A thermodynamic model for agglomeration of DNA-looping proteins

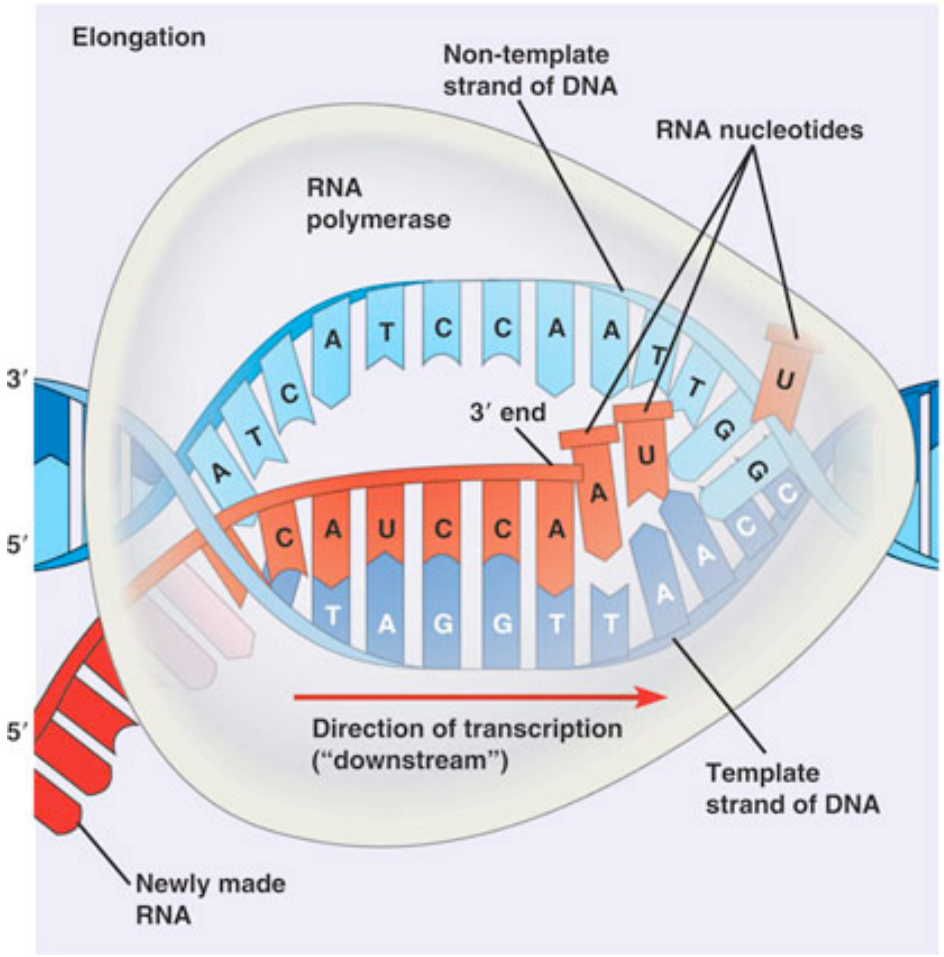
Sumedha
I.S.I Foundation, Torino, Italy.

Work in collaboration with: **Martin Weigt**(ISI, Torino, Italy).

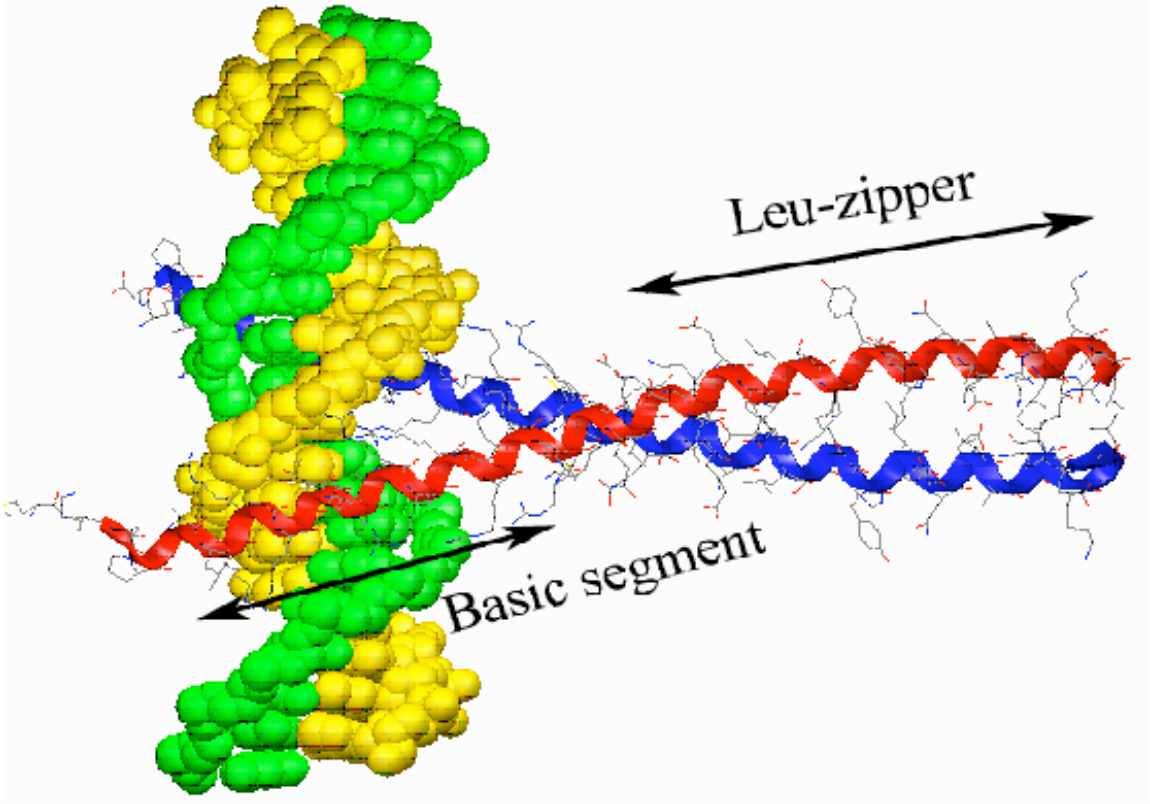
Outline

- Transcription
- Models and mechanism for single DNA looping.
- Transcription factories
- Our model for agglomeration of DNA-loops (transcription factories).
- Conclusions.

Transcription



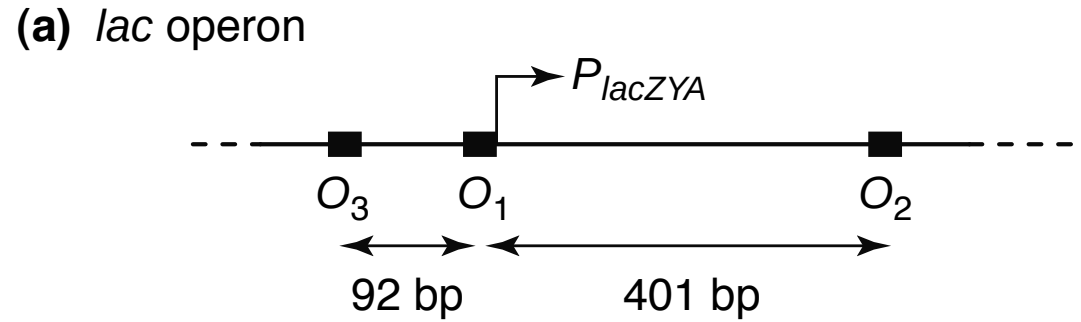
Structure of transcription factors



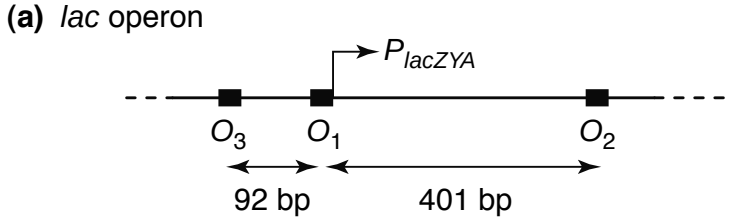
➡ mostly homo-dimers of two protein molecules

