Transcriptomania

Bud Mishra

Courant

NYU

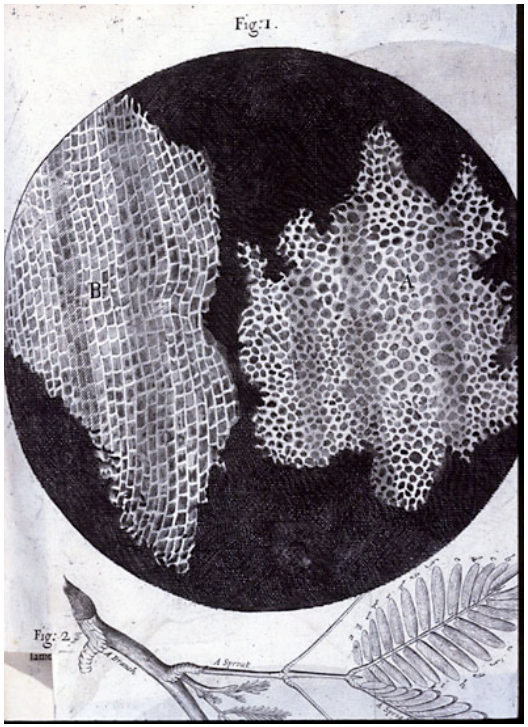
Robert Hooke (1635-1703) was an **experimental scientist**, **mathematician**, **architect**, and **astronomer**. Secretary of the Royal Society from 1677 to 1682, …

"**England's Da Vinci**" because of his wide range of interests.

His work **Micrographia** of 1665 contained his microscopical investigations, which included the first identification of biological cells.

Micrographia

“The Brain & the Fancy”



- “The truth is, the science of Nature has already been too long made only a work of the brain and the fancy. It is now high time that it should return to the plainness and soundness of observations on material and obvious things.”
  - Robert Hooke. (1635 - 1703), *Micrographia* 1665

Truth
Glimpsed or Demonstrated

The great distance between them…

In his drafts of Book II, Newton had referred to Hooke as the most illustrious Hooke— "**Cl[arissimus] Hookius**."

Hooke became involved in a dispute with Isaac Newton over the priority of the discovery of the inverse square law of gravitation.

- “[Huygen’s Preface] is concerning those properties of gravity which I myself first discovered and showed to this Society and years since, which of late Mr. Newton has done me the favour to print and publish as his own inventions.“
  - Hooke to Halley

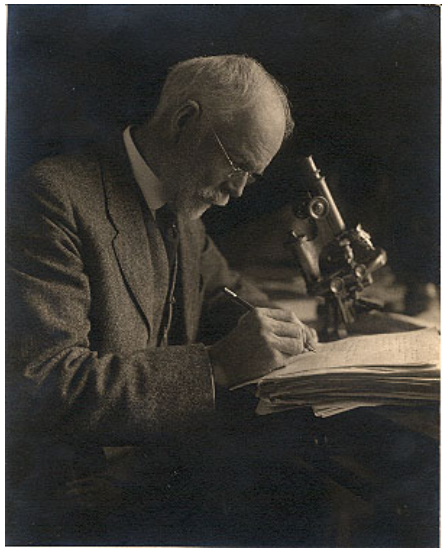
- "Now is this not very fine? Mathematicians that find out, settle & do all the business must content themselves with being nothing but dry calculators & drudges & another that does nothing but pretend & grasp at all things must carry away all the inventions…
- "**I beleive you would think him a man of a strange unsociable temper.**"
  - Newton to Halley

- “If I have seen further than other men, it is because I have stood on the shoulders of giants and **you my dear Hooke, have not**."
  – Newton to Hooke

- The great distance between
  - a glimpsed truth and
  - a demonstrated truth
    - Christopher Wren/Alexis Claude Clairaut

“Axioms of Platitudes”

-E.B. Wilson



1. Science need not be mathematical.
2. Simply because a subject is mathematical it need not therefore be scientific.
3. Empirical curve fitting may be without other than classificatory significance.
4. Growth of an individual should not be confused with the growth of an aggregate (or average) of individuals.
5. Different aspects of the individual, or of the average, may have different types of growth curves.

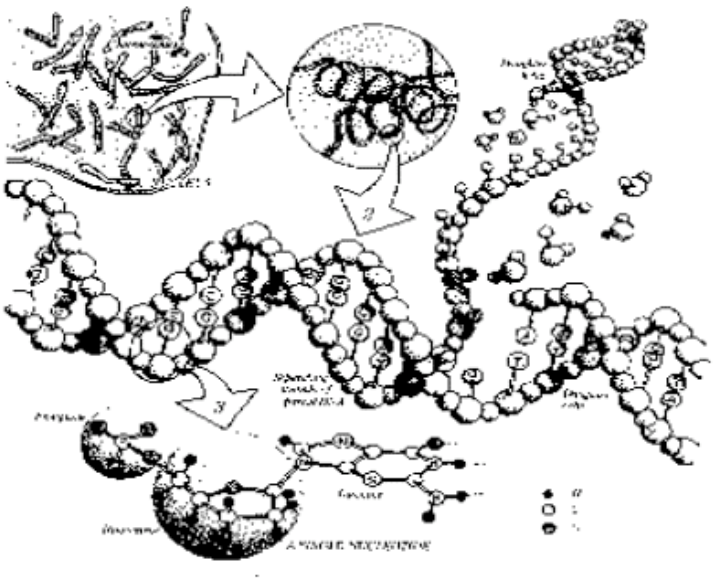
“The Brain & the Fancy”

“Work on the mathematics of growth as opposed to the statistical description and comparison of growth, seems to me to have developed along two equally unprofitable lines… It is futile to conjure up in the imagination a system of differential equations for the purpose of accounting for facts which are not only very complex, but largely unknown,… What we require at the present time is **more measurement and less theory**.”

– Eric Ponder, Director, CSHL (LIBA), 1936-1941.

More Measurement & Less Theory

What can be measured?

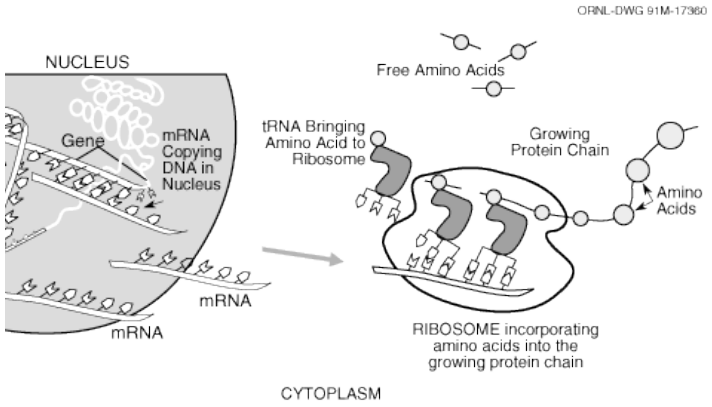


Genome: All the hereditary information of an organism, encoded in its DNA

Very long sequence of **nucleotides** or **bases**:

$\Sigma = \{A, T, C, G\}$

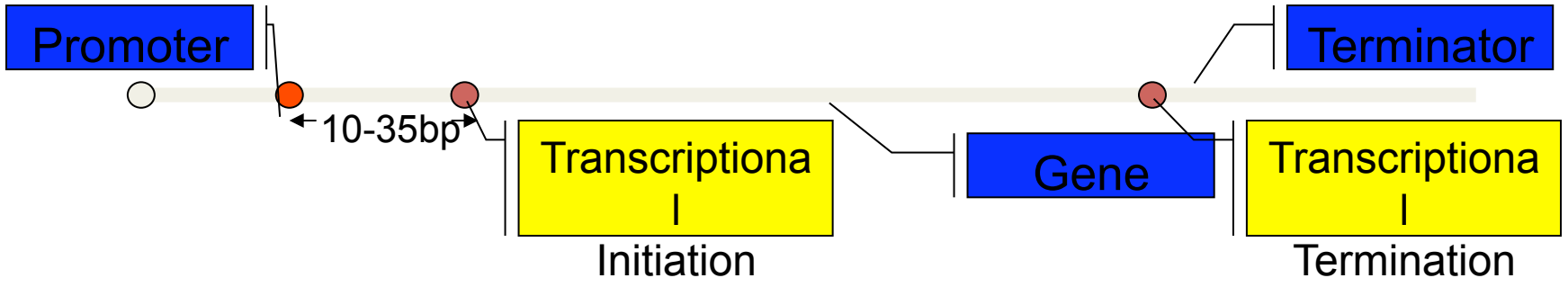
The Central Dogma (due to Francis Crick in 1958)



"The central dogma states that once `information' has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein."

DNA Transcription **RNA** Translation **Protein**

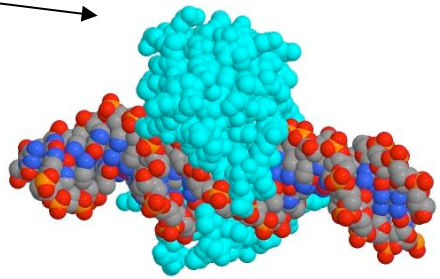
- A specific region of DNA that determines the synthesis of proteins (through the transcription and translation) is called a **gene**
- Transcription of a gene to a **messenger RNA**, **mRNA**, is keyed by a transcriptional activator/factor, which attaches to a **promoter** (a specific sequence adjacent to the gene).



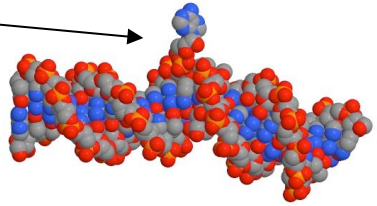
Measurements

- Cellular State:
  - Genome
  - Epigenome
  - **Transcritome**
  - Proteome
  - Metabolome
- Single-Cell & Single Molecules
  - Focus on RNA (dynamic & highly variable; yet quantifiable)

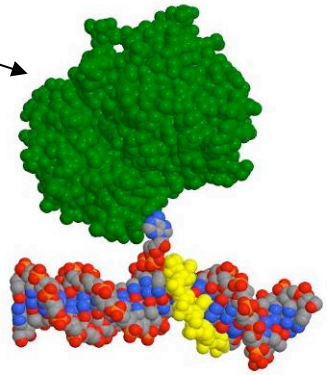
Nicking endonuclease

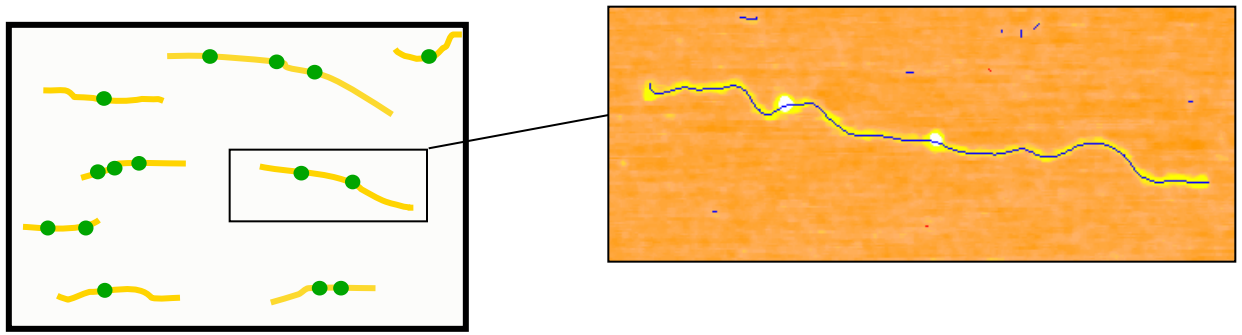


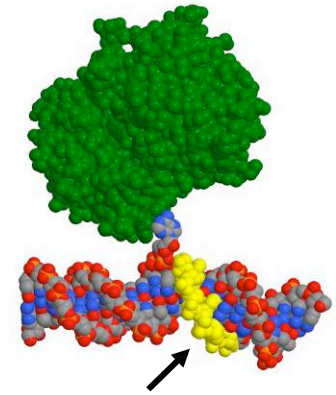
Biotin label



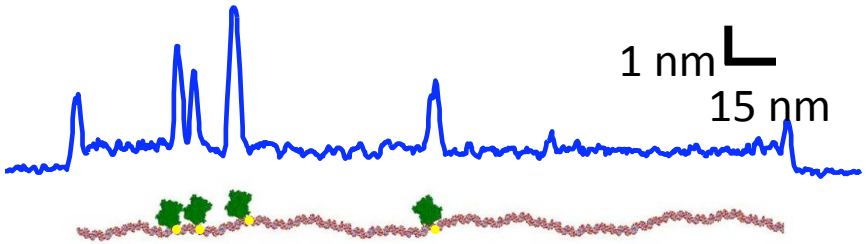
Streptavidin tag

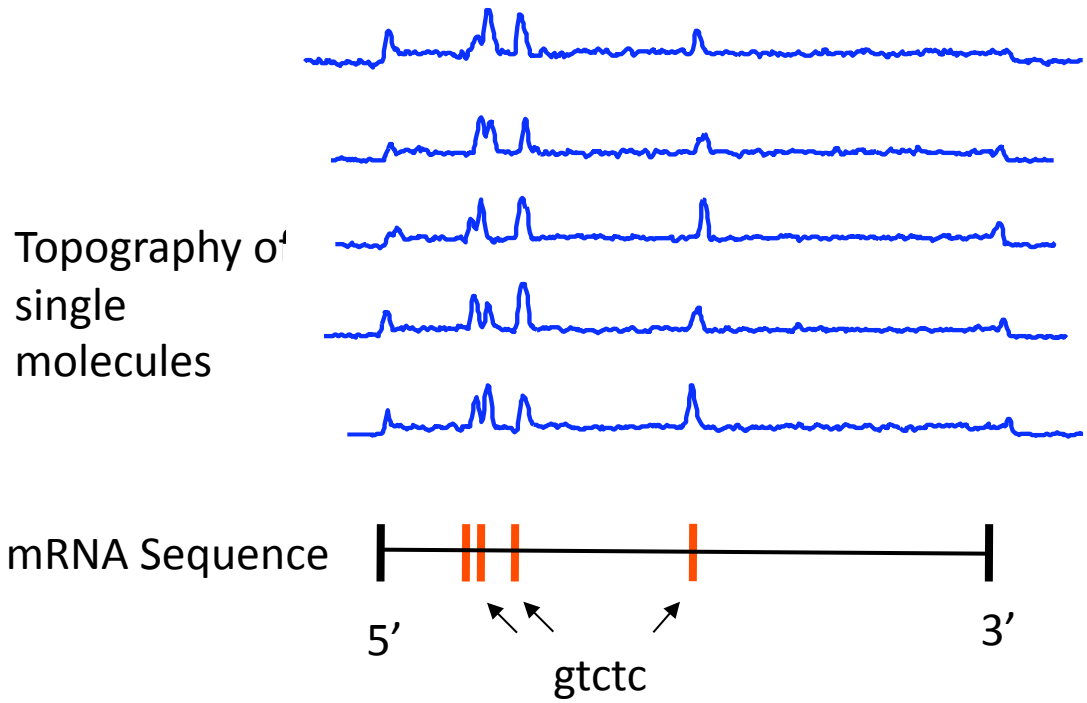


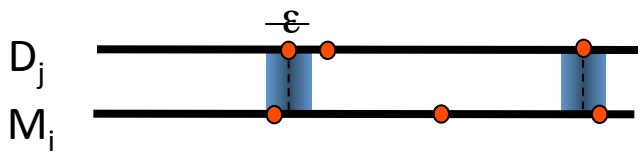




…gtctc..




