



Improved database searches for orthologous sequences by conditioning on outgroup sequences*

Philip J. Cotter, Daniel R. Caffrey and Denis C. Shields[†]

Department of Clinical Pharmacology, Royal College of Surgeons in Ireland,
123 Stephen's Green, Dublin 2, Ireland

Received on December 8, 2000; revised on May 10 and July 18, 2001; accepted on August 15, 2001

ABSTRACT

Motivation: Searches of biological sequence databases are usually focussed on distinguishing significant from random matches. However, the increasing abundance of related sequences on databases present a second challenge: to distinguish the evolutionarily most closely related sequences (often orthologues) from more distantly related homologues. This is particularly important when searching a database of partial sequences, where short orthologous sequences from a non-conserved region will score much more poorly than non-orthologous (outgroup) sequences from a conserved region.

Results: Such inferences are shown to be improved by conditioning the search results on the scores of an outgroup sequence. The log-odds score for each target sequence identified on the database has the log-odds score of the outgroup sequence subtracted from it. A test group of *Caenorhabditis elegans* kinase sequences and their identified *C.elegans* outgroups were searched against a test database of human Expressed Sequence Tag (EST) sequences, where the sets of true target sequences were known in advance. The outgroup conditioned method was shown to identify 58% more true positives ahead of the first false positive, compared to the straightforward search without an outgroup. A test dataset of 151 proteins drawn from the *C.elegans* genome, where the putative 'outgroup' was assigned automatically, similarly found 50% more true positives using outgroup conditioning. Thus, outgroup conditioning provides a means to improve the results of database searching with little increase in the search computation time.

Availability: Perl scripts for the Outgroup Conditioned Score (OCS) method are available without charge for non-profit academic use from <http://www.bioinf.org/vibe/>

software/OCS/. Scripts have been optimized for Linux or OSF with a Perl v5 interpreter.

Contact: dshields@rcsi.ie.

INTRODUCTION

The searching of sequence databases is one of the principal tools that molecular biologists use to relate a sequence of interest to all other potentially related sequences that are already known. These search methods calculate the probability that the observed match with the most similar sequence in the database would occur at random, and rank the results in the order of the estimated probability, or some alternative scoring system (Altschul *et al.*, 1990; Altschul and Gish, 1996; Karlin and Altschul, 1990, 1993). These methods are extremely powerful when the sequence being searched is of unknown function, and the similarity to known sequences on the database indicate that the sequences are homologous (having arisen from a common ancestor at some point in their evolutionary history). This observation allows a prediction that the sequence and its best match are likely to have related functions. However, increasingly, sequence databases contain very many members of multi-gene families and the task is not simply to define the broad family that a sequence belongs to, but to define which particular database sequence(s) the sequence of interest resembles most closely. Assuming that evolutionary rates are constant in different lineages and that the database contains complete, and not partial, sequences, the top match in a database search is the best estimate of the most closely related sequence. However, evolutionary rates are not constant for all sequences within a family, and many databases, in particular Expressed Sequence Tag (EST) databases, include many partial sequences. A local alignment score of an EST from a paralogous sequence might be higher than the score resulting from a true orthologous EST, if this EST only contains less conserved parts of the gene. This is the problem that we seek to address.

*Non-standard abbreviations: OCS: Outgroup Conditioned Score, HOCS: Heuristic Outgroup Conditioned Score, HOCP: Heuristic Outgroup Conditioned *P*-value ratio, EST: Expressed Sequence Tag.

[†] To whom correspondence should be addressed.

Information about evolutionary relationships is incorporated into some search methods. An aligned group of homologous sequences may be used to search a database by converting to a set of 20 amino acid probabilities at each residue (a 'profile'). This allows greater weighting of matches to the more conserved regions (Gribskov *et al.*, 1987) which increases the power to detect very distantly related sequences. Position-specific iterative searches of single sequences similarly use the best hits on the database to help find more weakly related sequences (Altschul *et al.*, 1997). Other heuristic approaches permit the combination of well aligned and poorly aligned segments (Grundy and Bailey, 1999) to increase power. However, when identifying the most closely related sequences within a family, the objective is exactly opposite: to screen out as many of the members of the family as possible.