- bind to palindromic binding motifs
- might introduce DNA loops

Protein mediated DNA loop

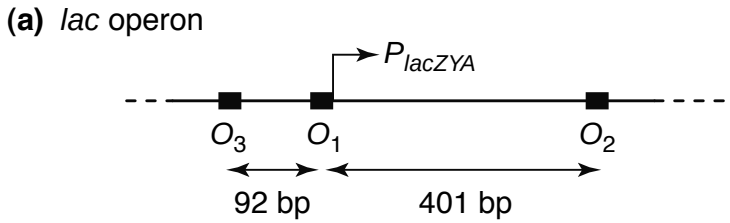


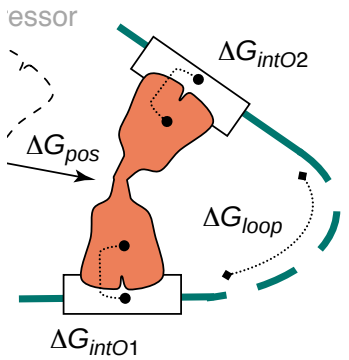
Protein mediated DNA loop



- Binding of lac repressor to O_1 represses transcription
- O_2 and O_3: auxillary operators(lac repressor can bind but cannot suppress expression).
- Simultaneous elimination of O_2 and O_3 reduces repression level almost 100 fold.

Protein mediated DNA loop

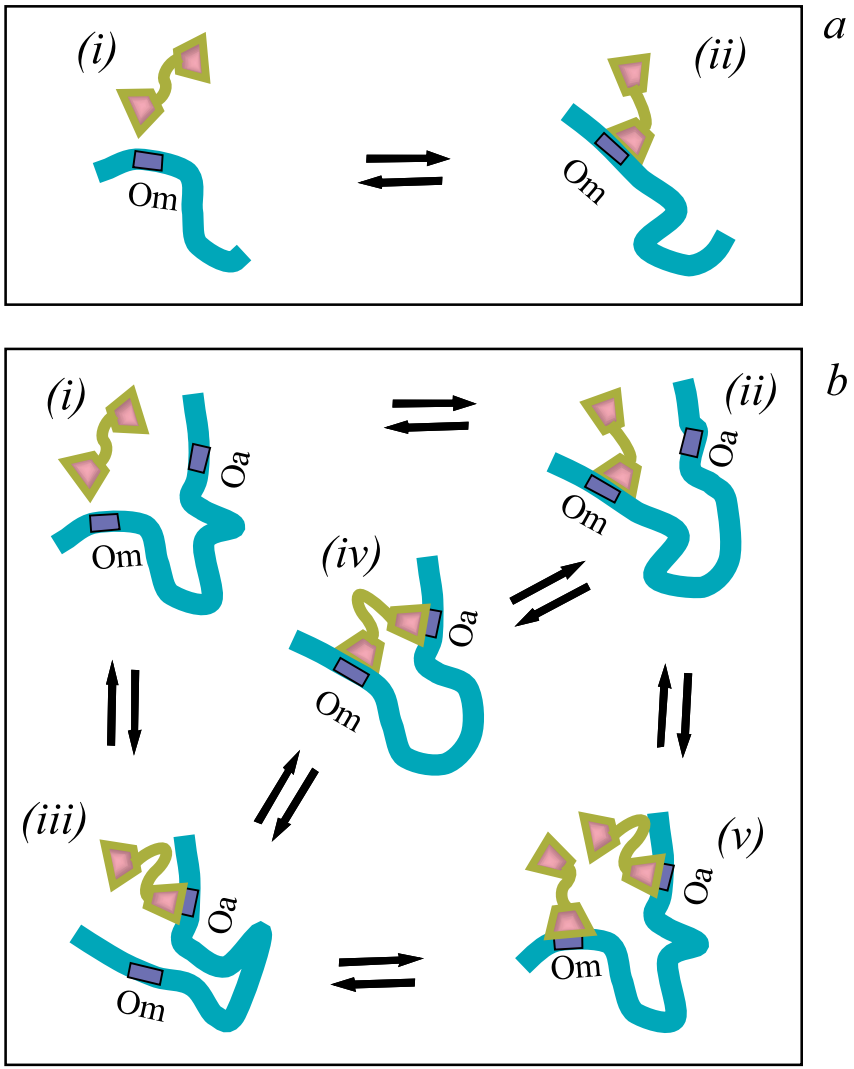




- Binding of lac repressor to O_1 represses transcription
- O_2 and O_3: auxillary operators(lac repressor can bind but cannot suppress expression).
- Simultaneous elimination of O_2 and O_3 reduces repression level almost 100 fold.
- DNA looping : involves entropy cost-but increases effective concentration of lac repressor near O_1

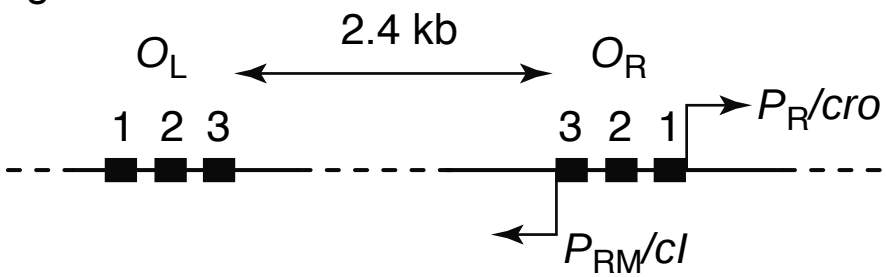
Leibler and Vilar, JMB 2003

lac repressor: a dimer protein, binds simultaneously to two operators, looping the intervening DNA.

