doi:10.1016/j.jmb.2010.05.012 J. Mol. Biol. (2010) 400, 605–617

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

T-RMSD: A Fine-grained, Structure-based Classification Method and its Application to the Functional Characterization of TNF Receptors

Cedrik Magis[1]*, François Stricher[1], Almer M. van der Sloot[1], Luis Serrano[1,3] and Cedric Notredame[2]

[1]*EMBL/CRG Systems Biology Research Unit Center for Genomic Regulation (CRG), UPF, Barcelona, Catalunya, 08003, Spain*

[2]*Bioinformatics and Genomics program, Center for Genomic Regulation (CRG), UPF, Barcelona, Catalunya, 08003, Spain*

[3]*Institucio Catalana de Recerca I Estudis Avancats (ICREA) professor, Center for Genomic Regulation (CRG), Barcelona, Catalunya, 08003, Spain*

Received 18 March 2010; received in revised form 21 April 2010; accepted 7 May 2010
Available online 13 May 2010

Edited by M. Sternberg

This study addresses the relation between structural and functional similarity in proteins. We introduce a novel method named tree based on root mean square deviation (T-RMSD), which uses distance RMSD (dRMSD) variations to build fine-grained structure-based classifications of proteins. The main improvement of the T-RMSD over similar methods, such as Dali, is its capacity to produce the equivalent of a bootstrap value for each cluster node. We validated our approach on two domain families studied extensively for their role in many biological and pathological pathways: the small GTPase RAS superfamily and the cysteine-rich domains (CRDs) associated with the tumor necrosis factor receptors (TNFRs) family. Our analysis showed that T-RMSD is able to automatically recover and refine existing classifications. In the case of the small GTPase ARF subfamily, T-RMSD can distinguish GTP- from GDP-bound states, while in the case of CRDs it can identify two new subgroups associated with well defined functional features (ligand binding and formation of ligand pre-assembly complex). We show how hidden Markov models (HMMs) can be built on these new groups and propose a methodology to use these models simultaneously in order to do fine-grained functional genomic annotation without known 3D structures. T-RMSD, an open source freeware incorporated in the T-Coffee package, is available online.

Keywords: structural classification; functional classification; multiple sequence alignment; Tumor Necrosis Factor Receptor; Cysteine-rich Domain

Introduction

Fold identification is a very active field of research, whose importance reflects the usefulness of structural information when elucidating protein function.[1] It is therefore not surprising to note that several structural classifications have been developed over time.[2–4] These resources, which are either used in a stand alone fashion or through their integration within domain databases,[5] tend to be focused on the identification of generic folds. In practice, these resources are mostly used to help assigning an anonymous sequence to some high-level fold/functional category. The subtle differences associated with fine-grain functional and structural variations are often overlooked. Refining such classification and showing that fine-grain structural clustering can have clear functional implications is precisely the object of the work presented here.

We have developed a novel method named tree based on RMSD (T-RMSD), which computes a clustering using a combination of structure comparison and tree reconstruction methods. The aim of the model is to identify subtle structural differences within a group of homologous protein sequences or domains using distance RMSD measures (dRMSD) as a similarity metrics. Contrarily to the standard

*Corresponding author. E-mail address: cedrik.magis@crg.es.

Abbreviations used: T-RMSD, tree based on RMSD; dRMSD, distance RMSD; CRD, cysteine-rich domain; TNFR, tumor necrosis factor receptor.

RMSD that measures the distance between superposed C^{α} equivalents, dRMSD[6] relies on a comparison between homologous intramolecular distances. It is therefore independent of any structural superposition, and depends only on how equivalences are defined. Dali[7] is probably the most popular dRMSD-based superposition procedure. Interestingly, the metrics can be used to evaluate MSAs independently of any reference or superposition,[8,9] a property that makes it an ideal measure for comparing the structural merits of alternative alignments using a meta-aligner such as M-Coffee[10] or 3D-Coffee.[11] Existing dRMSD-based metrics, however, can be used only to estimate a global structural similarity, thus making it hard to properly quantify the reliability of clusters built upon them. Here, we show that one can take advantage of the distance-based nature of the dRMSD to build distinct clusters on each individual column of an alignment, thus producing a collection of alternative clusters that can be compared, combined and assessed for robustness, in a way reminiscent of bootstrap support analysis in phylogeny. In practice, a bootstrap value is associated with each node. It indicates the fraction of ungapped columns supporting the split associated with the considered node (100 indicates a full support from all considered columns). This approach, outlined in Fig. 1, forms the backbone of the T-RMSD.

Clustering strategies are mostly exercises in description, and their full power materializes only when clusters prove useful to reveal functional relationships. One of the main limitations when estimating the relative merits of fine-grain classifiers is the scarcity of experimental data lending itself to such analysis. Indeed, one needs a family of sequences with several well characterized members both from a structural and a functional point of view. One also requires these members to be sufficiently functionally distinct to allow an estimation of how well a clustering manages to correlate functional and structural variations. In practice, the

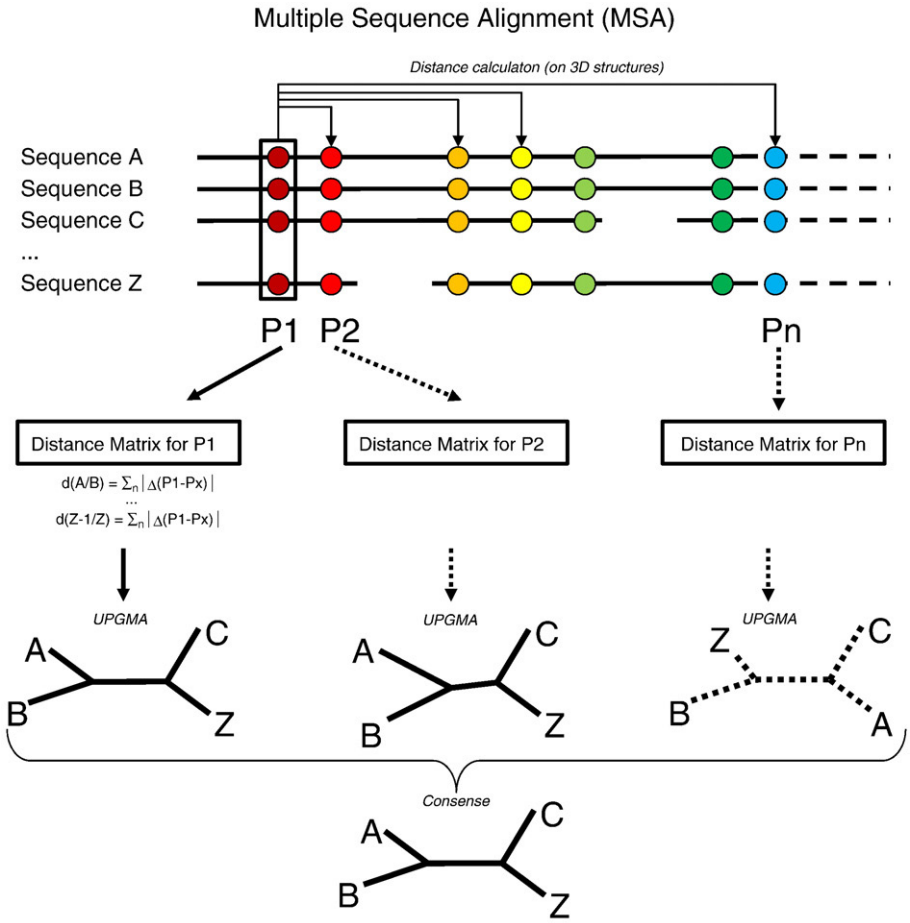

Fig. 1. A schematic display of the T-RMSD procedure. A multiple sequence alignment is used to declare equivalences among aligned residues. Each sequence must have a known 3D structure. For a given ungapped position (P1) and a given sequence A, the C^{α}–C^{α} distance between the residue at position P1 and every other residue of A (occurring in an ungapped column) is estimated on the 3D structure. A similar set of distances is measured on structure B. The divergence between A and B is estimated by summing the absolute differences between equivalent pairs on A and B. A divergence matrix can be computed for position P1 by considering every possible pair of sequence. Such matrices are computed for every ungapped position of the MSA and are then turned into a UPGMA tree. The collection of trees thus obtained is combined into a consensus tree using the consense program of the PHYLIP package (see Methods).

