A Generalized Profile Syntax for Biomolecular Sequence Motifs and its Function in Automatic Sequence Interpretation

Philipp Bucher
Swiss Institute for Experimental Cancer Research
Ch. des boveresses 155
CH–1066 Epalinges s/Lausanne, Switzerland
pbucher@isrec-sun1.unil.ch

Amos Bairoch
Medical Biochemistry Department
Centre Médical Universitaire
CH-1211 Geneva, Switzerland
bairoch@cmu.unige.ch

Abstract

A general syntax for expressing biomolecular sequence motifs is described, which will be used in future releases of the PROSITE data bank and in a similar collection of nucleic acid sequence motifs currently under development. The central part of the syntax is a regular structure which can be viewed as a generalization of the profiles introduced by Gribskov and coworkers. Accessory features implement specific motif search strategies and provide information helpful for the interpretation of predicted matches. Two contrasting examples, representing E. coli promoters and SH3 domains respectively, are shown to demonstrate the versatility of the syntax, and its compatibility with diverse motif search methods. It is argued, that a comprehensive machine-readable motif collection based on the new syntax, in conjunction with a standard search program, can serve as a general-purpose sequence interpretation and function prediction tool.

Introduction

Nucleic acid and protein sequence motifs are popular research objects of computational biologists for various reasons. Machine-readable motif descriptions can be used for automatic structure and function prediction. The exercise of defining a motif may provide insights into molecular mechanisms of gene expression, from transcriptional activation via RNA processing and protein folding to physiological activity. Finally, there are exciting potentials of synergism with other fields such as speech recognition, exemplified by dynamic programming algorithms and hidden Markov models.

The concepts of a sequence motif itself evades exact definition. It necessarily implies some kind of structured similarity but may have functional aspects too. In the biological literature, the term motif often refers to short regions of sequence similarity. Here, it is used in a broader sense encompassing also larger objects such as protein families.

The components of a complete research methodology are diagramed in Fig. 1. The process of defining of a sequence motif starts with a set of data, alternatively called observations, and ends with a formal description of the motif. The roles of data and description are reversed in a motif search method. The observations may consist of sequences only, or include structural and functional information as well. Functional information may have the form of numbers or qualitative labels attached to sequence residues, or be implicit in a given classification. The formal motif description can be a specific motif search algorithm. More frequently, it is a separate data structure, known under names like pattern, weight matrix, or profile. In order to serve as input to a motif search algorithm, the description must conform to syntactic standards. In practice, the boundary between motif description and search algorithm is not always obvious.

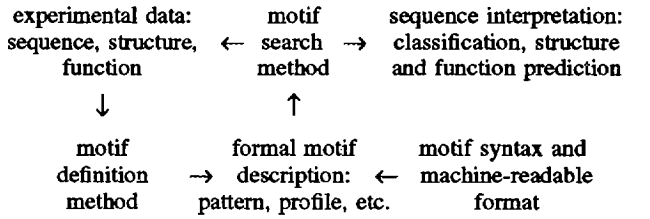


Figure 1. Components of a motif research methodology.

There is a great diversity of motif definition methods relying on different mathematical, physical and biological theories. Individual techniques are more or less specific for certain classes of biological objects, or input data structures. A large subgroup of methods proceeds via the intermediate step of a multiple sequence alignment. Other approaches exploit the self-loop on the input data visualized in Fig. 1, by minimizing the discrepancies between observation and prediction with machine learning techniques. In contrast, the methods for motif search are rather uniform. Most of them can be understood as algorithms solving an optimal alignment problem, either by a rigorous or a heuristic approach. However, this is not always obvious in the original descriptions. The E. coli promoter example will illustrate this point.

Not all motifs have a known biological function. Often the discovery of significant sequence similarity precedes the identification of a physiological role. In this case, the motif search method merely serves to locate new motif occurrences. In the other case where functional information is available, the motif search method becomes a full-fledged structure or function prediction

This work was supported in part by grant 31-37687.93 from the Swiss National Science Foundation.

method, provided that the biological knowledge is passed on to a software tool in appropriate form. The mapping of physiological properties to individual residues of a sequence occurs via an implicit or explicit motif-to-sequence alignment.

In this paper we describe and discuss a general syntax to express biomolecular sequence motifs based on a quantitative descriptor similar to weight matrices or profiles. The syntax and corresponding format conventions will be used in the PROSITE data bank and in a similar compilation of nucleic acid sequence motifs currently under development. PROSITE is an annotated collection of protein sites and patterns, distributed as a human and machine readable text file. With over 1000 entries, it is the most comprehensive of its kind (Bairoch 1993). The new syntax will allow description of a wider range of biological objects, including highly divergent protein domains which escape the regular expression-like pattern syntax exclusively used so far. The future nucleic acid sequence motif collection will mimic the format of PROSITE and cover objects like gene expression signals, DNA-protein binding sites, and interspersed repetitive elements.

The definition of a general motif syntax is motivated in part by a software concept. Underlying is the belief that a comprehensive motif library in conjunction with a standard search algorithm can function as a general-purpose sequence interpretation and function prediction tool. From a computational viewpoint, the motif syntax approaches the role of a specialized symbolic programming language, the search tool that of an interpreter. In this design, a text file format transports the knowledge arising from specialized motif definition efforts to the bench biologist's sequence analysis programs, eliminating the need for new software development. The experience with PROSITE proves that this scheme is successful in making motifs rapidly available to users.

