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
1 Introduction

2 DNASITE algorithm
   ■ Exploring existing complexes
   ■ DNASITE flowchart

3 Example

4 Benchmark

5 Summary

6 Acknowledgements
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:
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
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:
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)
Introduction

Protein-DNA recognition matrices

\[
\ln \left( \frac{f_{ij}}{f_i f_j} \right)
\]

Mandel-Gutfreund and Margalit (1998) NAR, 26: 2306-2312

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLY</td>
<td>-3.93</td>
<td>-3.93</td>
<td>-3.93</td>
<td>-3.93</td>
</tr>
<tr>
<td>ALA</td>
<td>-3.93</td>
<td>-3.93</td>
<td>0.66</td>
<td>-3.72</td>
</tr>
<tr>
<td>VAL</td>
<td>-3.93</td>
<td>-3.93</td>
<td>-0.17</td>
<td>-3.57</td>
</tr>
<tr>
<td>ILE</td>
<td>-3.93</td>
<td>-3.93</td>
<td>0.65</td>
<td>-3.44</td>
</tr>
<tr>
<td>LEU</td>
<td>-3.93</td>
<td>-3.93</td>
<td>-0.94</td>
<td>-3.93</td>
</tr>
<tr>
<td>PHE</td>
<td>-3.93</td>
<td>-3.93</td>
<td>-0.81</td>
<td>-0.12</td>
</tr>
<tr>
<td>TRP</td>
<td>-1.96</td>
<td>-3.93</td>
<td>-1.96</td>
<td>-3.93</td>
</tr>
<tr>
<td>TYR</td>
<td>-2.87</td>
<td>-2.87</td>
<td>0.54</td>
<td>0.13</td>
</tr>
<tr>
<td>MET</td>
<td>-2.58</td>
<td>-0.28</td>
<td>0.42</td>
<td>-0.28</td>
</tr>
<tr>
<td>CYS</td>
<td>-2.23</td>
<td>0.07</td>
<td>-2.23</td>
<td>0.07</td>
</tr>
<tr>
<td>THR</td>
<td>-3.46</td>
<td>-0.06</td>
<td>-0.06</td>
<td>-1.16</td>
</tr>
<tr>
<td>SER</td>
<td>0.42</td>
<td>-0.68</td>
<td>-0.28</td>
<td>-0.68</td>
</tr>
<tr>
<td>GLN</td>
<td>-0.09</td>
<td>1.16</td>
<td>0.31</td>
<td>-3.09</td>
</tr>
<tr>
<td>ASN</td>
<td>0.48</td>
<td>1.93</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>GLU</td>
<td>-3.93</td>
<td>-1.24</td>
<td>-3.93</td>
<td>0.55</td>
</tr>
<tr>
<td>ASP</td>
<td>-3.93</td>
<td>-3.37</td>
<td>-3.93</td>
<td>1.01</td>
</tr>
<tr>
<td>HIS</td>
<td>1.56</td>
<td>0.46</td>
<td>0.87</td>
<td>-0.23</td>
</tr>
<tr>
<td>ARG</td>
<td>2.74</td>
<td>0.34</td>
<td>1.25</td>
<td>-3.93</td>
</tr>
<tr>
<td>LYS</td>
<td>2.16</td>
<td>-0.08</td>
<td>0.21</td>
<td>-3.93</td>
</tr>
<tr>
<td>PRO</td>
<td>-3.93</td>
<td>-3.93</td>
<td>-0.30</td>
<td>-3.29</td>
</tr>
</tbody>
</table>
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 (< 12Å):
  - (+) CA
  - (-) N1/N9
Interface comparison

- Interface atoms (< 12Å):
  - (+) CA
  - (-) N1/N9

- RMSD calculated over MAMMOTH superimpositions
Homologous protein-DNA interfaces are conserved

Median values for 442 pairs of superimposed PDB complexes.
DNASITE: Comparative footprinting of DNA-binding proteins

- DNASITE algorithm
- Exploring existing complexes

SCOP folds show different interface conservation
Contact side chains can be modelled

987 base H-bonding residues modelled by SCWRL with templates $\geq 30\%$ID
**Can we model protein-DNA complexes?**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>do DNA-binding proteins conserve their docking geometry?</td>
<td>YES, as a function of % sequence identity</td>
</tr>
<tr>
<td>can we identify modelled protein residues that contact DNA?</td>
<td>YES, at least we can model most H-bonding residues</td>
</tr>
</tbody>
</table>
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

- 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

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

\[
s_i + N_{template} + P_{model} = PN_i
\]
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?)
How DNASITE builds comparative models

\[
deform(s_i, N_{\text{template}}) = f(s_i, Olson, \text{geom}(N_{\text{template}}))
\]

- 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)
Example:

**DNASITE example: E.coli SoxS**

model 1bl0_A 116 DNACOMPLEX 41 9e-25
_query SKWYLQRMFRTVTHQTLDGYIRQRLLLAVELRTTERPIFDIAMDLGYVSQQTFSRVF
_template SKWHLQRMFKETGHSLGQYIRSRKMTEIAQKLKESNEPILYLAERYGFESQQLTTLRTFK
_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):
= NNNNNTTNGCCNNNNGTGCNNN +2.60 0.67 2.50e-01
= NNNNNTTNGCANNNNGTGCNNN +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  SKWYLQRMFRTVTHQTLGDYIRQRLLLAAVELRTTERPIFDIAMLGYVSQQTFSRVR
  _template SKWHLQRMFKETGHSQYIRSRKMTESIQLKESNEPILYGYEFESQQTTLRTFK
  _contacts ..*..*...............................**..**...

