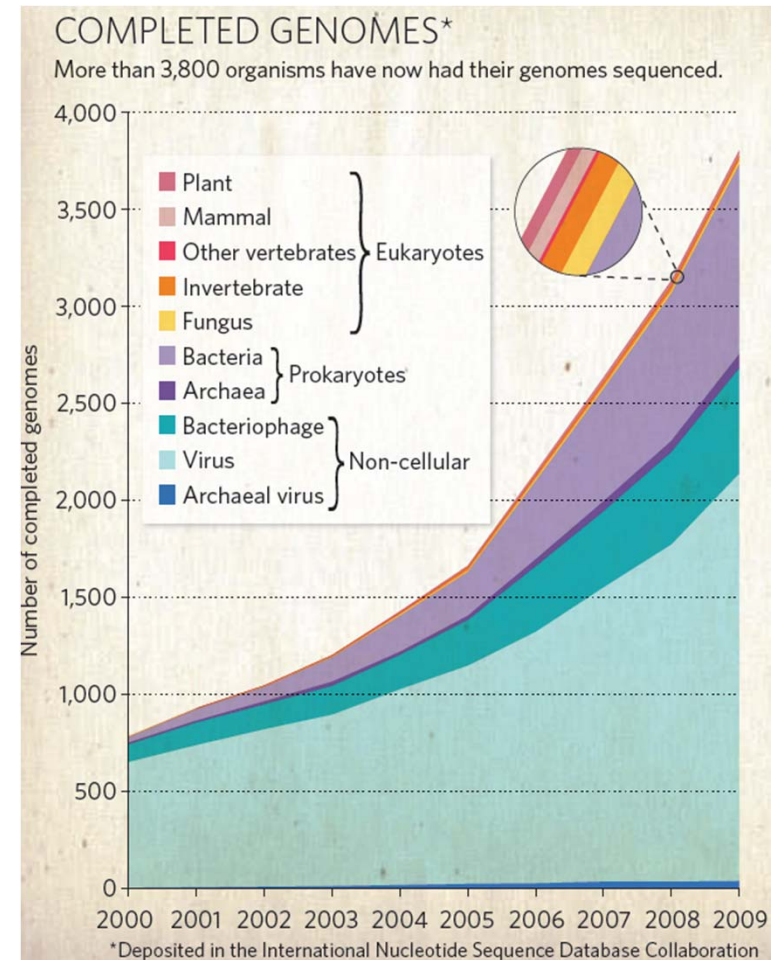
.....CCCTGGAAGCACAGGAATTTTTGTGTGGGAAAATCTAGAGGAGCTGCAACTGGAGCAAAG  
 AGTGGTGTAGTTCATTATCGTTATGGAGACAAGATACTTAATTTAATGGCATCCATTCCATATGATTGGAAAG  
 AACTTGTATCGTGGTGTAGTTCATTATCGTTATGGAGACAAGATACTTAATTTAATGGCATCCATTCCATATG  
 ATTGAAAGAACTTGTATCGTGGTGTAGTTCATTATCGTTATGGAGACAAGATACTTAATTTAATGGCATCCA

“The genome  
 revolution is only  
 just beginning.”

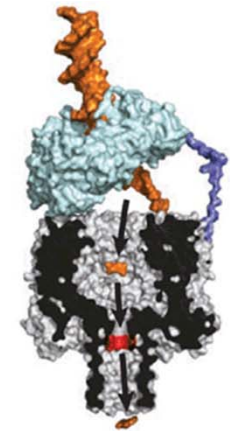
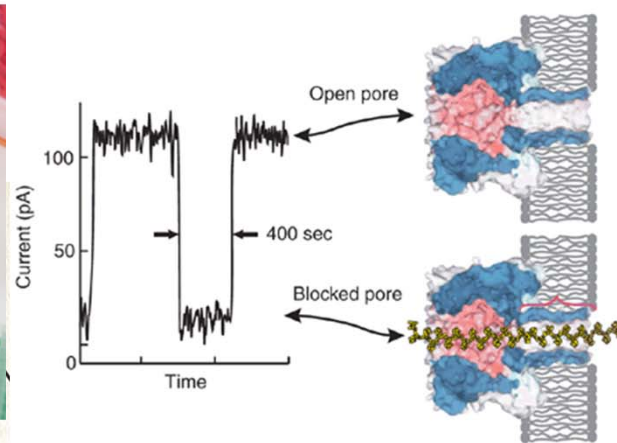
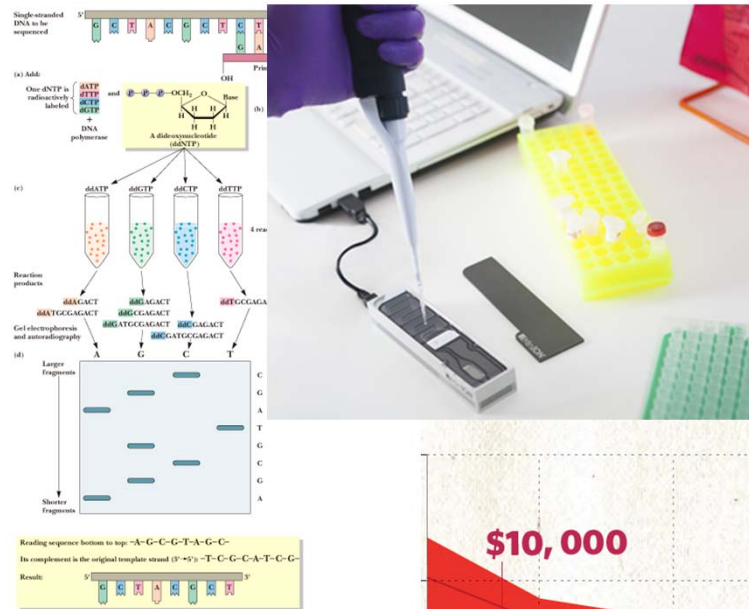
Venter JC (2010) Nature 464, 676-677.

“My students can gather certain types of  
 experimental data 1.000 and even 10.000  
 times faster than I could 40 years ago.”

Weinberg RA (2010) Nature 464, 678.



