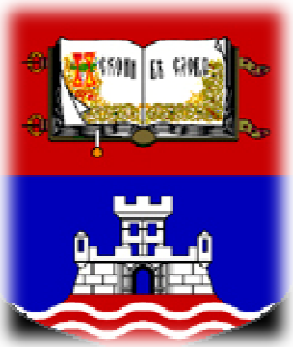
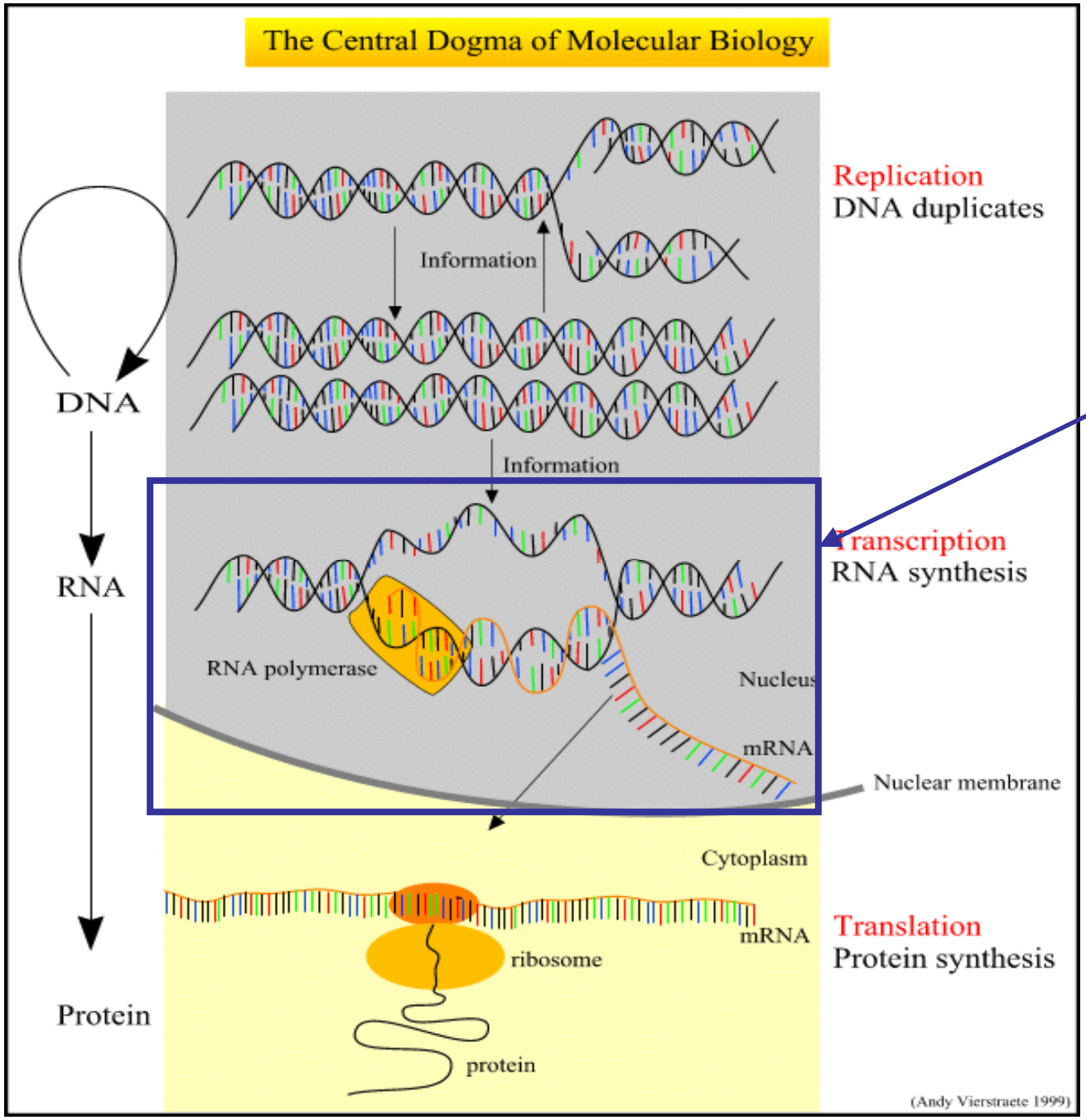
Transcription start site identification in bacteria

Marko Djordjevic

Faculty of Biology, University of Belgrade

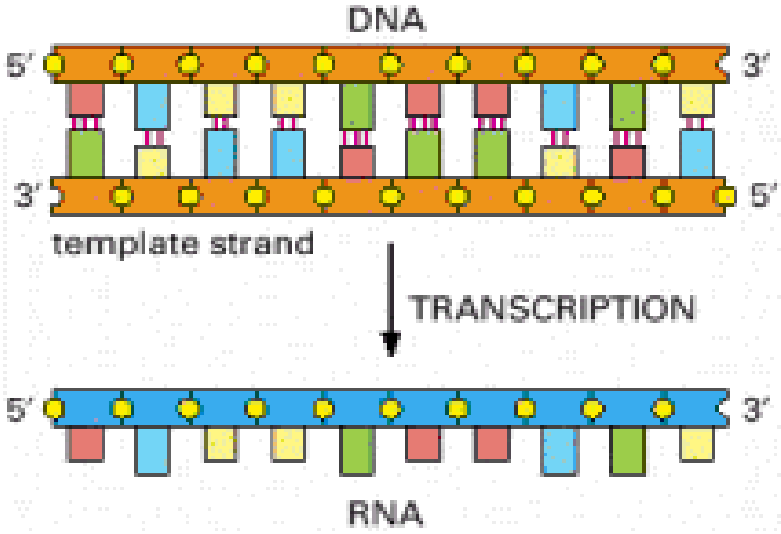




How is transcription initiated?

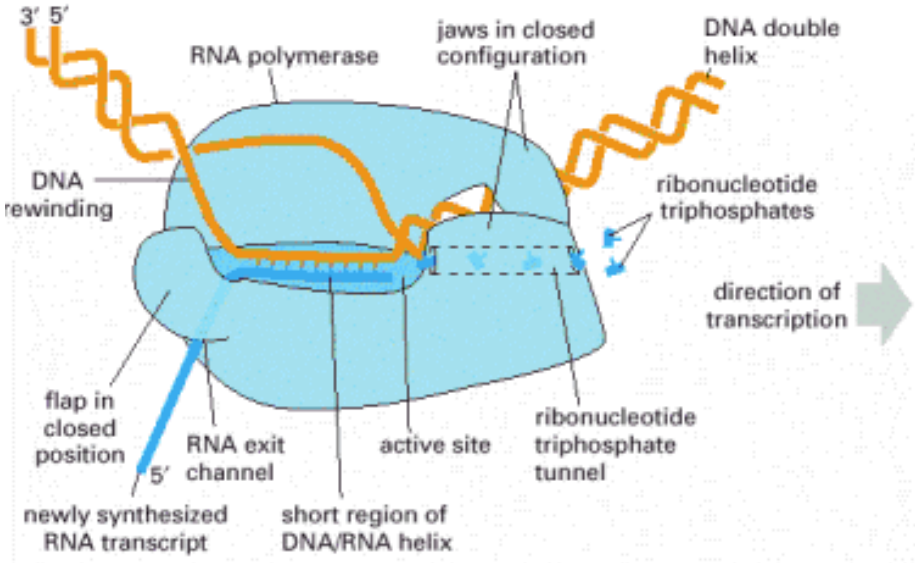
How to predict transcription start sites?

RNA polymerase



Synthesis of RNA from DNA template is called transcription.

Basic idea behind transcription is complementarity.

