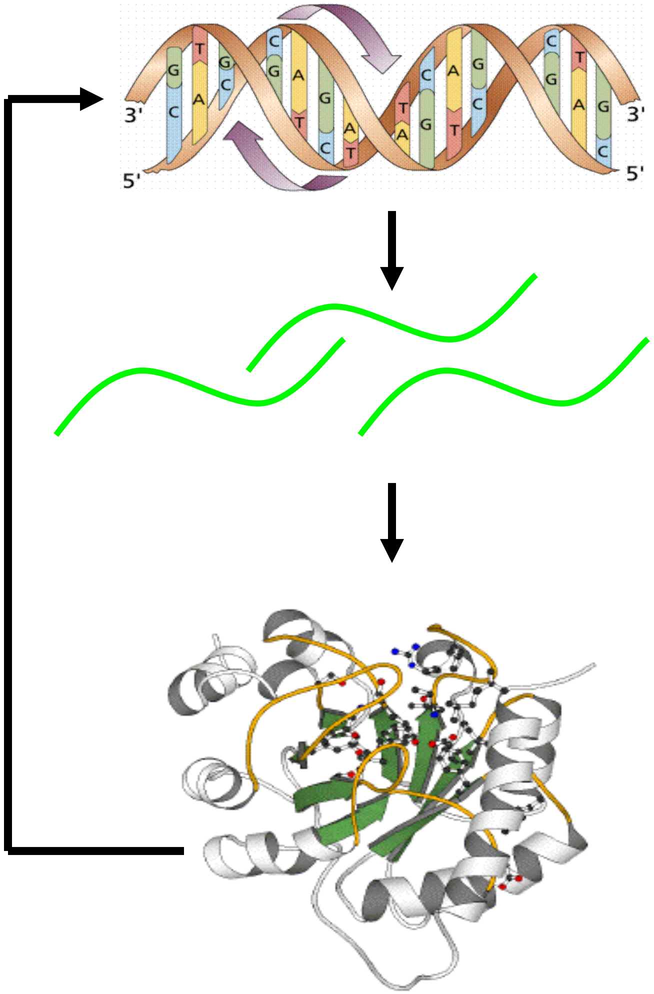
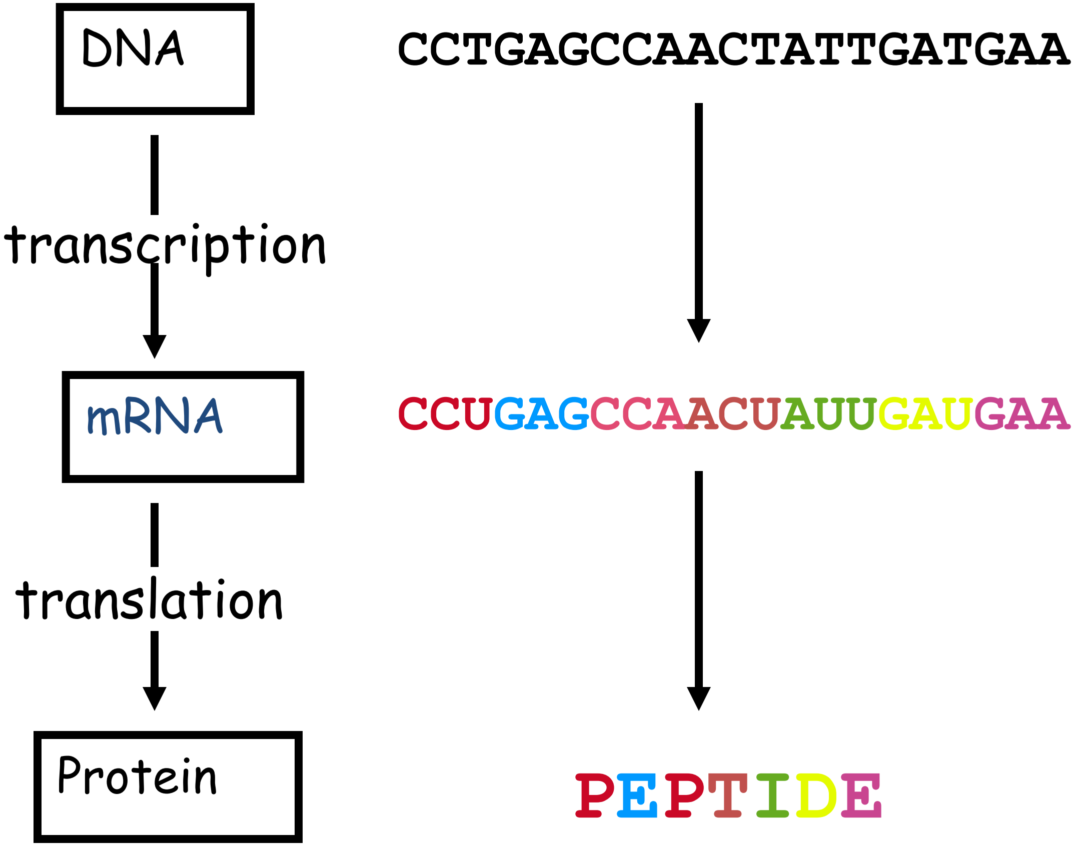
10-601
Machine Learning

HMM applications in computational biology

Central dogma





DNA

transcription

mRNA

translation

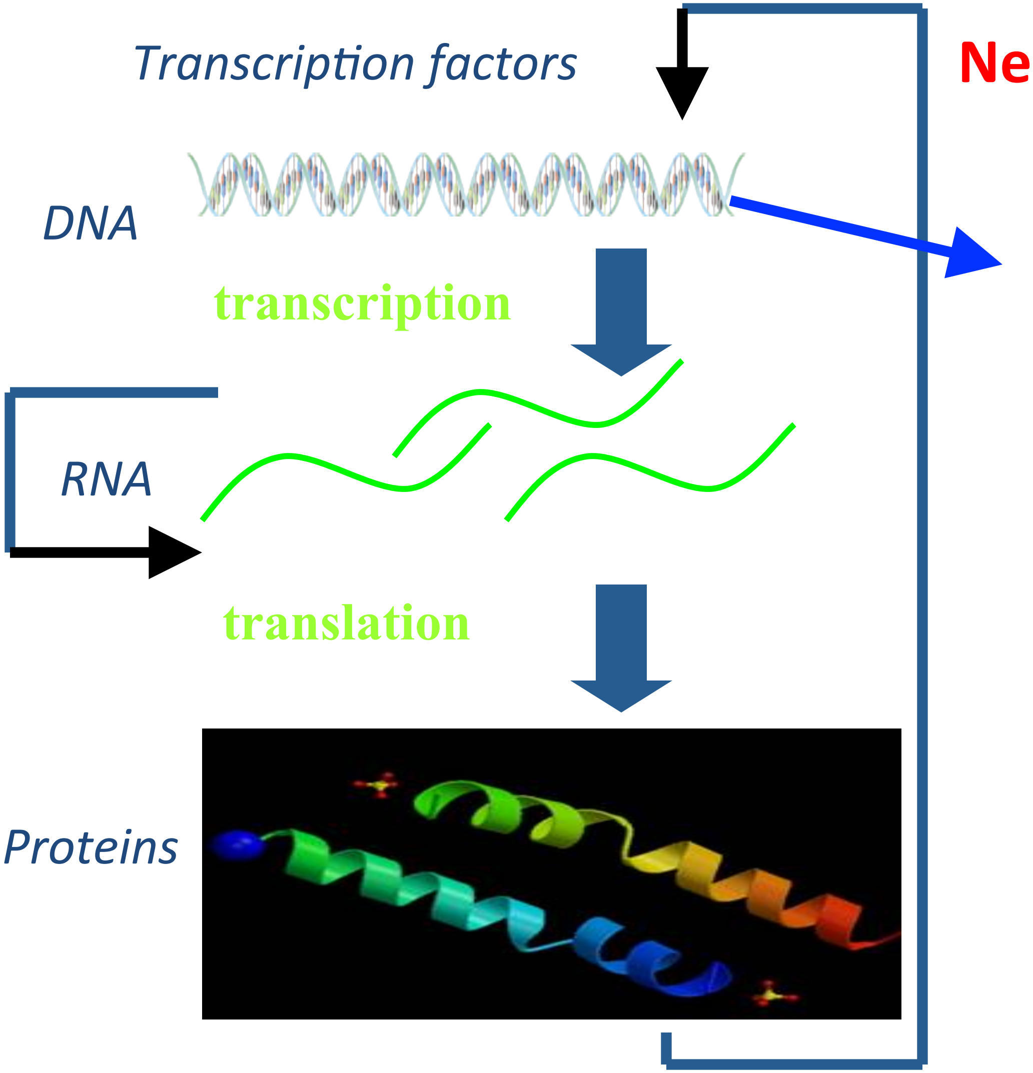
Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

