BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 14 2013, pages 1726–1733
doi:10.1093/bioinformatics/btt271

Sequence analysis

Advance Access publication May 14, 2013

Identification of hidden relationships from the coupling of Hydrophobic Cluster Analysis and Domain Architecture information

Guilhem Faure and Isabelle Callebaut*

IMPMC, UMR7590, CNRS, Université Pierre et Marie Curie-Paris6, Paris Cedex 05, France

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Describing domain architecture is a critical step in the functional characterization of proteins. However, some orphan domains do not match any profile stored in dedicated domain databases and are thereby difficult to analyze.

Results: We present here an original novel approach, called TREMOLO-HCA, for the analysis of orphan domain sequences and inspired from our experience in the use of Hydrophobic Cluster Analysis (HCA). Hidden relationships between protein sequences can be more easily identified from the PSI-BLAST results, using information on domain architecture, HCA plots and the conservation degree of amino acids that may participate in the protein core. This can lead to reveal remote relationships with known families of domains, as illustrated here with the identification of a hidden Tudor tandem in the human BAHCC1 protein and a hidden ET domain in the *Saccharomyces cerevisiae* Taf14p and human AF9 proteins. The results obtained in such a way are consistent with those provided by HHPRED, based on pairwise comparisons of HHMs. Our approach can, however, be applied even in absence of domain profiles or known 3D structures for the identification of novel families of domains. It can also be used in a reverse way for refining domain profiles, by starting from known protein domain families and identifying highly divergent members, hitherto considered as orphan.

Availability: We provide a possible integration of this approach in an open TREMOLO-HCA package, which is fully implemented in python v2.7 and is available on request. Instructions are available at http://www.impmc.upmc.fr/~callebau/tremolohca.html.

Contact: isabelle.callebaut@impmc.upmc.fr

Supplementary information: Supplementary Data are available at *Bioinformatics* online.

Received on March 15, 2013; revised on May 2, 2013; accepted on May 7, 2013

1 INTRODUCTION

Orphan domains are segments of proteins forming autonomous folding units that cannot be assigned to a known domain family, as stored in dedicated domain databases (Ekman *et al.*, 2005). They may be included in large unassigned regions, which make up at least 10% of the residues in a typical proteome (Ekman *et al.*, 2005). These orphan domains have either evolved too far from the nearest neighbors to be assigned to a domain, or they have been created by some *de novo* mechanisms. Most of the solved 3D structures of orphan domains, however, show structural similarity to already known protein domains, suggesting that the fraction of orphan domains that have distant homologs is high (Siew and Fischer, 2004). This is consistent with earlier theoretical studies, which have suggested that protein domains fall into a limited number of protein folds and families (Wolf *et al.*, 2000). The distant homolog theory is also supported by the identification of remote homologs to orphan domains through sequences from environmental sequencing projects, although these metagenomics studies also revealed many novel orphans (Rusch *et al.*, 2007). A lot of specific analyses led to link orphan domains to already known families or to identify new families of domains (Aravind and Koonin, 1999). Among these studies are those we performed using the fold signatures defined through the Hydrophobic Cluster Analysis (HCA) approach [e.g. (Callebaut and Mornon, 1997a, b; Callebaut *et al.*, 1999, 2002, 2005, 2006) for some examples]. HCA is based on a bidimensional representation of the sequence, in which hydrophobic amino acids congregate into clusters (Callebaut *et al.*, 1997; Gaboriaud *et al.*, 1987), which are statistically centered on regular secondary structures (Hennetin *et al.*, 2003; Woodcock *et al.*, 1992). Hydrophobic clusters associated with core secondary structures are stable relative to evolution, offering a way to efficiently compare sequences at very low levels of sequence identity (see Supplementary Data S1 for a detailed description of the method and its practical use).

Here, we wished to develop a methodology inspired from our experience in deciphering orphan domains, which facilitates the detection of distant relationships with already known families of domains by automatic inference from the PSI-BLAST results. This methodology, called TREMOLO-HCA (after Travel into REmote HOmoLOgy with HCA), allows an easier exploitation of results from sequence similarity search programs (e.g. PSI-BLAST), by providing contextual (domain architecture) and structural (conserved hydrophobic core) information. It furnishes for each sequence aligned with the query the domain architecture of the whole protein, as well as indicates the conservation degree of amino acids that may participate in the protein fold. The interest of the methodology is illustrated with two examples of conserved domain database (CDD)-orphan domains, which were first delineated from the whole-protein sequence by using our experience in HCA coupled with a home-made program called SEG-HCA (our unpublished data), developed for

*To whom correspondence should be addressed.

