Issues in searching molecular sequence databases

Stephen F. Altschul, Mark S. Boguski, Warren Gish & John C. Wootton

Sequence similarity search programs are versatile tools for the molecular biologist, frequently able to identify possible DNA coding regions and to provide clues to gene and protein structure and function. While much attention had been paid to the precise algorithms these programs employ and to their relative speeds, there is a constellation of associated issues that are equally important to realize the full potential of these methods. Here, we consider a number of these issues, including the choice of scoring systems, the statistical significance of alignments, the masking of uninformative or potentially confounding sequence regions, the nature and extent of sequence redundancy in the databases and network access to similarity search services.

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Correspondence should be addressed to M.S.B.

The advent of rapid DNA sequencing technology in the mid-1970s led to an information explosion that continues unabated today. Molecular sequence data have become the common currency of biomedical research and often provide unexpected links among diverse biological systems. These connections accelerate research progress and may even open up entirely new fields of inquiry. One approach to discovering such connections, database "homology" searching, has been executed countless times, often with surprising results and has become an essential method for the molecular biologist. While the particular algorithm used is of course important, the effectiveness of database searches is dependent as well on a large number of correlative factors, many of which tend to be overlooked or dealt with an an inefficient or *ad hoc* manner. These include the following:

Scoring systems. Most database search algorithms rank alignments by a score, whose calculation is dependent upon a particular scoring system. Usually there is a default system, but it may not be ideal for a user's particular problem. For example, haemoglobin subunits used to be regarded as "typical" proteins and are often still used as benchmark query sequences for evaluating new database search techniques and scoring systems. However today it is more common to encounter much larger and more complex sequences (see below) and methods developed and optimized for small, uniformly-conserved, single-domain proteins are inadequate. Scores that are best for detecting similarities between greatly diverged sequences differ from those best for detecting short but nearly identical segments[1,2]. Optimal strategies for detecting similarities between DNA protein-coding regions differ from those for non-coding regions[3,4]. Special scoring systems for detecting frame-shift errors in the databases have recently been described[5]. A database search program should therefore make a variety of scoring systems available and users should be aware of which ones are best suited to their problems.

Alignment statistics. Given a query sequence, most database search programs will produce an ordered list of imperfectly matching database similarities, but none of them need have any biological significance. An important question is how strong a similarity is necessary to be considered surprising. United by a common theory, a number of analytic[6–9] and empirical results[2,10–13] are now available for assessing database search results. However, one still sees occasional extravagant claims in the literature, usually springing either from misapplication of the normal distribution or from an absence of critical statistical analysis.

Databases. The use of an up-to-date sequence database is clearly a vital element of any similarity search. Sequence relationships critical to important discoveries have on occasion been missed because old or incomplete databases were employed. However, the variety of databases available, and their overlapping coverage, has the potential to render similarity searching cumbersome and inefficient. This no longer need be the case. Timely access to complete and "nonredundant" sequence databases has become relatively simple and inexpensive.

Database redundancy and sequence repetitiveness. Surprisingly strong biases exist in protein and nucleic acid sequences and sequence databases. Many of these reflect fundamental mosaic sequence properties that are of considerable biological interest in themselves, such as segments of low compositional complexity or short-period

Table 1 The BLAST family of programs

Program[a]	Query sequence	Database sequences	Comments
BLASTP	protein	protein	• Default scoring matrix[b] is BLOSUM62; change with command line option "M=PAM250", for example • Low-complexity masking with "-filter" option; choice of either the SEG[87] and XNU[74] algorithms
BLASTN	nucleotide (both strands)	nucleotide	• Parameters optimized for speed, not sensitivity; not intended for finding distantly-related, coding sequences • Automatically checks complementary strand of query
BLASTX	nucleotide (six-frame translation)	protein	• Very useful for preliminary data containing potential frameshift errors[4] • Nine different genetic codes available[c]; change with command line "C=1" (vertebrate mitochondrial) for example • Low-complexity filter option as for BLASTP
TBLASTN	protein	nucleotide (six-frame translations)	• Essential for searching protein queries against dbEST[60] • Often useful for finding undocumented open reading frames or frameshift errors in database sequences • Same genetic code options as for BLASTX

[a] These programs are available through the BLAST Network and e-mail servers (see text) and the source codes are available by anonymous ftp on ncbi.nlm.nih.gov.

[b] More than 65 different PAM[1,2,35,36,40], BLOSUM[41,45] and other scoring matrices are available. PAM120 or BLOSUM62 are best for general purposes but a useful combination for detecting strong and short to long and weak similarities consists of PAM30, PAM120 and PAM250 (ref. 2).

[c] Default genetic code (C=0) is "standard" or "universal" code. Other codes available include: 1, Vertebrate mitochondrial; 2, Yeast mitochondrial; 3, Mold mitochondrial and mycoplasma; 4, Invertebrate mitochondrial; 5, Ciliate macronuclear; 6, Protozoan mitochondrial; 7, Plant mitochondrial; and 8, Echinodermate mitochondrial.

repeats. Databases also contain some very large families of related domains, motifs or repeated sequences, in some cases with hundreds of members. In other cases there has been a historical bias in the molecules that have been chosen for sequencing. In practice, unless special measures are taken, these biases very commonly confound database search methods and interfere with the discovery of interesting new sequence similarities. Problems include the occurrence of misleading, spuriously-high scores, ambiguities in the phase of sequence alignments and overwhelmingly large output lists in which interesting results may be inconspicuously buried. We shall describe some recently developed methods that largely solve these problems by automatically detecting and masking potentially confounding subsequences.

Failure to deal properly with the factors described above can result in chance similarities being claimed significant, or biologically important relationships being overlooked. Here, we shall discuss these and several other issues in database searching. While we will frequently use the BLAST programs[4,14] (Table 1) as examples, most of the questions considered have quite general relevance.

Algorithms and programs

The earliest sequence comparison studies focussed on the alignment of complete sequences[15–17]. However, with the recognition that proteins frequently share only isolated regions of similarity, corresponding for instance to structural motifs or active sites, attention shifted to algorithms for local alignment[18–21]. Essentially all database search methods have been based upon measures of local sequence similarity.

In general, local alignments are assessed by means of a score, which is computed as the sum of scores for aligned pairs of residues and scores for gaps[18]. How these scores are chosen, and what they signify, is discussed below. The time necessary to find alignments that optimize such scores is sufficiently great that, for most practical purposes, either parallel architecture machines[22–26] or heuristic methods such as Fasta[27,28] are required. The problem may be simplified by forbidding gaps. This leads to faster heuristic methods such as the BLAST algorithms[4,14] (Table 1), as well as to efficient hardware implementations[29]. While some sensitivity to weak similarities may be lost by eschewing gaps[30], easier generalization[31] and rigorous statistical results[6–9] become available. Alternatively, local alignments may be assessed in a more sophisticated manner than by the simple sum of substitution and gap scores[32]. This may lead to more sensitive detection of weak similarities, but at the price of greatly increased computation time[33].

In general, the relevant considerations in choosing a particular algorithm are hardware requirements, speed and sensitivity to biological relationships. The tensions between these competing claims are resolved variously by programs such as Fasta[28], BLAST[14] and Blaze[25]. The relative merits of these and the other programs have been discussed at length elsewhere[30,33]. The idea of optimizing a measure of local similarity is common to virtually all popular programs, and the results they produce therefore do not differ in any truly essential way.

Local alignment statistics

Not all biologically important sequence relationships will be detected by sequence similarity search programs and, even when found, they may be lost among irrelevant or chance similarities. While experiment is the ultimate arbiter of biological significance, mathematical analysis can indicate which similarities are unlikely to have arisen by chance and therefore merit special attention. Thus an important question concerning alignments produced by any database search is whether they can be considered statistically significant.

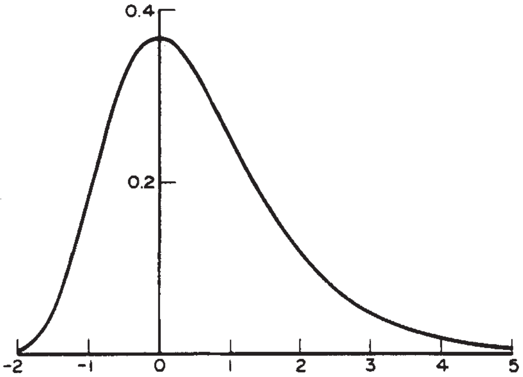


Fig. 1 The probability density function of the extreme value distribution with characteristic value u=0 and decay constant λ=1.

One approach sometimes taken is to record an optimal local alignment score for each database sequence and then to report these scores as standard deviations from the mean. There are several serious and frequently unrecognized pitfalls to this procedure. First, the optimal scores for the comparison of a query sequence to different database sequences can not be assumed to be drawn from the same distribution. The longer a given database sequence, the greater the score expected by chance. Also, variation in residue composition among sequences can yield different score distributions. Second, unless a rigorous optimization algorithm is employed, the true

Box 1 The extreme value distribution and local sequence similarities

Just as the *sum* of many independent random variables results naturally in a normal distribution, the *maximum* of many independent random variables yields an *extreme value* distribution[78]. (For rigour, this statement must be qualified in many ways, but we will omit the technicalities here.) Because the score of an optimal local alignment is, for practical purposes, the maximum of many essentially independent alignment scores, the extreme value distribution plays a central role in the statistics of local sequence alignments. This distribution may be described by two parameters, the *characteristic value, u*, and the *decay constant*, λ; the probability of observing a score greater than or equal to x purely by chance is given by the formula

$$1\text{-}\exp(-e^{-\lambda(x-u)})$$

The probability density of the standard extreme value distribution, with $u=0$ and $\lambda=1$, is shown in Fig. 1. For random sequences, the maximal segment pair scores used by the BLAST algorithms[4,14,31] can be shown to obey an extreme value distribution[1,6–8]. While analysis is not available for the scores of alignments with gaps, experiment[10–12] and analogy[6–8,46,79–81] strongly suggest that they too should obey this type of distribution.

