DNASITE: Comparative footprinting of DNA-binding proteins

Bruno Contreras-Moreira
contrera@ccg.unam.mx

Centro de Ciencias Genómicas
Universidad Nacional Autónoma de México

ISMB 2006, Fortaleza, Brasil

Purpose of this work

- **Idea:** identification of regulatory sequences by comparative modelling of protein-DNA complexes.

Purpose of this work

- **Idea:** identification of regulatory sequences by comparative modelling of protein-DNA complexes.
- **Motivation:**

Purpose of this work

- **Idea:** identification of regulatory sequences by comparative modelling of protein-DNA complexes.
- **Motivation:**
  - design experiments
  - improve description of regulatory networks

Background

- Related methods:

Background

- Related methods:
  - use collections of known binding sites (MEME,consensus)

Background

- Related methods:
  - use collections of known binding sites (MEME,consensus)
  - do not require previous knowledge of sites:

Background

- Related methods:
  - use collections of known binding sites (MEME,consensus)
  - do not require previous knowledge of sites:
    - phylogenetic footprinting

Background

- Related methods:
  - use collections of known binding sites (MEME,consensus)
  - do not require previous knowledge of sites:
    - phylogenetic footprinting
    - oligo analysis

Background

- Related methods:
  - use collections of known binding sites (MEME,consensus)
  - do not require previous knowledge of sites:
    - phylogenetic footprinting
    - oligo analysis
- DNASITE exploits the Protein Data Bank and builds on:

Background

- Related methods:
  - use collections of known binding sites (MEME,consensus)
  - do not require previous knowledge of sites:
    - phylogenetic footprinting
    - oligo analysis
- DNASITE exploits the Protein Data Bank and builds on:
  - previous work on crystallographic complexes (Kono & Sarai, Paillard & Lavery)

Background

- Related methods:
  - use collections of known binding sites (MEME,consensus)
  - do not require previous knowledge of sites:
    - phylogenetic footprinting
    - oligo analysis
- DNASITE exploits the Protein Data Bank and builds on:
  - previous work on crystallographic complexes (Kono & Sarai, Paillard & Lavery)
  - protein-DNA recognition codes (Mandel-Gutfreund & Margalit, Luscombe & Thornton)

Protein-DNA recognition matrices

```
# ln[fij/(fi x fj)]
#Mandel-Gutfreund and Margalit (1998) NAR,26: 2306-2312
#       G      A      T      C
GLY  -3.93  -3.93  -3.93  -3.93
ALA  -3.93  -3.93   0.66  -3.72
VAL  -3.93  -3.93  -0.17  -3.57
ILE  -3.93  -3.93   0.65  -3.44
LEU  -3.93  -3.93  -0.94  -3.93
PHE  -3.93  -3.93  -0.81  -0.12
TRP  -1.96  -3.93  -1.96  -3.93
TYR  -2.87  -2.87   0.54   0.13
MET  -2.58  -0.28   0.42  -0.28
CYS  -2.23   0.07  -2.23   0.07
THR  -3.46  -0.06  -0.06  -1.16
SER   0.42  -0.68  -0.28  -0.68
GLN  -0.09   1.16   0.31  -3.09
ASN   0.48   1.93   0.71   0.71
GLU  -3.93  -1.24  -3.93   0.55
ASP  -3.93  -3.37  -3.93   1.01
HIS   1.56   0.46   0.87  -0.23
ARG   2.74   0.34   1.25  -3.93
LYS   2.16  -0.08   0.21  -3.93
PRO  -3.93  -3.93  -0.30  -3.29
```

Comparative modelling of protein-DNA complexes

- Previous structural approaches require crystallographic protein-DNA complexes.

Comparative modelling of protein-DNA complexes

- Previous structural approaches require crystallographic protein-DNA complexes.
- We ask whether comparative/homology models can also be used:

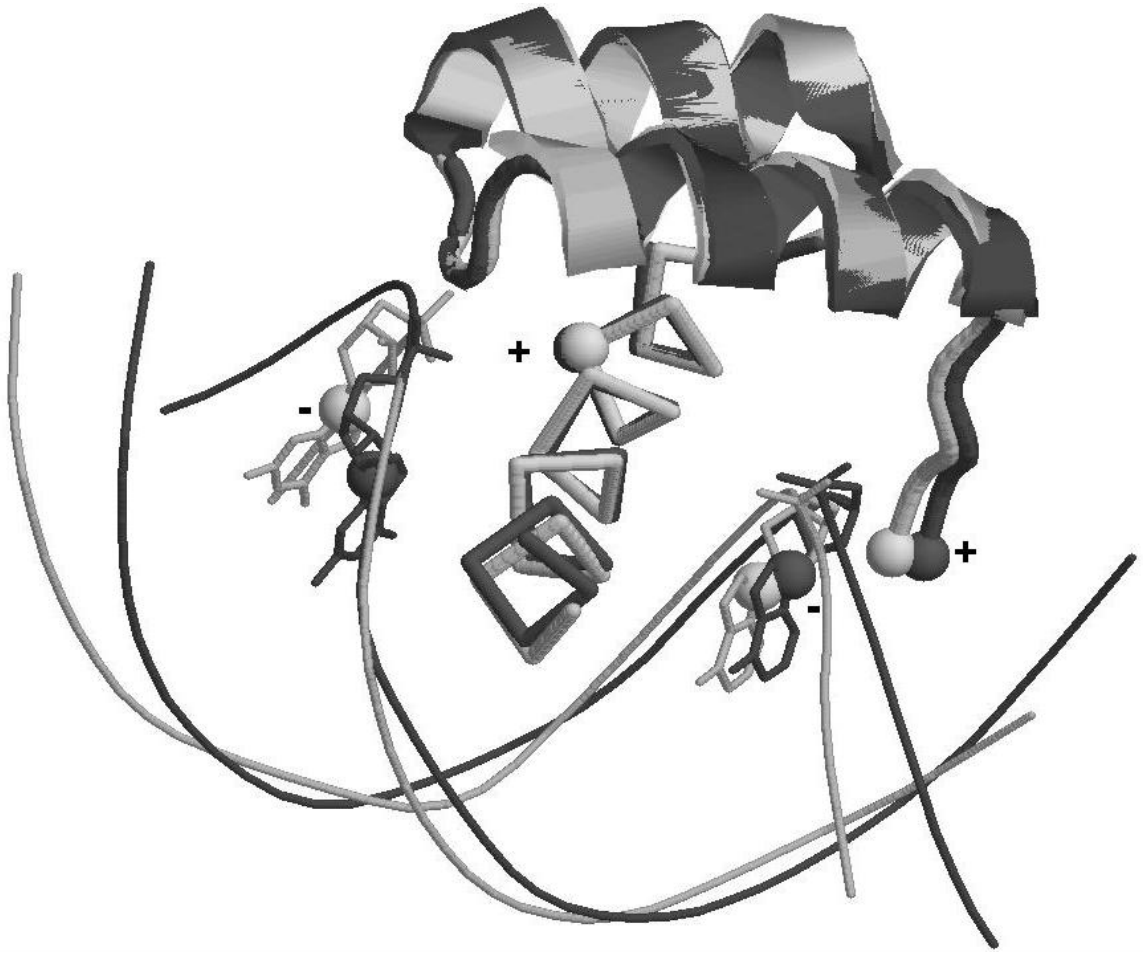
Comparative modelling of protein-DNA complexes