Biological data is rapidly accumulating



Transcription factors

DNA

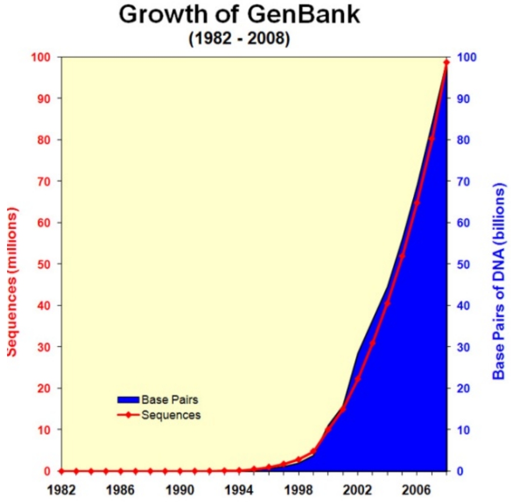
transcription

RNA

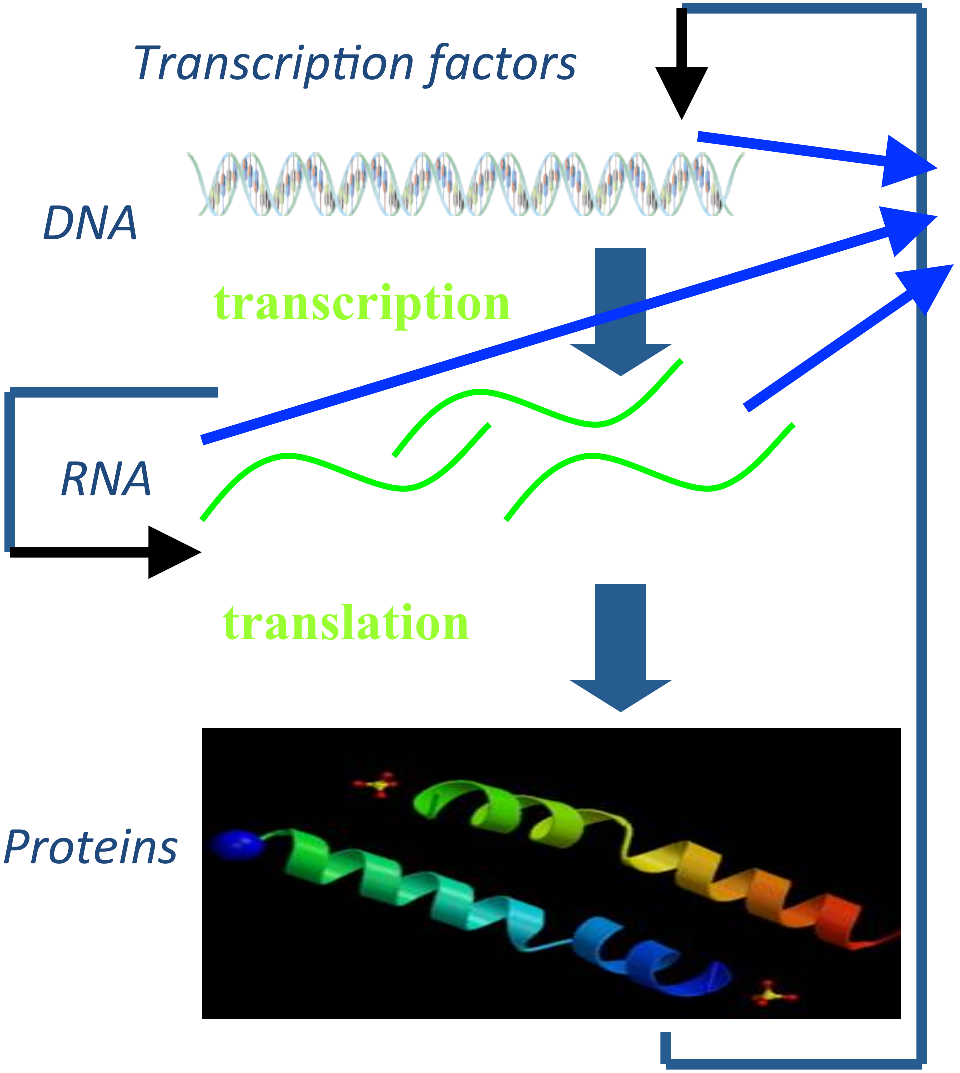
translation

Proteins

Next generation sequencing



Biological data is rapidly accumulating



Transcription factors

DNA

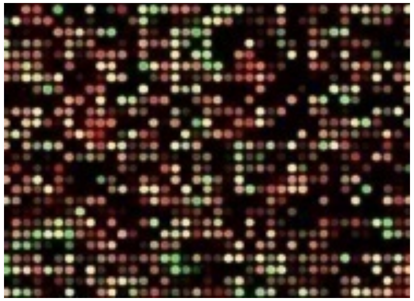
transcription

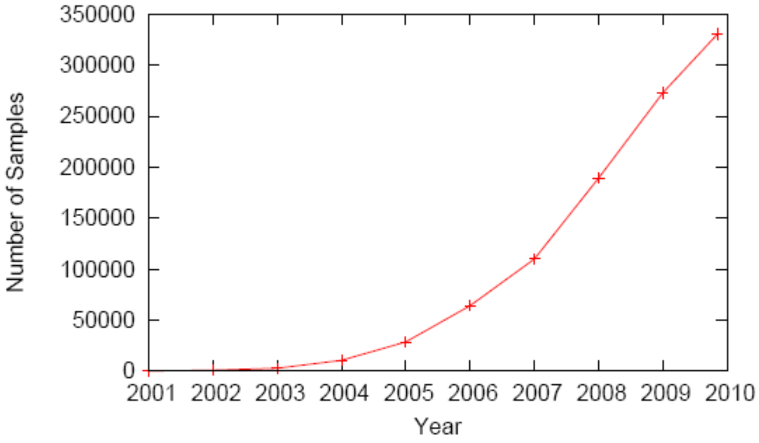
RNA

translation

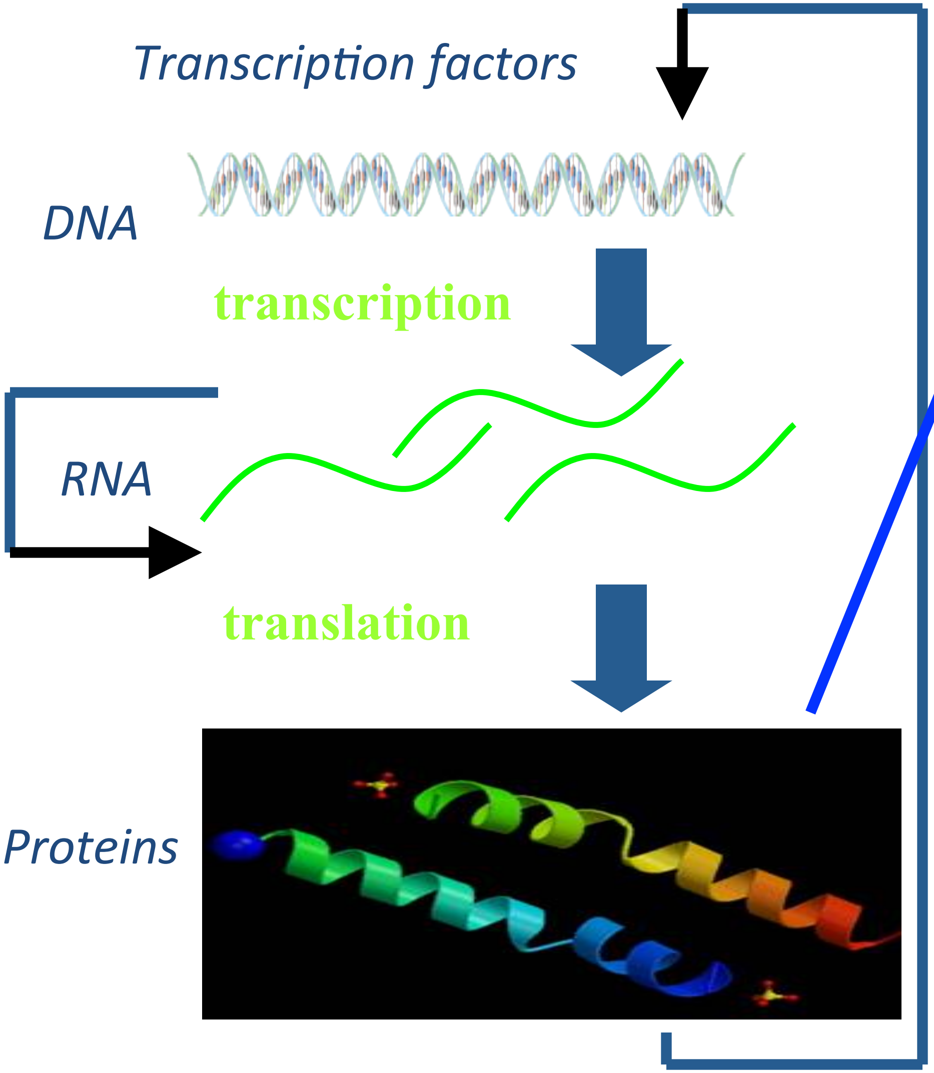
Proteins

Array / sequencing technology





Biological data is rapidly accumulating



Transcription factors

DNA

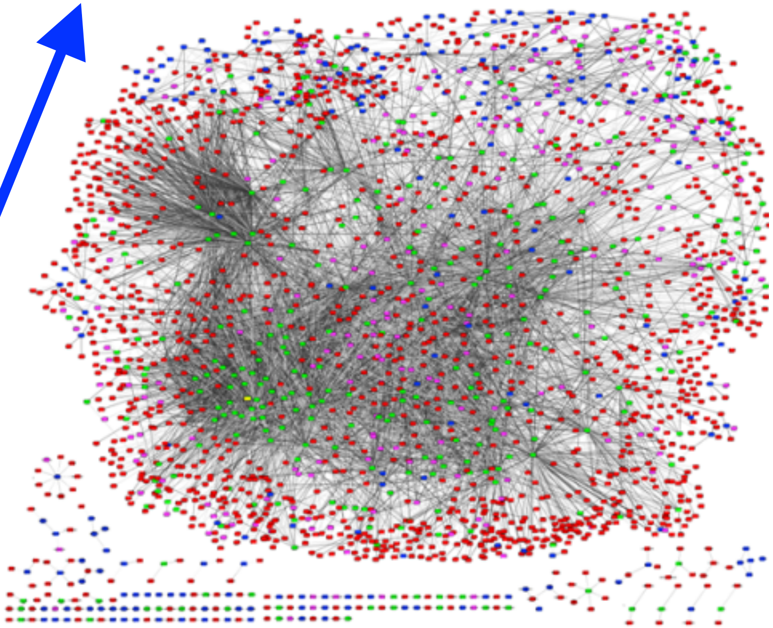
transcription

RNA

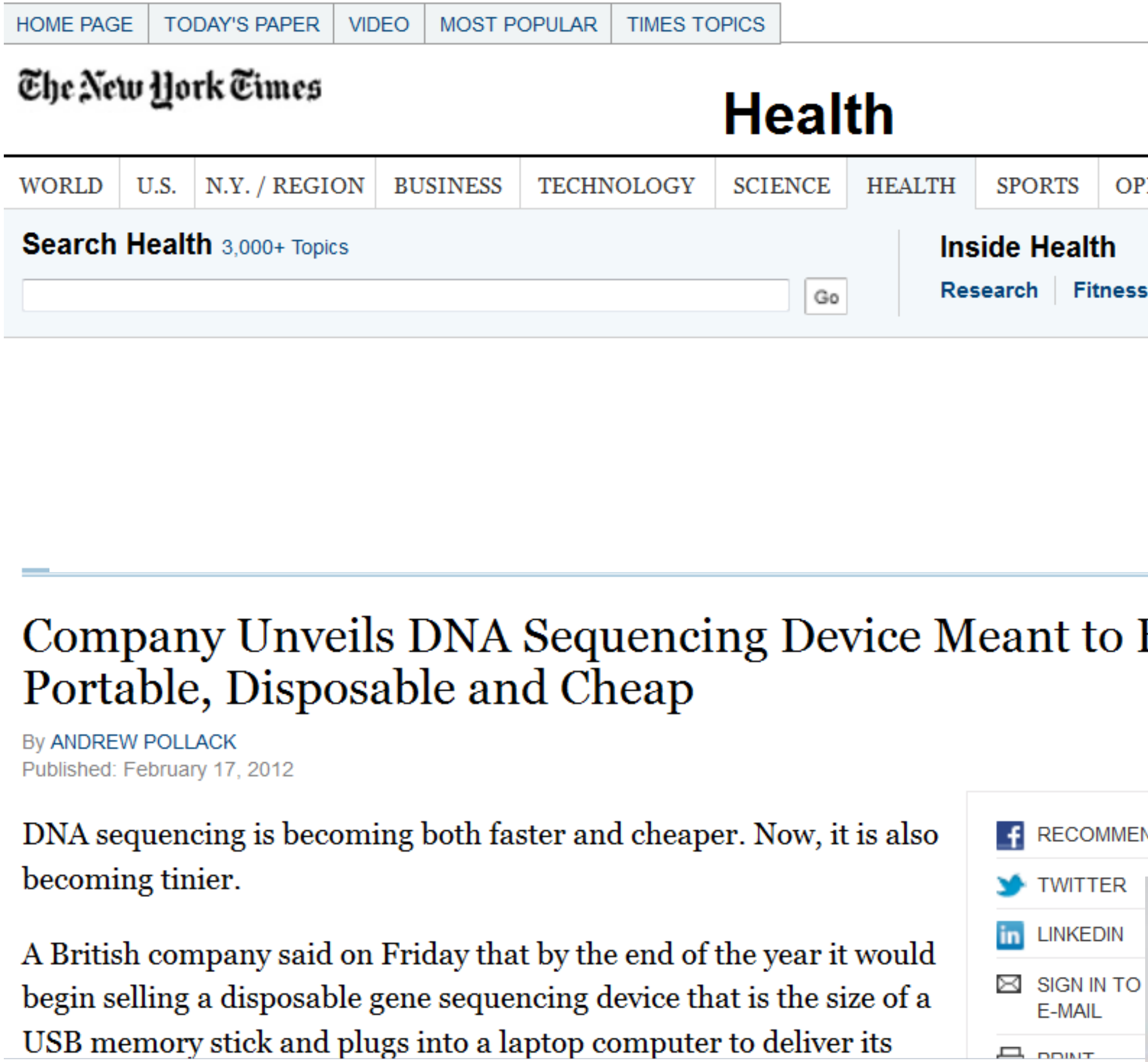
translation

Proteins

Protein interactions



- 38,000 identified interactions
- Hundreds of thousands of predictions



Company Unveils DNA Sequencing Device Meant to Be Portable, Disposable and Cheap

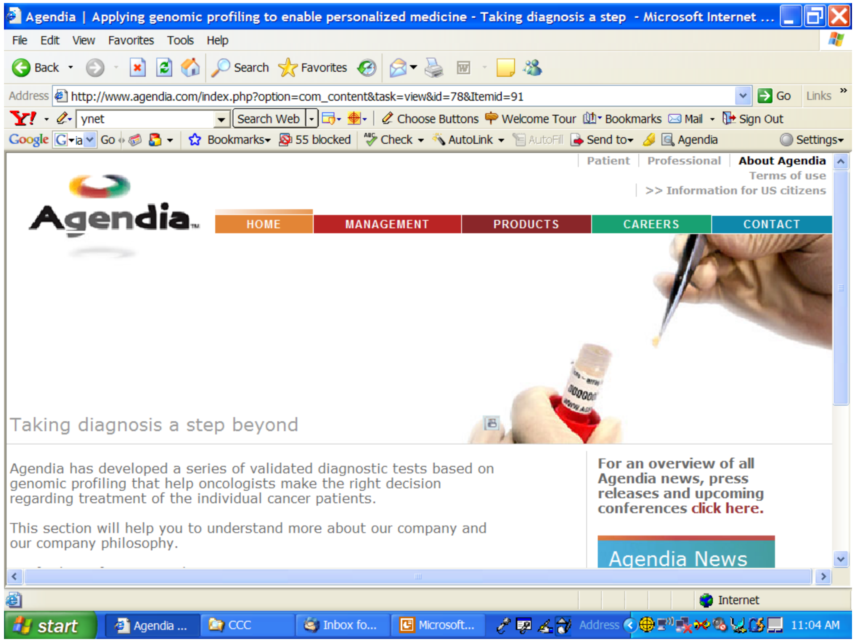
By ANDREW POLLACK
Published: February 17, 2012

DNA sequencing is becoming both faster and cheaper. Now, it is also becoming tinier.

A British company said on Friday that by the end of the year it would begin selling a disposable gene sequencing device that is the size of a USB memory stick and plugs into a laptop computer to deliver its

FDA Approves Gene-Based Breast Cancer Test*

" MammaPrint is a DNA microarray-based test that measures the activity of 70 genes in a sample of a woman's breast-cancer tumor and then uses a specific ***formula*** to determine whether the patient is deemed low risk or high risk for the spread of the cancer to another site."


Agendia | Applying genomic profiling to enable personalized medicine - Taking diagnosis a step - Microsoft Internet ...
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.agendia.com/index.php?option=com_content&task=view&id=78&Itemid=91
Go Links
ynet Search Web Choose Buttons Welcome Tour Bookmarks Mail Sign Out
Google Go Bookmarks 55 blocked Check AutoLink AutoFill Send to Agendia Settings
Patient | Professional | About Agendia
Terms of use
>> Information for US citizens
Agendia
HOME
MANAGEMENT
PRODUCTS
CAREERS
CONTACT
Agendia News
Internet
start
Agendia ...
CCC
Inbox fo...
Microsoft...
Address
11:04 AM

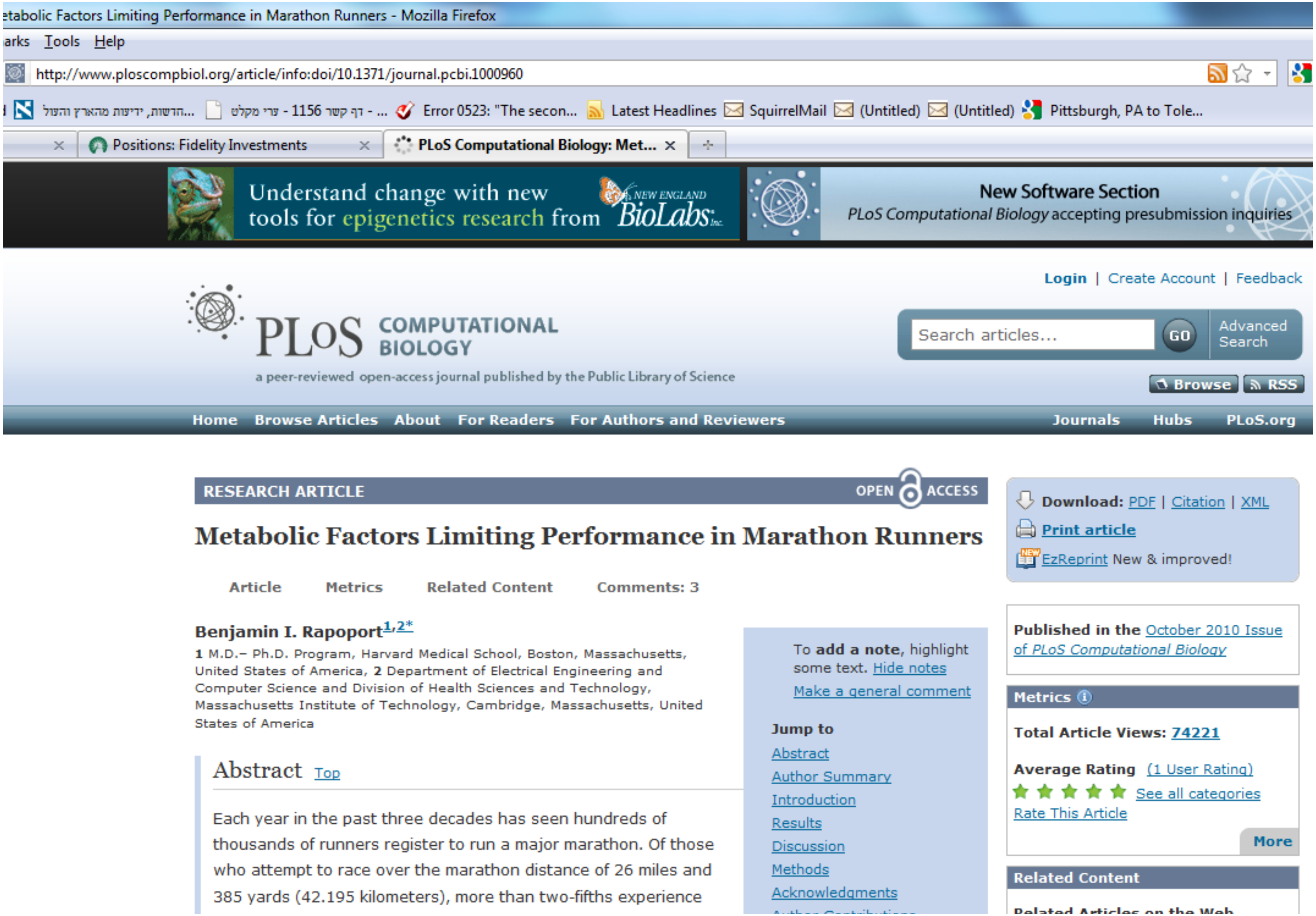
Taking diagnosis a step beyond

Agendia has developed a series of validated diagnostic tests based on genomic profiling that help oncologists make the right decision regarding treatment of the individual cancer patients.

This section will help you to understand more about our company and our company philosophy.

For an overview of all Agendia news, press releases and upcoming conferences click here.

*Washington Post, 2/06/2007


etabolic Factors Limiting Performance in Marathon Runners - Mozilla Firefox
arks Tools Help
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000960
הדשות, ידיעות מהארץ והעול...
דף קשר 1156 - ערי מקלט - ...
Error 0523: "The secon...
Latest Headlines
SquirrelMail
(Untitled)
(Untitled)
Pittsburgh, PA to Tole...
Positions: Fidelity Investments
PLoS Computational Biology: Met...
Understand change with new tools for epigenetics research from NEW ENGLAND BioLabs Inc.
New Software Section
PLoS Computational Biology accepting presubmission inquiries
Login | Create Account | Feedback
PLoS COMPUTATIONAL BIOLOGY
a peer-reviewed open-access journal published by the Public Library of Science
Search articles...
GO
Advanced Search
Browse
RSS
Home
Browse Articles
About
For Readers
For Authors and Reviewers
Journals
Hubs
PLoS.org
RESEARCH ARTICLE
OPEN ACCESS
Article
Metrics
Related Content
Comments: 3
Download: PDF | Citation | XML
Print article
EzReprint New & improved!
Jump to
Abstract
Author Summary
Introduction
Results
Discussion
Methods
Acknowledgments
Metrics
Rate This Article
See all categories
More
Related Content
Related Articles on the Web

Metabolic Factors Limiting Performance in Marathon Runners

Benjamin I. Rapoport[1,2*]

1 M.D.– Ph.D. Program, Harvard Medical School, Boston, Massachusetts, United States of America, **2** Department of Electrical Engineering and Computer Science and Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

To **add a note**, highlight some text. Hide notes

Make a general comment

Published in the October 2010 Issue of *PLoS Computational Biology*

Total Article Views: 74221

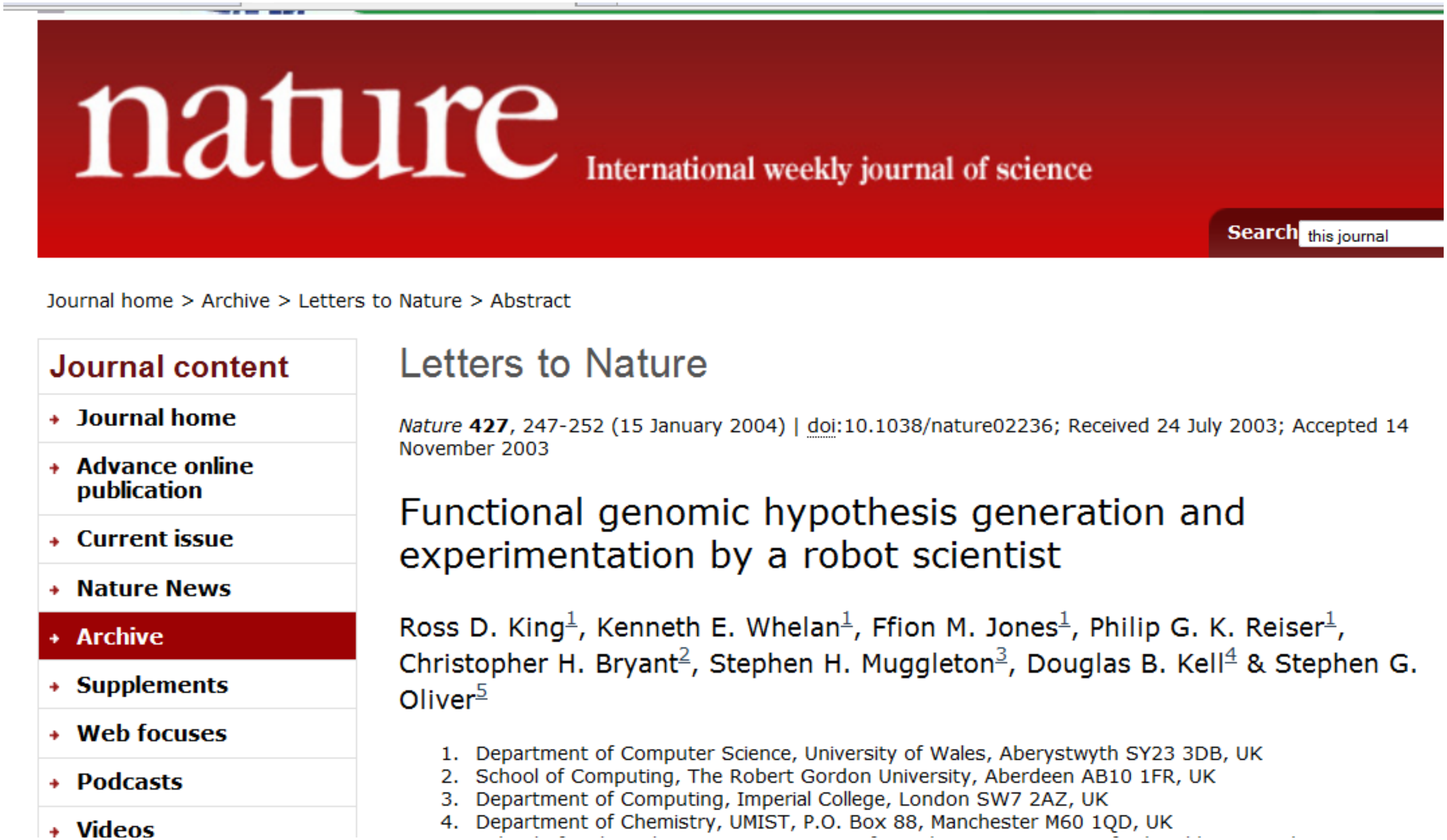
Average Rating (1 User Rating)

★ ★ ★ ★ ★

Abstract Top

Each year in the past three decades has seen hundreds of thousands of runners register to run a major marathon. Of those who attempt to race over the marathon distance of 26 miles and 385 yards (42.195 kilometers), more than two-fifths experience

Active Learning


nature
International weekly journal of science
Search
this journal
Journal home > Archive > Letters to Nature > Abstract
Journal content
Journal home
Advance online publication
Current issue
Nature News
Archive
Supplements
Web focuses
Podcasts
Videos

Letters to Nature

Nature **427**, 247-252 (15 January 2004) | doi:10.1038/nature02236; Received 24 July 2003; Accepted 14 November 2003

Functional genomic hypothesis generation and experimentation by a robot scientist