In order to use the formula above, one needs to estimate the relevant parameters u and λ for a given sequence comparison. These will, in general, depend upon the composition and length of the sequences being compared, and upon the particular scoring system used. For alignments with gaps, the parameters may be estimated by random simulation[13], or by examining optimal local alignment scores from unrelated sequences[10,12]. For ungapped alignments, the parameters may be calculated directly[6–8]. In this case, the parameter u may be written as

$$u = \frac{\ln Kmn}{\lambda}$$

where m and n are the sequences' lengths and K and λ may be calculated from the substitution scores and sequence compositions[6–8].

We have described how to calculate the probability, p, that a given local-alignment score would arise from the comparison of two random sequences. This probability must be adjusted for the multiple comparisons performed in a database search (see text). The applicable Poisson distribution implies that the probability of observing at least one alignment with pairwise p-value p from a search of a database containing D sequences may be estimated as

$$P \approx 1-e^{-Dp}$$

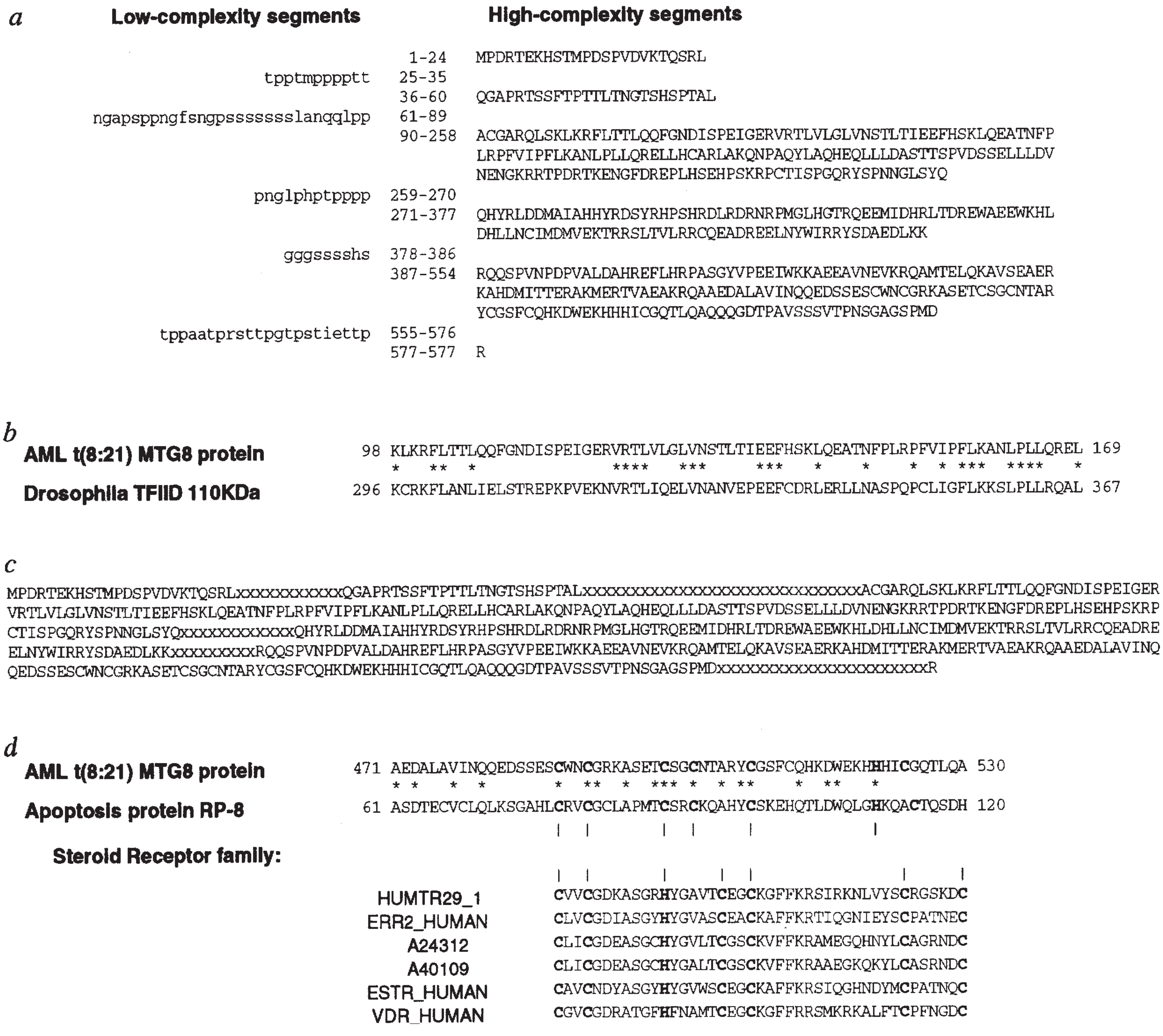
When $P<0.1$, it may be well approximated as simply Dp. This approach makes the implicit assumption that all sequences in the database are *a priori* equally likely to share some relationship with the query. An alternative view, based on the idea that many proteins possess multiple domains, is that all equal-length protein segments in the database are *a priori* equally likely to be related to the query. This approach implies a different normalization. Assume that the alignment of interest involves a database sequence of length n residues, and that the complete database has N residues. Then, in the equation above, D should be replaced by N/n. This is the default normalization currently employed by the BLAST programs. (In the context of DNA as opposed to protein database searches, it is the only normalization that really makes sense.) Reasons for calculating significance in the context of pairwise protein comparisons in the first place, rather than sequence-database comparisons, are to allow for multiple high-scoring alignments and for protein compositional heterogeneity.

The BLAST programs[4,14] (Table 1) may generate several high-scoring alignments for a given pair of sequences. While the significance of these alignments may be assessed individually, it is frequently of value to construct a combined assessment. One method uses the fact that the *number* of segment pairs expected by chance to have score at least x is approximately Poisson distributed, with parameter $e^{-\lambda(x-u)}$ (refs 6–8). Thus, if three distinct segment pairs with scores 50, 45 and 40 are found in a given pairwise comparison, one may calculate the probability p that at least three pairs, all with score at least 40, would appear by chance. This approach has the weakness of depending upon only the lowest among the r greatest scores. Alternatively, one may calculate the sum S_r of the r highest scores. The random distribution of such sums has been derived and the appropriate tail probability is available numerically as a double integral[9].

The BLAST programs currently use the former, Poisson method, of assessing multiple high-scoring segment pairs. Not all sets of segment pairs, however, warrant a joint assessment. Only when such a set may be combined into a consistent, gapped alignment is it really appropriate to consider the separate segment pairs as parts of a greater whole. Accordingly, as a default, the BLAST programs require such consistency before calculating a joint statistical assessment. The imposition of such consistency has the further advantage of sharpening the joint statistics[9].

The problem of multiple tests arises again in using either the Poisson or sum p-values described above. For example, while the probability for finding at least three segment pairs with score at least 40 may be valid, in practice one has considered as well the single best segment pair in isolation, the two best segment pairs, etc. These multiple tests can result in too optimistic a significance claim for the best overall result. P. Green (personal communication) has suggested a simple solution to this difficulty: dividing the p-value for a result involving r segment pairs by the factor $(1-a)a^{r-1}$, where a is a constant between 0 and 1, yields a conservative p-value for the multiple tests. The parameter a can be viewed as a "gap penalty." Choosing a near 0 greatly favours results involving a single segment pair. Choosing a near 1 favours results with fewer segment pairs only slightly, but may underestimate significance because of the actual non-independence of the multiple tests. The p-values reported by the BLAST programs implement this multiple test discount procedure, with a default of a=0.5. □

optimal pairwise scores will be systematically underestimated and the shape of the true distribution will be ill-determined. Third, comparing a query sequence to a set of uniform length random sequences yields scores that obey not a normal but an *extreme value* distribution (Box 1 and Fig. 1). The tail of this distribution decays exponentially in x rather than x^2, so assuming normality tends grossly to exaggerate an alignment's significance. Finally, a database search involves many essentially independent trials. If the database contains 50,000 sequences, a score with probability 10^{-6} of arising from a single comparison is only marginally significant in the context of the complete search. The last two points alone imply that an alignment may easily achieve a score over



a

```
Low-complexity segments                   High-complexity segments

                               1-24    MPDRTEKHSTMPDSPVDVKTQSRL
                tpptmppppptt  25-35
                              36-60    QGAPRTSSFTPTTLTNGTSHSPTAL
ngapsppngfsngpssssssslanqqlpp 61-89
                              90-258   ACGARQLSKLKRFLTTLQQFGNDISPEIGERVRTLVLGLVNSTLTIEEFHSKLQEATNFP
                                       LRPFVIPFLKANLPLLQRELLHCARLAKQNPAQYLAQHEQLLLDASTTSPVDSSELLLDV
                                       NENGKRRTPDRTKENGFDREPLHSEHPSKRPCTISPGQRYSPNNGLSYQ
                 pnglphptpppp 259-270
                              271-377  QHYRLDDMAIAHHYRDSYRHPSHRDLRDRNRPMGLHGTRQEEMIDHRLTDREWAEEWKHL
                                       DHLLNCIMDMVEKTRRSLTVLRRCQEADREELNYWIRRYSDAEDLKK
                    gggssssshs 378-386
                              387-554  RQQSPVNPDPVALDAHREFLHRPASGYVPEEIWKKAEEAVNEVKRQAMTELQKAVSEAER
                                       KAHDMITTERAKMERTVAEAKRQAAEDALAVINQQEDSSESCWNCGRKASETCSGCNTAR
                                       YCGSFCQHKDWEKHHHICGQTLQAQQQGDTPAVSSSVTPNSGAGSPMD
        tppaatprsttpgtpstiettp 555-576
                              577-577  R
```