Topography o
single
molecules
mRNA Sequence
5’
3’
gtctc



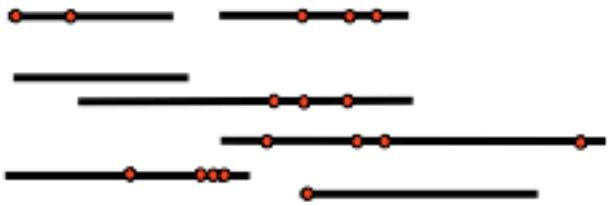
p(M | D,I) = p(length) x p(alignment) x p(true label) x p(false label)

p(length) ~ normal

p(alignment) ~ normal

p(true label) ~ binomial

p(false label) ~ Poisson



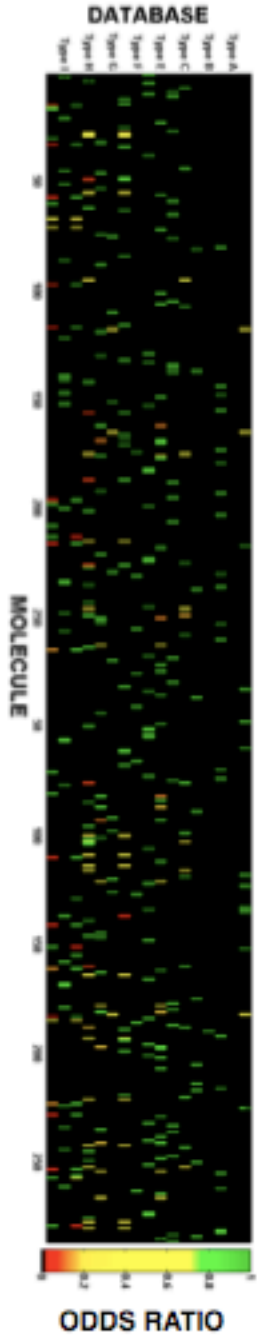
Unknown Molecules (M_i)

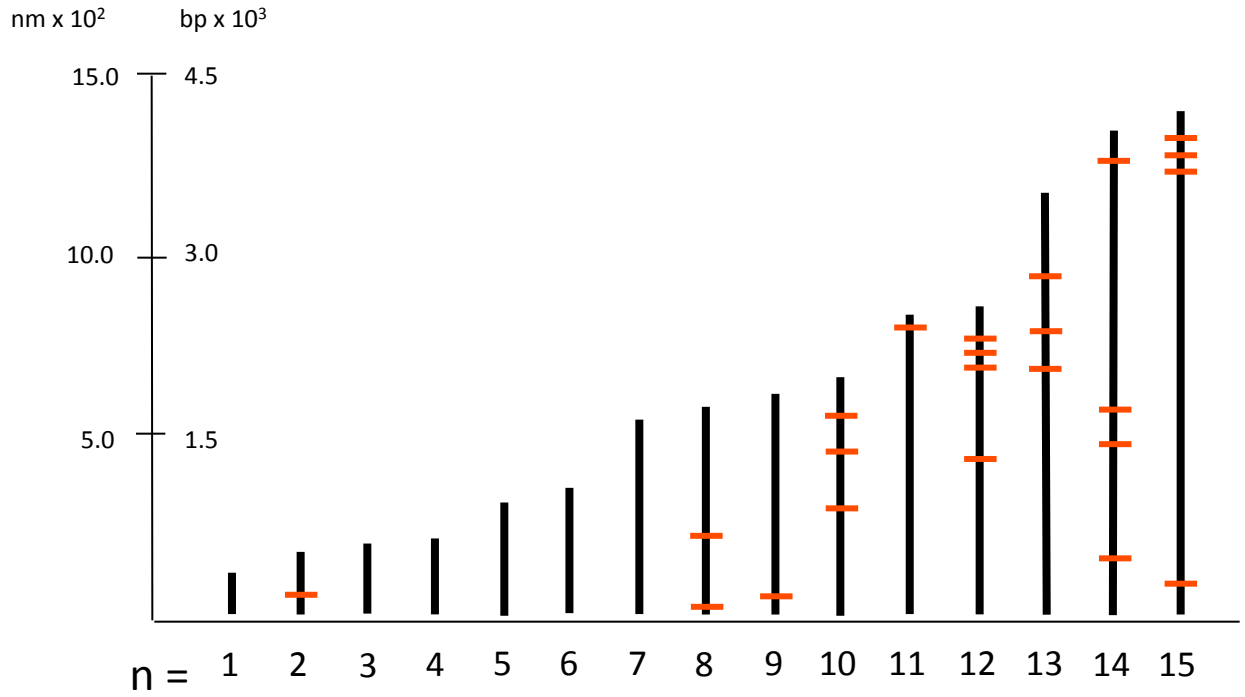
Bayes' Theorem $$p(D \mid M,I) = \frac{p(D \mid I)\, p(M \mid D,I)}{p(D \mid I)}$$

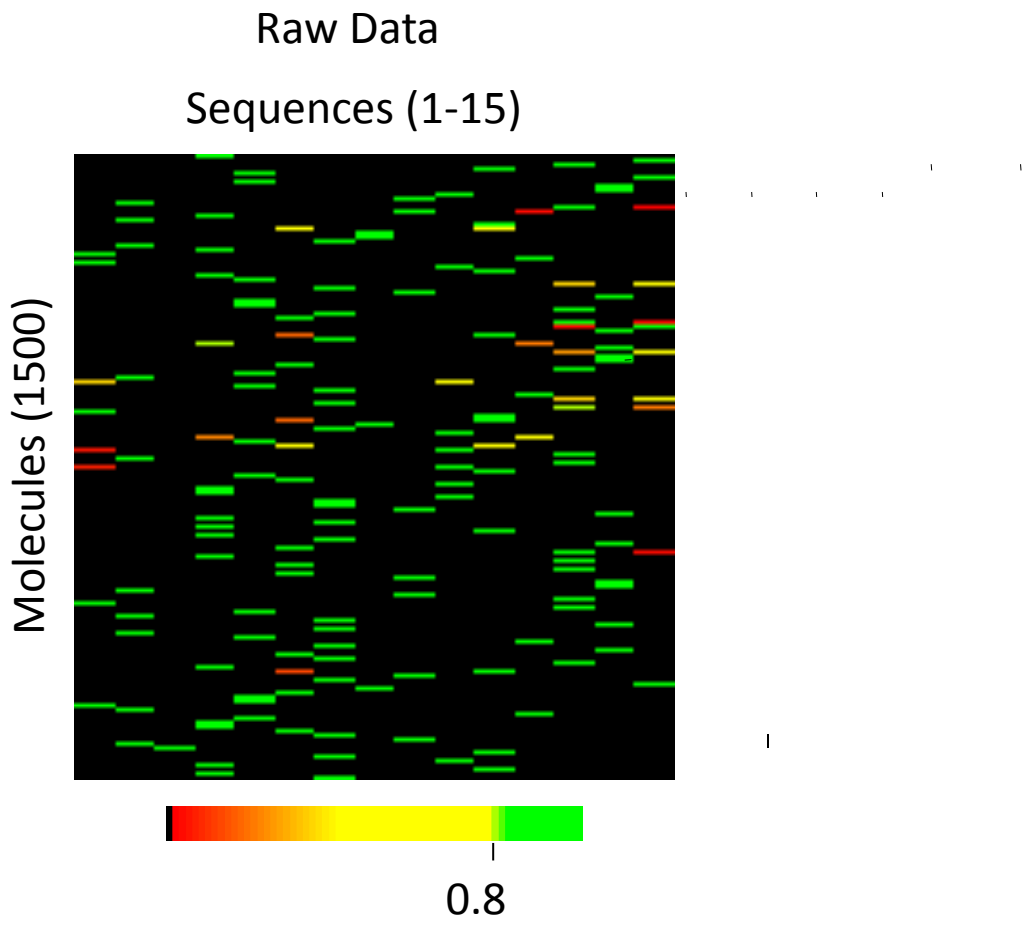
Odds Ratio $$O_{ijk} = \frac{p(D_j \mid I)\, p(M_i \mid D_j,I)}{p(D_k \mid I)\, p(M_i \mid D_k,I)}$$

transcript 1: *caatattccgtctctccgtacttcccagagtctcgcttc*
transcript 2: *ttatcttatatcggca aatgtctcctccaactctg….*
transcript 3: *ctcgtctcaactgataaaatgtctccttcccagcc….*
transcript 4*:* *atatcggcaatagtctctcggcaatatcggcaaatatc…*
……

Sequence Database (D_j)




nm x 10²
bp x 10³
15.0
4.5
10.0
3.0
5.0
1.5
n =
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


Raw Data
Sequences (1-15)
Molecules (1500)
0.8

Clustered Data
Sequences (1-15)
Molecules (1500)
0.8
1
0.8
0.6
0.4
0.2
0

In this analysis, we treat each map as a unique 'molecular signature.' The first step in determining this probability is to calculate the Hamming distance between molecular signatures, HamDist, assuming a total number of 'good signatures', *S.*

Each signature is randomly selected from the set of all possible binary vectors, with a probability π. The computation of this probability proceeds as follows: start with a selected signature *f0* from the set *S,* and compute all the possible signatures whose Hamming distances from *f0 range between 1 and HamDist; there are*

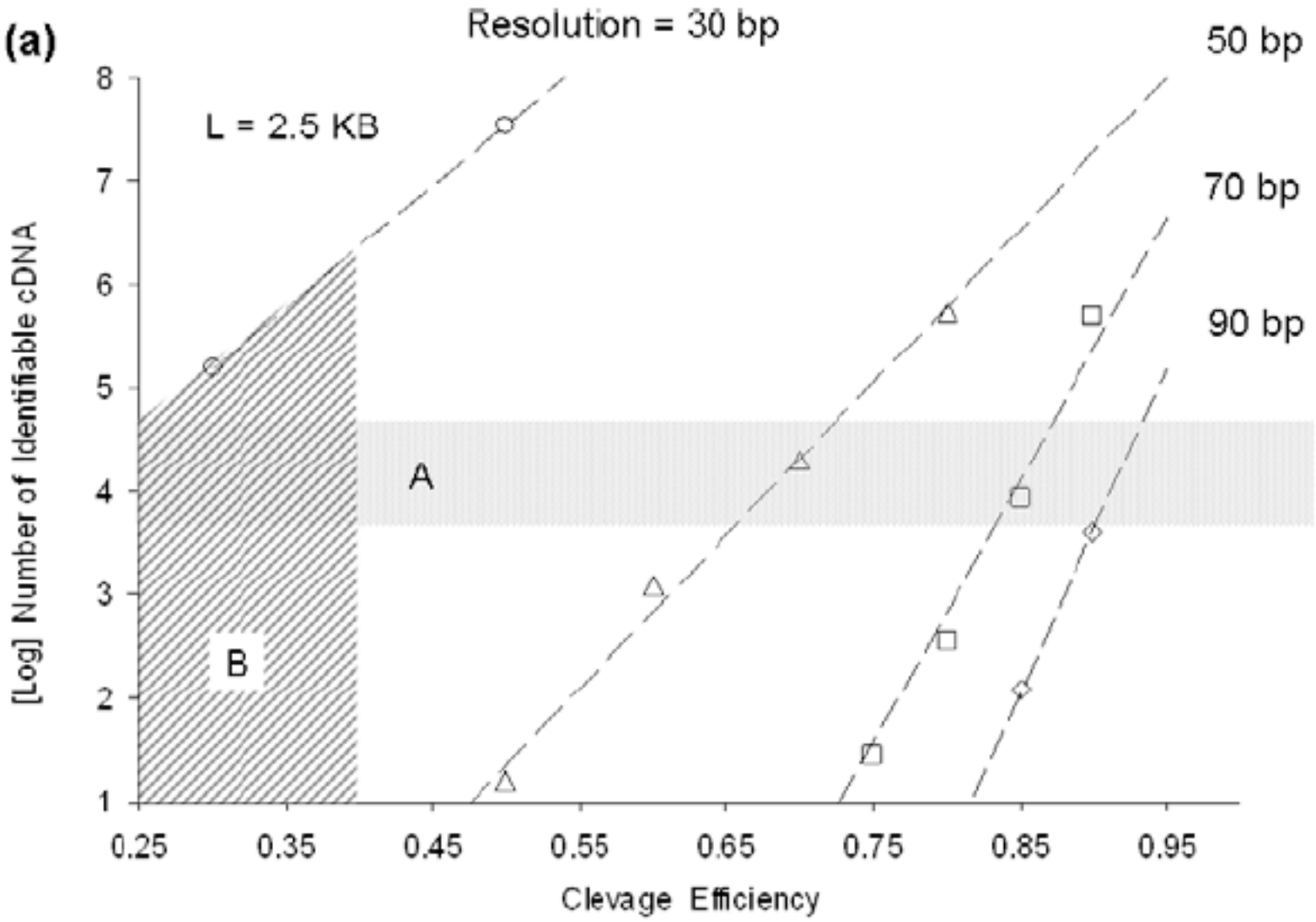
$$\sum_{k=0}^{\text{HamDist}-1} \text{Binomial}[M, k]$$

such signatures, and with high probability, they do not contain even a single signature from the set *S (*probability *> (1 −10^{−12}) > (1 − π)^vol)*

We compute the uniqueness of the identification probability, given a fixed sizing accuracy, *α,* enzyme recognition site frequency, *p_c,* and cleavage rate, *p_d:* we compute this probability as follows

$$\sum_{b=0}^{\text{Floor(HamDist/2)}} \sum_{a=0}^{M-b} \text{Multinomial}[a, b, M-a-b](\alpha\ p_{\text{c}}\ p_{\text{d}})^{a} \times (\alpha\ p_{\text{c}}\ (1-p_{\text{d}}))^{b}(1-\alpha\ p_{\text{c}})^{(M-a-b)}. \quad (2)$$

That is, we sum the probabilities that starting with a signature with *(a + b)* unit bits, exactly *b* unit bits are lost from the mapped signature as a consequence of incomplete cleavage


(a)
Resolution = 30 bp
50 bp
70 bp
90 bp
L = 2.5 KB
A
B
[Log] Number of Identifiable cDNA
Clevage Efficiency
1
2
3
4
5
6
7
8
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95

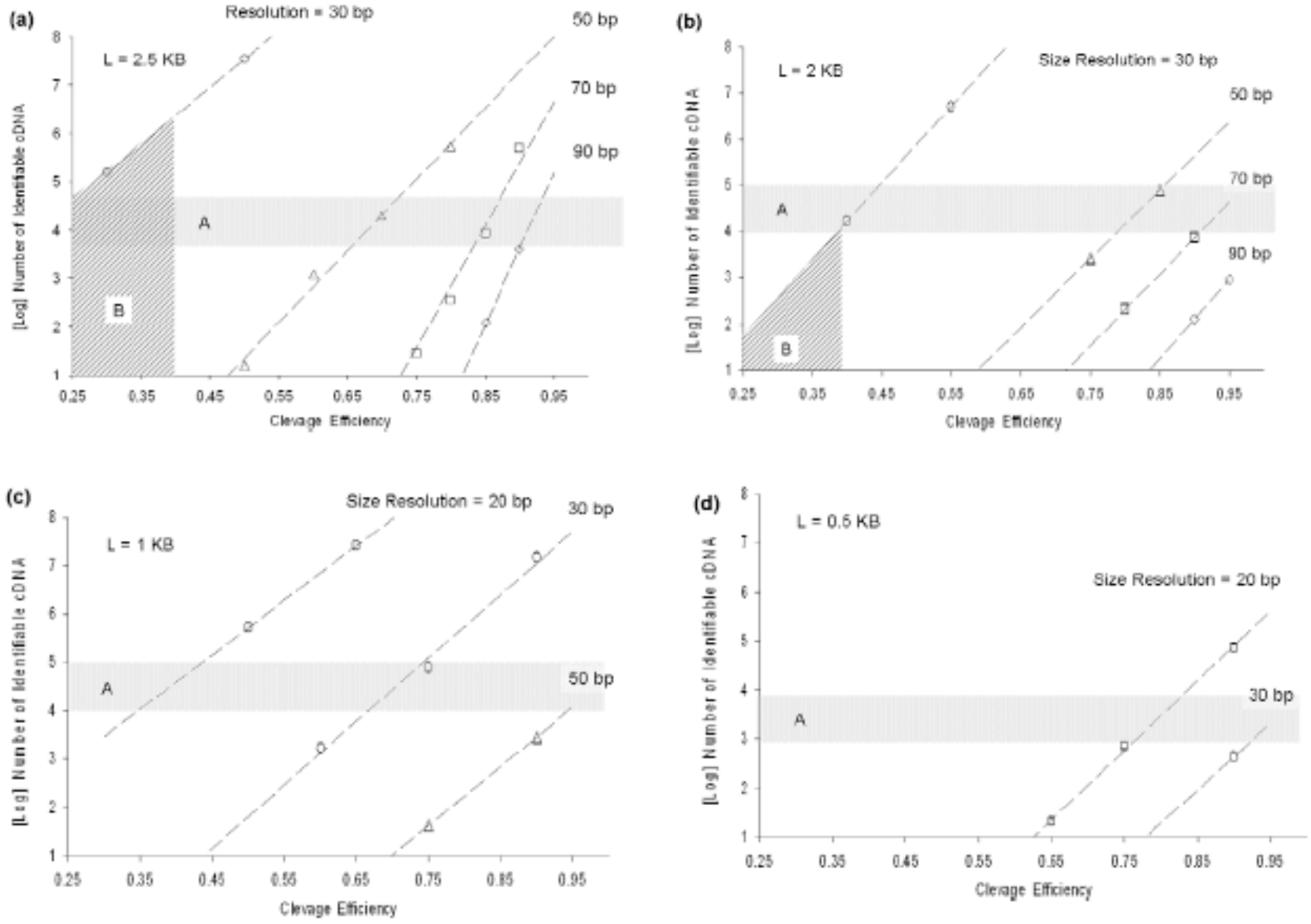
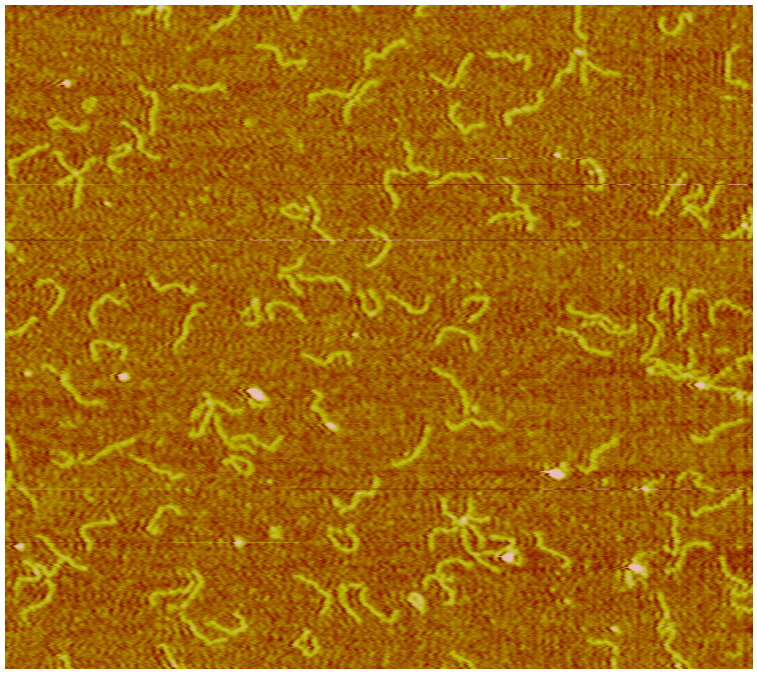
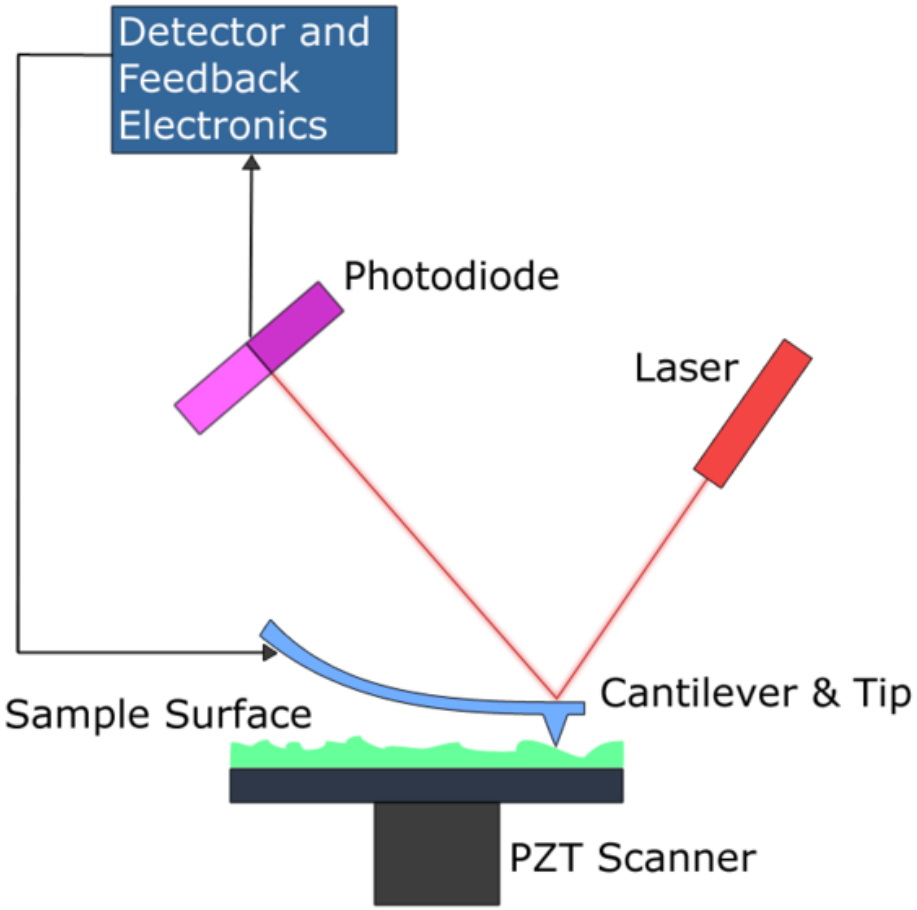


