

Research article

Open Access

## SPOC: A widely distributed domain associated with cancer, apoptosis and transcription

Luis Sánchez-Pulido\*<sup>1</sup>, Ana M Rojas<sup>1</sup>, Karel H van Wely<sup>2</sup>, Carlos Martínez-A<sup>2</sup> and Alfonso Valencia<sup>1</sup>

Address: <sup>1</sup>Protein Design Group, Centro Nacional de Biotecnología (CNB-CSIC). Cantoblanco, E-28049 Madrid, Spain and <sup>2</sup>Department of Immunology and Oncology, Centro Nacional de Biotecnología (CNB-CSIC). Cantoblanco, E-28049 Madrid, Spain

Email: Luis Sánchez-Pulido\* - [sanchez@cnb.uam.es](mailto:sanchez@cnb.uam.es); Ana M Rojas - [arojas@cnb.uam.es](mailto:arojas@cnb.uam.es); Karel H van Wely - [kvanwely@cnb.uam.es](mailto:kvanwely@cnb.uam.es); Carlos Martínez-A - [cmartineza@cnb.uam.es](mailto:cmartineza@cnb.uam.es); Alfonso Valencia - [valencia@cnb.uam.es](mailto:valencia@cnb.uam.es)

\* Corresponding author

Published: 07 July 2004

Received: 05 March 2004

BMC Bioinformatics 2004, 5:91 doi:10.1186/1471-2105-5-91

Accepted: 07 July 2004

This article is available from: <http://www.biomedcentral.com/1471-2105/5/91>

© 2004 Sánchez-Pulido et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

### Abstract

**Background:** The Split ends (Spen) family are large proteins characterised by N-terminal RNA recognition motifs (RRMs) and a conserved SPOC (Spen paralog and ortholog C-terminal) domain. The aim of this study is to characterize the family at the sequence level.

**Results:** We describe undetected members of the Spen family in other lineages (*Plasmodium* and Plants) and localise SPOC in a new domain context, in a family that is common to all eukaryotes using profile-based sequence searches and structural prediction methods.

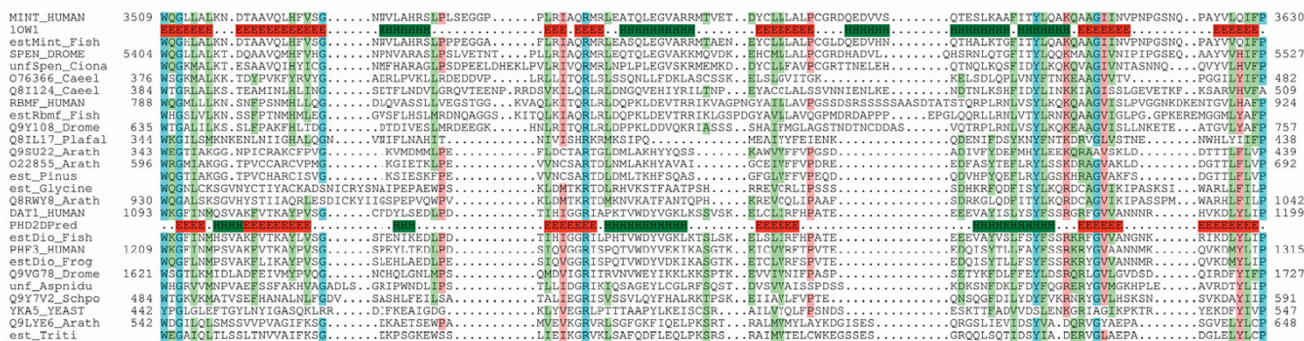
**Conclusions:** The widely distributed DIO (Death inducer-obliterator) family is related to cancer and apoptosis and offers new clues about SPOC domain functionality.

