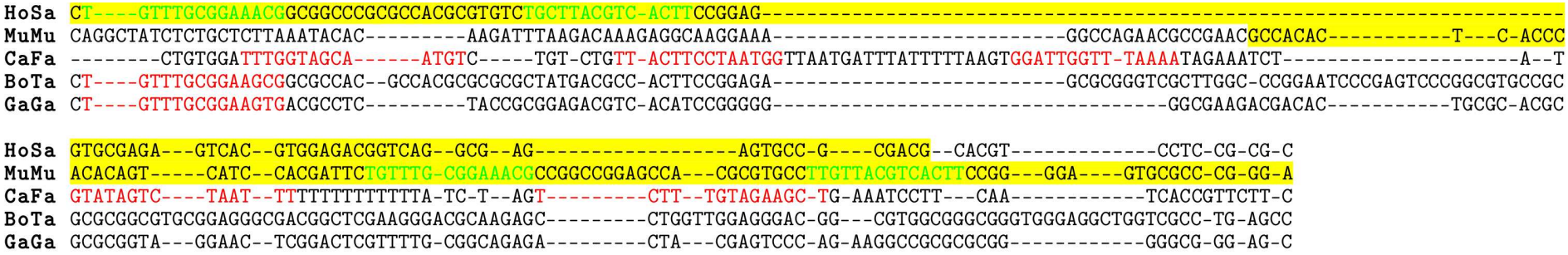
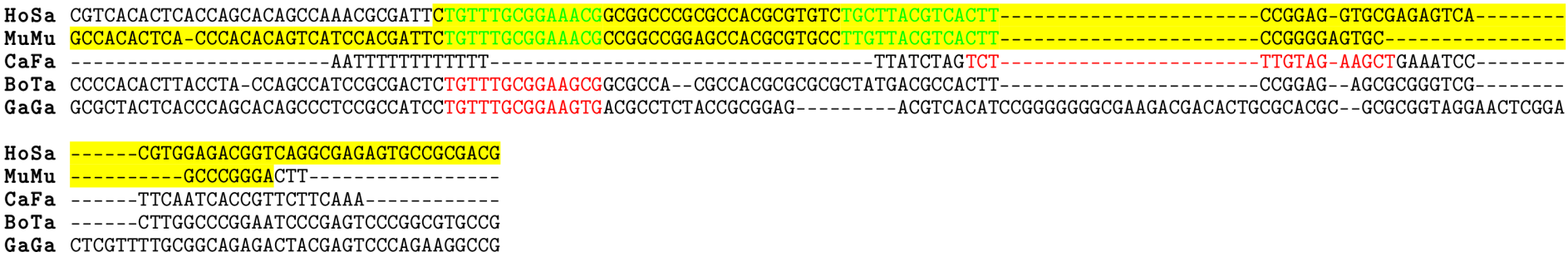
Improved Promoter Alignments using Pro-Coffee