- Previous structural approaches require crystallographic protein-DNA complexes.
- We ask whether comparative/homology models can also be used:
  - do homologous DNA-binding proteins conserve their docking geometry?

Comparative modelling of protein-DNA complexes

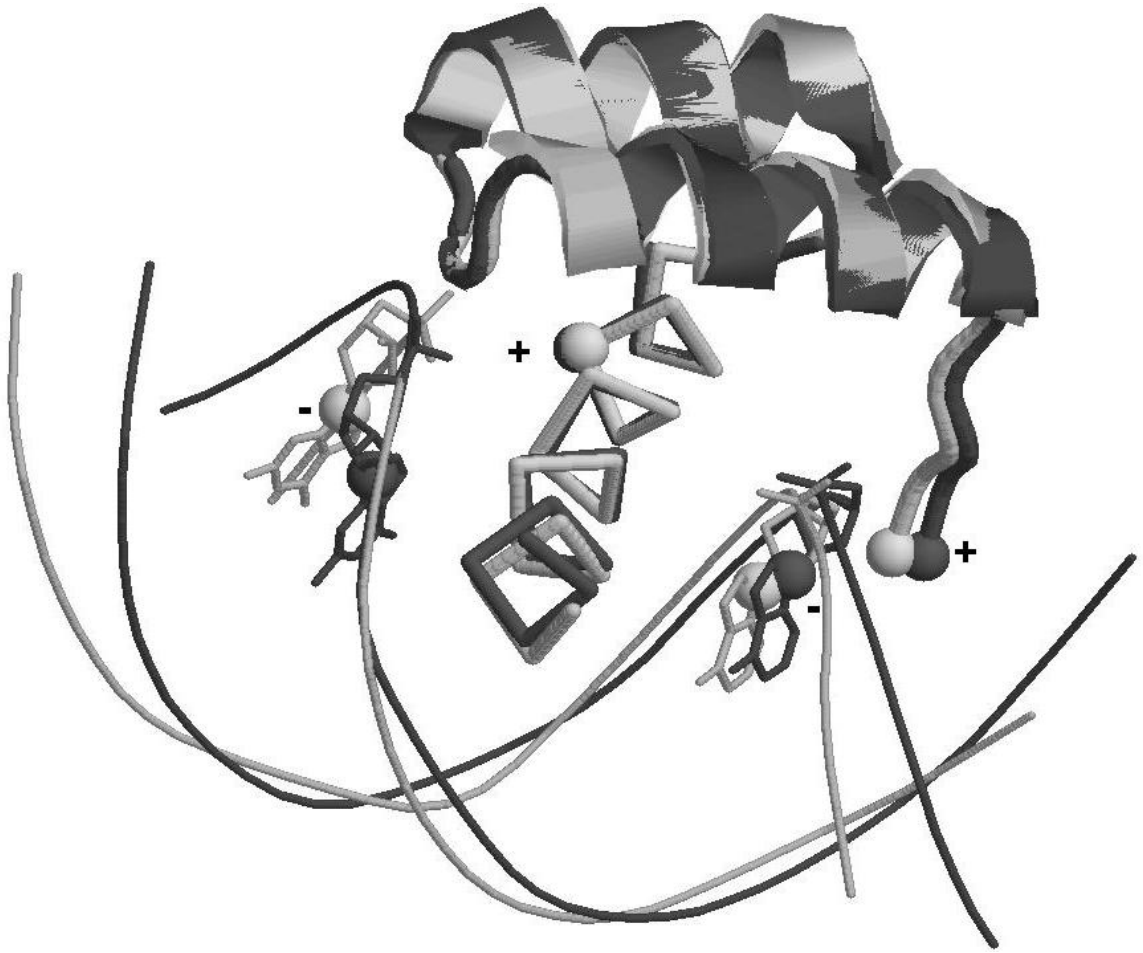
- Previous structural approaches require crystallographic protein-DNA complexes.
- We ask whether comparative/homology models can also be used:
  - do homologous DNA-binding proteins conserve their docking geometry?
  - can we identify modelled protein residues that contact DNA?

