# DNA familial binding profiles made easy : Shaun Mahony, Philip Auron, Panayiotis Benos

BBSI 2007 @ Pitt, Group 1, Journal club presentation

Ankur Agarwal

# What is the paper about ?

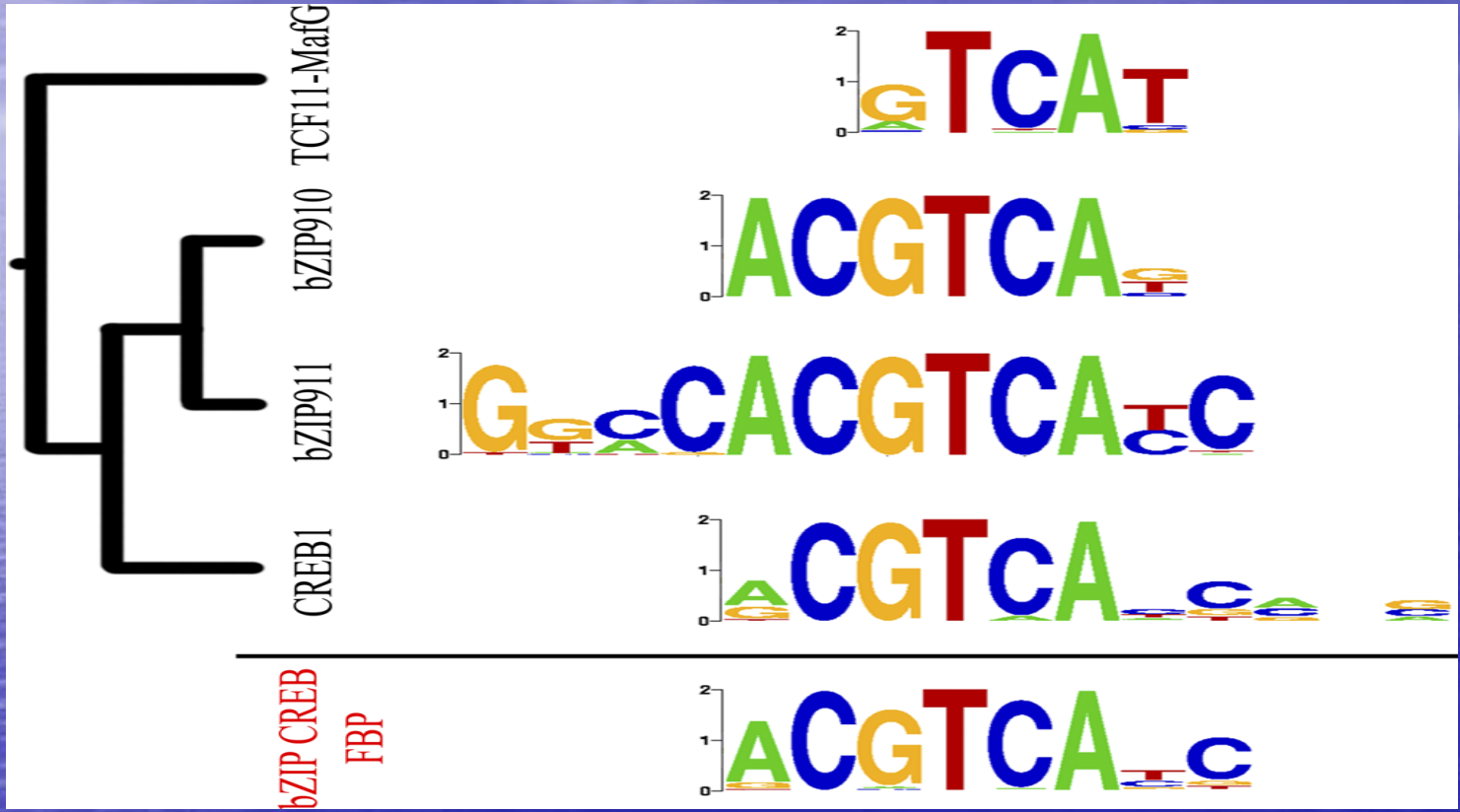
- This paper is about a software platform (STAMP) that classifies (builds clusters) TF proteins based on their DNA binding motifs.
- The clusters represent FBPs (Familial Binding Profiles) for the proteins.
- This platform leads to a more accurate TF classification so much so that TF proteins belonging to same structural class or family can go to different clusters and vice versa.

