critical review of the manuscript and many helpful suggestions we thank Steve Bryant, John Wilbur, Dennis Benson, Mark Boguski, and Jim Ostell.

The production of *Entrez* requires the efforts of many talented individuals. There is insufficient space to list them all here, but virtually the entire staff of the NCBI is involved in one way or another. We thank them all for a job well done.

[11] Applying Motif and Profile Searches

By PEER BORK and TOBY J. GIBSON

Introduction

The demonstration of homology, meaning descent from a common ancestor, is an essential tool in gaining understanding of gene function, whether one wants to obtain an overview of the functions for all the genes described during a genomic sequencing project or to focus on a particular protein. With the expansion of sequence databases, similarity searches have a steadily increasing chance of providing a clue toward functional characterization. The likelihood of identifying homologs is currently higher than 80% for bacteria, 70% for yeast, and about 60% for animal sequence queries.[1,2]

On the basis of experience with large-scale sequence analysis,[3-5] we estimate that at present about 10–20% of identifiable similarities cannot be retrieved automatically by standard database search programs such as BLASTP[6,7] and FASTA[8] alone. The proportion of missed similarities is even higher when considering modular proteins that are composed of several (often small) functionally and structurally independent domains. The significance of "twilight zone" matches (i.e., tempting pairwise similarities below widely used thresholds of approximately 25% identity, depending

[1] P. Bork, C. Ouzounis, and C. Sander, *Curr. Opin. Struct. Biol.* **4,** 393 (1994).

[2] E. V. Koonin, R. L. Tatusov, and K. E. Rudd, this volume [18].

[3] P. Bork, C. Ouzounis, C. Sander, M. Scharf, R. Schneider, and E. Sonnhammer, *Protein Sci.* **1,** 1677 (1992).

[4] E. V. Koonin, P. Bork, and C. Sander, *EMBO J.* **13,** 493 (1994).

[5] P. Bork, C. Ouzounis, G. Casari, R. Schneider, C. Sander, M. Dolan, W. Gilbert, and P. M. Gillevet, *Mol. Microbiol.* **16,** 955 (1995).

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *J. Mol. Biol.* **215,** 403 (1990).

[7] S. F. Altschul, M. S. Boguski, W. Gish, and J. C. Wootton, *Nat. Genet.* **6,** 119 (1994).

[8] W. R. Pearson and D. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85,** 2444 (1988).

on length) has to be assessed using information provided by inspection of multiple alignments of protein families, as well as by deploying motif and profile alignment strategies based on these alignments. Searches with these tools, in turn, often lead to the identification of even more divergent homologs. The purpose of this chapter is to outline the main strategies currently in use, making clear both their powers and pitfalls, and to demonstrate their usage with two well-known domains as examples.

Terminology: Motifs, Patterns, and Profiles

The meaning of the terms motif, block, pattern, and profile need to be clarified. The search for motifs, namely, small conserved regions within larger entities, implies that only some of the information contained in a protein or domain is used. Sometimes, motifs are applied to certain functional features (e.g., glycosylation sites, SH3-binding sites) that develop independently from the surrounding context; for this (minority of) motifs the concept of homology may be irrelevant. As insertions and deletions (gaps) within a motif are not easy to handle from the mathematical point of view, a more technical term, alignment block, has been introduced that refers to conserved parts of multiple alignments containing no insertions or deletions. Patterns can be used in a more broader sense, as they can describe small motifs or larger regions containing several motifs and can also contain gaps. (Some authors use the term pattern to mean compositionally biased segments or low complexity regions, e.g., runs of aspartate residues; this usage is not meant here.) Profile is usually used to mean a full representation of features in the aligned sequences and normally implies position-dependent weights/penalties for all 20 amino acids (as well as for insertion and deletion).

Thus, there need be no contradiction of the terms motif and profile, as profiles can also be restricted to smaller regions. Nevertheless, the terms motif and profile mirror two different ideologies in the field of using family information for improving the sensitivity of homology searches: (1) restriction to key conserved features to reduce the higher "noise" level of the more variable regions, in contrast to (2) inclusion of all possible information to maximize the overall signal of the entity (protein/domain). Both approaches are valid, as documented by many successful applications of each. It is worth noting that motifs are usually harnessed to fast word search algorithms, which can be used despite limited resources, whereas profiles often use exhaustive but slow dynamic programming algorithms. Therefore it is best to use the method most appropriate given the resources and the nature of the protein family under study. It can also be advisable to do the earlier searches using motifs first, and to follow up with the slower profiles at the end.[9]

Numerous methods exist for both motif and profile searches, and the different methods grade into each other. No exact formula can be given for the choice of the method, as each protein family has a different conservation pattern. Different functional and structural constraints lead to a certain distribution of conserved and variable regions within a multiple alignment that may better suit one of the approaches.

Note that we have specifically excluded here methods that can be combined under the term "threading," that is, that try to derive potentials (which can be translated into profiles) from known three-dimensional structures, in order to recognize the compatibility of a given sequence to one of the known three-dimensional folds. Such methods have been reviewed extensively (e.g., Ref. 10 and references therein). Although they should have great potential in finding distant homologies, as yet only a few predictions based on such approaches have been published. Such successes as there have been could, in our view, have been achieved with conventional motif and profile searches using solely sequence information.

Procedures

In the hope of predicting a function for a protein under study, fast homology search programs are almost universally used. The current standard seems to be the BLAST series of programs, accessible via several World Wide Web (WWW) servers, although both FASTA and BLITZ are also frequently used. These programs undertake a database search for a query sequence and are usually the sole search undertaken. However, the results of such searches make a logical starting point for motif and profile searches (Fig. 1). These can be divided into two steps: (1) derivation of a pattern and (2) searching for the pattern (motif/profile). For the first step, programs such as CAP (consistent alignment parser; see Ref. 11 and references therein) have been developed that are able to parse outputs of "one against all" initial database search programs and that automatically create multiple alignments of conserved regions shared by the query sequence and some of the database proteins. Other methods have been developed that can be used to find conserved regions in a set of unaligned sequences (e.g., Gibbs sampler[12]; blockfinder: S. Henikoff, personal communication,

[9] P. Bork, J. Gellerich, H. Groth, R. Hooft, and F. Martin, *Protein Sci.* **4,** 268 (1995).

[10] F. Eisenhaber, B. Persson, and P. Argos, *Crit. Rev. Biochem. Mol. Biol.* **30,** 1 (1995).

[11] R. L. Tatusov, S. F. Altschul, and E. V. Koonin, *Proc. Natl. Acad. Sci. U.S.A.* **91,** 12091 (1994).

[12] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, *Science* **262,** 208 (1993).

