Benefits of a Generalized Profile Syntax for Biomolecular Sequence Motifs

Kay Hofmann and Philipp Bucher

Bioinformatics Group, Swiss Institute for Experimental Cancer Research, CH-1066 Epalinges s/Lausanne, Switzerland. (INTERNET: khofmann@isrec-sun1.unil.ch ; pbucher@isrec-sun1.unil.ch)

For the detection of distantly related members of protein families and superfamilies, the profile method as introduced by Gribskov et al. [1] is more sensitive than sequence comparison algorithms using single query sequences.
Sequence profiles, which are usually derived from multiple alignments of sequences with a known relationship, consist of tables of position-specific scores and gap-penalties. According to the profile syntax, as defined in [1], each position in a protein profile contains scores for all of the possible amino acids as well as one penalty score for opening and one for continuing a gap at the specified position. For the successful application of the profile method, it is necessary that a set of related sequences is known in advance. The enhanced sensitivity of the profile method compared to methods using single query sequences is reached by two properties:

- Positions in the alignment that are better conserved usually get higher weights in the comparison scoring
- The position-dependent scoring scheme allows the adaption of the amino acid scores to the specific requirements of that position. For example, the residues allowed at a particular position can be restricted to a small subset of amino acids, even if those show no general tendency to be substituted for each other.

Recently, several attempts have been made to further improve the sensitivity of the profile method [2,3]. It has been shown that refining the procedures to construct a profile starting from a given multiple alignment can lead to a significantly enhanced sensitivity of the profile method. In addition, several other representations for sequence domains or motifs have been employed that do not necessarily require the presence of a correct and complete multiple alignment, among these are the emerging techniques of Gibbs-sampling [4] and Hidden Markov Models [5].
The description of protein motifs and domains by short patterns has become very important since there is an extensive set of such patterns available in the PROSITE database[6]. Since the syntax of the PROSITE patterns is too restrictive to describe all relevant motifs, incorporation of profile-entries into the PROSITE has started with SwissProt release 29. For the representation of the profile entries in PROSITE, we chose a generalized profile syntax that has been designed to overcome the major restrictions of the currently used profile format. The exact syntax description of the new profile format is given in the PROSITE documentation [6], the main differences compared to the currently used format are as follows:

- The new profile format can accommodate various domain/motif description methods like Gribskov-type profiles, PROSITE patterns, Hidden Markov models, weight matrices etc.
- It consists of an alternating series of match and insert positions, thus allowing asymmetrical gap-opening and gap-closing penalties. It also allows different penalties for gaps in the sequence and gaps in the profile.
- It actively supports the evaluation of multiple matches in a single sequence by specifying 'disjointedness'-modes and parameters.
- Since the new profile format is intended as a basis for an extensive collection of profiles, all information necessary for scanning a new sequence against a database of profiles have to be stored within the profile description. This includes scaling modes and parameters, cut-off levels for various degrees of significance etc.

This poster describes and explains the syntax of the new profile format. Also shown are several examples for the usefulness of a generalized profile description by comparing the results of a profilesearch with conventional database searches using single query sequences.

[1] Gribskov, M.; McLachlan, M.; and Eisenberg, D. *Proc.Natl.Acad.Sci.USA* **84**:4355-4358 (1987).
[2] Lüthy, R.; Xenarios I.; and Bucher, P. *Prot. Sci.* **3**:139-146 (1994).
[3] Thompson, J.D.; Higgins, D.; and Gibson, T.J. *Comput. Appl. Biosci.* **10**:19-29 (1994).
[4] Lawrence, C.E.; Altschul, S.F.; Boguski, M.S.; Liu, J.S.; Neuwald, A.F.; Wooton, J.C. *Science* **262**:208-214 (1993)
[5] Krogh, A.; Brown, M.; Mian, I.S.; Sjölander, K.; and Haussler, D. *J. Mol. Biol.* **235**:1501-1531 (1994).
[6] Bairoch, A. Nucl. Acid. Res. 21:3097-3103 (1993)
(database available by anonymous FTP from expasy.hcuge.ch)

What is a profile? And why are profiles 'better' than sequences?

A profile is a two-dimensional table of position specific match- gap-, and insert-scores that is representative of a set of related sequences. A profile is a kind of 'generalised sequence', it can be aligned to a sequence by the usual dynamic programming algorithms and can be used for database searches as well.

A sequence contains only information on what residue occurs at a certain position. In order to align this sequence to another one, a substitution matrix has to be applied. This substitution matrix assigns matching scores to pairs of residues depending on their similarity. Additionally there must be scores for the situation that a residue in one sequence has no counterpart in the other sequence. In typical sequence alignment programs, one substitution matrix and two different gap-scores are applied. One of the gap scores is applied when a gap has to be introduced into one of the sequences, the other (lower) score is applied when an existing gap has to be extended by one residue. The substitution matrix as well as the gap-scores are valid for all positions of the alignment.

