BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 13 2007, pages 1694–1696
doi:10.1093/bioinformatics/btm148

Structural bioinformatics

TFmodeller: comparative modelling of protein–DNA complexes

Bruno Contreras-Moreira*, Pierre-Alain Branger and Julio Collado-Vides

Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av.Universidad, s/n, 62210 Cuernavaca, Morelos, México

Received on March 15, 2007; revised on April 11, 2007; accepted on April 12, 2007

Advance Access publication April 25, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Interactions between proteins and DNA molecules lie at the core of the 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. TFmodeller exploits the current knowledge about protein–DNA interfaces contained in the Protein Data Bank and uses it to model similar interfaces related by homology. Results are emailed to the user and include an evolutionary contact matrix, a schematic representation of the putative binding interface and atomic coordinates of the modelled complex. The library of complexes used by TFmodeller is updated on a weekly basis and is available for download.

Availability: TFmodeller and its web service interface are free for academic users at http://www.ccg.unam.mx/tfmodeller

Contact: contrera@ccg.unam.mx

1 INTRODUCTION

Fundamental cellular processes such as transcriptional regulation can be decomposed into a series of molecular interactions, such as protein–nucleic acid complexes. Structural studies have been very important in unveiling atomic details of these interactions, improving our understanding of cellular biology. A notable example is the structure of the 30S ribosome subunit, which was very helpful for understanding the mechanism of some antibiotics (Wimberly *et al*., 2000).

Atomic-scale descriptions of such molecular complexes are routinely deposited in the Protein Data Bank (PDB) (Berman *et al*., 2000) and can be exploited in a variety of ways. The computer program presented here, TFmodeller, takes advantage of the collection of protein–DNA complexes contained in the PDB by using comparative modelling, a procedure that approximates the 3D arrangement of a protein sequence given an alignment to template proteins of known structure (Marti-Renom *et al*., 2000).

Comparative modelling tools such as SWISS-MODEL (Guex *et al*., 1999) are widely used by the experimental community. However, there is currently no equivalent tool for the comparative analysis of protein–DNA complexes. TFmodeller aims to fill that gap, supported by recent observations about the conservation of protein–DNA docking geometries (Contreras-Moreira and Collado-Vides, 2006; Siggers *et al*., 2005) and the fact that homologous transcription factors bind to similar DNA sequences (Sandelin and Wasserman, 2004).

2 MODELLING PROCEDURE

The steps followed by TFmodeller, as in Figure 1A, are as follows:

(1) Search for homologous protein–DNA complexes in a library of complexes with three iterations of PSI-BLAST (Altschul *et al*., 1997). Alternatively, the program can use templates provided by the user.

