Motifer, a search tool for finding amino acid sequence patterns from nucleotide sequence databases

Henrik Jörnvall*

Laboratory of Molecular Neurobiology, Department of Neuroscience, Karolinska Institutet, S-171 77 Stockholm, Sweden

Received 17 June 1999

Abstract Motifer is a software tool able to find directly in nucleotide databases very distant homologues to an amino acid query sequence. It focuses searches on a specific amino acid pattern, scoring the matching and intervening residues as specified by the user. The program has been developed for searching databases of expressed sequence tags (ESTs), but it is also well suited to search genomic sequences. The query sequence can be a variable pattern with alternative amino acids or gaps and the sequences searched can contain introns or sequencing errors with accompanying frame shifts. Other features include options to generate a searchable output, set the maximal sequencing error frequency, limit searches to given species, or exclude already known matches. Motifer can find sequence homologues that other search algorithms would deem unrelated or would not find because of sequencing errors or a too large number of other homologues. The ability of Motifer to find relatives to a given sequence is exemplified by searches for members of the transforming growth factor-β family and for proteins containing a WW-domain. The functions aimed at enhancing EST searches are illustrated by the 'in silico' cloning of a novel cytochrome P450 enzyme.

Key words: Database search; Nucleotide database; Software; Amino acid sequence motif; Motifer

1. Introduction

Bioinformatics is an essential tool to evaluate the vast amount of sequence information in databases. The rapid increase in known sequences is illustrated by the EMBL database. Its first release in 1982 contained 568 sequence entries with a total of 585 433 nucleotides. Today, this database has reached its 58th release with 3 272 064 sequences and 2 355 200 790 nucleotides. During the last year, the database has grown with an average of more than 3000 sequences each day (http://www.ebi.ac.uk). The human genome, together with more than 50 other genome projects, will soon yield more information than the current total of sequences so far collected. Similarly, expressed sequence tags (ESTs) constitute comprehensive data sets for the genomes of many organisms. However, since ESTs are derived from mRNA and many proteins are expressed at low levels or in a defined temporal and spatial pattern, they offer special problems. A priority when constructing EST databases is a high throughput, with small possibilities for annotation or quality control. Moreover, the purified mRNA is often truncated and sequencing is limited to single analyses from each end. The resulting EST is therefore often not complete and with frequent mutations and frame shifts.

Various tools have been designed to search sequence databases. The most frequently used programs are FASTA [1,2] and basic local alignment search tool (BLAST) [3]. They are powerful and can usually find homologues to a given sequence. However, in searches for very distant homologies, where only a non-contiguous amino acid pattern is conserved, or in searches of the EST databases, these programs are not always capable of matching relevant sequences. BLAST searches for sequences with a high local similarity to the query sequence, thus missing sequences with only distantly spaced amino acids conserved. Gapped BLAST (BLAST 2.0), position-specific iterated BLAST (PSI-BLAST) and pattern-hit initiated BLAST (PHI-BLAST) [4,5] are later versions that address these issues. However, the underlying algorithm is essentially the same and they, therefore, still rely on local similarities when searching databases. Gapped BLAST can introduce gaps between separate BLAST hits within a given sequence. PSI-BLAST iterates searches and integrates information from sequences earlier found to a position-specific score matrix that is used as query for subsequent searches. PHI-BLAST takes an amino acid pattern and a protein sequence as input, finds all instances of the amino acid pattern in a protein database and then scores the sequences found, based on similarity to the input protein sequence. Both PSI- and PHI-BLAST are limited to searches of protein databases and do not address the specific issues when searching nucleotide databases. Furthermore, all programs searching EST databases must handle the frequent frame shifts from sequencing errors. Genomic sequences should display a higher level of integrity but instead, they contain introns.

With these problems in mind, Motifer was constructed with the aim of finding sequence relationships that can be searched for in a variable manner. In the case of closely related structures, Motifer is an alternative to FASTA and BLAST, but the true power of the present program lies in its ability to search ESTs or genomic databases for distant homologues.

