Computer Methods for Identifying Significant Features in Protein Sequences

David Neil Perkins

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others

The University of Leeds
Department of Biochemistry and Molecular Biology
September 1994

Abstract

The research described in this thesis can be easily and conveniently separated under two broad headings, the definition of discriminating motif sets for protein families and software development. In this instance the phrase motif set refers to a combination of features in the amino acid sequences of a family of proteins that is diagnostic of family membership and therefore has predictive value in identifying new family members.

Under the first heading, a number of sets of motifs are described in detail while a number of others are included as an appendix in a format compatible with the PRINTS motif database. All these studies involved the multiple alignment of protein sequences extracted from the database and the use of database scanning techniques. From these motif sets it has been possible to identify new members of protein families and they may also supply valuable information for the exploration of the possible function and structure of the protein families.

A number of sequence analysis software packages are also described. They include both novel software and also the reworking of old algorithms with additions to make them more efficient, more useful for modern requirements and to fix existing problems. In the former category, new sequence alignment programs have been developed which integrate structural information (if any is available) with sequence and physicochemical properties. A number of programs are also discussed that allow the display and manipulation of a variety of sequence parameters, such as hydropathy and positional variability, which are very useful tools for motif definition. All these programs are written in C and the majority make use of the X/Motif programming libraries, where appropriate, and are available on a variety of different hardware platforms.

The ADSP system has also been rewritten to make it more efficient and it has been ported to the UNIX operating system to make it more accessible to a larger number of users.

Acknowledgements

I should like to thank my supervisors, Prof. A.C.T. North and Prof. J.B.C. Findlay, for allowing me to continue at my own pace. Thanks to Jasbinder Singh and Don Akrigg for the maintenance of the computer systems essential to this work and also to the MRC for providing the funding.

I am also indebted to the following people;

Dr T.K Attwood and Dr A.J. Bleasby for all the advice and encouragement they have given and the late Dr. I. Haneef for helpful discussions.

Dr A.K.H. Holzenburg for his advice and allowing me to work with him at the Lawrence Berkeley Laboratories.

All the friends and colleagues in the department who have put up with me during the past few years, particularly Drs S.S.B. Glover, E.A. Yates and A. Rashid. Also thanks to A. Rachedi for many useful discussions.

Dominic Steele for his help with the final stage of this thesis.

I am especially grateful to my parents and Dr. K.D. Parker for their help and understanding during my time in Leeds.

Table of Contents

List of Appendices

List of figures and tables

Chapter One
Introduction

Proteins are amongst the most important and diverse of all the macromolecules needed to sustain life. They perform a multitude of functions within the cell, from structural support to the catalysis of essential biochemical reactions. The majority of the information available for the study of proteins is derived from amino acid composition and biochemical studies, with only a relatively small number of three-dimensional structures being known. With recent advances in DNA sequencing techniques and the advent of the Human Genome Mapping project (and other genomic studies of various organisms such as yeast) the amount of new sequence information becoming available is sure to outstrip structural information by many orders of magnitude. Indeed even at the present time when such projects are still at fairly early stages the number of protein sequences known is very much larger than the number of three dimensional structures solved. Therefore, as the information supplied by the sequence of a protein is often the only clue to possible function and structure, sequence analysis techniques and software are very useful tools.

1.2 Databases for Molecular Biology Research

When a protein or DNA sequence has been established they are, in the majority of cases, deposited in one or more sequence databases; the most notable of these databases are discussed below.

1.2.1 NEWAT Database

This database was originally compiled as a supplement to the Atlas of Protein Sequence and Structure (Dayhoff, M.O. (1978)) and also as a resource that would be useful for examining protein relationships. The sequences were collected from literature surveys and a number of programs are available to manipulate and interrogate the data. The database is divided into six sections based on taxonomic classifications and protein function and was most notably used to demonstrate a link between platelet derived growth factor and a viral oncogene (Wheatfield, M.D. et al.

(1983)). This database was enlarged by Doolittle (1981) who also removed some of the redundant sequences.

1.2.2 NBRF/Protein Information Resource (PIR) Databases

The NBRF (Orcutt, B.C. et al. (1983)) has been collecting and collating protein sequence information for a number of years using literature searches to identify new sequences. A number of programs are also distributed with the database, the most notable being the Protein Sequence Query (PSQ) program which is used to interrogate the database. The database is split into three sections. PIR1 includes those sequences that have been classified into families based on sequence similarity and have annotated database entries, sequences in PIR2 have annotation only while PIR3 includes unverified sequence entries.

1.2.3 SWISS-PROT

Sequence data in the SWISS-PROT database is derived from three different sources, these being the NBRF/PIR database, translation of entries from the EMBL nucleotide sequence database and from literature surveys (Bairoch, A. et al. (1991)). SWISS-PROT is distinguished from other databases by the amount of annotation that is included with each entry, for instance similarities to other sequences, diseases associated with the protein, domains and sites and post-translational modifications. The annotations for each of the sequence entries are updated regularly using both information provided by literature searches and also by external experts. As redundancy of sequence information is reduced to a minimum and the annotation is of such a high quality the SWISS-PROT database is, at present, the highest priority source database for the compilation of the OWL composite database. The SWISS-PROT database is also used to produce the PROSITE pattern database described below.

1.2.4 GenBank Nucleic Acid Database

The GenBank database is a computer based collection of all the published RNA and DNA sequences along with the appropriate biological annotation (Burks, C. et al. (1985)). The sequences are entered into the databases by both direct submissions from molecular biologists and as the result of literature searches. The entries are divided into a number of taxonomical classes, for instance primates and invertebrates. The PGtrans protein sequence database is a translated version of the

GenBank database, a computer algorithm developed by Claverie and Sauvaget (1985) is used for the translation. A translated version of GenBank is also used to build the OWL composite protein sequence database.

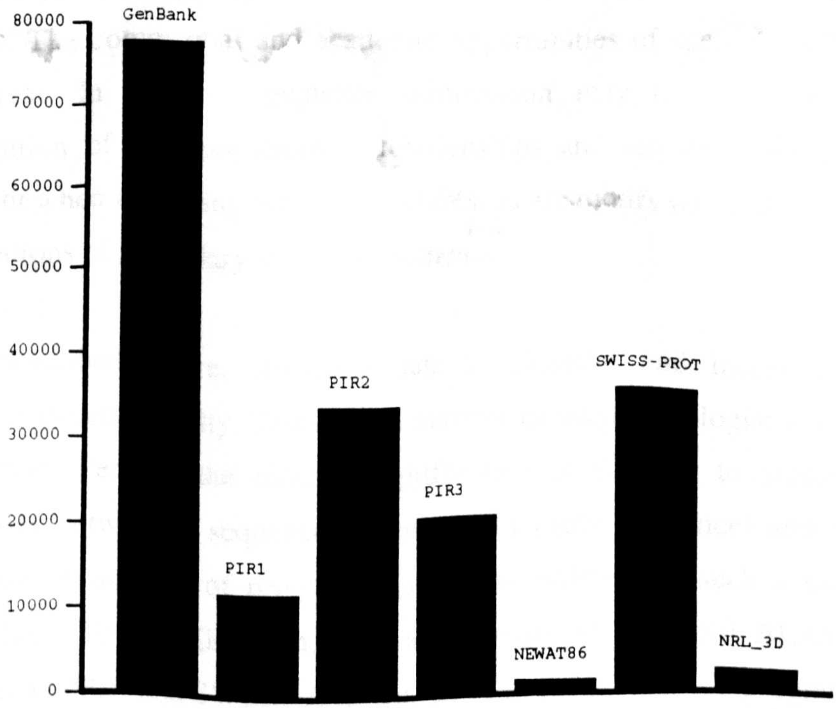
1.2.5 Protein Data Bank

This database is a computer based archival file for macromolecular structures (Bernstein, F.C et al. (1977)). Structures submitted by various research groups are entered into the database using a standard format and a number of FORTRAN programs for the manipulation of the data are also distributed along with the database. The protein sequences from the three dimensional protein structures present in this database have been collated into a sequence database known as NRL_3D (Namboodiri., K. et al. (1989)).

1.2.6 Non-Redundant Composite Databases

As sequences often appear in a number of different databases and sometimes may even appear more than once in the same database with both the protein and translated nucleic acid sequence being present, a number of non-redundant composite databases have been developed which are derived from a number of source databases. All the sequence data described in this thesis has been derived from the OWL protein sequence database (which is described in detail in the following chapter), the largest and most rigorously defined composite protein sequence database know to the author. Claverie and Bricault (1986) have also described a composite database, PseqIP, but this has fewer source databases.

Figure 1.1 illustrates the present size of the databases described above.



Database	Version
GenBank (translated)	83
PIR1	40
PIR2	40
PIR3	40
NEWAT86	1986 release
SWISS-PROT	28
NRL_3D	14

Figure 1.1 - The present size of the more notable sequence databases. The values on the y-axis refers to the number of sequence entries in each database.

1.3 Manipulation of Sequence Information

A large amount of research effort has been applied to the field of sequence analysis as the potential benefits of a useful system are enormous. For instance if a motif is devised for a particular ligand binding site this information may be applied to drug design to improve the binding abilities for this ligand. Sequence analysis may also provide an insight into the function and structure of newly sequenced proteins. The commercial and academic opportunities of such work is potentially very great. In addition, sequence information may be useful in aiding the investigation of structure-function relationships and sequence similarity is very important when modelling protein structures, as are motifs which describe particular combinations of secondary structure elements.

As described above, sequence data is plentiful and increasingly easy to determine experimentally. There are a number of ways a biologist may exploit this information. Perhaps the most straightforward of these is to search for global similarities between a sequence of interest (a probe sequence) and the database sequences. A number of programs have been written for such a task, the most notable being FASTA (Pearson, W.R. and Lipman, D.J. (1988)), BLAST (Altschul, S.F. et al. (1990)) and SWEEP (Akrigg, D. et al. (1992)). These programs are based on the initial work of Needleman and Wunsch (1970).

While such global searches may produce invaluable information, often interesting similarities may be hidden by the rest of the sequence. For instance, it is known that there are three sequence segments that are involved in the binding of GTP and GDP (Dever, T.E. et al. (1987)). These three segments have a total length of fourteen residues, therefore global searches may not indicate that a probe sequence may bind GTP as the rest of the sequence may be dissimilar, ie the signal to noise ratio is very low. A method is therefore required that can represent the important structural and functional information (henceforth referred to as sequence features) contained within the primary structure of a protein. These regions of the sequence that are characteristic of a particular protein family are know as motifs.

Also, as the growth of databases is rapid, there is an increasing need to make the data more manageable. This may be achieved using databases of conserved motifs instead of whole sequences or by clustering database sequences into broad families.

This latter process has been carried out using both classical sequence similarity calculations (Gonnet, G.H. et al. (1992)) and also by the application of novel mathematical techniques (van Heel, M. (1991)).

In both these cases, database searches with sequences of unknown structure and function are faster and generally produce less noise when comparisons are made with groups of motifs or proteins rather than individual sequence database entries as many fewer comparisons are required. For instance SWEEP takes a number of hours to compare an 'unknown' lysozyme sequence with the full OWL database, whereas just a few seconds is needed to identify the sequence as a lysozyme using a database of motifs and software written by the author. Unfortunately it will be a considerable time before motifs are available for all the known protein families, although the number of entries in both the PROSITE and PRINTS databases are increasing rapidly and database clustering techniques may relieve the problem to some extent. If the suggestion that there are only around 1000 to 2000 protein families (Chothia, C. (1992)) is found to be true, the problems are not insurmountable.

1.4 Motif Concepts

Using the scheme devised by Hodgman (1989) there are three basic types of motifs and thus three general methods for their definition and comparison with database sequences. These are described below.

1) The first of these, **sequence similarity**, is perhaps of the most relevance to this thesis. In this case the actual residue identity of each position in the motif is compared with the database sequences, when a motif has been derived from a multiple sequence alignment then usually some method is applied to take into account the frequencies of residue types at each position in a motif. More distant sequence relationships may also be detected if this technique is used in conjunction with one of the many matrices of amino acid substitutions that are available. Figure 1.2 illustrates the different methods that may be used to represent this type of motif.

```
GHVDHGKTT
```

1.2a

```
GHVDSGKST
GHVDSGKST
GAGESGKST
GHVDHGKST
GAGGVGKSA
GAGGVGKSA
GAGGVGKSC
GHIDHGKST
GHVDHGKTT
GPGGVGKSA
GHVDHGKTT
GDQSSGKSS
GRSNAGKSS
AHIDAGKTT
AHIDAGKTT
```

1.2b

```
[AG]-x(4)-G-K-[ST]
```

1.2c

All the motifs and patterns shown above describe one of the three sequence segments that have been shown to be responsible for the binding of GTP. Figure 1.2a illustrates a simple type one motif. Figure 1.2b shows a motif which consists of aligned segments from a number of proteins aligned by the author - such a motif may be described as a motif set or feature, compound features consist of two or more of these types of motif. Figure 1.2c shows the equivalent PROSITE pattern.

2) **Computer plots of amino acid properties.** Protein sequence motifs may also be defined by examining graphs of amino acid properties. For instance hydropathy plots can indicate the location of transmembrane segments and hydrophobic moment plots (Eisenberg, D. et al. (1982)) may be used to elucidate the amphiphilic nature of a sequence segment. Examples of motifs defined by the author using such methods are described in later chapters.

3) **Helical Wheels.** This method involves the projection of a protein sequence onto a representation of an alpha-helix. Using this technique it is possible to identify amphiphilic regions of a sequence, although it probably belongs more in the realm of secondary structure prediction rather than motif definition. An example of a helical wheel is shown in figure 1.3 (Donnely, D. et al. (1993)).

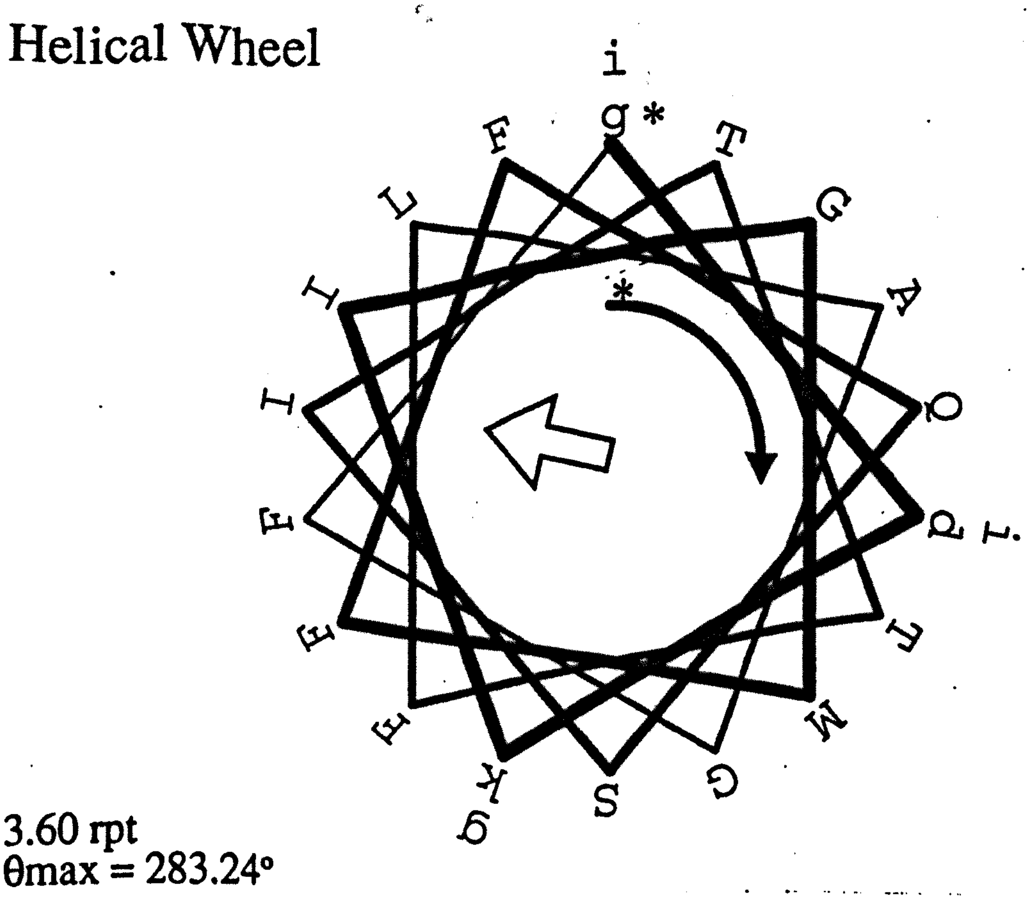


Figure 1.3 - A helical wheel of of one of the putative transmembrane segments of the human multidrug resistance protein.

1.5 Use of Sequence Identity and Similarity Motifs

As stated above, the sequence similarity approach is more relevant to the work carried out by the author. A number of software packages have been developed using this approach, some of the more notable and relevant being described in detail below.

1.5.1 LUPES (Leeds University Protein Engineering Software)

This software system takes as input motif files and then presents this information as weight matrices. The rows and columns that make up these matrices represent positions in the motif and residue frequencies respectively. Each element in the matrix thus represents the weight for a particular residue type at each location in a motif, initially this value is calculated using the motif file alone although LUPES allows a user to modify the weight interactively and even negative weights can be assigned if desired. While such manual intervention has been criticised by a number of workers because of the subjective element it introduces there may be some cases, albeit probably only a limited number, where such a technique may be useful. A program contained within the LUPES package, MEGASCAN, is then used to compare a sequence database with the weight matrix, output is produced in the form of a list of matches with the highest scoring sequences at the top of the list. Other scanning methods are also available including SPACESCAN in which the relative spacing between residue types is taken into account and NLWSCAN which uses groups of 2 and 3 residues for scanning the database.

A database of motifs derived using the LUPES system has been compiled (the Features Database) which is interrogated using the SYBIL program (Bleasby, A.J., Nicholson, R. personal communication). This has now been superseded to a large extent by the PRINTS database described below.

1.5.2 Dictionary of Sequence Motifs

Ogiwara A. et al. (1992) have devised an automatic method for the identification of conserved motifs that are exclusive to functionally related proteins. As an initial step, all the proteins within a family (the NBRF/PIR superfamily classification is used to define a family) are scanned for short motifs that are well conserved and exclusive to the group of interest. When these unique (or almost unique) motifs have been located, they are converted into peptide sentences which represent the multiple motifs and their separation (this is analogous to the concept of an ADSP Compound Feature which is described below). A consensus peptide sentence is then produced. This procedure was used to define motifs that characterise over 50% of the superfamilies within the PIR 26.0 database. The initial reliance on the NBRF/PIR superfamily classification, described above, may however limit the usefulness of this system.

1.5.3 Consensus Template Alignment

A pattern-matching procedure has been devised (Taylor, W.R. (1986)) based on fitting templates to a protein sequence, allowing certain structural constraints to be applied to the identified patterns. Templates are initially defined using an alignment of protein sequences with known structure and are further refined by adding the sequences of related proteins of unknown structure. The conserved sections of these alignments are then chosen to create templates, each position in the template is assigned a property such as a residue type or hydrophobicity. Thus a template may contain information regarding absolute amino acid identity in addition to the physicochemical properties of residues. The sequences used in the initial alignment are given as input to the SETEM program which identifies the initial templates. FITEM fits the templates to a database of sequences, those sequences that are successfully fitted are included in the initial alignment and another cycle carried out until no new sequences are identified. The final templates produced by this iterative method are known as the search templates. These search templates were shown by Taylor to identify the conserved features in known immunoglobulin and related sequences but not in other non-immunoglobulin sequences.

1.5.4 Consensus Patterns

The programs (MOTIF and PATTERN) described by Cockwell and Giles (1989) are used to compare user-defined motifs with a database of test sequences using a special method to represent motifs and are designed for the application and refinement of motifs rather than their initial definition. If more than one residue is allowed at a particular position then square brackets are used, a caret symbol (^) indicates that a residue type is not allowed at a particular position, X allows any residue type at a position while dots are used to restrict motif searches to the N or C terminus of a protein. A motif defined using this notation is illustrated below.

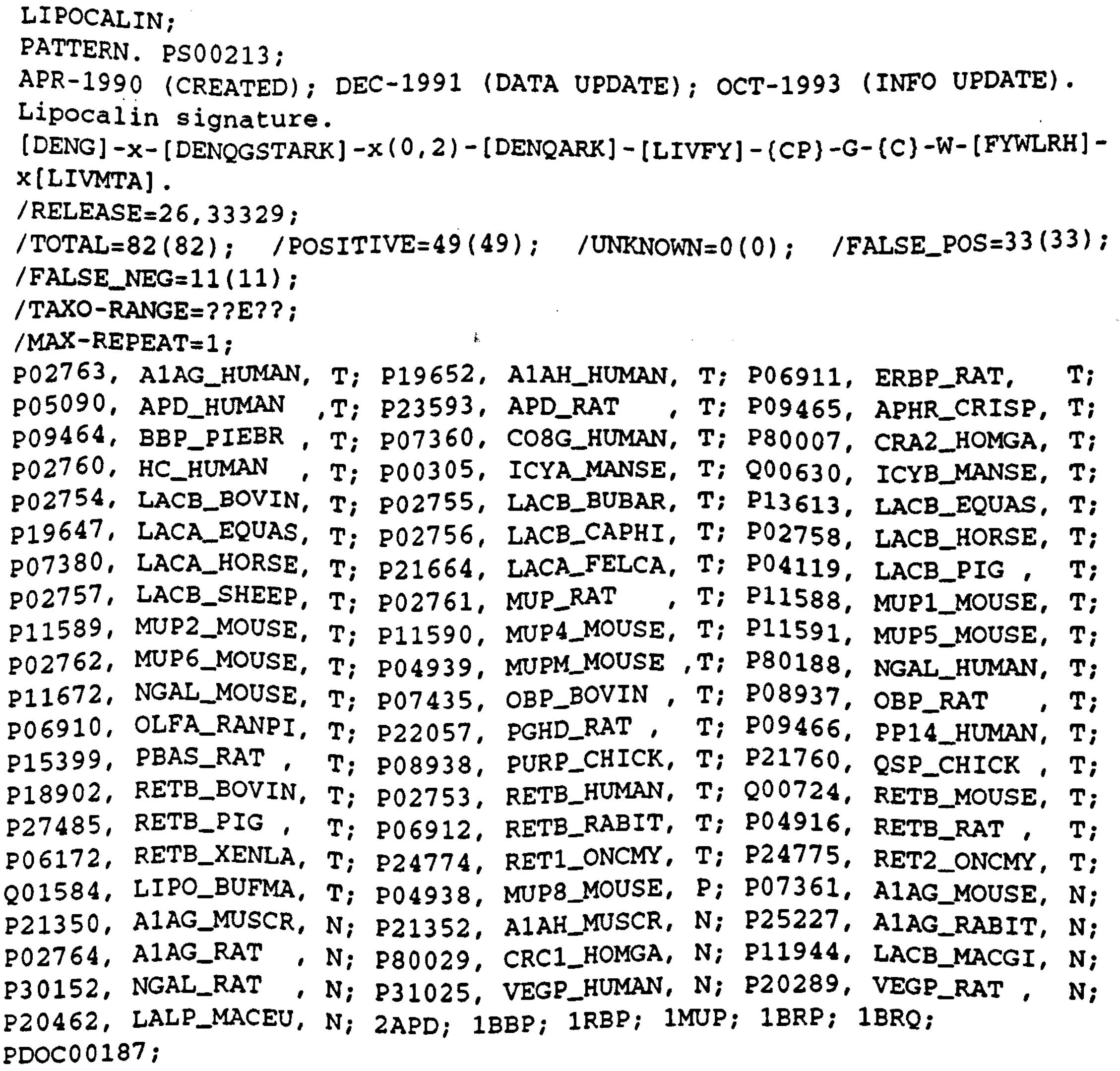
E[QN]A^S.

Thus E is the first residue, the second may be Q or N then A while the last residue must not be S. The dot at the end of the motif indicates that searching should be confined to the C-terminus of a protein. The authors also describe patterns which are made up of a number of motifs together with their relative spacings.

1.5.5 PROSITE

This is a very large database of motifs (or patterns) that, at the present release, contains 926 patterns (Bairoch, A.). The motifs are grouped into broad categories, for example patterns which relate to domains and enzymes, and are represented in a format similar to the one used by Cockwell and Giles as described above. Entries are derived as a result of literature searches, the motifs described are then tested using the SWISS-PROT database to see if tuning is required. If the latter is found necessary then the pattern is modified by increasing its length to make it more specific. Although the PROSITE database is widely used there are some entries which seem to this author to be of little value, for example some patterns are only three residues long so the chance of random matches is significant.

An example PROSITE entry (in this case for the lipocalin family) is shown in figure 1.4.



```
LIPOCALIN;
PATTERN. PS00213;
APR-1990 (CREATED); DEC-1991 (DATA UPDATE); OCT-1993 (INFO UPDATE).
Lipocalin signature.
[DENG]-x-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-
x[LIVMTA].
/RELEASE=26,33329;
/TOTAL=82(82);  /POSITIVE=49(49);  /UNKNOWN=0(0);  /FALSE_POS=33(33);
/FALSE_NEG=11(11);
/TAXO-RANGE=??E??;
/MAX-REPEAT=1;
P02763, A1AG_HUMAN, T; P19652, A1AH_HUMAN, T; P06911, ERBP_RAT,   T;
P05090, APD_HUMAN  ,T; P23593, APD_RAT    , T; P09465, APHR_CRISP, T;
P09464, BBP_PIEBR , T; P07360, CO8G_HUMAN, T; P80007, CRA2_HOMGA, T;
P02760, HC_HUMAN  , T; P00305, ICYA_MANSE, T; Q00630, ICYB_MANSE, T;
P02754, LACB_BOVIN, T; P02755, LACB_BUBAR, T; P13613, LACB_EQUAS, T;
P19647, LACA_EQUAS, T; P02756, LACB_CAPHI, T; P02758, LACB_HORSE, T;
P07380, LACA_HORSE, T; P21664, LACA_FELCA, T; P04119, LACB_PIG ,  T;
P02757, LACB_SHEEP, T; P02761, MUP_RAT    , T; P11588, MUP1_MOUSE, T;
P11589, MUP2_MOUSE, T; P11590, MUP4_MOUSE, T; P11591, MUP5_MOUSE, T;
P02762, MUP6_MOUSE, T; P04939, MUPM_MOUSE ,T; P80188, NGAL_HUMAN, T;
P11672, NGAL_MOUSE, T; P07435, OBP_BOVIN , T; P08937, OBP_RAT    , T;
P06910, OLFA_RANPI, T; P22057, PGHD_RAT ,  T; P09466, PP14_HUMAN, T;
P15399, PBAS_RAT ,  T; P08938, PURP_CHICK, T; P21760, QSP_CHICK , T;
P18902, RETB_BOVIN, T; P02753, RETB_HUMAN, T; Q00724, RETB_MOUSE, T;
P27485, RETB_PIG ,  T; P06912, RETB_RABIT, T; P04916, RETB_RAT ,  T;
P06172, RETB_XENLA, T; P24774, RET1_ONCMY, T; P24775, RET2_ONCMY, T;
Q01584, LIPO_BUFMA, T; P04938, MUP8_MOUSE, P; P07361, A1AG_MOUSE, N;
P21350, A1AG_MUSCR, N; P21352, A1AH_MUSCR, N; P25227, A1AG_RABIT, N;
P02764, A1AG_RAT  , N; P80029, CRC1_HOMGA, N; P11944, LACB_MACGI, N;
P30152, NGAL_RAT  , N; P31025, VEGP_HUMAN, N; P20289, VEGP_RAT ,  N;
P20462, LALP_MACEU, N; 2APD; 1BBP; 1RBP; 1MUP; 1BRP; 1BRQ;
PDOC00187;
```

Figure 1.4 - An example PROSITE entry, in this case the pattern for the lipocalin family of proteins. The actual pattern is shown towards the beginning of the entry, the codes (for example A1AG_HUMAN) relate to entries in the SWISS-PROT database. This example also shows a large number of false positives (ie proteins which match with the pattern but are not members of the lipocalin family), indicating that this particular pattern does not possess a significant degree of discriminating ability.

1.5.6 Profile Analysis

The authors (Gribskov, G. et al. (1987)) describe a system designed to detect distantly related proteins using a position specific scoring table which they refer to as a profile. An alignment of sequences is initially prepared using structural information (if any is available) which is then used, along with the Dayhoff mutational data matrix, to construct a profile based on both residue identity and their relative substitution values. Gap penalties may also be applied to the profile if desired. The profile is then compared with test sequences using a modified form of the dynamic programming algorithm (The dynamic programming algorithm is a recursive procedure that attempts to produce the best alignment possible between two sequences). The authors have demonstrated the efficiency of their programs using the globin fold as an example, although in situations where no structural information is available or sequence similarity is low the technique may be of less use.

1.5.7 Primary Sequence Patterns from Sets of Related Protein Sequences

This method (Smith, R.F. and Smith, T.F. (1990)) involves calculating the pairwise similarity of a set of sequences to generate a tree (dendrogram). This tree is then decomposed by replacing the node connecting the two most similar termini until only a single common pattern remains. A pattern is produced at each node by applying the dynamic programming algorithm to align the pair of sequences or patterns connected by each node. The authors have used this technique to produce a library of patterns for homologous protein families in the NBRF/PIR database.

1.5.8 Flexible Patterns

This technique derives patterns from a multiple alignment of sequences, each pattern contains information regarding conserved residues and also the number of gaps between each residue (Barton, G.F. and Sternberg, M.J.E. (1990)). The dynamic programming algorithm is then used to align these patterns with test sequences. The authors have demonstrated that a pattern derived from an alignment of seven globins was able to discriminate for all the the globins in the NBRF/PIR database.

1.5.9 SCRUTINEER

Scrutineer (Sibbald, P.R. and Argos, P. (1990)) is an interactive package that is designed to search for motifs in the SWISS-PROT and SeqDb databases. It has the capability to search for strings of amino acids with a number of possible identities in each position, variable length motifs and can take into account the physicochemical properties associated with amino acids. In addition, Scrutineer may also be used to search databases with aligned motifs. A number of these scanning methods may also be combined in one search but, in contrast to the ADSP system described below, Scrutineer is only really a useful tool when motifs have already been defined.

1.5.10 ADSP

This system is the most relevant to this thesis as all the discriminating motif sets that are described by the author have been defined using the ADSP algorithms (Attwood, T.K. and Parry-Smith, D.J. (1992)). In addition most of the software written by the author has been written with the intention of extending and interfacing to the algorithms of ADSP. The system incorporates a powerful method for characterising and predicting the occurrence of protein families and sub-families. It is also entirely objective as sequence information alone is used for the definition of motifs, pre-existing structural and functional information is not required in contrast to some of the other methods described above. A good sequence alignment is needed initially, from this alignment conserved motifs are identified and written to files. These files are then used to scan a protein sequence database iteratively. The motifs defined for a protein family are known collectively as compound features. A more detailed description of the implementation and application of the ADSP algorithms is given in the following chapter.

A large number of motif sets have been defined using the ADSP system, many of which have been incorporated into the PRINTS database. This database not only contains the relevant motifs but also includes a large amount of other information such as references and commentaries on each entry. The PRINTS database is interrogated using SMITE (Bleasby, A.J. personal communication) and also many of the programs written by the author offer powerful interfaces to PRINTS.

1.6 Secondary Structure Prediction

The fact that proteins may spontaneously renature indicates that all the information required for folding is also contained within the amino acid sequence, although in a few cases specialised enzymes known as chaperonins have been shown to be involved in this process. The methods described above may be used to devise sequence motifs that describe particular structural conformations but, in addition, there is also a separate branch of sequence analysis that attempts to deal with structure prediction. Although not of direct relevance to most of the work described in this thesis, three of the most widely used techniques are described briefly below for the sake of completeness.

1.6.1 Chou-Fasman

This technique (Chou, P.Y. and Fasman, G.D. (1974)) was originally designed to be used without access to a computer, although many computer based applications are now available. The method calculates a moving average of values that indicate the propensity of a residue to adopt one of three conformations, ie alpha helix, beta strand or turn. The values used are initially calculated from the observed frequencies of a given residue type to be found in a particular secondary structure. Normalisation is then carried out by calculating the frequency of occurrence by chance. Various rules designed by the authors are then used to attempt to define the exact ends of the secondary structure elements. These rules appear to be rather arbitrary and are perhaps the major drawback of the method.

1.6.2 Garnier-Osguthorpe-Robson

The method of Garnier et al. (1978) is more sophisticated than that described above in that its background lies in the application of information theory, despite this the method is also easier to code for a computer. The algorithm described by the authors involves calculating the secondary structure propensities by taking into account the eight residues preceding and following the residue of interest, a window length of 17 is thus used. The authors also describe the use of Decision Constants to improve the accuracy of prediction for proteins that are composed of almost entirely one sort of secondary structure. This method is probably the most widely employed of the available secondary structure prediction methods and it's efficiency may also be increased by using alignments instead of single sequences.

1.6.3 Pattern Recognition Methods

In addition to the above widely used algorithms which are based on the observed frequencies of the occurrence of a particular residue type in each secondary structure conformation, a number of secondary structure prediction techniques are available that use pattern searching methods such as that described by Lim (1974). This method searches for local hydrophobicity patterns which correspond to those expected with secondary structure elements of an amphiphillic nature. For instance, such an alpha helix could be expected to have a hydrophobic or hydrophillic residue approximately every 3.5 residues. The problem with such techniques is that they rely on the secondary structure elements to be amphiphillic, which may not always be the case.

1.7 Further Applications of sequence analysis

Another area of sequence analysis that has become increasingly important is the use of sequence alignment programs and motifs to define probes for use with DNA libraries when attempting to isolate the nucleic acid sequences of similar proteins. Using such techniques, the author has been involved with designing probes to isolate and sequence the lipoxygenase gene from Tomatoes. Thus not only does the field of sequence analysis offer an invaluable insight into the study and exploitation of proteins, it may also be used to increase the amount of sequence data available.

1.8 Conclusion

A number of important conclusions can be drawn from the above review of sequence analysis procedures:

1) There is an abundance of sequence information that threatens to overwhelm both users and the algorithms that manipulate this data, therefore software for the definition of motifs is a very important tool for making sequence information more manageable.

2) There is a relative shortage of structural information, sequence data is often the only means of deducing the possible structure and function of a protein. Also in those cases when structural information is available as much data should be extracted as possible, the VISTAS program described in a later chapter is designed to facilitate this by integrating structural and sequence data.

The first section of this thesis is concerned with the detailed description of a number of motifs defined by the author, a number of others that have been entered in the PRINTS database are also included as an appendix. Chapter three describes the use of motifs that give clues to the possible ligand binding properties of some members of a protein family while chapter four illustrates the use of motifs to identify new members of a family. A number of other motifs defined by the author are shown in appendix C. Later chapters will describe software written by the author with the specific intention to produce user-friendly, yet powerful, tools for sequence analysis.

1.9 References

Akrigg, D., Attwood, T.K, Bleasby, A.J., Findlay, J.B.C., North, A.C.T., Maughan, N.A., Parry-Smith, D.J, Perkins, D.N. SERPENT - an information storage and analysis resource for protein sequences. CABIOS **8** (1992) pp295-296

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., Basic local alignment search tool. J. Mol. Biol. **215** (1990) pp403-410

Attwood, T.K., Parry-Smith, D.J., ADSP - a new package for computational sequence analysis. CABIOS **8** (1992) pp451-459

Bairoch, A., PROSITE: a Dictionary of Protein Sites and Patterns. University of Geneva.

Bairoch, A., Boeckmann, B., The SWISS-PROT protein sequence data bank. Nucleic Acids Res. **19** (1991) pp2247-2249

Barton, G.J, Sternberg, M.J.E., Flexible Protein Sequence Patterns; A Sensitive Method to Detect Weak Structural Similarities. J. Mol. Biol. **212** (1990) pp389-402

Berstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M., The protein databank: A computer-based archival file for macromolecular structures. J. Mol. Biol. **112** (1977) pp535-542

Burks, C., Fickett, J.W., Goad, W.B., Minoru, K., Lewitter, F.I, Rindomw, W.P, Swindell, C.D., Tung, C., Bilosky, H.S., The GenBank nucleic acid sequence database. CABIOS **1** (1985) pp225-233

Chothia, C., One thousand families for the molecular biologist. Nature **347** (1992) pp543-544

Chou, P.Y., Fasman, G.D., Prediction of Protein Conformations. Biochemistry **13** (1974) pp212-245

Claverie, J.M., Bricault, L., PseqIP: a nonredundant and exhaustive protein sequence data bank from four major existing collections. Proteins **1** (1986) pp60-65

Claverie, J.M, Sauvaget, I., A new protein sequence data bank. Nature **318** (1985) pp19

Cockwell, K.Y., Giles, I.G., Software Tools for Motif and Pattern Scanning: Program descriptions including a Universal Sequence Reading Algorithm CABIOS **5** (1989) pp227

Dayhoff, M.O (ed.), Atlas of Protein Sequence and Structure (1978), National Biomedical Research Foundation, Washington D.C., USA.

Dever, T.E, Glynias, M.J., Merrick, W.C., GTP binding domains: Three consensus sequence elements with distinct spacing. Proc. Natl. Acad. Sci. (USA) **84** (1987) pp1834-1818

Donnelly, D., Overington, J.P., Ruffle, S.V., Nugent, J.H.A., Blundell, T.L. Modeling alpha-helical transmembrane domains - the calculation and use of substitution tables for lipid-facing residues. Protein Sci. **2** (1993) pp55-70

Doolittle, R.F., Similar amino acid sequences: chance or common ancestry ? Science **214** (1981) pp149-159

Eisenberg, D., Weiss, R.M., Terwilliger, T.C., The Helical hydrophobic moment: a measure of the amphiphillicity of a helix. Nature **299** (1982) pp371-374

Garnier, J., Osguthorpe, D.J, Robson, B. Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins. J. Mol. Biol. **120** (1978) pp97-120

Gonnet, G.H., Cohen, M.A., Benner, S.A, Exhaustive matching of the entire protein sequence database. Science **256** (1992) pp1443-1445

Gribskov, M., McLachlan, A.D., Eisenberg, D. Profile Analysis: Detection of Distantly Related Proteins. Proc. Natl. Acad. Sci. (USA) **84** (1987) pp4355-4358

Hodgman, T.C., The Elucidation of Protein Function by Sequence Motif Analysis. CABIOS **5** (1989) pp1-13

Lim, V., Algorithms for prediction of alpha-helical and beta structural regions in globular proteins. J. Mol. Biol. **80** (1974) pp873-894

Namboodiri, K., Pattabiramam, N., Lowrey, A., Gaber, B., George, D.G., Barker, W.C., NRL_3D: A sequence structure database, PIR Newsletter **8** (1989)

Needleman, S.B., Wunsch, C.D., A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. **48** (1970) pp443-453

Ogiwara, A., Uchiyama, I., Seto, Y., Kanehisa M., Construction of a Dictionary of Sequence Motifs that Characterise Groups of Related Proteins. Protein Engineering **5** (1992) pp479-488

Orcutt, B.C., George, D.G., Dayhoff, M.O., Protein and Nucleic Acid Sequence Database Systems. Ann. Rev. Biophys. Bioeng. **12** (1983) pp419-441

Pearson, W.R., Lipman, D.J., Improved tools for biological sequence comparison. Proc. Antl. Acad. Sci. (USA) **85** (1988) pp2444-2448

Sibbald, P.R., Argos, P., Scrutineer: a computer program that flexibly seeks and describes motifs and profiles in protein sequence databases. CABIOS **6** (1990) pp279-288

Smith, R.F., Smith, T.F., Automatic Generation of Primary Sequence Patterns from Sets of Related Protein Sequences. Proc. Natl. Acad. Sci. (USA) **87** (1990) pp118-122

Taylor, W.R., Identification of Protein Sequence Homology by Consensus Template Alignment. J. Mol. Biol. **188** (1986) pp233-258

Van Heel, M, A new family of powerful multivariate statistical sequence analysis methods. J. Mol. Biol. **220** (1991) pp877-887

Wheatfield, M.D., Scrace, G.T, Whittle, N., Stroobant, P., Johnsson, A., Wasteson, A., Westermark, B., Heldin, C.H., Huang, J.S., Deuel, T.F., Platelet-derived growth-factor is struturally similar to the putative transforming protein P28sis of Simian sarcoma virus. Nature **304** (1983) pp35-39

Chapter Two

Materials and Methods used for Motif Definition

The initial section of this thesis will describe the definition of a number of sets of discriminating motifs. All of these studies used the same methodology and algorithms, which are described below. Most of the programs used are either updated and rewritten ADSP routines or new routines, in both cases written in portable C by the author. Figure 2.1 shows a flow diagram illustrating the process of motif definition using the ADSP based system, these algorithms and methods are described in more detail below.

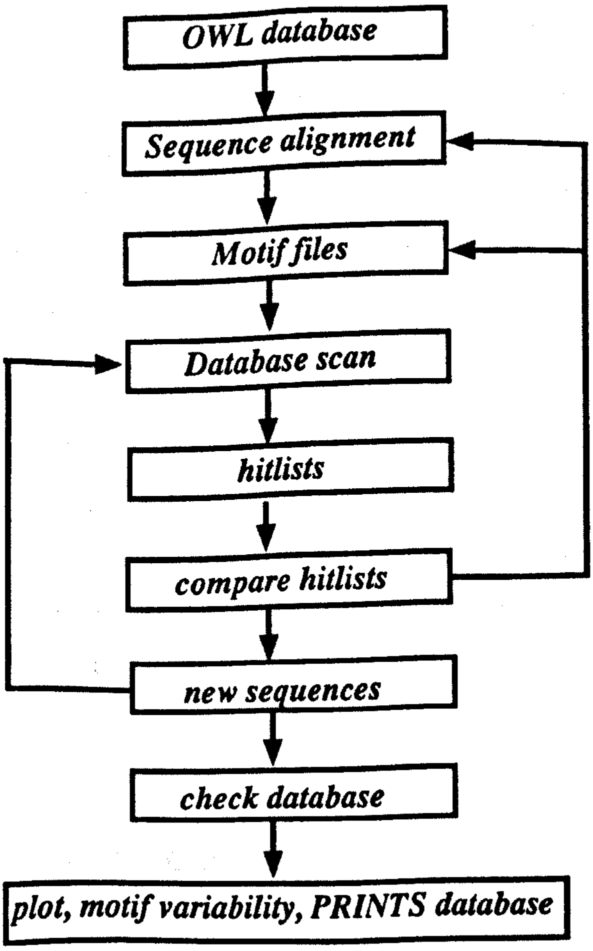


Figure 2.1 Motif definition system flow diagram

2.1 The OWL Database

The OWL database is a largely non-redundant database produced from a number of source databases (Bleasby, A.J. and Wootton, J.C. (1990)). At the moment these source databases are SWISS-PROT, NBRF/PIR, GenBank and NRL_3D. The nucleic acid entries from the GenBank database are translated using software written at Leeds before inclusion. All the sequences from the source databases are compared with each other and those sequences which are identical or have only trivial mis-matches are discarded. This process is carried out by assigning priorities to databases, the sequence from the source database with the highest priority being preferentially retained. This priority is mainly dependent on the quality and the amount of information given for each sequence entry, currently the SWISS-PROT database has the highest priority. The only redundant sequences found in the OWL database are from the NRL_3D database which is retained in its entirety to aid the interface to the Brookhaven (PDB) structural database. The OWL database is updated at regular intervals (approximately every three months) and is the largest and most up to date composite database available, the current version (23.2) containing over 83,000 sequence entries (over 26,000,000 residues). Figure 2.2 shows the rapid growth of the OWL database from version 1.1 to the present day while figure 2.3 shows the proportion of the total number of sequences derived from each source database.

The OWL database consists of a number of files. These include binary indexing files, which are used by application programs to quickly find the location of a particular sequence or text string, and also ASCII files which contain the sequence entries and entry descriptions. The sequences in the database are stored in the NBRF/PIR format, ie each entry has :-

```
>P1;DATABASE_CODE
Short description
SEQUENCE_HERE_IN_SINGLE_LETTER_CODE
*
```

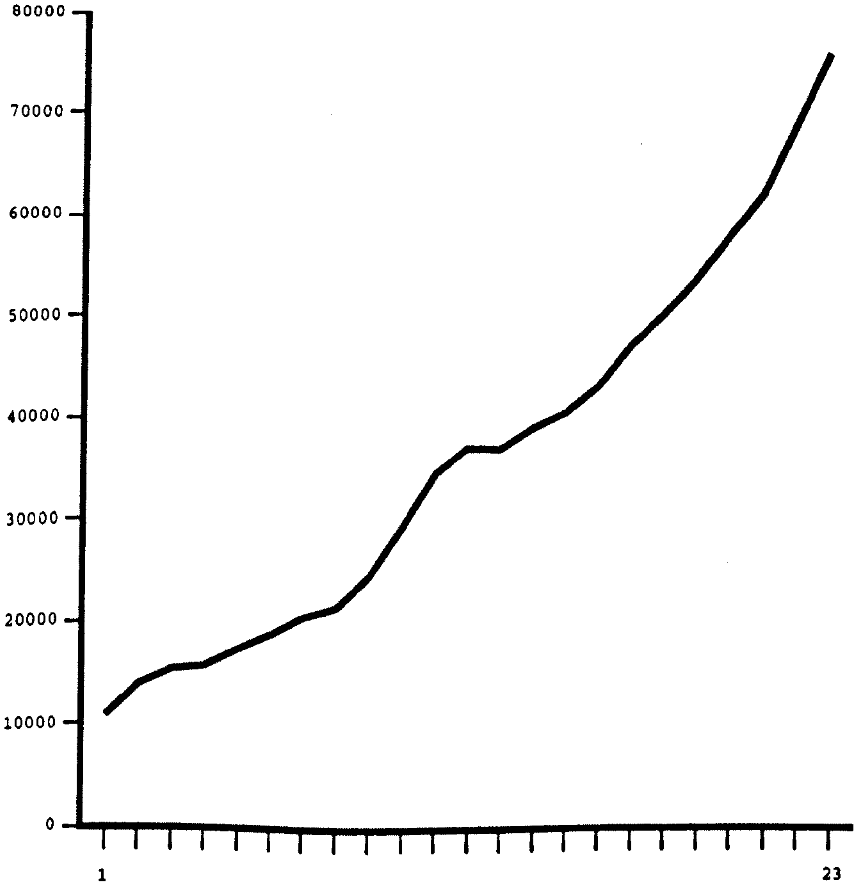
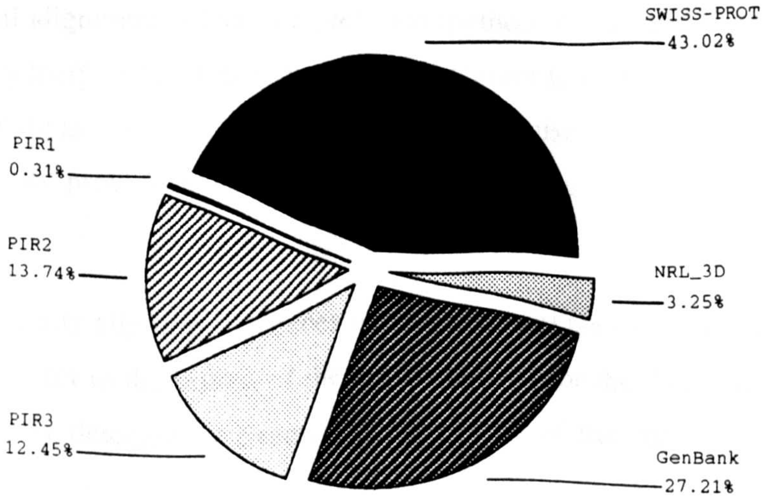


Figure 2.2 - The growth of the OWL database from version 1 (May 1988) to version 23 (March 1994). The x-axis represents the database version while the y-axis illustrates the number of sequence entries.



Database	Number of sequences	Number of residues
PIR1 (v. 40)	261	1651
PIR2 (v. 40)	11499	2677032
PIR3 (v. 40)	10424	2976511
SWISS-PROT (v. 28)	35998	12495819
NRL_3D (v. 14)	2722	484598
GenBank (v.83.0)	22768	7292381

figure 2.3 - The contribution of the source databases to OWL version 23.2

2.2 Sequence Alignment

A good, accurate alignment is essential as a first step for the definition of discriminating motifs. While automatic alignment techniques, such as CLUSTALV (Higgins, D.G. et al. (1992)), allow the production of objective alignments where the similarity of sequences is low these alignments are usually very poor and therefore manual alignment is then the preferred method. Automatic alignments are also usually very inefficient when sequences of differing lengths are used and often insert an inordinate number of gaps in attempts to optimise an alignment, although they may however provide a useful starting alignment which can be improved manually.

Although manually aligning sequences introduces a degree of subjectivity, if the alignment is incorrect in the region of the motif selected for the database scan, this inaccuracy is easily detected by examining the results of the database searches.

A number of manual alignment methods are available, some of these are reviewed in a later chapter of this thesis. Of particular relevance to sequence analysis at Leeds are SOMAP (Parry-Smith, D.J. and Attwood, T.K. (1991)), MANALIGN and also two new alignment programs written by the author (ALIGN and XALIGN) which will be described in a later section of this thesis. In the case of ALIGN and XALIGN, colour blocks are used as standard to facilitate the alignment of sequences which have low homology and also to ensure the highest possible accuracy. Colour alignments are also available from SOMAP on a limited number of character cell terminals.

A small section of an alignment of ATP synthase c subunits is shown in figure 2.4 along with the corresponding alignment coloured by residue type. This alignment, initially produced using CLUSTALV and refined manually, illustrates how much easier areas of homology may be identified using colour sequence alignments rather than simple monochrome representations. It has been suggested that these proteins have two transmembrane segments (Fragar, D. et al. (1994)), these are easily identified using the colour alignment as hydrophobic residues are coloured grey.

```
ATPL_RHORU ----DAEAAKMIGAGLAAIGMIGSGIGVGNIWANLIATVGRNPAAKST
ATPL_BACME -----------ASAIAIGLAALGAGIGNGLIVSKTIEGTARQPEARGT
ATPL_ECOLI -----------AAAVMMGLAAIGAAIGIGILGGKFLEGAARQPDLIPL
ATPH_SPIOL -----------AAGLAVGLASIGPGVGQGTAAGQAVEGIARQPEAEGK
ATPL_PROMO -------AASAVGAGAAMIAGIGPGVGQGYAAGKAVESVARQPEAKGD
ATPL_SULAC -----FEGLNIGAGLAIGLAAIGAGVAVGMAAAAGIGVLTERRD----
ATPL_BACFI -----------GAAIAAGLAAVAGAIAVAIIVKATIEGTTRQPELRGT
ATPL_VIBAL -----------AVGIIVGLASLGTAIGFALLGGKFLEGAARQPEMAPM

ATPL_RHORU VELYGWIGFAVTEAIALFALVVALILLFAA
ATPL_BACME LTSMMFVGVALVEALPIIAVVIAFMVQGK
ATPL_ECOLI LRTQFFIVMGLVDAIPMIAVGLGLYVMFAVA
ATPH_SPIOL IRGTLLLSLAFMEALTIYGLVVALALLFANPFV
ATPL_PROMO IISTMVLGQAIAESTGIYSLVIALILLYANPFVGLLG
ATPL_SULAC MFGTILIFVAIGEGIAVYGILFAVLMLFGKF
ATPL_BACFI LQTLMFIGVPLAEAVPIIAIVISLLILF
ATPL_VIBAL LQVKMFIIAGLLDAVPMIGIVIALLFTFANPFVGQLG
```

ATPL_RHORU ATP synthase c - *Rhodospirillum rubrum*
ATPL_BACME ATP synthase c - *Bacillus megaterium*
ATPL_ECOLI ATP synthase c - *Escherichia coli*
ATPH_SPIOL ATP synthase c - Spinach
ATPL_PROMO ATP synthase c - *Propionigenium modestum*
ATPL_SULAC ATP synthase c - *Sulfolobus acidocadarius*
ATPL_BACFI ATP synthase c - *Bacillus firmus*
ATPL_VIBAL ATP synthase c - *Vibrio alginolyticus*

Figure 2.4 - An alignment of ATP synthase c proteins. The equivalent colour alignment produced using the SOCOL programme (Parry-Smith, D.J. personal communication) is shown on the following page. The key to the colours used is shown in appendix D.

```
ATPL_RHORU    ----DAEAAKMIGAGLAAIGMIGSGIGVGNIWANLIATVGRNPAAKSTVELYGWIGFAVTEAIALFALVVALILLFAA
ATPL_BACME    -----------ASAIAIGLAALGAGIGNGLIVSKTIEGTARQPEARGTLTSMMFVGVALVEALPIIAVVIAFMVQGK
ATPL_ECOLI    -----------AAAVMMGLAAIGAAIGIGILGGKFLEGAARQPDLIPLLRTQFFIVMGLVDAIPMIAVGLGLYVMFAVA
ATPH$SPIOL    -----------AAGLAVGLASIGPGVGQGTAAGQAVEGIARQPEAEGKIRGTLLLSLAFMEALTIYGLVVALALLFANPFV
ATPL$PROMO    --------AASAVGAGAAMIAGIGPGVGQGYAAGKAVESVARQPEAKGDIISTMVLGQAIAESTGIYSLVIALILLYANPFVGLLG
ATPL_SULAC    -----FEGLNIGAGLAIGLAAIGAGVAVGMAAAAGIGVLTERRD-----MFGTILIFVAIGEGIAVYGILFAVLMLFGKF
ATPL_BACFI    -----------GAAIAAGLAAVAGAIAVAIIVKATIEGTTRQPELRGTLQTLMFIGVPLAEAVPIIAIVISLLILF
ATPL_VIBAL    -----------AVGIIVGLASLGTAIGFALLGGKFLEGAARQPEMAPMLQVKMFIIAGLLDAVPMIGIVIALLFTFANPFVGQLG
```

2.3 Motif Selection

After the sequence alignment has been prepared, it can be examined manually for the areas of highest conservation. This process can also be carried out by producing graphs of the positional variability of alignments using programs written by the author which will be described in a later section of this thesis. When these areas have been identified, motifs can be selected and written to a file, a typical motif file is shown in figure 2.5. As can be seen, a motif is written to the file from each sequence in the alignment. These files are then submitted to a database scanning routine.

```
% from XALIGN
Motif number 1
12
ADWVCLAQHESN
AEWICIIFHMSG
ANWVCMAEYESN
GNWVCAAKFESN
GNWVCAAKYESN
GNWVCAANYESG
GNWVCAANYESS
GNWVCAARYESN
GNWVCVAKFESN
LEWTCVLFHTSG
PEWVCTAFHTSG
PEWVCTTFHTSG
SEWICTLFHTSG
SNWVCLVENESG
*
```

Figure 2.5 - A typical motif file, in this case derived from an alignment of α-lactalbumins and c-type lysozymes.

2.4 Database Scanning

For the definition of the motifs described in the following chapters the SCAN program was used. The score for each position in a motif is calculated from the residue frequency of the original motif files and in all cases this was the single positional frequency rather than that based on pairwise separation. The scoring method is illustrated below.

If there were three sequences in an alignment then a typical motif file might contain the following motifs :-

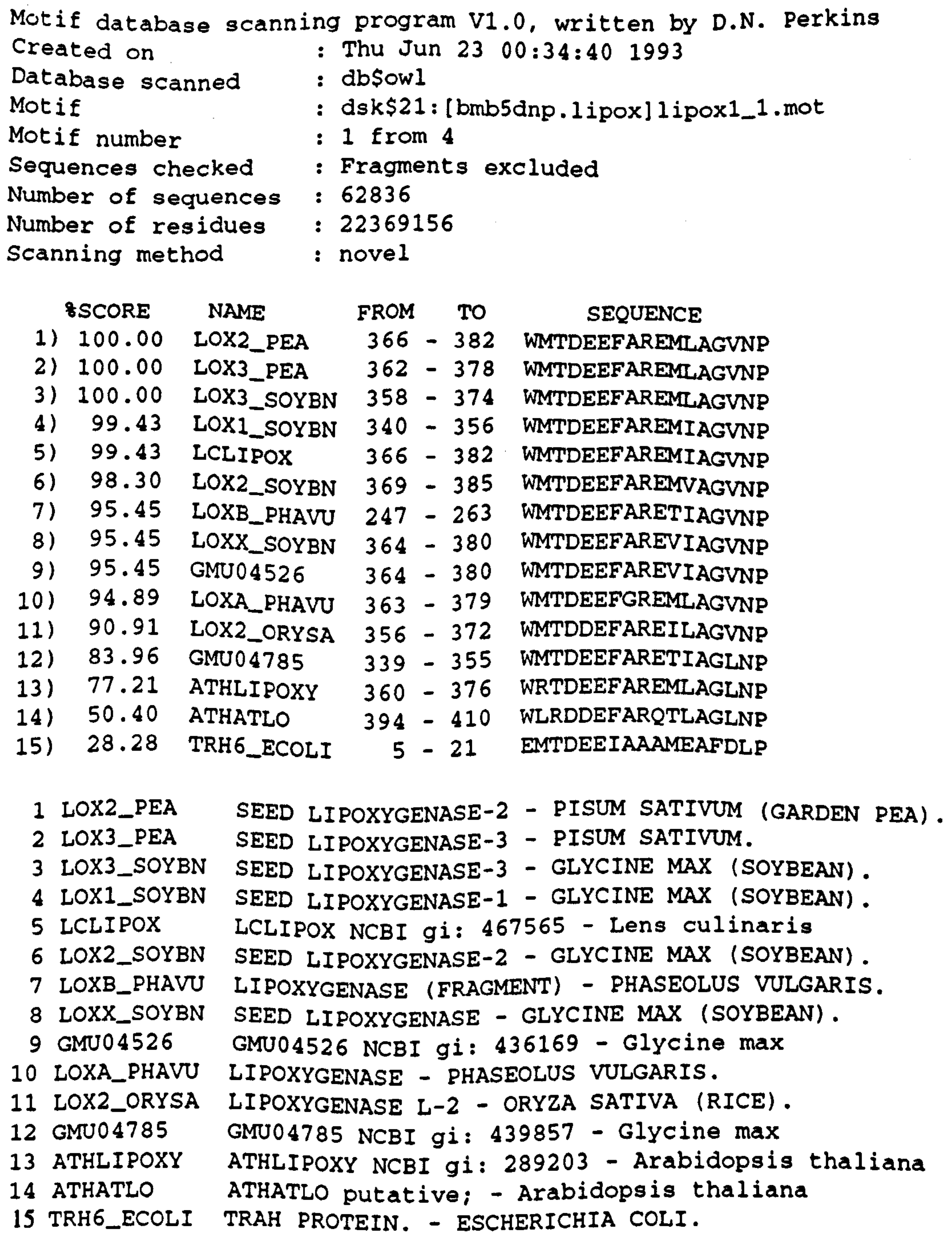
```
VFGRCELAAA
IFERCELAAI
FFERCELAII
```

This motif set is slid along a test sequence derived from the database. If the residue at position one in the test sequence is V then the score for that position is :-

Number of times V occurs in the motif file / number of motifs

In this case the score would be 33%. An arginine residue at position four would thus score 100% and so on.