To determine the precise relationships between a group of sequences the most thorough approach is to create a model of their relationships, represented by an evolutionary tree connecting the sequences together (e.g. Figure 1). Such an approach has been taken in the characterization of clusters of orthologous sequences (Tatusov *et al.*, 1997, 2000). Retief *et al.* (1999) have extended this evolutionary approach to precisely the problem that we seek to address: how to identify more closely related orthologues. Their method is to use a stringent substitution matrix that places greater weight on more closely related sequences than on more distantly related ones. Choice of substitution matrix alone does not account for rate variation or partial sequences with low scores, but their approach of reconstructing the phylogeny of the sequences will be the best guide in identifying potential orthologues. Unfortunately, the target database may contain many partial and error-ridden sequences. Phylogenetic reconstruction is only reliable from relatively complete sequences, and even complete and error-free sequences can be laborious to align without substantial errors. Therefore, a more robust and cruder method is required in order to investigate the data, making the minimum of prior assumptions regarding accuracy and completeness of the sequences in the target database.

Here we present a method that permits the inclusion of a limited amount of evolutionary information (an 'outgroup' which is evolutionarily more distant than the orthologous sequences that we are interested in detecting) to assist in screening out more evolutionarily distant sequences. In searches of EST databases, we show that this method improves the power to detect orthologous sequences.

METHODS AND ALGORITHM

The problem

The problem is represented by the following example, where genes from a complete genome of species C are

being searched against partial sequences from genome of species H. A family of sequences is well characterized in species C, including sequences $x_C, y_C, z_C \dots$ for which the evolutionary relationships are represented by a tree (Figure 1). A database is available containing many partial (incomplete) sequences from another species H. The objective is to identify the sequence(s) from species H which are orthologous with a particular sequence x_C , and to downweight sequences from H which are orthologous to outgroup sequences, $y_C, z_C \dots$. Since partial sequences may be from conserved or rapidly evolving regions, it is expected that sometimes higher scores will be obtained for matches to the outgroup sequences which include conserved regions, than to the truly orthologous partial sequences derived from less conserved regions. While not all problems will involve identifying orthologous sequences from different species, for convenience we will here refer to the closely related sequences of interest as orthologues.

Theory

The query sequence is aligned to a database target sequence over particular regions of each sequence (usually by the database searching program). The outgroup sequence is aligned to the query sequence independently. An alignment of all three sequences is generated by merging these two alignments. This is performed automatically without user intervention by taking the two alignments and merging them on the basis of the common (query) sequence, inserting gaps as appropriate. While in general a multiple alignment approach will give a superior alignment of three sequences, we prefer to control the alignment process to avoid errors in alignment arbitrarily inflating or deflating scores. Since the same query-outgroup alignment is used in every score comparison, there is less room for such error. Residues at which there is a gap in any of the three sequences are excluded from consideration. Target sites that match neither query nor outgroup ('.' in Figure 2), or which match both equally ('*' in Figure 2) are uninformative. True orthologues should have a greater number of target sites y which match the query and do not match the outgroup ('+' in Figure 2), compared to target sites n which match the outgroup but not the query ('-' in Figure 2).

Scoring of protein similarities usually allows for amino acid similarity as well as identity, using a pre-defined log-odds substitution matrix between amino acids (Karlin and Altschul, 1990; Henikoff and Henikoff, 1992). The log-odds score S_q is calculated as the sum of the scores between query and target for the residues that are aligned by the search programme, limited to those residues at which the query and the outgroup protein are also aligned without a gap. The log-odds score S_o may be calculated between outgroup and target for all the residues (Altschul



Fig. 2. Illustration of OCS calculation with a database target sequence from search of *C.elegans SUR1* query with outgroup *F42G8.3*.

sequence is not identified in the outgroup search, the P -value is arbitrarily assigned as 1.0. When the target sequences in a database identified by query and outgroup searches are identical, then the ranking provided by HOCS and HOCP are identical, assuming that P -values are estimated from scores in a manner similar to that implemented in the BLAST program. Sequences for which there is no outgroup available will be arbitrarily demoted since the true P -value is probably less than 1; the extent of the demotion will be partly influenced by constants relating to database and query size used in estimating P . We have no theoretical justification supporting the heuristics HOCS and HOCP and only suggest them here as potential crude substitutes for the OCS statistic. In principle, the HOCP will account for the length of the target sequences; if all the outgroup scores were absent, it would provide a better measure than HOCS. For both HOCS and HOCP, the query and target scores are not derived from equivalent alignments, and therefore these statistics are less reliable than the OCS score.