Ross D. King[1], Kenneth E. Whelan[1], Ffion M. Jones[1], Philip G. K. Reiser[1], Christopher H. Bryant[2], Stephen H. Muggleton[3], Douglas B. Kell[4] & Stephen G. Oliver[5]

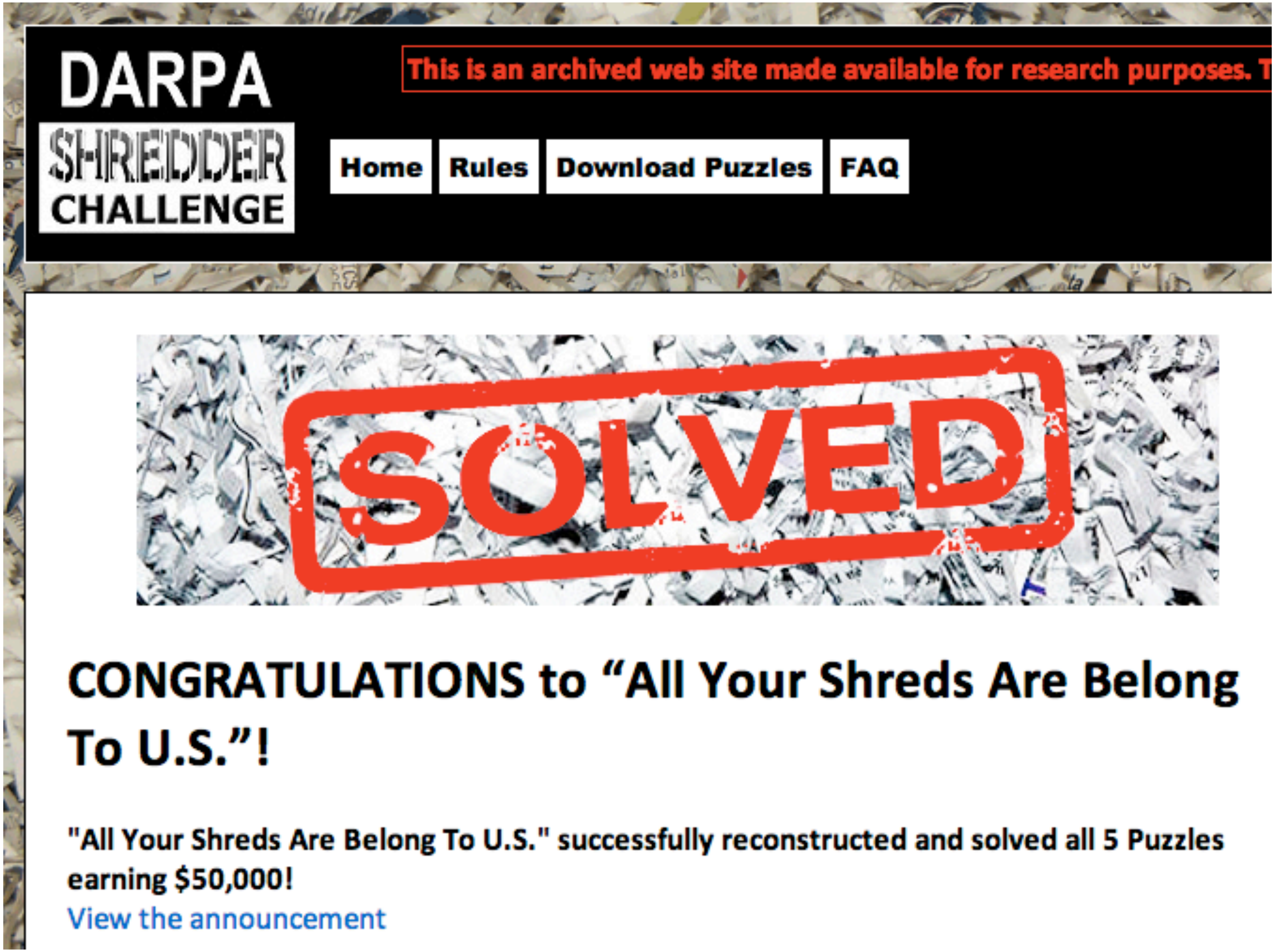
1. Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK
2. School of Computing, The Robert Gordon University, Aberdeen AB10 1FR, UK
3. Department of Computing, Imperial College, London SW7 2AZ, UK
4. Department of Chemistry, UMIST, P.O. Box 88, Manchester M60 1QD, UK

Sequencing DNA

First human genome draft in 2001

Due to *accumulated errors*, we could only reliably read at most **100-200 nucleotides.**

DARPA Shredder Challenge



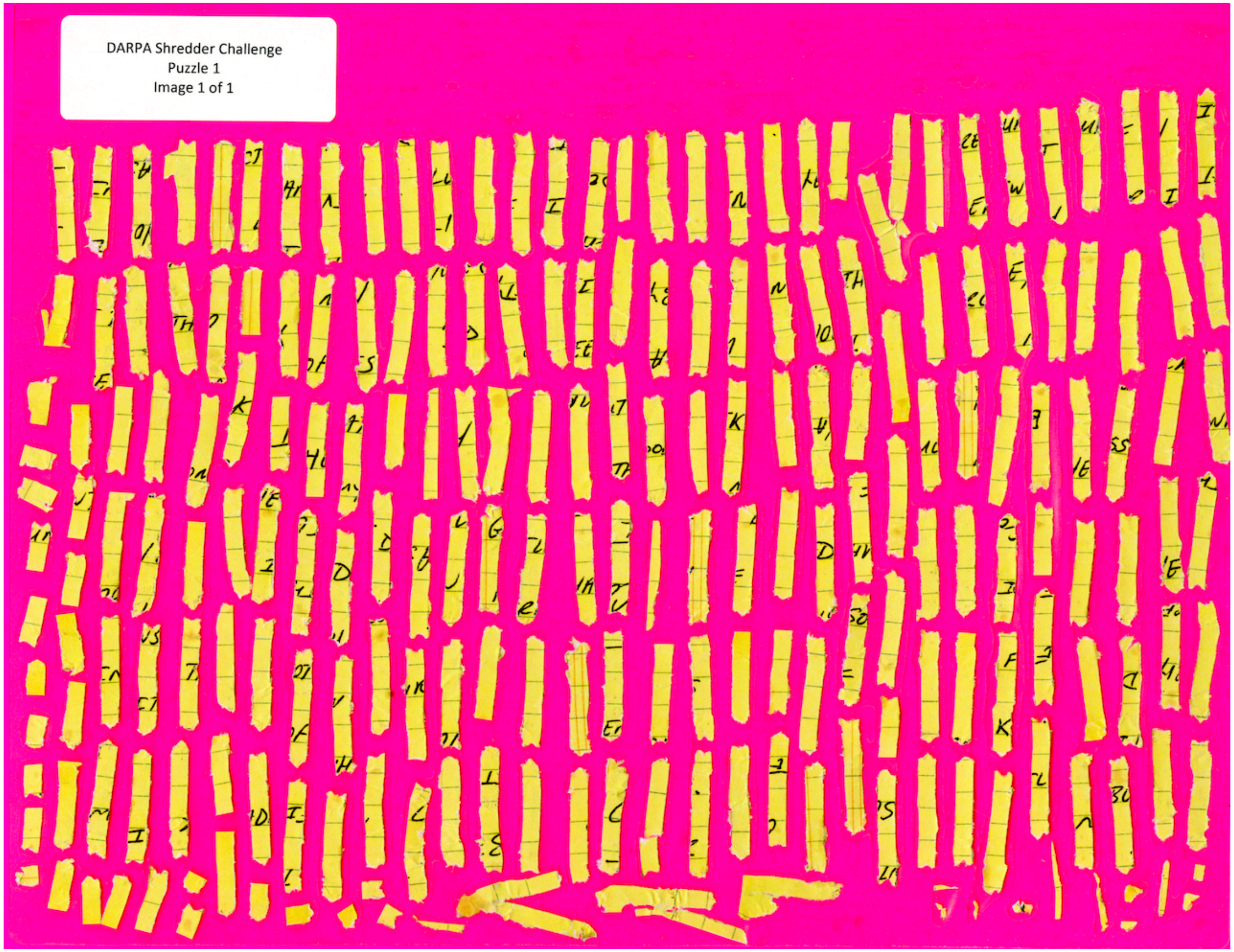
DARPA
SHREDDER
CHALLENGE
This is an archived web site made available for research purposes. T
Home
Rules
Download Puzzles
FAQ
SOLVED

CONGRATULATIONS to "All Your Shreds Are Belong To U.S."!

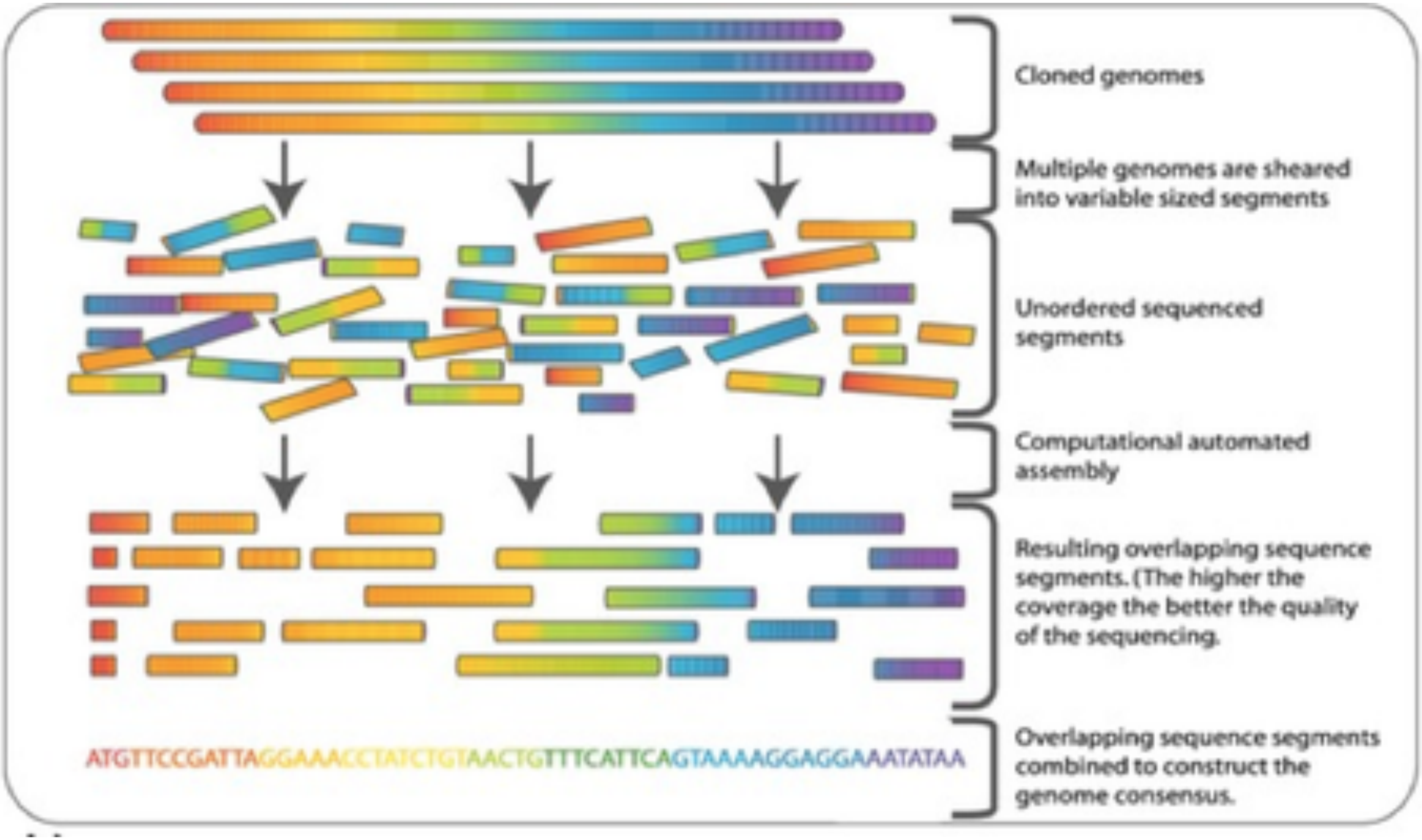
"All Your Shreds Are Belong To U.S." successfully reconstructed and solved all 5 Puzzles earning $50,000!

View the announcement

DARPA Shredder Challenge



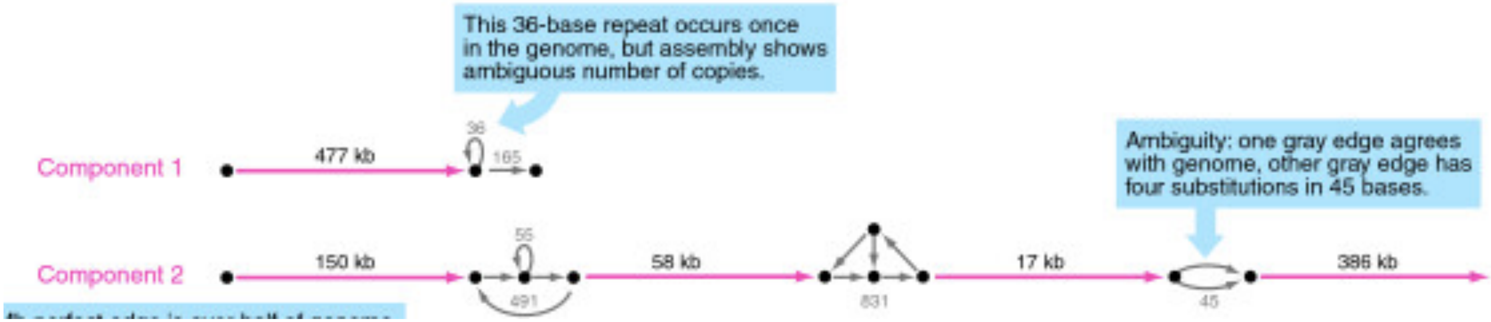
Shotgun Sequencing


Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequence segments combined to construct the genome consensus.

Wikipedia

Caveats

- Errors in reading
- Non-trivial assembly task: repeats in the genome



This 36-base repeat occurs once in the genome, but assembly shows ambiguous number of copies.

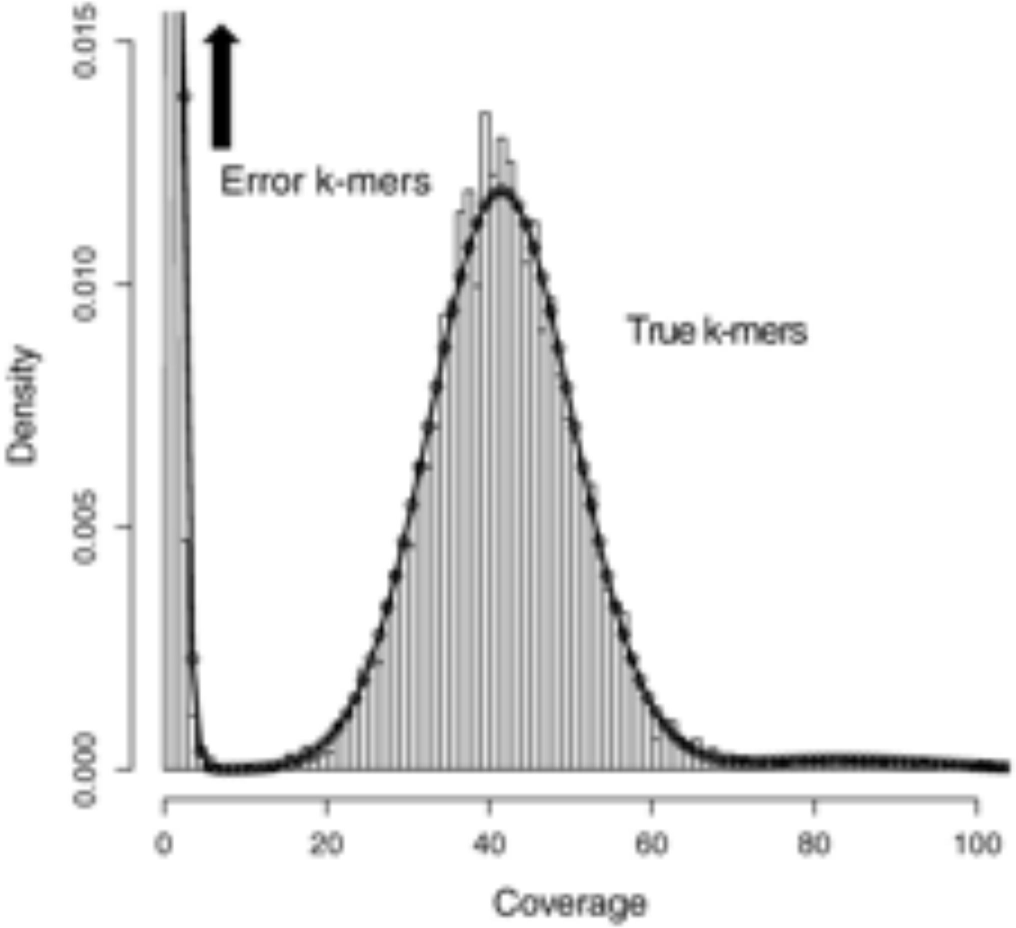
Ambiguity: one gray edge agrees with genome, other gray edge has four substitutions in 45 bases.

