Lecture 4 Random Walks

What is a structure?

i) Atomic level: PDB, crystallography.
needed for chemical details eg ezymes
eg protein - DNA recognition

ii) Generic level enough for non-folded structures: | dynamic
R_G = radius of gyration.
represents ensemble average size of polymer.
eg chromosome (unknown at high resolution)
every time it folds we may get different shape
nucleosome structure well defined (K rich)
but DNA sequence may be different.
eg virus has ordered protein shell
+disordered RNA-binding domains (K-rich)

slides

i) Pettsko-Ringe (2004) sequence ↔ consequence.
Sequence → structure → function.

ii) Universality: $R_G \propto N^{\nu}$ scaling.

"Statistical measures of structure"

1. 1D random walk.
N steps of $x_i = \pm a$
All 2^N configurations equally probable $p = 1/2^N$
Distance travelled $\langle R \rangle = \left\langle \sum_{i=1}^{N} x_i \right\rangle$
Variance $\langle R^2 \rangle = \left\langle \sum_i \sum_j x_i x_j \right\rangle$ ← average over all configs.

Summation and ensemble average commute:

$$\langle R \rangle = \sum \langle x_i \rangle = 0$$

$$\langle R^2 \rangle = \sum_{i=j} \langle x_i^2 \rangle + \sum_{i \neq j} \langle x_i x_j \rangle$$

$\langle x_i^2 \rangle \downarrow (\pm a)^2$; $\langle x_i x_j \rangle \downarrow$ averages to zero.

$\Rightarrow \langle R^2 \rangle = Na^2$ "width" of distribution $\sqrt{\langle R^2 \rangle} = \sqrt{N} \cdot a$.

2. Probability Distribution

n_r steps right, n_l steps left $\Big\}$ $N = n_r + n_l$ $\quad p_r = p_l = \frac{1}{2}$ each step.

$$W(n_r; N) = \frac{N!}{(N-n_r)!\, n_r!}$$ a binomial coefficient

slide eg $N = 3$ steps

$n_r = 0, 1, 2, 3 \quad W = 1, 3, 3, 1 \quad \sum W = 2^3$

$$p_r = \tfrac{1}{2} \Rightarrow p(n_r; N) = \frac{N!}{(N-n_r)!\, n_r!} \left(\tfrac{1}{2}\right)^N$$

Can see that probabilities are normalised:

$$\sum_{n_r=0}^{N} W(n_r; N) = 2^N \Rightarrow \sum_{n_r=0}^{N} p(n_r; N) = 1$$

(binomial)

More configurations in the middle of dist.

$\Rightarrow$ higher entropy, more likely

Textbook p317 evaluates p using:

i) Stirling $\ln N! \simeq N \ln N - N + \frac{1}{2} \ln(2\pi N)$

ii) $\ln(1+x) \simeq x - \frac{1}{2}x^2$

$R = (n_r - n_l)a \Rightarrow n_r = \frac{N}{2} + \frac{R}{2a} \quad n_l = \frac{N}{2} - \frac{R}{2a}$

result $\ln p(R; N) = \ln 2 - \frac{1}{2}\ln(2\pi N) - \frac{R^2}{2Na^2}$

(quite a lot of work)

$$\Rightarrow p(R; N) = \frac{2}{\sqrt{2\pi N}}\, e^{-R^2/2Na^2}$$

This is still the discrete form, $R = (n_0)a$. $\Delta R = 2a$.

Continuous form,

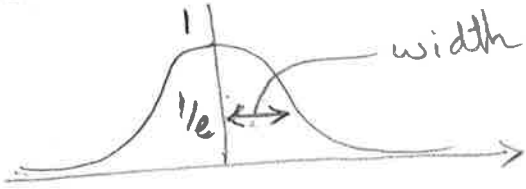
$$p(R; N)\, dR = \frac{1}{\sqrt{2\pi N a^2}}\, e^{-R^2/2Na^2}\, dR.$$

Normalised Gaussian of width $\sqrt{N}a$

This is what we expected from the central limit theorem:

Binomial dist → Gaussian dist.

slide
100 segs
bin/Gauss.