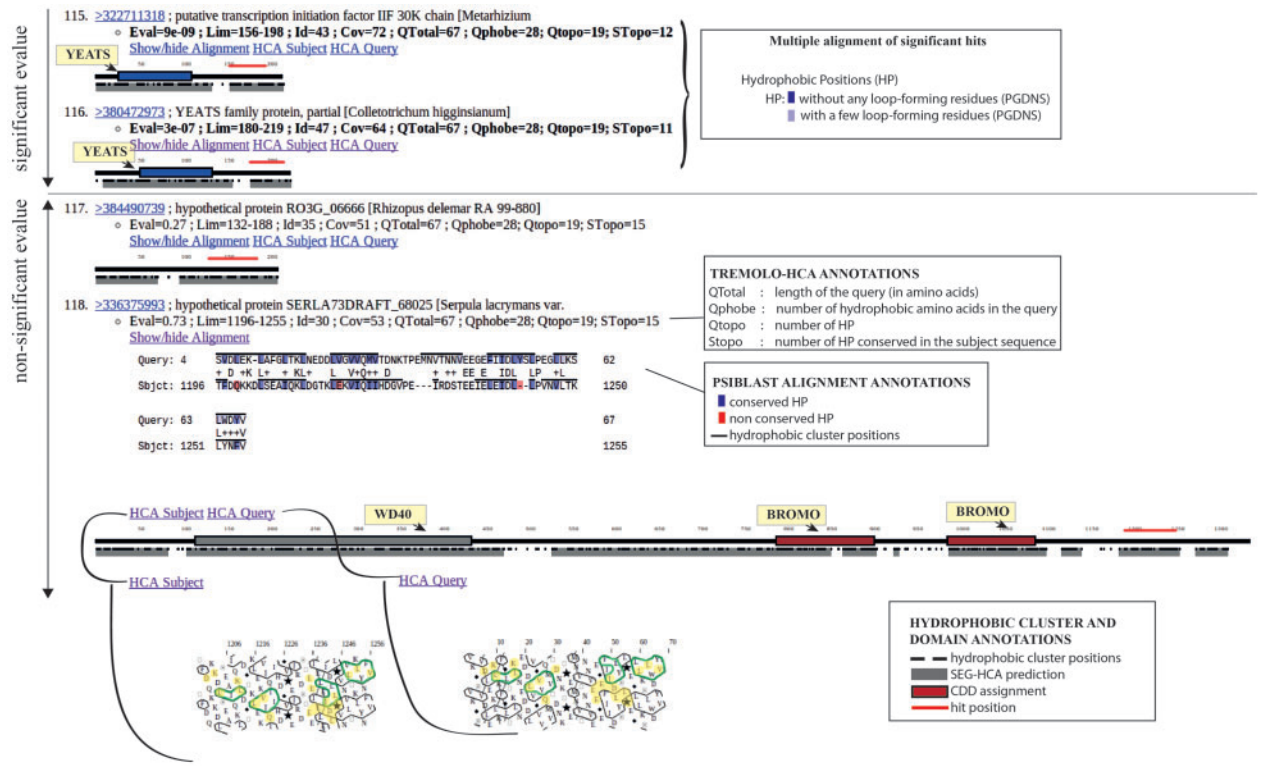

Fig. 1. Example of a TREMOLO-HCA output. The sequences of the whole proteins with which the query is aligned are shown as lines and domains as boxes [colored when assigned from the CDD and gray when predicted by SEG-HCA, an automatic procedure for predicting foldable segments which have high density in hydrophobic clusters (our unpublished data)]. The positions of the segments aligned with the query sequence are added to this schematic representation as red lines. The corresponding 1D alignments, as given in the PSI-BLAST results, are also provided. For these 1D alignments, conserved hydrophobic positions (positions that are always occupied by strong hydrophobic amino acids in significant pairwise alignments) are reported in blue on the query sequence. Their numbers are indicated and should correspond to one half of the total number of hydrophobic amino acids of the query sequences for an accurate prediction of the hydrophobic core of globular domains. The corresponding positions are colored in the sequence aligned with the query sequence: (i) also in blue if occupied by a strong hydrophobic amino acid and (ii) in red if not. This coloring scheme thus allows the quick evaluation of the conservation of predicted core hydrophobic positions. Beside the domain architecture of the aligned sequence and the pairwise sequence alignment, the HCA plot of the aligned segment is also made available, to allow a quick evaluation of the hydrophobic cluster compatibility. For further analysis outside the 1D alignment limits, the user can get the HCA plots of the whole-protein sequence through the HCA plot web server (http://bioserv.impmc.upmc.fr/hca-form.html)

the automatic HCA-based delineation of globular domains from the analysis of single sequences.

2 METHODS

The TREMOLO-HCA procedure adds to the PSI-BLAST results information: (i) about the domain architecture of proteins with which the query sequence is aligned and (ii) about the conservation rate of the hydrophobic amino acids, which are conserved in the considered family of proteins and likely participate in its hydrophobic core (Fig. 1).

The implemented procedure uses a standard PSI-BLAST output. In the application examples presented here, the sequences used as queries in PSI-BLAST were first delineated using HCA (Supplementary Data S1). These are rich in hydrophobic clusters, thus constituting potential globular domains, but they did not mach any CDD profiles and are thus considered as CDD-orphans. The non-redundant (nr) database at National Centre of Biological Information (NCBI) was used as reference database, and PSI-BLAST was run with default parameters until convergence is obtained (or after eight iterations if convergence is not reached at this stage). However, other settings of PSI-BLAST parameters and intermediate PSI-BLAST results during the iterative process can be considered by the user.

Significant (E-value $<5 \times 10^{-3}$) and non-significant (E-value $>5 \times 10^{-3}$) results are both considered. The whole sequence of each protein aligned with the query is then used for identifying already known domains through the RPS-BLAST program, run on the widely used NCBI's CDD, which includes NCBI-curated domains and domain models from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAM) (Marchler-Bauer *et al.*, 2013). Information on identified conserved footprints is summarized schematically (Fig. 1), allowing the localization of the hit within the context of the full domain architecture of the protein and thereby helping the interpretation of the PSI-BLAST results. The redundancy can be reduced at the level of domain architecture and/or of sequence identity.

To provide criteria for assessing the reliability of alignments at low levels of sequence identity (thus typically below the threshold value), we provide access to the HCA plots of the two aligned sequences (Fig. 1), allowing the evaluation of hydrophobic cluster compatibilities. A guideline to the use of HCA is available in Supplementary Data S1. We also aim, when possible, at highlighting core positions, i.e. positions participating in the hydrophobic core of globular domains (Fig. 1). These

positions constitute approximately one half of the total number of strong hydrophobic amino acids (Eudes *et al.*, 2007). They can be evaluated with accuracy from the significant pairwise alignments (E-value $<5 \times 10^{-3}$), if they are deduced from a set of a sufficient number of sufficiently distant sequences (typically at least 10 non-redundant sequences, which do not share >40% pairwise identity). We thus identify, within the significant pairwise alignments (E-value $<5 \times 10^{-3}$) from which redundancy can be reduced (by default at 70% sequence identity), those positions that are mainly occupied by strong hydrophobic amino acids [at least 75% of hydrophobic amino acids (V, I, L, F, M, Y and W), without loop-forming residues (P, G, D, N and S) (Poupon and Mornon, 1998)]. The number of these conserved hydrophobic positions is given as a reference, allowing an estimation of the quality of the prediction. Ideally, this should represent 50% of the total number of hydrophobic amino acids of the query sequence. We provide a possible implementation of the TREMOLO-HCA procedure into a python software. This latter can be customized and modified by users. Package is available on request, and explanations are available at http://www.impmc.upmc.fr/~callebau/tremolohca.html. This software was used to investigate the two examples presented later in the text.

