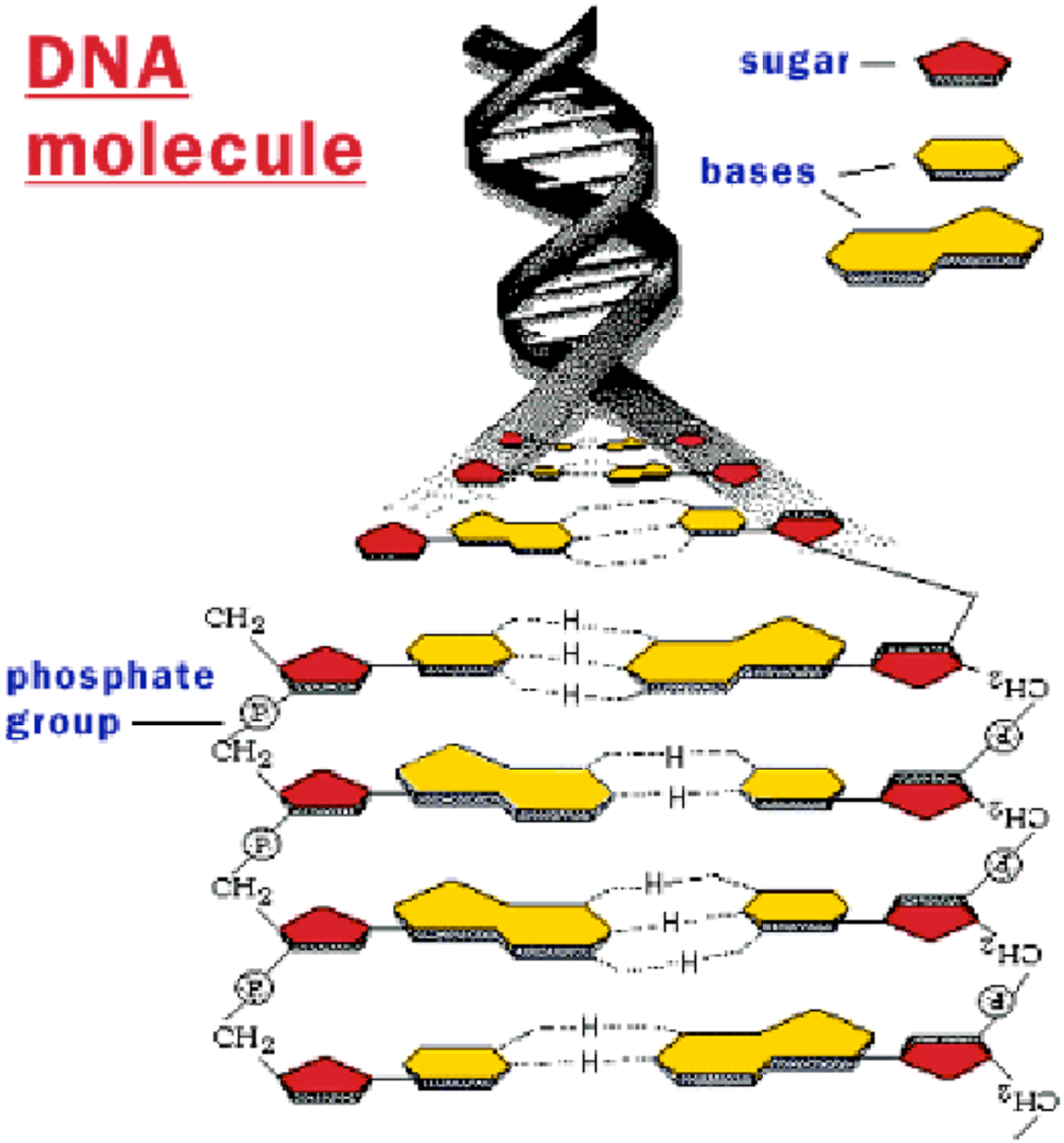
Lecture 4: Plan B from Outer Space: Nanochannels and Nanopores

Question from yesterday (I think I either totally bored you or totally lost you or both):

If shotgun sequencing needs to sequence only short pieces of DNA (up to 1000 bp) by sorting by length, why bother to look at big pieces of DNA at all?

A little biology about the importance of both gene sequence and transcription factors.

DNA molecule



DNA contains the code for you. There are about 3 billion basepairs in the human genome, or about 1 meter of DNA.

earth

country

state

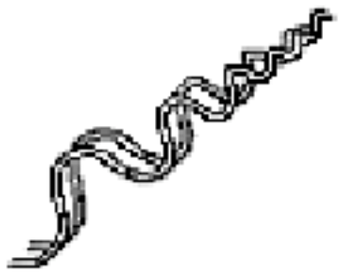
city

people

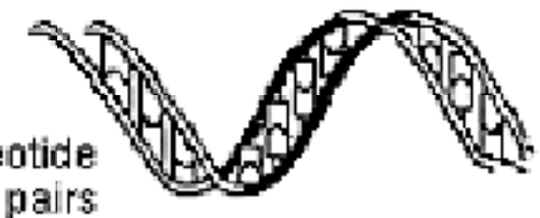
cell

chromosome

chromosome fragment



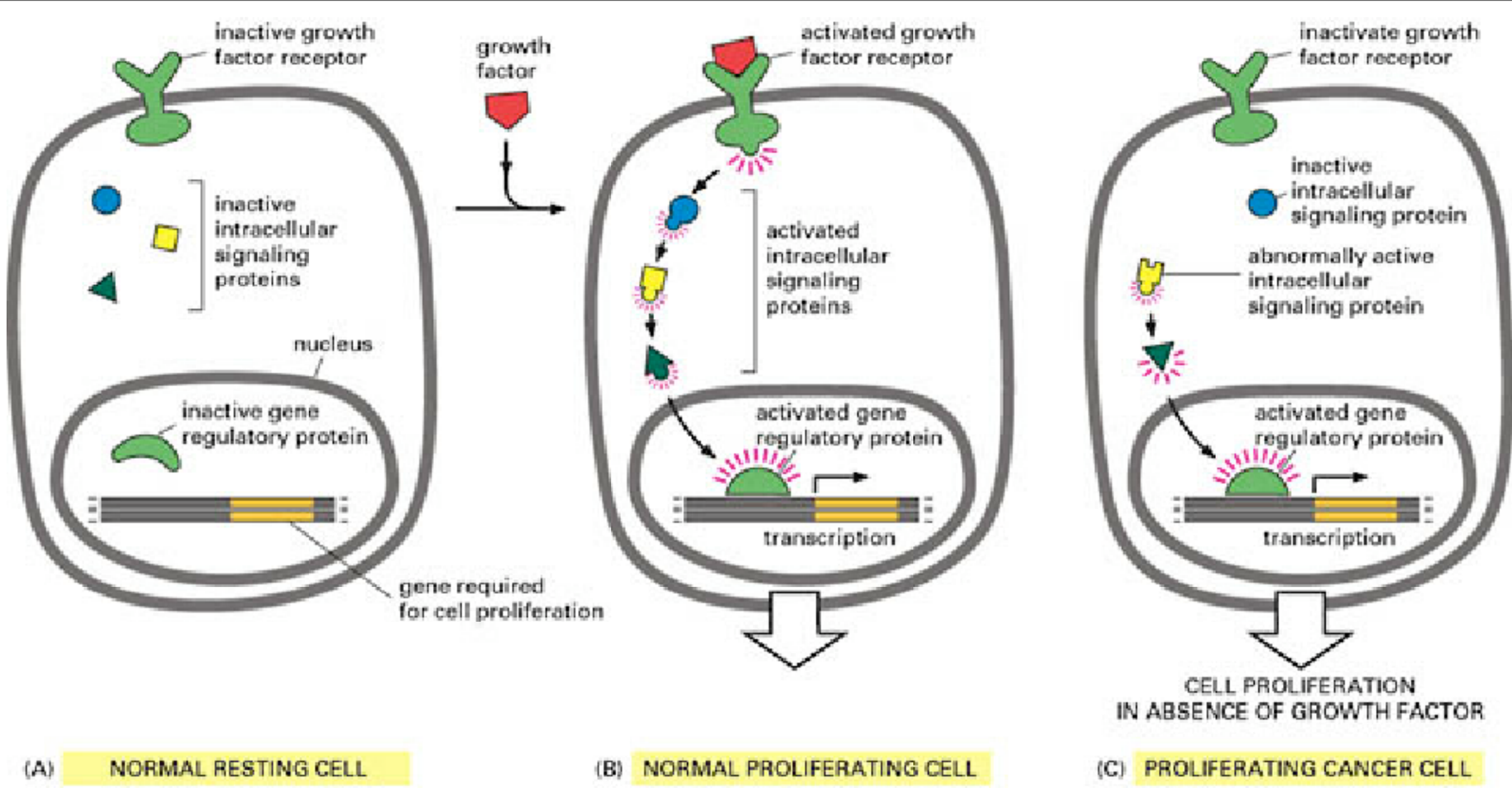
gene



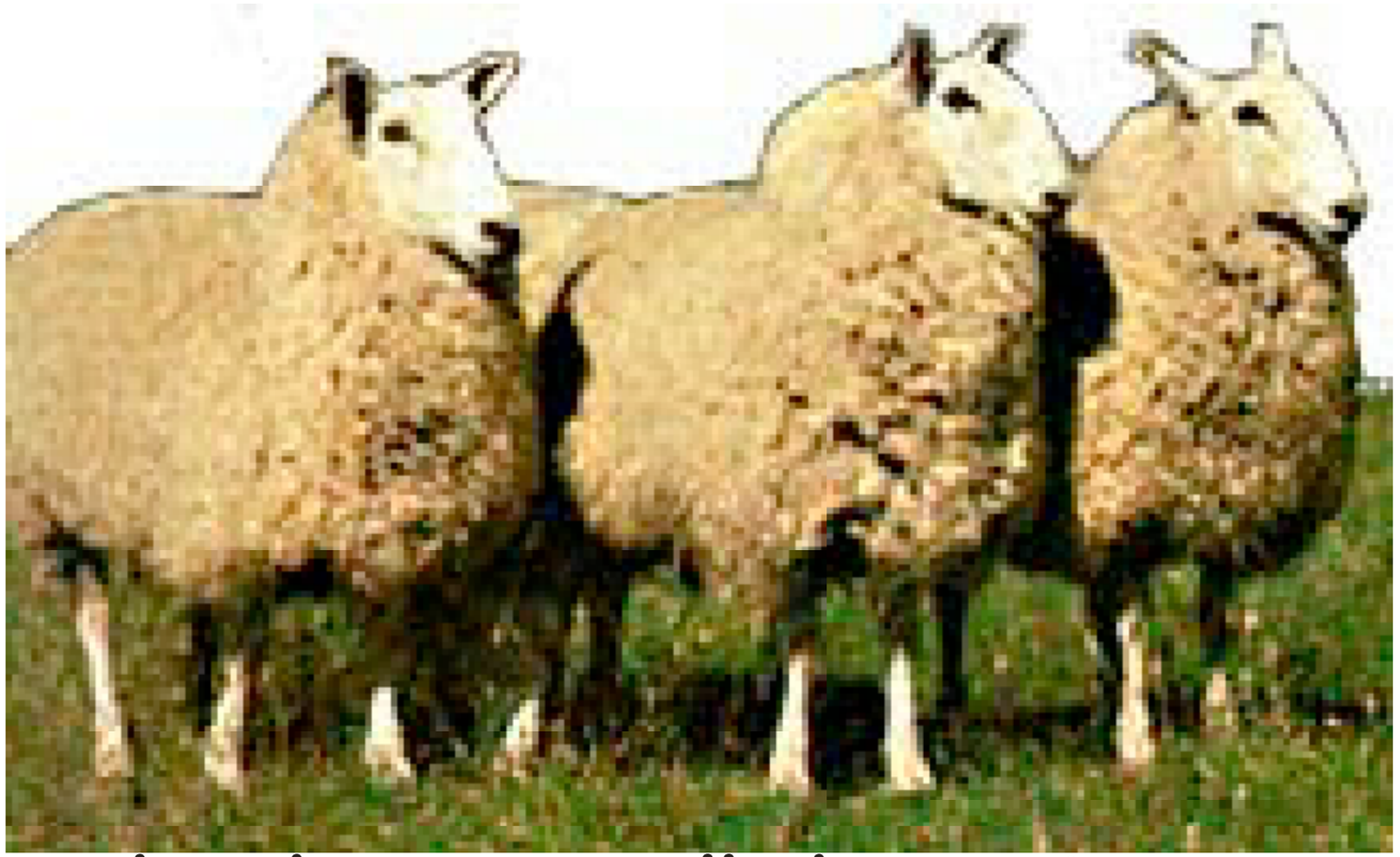
nucleotide base pairs

Comparative Scale of Mapping

Nearly all the 10^{14} cells in your body contain the SAME genome. Probably each cell has a slightly different genetic sequence. But, it isn't the genome that makes each cell different.



Promoter and repressor proteins, which bind to specific parts of the genome, control expression. The CYTOPLASM of the cells contains the DYNAMIC control information. The DNA is the ROM, the proteins are the OS that makes a liver cell a liver cell.



That's why there was all the excitement about Dolly: they took the NUCLEUS from a fully differentiated cell in the udder, put it into the egg cell of another sheep which presumably had the right protein content to reset the clock, and transformed a mature cell into a "fertilized" egg. In some respects, this is what cancer does: goes BACK in time, to embryo level.

For example, in the case of metastatic cancer, the transformed cells that have lost control and think they are immortal, spread through the blood stream and form little colonies. These transformed cells are different than normal cells:

(1) genome is probably different (mistakes)
(2) control proteins are different (wrong controls)
(3) cytoskeleton is different (growing too fast)
(4) surface proteins different (not a differentiated cell)

The problem is, it only takes 1 transformed cell to kill you. You need to FIND and CHARACTERIZE those rare cells, perhaps a SINGLE cell.

OK, that is the biology lesson for today.
Hopefully you are motivated now to
see why we want to:

(1) FIND rare cells circulating in the blood
(2) EXTRACT the DNA
(3) FRACTIONATE the chromatin by length
(4) SCAN the protein pattern on the DNA
(5) MAP the genome

The logical progression is: use blood, find out how to sort blood cells, find out how to capture chromatin, find how to analyze DNA at several levels.

Once you get the megabase DNA molecules extracted and sized, the next major task is to analyze the DNA for two things:

(1) The sequence of bases on the DNA

(2) The control proteins that bind tightly to the DNA

Both problems require scanning with very high precision the DNA molecule.

III. Some recent work in our lab to use nanochannels to analyze dsDNA at high spatial resolution.

"SEQUENCING AND RESEQUENCING THE HUMAN GENOME": A "RECENT" MEETING AT THE NIH.

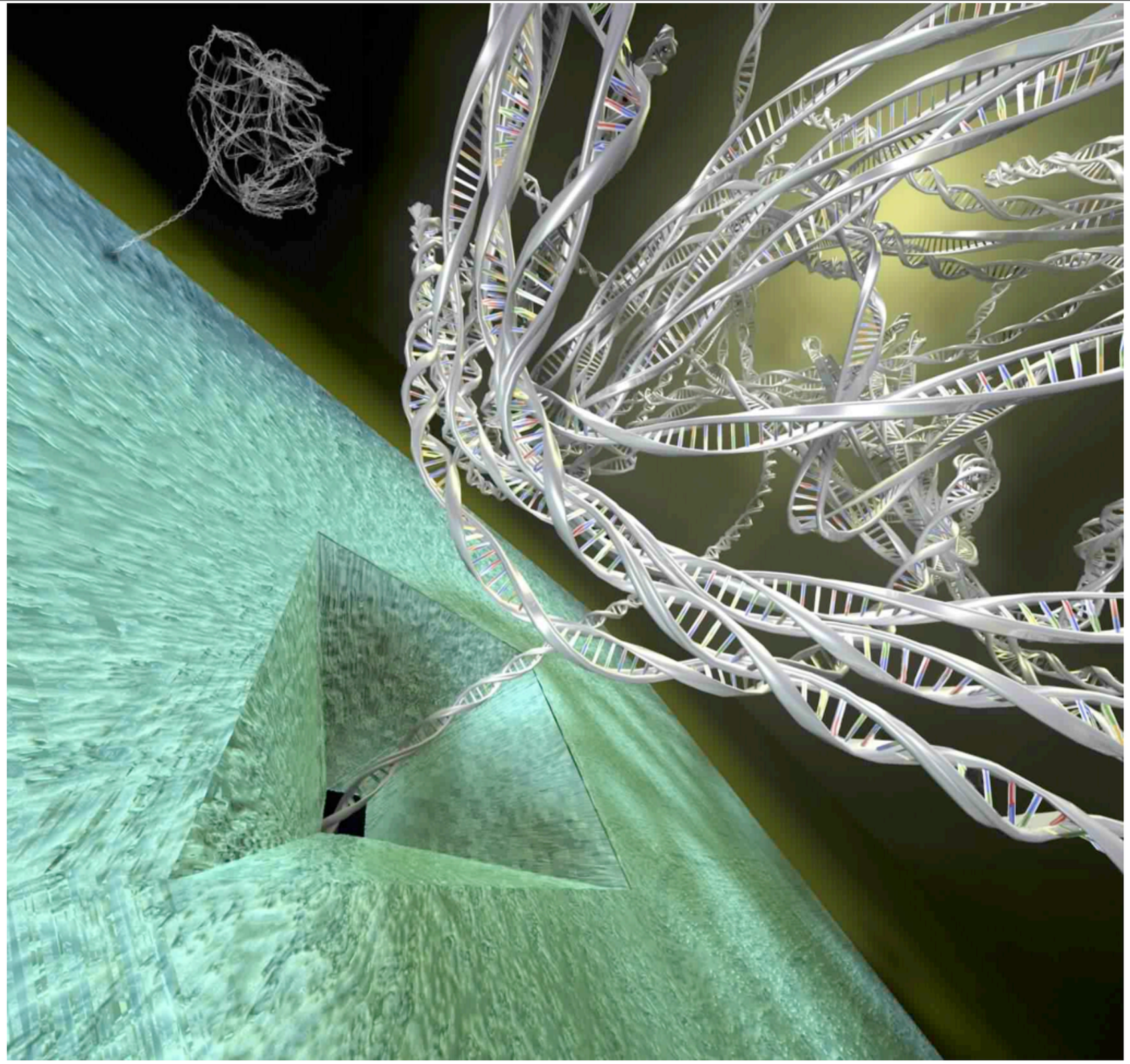
THERE WAS A CLEAR CONCENSUS: WE NEED TO DROP THE PRICE OF SEQUENCING A LARGE GENOME FROM ABOUT $100,000,000 TO $1,000: A DROP IN 5 ORDERS OF MAGNITUDE.

HOW CAN WE DO IT?

SEQUENCING DNA USING SEMI-CONVENTIONAL TECHNOLOGIES REQUIRES BEING ABLE SEPARATE DNA MOLECULES THAT ARE 0.3 NM (1 BASEPAIR) DIFFERENT IN LENGTH.

