Nucleic Acids Research, 2004, Vol. 32, Database issue **D171–D173**
DOI: 10.1093/nar/gkh101

TrSDB: a proteome database of transcription factors

Antoni Hermoso, Daniel Aguilar, Francesc X. Aviles and Enrique Querol*

Institut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

Received August 13, 2003; Revised and Accepted October 13, 2003

ABSTRACT

TrSDB—TranScout Database—(http://ibb.uab.es/trsdb) is a proteome database of eukaryotic transcription factors based upon predicted motifs by TranScout and data sources such as InterPro and Gene Ontology Annotation. Nine eukaryotic proteomes are included in the current version. Extensive and diverse information for each database entry, different analyses considering TranScout classification and similarity relationships are offered for research on transcription factors or gene expression.

INTRODUCTION

Transcription factors have a major role in the complex regulation of gene expression. Their identification, characterization and the exploration of their diversity is an actual important step towards understanding determinant biological processes such as cell development, tissue differentiation and apoptosis. Knowledge of the triggers and effectors that link these latter processes with their expression basis have evident biological and biomedical implications. Classifying transcription factors in related groups among different organisms (some used as model organisms) may help us to highlight evolutionarily conserved or dissimilar strategies and propose shared solutions to different problems where related transcription factors occur (1–3).

In recent years, a large amount of information has been accumulated about proteins with transcriptional regulatory activity. This knowledge has been progressively integrated in some comprehensively annotated biological databases such as TRANSFAC (4). In our approach, TranScout classification and algorithm (5) have been used for analyzing, by the time of writing, nine eukaryotic non-redundant proteomes (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Guillardia theta*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*). Predictive analyses with ProtLoc (cellular localization assignment tool) (6), TransMem (transmembrane domains prediction program) (7), InterPro (8) and Gene Ontology Annotation (GOA) (9) are also integrated in our database, providing additional validation to our results.

BROWSER FACILITIES

TrSDB can be freely accessed at the website: http://ibb.uab.es/trsdb. Users may query TrSDB after selecting an organism by entering keywords to be searched for in the sequence annotation (e.g cancer), by sequence accession code (e.g. Q9Y5A6) or by accession codes that match InterPro (e.g. IPR001356), GOA (e.g. GO:0003700) or TranScout analyses or predictions (e.g. TRS0023). If preferred, queries may be restricted to two subsets: those entries showing positive TranScout matches (which are more likely to be transcription factors) or those that are annotated by InterPro. After submitting a query, a list of matching entries with a brief description (alternative sequence accession codes and annotation) is displayed (Fig. 1).

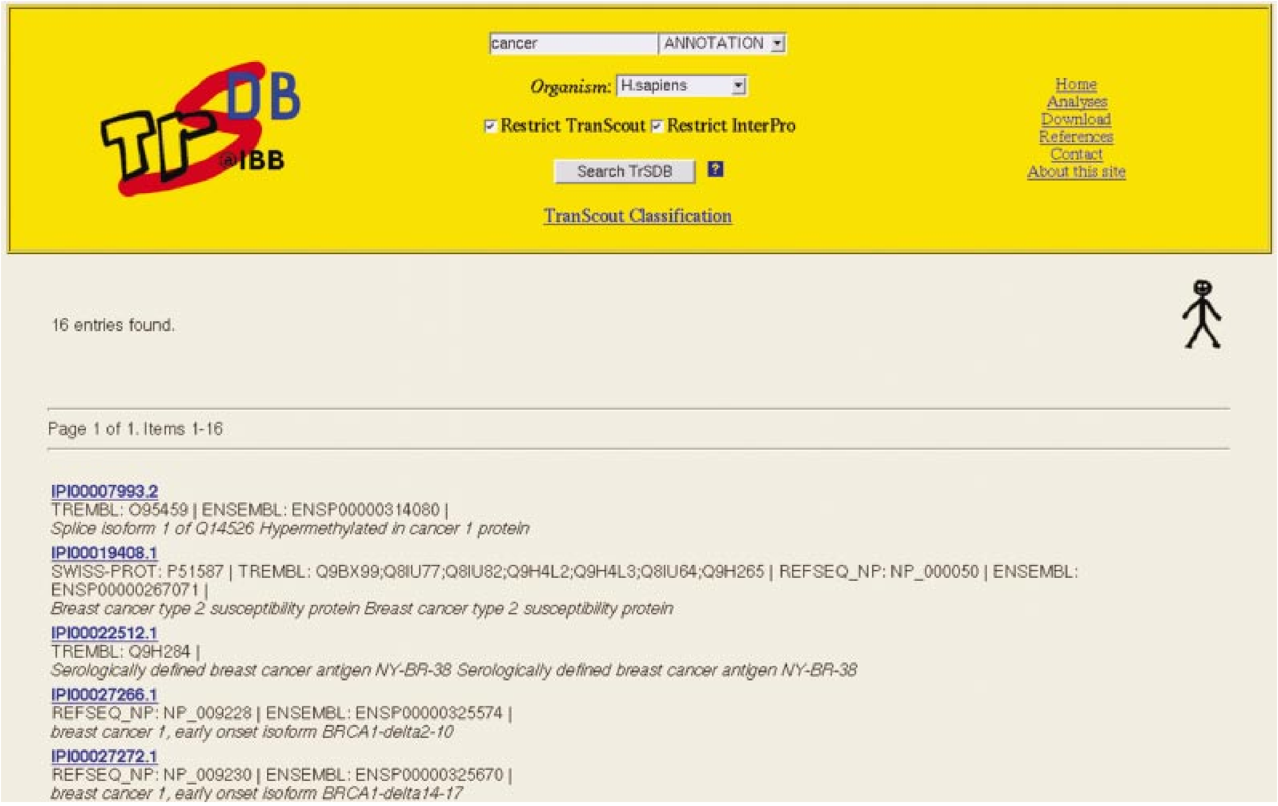
When clicking one entry, all TrSDB available information about this sequence is shown, ranging from relevant external database crosslinks to matches of used predictive programs. A graphic is available with all the assumed protein signatures located along the sequence. Each matching signature is displayed in a different colour together with the associated signature code of the prediction (Fig. 2).