Figure 2. Computations of the number of unambiguously identifiable cDNA species (>95% probability) for a given bp sizing accuracy as a function of cleavage efficiency and cDNA size: 2.5 kb, (b) 2 kb, (c) 1 kb and (d) 0.5 kb. For cDNA length 2kb, as sizing resolution degrades from 50 to 90 bp, difficult-to-achieve cleavage efficiency (>80%) is needed to distinguish many species ($>10^4$). As sizing resolution approaches 30 bp, 10^4 to 10^6 species can be detected, even at very low cleavage rates (30%–50%). Region B indicates the parametric space accessible given the resolution (~30 bp) and cleavage efficiency demonstrated (~40%) in our experiments.

Analysis of Atomic Force Micrographs to Measure RNA and DNA Length with High Precision



A Brief Introduction to AFM

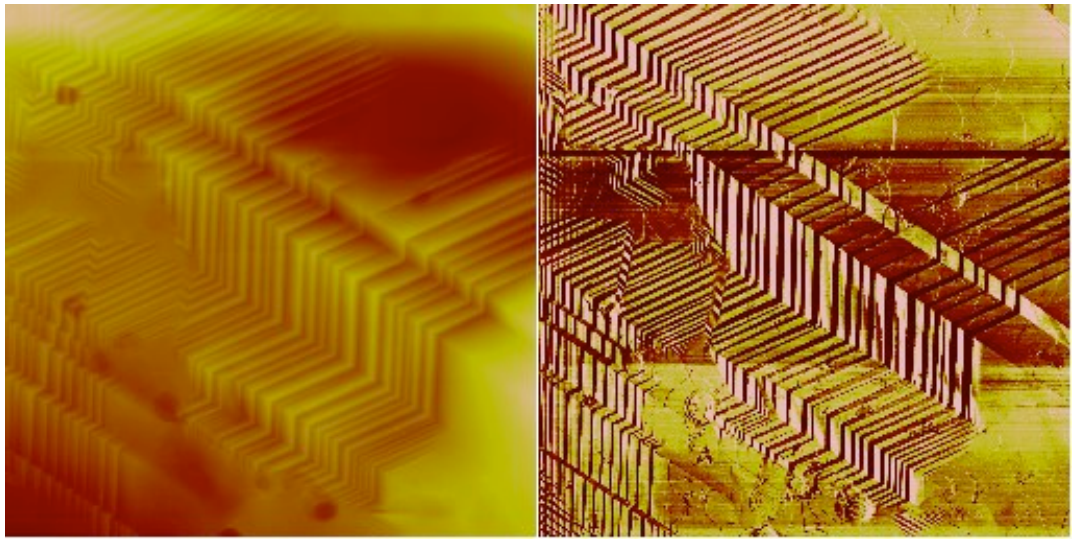


Two Basic Measurements from AFM

- At each point, (x,y), in an area, we can measure:
  - the displacement in the z-direction for height
  - the change in oscillation frequency for softness

Height map

Deflection map



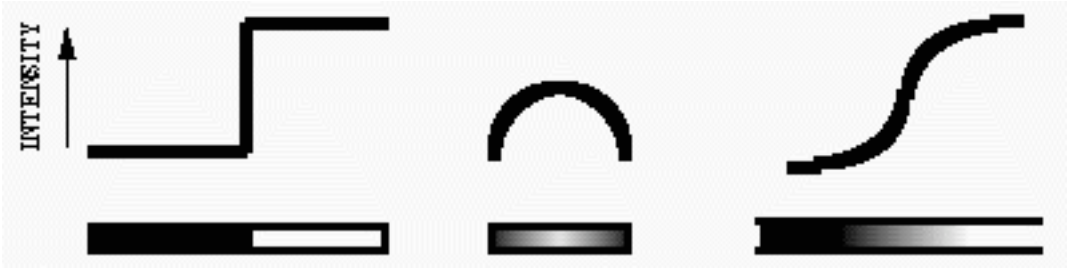
Two Intrinsic Problems with AFM

- Tip convolution effects, especially on the ends of molecules
- Thermal drift

Tip Convolution (continued)

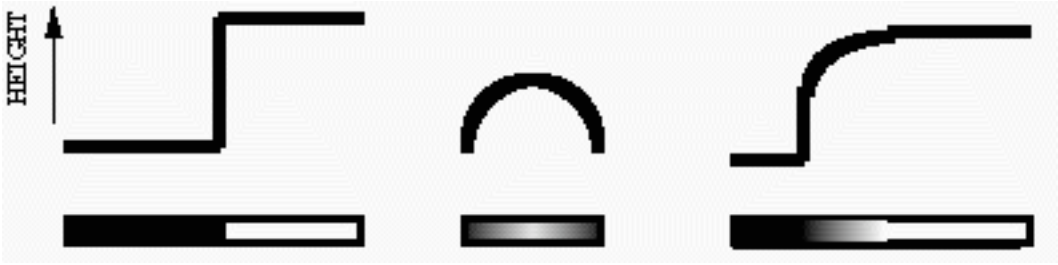
In an electron microscope:

this surface… …and this function… …give this image.



In an atomic force microscope:

this surface… …and this function… …give this image.

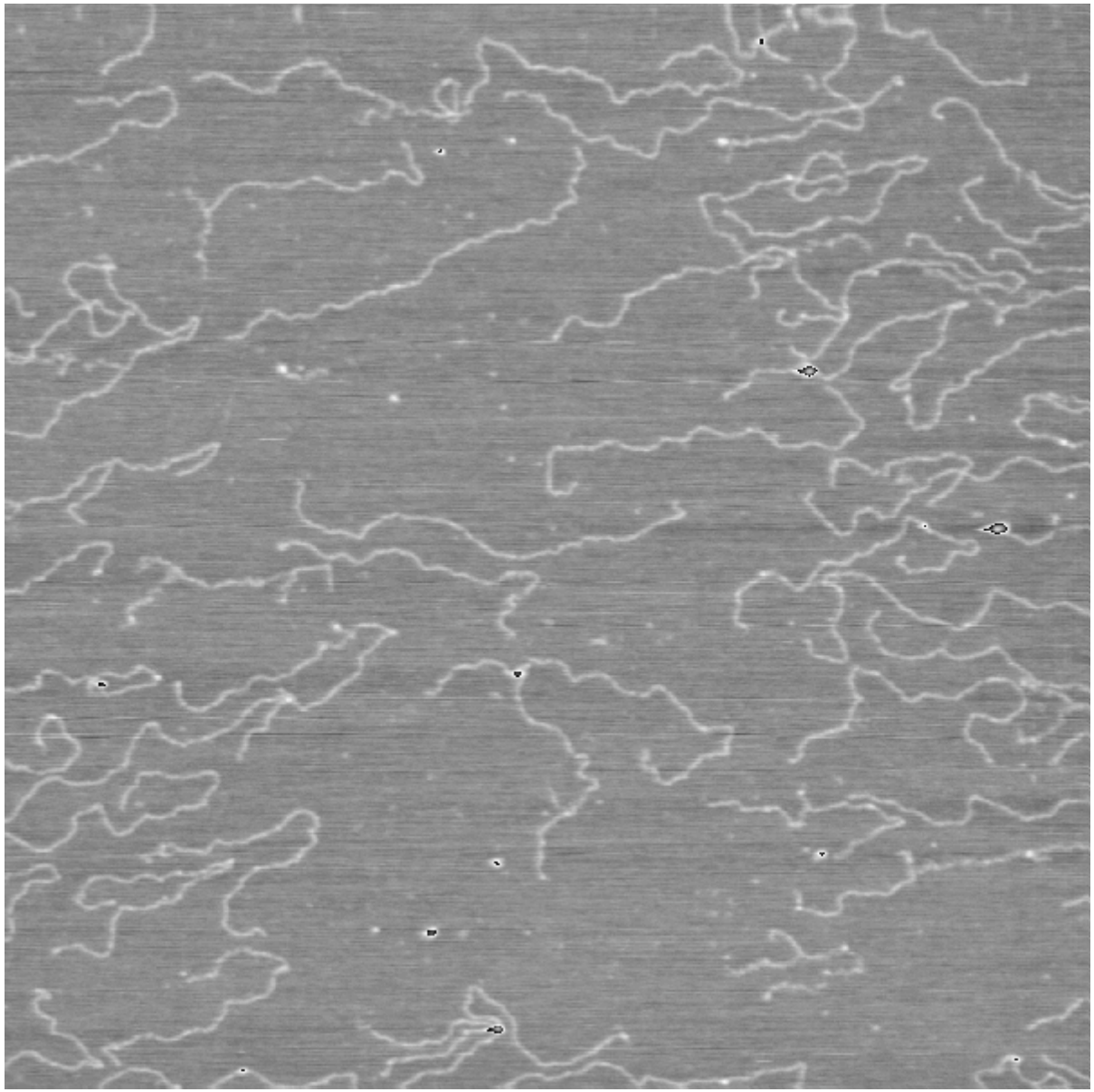


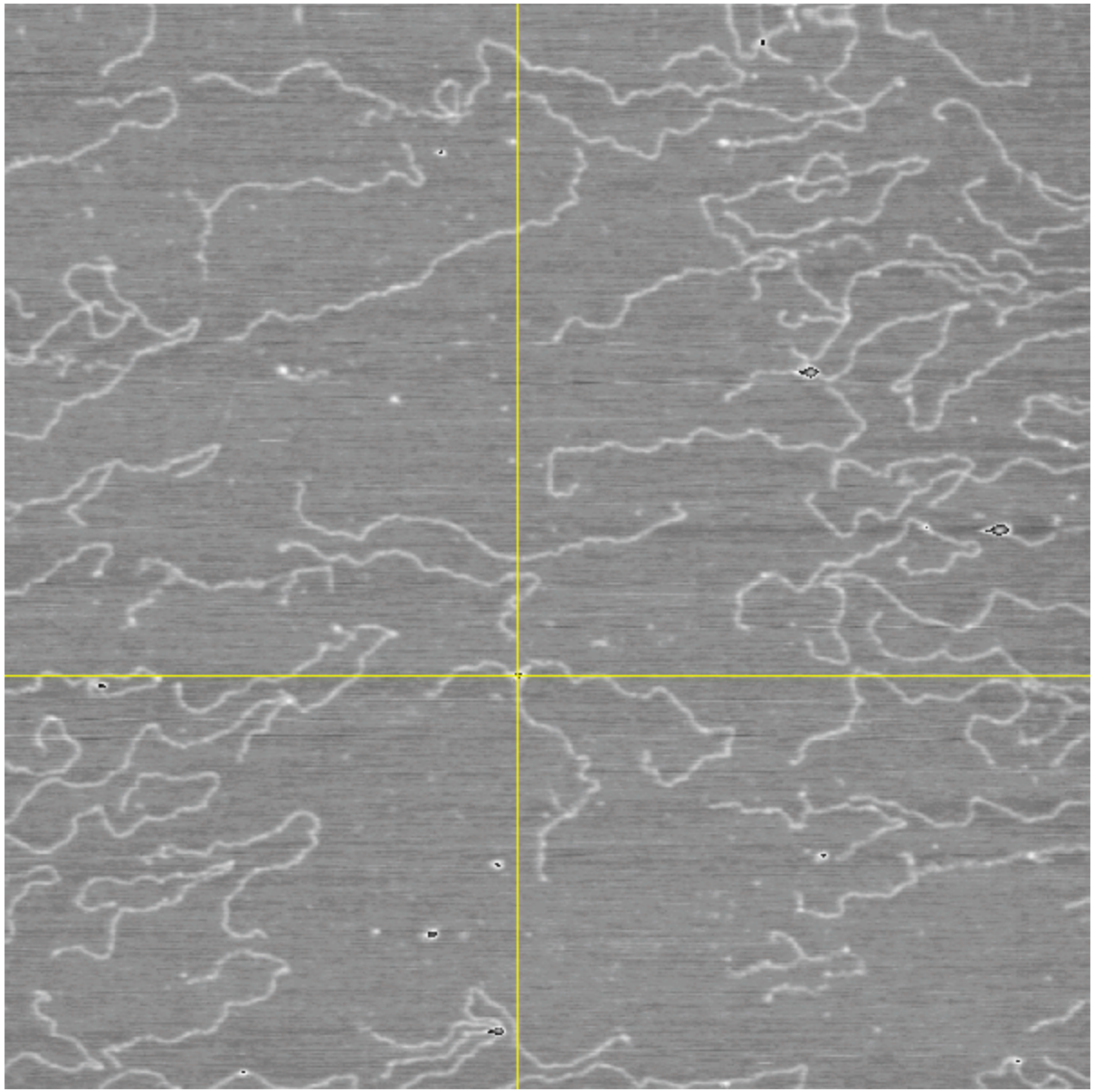
Problem: Thermal Drift

- Each component of the AFM (tip, cantilever arm, sample, stage, pezioelectronics) has its own coefficient of thermal expansion.
- Even minute fluctuations in ambient temperature lead to an aggregate displacement of the materials with respect to each other, hence drift.

An Example of Thermal Drift

- 8 images of pUC19 DNA plasmids taken in sequence
- The mean scan period is 33.7 min/img.
- Each image is 1408x1408 pixels (2x2 microns)
- The resolution is 1.42 nm/pix
- The scan rate is 0.001 sec/pix, 980 pix/sec, 1392 nm/sec
- The images were scanned in the top-to-bottom direction.





