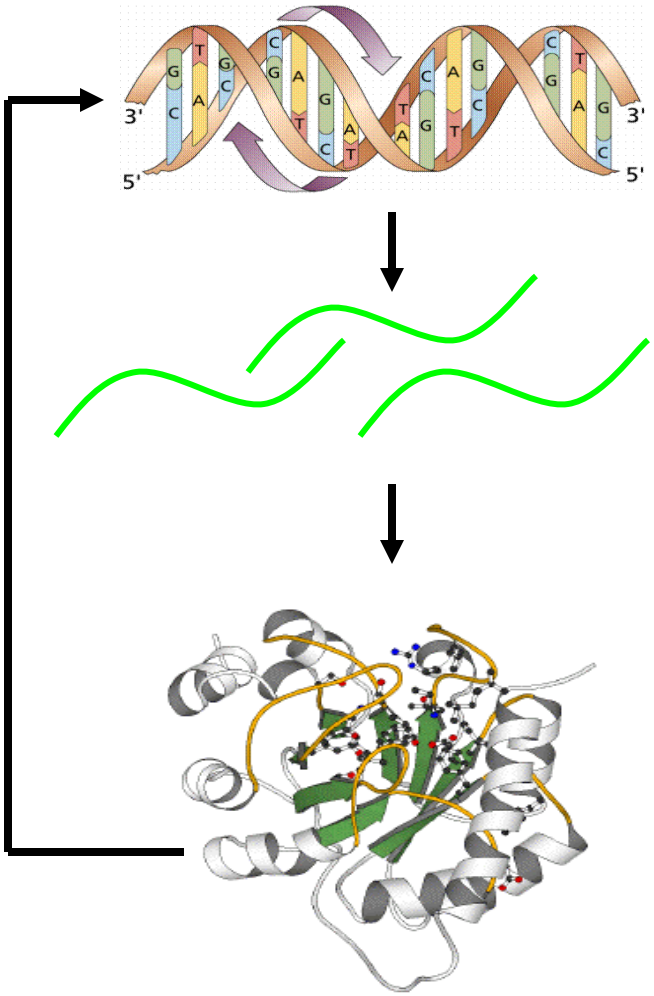
10-701
Machine Learning

HMM applications in computational biology

Central dogma



DNA

transcription

mRNA

translation

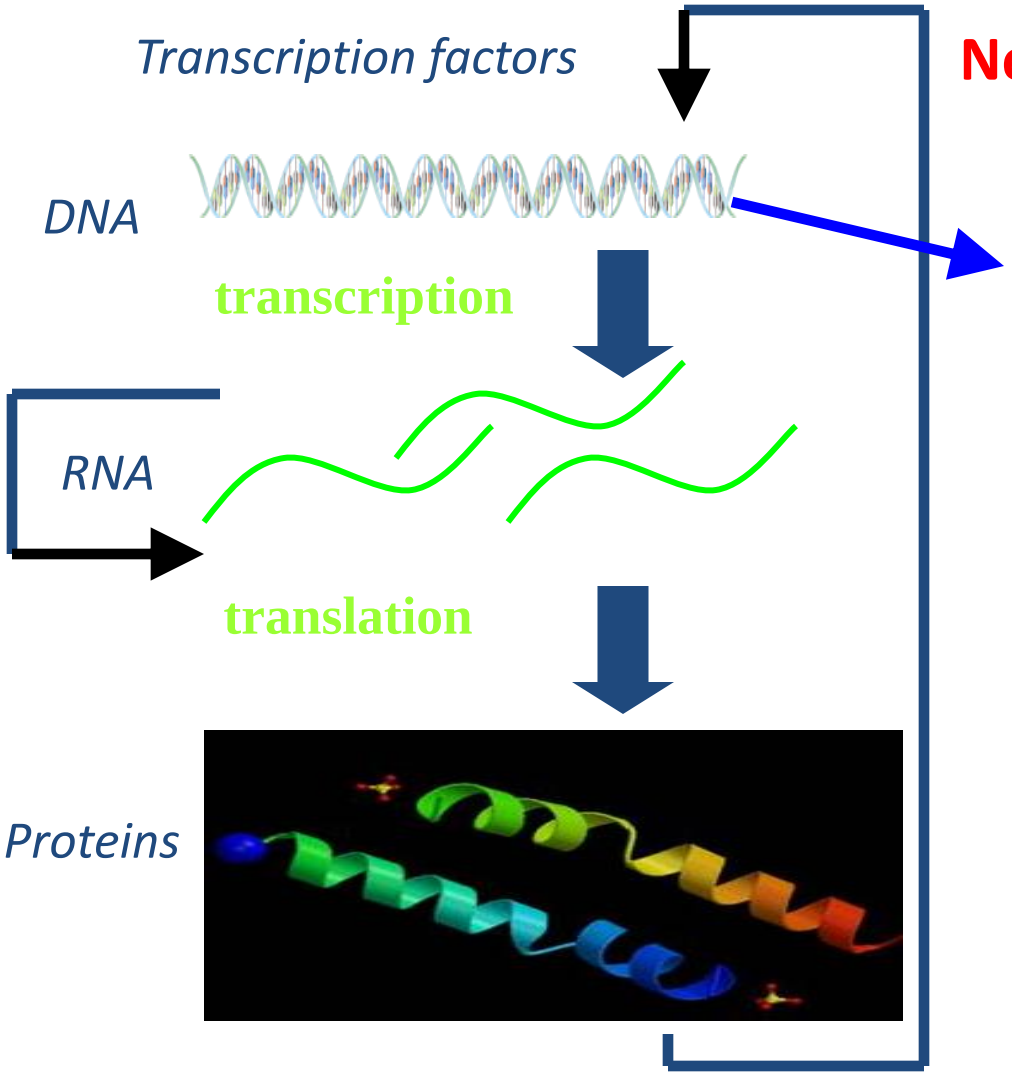
Protein

CCTGAGCCAACTATTGATGAA

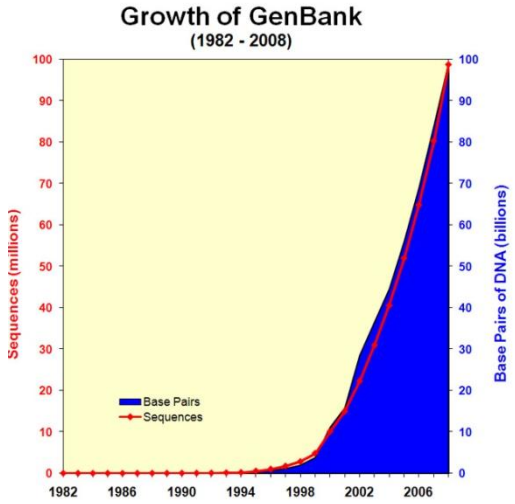
CCUGAGCCAACUAUUGAUGAA

PEPTIDE

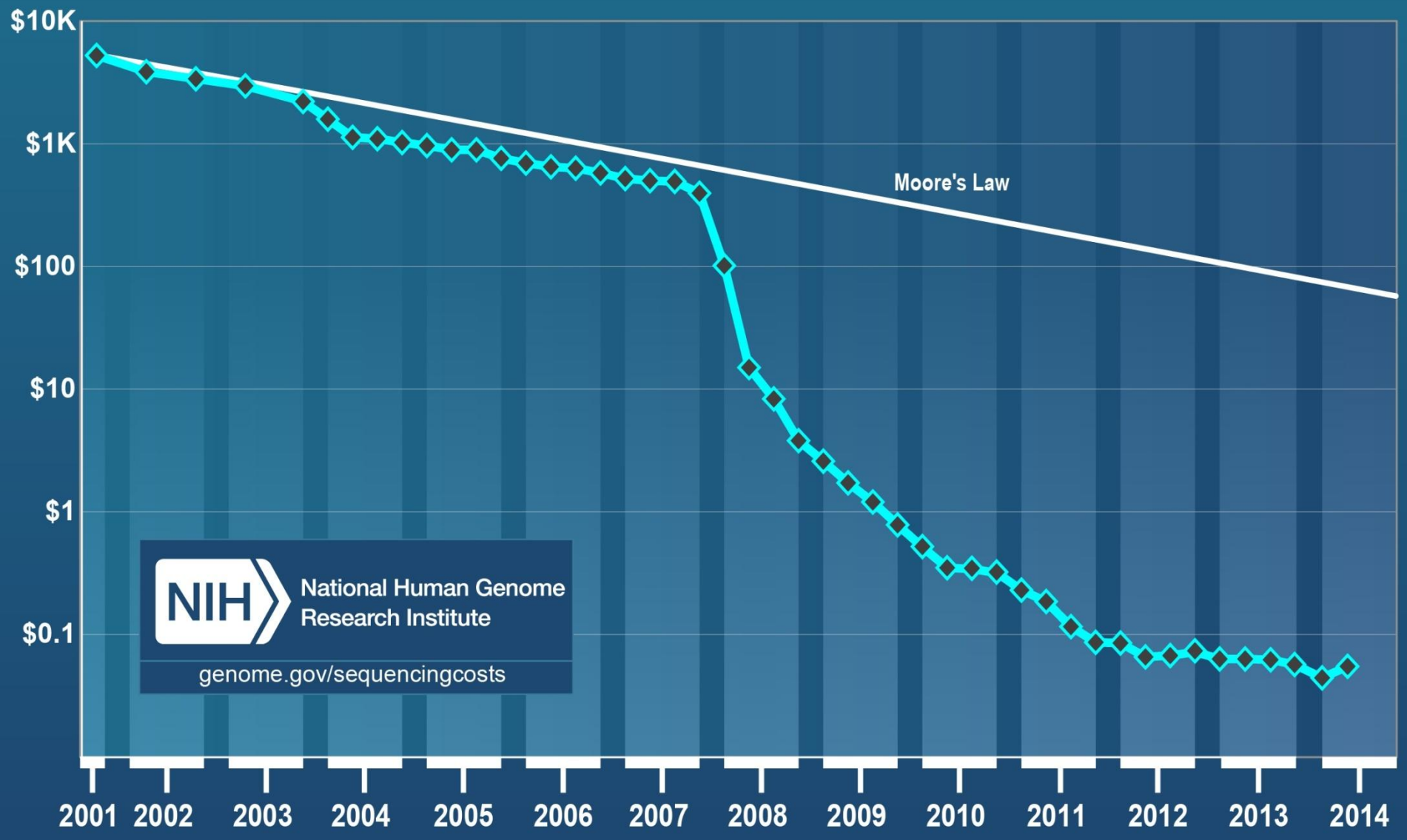
Biological data is rapidly accumulating



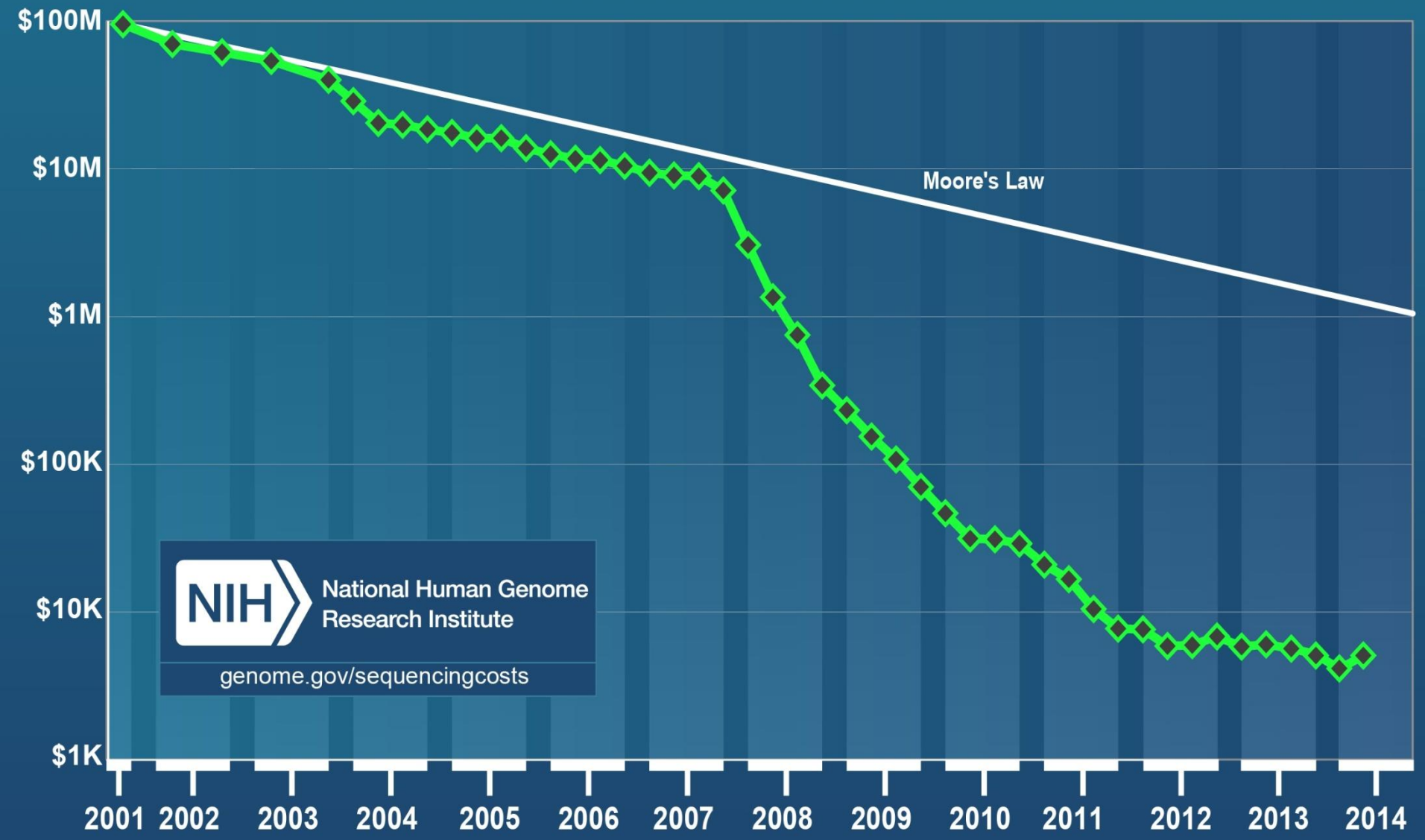
Next generation sequencing



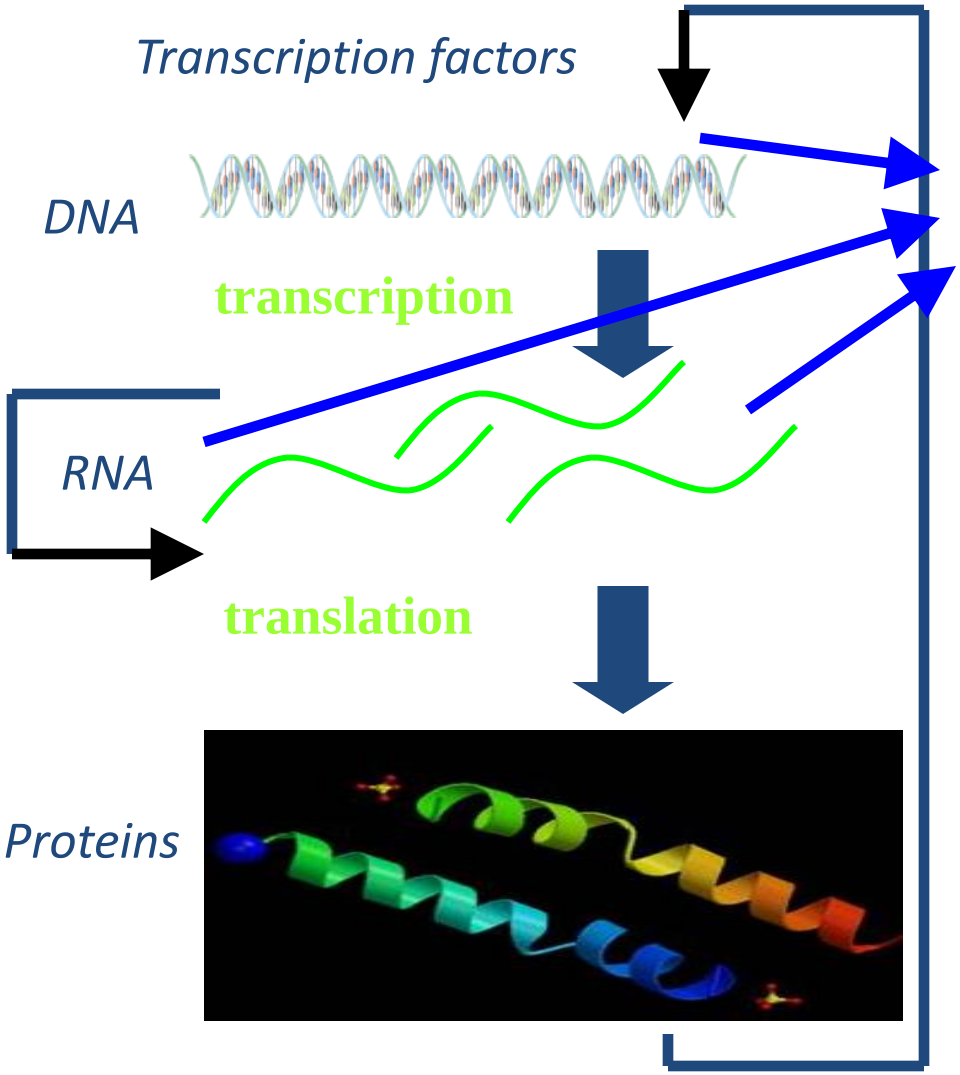
Cost per Raw Megabase of DNA Sequence



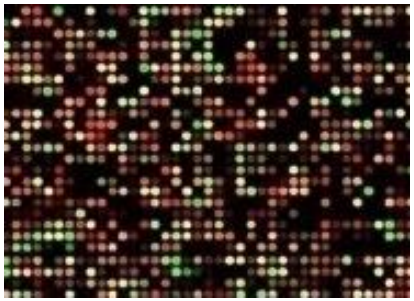
Cost per Genome

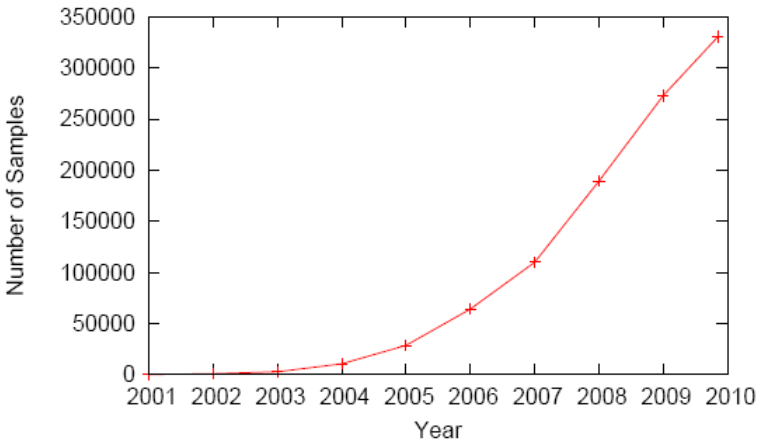


Biological data is rapidly accumulating

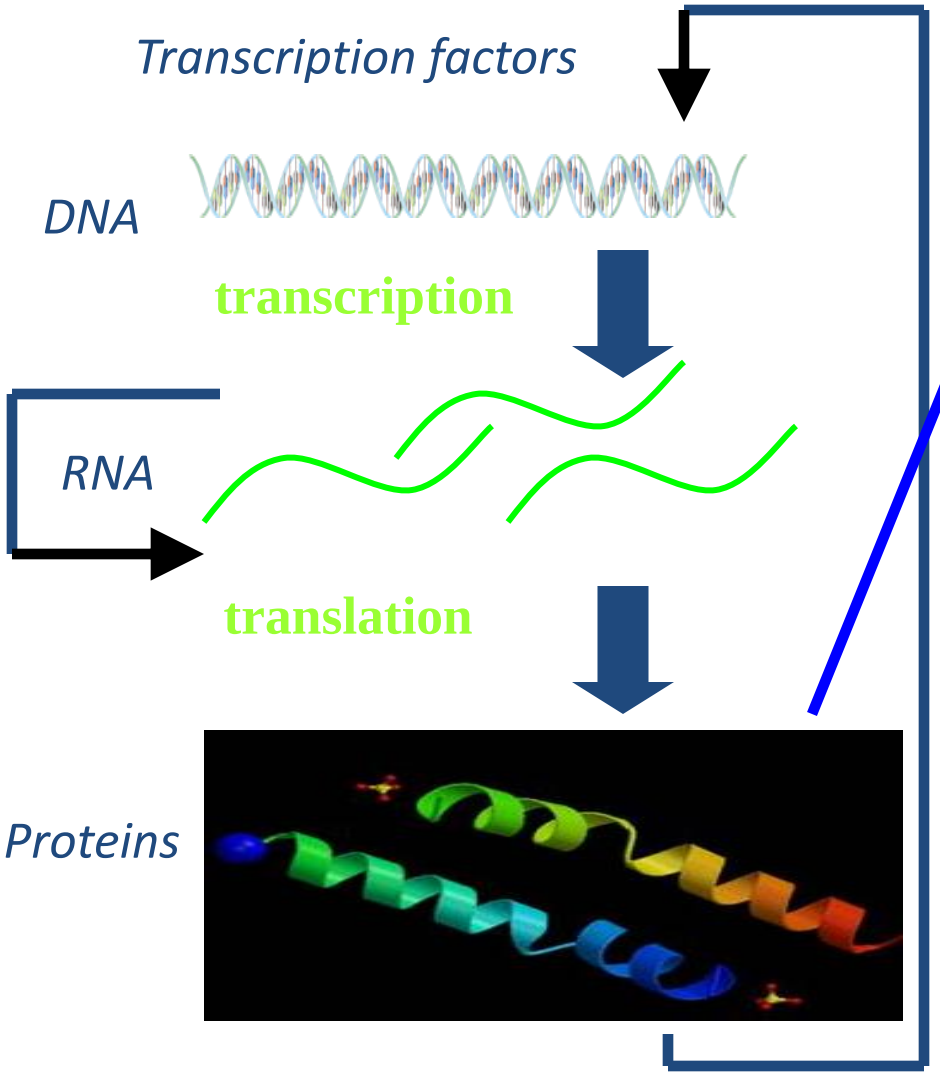


Array / sequencing technology

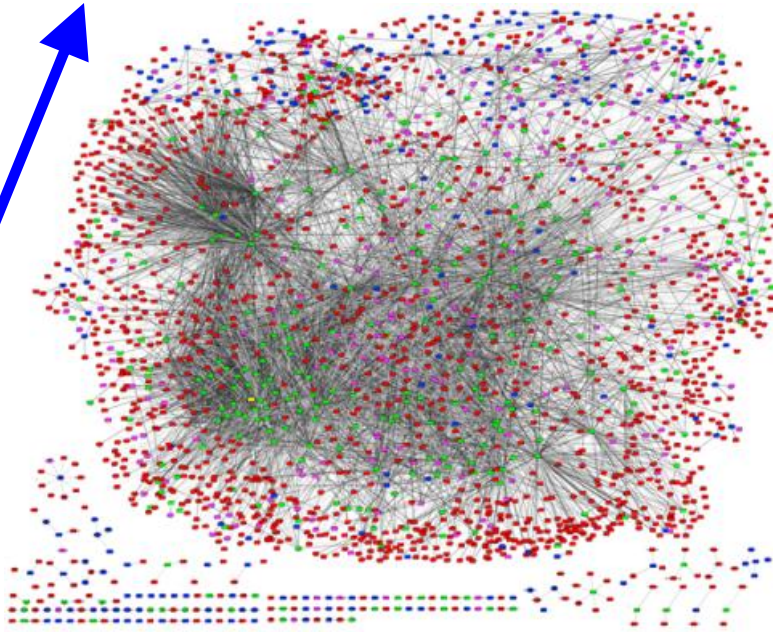




Biological data is rapidly accumulating



Protein interactions