The top scoring regions from all the test sequences in the database are output in the form of a hitlist, which is ordered with the highest scoring sequences in the upper regions of the file. Each entry in the hitlist (a hit) consists of the protein name, the position in the sequence where the motif matches and the score. A typical (although much shortened) hitlist is shown in figure 2.6.



```
Motif database scanning program V1.0, written by D.N. Perkins
Created on             : Thu Jun 23 00:34:40 1993
Database scanned       : db$owl
Motif                  : dsk$21:[bmb5dnp.lipox]lipox1_1.mot
Motif number           : 1 from 4
Sequences checked      : Fragments excluded
Number of sequences    : 62836
Number of residues     : 22369156
Scanning method        : novel

     %SCORE    NAME          FROM    TO          SEQUENCE
  1) 100.00   LOX2_PEA       366  -  382   WMTDEEFAREMLAGVNP
  2) 100.00   LOX3_PEA       362  -  378   WMTDEEFAREMLAGVNP
  3) 100.00   LOX3_SOYBN     358  -  374   WMTDEEFAREMLAGVNP
  4)  99.43   LOX1_SOYBN     340  -  356   WMTDEEFAREMIAGVNP
  5)  99.43   LCLIPOX        366  -  382   WMTDEEFAREMIAGVNP
  6)  98.30   LOX2_SOYBN     369  -  385   WMTDEEFAREMVAGVNP
  7)  95.45   LOXB_PHAVU     247  -  263   WMTDEEFARETIAGVNP
  8)  95.45   LOXX_SOYBN     364  -  380   WMTDEEFAREVIAGVNP
  9)  95.45   GMU04526       364  -  380   WMTDEEFAREVIAGVNP
 10)  94.89   LOXA_PHAVU     363  -  379   WMTDEEFGREMLAGVNP
 11)  90.91   LOX2_ORYSA     356  -  372   WMTDDEFAREILAGVNP
 12)  83.96   GMU04785       339  -  355   WMTDEEFARETIAGLNP
 13)  77.21   ATHLIPOXY      360  -  376   WRTDEEFAREMLAGLNP
 14)  50.40   ATHATLO        394  -  410   WLRDDEFARQTLAGLNP
 15)  28.28   TRH6_ECOLI       5  -   21   EMTDEEIAAAMEAFDLP

  1 LOX2_PEA     SEED LIPOXYGENASE-2 - PISUM SATIVUM (GARDEN PEA).
  2 LOX3_PEA     SEED LIPOXYGENASE-3 - PISUM SATIVUM.
  3 LOX3_SOYBN   SEED LIPOXYGENASE-3 - GLYCINE MAX (SOYBEAN).
  4 LOX1_SOYBN   SEED LIPOXYGENASE-1 - GLYCINE MAX (SOYBEAN).
  5 LCLIPOX      LCLIPOX NCBI gi: 467565 - Lens culinaris
  6 LOX2_SOYBN   SEED LIPOXYGENASE-2 - GLYCINE MAX (SOYBEAN).
  7 LOXB_PHAVU   LIPOXYGENASE (FRAGMENT) - PHASEOLUS VULGARIS.
  8 LOXX_SOYBN   SEED LIPOXYGENASE - GLYCINE MAX (SOYBEAN).
  9 GMU04526     GMU04526 NCBI gi: 436169 - Glycine max
 10 LOXA_PHAVU   LIPOXYGENASE - PHASEOLUS VULGARIS.
 11 LOX2_ORYSA   LIPOXYGENASE L-2 - ORYZA SATIVA (RICE).
 12 GMU04785     GMU04785 NCBI gi: 439857 - Glycine max
 13 ATHLIPOXY    ATHLIPOXY NCBI gi: 289203 - Arabidopsis thaliana
 14 ATHATLO      ATHATLO putative; - Arabidopsis thaliana
 15 TRH6_ECOLI   TRAH PROTEIN. - ESCHERICHIA COLI.
```

Figure 2.6 - A shortened hitlist produced by the SCAN program

SCAN also includes a system of modified scanning which produces a hitlist with a greater portion of true positive hits in the upper parts of each list. If a residue in the test sequence matches with any residue in the motif set, then a counter is incremented by one. If there is no match the counter retains its value. This counter value is then multiplied by the total score for the entire motif.

For instance with the following motif set :-

	V	F	G	R	C	E	L	A	A	A
	I	F	E	K	C	E	I	A	A	I
	F	F	E	K	C	E	I	V	I	I
Highest attainable score at position	.33	1	.66	.66	1	1	.66	.66	.66	.66

Total is 7.33, counter value is 10

part of test sequence	V	F	D	R	C	E	L	V	A	A
	|	|		|	|	|	|	|	|	|
counter value	1	2	2	3	4	5	6	7	8	9

End counter value is therefore 9.

score for normal frequency scanning is 5.33 / 10 (53.3%)

score for modified scanning technique is (5.33 x 9) / (7.33 x 10) (65.44%)

The score produced by the modified database scanning method is greater as it takes into account the number of positions matched as well as the simple residue frequency score. In the case above, a large number of residues in the test segment contribute to the score, thus this value is higher.

As the ADSP method of motif definition is iterative (described below), then a number of database searches may be needed before the final motif sets are defined. In practice, this means that at least two scanning operations are carried out for each motif set.

A new version of the SCAN program has been written by the author which allows the use of the techniques described here along with the application of substitution matrices and simple statistics. The user may select an option from this new program that outputs score frequencies allowing other statistical approaches to be applied if desired. The new SCAN program is much more portable, being successfully compiled and run on a number of platforms and is also substantially faster than the previous program, the greatest increase in speed being on Silicon Graphics platforms where the time taken for a typical search of the OWL database (version 23.0) was reduced from over eight hours to under thirty minutes.

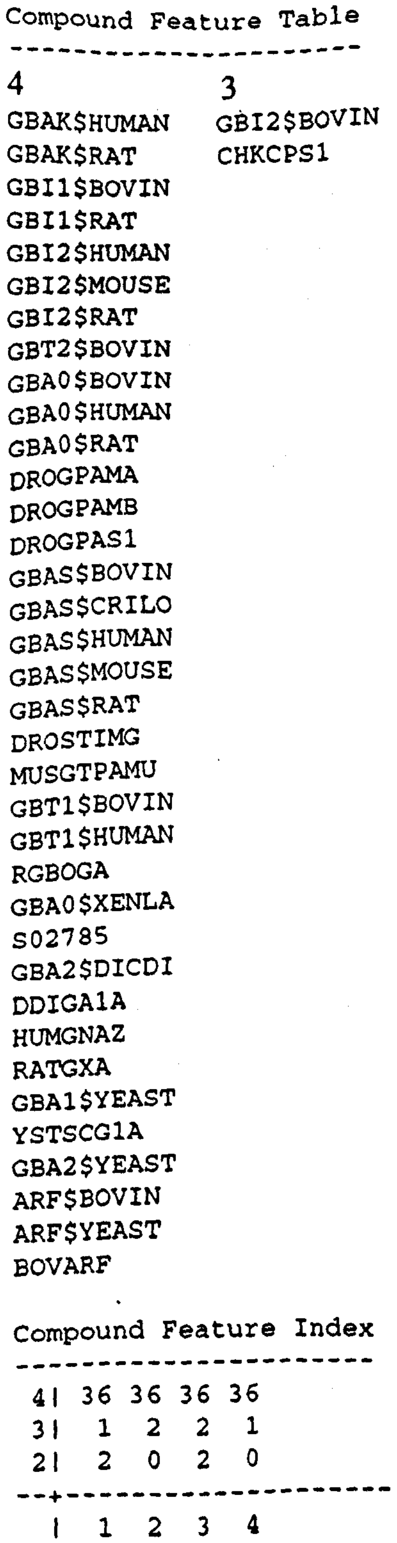
The new SCAN program also may be used to scan any nucleic acid or protein sequence database which conforms to the NBRF/PIR file format, for example GenBank.

2.5 Comparison of Hitlists

The COMPARE program is used to analyse the hitlist files produced by the database scanning programs as manual analysis of such files would be very time consuming, tedious and prone to error. Attempts had been made to port the VMS version of COMPARE (Parry-Smith, D.J (1990)) from the original ADSP system to UNIX platforms, but this was largely unsuccessful. Therefore, a complete rewrite was undertaken by the author which resulted in a much more efficient and portable program. This routine uses the dynamic memory allocation facilities provided by the C programming language to ensure the maximum efficiency of machine usage, allowing a large number of long files to be analysed, limited only by the memory of the host machine rather than fixed array bounds. The new COMPARE also allows for the manipulation of the statistical data produced by the SCAN program.

COMPARE allows a large number of hitlists (restricted only by the memory capacity of the machine used) to be checked for identical sequences codes. A file is then produced, the Compound Feature Index (or CFI), which contains a table illustrating the number of hitlists a particular sequence is found in. Another file, the LIS file, may also be produced which gives a more detailed representation of the results as the matching motif from the database sequences are also shown. Figure 2.7 illustrates a typical CFI. The sequences must be found in the correct order, ie motif one must be nearer the N-terminus than motif two, and must not be overlapping otherwise COMPARE will discard the entry from the hitlist files.

The sequences which are shown to match with all the motifs (or features) used to scan the database are considered to be the true set of hits. The motifs from this true set are written to new motif files by the COMPARE program and are then used to scan the database again to produce more hitlists. The new hitlists are again analysed with COMPARE and if any extra sequences are shown to match with all the motifs then these are added to the motif files and another database scan carried out. This process is repeated until no new sequences are evident in the true set. In this way the motifs originally selected are refined in an iterative and objective manner with no user input. The weighting value for a particular residue type is defined only by the sequence data present in the database and not by manual intervention as in a number of other sequence analysis methods such as MEGASCAN from the LUPES package.



```
Compound Feature Table
----------------------
4              3            2
GBAK$HUMAN     GBI2$BOVIN   YOR1$PVX
GBAK$RAT       CHKCPS1      PVXX3
GBI1$BOVIN
GBI1$RAT
GBI2$HUMAN
GBI2$MOUSE
GBI2$RAT
GBT2$BOVIN
GBA0$BOVIN
GBA0$HUMAN
GBA0$RAT
DROGPAMA
DROGPAMB
DROGPAS1
GBAS$BOVIN
GBAS$CRILO
GBAS$HUMAN
GBAS$MOUSE
GBAS$RAT
DROSTIMG
MUSGTPAMU
GBT1$BOVIN
GBT1$HUMAN
RGBOGA
GBA0$XENLA
S02785
GBA2$DICDI
DDIGA1A
HUMGNAZ
RATGXA
GBA1$YEAST
YSTSCG1A
GBA2$YEAST
ARF$BOVIN
ARF$YEAST
BOVARF

Compound Feature Index
----------------------
 4| 36 36 36 36
 3|  1  2  2  1
 2|  2  0  2  0
--+-------------------
  |  1  2  3  4
```

Figure 2.7 A typical Compound Feature Index (CFI) produced by the COMPARE program, in this case for four motifs derived from G protein α chains. The sequences on the far left of the table match with all motifs, those on the far right match with only two motifs. The table at the bottom of the CFI indicates how many sequences match with each motif.

The number of the motif matched is shown on the bottom line of the table (in this case 1 to 4). The column of figures to the left of the table shows the number of motifs matched while the numbers in the table itself show which motifs are matched by the sequence codes.

COMPARE also allows the user to determine whether the initial motifs are flawed and need redefining, for instance if a large number of sequences that are know members of the family being studied are shown to be only partial matches. However, database entries which are only fragments of the whole sequence will be shown to be partial matches so some vigilance is required. Also, the user should be very wary of motifs if a sequence that is shown to match with all motifs initially is then found to be only a partial match after the next database scan.

After the final database scan, a motif set with good discriminating efficiency should show that all true hits match with all the motifs in the set, no hits in the lower columns of the CFI and the only hits in the two feature column should be attributable to noise (assuming more than two motifs are used initially).

The COMPARE program additionally allows the use of distance criteria, the user supplying the maximum number of residues allowable between each motif. This option is useful for reducing the amount of noise in a hitlist but should be used with caution to ensure that true hits are not excluded.

2.6 Graphical Interpretation of the Final Motif Files.

The PLOT program has also been totally rewritten by the author to allow greater portability and efficiency. The original version required a complicated input file which had to be edited each time it was used, whereas the new versions take only command line parameters. The GKS library was used to produce graphical output from the original program which meant that the program could only be run on machines with an appropriate licence and was also VMS specific. The new version of PLOT uses the standard X11 and Motif libraries and is written in portable C. The author has also written a version that uses the GL graphics libraries which takes advantage of the extra graphics capabilities of Silicon Graphics machines.

The postscript drivers used by the new PLOT were also written by the author, in contrast to the previous version which used the GKS drivers, allowing greater flexibility.

The PLOT program takes the motif files produced by the final database search and then scans a single sequence using either the modified scoring technique described above or simple residue frequency scoring. The graph produced shows all the motifs on a single screen or sheet of paper, the areas of the test sequence which have a high degree of similarity to the motifs are indicated by peaks in the graph. Some idea of the discriminating efficiency of the motifs may also be obtained as good motifs should show clearly defined peaks of maximum value, whereas poor motifs show only poorly defined peaks.

PLOT is also particularly useful for quickly identifying similarities between new database sequence entries and existing motifs, such as those contained within the PRINTS database. Other methods of manipulating the motif information contained in the PRINTS database is described in a later chapter. Typical output from the PLOT program is shown in figure 2.8.

2.7 Consensus Motifs and Motif Variability

These programs are useful tools for extracting information from motif files and also for making data from large motif sets more manageable. They are discussed fully in a later chapter of this thesis.

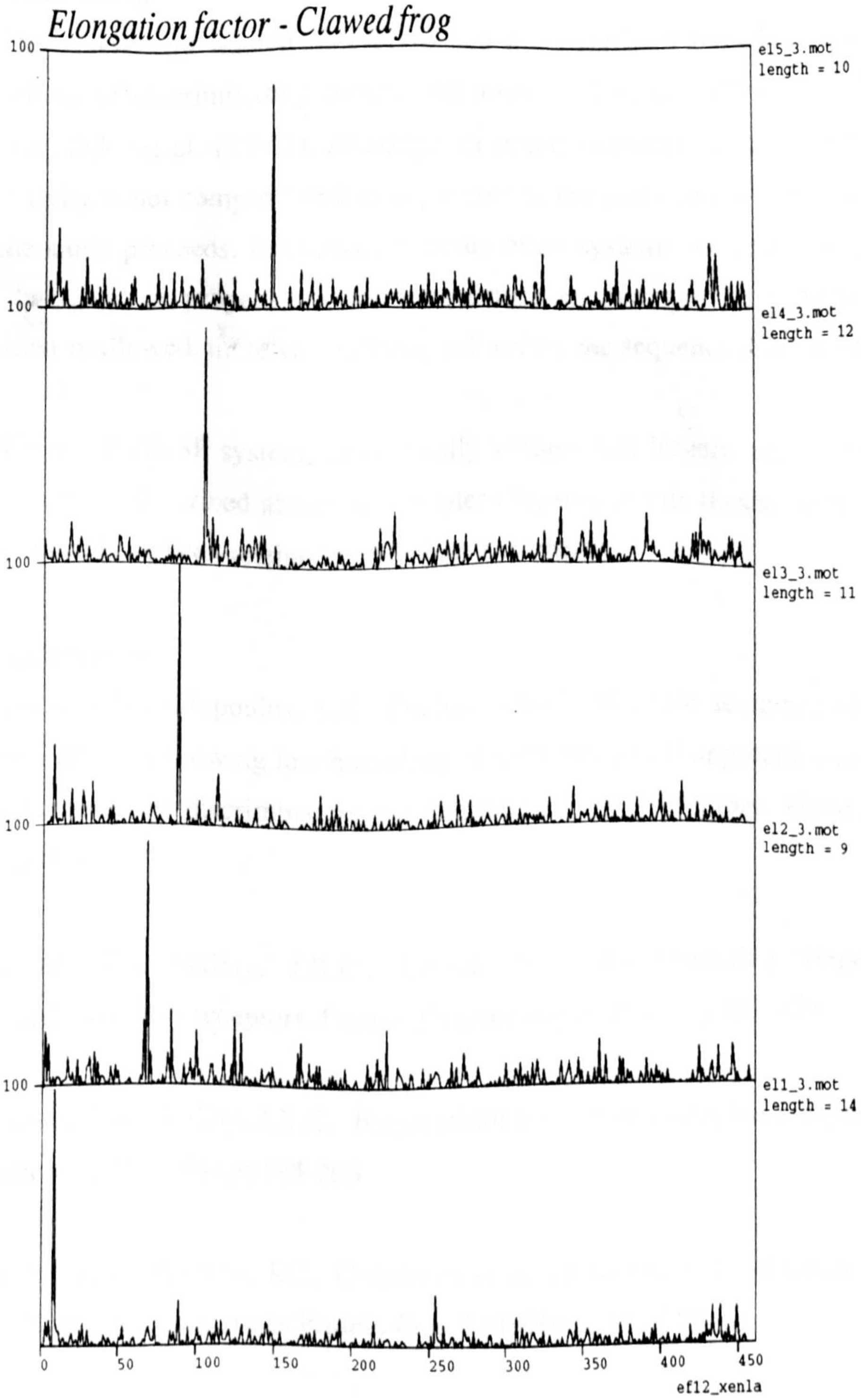


Figure 2.8 - Typical Postscript output from the XPLOT program. This example shows an elongation factor scanned with motifs defined by the author. The x-axis represents the residue number while the percentage score for each motif is shown on the y-axis.

2.8 Conclusion

The ADSP and related algorithms are a proven and reliable method for the definition of discriminating motifs (Attwood, T.K et al. (1991), (1993), (1994), Flower, D.R. et al. (1991)). Although an initial alignment is produced manually, objectivity is not compromised as any errors in the alignment will become obvious as the study proceeds. In contrast to many other systems, no user manipulation of the 'weights' or propensities for a particular residue type at a particular motif position is allowed, these values being defined by the sequence data alone.

While the ADSP system, as originally written, has largely been superseded by the programs described above and in later chapters of this thesis, most of the core algorithms have been retained with minor alterations.

2.9 References

Attwood, T.K., Eliopoulos, E.E., Findlay, J.B.C., Multiple sequence alignment of protein families showing low homology: a methodological approach using database pattern-matching discriminators for G-protein-linked receptors. Gene **98** (1991) pp153-159

Attwood, T.K.,Findlay, J.B.C., Design of a discriminating fingerprint for G-protein-coupled receptors. Protein Engineering **6** (1993) pp167-176

Attwood, T.K.,Findlay, J.B.C., Fingerprinting G-protein coupled receptors. Protein Engineering **7** (1994) pp195-203

Bleasby, A.J, Wootton, J.C., Construction of validated, non-redundant composite protein sequence databases Protein Eng. **3** (1990) pp153-155

Flower, D.R.,North,A.C.T.,Attwood, T.K., Mouse oncogene protein 24p3 is a member of the lipocalin protein family. Biochem. Biophys. res. comm. **180** (1991) pp69-74

Fraga, D., Hermolin, J., Oldenburg, M., Miller, M.J., Fillingame, R.H., Arginine 41 of the subunit c of Escherichia coli H^+-ATP synthase is essential in binding and coupling of F1 to F0. J. Biol. Chem. **269** (1994) pp7532-7537

Higgins, D.G., Bleasby, A.J., Fuchs, R., CLUSTALV: improved software for multiple sequence alignment. CABIOS **8** (1992) pp189-191

Parry-Smith, D.J. Algorithms and data structures for protein sequence analysis. Thesis (1990), University of Leeds

Parry-Smith, D.J., Attwood, T.K., SOMAP: a novel interactive approach to multiple protein sequence alignment. CABIOS **7** (1991) pp233-235

Chapter 3

C-type Lysozymes

3.1 Summary

Three sets of composite motifs have been assembled for c-type lysozyme, lactalbumin and super-family definition (in this case super-family refers to the set of sequences that is composed of all the c-type lysozymes and lactalbumins). An important region for discrimination was shown to be found in the calcium binding section of the lactalbumin sequences. From scans of the OWL protein sequence database, the diagnostic capacity of these motifs was confirmed as all sequences of the correct type were identified. Seventeen lactalbumin sequences were eventually used to construct the lactalbumin composite motifs, sixty-two c-type lysozyme sequences were used to create the final c-type lysozyme motifs and a total of eighty-one sequences were used for the super-family diagnostic motifs.

3.2 Introduction

Lysozymes are ubiquitous enzymes that have been isolated from the different organs or secretions of organisms as diverse as vertebrates, invertebrates, phages, and bacteria. Much of the early lysozyme data was collected from birds, including hen egg-white lysozyme which was the first enzyme structure to be elucidated (Blake, C.C.F. et al. (1965)). This family of lysozymes is discussed in this chapter and is known as chicken-type (or c-type) lysozymes although they have now been characterised in many other animals besides birds, for instance insects and mammals (Jolles, P. and Jolles, J. (1984)). The c-type lysozyme super-family is, however, distinct from the goose-type and T4 phage lysozyme families.

Lactalbumins have been shown to possess strong sequence and three-dimensional structural similarities to the c-type lysozymes and are thought to have evolved from a common ancestor (Nitta, K. and Sugai, S. (1989)). The intron-exon constitution of their respective genes are also virtually identical (Kumagi, I. et al. (1992)) lending support to this theory. The high degree of similarity between the primary structures means that the c-type lysozyme and lactalbumin families provide an excellent protein family for the validation of a sequence analysis technique as efficient and effective algorithms should ideally be able to distinguish not only the whole super family from the other database sequences but also lysozyme from lactalbumin sequences.

Despite the similarities mentioned above lysozyme and lactalbumin perform very different biological roles, although it is thought that there may be some similarity between their respective ligands. Lysozyme is responsible for the lysis of bacterial cell walls by catalysing the hydrolysis of the beta-1,4 glycosidic linkage between N-acetyl-D-glucosamine and n-acetyl-D-muramic acid in polysaccharides, they thus function as the first line of host defence against bacterial infection. In addition to this role, lysozyme has also been recruited as a digestive enzyme in a number of species such as cattle, deer and colombine monkey. The enzymes from these animals share a number of properties not shown in other lysozymes, ie a low optimum pH and resistance to pepsin. It is thought that these enzymes degrade the cell walls of bacteria in the gut, making the cell contents available for digestion (Irwin D.M and Wilson, A.C. (1989)).

In contrast to lysozyme, lactalbumin is found only in mammary glands and comprises fifteen percent of the total protein content of human milk. Here, this protein plays an important part in the production of lactose by modulating the carbohydrate binding properties of beta-galactosyltransferase in the lactating mammary gland through a protein-protein interaction, the resulting complex catalyses the addition of galactose to glucose to produce lactose. Lactalbumins have also been shown to be able to bind calcium, whereas this property is absent from the vast majority of lysozymes. It has been suggested that lactalbumin diverged from an ancestral lysozyme and that this involved the development of this calcium binding ability. It is uncertain when the gene duplication event which led to the development of lactalbumin from lysozyme occurred, some authors suggest that the event occurred before the divergence of birds and mammals (Prager, E.M. and Wilson, A.C. (1988)) while other data indicates the divergence was more recent (Shewale, J.G. et al. (1984)).

3.3 Motif Definition

To create motifs with which to scan the OWL database, a number of multiple alignments of protein sequences were prepared. Within both the lactalbumin and lysozyme families homology is relatively high and so the alignment process was quite straightforward. In the case of the lysozymes, twelve sequences were used to create a multiple alignment (figure A.1.1). Eleven sequences featured in the lactalbumin multiple alignment (figure A.1.2). For the definition of the super-family motifs an alignment of six lysozymes and six lactalbumins (ie a total of twelve sequences) was produced (figure A.1.3). Plots of the alignments were produced, coloured by positional variability, and examined for the regions of highest conservation.

After the first database scans, compound feature indices were produced by examining hitlists of one hundred for the lactalbumin discriminators, one hundred and fifty for the lysozyme discriminators and a hitlist of two hundred for the super-family discriminators. All hitlists had distance criteria applied which involved the use of a program which calculates the relative distance between motifs from the initial alignment. This technique removes the noise from the two features column of the compound features index, ie the signal to noise ratio is improved.

3.3.1 Lactalbumin Discriminators

Six motifs (figure 3.1) were selected from the most conserved sections of the lactalbumin alignment and used to scan the OWL sequence database. The first iteration produced seventeen sequences that matched with all six motifs. Eleven of these sequences had been used to create the original motifs. The appropriate motifs from the six additional sequences were added to the initial motif files and another database scan was carried out. The second scan showed seventeen sequences in the six features column, indicating that convergence had been reached as no extra sequences were found.

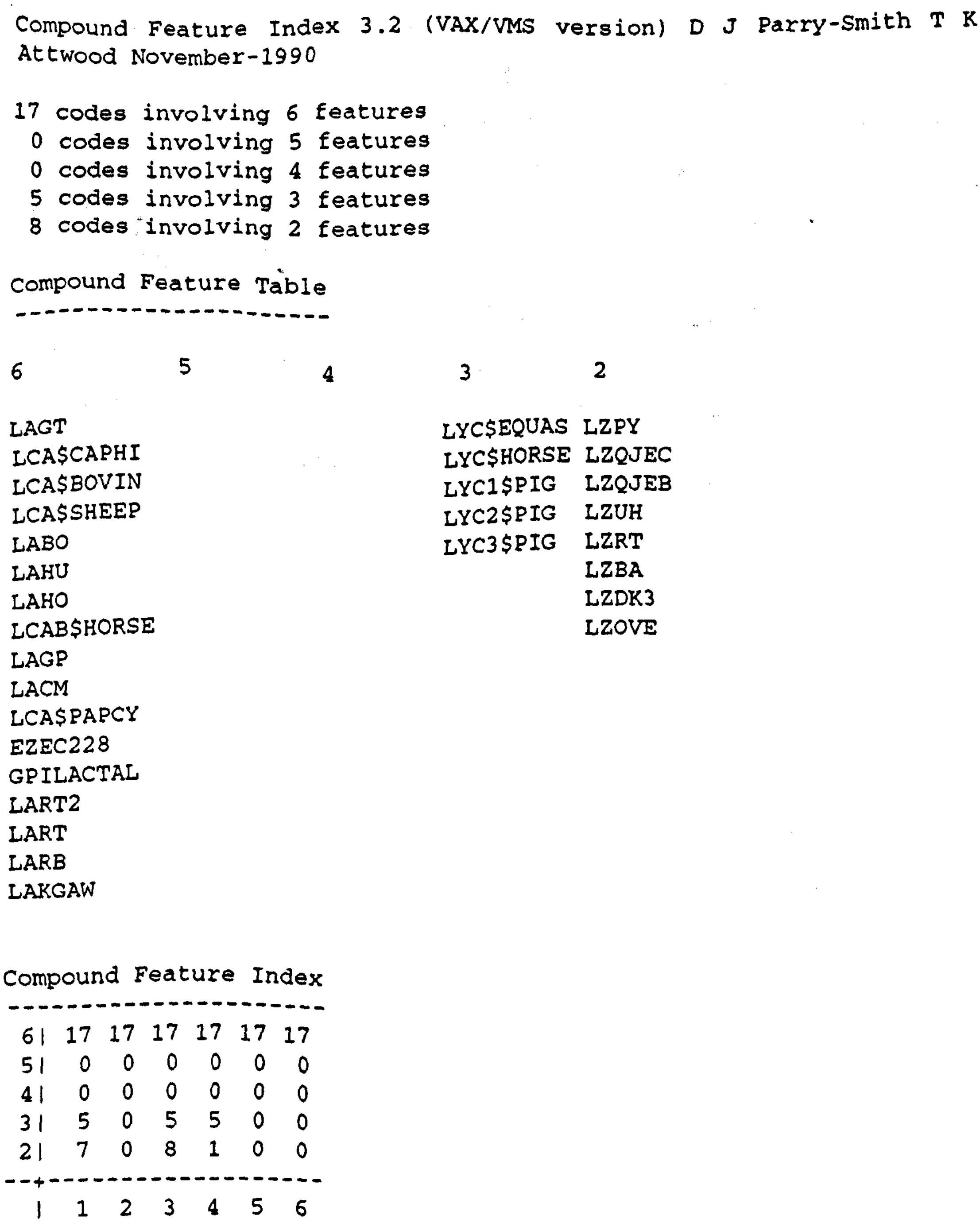
These seventeen sequences were found to be all the lactalbumin sequences contained within the OWL protein sequence database (version 9.0). The final Compound Feature Index is shown in figure 3.2. A number of lysozymes were also shown to match with two or three motifs.

Pcode	Motif 1	Motif 2	Motif 3	Motif 4
LABO	EVFRELKDLKGYGGVSLPEWV	FHTSGYDTEAIV	HSSNICNISC	KFLDDDLTDD
LCA$BOVIN	EVFRELKDLKGYGGVSLPEWV	FHTSGYDTQAIV	HSSNICNISC	KFLDDDLTDD
LAGT	EVFQKLKDLKDYGGVSLPEWV	FHTSGYDTQAIV	HSRNICNISC	KFLDDDLTDD
LCA$CAPHI	EVFQKLKDLKDYGGVSLPEWV	FHTSGYDTQAIV	HSRNICNISC	KFLDDDLTDD
LAHO	ELSEVLKSMDGYKGVTLPEWI	FHSSGYDTQTIV	PSRNICGISC	KFLDDDLTDD
LCAB$HORSE	QLSQVLKSMDGYKGVTLPEWI	FHNSGYDTQTIV	PSRNICGISC	KFLDDDLTDD
LART2	EVSHAIEDMDGYEGVSLPEWT	FHTSGYDTEASV	ESENICDISC	KFLDDELADD
LART	EVSHAIEDMDGYQGISLLEWT	FHTSGYDSQAIV	ESENICDISC	KFLDDELADD
LACM	KLSDELKDMNGHGGITLAEWI	FHMSGYDTETVV	QSRNICDISC	KFLDDDLTDD
LARB	ELTEKLKELDGYRDISMSEWI	FHTSGLDTKITV	QSKNICDTPC	NFLDDNLTDD
LAKGAW	QASQILKEHGMDKVIPLPELV	FHISGLSTQAEV	VANSVCGILC	KFLDDDITDD

Pcode	Motif 5	Motif 6
LABO	VGINYWLAH	CSEKLDQWLC
LCA$BOVIN	VGINYWLAH	CSEKLDQWLC
LAGT	VGINYWLAH	CSEKLDQWLC
LCA$CAPHI	VGINYWLAH	CSEKLDQWLC
LAHO	EGIDYWLAH	CSEKLEQWLC
LCAB$HORSE	EGIDYWLAH	CSEKLEQWLC
LART2	KGINYWLAH	CSEKLEQWRC
LART	KGIDYWKAH	CSEKLEQWRC
LACM	EGIDYWLAH	CSEKLEQWQC
LARB	EGIDHWLAH	CSENLEQWVC
LAKGAW	EGLGYWKAH	CLEDLDQWRC

LABO	Alpha-lactalbumin - Bovine
LCA$BOVIN	Alpha-lactalbumin precursor - Bovine
LAGT	Alpha-lactalbumin - Goat
LCA$CAPHI	Alpha-lactalbumin precursor - Goat
LAHO	Alpha-lactalbumin - Horse
LCAB$HORSE	Alpha-lactalbumin b and c - Horse
LART2	Alpha-lactalbumin (version 2) - Rat
LART	Alpha-lactalbumin - Rat
LACM	Alpha-lactalbumin - Arabian camel
LARB	Alpha-lactalbumin - Rabbit
LAKGAW	Alpha-lactalbumin - Red-necked wallaby

Figure 3.1 - The initial lactalbumin motifs.



```
Compound Feature Index 3.2 (VAX/VMS version) D J Parry-Smith T K
Attwood November-1990

17 codes involving 6 features
 0 codes involving 5 features
 0 codes involving 4 features
 5 codes involving 3 features
 8 codes involving 2 features

Compound Feature Table
----------------------

6            5           4          3           2

LAGT                               LYC$EQUAS   LZPY
LCA$CAPHI                          LYC$HORSE   LZQJEC
LCA$BOVIN                          LYC1$PIG    LZQJEB
LCA$SHEEP                          LYC2$PIG    LZUH
LABO                               LYC3$PIG    LZRT
LAHU                                           LZBA
LAHO                                           LZDK3
LCAB$HORSE                                     LZOVE
LAGP
LACM
LCA$PAPCY
EZEC228
GPILACTAL
LART2
LART
LARB
LAKGAW


Compound Feature Index
----------------------
  6|  17  17  17  17  17  17
  5|   0   0   0   0   0   0
  4|   0   0   0   0   0   0
  3|   5   0   5   5   0   0
  2|   7   0   8   1   0   0
--+-------------------------
   |   1   2   3   4   5   6
```

Figure 3.2 - The final compound feature index produced by the lactalbumin motifs. The motifs from and descriptions of all these sequences is shown in appendix B.1.1.

3.3.2 C-type Lysozyme Discriminators

Six motifs were selected from the most conserved regions of an alignment of twelve lysozyme sequences (figure 3.3). After the first iteration, sixty-two sequences were shown to display all six motifs. New files were prepared that contained all the motifs from these sequences and another iteration carried out. The second database search showed no extra sequences indicating that convergence had been reached.

These sixty-two sequences consisted of all the complete c-type lysozyme sequences in version 11.0 of the OWL composite database. Lactalbumin sequences where shown as a sub-family in the two features column of the final Compound Feature Index (figure 3.4).

Pcode	Motif 1	Motif 2	Motif 3
LZCH	VFGRCELAAAMKRHGLDN	KFESNFNTQATNR	PGSRNLCNIPC
LZQJEC	VFGRCELAAAMKRHGLDN	KFESNFNSQATNR	PGSRNLCNIPC
LZQJEB	VFGRCELAAAMKRHGLDN	KFESNFNSQATNR	PGSRNLCNIPC
N$2HFLY	VFGRCELAAAMKRHGLDN	KFESNFNTQATNR	PGSRNLCNIPC
N$3LYM	VFGRCELAAAMKRHGLDN	KFESNFNTQATNR	PGSRNLCNIPC
LZDK3	VYERCELAAAMKRLGLDN	NYESSFNTQATNR	PRAKNACGIPC
LZOVE	IYKRCELAAAMKRYGLDN	RYESNYNTQATNR	PGTKNLCHISC
LZBA	IFERCELARTLKRLGLDG	KWESDYNTQATNY	PGAVNACHISC
LZBO	VFERCELARTLKKLGLDG	KWESSYNTKATNY	PNAVDGCHVSC
N$1LZ1	VFERCELARTLKRLGMDG	KWESGYNTRATNY	PGAVNACHLSC
LYC1$PIG	VYDRCEFARILKKSGMDG	KWESDFNTKAINR	PKAVNACHISC

Pcode	Motif 4	Motif 5	Motif 6
LZCH	SALLSSDITASVNCAK	NGMNAWVAWR	NRCKGTDVQAWIRG
LZQJEC	SALLSSDITATVNCAK	NGMNAWVAWR	NRCKGTDVHAWIRG
LZQJEB	SALLSSDITATVNCAK	BGMNAWVAWR	NRCKGTDVQAWIRG
N$2HFLY	SALLSSDITASVNCAK	DGMNAWVAWR	NRCKGTDVQAWIRG
N$3LYM	SALLSSDITASVNCAK	NGMNAWVAWR	NRCKGTDVQAWIRG
LZDK3	SVLLRSDITEAVKCAK	DGMNAWVAWR	NRCKGTDVSRWIRG
LZOVE	SALMGADIAPSVRCAK	DGMNAWVAWR	KHCKGTDVSTWIKD
LZBA	NALLQDNITDAVACAK	QGIRAWVAWR	NHCQNRDVSQYVQG
LZBO	SELMENDIAKAVACAK	QGITAWVAWK	SHCRDHDVSSYVEG
N$1LYZ	SALLQDNIADAVACAK	QGIRAWVAWR	NRCQNRDVRQYVQG
LYC1$PIG	KVLLDDDLSQDIECAK	QGIKAWVAWR	THCQNKDVSQYIRG

LZCH	Lysozyme c precursor - Chicken
LZQJEC	Lysozyme c - California quail
LZQJEB	Lysozyme c - Common bobwhite
N$2HFLY	Lysozyme c - Chicken
N$3LYM	Lysozyme c - Hen egg
LZDK3	Lysozyme c III - Duck
LZOVE	Lysozyme c - Plain chachalaca
LZBA	Lysozyme c - Baboon
LZBO	Lysozyme c 2 - Bovine
N$1LYZ	Lysozyme c - Hen egg white
N$1LZ1	Lysozyme c - Human
LYC1$PIG	Lysozyme c I - Pig

Figure 3.3 - The six initial lysozyme motifs

```
Compound Feature Index

62 codes involving 6 features
0  codes involving 5 features
0  codes involving 4 features
0  codes involving 3 features
17 codes involving 2 features

Compound Feature Table

6                  5                4             3          2

LZCH                                                         LAGT
N$1LYMA                                                      LCA$CAPHI
N$1LYMB                                                      LCA$SHEEP
N$1LYZ                                                       LAHU
N$1LZHA                                                      LCAB$HORSE
N$1LZHB                                                      LABO
N$2HFMY                                                      LCA$BOVIN
N$2LYM                                                       LCA$PAPCY
N$2LYZ                                                       N$1ALC
N$2LZH                                                       LAHO
N$2LZT                                                       LAKGAW
N$3HFMY                                                      LAGP
N$3LYM                                                       GPILACTAL
N$3LYZ                                                       LCA$PIG
N$4LYZ                                                       LACM
N$5LYZ                                                       LART
N$6LYZ                                                       LARB
N$7LYZ
N$8LYZ
S05657
N$2HFLY
!LCOT
JT0526
EZEC462
N$1LZ2
LYC$MELGA
N$2LZ2
EZEC471
LZQJEC
LZQJEB
LZFER
EZEC470
LZUH
EZEC465
EZEC466
LZQJE
```

(Continued on the next page)

```
Compound Feature Table

6                      5                   4                3            2


LZDK3
LZDK
LZTK
LZOVE
LZBA
LZHU
HUMLSZA
N$1LZ1
LYC$RABIT
HUMLYZ
LYC$PREEN
LYCP$MOUSE
LYCM$MOUSE
LYC3$PIG
LYC1$PIG
LZRT
BOVLSZ3A
LZBO
LYC$AXIAX
LYC$SHEEP
BOVLSZ1A
LYC2$PIG
LYC$BOVIN
LYC$EQUAS
LYC$HORSE
LZPY
```

```
Compound Feature Index
----------------------
 6|  62  62  62  62  62  62
 5|   0   0   0   0   0   0
 4|   0   0   0   0   0   0
 3|   0   0   0   0   0   0
 2|   1   0  16  17   0   0
--+-----------------------------
  |   1   2   3   4   5   6
```

Figure 3.4 The final lysozyme compound feature index. The descriptions of and motifs from all these sequences is shown in appendix B.1.2.

3.3.3 Super-family Discriminators

Six motifs were defined from an initial alignment of twelve sequences comprising representatives of both lactalbumins and lysozymes (figure 3.5). These lysozyme c super-family discriminators produced eighty-one sequences in the six feature column after the first iteration. All these sequences were members of the super-family. New motif files were produced and another database scan carried out. This iteration produced no additional sequences, therefore convergence had been reached.

The eighty-one sequences shown to match with all six motifs were found to be all the complete lysozyme and lactalbumin sequences in the OWL database (version 11.0). Only one sequence was shown to match with two motifs. Figure 3.6 shows the final Compound Feature Index (CFI) produced.

Pcode	Motif 1	Motif 2	Motif 3	Motif 4
LZQJEB	FGRCELAAAMK	YSLGNWVCAA	STDYGVLQINSRWWCND	NLCNIPCSAL
LZUH	FGRCELAAAMK	YSLGNWVCAA	STDYGVLQINSRWWCND	NLCNIPCSAL
LAHO	FTKCELSEVLK	VTLPEWICTI	KTEYGLFQINNKMWCRD	NICGISCDKF
LYC$BOVIN	FERCELARTLK	VSLANWLCLT	STDYGIFQINSKWWCND	DGCHVSCREL
!LCOT	YGRCELAAAMK	YSLGNWVCAA	STDYGILQINSRWWCND	NLCNIPCSAL
LARB	LTRCELTEKLK	ISMSEWICTL	STEYGIFQINSKLWCVS	NICDTPCENF
LCA$PAPCY	FTKCELSQNLY	IALPELICTM	STEYGLFQISNALWCKS	NICDITCDKF
LYC3$PIG	YDRCEFARILK	VSLANWVCLA	STDYGIFQINSRYWCND	NACHISCKVL
LCA$SHEEP	LTKCEAFQKLK	VSLPEWVCTA	STEYGLFQINNKIWCKD	NICNISCDKF
LZPY	IPRCELVKILR	KTVANWVCLV	SRDYGIFQINSKYWCND	NACNINCSKL
LART2	FTKCEVSHAIE	VSLPEWTCVL	STEYGLFQISNRDWCKE	NICDISCDKF
LAKGAW	YRKCQASQILK	IPLPELVCTM	NKEYGIFQISNDGWCAE	SVCGILCSKF

Pcode	Motif 5	Motif 6
LZQJEB	LSSDITATVNCAKKIV	GMNAWVAWRNRC
LZUH	QSSDITATANCAKKIV	GMNAWVAWRKHC
LAHO	LDDDLTDDVMCAKKIL	GIDYWLAHKPLC
LYC$BOVIN	MENDIAKAVACAKHIV	GITAWVAWKSHC
!LCOT	LSSDITASVNCAKKIV	GMNAWVAWRNRC
LARB	LDDNLTDDVKCAMKIL	GIDHWLAHKPLC
LCA$PAPCY	LDDDITDDIMCAKKIL	GIDYWIAHKALC
LYC3$PIG	LDDDLSQDIECAKRVV	GIKAWVAWKAHC
LCA$SHEEP	LDDDLTDDIVCAKKIL	GINYWLAHKALC
LZPY	RDDNIADDIQCAKKIA	GLTPWVAWKKYC
LART2	LDDELADDIVCAKKIV	GINYWLAHKPMC
LAKGAW	LDDDITDDIECAKKIL	GLGYWKAHETFC

LZQJEB	Lysozyme c - Common bobwhite
LZUH	Lysozyme c - Helmeted guineafowl
LAHO	Alpha-lactalbumin - Horse
LYC$BOVIN	Lysozyme c precursor - Bovine
!LCOT	Lysozyme - Coturnix
LARB	Alpha-lactalbumin - Rabbit
LCA$PAPCY	Alpha-lactalbumin - Yellow Baboon
LYC3$PIG	Lysozyme c-3 - Pig
LCA$SHEEP	Alpha-lactalbumin precursor - Sheep
LZPY	Lysozyme c - Pigeon
LART2	Alpha-lactalbumin (version 2) - Rat
LAKGAW	Alpha-lactalbumin - Red-necked wallaby

Figure 3.5 - The six initial super-family motifs

```
Compound Feature Index

81 codes involving 6 features
 0 codes involving 5 features
 0 codes involving 4 features
 0 codes involving 3 features
 1 code involving 2 features

Compound Feature Table

6                5                4                3                2

LZCH                                                                LYC$SALGA
N$1LYMA
N$1LYMB
N$1LYZ
N$1LZHA
N$1LZHB
N$2HFLY
N$2HFMY
N$2LYM
N$2LYZ
N$2LZH
N$2LZT
N$3HFMY
N$3LYM
N$3LYZ
N$4LYZ
N$5LYZ
N$6LYZ
N$7LYZ
N$8LYZ
JT0526
!LCOT
LZQJEC
LZQJEB
EZEC471
S05657
LYC$MELGA
N$2LZ2
LZFER
EZEC462
N$1LZ2
LZUH
EZEC470
LZDK3
LZDK
EZEC465
EZEC466
LZOVE
LZQJE
LYC$RABIT
LZBA
LZTK
LYC$PREEN
```

(Continued on the next page)

```
Compound Feature Table

6                5                4                3                2

LZHU
N$1LZ1
HUMLSZA
HUMLYZ
LYCM$MOUSE
LZRT
LYCP$MOUSE
LYC1$PIG
LYC3$PIG
LYC2$PIG
LZBO
BOVLSZ3A
LYC$EQUAS
LYC$SHEEP
BOVLSZ1A
LYC$HORSE
LCA$SHEEP
LAGT
LCA$CAPHI
LYC$BOVIN
LYC$AXIAX
LCA$BOVIN
LAHO
LCA$PIG
LABO
LZPY
EZEC228
LAHU
LCAB$HORSE
LART2
LCA$PAPCY
N$1ALC
LARB
LART
LACM
LAKGAW
LAGP
GPILACTAL
```

```
Compound Feature Index
----------------------
  6|  81  81  81  81  81  81
  5|   0   0   0   0   0   0
  4|   0   0   0   0   0   0
  3|   0   0   0   0   0   0
  2|   1   1   0   0   0   0
--+--------------------
   |   1   2   3   4   5   6
```

Figure 3.6 - the final CFI produced after the second database scan. The motifs from and the descriptions of all these sequences is shown in appendix B.1.3.

3.4 Individual Sequence Analysis Using the Converged Lactalbumin, C-type Lysozyme and Super-Family Motifs

The final motif files produced by the iterative process were then used to scan individual sequences, the results were plotted using the XPLOT program written by the author and described in the previous chapter. This process allows the graphical illustration of the diagnostic efficiency of the selected motifs. The x-axis of each graph shows the residue number while the percentage score for each motif is shown on the y-axis.

Figure 3.7 shows the lactalbumin motifs used to scan c-type lysozyme and lactalbumin sequences. All the motifs are clearly shown in the lactalbumin sequences whereas the c-type lysozyme sequences achieve a relatively low score.

Figure 3.8 illustrates lactalbumin and c-type lysozyme sequences scanned with the c-type lysozyme motifs. The c-type lysozyme sequences show very high scores for all the motifs. In contrast, the scores for the lactalbumin sequences are relatively low for all or most of the discriminating motifs.

Figure 3.9 shows the super-family motifs used to scan individual c-type lysozyme and lactalbumin sequences. As can be seen, both c-type lysozyme and lactalbumin sequences score highly for these discriminators.

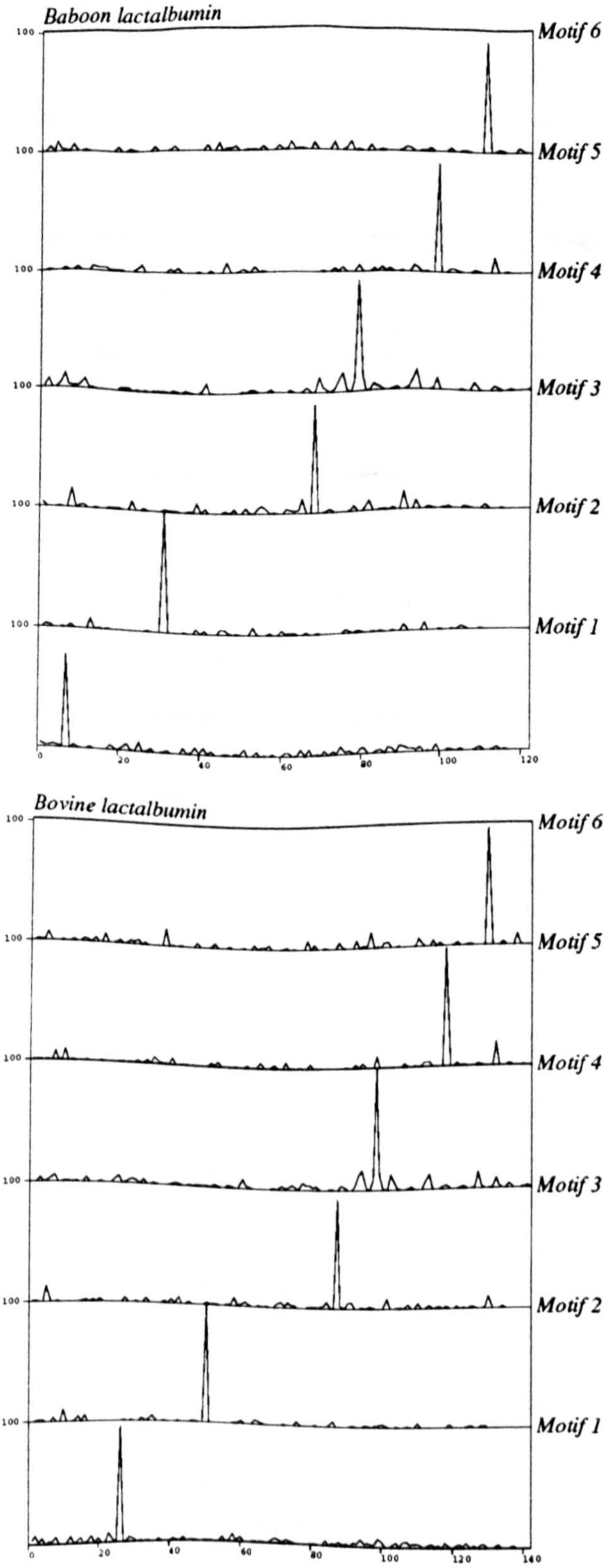


Figure 3.7a Individual lactalbumin sequences scanned with the final converged lactalbumin motifs

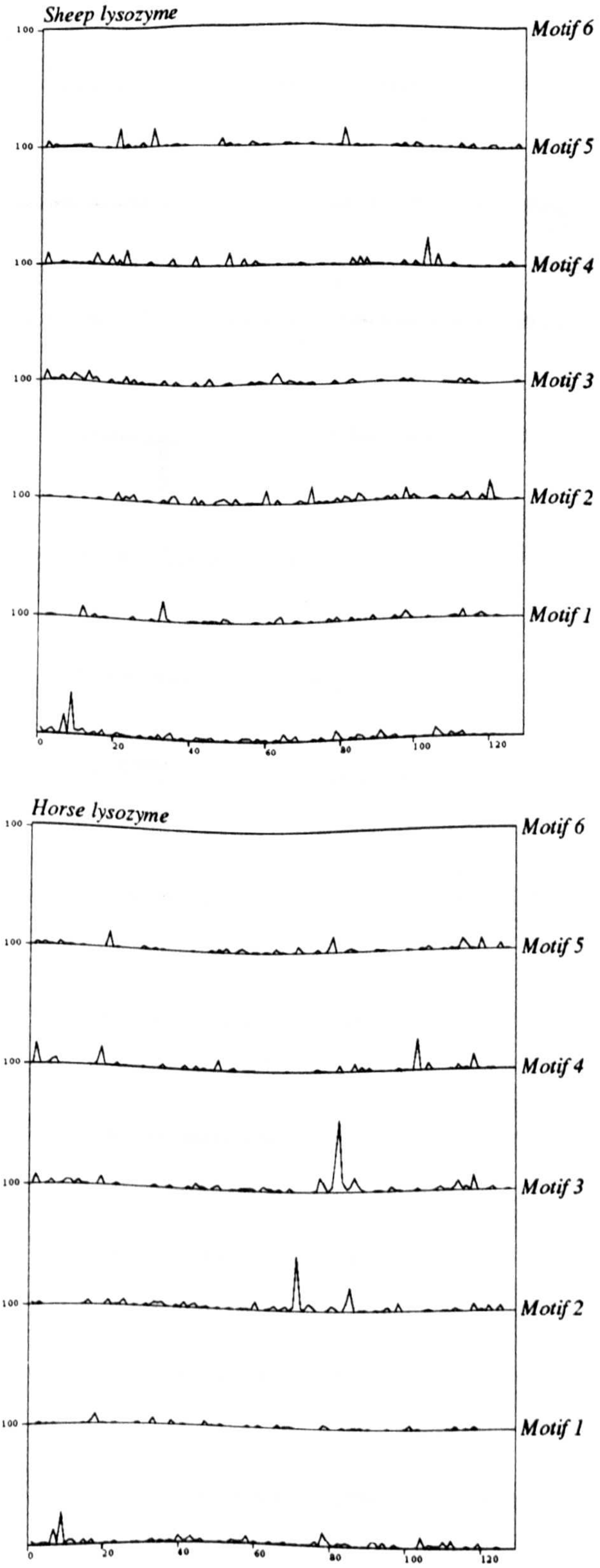
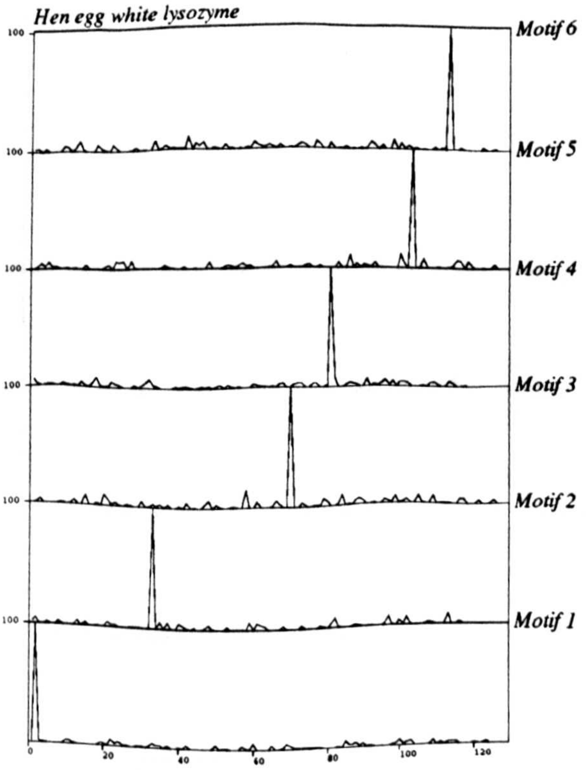


***Figure 3.7b** Lysozyme sequences scanned with the final lactalbumin motifs. Note the higher degree of similarity between the horse lysozyme and motif four. This is discussed later in this chapter.*



