Machines Learning what makes Biology tick

Panagiotis Alexiou

CORE019 Pokroky a výzvy v moderní biologii (podzim 2021)

Gregor Mendel: The Friar Who Grew Peas

Gregor Mendel 1822-18

Rules of Classical Gen

Crossed and counted p

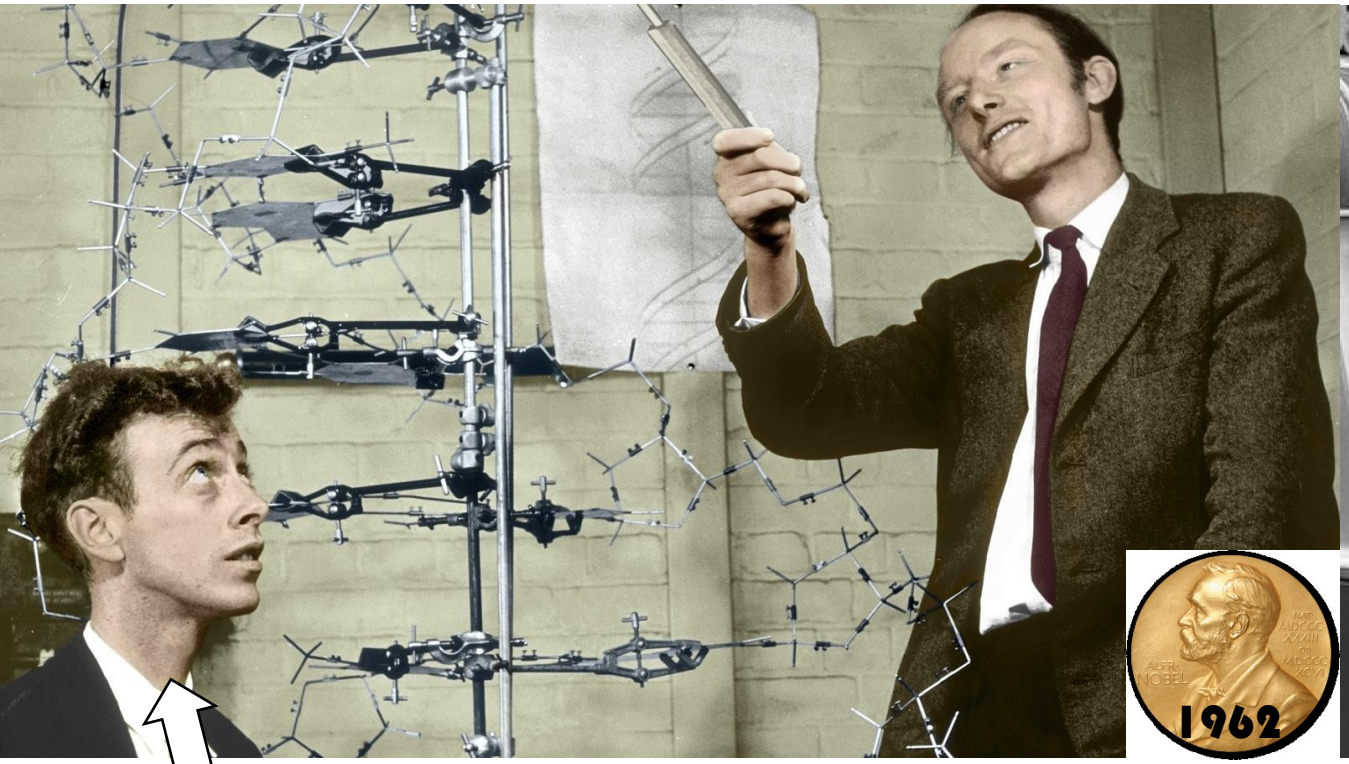
His findings forgotten
50 years

In the 1910s-30s new
disciplines emerged

Interdisciplinary resea

- Chemistry
- Physics
- Biology

- ➢ Genetics
- ➢ Biochemistry
- ➢ Biophysics



Francis Crick identified himself as a molecular biologist as a way of shortening his previous description of himself as "a mixture of a crystallographer, biophysicist, biochemist, and geneticist."

1950s – Discovery of DNA structure

Arthur Samuel of IBM developed a computer program for playing checkers. The program used a scoring function to assess moves, and learned from previous games.

1950s – First Machine Learning

Machine Learning

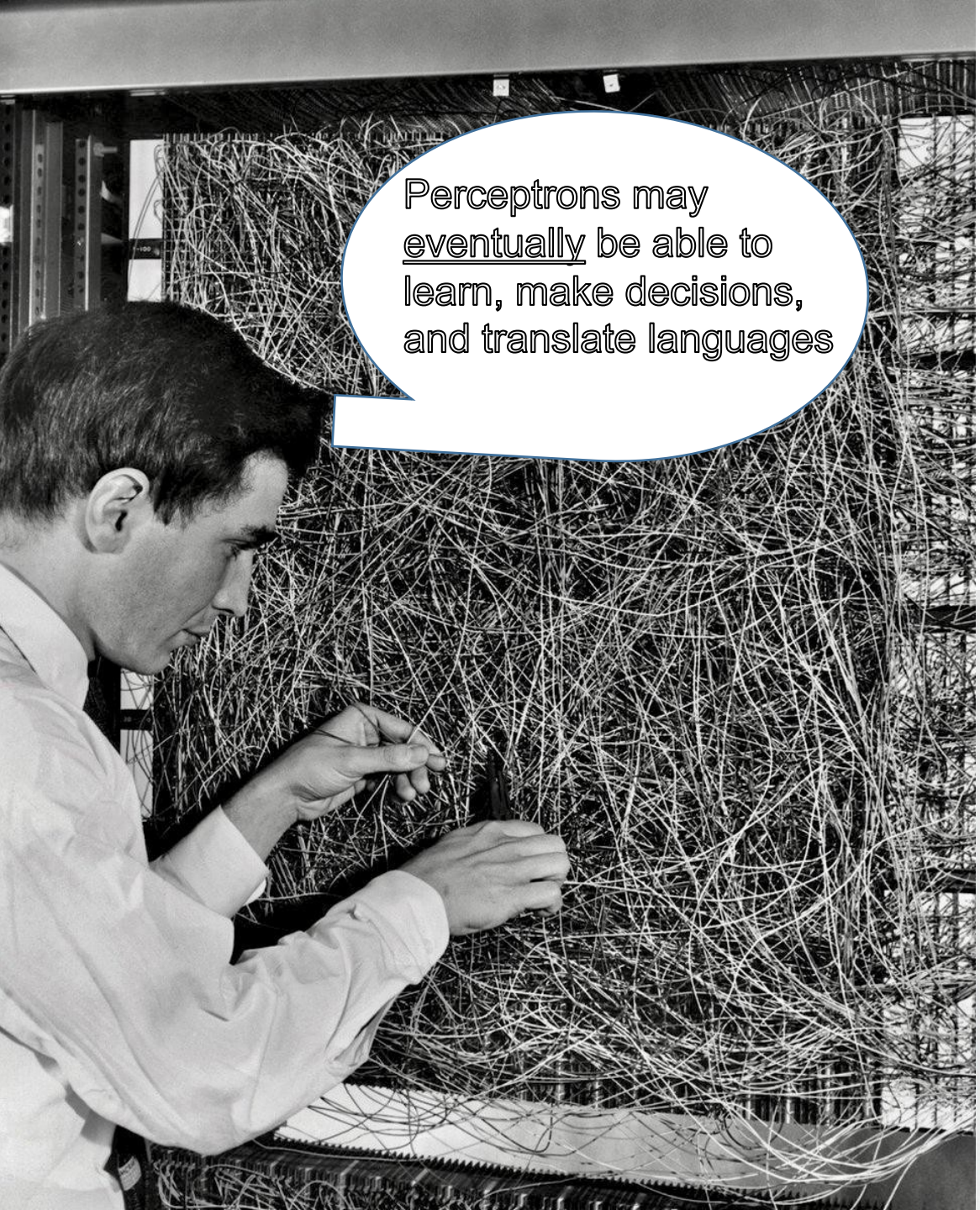
The use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.

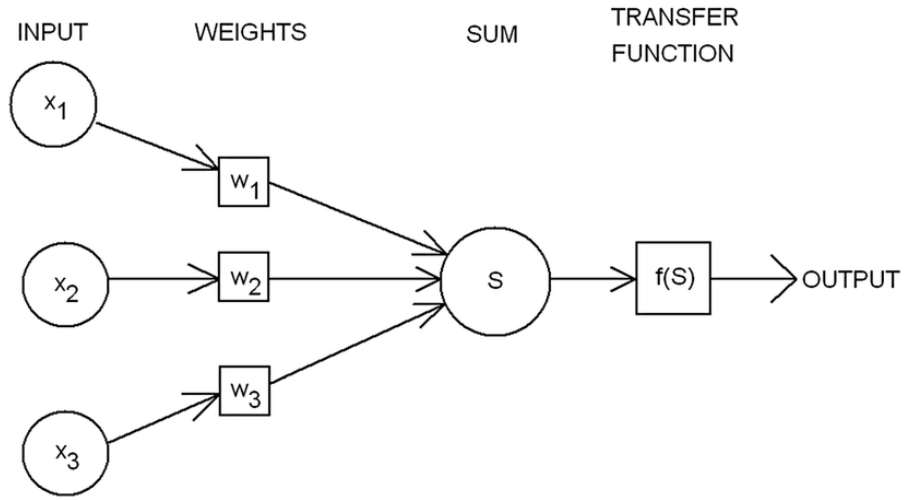
Machine Learning

Make

The use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data.

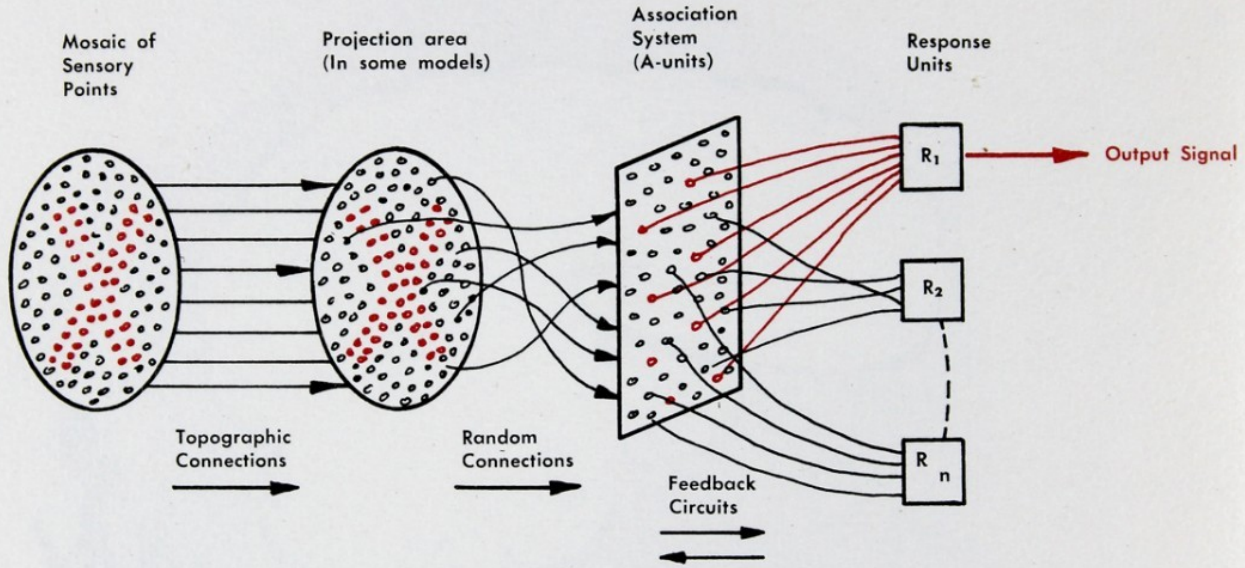
Make Computer Systems that learn from patterns in data

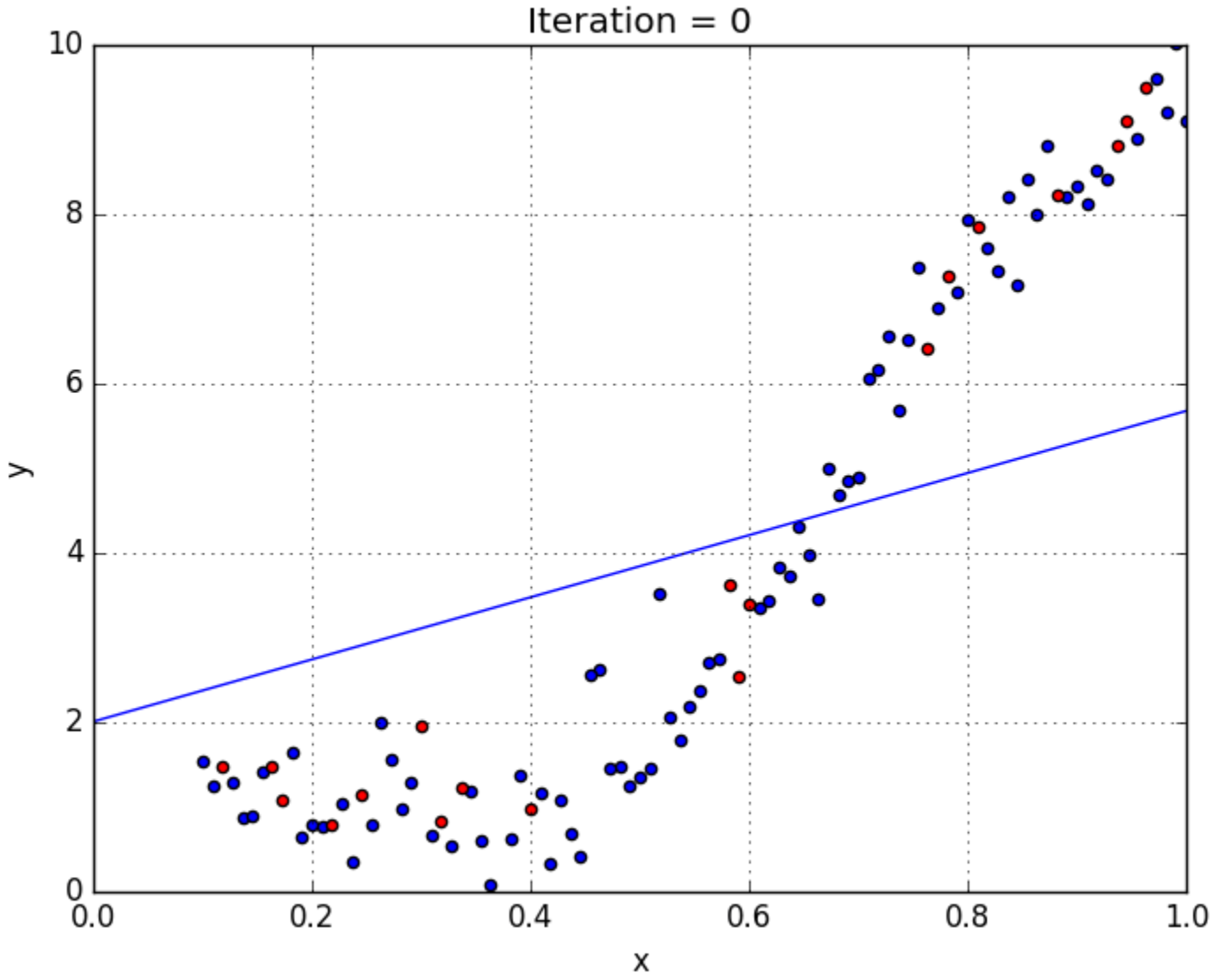


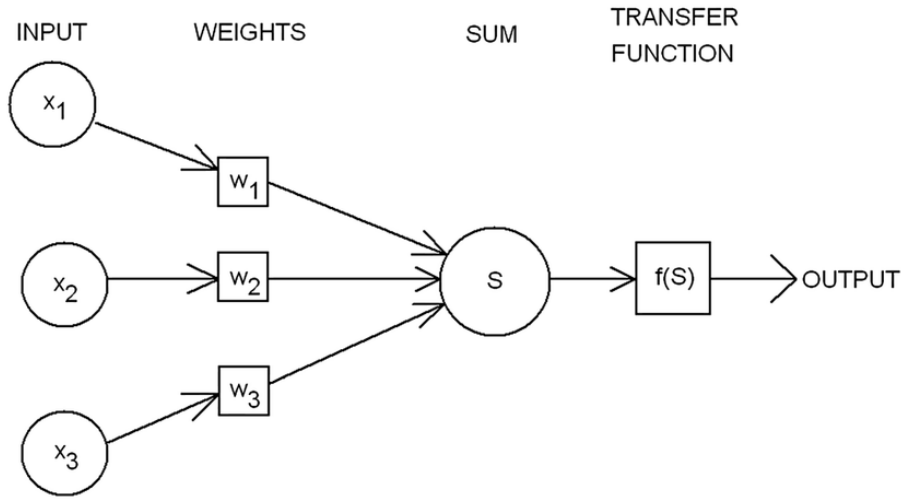


1950s – Mark I Perceptron
First Artificial Neural Networ

FIG. 1 — Organization of a biological brain. (Red areas indicate active cells, responding to the letter X.)



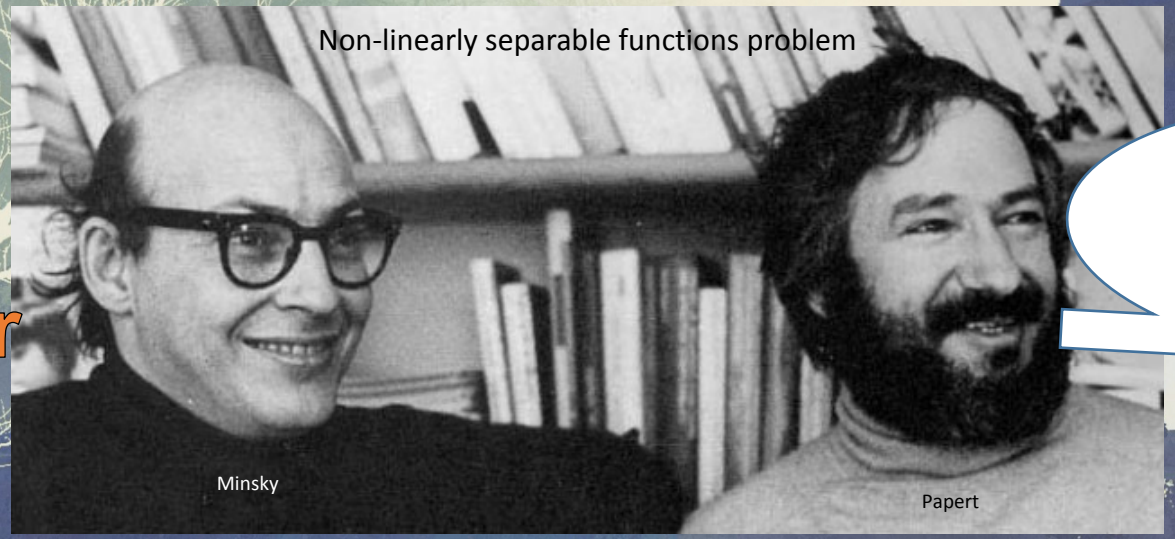


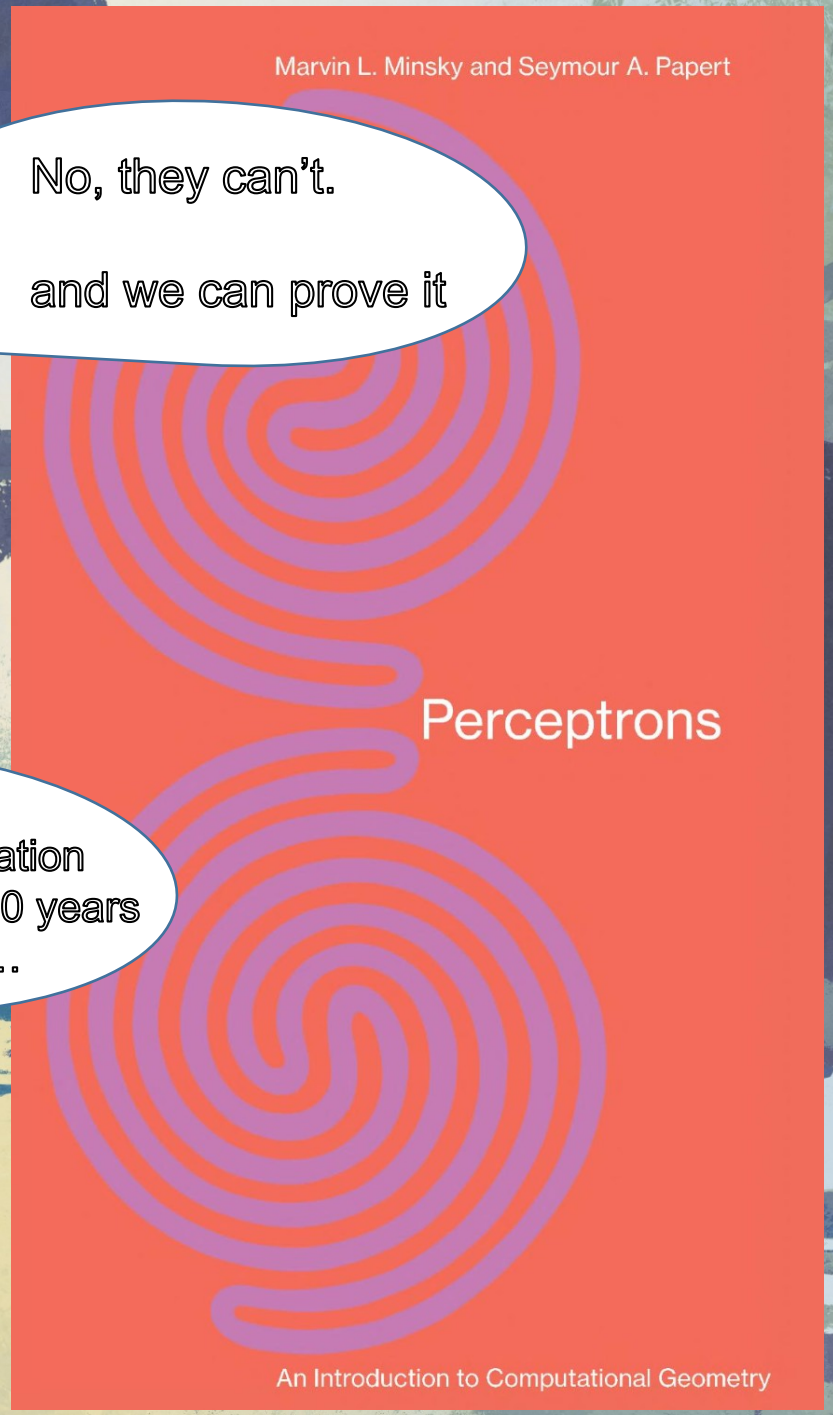


1960s:
The perceptron is able to solve simple problems such as linear regression.