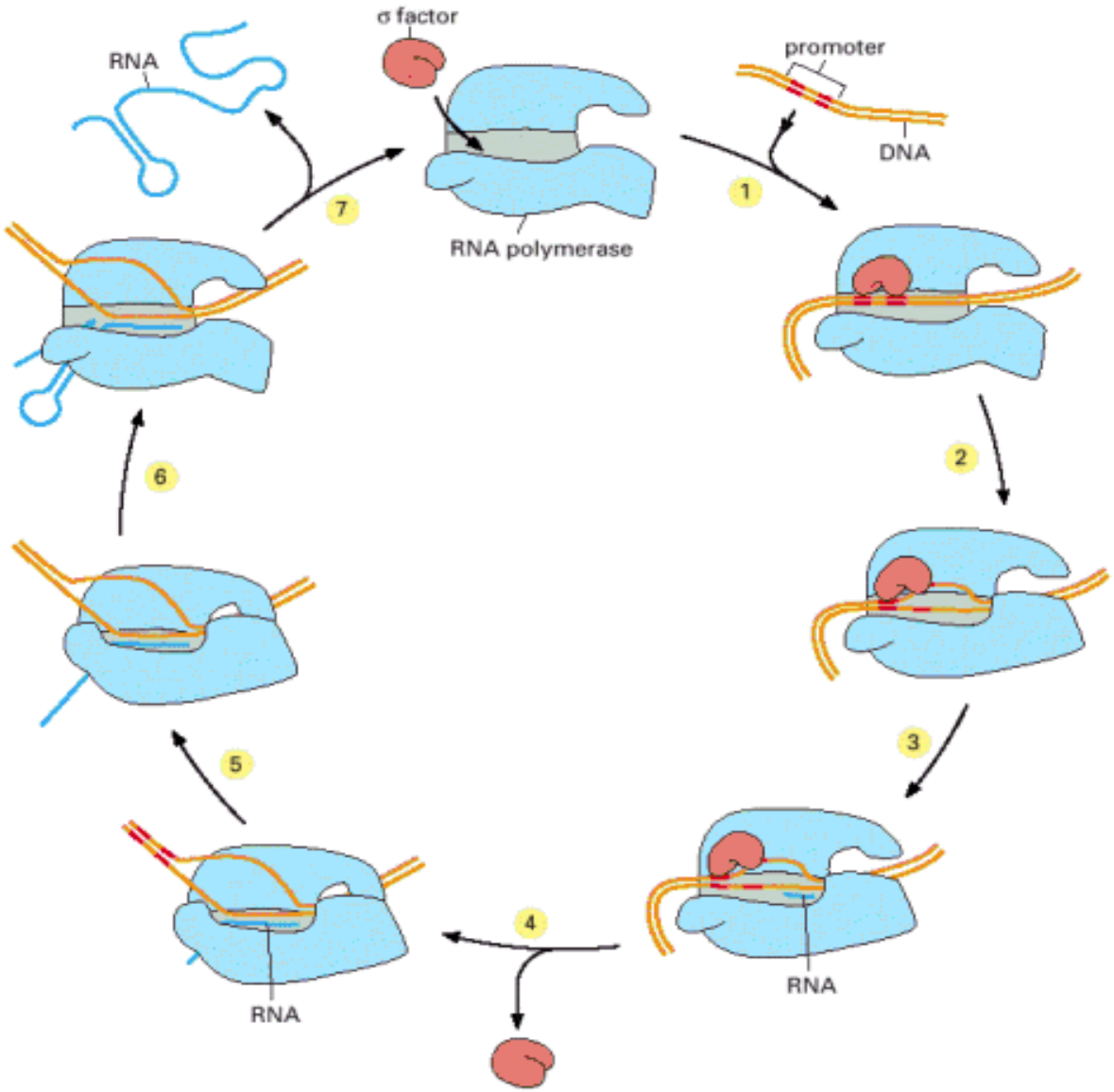


Molecular machine that exhibits transcription is RNA polymerase.

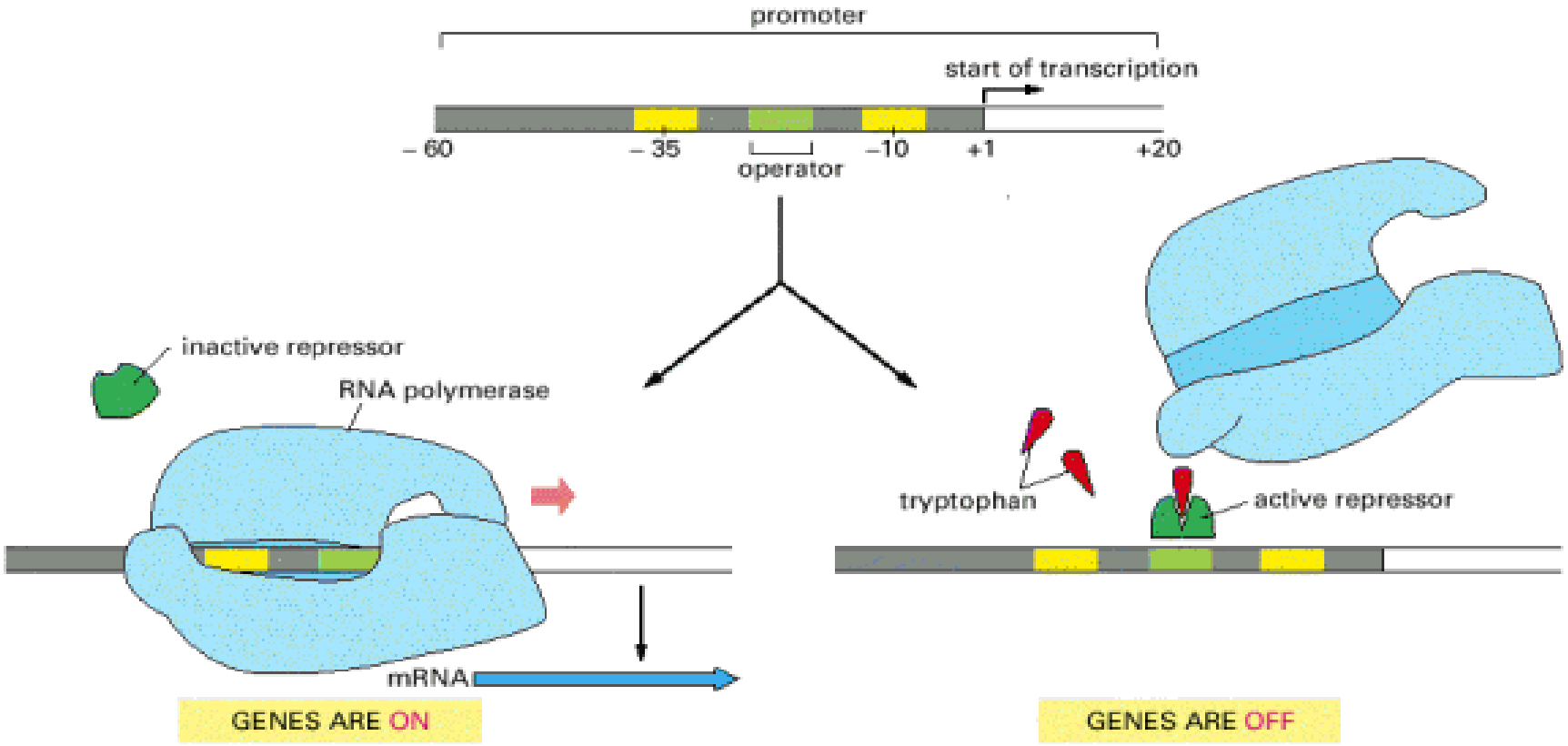
RNA polymerase is essential for life and is found in all living organisms.

2006 Nobel prize in Chemistry for RNA Polymerase awarded to Roger Kornberg.

Stages of transcription by RNA polymerase



Control of Gene Expression by Transcription Factors



Alberts *et al,* Molecular Biology of the cell.

Transcription start sites (TSS)

A starting point to understand transcription regulation

Necessary for gene and operon prediction

TSS detection in genome

Classical bioinformatic problem

Existing methods show poor accuracy
(a huge number of false positives)

RpoD15	27	37.4	1082	47	32,905	35
RpoD16	48	34.9	945	50	45,334	35
RpoD17	116	37.3	3138	51	138,293	30
RpoD18	34	38.0	394	50	31,666	32
RpoD19	25	38.2	877	43	50,286	30

Bacterial promoter structure

promoterxxxstrand	-35	spacer	-15	short -10
'accApxxxforward'	'TTGCTA'	[17]	'AGGC'	'AAATT'
'accBpxxxforward'	'TTGATT'	[17]	'GACC'	'AGTAT'
'accDpxxxreverse'	'TATCCA'	[19]	'TGTT'	'TTAAT'
'aceBpxxxforward'	'TTGATT'	[16]	'GAGT'	'AGTCT'
'acnAp1xxxforward'	'CTAACA'	[15]	'GCCT'	'TTATA'
'acnAp2xxxforward'	'TCAAAT'	[19]	'TGTT'	'ATCTT'
'acnBxxxforward'	'TTAACA'	[17]	'TGCT'	'ATTCT'
'adhEp1xxxreverse'	'CTAATG'	[17]	'TACT'	'ACAAT'

CAAATT
CAGTAT
TTTAAT
TTTATA
TATCTT

TATAAT ⟵ consensus sequence

Weight matrix

A	-38	19	1	12	10	-48
C	-15	-38	-8	-10	-3	-32
G	-13	-48	-6	-7	-10	-48
T	17	-32	8	-9	-6	19

Basic difficulty: motifs that bacterial promoter are highly degenerated

What are possible problems?