2. Materials and methods

2.1. General concept

The sequences to be searched are supplied to Motifer in EMBL or FASTA format in one or multiple files. The nucleotide databases are first translated into all six reading frames, whereby each sequence generates three forward translations and three translations of the reverse complement. The subsequent search of the translated sequences is dynamic and the algorithm can scan all reading frames in a given direction simultaneously as allowed by the search parameters. The core of the search algorithm is a recursive function that compares the query pattern in all possible combinations with the translated nucleotide sequences. Because of the continuous scanning of all read-

*Corresponding author. Fax: (46) (8) 337 462.
E-mail: henjor@cajal.mbb.ki.se

PII: S0014-5793(99)00941-2

ing frames and the inherent variability in the query pattern, the algorithm gives a slow search. However, it can scan the EMBL EST databases with a query of 15 positions overnight on a standard desktop PC, but the exact search time is dependent on the sequences searched, the complexity of the search pattern and allowances in amino acid spacing. Motifer is available in both Windows NT and Linux versions. Non-commercial users may obtain a copy by E-mail request.

2.2. Search parameters

A query amino acid pattern and one or more nucleotide sequences to search are the two requirements when using Motifer. The syntax of the query permits flexibility in the design of the amino acid pattern with which to search the nucleotide sequence(s). The amino acids defining the pattern and the distances spacing them form the base of the query, but alternative amino acids, variable spacing intervals and allowance for frame shifts can be specified at each position of the query pattern. The construction of a query and its syntax is illustrated in Fig. 1. Note that not only the residues defining the pattern are evaluated during the search, but intervening ones as well. Thus, the query C(3)C denotes two cysteines spaced by three residues that by default cannot be cysteine or stop, while C(3,C)C allows for a cysteine between the two outer cysteines. The query pattern can be degenerate and FY(4–6)V thus denotes a phenylalanine or tyrosine and a valine spaced by 4–6 intervening residues.

2.3. Scoring system

All hits are scored according to the probability that they would occur by chance, based on the codon frequencies of the non-coding reading frames as investigated on the vertebrate sequences in the EMBL database. The observed codon usage frequencies in the coding reading frames correspond well to those previously reported [6]. To exemplify how the scoring calculations are performed, consider Cys with a frequency of 2.25% in the coding reading frames of vertebrate EMBL sequences. In contrast, the codons for Cys (TGT and TGC) were found to occur in 3.94% of the five non-coding reading frames of the same database. Similarly, 0.18% of the codons in coding reading frames code for a stop, versus 4.06% in non-coding reading frames. Therefore, the pattern Cys-Xxx-Cys has a calculated probability of 1.43×10^{-3} to occur by chance in a non-coding reading frame (i.e. the first and third Cys have a probability of 3.94×10^{-2} and the probability for Xxx being anything but Cys or stop calculates to 9.20×10^{-1}). The score for a given hit therefore reflects the probability for the pattern to occur in a non-coding reading frame and thus to be a false hit.

2.4. Output refinements

Upon finishing a search, Motifer combines the hits found within the sequence to a longer, continuous hit that still fulfills the constraints set by the query pattern and other search parameters. By this combination, an amino acid that is erroneously assumed to be conserved and thus included in the search pattern does not ruin an otherwise perfect pattern. Furthermore, this function allows for differences in the distances separating the amino acids forming the search pattern. The user can set several parameters that govern how, or if, these combinations are made and can thereby optimise the search to any discrepancies expected between a specific query and its homologues at the nucleotide level in the database searched. For example, the limits within which the query can vary can be specified, as can the allowance to introduce frame shifts when combining two hits into one.

Several features are aimed at enhancing the results of searches in EST databases and genomic sequences. Frame shifts resulting from sequencing errors can either be handled by allowing for them at specific positions in the query sequence or by virtue of the combination of two separate hits as described above. By allowing for frame shifts in the query sequence, the possibility for false hits is greatly increased. This can be prevented in Motifer by setting the maximum allowed sequencing error and thereby limiting the number of jumps between different reading frames.