Late 1960s – First AI Winter

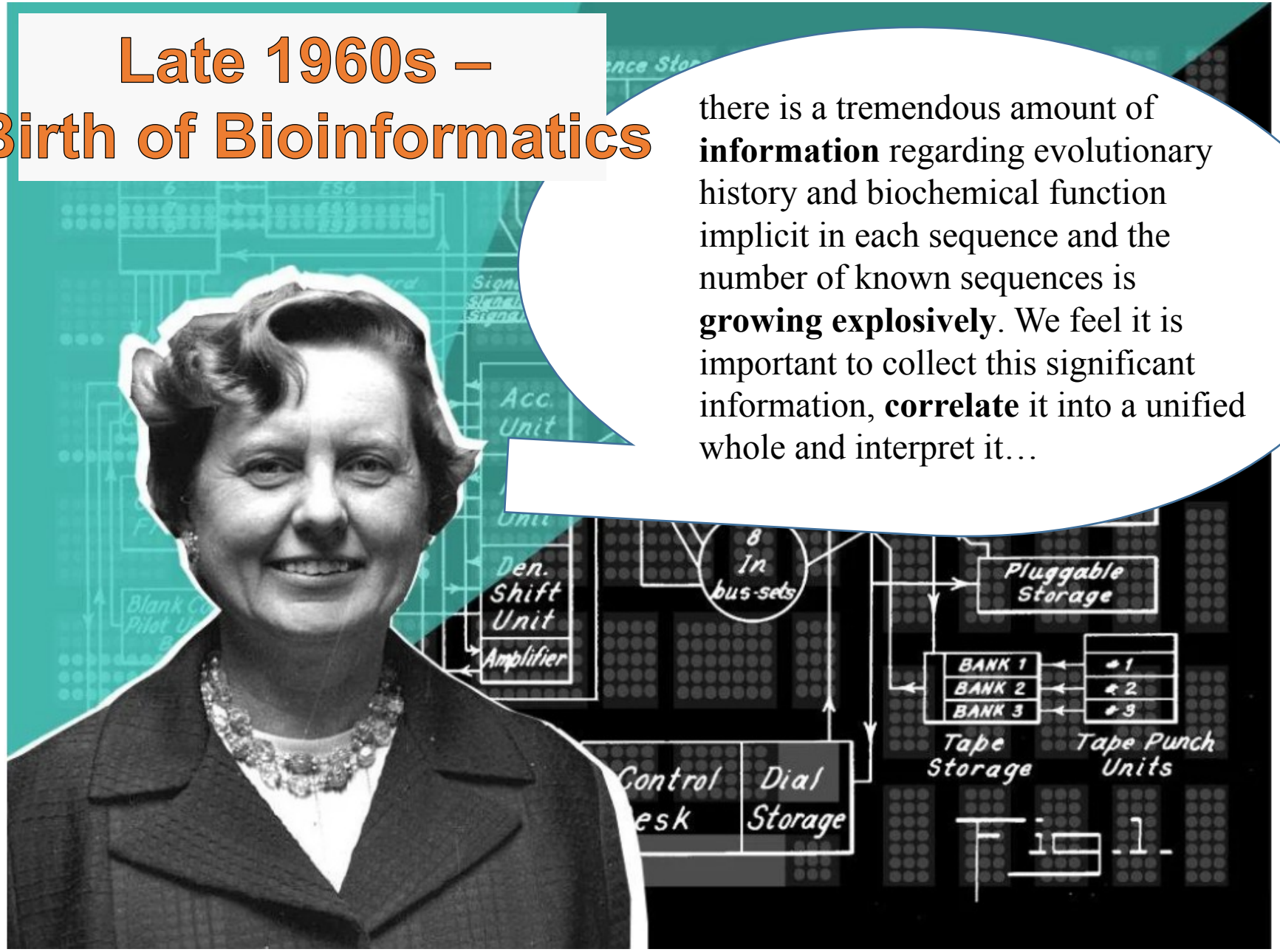
Non-linearly separable functions problem



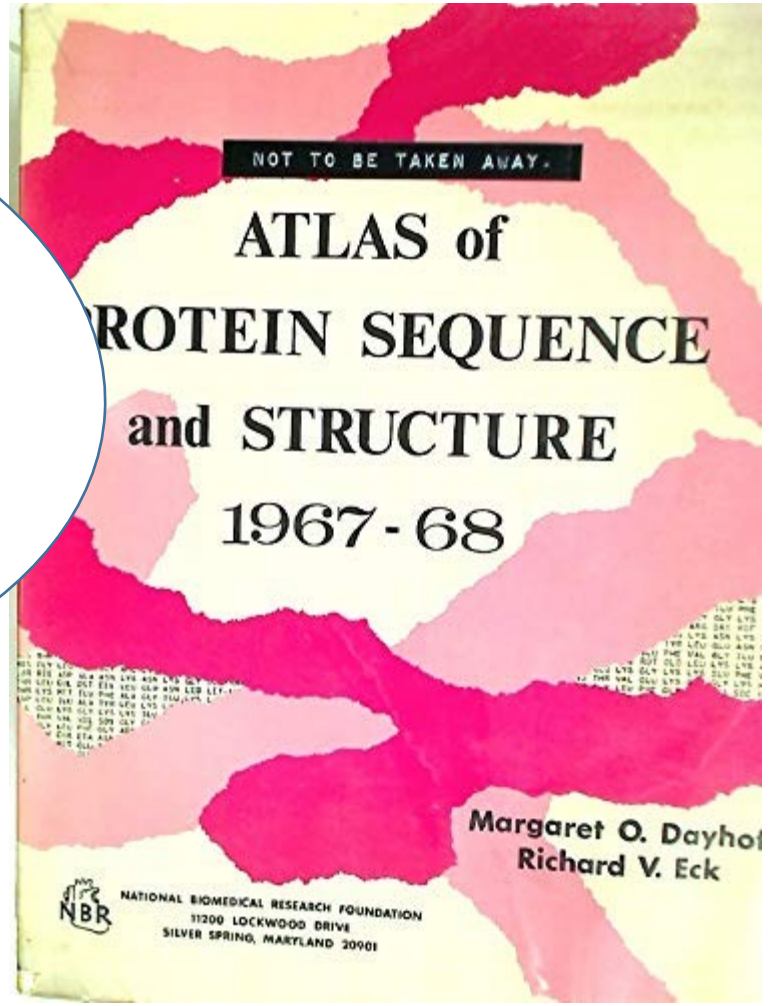


Late 1960s – Birth of Bioinformatics

there is a tremendous amount of **information** regarding evolutionary history and biochemical function implicit in each sequence and the number of known sequences is **growing explosively**. We feel it is important to collect this significant information, **correlate** it into a unified whole and interpret it…



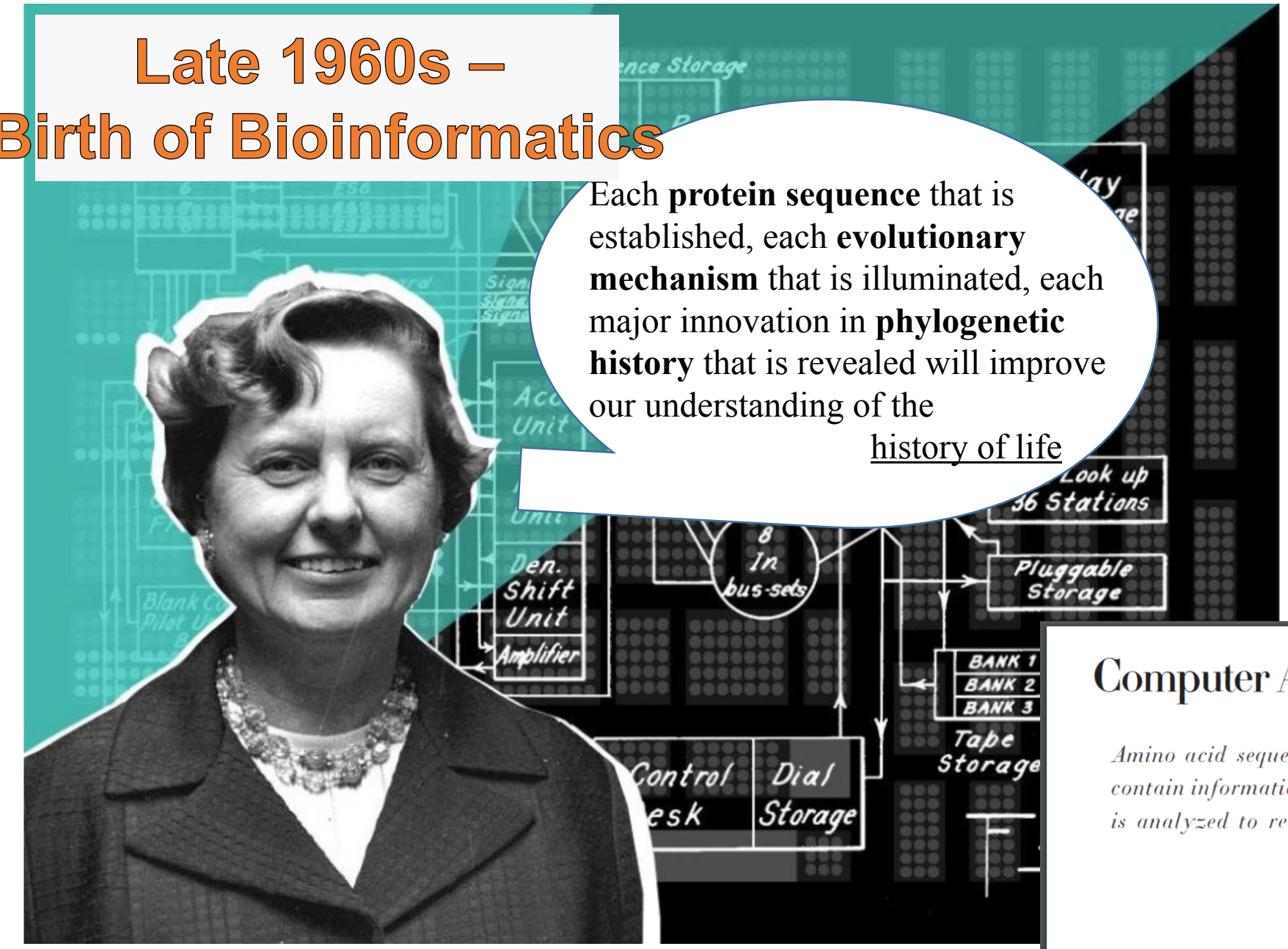
Margaret Dayhoff



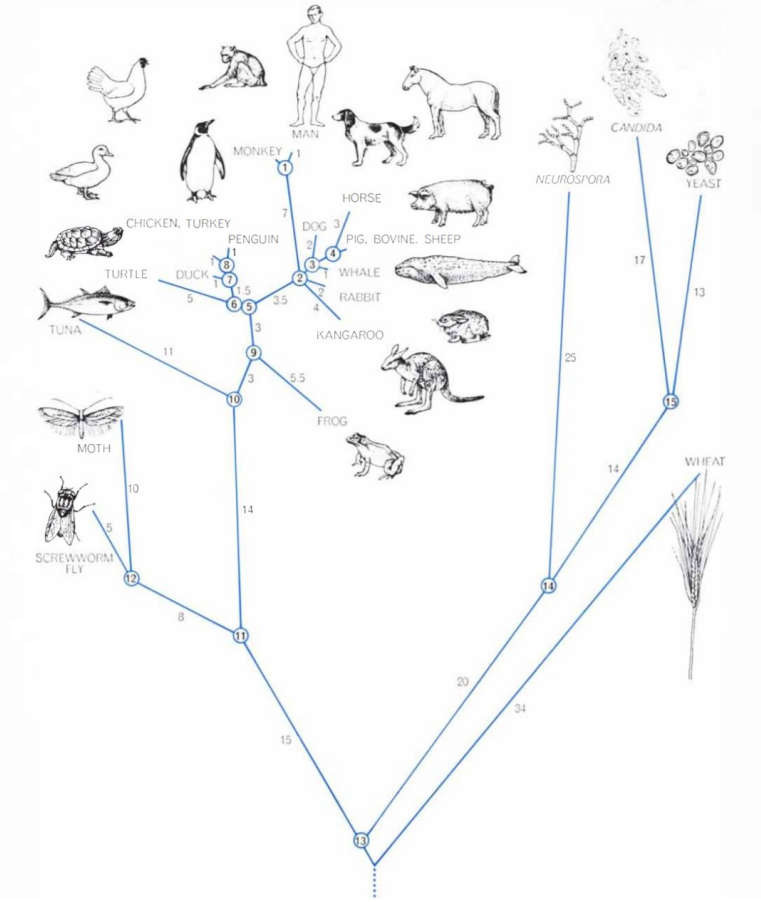
Book of Protein Sequences
Contained 65 protein sequences
from various species

Late 1960s – Birth of Bioinformatics

Each **protein sequence** that is established, each **evolutionary mechanism** that is illuminated, each major innovation in **phylogenetic history** that is revealed will improve our understanding of the history of life



Margaret Dayhoff



Computer Analysis of Protein Evolution

Amino acid sequences of similar proteins in different organisms contain information on relations among species. This information is analyzed to reconstruct in detail the history of living things

by Margaret Oakley Dayhoff

The protein molecules that determine the form and function of every living thing are intricately

sequences is something fundamentally new in biology and biochemistry, unprecedented in quantity, in concentrated

tions in the organisms in which they are found, and they can often be substituted for one another in laboratory experi-

1970s – Nucleic Acid Sequencin

J Mol Biol. 1977 Oct 15;116(1):29-30. doi: 10.1016/0022-2836(77)90116-4.

Computer processing of DNA sequence data

D McCallum, M Smith

PMID: 592383 DOI: 10.1016/0022-2836(77)90116-4

APPENDIX

Computer Processing of DNA Sequence Data

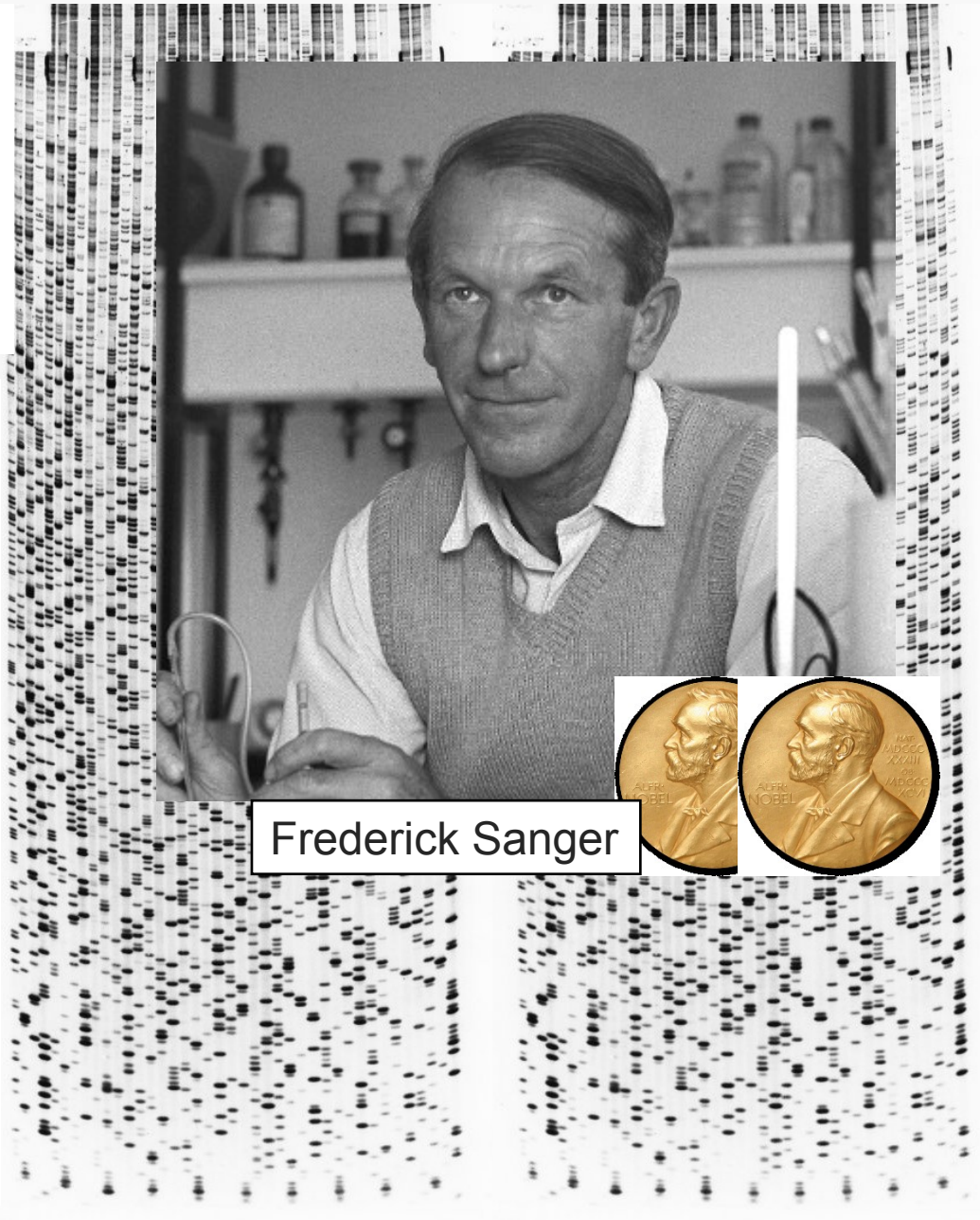
DUNCAN MCCALLUM

Management Services, Ciba-Geigy Ltd, Duxford, Cambridge, England

AND MICHAEL SMITH

Department of Biochemistry, Faculty of Medicine
University of British Columbia, 2075 Westbrook Place
Vancouver, B.C., Canada V6T 1W5

The sequence of ϕX174 DNA contains approximately 5400 nucleotides. Therefore, it was desirable to have automated procedures to process this large amount of data, both to eliminate errors and to save time. In this Appendix we describe the basic features of the computer programs used in this study.



1980s – Automated Genome Sequencing

GENOMICS 1, 201–212 (1987)

REVIEW

Automated DNA Sequencing and Analysis of the Human Genome

LEROY E. HOOD, MICHAEL W. HUNKAPILLER,* AND LLOYD M. SMITH[1]

*California Institute of Technology, Pasadena, California 91125, and *Applied Biosystems, Inc., 850 Lincoln Centre Drive, Foster City, California 94404*

Received October 14, 1987

In the past few years, striking advances have been made in automating DNA sequence analysis. Currently, efforts are underway to automate and improve DNA purification, mapping, and data processing procedures. The predictable advances in these technologies should soon place us in a position to sequence the entire human genome. The information derived from this project will have profound implications for basic biology and clinical medicine alike.

INTRODUCTION

A proposal to undertake the detailed mapping and sequence analysis of the human genome has developed within the biological community in the last 2 years. This proposal has met with enthusiasm on the part of some and skepticism on the part of others. The complementary strands of DNA. These strands are long, linear arrays of four different nucleotides (A, G, C, and T), and complementarity is achieved by the A's and C's on one strand always pairing with the T's and G's, respectively, on the other. These chromosomes contain most of the information necessary for the construction of a human organism. The one-dimensional information of the nucleotide sequence in the chromosomes, encoded in discrete segments called genes, is transcribed and translated into linear protein polymers composed of 20 different amino acid subunits. The linear amino acid sequences of proteins direct their folding into the three-dimensional structures that give our body size and shape and catalyze the chemical reactions of life. The human genome contains about 3 billion nucleotides per haploid set of chromosomes (a haploid genome is one in which there is only one member of each chromosome pair), a

Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2444-2448, April 1988
Biochemistry

Improved tools for biological sequence comparison

(amino acid/nucleic acid/data base searches/local similarity)

WILLIAM R. PEARSON* AND DAVID J. LIPMAN†

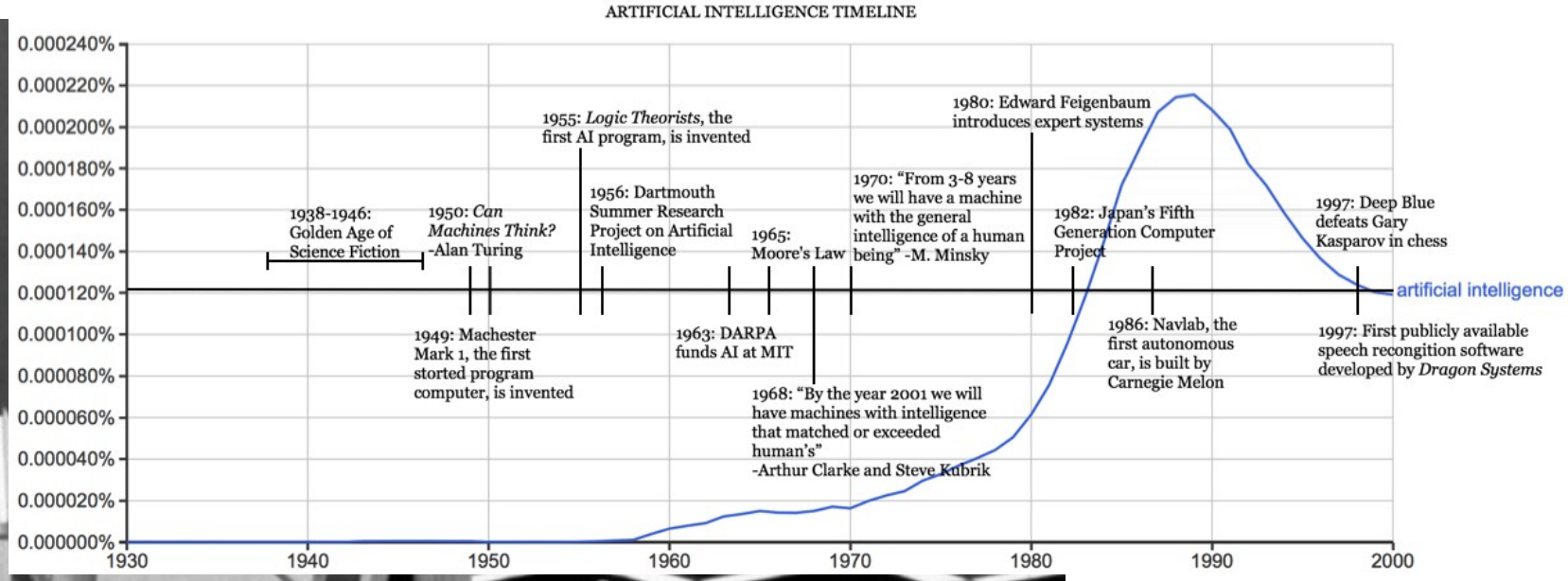
*Department of Biochemistry, University of Virginia, Charlottesville, VA 22908; and †Mathematical Research Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892

Communicated by Gerald M. Rubin, December 2, 1987 (received for review September 17, 1987)

ABSTRACT **We have developed three computer programs for comparisons of protein and DNA sequences. They can be used to search sequence data bases, evaluate similarity scores, and identify periodic structures based on local sequence similarity. The FASTA program is a more sensitive derivative of the FASTP program, which can be used to search protein or DNA sequence data bases and can compare a protein sequence to a DNA sequence data base by translating the DNA data base as it is searched. FASTA includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences. The RDF2 program can be used to evaluate the significance of similarity scores using a shuffling method that preserves local sequence composition. The LFASTA program can display all the regions of local similarity between two sequences with scores greater than a threshold, using the same scoring parameters and a similar alignment algorithm; these local similarities can be displayed as a "graphic matrix" plot or as individual alignments. In addition, these programs have been generalized to allow comparison of DNA or protein sequences based on a variety of alternative scoring matrices.**

FASTP and FASTA achieve much of their speed and selectivity in the first step, by using a lookup table to locate all identities or groups of identities between two DNA or amino acid sequences during the first step of the comparison (2). The *ktup* parameter determines how many consecutive identities are required in a match. For example, if *ktup* = 4 for a DNA sequence comparison, only those identities that occur in a run of four consecutive matches are examined. In the first step, the 10 best diagonal regions are found using a simple formula based on the number of *ktup* matches and the distance between the matches without considering shorter runs of identities, conservative replacements, insertions, or deletions (1, 3).

In the second step of the comparison, we rescore these 10 regions using a scoring matrix that allows conservative replacements and runs of identities shorter than *ktup* to contribute to the similarity score. For protein sequences, this score is usually calculated using the PAM250 matrix (4), although scoring matrices based on the minimum number of base changes required for a replacement or on an alternative measure of similarity can also be used with FASTA. For each of these best diagonal regions, a subregion with maxi-



1980s – 'Artificial Intelligence' Expert Syster

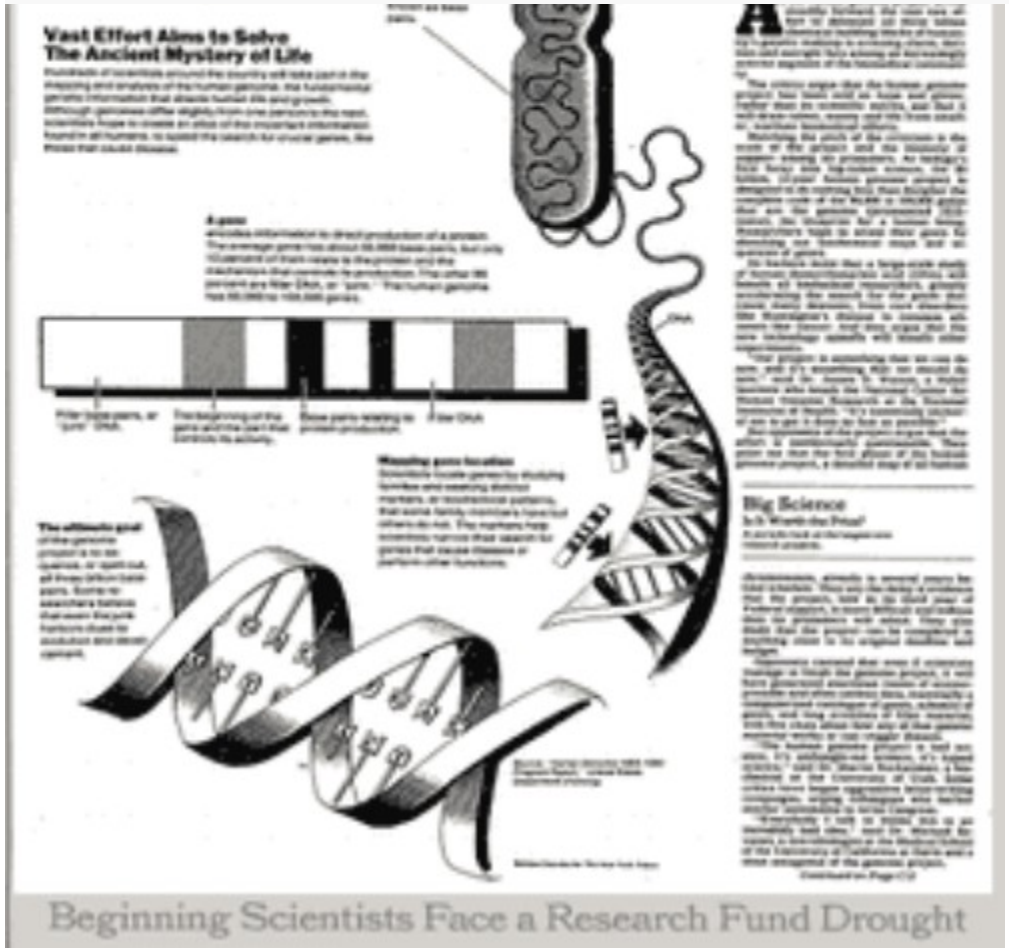
> *"… very limited success in particular areas, followed immediately by failure to reach the broader goal at which these initial successes seem at first to hint…".*


Science Times

The New York Times

Great 15-Year Project To Decipher Genes Stirs Opposition

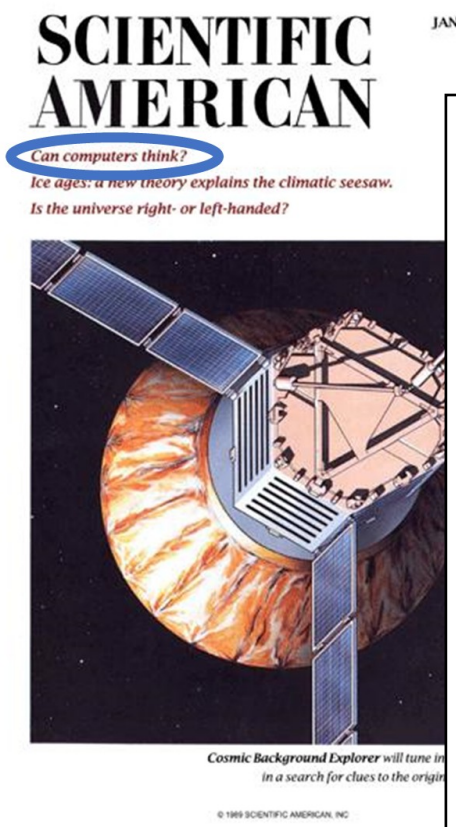
1990 – Start of the Human Genome Project


Vast Effort Aims to Solve The Ancient Mystery of Life

Big Science

Beginning Scientists Face a Research Fund Drought

1990 – Boosting


SCIENTIFIC AMERICAN

JANUARY 1990
$2.95

Can computers think?