Interface comparison



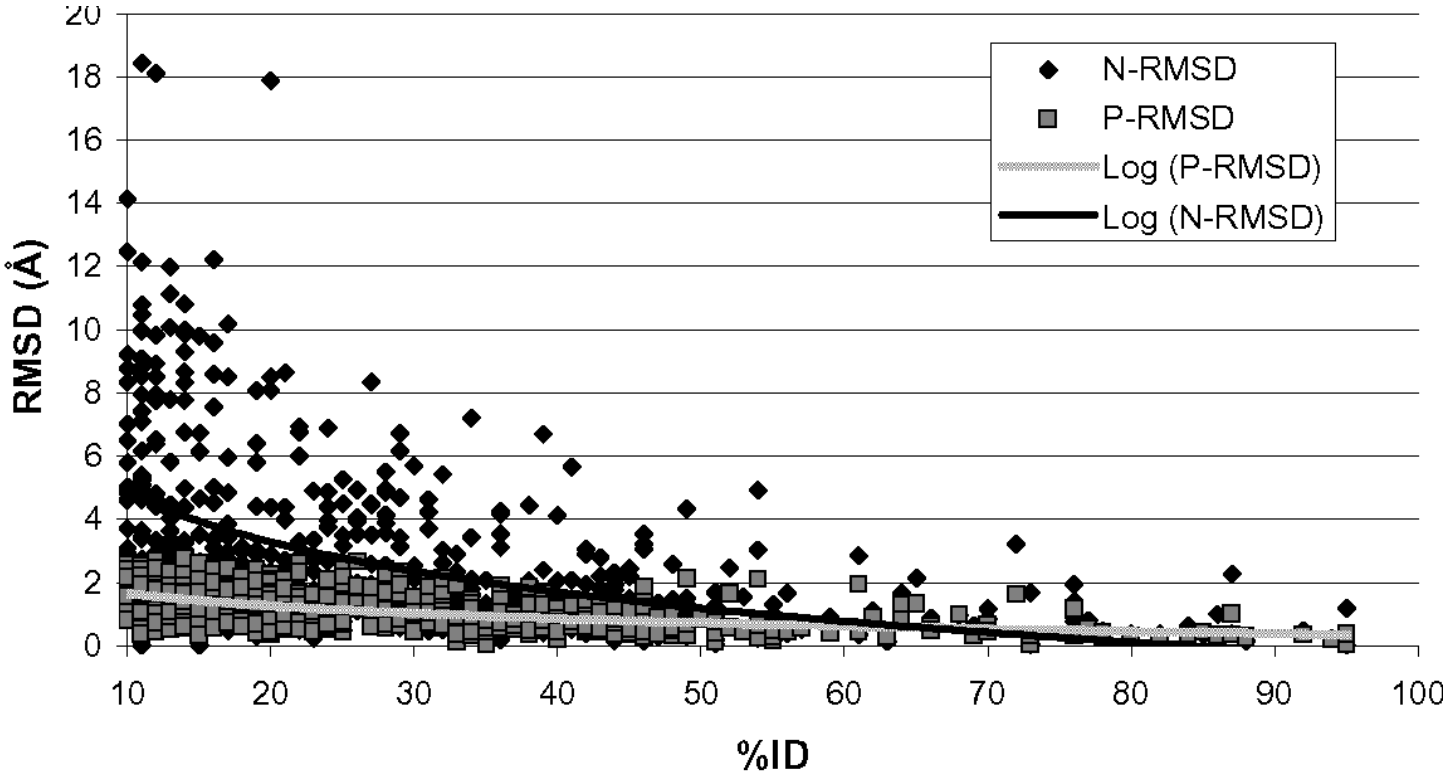
- interface atoms ($< 12\mathring{A}$):
  - (+) CA
  - (-) N1/N9

Interface comparison



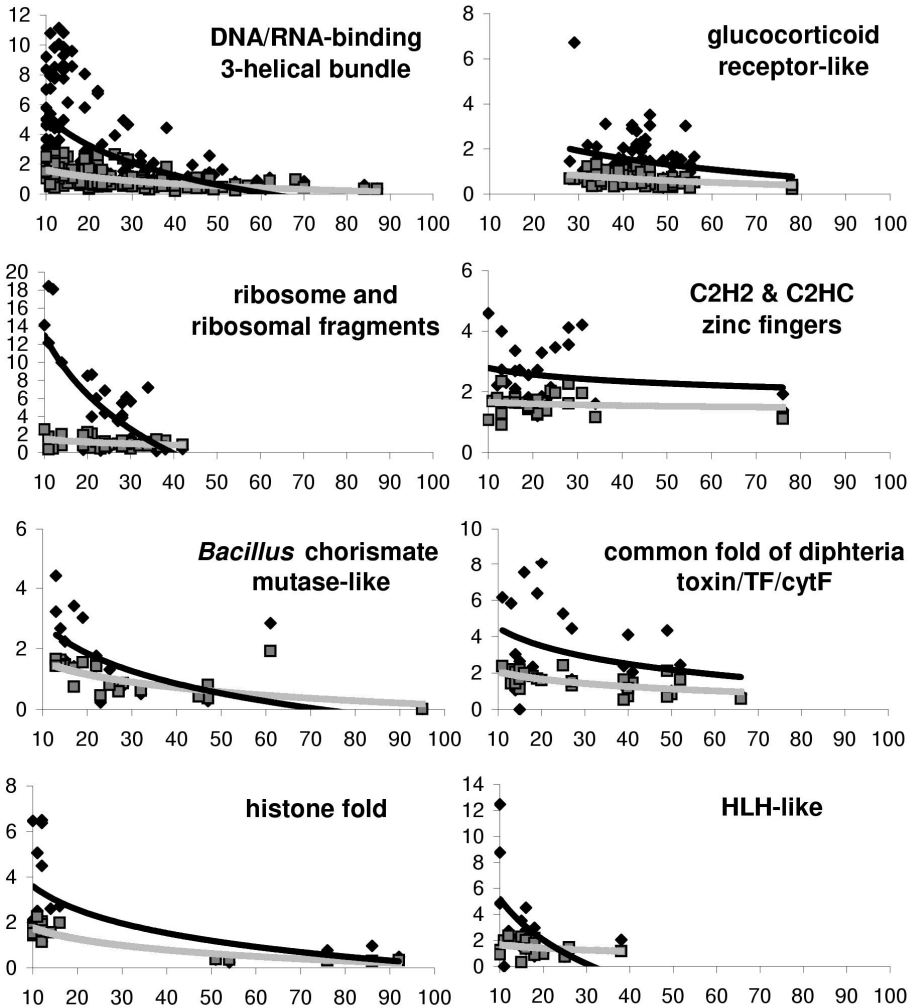
- interface atoms ($< 12\mathring{A}$):
  - (+) CA
  - (-) N1/N9
- RMSD calculated over MAMMOTH superimpositions