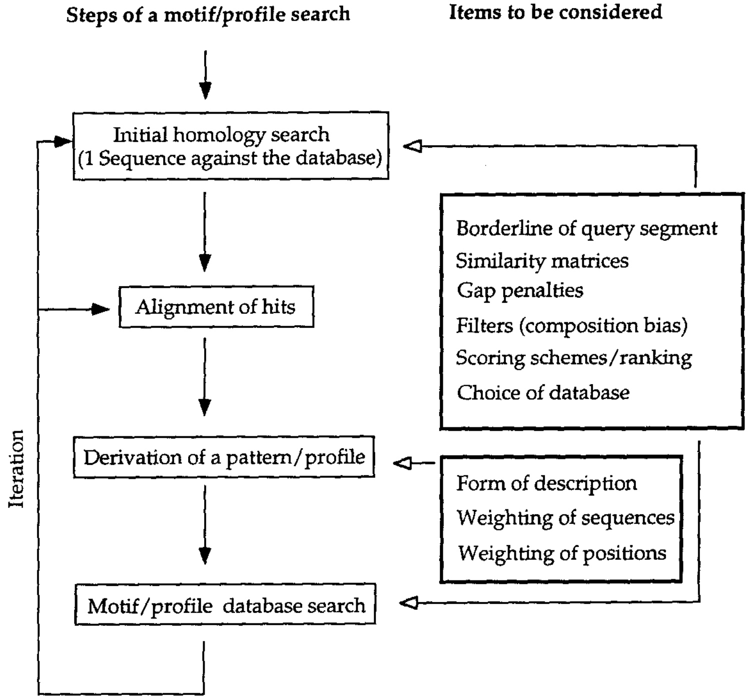

FIG. 1. Flowchart of steps in motif/profile searches and items to be considered. Note that some items may not apply to all the methods (e.g., gap parameters are not needed in MoST). Also, some of the programs combine several steps (e.g., in SOM, a starting alignment is not necessary because the detection of the conserved features in a set of sequences is part of the pattern recognition algorithm itself, Ref. 33a).

see Fig. 2; pattern extraction[13]; SOM: Ref. 33a, see Fig. 2). Even if such motifs are defined, it remains non-trivial to develop a proper description of these regions in order to be as sensitive as possible in the second step, the pattern search itself.

Available Programs

It is impossible to give a comprehensive overview of the numerous methods that exist for pattern (motif/profile) searches, especially given the explosion of WWW activities. The Internet provides access to a broad variety of programs/servers from simple string searches to sophisticated profile descriptions (for a small collection of recommended servers, see

[13] R. F. Smith and T. F. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **87,** 118 (1990).

WWW: searchable motif and pattern databases

PROSITE	[Geneva's Expasy]	http://expasy.hcuge.ch/sprot/prosite.html
Motif search, ICR	[Kyoto]	http://www.genome.ad.jp/SIT/MOTIF.html
Scan of profiles, ISREC	[Lausanne]	http://ulrec3.unil.ch/software/PFSCAN_form.html
BLOCKS, Hutchinson	[Seattle]	http://www.blocks.fhcrc.org/
PRINTS , UC	[London]	http://www.biochem.ucl.ac.uk/~attwood/PRINTS/PRINTS.html
PIMA, BCM	[Houston]	http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html
PRODOM	[Toulouse]	http://protein.toulouse.inra.fr/prodom.html

WWW: Motif and profile searches

Regular expressions,Univ	[Washington]	http://ibc.wustl.edu/fpat/
PROFILE, Weizmann	[Tel Aviv]	http://sgbcd.weizmann.ac.il/Bic/ExecAppl.html
PATSCAN motif search	[Argonne]	http://www.mcs.anl.gov/home/papka/ROSS/patscan.html
PatternFind, ISREC	[Lausanne]	http://ulrec3.unil.ch/software/PATFND_mailform.html
Pmotif (protein >DNA)	[Minneapolis]	http://alces.med.umn.edu/pmotif.html
HMM descriptions	[St. Louis]	http://genome.wustl.edu/eddy/hmm.html
Discover (email server!)	[New York]	http://hertz.njit.edu/~jason/help.html

FTP: addresses for some motif and profile search programs

Barton's flexible patterns	[Oxford]	ftp://geoff.biop.ox.ac.uk
Propat (property pattern)	[Berlin]	ftp://ftp.mdc-berlin.de/pub/makpat
SOM (neural networks)	[Berlin]	ftp://ftp.mdc-berlin.de/pub/neural
SearchWise	[Oxford]	http://www.ocms.ox.ac.uk/~birney/wise/topwise.html
PROFILE, EBI	[Cambridge]	ftp://ftp.ebi.ac.uk/pub/software/unix
TPROFILESEARCH, EBI	[Cambridge]	ftp://ftp.ebi.ac.uk/pub/vax/egcg
MoST (motif search tool)	[Bethesda]	ftp://ncbi.nlm.nih.gov/pub/koonin/most
CAP (blast output parser)	[Bethesda]	ftp://ncbi.nlm.nih.gov/pub/koonin/cap

FIG. 2. Some programs that are accessible on the Internet via WWW pages or ftp addresses. Note that this list is not comprehensive but that the authors have tested the programs referred to in this table. A more detailed list and respective pointers can be obtained via the WWW from http://www.embl-heidelberg.de/~bork/pattern.html.

Fig. 2, or go directly to the WWW at http://www.embl-heidelberg.de/~bork/pattern.html). In this section we will briefly summarize proven methods that have been extensively and successfully used in the identification of distant homology.

Most of the methods currently used have their origins in approaches and ideas developed in the 1970s and 1980s. The thorough review by Taylor[14] presents the earlier history of motif and profile searches. The template pattern matching method of Taylor and colleagues is based on combinations of profiles of amino acid properties and secondary structure propensities implemented in a flexible controlled environment.[15] Although the template methods require a fairly high level of user understanding and are not being widely used, they did achieve a notable success. On the basis of very weak similarities to the subdomain fold of aspartic proteases that were picked up by the template method, Pearl and Taylor[16] built a model structure for the human immunodeficiency virus (HIV) protease. The model was refuted

[14] W. R. Taylor, *Protein Eng.* **2,** 77 (1988).

[15] W. R. Taylor, *Prog. Biophys. Mol. Biol.* **54,** 159 (1989).

[16] L. H. Pearl and W. R. Taylor, *Nature* (*London*) **328,** 351 (1987).

by the first solved HIV protease structure, but the second solved structure found the first one to have been incorrectly built, demonstrating that the prediction was essentially correct.[17]

Searches with Regular Expressions

The simplest method to search for a motif makes use of regular expressions that can be combined with logical operators to identify a repetitive or especially conserved motif in other proteins. This method is usually very fast, is implemented in various software packages, and can be recommended as a first scan in order to estimate the number of occurrences, as well as the background noise, for a motif of interest. As the search does not provide any significance estimates, conclusions drawn from matches have to be evaluated very carefully.

Motif Databases

Signatures, based on regular expressions, for numerous protein families and functional sites are stored in the PROSITE database[18]; the database annotation provides an excellent description of the respective families/motifs. As a simple description by regular expressions has obvious limitations, more flexible descriptors have been developed,[19] and thus the PROSITE database will become even mor useful in the near future. Another approach to increase the utility of the collected motifs was taken by Henikoff and Henikoff,[20] who reconstructed alignments from the PROSITE entries, providing a database of core alignment "blocks" which is searched with tools that are more sophisticated than regular expressions. The Blocks server can be accessed by E-mail or WWW (Fig. 2).

Consensus Patterns

Patthy has developed and applied a method[21,22] that assigns a pattern to an alignment using the concept of a consensus sequence. The pattern is a string of amino acids that are conserved according to a user-defined threshold, are separated by "unimportant" positions, and are given position-dependent gap penalties. Although this method does not use the full

[17] A. Wlodawer, M. Miller, M. Jaskólski, B. K. Sathayanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider, and S. B. H. Kent, *Science* **245,** 616 (1989).

[18] A. Bairoch and P. Bucher, *Nucleic Acids Res.* **22,** 3583 (1994).

[19] P. Bucher, K. Karplus, M. Mooeri, and K. Hofmann, *Comput. Chem.* **20,** in press (1996).

[20] S. Henikoff and J. G. Henikoff, *Nucleic Acids Res.* **19,** 6565 (1991).

[21] L. Patthy, *J. Mol. Biol.* **198,** 567 (1987).