PRESENT TECHNIQUES CAN DISTINGUISH BASEPAIRS OUT TO ABOUT 700 BP. PRESENT MACHINES CAN DO THIS AT ABOUT 1 BASEPAIR/ SEC.

IT'S GREAT, BUT NOT GOOD ENOUGH. CAN WE DO BETTER USING NANOFABRICATED TECHNIQUES?



http://www.mb.tn.tudelft.nl/image_gallery/big/dnatranslocatie_big.jpg

Voltage-Driven DNA Translocations through a Nanopore

Amit Meller,[1,2,*] Lucas Nivon,[1,3] and Daniel Branton[3]

[1]*The Rowland Institute for Science, Cambridge, Massachusetts 02142*

[2]*Department for Biomedical Engineering, Boston University, Boston, Massachusetts 02215*

[3]*Department of Molecular & Cellular Biology, Harvard University, Cambridge, Massachusetts 02138*

(Received 8 December 2000)

The NIHGRI is betting pretty heavily on nanopores in some guise to do single molecule sequencing of DNA. I think some of you are interested in this technique.

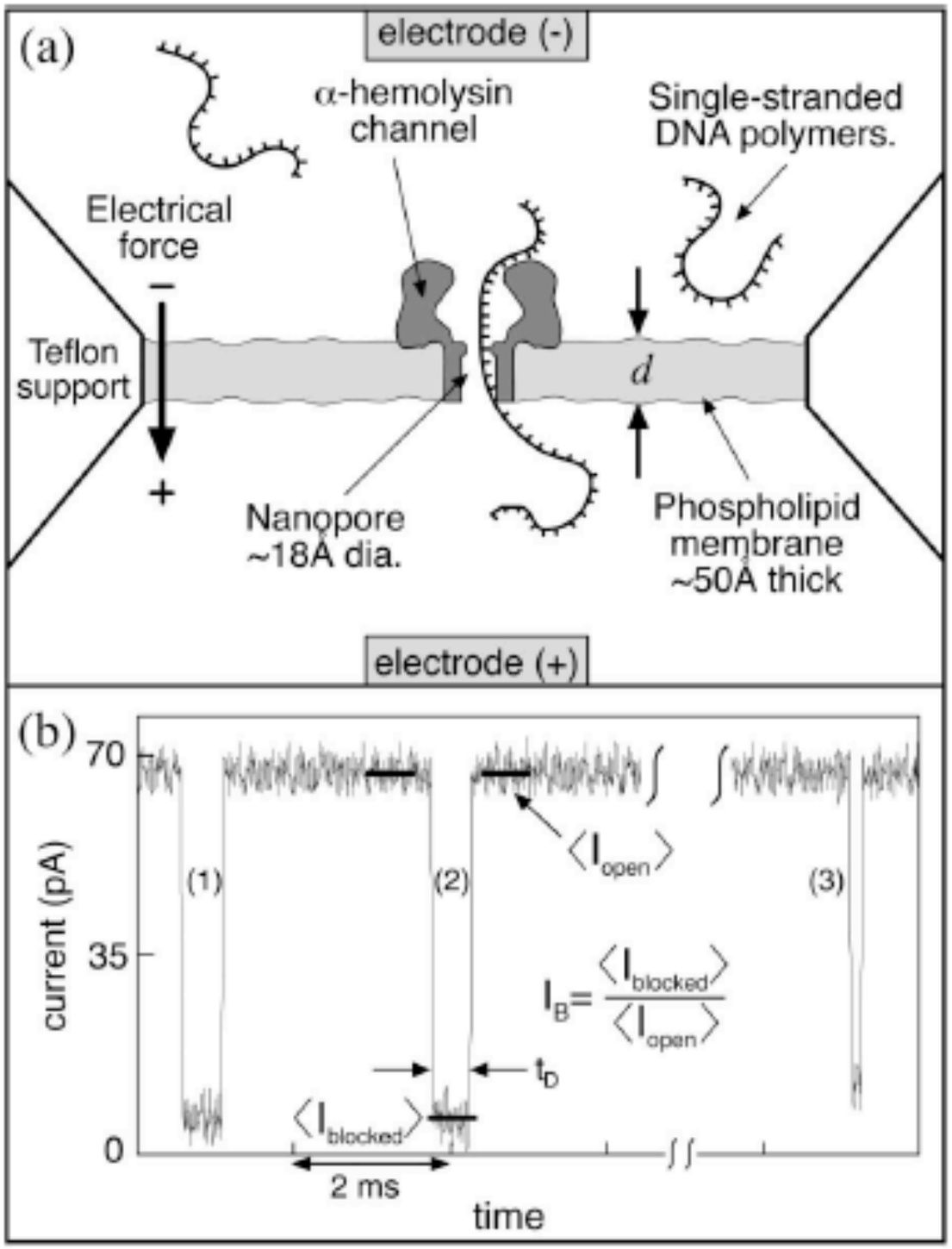


FIG. 1. (a) Single-stranded DNA molecules (negatively charged) and salt ions are electrically driven through a single α-hemolysin protein pore embedded in a phospholipid membrane. Most of the ionic current through the pore is blocked during DNA passage. (b) Three representative translocation current blockades are shown for 15mer ("1" and "2") and for 7mer ("3") poly(dA). For each event we measured the translocation time, t_D, and the average event blockade, $\langle I_B \rangle$.

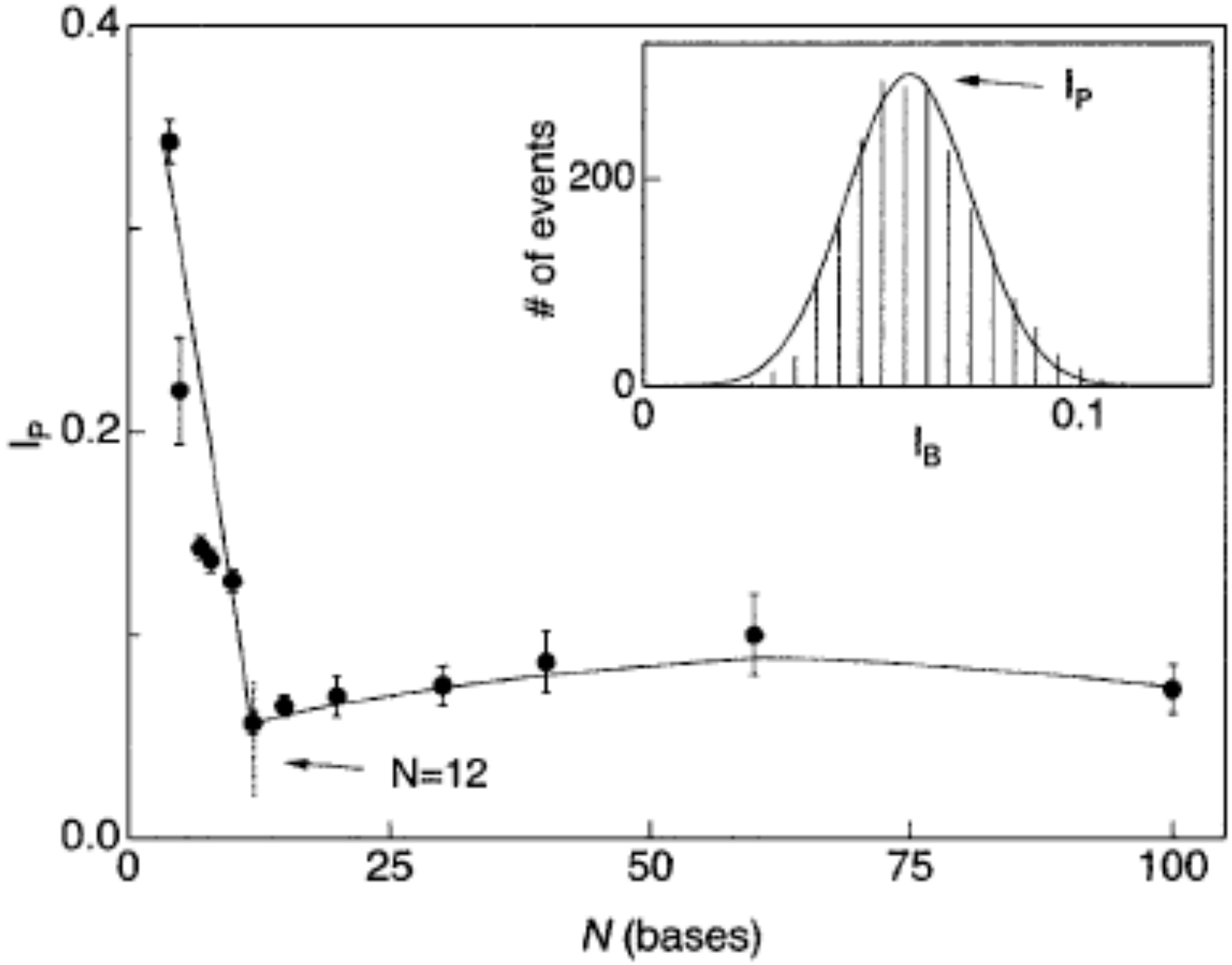
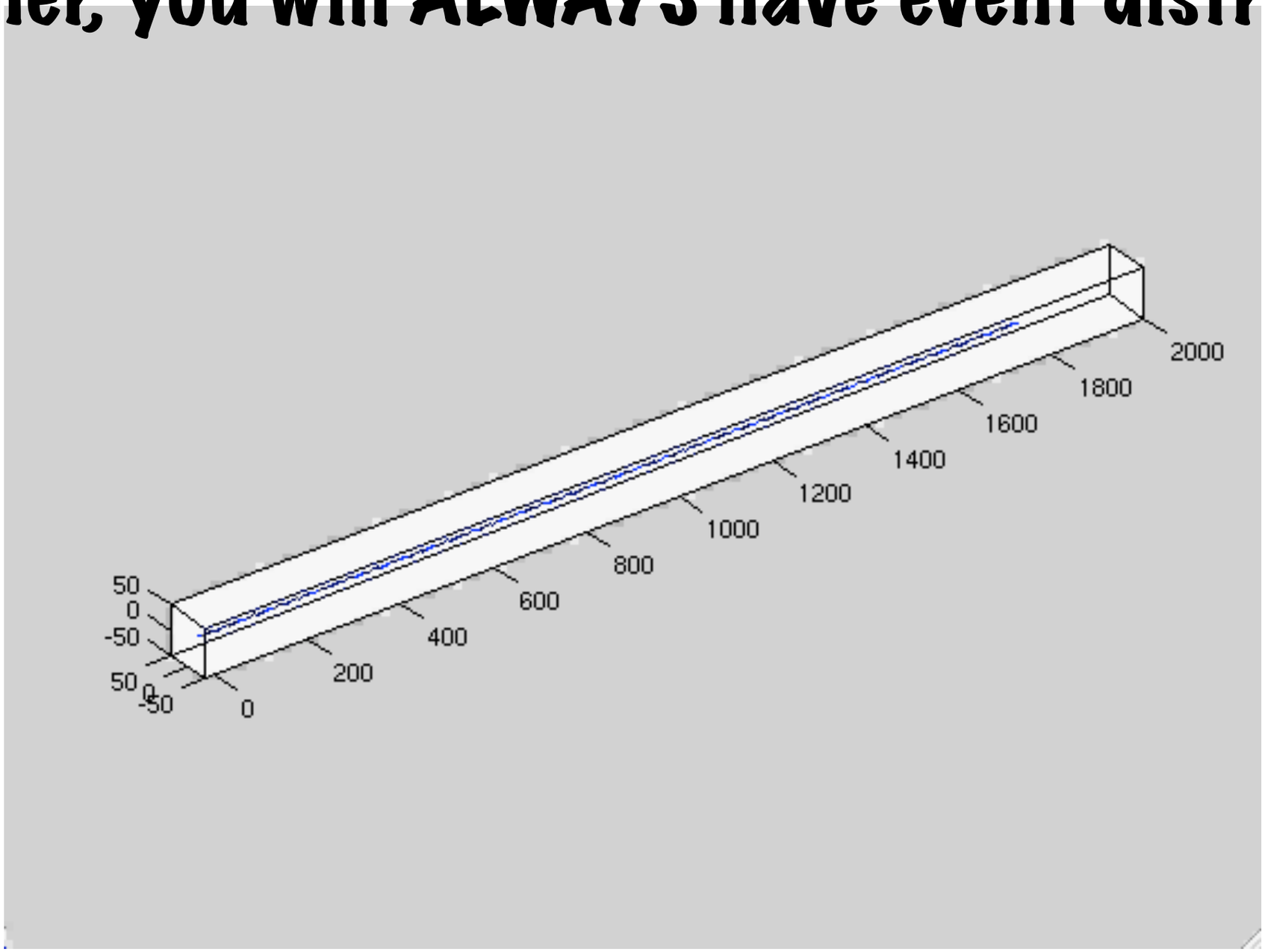


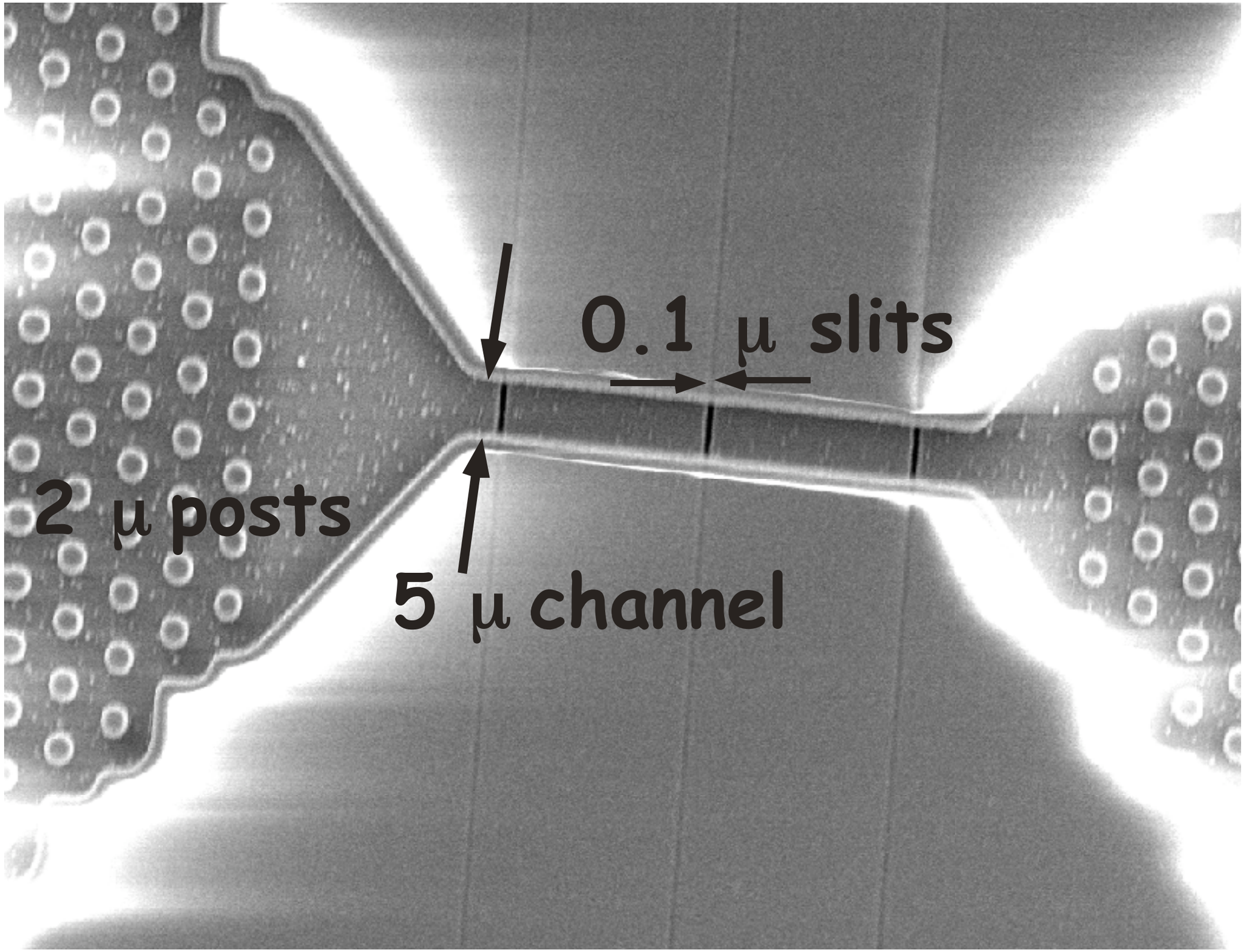
FIG. 2. The most probable blockade level, I_P, as a function of N for poly(dA). I_P depends weakly on N for long polymers ($N > 12$). In contrast, I_P has a steep dependence on N for shorter polymers ($N < 12$). The transition point ($N \approx 12$) corresponds to polymer contour length of ~48 Å. Inset: I_P is extracted from the translocation blockade distribution of the individual events, which is well fit by a Gaussian function. Error bars (standard error of the mean) are determined by evaluating I_P for 3–5 data sets of the same polymer. The line is drawn to guide the eye.

I think nanopores will run into polymer physics problems like I discussed yesterday: if you don't wring out the entropy of the polymer, you will ALWAYS have event distributions.

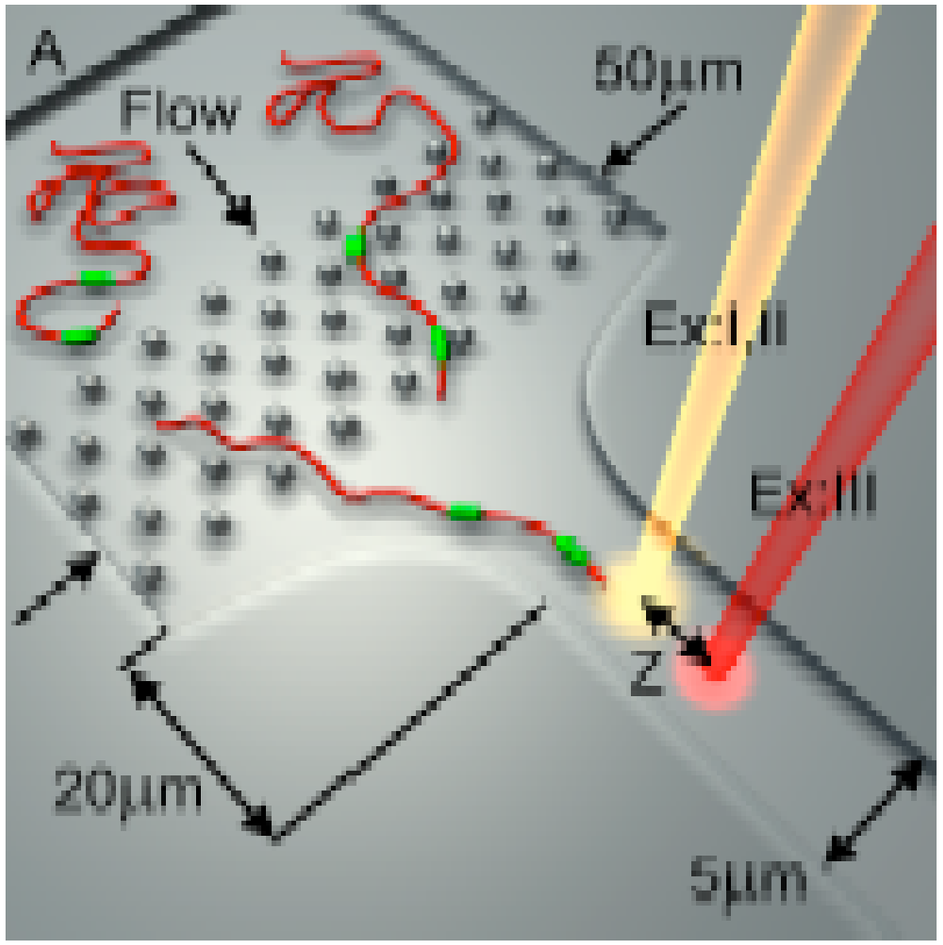


Nanochannels wring out entropy

A tempting approach is to use near-field single molecule excitation techniques, which combine the sensitivity of optical techniques with the high spatial resolution of near-field aperatures. Normally this is done by scanning the tip over a molecule...in our case the molecules are already moving along in a wafer, so we have Mohammed come to the mountain: we let the molecules move past a fixed nanofabricated slit.