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

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

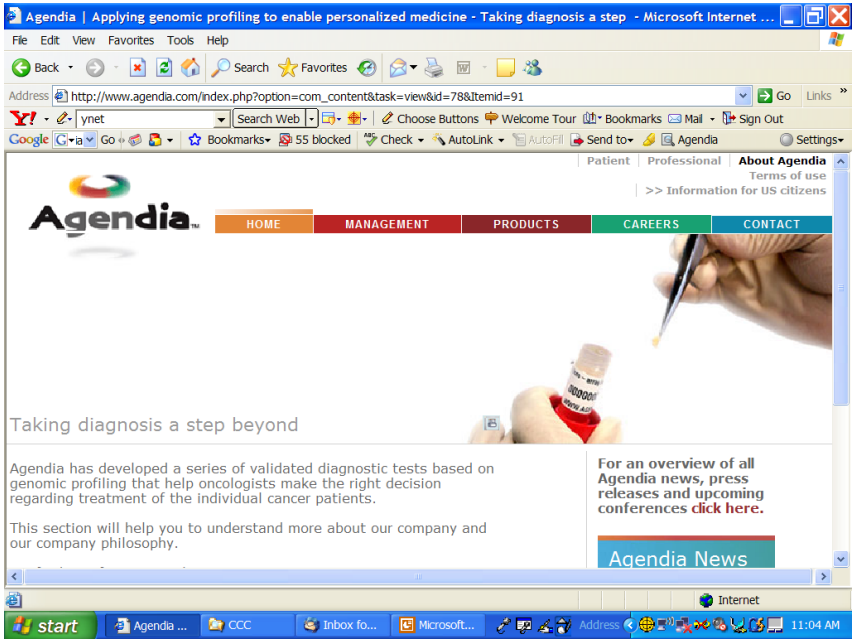
By ANDREW POLLACK
Published: February 17, 2012

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

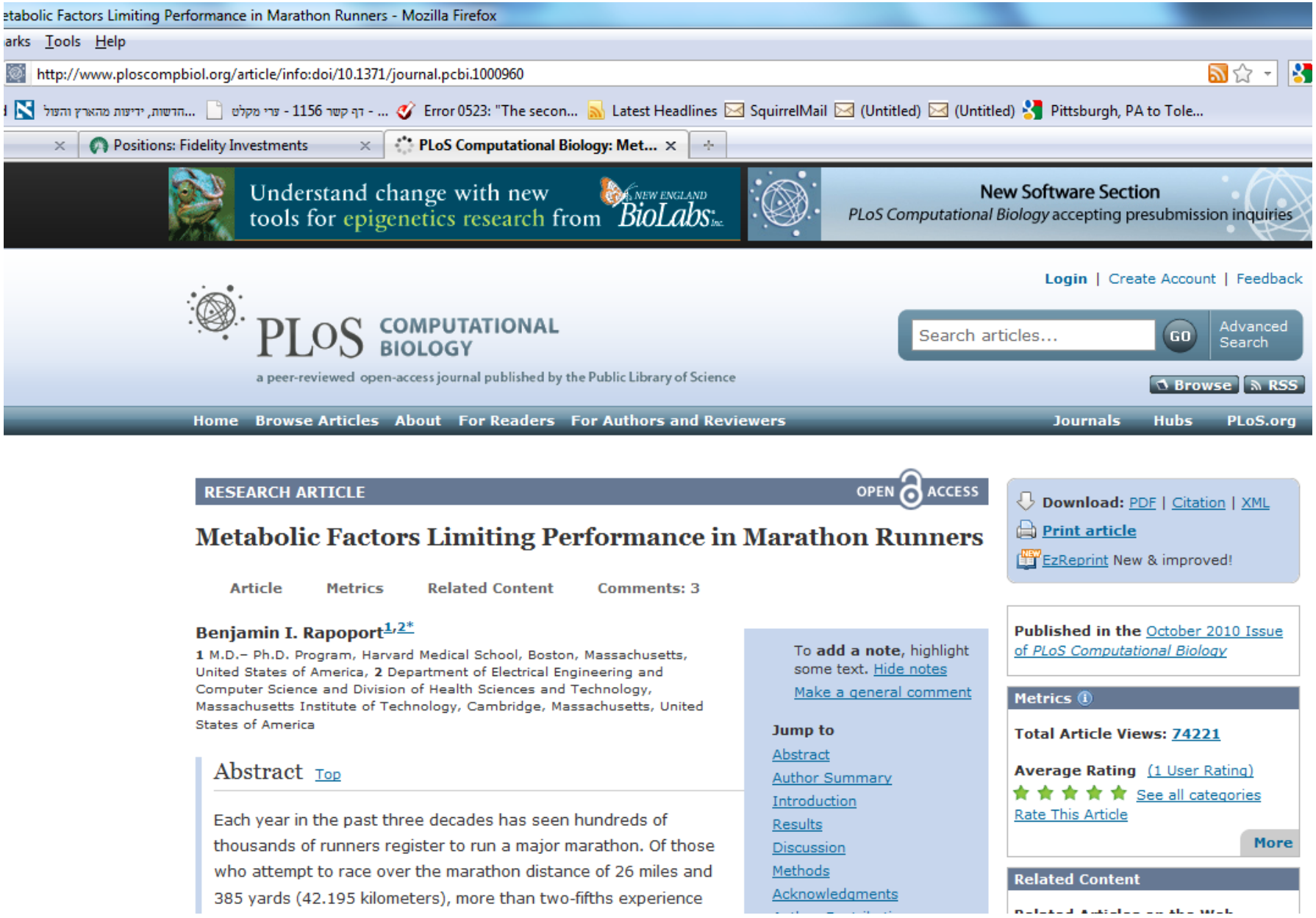
A British company said on Friday that by the end of the year it would begin selling a disposable gene sequencing device that is the size of a USB memory stick and plugs into a laptop computer to deliver its

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."



*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
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
Metabolic Factors Limiting Performance in Marathon Runners
Article Metrics Related Content Comments: 3
Benjamin I. Rapoport
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
Jump to
Abstract
Author Summary
Introduction
Results
Discussion
Methods