A profile contains much more information than a single sequence since it is based on a group of sequences. A profile contains for each position the scoring information for every possible residue that may occur in an alignment. Scoring information on gap creation and gap extension is also contained within a profile. For the alignment of a profile to a sequence, no substitution matrix is needed, the information of the matrix is usually used in the construction of a profile. The simplest profile can thus be created from one sequence and one substitution matrix: the matrix is applied to each position of the sequence and gives a score for every possible residue in the alignment.

But to exploit the advantages of a profile, more than one sequence is necessary. If a multiple alignment of sequences is used for the construction of a profile, each position contains information on what type of residue is preferred at that particular position. Since multiple alignment often contain gaps, there is also the information on what positions of the profiles are more likely to tolerate gaps. The most valuable information stored in a profile is caused by the fact that usually in multiple alignments the sequence similarity is not evenly distributed. Instead, there are some positions that are highly divergent while other positions are highly conserved. A reason for high divergence at a certain position in proteins could be that these residues are exposed to the solvent that can accommodate a broad variety of side chains. In contrast, amino acids at highly conserved positions could either be important for stabilising the tertiary structure of the protein or participate in the binding of substrates, cofactors or other subunits of the protein. This uneven distribution of the residue variability can be exploited in the process of profile construction by giving higher weights to conserved positions than to variable ones.

Generally, the main advantage of a profile compared to a sequence is based on the fact that in profiles all scores are position-specific while in sequence comparisons scoring schemes and gap penalties are invariant.

The use of patterns for similarity searching is an approach that is complementary to sequence alignments as described above. Patterns, which are like profiles based on multiple sequences, focus only on the most conserved residues of a sequence family and neglect the more divergent parts of the sequence. This approach has proved to be very successful in some cases, and has lead to the collection of large pattern libraries like PROSITE. However, the total neglection of medium and weakly conserved positions leads to a loss of information that has proved to be disadvantageous in several cases.

Therefore, we decided to improve the usefulness of PROSITE by adding profile entries where the standard pattern repertoire of PROSITE seems to be inappropriate to describe a protein family. For doing this as effectively as possible, we created a novel enhanced profile format that is particularly suited for the construction of a profile database. It also includes all necessary features for the construction of NUCSITE, the collection of nucleic acid patterns and profiles that is intended to coexist with PROSITE.

The general parameters

Topology	may be either ‘linear’ or ‘circular’.
Alphabet	arbitrary alphabets allowed, both for proteins and nucleic acids.
Disjointness	either ‘unique’ (only single match allowed) or ‘protected region’ that must not overlap in multiple matches.
Normalization	various methods are defined, the required parameters have to be present.
Cut-off	Cut-off parameters may be given either in terms of ‘raw score’ or for one or more of the above normalizations. Several Cut-off levels (corresponding to various degrees of significance) can be specified.

The position-specific profile scores

Insert Positions:

B^e	external initiation score
B^i	internal initiation score
E^e	external termination score
E^i	internal termination score

$T_{B->M}$	state transition score beginning to match
$T_{B->I}$	state transition score beginning to insert
$T_{B->D}$	state transition score beginning to deletion
$T_{B->E}$	state transition score beginning to end state
$T_{M->M}$	state transition score match to match
$T_{M->I}$	state transition score match to insert
$T_{M->D}$	state transition score match to deletion
$T_{M->E}$	state transition score match to end state
$T_{I->M}$	state transition score insert to match
$T_{I->I}$	state transition score insert to insert
$T_{I->D}$	state transition score insert to deletion
$T_{I->E}$	state transition score insert to end
$T_{D->M}$	state transition score deletion to
$T_{D->I}$	state transition score deletion to
$T_{D->D}$	state transition score deletion to deletion
$T_{D->E}$	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 Positions:

M(X)	match extension score for residue X
M(*)	match extension score for residue not contained in the alphabet
D	deletion extension score