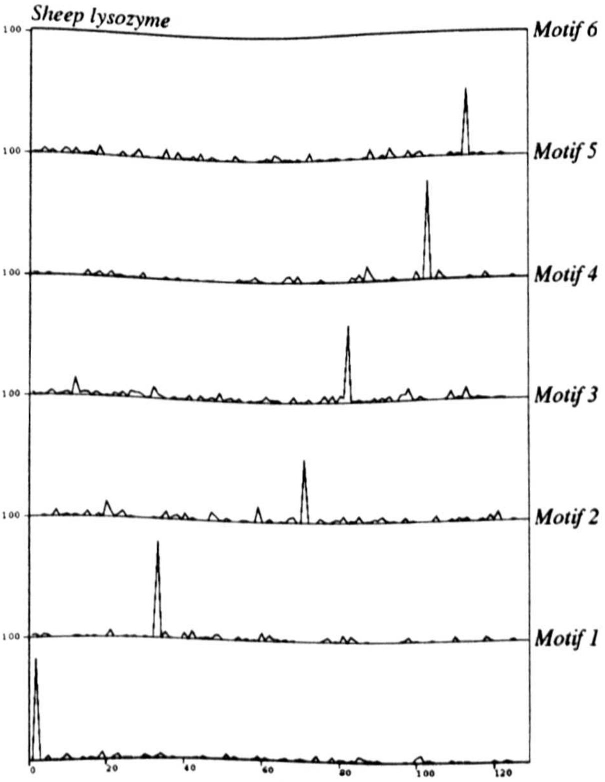


Figure 3.8a Lysozyme sequences scanned with the final lysozyme motifs

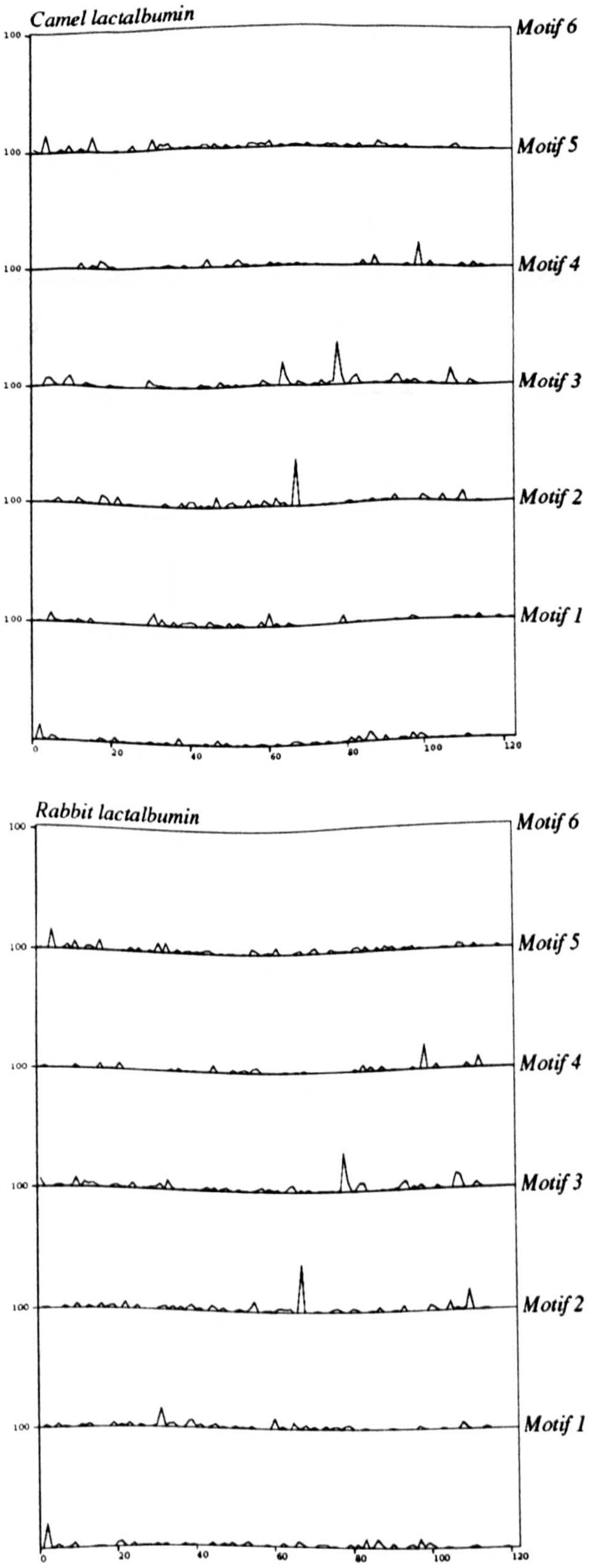
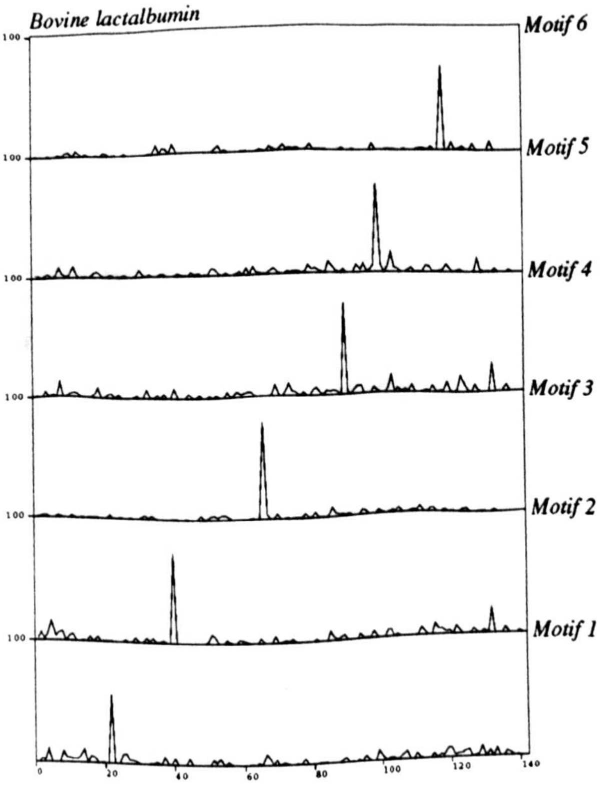


Figure 3.8b Lactalbumin sequences scanned with the converged lysozyme motifs



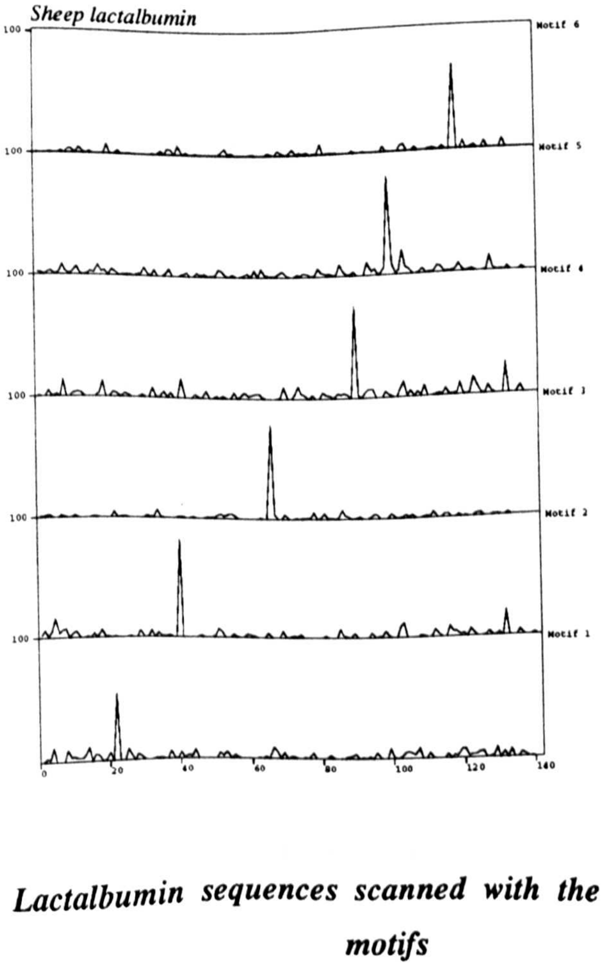


Figure 3.9a Lactalbumin sequences scanned with the converged super-family motifs

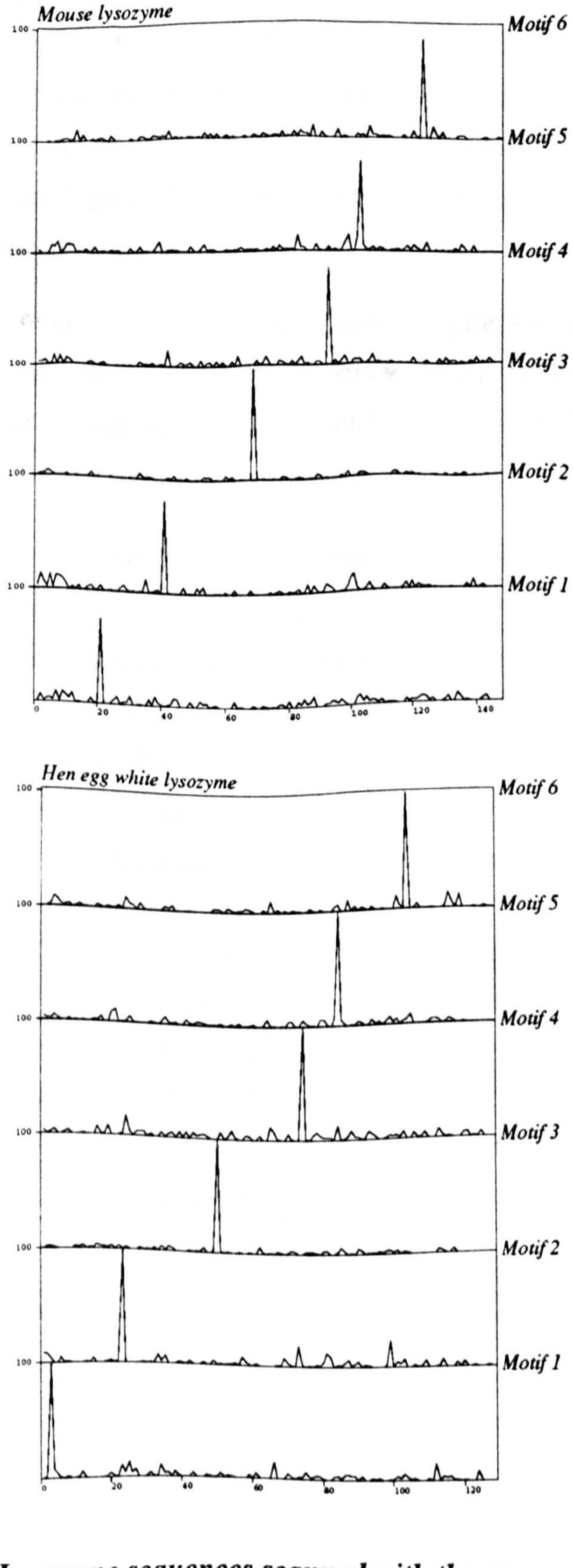


Figure 3.9b Lysozyme sequences scanned with the converged super-family motifs

3.5 Discussion

From figures 3.7 to 3.9 above it can be seen that the motifs chosen are very efficient discriminators for the appropriate sequences. The clear cut-off in each of the compound feature indices also indicates good discriminating power. Thus, each set of motifs are highly diagnostic for its own family and also show the other as a related sub-family.

The super-family compound feature index (figure 3.6) shows only one sequence in the two features column, the lysozyme sequence from the Rainbow Trout (LYC$SALGA) which matches with motifs one and two. The other four motifs are not seen in this sequence as it is a fragment.

In the case of the c-type lysozyme discriminators, seventeen lactalbumin sequences were shown as a sub-family which share two of the six selected motifs (figure 3.4). The proteins shown in the two feature all matched with motif four. Sixteen of the seventeen also matched with motif three while one sequence (the lactalbumin from Red-Necked Wallaby, database code LAKGAW) matched with motif one. Motifs two, five and six thus appear to have the greatest discriminating efficiency as no lactalbumins matched with these motifs.

With the lactalbumin motifs, the issue of sub family is slightly more involved. As figure 3.2 shows, a number of c-type lysozyme sequences are shown in the two features column but some also appear in the three feature column. The lysozyme sequences in the three feature column all show motif four along with motifs one and three. The lysozyme sequences shown in the two features all match with motifs one and three, apart from pigeon lysozyme (LZPY) that matches with motifs three and four.

Chachalaca lysozyme	ALMGADIAPS	Conventional lysozymes
Chicken lysozyme	ALLSSDITAS	
Donkey lysozyme	KLLDDNIDDD	Calcium binding lysozymes (including pig sequence)
Horse lysozyme	KLLDENIDDD	
Pig lysozyme	VLLDDDLSQD	
Pigeon lysozyme	KLRDDNIADD	
Bovine lactalbumin	KFLNNDLTNN	Lactalbumins
Rabbit lactalbumin	NFLDDNLTDD	
Rat lactalbumin	KFLDDELADD	
Human lactalbumin	KFLDDDITDD	

↑ ↑↑

Asp residues involved in calcium binding

Figure 3.10 Lactalbumin motif four. The bovine lactalbumin sequence includes Asn residues that are probably mis-identified Asp residues.

Figure 3.10 shows lactalbumin motif four from a number of c-type lysozyme and lactalbumin sequences. It can be seen that the horse, pig, donkey, and pigeon sequences are much more similar to the lactalbumin sequences than the other lysozyme sequences. In the lactalbumins the section of the sequence shown in figure 3.9 has been shown to be the region that is involved with the binding of calcium. The binding site was deduced using high resolution X-ray structure analysis (Stuart D.I. et al. (1986)) and was shown to consist of three aspartic acid residues (Residues 82, 87 and 88 using human lactalbumin sequence numbering). It was first suggested that the calcium bound to lactalbumin stabilised the structure, but recently it has been claimed that calcium controls the release of lactalbumin from the golgi membrane and that the pattern of ion binding may also affect the catalytic properties of the lactose synthetase complex.

In the case of horse lysozyme, the similarity with lactalbumins is at a functional level as this protein has been shown to be able to bind calcium (Nitta K. et al. (1987)). This functional similarity is also true of the Donkey lysozyme (Godovac-Zimmerman J. et al (1988)). Pigeon lysozyme also has been shown to have the ability to bind calcium (Nitta, K. et al. (1988)), and matches with motif four, but is less similar to the N-terminal region of the lactalbumins and thus lacks motif one.

The calcium binding site in the pig lysozymes (Jolles, J. et al. (1989)) is not as highly conserved and appears to be partially formed or destroyed as all three of the pig lysozyme sequences lack the second of the aspartic acid residues which have been shown to be involved in conferring the ability to bind calcium (this residue being replaced by a glutamine residue), although it is plain that these sequences share a higher similarity with lactalbumins than the other lysozymes in this region of the sequence. However, horse lysozyme which has the ability to bind calcium has an aspartic acid residue that is conserved in all the lactalbumins replaced by a glutamic acid residue in motif four indicating a possible exchangeability between the two types of acid residue and it may also be possible that the glutamine residue in the pig sequences has been wrongly identified as all three pig lysozymes were sequenced by the same laboratory at the same time. Such an example of possible mis-sequencing is seen in a bovine lactalbumin (database code EZEC228) where all three of the important aspartate residues are identified as asparagines. The sequencing authors describe all three pig sequences as conventional lysozymes (ie not of the same class as the horse milk and other calcium binding lysozymes), although they also mention that the pig lysozymes share few properties in common with the other ruminant stomach lysozymes being studied. All three pig sequences, however, do have the three conserved aspartic acid residues in the region of the calcium binding site which are also seen in almost all lactalbumins but not in any other lysozymes. In addition, the three pig lysozymes are all found in the stomach where calcium binding may be important to stabilise the structure in a harsh acid, protease rich environment. Other lysozymes which are found in the stomach do not show calcium binding properties although it is known that the pig stomach lysozymes have different properties than the ruminant stomach sequences, for instance the highest concentrations of the pig enzymes are found in the posterior stomach rather than the anterior stomach as in the case of the deer and cattle stomach lysozymes. It has also been postulated that the pig stomach lysozymes have a different role in that they protect against bacterial infection from the faeces that pigs sometimes eat as well as liberating the bacterial cell contents for digestion.

More recently, lysozyme mutants have been prepared which have the ability to bind calcium with a binding site similar to a EF-hand structure, these proteins having enhanced structural stability (Inaka K. et al. (1991)). These sequences also match with motif four. The bound calcium has been shown to enhance the structural stability of the protein and the mutant lysozymes were also shown to be more resistant to protease digestion (Kuroki, R. et al. (1989)).

The backbone of the other non-calcium binding lysozymes is similar to that of lactalbumins in the calcium binding region, but the side chains are generally radically different and are less well conserved. This is shown in figure 3.11 which illustrates the positional variability of the calcium binding region of lysozymes, calcium binding lysozymes and lactalbumins. The plot clearly shows that the degree of sequence conservation in this region of lysozyme sequences is at a much lower level than that of the lactalbumins, while the calcium binding lysozymes represent 'a half-way house' between the two. In this study a similarity matrix based on the superimposition of three-dimensional protein structures was used (Risler, J.L. et al. (1988)) lower values indicate a higher degree of conservation. When the positional variability residues for the entire motif are summed and then divided by the number of residues the following values are produced, further illustrating the higher degree of conservation shown by the lactalbumins in this region of the sequence. The number of different residues at each position in the motif divided by the length of the motifs is also shown :-

	Similarity matrix	**Different residues / length of motif**
1) Lysozyme	5.71	3.60
2) Horse/Donkey/Pig/Pigeon lysozymes	5.18	1.70
3) Lactalbumin	1.01	1.50

(In the case of the similarity score, 0 indicates total conservation, a value of 50 indicates the lowest degree of conservation. For the identity score, a value of 1 would be attained by a completed conserved motif, a value of 20 would indicate that the motif has no conserved positions.)

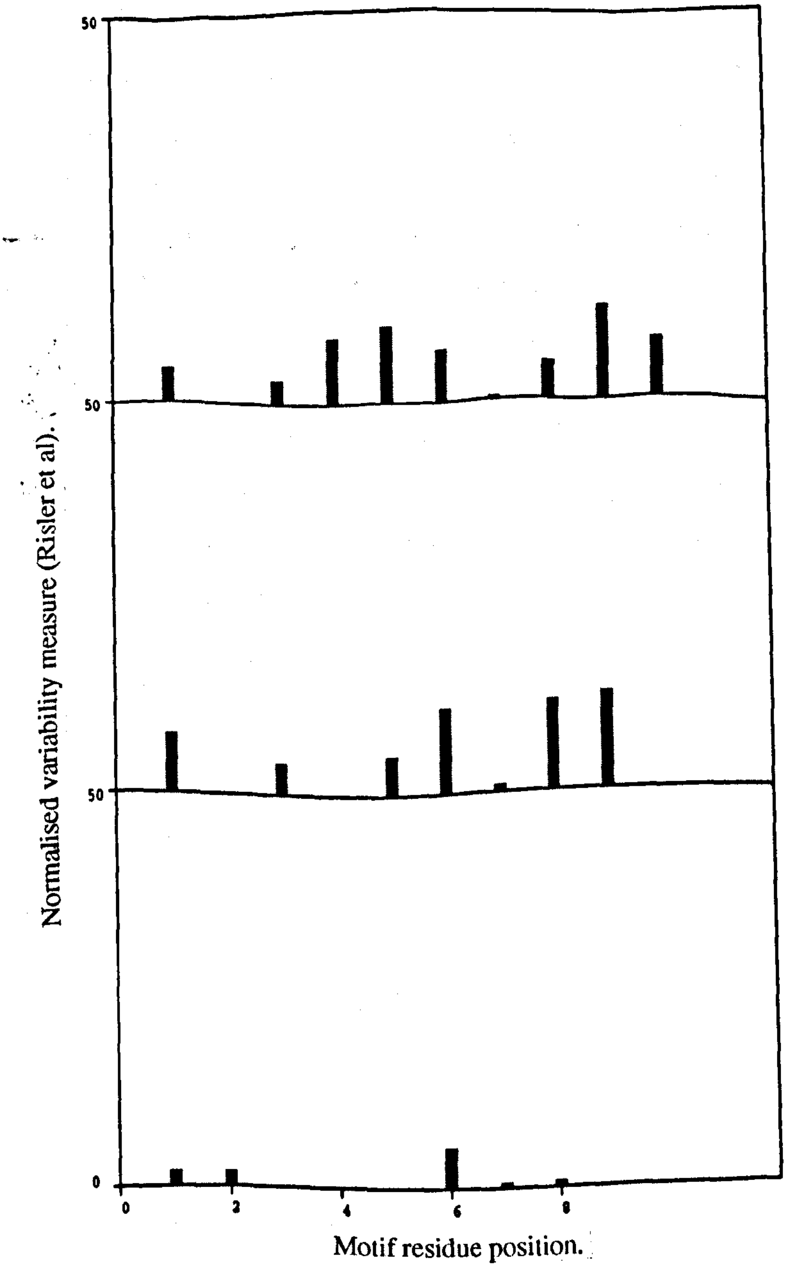


Figure 3.11 The positional variability of the lactalbumin calcium binding motif from lysozymes (top), calcium-binding lysozymes (middle) and lactalbumins (bottom). The higher bars indicate the more variable residues, residue number is on the x-axis.

In addition to the c-type lysozymes described above there are other classes of lysozymes which have been reported to show some similarity in their three-dimensional structures (Gutter et al. (1983)), these being the bacteriophage T4 type and goose type lysozymes. The author has produced alignments and carried out some sequence analysis with the latter type but, as so few sequences are available, this could not be profitably extended. The study carried out by the author described above however did not show any sequence similarity between the different types of lysozyme suggesting that the classes may have arisen by the process of convergent evolution, ie that there is no common ancestor.

Weaver et al. (1985), however, suggest that the structural similarity is too great for convergent evolution and that there must have been some distant common ancestor, although they admit that this theory is not supported by the sequence evidence or the intron-exon organisation of the appropriate genes.

3.6 Conclusion

This study has shown that the ADSP method of sequences analysis is a useful technique, in that it has the ability to distinguish between the lysozymes and lactalbumins. This efficiency is even more emphasised when the similarity of c-type lysozyme and lactalbumin sequences is considered.

In addition to validating the technique used to create discriminating motifs, the study has drawn attention to the similarity between the calcium binding region of lactalbumin sequences and similar areas of some c-type lysozyme sequences. In the case of all of these lysozymes, apart from the pig proteins, this similarity has been confirmed as the ability to bind calcium has been demonstrated experimentally. These motifs may thus be of use not only to predict new members of the lactalbumin and lysozyme families but also to identify those lysozyme sequences that possess possible calcium binding properties.

3.7 References

Blake, C.C.F., Koenig, D.F., Mair, G.A., North, A.C.T., Phillips, D.C., Sarma, V.R., Structure of hen egg-white lysozyme, A three-dimensional fourier synthesis at 2 angstroms resolution. Nature (1965) **206** pp757

Godovac-Zimmerman J., Conti A., Napolitano L., The primary structure of Donkey (*Equas asinus*) c-type lysozyme contains the Ca(II) binding site of alpha lactalbumin. Biol. Chem. **369** (1988) pp1109-1115

Gutter M.G., Weaver, L.H., Matthews, B.W., Goose lysozyme structure: an evolutionary link between hen and bacteriophage lysozymes ? Nature **303** (1983) pp828-831

Inaka K., Kuroki R., Kikuchi M., Matsushima M., Crystal Structures of the apoand holomutant human lysozymes with an introduced calcium binding site. J. Biol. Chem. **266** (1991) pp20666-20671

Irwin D.M., Wilson A.C., Multiple cDNA sequences and the evolution of bovine stomach lysozyme. J. Biol. Chem. **264** (1989) pp11387-11393

Jolles J., Jolles P., Whats new in lysozyme research - always a model system, today as yesterday. Mol. Cell. Biochem. **63** (1984) pp165-189

Jolles J., Jolles P., Bowman B.H., Prager E.M, Stewart C., Wilson A.C., Episodic evolution in the stomach lysozymes of ruminants. J. Mol. Evol. **28** (1989) pp528-535

Kumagi, I., Takeda, S., Miura, K.I., Functional conversion of the homologous proteins alpha-lactalbumin and lysozyme by exon exchange. Proc. Natl. Acad. Sci USA (1992) **89** pp5887-5891

Kuroki R., Taniyama Y., Seko C., Nakamura H., Kikuchi M., Ikehara M., Design and creation of a calcium binding site in human lysozyme to enhance structural stability. Proc. Natl. Acad. Sci. **86** (1989) pp6903-6907

Nitta, K., Sugai, S., The evolution of lysozyme and alpha-lactalbumin. Eur. J. Biochem. **182** (1989) pp111-118

Nitta, K., Tsuge,H., Shimazaki, K., Sugai, S., Calcium-binding lysozymes. Biol. Chem. Hoppe-Seyler **369** (1988) pp671-675

Nitta K., Hideaki T., Shintaro S., Shimazaki K., The calcium binding property of equine lysozyme FEBS Letters **223** (1987) pp405-408

Prager, E.M., Wilson, A.C., Ancient origin of lactalbumin from lysozyme: Analysis of DNA and amino acid sequences. J. Mol. Evol. **27** (1988) pp326-335

Risler, J.L., Delorme, M.O., Delacroix, H., Henaut, A., Amino acid substitutions in structurally related proteins, a pattern recognition approach. J. Mol. Biol. **204** (1988) pp1019-1029

Shewale J.G., Sinha S.K., Brew K., Evolution of alpha-lactalbumins. J. Biol. Chem. **259** (1984) pp4947-4956

Stuart D.I., Acharya K.R., Walker N.P.C., Smith S.G., Lewis M., Phillips D.C., alpha-lactalbumin possesses a novel calcium binding loop. Nature **324** (1986) pp84-87

Weaver, L.H., Grutter, M.G., Remington, S.J., Gray, T.M., Isaacs, N.W., Matthews, B.W., Comparison of goose-type, chicken-type, and phage-type lysozymes illustrates the changes that occur in both amino acid sequence and three-dimensional structure during evolution. J. Mol. Evol. **21** (1985) pp97-111

Chapter Four
Proton Symport/Antiport proteins

4.1 Summary

Two sets of composite motifs have been defined for this large and varied family. Five motifs were found to be good discriminators for those proteins responsible for sugar uptake while two motifs were found to be diagnostic of a wide range of symporters/antiporters with functions such as conferring antibiotic resistance and sugar transport across the cell membrane. Checks of the OWL protein sequence database confirmed that the sugar transporter motifs had identified all the known sequences in the family. In the case of the full symport/antiport family all known members of the family, with one exception, were found along with several new additions. Forty nine sugar transporters featured in one of the final motif sets, while seventy six sequences made up the final motif sets for the full family. Both studies were carried out using version 19.0 of the OWL protein sequence database.

4.2 Introduction

Owing to its essentially lipid character the cell membrane represents an effective barrier to the passage of hydrophilic molecules, therefore transport mechanisms play a vital role in the maintenance of the cell environment as they allow the influx of essential substrates consumed during cell growth and replication and maintain the ionic balance of the cell. The efflux of molecules also allows the cell to dispose of potentially toxic end metabolites.

Cells have developed three main methods of selective transport, ie molecules may diffuse down a concentration gradient (facilitated diffusion), be coupled to the concentration gradient of another ion (cotransport) or molecule transport may be driven by an energy dependent process (active transport). In all three cases, conformational change is thought to be the basic mechanism of transport (Walmsley, A.R. (1988)).

Cotransporters have been shown to couple molecule transport with a number of different ion species, for instance sodium and potassium. The majority of the proteins described in this study link the transport of molecules with the movement of protons, a process first suggested by Mitchell (1963). The proton gradient used to drive the uptake or efflux of molecules is produced by respiration or by the hydrolysis of ATP.

Those proteins which provide for the influx of molecules are known as symports as both the proton and molecule travel into the cell. Antiports catalyse the efflux of molecules, in this case the proton travels in the opposite direction to the molecule. The mammalian erthyrocyte glucose transport proteins described in this study, however, accumulate glucose in a facillitative manner. This is possible as the metabolic rate of red blood cells is great enough to ensure a very favourable concentration gradient (Walmsley, A.R. (1988)). Other members of this family may also work in a similar manner as all the proteins concerned are yet to be fully characterised. The majority of those family members that are found in eukaryotes probably operate in a facillitative manner. These proteins are known as uniports as only one molecule is transported.

As carbohydrates provide the main source of energy for a cell, sugar transport systems are very widespread. In mammals there are tissue specific glucose transporters that are members of this family, for instance GLUT 1 is found in red blood cells, GLUT 2 in the liver (the organ with the most crucial role in maintaining the correct blood sugar level), GLUT 3 in the brain and GLUT 4 which is found in insulin sensitive tissues (Kayano, T. et al. (1988)).

In addition to animals, sugar symporters from this family are very widespread being found in plants, fungi and also prokaryotes (Henderson, P.J.F et al. (1992)). Members of this family are also involved with the transport of citrate and alpha-ketoglutarate, both of these molecules are Krebs cycle intermediates.

The similarity between these eukaryote and prokaryote proteins is too great to be explained by convergent evolution and it is thought that the ancient ancestral protein was present in an organism that predates the divergence of the two groups (Maiden, M.C.J et al. (1987)).

In addition to their roles in nutrient uptake, the antiport proteins of this family are responsible for the antibiotic resistance properties seen in some microorganisms (Levy, S.B. (1992)). This diverse subset of proteins catalyse the efflux of such molecules as quinolone, tetracycline, methylenomycin, antiseptics and other drugs as well as aminotriazole and cycloheximide from yeast cells. The antiport proteins are thus of great medical significance.

4.3 Motif Definition

4.3.1 Sugar Transporters

Two alignments were produced to provide motif sets for these studies. In the case of the sugar transporters, sixteen sequences were aligned as shown in appendix A.2.1. This alignment was relatively easy to produce using a colour sequence alignment package, even though some sequences have quite low similarity.

It has been suggested that these proteins have twelve putative transmembrane regions (Mueckler, M. et al. (1985)) so a hydropathy plot was produced from this alignment to identify the probable transmembrane segments. These areas seem to be the most conserved whereas parts of the intracellular loops appeared to be generally only conserved across subsets of the family.

Five motifs were selected (figure 4.1), these being from regions of the sequences thought to contain the transmembrane segments. Motif one corresponds to putative transmembrane segment one, motif two corresponds to transmembrane segment four, motif three to transmembrane segment five, motif four to segment ten and motif five corresponds to transmembrane segment eleven. Figure 4.2 shows a hydropathy plot with the locations of the five motifs indicated.

A hitlist length of six hundred was used when comparing the five hitlists produced by the database scan. After the first iteration, forty eight sequences were shown to match with all the motifs. The extra motifs were added to the motif sets and another iteration carried out. The second compound feature index showed forty nine sequences that matched with all five motifs. New motif sets were prepared and another database scan carried out. This showed no extra sequences that matched with all the motifs indicating that convergence had been reached.

The final Compound Feature Index is shown in Figure 4.3. Distance criteria was applied to the final compound feature index to remove noise from the two features column. This distance criteria was defined using the initial alignment to ensure the required degree of objectivity.

Pcode	Motif 1	Motif 2	Motif 3
GAL2_YEAST	GFMFGWDTSTI	FIGRIISGLGVGGIAVLCPM	VSCYQLMITAGIFLGYCTNY
RATGLTP	SFQFGYDIGVI	IAGRSVSGLYCGLISGLVPM	GTLLQLGITVGIIISQILGL
MAL6_SACCA	LIQEGYDTAIL	AVGQALCGMPWGCFQCLTVS	TTYSNLCWTFGQLFAAGIMK
LACP_KLULA	ATMQGYDGALM	IGGRWFVAFFATIANAAAPT	AGLYNTLWSVGSIVAAFSTY
SNF3_YEAST	GFLFGYDTGLI	IVGRVISGIGIGAISAVVPL	ISTYQWAITWGLLVSSAVSQ
ARAE_ECOLI	GLLFGLDIGVI	IAARVVLGIAVGIASYTAPL	ISMYQLMVTLGIVLAFLSDT
GTR1_MOUSE	SLQFGYNTGVI	ILGRFIIGVYCGLTTGFVPM	GTLHQLGIVVGILIAQVFGL
QAY_NEUCR	SCMIGYDSAFI	IAGRVLAGIGVGGASNMVPI	VGIYELGWQIGGLVGFWINY
GTR4_MOUSE	SLQFGYNIGVI	ILGRFLIGAYSGLTSGLVPM	GTLNRLAIVIGILVAQVLGL
GTR5_HUMAN	SFQYGYNVAAV	IISRLLVGICAGVSSNVVPM	GVVPQLFITVGILVAQIFGL
HUP1_CHLKE	GLLLGYDNGVT	IVGRVLLGFGVGLGSQVVPQ	NIGYQLFVTIGILIAGLVNY
CIT1_ECOLI	FFLFGFYATYI	LVGRLLQGFSAGVELGGVSV	SASQQVAIVVAALIGYGLNV
CIT_KLEPN	FFLFGFYATYI	LIGRLLQGFSAGAELGGVSV	SGSQQVAIMVAAAMGFALNA
CITA_SALTY	FFLFGFYATYI	LLGRLLQGFSAGVELGGVSV	SASQQVAIVVAALIGYSLNI
LEID2TRA	PLLYGYNLGFV	FVARIVLGFPLGWQSITSSH	GTLFQVSVSTGIFVTSFFGL
PRO1_LEIEN	GSLNGYSIGFV	IVGRFVIGLFLGVICVACPV	GVMFQVFTTLGIFVAALMGL

Pcode	Motif 4	Motif 5
GAL2_YEAST	NCMIVFTCFYIFCYATTWAPVAWV	AESFPLRVKSKCM
RATGLTP	YVSMTAIFLFVSFFEIGPIPIPFF	REWFTQIWRPGAI
MAL6_SACCA	MGSGALLMVVAFFYNLGIAPVVFC	SEMPSSRLRTKTI
LACP_KLULA	NGALVFIYLFGGIFSFAFTPMQSM	TEVSTNLTRSKAQ
SNF3_YEAST	KVMIAFICLFIAAFSATWGGVVWV	AELYPLGVRSKCT
ARAE_ECOLI	WLSVGMTMMCIAGYAMSAAPVVWI	SEIQPLKCRDFGI
GTR1_MOUSE	YLSIVAIFGFVAFFEVGPGPIPWF	AELFSQGPRPAAI
QAY_NEUCR	IAAIFFFYLWTAFYTPSWNGTPWV	SEMFDQNTRSLGQ
GTR4_MOUSE	YVSIVAIFGFVAFFEIGPGPIPWF	AELFSQGPRPAAM
GTR5_HUMAN	YISIVCVISYVIGHALGPSPIPAL	TEIFLQSSRPSAF
HUP1_CHLKE	SGILAVICIFISGFAWSWGPMGWL	SEIFTLETRPAGT
CIT1_ECOLI	FTRMTLVLLWFSFFFGMYNGAMVA	TEVMPVYVRTVGF
CIT_KLEPN	FLMMLSVLLWLSFIYGMYNGAMIP	TEIMPAEVRVAGF
CITA_SALTY	FTRMTLVLLWFSFFFGMYNGAMVA	TEVMPVYVRTVGF
LEID2TRA	GIAITGIAIFIALYEMGVGPCFYV	VDVFPESFRPIGS
PRO1_LEIEN	GVAITGILLFILGFEVCVGPCYYV	QDMFPPSFRPRGA

GAL2_YEAST	Galactose permease - Yeast
RATGLTP	Glucose transporter - Rat
MAL6_SACCA	Maltose permease - Yeast
LACP_KLULA	Low-affinity glucose transporter - Yeast
SNF3_YEAST	High-affinity glucose transporter SNF3 - Yeast
ARAE_ECOLI	Arabinose-proton symport - Yeast
GTR1_MOUSE	Glucose transporter protein - Mouse
QAY_NEUCR	Quinate transporter - *Neurospora crassa*
GTR4_MOUSE	Glucose transporter - Mouse
GTR5_HUMAN	Glucose transporter - Human
HUP1_CHLKE	Hexose cotransporter - *Chlorella kessleri*
CIT1_ECOLI	Citrate-proton symport - *E. coli*
CIT_KLEPN	Citrate-proton symport - *Klebsiella pneumoniae*
CITA_SALTY	Citrate-proton symport - *Salmonella typhimurium*
LEID2TRA	Glucose transporter - *Leishmania donovani*
PRO1_LEIEN	Probable transport protein - *Leishmania enriettii*

Figure 4.1 - The five motifs defined for the sugar transport and related proteins

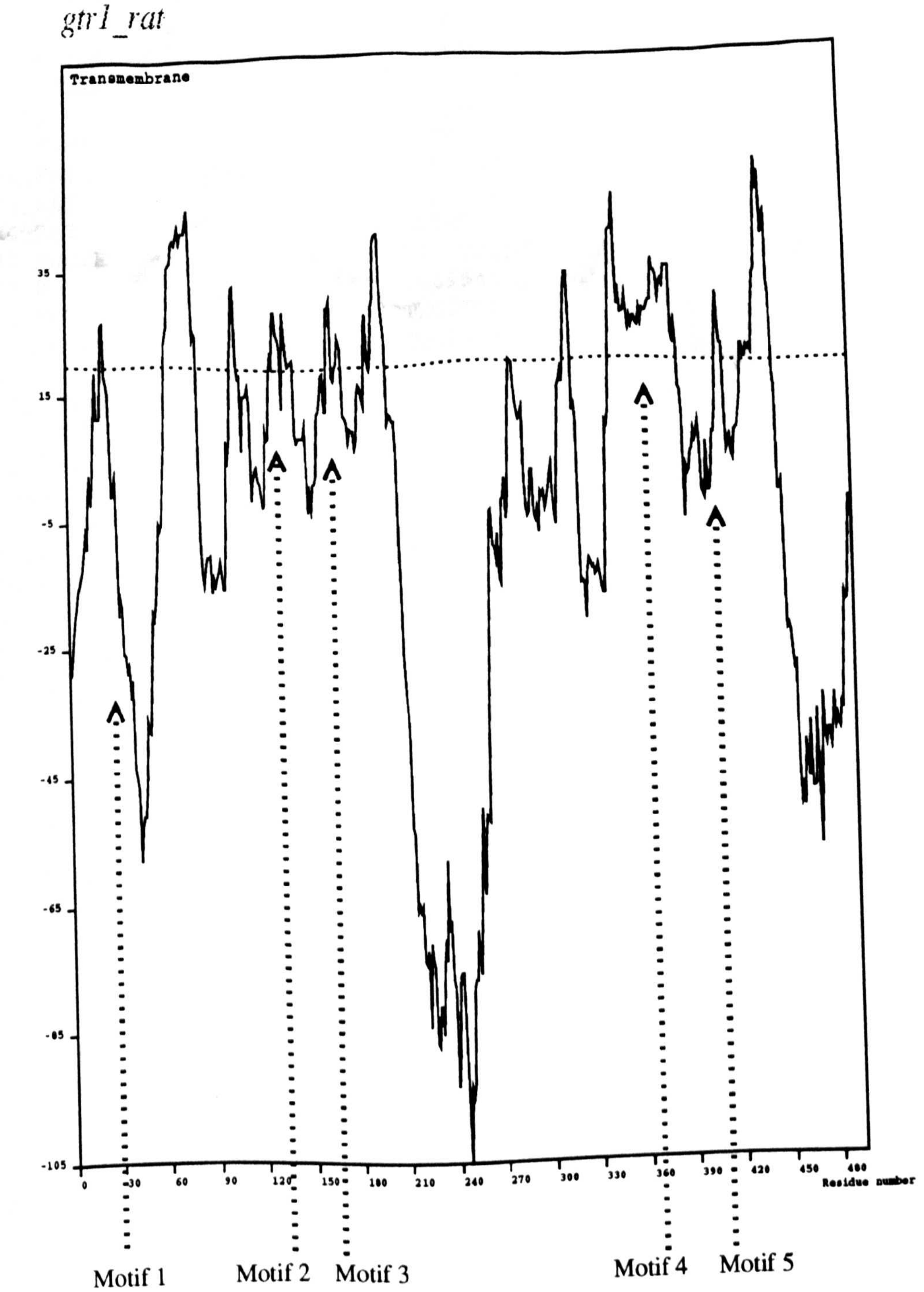


Figure 4.2 The location of the five motifs with regard to the twelve putative transmembrane segments. The x-axis represents residue number while the hydropathy value is indicated on the y-axis. Parts of the graph above the dotted line indicate significantly hydrophobic segments (on the scale defined by Engelman et al. (1986)). The sequence used for the graph is GTR1_RAT (Rat type 1 glucose transporter).

```
49 codes involving 5 features, 1 code involving 4 features, 0 codes
involving 3 features, 11 codes involving 2 features

Compound Feature Table
5          4           3         2
GTR1_BOVIN GTR1_PIG              STMBAHBRP
GTR1_HUMAN                       KGTP_ECOLI
S09705                           TCR1_ECOLI
GTR4_HUMAN                       ECOTN10
GTR1_RABIT                       HYIN_PSESS
GTR1_RAT                         VGLM_PHV
GTR4_RAT                         ECOPNT1
GTR1_MOUSE                       PNTB_ECOLI
A30310                           S1049541
GTR4_MOUSE                       STYCITCB
GTR3_CHICK                       STYCITCA
A41751
GTR3_HUMAN
GTR2_HUMAN
GTR2_RAT
S05319
GTR2_MOUSE
RATGLTP
STP1_ARATH
TOBMST1
SNF3_YEAST
HUP1_CHLKE
CHLHUP1G
A40538
B40538
YSCHXT4A
XYLE_ECOLI
HXT2_YEAST
RAG1_KLULA
A39728
GLCP_SYNY3
GAL2_YEAST
JQ0383
ATHSTP4
GTR5_HUMAN
GLF_ZYMMO
LEID1TRA
QAY_NEUCR
S108238
PRO1_LEIEN
ARAE_ECOLI
QUTD_ASPNI
LEID2TRA
CIT1_ECOLI
CIT2_ECOLI
CITA_SALTY
CIT_KLEPN
MAL6_YEAST
LACP_KLULA
```

```
Compound Feature Index
 ------------------
 5| 49 49 49 49 49
 4|  0  1  1  1  1
 3|  0  0  0  0  0
 2|  4  7  6  3  3
--+---------------
  | 1  2  3  4  5
```

Figure 4.3 The final Compound Feature index for the sugar transporters. The key to the database codes and the full motifs are shown in appendix B.2.1

4.3.2 Full Family

For the full family discriminators an alignment of twenty two sequences, including some symport and antiport proteins, was produced as shown in figure A.2.2. All the transmembrane sections that had a relatively high degree of similarity were used to scan the database, but only two motifs were shown to have discriminating ability. Of these two motifs, motif one corresponds to transmembrane segment four while motif two corresponds to transmembrane segment five (figure 4.4). As only two motifs were shown to have discriminating ability, the ADSP technique had to be extended as the COMPARE module used to produce compound feature indices is only really of use with three motifs or more. With only two motifs there is the strong possibility that 'noise', in the form of randomly matched motifs, would appear in the compound feature indices. Therefore a distance criteria was imposed after each database scan, so not only must motifs match with a sufficiently high score but they must also be in the correct region of the sequence. To maintain the objectivity of the study, this distance criteria was set using the initial alignment as a template. Motifs from the amino acid transporters from *E. coli, S. typhimurium* and *P. aeruginosa* were also removed as hydropathy plots indicated that the motifs were located in the wrong transmembrane segments, these proteins are the membrane channel components of their respective periplasmic binding protein-dependent systems (Nazos, P.M. et al. (1986)) and are probably unrelated.

A hitlist length of eight hundred and fifty was used for the comparison of the hitlists. After the first iteration, seventy one sequences were shown to match with both motifs. The additional motifs were added to the initial motif sets and another database scan carried out. This iteration added five more additional sequences making a total of seventy six sequences. After the third database scan seventy six sequences were shown to match with all motifs, the lack of any additional sequences indicated that convergence had been reached. Figure 4.5 shows the final Compound Feature Index.

Pcode	Motif 1	Motif 2
RAG1_KLULA	QYFIGRIISGLGVGGITVLSP	SCYQLMITFGIFLGYCTNYGTK
ATR1_YEAST	FFIISRAFQGLGIAFVLPNVL	SFVGAMAPIGATLGCLFAGLIG
GTR1_RAT	MLILGRFIIGVYCGLTTGFVP	TLHQLGIVVGILIAQVFGLDSI
GTR5_HUMAN	LIIISRLLVGICAGVSSNVVP	VVPQLFITVGILVAQIFGLRNL
ARAE_ECOLI	MLIAARVVLGIAVGIASYTAP	SMYQLMVTLGIVLAFLSDTAFS
QAY_NEUCR	PIIAGRVLAGIGVGGASNMVP	GIYELGWQIGGLVGFWINYGVN
SNF3_YEAST	LLIVGRVISGIGIGAISAVVP	STYQWAITWGLLVSSAVSQGTH
M22563381	AIVVFRVLQGLFGALMQPSAL	GVVGASTAAGPIIGGLLVQHVG
S19863	LLVLARFGQGAGEALSLPAAM	SVASVGLVLGFLLSGVITQLFS
CITA_SALTY	LVLLGRLLQGFSAGVELGGVS	ASQQVAIVVAALIGYSLNITLG
CIT_KLEPN	LVLIGRLLQGFSAGAELGGVS	GSQQVAIMVAAAMGFALNAVLE
JQ1479	VLYIGRIVAGITGATGAVAGA	ACFGFGMVAGPVLGGLMGGFSP
TCR1_ECOLI	MLYLGRLLSGITGATGAVAAS	ASFGLGLIAGPIIGGFAGEISP
TCR_BACST	LLIMARFIQGAGAAAFPALVM	SIVAMGEGVGPAIGGMIAHYIH
STMBAHBRP	VLIAARLVQGFSLGGEYGAAT	SFQYVASSVGHILAGLSTLAAS
PRO1_LEIEN	VLIVGRFVIGLFLGVICVACP	VMFQVFTTLGIFVAALMGLALG
S18593	VLVACRVVAALANAGFLAVAL	SGTTVATVAGVPGGSLLGTWLG
GTR1_HUMAN	MLILGRFIIGVYCGLTTGFVP	TLHQLGIVVGILIAQVFGLDSI
S18539	QLIAARACMGVSGAAVLPSTL	ASVGFALGIGPVTGGILLAHFW
JQ1201	VFLGLRILQACGASACLVSTF	SMLAMVPAVGPLLGALVDMWLG
S21395	VLLVTRIVGALANAGFLAVAL	GGVTIACVVGVPGGALLGELWG
B40046	MLTAARFLQGGLGALMIPQGL	PAIGLGAVLGPIVAGFLVDADL

RAG1_KLULA	Glucose transporter - *Kluyveromyces lactis* (Yeast)
ATR1_YEAST	Aminotriazole resistance protein - Yeast
GTR1_RAT	Glucose transporter protein, type 1 - Rat
GTR5_HUMAN	Glucose transporter, type 5 - *Homo sapiens* (Human)
ARAE_ECOLI	Arabinose-proton symport - *Escherichia coli*
QAY_NEUCR	Quinate transporter - *Neurospora crassa*
SNF3_YEAST	Glucose transporter - *Saccharomyces cerevisiae*
M22563381	tmcA protein - *Streptomyces glaucescens*
S19863	Lincomycin resistance protein - *Streptomyces licolnensis*
CITA_SALTY	Citrate-proton symport - *Salmononella typhimurium*
CIT_KLEPN	Citrate-proton symport - *Klebsiella pneumoniae*
JQ1479	Tetracycline resistance protein - *Escherichia coli*
TCR1_ECOLI	Tetracycline resistance protein - *Escherichia coli*
TCR_BACST	Tetracycline resistance protein - *B. stearothermophilus*
STMBAHBRP	STMBAHBRP ORF3 - *Streptomyces hygroscopicus*
PRO1_LEIEN	Probable transport protein (LTP) - *Leishmania enriettii*
S18593	Chloramphenicol resistance protein - *Streptomyces lividans*
GTR1_HUMAN	Glucose transporter protein, type 1 - *Homo sapiens*
S18539	actVA-1 protein - *Streptomyces coelicolor*
JQ1201	CmlA protein - *Pseudomonas* sp.
S21395	Chloramphenicol resistance protein - *Rhodococcus fascians*
B40046	Tetracycline resistance homolog - *Streptomyces coelicolor*

Figure 4.4 The initial motifs used to define the super-family motifs

```
76 codes involving 2 features
```

Compound Feature Table

```
    2
GTR4_HUMAN
CIT2_ECOLI
GTR5_HUMAN
MMR_STRCO
GLCP_SYNY3
TCR1_ECOLI
GTR1_BOVIN
GTR1_HUMAN
GTR1_MOUSE
GTR1_PIG
GTR1_RABIT
GTR1_RAT
S09705
A30310
GTR4_MOUSE
GTR4_RAT
GTR3_CHICK
A41751
HUP1_CHLKE
CHLHUP1G
GTR3_HUMAN
SNF3_YEAST
GTR2_HUMAN
GTR2_MOUSE
GTR2_RAT
S05319
RATGLTP
STP1_ARATH
TOBMST1
ATHSTP4
S22742
QAY_NEUCR
TCR1_BACSU
PRO1_LEIEN
TCR_BACST
TCR_STRAG
TCR_STRPN
RAG1_KLULA
A39728
YSCHXT4A
GAL2_YEAST
JQ0383
HXT2_YEAST
ARAE_ECOLI
TCR2_BACSU
TCR_STAAU
QQSABT
CIT_KLEPN
CITA_SALTY
QUTD_ASPNI
S108238
CIT1_ECOLI
ECOTN10
XYLE_ECOLI
STMBAHBRP
S19863
RATCGAT
TCR3_ECOLI
JQ1479
TCR2_ECOLI
ACCPCAOP3
GLF_ZYMMO
RATSVAT
B40046
A39705
QACA_STAAU
ATR1_YEAST
LEID2TRA
S18539
S108506
YSACYHR
BMR_CANAL
M225633S1
S21395
S18593
JQ1201
```

```
2| 76 76
--+------
 |  1  2
```

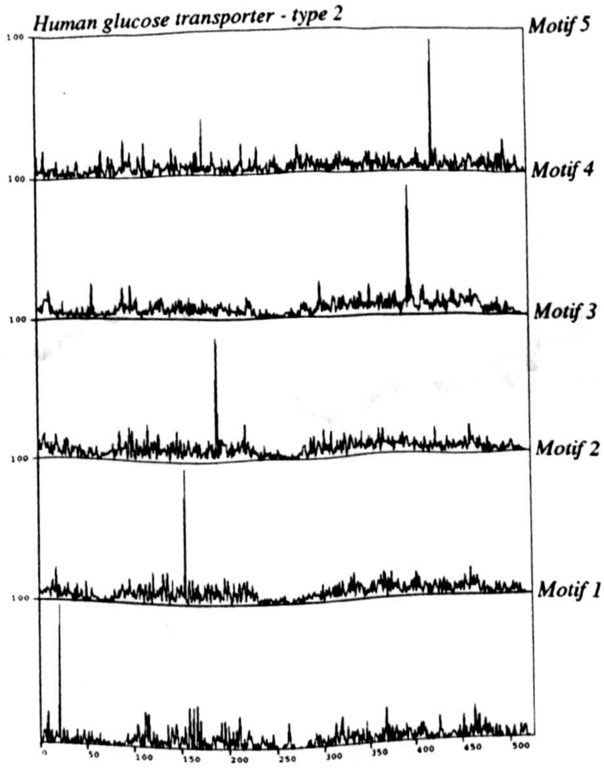
Figure 4.5 - The final Compound Feature Index for the super-family motifs. The key to the database codes and the full motifs are shown in appendix B.2.2

4.4 Discriminator efficiency

To demonstrate the effectiveness of the discrimination provided by the final motif sets, individual sequences were scanned. Figure 4.6 illustrates a number of sugar transport proteins scanned with the sugar transporter motifs. As can be seen, peaks appear in regions of the sequence that match a particular motif. The height of these peaks and the fact that they appear in the correct spacing indicate a high discrimination efficiency. Another transmembrane protein, in this case the cystic fibrosis conductance regulator, is also shown as a control to illustrate how poorly unrelated sequences score. This protein is a particularly useful control as it also contains twelve putative transmembrane segments.

The same procedure was carried out for the super-family motifs. Again the cystic fibrosis conductance regulator was used as a control. The three members of the family are shown to score well, while the control protein matches only poorly (Figure 4.7).

The discriminating efficiency of the motifs selected is thus confirmed by these graphs, as true sequences show high scores while the control sequence scores poorly.



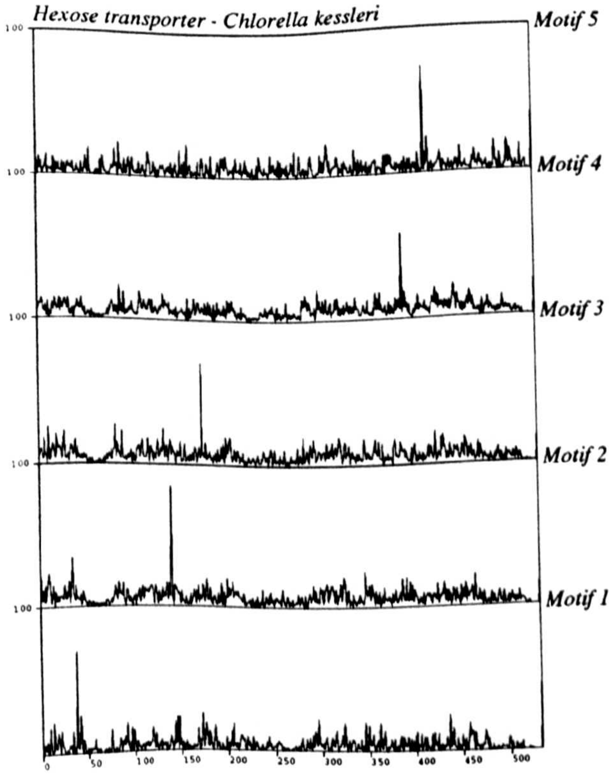


Figure 4.6a Sugar transport sequences scanned with the final sugar transporter motifs

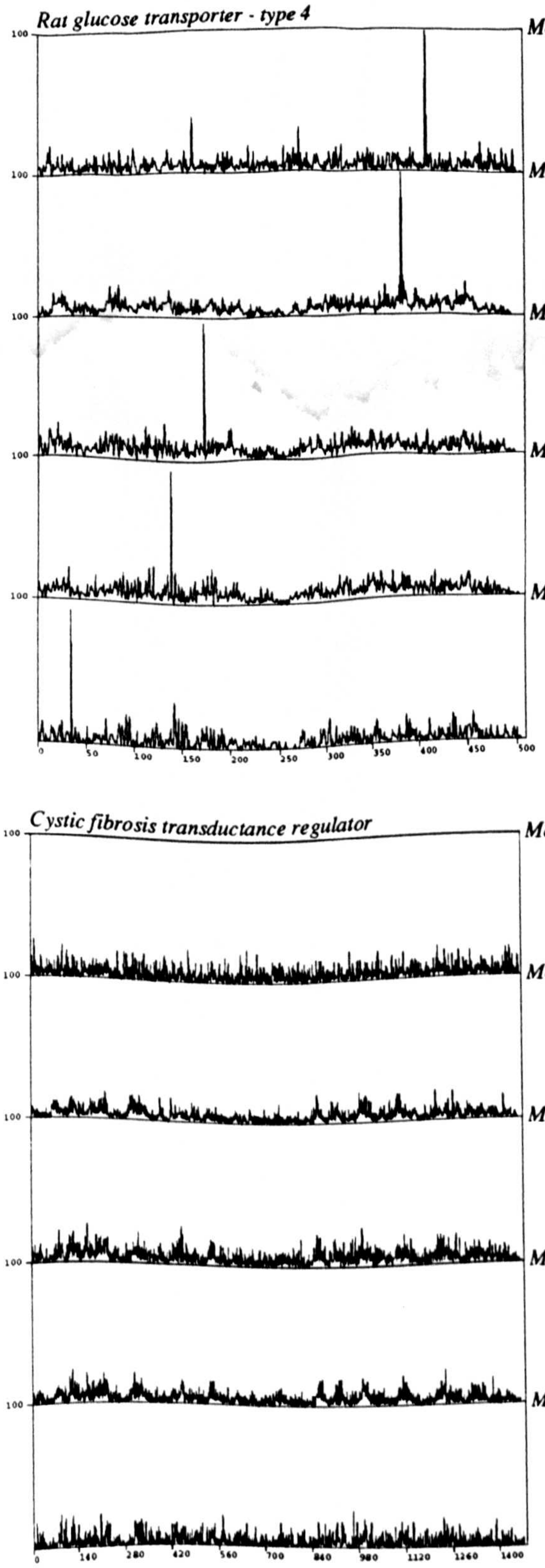


Figure 4.6b - Individual sequences scanned with the sugar transporter motifs. The lower graph is included as a control.