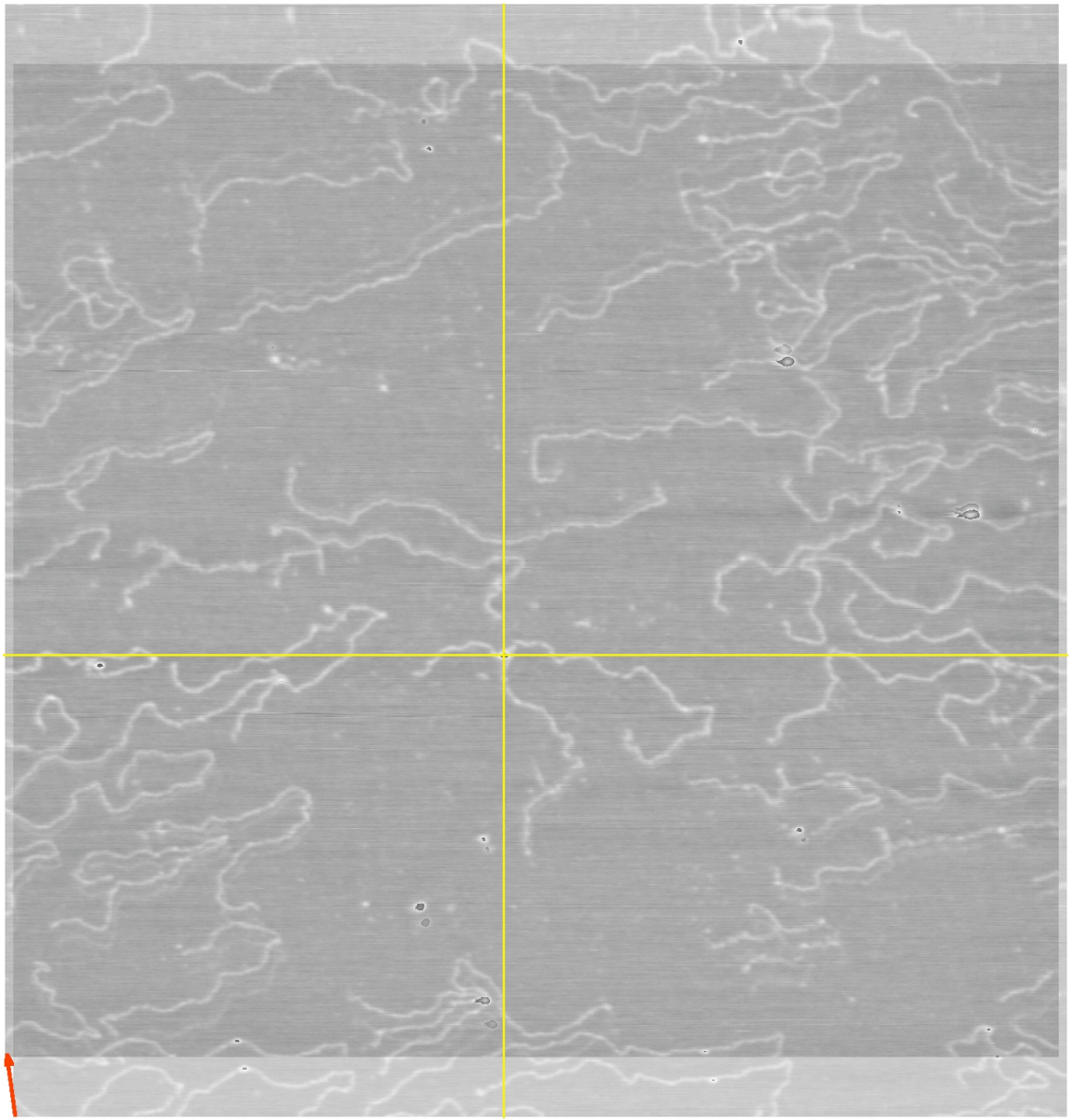


Figure 1: Alignment of the first two AFM top-to-bottom scans (at $t + 0$ min and $t + 34$ min), with displacement vector (red) added. The vector magnitude is 80 pixels, which, at 1.42 $\frac{nm}{pixel}$ resolution, gives a displacement of 113 nm, at a rate of 3.3 $\frac{nm}{min}$.

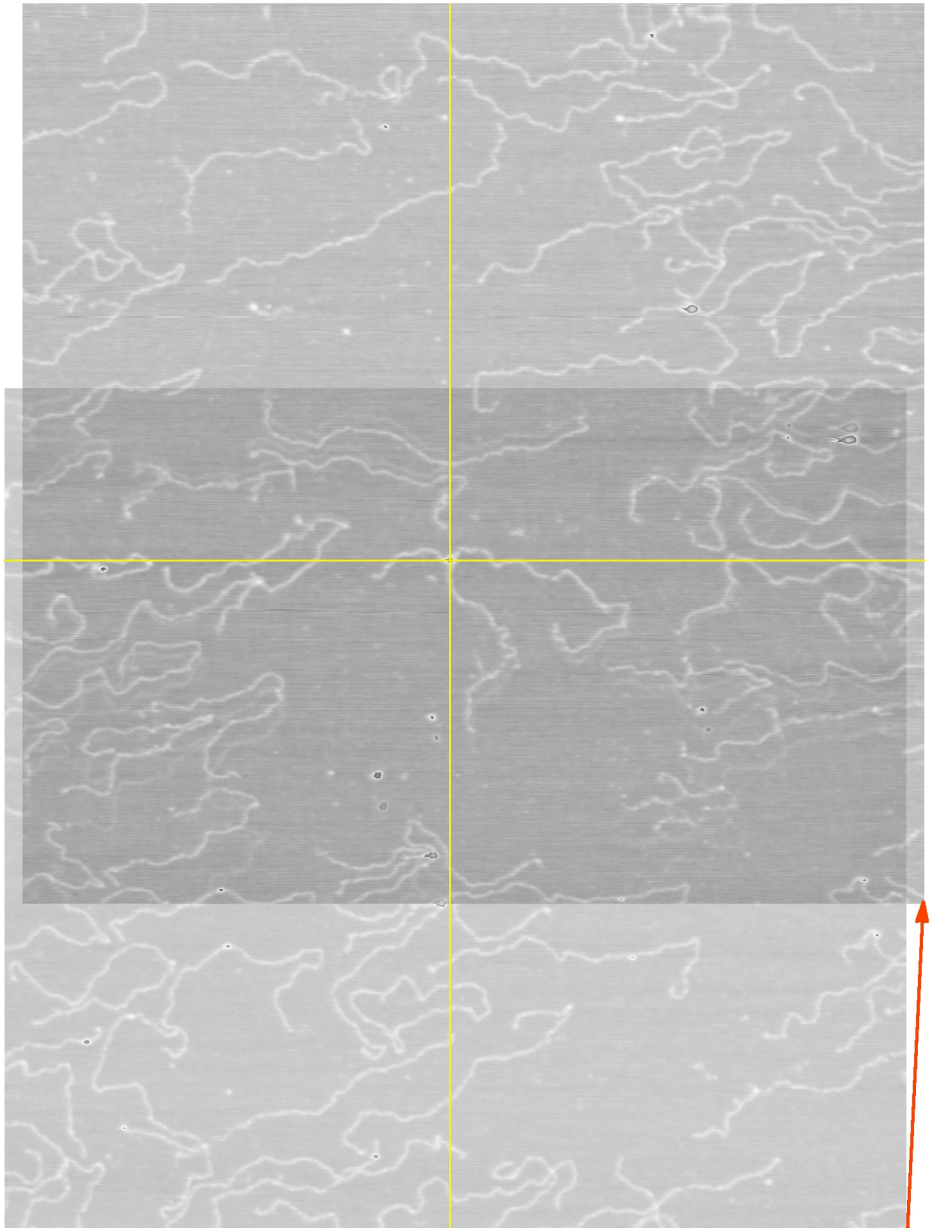
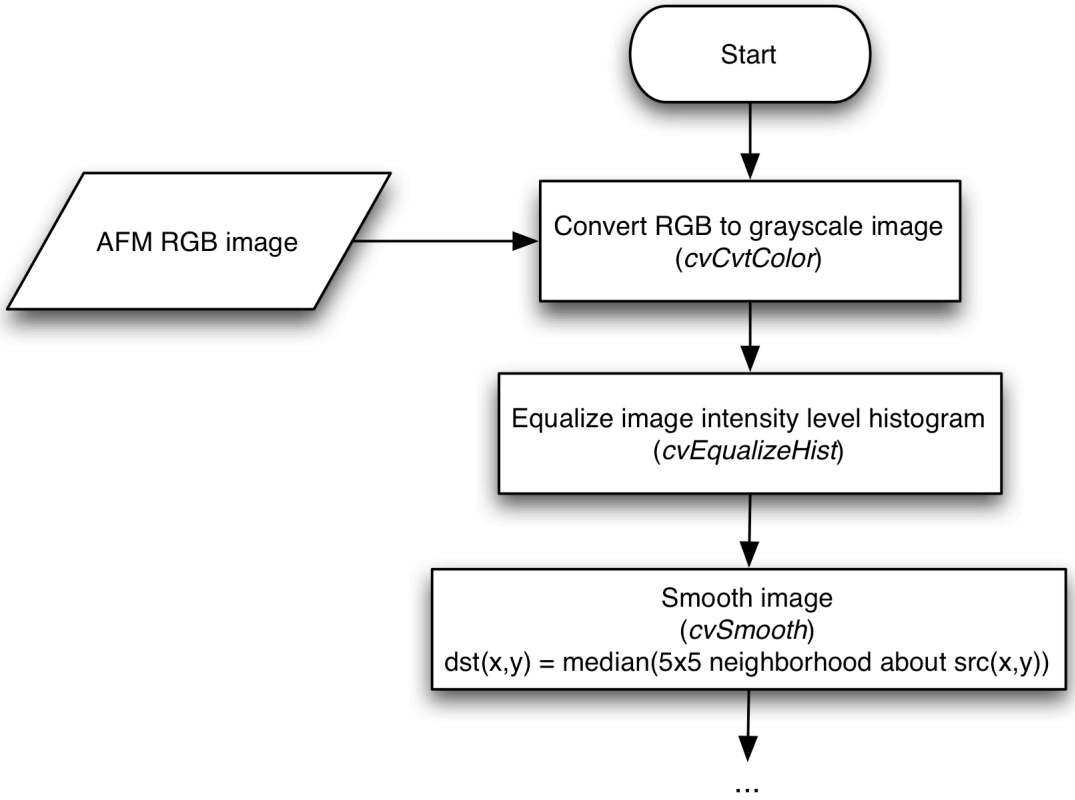


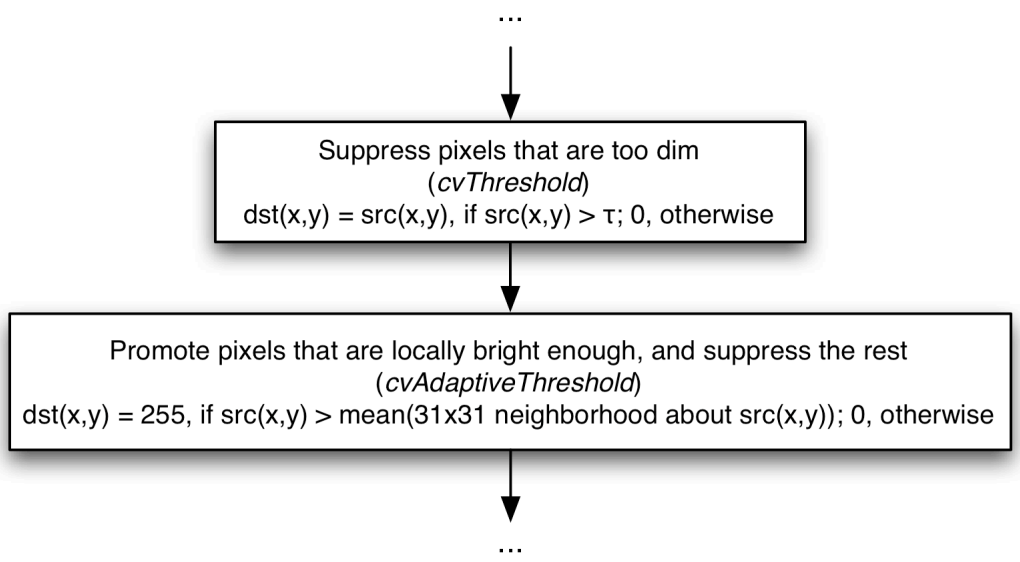
Figure 2: Alignment of the first and last AFM top-to-bottom scans (at $t+0$ min and $t+404$ min), with displacement vector (red) added. The vector magnitude is 593 pixels, which, at 1.42 $\frac{nm}{pixel}$ resolution, gives a displacement of 842 nm, at a rate of 2.1 $\frac{nm}{min}$, representing the average drift rate over the net displacement in 404 min.

Image Analysis

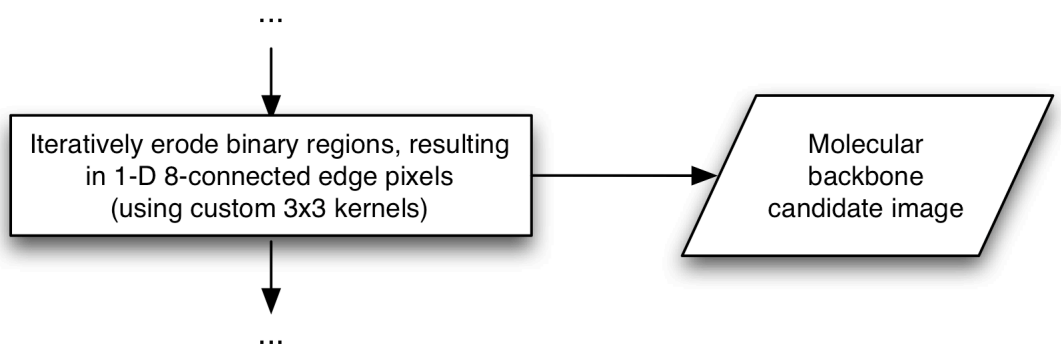
Set up the Image for Processing



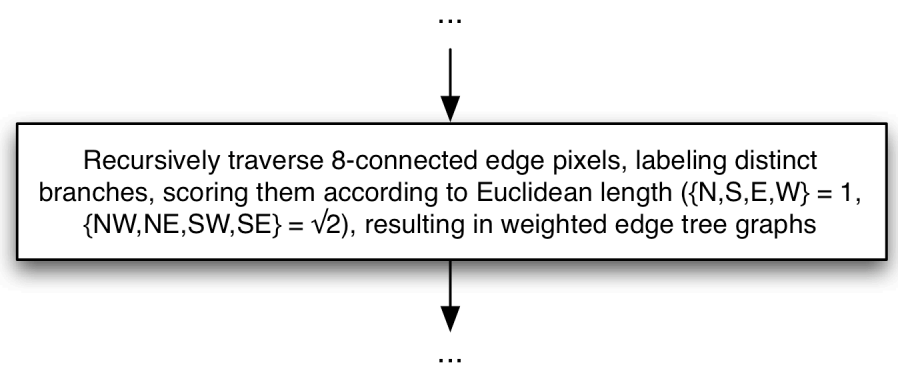
Extract Foreground from Background



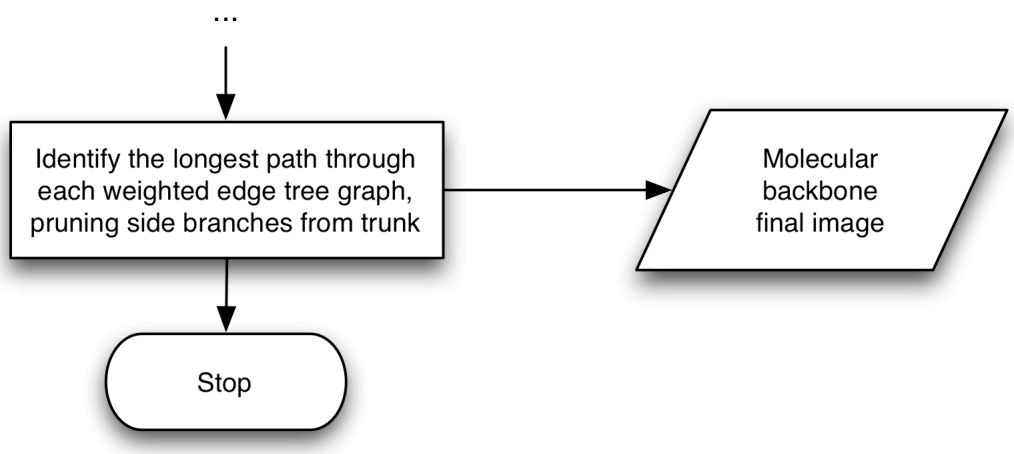
Reduce the Image to Its Essential Morphology



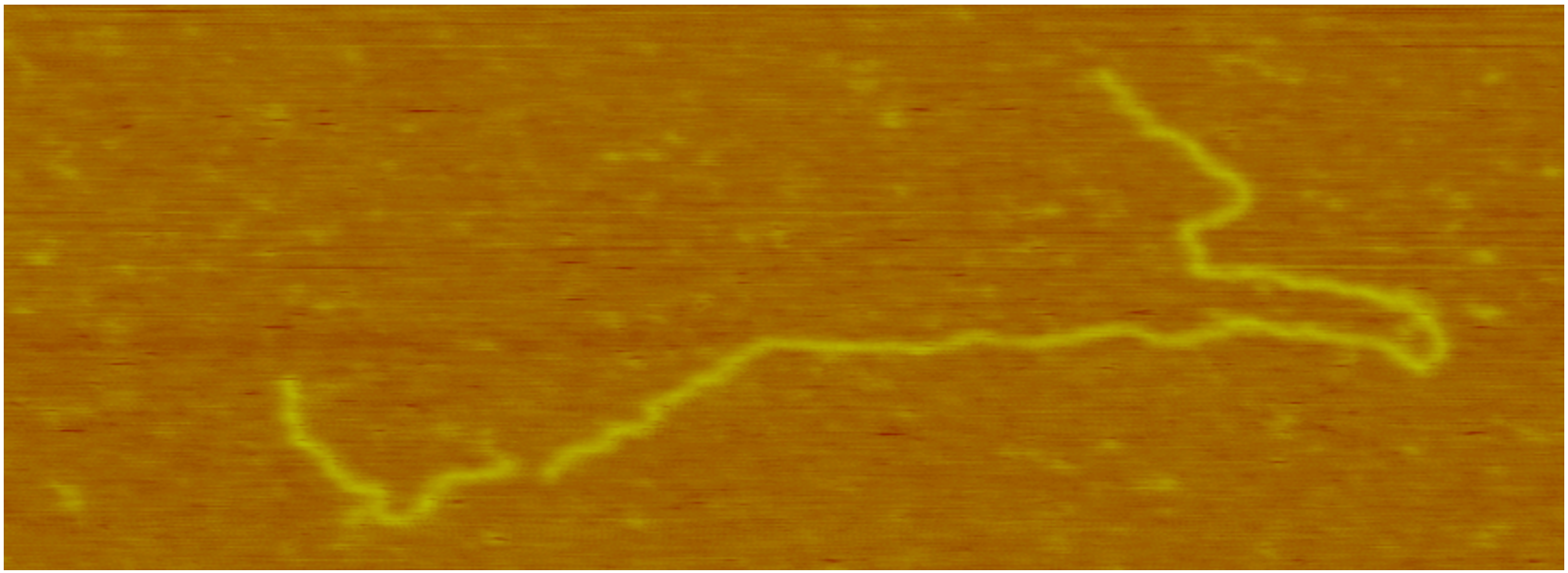
Transform the Morphological Features into a Graph