Ice ages: a new theory explains the climatic seesaw.

Is the universe right- or left-handed?

Cosmic Background Explorer will tune in

in a search for clues to the origin

Machine Learning, 5, 197–227 (1990)

The Strength of Weak Learnability

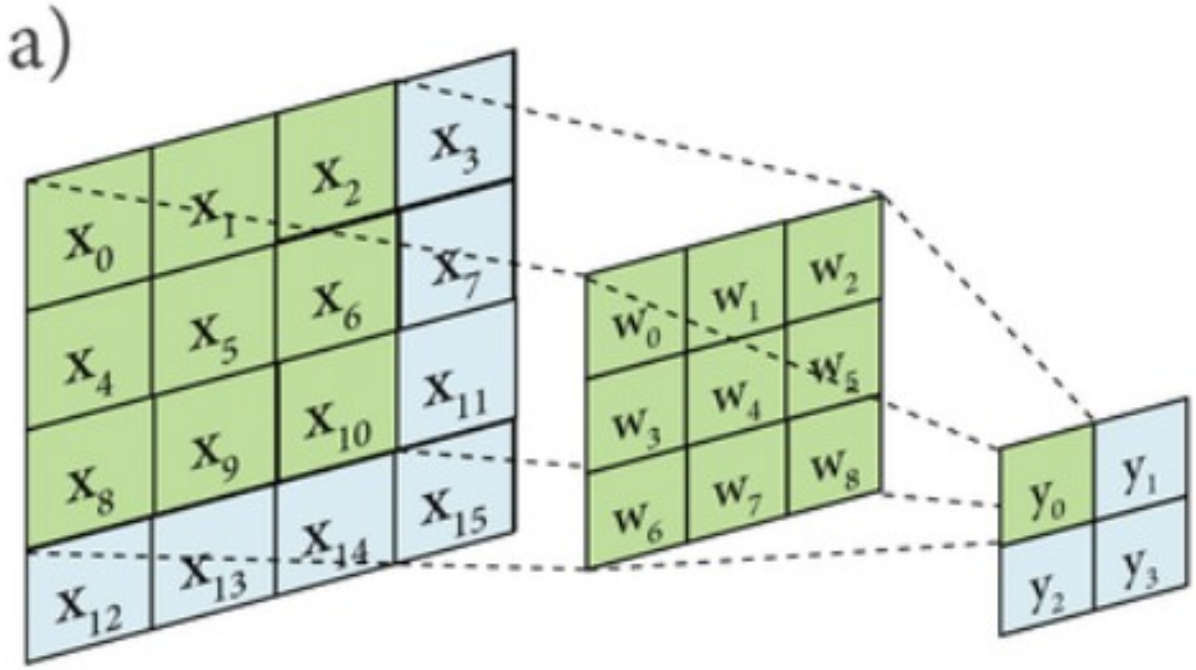
ROBERT E. SCHAPIRE (rs@theory.lcs.mit.edu)

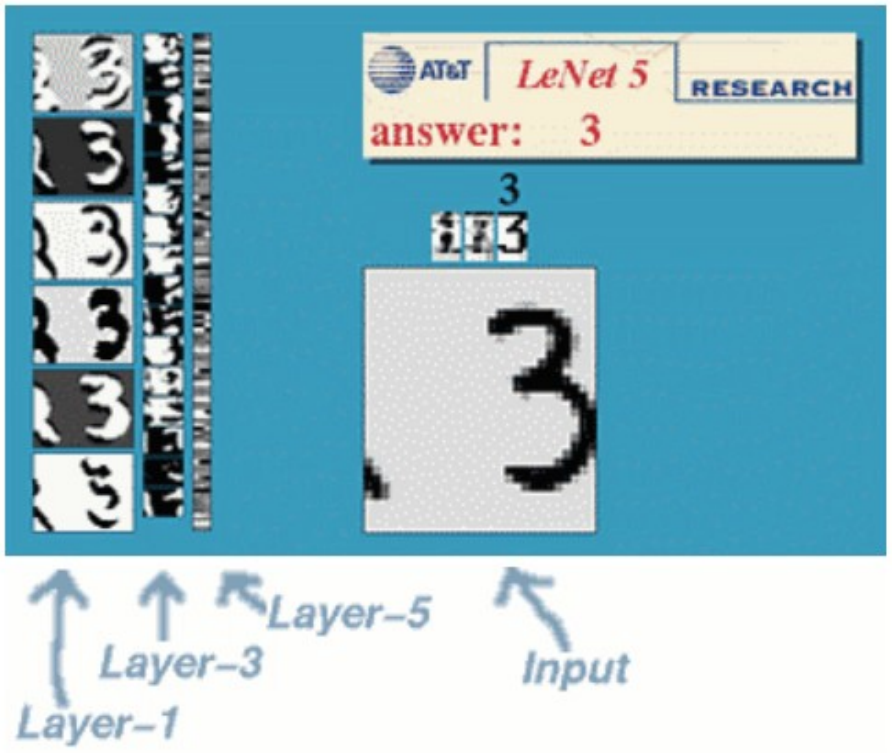
MIT Laboratory for Computer Science, 545 Technology Square, Cambridge, MA 02139

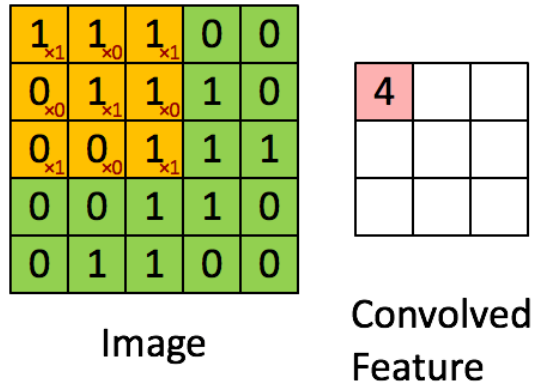
Abstract. This paper addresses the problem of improving the accuracy of an hypothesis output by a learning algorithm in the distribution-free (*PAC*) learning model. A concept class is *learnable* (or *strongly learnable*) if, given access to a source of examples of the unknown concept, the learner with high probability is able to output an hypothesis that is correct on all but an arbitrarily small fraction of the instances. The concept class is *weakly learnable* if the learner can produce an hypothesis that performs only slightly better than random guessing. In this paper, it is shown that these two notions of learnability are equivalent.

A method is described for converting a weak learning algorithm into one that achieves arbitrarily high accuracy. This construction may have practical applications as a tool for efficiently converting a mediocre learning algorithm into one that performs extremely well. In addition, the construction has some interesting theoretical consequences, including a set of general upper bounds on the complexity of any strong learning algorithm as a function of the allowed error ϵ.

990s – Convolutional Neural Network








Yann LeCun


Yoshua Bengio


Geoffrey Hinton

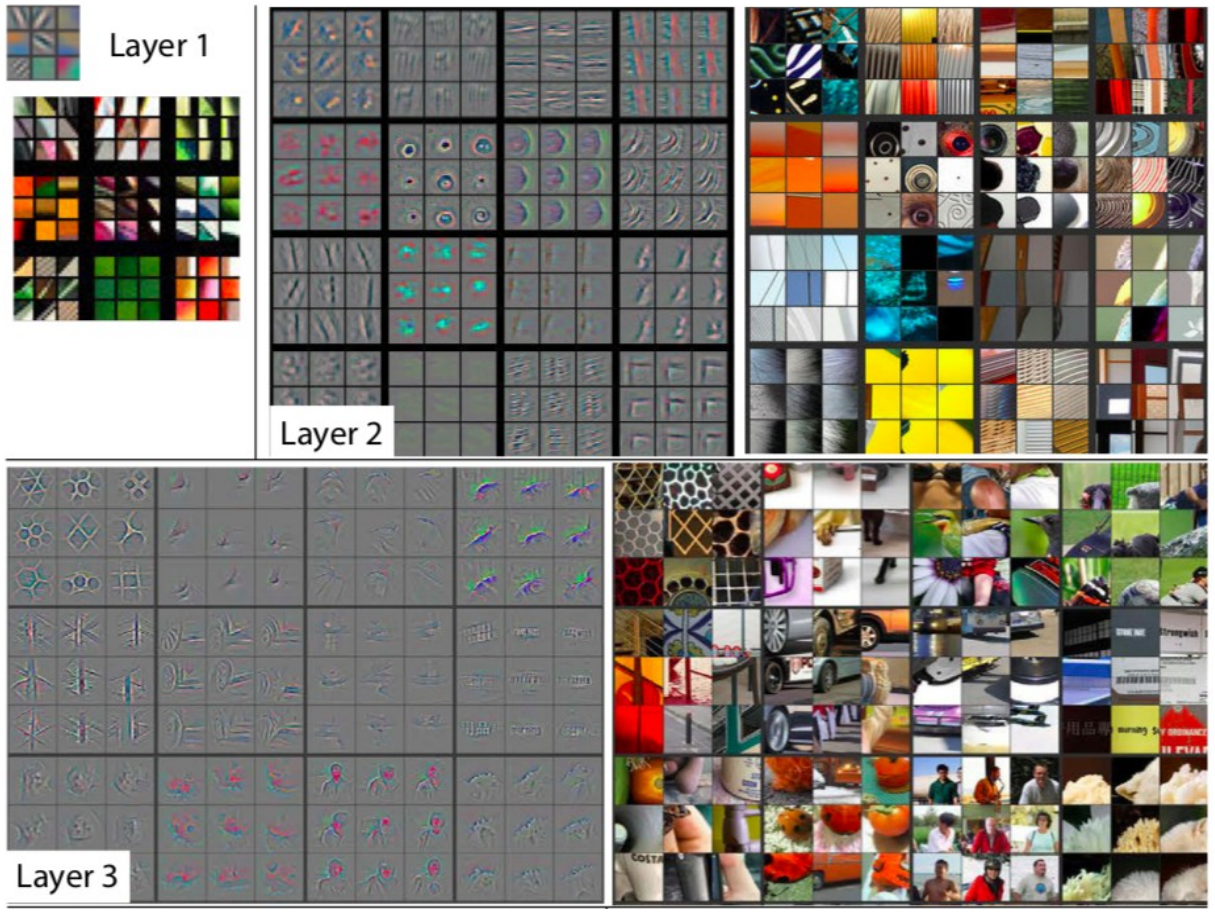
https://playground.tensorflow.org/

*

-1	-2	-1
0	0	0
1	2	1

*

-1	0	1
-2	0	2
-1	0	1



2015 – Deep Learning Revolution

Yann LeCun Yoshua Bengio Geoffrey Hinton

REVIEW

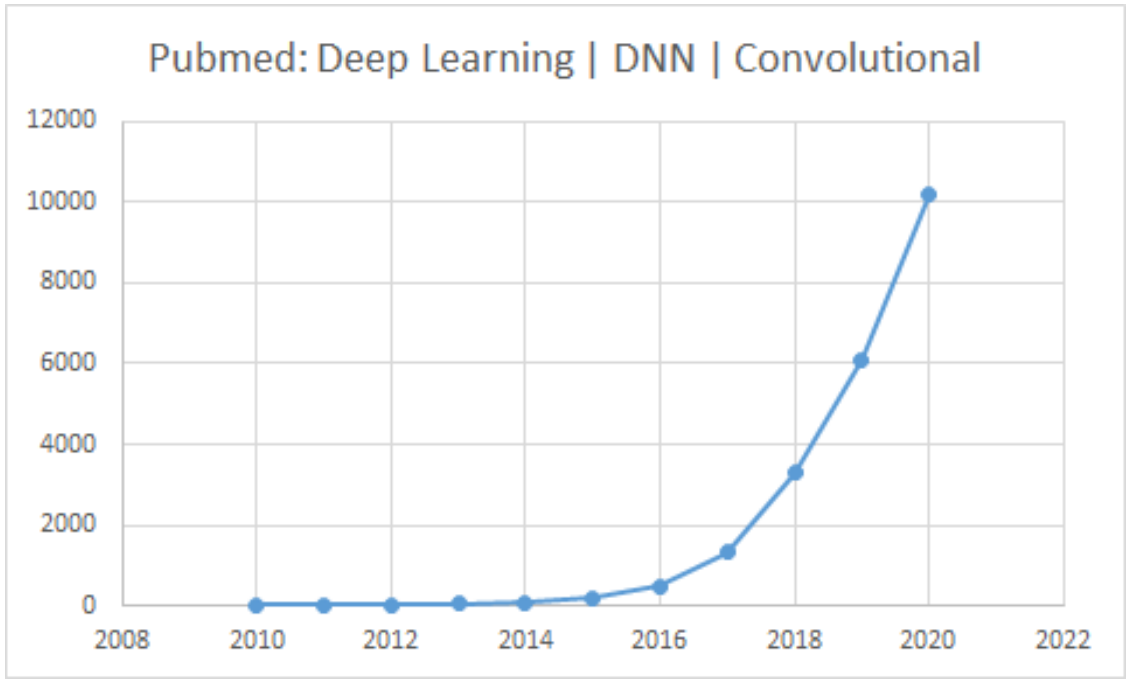
doi:10.1038/nature14539

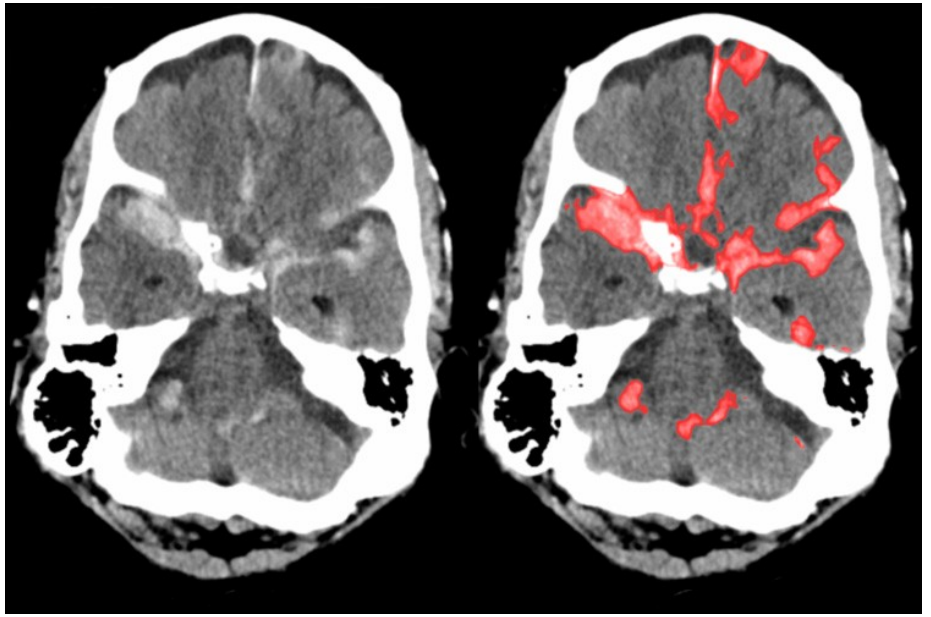
Deep learning

Yann LeCun[1,2], Yoshua Bengio[3] & Geoffrey Hinton[4,5]

Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.


M25-35/W
F25-35/B
F20-30/A
>MARY_SMITH*
>FEM/AGE_35
>ID_VIP++
>IN_STORE:
><30:21>
F25-35/W
F45-55/W
>JOHN_SMITH*
>MALE/AGE_41
>ID_VIP++
>IN_STORE:
><30:21>
OPTICIANS
LIGHT TRUCK
TRUCK
TRUCK
ВЫХОД В ГОРОД
ACCESS TO CITY
ALPHAGO



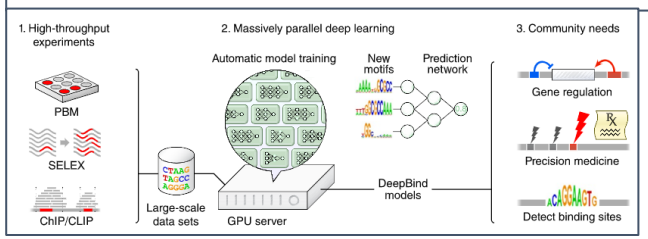


_computational BIOLOGY

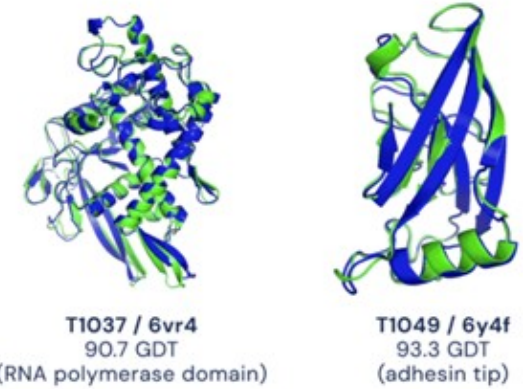
ANALYSIS

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi[1,2,6], Andrew Delong[1,6], Matthew T Weirauch[3–5] & Brendan J Frey[1–3]



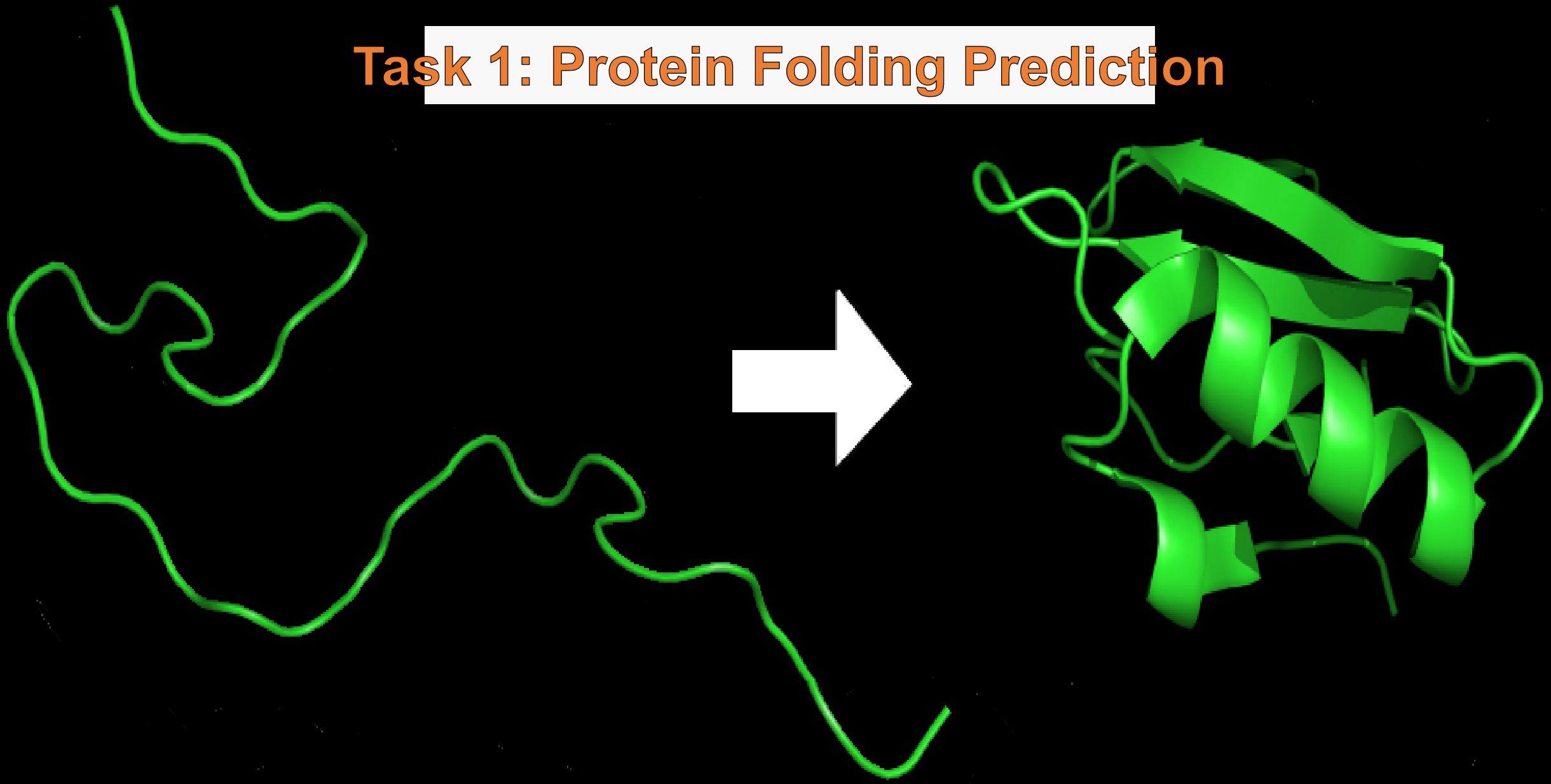
DeepMind



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Task 1: Protein Folding Prediction



1972 – Protein sequence and structure

STUDIES ON THE PRINCIPLES THAT GOVERN THE FOLDING OF PROTEIN CHAINS

Nobel Lecture, December 11, 1972

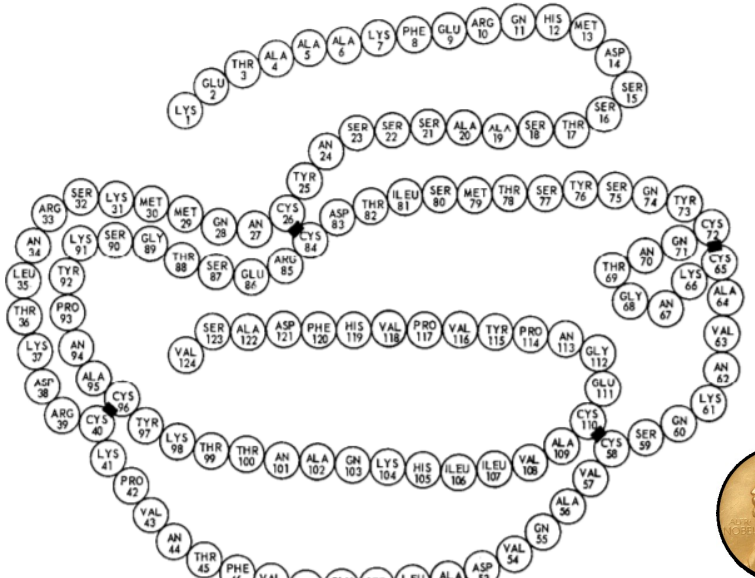
by

CHRISTIAN B. ANFINSEN

National Institutes of Health
Bethesda, Maryland

The telegram that I received from the Swedish Royal Academy of Sciences specifically cites ". . . studies on ribonuclease, in particular the relationship between the amino acid sequence and the biologically active conformation..." The work that my colleagues and I have carried out on the nature of the process that controls the folding of polypeptide chains into the unique three-dimensional structures of proteins was, indeed, strongly influenced by observations on the ribonuclease molecule, Many others, including Anson and Mirsky (1) in the '30s and Lumry and Eyring (2) in the '50s, had observed and discussed the reversibility of denaturation of proteins. However, the true elegance of this consequence of natural selection was dramatized by the ribonuclease work, since the refolding of this molecule, after full denaturation by reductive cleavage of its four disulfide bonds (Figure 1), required that only one of the 105

BOVINE PANCREATIC RIBONUCLEASE



1	MEPRVVKPPGQDLVVESLKSRYGLGGSCPDEYDFSNFYQSKYKRRTLTSP	50
51	GDLDIYSGDKVGSSLKYSDESKHCRTPLGSLFKHVNVNCLDDELDSFHDL	100
101	KKQETEEELIENDYRVSTSKITKQSFKEIEKVALPTNTTSSRPRTECCSD	150
151	AGDSPLKPVSCPKSKASDKRSLLPHQISQIYDELFQIHLKLQCETAAQQK	200
201	FAEELQKRERFLLEREQLLFRHENALSKIKGVEEEVLTRFQIIKEQHDAE	250
251	VEHLTEVLKEKNKETKRLRSSFDALKELNDTLKKQLNEASEENRKIDIQA	300
301	KRVQARLDNLQRKYEFMTIQRLKGSSHAVHEMKSLKQEKAPVSKTYKVPL	350
351	NGQVYELLTVFMDWISDHHLSKVKHEESGMDGKKPQLKFASQRNDIQEKC	400
401	VKLLPLMTEQLQWMPFVNIKLHEPFVKFIYWSLRQLDAGAQHSTMTSTLR	450
451	RLGEDIFKGVVTKGIQDNSPQHSVENKPKTAAFFKSSNLPLRFLSTLIVL	500
501	KTVTQADYLAQAFDSLCLDLKTEEGKTLFLEYQAVPVILSHLRISSKGLL	550
551	SNVIDSLLQMTVESKSLQPFLEACSNSLFFRTCSVLLRAPKLDLQILEKL	600
601	SIILQKLSKIKSNKKLFELFTIHLMLQEIQRTTNPEHAFLCINLNSTLFN	650
651	LGLTKCNSLVSSASP	700

In theory,
a protein's amino acid sequence
should fully determine its structure.