analysis of the Pfam[12] domain database (containing >11,000 families) reveals about 100 protein families associated with sufficient structural information (>20 distinct UniProt ID). These families include protein-binding domains (PDZ, SH2 and SH3), nucleic acid-binding protein (RNA recognition motifs, zinc fingers and U-box) and enzymes (protein kinases, histone acetyltransferases and proteases). Unfortunately, and although a large literature corpus is available for many of these families, the required information is not yet integrated with appropriate levels of detail and the establishment of reference datasets such as those used here remains a labor-intensive process. In the context of this work, we have focused the validation of our method (T-RMSD) on two well-described domain families: the small GTPase RAS superfamily[13] and the cysteine-rich domains (CRDs) associated with tumor necrosis factor receptors[14] (TNFRs). Both of these domain families have in common the availability of a large number of experimental 3D structures (between 20 and 50 non-redundant structures), along with enough experimental data to support the existence of significantly different functional subgroups.

These two families also define the two extremes of our functional/structural spectrum. On the one hand, RAS, a super family made of related small GTPases with subcategories defined well enough at the sequence level to be recovered using profile collections. These subcategories are associated unambiguously with distinct features such as cellular localization, pathway and conformational switches related to GTP hydrolysis (Table 1). Altogether, this makes the GTPase RAS superfamily an ideal validation target. The CRDs, on the other hand, exhibit a less clear functional/structural relationship. CRDs constitute the ligand-binding region of the TNFRs, a class of receptors involved in many key cellular processes (cell survival, programmed cell death, proliferation) and whose association with important pathologies is well documented (cancer, viral infection, autoimmune disease).[15–17] While TNFRs exhibit very different signaling capacities, sizes and a variety of domain organization, most of them seem to use similar binding modes.[18] Because of its biological importance, this process has received a lot of attention,[19–23] although the precise determinants shaping CRD specificity toward a given TNFL remain to be determined. Sequence analysis alone brings little information, with the main domain databases (Pfam, Smart[24] and PROSite[25]) showing, on average, 40% agreement with UniProt annotation.[26] This disagreement should not come as a surprise given the wide sequence diversity measured across CRDs (about 18±8% of identity). The current CRD classification is therefore structure-based[27,28] and relies on a module decomposition of each domain (N and C terminal), with each module making up about half of a CRD. This approach conveniently describes all existing CRDs but it fails at defining functionally homogeneous groups and is therefore not really informative. Furthermore, the most common module combination puts together CRDs that can be structurally diverse (RMSD up to 5 Å) and probably just as functionally different, given their tendency to occur in any part of the domain architecture. In the context of this work, one of our goals was to assess whether systematic structure-based clustering can yield new more homogeneous categories. The resulting methodology is very generic and can be applied to any protein family for which the definition of structural subgroups makes biological sense (provided enough structural data are available).

Table 1. Small GTPase RAS superfamily functional classification

Domain	PDB	ID (%)	Function/Localization	Reference
RAS	9	46±10	Specific association with plasma membrane, Golgi or ER after post-translational modification.	25
RHO	5	52±13	Rho GTPases are key regulators of cytoskeletal dynamics and affect many cellular processes, including cell polarity, migration, vesicle trafficking and cytokinesis	26
RAB	19	42±9	Organise membrane trafficking processes, vesicle budding, uncoating, mobility and fusion	27
RAN	1	n.a.	Ran is located predominantly in the nucleus of eukaryotic cells and is involved in the nuclear import of proteins as well as in control of DNA synthesis and of cell-cycle progression, nuclear localization, and its lack of sites required for post-translational lipid modification	28
SAR	0	n.a.	Recruitment of proteins or complexes to initiate vesicle budding.	29
ARF	6	48±10	Regulation of vesicle biogenesis in intracellular traffic/Phospholipase D activator	29,30
RGK	1	n.a.	Cytoskeletal remodeling via RHO interaction, voltage-gated calcium channel activity modulation.	31
Ambiguous	2	n.a.	Undefined	—
Overall	50	29±12		

The RAS superfamily is dissected into seven domain families, each assuming a specific function and cellular localization. For the overall superfamily, as for each subfamily, the average identity and its standard deviation are indicated, calculated for the sequences included in our reference set.

Results

Validation of T-RMSD for functional classification of RAS subfamilies

We first estimated the relevance of the structure-based clustering produced by the T-RMSD (Fig. 1; Materials and Methods) on the classification of the RAS GTPase subfamilies (Table 1).[29-35] For that purpose, non-redundant sequences with a known structure (Supplementary Data Table 1) were multiply aligned with 3D-Coffee and the resulting MSA was used to derive a neighbor-joining (NJ) tree (Supplementary Data Fig. 2A). As expected, this clustering roughly coincides with the widely accepted RAS classification,[13] with the exception of centaurin γ1 (2IWR) and centaurin γ3 (3IHW). This does not come as a surprise because these two proteins are hard to classify, which is reflected by the disagreement among domain databases. The same 3D-Coffee MSA was then fed into T-RMSD in order to produce a structure-based clustering (Fig. 2b). The topology thus obtained is roughly similar to the sequence-based analysis with one notable exception, its capacity to distinguish the GTP-bound from the GDP-bound states of the ARF subfamily.[36,37] The split separating these two subgroups (ARF-GTP and ARF-GDP) has a bootstrap value of 38, which means that 38% of the columns support this dichotomy. However, discrimination cannot be achieved on the basis of sequence alone because each RAS GTPase is able to bind both GTP and GDP. This example therefore illustrates well the advantage of a structure-based clustering over its sequence-based counterpart. Moreover, obtaining the reliability value (bootstrap) also shows the benefits of basing an analysis on the simultaneous use of many possible models (one per column) rather than relying on a global measure. It is worth mentioning that discrimination of GTP and GDP forms could be achieved only on the ARF subfamily, which suggests a more concerted conformational change related to GTP or GDP binding than the rest of the GTPases of the RAS superfamily.

Structure-based classification of CRDs associated with the TNFR family

We applied the same procedure to the CRD dataset. In that context, the goal was to validate the method and to estimate whether the current classification could be refined to become more informative. A set of 22 non-redundant sequences was identified (see Materials and Methods) and multiply aligned using 3D-Coffee. The resulting MSA was used to compute the unrooted neighbor-joining phylogenetic tree shown in Fig. 3a. In theory, such a tree could be used to reveal subclasses (as happens with the RAS family) but in practice the evolutionary distance between the considered domains is too great, resulting in an unrooted tree poorly resolved with no significant bootstrap support for any deep node (Fig. 3a). Furthermore, the subgroups it defines show no significant agreement with the current module based classification.[27,28]

The current CRD classification is based on the dissection and comparison of five PDB structures suggesting nine different modules (A1, A2, B1, B2, C2, X2, D2, N and N1), the combination of which can be used to describe eight CRD types. The 22 CRDs considered in the dataset are predominantly A1-B2 types (underlined in red in Fig. 3a). They appear to be evenly spread and associated with all the other module combinations such as A1-B1, A2-B1, A1-C2, A1-D2 or X2-N. Similar results were obtained even when excluding columns containing gaps. When applying the T-RMSD analysis to the same dataset (Fig. 3b), we obtain a much more informative topology, clearly supporting three well defined groups with one outlier. The bootstrap values indicate the fraction of ungapped columns (13 in total) supporting a given split and suggest a good support for these three groups. For instance, the separation between group II and the other CRDs is supported by 100% of ungapped columns, while 53% support the separation between group I and III. Interestingly, the groups thus defined are fairly homogeneous from a module point of view, with groups I and II composed almost exclusively by A1-B2 CRDs and group III perfectly overlapping a previously described CRDs category: the small TNFR (made of one or two CRDs).[38] We compared these results with those obtained when using the Dali *Z*-score as a clustering metrics (Materials and Methods) and found the resulting clusters to be in broad agreement (Supplementary Data Fig. 1) even though they lack the bootstrap value associated with the T-RMSD clusters.