Homologous protein-DNA interfaces are conserved

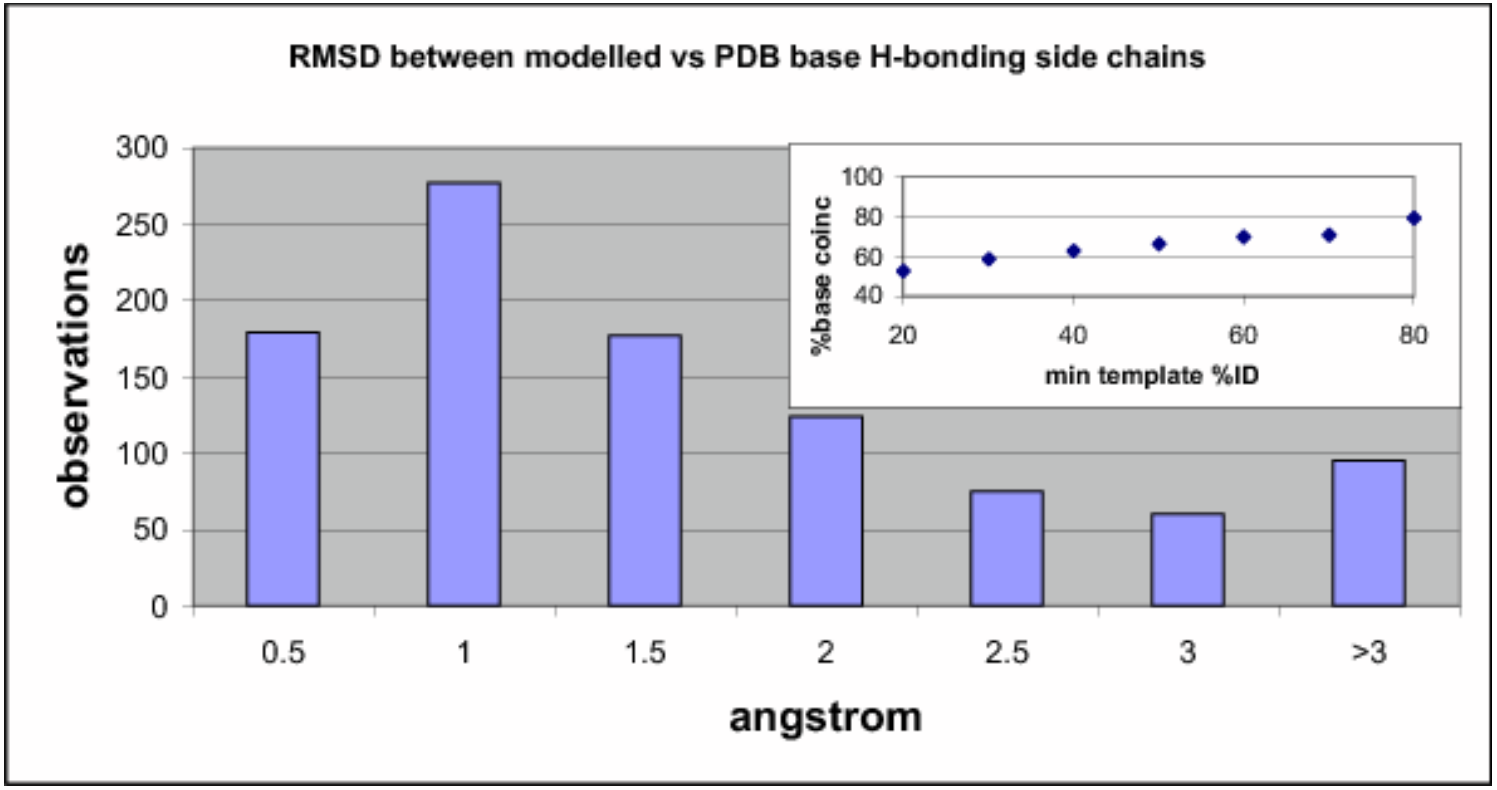


Median values for 442 pairs of superimposed PDB complexes.

SCOP folds show different interface conservation



Contact side chains can be modelled



987 base H-bonding residues modelled by SCWRL with templates >= 30%ID

Can we model protein-DNA complexes?

do DNA-binding proteins conserve their docking geometry?

YES, as a function of % sequence identity

can we identify modelled protein residues that contact DNA?