$e^{-R^2/2\sigma^2}$ $\sigma = \sqrt{N}a$.

3. Three-dimensional Random Walk.

$\vec{R} = (x, y, z)$ becomes a vector.

Each component obeys a separate, indep. rand walk

total distance from origin $R^2 = x^2 + y^2 + z^2$

Step is also resolved into 3 components which are equal on average:

$a_x = a_y = a_z$ with $a^2 = a_x^2 + a_y^2 + a_z^2 = 3a_x^2$

$$p(\vec{R}; N)\, dx\, dy\, dz = \left(\frac{1}{\sqrt{2\pi N a_x^2}}\right)^3 \underbrace{e^{-(x^2+y^2+z^2)/2Na_x^2}}_{\text{product of three dists.}} dx\, dy\, dz.$$

$$= \left(\frac{3}{2\pi N a^2}\right)^{3/2} e^{-3R^2/2Na^2}\, dx\, dy\, dz.$$

Kuhn Length, a = rigid step size in 3D.
= length over which polymer is assumed to stay rigid

Useful probability distribution for a random walk polymer.

Two defining variables:

a = segment (Kuhn) length

N = number of links.

Length of polymer $L = Na$ is not enough information to specify its distribution/size.

4. Persistence Length, ξ_p.

Closely related concept to Kuhn length.

Defined as tangent – tangent correlation funct.

s, u distance along polymer

$$\langle \vec{t}(s) . \vec{t}(u) \rangle = e^{-|s-u|/\xi_p}$$

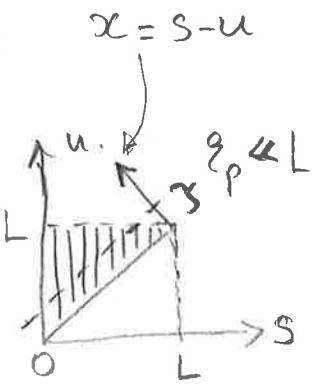
assumes exponential decay of correlations of direction persisting for length ξ_p

$\vec{R}$ = end to end length.

$$\vec{R} = \int_0^L \vec{t}(s)\,ds \quad \text{integrating along polymer.}$$

$$\langle \vec{R}^2 \rangle = \left\langle \int_0^L \vec{t}(s)\,ds \cdot \int_0^L \vec{t}(u)\,du \right\rangle$$

integration and average commute.



$$\langle \vec{R}^2 \rangle = \int_0^L ds \int_0^L du\, e^{-|s-u|/\xi_p}$$

$$= 2 \int_0^L ds \int_s^L du\, e^{-(s-u)/\xi_p}$$

$$= 2 \int_0^L ds \int_0^\infty dx\, e^{-x/\xi_p}$$

$$= 2 L \xi_p \quad \text{assuming } \xi_p \ll L$$

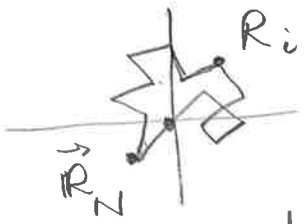
$$= Na^2 = La \quad \text{from before.}$$

2.2.2015. Hence $\boxed{a = 2\xi_p}$ relation between Kuhn length a and persistence length ξ_p.

Slide ξ_p = 50 nm DNA
30 nm 10 nm fibre : DNA + nucleosomes.

5. Radius of Gyration. [Sam Edwards]

Important concept, but tricky derivation. Polymer of length $L = Na$ undergoes a 3D random walk, starting at the origin and ending at $\vec{R}_N$, given by Gaussian dist.



$\vec{R}_i$ is position of i'th link.

