Binding Site with `transfac`
Can we do better?

Peter F. STADLER

Lehrstuhl f. Bioinformatik; Institute f. Informatik,
University of Leipzig, Gemany
University of Vienna & Santa Fe Institute (external faculty)

`http://www.bioinf.uni-leipzig.de/~studla`

Bled, Feb 16 2003

Transfac matrices

```
AC  <A HREF=/dbget-bin/get_linkdb?tfmatrix+M00201>M00201</A>
XX
ID  V\$CEBP_C
XX
DT  05.09.1995 (created); dbo.
DT  30.11.1995 (updated); ewi.
XX
NA  C/EBP
XX
DE  C/EBP binding site
XX
XX
PO       A      C      G      T
01    2.57   1.73   3.05   1.87      N
02    0.90   0.76   6.74   0.82      G
03    3.20   0.79   0.90   4.33      W
04    3.96   2.75   2.51   0.00      N
05    0.00   0.00   0.00   9.22      T
06    0.00   0.00   4.03   5.19      K
07    3.29   1.67   0.85   3.41      N
08    0.00   2.20   3.53   3.49      K
09    0.00   0.91   8.31   0.00      G
10    0.90   4.78   0.00   3.54      Y
11    9.22   0.00   0.00   0.00      A
12    9.22   0.00   0.00   0.00      A
13    0.00   0.85   4.03   4.33      K
14    2.49   1.75   4.08   0.90      N
15    0.00   1.73   0.90   0.82      S
16    6.63   1.69   0.90   0.00      A
17    0.00   5.80   0.00   3.42      Y
18    4.72   0.90   1.82   1.78      A
XX
BA  total weight of sequences: 9.22
XX
CC  consind generated matrix (random_expectation:    0.27)
XX
```

Scoring Potential Binding Sites

`transfac` table entries $\pi_\alpha(k)$

The simple method:

$$\Sigma(i) = \sum_{k=1}^{L} \left(\sum_{\alpha=\{G,C,A,T\}} \pi_\alpha(k)\delta[x_{i+k}, \alpha] \right)$$

$$\Sigma_{\max} = \sum_{k=1}^{L} \max_{\alpha} \pi_\alpha(k)$$

$$\langle \Sigma(i) \rangle = \frac{1}{4} \sum_{k=1}^{L} \sum_{\alpha} \pi_\alpha(k)$$

and then normalize

$$\text{score}(i) = \frac{\Sigma(i) - \langle \Sigma(i) \rangle}{\Sigma_{\max} - \langle \Sigma \rangle}$$

Latest addition to the `tracker` program: `transfucker`

```
oerzl> transfucker -R Cdx -I -t 0.6 -f cluster473.al
    33 0.6847 M00100 "DE  CdxA"
   136 0.6552 M00100 "DE  CdxA"
    10 0.6524 M00101 "DE  CdxA"
    33 0.7805 M00101 "DE  CdxA"
HfM,            tgaacTTGATGTCATGagctacc--------TGTTAATGTAATTTTTtcct--TGTGGTT
HsA,            _____TGGATATAATGtattttcagctcaatTGTTAATGTGATGTATattttgTGTTATT
PsA,            ___________________________________________TActga-----TTAAT
DrAa,           TGT-----------------------------------------------------
FrAa,           TGT-----------------------------------------------------
M00100          ..................................MTTTATR.....................
M00101          .........WWTWMTR.................WWTWMTR.....................
~
~
HfM,            TGTGTCTGGAATTTTGCATGATACAA------------TGCTGTCAGTgtGTACCtact
HsA,            CGTGAATAGTCTTTTGCATGTCGCACAATGTT------TgatgtccccaaaGTACCA---
PsA,            CCTGAA--GTAGTTTGCATGGTGCAAAATGCT------TTGCTTTTACTTG-----A---
DrAa,           --------GAGTTTTGCAGCATACGCAATtgtctgaaGTTGTTCTCCGGTG-----AT--
FrAa,           --------GAGATTTGCATGATCCGt-----------GTTGCTTTCACGTG-----TTat
M00100          ..........................................................
M00101          ..........................................................
~
~
HfM,            catt-----ACTGAGTTTTATCAGTTCTACTGTGTGACTAGATGTTAT
HsA,            --------CACTGAGTTCTATCAGTTATCCTTTGTGAGCCTATGATAT
PsA,            --------CACTGAGTTTTATCAGTTAatacctctta___________
DrAa,           ---------GCTGAGTTTTATCAGTCGGaat_________________
FrAa,           tagcggggCGCTGAGTTTTATCAGTTGG-------GAGCCcttc____
M00100          ..............MTTTATR.........................
M00101          ..............................................
```

```
#!/bin/participant
wakeup -now
/etc/rc.d/init.d/brain start
```

WANTED: Your ideas for a better scoring function

That it — we can have beer now ...