Structural characterization of the new CRDs classification

The differences captured here, which clearly reflect genuine structural variation, might be anecdotal and non-biologically informative. In order to rule out this possibility we performed a detailed analysis of the clusters, exploring all three layers of information associated with the considered sequences: homology, structure and function. Our goal was to establish whether each cluster is specifically associated with a structural or functional property.

Each subgroup was realigned using 3D-Coffee and analyzed for sequence conservation. As expected, the degree of conservation is fairly poor aside from a few highly conserved residues: one highly conserved aromatic/hydrophobic position whose involvement in the domain packing has been previously described through structural analysis[19,21] and the disulphide bridges network (C1–C2, C3–C5 and C4–C6). Nonetheless, it is interesting to note how the three groups identified at the structural level might have been easily defined on the basis of their insertion-deletion (indel) patterns (Fig. 4a). For

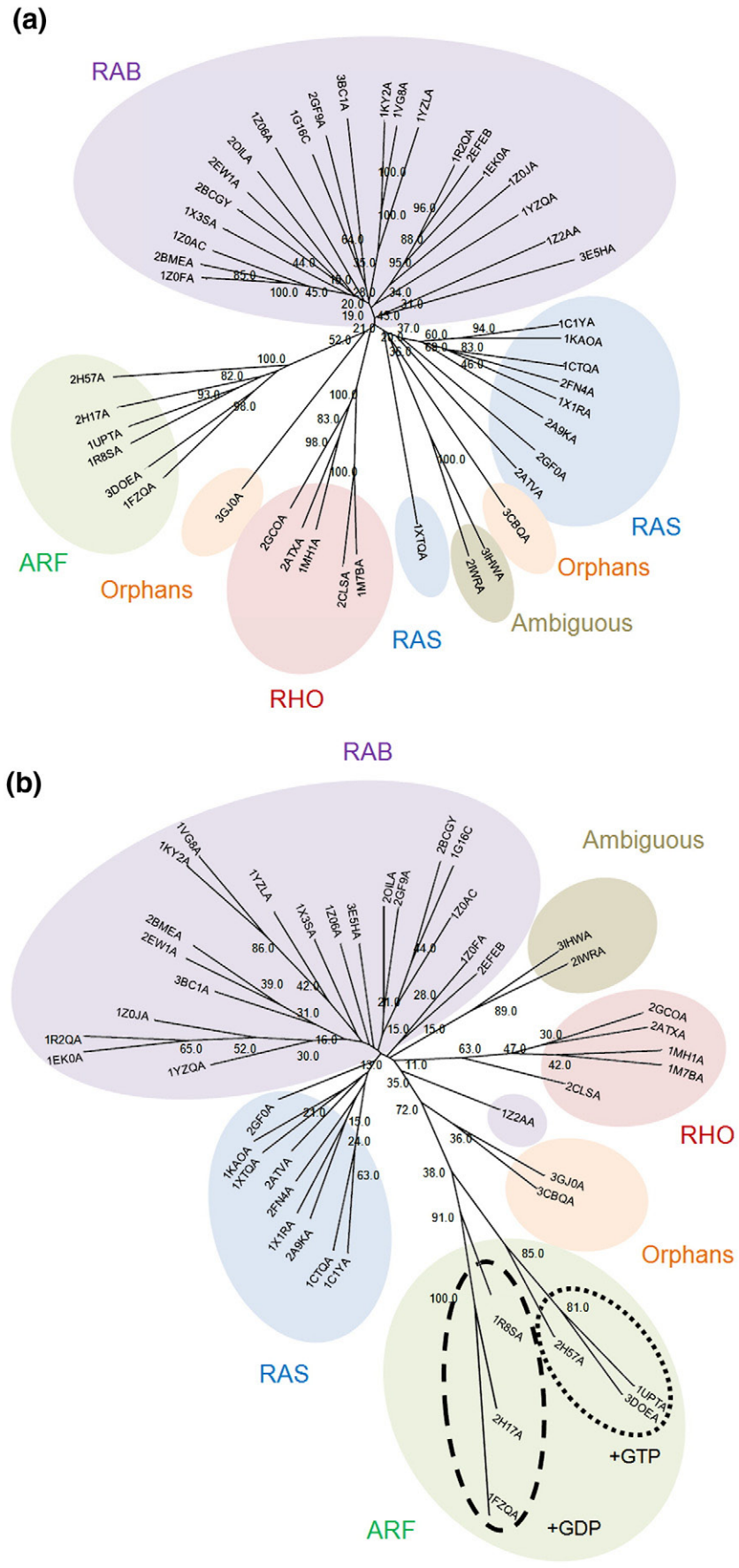

Fig. 2. Tree analysis of current sequence/function clustering of small GTPase RAS superfamily compared to a new clustered based classification. (a) Phylogenetic analysis of the 43 structures of the non redundant GTPase RAS (see Methods). The neighbor-joining tree was derived from MSA of CRDs using 3D-Coffee. For clarity, bootstrap values under 10.0 are not indicated on the nodes. (b) T-RMSD structural clustering derived from MSA obtained by 3D-Coffee. RAB domains are indicated in light violet, RAS domains in light blue, ARF domains in light green, orphans domains (RAN and RGK) in light orange, RHO in light red and ambiguous domains in khaki.