model 1d5y_A 288 DNACOMPLEX 55 2e-27
  _query  SKWYLQRMFRTVTHQTLGDYIRQRRLLAAVELRTTERP
  _template SKWHLQRMFKDVTGAHAYIRARRLSKSAVALRTARP
  _contacts *.**.................................
DNASITE: Comparative footprinting of DNA-binding proteins

Example

SoxS consensus of two models (2)

> SoxS number of comparative complexes = 2

= NNTTTNGCCNNNNGTGGCNNN +2.60 0.67 2.50e-01
= NNTTTNGANCNNNNGTGGCNNN +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

= NNNNNNNNNNNNNNGTGCTGNN +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
= NNNNNNNNNNNNGTGCTGNN
= NNNNNNNNNNNNGTGCTGNN
DNASITE: Comparative footprinting of DNA-binding proteins

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 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  C | 0 0 0 0 0 0 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 2
  G | 0 0 0 0 2 0 0 0 0 0 0 0 0 2 0 2 2 0
  T | 2 2 2 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0
Comparing DNASITE footprints to known binding sites

- _patser_ DNASITE matrix for SoxS
  A | 0 0 0 0 0 0 0 1 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 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 2 0 0 0 0

- PATSER search
  activator -72.5 tgcgctttcttGTTTGGTTTTTCGTGCCAtatgtttcggtg
  activator -61.5 tccactttcaTGTAGCACAGTGTGCAGTcctgctcgtt
  activator -56.5 gttaacctgTTGCATTAATTGCTAAAAGctataactg
  activator -60.5 tcatcgggctATTTAACCGTTAGTGCTcctttctctc
  activator -40  cgcggcaaaaaGCAGAAACTGTAACGCagcagtagca
  ...

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

PATSER search

- activator -72.5 tgcgctttctgttttGTTTTTTCGTCGAtatatgttctgg
- activator -61.5 tccactttcaTTGAGCAGTGTGCTGCACTcttctgtttc
- activator -56.5 gtttaacctgTTGAGCAGTGTGCTGCACTcttctttg
- activator -60.5 tcatcgggctATTAAACGTGTAGCTCCTctttttctttc
- 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 0 0
C | 0 0 0 0 0 0 2 1 0 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 0 0
T | 2 2 2 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0

**PATSER search**

- activator -72.5 tgcgctttcttGTTCACCCGCAATatgtttcggtg
- activator -61.5 tccactttcaTGTAGCACAGTGTGCAGTcctgctcggtt
- activator -56.5 gtttaacctgTTGCATTAATTGCTAAAAAgctataactg
- activator -60.5 tcacgctgtATTTAACCGTTAGTGCTcctttctctc
- activator -40  cgcggcaaaaGCAGAAACTGTAAAAGGCagcagtagca
...

**how many sites are recovered?**
Benchmark results

<table>
<thead>
<tr>
<th>params</th>
<th>def</th>
<th>CM</th>
<th>sc3</th>
<th>Df1</th>
<th>Df2</th>
<th>c1</th>
<th>M</th>
<th>F</th>
<th>P10^{-2}</th>
<th>P10^{-3}</th>
<th>MF</th>
<th>FP10^{-4}</th>
</tr>
</thead>
<tbody>
<tr>
<td>%sites</td>
<td>94</td>
<td>90</td>
<td>94</td>
<td>95</td>
<td>94</td>
<td>98</td>
<td>97</td>
<td>93</td>
<td>93</td>
<td>94</td>
<td>96</td>
<td>97</td>
</tr>
<tr>
<td>−lnP</td>
<td>4.7</td>
<td>4.5</td>
<td>4.6</td>
<td>4.7</td>
<td>4.6</td>
<td>4.3</td>
<td>4.6</td>
<td>4.8</td>
<td>4.5</td>
<td>4.4</td>
<td>4.6</td>
<td>4.4</td>
</tr>
<tr>
<td>signif</td>
<td>1.5</td>
<td>1.3</td>
<td>1.7</td>
<td>1.9</td>
<td>1.5</td>
<td>2.1</td>
<td>2.4</td>
<td>1.8</td>
<td>1.6</td>
<td>2.0</td>
<td>2.5</td>
<td>2.9</td>
</tr>
</tbody>
</table>

R^2 = 0.3249
Benchmark logos (1)

AraC (25, 20)

FruR (20, 38)

Fur (20, 60)

<table>
<thead>
<tr>
<th>P10⁻⁴</th>
<th>MF</th>
<th>FP10⁻⁴</th>
<th>wconsensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>x, y</td>
<td>x, y</td>
<td>x, y</td>
<td>x=%sites,y=score</td>
</tr>
<tr>
<td>(%ID,%IID)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Benchmark logos (2)

DNASITE: Comparative footprinting of DNA-binding proteins

Benchmark logos (2)

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>P10−4 x, y</th>
<th>MF x, y</th>
<th>FP10−4 x, y</th>
<th>wconsensus x=%sites,y=score (%ID,%IID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoxS (55, 80)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100, 4.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100, 4.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100, 5.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SoxS (41, 86)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>33, 6.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100, 5.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>33, 6.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRP (100, 100)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100, 4.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100, 5.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100, 5.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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...
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.
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 performance improves when the conserved part of templates is inherited.
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