Kinetic effects are important?

Poised promoters: Sites where RNAP binds with high affnity, but opens the two DNA strands too slowly for functional transcription.

What kinetic parameters are relevant for promoter recognition?

Alignment is not acurate?

Additional motifs determine specificity?

Talk Overview

PART I

A biophysical model of transcription initiation in bacteria

(Biophys J. 2008;94(11):4233)

PART II

Estimate importance of kinetic effects

(Integrative Biol. 2013; 5(5):796)

PART III

More accurate alignment of promoter elements

(J Bacteriol. 2011;193(22):6305)

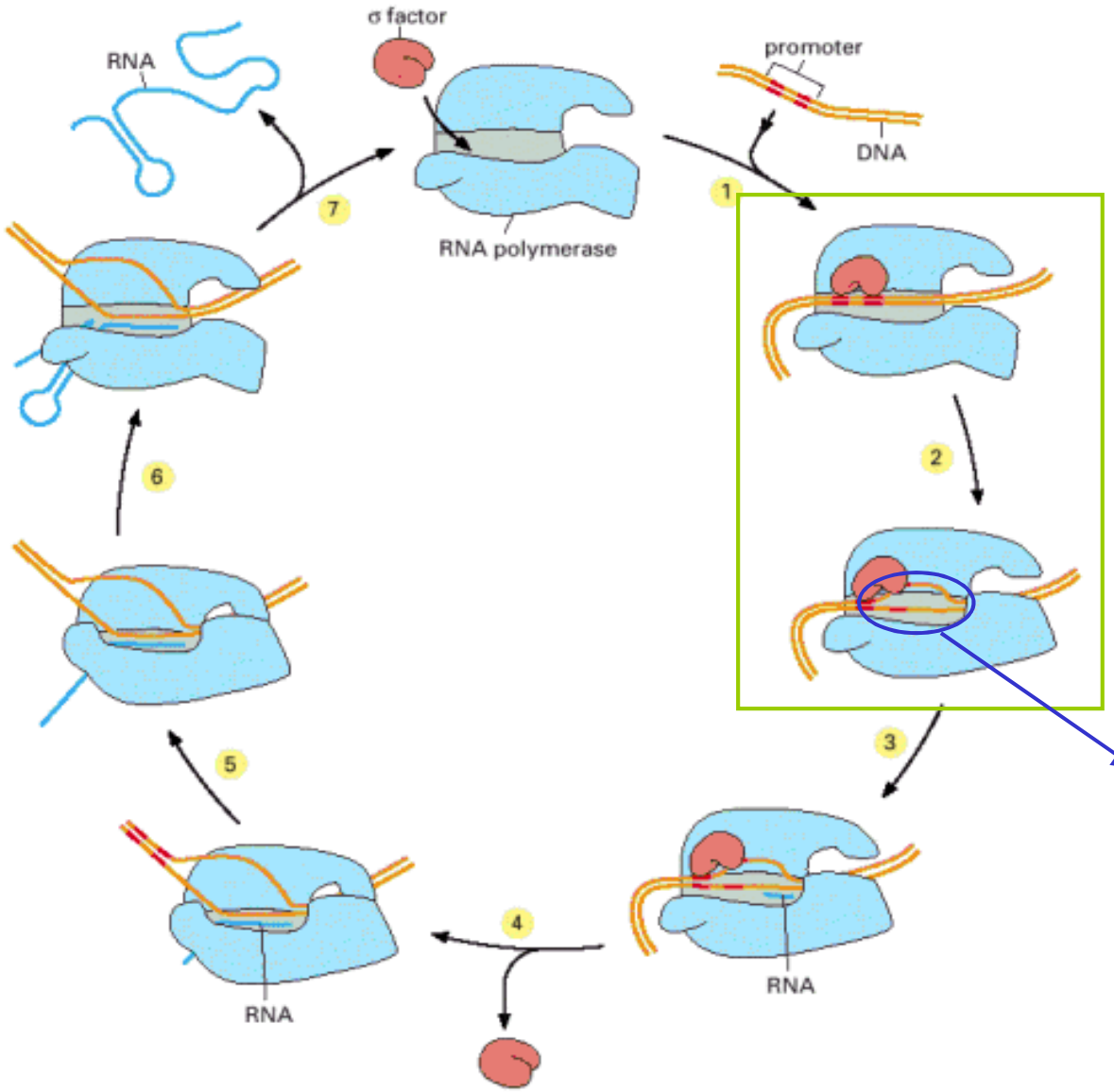
Beginning of an algorithm

(J Mol Biol. 2012;416(3):389)

PART I

A biophysical model of transcription initiation

Stages of transcription by RNA polymerase

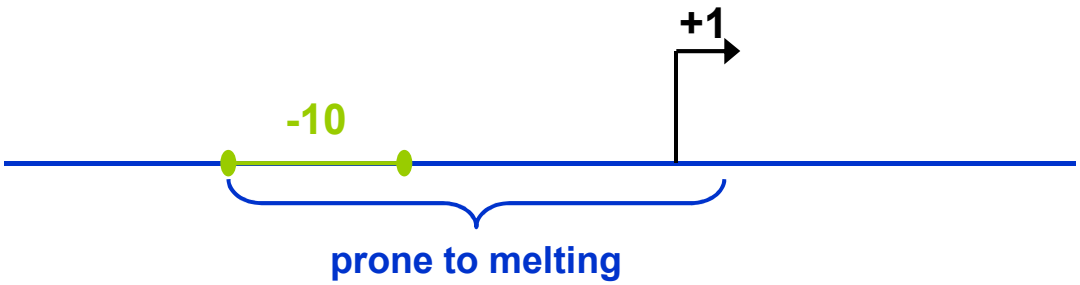


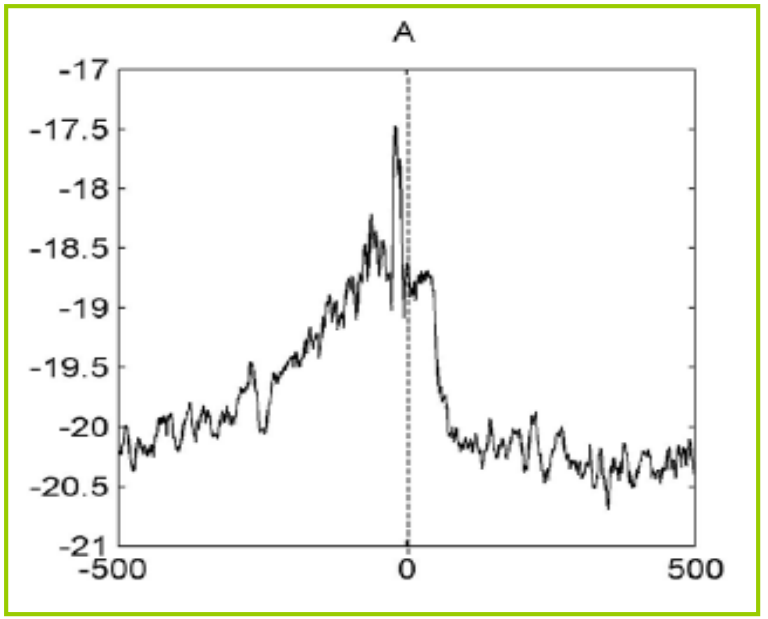
The open complex formation is the first step in transcription initiation.

RNAP opens two strands of DNA, so that a transcription bubble of ~15bps is formed.