Outgroup choice

There are two means by which the outgroup may be chosen. The first is based on an individual evaluation of the phylogenetic tree of the protein family in question, which will require alignment of the sequences, phylogenetic tree construction and ideally might include some statistical measures of tree reliability, such as bootstrap measures. We investigated the utility of this approach using a kinase validation dataset.

The second approach is to define the outgroup automatically without alignment, based for example on BLAST sequence search comparisons among the protein family members. For example, if the objective is to identify potential new sequences on a human EST database which are orthologous with a given *Caenorhabditis elegans* protein, the *C.elegans* outgroup could be chosen by defining an arbitrary degree of similarity between query

and outgroup sequences. The degree of similarity chosen depends on the anticipated rate of evolution. If there are known orthologues within the protein family, the query and outgroup should be typically more dissimilar than those known human–*C.elegans* orthologues. We investigated the utility of this approach in a second multiprotein validation dataset of *C.elegans* proteins by arbitrarily setting a query–outgroup similarity cut-off as 90% of the BLAST query–query score to define outgroups.

Application

Since rapid search results are generally desirable, the methods were applied to search outputs from the efficient TBLASTN program (<ftp://ncbi.nlm.nih.gov/blast/executables/>). The program performs pair-wise alignment of the protein sequences against each sequence within a nucleotide database dynamically translated into all six reading frames (Altschul *et al.*, 1990; Karlin and Altschul, 1990). The program returns the top hits ranked according to the P -value, and with alignments of the regions of similarity contributing to the score. Performance of the exact (OCS) and heuristic (HOCS, HOCP) scoring methods were compared to an exhaustive search of each partial sequence against every target, as well as to a straightforward search of the query without considering the outgroup. Results were evaluated by summarizing statistics relating to the distribution of true positives and false positives in search outputs. These statistics are not ideal, since they are quite strongly dependent on the particular score assigned to the first false positive; however, they represent the best means whereby we could assess the relative rankings of the scores at the most important upper region of the score ranks.

Kinase validation dataset. The utility of any alternative algorithm requires assessment by some objective means. We chose a group of proteins that illustrate this problem well, for which the evolution of the multigene family is sufficiently well characterized to permit a comparison of the observed performance of the methods with expectations. The test database to be searched was the human EST database. The test queries were taken as a number of different kinases from the nematode worm *C.elegans* from the MAP and Tyrosine kinase families, based on previous evolutionary analyses (Caffrey *et al.*, 1999; Rikke *et al.*, 2000). Only the kinase domain of the proteins was used in the searches. The test target database was limited to a subset of the human Unigene database (Benson and Rapp, August 1997 Research: the UniGene collection. NCBI News, 96-3272) taken from <ftp://ncbi.nlm.nih.gov/repository/UniGene/Hs.seq.uniq.Z>. Extracting all clusters whose annotation included the word 'kinase' substantially reduced the size of the test target database. These were then concatenated into a

single file of 349 kinase clusters, containing 21 979 sequences in total. Each EST was modified during the build procedure so that its annotation line included reference to the accession number of the cluster of which it was a member. This test database is available at www.bioinf.org/vibe/datasets/est_kinase_set. The search of all EST sequences was carried out against the non-redundant *C.elegans* protein database, WORMPEP release 17 (obtained from ftp://ftp.sanger.ac.uk/pub/databases/wormpep/old_wormpep17/).

Automatic-outgroup protein validation set. A second test data set was defined as follows: in order to evaluate how useful OCS might be in practise, we imitated the choice of outgroup that might be used in practise based on limited knowledge from BLAST search results. Groups of paralogues within *C.elegans* were defined as all proteins of a set where the BLAST scores between them was at least 90% of the BLAST score of the sequences to themselves. Parologue sets were similarly defined for the human Unigene 'unique' cluster representative sequences. Many of these sets contained a single protein. Then, for a 'query' representative from each *C.elegans* set we defined a human 'target' sequence which was the top hitting sequence. If a search of the human target against WORMPEP identified the query protein set, they were defined as true positive matches. An 'outgroup' sequence was then defined as the WORMPEP protein set most closely related to the query. In order to eliminate noise caused by very low matching sequences, we excluded all proteins where the query–target, query–outgroup BLAST scores were less than 100. In order to include only the more likely outgroups, we excluded all those where the query–target BLAST score minus the outgroup–target BLAST score was less than 100, and the query–target score minus the query–outgroup score was less than 100. This process defined 151 query proteins with outgroups for validation.