Jonas Tegenfeldt, great post-doc



US Genomics, commerical scanner. Lacks optical resolution, sensitivity, elongation. My experience with US Genomics has taught me that Universities should stay the hell out of patenting and launching companies.

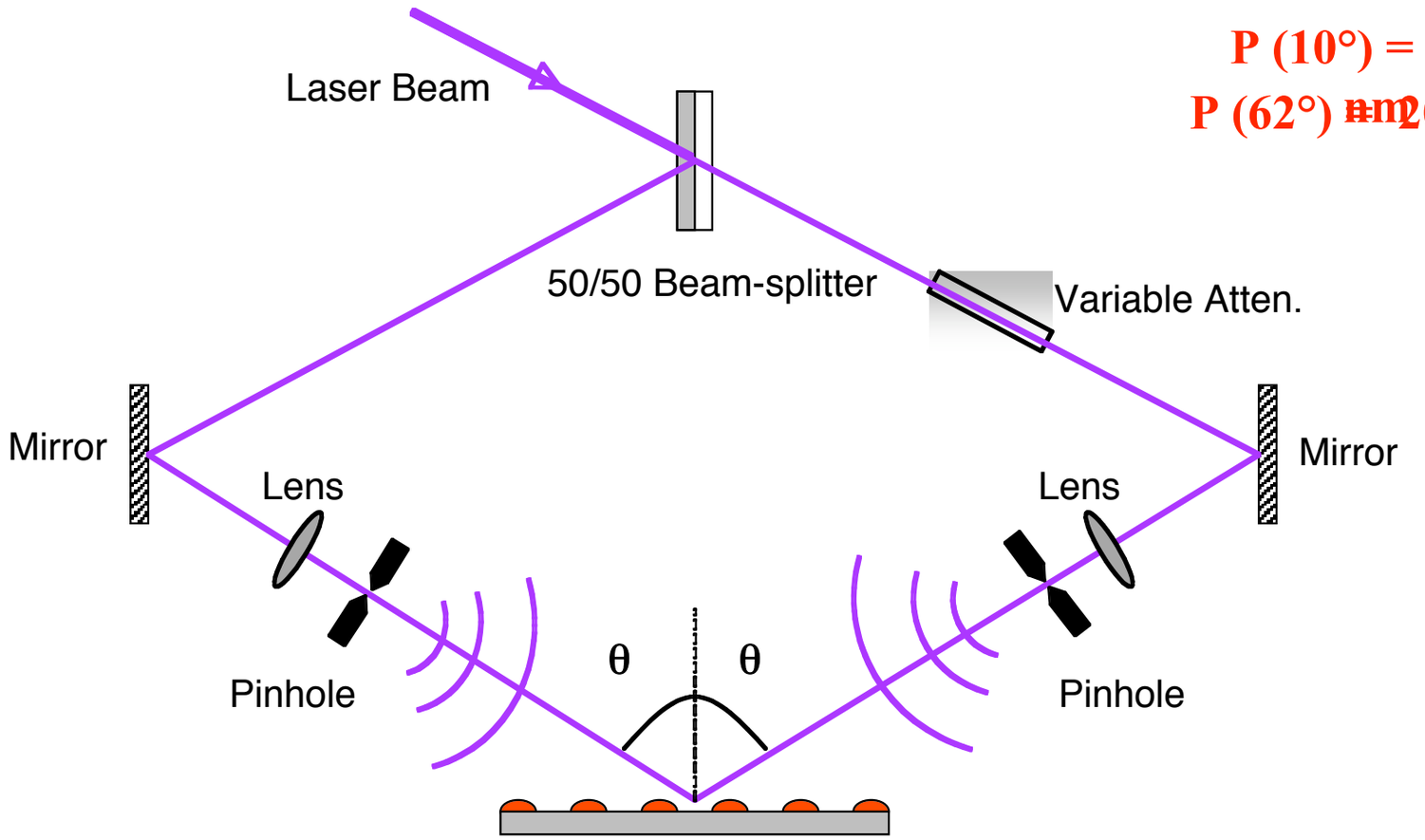
Nanoimprint Molds: Interference Lithography

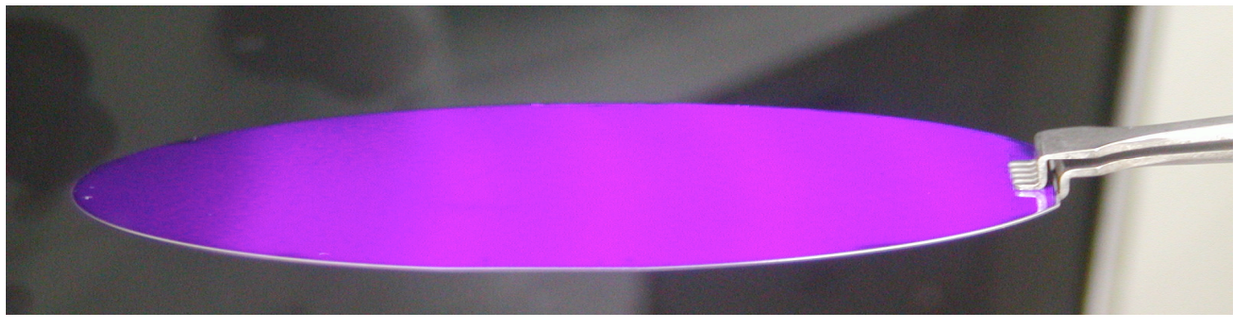
Argon Ion Laser (λ = 351 nm)

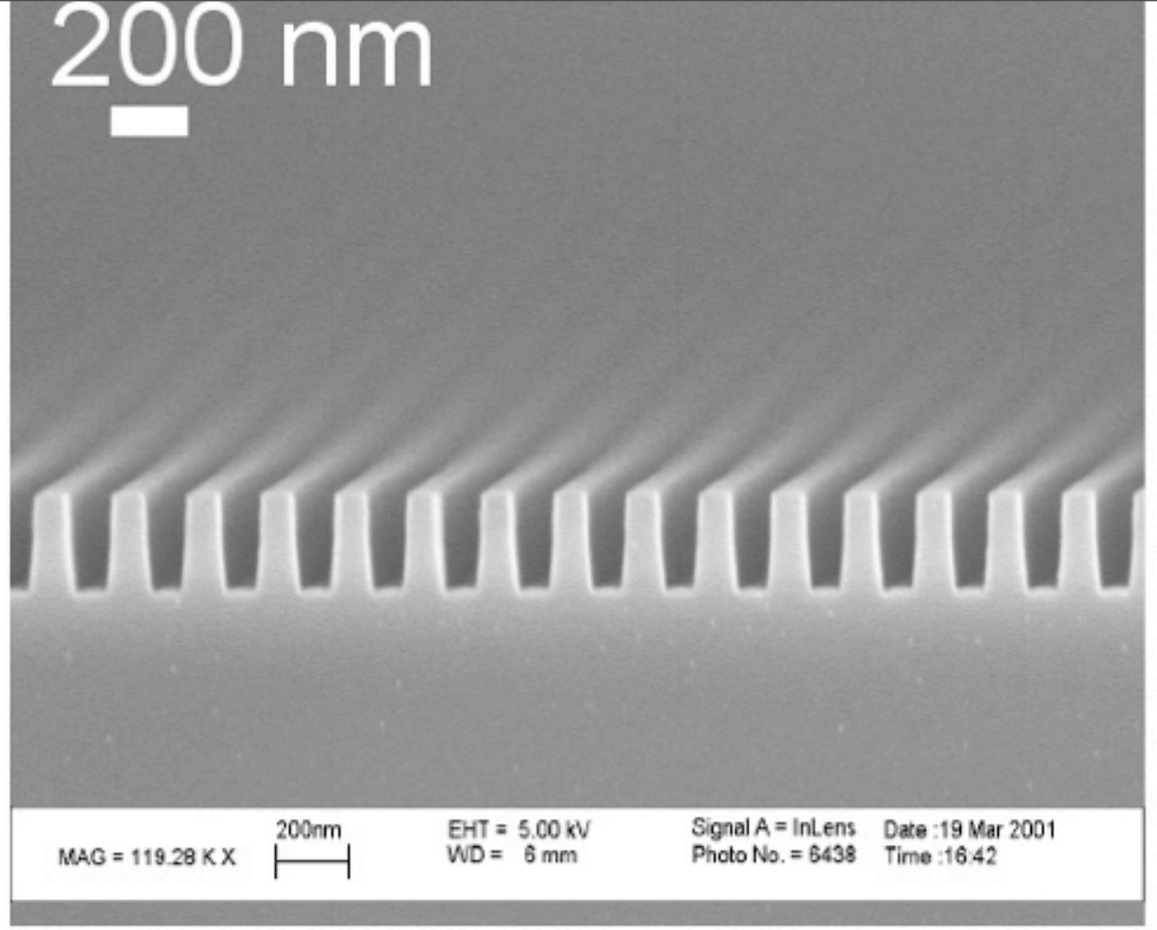
$$P(\theta) = \frac{\lambda}{2 \sin \theta}$$

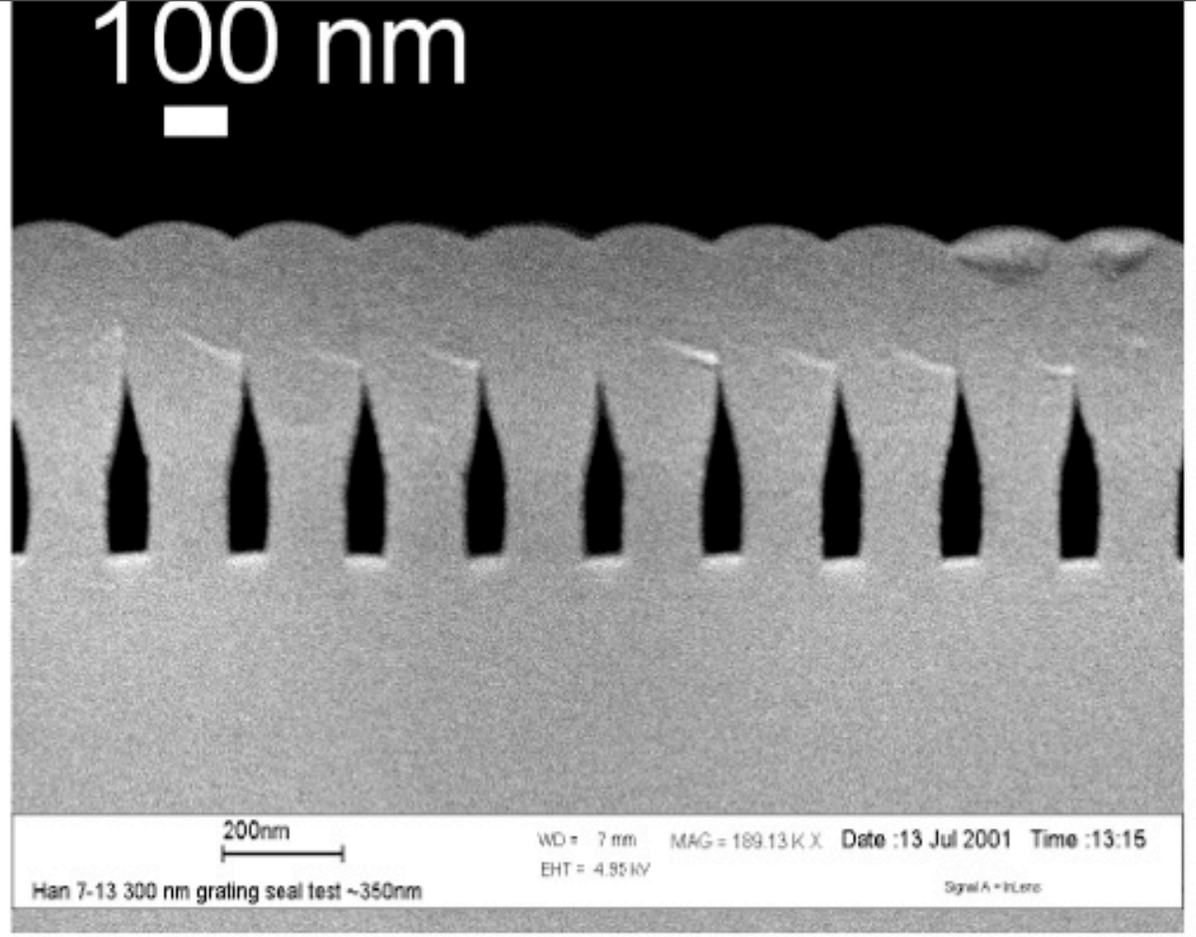
P (10°) = 1000 nm

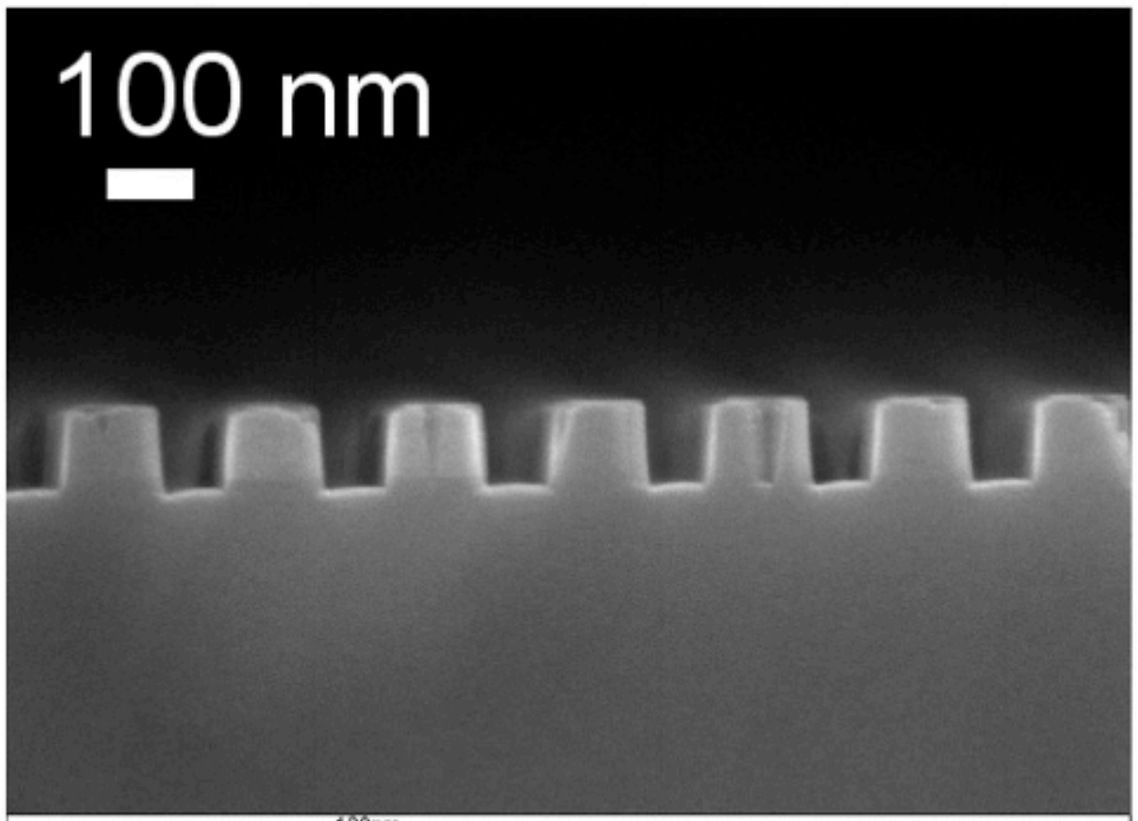
P (62°) ~~nm~~ 200 nm

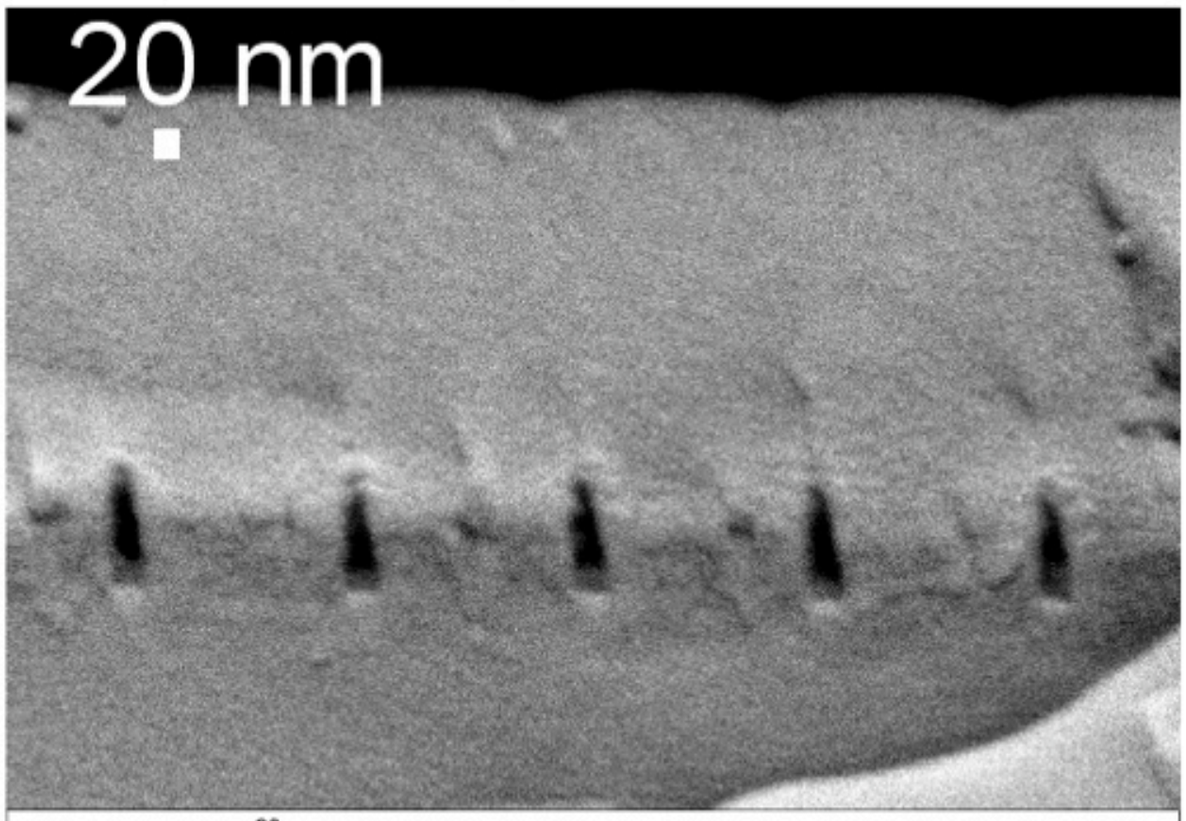





200 nm
MAG = 119.28 K X
200nm
EHT = 5.00 kV
WD = 6 mm
Signal A = InLens
Photo No. = 6438
Date :19 Mar 2001
Time :16:42