3 RESULTS AND DISCUSSION

We analyzed globular-like domains delineated through HCA, which are not assigned to an already known domain of the CDD. Such segments cover all fold classes (α, β, α and β) and have generally relatively small lengths, although large domains (up to 100 amino acids) can also be found. We focused our interest on CDD-orphan domains found within some particular protein families that we have identified previously and/or play key roles in the DNA damage response (DDR) and/or in epigenetic regulation. TREMOLO-HCA analyses showed that some of these predicted CDD-orphan domains can be linked to already known families of domains, by direct inference from the PSI-BLAST significant results. This is exemplified later in the text, for an orphan domain found in the BAHCC1/TNRC18 proteins, which can be linked to the Tudor family. Another interesting case is an OB-fold, which is detected in the human Tudor domain protein TDRD3. This OB-fold can also be linked to the human and yeast RMI1 proteins (Yin *et al.*, 2005), which are subunits of the RecQ (Sgs1p)—Top III (Top3p) complex and are involved in the processing of homologous recombination intermediates (Supplementary Data S2). The other cases, which can not be predicted by direct inference from the PSI-BLAST significant results, may either constitute new domains or be linked at very high level of divergence to already known domains and require for their characterization a sensitive analysis of the PSI-BLAST background noise (non-significant E-values). This last analysis can also be helped by the consideration of the hydrophobic cluster conservation, as well as by the knowledge of the architecture of the aligned proteins. Such an analysis is also illustrated later in the text, for an orphan domain found in several members of the YEATS family.

3.1 A tandem of Tudor domains in the BAHCC1/TNRC18 proteins

The BAH (after bromo adjacent homology) family of domains includes diverse DNA- and chromatin-associated proteins such as the CpG–DNA methylase DNMT1 and the replication origin complex subunit 1 ORC1 (Callebaut *et al.*, 1999). The BAH domain of the silent information regulator Sir3p, a key silencing protein in *Saccharomyces cerevisiae*, which has evolved from Orc1 by gene duplication (Hickman and Rusche, 2010), interacts with multiple surfaces of the nucleosome, suggesting a possible involvement in nucleosome compaction that would be disrupted by post-translational modifications on histones (Armache *et al.*, 2011). In contrast, the BAH domain of mouse ORC1 interacts with the histone H4 dimethylated at lysine 20 (Kuo *et al.*, 2012), through an aromatic dimethyl-lysine–binding cage. In *S.cerevisiae*, the interaction of the Sir1 protein with the BAH domain of Orc1 plays a key role in the establishment of a silent chromatin structure at the cryptic mating-type loci HMR and HML (Hou *et al.*, 2005).

A segment predicted by SEG-HCA, which stay orphan when CDD is searched, is found in the human BAHCC1 (after BAH and coiled-coil domain-containing protein 1), a large protein (2608 amino acids), including at its C-terminal end a BAH domain (amino acids 2482–2602) and whose function remains poorly understood. This 'orphan' segment, including hydrophobic clusters typical of regular secondary structures, is found between amino acids 1868 and 2028. Using this fragment as query in a TREMOLO-HCA analysis, we found significant similarities with the orphan region of the related trinucleotide repeat-containing gene 18 protein TNRC18, which also possesses a C-terminal BAH domain. Significant similarities are also observed with an as yet uncharacterized protein c11orf16 and with long isoform homologs of the *Drosophila* capicua protein, an HMG-box containing transcriptional repressor, which is involved in cancer and neurodegeneration (Bettegowda *et al.*, 2011; Jiménez *et al.*, 2012).

A relationship to Tudor domains, as well as the obvious duplication of the Tudor domain (tandem) in the query sequence, can significantly be deduced from similarities with Tudor domains of several hypothetical proteins, as well as with the tandem of Tudor domains found in the histone-lysine-*N*-methyltransferase SETDB1, in the histone lysine-specific demethylase 4 (KDM4, also known as JMJD2) and in PHF20, albeit sequence identities are very low (<15%) (Fig. 2). The Tudor domain is a small four or five stranded β-barrel fold, possessing the ability to interact with methylated partners (Adams-Cioaba and Min, 2009; Taverna *et al.*, 2007; Yap and Zhou, 2010). Tandem Tudor domains have been characterized in several proteins, among which 53BP1 and the JMJD2A histone demethylase. The individual Tudor domains can be organized as two independent domains (as in 53BP1, Crb2, SETDB1 and PHF20) or as inter-digitated domains (as in the JMJD2/KMD4), with each lobe of a saddle-shaped structure resembling isolated Tudor domains. These last domains, formed by the exchange of strands β3 and β4, are called hybrid Tudor domains.