MacCallum et al., GB 2009

Error Correction in DNA sequencing

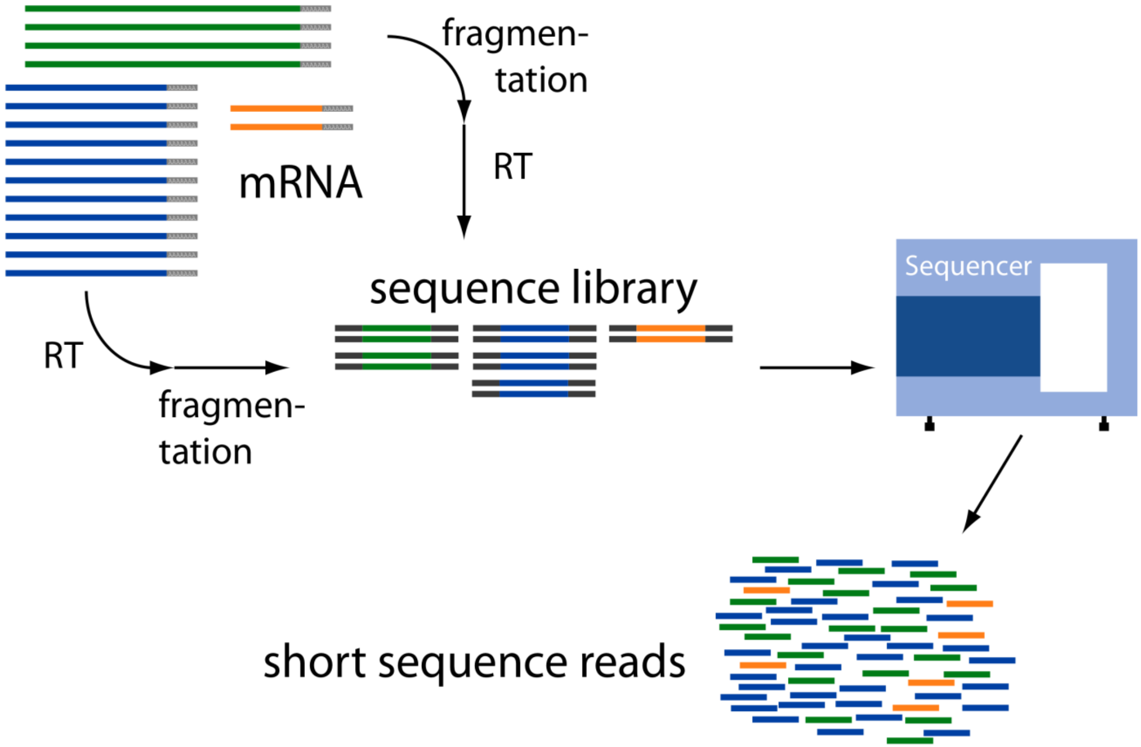
- The fragmentation happens at random locations of the molecules. We expect all positions in the genome to have the same # number of reads

K-mers = substrings of length K of the reads. Errors create error k-mers.



Kellly et al., GB 2010

Transcriptome Shotgun Sequencing (RNA-Seq)



@Friedrich Miescher Laboratory

Sequencing RNA transcripts.

Reminder:

- (mRNA) Transcripts are "expression products" of genes.
- Different genes having different expression levels so some transcripts are more or less abundant than others.

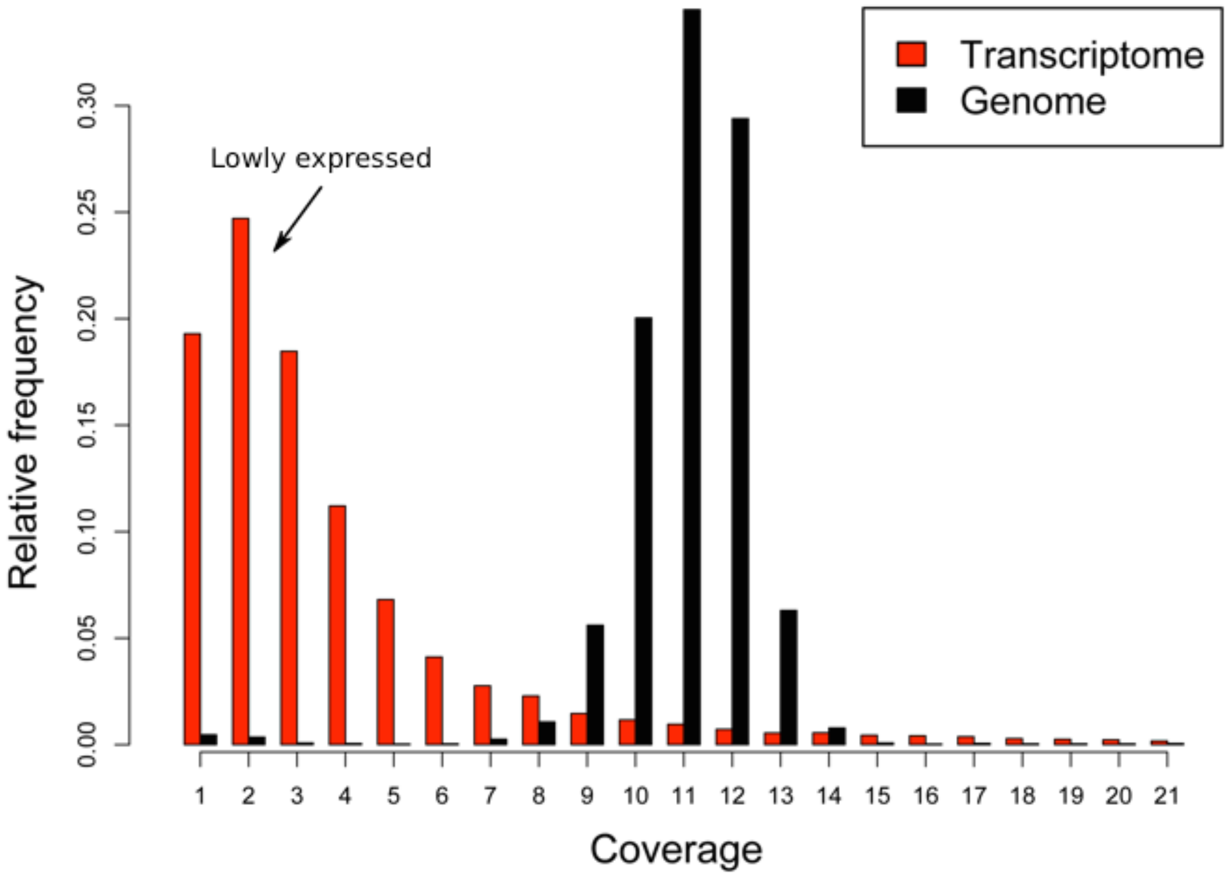
Challenges

- Large datasets: 10-100 millions reads of 75-150 bps.
- Memory efficiency: Too time consuming to perform out-memory processing of data.

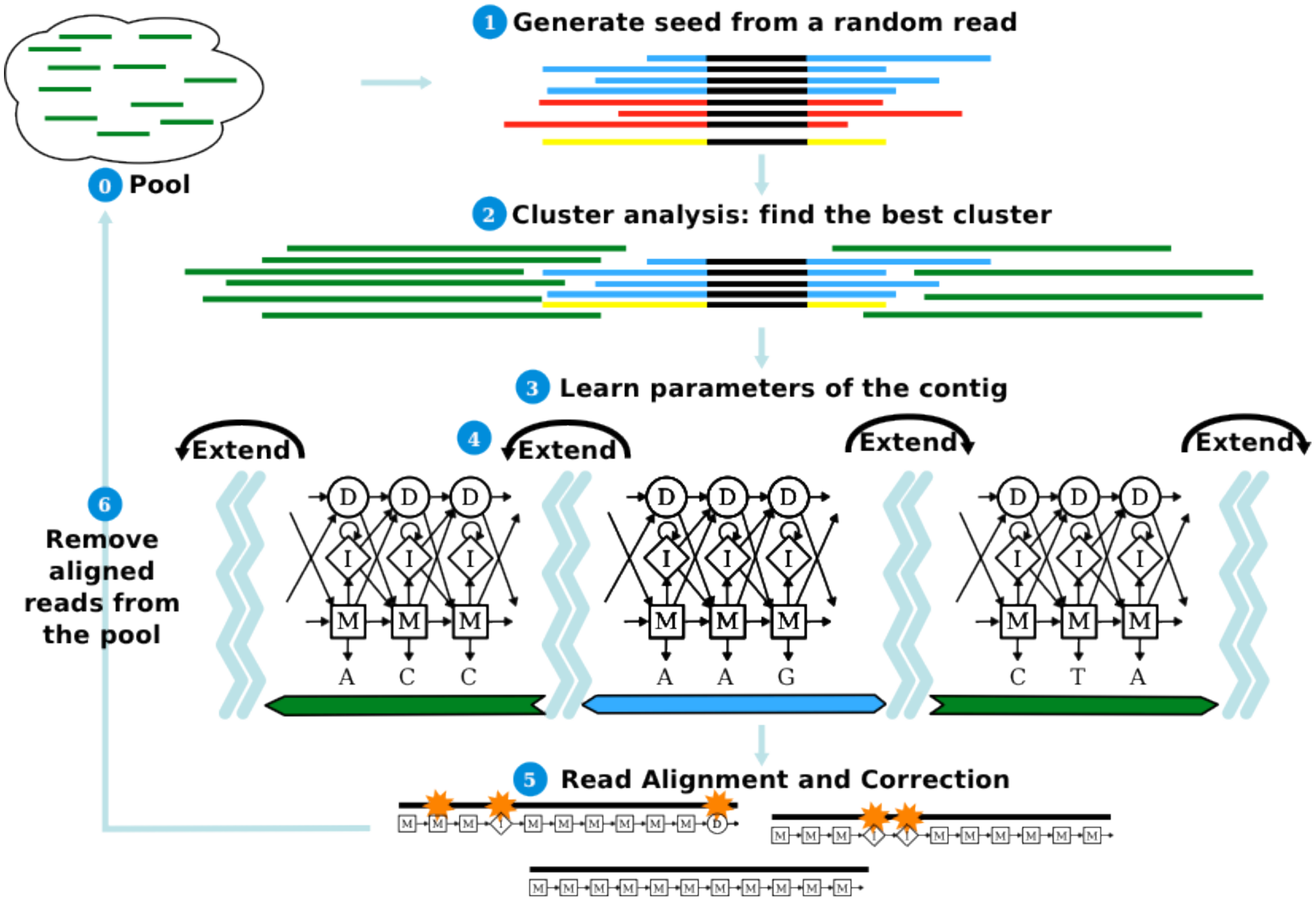
DNA Sequencing + **others :** alternative slicing, RNA editing, post-transcription modification.

Errors are non uniformly distributed

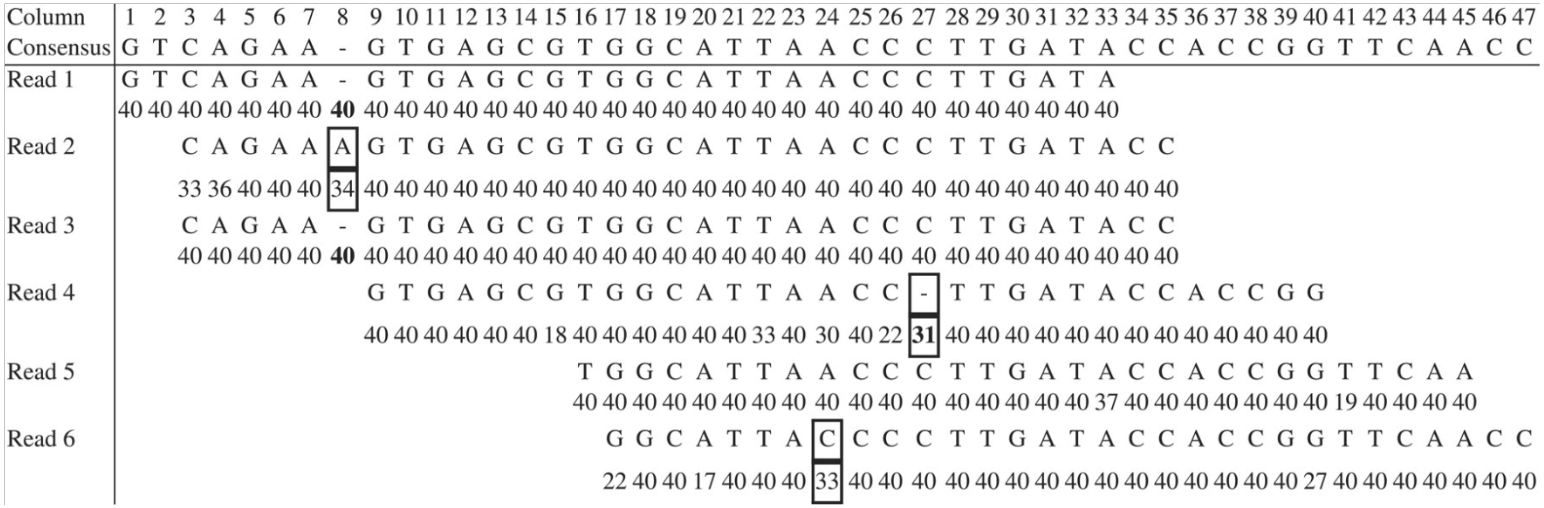
- Some transcripts are more prone to errors
- Errors are harder to correct in reads from lowly expressed transcripts



SEECER
Error Correction + Consensus sequence estimation for RNA-Seq data


0 Pool
1 Generate seed from a random read
2 Cluster analysis: find the best cluster
3 Learn parameters of the contig
4
Extend
Extend
Extend
Extend
D
I
M
A C C
A A G
C T A
5 Read Alignment and Correction
6 Remove aligned reads from the pool

Key idea: HMM model

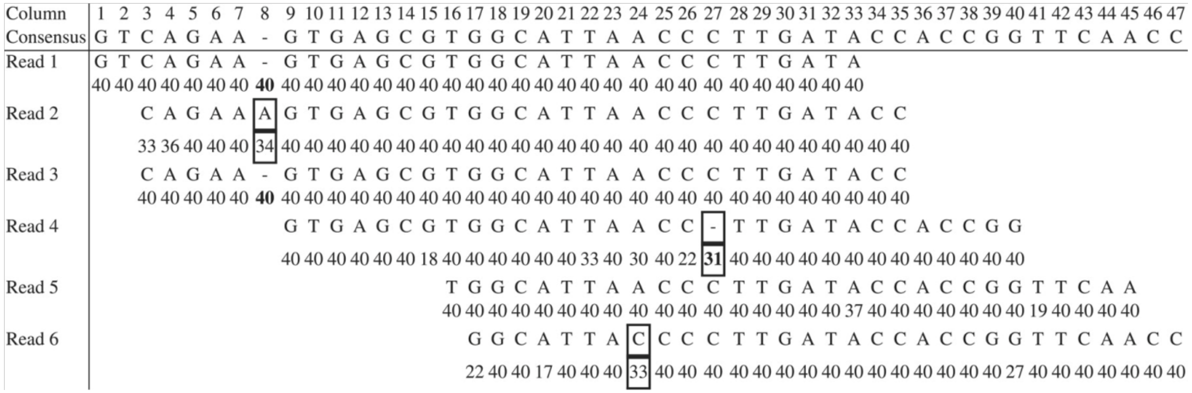


Column	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
Consensus	G	T	C	A	G	A	A	-	G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A	C	C	C	T	T	G	A	T	A	C	C	A	C	C	G	G	T	T	C	A	A	C	C
Read 1	G	T	C	A	G	A	A	-	G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A	C	C	C	T	T	G	A	T	A														
	40	40	40	40	40	40	40	**40**	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40														
Read 2			C	A	G	A	A	A	G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A	C	C	C	T	T	G	A	T	A	C	C												
			33	36	40	40	40	34	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40												
Read 3			C	A	G	A	A	-	G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A	C	C	C	T	T	G	A	T	A	C	C												
			40	40	40	40	40	**40**	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40												
Read 4									G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A	C	C	-	T	T	G	A	T	A	C	C	A	C	C	G	G							
									40	40	40	40	40	40	18	40	40	40	40	40	40	33	40	30	40	22	**31**	40	40	40	40	40	40	40	40	40	40	40	40	40							
Read 5																T	G	G	C	A	T	T	A	A	C	C	C	T	T	G	A	T	A	C	C	A	C	C	G	G	T	T	C	A	A		
																40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	37	40	40	40	40	40	40	40	19	40	40	40	40		
Read 6																	G	G	C	A	T	T	A	C	C	C	C	T	T	G	A	T	A	C	C	A	C	C	G	G	T	T	C	A	A	C	C
																	22	40	40	17	40	40	40	33	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	27	40	40	40	40	40	40	40

Salmela et al., Bioinformatics 2011

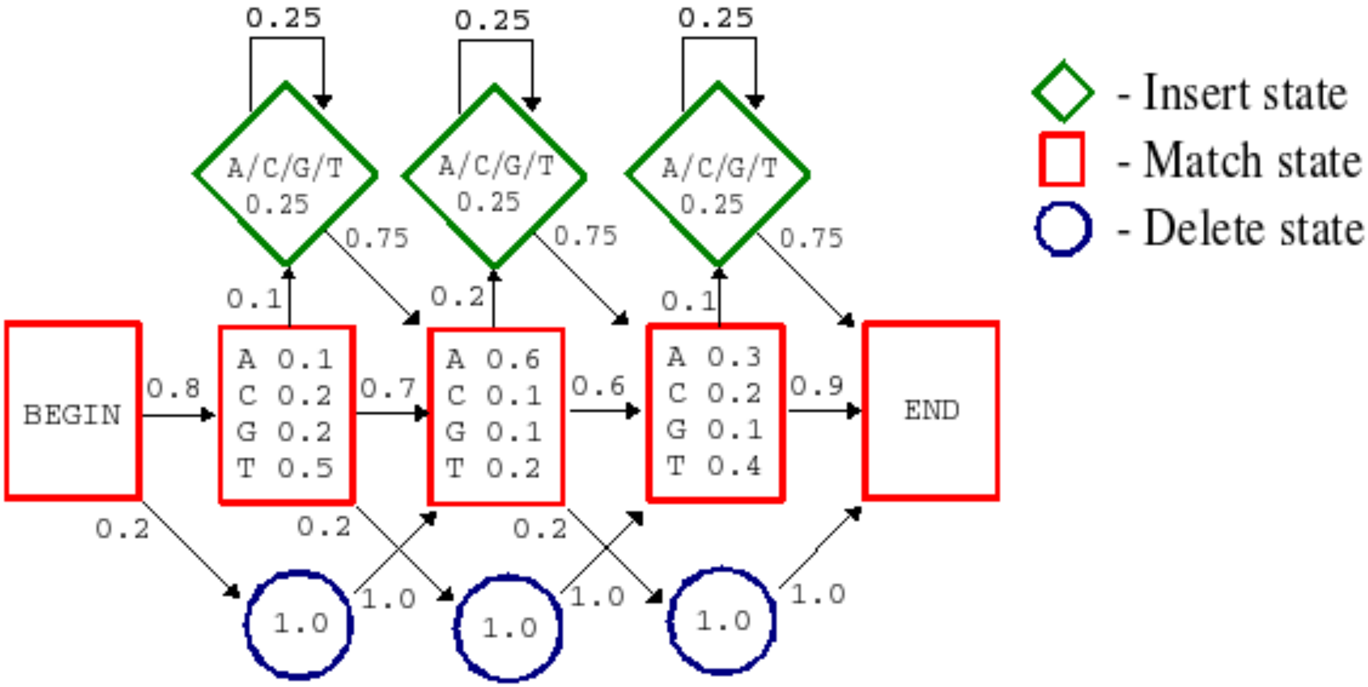
The way sequencers work:

- Read letter by letter sequentially
- Possible errors: Insertion , Deletion or Misread of a nucleotide



Column	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Consensus	G	T	C	A	G	A	A	-	G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A
Read 1	G	T	C	A	G	A	A	-	G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A
	40	40	40	40	40	40	40	**40**	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Read 2			C	A	G	A	A	A	G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A
			33	36	40	40	40	34	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Read 3			C	A	G	A	A	-	G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A
			40	40	40	40	40	**40**	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Read 4									G	T	G	A	G	C	G	T	G	G	C	A	T	T	A	A
									40	40	40	40	40	40	18	40	40	40	40	40	40	33	40	30
Read 5																T	G	G	C	A	T	T	A	A
																40	40	40	40	40	40	40	40	40
Read 6																	G	G	C	A	T	T	A	C
																	22	40	40	17	40	40	40	33

Column	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
Consensus	C	C	C	T	T	G	A	T	A	C	C	A	C	C	G	G	T	T	C	A	A	C	C
Read 1	C	C	C	T	T	G	A	T	A														
	40	40	40	40	40	40	40	40	40														
Read 2	C	C	C	T	T	G	A	T	A	C	C												
	40	40	40	40	40	40	40	40	40	40	40												
Read 3	C	C	C	T	T	G	A	T	A	C	C												
	40	40	40	40	40	40	40	40	40	40	40												
Read 4	C	C	-	T	T	G	A	T	A	C	C	A	C	C	G	G							
	40	22	**31**	40	40	40	40	40	40	40	40	40	40	40	40	40							
Read 5	C	C	C	T	T	G	A	T	A	C	C	A	C	C	G	G	T	T	C	A	A		
	40	40	40	40	40	40	40	40	37	40	40	40	40	40	40	40	19	40	40	40	40		
Read 6	C	C	C	T	T	G	A	T	A	C	C	A	C	C	G	G	T	T	C	A	A	C	C
	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	27	40	40	40	40	40	40	40



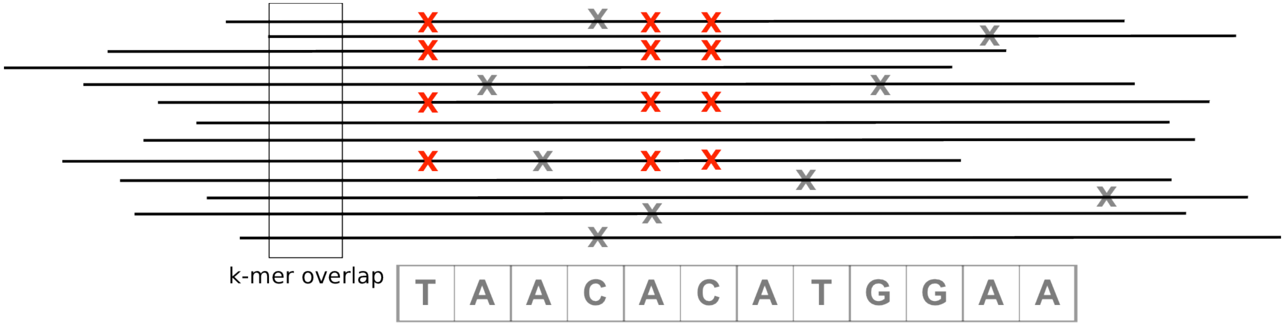
Building (Learning) the HMMs and Making Corrections (Inference)

Learning = Expectation-Maximization
Inference = Viterbi algorithm

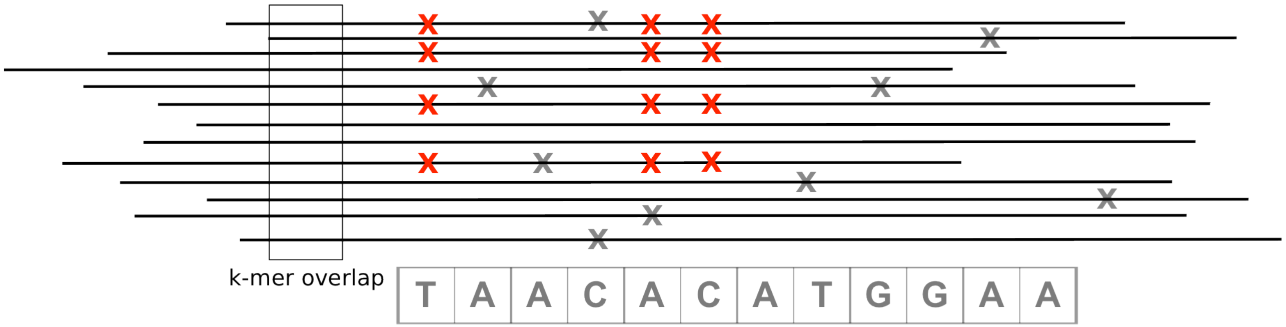
Seeding:
Guessing possible reads using k-mer overlaps.
Constructing the HMM from these reads.

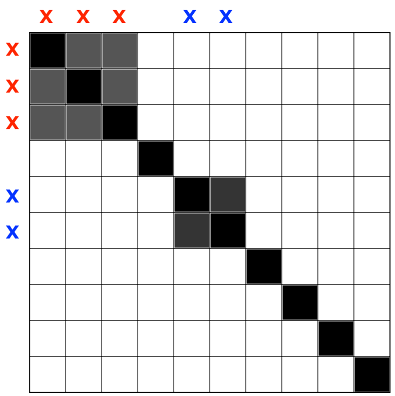
Speed up:
The k-mer overlaps yield approximate multiple alignments of reads.
We can learn HMM parameters from this directly.

Clustering to improve seeding



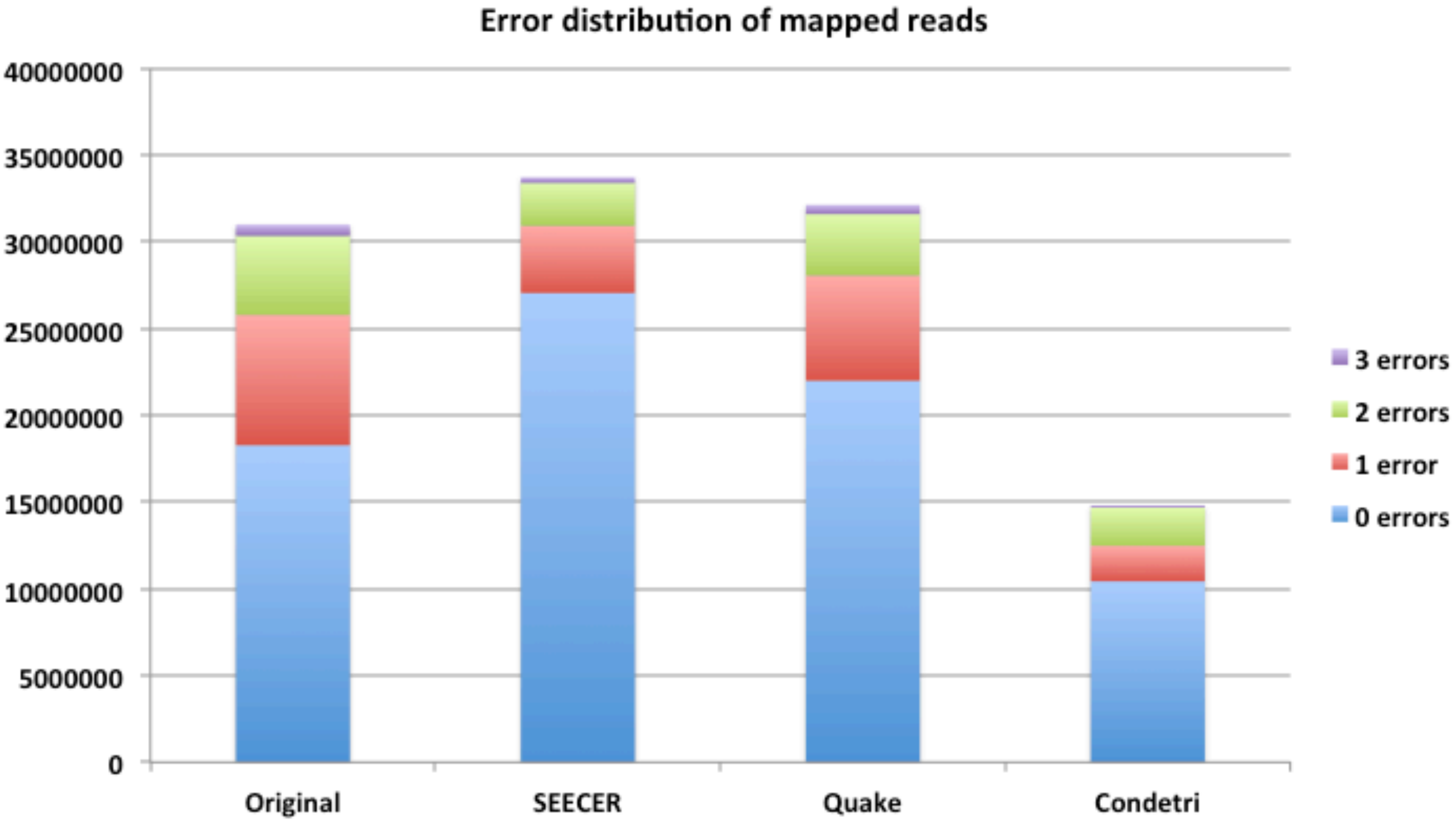
Real biological differences should be supported by a set of reads with similar mismatches to the consensus



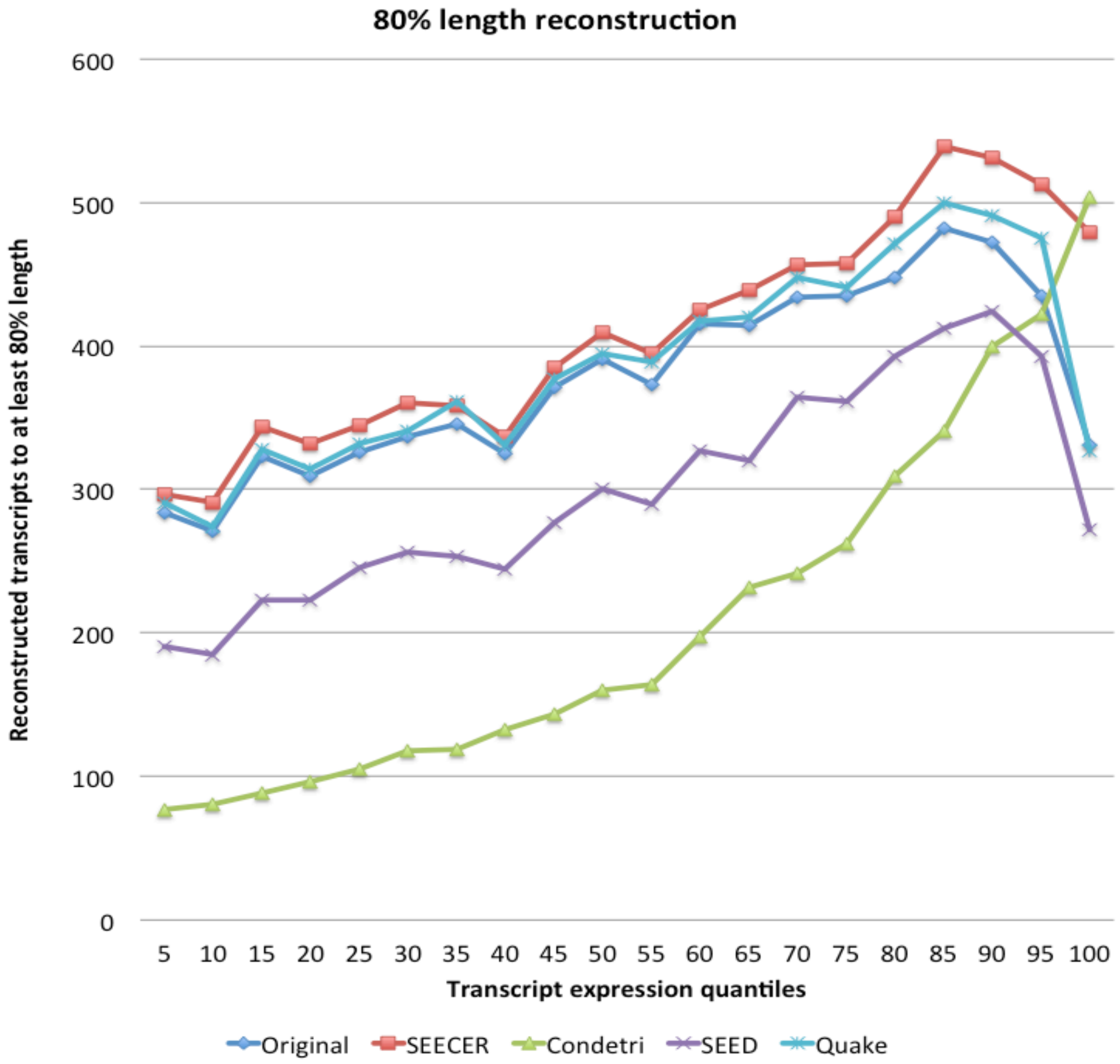


1. Clustering positions with mismatches to identify clusters of correlated positions.
2. Build a similarity matrix between these positions.
3. Use Spectral clustering to find clusters of correlated positions.
4. Filter reads have mismatches in these clusters.

Comparison to other methods



Using the corrected reads, the assembler can recover **more** transcripts

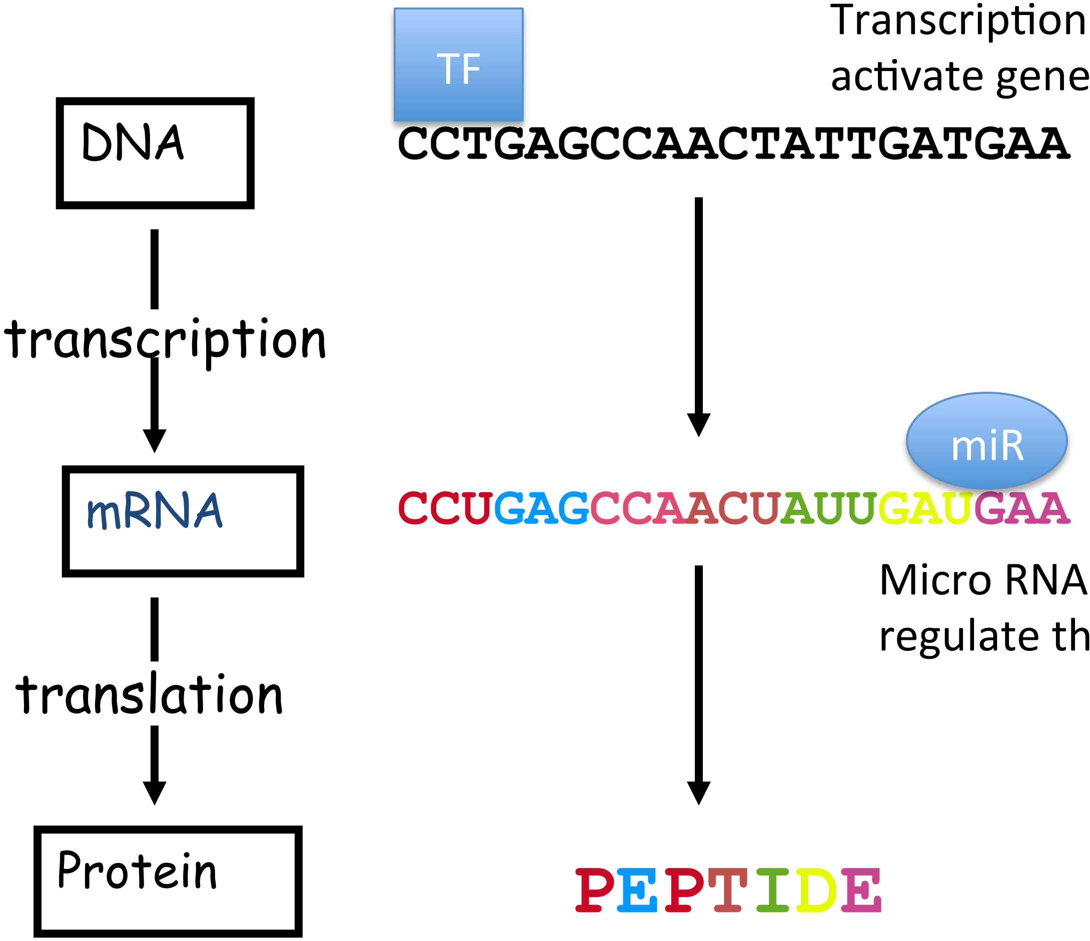


Things that work

- Approximate learning to speed up on large datasets.
- In real world, one technique is not enough. A solution involves using many techniques.
- Precision and Recall are trade-offs.