1994 – CASP: Critical Assessment of protein Structure Predictio

Establishes ‘Protein Folding’ problem as holy grail of machine learning in biology

Participants must blindly predict
the structure of the proteins, and these predictions
are subsequently compared to the ground truth
experimental data when they become available.

Given an amino-acid sequence predict protein structure

PROTEINS: Structure, Function, and Genetics 23:ii–iv (1995)

INTRODUCTION

A Large-Scale Experiment to Assess Protein Structure Prediction Methods

Methods for obtaining information about structure from amino acid sequence have apparently been advancing rapidly. But just what can these methods currently deliver? The following papers present the results of a large scale experiment that we have orchestrated to determine the current state of the art in protein structure prediction. We consider that the only way to objectively assess the used. The prediction challenge is then in devising techniques that can determine the detailed structural differences between the target and the known related structures. These techniques deal with the alignment of the target sequence on the templates, the best choice of template structure for each part of the chain, small (of the order of 1 or 2 Å) adjustments of main chain position, the orientation of side

John Moult

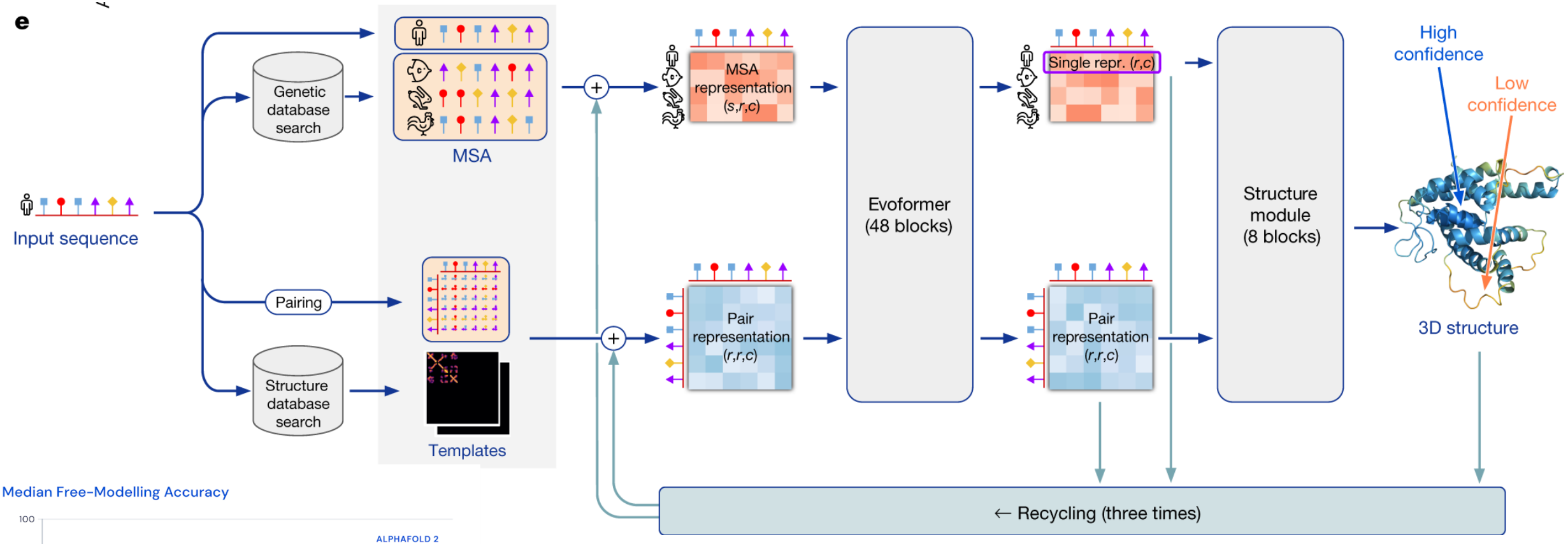
2020 – CASP ‘solved’ by Alphafold2

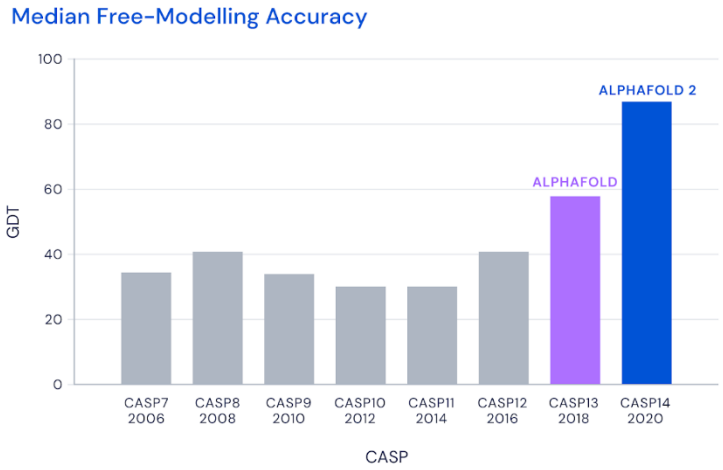
We have been stuck on this one problem – how do proteins fold up – for nearly 50 years. To see DeepMind produce a solution for this, having worked personally on this problem for so long and after so many stops and starts, wondering if we’d ever get there, is a very special moment.

PROFESSOR JOHN MOULT

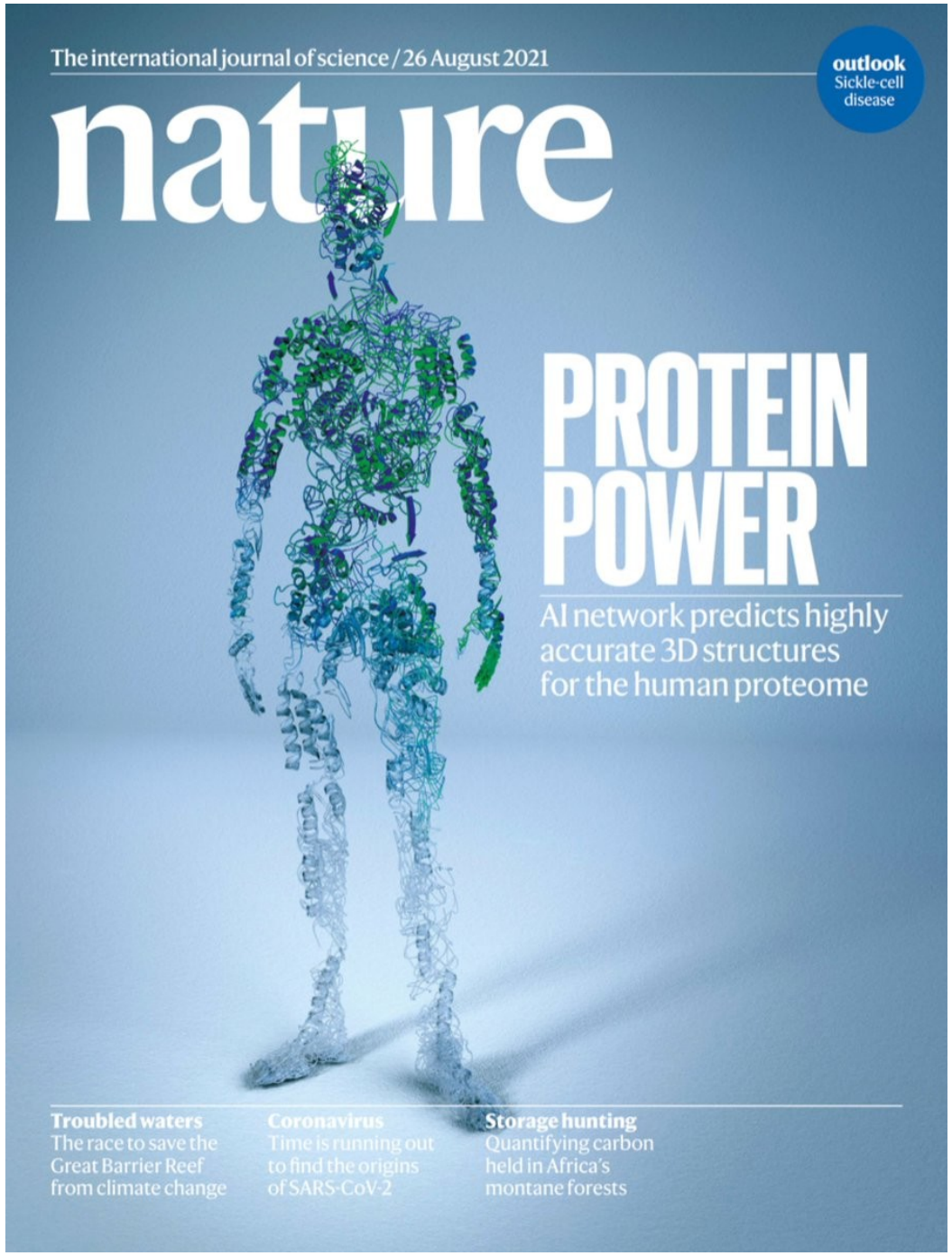
CO-FOUNDER AND CHAIR OF CASP, UNIVERSITY OF MARYLAND

2020 – CASP ‘solved’ by Alphafold2





https://www.nature.com/articles/s41586-021-03819-2



2021 –Alphafold2 changes Structural Biology

After decades of effort, only ~18% of the total residues in human protein sequences are covered by experimentally determined structures at this time. Alphafold doubles this number overnight.

In the near future, machine learning should be explored for predicting structures of protein–nucleic acid complexes... experimentally resolved protein–RNA complex structures remain low in number, and training sets are thus small, which may impair success at this time.

correspondence

Check for updates

AlphaFold2 and the future of structural biology

To the Editor — AlphaFold2 is a machine-learning algorithm for protein structure prediction that has now been used to obtain hundreds of thousands of protein models. The resulting resource is marvelous and will serve the community in many ways. Here I discuss the implications of this breakthrough achievement, which changes the way we do structural biology.

Imagine a website where you could already been applied to predict structures of several protein complexes. Like AlphaFold2, RoseTTAFold is available to the community and can now be used as an alternative route to predict protein structure from sequence.

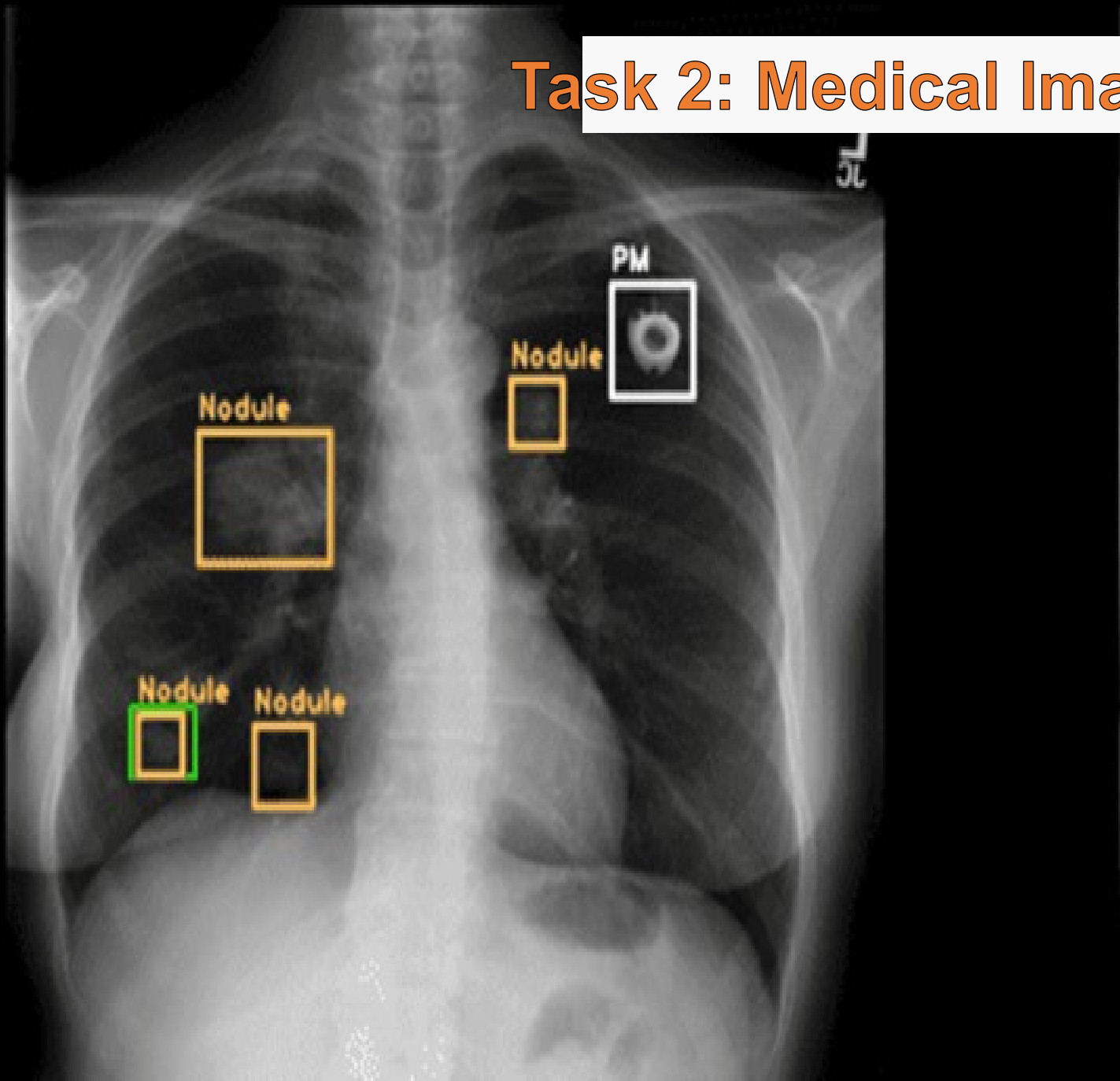
AlphaFold2 and the community

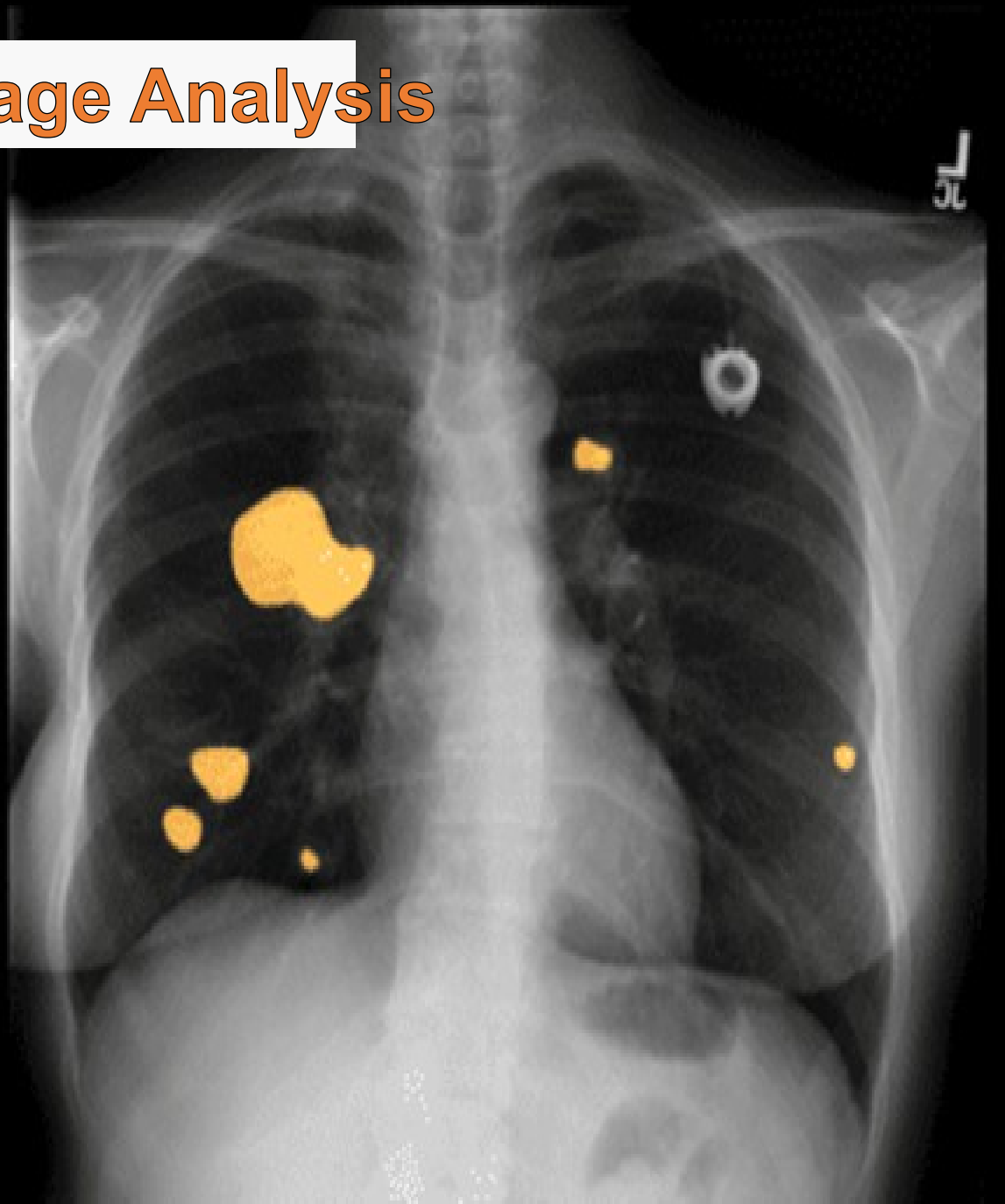
Half a century ago, the structural biology community had decided that all experimentally resolved macromolecular

solution of domain structures by NMR may be replaced by fast predictions so that the unique advantages of NMR in investigating protein folding and dynamics and the binding of ligands and nucleic acids can be utilized more readily.

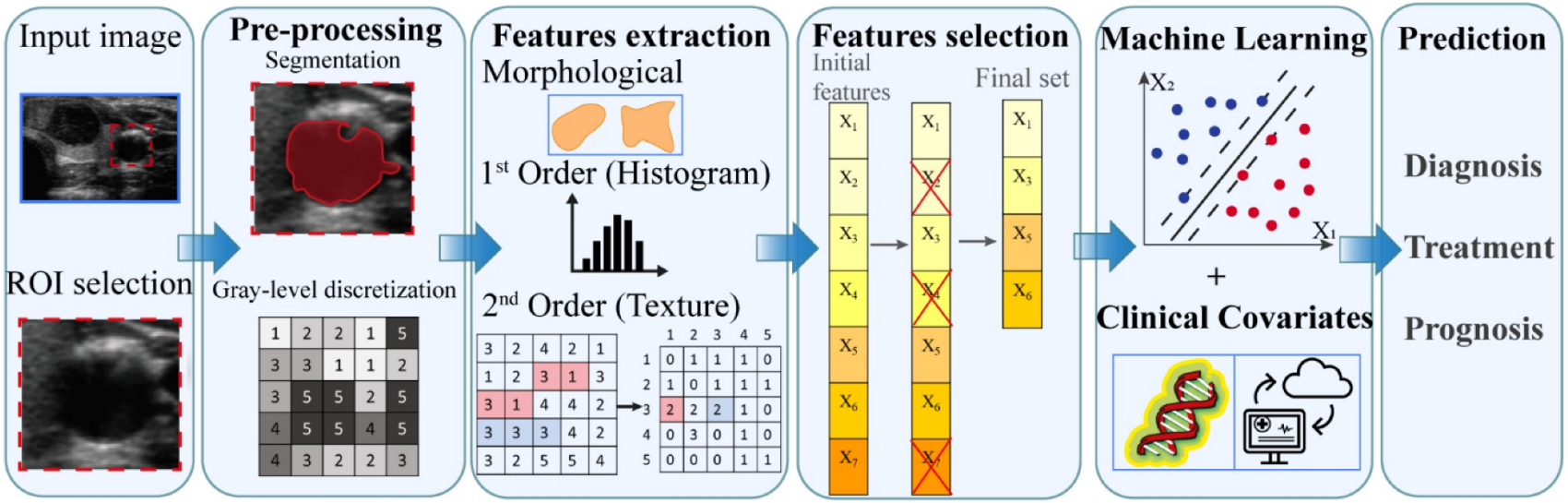
The new prediction algorithms should also improve automated model building. This will not change the general approach in structural biology, which has always

Task 2: Medical Image Analysis

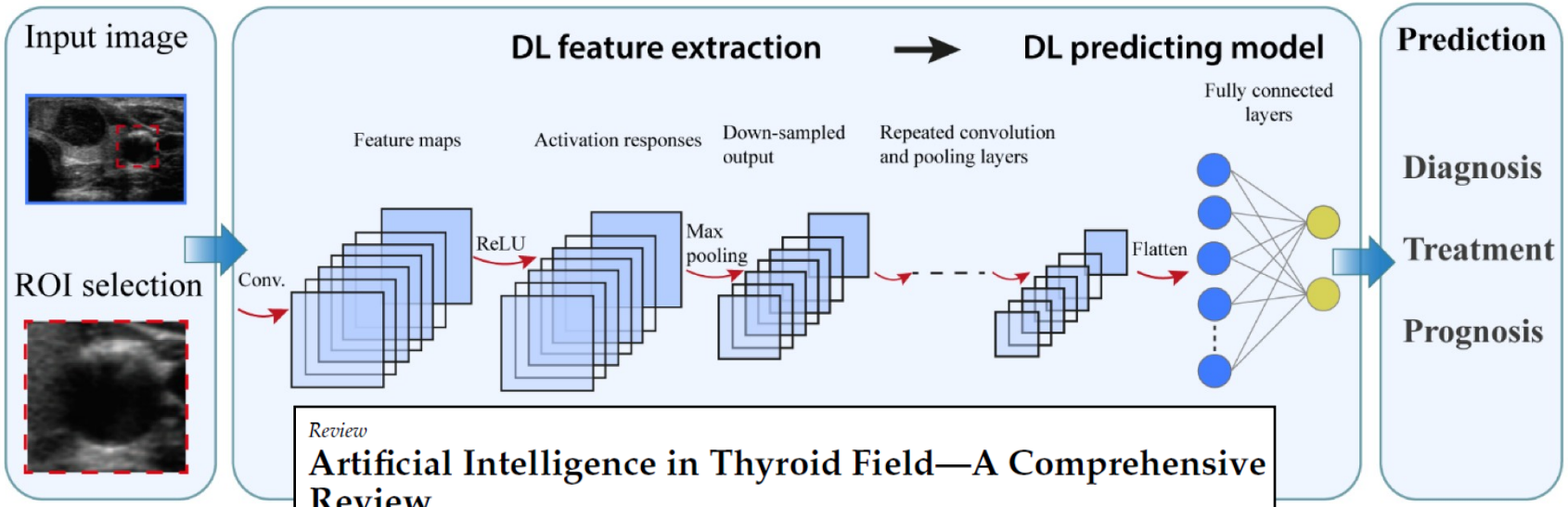




I. Conventional radiomics



II. Deep Learning based radiomics



Review

Artificial Intelligence in Thyroid Field—A Comprehensive Review

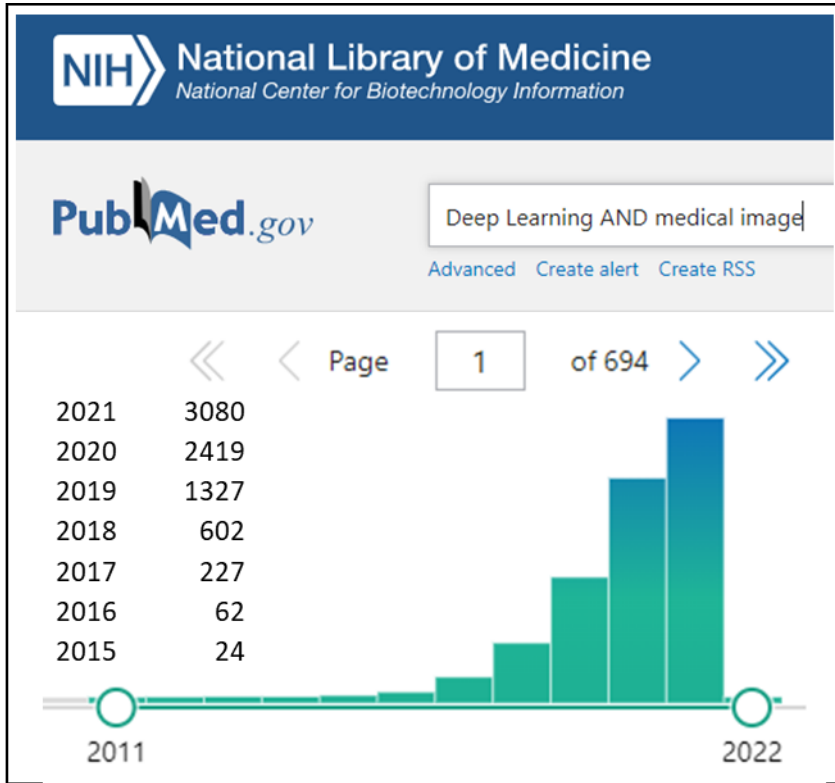
Fabiano Bini [1,*], Andrada Pica [1], Laura Azzimonti [2], Alessandro Giusti [2], Lorenzo Ruinelli [3,4], Franco Marinozzi [1] and Pierpaolo Trimboli [5,6]

