Crash course on Computational Biology for Computer Scientists

Bartek Wilczyński
bartek@mimuw.edu.pl
http://regulomics.mimuw.edu.pl

Phd Open lecture series

17-19 XI 2016

Topics for the course

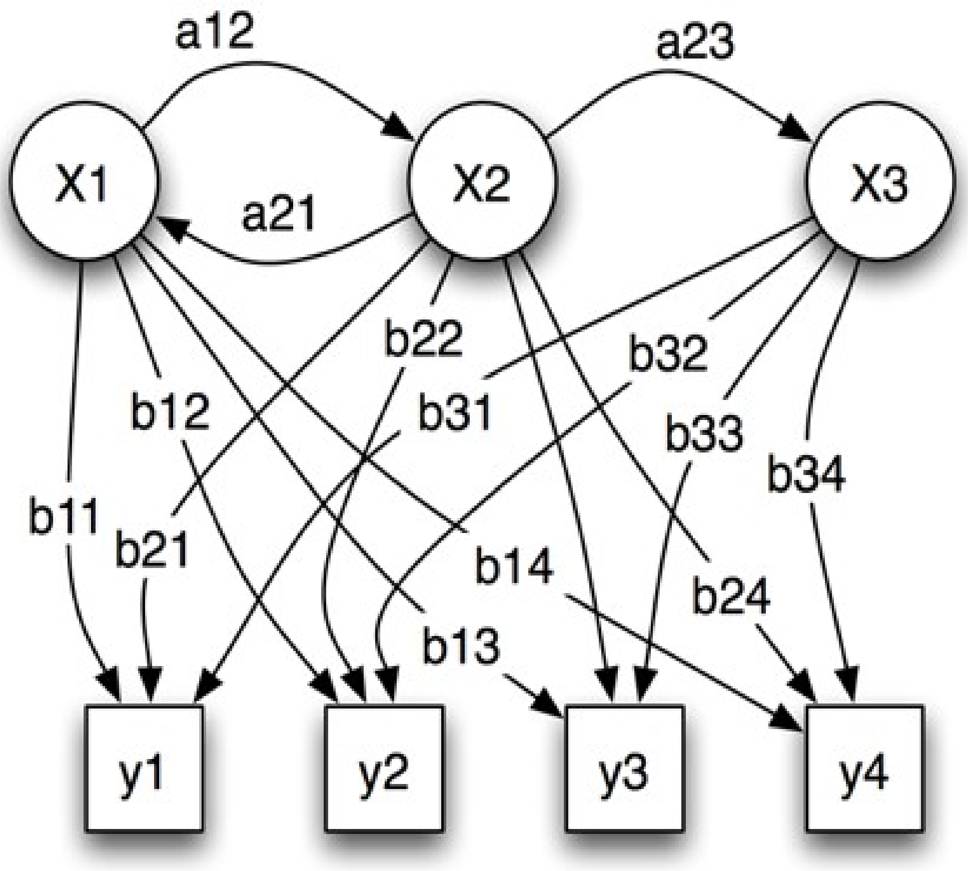
- Sequences in Biology – what do we study?
- Sequence comparison and searching – how to quickly find relatives in large sequence banks
- Tree-of-life and its construction(s)
- Short sequence mapping – where did this word come from
- DNA sequencing and assembly – puzzles for experts
- Sequence segmentation – finding modules by flipping coins
- Data storage and compression – from DNA to bits and back again
- Structures in Biology – small and smaller

Markov Models

- The model consists of a state space $Q \neq \emptyset$ (for our purposes Q is finite)
- and a transition probability matrix p_{ij} where $i, j \in Q$
- The model has no memory, the probability of moving from state i to j depends only on the state i.
- multiplying the matrix P, we can compute the change of the probability distribution as the model "steps" forward
- We are usually interested in stationary distributions π, such that $\pi \cdot P = P$

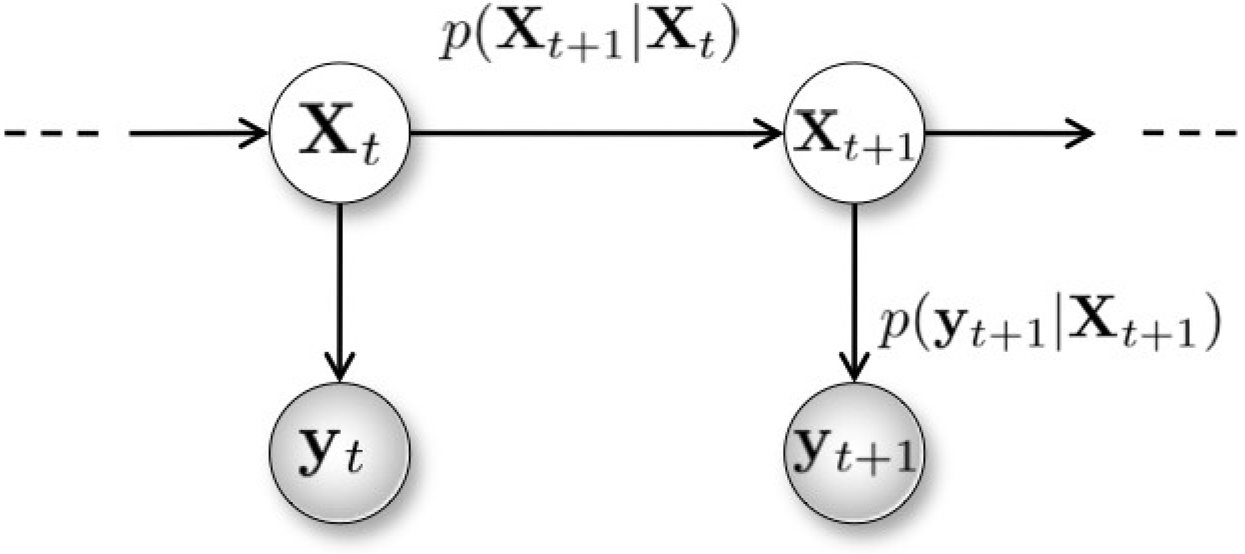
Hidden Markov Models

- Now the Markov Chain is not observable
- We only observe some emitted signals, probabilisticly depending on the chain state
- So in addition to the transition matrix, we have a emission matrix

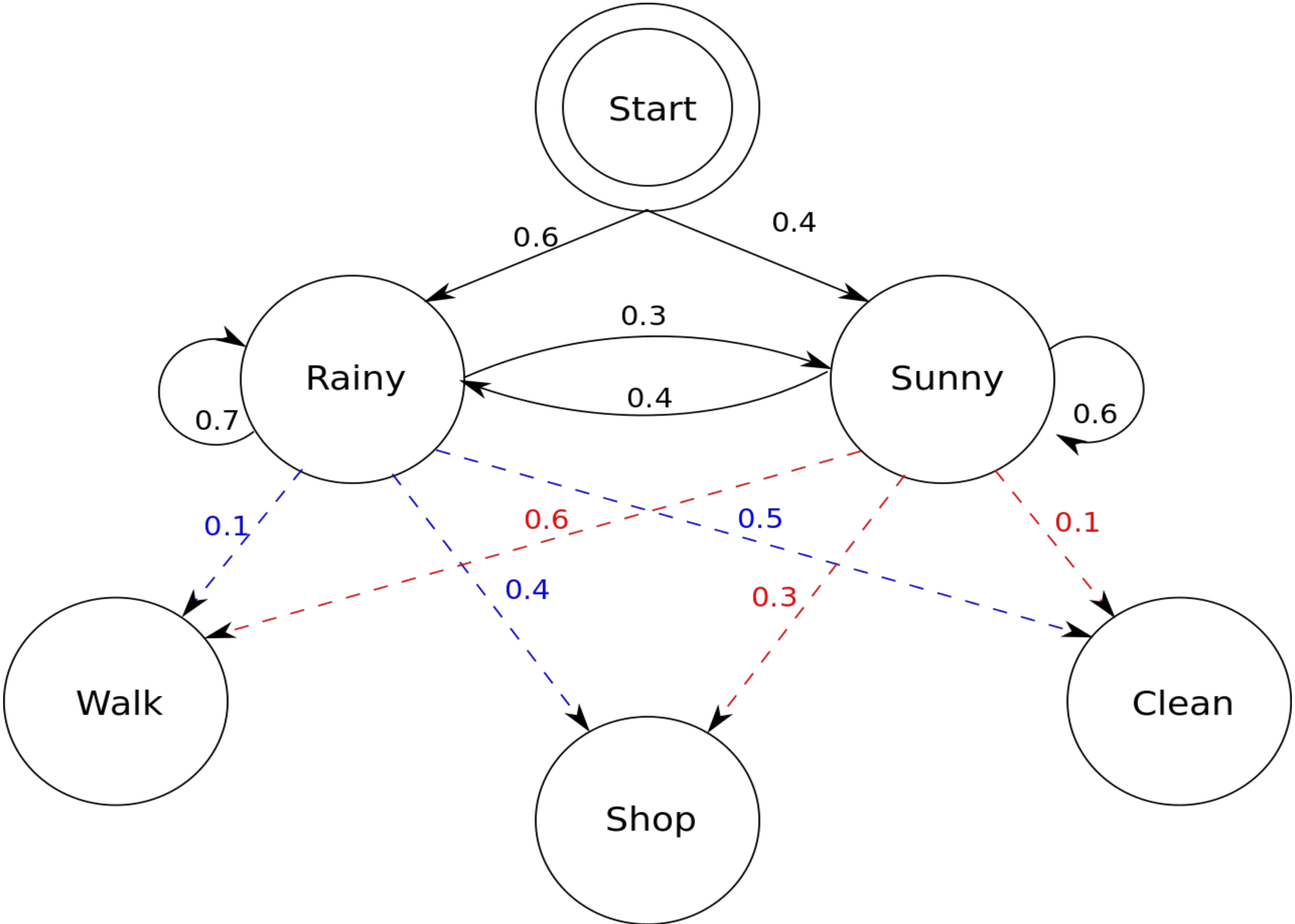


Trajectories of HMMs

- The Markov model changes states (Xs) over time using transition matrix
- At each state a random symbol is emitted based on the emission probabilities


$p(\mathbf{X}_{t+1}|\mathbf{X}_t)$
$\mathbf{X}_t$
$\mathbf{X}_{t+1}$
$p(\mathbf{y}_{t+1}|\mathbf{X}_{t+1})$
$\mathbf{y}_t$
$\mathbf{y}_{t+1}$

HMM example


Start
0.6
0.4
Rainy
0.3
0.4
Sunny
0.7
0.6
0.1
0.4
0.6
0.5
0.3
0.1
Walk
Shop
Clean

Reconstructing trajectory states

For any trajectory π, we can calculate the probability of emiting S

$$P(S, \pi) = \prod_{t=0}^{n-1} e_{\pi(t+1)}(S(t+1)) \cdot p_{\pi(t),\pi(t+1)},$$

Can we find the optimal trajectory π, given S?

$$P(S, \pi_*) = max\{P(S, \pi) \mid \pi \in Q^*, |\pi| = |S|\}.$$

Viterbi algorithm

We can use dynamic programming, filling in the $v(i,k)$ matrix

$$v(i,k) = max\{P(S[1..i],\pi) \mid \pi \in Q^i, \quad \pi(i) = k\}.$$

with the initial condition:

$$v(0,k) = \begin{cases} 1 & \text{gdy } k = k_0, \\ 0 & \text{gdy } k \neq k_0. \end{cases}$$

and step function:

$$v(i,k) = e_k(S(i)) \cdot \max_{l \in Q}[v(i-1,l) \cdot p_{l,k}].$$

To finally read out the seeked probability:

$$P(S,\pi_*) = \max_{k \in Q}[v(|S|,k)].$$

The forward and backward probabilities of trajectories

Now, we can calculate the probability of emitting S, over all possible trajectories, with the Forward-method. The initial step is as follows:

$$f(0,k) = \begin{cases} 1 & \text{gdy } k = k_0, \\ 0 & \text{gdy } k \neq k_0. \end{cases}$$

Then, we make similar steps:

$$f(i,k) = e_k(S(i)) \cdot \sum_{l \in Q} f(i-1,l) \cdot p_{l,k}.$$

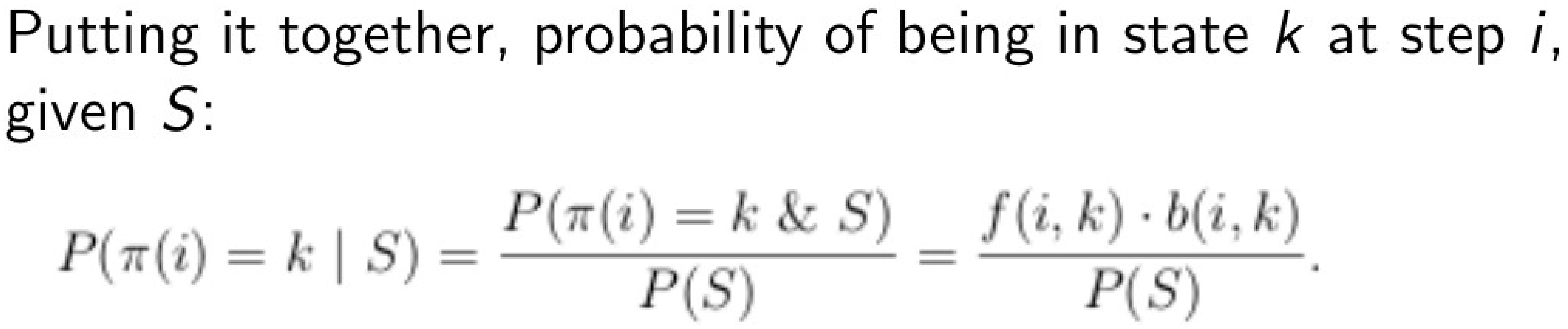
and finally we can calculate the total probability at the end:

$$P(S) = \sum_{k \in Q} f(|S|,k).$$

The same works backwards:

$$b(i,k) = \sum_{l \in Q} p_{k,l} \cdot e_l(S(i+1)) \cdot b(i+1,l).$$

Where were we at time t?



Putting it together, probability of being in state k at step i, given S:

$$P(\pi(i) = k \mid S) = \frac{P(\pi(i) = k \;\&\; S)}{P(S)} = \frac{f(i,k) \cdot b(i,k)}{P(S)}.$$

Given the sequence of emitted symbols, we can estiimate the likely states of the hidden system

The emission matrix can be then estimated

Estimate of the Emission matrix:

$$e_k(x) = \frac{E_k(x)}{\sum_{y \in \Sigma} E_k(y)}.$$

Can be calculated using f and b

$$E_k(x) = \sum_{j=1}^{n} \sum_{i \in I_j(x)} \frac{f_{\mathcal{M}}^{(j)}(i,k) \cdot b_{\mathcal{M}}^{(j)}(i,k)}{P_{\mathcal{M}}(S_j)},$$

As well as the transition matrix

Similarly the transition matrix:

$$p_{k,l} = \frac{P_{k,l}}{\sum_{q \in Q} P_{k,q}},$$

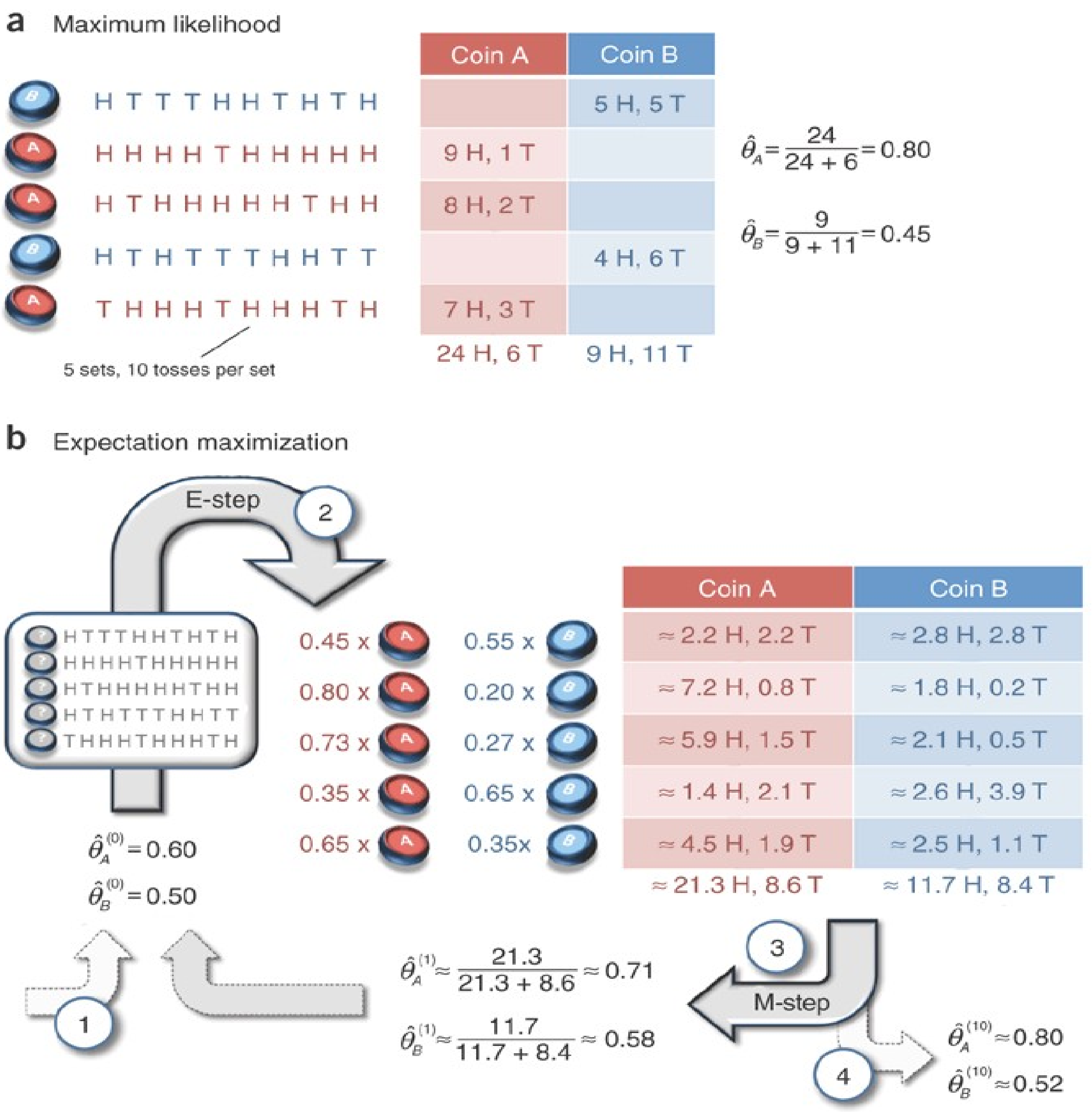
depends on f and b

$$P_{k,l} = \sum_{j=1}^{n} \sum_{i=1}^{|S_j|} \frac{f_{\mathcal{M}}^{(j)}(i,k) \cdot p_{k,l}^{\mathcal{M}} \cdot e_l^{\mathcal{M}}(S_j(i+1)) \cdot b_{\mathcal{M}}^{(j)}(i+1,l)}{P_{\mathcal{M}}(S_j)}.$$

Baum Welch algorithm

- Suppose, we only know the word S and the sets Q and Σ. Can we estimate both p_{ij} and e_{ij}?
- We can start with random p_{ij}, e_{ij} and iteratively proceed as follows:
  - Calculate the estimates of being in each of states at each step using f, b and current estimates of e, p.
  - Find the optimal e, p, given current e, p, f, b
- This is an example of a known procedure called expectation-maximization
- It is proven to converge to a local optimum.