A too large number of known species variants is a problem when searching databases for new members of a protein family. These variant sequences can, by their sheer number, hide new sequences in the search output. To circumvent this, Motifer can take a list of accession numbers of already known or otherwise uninteresting sequences and exclude such sequences from the output. Furthermore, if the database to be searched is in EMBL format, the user can limit the search to specific species and thus further enhance the output. The output itself can also be searched with a completely new query and the result of an initial broad search can then be reduced in an attempt to minimise the number of false hits.

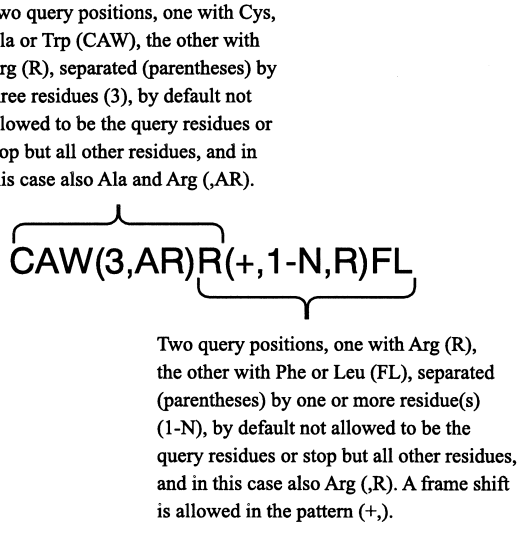

Fig. 1. Query syntax of Motifer.

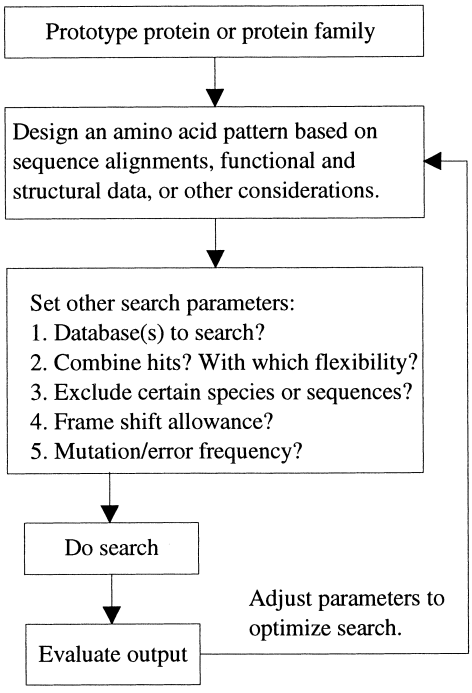

Fig. 2. General approach to a search using Motifer.

3. Results and discussion

3.1. General approach to a search

Motifer can find nucleotide sequences corresponding to an amino acid pattern given as query. Therefore, much attention must be focused on obtaining a valid query that contains as many relevant and for the protein function critical residues as possible. One strategy to find those pertinent residues is to align all known proteins of the family under investigation. However, when few or no other family members are known, functional and structural data must dictate the query pattern.

A

```
TGF-β1               ALDTNYCFSSTEKNCCVR.QLYIDFRKDLGWK.WIHEPKGYHANFCLGPCPY.....IWSLDTQYSKVLALY..NQHNPGASASPCCVPQALEPLPIVYY.VGRKPKVEQLSNMIVRSCKCS
Activin β A-chain       GLECDGKVNICCKK.QFFVSF.KDIGWNDWIIAPSGYHANYCEGECPSHIAGTSGSSLSFHSTVINHYRMRGHSPFANLKSCCVPTKLRPMSMLYYDDGQNIIKKDIQNMIVEECGCS
BMP-2                 AKHKQRKRLKSS.CKRHPLYVDF.SDVGWNDWIVAPPGYHAFYCHGECPFPLADHLNST..NHAIVQTLVNSVNSKIP...KACCVPTELSAISMLYLDENEKVVLKNYQDMVVEGCGCR
                                   25-28                3             28-32              0            30-34              1
```