End-to-end distance overestimates the size (variance $\langle R_N^2 \rangle = Na^2$). We are more interested in the distribution of matter along the chain. Two definitions:

$$\vec{R}_{CM} = \frac{1}{N}\sum_{i=1}^{N} \vec{R}_i = \text{centre of mass.}$$

$$R_G^2 = \frac{1}{N}\sum_{i=1}^{N} \left\langle (\vec{R}_i - \vec{R}_{CM})^2 \right\rangle. \qquad \text{I}$$

$$= \frac{1}{2}\frac{1}{N^2}\sum_{i}^{N}\sum_{j}^{N} \left\langle (\vec{R}_i - \vec{R}_j)^2 \right\rangle. \qquad \text{II}$$

$\frac{1}{2}$ avoids double counting of distances

Second definition shows that this R_G is what will be seen in a scattering expt which measures point-to-point correlations.

Polymer is a random walk from i to j given by the same Gaussian probability:

$$\left\langle (R_i - R_j)^2 \right\rangle = \int (R_i - R_j)^2 P(R_i - R_j, |i-j|)\, d^3R.$$

$$= |i-j| a^2 \text{ , the variance}$$

$$\Rightarrow R_G^2 = \frac{1}{2}\frac{a^2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} |i-j| = \frac{1}{2}\frac{a^2}{N^2} 2 \sum_{i=1}^{N}\sum_{j=1}^{i} (i-j)$$

$$= \frac{a^2}{N^2}\sum_{i=1}^{N} \frac{i^2}{2} = \frac{a^2}{2N^2}\,\frac{N(N+1)(2N+1)}{6} \simeq \frac{Na^2}{6} \qquad N \gg 1$$

using sum of integers & sum of squares of integers.

6. Radius of gyration:

$$\sqrt{\langle R_G^2 \rangle} = \sqrt{\frac{La}{6}} = \sqrt{\frac{L\xi_p}{3}} = \sqrt{\frac{\langle R^2 \rangle}{6}}$$

Can calculate the expected radius of a strand of ds DNA from its Length, N_{bp}.

$L = N_{bp} \times 0.34\,nm$ free DNA.

$\xi_p = 50\,nm$ ds DNA.

slide

λ phage	$N_{bp} = 6 \times 10^4$	$R_G = 0.5\,\mu m$
E Coli	6×10^6	$5\,\mu m$
Human Chr 1	2.5×10^8	$50\,\mu m$.

Conclusion: phage has no problem, except it needs to actively pack DNA into capsid (100nm) for transmission. Discuss later.

E Coli has no problem: $1 * 2\,\mu m$ cell size.

Human needs a different way to pack:

i) tethering.

slides ii) folding onto nucleosomes

protein core, histone octamer

2 turns of ds DNA ~ 170 bp. + spacer

⟷ 11 nm diam x 6nm thick.

"beads on a string" → 10nm fibre

Expt shows $\xi_p = 30\,nm$ for 10 nm fibre

7. Linear density parameter γ

γ = N_{bp} per nm length (base pair spacing L 0.34 nm)

$\doteq 3$ for naked DNA

$\doteq$ 200bp / 6nm = 33 for 10 nm fiber

$\sim$ 100 for 30 nm fiber in vitro

$\sim 2.5 \times 10^8 / 2000\,nm = 125{,}000$ for entire chromosome.

Repeat calculation for 10nm fibre:

$L = 2.5 \times 10^8 \div 33 = 8 \times 10^6$ nm. chr 1

$R_G = \sqrt{L l_p / 3} = 9\mu m$. more managable

↑ 30 nm. for 10 nm fibre

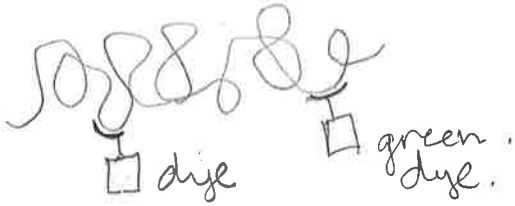
Actual chr 1 size is about 2μm in its most condensed state. Can be decondensed by action of protease to ~ 10μm. or more.