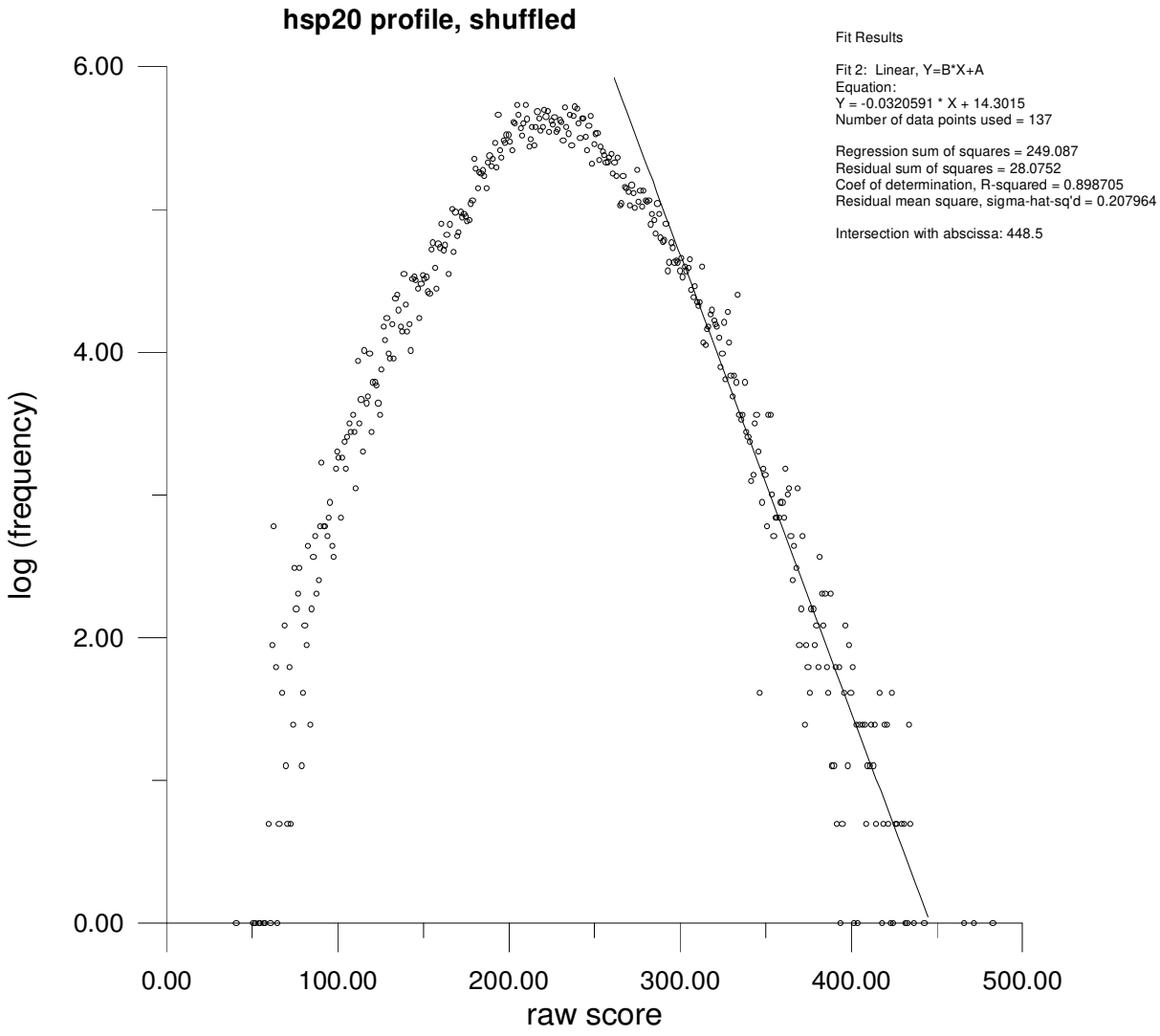
```
ID   HSP20; MATRIX.
AC   PS01031;
DT   JUN-1994 (CREATED); JUN-1994 (DATA UPDATE); JUN-1994 (INFO UPDATE).
DE   Heat shock hsp20 proteins family profile.
MA   /GENERAL_SPEC: ALPHABET='ACDEFGHIKLMNPQRSTVWY'; LENGTH=97;
MA   /DISJOINT: DEFINITION=PROTECT; N1=2; N2=96;
MA   /NORMALIZATION: MODE=1; FUNCTION=GLE_ZSCORE;
MA    R1=239.0; R2=-0.0036; R3=0.8341; R4=1.016; R5=0.169;
MA   /CUT_OFF: LEVEL=0; SCORE=400; N_SCORE=10.0; MODE=1;
MA   /DEFAULT: MI=-210; MD=-210; IM=0; DM=0; I=-20; D=-20;
MA   /M: SY='R'; M=-12,-44,-11,-13,-13,-22,-2,-7,18,-12,5,-3,-11,0,21,-6,-5,-11,-16,-34;
MA   /M: SY='D'; M=1,-41,17,16,-41,-3,3,-11,-1,-22,-12,8,-7,12,-7,0,-2,-19,-53,-36;
MA   /M: SY='D'; M=2,-37,15,13,-36,2,5,-15,-3,-26,-17,10,-6,7,-10,3,2,-17,-53,-28;
MA   /M: SY='P'; M=1,-41,6,8,-38,-4,2,-20,9,-30,-14,6,13,9,8,3,0,-22,-48,-45;
MA   /M: SY='D'; M=2,-43,23,20,-42,2,9,-18,2,-30,-18,14,-5,14,-6,2,0,-21,-57,-35;
MA   /M: SY='D'; M=4,-34,9,8,-34,6,0,-17,5,-29,-14,8,-1,5,1,5,2,-17,-47,-38;
MA   /M: SY='F'; M=-28,-32,-38,-38,50,-42,-1,2,-11,6,-6,-21,-35,-27,-27,-24,-23,-14,-3,47;
MA   /M: SY='Q'; M=0,-33,-2,-7,-26,-9,-4,1,1,-10,1,-1,-5,2,0,-2,1,0,-44,-37;
MA   /M: SY='L'; M=-13,-36,-34,-37,23,-31,-21,28,-15,29,24,-24,-25,-24,-27,-20,-10,22,-33,0;
MA   /M: SY='K'; M=-8,-32,-5,-5,-19,-16,3,-11,13,-19,-2,1,-9,2,12,-3,-3,-15,-32,-28;
MA   /M: SY='L'; M=-10,-39,-30,-32,15,-26,-20,20,-16,27,20,-21,-20,-21,-27,-17,-9,16,-32,-5;
MA   /M: SY='D'; M=3,-48,33,27,-51,4,6,-19,0,-35,-22,18,-10,13,-13,2,0,-16,-65,-41;
MA   /I: MI=-55; MD=-55; I=-5;
MA   /M: SY='V'; D=-5; M=-3,-33,-23,-32,-5,-19,-21,28,-16,26,30,-17,-14,-15,-19,-12,-1,30,-48,-28;
MA   /I: MI=-55; MD=-55; I=-5;
MA   /M: SY='P'; D=-5; M=1,-2,-1,0,-3,0,0,-1,-1,-2,-2,0,4,0,0,1,0,-1,-4,-4;