STAMP stands for Similarity Tree building and Alignment of DNA Motifs and Profiles

# Why is it important ?

- If unknown TF proteins can be (correctly) classified based on the (known) DNA motifs they bind to then it would lead to better identification and estimation of regulatory elements and circuits in the organism's DNA and would also lead to identification of hitherto unknown TF proteins.

# What is FBP ?




In this example, the binding motifs for four bZIP–CREB transcription factors are aligned in a multiple-motif alignment. The generalized familial binding profiles correspond to the weighted average of the individual profiles.

Source : Same paper by Dr Benos et al, Figure 1, Pg 2

# Protein Families

- Comprises of a group of evolutionarily related proteins e.g. the zinc finger family of proteins.
- They may or may not have same structure however they will share some of the protein domains.
- More often than not they will have the same biological function.

# Methods

- TF DNA-binding preferences are modeled using PSSMs aka position specific scoring matrices
- PSSMs are generated from frequency matrices by converting frequencies to scores (by taking logs etc, a formula can be applied)

# Frequency matrices

- For a motif of length m using an alphabet of n characters , a frequency matrix is an n by m matrix in which each element contains the frequency at which a given member of the alphabet is observed at a given position in an aligned set of sequences containing the motif.

Taken from :
http://murphylab.web.cmu.edu/presentations/MurphyBioJClub19991201/sld007.htm

# Frequency matrix example

TGGGGGA
TGAGAGA
TGGGGGA
TGAGAGA
TGAGGGA

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 4 | 1 | 3 | 1 | 6 |
| T | 6 | 1 | 1 | 1 | 1 | 1 | 1 |
| G | 1 | 6 | 3 | 6 | 4 | 6 | 1 |
| C | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

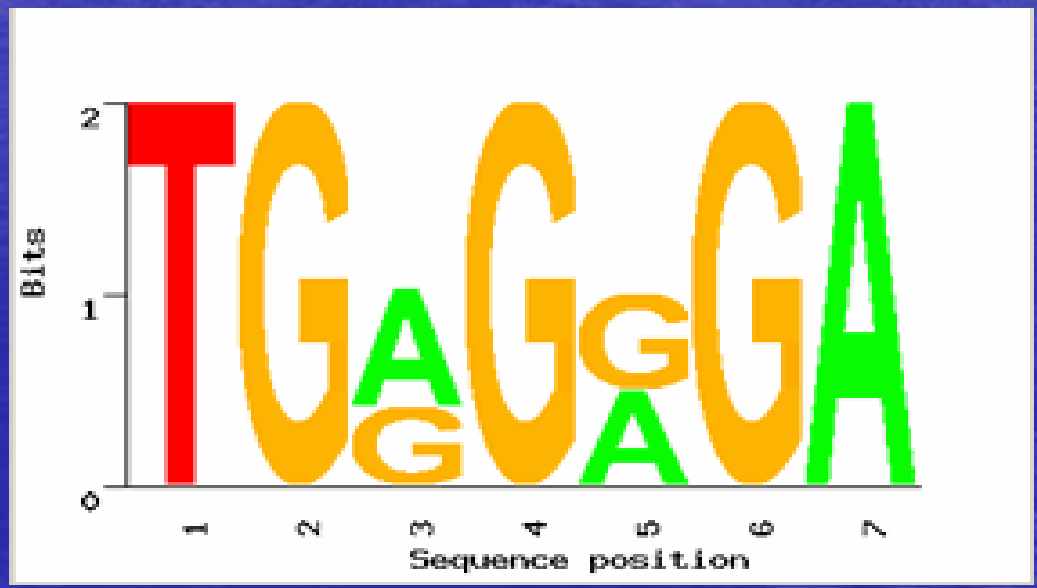
To generate a PSSM from a frequency matrix pseudo counts are added to base frequencies to avoid zero probabilities and avoid other errors.

# PSSM example

- Formula used :

$$I_i = 2 + \Sigma f_{b,i} \log_2 f_{b,i}$$

http://weblogo.berkeley.edu/




The values in PSSM columns reflect preference of the TF for the corresponding base in the position.

# PSSMs

- PSSMs assume independency between base positions which is a simplifying & valid assumption in most of the situations.

# Motivation for STAMP

- FBPs are useful in DNA sequence analysis but there is not a lot of work done on the methods used to align DNA motifs.
- Current FBPs are built semi-empirically (part algorithmic and part experimental) and the approach cannot be extended to a big collection of sequences.
- Motif families are not very well defined in current motif databases.

