BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 2 2014, pages 274–281
doi:10.1093/bioinformatics/btt379

Sequence analysis

Advance Access publication July 4, 2013

Rapid similarity search of proteins using alignments of domain arrangements

Nicolas Terrapon[1,†,‡], January Weiner[2,†], Sonja Grath[1,§], Andrew D. Moore[1] and Erich Bornberg-Bauer[1,*]

[1]Westfalian Wilhelms University, Institute of Evolution and Biodiversity, Huefferstr. 1, 48149 Muenster, Germany and [2]Max Planck Institute for Infection Biology, Charitéplatz 1, 10117 Berlin, Germany

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Homology search methods are dominated by the central paradigm that sequence similarity is a proxy for common ancestry and, by extension, functional similarity. For determining sequence similarity in proteins, most widely used methods use models of sequence evolution and compare amino-acid strings in search for conserved linear stretches. Probabilistic models or sequence profiles capture the position-specific variation in an alignment of homologous sequences and can identify conserved motifs or domains. While profile-based search methods are generally more accurate than simple sequence comparison methods, they tend to be computationally more demanding. In recent years, several methods have emerged that perform protein similarity searches based on domain composition. However, few methods have considered the linear arrangements of domains when conducting similarity searches, despite strong evidence that domain order can harbour considerable functional and evolutionary signal.

Results: Here, we introduce an alignment scheme that uses a classical dynamic programming approach to the global alignment of domains. We illustrate that representing proteins as strings of domains (domain arrangements) and comparing these strings globally allows for a both fast and sensitive homology search. Further, we demonstrate that the presented methods complement existing methods by finding similar proteins missed by popular amino-acid–based comparison methods.

Availability: An implementation of the presented algorithms, a web-based interface as well as a command-line program for batch searching against the UniProt database can be found at http://rads.uni-muenster.de. Furthermore, we provide a JAVA API for programmatic access to domain-string–based search methods.

Contact: terrapon.nicolas@gmail.com or ebb@uni-muenster.de

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

Received on December 30, 2012; revised on June 12, 2013; accepted on June 27, 2013

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡Present address: Architecture et Fonction des Macromolécules Biologiques, Aix-Marseille Université, CNRS UMR 7257, 162 av. de Luminy, 13288 Marseille, France.
§Present address: Department of Biology II, University of Münich (LMU), Grosshaderner Str. 2, 82152 Planegg-Martinsried, Germany.

1 INTRODUCTION

Accurate identification of sequence similarity is a crucial and integral part of modern biological analysis. Sequence similarity is often used to infer functional similarity when comparing an unknown protein sequence to a set of well described sequences and can be the result of conservation or convergent evolution. Homologous sequences are often identified by comparing a query sequence with a large collection of sequences (subjects) in public databases. In such comparisons, a pairwise alignment attempts to match the homologous characters (nucleotides or amino-acid residues) of query and subject sequences according to an evolutionary model. Pairwise alignments can be evaluated in terms of similarity or distance scores (Spang and Vingron, 1998). Furthermore, the probability to obtain an identical or better score by chance alone is computed (Karlin and Altschul, 1993) and a threshold is chosen to accept or reject potential homology.

Given an underlying evolutionary model, calculating the optimal pairwise alignment is possible (Needleman and Wunsch, 1970) but computationally demanding. Therefore, many similarity searches use heuristic approaches. Table 1 shows an overview of the main methods used for sequence homology search.

Local alignment algorithms such as BLAST (Altschul *et al.*, 1990) focus on conserved regions only, and hence do not enforce an alignment of entire sequences. This approach has advantages in finding homology. First, homologous proteins can have regions that are not similar, for example, if these regions have not been under selectional constraints or are fast-evolving. Second, proteins often have regions of local similarity, which correspond to protein domains. Finally, heuristics are much faster because only stretches of local similarities are considered (Altschul *et al.*, 1990).

However, using local alignments also bears drawbacks. The most obvious drawback is that multi-domain proteins that share one domain may exhibit a high degree of local similarity, which can mask an overall low degree of global similarity (Song *et al.*, 2008). Ergo, applying a local alignment approach for finding functional homologs may be misleading. In contrast, global similarity suffers the opposite extreme: remote homologs of a query can be missed owing to the low amount of global sequence similarity. To address these problems, the protein's domain composition or *domain arrangement* (i.e. the N to C terminal order of domains) can be taken into account (Table 1). Protein domains are the units of protein structure, function and evolution

Table 1. Overview of selected algorithms used for protein homology search

Algorithm	Type	Reference
Optimal sequence alignment algorithms		
Needleman-Wunsh	Global alignment algorithm	Needleman and Wunsch (1970)
Smith–Waterman	Local alignment algorithm	Smith and Waterman (1981)
Heuristic sequence alignment algorithms		
FASTA	Fast on-the-fly heuristic sequence similarity search	Pearson and Lipman (1988)
BLAST	Heuristic search algorithm relying on a preformatted database	Altschul *et al.* (1990)
DELTA-BLAST	Domain enhanced lookup time accelerated BLAST	Boratyn *et al.* (2012)
Profile-based sequence alignment algorithms		
PSI-BLAST	Profile-based BLAST	Altschul *et al.* (1997)
HMMER	Profile-based search with HMMs	Eddy (1998)
JACKHMMER	Iterative profile-HMM search	Johnson *et al.* (2010)
Domain-based algorithms		
CDART	Number of domains common to two proteins	Marchler-Bauer *et al.* (2007)
DomainTeams	Local microsyntenies irrespective of protein boundaries	Pasek *et al.* (2005)
Domain Distance	Number of mismatches between two arrangements	Björklund *et al.* (2005)
RASPODOM	Dynamic-programming–based algorithm for search of circularily permutated homologs	Weiner *et al.* (2005)
DAhunter, WDAC	Measure based on a weighted sum of three similarity scores	Lee and Lee (2008, 2009)
Neighbourhood correlation	Combination of local sequence similarity and domain information	Song *et al.* (2008)
RADS/RAMPAGE	Dynamic-programming–based algorithm for domain alignment and global protein alignment	*This study*