Acknowledgments
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
Download: PDF | Citation | XML
Print article
EzReprint New & improved!
Published in the October 2010 Issue of PLoS Computational Biology
Metrics
Total Article Views: 74221
Average Rating (1 User Rating)
See all categories
Rate This Article
More
Related Content

Active Learning

nature International weekly journal of science

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]

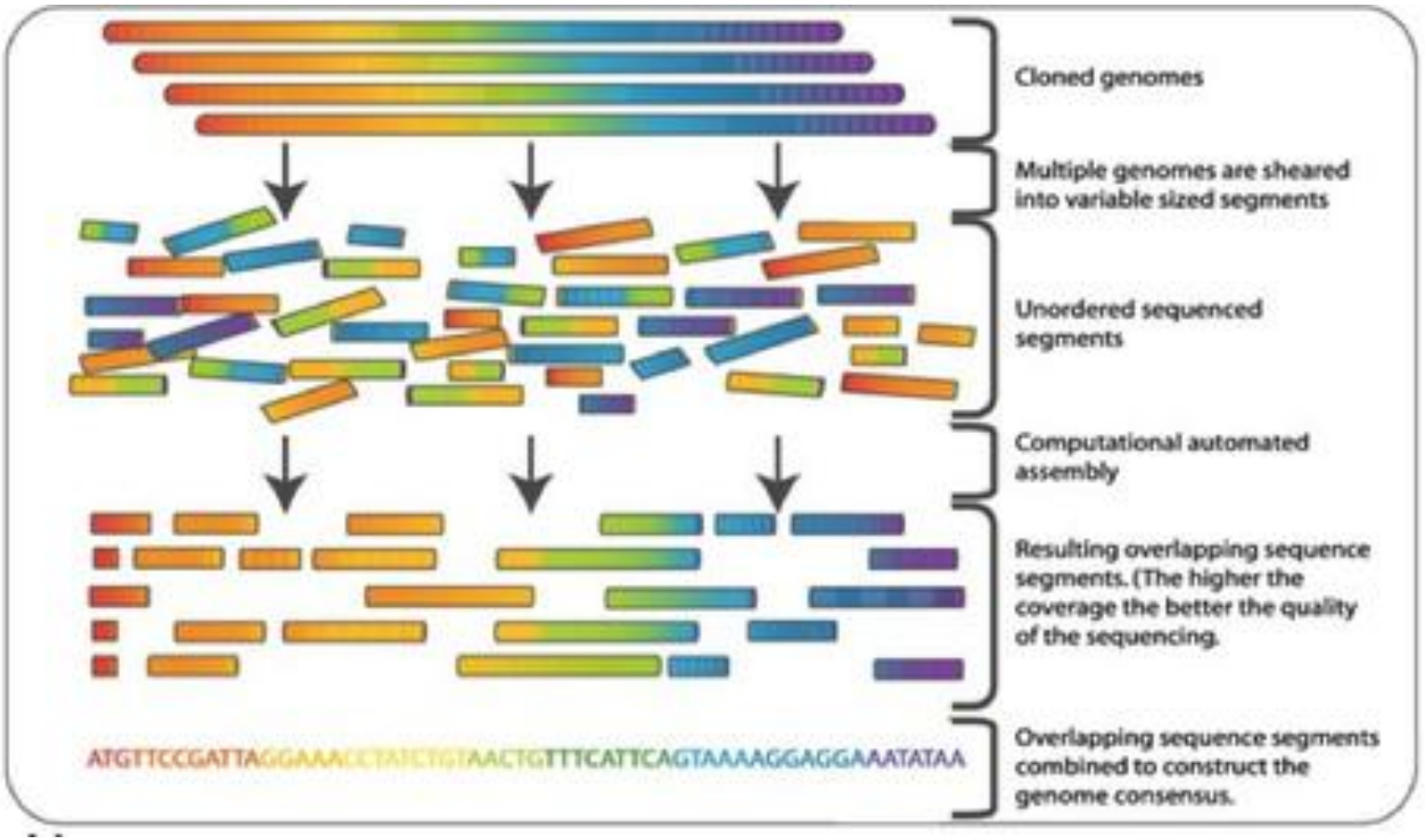
1. Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK
2. School of Computing, The Robert Gordon University, Aberdeen AB10 1FR, UK
3. Department of Computing, Imperial College, London SW7 2AZ, UK
4. Department of Chemistry, UMIST, P.O. Box 88, Manchester M60 1QD, UK

Sequencing DNA

First human genome draft in 2001

Due to *accumulated errors*, we could only reliably read at most **300-500 nucleotides.**