Ionas Erb

Centre de Regulació Genòmica, Barcelona
Cédric Notredame's Group

26th TBI Winterseminar in Bled, Slovenia
13th – 20th February 2011


HoSa CT----GTTTGCGGAAACGGCGGCCCGCGCCACGCGTGTCTGCTTACGTC-ACTTCCGGAG----------
MuMu CAGGCTATCTCTGCTCTTAAATACAC---------AAGATTTAAGACAAAGAGGCAAGGAAA----------GGCCAGAACGCCGAACGCCACAC----------T---C-ACCC
CaFa --------CTGTGGATTTGGTAGCA------ATGTC-----TGT-CTGTT-ACTTCCTAATGGTTAATGATTTATTTTTAAGTGGATTGGTT-TAAAATAGAAATCT----------A--T
BoTa CT----GTTTGCGGAAGCGGCGCCAC--GCCACGCGCGCGCTATGACGCC-ACTTCCGGAGA----------GCGCGGGTCGCTTGGC-CCGGAATCCCGAGTCCCGGCGTGCCGC
GaGa CT----GTTTGCGGAAGTGACGCCTC---------TACCGCGGAGACGTC-ACATCCGGGGG----------GGCGAAGACGACAC----------TGCGC-ACGC
HoSa GTGCGAGA---GTCAC--GTGGAGACGGTCAG--GCG--AG----------AGTGCC-G----CGACG--CACGT----------CCTC-CG-CG-C
MuMu ACACAGT-----CATC--CACGATTCTGTTTG-CGGAAACGCCGGCCGGAGCCA---CGCGTGCCTTGTTACGTCACTTCCGG---GGA----GTGCGCC-CG-GG-A
CaFa GTATAGTC----TAAT--TTTTTTTTTTTTA-TC-T--AGT---------CTT--TGTAGAAGC-TG-AAATCCTT---CAA----------TCACCGTTCTT-C
BoTa GCGCGGCGTGCGGAGGGCGACGGCTCGAAGGGACGCAAGAGC---------CTGGTTGGAGGGAC-GG---CGTGGCGGGCGGGTGGGAGGCTGGTCGCC-TG-AGCC
GaGa GCGCGGTA---GGAAC--TCGGACTCGTTTTG-CGGCAGAGA---------CTA---CGAGTCCC-AG-AAGGCCGCGCGCGG----------GGGCG-GG-AG-C


HoSa CGTCACACTCACCAGCACAGCCAAACGCGATTCTGTTTGCGGAAACGGCGGCCCGCGCCACGCGTGTCTGCTTACGTCACTT----------CCGGAG-GTGCGAGAGTCA--------
MuMu GCCACACTCA-CCCACACAGTCATCCACGATTCTGTTTGCGGAAACGCCGGCCGGAGCCACGCGTGCCTTGTTACGTCACTT----------CCGGGGAGTGC----------
CaFa ----------AATTTTTTTTTTTT----------TTATCTAGTCT----------TTGTAG-AAGCTGAAATCC--------
BoTa CCCCACACTTACCTA-CCAGCCATCCGCGACTCTGTTTGCGGAAGCGGCGCCA--CGCCACGCGCGCGCTATGACGCCACTT----------CCGGAG--AGCGCGGGTCG--------
GaGa GCGCTACTCACCCAGCACAGCCCTCCGCCATCCTGTTTGCGGAAGTGACGCCTCTACCGCGGAG---------ACGTCACATCCGGGGGGGCGAAGACGACACTGCGCACGC--GCGCGGTAGGAACTCGGA
HoSa ------CGTGGAGACGGTCAGGCGAGAGTGCCGCGACG
MuMu ----------GCCCGGGACTT----------
CaFa ------TTCAATCACCGTTCTTCAAA----------
BoTa ------CTTGGCCCGGAATCCCGAGTCCCGGCGTGCCG
GaGa CTCGTTTTGCGGCAGAGACTACGAGTCCCAGAAGGCCG

Problems when aligning DNA

(i) lack of informative structural constraints
(ii) small alphabet, low information content
(iii) heterogeneity of functional features, no uniform model
(iv) more challenges in promoters (duplications, high turnover), loss of colinearity

What we want (and don't want) to do

- Want a 'classical' approach applicable to whole genome alignments
- Use no information about motifs that might occur
- We want to get more footprints than off-the shelf methods

An idea from another project (BlastR)

Use nearest-neighbour correlation structure in sequence.

Transform the DNA alphabet into a pseudo amino-acid alphabet where one letter codes for two neigbouring nucleotides.

$\Rightarrow$ sequence of overlapping di-nucleotides

How do we evaluate substitution costs?