[see Moore *et al.* (2008) for a review]. Tools that consider domains to find functionally similar, or even homologous, proteins, do exist. For example, CDART (Geer *et al.*, 2002) measures similarity by considering the number of domains shared between two proteins. CDART does not consider the order of domains, the number of repeats or the non-shared domains when determining similar proteins; in essence, CDART treats proteins as a 'bag of domains'. Several approaches have been developed that attempt to circumvent the problems with local similarity, which modern search and alignment algorithms face. Song *et al.* (2008) defined local sequence similarity scores with the analysis of a local neighbour similarity network into a 'Neighbourhood Correlation Score', thus extending the local similarity score to include a global perspective. Lin *et al.* (2006) defined a similarity measure between proteins as the weighted average of three similarity scores, which consider the number of shared domains and the number of domains that are specific for one protein, domain duplications and the local order of domains (incidence of ordered domain pairs). A similar approach is used by DAhunter and WDAC, which also incorporates domain versatility and domain frequency into the search for similar domain arrangements (Lee and Lee, 2008, 2009). Björklund *et al.* (2005) defined the domain distance, a protein distance measure based on the number of domains that are not matched between two proteins. In essence, the domain distance is an edit distance as it measures the number of edit operations (domain additions or deletions) needed to move from one arrangement to another.

In addition to similarity search, domain-based approaches have been also used for protein clustering [e.g. Enright and Ouzounis (2000)]. The DomainTeams tool, for example, analyses domainwise local similarities (microsyntenies) on bacterial chromosomes, irrespective of gene boundaries. By using microsyntenies, DomainTeams is able to detect clusters of orthologous genes even if domain arrangements are variable (Pasek *et al.*, 2005).

Several approaches demonstrated that domain content can be used to reconstruct phylogenies (Werren *et al.*, 2010). Furthermore, beyond domain content, the order of domains in a protein can help elucidate functional and structural constraints (Kummerfeld and Teichmann, 2009). While aforementioned domain-based algorithms are promising, to date none takes advantage of such a context-based analysis, i.e. the knowledge that the linear combinations of domains bear a strong phylogenetic and functional signal.

Here, we present two algorithms for finding homologous proteins that exploit this signal of conserved domain arrangements. The first algorithm, RADS (**R**apid **A**lignment of **D**omain **S**trings), represents proteins as domain arrangements and does not require any sequence information. The similarity between two proteins is determined by aligning the proteins' domain arrangements to each other using a dynamic programming algorithm. A key advantage to this approach is the reduction in time complexity—while proteins in UniProt contain on average 324 amino acids, they harbour an average of only 1.5 domains (2.63 for multi-domain arrangements), making the number of required comparisons considerably smaller. Further, by only considering proteins that share at least one common domain, the search space can be drastically reduced. While abandoning the information contained within the sequences does come at the cost of sensitivity, it is well established that domains are the units of protein evolution and are a good proxy for sequence similarity (Buljan and Bateman, 2010; Chothia and Gough, 2009;

Hunter *et al*., 2012; Moore *et al*., 2008; Sjölander *et al*., 2011; Wang and Caetano-Anollés, 2009). It follows that the alignment of domains can be used for swift detection of potential homologs.

The second presented algorithm, RAMPAGE (**R**apid **A**lignment **M**ethod of **P**roteins based on domain **A**rran**GE**ments) complements RADS and addresses the need for increased sensitivity. RAMPAGE creates global alignments of amino-acid sequences using the domainwise alignments provided by RADS as guideline, similar to previously described approaches such as the segment-based approach applied in DIALIGN-T (Subramanian *et al*., 2005). We demonstrate that methods searching for domain arrangements yield biologically meaningful results, which are often complementary to sequence-based methods and, at the same time, work at a speed that is significantly faster than local alignment tools.

2 MATERIALS AND METHODS

2.1 Data

For domain annotation, we used the current version (26.0) of the Pfam database (Punta *et al*., 2012). Pfam is based on a sequence database—*Pfamseq*—which is currently based on UniProt release 2011_06 [**Uni**versal **Prot**ein Resource, UniProt Consortium (2012)]. The Pfam database consists of two sections: Pfam-A (manually curated) and Pfam-B (automatically defined). Throughout the complete study, we used Pfam-A annotations. All sequences have been annotated with domains using HMMER 3.0 (Eddy, 2011). The current data consist of 15 929 002 non-redundant proteins, 79.4% (11 769 563) of which are assigned to at least one Pfam domain.