Identify the Longest Path through the Graph



Original AFM Image

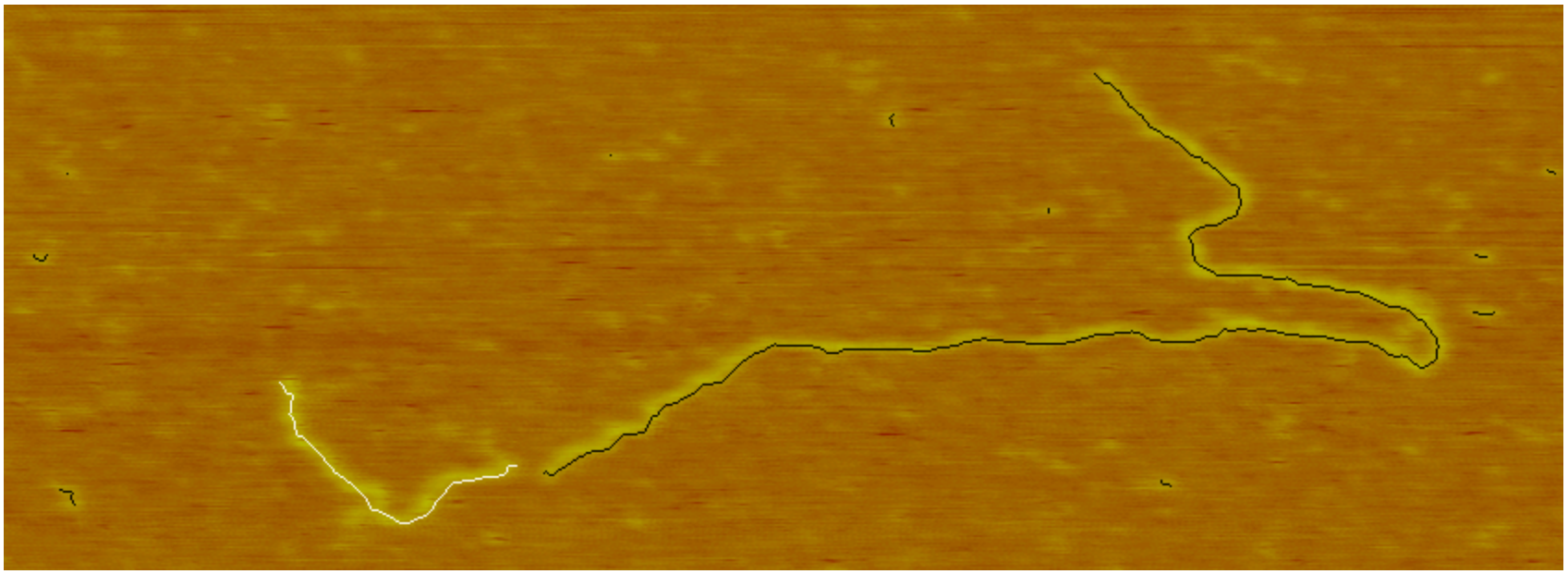


Filtered Image Showing 1-D Edge Trees

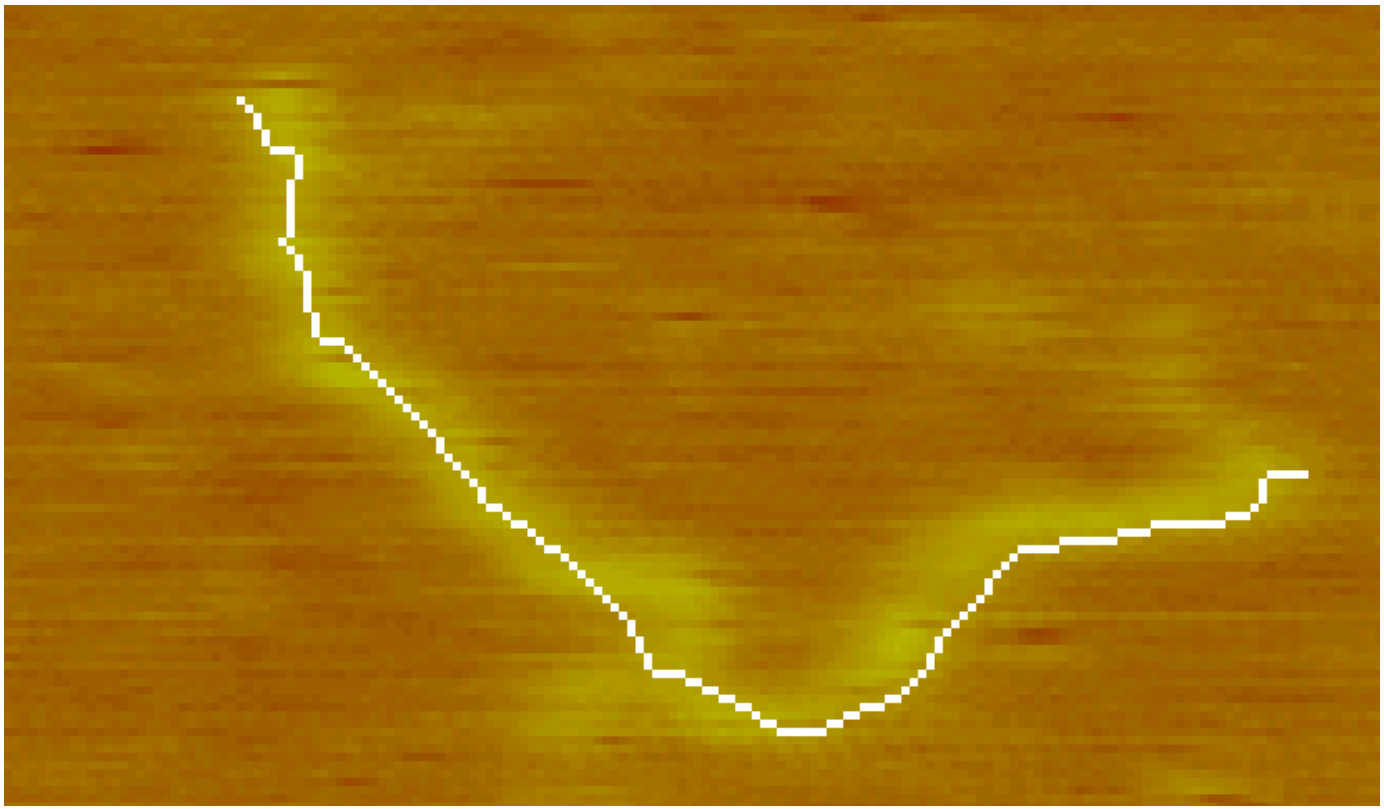
Final Backbone Image

The Backbone in Edge Tree Context

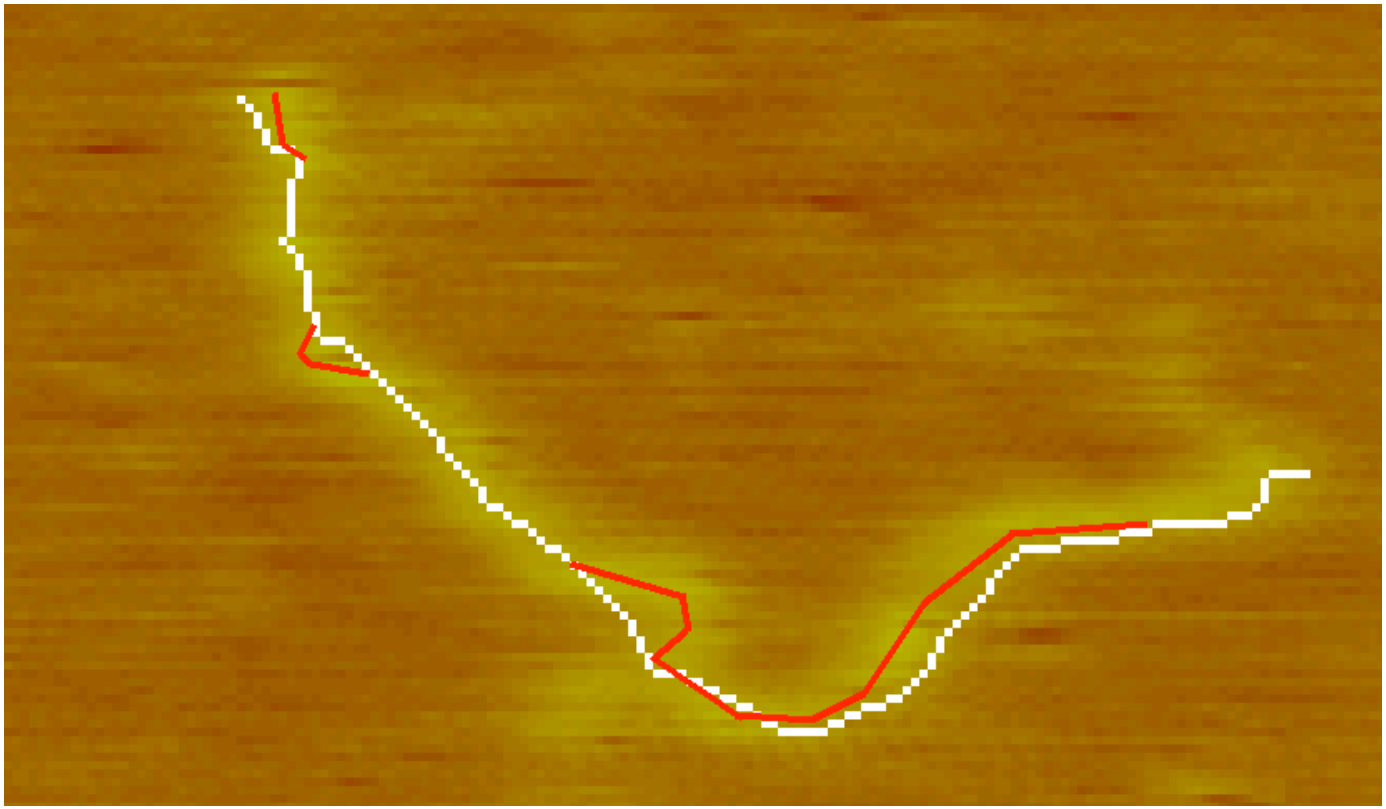
The Backbone in Original Context



Magnified

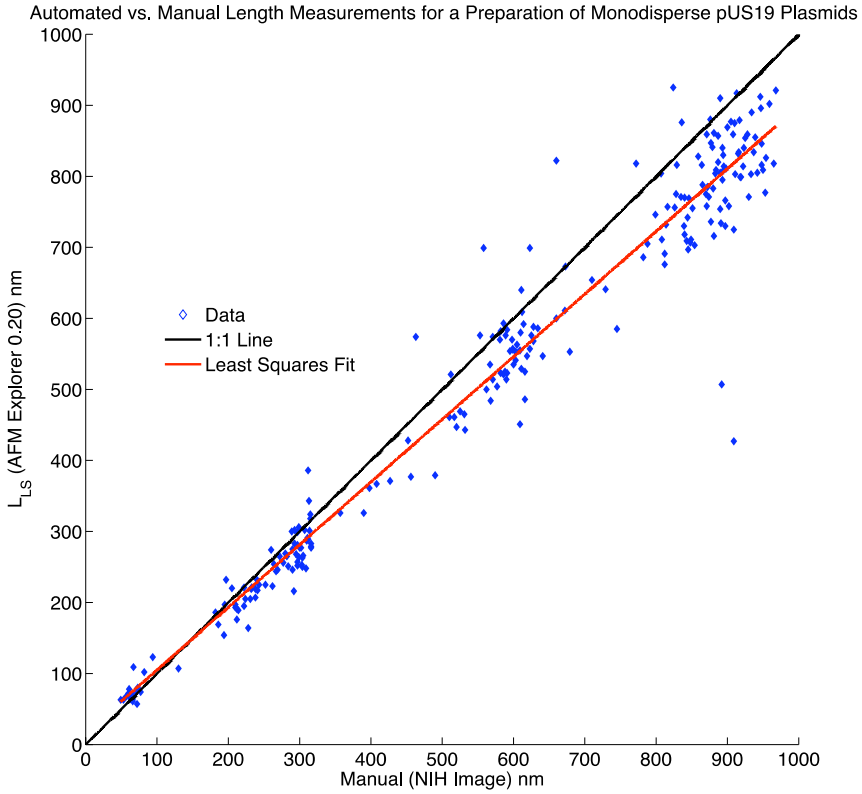


The Obvious Problem

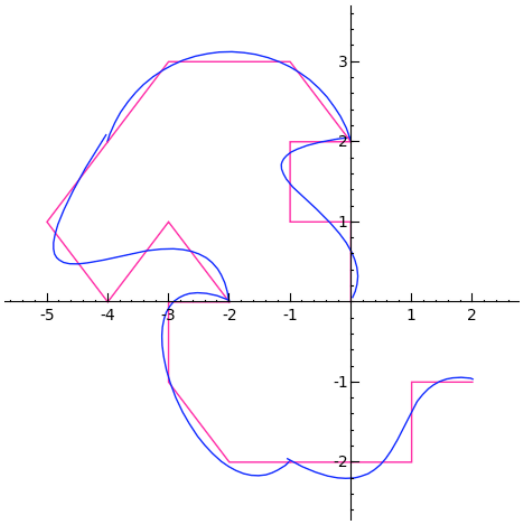


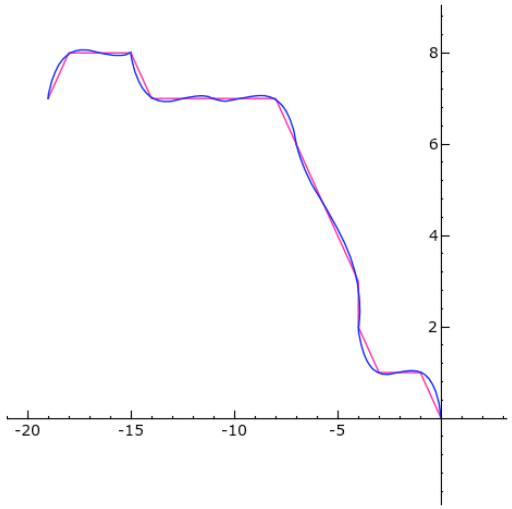
And Yet, Early Comparative Results Show the AFME Does Very Well

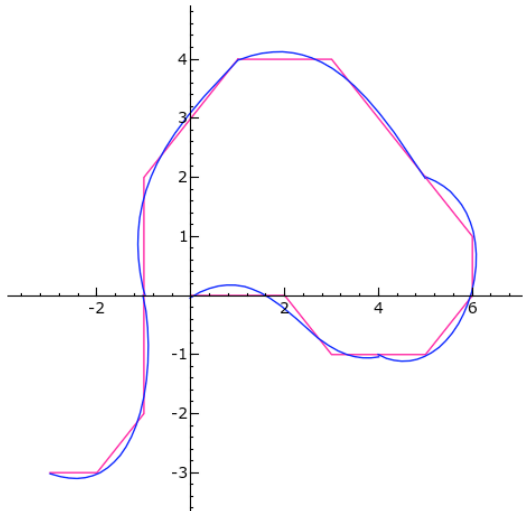
- 245 molecular fragments from 50 images of digested pUC19 (automated AFME vs manual NIH Image)
- As molecule fragment length increased, AFME progressively underestimated length
- AFME's initial length estimation error is below 2%
- Note the clusters: they indicate fragment lengths matching the restriction map: 75, 223, and 584 nm

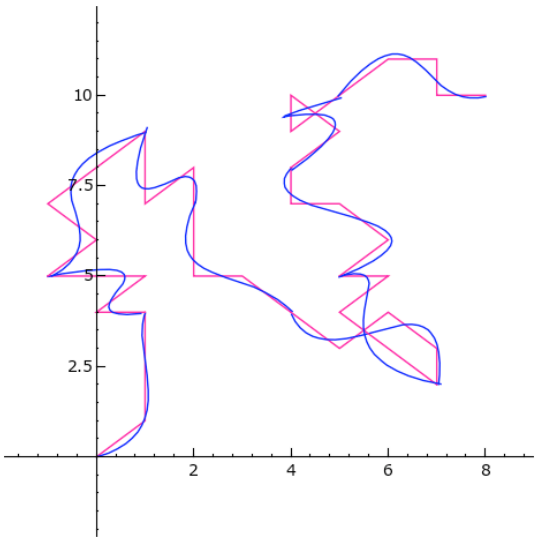


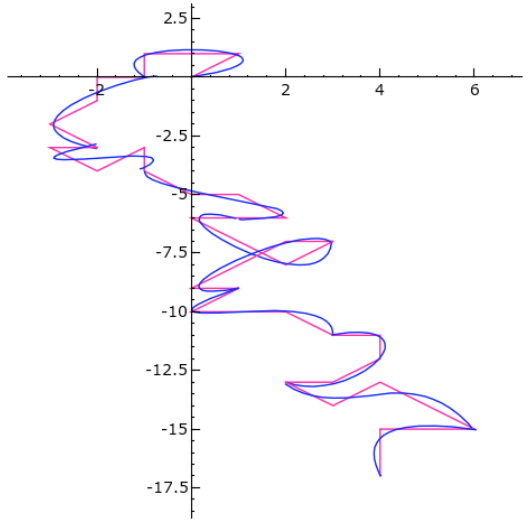
Example 5- and 10-spline Simulated Molecules