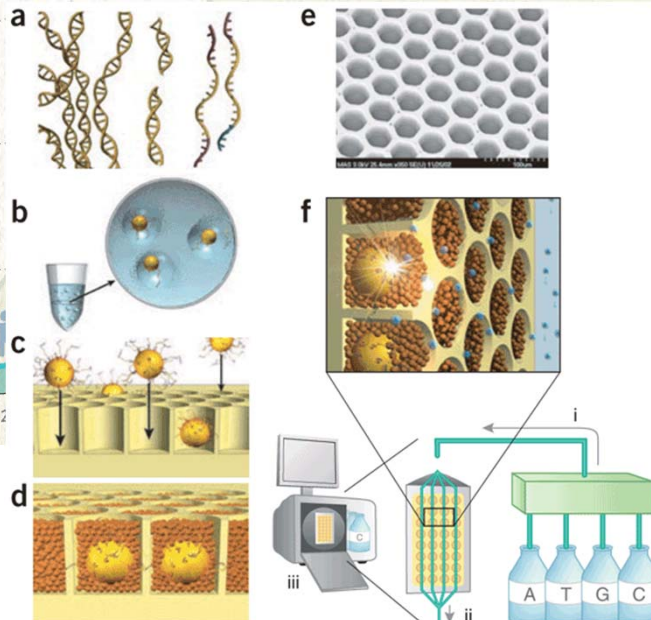
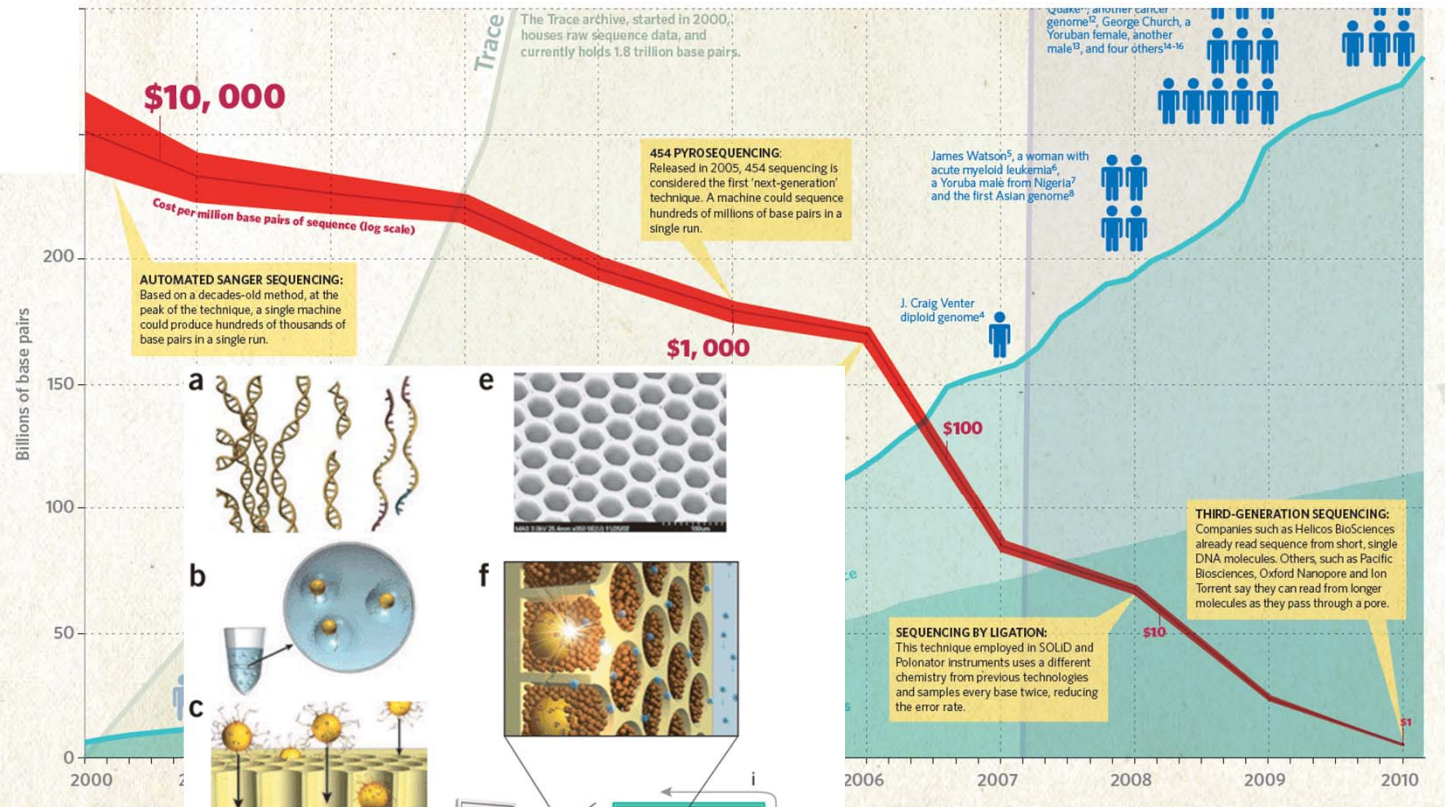
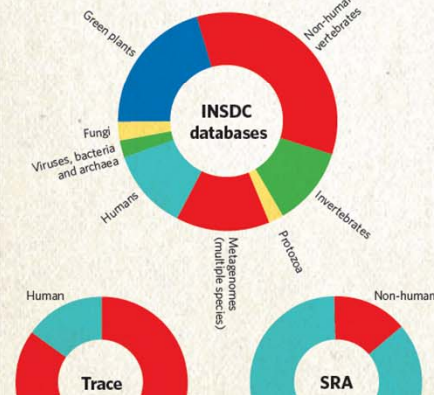
podatki, podatki, podatki, podatki, podatki, podatki,  
 podatki, podatki, podatki, podatki, podatki, podatki,  
 podatki, podatki, podatki, podatki, podatki, podatki,



roughly every 18 months. But this number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace archive and Sequence Read Archive (SRA). See Editorial, page 649, and human genome special at [www.nature.com/humangenome](http://www.nature.com/humangenome)

### DNA SEQUENCES BY TAXONOMY

International Nucleotide Sequence Database Collaboration: The main repositories of 'finished' sequence span a wide range of organisms, representing the many priorities of scientists worldwide.