                    88 lines deleted

MA   /M: SY='L'; M=-13,-36,-29,-26,3,-30,-19,15,-20,29,26,-21,-21,-17,-22,-17,-11,10,-34,-11;
MA   /I: MI=-59; MD=-59; I=-6;
MA   /M: SY='D'; D=-6; M=0,-4,3,2,-4,0,1,-2,0,-3,-2,2,-1,1,-1,0,0,-2,-5,-3;
MA   /I: MI=-59; MD=-59; I=-6;
MA   /M: SY='S'; D=-6; M=2,-3,-3,-4,-9,-1,-5,-5,-4,-11,-6,1,-1,-6,-5,4,4,-6,-14,-6;
MA   /M: SY='E'; M=3,-37,13,14,-37,0,3,-17,6,-28,-13,9,2,11,2,4,2,-19,-49,-39;
MA   /M: SY='D'; M=2,-46,31,26,-46,4,10,-21,6,-35,-21,19,-8,14,-6,4,0,-23,-58,-35;
MA   /M: SY='G'; M=13,-34,6,2,-48,48,-21,-26,-17,-41,-28,3,-5,-12,-26,11,0,-14,-70,-52;
MA   /M: SY='V'; M=0,-25,-20,-36,-9,-16,-20,31,-16,17,21,-15,-13,-16,-19,-10,2,34,-54,-33;
MA   /M: SY='L'; M=-19,-59,-40,-34,18,-41,-21,24,-29,58,37,-29,-25,-18,-30,-28,-17,19,-19,-9;
MA   /M: SY='T'; M=4,-25,-6,-8,-18,-8,-10,0,-1,-12,-2,0,-4,-8,-8,5,14,-1,-43,-20;
MA   /M: SY='I'; M=-7,-31,-29,-37,8,-25,-22,34,-22,29,22,-21,-19,-21,-27,-17,-5,32,-43,-15;
MA   /M: SY='T'; M=7,-20,4,2,-33,0,-8,-8,1,-24,-12,5,0,-2,-6,10,14,-10,-49,-30;
MA   /M: SY='V'; M=5,-20,-14,-24,-18,-3,-21,20,-21,4,6,-11,-8,-16,-23,-4,3,24,-58,-32;
MA   /M: SY='P'; M=9,-30,-6,-2,-45,-5,2,-20,-9,-25,-20,-2,50,5,-1,8,2,-14,-55,-46;
MA   /M: SY='K'; M=-11,-52,1,-1,-1,-17,2,-18,43,-28,3,9,-10,8,33,-2,-1,-23,-33,-43;
MA   /I: MI=*; MD=*; I=0;
NR   /RELEASE=29,38303;
NR   /TOTAL=117(116); /POSITIVE=117(116); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=0(0);
CC   /TAXO-RANGE=??EP?; /MAX-REPEAT=2;
DR   P06904, CRAA_ALLMI,T; P02482, CRAA_ARTJA,T; P02474, CRAA_BALAC,T;
DR   P02470, CRAA_BOVIN,T; P02487, CRAA_BRAVA,T; P02472, CRAA_CAMDR,T;

                    38 lines deleted

DR   Q06823, SP21_STIAU,T; P12812, P40_SCHMA ,T;
DR   P30220, HS3E_XENLA,P;
DO   PDOC00791;
//
```



