Comparative modelling and footprinting of protein-DNA complexes

Bruno Contreras-Moreira and Julio Collado-Vides

contrera@ccg.unam.mx
Centro de Ciencias Genómicas.
Universidad Nacional Autónoma de México
Cuernavaca, Morelos, 62210, México

Motivation:
Interactions between proteins and DNA molecules lie at the core of fundamental cellular processes such as transcriptional regulation. Some of these interactions have been experimentally described at atomic scale, but the molecular details of many others remain to be discovered. This work investigates ways to exploit the current knowledge about protein-DNA interfaces contained in the Protein Data Bank, with two aims: 1) modelling similar interfaces related by homology and 2) identifying genomic regulatory motifs.

Results:
First, we describe the structural and evolutionary conservation of protein-DNA interfaces, and the limits they impose on modelling accuracy. Second, we estimate the error rate associated to modelling interface amino acid side chains, those likely to be contacting nucleotide N-bases. Third, a simplistic protocol is implemented and different parameters are benchmarked on a set of 85 regulators from *Escherichia coli* , . Finally, this approach is made available via the TFmodeller web server.

protein-DNA interface conservation and modelling limitations

Fig1

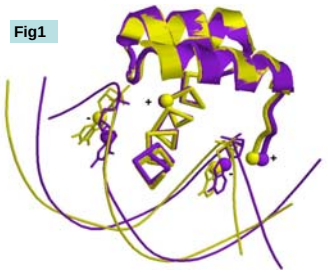


Fig2

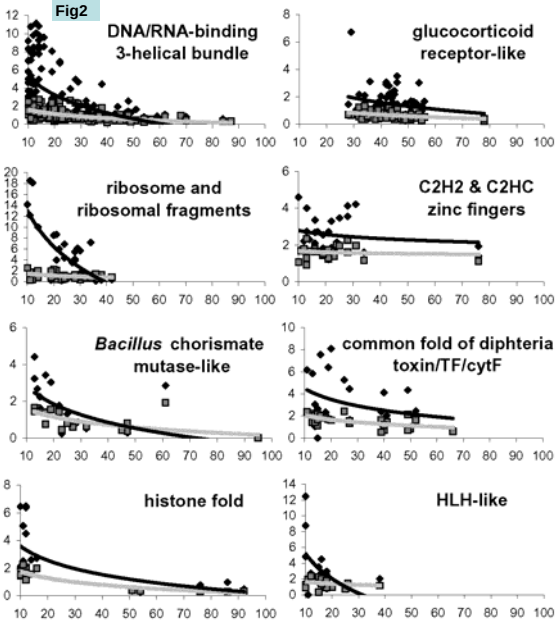


Fig3

REL	0.2	0.4	0.6	0.8	RMSD	0.2	0.4	0.6	0.8	obs	0.2	0.4	0.6	0.8
ASP	0.50	0.57	0.00	----	ASP	1.78	1.35	1.60	----	ASP	028	021	003	---
ASN	0.63	0.75	0.86	0.86	ASN	1.11	0.70	1.24	0.28	ASN	252	061	014	022
LYS	0.52	0.65	0.90	0.70	LYS	1.47	1.07	0.66	0.68	LYS	204	175	010	010
TYR	0.40	0.38	0.50	----	TYR	2.44	2.00	3.54	----	TYR	015	008	002	---
GLU	0.50	0.70	0.60	0.80	GLU	1.51	1.10	0.76	1.18	GLU	115	122	005	005
ARG	0.48	0.73	0.63	0.71	ARG	2.13	1.41	1.20	1.16	ARG	507	283	054	021
CYS	1.00	1.00	----	----	CYS	0.90	0.99	----	----	CYS	002	001	---	---
THR	0.23	0.40	0.40	0.80	THR	1.28	0.90	0.70	0.39	THR	013	005	005	010
HIS	0.42	0.18	0.33	----	HIS	1.88	2.10	2.00	----	HIS	036	011	006	---
GLN	0.31	0.67	0.90	0.71	GLN	1.93	1.20	1.00	0.39	GLN	051	039	010	007
SER	0.39	0.18	0.56	----	SER	2.05	1.57	1.56	----	SER	038	011	009	---
TRP	----	----	1.00	----	TRP	----	----	0.46	----	TRP	---	---	002	---

Total observations = 2193

Figure 1 shows two superposed *Drosophila* TFs with protein sequences that are 40% identical, highlighting their structural conservation. Figure 2 shows the conservation trends observed with 442 superposed DNA-binding domains annotated in SCOP, with grey data corresponding to the conservation of the protein interface (+) and the black points to the DNA interface (-). This data is derived from 442 superposed protein-DNA complexes, that were also analyzed to generate Table 3, that shows how accurately different H-bonding residues are modelled by SCWRL. Accuracy is measured in terms of RMSD and in terms of coincident DNA contacts (REL).