(i) 425 vertebrate TF binding sites alignments from TRANSFAC
(ii) build one big pairwise alignment with spacers between sites
(iii) translate to new alphabet and evaluate (BLOSUM style)

$$\log \frac{p\binom{x}{y}}{p(x)p(y)}$$

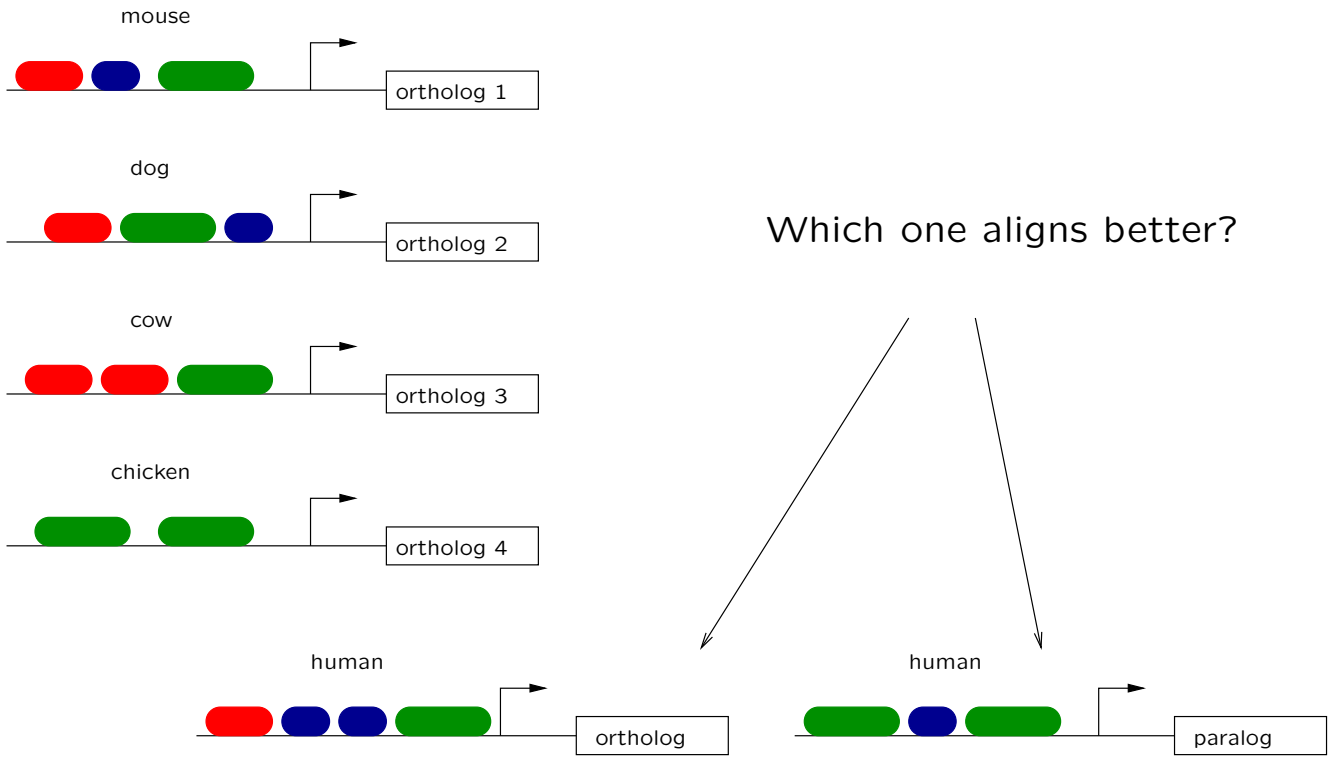
substitution matrix with these entries for translated sequences

Questions

What are the gap costs?

How can we know what is a good alignment?