2.2 Algorithms

2.2.1 RADS—Rapid Alignment of Domain Strings In the first approach, proteins are represented by domain arrangements, i.e. N- to C-ordered strings of domains. Subsequently, domain arrangements are compared using a version of the Needleman–Wunsch (NW) algorithm (Needleman and Wunsch, 1970) with values for match (same domain ID), mismatch (different domain IDs) and affine gap costs. In the RADS implementation with default parameters, a matched pair of domains receives a score of 150, a mismatch or internal gap opening a score of −100, and the cost for internal gap extension is set to −50. The costs for terminal gap opening is set to −50, and for terminal gap extension to −25. Low terminal gap penalties correspond to the finding that indels are more frequent on protein termini (Weiner *et al*., 2006). Additionally, we introduce a score normalization scheme that considers the following:

- Domain length, because a match between two larger domains corresponds to a conservation between longer stretches of amino acids. The match score between domains is weighted by the length of the shortest domain occurrence within the pair, and therefore, longer matching domains get higher scores.
- Difference in length between domains, as strong variation in size might indicate functional divergence. For example, if, in a given protein, only a short fragment of a certain domain is found, the extent to which this fragment is still fully functional is uncertain. Furthermore, the fragmentation of a domain can hold phylogenetic signal. Ergo, matching a fragment to a more complete domain receives a lower score than two complete domains or two fragmented domain matches.

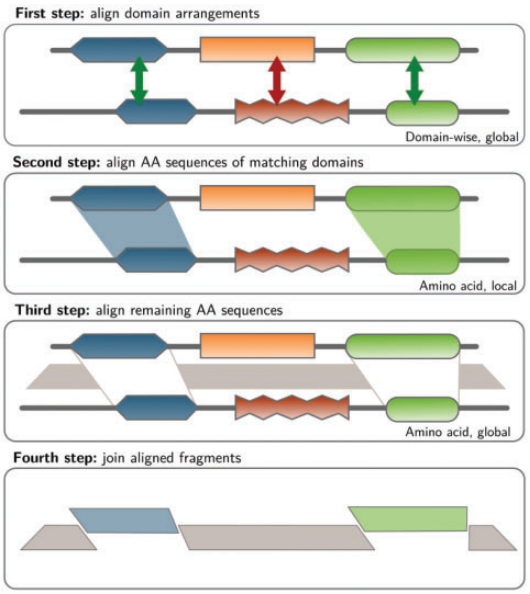


Fig. 1. Scheme of the domain-driven amino-acid alignment (RAMPAGE). In the first step, the matching homologous domains are identified by RADS. Second, local amino-acid sequence alignments of the matching domains are created. In the third step, all remaining fragments of the protein sequences are aligned globally. Finally, all subalignments are assembled to produce a full alignment

The normalized match score $S^*_{(d_1,d_2)}$ between the two domains d_1 and d_2 is defined as follows:

$$S^*_{(d_1,d_2)} = S_{(d_1,d_2)} \cdot \frac{L_{min}(d_1,d_2)}{\bar{x}_{(L)}} \cdot \frac{L_{min}(d_1,d_2)}{L_{max}(d_1,d_2)}, \quad (1)$$

where S_{d_1,d_2} is the original score of the matching domains d_1 and d_2, $\bar{x}_{(L)}$ is the average length of a Pfam domain in UniProt and is used to have a per-position score in matching domains and $L_{min}(d_1,d_2)$ and $L_{max}(d_1,d_2)$ represent the short and long domain in a matching pair, respectively. We find that the application of this simple normalization scheme improved the classification power of RADS (Supplementary Material S1).

2.2.2 RAMPAGE—Rapid Alignment Method of Proteins based on domain ArranGEments RAMPAGE produces full sequence-based alignments between query and hits based on a preceding RADS search. This can be relevant if large stretches between proteins are not aligned by RADS, such as may result from poor domain-annotation coverage or strong sequence divergence. The procedure is described in Figure 1. After conducting a domain string alignment using RADS, the pairs of sequence fragments corresponding to matched domains are locally aligned using the Smith–Waterman algorithm (Smith and Waterman, 1981). The remaining fragments are terminal sequences and inter-match fragments. These fragments are now aligned using the global NW algorithm. Finally, all alignments are joined together to give a full global alignment.

2.3 Estimation of the false-negative rate

To estimate the probability that two proteins contain a detectable region of similarity without sharing any domains, we randomly selected one million pairs of proteins from UniProt that did not share any domain. For each pair, we searched for homologous regions using the program bl2seq from the BLAST package. Furthermore, for each pair, we

recorded information about the number of domains and the fraction of sequences not assigned to any domain to test for possible bias.

2.4 Correlation of alignment scores for NW with RADS and RAMPAGE

To test for a possible influence of incorrect domain matches introduced by the RADS algorithm on the overall alignment quality, we determined correlations between RADS and RAMPAGE scores and NW scores using the Spearman's rank correlation. We tested different datasets. First, correlations were assessed for 10 000 pairs of randomly selected sequences from UniProt. Each pair consisted of two proteins that had at least one domain in common. Second, we randomly selected pairs of proteins that share a certain number (one, two, three, four or more than four) of domains. Each group contained 100 protein pairs. The second dataset allowed us to illustrate the impact of arrangement length on RADS and RAMPAGE.

