BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 15 2009, pages 1869–1875
doi:10.1093/bioinformatics/btp342

Sequence analysis

Rapid detection, classification and accurate alignment of up to a million or more related protein sequences

Andrew F. Neuwald

Department of Biochemistry & Molecular Biology and The Institute for Genome Sciences, University of Maryland, School of Medicine, 801 West Baltimore Street, BioPark II, Room 617, Baltimore, MD 21201, USA

Received on February 9, 2009; revised and accepted on May 27, 2009

Advance Access publication June 8, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The patterns of sequence similarity and divergence present within functionally diverse, evolutionarily related proteins contain implicit information about corresponding biochemical similarities and differences. A first step toward accessing such information is to statistically analyze these patterns, which, in turn, requires that one first identify and accurately align a very large set of protein sequences. Ideally, the set should include many distantly related, functionally divergent subgroups. Because it is extremely difficult, if not impossible for fully automated methods to align such sequences correctly, researchers often resort to manual curation based on detailed structural and biochemical information. However, multiply-aligning vast numbers of sequences in this way is clearly impractical.

Results: This problem is addressed using Multiply-Aligned Profiles for Global Alignment of Protein Sequences (MAPGAPS). The MAPGAPS program uses a set of multiply-aligned profiles both as a query to detect and classify related sequences and as a template to multiply-align the sequences. It relies on Karlin–Altschul statistics for sensitivity and on PSI-BLAST (and other) heuristics for speed. Using as input a carefully curated multiple-profile alignment for P-loop GTPases, MAPGAPS correctly aligned weakly conserved sequence motifs within 33 distantly related GTPases of known structure. By comparison, the sequence- and structurally based alignment methods hmmalign and PROMALS3D misaligned at least 11 and 23 of these regions, respectively. When applied to a dataset of 65 million protein sequences, MAPGAPS identified, classified and aligned (with comparable accuracy) nearly half a million putative P-loop GTPase sequences.

Availability: A C++ implementation of MAPGAPS is available at http://mapgaps.igs.umaryland.edu.

Contact: aneuwald@som.umaryland.edu

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

1 INTRODUCTION

Protein sequence patterns that have been conserved for a billion years or more contain implicit information regarding structural, functional and mechanistic features shared by evolutionarily related proteins. Divergent sequence patterns—i.e., patterns conserved only within descendent proteins performing a particular divergent function—likewise contain implicit information regarding functional distinctions between related proteins. Hence molecular features responsible for important functional similarities and differences can be revealed by mapping patterns of similar and divergent residues to proteins of known structure. With this in mind, we developed (Neuwald, 2007a; Neuwald *et al*., 2003) and applied (Kannan and Neuwald, 2004, 2005; Kannan *et al*., 2007; Neuwald, 2006, 2007b; Neuwald *et al*., 2003) a Bayesian approach, termed CHAIN analysis, for functionally categorizing proteins based on divergent sequence patterns [reviewed in (Neuwald, 2006)]. These patterns are interpreted by mapping them to corresponding structural features (Neuwald, 2007a).

In order to statistically characterize protein functional divergence in this way, it is important to identify and accurately align a large number of related protein sequences inasmuch as subtle, yet statistically significant patterns often become evident only within vast amounts of data. For this reason, the genomics and metagenomics initiatives will provide unprecedented statistical power to extract functional and mechanistic information regarding major protein classes. Such classes include, for example, eukaryotic protein kinases (Hanks and Hunter, 1995), glycosyltransferases (Coutinho *et al*., 2003), α,β-hydrolase fold enzymes (Holmquist, 2000), the α-family of pyridoxal-phosphate-dependent enzymes (Christen and Mehta, 2001), members of the haloacid dehalogenase superfamily (Koonin and Tatusov, 1994), and various subclasses of phosphate-binding loop (P-loop) NTPases, including ATP binding cassette (ABC) transporter-related proteins (Davidson and Maloney, 2007), P-loop kinases (Leipe *et al*., 2003), AAA+ ATPases (Neuwald *et al*., 1999), DEAD/H helicases (Bork and Koonin, 1993) and P-loop GTPases (Leipe *et al*., 2002).

Unfortunately, highly accurate alignment of a vast number of sequences is often very difficult to achieve due to residue dissimilarities associated with protein functional divergence, short insertions and deletions, and larger sequence restructuring events. Such restructuring events include, for example: (i) the insertion of one domain within another domain (as occurs, for instance, in structural maintenance of chromosomes (SMC) proteins, where a long-coiled coil domain is inserted within an ABC-like 'head' domain) (Melby *et al*., 1998); (ii) the obscuring of a protein domain at the sequence level through the incorporation of 'inteins' (Pietrokovski, 2001)—'protein introns' that can break up a domain-encoding sequence at multiple sites and that can excise themselves to produce a functional domain; (iii) rare insertions within an otherwise conserved secondary structural element (such as the insertion of a

β-hairpin loop within an α-helix of the enhancer binding protein PspF) (Rappas *et al*., 2005); and (iv) domain circular permutation (as occurs in certain P-loop GTPases, such as YjeQ) (Shin *et al*., 2004). Moreover, many multiple alignment programs rely on heuristic procedures that—though helpful in improving the alignment of small sets of sequences—can have undesirable side effects. For example, often such methods will multiply align randomly generated sequences, which is clearly incorrect from both an evolutionary and a statistical perspective.