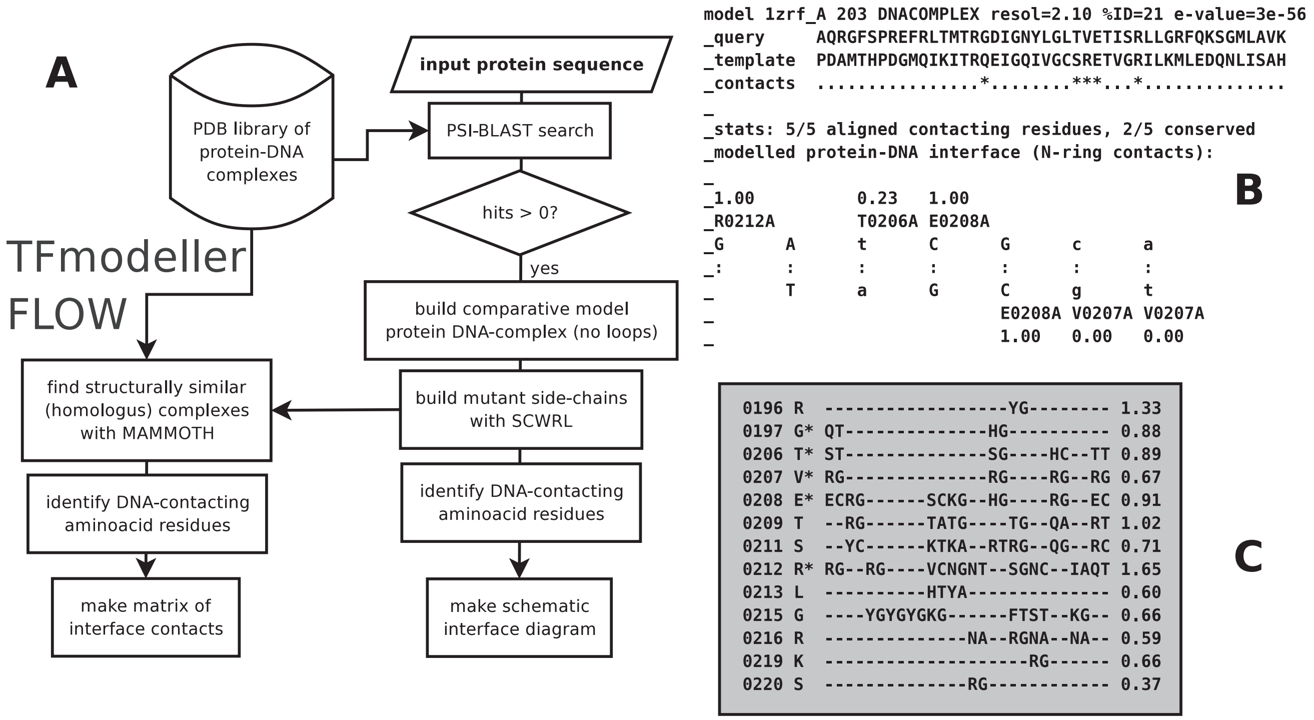
(2) Use local PSI-BLAST alignments to build the protein backbone of the modelled complex. Missing loops are not built. Users can also provide sequence alignments for their uploaded templates.

(3) Add mutated amino acid side-chains keeping the template DNA in frame, using SCWRL (Canutescu *et al*., 2003), take the rest from the template. The final product is called a model, and will be eventually emailed to the user.

For each model a schematic interface diagram is calculated (Fig. 1B), assigning a statistical contact reliability measure to every modelled side-chain (derived from a benchmark that included 2193 modelled H-bonding interface residues). The diagram also highlights those parts of the DNA motif most likely to have changed because of interface mutations. Only nitrogen base contacts are considered, for only those can be sequence specific. The distance threshold for contacts is set to 4.1 Å. Indirect readout mechanisms are not currently supported by this software.

In addition, for every successfully modelled input sequence, a matrix of structurally homologous interface contacts is built, by scanning the library of complexes and aligning all similar complexes found by the program MAMMOTH (Ortiz *et al*., 2002). These matrices are multiple alignments of protein–DNA interfaces in which equivalent contacts are aligned, as shown in Figure 1C. This interface alignment is also used to estimate the putative binding specificity of the input sequence, calculated as the ratio specific contacts/family contacts, using a similar

*To whom correspondence should be addressed.


A
PDB library of protein-DNA complexes
input protein sequence
PSI-BLAST search
hits > 0?
yes
TFmodeller FLOW
build comparative model protein DNA-complex (no loops)
find structurally similar (homologus) complexes with MAMMOTH
build mutant side-chains with SCWRL
identify DNA-contacting aminoacid residues
identify DNA-contacting aminoacid residues
make matrix of interface contacts
make schematic interface diagram
B
C

```
model 1zrf_A 203 DNACOMPLEX resol=2.10 %ID=21 e-value=3e-56
_query     AQRGFSPREFRLTMTRGDIGNYLGLTVETISRLLGRFQKSGMLAVK
_template  PDAMTHPDGMQIKITRQEIGQIVGCSRETVGRILKMLEDQNLISAH
_contacts  ................*........***...*.............
_
_stats: 5/5 aligned contacting residues, 2/5 conserved
_modelled protein-DNA interface (N-ring contacts):
_
_1.00         0.23   1.00
_R0212A       T0206A E0208A
_G      A      t      C      G      c      a
_:      :      :      :      :      :      :
_       T      a      G      C      g      t
_                            E0208A V0207A V0207A
_                            1.00   0.00   0.00
```

```
0196 R  ------------------YG-------- 1.33
0197 G* QT--------------HG---------- 0.88
0206 T* ST--------------SG----HC--TT 0.89
0207 V* RG--------------RG----RG--RG 0.67
0208 E* ECRG------SCKG--HG----RG--EC 0.91
0209 T  --RG------TATG----TG--QA--RT 1.02
0211 S  --YC------KTKA--RTRG--QG--RC 0.71
0212 R* RG--RG----VCNGNT--SGNC--IAQT 1.65
0213 L  ----------HTYA-------------- 0.60
0215 G  ----YGYGYGKG------FTST--KG-- 0.66
0216 R  --------------NA--RGNA--NA-- 0.59
0219 K  --------------------RG------ 0.66
0220 S  --------------RG------------ 0.37
```