Recent findings

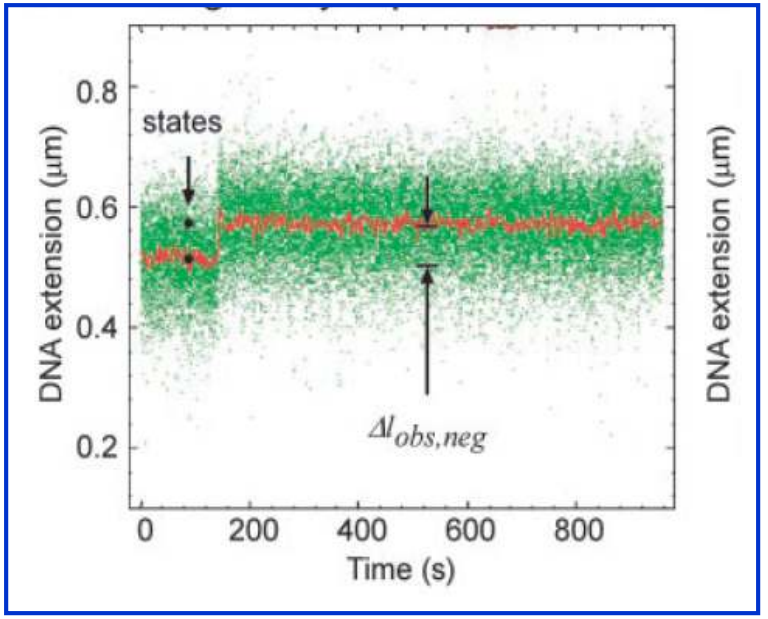
Bioinformatic study shows that region of ~15bps immediately upstream of transcription start site is prone for melting.



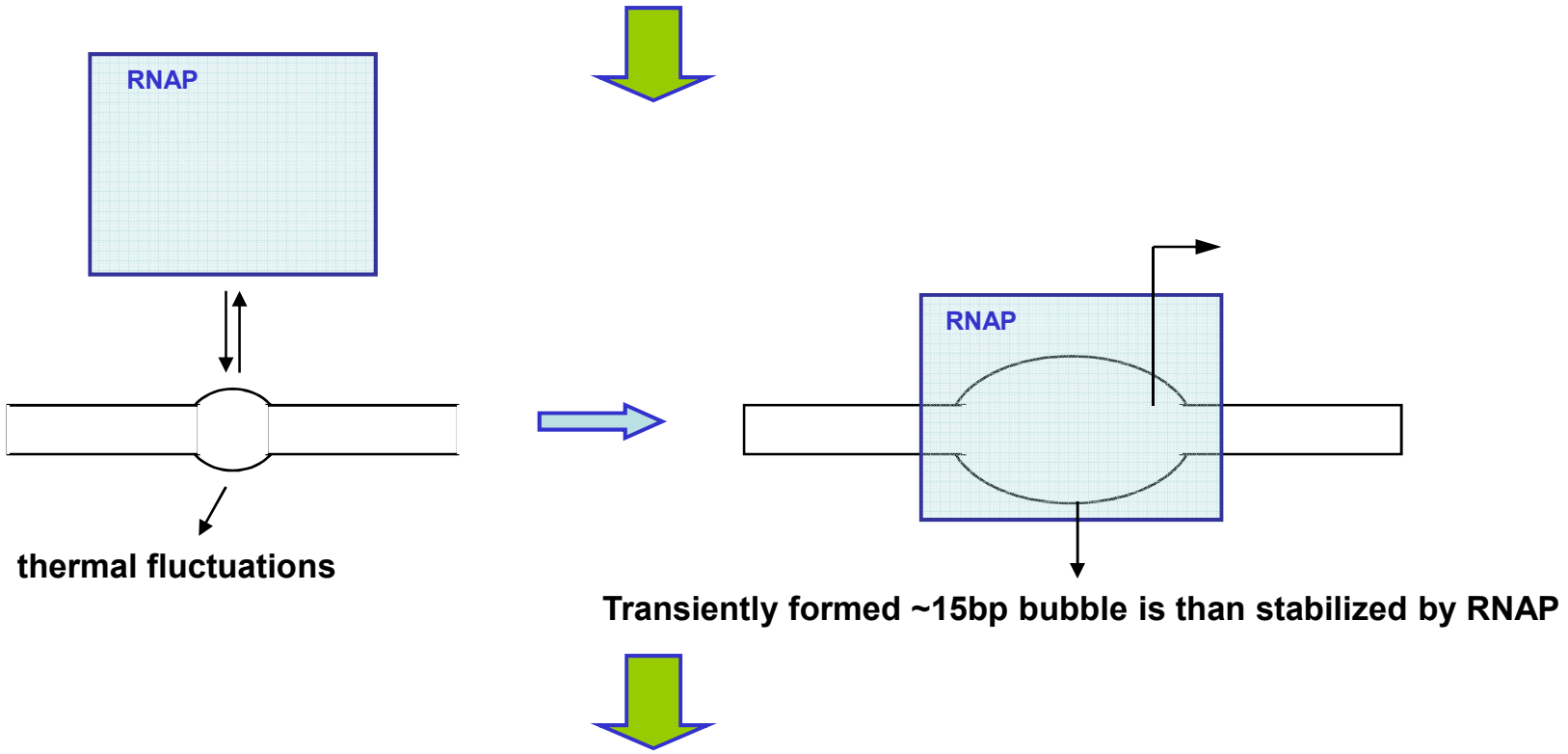


Single molecule experiments show that promoter region is melted in one step at least at the time resolution of 1s.

Since only short living intermediates (if any) exist, it is hard to directly experimentally test different hypothesis.



Bubble is formed in one step, through thermal fluctuations which transiently break bonds in dsDNA (DNA breathing).



In this simple model, the bubble formation is independent from RNAP, i.e. the role of RNAP is only to stabilize the final bubble.

M. Djordjevic and R. Bundschuh, *Biophys. J* 94 (11): 4223 (2008)

Biophysics of bubble formation in dsDNA

$$\Delta G_m(S) = \gamma + c\ln(l+1) + \Delta\tilde{G}_m(S)$$

Energy required to melt a bubble in DNA

Due to high initiation energy, bubble is formed cooperatively, i.e. as a zipper.

$$\frac{dp_l(t)}{dt} = k_- p_{l+1}(t) + k_+ p_{l-1}(t) - (k_+ + k_-) p_l(t)$$

Kinetics of bubble formation:

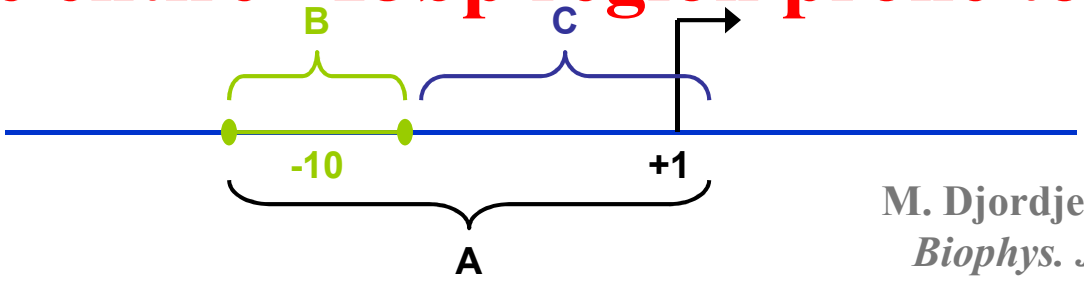
$$k_o = \frac{k_-}{l_0}\exp\left(\Delta G_m(S)/k_B T\right) \sim 10^{-8} - 10^{-11} 1/s$$

The rate of bubble opening

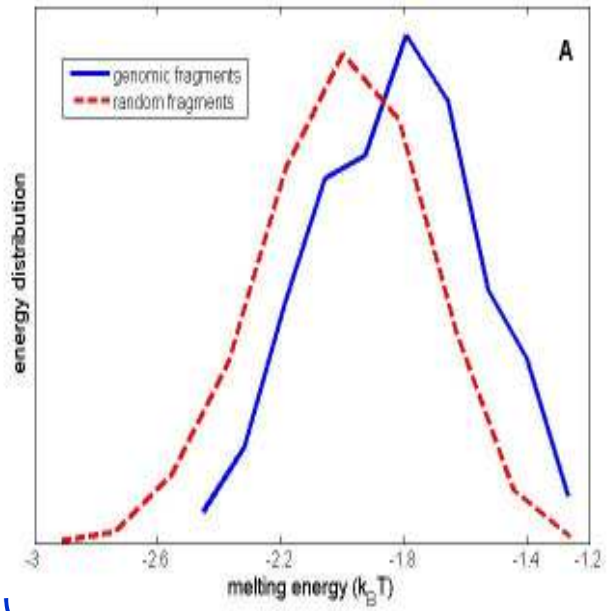
Between five and eight orders of magnitude larger compared to experimentally measured rates of bubble formation.