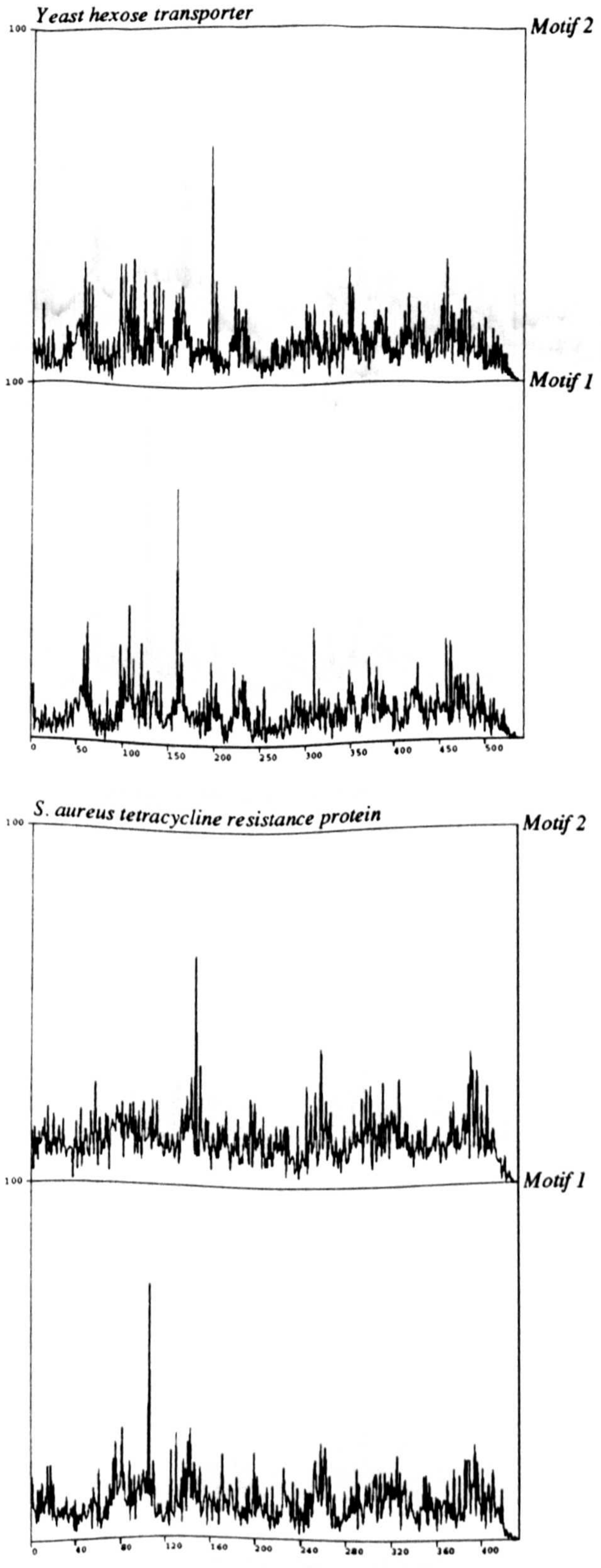


Figure 4.7a The super-family motifs used to scan individual sequences

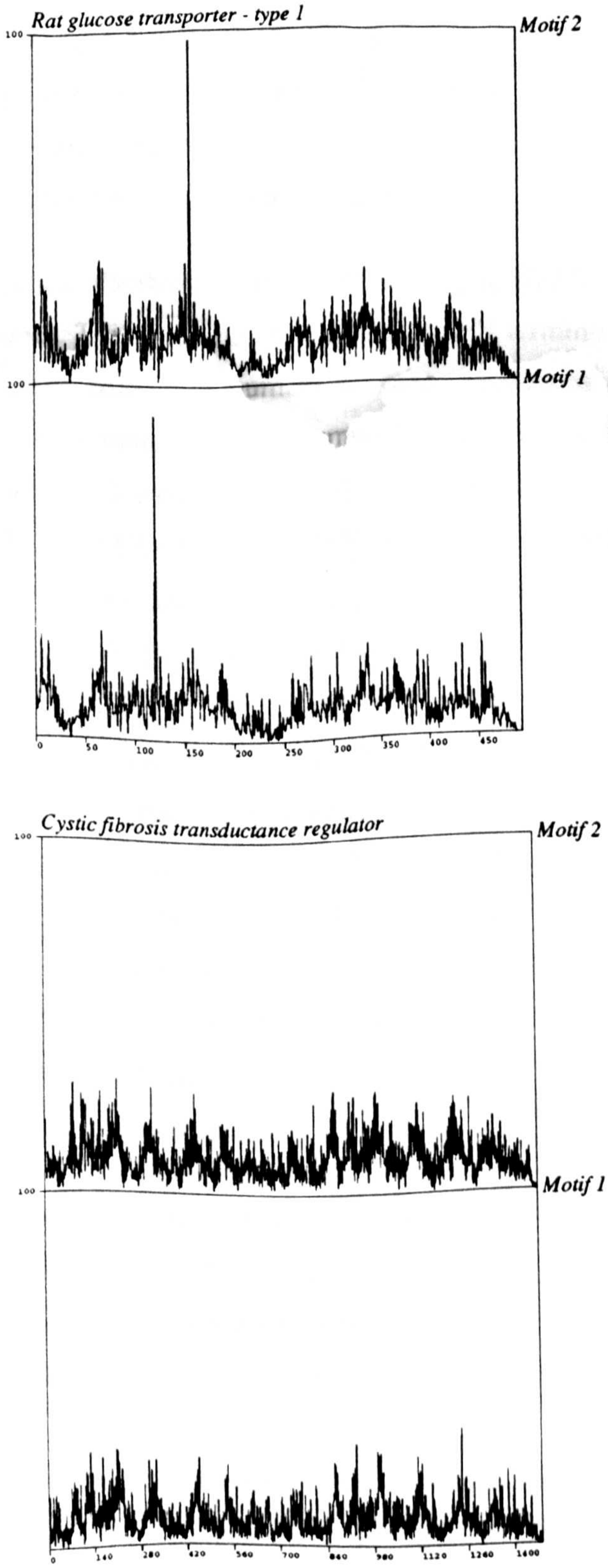


Figure 4.7b Individual sequences scanned with the final super-family motifs. The lower graph is included as a control.

4.5 Discussion

Figures 4.6 and 4.7 indicate the strong discriminating efficiency of the selected motifs, in that members of the family score highly while non-members do not. In the case of the sugar transporters, all the sequences known to belong to the family were shown in the final compound feature index.

One sequence, database code GTR1_PIG (pig GLUT1) is shown in the four features column. This sequence has a truncated N terminus so lacks motif one. In the two features column a number of other sequences are also shown. These include other proton antiport/symports, ie STMBAHBRP (*Streptomyces hygroscopicus* ORF3 transport protein), KGTP_ECOLI (*E. coli* alpha-ketoglutarate permease), TCR1_ECOLI and ECOTN10 (*E. coli* tetracycline resistance protein). The other proteins are also membrane proteins but are not thought to be members of this family, ie HYIN_PSESS (indoleacetamide hydrolase from *Pseudomonas syringae*), VGLM_PHV (Prospect Hill virus M polyprotein), S1049541 (Polysulphide reductase chain c) and ECOPNT1 and PNTB_ECOLI (both *E. coli* transhydrogenases). This latter sequence was further examined to check whether there was any significant relationship between this protein and the sugar transporters, a significantly hydrophilic C-terminal region suggested that any similarity was at a low level. Also shown in the two features column are the sequences STYCITCA and STYCITCB (Citrate/Sodium symport proteins from *Salmonella dublin* and *Salmonella pullorum* respectively), both of these sequences match with motifs two and four.

All these sequences, apart from PNTB_ECOLI, ECOPNT1, VGLM_PHV, and S1049541, match with motif two which is the first of the hydrophobic segments. The two related permeases also match with motif five, while both tetracycline resistance proteins also match with motif three.

A comparison between the PROSITE codes (Bairoch, A.) SUGAR_TRANSPORT_1 and SUGAR_TRANSPORT_2 and the final compound feature index for the sugar transporters (Figure 4.3) suggests that the motifs selected are more efficient than the PROSITE patterns. While the patterns have a large number of false positives and some false negatives, this is not true of the motifs used in this study. The PROSITE patterns do, however, suggest that the sequences PH84_YEAST (Yeast phosphate transporter), R137_YEAST (Yeast metabolite transporter) and KGTP_ECOLI (alpha ketoglutarate transporter from *E. coli*) are members of the sugar transport family while the results presented here show these sequences as members of the symport/antiport super-family, but not as part of the sugar transporter subset. This may be due to the PROSITE pattern being too flexible to restrict true matches to the sugar transporter subset of sequences while being too flawed to include the large range of sequences shown in the super family. A number of sequences are also shown in this study that do not appear in the PROSITE pattern description, LEID1TRA and LEID2TRA (glucose transporters from *Leishmania donovani*) being two examples. However, some of these may not be shown in the patterns because the PROSITE database is compiled using the smaller subset of sequences found in SWISS-PROT rather than the composite OWL protein sequence database.

With the full antiport/symport, one known member of the family was not shown in the final hitlists (MAL6_YEAST - Yeast maltose permease). This sequence lacks the conserved arginine residue in motif one shown in the other members of the family, this residue being exchanged for a asparagine. In addition, a number of previously identified family members were only shown to match with motif one. These sequences were database codes KGTP_ECOLI (alpha-ketoglutarate permease from *E. coli*), PH84_YEAST (Yeast phosphate transporter), R137_YEAST (Yeast probable metabolite transporter), NORA_STAA (quinolone resistance protein from *S. aureus*), BICA_ECOLI (*E. coli* bicyclomycin resistance protein), LEID1TRA (*Leishmania donovani* glucose transport protein), LACP_KLULA (*Kluyveromyces lactis* lactose permease), the myo-insitol transporters from yeast (database codes A40538 and B40538) and the multidrug resistance protein from *E. coli* (database code EMRB_ECOLI).

Distance criteria was imposed after every database scan during this study as only two motifs were selected. Using this method, in addition to the use of the hydropathy plots mentioned above, it was found that all noise (ie false hits) were removed from the motif sets but all the true hits were retained. As mentioned previously, these distance rules were derived from the initial alignment and were not user-defined to ensure objectivity.

Examination of the final compound feature index for the super-family motifs shows the wide occurrence of these proton symport/antiport proteins across a range of organisms from eukaryotes to prokaryotes and from animals to plants. They perform a particularly large number of functions including sugar transport and the efflux of antibiotics leading to drug resistance. A number of sequences are also shown in figure 4.5 that were not previously reported to be members of this family. These include the *Nicotiana tabacum* monosaccharide transporter (database code TOBMST1) (Sauer, N. et al. (1992)), the methylenomycin A resistance protein from *Bacillus subtilis* (database code S22742) (Putzer, H. et al. (1992)), and the tetracenomycin C resistance protein from *Streptomyces glaucescens* (database code M225633S1) (Guilfoile, P.G. and Hutchinson, C.R. (1992)). In addition the rat amine transporters, database codes RATCGAT and RATSVAT, are shown in the final compound feature index. These proteins are responsible for the ATP-dependent accumulation of biogenic amine neurotransmitters into the secretory organelles of neurons and a number of other cells (Erikson, J.D. et al. (1992)). At the time of this study, these proteins were also new to this family and have since been confirmed by other workers (Henderson, P.J.F personal communication, Linial, M. (1993)). Also shown to match with both motifs is the *Candida maltosa* cycloheximide resistance protein (database code YSACHRA) which the sequencing authors claimed had no significant similarity with any other database sequences (Sasnauskas, K. et al. (1992)) and also a putative transport protein from Acinetobacter calcoaceticus. In the case of the former sequence this similarity has also been confirmed by other workers.

The two studies have also drawn attention to the fact that the fourth and fifth putative transmembrane regions seem to be very significant, as both are quite well conserved across the whole family (particularly the fourth transmembrane segment which contains a conserved arginine residue). This suggests that they have a major structural or functional role in the protein. Other workers have also suggested that the N-terminal region is involved with the basic function of the protein while the C-terminal region is involved in specificity (Rouch, D.A. et al. (1990)).

The study described above which was limited to the sugar transporters also suggests this may be true. In this case, motifs two and three were selected from the fourth and fifth transmembrane segments while motifs one, four and five were from the first, tenth and eleventh transmembrane regions respectively. These latter two motifs may have a role in the specificity of the sugar transporters as experimental data from inhibitor and photo-affinity labelling techniques have also suggested that the sugar binding sites are located in the C-terminal region of the proteins (Baldwin, S. and Henderson, P.J.F. (1989)). The sequences shown in the two features column that are members of the family all match with motif two (the fourth transmembrane segment) and one other of the motifs, demonstrating the commonality of the fourth transmembrane segment. Motif one (derived from transmembrane segment one) also may have an important structural and or functional role as it is conserved only across the sugar transporters.

It has also been suggested that the symmetrical nature of the proposed structure of these proteins may be due to gene duplication (Rubin, R.A. et al. (1990)), ie each protein is composed of two copies of the same six membrane spanning regions. While there are some repeats to be seen, the PESPRY motif being particularly noticeable in the hydrophilic loop between the sixth and seventh putative transmembrane segment and in the C-terminus of some sugar transporters, sequences generally appear only once in the hitlists produced for each motif. If duplication had occurred, it would be expected that each sequence would appear twice in the hitlists, once for the N-terminus and once for the C-terminus. This suggests that the second set of six membrane spanning segments are either not simple repeats of the first, or that the degree of similarity between the two has become significantly lower over evolutionary time. If the latter is the case, the results above suggest that the N-terminus of the protein has remained the most conserved while the C-terminus has possibly evolved for different substrate specifities as the motifs that are conserved over the whole family are from the n-terminal region.

A number of transport proteins have also been shown to have hydropathy profiles that are very similar to that of the proteins described here in that they appear to have twelve transmembrane regions. Probably the best known of these is the LACY lactose transporter from *E. coli*, which was the first proton symport to be well characterised. Some authors have suggested that this protein is a member of the family described in this chapter (albeit with only slight similarity) (Marger, M.D. and Saier, M.H. (1993)), while others have found no evidence for its inclusion (Bairoch, A., Griffith, J.K. et al. (1992)). During the course of this study LACY was not shown in any of the hitlists produced by the database scans which suggests that there are at least two families of proteins which have similar transport mechanisms. The results presented here suggest that these two families probably evolved by the process of convergent evolution, ie there was no common ancestor protein.

4.6 Conclusion

This research has indicated that the fourth and fifth transmembrane segments probably have a crucial structural or functional role as they are are relatively well conserved across the whole family. This is especially true of the fourth putative transmembrane segment.

In the case of the sugar transporter subset, the results suggest that the tenth and eleventh transmembrane regions also have an important function, this is supported by biochemical evidence as the sugar binding site has been shown to be in this region of the sequence. It may be also be possible, as these comprise almost all of the symporter subset of this family, that these regions of the sequence may also have a role in defining whether a protein is an antiporter or a symporter as they do not appear to be conserved in the antiporters. The same may also be true for transmembrane segment one.

It is also clear from the results described above that this family of transporters is very diverse and widespread being present in both prokaryotes and eukaryotes, a number of new members are also identified extending the family further.

The hitlists produced for each motif indicate that the C-terminal region is probably not a simple repeat of the N-terminus, if there was an ancestral gene duplication then the sequence similarity has decreased over evolutionary time and is now imperceptible in most family members.

4.7 References

Bairoch, A. PROSITE database, University of Geneva

Baldwin, S., Henderson, P.J.F., Homologies between sugar transporters from eukaryotes and prokaryotes. Ann. Rev. Physiol. **51** (1989) pp459-471

Engelman, D.M., Steitz, T.A., Goldman, A., Identifying non-polar transbilayer helices in amino acid sequences of membrane proteins. Ann. Rev. Biophys. & Biophys. Chem. **15** (1986) pp321-353

Erickson, J.D, Eiden, L.E., Hoffman, B.J., Expression of a reserpine-sensitive vesicular monoamine transporter. Proc. Natl. Acad. Sci. USA **89** (1992) pp10993-10997

Griffith, J.K., Baker, M.E., Rouch, D.A., Page, M.G.P., Skurray, R.A., Paulsen, I.T., Cahter, K.F., Baldwin, S.A.,Henderson, P.J.F, Membrane transport proteins: implications of sequence comparisons. Curr. Opin. Cell Biol. **4** (1992) pp684-695

Henderson, P.J.F., Baldwin, S.A., Cairns, M.T., Bambos, M., Charalambous, H., Dent, C., Gunn, F., Liang, W., Lucas, V.A., Martin, G.E., McDonald, T.P., McKeown, B.J., Muiry, J.A.R., Petro, K.R., Roberts, P.E., Shatwell, K.P., Smith, G., Tate, C.G., Sugar-cation symport systems in bacteria. Int. Rev. Cytol. **137** (1992) pp149-207

Guilfoile, P.G., Hutchinson, C.R., Submitted to EMBL data library, February 1992

Kayano, T., Fukumoto, H., Eddy, R.L., Fan, Y., Byers, M.G., Shows, T.B., Bell, G.I., Evidence for a family of human glucose transporter like proteins. J. Biol. Chem. **263** (1988) pp15245-15248

Levy, S.B., Active Efflux Mechanisms for Antimicrobial Resistance, Antimicrob. Agents and Chemotherapy **36** (1992) pp695-703

Linial, M., Vesicular transporter joins the major facilitator superfamily (MFS). TIBS letters **18** (1993) pp248-249

Maiden, M.C.J., Davies, E.O., Baldwin, S.A., Moore, D.C.M, Henderson, P.J.F, Mammalian and bacterial sugar transport proteins are homologous. Nature **325** (1987) pp641

Marger, M.D., Saier, M.H., A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport. TIBS **18** (1993) pp13-20

Mitchell, P., Molecule, group and electron translocation through natural membranes. Biochem. Soc. Symp **22** (1963) pp142-169

Muekler, M., Caruso, C., Baldwin, S.A., Panico, M., Blench, I., Morris, H.R., Allard, W.J., Lienhard, G.E., Lodish, H.F., Sequence and Structure of a Human Glucose Transporter. Science **22** (1985) pp941-945

Nazos, P.M., Antonucci, T.K., Landick, R., Oxender, D.L., Cloning and characterisation of livH, the structural gene encoding a component of the leucine transport system in *Escherichia coli*. J. of Bacteriol. **166** (1986) pp565-573

Putzer,H., Gendron, N., Grunberg-Manago, M. Submitted to GenBank data bank, June 1992

Rouch, D.A., Cram, D.S., DiBerardino, D., Littlejohn, T.G., Skurray, R.A., Efflux mediated antiseptic resistance gene qacA from *Staphylococcus aureus*: common ancestry with tetracycline and sugar transport proteins. Molec. Microbiol. **4** (1990) pp2051-2062

Rubin, R.A, Levy, S.B., Heinrikson, R.L., Kezdt,F.J., Gene duplication in the evolution of the two complementing domains of gram-negative bacterial tetracycline efflux proteins. Gene **87** (1990) pp7-13

Sasnauskas, K., Jomantiene, R., Lebediene, E., Lebedys, J., Januska, A., Janulaitis, A., Cloning and sequence analysis of a Candida maltosa gene which confers resistance to cycloheximide. Gene **116** (1992) pp105-108

Sauer, N., Stadler, R. Submitted to EMBL data library, June 1992.

Walmsley,A.R., The dynamics of the glucose transporter, TIBS **13** (1988) pp226-231

Chapter Five

5.1 Introduction

During the course of the research described in this thesis it became necessary to develop software which, while not being directly connected to the database scanning procedures, was essential for the definition of discriminating motifs or for sequence alignment. This chapter describes the most pertinent of these programs while figure 5.1 illustrates their place in the sequence analysis scheme.

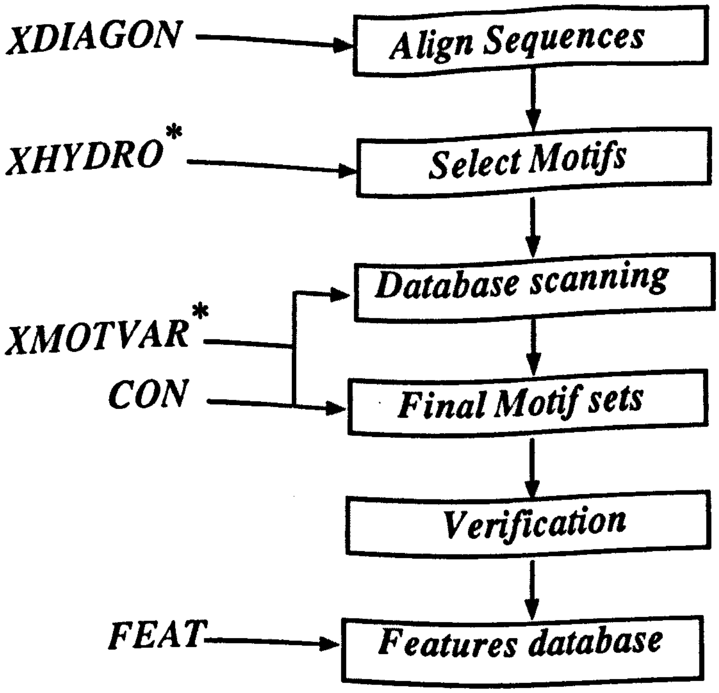


Figure 5.1 illustrates the stages of discriminator definition and refinement where the programs described in this chapter are most useful. An asterix indicates that a GL version of the program is available for Silicon Graphics machines.

5.2 Programming Details

All the programs described were written in C and are portable across most platforms but in Leeds are generally only available on VAX, Silicon Graphics and SUN clusters. Where applicable, graphics displays are produced using the standard X and Motif libraries, although versions of some programs are also available that take advantage of the advanced graphics capabilities of the GL library found on Silicon Graphics machines. Postscript output in monochrome and colour is produced from drivers written by the author. The individual programs with example output are described below.

5.3 Diagon Plots

The simplest, yet probably the most useful, way of comparing two sequences is to compare every residue from one sequence with every residue from the other, then plotting this data in the form of a two-dimensional array of points (a Diagon plot). Stretches of residues that are common to both sequences are represented by diagonal lines on this plot, insertions and deletions are indicated by offsets from the main diagonal. These plots are particularly useful when initially aligning two sequences and give an indication of how many gaps and where the gaps should be inserted into each sequence. XDIAGON produces such diagon plots for both nucleic acid and protein sequences.

When the user initiates the XDIAGON program, a prompt is produced asking for two sequence names. These may be files (in NBRF/PIR format) or database codes, in the latter case the sequence is extracted directly from the OWL database (Bleasby, A.J., personal communication). The whole length of the sequences may be compared, or the user may define the start and end residues of the segment of interest, XDIAGON has the ability to produce square or rectangular graphs reducing the inherent distortion in the plot when one sequence is much longer than the other.

A number of different comparison methods are also available, these being reduced alphabet, identities, MDM78 (Dayhoff, M.O. (1978)) and user defined. If the reduced alphabet option is selected then residues are grouped together and individual residues within each group are treated as being identical, for instance both aspartic acid and glutamic acid are of the same group (Figure 5.2 illustrates the full residue groupings used). This option is particularly useful when sequences have very low similarity. The identity option only generates a point if the individual residues being compared are identical while the MDM78 option uses a substitution matrix with a user-defined threshold level of similarity to generate points. In addition to the MDM78 matrix, the user-defined option allows the user to supply a matrix of their choice.

Asp (D), Glu (E)	- Acidic group
Cys (C)	- Cysteine
Pro (P)	- Proline
Gly (G)	- Glycine
Ala (A), Ile (I), Leu (L), Met (M), Val (V)	- Hydrophobic group
Asn (N), Gln (Q), Ser (S), Thr (T)	- Polar group
Phe (F), Trp (W), Tyr (Y)	- Aromatic group
His (H), Lys (K), Arg (R)	- Basic group

Figure 5.2 XDIAGON residue groupings. Those residues in the same group are treated as being identical for the reduced alphabet option.

A windowing facility is also included within XDIAGON to reduce the noise in a plot. For instance if a window length of ten residues is specified with a stringency of four, points are only plotted if there are at least four other residue matches from the section of sequence five residues either side of the position being compared. The stringency and window values are often very difficult to define optimally, a stringency that is too high will result in possibly interesting data being missed while a stringency that is too low will allow noise to mask data. To overcome this problem, the author has developed a colour option for XDIAGON. The user defines the window length, then all the points in that window are plotted in a colour dependent on how many matches were found in the window. If the default colours are used, then red is used for points in windows where there are a lot of matches through to blue which is used to plot points with only a small number of matches within the window. The number of colours and the actual colours used may be defined by the user if desired. Figure 5.3 illustrates a colour diagon plot produced by the self comparison of the human multidrug resistance protein, the regions of similarity (red and yellow) are easily identifiable from the noise which is coloured blue. Normally when a sequence is compared with itself a single line on the main diagonal is produced but, as the sequence used for figure 5.3 has internal repeats, there are also lines shown offset from the main diagonal.

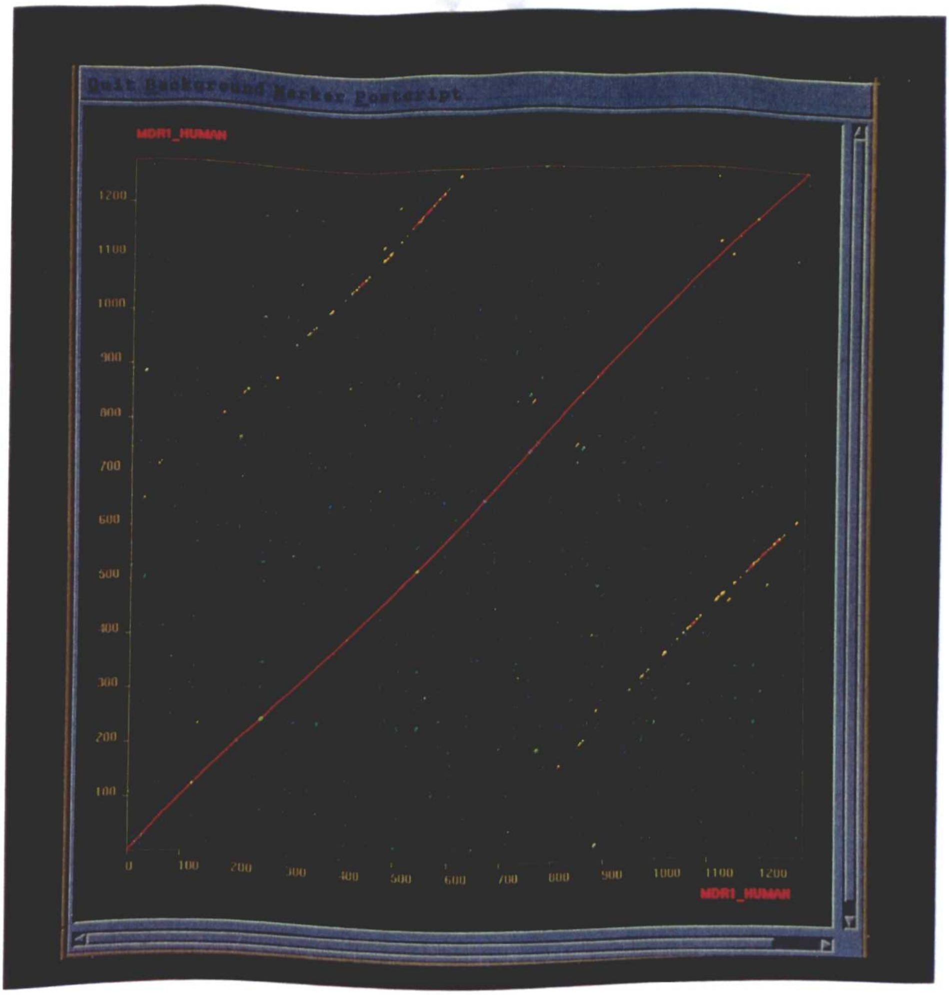


Figure 5.3 A XDIAGON plot illustrating the self-comparison of the human multidrug resistance protein.

5.4 Amino Acid Physicochemical Properties.

Many properties that are inherent in protein sequences may be of use when initially selecting motifs from a sequence alignment, particularly in those situations where no three-dimensional structure of the proteins of interest is know. These include hydropathy measurements that allow the putative transmembrane segments and core regions of protein sequences to be identified with reasonable confidence and secondary structure propensities. XHYDRO allows a number of graphs (up to a maximum of four) to be plotted on the same sheet of paper or screen so that properties can be related easily to each other. The program takes as input either a file name or database code. If the file contains multiple, pre-aligned sequences, then the whole alignment is used to calculate the desired properties. Other programs are available that plot amino acid properties (for example Mandler, J. (1988)), but these tend to be limited by little portability, poor hard copy facilities, usually only take a single sequence as input rather than an alignment and have a limited number of graph types. The types of graph that can plotted by XHYDRO were carefully chosen to extract the maximum possible useful information from a sequence or alignment and are described below, Figure 5.4 shows typical XHYDRO output.

5.4.1 Hydropathy.

Hydrophobic interactions are a major factor in the structural stability of proteins since the interactions between non-polar residues and water are so unfavourable there is a strong tendency for these residues to aggregate together at the core of the molecule in globular proteins or to cluster in the hydrophobic environment of the membrane in the case of membrane transport proteins. In both cases, especially the latter, hydropathy plots are particularly useful in determining the possible structure of a protein, for instance hydropathy plots may indicate how many putative transmembrane segments a protein may have. In some cases similar hydropathy plots may also indicate whether two proteins are structurally related even though their sequence homology may be very low or non-existent, as in the case of the *E. coli* lactose transporter and human erthyrocyte glucose transporter mentioned in an earlier chapter. Also, as these transmembrane segments are so crucial to the structure of channels, they are very often the most conserved regions in a sequence alignment. There is some debate about which hydropathy scale is most appropriate in a given situation (Crimi, M. and Esposti, M. D. (1991)), therefore a number of different scales are provided as described below. All the hydropathy options utilise a windowing algorithm to reduce the amount of noise in the final graphs.

i) Eisenberg hydropathy scale (Sweet R.M. and Eisenberg, D. (1983)). This is a consensus scale derived from five other scales, the most notable being the scale defined by Wolfenden et al based on the vapour pressures of side chain analogues and the scale of Janin which was derived by counting the buried and exposed residues in globular proteins. A window length of nine residues is used.

ii) Kyte and Doolittle scale (Kyte, J. and Doolittle, R.F. (1982)). The hydropathy scale described by the above authors was based on an amalgam of data derived from experimental measurements of the free energy of transfer between various phases along with some intuitive adjustment of the final values when these were contradictory. A window length of twenty is used rather than that suggested by the authors as this has been shown to be the most effective for identifying transmembrane segments.

ii) Transmembrane hydropathy scale (Engelman, D.M. et al. (1986)). This scale was specifically designed to identify the transmembrane sections of proteins and takes into account the specific conditions that occur in an α-helical polypeptide in a low dielectric environment, for example the water-accessible surface area of each residue type in such a conformation. The resultant hydrophobicity values are generally the first choice of the author of this thesis when studying possible membrane proteins, although if transmembrane sequences do exist that are of beta conformation the scale may be less useful due to the way it was initially calculated.

5.4.2 Positional Variability.

This option allows for the identification of the most conserved parts of a sequence alignment (by comparing every residue at a particular position with every other residue at that position), these regions being the segments of the alignment most suitable for database searching. Most of the widely used substitution matrices are based on observations made from sequence alignments which may give biased results as distantly related sequences are often difficult to align, therefore it was decided to use a matrix derived from the superimposition of three-dimensional structures (Risler J.L. et al. (1988)) as the default, although this can be changed as desired by the user. To produce this matrix, Risler et al. superimposed the three-dimensional structures from eleven protein families and if the c-alpha atoms from each chain were less than 1.2 angstroms apart at a particular position the appropriate amino acids were considered to be substitutable by the other. This matrix has been shown by Risler et al. to produce more accurate alignments of distantly related proteins than other widely used matrices. If only this option is selected, then XHYDRO displays the graph as a histogram otherwise the positional variability is displayed as a normal graph, the positions with the highest values being the most variable.

5.4.3 Solvent Accessible Area.

This option utilises a scale derived from measuring the mean solvent accessible surface area of each of the residue types in twenty three folded proteins (Rose, G.D. et al. (1985)). This option may be used in conjunction with the hydropathy scales as it has been demonstrated that those residues that have the highest solvent accessible surface area are the most hydrophilic, hydrophobic residues tending to have low solvent accessible surface area values. A window length of five residues is used to reduce the noise in the final plot.

5.4.4. Flexibility.

The dynamic properties of a protein are essential for its function, for instance in substrate recognition and in conformational changes after substrate binding. The amino acid residues that are located in the most mobile regions of a protein are generally the most hydrophilic and have the smallest volumes (so are less likely to be involved in interactions with surrounding residues). The scale used (Ragone, R. et al. (1989)) takes advantage of these properties and is designed to identify the most flexible residues in an amino acid sequence. Plots produced using this scale where shown by those authors to be very similar to the appropriate graphs of B factors, indicating a high degree of accuracy. A window length of ten residues is used, rather than five as suggested by the authors, to reduce the amount of noise in a plot.

5.4.5 Garnier-Osguthorpe-Robson Secondary Structure Prediction.

Although secondary structure prediction techniques tend to be notoriously unreliable, the Garnier-Osguthorpe-Robson (GOR) technique (Garnier, J. et al. (1978)) is generally considered to be amongst the more accurate and may be useful in a number of situations, for instance if a user needed to select the putative alpha-helical segments of a sequence as motifs. The GOR algorithm is based on moving sixteen-residue, overlapping windows along the test sequence and calculating the propensity for the possible conformations of a particular residue type at a particular location in this window. XHYDRO has the ability to produce a secondary structure prediction using an alignment as well as a single sequence, the accuracy of the prediction has been shown to be improved if the former is used (Thornton, J.M. et al. (1991), Zvelebil, M.J. et al. (1987)). The graph produced displays the propensities for the four possible conformations as lines drawn in different colours.

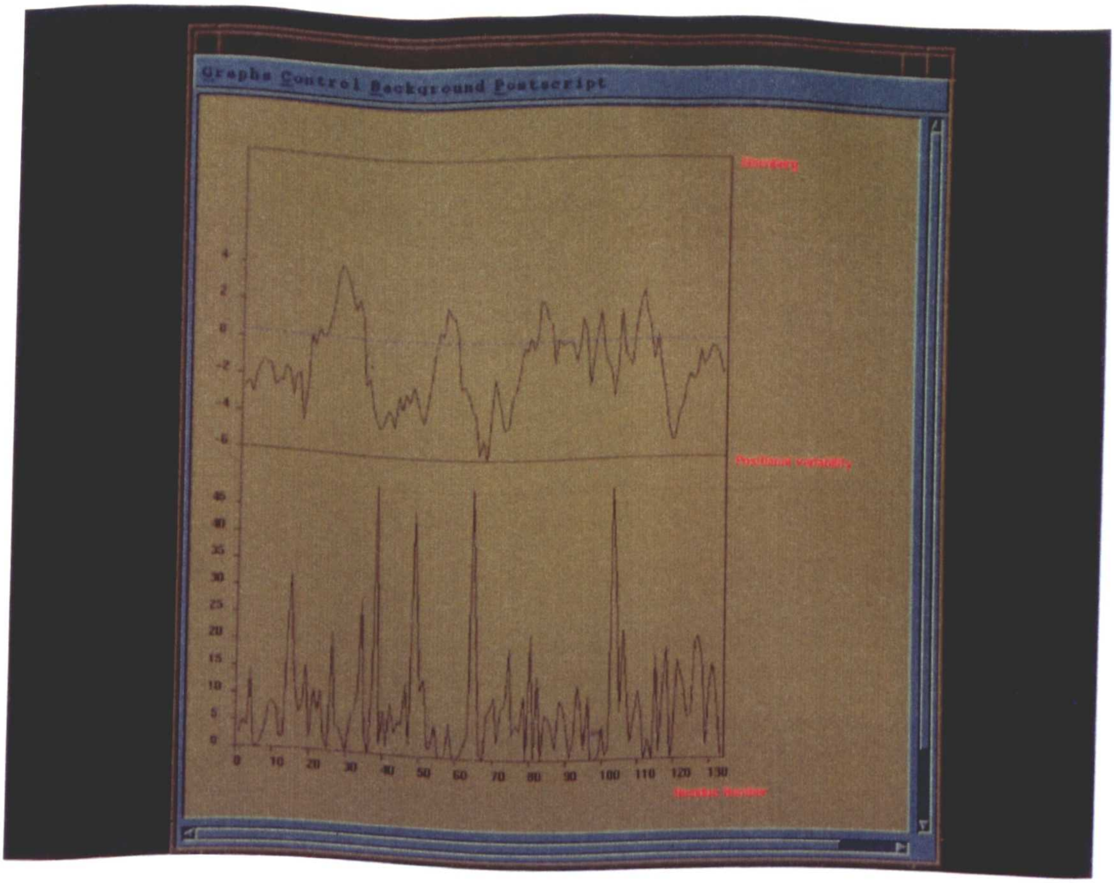


Figure 5.4 XHYDRO output from an alignment of lysozyme and lactalbumin sequences. The top graph shows the Eisenberg hydropathy graph while the bottom graph shows positional variability.

5.5 Motif Positional Variability and Consensus Sequence

These two programs were written to aid the visualisation and manipulation of the sequence information contained within large motif files of the format described in chapter two.

XMOTVAR displays the positional variability of a number of motifs on the same screen or piece of paper in the form of a coloured histogram or ordinary graph, the residues showing the lowest and highest degrees of conservation are thus easily identified. The variability can be calculated either by using a substitution matrix (the matrix defined by Risler et al. (as described above) is the default, although this can be changed by the user) or by simply counting the number of different residue types at each position. This program is especially useful for identifying those residues that contribute the most to the discriminating efficiency of a motif set and also for refining motifs during the iterative database scanning process as variable residues near the ends of a motif may be identified and removed. Example output from this program is shown in figure 5.5.

CON also takes a list of motif file names as input and then produces a consensus sequence from each motif. This program is useful for converting motif files into PROSITE-style patterns and also for reducing the data contained in large motif files into a more manageable format. An example motif file and CON output is shown in figure 5.6.

```
lac_casite.mot
CDITCDKFLD
CGISCDKFLD
CDITCDKFLD     ******************************************
CGISCDKFLD     1) Motif = lac_casite.mot
CGISCNKFLD     [C] [D G N] [I] [S T] [C] [D N] [K] [F L] [L] [D]
CDISCDKFLD     ******************************************
CNISCDKFLD
CDISCDKLLD
```

Figure 5.6 illustrates a sample motif file (derived from nine lactalbumin sequences) on the left with corresponding CON output on the right. The square brackets in the CON output delimit each residue position.

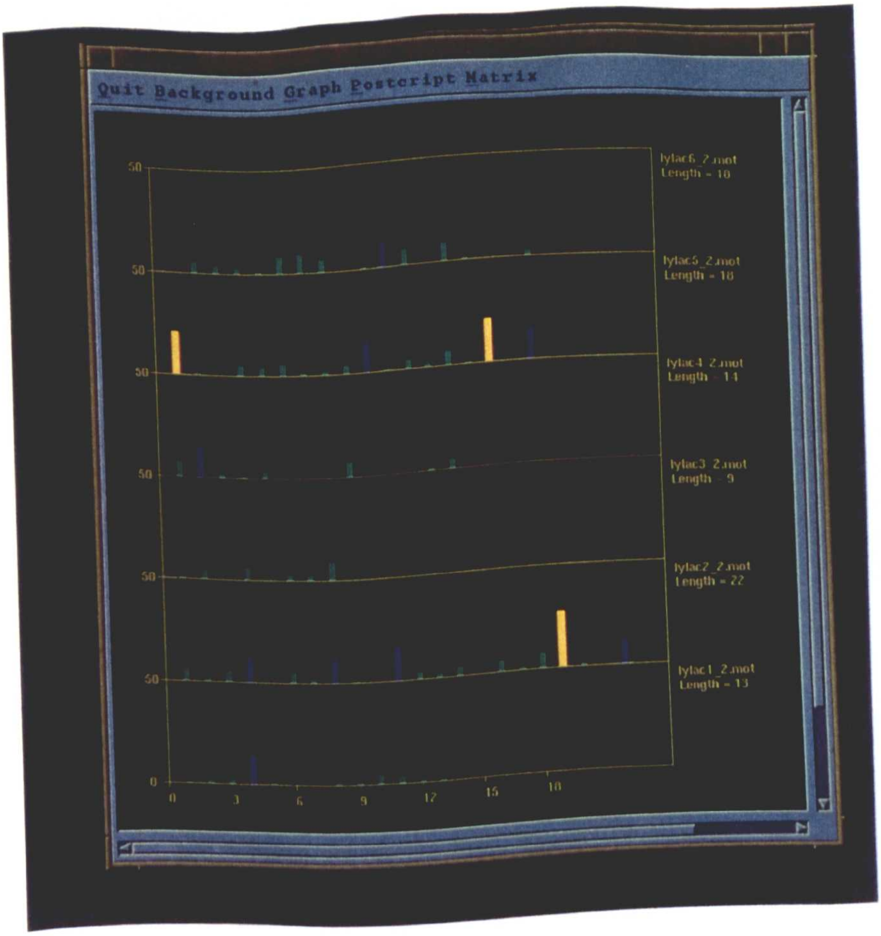


Figure 5.5 Typical output from the XMOTVAR program.

5.6 PRINTS Database Scanning

When the motif sets have been verified and checked then they are ultimately entered into the PRINTS database. There exists a special interrogation language for this database (SMITE) which allows users to view PRINTS entries but there was still a need for a algorithm which compared a test sequence with all the motifs in the features database, the FEAT program fulfils this need. FEAT initially prompts the user for a sequence file name or database code. As very large numbers of sequences can be processed the file name given may be that of a large database such as OWL, FEAT is therefore a useful tool for the rapid identification of any new member of a family whose discriminating motifs exist in the PRINTS database and may also be of use in updating the PRINTS database with each new release of the OWL sequence database. The scanning method is user-defined and may be novel or simple (see chapter two), the user can also select other parameters such as allowing motifs to overlap or which PRINTS database entries are to be checked. Figure 5.7 illustrates typical FEAT output.

```
Sequence code is >P1;LYC_BOVIN
sequence number = 1
Sequence length is 147 residues
Motif = 1
Motif length = 11
Code = LYSLACT
Sequence = FERCELARTLK
Position = 20 to 30
Score = 94.98%

Motif = 2
Motif length = 13
Code = LYSOZYME
Sequence = KWESSYNTKATNY
Position = 50 to 62
Score = 91.73%
```

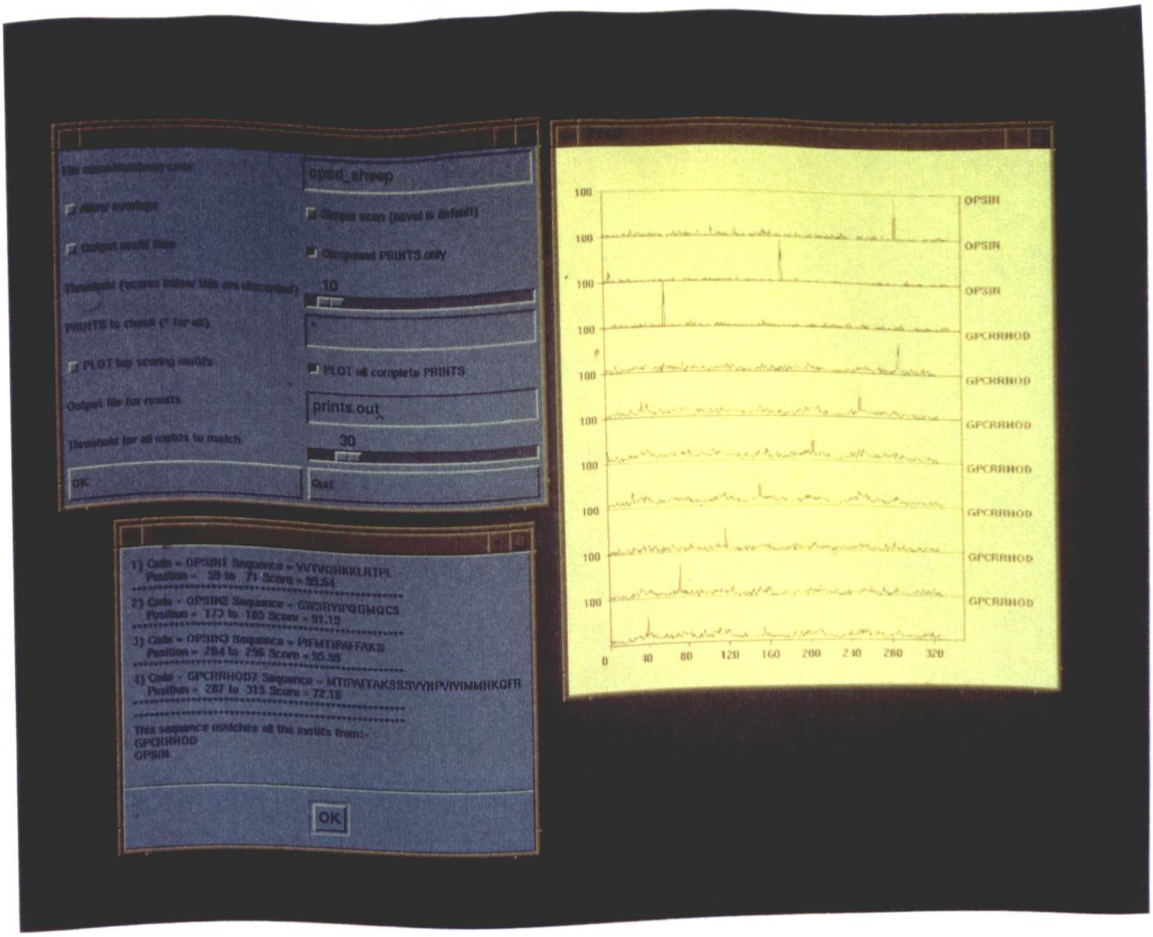
(Continued on the next page)

```
Motif = 3
Motif length = 17
Code = LYSLACT
Sequence = STDYGIFQINSKWWCND
Position = 68 to 84
Score = 98.12%

Motif = 4
Motif length = 16
Code = LYSOZYME
Sequence = RELMENDIAKAVACAK
Position = 99 to 114
Score = 70.44%

Motif = 5
Motif length = 10
Code = LYSOZYME
Sequence = QGITAWVAWK
Position = 120 to 129
Score = 85.81%

Motif = 6
Motif length = 12
Code = LYSLACT
Sequence = GITAWVAWKSHC
Position = 121 to 132
Score = 88.26%
```

Figure 5.7 The output produced by searching the PRINTS database (version 4) with the sequence LYC_BOVIN (Bovine lysozyme).

The figure above indicates that the bovine lysozyme has significant similarity with two entries in the PRINTS database, ie LYSOZYME and LYSLACT (lysozyme-c super-family motifs). FEAT allows the user to specify a PRINTS database entry to search, therefore it would be possible to examine each of these two entries separately in more detail.

The author has also written a Xlib/Motif based version of the PRINTS database scanning program which produces PLOT output to graphically illustrate how well each matching motif scores. This program has the ability to extract and write to a file motifs from the PRINTS database allowing a user to manipulate this data further if desired. Example output from this program is shown in figure 5.8.



File name/database code: opsd_sheep

☐ Allow overlaps

☐ Output motif files

☐ Simple scan (novel is default)

☑ Compound PRINTS only

Threshold (scores below this are discarded): 10

PRINTS to check (* for all): *

☐ PLOT top scoring motifs

☑ PLOT all complete PRINTS

Output file for results: prints.out

Threshold for all motifs to match: 30

```
1) Code = OPSIN1 Sequence = VVTVQHKKLRTPL
   Position =  59 to  71 Score = 99.64
*************************************************
2) Code = OPSIN2 Sequence = GWSRYIPQGMQCS
   Position =  173 to  185 Score = 91.15
*************************************************
3) Code = OPSIN3 Sequence = PIFMTIPAFFAKS
   Position =  284 to  296 Score = 95.99
*************************************************
4) Code = GPCRRHOD7 Sequence = MTIPAFFAKSSSVYNPVIYIMMNKQFR
   Position =  287 to  313 Score = 72.16
*************************************************
*************************************************
This sequence matches all the motifs from:-
GPCRRHOD
OPSIN
```

Figure 5.8 Output from XPRINTS. In this case the sequence of sheep rhodopsin was given as input.

5.7 References

Crimi, M., Esposti M.D., Structural predictions for membrane proteins: the dilemma of hydropathy scales. TIBS **16** (1991) pp119

Dayhoff, M.O. (ed.) Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington D.C., U.S.A.

Engelman D.M., Steitz, T.A., Goldman, A. Identifying non-polar transbilayer helices in amino acid sequences of membrane proteins. Ann. Rev. Biophys. & Biophys. Chem. **15** (1986) pp321-353

Garnier, J., Osguthorpe, D.J, Robson, B. Analaysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. **120** (1978) pp97-120

Kyte, J., Doolittle, R.F., A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157** (1982) pp105-132

Mandler, J. Hystruc - Hydropathy and secondary structure prediction. CABIOS **4** (1988) pp309

Ragone, R., Facchiano, F., Facchiano, A., Facchiano, A.M., Colonna, G. Flexibility plot of proteins. Protein Engineering **2** (1989) pp497-504

Risler J.L., Delorme M.O., Delacroix H., Henaut A. amino acid substitutions in structurally related proteins, a pattern recognition approach. J. Mol. Biol. **204** (1988) pp1019-1029

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H., Hydrophobicity of amino acid residues in globular proteins. Science **229** (1985) pp834-838

Sweet, R.M., Eisenberg D., Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J. Mol. Biol. **171** (1983) pp479-488

Thornton, J.M., Flores, T.P., Jones, D.T., Swindells, M.B., Prediction of progress at last. Nature **354** (1991) pp105-106

Zvelebil, M.J, Barton, G.J., Taylor, W.R., Sternberg, M.J.E. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J. Mol. Biol. **195** (1987) pp957-961

Chapter 6
VISTAS

6.1 Summary

This chapter will describe the methodology and the use of the VISTAS package written by the author, along with mention of the related programs ALIGN and XALIGN.

6.2 Introduction

The pre-existing sequence analysis software at Leeds was not only tied to a VMS cluster, but in the authors opinion was very user unfriendly as several different programs were required to perform relatively simple procedures. Users were also required to edit a number of files for input to each routine. With the availability of modern workstations it is possible to produce very user friendly and powerful programs for sequence analysis and this is the avenue the author decided to pursue, although at an early date it was decided that any source code produced was to be made as portable as possible. In fact the only restrictions to portability are related to the graphics libraries used. This portability issue will be discussed later.

During the research of protein sequence motifs, it was noticeable that there was very little interface between multiple sequence alignments and the increasing amount of structural information that was becoming available. For instance, it would be useful to display sequence motifs on the three-dimensional structure of a protein, if one was available. Also, in some cases it would be very informative to select motifs from specific areas of structure, for example all the alpha structure regions or the active site of a protein. Other information, such as positional variability and secondary structure prediction, may be invaluable during motif definition but, to the authors knowledge, there was no package available that integrated the display of these properties with multiple sequence alignments and three-dimensional structures in an interactive way.

The VISTAS and ALIGN programs were written to answer all of these needs, ie the integration of structural information when available, a high degree of user-friendliness and the ability to take into account a number of other alignment properties such as hydropathy. In fact, as will be described below, the VISTAS and ALIGN systems integrate all the functions an average user would require in a single mouse driven package, keyboard input being required only for defining file names and a small number of other actions.

6.3 VISTAS and ALIGN Programming Details

Both programs are written in C and both use the GL graphics libraries. These graphics libraries are only available at the present time on SUN and Silicon Graphics machines, but OpenGL may become the de facto three-dimensional graphics language being supported on a large number of platforms such as MicroSoft Windows and Digital ALPHA machines. Porting from GL to OpenGL is a reasonably easy process, the author already has a GLX/Motif version of VISTAS at a late stage of development. There is also a version of ALIGN (known as XALIGN) written using the standard Xlib/Motif libraries which compiles and runs, with no alteration to the source code, on VAX/VMS, ALPHA/VMS, Silicon Graphics and SUN platforms. There is no reason to believe that this would not be true of any platform which supports X/Motif.

Although, as stated above, GL is not available on every machine in Leeds, the software has been carefully written so that graphics calls are isolated from other sections of the code. It is thus a relatively simple process to retain the vast majority of the source code while just replacing the drawing routines, linking in the correct object files. This avenue is being pursued with PEX based graphics routines being developed.

VISTAS and ALIGN are written in a very modular way, thus aiding debugging and also allowing the application of extra algorithms without the need to alter large sections of source code. The modules are compiled separately and are then linked together to produce a final executable image. Currently VISTAS consists of nine and a half thousand lines of C (excluding comment lines), and ALIGN has almost six thousand lines of code (excluding comments). However, as ALIGN and VISTAS share some routines there are some fourteen thousand lines of unique C in total for the two packages.

In conclusion, though the use of some graphics libraries may restrict the portability of an application to other platforms, with careful planning this may be overcome. In the author's opinion, both ALIGN and VISTAS are written in such a way that made the porting of ALIGN to XALIGN relatively easy.

6.4 Internal organisation of data

The two main types of data which VISTAS must manipulate are those related to structural and sequence information. Data for the c-alpha display of a protein is maintained in a pre-defined three-dimensional array corresponding to the scaled x, y and z coordinates of the protein. The data required to draw the van der Waals and full-bond representation of a protein is stored in a linked list, each structure being assigned when needed using the dynamic memory allocation facilities of the C programming language. The types of structure used by VISTAS for storing this data are shown below :-

```
/* Structure for van der Waals data */
struct {
        float coords[3]; /* scaled X,Y,Z coordinates */
        float dia;        /* diameter of sphere to be drawn */
        int res_col;      /* index value for atom-type colour */
        int col;     /* index value for residue/property colour */
        v_w *next;   /* pointer to next structure in list */
        } v_w;
```

```
/* Structure for full atom display */
struct {
        float f_co[3]; /* scaled X,Y,Z coordinates of one end of
the bond */
       float t_co[3]; /* scaled X,Y,Z coordinates of the other end
of the bond *
        int f_col; /* index value for the residue/property colour
at one end of the bond */
        int t_col; /* index value for the residue/property colour
at the other end of the        bond */
       f_s *next; /* pointer to next structure in the list */
       } f_s;
```

The VISTAS and ALIGN programs maintain sequence data in pre-defined structures rather than linked lists. While this may be wasteful of memory and apply limits to the length and number of sequences that may be used as input it allows the program to very quickly locate selected sequences and residues without having to move along a linked list.

While the memory used should present no problems on modern machines, the author intends to modify the sequence data structures of VISTAS and ALIGN to conform to that used by the XALIGN program. In the case of the latter a large array of pointers to structures is used, the structures and pointers being initialised only when needed. Using this system of data management it is possible to maintain a compromise between program efficiency and memory usage.

6.5 Using VISTAS

The next section of this chapter will be in the form of a user-guide, which is probably the best method to describe the software. VISTAS will be discussed first, then XALIGN and ALIGN will be mentioned.

6.5.1 VISTAS, ALIGN and XALIGN Defaults Files

VISTAS uses a number of default files, the locations of which are defined by environment variables so that the program is not limited to a particular directory configuration. Users may copy these files to their own directories and modify them at will or supply alternative files.
These files are :-

1) A file containing residue colouring information (SOM_COL environment variable). This file contains the colours to be used for a particular residue type in an alignment or structure display in the standard RGB format. A typical entry in this file would be :-

```
A 150 150 150
```

In this case all alanine residues would be coloured dark grey.

The following template is used for the default file :-

a) Hydrophobic residues - grey (A,I,L,M,V).
b) Acidic residues - red (D,E).
c) Cystine/cysteine residues - yellow (C).
d) Basic residues - blue (H,K,R).
e) Aromatic residues - purple (F,Y,W).
f) Proline and glycine - brown (P,G).
g) Polar residues - green (T,S,N,Q).

These colours are used for all the sequence alignments coloured by residue type in this thesis and were originally defined by Dr. T.K. Attwood.

2) Colour information for calculation routines (DIV_TXT environment variable). This file contains the colour information and divisions to be used when colouring an alignment or structure by a calculated property such as hydropathy or positional variability. The divisions may be dynamic, in this case the user simply supplies a number of colours in the standard RGB format and the program divides the range of calculated values by this number. Fixed divisions may also be specified, where a colour is assigned to a particular range of values. The default file uses dynamic range definition for all the calculated physicochemical properties apart from positional variability which has defined colours for particular values.

3) Colour information for postscript output (PS_COL environment variable). This file contains the residue type colours used for the postscript output of alignments in the standard RGB format. By default, these colours are the same as those used for the colour alignment and structure but may be changed independently by the user.

4) Positional variability data (VAR_DATA environment variable). This file contains the substitution matrix defined by Risler et al. (1988) with the values normalised to a range from zero to one hundred. This file is usually transparent to the user, but redefining VAR_DATA allows different substitution matrices to be used.

5) Secondary structure prediction data (GAR_DATA environment variable). Another file usually transparent to the user, this file contains the secondary structure propensity data as described by Garnier-Osguthorpe-Robson (1987).

6) PRINTS database indexing file (PRINTS_NAME environment variable). This file contains the code and a line of description for each entry in the PRINTS database. In addition, the file contains the offset from the beginning of the PRINTS database for each entry allowing VISTAS to rapidly locate the appropriate motif information.

When used on a VMS platform XALIGN uses the same file identifiers, but in this case these are logicals rather than environment variables.

6.5.2 Running VISTAS

When the user types in the command VISTAS, an optional logo is displayed. This may be disabled by resetting the LOGO environment variable to a null value. Pressing the middle mouse button begins the program proper.

The user is then prompted for a number of input files. The underlined text below represents the prompts supplied by the computer.

Enter PDB file name of structure to be displayed (return for default) >

This prompt is repeated until an existing PDB file name is given or the default taken. The default structure to display is defined by the environment variable PDB_DEF. The structure file must be in a format which conforms to the PDB standard.

Molecule identifier (A is the default) >

Here, the user enters the chain identifier of the structure within the PDB file. The default is to extract chain A from the file. Any ligands in the PDB file will also be read by the program.

Enter name of file containing motif information >

The motif information file contains the residue positions of the start and end of a motif along with the colour in which it is to be displayed (in RGB format). Pressing return without a file name produces the next prompt, the menu options relating to motifs are greyed out until motifs are selected from within the program.

Enter name of file containing sequence alignment >

The user should enter the name of a NBRF/PIR format file containing a sequence alignment. In the case of VISTAS and ALIGN, a maximum of five hundred sequences may be read at one time. XALIGN uses dynamic memory allocation because of its wider range of platforms, with the maximum number of sequences being limited by the memory available on the machine being used. If no alignment is specified then VISTAS will extract the sequence from the PDB file given at the first prompt. If a file name is given, then the following prompt is produced :-

Enter code of the sequence that corresponds to the 3D structure >

Here the user should give the name of the sequence in the alignment whose structure is displayed. This is necessary as VISTAS needs to be able to take into account the number of gaps in the sequence when applying the colouring subroutines.

Number of sequences to display (default is 10) >

The user can specify the number of sequences to be displayed in the sequence alignment window, the default being ten. A maximum of twenty is allowed for VISTAS as a structure window also has to be displayed. ALIGN has a dynamic maximum number of displayed sequences, the program interrogates the host machine for the screen size and then calculates the largest window that can be displayed.