In either of these configurations (standard or hybrid), the methyl-binding cages include two highly conserved aromatic amino acids in strands β1 (a tryptophan) and β2 (a tyrosine). The site is completed by a phenylalanine and an aspartic acid, located at the end of strand β3 (FxD motif) from the same or from the other Tudor motif, depending on the double Tudor motifs forming a standard or hybrid (inter-digitated) structure, respectively. Examination of the alignment (Fig. 2) suggests that the Tudor tandem domains of BAHCC1/TNRC18 would adopt a standard, non-inter-digitated configuration, as all the elements

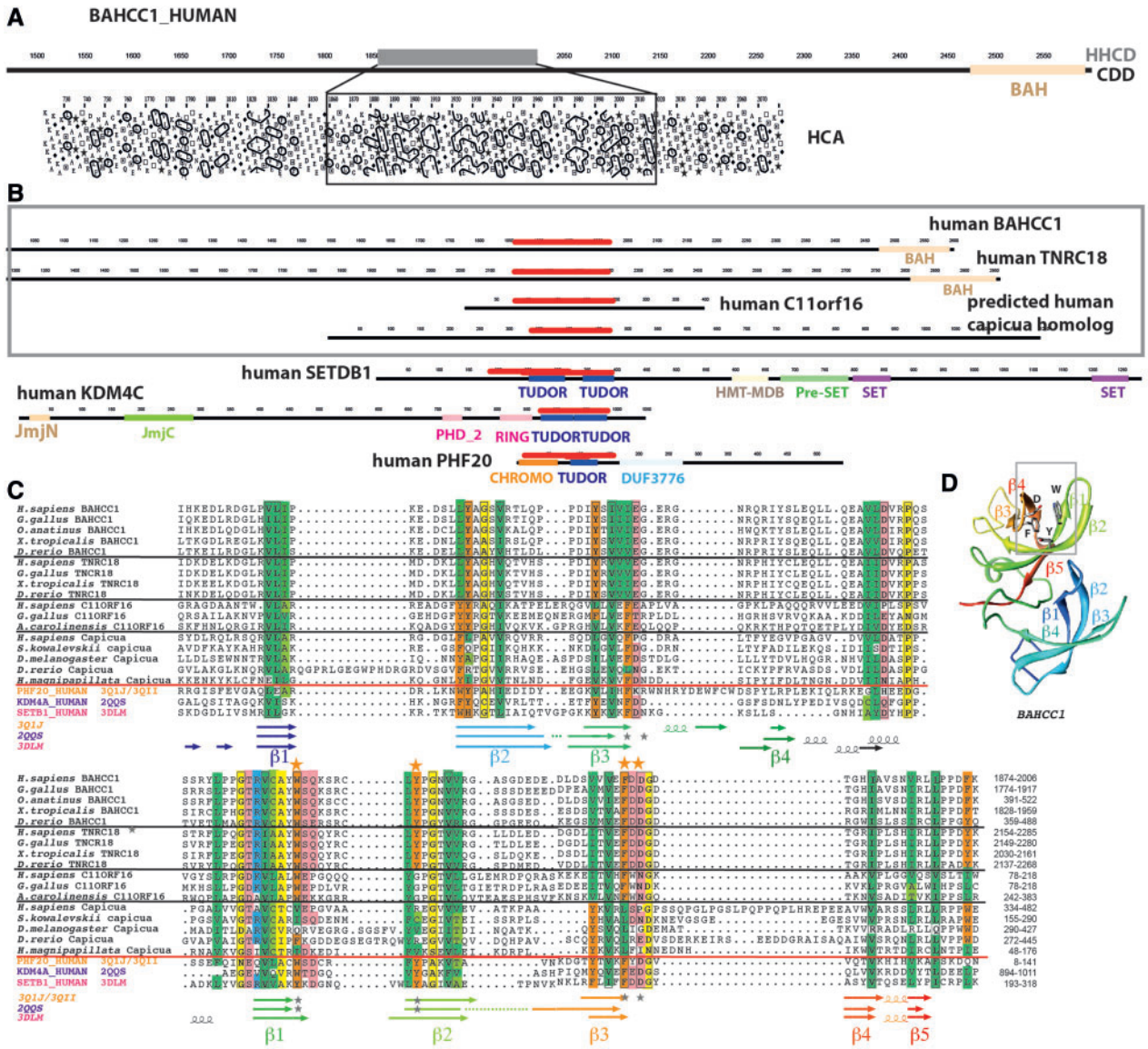

Fig. 2. A hidden double Tudor domain in BAHCC1/TNRC18, C11ORF16 and capicua proteins. (**A**) The predicted orphan segment of human BAHCC1 was used as query in PSI-BLAST searches and analyzed through the TREMOLO-HCA procedure. (**B**) Significant similarities were found with orphan domains from the TNRC18, C11ORF16 and capicua proteins, as well as with double tudor domains from several proteins (e.g. human SETDB1, human KDM4A and human PHF20). (**C**) Multiple alignment of the Tudor tandem domains of the BAHCC1/TNRC18, C11ORF16 and Capicua proteins with those of human PHF20, KDM4A and SETDB1. The observed secondary structures are indicated below the alignment, together with the pdb identifiers. Conserved hydrophobic amino acids are highlighted in green, whereas other colors are used for other similarities [orange: aromatic, pink: acidic, blue: basic, yellow: small (A, G and V) or loop-forming (P, G, D, N and S) amino-acids]. Some amino acids (A, C and T), which may integrate the hydrophobic alphabet following the context, are indicated in light green. The double Tudor domains of KDM4A adopts an interdigitated structure, with strands $\beta 2$ and $\beta 3$ forming a long strand and allowing the exchange of strands $\beta 3$ and $\beta 4$ between the two tudor-like lobes. Amino acids identified as participating in the methyl-binding pockets in the different structures are indicated with gray stars, with our prediction for BAHCC1 indicated with yellow stars, above the alignment. Genbank identifiers (gi): BAHCC1: 205371795 (*Homo sapiens*), 363740891 (*Gallus gallus*), 345318278 (*Ornithorhynchus anatinus*), 301621606 (*Xenopus tropicalis*), 326666283 (*Danio rerio*); TNRC18: 187608897 (*H.sapiens*), 363739553 (*G.gallus*), 301603727 (*X.tropicalis*), 326666134 (*D.rerio*); C11ORF16: 20381227 (*H.sapiens*), 118091239 (*G.gallus*), 327272286 (*Anolis carolinensis*); Capicua: 341915867 (*H.sapiens*), 291242109 (*Saccoglossus kowalevskii*), 386766127 (*Drosophila melanogaster*), 326676456 (*D.rerio*), 221132917 (*Hydra magnipapillata*); PHF20: 32699605 (*H.sapiens*); KDM4A: 308153453 (*H.sapiens*), SETB1: 25091210 (*H.sapiens*). (**D**) Model of the 3D structure of the BAHCC1 double Tudor tandem, made on the basis of the human SETDB1 structure (pdb 3DLM) and on which is highlighted the putative methyl-binding pocket (box), which is highly similar to that found in the second domain of PHF20 (pdb 3Q11)