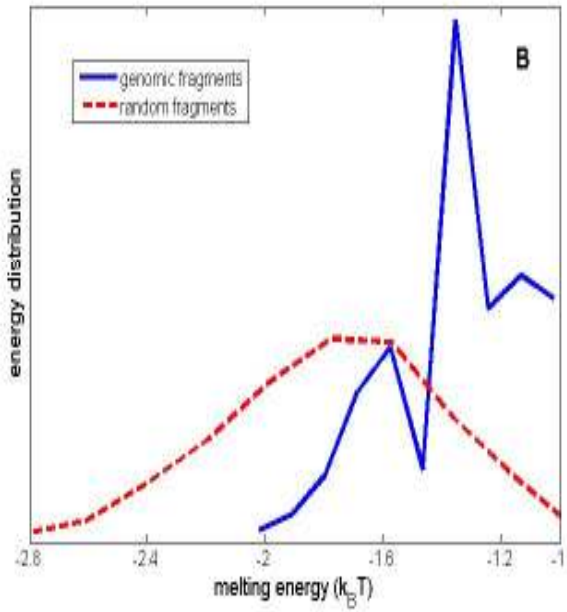
Poor agreement with the experiment!

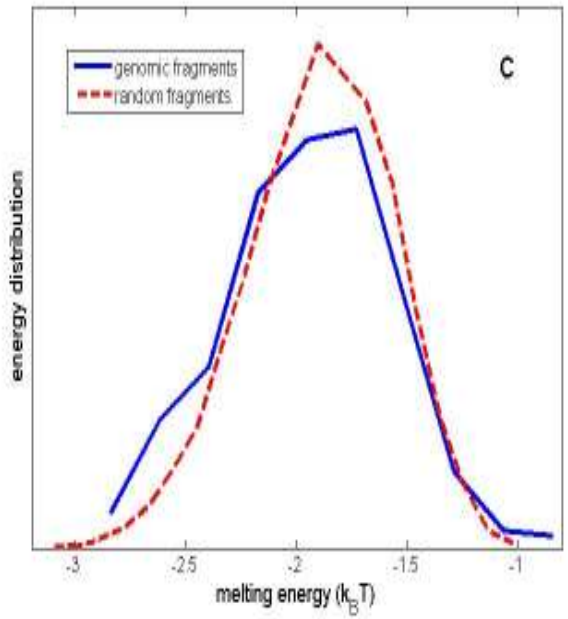
Why is the entire ~15bp region prone to melting?



M. Djordjevic and R. Bundschuh,
Biophys. J 94 (11): 4223 (2008)

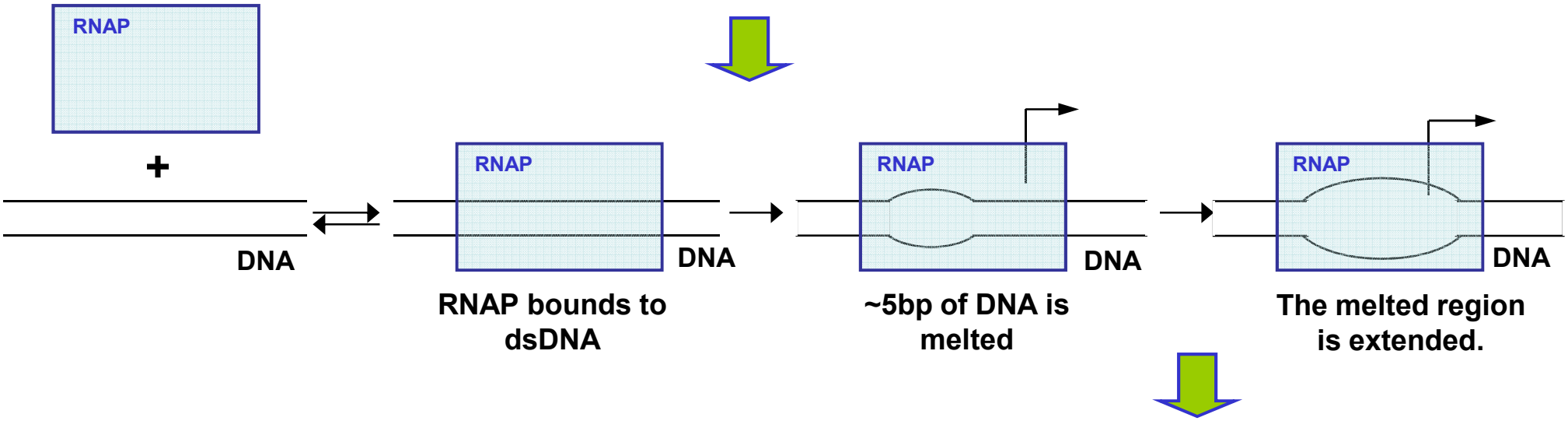






Reported melting destabilization of entire ~15bp transcription bubble is an artificial consequence of the fact that only -10 region is prone to melting!

In the first step, only -10 region is melted through thermal fluctuations facilitated by RNAP-ssDNA interactions.



M. Djordjevic and R. Bundschuh,
Biophys J 94 (11): 4223 (2008)

In the second step, the bubble extends towards the transcription start site.

First step has to be rate limiting (from the single-molecule experiment).

The hypothesis is consistent with recent structural data, indicating that aromatic residues of RNAP sigma subunit are ideally positioned to interact with transiently exposed -10 element single-stranded bases.

The rate of transition from closed to open complex

$$[RNAP]+[DNA] \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} [RNAP-DNA]_c \xrightarrow{k_{f1}} [RNAP-DNA]_{o1} \xrightarrow{k_{f2}} [RNAP-DNA]_o$$

Melting of -10 region is rate determining

$$k_f\left(S_{(-10)}\right) \approx k_{f1}\left(S_{(-10)}\right) \sim \exp\left(\frac{\Delta G_m\left(S^*_{(-10)}\right) + \Delta G_{ds}\left(S^*_{(-10)}\right) - \Delta G_{ss}\left(S^*_{(-10)}\right)}{k_B T}\right)$$

energy to melt -10 region

interaction with dsDNA

interaction with ssDNA

$\Delta G_m\left(S^*_{(-10)}\right)$

DNA melting energy parameters extensively measured (Santa Lucia)

$\Delta G_{ds}\left(S^*_{(-10)}\right)$

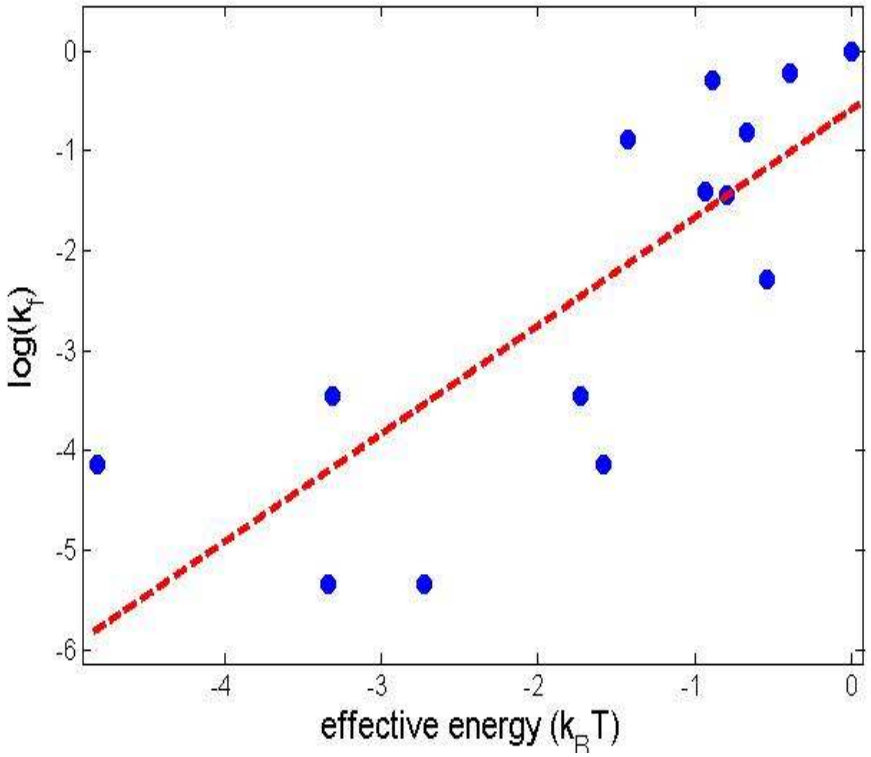
Measured at lower temperature to prevent DNA melting