[22] L. Patthy, this volume [12].

information of the multiple alignment (learning set), the success of the method in finding distant relationships[22] shows that filtering important positions can be used to increase the signal-to-noise ratio (although the signal itself becomes weaker).

Property Patterns

The program PROPAT, developed by Bork and Grunwald[23] and later improved by Rohde and Bork,[24] has been applied many times in the detection of distant homologies. This method is able to generalize a pattern, even from a rather small learning set, by automatically deriving distinct combinations of physicochemical properties for each position; a vector of such properties is assigned to each amino acid (for details, see Ref. 24). It can be used for a single motif, combinations of motifs, or for whole domains, and is already a step toward profile searching, since a vector of weights (in this case penalties) is assigned to each position of the alignment (including gaps). PROPAT can search six-frame translations of DNA databases.

Flexible Patterns

The flexible patterns of Barton and Sternberg[25] combine features of motifs and profiles. The patterns can be set up in various ways but are essentially permutations of conserved blocks, separated by gaps of specified ranges, and are compared to sequences using a dynamic programming approach. The Barton approach has been applied, for example, in a survey of the DHR domain distribution.[26]

Classical Profile Method

Profile analysis as implemented by Gribskov *et al.*[27,28] is used to perform exhaustive alignment by dynamic programming of a family-based scoring matrix against test sequences. The profile is comprised of two components for each position in the alignment: (1) scores for the 20 amino acids and (2) variable gap opening and extension penalties. The amino acid substitution scores are created by summing Dayhoff exchange matrix values according

[23] P. Bork and C. Grunwald, *Eur. J. Biochem.* **191,** 347 (1990).

[24] K. Rohde and P. Bork, *Comput. Appl. Biosci.* **9,** 183 (1993).

[25] G. J. Barton and M. J. E. Sternberg, *J. Mol. Biol.* **212,** 389 (1990).

[26] C. P. Ponting and C. Phillips, *Trends Biochem. Sci.* **20,** 102 (1995).

[27] M. Gribskov, A. D. McLachlan, and D. Eisenberg, *Proc. Natl. Acad. Sci. U.S.A.* **84,** 4355 (1987).

[28] M. Gribskov and S. Veretnik, this volume [13].

to the observed amino acids in each column of the alignment. Gap penalties are reduced at positions with gaps, according to the length of the longest insertion spanning that point in the alignment. The programs PROFILEMAKE, PROFILESEARCH, and PROFILEGAP are widely available through the GCG (Genetics Computer Group) sequence analysis package,[29] making them the most frequently used programs in the field of motif and profile searches. However, the GCG PROFILESEARCH (including version 8.0) does not handle current database sizes and fails to warn the user that the search is incomplete. The TPROFILESEARCH version [available from P. Rice, EBI (European Bioinformatics Institute), Hinxton, UK] corrects this problem.

Improved Profile Methods

A number of modifications have been suggested for improving the creation of profiles that increase the sensitivity of the method. Several of these improvements have been incorporated into programs such as PROFILEWEIGHT[30] and the method of Lüthy *et al.*[31] For example, an alignment often consists of many closely related sequences together with a few rather divergent ones. The closely related sequences in the multiple alignment (learning set) offer little additional information, yet bias the profile residue scores. Sequence weighting schemes which upweight divergent sequences while downweighting closely related groupings have been found to improve profile sensitivity.[30,31]

Noise is also reduced in database searches by gap excision[30] as long insertions are sites of breakdown in homology within the family and typically lack meaningful conservation. Release 2 of PROFILEWEIGHT will also bring in new gap penalty reductions based on average gap length, rather than the single longest sequence, to better match observed gap properties in alignments.

Both TPROFILESEARCH (P. Rice, EBI, Hinxton, see Fig. 2) and the PairWise/SearchWise package (E. Birney, J. Thompson, and T. Gibson, see Fig. 2) are able to perform protein profile alignments to six-frame translations of DNA sequences. The latter programs use an extension to dynamic programming to compare the profile simultaneously to the three translation frames of a DNA strand, allowing frame jumping.[32]

[29] J. Devereux, P. Haeberli, and O. Smithies, *Nucleic Acids Res.* **12,** 387 (1984).

[30] J. D. Thompson, D. G. Higgins, and T. J. Gibson, *Comput. Appl. Biosci.* **10,** 19 (1994).

[31] R. Lüthy, I. Xenarios, and P. Bucher, *Protein Sci.* **3,** 139 (1994).

[32] T. J. Gibson, E. Birney, M. Hyvönen, A. Musacchio, and M. Saraste, *Trends Biochem. Sci.* **19,** 349 (1994).

Automated Iterative Motif Search

The motif search method MoST (for Motif Search Tool[2,11]) follows the BLAST strategy (in which gaps are not treated) in order to be able to handle the resulting alignment blocks in a proper mathematical sense. These blocks are converted to a position-dependent weight matrix following a log-odds, Bayesian-based approach and incorporating prior residue probabilities calculated from a mixture of Dirichlet distributions. MoST combines an extremely fast block search with good sensitivity.[11] In addition, automatic iterations have been incorporated, that is, database sequences scoring above a user-defined threshold are incorporated in the block alignment and the weighting for the next iteration is adapted to the new alignment. To allow for the different behaviors of protein families, manual intervention is possible at several levels. The drawback of excluding gaps will be circumvented in an improved version that handles the statistics of several blocks (E. V. Koonin, personal communication). Thus, this method is highly recommended, particularly if functionally conserved residues are surrounded by semiconserved but structurally important positions, as can be observed in distantly related enzymes.

Other Methods

In addition to improvements in techniques related to the ones mentioned above, other approaches that might prove valuable in identifying distant homologies include the application of neural networks[33,33a] and methods that try to tackle the automatic iteration of the search procedure.[34] The use of hidden Markov models (HMMs) offers the prospect of a more formal mathematical treatment for profiles. For example, using a profile description that incorporates HMMs,[19] Bucher and colleagues have been able to identify several divergent intracellular domains. A number of groups are currently working on the application of HMMs to profile searches (see Fig. 2 for a WWW pointer to one of the HMM alignment descriptions).

Parameters and Pitfalls

Multiple factors influence the choice for a certain program, such as computer resources, sensitivity, local availability, and user-friendliness. Ideally, one should know the powers and pitfalls of each of them and also how to handle the optional parameters. To help the user set up and run

[33] D. Frischman and P. Argos, *J. Mol. Biol.* **228,** 951 (1993).

[33a] J. Hanke, G. Beckmann, P. Bork, and J. G. Reich, *Protein Sci.* **5,** in press.

[34] T. M. Yi and E. S. Lander, *Protein Sci.* **3,** 1315 (1993).

pattern searches effectively, we have provided a checklist of points, which we routinely apply (Table I).

The GCG profile programs illustrate why it is important for the user to review the parameter set up. The default normalization for database entry length, which downweights the scores against long sequences, can be highly deleterious for domains in large proteins. The default normalization for amino acid bias also lowers sensitivity as it downweights characteristic residues in database entries.

Another important point to consider is the report of only the optimal hit; that is, it is easy to overlook internal repeats. Most of the programs discussed here have ways to handle such internal repeats. With the GCG programs, the user must remember to look for repeats themselves, for example, by iteratively searching the regions in a protein on either side of a match against a domain profile. Dot-plot self-comparison can provide a second independent check.