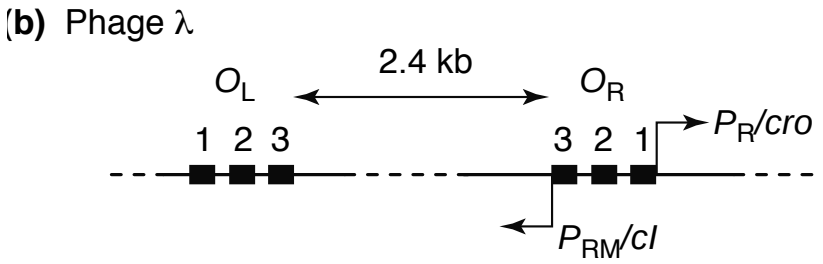

a
(i)
(ii)
Om
Om
b
(i)
(ii)
(iii)
(iv)
(v)
Oa
Om

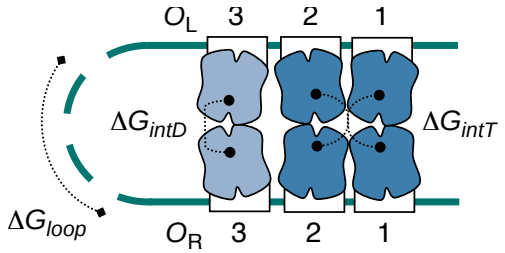
Protein mediated DNA loop

(b) Phage λ



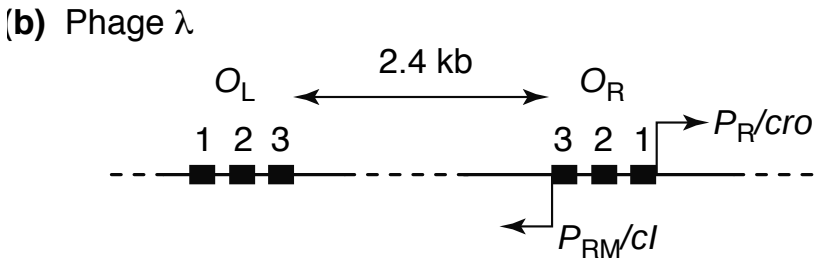
Protein mediated DNA loop

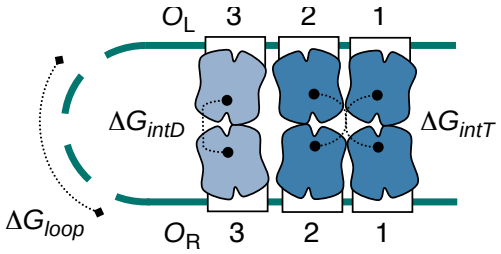




Looping cost is higher, concentrations of TFs is low.

Protein mediated DNA loop





Looping cost is higher, concentrations of TFs is low.

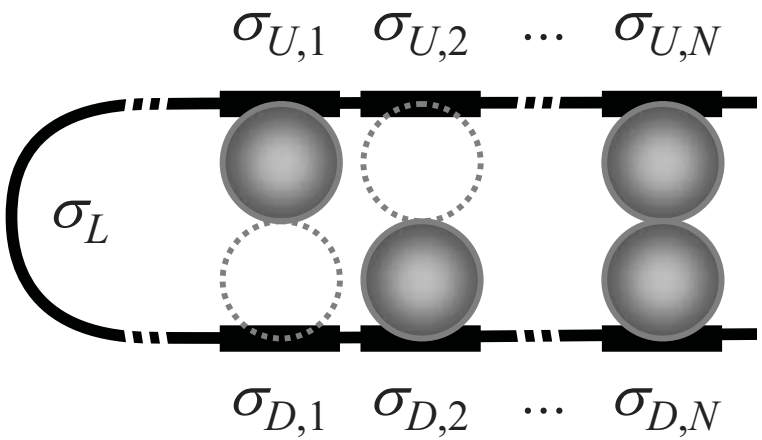
3 important contributions:

- binding affinity of TF to binding site
- entropy losses due to DNA looping (for short loops also bending stiffness)
- attractive short range protein-protein interactions

Vilar and Saiz, Current Op. in Gen. and Dev. 2005.

Modeling multi-protein DNA-loop

Vilar and Saiz, PRL 2006



$$H = (c + e\sum_{i=1}^{N}\sigma_{U,i}\sigma_{D,i})\sigma_L + g\sum_{i=1}^{N}(\sigma_{U,i} + \sigma_{D,i}) \ , \ = \sum_{i=1}^{N} H_{P,i}$$

e : interaction energy between two proteins.

g : protein-DNA binding energy

c : cost of looping

Modeling multi-protein DNA-loop

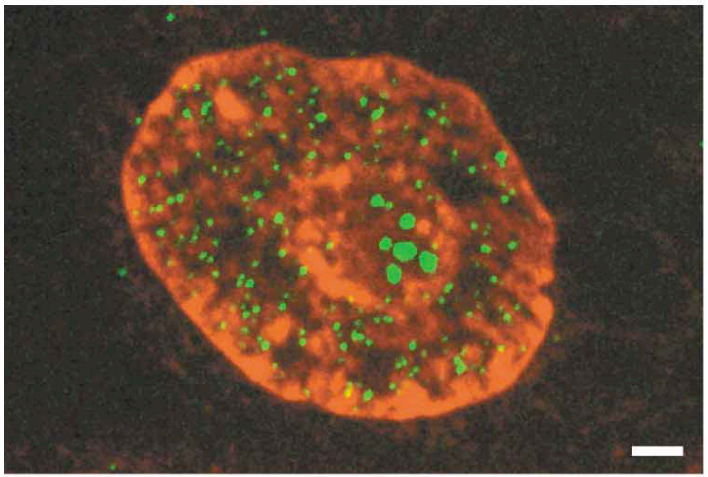
Critical protein concentration for unbound-bound transition:

$$\tilde{n} = e^{\beta g^o} \sqrt{\frac{c\beta}{N\left(e^{-e\beta} - 1\right)}}$$

Increasing the number of binding sites, system reaches the looped phase at arbitrary small protein concentrations.