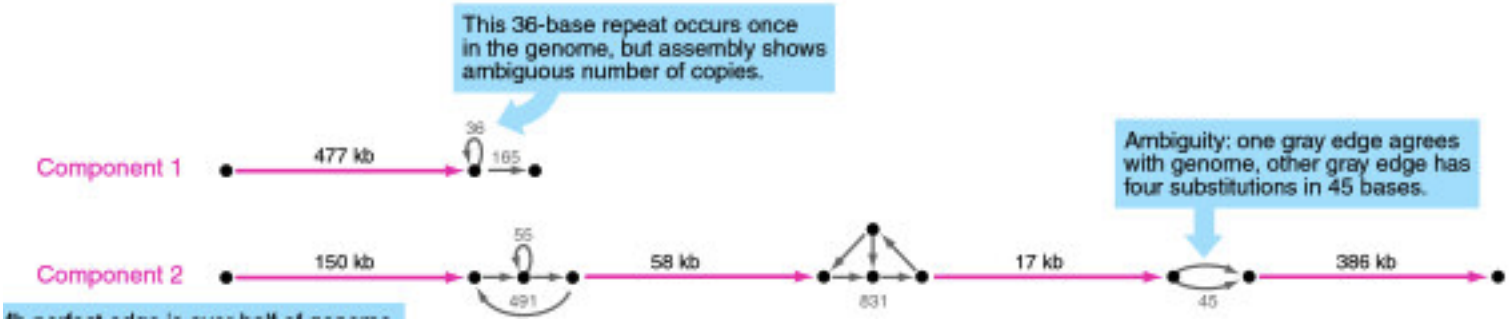
Shotgun Sequencing



Wikipedia

Caveats

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

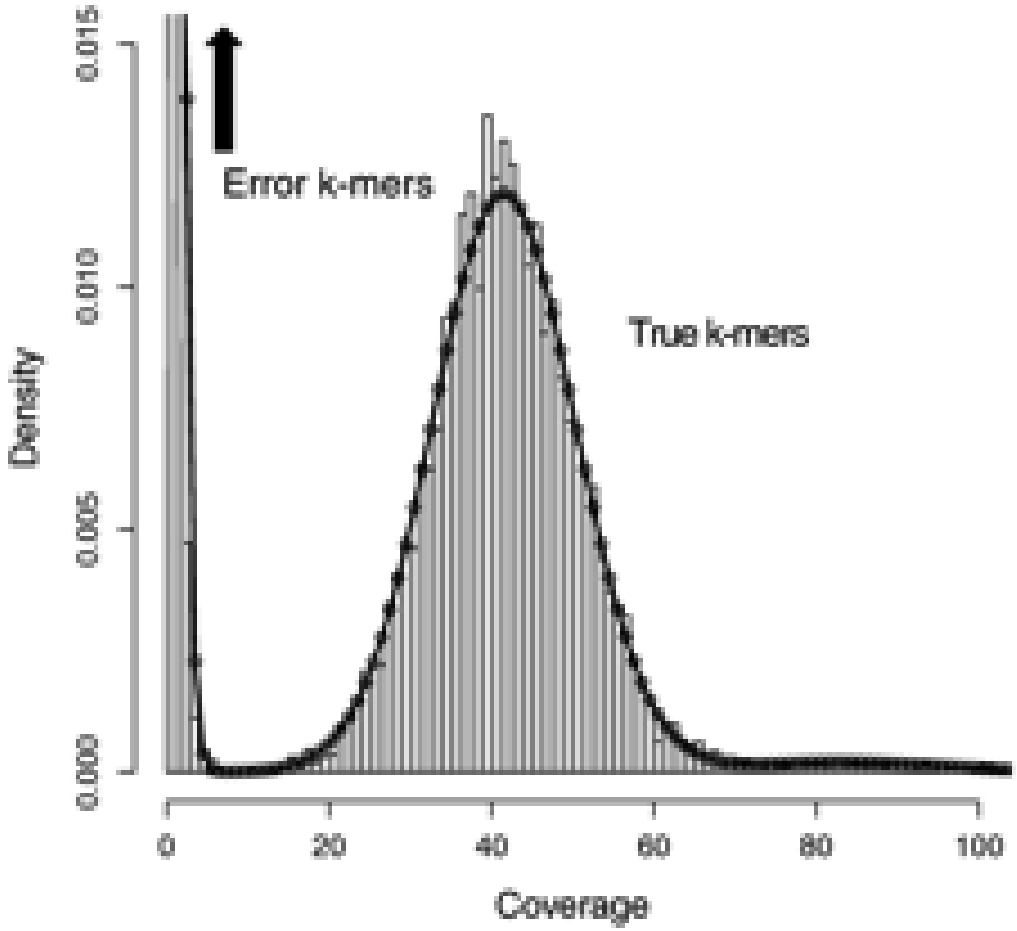


MacCallum et al., GB 2009

Error Correction in DNA sequencing

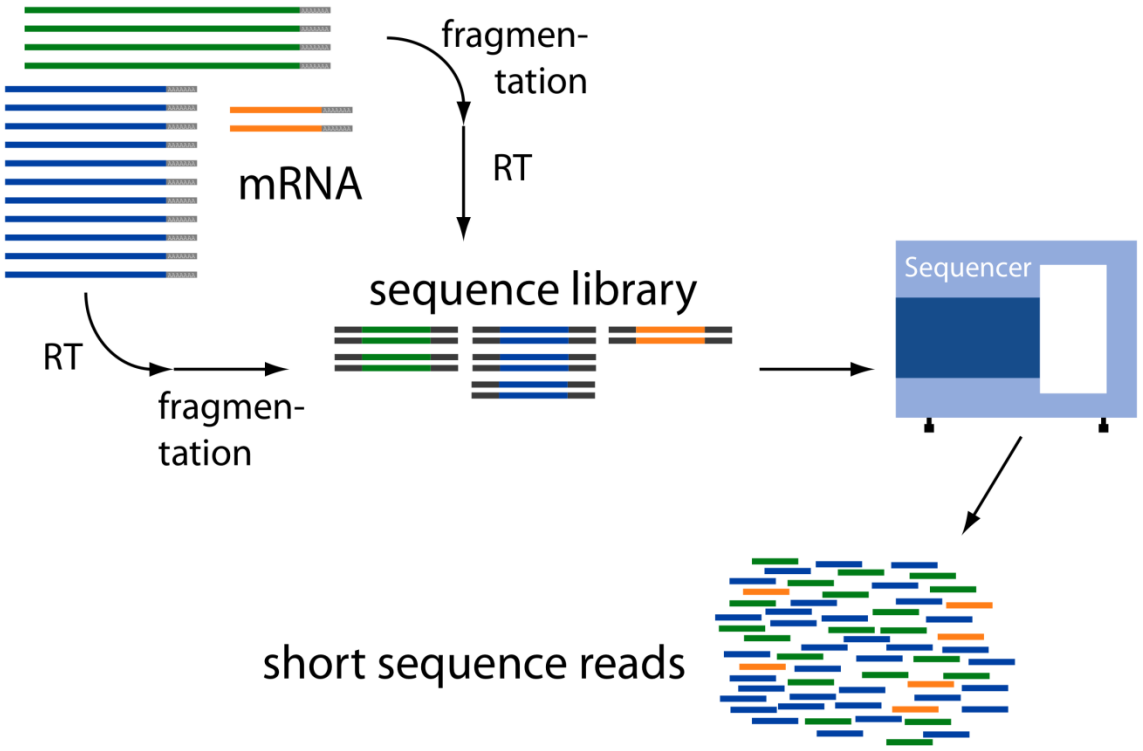
- The fragmentation happens at random locations of the molecules. We expect all positions in the genome to have the same # number of reads

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 molecule 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.

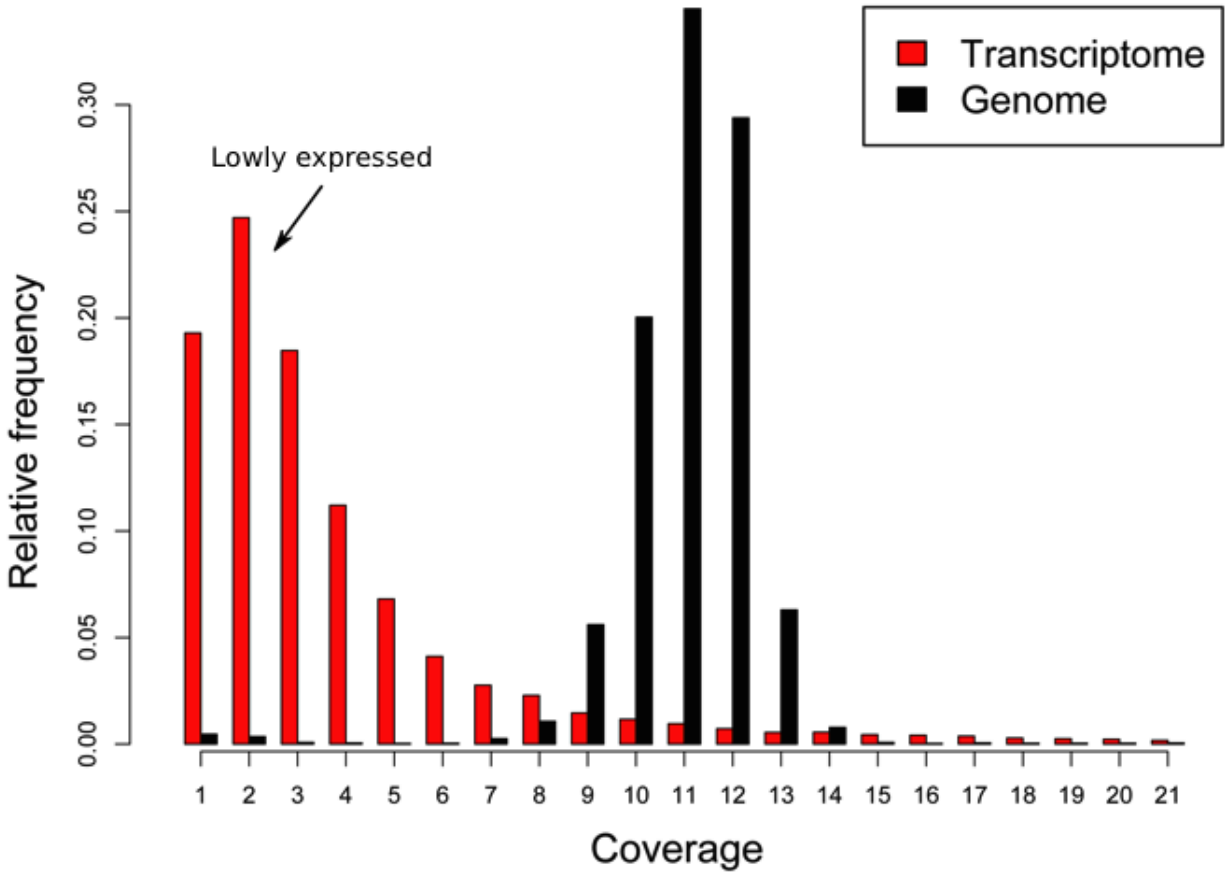
Challenges

- 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 splicing, RNA editing, post-transcription modification.