Domain Borders

As in database searches with a single sequence, choice of the proper length of the query is of great importance. Three-dimensional structures increasingly reveal how to refine the domain borders, adjustment of which often leads to increased sensitivity. One particular problem is the phasing of successive repeats; in the literature there are numerous examples in which artificial domains are published spanning the C terminus of one repeat and the N terminus of the next. These "repeats" will thus never be detected when occurring as a single domain in other proteins. Contrary to popular perception, intron boundaries are extremely unreliable guides to intracellular domain boundaries, because traceable exon shuffling appears to involve extracellular domains exclusively. Intron boundaries have been more useful in demarcating extracellular modules, but they must show cross-consistency between repeats (or proteins), as the intron positions have often rearranged subsequently, particularly in *Caenorhabditis elegans* (see review by Patthy[35] for discussion of exon shuffling and module boundaries).

Gap Treatment and Sequence Weights

Gapped regions of alignments usually contain little useful information. They are typically deleted in motif searches and specified as allowed insertion ranges. In profiles, they are given reduced gap penalties and may also be deleted. Gaps vary in their tolerance of long insertions, and this ought to be reflected in the cost assigned to each gap. As yet there is no satisfactory

[35] L. Patthy, *Curr. Opin. Struct. Biol.* **4,** 383 (1994).

TABLE I
CHECKLIST FOR MOTIF AND PROFILE SEARCHES

Point to be considered	Reason
Is multiple alignment correct?	Correct alignment is prerequisite for any search method: misalignments and frameshifts seriously degrade the signal
Is given set of sequences representative?	Some programs lack effective downweighting schemes for close relatives in learning set: sometimes one should omit very redundant sequences and select a representative set of divergent sequences
Are borderlines correctly defined or does enlargement/reduction of the studied segment make more sense?	Arbitrarily truncated segments will have a weaker signal, whereas artificially enlarged ones add noise
Is appropriate method being used for the sequence family?	For highly gapped alignments, block searches are not advisable; profiles spanning full proteins are sometimes less sensitive than restriction to a few short conserved motifs, when there is no other strong conservation
Is chosen amino acid substitution matrix appropriate?	Depending on the family, another matrix might lead to clearer/different results. For example, "soft" matrices tend to be inappropriate for short motifs
Are gaps considered appropriate?	Large insertions might occur between more conserved regions; small extracellular domains with cores mainly of disulfide bridges have more freedom for insertions/deletions than, for example, enzymes
For profile searches, have the gap penalties been optimized on trial runs?	Appropriate penalties vary with divergence of query set. Set too strong: gaps cannot be crossed. Set too weak: query profile spreads out over false positives, giving higher scores
Are apparently essential positions (e.g., required for catalysis) set to be required in pattern/profile?	Weight/penalty for such positions is often not high enough in available programs and additional information should be manually included, e.g., by stronger weight/penalties
Have all databases been searched?	Many programs are unable to search in DNA databases or six-frame translations which usually harbor additional hits; some network servers might offer out-of-date databases
Is output interpreted correctly?	Be familiar with parameters and scoring systems, to be sure of the resulting scores (e.g., normalized Z scores in PROFILESEARCH are misleading when searching for small domains in larger proteins as they upweight small sequences)
Is knowledge about putative target sequences applied appropriately?	Searching with core metabolic enzyme suggests downweighting hits with extracellular proteins, as biological context is different, but be careful as all kinds of exceptions exist and proteins with unrelated functions can be homologous

TABLE I (*continued*)

Point to be considered	Reason
Have databases been searched with putative novel members before inclusion into alignment for next iteration?	If putative novel member belongs to a family that is well-characterized and distinct or which has different conserved regions, it is likely a false positive
Has reciprocity of detections been checked?	If profile of family A identifies family B as similar, does profile of family B find family A? Caution is needed as in some cases profiles of two artificially aligned families might identify both families before noise

way of calculating these costs. The user should conduct several trial searches varying the gap parameters in order to optimize alignment of the query profile with the database proteins (especially to eliminate dramatic spreading of the profile against false positives) and the detection signal-to-noise ratio.

Sequence weighting should be used in preparing profiles. See Higgins *et al.*[36] for a discussion of the resulting benefits.

Appropriate Use of Amino Acid Similarity Matrices

New residue substitution matrices, particularly the BLOSUM[37] and Gonnet series,[38] have been found to improve the signal-to-noise ratio in profiles.[30,31] BLOSUM 45, a moderate to high divergence matrix, works well and is a good starting point.[31] However, several matrices should always be tested. As in all sequence searches, the length of the query affects the tolerance of the noise introduced by high divergence matrices. Small domains may need more stringent matrices such as BLOSUM 62 (or Gonnet Pam120/Pam160), whereas profiles for larger domains are often most sensitive at higher divergence, in which case the Gonnet series in the range Pam250–Pam350 is useful. The older Dayhoff Pam matrices and the GCG default normalized matrix perform considerably less well. Gap penalties need to be recalibrated whenever a different matrix is used.

Errors and Expressed Sequence Tags

Errors in sequence databases, particularly shifts in the translation frame, can be caused by sequencing mistakes and wrongly predicted introns and

[36] D. G. Higgins, J. D. Thompson, and T. J. Gibson, this volume [22].

[37] S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. U.S.A.* **89,** 10915 (1992).

[38] S. A. Benner, M. A. Cohen, and G. H. Gonnet, *Protein Eng.* **11,** 1323 (1994).

exons. Frameshift errors are known to be widespread, and it is likely that some 5% of entries in SWISS-PROT, the most carefully annotated sequence database, have frameshifted segments. They are so common that we now expect to find frameshifts in each new protein family being investigated, as was true for both test examples given below.

If there are frameshifts in the learning set, they can lead to the introduction of gaps at erroneous positions and affect the apparent residue conservation. Both of these effects will degrade the query pattern. If there are frameshifts (due to errors or introns) in database entries, at the protein level only part of a pattern will match, while at the DNA level the frames jump. SearchWise can track the frame jump and reveal these in the search outputs, whereas its interactive partner PairWise allows suspicious sequences to be investigated in more detail.

Another increasing problem concerns fragmented and truncated entries. Truncated open reading frames (ORFs) are often adjacent to a targeted gene sequence. Similarly, the ends of cosmid entries from genomic sequencing projects have unreliable annotation. Often, cDNAs are reported as full length but are truncated at their N termini (and occasionally C termini). Even more complicated to deal with are the EST entries (expressed sequence tags), single gel reads (usually ~300 bases) of random cDNAs. Not only is there a high frameshift error rate ($\geqslant$10% of entries, sometimes multiple), but the hits are, by definition, fragments of the query and so search scores will be less than for full-length sequences. To deal with such a situation, parameters might be specially tuned as, for example, implemented in SearchWise (and applied by Aasland *et al.*[39]), in which case the EST databases should be searched separately.

Significance Assessments

Having discussed major sources for improvement of sensitivity, the assessment of significances for tempting hits remains problematic. Most of the programs provide some significance estimates, but all these calculations have to make certain assumptions that lead in practice to limitations. The calculation of Z scores (normal deviates) is estimated from the quality of the input sequences compared to the total distribution of scores. However, the score distribution does not follow a Gaussian distribution, as assumed in many Z-score calculations; Dirichlet mixtures are currently used to describe phenomena such as tails in the distributions. Another scoring system is based on p values that give the probability of a match by chance. The probability may not take into account sequence and residue biases, and is

[39] R. Aasland, T. J. Gibson, and A. F. Stewart, *Trends Biochem. Sci.* **20,** 56 (1995).

dependent on database size, so that a given value (e.g., 10^{-6}) will become less significant in time as the databases expand.