Profile-based alignment methods, such as reverse-position-specific BLAST, a variant of PSI-BLAST (Altschul *et al*., 1997), and HMM-aligners, such as the hmmsearch and hmmalign programs within the HMMer package, improve the alignment process by using a profile derived from a curated multiple sequence alignment. Here I build upon these approaches by describing Multiply-Aligned Profiles for Global Alignment of Protein Sequences (MAPGAPS). MAPGAPS uses a multiple-profile alignment to 'map the gaps' (i.e. the insertions and deletions, both large and small) between distantly related proteins. The multiple-profile alignment serves both as a query for detecting and classifying related sequences and as a template for globally aligning the sequences to each other.

Creating and maintaining multiple-profile alignments and searching with them in this way has several advantages. In particular, this facilitates rapid detection and accurate alignment of up to a million or more related protein sequences, yet is equally useful and accurate for alignment of small sequence sets. With each new database release it facilitates rapid realignment and classification either of specific subgroups of sequences or of all the members of a major protein class. It allows the incorporation of detailed structural, biochemical and mechanistic information into the alignment process, and it allows alignments to be continuously refined and corrected, as new information comes to light. Here these and other advantages are illustrated by applying MAPGAPS to P-loop GTPases.

2 MATERIALS AND METHODS

A MAPGAPS search involves two steps requiring two procedures: (i) the Multiple Alignment of Profiles (MAP) procedure and (ii) the Global Alignment of Protein Sequences (GAPS) procedure (Fig. 1). The GAPS procedure requires two input files (green boxes in Fig. 1), a multiple-profile alignment, which is created by the MAP procedure, and a protein sequence database; it can also utilize an optional input file (orange box in Fig. 1) that specifies a pruning tree (see below). The GAPS procedure uses search heuristics to detect matching sequences, which it then classifies and multiply aligns.

2.1 Input files

The MAP procedure requires as input two files (blue boxes in Fig. 1): (i) a 'template' alignment and (ii) an array of multiple sequence alignments, each of which corresponds to a subgroup of (typically closely related) proteins within a larger set of related proteins. The MAP procedure creates a multiple-profile alignment by first generating profiles (i.e. position-specific scoring matrices) corresponding to the input multiple alignments and then multiply aligns the profiles based on the template alignment. Profiles of circularly permuted domains can be created by applying a MAPGAPS routine that converts a permuted alignment into an un-permuted alignment that is consistent with the template. The template consists of a set of multiply-aligned consensus sequences—one sequence for each profile with the first sequence representing the consensus sequence for the template itself.

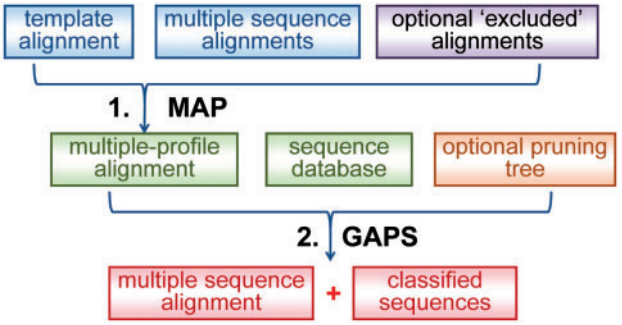

Fig. 1. Flowchart for a MAPGAPS search.

2.1.1 The template alignment The sequences that make up the template alignment typically are very difficult to align accurately and, as a result, require manual curation in the light of sequence motif analysis and of structural and biochemical information. For the analysis here, Bayesian methods (Lawrence *et al*., 1993; Liu *et al*., 1995, 1999; Neuwald and Liu, 2004; Neuwald *et al*., 1995, 1997) were used to identify very weakly conserved sequence motifs. Bayesian partitioning with pattern selection (Neuwald, 2007a; Neuwald *et al*., 2003) was used to identify divergent sequence patterns characteristic of specific protein subgroups and, at the same time, assign sequences to subgroups based on those patterns. The CHAIN program (Neuwald, 2007a) was used to verify the structural locations of conserved residues by automatically mapping them to available protein structures. Finally, application of the MAPGAPS program itself was required to iteratively expand and improve the template alignment.

2.1.2 Subgroup alignments The construction of input multiple sequence alignments is fairly straightforward because (by design) each of these consists of sequences that belong to a relatively closely related protein subgroup that thus can be accurately aligned using standard methods. For the analysis here, either MUSCLE (Edgar, 2004a and b) or PSI-BLAST (Altschul *et al*., 1997) were used. As a starting point for obtaining such alignments, a search can also be performed using only a template alignment, in which case MAPGAPS returns a set of multiple alignments, each of which consists of those sequences whose highest (statistically significant) scores are to a specific template sequence. For such a search the template consists of actual sequences rather than consensus sequences. These alignments and the template alignment can then be edited and used as input for subsequent searches.