Alignment evaluation based on homology



Summary of procedure

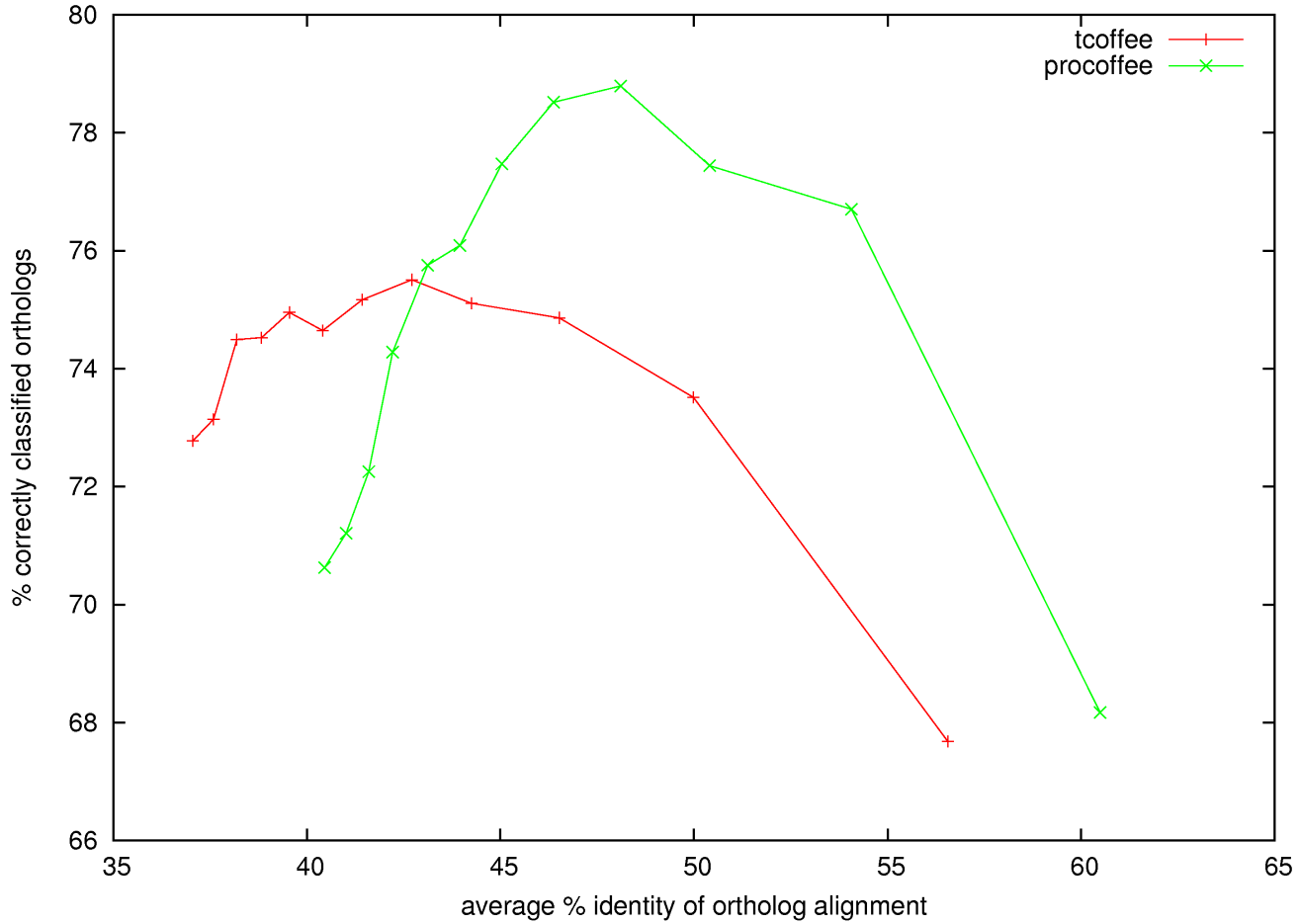
(i) - collect unique orthologs to human genes
in mouse, dog, cow and chicken using ENSEMBL
- only use genes forming cliques of unique orthologs
- for these also collect human paralogs
$\Rightarrow$ 3258 genes with 4 paralogs each on average
- get 500 bp upstream sequences of all ortholgs and paralogs

(ii) compare percent identity of ortholog alignment with
percent identity of each paralog alignment

A sanity check

It works for amino-acid sequence alignments of the gene:
(Default) T-Coffee classifies 98% of the orthologs correctly.