Central dogma

Different regulators control the information flow from DNA to protein



Transcription factors (TFs) bind to DNA and activate genes

TF

DNA

CCTGAGCCAACTATTGATGAA

transcription

miR

mRNA

CCUGAGCCAACUAUUGAUGAA

Micro RNAs (miRs) bind to mRNA to down regulate their expression

translation

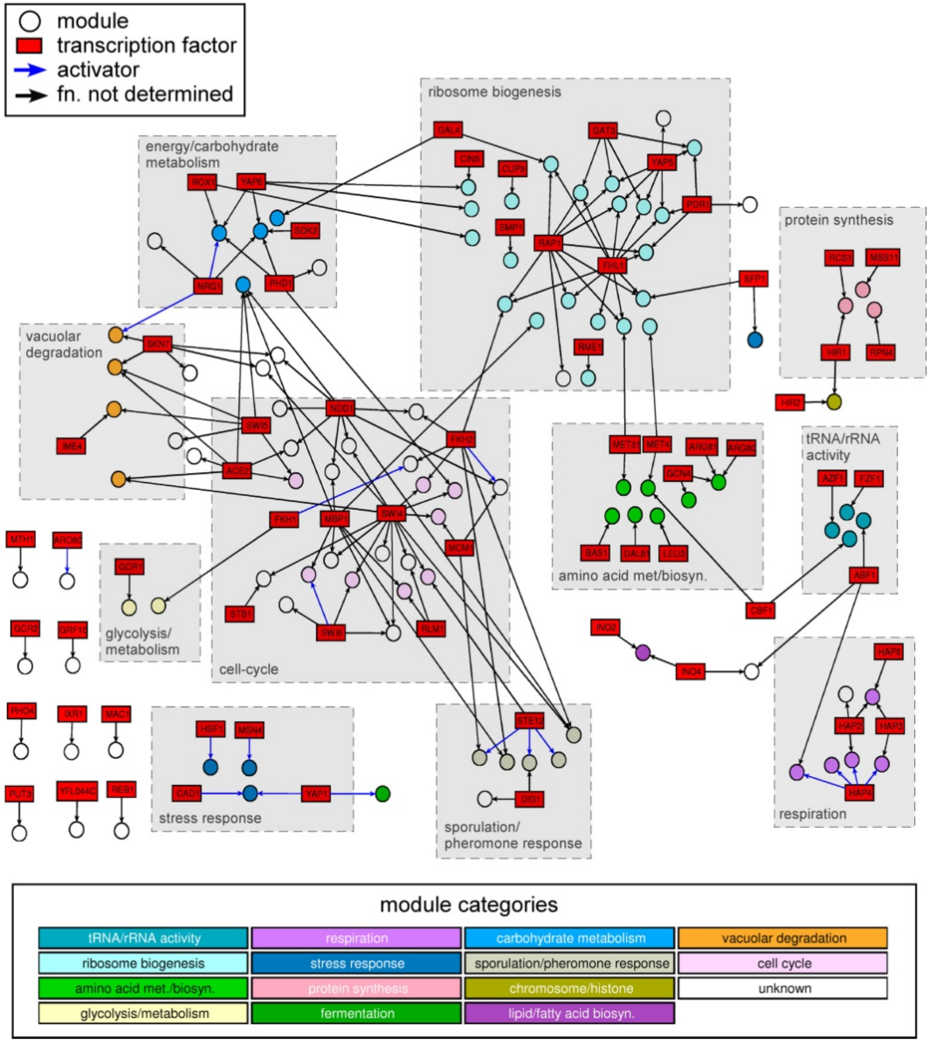
Protein

PEPTIDE

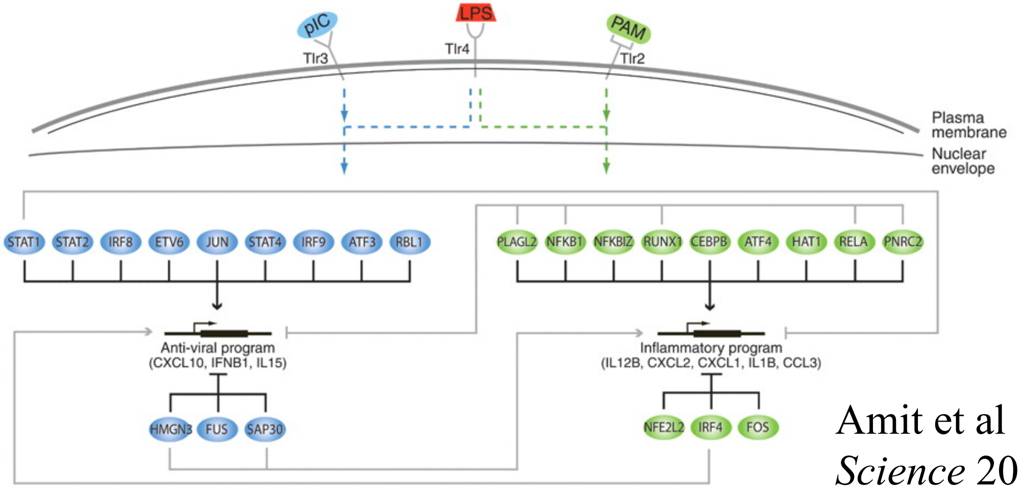
Integrating expression and protein-DNA interaction data

Lee *et al Science* 2002

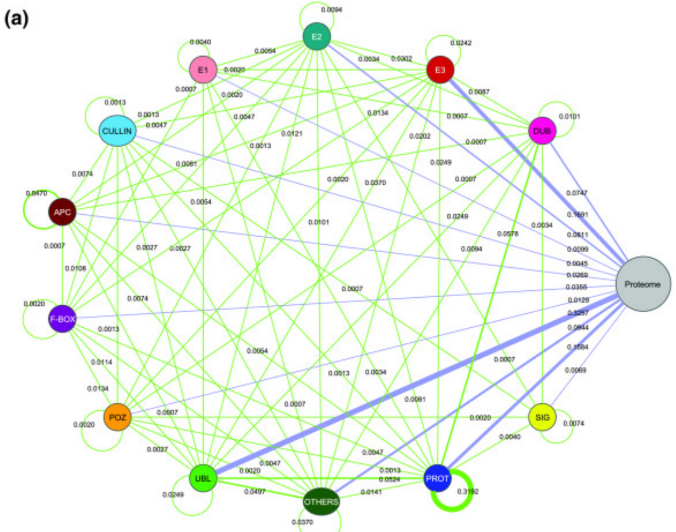
Bar-Joseph *et al Nature Biotechnology* 2003



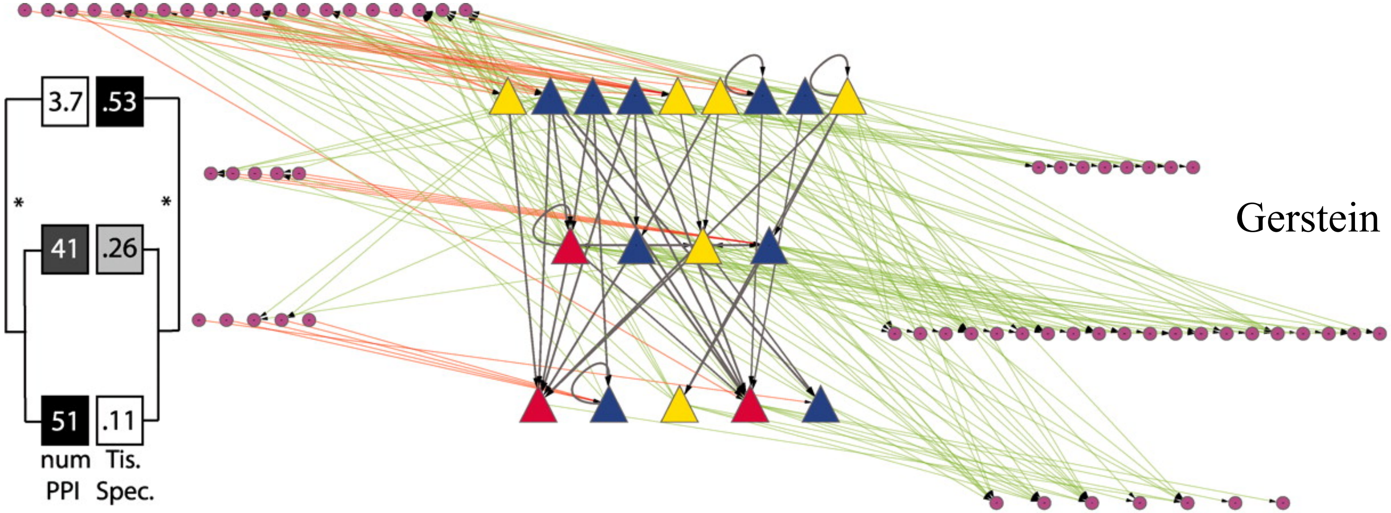
Methods for reconstructing networks in cells



