UNDERSTANDING TRANSCRIPTION FACTOR – DNA RECOGNITION USING IN SILICO, IN VITRO AND IN VIVO APPROACHES: THE CSL CASE

UNIVERSITY OF CAMBRIDGE

Rubben Torella, Robert C. Glen*

In collaboration with the Bray and the Adryan groups in Cambridge, UK and the Kovall group in Cincinnati, USA.

Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, UK

*Corresponding author email address rcg28@cam.ac.uk

Unilever Cambridge Centre for Molecular Informatics

Introduction

The regulation of transcription is fundamental to development and physiology, and occurs through binding of transcription factor proteins to specific DNA sequences in the genome. The CSL (CBF1/RBP-J/Suppressor of Hairless/LAG-1) transcription factor is a core component of the Notch signaling pathway and acts in concert with co-activator or co-repressor proteins to control the activity of the associated target genes. One fundamental question is how CSL can recognize and select among different DNA sequences available *in vivo* and what influence these different sequences have on its function. We have investigated CSL-DNA recognition using computational approaches to analyze the energetics of CSL bound to different DNAs and testing the *in silico* predictions with *in vitro* and *in vivo* assays.

In silico CSL-DNA binding prediction

We have investigated whether protein structure-based *in silico* approaches could be used to provide information about the full repertoire of binding sites. To achieve this, the FOLDX software [1] was used to calculate changes in binding energy when varying the nucleotides within the constraints of the CSL-DNA co-crystal structure. Starting from the X-ray structure in which CSL binds the (current) highest-affinity DNA motif comprised of eight nucleotides CGTGGGAA (PDB code 3BRG,[2]), all 4^8 permutations of an 8-nt motif were tested and the resulting 65536 relative binding energies calculated. A threshold of 3 *kcal/mol* difference from the top predicted DNA sequence was used as a cut-off, separating 220 putative "bound" motifs from the residual 65316 DNA sequences. These 220 "bound" sequences were used to generate a binding logo and this was compared with the previously used TRANSFAC logo [3].

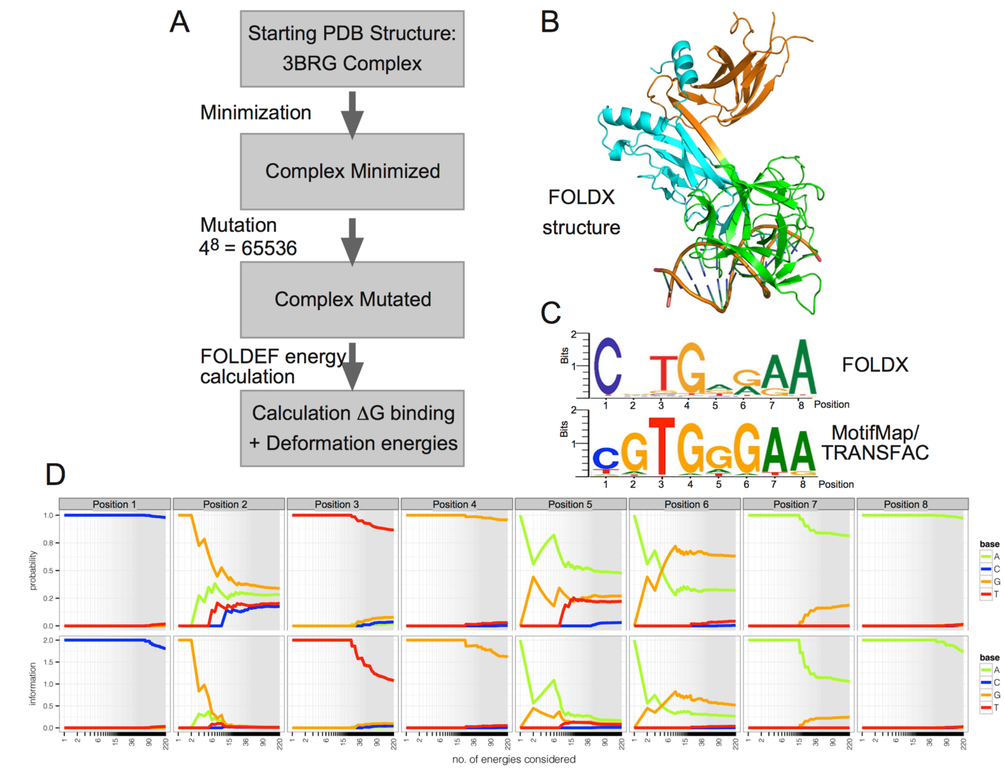

Figure 1: Overview of the FOLDX strategy and results.
A: Flow chart summarizing the FOLDX computational strategy. **B:** CSL-DNA structure used for the analysis (CSL domains NTD cyan; BTD green, CTD orange). The position of the residues that were mutated to perturb inter-domain communication are indicated by orange spheres. **C:** Comparison of energy logos obtained from FOLDX predictions and from empirical binding analysis (RBPJ M01112 MotifMap/Transfac) **D:** At each position, the nucleotide frequencies for different energy thresholds are plotted according to probability and information content.