Training gap opening penalties on upstream regions



Results for 2000 bp upstream

Method	% correctly classified orthologs
t-coffee	73.8
probcons	78.7
clustalw	81.8
muscle	82.1
t-coffee trained	82.4
mafft	84.2
pro-coffee trained	86.8

Nice, but we'd like an 'experimental' validation

Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding

Dominic Schmidt,[1,2]* Michael D. Wilson,[1,2]* Benoit Ballester,[3]* Petra C. Schwalie,[3]
Gordon D. Brown,[1] Aileen Marshall,[1,4] Claudia Kutter,[1] Stephen Watt,[1] Celia P. Martinez-Jimenez,[5]
Sarah Mackay,[6] Iannis Talianidis,[5] Paul Flicek,[3,7]† Duncan T. Odom[1,2]†

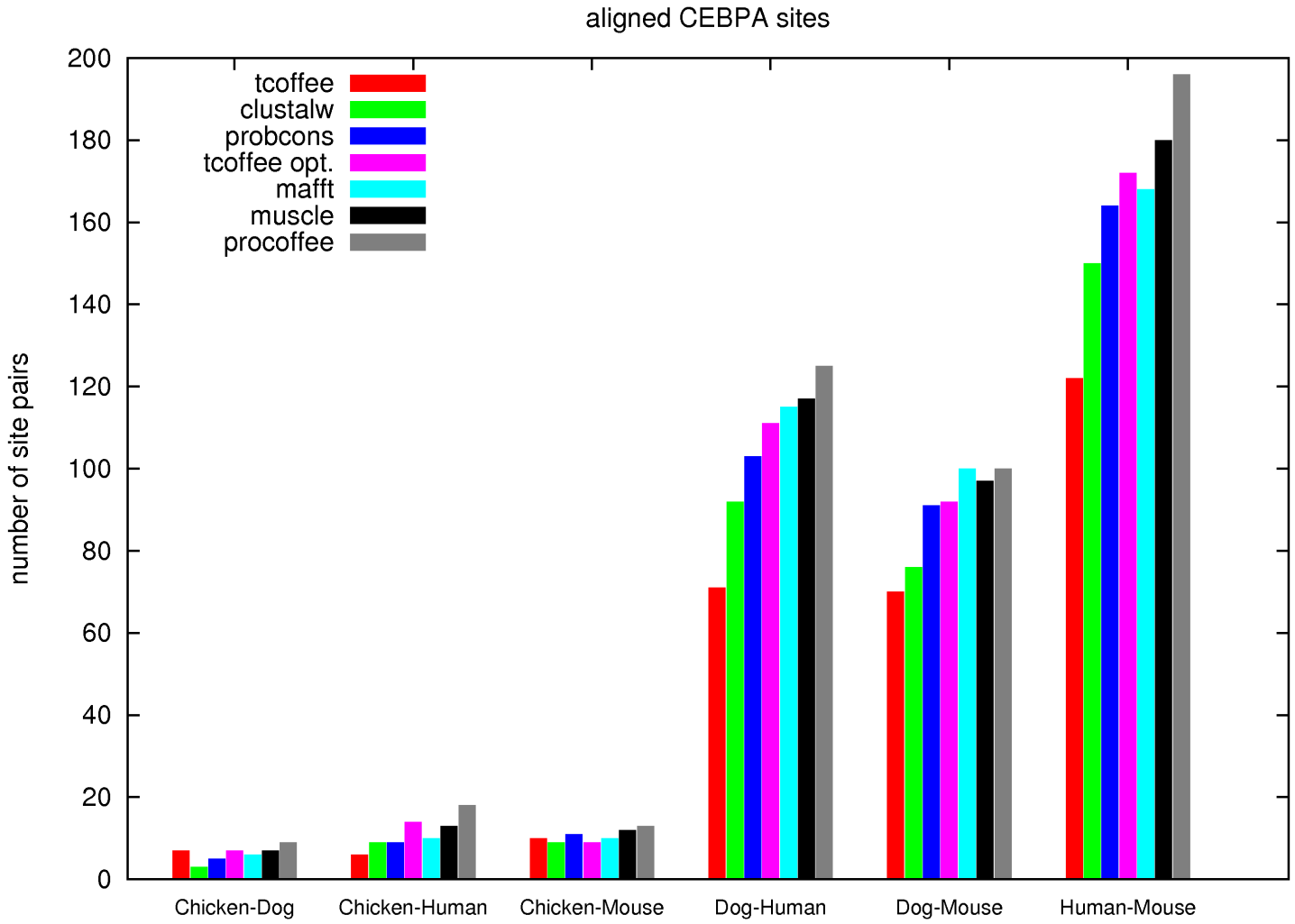
Science **328** (2010)