b

```
AML t(8:21) MTG8 protein         98 KLKRFLTTLQQFGNDISPEIGERVRTLVLGLVNSTLTIEEFHSKLQEATNFPLRPFVIPFLKANLPLLQREL 169
                                    *   **  *              ****   ***     ***   *    *    *  * ***  ****   *
Drosophila TFIID 110KDa         296 KCRKFLANLIELSTREPKPVEKNVRTLIQELVNANVEPEEFCDRLERLLNASPQPCLIGFLKKSLPLLRQAL 367
```

c

```
MPDRTEKHSTMPDSPVDVKTQSRLxxxxxxxxxxxQGAPRTSSFTPTTLTNGTSHSPTALxxxxxxxxxxxxxxxxxxxxxxxxxxxxxACGARQLSKLKRFLTTLQQFGNDISPEIGER
VRTLVLGLVNSTLTIEEFHSKLQEATNFPLRPFVIPFLKANLPLLQRELLHCARLAKQNPAQYLAQHEQLLLDASTTSPVDSSELLLDVNENGKRRTPDRTKENGFDREPLHSEHPSKRP
CTISPGQRYSPNNGLSYQxxxxxxxxxxxxQHYRLDDMAIAHHYRDSYRHPSHRDLRDRNRPMGLHGTRQEEMIDHRLTDREWAEEWKHLDHLLNCIMDMVEKTRRSLTVLRRCQEADRE
ELNYWIRRYSDAEDLKKxxxxxxxxxRQQSPVNPDPVALDAHREFLHRPASGYVPEEIWKKAEEAVNEVKRQAMTELQKAVSEAERKAHDMITTERAKMERTVAEAKRQAAEDALAVINQ
QEDSSESCWNCGRKASETCSGCNTARYCGSFCQHKDWEKHHHICGQTLQAQQQGDTPAVSSSVTPNSGAGSPMDxxxxxxxxxxxxxxxxxxxxxxR
```

d

```
AML t(8:21) MTG8 protein    471 AEDALAVINQQEDSSESCWNCGRKASETCSGCNTARYCGSFCQHKDWEKHHHICGQTLQA 530
                                * *   *  *       *  **  *  *** *  * **    *  **   *
Apoptosis protein RP-8       61 ASDTECVCLQLKSGAHLCRVCGCLAPMTCSRCKQAHYCSKEHQTLDWQLGHKQACTQSDH 120
                                                 |  |       |  |     |            |

Steroid Receptor family:
                                                 |  |       |     |  |               |     |
                   HUMTR29_1                     CVVCGDKASGRHYGAVTCEGCKGFFKRSIRKNLVYSCRGSKDC
                  ERR2_HUMAN                     CLVCGDIASGYHYGVASCEACKAFFKRTIQGNIEYSCPATNEC
                      A24312                     CLICGDEASGCHYGVLTCGSCKVFFKRAMEGQHNYLCAGRNDC
                      A40109                     CLICGDEASGCHYGALTCGSCKVFFKRAAEGKQKYLCASRNDC
                  ESTR_HUMAN                     CAVCNDYASGYHYGVWSCEGCKAFFKRSIQGHNDYMCPATNQC
                  VDR_HUMAN                      CGVCGDRATGFHFNAMTCEGCKGFFRRSMKRKALFTCPFNGDC
```

ten standard deviations from the mean yet fail to be statistically significant.

Box 1 discusses the extreme value distribution and how it may be used to calculate the probability that a gap-free local alignment with a given score would arise from the comparison of two random sequences. It also describes how to modify this probability to account for the "multiple tests" of a database search. Such a search can itself generate data which provide an alternative to the analytic method (Box 1) for estimating alignment statistical significance[12]. For a given query, one records the best alignment score to each database sequence. If score S is observed $f(S)$ times, then plotting log $f(S)$ versus S tends to produce a straight line; extrapolation of this line can yield estimates of statistical significance[12].

One advantage of this approach is that it is applicable to cases for which no rigorous theory is available, such as scores from gapped alignments. Thus heuristic programs such as Fasta[28], or parallel implementations of the Smith-Waterman algorithm[18] such as Blaze[25] or Blitz[26], can estimate statistical significance using this method. Furthermore, because the scores generated derive from comparisons of real sequences, no "random protein" model is needed. A disadvantage of the method is the need to generate optimal alignment scores for a substantial fraction of database sequences in order to calculate statistical significance. Potential inaccuracy arises from variation in database sequence size and composition, which implies that each data point is really drawn from a separate distribution[6,10,13]. Also, if many sequences related to the query are present (see discussion on database redundancy below), it may be difficult to base the plotted line upon only unrelated sequences. An alternative "curve fitting" approach is to estimate the parameters of the implicit extreme value distribution for the scoring system at hand[2,10,11,13]. In one form or another, curve fitting will generally be necessary to calculate the statistical significance of scores derived from gapped alignments or other complex scoring systems[2,10–13].

The most important "failure" of the local alignment statistics discussed here is on comparisons of regions with restricted or unusual amino acid or nucleotide composition. Such regions are quite common in proteins, but are clearly not well described by the same random model used for other sequence regions (see below). Because an alignment of such "low complexity" regions has little real meaning, it is best simply to note their existence, but exclude them from alignments produced in database searches (see Figs 2 and 3 for examples).

Scoring matrices and gap costs

Many different amino acid substitution score matrices have been proposed over the years for use with sequence comparison and database search programs[1,3,34–43], and a variety of rationales have been used for their construction. However, it is possible to show that in the context of seeking high-scoring segment pairs without gaps, any such matrix has an implicit amino acid pair frequency distribution that characterizes the alignments it is optimized for finding. More precisely, let p_i be the frequency with which amino acid i occurs in proteins sequences and, within the class of alignments sought, let q_{ij} be the frequency with which amino acids i and j are aligned. Then the scores that best distinguish these alignments from chance are given by the formula

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

The base of the logarithm is arbitrary, affecting only the scale of the scores. Any set of scores useful for local alignment can be written in this form, so a choice of substitution matrix can be viewed as an implicit choice of "target frequencies" q_{ij} (refs 1,6).

The target frequencies characterizing alignments of closely related sequences clearly differ from those for alignments of sequences that are greatly diverged. Therefore a single matrix can not be optimal for recognizing relationships at all evolutionary distances[1,2,12]. It has been argued that for most practical purposes, three separate matrices should be adequate for locating all alignments containing sufficient information to rise above background noise[1,2]. The question remains how best to estimate the appropriate corresponding target frequencies.

Estimating the frequencies with which the various amino acids tend to mutate into one another is a necessarily empirical problem. The first approach to the question was taken by Dayhoff and coworkers[35,36]. Their "PAM" model of molecular evolution allowed target frequencies and the corresponding score matrices to be

◀Fig. 2 Significant sequence matches of the human MTG8 product: the effect of low-complexity masking. MTG8 (ref. 84) is the translated product of a chromosome 8 gene involved in a t(8:21) translocation that results in an AML1-MTG8 fusion transcript in a case of acute myeloid leukaemia (GenBank accession number D14820). *a*, Automated segmentation of low-complexity sequences in MTG8 at relatively high stringency. To be defined as low-complexity in this run of the SEG algorithm (Box 2), a sequence region must contain at least one 12-residue window with complexity (*K*, Box 2) less than 0.315. SEG then finds the minimally probable (lowest P_0, Box 2) low-complexity subsequence, of any length, within the overlapping windows of this region. The sequence segments read from left to right and their order in the polypeptide runs from top to bottom, as shown by the central column of residue numbers. *b*, The strong match, which emerges clearly without masking (Poisson p-value 2.5 × 10^{-6}), between sections of MTG8 and *Drosophila melanogaster* transcription factor TFIID 110-kDa subunit[85–86]. *c*, MTG8 filtered as in (*a*) but with the low-complexity segments masked by "x" characters, for use as a query sequence in database searches. *d*, The significant match between a region of MTG8 containing a cysteine cluster and rat apoptosis protein RP-8. RP-8 (ref. 87) is a gene expressed early in the process of programmed cell death (apoptosis) following glucocorticoid induction in rat thymocytes (GenBank accession number M80601). This match[84], had a Poisson p-value of 0.0036 for a BLASTP search of the NCBI non-redundant database of 13th September 1993. *, Identical amino acids; I, Conserved Cys or His residues. Also shown is a sample of the class of zinc-fingers that occur in the DNA binding domain of the steroid receptor family[88], indicating a suggestive similarity (which is not statistically significant by pairwise alignment statistics and would require experimental confirmation) in the positions of most of the Cys or His residues.

Before low-complexity filtering, MTG8 generated an output list from the NCBI non-redundant database of greater than 400 Kbytes containing 599 database sequences scoring above the BLASTP default threshold. The significant match to apoptosis protein was an inconspicuous 62nd in this list and scored much lower than many spurious low-complexity matches. After masking of MTG8 as in (*b*), this match was 6th in a list of 83 sequences. The latter list contained many matches to a "medium complexity" region of MTG8 which is tentatively predicted to be alpha helical coiled coil (residues 416–476). Further filtering with SEG at lower stringency (*K* < 0.365 for a 14-residue window) effectively masked this region, and resulted in a BLASTP output list of only 9 sequences, in which the apoptosis protein was ranked in score only below the MTG8 self-matches and the match to TFIID 110-kDa subunit.