Expectation-Maximization



a Maximum likelihood

		Coin A	Coin B
B	H T T T H H T H T H		5 H, 5 T
A	H H H H T H H H H H	9 H, 1 T	
A	H T H H H H H T H H	8 H, 2 T	
B	H T H T T T H H T T		4 H, 6 T
A	T H H H T H H H T H	7 H, 3 T	
		24 H, 6 T	9 H, 11 T

5 sets, 10 tosses per set

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

b Expectation maximization

		Coin A	Coin B
0.45 x A	0.55 x B	≈ 2.2 H, 2.2 T	≈ 2.8 H, 2.8 T
0.80 x A	0.20 x B	≈ 7.2 H, 0.8 T	≈ 1.8 H, 0.2 T
0.73 x A	0.27 x B	≈ 5.9 H, 1.5 T	≈ 2.1 H, 0.5 T
0.35 x A	0.65 x B	≈ 1.4 H, 2.1 T	≈ 2.6 H, 3.9 T
0.65 x A	0.35x B	≈ 4.5 H, 1.9 T	≈ 2.5 H, 1.1 T
		≈ 21.3 H, 8.6 T	≈ 11.7 H, 8.4 T

$$\hat{\theta}_A^{(0)} = 0.60$$

$$\hat{\theta}_B^{(0)} = 0.50$$

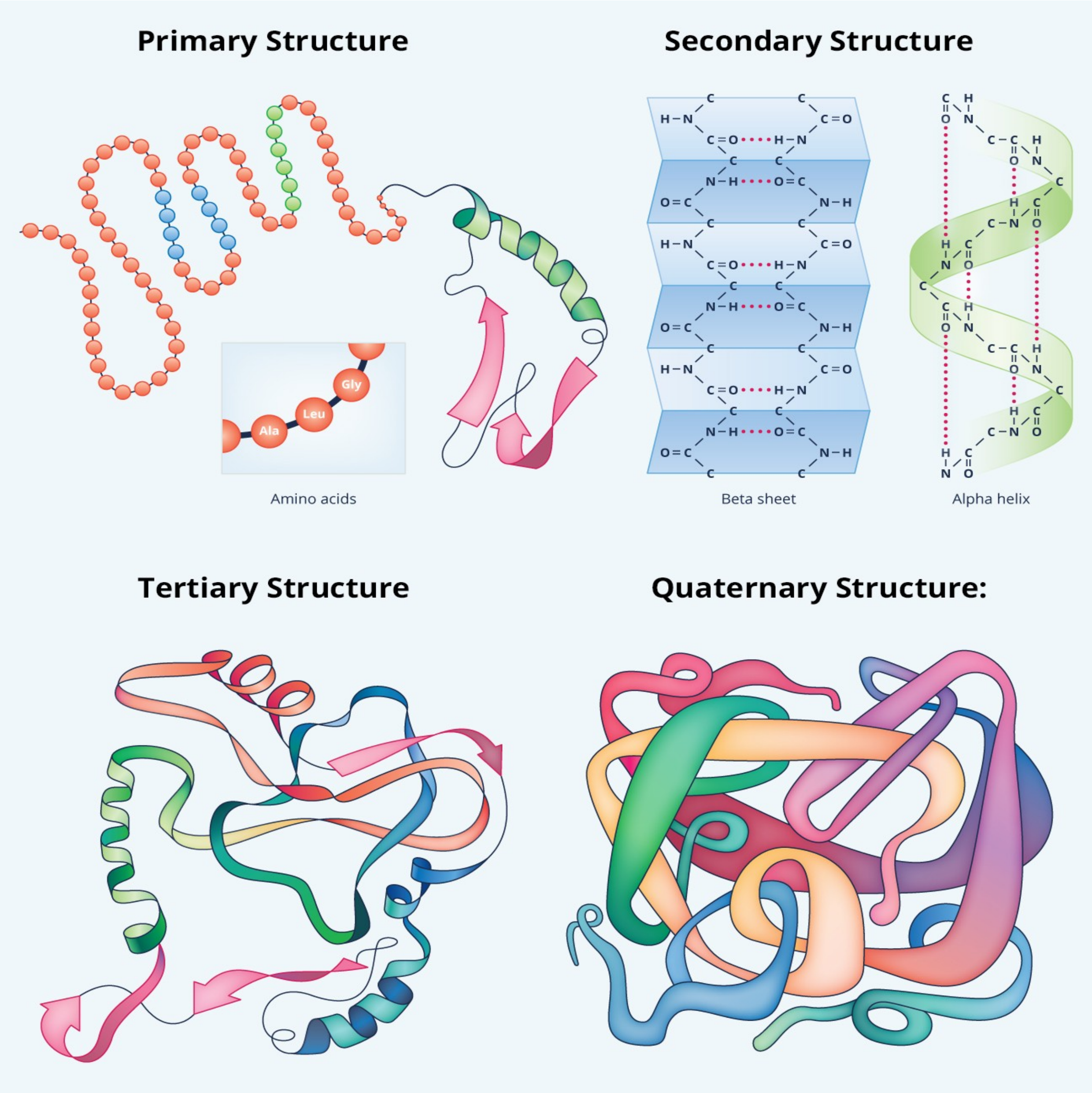
$$\hat{\theta}_A^{(1)} \approx \frac{21.3}{21.3+8.6} \approx 0.71$$

$$\hat{\theta}_B^{(1)} \approx \frac{11.7}{11.7+8.4} \approx 0.58$$

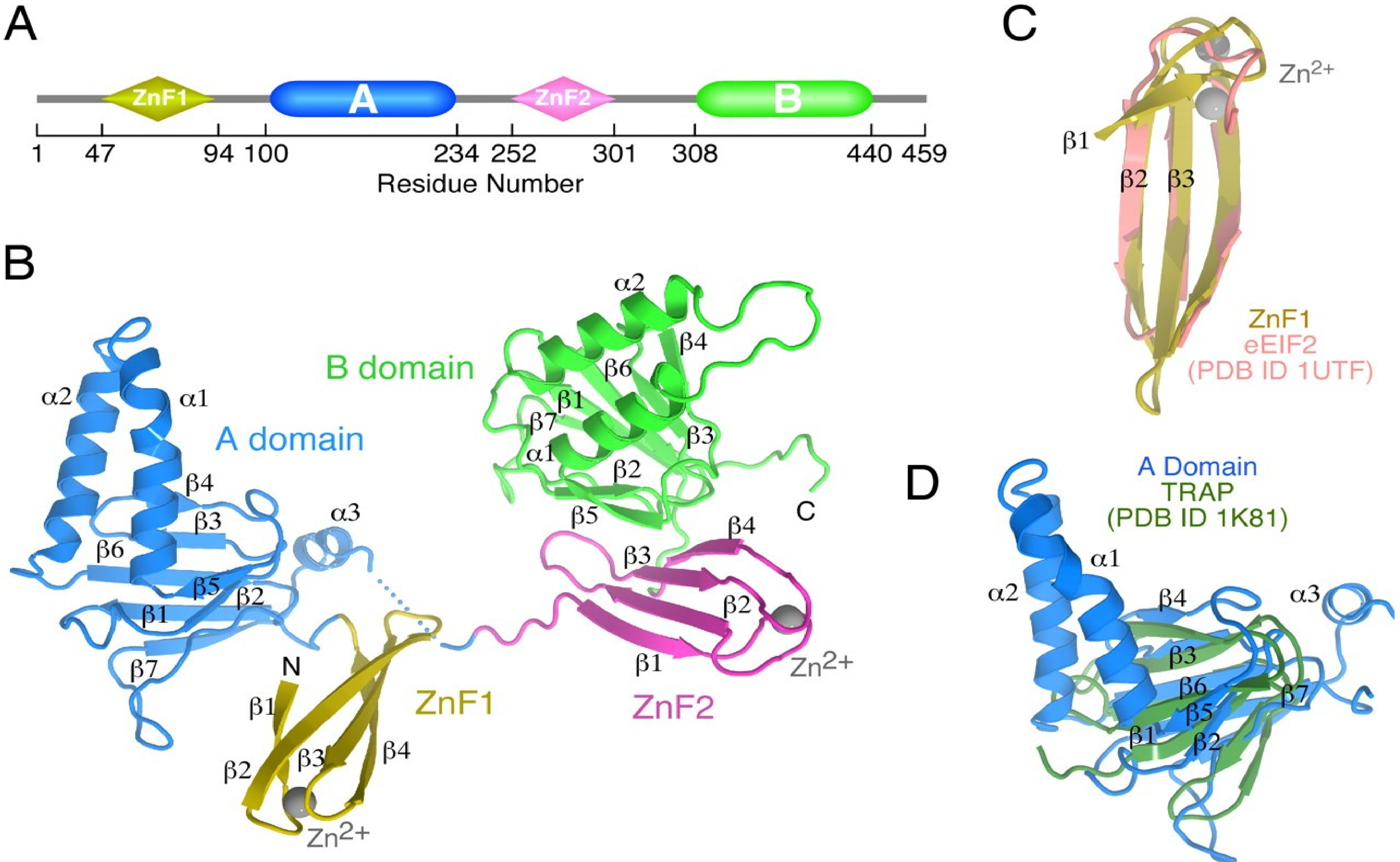
$$\hat{\theta}_A^{(10)} \approx 0.80$$

$$\hat{\theta}_B^{(10)} \approx 0.52$$

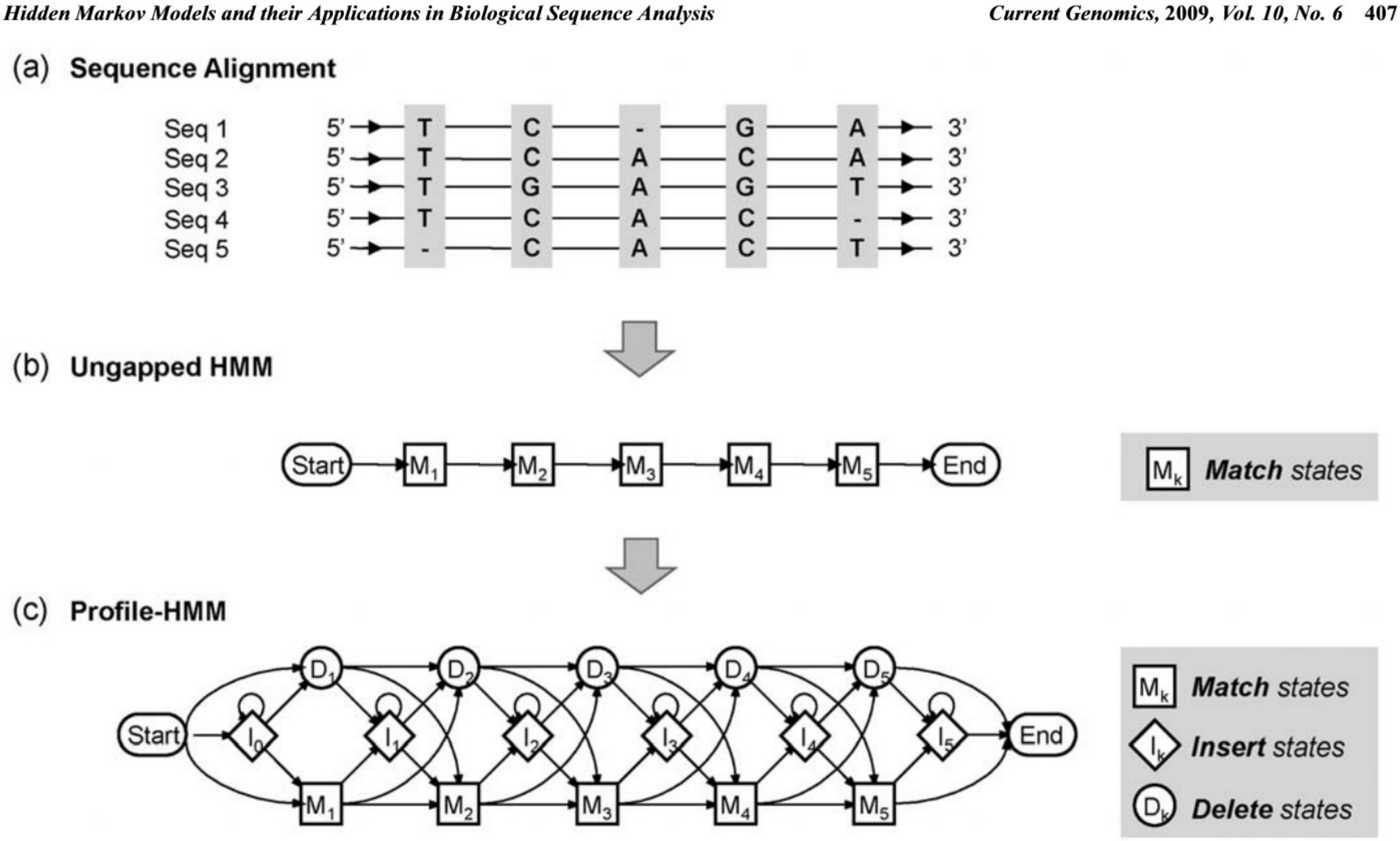
Protein structure



Protein domains



Profile HMMs



Hidden Markov Models and their Applications in Biological Sequence Analysis *Current Genomics*, 2009, *Vol. 10, No. 6* 407

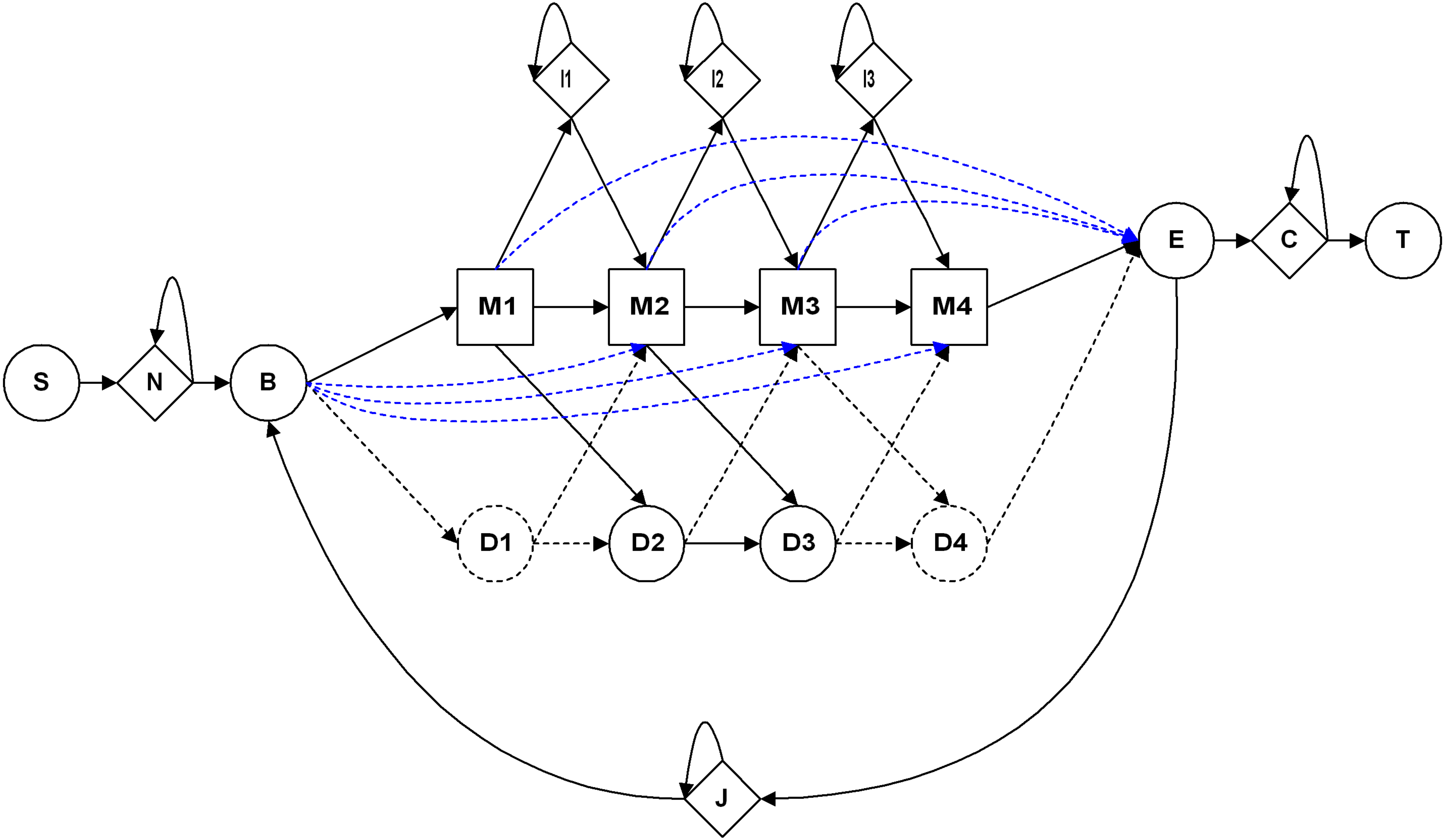
(a) **Sequence Alignment**

Seq 1	5'	T	C	-	G	A	3'
Seq 2	5'	T	C	A	C	A	3'
Seq 3	5'	T	G	A	G	T	3'
Seq 4	5'	T	C	A	C	-	3'
Seq 5	5'	-	C	A	C	T	3'

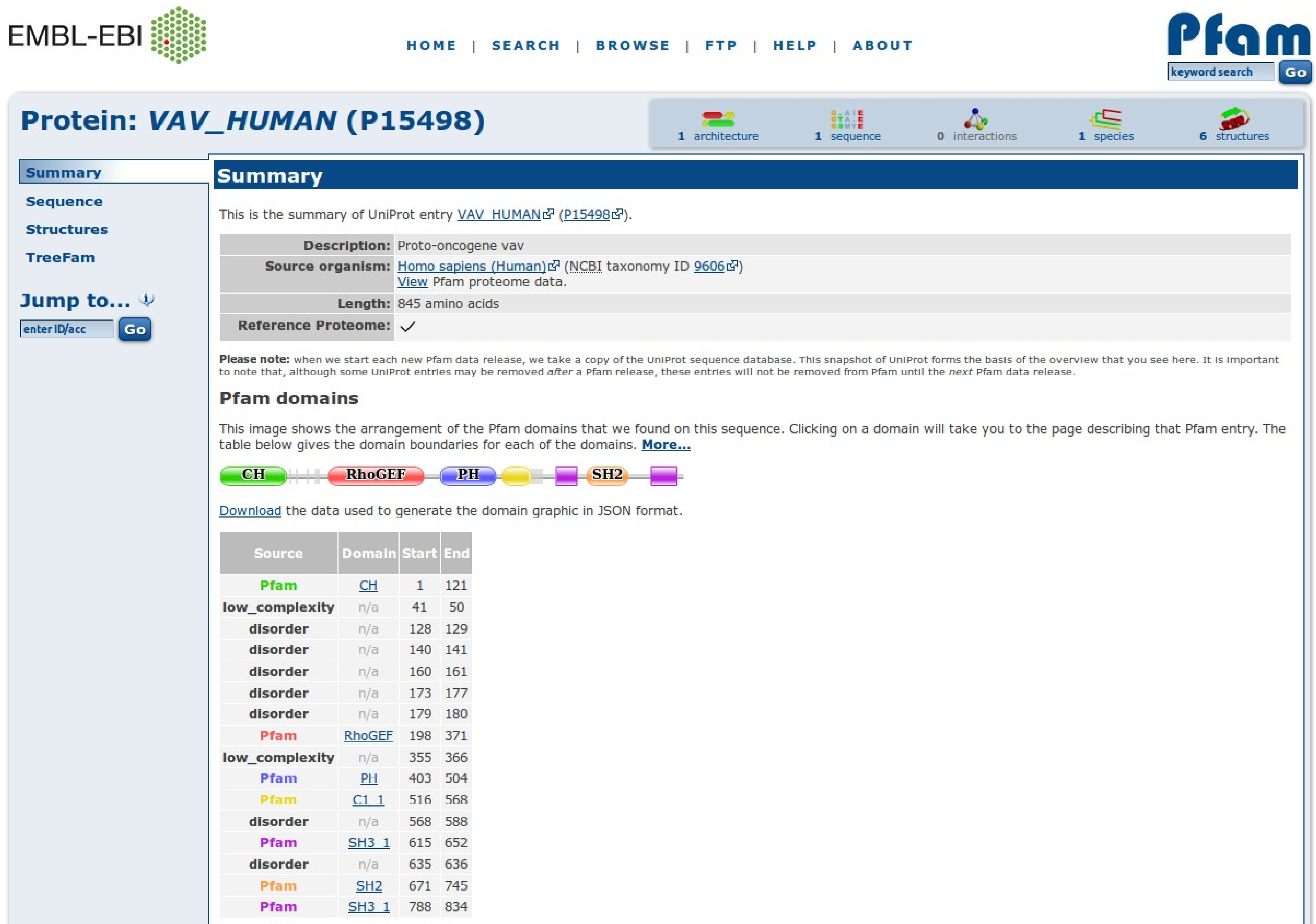
(b) **Ungapped HMM**

(c) **Profile-HMM**

Finding a domain in a longer protein sequence



PFAM sequence annotation



Protein: *VAV_HUMAN* (P15498)

Summary

This is the summary of UniProt entry VAV_HUMAN (P15498).

Description:	Proto-oncogene vav
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606) View Pfam proteome data.
Length:	845 amino acids
Reference Proteome:	✓

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

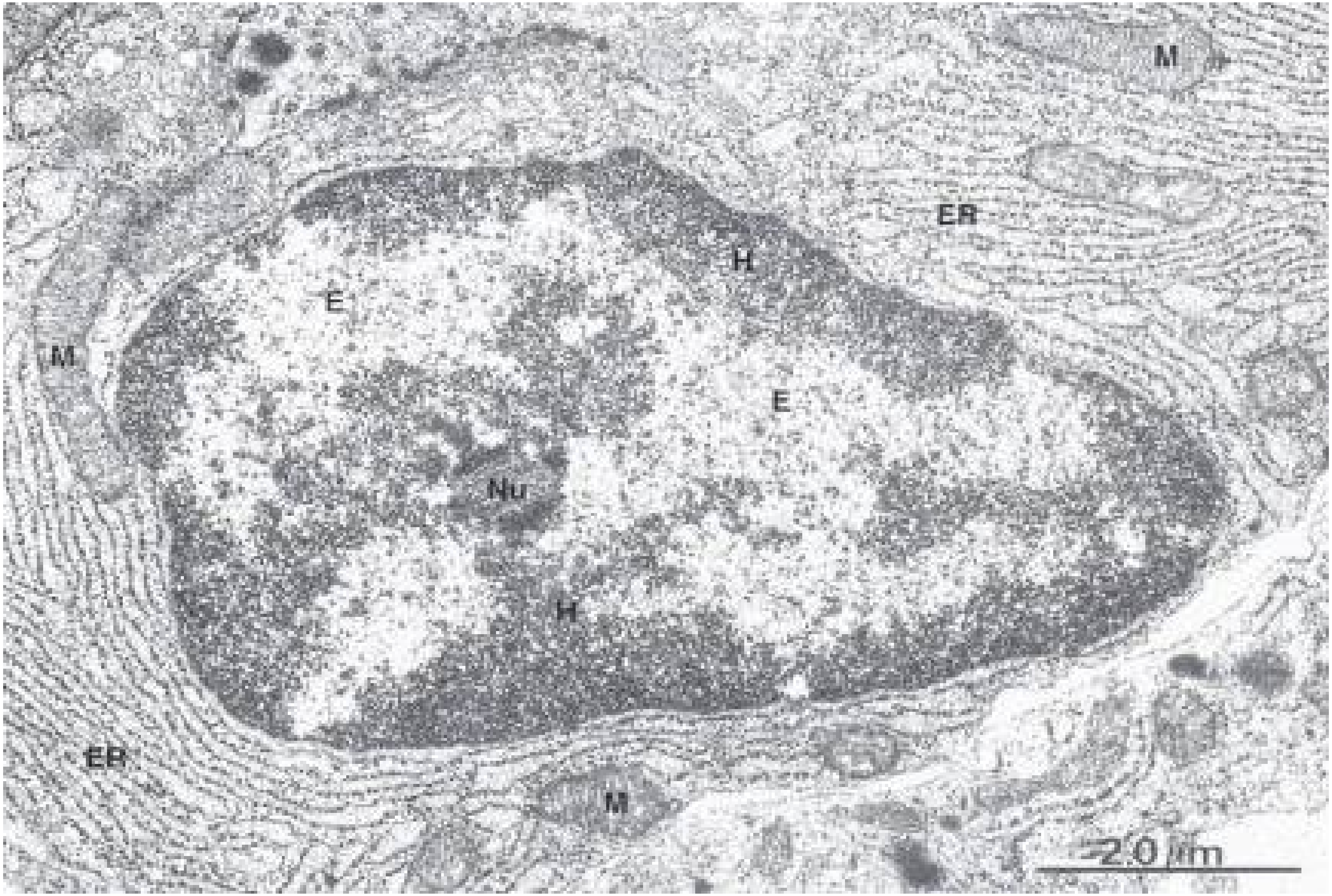
Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains. **More...**

Download the data used to generate the domain graphic in JSON format.