The better the statistical description, the more reliable will be the resulting scores. (Thus, here is another hidden parameter that has an influence on the results but differs from family to family.) Deficiencies in the current schemes are widely acknowledged, and efforts are being made to improve statistical assessments. Claverie[40] has suggested statistics aimed at better discrimination for the threshold between true and false hits for ungapped profiles of fixed length, assuming that random matches behave according to the extreme value distribution. The distribution of the family members is considered in MoST when estimating the ratio of the expected number of sequence segments with a given score to the observed number (parameter r in MoST), yielding a rather accurate estimate for a given alignment block.

In any event, the significance values of the different methods only try to eliminate false-positive hits. In this they are largely successful for globular sequences but are less reliable for reduced complexity sequences. As judged from the numerous very similar three-dimensional structures without obvious sequence similarity, many homologs do not meet the significance criteria of the particular method and are likely to escape attention when applying motif and profile searches.

Assessing Statistically Insignificant Hits

The way to assess weak candidate hits is to consider what are the constraints which must apply between sequences as a consequence of homology between them. These constraints are predominantly structural in nature, and yet they can often be applied even in the absence of a solved structure. Table II provides a checklist designed to help in weeding out the false hits: again we routinely use these checks. This logic was used, for example, in detecting highly divergent PH domains.[41] Conserved hydrophobic residues were assumed to be in the core, whereas runs of hydrophilic residues tolerating gaps were assumed to be in exposed loops. From the periodicity of the conserved residues it was possible to infer the number of α helices and β strands in the domain. Borderline hits from profile searches were then only accepted when the sequence was fully compatible with all the predicted secondary structures, the hydrophobic core residues, the predominantly hydrophilic surface residues, and the absence (or rarity) of Pro and Gly in helices and strands. Later, solved PH domain structures

[40] J.-M. Claverie, *Comput. Chem.* **18,** 287 (1994).

[41] A. Musacchio, T. J. Gibson, P. Rice, J. Thompson, and M. Saraste, *Trends Biochem. Sci.* **18,** 343 (1993).

TABLE II
CHECKLIST FOR EVALUATING PLAUSIBILITY OF WEAK HITS

Examinations on hit	Reason
Is amino acid distribution consistent with globular (cytosolic, extracellular), integral membrane, coiled-coil, fibrous, or random-coil structure?	These are mutually exclusive structural classes that should not overlap within a domain (although they can be juxtaposed in multidomain proteins). High scoring random coil is not a good indicator of homology
Is there structural information known for query or hit?	Knowledge of three-dimensional structure greatly facilitates evaluation, as constraints from hydrophic core, catalysis, etc., can be included
Is there a partial overlap of hit with established domain class in reciprocal search?	Partial overlap immediately rules out potential similarity. By definition, globular domains do not overlap (although they can be inserted into loops in other domains)
Is full domain potentially present?	Globular structure is stabilized by interactions in hydrophobic core. The presence of only half a globular domain is thus very rarely observed
Is there match to all conserved alignment blocks?	Conserved blocks usually indicate secondary structural elements
Is there match to all highly conserved hydrophobic residues?	Conserved hydrophobic residues are essential to given hydrophobic core. Very few exceptions are tolerated
Do most positions that are aligned to unconserved positions in query have hydrophilic residues?	Surface residues are usually hydrophilic and are unconserved unless binding other molecules. Multiple mismatched hydrophobic residues are contrary indicators. (Surprisingly frequently, transmembrane regions are erroneously aligned to cytosolic proteins)
Has Pro been aligned to position in block where it was not seen before?	Pro is favored in the N-terminal 3 residues of α helices. Any deeper and it breaks H bonds. It is allowed on edge β strands. It breaks H bonds on internal strands. Exceptions are rare and cannot be arbitrarily invoked for weak hits
Has Gly been aligned to position in block where it was not seen before?	The lack of a side chain reduces helix and strand stability. Gly aligned to small hydrophobic residues (Ala, Val, Cys) may indicate a plausible tight packing arrangement; otherwise, only occasional exceptions may be tolerated
Is segment rich in Gly, Pro, Asn, Ser aligned to block poor in these residues?	Such alignment indicates that a loop region is erroneously aligned to a secondary structure element

TABLE II (*continued*)

Examinations on hit	Reason
Are matches to blocks consistent with block secondary structure?	Secondary structures of matched blocks should be identical. In addition to above rules, amino acid preferences may be indicative: e.g., aligning a sequence composed of β-preferring residues like Ile, Val, Thr, Ser onto an α helix would be highly implausible (unless these were already favored in aligned sequences)
Have new insertions/deletions appeared in conserved regions?	Alignment blocks are usually conserved due to structural or functional constraints; therefore, large or frequent insertions and deletions are unlikely
For Cys-rich sequences, do Cys patterns match?	Number and spacing of Cys residues distinguish between classes of extracellular disulfide-rich modules, as well as (often with His) intracellular zinc fingers, e.g., GAL4
Are functions of hits compatible?	On one hand do not overinterpret results to fit a tempting functional context; on the other hand, some functional aspects should be considered (e.g., query proteins are extracellular, hit is a metabolic enzyme)
Does additional functional or biochemical information provide some clues for homology?	Already identified catalytic residues, disulfide bridges, mutation data, etc., add constraints that can be helpful in excluding false positives

vindicated the structural assignments, thereby lending credence to the PH domain detections.[32]

Secondary structure predictions can help in clarifying whether two or more sequence families are structurally related. It is now recognized that secondary structure predictions based on multiple alignments are often quite accurate. The server of Rost and Sander[42] usually gives good results, in particular when the reliability scores are taken into account. For example, pattern searches with a phosphate-binding motif present in certain β/α-barrel enzymes, such as glycolate oxidase, weakly picked up a surprising number of different enzyme families with unknown structures and, mostly, poorly investigated mechanisms. These families had no apparent sequence similarity to each other. Secondary structure predictions,[42] based on alignments of the individual families, gave alternating β/α predictions, consistent with the β/α-barrel structure. This information, together with the length of the proteins and the C-terminal location of the phosphate-binding motif, was critical to assigning homology between a large set of enzymes acting

[42] B. Rost and C. Sander, *J. Mol. Biol.* **232,** 584 (1993).

on heterocyclic compounds in diverse pathways, including histidine biosynthesis, purine metabolism, and thiamin biosynthesis.[9]

To illustrate the strategies explained above, we present two examples in which we are able to identify new members of well-established domain families using several different pattern search programs. In the process, we identified two frameshifted entries, and one with a wrongly predicted spliced exonic sequence, which undoubtedly hindered earlier detection of the domains.

Retrieving GAL4 Domains

Our first example is the retrieval of fungal zinc binuclear cluster domains, a specific DNA-binding domain found in numerous fungal transcription factors. The three-dimensional structure has been determined for the corresponding domains in PPR1 and in GAL4 (see Marmorstein and Harrison[43] and references therein). GAL4 activates galactose catabolism in yeast and is the best characterized member of the family. GAL4-like DNA-binding domains are extremely widespread in fungi and are involved in numerous transcription regulation pathways, and yet not a single one of these domains has been found in any other eukaryote, nor in bacteria. With more than 50% of the genomic sequence of yeast available in public databases, an up-to-date collection would yield a first realistic estimate of their total number. Another interesting problem is the mode of spreading of this domain: Is it simple gene duplication and subsequent modification, or, as suggested by the considerable number, is a sort of domain shuffling the reason for their frequency? In the latter case, the location of the domain should vary within the proteins. So far, it has been mostly found at the N terminus of transcription regulators, including GAL4 itself.

Given the defined GAL4 domain borderlines (verified by the three-dimensional structure) and assuming a nonbiased amino acid distribution within the domain, a first quick homology search can be performed by BLASTP using only the domain itself as the query, in order to reduce the noise caused by matches of other parts of the GAL4 protein. A scan of the well-annotated SWISS-PROT database (release 31) records 9 hits with significant matches below a probability of matching by chance of $p = 10^{-7}$. (Although this is a strict value for assessing BLAST outputs, the first false positive, a viral coat protein, had a p value of 9.4×10^{-6}.)