B

```
C(25-28)C(3)C(28-32)C(0)C(30-34)C(1)C
```

C

#	Probability	ID	RF	1st Aa	Pattern: C	25-28	C	3	C	28-32	C	0	C	30-34	C	1	C	Description
1	3.78108e-013	MMINAS	2	215	C	28	C	3	C	30	C	0	C	34	C	1	C	M.musculus mRNA for inhibin alpha subunit
2	3.78108e-013	RNINHBAB	1	180	C	28	C	3	C	30	C	0	C	34	C	1	C	Rat inhibin alpha-subunit gene, exon 2.
3	3.78108e-013	MMINHIBB	3	182	C	28	C	3	C	30	C	0	C	34	C	1	C	Mouse inhibin alpha-subunit exon 2, complete cds.
4	3.78108e-013	RRINHA	1	348	C	28	C	3	C	30	C	0	C	34	C	1	C	Rat inhibin alpha-subunit mRNA, complete cds.
5	3.78108e-013	MMALPHA	3	270	C	28	C	3	C	30	C	0	C	34	C	1	C	M.musculus mRNA for inhibin alpha-subunit.
6	4.10151e-013	S77492	1	258	C	28	C	3	C	32	C	0	C	31	C	1	C	BMP-3=bone morphogenetic protein 3 [rats, FRC, mRNA Partial, 1292nt].
7	4.10151e-013	RRPBIP	3	453	C	28	C	3	C	32	C	0	C	31	C	1	C	Rat mRNA for prepro bone inducing protein, complete cds.
8	4.10151e-013	S82648	3	415	C	28	C	3	C	32	C	0	C	31	C	1	C	growth-differentiation factor-10=bone morphogenetic protein-3homolog
9	4.10151e-013	RNPBMP3	2	567	C	28	C	3	C	32	C	0	C	31	C	1	C	Rat mRNA for prepro bone morphogenetic protein-3.
10	4.44909e-013	MM08337	1	501	C	28	C	3	C	31	C	0	C	31	C	1	C	Mus musculus CD-1 putative growth factor GDF5 (Gdf5) mRNA, completecds.
11	4.44909e-013	MM08339	1	50	C	28	C	3	C	31	C	0	C	31	C	1	C	Mus musculus BALB/c putative growth factor GDF7 (Gdf7) gene,partial cds.
12	4.44909e-013	RNVGR	1	106	C	28	C	3	C	31	C	0	C	31	C	1	C	Rat vgr mRNA
13	4.44909e-013	MMBMP5M	1	351	C	28	C	3	C	31	C	0	C	31	C	1	C	Mus musculus bone morphogenetic protein 5 (BMP5) mRNA, completecds.
14	4.44909e-013	MM08338	1	24	C	28	C	3	C	31	C	0	C	31	C	1	C	Mus musculus BALB/c putative growth factor GDF6 (Gdf6) gene,partial cds.
15	4.44909e-013	MMVGR1	1	481	C	28	C	3	C	31	C	0	C	31	C	1	C	M.musculus Vgr-1 mRNA
16	4.44909e-013	MMVG1A	1	342	C	28	C	3	C	31	C	0	C	31	C	1	C	Mouse Vgr-1 protein mRNA, complete cds.
17	4.44909e-013	MMOP1G	2	363	C	28	C	3	C	31	C	0	C	31	C	1	C	Mouse OP-1 mRNA for osteogenic protein 1
18	4.44909e-013	AF101440	3	208	C	28	C	3	C	31	C	0	C	31	C	1	C	Mus musculus bone morphogenetic protein 10 (Bmp10) gene, exon 2 andcomplete cds.