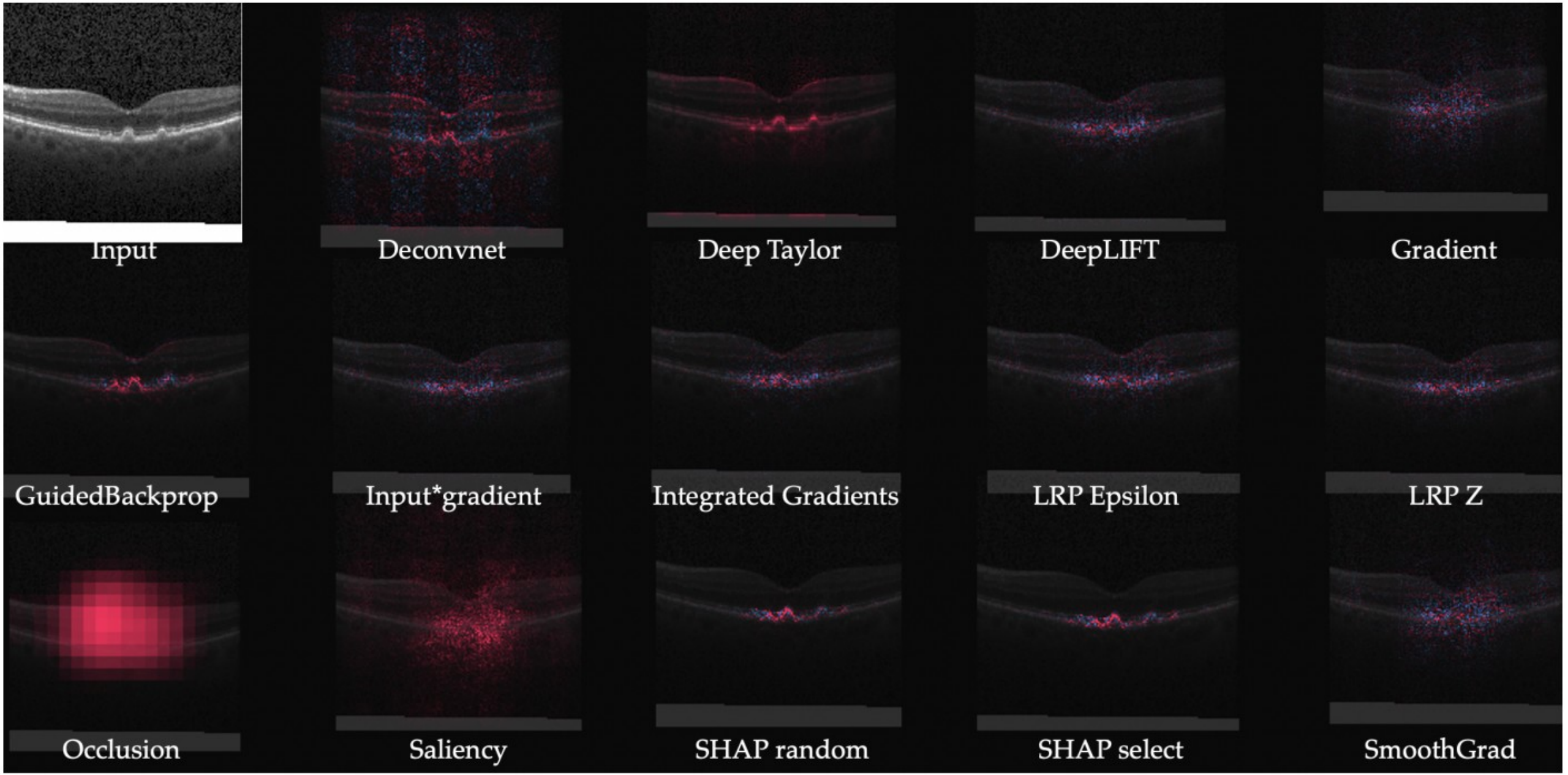
Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints

Ohad Oren, Bernard J Gersh, Deepak L Bhatt

Difficult circumstances might ensue in which a recommendation for treatment might be given in the absence of a well defined abnormality detected by routine imaging

Explainability





xplainability of medical diagnostics

On balance, it is likely that more and more microcomputer-based medical expert systems will become available. One can already find surprisingly complex expert systems that run on a microcomputer, although the scope is usually narrow…

Clinicians with an interest in expert systems should find that there are many opportunities to examine them through the increasing number of publications and conferences devoted to all facets of medicine and computing, including medical expert systems.

In April 2018, the US Food and Drug Administration approved **the first AI-based diagnostic, IDx-DR**, which detects diabetic retinopathy in people with diabetes by analyzing retinal images. Machine learning will soon be applied to many other medical conditions, from **cardiology** to **neurodegenerative** diseases and beyond…

1986 – Expert Systems

Today – Autonomous AI diagnostic

Medical Informatics

Medical Expert Systems—Knowledge Tools for Physicians

EDWARD H. SHORTLIFFE, MD, PhD, *Stanford, California*

Recent advances in the field of artificial intelligence have led to the emergence of expert systems, *computational tools designed to capture and make available the knowledge of experts in a field. Although much of the underlying technology available today is derived from basic research on biomedical advice systems during the 1970s, medical application packages are thus far generally unavailable from the young artificial intelligence industry. Medical expert systems will begin to appear, however, as researchers in medical artificial intelligence continue to make progress in key areas such as knowledge acquisition, model-based reasoning and system integration for clinical environments. It is accordingly important for physicians to understand the current state of such research and the theoretic and logistic barriers that remain before useful systems can be made available. One experimental system, ONCOCIN, provides a glimpse of the kinds of knowledge-based tools that will someday be available to physicians.*

(Shortliffe EH: Medical expert systems—Knowledge tools for physicians, *In* Medical informatics [Special Issue]. West J Med 1986 Dec; 145:830-839)

ARTICLE OPEN

Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices

Michael D. Abràmoff[1,2,3,4], Philip T. Lavin[5], Michele Birch[6], Nilay Shah[7] and James C. Folk[1,2,3]

Artificial Intelligence (AI) has long promised to increase healthcare affordability, quality and accessibility but FDA, until recently, had never authorized an autonomous AI diagnostic system. This pivotal trial of an AI system to detect diabetic retinopathy (DR) in people with diabetes enrolled 900 subjects, with no history of DR at primary care clinics, by comparing to Wisconsin Fundus Photograph Reading Center (FPRC) widefield stereoscopic photography and macular Optical Coherence Tomography (OCT), by FPRC certified photographers, and FPRC grading of Early Treatment Diabetic Retinopathy Study Severity Scale (ETDRS) and Diabetic Macular Edema (DME). More than mild DR (mtmDR) was defined as ETDRS level 35 or higher, and/or DME, in at least one eye. AI system operators underwent a standardized training protocol before study start. Median age was 59 years (range, 22–84 years); among participants, 47.5% of participants were male; 16.1% were Hispanic, 83.3% not Hispanic; 28.6% African American and 63.4% were not; 198 (23.8%) had mtmDR. The AI system exceeded all pre-specified superiority endpoints at sensitivity of 87.2% (95% CI, 81.8–91.2%) (>85%), specificity of 90.7% (95% CI, 88.3–92.7%) (>82.5%), and imageability rate of 96.1% (95% CI, 94.6–97.3%), demonstrating AI's ability to bring specialty-level diagnostics to primary care settings. Based on these results, FDA authorized the system for use by health care providers to detect more than mild DR and diabetic macular edema, making it, the first FDA authorized autonomous AI diagnostic system in any field of medicine, with the potential to help prevent vision loss in thousands of people with diabetes annually. ClinicalTrials.gov NCT02963441

npj Digital Medicine (2018)1:39 ; doi:10.1038/s41746-018-0040-6

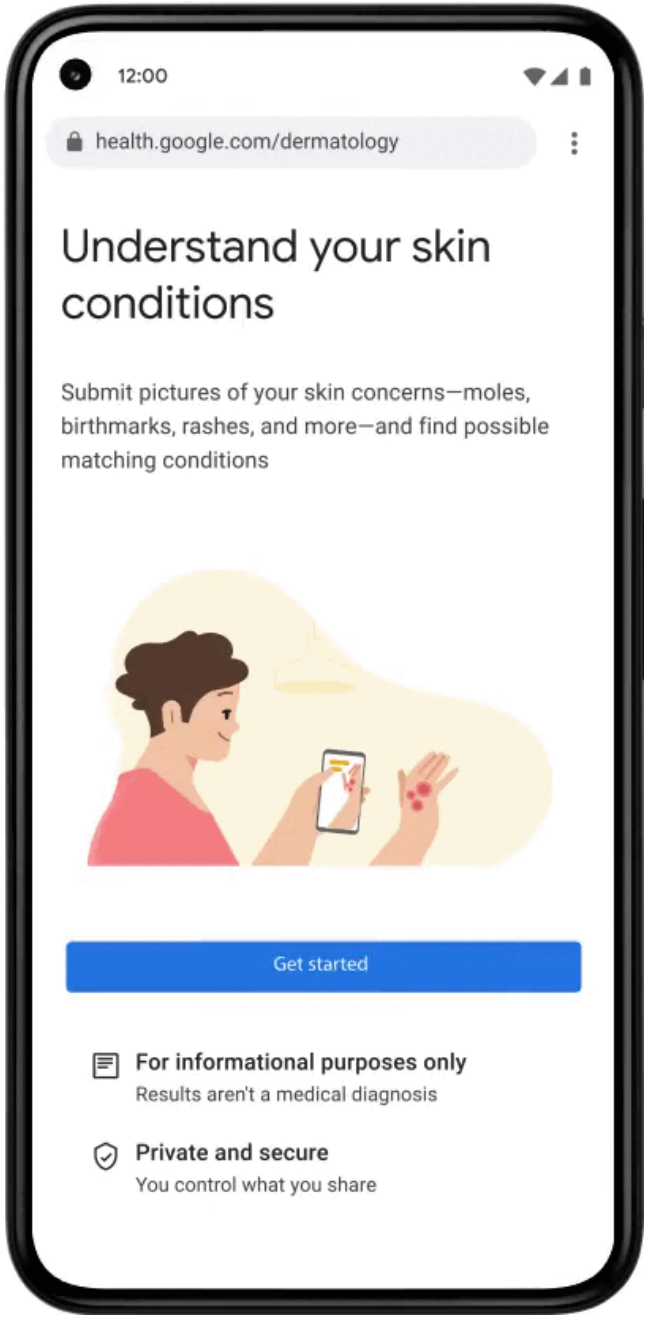
Google Health

AI-enabled imaging and diagnostics previously thought impossible

In partnership with healthcare organizations globally, we're researching robust new AI-enabled tools focused on diagnostics to assist clinicians. Drawing from diverse datasets, high-quality labels, and state-of-the-art deep learning techniques, we are making models that we hope will eventually support medical specialists in diagnosing disease. We're excited to further develop this research towards new frontiers—and to demonstrate that AI has the ability to enable novel, transformative diagnostics.

Improving access to skin disease information

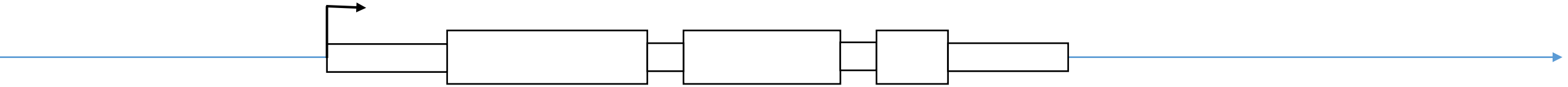
Through computer vision AI and image search capabilities, we are developing a tool to help individuals better research & identify their skin, hair, and nail conditions. The tool supports hundreds of conditions, including more than 80% of the conditions seen in clinics and more than 90% of the most commonly searched conditions. The work was highlighted in both Nature Medicine and JAMA Network Open.



This product has been CE marked as a Class I medical device in the EU. It is not available in the United States.

Task 3: Genomic Functional Annotation

Late 1990s – Genomic Annotation



Finding genes by computer: the state of the art

JAMES W. FICKETT

Discovering new genes, and their functions, can be aided not only by special purpose gene (and coding region) finding software, but also by searches in key databases, and by programs for finding particular sites relevant to gene expression, such as promoters and splice sites. No one software package includes all the necessary tools. I describe here the main kinds of tools; their working principles, strengths and limitations; and how combined evidence from multiple tools can aid in optimum gene identification.

Finding the genes in genomic DNA

Christopher B Burge* and Samuel Karlin†

Genome sequencing efforts will soon generate hundreds of millions of bases of human genomic DNA containing thousands of novel genes. In the past year, the accuracy of computational gene-finding methods has improved significantly, to the point where a reasonable approximation of the gene structures within an extended genomic region can often be predicted in advance of more detailed experimental studies.

Addresses
*Center for Cancer Research and Department of Biology, Massachusetts Institute of Technology, 40 Ames Street, E17-526 Cambridge, MA 02139, USA; e-mail: cburge@mit.edu
†Department of Mathematics, Stanford University, 450 Serra Mall, Stanford, CA 94305, USA; e-mail: sam@galois.stanford.edu
Correspondence: Samuel Karlin

Current Opinion in Structural Biology 1998, **8**:346–354

sequences, owing to the higher gene density typical of prokaryotes and the absence of introns in their protein coding genes. These properties generally imply that most open reading frames (ORFs) encountered in a prokaryotic sequence that are longer than some reasonable threshold, such as 300 or 500 base pairs (bp) will likely correspond to genes. The primary difficulties arising from this simple approach are that very small genes will be missed and that the occurrence of overlapping long ORFs on opposite DNA strands (genes and 'shadow genes') often leads to ambiguities. To resolve these problems, several methods have been devised that use different types of Markov models (see below) in order to capture the compositional differences among coding regions, 'shadow' coding regions (coding on the opposite DNA

2001 – Human Genome Sequenced

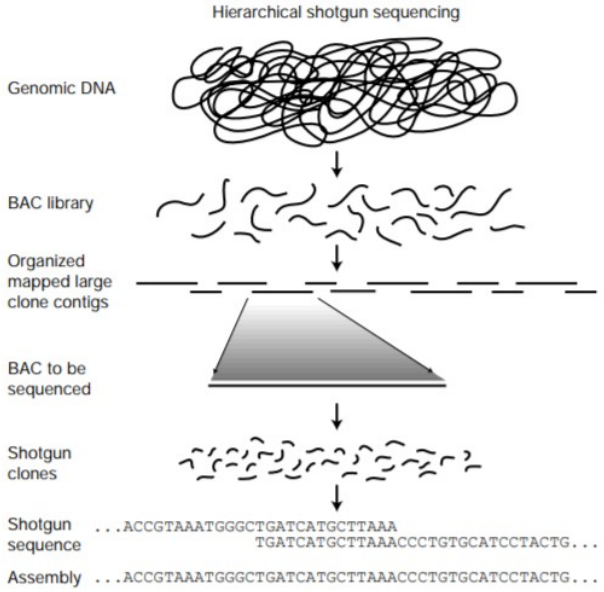
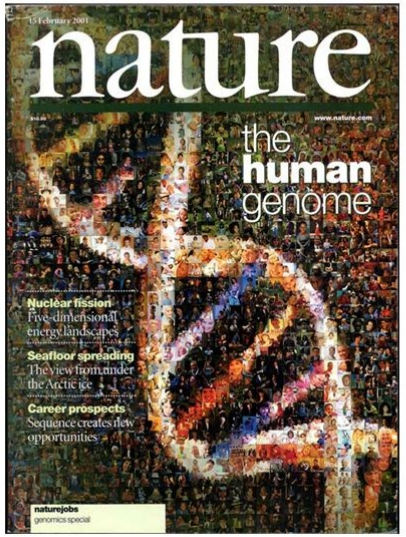
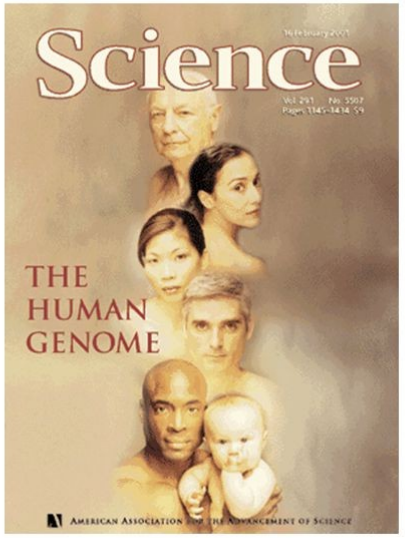


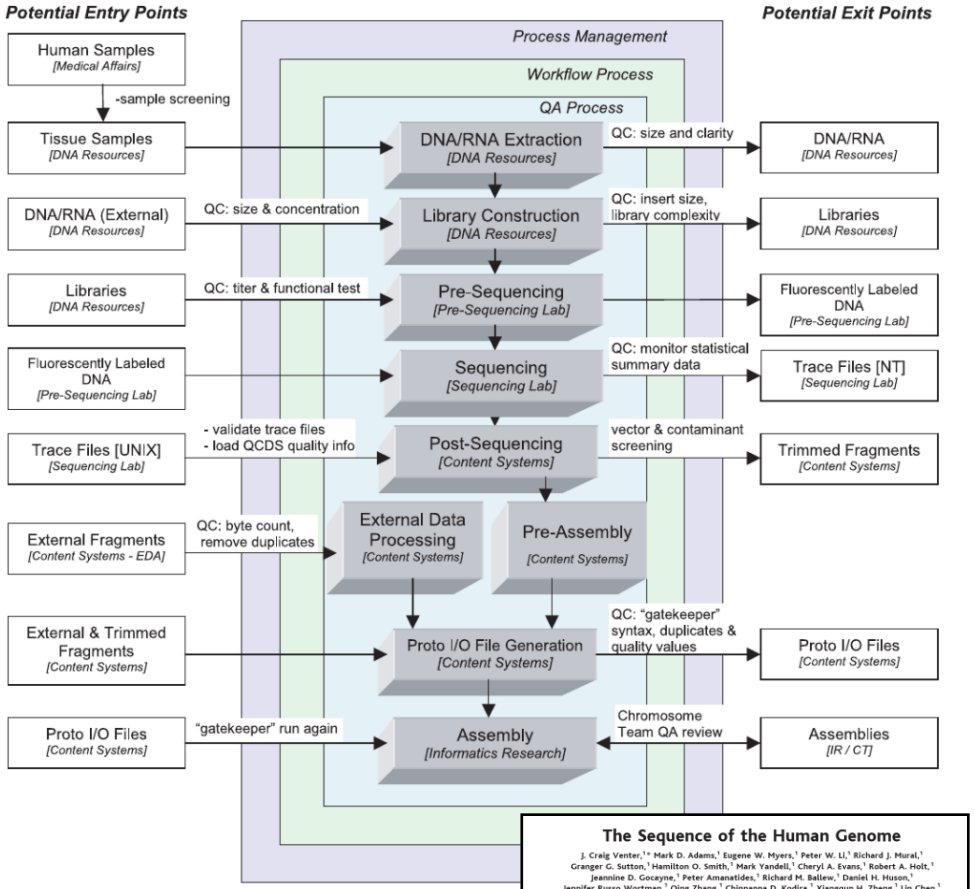
Figure 2 Idealized representation of the hierarchical shotgun sequencing strategy. A library is constructed by fragmenting the target genome and cloning it into a large-fragment cloning vector; here, BAC vectors are shown. The genomic DNA fragments represented in the library are then organized into a physical map and individual BAC clones are selected and sequenced by the random shotgun strategy. Finally, the clone sequences are assembled to reconstruct the sequence of the genome.

Genome sequencing

February 2001 - Publication of the first draft of the human genome







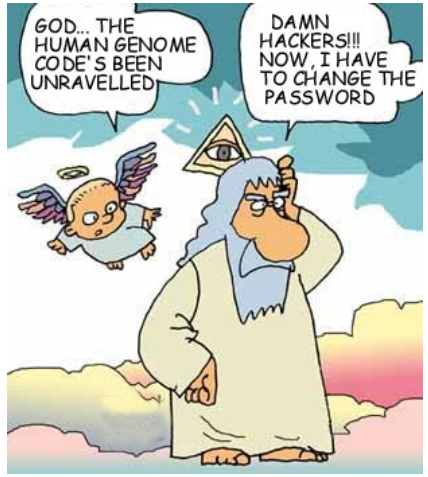
Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

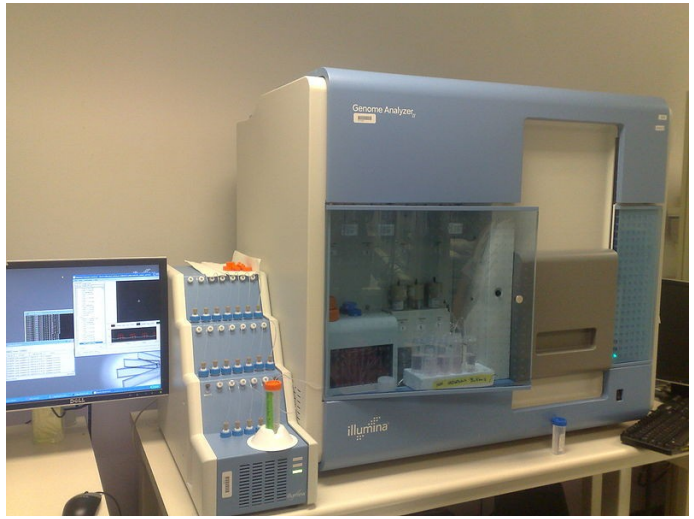
The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