of the methyl-binding signature (W in strand β1, Y in strand β2 and FxD in strand β3) are found within the second Tudor motif. A methyl-binding site in the second Tudor domain of the BAHCC1/TNRC18 Tudor tandem is thus likely, as also observed in PHF20, where second Tudor domain has been shown to interact with dimethylated lysines from p53 peptides (Adams-Cioaba *et al.*, 2012; Cui *et al.*, 2012). The prediction is less obvious for the Tudor tandem of C11ORF16 and capicua, for which no obvious standard methyl-binding sites could be identified.

The discovery of Tudor tandems in the BAHCC1/TNRC18 proteins, together with a BAH domain, underlines their potential importance in combinatorial post-translational modification readout at the histone or nucleosomal level.

3.2 The C-terminal end of the yeast Taf14p transcription factor belongs to the ET family of domains

Yeast transcription factor Taf14p belongs to the YEATS family of proteins, whose members are found in several chromatin-modifying and transcriptional complexes (Schulze *et al.*, 2010).

Taf14p, also known as Anc1, is associated with several chromatin remodeling and transcription complexes (Swi/Snf, RSC, INO80 and NuA3). Taf14p is also a subunit of the transcription factor TFIID and TFIIF in *S.cerevisiae* [reviewed in Zhang *et al.* (2011)]. The function of YEAST domains, named after some of the proteins containing it (i.e. Yaf9, ENL, AF9, Taf14p and Sas5), remains poorly understood (Schulze *et al.*, 2010). It is located at the N-terminal extremity of the Taf14p protein, a shared property of YEAST domains within the YEAST family. An orphan segment is identified at the C-terminal end of the protein, clearly separated from the YEAST domain by a hinge region, lacking hydrophobic amino acids.

Using this C-terminal domain of yeast Taf14p as a query (amino acids 173–239), we identified significant similarities at PSI-BLAST convergence by iteration 3 with the Taf14p orthologs from various fungi species, possessing similar domain architectures, as revealed in the TREMOLO-HCA output. Marginal similarities (E-value 318) were, however, also readily observed with the C-terminal domain of the yeast Sas5p (*something about silencing 5*) protein, another member of the YEATS family (Fig. 3). Examination of the corresponding alignment indicated that the positions in which hydrophobicity is conserved in the Taf14p orthologs are mostly also occupied by hydrophobic amino acids in the yeast Sas5p sequence (gray squares in Fig. 3). Consistently, the Taf14p/Sas5p relationship is supported by comparison of the HCA plots, indicating a good conservation of the hydrophobic clusters, associated with the regular secondary structures (Fig. 3). No other protein outside yeast Taf14p and Sas5p were, however, highlighted with PSI-BLAST significant E-values, when the sequences of Sas5p were included in the PSSM and PSI-BLAST run again to convergence.

Interestingly, in the same PSI-BLAST output and just below the threshold value ($E > 0.001$), marginal similarities were also observed with the C-terminal regions of proteins possessing multiple BROMO domains and belonging to the BET family of proteins (Florence and Faller, 2001) (e.g. *Stereum hirsutum* hypothetical protein, shown in Figure 3A and BDF1, BDF2, BRD2, BRD3 and BRD4 from human, as well as GTE6 from *Arabidopsis thaliana* and FSH from *D. melanogaster*, in Fig. 3D). These C-terminal regions correspond to ET (after extra-terminal) domains, which are specific of proteins of this BET family. Information about ET domains is, however, not yet stored in domain databases, even though the experimental 3D structure of the mouse BRD4 ET domain has been solved (Lin *et al.*, 2008). Examination of the HCA plots indicated similarities between hydrophobic clusters, and the Taf14p/ET sequence alignment showed that the positions for which hydrophobicity is conserved correspond to the conserved hydrophobic positions of the Taf14p/Sas5p alignment. These observations support the hypothesis of an ET fold for the yeast Taf14p/Sas5p C-terminal domains. The conserved hydrophobic positions deduced for the Taf14p/Sas5p alignment well correspond to amino acids that are buried within the 3D structure of the BRD4 ET domain (Fig. 3).

Finally, we also examined the C-terminal domains of all members of the YEATS family, and in addition of *S. cerevisiae* Sas5p, we identified the ET signature within the highly conserved C-terminal domains of two homologous human proteins*:* AF9 (ALL1-fused gene from chromosome 9) and ENL (Eleven–Nineteen Leukemia), which are common Mixed Lineage Leukemia (MLL) fusion partners (Fig. 3). In contrast, the C-terminal coiled-coil domain of human GAS41 and *S. cerevisiae* YAF9, also containing an N-terminal YEATS domain, did not match the enlarged ET domain profile. This observation does not support a previous analysis (Le Masson *et al.*, 2003), reporting a potential relationship between the GAS41/YAF9 and AF9 C-terminal domains. The relationship of yeast Taf14p, yeast Sas5p and human AF9/ENL to ET domains was finally further supported by using profile–profile comparison methods [HHPRED (Söding *et al.*, 2005) probabilities to share significant similarity with the ET domain of mouse BDRD4 (pdb 2jns): 97.5, 97.8 and 98.2, respectively]. Interestingly, although this manuscript was submitted for publication, the experimental 3D structure of the ET-like domain of human AF9 (named AHD after ANC1 homology domain) was solved in complex with an AF4 peptide. However, the authors did not report the obvious structural relatedness with ET domains, which remained thus uncovered (Leach *et al.*, 2013). The two structures are indeed similar (Fig. 3C), confirming our analysis. Interestingly, although the ET domain of BDR4 forms an autonomous stable structure, the ET-like/AHD domain of AF9 only folds on AF4 binding.