Box 2 Low-complexity sequences and short-period tandem repeats

To study low-complexity sequences and short-period tandem repeats, we first consider sequences as mixtures of regions with unknown statistical properties and then attempt to infer these properties. In order to put all possible low-complexity segments on an equal footing, we define local compositional complexity ignoring prior probabilities for the 20 amino acids or 4 nucleotides[67,82,83]. Complexity is a function of the compositional *state* of a sequence segment or window. For example, the numbers (3,2,2,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0), representing, in decreasing order of abundance, counts for the various amino acids, describe one of the 77 possible complexity states of a 12-residue peptide window. Many possible sequences and amino acid compositions, with different residue types corresponding to the 20 numbers, share this complexity state. Formally, we define the local compositional complexity K of a sequence window of length L, as

$$K = \frac{1}{L}\log_{20}\left[\frac{L!}{\prod_{i=1}^{20} n_i!}\right]$$

where the n_i are the 20 numbers in the compositional state vector described above. Analogous to the enumeration of microstates in statistical mechanics, K measures the information per position needed, given the window's composition, to specify a particular residue order. Assuming uniform prior probabilities for the appearance of the various residues, the probability P_0 for the occurrence of a given compositional state is

$$P_0 = \frac{1}{20^L}\left[\frac{L!}{\prod_{i=1}^{20} n_i!}\right]\left[\frac{20!}{\prod_{k=0}^{L} r_k!}\right]$$

where r_k is the count of the number of times the number k occurs among the n_i. K and P_0 are functions of only the complexity state vector; they do not depend on which amino acids correspond to the 20 numbers in the vector or on the actual probabilities of the various amino acids. For the DNA alphabet, 4 replaces 20 in the above equations[82,83].

SEG[67] is an optimal segmentation algorithm based on the theory described above. It identifies, at a defined level of stringency, all the low-complexity segments in a sequence that minimize P_0 within a local region of low K. A similar approach may be used to identify tandemly repeated segments of any defined period; methods for the purpose are under development. A heuristic algorithm, XNU[74], for identifying and masking short-period repeats finds self-matching segments that yield high PAM or BLOSUM scores when offset by a small number of residues, regardless of local compositional complexity. With appropriate parameterization, XNU and SEG are complementary.

Programs such as SEG and XNU may be used to mask appropriate query sequence segments prior to database searching, replacing the residues in these segments by "x" characters (see Fig. 1c). The score for "x" in each row or column of a PAM or BLOSUM amino acid matrix may be calculated as the mean of the 20 residue–pair scores in that row or column, insuring that the impact of the masking character on the distribution of matching segment scores is minimal. □

calculated for any desired amount of evolutionary change. The details of the PAM model have been criticized[44], and the vast increase in available sequence data has prompted recalculation of the model's parameters[40,42]. Scores for DNA sequence comparison based on a PAM-like mutational model have also been described[3]. A different approach to estimating appropriate target frequencies relies not on fitting an evolutionary model, but rather on the direct observation of relatively distant, but nevertheless presumed largely correct, sequence alignments[41]. A variety of empirical tests have been claimed to support the superiority of the resulting "BLOSUM" matrices for detecting sequence homology[41,45]. Lacking an evolutionary model, however, this approach is less adaptable to generating matrices tailored to specific applications[3,5].

The theory linking substitution matrices with target frequencies is rigorously established only for local alignments lacking gaps. Therefore the development above is generally valid only for the BLAST and related algorithms[4,14,29]. A more general theory for alignments with gaps should, however, have the same broad outlines[10,46], and target frequency based substitution matrices are perhaps nearly optimal for this more general case. Gapped alignments present the additional problem of choosing appropriate gap costs[47]. The simplest algorithms require these costs to be a linear function of gap length[48–50], but efficient algorithms for more general gap costs are also available[51]. Because no theory exists, appropriate gap costs have generally been chosen by trial and error, although there have been some recent efforts to give this problem a sounder empirical footing[52,53].

The user of database search programs should recognize that the default substitution scores and, where applicable, gap costs, have generally been chosen to be appropriate for the most frequent sort of query. These scores may not, however, be optimal for a specific problem. In particular, matrices such as PAM-120 or BLOSUM-62 (the current BLASTP default)[41] are tailored for alignments of moderately diverged sequences. Very strong but short similarities, or very long but weak ones may easily be missed by these matrices[1–3]. A fully functional database search system should therefore provide a range of scoring systems to its users, so that the algorithm can be adapted to the problem at hand.

Databases and access

The most important requirement for database searching is a comprehensive, up-to-date database. Full releases of GenBank® now occur every two months, and daily updates are available for downloading or direct searching by e-mail and network services[54]. GenBank has undergone a major expansion in data coverage and now includes, in addition to nucleotide sequences, data from the major protein sequence and protein structure databases, as well as data from U.S. and European patents[54]. Approximately 36% of the records in GenBank are produced by the international collaborators, EMBL Data Library[55] and the DNA Database of Japan (DDBJ), with whom database updates are exchanged daily. Copies of the databases are available at many sites worldwide[54,55].

GenBank (release 80.0) contains 164 megabases of sequence and is doubling in size every 21 months (D. Benson, personal communication). This rate can only increase as a result of genome projects and automated sequencing technology. As mentioned above, special purpose computers have a role in maintaining reasonable search performance in the wake of this data deluge, but considerable improvements in search efficiency can be obtained by considering the nature of the data itself.

Many sequence databases have a large degree of internal "redundancy" for historical reasons related to available technology and research trends, and also due to the

Fig. 3

a

```
                                                                       Smallest
                                                                         Poisson
                                                                  High  Probability
Sequences producing High-scoring Segment Pairs:                  Score  P(N)       N

*pir|S21391|S21391 Hypothetical protein - Mouse | 0.0 0.0 ...     7006  0.0         1
*pir|S25716|S25716 Hypothetical protein 1 - Mouse | 0.0 0....     7006  0.0         1
*pir|S25714|S25714 Son of sevenless 2 - Mouse (fragment) |...     2866  0.0         1
*pir|A41216|A41216 Guanine nucleotide exchange activator -...      614  4.4e-167    6
*pir|PC1114|PC1114 SKCDC25 protein - Yeast (Saccharomyces ...      157  2.0e-28     4
 pir|A33647|A33647 Sulfated surface glycoprotein 185 - Vol...       95  2.7e-28     5
*pir|S28407|S28407 CDC25 protein homolog - mouse | 0.0 0.0...      191  1.3e-21     5
 pir|S06733|S06733 hydroxyproline-rich glycoprotein precur...      113  7.3e-21     5
 pir|S22697|S22697 extensin - Volvox carteri (fragment) | ...      151  1.4e-19     3
*pir|S22693|S22693 CDC25 protein homolog - mouse | 0.0 0.0...      191  1.7e-17     1
*pir|S20730|S20730 Hypothetical protein - Mouse | 0.0 0.0 ...      191  1.7e-17     1
*pir|A26596|RGBYC5 cell division control protein CDC25 - y...      145  2.5e-16     5
*pir|PS0040|PS0040 SCD25 protein - yeast (Saccharomyces ce...      131  3.7e-15     3
*pir|S12942|S12942 SDC25 protein - Yeast (Saccharomyces ce...      128  2.1e-14     3
 pir|S20790|S20790 Extensin - Almond | 0.0 0.0 0.0 0.0 0.0          81  7.4e-14     4
 pir|S13383|S13383 Hydroxyproline-rich glycoprotein - Sorg...       95  3.2e-12     2
 pir|S14449|S14449 Hydroxyproline-rich glycoprotein - Sorg...       95  3.2e-12     2
 pir|A43932|A43932 mucin - human (fragment) | 0.0 0.0 0.0 ...      101  4.8e-12     2
 pir|JQ1663|JQ1663 hybrid proline-rich protein - maize | 0...       74  1.1e-11     3
 pir|S25299|S25299 Extensin (clone Tom L-4) - Tomato | 0.0...       86  1.3e-11     3
 pir|PQ0479|PQ0479 extensin-like protein (clone pMG14) - c...       96  5.5e-11     3
 pir|S24620|S24620 Pistil extensin like protein - Common t...       96  5.5e-11     3
 pir|S08315|S08315 Cell wall protein - Maize | 0.0 0.0 0.0...       88  1.0e-10     2
*pir|S28415|S28415 Guanine nucleotide dissociation stimula...       92  1.3e-10     3
 pir|D29149|D29149 proline-rich protein - mouse (fragment)...       59  4.0e-10     5

Local alignments:

>pir|A33647|A33647 Sulfated surface glycoprotein 185 - Volvox carteri
          Length = 485

 Score = 95 (43.5 bits), Expect = 0.00038, P = 0.00038
 Identities = 28/88 (31%), Positives = 29/88 (32%)

Query:  1210 PPESPPLLPPREPVRTPDVFSSSPLHLQPPPLGKKSDHGNAFFPNSPSPFTPPPPQTPSP 1269
             PP   P    R P   P     SP    P P            P  PSP  PPPP  P P
Sbjct:   230 PPSPQPTASSRPPSPPPSPRPPSPPPPSPSPPPPPPPPPPPPPPPPPPSPPPPPPPPPPPP 289

Query:  1270 HGTRRHLPSPPLTQEMDLHSIAGPPVPP 1297
                    PSPP         +  PP PP
Sbjct:   290 PPPPPPSPSPPRKPPSPSPPVPPPPSPP 317

 Score = 80 (36.6 bits), Expect = 2.4e-15, Poisson P(3) = 2.4e-15
 Identities = 25/90 (27%), Positives = 31/90 (34%)

Query:  1151 PVPPPVPPRRRPESAPAESSPSKIMSKHLDSPPAIPPRQPTSKAYSPRYSISDRTSISDP 1210
             P   P+PP  +P ++    SP          PP+  P  P      P            P
Sbjct:   224 PNNSPLPPSPQPTASSRPPSPPPSPRPPSPPPPSPSPPPPPPPPPPPPPPPPPPSPPPPPP 283

Query:  1211 PESPPLLPPREPVRTPDVFSSSPLHLQPPP 1240
             P  PP  PP  P  +P     SP    PPP
Sbjct:   284 PPPPPPPPPPPPSPSPPRKPPSPSPPVPPP 313
```