# Genomika po 2000

2001 Osnutek človeškega genoma

2004 Človeški genom dokončan

## *Metagenomika sargaškega morja*

2006 *Neandertalčeva DNA*

2007 Venterjev diploidni genom

2008 J. Watson, bolnica z akutno mieloidno levkemijo, Yoruba, genom azijsca

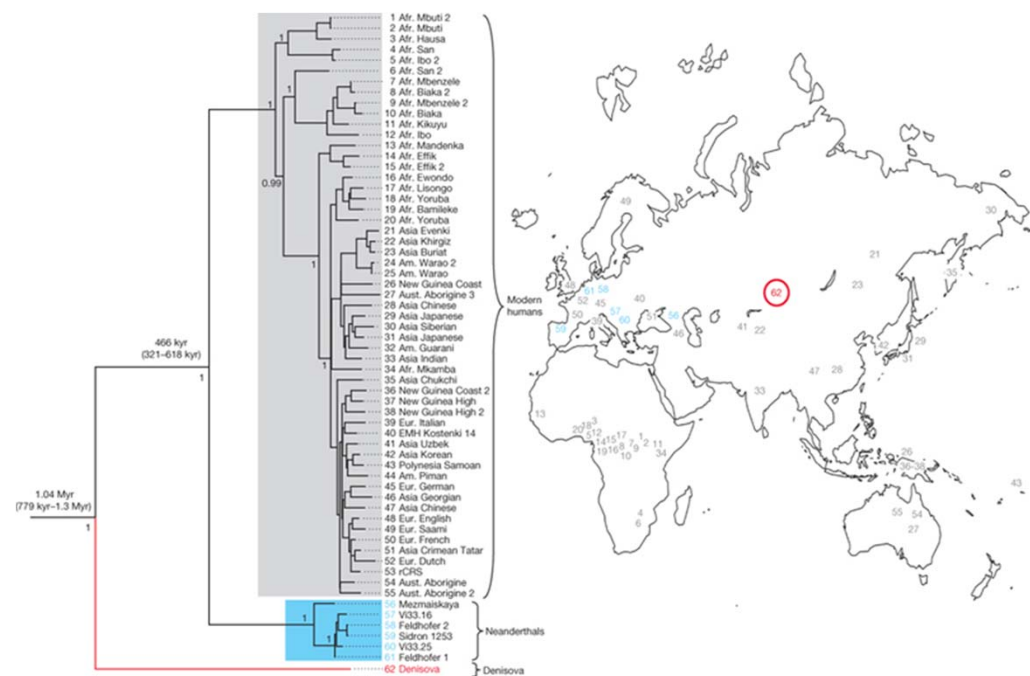
## *Genom mamuta*

2009 Dva genoma korejcev, genom bolnika z rakom, še osem posameznih genomov

2010 Gliomska celična linija, **Inuk**, še sedem posameznih genomov

*3800 organizmov z znanim genomom*

## *Nov predstavnik rodu Homo*

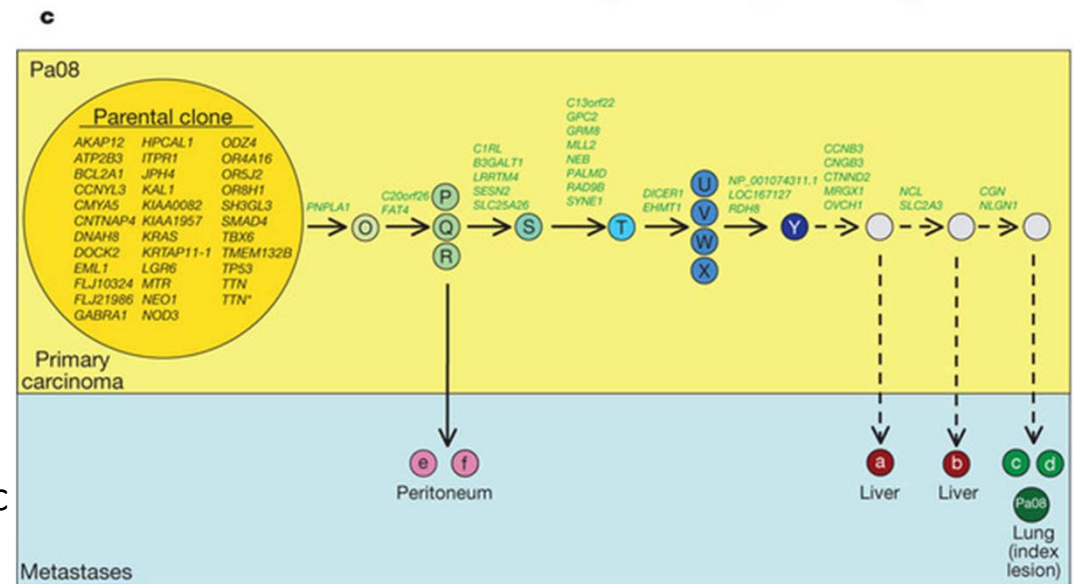
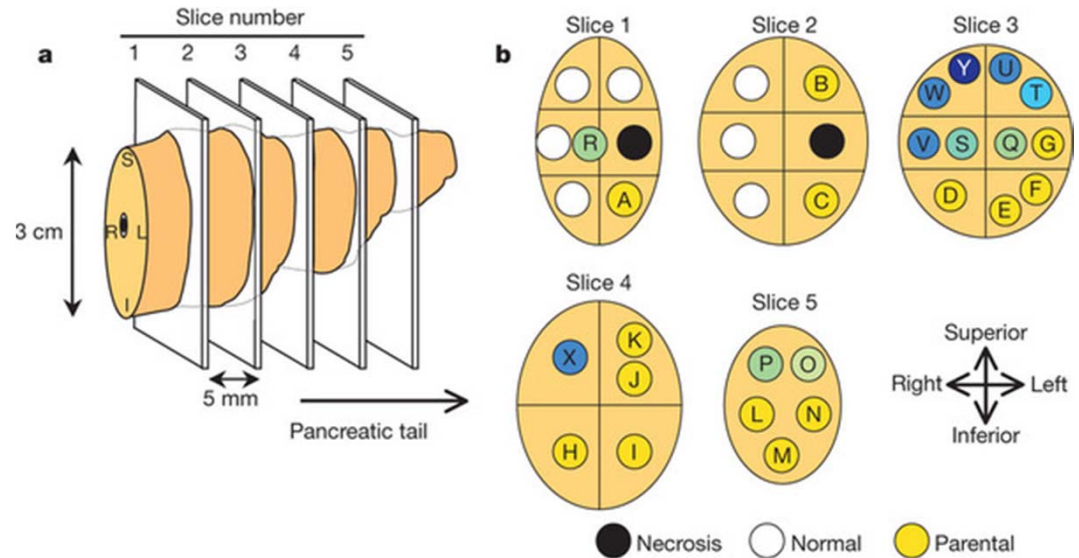
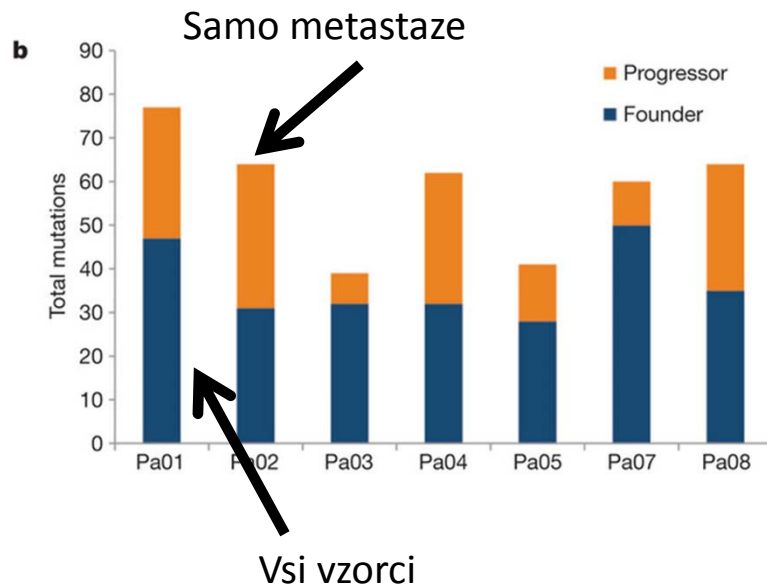