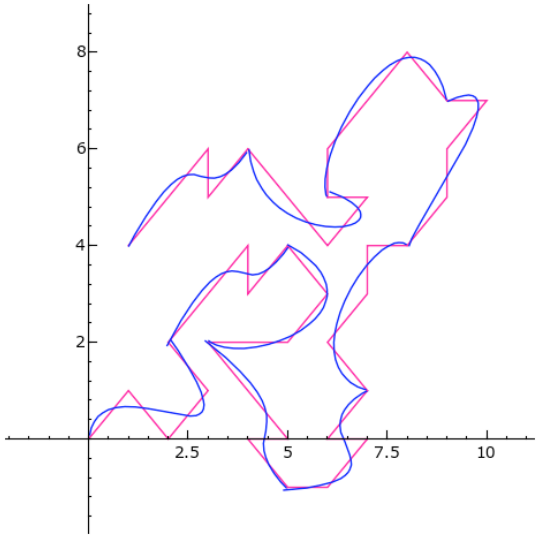






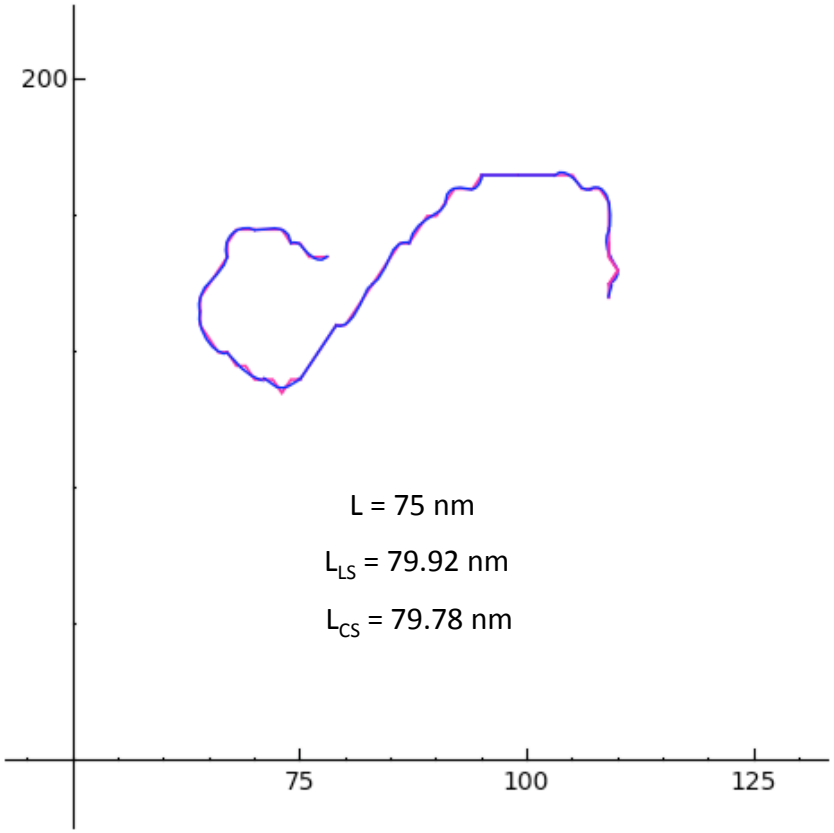






Example Real Molecule

18-spline calibration molecule



L = 75 nm

L_{LS} = 79.92 nm

L_{CS} = 79.78 nm

Solve for L_{CS} Length Correction Without Weighting

The model We train a linear regression model on $q \geq 5$ calibrating molecule backbones, $\vec{b}' \in \mathcal{B}'$, having known theoretical length $\mathcal{L}$, using values from these 5 features: $\{n_{horz}, n_{vert}, n_{diag}, n_{perp}, n_{tkav}\}$, giving

$$\begin{bmatrix} n_{horz}(\vec{b}'_1) & n_{vert}(\vec{b}'_1) & n_{diag}(\vec{b}'_1) & n_{perp}(\vec{b}'_1) & n_{tkav}(\vec{b}'_1) \\ n_{horz}(\vec{b}'_2) & n_{vert}(\vec{b}'_2) & n_{diag}(\vec{b}'_2) & n_{perp}(\vec{b}'_2) & n_{tkav}(\vec{b}'_2) \\ n_{horz}(\vec{b}'_3) & n_{vert}(\vec{b}'_3) & n_{diag}(\vec{b}'_3) & n_{perp}(\vec{b}'_3) & n_{tkav}(\vec{b}'_3) \\ n_{horz}(\vec{b}'_4) & n_{vert}(\vec{b}'_4) & n_{diag}(\vec{b}'_4) & n_{perp}(\vec{b}'_4) & n_{tkav}(\vec{b}'_4) \\ n_{horz}(\vec{b}'_5) & n_{vert}(\vec{b}'_5) & n_{diag}(\vec{b}'_5) & n_{perp}(\vec{b}'_5) & n_{tkav}(\vec{b}'_5) \\ & & \cdots & & \\ n_{horz}(\vec{b}'_q) & n_{vert}(\vec{b}'_q) & n_{diag}(\vec{b}'_q) & n_{perp}(\vec{b}'_q) & n_{tkav}(\vec{b}'_q) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} = \begin{bmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ l_5 \\ \cdots \\ l_q \end{bmatrix} \tag{1}$$

$$\iff N\vec{a} = \vec{l},$$

where N is the $q \times 5$ feature matrix, $\vec{a}$ is the correction coefficient 5-vector to solve for, and $\vec{l}$ is the length estimate error q-vector $[..., (\mathcal{L} - L_{CS}(\vec{b}'_i)), ...]$, where $i = 1, ..., q$. The model has the analytic solution

$$\vec{a} = (N^T N)^{-1} N^T \vec{l}. \tag{2}$$

Then each $\vec{b}' \in \mathcal{B}'$ obtains its final estimate, $\mathcal{L}' \in \{\mathcal{L}'_T, \mathcal{L}'_W\}$, from the correction function

$$\begin{aligned} C &: \mathcal{B}' \to \mathbb{R} \\ &: \vec{b}' \mapsto \\ &\quad a_1 n_{horz}(\vec{b}') + \\ &\quad a_2 n_{vert}(\vec{b}') + \\ &\quad a_3 n_{diag}(\vec{b}') + \\ &\quad a_4 n_{perp}(\vec{b}') + \\ &\quad a_5 n_{tkav}(\vec{b}'), \end{aligned} \tag{3}$$

and is given by

$$\mathcal{L}'(\vec{b}') = L_{CS}(\vec{b}') + C(\vec{b}'). \tag{4}$$

Outliers

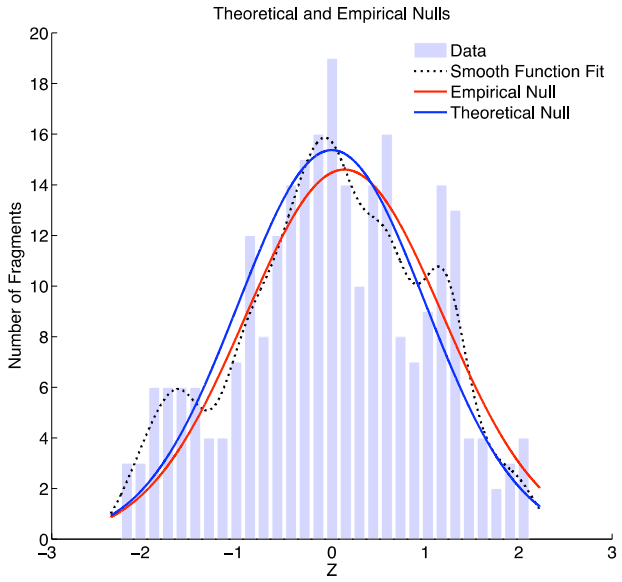
- Upon taking into consideration the difference between the empirically measured null distribution and the actual shape of the L_{CS} measurement distribution, certain observations appear to be false positives and others false negatives
- Use the empirical local false discovery rate (fdr)
- This suggests a weighted formulation of the error minimization problem given by

$$min\ ||\vec{r}||_W^2 = min \sum_{\vec{b'} \in \mathcal{B}'} W(\vec{b'}) r_{\vec{b'}}^2,$$

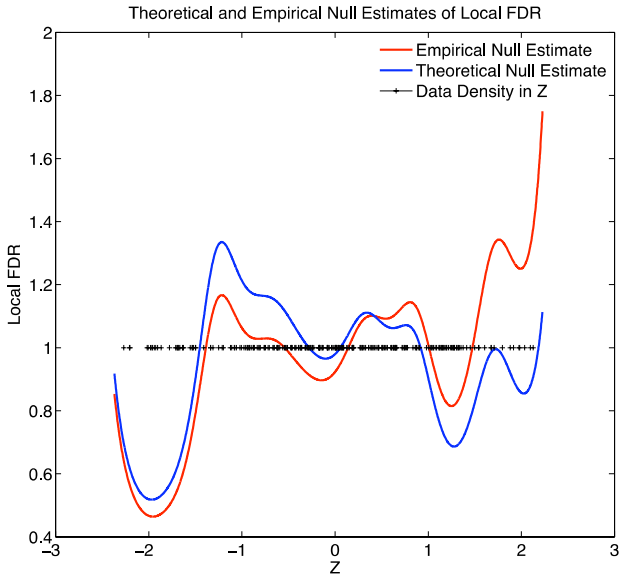
where $r_{\vec{b'}} = \mathcal{L} - L_{CS}(\vec{b'})$

and $W : \mathbb{R} \to \mathbb{R}$

is the local fdr weighting function.



(a) Theoretical and empirical null distributions of L_{CS} values of *Train*. $N = 263$, $\mu_T = 85.49$ nm, $\sigma_T = 6.73$ nm, $c_{v_T} = 0.08$, $\mu_E = 86.39$ nm, $\sigma_E = 7.09$, $c_{v_E} = 0.08$ nm. We obtain the empirical null by using the characteristic function approach taken by J. Jin and T. Cai. The smooth function fit, $f(Z)$, was created using Matlab's *ksdensity* function with a kernel width of 0.2.



(b) Local FDR curves derived from the theoretical and empirical null distributions of L_{CS} values of *Train*, with respect to $f(Z)$. The line at Local FDR = 1 indicates the data density along the Z axis. The local FDR curve derived from the empirical null distribution is used to weight *Train* data during the training phase.

Solving for the fdr-Weighted Correction Coefficients

The new weighted formulation of the estimator, $\mathcal{L}'_W$, is obtained by solving for $\vec{a}$ using the following Matlab pseudocode.

$$
\begin{aligned}
&N = diag(W) * N; \\
&\vec{l} = diag(W) * (\mathcal{L} - L_{CS}); \\
&\vec{a} = N \backslash \vec{l};
\end{aligned}
$$

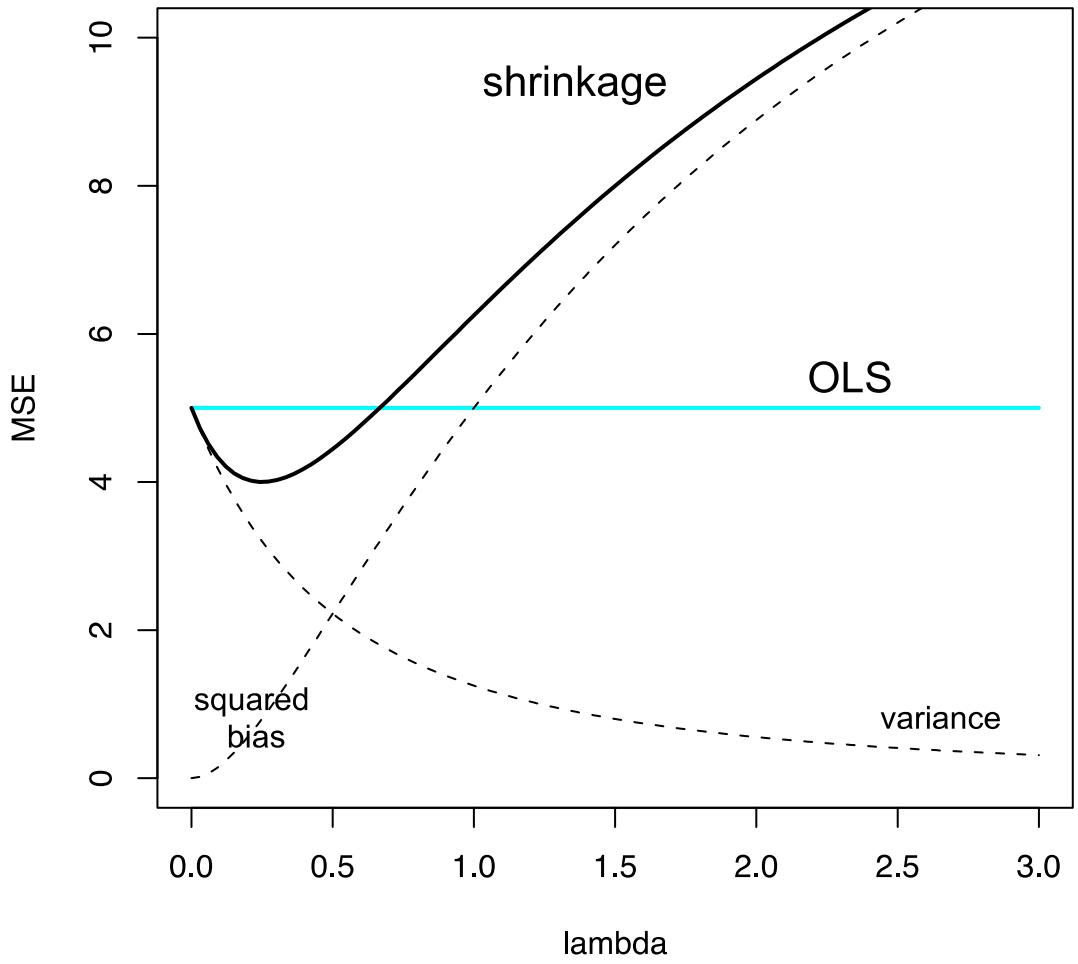
James-Stein Shrinkage

- In our modeling of estimation error so far, one or more features in training may introduce too much variance (systematic error) or dependence (model error)
- In 1961, James and Stein published their seminal paper describing a method to improve estimating a multivariate normal mean

$$\vec{\mu} = [\mu_1, ..., \mu_k]$$

under expected sum of squares loss, provided the degree of freedom, k, is at least 3

Background for James-Stein Shrinkage (continued)



Spherical James-Stein Shrinkage

Let $\vec{a} = [a_i, ..., a_k]$ have a k-variate normal distribution with mean vector $\vec{\mu}$ and covariance matrix $\sigma^2 I$, which we measure empirically in train mode. We would like to estimate $\vec{\mu}$ using an estimator

$$\delta(\vec{a}) = [\delta_1(\vec{a}), ..., \delta_k(\vec{a})] \tag{1}$$

under the sum of squares error loss

$$L(\vec{\mu}, \delta) = \sum_{i=1}^{k} (\mu_i - \delta_i)^2 \tag{2}$$

In terms of expected loss,

$$R(\vec{\mu}, \delta) = E_\mu[L(\vec{\mu}, \delta(\vec{a}))], \tag{3}$$

James and Stein show that when $k \geq 3$, an improved estimator is obtained by a symmetric (or spherical) shrinkage in $\vec{a}$ given by

$$\delta(\vec{a}) = \left[1 - \frac{\kappa(q-k)s^2}{\sum_{i=1}^{q} (N\vec{a})_i^2}\right]^+ \vec{a}, \tag{4}$$

where

$$\kappa = \frac{(k-2)}{(q-k+2)}, \tag{5}$$

and s^2 is the empirical estimate of variance, σ^2, given by

$$s^2 = \frac{1}{(q-k)} \sum_{i=1}^{q} (\mathcal{L} - L_{CS_i} - (N\vec{a})_i)^2. \tag{6}$$

and where $[x]^+ \equiv max\{0, x\}$.