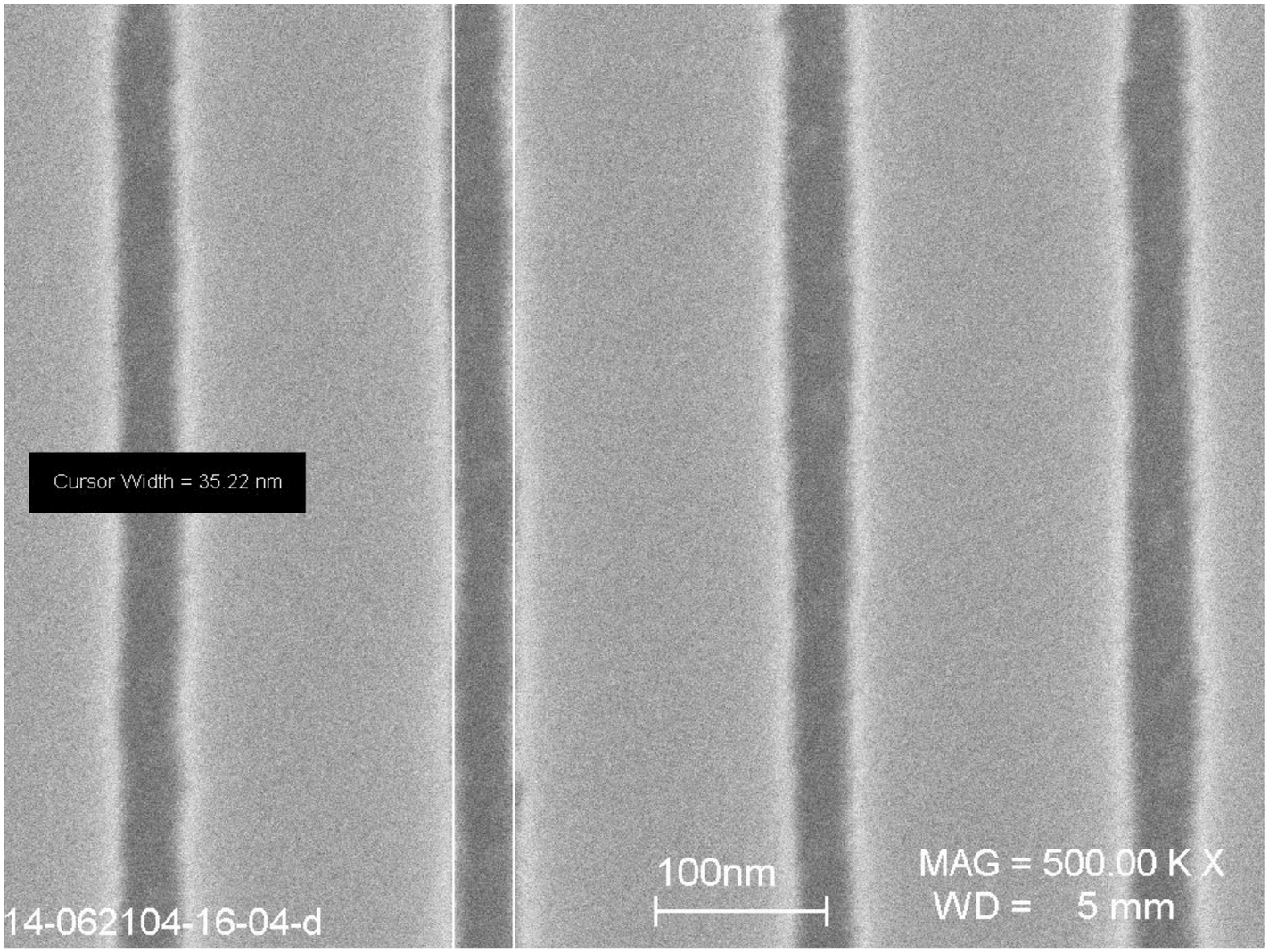

100 nm
200nm
WD = 7 mm
EHT = 4.95 kV
MAG = 189.13 K X
Date :13 Jul 2001
Time :13:15
Han 7-13 300 nm grating seal test ~350nm


100 nm
MAG = 265.53 K X
100nm
WD = 6 mm
EHT = 4.20 kV
Signal A = InLens
Date :11 Jul 2001
Time :15:16


20 nm
20nm
WD = 10 mm
EHT = 4.70 kV
MAG = 364.35 K X
Date :16 Jul 2001
Time :15:37
Han 7-16 200 nm grating seal test 80 nm deep~350nm

35 nm Channels

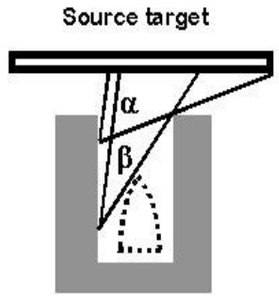
Tolerances: ~ 10 nm



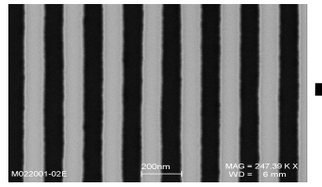
BIO NANO MATRIX

BIO NANO MATRIX

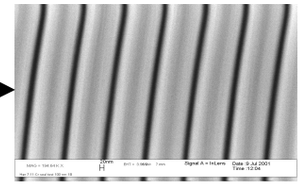
New fabrication Concept for Micro/nanofluidics



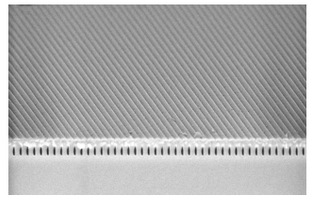
Differential Deposition (DD)

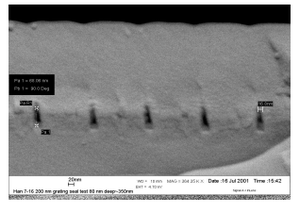


Channels ~100nm



Channels reduced to ~20 nm

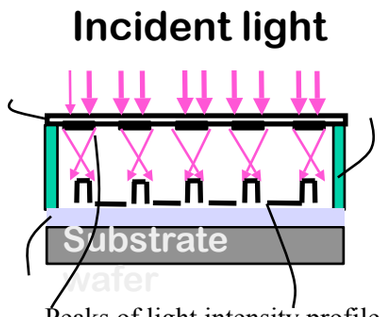




Single step, fast fabrication of enclosed nanochannel ~10nm

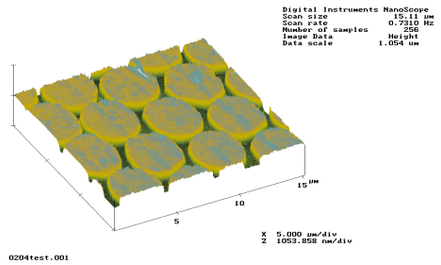
Cao H et al Fabrication of 10 nm Enclosed nanofluidic Channels (2002) Applied Physics Letters, Vol. 81, No. 1, pp174
Austin RH et al Scanning the Controls: Genomics and Nanotechnology (2002) IEEE Transactions On Nanotechnology,

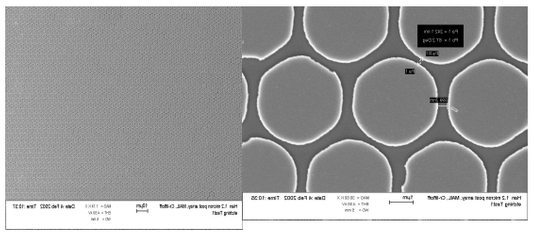
Vol. 1, No. 1, pp12



Peaks of light intensity profile on resist shifted after interference

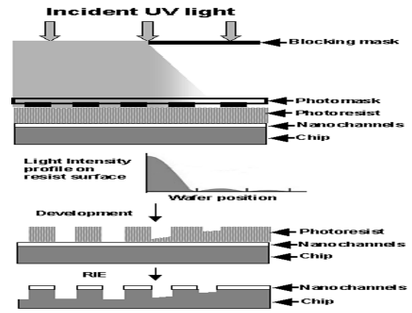
Diffraction Interference Lithography (DIL)



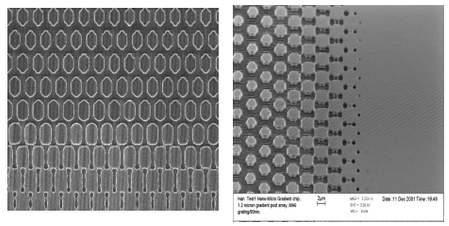


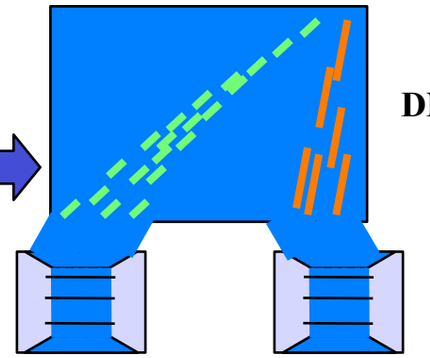
2. From Interference lithography, nanoimprinting lithography to integration of DD, DIL and DGL

New Level of Integration With Unconventional Fabrication Techniques



Gradient Nanochip





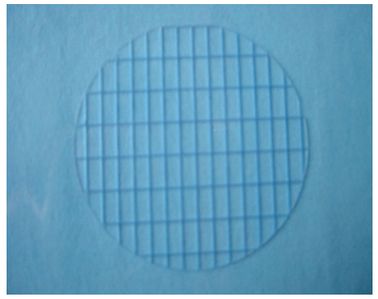
DNA Sorting

Nanochannel NF scanner

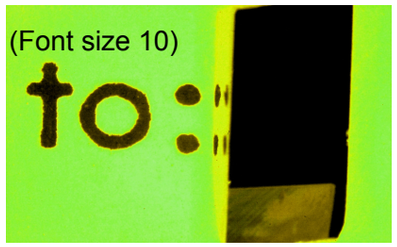
Diffraction Gradient Lithography (DGL)

Dicing

Nanosystems



Wafer Scale Fab



Cao H et al Fabrication of Gradient Fluidic Structures Interfacing Microfluidics and Nanofluidics (2002) Applied Physics Letters, Vol. 81, No. 16, p3058

3. **From single feature fluidics to hybrid gradient micro/nanofluidics**
   - **Integration of multiple fab methods**
   - **Feature sizes from 1-5μm, 100nm to 10 nm and gradient**

From Bio to Nano to BioNano
Path to Commercialization

Han Cao, Paru Deshpande, Zhaoning Yu, Stephen Y. Chou
NanoStructure Lab, Department of Electrical Engineering

In collaboration with

Jonas Tegenfeldt, Robert H. Austin
Department of Physics

Princeton University

But I think the cart is in front of the horse here. What IS the physics of polymers in nanochannels? Is it unusal? Showstoppers?

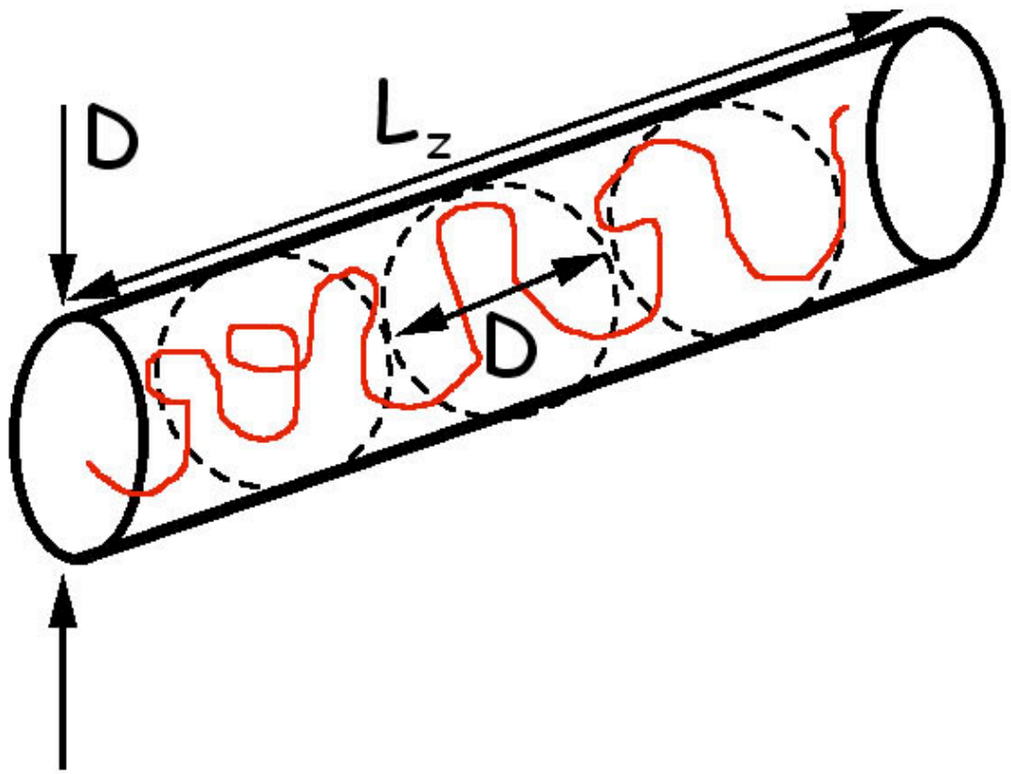
Here's an interesting polymer problem: what happens when you put a long polymer of persistence length p in a nanochannel? Lot's of surprises.

Suppose the channel is say 200 nm wide, and the polymer has a persistence length of 50 nm. The diameter of the dsDNA molecule is only about 2 nm, so most of the volume of the channel is water, since the diameter of the polymer is much less than the persistence length or the channel dimension.

You might think that the self-avoiding random walk would be an unnecessary complication.

That's wrong! Without self avoidance, the radius of gyration of a long polymer in a channel is independent of the channel diameter.

As usual, P. de Gennes worked this out long ago. The idea is very simple: a self-avoiding polymer forms incompressible "blobs" in a channel, each "blob" has a diameter equal to the diameter of the channel.



If the polymer has contour length L, the end-end length in a tube of diameter D for a polymer of width w is:

$$L_z = L\frac{(pw)^{1/3}}{D^{2/3}}$$

The entropic spring constant k of this confined polymer is:

$$k \simeq \frac{15}{4}\frac{k_B T}{L}\left[\frac{1}{pwD}\right]^{1/3}$$

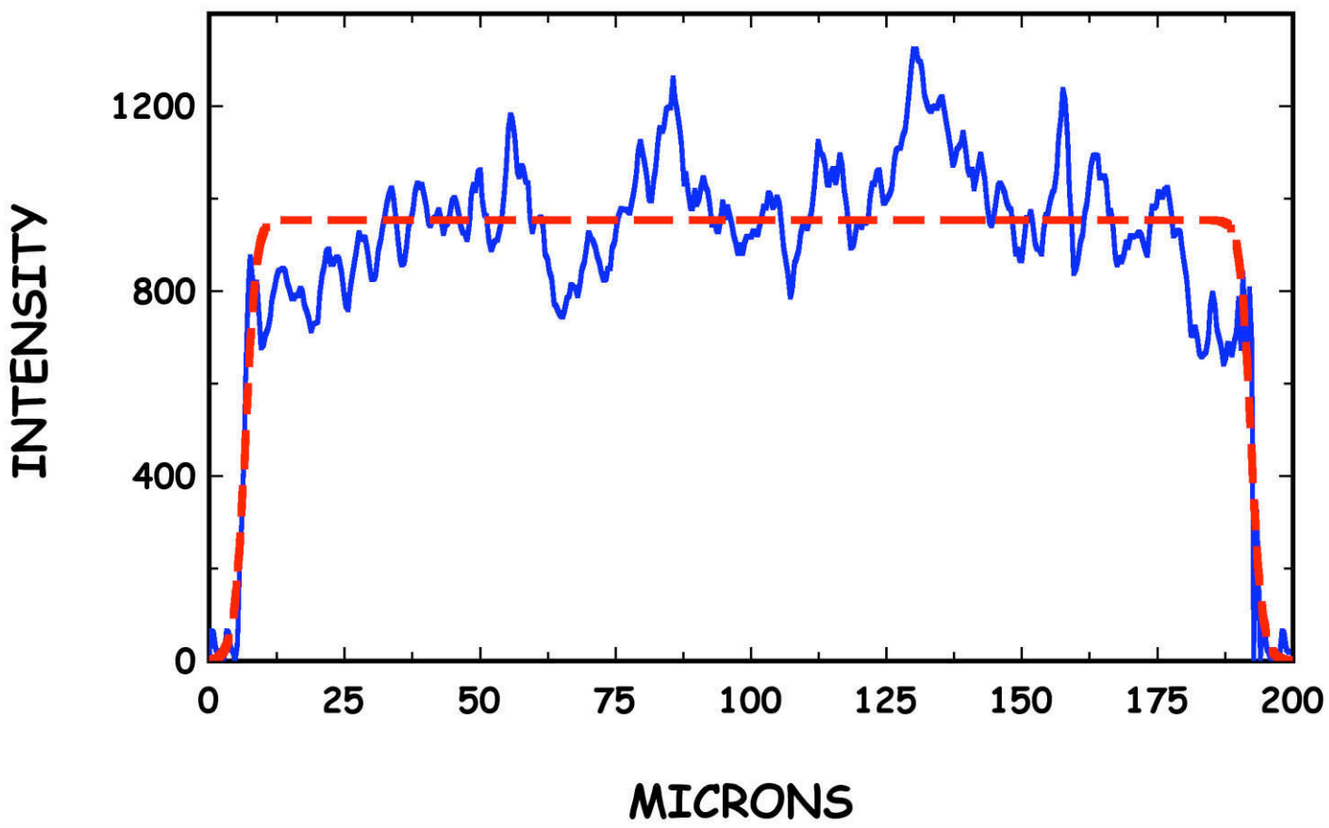
$$k \simeq \frac{15}{4}\frac{k_B T}{L}\left[\frac{1}{pwD}\right]^{1/3}$$