In order to meet the expectations outlined above, a general motif syntax must be versatile in order to represent a large variety of biological objects. Furthermore, it should be capable of accommodating the output of many different motif definition methods, and of translating it into precise search instructions. Moreover, it should should serve as a vector for structural and functional information. From yet another viewpoint, a clear structure and conceptual parsimony are qualities which make a standard format acceptable to others. The design of the new syntax was guided by such considerations.

The emphasis in this paper is on concepts and examples. Space limitations preclude a detailed description of all syntactic features. The specific file format used in PROSITE is described in the documentation distributed along with this data bank.

Structure and Function of the Generalized Profile Syntax

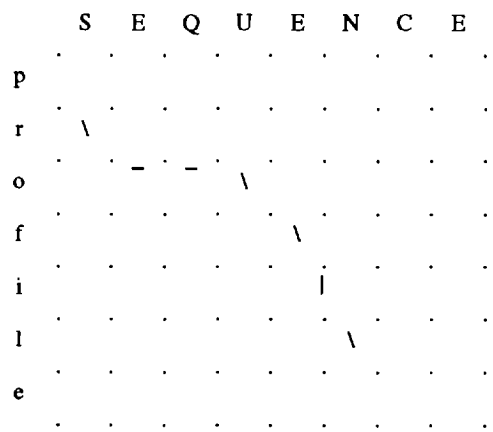
A motif description based on the generalized profile syntax will be called a profile. It consists of two parts: a regularly structured basic profile, and so-called accessories. The exclusive function of the basic profile is to assign a number to an alignment between itself and a sequence. This number will be referred to as a similarity score. The accessories provide additional information guiding motif search operations and interpretation of the results.

Basic profile structure

The name indicates that the new data structure can be viewed as a generalization of the profiles introduced by Gribskov, McLachlan, & Eisenberg (1987). Similar motif descriptors are also known under names like weight matrices (e.g. Staden 1990) or flexible patterns (Barton & Sternberg 1990).

A profile consists of an alternating sequence of match and insert positions. The two types of positions contain complementary sets of numeric parameters called profile scores. Match positions correspond to residues which typically occur in a sequence motif. Insert position are places where additional residues can be inserted.

Profile scores serve to compute a similarity score. In order to make clear what kind of scores are needed at which type of position, the notion of a profile-sequence alignment needs first to be introduced. The path matrix representation known from pairwise sequence alignments is helpful to this end.



In the above diagram, the capital letters represent sequence residues, the lower-case letters represent profile match positions. Profile insert positions are not marked by symbols. The path indicated by horizontal, vertical, and diagonal bars defines the following alignment:

S E Q U E – N
r – – o f i l

Each possible alignment corresponds to a path in the path matrix with a unique coordinate sequence. The above alignment has the following coordinate sequence:

(1,0), (2,1), (2,2), (2,3), (3,4), (4,5) ,(5,5) ,(6,6) .

By convention, the upper left corner of the matrix is assigned coordinates (0,0). Note that path matrix coordinates coincide with profile insert positions rather than with match positions, and fall between consecutive residues of the sequence. This observation has a bearing on the profile structure.

A coordinate sequence (i_0,j_0), (i_1,j_1) ... (i_L,j_L) represents a valid alignment between a profile of length N and a sequence of length M if, and only if:

$$\{\ 0 \le i_k \le N \text{ AND } 0 \le j_k \le M\ \} \text{ for } 0 \le k <= L$$

$$\text{AND } \{\ (\ i_k+1=i_{k+1} \text{ AND } j_k+1=j_{k+1}\ )$$

$$OR\ (\ i_k=i_{k+1} \text{ AND } j_k+1=j_{k+1}\ )$$

$$OR\ (\ i_k+1=i_{k+1} \text{ AND } j_k=j_{k+1}\ )\ \} \text{ for } 0 \le k \le L-1\ .$$

Note that this definition encompasses both global and local types of alignments. In the following, it is not necessary to distinguish between these two alternatives. A global alignment may simply be viewed as a limit case of a local alignment.

A profile assigns a number to any possible alignment. The concept of an optimal alignment is of no importance in this context. The similarity score is defined as the sum of the profile scores assigned to all scorable components of the alignment. These components are: the *beginning*, all *extension steps*, all *state transitions*, and the *end*. Some of these terms need further explanation. An extension step occurs between any consecutive elements of the coordinate sequence. There are three different types of extension steps: match, insert, and deletion extension steps. In the above path matrix diagram, diagonal bars represent match steps, horizontal bars represent insert steps, and vertical bars represent deletion steps. The number of extension steps defines the length of the alignment. The types of extension steps are also called states; this term is borrowed from hidden Markov models (Krogh et al. 1994). Each extension step is thus associated with a state. The beginning and the end of an alignment are also considered states. A state transition occurs between any two consecutive states, also between identical states. In summary, the similarity score of an alignment of length L is the sum of $2L+1$ component scores: 1 initiation, 1 termination, L extension, and $L+1$ state transition scores. All scores are provided by the profile itself in a position-specific manner. As a consequence, the similarity score does not depend on additional parameters intrinsic to an alignment method. The different types and functions of profile scores are now explained.

The scores for the beginning and for the end of the alignment are called initiation scores and termination scores. There are two types of scores for each class. The external initiation score applies to alignments starting at the beginning of the sequence. The internal initiation score applies to alignments starting at a sequence internal position. External and internal termination scores are defined analogously. The function of initiation and termination scores is to flexibly encode local or global alignment scoring modes. In addition, they may serve to anchor a motif at the beginning or at the end of a sequence.