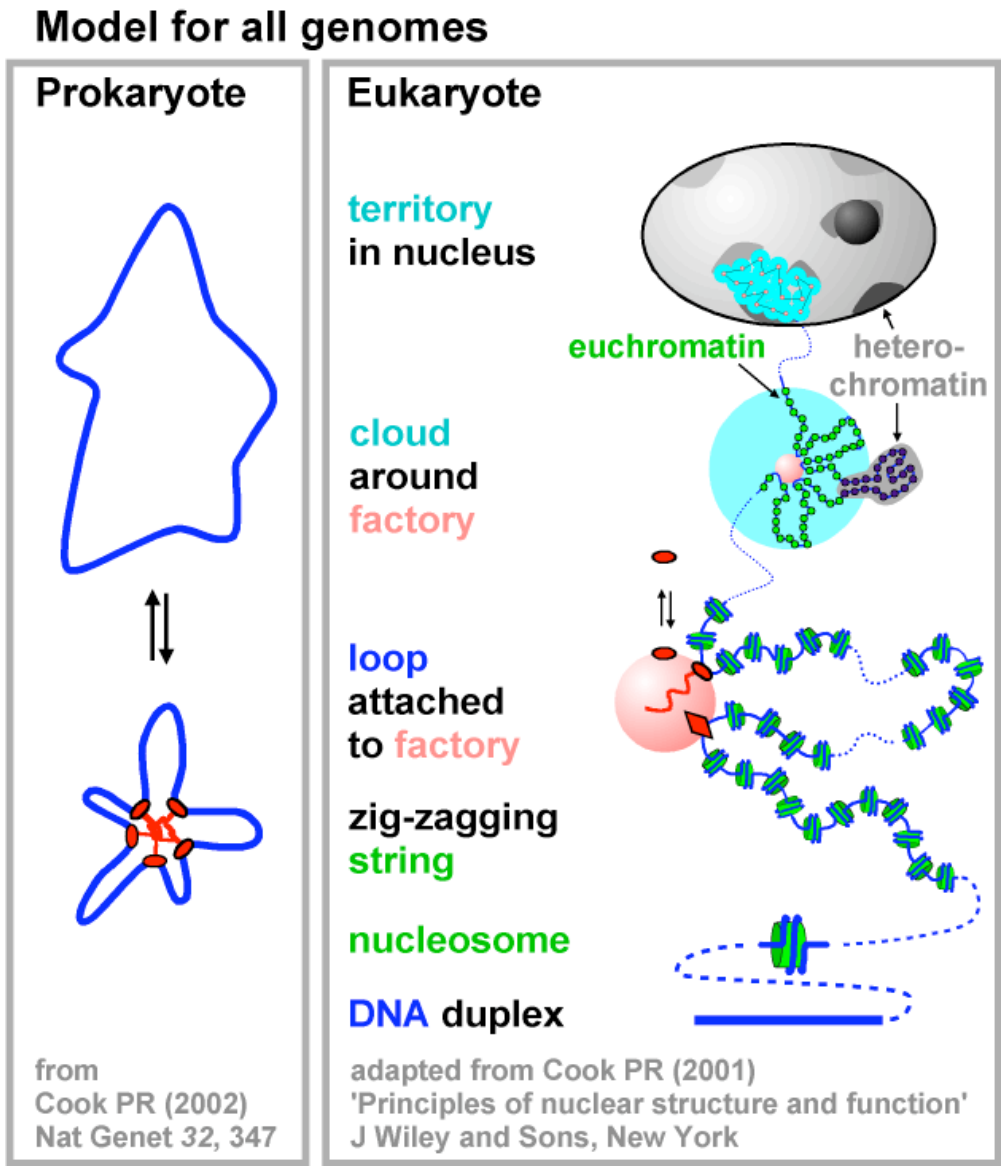
Transcription Factory

RNA polymerase occurs in highly enriched foci:
Transcription factories



living Bacillus subtillis, P. Cook(Nature Genetics,2002)

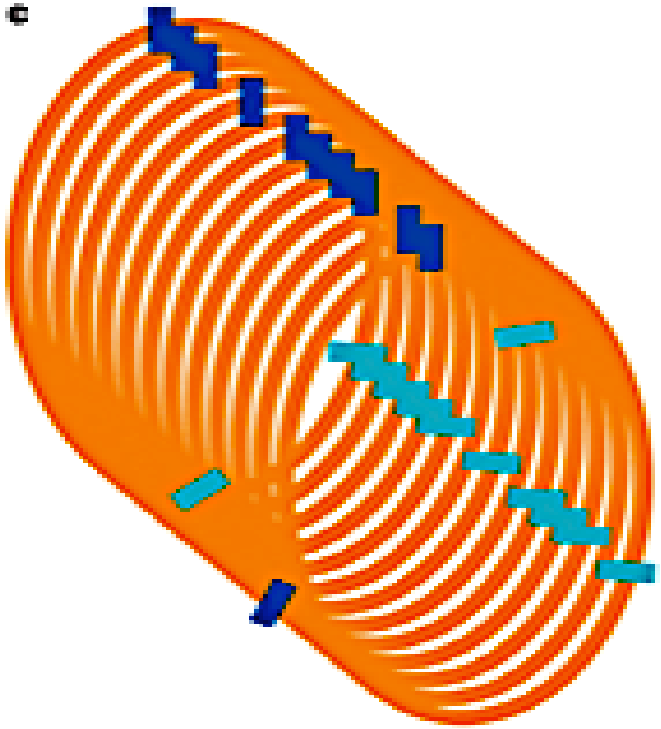
Cook's model of transcription factories



- formation of complexes containing RNA polymerase and transcription factors
- DNA / chromatin loops back due to affinity between DNA and proteins

mechanism of factory formation?

The solenoid model



- Co-regulated genes are regularly spaced on DNA/chromatin.
- Transcription factors induce loops by linking co-regulated genes
- Spontaneous organization of focal points of transcriptional activity
- Spontaneous organization of DNA / chromatin in solenoid

F. Kepes, JMB, 2003

Transcription foci

- Multiple active genes and gene clusters are often located together at places in the nucleus that have high local concentration of transcription and m-RNA machinery.

Transcription foci

- Multiple active genes and gene clusters are often located together at places in the nucleus that have high local concentration of transcription and m-RNA machinery.
- Actively transcribed genes that are separated by up to 40Mb of chromosomal sequence frequently co-localize in the same transcription factory.

Transcription foci

- Multiple active genes and gene clusters are often located together at places in the nucleus that have high local concentration of transcription and m-RNA machinery.
- Actively transcribed genes that are separated by up to 40Mb of chromosomal sequence frequently co-localize in the same trasncription factory.
- Genome is organized into loops and looping plays an important role in controlling gene activity by bringing distant genes together so that they can bind to local concentrations of local proteins.

Chaklova et al, Nature Reviews, Genetics,2005
Fraser and Bickmore, Nature 2007

Modeling agglomeration of transcription factors

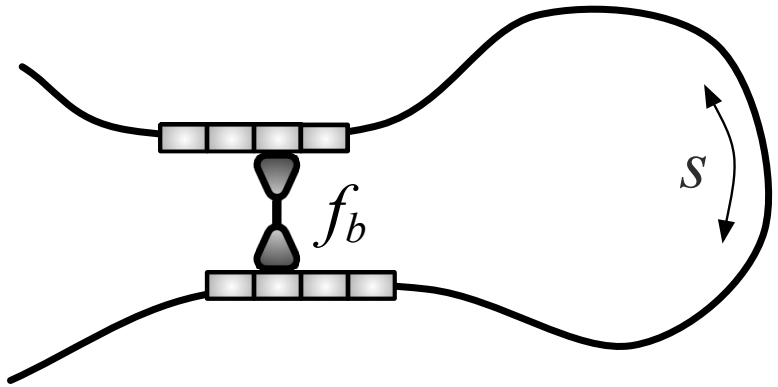
We assume:

- Many protein binding domains (N) on DNA.
- Each binding domain has K binding sites.
- Bivalent transcription factor(TF) can bind to two binding sites from two different domains.

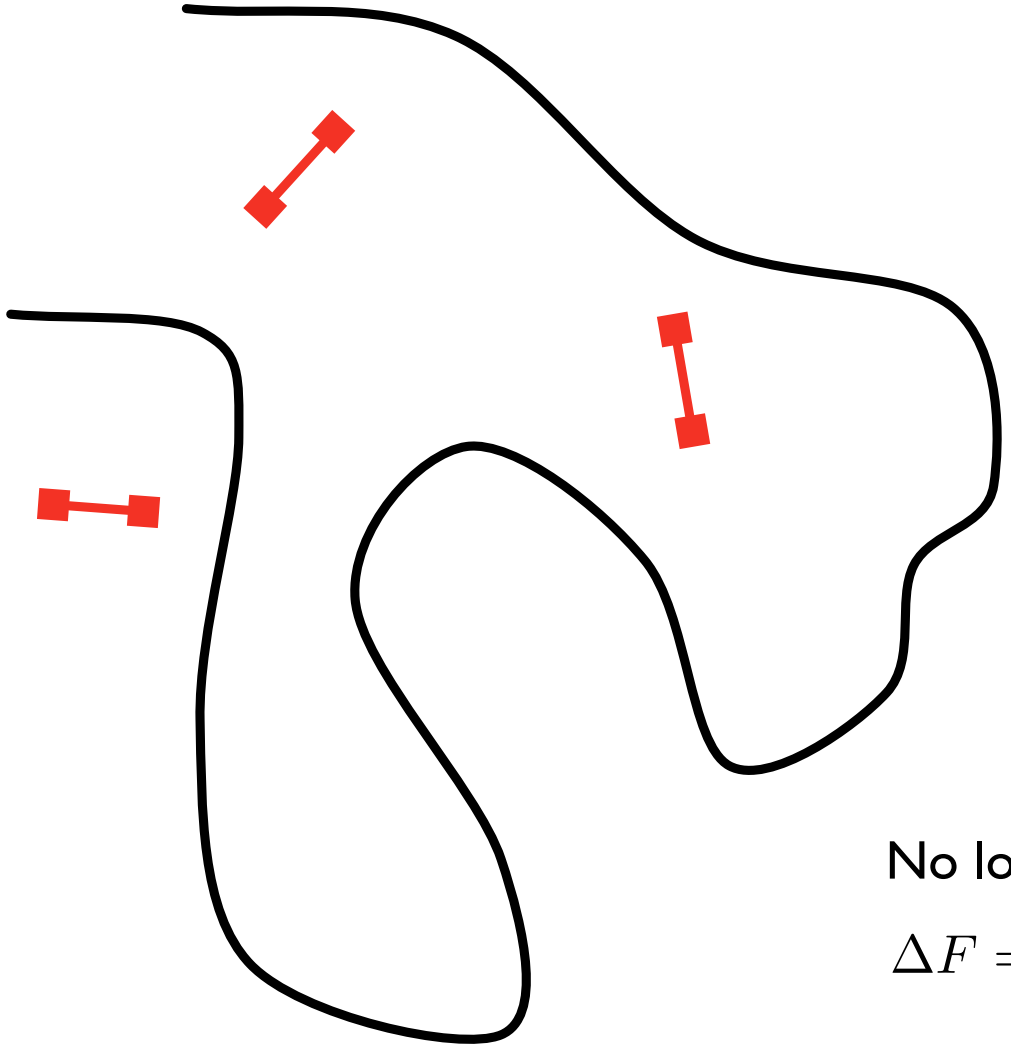
Modeling agglomeration of transcription factors