2.5 Receiver-operating-characteristics curves

We compared the aforementioned approaches for detecting homology: (i) based on domain annotation: RADS, CDART, domain distance and DAhunter; (ii) based on amino-acid sequence alignement: RAMPAGE, BLAST and DELTA-BLAST; (iii) based on profile search: PSIBLAST and JACKHMMER. For each of these methods, we calculated receiver-operating-characteristics (ROC) curves to evaluate their performance in retrieving close orthologs. We used data from the OMA (**O**rthologous **MA**trix) project (Altenhoff *et al.*, 2011) obtained via the OMA browser (http://omabrowser.org/). The OMA browser offers orthologous relationships for proteins from more than 1000 species. We were able to match 3 322 654 sequences from the OMA browser (version May 2011) to UniProt identifiers (unique protein IDs). Then, we randomly selected 1000 proteins and performed searches with all six algorithms. For each algorithm, we determined for each hit whether it appears with the query in the list of homologous pairs (http://omabrowser.org/All/oma-pairs.txt.gz); if so, the hit was counted as true positive, otherwise as false positive. For a given algorithm, ROC curves for both single- and multi-domain queries were then derived from the pooled and sorted results of all searches. We further tested performance of the different methods against more difficult datasets, notably with short repeated domains or long amino-acid stretches without domain annotation.

2.6 Implementation

The core algorithms RADS and RAMPAGE are implemented in ANSI C, and are provided as a command-line application, which can be used to scan any type of non-overlapping domain annotation against any custom database. For convenience, we implemented a web application that can be used to perform searches against UniProt using Pfam domain definitions (Supplementary Material S2). Domain arrangements of UniProt proteins were precomputed and indexed to ensure swift identification of proteins that share at least one domain as only these are further considered in RADS and RAMPAGE searches (not true for BLAST). For database access, query and management as well as for visualization purposes, a set of Perl modules was developed. To achieve maximum portability, simplicity and speed, the databases were created using SQLite and stored on a solid state drive. An average RADS web interface search against the approximately 15 000 000 proteins of PfamSeq takes ~1.5 s on an Intel Core i3-2120 machine (3.30 GHz, 4 GB RAM). An additional 0.3 s are required if the query must first undergo domain annotation. The web interface for RADS and RAMPAGE searches was combined with an interface for BLAST and an internal HMMSCAN pipeline (Eddy, 2011), making parallel sequence- and domain-based searches as well as on-the-fly sequence annotation possible. This allows users to specifically target results that were missed by one of the search algorithms. Results can be sorted by RADS, RAMPAGE or BLAST scores and are presented along with a graphical representation of domain arrangements. Because the underlying UniProt release corresponds to the Pfam release (used for domain annotation), updates of the web interface will follow Pfam releases.

Besides the C command-line application and the web interface, we provide an initial JAVA archive, which can be used as a command-line application to batch-query the web interface or as a JAVA API for programmatic access to the web interface. We have recently incorporated the JAVA API into DoMosaics, a graphical tool for domainwise analysis of proteins (Moore *et al.*, 2013).

The web interface, the command-line implementation as well as the JAVA jar can be obtained under http://rads.uni-muenster.de. DoMosaics is available under http://www.domosaics.uni-muenster.de. All programs and libraries can be used on all major platforms and are freely availability under the GNU public licence.

3 RESULTS AND DISCUSSION

3.1 Impact of domain misannotation

Domain-based similarity search algorithms can only be applied to pairs of proteins that have at least one domain in common. As one of the first steps of RADS, the search space is reduced by removing all proteins that do not share any domains with the query. Domains that are wrongly annotated (or wrongly absent) (Beaussart *et al.*, 2007; Terrapon *et al.*, 2009) may lead to filter out sequences that harbour significant similarity to the query. To quantify this possible error, we randomly selected one million sequence pairs that do not share any domain and aligned them. Less then 0.01% (60 pairs) exhibited similarity at an *e* value threshold $<10^{-5}$, a widely used default threshold for database searches; even relaxing to a threshold of 10^{-3} results in <0.1% error (682 pairs). Moreover, distinct domain families belonging to the same Pfam clan explain 24 cases out of 60 at 10^{-5} (and 92 cases out of 682 at 10^{-3}). As the definition of subfamilies is a central aspect of the recent Pfam annotation strategy, we believe such errors will become increasingly rare. Ergo, the error caused by the initial filtering step in which only sequences are retained that share at least one domain with the query is negligible.

3.2 Correlation of sequence- and domain-based alignments

Next, we wanted to know whether domain-based alignments are reliable in the sense that they reflect an exact global amino-acid alignment. To investigate this, we tested if scores from an alignment of domain arrangements are correlated with scores from pairwise amino-acid sequence alignments. There are at least three possible explanations why scores from an alignment of domain arrangements may not perfectly correlate with scores from pairwise amino-acid sequence alignments. First, by aligning domains as characters (as opposed to aligning amino acids as characters), one might expect a lower resolution in closely related proteins. At the domain level, closely related proteins tend to have similar, if not identical, arrangements (Forslund *et al.*, 2011). In comparison, closely related proteins can still exhibit variation at the sequence level, which can provide higher resolution in elucidating the exact relationships. Hence, aligning domains alone makes it difficult to correctly rank close homologs, as they may share the

same arrangements and hence the same score. Second, a domain-based homology search can provide relevant results for divergent proteins by the use of accurate annotations delivered through domain family profiles. Such remote domain homologs are usually not recognized with amino-acid alignments (Gonzalez and Pearson, 2010). Third, domain arrangements might be incorrectly aligned, e.g. matching domains are not aligned optimally or different domains are wrongly aligned with each other. In this case, the score of a domain-based homology search will strongly differ from the score obtained in an amino-acid sequence comparison.