### Background

The aim of this study was to characterize the Spen family at the sequence level. The Spen family of proteins participates in various biological processes. It is involved in neuronal cell fate, survival and axonal guidance [1-3], cell cycle regulation [4], and repression of head identity in the embryonic trunk [5]. More recently it has been shown that the split ends gene participates in Wingless signalling in the eye, wing and leg imaginal discs [6] in the fly. The human Spen protein SHARP (SMRT/HDAC1-associated repressor protein) has been identified as a component of transcriptional repression complexes in both nuclear receptor and Notch/RBP-Jkappa signalling pathways [7-9]. Therefore, the Spen family of proteins appears to regulate transcription in several signalling pathways. In addition, the human Spen protein RBM15 (RNA-binding motif protein 15) is involved in the recurrent t(1;22)

translocation whose product is the RBM15-MKL1 fusion protein. This aberrant protein is related to the megakaryoblastic leukemia 1 (MKL1) [10,11].

On the other hand, DIO and its homologue PHF3 (PHD finger protein 3) are human proteins that contain a PHD (Plant Homeodomain) finger and a TFIIS (Transcription Factor S-II) domain: both domains are usually associated with transcription [12,13]. The DIO-1 protein regulates the early stages of cell death in mouse and humans [14,15]. Experimental evidence shows that the PHF3 protein is ubiquitously expressed in normal tissues, including brain. However, its expression is dramatically reduced or lost in glioblastoma, the most frequent tumour reported in human brain [12,13]. Although this family has been shown to be involved in apoptosis and cancer, the underlying molecular mechanisms are unclear.



**Figure 1**  
**Representative alignment of the SPOC domain.** The colouring scheme indicates average BLOSUM62 score (correlated to amino acid conservation) in each alignment column: cyan (greater than 3), light red (between 3 and 1.5) and light green (between 1.5 and 0.5). The limits of the domains are indicated by the residue positions on each side. The limits of proteins from partially sequenced genes whose full-length proteins are not available are not shown. X-Ray determined structure of the SPOC domain [9], pdb-code: IOW1 is shown below the SHARP sequence (Swissprot-ID:MINT\_HUMAN). PHD secondary structure prediction [29] for DIO family is shown below the DIO-I human sequence (Swissprot-Id: DAT1\_HUMAN), with E indicating a  $\beta$  strand (in red) and H an  $\alpha$  helix (in green). The asterisks below the alignment marks the conserved pair arginine-tyrosine mentioned in the text. The sequences are named with their swissprot or sptrembl identifications, and also, if necessary, with their common species name: Human, *Homo sapiens*; Frog, *Xenopus laevis*; Drome, *Drosophila melanogaster*; Caeel, *Caenorhabditis elegans*; Arath, *Arabidopsis thaliana*; Ciona, *Ciona intestinalis*; Yeast, *Saccharomyces cerevisiae*; Fish, *Brachydanio rerio*; Plafal, *Plasmodium falciparum*; Aspidu, *Aspergillus nidulans*; Pinus, *Pinus taeda*; Glycine, *Glycine max*; Schpo, *Schizosaccharomyces pombe*; Triti, *Triticum aestivum*. The "est" prefix identifies consensus sequences manually reconstructed by assembling highly similar expressed sequence tags from identical species (conceptual translations). The "unf" prefix identifies sequences obtained from Genome BLAST server at NCBI [24]. Complementary information is accessible at: <http://www.pdg.cnb.uam.es/SPOC>.

**Results**

Sequence profiles of the C-terminal conserved region of DIO family found the SPOC domain of Spen family at E-values of 0.083. Reciprocally, the profile of the SPOC domain of Spen detected the DIO family with an E-value of 0.05. We localised new members of the Spen and DIO families in different eukaryote lineages (Figures 1 and 2). Statistically significant E-values connected all the SPOC domain-containing families. None of these HMMer profile searches retrieved any new unrelated sequences and, as stated above, reciprocal searches produced convergent results.