19	4.82614e-013	RNBOMPR4A	2	441	C	28	C	3	C	30	C	0	C	31	C	1	C	R.norvegicus bone morphogenetic protein 4 mRNA
20	4.82614e-013	RNBMP2	2	314	C	28	C	3	C	30	C	0	C	31	C	1	C	R.norvegicus mRNA for bone morphogenetic protein 2
21	4.82614e-013	MMBMP4	3	426	C	28	C	3	C	30	C	0	C	31	C	1	C	M.musculus mRNA for bone morphogenetic protein 4 (BMP-4)
22	4.82614e-013	S65032	2	397	C	28	C	3	C	30	C	0	C	31	C	1	C	bone morphogenetic protein-4 [mice, Dunn osteosarcoma, mRNA, 1692nt]
23	4.82614e-013	RNMORPH2A	2	314	C	28	C	3	C	30	C	0	C	31	C	1	C	Rat bone morphogenetic protein 2 related mRNA sequence.
24	4.82614e-013	MMNOD	1	363	C	28	C	3	C	31	C	0	C	30	C	1	C	M. musculus nodal gene, a TGF-beta-like gene
25	4.82614e-013	MMBMP2A	2	2878	C	28	C	3	C	30	C	0	C	31	C	1	C	Mus musculus bone morphogenetic protein 2 (BMP-2) gene, completecds.
26	5.67879e-013	S98336	1	753	C	25	C	3	C	32	C	0	C	30	C	1	C	Mullerian inhibiting substance [rats, Genomic, 2760 nt]
27	5.67879e-013	MMAMH	3	830	C	25	C	3	C	32	C	0	C	30	C	1	C	M.musculus mAmh gene for anti-Mullerian hormone
28	6.68208e-013	RN03491	3	448	C	27	C	3	C	28	C	0	C	30	C	1	C	Rattus norvegicus Wistar transforming growth factor beta-3 mRNA,complete cds.
29	6.68208e-013	MMTGFRNA	2	411	C	27	C	3	C	28	C	0	C	30	C	1	C	Mouse transforming growth factor beta mRNA (TGF-beta), completecds.
30	6.68208e-013	MMTGFB3	2	517	C	27	C	3	C	28	C	0	C	30	C	1	C	mouse transforming growth factor beta-3 mRNA, complete cds.
31	6.68208e-013	RNTGFB1	2	431	C	27	C	3	C	28	C	0	C	30	C	1	C	Rat mRNA for transforming growth factor-beta 1
32	6.68208e-013	MMU009862	1	583	C	27	C	3	C	28	C	0	C	30	C	1	C	Mus musculus mRNA for transforming growth factor-beta 1
33	6.68208e-013	MMTGFB2	3	723	C	27	C	3	C	28	C	0	C	30	C	1	C	Mouse mRNA for transforming growth factor-beta2
34	1.97121e-011	MATGFB1	3	41	C	27	C	3	C	28	C	0	C	32	C			M.auratus mRNA for transforming growth factor beta-1
35	1.77251e-010	MMGDF1	2	323	C	28	C	3	C	36	C	0	C	31	C	1	C	Mouse transforming growth factor-beta (GDF-1) mRNA, complete cds.
36	1.77251e-010	MMOP2A	3	328	C	28	C	3	C	31	C	0	C	31	C	1	C	Mus musculus osteogenic protein-2 (OP-2) mRNA, complete cds.
37	1.77251e-010	MMGDF1A	1	760	C	28	C	3	C	36	C	0	C	31	C	1	C	Mus musculus growth/differentiation factor 1 (GDF-1) mRNA, completecds.