# Rak trebušne slinavke

7 pacientov

Metastaze v jetrih, pljučih, peritoneju

Določanje zaporedja eksoma 20661 genov



## Kombinacija

- Mikromrež
- Traskriptoma
- Določanja dela zaporedja genoma

Jones et al. (2008) Science 321, 1801-1806

Yachida et al. (2010) Nature 467, 1114-1117

Distant metastasis occurs late during the genetic evolution of pancreatic cancer

## Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,<sup>1\*</sup> Karin Remington,<sup>1</sup> John Heidelberg,<sup>3</sup> Aaron L. Halpern,<sup>2</sup> Doug Rusch,<sup>2</sup> Jonathan A. Eisen,<sup>3</sup> Dongying Wu,<sup>3</sup> Ian Paulsen,<sup>3</sup> Karen E. Nelson,<sup>3</sup> William Nelson,<sup>3</sup> Derrick E. Fouts,<sup>3</sup> Samuel Levy,<sup>2</sup> Anthony H. Knap,<sup>6</sup> Michael W. Lomas,<sup>6</sup> Ken Nealson,<sup>5</sup> Owen White,<sup>3</sup> Jeremy Peterson,<sup>3</sup> Jeff Hoffman,<sup>1</sup> Rachel Parsons,<sup>6</sup> Holly Baden-Tillson,<sup>1</sup> Cynthia Pfannkoch,<sup>1</sup> Yu-Hui Rogers,<sup>4</sup> Hamilton O. Smith<sup>1</sup>

<sup>1</sup>The Institute for Biological Energy Alternatives, <sup>2</sup>The Center for the Advancement of Genomics, 1901 Research Boulevard, Rockville, MD 20850, USA. <sup>3</sup>The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.

<sup>4</sup>The J. Craig Venter Science Foundation Joint Technology Center, 5 Research Place, Rockville, MD 20850, USA. <sup>5</sup>University of Southern California, 223 Science Hall, Los Angeles, CA 90089-0740, USA. <sup>6</sup>Bermuda Biological Station for Research, Inc., 17 Biological Lane, St George GE 01, Bermuda.

