Structural (and sequence-based) analysis of transcriptional regulation

Bruno Contreras-Moreira

.

bcontreras@eead.csic.es
http://www.eead.csic.es/compbio

Estación Experimental de Aula Dei, CSIC, Zaragoza, España

JNB08, Valencia

Credits

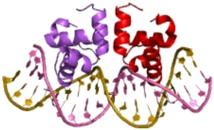
- **Universidad de Zaragoza, España**
  Vladimir Espinosa Angarica

Credits

- **Universidad de Zaragoza, España**
  Vladimir Espinosa Angarica
- **UNAM, México**
  Irma Lozada-Chávez
  Julio Collado-Vides

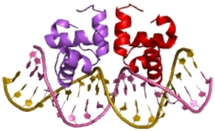
Purpose of this work

- **Motivation:** characterization of regulatory sequences by exploiting protein-DNA complexes at the Protein Data Bank

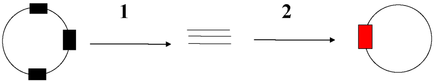


Purpose of this work

- **Motivation:** characterization of regulatory sequences by exploiting protein-DNA complexes at the Protein Data Bank



- **Value:** adding 3D information to sequence-based protocols



Dissecting protein-DNA interfaces in 3D

- Direct readout:

Dissecting protein-DNA interfaces in 3D

- Direct readout:
    - Hydrogen bonds between amino acid sidechains and N bases:

Dissecting protein-DNA interfaces in 3D

- Direct readout:
  - Hydrogen bonds between amino acid sidechains and N bases:
    - direct bonds

Dissecting protein-DNA interfaces in 3D

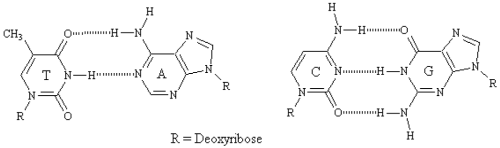
- Direct readout:
  - Hydrogen bonds between amino acid sidechains and N bases:
    - direct bonds
    - water-mediated bonds

Dissecting protein-DNA interfaces in 3D

- Direct readout:
  - Hydrogen bonds between amino acid sidechains and N bases:
    - direct bonds
    - water-mediated bonds
  - Hydrophobic interactions:

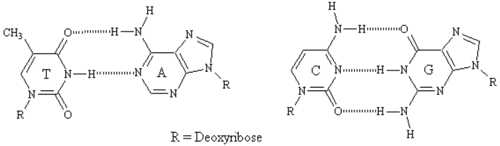
Dissecting protein-DNA interfaces in 3D

- Direct readout:
  - Hydrogen bonds between amino acid sidechains and N bases:
    - direct bonds
    - water-mediated bonds
  - Hydrophobic interactions:
    - C5M of thymine N-ring (Kono & Sarai)



Dissecting protein-DNA interfaces in 3D

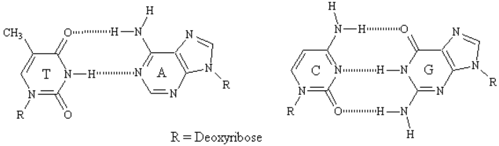
- Direct readout:
  - Hydrogen bonds between amino acid sidechains and N bases:
    - direct bonds
    - water-mediated bonds
  - Hydrophobic interactions:
    - C5M of thymine N-ring (Kono & Sarai)



- Indirect readout:

Dissecting protein-DNA interfaces in 3D

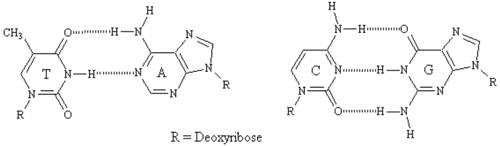
- Direct readout:
  - Hydrogen bonds between amino acid sidechains and N bases:
    - direct bonds
    - water-mediated bonds
  - Hydrophobic interactions:
    - C5M of thymine N-ring (Kono & Sarai)



- Indirect readout:
  - sequence-specific deformation of DNA base steps (Olson)

Dissecting protein-DNA interfaces in 3D

- Direct readout:
  - Hydrogen bonds between amino acid sidechains and N bases:
    - direct bonds
    - water-mediated bonds
  - Hydrophobic interactions:
    - C5M of thymine N-ring (Kono & Sarai)