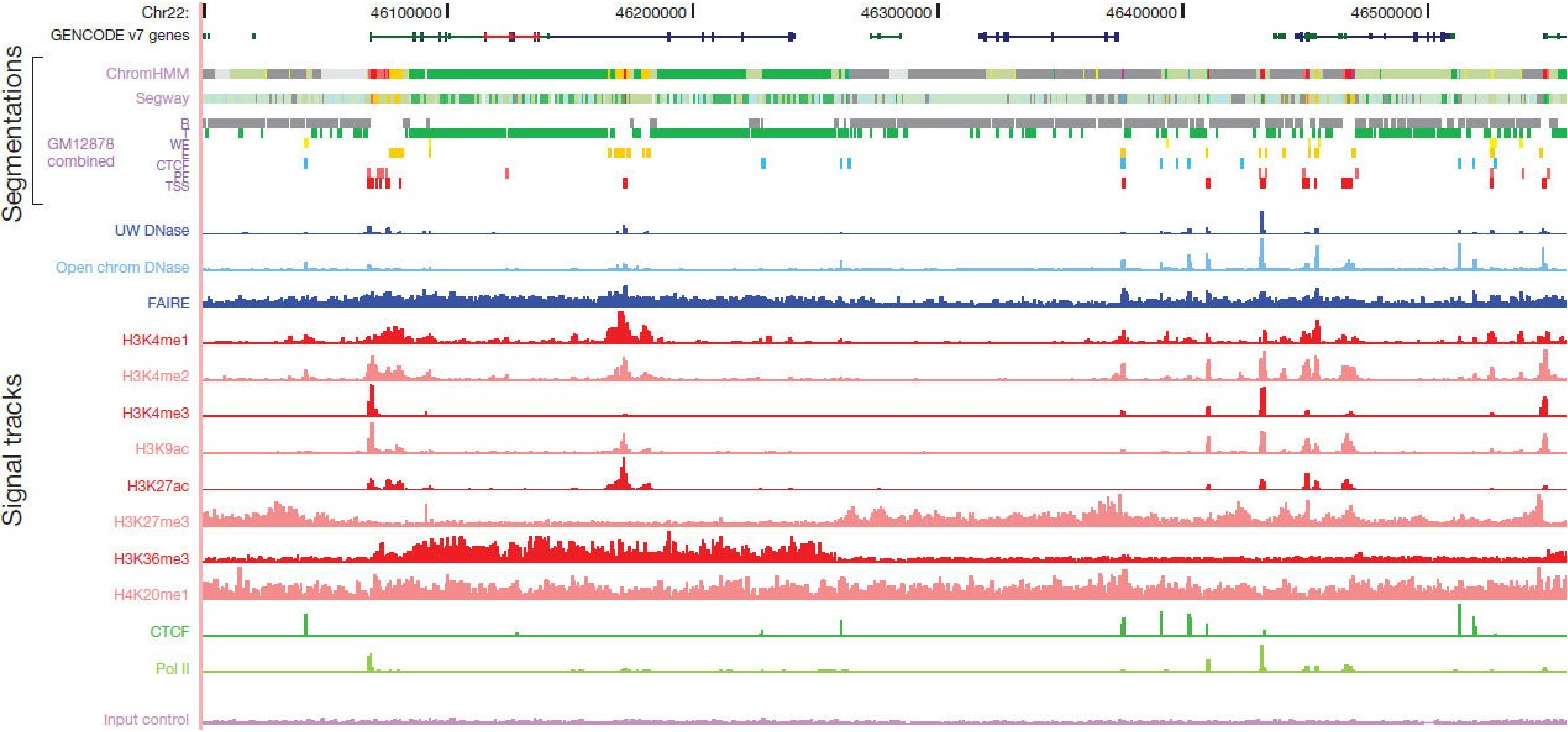
Source	Domain	Start	End
Pfam	CH	1	121
low_complexity	n/a	41	50
disorder	n/a	128	129
disorder	n/a	140	141
disorder	n/a	160	161
disorder	n/a	173	177
disorder	n/a	179	180
Pfam	RhoGEF	198	371
low_complexity	n/a	355	366
Pfam	PH	403	504
Pfam	C1_1	516	568
disorder	n/a	568	588
Pfam	SH3_1	615	652
disorder	n/a	635	636
Pfam	SH2	671	745
Pfam	SH3_1	788	834

What is the chromatin state?

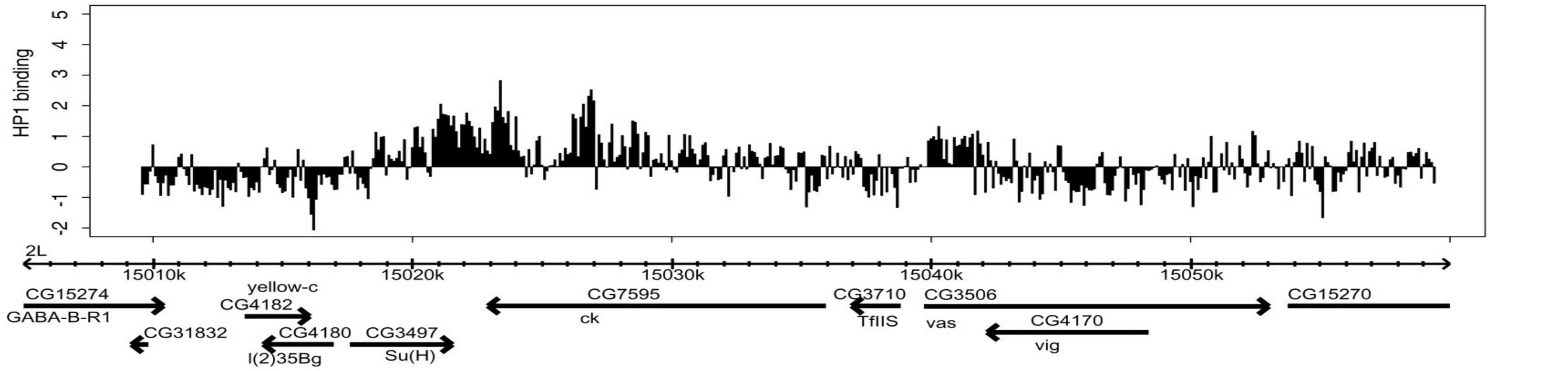


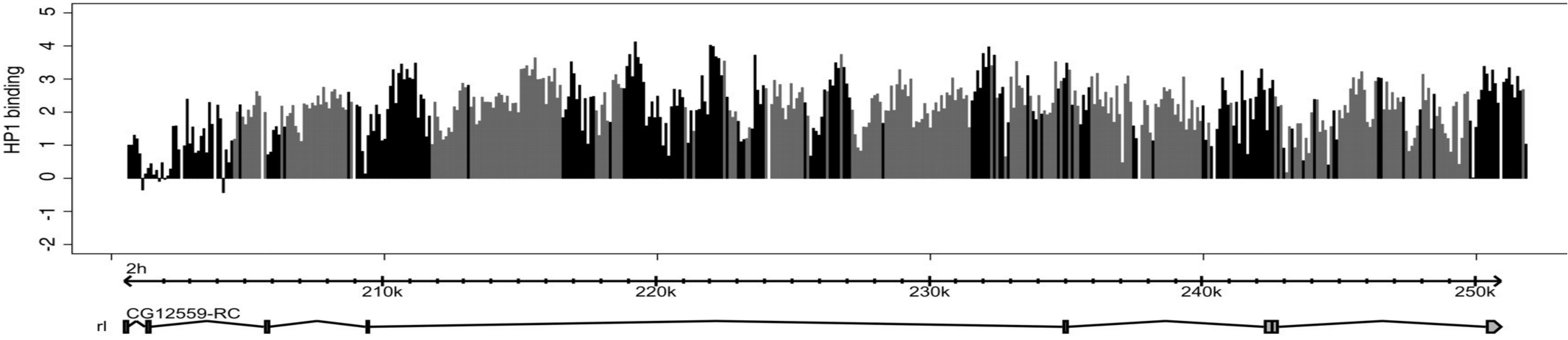
UCSF School of medicine

ChIP data from ENCODE project



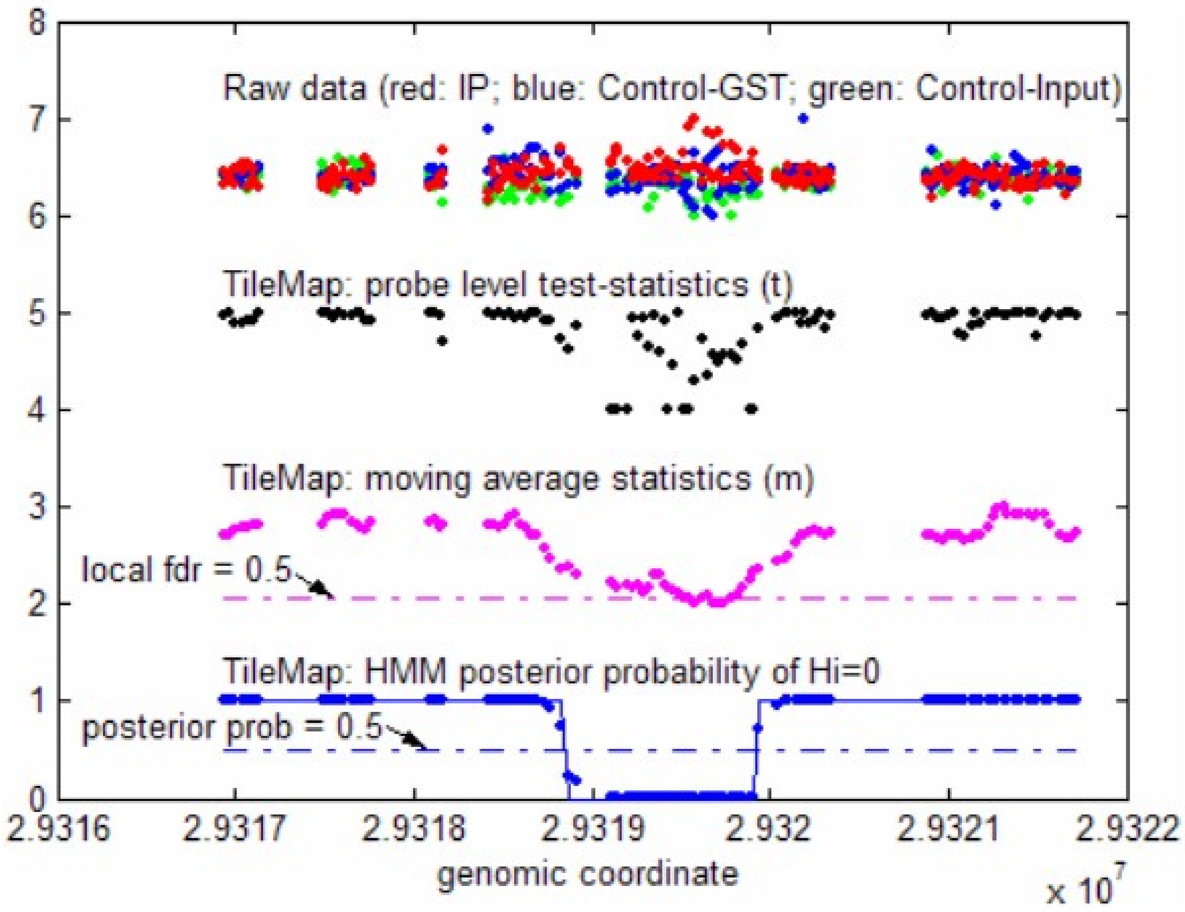
Chromatin Immunoprecipitation data



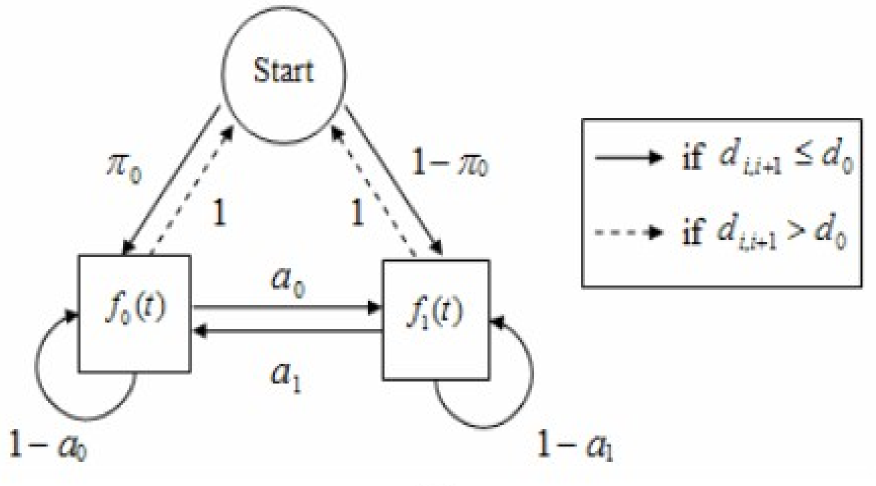


- Considereble noise level

HMM model



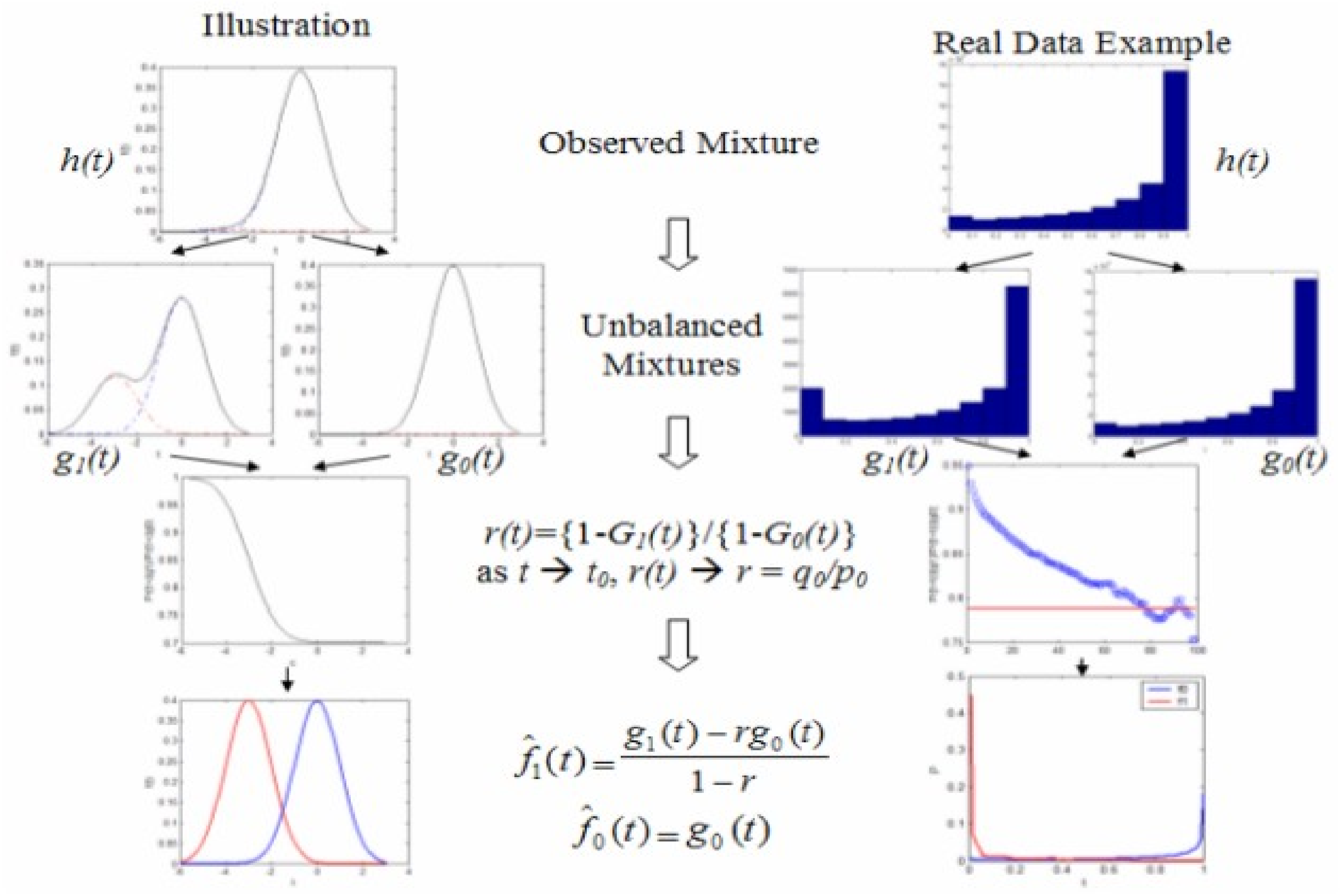
(a)



(b)

- TileMap method (Ji&Wong 2005, Bioinfiormatics)
- Hidden Markov model for segmentation of ChIP data with 2 states:
  - 0 – no enrichment
  - 1 - enrichment
- Emissions are Gaussian