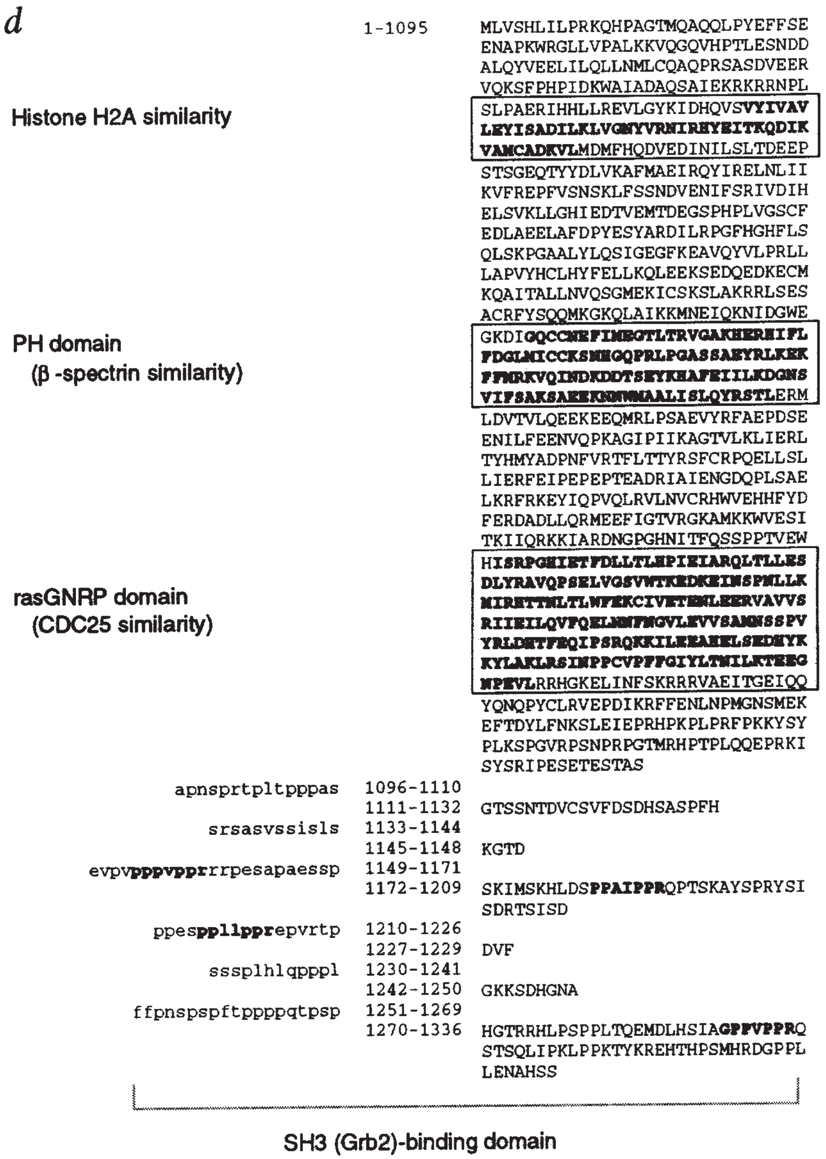
b

```
                                                                       Smallest
                                                                         Poisson
                                                                  High  Probability
Sequences producing High-scoring Segment Pairs:                  Score  P(N)       N

*pir|S21391|S21391 Hypothetical protein - Mouse | 0.0 0.0 ...     6368  0.0         1
*pir|S25716|S25716 Hypothetical protein 1 - Mouse | 0.0 0....     6368  0.0         1
*pir|S25714|S25714 Son of sevenless 2 - Mouse (fragment) |...     2866  0.0         1
*pir|A41216|A41216 Guanine nucleotide exchange activator -...      614  1.3e-167    6
*pir|PC1114|PC1114 SKCDC25 protein - Yeast (Saccharomyces ...      157  1.2e-28     4
*pir|S28407|S28407 CDC25 protein homolog - mouse | 0.0 0.0...      191  7.6e-22     5
*pir|S22693|S22693 CDC25 protein homolog - mouse | 0.0 0.0...      191  1.2e-17     4
*pir|S20730|S20730 Hypothetical protein - Mouse | 0.0 0.0 ...      191  1.2e-17     4
*pir|A26596|RGBYC5 cell division control protein CDC25 - y...      145  1.4e-16     5
*pir|PS0040|PS0040 SCD25 protein - yeast (Saccharomyces ce...      131  2.6e-15     3
*pir|S12942|S12942 SDC25 protein - Yeast (Saccharomyces ce...      128  1.5e-14     3
*pir|S28415|S28415 Guanine nucleotide dissociation stimula...       92  9.1e-11     3
*pir|S28098|S28098 ste6 protein - yeast (Schizosaccharomyc...      119  1.7e-07     1
 pir|A45564|A45564 histone 2A, H2A - Plasmodium falciparum...       92  0.00045     1
 pir|A44159|A44159 beta-spectrin general isoform, beta G-s...       92  0.0010      1
 pir|A25077|A25077 histone H2A.2 - sea urchin (Psammechinu...       89  0.0011      1
 pir|C25381|HSURA1 histone H2A-beta, sperm - sea urchin (S...       86  0.0015      1
 pir|S07392|S07392 Histone H2A.F/Z - Sea urchin (Strongylo...       88  0.0016      1
 pir|A02599|HSCH2F histone H2A.F, embryonic - chicken | 10...       88  0.0016      1
 pir|S03642|S03642 histone H2A.Z - bovine | 1085.0 0.0 0.0...       88  0.0017      1
 pir|S03644|S03644 histone H2A.Z - rat | 1085.0 0.0 0.0 0....       88  0.0017      1
 pir|S03282|S03282 histone H2A.F - chicken | 1085.0 0.0 0....       88  0.0017      1
 pir|A35881|A35881 histone H2A.Z - human | 1085.0 0.0 0.0 ...       88  0.0017      1
 pir|S03643|S03643 histone H2A.Z - human | 1085.0 0.0 0.0 ...       88  0.0017      1
 pir|JQ0794|JQ0794 histone H2A.III - Volvox carteri | 0.0 ...       88  0.0017      1

Local alignments:

>pir|S28098|S28098 ste6 protein - yeast (Schizosaccharomyces pombe)
          Length = 911

 Score = 119 (54.6 bits), Expect = 1.7e-07, P = 1.7e-07
 Identities = 25/82 (30%), Positives = 46/82 (56%)

Query:   845 TTNLTLWFEKCIVETENLEERVAVVSRIIEILQVFQELNNFNGVLEVVSAMNSSPVYRLD 904
             + +L  +  + IV+ E    R  V++  I++    +ELNNF  +  ++SA+NSSP++RL
Sbjct:   710 SNHLVNFVTETIVQEEEPRRRTNVLAYFIQVCDYLRELNNFASLFSIISALNSSPIHRLR 769

Query:   905 HTFEQIPSRQKKILEEAHELSE 926
              T+  + S+     E  + L+E
Sbjct:   770 KTWANLNSKTLASFELLNNLTE 791

 Score = 67 (30.7 bits), Expect = 0.00011, Poisson P(2) = 0.00011
 Identities = 12/36 (33%), Positives = 20/36 (55%)

Query:   928 HYKKYLAKLRSINPPCVPFFGIYLTNILKTEEGNPE 963
             ++  Y   L +   PCVPF G+Y T++   + GN +
Sbjct:   795 NFSNYRDCLENCVLPCVPFLGVYFTDLTFLKTGNKD 830
```

c

```
                                                                          Smallest
                                                                           Poisson
                                                                   High  Probability
Sequences producing High-scoring Segment Pairs:                   Score  P(N)       N

*pir|S21391|S21391 Hypothetical protein - Mouse | 0.0 0.0 ...  2957  0.0        1
*pir|S25716|S25716 Hypothetical protein 1 - Mouse | 0.0 0....  2957  0.0        1
*pir|S25714|S25714 Son of sevenless 2 - Mouse (fragment) |...   833  1.8e-108   1
*pir|A41216|A41216 Guanine nucleotide exchange activator -...   258  5.6e-53    2
 pir|A45564|A45564 histone 2A, H2A - Plasmodium falciparum...    92  0.00027    1
 pir|A44159|A44159 beta-spectrin general isoform, beta G-s...    92  0.00061    1
 pir|A25077|A25077 histone H2A.2 - sea urchin (Psammechinu...    89  0.00067    1
 pir|C25381|HSURA1 histone H2A-beta, sperm - sea urchin (S...    86  0.00085    1
 pir|S07392|S07392 Histone H2A.F/Z - Sea urchin (Strongylo...    88  0.00095    1
 pir|A02599|HSCH2F histone H2A.F, embryonic - chicken | 10...    88  0.00097    1
 pir|S03642|S03642 histone H2A.Z - bovine | 1085.0 0.0 0.0...    88  0.00098    1
 pir|S03644|S03644 histone H2A.Z - rat | 1085.0 0.0 0.0 0....    88  0.00098    1
 pir|S03282|S03282 histone H2A.F - chicken | 1085.0 0.0 0....    88  0.00098    1
 pir|A35881|A35881 histone H2A.Z - human | 1085.0 0.0 0.0 ...    88  0.00098    1
 pir|S03643|S03643 histone H2A.Z - human | 1085.0 0.0 0.0 ...    88  0.00098    1
 pir|JQ0794|JQ0794 histone H2A.III - Volvox carteri | 0.0 ...    88  0.00099    1
 pir|JQ0796|JQ0796 histone H2A.IV - Volvox carteri | 0.0 0...    88  0.00099    1
 pir|S08118|S08118 histone H2A.vD - fruit fly (Drosophila ...    88  0.0011     1
 pir|S04565|S04565 histone H2A (clone pCH3.5E) - chicken (...    86  0.0013     1
 pir|A38054|HSUR9M histone H2A, gonadal - sea urchin (Psam...    86  0.0018     1
 pir|A37574|HSUR9P histone H2A, gonadal - sea urchin (Pare...    86  0.0018     1
 pir|S18004|S18004 Histone H2A - Midge (Chironomus thummi ...    86  0.0018     1
 pir|A27790|HSKW2A histone H2A - Caenorhabditis elegans | ...    86  0.0019     1
 pir|A92959|HSTR21 histone H2A, gonadal - rainbow trout | ...    86  0.0019     1
 pir|S04238|S04238 histone H2A - Caenorhabditis elegans | ...    86  0.0019     1

Local alignments:

>pir|A45564|A45564 histone 2A, H2A - Plasmodium falciparum
            Length = 132

 Score = 92 (42.0 bits), Expect = 0.00027, P = 0.00027
 Identities = 20/45 (44%), Positives = 31/45 (68%)

Query:   145 VYIVAVLEYISADILKLVGNYVRNIRHYEITKQDIKVAMCADKVL 189
             VY+ AVLEY+ A+IL+L GN  R+ +    IT + I++A+   D+ L
Sbjct:    50 VYLAAVLEYLCAEILELAGNAARDNKKSRITPRHIQLAVRNDEEL 94


>pir|A44159|A44159 beta-spectrin general isoform, beta G-spectrin - human
            Length = 2364

 Score = 92 (42.0 bits), Expect = 0.00061, P = 0.00061
 Identities = 17/38 (44%), Positives = 23/38 (60%)

Query:   523 DTSEYKHAFEIILKDGNSVIFSAKSAEEKNNWMAALIS 560
             D  + KH F++ L DGN  +F AK  EE N W+ A+ S
Sbjct:  2267 DYKKKKHVFKLRLNDGNEYLFQAKDDEEMNTWIQAISS 2304
```