Summary of procedure

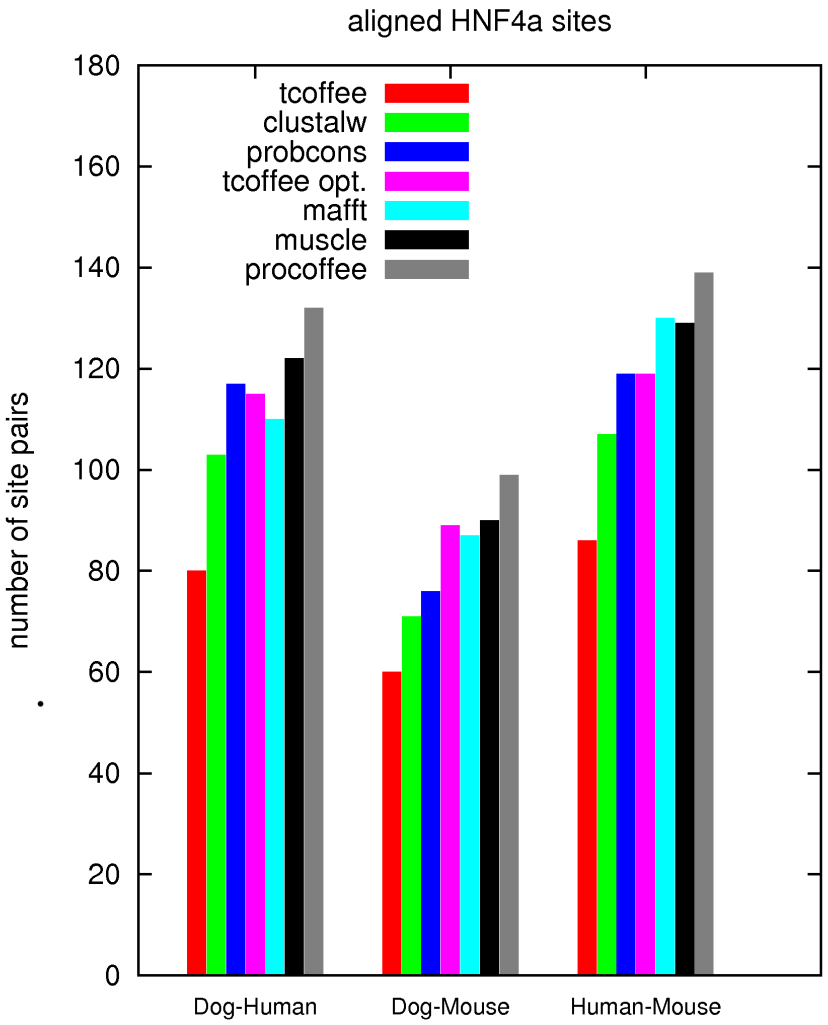
(i) - ChIP-seq raw data available, mapping and peak finding done by J. Gonzáles-Vallinas and E. Eyras
- 100 bp binding regions for two transcription factors:
CEBPA in human, mouse, dog, chicken
HNF4a in human, mouse, dog

(ii) - do motif scan in regions to get validated binding sites
- map regions onto 2477 alignments and count gaplessly aligned sites in pairs of species