2.1.3 Eliminating irrelevant sequences The MAPGAPS program can also trap and eliminate related but irrelevant sequences, that is proteins that do not belong to the category of interest but that share sufficient sequence similarity to obtain significant alignment scores. This is done by including, as (optional) input to the MAP procedure, an array of multiple sequence alignments (purple box in Fig. 1)—one alignment for each irrelevant subgroup. Because database matches to the corresponding profiles are simply discarded, these 'excluded profiles' are not represented by consensus sequences in the template alignment. This feature is useful for eliminating certain subgroups from the output alignment when the goal is to analyze functionally divergent residues that distinguish one specific subgroup from another specific subgroup within a larger protein class.

2.2 Search heuristics

Using a brute force approach, a MAPGAPS search would involve computing an alignment score for each database sequence against each of the profiles. As a first step in enhancing search speed, MAPGAPS incorporates the search heuristics used in PSI-BLAST (Altschul *et al*., 1997). To ensure alignment of only those regions sharing significant sequence similarity, it relies on the

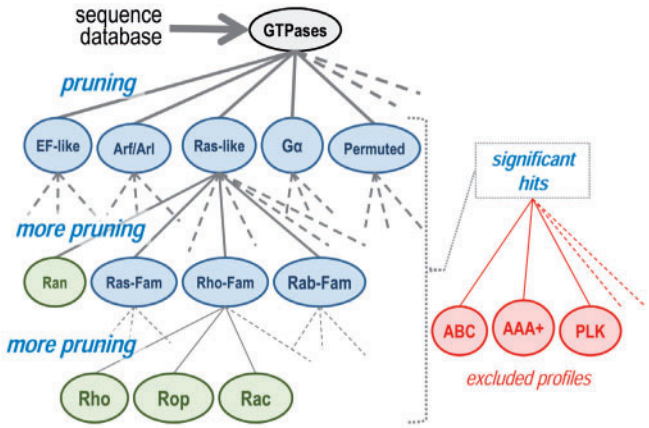

Fig. 2. A MAPGAPS search with heuristic pruning and filtering of irrelevant sequences. Each oval-shaped node represents the profile for a P-loop GTPase subgroup (root node, gray; non-terminal pruning nodes, blue; terminal nodes, green) or for an irrelevant subgroup (red nodes). The thinning of line widths as one goes down the tree denotes pruning of database sequences based on parent node threshold trigger scores. After the search, the program discards each significant sequence hit with a highest score against an excluded profile.

gapped-based statistics (Karlin and Altschul, 1990) used in PSI-BLAST. For additional speed MAPGAPS prunes the search space as follows.

The MAPGAPS program first scores each database sequence against a profile of the template alignment itself. Only those sequences that attain an ungapped PSI-BLAST high-scoring segment pair (HSP) (Altschul *et al.*, 1997) score above a specified cutoff (22 bits by default) (this is termed the 'threshold trigger score') are compared against the other profiles. Pruning of the input sequence set in this way is implemented using a bitwise data structure for efficient set operations (Neuwald and Green, 1994). As an additional, optional search heuristic, the profiles can be arranged into a tree, in which case the template profile serves as the root node and each subgroup profile serves as either a non-terminal or a terminal (leaf) node (Fig. 2). The tree structure is specified using a 'depth-first traversal' representation, as described in Appendix 1 (see Supplementary Data A). (If the pruning tree input file is omitted, then all profiles except the root node are treated as child nodes of the root.) Only those sequences that obtain threshold trigger scores against a particular non-terminal profile are searched against the associated child profiles. In addition, for parent profiles sharing greater similarity to child profiles—i.e., when searching further down the tree— more stringent PSI-BLAST word, trigger and extension threshold scores (Altschul *et al.*, 1990, 1997) are used. Finally, those sequences with significant hits against the multiple-profile alignment are scored against excluded profiles in order to filter out irrelevant sequences. Altogether, these various pruning steps typically provide an additional 10–30-fold speed up without a significant loss of sensitivity. (Mainly short sequence fragments and members of very distantly related subgroups not represented in the template alignment escape detection.)

2.3 Post-search processing

The MAPGAPS program utilizes post-search processing routines to classify and multiply align matching sequences.

2.3.1 Sequence classification and identification of unclassified sequences Each database sequence with a significant match to at least one profile is classified by assigning it to the subgroup represented by its highest scoring profile. Multiple domains within a single sequence may be classified, of course, into distinct categories based on the highest profiles scores for each domain. In some cases, a sequence or domain will be assigned, not to a specific family or subfamily, but into a superfamily or subclass or even the class as a whole. Such sequences—or, for that matter any sequences with low, but significant (highest) profile scores—correspond to protein subgroups that are not explicitly modeled by a profile. MAPGAPS can flag these sequences using an option for reporting low scoring, yet significant matches. Once new protein subgroups are identified in this way, alignments and corresponding profiles can be created and incorporated into the template alignment.