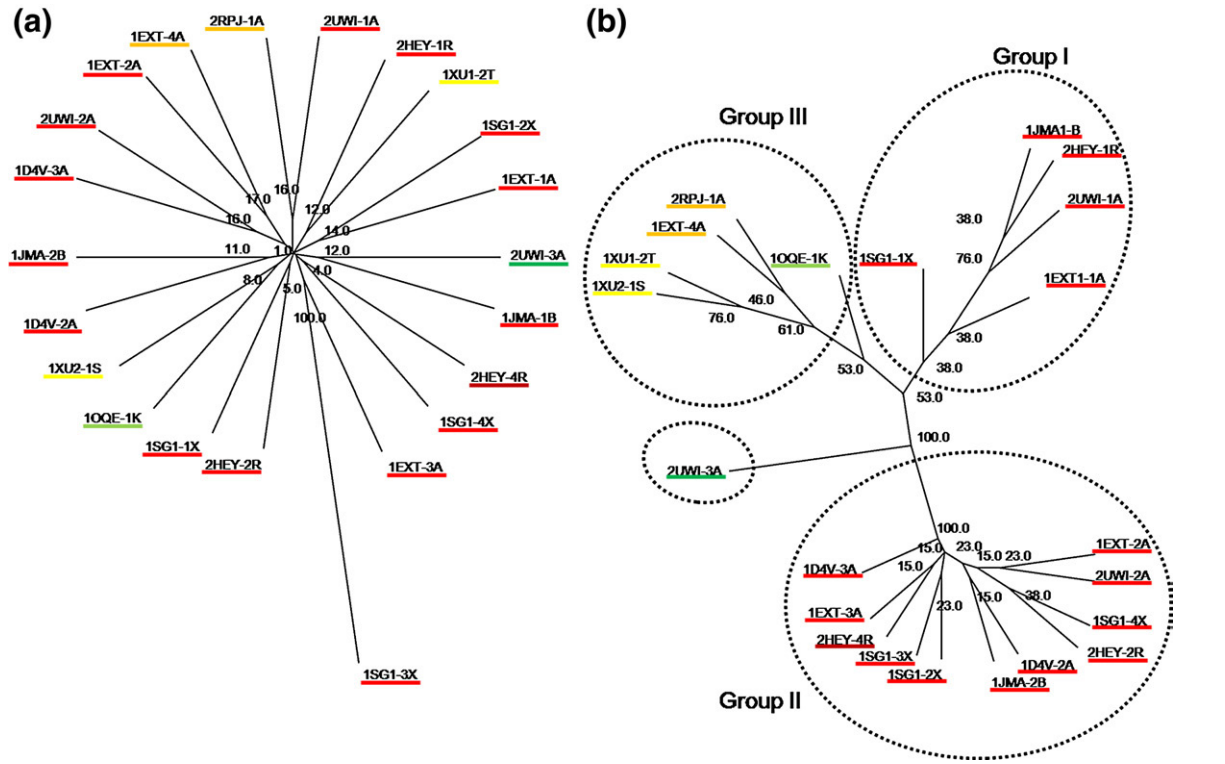

Fig. 3. Tree analysis of current module-based CRD classification compared to a new clustered based classification. (a) Phylogenetic analysis of the 22 structures of non redundant CRDs (see Methods). The neighbor-joining tree was derived from MSA of CRDs using 3D-Coffee. (b) T-RMSD structural clustering derived from MSA obtained by 3D-Coffee. For a and b, each CRD is represented by its PDB identifier, followed by its position number within the architecture of its corresponding TNFR and the chain identifier of the corresponding PDB file. The module composition is indicated by colored underlines: A1-B1, dark red; A1-B2, red; A1-C2, orange; A1-D2, yellow; X2-N, light green; and A2-B1, dark green. Bootstrap values are given for each node.

instance, in group I, the block defined by the C2–C5 cysteines is fairly compact and does not seem to tolerate insertions or deletions. By contrast, the corresponding region within groups II and III is much more variable. Conversely, the C5–C6 block of group II is more constrained than its counterparts in groups I and III. As a consequence, the overall length is fairly conserved within each group but variable across groups. Group II is the largest, 5–10 residues longer than groups I and III (Table 2). Interestingly, the very nature of this variability probably explains the failure of sequence-based clustering methods, whose metrics are based on identity rather than a comparison of indel patterns.

We then asked if such conservation patterns could reflect structural homogeneity. We measured the average standard RMSD in each group (Table 2) and found them highly homologous with average values below 1.0 Å within each group (see Materials and Methods). This observation is not surprising given the structure-based nature of the clustering, but it nonetheless confirms the adequacy of dRMSD as a distance metric. As suggested by their common A1-B2 module composition, groups I and II contain the most similar members (Fig. 4b, left and middle respectively), with precisely conserved topologies (loop 1 & 2, turn 1 & 2). Despite a smaller structural core (black dotted lines in Fig. 4b, right) group III is a bit more diverse albeit homogeneous enough to also yield a standard RMSD of less than 1.0 Å.

Homology-based validation of the new CRD classification

Another way to estimate the biological meaning of our three clusters is to assess the discriminative power of sequence homology models based upon

Fig. 4. Sequence and structure analysis of CRD structural groups (a) MSAs of CRDs type I, II and III obtained from 3D-Coffee. Alignments are illustrated using JalView 2.4 and conserved residues are colored from light blue to dark blue. Conserved cysteine (dark blue columns) residues connectivity is indicated by black lines. Structural elements described after structural superposition (Fig. 3b) are indicated under each alignment. Consensus sequences indicated for each type were derived from MSAs. (b) Structural superimposition of CRDs according to their structural type I, II or III. CRD structures are represented as ribbon and the disulfide bridge is shown as a black stick. CRDs are colored using a rainbow scale according to the alphabetical order given by the PDB names and the position of the domain in the corresponding TNFR architecture (from ***1D4V-2*** to ***2HEY-4***). Structurally variable regions are circled by broken black lines and structurally conserved regions are circled by black dotted lines.

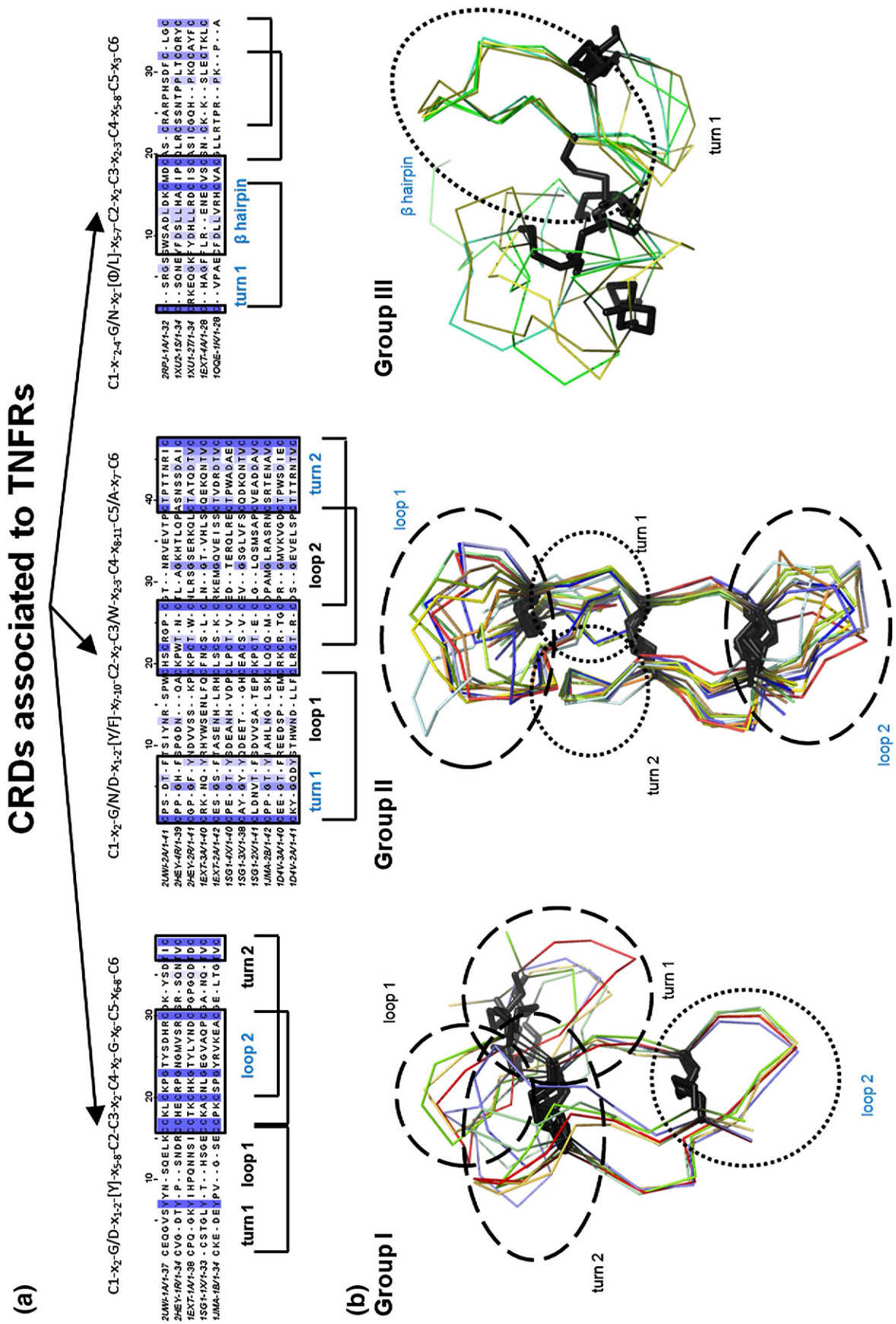

Fig. 4 (*legend on previous page*)

Table 2. Characterization of CRDs according to structural type.