d



```
                                1-1095   MLVSHLILPRKQHPAGTMQAQQLPYEFFSE
                                         ENAPKWRGLLVPALKKVQGQVHPTLESNDD
                                         ALQYVEELILQLLNMLCQAQPRSASDVEER
                                         VQKSFPHPIDKWAIADAQSAIEKRKRRNPL
                                         SLPAERIHHLLREVLGYKIDHQVSVYIVAV
                                         LEYISADILKLVGNYVRNIRHYEITKQDIK
                                         VAMCADKVLMDMFHQDVEDINILSLTDEEP
                                         STSGEQTYYDLVKAFMAEIRQYIRELNLII
                                         KVFREPFVSNSKLFSSNDVENIFSRIVDIH
                                         ELSVKLLGHIEDTVEMTDEGSPHPLVGSCF
                                         EDLAEELAFDPYESYARDILRPGFHGHFLS
                                         QLSKPGAALYLQSIGEGFKEAVQYVLPRLL
                                         LAPVYHCLHYFELLKQLEEKSEDQEDKECM
                                         KQAITALLNVQSGMEKICSKSLAKRRLSES
                                         ACRFYSQQMKGKQLAIKKMNEIQKNIDGWE
                                         GKDIGQCCNEFIMEGTLTRVGAKHERHIFL
                                         FDGLMICCKSNHGQPRLPGASSAEYRLKEK
                                         FFMRKVQINDKDDTSEYKHAFEIILKDGNS
                                         VIFSAKSAEEKNNWMAALISLQYRSTLERM
                                         LDVTVLQEEKEEQMRLPSAEVYRFAEPDSE
                                         ENILFEENVQPKAGIPIIKAGTVLKLIERL
                                         TYHMYADPNFVRTFLTTYRSFCRPQELLSL
                                         LIERFEIPEPEPTEADRIAIENGDQPLSAE
                                         LKRFRKEYIQPVQLRVLNVCRHWVEHHFYD
                                         FERDADLLQRMEEFIGTVRGKAMKKWVESI
                                         TKIIQRKKIARDNGPGHNITFQSSPPTVEW
                                         HISRPGHIETFDLLTLHPIEIARQLTLLES
                                         DLYRAVQPSELVGSVWTKEDKEINSPNLLK
                                         MIRHTTNLTLWFEKCIVETENLEERVAVVS
                                         RIIEILQVFQELNNFNGVLEVVSAMNSSPV
                                         YRLDHTFEQIPSRQKKILEEAHELSEDHYK
                                         KYLAKLRSINPPCVPFFGIYLTNILKTEEG
                                         NPEVLRRHGKELINFSKRRRVAEITGEIQQ
                                         YQNQPYCLRVEPDIKRFFENLNPMGNSMEK
                                         EFTDYLFNKSLEIEPRHPKPLPRFPKKYSY
                                         PLKSPGVRPSNPRPGTMRHPTPLQQEPRKI
                                         SYSRIPESETESTAS
          apnsprtpltpppas   1096-1110
                            1111-1132   GTSSNTDVCSVFDSDHSASPFH
             srsasvssisls   1133-1144
                            1145-1148   KGTD
  evpvpppvppprrrpesapaessp   1149-1171
                            1172-1209   SKIMSKHLDSPPAIPPRQPTSKAYSPRYSI
                                         SDRTSISD
        ppespppllpprepvrtp   1210-1226
                            1227-1229   DVF
             sssplhlqpppl   1230-1241
                            1242-1250   GKKSDHGNA
      ffpnspspftppppqtpsp   1251-1269
                            1270-1336   HGTRRHLPSPPLTQEMDLHSIAGPPVPPRQ
                                         STSQLIPKLPPKTYKREHTHPSMHRDGPPL
                                         LENAHSS
```

existence of clusters of closely-related sequences from multigene families. Also, equivalent gene products have frequently been sequenced in a number of different species or organisms. In release 36.0 of PIR International[56], for example, there were 653 members of the globin superfamily, 349 cytochromes c, 583 sequences with immunoglobulin domains and 274 protein kinases. Considering only perfectly matching sequences, among the 52,257 protein sequences in this database, there are over 3,900 duplicate entries and over 3,800 perfect substrings of longer entries that together comprise about 10% of the total amino acid residues. Among nucleic acid sequences there are thousands of Alu variants in GenBank. And the problem of redundancy is only getting worse: as a result of projects designed to sample expressed genes rapidly[57–59], tens of thousands of sequence fragments are being added to the databases[60]; many of these sequences represent small pieces of known genes. Due to the error-prone nature of these sequence fragments[60], identifying redundancy in these collections is a more difficult task.

As well as decreasing the speed of database searches, redundancy can obscure novel matches in the output, by yielding slews of similar or identical alignments. Practically, there are two simple ways to avoid this problem: i) construct a smaller "nonredundant" database[61]; ii) preprocess the query sequence for the presence of known domains and mask these prior to searching. (The concept of query masking is discussed in the next section.)

NCBI[62] maintains two quasi-nonredundant sequence collections (NRDB), one for proteins and one for nucleic acids. For example, the protein NRDB is constructed iteratively starting with SWISS-PROT[63], which is the smallest and least redundant of the major protein databases. All of the proteins in PIR International[56] are compared to those in SWISS-PROT, and identical sequences are excluded from the former while maintaining pointers to relevant annotation. Next, all of the protein translations from GenBank coding sequences ("GenPept") are compared to the merged SWISS-PROT plus PIR. Likewise, protein sequences from the Brookhaven structure database (PDB) and other sources are incorporated into NRDB. (The OWL nonredundant sequence database[61] is constructed from the same sources.) This simple procedure reduces the size of the combined databases by 50%, yet ensures that all sequences are represented. More sophisticated methods for creating derived, composite views of protein and DNA sequence data promise even further reductions[54].

Another key issue is access to the databases. Researchers may perform database similarity searches remotely by sending their queries, via electronic mail, to centralized "server" computers, where large and frequently updated databases are maintained, and where fast processors and sophisticated software are available. E-mail services of this sort have been available from various sources for several years. For example, NCBI provides the BLAST e-mail server (for more information, send a "help" message to the Internet address blast@ncbi.nlm.nih.gov), and EMBL provides Blitz (nethelp@embl-heidelberg.de). Additional sites and services are given in ref. 64. In addition to database search and retrieval services, such sites maintain repositories of public domain software and specialized datasets that may be accessed via "anonymous ftp" over the Internet[65]. The existence of high-performance networks is also giving rise to a new generation of "client-server applications" that make possible direct, real-time user interactions with remote servers. NCBI's BLAST network service and *Entrez* retrieval system are two examples. For users of the many excellent commercial software packages for sequence analysis, we would anticipate the development of network client-server capabilities in the near future.

Masking of low-complexity sequences

Interspersed local regions of very simple amino acid composition are surprisingly abundant in protein sequences[67]. Some of these regions are homopolymers or short-period repeats, but most are not periodic and appear as mosaics of predominantly one or a few types of residue. Their compositional bias is in marked contrast to the structural domains and motifs of globular proteins familiar from crystal and NMR structures. Based on a relatively stringent definition of low-complexity[67], more than half of the sequences in the database contain at least one such region, and 14% of the amino acids occur in clusters of highly biased local composition. Moreover, a large excess of "medium-complexity" regions may be defined using a less stringent definition of complexity: these are found in many recently-deduced protein sequences that lack true homologues and do not belong to the class of "ancient conserved sequences"[68]. Very little is known about the molecular structures, dynamics, interactions and evolution of most low- and medium-complexity protein segments.