Enter name of file containing residue colours >

If the default is taken at this prompt, then the SOM_COL environment variable is interrogated for the path of the colour file described above. The user may supply the name of a personalised file if required

Enter colour information file >

The default is the file defined by the DIV_TXT environment variable. Again the user may supply a different file if desired.

When all the file names have been given and checked by the program a window is opened up on the screen for the display of the structure. The user can specify the size and location of this window by the manipulation of the mouse, although the window always retains a predefined width to height ratio to ensure that the protein is displayed in a sensible manner. After the structure window has been sized, the sequence alignment window is produced. This may be placed anywhere on the screen by the user, but the size is fixed and cannot be altered. The same is true of the window produced by ALIGN, but the XALIGN sequence alignment window may be resized at will with dynamic vertical and horizontal scroll bars allowing the user to move around the work area. A typical VISTAS screen at a beginning of a session is shown in figure 6.1

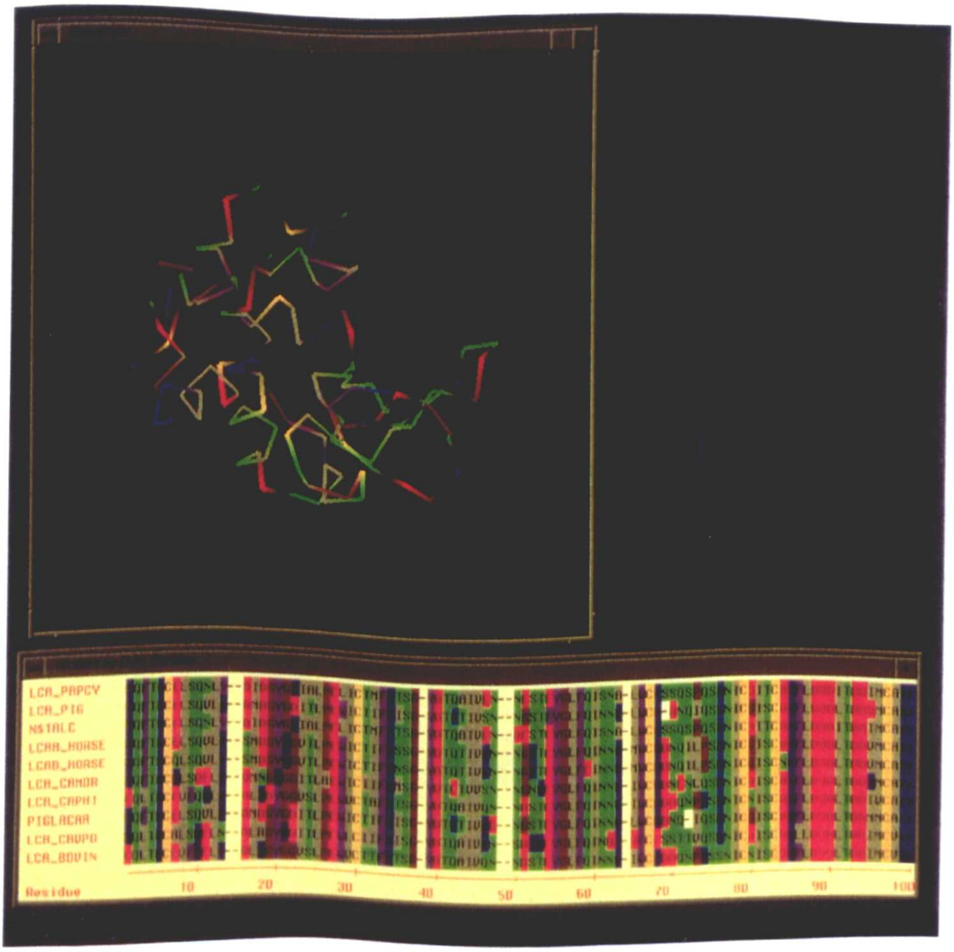


Figure 6.1 A typical VISTAS session

6.5.3 Mouse Menus

Right Mouse Button

The program is almost entirely mouse driven. The right mouse button brings up the graphics menu, which provides functions to manipulate the structure display window. A menu item is selected by releasing the right mouse button over the required option. Pressing the left mouse menu brings up the functions menu. This deals with the manipulation of the sequence alignment, the calculation of various physicochemical properties and the interface to other programs. Selections from this menu are made by pressing the right mouse button over the required option. A representation of the menu displayed when the right button is pressed is shown in figure 6.2. Each menu option is described below in detail.

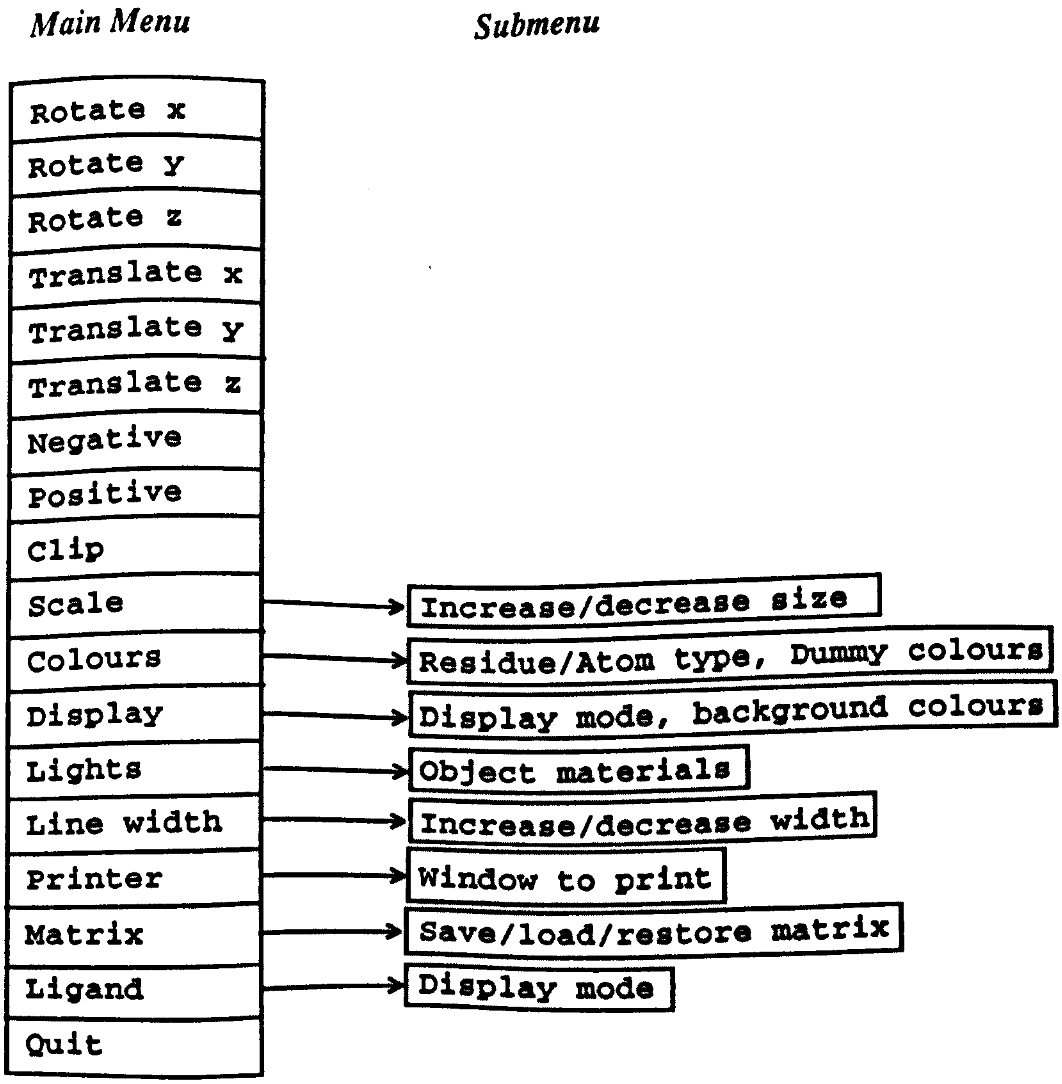


Figure 6.2 A representation of the menu invoked by the right mouse button

6.5.3.1 Rotate x,y and z

These options control the rotation of the structure displayed in the graphics window. It was decided to use such control rather than mouse dragging to allow finer control of the structure orientation. MIDAS is an example of a program which uses mouse dragging to rotate and place structures, it being quite a difficult and protracted procedure to get the structure in the desired position. Unfortunately, the standard GL library uses post-multiplication of translation matrices as standard which means that the axis of rotation and translation are retained from the displayed object rather than the screen. This means that a rotation of ninety degrees around the x axis would make subsequent rotations around the y axis appear to be around the z axis. This can be overcome by forcing the program to pre-multiply the translation matrices as illustrated by the C routines below :-

```
/* Initialise matrices (4x4 arrays) */
 Matrix temp,compound_matrix;


pushmatrix(); /* Push down matrix stack, leaving a copy of matrix at the top */
loadmatrix(temp); /* Load matrix on stack */
rotate(); /* Rotate structure */


multmatrix(compound_matrix); /* premultiply matrices */
getmatrix(compound_matrix); /* Get copy of matrix */
popmatrix(); /* Pop matrix stack */
```

When a particular rotation option is chosen, the structure rotates around the selected screen axis until the middle mouse button is pressed.

6.5.3.3 Translate x and y

These options control translations on the specified axis. The translation continues until the middle mouse button is pressed.

6.5.3.3 Negative and Positive

These options control the direction of translation and rotation around a particular axis. A negative rotation rotates the displayed structure in a anti-clockwise direction and translates to the bottom left of the screen. Positive translations and rotations occur in the opposite direction.

6.5.3.4 Clip

This option allows the user to clip parts of the displayed structure. The direction of clip is controlled by the negative/positive menu option above. Again, clipping continues until the middle mouse button is pressed.

6.5.3.5 Scale

When this option is selected, a submenu is displayed. This allows the user to increase or decrease the size of the displayed structure. The initial clipping planes are set at a very wide range, allowing the user to greatly magnify a particular region of a structure. Scaling continues until the middle mouse button is pressed. An option in this submenu also allows the user to increase or decrease the size of spheres used in the ball and stick displays described later. If the structure is not displayed using spheres, these latter options are greyed out.

6.5.3.6 Colours

This option brings up a submenu with options for different colouring algorithms. Residue colouring colours the structure and alignment according to residue type, the colour values being defined in the file referred to by the environment variable SOM_COL. Dummy colouring simply colours the alignment and structure with a repeating red-green-blue pattern and atom colours colour the structure only with atom-specific colours.

6.5.3.7 Display

This submenu is probably the most important controlled by the right mouse button. One of the options allows the user to select the method used to display the structure, at the moment modes available are C-alpha trace, skeletal, C-alpha spheres with bonds, full spheres with bonds, space filling and space filling with dots. Examples of all these display modes are shown in figure 6.3. A ribbon display will also be added when time allows.

The retain colour and colour options in this submenu allow the user to switch the colouring algorithms for the structure and or sequence displays on and off. Using these it is possible to display, for instance, the structure coloured by residue type with the sequence alignment coloured by some physicochemical property.

The colour motifs option allows the user to colour the structure and sequence according to the motifs selected, if no motifs have been selected this option is greyed out.

The background options allows the manipulation of the colours for the structure window background. Options supplied are plain orange, gouraud shaded orange, black and user defined. This latter option allows users to enter their own choice of gouraud shaded colours for the background.

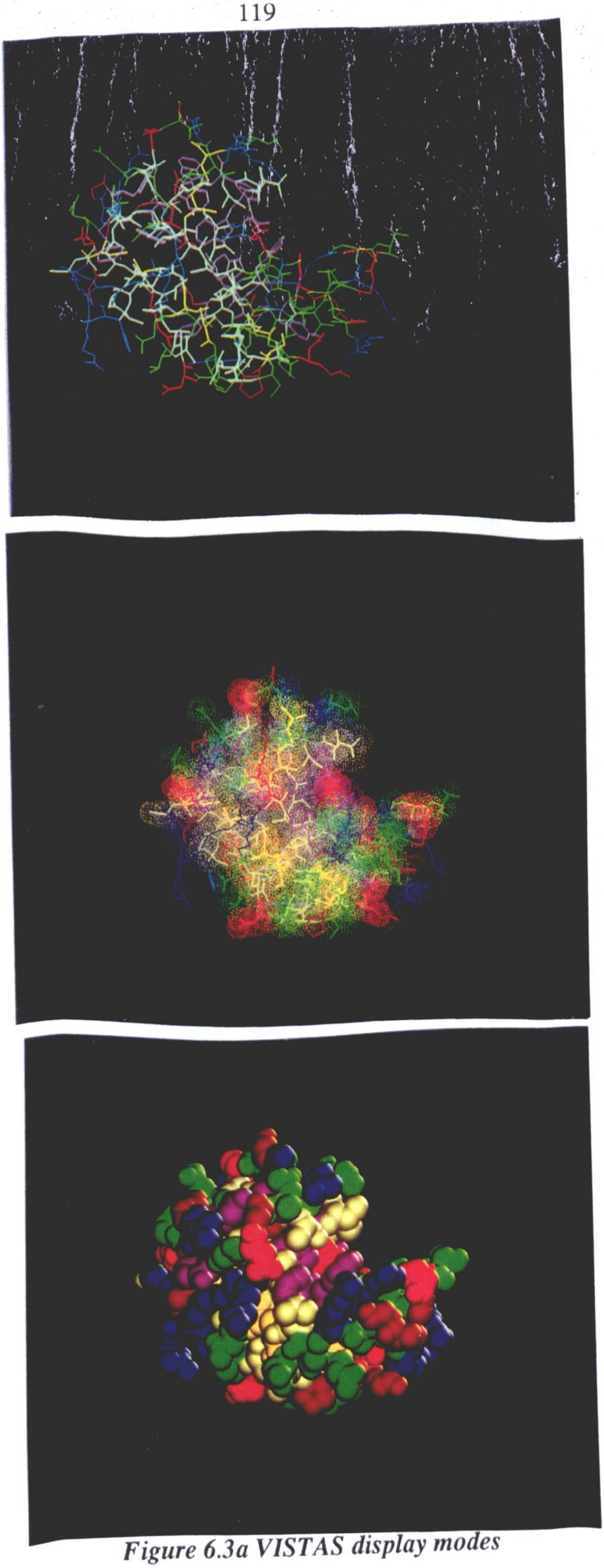


Figure 6.3a VISTAS display modes

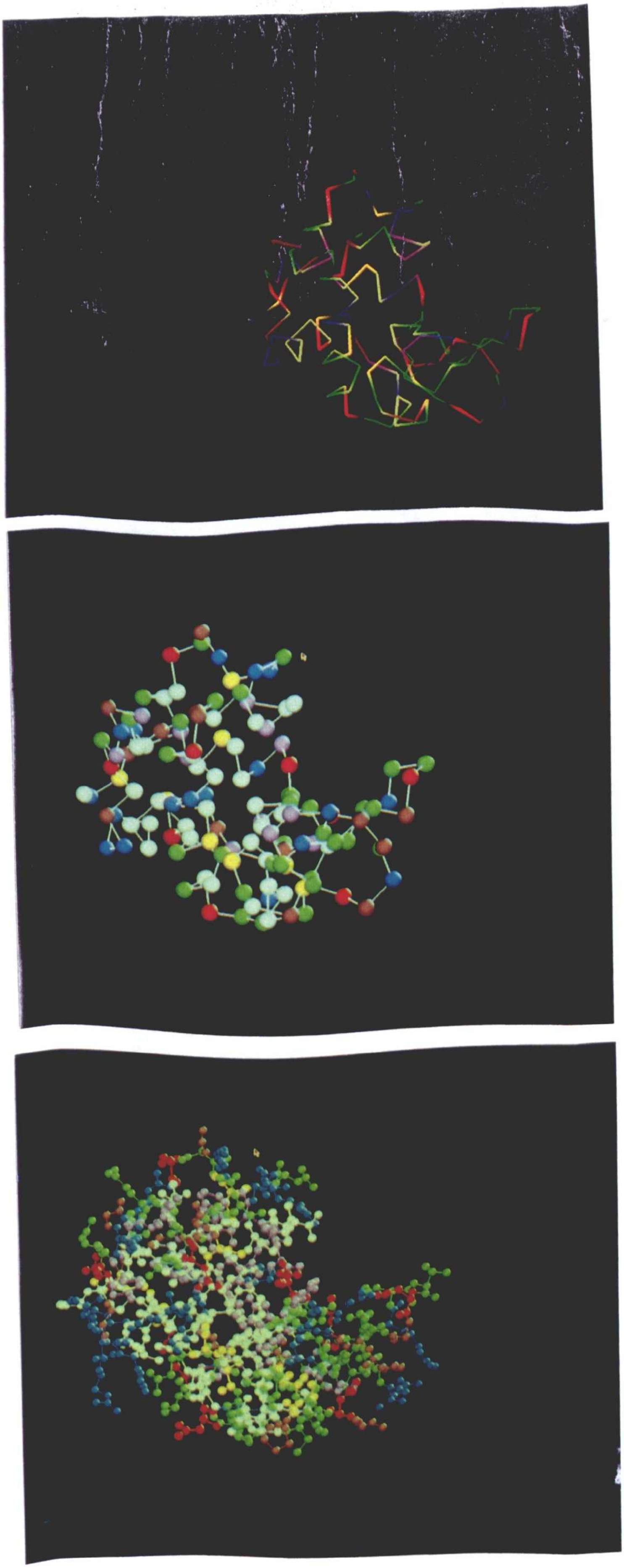


Figure 6.3b VISTAS display modes

6.5.3.8 Lights

The GL library allows the programmer to specify a number of lighting properties to be used when displaying an object, for instance ambient and specular lights. With the proper use of these parameters it is possible to give the illusion that a displayed object is made of a particular material. This submenu allows a user to select a material or to toggle the lighting on and off. The materials currently available are plastic (three types), steel, glass, brass, pewter, silver, gold, plaster, bronze and rubber. While most of these are largely cosmetic they allow the customisation of displays and the transparency offered by the glass material may be useful in identifying atom types behind the front of the display.

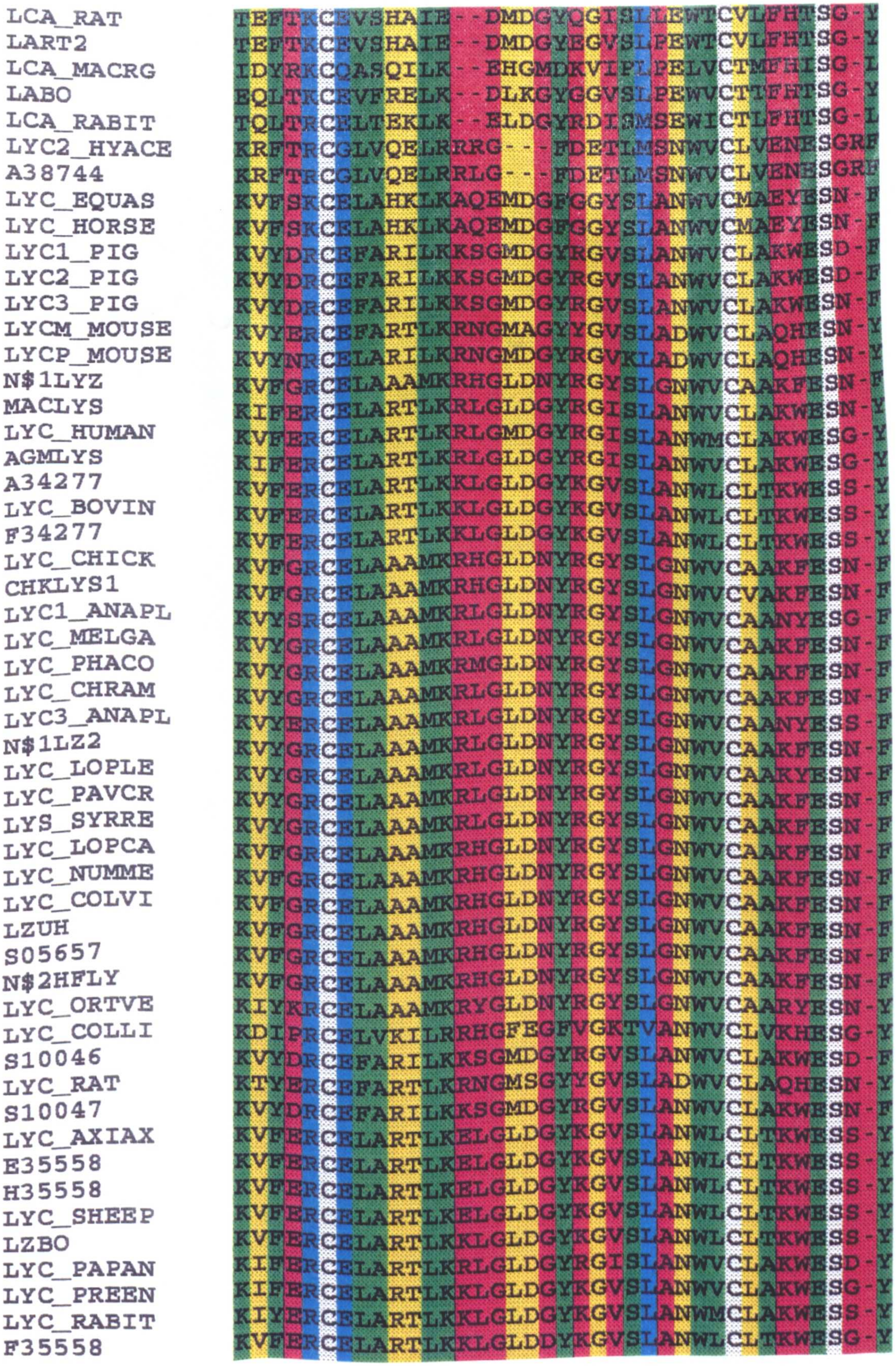
While lighting models provide a three dimensional impression they are particularly processor intensive, so the option to switch lighting off is also provided. Space filling models displayed with dots are also easier to interpret with no lighting.

6.5.3.9 Line Width

This submenu controls the width of the lines drawn to represent bonds, which can be increased or decreased at will.

6.5.3.10 Printer

VISTAS provides hardcopy facilities for all the windows it displays, controlled by this submenu. The structure display window, graph display and plot display are saved to files in the Silicon Graphics RGB format, while the sequence alignment is in postscript. The use of the RGB file format allows the incorporation of screen displays into Explorer documents, an option unlikely to be required for sequence alignments. If it is found to be necessary to convert RGB files into postscript, Silicon Graphics have provided an appropriate utility in the form of the TOPS program. When the alignment option is selected from the menu, the user is prompted for an output file name and also a file containing the colour information for particular residue types if appropriate. The path for this latter file is provided by the PS_COL environment variable, but the user can specify a personalised file if required. Part of a sequence alignment produced by VISTAS and coloured by positional variability is shown in figure 6.4.



```
LCA_RAT       TEFTKCEVSHAIE--DMDGYQGISLLEWTCVLFHTSG-Y
LART2         TEFTKCEVSHAIE--DMDGYEGVSLPEWTCVLFHTSG-Y
LCA_MACRG     IDYRKCQASQILK--EHGMDKVIPLPELVCTMFHISG-L
LABO          EQLTKCEVFRELK--DLKGYGGVSLPEWVCTTFHTSG-Y
LCA_RABIT     TQLTRCELTEKLK--ELDGYRDISMSEWICTLFHTSG-L
LYC2_HYACE    KRFTRCGLVQELRRRG---FDETLMSNWVCLVENESGRF
A38744        KRFTRCGLVQELRRLG---FDETLMSNWVCLVENESGRF
LYC_EQUAS     KVFSKCELAHKLKAQEMDGFGGYSLANWVCMAEYESN-F
LYC_HORSE     KVFSKCELAHKLKAQEMDGFGGYSLANWVCMAEYESN-F
LYC1_PIG      KVYDRCEFARILKKSGMDGYRGVSLANWVCLAKWESD-F
LYC2_PIG      KVYDRCEFARILKKSGMDGYRGVSLANWVCLAKWESD-F
LYC3_PIG      KVYDRCEFARILKKSGMDGYRGVSLANWVCLAKWESN-F
LYCM_MOUSE    KVYERCEFARTLKRNGMAGYYGVSLADWVCLAQHESN-Y
LYCP_MOUSE    KVYNRCELARILKRNGMDGYRGVKLADWVCLAQHESN-Y
N$1LYZ        KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESN-F
MACLYS        KIFERCELARTLKRLGLDGYRGISLANWVCLAKWESN-Y
LYC_HUMAN     KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESG-Y
AGMLYS        KIFERCELARTLKRLGLDGYRGISLANWVCLAKWESG-Y
A34277        KVFERCELARTLKKLGLDGYKGVSLANWLCLTKWESS-Y
LYC_BOVIN     KVFERCELARTLKKLGLDGYKGVSLANWLCLTKWESS-Y
F34277        KVFERCELARTLKKLGLDGYKGVSLANWLCLTKWESS-Y
LYC_CHICK     KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESN-F
CHKLYS1       KVFGRCELAAAMKRHGLDNYRGYSLGNWVCVAKFESN-F
LYC1_ANAPL    KVYSRCELAAAMKRLGLDNYRGYSLGNWVCAANYESG-F
LYC_MELGA     KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESN-F
LYC_PHACO     KVYGRCELAAAMKRMGLDNYRGYSLGNWVCAAKFESN-F
LYC_CHRAM     KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESN-F
LYC3_ANAPL    KVYERCELAAAMKRLGLDNYRGYSLGNWVCAANYESS-F
N$1LZ2        KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESN-F
LYC_LOPLE     KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKYESN-F
LYC_PAVCR     KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESN-F
LYS_SYRRE     KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESN-F
LYC_LOPCA     KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESN-F
LYC_NUMME     KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESN-F
LYC_COLVI     KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESN-F
LZUH          KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESN-F
S05657        KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESN-F
N$2HFLY       KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESN-F
LYC_ORTVE     KIYKRCELAAAMKRYGLDNYRGYSLGNWVCAARYESN-Y
LYC_COLLI     KDIPRCELVKILRRHGFEGFVGKTVANWVCLVKHESG-Y
S10046        KVYDRCEFARILKKSGMDGYRGVSLANWVCLAKWESD-F
LYC_RAT       KTYERCEFARTLKRNGMSGYYGVSLADWVCLAQHESN-Y
S10047        KVYDRCEFARILKKSGMDGYRGVSLANWVCLAKWESN-F
LYC_AXIAX     KVFERCELARTLKELGLDGYKGVSLANWLCLTKWESS-Y
E35558        KVFERCELARTLKELGLDGYKGVSLANWLCLTKWESS-Y
H35558        KVFERCELARTLKELGLDGYKGVSLANWLCLTKWESS-Y
LYC_SHEEP     KVFERCELARTLKELGLDGYKGVSLANWLCLTKWESS-Y
LZBO          KVFERCELARTLKKLGLDGYKGVSLANWLCLTKWESS-Y
LYC_PAPAN     KIFERCELARTLKRLGLDGYRGISLANWVCLAKWESD-Y
LYC_PREEN     KIFERCELARTLKKLGLDGYKGVSLANWVCLAKWESG-Y
LYC_RABIT     KIYERCELARTLKKLGLDGYKGVSLANWMCLAKWESS-Y
F35558        KVFERCELARTLKKLGLDDYKGVSLANWLCLTKWESG-Y
```

Figure 6.4 Part of an alignment of lysozyme and α lactalbumin sequences coloured by positional variability. White indicates totally conserved residues, blue well conserved through green and yellow to red which indicates the most variable residues.

6.5.3.11 Matrix

The matrix submenu allows the manipulation of the translation and rotation matrices used for the structure display. These can be written to files when a structure is in the required orientation and can be read at a later date when the program is restarted. Another option in this submenu allows the translation and rotation matrix to be restored to its initial state. Matrix files contain five lines, the first four lines are the 4x4 translation matrix while the fifth line represents the x, y and z axis translation values and scale value.

6.5.3.12 Ligand

Selecting the ligand option produces a submenu which allows the ligand display to be switched on and off as desired and also the ligand display mode to be defined. Three options are allowed, these being stick, space filling and double space filling. This latter option displays the double van der Waals radii of the ligand atoms which, when combined with the skeletal display, allows the identification of possible close contacts. Figure 6.5 illustrates the different ligand display modes. When no ligand is present, all these options are greyed out.

The display of ligands is a useful feature as it allows the identification of the residues involved in interactions with other molecules, these residues can then be selected using the options described below and used to search the sequence database.

6.5.3.13 Quit

This menu option quits the program immediately and returns the user to the computer's operating system.

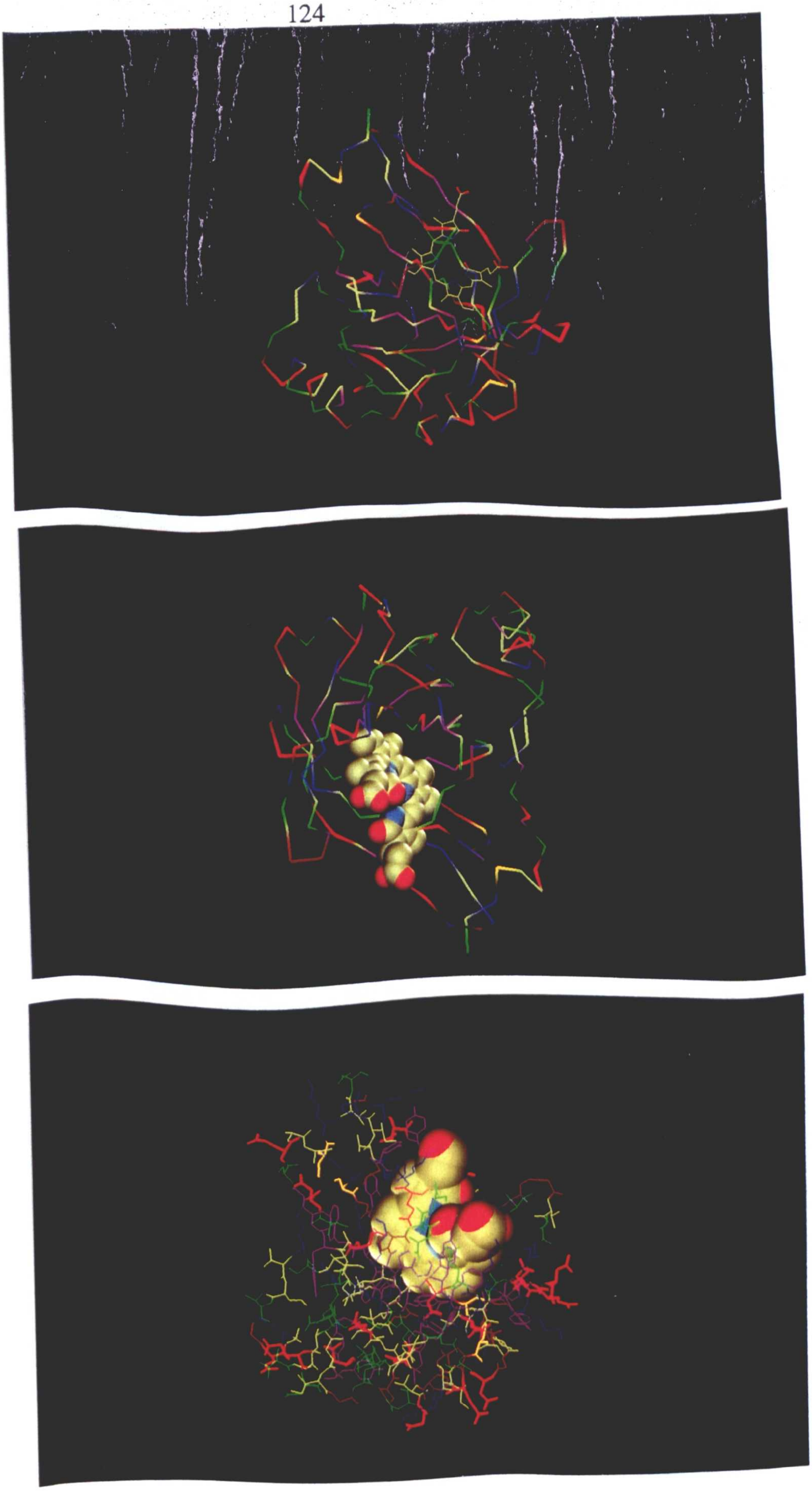


Figure 6.5 Ligand display modes

6.5.4 Left Mouse Button

The left mouse button controls those functions not linked directly to the structure window and within which lies VISTAS main strength. Figure 6.6 shows a stylised diagram of the menus invoked by pressing the left mouse button, these are described below.

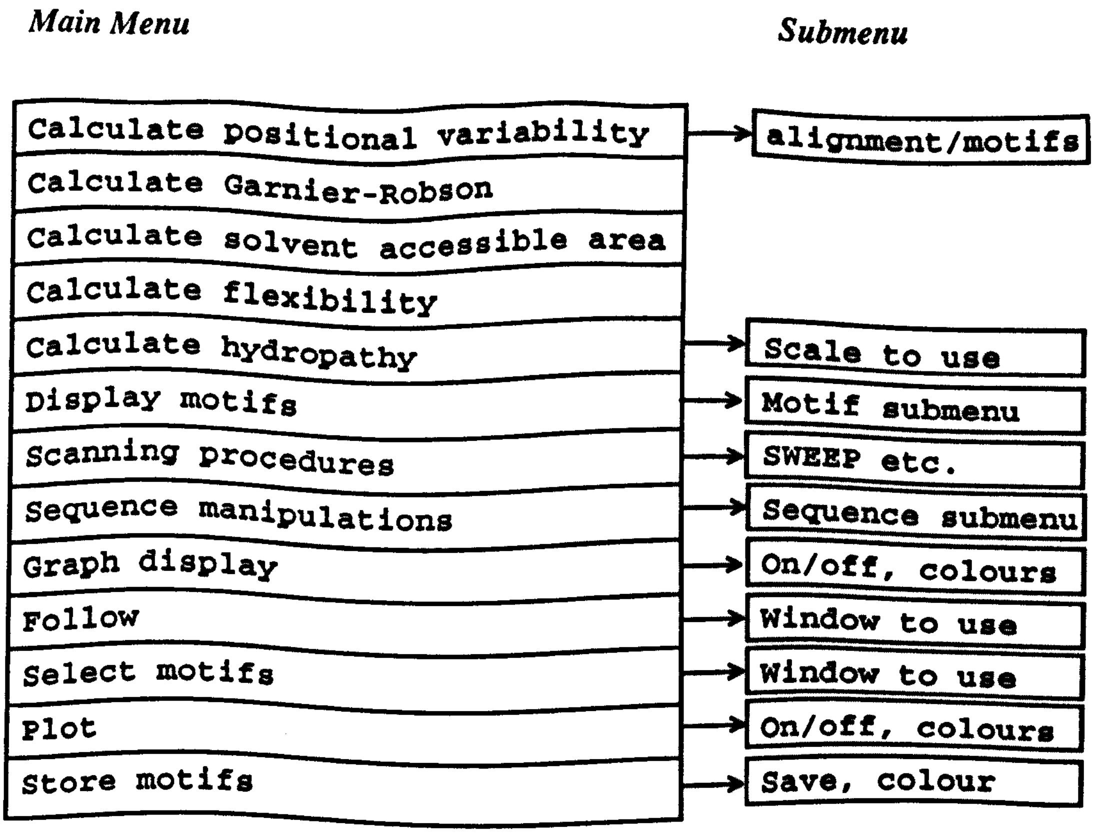


Figure 6.6 The menu invoked by the left mouse button

6.5.4.1 Amino Acid Properties

As more thorough descriptions of the following properties are given in the previous chapter, the algorithms and scales used will only be discussed briefly here.

Positional Variability

The positional variability of an alignment can be calculated either from the alignment displayed or from motif files. This latter option allows the user to display a manageable alignment while displaying data from a very large number of sequences. When the motif file option is selected, the user is prompted for a file which contains the list of motif file names. The positional variability is calculated by comparing every residue at a particular position with every other residue at that position, the similarity values are taken from the substitution matrix defined by Risler et al. (1988). This matrix may be replaced by redefining the environment variable VAR_DATA. When the values have been calculated, the alignment and structure are coloured using the information from the file defined by the DIV_TXT environment variable. When motif files are used only those areas of the alignment that correspond to a particular motif are coloured by positional variability, the rest of the alignment is coloured cyan.

Garnier-Osguthorpe-Robson Secondary Structure Prediction

This option uses the Garnier-Osguthorpe-Robson (1978) algorithm to predict the possible secondary structure of the sequence alignment. The sequence alignment and structure are then coloured by the four possible structural conformations, ie turn, coil, alpha and beta. While secondary structure prediction is notoriously unreliable, the use of an alignment does produce more accurate results and it can be informative to compare the results with the known structure. With ALIGN and XALIGN, this option is more worthwhile as it is the only indication of secondary structure and it may be useful to be able to select motifs from, for instance the putative transmembrane helices of a membrane transport protein.

Solvent Accessible Area

The scale used by this option was derived from measuring the mean solvent accessible surface area of each of the residue types in twenty three folded proteins (Rose, G.D. et al. (1985)). This option is designed to be used in conjunction with the hydropathy scales as it has been demonstrated that those residues that have the highest solvent accessible surface area are the most hydrophilic. A window length of five residues is used.

Flexibility

This option allows the identification of those residues of a sequence or sequence alignment that are the most flexible (Ragone, R. et al. (1989)). The scale used exploits the fact that the residues most likely to be found in such regions are generally the most hydrophilic and have the smallest volumes. A window length of ten residues is used.

Hydropathy

This submenu contains three options, each one a different hydropathy scale. The Kyte and Doolittle (1982)· method uses a window length of twenty residues, the Eisenberg (Sweet, R.M. and Eisenberg, D. (1983)) method a nine residue window and the transmembrane method (Engelman, D.M. et al. (1986)) a window length of twenty. These algorithms, especially the latter, are very useful for detecting the possible transmembrane regions in proteins where no three-dimensional structure is available. The whole alignment is used to calculate the hydropathy values, not just a single sequence, and the sequence alignment and structure are coloured according to the information from the file defined by the DIV_TXT environment variable.

6.5.4.2 Display Motifs

When this option is selected a separate menu is produced, invoked by the left mouse button, which displays the active motifs and also whether the display of these motifs is switched on or off. The user toggles a motif on or off by pressing the right mouse button when the cursor is over the required menu option, the part of the structure which corresponds to the motif is then displayed as appropriate. The display may be produced in any of the possible modes, for instance spaçe filling. A separate menu was used for the control of motif displays as, while being a little more confusing than when integrated with the main menus, the number of active motifs is dynamic and the full menus would need to be refreshed after each operation. Figure 6.7 illustrates the display of a number of lysozyme motifs along with the motif menu.

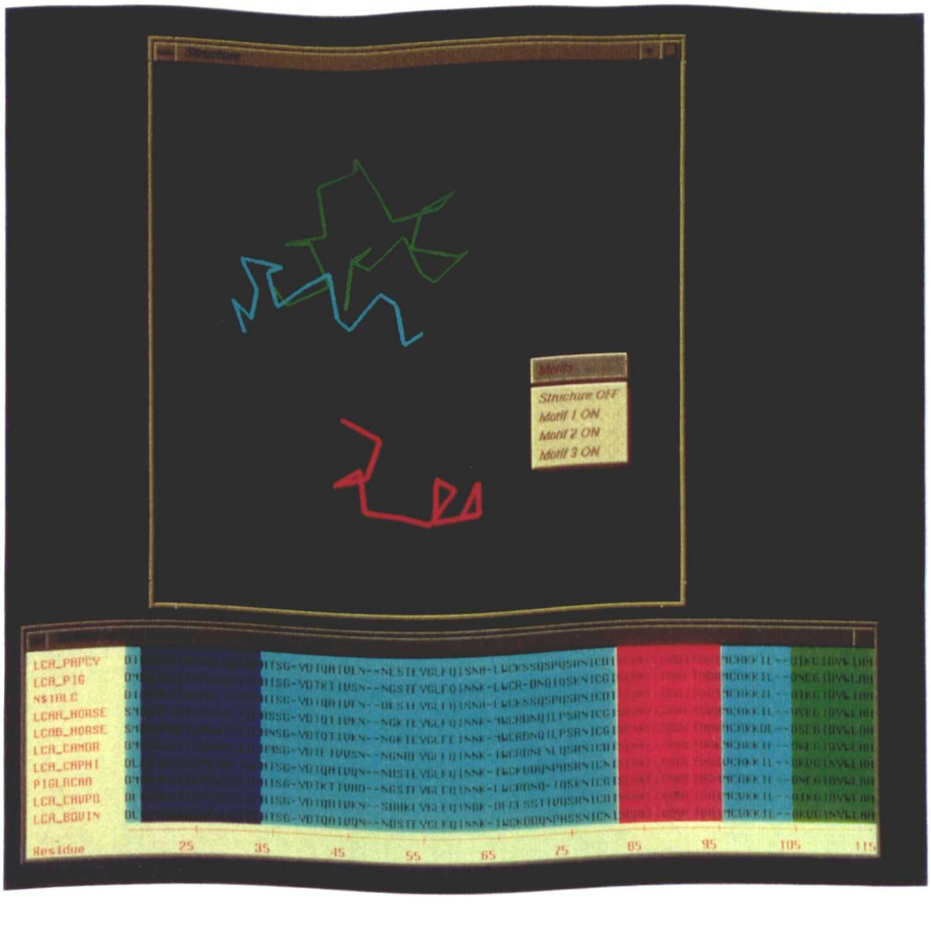


Figure 6.7 A VISTAS session illustrating the use of the motif submenu.

6.5.4.3 Scanning Procedures

One of the most powerful features of VISTAS, ALIGN and XALIGN are the direct interfaces to a number of other programs which allow a user to perform the whole process of motif definition and database scanning in a seamless manner. The following programs are available from this menu :-

1) SWEEP (Akrigg, D. et al. (1988)) and FASTA (Pearson, W.R. and Lipman, D.J. (1988)) global sequence searching programs. The user may either pass a sequence selected from the alignment window or give a database code. A number of prompts are then produced requiring more information, such as the database to search. VISTAS submits the FASTA or SWEEP job as a background process when all the prompts have been answered, a log file is produced for any information passed back by the machine. SWEEP searches the sequence database with the given sequence by considering matches from the database to the whole length of the probe sequence and also to overlapping sub-sequences of the probe sequence. FASTA uses a technique based on producing dot-plots of the probe and database sequences, areas of similarity being shown as diagonal lines on such a plot. The final similarity score between the two sequences is calculated by joining these regions.

2) SCAN. The motifs selected by the user from VISTAS, ALIGN or XALIGN can be submitted directly to the motif database scanning routine described in chapter two. The user is prompted for information, for instance the scoring method to be used, then the SCAN job is submitted as a background process. A log file is produced for any computer generated messages.

3) COMPARE. This is the same algorithm as described in chapter two and allows the analysis of hitlists produced by database scans with motifs. This option, together with SCAN, allows the user to define motifs, scan the database, analyse hitlists and refine motifs all within the same program with no file editing required. Again prompts are produced for the user to provide the appropriate parameters to COMPARE.

4) SMITE and DELPHOS (Akrigg, D. et al. (1988)). SMITE is the PRINTS database interrogation language. The PRINTS database contains motifs which discriminate for specific protein families in a similar manner to the entries in the PROSITE database. DELPHOS is the OWL sequence database query language. Both SMITE and DELPHOS are run in a separate window, the main VISTAS program pauses until the user leaves the query software being used by typing "quit".

5) PRINTS database scanning. All the motifs that are defined at Leeds are entered, after a rigorous checking procedure, into the PRINTS database. This is similar in concept to the PROSITE database but contains a more thorough description of each entry, some examples of PRINTS database entries are shown in appendix C. VISTAS allows the user to scan the entire database or just a named entry. The results are then presented to the user, who may add the motifs to the display list of motifs and write new motif files if desired. In contrast to the other programs described above, the PRINTS database scanning module is an integral part of VISTAS and is linked in to produce the final executable image.

The scanning procedures submenu also contains two options which allow the manipulation of the motif lists used by the display and database scanning routines. The clear motif list option removes all the present motifs while the write motif list option writes the names of all the active motifs to a file, the name of which is defined by the user. This file can then be used as input for other sequence analysis programs.

6.5.4.4 Sequence Manipulations

In addition to the structure display and database scanning options, VISTAS has a very powerful sequence alignment and editing capability superior to many packages produced purely for this purpose alone. At the present only manual alignment is supported, although automatically aligned sequences can be imported. The author intends to introduce a hybrid automatic and manual alignment system where the user may fix parts of the alignment while the rest of the sequences are aligned automatically. Automatic alignment alone, while producing objective output, tends to be limited by a number of factors, for instance if the sequences to be aligned are not of similar lengths large numbers of gaps are often inserted.

6.5.4.4.1 Alignment Navigation

If the alignment is large, the sequence window will only display part of the alignment. The 'r' key scrolls the alignment a complete window (100 residues) to the right while the 'R' key scrolls the alignment to the right by 10 residues. The 'L' and 'l' keys perform similar functions but scroll to the left. The 'd' key scrolls the alignment down a complete screen while pressing 'D' scrolls the window 5 sequences down. The 'u' and 'U' keys are used for scrolling the alignment up by similar amounts.

6.5.4.4.2 Insert/Delete gaps

When this option is selected gaps may be introduced into a sequence by moving the mouse pointer to the desired position and then pressing the right mouse button. Pressing the left mouse button deletes gaps at the cursor position. Single or multiple gaps may be inserted, the number can be user-defined by pressing the 'i' or 'I' keys to produce a prompt. The gaps algorithm makes extensive use of the C string handling functions which, while being inherently slow, are much faster than updating a sequence in memory one residue at a time.

6.5.4.4.3 Write Sequence Set

After selecting this option the user is prompted for a file name. All the sequences present in the alignment, including any gaps, are then written to this file using the standard NBRF/PIR format.

6.5.4.4.4 Write Part of the Sequence Set

After selecting this option the user defines the beginning and end of the section of the sequence alignment required by clicking on the right mouse button when the cursor is in the correct position. A file name prompt is then displayed. Only the selected section of the alignment is written to the specified file allowing the user to discard unwanted sections of the alignment. This option is particularly useful in those situations where some sequences in the alignment have long unwanted leader sequences.

6.5.4.4.5 Write Identity Matrix

When this option has been selected, the beginning and end residues of the appropriate section of the alignment are defined by clicking on the right mouse button when the cursor is in the correct positions. After responding to the file name prompt, the frequencies of the residue types at each position is then written in the form of a matrix.

6.5.4.4.6 Add Sequences

When this menu option is selected, a prompt is produced which requests a file name or database code. VISTAS first checks the user's directory for the file and, if this search is unsuccessful, then checks the OWL database index files for a protein of the given name. If a file name is given, that file must contain sequences in the NBRF/PIR format. When the file has been located or the sequence extracted from the OWL database the user then clicks on the right mouse button when the cursor is in the position in the alignment where the sequence or sequences are to be inserted. This whole procedure, from prompt to selecting the insertion position, is repeated until the middle mouse button is pressed allowing multiple sequence additions.

6.5.4.4.7 Delete Sequences

After selecting this option the uses clicks on the right mouse button when the cursor is over the chosen sequence which is then removed from the alignment. Pressing the middle mouse button leaves the sequence deletion mode.

6.5.4.4.8 Swap Sequences

The two sequences whose positions in the alignment are to be swapped are selected by clicking on the right mouse button when the mouse cursor is in the required positions. Again, pressing the middle mouse button returns the user to the main menus.

6.5.4.4.9 Go to Residue

After selecting this option, the user is prompted for a residue number. The alignment is then redisplayed with the selected residue being in the first column of the alignment window.

6.5.4.4.10 Find Motif

The user is first prompted for a motif which is entered from the keyboard. The sequence to be searched is then selected using the cursor and the right mouse button. A fuzzy search is carried out, the highest scoring segments from the selected sequence being displayed to the user. If required, the user can then reset the display so that the first residue of the highest scoring segment is in the first column of the alignment window. Pressing the middle mouse button returns to the main menus.

6.5.4.4.11 Make Group

Sequences from the alignment may be grouped together so that an insertion or deletion in one group member produces a similar insertion or deletion in all the other group members, thus making the alignment process much less time-consuming and tedious. The sequences to be grouped are selected by pressing the right mouse button when the cursor is in the required position. More sequences may be added to a group until the middle mouse button is pressed, when the names of the sequences in the group are displayed in the same colour in the alignment window.

6.5.4.4.12 Groups On/Off

A sequence group may be toggled on or off by pressing the right mouse button when the cursor is over any member of the required group. When a group is switched off, the members of that group are treated as individuals and insertions and deletions are not mirrored in the other group members.

6.5.4.4.13 Reset Group

A sequence group may be reset by pressing the right mouse button when the cursor is over any member of the desired group. This option removes the group from the computer's memory and none of the group functions then apply to the former members.

6.5.4.4.14 Add to Group

This option allows new members to be added to a predefined group. The group is first selected by pressing the right mouse button when the cursor is over any member of the desired group, then the sequences to be added are selected in a similar manner. Pressing the middle mouse button returns to the main menus.

6.5.4.4.15 Sequence Editor

When this option is selected, the user defines the residue or residues to be changed by positioning the cursor in the appropriate place and then pressing the right mouse button. The user is then prompted for the residue or string of residues to replace the previous sequence. This option is particularly useful for studying the effects of mutations on secondary structure prediction.

6.5.4.4.16 Define Anchor Point

An anchor point allows the user to insert or delete gaps but retain a particular part of the alignment intact, gaps are inserted and deleted either side of the anchored sequence to ensure that it stays in the same position.

6.5.4.4.17 Reset Anchor Point

This option removes any anchor points, gaps subsequently being inserted or deleted in the normal fashion.

6.5.4.4.18 Select Ruler Sequence

The user may select the sequence that is used to define the ruler that is displayed at the bottom of the alignment window. This sequence is selected by pressing the right mouse button when the cursor is the required position. The sequence name is then coloured blue to aid identification and the residue numbers displayed on the ruler relate to that particular sequence, gaps being disregarded. This option allows the easy identification of particularly significant residues, for instance it would be straightforward for a user to locate residue Ser 134 in a sequence even if that sequence contained a large number of gaps.

6.5.4.4.19 Alignment Ruler

When this option is selected the residue numbers displayed on the ruler relate to the whole alignment including all the gaps.

6.5.4.4.20 Go to End

After this menu option has been selected, the user selects a particular sequence with the cursor and right mouse button. The alignment is then redisplayed, with the last residue of the selected sequence being in the first column of the alignment window.

6.5.4.4.21 Go to Start

This option is identical to that described above except that the first residue of the selected sequence is displayed in the first column of the sequence display.

6.5.4.5 Graph Display

Another feature of VISTAS is the ability to display physicochemical properties in an interactive manner both by colour coding the sequence and structure and also by displaying the data as a graph in a separate window. The graph display submenu allows the user manipulation of this graph display, the individual options being described below.

6.5.4.5.1 Graph On/Off

The graph display may be switched on and off as desired by the use of these menu options. When the graph window is produced the size is fixed so that a correct aspect ratio is retained. To the right of the graph window is a colour bar which indicates the value ranges represented by the colours used for the structure and alignment displays. A typical graph window is shown in figure 6.8.

6.5.4.5.2 Reset Colours

This option is used to reset the graph display after the Follow options (described below) have been used.

6.5.4.5.3 White Background/Orange Background

The background colour of the graph display may be changed as desired.

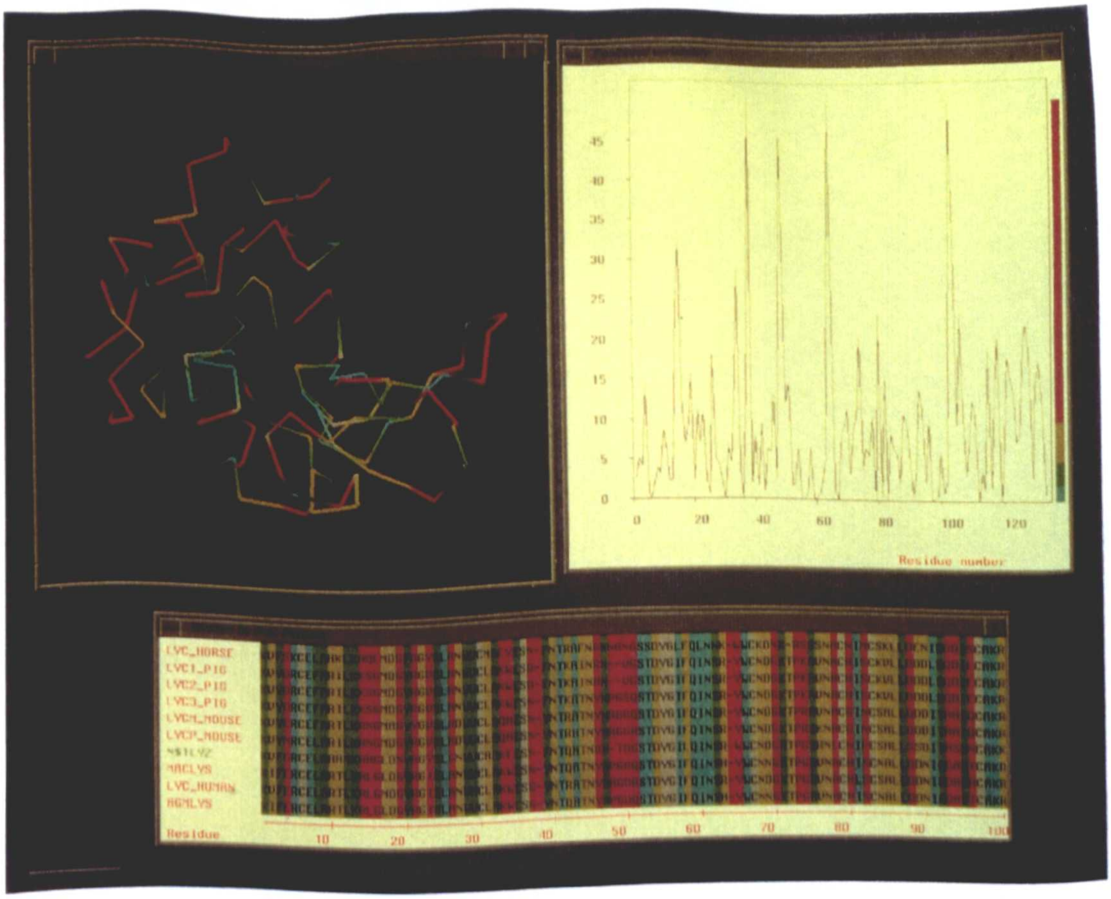


Figure 6.8 A VISTAS session with a graph window displayed, in this case showing the positional variability of an alignment of lysozymes and α lactalbumins.

6.5.4.6 Follow

The Follow submenu allows the integration of sequence, structure, graph and plot windows. The user selects a segment or a single residue from any of the displayed windows by pressing the right mouse button when the cursor is at the appropriate positions, the corresponding areas of the other displays are then indicated. For instance if part of the protein structure is selected, the segment of the sequence alignment that corresponds to this region is coloured red while the appropriate parts of the graph and plot windows are indicated by dotted lines. Pressing the middle mouse button returns to the main menus.

6.5.4.7 Select Motifs

This sub menu allows motifs to be selected from any of the display windows. The user selects the desired region of the window by pressing the right mouse button on the beginning and end of the segment and is then prompted for a file name. The corresponding motif from the sequence alignment is then written to a file for use with the database scanning or general motif manipulation algorithms. This option makes it easy for users to select particular areas of secondary structure, for instance all the alpha helices, or the active site from the structure display and then scan the database with these sequence motifs. The graph display is also useful as users may, for instance, select the significantly hydrophobic sections. The motifs selected may be used to scan the sequence database without leaving the program by selecting the appropriate option from the Scanning Procedures submenu described above.

6.5.4.8 Plot

This submenu interfaces to the PLOT routine mentioned in earlier chapters, the output is then displayed in a separate window. PLOT uses the motifs that the user has previously defined, if no motifs have been selected then the Plot submenu is greyed out and is non-selectable. If the Plot from Alignment option is selected then the user must define the sequence to be used by pressing the right mouse button when the cursor is at the correct position. The Plot from Database option first prompts the user for a protein name, the sequence is then extracted directly from the OWL database. The Close plot window option removes the PLOT window, as with the graph display the background colour for the PLOT window may be changed. Figure 6.9 illustrates a VISTAS session with a PLOT window.