Sequence characteristics	Group I	Group II	Group III
Overall size	35±2	40±2	31±3
Average identity (%)	33±4	30±7	26±7
Core block length	18	22	13
Core block length after superposition	16.0±1.2	19.3±1.9	10.5±2.8
Ratio size[a] (%)	44.8±5.0	48.4±6.2	34.2±11.2
Core block RMSD (Å)	0.64±0.13	0.66±0.32	0.75±0.75

Overall size corresponds to the number of residues within CRDs (average). Core block size corresponds to the number of structurally conserved residues. The Core block/Overall ratio corresponds to the ratio of the number of structurally conserved residues (Core block) to the total number of residues (Overall, average). Structural similarity of core blocks estimated by average RMSD values, indicated (in Å) for each cluster (see Methods).

[a] Core block/Overall.

them. In order to do this analysis, the MSAs associated with each of the three groups were turned into hidden Markov model (HMM) profiles (Materials and Methods) and used to scan the CRD database, the rationale being that the HMM giving the best score is more likely to define the true nature of considered CRDs. This methodology was validated beforehand using a leave-one-out strategy to assess each HMM specificity/sensitivity (Table 3). The leave-one-out was done by removing in turn each sequence from its corresponding MSA, re-estimating the associated HMM and measuring the *E*-value of the left out sequence when aligned with the HMMs of the three groups (i.e. the re-estimated HMM and the other two). When applied in turn to each of the 22 CRDs this validation shows that 100% of the sequences achieve a better score (lower *E*-value) when aligned to the HMM corresponding to their original group (Table 3). This high specificity is unlikely to result from any bias in sequence similarity given the similar level of identity across the three CRD types (28±7%) or within each of the three groups (26–33%).

The three HMMs built on the full dataset for each group were used to annotate the 80 remaining human sequences annotated as CRDs or atypical CRDs in UniProt but for which no experimental structure is available. Domains with an *E*-value greater than 0.1 with each of the HMMs of the three groups were left unclassified (22 in total). The remainder were labeled (i.e. assigned to group I, II or III) by the HMM yielding the lowest *E*-value (Supplementary Data Table S2). Newly and non-ambiguously automatically classified CRDs were then incorporated in the groups MSAs and used to update their respective HMM profiles (Fig. 5). The three HMMs thus produced were used to label all the CRDs with UniProt sequences annotated as TNFRs. In 98% of the cases, orthologous CRDs received an identical label across all vertebrate species.

Functional evaluation of the new classification

The three CRD groups identified through structural analysis were analyzed for known functions associated with their members. CRD repeats have been shown to play distinct roles in TNFRs. For instance, the N-terminal CRD (known as the pre-ligand assembly domain or PLAD) has a very specific feature that sets it apart from other CRDs: it does not necessarily interact with the ligand but forms a specific pre-ligand assembly complex (homo

Table 3. Leave-one-out strategy performed on CRD sequences using HMM profiles built from MSA according to structural clustering described earlier

Domain information				*E* value		
Name	PDB	Gene	Domain	Gr I	Gr II	Gr III
TNF-R1	1EXT	TNFRSF1A	1	3,7E-05	6,4E-05	>0,1
Ox40	2HEY	TNFRSF4	1	3,5E-07	1,2E-03	>0,1
HVEM	1JMA	TNFRSF14	1	5,9E-07	5,4E-05	1,0E-01
NGFR	1SG1	TNFRSF16	1	4,2E-08	3,1E-06	>0,1
CRME	2UWI	CRME	1	5,9E-06	8,2E-05	9,1E-04
TNF-R1	1EXT	TNFRSF1A	2	1,4E-03	1,3E-14	4,0E-03
TNF-R1	1EXT	TNFRSF1A	3	1,4E-03	2,7E-06	9,5E-03
Ox40	2HEY	TNFRSF4	2	2,2E-05	6,1E-10	>0,1
Ox40	2HEY	TNFRSF4	4	>0,1	1,7E-06	>0,1
DR5	1D4V	TNFRSF10B	2	6,2E-05	3,3E-13	>0,1
DR5	1D4V	TNFRSF10B	3	3,7E-03	3,6E-08	>0,1
HVEM	1JMA	TNFRSF14	2	1,3E-05	8,9E-12	2,3E-03
NGFR	1SG1	TNFRSF16	2	1,1E-02	5,1E-07	>0,1
NGFR	1SG1	TNFRSF16	3	2,8E-05	3,9E-08	>0,1
NGFR	1SG1	TNFRSF16	4	4,2E-03	9,7E-16	>0,1
CRME	2UWI	CRME	2	>0,1	2,1E-05	>0,1
TNF-R1	1EXT	TNFRSF1A	4	>0,1	>0,1	7,0E-02
TACI	1XU1	TNFRSF13B	1	>0,1	>0,1	2,2E-08
BCMA	1OQD	TNFRSF17	1	3,1E-02	1,7E-02	8,6E-05
TWEAKR	2RPJ	TNFRSF12A	1	>0,1	>0,1	3,7E-02
BAFF-R	1OQE	TNFRSF13C	1	>0,1	>0,1	3,1E-02

CRDs are indicated by the common name of the receptor they belong to, the PDB file used for the structural analysis, the gene name of the corresponding receptor and the position in the receptor architecture from N to C terminus. Non-significant *E* values are >0.1.

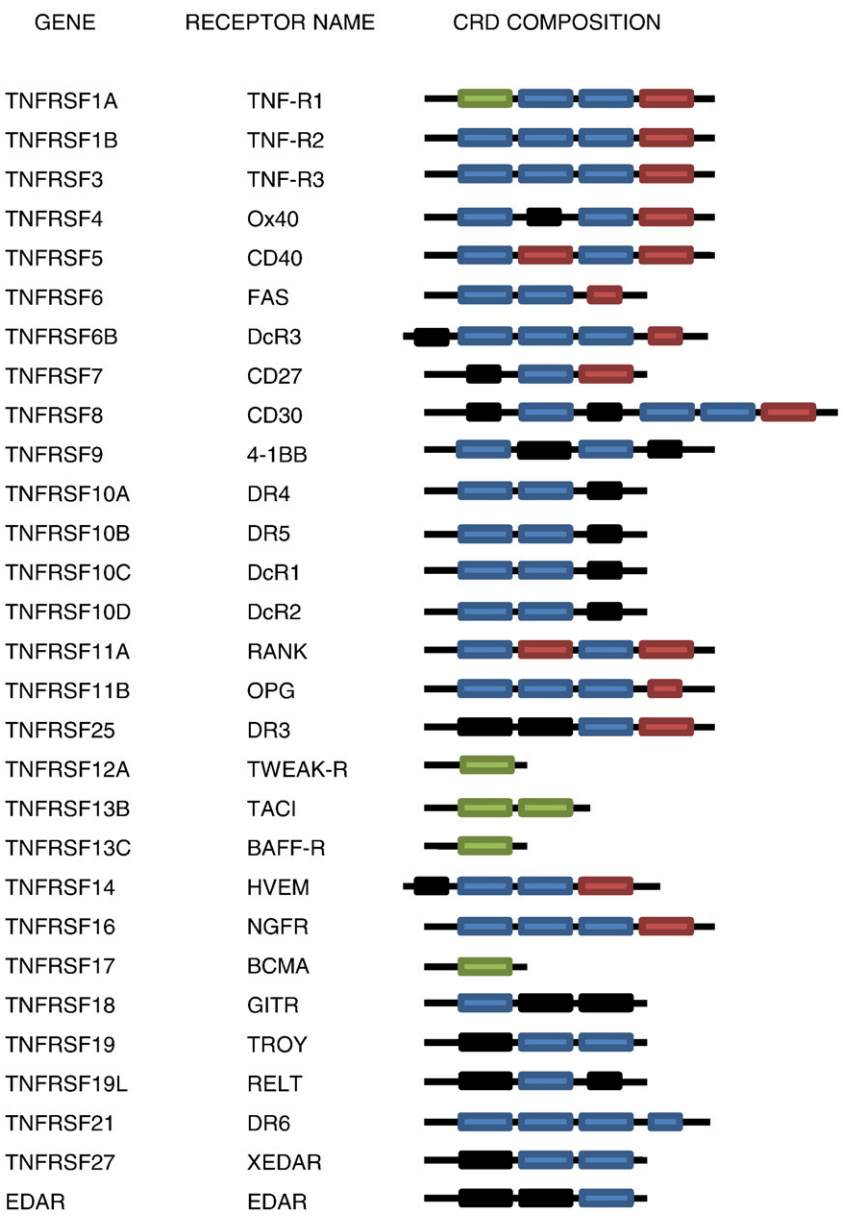

Fig. 5. Classification of human putative CRDs associated with TNFRs according to the structural types I, II and III defined here. Type I, red box; type II, blue box; and type III, green box. Unclassified CRDs are represented as a black box. Half CRDs are identified by smaller boxes.