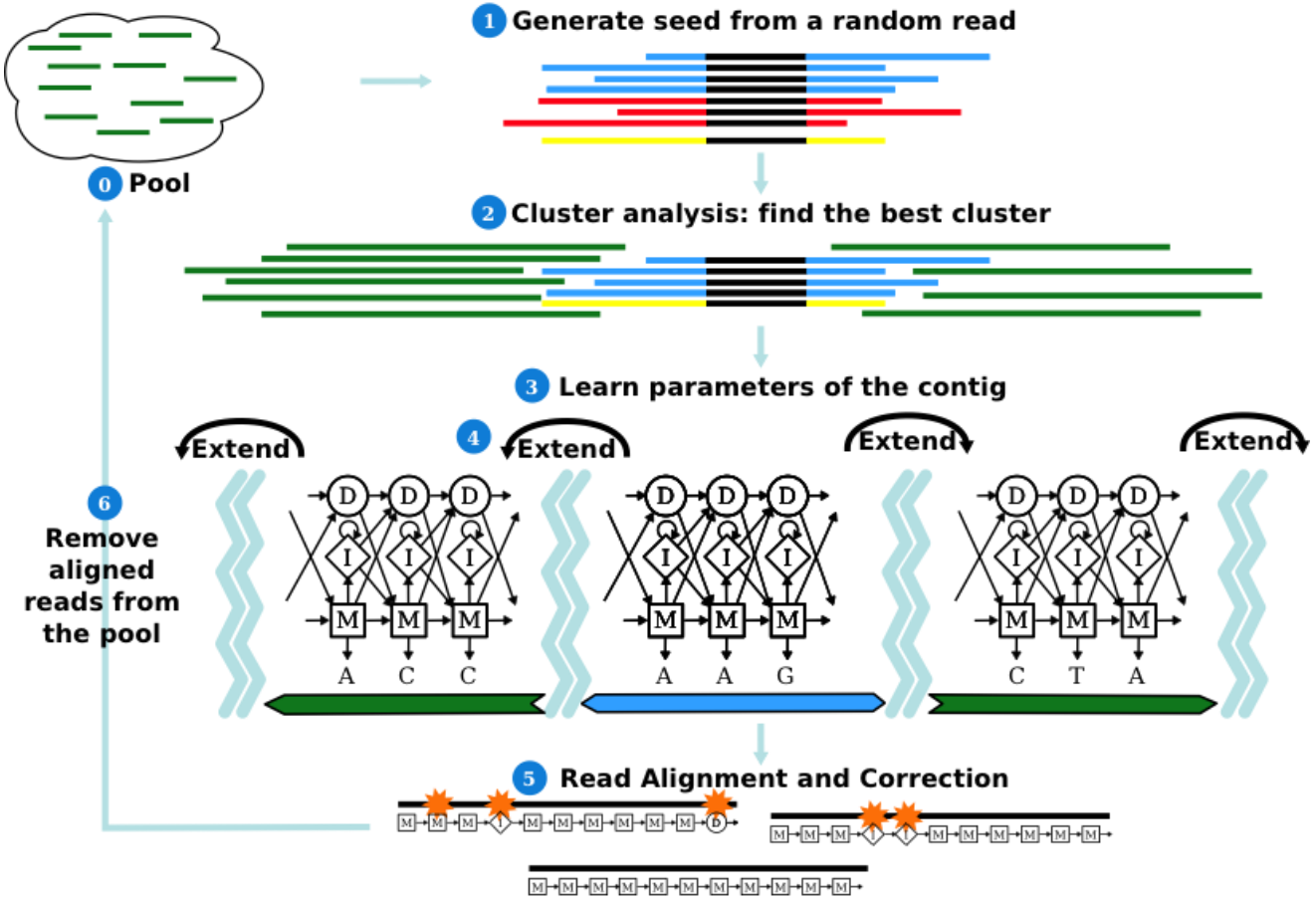
Errors are non uniformly distributed

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

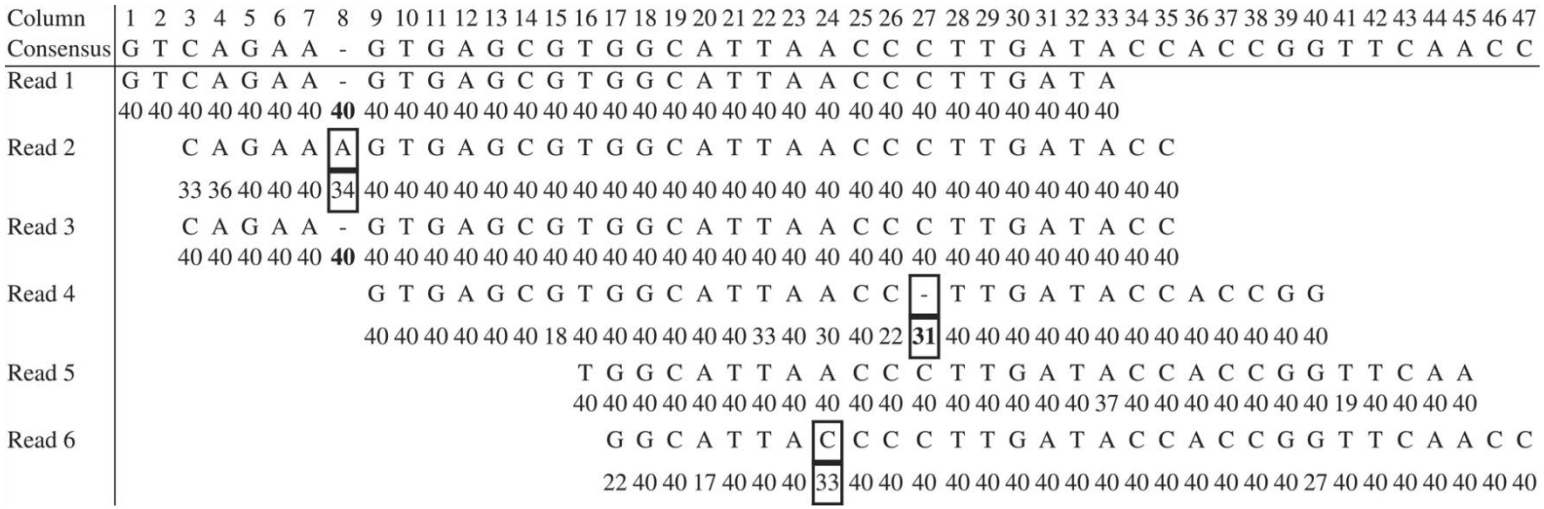


SEECER

Error Correction + Consensus sequence estimation for RNA-Seq data



Key idea: HMM model

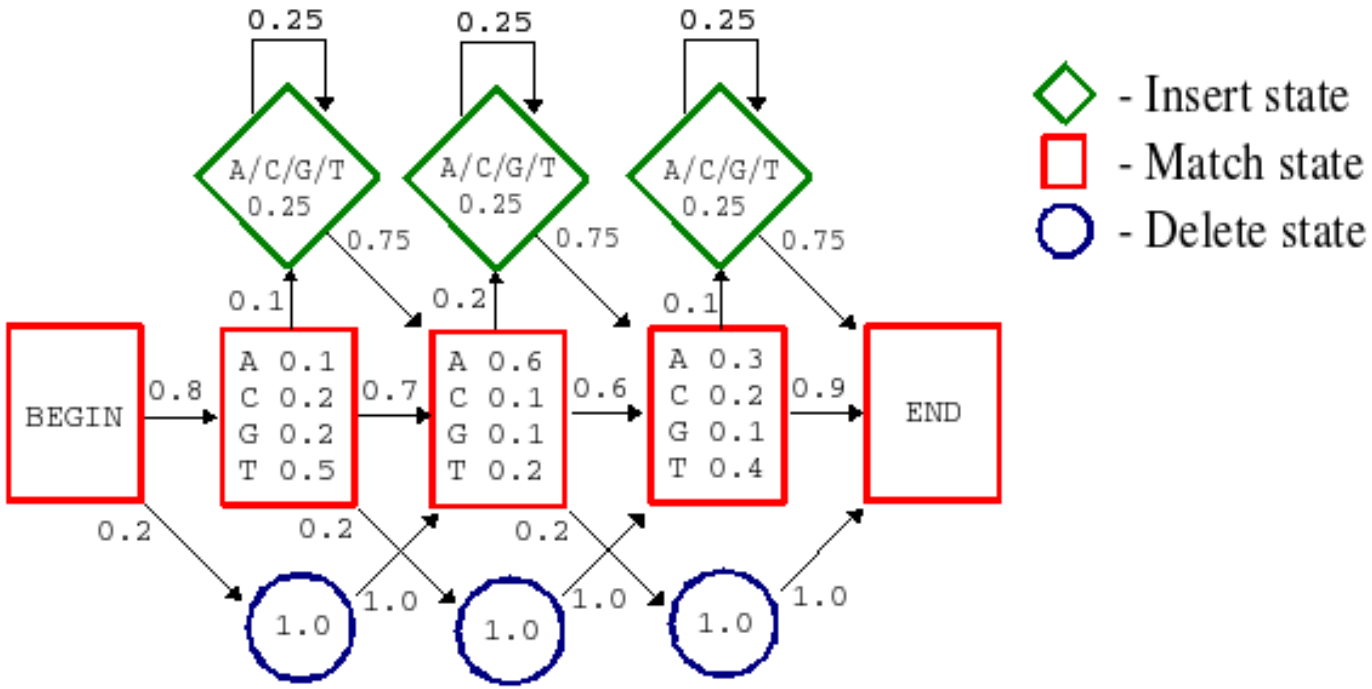


Salmela et al., Bioinformatics 2011

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	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																																															
Read 5																																															
Read 6																																															



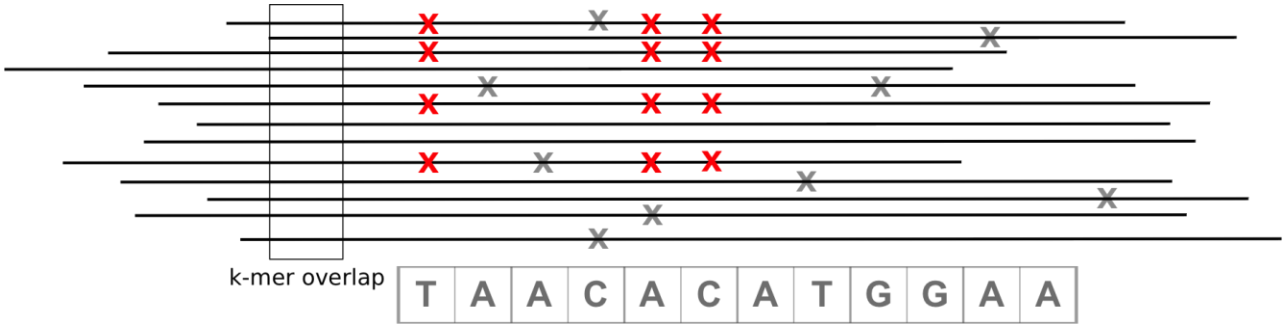
Building (Learning) the HMMs and Making Corrections (Inference)

Learning = Expectation-Maximization
Inference = Viterbi algorithm

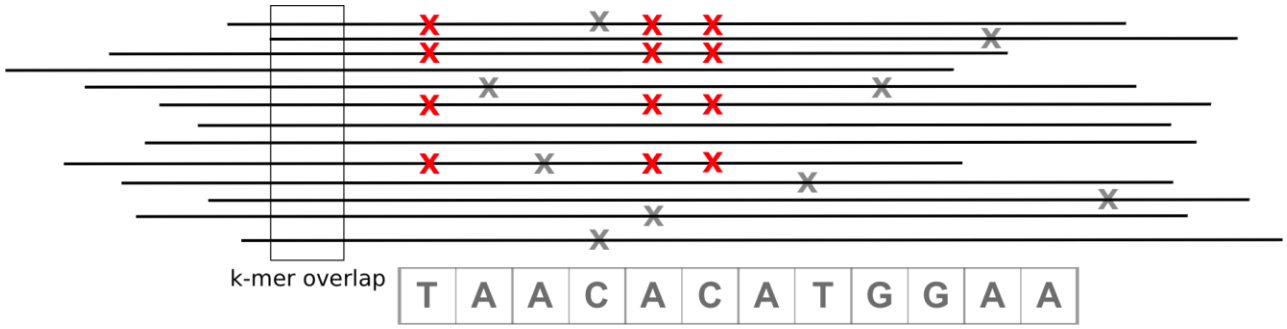
Seeding:
Guessing possible reads using k-mer overlaps.
Constructing the HMM from these reads.