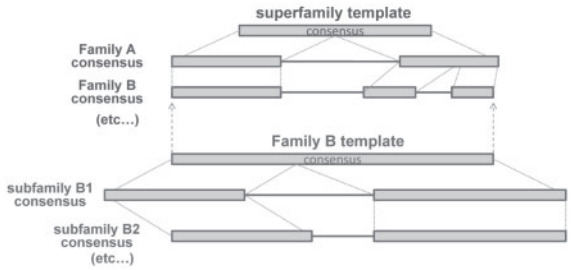

Fig. 3. Mapping a family template onto a superfamily template.

2.3.2 Linking together local alignments Certain sequences may share with their highest scoring profile, two or more locally aligned regions of significant similarity separated by unaligned regions. It is desirable to join these locally aligned regions together into an alignment that more fully spans the protein domain that is modeled by the profile. The GAPS procedure does this by applying the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970) with the constraint that only those regions of the dynamic programming matrix corresponding to the previously identified local alignments are considered and with the penalty for gaps between local alignments set to zero. This identifies the highest scoring sequence-to-profile alignment passing through the local alignments and allows overlaps between locally aligned regions to be resolved. Note that by not imposing a gap penalty between local alignments, this procedure may allow very large insertions to occur within a protein domain.

2.3.3 Multiple alignment of detected sequences The multiple-profile alignment serves as a template for aligning all of the matching database sequences to each other. This is accomplished by using both the pairwise alignment of each matching database sequence against its highest scoring profile and that profile's alignment against the template alignment: If position x in a matching sequence aligns with position y in profile Y and if position y in profile Y aligns with position z in the template alignment, then position x in the sequence is aligned with position z in the all-inclusive multiple alignment. In this way, as long as the template alignment is accurate, corresponding residues that would otherwise be impossible to align using standard methods, can be aligned correctly and automatically.

This procedure can also be used to align lower-level template alignments to each other using a higher-level template (Fig. 3). This is implemented within another MAPGAPS procedure (the 'convert' program), which thus can align various subfamily templates to a common family template, various families templates to a common superfamily template, etc. Such output files can then be merged into a single high-level template prior to a MAPGAPS search. The construction of low-level template alignments allows MAPGAPS to detect and align only those database sequences belonging to a specific subgroup within a protein class. In such a search the family B profile in Figure 3, for example, would then become the root in Figure 2 and all profiles outside of family B's subtree would then serve as excluded profiles.

3 RESULTS

The MAPGAPS program was implemented in C++ and applied to a set of 33 distantly related proteins containing P-loop GTPase

Table 1. The numbers of weakly conserved regions within 33 P-loop GTPases of known structure that are misaligned by various methods

Conserved region	Function	G-domain structure[a]	Typical pattern[b]	MAPGAPS	MUSCLE	PROMALS3D w/ sequences	PROMALS3D w/ structures[c]	hmmalign local	hmmalign global
Walker A	Phosphate binding	$\beta 1$ to $\alpha 1$	G….GK[ST]	0	13	1	0	1	0
Switch I[d]	Mg++ binding	loop2	T	0	>20	4	>20	5	4
Walker B	Coordinates Mg^{++}	$\beta 3$	hhD..G	0	>20	2	2	2	0
$\beta 4-$strand	Interacts with NK.D	$\beta 4$	hhhh[DSN]	0	>20	4	4	5	0
NK.D motif	Binds guanine & ribose	$\beta 5$, loop8	[NT]K.D	0	>20	3	3	5	0
NK.D-Arg[e]	Forms salt-bridge	loop8	R	0	7	3	3	2	2
SAK motif	Interacts with guanine & NK.D	$\beta 6$, loop10, $\alpha 5$	[SC][AG]K	0	>20	9	6	8	5
			Total	0	>120	23	>38	26	11
			Shuffled[f]	0	33	33	ND	33	33

[a] Structural features are according to (Wittinghofer, 2000).
[b] The symbol 'h' represents a hydrophobic residue.
[c] Only 30 proteins were tested in this case due to a program-imposed limitation.
[d] The switch I motif is absent from and thus was not scored for five of the input sequences.
[e] This arginine is located just beyond the NK.D motif and is conserved only within Ran, Rab, Ras and Rho GTPases, in which it forms a salt bridge with a conserved acidic residue in the $\beta 6$ strand (Neuwald *et al*., 2003); seven of these GTPases were included in the input sequence set.
[f] In a separate test, randomly shuffled versions of the input sequences were included in the input set; shown are the number of these shuffled sequences included by each program in the output alignment (along with the 33 P-loop GTPases sequences).
ND: Not done.

domains of known structure, and the resulting alignment was compared with those obtained using several other multiple alignment methods. To illustrate the ability of MAPGAPS to rapidly detect, classify and align large numbers of sequences, a search for P-loop GTPases within a set of 65-million sequences was performed.