Several striking features are evident when comparing the FOLDX binding logo with the energy logo obtained from empirical binding analysis (**Fig. 1C**). First, although it is generally considered that C and T are equally tolerated at position 1, FOLDX indicates a strong preference for a cytosine at this location. This difference could be explained by the contacts made by a glutamine residue within the BTD of CSL with the complementary base in position 1 (**Fig.2A**). Second, while there is thought to be a strong preference for guanine at positions 2 and 6, FOLDX indicated much greater sequence tolerance at these positions with little preference at position 2 and tolerance for G or A at position 6.

In vitro and *in vivo* validation of the *in silico* prediction: nucleotide preference in position 1

The FOLDX results suggest that CSL motifs with C at position 1 will have higher binding energies, in contrast to the canonical PWM and to the logo derived from the protein binding microarrays (PBMs). To investigate CSL binding characteristics further, we compared the C/T motif variants in two assays. We tested the binding of CSL to CGTGGGAA and TGTGGGAA via isothermal calorimetry (ITC, **Fig.2C-D**) and then the *in vivo* response to the binding to these two DNA sequences using a luciferase reporter assay (**Fig.2B**).

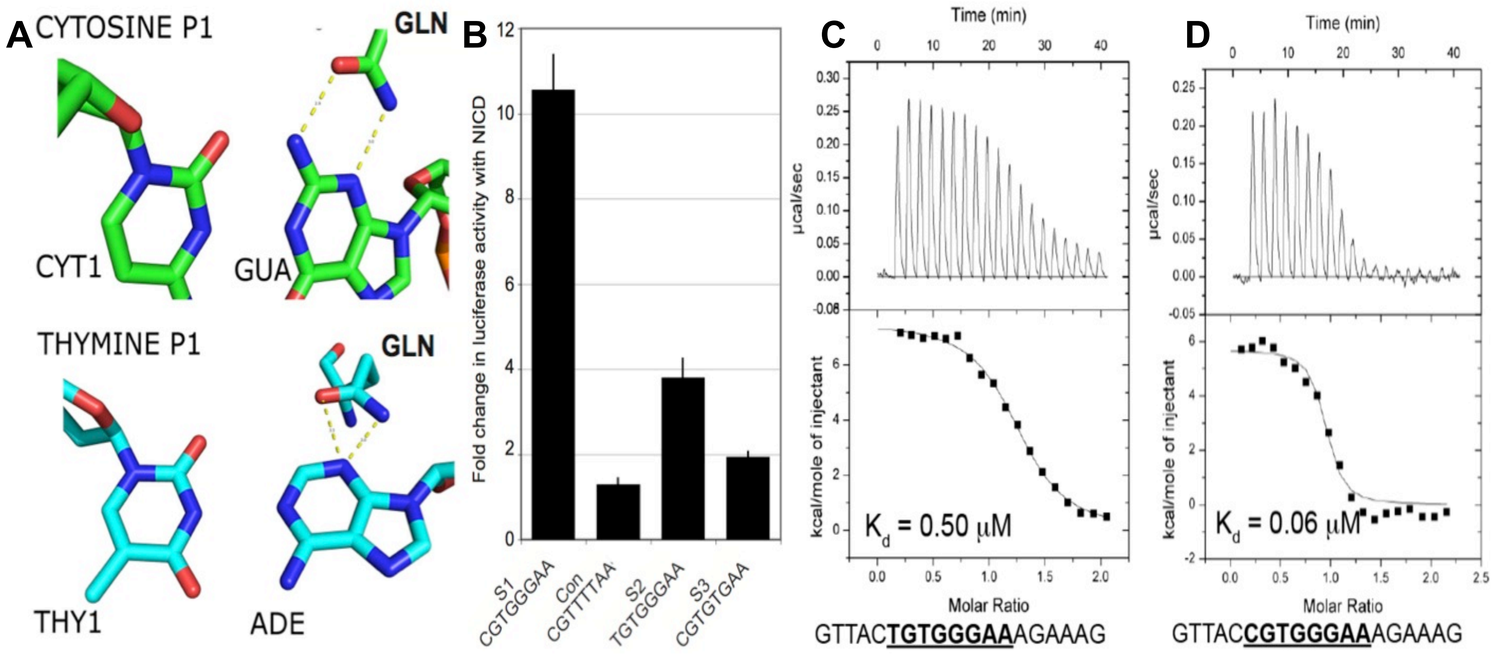

Figure 2: Experimental validation of the in silico prediction of the position 1 CSL DNA preference. A: Chemical explanation for the preference of cytosine over thymine in position 1: the complementary guanine can offer two groups for making H-bond contacts (NH_2 and aromatic N), while the adenine can only offer one group (aromatic N) for the interaction with glutamine. B: Response of reporters containing the indicated oligonucleotides to NICD, measured as fold change in luciferase activity in extracts from transfected cells. C and D: Representative thermograms for CSL binding to TGTGGGAA (C) and CGTGGGAA (D) DNA sequences. Relative affinities and specific DNA sequences are shown for each experiment.

References

[1]Nadra AD *et al.,* **DNA-binding specificity prediction with FoldX.** 2011, Methods Enzymology; 498:3-18.