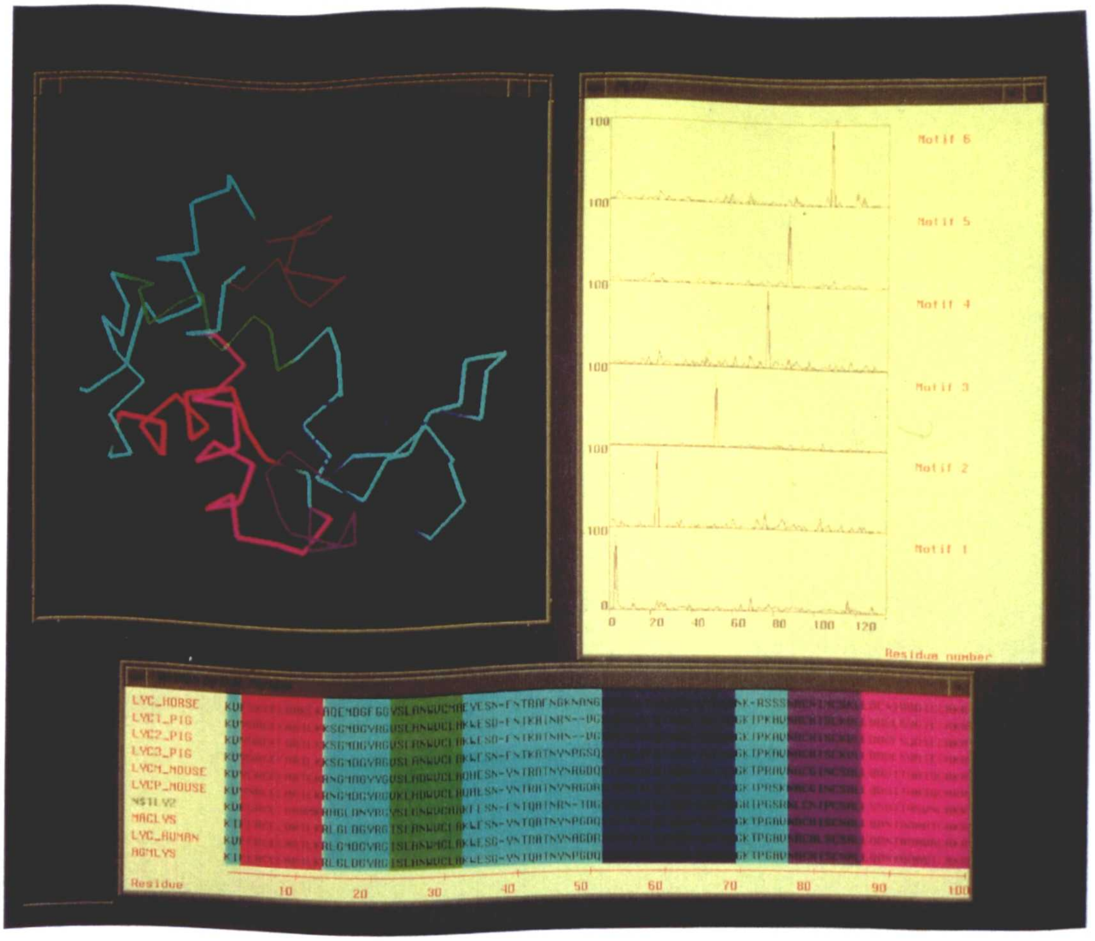


Figure 6.9 A VISTAS session with a PLOT window displayed.

6.5.4.9 Store Motifs

This option allows the user to save the predefined motifs in memory before clearing the motif list. The Colour motifs option from this submenu then colours the structure and alignment using the presaved motif definitions and the present motif definitions, the areas of overlap being coloured red. This option is useful when comparing the output from two PRINTS database searches.

6.6 ALIGN and XALIGN

Both these programs contain the options described above with the exception of those specifically concerned with the manipulation of the structure display window. These programs were written to account for the majority of situations when a structure that corresponds to the sequences being aligned is not yet available. As mentioned above, XALIGN makes use of the Xlib/Motif programming libraries and therefore has a user-interface that conforms to the Motif standard with pulldown rather than popup menus. XALIGN also has the ability to produce helical wheel displays. A typical XALIGN session is shown in figure 6.10.

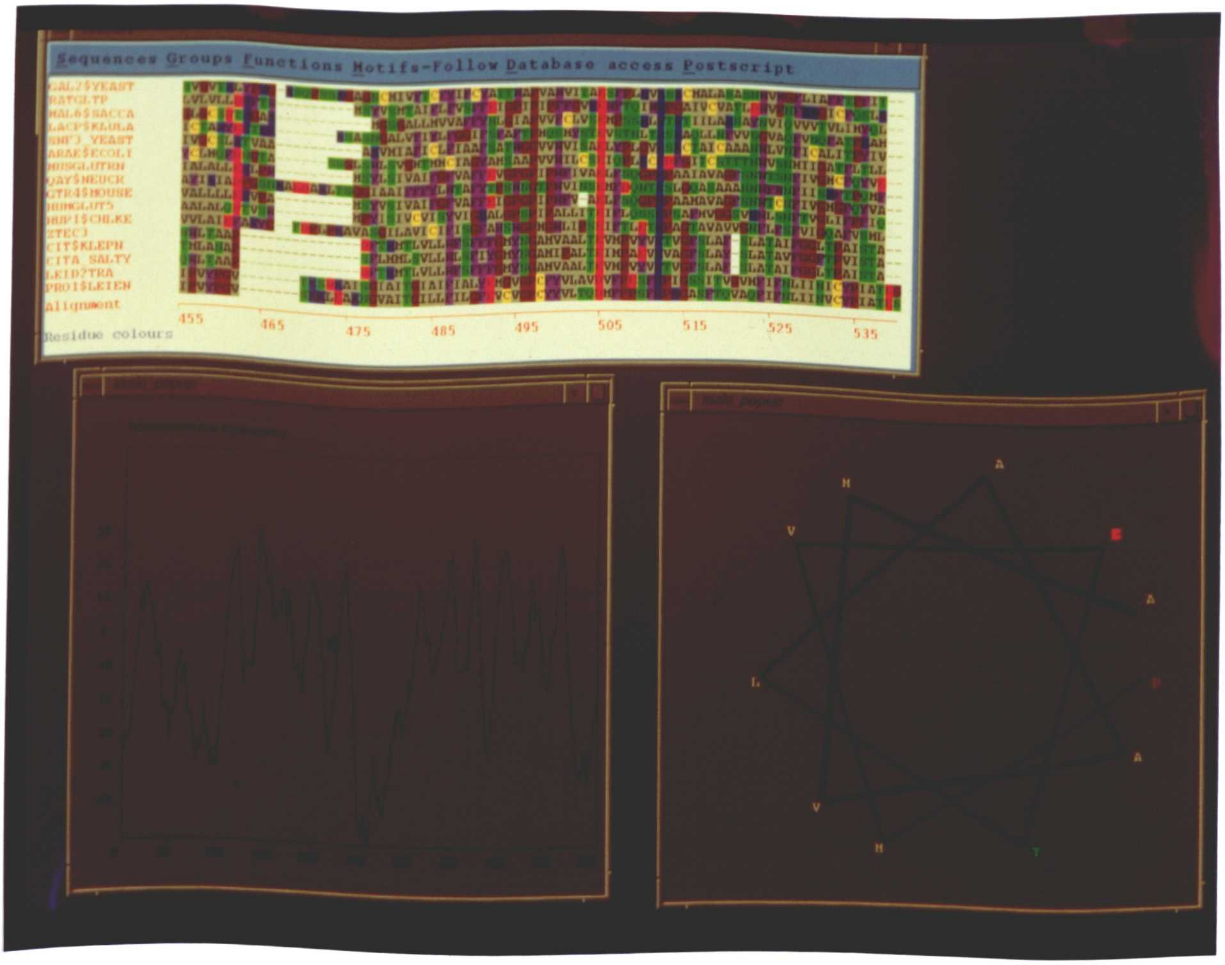


Figure 6.10 A XALIGN session with a graph window and a helical wheel displayed

6.7 Comparison With Other Software

Only two software packages are known to the author that have a similar range of functions to VISTAS and these are described below. Other relevant sequence alignment and sequence editing programs are also discussed.

6.7.1 CAMELEON

This package (Oxford Molecular (1990)) allows the user to display two sequences and, if desired, the three-dimensional structure of one of them. The Gascuel-Golmard (1988) secondary structure prediction algorithm has been implemented along with routines to display a number of properties (for instance hydropathy) and to identify regions of similarity between the two sequences. CAMELEON is basically the program described by Morris (1988) with the addition of a simple c-alpha display. VISTAS has numerous advantages over the CAMELEON software. For instance VISTAS allows the manipulation of large alignments rather than single sequences and also allows the tertiary structure to be displayed in a number of different modes instead of just the simple c-alpha stick display of CAMELEON. The biggest advantages offered by VISTAS however are the routines which allow the integration of the sequence and structure displays along with the direct access to the database scanning and interrogation programs. In contrast, CAMELEON provides no interface to sequence databases and the structure display only allows simple translations and rotations rather than full sequence-structure interactions.

6.7.2 Integrated Structure and Sequence Displays

This package (Schnobel, R. (1991)) was written for SUN machines and allows the display of a sequence alignment and tertiary structure. The package allows the translation and rotation of the three-dimensional structure and a user may also redefine the colours used for each residue type. The structure and sequence displays are integrated, in that selecting a part of the sequence will lead to the appropriate part of the tertiary structure to be highlighted and vice-versa. A side-by-side stereo mode is also available for the three-dimensional structure display. This package is designed only for the visualisation of data and includes none of the database and amino acid property exploration routines of VISTAS.

6.8 Comparison of Sequence Alignment and Editing Programs.

As stated above, VISTAS has very powerful sequence alignment and editing capabilities. A number of the sequence alignment and sequence editing packages known to the author are described below to provide a comparison with the VISTAS program.

6.8.1 MANALIGN

This is perhaps the simplest of all the packages to be discussed as it was written specifically to be used on any terminal type, therefore no screenmode features could be included. This program is part of the LUPES package which was developed at Leeds (Akrigg, D. et al. (1988)). A maximum of ten sequences are displayed as lines of ASCII characters and gaps may be inserted by choosing an option from the menu, the latter is displayed as the last line on the terminal screen. A disadvantage with this system is that the top of the sequence alignment is lost as the screen scrolls to accommodate the menu. Symbols are used to display residue similarities between each sequence. MANALIGN performs a useful purpose as it can be run on any terminal screen, but as it lacks a screenmode any comparison with programs such as VISTAS and ALIGN are unhelpful.

6.8.2 HOMED

In contrast to MANALIGN, HOMED (Stockwell P.A. and Petersen, G.B. (1987)) allows sequences to be edited and listed in parallel as a screenmode display is used. This display is based on the EDT and KED text editors found on computers with VMS operating systems as the authors contend that as most users are familiar with text editors it is simple to use a sequence alignment program that behaves in a similar manner. Later versions are based on the EMACS editor found on machines running the UNIX operating system (Stockwell, P.A. (1988)). On a VAX computer HOMED may be used to edit up to 50 sequences, each with a maximum of 10240 residues. The program generates a consensus sequence showing the predominant residue type at each position in an alignment and also displays the residue type ('oily' or polar).

6.8.3 MASE

MASE (Faulkner, D.V and Jurka, J. (1988)) is designed to run on machines running the Berkeley UNIX (BSD) operating system and provides full-screen displays on a number of terminal types including VT100 compatibles. The number of sequences that the program can manipulate is limited only by the amount of memory available on the host machine. MASE has a number of basic operating modes (ie cursor movement and pattern searching, sequence modifications, window manipulations, output and sequence analysis) all directed by keyboard input. In this instance sequence analysis refers to functions such as the computation of consensus sequences and identity matrices. MASE also has a facility which allows particular residue types to be highlighted which aids the alignment of multiple sequences.

6.8.4 MALIGNED

MALIGNED (Clark, S.P. (1992)) is a sequence alignment and editing tool designed to run on VAX/VMS systems and a maximum of 199 sequences can be aligned at any one time. This program is again based on the VAX EDT editor and has a display that is designed to assist in aligning multiple sequences by variously highlighting residues. The simplest of these highlighting modes shows the most abundant residue type at a particular position in the primary highlight (bold), the second most abundant in the secondary highlight (intermediate), and the third most abundant residue type in tertiary highlight (least bold). Less frequent residue types have no highlight. MALIGNED also allows the user to group residue types, eg aromatic, and then uses these groups to perform highlighting instead of individual residue types. In addition, consensus sequences may be produced.

6.8.5 LINEUP

LINEUP is part of the GCG package produced at the University of Wisconsin (Devereux, J. et al. (1984)) and is a screenmode multiple alignment editor. A maximum of thirty sequences may be displayed at one time and a consensus sequence can also be produced, although it is not possible to display similarities between each sequence. Also, only limited pattern searching routines are available. LINEUP runs on both VMS and UNIX systems and requires a VT52 or compatible terminal.

6.8.6 SOMAP

SOMAP (Parry-Smith, D.J. and Attwood, T.K. (1991)) is a screenmode sequence alignment editor that was developed as part of the ADSP software package. Extensive use is made of the C curses library allowing rudimentary menus to be displayed, although all input is keyboard-based. SOMAP has no internal limitation on the number or length of the sequences to be aligned, the only constraint being the available memory on the host machine. A number of display options are available, simple sequences, sequences with similarities and a colour display. The colour display is perhaps the most useful features of SOMAP, although it is only available on the VAX/VMS version. Screen scrolling and update is also rather slow in any display mode. Comprehensive pattern searching routines are supported along with the ability to output alignments in a format suitable for monochrome laser printing. A post-processing program is available to produce hard copy of colour alignments from SOMAP output.

6.9 Conclusions from Comparisons

It may be noted that almost all the packages described above are limited to a particular platform, usually VMS, and have only a limited number of features. The sequence alignment part of the VISTAS and ALIGN packages were designed to incorporate as many functions as possible without appearing to be confusing to potential users while XALIGN allows complete portability, the X/Motif interface being consistent across all platforms means that the operating system is transparent to the user while using the program. Of the above software only SOMAP has a colour display which is an invaluable aid to sequence alignment, this facility is taken further in VISTAS and ALIGN by allowing the interactive colouring of alignments by various amino acid properties as well as residue type. VISTAS and ALIGN are also the only sequence alignment programs known to the author that are almost completely mouse driven, keyboard input being limited as much as possible (mainly just for filename definition).

6.10 Conclusion

It is the authors belief that VISTAS and ALIGN provide a rich functionality coupled with ease of use and that the VISTAS package address the issue of integrating primary, secondary and tertiary structure along with physicochemical measurements in a rational and user-friendly manner. VISTAS provides all the options a biologist would require to perform a sequence analysis study and is almost totally mouse driven. It can also be used as a tool for displaying the results of a sequence study, relating the sequence information produced to the structural information via the PRINTS and ,in due course, PROSITE database interfaces.

6.11 References

Akrigg, D., Bleasby, A.J., Dix, N.I.M, Findlay, J.B.C., North, A.C.T., Parry-Smith, D.J., Wootton, J.C., Blundell, T.L., Gardner, S.P., Hayes, F., Sternberg, M.J.E., Thornton, J.M., Tickle, I.J., A protein sequence/structure database. Nature **335** (1988) pp745-746

CAMELEON software package, Oxford Molecular Ltd. (1990)

Clark, S.P., MALIGNED: A multiple sequence alignment Editor. CABIOS **8** (1992) pp535-538

Devereux, J., Haeberli, P., Smithies, O., A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Research **12** (1984) pp387-395

Engelman, D.M., Steitz, T.A., Goldman, A., Identifying non-polar transbilayer helices in amino acid sequences of membrane proteins. Ann. Rev. Biophys. & Biophys. chem. **15** (1986) pp321-353

Faulkner, D.V., Jurka, J., Multiple aligned sequence editor (MASE). TIBS **13** (1988) pp321-322

Garnier, J., Osguthorpe, D.J., Robson, B., Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. **120** (1978) pp97-120

Gascuel, O., Golmard, J.L., A simple method for predicting the secondary structure of globular proteins - implications and accuracy. CABIOS **4** (1988) pp357-365

Kyte, J., Doolittle, R.F., A simple method for displaying the hydrophathic character of a protein. J. Mol. Biol. **157** (1982) pp105-132

Morris, G.M., The matching of protein sequences using colour intrasequence homology. J. Mol. Graph. **6** (1988) pp135-142

Parry-Smith, D.J., Attwood, T.K., SOMAP: A novel interactive approach to multiple sequences alignment. CABIOS **7** (1991) pp233-235

Pearson, W.R., Lipman, D.J., Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. **85** (1988) pp2444-2448

Rangone, R., Facchiano, F., Facchiano, A., Facchiano, A.M.,Colonna, G., Flexibility plot of proteins. Protein Engineering **2** (1989) pp497-504

Risler, J.L., Delorme, M.O., Delacroix, H., Henaut, A. Amino acid substitutions in structurally related proteins, A pattern recognition approach. J. Mol. Biol. **204** (1988) pp1019-1029

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehus, M.H., Hydrophobicity of amino acid residues in globular proteins. Science **229** (1985) pp834-838

Schnobel, R. Integrated displays of aligned amino acid sequences and protein strutures. CABIOS **7** (1991) pp341-346

Stockwell, P.A. HOMED: A Homologous sequence editor. TIBS **13** (1988) pp322-324

Stockwell, P.A., Petersen, G.B., HOMED: A homologous sequence editor. CABIOS **3** (1987) pp37-43

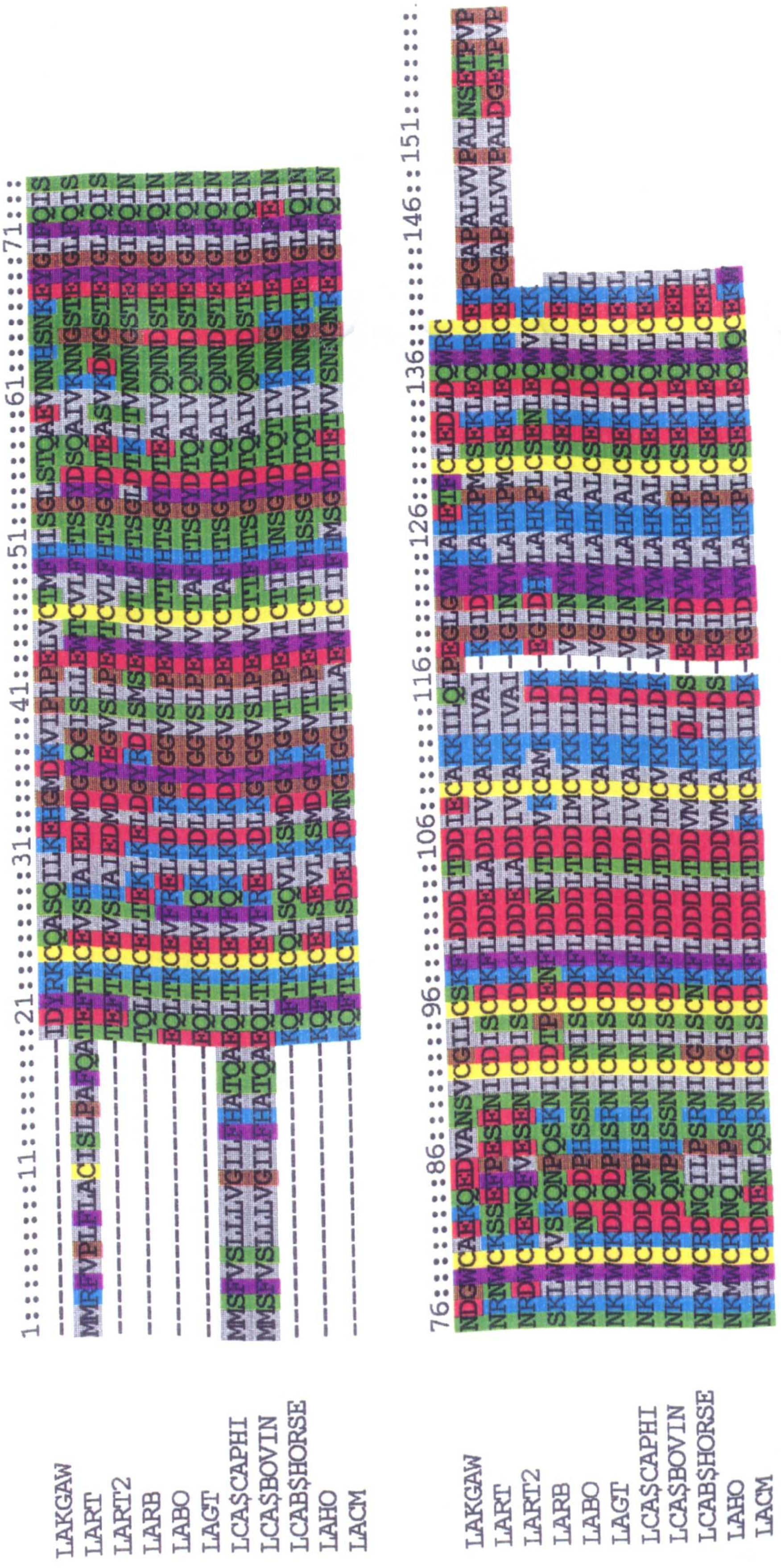
Sweet, R.M., Eisenberg, D., Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J. Mol. Biol. **171** (1983) pp479-488

Appendix A

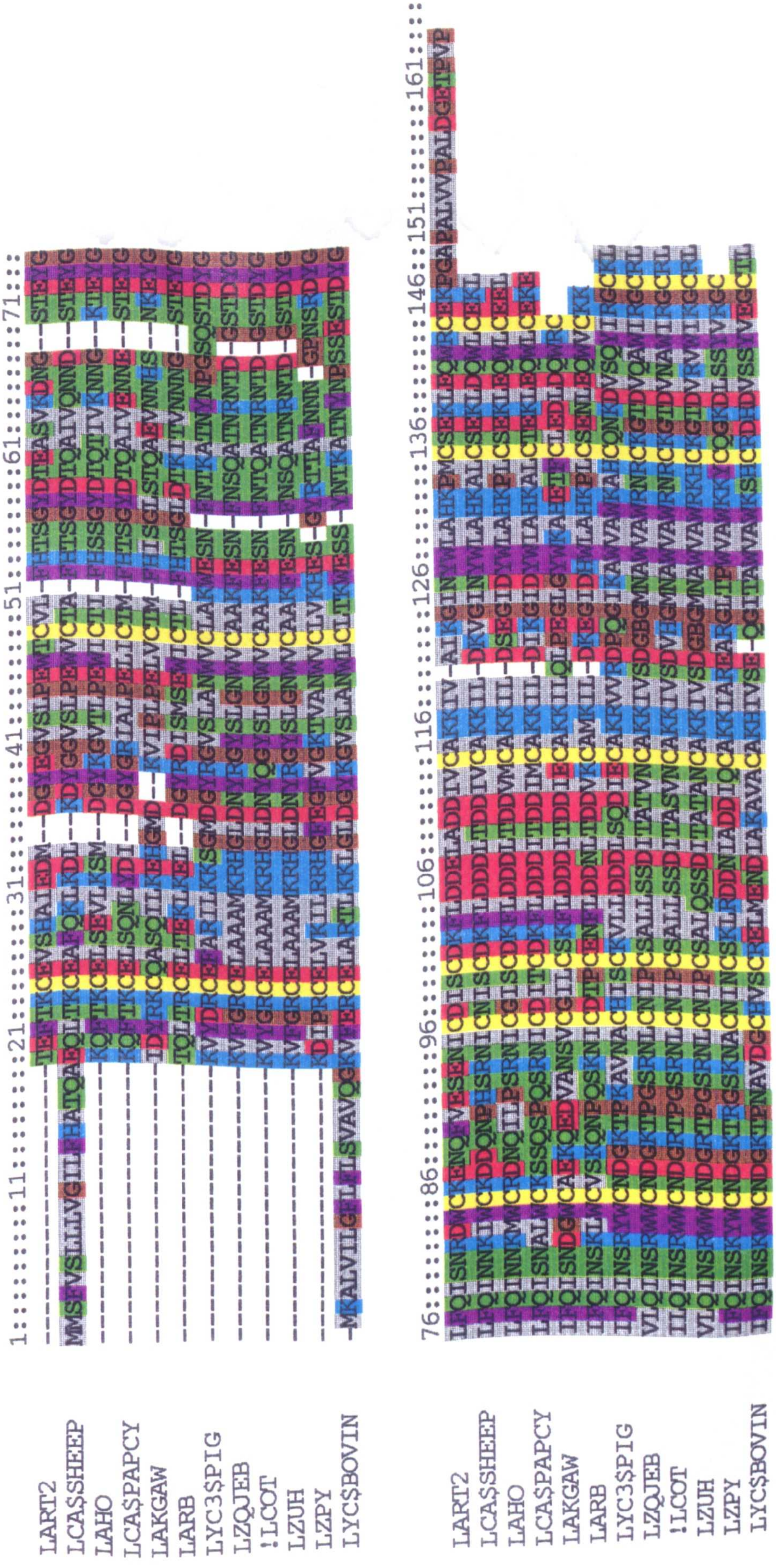
Sequence alignments used to define the motifs described in chapters three and four. Alignments were initially prepared using XALIGN, the SOCOL program was used to produce colour hard copy (Parry-Smith, D.J. personal communication). The key to the colours used for the alignments is shown in appendix D.

```
           1:::::::::11::::::::21::::::::31::::::::41::::::::51::::::::61::::::::71:::
LZCH       MRSLLILVLCFLPLAALGKVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNT-DGSTDYGIL
N$1LYM1    ------------------KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNT-DGSTDYGIL
N$2HFL3    ------------------KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNT-DGSTDYGIL
LZQJEB     ------------------KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNSQATNRNT-DGSTDYGVL
LZQJEC     ------------------KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNSQATNRNT-DGSTDYGVL
LZDK3      ------------------KVYERCELAAAMKRLGLDNYRGYSLGNWVCAANYESSFNTQATNRNT-DGSTDYGIL
N$1LZ11    ------------------KVFERCELARTLKRLGMDGYRGISLANWMCLAKWESGYNTRATNYNAGDRSTDYGIF
LYC1$PIG   ------------------KVYDRCEFARILKKSGMDGYRGVSLANWVCLAKWESDFNTKAINRNV--GSTDYGIF
LZBO       ------------------KVFERCELARTLKKLGLDGYKGVSLANWLCLTKWESSYNTKATNYNPSSESTDYGIF
LZOVE      ------------------KIYKRCELAAAMKRYGLDNYRGYSLGNWVCAARYESNYNTQATNRNSN-GSTDYGIL
LZBA       ------------------KIFERCELARTLKRLGLDGYRGISLANWVCLAKWESDYNTQATNYNPGDQSTDYGIF

           76::::::::86::::::::96::::::::106:::::::116:::::::126:::::::136:::::::146::
LZCH       QINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
N$1LYM1    QINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL
N$2HFL3    QINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGDGMNAWVAWRNRCKGTDVQAWIRGCRL
LZQJEB     QINSRWWCNDGKTPGSRNLCNIPCSALLSSDITATVNCAKKIVSDGBGMNAWVAWRNRCKGTDVQAWIRGCRL
LZQJEC     QINSRWWCNDGRTPGSRNLCNIPCSALLSSDITATVNCAKKIVSDGNGMNAWVAWRNRCKGTDVHAWIRGCRL
LZDK3      EINSRWWCDNGKTPRAKNACGIPCSVLLRSDITEAVKCAKRIVSDGDGMNAWVAWRNRCKGTDVSRWIRGCRL
N$1LZ11    QINSRYWCNDGKTPGAVNACHLSCSALLQDNIADAVACAKRVVRDPQGIRAWVAWRNRCQNRDVRQYVQGCGV
LYC1$PIG   QINSRYWCNDGKTPKAVNACHISCKVLLDDDLSQDIECAKRVVRDPQGIKAWVAWRTHCQNKDVSQYIRGCKL
LZBO       QINSKWWCNDGKTPNAVDGCHVSCSELMENDIAKAVACAKKIVSE-QGITAWVAWKSHCRDHDVSSYVEGCTL
LZOVE      QINSRWWCNDGRTPGTKNLCHISCSALMGADIAPSVRCAKRIVSDGDGMNAWVAWRKHCKGTDVSTWIKDCKL
LZBA       QINSHYWCNDGKTPGAVNACHISCNALLQDNITDAVACAKRVVSDPQGIRAWVAWRNHCQNRDVSQYVQGCGV
```

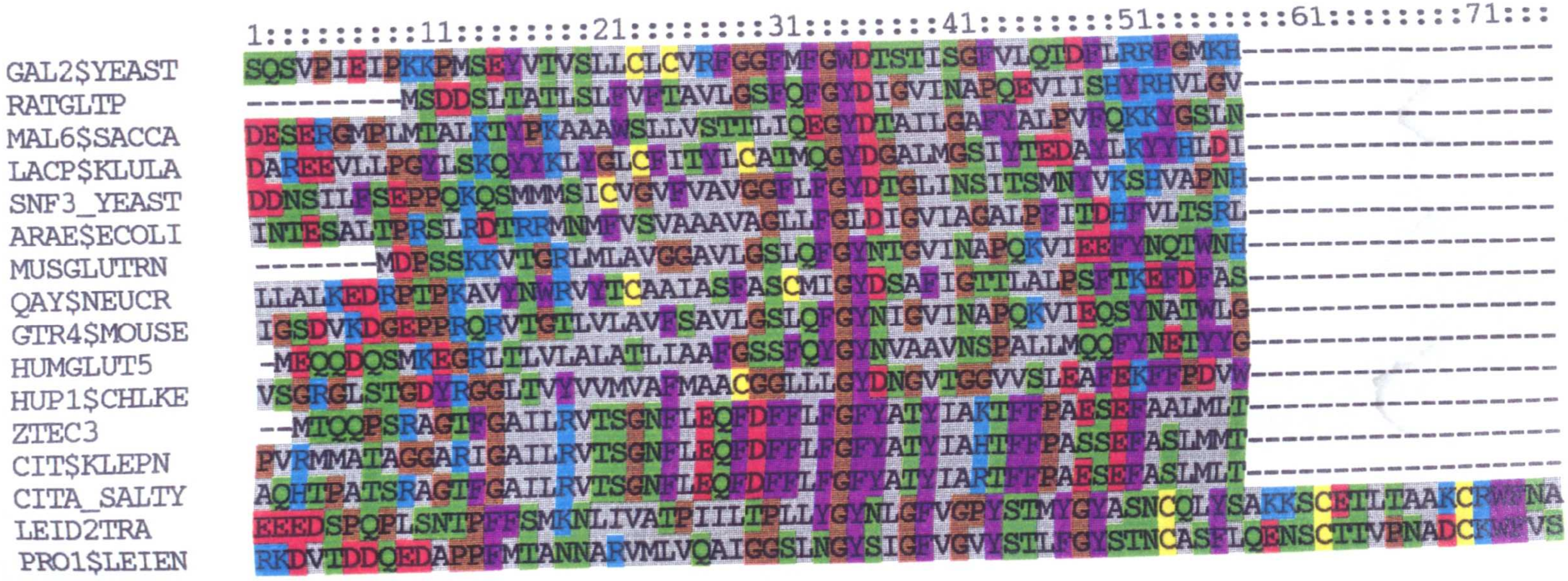
A.1.1 Initial alignment of lysozymes



A.1.2 Initial alignment of α lactalbumins



A.1.3 Initial alignment of lysozymes and α lactalbumins



```
              1:::::::::11::::::::21::::::::31::::::::41::::::::51::::::::61::::::::71:::
GAL2$YEAST    SQSVPIEIPKKPMSEYVTVSLLCLCVRFGGFMFGWDTSTISGFVLQTDFLRRFGMKH------------------
RATGLTP       ---------MSDDSLTATLSLFVFTAVLGSFQFGYDIGVINAPQEVIISHYRHVLGV------------------
MAL6$SACCA    DESERGMPLMTALKTYPKAAAWSLLVSTTLIQEGYDTAILGAFYALPVFQKKYGSLN------------------
LACP$KLULA    DAREEVLLPGYLSKQYYKLYGLCFITYLCATMQGYDGALMGSIYTEDAYLKYYHLDI------------------
SNF3_YEAST    DDNSILFSEPPQKQSMMMSICVGVFVAVGGFLFGYDTGLINSITSMNYVKSHVAPNH------------------
ARAE$ECOLI    INTESALTPRSLRDTRRMNMFVSVAAAVAGLLFGLDIGVIAGALPFITDHFVLTSRL------------------
MUSGLUTRN     -------MDPSSKKVTGRLMLAVGGAVLGSLQFGYNTGVINAPQKVIEEFYNQTWNH------------------
QAY$NEUCR     LLALKEDRPTPKAVYNWRVYTCAAIASFASCMIGYDSAFIGTTLALPSFTKEFDFAS------------------
GTR4$MOUSE    IGSDVKDGEPPRQRVTGTLVLAVFSAVLGSLQFGYNIGVINAPQKVIEQSYNATWLG------------------
HUMGLUT5      -MEQQDQSMKEGRLTLVLALATLIAAFGSSFQYGYNVAAVNSPALLMQQFYNETYYG------------------
HUP1$CHLKE    VSGRGLSTGDYRGGLTVYVVMVAFMAACGGLLLGYDNGVTGGVVSLEAFEKFFPDVW------------------
ZTEC3         --MTQQPSRAGTFGAILRVTSGNFLEQFDFFLFGFYATYIAKTFFPAESEFAALMLT------------------
CIT$KLEPN     PVRMMATAGGARIGAILRVTSGNFLEQFDFFLFGFYATYIAHTFFPASSEFASLMMT------------------
CITA_SALTY    AQHTPATSRAGTFGAILRVTSGNFLEQFDFFLFGFYATYIARTFFPAESEFASLMLT------------------
LEID2TRA      EEEDSPQPLSNTPFFSMKNLIVATPIILTPLLYGYNLGFVGPYSTMYGYASNCQLYSAKKSCETLTAAKCRWFNA
PRO1$LEIEN    RKDVTDDQEDAPPFMTANNARVMLVQAIGGSLNGYSIGFVGVYSTLFGYSTNCASFLQENSCTTVPNADCKWFVS
```

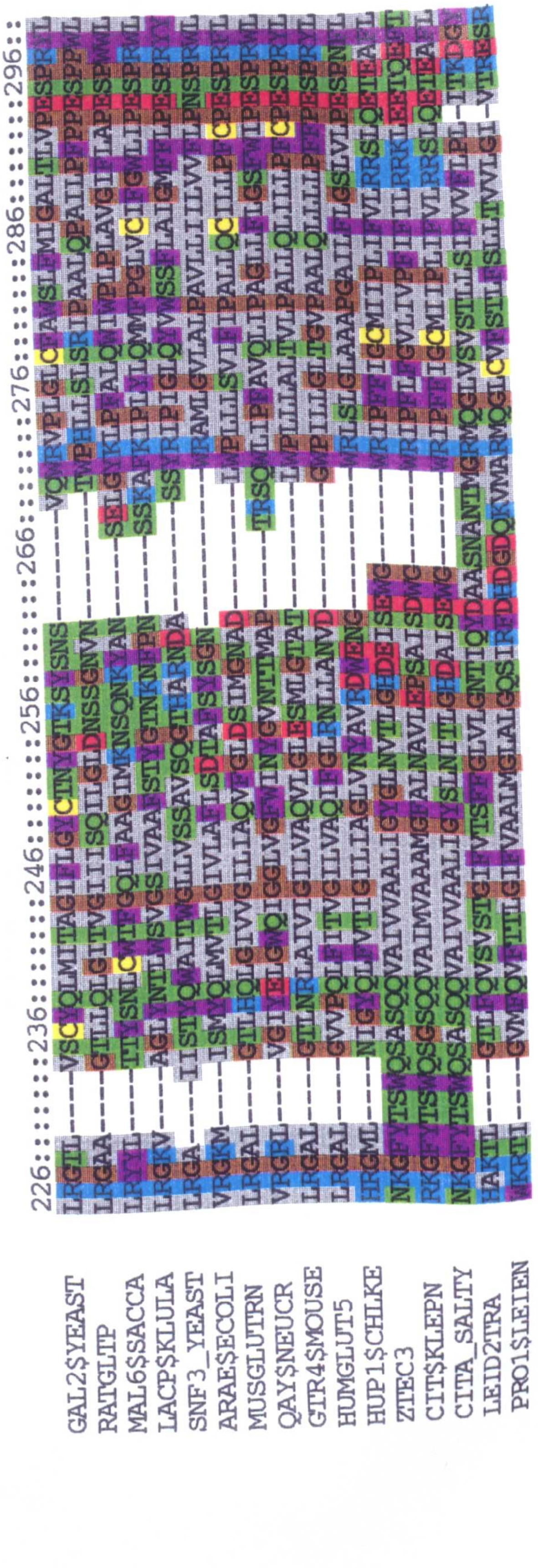
A.2.1 Sugar transporter and related proteins. For brevity, the extreme N and C-termini of some sequences have been removed

```
            76::::::::86::::::::96::::::::106:::::::116:::::::126:::::::136:::::::146::
GAL2$YEAST  -----------------------------------------------KDGTHYLSNVRTGLIVAIFNIGCAFGGI
RATGLTP     ------------PLDDRRATINYDINGTDTPLIVTPAHTTPDAWEEETEGSAHIVTMLWSLSVSSFAVGGMVASF
MAL6$SACCA  ----------------------------------------------SNTGDYEISVSWQIGLCLCYMAGEIVGLQ
LACP$KLULA  -----------------------------------------------------NSSSGTGLVFSIFNVGQICGAF
SNF3_YEAST  ---------------------------------------------------DSFTAQQMSILVSFLSLGTFFGAL
ARAE$ECOLI  ----------------------------------------------------------QEWVVSSMMLGAAIGAL
MUSGLUTRN   --------------------------------------------RYGEPIPSTTLTTLWSLSVAIFSVGGMIGSF
QAY$NEUCR   -------------------------------------------------YTPGALALLQSNIVSVYQAGAFFGCL
GTR4$MOUSE  ----------------------------------------RQGPGGPDSIPQGTLTTLWALSVAIFSVGGMISSF
HUMGLUT5    --------------------------------------------RTGEFMEDFPLTLLWSVTVSMFPFGGFIGSL
HUP1$CHLKE  --------------------------------------AKKQEVHEDSPYCTYDNAKLQLFVSSLF-LAGLVSCL
ZTEC3       ----------------------------------------------------------FAVFGSGFLMRP-IGAV
CIT$KLEPN   ----------------------------------------------------------FAVFGAGFLMRP-IGAI
CITA_SALTY  ----------------------------------------------------------FAVFGSGFLMRP-VGAI
LEID2TRA    STYVSNTTYGE-VCGWADRTTCFLKYSDE-AGCLSDSACKWSYSANTCGNQVGYSSIQSGVFAGSLVIGSTMGAL
PRO1$LEIEN  PTGSSYCGWPEVTCRKEYAYSSPAEMPGALARCEADSRCRWSYSDEECQNPSGYSSSESGIFAGSMIAGCLIGSV
```

A.2.1 continued

```
            151:::::::161:::::::171:::::::181:::::::191:::::::201:::::::211:::::::221::
GAL2$YEAST  ILSKGGDMYGRKKGLS-------IVVSVYIVGIIIQIASINKWYQYFIGRIISGLGVGGIAVLCPMLISEIAPKH
RATGLTP     FGGWLGDKLGRIKAML-----AANSLSLTGALLMGCSKFGPSHALIIAGRSVSGLYCGLISGLVPMYIGEISPHT
MAL6$SACCA  VTGPSVDYMGNRYILI--------MALFFLAAFIFILYFCKSLGMIAVGQALCGMPWGCFQCLTVSYASEICPLA
LACP$KLULA  FVP-LMDWKGRKPAIL--------IGCLGVVIGAIISSLTTTKSALIGGRWFVAFFATIANAAAPTYCAEVAPAH
SNF3_YEAST  TAPFISDSYGRKPTII-------FSTIFIFSIGNSLQVGAGGITLLIVGRVISGIGIGAISAVVPLYQAEATHKS
ARAE$ECOLI  FNGWLSFRLGRKYSLM--------AGAILFVLGSIGSAFATSVEMLIAARVVLGIAVGIASYTAPLYLSEMASEN
MUSGLUTRN   SVGLFVNRFGRRNSMLM-----MNLLAFVAAVLMGFSKLGKSFEMLILGRFIIGVYCGLTTGFVPMYVGEVSPTA
QAY$NEUCR   FAYATSYFLGRRKSLIA----FSVVFIIGAAIMLAADGQGRGIDPIIAGRVLAGIGVGGASNMVPIYISELAPPA
GTR4$MOUSE  LIGIISQWLGRKRAML-----ANNVLAVLGGALMGLANAVASYEILILGRFLIGAYSGLTSGLVPMYVGEIAPTH
HUMGLUT5    LVGPLVNKFGRKGALL-----FNNIFSIVPAILMGCSRVATSFELIIISRLLVGICAGVSSNVVPMYLGELAPKN
HUP1$CHLKE  FASWITRNWGRKVTMG--------IGGAFFVAGGLVNAFAQDMAMLIVGRVLLGFGVGLGSQVVPQYLSEVAPFS
ZTEC3       VLGAYIDRIGRRKGLMVTLAIMGCGTLLIALVPGYQTIGLLAPVLVLVGRLLQGFSAGVELGGVSVYLSEIATPG
CIT$KLEPN   VLGAYIDKVGRRKGLIVTLSIMATGTFLIVLIPSYQTIGLWAPLLVLIGRLLQGFSAGAELGGVSVYLAEIATPG
CITA_SALTY  VLGAYIDRIGRRKGLMVTLAIMGCGTLLIALVPGYQTIGLAAPALVLLGRLLQGFSAGVELGGVSVYLSEIATPG
LEID2TRA    MGGYLTKRLDYCKSFL-----FIGLLSVIGNVLTHVATGLFHYWVLFVARIVLGFPLGWQSITSSHYTDKFAPAN
PRO1$LEIEN  FAGPLASKIGARLSFL-----LVGLVGVVASVMYHASCAADEFWVLIVGRFVIGLFLGVICVACPVYIDQNAHPK
```

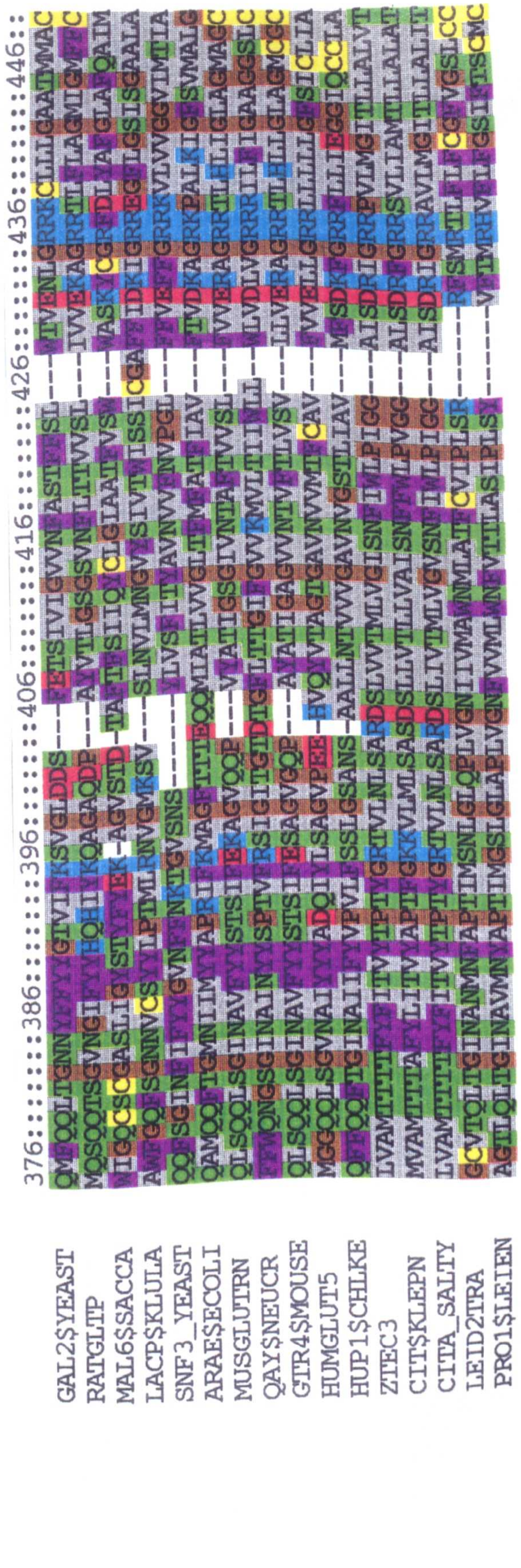
A.2.1 continued



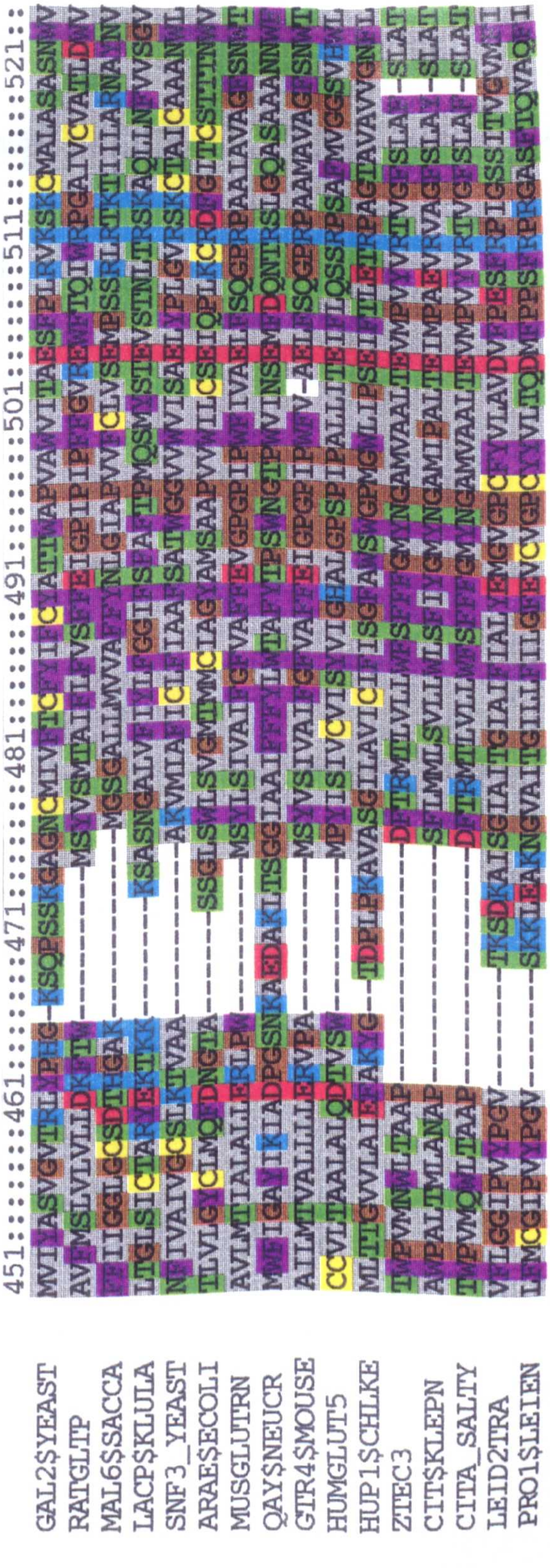
A.2.1 continued

```
              301:::::::311:::::::321:::::::331:::::::341:::::::351:::::::361:::::::371::
GAL2$YEAST    CEVNKVED-AKRSIAKSNKVSPEDPAV--QAELDLIMAGIEAEKLAGNAS-----WGELFSTKTKVFQRLLMGVFV
RATGLTP       TIDIDDEGNAKRILQSLQGYDEVSHE----LQEIKDESQKEEAETFLTLIE-----------LLR-SRDTRWPSLLIAFL
MAL6$SACCA    VKKGRIDQ-ARRSLERILSGKGPEKELLVSMELDKIKTTIEKEQKMSDEG----TYWDCVK-DGINRRRRTRIACLC
LACP$KLULA    VGVGREEE-AREFIIKYHLNGDRTHPLL-DMEMAEIIESFHGTDLSNPLE-MLDVRSLFR-TRSDRYRAMLVILM
SNF3_YEAST    VLKDKLDEAAKSLSFLRGVPVHDSGLLEELVEIKATYDYEASFGSSNFIDCFISSKS-RPKQTLRMFTGIALQAF
ARAE$ECOLI    AEKGRHIEAEEVLRMLRDTSEKAREE----LNEIRESLKLKQGG-------------WALFKINRNVRRAVFLGMLL
MUSGLUTRN     LINRNEENRAKSVLKKLRGTADVTRD----LQEMKEEGRQMMREKKVTIL---------ELFR-SPAYRQPILIAVVL
QAY$NEUCR     YANGKREEAMKVLCWIRNLEPTDRYI----VQEVSFIDADLERYTRQVGN-GFWKPFLSLK-QRKVQWRFFLGGML
GTR4$MOUSE    YIIRNLEGPARKSLKPLTGWADVSDA----LAELKDEKRKLERERPMSLL---------QLLG-SRTHRQPLIIAVVL
HUMGLUT5      LIQKKDEAAAKKALQTLRGWDSVDRE----VAEIRQEDEAEKAAGFISV---------LKLFR-MRSLRWQLLSIIVL
HUP1$CHLKE    VEKGKTE-KGREVLQKLRGTSEVDAE----FADIVAAVEIARPITMRQSWA---------SLFT--RRYMPQLLTSFVI
ZTEC3         QRKHRPD--TREIFTT--------------------------------------------IAKNWRIITAGIL
CIT$KLEPN     ARRHHLA--MRQVFAT--------------------------------------------LLANWQVVIAGMM
CITA_SALTY    QRKHRPD--TREIFAT--------------------------------------------IAKNWRIITAGIL
LEID2TRA      SKSRRGDYEGENSEDASRKAA---------------------------------EEYTMTQMIGPILNGVAM
PRO1$LEIEN    AKFDGGEEGRAELN---------------------------------------PSEYGYVEMIPRLLMGCVM
```

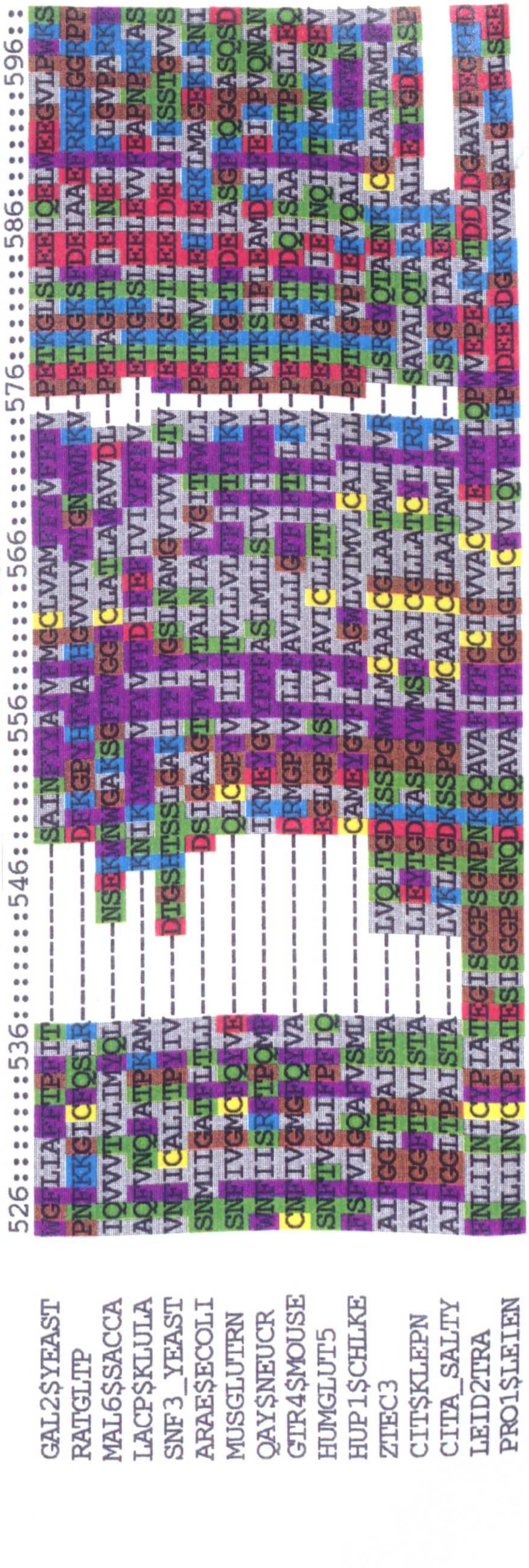
A.2.1 continued



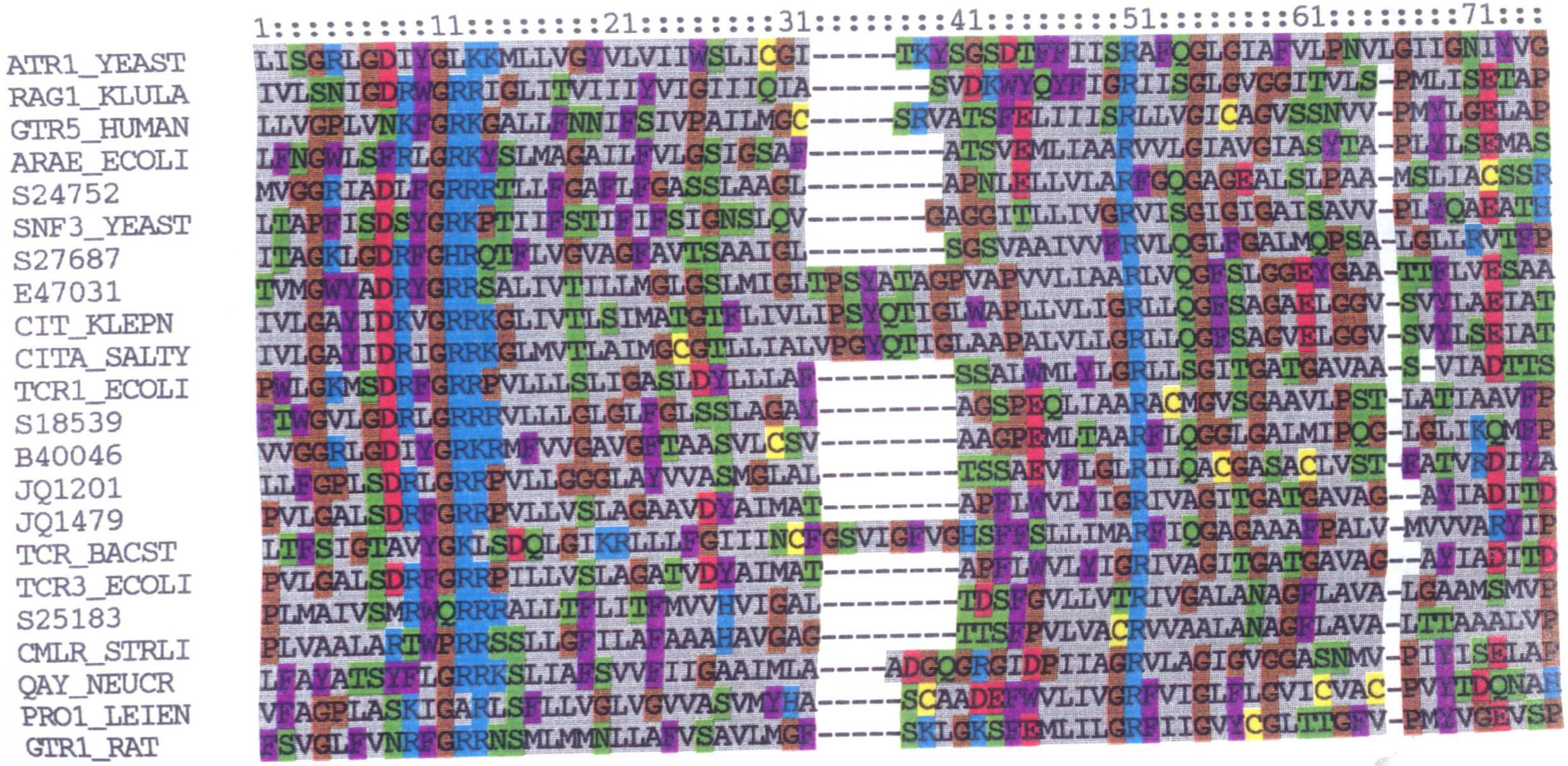
A.2.1 continued



A.2.1 continued

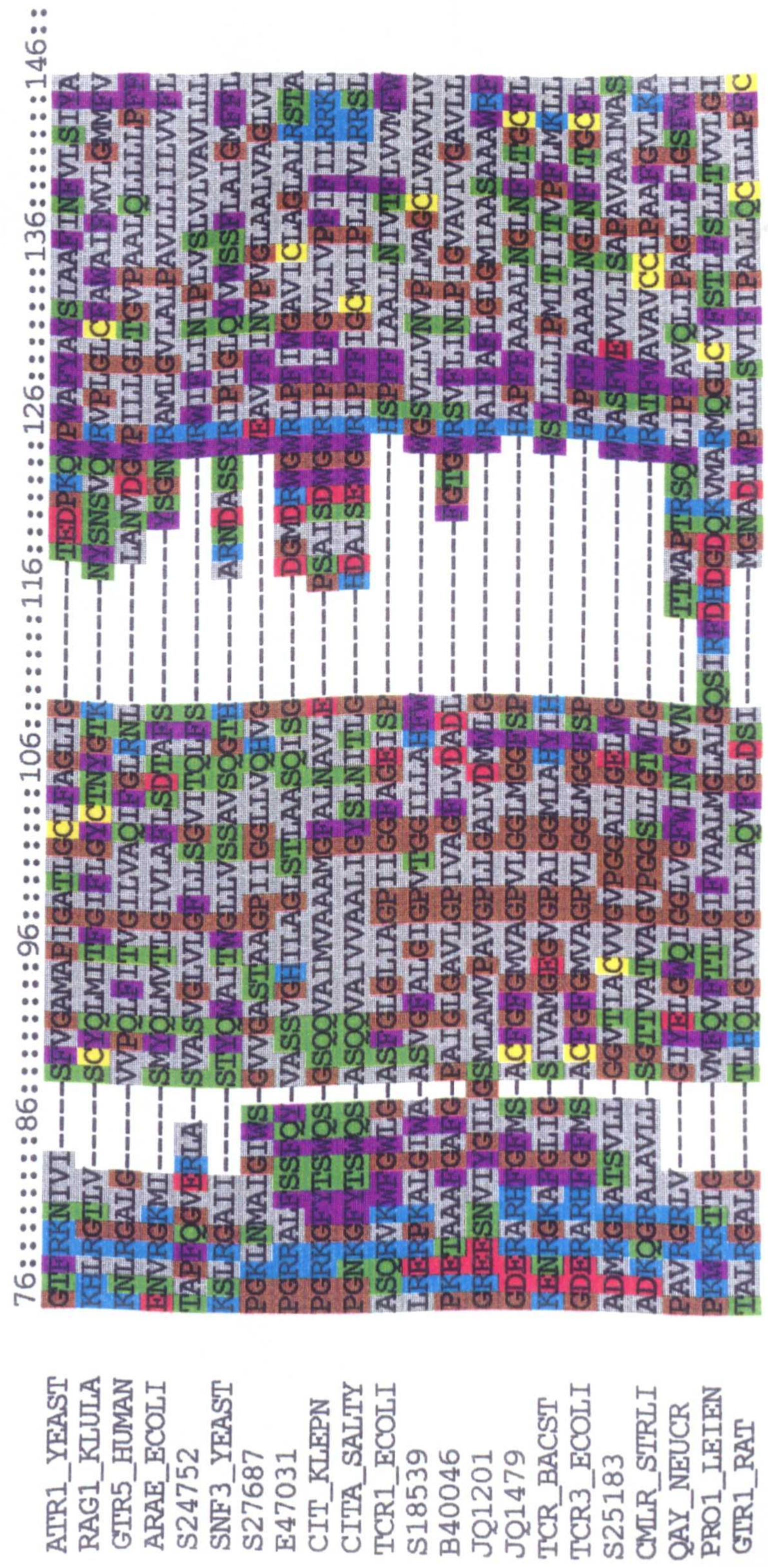


A.2.1 continued



```
             1:::::::::11:::::::::21:::::::::31:::::::::41:::::::::51:::::::::61:::::::::71:::
ATR1_YEAST   LISGRLGDIYGLKKMLLVGYVLVIIWSLICGI------TKYSGSDIFFIISRAFQGLGIAFVLPNVLGIIGNIYVG
RAG1_KLULA   IVLSNIGDRWGRRIGLITVIIIYVIGIIIQIA--------SVDKWYQYFIGRIISGLGVGGITVLS-PMLISETAP
GTR5_HUMAN   LLVGPLVNKFGRKGALLFNNIFSIVPAILMGC------SRVATSFELIIISRLLVGICAGVSSNVV-PMYLGELAP
ARAE_ECOLI   LFNGWLSFRLGRKYSLMAGAILFVLGSIGSAF---------ATSVEMLIAARVVLGIAVGIASYTA-PLYLSEMAS
S24752       MVGGRIADLFGRRRTLLFGAFLFGASSLAAGL---------APNLELLVLARFGQGAGEALSLPAA-MSLIACSSR
SNF3_YEAST   LTAPFISDSYGRKPTIIFSTIFIFSIGNSLQV--------GAGGITLLIVGRVISGIGIGAISAVV-PLYQAEATH
S27687       ITAGKLGDRFGHRQTFLVGVAGFAVTSAAIGL---------SGSVAAIVVFRVLQGLFGALMQPSA-LGLLRVTFP
E47031       TVMGWYADRYGRRSALIVTILLMGLGSLMIGLTPSYATAGPVAPVVLIAARLVQGFSLGGEYGAA-TTFLVESAA
CIT_KLEPN    IVLGAYIDKVGRRKGLIVTLSIMATGTFLIVLIPSYQTIGLWAPLLVLIGRLLQGFSAGAELGGV-SVYLAEIAT
CITA_SALTY   IVLGAYIDRIGRRKGLMVTLAIMGCGTLLIALVPGYQTIGLAAPALVLLGRLLQGFSAGVELGGV-SVYLSEIAT
TCR1_ECOLI   PWLGKMSDRFGRRPVLLLSLIGASLDYLLLAF----------SSALWMLYLGRLLSGITGATGAVAA-S-VIADTTS
S18539       FTWGVLGDRLGRRRVLLLGLGLFGLSSLAGAY----------AGSPEQLIAARACMGVSGAAVLPST-LATIAAVFP
B40046       VVGGRLGDIYGRKRMFVVGAVGFTAASVLCSV----------AAGPEMLTAARFLQGGLGALMIPQG-LGLIKQMFP
JQ1201       LLFGPLSDRLGRRPVLLGGGLAYVVASMGLAL----------TSSAEVFLGLRILQACGASACLVST-FATVRDIYA
JQ1479       PVLGALSDRFGRRPVLLVSLAGAAVDYAIMAT----------APFLWVLYIGRIVAGITGATGAVAG--AYIADITD
TCR_BACST    LTFSIGTAVYGKLSDQLGIKRLLLFGIIINCFGSVIGFVGHSFFSLLIMARFIQGAGAAAFPALV-MVVVARYIP
TCR3_ECOLI   PVLGALSDRFGRRPILLVSLAGATVDYAIMAT----------APFLWVLYIGRIVAGITGATGAVAG--AYIADITD
S25183       PLMAIVSMRWQRRRALLTFLITFMVVHVIGAL----------TDSFGVLLVTRIVGALANAGFLAVA-LGAAMSMVP
CMLR_STRLI   PLVAALARTWPRRSSLLGFILAFAAAHAVGAG----------TTSFPVLVACRVVAALANAGFLAVA-LTTAAALVP
QAY_NEUCR    LFAYATSYFLGRRKSLIAFSVVFIIGAAIMLA----ADGQGRGIDPIIAGRVLAGIGVGGASNMV-PIYISELAF
PRO1_LEIEN   VFAGPLASKIGARLSFLLVGLVGVVASVMYHA------SCAADEFWVLIVGRFVIGLFLGVICVAC-PVYTDQNAH
GTR1_RAT     FSVGLFVNRFGRRNSMLMMNLLAFVSAVLMGF------SKLGKSFEMLILGRFIIGVYCGLTTGFV-PMYVGEVSP
```

A.2.2 Proton symport/antiport sequences. *Only the section of the alignment from putative transmembrane segments 3 to 6 is shown.*



A.2.2 continued

Appendix B

The final motif sets and key to the sequences shown in chapters three and four. The numbers shown after the motifs indicate the number of residues from the N-terminus and the number of residues from the previous motif respectively.