Cost
~ 300M
USD

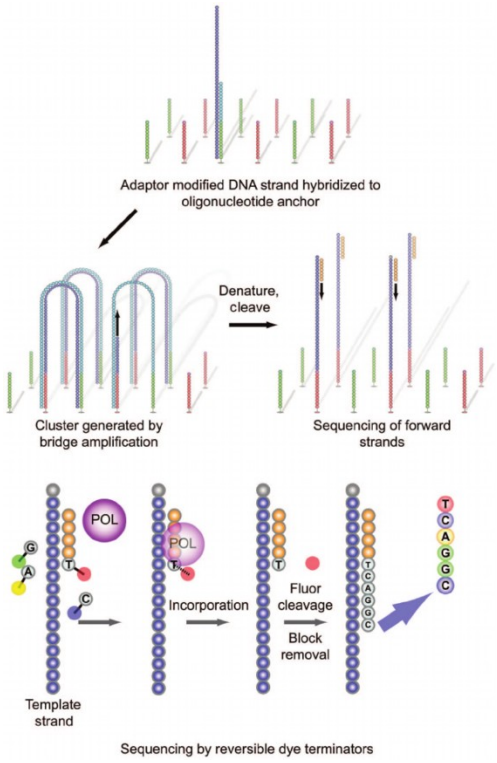


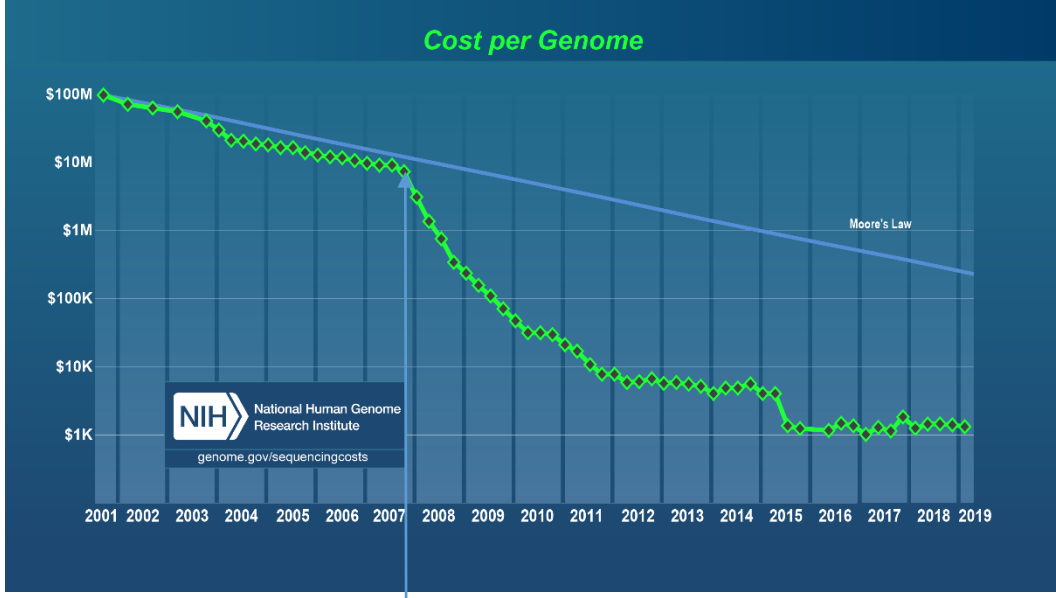

The Sequence of the Human Genome

2000s – Sequencing gets cheap

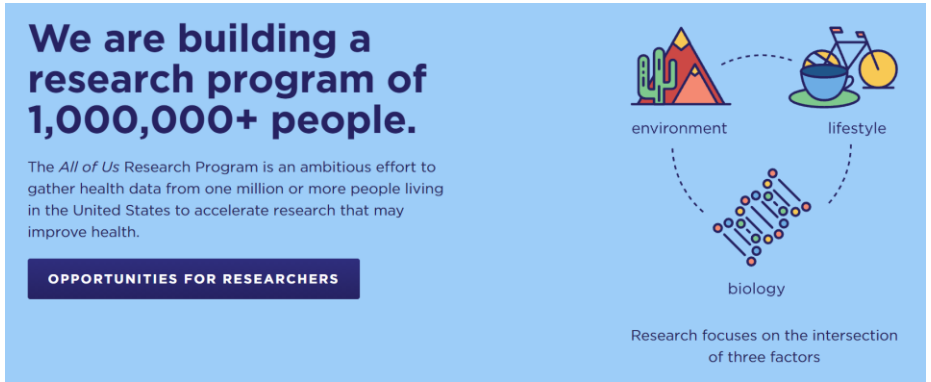


2006 – Solexa Genome Analyser
2007 – Solexa bought by Illumina





Next Generation Sequencing
New Generation Sequencing
NGS



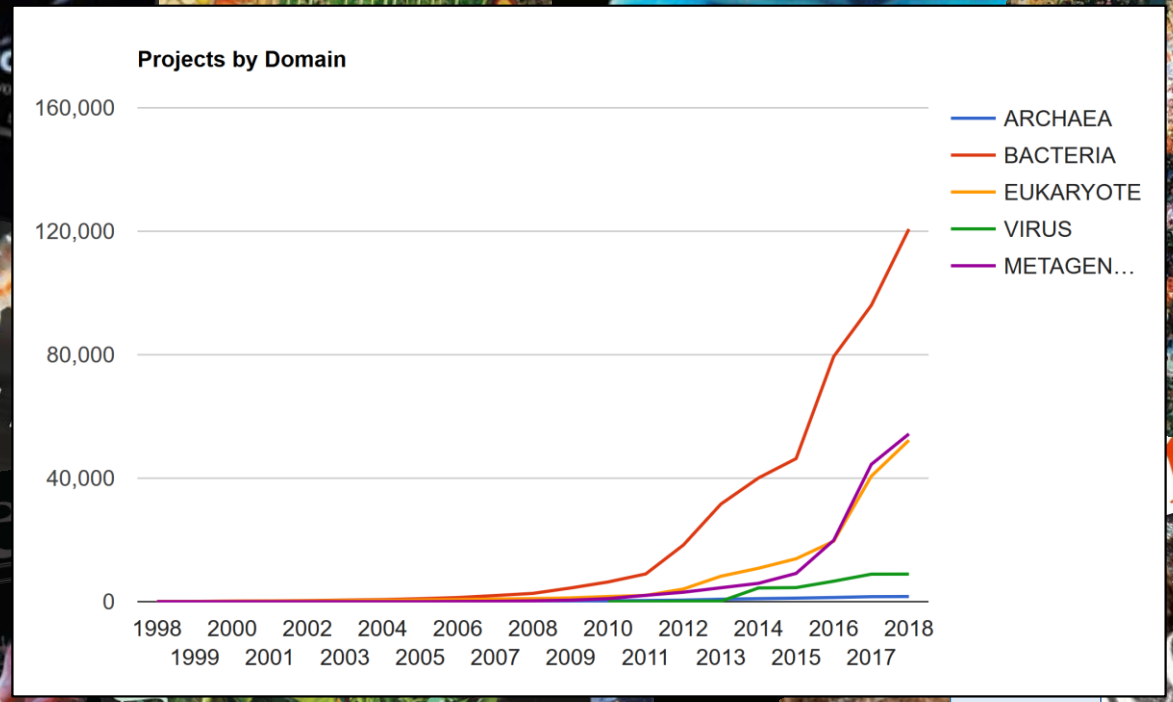
Realistic goal in three-five years

Sequence the entire human genome in a few days for $1000 (Era of Personal Genomics)

HOWEVER, speed of sequencing does not necessarily mean an **understanding** of the genetic information or DNA structure!

2015

The Genomic Era (2000-)



JGI GOLD

010s – Sequencing gets diversified

RNA Transcription
- Chromatin Isolation by RNA Purification (ChIRP-Seq)
- Global Run-on Sequencing (GRO-Seq)
- Ribosome Profiling Sequencing (Ribo-Seq)/ARTseq™
- RNA Immunoprecipitation Sequencing (RIP-Seq)
- High-Throughput Sequencing of CLIP cDNA library (HITS-CLIP) or
- Crosslinking and Immunoprecipitation Sequencing (CLIP-Seq)
- Photoactivatable Ribonucleoside–Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP)
- Individual Nucleotide Resolution CLIP (iCLIP)
- Native Elongating Transcript Sequencing (NET-Seq)
- Targeted Purification of Polysomal mRNA (TRAP-Seq)
- Crosslinking, Ligation, and Sequencing of Hybrids (CLASH-Seq)
- Parallel Analysis of RNA Ends Sequencing (PARE-Seq) or
- Genome-Wide Mapping of Uncapped Transcripts (GMUCT)
- Transcript Isoform Sequencing (TIF-Seq) or
- Paired-End Analysis of TSSs (PEAT)

RNA Structure
- Selective 2'-Hydroxyl Acylation Analyzed by Primer Extension Sequencing (SHAPE-Seq)
- Parallel Analysis of RNA Structure (PARS-Seq)
- Fragmentation Sequencing (FRAG-Seq)
- CXXC Affinity Purification Sequencing (CAP-Seq)
- Alkaline Phosphatase, Calf Intestine-Tobacco Acid Pyrophosphatase Sequencing (CIP-TAP)
- Inosine Chemical Erasing Sequencing (ICE)
- m6A-Specific Methylated RNA Immunoprecipitation Sequencing (MeRIP-Seq)

Low-Level RNA Detection
- Digital RNA Sequencing
- Whole-Transcript Amplification for Single Cells (Quartz-Seq)
- Designed Primer–Based RNA Sequencing (DP-Seq)
- Switch Mechanism at the 5' End of RNA Templates (Smart-Seq)
- Switch Mechanism at the 5' End of RNA Templates Version 2 (Smart-Seq2)
- Unique Molecular Identifiers (UMI)
- Cell Expression by Linear Amplification Sequencing (CEL-Seq)
- Single-Cell Tagged Reverse Transcription Sequencing (STRT-Seq)

Low-Level DNA Detection
- Single-Molecule Molecular Inversion Probes (smMIP)
- Multiple Displacement Amplification (MDA)
- Multiple Annealing and Looping–Based Amplification Cycles (MALBAC)
- Oligonucleotide-Selective Sequencing (OS-Seq)
- Duplex Sequencing (Duplex-Seq)

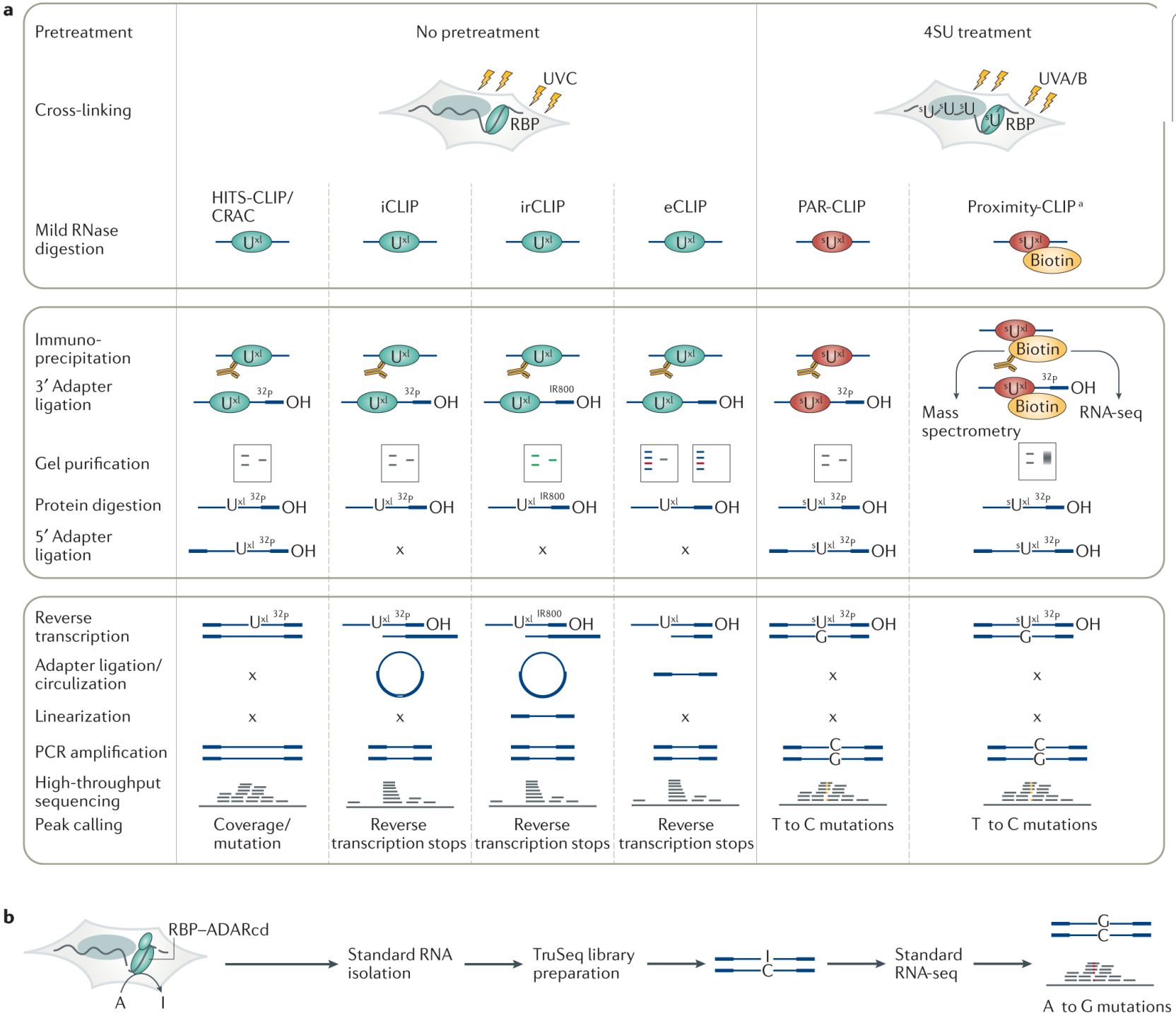
DNA Methylation
- Bisulfite Sequencing (BS-Seq)
- Post-Bisulfite Adapter Tagging (PBAT)
- Tagmentation-Based Whole Genome Bisulfite Sequencing (T-WGBS)
- Oxidative Bisulfite Sequencing (oxBS-Seq)
- Tet-Assisted Bisulfite Sequencing (TAB-Seq)
- Methylated DNA Immunoprecipitation Sequencing (MeDIP-Seq)
- Methylation-Capture (MethylCap) Sequencing or
- Methyl-Binding-Domain–Capture (MBDCap) Sequencing
- Reduced-Representation Bisulfite Sequencing (RRBS-Seq)

DNA-Protein Interactions
- DNase I Hypersensitive Sites Sequencing (DNase-Seq)
- MNase-Assisted Isolation of Nucleosomes Sequencing (MAINE-Seq)
- Chromatin Immunoprecipitation Sequencing (ChIP-Seq)
- Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-Seq)
- Assay for Transposase-Accessible Chromatin Sequencing (ATAC-Seq)
- Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)
- Chromatin Conformation Capture (Hi-C/3C-Seq)
- Circular Chromatin Conformation Capture (4-C or 4C-Seq)
- Chromatin Conformation Capture Carbon Copy (5-C)

Sequence Rearrangements
- Retrotransposon Capture Sequencing (RC-Seq)
- Transposon Sequencing (Tn-Seq) or Insertion Sequencing (INSeq)
- Translocation-Capture Sequencing (TC-Seq)

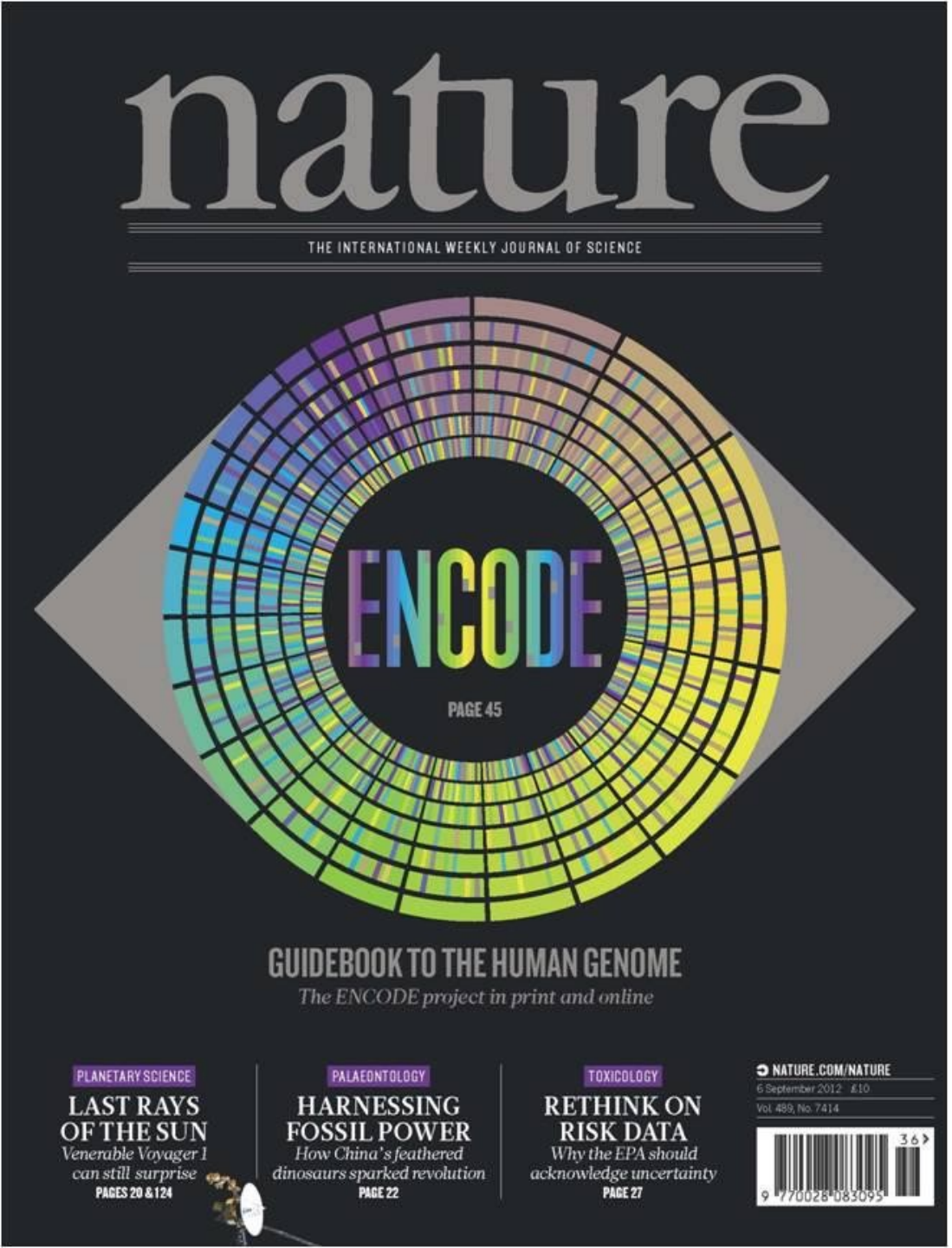
CLIP and complementary methods

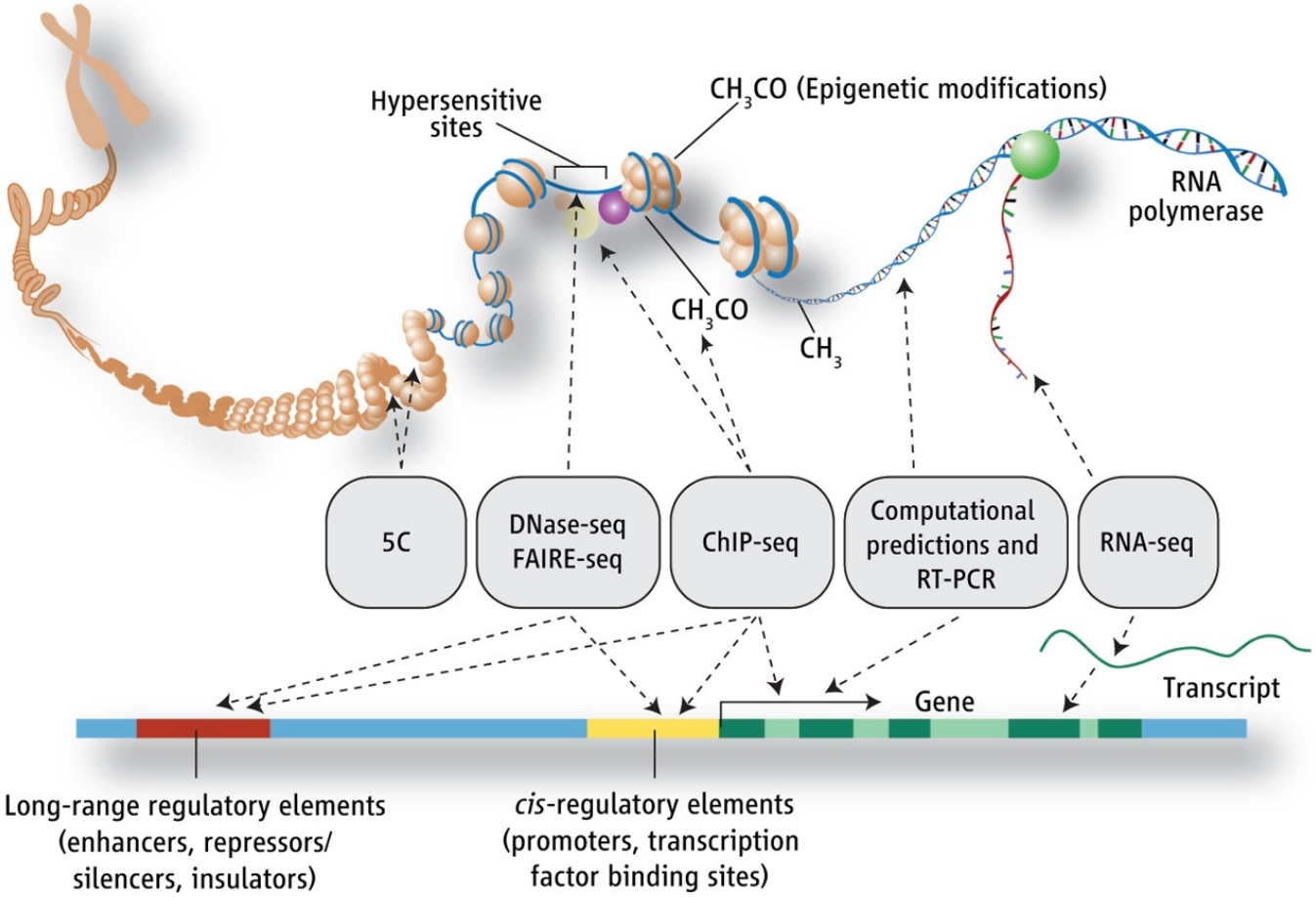
Markus Hafner[1], Maria Katsantoni[2,3], Tino Köster[4], James Marks[1], Joyita Mukherjee[5,6], Dorothee Staiger[4], Jernej Ule[5,6,7]✉ and Mihaela Zavolan[2,3]



https://www.nature.com/articles/s43586-021-00018-1

2012 – ENCODE





30 papers representing the integration and analysis of ENCODE data

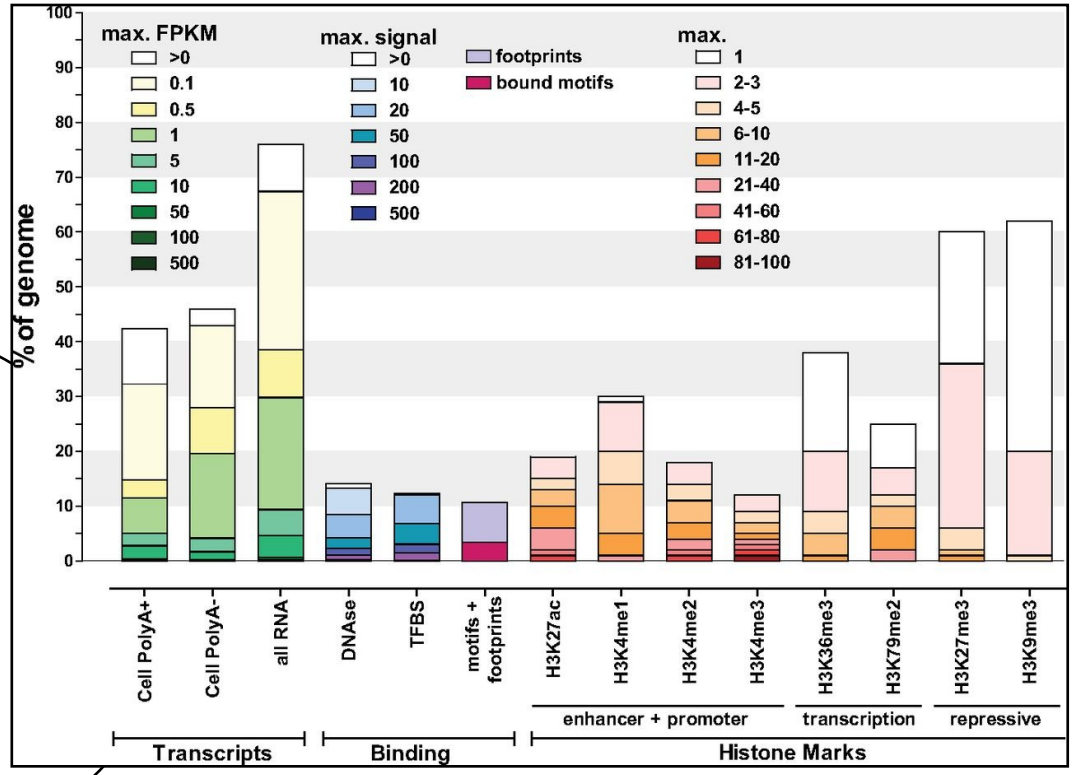
Defining functional DNA elements in the human genome

Manolis Kellis[a,b,1,2], Barbara Wold[c,2], Michael P. Snyder[d,2], Bradley E. Bernstein[b,e,f,2], Anshul Kundaje[a,b,3], Georgi K. Marinov[c,3], Lucas D. Ward[a,b,3], Ewan Birney[g], Gregory E. Crawford[h], Job Dekker[i], Ian Dunham[g], Laura L. Elnitski[j], Peggy J. Farnham[k], Elise A. Feingold[l], Mark Gerstein[l], Morgan C. Giddings[m], David M. Gilbert[n], Thomas R. Gingeras[o], Eric D. Green[j], Roderic Guigo[p], Tim Hubbard[q], Jim Kent[r], Jason D. Lieb[s], Richard M. Myers[t], Michael J. Pazin[j], Bing Ren[u], John A. Stamatoyannopoulos[v], Zhiping Weng[i], Kevin P. White[w], and Ross C. Hardison[x,1,2]

How much of our DNA is 'junk'?

OR

Can we identify the location of functional genomic elements?