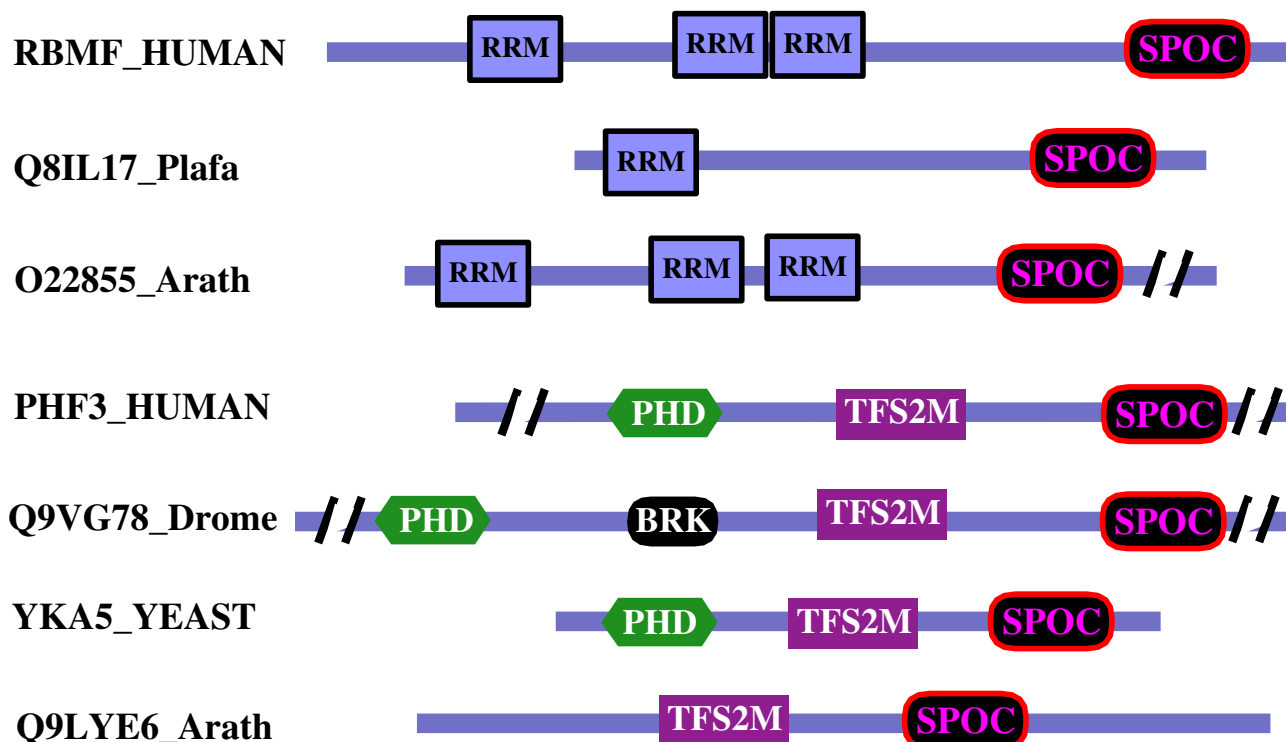
For the DIO family of sequences, secondary structure predictions were performed for the SPOC. These predictions showed good agreement with the crystal structure of the SPOC domain of SHARP [9] (Figure 1).

To investigate whether fold recognition analysis generated consistent results, we submitted the SPOC domain of DIO-1 (swissprot-id: DAT1\_HUMAN, residues 1093 to 1199) as a query to an independent fold assignment sys-

tem (see methods). The template IOW1 (the SPOC domain of SHARP protein) was found with a Z-score of -12.2 (estimated error rate <1%) despite its low sequence homology (16%).

Considering the E-values of the HMMer searches, the reliability of secondary structure predictions, and the fold assignment results, we are confident that the SPOC domain is present in the DIO family of proteins.

To highlight the degree of fold-conservation, we generated a structural model (Fig. 3B) of the SPOC domain of DIO. In the sequence alignment showed in figure 1, the C-terminal region of SPOC is missing. This region includes: two small helices (named E and F), which do not form part of the core and are not well conserved within the SHARP family, and the  $\beta$  sheet 7, which is part of the  $\beta$ -barrel core. The high sequence divergence in these region, made impossible to extend the alignment for automatic methods. However, for modelling purposes, the alignment was carefully extended to the C-terminal region and a beta sheet was detected in DIO, while the two helices



**Figure 2**  
**Schematic representation of the domain architecture and common features in representative members of the SPOC domain contained families.** The sequences are named with their swissprot or sptr embl identifications, and also, if necessary, with their common species name. The slashes represent inserts that are not shown. The proteins are drawn approximately to scale. The localization of the other domains PHD, TFIIS, BRK (BRM and KIS domain) and RRM is according to Pfam and SMART families databases [35,36].

were missing. Therefore, SPOC domain of DIO adopts a similar fold than the SPOC domain of SHARP and the seven strands  $\beta$ -barrel core is maintained (Figure 3).