We assume:

- Many protein binding domains (N) on DNA.
- Each binding domain has K binding sites.
- Bivalent transcription factor(TF) that can bind to two binding sites from two different domains.



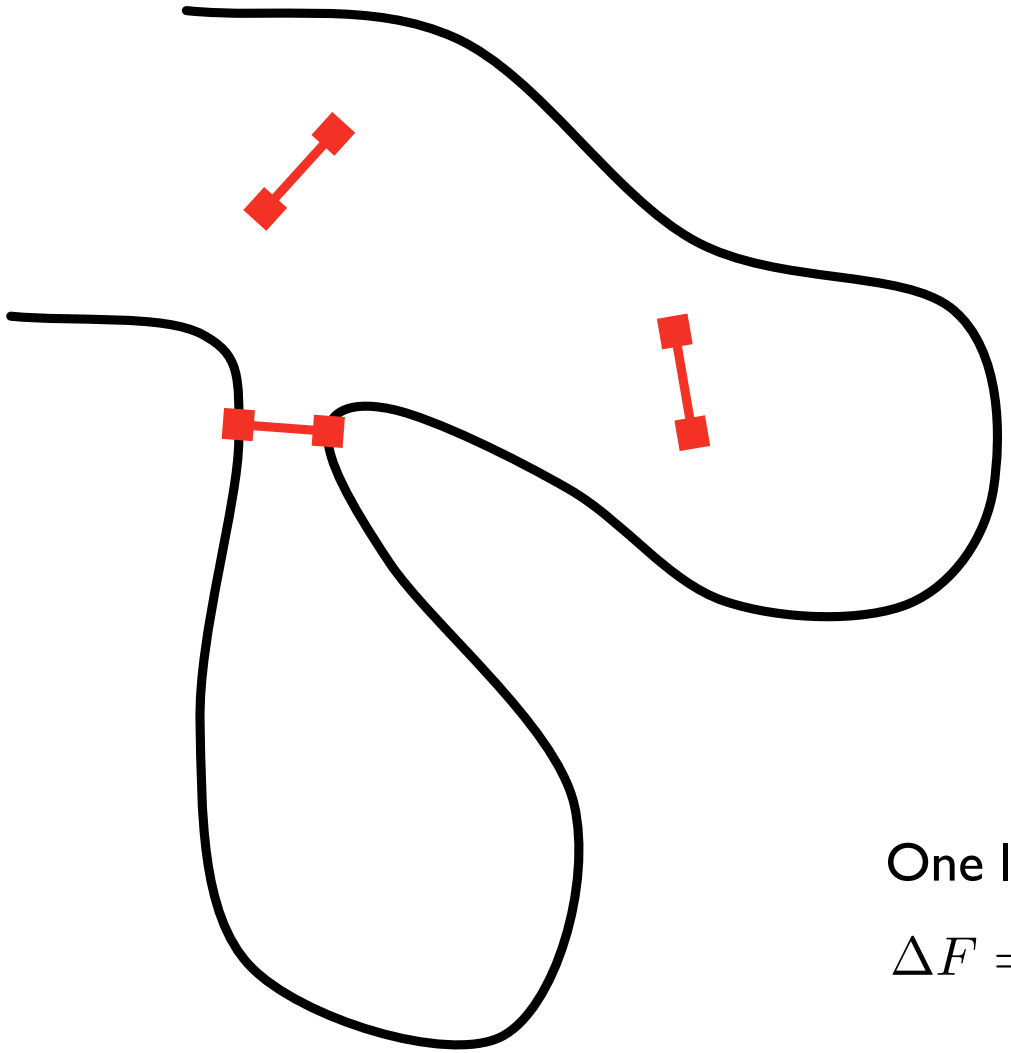
Agglomeration via DNA looping



No loop, no linker bound:

$\Delta F = 0$

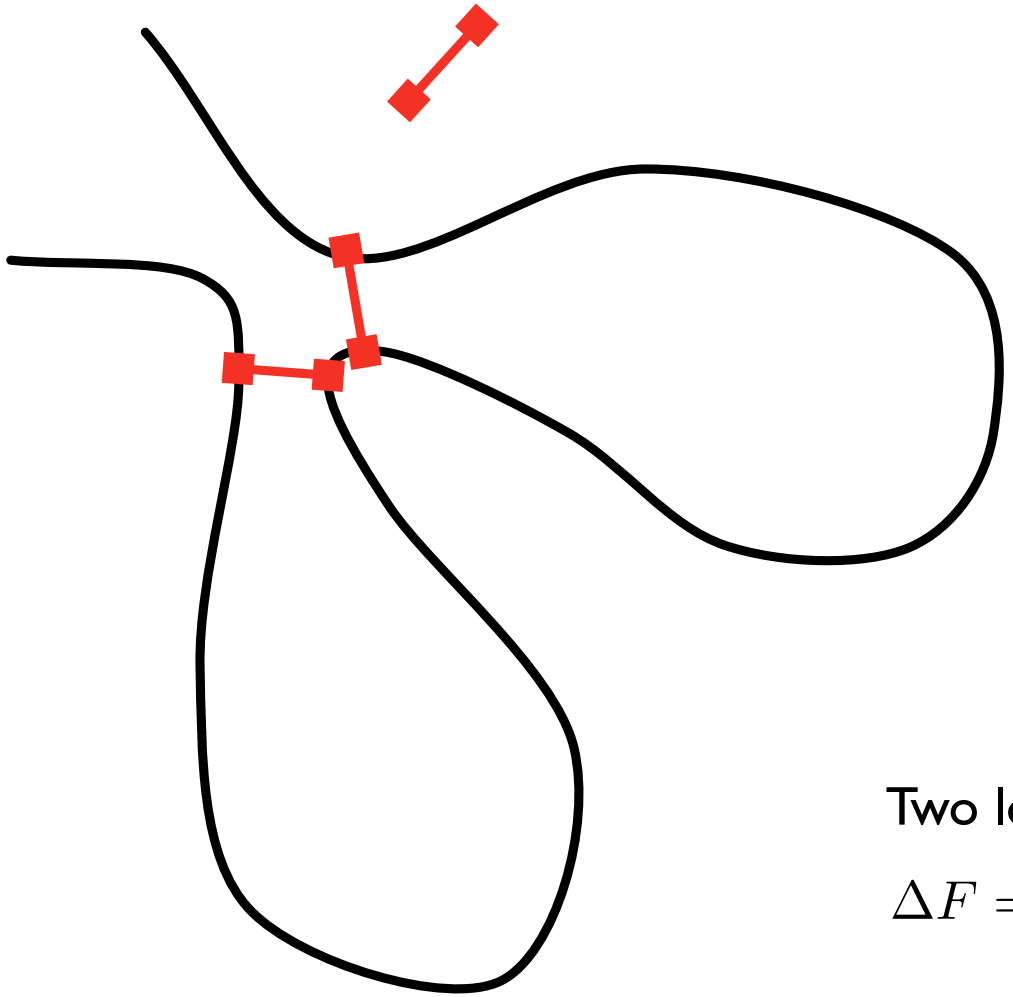
Agglomeration via DNA looping



One loop, one linker bound:

$$\Delta F = -E_b + T\Delta S$$

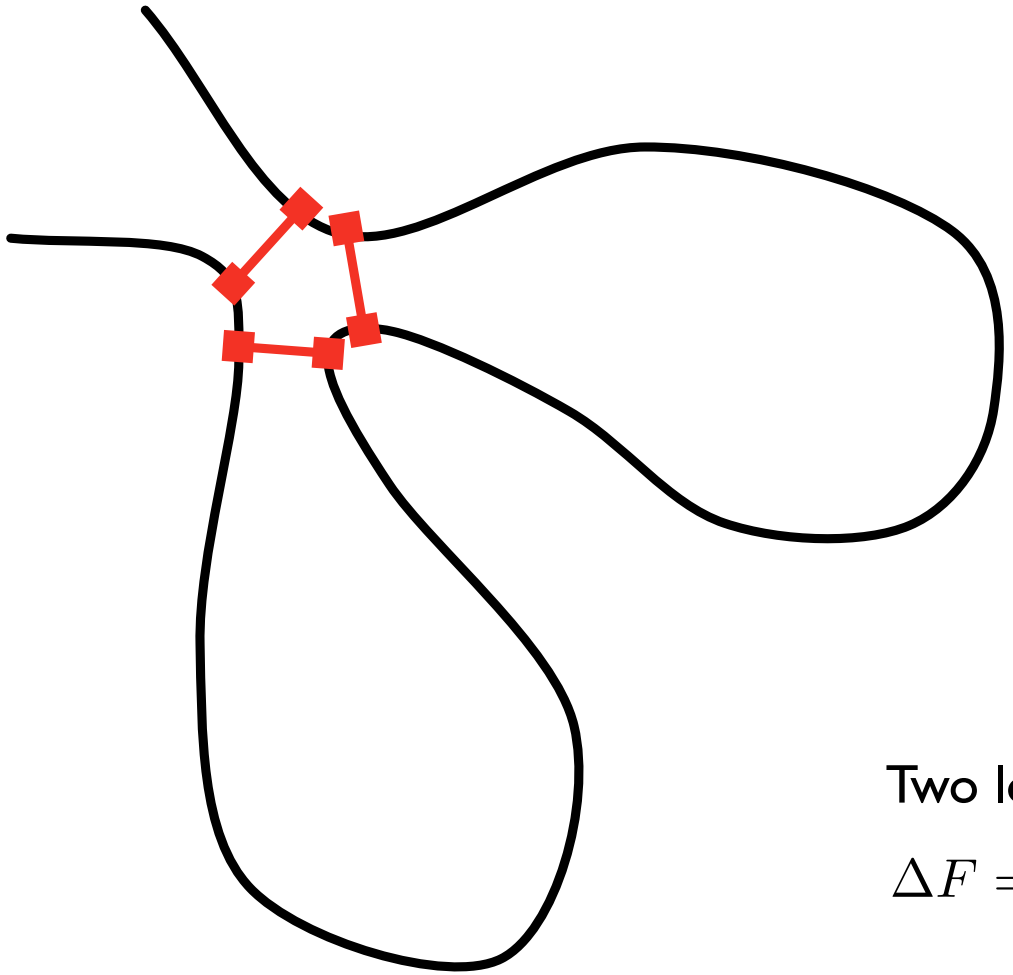
Agglomeration via DNA looping



Two loops, two linkers bound:

$$\Delta F = -2E_b + 2T\Delta S$$

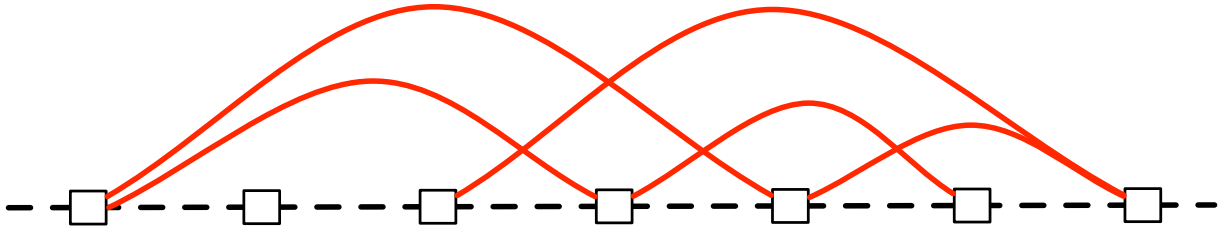
Agglomeration via DNA looping



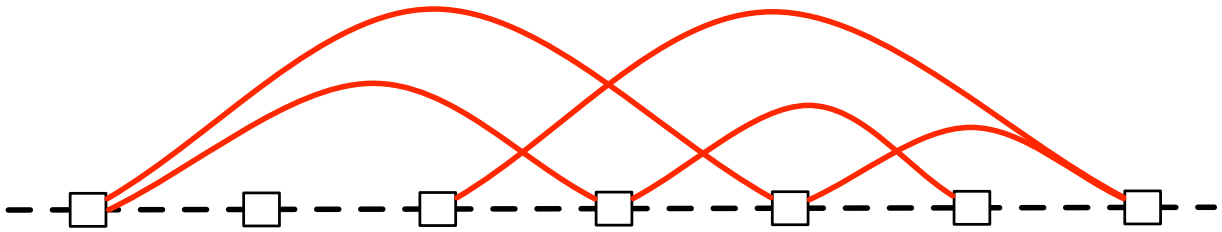
Two loops, three linkers bound:

$$\Delta F = -3E_b + 2T\Delta S$$

Mean-field modeling



Mean-field modeling

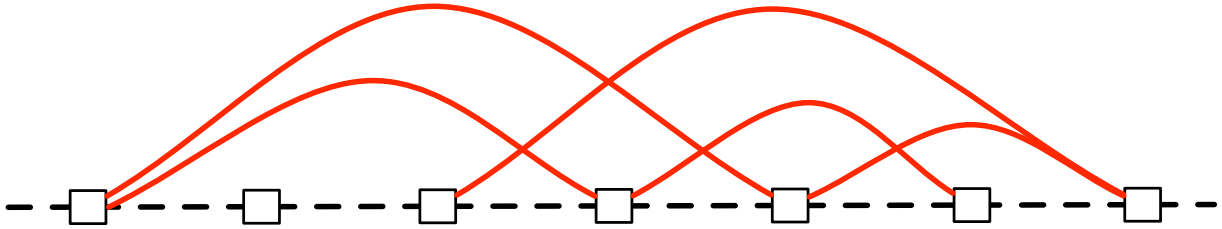


- can be modeled via the network of protein contacts
  - N binding sites are nodes
  - M doubly bound proteins are links
- each link contributes binding energy $-E_b$
- each connected component of n nodes contributes $(n-1)T\Delta S$