Emission model in TileMap



Observed Mixture

⇩

Unbalanced Mixtures

⇩

$r(t)=\{1\text{-}G_1(t)\}/\{1\text{-}G_0(t)\}$

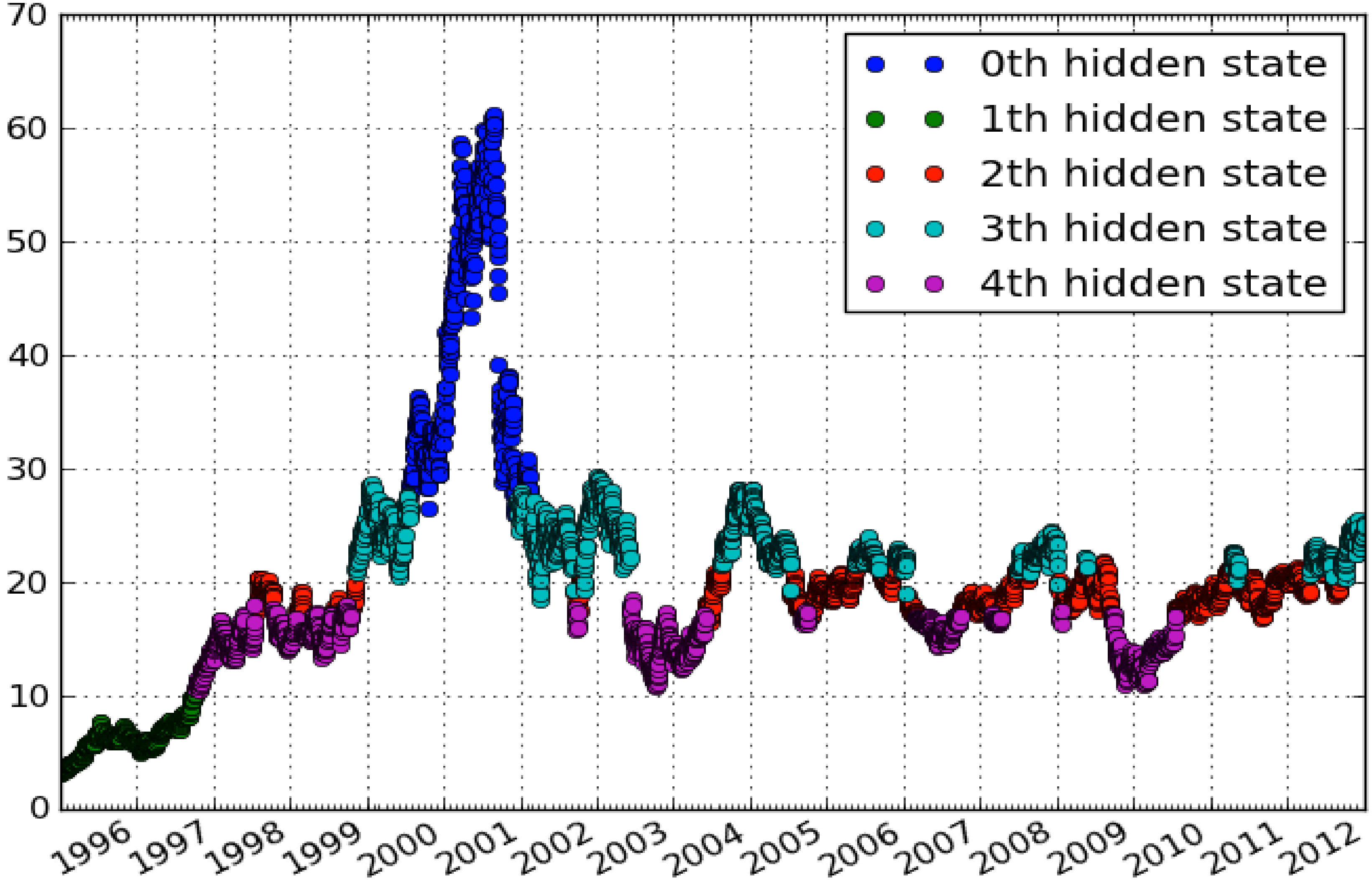
as $t \to t_0$, $r(t) \to r = q_0/p_0$

⇩

$$\hat{f}_1(t)=\frac{g_1(t)-rg_0(t)}{1-r}$$

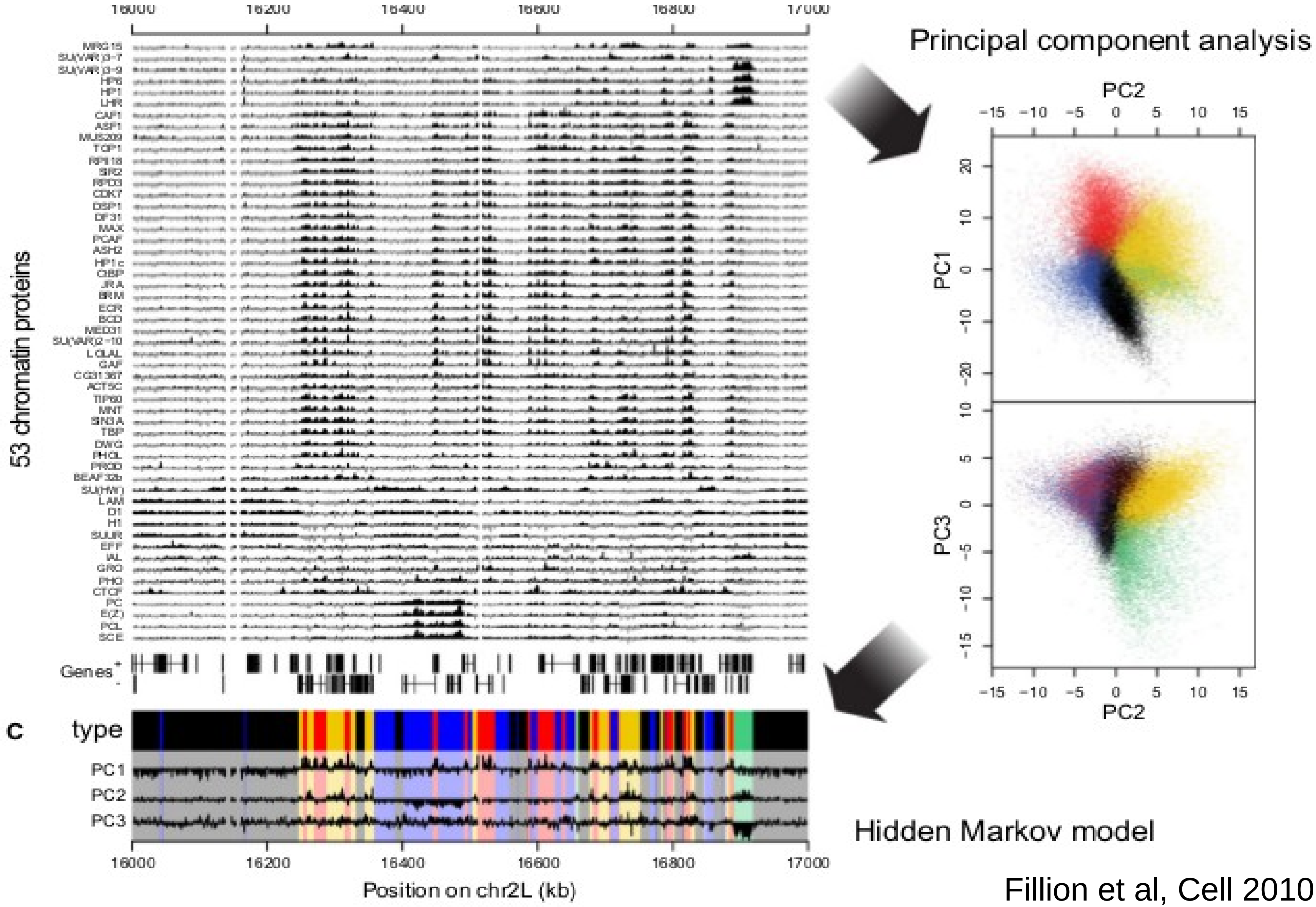
$$\hat{f}_0(t)=g_0(t)$$

Using Gaussian HMM for Stock Market



From scikit.learn documentation

You can use HMMs for chromatin



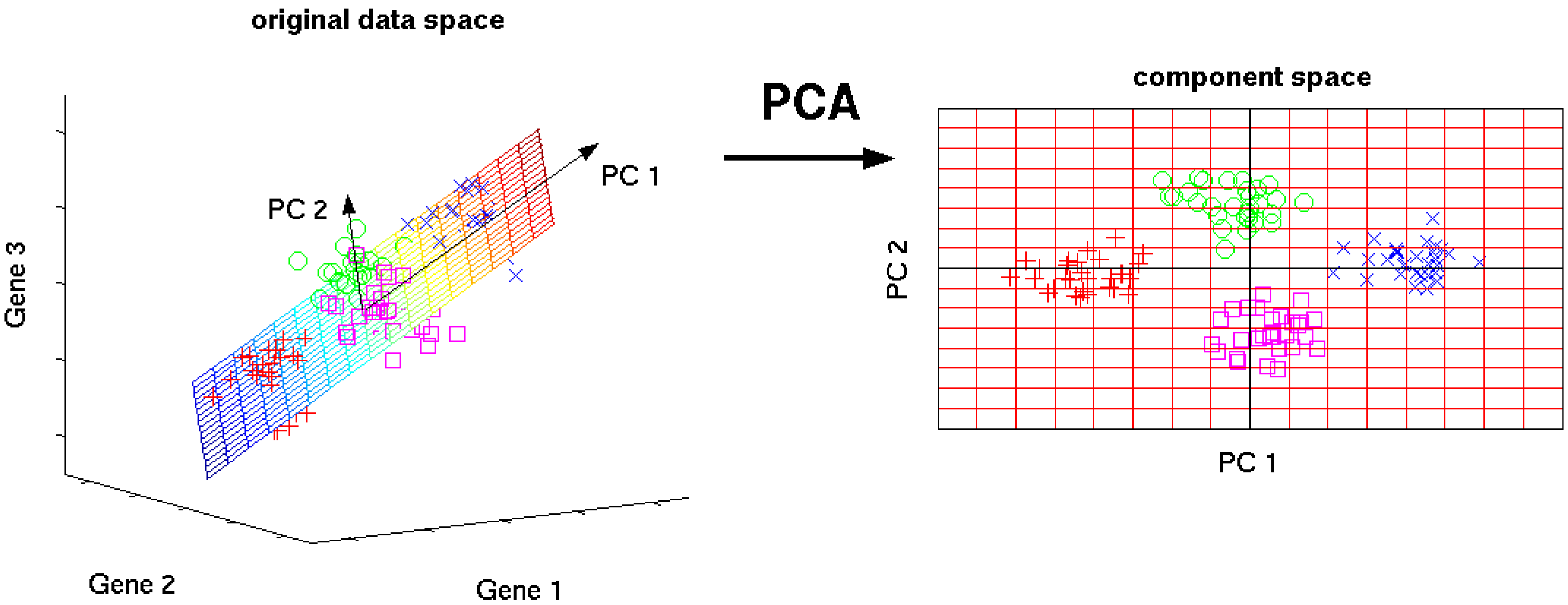
Principal component analysis

Hidden Markov model

Fillion et al, Cell 2010

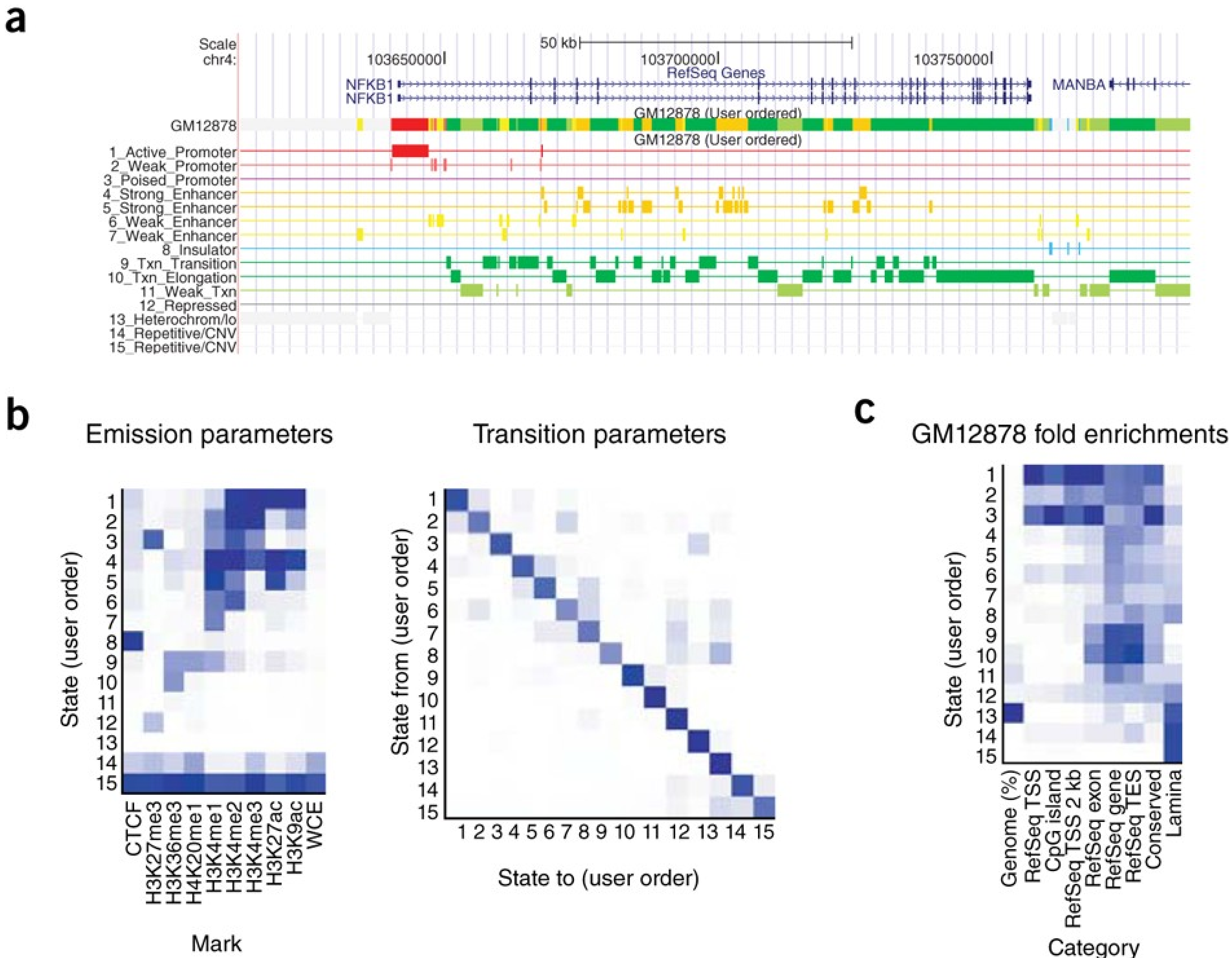
Using PCA to limit the emission space dimension

- Principal component analysis is a method of identifying orthogonal vectors with maximal variance in the multidimensional data



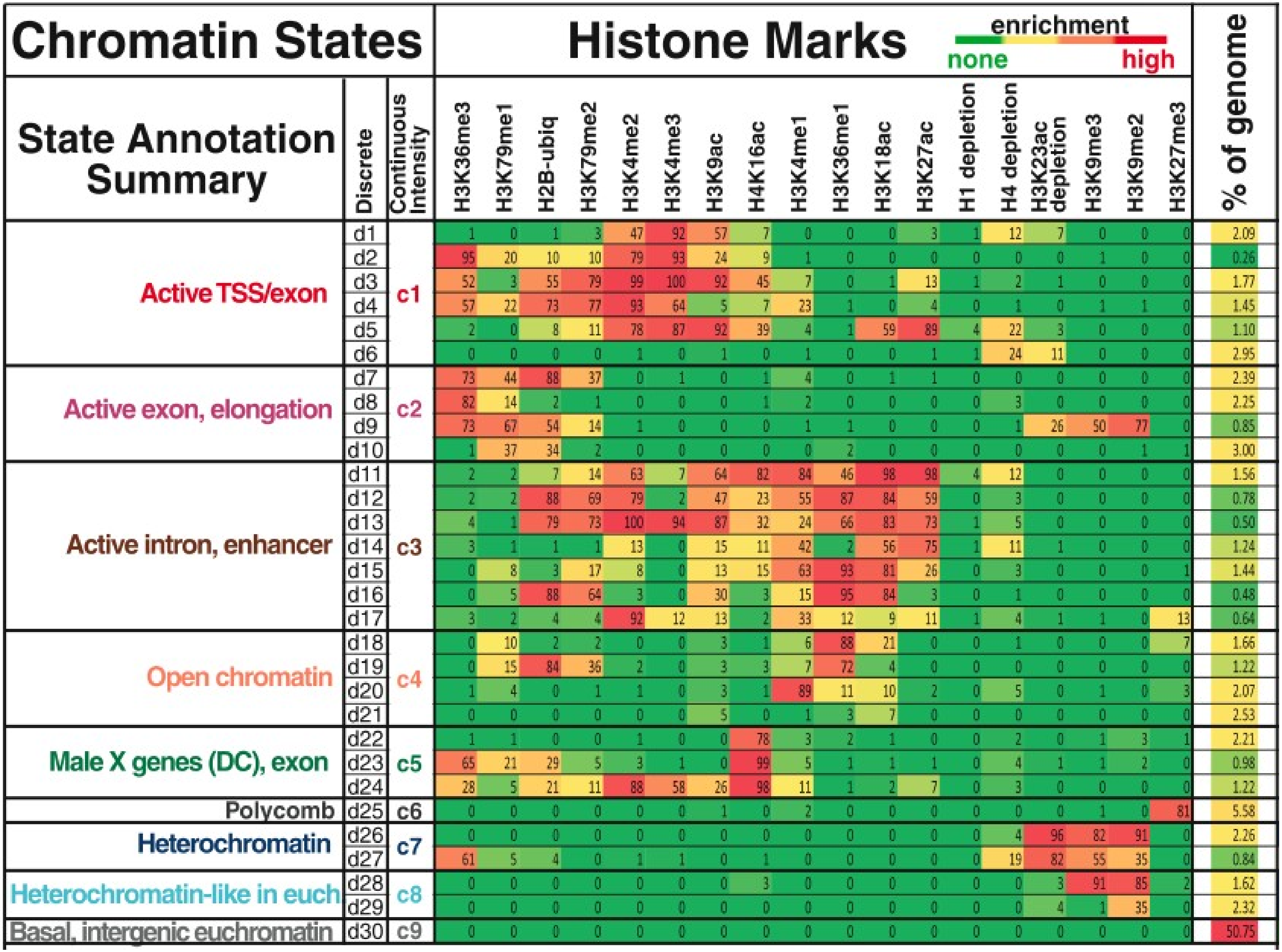
Independent multidimensional emissions

- ChromHMM is taking a different approach
- One can assume that all of the different ChIP measurements are independent of each other
- Then instead of exponential emission explosion, we have a matrix of emission probabilities for each state
- For each observable ChIP we need the probabilities vector for each hidden state
- This is even extendable to Gaussian emissions



Ernst&Kellis, 2012, Nat Biotech

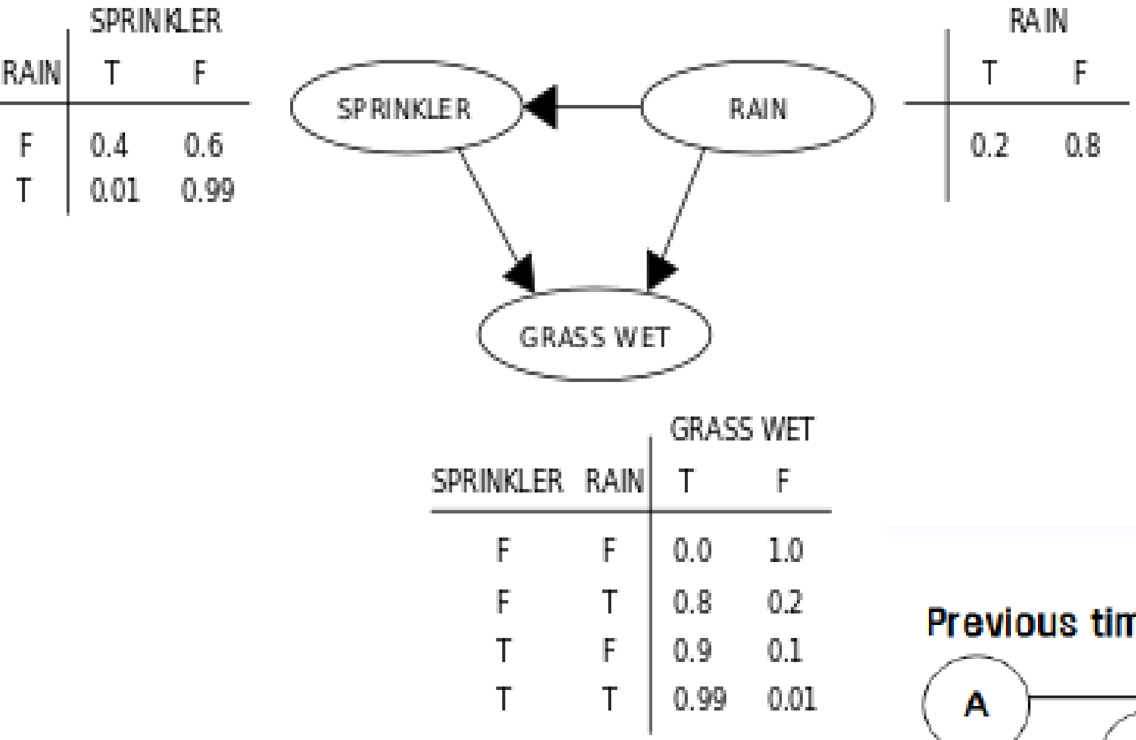
Emission matrix for Drosophila



Chromatin States			Histone Marks												enrichment none			high			
State Annotation Summary	Discrete	Continuous Intensity	H3K36me3	H3K79me1	H2B-ubiq	H3K79me2	H3K4me2	H3K4me3	H3K9ac	H4K16ac	H3K4me1	H3K36me1	H3K18ac	H3K27ac	H1 depletion	H4 depletion	H3K23ac depletion	H3K9me3	H3K9me2	H3K27me3	% of genome
Active TSS/exon	d1	c1	1	0	1	3	47	92	57	7	0	0	0	3	1	12	7	0	0	0	2.09
	d2		95	20	10	10	79	93	24	9	1	0	0	0	0	0	0	1	0	0	0.26
	d3		52	3	55	79	99	100	92	45	7	0	1	13	1	2	1	0	0	0	1.77
	d4		57	22	73	77	93	64	5	7	23	1	0	4	0	1	0	1	1	0	1.45
	d5		2	0	8	11	78	87	92	39	4	1	59	89	4	22	3	0	0	0	1.10
	d6		0	0	0	0	1	0	1	0	1	0	0	1	1	24	11	0	0	0	2.95
Active exon, elongation	d7	c2	73	44	88	37	0	1	0	1	4	0	1	1	0	0	0	0	0	0	2.39
	d8		82	14	2	1	0	0	0	1	2	0	0	0	0	3	0	0	0	0	2.25
	d9		73	67	54	14	1	0	0	0	1	1	0	0	0	1	26	50	77	0	0.85
	d10		1	37	34	2	0	0	0	0	0	2	0	0	0	0	0	0	1	1	3.00
Active intron, enhancer	d11	c3	2	2	7	14	63	7	64	82	84	46	98	98	4	12	0	0	0	0	1.56
	d12		2	2	88	69	79	2	47	23	55	87	84	59	0	3	0	0	0	0	0.78
	d13		4	1	79	73	100	94	87	32	24	66	83	73	1	5	0	0	0	0	0.50
	d14		3	1	1	1	13	0	15	11	42	2	56	75	1	11	1	0	0	0	1.24
	d15		0	8	3	17	8	0	13	15	63	93	81	26	0	3	0	0	0	1	1.44
	d16		0	5	88	64	3	0	30	3	15	95	84	3	0	1	0	0	0	0	0.48
	d17		3	2	4	4	92	12	13	2	33	12	9	11	1	4	1	1	0	13	0.64
Open chromatin	d18	c4	0	10	2	2	0	0	3	1	6	88	21	0	0	1	0	0	0	7	1.66
	d19		0	15	84	36	2	0	3	3	7	72	4	0	0	0	0	0	0	0	1.22
	d20		1	4	0	1	1	0	3	1	89	11	10	2	0	5	0	1	0	3	2.07
	d21		0	0	0	0	0	0	5	0	1	3	7	0	0	0	0	0	0	0	2.53
Male X genes (DC), exon	d22	c5	1	1	0	0	1	0	0	78	3	2	1	0	0	2	0	1	3	1	2.21
	d23		65	21	29	5	3	1	0	99	5	1	1	1	0	4	1	1	2	0	0.98
	d24		28	5	21	11	88	58	26	98	11	1	2	7	0	3	0	0	0	0	1.22
Polycomb	d25	c6	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	1	0	81	5.58
Heterochromatin	d26	c7	0	0	0	0	0	0	0	0	0	0	0	0	0	4	96	82	91	0	2.26
	d27		61	5	4	0	1	1	0	1	0	0	0	0	0	19	82	55	35	0	0.84
Heterochromatin-like in euch	d28	c8	0	0	0	0	0	0	0	3	0	0	0	0	0	0	3	91	85	2	1.62
	d29		0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	35	0	2.32
Basal, intergenic euchromatin	d30	c9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50.75

Modencode, Roy et al, Science 2010

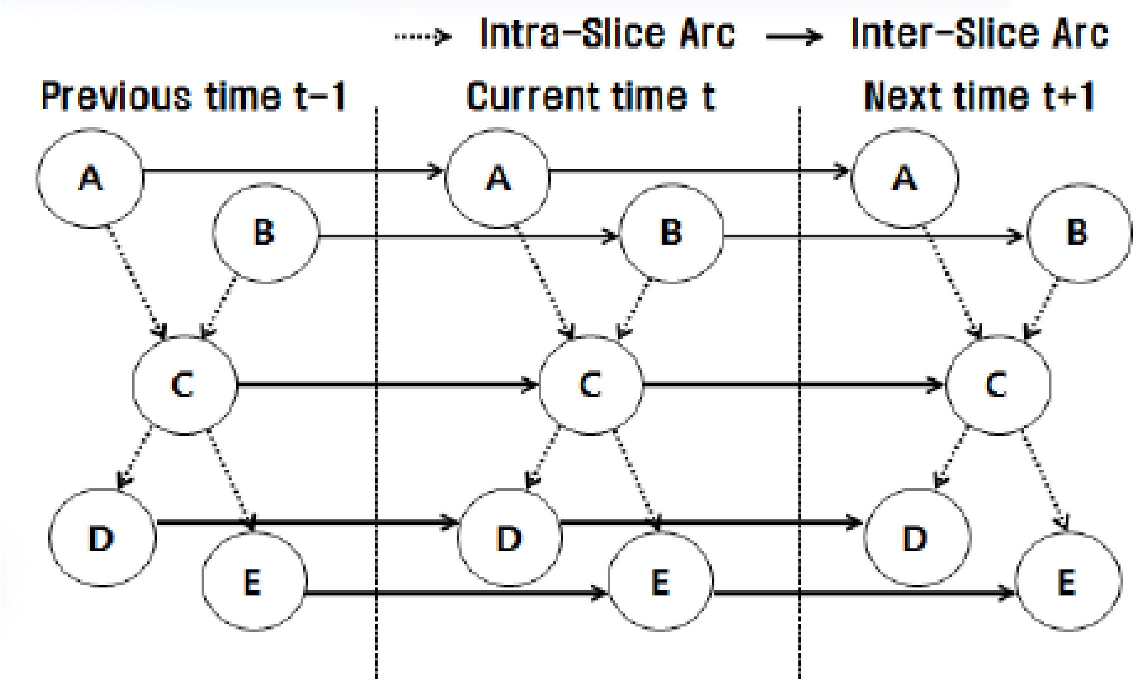
Bayesian Networks and Dynamic Bayesian Networks