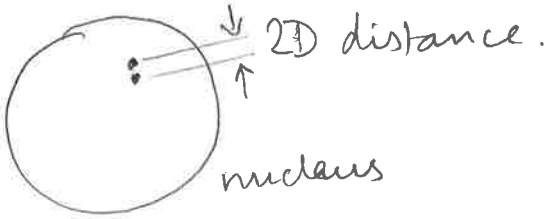
slides Chromosome Territories
Multicolor Fluorescence in situ Hybridisation
M-FISH coding of "paint"
DNA of isolated chromosomes chopped up and PCR amplified + fluorescent codes added.
Interphase: nucleus filled, non overlapping
preferred location 18 outside
(gene rich) 19 center of nucleus

8. Tethering.
Proposed mechanism for avoiding too large spreading out of DNA polymer. Wanner model of chrom. architecture has loops of 100kb to 1Mb with attachment to a scaffold:

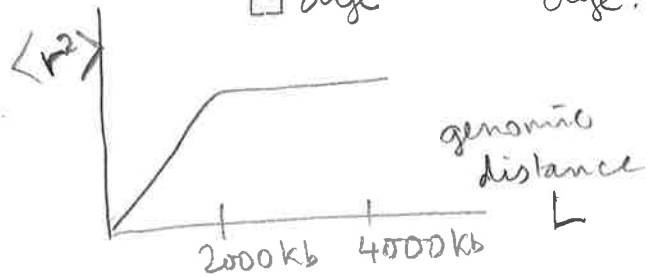
$L = 10^5 \div 33 = 3\mu m$ $R_G = 170$ nm ~ observed size of "chromomeres"

FISH probes for specific sequences separated by a known distance!





slide



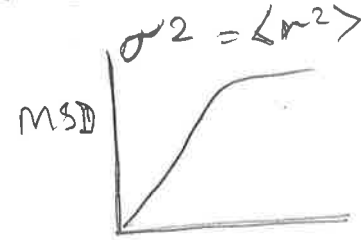
MSD is linear
$\langle r^2 \rangle \propto L$.

Measure linear density parameter:

slide $P(r) \propto e^{-3r^2/2Na^2} \Rightarrow \sigma^2 = Na^2/3$

a = Kuhn length $= 2\xi_p$ r^2 = distance^2 btw dyes

N = number of segments btw markers.



Physical length $L = N_{bp}/\nu$ [ν = N_{bp} per nm]

$= Na$

Graph shows $\sigma^2 = \frac{Na^2}{3} = \frac{La}{3} = \frac{N_{bp}a}{3\nu}$

See 1.8 μm^2 per 1000 kb distance

$$\frac{a}{\nu} = \frac{3\sigma^2}{N_{bp}} = \frac{3 \times 1.8\ \mu m^2}{10^6} = 5.4\ nm^2/bp$$

Naked DNA $a = 2\xi_p = 100$ nm $\Rightarrow \nu = 18$ bp/nm.

10nm fiber $a = 2\xi_p = 60$ nm $\Rightarrow \nu = 11$ bp/nm.

Numbers are more consistent with packing in nucleosomes, as expected.

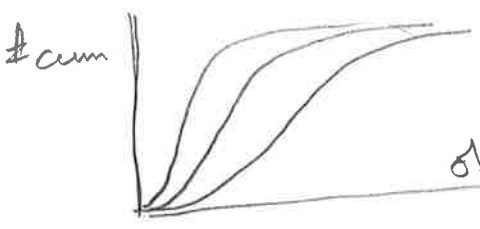
Distance r is $|\vec{r}|$ distributed in 3D.