Truncated James-Stein Shrinkage

When extreme μ_i are likely, then spherical shrinkage may give little improvement. This may occur, for instance, when the μ_i arise from a prior distribution with a long tail. A property of spherical shrinkage is that its performance is guaranteed only in a small subspace of parameter space, requiring that one select an estimator designed with some notion of where $\vec{\mu}$ is likely to be, such that the estimator shrinks toward it. An extreme μ_i will likely be outside of any small selected subspace, implying a large denominator and so little, if any, shrinkage in $\vec{a}$, thereby giving no improvement. To address this problem, Stein proposed a coordinate-based (or truncated) shrinkage method, given by

$$\delta_i^{(f)}(\vec{a}) = \left[1 - \frac{(f-2)s^2 min\{1, \frac{z_{(f)}}{|a_i|}\}}{\sum_{j=1}^{q} (N\vec{m})_j^2}\right]^+ a_i, \tag{1}$$

where f is a "large fraction" of k, $z_i = |a_i|, i = 1, ..., k$, $z_{(1)} < z_{(2)} < ... < z_{(f)} < ... < z_{(k)}$ forms a strictly increasing ordering on $z_1, ..., z_k$, s^2 is the empirical estimate of variance, σ^2, given by

$$s^2 = \frac{1}{(q-k)} \sum_{i=1}^{q} (\mathcal{L} - L_{CS_i} - (N\vec{a})_i)^2, \tag{2}$$

and $\vec{m}_i = min\{a_i, z_{(f)}\}, i = 1, ..., k.$ tein shows this estimator is minimax if $f \geq 3$. Observe that the denominator is small even when $(k - f)$ of the μ_i are extreme.

Shrinking Did Little to Our Feature Space

$\mathbf{i}$	1	2	3	4	5
train	1.000000	1.000000	1.000000	1.000000	1.000000
$\mathbf{a_i}$	-0.258699	-0.316009	-0.197179	-0.742293	1.637360
spherical	0.997422	0.997422	0.997422	0.997422	0.997422
$\delta_\mathbf{i}(\vec{\mathbf{a}})$	-0.258037	-0.315201	-0.196675	-0.740394	1.633170
truncated $(\mathbf{f=5})$	0.997422	0.997422	0.997422	0.997422	0.997422
$\delta_\mathbf{i}^{(\mathbf{5})}(\vec{\mathbf{a}})$	-0.258037	-0.315201	-0.196675	-0.740394	1.633170
truncated $(\mathbf{f=4})$	0.999108	0.999108	0.999108	0.999108	0.999596
$\delta_\mathbf{i}^{(\mathbf{4})}(\vec{\mathbf{a}})$	-0.258468	-0.315727	-0.197003	-0.741631	1.636700
truncated $(\mathbf{f=3})$	0.999655	0.999655	0.999655	0.999853	0.999933
$\delta_\mathbf{i}^{(\mathbf{3})}(\vec{\mathbf{a}})$	-0.258610	-0.315900	-0.197111	-0.742184	1.637250

- In our experiments, James-Stein shrinkage factors were nearly 1, indicating our 5 features had little noise or dependence
- Hence, we were confident our linear regression model did not overfit

Training and Test Data Sets Used in Experiments

Data Set	Images	Fragments	τ (nm)
Train	5	263	75
Test Knowns	14	2,452	75
Test Unknowns A	44	15,477	223
Test Unknowns B	101	54,093	584

Experimental Results (continued)

	Test Knowns N=2,452 τ=75 nm			Test Unknowns A N=15,477 τ=223 nm			Test Unknowns B N=15,093 τ=584 nm		
	μ	σ	c_v	μ	σ	c_v	μ	σ	c_v
$\mathbf{L_{CS}}$	89.05	8.27	0.09	278.91	14.75	0.05	669.66	87.21	0.13
$\mathcal{L}'_{\mathbf{W}}$	78.54	7.91	0.10	233.57	10.85	0.05	553.42	36.71	0.07

Experiments

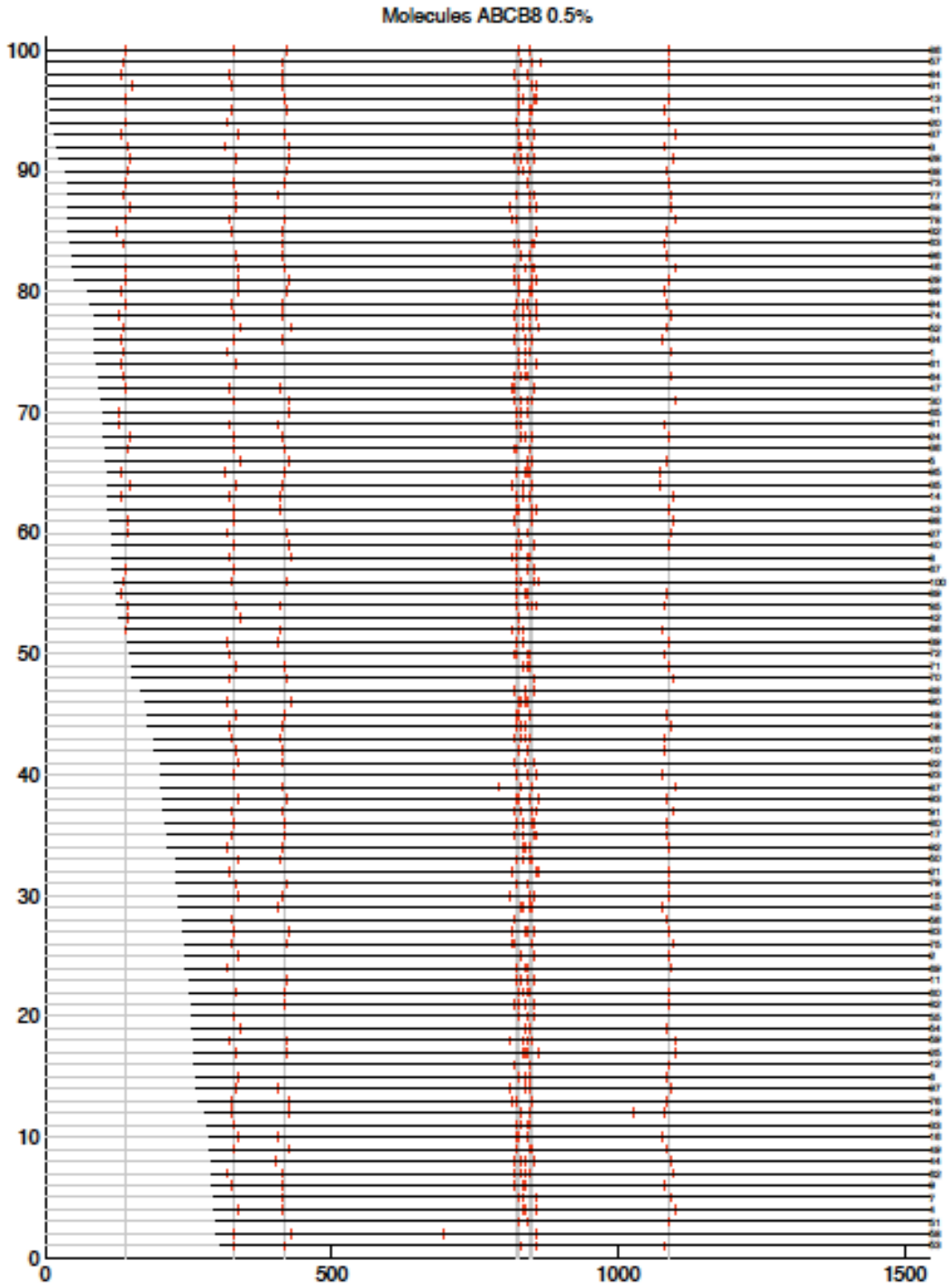
Gene Family	# Members	# Variants	# Unique Variants	Percentage
ABC	81	54	1	67.50
ABHD	26	15	0	57.69
ADAMTS	23	7	0	30.43
ALDH	33	25	1	78.13
ATP	222	153	2	69.55
CACN	88	74	0	84.09
CATSPER	8	3	0	37.50
CTS	26	15	0	57.69
CYP	75	32	0	42.67
DNAJ	58	27	1	47.37
FOX	64	26	0	40.63
FZD	13	3	0	23.08
GJ	27	12	0	44.44
GPR	207	92	2	44.88
IFT	32	26	0	81.25
IL	157	99	2	63.87
KCN	154	98	0	63.64
KIF	57	24	0	42.11
KRT	60	9	0	15.00
NLR	38	23	0	60.53
PAX	27	23	0	85.19
PDI	23	7	1	31.82
PG	64	36	0	56.25
PTP	227	177	0	77.97
RAB	84	38	1	45.78
SCN	38	29	0	76.32
SERPIN	59	35	0	59.32
SLC	600	368	7	62.06
SMAD	18	16	0	88.89
SMC	11	7	0	63.64
SOX	25	9	0	36.00
TBX	27	16	0	59.26
TNFRSF	43	25	0	58.14
TNFSF	85	45	0	52.94
TRIM	125	79	1	63.71
UBA	10	6	0	60.00
USP	75	36	0	48.00
WNT	22	7	1	33.33
ZFYVE	55	35	0	63.64
ZNF	904	467	0	51.66
All families	**3971**	**2278**	**20**	**57.66**
Whole database	**29563**	**16885**	**267**	**57.64**

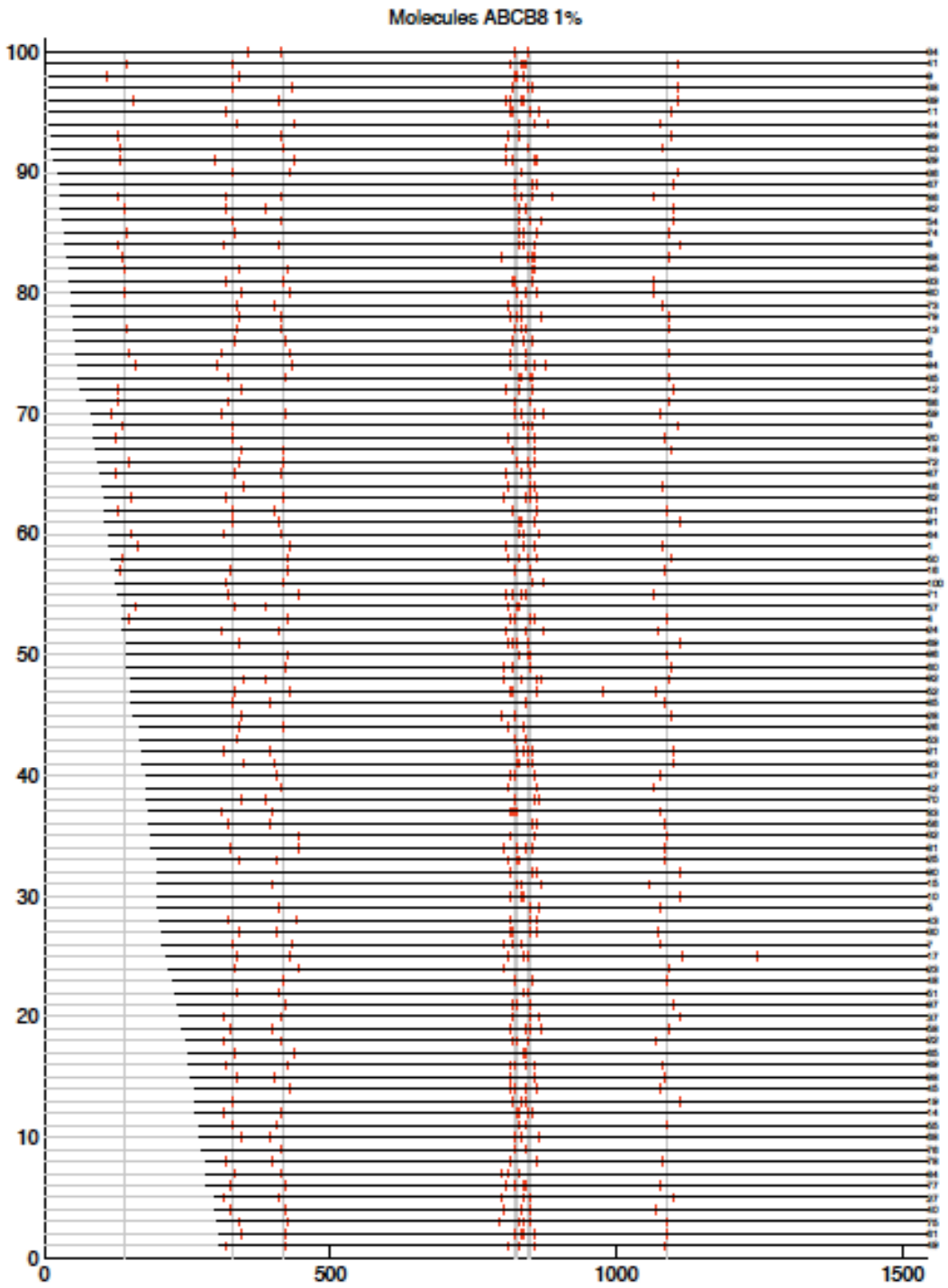
Simulated molecules

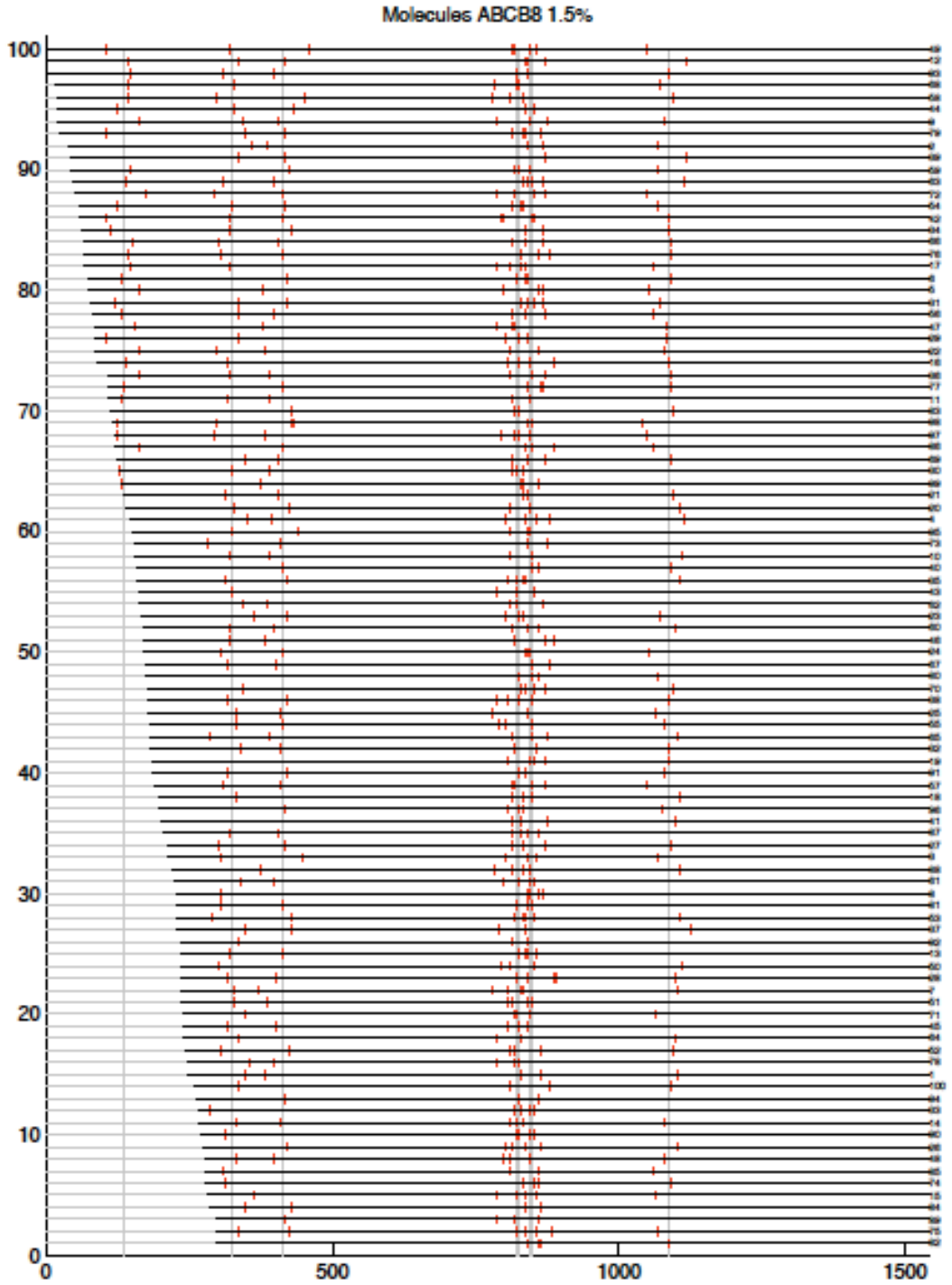
- labeling rate: 80%
- labeling position error
  - normal distribution
  - s.d. = 0.5%/1%/1.5%/2% of length
  - bounded to +/- 2 s.d.
- false label probability: 3%
- length truncated from 5’ end
  - uniform distribution
  - 0 - 20%