or hetero receptor–receptor complexes), prerequisite in the TNFL binding process.[39] The existence of such domains, which are essential for binding, has been shown experimentally in several human TNFRs, including CD40, TNFR1, TNFR2 and FAS.[40,41] The core CRDs, however, are directly involved in ligand binding activities as revealed by the structural analysis of TNFR/TNFL complexes (Materials and Methods). The annotation resulting from the scanning of human TNFRs with our new collection of CRD HMMs is shown in Fig. 5. A total of 88% of the type I CRDs occur on the N terminus, thus labeling the termination of 56% of the large receptors (three to six CRDs). In contrast, type II CRDs are found predominantly (about 92%) in the TNFR core. In agreement with the literature, we found type III to be mostly associated with small TNFRs made of one or two CRD repeats for which a single domain is sufficient for a high level of interaction specificity.[38] These findings suggest that the two new subcategories identified within the A1-B2 module group correspond to two genuinely distinct classes of CRDs: TNFL binders and PLAD CRDs. Unassigned CRDs present us with the possibility that other categories might exist, yet unreported for lack of structural information.

Discussion

Here, we describe T-RMSD, a novel method for fine-grained structural classification. Given an MSA where each sequence has a known structure, the T-RMSD uses dRMSD measurements to estimate a

clustering for each column of the MSA. The resulting clusters are combined into one unique consensus cluster, whose nodes are scored for their frequency among the combined clusters. We show that this algorithm can recover existing and validated classifications (small GTPase RAS superfamily) or allow the refinement of existing classifications (CRD family). But more importantly, we show that these fine-grain structural classes correspond to well-defined functional variations, which justify the systematic usage of this approach for refined domain annotation.

The T-RMSD could be described as DALI clustering with bootstrap values. It is therefore not surprising in any way that our method delivers accurate results, but it must be pointed out that providing meaningful reliability values for cluster nodes is just as important as providing a clustering in the first place. Our approach provides reliability estimation, and it gives an indication of which intramolecular distances support better a given split in the clustering (i.e. the column setting apart one subgroup from the others). We do not use this information in the current work, but it could certainly find important applications in homology modeling and drug design. The clusters produced by T-RMSD are very precise and reliable. Even though it is fairly obvious that large families can be divided into smaller subfamilies, we show here that these refined classifications are sophisticated usage of structural information, and that they make biological sense, probably because the structure/function relationship extends deep, at least within the small GTPase RAS superfamily and CRD family.

Validation is achieved on the RAS superfamily. We show how the T-RMSD recovers the known classification and manages to precisely identify within ARF the structural differences associated with the two states of GTPase proteins: GTP and GDP binding. We also tested the behavior of the T-RMSD on a more challenging dataset: the TNFR CRDs. As illustrated by the lack of a clear consensus among domain databases and large literature corpus, classifying CRDs is a challenging endeavor. Our results suggest that the T-RMSD approach is able to recover a classification roughly similar to that produced by Dali. However, the bootstrap support makes it possible to use well supported deep branching nodes to define three groups unambiguously and in a non-arbitrary way. One difficulty when classifying CRDs stems from the poor support offered by sequence similarity. In order to validate our new classification, we therefore used all the available information associated with the considered sequences. We first showed that their alignment reveals well conserved spacing between disulphide bridges, in a way that appears to be group specific. We showed that all subgroups considered are reasonably homogenous from a structural point of view. In order to further build on this descriptive validation, we looked into the predictive power of the three newly defined groups. We did so by estimating an HMM for each group and by establishing the preference of each HMM for its group members. While this provided clear support for the notion of the homogeneity within each group, the strongest support for the new categories presented here probably stems from the capacity of their associated HMMs to identify CRDs by functional criteria, such as PLAD (type I), ligand binding capacity (type II) or small TNFR-associated CRDs (type III). Although this third category appears to be the most heterogeneous, detailed structural analysis reveals that all its members exhibit a highly conserved hairpin structure known to interact with a conserved binding pocket present in both APRIL and Baff (Baff/Baff-R, Baff/BCMA, APRIL/BCMA and APRIL/TACI co-crystals). This finding is consistent with the reported capacity of several small TNFRs to cross-react (TACI, Baff-R and BCMA) with their cognate ligands (APRIL and Baff).[42] In Fig. 5, most of the N-terminal CRDs are labeled as type I, but it is worth mentioning that several of N-terminal domains could not be labeled by any of the three HMMs. This observation raises the possibility that further expansion of the classification might be needed, either by defining a new typology or by expanding the current one. Unfortunately, the T-RMSD approach cannot help here, due to a lack of structural information, but these observations might provide an incentive to prioritize these proteins for structure determination.

To conclude, our detailed validation of T-RMSD on the small GTPase and the CRD classifications provide a new basis for using structural and functional information when performing genome annotation at a detailed functional level. It must be stressed that expanding the validation reported here would be mostly hampered by the lack of functional evidence. All known large families are easily clustered using our approach, but when doing so, one will often be at a loss to estimate whether these clusters make any biological sense. We have shown here that structure-based clusters were biologically meaningful on two well studied cases and we are now planning to explore further in this direction by applying this methodology to other important protein families, such as metabolic enzymes, protein-binding domains involved in signaling pathways and DNA-binding proteins.

Materials and Methods

Hardware and software

All experiments ran on a standard 2 GHz dualcore PC with 4 Gbytes of RAM. The T-RMSD is incorporated in the T-Coffee package, an open source freeware†.

† www.tcoffee.org

Multiple sequence alignment and phylogenetic tree computation

Multiple sequence alignments were carried out using the T-Coffee[43] alignment package (version 7.68). Structure-based multiple sequence alignments were done with the 3D-Coffee mode of T-Coffee that makes it possible to combine alternative pairwise structural and sequence alignment methods.[11] 3D-Coffee was used to combine three different types of aligners: SAP (structural aligner), slow_pair (global aligner) and lalign_id_pair (local aligner). For each sequence, the corresponding PDB was used as template. This alignment was used to compute a neighbor-joining[44] phylogenetic tree with 100 bootstrap replicates using the PHYLIP package‡.[45] Phylogenetic and structural trees were represented using Phylodendron version 0.8d.

Structural clustering using DaliLite version 3.0

Structural clustering of CRDs was done with DaliLite version 3.0. The *Z*-scores obtained from pairwise comparison using DaliLite v3.0 were normalized to generate a symmetrical square matrix of distances. The matrix was then used to generate a structural tree using the NEIGHBOR executable (an implementation of Neighbour Joining and of the UPGMA methods) from the PHYLIP package.