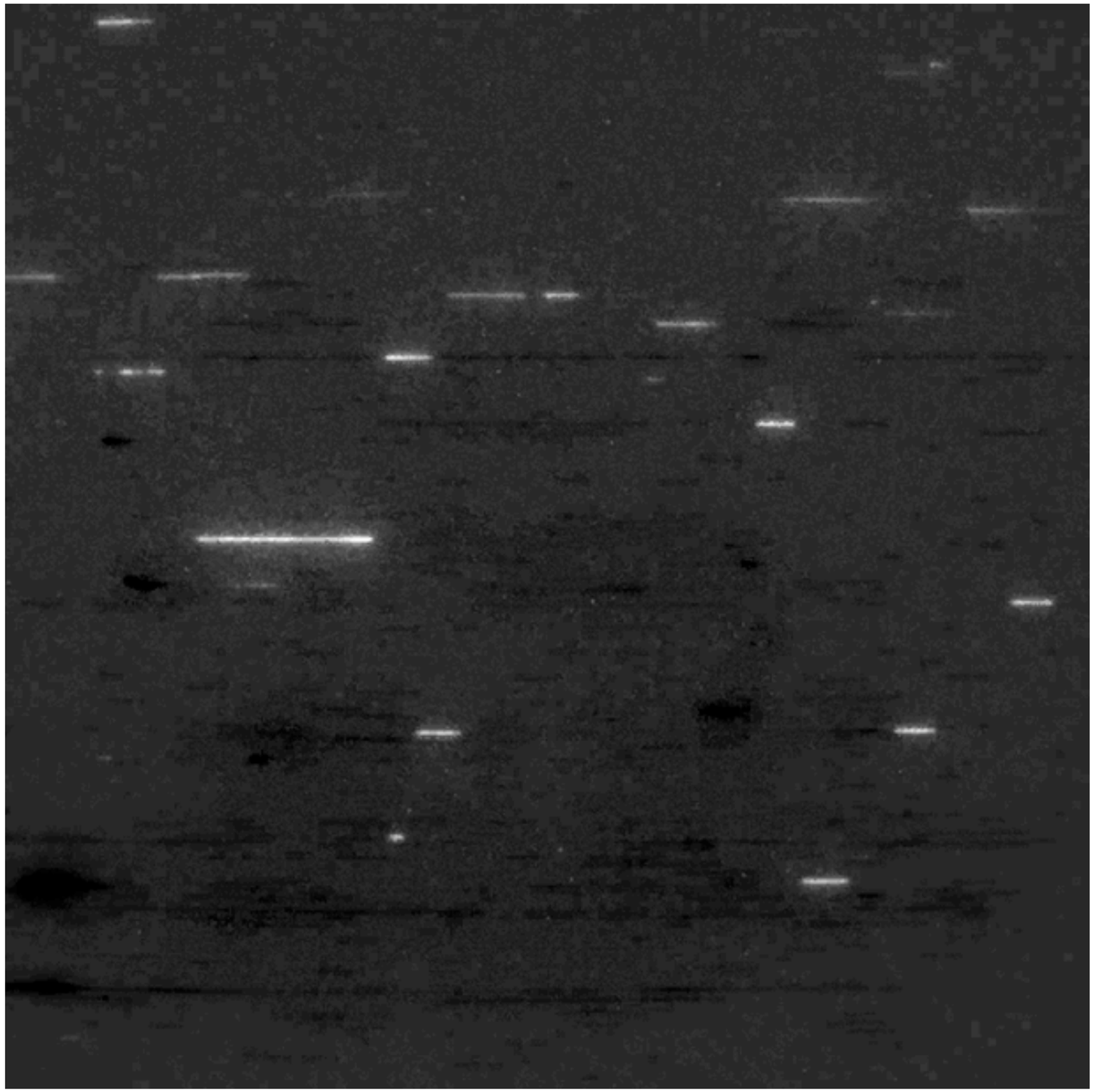
This is an interesting expression, and it addresses some issues about attempts to use nanopores to size/sequence DNA molecules.

As D, the width of the nanochannel decreases, the spring constant gets larger and larger: it gets stiffer. This means that the amplitude of the length thermal "noise" as the DNA molecule flickers back and forth decreases. This is good.

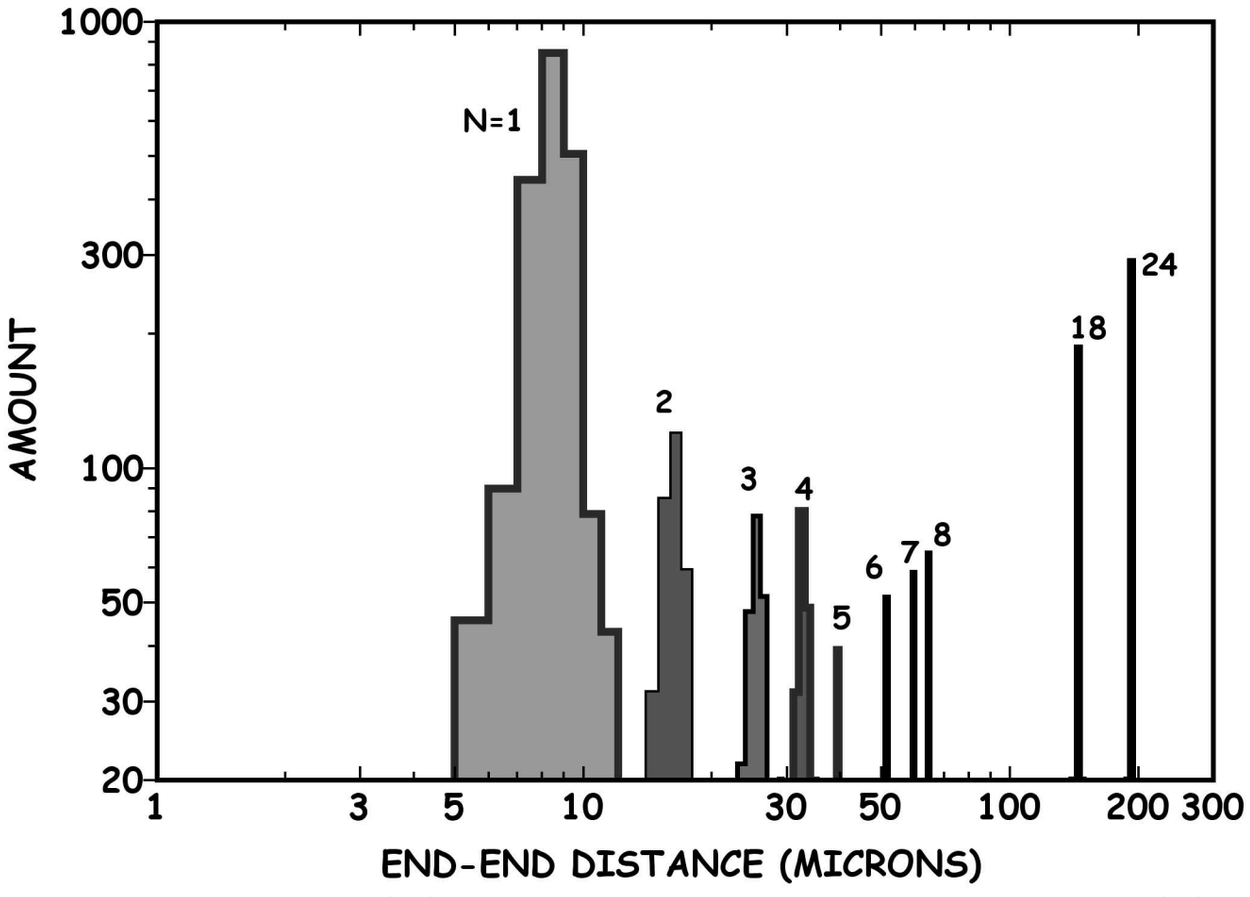
The dynamics of confined polymers is extremely important. Why? Because, the longer you stare at the dynamics of these molecules, the more you know about the standard deviation of the mean and hence the length, to much better than the wavelength of the observing light.

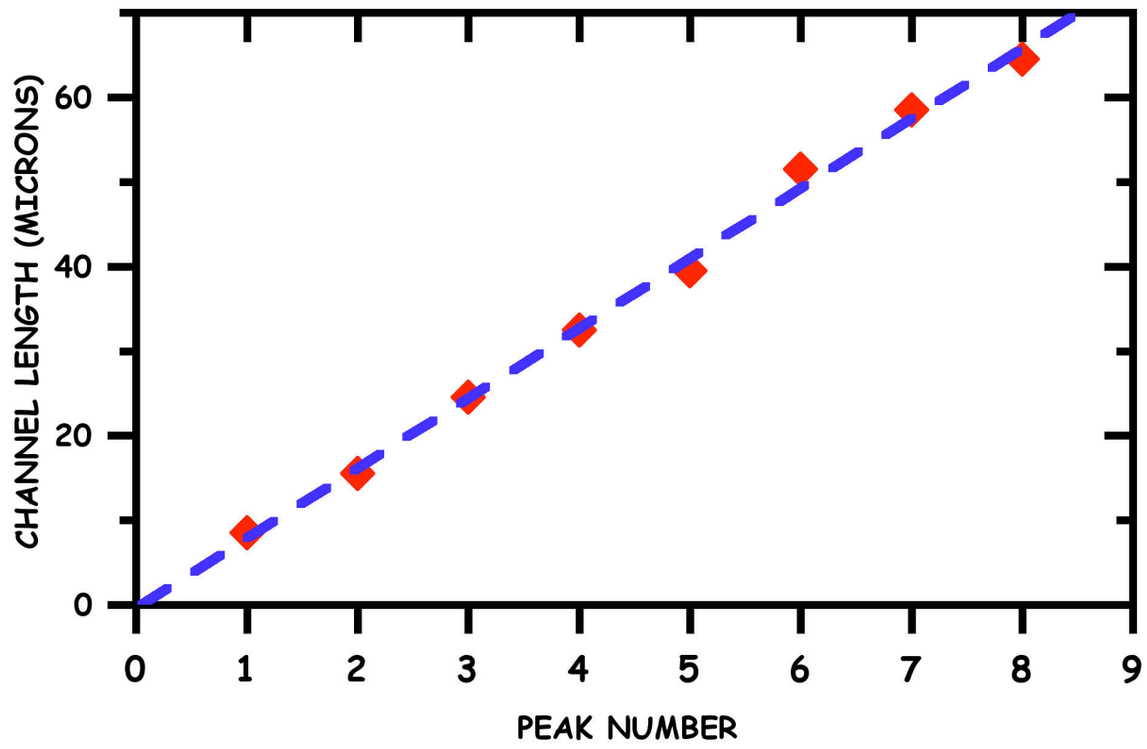
$$I(z) = \frac{I_o}{2}\left[Erf\left(\frac{z}{\sigma_o\sqrt{2}}\right) - Erf\left(\frac{z-L_z}{\sigma_o\sqrt{2}}\right)\right],$$

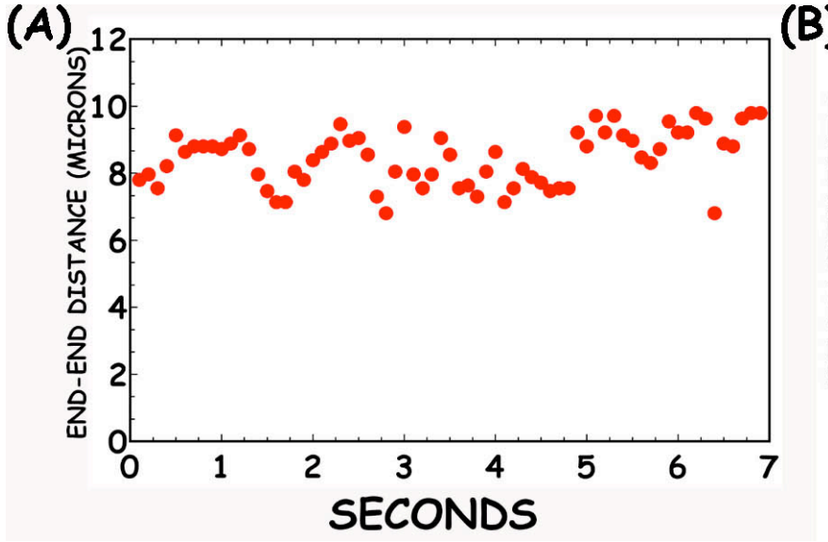


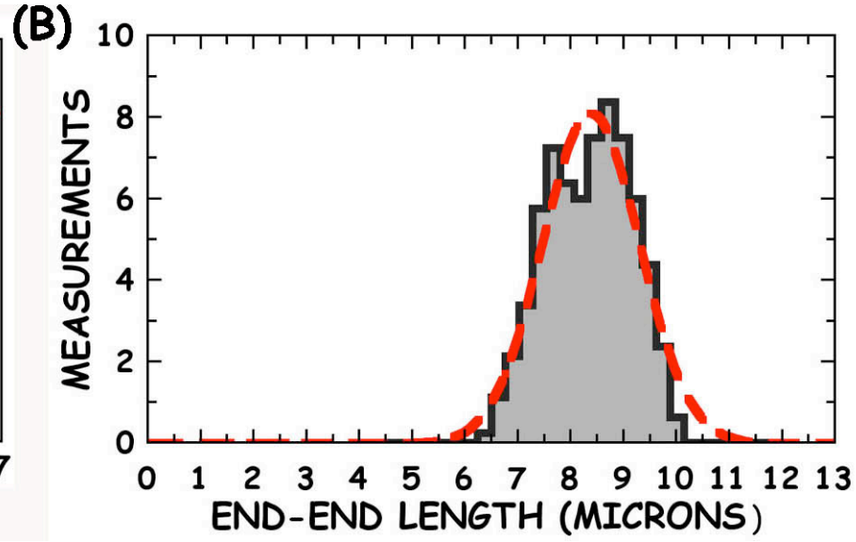


The idea is to run the molecules in, stare, get the length, put new ones in, and build histograms quickly.



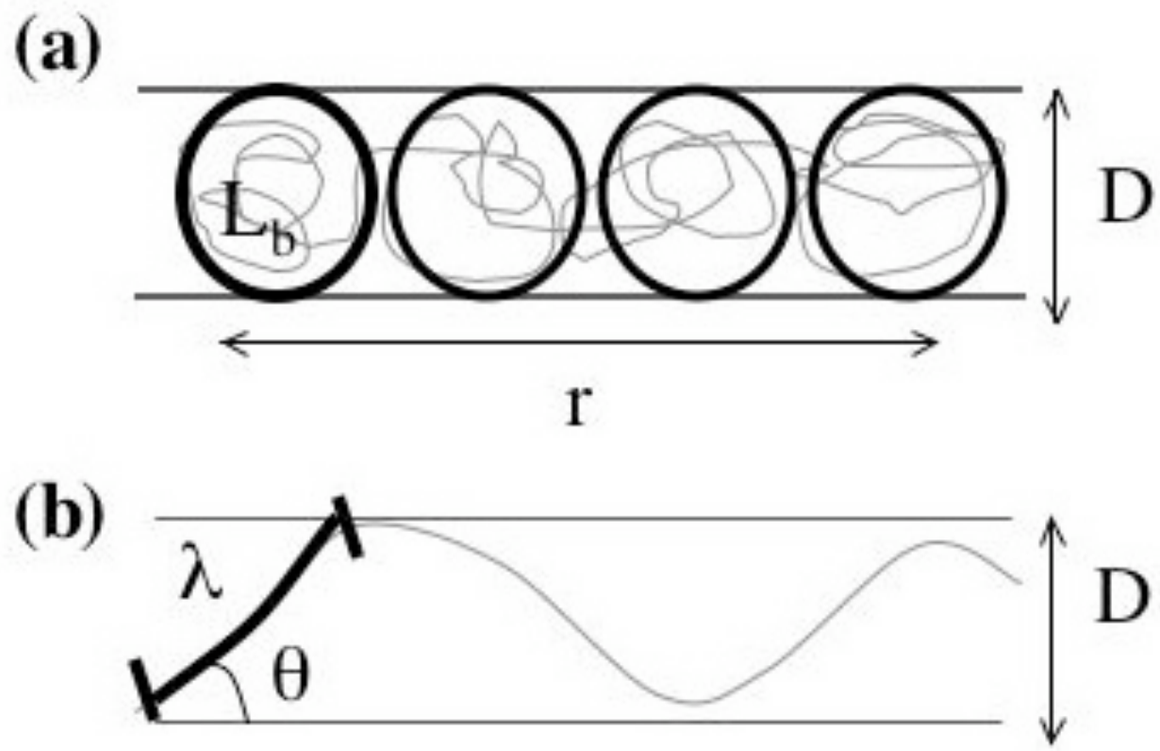




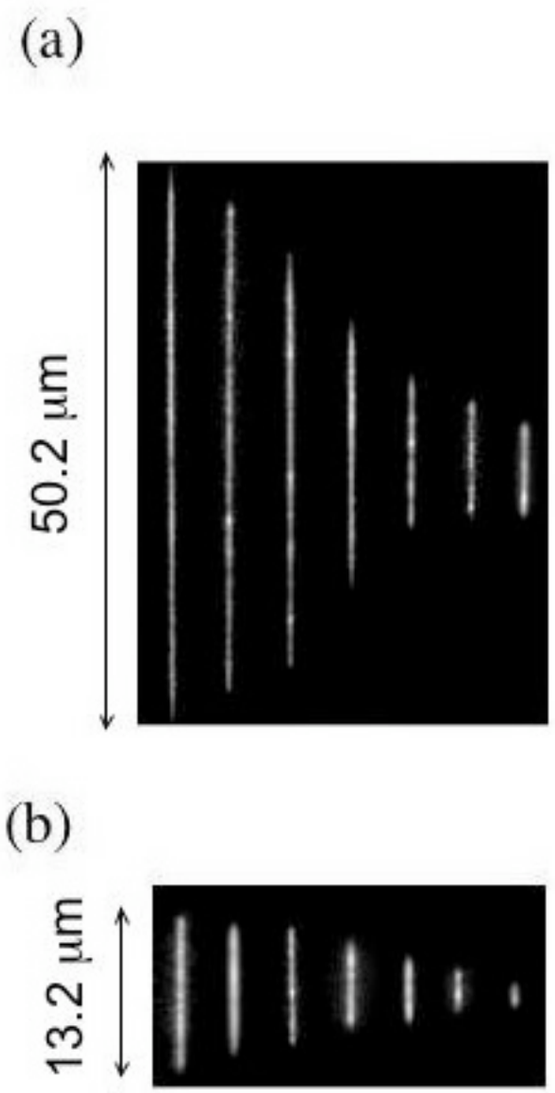


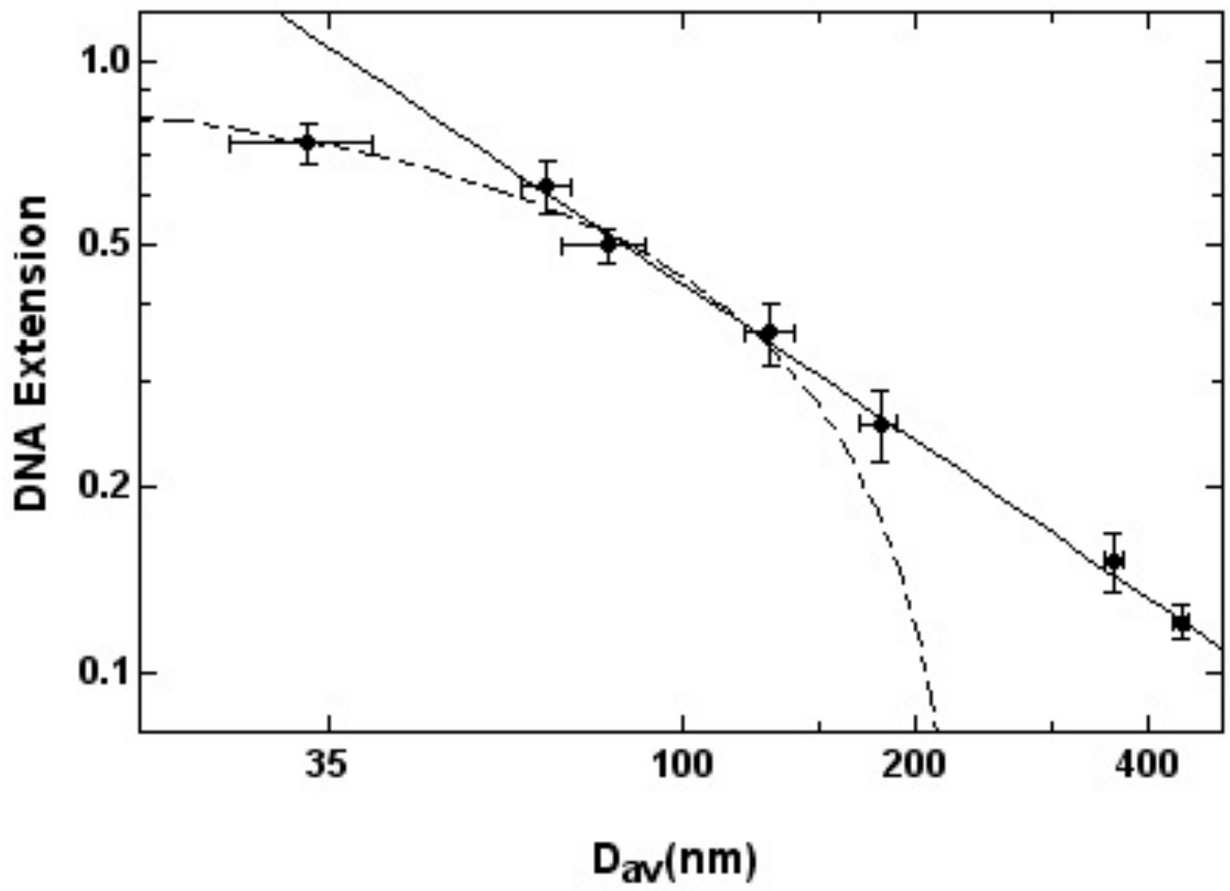
In principle, in a few seconds you can know the length of the dsDNA molecule to 100 bps. Not sequencing, but mapping.