YES, at least we can model most H-bonding residues

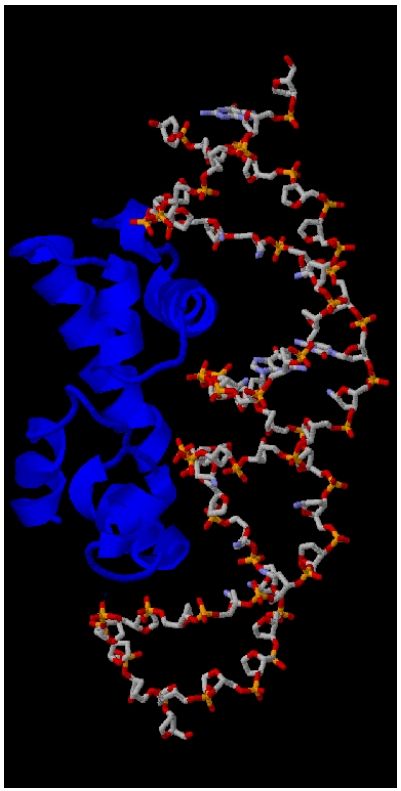
How DNASITE builds comparative models

- scan input protein sequence against library of PDB complexes (PSI-BLAST)

How DNASITE builds comparative models

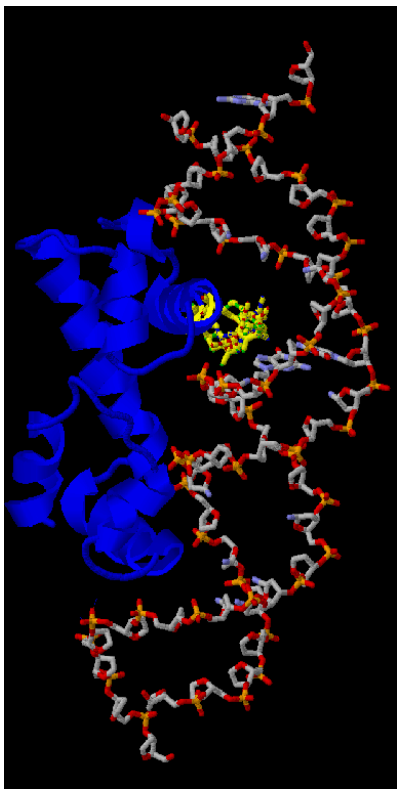
- scan input protein sequence against library of PDB complexes (PSI-BLAST)
- for each template PDB:

How DNASITE builds comparative models



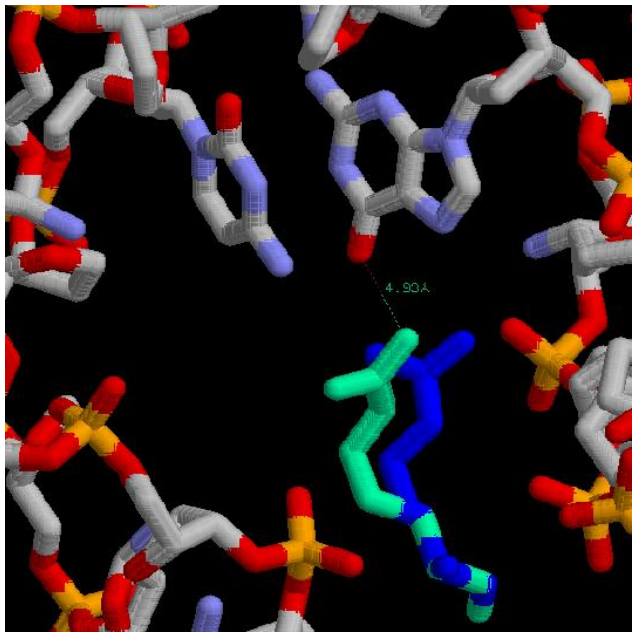
- scan input protein sequence against library of PDB complexes (PSI-BLAST)
- for each template PDB:
  - build comparative complex core

How DNASITE builds comparative models



- scan input protein sequence against library of PDB complexes (PSI-BLAST)
- for each template PDB:
  - build comparative complex core
  - model mutant protein side-chains (SCWRL)

How DNASITE builds comparative models



distance < 4.5Å from pur/pyr ring atoms, PSI-BLAST IC > 0.3

- scan input protein sequence against library of PDB complexes (PSI-BLAST)
- for each template PDB:
  - build comparative complex core
  - model mutant protein side-chains (SCWRL)
  - identify DNA-contacting residues

How DNASITE builds comparative models

$$s_i + N_{template} + P_{model} = PN_i$$

- scan input protein sequence against library of PDB complexes (PSI-BLAST)
- for each template PDB:
  - build comparative complex core
  - model mutant protein side-chains (SCWRL)
  - identify DNA-contacting residues
  - thread all? possible DNA sequences:

How DNASITE builds comparative models

$$score(P, N_i) = \sum\sum match(P_j, N_{ik}, matrix)$$

- scan input protein sequence against library of PDB complexes (PSI-BLAST)
- for each template PDB:
  - build comparative complex core
  - model mutant protein side-chains (SCWRL)
  - identify DNA-contacting residues
  - thread all? possible DNA sequences:
    - calculate protein-DNA agreement score (family corrected?)