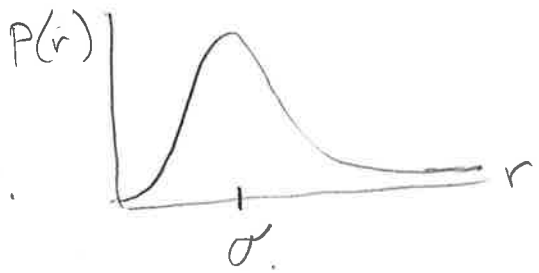
Integrating $P(\vec{r}; N)$ over angular variables

$$P(r; N)\,dr = \left(\frac{3}{2\pi Na^2}\right)^{3/2} 4\pi r^2 e^{-3r^2/2Na^2}\,dr$$

Actual data: van den Engh et al Science (1992)

Cumulative probability:

slide 

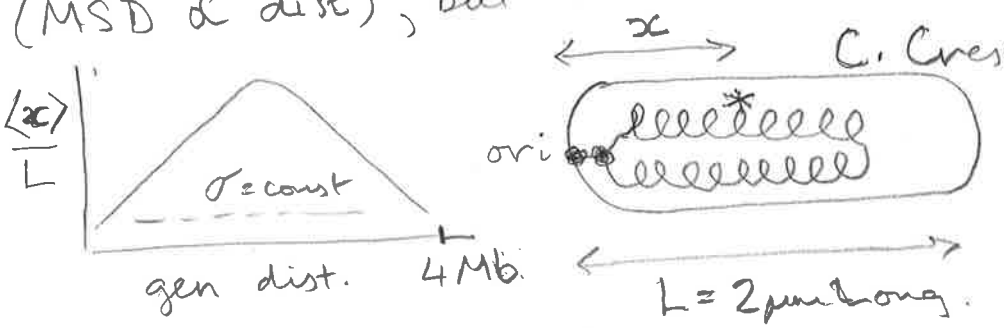


9. Bacterial Chromosomes (p327)

i) Caulobacter crescentus

[Vibrio cholerae] →

slide No longer see random walk behaviour (MSD of dist), but linear trend. $\langle x \rangle$ of dist.



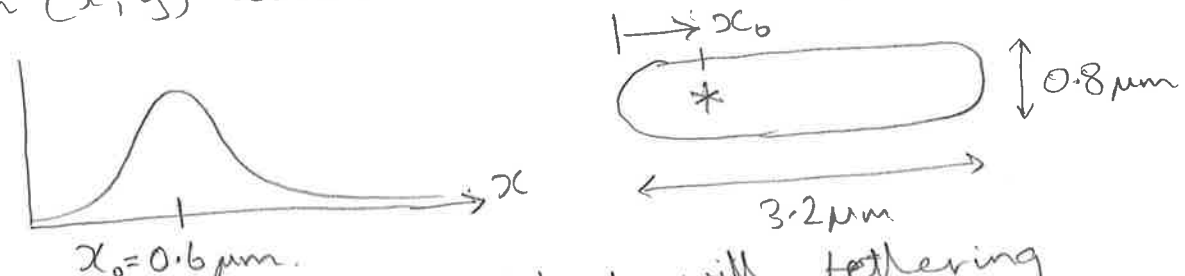
Suggests model of partitioning into loops to give linear trend. Loop ~ 8 kb (4 Mb total)

Free naked DNA 4 Mb = 1.4 mm long!

random walk $R_G = \sqrt{L \ell_p / 3} \approx 5 \mu m$.

ii) Vibrio cholerae

Similar story. Location of marker in (x, y) coords of cylinder-shaped cell.



Gaussian dist consistent with tethering at $x = x_0$, 0.6 μm from pole of bacterium.

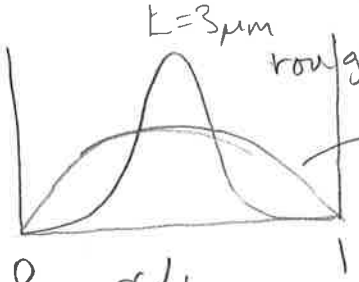
y-direction is non-Gaussian.