$\Delta G_{ss}\left(S^*_{(-10)}\right)$

Measured with DNA construct mimicking open complex

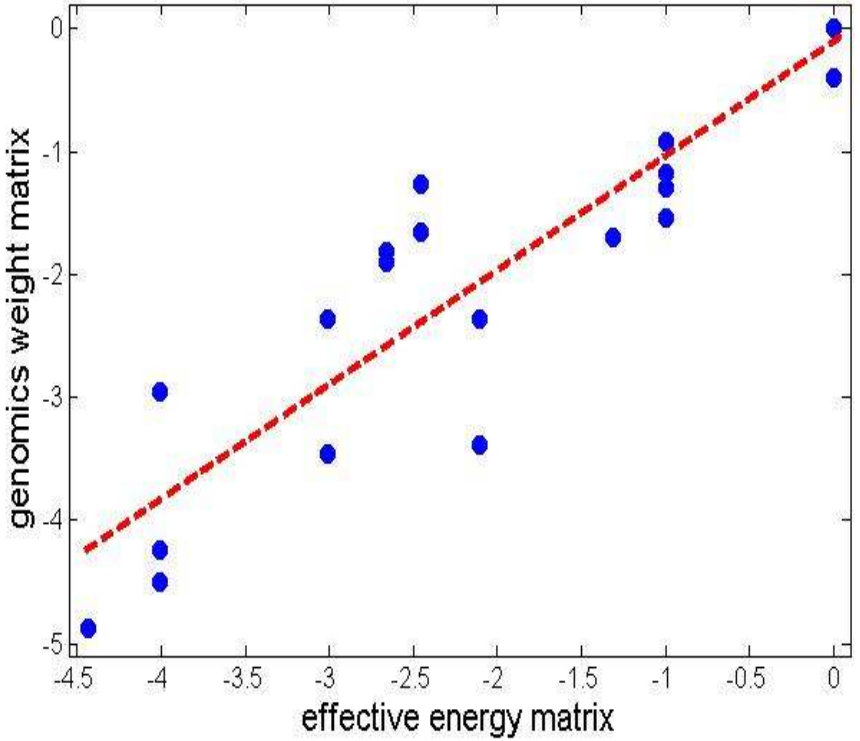
Comparison of the model with experimental data

Biochemical data



Reasonably high correlation constant (0.79) and statistically highly significant (P~10^{-3}).

Genomics data



Very good agreement with high correlation constant (0.93) and highly significant P value (10^{-11}).

Conclusion I

- **The results strongly support qualitative hypothesis, by which the open complex is formed as a two step process, where the first rate-limiting step consists of melting the upstream part of the transcription bubble through DNA breathing facilitated by RNAP-DNA interactions.**

- **We derived an explicit (simple) relationship connecting transcription initiation rate with measured physical properties of promoter-DNA and RNAP-DNA interactions (DNA melting energy and RNAP-DNA interaction energy in closed and open complex).**

- **Bioinformatic applications: allow efficient analysis of kinetic properties of DNA sequences on the whole genome scale.**

M Djordjevic and R Bundschuh, *Biophys. J* 94 (11): 4223 (2008)

PART II

Estimating kinetic effects

Kinetics of transcription initiation

Poised promoters - Locations in genome where RNAP binds with high affinity, but has a low rate of transcription initiation.

Is RNA polymerase kinetically poised at many locations in genome?

If yes, taking into account kinetic effects is likely necessary for accurate transcription start site detection

M Djordjevic, Integrative Biol. 2013; 5(5):796

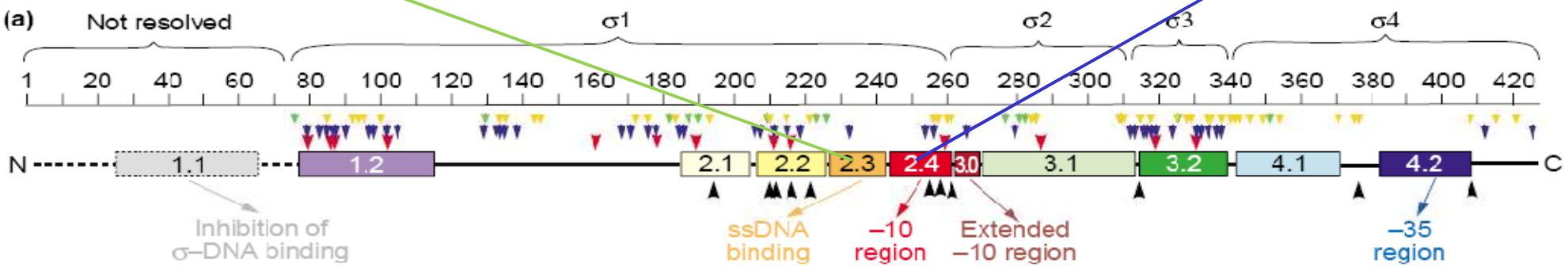
Poised promoters are determined by high binding affinity and low transcription initiation rate.

Rate of transcription initiation

Binding affinity

Depends on interaction energy of RNAP with ssDNA, and on DNA melting energy.

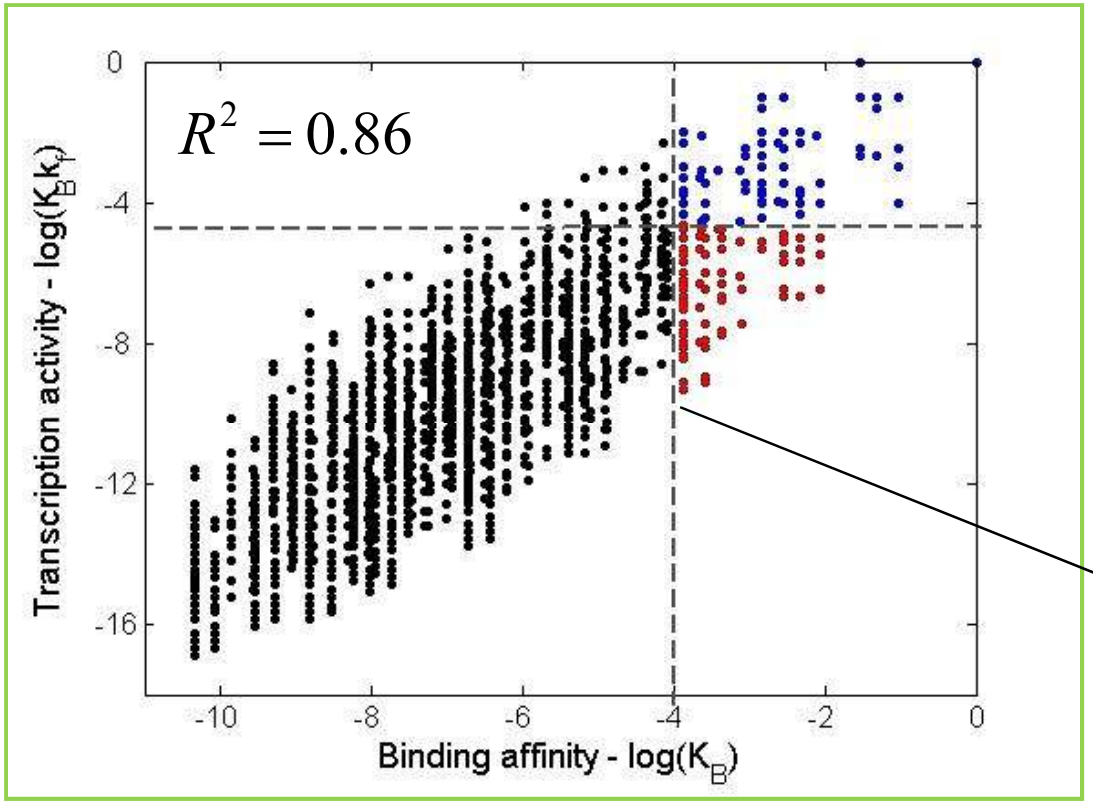
Depends on interaction energy of RNAP with dsDNA.