To test the correlation of domain-based and sequence-based scores, we first ran RADS on random protein pairs. Next, applying RAMPAGE, we build global amino-acid alignments from domainwise RADS searches (Fig. 1 and 'Materials and Methods' section). Because the alignment reported by RAMPAGE is a suboptimal solution of the NW alignment for a pair of sequences, the maximum score that can be achieved by RAMPAGE will be the NW score. If RADS misaligns domains, then scores from RAMPAGE should be significantly worse than scores reported by the NW algorithm. We show that both RADS and RAMPAGE scores correlate significantly with NW scores on the 10 000 protein-pair set (Spearman's rank correlation, RADS versus NW: $\rho = 0.42$, $P < 0.001$; RAMPAGE versus NW: $\rho = 0.98$, $P < 0.001$). Additionally, we investigated the effect of domain number per protein by using a second set of random domain pairs. As to be expected, we find that correlations for RADS are higher when more domains are shared between the protein pairs. This indicates that RADS performs better when applied to multi-domain searches than to single-domain searches (see Supplementary Material S1 for details).

3.3 Using ROC curves to evaluate algorithm performance

We used the OMA database to evaluate the performance of RADS, RAMPAGE and other methods (see 'Materials and Methods' section for details). Whereas BLAST and RAMPAGE generally perform equally well, we find an overall lower performance for RADS and the other domain-based methods. An in-depth manual investigation of the results shows that the main problem with pure domain-based approaches is the low resolution resulting for queries which contain only a single annotated domain (Fig. 2, top). Domain-based methods treat single domain hits almost equally if they consist of the same domain arrangement as the query. Again, this indicates that domain-based approaches are better suited for multi-domain proteins. Indeed, for proteins that contain at least two domains, RADS and the other domain-based methods exhibit increased performance (Fig. 2, bottom). A manual analysis of the results reveals that the domain-based approaches achieve their better performance not necessarily by providing exactly the same results as BLAST, but rather by being able to detect distant homologs, which are missed by BLAST. While the discriminative performance of the domain-based methods grows with the number of domains in the query, the reverse is true for BLAST. The high predictive power of domain-based approaches has been further demonstrated using proteins with long non-repeated domains. In contrast, proteins with short repeated domains challenge all

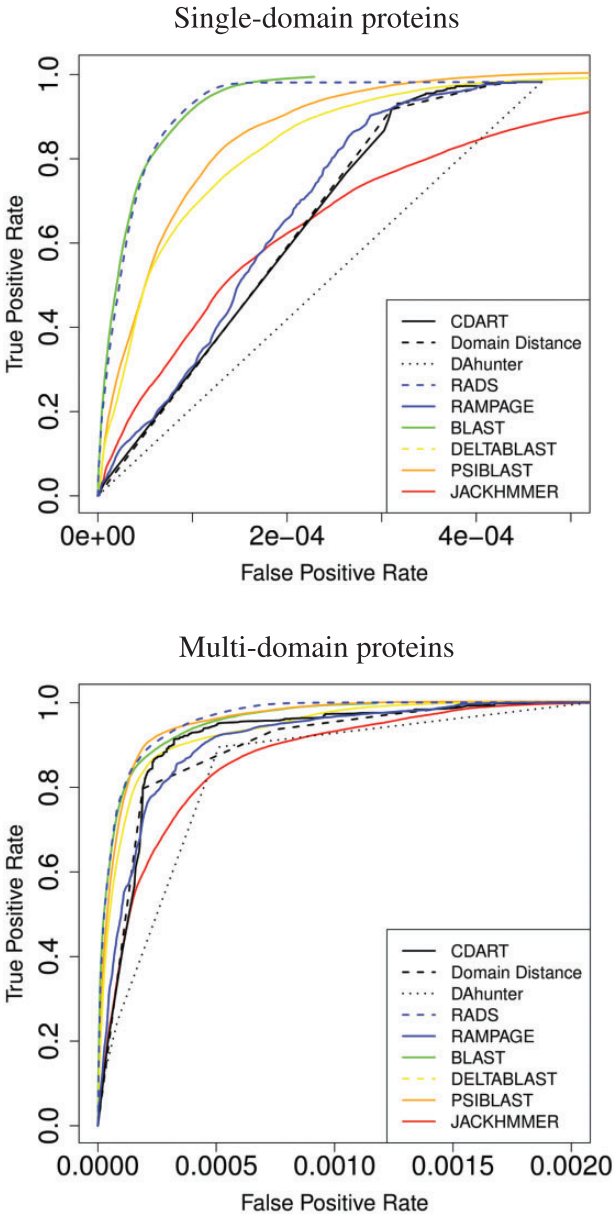
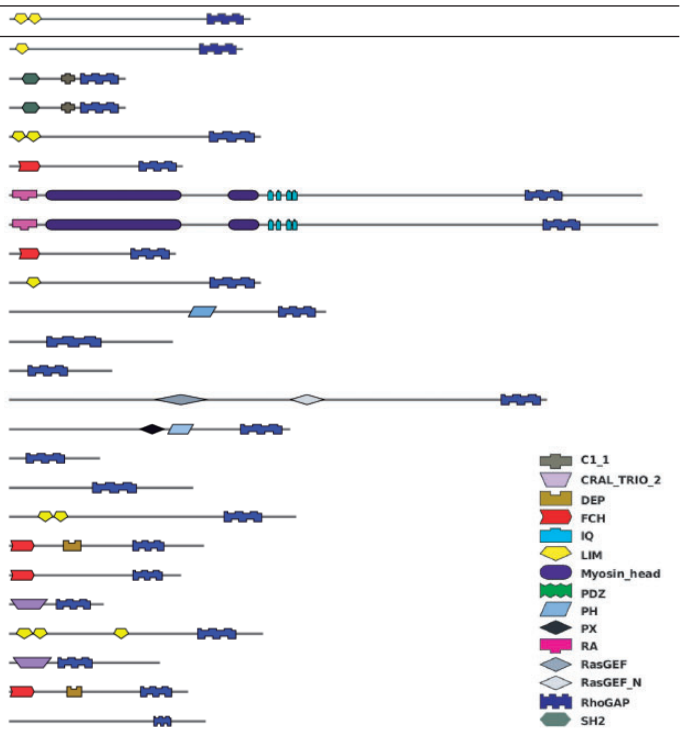