The relationship highlighted here between ET domains from the BET family and the C-terminal domains of some members of the YEATS family opens new perspectives for the characterization of this domain in both families of proteins. It is thus likely that the ET domains play a key role as a protein–protein interaction module, involved in transcriptional and chromatin regulation. This is supported in particular by the fact that the ET-like domains of the related proteins, AF9 and ENL, have been shown to be sufficient for transformation (Slany *et al.*, 1998) and interact with several proteins [for review (Muntean and Hess, 2012)]. Indeed, the ET-like domain of these proteins is responsible for the interaction with the scaffolding protein AF4 (Zeisig *et al.*, 2005), assembling the different components of Super Elongation Complexes (SEC) (He *et al.*, 2011; Luo *et al.*, 2012). It is also directly responsible for the binding of another MLL fusion

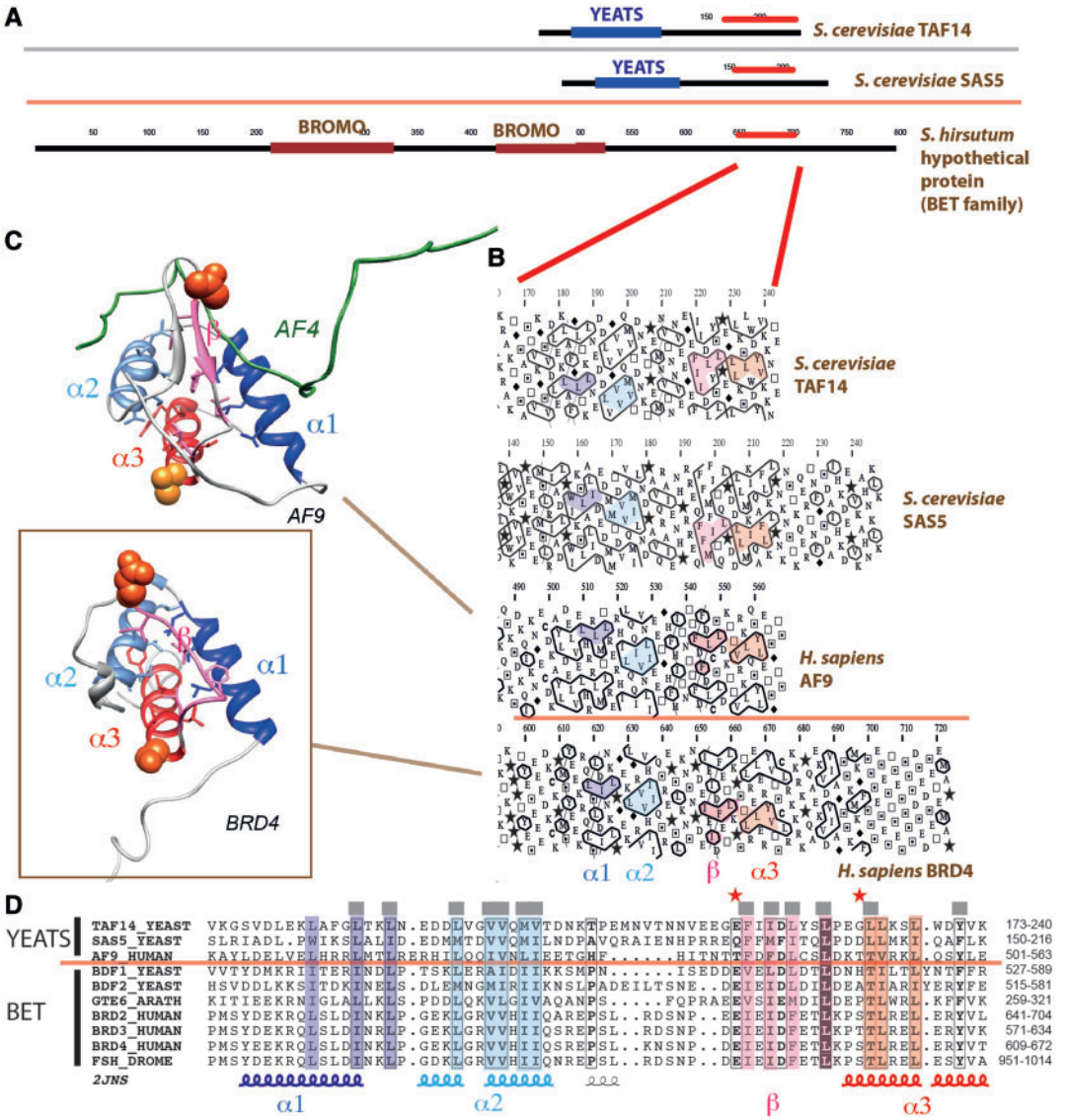

Fig. 3. A hidden ET domain in the C-terminal end of proteins of the YEATS family. (**A**) The predicted orphan segment of yeast Taf14p was used as a query in PSI-BLAST searches and analyzed through the TREMOLO-HCA procedure. Marginal similarities found with proteins of the YEATS family (Sas5p/AF9) were assessed through the comparison of HCA plots, highlighting compatibility of hydrophobic clusters [colored in panel (**B**)]. Other marginal similarities observed with ET domains from proteins of the BET family were also supported by the comparison of the HCA plots (B). (**C**) Bottom: experimental 3D structure of the ET domain of mouse BRD4 (pdb 2jns), in which hydrophobic amino acids that are conserved between the C-terminal domains of some proteins of the YEATS family and ET domains are highlighted in colors, as on the HCA plots (B). These participate in the hydrophobic core. The recently published AF9 ADH domain 3D structure (top: pdb 2lm0) confirms the similarity with the ET domain fold. Red spheres represent amino acids of human AF9, whose substitution in alanine disrupts the interaction with CBX8 (Tan *et al.*, 2011) [stars in panel (**D**)]. (D) Multiple alignment of ET-like domains of proteins of the YEATS family (yeast Taf14p and Sas5p, human AF9/ENL) and ET domains of the BET family. Conserved hydrophobic amino acids are highlighted with colored boxed according to the regular secondary structures in which they are included [gray squares indicate those which are conserved in the PSI-BLAST significant pairwise alignments (HP in Fig. 1)]. These, as deduced from the experimental structure of mouse BRD4 [panel (D)], are reported below the alignment. UniProt accession numbers: P35189 (Taf14p_YEAST), Q99314 (Sas5p_YEAST), P42568 (AF9_HUMAN), P35817 (BDF1_YEAST), Q07442 (BDF2_YEAST), Q9FT54 (GTE6_ARATH), P25440 (BRD2_HUMAN), Q15059 (BRD3_HUMAN), O60885 (BRD4_HUMAN), P13709 (FSH_DROME)

partner, ABI-1 (García-Cuéllar *et al.*, 2000) and of the chromobox homolog 8 (CBX8, also known as human Polycomb 3), which facilitates the transcriptional activation of MLL-AF9 target genes (García-Cuéllar *et al.*, 2001; Tan *et al.*, 2011). Finally, it is worth noting that MLL fusions, including MLL-AF9, were recently shown to be associated with the BET family of chromatin adaptor proteins within Super Elongation Complexes, and that a small inhibitor of these has profound

efficacy against MLL-fusion leukemic cell lines (Dawson *et al.*, 2011). Whether the ET domains of both families of proteins are involved in the binding of common SEC partners deserve further investigations.

4 CONCLUSION