The scores for extension steps comprise three classes: match extension scores, insert extension scores, and deletion extension scores. Match and insert extension scores are residue-specific because the corresponding alignment steps involve one sequence residue. The deletion extension score, which does not involve a sequence residue, is constant. There are 16 different types of state transition scores for all meaningful successions (see Table 1). State transition scores serve similar functions as gap opening and gap extension weights in the scoring of a pairwise sequence alignment.

Table 1. Position-specific profile scores

Insert position:	
B^e	external initiation score
B^i	internal initiation score
E^e	external termination score
E^i	internal termination score
$T_{B\to M}$	state transition score beginning to match
$T_{B\to I}$	state transition score beginning to insert
$T_{B\to D}$	state transition score beginning to deletion
$T_{B\to E}$	state transition score beginning to end
$T_{M\to M}$	state transition score match to match
$T_{M\to I}$	state transition score match to insert
$T_{M\to D}$	state transition score match to deletion
$T_{M\to E}$	state transition score match to end
$T_{I\to M}$	state transition score insert to match
$T_{I\to I}$	state transition score insert to insert
$T_{I\to D}$	state transition score insert to deletion
$T_{I\to E}$	state transition score insert to end
$T_{D\to M}$	state transition score deletion to match
$T_{D\to I}$	state transition score deletion to insert
$T_{D\to D}$	state transition score deletion to deletion
$T_{D\to E}$	state transition score deletion to end
$I(X)$	insert extension score for residue X
$I(*)$	insert extension score for residue not contained in the alphabet
Match position:	
$M(X)$	match extension score for residue X
$M(*)$	match extension score for residue not contained in the alphabet
D	deletion extension score

A profile is based on a specific alphabet, usually the twenty-letter amino acid or the four-letter nucleotide alphabet. The alphabet is considered a basic constituent of the profile because it determines the number of profile scores per insert and match position. There is one insert extension score and one match extension score for each character in the alphabet. In practice, it is useful to define an additional insert and match extension score to deal with unexpected characters appearing in real sequences. All other types of scores are residue independent.

A look at the path matrix diagram makes clear which type of score is associated with which type of profile position. Initiation, termination, state transition, and insert extension scores belong to insert positions. Match extension and deletion extension scores belong to match positions. Profile scores may be integer or real numbers. In addition they may assume a special value representing a forbidden residue or a forbidden alignment operation.

A last point needs to mentioned. A profile has a defined topology, either linear or circular. Molecular sequences can also be linear or circular. A linear profile starts and ends with an insert position. The previously presented definition of a sequence alignment applies only to linear profiles and sequences. Alignments involving circular profiles or circular sequences, or both, correspond to paths on a cylindrical surface, or on a torus. Adaptation of the alignment definitions is straightforward.

Circular profiles may represent motifs which occur as a variable number of tandemly repeated units. A special type of circular profiles, consisting of a single match position, may be used to represent protein domains of biased amino acid composition.

Profile accessories

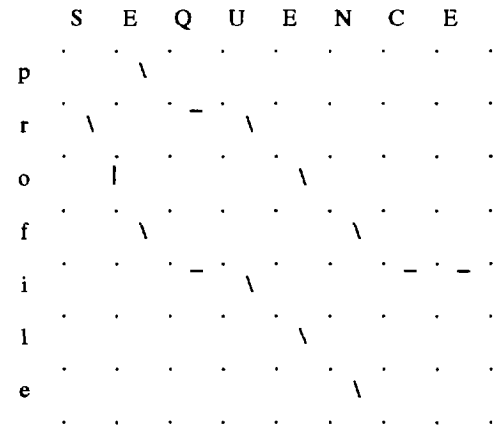
The primary purpose of a profile is to identify as reliably as possible biologically relevant motif occurrences in new sequences. The specifications contained in the basic profile structure are not sufficient to define a rational search strategy to this end. The so-called profile accessories fill this gap and provide additional information guiding the interpretation of profile matches. Four types of accessories are distinguished: cut-off values, disjointness definitions for multiple matches, score normalization modes, and feature tables. The first two are essential for delivering determinative search instructions to a profile search algorithm, and thus are obligatory components of a profile. The third and the fourth are optional. In contrast to the components of the basic profile structure, which will be kept stable, profile accessories are conceived as a growing collection of features. The assumption is that incorporation of new accessories will not immediately affect the basic operations of a motif search algorithm.

The function of a cut-off value is to a priori exclude a large number possible alignments between a profile and a sequence from further consideration by a motif search algorithm. The fate of the remaining sequences, with similarity scores higher than or equal to the cut-off value, depends on the disjointness definition applied. Another aspect of a cut-off value is that it gives a profile a qualitative meaning. The qualitative interpretation may be used to assess the performance of a motif description by statistics of false positives and false negatives. A two state prediction assay may also be at the center of a motif definition method relying on machine learning techniques. In certain situations, it may be appropriate to supply more than one cut-off value, partitioning the range of alignment scores into multiple areas. The areas may correspond to different degrees of certainty, ranges of evolutionary distance, or levels of physiological activity. The syntax allows descriptions to be attached to each cut-off level.