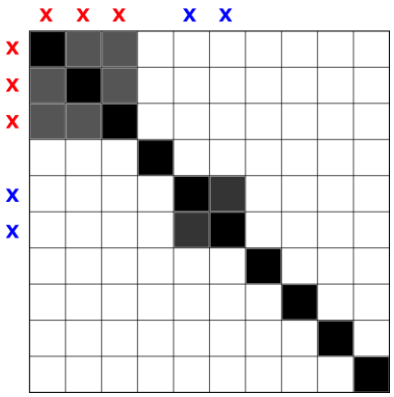
Speed up:
The k-mer overlaps yield approximate multiple alignments of reads.
We can learn HMM parameters from this directly.

Clustering to improve seeding



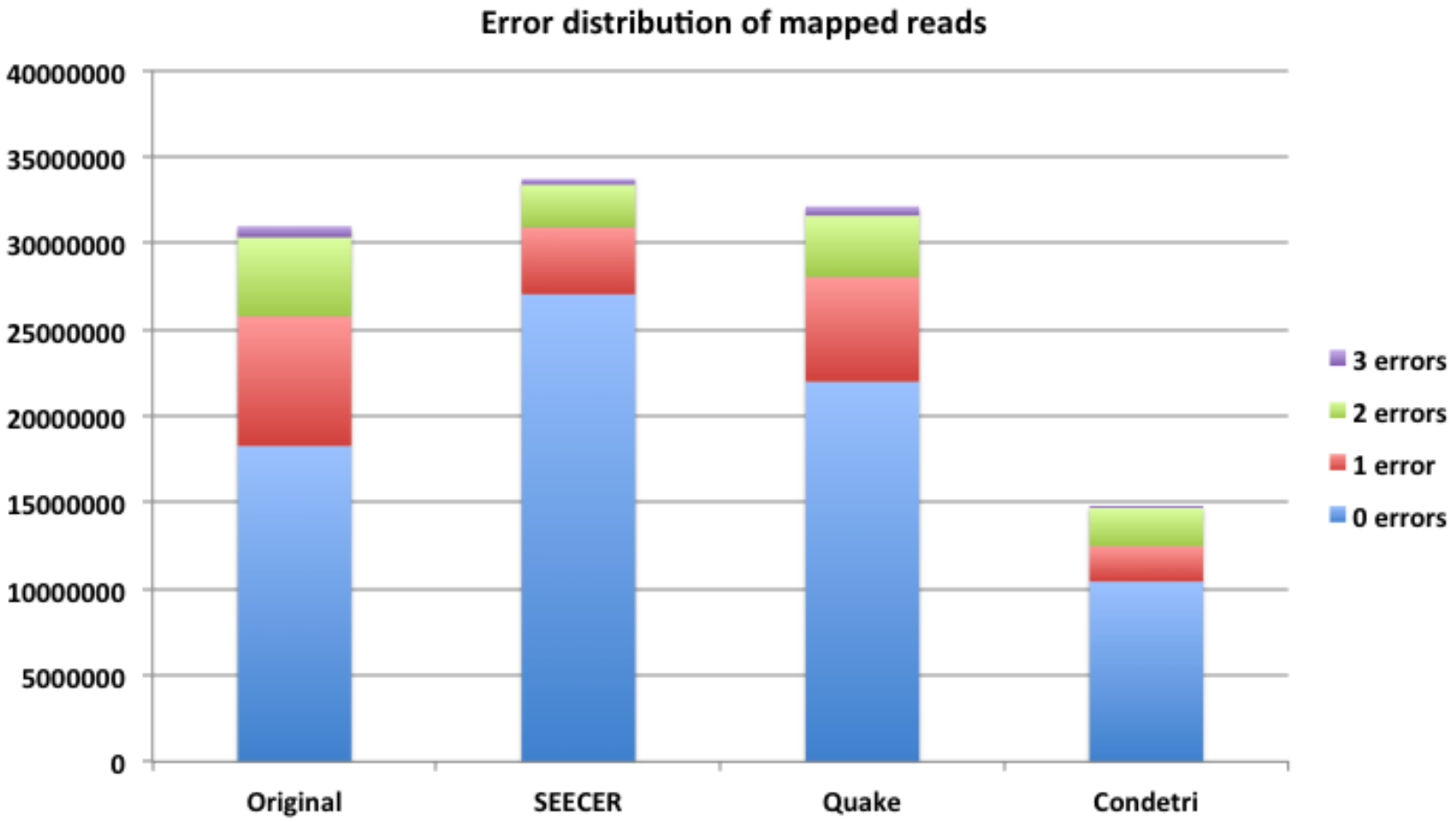
Real biological differences should be supported by a set of reads with similar mismatches to the consensus



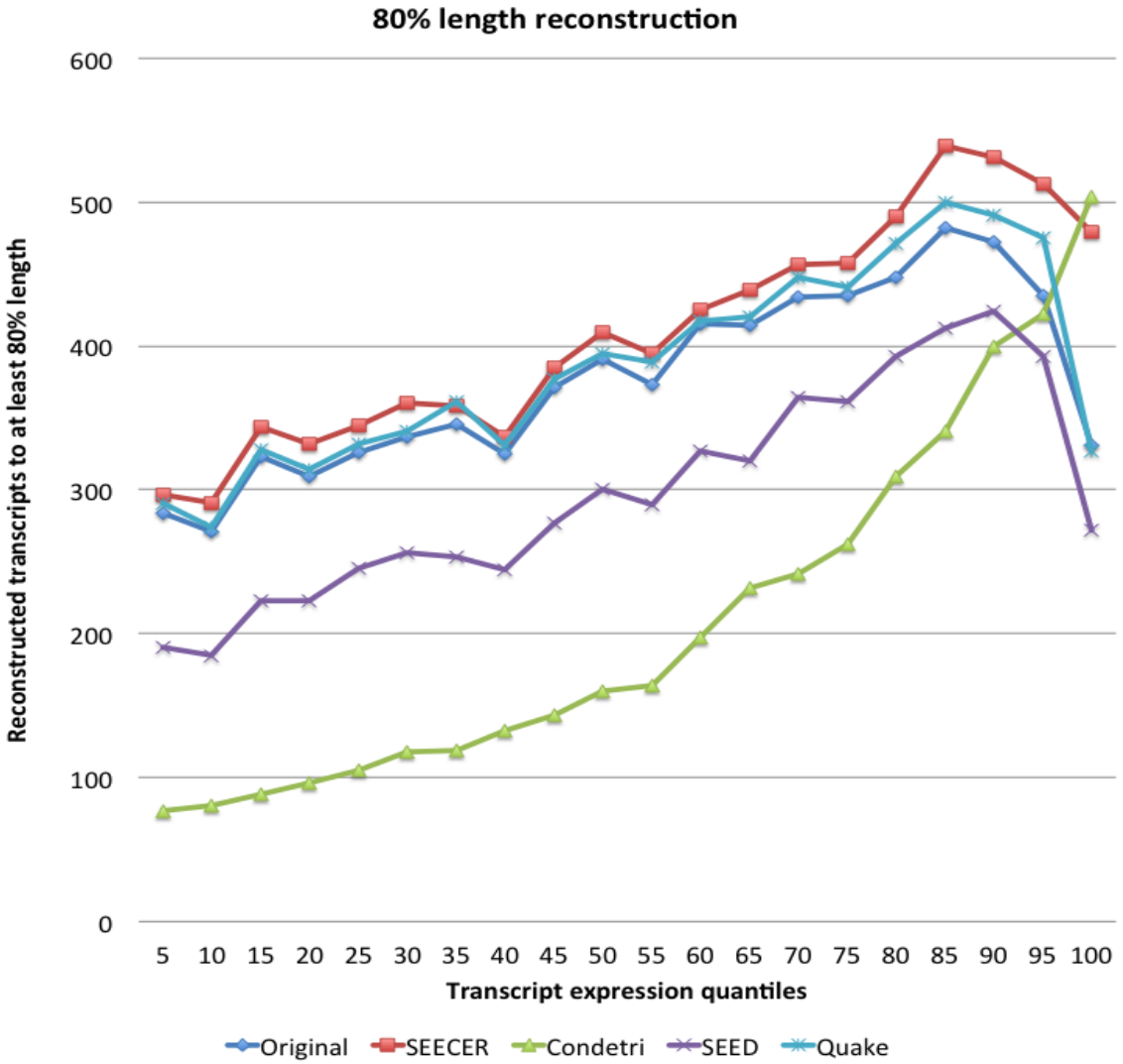


1. Clustering positions with mismatches to identify clusters of correlated positions.
2. Build a similarity matrix between these positions.
3. Use Spectral clustering to find clusters of correlated positions.
4. Filter reads have mismatches in these clusters.