With the completion of the human genome sequence, attention turned to identifying and annotating its functional DNA elements. As a complement to genetic and comparative genomics approaches, the Encyclopedia of DNA Elements Project was launched to contribute maps of RNA transcripts, transcriptional regulator binding sites, and chromatin states in many cell types. The resulting genome-wide data reveal sites of biochemical activity with high positional resolution and cell type specificity that facilitate studies of gene regulation and interpretation of noncoding variants associated with human disease. [redacted] raising the question of whether nonconserved but biochemically active regions are truly functional. Here, we review the strengths and limitations of biochemical, evolutionary, and genetic approaches for defining functional DNA segments, potential sources for the observed differences in estimated genomic coverage, and the biological implications of these discrepancies. We also analyze the relationship between signal intensity, genomic coverage, and evolutionary conservation. Our results reinforce the principle that each approach provides complementary information and that we need to use combinations of all three to elucidate genome function in human biology and disease.

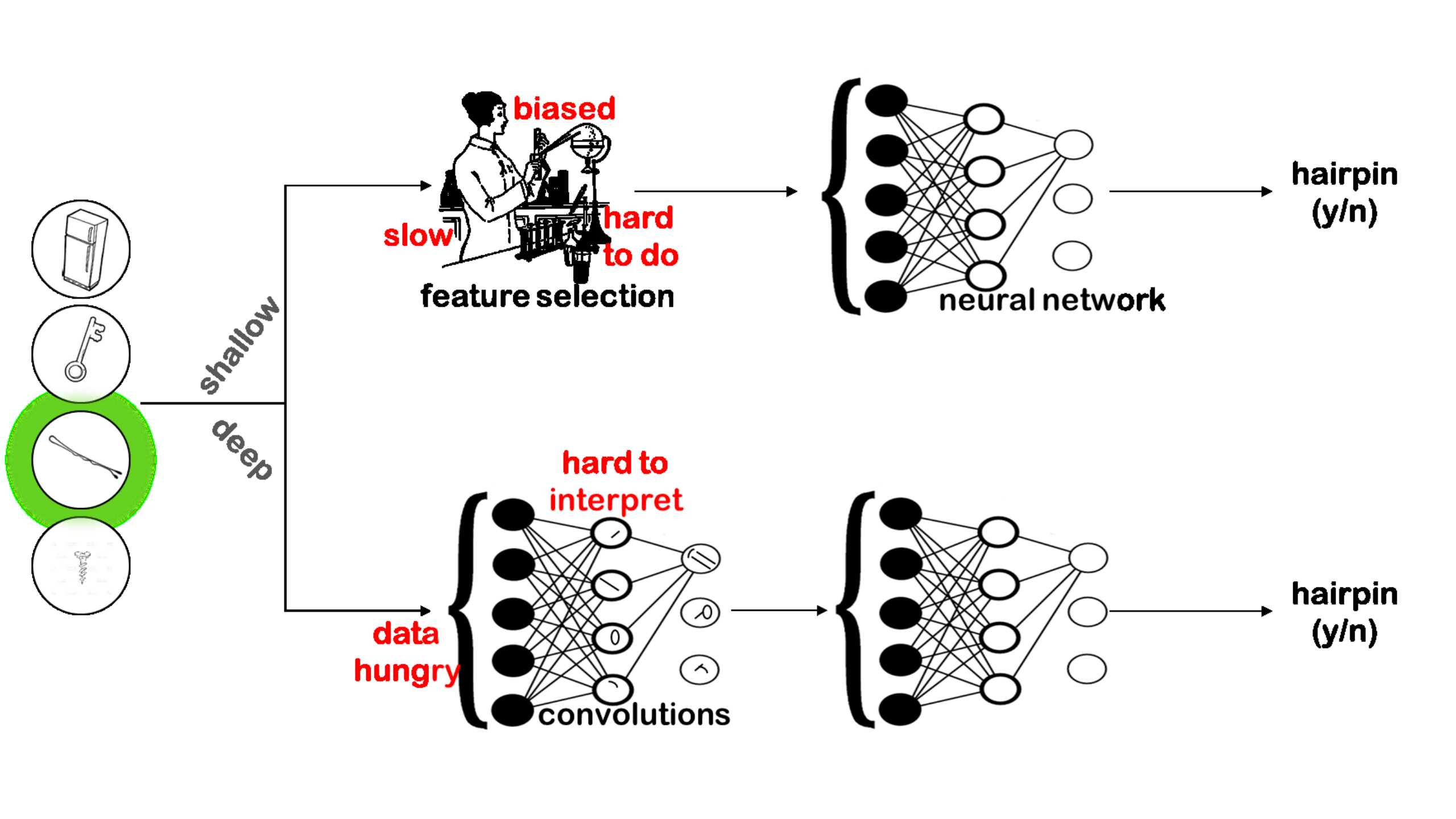
CTGTGGTGCTCAACTGTGATTCCTTTTCACA
TTCACCCTGGATGTTCTCTTCACTGTGGGAT
GAGGTAGTAGGTTGTATAGTTTTAGGGTCA
CACCCACCACTGGGAGATAACTATACAATCT
ACTGTCTTTCCTAACGTGATAGAAAAGTCTG
CATCCAGGCGGTCTGATAGAAAGTCAGTTA
ACTAATTGTACAATATCTGTGGTGCTCAACT
GTGATTCCTTTTCACCATTCACCCTGGATGTT
CTCTTCACTGTGGGATGAGGTAGTAGGTTGT
ATAGT**TTTAGGGTCACACCCACCAC**TGGGA
GATAACTATACAATCTACTGTCTTTCCTAACG
TGATAGAAAAGTCTGCATCCAGGCGGTCTG
ATAGAAAGTCAGTTAACTAATTGTACAATA
TCTGTGGTGCTCAACTGTGATTCCTTTTCAC
CATTCACCCTGGATGTTCTCTTCACTGTGGG
ATGAGGTAGTAGGTTGTATAGTTTTAGGGTC
ACACCCACCACTGGGAGATAACTATACAATC
TACTGTCTTTCCTAACGTGATAGAAAATGCA
GTCTGCATCCAGGCGGTCTGATAGAAAGGG
AGTCAGTTAACTAATTGTACAACTCCTTATAT
ATATTCTGCATCCAGGCGGTCTCTTATAAGC
CTGCATCCAGGCGGTCGCGGTAGTATTAGT
TTAGGGTCATTAGGGTCAGTCCTATTAGTAC

Find a hairpin in a junkyard

background
Variability

class
Imbalance

unknown
Features

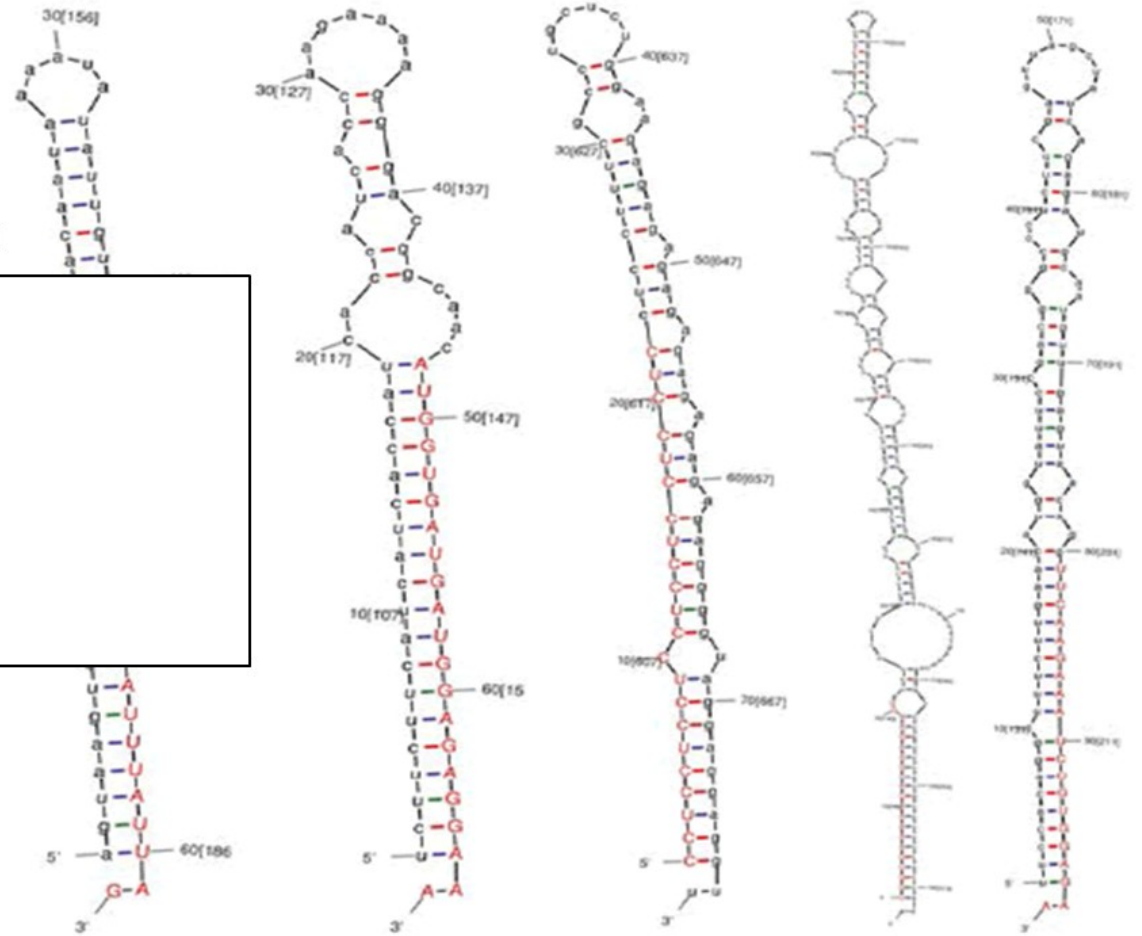

shallow
deep
biased
slow
hard
to do
feature selection
neural network
hairpin
(y/n)
hard to
interpret
data
hungry
convolutions
hairpin
(y/n)

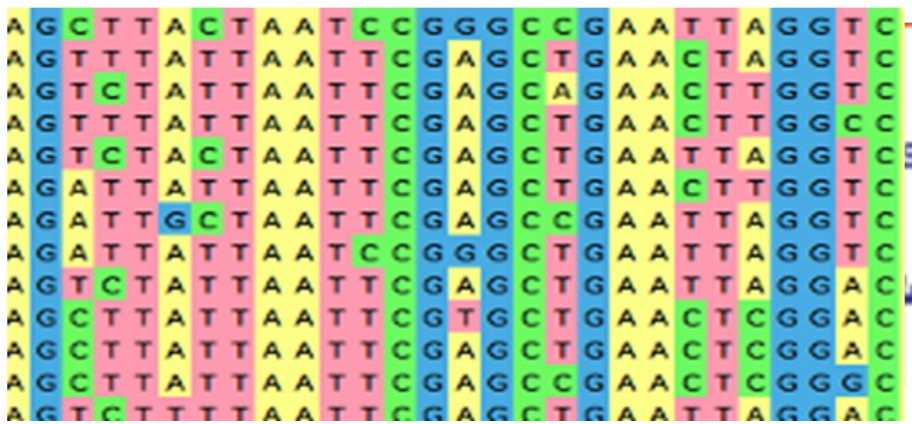
pre-miRNA genomic locus identification

- Known hairpin structure
- High conservation with distinct pattern
- Some sequence preferences
- Approx. 2000 known in human

=

Should be an easy task!?





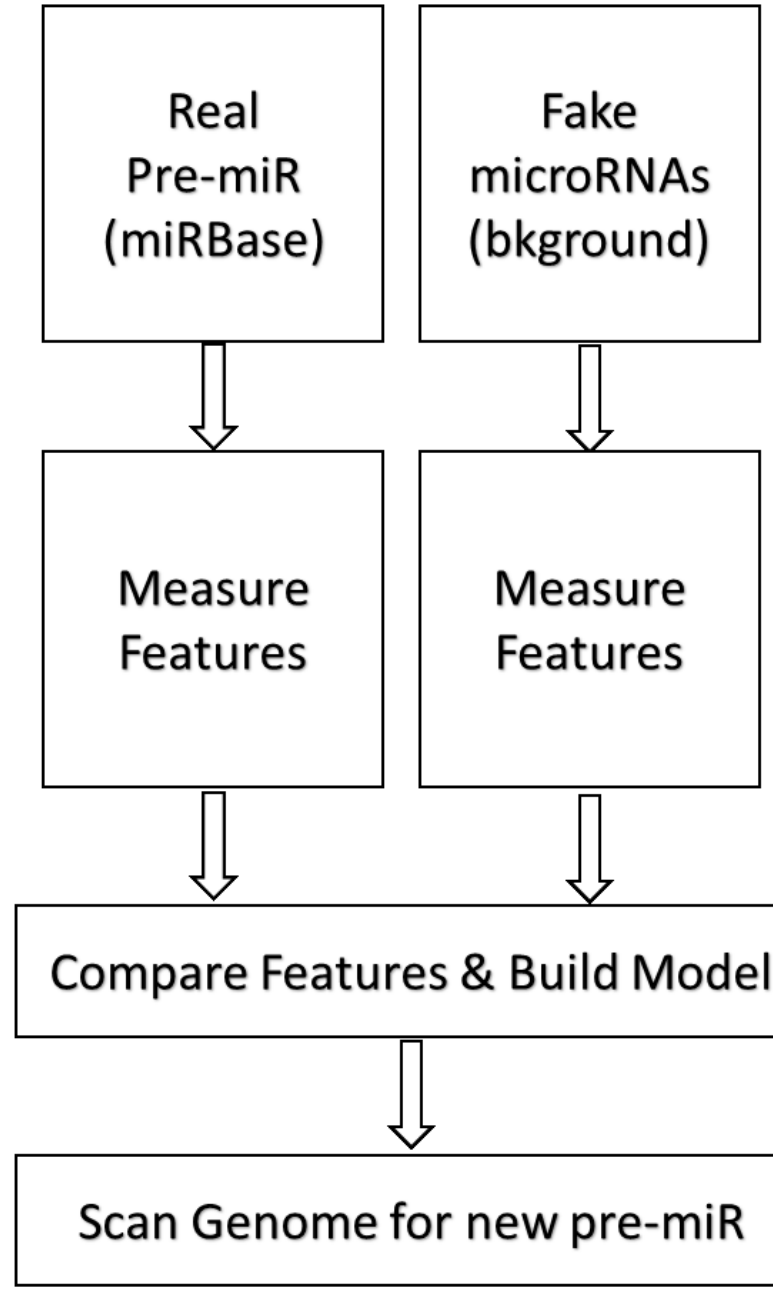
Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W339–W344***
doi:10.1093/nar/gkm368

MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features

Peng Jiang, Haonan Wu, Wenkai Wang, Wei Ma, Xiao Sun and Zuhong Lu*

State Key Laboratory of Bioelectronics, Department of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096, P. R. China

Received January 18, 2007; Revised and Accepted April 26, 2007



The *P*-value of randomization test feature

In order to determine if the MFE value is significantly different from that of random sequences, a Monte Carlo randomization test was used (22). The test can be summarized as follows:

i. Compute MFE of the secondary structure inferred from the original sequence.

ii. Randomize the order of the nucleotides in the original sequence while keeping the dinucleotide distribution (or frequencies) constant. Then compute the MFE for the inferred structure based on the shuffled sequence.

iii. Repeat step 2 a great number of times (1000) in order to build the distribution of MFE values.

iv. If N is the number of iterations and R the number of randomized sequences that have a MFE value less or equal to the original value, then P-value is defined as:

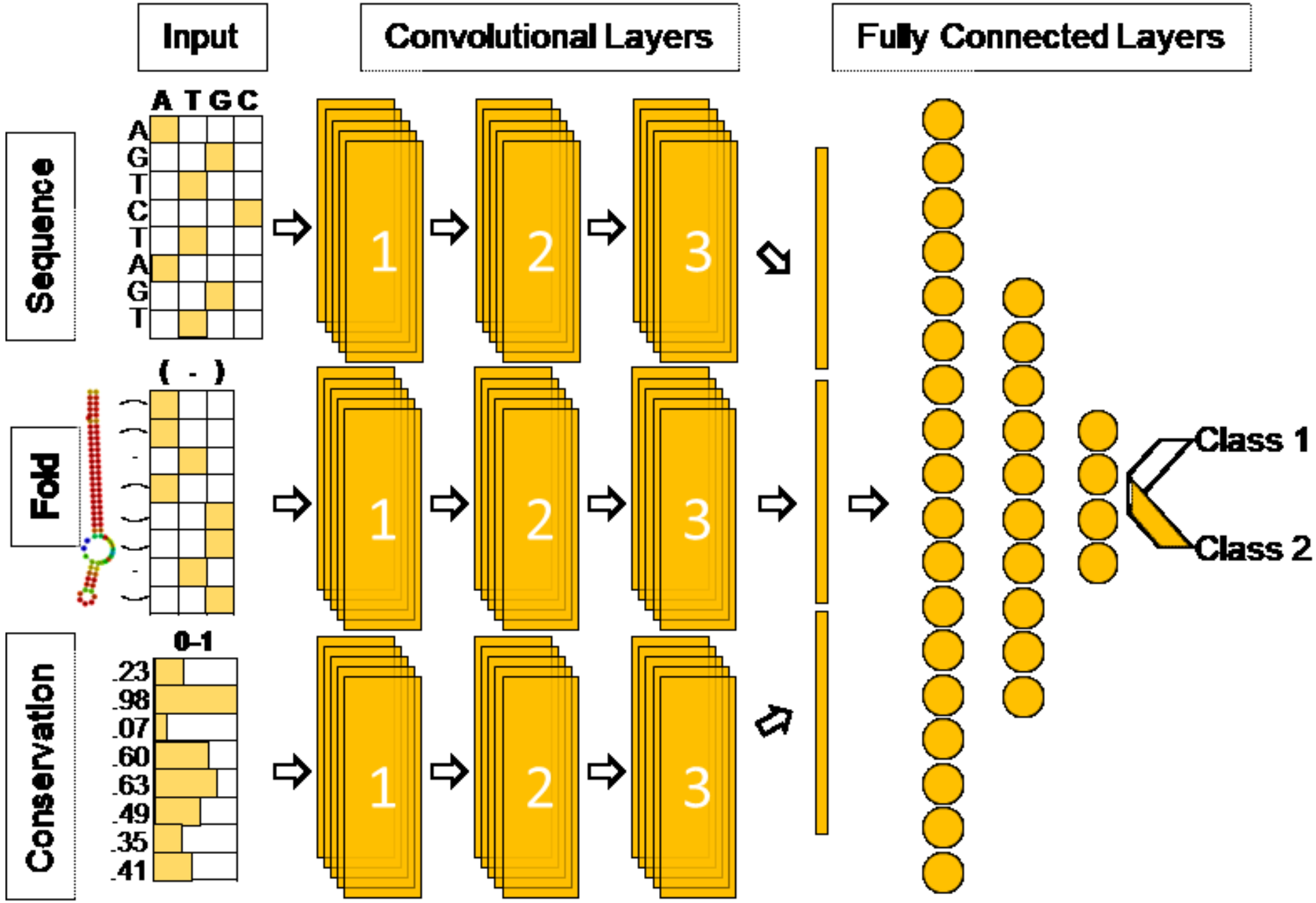
$$P = \frac{R}{N+1}$$

Features	Sp (%)	Se (%)	ACC (%)	MCC
A	90.48	85.89	88.21	0.77
A + B	95.24	91.41	93.35	0.87
A + C	97.62	94.47	96.07	0.92
A + B + C	98.21	95.09	96.68	0.94

A: local contiguous triplet structure composition;

B: Minimum of free energy (MFE) of the secondary structure;

C: *P*-value.


Input
Convolutional Layers
Fully Connected Layers
Sequence
A T G C
A
G
T
C
T
A
G
T
Fold
(-)
Conservation
0-1
.23
.98
.07
.60
.63
.49
.35
.41
1
2
3
Class 1
Class 2

Nucleic Acids Research, 2007, Vol. 35, Web Server issue **W339–W344**
doi:10.1093/nar/gkm368

MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features

Peng Jiang, Haonan Wu, Wenkai Wang, Wei Ma, Xiao Sun and Zuhong Lu*

State Key Laboratory of Bioelectronics, Department of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096, P. R. China

Received January 18, 2007; Revised and Accepted April 26, 2007

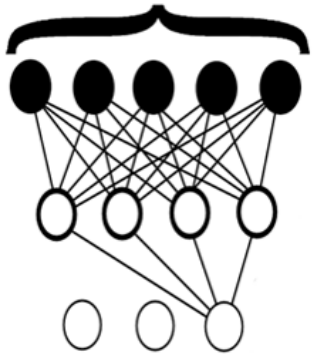
440 citations

Features	Sp (%)	Se (%)	ACC (%)	MCC
A	90.48	85.89	88.21	0.77
A + B	95.24	91.41	93.35	0.87
A + C	97.62	94.47	96.07	0.92
A + B + C	98.21	95.09	96.68	0.94

A: local contiguous triplet structure composition;

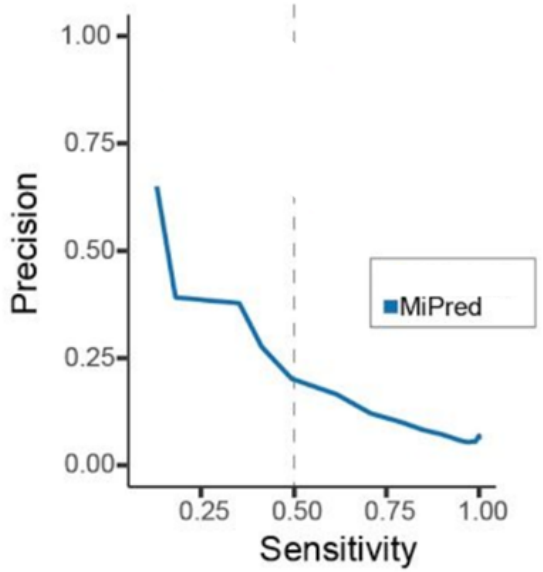
B: Minimum of free energy (MFE) of the secondary structure;

C: *P*-value.