The methodology developed here for easier domain inference from the PSI-BLAST results provides contextual (domain architecture) and structural (conserved hydrophobic core) information. Protein domain architectures can indeed give useful information about the functional context in which the query sequence can be present. The remote relationships of orphan sequences to already known families of domains revealed here were also highlighted by the sensitive HHPRED program, based on pairwise comparisons of HHMs. Our approach can, however, be applied even in absence of CDD profiles or known 3D structures for identifying novel families of domains. It can also be used in a reverse way for refining CDD profiles, by starting from known protein domain families and identifying highly divergent members, hitherto considered as orphan. Hence, starting with the TUDOR profile or with the sequence of the BDRD4 ET domain, for which the experimental 3D structure is known, we were able to identify the remote relationships to human BAHCC1 and yeast Taf14p. Revisiting systematically the CDD profiles using such an approach could thus lead to significant enhancement of their sensitivity for detecting remote relationships.

Information about positions of the alignment for which hydrophobicity is conserved in homologous sequences is especially useful to get insight into the likelihood of structural relationships for alignments that are reported with non-significant E-values in the PSI-BLAST results, which have generally to be further supported by other information. The conservation of core secondary structures over the whole domain (and not limited to one or a few ones), as assessed by considering these positions and comparing the corresponding hydrophobic clusters, are thus an important factor for assessing the relevance of marginal similarities. In this context, HCA can be rewarding for detecting similarities outside the first limits detected by PSI-BLAST, independently of the presence of variable indels that are generally difficult to handle by alignment procedures (see Supplementary Data S1 and S2 for application examples).

Of course, the methodology described here and applied to the analysis of PSI-BLAST results may also be adapted to process results coming from other methods for remote homology detection, such as HHpred (Söding *et al.*, 2005), HHblits (Remmert *et al.*, 2011) or HMMER (Finn *et al.*, 2011).

ACKNOWLEDGEMENTS

The authors thank Dr Jean-Paul Mornon for insightful developments associated with Hydrophobic Cluster Analysis, his enthusiastic suggestions and comments. They also thank Dr Raphaël Guerois for critical reading of this manuscript. They are grateful to Dr Joannes Söding and Dr Aron Marchler-Bauer for allowing them to include their tools in the TREMOLO-HCA package.

Funding: Agence Nationale de la Recherche (ANR Blanc-SVSE-8-2011-TELO&DICENs) and Institut National du Cancer (INCa-DiREP).

Conflict of Interest: none declared.

REFERENCES

Adams-Cioaba,M.A. *et al.* (2012) Crystal structures of the Tudor domains of human PHF20 reveal novel structural variations on the Royal Family of proteins. *FEBS Lett.*, **586**, 859–865.

Adams-Cioaba,M.A. and Min,J. (2009) Structure and function of histone methylation binding proteins. *Biochem. Cell Biol.*, **87**, 93–105.

Aravind,L. and Koonin,E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.*, **287**, 1023–1040.

Armache,K.J. *et al.* (2011) Structural basis of silencing: Sir3 BAH domain in complex with a nucleosome at 3.0 Å resolution. *Science*, **334**, 977–982.

Bettegowda,C. *et al.* (2011) Mutations in CIC and FUBP1 contribute to human oligodendroglioma. *Science*, **333**, 1453–1455.

Callebaut,I. and Mornon,J.P. (1997a) From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair. *FEBS Lett.*, **400**, 25–30.

Callebaut,I. and Mornon,J.P. (1997b) The human EBNA-2 coactivator p100: multidomain organization and relationship to the staphylococcal nuclease fold and to the tudor protein involved in *Drosophila melanogaster* development. *Biochem. J.*, **321**, 125–132.

Callebaut,I. *et al.* (1997) Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol. Life Sci.*, **53**, 621–645.

Callebaut,I. *et al.* (1999) The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. *FEBS Lett.*, **446**, 189–193.

Callebaut,I. *et al.* (2002) Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Res.*, **30**, 3592–3601.