	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

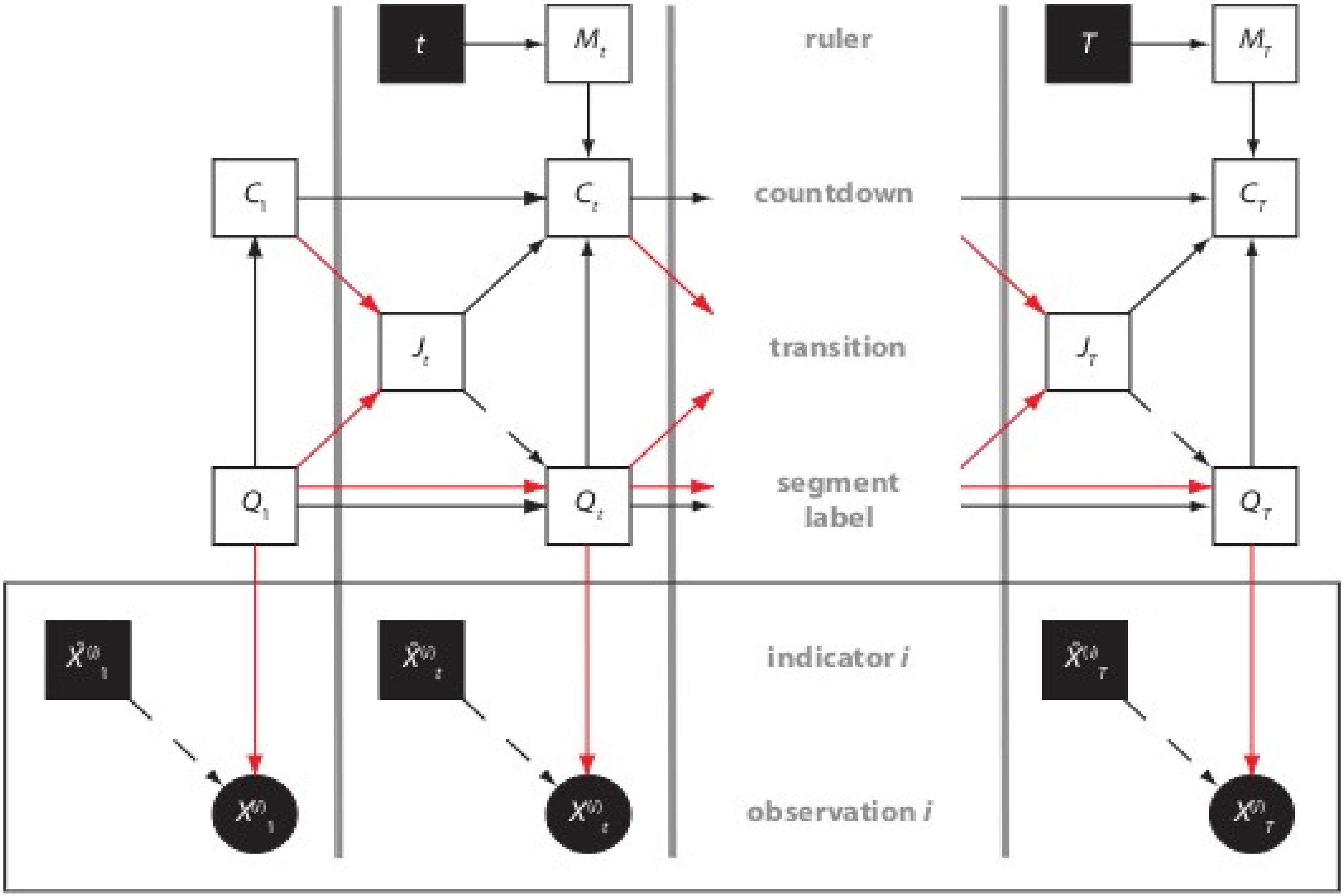
RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

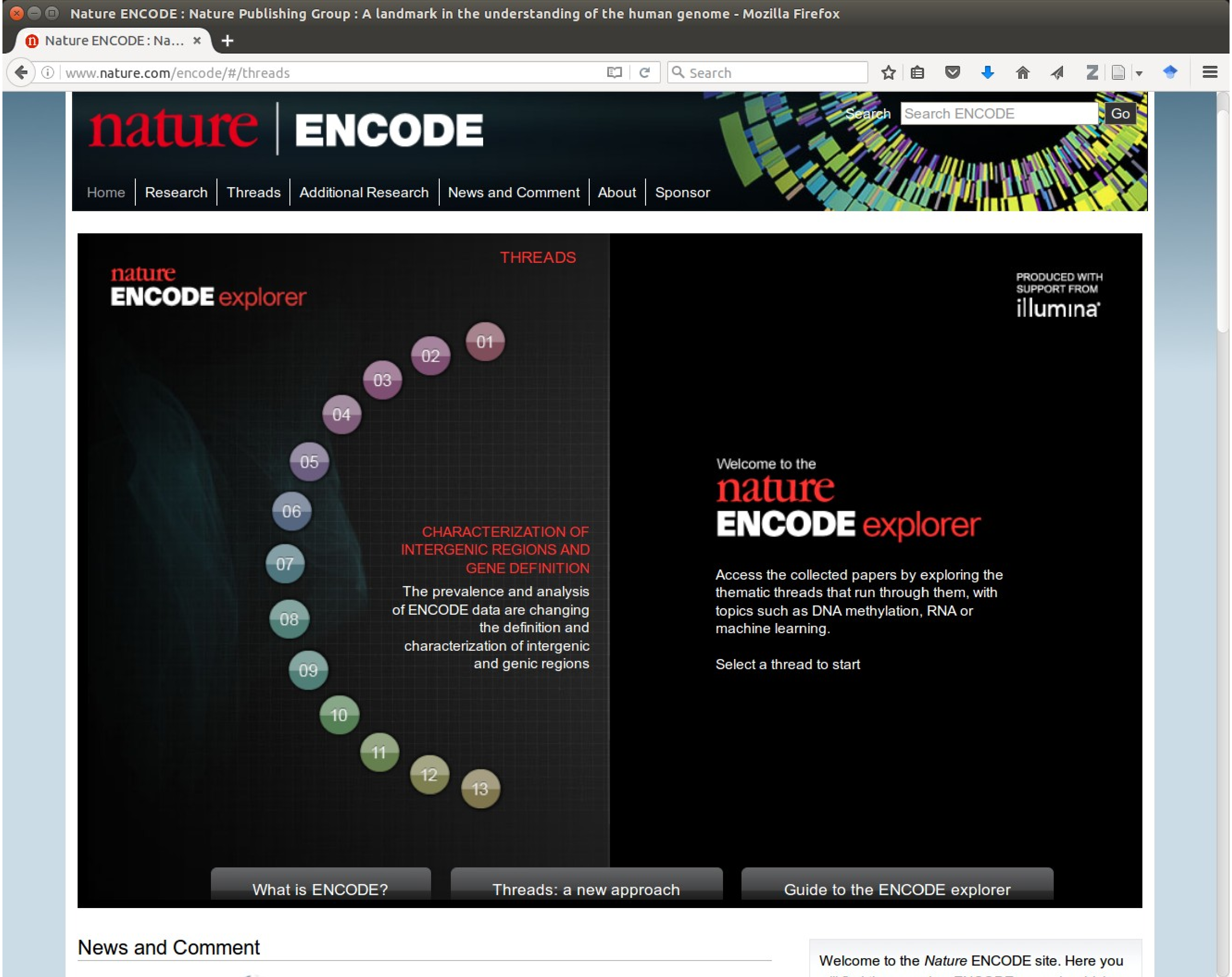


Segway Dynamic Bayesian Network

Supplementary Fig. 11: Graphical model representation of the default Segway DBN.



Hoffman et al. Nat. Methods 2012



CHARACTERIZATION OF INTERGENIC REGIONS AND GENE DEFINITION

The prevalence and analysis of ENCODE data are changing the definition and characterization of intergenic and genic regions

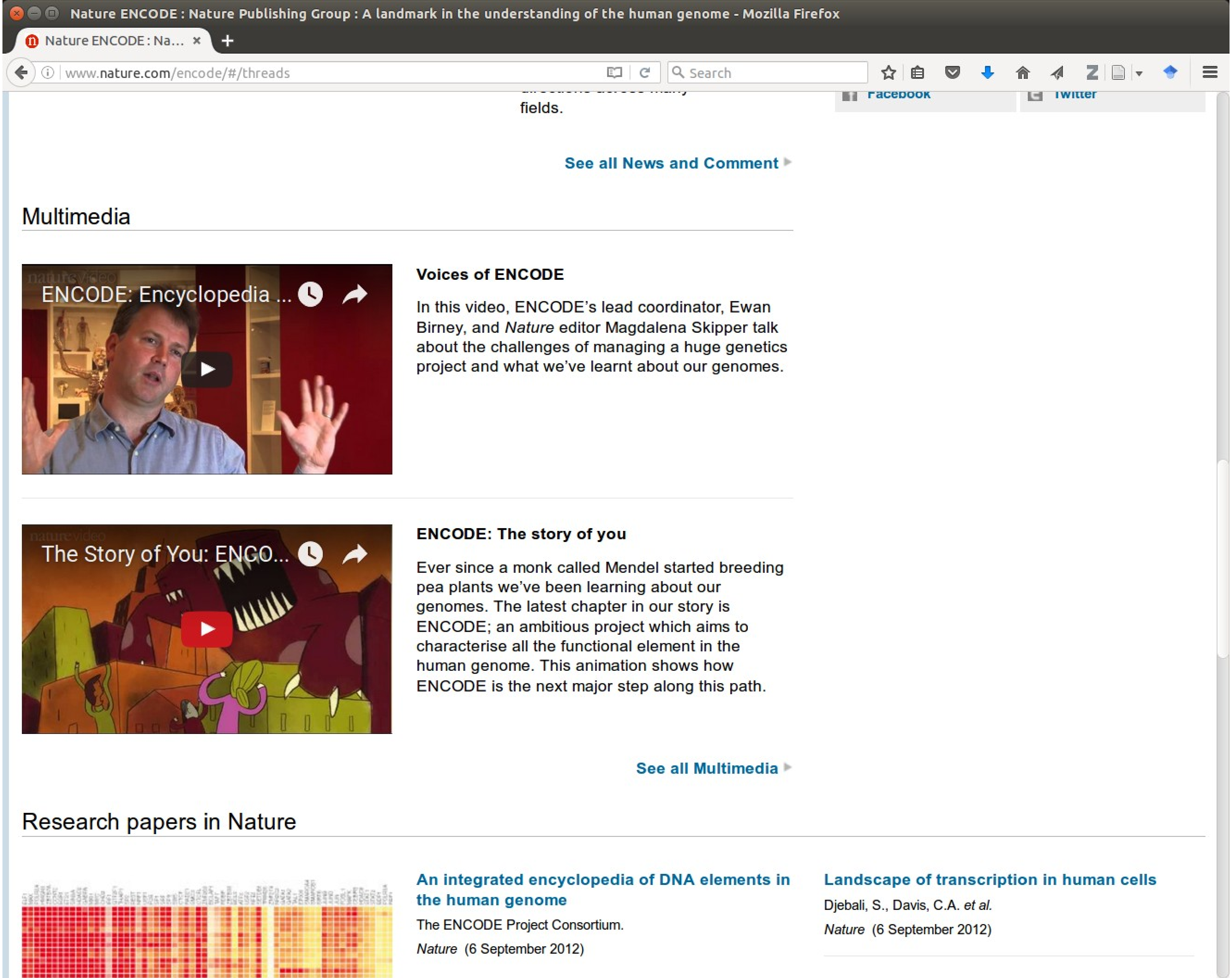
Welcome to the nature ENCODE explorer

Access the collected papers by exploring the thematic threads that run through them, with topics such as DNA methylation, RNA or machine learning.

Select a thread to start

News and Comment

Welcome to the *Nature* ENCODE site. Here you


Nature ENCODE : Nature Publishing Group : A landmark in the understanding of the human genome - Mozilla Firefox
Nature ENCODE : Na...
www.nature.com/encode/#/threads
Search
Facebook
Twitter
naturevideo
ENCODE: Encyclopedia ...
naturevideo
The Story of You: ENCO...

fields.

See all News and Comment

Multimedia

Voices of ENCODE

In this video, ENCODE's lead coordinator, Ewan Birney, and *Nature* editor Magdalena Skipper talk about the challenges of managing a huge genetics project and what we've learnt about our genomes.

ENCODE: The story of you

Ever since a monk called Mendel started breeding pea plants we've been learning about our genomes. The latest chapter in our story is ENCODE; an ambitious project which aims to characterise all the functional element in the human genome. This animation shows how ENCODE is the next major step along this path.

See all Multimedia

Research papers in Nature

An integrated encyclopedia of DNA elements in the human genome

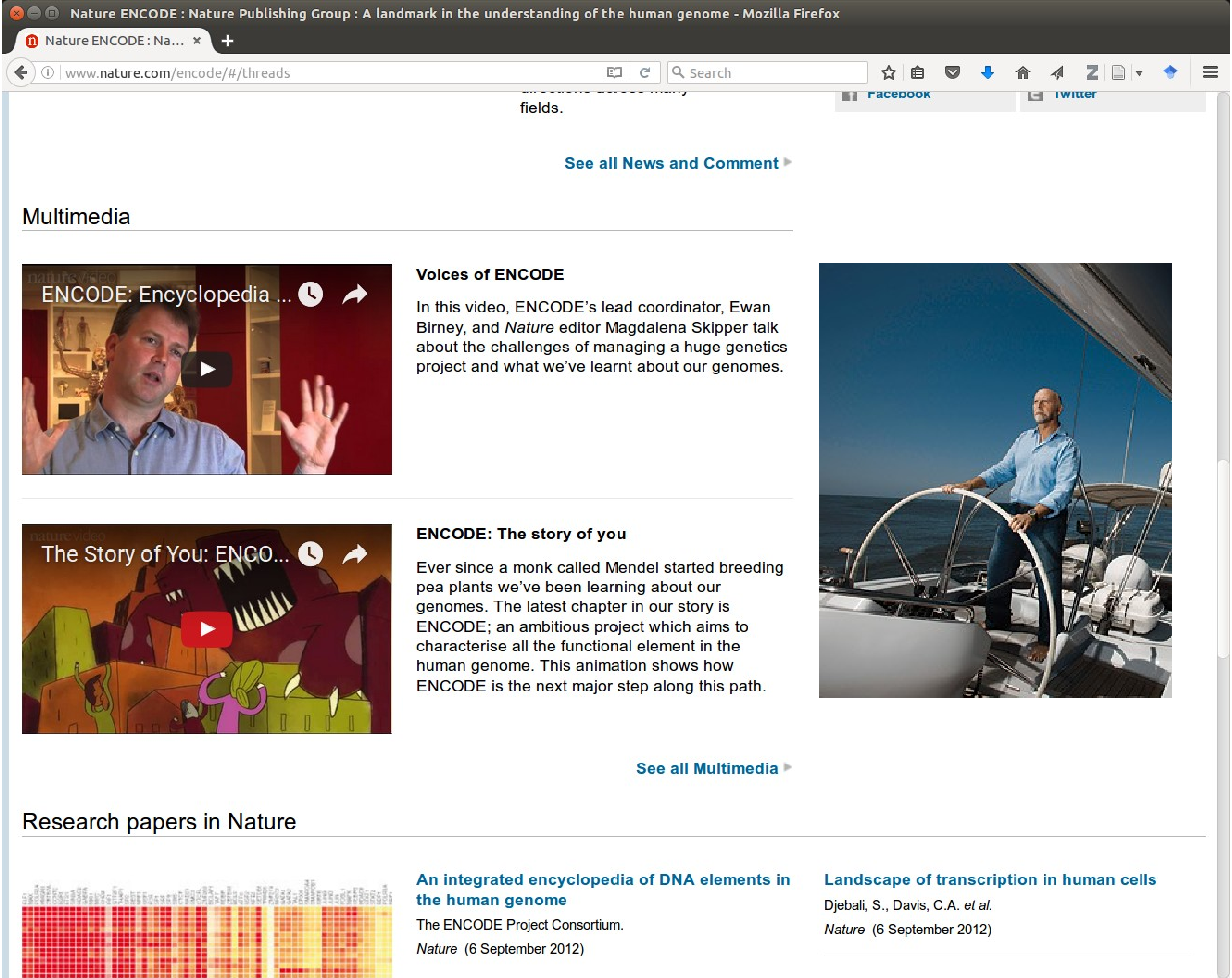
The ENCODE Project Consortium.

Nature (6 September 2012)

Landscape of transcription in human cells

Djebali, S., Davis, C.A. *et al.*

Nature (6 September 2012)


Nature ENCODE : Nature Publishing Group : A landmark in the understanding of the human genome - Mozilla Firefox
Nature ENCODE : Na...
www.nature.com/encode/#/threads
Search
Facebook
Twitter
ENCODE: Encyclopedia ...
The Story of You: ENCO...

fields.

See all News and Comment

Multimedia

Voices of ENCODE

In this video, ENCODE's lead coordinator, Ewan Birney, and *Nature* editor Magdalena Skipper talk about the challenges of managing a huge genetics project and what we've learnt about our genomes.

ENCODE: The story of you

Ever since a monk called Mendel started breeding pea plants we've been learning about our genomes. The latest chapter in our story is ENCODE; an ambitious project which aims to characterise all the functional element in the human genome. This animation shows how ENCODE is the next major step along this path.

See all Multimedia

Research papers in Nature

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium.

Nature (6 September 2012)

Landscape of transcription in human cells

Djebali, S., Davis, C.A. *et al.*

Nature (6 September 2012)

On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE

Dan Graur[1,*], Yichen Zheng[1], Nicholas Price[1], Ricardo B.R. Azevedo[1], Rebecca A. Zufall[1], and Eran Elhaik[2]

[1]Department of Biology and Biochemistry, University of Houston

[2]Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health

***Corresponding author:** E-mail: dgraur@uh.edu.

Accepted: February 16, 2013

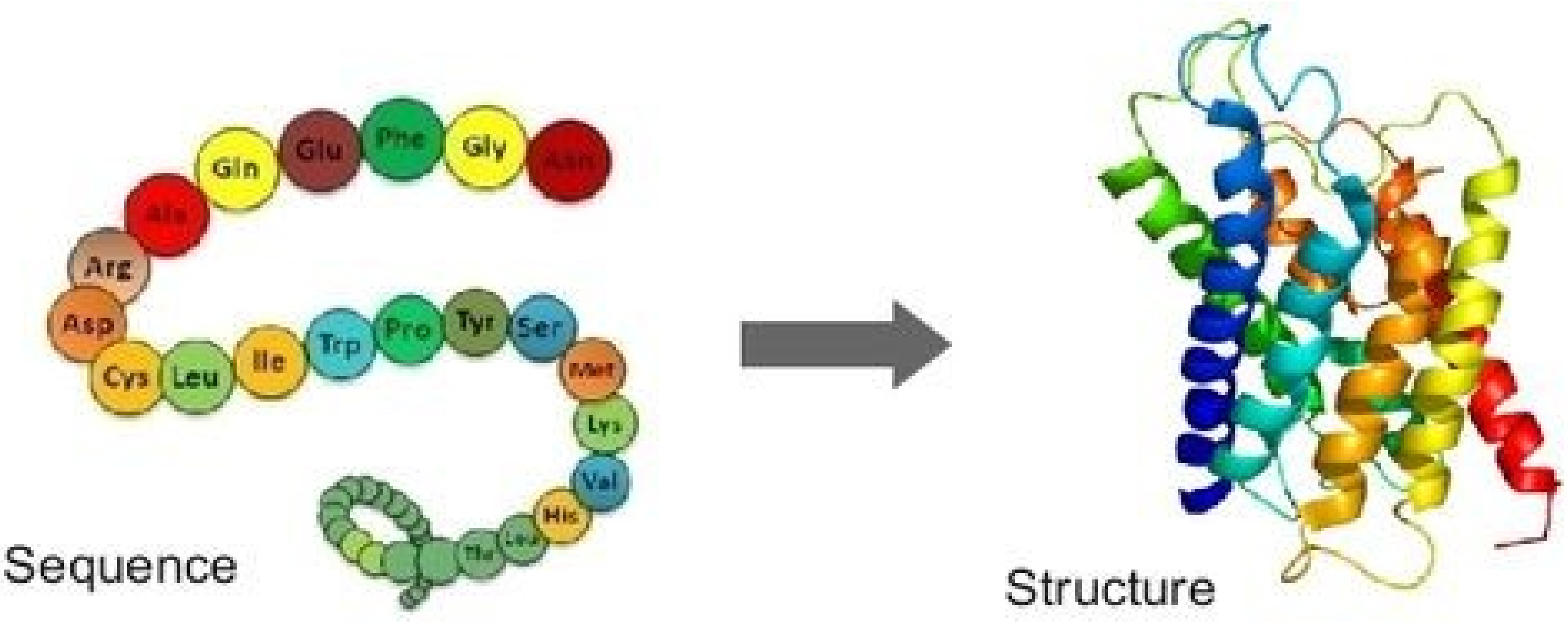
Abstract

A recent slew of ENCyclopedia Of DNA Elements (ENCODE) Consortium publications, specifically the article signed by all Consortium members, put forward the idea that more than 80% of the human genome is functional. This claim flies in the face of current estimates according to which the fraction of the genome that is evolutionarily conserved through purifying selection is less than 10%. Thus, according to the ENCODE Consortium, a biological function can be maintained indefinitely without selection, which implies that at least 80 – 10 = 70% of the genome is perfectly invulnerable to deleterious mutations, either because no mutation can ever occur in these "functional" regions or because no mutation in these regions can ever be deleterious. This absurd conclusion was reached through various means, chiefly by employing the seldom used "causal role" definition of biological function and then applying it inconsistently to different biochemical properties, by committing a logical fallacy known as "affirming the consequent," by failing to appreciate the crucial difference between "junk DNA" and "garbage DNA," by using analytical methods that yield biased errors and inflate estimates of functionality, by favoring statistical sensitivity over specificity, and by emphasizing statistical significance rather than the magnitude of the effect. Here, we detail the many logical and methodological transgressions involved in assigning functionality to almost every nucleotide in the human genome. The ENCODE results were predicted by one of its authors to necessitate the rewriting of textbooks. We agree, many textbooks dealing with marketing, mass-media hype, and public relations may well have to be rewritten.