$$\Delta F = -ME_b + T\Delta S \sum_n (n-1) N_{cc}(n)$$

- each CC of n nodes might contain up to $n(n-1)/2$ links

Mean-field modeling

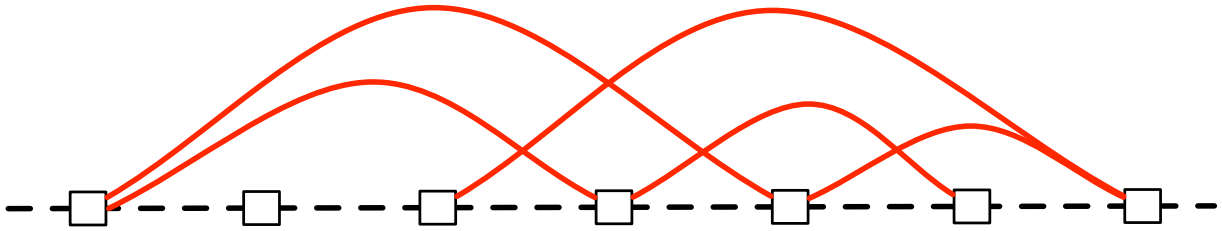


$$\begin{aligned}\Delta F &= -ME_b + T\Delta S \sum_n (n-1)N_{cc}(n) \\ &= -ME_b - TN_{cc}\Delta S + const\end{aligned}$$

Partition function:

$$Z = \sum_{\text{graphs}} \exp\left\{\frac{1}{T}ME_b + N_{cc}\Delta S\right\}$$

An equivalent random-graph ensemble



- a priori distribution of a sparse random graph

$$P_0(G|\gamma, N) = e^{-\gamma N} \left(\frac{\gamma}{N}\right)^M$$

- weighted with respect to number of connected components

$$P(G|\gamma, q, N) = \frac{1}{Z(\gamma, q, N)} P_0(G|\gamma, N)\ q^{N_{cc}}$$

Random graph with maximum degree K

K binding sites per binding domain.

Number of ways of binding d proteins to a binding domain:

$$p_d = \frac{1}{\mathcal{N}} \binom{N}{d} \left(\frac{\gamma}{KN}\right)^d \left(1 - \frac{\gamma}{KN}\right)^{N-d} \Theta(K-d) M_{K,d}$$

Probability of a graph G, with L(G) edges:

$$P(\{J_{i,j}\}, \gamma, N) = \frac{1}{\mathcal{N}} \prod_{i<j} \left[(1 - \frac{\gamma}{KN})\delta_{J_{i,j},0} + \frac{\gamma}{KN}\delta_{J_{i,j},1}\right] \prod_i \frac{K!}{K - \sum_j J_{i,j}} \Theta(\kappa - \sum_j J_{i,j})$$

Random graph with maximum degree K

For finite K, we assume: adding a typical vertex/edge to a large random graph, does not change its properties significantly.

$$P(C, d; \gamma, N+1) = \sum_{\Delta C} K_d(\Delta C) P(C + \Delta C; \gamma, N)$$

$K_d(\Delta C)$: extra weight of adding a new vertex

Random graph with maximum degree K

For finite K, we assume: adding a typical vertex/edge to a large random graph, does not change its properties significantly.

$$P(C, d; \gamma, N+1) = \sum_{\Delta C} K_d(\Delta C) P(C + \Delta C; \gamma, N)$$

$K_d(\Delta C)$: extra weight of adding a new vertex

ν : fraction of binding domains inside the giant component

π : probability of finding free binding site inside the giant component.

$$\pi = \frac{(K - \bar{d_{in}})\nu}{K - \bar{d}}$$

Random graph with maximum degree K

For finite K, we assume: adding a typical vertex/edge to a large random graph, does not change its properties significantly.

$$P(C, d; \gamma, N+1) = \sum_{\Delta C} K_d(\Delta C) P(C + \Delta C; \gamma, N)$$

$K_d(\Delta C)$: extra weight of adding a new vertex

π : probability of finding free binding site inside the giant component.

$$\pi = \frac{(K - \bar{d_{in}})\nu}{K - \bar{d}}$$

$$K_d(\Delta C) \propto p_d \sum_{d_0=0}^{d} \pi^{d_0} (1-\pi)^{d-d_0} \delta(\Delta C, d - d_0 - \delta(d_0, 0))$$

Three order parameters in the system π, ν, γ , given by:

$$\nu = 1 - \frac{q}{q - 1 + \left(1 + \frac{\gamma\pi}{K(1+x)}\right)^K}$$

$$\pi = \frac{1}{K - c}\left(\nu K - c + \frac{(1-\nu)Kx}{1+x}\right)$$

$$c = \frac{K(1-\nu)x}{q(1+x)}\left[q - 1 + \frac{1+(q-1)\pi}{1-\pi}\left(\frac{1+\nu(q-1)}{1-\nu}\right)^{\frac{K-1}{K}}\right]$$

ν : fraction of binding domains inside the giant component.

π : probability of finding free binding site inside the giant component.

c : average number of transcription factors/binding domain.

Expanding around the fixed point :