Results of PROFILESEARCH

score,	GLE_Zscore,	ID ,	from,	to,	Description
1468,	42.412,	"HS11_LYCES",	52,	139,	"17.8 KD CLASS I HEAT SHOCK PROTEIN."
1466,	42.521,	"HS11_WHEAT",	49,	136,	"PROTEIN) (HEAT SHOCK PROTEIN 17) (HSP 16"
1449,	41.786,	"HS12_ORYSA",	52,	139,	"17.4 KD CLASS I HEAT SHOCK PROTEIN."
1444,	41.230,	"HS16_SOYBN",	59,	146,	"18.5 KD CLASS I HEAT SHOCK PROTEIN (HSP "
1442,	42.202,	"HS11_MEDSA",	41,	128,	"18.1 KD CLASS I HEAT SHOCK PROTEIN (FRAG"
1437,	41.110,	"HS12_DAUCA",	57,	144,	"18.0 KD CLASS I HEAT SHOCK PROTEIN."
		100 lines deleted			
823,	17.180,	"P40_SCHMA",	136,	213,	"MAJOR EGG ANTIGEN (P40)."
817,	18.994,	"HS30_NEUCR",	61,	216,	"30 KD HEAT SHOCK PROTEIN."
782,	19.877,	"HS18_CLOAB",	50,	137,	"18 KD HEAT SHOCK PROTEIN (HSP 18)."
725,	18.229,	"14KD_MYCTU",	47,	126,	"14 KD ANTIGEN (16 KD ANTIGEN) (HSP 16.3)"
721,	14.306,	"P40_SCHMA",	267,	348,	"MAJOR EGG ANTIGEN (P40)."
720,	18.247,	"IBPA_ECOLI",	43,	123,	"16 KD HEAT SHOCK PROTEIN A."
624,	14.878,	"IBPB_ECOLI",	39,	109,	"16 KD HEAT SHOCK PROTEIN B."
554,	11.949,	"HS6C_DROME",	76,	134,	"HEAT SHOCK PROTEIN 67B3 (HEAT SHOCK 18 K"
498,	11.993,	"HS3A_XENLA",	47,	87,	"HEAT SHOCK PROTEIN 30A."
462,	5.426,	"DYHC_DICDI",	118,	220,	"DYNEIN HEAVY CHAIN, CYTOSOLIC (DYHC)."
432,	5.229,	"KFES_FSVGA",	163,	233,	"TYROSINE-PROTEIN KINASE TRANSFORMING PRO"
431,	5.561,	"KFES_FSVST",	54,	101,	"TYROSINE-PROTEIN KINASE TRANSFORMING PRO"
431,	4.907,	"KFES_FELCA",	397,	444,	"PROTO-ONCOGENE TYROSINE-PROTEIN KINASE F"
431,	4.905,	"KFES_HUMAN",	399,	446,	"PROTO-ONCOGENE TYROSINE-PROTEIN KINASE F"
427,	4.998,	"MS16_YEAST",	66,	119,	"ATP-DEPENDENT RNA HELICASE MSS116."
420,	7.729,	"HBPL_TRETO",	59,	126,	"HEMOGLOBIN."
416,	5.827,	"G3P1_TRIKO",	41,	110,	"GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE"
407,	4.203,	"SUWA_DROME",	749,	826,	"SUPPRESSOR OF WHITE APRICOT PROTEIN."

The inositol monophosphatase/fructose 1,6-bisphosphatase family
An example for the usefulness of the profile method

What is known about this protein family?

Inositol monophosphatase hydrolyses the phosphate group of inositol-1-monophosphate and plays an important role in the phosphatidyl-inositol second messenger pathway. It has been shown to be inhibited by lithium ions and is thought to be the main target of lithium therapy in several cerebral disorders.
The sequence of the known inositol monophosphatases shows considerable overall similarity to a few prokaryotic and eukaryotic proteins of unclear function. The 3D-structure of the homodimer has been determined by X-ray crystallography, the active site and the lithium binding site are both characterised.