3.1 Multiply-aligning a small set of sequences

Columns 1–4 of Table 1 describe several short regions of sequence similarity that are weakly conserved across distantly related subgroups of P-loop GTPases (Wittinghofer, 2000) and that thus are particularly difficult to multiply-align correctly. One hundred and ninty-one profiles of various GTPase subgroups (Leipe *et al*., 2002) and a corresponding template alignment were created and used as input to the MAP procedure. The resulting output files were used by the GAPS procedure to align a set of 33 distantly related GTPases of known structure; the same sequences were also aligned using MUSCLE (version 3.7) (Edgar, 2004a and b), PROMALS3D (Pei *et al*., 2008a and b), and the hmmalign program within the HMMER software package (version 2.3.2) (http://hmmer.janelia.org/). For the hmmalign program, a profile HMM of the P-loop GTPase class was constructed using the MAPGAPS template alignment as input to hmmbuild (a program within the HMMER package for creating an HMM profile from an input alignment). All programs were run using their default parameter settings and the results are summarized in Table 1. (Output alignments are available as Supplementary Data B.)

As shown in column 5 of Table 1, MAPGAPS correctly aligns all of the short, weakly conserved regions within these GTPase domains, which is of course not surprising given that it is informed by a carefully curated multiple-profile alignment. This comparison nevertheless demonstrates how MAPGAPS can work around otherwise impossible obstacles hindering accurate alignment of divergent sequences. As shown in columns 6–10 of Table 1, unsupervised and less-supervised methods misalign these regions to varying degrees. MUSCLE, which relies on sequence data only, fails to correctly align nearly all of these regions. On the other hand, the sequence-based version of PROMALS3D, which uses homology to known structures in conjunction with a progressive alignment procedure, performs significantly better, but still misaligns 23 regions. The structurally based version of PROMALS3D performs slightly better still except on the switch I region, for which it performs dramatically worse—presumably due to the inherent structural flexibility of this region.

The comparison of MAPGAPS to hmmalign is most appropriate here inasmuch as both programs rely on a curated multiple alignment as the query. Indeed, for this test, both hmmalign and MAPGAPS utilized the same curated template alignment. Nevertheless, the local alignment version of hmmalign either misaligns or fails to aligned about as many regions as does the sequence-based version of PROMALS3D. The global alignment version of hmmalign comes closest to matching the performance of MAPGAPS—though it still misaligns 11 sequences corresponding to three distinct regions within these GTPases: (i) the switch I region; (ii) a loop region directly following the NK.D motif that, within certain GTPases, conserves an arginine residue; and (iii) the SAK motif region. Thus it is worthwhile considering these regions in greater detail.

Presumably hmmalign and other programs misalign the Switch I region for certain GTPases due to its high structural and sequence variability: this region merely conserves a single threonine (or rarely a serine) residue that, in the canonical structural conformation, precedes the $\beta 2$ strand by three residues. As a result, there is often insufficient sequence and structural information to correctly align this residue except through manual curation.

The arginine following the NK.D motif tends to be misaligned for two reasons: first, both this arginine and an associated acidic residue, with which it forms a salt bridge, are conserved within Ras-, Rab-, Ran- and Rho-related sequences, but not within other P-loop GTPases. As a result, using a single profile HMM to model

all GTPases only weakly favors an arginine at this position. Second, within certain of these GTPases a deletion directly precedes this arginine and an insertion (up to 15-residue in length) directly follows it (Neuwald *et al*., 2003). Hence these sequence characteristics make it extremely difficult to align this arginine correctly without manual curation of multiply aligned profiles.

Likewise, misalignment of the SAK motif region by hmmalign or other methods appears due to sequence (and presumably corresponding functional) divergence. Consider, for example, this region within human interferon-induced guanylate-binding protein 1 (pdb_id: 2b8w) (Ghosh *et al*., 2006). In this case, the SAK motif serine residue—the sidechain OH group of which normally forms a hydrogen bond with a sidechain oxygen atom of the NK.D aspartate residue—is replaced by an arginine. This arginine's C_β hydrogen atom, which is structurally analogous to the serine sidechain OH hydrogen, contacts the sidechain of the NK.D aspartate just as does the canonical serine—thereby indicating a clear-cut structural correspondence between these divergent residues. Of course, when classifying and characterizing proteins based on functionally divergent sequence features, it is essential to correctly align dissimilar residues at structurally corresponding locations. This again requires manual curation of multiply aligned profiles.

The relatively impressive performance of the global version of hmmalign is muted somewhat by the fact that the high quality of the input HMM profile was achieved through iterative applications of the CHAIN and MAPGAPS programs. (The Pfam and SMART databases lack a curated GTPase profile of comparable quality.) In particular, during creation of the template alignment, MAPGAPS was required for gathering distantly related GTPases and for identifying new subgroups for incorporation into and refinement of the evolving template alignment, whereas the CHAIN program was required for classification of subgroups based on statistically significant divergent sequence patterns. Thus it is difficult to decouple hmmalign's performance from important contributions made by the MAPGAPS program to the construction of the HMM profile.