That work came from 100 nm channels, 2 persistence lengths wide. We have pushed down since them to channels of diameter D below 1 persistence length. At this scale, de Gennes scaling breaks down, and we enter a new area, the Odijk limit where now it is now entropy but elastic deformation that determines the statistics.

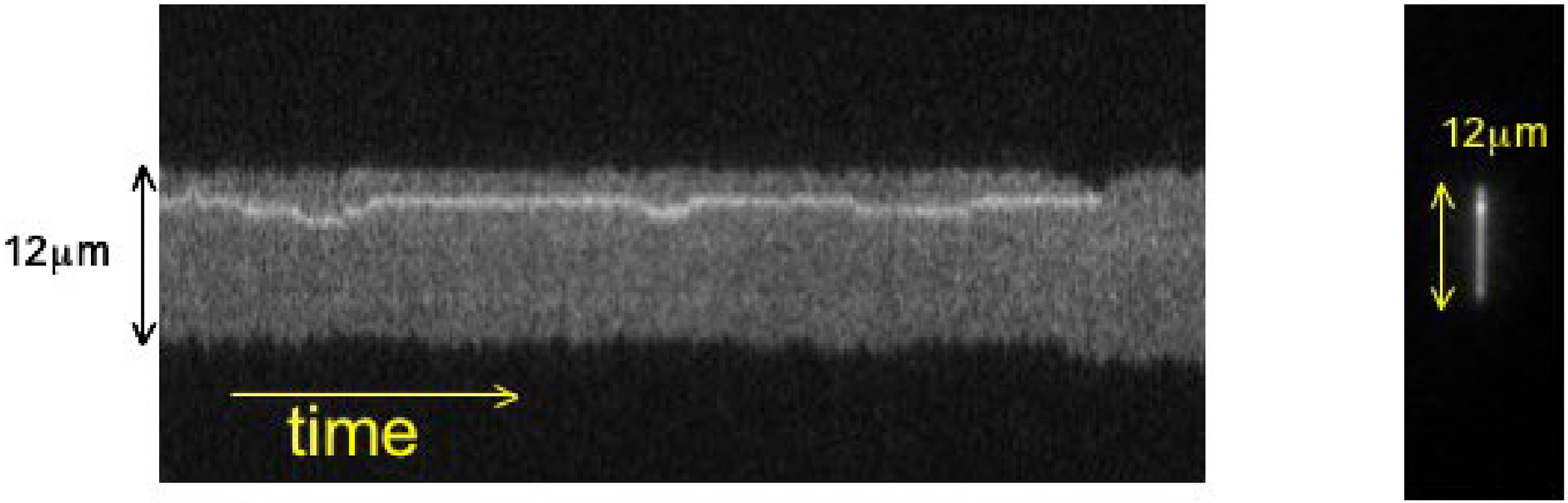


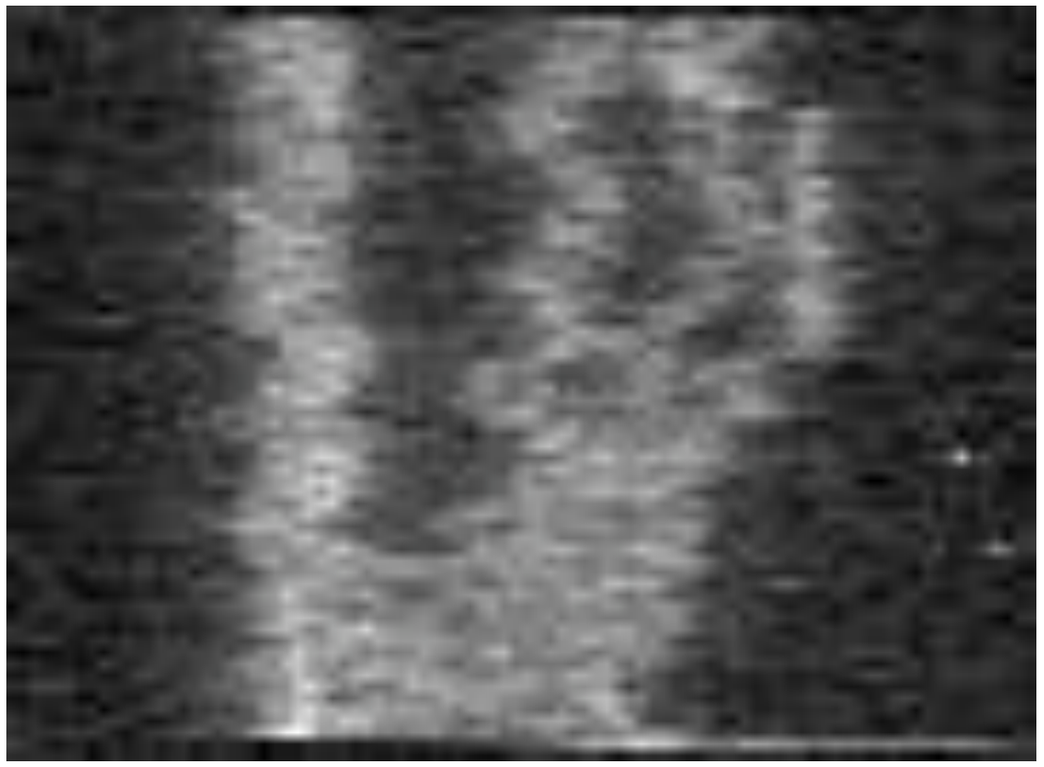
$$L_z = Lcos(\theta) = L\left[1 - A\left(\frac{D}{P}\right)^{2/3}\right].$$





The dynamics of confinement and the influence of self-avoidance at this length scale may have some real surprises for us....solitons, anyone?



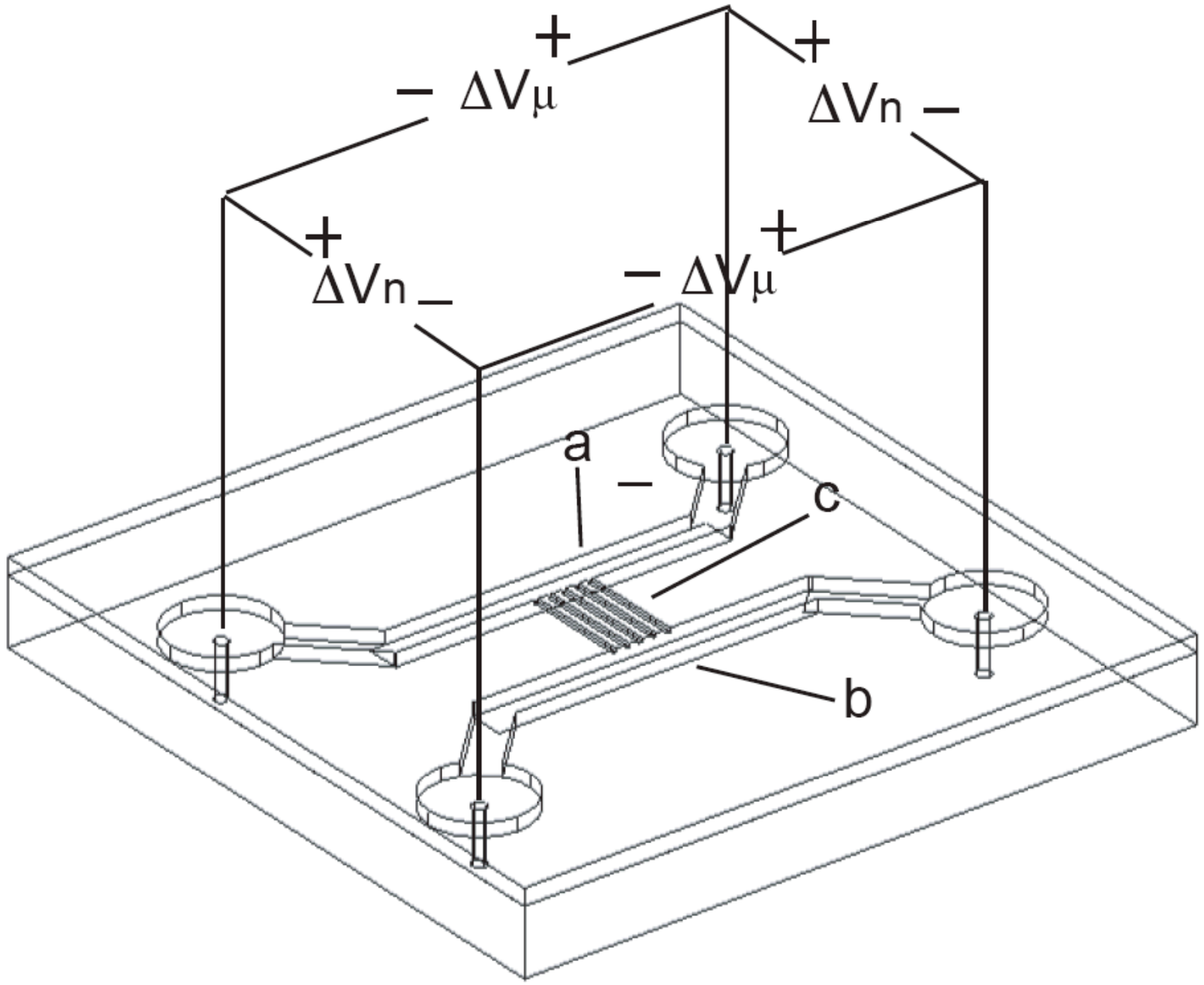


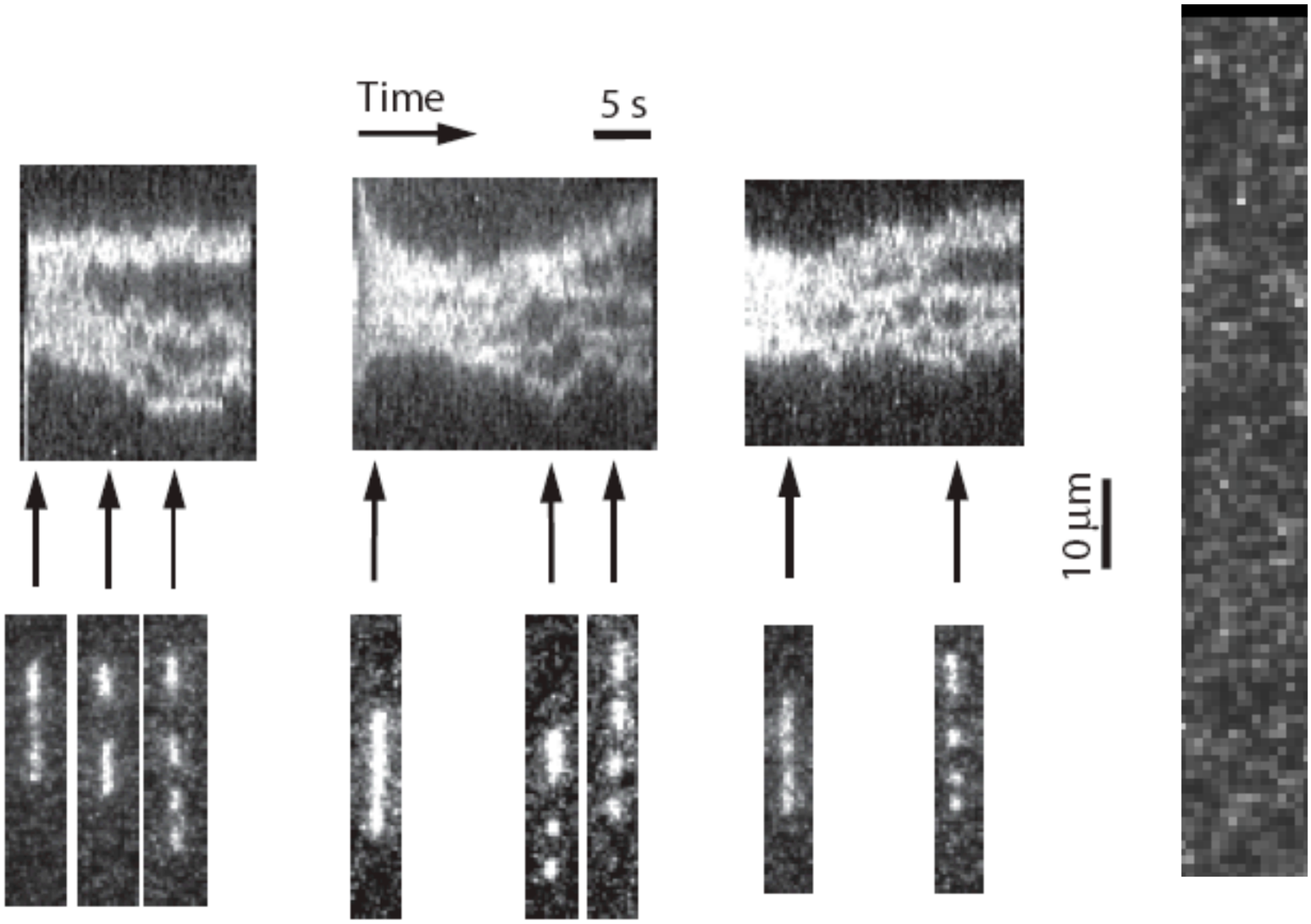
My post-doc Robert Riehn with help from a molecular biologist Manchun Lu has even started watching the time-resolved cutting of genomic length DNA by restriction enzymes, in nanochannels.

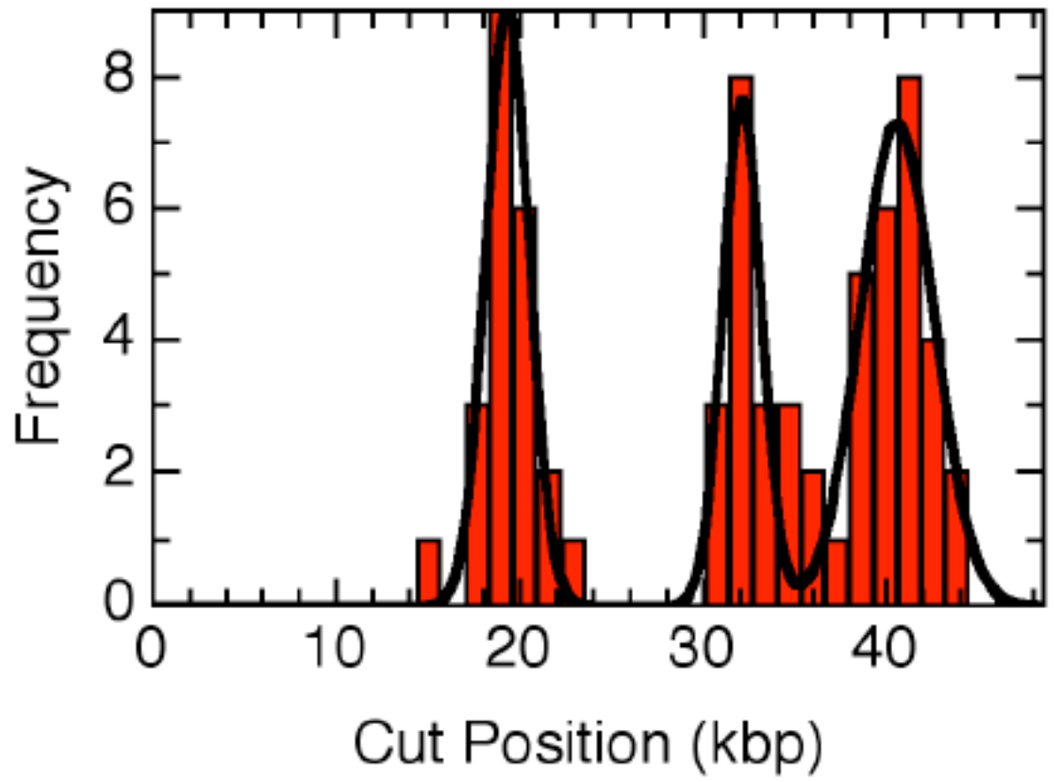
Restriction mapping of DNA with endonucleases is a central method of modern molecular biology. It is based on the measurement of fragment lengths after digestion, while possibly maintaining the respective order.

We decided that perhaps we could bring restriction enzymes into these nanochannels and cut genomic length DNA molecules at precise sites. Since we would observe the cutting directly, there would be no scrambling of the order of the cut sites, and so we could do a direct physical map of a DNA of genomic length.