Fructose 1,6-bisphosphatase hydrolyses fructose 1,6-bisphosphate to fructose 6-phosphate. The enzyme also binds lithium, the homodimeric 3D-structure of the enzyme is very closely related to the 3D-structure of the inositol monophosphatase. There is, however, no significant sequence similarity between those two enzymes. The only detectable relatives of the mammalian fructose 1,6-bisphosphatase sequence are its bacterial, fungal and plant homologues as well as sedoheptulose 1,6-bisphosphatases. Considering the 3D-structure, the amino acids participating in the lithium binding site are quite well conserved between the two sequence families, but they are dispersed over the primary structure and not useful for sequence similarity analysis. There is one short stretch of 5 contiguous amino acids involves in lithium binding that is conserved in the inositol mono-phosphatase- and the fructose 1,6-bisphosphatase families, but this stretch is too short to be used in database searches.

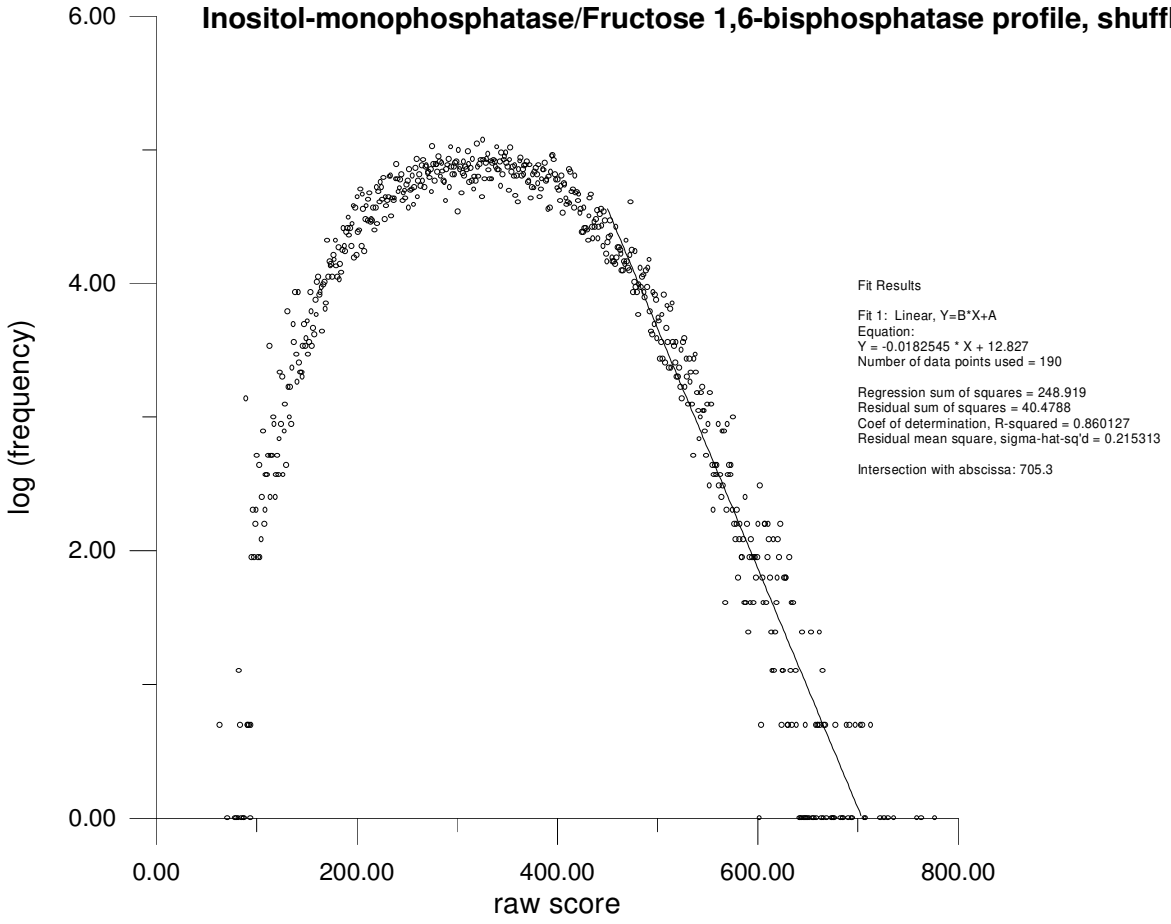
Inositol polyphosphate 1-phosphatase has not been as extensively studied as the inositol monophosphatase or the fructose 1,6-bisphosphatase but it shares some properties with these two enzymes. It has also been reported to bind lithium and shares the 5-residue pattern with the two sequence families mentioned above. It shows no detectable sequence similarity to any other protein.

What can be found by other sequence analysis methods?