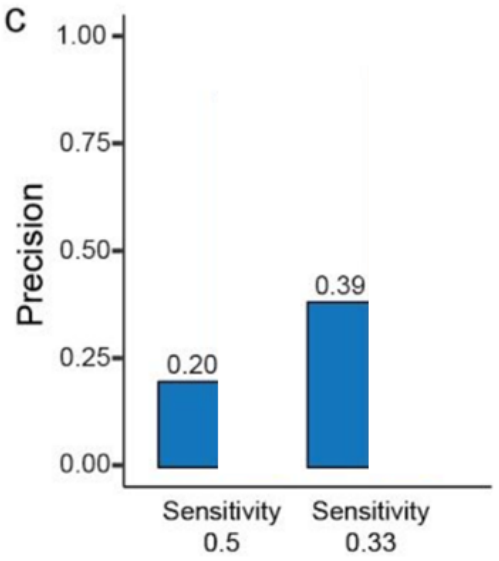


a

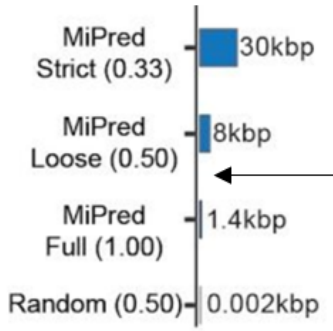
b



c

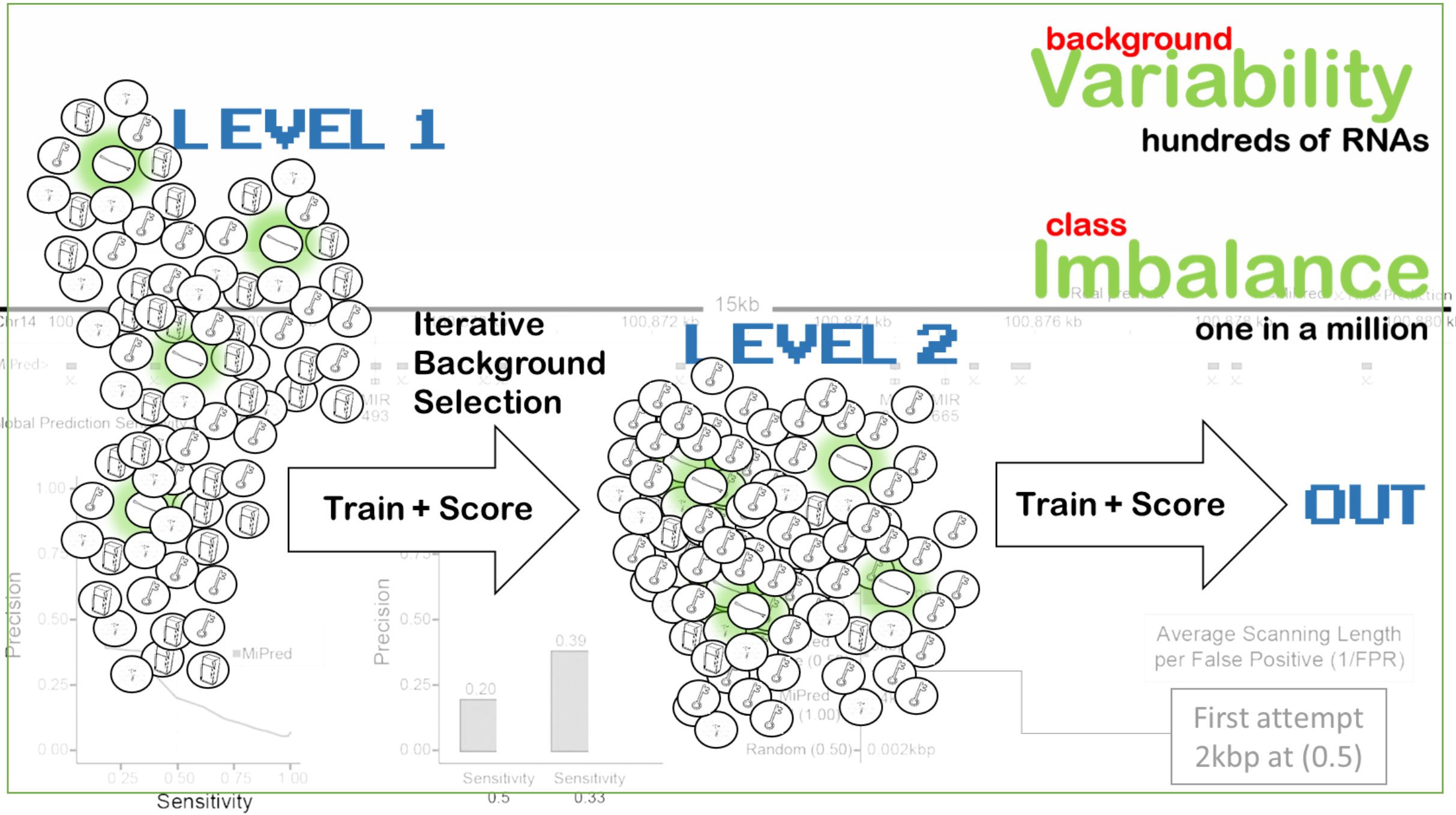


d



Average Scanning Length per False Positive (1/FPR)

First attempt 2kbp at (0.5)


background
Variability
hundreds of RNAs
class
Imbalance
one in a million
LEVEL 1
LEVEL 2
Iterative
Background
Selection
Train + Score
Train + Score
OUT
a
b
15kb
Chr14
100,872 kb
100,874 kb
100,876 kb
Precision
Sensitivity
MiPred
0.20
0.39
Sensitivity 0.5
Sensitivity 0.33
Average Scanning Length
per False Positive (1/FPR)
First attempt
2kbp at (0.5)
Random (0.50)
0.002kbp

Nucleic Acids Research, 2007, Vol. 35, Web Server issue W339–W344
doi:10.1093/nar/gkm368

MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features

Peng Jiang, Haonan Wu, Wenkai Wang, Wei Ma, Xiao Sun and Zuhong Lu*

State Key Laboratory of Bioelectronics, Department of Biological Science and Medical Engineering, Southeast University, Nanjing, 210096, P. R. China

Received January 18, 2007; Revised and Accepted April 26, 2007

www.nature.com/scientificreports

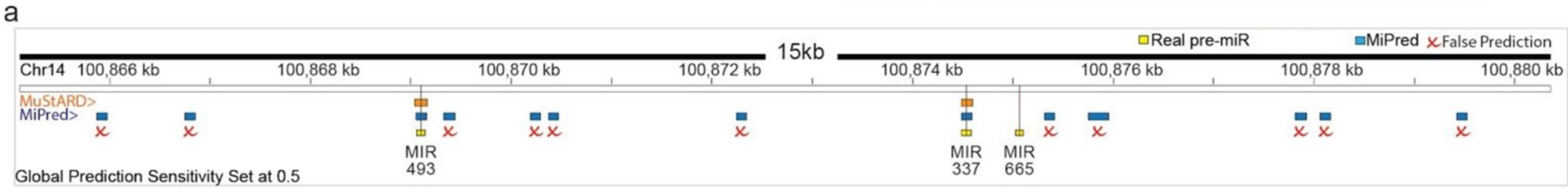
SCIENTIFIC REPORTS
nature research

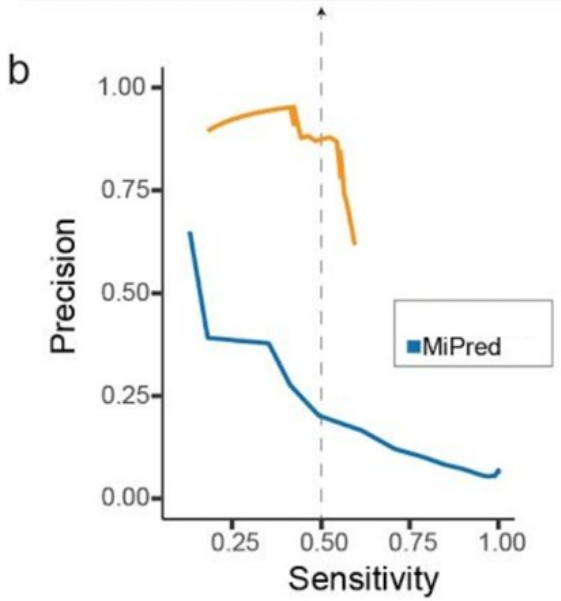
Check for updates

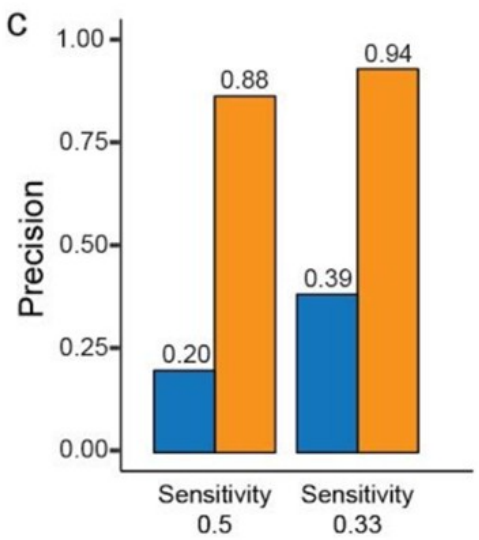
OPEN

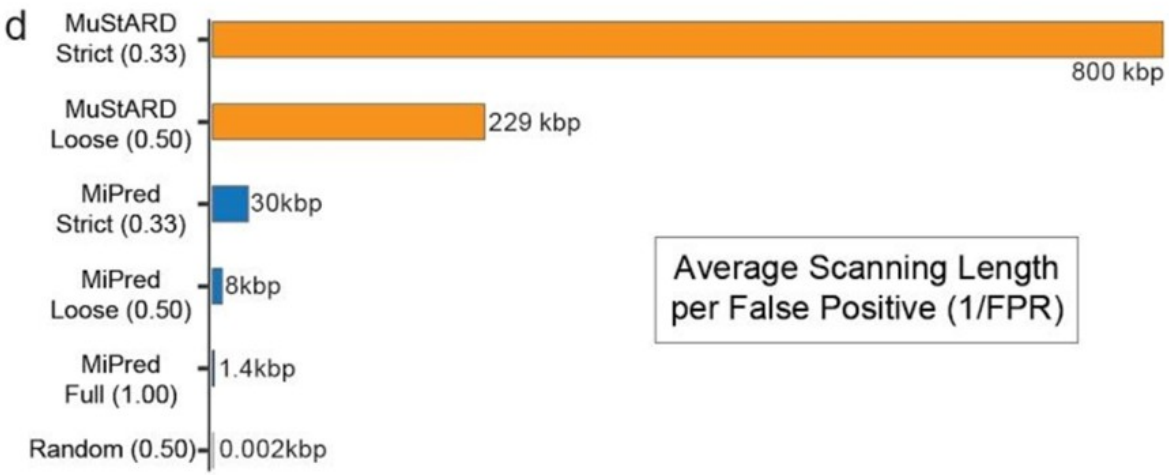
Multi-branch Convolutional Neural Network for Identification of Small Non-coding RNA genomic loci

Georgios K. Georgakilas[1], Andrea Grioni[1], Konstantinos G. Liakos[3], Eliska Chalupova[2], Fotis C. Plessas[3] & Panagiotis Alexiou[1]✉

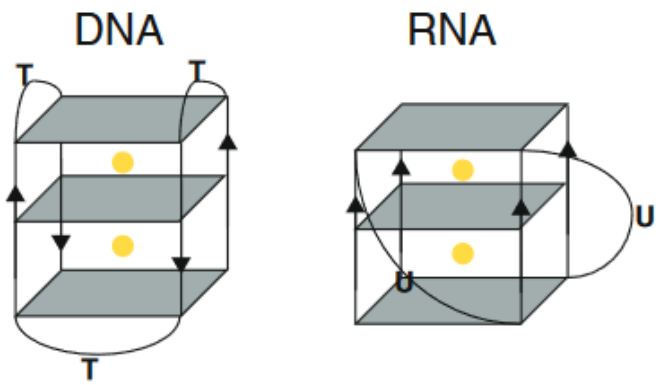


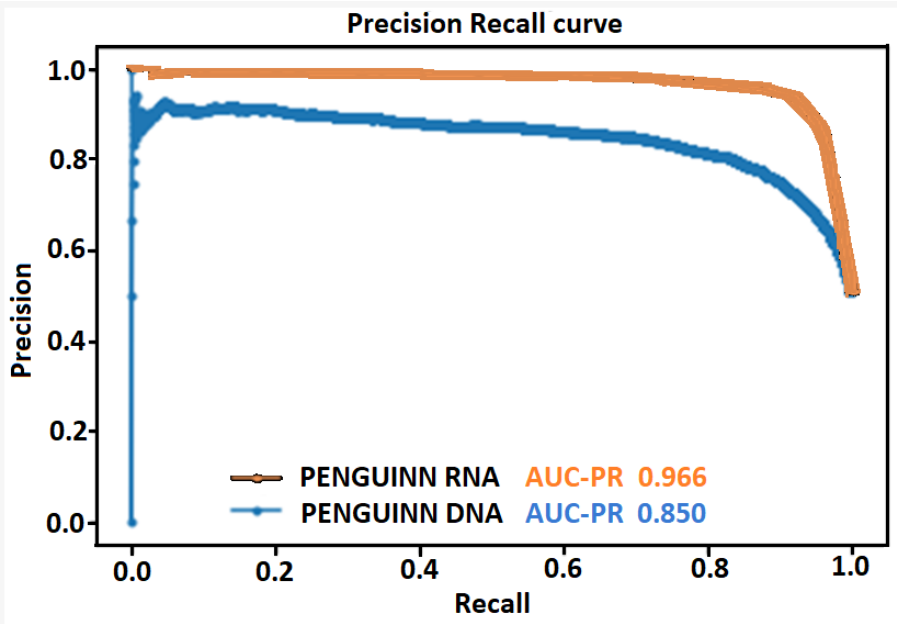






GACGGCAGGGGAGGAGCCGGGTCCACTGCCGGGTGGAGGGCaAGGCGAGTGTGTGTCCTTATCCTAGCAATTGGGG
CGCGGGCCTGTGAGCCAGTTGGAGTTGCGGCGGCGGGAACGATTGGGCTGAGCAGAGGACGACATGTTGCTTTTCGT
GGAGGTGAGTGCATTATGCTAGTCTCGTCCTGCTCTTAGGAGAGCA



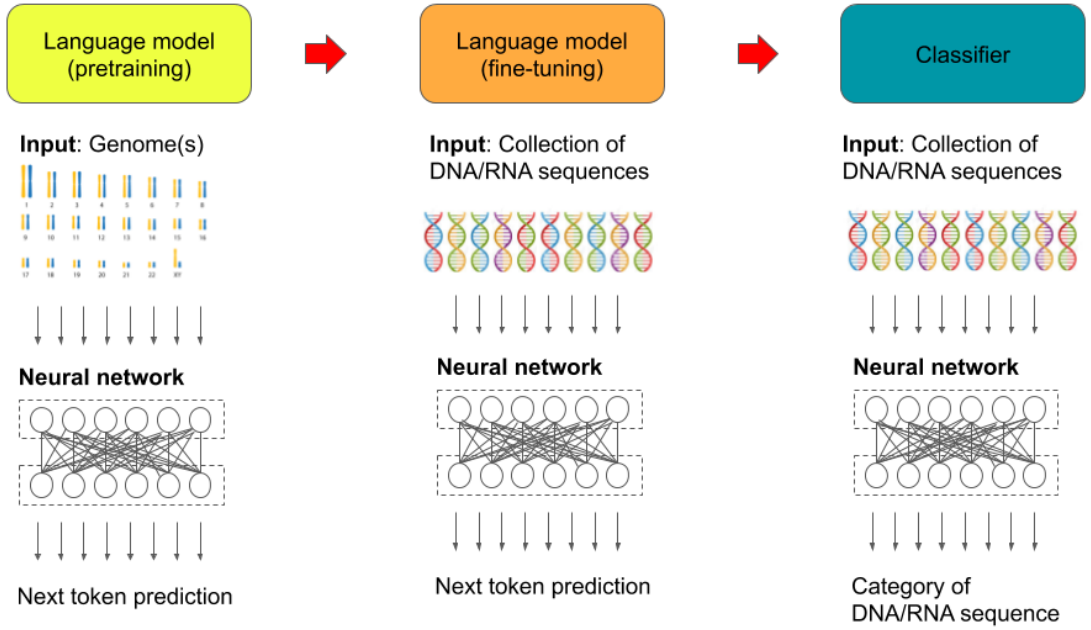


Genomic Annotation Benchmarks

- Ready to use genomic classification datasets (cleaned, train/test split)
- Get the benchmark to your machine with **one line of Python code**
- Pre-trained models can be used for transfer learning

(bit.ly/genbench)

Name	Number of seqs	Seq length	Baseline model accuracy
Human non-TATA promoters	36131	251	84.5%
Human enhancers	28000	500	87.9%
Coding vs. intergenic	100000	200	84.8%



Friday 19th November 2021: Hackathon! (in hybrid mode) – ALL are welcome – Email Panagiotis Alexiou for details

A)

ENNGene

Select a task to be run:

Preprocessing

Documentation

FAQ

GitHub

CEITEC

MASARYK UNIVERSITY

Preprocessing

Load parameters from a previous run

Output folder (result files will be exported here; home directory used as default)

/home/eliska/enngene_output

Use already preprocessed file from a previous run

Branches

Choose an option

Window size

100 - +

Seed for semi-random window placement upon the sequences

42 - +

Input Coordinate Files

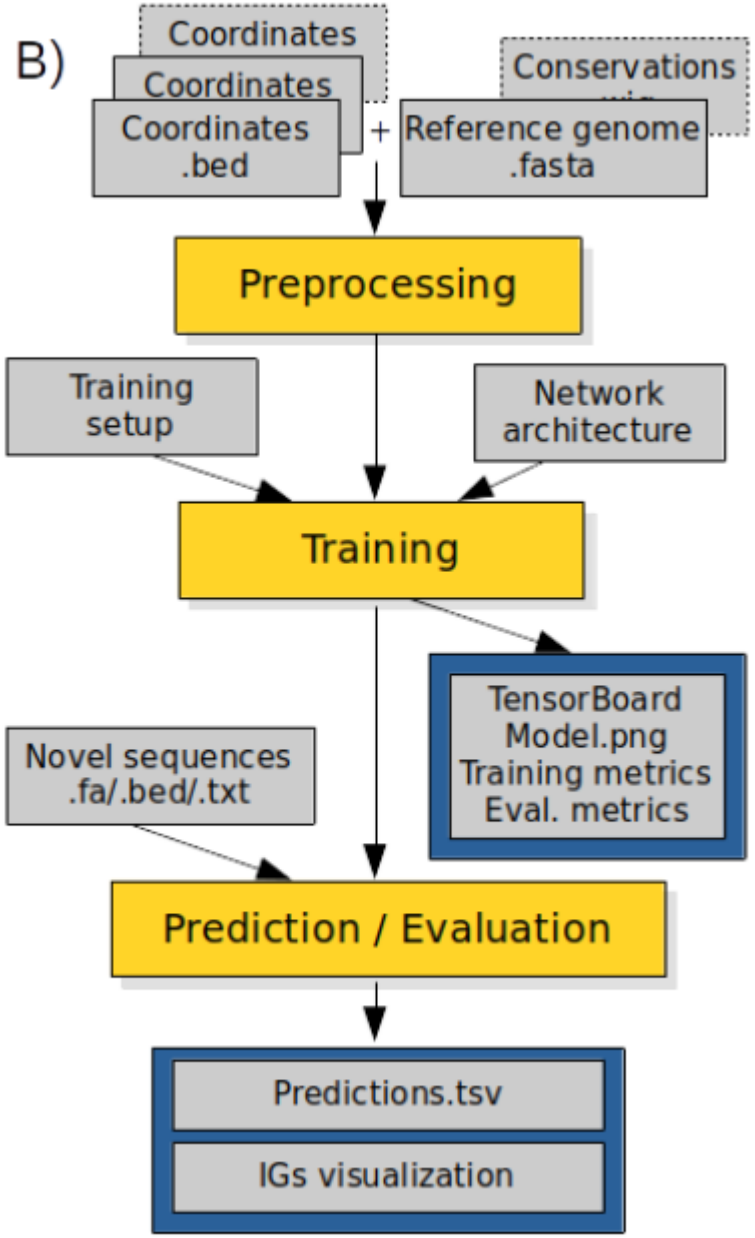
Number of input files (= no. of classes):

2 - +

File no. 1 (.bed)

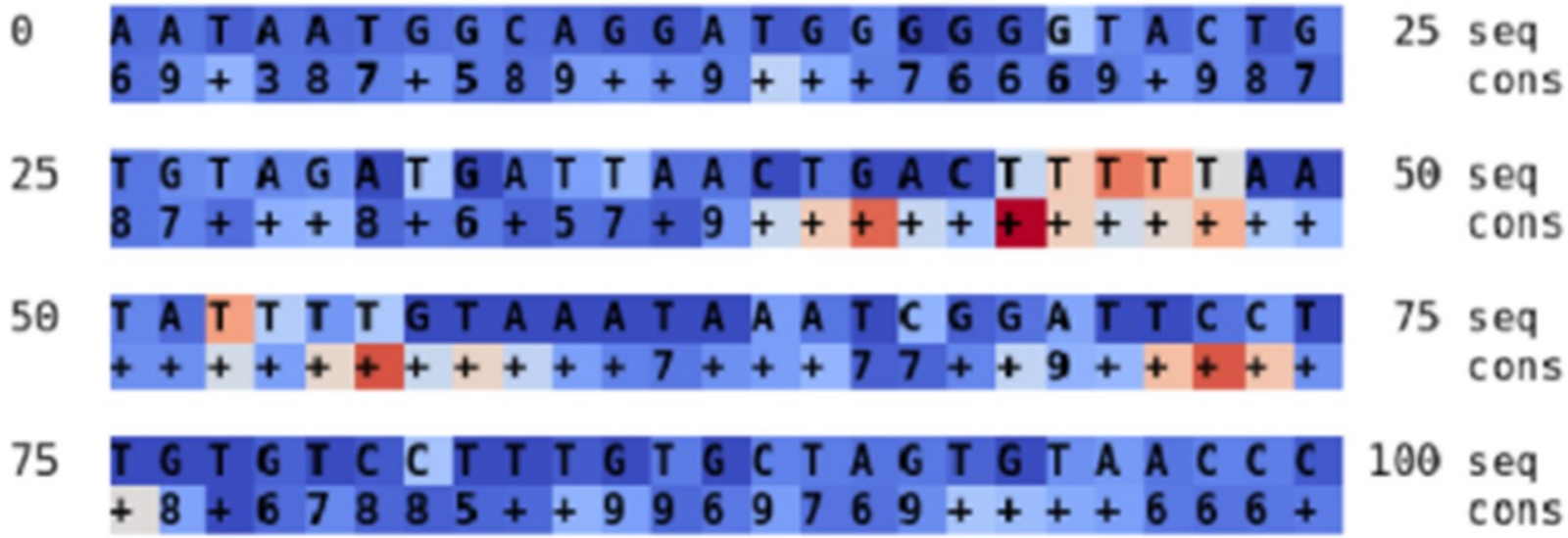
File no. 2 (.bed)

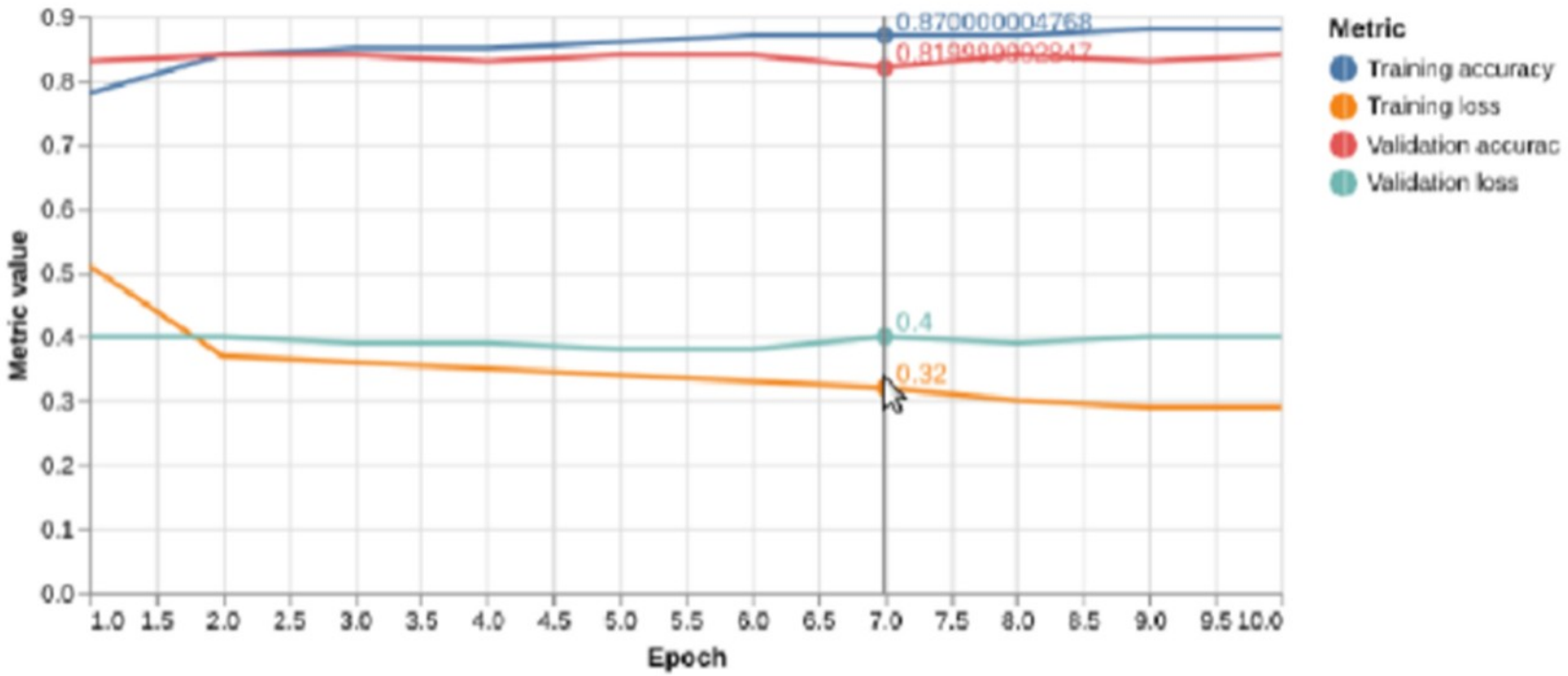
B)



https://bit.ly/ENNGene

Integrated Gradients Interpretation





Evaluation of trained model

https://bit.ly/ENNGene

Genomic or Transcriptomic functional elements in need of identification

RNA Binding Proteins

miRNA targets

Small RNA Loci

Enhancers

Transcription Factor Binding Sites

RNA Modification Sites

Non-coding RNAs

Machines Learning what makes Biology tick

Thank you for your attention!

Panagiotis Alexiou
CEITEC-MU
Brno, CZ

Group Leader

Panagiotis Alexiou

Postdoc

Petr Simecek

postDoc

NEW MEMBER

Ilektra Giassa

postDoc

NEW MEMBER

Tomas Majtner

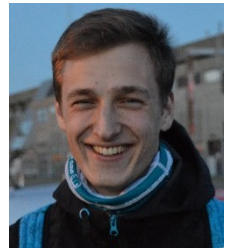
PHD Student

NEW MEMBER

Vlastimil Martinek

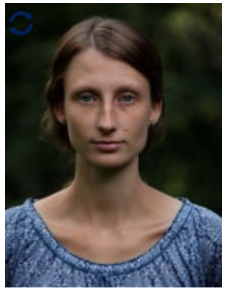
PHD Student

NEW MEMBER



David Cechak

PHD Student

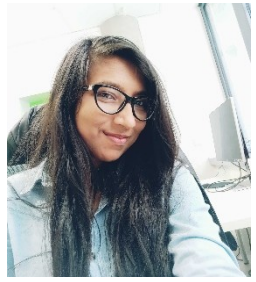


Eliska Chalupova

PHD Student

Ondrej Vaculik

PHD Student



Kriti Bhaghat

PHD Student

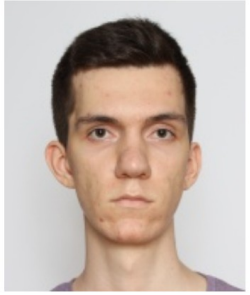
NEW MEMBER

Katarina Gresova

student

Eva Klimentova

student



Jakub Polacek