IMPLEMENTATION

Implementation of OCS searching method

Twelve *C.elegans* kinases were selected for which an outgroup sequence was known and for which the human orthologue was identified (Caffrey *et al.*, 1999; Rikke *et al.*, 2000). Choice of the outgroup is the only step in this search method that requires some care, since interpretation of large family trees of sequences containing many sequences from a number of organisms is not always straightforward. We limited the choices of query, outgroup and human orthologue to those sequences where previous analyses (Caffrey *et al.*, 1999; Rikke *et al.*, 2000) provided reasonable evidence supporting those choices.

Each query protein sequence was searched against the EST_kinase database, using the TBLASTN program

version 2.0.10. Alignments were returned for every sequence scoring above the default BLAST (Altschul *et al.*, 1990) threshold (expected number of hits by chance, $E = 10$). CLUSTALW (Higgins *et al.*, 1996) was used to create an alignment of the query and outgroup sequences. The Perl script then generated a 3-sequence alignment of query, outgroup and database target sequence. This alignment simply merged the two pre-existing alignments (Query–target and query–outgroup) without changing their alignments in any way. The OCS statistic was then calculated using the aligned residues for which there were no gaps, using the Blosum 62 substitution matrix (Henikoff and Henikoff, 1992) to calculate scores. The search output was then ranked according to OCS.

Calculation of the heuristic statistics

HOCS was calculated by merging the results of the two independent BLAST searches of the query and outgroup sequences against the target database. This then permitted ranking of the results according to HOCS. The HOCP ranking was obtained similarly.

Exhaustive searching of every EST versus protein database

An exhaustive search of each EST sequence from the EST_kinase database against the *C.elegans* protein database was performed. In principle, this method should avoid the problem of a partial orthologue missed in a standard search (since it scores much lower than a complete non-orthologue); such a sequence should rank its orthologue at the top of its search output. Each search was ranked on P -value.

Statistics

The Wilcoxon matched pairs signed-rank test was performed using STATA version 6 software (STATA statistical corporation, Texas).

RESULTS

Kinase validation dataset

The 12 sequences used to test this method are shown in Table 1, along with their outgroups and predicted known human orthologues. The primary purpose of the method is to help distinguish among a large list of sequences, all of which have a score above a given threshold. Therefore, we defined a true positive as an EST from an orthologous UniGene cluster that had a score above the BLAST threshold when searched with the *C.elegans* query. Thus, the true positives are defined *a priori* on the basis of the UniGene cluster membership and their location within phylogenetic trees which were previously constructed using complete protein sequences without reference to this method.

Table 1. Protein kinase query sequences used in comparison of search methods

Query group	<i>C.elegans</i> querysequence ^a	<i>C.elegans</i> outgroup sequence ^a	Orthologous (<i>Homo sapiens</i>) UniGene clusters ^b
ERK 1/2	SUR1 P39745 (SP)	F42G8.3 AF038618 (GB)	ERK1: Hs#S4664 ERK2: Hs#S269478
JNK 1/2/3	B0478.1 (wp17)	F42G8.3 AF038618 (GB)	JNK1: Hs#S1197 JNK2: Hs#S2612 JNK3: Hs#S3573
MEK 1/2	MEK2 A56466 (GB)	R03G5.3 (wp17)	MEK1: Hs#S551621 MEK2: Hs#S551622
MKK 3/6	R03G5.3 (wp17)	K08A8 (wp17)	MKK3: Hs#S554522 MKK6: Hs#S342588
MKK 4	F42G10.2 (wp17)	K08A8 (wp17)	MKK4: Hs#S3608
MKK 7	K08A8 (wp17)	F42G10.2 (wp17)	MKK7: Hs#S1055089
PAK 1/2/3	CEPAK AAC47308.1 (GB)	T19A5 U53153 (GB)	PAK3: Hs#S1263109 PAK1: Hs#S4134 PAK2: Hs#S4135
RYK	C16B8.1 (GB)	DTK F11E6.8 (GB)	RYK: Hs#S554009
DTK/RON	DTK F11E6.8 (GB)	RYK C16B8.1 (GB)	DTK: Hs#S3274 RON ~ MST1R: Hs#S5366
EPH group	EPHB2 ^c M03A1.1 (GB)	HER2 ^c ZK1067.1 (GB)	HEK: Hs#S2638 HEK11: Hs#S3240 EHK-1 receptor: Hs#S305539 EPHB2: Hs#S952865 Eph-like receptor: Hs#S1368505 Hs#S2367 HTK: Hs#S630 HEK2: Hs#S5279 Hs#S1629
ALK	ALK T10H9.2 (GB)	DDR ^c DAF-2 (GB)	ALK: Hs#S876248 ALK-like leukocyte TK: Hs#S277
ERB/HER	HER2 ^c ZK1067.1 ^c (GB)	EPHB2 ^c M03A1.1 ^c (GB)	ERBB4: Hs#S3218 HER2: Hs#S1122