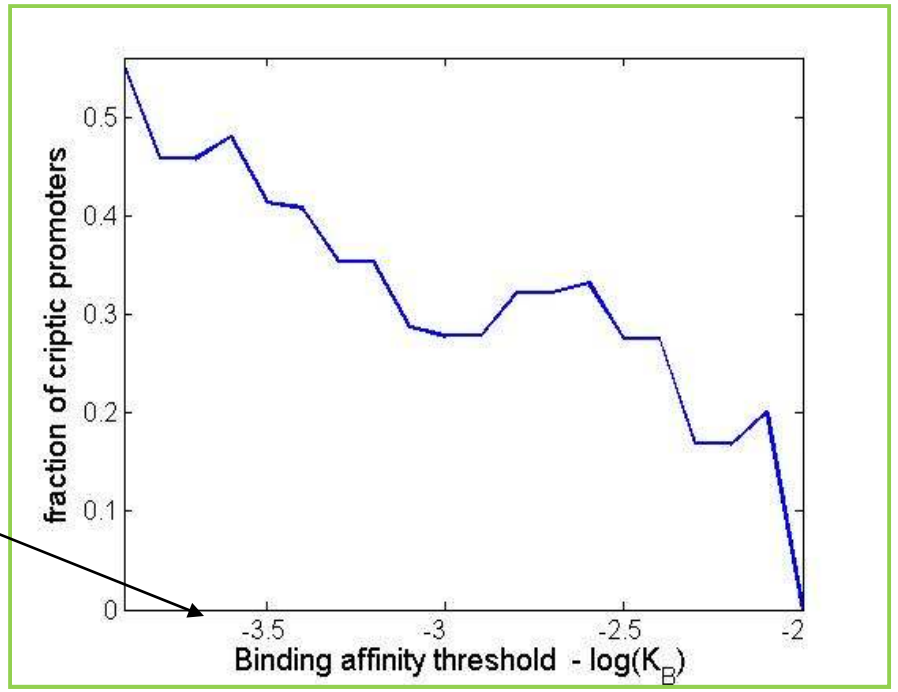


There is no a-priori reason for why binding affinity and the rate of transcription initiation should be related to each other.

For every sequence in E. coli intergenic regions we calculate transcription activity and binding affinity

M Djordjevic, Integrative Biol. 2013; 5(5):796

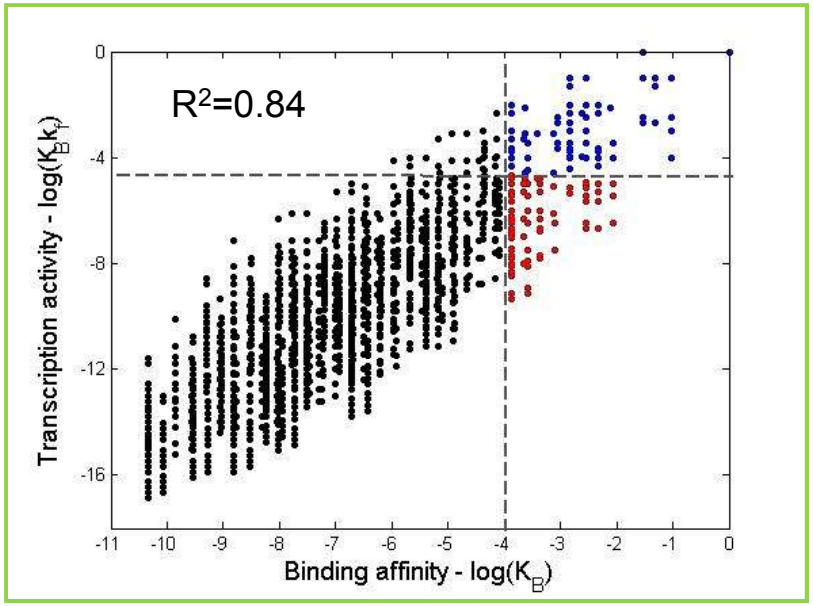




As we go to higher binding affinities, most (or all) of these strong binders correspond to functional promoters (i.e. to detectable levels of transcription).

What are the causes of good correlation between the binding affinity and the rate of transcription initiation?

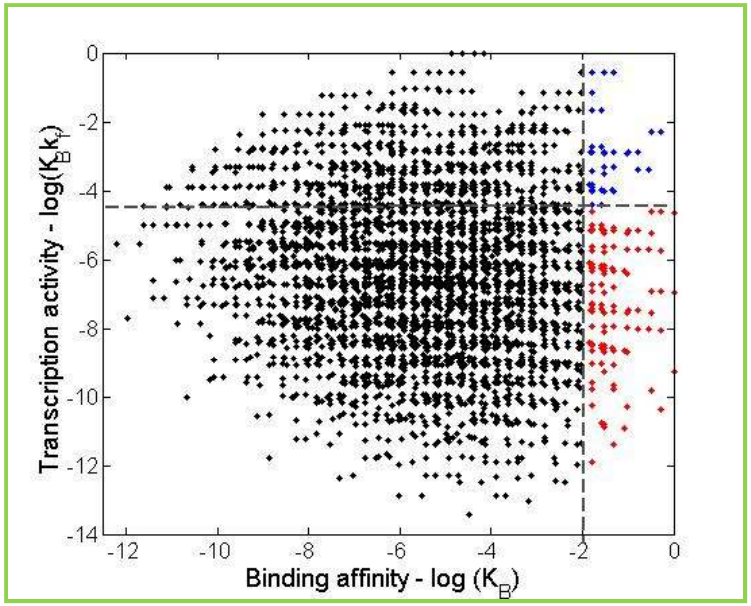
If DNA sequence in intergenic regions is randomized



Negligible decrease of correlation

Good correlation is not due to genome sequence!

If interaction energies of RNAP binding domains are randomly permuted

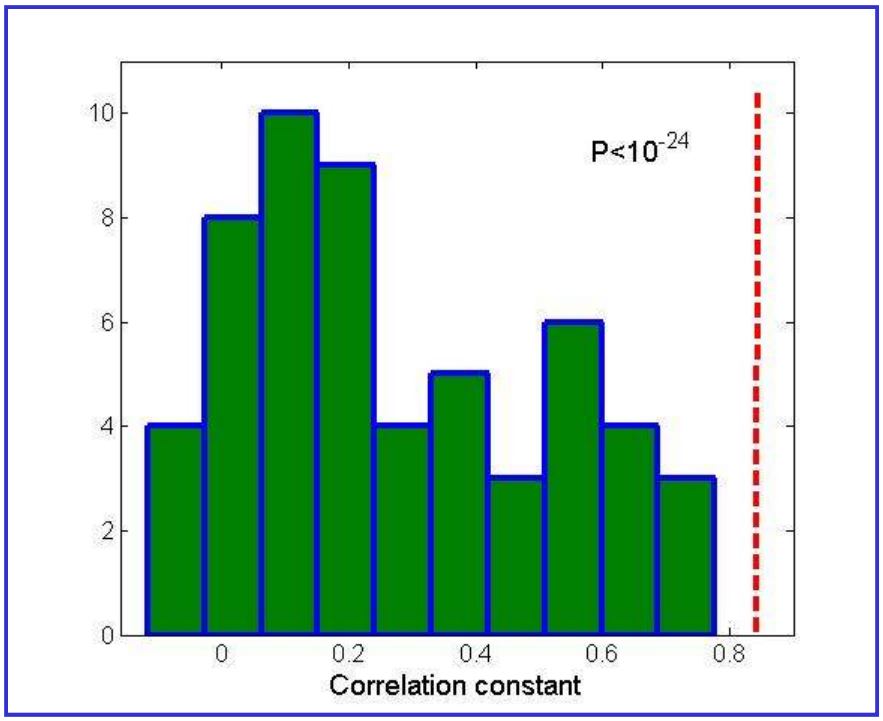


Correlation is completely lost

Good correlation is entirely on the level of RNAP protein domains.

Is the good correlation due to some generic property of DNA binding domains?

Substitute specificities binding domains 2.3 (ssDNA interactions) and 2.4 (dsDNA interactions), with those of different *E coli* DNA binding proteins.



Correlation constant corresponding to actual RNAP binding domains is larger compared to the correlation constants for other E Coli DNA binding domains.

Interaction domains of RNAP are `hardwired' so as to ensure evading poised promoters

M Djordjevic, Integrative Biol. 2013; 5(5):796

Conclusion II

RNAP DNA binding domains are disigned so as to reduce the extent of RNAP poising in genome.

There is still a substantial number of poised promoters in genome.

Kinetic effects should be taken into account in both experimental and bioinformatics searches of TSS

(M. D. and M. Djordjevic, in preparation).