Key words: junk DNA, genome functionality, selection, ENCODE project.

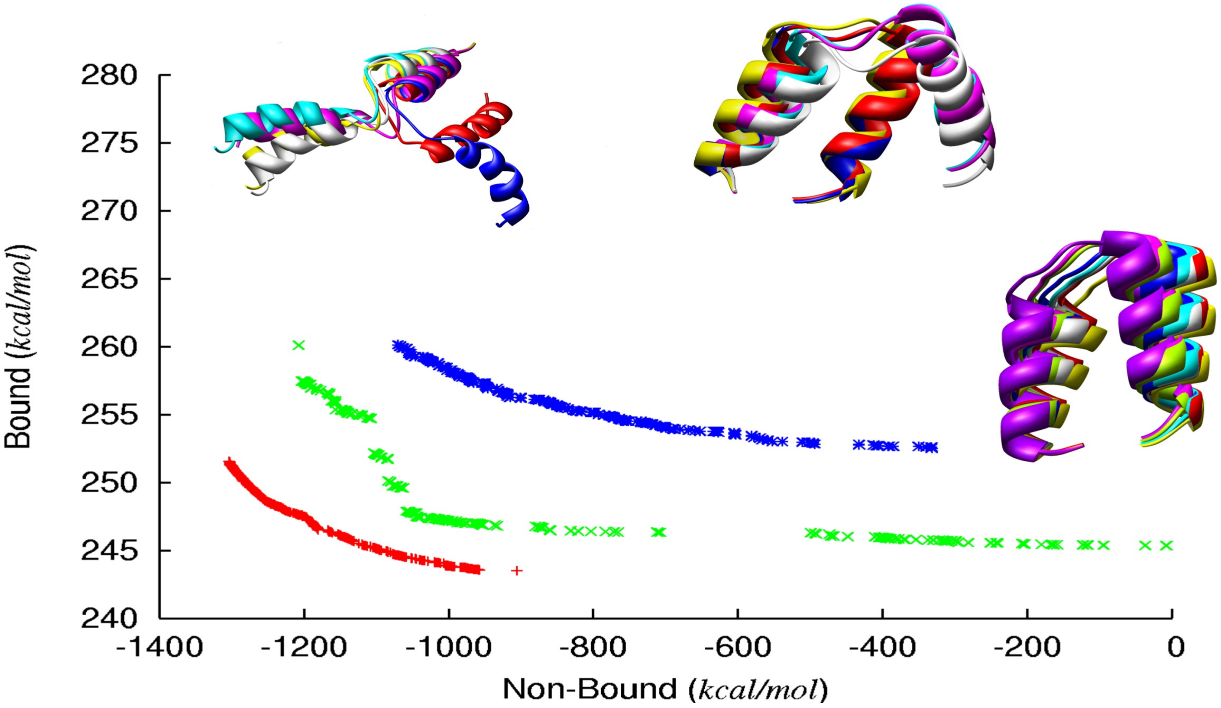
Protein structure prediction

- We can predict the protein sequence from reading DNA, but we do not know how it will fold to perform its function

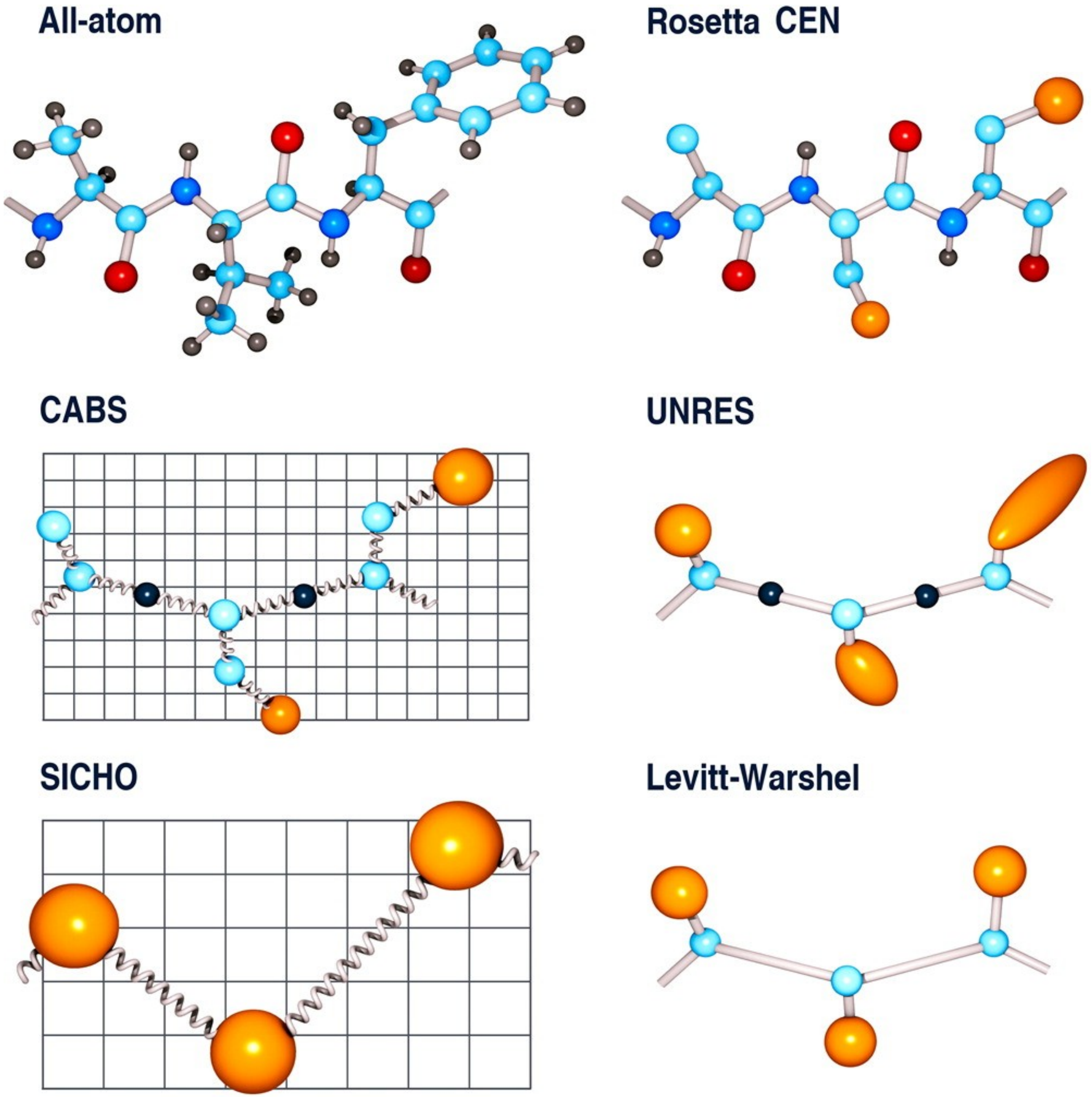


Protein structure energy function

- Given our understanding of molecular dynamics, we should be able to score different conformations of the same protein chain
- This is expensive, as proteins contain thousands of atoms

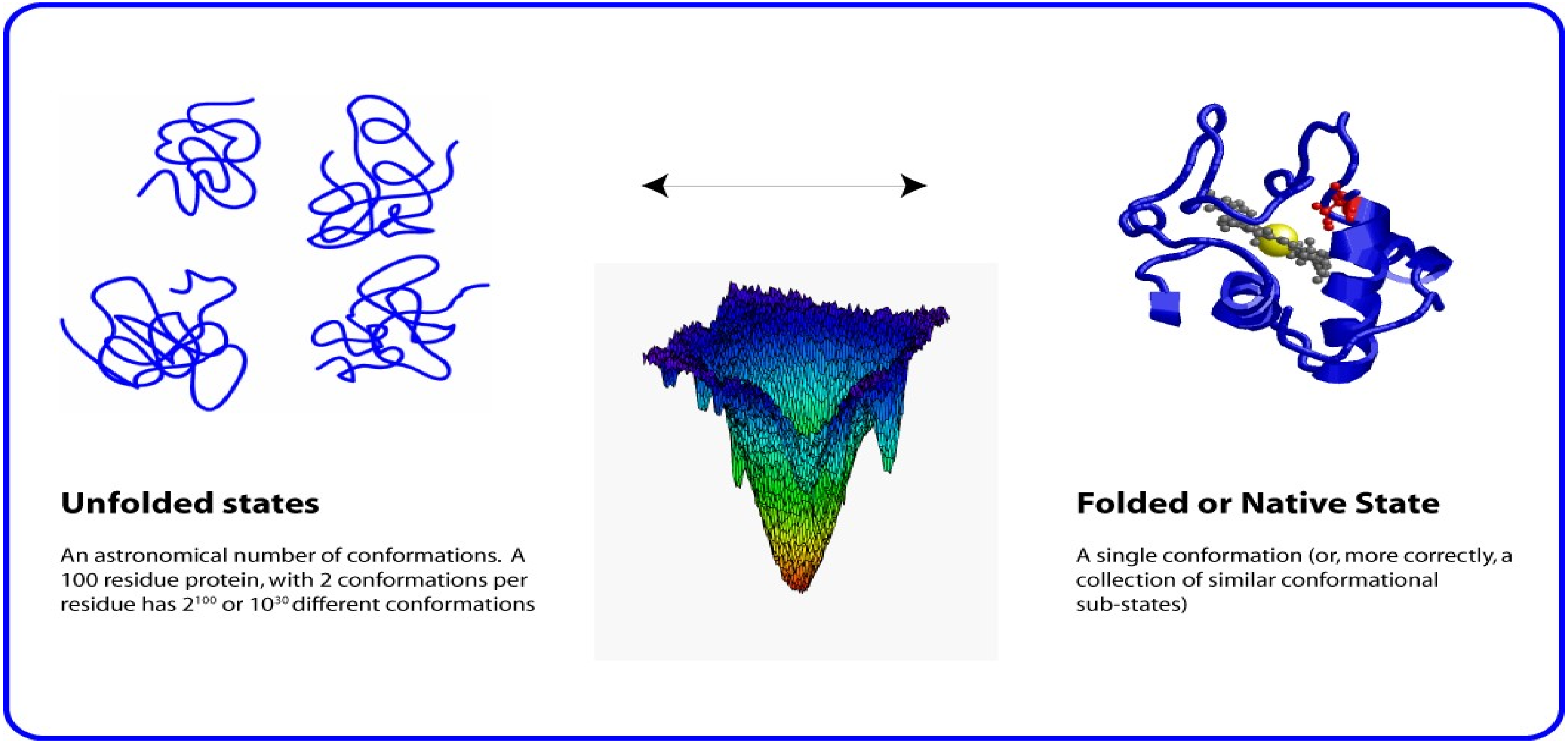


Simplified Computational models of protein structure



Anfinsen's „conjecture”

- Since proteins can fold in the real world, the energy landscape should have a very strong global optimum



Unfolded states

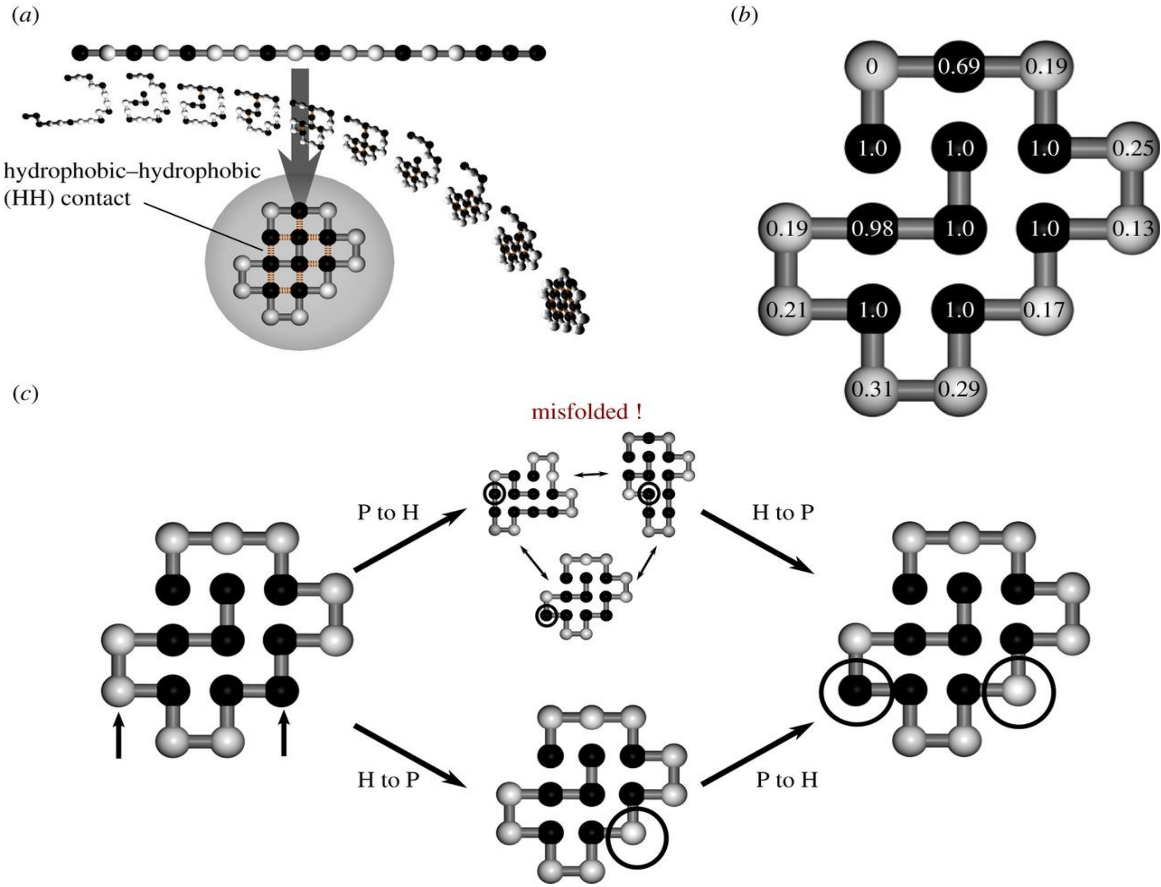
An astronomical number of conformations. A 100 residue protein, with 2 conformations per residue has 2^{100} or 10^{30} different conformations

Folded or Native State

A single conformation (or, more correctly, a collection of similar conformational sub-states)

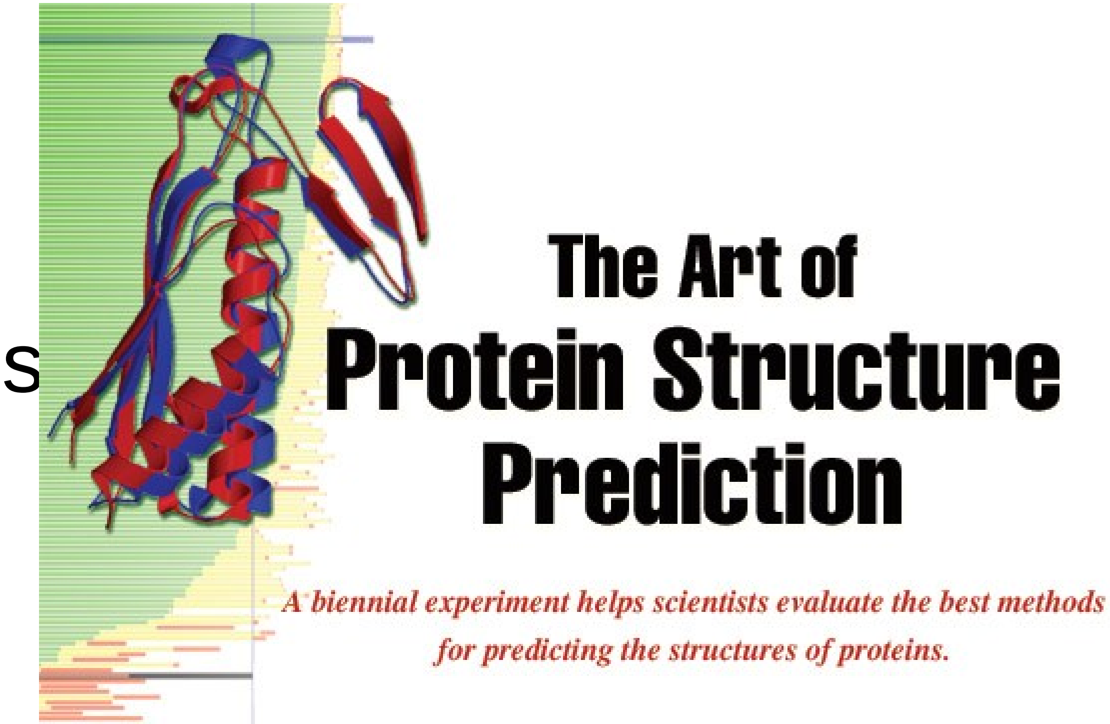
Computationally this is difficult

- Even the simplest model:
  - hydrophobic/polar representation of residues
  - On a rectangular lattive
- leads to a NP-hard problem of finding the optimal configuration



CASP experiment

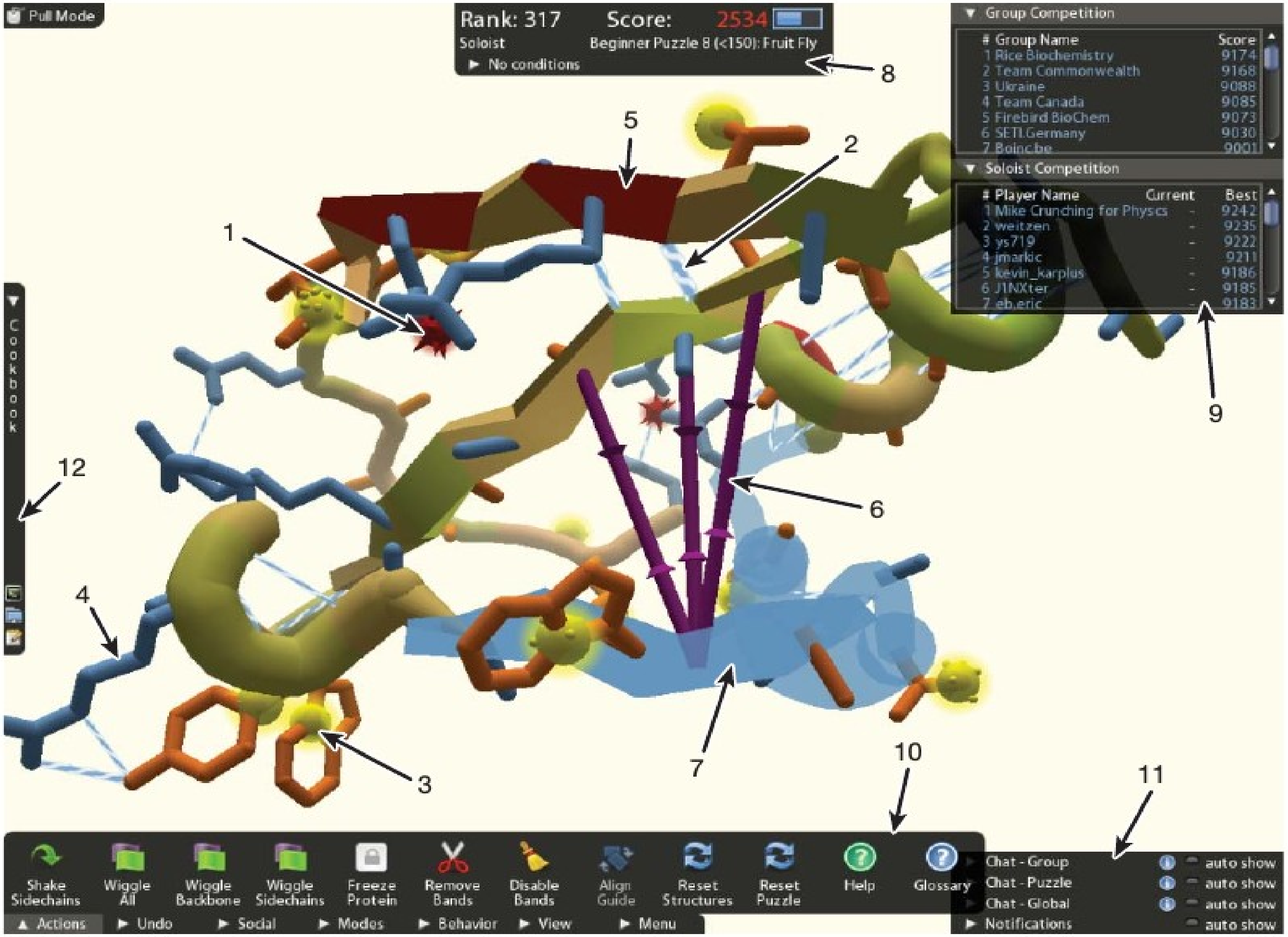
- Critical Assessment of Structure Prediction methods
- Crystallographers solve structures and release sequences to scientists so that they can make blind predictions



The Art of Protein Structure Prediction

A biennial experiment helps scientists evaluate the best methods for predicting the structures of proteins.