^aGB: Genbank; SP: SwissProt; wp17: Wormpep version 17, Sequence names, i.e. 'SUR1', as presented in Caffrey *et al.* (1999).

^bSequence identifiers are the UniGene accession number of the representative sequence.

^cSequence has been edited (limited to kinase domain).

Table 2 indicates the number of true positives that ranged between 4 and 105 sequences for various queries. It also indicates the number of other (non-orthologous) sequences appearing in the search output above the threshold. Any improvement in interpretation should assist in distinguishing the orthologous from the non-orthologous sequences within the output.

The outgroup conditioned search method essentially reorders the ranks of the TBLASTN query search output, relegating those with relatively strong similarity to the outgroup sequence and in consequence promoting those with strong similarity to the query. In order to quantify this, we evaluated the number of contiguous top-ranking true positives whose rank was greater than the rank of the highest-

ranking false positive. This focuses on the performance most relevant to researchers, who will inspect a search output from the first sequence and work their way down the list. Over all 12 kinase queries, the number of contiguous top-ranking true positives was 89 for the straightforward BLAST search, but rose to 141 for the outgroup conditioned method (OCS). This 58% improvement represents a substantial improvement in sensitivity with only an effective doubling in computing time. There is considerable variability between the proteins in the relative efficiency of the search methods. This increase in contiguous top-ranking true positives seen for different proteins ranges between 1 and over 3-fold. This is not surprising, given the dependence on the distribution of ESTs representing each

Table 2. Comparison of accuracy between search methods: application to MAP kinase and tyrosine kinase sequences

Sequence group	BLAST regular search query versus EST database			OCS method True positives ^d ESTs ranked higher than first false positive	BLAST exhaustive: each ESTs versus <i>C.elegans</i> protein database		
	Total no. of true positives ^a ESTs in full results	Total no. of false positives ^a ESTs in full results	True positives ^d ESTs ranked higher than first false positive		True positives ESTs with <i>C.elegans</i> query ranked #1	True positives ESTs with <i>C.elegans</i> query ranked #2-5	False positive ESTs with <i>C.elegans</i> query ranked #1
ERK 1/2 ^b	64	237	6	20	35	14	24
JNK 1/2/3 ^b	30	231	12	20	39	1	3
MEK 1/2 ^b	105	289	10	35	94	23	2
MKK 3/6 ^b	20	329	11	11	18	7	10
MKK 4 ^b	5	363	3	3	4	1	0
MKK 7 ^b	39	330	5	10	32	8	13
PAK 1/2/3 ^b	31	301	6	6	18	13	7
RYK ^c	30	196	4	8	17	3	2
DTK/RON ^c	4	243	1	1	5	0	10
EPH group ^c	66	251	18	18	43	3	1
ALK ^c	12	305	6	6	6	0	10
ERB/HER ^c	8	275	3	3	8	0	30
Total counts	424	3340	89	141	325	73	97

^aTrue positives defined as orthologues (and false positives as non-orthologues) from inspection of phylogenetic trees produced by Caffrey *et al.* (1999) and Rikke *et al.* (2000).

^bSequences taken from Caffrey *et al.* (1999).

^cSequences taken from Rikke *et al.* (2000).

sequence on the database, and given the different particular evolutionary histories of each kinase group. All of the proteins here showed either an improvement or no change, with no instance of the OCS performing more poorly than the straightforward BLAST search; however, it is likely that in other cases that this might occasionally happen by chance.