PART III

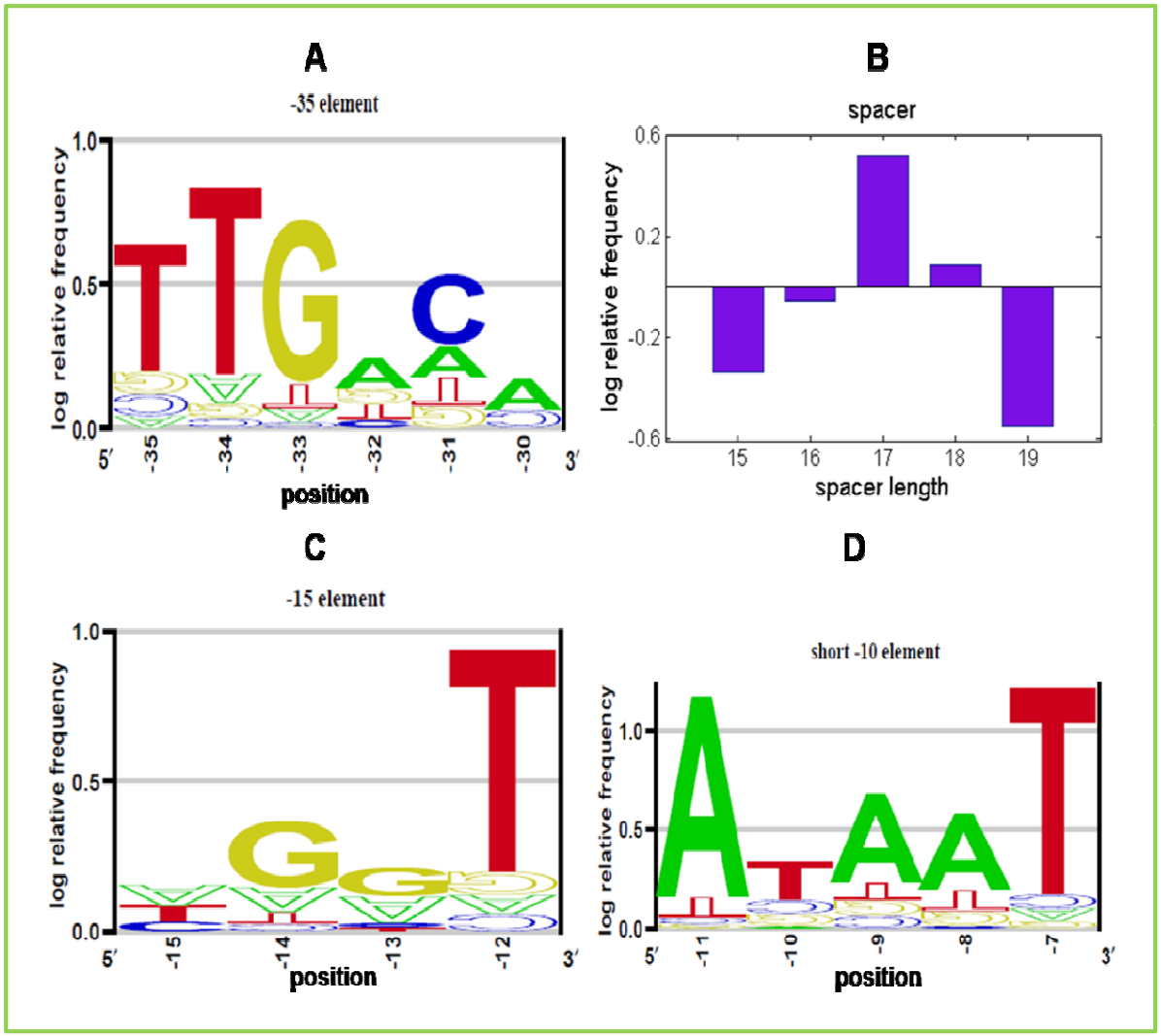
Redefining promoter sequence specificity

Alignment of promoter elements

- **Align promoter elements of ~300 experimentally detected TSS**
- **First align -10 elements through Gibbs search**
- **Use them as anchor to align -35 elements**
- **Perform iterative supervised search to improve the alignment**

Djordjevic M. J Bacteriol. 2011;193(22):6305

Specificity of promoter elements

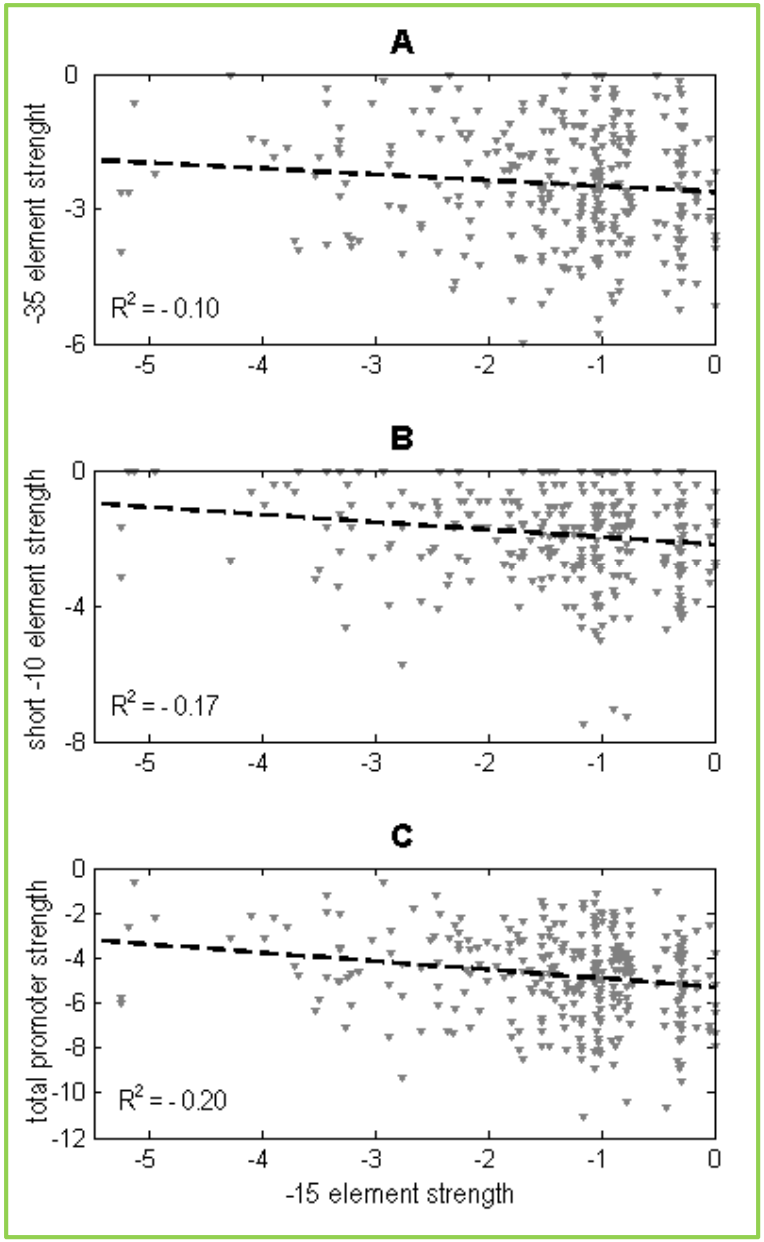


Qualitative differences with previously published alignments

A careful alignment allows detecting and constructing weight matrices for sequences outside of -10 and -35 element.

Djordjevic M. J Bacteriol. 2011;193(22):6305

Element strength correlation



-15 element and -35 element interact with RNAP in dsDNA form

-10 element interacts with RNAP in ssDNA form

Surprisingly, -15 element exhibits a significantly stronger negative correlation with total promoter strength than with -35 element.

Total promoter strength rather than binding affinity of RNAP to dsDNA determines functional promoter.

Djordjevic M. J Bacteriol. 2011;193(22):6305

Predictions with new alignment

Standard weight-matrix algorithm with new alignment can detect all experimentally found promoters in E. coli bacteriophage phiEco32.

(Pavlova O, *et al*., J Mol Biol. 2012,416(3):389)

Note: Bacteriophages have short genome sequence and strong promoters – relatively easy problem.

Conclusion and outlook

Explicit biophysical modeling is likely a proper framework for accurate TSS prediction.

Kinetic effects have to be taken into account

More careful alignments should increase search specificity.

Challenge: how to accurately parametrize the biophysical models

Acknowledgements

Ministry of Science and Education of Serbia

FP7 Marie Curie International Reintegration grant