- Indirect readout:
  - sequence-specific deformation of DNA base steps (Olson)
- Stabilizing interactions, not sequence-specific

Example of interface: E.coli NarL (1je8)

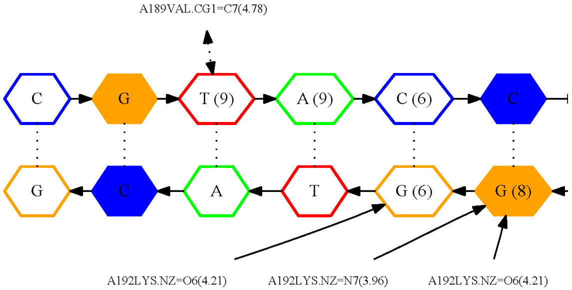


Figure: atomic interface dissected with modified version of HBPLUS (numbers in bases are total 4.5Åcontacts, Mirny)

Derivation of weight matrices for direct readout

- HBPLUS + nr50 library of protein-DNA complexes

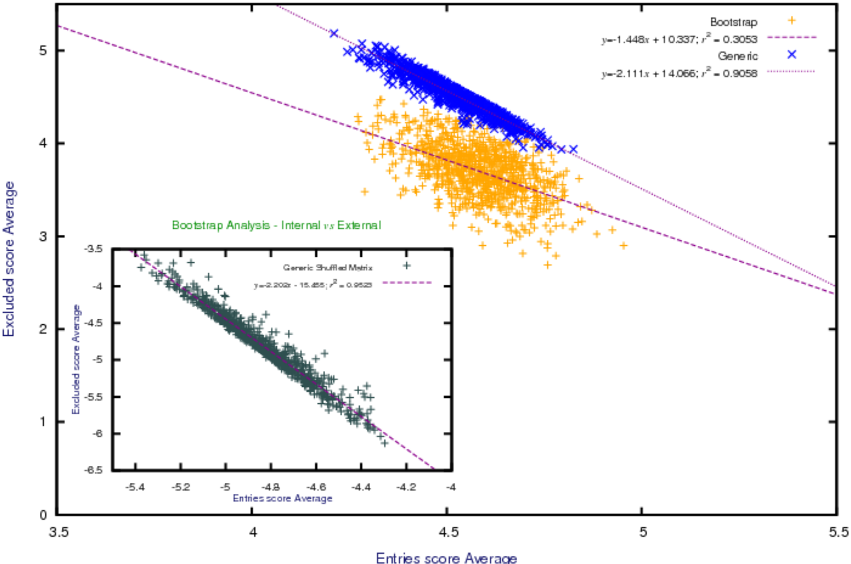
Derivation of weight matrices for direct readout

- HBPLUS + nr50 library of protein-DNA complexes
- weights calculated as $w_{pn} = ln(\frac{(n_{pn}+exp_{pn})/(N_p+1)}{exp_{pn}})$

H BONDS

		THYMINE			
		N1	O2	N3	O4
R					
	NE	-6.497	1.645(5)	-6.497	1.135(3)
	NH1	-6.497	3.127(22)	-6.497	1.135(3)
	NH2	-6.497	2.926(18)	-6.497	2.338(10)
K					
	NZ	-5.591	2.489(16)	-5.591	2.019(10)
S					
	OG	-4.625	2.545(5)	-4.625	1.629(2)
T					
	OG1	-4.382	2.408(3)	1.311(1)	2.695(4)
N					
	OD1	-5.509	-5.509	1.097(1)	-5.509
	ND2	-5.509	3.293(9)	-5.509	3.987(18)

Bootstraps of atomic interface matrices



Biological relevance of contact-based binding scores

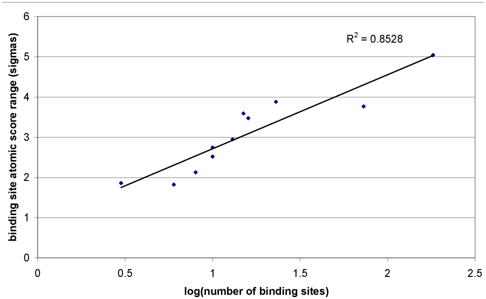
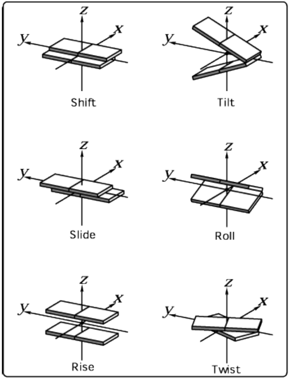