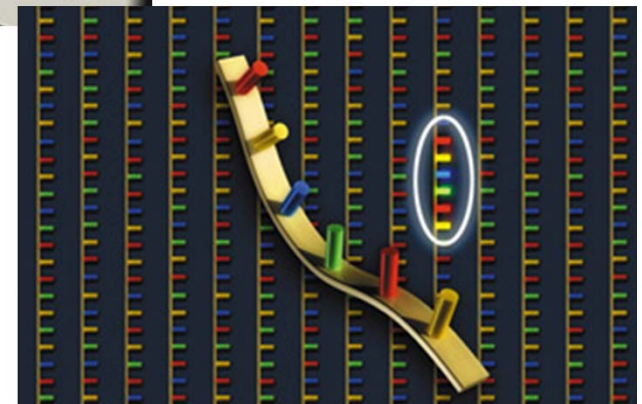
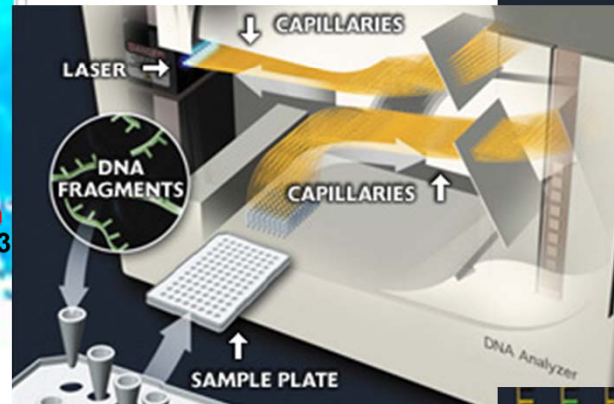
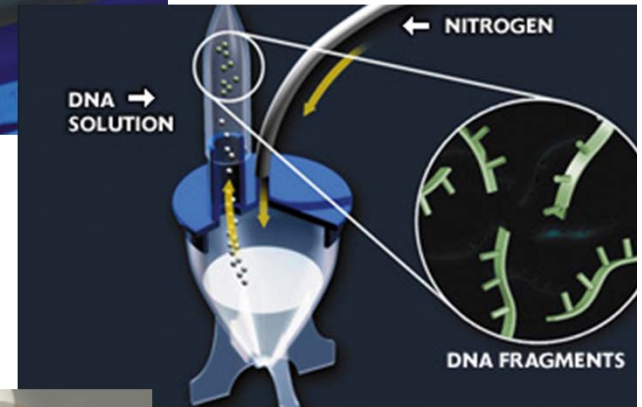
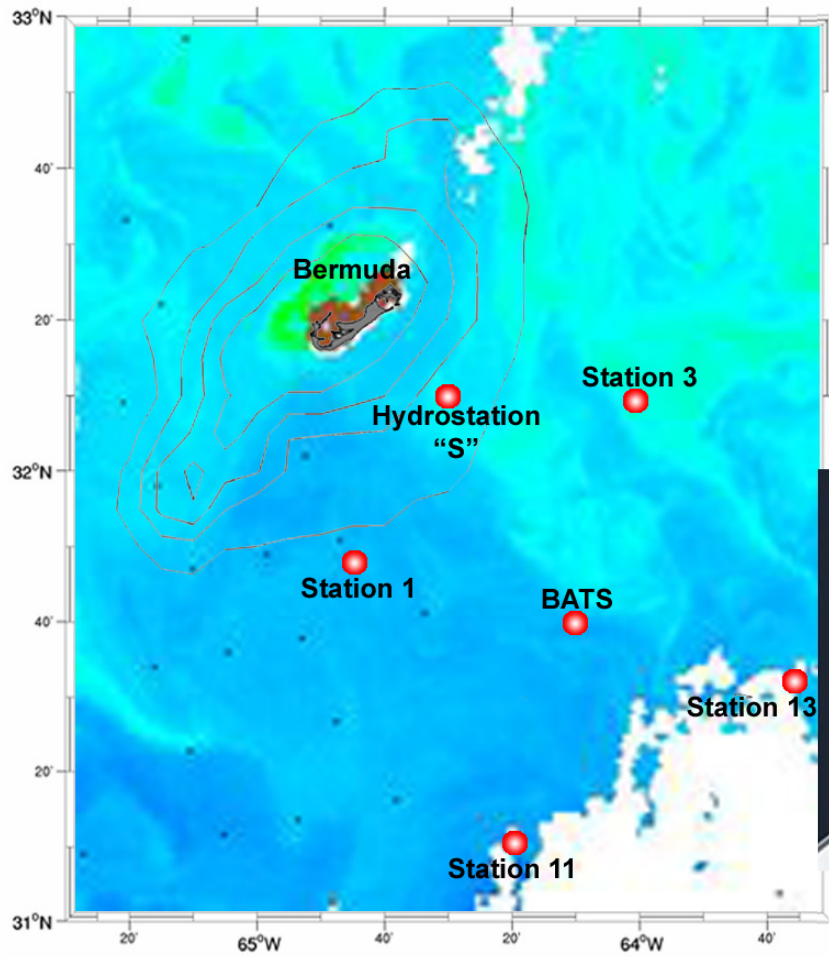
\*To whom correspondence should be addressed. E-mail: jcventer@tcag.org

**We have applied “whole genome shotgun sequencing” to microbial populations collected *en mass* on tangential flow and impact filters from sea water samples collected from the Sargasso Sea near Bermuda. A total of 1.045 billion basepairs of non-redundant sequence was generated, annotated and analyzed to elucidate the gene content, and diversity and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1800 genomic species based on sequence relatedness, including 148 novel bacterial phylotypes. We have identified over 1.2 million new genes represented in these samples, including more than 782 new rhodopsin-like photoreceptors. Variation in species present and stoichiometry suggests substantial oceanic microbial diversity.**

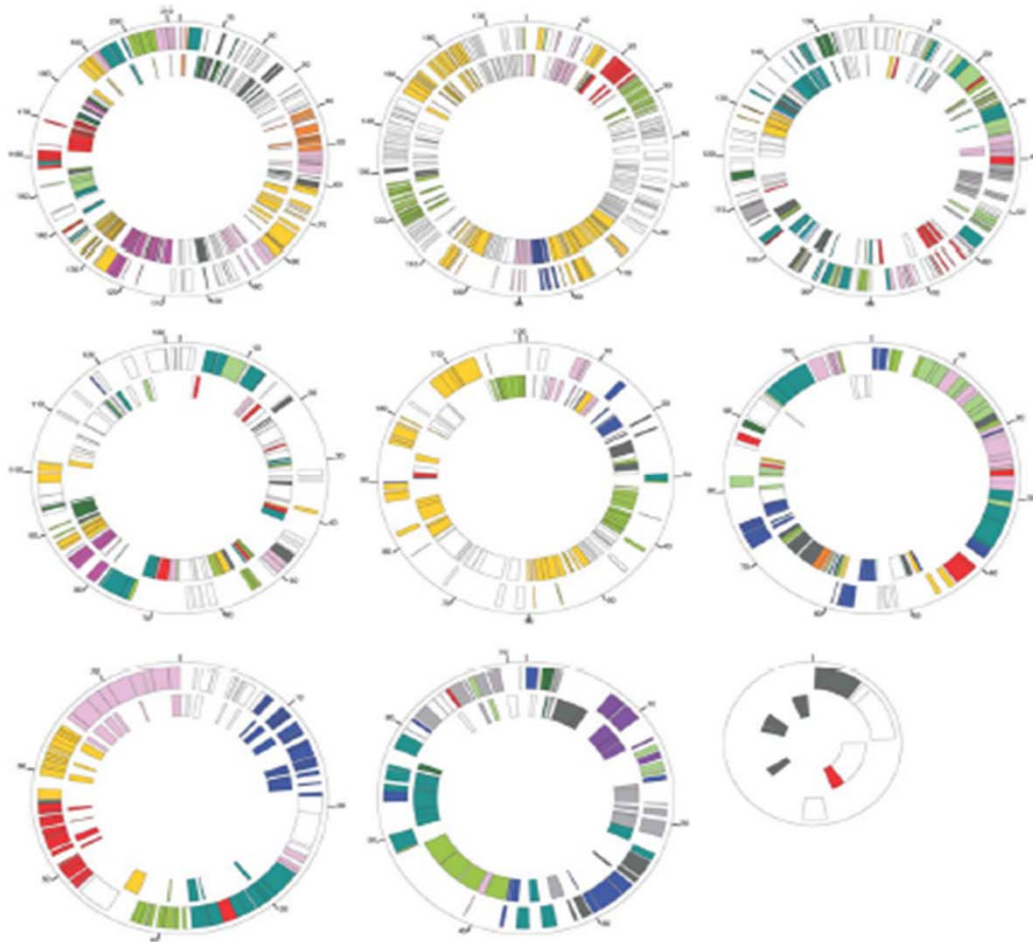
environment. Further, we concentrated on the genetic material captured on filters sized to isolate primarily microbial inhabitants of the environment, leaving detailed analysis of dissolved DNA and viral particles on one end of the size spectrum, and eukaryotic inhabitants on the other, for subsequent studies.