The notion of disjoint alignments is subtle. There are situations where only a single best alignment and its similarity score are of interest. This arises for instance with a profile serving exclusively as a signature for a protein family. More frequently, the same motif may occur more than once in a given sequence, and each occurrence will be of interest. In the first case, the motif search problem is simple and can be solved by a standard optimal alignment algorithm such as described by Gribskov, Lüthy, & Eisenberg (1990). In the second case, the task is more difficult to define and involves the notion of disjointness.

The conceptually simplest approach would be to list all profile-sequence alignments with similarity scores greater than or equal to the cut-off value. However, such proceeding would not yield useful results because high scoring alignments typically occur as clusters of numerous overlapping alignments with comparable scores. Two members of such a group may differ only by one extension step at the end of one alignment. In pairwise comparisons, a group of overlapping alignments is represented by a single highest scoring member (for a review, see Pearson & Miller, 1992). This seems a reasonable approach for profiles too.

The technical term for the opposite of alignment overlap overlap is disjointness. It is important to recognize that there are many ways to define this relation. The methods developed for finding multiple, locally optimal alignments between two sequences consider two alignments disjoint if they have no extension step in common (Waterman & Eggert 1987). The alignments shown in the path matrix below illustrate this specific notion.



However, such a definition may not be adequate in many motif search application, as it allows the same sequence residue to be matched with multiple profile positions, which may have mutually exclusive functions. Imagine the case of a protein structural domain. There, it is inconceivable that the same residue simultaneously participates in the formation of two physically distinct domains, occupying different places within these domains. There may be no single disjointness definition adequate for all kinds of biological sequence motifs amenable to profile representation. Therefore, a specific notion of disjointness is viewed and implemented as a profile-inherent property rather than as a variable of an alignment method. The profile syntax provides a list of parameterized disjointness definitions. Two types are currently used. The first one requires that a certain area of the profile is not matched with overlapping segments of the same sequence. The parameters of this definition

are the starting and end points of the protected area. The second represents a special case in that it allows only one optimal alignment per sequence. This definition implements a conventional optimal alignment search strategy.

The ensemble of profile, cut-off value, and disjointness definition, in principle, amounts to determinative instructions for a motif search algorithm. In practice, this is only approximately true, because of ambiguities in the statement of the multiple local alignment problem. The available algorithms for multiple local alignments proceed sequentially by first selecting a best alignment, and then iteratively selecting a next best one from the pool of remaining alignments disjoint from those already accepted, until there are no more alignments passing the cut-off criterion (e.g. Waterman and Eggert 1987). This principle is compatible with profile search and with a large variety of disjointness definitions. We are in the process of developing a space-efficient algorithm for profile search along the lines of Huang & Miller (1991), with generalizations for circular profiles and sequences. It needs to be pointed that the information contained in a motif description only attempts to state the goal of a profile-sequence comparison, leaving space for alternative algorithmic solutions to achieve this goal.

Normalization modes translate the raw-scores computed directly from the profiles scores into more easily interpretable units. There may be multiple normalization modes for the same profile, each one associated with a different mathematical, physical, or biological interpretation. Normalized units can be given a name which may appear in a program output.

Normalization functions are required to preserve the ranking of scores pertaining to alternative alignments between the same profile and the same sequence. However, since normalization functions may depend on sequence parameters such as length and residue composition, they will generally not preserve the order of scores pertaining to matches form different sequences arising in a database search. The profile syntax offers an expandable list of parameterized normalization functions. Different normalization modes of the same profile may be assigned a priority controlling certain operations of a motif search program, e.g. the output sorting of accepted matches. Cut-of values may be specified in normalized score units rather than raw score units.

Feature tables contain information on structural and functional properties of profile positions and regions. Programs may map these properties to residues in the sequence via a profile-sequence alignment and report the result of this mapping in an output listing. Although feature table information clearly belongs into a motif description, there was no need to develop new syntactic standards for it. Feature tables of profiles are analogous to feature tables of sequences, and therefore may rely on already existing conventions. PROSITE provides its own format for this kind of information.

Examples

E. coli Promoters

The profile shown in Table 2 describes the major class of E. coli promoters recognized by the RNA polymerase σ^{70} complex. It is based on work by Mulligan et al. (1984). The upper and middle parts of the Table contain the basic profile; the lower section presents the accessories.

The profile is substructured into four operationally distinct modules. Positions 1 to 16 contain a weight matrix characterizing the −35 region of E. coli σ^{70} promoters. The consensus box TTGACA starts at position 10. (The term weight matrix is used in the same sense as in Staden (1990), referring to a weighting table for fixed length sequences, not allowing for insertions or deletions.) Positions 17 to 25 represent a fixed-length spacer with no base preferences. Positions 26 to 31 constitute a variable length linker scoring module, the functioning of which will be explained below. Positions 32 to 45 contain a weight matrix characterizing the −10 region. The consensus box TTGACA starts at position 37.

The original authors defined a raw score which is the sum of the scores assigned by the two weight matrices to the putative −35 and −10 regions, plus a score assigned to the spacing between the core hexamer boxes. The scores for variable linker lengths are: 14 for the most preferred spacing of 17; 6 for spacings of 16 and 18; 1 for spacings of 15, 19, 20 and 21. All other spacings are forbidden. The linker length scoring module of the profile achieves this scheme as follows. The constant prohibitive values of most state transitions score defined in the upper part of Table 2, together with the prohibitive values assigned to the other state transition scores at most profile positions, make sure that insertions can occur nowhere, and that a deletion gap can only be opened at insert position 25 and must be closed before or at insert position 31. A promoter with the maximal linker length of 21 can be aligned without gap to the profile. In this case, the linker length score is provided by $T_{M\to M}$=1 at insert position 25. Promoters with linker lengths 15 to 20 require a deletion gap in their alignment to the profile. The corresponding scores are provided by the values of $T_{D\to M}$ at insert positions 26 to 31. The values of the initiation and termination scores implement a global alignment scoring mode with free endgaps in the profile as well as in the sequence.