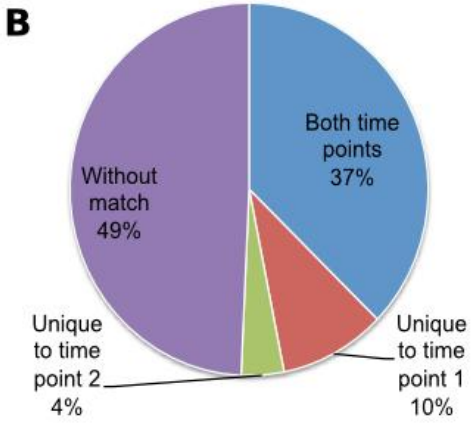
Comparison to other methods

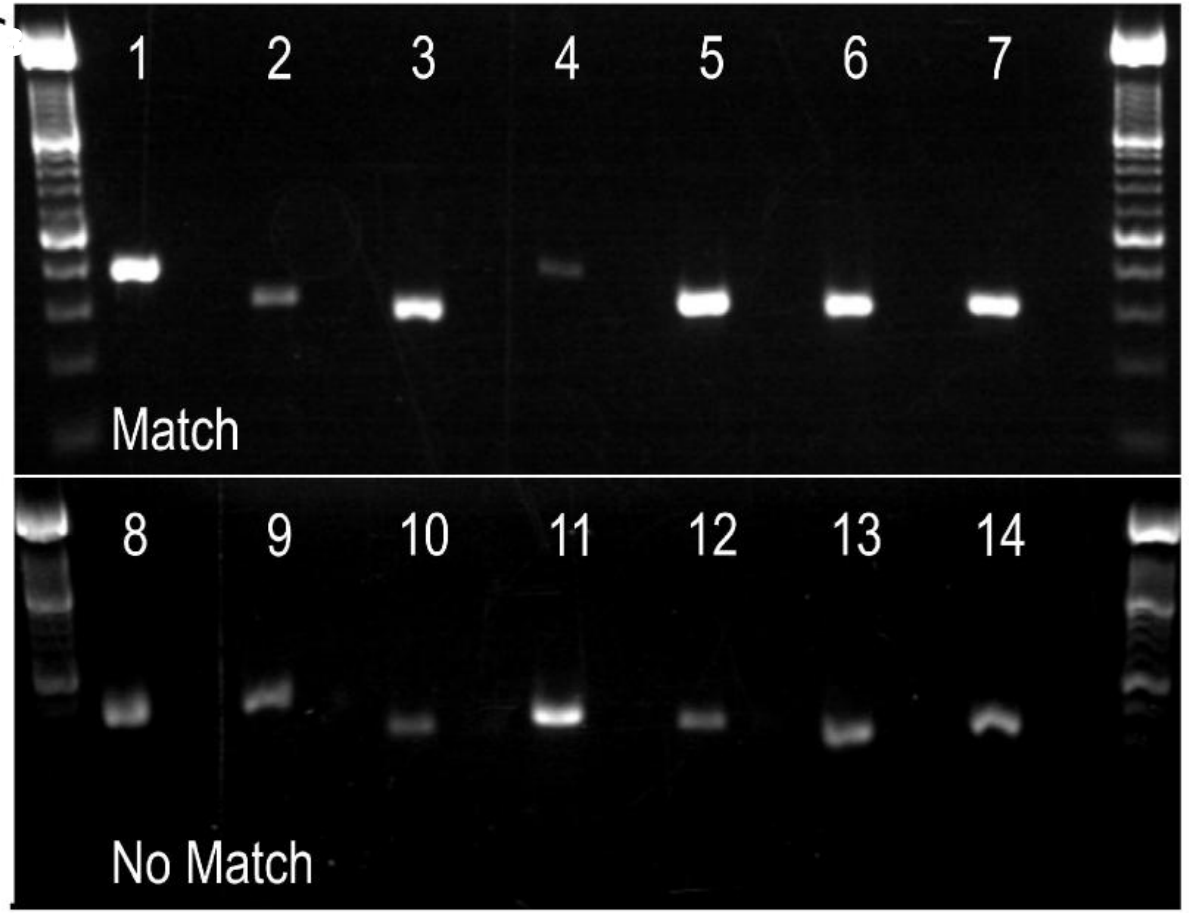


Using the corrected reads, the assembler can recover more transcripts



Analysis of sea cucumber data

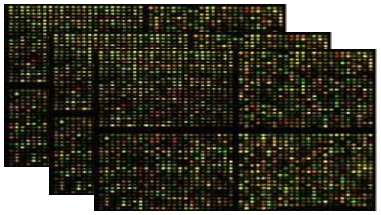


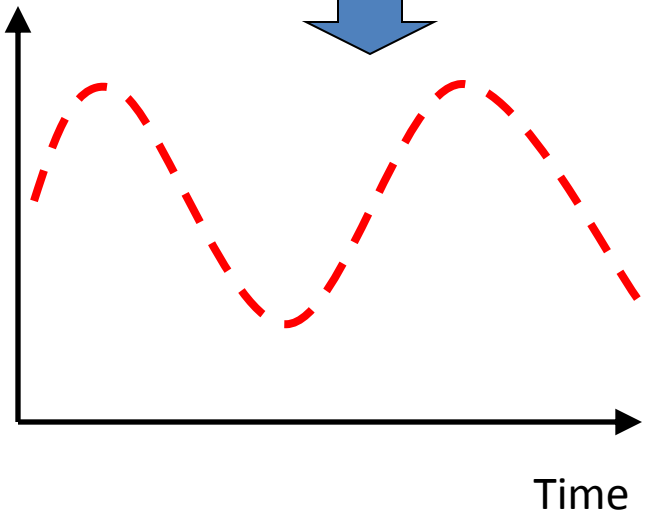


Data integration in biology

Key problem: Most high-throughput data is static

Time-series measurements



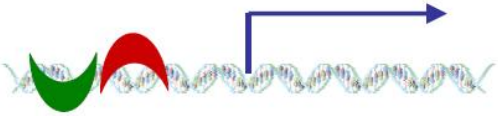


Static data sources

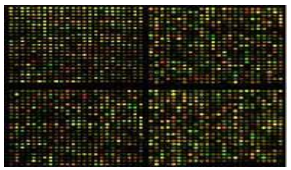
Sequencing

motif

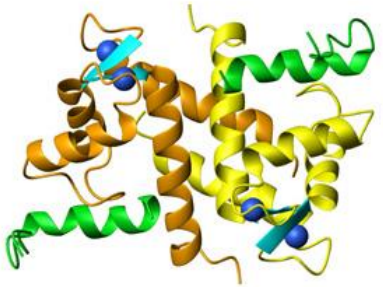
CHIP-chip



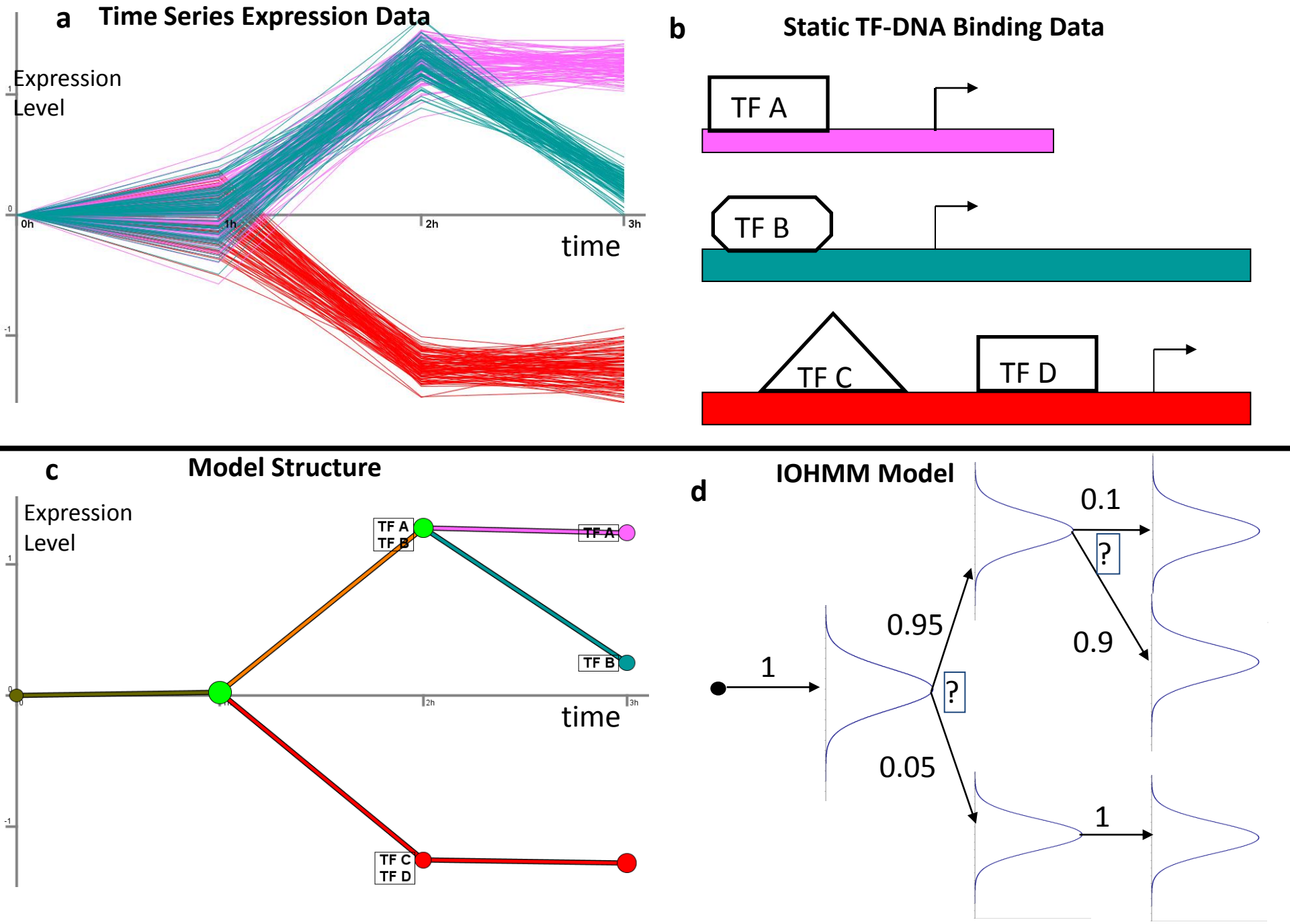
microarray



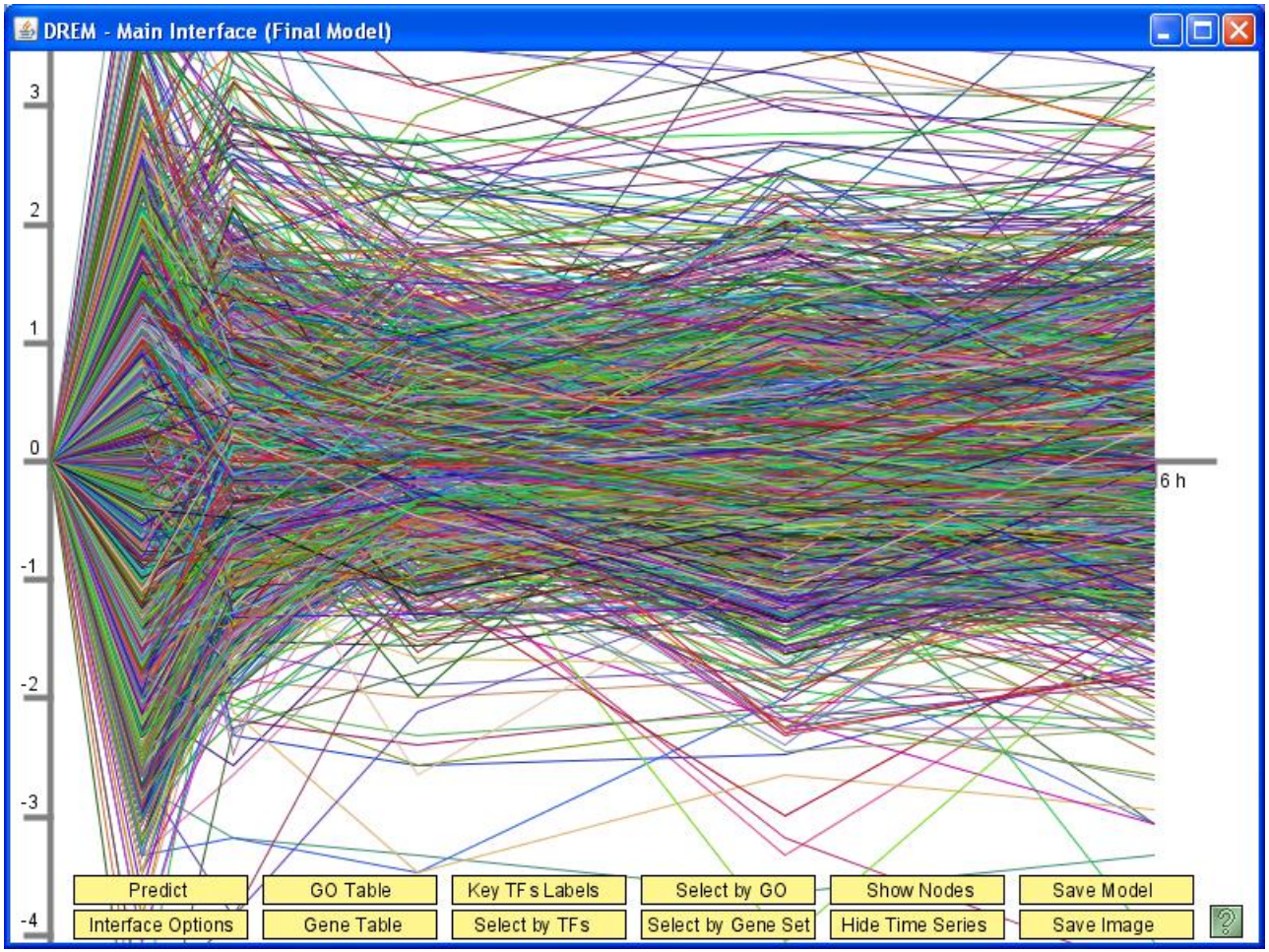
PPI



DREM: Dynamic Regulatory Events Miner


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

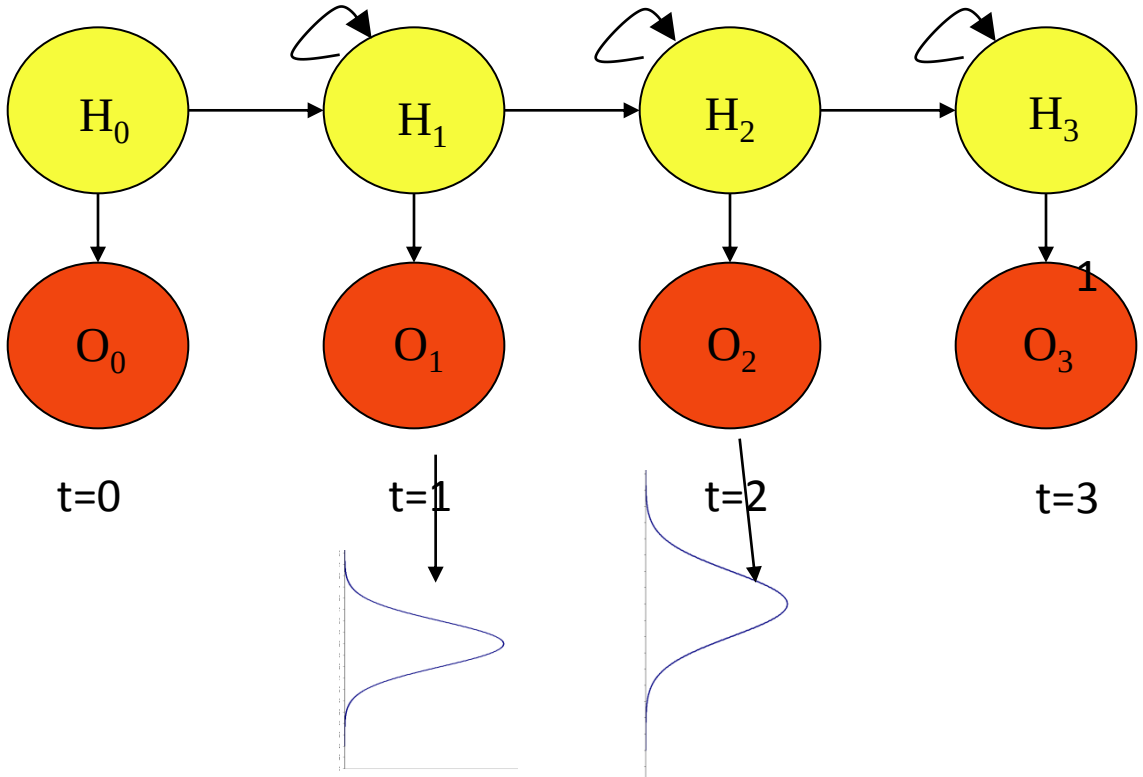
Things are a bit more complicated: Real data



A Hidden Markov Model

Hidden States

