3D-footprint: a database for the structural analysis of protein-DNA complexes

Bruno Contreras-Moreira
bcontreras@eead.csic.es
http://floresta.eead.csic.es/3dfootprint

Estación Experimental de Aula Dei, CSIC, Zaragoza, España

October 27, 2010
1 Background: experimental and computational methodologies
   ■ 1st: an algorithm that enumerates contacts
   ■ 2nd: a readout algorithm

2 The 3D-footprint database: design and analysis
   ■ Analysis of the content of the database

3 Using the database
DNA footprinting, a molecular/genomic protocol

Experiments often find site variants across promoters:

ACGTTTAG

TCGTTAAG

GCGATAAG

Position weight matrices (PWM) condense aligned sites:

\[
\begin{align*}
\text{A} & | 1 & 0 & 0 & 1 & 0 & 2 & 3 & 0 \\
\text{C} & | 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{G} & | 1 & 0 & 3 & 0 & 0 & 0 & 0 & 3 \\
\text{T} & | 1 & 0 & 0 & 2 & 3 & 1 & 0 & 0 \\
\end{align*}
\]
DNA footprinting, a molecular/genomic protocol

Experiments often find site variants across promoters:

- ACGTTTAG
- TCGTTAAG
- GCGATAAG

Position weight matrices (PWM) condense aligned sites:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

- geneA ACGTTTAG
- geneB TCGTTAAG
- geneCD GCGATAAG

(sequence logo)
The Protein Data Bank as a repository of footprint experiments

Example: NarL dimer (1JE8) from *Escherichia coli*

The PDB currently contains over 3600 protein-DNA complexes
Dissecting a protein-DNA interface (3D)

(click to see video)

credit: A. Sebastián
Dissecting a protein-DNA interface (2D)

A189VAL.CG1=C7(4.78)

A182LYS.NZ=O6(4.21)  A182LYS.NZ=N7(3.96)  A192LYS.NZ=O6(4.21)  A188LYS.CE=C7(2.14)
3D-footprint: a database for the structural analysis of protein-DNA complexes

Background: experimental and computational methodologies

Calculating a PWM from an interface

How can we go from this?
Calculating a PWM from an interface

How can we go from this?

<table>
<thead>
<tr>
<th>A</th>
<th>13</th>
<th>56</th>
<th>9</th>
<th>14</th>
<th>23</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>14</td>
<td>14</td>
<td>69</td>
<td>52</td>
<td>31</td>
<td>20</td>
</tr>
<tr>
<td>G</td>
<td>13</td>
<td>13</td>
<td>9</td>
<td>14</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>T</td>
<td>56</td>
<td>13</td>
<td>9</td>
<td>16</td>
<td>21</td>
<td>19</td>
</tr>
</tbody>
</table>
Calculating a PWM from an interface

How can we go from this?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>13</th>
<th>56</th>
<th>9</th>
<th>14</th>
<th>23</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>14</td>
<td>14</td>
<td>69</td>
<td>52</td>
<td>31</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>13</td>
<td>13</td>
<td>9</td>
<td>14</td>
<td>21</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>56</td>
<td>13</td>
<td>9</td>
<td>16</td>
<td>21</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

to this?
Morozov’s algorithm enumerates atomic contacts

Morozov & Siggia (2007)
DNAPROT algorithm estimates direct + indirect readout

Angarica et al. (2008)
The 3D-footprint database: design and analysis
Flowchart: combining readout and contact PWMs

1) analyze interface

2.1) get contact PWM

Cumulative contact algorithm

19 26 12 12 25

2.2) get readout PWM

DNAPROT algorithm

{T, A, G, C}

2 2

2.3) compare to others

4) find related motifs

PubMed

3) mean PWM

GCC CG GC CG C T

G CG GC CG C GC
Flowchart: combining readout and contact PWMs

3D-footprint: a database for the structural analysis of protein-DNA complexes

- The 3D-footprint database: design and analysis

<table>
<thead>
<tr>
<th>total complexes</th>
<th>nr</th>
<th>multimers</th>
<th>redundant</th>
</tr>
</thead>
<tbody>
<tr>
<td>2642</td>
<td>566 (196 with PubMed sites)</td>
<td>814</td>
<td>1262</td>
</tr>
</tbody>
</table>
Typical attributes of a 3D-footprint entry

A

B

C

# IC=14.65  IC/col=1.22  n_of_columns=12  specificity:

A | 24 0 0 96 0 0 0 0
C | 24 96 96 0 96 0 0
G | 24 0 0 0 0 96 0
T | 24 0 0 0 0 0 0

scan!
Interface residues are also annotated

- `>1nlw_D interface=5,6,8,9,12,13`
  `SRSTHNMEEKNRRAHLRLSLEKLGVLGPLGDSSR...`
  ** ** **
Interface residues are also annotated

- >1nlw_D interface=5,6,8,9,12,13
  SRSTHNEMEKRNRAHLRLSLEKLKGLVPLGPDSSR...
  ** ** **

- protein-DNA complexes are clustered based on structural similarity, highlighting interface side-chains

  RG--HGKT--EARCRC-- L +-----1a0a_B
  +--------2
  ----HGNTEARCGR-- L ! !  +1nlw_D
  ! +-----1
  ----HGNTECRCRG-- L --4 +1nkp_E
  !
  --QG----VTECRGQG-- L ! +--------1an4_A
  +----3
  ----NT--ETRCMT-- +--------2ql2_D
Interface residues are also annotated