How DNASITE builds comparative models

$$deform(s_i, N_{template}) = f(s_i, Olson, geom(N_{template}))$$

- scan input protein sequence against library of PDB complexes (PSI-BLAST)
- for each template PDB:
  - build comparative complex core
  - model mutant protein side-chains (SCWRL)
  - identify DNA-contacting residues
  - thread all? possible DNA sequences:
    - calculate protein-DNA agreement score (family corrected?)
    - estimate DNA deformation cost (X3DNA)

How DNASITE builds comparative models

- scan input protein sequence against library of PDB complexes (PSI-BLAST)
- for each template PDB:
  - build comparative complex core
  - model mutant protein side-chains (SCWRL)
  - identify DNA-contacting residues
  - thread all? possible DNA sequences:
    - calculate protein-DNA agreement score (family corrected?)
    - estimate DNA deformation cost (X3DNA)
  - rank DNA sequences (p-value)

DNASITE example: *E.coli* SoxS

```
model 1bl0_A 116 DNACOMPLEX 41 9e-25
_query    SKWYLQRMFRTVTHQTLGDYIRQRRLLLAAVELRTTERPIFDIAMDLGYVSQQTFSRVFR
_template SKWHLQRMFKKETGHSLGQYIRSRKMTEIAQKLKESNEPILYLAERYGFESQQTLTRTFK
_contacts ..*..**.....................................**..**...
_
_stats: 7/7 aligned contacting residues, 6/7 conserved  <- interface identity
_predicted contacting residues in this model:
_contact GLN A  92  (0) 6 T
_contact ARG A  96  (0) 39 G
_contact SER A  95  (1) 7 T
_contact ARG A  96  (0) 9 G
_contact GLN A  45  (0) 17 T
_contact ARG A 100  (1) 38 T
_contact GLN A  92  (0) 42 A
_contact ARG A  46  (0) 30 C
_contact ARG A  46  (0) 19 G
_contact GLN A  45  (0) 16 G
_contact TRP A  42  (0) 31 C
_contact ARG A  46  (0) 29 G
_contact GLN A  91  (0) 5 T
_oligo length = 1 (9), possible mutations = 4
_template reference: S.RHEE et al. PROC.NAT.ACAD.SCI.USA V. 95 10413 1998
_
_predicted binding sites and their scores (MAXPVALUE=0.1):
= NNNNNTTTNGCCNNNNGTGGCNNN +2.60 0.67 2.50e-01
= NNNNNTTTNGCANNNNGTGGCNNN +1.12 0.00 5.00e-01 # original complex DNA sequence
__.....|||.||+....|||||... residues c84,c85,c88,c89,c89,m93,c38,c38,c35,c39,c39, DNAID 9/11
```

SoxS consensus of two models (1)

```
> SoxS number of comparative complexes = 2

model 1bl0_A 116 DNACOMPLEX 41 9e-25
_query    SKWYLQRMFRTVTHQTLGDYIRQRRLLLAAVELRTTERPIFDIAMDLGYVSQQTFSRVFR
_template SKWHLQRMFKKETGHSLGQYIRSRKMTEIAQKLKESNEPILYLAERYGFESQQTLTRTFK
_contacts ..*..**.....................................**..**...

model 1d5y_A 288 DNACOMPLEX 55 2e-27
_query    SKWYLQRMFRTVTHQTLGDYIRQRRLLLAAVELRTTERP
_template SKWHLQRMFKDVTGHAIGAYIRARRLSKSAVALRLTARP
_contacts *.**.**.............................
```

SoxS consensus of two models (2)

```
> SoxS number of comparative complexes = 2

= NNTTTNGCCNNNNGTGGCNNN +2.60 0.67 2.50e-01
= NNTTTNGCANNNNGTGGCNNN +1.12 0.00 5.00e-01 # original complex DNA sequence
__..|||.||+....|||||... residues c84,c85,c88,c89,c89,m93,c38,c38,c35,c39,c39, DNAID 9/11
_
= NNNNNNNNNNNNNGTGCTGNN +0.00 0.00 5.00e-01 # original complex DNA sequence
__............|||||+.. residues c38,c38,c39,c39,c33,m36, DNAID 5/6



consensus superposition of 2 best comparative footprints
_PDB consensus superposition file SoxS_consensus.pdb
= NNNNNNNNNNNNNGTGGCNNN
= NNNNNNNNNNNNNGTGCTGNN
```

Benchmark with *E.coli* regulators in RegulonDB

Data set

85 DNASITE complexes with reported sites (9 SCOP folds)

DNASITE parameter sets

- **default:** MG matrix, $3 contacts/res$,deform $1.6 kcal/mol$
- **CM:** matrix built by the author based only on distance cut-offs
- **sc3:** uses SCWRL3.0 instead of version 2.7
- **Df1, Df2, Df3:** deform $1, 2, 3 kcal/mol$
- **C1:** $1 contact/res$
- **M:** conservative, models only mutated side chains
- **F:** uses family-specific correction
- **P:** P-value cut-off for threaded sequences, original DNA kept

Comparing DNASITE footprints to known binding sites

- _patser DNASITE matrix for SoxS

```
A |   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
C |   0   0   0   0   0   2   1   0   0   0   0   0   0   0   0   2
G |   0   0   0   0   2   0   0   0   0   0   0   2   0   2   2   0
T |   2   2   2   0   0   0   0   0   0   0   0   0   2   0   0   0
```

Comparing DNASITE footprints to known binding sites

- _patser DNASITE matrix for SoxS

A	|	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
C	|	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	2
G	|	0	0	0	0	2	0	0	0	0	0	0	2	0	2	2	0
T	|	2	2	2	0	0	0	0	0	0	0	0	0	2	0	0	0