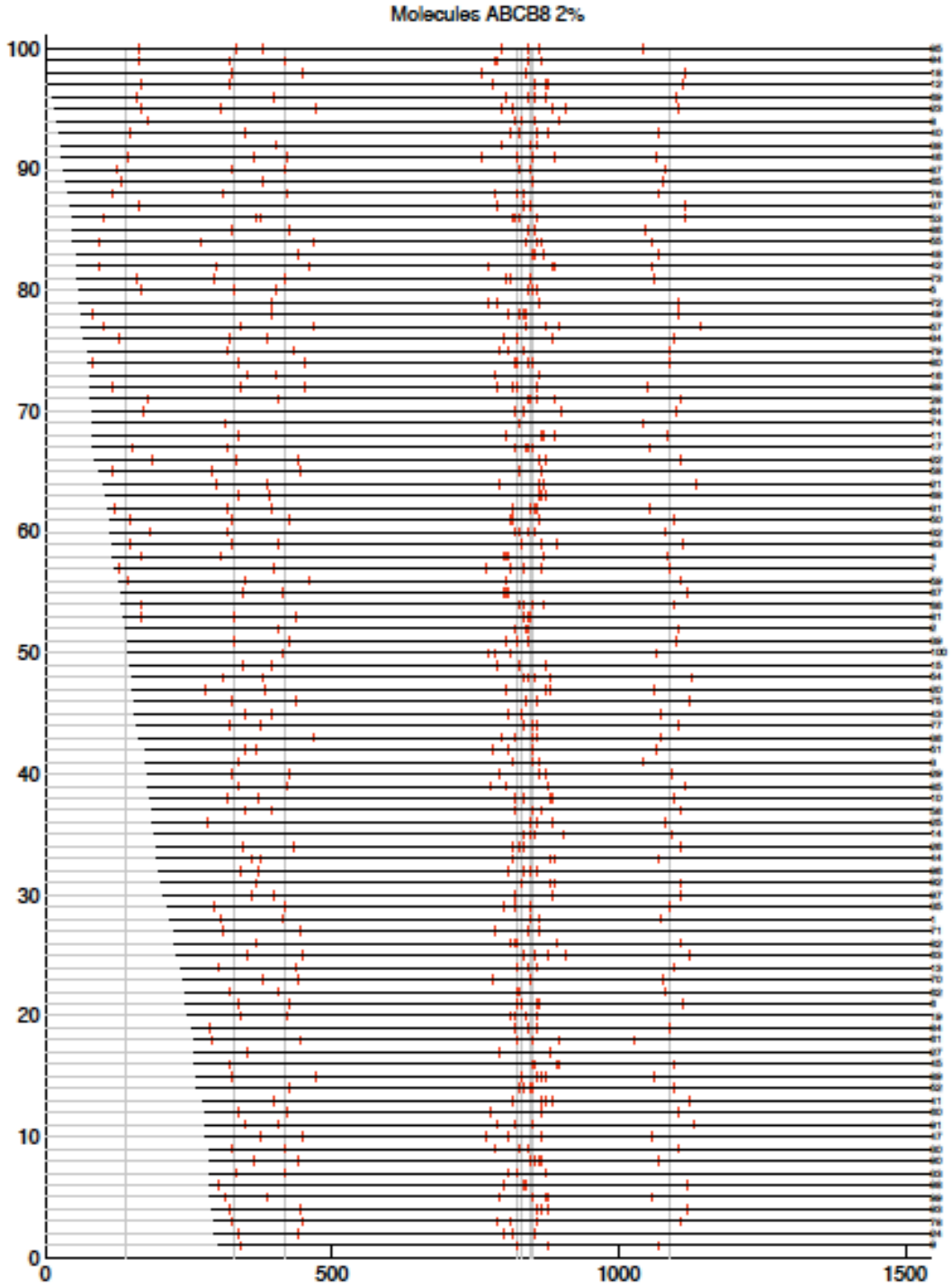
Notations

- Total number of genes in database n = 29563
- Hypothesis H_k, k=1 to n
- Molecule M_{ij} i=1 to n, j=1 to 100
- 81 genes in database belongs to ABC gene family
  - 69 completed


Molecules ABCB8 0.5%
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500


Molecules ABCB8 1%
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500


Molecules ABCB8 1.5%
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500


Molecules ABCB8 2%
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500

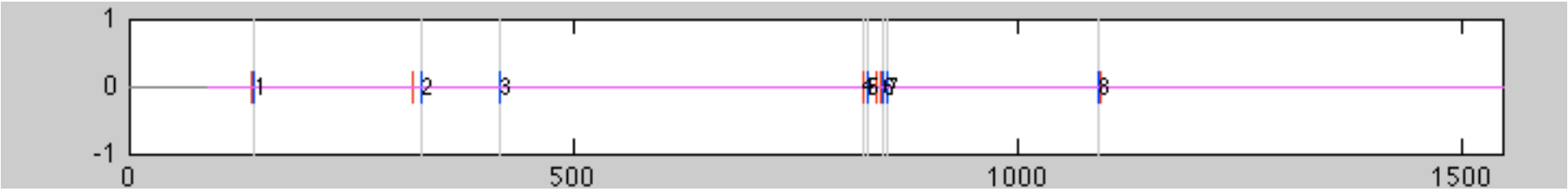
Alignment Score

- Compare length of H_k and M_{ij}, continue only if $L(M_{ij})$ is within 80-100% of $L(H_k)$
- Align H_k and M_{ij} at the 3' end, discard the labels on H_k those exceed length $L(M_{ij})$
- Continue only if $n_{ij}-n_k \leq 3$ (no more than 3 false labels)
- Compute alignment table by given errors
- Generate all *possible* alignment combinations
- Select a subset of combinations that contains the minimum number of missing labels
- Calculate alignment scores (probability) for the selected combinations and save the maximum score

Alignment Example 1

- ABCB8 molecule 1: H_{2958} $M_{2958,1}$
- Align Table

	m_1	m_2	m_3	m_4	m_5	m_6	Candidates
h_1	1	0	0	0	0	0	m_1
h_2	0	1	0	0	0	0	m_2
h_3	0	0	0	0	0	0	
h_4	0	0	1	1	1	0	m_3, m_4, m_5
h_5	0	0	1	1	1	0	m_3, m_4, m_5
h_6	0	0	1	1	1	0	m_3, m_4, m_5
h_7	0	0	1	1	1	0	m_3, m_4, m_5
h_8	0	0	0	0	0	1	m_6

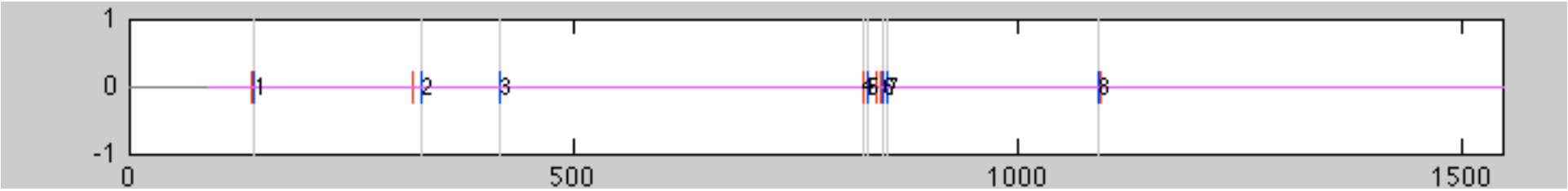


Alignment Example 1

- Generate all possible combinations (# false labels≤3, no repetition, sorted order) (261 found)
- Select combinations that have minimum # missing labels

h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	
m_1	m_2	0	m_3	m_4	m_5	0	m_6	
m_1	m_2	0	m_3	m_4	0	m_5	m_6	
m_1	m_2	0	m_3	0	m_4	m_5	m_6	*
m_1	m_2	0	0	m_3	m_4	m_5	m_6	

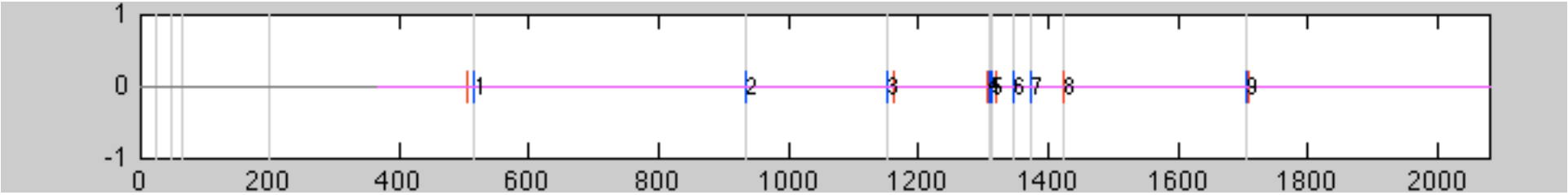
- Calculate scores for all selected combinations, save the maximum one



Alignment Example 2

- ABCA9 molecule 3: H_{3107} $M_{3107,3}$
- Align Table

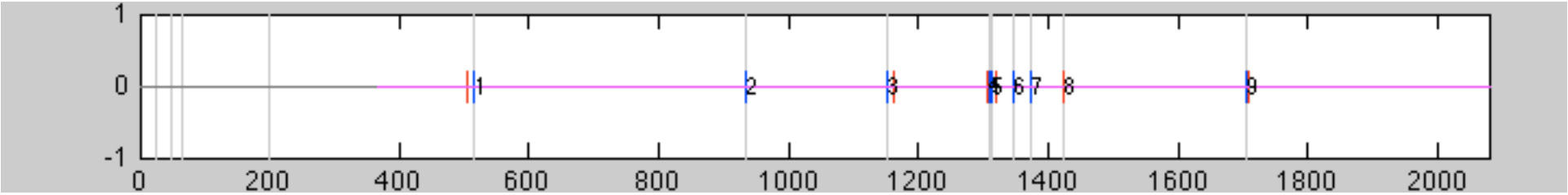
	m_1	m_2	m_3	m_4	m_5	m_6	Candidates
h_1	1	0	0	0	0	0	m_1
h_2	0	0	0	0	0	0	
h_3	0	1	0	0	0	0	m_2
h_4	0	0	1	1	0	0	m_3, m_4
h_5	0	0	1	1	0	0	m_3, m_4
h_6	0	0	1	1	0	0	m_3, m_4
h_7	0	0	1	1	1	0	m_3, m_4, m_5
h_8	0	0	0	0	1	0	m_5
h_9	0	0	0	0	0	1	m_6



Alignment Example 2

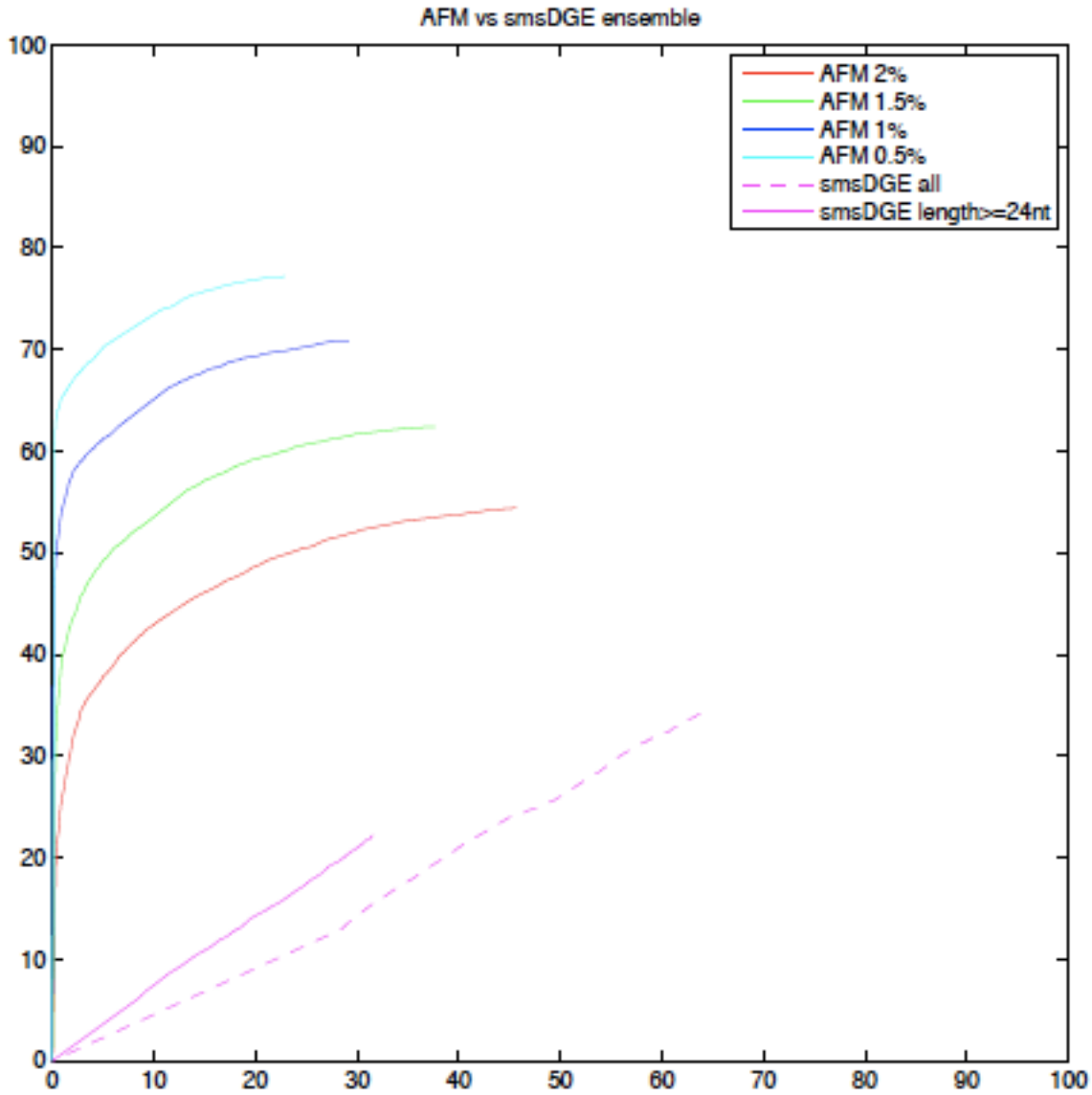
- 305 possible combinations found
- Select combinations that have minimum # missing labels

h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	
m_1	0	m_2	m_3	m_4	0	m_5	0	m_6	
m_1	0	m_2	m_3	0	m_4	m_5	0	m_6	
m_1	0	m_2	0	m_3	m_4	m_5	0	m_6	
m_1	0	m_2	m_3	m_4	0	0	m_5	m_6	*
m_1	0	m_2	m_3	0	m_4	0	m_5	m_6	
m_1	0	m_2	0	m_3	m_4	0	m_5	m_6	
m_1	0	m_2	m_3	0	0	m_4	m_5	m_6	
m_1	0	m_2	0	m_3	0	m_4	m_5	m_6	
m_1	0	m_2	0	0	m_3	m_4	m_5	m_6	



ROC

- Every gene i has a score matrix $S_i=\{s_{jk}\}$, j=1 to 100, k=1 to n
- Calculate number of True Positives and False Positives while varying threshold θ from 0 to 1
  - True positive: For each row of S_i (fixed j, k=1 to n), if $\max(\{s_{jk}\}) = s_{ik} \geq \theta$ and there exists only one max, i.e. molecule M_{ij} matches with H_i
  - False positive: otherwise (M_{ij} matches with H_k, $i \neq k$ or more than one max score or max score$<\theta$)


AFM vs smsDGE ensemble
AFM 2%
AFM 1.5%
AFM 1%
AFM 0.5%
smsDGE all
smsDGE length>=24nt
0
10
20
30
40
50
60
70
80
90
100

How do we make sense of the massive amount of such single-cell single-molecule data?

More Theory & Less Measurement…

Did Hooke never get it?

Hooke

Thursday 25 May 1676

Damned Doggs.

Vindica me deus.

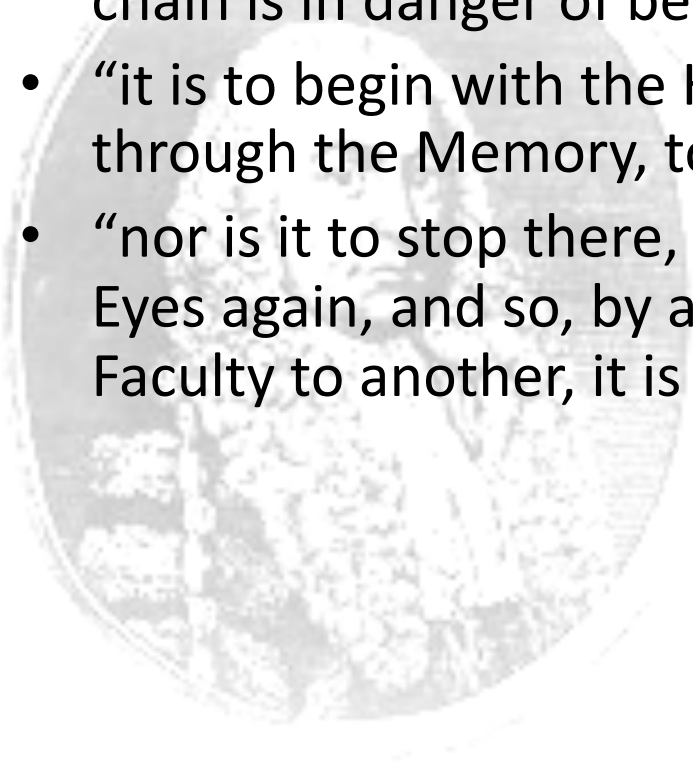
- Commenting on

Sir Nicholas Gimcrack character in

The Virtuoso, a play by Thomas Shadwell.

Hooke…

- “So many are the links, upon which the true Philosophy depends, of which, if any can be loose, or weak, the whole chain is in danger of being dissolved;
- “it is to begin with the Hands and Eyes, and to proceed on through the Memory, to be continued by the Reason;
- “nor is it to stop there, but to come about to the Hands and Eyes again, and so, by a continuall passage round from one Faculty to another, it is to be maintained in life and strength.”



Hooke

in the Royal Society, 26 June 1689

- “I have had the misfortune either not to be understood by some who have asserted I have done nothing…
- “Or to be misunderstood and misconstrued (for what ends I now enquire not) by others…
- “And though many things I have first Discovered could not find acceptance yet I finde there are not wanting some who pride themselves on arrogating of them for their own…
- “—But I let that passe for the present.”