◀Fig. 3 The mouse protein Sos1 functions as a key intermediate in transmitting signals from receptor tyrosine kinases to ras via protein-protein interactions[89,90]. Sos1 (PIR accession S21391) is a member of a family of ras guanine nucleotide-releasing proteins (GNRP) that also includes *S. cerevisiae* CDC25 and SDC25, *S. pombe* Ste6, and the *Drosophila* gene, Son of sevenless[91]. Mouse Sos1 is a large, mosaic protein with several different domains, including a rasGNRP domain and a low complexity region that binds to an "adapter" protein called Grb2[92]. *a*, Results of a BLASTP search using an Sos1 query sequence without any masking applied. In addition to several "self hits" in the output, we see significant matches to some *S. cerevisiae* proteins, but Ste6 does not appear in the top 25 matches despite its presence in the database (PIR International, release 37). Moreover, the true positive matches are interspersed with many false positives, consisting of a number of functionally unrelated proline-rich proteins. These artifactual matches are highly significant in the statistical sense, but a glance at some of the local alignments shows that one is not justified in inferring similar function despite the high scores and low p-values. *b*, An identical search, except that in this case the Sos1 query has been pre-processed using SEG masking with default parameters. Note that the top of the "hit list" is now populated only by *bona fide* members of the rasGNRP family and that all artifactual matches against proline-rich proteins have disappeared. Furthermore, a match to *S. pombe* Ste6 is now obvious; a local alignment between this protein and Sos1 is shown. Interestingly, Sos1 shows significant local similarities to histone H2A and β-spectrin (see below). *c*, Results of another search with masking of both low complexity regions (*b*) and the rasGNRP domain. The top four matches now consist only of those proteins that share more extensive, or global, similarity with the query beyond the rasGNRP domain. In this example, the additional information gained by this extra masking step is not striking. But one can imagine the dramatic effect this would have in shrinking the "hit list" if the query possessed a kinase domain, of which there are hundreds of examples in the database. (See ref. 74 for an example involving immunoglobulin domains). *d*, The query sequence, mouse Sos1, annotated with the various domains indentifiable by BLASTP searching. The rasGNRP domain is according to Boguski & McCormick[91]. The proline-rich carboxy terminal region is known to interact with Src homology (SH3) domains in Grb2[92]. With regard to the local similarities between Sos1 and histone H2A and β-spectrin, it has recently been shown that Sos1, β-spectrin and a number of other proteins possess "pleckstrin homology" or PH domains[93]. The local alignment produced by BLASTP (*c*) corresponds to these PH domains. The similarity between Sos1 and histone H2A has not previously been reported and is difficult to interpret biologically. Nonetheless, the similarity is as significant as that of the PH domain and may have structural, as opposed to functional, implications[94].

Low-complexity segments confound database search algorithms in two ways. First, most of these segments do not generally give meaningful alignments position by position in ways that reflect actual structure and mutational history: they evidently evolve relatively rapidly by processes such as replication slippage and repeat expansion[67]. (At the DNA sequence level, trinucleotide and dinucleotide repeat polymorphisms provide a familiar example[69,70].) Permutations, shuffles or reversals of low-complexity amino acid sequences generally give alignment scores similar to the original sequence. Second, the residue compositions of low-complexity segments are very different from that of the database as a whole. This is evident if all low-complexity segments in the database are grouped into a single class: a strong excess of alanine, glycine, proline, serine, glutamate and glutamine results. However, this lumped class is itself heterogeneous, containing for example glutamine-rich and proline-rich subclasses. These statistical biases contrast with those that characterize the bulk of most query and database sequences, and on which score-based alignment statistics are founded. Thus the high scores of alignments of low-complexity segments are due primarily to their compositional biases and do not necessarily reflect significant positional similarity.

Several classes of low-complexity residue clusters have been analysed for statistical significance by Karlin and coworkers[71–73]. Their methods, which use the contrasting residue frequencies of specific clusters and those of complete proteins or databases, are embodied in the SAPS software[73]. SEG[67], the algorithm employed by the BLAST programs for filtering low-complexity segments from query sequences prior to database searching (Figs 2 and 3), employs instead optimal segmentation methods applied to a more general definition of compositional complexity (see Box 2).

Masking of highly abundant sequences

Database searching can be performed efficiently in phases, with a query first compared to a small database containing domains representative of large sequence families. Subsequences of a query that match one or more of these domains can then be masked prior to full-scale searching, thereby eliminating most of the redundant output[74]. Annotated collections of prototypic human repetitive sequences[75], such as Alu and protein kinase catalytic domains[76], exist and can be used to pre-filter a query (Fig 3*c*). (Both of these data sets are available from the NCBI Data Repository on CD-ROM and by anonymous ftp. See /repbase/alu, /repbase/humrep and /pkinases/pkcdd.fa at ncbi.nlm.nih.gov.) For proteins, a more comprehensive solution to the problem is approached by building a small, representative set of protein superfamilies or motifs and using this as a screening database with automatic masking of matching query subsequences (unpublished results). This technology is still under development but recent studies indicate that a representative set of only 1,000–3,000 sequences may suffice[68]; such a database can be searched in seconds. The first large-scale implementation of this strategy has been performed for a specialized database of "expressed sequence tags" or ESTs[60] where such pre-filtering is also employed to detect contamination by vector sequences.

Conclusions

The stated goals of the U.S. Genome Project include the production of 50 megabases of DNA sequence data per year by 1998 and the identification and correlation of genes in humans and model organisms[77]. Database similarity searching will be one of the major informatics tools used in this endeavor. Not only efficient algorithms, but also a choice of appropriate scoring systems, well-defined measures of statistical significance and a better understanding of the sequences themselves, are critical for the automated analysis schemes that this amount of data will inevitably require.

Special purpose and faster general purpose computers will have roles in sifting through this increasing volume of sequence data. But large improvements in the efficiency of searching can be obtained by considering the nature of the data and implementing new strategies that capitalize on this knowledge. One of these strategies is to preprocess a query sequence to identify known domains and motifs, dispersed repeats, low complexity segments and other regions of compositional bias such as potential membrane-spanning and α-helical coiled-coil regions. We have described several preprocessing techniques that are suitable for automation and have demonstrated their practical utility with examples. Foreknowledge of query features enables one to perform faster and more effective searches better and to evaluate search results.

Another, complementary strategy is to reduce the redundancy in the target database(s) to be searched. We have outlined one simple but useful approach to the reductive merging of diverse, but overlapping, source databases. But newer, cleaner and richer views of the sequence data, optimized for gene discovery, are on the horizon.

Note added in proof: NCBI has recently established a GenBank® World Wide Web server (the URL is http://www.ncbi.nlm.nih.gov) that provides network access to many of the software tools and data sources described in this review.

Acknowledgements

GenBank is a registered trademark of the U.S. Department of Health and Human Services.

1. Altschul, S.F. Amino acid substitution matrices from an information theoretic perspective. *J. molec. Biol.* **219**, 555–565 (1991).
2. Altschul, S.F. A protein alignment scoring system sensitive at all evolutionary distances. *J. molec. Evol.* **36**, 290–300 (1993).
3. States, D.J., Gish, W. & Altschul, S.F. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods* **3**, 66–70 (1991).
4. Gish, W. & States, D.J. Identification of protein coding regions by database similarity search. *Nature Genet*. **3**, 266–272 (1993).
5. Claverie, J.-M. Detecting frameshifts by amino acid sequence comparison. *J. molec. Biol*. **234**, 1140–1157 (1993).
6. Karlin, S. & Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. natn. Acad. Sci. U.S.A.* **87**, 2264–2268 (1990).
7. Karlin, S., Dembo, A. & Kawabata, T. Statistical composition of high-scoring segments from molecular sequences. *Ann. Stat.* **18**, 571–581 (1990).
8. Dembo, A. & Karlin, S. Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Ann. Prob*. **19**, 1737–1755 (1991).
9. Karlin, S. & Altschul, S.F. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. natn. Acad. Sci. U.S.A.* **90**, 5873–5877 (1993).
10. Smith, T.F., Waterman, M.S. & Burks, C. The statistical distribution of nucleic acid similarities. *Nucl. Acids Res*. **13**, 645–656 (1985).
11. Altschul, S.F. & Erickson, B.W. A nonlinear measure of subalignment similarity and its significance levels. *Bull. math. Biol.* **48**, 617–632 (1986).
12. Collins, J.F., Coulson, A.F.W. & Lyall, A. The significance of protein sequence similarities. *CABIOS* **4**, 67–71 (1988).