**The Sargasso Sea.** The northwest Sargasso Sea, at the Bermuda Atlantic Time-series Study site (BATS), is one of the best-studied and arguably most well-characterized regions of the global ocean. The Gulf Stream represents the western and northern boundaries of this region and provides a strong physical boundary separating the low nutrient, oligotrophic, openocean from the more nutrient-rich waters of the U.S. continental shelf. The Sargasso Sea has been intensively studied as part of the 50-year time-series of ocean physics and biogeochemistry (3, 4), and provides an opportunity for





<http://www.sorcerer2expedition.org/version1/HTML/main.htm>



Environmental genome shotgun sequencing of the Sargasso Sea. Venter, J.C. et al. 2004. Science 304:66-74.

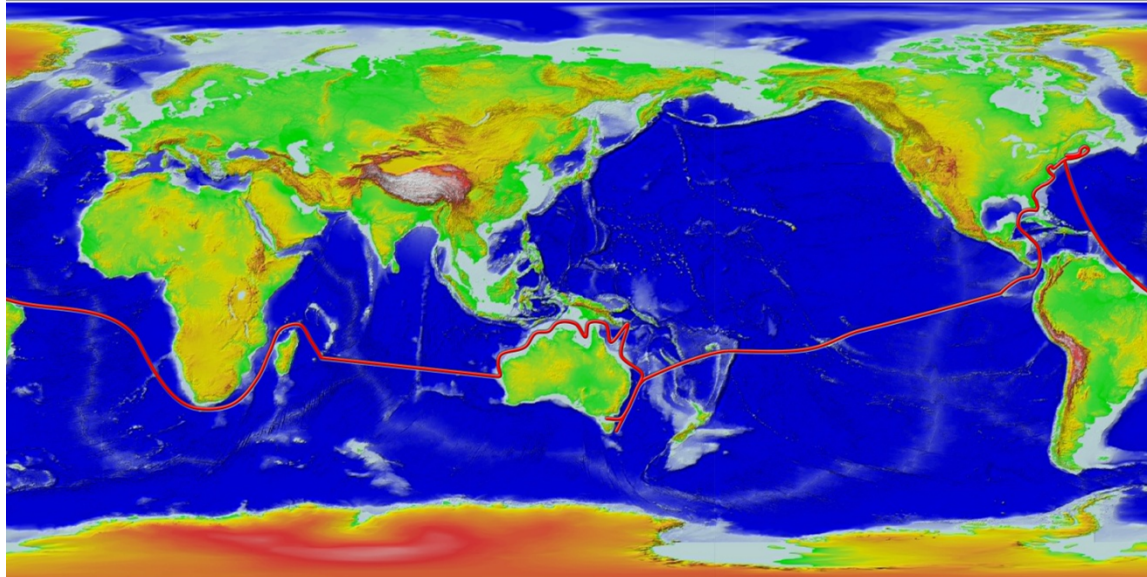
- 1.66 milijona zaporedij,
- ki so predstavljala več kot 1.2 milijona novih genov
- z računalniško analizo so uspeli sestaviti 12 kompletnih bakterijskih genomov.
- iz podatkov ocenjujejo, da je v vzorcu Sargaškega morja vsaj 1800 bakterijskih vrst in preko 70.000 povsem novih genov.
- od teh je pet vrst zastopanih v večjem številu.

# Katalog vseh genov na zemeljski obli?

<http://www.sorcerer2expedition.org/version1/HTML/main.htm>



## Sorcerer II Expedition World Sampling Route



J. Craig Venter  
INSTITUTE


**Sargasso samples generated over 1 billion bp of ocean genomes and 1.3 million new genes**

**Sampling from 300 additional sites can provide hundreds of billions of base pairs of DNA sequence and up to 1 billion new genes that will permit cost effective analysis of ocean microbial diversity and environmental impact**

**One time cost for DNA sequencing and database \$50-70 Million (\$15 million in hand)**



*This article is  
part of the  
Oceanic  
Metagenomics  
collection.*

 PLoS Biology | [www.plosbiology.org](http://www.plosbiology.org)