[2]Friedmann DR *et al.,* **RAM-induced allostery facilitates assembly of a notch pathway active transcription complex.**2008,J.Biol.Chem.; 283(21) 14781-91
[3]Matys V *et al.,* **TRANSFAC®: transcriptional regulation, from patterns to profiles.** 2003, NAR; 31(1): 374–378.

We thank Unilever for funding

CSL can regulate transcription at intermediate levels

The same protocol (ITC + luciferase assay) has been used to test DNA sequences that present mutations in positions 2 and 6 (**Fig. 3**), that the *in silico* prediction suggests are more flexible in their ability to accept different nucleotides.

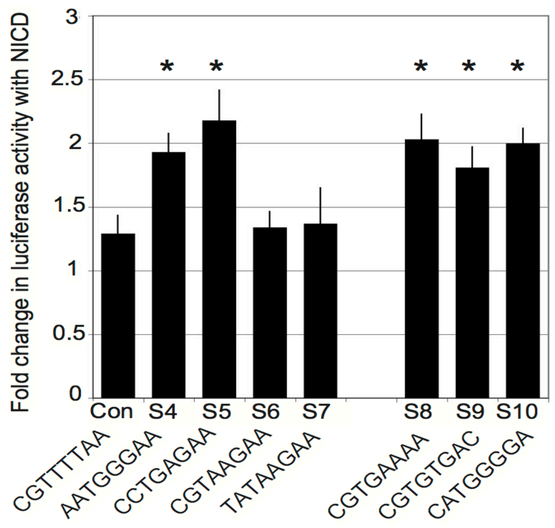

Figure 3: Functional relevance of FOLDX predictions.
Response of reporters containing the indicated oligonucleotides to NICD, measured as fold change in luciferase activity in extracts from transfected cells. Activities were normalized to co-transfected renilla plasmid to control for transfection efficiencies. Error bars depict standard error of the mean from >3 biological replicate experiments. * indicates that the response was significantly different from the control ($p<0.05$, paired t-test).

CSL dynamical response directly affects the transcription

Surprisingly, analysis shows that a range of sequences representing a broad repertoire of motifs can be bound by CSL (**Fig. 2, 3**), with many having similar functional activities. Subtle differences in binding versus functional activities prompted us to investigate the effects of different DNA sequences on CSL dynamical behavior, by performing MD simulations of CSL in the presence of CGTGGGAA, TGTGGGAA, CGTGTGAC, which exhibit intermediate binding and transcriptional regulation and CGTAAGAA, which exhibits little/no binding or activity but has a K_d significantly different from negative controls.
Internal coordination analyses showed (**Fig. 4**) differences in intra-domain correlation between domains, indicating how CSL can recognize different DNA sequences and by changing its internal correlation between domains, transduce a dynamic signal that influences binding of ancillary proteins.