Callebaut,I. *et al.* (2005) Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. *BMC Genomics*, **6**, 100.

Callebaut,I. *et al.* (2006) Cernunnos interacts with the XRCC4 x DNA-ligase IV complex and is homologous to the yeast nonhomologous end-joining factor Nej1. *J. Biol. Chem.*, **281**, 13857–13860.

Cui,G. *et al.* (2012) PHF20 is an effector protein of p53 double lysine methylation that stabilizes and activates p53. *Nat. Struct. Mol. Biol.*, **19**, 916–924.

Dawson,M.A. *et al.* (2011) Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature*, **478**, 529–533.

Ekman,D. *et al.* (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.*, **348**, 231–243.

Eudes,R. *et al.* (2007) A generalized analysis of hydrophobic and loop clusters within globular protein sequences. *BMC Struct. Biol.*, **7**, 2.

Finn,R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.

Florence,B. and Faller,D.V. (2001) You bet-cha: a novel family of transcriptional regulators. *Front Biosci.*, **6**, D1008–D10018.

Gaboriaud,C. *et al.* (1987) Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Lett.*, **224**, 149–155.

García-Cuéllar,M.P. *et al.* (2000) ENL, the MLL fusion partner in t(11;19), binds to the c-Abl interactor protein 1 (ABI1) that is fused to MLL in t(10;11)+. *Oncogene*, **19**, 1744–1751.

García-Cuéllar,M.P. *et al.* (2001) The ENL moiety of the childhood leukemia-associated MLL-ENL oncoprotein recruits human Polycomb 3. *Oncogene*, **20**, 411–419.

He,N. *et al.* (2011) Human Polymerase-Associated Factor complex (PAFc) connects the Super Elongation Complex (SEC) to RNA polymerase II on chromatin. *Proc. Natl Acad. Sci. USA*, **108**, E636–E645.

Hennetin,J. *et al.* (2003) Non-intertwined binary patterns of hydrophobic/nonhydrophobic amino acids are considerably better markers of regular secondary structures than nonconstrained patterns. *Proteins*, **51**, 236–244.

Hickman,M.A. and Rusche,L.N. (2010) Transcriptional silencing functions of the yeast protein Orc1/Sir3 subfunctionalized after gene duplication. *Proc. Natl Acad. Sci. USA*, **107**, 19384–19389.

Hou,Z. *et al.* (2005) Structural basis of the Sir1-origin recognition complex interaction in transcriptional silencing. *Proc. Natl Acad. Sci. USA*, **102**, 8489–8494.
Jiménez,G. *et al.* (2012) The Capicua repressor—a general sensor of RTK signaling in development and disease. *J. Cell Sci.*, **125**, 1383–1391.
Kuo,A.J. *et al.* (2012) The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature*, **484**, 115–119.
Le Masson,I. *et al.* (2003) Yaf9, a novel NuA4 histone acetyltransferase subunit, is required for the cellular response to spindle stress in yeast. *Mol. Cell. Biol.*, **23**, 6086–6102.
Leach,B.I. *et al.* (2013) Leukemia fusion target AF9 is an intrinsically disordered transcriptional regulator that recruits multiple partners via coupled folding and binding. *Structure*, **21**, 176–183.
Lin,Y.J. *et al.* (2008) Solution structure of the extraterminal domain of the bromodomain-containing protein BRD4. *Protein Sci.*, **17**, 2174–2179.
Luo,Z. *et al.* (2012) The super elongation complex (SEC) family in transcriptional control. *Nat. Rev. Mol. Cell. Biol.*, **13**, 543–547.
Marchler-Bauer,A. *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.*, **41**, D348–D352.
Muntean,A.G. and Hess,J.L. (2012) The pathogenesis of mixed-lineage leukemia. *Annu. Rev. Pathol.*, **7**, 283–301.
Poupon,A. and Mornon,J.P. (1998) Populations of hydrophobic amino acids within protein globular domains: identification of conserved ''topohydrophobic'' positions. *Proteins*, **33**, 329–342.
Remmert,M. *et al.* (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **25**, 173–175.
Rusch,D.B. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77.
Schulze,J.M. *et al.* (2010) Reading chromatin. Insights from yeast into YEATS domain structure and function. *Epigenetics*, **5**, 573–577.
Siew,N. and Fischer,D. (2004) Structural biology sheds light on the puzzle of genomic ORFans. *J. Mol. Biol.*, **342**, 369–373.
Slany,R.K. *et al.* (1998) The oncogenic capacity of HRX-ENL requires the transcriptional transactivation activity of ENL and the DNA binding motifs of HRX. *Mol. Cell. Biol.*., **18**, 122–129.
Söding,J. *et al.* (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
Tan,J. *et al.* (2011) CBX8, a polycomb group protein, is essential for MLL-AF9-induced leukemogenesis. *Cancer Cell*, **20**, 563–575.
Taverna,S.D. *et al.* (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.*, **14**, 1025–1040.
Wolf,Y.I. *et al.* (2000) Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.*, **299**, 897–905.
Woodcock,S. *et al.* (1992) Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng.*, **5**, 629–635.
Yap,K.L. and Zhou,M.M. (2010) Keeping it in the family: diverse histone recognition by conserved structural folds. *Crit. Rev. Biochem. Mol. Biol.*, **45**, 488–505.
Yin,J. *et al.* (2005) BLAP75, an essential component of Bloom's syndrome protein complexes that maintain genome integrity. *EMBO J.*, **24**, 1465–1476.
Zeisig,D.T. *et al.* (2005) The eleven-nineteen-leukemia protein ENL connects nuclear MLL fusion partners with chromatin. *Oncogene*, **24**, 5525–5532.
Zhang,W. *et al.* (2011) Solution structure of the Taf14 YEATS domain and its roles in cell growth of *Saccharomyces cerevisiae*. *Biochem. J.*, **436**, 83–90.