Observed outputs
(expression levels)



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

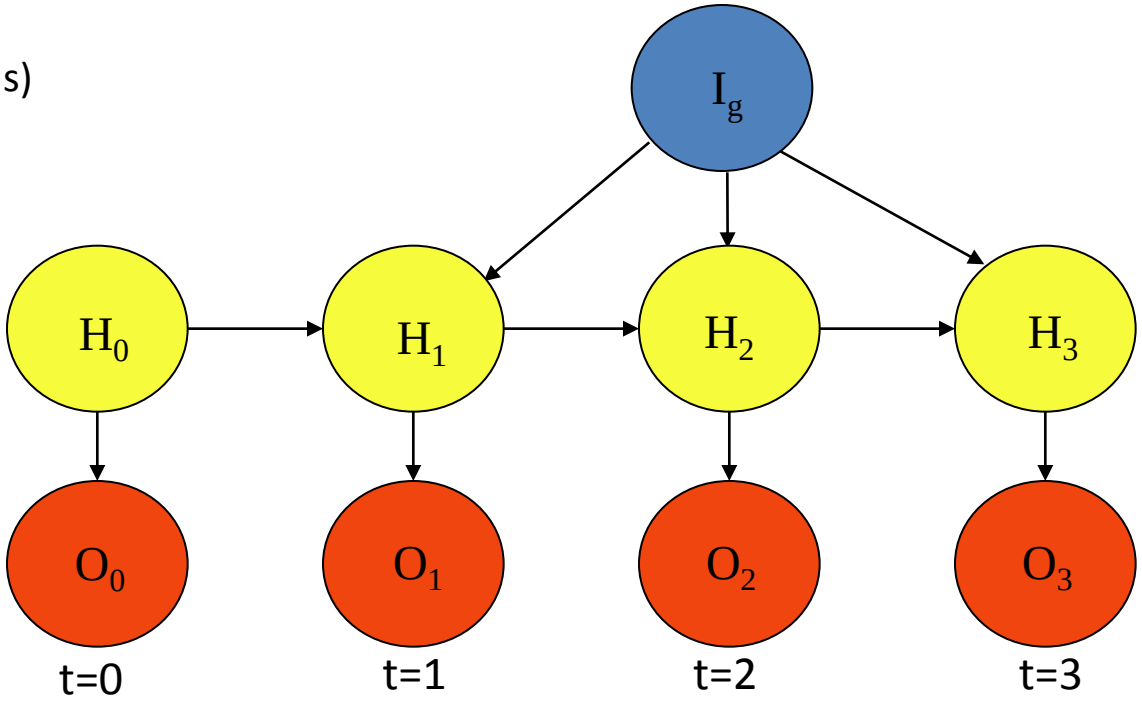
Schliep et al *Bioinformatics* 2003

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

$$r(G|M) = \sum_{g \in G} \log \sum_{q \in Q} \prod_{t=1}^{n-1} f_{q(t)}(o_g(t)) \prod_{t=1}^{n-1} P(H_t = q(t) | H_{t-1} = q(t-1), I_g)$$

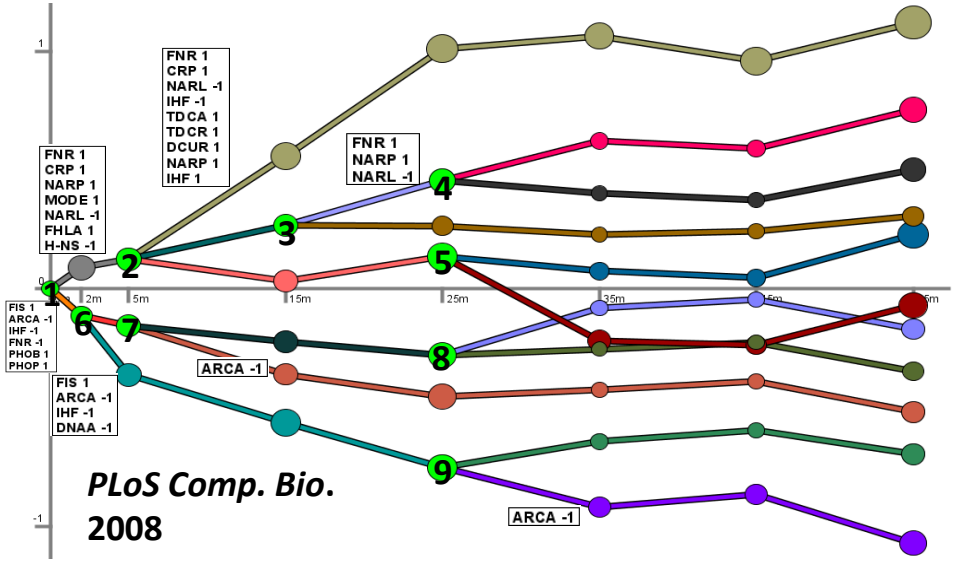
Sum over all genes

Sum over all paths *Q*

Product over all Gaussian emission density values on path

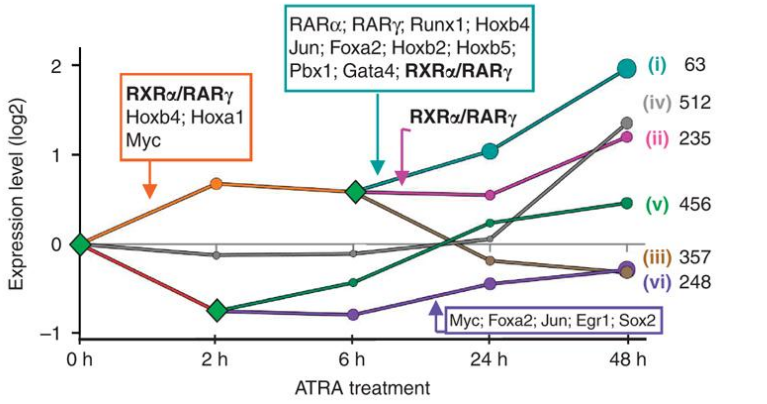
Product over all transition probabilities on path

E. coli. response



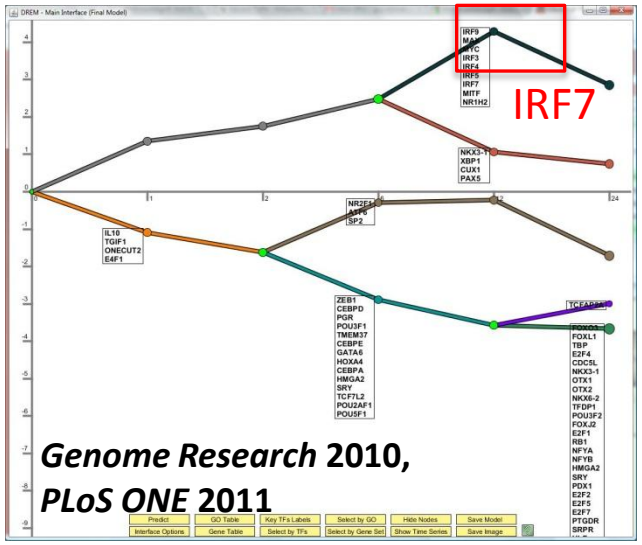
PLoS Comp. Bio. 2008

Stem cells differentiation



Nature MSB 2011

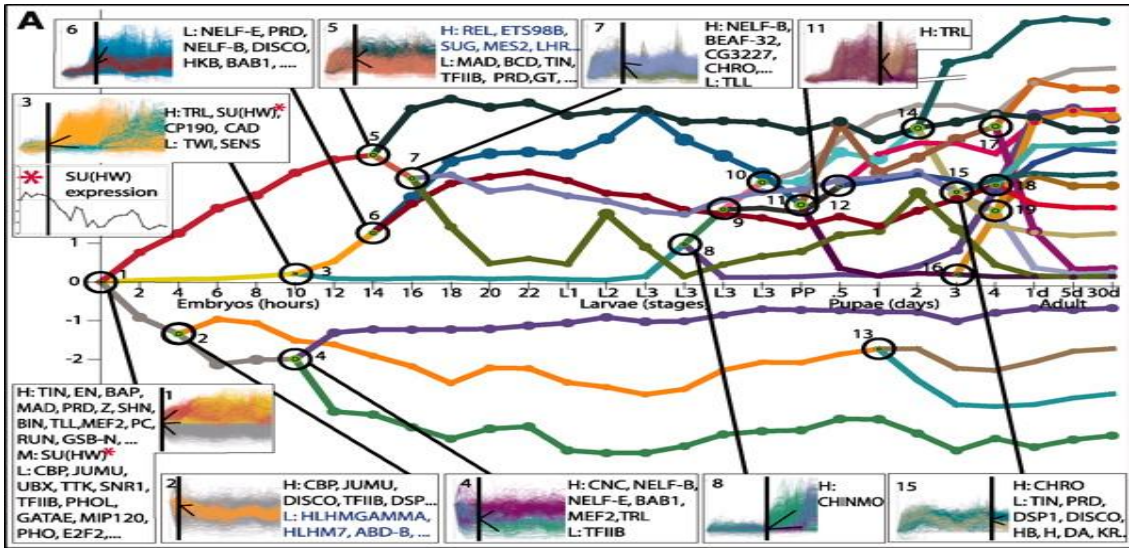
Mouse Immune response



Genome Research 2010,
PLoS ONE 2011

Fly development

Science 2010



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.