At the bottom of the page, NCBI BLAST (10) premade analyses against the proteomes in the current database (using default parameters and a E-value threshold of 10^{-20}) are available in order to allow users to browse similar entries within the same organism or in others.

PROTEOME ANALYSES

Besides proteome browser facilities, TrSDB offers a set of compiled analyses and statistics of TranScout predictions with significant annotation from other resources. Derived sets of analyses may help researchers to test previously annotated or reannotated proteins that may have a novel and/or conflicting function. A selected list of GO entries are used as a criterion for considering which entries may have been previously annotated as transcription factors by cross-checking against the InterPro (through InterPro2GO) and/or GOA databases. Previously poorly annotated entries are highlighted since they are hot candidates for which our predictions could be particularly useful. Furthermore, a list of proteome entries detected as positive by TranScout, which have high similarity (BLAST with default parameters and a E-value threshold of 10^{-20}) to other TranScout positive protein entries in TrSDB are offered for each organism in the database.

*To whom correspondence should be addressed. Tel: +34 93 5811429; Fax: +34 93 5812011; Email: Enric.Querol@uab.es



cancer ANNOTATION

Organism: H.sapiens

☑ Restrict TranScout ☑ Restrict InterPro

16 entries found.

Page 1 of 1. Items 1-16

IPI00007993.2
TREMBL: O95459 | ENSEMBL: ENSP00000314080 |
Splice isoform 1 of Q14526 Hypermethylated in cancer 1 protein