$$\{\nu, \pi, \gamma\} = \{0, 0, cqK/(K-c)\}$$

$$\pi = -\frac{2K[c(K-1)-K]}{c^2(K-1)[2+K(q-2)]}$$

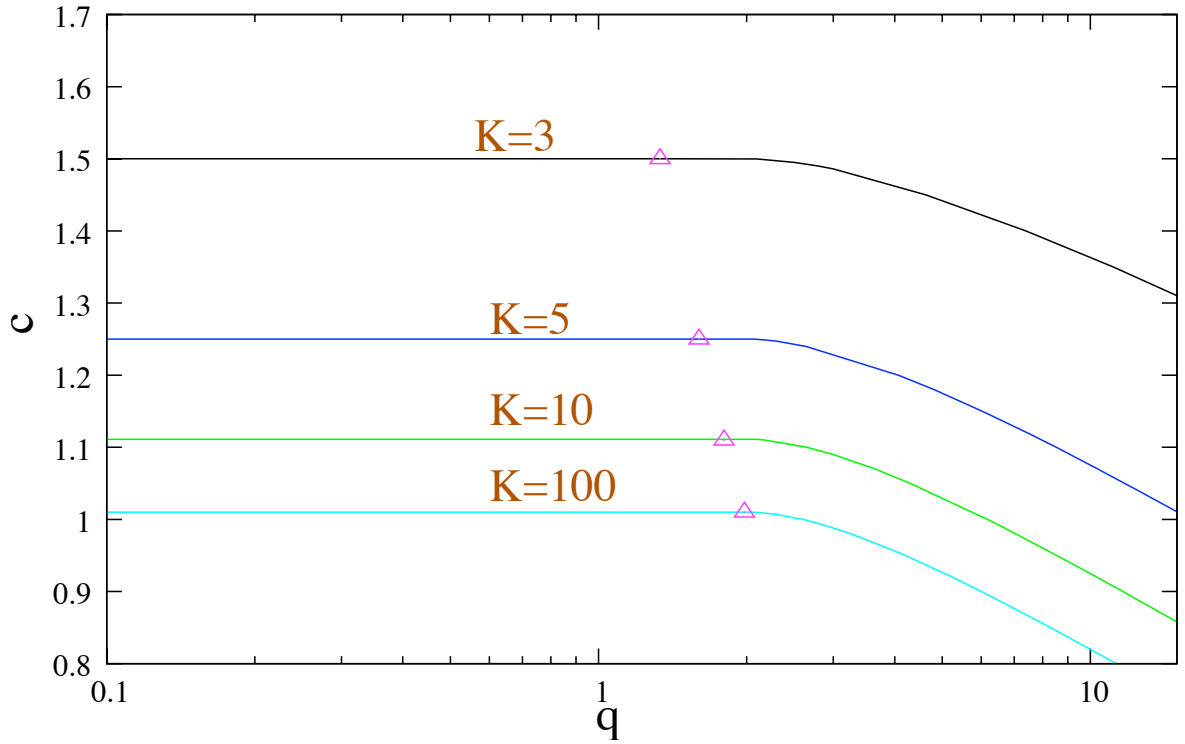
Percolation transition for

$$q < q_{critical} = 2 - 2/K$$

at:

$$c_{critical} = K/(K-1)$$

Phase Diagram:

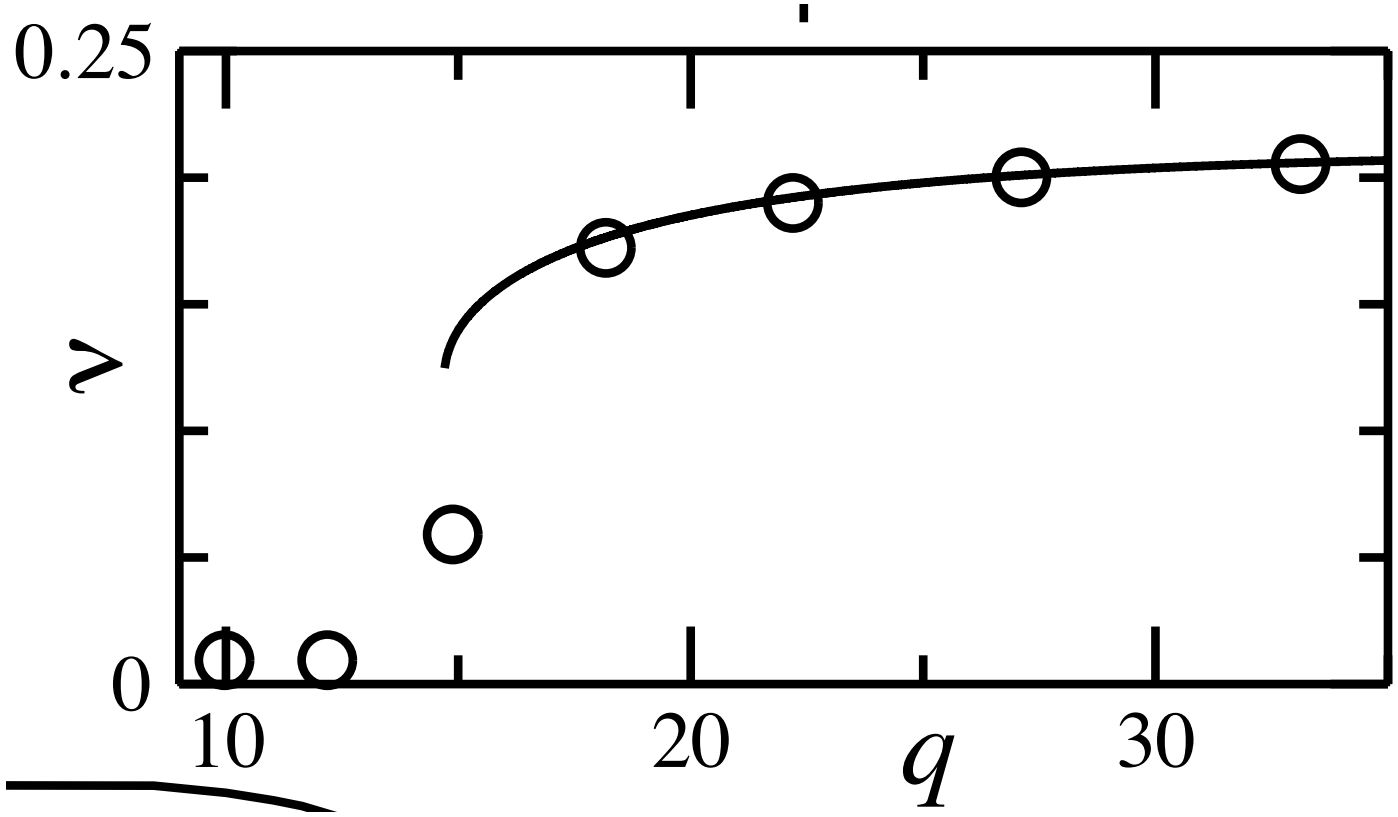


$K \to \infty$: usual random graph with poisson degree distribution.

$K = 2$: cont. transition for q<1 for c = 2

Comparison with simulations

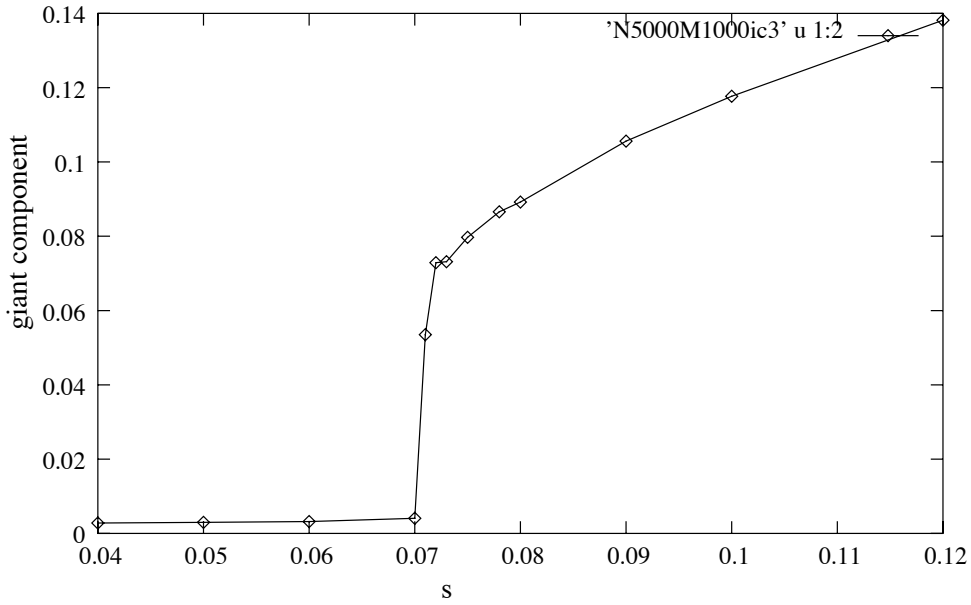
K=20, N=5000,M=2000



Length dependence

Assuming DNA to be a Gaussian chain, entropy loss:

$$s_l = \frac{3}{2} \ln(l/l_0)$$



Summary

- Transcription factories are a result of multiplicity of binding domains (with K>2).
- As entropy cost of loops increases, the formation of cluster happens at lower and lower concentration of TFs (consistent with experimental observations for λ phage).
- We obtained the full phase diagram for atypical random-graphs (with weight associated with number of clusters) for finite degree cutoff.