Amit et al *Science* 2009



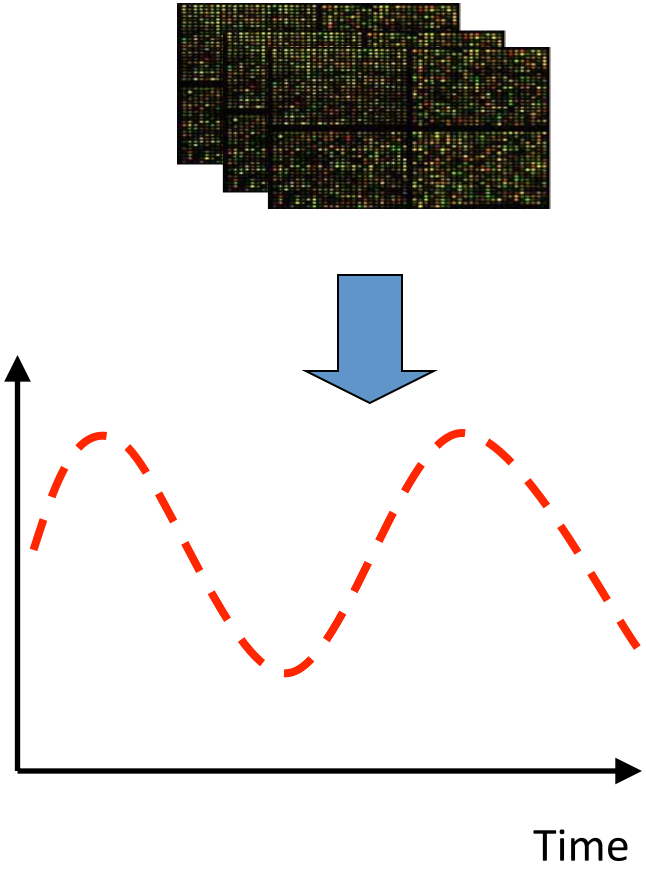
Venancio et al *Genome Biology* 2009



Gerstein et al *Science* 2010

Key problem: Most high-throughput data is static

Time-series measurements

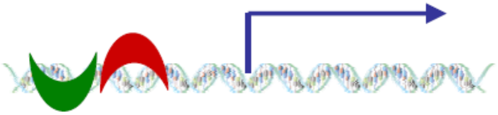


Static data sources

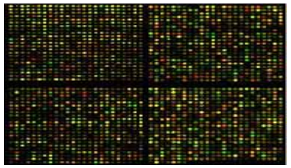
Sequencing

motif

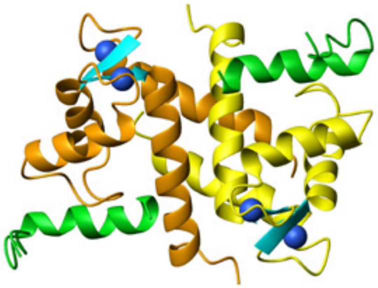
CHIP-chip



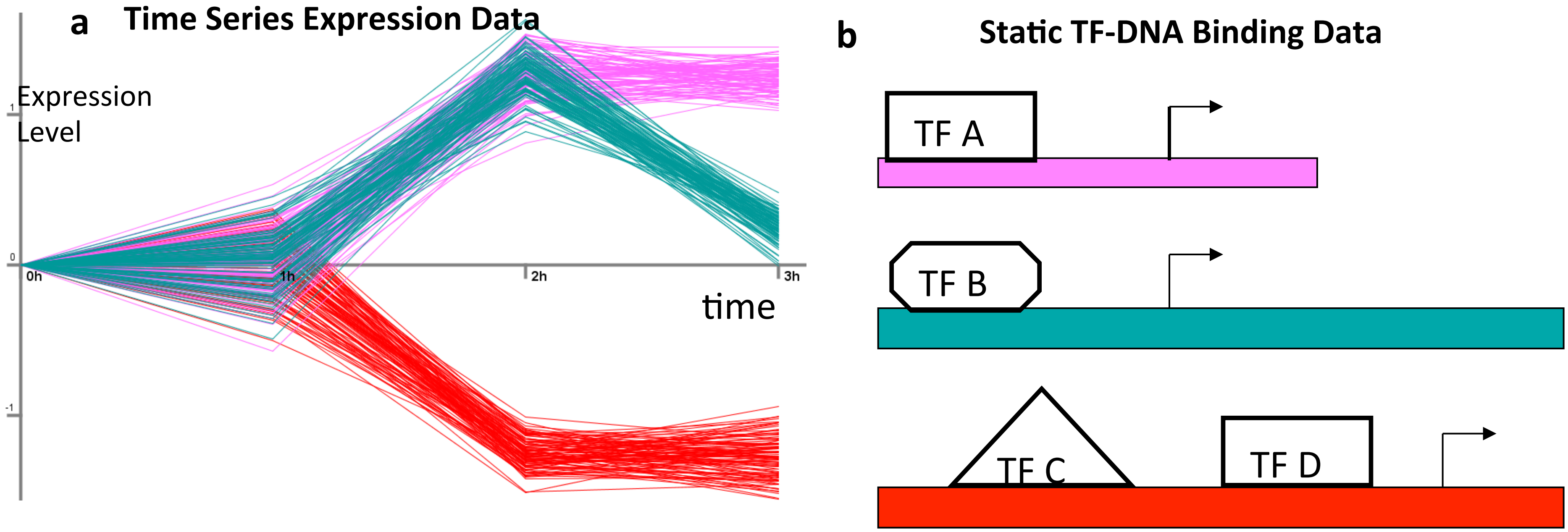
microarray



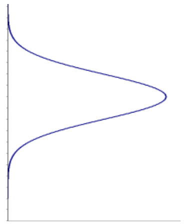
PPI

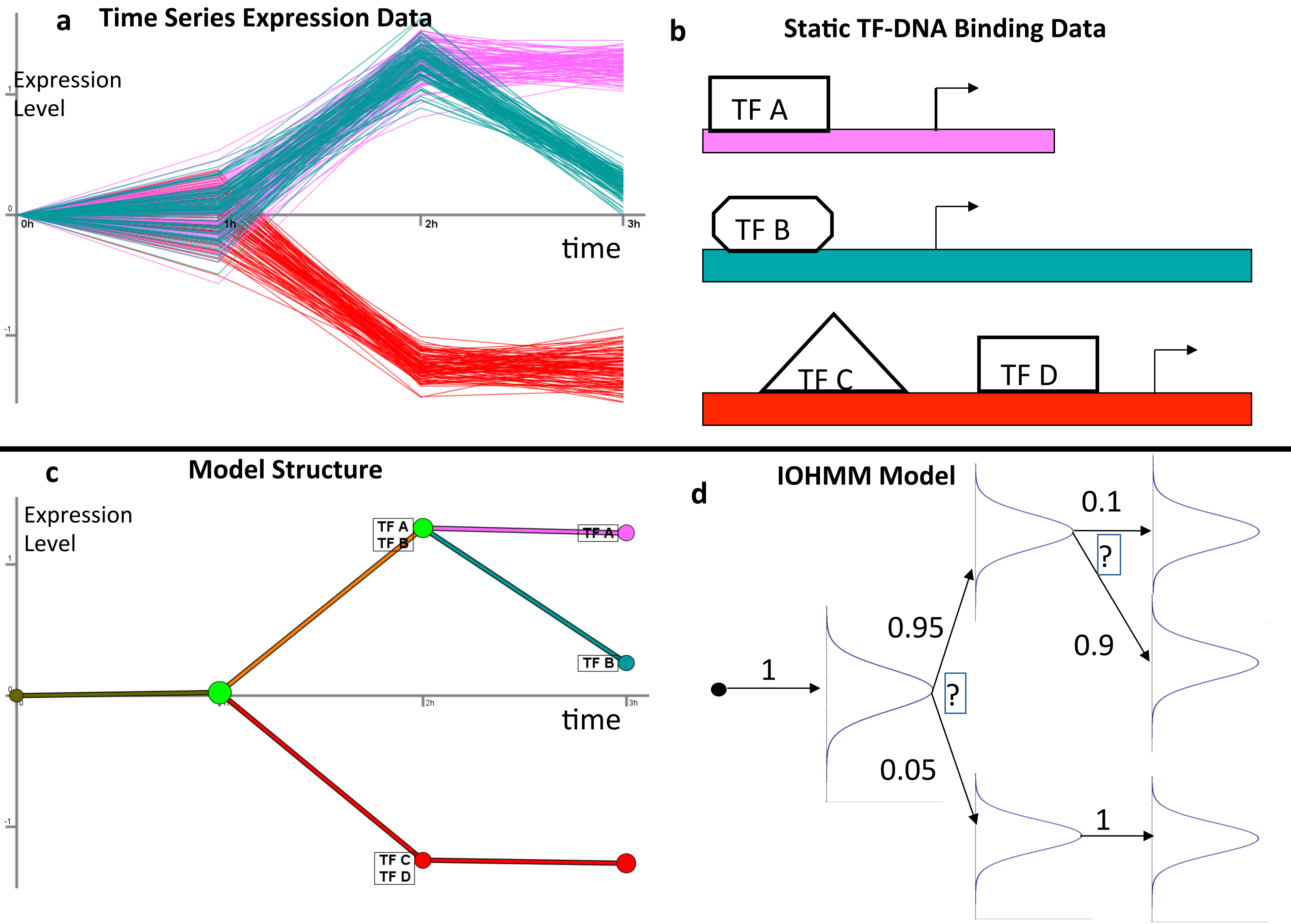


DREM: Dynamic Regulatory Events Miner

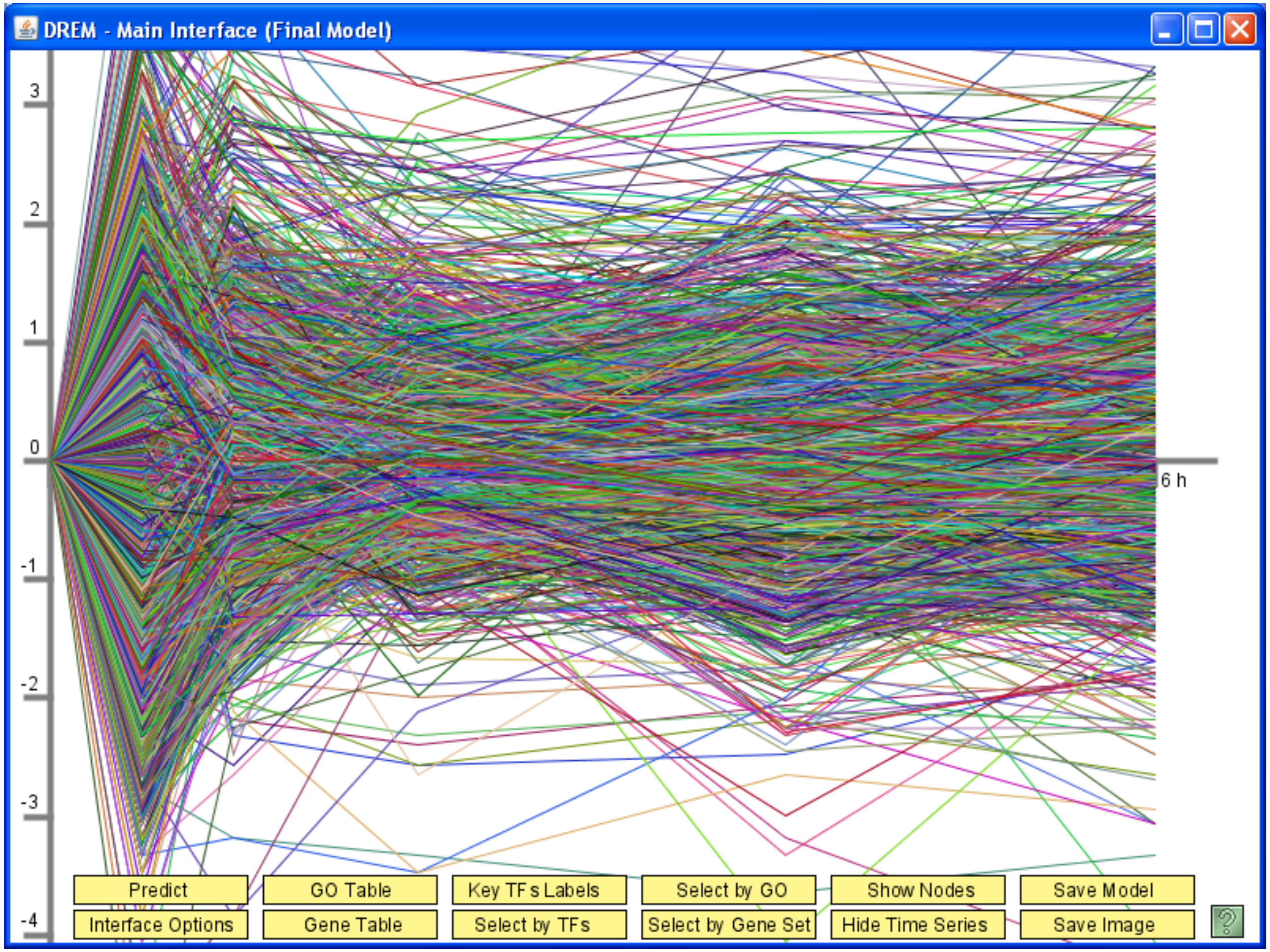

a Time Series Expression Data

b Static TF-DNA Binding Data




a
Time Series Expression Data
Expression Level
1
0
-1
0h
1h
2h
3h
time
b
Static TF-DNA Binding Data
TF A
TF B
TF C
TF D
c
Model Structure
Expression Level
1
0
-1
2h
3h
TF A
TF B
TF A
TF B
TF C
TF D
time
d
IOHMM Model
1
0.95
?
0.05
0.1
?
0.9
1

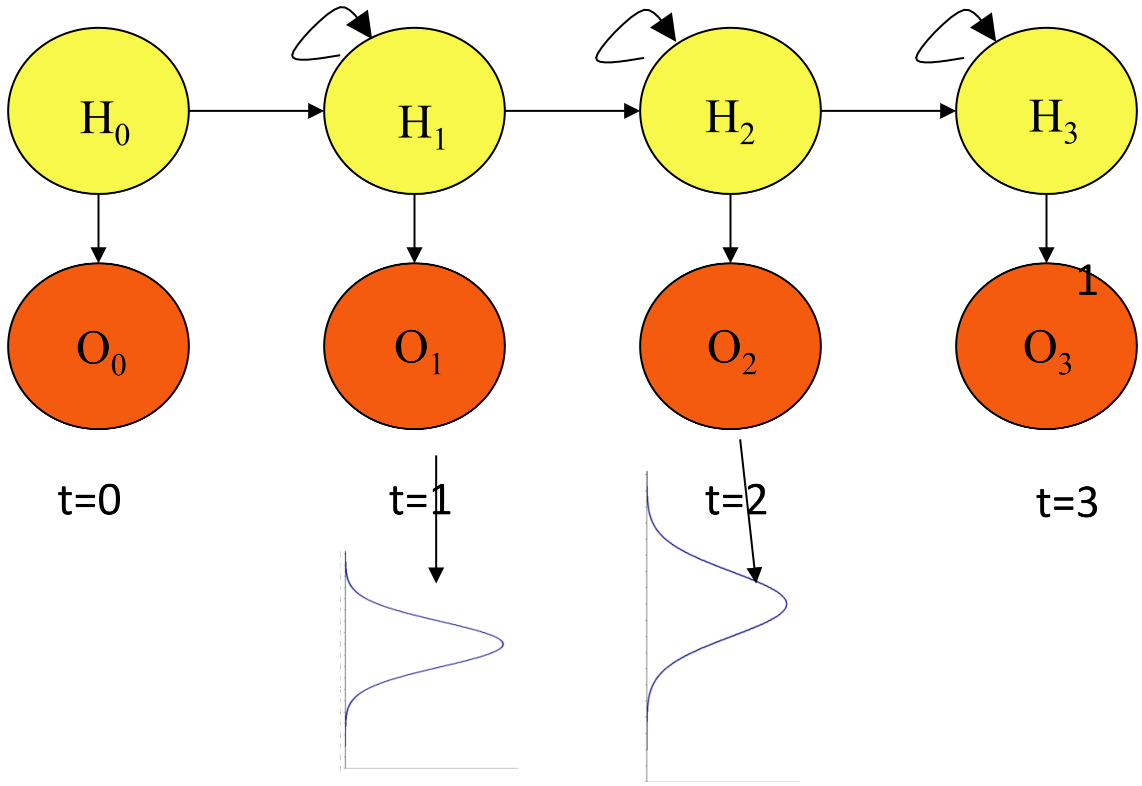
Things are a bit more complicated: Real data



A Hidden Markov Model

Hidden States

Observed outputs
(expression levels)



$$L(H,O;\Theta) = \prod_{i=1}^{n}\left[\prod_{t=1}^{T} p(O_t(i)\,|\,H_t(i))\right]\left[\prod_{t=2}^{T} p(H_t(i)\,|\,H_{t-1}(i))\right]$$

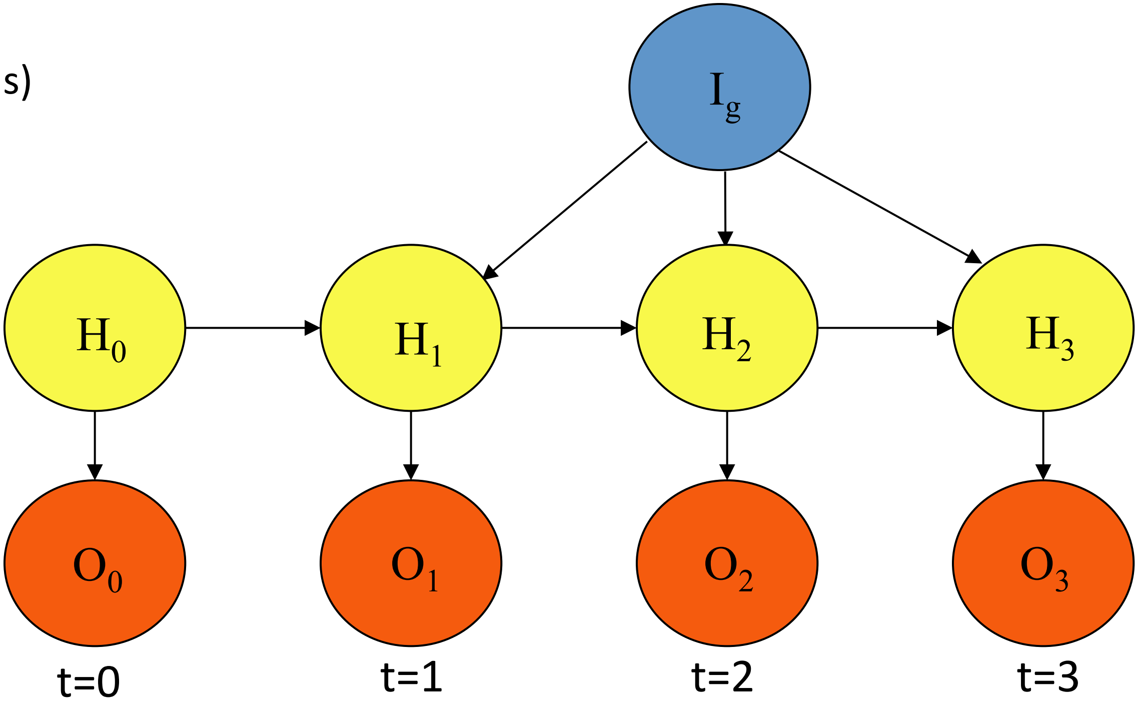
Schliep et al *Bioinformatics* 2003

Input – Output Hidden Markov Model

Input (Static TF-gene interactions)

Hidden States (transitions between states form a tree structure)

Emissions (Distribution of expression values)



Log Likelihood

But how do we express these conditional probabilities?

$$r(G|M) = \underbrace{\sum_{g\in G}}_{\text{Sum over all genes}} \log \underbrace{\sum_{q\in Q}}_{\text{Sum over all paths } Q} \prod_{t=1}^{n-1} \underbrace{f_{q(t)}(o_g(t))}_{\text{Product over all Gaussian emission density values on path}} \boxed{\prod_{t=1}^{n-1} \underbrace{P(H_t = q(t)|H_{t-1} = q(t-1), I_g)}_{\text{Product over all transition probabilities on path}}}$$

Input-Output Hidden Markov Model

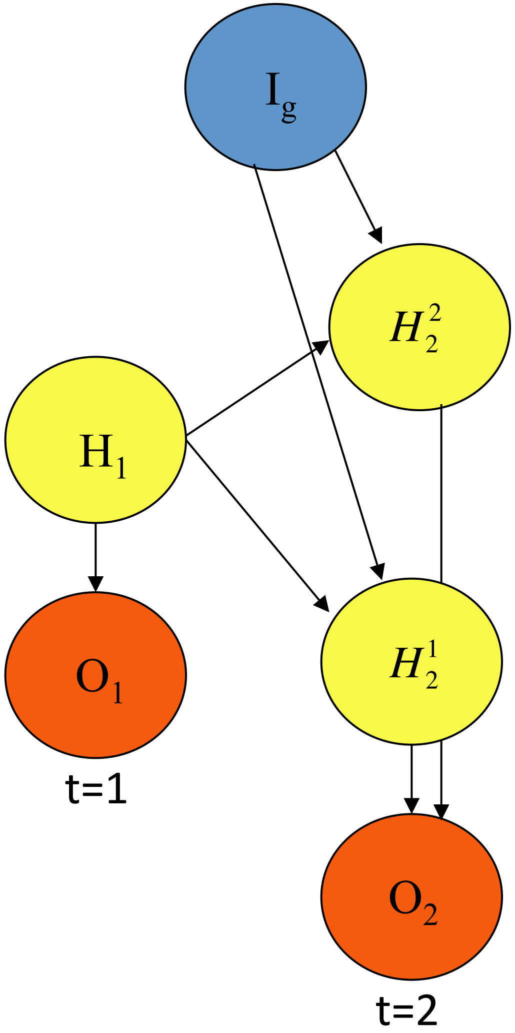
learning the transition probabilities

How do compute *P* for a state with 2 children?

We can write it as a logistic regression classification problem!

$$P(H_2^1 = q(2) \mid H_1 = q(1), I_g) = ?$$

$$P(H_2^2 = q(2) \mid H_1 = q(1), I_g) = ?$$



Input-Output Hidden Markov Model

learning the transition probabilities

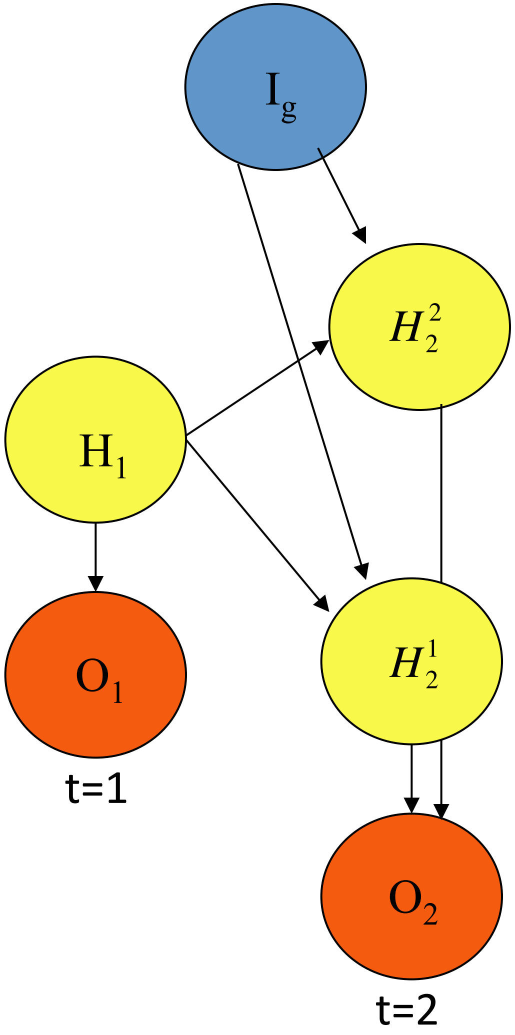
How do compute *P* for a state with 2 children?

We can write it as a logistic regression classification problem!

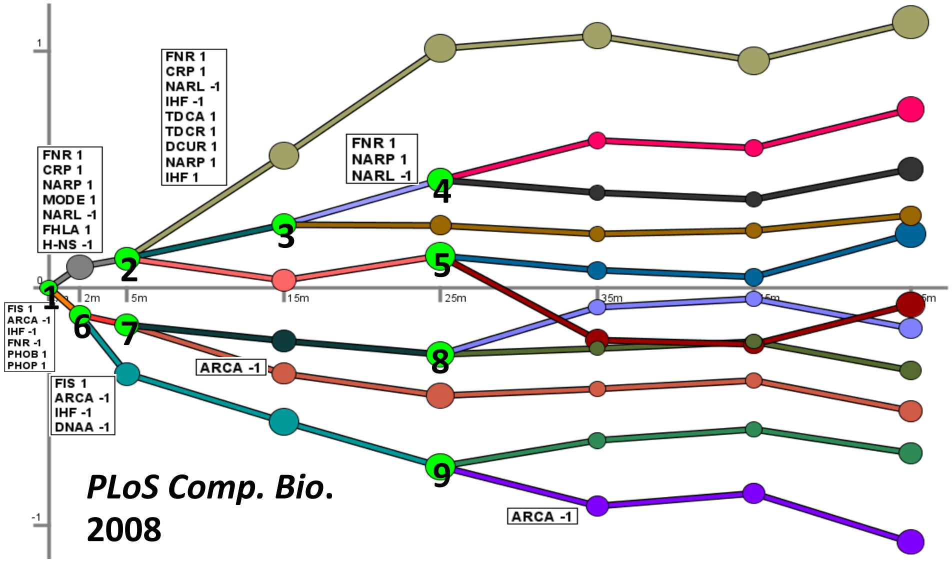
$$P(H_2^1 = q(2) \mid H_1 = q(1), I_g) = \frac{1}{1+\exp(w_0 + \underbrace{\sum_r}_{\text{Sum over all regulators}} I_{g,r} w_r)}$$

Optimize $w_{i'}$’s with a logistic regression classifier

$$\hat{W} = \underset{W}{\operatorname{argmax}} \left[\underbrace{l(W)}_{\text{likelihood with parameters } W} + \underbrace{\ln p(W)}_{L_1\text{-penalty to promote sparsity}} \right]$$