It is also important to note in this context that, when the input set includes randomly shuffled versions of the input sequences, MUSCLE, PROMALS3D and hmmalign incorporate these into the output alignment (last row in Table 1). Thus, although these other methods are able to correctly align certain weakly conserved regions, this ability appears to be associated with a substantial loss of specificity. Moreover, for MUSCLE and PROMALS3D inclusion of these random sequences degrades the alignment quality of the original GTPase sequences. For MAPGAPS, however, sequences lacking significant sequence similarity to the query profiles are excluded from and have no effect upon the output alignment. Also, unlike these other methods, when a permuted GTPase sequence (such as YjeQ) (Shin *et al*., 2004) is added to the input set, MAPGAPS can correctly align the sequence while again maintaining the alignment quality of the original sequences (data not shown).

Taken together, this analysis demonstrates that, given a carefully curated multiple profile alignment, the MAPGAPS program is able to correctly align subtly conserved regions and, indeed, even strikingly divergent, yet structurally related regions that would be impossible to align using more fully automated methods. It also demonstrates how MAPGAPS is guaranteed to multiply align a particular set of protein sequences in precisely the same way, whether or not additional (related or unrelated) sequences are included in the input set.

3.2 Identifying and aligning database sequences

Because the MAPGAPS program will align only those sequences with significant similarity to the 'query' multiple-profile alignment and because it does this in a manner that is independent of the makeup of the sequence set being searched, it is useful for identifying, classifying and accurately aligning large numbers of related sequences within a large database. Such database searches are also useful for identifying related protein subgroups that currently are not modeled by the multiple-profile alignment. Such applications are illustrated here for P-loop GTPases.

3.2.1 Detection, alignment and classification of putative GTPase domains MAPGAPS was used to search the December 11, 2008 releases of the NCBI nr , env_nr and translated est databases (containing 7 463 447, 6 028 191 and 51 914 606 sequences, respectively) against 181 GTPase profiles. (For the est databases only translated open reading frames $\geq$100 residues in length were searched.) The search took about an hour using a dozen grid nodes (and using a three-level heuristic pruning tree). This identified and aligned 475 038 putative GTPase domains (403 340 unique domains in 400 859 sequences). A total of 171 465 of the unique domains aligned with the entire GTPase domain or nearly so; after purging this set to remove domains with $\geq$98% identity, 53 301 nearly fully aligned GTPase domains remained.

MAPGAPS classified the 403 340 unique domains into the major GTPase categories represented by the pie chart in Figure 4a and into 181 subcategories (not shown). In addition, 4369 significant hits were discarded as irrelevant. Excluded profiles were included for the following P-loop NTPase subgroups: AAA+ ATPases, helicases, P-loop kinases, ABC transporters and six other minor categories of (typically poorly characterized) NTPases. Figure 4b shows a breakdown of the domain-containing sequences by major taxonomic groups.

3.2.2 Identifying unclassified subgroups Figure 5 shows a hypothetical protein sequence aligned against a profile of dynamin-like domains. This alignment was generated using a MAPGAPS option for identifying unclassified sequences (based on weak, yet significant scores against highest scoring profiles). This protein was detected as two local, overlapping alignments that were trimmed

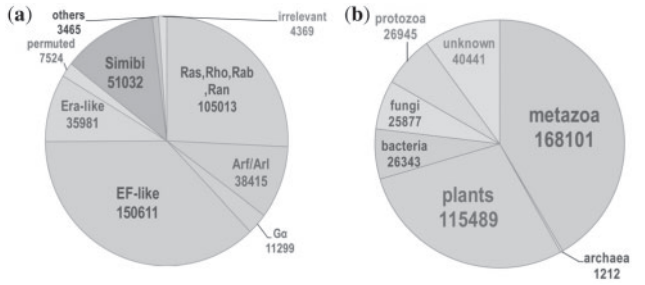

Fig. 4. Pie charts showing numbers of putative GTPase domains and sequences identified in a search of available protein sequences. (**a**) Numbers of unique domains detected within major subgroups. (**b**) Numbers of detected sequences classified by major taxonomic categories.