These sequences will normally be aligned in the next step, an option either provided within the motif/search/programs or performed with a multiple alignment program. The GAL4 alignment (Fig. 3) shows six invariant cysteines, essential for structural zinc binding, and strong conservation of

[43] R. Marmorstein and S. C. Harrison, *Genes Dev.* **8**, 2504 (1994).

several other residues, in particular positively charged residues between the second and third cysteine, that are important in DNA binding. Only one gap was opened in the starting set. We compared here three methods (profile/SearchWise, PROPAT, MoST); all three collect the same number of sequences in the protein databases, with no false positives in SWISS-PROT (a few difficult cases are discussed below). All methods performed much better than iterative BLASTP database searches. Thus, another application of motif and profile search programs becomes apparent with the presence of large families in databases, namely, the fast collection of all members. In SWISS-PROT release 31 we found 42 GAL4-like domains, whereas only 37 of them will be found using BLASTP (with the threshold $p = 10^{-7}$). Even if the remaining 5 can be pulled out of the twilight zone by various output evaluation procedures, preparation and evaluation of many BLASTP runs require a considerable effort compared, for example, to a single run (after the initial BLASTP search with GAL4) of MoST.

When extending the search to other data resources, including nucleotide sequence databases, 60 sequences were retrieved (Fig. 3). Fifty entries contain annotation hinting at the presence of GAL4-like domains: of the remainder, only two had been published at the time of writing, but the domain presumably escaped attention in several other cases. During the course of the searches, two erroneous sequences were detected, both of which were already entered in SWISS-PROT: (1) Yhl6_Yeast lacks the N-terminal GAL4 domain because it contains a frameshift (detected by SearchWise) and so was not placed in the predicted ORF YHR056c by automatic translation procedures (Fig. 4a), and (2) Alcr_Emeni has a missing exon (detected by several programs) due to incorrect interpretation of the splicing pattern (Fig. 4b).

Thus, the example of GAL4 underlines some of the points discussed. The pattern searches led to the identification of previously undetected GAL4 family members and also revealed two common kinds of errors (one in the sequence itself and one in the interpretation of the raw DNA sequence). From the 44 GAL4 domains we found in yeast, spread over all the chromosomes, we can extrapolate to the whole genome and expect more than 80 GAL4 domains in total. The majority of the proteins seem to be colinear, sharing another (more weakly) conserved domain (P. Bork, unpublished results), so that the frequency in yeast seems to be the result of extensive gene duplication rather than domain shuffling.

Jak Kinases Contain SH2 Domains

We also conducted pattern searches with SH2, a well-known domain in signaling proteins that recognizes and binds phosphotyrosine-containing