The authors of the motif have developed a heuristic algorithm to identify high scoring promoter sites. The method, which has been implemented in a program named TARGSEARCH, proceeds in two steps. It first finds all three-out-of-six matches to the consensus sequences TTGACA and TATAAT, and then identifies all appropriately spaced heterologous pairs, for which it computes a similarity score. By contrast, the profile representation proposes a rigorous (and perhaps also more efficient) dynamic programming algorithm to achieve essentially the same goal. This illustrates that the exercise of reformulating an existing specialized method in terms of a general concept can lead to unexpected technical improvements.

Table 2. E. coli RNA polymerase sigma[70] promoters

Alphabet: {*A,C,G,T*} Toplogy: *linear*

Constant scores: B^e=0 E^e=0 $I(A)$ ... $I(T)$=0 $I(*)$=0 $M(*)$=0 D=0 $T_{B\to M}$=0 $T_{B\to I}$=* $T_{B\to D}$=*
$T_{B\to E}$=* $T_{M\to I}$=* $T_{M\to E}$=0 $T_{I\to M}$=* $T_{I\to I}$=0 $T_{I\to D}$=* $T_{I\to E}$=* $T_{D\to I}$=* $T_{D\to D}$=0 $T_{D\to E}$=*

Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
$M(A)$	11	8	8	7	8	7	3	5	5	0	1	0	14	5	9	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	5	3	4	4	0	20	5	12	11	0	7	4	6
$M(C)$	3	4	2	4	4	3	5	2	8	1	1	2	3	11	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	4	5	4	5	2	0	3	3	4	1	2	7	6
$M(G)$	3	2	4	2	4	5	5	5	5	2	1	17	1	2	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	5	8	7	2	0	3	3	3	0	6	5	6
$M(T)$	4	7	7	8	5	6	8	9	3	17	18	2	4	5	7	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	6	8	5	6	17	1	9	3	4	20	6	5	4
$T_{M\to D}$	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	0	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
$T_{M\to M}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$T_{D\to M}$	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	1	1	6	14	6	1	*	*	*	*	*	*	*	*	*	*	*	*	*	*
B^i	0	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
E^i	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	0

Normalization: Mode 1: Homology score (%) = −90.56 + 0.5773 × *raw score*
Mode 2: log KBk_2 = −10.12 + 0.0622 × *raw score*

Cut-off value: Homology score ≥ 45.0%

Disjoint definition: Two profile-sequence alignments are disjoint if and only if the sequence segments matched with profile pos. 37 to 42 do not overlap.

Features: pos. 10-15 TTGACA box (−35 region)
pos. 37-42 TATAAT box (−10 region)

* represents a prohibitive profile score.

The profile accessories define two normalization modes, a cut-off value and a disjointness definition. In addition, the position of the core −35 and −10 hexamer boxes are indicated as feature tables. The first normalization procedure, defining a so-called "Homology score", simply maps the range of possible raw score values to numbers between 0 and 100. The second normalization mode is more interesting. It provides an estimate of the second-order rate constant for open complex formation between RNA polymerase and the promoter. The parameters of the linear normalization function were derived by a linear regression analysis based on 31 transcriptionally assayed promoters. The resulting correlation coefficient was 0.83. The cut-off value reflects a recommendation made in the original paper. Note that the disjointness definition protects only the TATAAT box regions from sequence overlap. This attempts to capture the biological fact that two adjacent TATAAT boxes can direct transcription from two distinct initiation sites as close as six bp apart from each other.

SH3 domain

Table 3 shows a profile characterizing the Src homology domain SH3 (Musacchio et al. 1992). Unlike the previous example, this motif was recognized by sequence similarity alone, and is still awaiting discovery of a function. The profile has been derived with an extension (Lüthy, Xenarios, and Bucher 1994) of the basic profile construction method described by Gribskov, Lüthy, & Eisenberg (1990). The gap regions are handcrafted. Similar extensions of the profile methodology have recently been reported by another group (Thompson, Higgins, & Gibson 1994) and applied with similar success to the SH3 domain. This or a very similar profile will appear in the next PROSITE release. The SH3 domain is an examples of a motif which could not be accurately described by the regular expression-like syntax exclusively used until now.

The SH3 is domain is modeled by three homology blocks separated by two gap regions. Within the homology blocks, small insertions and deletions are not totally forbidden but strongly impeded by high gap costs defined by the four state transition scores shown below the match extension scores. These costs are substantially lower in the gap regions, which in addition offer a dummy match position that can be skipped freely. The constant initiation and termination scores set to zero define a local optimal alignment algorithm. This profile exclusively relies on syntactic elements compatible with the profile search and alignment methods described by Gribskov Lüthy, & Eisenberg (1990), and therefore can be automatically reformatted for use with GCG profile analysis programs. The only feature which cannot be emulated by these programs is the disjointness definition allowing for multiple matches in the same sequence, illustrating an important methodological improvement inherent in the new syntax.