BLAST finds no significant similarity between any member of two different sub-families. This holds for different substitution matrices employed, PAM40, BLOSUM62 and PAM200. The members within a subfamily are easily detected.
The same applies to Smith/Waterman type search algorithms as implemented in MPsearch and FLASH.
The short patch of amino acid similarity is not sufficient to result in a satisfactory PROSITE pattern, the current patterns (PS00629 and PS00630) find the inositol monophosphatase subfamily and inositol-polyphosphate 1-phosphatase but fail to detect the fructose-1,6-bisphosphatases. If the pattern rules would be relaxed to also include the 1,6-bisphosphatases, they would also pick up a large amount of false positives.

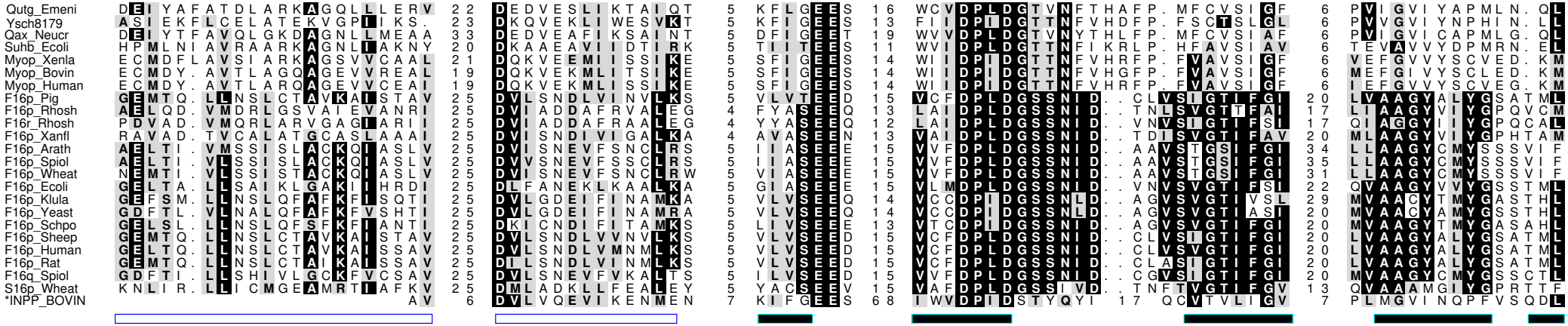
What can be found by the profile method?

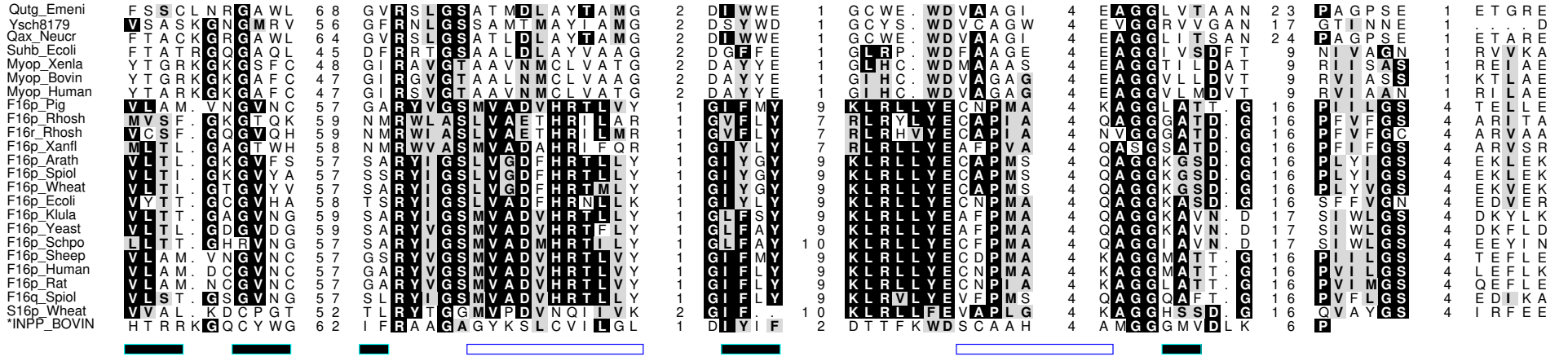
Admittedly, a profile constructed from one subfamily alone is also not able to reliably detect members of a different subfamily.
However, given the similarity of the 3D-structures of the inositol-monophosphatase and the fructose 1,6-bisphosphatase, it is possible to make a 'structural alignment' of these two enzymes, even if the sequences align only very poorly. For accomplishing that, the two 3D-structures were subjected to a rigid-body superposition algorithm (as implemented in PROMOD), and corresponding residues were determined by visual inspection. Residues that had no clear partner in the superposition were not further considered. The resulting 'structural alignment' was enlarged by automatically aligning the members of the two subfamilies to their structurally characterised representative. The resulting alignment is shown below, the regions that gave no satisfactory superposition are omitted.
Based on this alignment a profile was created, now comprising members of the inositol-monophosphatase and fructose 1,6-bisphosphatase subfamilies. When running this profile against release 29 of SwissProt, all members of the two mentioned subfamilies gave significant scores, as did the inositol polyphosphate 1-phosphatase.
This result shows for the first time a relationship between the inositol-polyphosphate 1-phosphatase and the fructose 1,6-bisphosphatase and inositol-monophosphatase families on the basis of the sequence alone. Up to now, this relationship has been proposed only because of their similar enzymatic activities and some common properties.