38	1.77251e-010	MM39545	1	360	C	28	C	3	C	31	C	0	C	31	C	1	C	Mus musculus bone morphogenetic protein 8B (Bmp8b) mRNA, completecds.
39	1.92272e-010	MMCTBC02	2	175	C	27	C	3	C	35	C	0	C	31	C	1	C	Mus musculus activin beta-c precursor gene, exon 2 and completecds.
40	1.92272e-010	MMU96386	3	319	C	27	C	3	C	33	C	0	C	31	C	1	C	Mus musculus activin beta E subunit mRNA, complete cds.
41	1.92272e-010	MMU95962	2	297	C	27	C	3	C	35	C	0	C	31	C	1	C	Mus musculus activin betaC subunit mRNA, complete cds.
42	1.92272e-010	MMINHACTS	3	291	C	27	C	3	C	35	C	0	C	31	C	1	C	M.musculus mRNA for inhibin/activin bC subunit
43	1.92272e-010	MMINBBCX2	2	230	C	27	C	3	C	35	C	0	C	31	C	1	C	M.musculus inbbc gene, exon 2
44	1.92272e-010	RRINHB	1	374	C	27	C	3	C	35	C	0	C	31	C	1	C	Rat inhibin beta-A-subunit mRNA, complete cds.
45	1.92272e-010	MMBETA	2	377	C	27	C	3	C	35	C	0	C	31	C	1	C	M.musculus mRNA for inhibin beta-A subunit
46	1.92272e-010	MMBETAB	1	19	C	27	C	3	C	34	C	0	C	31	C	1	C	M.musculus mRNA for inhibin beta-B subunit
47	1.92272e-010	AF089825	3	311	C	27	C	3	C	33	C	0	C	31	C	1	C	Rattus norvegicus activin beta E mRNA, complete cds.
48	2.08566e-010	MMU84005	2	317	C	26	C	3	C	25	C	0	C	31	C	1	C	Mus musculus growth/differentiation factor-8 mRNA, complete cds.
49	2.08566e-010	AF019624	1	283	C	26	C	3	C	25	C	0	C	31	C	1	C	Rattus norvegicus myostatin (MSTN) mRNA, complete cds.
50	2.45415e-010	MMU78109	1	217	C	26	C	3	C	29	C	0	C	28	C	1	C	Mus musculus prepro-neurturin mRNA, complete cds.
51	2.66212e-010	MMGDNF	1	118	C	26	C	3	C	28	C	0	C	28	C	1	C	Mouse mRNA for glial cell line-derived neurotrophic factor(GDNF),complete cds.
52	2.66212e-010	MM37459	2	161	C	26	C	3	C	28	C	0	C	28	C	1	C	Mus musculus glia-derived neurotrophic growth factor (GDNF) mRNA,complete cds.
53	2.66212e-010	MMD350S3	1	91	C	26	C	3	C	28	C	0	C	28	C	1	C	Mus musculus gene for neurotrophic factor, exon 3 and complete cds.
54	2.66212e-010	S75585	1	92	C	26	C	3	C	28	C	0	C	28	C	1	C	glial cell line-derived neurotrophic factor GDNF555 {alternativelyspliced}
55	2.66212e-010	S75583	1	118	C	26	C	3	C	28	C	0	C	28	C	1	C	glial cell line-derived neurotrophic factor GDNF633