13. Mott, R. Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. math. Biol.* **54**, 59–75 (1992).
14. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. molec. Biol.* **215**, 403–410 (1990).
15. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. molec. Biol.* **48**, 443–453 (1970).
16. Sellers, P.H. On the theory and computation of evolutionary distances. *SIAM J. appl. Math.* **26**, 787–793 (1974).
17. Sankoff, D. & Kruskal, J.B. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* (Addison-Wesley, Reading, MA, 1983).
18. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J. molec. Biol.* **147,** 195–197 (1981).
19. Goad, W.B. & Kanehisa, M.I. Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucl. Acids Res.* **10**, 247–263 (1982).
20. Sellers, P.H. Pattern recognition in genetic sequences by mismatch density. *Bull. math. Biol.* **46**, 501–514 (1984).
21. Waterman, M.S. & Eggert, M. A new algorithm for best subsequence alignments with applications to tRNA-rRNA comparisons. *J. molec. Biol.* **197**, 723–728 (1987).
22. Coulson, A.F.W., Collins, J.F. & Lyall, A. Protein and nucleic acid database searching: a suitable case for parallel processing. *Comp. J.* **30**, 420–424 (1987).
23. Chow, E.T., Hunkapiller, T., Peterson, J.C., Zimmerman, B.A. & Waterman, M.S. in *Proc. 1991 Int. Conf. on Supercomputing*, 216–223 (ACM Press, New York, 1991).
24. Jones, R. Sequence pattern matching on a massively parallel computer.*CABIOS* **8**, 377–383 (1992).
25. Brutlag, D.L. *et al.* BLAZE: an implementation of the Smith-Waterman sequence comparison algorithm on a massively parallel computer. *Comput. Chem.* **17**, 203–207 (1993).
26. Sturrock, S.S. & Collins, J.F. MPsrch version 1.3. (Biocomputing Research Unit, University of Edinburgh, 1993).
27. Lipman, D.J. & Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441 (1985).
28. Pearson, W.R. & Lipman, D.J. Improved tools for biological sequence comparison. *Proc. natn. Acad. Sci. U.S.A.* **85**, 2444–2448 (1988).
29. White, C.T. *et al.* in *Proc. 1991 IEEE Int. Conf. Comp. Design: VLSI in Computers and Processors*, 504–509 (IEEE Comp. Soc. Press, Los Alamitos, CA, 1991).
30. Pearson, W.R. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**, 635–650 (1991).
31. Altschul, S.F. & Lipman, D.J. Protein database searches for multiple alignments. *Proc. natn. Acad. Sci. U.S.A.* **87**, 5509–5513 (1990).
32. Argos, P. A sensitive procedure to compare amino acid sequences. *J. molec. Biol.* **193**, 385–396 (1987).
33. Vogt, G. & Argos, P. Searching for distantly related protein sequences in large databases by parallel processing on a transputer machine. *CABIOS* **8**, 49–55 (1992).
34. McLachlan, A.D. Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c_{551}. *J. molec. Biol.* **61**, 409-424 (1971).
35. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. in *Atlas of Protein Sequence and Structure* vol. 5, suppl. 3 (ed. M.O. Dayhoff) 345–352 (Natn. Biomed. Res. Found., Washington, 1978).
36. Schwartz, R.M. & Dayhoff, M.O. in *Atlas of Protein Sequence and Structure* vol. 5, suppl. 3 (ed. M.O. Dayhoff) 353–358 (Natn. Biomed. Res. Found., Washington, 1978).
37. Feng, D.F., Johnson, M.S. & Doolittle, R.F. Aligning amino acid sequences: comparison of commonly used methods. *J. molec. Evol.* **21,** 112–125 (1985).
38. Rao, J.K.M. New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. *Int. J. peptide protein Res.* **29,** 276–281 (1987).
39. Risler, J.L., Delorme, M.O., Delacroix, H. & Henaut, A. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J. molec. Biol.* **204**, 1019–1029 (1988).
40. Gonnet, G.H., Cohen, M.A. & Benner, S.A. Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445 (1992).
41. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. natn. Acad. Sci. U.S.A.* **89**, 10915–10919 (1992).
42. Jones, D.T., Taylor, W.R. & Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *CABIOS* **8**, 275–282 (1992).
43. Overington, J., Donnelly, D., Johnson, M.S., Sali, A. & Blundell, T.L. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Prot. Sci.* **1**, 216–226 (1992).
44. Wilbur, W.J. On the PAM matrix model of protein evolution. *Molec. Biol. Evol.* **2**, 434–447 (1985).
45. Henikoff, S. & Henikoff, J.G. Performance evaluation of amino acid substitution matrices. *Proteins* **17**, 49–61 (1993).
46. Waterman, M.S., Gordon, L. & Arratia, R. Phase transitions in sequence matches and nucleic acid structure. *Proc. natn. Acad. Sci. U.S.A.* **84**, 1239–1243 (1987).
47. Fitch, W.M. & Smith, T.F. Optimal sequence alignments. *Proc. natn. Acad. Sci. U.S.A.* **80**, 1382–1386 (1983).
48. Gotoh, O. An improved algorithm for matching biological sequences. *J. molec. Biol.* **162**, 705–708 (1982).
49. Altschul, S.F. & Erickson, B.W. Optimal sequence alignment using affine gap costs. *Bull. math. Biol.* **48**, 603–616 (1986).
50. Myers, E.W. & Miller, W. Optimal alignments in linear space. *CABIOS* **4**, 11–17 (1988).
51. Miller, W. & Myers, E.W. Sequence comparison with concave weighting functions. *Bull. math. Biol.* **50**, 97–120 (1988).
52. Pascarella, S. & Argos, P. Analysis of insertions/deletions in protein structures. *J. molec. Biol.* **224**, 461–471 (1992).
53. Benner, S.A., Cohen, M.A. & Gonnet, G.H. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. molec. Biol.* **229**, 1065–1082 (1993).
54. Benson, D., Lipman, D.J. & Ostell, J. GenBank. *Nucl. Acids Res.* **21**, 2963–2965 (1993).
55. Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J. & Cameron, G.N. The EMBL data library. *Nucl. Acids Res.* **21**, 2967–2971 (1993).
56. Barker, W.C., George, D.G., Mewes, H.-W., Pfeiffer, F. & Tsugita, A. The PIR-International databases. *Nucl. Acids Res.* **21**, 3089–3092 (1993).
57. Adams, M.D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–1656 (1991).
58. Sikela, J.M. & Auffray, C. Finding new genes faster than ever. *Nature Genet.* **3**, 189–191 (1993).
59. Davies, K. The EST express gathers steam. *Nature* **364**, 554 (1993).
60. Boguski, M.S., Lowe, T.M.J. & Tolstoshev, C.M. dbEST — database for "expressed sequence tags". *Nature Genet.* **4**, 332–333 (1993).
61. Bleasby, A.J. & Wootton, J.C. Construction of validated, non-redundant composite sequence databases. *Protein Eng.* **3**, 153–159 (1990).
62. Benson, D., Boguski, M., Lipman, D.J. & Ostell, J. The national center for biotechnology information. *Genomics* **6**, 389–391 (1990).
63. Bairoch, A. & Boeckmann, B. The SWISS-PROT protein sequence data bank, recent developments. *Nucl. Acids Res.* **21**, 3093–3096 (1993).
64. Henikoff, S. Sequence analysis by electronic mail server. *Trends biochem. Sci.* **18**, 267–268 (1993).
65. Krol, E. *The Whole Internet User's Guide & Catalog.* (O'Reilly & Assoc., Inc., Sebastopol, CA, 1992).
66. Network Entrez. *NCBI News* **2**(2), 1 (National Library of Medicine, Bethesda, MD, 1993).
67. Wootton, J.C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149–163 (1993).
68. Green, P., Lipman, D., Hillier, L., Waterston, R., States, D.J. & Claverie, J.-M. Ancient conserved regions in new gene sequences. *Science* **259**, 1711–1716 (1993).
69. Riggins, G.J. *et al.* Human genes containing polymorphic trinucleotide repeats. *Nature Genet.* **2,** 186–191 (1992).
70. Harding R.M., Boyce A.J. & Clegg, J.B. The evolution of tandemly repetitive DNA: recombination rules. *Genetics* **132**, 847–859 (1992).
71. Karlin, S. & Brendel, V. Charge configurations in viral proteins. *Proc. natn. Acad. Sci. U.S.A.* **85**, 9396–9400 (1988).
72. Karlin, S. & Brendel, V. Charge and statistical significance in protein and DNA sequence analysis. *Science* **257**, 39–49 (1992).
73. Brendel, V., Bucher, P., Nourbakhsh, I.R., Blaisdell, B.E. & Karlin, S. Methods and algorithms for statistical analysis of protein sequences. *Proc. natn. Acad. Sci. U.S.A.* **89**, 2002–2006 (1992).
74. Claverie, J.-M. & States, D.J. Information enchancement methods for large scale sequence analysis. *Comput. Chem.* **17**, 191–201 (1993).
75. Jurka, J., Walichiewicz, J. & Milosavljevic, A. Prototypic sequences for human repetitive DNA. *J. molec. Evol.* **35**, 286–291 (1992).
76. Hanks, S.K. & Quinn, A.M. Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Meth. Enzymol.* **200**, 38–62 (1991).
77. Collins, F. & Galas, D. A new five-year plan for the U.S. human genome project. *Science* **262**, 43–46 (1993).
78. Gumbel, E.J. *Statistics of extremes.* (Columbia Univ. Press, New York, 1958).
79. Arratia, R., Gordon, L. & Waterman, M.S. An extreme value theory for sequence matching. *Ann. Stat.* **14**, 971–993 (1986).
80. Arratia, R., Morris, P. & Waterman, M.S. Stochastic scrabble: large deviations for sequences with scores. *J. appl. Prob.* **25**, 106–119 (1988).
81. Arratia, R. & Waterman, M.S. The Erdos-Renyi strong law for pattern matching with a given proportion of mismatches. *Ann. Prob.* **17**, 1152–1169 (1989).
82. Salamon, P. & Konopka, A.K. A maximum entropy principle for distribution of local complexity in naturally occurring nucleotide sequences. *Comput. Chem.* **16**, 117–124 (1992).
83. Salamon, P., Wootton, J.C., Konopka, A.K. & Hansen, L. On the robustness of maximum entropy relationships for complexity distributions of nucleotide sequences. *Comput. Chem.* **17**, 135–148 (1993).
84. Miyoshi, H. *et al.* The t(8:21) translocation in acute myeloid leukemia results in production of an AML1-MTG8 fusion transcript. *EMBO J.* **12**, 2715–2721 (1993).
85. Kokubo, T., Gong, D-W., Roeder, R.G., Horikoshi, M. & Nakatani, Y. The Drosophila 110-kDa TFIID subunit directly interacts with the N-terminal region of the 230-kDa subunit. *Proc. natn. Acad. Sci. U.S.A.* **90**, 5896–5900 (1993).
86. Hoey, T. *et al.* Molecular cloning and functional analysis of Drosophila TAF110 reveal properties expected of coactivators. *Cell* **72**, 247–260 (1993).
87. Owens, G.P., Hahn, W.E. & Cohen, J.J. Identification of mRNAs associated with programmed cell death in immature thymocytes. *Mol. cell. Biol.* **11**, 4177–4188 (1991).
88. Schwabe, J.W., Neuhaus, D. & Rhodes, D. Solution structure of the DNA-binding domain of the oestrogen receptor. *Nature* **348**, 458–461 (1990).
89. Feig, L.A. The many roads that lead to Ras. *Science* **260**, 767–768 (1993).
90. McCormick, F. How receptors turn Ras on. *Nature* **363**, 15–16 (1993).
91. Boguski, M.S. & McCormick, F. Proteins regulating Ras and its relatives. *Nature* **366**, 643–654 (1993).
92. Rozakis-Adcock, M., Fernley, R., Wade, J., Pawson, T. & Bowtell, D. The SH2 and SH3 domains of mammalian Grb2 couple the EGF receptor to the Ras activator mSos1. *Nature* **363**, 83–85 (1993).
93. Musacchio, A., Gibson, T., Rice, P., Thompson, J. & Saraste, M. The PH domain is a common piece in the structural patchwork of signalling (and other) proteins. *Trends biochem. Sci.* **18**, 343–348 (1993).
94. Arents, G., Burlingame, R.W., Wang, B.C., Love, W.E. & Moudrianakis E.N. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. *Proc. natn. Acad. Sci. U.S.A.* **88**, 10148–10152 (1991).