AFLR_ASPFL	23	RKLRDSCTSCASSKVRCTKEKP- 1-CARCIERGL--ACQYMVSKRMGRN	P41765	
ALCR_EMENI	6	RRQNHSCDPCRKGKRRCDAPEN-11-CSNCKRWNK--DCTFNWLSSQRSK	P21228	a
AMDR_ASPOR	13	GNGSAACIHCHRRKVRCDARIV- 3-CSNCRSAGK-ADCRIHEKKKRLAV	Q06157	
AMDR_EMENI	13	GNGSAACVHCHRRKVRCDARLV- 3-CSNCRSAGK-TDCQIHEKKKKLAV	P41044	
ARG2_YEAST	15	AKTFTGCWTCRGRKVKCDLRHP- 1-CQRCEKSNL--PCGGYDIKLRWSK	P05085	
CAT8_YEAST	64	YRIAQACDRCRSKKTRCDGKRP- 1-CSQCAAVGF--ECRISDKLLRKAY	P39113	
CB32_YEAST	8	LKSKHPCSVCTRRKVKCDRMIP- 0-CGNCRKRGQDSECMKSTKLITASS	P40969	
CYP1_YEAST	58	NRIPLSCTICRKRKVKCDKLRP- 1-CQQCTKTGVAHLCHYMEQTWAEEA	P12351	
CZF1_CANAL	312	KRSRMGCLTCRQRKKRCCETRP- 1-CTECTRLRL--NCTWPKPGTEHKN	P28875	
DA81_YEAST	144	GNLMGSCNQCRLKKTKCNYFPD- 3-CLECETSRT--KCTFSIAPNYLKR	P21657	
GAL4_YEAST	5	SSIEQACDICRLKKLKCSKEKP- 1-CAKCLKNNW--ECRYSPKTKRSPL	P04386	
LAC9_KLULA	89	EVMHQACDACRKKKWKCSKTVP- 1-CTNCLKYNL--DCVYSPQVVRTPL	P08657	
LEUR_YEAST	31	RKRKFACVECRQQKSKCDAHER- 4-CTKCAKKNV--PCILKRDFRRTYK	P08638	
LY14_YEAST	153	KYSRNGCSECKRRRMKCDETKP- 1-CWQCARLNR--QCVYVLNPKNKKR	P40971	
MA3R_YEAST	2	TLVKYACDYCRVRRVKCDGKKP- 0-CSRCIEHNF--DCTYQQPLKKRGS	P38157	
MA6R_YEAST	2	GIAKQSCDCCRVRRVKCDRNKP- 0-CNRCIQRNL--NCTYLQPLKKRGP	P10508	
NIRA_EMENI	36	RCVSTACIACRRRKSKCDGNLP- 1-CAACSSVYH-TTCVYDPNSDHRRK	P28348	
NIT4_NEUCR	47	RCVSTACIACRRRKSKCDGALP- 1-CAACASVYG-TECIYDPNSDHRRK	P28349	
PDR1_YEAST	40	SKVSKACDNCRKRKIKCNGKFP- 0-CASCEIYSC--ECTFSTRQGGARI	P12383	
PDR3_YEAST	9	SKVSTACVNCRKRKIKCTGKYP- 0-CTNCISYDC--TCVFLKKHLPQKV	P33200	
PPR1_YEAST	28	SKSRTACKRCRLKKIKCDQEFP- 1-CKRCAKLEV--PCVSLDPATGKDV	P07272	
PUT3_YEAST	28	QRSSVACLSCRKRHIKCPGGNP- 0-CQKCVTSNA--ICEYLEPSKKIVV	P25502	
QA1F_NEUCR	70	QRVSRACDQCRAAREKCDGIQP- 1-CFPCVSQGR--SCTYQASPKKRGV	P11638	
QUTA_EMENI	43	QRVSRACDSCRSKKDKCDGAQP- 1-CSTCASLSR--PCTYRANPKKRGL	P10563	
SUC1_CANAL	7	APYTRPCDSCSFRKVKCDMKTP- 0-CSRCVLNNL--KCTNNRIRKKCGP	P33181	
THI1_SCHPO	33	RRVFRACKHCRQKKIKCNGGQP- 0-CISCKTLNI--ECVYAQKSQNKTL	P36598	
UGA3_YEAST	11	KYSKHGCITCKIRKKRCSEDKP- 1-CRDCRRLSF--PCIYISESVDKQS	P26370	
UME6_YEAST	765	TRSRTGCWICRLRKKKCTEERP- 1-CFNCERLKL--DCHYDAFKPDFVS	P39001	d
YAF1_YEAST	60	NRILFVCQACWKSKTKCDREKP- 1-CGRCVKHGL--KCVYDVSKQPAPR	P39720	
YB00_YEAST	101	SRVTKACDYCRKRKIRCTEIEP- 4-CRNCIKYNK--DCTFHFHEELKRR	P38114	
YB89_YEAST	34	KNTNVACVNCRRLHVSCEAKRP- 0-CLRCISKGLTALCVDAPRKKSKYL	P38140	
YB90_YEAST	24	GRTFTGCWACRFKKRRCDENRP- 1-CSLCAKHGD--NCSYDIRLMWLEE	P38141	
YBG6_YEAST	51	HRPVTSCTHCRQHKIKCDASQN- 4-CSRCEKIGL--HCEINPQFRPKKG	P34228	
YBO3_YEAST	50	KKASHACDQCRRKRIKCRFDKH- 3-CQGCLEVGE--KCQFIRVPLKRGP	P38073	
YCO1_YEAST	11	SKAFKTCLFCKRSHVVCDKQRP- 0-CSRCVKRDIAHLCREDDIAVPNEM	P19541	
YCZ6_YEAST	9	PRLRLVCLQCKKIKRKCDKLRP- 1-CSRCQQNSL--QCEYEERTDLSAN	P25611	
YE14_YEAST	12	SRVTKACDRCHRKKIKCNSKKP- 0-CFGCIGSQS--KCTYRNQFREPIE	P39961	
YHL6_YEAST	9	VRKPPACTQCRKRKIGCDR!KP- 1-CGNCVKYNK-PDCFYPDGPGKMVA	P38781	b
YHX8_YEAST	16	TTELYSCARCRKLKKKCGKQIP- 1-CANCDKNGA--HCSYPGRAPRRTK	P38699	
YIN0_YEAST	15	RRVTRACDECRKKKVKCDGQQP- 0-CIHCTVYSY--ECTYKKPTKRTQN	P40467	
YJ16_YEAST	41	GRAHRACIACRKRKVRCSGNIP- 0-CRLCQTNSY--ECKYDRPPRNSSV	P39529	
YJX9_YEAST	14	KSIQTACEFCHTKHIQCDVGRP- 0-CQNCLKRNIGKFCRDKKRKSRKRI	P42950	
YK44_YEAST	13	HRITVVCTNCKKRKSKCDRTKP- 0-CGTCVRLGDVDSCVYLTDSSGQPE	P36023	
YKD8_YEAST	41	TKASRACDQCRKKKIKCDYKDE- 3-CSNCQRNGD--RCSFDRVPLKRGP	P32862	
YKW2_YEAST	18	RKPAKSCHFCRVRKLKCDRVRP- 1-CGSCSSRNR-KQCEYKENTSAMED	P35995	
MALX/YEAST	7	TCAKQACDCCRIRRVKCDGKRP- 0-CSSCLQNSL--DCTYLQPSRKRGP	L12223_1	
PRIB/LINED	14	VRGARACTTCRAAKMKCVGAED- 4-CQRCKRANV--QCIFEKHRRGRKP	D14489_1	
PAH2/PICAN	?	RKVGAACVICHRRKIKCDIGTA- 3-CSKCKELKVESQCVLHKRRRKTDG	U22930_1	d
SIP4/YEAST	40	VRKAHACDRCRLKKIKCDGLKP- 1-CSNCAKIDF--PCKTSDKLSRRGL	U17643_1	
UAY/ASPNI	61	FRNVSACNRCRQRKNRCDQRLP- 1-CQACEKAGV--RCVGYDPITKREI	X84015_1	d
YX1/CANPE	194	KGNPNPCDHCRRRQIKCITVPN- 3-CVQCETKGI--KCTHSESPSNPAL	X02903_3	c
YFX1/YEAST	2	ARNRQACDCCCIRRVKCDRKKP- 0-CKCCLQHNL--QCTYLRPLKKRGP	D50617	
YLX1/YEAST	38	NKSKTGCDNCKRRRVKCDEGKP- 1-CKKCTNMKL--DCVYSPIQPRRRK	U19027_9	d
YLX2/YEAST	9	VKPSFVCLRCKQRKIKCDKLWP- 1-CSKCKASSS--ICSYEVEPGRINK	Z47973_17	d
YLX3/YEAST	35	KGRSRSCLLCRRRKQRCDHKLP- 1-CTACLKAGI--KCVQPSKYSSSTS	U17243_1	d
YMX1/YEAST	81	LRVQKACELCKKRKVKCDGNNP- 0-CLNCSKHQK--ECRYDFKATNRKR	Z49211_7	
YMX2/YEAST	25	RKVIKSCAFCRKRKLKCSQARP- 1-CQQCVIRKL-PQCVYTEEFNYPLS	U17244_9	
YMX3/YEAST	70	KRNSFACVCCHSLKQKCEPSDV- 7-CRRCLKHKK--LCKFDLSKRTRKR	Z46373_6	d
YOX1/YEAST	130	KRVSKACDHCRKRKIRCDEVDQ- 4-CSNCIKFQL--PCTFKHRDEILKK	X83121_8	d
YPX1/YEAST	2	SIVRSQCDCCRVRRVKCDRNRP- 0-CDRCRQRNL--RCTYLQPLRKRGP	U25841_11	

FIG. 3. GAL4 domain alignment. From left to right: first column, names (SWISS-PROT codes are given when available); second column, position of the first amino acid; third column, sequence; fourth column, database accession number (an underscore indicates the number of the ORF in larger cosmids). Numbers within the alignments indicate omitted amino acids

peptides. Solved structures of SH2/peptide complexes[44] have revealed the SH2 core and functional residues, of which the most conserved is a phosphate-binding arginine. We began with a search of SWISS-PROT (release 31) using BLASTP and the archetypal chicken Src SH2 domain. The BLASTP search detected numerous SH2 domains before the highest false positive ($p = \sim 10^0$), but many were well below our very conservative threshold of $p = 10^{-7}$. Many of the top hits were also tyrosine kinases almost identical to Src, which would add little or no value to the patterns. Therefore, a representative set of 27 SH2 domains above the threshold of 10^{-7} was used to initiate the different motif and profile searches. Again, all three methods performed well, with the profiles being slightly ahead as they can cope best with many (sometimes large) insertions (see alignment of SH2 domains in Higgins *et al.*[36]). Although the number of iterations and the computer time used are different for each program, all programs could detect known divergent members of the family such as the only described yeast SH2 domain (Spt6), its *C. elegans* homolog Emb5, and the STAT family of transcription factors.

In addition to the expected hits, both PROPAT and SearchWise indicated the presence of divergent SH2 domains in the Jak group of tyrosine kinases, which were not assigned SH2 domains in the annotation. SH2 in Jak kinases would be of clear biological significance since the great majority of cytoplasmic tyrosine kinases possess the domains. Both programs picked up the Jak entries more strongly in the later iterations, as increasingly divergent SH2 domains were added to the alignments. The Jak entries scored better than the STATs. The core structural and surface functional positions, including the critical conserved phosphate-binding positive charge, were all satisfied by the Jaks. Reciprocal searches (as recommended in Table I) with SearchWise and a profile of the putative Jak kinase SH2 domains listed 55 SH2-containing entries before the first false positive in SWISS-PROT. Therefore, the Jak kinases fulfill the criteria in Table II for divergent homologs.