D

```
GDNF SPDKQAAALPRRERNRQAAAASPENSRGKGRRGQRGKNRGCVLTAIHLNVTDLGLG.YETKEELIFRYCSGSCESAETMYDKILKNLSRSRRLTSDKV.GQACCRPVAFD..DDLSFLDDNLVYHILRKHSAKRCGCI
NTN                                           PGARPCGLRELEVRVSELGLG.YTSDETVLFRYCAGACEAAIRIYDLGLRRLRQRRRVRRERARAHPCCRPTAYE..DEVSFLDVHSRYHTLQELSAREСACV
TGF-β1                                ALDTNYCFSSTEKNCCVRQLYIDFRKDLGWKWIHEPKGYHANFCLGPC.PYIWSLDTQYSKVLALYNQHNPGASASPCCVPQALEPLPIVYYVGRKPKVEQLSNMIVRSCKCS
                                                       26-27          3        28-29           0          28-30          1
```

Fig. 3. (A) Alignment of the primary structure of the three members of the TGF-β family, (B) the conserved Cys pattern formatted as a Motifer query, (C) the output of Motifer after a search on the EMBL rodent database with the conserved Cys pattern of the TGF-β family as query and an alignment of TGF-β1 to cell-line-derived GDNF and NTN, two distant members of the TGF-β family (D). In A, the alignment of TGF-β1 (SWISS-PROT P04202), activin-β A-chain (SWISS-PROT Q04998) and BMP-2 (SWISS-PROT P21274) shows the level of similarity and the identity within the TGF-β family consisting of more than 30 members. Conserved residues are shaded and numbers indicate the Cys distances of the TGF-β family. In C, the output of a search of the EMBL rodent database with the conserved Cys pattern of B shows the occurrences of NTN (ID: MMU78109) and GDNF (ID: MMGDNF) at rows 50 and 51, respectively. In D, NTN and GDNF found in C are aligned. Note the mismatch between the conserved Cys pattern of the TGF-β family and that of GDNF and NTN at the penultimate inter-Cys distance. The search parameters used allowed Motifer to correct for the mismatch and provide GDNF and NTN with a high score in spite of an erroneous query pattern.

The choice of database to search must also be taken into consideration. EST or genomic databases would be used to find novel proteins, while annotated databases would be used to describe new relationships or functions to already cloned sequences. However, the sequence information in the different databases is not comparable. The EST databases contain frequent sequencing errors and the probability for a frame shift between two residues separated by a large distance is high. Because of this, frame shifts have to be tolerated to a higher degree when searching EST databases than when searching annotated databases. In addition, ESTs are generally sequenced and submitted to the databases without any directional information. Therefore, all reading frames must be examined, while the reverse reading frames can generally be omitted when searching annotated databases. Several features of Motifer are aimed at handling and reducing the impact of those problems, but the user must be aware of the inherent properties of the database under investigation. Please refer to Fig. 2 for a suggested approach to database searches using Motifer.

3.2. Distant family assignments

To assess the potency of the program, the EMBL rodent database was searched with the conserved Cys pattern of the

transforming growth factor-β (TGF-β) family [7] as query (Fig. 3). As expected, all close family members of TGF-β were found among the top hits, a result similar to that obtained with FASTA, BLAST and other search tools using full-length TGF-β as query. However, among the top hits reported by Motifer were also glial cell-line-derived neurotrophic factor (GDNF) [8] and neurturin (NTN) [9], two distantly related members of the TGF-β family with just 16 and 23% residue identity to TGF-β1, respectively. The observed amino acid identity is largely due to the strictly conserved Cys pattern, with only the penultimate inter-Cys distance showing a discrepancy between GDNF and NTN on the one hand and the other TGF-β family members on the other (Fig. 3D). Because of this mismatch, the hits for GDNF and NTN are interrupted and split into two parts, but Motifer is able to correctly combine the two separate hits into one and thereby to give the sequence a high score in spite of an erroneous query pattern (Fig. 3C). Furthermore, with the features aimed at reducing background hits, the number of false hits is quite low, despite the allowance and correction for an erroneous and short query pattern.

In general, any highly conserved pattern can be used to identify members of a protein family. However, since relatively few amino acids are used as query, the probability for a false random hit is high. To minimise the number of false hits, it is suitable to use relatively rare amino acids when constructing the query pattern and to extend the query pattern as far as possible. A search using the seven conserved cysteines of the TGF-β family gives both a high specificity and a high sensitivity as shown above. Another approach is to use rare amino acid residues such as tryptophan in the search pattern. To illustrate this, the rodent and human annotated EMBL databases were searched with the simplified WW-domain consensus pattern W(7–9,G)GK(13н16,K)W(2)P. The WW-domain is involved in protein-protein interactions and it shows functional similarity to the SH3-domain that binds proline-rich ligands [10]. To reduce the background of the search, all hits in reverse reading frames and in sequences longer than 1000 amino acids were excluded. In spite of the high probability for a random hit (in the order of 10^{-7}), there were only 11 hits among the top 30 of the output that were judged as non-relevant because of adjacent stop codons. A further 11 hits corresponded to known proteins with WW-domains as expected, while eight sequences matched the WW-domain query but lacked other homologies to the WW-domain and had not been described as WW-proteins. The validity of the search was confirmed by the finding of YAP, CSF1-R, dystrophin, NEDD4, PDGFR, c-Kit and other WW-domain proteins. The eight novel proteins found may constitute thus far unknown WW-domain forms or be co-incidental, but in either case, the WW-domain search, together with the TGF-β search, serves to illustrate the ability of Motifer to identify very remote protein relationships.