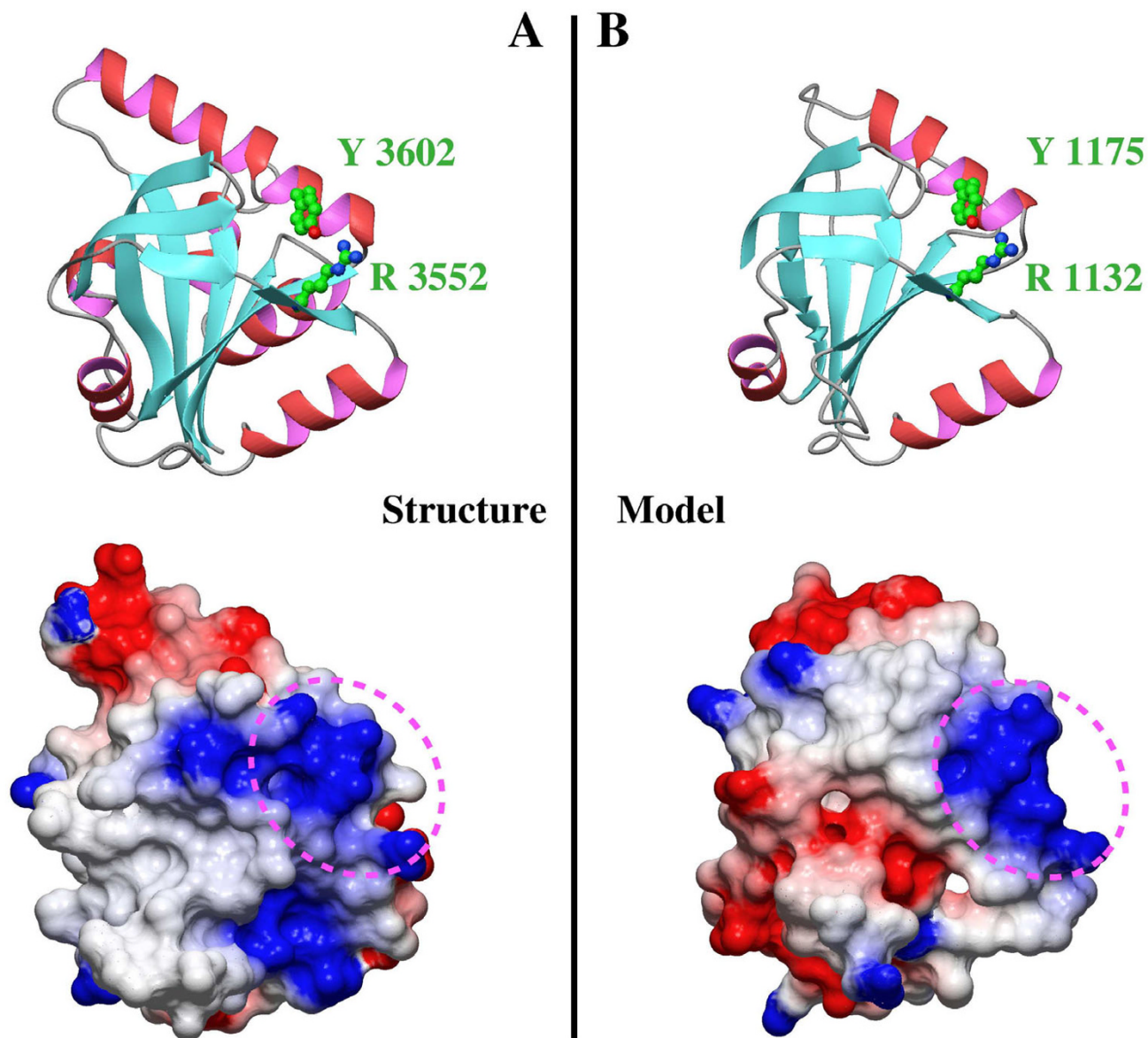
**Discussion**

Reported as a protein-protein interaction domain, the structure of the SPOC domain of SHARP contains a basic cluster essential for interaction with SMRT (silencing mediator for retinoid and thyroid receptors) the co-repressor [9]. The Arg 3552 of SHARP is localised in this basic cluster. Interestingly, this arginine is fully conserved in the SPOC alignment (Figure 1 and 3). The full conservation of Arg 3552 may be important, especially when equivalent substitutions in the protein-protein interaction interface (e.g. Lysine) might have little effect on the binding capability of this cluster. Therefore, one explanation for this fully conserved arginine could be its specific post-translational modification. Arginine methylation is a common post-translational modification in transcription regulation proteins that are catalysed by type I and II protein

arginine methyltransferases [16]. Arginine methylation modulates transcriptional activity and it has recently been related with a wide range of cellular processes [17].

For instance, proteins like the coactivator of nuclear receptors CBP (CREB-binding protein) follow this schema. In this protein there is a methylation site at Arg 600. This residue is essential for stabilising the structure of the domain implicated in CREB recruitment. There is a critical interaction between Arg 600 and Tyr 640 [18]. Disruption of this interaction by arginine methylation could lead to conformational changes [19]. Analogously, a similar interaction is observed on the surface of SPOC between the conserved residues Arg 3552 and Tyr 3602 (Figure 3) [9].

As arginine is one of the aminoacids most frequently found at the active sites of enzymes [20], alternative functional hypothesis for this conserved arginine is that it might form part of an active site of unknown function in the SPOC domain, or contribute to a specific catalytic



**Figure 3**

**Comparison of SPOC domains from SHARP and DIO proteins.** (A) Ribbon representation and electrostatic surface potential map of structure of the SPOC domain of SHARP protein (PDB code: 1OW1). (B) Ribbon representation and electrostatic surface potential map of homology model of the SPOC domain of DIO protein. Blue indicates positively charged regions, whereas red shows negatively charged regions. All the molecules are in the same orientation. Dotted circles indicates the conserved basic cluster, where is located the conserved arginine-tyrosine pair.

function in other proteins, analogous to the "Arginine Finger" in Ras-GAP proteins [21].