- PATSER search

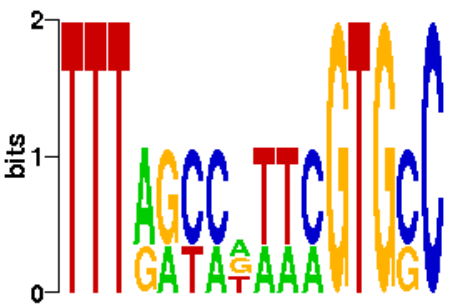
Comparing DNASITE footprints to known binding sites

- ```
  _patser DNASITE matrix for SoxS
  A |   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
  C |   0   0   0   0   0   2   1   0   0   0   0   0   0   0   0   2
  G |   0   0   0   0   2   0   0   0   0   0   0   2   0   2   2   0
  T |   2   2   2   0   0   0   0   0   0   0   0   0   2   0   0   0
  ```
- PATSER search
- ```
  activator -72.5 tgcgcttcttGTTTGGTTTTTCGTGCCAtatgttcgtg
  activator -61.5 tccactttcaTGTAGCACAGTGTGCAGTcctgctcgtt
  activator -56.5 gtttaacctgTTGCATTAATTGCTAAAAgctataactg
  activator -60.5 tcatcgggctATTTAACCGTTAGTGCCTcctttctctc
  activator -40   cgcggcaaaaGCAGAAACTGTAAAACGCagcagtagca
  ...
  ```

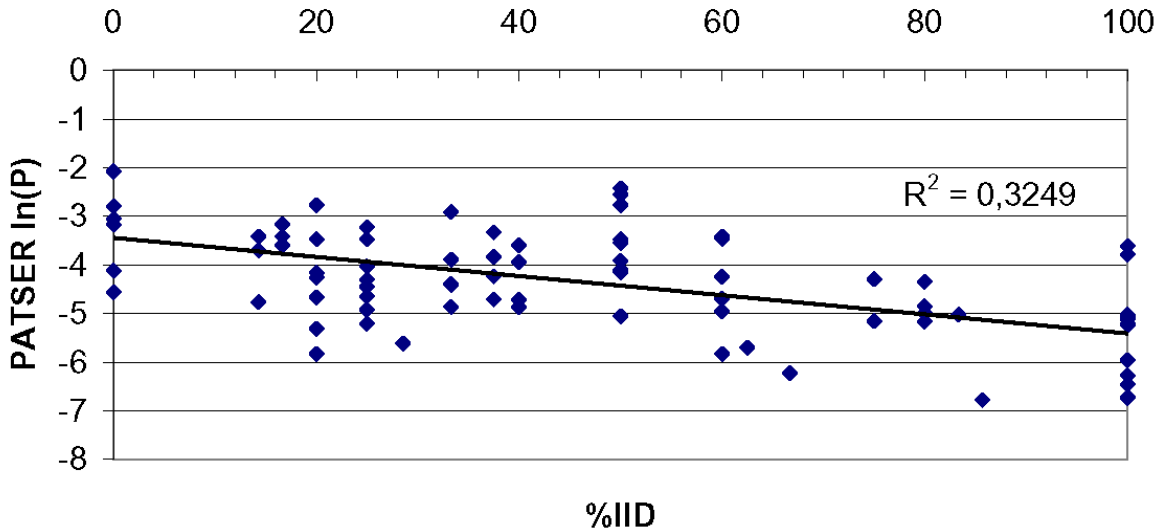
Comparing DNASITE footprints to known binding sites

- ```
  _patser DNASITE matrix for SoxS
  A |   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
  C |   0   0   0   0   0   2   1   0   0   0   0   0   0   0   0   2
  G |   0   0   0   0   2   0   0   0   0   0   0   2   0   2   2   0
  T |   2   2   2   0   0   0   0   0   0   0   0   0   2   0   0   0
  ```
- PATSER search
- ```
  activator -72.5 tgcgcttcttGTTTGGTTTTTCGTGCCAtatgttcgtg
  activator -61.5 tccactttcaTGTAGCACAGTGTGCAGTcctgctcgtt
  activator -56.5 gtttaacctgTTGCATTAATTGCTAAAAgctataactg
  activator -60.5 tcatcgggctATTTAACCGTTAGTGCCTcctttctctc
  activator -40   cgcggcaaaaGCAGAAACTGTAAAACGCagcagtagca
  ...
  ```
- how many sites are recovered?

Comparing DNASITE footprints to known binding sites

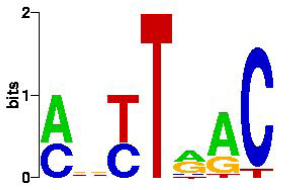
- ```
  _patser DNASITE matrix for SoxS
  A |   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
  C |   0   0   0   0   0   2   1   0   0   0   0   0   0   0   0   2
  G |   0   0   0   0   2   0   0   0   0   0   0   2   0   2   2   0
  T |   2   2   2   0   0   0   0   0   0   0   0   0   2   0   0   0
  ```
- PATSER search
- ```
  activator -72.5 tgcgcttcttGTTTGGTTTTTCGTGCCAtatgttcgtg
  activator -61.5 tccactttcaTGTAGCACAGTGTGCAGTcctgctcgtt
  activator -56.5 gtttaacctgTTGCATTAATTGCTAAAAgctataactg
  activator -60.5 tcatcgggctATTTAACCGTTAGTGCCTcctttctctc
  activator -40   cgcggcaaaaGCAGAAACTGTAAAACGCagcagtagca
  ...
  ```
- how many sites are recovered?
- 