The first normalization mode simply describes the integer to real conversion performed by the GCG programs. The second normalization modes, which produces length-adjusted Z-scores, relies on a formula proposed by Gribskov, Lüthy, and Eisenberg (1990). The parameters have been estimated by a search for SH3 domains in SWISS-PROT using the GCG program ProfileSearch. The cut-off value proposed is based on the same analysis. A lower priority is assigned to the length-dependent normalization mode because the corresponding scores are less efficient in separating true positives from true negatives than the original scores.

Discussion

The profile concept underlying the new syntax is purely operational. Its sole function is to assign a similarity score to a specifically aligned sequence segment. Besides this function, the profile scores are deliberately stripped of any other meaning that could impede its application in a particular context. Our goal was to unify the concepts and techniques relevant to the bench biologist, namely the methods to search for already defined sequence motifs, and notions guiding the interpretation of potential occurrences. With regard to the problem of defining a sequence motif ab initio, which primarily concerns computational biologists, methodological diversity is mandated by the even greater diversity of biological objects and the multiplicity of experimental procedures. In the past, different groups have used different approaches to define sequence motifs, different structures to represent motifs, and different algorithms to search for motifs. In the future, the generalized profile syntax will permit uniform representation of many motifs and thereby facilitate the task of software producers providing corresponding search tools to end users.

New algorithms are not described this paper. We don't think this is a requirement for understanding the new syntax. The problem which has to be tackled by a such algorithms is precisely stated by: (i) the definition of a profile-sequence alignment, (ii) the definition of the similarity score, and (iii) the definition of disjointness of two alternative alignments. The task of finding multiple profile matches is exactly analogous to the the task of finding multiple best sequence alignments. This problem is thoroughly discussed in (Waterman & Eggert 1987). The definition of the similarity score for a profile match is such that all algorithms designed for pairwise sequence alignment and using a conventional scoring system are applicable to to profiles with minor modifications. This also holds for the newly introduced disjointness definition which is compatible with the principle of successive removal of path matrix connections after acceptance of a new alignment. Current methods to find multiple best sequence alignments are based in this principle. The extensions necessary to deal with circular sequences or profiles may not be described in the extant literature, but are conceptually straightforward.

The basic profile structure we propose is most closely related to its predecessor, the profiles introduced by Gribskov, McLachlan and Eisenberg (1987), and to the kind of Hidden Markov models described in Krogh et al. (1994). A brief survey of the shared properties and differences vis-a-vis to these descriptors suggests itself.

Table 3. Src homology domain SH3

Alphabet: {*A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y*} Toplogy: *linear*

Constant scores: B^e=0 B^i=0 E^e=0 E^i=0 $I(A)$ ... $I(Y)$=0 $I(*)$=0 $M(*)$=0 D=0 $T_{B\to M}$=0
$T_{B\to I}$=* $T_{B\to D}$=* $T_{B\to E}$=* $T_{M\to M}$=0 $T_{M\to E}$=0 $T_{I\to M}$=0 $T_{I\to D}$=* $T_{I\to E}$=* $T_{D\to M}$=0 $T_{D\to I}$=* $T_{D\to E}$=*

Pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53
$M(A)$	-1	-1	2	-2	-3	0	-4	-1	1	0	0	1	0	0	-2	1	-2	-1	-1	2	1	0	-2	0	0	-1	0	-1	1	0	1	1	-7	-6	-1	1	-1	0	0	0	0	0	-1	1	-4	-2	1	1	1	-4	0	0	0
$M(C)$	-3	-4	-2	-6	-1	-5	-2	-5	-3	-4	-4	-4	-4	-6	-6	-3	-5	-5	-6	-4	-6	-4	-5	-4	-3	-5	-5	-5	-3	0	-4	-5	-9	-8	-6	-2	-5	0	-3	-4	-4	-5	-5	-4	-4	-4	-3	-3	-4	-1	-3	-5	-4
$M(D)$	-2	-2	0	-4	-5	3	-5	1	0	0	1	1	2	3	-4	1	-5	0	0	1	4	-1	-4	0	-2	-2	2	0	1	0	0	2	-8	-7	0	0	0	0	-1	0	0	1	-1	0	-4	-4	-1	0	2	-5	-2	2	-1
$M(E)$	-3	-2	0	-3	-5	2	-5	1	0	0	1	1	2	3	-4	1	-5	0	1	0	3	-1	-4	-1	-3	-3	2	0	0	0	0	2	-8	-7	0	-1	0	0	-1	0	0	1	-1	0	-4	-5	-1	0	1	-5	-4	2	-1
$M(F)$	2	-2	-4	2	7	-5	8	-3	-4	-4	-5	-5	-4	-5	1	-4	5	-2	-2	-6	-6	-2	2	-2	-1	0	-5	-1	-4	0	-4	-6	0	0	-3	-5	-2	0	-3	-2	-4	-4	-4	-5	2	5	-5	-3	-4	8	-2	-5	-2
$M(G)$	-3	-2	1	-4	-5	0	-6	-1	0	-1	0	0	0	0	-4	0	-4	-2	-2	5	1	-2	-4	-2	-2	-2	0	-1	0	0	1	2	-8	-7	-2	3	-2	0	0	-1	2	-1	-2	4	-5	-4	-1	1	0	-6	-1	0	-2
$M(H)$	-1	0	-1	-1	0	0	-1	0	-1	0	1	0	1	0	-3	0	-2	0	0	-2	0	-1	-3	0	-2	-2	1	0	0	0	-1	0	-3	-3	0	-2	1	0	-2	-1	-1	1	1	-2	-2	-2	0	-1	1	0	-3	0	0
$M(I)$	1	1	-1	2	-1	-2	-1	-2	-1	-1	-2	-2	-2	-2	3	-2	2	-1	-2	-3	-3	1	3	0	4	2	-2	-2	-1	0	-1	-3	-6	-4	-2	-1	-1	0	0	-1	-2	-2	-1	-3	-2	3	-2	-1	-1	-1	4	-2	-1
$M(K)$	-2	0	-1	-3	-4	0	-2	1	0	1	1	-1	0	0	-3	0	-1	1	2	-2	0	0	-3	0	-2	-2	1	2	0	0	-1	-1	-4	-4	2	-2	0	0	-1	1	0	2	1	-2	-3	-1	-1	0	1	-4	-3	1	-1
$M(L)$	0	-1	-3	5	-1	-3	0	-3	-3	-3	-3	-3	-4	-3	5	-3	4	-1	-2	-5	-4	0	5	-1	2	3	-4	-3	-3	0	-2	-4	-2	-1	-3	-3	-1	0	-2	-2	-4	-3	-2	-4	-1	2	-3	-3	-3	0	2	-3	-2
$M(M)$	-1	0	-1	3	-2	-2	-1	-1	-1	-1	-2	-2	-2	-2	4	-2	2	0	0	-3	-3	1	3	0	2	2	-2	-1	-1	0	-1	-3	-5	-3	-1	-2	-1	0	0	-1	-2	-1	-1	-3	-2	1	-3	-2	-2	-2	2	-2	-1
$M(N)$	-1	-1	0	-3	-3	1	-2	1	0	0	1	1	1	1	-3	1	-3	0	1	0	2	0	-3	0	-2	-2	2	1	1	0	0	1	-5	-5	0	0	0	0	0	1	0	1	0	0	-3	-3	0	0	2	-3	-2	1	0
$M(P)$	-3	-1	1	-3	-4	0	-5	-1	0	0	0	1	-1	-1	-3	0	-4	-1	0	0	-1	-1	-3	-1	-2	-2	-1	-1	0	0	0	-1	-7	-6	-1	0	-1	0	-1	-1	-1	-1	-1	-1	-5	-3	7	0	0	-6	-1	0	1
$M(Q)$	-2	-1	0	-2	-4	1	-4	1	0	1	1	1	1	2	-2	0	-3	1	1	-1	2	0	-2	0	-2	-1	2	0	0	0	0	0	1	-6	-5	1	-2	1	0	-1	0	-1	2	1	-1	-4	-4	0	0	0	-5	-2	2	0
$M(R)$	-3	0	-1	-3	-4	-1	-4	1	0	1	1	-1	-1	-2	-3	0	-4	0	2	-3	-1	0	-3	0	-3	-2	0	1	0	0	-2	-2	3	1	1	-3	0	0	-1	0	-1	1	2	-3	-2	-3	0	-1	0	-4	-3	0	-1
$M(S)$	-1	-1	1	-3	-3	0	-3	0	1	0	0	0	0	0	-3	1	-3	-1	0	1	0	-1	-3	0	-1	-1	0	0	1	0	0	0	-3	-3	0	1	-1	0	0	0	0	0	0	1	-2	-3	1	1	1	-3	-1	0	0
$M(T)$	-1	0	1	-2	-3	0	-3	0	1	0	0	0	0	0	-2	1	-2	0	0	0	0	0	-1	0	0	-1	0	0	1	0	0	0	-6	-5	0	0	-1	0	1	1	0	0	-1	0	-3	-2	0	1	1	-3	0	0	0
$M(V)$	0	0	0	1	-3	-1	-3	-2	-1	-2	-2	-2	-3	-3	2	-2	1	-1	-3	-2	-3	1	3	-1	4	2	-3	-2	-1	0	-1	-3	-7	-5	-3	0	-2	0	0	-1	-2	-2	-2	-2	-3	2	-1	-1	-1	-4	4	-3	-2
$M(W)$	-4	-3	-6	-3	-1	-7	-1	-5	-6	-4	-4	-6	-6	-7	-3	-5	-2	-5	-5	-8	-8	-5	-4	-5	-6	-4	-5	-5	-5	0	-6	-8	21	14	-4	-8	-4	0	-6	-5	-6	-5	-2	-8	3	-3	-7	-5	-5	0	-7	-6	-5
$M(Y)$	1	-3	-4	0	9	-5	8	-4	-5	-4	-4	-5	-3	-4	-2	-4	2	-3	-5	-6	-5	-4	0	-2	-3	-2	-4	-3	-4	0	-4	-5	0	0	-4	-5	-2	0	-4	-3	-5	-4	-4	-5	3	2	-6	-4	-3	10	-4	-5	-2
$T_{M\to I}$	*	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	0	-7	-7	-7	-7	-7	-7	-7	0	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	*
$T_{I\to I}$	*	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-2	-7	-7	-7	-7	-7	-7	-7	-4	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	*
$T_{M\to D}$	*	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	0	-7	-7	-7	-7	-7	-7	-7	0	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	*
$T_{D\to D}$	*	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-2	-7	-7	-7	-7	-7	-7	-7	-4	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	-7	*

Normalization: Mode 1: OrigScore = 0.1 × *raw score* (*priority*=1)

Mode 2: $Z\text{-}score = \dfrac{raw\ score\ /\ \{31.7 \times [1 - \exp(-0.0039 \times SeqLength - 0.79)]\ \} - 1.007}{0.199}$ (*priority*=2)

Cut-off value: OrigScore ≥ 6.3

Disjoint definition: Two profile-sequence alignments are disjoint if and only if the sequence segments matched with profile positions 1 to 53 do not overlap.