Fig. 2. ROC curves for sequence- and domain-based approaches. ROC curves are obtained from ortholog-group assignments for 1000 random query proteins in the OMA database

tested methods (Supplementary Material S3.1). We additionally scrutinized the performance of methods when applied to proteins with long stretches of amino acids that lack domain assignment (Supplementary Material S3.2). Here, we find that all methods achieve a good sensitivity/specificity trade-off, even if such a stretch is found in the center of a given protein and splits a domain arrangement. For the latter, we find that RAMPAGE performs well despite its global-alignment strategy. It is important to note that these ROC analysis are biased in favour of sequence-based methods, as the definition of orthologous groups in the OMA database is based on BLAST results.

Table 2. Combined BLAST and RADS search with RGA3_SCHPO as query

Identifier	BLAST *e* value	RADS score	Domain Arrangement
RGA3_SCHPO	0.0	379	
RGA4_SCHPO	1e-23	248	
CHIN_HUMAN	7e-18	118	
CHIN_MOUSE	6e-17	118	
RGA2_YEAST	4e-13	326	
RGA7_SCHPO	2e-12	128	
MYO9A_HUMAN	5e-12	30	
MYO9A_MOUSE	5e-12	30	
RGD1_YEAST	2e-11	118	
RGA1_YEAST	2e-10	213	
RGA2_SCHPO	6e-08	135	
SAC7_YEAST	1e-04	-4	
BAG7_YEAST	0.001	53	
BEM2_YEAST	0.002	168	
BEM3_YEAST	0.004	135	
RGA5_SCHPO	0.12	58	
RGA6_SCHPO	0.16	32	
RGA1_SCHPO	9.6	357	
RGA8_SCHPO	> 10	94	
RGA9_SCHPO	> 10	57	
YIL2_SCHPO	> 10	116	
LRG1_YEAST	> 10	297	
ECM25_YEAST	> 10	125	
RGD2_YEAST	> 10	84	
YHY2_YEAST	> 10	-102	

Shown are all RhoGAP-proteins in *S.pombe* and *S.cerevisiae*, additional human and mouse chimaerins (CHIN_HUMAN and CHIN_MOUSE) and non-typical 9α myosins (MYO9_HUMAN and MYO9_MOUSE). RGA1_SCHPO—the best RADS hit—is missed by BLAST. In contrast, BLAST reports the human and mouse chimaerins as highly significant hits.

3.4 Case studies

Next, we investigated hits where RADS obtains a high score, while BLAST does not show a significant hit or, vice versa, hits where RADS reports a poor score, while BLAST reports significant similarity. In the following, we demonstrate two selected examples.

3.4.1 Yeast Rho-GTPase-activating proteins The yeast species *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* both contain paralogs of yeast Rho-GTPase-activating (RGA) proteins. RGA proteins are characterized by an N-terminal LIM domain, a repetitive central region and a C-terminal RhoGAP domain (Table 2). We performed a BLAST search (*e* value $<10^{-3}$) against UniProt using the protein RGA3_SCHPO as query. As the RGA protein family exhibits a high degree of sequence divergence (Tcherkezian and Lamarche-Vane, 2007), BLAST does not find all subjects with a similar or identical domain arrangement. In contrast, some proteins in human and mouse with a different function and domain arrangement such as chimaerins (CHIN_HUMAN and CHIN_MOUSE) or non-typical 9α myosins (MYO9_HUMAN and MYO9_MOUSE) are reported as significant BLAST hits owing to high similarity within the RhoGAP domain. BLAST correctly identifies several fungal proteins with a domain arrangement similar to that of the query (e.g. *RGA4_SCHPO* and *RGA2_YEAST*), but fails to identify distant homologs. BLAST does not report, for instance, the RGA1_SCHPO protein, which has exactly the same arrangement as the query and comes from the same species *S.pombe*. Similarly, BLAST fails to identify the *S.cerevisiae* protein LRG1_YEAST, which exhibits an arrangement that only differs from the query by an additional third LIM domain. In contrast, RADS reports this protein and other identical arrangements with high scores and is therefore able to find more remote homologs. The identification of remote homologs that escape BLAST searches but have similar domain arrangements is of high importance because they are usually considered to have a more similar function (Gerstein and Hegyi, 2001). We report *e* values and RADS scores for the query and relevant hits in Table 2 (an exhaustive list is provided in the Supplementary Material S4.1).