# Motivation cont..

- In the past many other authors have worked on building better and reliable FBP construction methods and the authors of this paper are expanding a similar study by doing a through and rigorous treatment of current distance measuring metrics, motif alignment, tree building and clustering approaches.

# Evaluating similarity metrics

- Authors are evaluating six distance metrics (PCC, pCS, AKL, SSD, ALLR, ALLR_LL).
- The six metrics are compared wrt their efficiency in capturing similarities in PSSM columns in aligning PSSM motifs.

# Evaluating motif alignment strategies

- To test the efficacy of column scoring metrics and alignment method combinations a “best hit” approach was used.
- In a database search the best match to a given motif is expected to be a motif associated with a member of the same structural class and the results would be considered good if the proportional of motifs that match another member of the same structural class is a good number (close to 1).

# Motif alignment strategies contd ..

- SW local alignments were found to be better and NW global alignments for motif alignments.
- The results of best hit approach using PCC column comparison metric and un gapped SW alignment method compared very well with a Bayesian algorithm (Narlikar and Hartemink) on the same dataset !

# Optimal clusters in a tree ??

- A new measure is developed for automatically determining the optimal number of clusters in a given motif tree : $CH_{log}$ (log modified Calinski and Harabasz)
- This measure is used on DNA motif tree built by a better combination of optimal distance metrics and alignment strategies to generate a new set of FBPs without any prior knowledge of TF structural class or families, a completely algorithmic classification!

# Evaluating Tree building methods

- SOTA and UPGMA were compared as tree building methods and UPGMA was found to be better.
- A DNA motif tree was built (using UPGMA) for a nonzinc-finger JASPER dataset consisting of 71 motifs. $CH_{log}$ gave 17 as the optimal number of clusters .
- When LOOCV (leave one out cross validation) check was performed only two misclassifications were found in addition to two singleton clusters. A classification efficiency of 67/71 = 94 %. This is higher than a different tree building approach attempted in the past with 87 % efficiency !

# Tree building methods contd..

- When zinc finger motifs (DOF and GATA ) were also included 15 clusters were similar across both the trees with classification efficiency (LOOCV test) of 91 % compared to just 76 % from an earlier study.
- Clusters formed to accommodate the new motifs were very reasonable and sound.

# STAMP development

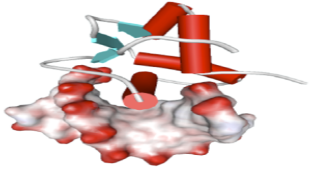
- STAMP incorporates a fully automated method for PSSM clustering based on the combinations of distance metrics, alignment strategies and tree building methods examined so far.
- A different (but very similar to $CH_{log}$) metric was also used to determine the optimal number of clusters.

# A note on advantages of automatic clustering

- By remaining ignorant of the structural class of the motif we can find interesting cases where diverse structural classes and families are more suitably grouped together if they have a similar DNA motif binding affinities.
- Now we can also find differences in DNA binding affinities between the sub families of proteins (which is also very interesting) whereas initially we were tempted to group such DNA motifs together.

# STAMP results

- 17 clusters were obtained for 71 JASPER PSSM models ( non-zinc finger family).
- Divides the dataset into homogeneous clusters with respect to structural group of the corresponding TFs whereas no structural information was fed into in the program initially.
- # 2 agrees with the notion that structurally similar TFs tend to have similar binding specificities – which is a good indication of the validity of the results.



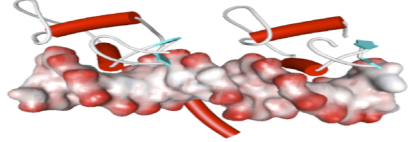
ETS

### ETS

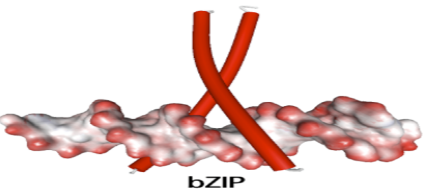
**Members:** *ETS:* c-ETS, SPI1, SPIB, ELK1, ELK4, E74A, GABPA