Confined random walk (pp 330-333) with tether at $x = x_0$

$$P(x) = \sqrt{\frac{1}{2\pi N a^2}}\, e^{-(x-x_0)^2/2Na^2}$$ random walk

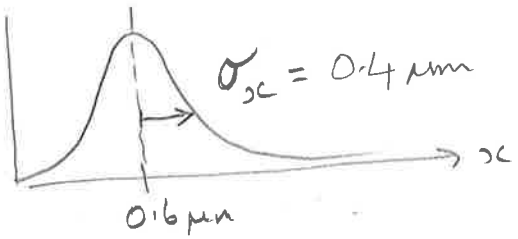
N = fluorescent marker position in units of Kuhn length $a = 100$ nm (naked DNA).

slide.



Quantitative analysis of data: [1D model]

slide



$\sigma_x^2 = Na^2$ [not ×3 because
$= La$
$a = 100$ nm naked DNA
$L = 1600$ nm
$N_{bp} = 5000$ base pairs

we conclude that this marker was placed 5kb away from the tethering location.

10. DNA looping (p334)

slide

A. DNA bubble, as in RNA Polymerase movie.
B. RNA hairpin. Most common secondary structure of RNA eg tRNA ribosome
C. Lac repressor.
D. Recombination event in chromosome especially associated with meiosis mechanism of exchanging genes between parental lines. p97 dog/rat/mouse.

→ Lac repressor = protein (tetramer)
Lac operon = DNA sequence it binds to.
most effective when bound to two operators simultaneously 2×2 dimers.
Fewer combinations available when DNA is looped.

Understand Lac repressor function as probability of loop formation:

Probability of random walk returning to origin (in one dim.)

$$P_0 = \frac{\text{number of looped configs}}{\text{total number of configs.}}$$

For N Kuhn segments (length a each).

$$P_0 = \frac{N!}{(N/2)!\,(N/2)!} \div 2^N$$

$$\ln N! = N\ln N - N + \tfrac{1}{2}\ln(2\pi N)$$

$$N! = N^N / e^N \cdot \sqrt{2\pi N} = \left(\frac{N}{e}\right)^N \sqrt{2\pi N}$$

$$P_0 = \frac{(N/e)^N \quad \sqrt{2\pi N}}{(N/2e)^{N/2}\,(N/2e)^{N/2} \quad (2\pi N/2)} \frac{1}{2^N}$$

$$= \sqrt{\frac{2}{\pi N}}$$

(except slightly less trend $\propto N^{-1/2}$ in 1D.

slide Cyclisation probability

Purely entropic considerations: better model (later) will include the energetics of bending the DNA into a loop.

Otherwise, just random walk considerations would suggest $N = 1$ most likely loop!

slide Data for 1EFA on PDB.

domain 1

domain 2.

α-helices inserted into major groove of DNA

side chains form H-bonds to edge of base pairs to detect sequence.

11. 3D random walk.

Cyclisation probability

$P_0 = \sqrt{\frac{2}{\pi N}}$ in one dimension.

3D can be thought of three independent random walks in x, y, z.

All of them have to return to the origin so product of probabilities.

$$P_0 = \left(\frac{2}{\pi N}\right)^{3/2} \text{ in 3D.}$$

Can elaborate by considering range of distance over which chain returns.

[But for lac repressor to be effective, we want $P_0 \sim \frac{1}{1000}$ $\frac{2}{\pi N} \sim \frac{1}{100}$ or $N \sim 65$

Kuhn length $a \sim 100\,nm = 300\,bp$.

$L = Na = 6.5\,\mu m$ or 20 kbp quite long.

Actual lac operons are spaced ~500 bp apart, so bending becomes the limiting factor.]