This has been a hard road to go down!


+
− ΔVμ
+
ΔVn −
+
ΔVn −
+
− ΔVμ
a
−
c
b


Time
5 s
10 µm



Sequence	19.4	31.6	39.9
Histogram	19.3 ± 1.2	32.1 ± 1.0	40.6 ± 2.0
Weighted Average	19.9 ± 1.3	32.8 ± 1.3	40.7 ± 1.7

As we push down to the true nanometer scale, we increasingly have to fight for signal/noise to do single molecule detection because of stray fluorescence, raman scattering, etc.

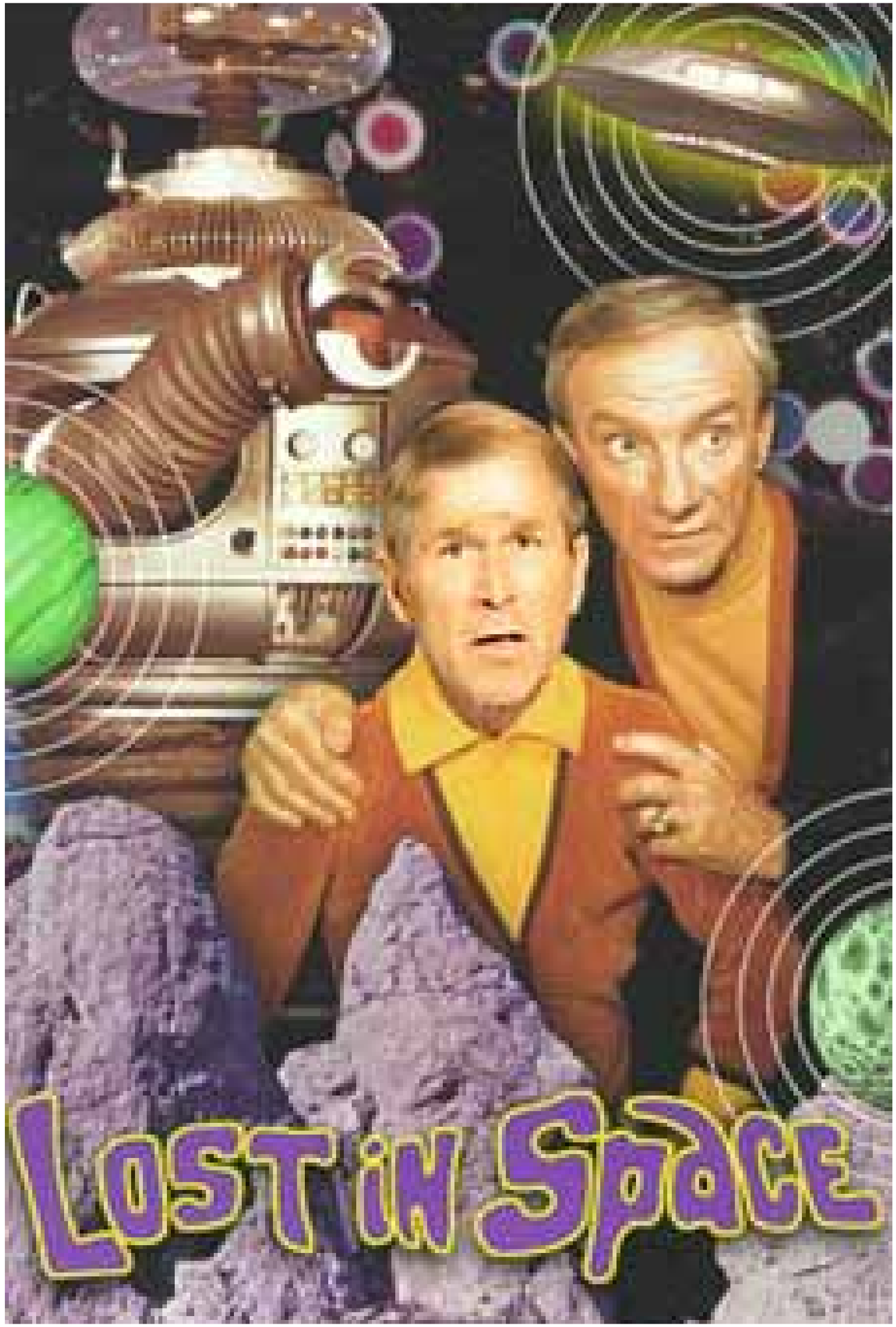
One really has to do near-field excitation in order to reject background pickup, even with the best cooled, intensified CCD cameras and high N.A. objectives.

IV. Some recent work on observing single transcription factors bound to DNA.

Of course, we would also like to image the transcription factors I talked about at the beginning of this talk with as high a resolution as we can measure the DNA in the nanochannels. My excellent post-doc Yan Mei Wang has been doing this work. ALL of these people by the way are converts to biological physics.

Absolutely critical collaborations here with Ted Cox in Molecular Biology, and Shirley Tilghman who has been lost to the Dark Side (there are several Dark Sides by the way).

One of the great challenges in protein-DNA interaction study is to obtain information from a single protein-DNA molecule, rather than an ensemble of millions of molecules, as is the case for present day technologies. In order to understand protein-DNA interactions at the single molecule level, the single protein molecules must be imaged with high resolution, as to resolve two key fundamental issues: one is to obtain the number of proteins bound; the other is to localize the bound proteins on DNA, which would require that the DNA be extended in a linear manner. We use our nanochannels for this.


LOST IN SPACE

Endonucleases and transcription factors (they control gene expression) bind very tightly (k_d about 10^{-11} M) to highly specific binding sites.

In E. coli, the doubling time of the cell during exponential growth is about 20 minutes. The E. coli genome is about 4 million bases.

Is there enough time to find the right site, or are they Lost in Space?

If the proteins do a random walk in 3-D space to find their binding sites, this is a bimolecular diffusion problem.

$$\frac{d[MB]}{dt} = -k[Mb][CO]$$

$$[\text{Mb}] + [\text{CO}] = [\text{Mb-CO}]_o + [CO]_o = Constant$$

$$k \sim 4\pi D_3 l_{target} \sim \frac{4k_B T l_{target}}{3\eta a}$$

Written in terms of liters/sec-moles we get:

$$k \sim 4\pi D_3 l_{target} \sim 10^8 liters/sec - moles$$

This is a problem for many proteins (such as transcription factors) which need to find a specific sequence. If the sequence is about as large as the diameter of the protein and we have on the order of 1 protein/cell (10 pM). then, the rate at which the protein can find the sites is roughly:

$$\frac{dN}{dt} \sim -k[P][N = 10^{-3} sec^{-1} N$$

This is bad: 10^3 seconds is about 1 hour.

Ok, suppose Nature is clever (is She intelligent?) and places genes for proteins "right next" to where they are needed. Doesn't this solve the problem?

The problem is, as Thomas Wolfe pointed out, you can never go home again, esp. if you can't find a job after grad school.

3-D space is really big, and random walks are very non-intuitive. For true POINTS, the probability of returning to the origin in a random walk is .34, in 3-D, even at t= infinity. That means that if you miss a point in your first pass, you never get another chance. You are Lost in Space. How do you get to Carnegie Hall? Stay out of 3-D.

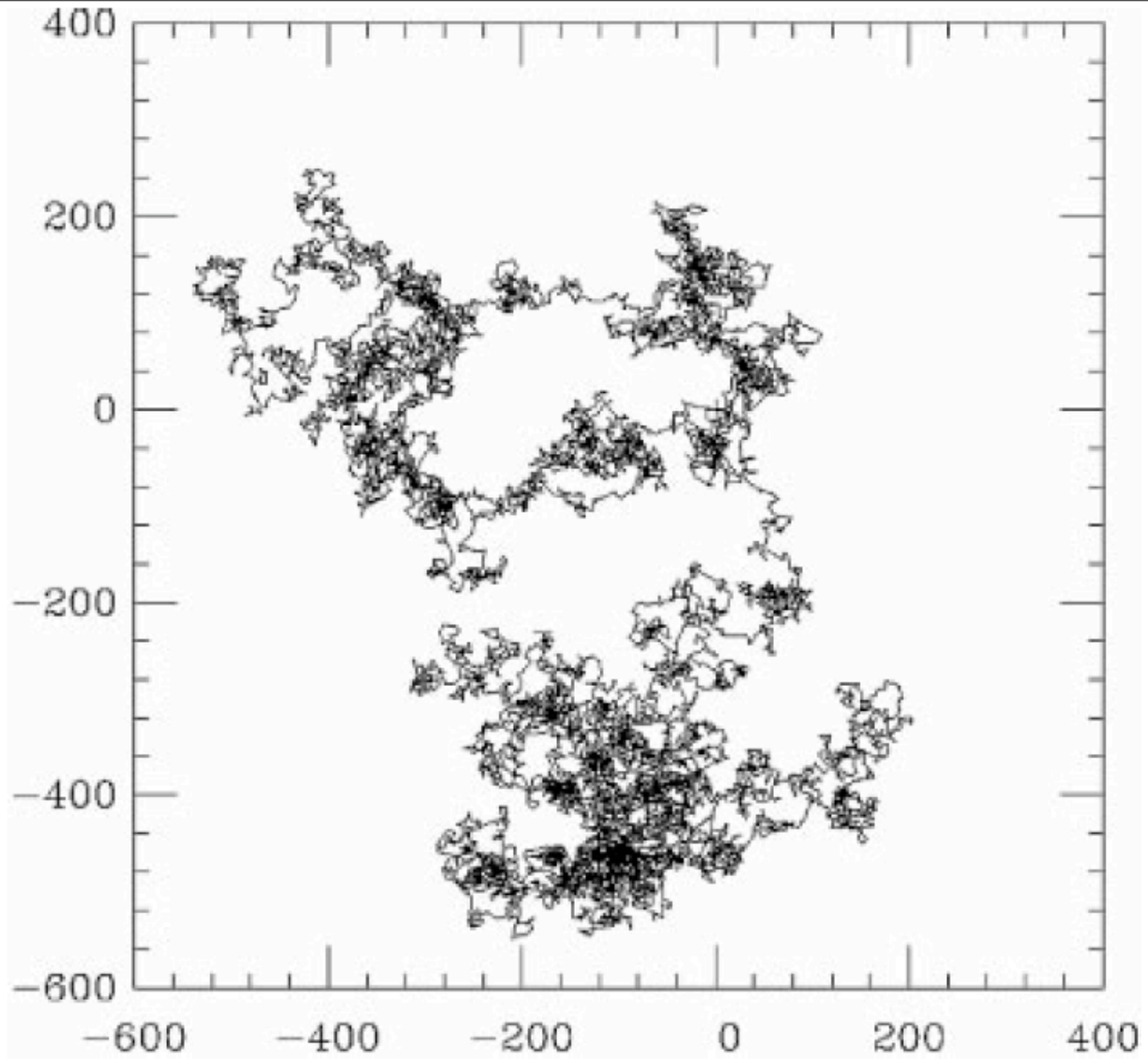


Figure 2. The trajectory of a diffusing protein is a 'random walk'. A 3-D random walk of 10^6 steps, each of length 1, is shown here as a projection on a 2-D plane. The end-points of the walk are at (0,0) and at (–300,–300). The overall size of the random walk is roughly the square root of the number of steps, or ~1000 steps, as expected from the mean square law $R^2 = Dt$. Although this law holds quantitatively only in statistical terms when applied to many random walks, the overall size of one random walk is usually reasonably well estimated by the mean square law. It is important to note that the random walk has many voids; it does not completely explore the 3-D region it extends through.

http://math.furman.edu/~dcs/java/rw.html

3040–3052 *Nucleic Acids Research, 2004, Vol. 32, No. 10*
DOI: 10.1093/nar/gkh624

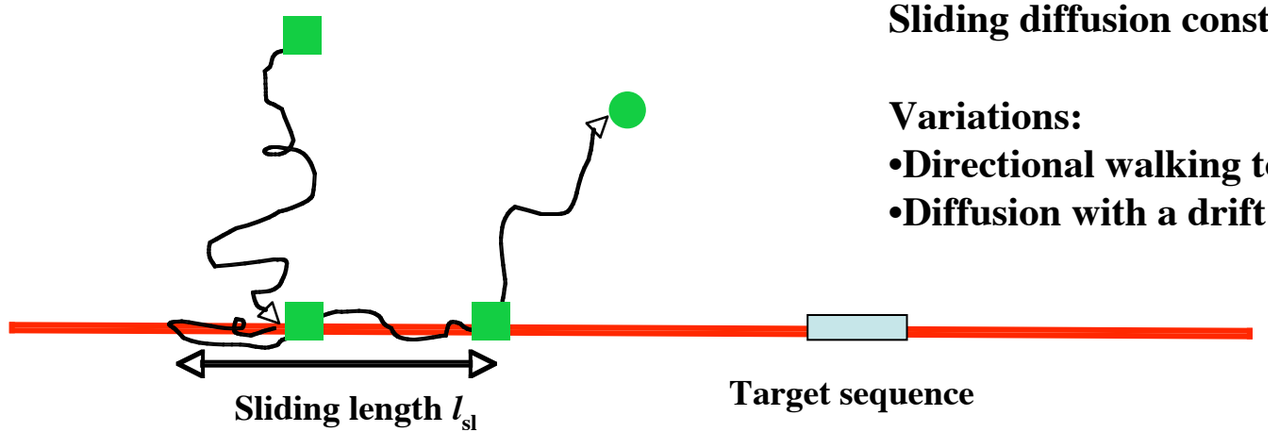
SURVEY AND SUMMARY

How do site-specific DNA-binding proteins find their targets?

Stephen E. Halford* and John F. Marko[1]

Two facilitated diffusion models:

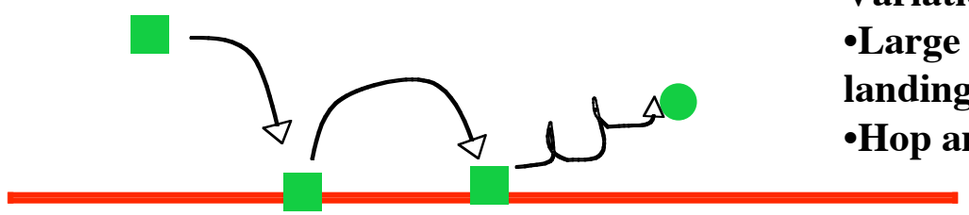
1. 1D-Sliding along nonspecific DNA (Berg, 1976)



Sliding diffusion constant D_1

Variations:
- **Directional walking towards target**
- **Diffusion with a drift**

2. Proteins hop along DNA, loosing contact with DNA during hops (Berg, 1976)



Effective 1D diffusion constant D_1

Variations:
- **Large hops with uncorrelated landing locations**
- **Hop and slide**

But, life does in fact go on, and in vitro measurements of lacI binding rates to specific sites give k up to x1000 GREATER than the diffusion limit! How can this be? It seems like moving faster than the speed of light.

Life is more secure in 2D and particularly in 1D unless you are a string theorist.

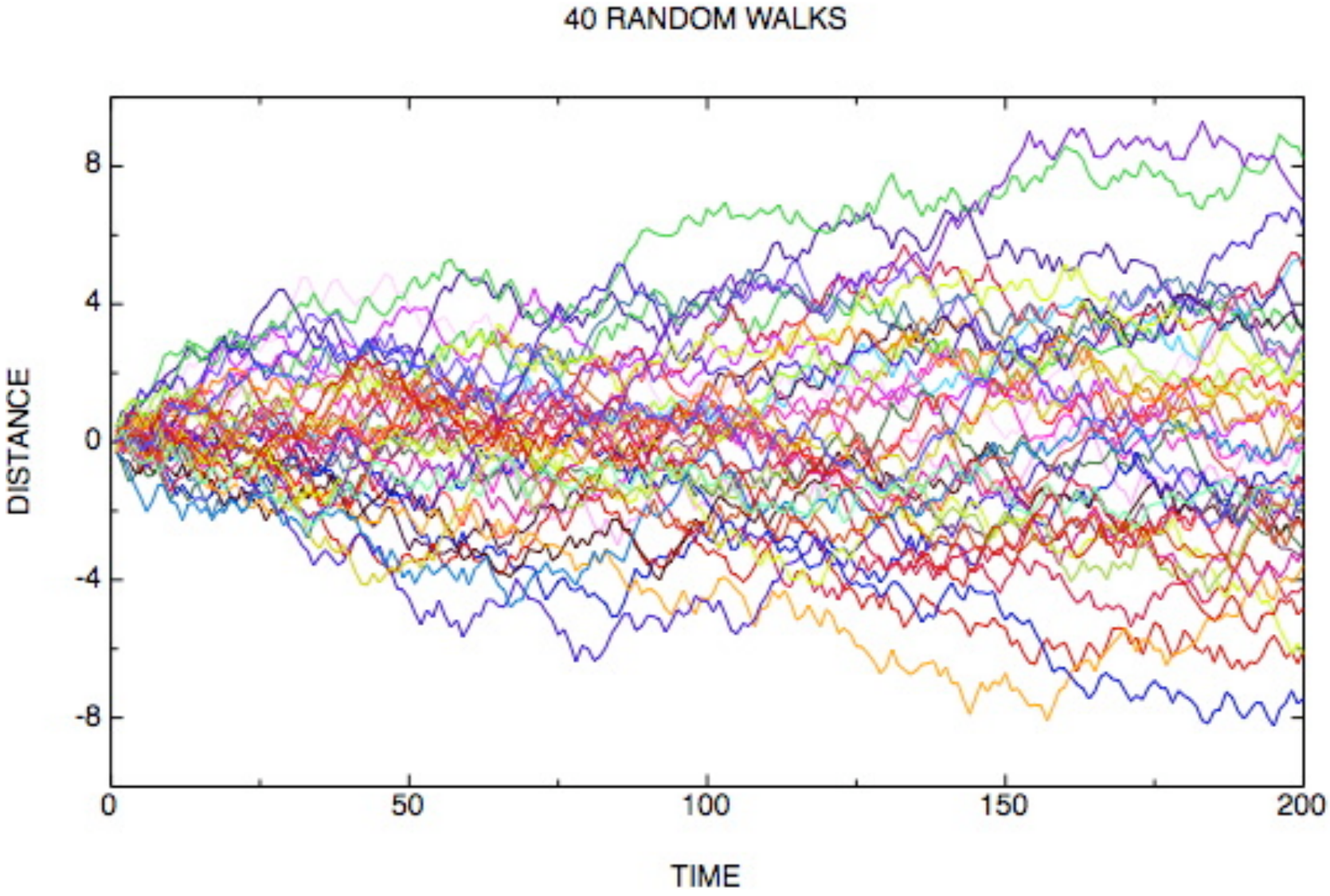
In 1D the odds of returning to the origin in a random walk are, not surprisingly, 1. A 1-D walks covers ALL of a line in infinite time.

So, it has been suggested that proteins find their sites by a combination of 3D diffusion (to get away from the site of production) and 1D diffusion along the DNA (because you CAN get home in 1D).