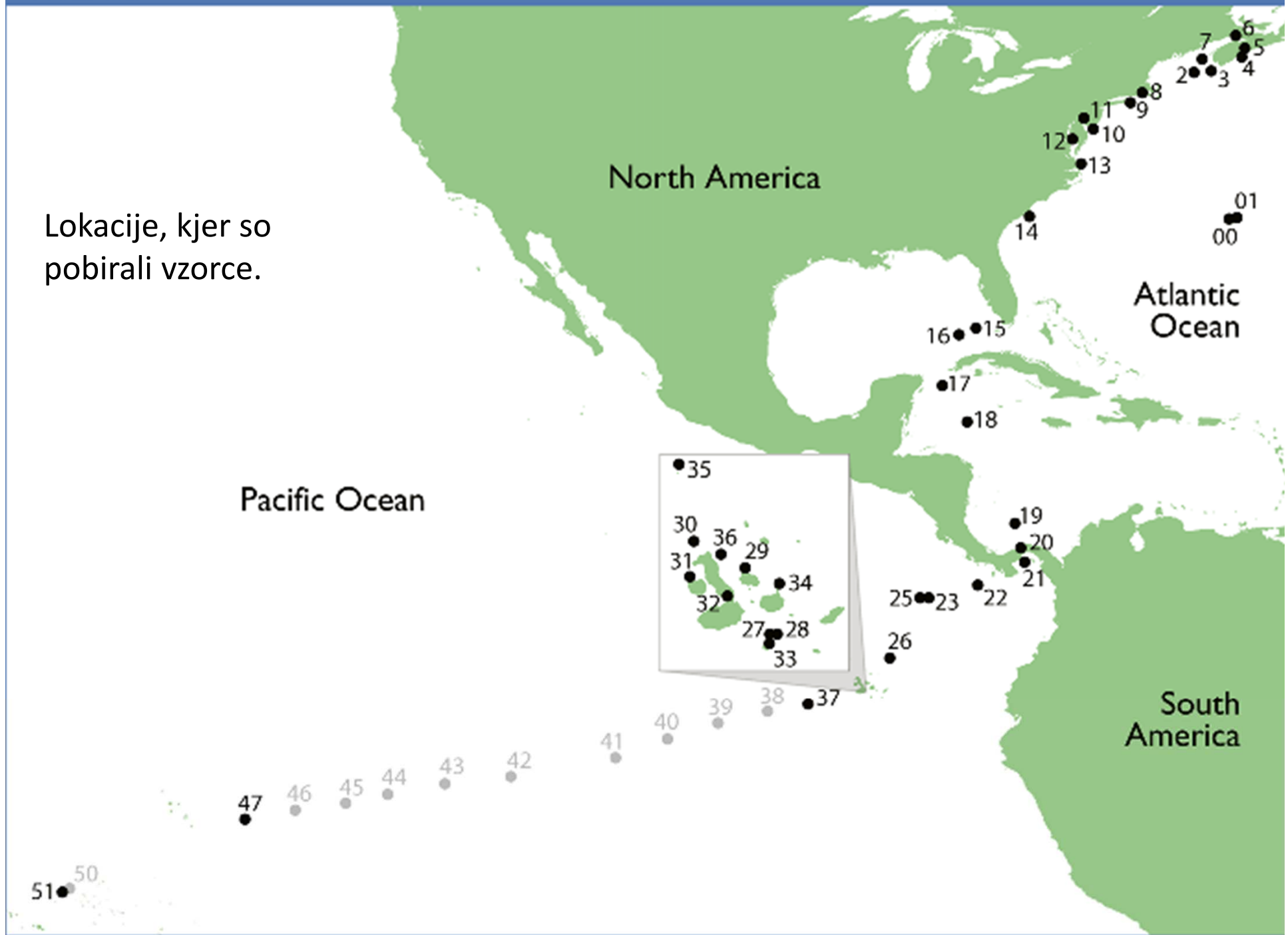
**The Sorcerer II Global Ocean Sampling expedition:  
northwest Atlantic through eastern tropical Pacific.**

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC.

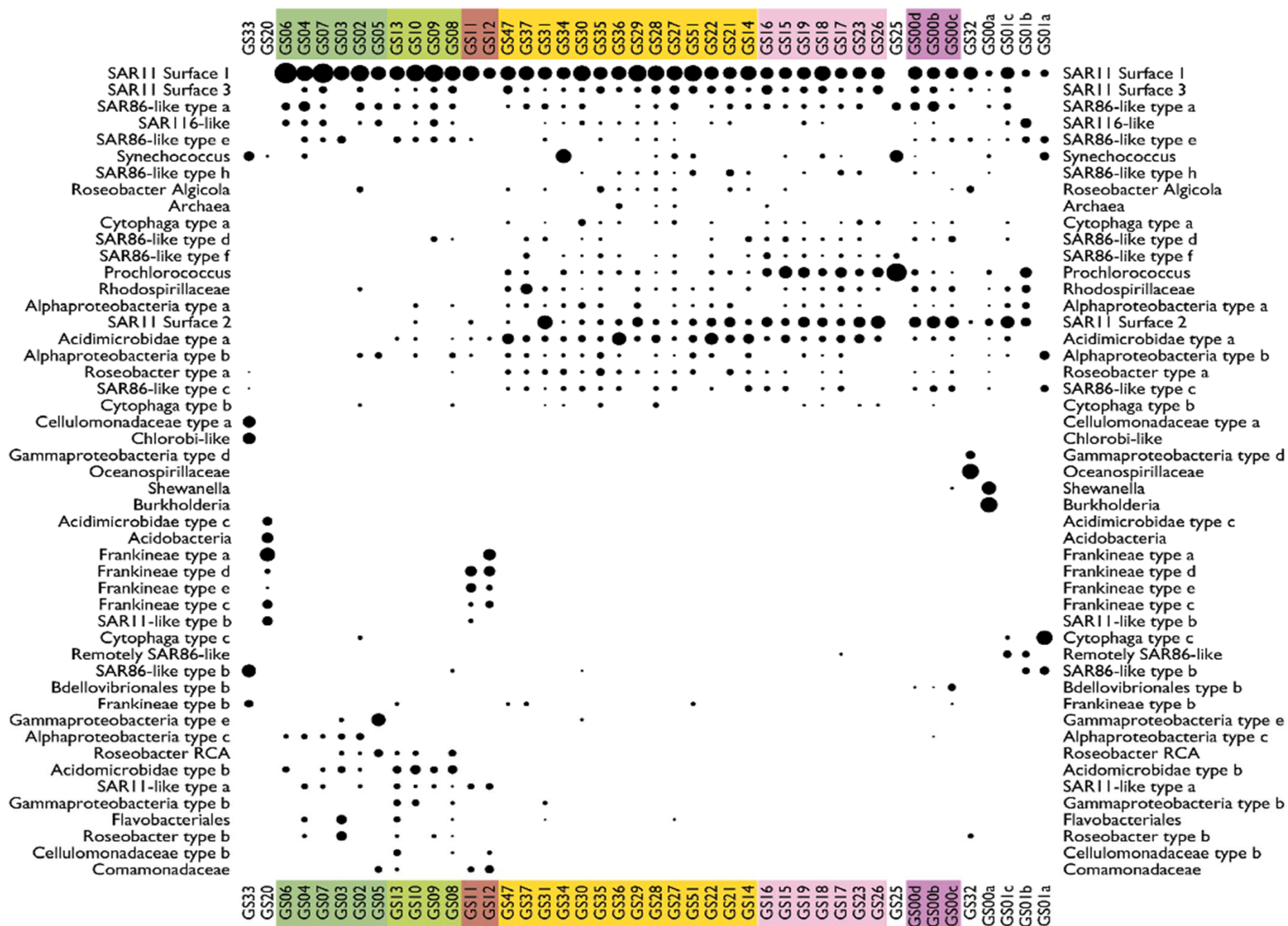
PLoS Biol. 2007 Mar;5(3):e77.