Fig4

	G	A	T	C
Gly	–3.93	–3.93	–3.93	–3.93
Ala	–3.93	–3.93	0.66	–3.72
Val	–3.93	–3.93	–0.17	–3.57

footprint benchmark and TFmodeller web server (www.ccg.unam.mx/tfmodeller)

Fig5

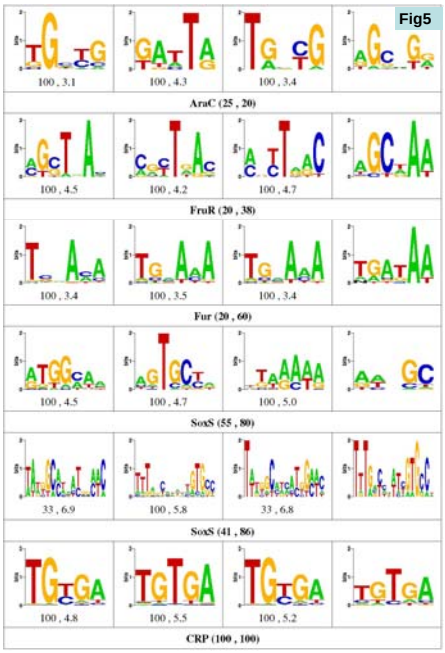


Fig6

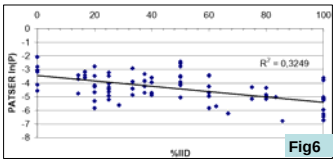
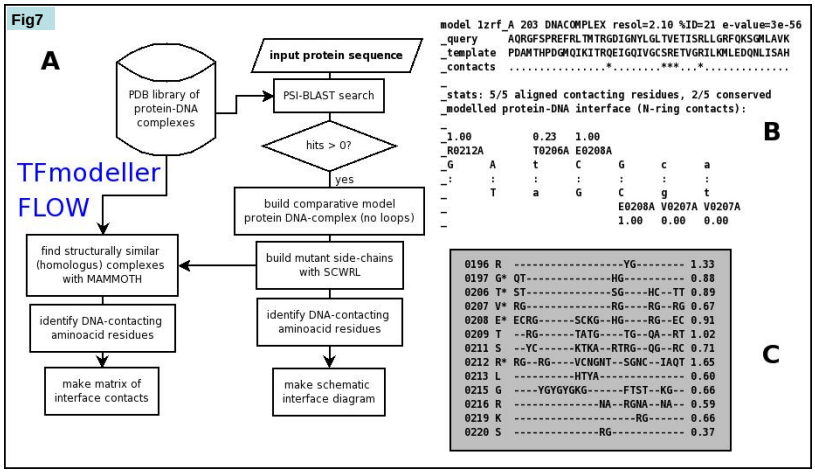


Fig7



By using a simplistic scoring scheme (Table 4) it is possible to estimate the preferred DNA sequences discriminated by a set of 85 modelled *E.coli* TFs. Results from selected examples are shown in Figure 5, including TFs modelled with remote templates, such as AraC, and perfect templates, such as CRP. Templates are evaluated with two measures (interface, and overall sequence identity). Three different predictions are shown for each TF (left), next to the sequence logo derived from annotated binding sites in RegulonDB (right). Figure 6 shows a significant correlation between the accuracy of the predicted DNA motifs and the interface identity od the template used to build the complex. Figure 7 summarizes how this approach was implemented in order to set up the TFmodeller server, and how the output looks like.

References

Contreras-Moreira, B., Branger, P.A. & Collado-Vides, J. (2007). TFmodeller: comparative modelling of protein-DNA complexes. Bioinformatics, doi: 10.1093/bioinformatics/btm148.

Contreras-Moreira, B. & Collado-Vides, J. (2006). Comparative footprinting of DNA-binding proteins. Bioinformatics, 22(14): e74.