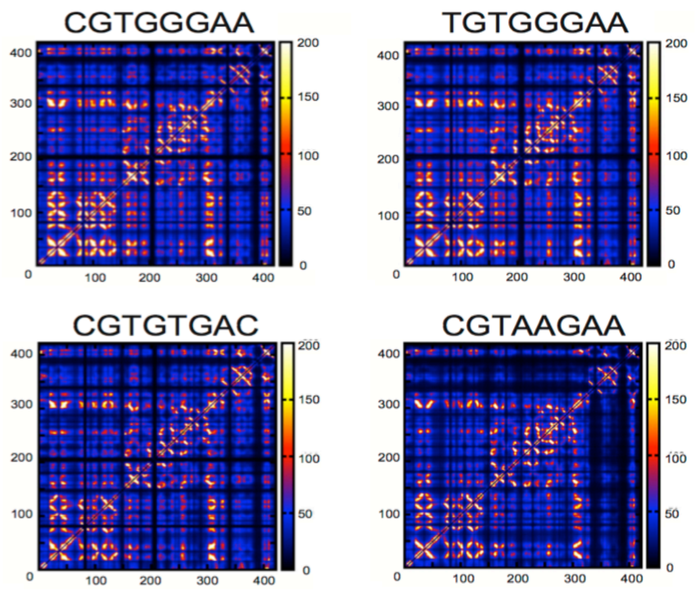

Figure 4: Inter-domain correlation in presence of different DNA sequences. A colour-code has been used to indicate strong residue correlation (yellow), medium correlation (red) weak/no correlation (blue/black).

Mutations on the correlation pathway affects *in vivo* response

The importance of domain-domain correlation has been demonstrated by mutations of residues involved in dynamic correlation between the residues that binds the DNA and residues that are involved in interactions with other proteins, using *in silico* and *in vivo* approaches (**Fig.5**).

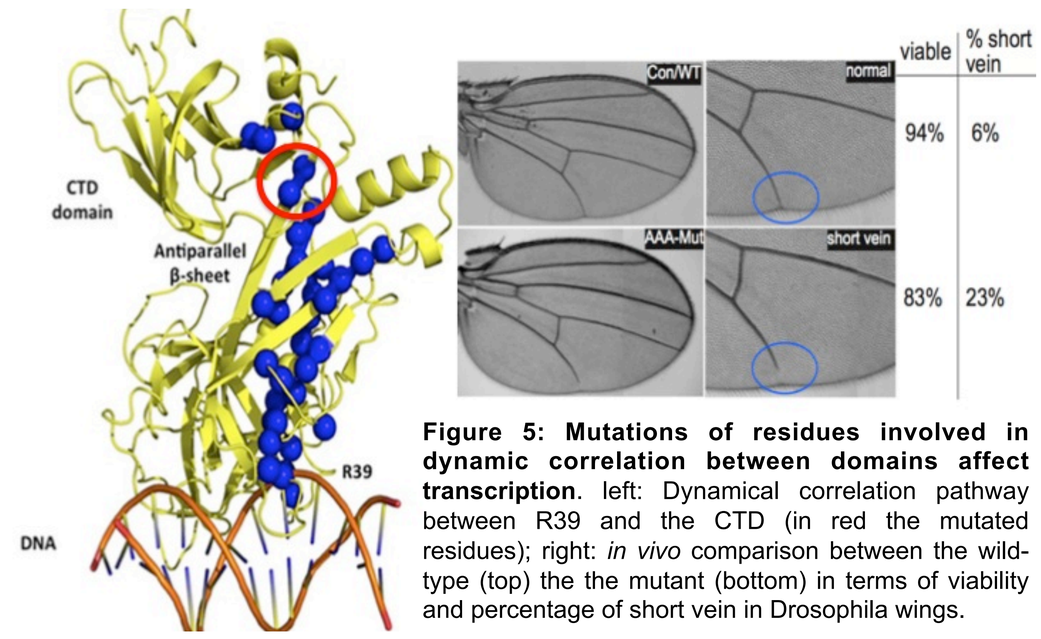

Figure 5: Mutations of residues involved in dynamic correlation between domains affect transcription. left: Dynamical correlation pathway between R39 and the CTD (in red the mutated residues); right: *in vivo* comparison between the wild-type (top) the the mutant (bottom) in terms of viability and percentage of short vein in Drosophila wings.

Conclusions

In silico approaches to investigate the mechanisms of CSL binding have revealed novel features, increasing our understanding of the repertoire of sequences that are functional *in vivo*.

Furthermore, our results predict a profound effect of DNA binding on the inter-domain correlated motions, with lower affinity sequences demonstrating a reduced correlation compared to high affinity sites. *In vivo* results show that mutations of residues involved in signal communication directly affect gene regulation, confirming that the dynamical response to different DNA sequences is critical for protein-DNA recognition and binding *in vivo*.