3.4.2 Phosphomannomutases Phosphomannomutases are a large family of enzymes containing a highly conserved domain

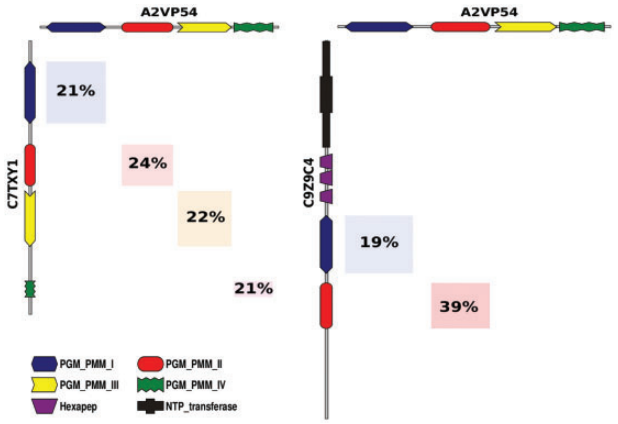


Fig. 3. Example results of a combined RADS/BLAST search in enzyme family of phosphomannomutases illustrating the differences between the two algorithms. **Left:** comparison between the original query (A2VP54_MYCTU) and a protein with the same domain arrangement (C7TXY1_SCHJA, a putative phosphoglucomutase) and high RADS score that is not detected by BLAST at an *e* value of 0.1; **Right:** comparison with a significant (e value $<10^{-6}$) BLAST hit (C9Z9C4_STRSW) with little agreement with the query in domain arrangement

arrangement of four domains (PGM_PMM I-IV, Fig. 3). Owing to high sequence divergence between members of the family, a sequence-based search using one family member as query may not necessarily find the entire family. With RADS or other domain-based approaches, all conserved arrangements are found in one step, and will furthermore be shown as more closely related than proteins containing only a strong, but local, similarity. Starting with one Phosphomannomutase (A2VP54_MYCTU) as query, we performed a BLAST and a RADS search against the whole UniProt database. The combined search (BLAST + RADS) reports 10 396 hits, which contain all 8078 proteins with the same domain arrangement as the query. BLAST detects 6758 proteins at an *e* value cut-off of 10^{-3}, 1065 of those have a different domain arrangement. Therefore, BLAST misses 2385 (29.5%) proteins with the same domain arrangement as the query (more details in the Supplementary Material S4.2).

4 CONCLUSION

It has been shown that proteins with identical multi-domain arrangements have similar function in 90% of all cases, whereas 35% of multi-domain proteins sharing only one domain are functional homologs (Gerstein and Hegyi, 2001). We therefore consider the RADS approach to identify all identical arrangements to a given query as a fast and valuable step in determining protein similarity across large datasets. For single-domain proteins and proteins that lack domain annotation, however, sequence-based algorithms provide more relevant results as they are able to identify local similarities, which are crucial in distinguishing close homologs. RAMPAGE alleviates this problem and perform with similar sensitivity but significantly faster than BLAST (Supplementary Material S5).

We propose that RADS is complementary to sequence similarity-based approaches and that exploiting domain arrangements for similarity search is, in particular for multi-domain proteins, an important means for identifying functional homologs. Although RADS explicitly disregards sequence information at the alignment stage, the comparison of domains is nonetheless driven by sequence information, as multiple sequence alignments are used to construct the profile. Indeed, such sequence profiles are sensitive, as they in essence describe the variation at the sequence level of a large set of related proteins. This is particularly valuable when dealing with the increasing amount of available sequence data, as the ability to use the strong discrimination power of hidden Markov models (HMMs) or other profile-based methods can be used to initially reduce search space. The sensitivity of RADS/RAMPAGE is theoretically comparable with profile-based searches such as PSI-BLAST (Schaffer *et al.*, 2001), yet a sensitive similarity search can be achieved without compromising the actual search time because the computationally heavy stage is shifted to protein annotation. However, recent advancements in the HMMER package (Eddy, 2011) have made on-the-fly annotation of protein domains feasible, such that a RADS search is swift even for raw sequence queries. Several published approaches use domain composition or arrangements for homology detection (Table 1). Nonetheless RADS/RAMPAGE will be a valuable resource for the community—for several reasons. First, CDART (Geer *et al.*, 2002) does not rank results based on domain arrangement similarity. Rather, it provides the user with the number of shared domains without taking any additional domain or sequence information into account. Moreover, because CDART uses custom-defined domain definitions, results are difficult to compare with more widespread domain definitions such as Pfam. Second, the domain distance introduced by Björklund *et al.* (2005) provides a simple but efficient metric for describing the dissimilarity of domain arrangements, which is defined as the number of domains that remain unaligned in a global alignment of domain arrangements. Forslund *et al.* (2011) recently published an analogous approach where domain arrangement similarity is expressed as the number of aligned domains. To the best of our knowledge, neither of these metrics has been implemented in a publicly available resource. Third, in the most recently described method (Lee and Lee, 2009), which includes a web interface, the authors proposed a complex measure based on information theory that combines the domain composition similarity with ordered pairs and domain duplications. However, the method is 'kingdom'-dependent (Eukaryota, Bacteria, Archea), limiting the possibility of large-scale, cross kingdom comparisons. Furthermore, the scoring system as implemented has not been tested on large datasets and we were unable to reproduce the results provided by the web interface, which only reports the 10 most similar arrangements and seems to be outdated (2009). Fourth, we find that RADS/RAMPAGE perform significantly faster than all other methods tested in this study (Supplementary Material S4). Finally, we provide a set of implementations to meet the needs of various groups. We provide a fast C-based command-line application for running custom domain-string comparisons, a web interface for querying UniProt with Pfam and a command-line JAVA application for querying the web interface in batch mode, which can also be used

as a JAVA library for programmatic access. Furthermore, we have integrated the RADS/RAMPAGE search into a GUI-based tool termed DoMosaics, which is available for download from http://domosaics.uni-muenster.de (Moore *et al.*, 2013).