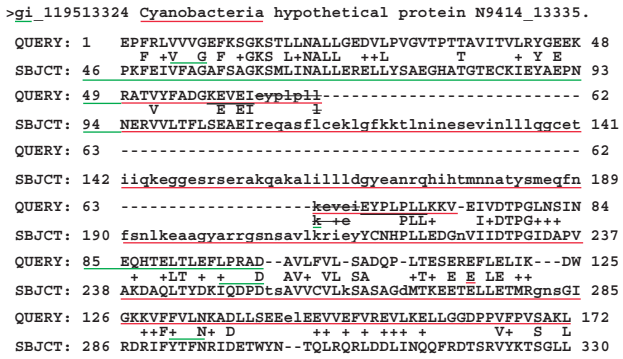

Fig. 5. MAPGAPS alignment of a hypothetical protein sequence (SBJCT) against a dynamin-like GTPase profile consensus sequence (QUERY). This putative GTPase domain was initially detected as two weakly significant overlapping local alignments (with *E*-values of 0.000016 and 0.00014) that were converted into the global alignment shown; this introduced the long insertion shown in green. Trimmed-back (previously overlapping) residues in the query are highlighted in red and crossed out; the corresponding, non-trimmed-back residues are underlined.

back and patched together into a single global alignment during post-search processing. It belongs to a putative GTPase family that currently includes 18 other cyanobacterial hypothetical proteins and that, of course, was not modeled by the multiple-profile alignment used as the query in the search. Thus this example illustrates how MAPGAPS can facilitate the detection of unclassified GTPase subgroups, for which profiles can then be created.

4 DISCUSSION

The MAPGAPS approach of using manually curated multiply aligned profiles as a query has a number of advantages over other, more fully automated alignment methods. For example, unlike these other methods, it eliminates both unrelated and related, but irrelevant sequences from the input set automatically, and the alignment it returns is not influenced by the makeup of the input set. It also allows the alignment process to be informed by the implicit biochemical, structural and sequence motif information that is incorporated into the curated template alignment. This, in turn, ensures correct alignment of regions that would otherwise be impossible to multiply-align accurately. Of course, for large, highly diverse protein classes, such as the P-loop GTPases, it is non-trivial to construct a comprehensive and perfectly correct template alignment, as well as a comprehensive set of profiles for eliminating irrelevant sequences. Nevertheless, lingering alignment errors and irrelevant sequences eventually will be eliminated through further refinement of the multiple-profile alignment as additional information becomes available. Moreover, once constructed, these input files eliminate the need to repeatedly reconstruct alignments starting from unaligned sequences, so that eventually a high-quality alignment of all currently available sequences within a protein class or of a specific subgroup within that class will be easy to obtain.

Because the amount of time that MAPGAPS requires to multiply align a set of sequences scales up linearly with the size of the set, it can detect, classify and accurately align vast numbers of sequences—up to a million or more in a few hours on a multiprocessor grid. (This is clearly implied by the analysis here, even though there were not quite enough GTPases in the current database to demonstrate this directly.) This approach also lends itself to further heuristic improvements. For example, using the template alignment, the positions at which a family profile shares similarity to a database sequence (when first detected during a search) could be mapped directly to the corresponding positions in closely related subfamily profiles. This would allow very rapid alignment of that sequence against these subfamily profiles.

The MAPGAPS approach has several other applications. It can be used to benchmark fully automated, multiple alignment methods, as was illustrated here. It also could be used to identify and annotate genomic sequences. Because curation of the template alignment is based on detailed structural analysis and because MAPGAPS can align protein sequences with high accuracy, it would be straightforward to incorporate routines to output detailed and specific structural annotations. In particular, this would allow functional annotation of specific codons within open reading frames; such an annotation might state, for example, that a specific codon corresponds to an arginine finger within a certain class of ATPases. Such assignments could be validated statistically by requiring that annotated codons have high posterior probabilities of being correctly aligned (Yu and Smith, 1999) to key positions within a profile. Likewise, it would be straightforward to allow Pfam (Finn *et al*., 2008) profiles, which are commonly used for genome annotation, to be multiply aligned and used as input to MAPGAPS in this way.

MAPGAPS also is useful for statistical analysis of protein functional divergence within the context of CHAIN analysis (Neuwald, 2006). Indeed, this was the main motivation behind development of the MAPGAPS program. CHAIN analysis, as implemented in the CHAIN program (Neuwald, 2007a), identifies functionally divergent subgroups of proteins based on sequence patterns that are most strikingly conserved within specific subgroups, but that are strikingly non-conserved outside of those subgroups. By accurately aligning vast numbers of protein sequences, the MAPGAPS program greatly enhances both the evolutionary scope and the sensitivity of CHAIN analysis. To illustrate this point, CHAIN analysis of the P-loop GTPase multiple alignment that is described here recently identified a highly distinguishing structural feature of Ras-like GTPases (Neuwald, 2009).

However, just as the MAPGAPS output is useful as input to the CHAIN program, the CHAIN output is useful as input to the MAPGAPS program. Because the CHAIN program uses rigorous statistical criteria to define functionally divergent protein subgroups, it facilitates the construction of statistically meaningful subgroup alignments, which can be used as input to the MAPGAPS program. Moreover, because the CHAIN program identifies those residues that most distinguish a specific subgroup from other, related subgroups and because it can identify the structural locations of those residues, it guides curation of the MAPGAPS template alignment. Thus the MAPGAPS and CHAIN programs can be used synergistically to construct multiple-profile alignments of various protein classes.

ACKNOWLEDGEMENTS

I thank Michelle Giglio, David Liebke and two anonymous referees for helpful comments and suggestions.