B.1.1 Final lactalbumin motifs

LAGT	Alpha-lactalbumin - Goat
LCA$CAPHI	ALPHA-LACTALBUMIN PRECURSORC - Goat
LCA$SHEEP	ALPHA-LACTALBUMIN PRECURSOR - Sheep
LCA$BOVIN	ALPHA-LACTALBUMIN PRECURSOR - Bovine
LAHU	Alpha-lactalbumin precursor - Human
LABO	Alpha-lactalbumin - Bovine
LAHO	Alpha-lactalbumin - Horse
LCAB$HORSE	ALPHA-LACTALBUMIN B AND C - Horse
LCA$PAPCY	ALPHA-LACTALBUMIN - Yellow baboon
LAGP	Alpha-lactalbumin precursor - Guinea pig
LACM	Alpha-lactalbumin - Arabian camel
GPILACTAL	GPILACTAL pre-alpha-lactalbumin - *Cavia porcellus*
EZEC228	ALPHA-LACTALBUMIN - Bovine
LART2	Alpha-lactalbumin (version 2) - Rat
LART	Alpha-lactalbumin (version 2) - Rat
LARB	Alpha-lactalbumin - Rabbit
LAKGAW	Alpha-lactalbumin - Red-necked wallaby

LYC$EQUAS	LYSOZYME C - Donkey
LYC$HORSE	LYSOZYME C - Horse
LYC1$PIG	LYSOZYME C-1 - Pig
LYC2$PIG	LYSOZYME C-2 - Pig
LYC3$PIG	LYSOZYME C-3 - Pig

Database Version - OWL9.0

<u>motif 1</u>

EVFRELKDLKGYGGVSLPEWV	LABO	7	7
EVFRELKDLKGYGGVSLPEWV	LCA$BOVIN	26	26
EVFRELKDLKGYGGVSLPEWV	EZEC228	7	7
ELSQLLKDIDGYGGIALPELI	LAHU	26	26
EVFQKLKDLKDYGGVSLPEWV	LAGT	7	7
EVFQKLKDLKDYGGVSLPEWV	LCA$CAPHI	26	26
EAFQKLKDLKDYGGVSLPEWV	LCA$SHEEP	26	26
ELSEVLKSMDGYKGVTLPEWI	LAHO	7	7
QLSQVLKSMDGYKGVTLPEWI	LCAB$HORSE	7	7
ELSQNLYDIDGYGRIALPELI	LCA$PAPCY	7	7
EVSHAIEDMDGYEGVSLPEWT	LART2	7	7
ALSHELNDLAGYRDITLPEWL	LAGP	26	26
ALSHELNDLAGYRDITLPEWL	GPILACTAL	26	26
KLSDELKDMNGHGGITLAEWI	LACM	7	7
EVSHAIEDMDGYQGISLLEWT	LART	26	26
ELTEKLKELDGYRDISMSEWI	LARB	7	7
QASQILKEHGMDKVIPLPELV	LAKGAW	7	7

motif 2

FHTSGYDTEAIV	LABO	31	3
FHTSGYDTQAIV	LCA$BOVIN	50	3
FHTSGYDTEAIV	EZEC228	31	3
FHTSGYDTQAIV	LAHU	50	3
FHTSGYDTQAIV	LAGT	31	3
FHTSGYDTQAIV	LCA$CAPHI	50	3
FHTSGYDTQAIV	LCA$SHEEP	50	3
FHSSGYDTQTIV	LAHO	31	3
FHNSGYDTQTIV	LCAB$HORSE	31	3
FHTSGYDTQAIV	LCA$PAPCY	31	3
FHTSGYDTEASV	LART2	31	3
FHISGYDTQAIV	LAGP	50	3
FHISGYDTQAIV	GPILACTAL	50	3
FHMSGYDTETVV	LACM	31	3
FHTSGYDSQAIV	LART	50	3
FHTSGLDTKITV	LARB	31	3
FHISGLSTQAEV	LAKGAW	31	3

motif 3

HSSNICNISC	LABO	68	25
HSSNICNISC	LCA$BOVIN	87	25
HSSNICNISC	EZEC228	68	25
QSRNICDISC	LAHU	87	25
HSRNICNISC	LAGT	68	25
HSRNICNISC	LCA$CAPHI	87	25
HSRNICNISC	LCA$SHEEP	87	25
PSRNICGISC	LAHO	68	25
PSRNICGISC	LCAB$HORSE	68	25
QSRNICDITC	LCA$PAPCY	68	25
ESENICDISC	LART2	68	25
QSRNICDISC	LAGP	87	25
QSRNICDISC	GPILACTAL	87	25
QSRNICDISC	LACM	68	25
ESENICDISC	LART	87	25
QSKNICDTPC	LARB	68	25
VANSVCGILC	LAKGAW	68	25

motif 4

KFLDDDLTDD	LABO	79	1
KFLDDDLTDD	LCA$BOVIN	98	1
KFLNNDLTNN	EZEC228	79	1
KFLDDDITDD	LAHU	98	1
KFLDDDLTDD	LAGT	79	1
KFLDDDLTDD	LCA$CAPHI	98	1
KFLDDDLTDD	LCA$SHEEP	98	1
KFLDDDLTDD	LAHO	79	1
KFLDDDLTDD	LCAB$HORSE	79	1
KFLDDDITDD	LCA$PAPCY	79	1
KFLDDELADD	LART2	79	1
KLLDDDLTDD	LAGP	98	1
KLLDDDLTDD	GPILACTAL	98	1
KFLDDDLTDD	LACM	98	1
KFLDDELADD	LART	98	1
NFLDDNLTDD	LARB	79	1
KFLDDDITDD	LAKGAW	79	1

motif 5

VGINYWLAH	LABO	99	10
VGINYWLAH	LCA$BOVIN	118	10
VGINYWLAH	EZEC228	99	10
KGIDYWLAH	LAHU	118	10
VGINYWLAH	LAGT	99	10
VGINYWLAH	LCA$CAPHI	118	10
VGINYWLAH	LCA$SHEEP	118	10
EGIDYWLAH	LAHO	99	10
EGIDYWLAH	LCAB$HORSE	99	10
KGIDYWIAH	LCA$PAPCY	99	10
KGINYWLAH	LART2	99	10
KGIDYWLAH	LAGP	118	10
KGIDYWFAH	GPILACTAL	118	10
EGIDYWLAH	LACM	99	10
KGIDYWKAH	LART	118	10
EGIDHWLAH	LARB	99	10
EGLGYWKAH	LAKGAW	100	11

motif 6

CSEKLDQWLC	LABO	111	3
CSEKLDQWLC	LCA$BOVIN	130	3
CSEKLDQWLC	EZEC228	111	3
CTEKLEQWLC	LAHU	130	3
CSEKLDQWLC	LAGT	111	3
CSEKLDQWLC	LCA$CAPHI	130	3
CSEKLDQWLC	LCA$SHEEP	130	3
CSEKLEQWLC	LAHO	111	3
CSEKLEQWLC	LCAB$HORSE	111	3
CTEKLEQWLC	LCA$PAPCY	111	3
CSEKLEQWRC	LART2	111	3
CSDKLEQWYC	LAGP	130	3
CSDKLEQWYC	GPILACTAL	130	3
CSEKLEQWQC	LACM	111	3
CSEKLEQWRC	LART	130	3
CSENLEQWVC	LARB	111	3
CLEDLDQWRC	LAKGAW	112	3

B.1.2 Final lysozyme motifs

LZCH	Lysozyme c precursor - Chicken
N$1LYMA	Lysozyme chain A - Hen egg white
N$1LYMB	Lysozyme chain B - Hen egg white
N$1LYZ	Lysozyme - Hen egg white
N$1LZHA	Lysozyme chain A - Hen egg white
N$1LZHB	Lysozyme chain B - Hen egg white
N$2HFMY	Lysozyme - Chicken
N$2LYM	Lysozyme - Hen egg white
N$2LYZ	Lysozyme - Hen egg white
N$2LZH	Lysozyme - Hen egg white
N$2LZT	Lysozyme - Hen egg white
N$3HFMY	Lysozyme - Hen egg white
N$3LYM	Lysozyme - Hen egg white
N$3LYZ	Lysozyme - Hen egg white
N$4LYZ	Lysozyme - Hen egg white
N$5LYZ	Lysozyme - Hen egg white
N$6LYZ	Lysozyme - Hen egg white
N$7LYZ	Lysozyme - Hen egg white

N$8LYZ	Lysozyme - Hen egg white
S05657	Lysozyme c - Chicken
N$2HFLY	Lysozyme c - Chicken
!LCOT	LYSOZYME - Coturnix
JT0526	Lysozyme c - Indian peafowl
EZEC462	LYSOZYME - Turkey
N$1LZ2	Lysozyme - Turkey egg white
LYC$MELGA	LYSOZYME C PRECURSOR - Turkey
N$2LZ2	Lysozyme - Turkey egg white
EZEC471	LYSOZYME - Bobwhite quail
LZQJEC	Lysozyme c - California quail
LZQJEB	Lysozyme c - Common bobwhite
LZFER	Lysozyme c precursor - Ring-necked pheasant
EZEC470	LYSOZYME - Guinea hen
LZUH	Lysozyme c - Helmeted guineafowl
EZEC465	LYSOZYME II - Kaki duck
EZEC466	LYSOZYME - Duck III
LZQJE	Lysozyme c - California quail
LZDK3	Lysozyme c III - Duck
LZDK	Lysozyme c precursor - Duck
LZTK	Lysozyme c precursor - Turkey
LZOVE	Lysozyme c - Plain chachalaca
LZBA	Lysozyme - Baboon
LZHU	Lysozyme - Human
HUMLSZA	Lysozyme precursor - *Homo sapiens*
N$1LZ1	Lysozyme - Human
LYC$RABIT	LYSOZYME C - Rabbit
HUMLYZ	HUMLYZ lysozyme - Artificial gene
LYC$PREEN	LYSOZYME C - *Hanuman langur*
LYCP$MOUSE	LYSOZYME C - Mouse
LYCM$MOUSE	LYSOZYME C - Mouse
LYC3$PIG	LYSOZYME C-3 - Pig
LYC1$PIG	LYSOZYME C-1 - Pig
LZRT	Lysozyme - Rat
BOVLSZ3A	lysozyme 3a precursor - *Bos taurus*
LZBO	Lysozyme c 2 - Bovine
LYC$AXIAX	LYSOZYME C 1 AND 2 - Axis deer
LYC$SHEEP	LYSOZYME C 1A TO 4B - Sheep
BOVLSZ1A	lysozyme 1a precursor - Bos taurus
LYC2$PIG	LYSOZYME C-2 - Pig
LYC$BOVIN	LYSOZYME C PRECURSOR - Bovine
LYC$EQUAS	LYSOZYME C - Donkey
LYC$HORSE	LYSOZYME C - Horse
LZPY	Lysozyme c - Pigeon

Database Version - OWL11.0

motif 1

VFGRCELAAAMKRHGLDN	LZCH	20	20
VFGRCELAAAMKRHGLDN	LZQJEC	2	2
VFGRCELAAAMKRHGLDN	LZQJEB	2	2
VFGRCELAAAMKRHGLDN	LZUH	2	2
VFGRCELAAAMKRHGLDN	EZEC470	2	2
VFGRCELAAAMKRHGLDN	EZEC471	2	2
VFGRCELAAAMKRHGLDN	S05657	2	2
VFGRCELAAAMKRHGLDN	N$1LYMA	2	2

VFGRCELAAAMKRHGLDN	N$1LYMB	2	2
VFGRCELAAAMKRHGLDN	N$1LYZ	2	2
VFGRCELAAAMKRHGLDN	N$1LZHA	2	2
VFGRCELAAAMKRHGLDN	N$1LZHB	2	2
VFGRCELAAAMKRHGLDN	N$2HFLY	2	2
VFGRCELAAAMKRHGLDN	N$2HFMY	2	2
VFGRCELAAAMKRHGLDN	N$2LYM	2	2
VFGRCELAAAMKRHGLDN	N$2LYZ	2	2
VFGRCELAAAMKRHGLDN	N$2LZH	2	2
VFGRCELAAAMKRHGLDN	N$2LZT	2	2
VFGRCELAAAMKRHGLDN	N$3HFMY	2	2
VFGRCELAAAMKRHGLDN	N$3LYM	2	2
VFGRCELAAAMKRHGLDN	N$3LYZ	2	2
VFGRCELAAAMKRHGLDN	N$4LYZ	2	2
VFGRCELAAAMKRHGLDN	N$5LYZ	2	2
VFGRCELAAAMKRHGLDN	N$6LYZ	2	2
VFGRCELAAAMKRHGLDN	N$7LYZ	2	2
VFGRCELAAAMKRHGLDN	N$8LYZ	2	2
VYGRCELAAAMKRHGLDN	!LCOT	2	2
VYGRCELAAAMKRLGLDN	LYC$MELGA	20	20
VYGRCELAAAMKRLGLDN	JT0526	2	2
VYGRCELAAAMKRLGLDN	EZEC462	2	2
VYGRCELAAAMKRLGLDN	N$1LZ2	2	2
VYGRCELAAAMKRLGLDN	N$2LZ2	2	2
VYERCELAAAMKRLGLDN	LZDK3	2	2
VYGRCELAAAMKRMGLDN	LZFER	20	20
VYGRCELAAAMKRHGLDK	LZQJE	20	20
VYSRCELAAAMKRLGLDN	LZDK	20	20
VYSRCELAAAMKRLGLDN	EZEC465	2	2
VYQRCELAAAMKRLGLDN	EZEC466	2	2
VYGRCELAAAMKRLGLBB	LZTK	20	20
IYKRCELAAAMKRYGLDN	LZOVE	2	2
VFERCELARTLKRLGMDG	LZHU	2	2
VFERCELARTLKRLGMDG	HUMLSZA	20	20
VFERCELARTLKRLGMDG	HUMLYZ	3	3
VFERCELARTLKRLGMDG	N$1LZ1	2	2
VFERCELARTLKKLGLDG	LZBO	2	2
VFERCELARTLKKLGLDG	LYC$BOVIN	20	20
VFERCELARTLKKLGLDG	BOVLSZ1A	20	20
VFERCELARTLKKLGLDG	BOVLSZ3A	20	20
VFERCELARTLKELGLDG	LYC$AXIAX	2	2
VFERCELARTLKELGLDG	LYC$SHEEP	2	2
IFERCELARTLKRLGLDG	LZBA	2	2
IFERCELARTLKKLGLDG	LYC$PREEN	2	2
VYNRCELARILKRNGMDG	LYCP$MOUSE	20	20
IYERCELARTLKKLGLDG	LYC$RABIT	2	2
VYERCEFARTLKRNGMAG	LYCM$MOUSE	20	20
VYDRCEFARILKKSGMDG	LYC1$PIG	2	2
VYDRCEFARILKKSGMDG	LYC2$PIG	2	2
VYDRCEFARILKKSGMDG	LYC3$PIG	2	2
TYERCEFARTLKRNGMSG	LZRT	2	2
VFSKCELAHKLKAQEMDG	LYC$EQUAS	2	2
VFSKCELAHKLKAQEMDG	LYC$HORSE	2	2
DIPRCELVKILRRHGFEG	LZPY	2	2

motif 2

KFESNFNTQATNR	LZCH	51	13
KFESNFNSQATNR	LZQJEC	33	13
KFESNFNSQATNR	LZQJEB	33	13
KFESNFNSQATNR	LZUH	33	13
KFESNFNSQATNR	EZEC470	33	13
KFESNFNSQATNR	EZEC471	33	13
KFESNFNTQATNR	S05657	33	13
KFESNFNTQATNR	N$1LYMA	33	13
KFESNFNTQATNR	N$1LYMB	33	13
KFESNFNTQATNR	N$1LYZ	33	13
KFESNFNTQATNR	N$1LZHA	33	13
KFESNFNTQATNR	N$1LZHB	33	13
KFESNFNTQATNR	N$2HFLY	33	13
KFESNFNTQATNR	N$2HFMY	33	13
KFESNFNTQATNR	N$2LYM	33	13
KFESNFNTQATNR	N$2LYZ	33	13
KFESNFNTQATNR	N$2LZH	33	13
KFESNFNTQATNR	N$2LZT	33	13
KFESNFNTQATNR	N$3HFMY	33	13
KFESNFNTQATNR	N$3LYM	33	13
KFESNFNTQATNR	N$3LYZ	33	13
KFESNFNTQATNR	N$4LYZ	33	13
KFESNFNTQATNR	N$5LYZ	33	13
KFESNFNTQATNR	N$6LYZ	33	13
KFESNFNTQATNR	N$7LYZ	33	13
KFESNFNTQATNR	N$8LYZ	33	13
KFESNFNTQATNR	!LCOT	33	13
KFESNFNTHATNR	LYC$MELGA	51	13
KFESNFNTHATNR	JT0526	33	13
KFESNFNTHATNR	EZEC462	33	13
KFESNFNTHATNR	N$1LZ2	33	13
KFESNFNTHATNR	N$2LZ2	33	13
NYESSFNTQATNR	LZDK3	33	13
KFESNFNTGATNR	LZFER	51	13
KFESBFBTZATBR	LZQJE	51	13
NYESGFNTQATNR	LZDK	51	13
NYESSFNTQATNR	EZEC465	33	13
NYESGFNTQATNR	EZEC466	33	13
KFZSNFNTHATNR	LZTK	51	13
RYESNYNTQATNR	LZOVE	33	13
KWESGYNTRATNY	LZHU	33	13
KWESGYNTRATNY	HUMLSZA	51	13
KWESGYNTRATNY	HUMLYZ	34	13
KWESGYNTRATNY	N$1LZ1	33	13
KWESSYNTKATNY	LZBO	33	13
KWESSYNTKATNY	LYC$BOVIN	51	13
KWESSYNTKATNY	BOVLSZ1A	51	13
KWESSYNTKATNY	BOVLSZ3A	51	13
KWESSYNTKATNY	LYC$AXIAX	33	13
KWESSYNTKATNY	LYC$SHEEP	33	13
KWESDYNTQATNY	LZBA	33	13
KWESGYNTEATNY	LYC$PREEN	33	13
QHESNYNTRATNY	LYCP$MOUSE	51	13
KWESSYNTRATNY	LYC$RABIT	33	13
QHESNYNTRATNY	LYCM$MOUSE	51	13
KWESDFNTKAINR	LYC1$PIG	33	13

KWESDFNTKAINH	LYC2$PIG	33	13
KWESNFNTKATNY	LYC3$PIG	33	13
QHESNYNTQARNY	LZRT	33	13
EYESNFNTRAFNG	LYC$EQUAS	33	13
EYESNFNTRAFNG	LYC$HORSE	33	13
KHESGYRTTAFNN	LZPY	33	13

motif 3

PGSRNLCNIPC	LZCH	88	24
PGSRNLCNIPC	LZQJEC	70	24
PGSRNLCNIPC	LZQJEB	70	24
PGSRNLCNIPC	LZUH	70	24
PGSRNLCNIPC	EZEC470	70	24
PGSRNLCNIPC	EZEC471	70	24
PGSRNLCNIPC	S05657	70	24
PGSRNLCNIPC	N$1LYMA	70	24
PGSRNLCNIPC	N$1LYMB	70	24
PGSRNLCNIPC	N$1LYZ	70	24
PGSRNLCNIPC	N$1LZHA	70	24
PGSRNLCNIPC	N$1LZHB	70	24
PGSRNLCNIPC	N$2HFLY	70	24
PGSRNLCNIPC	N$2HFMY	70	24
PGSRNLCNIPC	N$2LYM	70	24
PGSRNLCNIPC	N$2LYZ	70	24
PGSRNLCNIPC	N$2LZH	70	24
PGSRNLCNIPC	N$2LZT	70	24
PGSRNLCNIPC	N$3HFMY	70	24
PGSRNLCNIPC	N$3LYM	70	24
PGSRNLCNIPC	N$3LYZ	70	24
PGSRNLCNIPC	N$4LYZ	70	24
PGSRNLCNIPC	N$5LYZ	70	24
PGSRNLCNIPC	N$6LYZ	70	24
PGSRNLCNIPC	N$7LYZ	70	24
PGSRNLCNIPC	N$8LYZ	70	24
PGSRNLCNIPC	!LCOT	70	24
PGSKNLCNIPC	LYC$MELGA	88	24
PGSRNLCNIPC	JT0526	70	24
PGSRNLCNIPC	EZEC462	70	24
PGSRNLCNIPC	N$1LZ2	70	24
PGSKNLCNIPC	N$2LZ2	70	24
PRAKNACGIPC	LZDK3	70	24
PGSKNLCHIPC	LZFER	88	24
PGSRBLCBIPC	LZQJE	88	24
PRSKNACGIPC	LZDK	88	24
PGSKNACGIPC	EZEC465	70	24
PGSKNACGIPC	EZEC466	70	24
PGSKBLCBIPC	LZTK	88	24
PGTKNLCHISC	LZOVE	70	24
PGAVNACHLSC	LZHU	71	25
PGAVNACHLSC	HUMLSZA	89	25
PGAVNACQLSC	HUMLYZ	72	25
PGAVNACHLSC	N$1LZ1	71	25
PNAVDGCHVSC	LZBO	71	25
PNAVDGCHVSC	LYC$BOVIN	89	25
PNAVDGCHVSC	BOVLSZ1A	89	25
PNAVDGCHVSC	BOVLSZ3A	89	25
PNAVDGCHVAC	LYC$AXIAX	71	25

PNAVDGCHVSC	LYC$SHEEP	71	25
PGAVNACHISC	LZBA	71	25
PGAVDACHISC	LYC$PREEN	71	25
PRSKNACGINC	LYCP$MOUSE	89	25
PRAVNACHIPC	LYC$RABIT	71	25
PRAVNACGINC	LYCM$MOUSE	89	25
PKAVNACHISC	LYC1$PIG	69	23
PKAVNACHISC	LYC2$PIG	69	23
PKAVNACHISC	LYC3$PIG	71	25
PRAKNACGIPC	LZRT	71	25
RSSSNACNIMC	LYC$EQUAS	70	24
RSSSNACNIMC	LYC$HORSE	70	24
RGSKNACNINC	LZPY	70	24

motif 4

SALLSSDITASVNCAK	LZCH	99	0
SALLSSDITATVNCAK	LZQJEC	81	0
SALLSSDITATVNCAK	LZQJEB	81	0
SALQSSDITATANCAK	LZUH	81	0
SALQSSDITATANCAK	EZEC470	81	0
SALLSSDITATVNCAK	EZEC471	81	0
SALLSSDITASVNCAK	S05657	81	0
SALLSSDITASVNCAK	N$1LYMA	81	0
SALLSSDITASVNCAK	N$1LYMB	81	0
SALLSSDITASVNCAK	N$1LYZ	81	0
SALLSSDITASVNCAK	N$1LZHA	81	0
SALLSSDITASVNCAK	N$1LZHB	81	0
SALLSSDITASVNCAK	N$2HFLY	81	0
SALLSSDITASVNCAK	N$2HFMY	81	0
SALLSSDITASVNCAK	N$2LYM	81	0
SALLSSDITASVNCAK	N$2LYZ	81	0
SALLSSDITASVNCAK	N$2LZH	81	0
SALLSSDITASVNCAK	N$2LZT	81	0
SALLSSDITASVNCAK	N$3HFMY	81	0
SALLSSDITASVNCAK	N$3LYM	81	0
SALLSSDITASVNCAK	N$3LYZ	81	0
SALLSSDITASVNCAK	N$4LYZ	81	0
SALLSSDITASVNCAK	N$5LYZ	81	0
SALLSSDITASVNCAK	N$6LYZ	81	0
SALLSSDITASVNCAK	N$7LYZ	81	0
SALLSSDITASVNCAK	N$8LYZ	81	0
SALLSSDITASVNCAK	!LCOT	81	0
SALLSSDITASVNCAK	LYC$MELGA	99	0
SALLSSDITASVNCAK	JT0526	81	0
SALLSSDITASVNCAK	EZEC462	81	0
SALLSSDITASVNCAK	N$1LZ2	81	0
SALLSSDITASVNCAK	N$2LZ2	81	0
SVLLRSDITEAVKCAK	LZDK3	81	0
SALLSSDITASVNCAK	LZFER	99	0
SALLSSBITASVBCAK	LZQJE	99	0
SVLLRSDITEAVRCAK	LZDK	99	0
SVLLRSDITEAVRCAK	EZEC465	81	0
SVLLRSDITEAVRCAK	EZEC466	81	0
SALLSSBITASVBCAK	LZTK	99	0
SALMGADIAPSVRCAK	LZOVE	81	0
SALLQDNIADAVACAK	LZHU	82	0
SALLQDNIADAVACAK	HUMLSZA	100	0

SALLQDNIADAVACAK	HUMLYZ	83	0
SALLQDNIADAVACAK	N$1LZ1	82	0
SELMENDIAKAVACAK	LZBO	82	0
RELMENDIAKAVACAK	LYC$BOVIN	100	0
SELMENEIAKAVACAK	BOVLSZ1A	100	0
SELMENDIAKAVACAK	BOVLSZ3A	100	0
SELMENNIDKAVTCAK	LYC$AXIAX	82	0
SELMENNIAKAVACAK	LYC$SHEEP	82	0
NALLQDNITDAVACAK	LZBA	82	0
SALLQNNIADAVACAK	LYC$PREEN	82	0
SALLQDDITAAIQCAK	LYCP$MOUSE	100	0
SDLLKDDITQAVACAK	LYC$RABIT	82	0
SALLQDDITAAIQCAK	LYCM$MOUSE	100	0
KVLLDDDLSQDIECAK	LYC1$PIG	80	0
KVLLDDDLSQDIECAK	LYC2$PIG	80	0
KVLLDDDLSQDIECAK	LYC3$PIG	82	0
SALLQDDITQAIQCAK	LZRT	82	0
SKLLDDNIDDDISCAK	LYC$EQUAS	81	0
SKLLDENIDDDISCAK	LYC$HORSE	81	0
SKLRDDNIADDIQCAK	LZPY	81	0

<u>motif 5</u>

NGMNAWVAWR	LZCH	121	6
NGMNAWVAWR	LZQJEC	103	6
BGMNAWVAWR	LZQJEB	103	6
BGMNAWVAWR	LZUH	103	6
DGMNAWVAWR	EZEC470	103	6
DGMNAWVAWR	EZEC471	103	6
DGMNAWVAWR	S05657	103	6
NGMNAWVAWR	N$1LYMA	103	6
NGMNAWVAWR	N$1LYMB	103	6
NGMNAWVAWR	N$1LYZ	103	6
NGMNAWVAWR	N$1LZHA	103	6
NGMNAWVAWR	N$1LZHB	103	6
DGMNAWVAWR	N$2HFLY	103	6
NGMNAWVAWR	N$2HFMY	103	6
NGMNAWVAWR	N$2LYM	103	6
NGMNAWVAWR	N$2LYZ	103	6
NGMNAWVAWR	N$2LZH	103	6
NGMNAWVAWR	N$2LZT	103	6
NGMNAWVAWR	N$3HFMY	103	6
NGMNAWVAWR	N$3LYM	103	6
NGMNAWVAWR	N$3LYZ	103	6
NGMNAWVAWR	N$4LYZ	103	6
NGMNAWVAWR	N$5LYZ	103	6
NGMNAWVAWR	N$6LYZ	103	6
NGMNAWVAWR	N$7LYZ	103	6
NGMNAWVAWR	N$8LYZ	103	6
HGMNAWVAWR	!LCOT	103	6
NGMNAWVAWR	LYC$MELGA	121	6
NGMNAWVAWR	JT0526	103	6
DGMNAWVAWR	EZEC462	103	6
DGMNAWVAWR	N$1LZ2	103	6
NGMNAWVAWR	N$2LZ2	103	6
DGMNAWVAWR	LZDK3	103	6
DGMNAWVAWR	LZFER	121	6
HGMNAWVAWR	LZQJE	121	6

DGMNAWVAWR	LZDK	121	6
DGMNAWVAWR	EZEC465	103	6
DGMNAWVAWR	EZEC466	103	6
BGMBAWVAWR	LZTK	121	6
DGMNAWVAWR	LZOVE	103	6
QGIRAWVAWR	LZHU	104	6
QGIRAWVAWR	HUMLSZA	122	6
QGIRAWVAWR	HUMLYZ	105	6
QGIRAWVAWR	N$1LZ1	104	6
QGITAWVAWK	LZBO	103	5
QGITAWVAWK	LYC$BOVIN	121	5
QGITAWVAWK	BOVLSZ1A	121	5
QGITAWVAWK	BOVLSZ3A	121	5
QGITAWVAWK	LYC$AXIAX	103	5
QGITAWVAWK	LYC$SHEEP	103	5
QGIRAWVAWR	LZBA	104	6
QGIRAWVAWR	LYC$PREEN	104	6
QGIRAWVAWR	LYCP$MOUSE	122	6
QGIRAWVAWR	LYC$RABIT	104	6
QGIRAWVAWR	LYCM$MOUSE	122	6
QGIKAWVAWR	LYC1$PIG	102	6
LGVKAWVAWR	LYC2$PIG	102	6
QGIKAWVAWK	LYC3$PIG	104	6
QGIRAWVAWQ	LZRT	104	6
KGMSAWKAWV	LYC$EQUAS	103	6
KGMSAWKAWV	LYC$HORSE	103	6
RGLTPWVAWK	LZPY	103	6

motif 6

NRCKGTDVQAWIRG	LZCH	131	0
NRCKGTDVHAWIRG	LZQJEC	113	0
NRCKGTDVQAWIRG	LZQJEB	113	0
KHCKGTDVRVWIKG	LZUH	113	0
KHCKGTDVRVWIKG	EZEC470	113	0
NRCKGTDVQAWIRG	EZEC471	113	0
NRCKGTDVQAWIRG	S05657	113	0
NRCKGTDVQAWIRG	N$1LYMA	113	0
NRCKGTDVQAWIRG	N$1LYMB	113	0
NRCKGTDVQAWIRG	N$1LYZ	113	0
NRCKGTDVQAWIRG	N$1LZHA	113	0
NRCKGTDVQAWIRG	N$1LZHB	113	0
NRCKGTDVQAWIRG	N$2HFLY	113	0
NRCKGTDVQAWIRG	N$2HFMY	113	0
NRCKGTDVQAWIRG	N$2LYM	113	0
NRCKGTDVQAWIRG	N$2LYZ	113	0
NRCKGTDVQAWIRG	N$2LZH	113	0
NRCKGTDVQAWIRG	N$2LZT	113	0
NRCKGTDVQAWIRG	N$3HFMY	113	0
NRCKGTDVQAWIRG	N$3LYM	113	0
NRCKGTDVQAWIRG	N$3LYZ	113	0
NRCKGTDVQAWIRG	N$4LYZ	113	0
NRCKGTDVQAWIRG	N$5LYZ	113	0
NRCKGTDVQAWIRG	N$6LYZ	113	0
NRCKGTDVQAWIRG	N$7LYZ	113	0
NRCKGTDVQAWIRG	N$8LYZ	113	0
NRCKGTDVNAWIRG	!LCOT	113	0
NRCKGTDVHAWIRG	LYC$MELGA	131	0

NRCKGTDVHAWIRG	JT0526	113	0
NRCKGTDVHAWIRG	EZEC462	113	0
NRCKGTDVHAWIRG	N$1LZ2	113	0
NRCKGTDVHAWIRG	N$2LZ2	113	0
NRCKGTDVSRWIRG	LZDK3	113	0
KHCKGTDVNVWIRG	LZFER	131	0
NRCKGTDVNAWIRG	LZQJE	131	0
NRCRGTDVSKWIRG	LZDK	131	0
NRCRGTDVSKWIRG	EZEC465	113	0
NRCRGTDVSKWIRG	EZEC466	113	0
NRCKGTBVHAWIRG	LZTK	131	0
KHCKGTDVSTWIKD	LZOVE	113	0
NRCQNRDVRQYVQG	LZHU	114	0
NRCQNRDVRQYVQG	HUMLSZA	132	0
NRCQNRDVRQYVQG	HUMLYZ	115	0
NRCQNRDVRQYVQG	N$1LZ1	114	0
SHCRDHDVSSYVEG	LZBO	113	0
SHCRDHDVSSYVEG	LYC$BOVIN	131	0
SHCRDHDVSSYVEG	BOVLSZ1A	131	0
SHCRDHDVSSYVQG	BOVLSZ3A	131	0
SHCRGHDVSSYVEG	LYC$AXIAX	113	0
SHCRDHDVSSYVEG	LYC$SHEEP	113	0
NHCQNRDVSQYVQG	LZBA	114	0
NHCQNKDVSQYVKG	LYC$PREEN	114	0
TQCQNRDLSQYIRN	LYCP$MOUSE	132	0
NHCQNQDLTPYIRG	LYC$RABIT	114	0
AHCQNRDLSQYIRN	LYCM$MOUSE	132	0
THCQNKDVSQYIRG	LYC1$PIG	112	0
AHCQNKDVSQYIRG	LYC2$PIG	112	0
AHCQNKDVSQYIRG	LYC3$PIG	114	0
RHCKNRDLSGYIRN	LZRT	114	0
KHCKDKDLSEYLAS	LYC$EQUAS	113	0
KHCKDKDLSEYLAS	LYC$HORSE	113	0
KYCQGKDLSSYVRG	LZPY	113	0

B.1.3 final super-family motifs

N$1ALC	Alpha-Lactalbumin - Baboon milk
LCA$PIG	ALPHA-LACTALBUMIN - Pig
LAGT	Alpha-lactalbumin - Goat
LCA$CAPHI	ALPHA-LACTALBUMIN PRECURSORC - Goat
LCA$SHEEP	ALPHA-LACTALBUMIN PRECURSOR - Sheep
LCA$BOVIN	ALPHA-LACTALBUMIN PRECURSOR - Bovine
LAHU	Alpha-lactalbumin precursor - Human
LABO	Alpha-lactalbumin - Bovine
LAHO	Alpha-lactalbumin - Horse
LCAB$HORSE	ALPHA-LACTALBUMIN B AND C - Horse
LCA$PAPCY	ALPHA-LACTALBUMIN - Yellow baboon
LAGP	Alpha-lactalbumin precursor - Guinea pig
LACM	Alpha-lactalbumin - Arabian camel
GPILACTAL	GPILACTAL pre-alpha-lactalbumin - *Cavia porcellus*
EZEC228	ALPHA-LACTALBUMIN - Bovine
LART2	Alpha-lactalbumin (version 2) - Rat
LART	Alpha-lactalbumin (version 2) - Rat
LARB	Alpha-lactalbumin - Rabbit
LAKGAW	Alpha-lactalbumin - Red-necked wallaby
LZCH	Lysozyme c precursor - Chicken
N$1LYMA	Lysozyme chain A - Hen egg white

N$1LYMB	Lysozyme chain B - Hen egg white
N$1LYZ	Lysozyme - Hen egg white
N$1LZHA	Lysozyme chain A - Hen egg white
N$1LZHB	Lysozyme chain B - Hen egg white
N$2HFMY	Lysozyme - Chicken
N$2LYM	Lysozyme - Hen egg white
N$2LYZ	Lysozyme - Hen egg white
N$2LZH	Lysozyme - Hen egg white
N$2LZT	Lysozyme - Hen egg white
N$3HFMY	Lysozyme - Hen egg white
N$3LYM	Lysozyme - Hen egg white
N$3LYZ	Lysozyme - Hen egg white
N$4LYZ	Lysozyme - Hen egg white
N$5LYZ	Lysozyme - Hen egg white
N$6LYZ	Lysozyme - Hen egg white
N$7LYZ	Lysozyme - Hen egg white
N$8LYZ	Lysozyme - Hen egg white
s05657	Lysozyme c - Chicken
N$2HFLY	Lysozyme c - Mouse
!LCOT	LYSOZYME - Coturnix
JT0526	Lysozyme c - Indian peafowl
EZEC462	LYSOZYME - Turkey
N$1LZ2	Lysozyme - Turkey egg white
LYC$MELGA	LYSOZYME C PRECURSOR - Turkey
N$2LZ2	Lysozyme - Turkey egg white
EZEC471	LYSOZYME - Bobwhite quail
LZQJEC	Lysozyme c - California quail
LZQJEB	Lysozyme c - Common bobwhite
LZPER	Lysozyme c precursor - Ring-necked pheasant
EZEC470	LYSOZYME - Guinea hen
LZUH	Lysozyme c - Helmeted guineafowl
EZEC465	LYSOZYME II - Kaki duck
EZEC466	LYSOZYME - Duck III
LZQJE	Lysozyme c - California quail
LZDK3	Lysozyme c III - Duck
LZDK	Lysozyme c precursor - Duck
LZTK	Lysozyme c precursor - Turkey
LZOVE	Lysozyme c - Plain chachalaca
LZBA	Lysozyme - Baboon
LZHU	Lysozyme - Human
HUMLSZA	Lysozyme precursor - *Homo sapiens*
N$1LZ1	Lysozyme - Human
LYC$RABIT	LYSOZYME C - Rabbit
HUMLYZ	HUMLYZ lysozyme - Artificial gene
LYC$PREEN	LYSOZYME C - *Hanuman langur*
LYCP$MOUSE	LYSOZYME C - Mouse
LYCM$MOUSE	LYSOZYME C - Mouse
LYC3$PIG	LYSOZYME C-3 - Pig
LYC1$PIG	LYSOZYME C-1 - Pig
LZRT	Lysozyme - Rat
BOVLSZ3A	lysozyme 3a precursor - *Bos taurus*
LZBO	Lysozyme c 2 - Bovine
LYC$AXIAX	LYSOZYME C 1 AND 2 - Axis deer
LYC$SHEEP	LYSOZYME C 1A TO 4B - Sheep
BOVLSZ1A	lysozyme 1a precursor - *Bos taurus*
LYC2$PIG	LYSOZYME C-2 - Pig
LYC$BOVIN	LYSOZYME C PRECURSOR - Bovine

LYC$EQUAS LYSOZYME C - Donkey
LYC$HORSE LYSOZYME C - Horse
LZPY Lysozyme c - Pigeon

Database version - OWL11.0

motif 1

FGRCELAAAMK	LZCH	21	21
FGRCELAAAMK	LZQJEC	3	3
FGRCELAAAMK	LZQJEB	3	3
FGRCELAAAMK	LZUH	3	3
FGRCELAAAMK	EZEC470	3	3
FGRCELAAAMK	EZEC471	3	3
FGRCELAAAMK	S05657	3	3
FGRCELAAAMK	N$1LYMA	3	3
FGRCELAAAMK	N$1LYMB	3	3
FGRCELAAAMK	N$1LYZ	3	3
FGRCELAAAMK	N$1LZHA	3	3
FGRCELAAAMK	N$1LZHB	3	3
FGRCELAAAMK	N$2HFLY	3	3
FGRCELAAAMK	N$2HFMY	3	3
FGRCELAAAMK	N$2LYM	3	3
FGRCELAAAMK	N$2LYZ	3	3
FGRCELAAAMK	N$2LZH	3	3
FGRCELAAAMK	N$2LZT	3	3
FGRCELAAAMK	N$3HFMY	3	3
FGRCELAAAMK	N$3LYM	3	3
FGRCELAAAMK	N$3LYZ	3	3
FGRCELAAAMK	N$4LYZ	3	3
FGRCELAAAMK	N$5LYZ	3	3
FGRCELAAAMK	N$6LYZ	3	3
FGRCELAAAMK	N$7LYZ	3	3
FGRCELAAAMK	N$8LYZ	3	3
YGRCELAAAMK	LZQJE	21	21
YGRCELAAAMK	LZFER	21	21
YGRCELAAAMK	LZTK	21	21
YGRCELAAAMK	LYC$MELGA	21	21
YGRCELAAAMK	JT0526	3	3
YGRCELAAAMK	EZEC462	3	3
YGRCELAAAMK	!LCOT	3	3
YGRCELAAAMK	N$1LZ2	3	3
YGRCELAAAMK	N$2LZ2	3	3
YERCELAAAMK	LZDK3	3	3
YSRCELAAAMK	LZDK	21	21
YSRCELAAAMK	EZEC465	3	3
YKRCELAAAMK	LZOVE	3	3
YQRCELAAAMK	EZEC466	3	3
FERCELARTLK	LZHU	3	3
FERCELARTLK	LZBA	3	3
FERCELARTLK	LZBO	3	3
FERCELARTLK	LYC$AXIAX	3	3
FERCELARTLK	LYC$BOVIN	21	21
FERCELARTLK	LYC$PREEN	3	3
FERCELARTLK	LYC$SHEEP	3	3
FERCELARTLK	BOVLSZ1A	21	21
FERCELARTLK	BOVLSZ3A	21	21
FERCELARTLK	HUMLSZA	21	21

FERCELARTLK	HUMLYZ	4	4
FERCELARTLK	N$1LZ1	3	3
YERCELARTLK	LYC$RABIT	3	3
YNRCELARILK	LYCP$MOUSE	21	21
FSKCELAHKLK	LYC$EQUAS	3	3
FSKCELAHKLK	LYC$HORSE	3	3
YERCEFARTLK	LZRT	3	3
YERCEFARTLK	LYCM$MOUSE	21	21
YDRCEFARILK	LYC1$PIG	3	3
YDRCEFARILK	LYC2$PIG	3	3
YDRCEFARILK	LYC3$PIG	3	3
FTKCELSQVLK	LCA$PIG	3	3
FTKCELSQLLK	LAHU	22	22
FTKCELSEVLK	LAHO	3	3
LTRCELTEKLK	LARB	3	3
FTKCELSQNLY	LCA$PAPCY	3	3
FTKCELSQNLY	N$1ALC	3	3
FTKCQLSQVLK	LCAB$HORSE	3	3
FTKCKLSDELK	LACM	3	3
LTKCEVFRELK	LABO	3	3
LTKCEVFRELK	LCA$BOVIN	22	22
LTKCEVFRELK	EZEC228	3	3
LTKCEVFQKLK	LAGT	3	3
LTKCEVFQKLK	LCA$CAPHI	22	22
LTKCEAFQKLK	LCA$SHEEP	22	22
IPRCELVKILR	LZPY	3	3
FTKCEVSHAIE	LART	22	22
FTKCEVSHAIE	LART2	3	3
YRKCQASQILK	LAKGAW	3	3
LTKCALSHELN	LAGP	22	22
LTKCALSHELN	GPILACTAL	22	22

motif 2

YSLGNWVCAA	LZCH	41	9
YSLGNWVCAA	LZQJEC	23	9
YSLGNWVCAA	LZQJEB	23	9
YSLGNWVCAA	LZUH	23	9
YSLGNWVCAA	EZEC470	23	9
YSLGNWVCAA	EZEC471	23	9
YSLGNWVCAA	S05657	23	9
YSLGNWVCAA	N$1LYMA	23	9
YSLGNWVCAA	N$1LYMB	23	9
YSLGNWVCAA	N$1LYZ	23	9
YSLGNWVCAA	N$1LZHA	23	9
YSLGNWVCAA	N$1LZHB	23	9
YSLGNWVCAA	N$2HFLY	23	9
YSLGNWVCAA	N$2HFMY	23	9
YSLGNWVCAA	N$2LYM	23	9
YSLGNWVCAA	N$2LYZ	23	9
YSLGNWVCAA	N$2LZH	23	9
YSLGNWVCAA	N$2LZT	23	9
YSLGNWVCAA	N$3HFMY	23	9
YSLGNWVCAA	N$3LYM	23	9
YSLGNWVCAA	N$3LYZ	23	9
YSLGNWVCAA	N$4LYZ	23	9
YSLGNWVCAA	N$5LYZ	23	9
YSLGNWVCAA	N$6LYZ	23	9

YSLGNWVCAA	N$7LYZ	23	9
YSLGNWVCAA	N$8LYZ	23	9
YSLGBWVCAA	LZQJE	41	9
YSLGNWVCAA	LZFER	41	9
YSLGNWVCAA	LZTK	41	9
YSLGNWVCAA	LYC$MELGA	41	9
YSLGNWVCAA	JT0526	23	9
YSLGNWVCAA	EZEC462	23	9
YSLGNWVCAA	!LCOT	23	9
YSLGNWVCAA	N$1LZ2	23	9
YSLGNWVCAA	N$2LZ2	23	9
YSLGNWVCAA	LZDK3	23	9
YSLGNWVCAA	LZDK	41	9
YSLGNWVCAA	EZEC465	23	9
YSLGNWVCAA	LZOVE	23	9
YSLGNWVCAA	EZEC466	23	9
ISLANWMCLA	LZHU	23	9
ISLANWVCLA	LZBA	23	9
VSLANWLCLT	LZBO	23	9
VSLANWLCLT	LYC$AXIAX	23	9
VSLANWLCLT	LYC$BOVIN	41	9
VSLANWVCLA	LYC$PREEN	23	9
VSLANWLCLT	LYC$SHEEP	23	9
VSLANWLCLT	BOVLSZ1A	41	9
VSLANWLCLT	BOVLSZ3A	41	9
MSLANWMCLA	HUMLSZA	41	9
ISLANWMCLA	HUMLYZ	24	9
ISLANWMCLA	N$1LZ1	23	9
VSLANWMCLA	LYC$RABIT	23	9
VKLADWVCLA	LYCP$MOUSE	41	9
YSLANWVCMA	LYC$EQUAS	23	9
YSLANWVCMA	LYC$HORSE	23	9
VSLADWVCLA	LZRT	23	9
VSLADWVCLA	LYCM$MOUSE	41	9
VSLANWVCLA	LYC1$PIG	23	9
VSLANWVCLA	LYC2$PIG	23	9
VSLANWVCLA	LYC3$PIG	23	9
ITLPEWICTI	LCA$PIG	21	7
IALPELICTM	LAHU	40	7
VTLPEWICTI	LAHO	21	7
ISMSEWICTL	LARB	21	7
IALPELICTM	LCA$PAPCY	21	7
IALPELICTM	N$1ALC	21	7
VTLPEWICTI	LCAB$HORSE	21	7
ITLAEWICII	LACM	21	7
VSLPEWVCTT	LABO	21	7
VSLPEWVCTT	LCA$BOVIN	40	7
VSLPEWVCTT	EZEC228	21	7
VSLPEWVCTA	LAGT	21	7
VSLPEWVCTA	LCA$CAPHI	40	7
VSLPEWVCTA	LCA$SHEEP	40	7
KTVANWVCLV	LZPY	23	9
ISLLEWTCVL	LART	40	7
VSLPEWTCVL	LART2	21	7
IPLPELVCTM	LAKGAW	21	7
ITLPEWLCII	LAGP	40	7
ITLPEWLCII	GPILACTAL	40	7

motif 3

STDYGILQINSRWWCND	LZCH	68	17
STDYGVLQINSRWWCND	LZQJEC	50	17
STDYGVLQINSRWWCND	LZQJEB	50	17
STDYGVLQINSRWWCND	LZUH	50	17
STDYGVLQINSRWWCND	EZEC470	50	17
STDYGVLQINSRWWCND	EZEC471	50	17
STDYGILQINSRWWCDN	S05657	50	17
STDYGILQINSRWWCND	N$1LYMA	50	17
STDYGILQINSRWWCND	N$1LYMB	50	17
STDYGILQINSRWWCND	N$1LYZ	50	17
STDYGILQINSRWWCND	N$1LZHA	50	17
STDYGILQINSRWWCND	N$1LZHB	50	17
STDYGILQINSRWWCND	N$2HFLY	50	17
STDYGILQINSRWWCND	N$2HFMY	50	17
STDYGILQINSRWWCND	N$2LYM	50	17
STDYGILQINSRWWCND	N$2LYZ	50	17
STDYGILQINSRWWCND	N$2LZH	50	17
STDYGILQINSRWWCND	N$2LZT	50	17
STDYGILQINSRWWCND	N$3HFMY	50	17
STDYGILQINSRWWCND	N$3LYM	50	17
STDYGILQINSRWWCND	N$3LYZ	50	17
STDYGILQINSRWWCND	N$4LYZ	50	17
STDYGILQINSRWWCND	N$5LYZ	50	17
STDYGILQINSRWWCND	N$6LYZ	50	17
STDYGILQINSRWWCND	N$7LYZ	50	17
STDYGILQINSRWWCND	N$8LYZ	50	17
STBYGILZIBSRWWCBB	LZQJE	68	17
STDYGILQINSRWWCND	LZFER	68	17
STBYGILZIBSRWWCBB	LZTK	68	17
STDYGILQINSRWWCND	LYC$MELGA	68	17
STDYGILQINSRWWCND	JT0526	50	17
STDYGILQINSRWWCDN	EZEC462	50	17
STDYGILQINSRWWCND	!LCOT	50	17
STDYGILQINSRWWCDN	N$1LZ2	50	17
STDYGILQINSRWWCND	N$2LZ2	50	17
STDYGILEINSRWWCDN	LZDK3	50	17
STDYGILQINSRWWCDN	LZDK	68	17
STDYGILEINSRWWCDN	EZEC465	50	17
STDYGILQINSRWWCND	LZOVE	50	17
STDYGILEINSRWWCDN	EZEC466	50	17
STDYGIFQINSRYWCND	LZHU	51	18
STDYGIFQINSHYWCND	LZBA	51	18
STDYGIFQINSKWWCND	LZBO	51	18
STDYGIFQINSKWWCDD	LYC$AXIAX	51	18
STDYGIFQINSKWWCND	LYC$BOVIN	69	18
STDYGIFQINSRYWCNN	LYC$PREEN	51	18
STDYGIFQINSKWWCND	LYC$SHEEP	51	18
STDYGIFQINSKWWCND	BOVLSZ1A	69	18
STDYGIFQINSKWWCND	BOVLSZ3A	69	18
STDYGIFQINSRYWCND	HUMLSZA	69	18
STDYGIFQINSRYWCND	HUMLYZ	52	18
STDYGIFQINSRYWCND	N$1LZ1	51	18
STDYGIFQINSRYWCND	LYC$RABIT	51	18
STDYGIFQINSRYWCND	LYCP$MOUSE	69	18
SYDYGLFQLNSKWWCKD	LYC$EQUAS	51	18
SSDYGLFQLNNKWWCKD	LYC$HORSE	51	18