In summary, the complementary approaches RADS and RAMPAGE and their implementations allow for fast similarity searches using domain arrangement comparisons. The web service provides a fine-grained arrangement scoring with user-adjustable parameters and comparison with BLAST. Results can be ordered, filtered by scores and/or reduced to non-redundant arrangements. Searches can be saved and downloaded, and we are currently developing new filters to allow for taxon- and species-specific searches. To ensure that the sequence database of the web service remains current, we are using Pfamseq annotations, and have developed routines to ensure that this resource is updated with each new Pfam release. Given the large amount of sequence data that has so much become an integral part of modern biological analysis and the increasing speed of profile-based annotation methods, we believe that both RADS and RAMPAGE will prove useful in identifying homology of interest, and that it can help to put the available data to work in pursuing interesting and relevant questions in molecular biology and evolution.

ACKNOWLEDGEMENT

E.B.B. gratefully acknowledges many useful and encouraging discussions with Spices from the Janet Thronton group during a sabbatical at EBI in 2009.

Funding: DFG (Deutsche Forschungs Gemeinschaft) (BO 2445/4-1 to E.B.B.).

Conflict of Interest: none declared.

REFERENCES

Altenhoff,A.M. *et al.* (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, **39**, D289–D294.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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

Beaussart,F. *et al.* (2007) Automated improvement of domain annotations using context analysis of domain arrangements (AIDAN). *Bioinformatics*, **23**, 1834–1836.

Björklund,A.K. *et al.* (2005) Domain rearrangements in protein evolution. *J. Mol. Biol.*, **353**, 911–923.

Boratyn,G.M. *et al.* (2012) Domain enhanced lookup time accelerated blast. *Biol. Direct*, **7**, 12.

Buljan,M. *et al.* (2010) Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.*, **11**, R74.

Chothia,C. and Gough,J. (2009) Genomic and structural aspects of protein evolution. *Biochem. J.*, **419**, 15–28.

Eddy,S. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Enright,A.J. and Ouzounis,C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.

Forslund,K. *et al.* (2011) Domain architecture conservation in orthologs. *BMC Genomics*, **12**, 326.

Geer,L.Y. *et al.* (2002) CDART: protein homology by domain architecture. *Genome Res.*, **12**, 1619–1623.

Gerstein,M. and Hegyi,H. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.

Gonzalez,M.W. and Pearson,W.R. (2010) Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.*, **38**, 2177–2189.

Hunter,S. *et al.* (2012) Interpro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

Johnson,L.S. *et al.* (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.

Karlin,S. and Altschul,S.F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA*, **90**, 5873–5877.

Kummerfeld,S.K. and Teichmann,S.A. (2009) Protein domain organisation: adding order. *BMC Bioinformatics*, **10**, 39.

Lee,B. and Lee,D. (2008) DAhunter: a web-based server that identifies homologous proteins by comparing domain architecture. *Nucleic Acids Res.*, **36**, W60–W64.

Lee,B. and Lee,D. (2009) Protein comparison at the domain architecture level. *BMC Bioinformatics*, **10**, S5.

Lin,K. *et al.* (2006) An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics*, **22**, 2081–2086.

Marchler-Bauer,A. *et al.* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.*, **35**, D237–D240.

Moore,A.D. *et al.* (2008) Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.*, **33**, 444–451.

Moore,A.D. *et al.* (2013) DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics* (in press).

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Pasek,S. *et al.* (2005) Identification of genomic features using microsyntenies of domains: domain teams. *Genome Res.*, **15**, 867–874.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Punta,M. *et al.* (2012) The pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Schaffer,A.A. *et al.* (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

Sjölander,K. *et al.* (2011) Ortholog identification in the presence of domain architecture rearrangement. *Brief. Bioinform.*, **12**, 413–422.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Song,N. *et al.* (2008) Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput. Biol.*, **4**.

Spang,R. and Vingron,M. (1998) Statistics of large-scale sequence searching. *Bioinformatics*, **14**, 279–284.

Subramanian,A.R. *et al.* (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.

Tcherkezian,J. and Lamarche-Vane,N. (2007) Current knowledge of the large rhogap family of proteins. *Biol. Cell*, **99**, 67–86.

Terrapon,N. *et al.* (2009) Detection of new protein domains by co-occurrence: application to *Plasmodium falciparum*. *Bioinformatics*, **23**, 3077–3078.

UniProt Consortium. (2012) Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Res.*, **40**, D71–D75.

Wang,M. and Caetano-Anollés,G. (2009) The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure*, **17**, 66–78.

Weiner,J. 3rd *et al.* (2005) Rapid motif-based prediction of circular permutations in multi-domain proteins. *Bioinformatics*, **21**, 932–937.

Weiner,J. 3rd *et al.* (2006) Domain deletions and substitutions in the modular protein evolution. *FEBS J.*, **273**, 2037–2047.

Werren,J.H. *et al.* (2010) Functional and evolutionary insights from the genomes of three parasitoid nasonia species. *Science*, **327**, 343–348.