# **The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific**

Lokacije, kjer so pobirali vzorce.



▼ PRESENCE AND ABUNDANCE OF DOMINANT RIBOTYPES



# The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific

•Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biol 5: e77 doi:[10.1371/journal.pbio.0050077](https://doi.org/10.1371/journal.pbio.0050077).

•Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The *Sorcerer II* Global Ocean Sampling expedition: Expanding the universe of protein families. PLoS Biol 5: e16 doi:[10.1371/journal.pbio.0050016](https://doi.org/10.1371/journal.pbio.0050016).

•Kannan N, Taylor SS, Zhai Y, Venter JC, Manning G (2006) Structural and functional diversity of the microbial kinome. PLoS Biol 5: e17 doi:[10.1371/journal.pbio.0050017](https://doi.org/10.1371/journal.pbio.0050017).



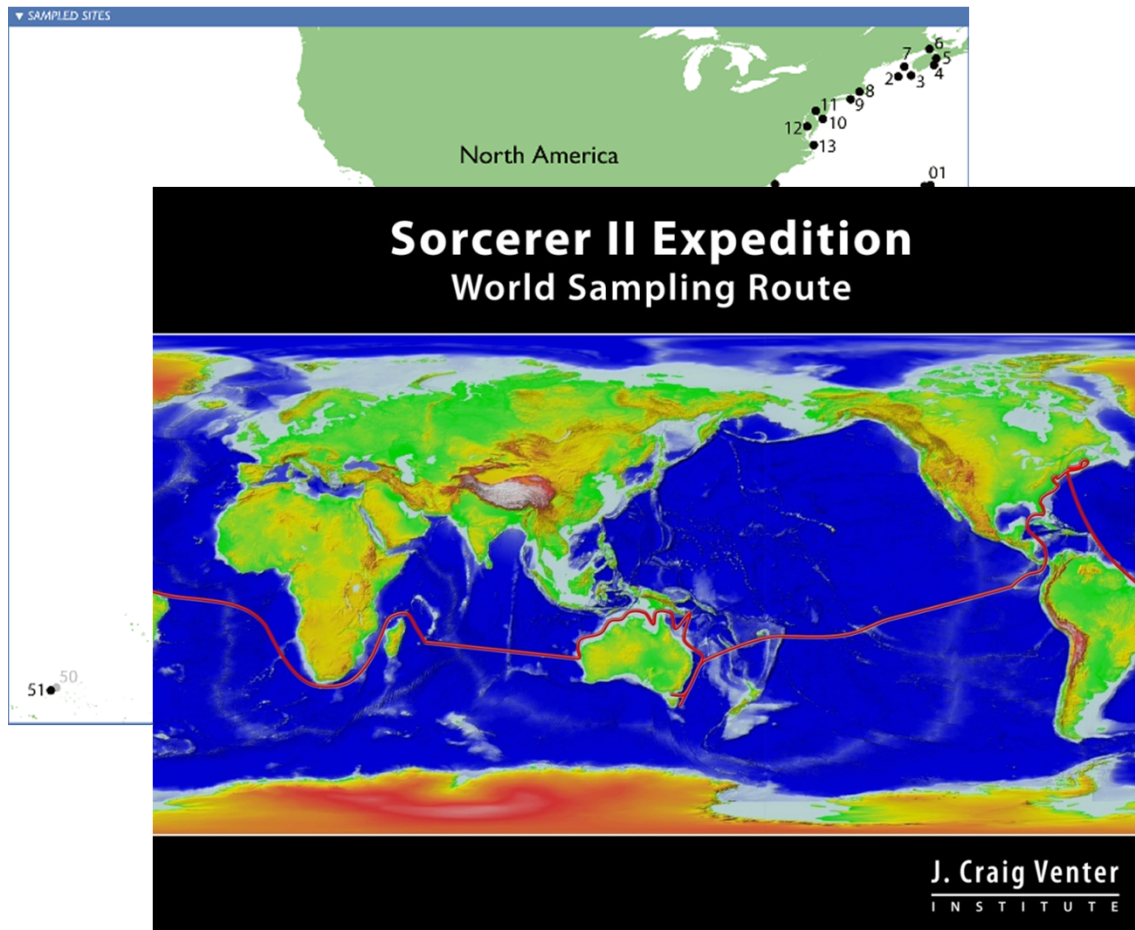
This article is part of the Oceanic Metagenomics collection.

PLoS Biology | www.plosbiology.org

33 strani

34 strani

11 strani  
skupaj 78



## Author Summary

The rapidly emerging field of metagenomics seeks to examine the genomic content of communities of organisms to understand their roles and interactions in an ecosystem. Given the wide-ranging roles microbes play in many ecosystems, metagenomics studies of microbial communities will reveal insights into protein families and their evolution. Because most microbes will not grow in the laboratory using current cultivation techniques, scientists have turned to cultivation-independent techniques to study microbial diversity. One such technique—shotgun sequencing—allows random sampling of DNA sequences to examine the genomic material present in a microbial community. We used shotgun sequencing to examine microbial communities in water samples collected by the *Sorcerer II* Global Ocean Sampling (GOS) expedition. Our analysis predicted more than six million proteins in the GOS data—nearly twice the number of proteins present in current databases. These predictions add tremendous diversity to known protein families and cover nearly all known prokaryotic protein families. Some of the predicted proteins had no similarity to any currently known proteins and therefore represent new families. A higher than expected fraction of these novel families is predicted to be of viral origin. We also found that several protein domains that were previously thought to be kingdom specific have GOS examples in other kingdoms. Our analysis opens the door for a multitude of follow-up protein family analyses and indicates that we are a long way from sampling all the protein families that exist in nature.

# BIOINFORMATIKA

Vmesnik med biologijo in računalništvom.

*Matematične, statistične in računalniške metode za reševanje bioloških problemov z uporabo DNA in proteinskih zaporedij in povezane informacije.*

*Organiziranje, analiza in distribucija biološke informacije pri opisovanju in reševanju bioloških problemov.*

*Zbiranje, organiziranje, shranjevanje in iskanje biološke informacije v podatkovnih zbirkah.*

***NI ISKANJE ČLANKOV IN BRSKANJE ZA REVIJAMI!!!***