Benchmark results



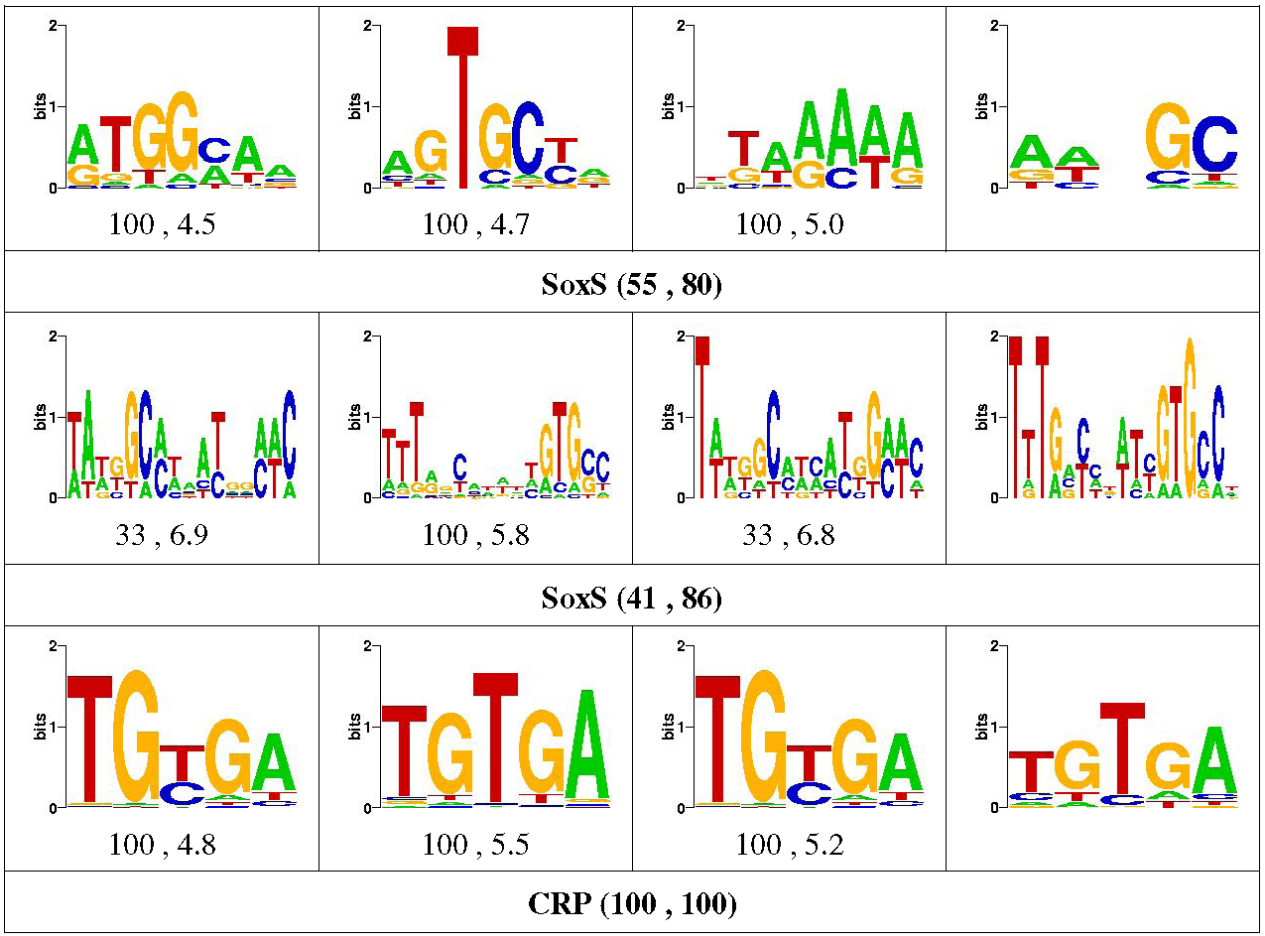
params	def	CM	sc3	Df1	Df2	c1	M	F	$P10^{-2}$	$P10^{-3}$	MF	$FP10^{-4}$
%sites	94	90	94	95	94	98	97	93	93	94	96	97
$-\bar{l}nP$	4.7	4.5	4.6	4.7	4.6	4.3	4.6	4.8	4.5	4.4	4.6	4.4
$\bar{signif}$	1.5	1.3	1.7	1.9	1.5	2.1	2.4	1.8	1.6	2.0	2.5	2.9

Benchmark logos (1)

 100 , 3.1	 100 , 4.3	 100 , 3.4	
AraC (25 , 20)			
 100 , 4.5	 100 , 4.2	 100 , 4.7	
FruR (20 , 38)			
 100 , 3.4	 100 , 3.5	 100 , 3.4	
Fur (20 , 60)			

$P10^{-4}$ x , y	MF x , y	$FP10^{-4}$ x , y	wconsensus x=%sites,y=score
(%ID,%IID)			

Benchmark logos (2)



$P10^{-4}$ x , y	MF x , y	$FP10^{-4}$ x , y	wconsensus x=%sites,y=score
(%ID,%IID)			

Summary

- Protein-DNA complexes are conserved in evolution; this allows us

Summary

- Protein-DNA complexes are conserved in evolution; this allows us
- to build comparative models of DNA-binding proteins that drive

Summary

- Protein-DNA complexes are conserved in evolution; this allows us
- to build comparative models of DNA-binding proteins that drive
- the prediction of their recognised DNA sequences

Summary

- Protein-DNA complexes are conserved in evolution; this allows us
- to build comparative models of DNA-binding proteins that drive
- the prediction of their recognised DNA sequences

However,

Summary

- Protein-DNA complexes are conserved in evolution; this allows us
- to build comparative models of DNA-binding proteins that drive
- the prediction of their recognised DNA sequences

However,

- DNASITE has many parameters that need tuning.

Summary

- Protein-DNA complexes are conserved in evolution; this allows us
- to build comparative models of DNA-binding proteins that drive
- the prediction of their recognised DNA sequences

However,

- DNASITE has many parameters that need tuning.
- Our prediction ability is limited, as the performace improves when the conserved part of templates is inherited.

URL and acknowledgements

I would like to thank:
Julio Collado-Vides
Marc Parisien
Xiangjun Lu
Cei Abreu-Goodger
Pierre-Alain Branger
Martín Peralta
Heladia Salgado
and
UNAM

http://www.ccg.unam.mx/dnasite