[44] G. Waksman, S. E. Shoelson, N. Pant, D. Cowburn, and J. Kuriyan, *Cell* (*Cambridge, Mass.*) **72,** 779 (1993).

in a loop region. Cys residues essential for the binuclear cluster are highlighted in boldface type. Marker letters: a, sequence differs from the protein in SWISS-PROT release 31 due to splicing revision (see Fig. 4a); b, symbol ! indicates frameshift, and sequence differs in SWISS-PROT release 31 (see Fig. 4; both errors have been corrected in subsequent SWISS-PROT releases, on the basis of this report); c, this protein (and the frameshifted YHL6) had not been reported to contain a GAL4 domain; d, GAL4 domains are not annotated in the database entries and thus some of these GAL4 domains were probably not detected earlier.

a

```
Score 9750
Aligned Ranges:
10-40 (profile)
5445-5541 (sequence)
Showing backward strand

                         **  *****  **    ! * *  * *     * *
Gal4 Profile:         10 ACDRCRKRKVKCDG+++KRPpCSRCAKRG.LECTY 40
DNA Translation:         ACTQCRKRKIGCDR^^^^KPICGNCVKYNKPDCFY
Embl:SCH8025:      5541' gtactaaaaagtgaGCCAacatgatgataacgttt
                         cgcaggagatggag    actggagtaaaacagta
                         tccacggagcgccg    agatgtccgtcggcttt 5445'
```

b

```
Score 9190
Aligned Ranges:
1-36 (profile)
1080-1277 (sequence)

                          *    ** *** *  **
Gal4 Profile:          1 KRVSTACDRCRKRKVKCDGKRP............................
DNA Translation:         RRQNHSCDPCRKGKRRCDAPVGCRYRLPSVCTDSR*DVTQENRNEANENG
Embl:ANALCR:        1080 cccacatgctcagacctggcggtctcccagtagactggacgaaaggagag
                         ggaaaggacggagagggacctgggagtcgtgcagggatcaaagaacaaag
                         acgttcctctcgcgacttcgatcatgcccgctctcataagatacgctacc


                            ** *      **
Gal4 Profile:         23 ..pCSRCAKRGLECTY 36
DNA Translation:         WVSCSNCKRWNKDCTF
Embl:ANALCR:        1230 tgtttatactaagtat
                         gtcgcagaggaaagct
                         gtgtatcgtgcgttcc 1277
```

FIG. 4. (a) Highest scoring local alignment produced by PairWise for the GAL4 profile against all three reverse frames of the yeast genomic cosmid embl:SCH8025. A single base insertion has caused a frameshift in the middle of a GAL4 domain, which would otherwise belong to the N terminus of the adjacent ORF YHR056c (predicted range −5426 to −2828). Translated DNA codons are shown in lowercase, and nucleotides spanned by the frameshift site are in uppercase and capped by the ^ symbol. Cys residues essential for the binuclear cluster are highlighted in boldface type. A WWW server that allows fast detection of frameshifts even in large cosmids (J. Boyle, N. P. Brown, and P. Bork, unpublished) will be available under http://www.embl-heidelberg.de/~boyle/errors/. (b) Highest scoring local alignment produced by PairWise for the GAL4 profile against the three forward frames of the *Aspergillus nidulans AlcR* gene in embl:ANALCR. An in-frame insertion including a stop codon separates two halves of a GAL4 domain, with only the latter part being in the predicted *AlcR* translation. Good matches to fungal splice donor, acceptor, and branch-point consensus sequences are shown in boldface italic type and reveal an overlooked exon containing the N-terminal portion of the GAL4 domain. Incorrect splicing inferences are not uncommon. In-frame introns (especially when lacking stop codons) are usually overlooked in fungi such as yeast due to the perception that introns are rare.

Surprisingly, two separate mouse Jak3 entries had different scores. One was well detected, but the other was below a number of false hits. Comparison of this sequence,[45] using the frameshifting algorithm of PairWise, to human, rat and the other mouse Jak3 revealed that it possessed seven frameshifted regions, of which two fell in the SH2 domain, accounting for the reduced profile score.

The exercise here reveals that the pattern methods can detect SH2 domains in Jak tyrosine kinases, which makes good sense, implying either inter- or intramolecular phosphotyrosine peptide recognition by these kinases. A review of the literature reveals that the Jak SH2 domains had been proposed earlier[46] yet were apparently rejected within the field, presumably to the detriment of Jak kinase research. The alignment of Jak and other SH2 domains is presented elsewhere in this volume in the Clustal alignment chapter,[36] where it is used as a divergent alignment test case.

Conclusion

There are numerous examples in which predictions based on motif and profile searches were useful as guides in further research, while themselves being verified by various experimental approaches. On the debit side, there are also a considerable number of "black sheep" that have caused resources to be squandered in exploration of wrong hypotheses. We wish to stress the two checklists provided here in Tables I and II, one for setting up and running the searches and the other for evaluating the borderline top hits. If these are followed, most of the suggestive, yet false, hits should be identified and discarded. Ultimately, solved three-dimensional structures are the arbiters of truth for homology predictions. Currently, the identification of new protein domains is followed almost immediately by the determination of the three-dimensional structure by NMR or X-ray. For more than 50% of the known intra- or extracellular modules (mostly defined by motif and profile searches), a three-dimensional structure is already available for at least one member of the family.[47] In addition to the identification of numerous domains[47] (e.g., the discovery of a nuclear domain superfamily present in cyclin, transcription factor IIB (TFIIB), and retinoblastoma proteins[48]), experimentally verified predictions with functional implications from our own work include the prediction of a type X polymerase in yeast,[3]

[45] S. G. Rane and E. P. Reddy, *Oncogene* **9,** 2415 (1994).

[46] A. G. Harpur, A.-C. Andres, A. Ziemiecki, R. R. Aston, and A. F. Wilks, *Oncogene* **7,** 1347 (1992).

[47] P. Bork and A. Bairoch, *Trends Biochem. Sci.* **20,** poster, March issue (1995).

[48] T. J. Gibson, J. D. Thompson, A. Blocker, and T. Kouzarides, *Nucleic Acids Res.* **22,** 946 (1994).

ATPase activities of several prokaryotic cell cycle proteins,[49] and homodimerization of the Norrie's disease protein via a specific disulfide bridge.[50]

Two additional points should be noted. (1) Soon essentially all sequences will have homologs in public databases. (2) Motif and profile search methods are being actively developed at a number of institutions and can be expected to be significantly improved. As a result these methods will continue to be very valuable tools.

Acknowledgments

We are indebted to Philipp Bucher, Eugene Koonin, and Joe Lewis for helpful comments on the manuscript and to the many colleagues who have also been involved in our cited work.

[49] P. Bork, A. Valencia, and C. Sander, *Proc. Natl. Acad. Sci. U.S.A.* **89,** 7290 (1992).

[50] T. Meitinger, A. Meindl, P. Bork, B. Rost, C. Sander, M. Haasemann, and J. Murken, *Nat. Genet.* **5,** 376 (1993).

[12] Consensus Approaches in Detection of Distant Homologies

By LASZLO PATTHY

Introduction

Recognition of homologies may provide important hints about the structure and function of proteins; therefore, there is a growing interest in methods of sequence comparison. The FASTA and FASTP programs use a rapid sequence comparison algorithm which, by identifying proteins with high similarity scores, is useful for the detection of related sequences.[1,2] When sequence similarity is low, however, it is difficult to decide whether this similarity is due to common ancestry or whether it merely reflects chance similarity of unrelated proteins. In such cases, statistical tests are used to decide whether two sequences are more similar than would be expected by chance. In this approach, the similarity score of the actual comparison is compared with the distribution of the scores determined for pairs of a large number of random permutations of the two sequences, and the standard deviation of the comparison above the mean of the randomized comparisons is calculated. The use of high cutoff values increases the confi-

[1] D. J. Lipman and W. R. Pearson, *Science* **227,** 1435 (1985).

[2] W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85,** 2444 (1988).