### Nuclear Receptor

**Members:** *Nuclear Receptor:* PPARG-RXRA, NR2F1, RORA1, RORA, RXR-VDR, PPARG, CFI-USP



Nuclear Receptor

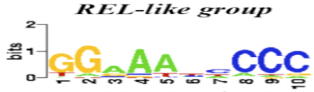


bZIP

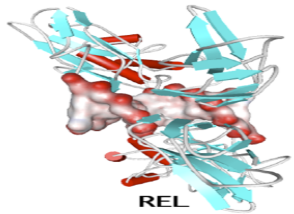
### bZIP CREB Subgroup

**Members:** *bZIP CREB:* TCF11-MafG, CREB1, bZIP911, bZIP910

### REL-like group



**Members:** *REL:* Dorsal_1, Dorsal_2, REL, NF-kB, RELA, NFKB1
*HOMEO:* En1
*bZIP:* Chop-cEBP



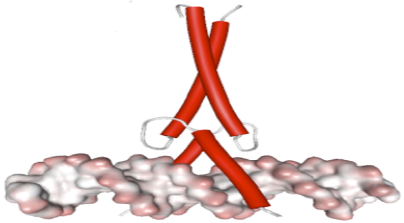
REL

### bHLH-ZIP

**Members:** *bHLH:* Arnt-Ahr, MYC-MAX, MAX, Arnt, USF1, Mycn

### bHLH Subgroup 1

**Members:** *bHLH:* TAL1-TCF3, HAND1-3



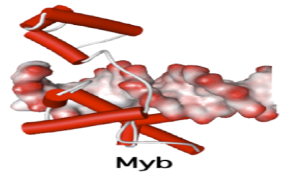
bHLH-ZIP

### bHLH Subgroup 2

**Members:** *bHLH:* Myf, NHLH1

### TRP-cluster, Myb Subgroup

**Members:** *TRP-cluster:* GAMYB, Myb, MYB.ph3



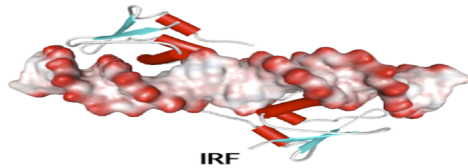
Myb

### bZIP cEBP Subgroup

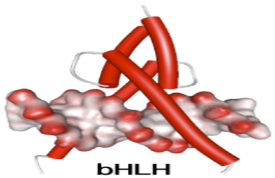
**Members:** *bZIP cEBP:* cEBP, NLIF3 HLF

### TRP-cluster, IRF Subgroup

**Members:** *TRP-cluster:* IRF1, IRF2



IRF



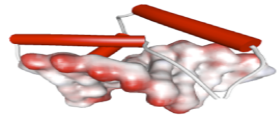
bHLH

### MADS

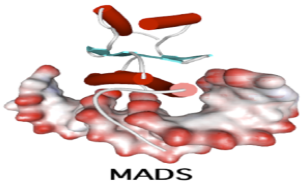
**Members:** *MADS:* Agamous, SRF, MEF2A, AGL3, SQUA

### HMG/Forkhead Group 1

**Members:** *Forkhead:* FOXD1, FOXF2, Foxq1, FOXI1, Foxa2, Foxd3
*HMG:* Sox17, Sox9,SRY, Sox5
*HOMEO:* Pbx



HMGB

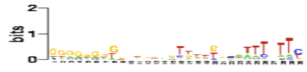


MADS

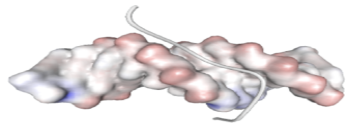
### HMG/Forkhead Group 2

**Members:** *HMG:* HMG-1
*Forkhead:* FOXC1

### HMG/HOMEO Mix.



**Members:** *HMG:* HMG-IY
*HOMEO:* Pax-4

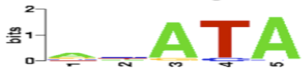


HMGA/HMG-I/Y

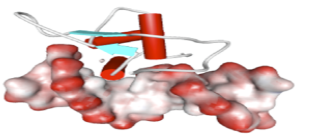
### HOMEO Subgroup

**Members:** *HOMEO:* Ubx, Nkx2-5, Prrx2, Athb-1, TCF1

### FOXL1 Singleton



**Members:** *Forkhead:* FOXL1

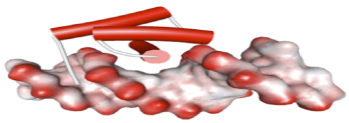


FOX/Forkhead

### Ar Singleton

**Members:** *Nuclear Receptor:* Ar

Pax



HOMEO

# Questions??

Thank you !