The two heuristic algorithms (HOCS and HOCP, data not shown) have the advantage that an alignment of query and outgroup is not necessary. The HOCS heuristic algorithm did not perform much better than a straightforward BLAST search, with a total of 97 true positives (over all 12 query sequences) at the top of the output compared to 89 for a straightforward BLAST. The HOCP performed with equivalent results to the more exact OCS method, with 146 true positives at the top of the output compared to 141 for the OCS (Table 2). Thus, the exact OCS statistic appears to have equivalent performance with this dataset, and since OCS scores are derived from a common alignment, this method is likely to be preferable for general use. However, for sequences with long insertions and deletions that may be critical in defining orthology, the heuristic approach (HOCP) may be superior since it will use this extra information. This information lies in the regions which are shared between the query and its orthologue, but are not shared by the outgroup. Since the query and the orthologue are more closely related, typically the inclusion

of these regions will raise the query score relative to the outgroup score.

A third exhaustive approach to identify orthology is more accurate. From these exhaustive searches, for each kinase the number of ESTs that correctly ranked the *C.elegans* orthologue #1 was calculated, giving a total of 325 across all proteins (Table 2). This was approximately equivalent to the number of orthologous ESTs identified within the entire BLAST output (a total of 424, Table 2). This method appears therefore to identify about three-quarters of the ESTs of interest. However, the exhaustive method is not completely perfect. A number of ESTs that were not orthologous identified the *C.elegans* sequence of interest as their #1 match (a total of 97), which therefore represent false positives. A number of orthologous ESTs identified a non-orthologous *C.elegans* sequence as their #1 match (a total of 73), which represent identified false negatives. While some of these non-orthologous #1 matches have low scores, this was not always the case. The exhaustive approach would be likely to show some improvement by considering the scores of an outgroup *C.elegans* target to assist in weeding out these incorrect matches.

Automatic outgroup validation dataset

The above examples of kinase searches illustrate how the method performs when good phylogenetic trees are avail-

able in order to define outgroups. However, frequently a researcher may make rough inferences about outgroups from inspection of BLAST results. We automatically generated a dataset of *C.elegans* 151 query, outgroup and target sequences drawn from the *C.elegans* database and defined according to relationships inferred from BLAST scores (see Section **Methods**). There is likely to be some falsely defined query/target/outgroup sets included in this definition, perhaps representing the level of error that users interpreting such relationships might make themselves. The average number of true positives until the first false positive using BLAST was 18.1, and the average for OCS was 29.9. This difference is statistically significant ($p = 0.003$, non-parametric signed-rank test using the ranks of the paired statistics on the 151 queries). Thus, 50% more true positives are identified using OCS. An important feature of this test data set was that to be included as a test case, outgroup–target BLAST scores are at least 100 less than the query–target scores. An additional set of 28 queries had outgroup–target and outgroup–query scores which were between 10 and 100 less than the query–target scores. For this subset, the OCS method did not provide any advantage (BLAST average number of true positives: 31.1, OCS: 27.3, $p = 0.14$). It is not surprising that the OCS method will not perform well if the query and outgroup are closely related.

The outgroup conditioned searching method was used within our laboratory to search for unknown kinase sequences. The utility of the method is illustrated by the following example. The *C.elegans* query sequence (accession number Z74029, gi: 3874986) is related but not orthologous to the PAK kinases, so a *C.elegans* PAK-orthologue CE-PAK (Caffrey *et al.*, 1999) was used as the outgroup. A candidate homologous human EST (accession number R18825) was identified at the top of the outgroup conditioned search output. This sequence was located much further down the ordinary BLAST query search output, with a rank of 40. Thus, the method identified a potential homologue that would probably have been missed by all but a very careful inspection of the straightforward BLAST output.

DISCUSSION AND CONCLUSION

The outgroup-conditioned search can greatly improve the sensitivity of a search of an incomplete database, such as an EST database. Two methods are proposed, one recommended exact method (OCS) that combines a search output with an alignment of query sequence and outgroup sequence, the other heuristic approach (HOCS or HOCP) that takes the search outputs of query and outgroup sequences as input. Thus the implementation is straightforward, and the computation time is only marginally increased or doubled. Evaluation of this method with the BLAST rapid search algorithm demon-