Gamification of protein folding



Rank: 317 Score: 2534
Soloist Beginner Puzzle 8 (<150): Fruit Fly
► No conditions

▼ Group Competition

#	Group Name	Score
1	Rice Biochemistry	9174
2	Team Commonwealth	9168
3	Ukraine	9088
4	Team Canada	9085
5	Firebird BioChem	9073
6	SETI.Germany	9030
7	Boinc.be	9001

▼ Soloist Competition

#	Player Name	Current	Best
1	Mike Crunching for Physics	-	9242
2	weitzen	-	9235
3	ys719	-	9222
4	jmarkic	-	9211
5	kevin_karplus	-	9186
6	J1NXter	-	9185
7	eb.eric	-	9183

Figure 1 | Foldit screenshot illustrating tools and visualizations. The visualizations include a clash representing atoms that are too close (arrow 1); a hydrogen bond (arrow 2); a hydrophobic side chain with a yellow blob because it is exposed (arrow 3); a hydrophilic side chain (arrow 4); and a segment of the backbone that is red due to high residue energy (arrow 5). The players can make modifications including 'rubber bands' (arrow 6), which add constraints to guide automated tools, and freezing (arrow 7), which prevents degrees of freedom from changing. The user interface includes information about the player's current status, including score (arrow 8); a leader board (arrow 9), which shows the scores of other players and groups; toolbars for accessing tools and options (arrow 10); chat for interacting with other players (arrow 11); and a 'cookbook' for making new automated tools or 'recipes' (arrow 12).

nature Vol 466|5 August 2010|**doi:10.1038/nature09304**

LETTERS

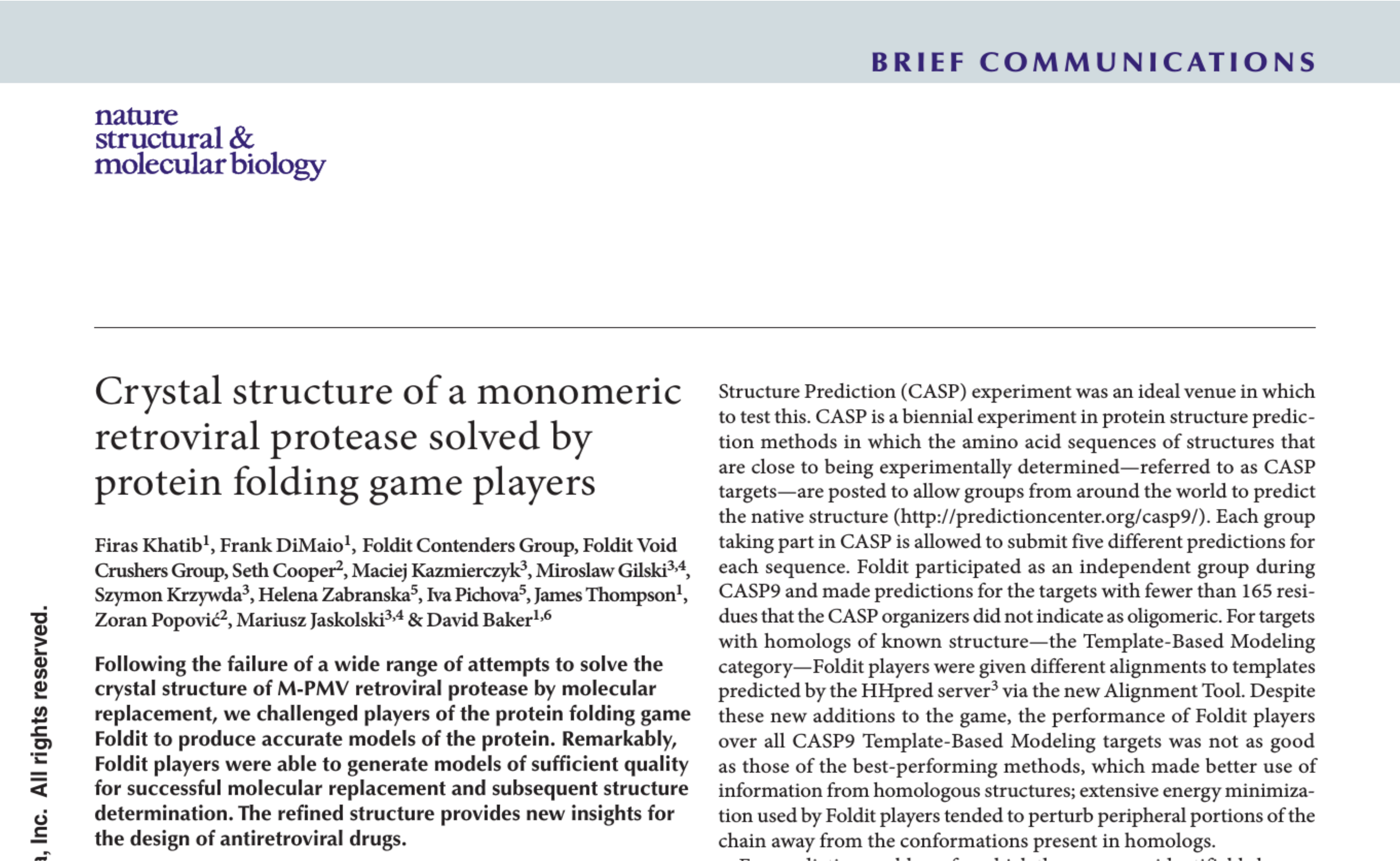
Predicting protein structures with a multiplayer online game

Seth Cooper[1], Firas Khatib[2], Adrien Treuille[1,3], Janos Barbero[1], Jeehyung Lee[3], Michael Beenen[1], Andrew Leaver-Fay[2]†, David Baker[2,4], Zoran Popović[1] & Foldit players

People exert large amounts of problem-solving effort playing computer games. Simple image- and text-recognition tasks have been successfully 'crowd-sourced' through games[1–3], but it is not clear if more complex scientific problems can be solved with human-directed computing. Protein structure prediction is one such problem: locating the biologically relevant native conformation of a protein is a formidable computational challenge given the very large size of the search space. Here we describe Foldit, a multiplayer online game that engages non-scientists in solving hard prediction problems. Foldit players interact with protein structures using direct manipulation tools and user-friendly versions of algorithms from the Rosetta structure prediction methodology[4], while they compete and collaborate to optimize the computed energy. We show that top-ranked Foldit players excel at solving challenging structure refinement problems in which substantial backbone rearrangements are necessary to achieve the burial of hydrophobic residues. Players working

retaining the deterministic Rosetta algorithms as user tools. We developed a multiplayer online game, Foldit, with the goal of producing accurate protein structure models through gameplay (Fig. 1). Improperly folded protein conformations are posted online as puzzles for a fixed amount of time, during which players interactively reshape them in the direction they believe will lead to the highest score (the negative of the Rosetta energy). The player's current status is shown, along with a leader board of other players, and groups of players working together, competing in the same puzzle (Fig. 1, arrows 8 and 9). To make the game approachable by players with no scientific training, many technical terms are replaced by terms in more common usage. We remove protein elements that hinder structural problem solving, and highlight energetically frustrated areas of the protein where the player can probably improve the structure (Fig. 1, arrows 1–5). Side chains are coloured by hydrophobicity and the backbone is coloured by energy. There are specific visual cues depicting hydrophobicity ('exposed hydrophobics'), interatomic

Solving new HIV protein structure



BRIEF COMMUNICATIONS

nature
structural &
molecular biology

Crystal structure of a monomeric retroviral protease solved by protein folding game players

Firas Khatib[1], Frank DiMaio[1], Foldit Contenders Group, Foldit Void Crushers Group, Seth Cooper[2], Maciej Kazmierczyk[3], Miroslaw Gilski[3,4], Szymon Krzywda[3], Helena Zabranska[5], Iva Pichova[5], James Thompson[1], Zoran Popović[2], Mariusz Jaskolski[3,4] & David Baker[1,6]

Following the failure of a wide range of attempts to solve the crystal structure of M-PMV retroviral protease by molecular replacement, we challenged players of the protein folding game Foldit to produce accurate models of the protein. Remarkably, Foldit players were able to generate models of sufficient quality for successful molecular replacement and subsequent structure determination. The refined structure provides new insights for the design of antiretroviral drugs.

Structure Prediction (CASP) experiment was an ideal venue in which to test this. CASP is a biennial experiment in protein structure prediction methods in which the amino acid sequences of structures that are close to being experimentally determined—referred to as CASP targets—are posted to allow groups from around the world to predict the native structure (http://predictioncenter.org/casp9/). Each group taking part in CASP is allowed to submit five different predictions for each sequence. Foldit participated as an independent group during CASP9 and made predictions for the targets with fewer than 165 residues that the CASP organizers did not indicate as oligomeric. For targets with homologs of known structure—the Template-Based Modeling category—Foldit players were given different alignments to templates predicted by the HHpred server[3] via the new Alignment Tool. Despite these new additions to the game, the performance of Foldit players over all CASP9 Template-Based Modeling targets was not as good as those of the best-performing methods, which made better use of information from homologous structures; extensive energy minimization used by Foldit players tended to perturb peripheral portions of the chain away from the conformations present in homologs.

Finding new algorithms

Algorithm discovery by protein folding game players

Firas Khatib[a], Seth Cooper[b], Michael D. Tyka[a], Kefan Xu[b], Ilya Makedon[b], Zoran Popović[b], David Baker[a,c,1], and Foldit Players

[a]Department of Biochemistry; [b]Department of Computer Science and Engineering; and [c]Howard Hughes Medical Institute, University of Washington, Box 357370, Seattle, WA 98195

Contributed by David Baker, October 5, 2011 (sent for review June 29, 2011)

Foldit is a multiplayer online game in which players collaborate and compete to create accurate protein structure models. For specific hard problems, Foldit player solutions can in some cases outperform state-of-the-art computational methods. However, very little is known about how collaborative gameplay produces these results and whether Foldit player strategies can be formalized and structured so that they can be used by computers. To determine whether high performing player strategies could be collectively codified, we augmented the Foldit gameplay mechanics with tools for players to encode their folding strategies as "recipes" and to share their recipes with other players, who are able to further modify and redistribute them. Here we describe the rapid social evolution of player-developed folding algorithms that took place in the year following the introduction of these tools. Players developed over 5,400 different recipes, both by creating new algorithms and by modifying and recombining successful recipes developed by other players. The most successful recipes rapidly spread through the Foldit player population, and two of the recipes became particularly dominant. Examination of the algorithms encoded in these

As the players themselves understand their strategies better than anyone, we decided to allow them to codify their algorithms directly, rather than attempting to automatically learn approximations. We augmented standard Foldit play with the ability to create, edit, share, and rate gameplay macros, referred to as "recipes" within the Foldit game (10). In the game each player has their own "cookbook" of such recipes, from which they can invoke a variety of interactive automated strategies. Players can share recipes they write with the rest of the Foldit community or they can choose to keep their creations to themselves.

In this paper we describe the quite unexpected evolution of recipes in the year after they were released, and the striking convergence of this very short evolution on an algorithm very similar to an unpublished algorithm recently developed independently by scientific experts that improves over previous methods.

Results

In the social development environment provided by Foldit, players evolved a wide variety of recipes to codify their diverse

Making improved enzymes

LETTERS

nature biotechnology

Increased Diels-Alderase activity through backbone remodeling guided by Foldit players

Christopher B Eiben[1,6], Justin B Siegel[1,6], Jacob B Bale[1,2], Seth Cooper[3], Firas Khatib[1], Betty W Shen[4], Foldit Players, Barry L Stoddard[4], Zoran Popovic[3] & David Baker[1,5]

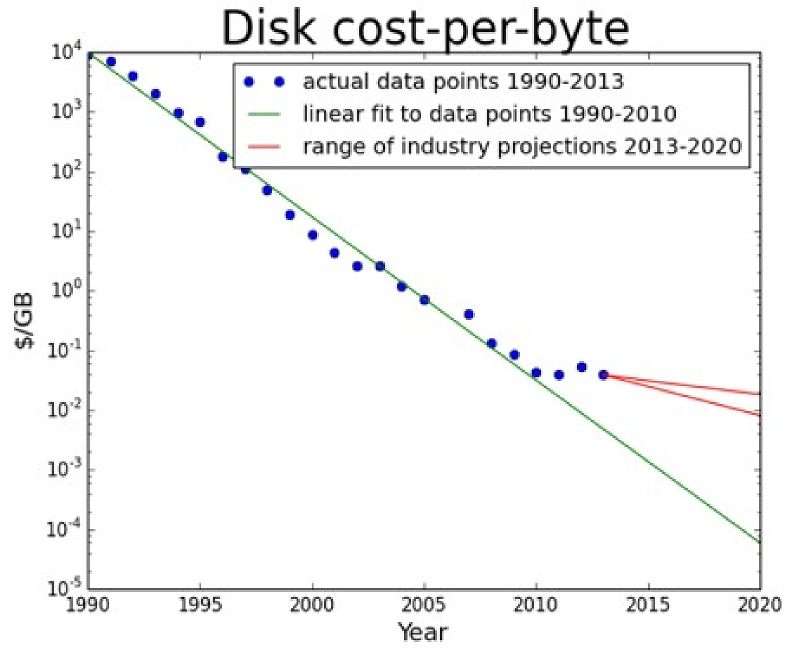
Computational enzyme design holds promise for the production of renewable fuels, drugs and chemicals. *De novo* enzyme design has generated catalysts for several reactions, but with lower catalytic efficiencies than naturally occurring enzymes[1–4]. Here we report the use of game-driven crowdsourcing to enhance the activity of a computationally designed enzyme through the functional remodeling of its structure. Players of the online game Foldit[5,6] were challenged to remodel the backbone of a computationally designed bimolecular Diels-Alderase[3] to enable additional interactions with substrates. Several iterations of design and characterization generated a 24-residue helix-turn-helix motif, including a 13-residue insertion, that increased enzyme activity >18-fold. X-ray crystallography showed that the large insertion adopts a helix-turn-helix structure positioned as in the Foldit model.

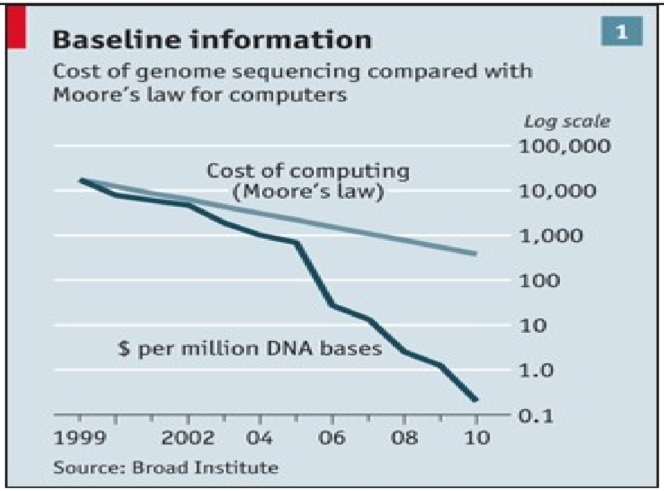
To explore whether human creativity can guide the search in this substantially larger space, we incorporated new tools allowing insertions, deletions and sequence substitutions into Foldit to supplement the existing tools available for manipulating protein conformation. To integrate players into the experimental design process, we presented them with a series of puzzles. To connect Foldit player iterative exploration with experimental testing, we established an advanced Foldit player as an intermediary between the Foldit community and the experimental laboratory, who presented players with puzzles at each stage of the design process. Using Foldit, the advanced player analyzed the top-ranking community designs and built sequence libraries around the structures to stabilize favorable interactions. The designs were then experimentally tested, and the best were used as input for the next puzzle posted to the online community (**Supplementary Fig. 1**).

We challenged Foldit players to remodel the active-site loops of a

Kryder's law

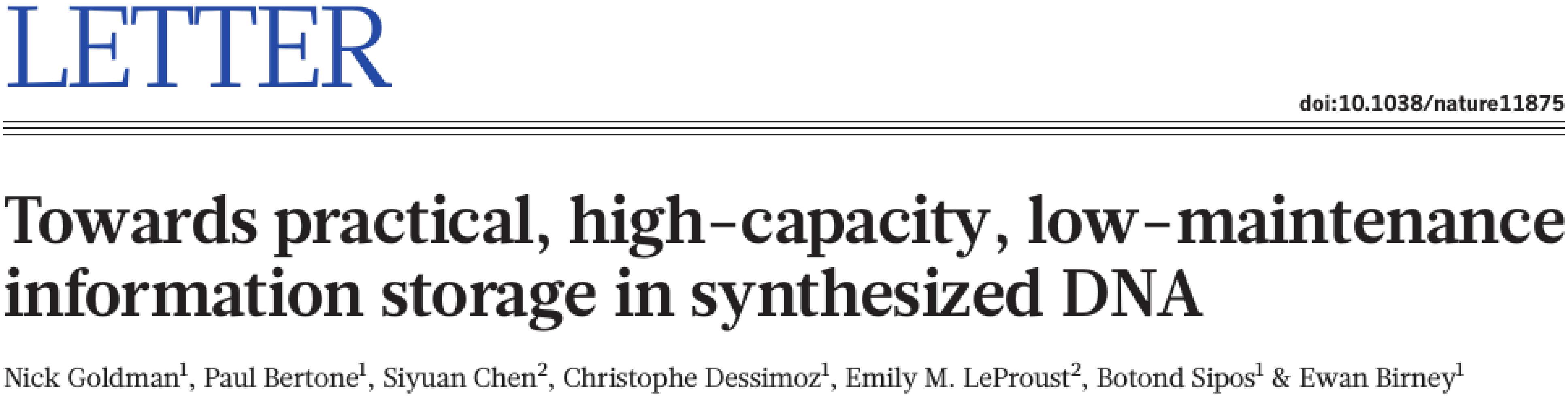
- For a long time the cost of magnetic storage was following Kryder's law of exponential reduction
- It is no longer the case
- It creates problems for storing all the sequencing data





Source: Broad Institute

Storing data in DNA



LETTER

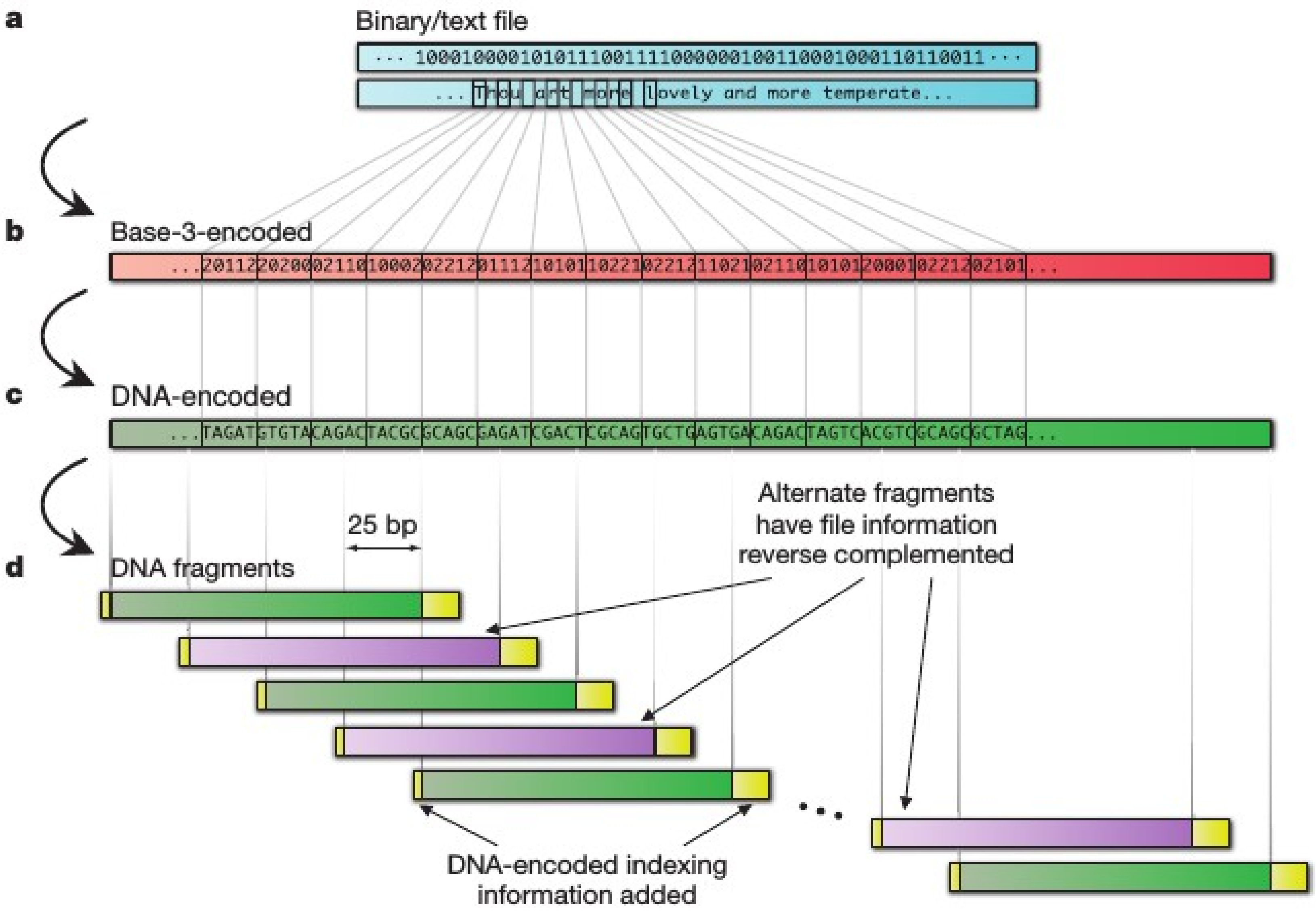
doi:10.1038/nature11875

Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

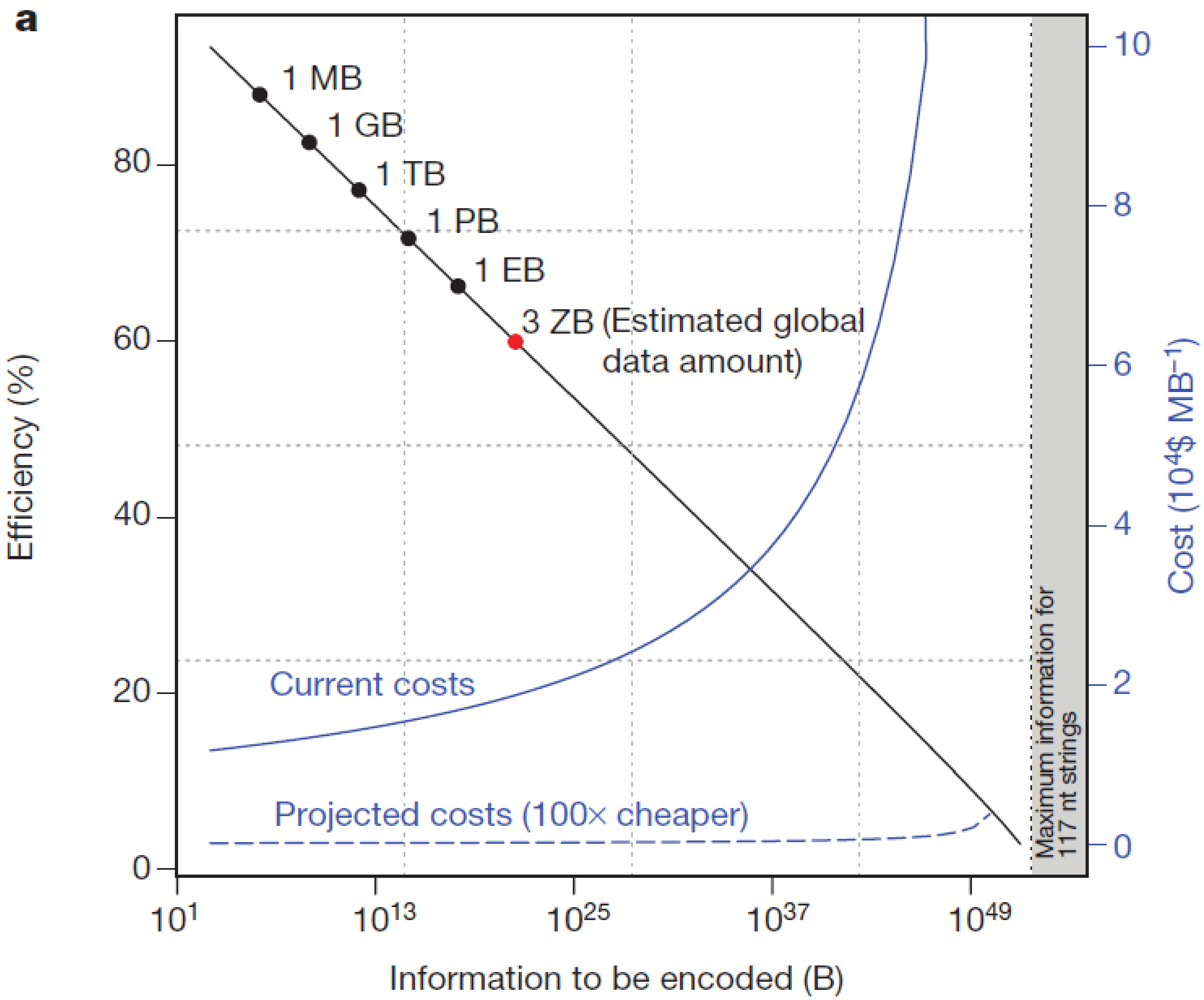
Nick Goldman[1], Paul Bertone[1], Siyuan Chen[2], Christophe Dessimoz[1], Emily M. LeProust[2], Botond Sipos[1] & Ewan Birney[1]