Structural clustering method: T-RMSD

A schematic display of the T-RMSD method is presented in Fig. 1. The structural clustering is obtained by comparing equivalent C^{α}–C^{α} intramolecular distances across the MSA of the considered structures. Two distances are considered equivalent when they separate a pair of equivalent residues. The list of residue equivalences is obtained from the alignment. Given a pair of aligned sequences I and J and the two aligned residues I_c and J_c, a global distance is associated with this pair by summing the difference of distances between every pair $d(I_cI_x)$ and $d(J_cJ_x)$, where I_x and J_x are two aligned residues of the corresponding sequence (Eq. (1)). The distance is then normalized by N, the number of contributing columns (i.e. the number of ungapped position x considered)

$$D_c(I,J) = \sum_{x}^{L} |d(I_cI_x) - d(J_cJ_x)| / N$$

Using this formula, one can estimate, on position c, a distance for every pair of sequences (provide there is no gap in the alignment column corresponding to position c). A set of such distances defines a distance matrix. Such matrices can be turned into clusters using an appropriate clustering algorithm. In the context of this work, we used the UPGMA algorithm to turn the matrices into unrooted trees. With C ungapped columns, one produces C trees. These trees are combined into one unique consensus tree using the *Consense* program (PHYLIP). *Consense* produces a consensus tree and labels each node with the number of trees supporting the considered split. This value can be normalized by C to indicate the fraction of columns supporting a given node. The T-RMSD procedure is implemented in T-Coffee (version 7.68 and higher).

Dataset of GTPase protein/domains

The small GTPase RAS superfamily is currently divided into six different domain subfamilies corresponding to specific functional differences: RAB, RAS, RHO, RAN, SAR and ARF as defined by SMART and PROSITE and RGK defined only by PIRSF.[46] We used this clustering as a basis for the evaluation of the T-RMSD on this superfamily. Structures associated with small GTPases were extracted from the Protein Data Bank[47] (PDB§) and selected by several criteria related to structure quality and non-redundancy: (1) only X-ray structures were used; (2) presenting a full structural coverage (no missing or undefined regions); (3) exhibiting resolution below 2.5 Å (with one exception at 2.65 Å); and (4) showing a limited sequence/structure redundancy using a threshold of 70%.[48] We thus defined a reference set of 43 domains (Supplementary Data Table S2): 19 RAB domains, nine RAS domains, six ARF domains, five RHO domains, one RAN, one RGK and no SAR domain because no structure presented full structural coverage. We included in this dataset two additional GTPases for which domain databases exhibit disagreement: AGAP2_HUMAN and AGAP3_HUMAN (centaurin γ1 and centaurin γ3), which contain a GTPase domain (Pfam: MIRO, PROSITE: RAB, SMART: small GTPase, and Pfam: MIRO, PROSITE: RAS, SMART: small GTPase, respectively).

Dataset of TNFR protein/domains

For sequence analysis, CRDs are referred to by the gene name corresponding to the receptor they belong to and by the number corresponding to their position within the receptor architecture (N-terminal to C-terminal). Sequences of CRDs associated with TNFR were obtained from UniProt annotation (called TNFR-Cys). It includes current described domain families corresponding to TNFR associated CRDs, unidentified CRDs, and half CRDs suggested in the current structural classification. Within UniProt, a few of these half-CRDs are annotated as atypical or truncated TNFR-Cys (from human, TNFRSF4-3, TNFRSF12A-1 and TNFRSF13C-1). However, a few CRDs and half CRDs are not described precisely regarding domain database or structural classification. For such cases, sequence boundaries were determined on the basis of consensus sequences of structural modules (from human, TNFRSF6B-5, TNFRSF14-4, TNFRSF19L-1 and TNFRSF19L-3) or according to domain databases (from *Vaccinia virus*, CRME-1, CRME-2 and CRME-3). A total of 97 CRDs or half CRDs were identified in the human TNFR family.

All structures of CRDs associated with TNFRs (from *Homo sapiens*, *Rattus norvegicus* and *Vaccinia virus*) used for analysis and classification performed in this study were extracted from the PDB. The selection of PDB files corresponding to similar domains was realized on the basis of the highest sequence coverage and highest resolution. PDB files corresponding to TNFRs (gene/receptor name) used in this work are: TNFRSF1A/TNF-

‡ Felsenstein, J. (2005). PHYLIP (Phylogeny Interference Package) version 3.6. *Distributed by the author. Departement of Genome Science, University of Washington, Seattle*

§ www.pdb.org

R1, 1EXT;[49] TNFRSF4/Ox40, 2HEY;[50] TNFRSF10B/DR5, 1D4V;[51] TNFRSF12A/TWEAK-R, 2RPJ;[52] TNFRSF13B/TACI, 1XU1;[42] TNFRSF13C/Baff-R, 1OQE;[38] TNFRSF14/HVEM, 1JMA;[53] TNFRSF16/NGFR, 1SG1;[54] TNFRSF17/BCMA, 1XU2;[42] CRME/CRME, 2UWI.[55]

Structural comparison of CRD subgroups

Global structural comparison using RMSD measurement was achieved after defining the structural core block within each subgroup. These structural core blocks include positions likely to be aligned across all pairwise structural alignment constructed using SAP version 0.5.[56] RMSD values, after core blocks superimposition, were obtained with the align command of PyMOL package version 1.1‖.[57] The number of aligned residues (despite the exclusion routine of the PyMOL align command) is globally similar with standard deviations between one and three residues (Table 2). Our results are presented by subgroup: average RMSD value, standard deviation and number of position aligned for each structural core block.

HMM profile for sequence classification

Hidden Markov models (HMM) have proven to be efficient for protein classification based on statistical relevance of specific position. We used HMMer version 2.0 to generate profiles corresponding to each structural subgroup we defined.[58,59] Profiles were built using the default mode, which corresponds to local alignments regarding domain sequences and global regarding the HMM. Each profile was calibrated empirically in order to increase sensitivity of searches using default parameters provided by HMMer version 2.0. Searches using HMM profiles were achieved on a reduced dataset corresponding to CRDs associated specifically with TNFRs.

Acknowledgements

This work was supported by European Union (EU) programs TRIDENT (grant number LSHC-CT-2006-037686), 3D REPERTOIRE (grant number LSHG-CT-2005-512028) and the Spanish Ministry of Education and Science through its Consolider programme. C.N. is supported by the CRG and the Plan Nacional (BFU 00419) A.M. v.d. S. was funded, in part, by a Juan de la Cierva fellowship of the Spanish Ministry of Education and Science. The authors wish to thank referees 1 and 2 for useful comments and suggestions.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2010.05.012

‖ www.pymol.org

References

1. Sadowski, M. I. & Jones, D. T. (2009). The sequence-structure relationship and protein function prediction. *Curr. Opin. Struct. Biol.* **19**, 357–362.
2. Orengo, C. A., Michie, A. D., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH: a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
3. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins databases for the inverstigation of sequence and structure. *J. Mol. Biol.* **247**, 536–540.
4. Dietmann, S., Park, J., Notredame, C., Heger, A., Lapp, M. & Holm, L. (2001). A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.* **29**, 55–57.
5. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D. *et al.* (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D224–D228.
6. Nishikawa, K. & Ooi, T. (1974). Comparison of homologous tertiary structures of proteins. *J. Theor. Biol.* **43**, 351–374.
7. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.
8. O'Sullivan, O., Zehnder, M., Higgins, D., Bucher, P., Grosdidier, A. & Notredame, C. (2003). APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics*, **19**, i215–i221.
9. Armougom, F., Moretti, S., Keduas, V. & Notredame, C. (2006). The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, **22**, e35–e39.
10. Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34**, 1692–1699.
11. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G. & Notredame, C. (2004). 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.* **340**, 385–395.
12. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, J. S., Hotz, H. R. *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288.
13. Wennerberg, K., Rossman, K. L. & Der, C. J. (2005). The RAS superfamily at a glance. *J. Cell Sci.* **118**, 843–846.
14. Bazan, J. F. (1993). Emerging families of cytokines and receptors. *Curr. Biol.* **3**, 603–606.
15. Hehlgans, T. & Pfeffer, K. (2005). The intriguing biology of the tumour necrosis factor/tumour necrosis factor receptor superfamily: players, rules and the games. *Immunology*, **115**, 1–20.
16. Balkwill, F. (2009). Tumour necrosis factor and cancer. *Nat. Rev. Cancer*, **9**, 361–371.
17. Locksley, R. M., Killeen, N. & Lenardo, M. J. (2001). The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell*, **104**, 487–501.
18. Zhang, G. (2004). Tumor necrosis factor family ligand-receptor binding. *Curr. Opin. Struct. Biol.* **14**, 154–160.
19. Idriss, H. T. & Naismith, J. H. (2000). TNFα and the TNF receptor superfamily: structure-function relationship(s). *Microsc. Res. Tech.* **50**, 184–195.
20. Cha, S. S., Sung, B. J., Kim, Y. A., Song, Y. L., Kim, H. J., Kim, S. *et al.* (2000). Crystal structure of TRAIL-DR5 complex identifies a critical role of the unique frame insertion in conferring recognition specificity. *J. Biol. Chem.* **275**, 31171–31177.