Figure: Atomic interface matrices yield scores that correlate with approximate measures of binding specificity for 11 E.coli TFs

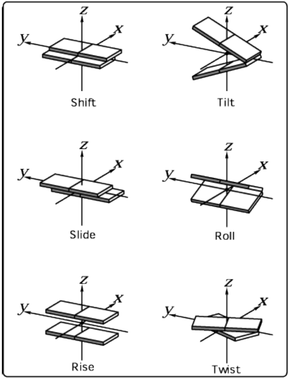
Calculating indirect readout contributions

- step geometry: rise, roll, shift, slide, tilt, twist (X3DNA)

Calculating indirect readout contributions

- step geometry: rise, roll, shift, slide, tilt, twist (X3DNA)
- 

Calculating indirect readout contributions

- step geometry: rise, roll, shift, slide, tilt, twist (X3DNA)
- 

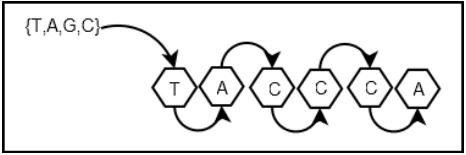
- $deformation = \sum_{st=0}^{n} \sum_{i=0}^{6} \sum_{j=0}^{6} spring_{ij} \Delta\theta_{i,st} \Delta\theta_{j,st}$ (Olson)

DNAPROT: in silico saturation mutagenesis of native DNA

- take coordinates of protein-DNA complex, with n base pairs

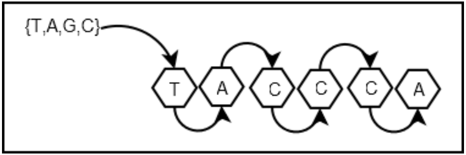
DNAPROT: in silico saturation mutagenesis of native DNA

- take coordinates of protein-DNA complex, with n base pairs
- starting from template DNA sequence, mutate $4n$ positions



DNAPROT: in silico saturation mutagenesis of native DNA

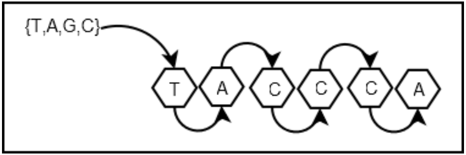
- take coordinates of protein-DNA complex, with n base pairs
- starting from template DNA sequence, mutate $4n$ positions



- score threaded mutations as $w(n) = e^{(-((1-D)*direct(n)+D*indirect(n)))}$ **(check D!)**

DNAPROT: in silico saturation mutagenesis of native DNA

- take coordinates of protein-DNA complex, with n base pairs
- starting from template DNA sequence, mutate $4n$ positions



- score threaded mutations as $w(n) = e^{(-((1-D)*direct(n)+D*indirect(n)))}$ **(check D!)**
- derive structure-based position weight matrix (PWM)

A	-1.38	0.36	-0.68	-2.23	-0.13	0.58
C	-0.95	0.09	1.07	1.28	0.12	-0.25
G	-1.10	-1.60	-0.91	-2.02	-0.08	-0.33
T	1.11	0.25	-1.79	-1.90	0.06	-0.34

Structure-based PWM for CRP (1cgp)

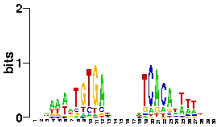


Figure: logo derived from cognate sites

Structure-based PWM for CRP (1cgp)