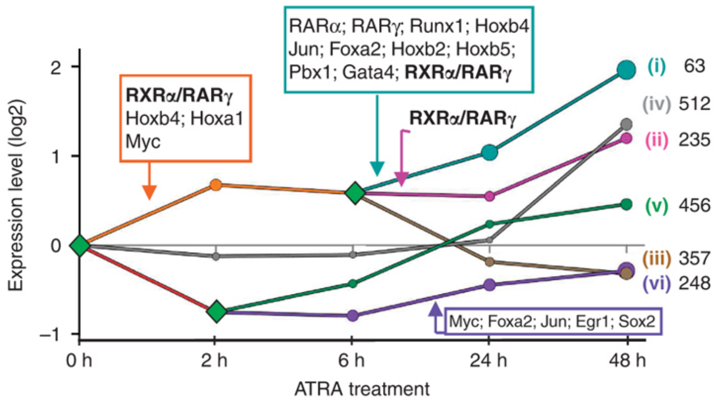


E. coli. response



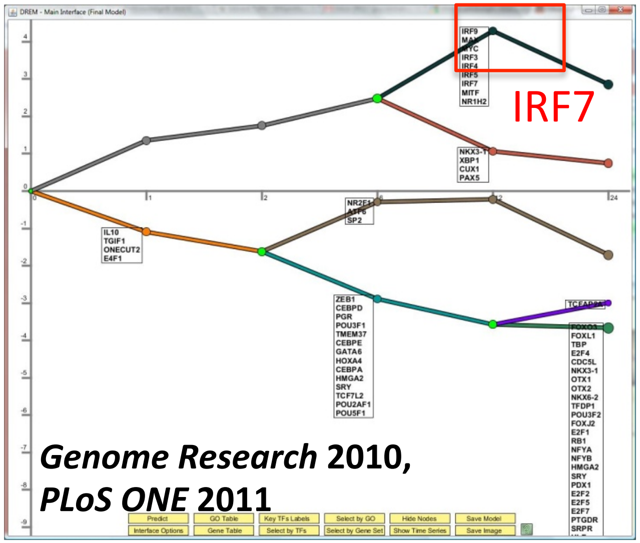
PLoS Comp. Bio. 2008

Stem cells differentiation



Nature MSB 2011

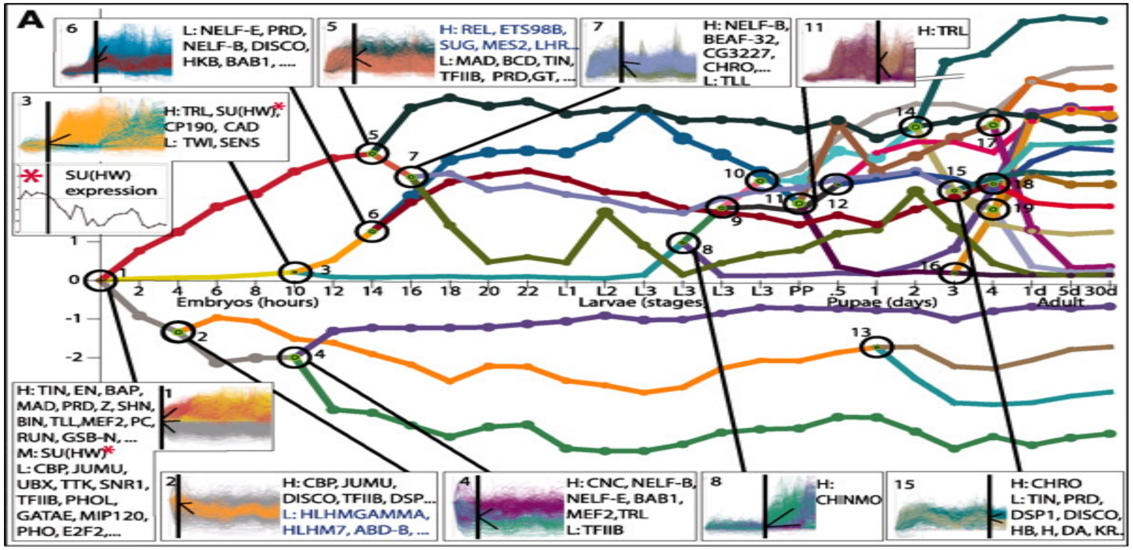
Mouse Immune response



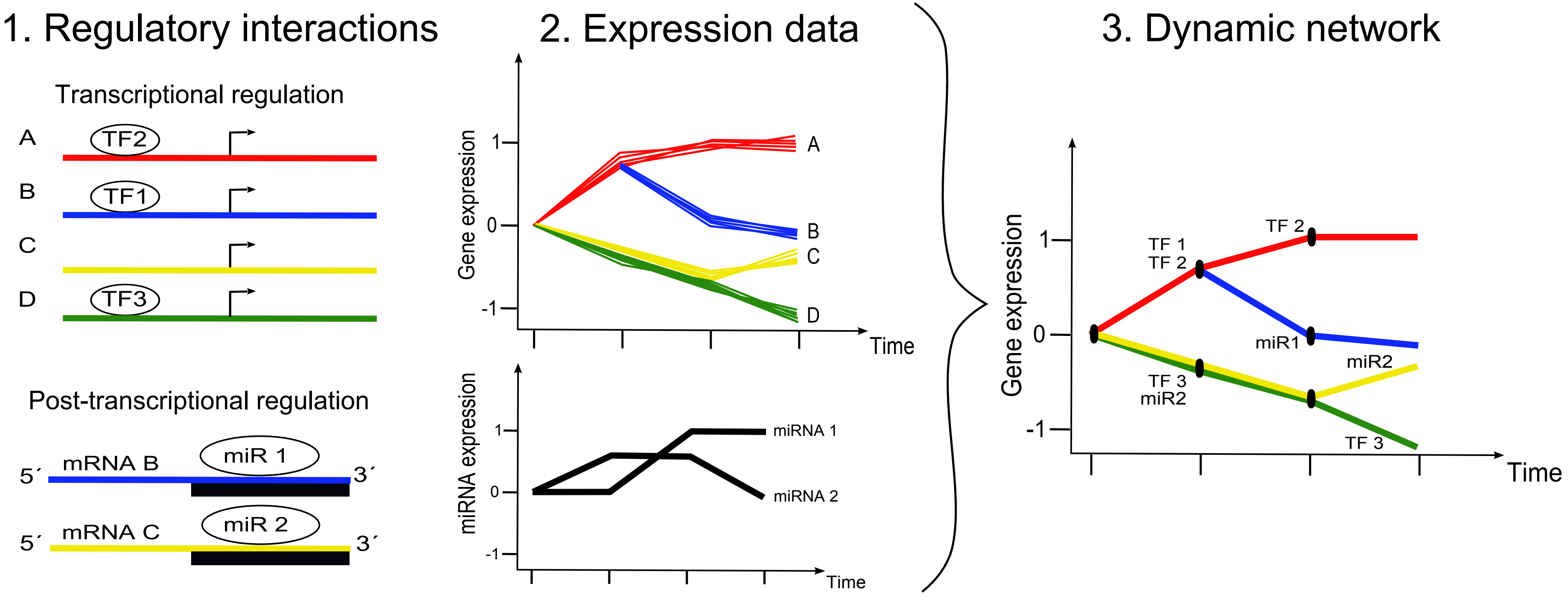
Genome Research 2010,
PLoS ONE 2011

Fly development

Science 2010



mirDREM



1. Regulatory interactions

2. Expression data

3. Dynamic network

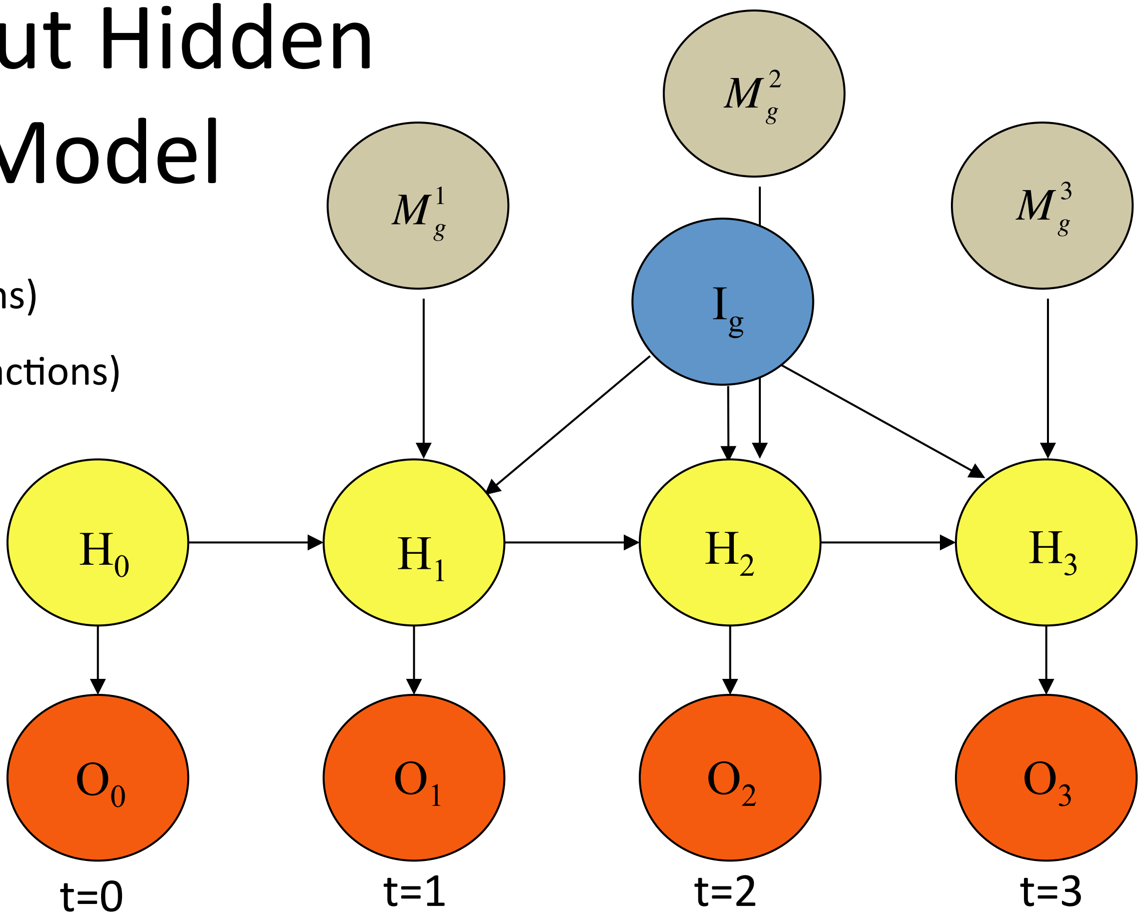
Input – Output Hidden Markov Model

Input (**Static** TF-gene interactions)

(**Dynamic** miR-gene interactions)

Hidden States (transitions between states form a tree structure)

Emissions (Distribution of expression values)



Log Likelihood

But how do we express these conditional probabilities?

$$r(G|M) = \underbrace{\sum_{g\in G}}_{\text{Sum over all genes}} \log \underbrace{\sum_{q\in Q}}_{\text{Sum over all paths } Q} \prod_{t=1}^{n-1} \underbrace{f_{q(t)}(o_g(t))}_{\text{Product over all Gaussian emission density values on path}} \boxed{\prod_{t=1}^{n-1} \underbrace{P(H_t = q(t)|H_{t-1} = q(t-1), I_g, M_g^t)}_{\text{Product over all transition probabilities on path}}}$$

Input-Output Hidden Markov Model

learning the transition probabilities

How do compute *P* for a state with 2 children?

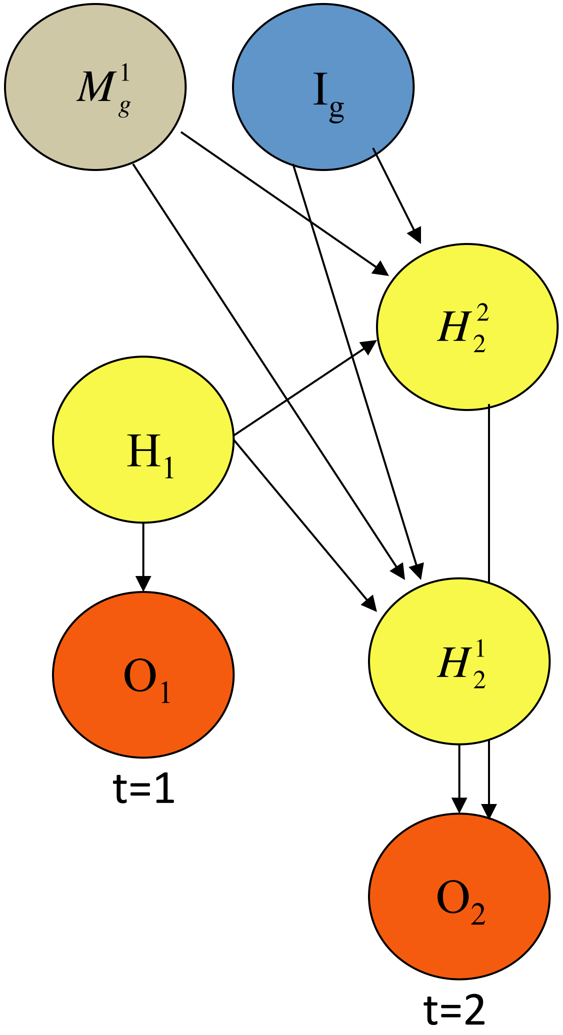
We can write it as a logistic regression classification problem!

$$P(H_2 = q(2) \mid H_1 = q(1), I_g) = \frac{1}{1 + \exp(w_0 + \underbrace{\sum_r}_{\text{Sum over all TFs}} I_{g,r} w_r + \underbrace{\sum_m}_{\text{Sum over all miRs}} M^1_{g,r} w_m)}$$

Optimize $w_{i'}$’s with a **constrained** logistic regression classifier

$$\hat{W} = \underset{W}{\operatorname{argmax}} \left[\underbrace{l(W)}_{\text{likelihood with parameters } W} + \underbrace{\ln p(W)}_{L_1\text{-penalty to promote sparsity}} \right]$$

s.t. $w_i \leq 0, i \in miRNAs$



Application to mouse data to understand human lung disease

Idiopathic pulmonary fibrosis (IPF)

Extensive lung fibrosis (source:wikipedia)

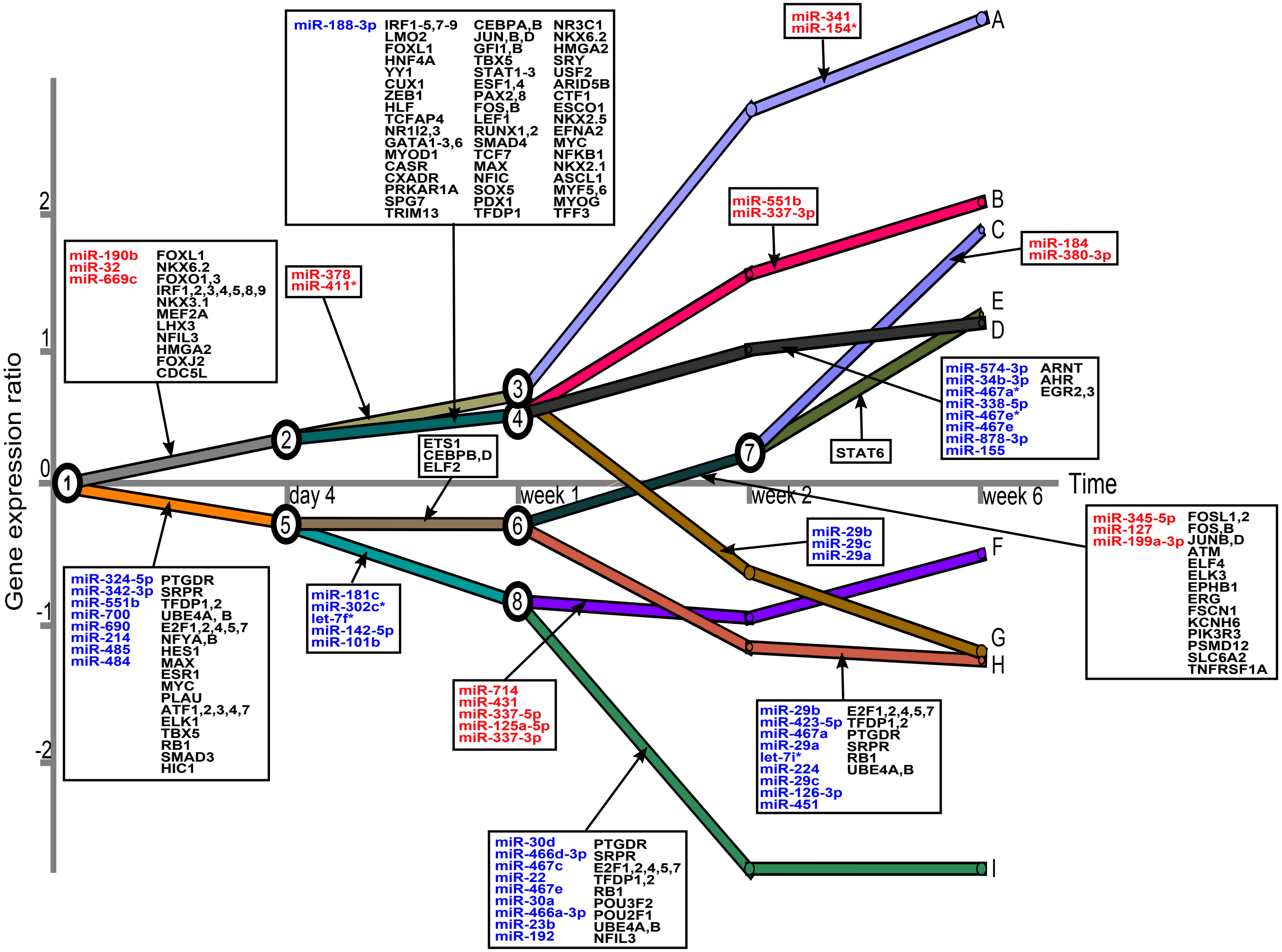
-100,000 people are affected (USA)

-about 30,000 new cases each year

-50% death rate after 3 years

-pathways for lung development appear activated in reverse direction during the disease

Joint dynamic network for lung development



miR-188-3p
IRF1-5,7-9 LMO2 FOXL1 HNF4A YY1 CUX1 ZEB1 HLF TCFAP4 NR1I2,3 GATA1-3,6 MYOD1 CASR CXADR PRKAR1A SPG7 TRIM13
CEBPA,B JUN,B,D GFI1,B TBX5 STAT1-3 ESF1,4 PAX2,8 FOS,B LEF1 RUNX1,2 SMAD4 TCF7 MAX NFIC SOX5 PDX1 TFDP1
NR3C1 NKX6.2 HMGA2 SRY USF2 ARID5B CTF1 ESCO1 NKX2.5 EFNA2 MYC NFKB1 NKX2.1 ASCL1 MYF5,6 MYOG TFF3

miR-341
miR-154*

miR-190b
miR-32
miR-669c
FOXL1 NKX6.2 FOXO1,3 IRF1,2,3,4,5,8,9 NKX3.1 MEF2A LHX3 NFIL3 HMGA2 FOXJ2 CDC5L

miR-378
miR-411*

miR-551b
miR-337-3p

miR-184
miR-380-3p

ETS1
CEBPB,D
ELF2

miR-574-3p
miR-34b-3p
miR-467a*
miR-338-5p
miR-467e*
miR-467e
miR-878-3p
miR-155
ARNT AHR EGR2,3

STAT6

miR-345-5p
miR-127
miR-199a-3p
FOSL1,2 FOS,B JUNB,D ATM ELF4 ELK3 EPHB1 ERG FSCN1 KCNH6 PIK3R3 PSMD12 SLC6A2 TNFRSF1A

miR-324-5p
miR-342-3p
miR-551b
miR-700
miR-690
miR-214
miR-485
miR-484
PTGDR SRPR TFDP1,2 UBE4A, B E2F1,2,4,5,7 NFYA,B HES1 MAX ESR1 MYC PLAU ATF1,2,3,4,7 ELK1 TBX5 RB1 SMAD3 HIC1

miR-181c
miR-302c*
let-7f*
miR-142-5p
miR-101b

miR-29b
miR-29c
miR-29a

miR-714
miR-431
miR-337-5p
miR-125a-5p
miR-337-3p

miR-29b
miR-423-5p
miR-467a
miR-29a
let-7i*
miR-224
miR-29c
miR-126-3p
miR-451
E2F1,2,4,5,7 TFDP1,2 PTGDR SRPR RB1 UBE4A,B

miR-30d
miR-466d-3p
miR-467c
miR-22
miR-467e
miR-30a
miR-466a-3p
miR-23b
miR-192
PTGDR SRPR E2F1,2,4,5,7 TFDP1,2 RB1 POU3F2 POU2F1 UBE4A,B NFIL3

- 22 out 56 miRNAs predicted by the method are differentially expressed in patient cohorts with the IPF lung disease
- different to development miR-30d is down regulated in IPF patients