3.3. 'In silico' cloning

In an effort to find new members of the cytochrome P450 superfamily [11], all known members of the family were aligned and the conserved residues were used to search the EMBL EST database with Motifer. To reduce the number of the many already known sequences in the output, BLAST searches were made with a number of known cytochrome P450 enzymes prior to the search with Motifer and the sequences thus found were given to Motifer to be excluded from the output of the search. The experiment was focused on finding human homologues and by limiting the species examined, the number of false hits was further restricted. In spite of these efforts to reduce the output, the search result still contained known sequences that had not been excluded and random false hits. However, an EST corresponding to a hitherto unknown cytochrome P450 enzyme was found (M. Oscarson, in preparation). The novel member belongs to the CYP2 family of cytochrome P450 enzymes and shares the same haeme-binding region (GXRXCXG).

This 'in silico' cloning of a novel enzyme illustrates the features of Motifer aimed at reducing the background of known sequences and the ability of the program to screen a large nucleotide database translated into all six reading frames for a given amino acid pattern. The novel cytochrome P450 was found at position 76 in the output of the search of the EMBL EST database, performed as described above, whereas a search done on the same database and with the same query, but not excluding known related sequences and not restricting the search to *Homo sapiens*, put the novel hit at position 1605. A graphical interpretation of the output and the positions of the regions detected with Motifer can be obtained by viewing the output text file in a font of size 1. This will in some instances give further indication as to which sequence to analyse or what part of a long query to focus on for a second search.

Combined, the results illustrate the ability of Motifer to handle frame shifts and minor discrepancies between wanted sequences and the query pattern. Motifer is concluded to extend existing tools with which to search the rapidly accumulating amount of nucleotide information.

Acknowledgements: I am grateful to Drs. Michael Fainzilber (The Weizmann Institute, Israel) for initial stimulation, Mikael Oscarson (Karolinska Institutet) for invaluable discussions and for collaboration on the cytochrome P450 searches, Rizaldy Scott (Karolinska Institutet) for information regarding WW-domains and Carlos F. Ibáñez (Karolinska Institutet) for constructive interest.

References

[1] Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA 85, 2444–2448.
[2] Lipman, D.J. and Pearson, W.R. (1985) Science 227, 1435–1441.
[3] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[4] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[5] Zhang, Z., Schäffer, A.A., Miller, W., Madden, T.L., Lipman, D.J., Koonin, E.V. and Altschul, S.F. (1998) Nucleic Acids Res. 26, 3986–3990.
[6] Nakamura, Y., Gojobori, T. and Ikemura, T. (1998) Nucleic Acids Res. 26, 334.
[7] Kingsley, D.M. (1994) Genes Dev. 8, 133–146.
[8] Lin, L.-F.H., Doherty, D.H., Lile, J.D., Bektesh, S. and Collins, F. (1993) Science 260, 1130–1132.
[9] Kotzbauer, P.T., Lampe, P.A., Heuckeroth, R.O., Golden, J.P., Creedon, D.J., Johnson Jr., E.M. and Milbrandt, J. (1996) Nature 384, 467–470.
[10] Sudol, M. (1996) Prog. Biophys. Mol. Biol. 65, 113–132.
[11] Nelson, D.R. et al. (1996) Pharmacogenetics 6, 1–42.