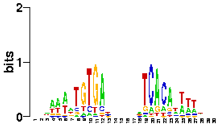


Figure: logo derived from cognate sites

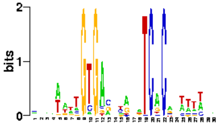
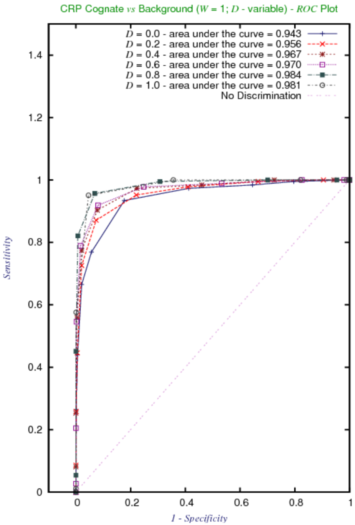


Figure: structure-based logo

Benchmark of DNAPROT with genomic-sized sequences



ROC curve of the structure-based PWM for CRP after scanning 10^6 random sequences (with similar %GC), mixed with 186 cognate sites extracted from RegulonDB 5.0

Comparison to a standard sequence-based method (CONSENSUS/PATSER)

- test set of binding sites in genomic sequences (true positives, TP)

Comparison to a standard sequence-based method (CONSENSUS/PATSER)

- test set of binding sites in genomic sequences (true positives, TP)
- score TPs and surrounding genomic sequences using

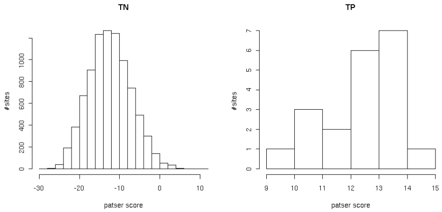
Comparison to a standard sequence-based method (CONSENSUS/PATSER)

- test set of binding sites in genomic sequences (true positives, TP)
- score TPs and surrounding genomic sequences using
  - sequence-based PWMs derived from *E.coli* cognate sites (PATSER)

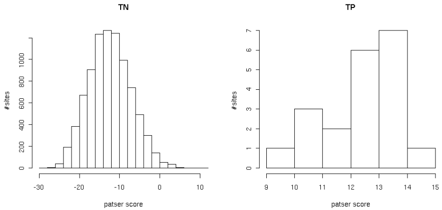
Comparison to a standard sequence-based method (CONSENSUS/PATSER)

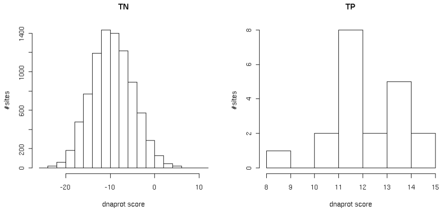
- test set of binding sites in genomic sequences (true positives, TP)
- score TPs and surrounding genomic sequences using
  - sequence-based PWMs derived from *E.coli* cognate sites (PATSER)
  - structure-based PWMs (DNAPROT + PDB complexes)

Comparison to PATSER predictions for PurR (20 TPs)



Comparison to PATSER predictions for PurR (20 TPs)

- 

- 

Comparison to a standard sequence-based method (CONSENSUS/PATSER)

TF	resol(A)	R(obs)	contacts	sites/genome	dnaprot50Z	ROC50Z	patserZ	corr
PurR(2puo)	2.9	0.16	6/6	20 / 8719	4.58	4.46	4.69	0.80
MetJ(1cma)	2.8	0.22	2/5	30 / 10631	2.05	2.17	2.9	0.35
NarL(1je8)	2.12	0.23	8/9	54 / 23327	2.42	1.42	2.79	0.42
CRP(1cgp)	3.0	0.24	6/8	613 / 202137	2.59	3.39	3.52	0.65
PhoB(1gxp)	2.5	0.25	4/4	17 / 6789	1.01	1.27	2.75	0.06
FadR(1h9t)	3.25	0.27	6/11	5 / 2169	1.99	1.66	4.73	-0.96

Table: Performance of DNAPROT compared to PATSER in terms of median Z-scores of true positive sites

Summary

- Given a good quality structural model of a TF-DNA complex,

Summary

- Given a good quality structural model of a TF-DNA complex,
- it is possible to build structure-based PWMs

Summary

- Given a good quality structural model of a TF-DNA complex,
- it is possible to build structure-based PWMs
- that drive the prediction of binding sites in genomes

Summary

- Given a good quality structural model of a TF-DNA complex,
- it is possible to build structure-based PWMs
- that drive the prediction of binding sites in genomes
- **Challenge:** can we use comparative models?