### Conclusions

The SPOC domain is present in different domain architectures among all the eukaryote lineages (Figures 1 and 2).

This study shows that, with the exponential growth of the sequence databases, sequence analysis sheds new light on biological function, even when structure is already available. The fact that this domain has been identified in cancer and apoptosis related proteins emphasises its importance in transcriptional regulation. Additional experimental

approaches using different members of the SPOC domain-containing families are required to confirm these hypotheses.

## Methods

### Sequence analyses

For the sequence analysis we related distant protein families via intermediate searches [22] using global hidden Markov model profiles (using *hmmsearch* of HMMer <http://hmm.wustl.edu/>) [23]. To improve the profile quality we followed two approaches: first, BLAST searches against unfinished genomes [24], and secondly, additional searches against EST (expressed sequence tags) databases [25]. This sequence enrichment improved the quality of the profile that was used to perform the searches against the non-redundant protein databases. We used *NAIL* to view and analyse the HMMer results [26]. The alignment was produced with HMMer [23] and *T-Coffee* software [27] using default parameters and was slightly refined manually. It is viewed with the *Belvu* program [28].

### Structural predictions and modeling

Secondary structure predictions were performed using *PHD* [29]. Fold recognition analyses were performed using the *FFAS* [30] server <http://ffas.ljcrf.edu/>. The model was based on the published crystal structure from the SPOC domain of *SHARP* protein [9] and obtained using *swiss-model* [31]. The model was evaluated using *PSQS* [32] and *WHATIF* [33] tools. Illustrations were generated with *MOLMOL* [34].

## Abbreviations

Spen, Split ends;

RRM, RNA Recognition motifs;

SPOC, Spen paralog and ortholog C-terminal;

DIO, Death inducer-obliterator;

SHARP, SMRT/HDAC1-associated repressor protein;

SMRT, silencing mediator for retinoid and thyroid receptors;

RBM15, RNA-binding motif protein 15;

MKL1, megakaryoblastic leukemia 1;

PHF3, PHD finger protein 3;

PHD, Plant Homeodomain;

TFIIS, Transcription Factor S-II;

CBP, CREB-binding protein;

GAP, GTPase Activating Protein;

HMM, hidden Markov model;

EST, expressed sequence tags;

BRK, BRM and KIS domain.