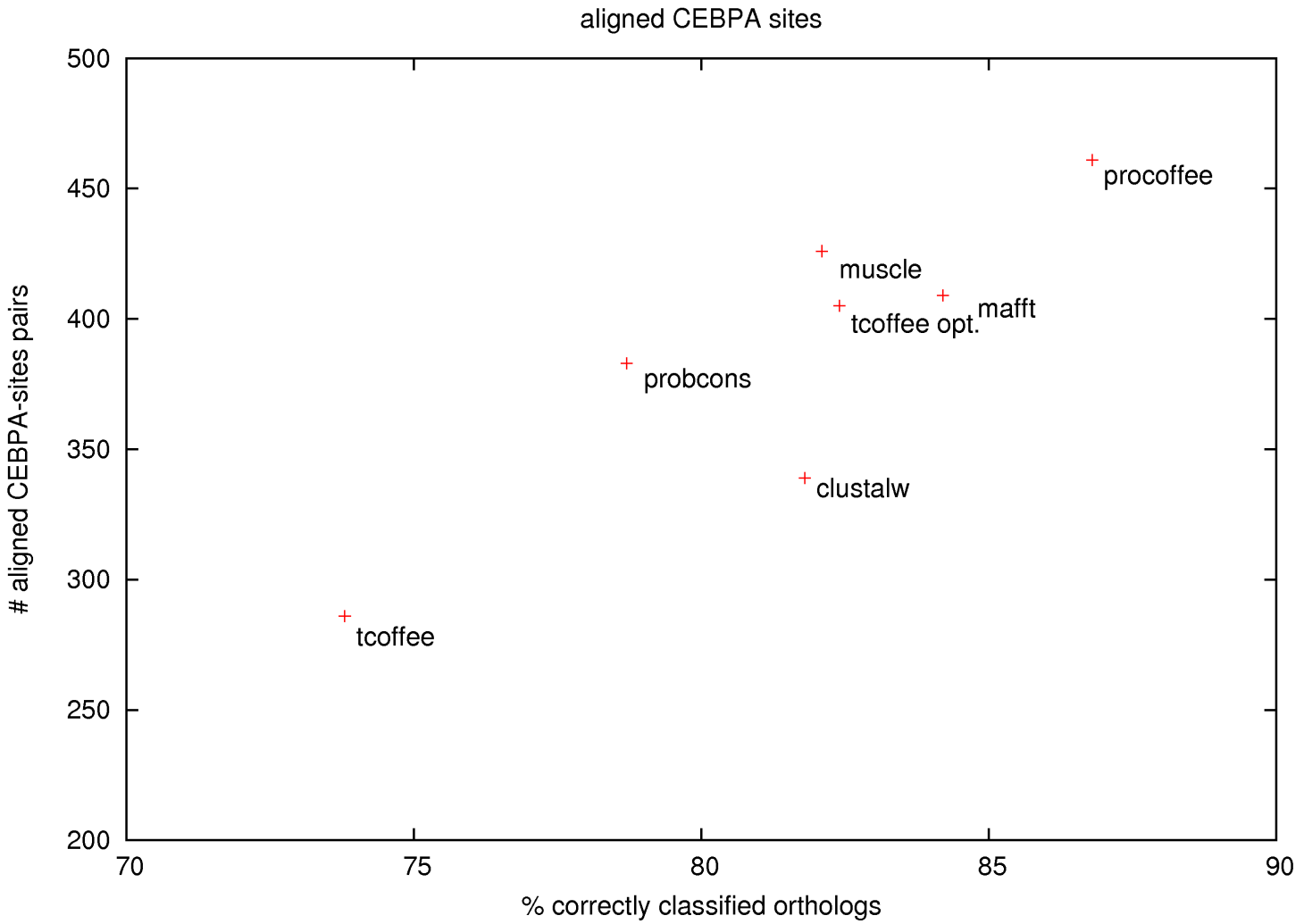
Results CEBPA



Results HNF4a



Correlation between ortholog test and ChIP-seq



Work in progress

- Combine alignments with different gap costs
- Tune competing methods, use our matrix with them
- How much is tuning, how much di-nucleotides?

Main results

(i) New method for promoter alignments
(ii) New validation framework for promoter alignments
(iii) Improvement on ortholog test also leads to better footprints
(iv) Good alignments manage trade-off between increasing identity and maintaining compact blocks

Check it out!

www.tcoffee.org/Projects_home_page/procoffee_home_page.html
command line: t_coffee yourfile.fa -mode=procoffee

Article in preparation

Acknowledgments

joint work with

Cédric Notredame (CRG),
Enrique Blanco (Universitat de Barcelona),
Juan González-Vallinas, Eduardo Eyras (GRIB)

Thank you!