21. Fellouse, F. A., Li, B., Compaan, D. M., Peden, A. A., Hymowitz, S. G. & Sidhu, S. S. (2005). Molecular recognition by a binary code. *J. Mol. Biol.* **348**, 1153–1162.
22. Lam, J., Nelson, C. A., Ross, F. P., Teitelbaum, S. L. & Fremont, D. H. (2001). Crystal structure of TRANCE/RANKL cytokine reveals determinants of receptor-ligand specificity. *J. Clin. Invest.* **108**, 971–979.
23. Hymowitz, S. G., Compaan, D. M., Yan, M., Ackerly, H., Dixit, V. M., Starovasnik, M. A. & de Vos, A. M. (2003). The crystal structure of EDA-A1 and EDA-A2: splice variants with distinct receptor specificity. *Structure*, **11**, 1513–1520.
24. Schultz, J., Milipetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **27**, 229–232.
25. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B., De Castro, E. *et al.* (2007). The 20 years of PROSITE. *Nucleic Acids Res.* **36**, D245–D249.
26. UniProt Consortium. (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* **37**, D169–D174.
27. Naismith, J. H. & Sprang, S. R. (1998). Modularity in the TNF-receptor family. *Trends Biochem. Sci.* **23**, 74–79.
28. Bodmer, J. L., Schneider, P. & Tschopp, J. (2002). The molecular architecture of the TNF superfamily. *Trends Biochem. Sci.* **27**, 19–26.
29. Hancock, J. F. (2003). Ras proteins: different signals from a different locations. *Nat. Rev. Mol. Cell Biol.* **4**, 373–384.
30. Etienne-Manneville, S. & Hall, A. (2002). Rho GTPases in cell biology. *Nature*, **420**, 529–635.
31. Stenmark, H. (2009). Rab GTPases as coordinators of vesicle traffic. *Nat. Rev. Mol. Cell Biol.* **10**, 513–525.
32. Sazer, S. & Dasso, M. (2000). The RAN decathlon: multiple roles of RAN. *J. Cell Sci.* **113**, 1111–1118.
33. Kahn, R. A., Cherfils, J., Elias, M., Lovering, R. C., Munro, S. & Schurmann, A. (2006). Nomenclature for the human Arf family of GTP-binding proteins: ARF, ARL, and SAR proteins. *J. Cell Biol.* **172**, 645–650.
34. Donaldson, J. G. & Honda, A. (2005). Localization and function of Arf family GTPases. *Biochem. Soc. Trans.* **33**, 639–642.
35. Kelly, K. (2005). The RGK family: a regulatory tail of small GTP-binding proteins. *Trends Cell Biol.* **15**, 640–643.
36. Vetter, I. R. & Wittinghofer, A. (2001). The guanine nucleotide-binding switch in three dimensions. *Science*, **294**, 1299–1304.
37. Pasqualato, S., Renault, L. & Cherfils, J. (2002). Arf, Arl, Arp and Sar proteins: a family of GTP-binding proteins with a structural device for 'front-back' communication. *EMBO Rep.* **3**, 1035–1041.
38. Liu, Y., Hong, X., Kappler, J., Jiang, L., Zhang, R., Xu, L. *et al.* (2003). Ligand-receptor binding revealed by the TNF family member TALL-1. *Nature*, **423**, 49–56.
39. Chan, F. K., Chun, H. J., Zheng, L., Siegel, R. M., Bui, K. L. & Lenardo, M. J. (2000). A domain in TNF receptors that mediates ligand-independent receptor assembly and signaling. *Science*, **288**, 2351–2354.
40. Chan, F. K. (2000). The pre-ligand binding assembly domain: a potential target of inhibition of tumour necrosis factor receptor function. *Ann. Rheum. Dis.* **59**, i50–i53.
41. Chan, F. K. (2007). Three is better than one: pre-ligand receptor assembly in the regulation of TNF receptor signaling. *Cytokine*, **37**, 101–107.
42. Hymowitz, S. G., Patel, D. R., Wallweber, H. J., Runyon, S., Yan, M., Yin, J. *et al.* (2005). Structures of APRIL-receptor complexes: like BCMA, TACI employs only a single cysteine-rich domain for high affinity ligand binding. *J. Biol. Chem.* **280**, 7218–7227.
43. Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.
44. Nei, M. & Saitou, N. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
45. Felsenstein, J. (2005). PHYLIP (Phylogeny Interference Package) version 3.6. Distributed by the author. Departement of Genome Science, University of Washington, Seattle.
46. Nikolskaya, A. N., Arighi, C. N., Huang, H., Barker, W. C. & Wu, C. H. (2007). PIRSF family classification system for protein functional and evolutionary analysis. *Evol. Bioinform.* **10**, 197–209.
47. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
48. Kosloff, M. & Kolodny, R. (2008). Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins: Struct. Funct. Genet.* **71**, *891–902.*
49. Naismith, J. H., Devine, T. Q., Kohno, T. & Sprang, S. R. (1996). Structures of the extracellular domain of the type I tumor necrosis factor receptor. *Structure*, **4**, 1251–1262.
50. Compaan, D. M. & Hymowitz, S. G. (2006). The crystal structure of the costimulatory OX40-OX40L complex. *Structure*, **14**, 1321–1330.
51. Mongkolsapaya, J., Grimes, J. M., Chen, N., Xu, X. N., Stuart, D. I., Jones, E. Y. & Screaton, G. R. (1999). Structure of the TRAIL-DR5 complex reveals mechanisms conferring specificity in apoptotic initiation. *Nat. Struct. Biol.* **6**, 1048–1053.
52. He, F., Dang, W., Saito, K., Watanabe, S., Kobayashi, N., Güntert, P. *et al.* (2009). Solution structure of the cysteine-rich domain in Fn14, a member of the tumor necrosis factor receptor superfamily. *Protein Sci.* **18**, 650–656.
53. Carfí, A., Willis, S. H., Whitbeck, J. C., Krummenacher, C., Cohen, G. H., Eisenberg, R. J. & Wiley, D. C. (2001). Herpes simplex virus glycoprotein D bound to the human receptor HveA. *Mol. Cell*, **8**, 169–179.
54. He, X. L. & Garcia, K. C. (2004). Structure of the nerve growth factor complexed with the shared neurotrophin receptor p75. *Science*, **304**, 870–875.
55. Graham, S. C., Bahar, M. W., Abrescia, N. G., Smith, G. L., Stuart, D. I. & Grimes, J. M. (2007). Structure of CrmE, a virus-encoded tumour necrosis factor receptor. *J. Mol. Biol.* **372**, 660–671.
56. Brown, N. P., Orengo, C. A. & Taylor, W. R. (1996). A protein structure comparison methodology. *Comput. Chem.* **20**, 359–380.
57. DeLano, W.L. (2002). The PyMOL Molecular Graphics System on World Wide Web www.pymol.org.
58. Durbin, R., Eddy, S., Krogh, A. & Mitchinson, G. (1998). *The theory behind profile HMMs.* Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, pp. 101–133, Cambridge University Press, Cambridge, UK.
59. Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.