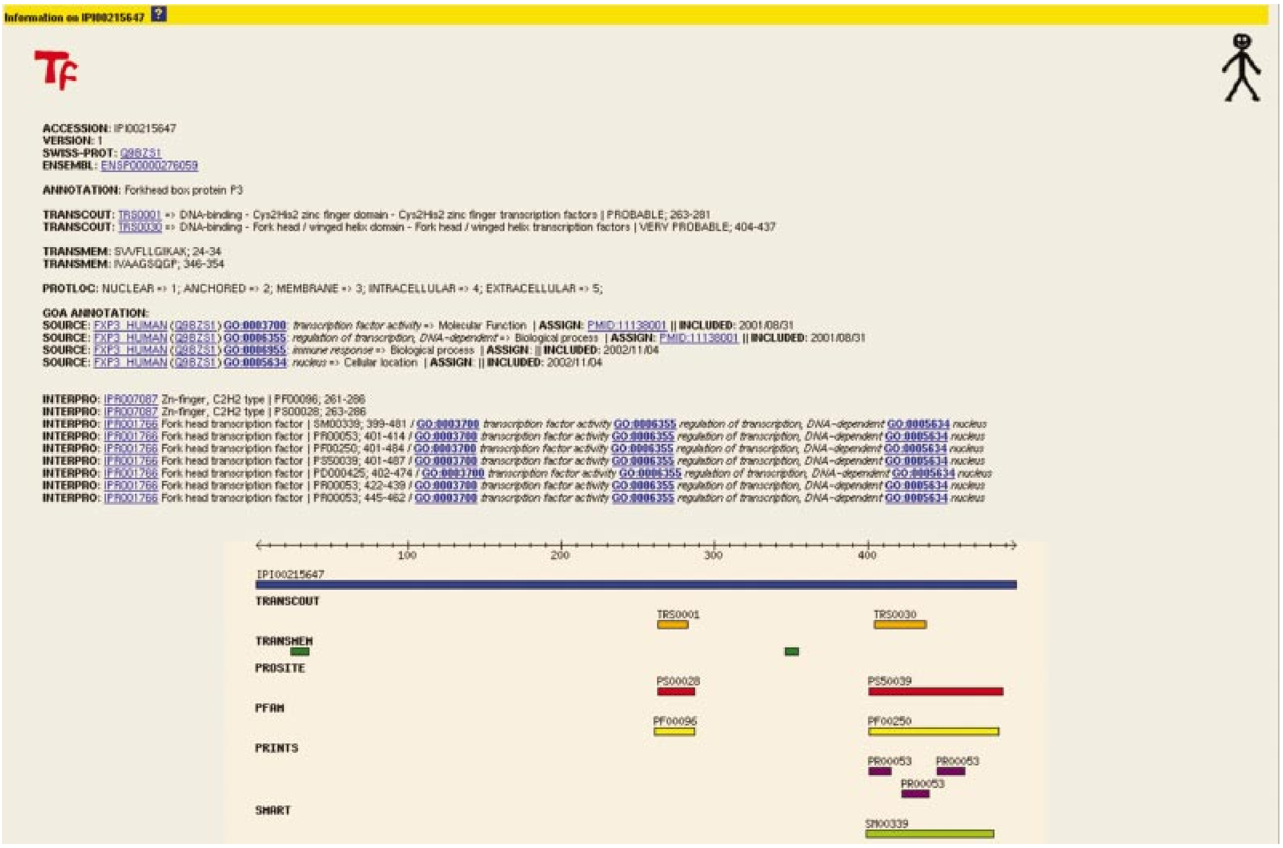
IPI00019408.1
SWISS-PROT: P51587 | TREMBL: Q9BX99;Q8IU77;Q8IU82;Q9H4L2;Q9H4L3;Q8IU64;Q9H265 | REFSEQ_NP: NP_000050 | ENSEMBL: ENSP00000267071 |
Breast cancer type 2 susceptibility protein Breast cancer type 2 susceptibility protein

IPI00022512.1
TREMBL: Q9H284 |
Serologically defined breast cancer antigen NY-BR-38 Serologically defined breast cancer antigen NY-BR-38

IPI00027266.1
REFSEQ_NP: NP_009228 | ENSEMBL: ENSP00000325574 |
breast cancer 1, early onset isoform BRCA1-delta2-10

IPI00027272.1
REFSEQ_NP: NP_009230 | ENSEMBL: ENSP00000325670 |
breast cancer 1, early onset isoform BRCA1-delta14-17

Figure 1. Detail of a query search output of the organism-based TrSDB browser.



ACCESSION: IPI00215647
VERSION: 1
SWISS-PROT: Q9BZS1
ENSEMBL: ENSP00000276059

ANNOTATION: Forkhead box protein P3

TRANSCOUT: TRS0001 => DNA-binding - Cys2His2 zinc finger domain - Cys2His2 zinc finger transcription factors | PROBABLE; 263-281
TRANSCOUT: TRS0030 => DNA-binding - Fork head / winged helix domain - Fork head / winged helix transcription factors | VERY PROBABLE; 404-437

TRANSMEM: SVVFLLGIKAK; 24-34
TRANSMEM: IVAAGSQGP; 346-354

PROTLOC: NUCLEAR => 1; ANCHORED => 2; MEMBRANE => 3; INTRACELLULAR => 4; EXTRACELLULAR => 5;

GOA ANNOTATION:
SOURCE: FXP3_HUMAN (Q9BZS1) GO:0003700: *transcription factor activity* => Molecular Function | **ASSIGN:** PMID:11138001 || **INCLUDED:** 2001/08/31
SOURCE: FXP3_HUMAN (Q9BZS1) GO:0006355: *regulation of transcription, DNA-dependent* => Biological process | **ASSIGN:** PMID:11138001 || **INCLUDED:** 2001/08/31
SOURCE: FXP3_HUMAN (Q9BZS1) GO:0006955: *immune response* => Biological process | **ASSIGN:** || **INCLUDED:** 2002/11/04
SOURCE: FXP3_HUMAN (Q9BZS1) GO:0005634: *nucleus* => Cellular location | **ASSIGN:** || **INCLUDED:** 2002/11/04