strates its utility in searching human EST databases with homologues from distantly related species, with a greater than 50% increase in the number of true positives ranked at the top of the search results output. The identification of most full cDNA sequences from all genes of the human genome will make this application less critical in the future. However, the need to distinguish orthologues from other homologues in the EST libraries of organisms for which there is no complete genome sequence is likely to remain an important application. For mammalian genomic genes for which no cDNA is known, if the exons are sufficiently scattered that the search method does not identify and assemble all the fragments, the genomic sequences effectively form a partial sequence database for which this method may prove useful. Domain rearrangements over evolutionary time will have a similar effect. Thus, a search of a *C.elegans* query against the human genome database might fail to correctly identify the orthologous human sequence if that human sequence has lost a conserved domain that is still shared by the *C.elegans* query, the *C.elegans* outgroup and the human orthologue of that outgroup. While such confusion may be best resolved by careful searching on individual domains, these are not always definable in advance. Finally, in the presence of very extreme variations in evolutionary rate the OCS method has the potential to identify orthologues more efficiently. This will be most beneficial when the sequence being searched for has a rapid rate of evolution, where the query sequence would otherwise assign a much higher rank to the more conserved non-orthologous sequences. Thus, in a variety of contexts, outgroup conditioned searching can provide a useful means to improve search results by adding a single piece of evolutionary information. Regularly updated exhaustive all-by-all comparisons among databases provide a sensible means to systematically address some of the problems, but the computational simplicity of the outgroup conditioned method provides a rapid means by which a sequence and its outgroup can be compared to a database.

The most difficult part of the OCS method is the choice of outgroup. Clearly, it would be ideal if the outgroup could be automatically chosen. However, the particular outgroup for a particular search will depend on exactly which evolutionary level is of interest, e.g. do I wish to identify the fungal orthologue of vertebrate cyclo-oxygenases, or alternatively the fish homologue of mammalian cyclo-oxygenase 2? It is possible that automated use of a number of alternative outgroups could assist in this, although there is no obvious sensible approach to wading through the many results files that would arise.

Incorporation of evolutionary information into search processes is not a new idea. Profile and position-specific iterative searches (Altschul *et al.*, 1997; Gribskov *et al.*,

1987) emphasize search similarities where the target sequence is similar at positions that are conserved in the sequences used for searching, in contrast to outgroup conditioning, which emphasizes similarities at residues that differ. The outgroup conditioned method could easily be generalized to a 'profile' or Hidden Markov Model subtracting the substitution probabilities of query and outgroup, but this would be computationally less efficient than the OCS algorithm presented here. A possible ultimate objective could be to create a dynamic database considering the entire set of evolutionary relationships between all sequences under study, including entire genome databases. A database search would essentially consist of adding sequences to the existing database. However, the establishment and maintenance of any such database is likely to be a considerable task. Therefore, the use of search tools such as profile searching and outgroup conditioned searching that incorporate certain facets of the overall evolutionary model to identify particular relationships are likely to be of continuing benefit.

ACKNOWLEDGEMENTS

This is a publication from the Biopharmaceutical Sciences Network, and is supported by grants from the Higher Education Authority (Ireland) and Enterprise Ireland. We thank James McInerney, Ken Wolfe and Cathal Seoighe for useful discussion.

REFERENCES

- Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Meth. Enzymol.*, **266**, 460–480.
- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Caffrey,D.R., O'Neill,L.A. and Shields,D.C. (1999) The evolution of the MAP kinase pathways: coduplication of interacting proteins leads to new signaling cascades. *J. Mol. Evol.*, **49**, 567–582.
- Gribkov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
- Grundy,W.N. and Bailey,T.L. (1999) Family pairwise search with embedded motif models. *Bioinformatics*, **15**, 463–470.
- Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
- Higgins,D.G., Thompson,J.D. and Gibson,T.J. (1996) Using CLUSTAL for multiple sequence alignments. *Meth. Enzymol.*, **266**, 383–402.
- 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*, **8**, 2264–2268.
- 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.
- Reich,J.G., Drabsch,H. and Daumler,A. (1984) On the statistical assessment of similarities in DNA sequences. *Nucleic Acids Res.*, **12**, 5529–5543.
- Retief,J.D., Lynch,K.R. and Pearson,W.R. (1999) Panning for genes: a visual strategy for identifying novel gene orthologs and paralogs. *Genome Res.*, **9**, 373–382.
- Rikke,B.A., Murakami,S. and Johnson,T.E. (2000) Paralogy and orthology of tyrosine kinases that can extend the life span of *Caenorhabditis elegans*. *Mol. Biol. Evol.*, **17**, 671–683.
- Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
- Tatusov,R.L., Galperin,M.T., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 671–683.