Results of PROFILESEARCH

```
 2898, "F16P_HUMAN",     33,    326, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2892,   "F16P_RAT",     33,    326, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2892,   "F16P_PIG",     33,    326, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2881, "F16P_SHEEP",     33,    326, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2857, "F16Q_SPIOL",     35,    329, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2794, "F16P_ARATH",    102,    411, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2787, "F16P_SPIOL",     42,    352, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2779, "F16P_SCHPO",     49,    341, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2764, "F16P_WHEAT",     95,    404, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2755, "F16P_KLULA",     41,    349, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2744, "F16P_YEAST",     42,    341, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2714, "F16P_ECOLI",     25,    324, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2655, "F16P_RHOSH",     18,    309, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2645, "F16P_XANFL",     46,    339, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2642, "F16R_RHOSH",     17,    307, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 2416, "MYOP_HUMAN",     11,    264, "INOSITOL MONOPHOSPHATASE"
 2397, "QUTG_EMENI",     11,    313, "QUTG PROTEIN"
 2393, "MYOP_XENLA",      9,    264, "INOSITOL MONOPHOSPHATASE"
 2355, "MYOP_BOVIN",     11,    264, "INOSITOL MONOPHOSPHATASE"
 2351, "SUHB_ECOLI",      4,    251, "EXTRAGENIC SUPPRESSOR PROTEIN."
 2304,  "QAX_NEUCR",     12,    325, "QA-X PROTEIN."
 1502, "F16R_ALCEU",      4,    180, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 1311, "HAL2_YEAST",      8,    327, "HALOTOLERANCE PROTEIN HAL2"
 1208, "F16P_ALCEU",     37,    133, "FRUCTOSE 1,6 BISPHOSPHATASE (FBPASE)"
 1159, "CYSQ_ECOLI",      1,    228, "CYSQ PROTEIN."
 1021, "F16P_RABIT",     33,    220, "FRUCTOSE1,6 BISPHOSPHATASE (FBPASE) (FRAGMENT)."
  880, "CYSQ_SALTY",      1,     91, "CYSQ PROTEIN (FRAGMENT)"
  819, "INPP_BOVIN",     46,    343, "INOSITOL POLYPHOSPHATE 1-PHOSPHATASE"
  790, "THER_BACST",    139,    484, "THERMOLYSIN PRECURSOR"
  787,  "CCA_ECOLI",     82,    382, "TRNA CCA-PYROPHOSPHORYLASE"
  771, "EXOQ_RHIME",    105,    314, "EXOQ PROTEIN."
  732, "GLDA_BACST",     36,    301, "GLYCEROL DEHYDROGENASE (EC 1.1.1.6)"
  708,  "IMP_BACSU",    153,    386, "INOSIN 5'-MONOPHOSPHATE DEHYDROGENASE"
  707, "G6PD_ZYMMO",    141,    360, "GLUCOSE-6-PHOSPHATE 1-DEHYDROGENASE"
  703, "BACS_HALHA",      1,    210, "SENSORY RHODOPSIN I (SR-I)"
```


Qutg_Emeni
Ysch8179
Qax_Neucr
Suhb_Ecoli
Myop_Xenla
Myop_Bovin
Myop_Human
F16p_Pig
F16p_Rhosh
F16r_Rhosh
F16p_Xanfl
F16p_Arath
F16p_Spiol
F16p_Wheat
F16p_Ecoli
F16p_Klula
F16p_Yeast
F16p_Schpo
F16p_Sheep
F16p_Human
F16p_Rat
F16q_Spiol
S16p_Wheat
*INPP_BOVIN
Helix
Strand


Qutg_Emeni
Ysch8179
Qax_Neucr
Suhb_Ecoli
Myop_Xenla
Myop_Bovin
Myop_Human
F16p_Pig
F16p_Rhosh
F16r_Rhosh
F16p_Xanfl
F16p_Arath
F16p_Spiol
F16p_Wheat
F16p_Ecoli
F16p_Klula
F16p_Yeast
F16p_Schpo
F16p_Sheep
F16p_Human
F16p_Rat
F16q_Spiol
S16p_Wheat
*INPP_BOVIN