Funding: The NIH Division of General Medicine Grant GM078541.

Conflict of Interest: none declared.

REFERENCES

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.

Bork,P. and Koonin,E.V. (1993) An expanding family of helicases within the 'DEAD/H' superfamily. *Nucleic Acids Res.*, **21**, 751–752.

Christen,P. and Mehta,P.K. (2001) From cofactor to enzymes. The molecular evolution of pyridoxal-5′-phosphate-dependent enzymes. *Chem. Rec.*, **1**, 436–447.

Coutinho,P.M. *et al*. (2003) An evolving hierarchical family classification for glycosyltransferases. *J. Mol. Biol.*, **328**, 307–317.

Davidson,A.L. and Maloney,P.C. (2007) ABC transporters: how small machines do a big job. *Trends Microbiol.*, **15**, 448–455.

Edgar,R.C. (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edgar,R.C. (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Finn,R.D. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

Ghosh,A. *et al*. (2006) How guanylate-binding proteins achieve assembly-stimulated processive cleavage of GTP to GMP. *Nature*, **440**, 101–104.

Hanks,S.K. and Hunter,T. (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.*, **9**, 576–596.

Holmquist,M. (2000) Alpha/Beta-hydrolase fold enzymes: structures, functions and mechanisms. *Curr. Protein Pept. Sci.*, **1**, 209–235.

Kannan,N. and Neuwald,A.F. (2004) Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2{alpha}. *Protein Sci.*, **13**, 2059–2077.

Kannan,N. and Neuwald,A.F. (2005) Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? *J. Mol. Biol.*, **351**, 956–972.

Kannan,N. *et al*. (2007) The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proc. Natl Acad. Sci. USA*, **104**, 1272–1277.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Koonin,E.V. and Tatusov,R.L. (1994) Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J. Mol. Biol.*, **244**, 125–132.

Lawrence,C.E. *et al*. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Leipe,D.D. *et al*. (2002) Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.*, **317**, 41–72.

Leipe,D.D. *et al*. (2003) Evolution and classification of P-loop kinases and related proteins. *J. Mol. Biol.*, **333**, 781–815.

Liu,J.S. *et al*. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.

Liu,J.S. *et al*. (1999) Markovian structures in biological sequence alignments. *JASA*, **94**, 1–15.

Melby,T.E. *et al*. (1998) The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge. *J. Cell Biol.*, **142**, 1595–1604.

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.

Neuwald,A.F. (2006) Bayesian shadows of molecular mechanisms cast in the light of evolution. *Trends Biochem. Sciences*, **31**, 374–382.

Neuwald,A.F. (2007a) The CHAIN program: forging evolutionary links to underlying mechanisms. *Trends Biochem. Sciences*, **32**, 487–493.

Neuwald,A.F. (2007b) Gα–G$\beta\gamma$ dissociation may be due to retraction of a buried lysine and disruption of an aromatic cluster by a GTP-sensing Arg–Trp pair. *Protein Sci.*, **16**, 2570–2577.

Neuwald,A.F. (2009) The charge-dipole pocket: a defining feature of signaling pathway GTPase on-off switches. *J. Mol. Biol.*, **390**, 142–153.

Neuwald,A.F. and Green P. (1994) Detecting patterns in protein sequences. *J. Mol. Biol.*, **239**, 698–712.

Neuwald,A.F. and Liu,J.S. (2004) Gapped alignment of protein sequence motifs through Monte Carlo optimization of a hidden Markov model. *BMC Bioinformatics*, **5**, 157.

Neuwald,A.F. *et al*. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, **4**, 1618–1632.

Neuwald,A.F. *et al*. (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Res.*, **25**, 1665–1677.

Neuwald,A.F. *et al*. (1999) AAA+: a class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res.*, **9**, 27–43.

Neuwald,A.F. *et al*. (2003) Ran's C-terminal, basic patch and nucleotide exchange mechanisms in light of a canonical structure for Rab, Rho, Ras and Ran GTPases. *Genome Res.*, **13**, 673–692.

Pei,J. *et al*. (2008a) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.

Pei,J. *et al*. (2008b) PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, W30–W34.

Pietrokovski,S. (2001) Intein spread and extinction in evolution. *Trends Genet.*, **17**, 465–472.

Rappas,M. *et al*. (2005) Structural insights into the activity of enhancer-binding proteins. *Science*, **307**, 1972–1975.

Shin,D.H. *et al*. (2004) Crystal structure of YjeQ from *Thermotoga maritima* contains a circularly permuted GTPase domain. *Proc. Natl Acad. Sci. USA*, **101**, 13198–13203.

Wittinghofer,A. (2000) The functioning of molecular switches in three dimensions. In Hall,A. (ed), *GTPases*. Oxford University Press, Oxford, pp. 244–310.

Yu,L. and Smith,T.F. (1999) Positional statistical significance in sequence alignment. *J. Comput. Biol.*, **6**, 253–259.