INTERPRO: IPR007087 Zn-finger, C2H2 type | PF00096; 261-286
INTERPRO: IPR007087 Zn-finger, C2H2 type | PS00028; 263-286
INTERPRO: IPR001766 Fork head transcription factor | SM00339; 399-481 / GO:0003700 *transcription factor activity* GO:0006355 *regulation of transcription, DNA-dependent* GO:0005634 *nucleus*
INTERPRO: IPR001766 Fork head transcription factor | PR00053; 401-414 / GO:0003700 *transcription factor activity* GO:0006355 *regulation of transcription, DNA-dependent* GO:0005634 *nucleus*
INTERPRO: IPR001766 Fork head transcription factor | PF00250; 401-484 / GO:0003700 *transcription factor activity* GO:0006355 *regulation of transcription, DNA-dependent* GO:0005634 *nucleus*
INTERPRO: IPR001766 Fork head transcription factor | PS50039; 401-487 / GO:0003700 *transcription factor activity* GO:0006355 *regulation of transcription, DNA-dependent* GO:0005634 *nucleus*
INTERPRO: IPR001766 Fork head transcription factor | PD000425; 402-474 / GO:0003700 *transcription factor activity* GO:0006355 *regulation of transcription, DNA-dependent* GO:0005634 *nucleus*
INTERPRO: IPR001766 Fork head transcription factor | PR00053; 422-439 / GO:0003700 *transcription factor activity* GO:0006355 *regulation of transcription, DNA-dependent* GO:0005634 *nucleus*
INTERPRO: IPR001766 Fork head transcription factor | PR00053; 445-462 / GO:0003700 *transcription factor activity* GO:0006355 *regulation of transcription, DNA-dependent* GO:0005634 *nucleus*

Figure 2. Partial entry output showing several crosslinks and all the protein signatures found for this sequence.

Interested users may also download MySQL dumps of core TrSDB data obtained from TranScout, ProtLoc and TransMem runs against considered proteome sets of the current TrSDB version, as well as TranScout classification definitions.

DATA SOURCES AND TOOLS

TrSDB relies upon proteome sets, derived InterPro analyses and GOA assignments maintained by EBI facilities (11). Most data are stored and accessed through a MySQL 4 relational database management system. Analyses and interchanging files are generated and stored in a suitably parseable way in specially formatted text or XML format. Many scripts for carrying out analyses, linking components and handling the web interface use BioPerl tools and modules (12).

FUTURE DIRECTIONS

TrSDB will continue to incorporate new organism proteome sources, extend predictions from TranScout and ProtLoc programs, adopting deeper information integration from novel and existing support resources and adding other approaches in the transcription factor prediction context.

ACKNOWLEDGEMENTS

We acknowledge Xavier Serrats for website design questions and Oscar Conchillos for server administration issues. This work has been supported by grants BIO2001-264 and BI02001-02046 from the MCYT (Ministerio de Ciencia y Tecnología, Spain) and by Centre de Referencia en Biotecnologia (Generalitat de Catalunya). A.H. acknowledges a fellowship from Ministerio de Ciencia y Tecnología.

REFERENCES

1. Latchman,D.S. (2001) Transcription factors: bound to activate or repress. *Trends Biochem. Sci.*, **26**, 211–213.
2. Coulson,R.M., Enright,A.J. and Ouzounis,C.A. (2001) Transcription-associated protein families are primarily taxon-specific. *Bioinformatics*, **17**, 95–97.
3. Coulson,R.M. and Ouzounis,C.A. (2003) The phylogenetic diversity of eukaryotic transcription. *Nucleic Acids Res.*, **31**, 653–660.
4. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
5. Aguilar,D., Oliva,B., Aviles,F.X. and Querol,E. (2002) TranScout: prediction of gene expression regulatory proteins from their sequences. *Bioinformatics*, **18**, 597–607.
6. Cedano,J., Aloy,P., Pérez-Pons,J.A. and Querol,E. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.*, **266**, 594–600.
7. Aloy,P., Cedano,J., Oliva,B., Avilés,F.X. and Querol,E. (1997) 'TransMem': a neural network implemented in Excel spreadsheets for predicting transmembrane domains of proteins. *Comput. Appl. Biosci.*, **13**, 231–234.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
9. Camon,E., Magrane M., Barrell D., Binns D., Fleischmann W., Kersey,P., Mulder N., Oinn T., Maslen J., Cox A. *et al.* (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.*, **13**, 662–672.
10. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
11. Pruess,M., Fleischmann,W., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E., Mittard,V., Mulder,N., Phan,I., Servant,F. *et al.* (2003) The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes. *Nucleic Acids Res.*, **31**, 414–417.
12. Stajich,J.E, Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The BioPerl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.