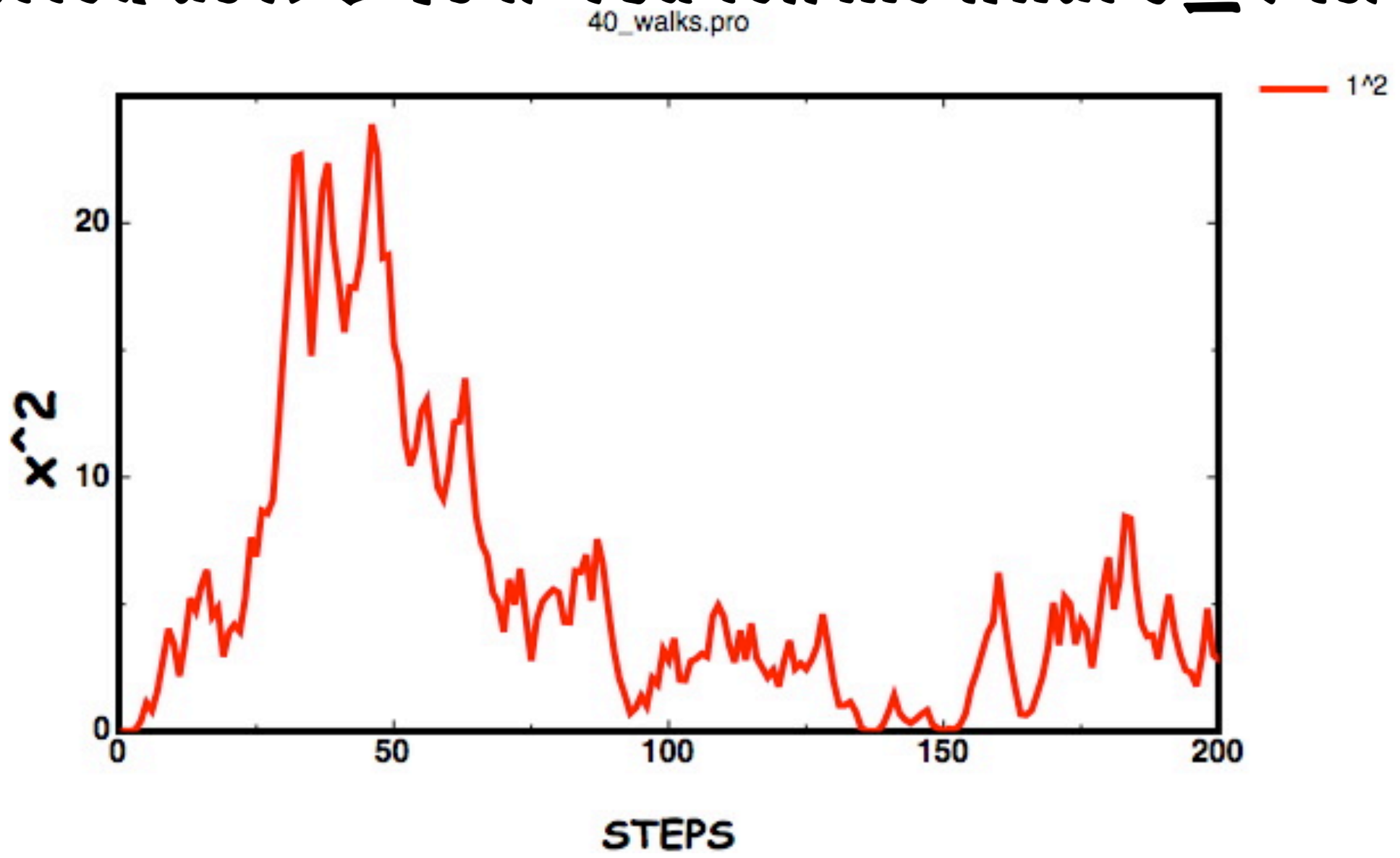
1-D Facilitated Diffusion!

I find the statistical mechanics of random walks to be quite counter-intuitive and amazingly noisy.

Here are 40 random walks in 1-D. There is a perfectly well-defined single value for D_1 here. Or is there?



Here is one random walk chosen at random, plotted as x^2 vs t. You tell me what D_1 is.



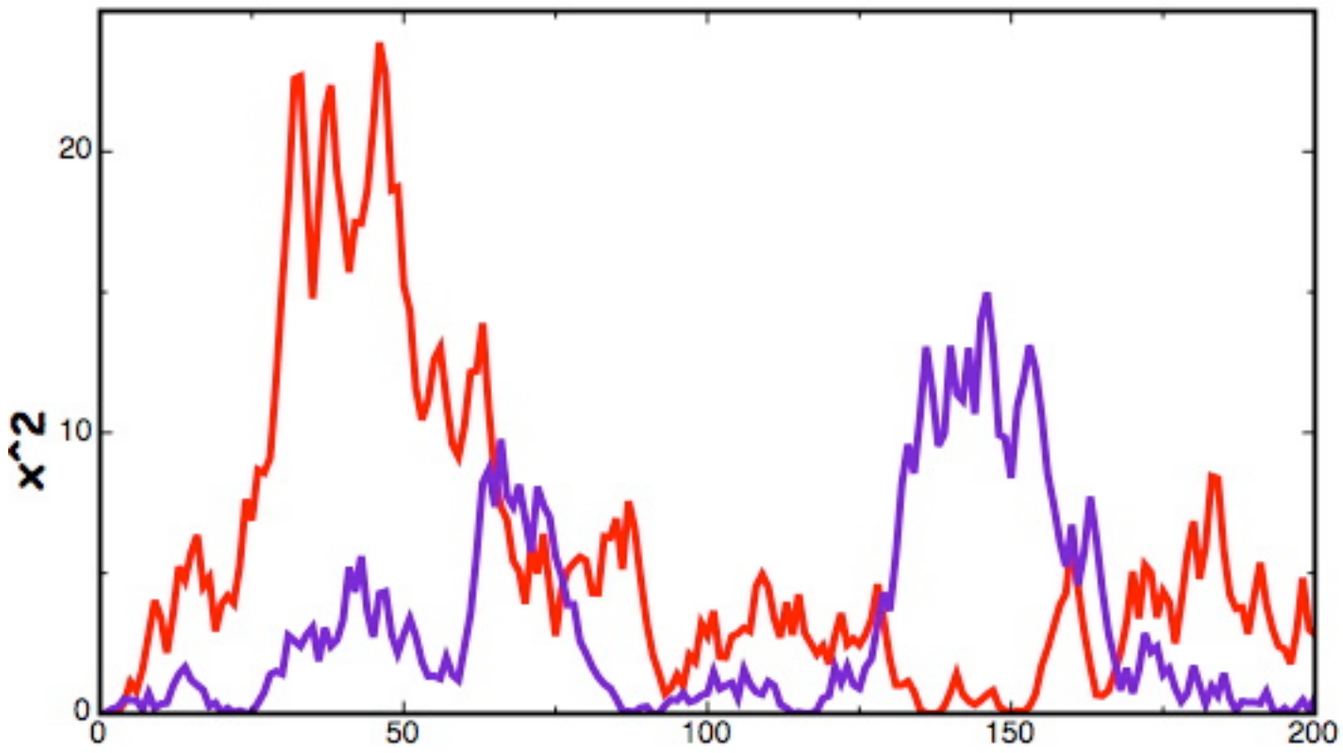
For a single 1-D walk, you are "always" rather close to the origin, and the path is VERY "noisy".

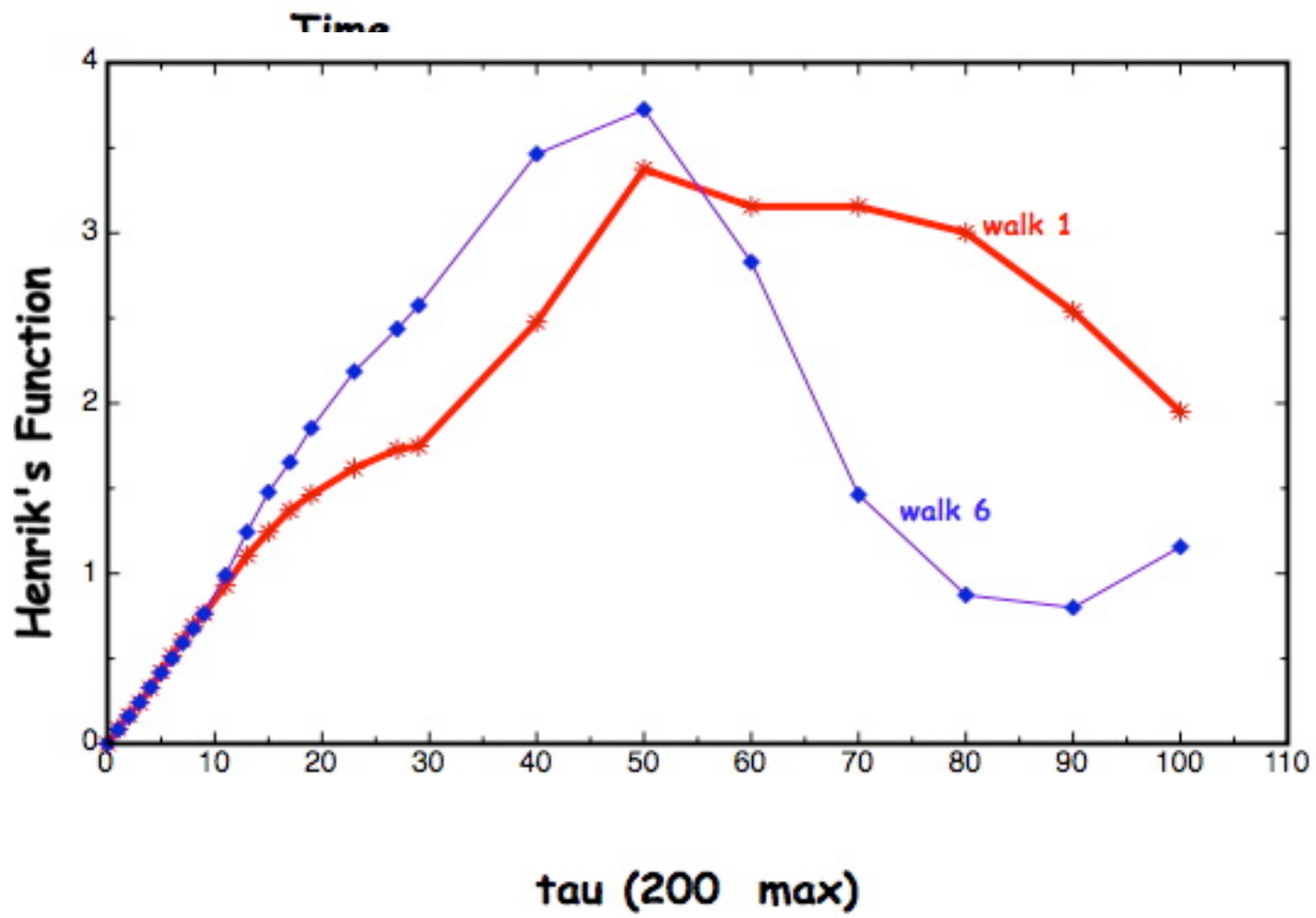
I usually get in losing, violent fights with theorists. Henrik Flyvbjerg pointed out my analysis is stupid.

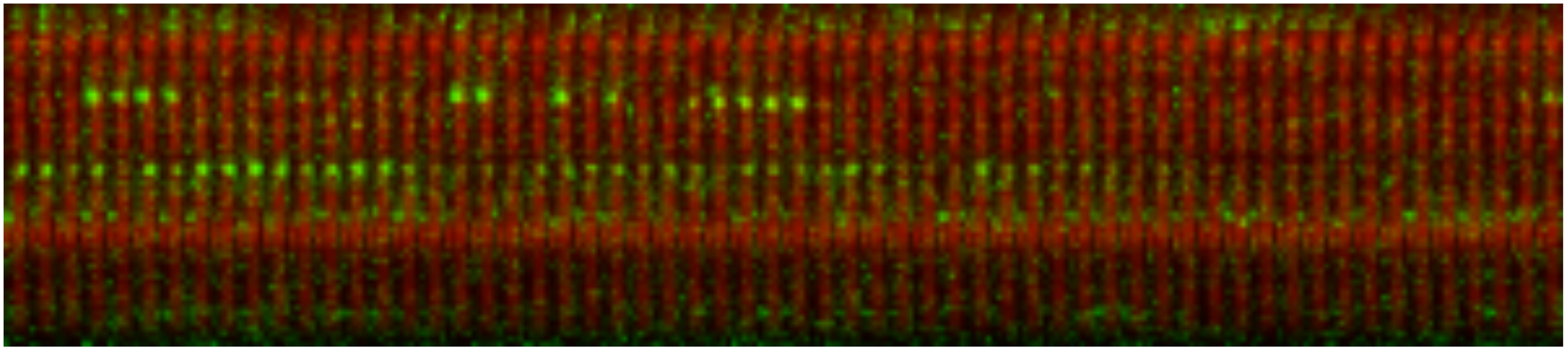
$$H_f(\tau) = \frac{1}{(N-\tau)} \Sigma_{j=1}^{N-\tau} \left[(x_j - x_{j+\tau})^2 \right]$$

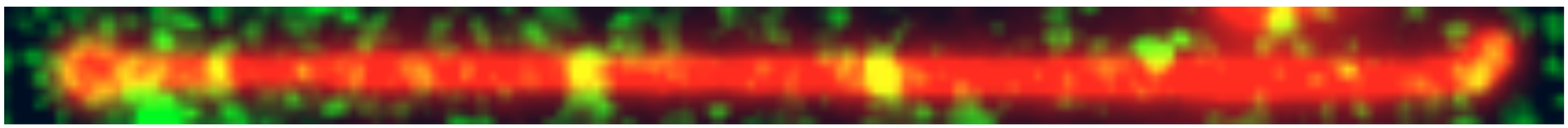
This is a funny kind of time-averaged mean-square displacement. If you plot $H_f(\tau)$ *vs.* τ you get a line whose slope is twice the diffusion constant as you might expect, but since you used ALL the points the error at least for $\tau << N$ vastly reduced.

It's just the auto-correlation of the walk, and "random walks" are highly correlated!

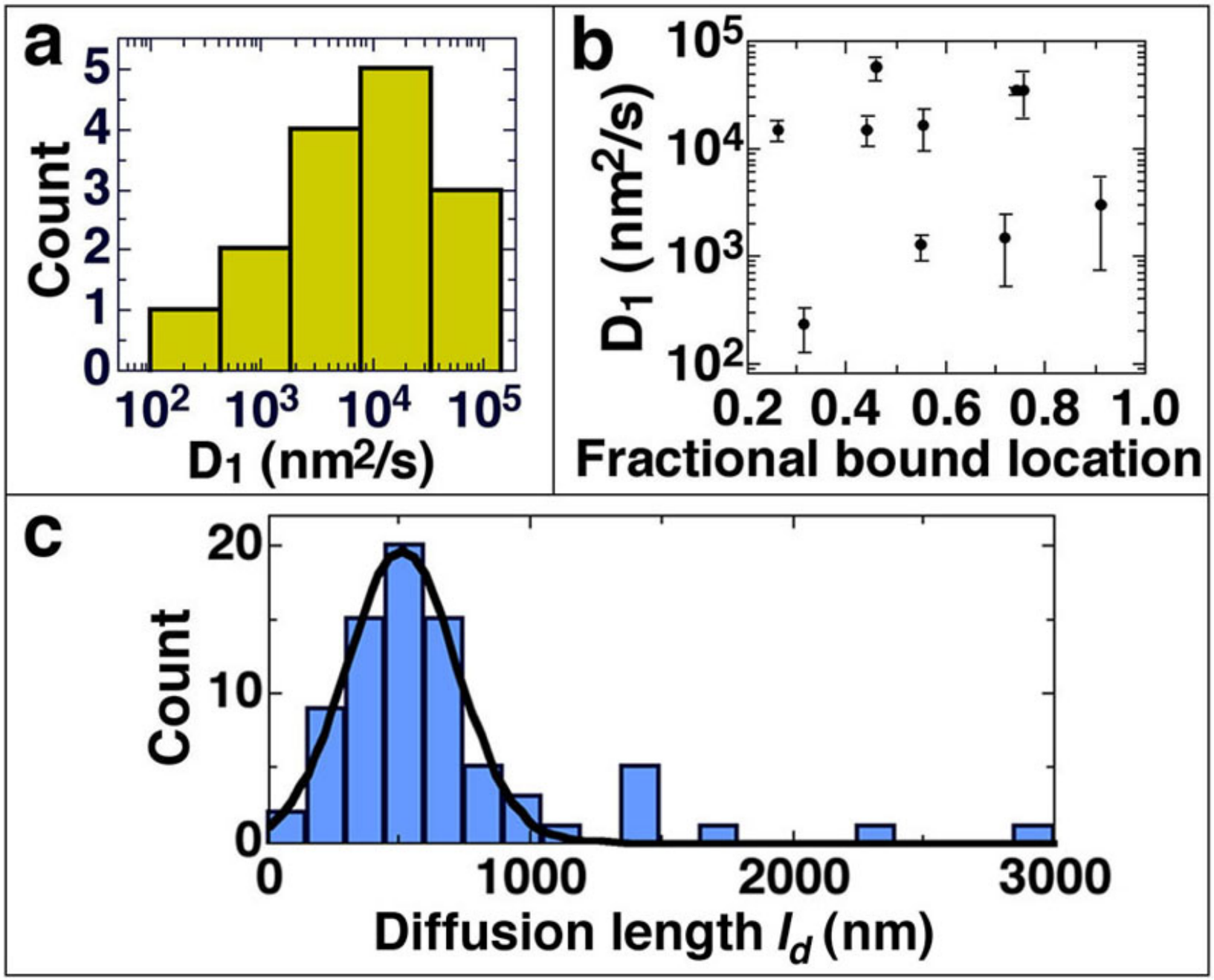

x^2
0
10
20
0
50
100
150
200
Time


Henrik's Function
0
1
2
3
4
0
10
20
30
40
50
60
70
80
90
100
110
walk 1
walk 6
tau (200 max)





There is no way for us to escape that there is a large DISTRIBUTION of 1-D diffusion coefficients for the lacI-GFP construct as it moves on non-specific sequence DNA.



The mean values of the 1-D coefficient lengths and the mean diffusional lengths l_d obtained from these single molecule meassurments indeed predict a LacI target binding rate 90 times faster than the 3D diffusion limit.

In some respects it is gratifying to get mean single molecule values that support ensemble averages, and it also a bit sobering to realize that much of biological knowledge revolves around the measurements of means of unknown ensembles of perhaps functionally very wide breadth.

1) There is a new world of nanoscale structures in biology that we are just beginning to explore. The statistical influence of restricted spatial dimensions is not trivial, is important to biology, and needs physics.

2) It is very, very difficult to both develop new techniques AND attack what biologists have decided is not a problem of general importance. Maybe the biologists are not the ones to lead always.

3) There is an enormously talented (well....I'm being PC and Deanspeak-like here) group of young physicists entering biology. There is a very conservative establishment they have to battle, and establishing supportive teams is critically important for survival.

My last talk, tomorrow, we'll have some fun and talk about demon bacteria.

Thanks!