- >1nlw_D interface=5,6,8,9,12,13
  SRSTHNMEMKNRRALRLSLLEKLLKVLGLVPLGPDSSR...
  ** ** **

- protein-DNA complexes are clustered based on structural similarity, highlighting interface side-chains
  RG--HGKT--EARCRC-- L +-----1a0a_B
    +--------2
  ----HGNMTEARCRG-- L ! ! ! +1nlw_D
    ! +---1
  ----HGNT--ECRCRG-- L --4 +1nkp_E
    !
  ---QG----VTECRCRGQG L ! +--------1an4_A
    +---3
  --------NT--ETRCMT-- +--------2ql2_D

- check poster #48 (C.Cantalapiedra) for recent results in interface alignment
Direct readout: hydrogen bonds dominate DNA recognition

(Contreras-Moreira et al., 2009)
Indirect readout is most important among restriction enzymes
PWM information content correlates with interface atomic interactions

(Contreras-Moreira, 2010)
Specificity varies across superfamilies...

DNA binding specificity of SCOP superfamilies

[Graph showing the specificity of SCOP superfamilies]
... and some bind DNA unspecifically

Unspecific DNA-binding SCOP superfamilies

- Lesion_bypass_DNA_polymerase_Y
- DNA-glycosylase
- Ribonuclease_H-like
- Histone-fold
- P-loop_containing_nucleoside_t
- DNA/RNA_polymerases

Observations
3D-footprints match with PubMed-mined motifs

...base substitutions that significantly reduced the NarL-mediated stimulation were restricted to a 6-base sequence, TACTCC, located at positions -193 to -198 in the narGHJI promoter. When a bases...
Using the database
Two main types of queries are supported

- query 1: input = DNA motif
  What kind of proteins might bind to such a DNA site?

DNA to protein

- Protein to DNA
Two main types of queries are supported

- query 1: input = DNA motif
  What kind of proteins might bind to such a DNA site?

- query 2: input = amino acid sequence
  What kind of motifs are recognized by proteins alike?
Query 1: DNA motif

Parameters:  -e 0.01
Input motif:
A | 0 0 100 100 0 0 0 3
C | 100 0 0 0 0 100 100 90
G | 0 0 0 0 0 0 0 4
T | 0 100 0 0 100 0 0 3

Results (sorted by STAMP expectation value):

<table>
<thead>
<tr>
<th>logo</th>
<th>E-value</th>
<th>organism</th>
<th>complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATTA</td>
<td>6.6613e-15</td>
<td>Homo sapiens</td>
<td>1yz8_P: SOLUTION STRUCTURE OF THE K50 CLASS</td>
</tr>
<tr>
<td>CTA TCC</td>
<td>2.6356e-10</td>
<td>Drosophila melanogaster</td>
<td>1zq3_P: NMR SOLUTION STRUCTURE OF THE BICOID</td>
</tr>
<tr>
<td>GATA</td>
<td>4.1593e-09</td>
<td>Drosophila melanogaster</td>
<td>2hos_A: PHAGE-SELECTED HOMEODOMAIN BOUND TO</td>
</tr>
</tbody>
</table>
Query 2: protein sequence

- *putative transcription factor*
  KSTAEARQSQLKIHSAAEKRERRERINAHLATLRR...
Query 2: protein sequence

- putative transcription factor
  KSTAEARQSALKIHSAAEKRERRERINAHLATLRR...

>lam9_A: HLH, helix-loop-helix DNA-binding domain; title=HUMAN
  SREBP-1A BOUND TO LDL RECEPTOR PROMOTER organism=HOMO SAPIENS
  interface=10,11,13,14,17,18, Length = 80

  Expect = 3e-04, Identities = 18/71 (25%)
  Interface identity = 3/6 (50%), Interface coverage = 6/6 (100%)

Query: 57  KSTAEARQALKIHSAAEKRERRERINAHLATLRRMIPDA-SQMDKATLLARV 106
  +S E R A H+A EKR R IN + L+ ++ ++++K+ +L +
  Sbjct: 1  QSRGEKRTA-- HNAIEKRYRSSINDKIELKDLVVGTEAKLNKSAVLRLKA 48
Query 2: protein sequence

- putative transcription factor
  KSTAERQSAHKIHSAEKRMRERINAHLATLRR...

- as shown in poster #47 (A. Sebastián), proteins with similar interfaces bind similar DNA motifs
3D-footprint: a database for the structural analysis of protein-DNA complexes

Using the database

Thanks for your attention

3D-footprint is a database of DNA-binding protein structures that is updated weekly with Protein Data Bank complexes, providing:

- structure-based binding specificities and sequence logos
- classification and clusters of protein-DNA interfaces
- downloads / stats
- complex of the week
- tutorial

Tools:
- footprint your structure
- consensus miner

help & credits

find DNA-binding proteins by name

search term: Submit demo
keyword,PD code,superfamily

find proteins that recognize a similar DNA motif

site to be searched:

match type: local E-value cutoff: 0.01
Submit demo sequence reset

BLAST similar DNA-binding proteins

protein sequence:

Submit demo reset

Created and maintained at the Laboratory of Computational Biology / Eead-CSIC

updated: Sun Mar 7 19:49:25 2010

http://floresta.eead.csic.es/3dfootprint
3D-footprint: a database for the structural analysis of protein-DNA complexes

Using the database