* represents a prohibitive profile score.

Relative to its predecessor, the new profile structure constitutes a generalization. For practical applications, the most useful extension is the option to search for multiple matches in the same sequence. This feature is contained in the parameterized disjointness definitions introduced among the accessories. Another important novelty is the introduction of an additional class of profile positions, namely the insert position. What may at first appear like an unnecessary complication, amounts in fact to a clarification. In the old format, it was not obvious whether gap opening and insert extension penalties apply to insertions and deletions initiated before or after the match position defined on the same line. The newly introduced profile scores allow differential scoring of insertions and deletions, symmetric representation of gaps by a combination of opening and closing penalties, and more flexible control over local and global alignment modes. Gap closing penalties are useful in preventing a deletion from running into a highly conserved block area. This was a relatively frequent complication arising with the old profile format, especially when low gap extension penalties were used. The residue-specific insert extension scores will be used to target a particular amino acid composition in a linker region. The practical relevance of these new feature is intuitively obvious but remains to be demonstrated by biological applications.

The analogy of the new profile structure with hidden Markov models (HMM) is even more striking. Virtually all profile scores can be mapped to parameters of the HMM models used by Krogh et al. (1994) to characterize protein domains. (The only exceptions are the initiation and termination scores at internal positions which relate to local alignment mode.) It is assumed that these and other HMM models applied in molecular biology can accurately be reflected by the profile syntax, provided that the scores are interpreted as log probabilities. Some of the transition scores, apparently redundant for the purpose of alignment scoring, turn out to be useful in such an exercise, e.g. the scores for transitions between identical states such as $T_{M \to M}$. The major difference between hidden Markov models and profiles lies in the interpretation attached to their parameters. While profiles are operational motif descriptors related to search algorithms, HMM models are integral components of a complete research methodology based on mathematical theory. Their parameters represent probabilities, with certain subsets necessarily summing up to one. No such restrictions apply to profiles. The theoretical overhead restricts the application of HMM models to a particular statement of the motif definition problem. In return for this, it renders them amenable to powerful analyses which are not generally applicable to profiles. Thus an HMM model converted into a profile would loose some of its application potential. A profile alignment algorithm can emulate a Viterbi algorithm for finding the single most likely alignment, but will not support use of negative log likelihood (NLL) scores for significance evaluation in a database search.

Finally, an important limitation of the new syntax needs to be discussed. The basic profile structure can only represent primary structure sequence motifs where the probability of finding a specific residue at a given position is independent of the residues occurring at all other positions. The generalized profile syntax will therefore not capture RNA hairpin motifs playing a role in translational regulation. This deliberate restriction is due to algorithmic considerations. We wanted to keep the basic profile structure amenable to standard sequence alignment techniques. Implementation of inter-residue dependencies will be envisaged when sufficiently versatile and efficient optimal algorithms for secondary structure motifs become available.

Acknowledgements

The authors thank Roland Lüthy, Michael Gribskov, and Stephen Altschul for helpful discussion and comments during the design of the new profile syntax. Ioannis Xenarios has contributed the SH3 domain profile shown in Table 3 with minor modifications.

References

Bairoch, A. 1993. The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucl. Acids. Res.* 21: 3097-3103.

Barton, G.J.; and Sternberg, M.J.E. 1990. Flexible protein sequence patterns: a sensitive method to detect weak structural similarities. *J. Mol. Biol.* 212: 389-402.

Gribskov, M.; McLachlan, M.; and Eisenberg, D. 1987. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.* 84: 4355-4358.

Gribskov, M.; Lüthy, R.; and Eisenberg, D. 1990. Profile analysis. *Meth. Enzymol.* 183: 146-159.

Huang, X.; and Miller, W. 1991. A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.* 12: 337-357.

Krogh, A.; Brown, M.; Mian, I.S.; Sjölander, K.; Haussler, D. 1994. Hidden Markov models in computational biology. *J. Mol. Biol.* 235: 1501-1531.

Lüthy, R.; Xenarios, I.; and Bucher, P. 1994. Improving the sensitivity of the sequence profile method. *Prot. Sci.* 3: 139-146.

Mulligan, E.M.; Hawley, D.K.; Entriken, R.; and McClure, W.R. 1984. Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. *Nucleic Acids Res.* 12: 789-800.

Musacchio, A.; Gibson, T.; Lehto, V.-P.; and Saraste, M. 1992. SH3 - an abundant protein domain in search of a function. *FEBS Letters* 307: 55-61.

Pearson. W.R.; and Miller, W. 1992. Dynamic programming algorithms for biological sequence comparison. *Meth. Enzymol.* 210: 575-601.

Staden, R. 1990. Searching for patterns in protein and nucleic acid sequences. *Meth. Enzymol.* 183: 193-211.

Thompson, J.D.; Higgins, D.G.; and Gibbson, J.D. 1994. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Applic. Biosci.* 10: 19-29.

Waterman, M.S.; and Eggert, M. 1987. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. Mol. Biol.* 197: 723-728.