Fig. 1. (**A**) Flow chart of TFmodeller. (**B**) Alignment to template complex 1zrf_A, with a total of five contacting residues, of which two are conserved in the query sequence. A schematic interface diagram is also shown, with threonine 206 from chain A assigned a contacting probability of 0.23. The thymine base contacted by T0206A is in lower case since this amino acid was originally a serine in the template and might have changed. Arginine 212 and glutamate 208 have a contact probability of 1.00, since they are conserved in 1zrf_A. (**C**) Matrix of homologous interface contacts with the query sequence on the left and subsequent columns corresponding to complexes from the library (query interface residues are marked with asterisks). For instance, threonine 206 is aligned to four interface contacts in four different templates: ST, SG, HC and TT, where the first letters stands for the amino acid residue and the second for the nitrogen base. The aligned threonine–thymine (TT) contact supports the previous prediction that T0206A might be contacting a T base. A fully explained example is presented in the tutorial found at http://www.ccg.unam.mx/tfmodeller.

approach to that proposed by (Luscombe and Thornton, 2002). The conservation of interface residues is also reported in the entropy column, using the position-specific scoring matrices generated by PSI-BLAST.

2.1 Library of complexes

Every week the 95% non-redundant clustering of protein chains is downloaded from the PDB, taking only chains complexed with DNA molecules. For each cluster the chain with best resolution is selected (this set is available for download). Finally, in order to perform BLAST searches, these selected chain sequences are merged with weekly Swiss-Prot updates (Wu *et al*., 2006).

2.2 Scope

TFmodeller can be used to model complexes from prokaryotes and eukaryotes. Previous work demonstrated the use of this tool for modelling bacterial regulators (Contreras-Moreira and Collado-Vides, 2006). Within eukaryotes, we find that 52–82% of curated transcriptional regulators from *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana* can be modelled with templates that conserve at least 50% of the interface amino acid residues. A benchmark on the performance with Zn-finger proteins remains to be done.

ACKNOWLEDGEMENTS

We thank Romualdo Zayas and Víctor del Moral for their technical support. This work has been supported by a fellowship from Universidad Nacional Autónoma de México awarded to B.C.M. and by NIH grant RO1-GM071962.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.

Canutescu,A.A. *et al*. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*., **12**, 2001–2014.

Contreras-Moreira,B. and Collado-Vides,J. (2006) Comparative footprinting of DNA-binding proteins. *Bioinformatics*, **22**, E74–E80.

Guex,N. *et al*. (1999) Protein modelling for all. *Trends Biochem Sci*., **24**, 364–367.

Luscombe,N.M. and Thornton,J.M. (2002) Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol*., **320**, 991–1009.

Marti-Renom,M.A. *et al*. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct*., **29**, 291–325.

Ortiz,A.R. *et al*. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*., **11**, 2606–2621.

Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol*., **338**, 207–215.

Siggers,T.W. *et al*. (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol*., **345**, 1027–1045.

Wimberly,B.T. *et al*. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.

Wu,C.H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–D191.