## Author's contribution

LSP and AR carried out the sequence and structural analysis of the domain.

KVW and CMA provided with the initial input of the research.

LSP, AR, KVW, CMA and AV authored the manuscript.

## Acknowledgements

We are grateful to M. Tress (CNB-Spain) and R. Rycroft for their helpful comments on the manuscript. This study was financed in part by the VI<sup>TH</sup> EU Framework Project QLGI-CT-2001-01536.

## References

- Chen F, Rebay I: **split ends, a new component of the *Drosophila* EGF receptor pathway, regulates development of midline glial cells.** *Curr Biol* 2000, **10**:943-946.
- Kuang B, Wu SC, Shin Y, Luo L, Kolodziej P: **split ends encodes large nuclear proteins that regulate neuronal cell fate and axon extension in the *Drosophila* embryo.** *Development* 2000, **127**:1517-1529.
- Rebay I, Chen F, Hsiao F, Kolodziej PA, Kuang BH, Laverty T, Suh C, Voas M, Williams A, Rubin GM: **A genetic screen for novel components of the Ras/Mitogen-activated protein kinase signaling pathway that interact with the yan gene of *Drosophila* identifies split ends, a new RNA recognition motif-containing protein.** *Genetics* 2000, **154**:695-712.
- Lane ME, Elend M, Heidmann D, Herr A, Marzodko S, Herzig A, Lehner CF: **A screen for modifiers of cyclin E function in *Drosophila melanogaster* identifies Cdk2 mutations, revealing the insignificance of putative phosphorylation sites in Cdk2.** *Genetics* 2000, **155**:233-244.
- Wiellette EL, Harding KV, Mace KA, Ronshaugen MR, Wang FY, McGinnis W: **spen encodes an RNP motif protein that interacts with Hox pathways to repress the development of head-like sclerites in the *Drosophila* trunk.** *Development* 1999, **126**:5373-5385.
- Lin HV, Doroquez DB, Cho S, Chen F, Rebay I, Cadigan KM: **Splits ends is a tissue/promoter specific regulator of Wingless signaling.** *Development* 2003, **130**:3125-3135.
- Shi Y, Downes M, Xie W, Kao HY, Ordentlich P, Tsai CC, Hon M, Evans RM: **Sharp, an inducible cofactor that integrates nuclear receptor repression and activation.** *Genes Dev* 2001, **15**:1140-1151.
- Oswald F, Kostezka U, Astrahantseff K, Bourteelle S, Dillinger K, Zechner U, Ludwig L, Wilda M, Hameister H, Knochel W, Liptay S, Schmid RM: **SHARP is a novel component of the Notch/RBP-Jkappa signalling pathway.** *EMBO J* 2002, **21**:5417-5426.
- Ariyoshi M, Schwabe JW: **A conserved structural motif reveals the essential transcriptional repression function of Spen proteins and their role in developmental signaling.** *Genes Dev* 2003, **17**:1909-1920.
- Mercher T, Coniat MB, Monni R, Mauchauffe M, Khac FN, Gressin L, Mugneret F, Leblanc T, Dastugue N, Berger R, Bernard OA: **Involvement of a human gene related to the *Drosophila* spen gene in the recurrent t(1;22) translocation of acute megakaryocytic leukemia.** *Proc Natl Acad Sci USA* 2001, **98**:5776-5779.