- Stored a text file, few images, a sound file in the DNA

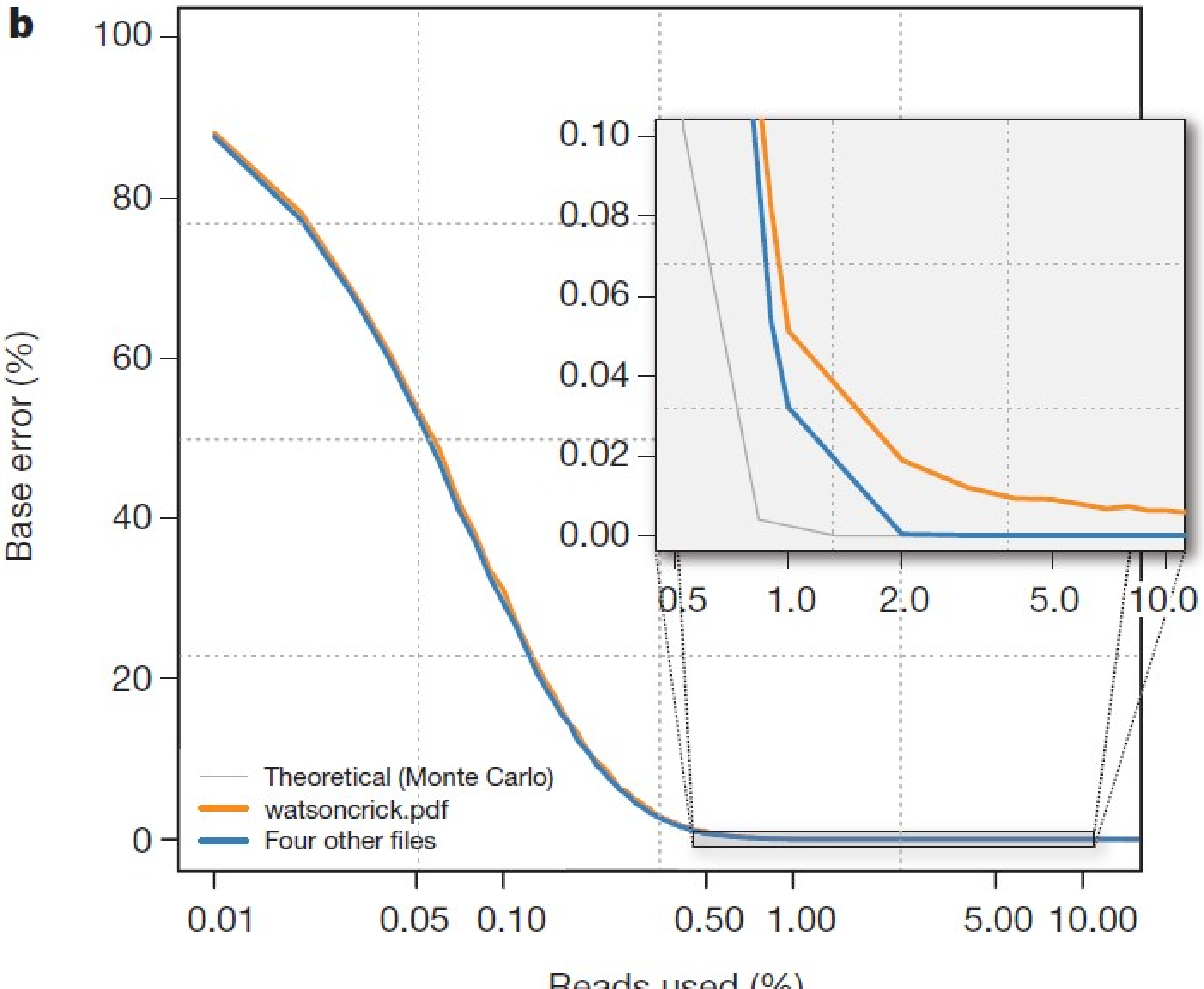
Encoding of a binary stream in a sequencable DNA



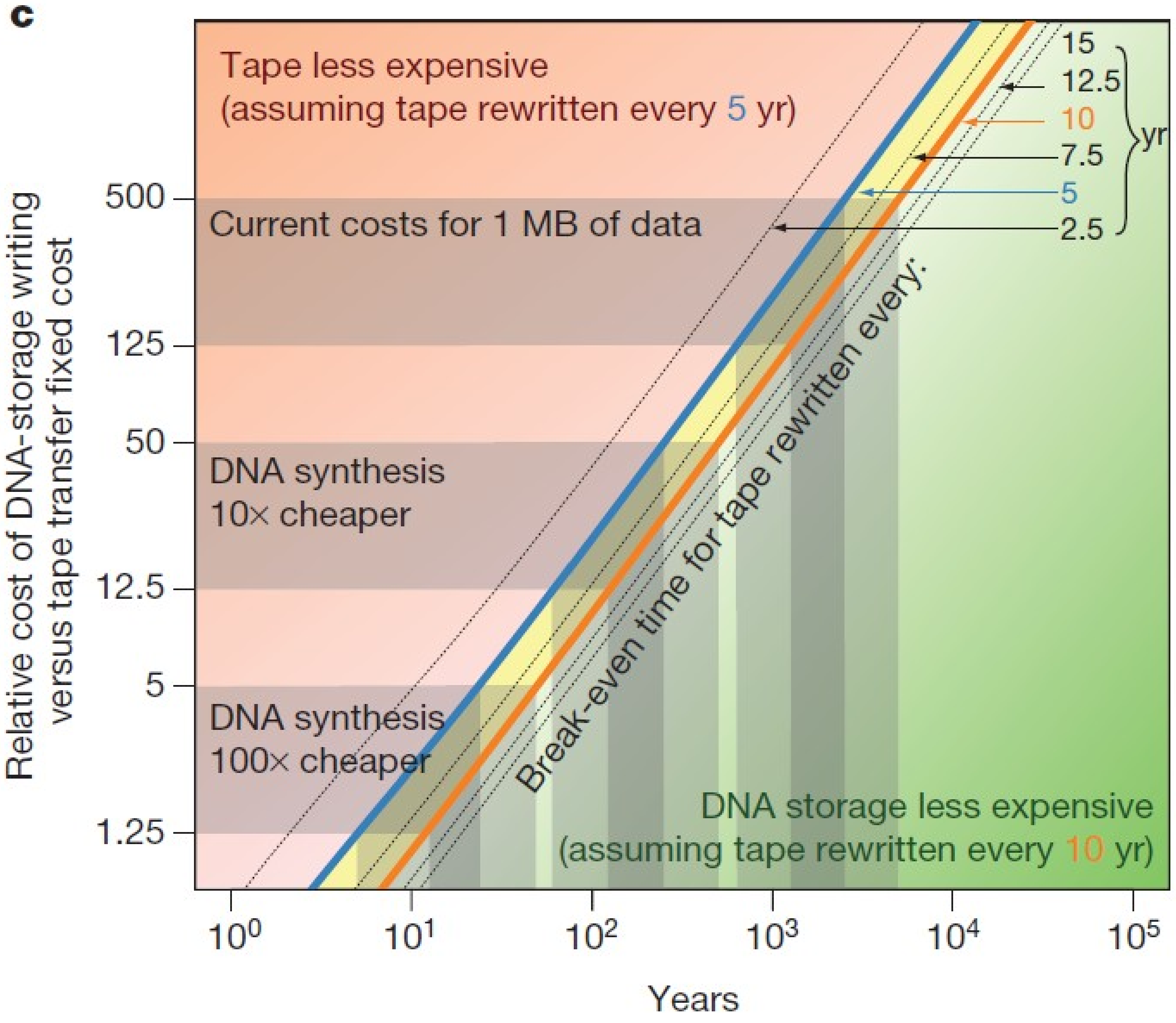
Cost of storing data in DNA



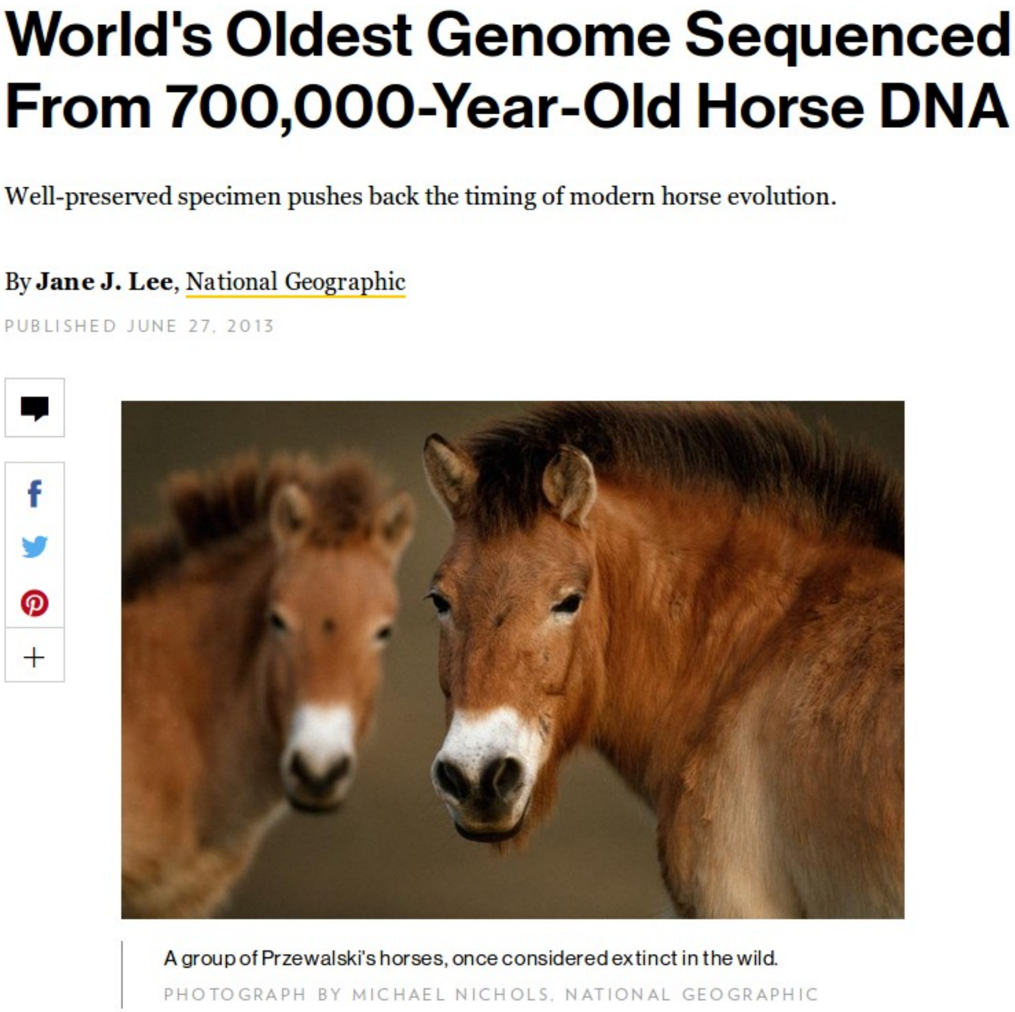
Cost of retrieving DNA stored data



Cost comparison with tape storage



DNA is not only small it's also extremely durable



World's Oldest Genome Sequenced From 700,000-Year-Old Horse DNA

Well-preserved specimen pushes back the timing of modern horse evolution.

By **Jane J. Lee**, National Geographic

PUBLISHED JUNE 27, 2013

A group of Przewalski's horses, once considered extinct in the wild.

PHOTOGRAPH BY MICHAEL NICHOLS, NATIONAL GEOGRAPHIC

TheScientist

JUNE 2015 | WWW.THE-SCIENTIST.COM

EXPLORING LIFE, INSPIRING INNOVATION

ANCIENT DNA REVOLUTION

NEW TECHNOLOGIES TURN BACK THE CLOCK ON SEQUENCING FOSSIL REMAINS

AUTOCATALYTIC SETS MODEL LIFE'S ORIGIN

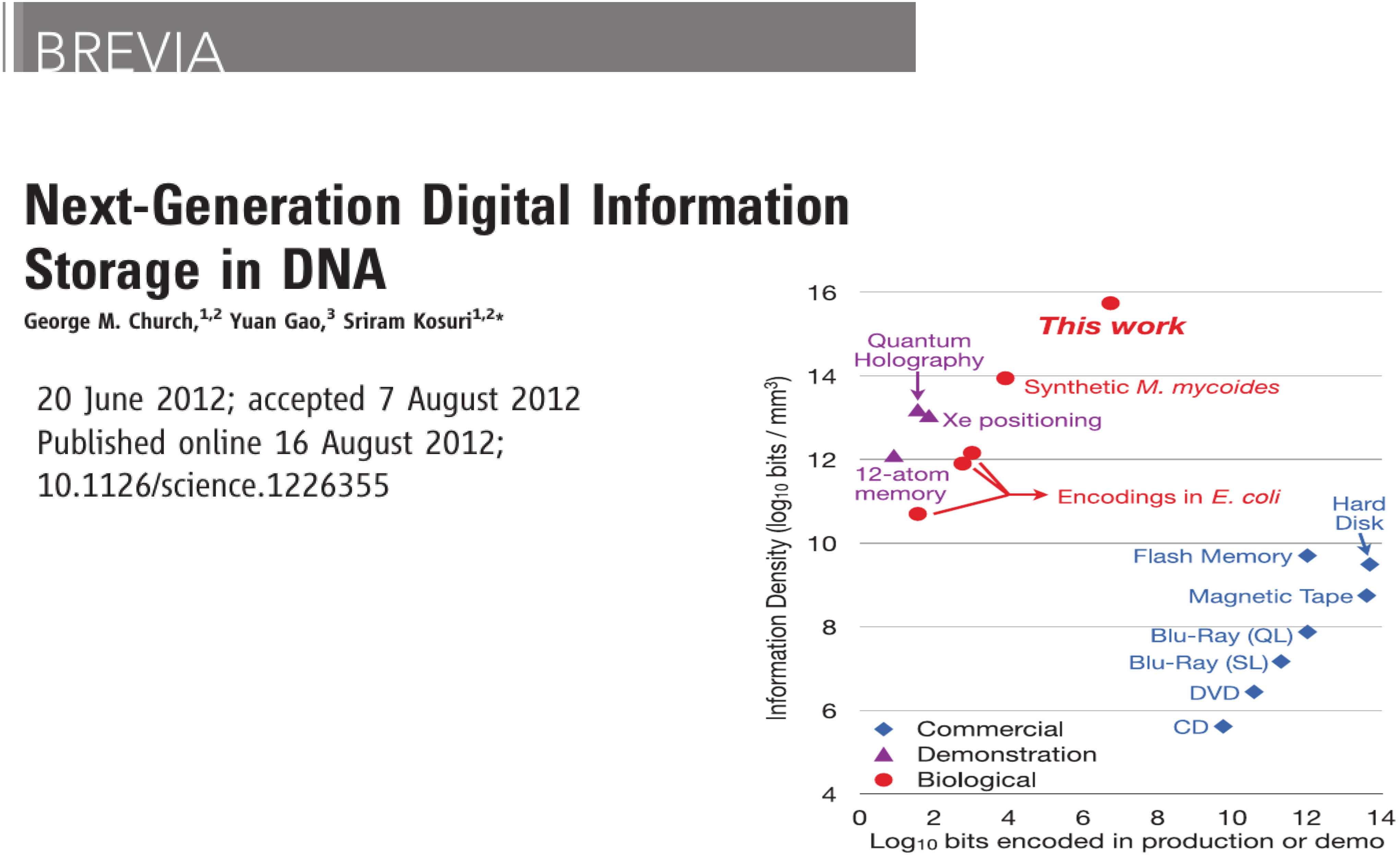
MOTION PERCEPTION ILLUSIONS AND THE BRAIN

CLINICAL TRIALS IN THE PRECISION MEDICINE ERA

PLUS

RNA-SEQ VERSUS MICROARRAYS

But they were not first to publish



BREVIA

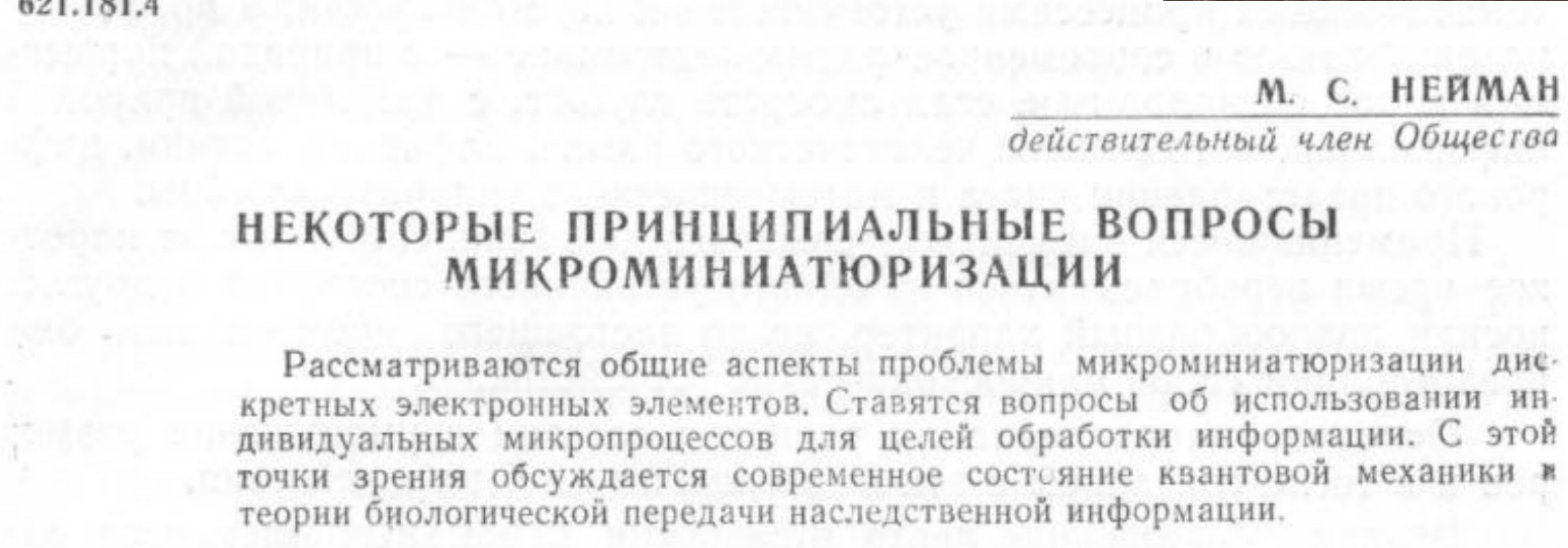
Next-Generation Digital Information Storage in DNA

George M. Church,[1,2] Yuan Gao,[3] Sriram Kosuri[1,2*]

20 June 2012; accepted 7 August 2012
Published online 16 August 2012;
10.1126/science.1226355

This is all petty dispute about months...

Радиотехника, т. 19, № 1, 1964 г.



621.181.4

М. С. НЕЙМАН

действительный член Общества

НЕКОТОРЫЕ ПРИНЦИПИАЛЬНЫЕ ВОПРОСЫ МИКРОМИНИАТЮРИЗАЦИИ

Рассматриваются общие аспекты проблемы микроминиатюризации дискретных электронных элементов. Ставятся вопросы об использовании индивидуальных микропроцессов для целей обработки информации. С этой точки зрения обсуждается современное состояние квантовой механики и теории биологической передачи наследственной информации.