11. Ma Z, Morris SW, Valentine V, Li M, Herbrick JA, Cui X, Bouman D, Li Y, Mehta PK, Nizetic D, Kaneko Y, Chan GC, Chan LC, Squire J, Scherer SW, Hitzler JK: **Fusion of two novel genes, RBM15 and MKL1, in the t(1;22) (p13;q13) of acute megakaryoblastic leukemia.** *Nat Genet* 2001, **28**:220-221.
12. Fischer U, Struss AK, Hemmer D, Michel A, Henn W, Steudel WI, Meese E: **PHF3 expression is frequently reduced in glioma.** *Cytogenet Cell Genet* 2001, **94**:131-136.
13. Struss AK, Romeike BF, Munnia A, Nastainczyk W, Steudel WI, König J, Ohgaki H, Feiden W, Fischer U, Meese E: **PHF3-specific antibody responses in over 60% of patients with glioblastoma multiforme.** *Oncogene* 2001, **20**:4107-4114.
14. Garcia-Domingo D, Ramirez D, Gonzalez de Buitrago G, Martinez-A C: **Death inducer-obliterato I triggers apoptosis after nuclear translocation and caspase upregulation.** *Mol Cell Biol* 2003, **23**:3216-3225.
15. Garcia-Domingo D, Leonardo E, Grandien A, Martinez P, Albar JP, Izpisua-Belmonte JC, Martinez-A C: **DIO-I is a gene involved in onset of apoptosis in vitro, whose misexpression disrupts limb development.** *Proc Natl Acad Sci USA* 1999, **96**:7992-7997.
16. McBride AE, Silver PA: **State of the arg: protein methylation at arginine comes of age.** *Cell* 2001, **106**:5-8.
17. Boisvert FM, Cote J, Boulanger MC, Richard S: **A Proteomic Analysis of Arginine-methylated Protein Complexes.** *Mol Cell Proteomics* 2003, **2**:1319-1330.
18. Xu W, Chen H, Du K, Asahara H, Tini M, Emerson BM, Montminy M, Evans RM: **A transcriptional switch mediated by cofactor methylation.** *Science* 2001, **294**:2507-2511.
19. Wei Y, Horng JC, Vendel AC, Raleigh DP, Lumb KJ: **Contribution to stability and folding of a buried polar residue at the CARM1 methylation site of the KIX domain of CBP.** *Biochemistry* 2003, **42**:7044-7049.
20. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM: **Analysis of catalytic residues in enzyme active sites.** *J Mol Biol* 2002, **324**:105-121.
21. Ahmadian MR, Stege P, Scheffzek K, Wittlinghofer A: **Confirmation of the arginine-finger hypothesis for the GAP-stimulated GTP-hydrolysis reaction of Ras.** *Nat Struct Biol* 1997, **4**:686-689.
22. Park J, Teichmann SA, Hubbard T, Chothia C: **Intermediate sequences increase the detection of homology between sequences.** *J Mol Biol* 1997, **273**:349-354.
23. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
24. Cummings L, Riley L, Black L, Souvorov A, Resenchuk S, Dondoshansky I, Tatusova T: **Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes.** *FEMS Microbiol Lett* 2002, **216**:133-138.
25. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST - database for "expressed sequence tags".** *Nat Genet* 1993, **4**:332-333.
26. Sanchez-Pulido L, Yuan YP, Andrade MA, Bork P: **NAIL-Network Analysis Interface for Linking HMMER results.** *Bioinformatics* 2000, **16**:656-657.
27. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**:205-217.
28. **BELVU** [[http://www.sanger.ac.uk/Software/Pfam/help/belvu\\_setup.shtml](http://www.sanger.ac.uk/Software/Pfam/help/belvu_setup.shtml)]
29. Rost B: **PHD: predicting one-dimensional protein structure by profile-based neural networks.** *Methods Enzymol* 1996, **266**:525-539.
30. Rychlewski L, Jaroszewski L, Li W, Godzik A: **Comparison of sequence profiles. Strategies for structural predictions using sequence information.** *Protein Sci* 2000, **9**:232-241.
31. Schwede T, Kopp J, Guex N, Peitsch MC: **SWISS-MODEL: an automated protein homology-modeling server.** *Nucleic Acids Res* 2003, **31**:3381-3385.
32. **PSQS** [<http://www.ljcs.org/psqs/>]
33. **WHATIF** [<http://www.cmbi.kun.nl/gv/servers/WWWVWI/>]
34. Koradi R, Billeter M, Wüthrich K: **MOLMOL: a program for display and analysis of macromolecular structures.** *J Mol Graphics* 1996, **14**:51-55.
35. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32 Database issue**:D138-141.
36. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration.** *Nucleic Acids Res* 2004, **32 Database issue**:D142-144.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