STDYGIFQINSRYWCND	LZRT	51	18
STDYGIFQINSRYWCND	LYCM$MOUSE	69	18
STDYGIFQINSRYWCND	LYC1$PIG	49	16
STDYGIFQINSRYWCND	LYC2$PIG	49	16
STDYGIFQINSRYWCND	LYC3$PIG	51	18
STFYGLFQINNKLWCRD	LCA$PIG	47	16
STEYGLFQISNKLWCKS	LAHU	66	16
KTEYGLFQINNKMWCRD	LAHO	47	16
STEYGIFQINSKLWCVS	LARB	47	16
STEYGLFQISNALWCKS	LCA$PAPCY	47	16
STEYGLFQISNALWCKS	N$1ALC	47	16
KTEYGLFEINNKMWCRD	LCAB$HORSE	47	16
NREYGLFQINNKIWCRD	LACM	47	16
STEYGLFQINNKIWCKN	LABO	47	16
STEYGLFQINNKIWCKD	LCA$BOVIN	66	16
STDYGLFQINNKIWCKN	EZEC228	47	16
STEYGLFQINNKIWCKD	LAGT	47	16
STEYGLFQINNKIWCKD	LCA$CAPHI	66	16
STEYGLFQINNKIWCKD	LCA$SHEEP	66	16
SRDYGIFQINSKYWCND	LZPY	50	17
STEYGLFQISNRNWCKS	LART	66	16
STEYGLFQISNRDWCKE	LART2	47	16
NKEYGIFQISNDGWCAE	LAKGAW	47	16
HKEYGLFQINDKDFCES	LAGP	66	16
HKEYGLFQINDKDFCDS	GPILACTAL	66	16

<u>motif 4</u>

NLCNIPCSAL	LZCH	92	7
NLCNIPCSAL	LZQJEC	74	7
NLCNIPCSAL	LZQJEB	74	7
NLCNIPCSAL	LZUH	74	7
NLCNIPCSAL	EZEC470	74	7
NLCNIPCSAL	EZEC471	74	7
NLCNIPCSAL	S05657	74	7
NLCNIPCSAL	N$1LYMA	74	7
NLCNIPCSAL	N$1LYMB	74	7
NLCNIPCSAL	N$1LYZ	74	7
NLCNIPCSAL	N$1LZHA	74	7
NLCNIPCSAL	N$1LZHB	74	7
NLCNIPCSAL	N$2HFLY	74	7
NLCNIPCSAL	N$2HFMY	74	7
NLCNIPCSAL	N$2LYM	74	7
NLCNIPCSAL	N$2LYZ	74	7
NLCNIPCSAL	N$2LZH	74	7
NLCNIPCSAL	N$2LZT	74	7
NLCNIPCSAL	N$3HFMY	74	7
NLCNIPCSAL	N$3LYM	74	7
NLCNIPCSAL	N$3LYZ	74	7
NLCNIPCSAL	N$4LYZ	74	7
NLCNIPCSAL	N$5LYZ	74	7
NLCNIPCSAL	N$6LYZ	74	7
NLCNIPCSAL	N$7LYZ	74	7
NLCNIPCSAL	N$8LYZ	74	7
BLCBIPCSAL	LZQJE	92	7
NLCHIPCSAL	LZFER	92	7
BLCBIPCSAL	LZTK	92	7
NLCNIPCSAL	LYC$MELGA	92	7

NLCNIPCSAL	JT0526	74	7
NLCNIPCSAL	EZEC462	74	7
NLCNIPCSAL	!LCOT	74	7
NLCNIPCSAL	N$1LZ2	74	7
NLCNIPCSAL	N$2LZ2	74	7
NACGIPCSVL	LZDK3	74	7
NACGIPCSVL	LZDK	92	7
NACGIPCSVL	EZEC465	74	7
NLCHISCSAL	LZOVE	74	7
NACGIPCSVL	EZEC466	74	7
NACHLSCSAL	LZHU	75	7
NACHISCNAL	LZBA	75	7
DGCHVSCSEL	LZBO	75	7
DGCHVACSEL	LYC$AXIAX	75	7
DGCHVSCREL	LYC$BOVIN	93	7
DACHISCSAL	LYC$PREEN	75	7
DGCHVSCSEL	LYC$SHEEP	75	7
DGCHVSCSEL	BOVLSZ1A	93	7
DGCHVSCSEL	BOVLSZ3A	93	7
NACHLSCSAL	HUMLSZA	93	7
NACQLSCSAL	HUMLYZ	76	7
NACHLSCSAL	N$1LZ1	75	7
NACHIPCSDL	LYC$RABIT	75	7
NACGINCSAL	LYCP$MOUSE	93	7
NACNIMCSKL	LYC$EQUAS	74	6
NACNIMCSKL	LYC$HORSE	74	6
NACGIPCSAL	LZRT	75	7
NACGINCSAL	LYCM$MOUSE	93	7
NACHISCKVL	LYC1$PIG	73	7
NACHISCKVL	LYC2$PIG	73	7
NACHISCKVL	LYC3$PIG	75	7
NICGISCDKF	LCA$PIG	70	6
NICDISCDKF	LAHU	90	7
NICGISCDKF	LAHO	71	7
NICDTPCENF	LARB	71	7
NICDITCDKF	LCA$PAPCY	71	7
NICDITCDKF	N$1ALC	71	7
NICGISCNKF	LCAB$HORSE	71	7
NICDISCDKF	LACM	71	7
NICNISCDKF	LABO	71	7
NICNISCDKF	LCA$BOVIN	90	7
NICNISCDKF	EZEC228	71	7
NICNISCDKF	LAGT	71	7
NICNISCDKF	LCA$CAPHI	90	7
NICNISCDKF	LCA$SHEEP	90	7
NACNINCSKL	LZPY	74	7
NICDISCDKF	LART	90	7
NICDISCDKF	LART2	71	7
SVCGILCSKF	LAKGAW	71	7
NICDISCDKL	LAGP	90	7
NICDISCDKL	GPILACTAL	90	7

Motif 5

LSSDITASVNCAKKIV	LZCH	102	0
LSSDITATVNCAKKIV	LZQJEC	84	0
LSSDITATVNCAKKIV	LZQJEB	84	0
QSSDITATANCAKKIV	LZUH	84	0

QSSDITATANCAKKIV	EZEC470	84	0
LSSDITATVNCAKKIV	EZEC471	84	0
LSSDITASVNCAKKIV	S05657	84	0
LSSDITASVNCAKKIV	N$1LYMA	84	0
LSSDITASVNCAKKIV	N$1LYMB	84	0
LSSDITASVNCAKKIV	N$1LYZ	84	0
LSSDITASVNCAKKIV	N$1LZHA	84	0
LSSDITASVNCAKKIV	N$1LZHB	84	0
LSSDITASVNCAKKIV	N$2HFLY	84	0
LSSDITASVNCAKKIV	N$2HFMY	84	0
LSSDITASVNCAKKIV	N$2LYM	84	0
LSSDITASVNCAKKIV	N$2LYZ	84	0
LSSDITASVNCAKKIV	N$2LZH	84	0
LSSDITASVNCAKKIV	N$2LZT	84	0
LSSDITASVNCAKKIV	N$3HFMY	84	0
LSSDITASVNCAKKIV	N$3LYM	84	0
LSSDITASVNCAKKIV	N$3LYZ	84	0
LSSDITASVNCAKKIV	N$4LYZ	84	0
LSSDITASVNCAKKIV	N$5LYZ	84	0
LSSDITASVNCAKKIV	N$6LYZ	84	0
LSSDITASVNCAKKIV	N$7LYZ	84	0
LSSDITASVNCAKKIV	N$8LYZ	84	0
LSSBITASVBCAKKIV	LZQJE	102	0
LSSDITASVNCAKKIV	LZFER	102	0
LSSBITASVBCAKKIA	LZTK	102	0
LSSDITASVNCAKKIA	LYC$MELGA	102	0
LSSDITASVNCAKKIV	JT0526	84	0
LSSDITASVNCAKKIA	EZEC462	84	0
LSSDITASVNCAKKIV	!LCOT	84	0
LSSDITASVNCAKKIA	N$1LZ2	84	0
LSSDITASVNCAKKIA	N$2LZ2	84	0
LRSDITEAVKCAKRIV	LZDK3	84	0
LRSDITEAVRCAKRIV	LZDK	102	0
LRSDITEAVRCAKRIV	EZEC465	84	0
MGADIAPSVRCAKRIV	LZOVE	84	0
LRSDITEAVRCAKRIV	EZEC466	84	0
LQDNIADAVACAKRVV	LZHU	85	0
LQDNITDAVACAKRVV	LZBA	85	0
MENDIAKAVACAKKIV	LZBO	85	0
MENNIDKAVTCAKQIV	LYC$AXIAX	85	0
MENDIAKAVACAKHIV	LYC$BOVIN	103	0
LQNNIADAVACAKRVV	LYC$PREEN	85	0
MENNIAKAVACAKHIV	LYC$SHEEP	85	0
MENEIAKAVACAKQIV	BOVLSZ1A	103	0
MENDIAKAVACAKHIV	BOVLSZ3A	103	0
LQDNIADAVACAKRVV	HUMLSZA	103	0
LQDNIADAVACAKRVV	HUMLYZ	86	0
LQDNIADAVACAKRVV	N$1LZ1	85	0
LKDDITQAVACAKRVV	LYC$RABIT	85	0
LQDDITAAIQCAKRVV	LYCP$MOUSE	103	0
LDDNIDDDISCAKRVV	LYC$EQUAS	84	0
LDENIDDDISCAKRVV	LYC$HORSE	84	0
LQDDITQAIQCAKRVV	LZRT	85	0
LQDDITAAIQCAKRVV	LYCM$MOUSE	103	0
LDDDLSQDIECAKRVV	LYC1$PIG	83	0
LDDDLSQDIECAKRVV	LYC2$PIG	83	0
LDDDLSQDIECAKRVV	LYC3$PIG	85	0

LDDDLTDDDMCAKKIL	LCA$PIG	80	0
LDDDITDDIMCAKKIL	LAHU	100	0
LDDDLTDDVMCAKKIL	LAHO	81	0
LDDNLTDDVKCAMKIL	LARB	81	0
LDDDITDDIMCAKKIL	LCA$PAPCY	81	0
LDDDITDDIMCAKKIL	N$1ALC	81	0
LDDDLTDDVMCAKKDL	LCAB$HORSE	81	0
LDDDLTDDKMCAKKIL	LACM	81	0
LDDDLTDDIMCVKKIL	LABO	81	0
LDDDLTDDIMCVKKIL	LCA$BOVIN	100	0
LNNDLTNNIMCVKKIL	EZEC228	81	0
LDDDLTDDIVCAKKIL	LAGT	81	0
LDDDLTDDIVCAKKIL	LCA$CAPHI	100	0
LDDDLTDDIVCAKKIL	LCA$SHEEP	100	0
RDDNIADDIQCAKKIA	LZPY	84	0
LDDELADDIVCAKKIV	LART	100	0
LDDELADDIVCAKKIV	LART2	81	0
LDDDITDDIECAKKIL	LAKGAW	81	0
LDDDLTDDIMCVKKIL	LAGP	100	0
LDDDLTDDIMCVKKIL	GPILACTAL	100	0

Motif 6

GMNAWVAWRNRC	LZCH	122	4
GMNAWVAWRNRC	LZQJEC	104	4
GMNAWVAWRNRC	LZQJEB	104	4
GMNAWVAWRKHC	LZUH	104	4
GMNAWVAWRKHC	EZEC470	104	4
GMNAWVAWRNRC	EZEC471	104	4
GMNAWVAWRNRC	S05657	104	4
GMNAWVAWRNRC	N$1LYMA	104	4
GMNAWVAWRNRC	N$1LYMB	104	4
GMNAWVAWRNRC	N$1LYZ	104	4
GMNAWVAWRNRC	N$1LZHA	104	4
GMNAWVAWRNRC	N$1LZHB	104	4
GMNAWVAWRNRC	N$2HFLY	104	4
GMNAWVAWRNRC	N$2HFMY	104	4
GMNAWVAWRNRC	N$2LYM	104	4
GMNAWVAWRNRC	N$2LYZ	104	4
GMNAWVAWRNRC	N$2LZH	104	4
GMNAWVAWRNRC	N$2LZT	104	4
GMNAWVAWRNRC	N$3HFMY	104	4
GMNAWVAWRNRC	N$3LYM	104	4
GMNAWVAWRNRC	N$3LYZ	104	4
GMNAWVAWRNRC	N$4LYZ	104	4
GMNAWVAWRNRC	N$5LYZ	104	4
GMNAWVAWRNRC	N$6LYZ	104	4
GMNAWVAWRNRC	N$7LYZ	104	4
GMNAWVAWRNRC	N$8LYZ	104	4
GMNAWVAWRNRC	LZQJE	122	4
GMNAWVAWRKHC	LZFER	122	4
GMBAWVAWRNRC	LZTK	122	4
GMNAWVAWRNRC	LYC$MELGA	122	4
GMNAWVAWRNRC	JT0526	104	4
GMNAWVAWRNRC	EZEC462	104	4
GMNAWVAWRNRC	!LCOT	104	4
GMNAWVAWRNRC	N$1LZ2	104	4
GMNAWVAWRNRC	N$2LZ2	104	4

GMNAWVAWRNRC	LZDK3	104	4
GMNAWVAWRNRC	LZDK	122	4
GMNAWVAWRNRC	EZEC465	104	4
GMNAWVAWRKHC	LZOVE	104	4
GMNAWVAWRNRC	EZEC466	104	4
GIRAWVAWRNRC	LZHU	105	4
GIRAWVAWRNHC	LZBA	105	4
GITAWVAWKSHC	LZBO	104	3
GITAWVAWKSHC	LYC$AXIAX	104	3
GITAWVAWKSHC	LYC$BOVIN	122	3
GIRAWVAWRNHC	LYC$PREEN	105	4
GITAWVAWKSHC	LYC$SHEEP	104	3
GITAWVAWKSHC	BOVLSZ1A	122	3
GITAWVAWKSHC	BOVLSZ3A	122	3
GIRAWVAWRNRC	HUMLSZA	123	4
GIRAWVAWRNRC	HUMLYZ	106	4
GIRAWVAWRNRC	N$1LZ1	105	4
GIRAWVAWRNHC	LYC$RABIT	105	4
GIRAWVAWRTQC	LYCP$MOUSE	123	4
GMSAWKAWVKHC	LYC$EQUAS	104	4
GMSAWKAWVKHC	LYC$HORSE	104	4
GIRAWVAWQRHC	LZRT	105	4
GIRAWVAWRAHC	LYCM$MOUSE	123	4
GIKAWVAWRTHC	LYC1$PIG	103	4
GVKAWVAWRAHC	LYC2$PIG	103	4
GIKAWVAWKAHC	LYC3$PIG	105	4
GIDYWLAHKALC	LCA$PIG	99	3
GIDYWLAHKALC	LAHU	119	3
GIDYWLAHKPLC	LAHO	100	3
GIDHWLAHKPLC	LARB	100	3
GIDYWIAHKALC	LCA$PAPCY	100	3
GIDYWIAHKALC	N$1ALC	100	3
GIDYWLAHKPLC	LCAB$HORSE	100	3
GIDYWLAHKPLC	LACM	100	3
GINYWLAHKALC	LABO	100	3
GINYWLAHKALC	LCA$BOVIN	119	3
GINYWLAHKALC	EZEC228	100	3
GINYWLAHKALC	LAGT	100	3
GINYWLAHKALC	LCA$CAPHI	119	3
GINYWLAHKALC	LCA$SHEEP	119	3
GLTPWVAWKKYC	LZPY	104	4
GIDYWKAHKPMC	LART	119	3
GINYWLAHKPMC	LART2	100	3
GLGYWKAHETFC	LAKGAW	101	4
GIDYWLAHKPLC	LAGP	119	3
GIDYWFAHKPLC	GPILACTAL	119	3

B.2.1 Sugar transporter final motifs

GTR1_BOVIN	GLUCOSE TRANSPORTER PROTEIN I - Bovine
GTR1_HUMAN	GLUCOSE TRANSPORTER PROTEIN I - *Homo sapiens*
S09705	Glucose transport protein - Mouse
GTR4_HUMAN	GLUCOSE TRANSPORTER IV - *Homo sapiens*
GTR1_RABIT	GLUCOSE TRANSPORTER PROTEIN I - Rabbit
GTR1_RAT	GLUCOSE TRANSPORTER PROTEIN I - Rat
GTR4_RAT	GLUCOSE TRANSPORTER IV - Rat
GTR1_MOUSE	GLUCOSE TRANSPORTER PROTEIN 1 - Mouse
A30310	Glucose transport protein GT1 - Mouse
GTR4_MOUSE	GLUCOSE TRANSPORTER IV - Mouse
GTR3_CHICK	GLUCOSE TRANSPORTER III - Chicken
A41751	Glucose-transport protein 3 - Mouse
GTR3_HUMAN	GLUCOSE TRANSPORTER-LIKE PROTEIN - Human
GTR2_HUMAN	GLUCOSE TRANSPORTER PROTEIN, LIVER - Human
GTR2_RAT	GLUCOSE TRANSPORTER PROTEIN, LIVER - Rat
S05319	Glucose transport protein, hepatic - Mouse
GTR2_MOUSE	GLUCOSE TRANSPORTER PROTEIN, LIVER - Mouse
RATGLTP	RATGLTP LOCUS RATGLTP - *Rattus norvegicus*
STP1_ARATH	GLUCOSE TRANSPORTER (SUGAR CARRIER) - Mouse-ear cress
TOBMST1	TOBMST1 LOCUS TOBMST1 - *Nicotiana tabacum*
SNF3_YEAST	HIGH-AFFINITY GLUCOSE TRANSPORTER SNF3 - Baker's yeast
HUP1_CHLKE	H(+)/HEXOSE COTRANSPORTER - *Chlorella kessleri*
CHLHUP1G	CHLHUP1G LOCUS CHLHUP1G - *Chlorella kessleri*
A40538	Myo-inositol transporter IRT1 - Yeast
B40538	Myo-inositol transporter IRT2 - Yeast
YSCHXT4A	YSCHXT4A LOCUS YSCHXT4A - *Saccharomyces cerevisia*
XYLE_ECOLI	XYLOSE-PROTON SYMPORT - *Escherichia coli*
HXT2_YEAST	HIGH-AFFINITY GLUCOSE TRANSPORTER HXT2 - Yeast
RAG1_KLULA	LOW-AFFINITY GLUCOSE TRANSPORTER - *Kluyveromyces lactis*
A39728	Hexose transport protein HXT1 - Yeast
GLCP_SYNY3	GLUCOSE TRANSPORT PROTEIN - *Synechocystis sp.*
GAL2_YEAST	GALACTOSE TRANSPORTER - Yeast
JQ0383	Galactose permease - Yeast
ATHSTP4	ATHSTP4 LOCUS ATHSTP4 - *Arabidopsis thaliana*
GTR5_HUMAN	GLUCOSE TRANSPORTER, SMALL INTESTINE - Human
GLF_ZYMMO	GLUCOSE FACILITATED DIFFUSION - *Zymomonas mobilis*
LEID1TRA	LEID1TRA LOCUS LEID1TRA - *Leishmania donovani*
QAY_NEUCR	QUINATE TRANSPORTER - *Neurospora crassa*
S108238	putative hexose transporter - *Trypanosoma brucei*
PRO1_LEIEN	PROBABLE TRANSPORT PROTEIN - *Leishmania enriettii*
ARAE_ECOLI	ARABINOSE-PROTON SYMPORT - *Escherichia coli*
QUTD_ASPNI	QUINATE PERMEASE - *Aspergillus nidulans*
LEID2TRA	LEID2TRA LOCUS LEID2TRA - *Leishmania donovani*
CIT1_ECOLI	CITRATE-PROTON SYMPORT - *E. coli*
CIT2_ECOLI	CITRATE-PROTON SYMPORT - *E. coli*
CITA_SALTY	CITRATE-PROTON SYMPORT - *Salmonella typhimurium*
CIT_KLEPN	CITRATE-PROTON SYMPORT - *Klebsiella pneumoniae*
MAL6_YEAST	MALTOSE PERMEASE - Baker's yeast
LACP_KLULA	LACTOSE PERMEASE - *Kluyveromyces lactis* (yeast)
GTR1_PIG	GLUCOSE TRANSPORTER PROTEIN I (FRAGMENT) - Pig

Database version - OWL19.0

Motif 1

GFLFGYDTGVI	LEID1TRA	13	13
SFQFGYDIGVI	GTR2_HUMAN	21	21
SFQFGYDIGVI	GTR2_MOUSE	21	21
SFQFGYDIGVI	GTR2_RAT	21	21
SFQFGYDIGVI	S05319	21	21
SFQFGYDIGVI	RATGLTP	21	21
SFQFGYNTGVI	GTR3_HUMAN	21	21
SFQFGYNTGVI	A41751	21	21
SLQFGYNTGVI	GTR1_BOVIN	23	23
SLQFGYNTGVI	GTR1_HUMAN	23	23
SLQFGYNTGVI	GTR1_MOUSE	23	23
SLQFGYNTGVI	GTR1_RABIT	23	23
SLQFGYNTGVI	GTR1_RAT	23	23
SLQFGYNTGVI	GTR3_CHICK	22	22
SLQFGYNTGVI	S09705	23	23
SLQFGYNTGVI	A30310	23	23
SLQFGYNIGVI	GTR4_HUMAN	35	35
SLQFGYNIGVI	GTR4_MOUSE	37	37
SLQFGYNIGVI	GTR4_RAT	35	35
GFLFGYDTGLI	SNF3_YEAST	108	108
GFMFGYDTGYI	A40538	97	97
GFMFGYDTGYI	B40538	123	123
GLLFGYDTAVI	XYLE_ECOLI	21	21
GLLFGLDIGVI	ARAE_ECOLI	33	33
GLLFGYDSAVI	GLF_ZYMMO	21	21
GFLFGFDTAVI	GLCP_SYNY3	28	28
GFIFGWDTGTI	A39728	75	75
GFVFGWDTGTI	HXT2_YEAST	67	67
GFVFGWDTGTI	RAG1_KLULA	74	74
GFVFGWDTGTI	YSCHXT4A	82	82
GLIFGYDIGIS	STP1_ARATH	34	34
GLIFGYDIGIS	TOBMST1	34	34
GFMFGWDTSTI	GAL2_YEAST	82	82
GFMFGWDTSTI	JQ0383	82	82
GLIFGYDLGIS	ATHSTP4	34	34
GLLLGYDNGVT	HUP1_CHLKE	38	38
GLLLGYDNGVT	CHLHUP1G	38	38
SCMIGYDSAFI	QAY_NEUCR	32	32
SCMIGYDSAFI	QUTD_ASPNI	32	32
FFLFGFYATYI	CIT1_ECOLI	28	28
FFLFGFYATYI	CIT2_ECOLI	28	28
FFLFGFYATYI	CITA_SALTY	31	31
FFLFGFYATYI	CIT_KLEPN	44	44
GTLNGYVIGYV	S108238	47	47
PLLYGYNLGFV	LEID2TRA	49	49
GSLNGYSIGFV	PRO1_LEIEN	55	55
SFQYGYNVAAV	GTR5_HUMAN	29	29
LIQEGYDTAIL	MAL6_YEAST	110	110
ATMQGYDGALM	LACP_KLULA	83	83

Motif 2

LVSRVIVGLAIGISSATIPV	LEID1TRA	98	74
IAGRSISGLYCGLISGLVPM	GTR2_HUMAN	155	123
IAGRSVSGLYCGLISGLVPM	GTR2_MOUSE	154	122
IAGRSVSGLYCGLISGLVPM	GTR2_RAT	153	121
IAGRSVSGLYCGLISGLVPM	S05319	154	122

IAGRSVSGLYCGLISGLVPM	RATGLTP	153	121
ILGRLVIGLFCGLCTGFVPM	GTR3_HUMAN	121	89
ILGRLLIGIFCGLCTGFVPM	A41751	121	89
ILGRFIIGVYCGLTTGFVPM	GTR1_BOVIN	123	89
ILGRFIIGVYCGLTTGFVPM	GTR1_HUMAN	123	89
ILGRFIIGVYCGLTTGFVPM	GTR1_MOUSE	123	89
ILGRFIIGVYCGLTTGFVPM	GTR1_RABIT	123	89
ILGRFIIGVYCGLTTGFVPM	GTR1_RAT	123	89
IIGRFIIGLFCGLCTGFVPM	GTR3_CHICK	122	89
ILGRFIIGVYCGLTTGFVPM	S09705	123	89
ILGRFIIGVYCGLTTGFVPM	A30310	123	89
ILGRFLIGAYSGLTSGLVPM	GTR4_HUMAN	139	93
ILGRFLIGAYSGLTSGLVPM	GTR4_MOUSE	141	93
ILGRFLIGAYSGLTSGLVPM	GTR4_RAT	139	93
IVGRVISGIGIGAISAVVPL	SNF3_YEAST	199	80
AVGRLIMGFGVGIGSLIAPL	A40538	183	75
AAGRLIMGFGVGIGSLISPL	B40538	209	75
VIYRIIGGIGVGLASMLSPM	XYLE_ECOLI	130	98
IAARVVLGIAVGIASYTAPL	ARAE_ECOLI	116	72
CFFRFLAGLGIGVVSTLTPT	GLF_ZYMMO	120	88
IFWRVLGGIGVGAASVIAPA	GLCP_SYNY3	111	72
FIGRIISGLGVGGITVLSPM	A39728	170	84
FIGRIISGMGVGGIAVLSPT	HXT2_YEAST	162	84
FIGRIISGLGVGGITVLSPM	RAG1_KLULA	169	84
FIGRIISGLGVGGIAVLSPM	YSCHXT4A	177	84
IVGRILLGFGIGFANQAVPL	STP1_ARATH	137	92
IVGRILLGFGIGFANQSVPL	TOBMST1	137	92
FIGRIISGLGVGGIAVLCPM	GAL2_YEAST	177	84
FIGRIISGLGVGGIAVLCPM	JQ0383	177	84
LIGRILLGFGVGFANQSVPV	ATHSTP4	136	91
IVGRVLLGFGVGLGSQVVPQ	HUP1_CHLKE	140	91
IVGRVLLGFGVGLGSQVVPQ	CHLHUP1G	141	92
IAGRVLAGIGVGGASNMVPI	QAY_NEUCR	128	85
YGGRVLAGIGVGAGSNICPI	QUTD_ASPNI	124	81
LVGRLLQGFSAGVELGGVSV	CIT1_ECOLI	118	79
LVGRLLQGFSAGVELGGVSV	CIT2_ECOLI	118	79
LLGRLLQGFSAGVELGGVSV	CITA_SALTY	121	79
LIGRLLQGFSAGAELGGVSV	CIT_KLEPN	134	79
CTGRVLIGLGVGILCSVCPM	S108238	179	121
FVARIVLGFPLGWQSITSSH	LEID2TRA	209	149
IVGRFVIGLFLGVICVACPV	PRO1_LEIEN	217	151
IISRLLVGICAGVSSNVVPM	GTR5_HUMAN	129	89
AVGQALCGMPWGCFQCLTVS	MAL6_YEAST	205	84
IGGRWFVAFFATIANAAAPT	LACP_KLULA	170	76

Motif 3

IVLNNLFLTGGQFVAAGFTA	LEID1TRA	98	74
GTFHQLAIVTGILISQIIGL	GTR2_HUMAN	189	14
GTLHQLALVTGILISQIAGL	GTR2_MOUSE	188	14
GTLHQLALVTGILISQIAGL	GTR2_RAT	187	14
GTLHQLALVTGILISQIAGL	S05319	188	14
GTLLQLGITVGIIISQILGL	RATGLTP	187	14
GTLNQLGIVVGILVAQIFGL	GTR3_HUMAN 1	55	14
GTLNQLGIVVGILVAQIFGL	A41751	155	14
GTLHQLGIVVGILIAQVFGL	GTR1_BOVIN	157	14
GTLHQLGIVVGILIAQVFGL	GTR1_HUMAN	157	14
GTLHQLGIVVGILIAQVFGL	GTR1_MOUSE	157	14

GTLHQLGIVVGILIAQVFGL	GTR1_RABIT	157	14
GTLHQLGIVVGILIAQVFGL	GTR1_RAT	157	14
GTLNQLGIVVGILVAQIFGL	GTR3_CHICK	156	14
GTLHQLGIVVGILIAQVFGL	S09705	157	14
GTLHQLGIVVGILIAQVFGL	A30310	157	14
GTLNQLAIVIGILIAQVLGL	GTR4_HUMAN	173	14
GTLNRLAIVIGILVAQVLGL	GTR4_MOUSE	175	14
GTLNQLAIVIGILVAQVLGL	GTR4_RAT	173	14
ISTYQWAITWGLLVSSAVSQ	SNF3_YEAST	233	14
TVINSLWLTGGQLVAYGCGA	A40538	217	14
TVINSLWLTGGQLIAYGCGA	B40538	243	14
VSFNQFAIIFGQLLVYCVNY	XYLE_ECOLI	164	14
ISMYQLMVTLGIVLAFLSDT	ARAE_ECOLI	150	14
VSGQQMAIVTGALTGYIFTW	GLF_ZYMMO	154	14
GSLQQLAIVSGIFIALLSNW	GLCP_SYNY3	145	14
VSCYQVMITLGIFLGYCTNF	A39728	204	14
VSFYQLMITLGIFLGYCTNY	HXT2_YEAST	196	14
VSCYQLMITFGIFLGYCTNY	RAG1_KLULA	203	14
VSCYQLMITLGIFLGYCTNY	YSCHXT4A	211	14
NIGFQLSITIGILVAEVLNY	STP1_ARATH	171	14
NLGFQLSITIGILVANVLNY	TOBMST1	171	14
VSCYQLMITAGIFLGYCTNY	GAL2_YEAST	211	14
VSCYQLMITAGIFLGYCTNY	JQ0383	211	14
NNGFQVAIIFGIVVATIINY	ATHSTP4	170	14
NIGYQLFVTIGILIAGLVNY	HUP1_CHLKE	174	14
NIGYQLFVTIGILIAGLVNY	CHLHUP1G	175	14
VGIYELGWQIGGLVGFWINY	QAY_NEUCR	162	14
VGVYELGWQIGGVVGFWINY	QUTD_ASPNI	158	14
SASQQVAIVVAALIGYGLNV	CIT1_ECOLI	156	18
SASQQVAIVVAALIGYGLNV	CIT2_ECOLI	156	18
SASQQVAIVVAALIGYSLNI	CITA_SALTY	159	18
SGSQQVAIMVAAAMGFALNA	CIT_KLEPN	172	18
GVLFQVFTTLGIMLAAMLGL	S108238	213	14
GTLFQVSVSTGIFVTSFFGL	LEID2TRA	243	14
GVMFQVFTTLGIFVAALMGL	PRO1_LEIEN	251	14
GVVPQLFITVGILVAQIFGL	GTR5_HUMAN	163	14
TTYSNLCWTFGQLFAAGIMK	MAL6_YEAST	239	14
AGLYNTLWSVGSIVAAFSTY	LACP_KLULA	204	14

Motif 4

GLFLALLAVFLALYAPGIGCIPWV	LEID1TRA	340	188
YVSMIAIFLFVSFFEIGPGPIPWF	GTR2_HUMAN	398	189
YVSMTAIFLFVSFFEIGPGPIPWF	GTR2_MOUSE	397	189
YVSMTAIFLFVSFFEIGPGPIPWF	GTR2_RAT	396	189
YVSMTAIFLFVSFFEIGPGPIPWF	S05319	397	189
YVSMTAIFLFVSFFEIGPIPIPFF	RATGLTP	396	189
FVCIGAILVFVAFFEIGPGPIPWF	GTR3_HUMAN	364	189
FVCIVAILIYVAFFEIGPGPIPWF	A41751	364	189
YLSIVAIFGFVAFFEVGPGPIPWF	GTR1_BOVIN	366	189
YLSIVAIFGFVAFFEVGPGPIPWF	GTR1_HUMAN	366	189
YLSIVAIFGFVAFFEVGPGPIPWF	GTR1_MOUSE	366	189
YLSIVAIFGFVAFFEVGPGPIPWF	GTR1_RABIT	366	189
YLSIVAIFGFVAFFEVGPGPIPWF	GTR1_RAT	366	189
YISIVATFGFVALFEIGPGPIPWF	GTR3_CHICK	363	187
YLSIVAIFGFVAFFEVGPGPIPWF	S09705	366	189
YLSIVAIFGFVAFFEVGPGPIPWF	A30310	366	189
YVSIVAIFGFVAFFEIGPGPIPWF	GTR4_HUMAN	382	189

YVSIVAIFGFVAFFEIGPGPIPWF	GTR4_MOUSE	384	189
YVSIVAIFGFVAFFEIGPGPIPWF	GTR4_RAT	382	189
KVMIAFICLFIAAFSATWGGVVWV	SNF3_YEAST	449	196
IVIIVFIIVFAAFYALGIGTVPWQ	A40538	445	208
IVIIVFIIVYAAFYALGIGTVPWQ	B40538	471	208
IVALLSMLFYVAAFAMSWGPVCWV	XYLE_ECOLI	370	186
WLSVGMTMMCIAGYAMSAAPVVWI	ARAE_ECOLI	359	189
VLPLASVLLYIAVFGMSWGPVCWV	GLF_ZYMMO	361	187
IIALVTANLYVFSFGFSWGPIVWV	GLCP_SYNY3	370	205
NCMIVFACFYIFCFATTWAPIAYV	A39728	426	202
NVMIVFTCLFIFFFAISWAPIAYV	HXT2_YEAST	418	202
NCMIVFACFYIFCFATTWAPIAYV	RAG1_KLULA	427	204
NCMIVFTCFYLFCFATTWAPIPFV	YSCHXT4A	433	202
IVVVTFICIYVAGFAWSWGPLGWL	STP1_ARATH	388	197
IVVVIFICVYVAGFAWSWGPLGWL	TOBMST1	386	195
NCMIVFTCFYIFCYATTWAPVAWV	GAL2_YEAST	433	202
NCMIVFTCFYIFCYATTWAPVAWV	JQ0383	433	202
NLIVALICIYVAGFAWSWGPLGWL	ATHSTP4	386	196
SGILAVICIFISGFAWSWGPMGWL	HUP1_CHLKE	390	196
SGILAVICIFISGFAWSWGPMGWL	CHLHUP1G	391	196
IAAIFFFYLWTAFYTPSWNGTPWV	QAY_NEUCR	393	211
IAAIFFFYLWTAFYTPSWNGTPWV	QUTD_ASPNI	389	211
FTRMTLVLLWFSFFFGMYNGAMVA	CIT1_ECOLI	328	152
FTRMTLVLLWFSFFFGMYNGAMVA	CIT2_ECOLI	328	152
FTRMTLVLLWFSFFFGMYNGAMVA	CITA_SALTY	331	152
FLMMLSVLLWLSFIYGMYNGAMIP	CIT_KLEPN	344	152
GVATTGIALFIAAFEFGVGSCFFV	S108238	399	166
GIAITGIAIFIALYEMGVGPCFYV	LEID2TRA	436	173
GVAITGILLFILGFEVCVGPCYYV	PRO1_LEIEN	438	167
YISIVCVISYVIGHALGPSPIPAL	GTR5_HUMAN	374	191
MGSGALLMVVAFFYNLGIAPVVFC	MAL6_YEAST	456	197
NGALVFIYLFGGIFSFAFTPMQSM	LACP_KLULA	428	204

Motif 5

GEIFPTHLRTSAA	LEID1TRA	366	2
AEFFSQGPRPAAL	GTR2_HUMAN	424	2
AEFFSQGPRSTAL	GTR2_MOUSE	423	2
AEFFSQGPRPTAL	GTR2_RAT	422	2
AEFFSQGPRPTAL	S05319	423	2
REWFTQIWRPGAI	RATGLTP	422	2
AELFSQGPRPAAM	GTR3_HUMAN	390	2
AELFSQGPRPAAI	A41751	390	2
AELFSQGPRPAAI	GTR1_BOVIN	392	2
AELFSQGPRPAAI	GTR1_HUMAN	392	2
AELFSQGPRPARI	GTR1_MOUSE	392	2
AELFSQGPRPAAV	GTR1_RABIT	392	2
AELFSQGPRPAAV	GTR1_RAT	392	2
AELFSQGPRPAAM	GTR3_CHICK	389	2
AELFSQGPRPAAI	S09705	392	2
AELFSQGPRPARI	A30310	392	2
AELFSQGPRPAAM	GTR4_HUMAN	408	2
AELFSQGPRPAAM	GTR4_MOUSE	409	1
AELFSQGPRPAAM	GTR4_RAT	408	2
AELYPLGVRSKCT	SNF3_YEAST	475	2
SELFPQNVRGIGT	A40538	470	1
SELFPQNVRGVGT	B40538	496	1
SEIFPNAIRGKAL	XYLE_ECOLI	396	2

SEIQPLKCRDFGI	ARAE_ECOLI	385	2
SEMFPSSIKGAAM	GLF_ZYMMO	387	2
GEMFNNKIRAAAL	GLCP_SYNY3	396	2
SECFPLRVKSKCM	A39728	452	2
AESYPLRVKNRAM	HXT2_YEAST	444	2
SESYPLRVKGKAM	RAG1_KLULA	453	2
SETFPLRVKSKCM	YSCHXT4A	459	2
SEIFPLEIRSAAQ	STP1_ARATH	414	2
SEIFPLEIRSAAQ	TOBMST1	412	2
AESFPLRVKSKCM	GAL2_YEAST	459	2
AESFPLRVKSKCM	JQ0383	459	2
SEISPLEIRSAAQ	ATHSTP4	412	2
SEIFTLETRPAGT	HUP1_CHLKE	416	2
SEIFTLETRPAGT	CHLHUP1G	417	2
SEMFDQNTRSLGQ	QAY_NEUCR	419	2
SEMFDPTVRSLAQ	QUTD_ASPNI	415	2
TEVMPVYVRTVGF	CIT1_ECOLI	354	2
TEVMPVYVRTVGF	CIT2_ECOLI	354	2
TEVMPVYVRTVGF	CITA_SALTY	357	2
TEIMPAEVRVAGF	CIT_KLEPN	370	2
QDLFPPSFRPKGG	S108238	425	2
VDVFPESFRPIGS	LEID2TRA	462	2
QDMFPPSFRPRGA	PRO1_LEIEN	464	2
TEIFLQSSRPSAF	GTR5_HUMAN	400	2
SEMPSSRLRTKTI	MAL6_YEAST	482	2
TEVSTNLTRSKAQ	LACP_KLULA	454	2

B.2.2 Super-family motifs

GTR4_HUMAN	GLUCOSE TRANSPORTER, INSULIN-RESPONSIVE - Human
GTR1_BOVIN	GLUCOSE TRANSPORTER PROTEIN I - Bovine
GTR1_HUMAN	GLUCOSE TRANSPORTER PROTEIN - Human
GTR1_MOUSE	GLUCOSE TRANSPORTER PROTEIN - Mouse
GTR1_PIG	GLUCOSE TRANSPORTER PROTEIN (FRAGMENT). - Pig
GTR1_RABIT	GLUCOSE TRANSPORTER PROTEIN - Rabbit
GTR1_RAT	GLUCOSE TRANSPORTER PROTEIN - Rat
S09705	Glucose transport protein - Mouse
A30310	Glucose transport protein GT1 - Mouse
GTR4_MOUSE	GLUCOSE TRANSPORTER - Mouse
GTR4_RAT	GLUCOSE TRANSPORTER - Rat
GTR3_CHICK	GLUCOSE TRANSPORTER TYPE 3 - Chicken
A41751	Glucose-transport protein 3 - Mouse
HUP1_CHLKE	H(+)/HEXOSE COTRANSPORTER. - *Chlorella kessleri*
CHLHUP1G	CHLHUP1G LOCUS CHLHUP1G - *Chlorella kessleri*
GTR3_HUMAN	GLUCOSE TRANSPORTER-LIKE PROTEIN - Human
SNF3_YEAST	HIGH-AFFINITY GLUCOSE TRANSPORTER SNF3 - Baker's yeast
GTR2_HUMAN	GLUCOSE TRANSPORTER PROTEIN, LIVER - Human
GTR2_MOUSE	GLUCOSE TRANSPORTER PROTEIN, LIVER. - Mouse
GTR2_RAT	GLUCOSE TRANSPORTER PROTEIN, LIVER. - Rat
S05319	Glucose transport protein, hepatic - Mouse
RATGLTP	RATGLTP LOCUS RATGLTP - *Rattus norvegicus*
STP1_ARATH	GLUCOSE TRANSPORTER - Mouse-ear cress
TOBMST1	TOBMST1 LOCUS TOBMST1 - *Nicotiana tabacum*
ATHSTP4	ATHSTP4 LOCUS ATHSTP4 - *Arabidopsis thaliana*
S22742	Methylenomycin A resistance protein - *Bacillus subtilis*
QAY_NEUCR	QUINATE TRANSPORTER - *Neurospora crassa*
TCR1_BACSU	TETRACYCLINE RESISTANCE PROTEIN - *Bacillus subtilis*
PRO1_LEIEN	PROBABLE TRANSPORT PROTEIN (LTP) - *Leishmania enriettii*

TCR_BACST	TETRACYCLINE RESISTANCE - *Bacillus stearothermophilus*
TCR_STRAG	TETRACYCLINE RESISTANCE - *Streptococcus agalactiae*
TCR_STRPN	TETRACYCLINE RESISTANCE - *Streptococcus pneumoniae*
RAG1_KLULA	LOW-AFFINITY GLUCOSE TRANSPORTER - *Kluyveromyces lactis*
A39728	Hexose transport protein HXT1 - Yeast
YSCHXT4A	YSCHXT4A LOCUS YSCHXT4A - *Saccharomyces cerevisia*
GAL2_YEAST	GALACTOSE TRANSPORTER - *Saccharomyces cerevisiae*
JQ0383	Galactose permease - Yeast
HXT2_YEAST	HIGH-AFFINITY GLUCOSE TRANSPORTER HXT2 - Yeast
ARAE_ECOLI	ARABINOSE-PROTON SYMPORT - *Escherichia coli*
TCR2_BACSU	TETRACYCLINE RESISTANCE PROTEIN - *Bacillus subtilis*
TCR_STAAU	TETRACYCLINE RESISTANCE PROTEIN - *Staphylococcus aureus*
QQSABT	Hypothetical protein B-295 - *Staphylococcus aureus*
CIT_KLEPN	CITRATE-PROTON SYMPORT - *Klebsiella pneumoniae*
CITA_SALTY	CITRATE-PROTON SYMPORT - *Salmononella thyphimurium*
QUTD_ASPNI	QUINATE PERMEASE - *Aspergillus nidulans*
S108238	putative hexose transporter - *Trypanosoma brucei*
CIT1_ECOLI	CITRATE-PROTON SYMPORT - *E. coli*
CIT2_ECOLI	CITRATE-PROTON SYMPORT - *E. coli*
GTR5_HUMAN	GLUCOSE TRANSPORTER, SMALL INTESTINE - Human
MMR_STRCO	METHYLENOMYCIN A RESISTANCE - *Streptomyces coelicolor*
GLCP_SYNY3	GLUCOSE TRANSPORT PROTEIN. - *Synechocystis sp.*
TCR1_ECOLI	TETRACYCLINE RESISTANCE PROTEIN - *Escherichia coli*
ECOTN10	ECOTN10 coding sequence - *Escherichia coli*
XYLE_ECOLI	XYLOSE-PROTON SYMPORT - *Escherichia coli*
STMBAHBRP	STMBAHBRP ORF3 - *Streptomyces hygroscopicus*
S19863	Lincomycin resistance - *Streptomyces lincolnensis*
RATCGAT	RATCGAT LOCUS RATCGAT - *Rattus norvegicus*
TCR3_ECOLI	TETRACYCLINE RESISTANCE PROTEIN - *Escherichia coli*
JQ1479	Tetracycline resistance protein - *Escherichia coli*
TCR2_ECOLI	TETRACYCLINE RESISTANCE PROTEIN - *Escherichia coli*
ACCPCAOP3	putative transport protein *Acinetobacter calcoaceticus*
GLF_ZYMMO	GLUCOSE FACILITATED DIFFUSION - *Zymomonas mobilis*
RATSVAT	RATSVAT LOCUS RATSVAT - *Rattus norvegicus*
B40046	Tetracycline resistance - *Streptomyces coelicolor*
A39705	Multidrug resistance protein - *Bacillus subtilis*
QACA_STAAU	ANTISEPTIC RESISTANCE PROTEIN - *Staphylococcus aureus*
ATR1_YEAST	AMINOTRIAZOLE RESISTANCE PROTEIN - Yeast
LEID2TRA	LEID2TRA LOCUS LEID2TRA - *Leishmania donovani*
S18539	actVA-1 protein - *Streptomyces coelicolor*
S108506	resistance to cycloheximide - *Candida maltosa*
YSACYHR	YSACYHR LOCUS YSACYHR - *Candida maltosa*
BMR_CANAL	BENOMYL/METHOTREXATE RESISTANCE - *Candida albicans*
M225633S1	export pump-tetracenomycin C - *Streptomyces glaucescens*
S21395	Chloramphenicol resistance - *Rhodococcus fascians*
S18593	Chloramphenicol resistance - *Streptomyces lividans*
JQ1201	CmlA protein - *Pseudomonas sp.*

Database version - OWL19.0

<u>Motif 1</u>

MLILGRFLIGAYSGLTSGLVP	GTR4_HUMAN	137	137
MLILGRFIIGVYCGLTTGFVP	GTR1_BOVIN	121	121
MLILGRFIIGVYCGLTTGFVP	GTR1_HUMAN	121	121
MLILGRFIIGVYCGLTTGFVP	GTR1_MOUSE	121	121
MLILGRFIIGVYCGLTTGFVP	GTR1_PIG	80	80
MLILGRFIIGVYCGLTTGFVP	GTR1_RABIT	121	121

MLILGRFIIGVYCGLTTGFVP	GTR1_RAT	121	121
MLILGRFIIGVYCGLTTGFVP	S09705	121	121
MLILGRFIIGVYCGLTTGFVP	A30310	121	121
ILILGRFLIGAYSGLTSGLVP	GTR4_MOUSE	139	139
ILILGRFLIGAYSGLTSGLVP	GTR4_RAT	137	137
MLIIGRFIIGLFCGLCTGFVP	GTR3_CHICK	120	120
MLILGRLLIGIFCGLCTGFVP	A41751	119	119
MLIVGRVLLGFGVGLGSQVVP	HUP1_CHLKE	138	138
MLIVGRVLLGFGVGLGSQVVP	CHLHUP1G	139	139
MLILGRLVIGLFCGLCTGFVP	GTR3_HUMAN	119	119
LLIVGRVISGIGIGAISAVVP	SNF3_YEAST	197	197
LIIAGRSISGLYCGLISGLVP	GTR2_HUMAN	153	153
LIIAGRSVSGLYCGLISGLVP	GTR2_MOUSE	152	152
LIIAGRSVSGLYCGLISGLVP	GTR2_RAT	151	151
LIIAGRSVSGLYCGLISGLVP	S05319	152	152
LIIAGRSVSGLYCGLISGLVP	RATGLTP	151	151
MLIVGRILLGFGIGFANQAVP	STP1_ARATH	135	135
MLIVGRILLGFGIGFANQSVP	TOBMST1	135	135
MLLIGRILLGFGVGFANQSVP	ATHSTP4	134	134
MLIAGRLIQGIGAALFMPSSL	S22742	107	107
PIIAGRVLAGIGVGGASNMVP	QAY_NEUCR	126	126
ILILARFIQGIGAAAFPALVM	TCR1_BACSU	105	105
VLIVGRFVIGLFLGVICVACP	PRO1_LEIEN	215	215
LLIMARFIQGAGAAAFPALVM	TCR_BACST	105	105
LLIMARFIQGAGAAAFPALVM	TCR_STRAG	105	105
LLIMARFIQGAGAAAFPALVM	TCR_STRPN	105	105
QYFIGRIISGLGVGGITVLSP	RAG1_KLULA	167	167
QYFIGRIISGLGVGGITVLSP	A39728	168	168
QYFIGRIISGLGVGGIAVLSP	YSCHXT4A	175	175
QYFIGRIISGLGVGGIAVLCP	GAL2_YEAST	175	175
QYFIGRIISGLGVGGIAVLCP	JQ0383	175	175
QYFIGRIISGMGVGGIAVLSP	HXT2_YEAST	160	160
MLIAARVVLGIAVGIASYTAP	ARAE_ECOLI	114	114
ILIFGRLVQGVGSAAFPSLIM	TCR2_BACSU	105	105
ILIFGRLVQGVGSAAFPSLIM	TCR_STAAU	105	105
ILIFGRLVQGVGSAAFPSLIM	QQSABT	105	105
LVLIGRLLQGFSAGAELGGVS	CIT_KLEPN	132	132
LVLLGRLLQGFSAGVELGGVS	CITA_SALTY	119	119
LIYGGRVLAGIGVGAGSNICP	QUTD_ASPNI	122	122
ALCTGRVLIGLGVGILCSVCP	S108238	177	177
LVLVGRLLQGFSAGVELGGVS	CIT1_ECOLI	116	116
LVLVGRLLQGFSAGVELGGVS	CIT2_ECOLI	116	116
LIIISRLLVGICAGVSSNVVP	GTR5_HUMAN	127	127
TLIAARLVQGAGAALFMPSSL	MMR_STRCO	116	116
DFIFWRVLGGIGVGAASVIAP	GLCP_SYNY3	109	109
MLYLGRLLSGITGATGAVAAS	TCR1_ECOLI	96	96
MLYLGRLLSGITGATGAVAAS	ECOTN10	96	96
EFVIYRIIGGIGVGLASMLSP	XYLE_ECOLI	128	128
VLIAARLVQGFSLGGEYGAAT	STMBAHBRP	124	124
LLVLARFGQGAGEALSLPAAM	S19863	121	121
LLFVARTLQGIGSSFSSVAGL	RATCGAT	189	189
VLYIGRIVAGITGATGAVAGA	TCR3_ECOLI	98	98
VLYIGRIVAGITGATGAVAGA	JQ1479	98	98
ILYAGRIVAGITGATGAVAGA	TCR2_ECOLI	98	98
SLVIFRFLTGIGLGAAMPNAT	ACCPCAOP3	126	126
IFCFFRFLAGLGIGVVSTLTP	GLF_ZYMMO	118	118
FLLIARSLQGIGSSCSSVAGM	RATSVAT	185	185

MLTAARFLQGGLGALMIPQGL	B40046	132	132
MLFISRMLGGISAPFIMPGVT	A39705	96	96
FVIAIRFLLGIAGALIMPTTL	QACA_STAAU	109	109
FFIISRAFQGLGIAFVLPNVL	ATR1_YEAST	163	163
VLFVARIVLGFPLGWQSITSS	LEID2TRA	207	207
QLIAARACMGVSGAAVLPSTL	S18539	114	114
GLSVLRVIAGFFAAPALSTGG	S108506	193	193
GLSVLRVIAGFFAAPALSTGG	YSACYHR	193	193
GLCILRFLGGFFASPCLATGG	BMR_CANAL	209	209
AIVVFRVLQGLFGALMQPSAL	M225633S1	116	116
VLLVTRIVGALANAGFLAVAL	S21395	93	93
VLVACRVVAALANAGFLAVAL	S18593	93	93
VFLGLRILQACGASACLVSTF	JQ1201	104	104

Motif 2

TLNQLAIVIGILIAQVLGLESL	GTR4_HUMAN	174	16
TLHQLGIVVGILIAQVFGLDSI	GTR1_BOVIN	158	16
TLHQLGIVVGILIAQVFGLDSI	GTR1_HUMAN	158	16
TLHQLGIVVGILIAQVFGLDSI	GTR1_MOUSE	158	16
TLHQLGIVVGILIAQVFGLDSI	GTR1_PIG	117	16
TLHQLGIVVGILIAQVFGLDSI	GTR1_RABIT	158	16
TLHQLGIVVGILIAQVFGLDSI	GTR1_RAT	158	16
TLHQLGIVVGILIAQVFGLDSI	S09705	158	16
TLHQLGIVVGILIAQVFGLDSI	A30310	158	16
TLNRLAIVIGILVAQVLGLESM	GTR4_MOUSE	176	16
TLNQLAIVIGILVAQVLGLESM	GTR4_RAT	174	16
TLNQLGIVVGILVAQIFGLEGI	GTR3_CHICK	157	16
TLNQLGIVVGILVAQIFGLDFI	A41751	156	16
IGYQLFVTIGILIAGLVNYAVR	HUP1_CHLKE	175	16
IGYQLFVTIGILIAGLVNYAVR	CHLHUP1G	176	16
TLNQLGIVVGILVAQIFGLEFI	GTR3_HUMAN	156	16
STYQWAITWGLLVSSAVSQGTH	SNF3_YEAST	234	16
TFHQLAIVTGILISQIIGLEFI	GTR2_HUMAN	190	16
TLHQLALVTGILISQIAGLSFI	GTR2_MOUSE	189	16
TLHQLALVTGILISQIAGLSFI	GTR2_RAT	188	16
TLHQLALVTGILISQIAGLSFI	S05319	189	16
TLLQLGITVGIIISQILGLDNS	RATGLTP	188	16
IGFQLSITIGILVAEVLNYFFA	STP1_ARATH	172	16
LGFQLSITIGILVANVLNYFFA	TOBMST1	172	16
NGFQVAIIFGIVVATIINYFTA	ATHSTP4	171	16
ALVSAASALGPFIGGVLVQLAG	S22742	149	21
GIYELGWQIGGLVGFWINYGVN	QAY_NEUCR	163	16
SLVAMGEGVGPAIGGMVAHYIH	TCR1_BACSU	146	20
VMFQVFTTLGIFVAALMGLALG	PRO1_LEIEN	252	16
SIVAMGEGVGPAIGGMIAHYIH	TCR_BACST	146	20
SIVAMGEGVGPAIVGMIAHYIH	TCR_STRAG	146	20
SIVAMGEGVGPAIGGMIAHYIH	TCR_STRPN	146	20
SCYQLMITFGIFLGYCTNYGTK	RAG1_KLULA	204	16
SCYQVMITLGIFLGYCTNFGTK	A39728	205	16
SCYQLMITLGIFLGYCTNYGTK	YSCHXT4A	212	16
SCYQLMITAGIFLGYCTNYGTK	GAL2_YEAST	212	16
SCYQLMITAGIFLGYCTNYGTK	JQ0383	212	16
SFYQLMITLGIFLGYCTNYGTK	HXT2_YEAST	197	16
SMYQLMVTLGIVLAFLSDTAFS	ARAE_ECOLI	151	16
SIVALGEGLGPSIGGIIAHYIH	TCR2_BACSU	146	20
SIVALGEGLGPSIGGIIAHYIH	TCR_STAAU	146	20
SIVALGEGLGPSIGGIIAHYIH	QQSABT	146	20

GSQQVAIMVAAAMGFALNAVLE	CIT_KLEPN	173	20
ASQQVAIVVAALIGYSLNITLG	CITA_SALTY	160	20
GVYELGWQIGGVVGFWINYGVD	QUTD_ASPNI	159	16
VLFQVFTTLGIMLAAMLGLILD	S108238	214	16
ASQQVAIVVAALIGYGLNVTLG	CIT1_ECOLI	157	20
ASQQVAIVVAALIGYGLNVTLG	CIT2_ECOLI	157	20
VVPQLFITVGILVAQIFGLRNL	GTR5_HUMAN	164	16
AIVATSSGLGPTVGGLMVSAFG	MMR_STRCO	158	21
SLQQLAIVSGIFIALLSNWFIA	GLCP_SYNY3	146	16
ASFGLGLIAGPIIGGFAGEISP	TCR1_ECOLI	136	19
ASFGLGLIAGPIIGGFAGEISP	ECOTN10	136	19
SFNQFAIIFGQLLVYCVNYFIA	XYLE_ECOLI	165	16
SFQYVASSVGHILAGLSTLAAS	STMBAHBRP	161	16
SVASVGLVLGFLLSGVITQLFS	S19863	161	19
GGLALGLLVGAPFGSVMYEFVG	RATCGAT	231	21
ACFGFGMVAGPVLGGLMGGFSP	TCR3_ECOLI	138	19
ACFGFGMVAGPVLGGLMGGFSP	JQ1479	138	19
ACFGVGMVAGPVAGGLLGAISL	TCR2_ECOLI	138	19
CGYNLGMAIGGFISSWLIPAFG	ACCPCAOP3	167	20
SGQQMAIVTGALTGYIFTWLLA	GLF_ZYMMO	155	16
GGLAMGVLVGPPFGSVLYEFVG	RATSVAT	227	21
PAIGLGAVLGPIVAGFLVDADL	B40046	173	20
GYMSAAISTGFIIGPGIGGFLA	A39705	133	16
IASSIGAVFGPIIGGALLEQFS	QACA_STAAU	151	21
AMAPIGATLGCLFAGLIGTEDP	ATR1_YEAST	206	22
TLFQVSVSTGIFVTSFFGLVLG	LEID2TRA	244	16
ASVGFALGIGPVTGGILLAHFW	S18539	155	20
GVWSIFAVAGPSIGPLIGAAVI	S108506	230	16
GVWSIFAVAGPSIGPLIGAAVI	YSACYHR	230	16
AAWSLGAVCGPSFGPFFGSILT	BMR_CANAL	246	16
GVVGASTAAGPIIGGLLVQHVG	M225633S1	157	20
GGVTIACVVGVPGGALLGELWG	S21395	134	20
SGTTVATVAGVPGGSLLGTWLG	S18593	134	20
SMLAMVPAVGPLLGALVDMWLG	JQ1201	146	21

Appendix C

Example entries from the PRINTS database

C.1 ANNEXIN

COMPOUND(7)
D. N. PERKINS 1/5/1991
ANNEXINS

1. BARTON, G.J., NEWMAN, R.H., FREEMONT, P.S., CRUMPTON, M.J. Amino acid sequence analysis of the annexin super-gene family of proteins.
EUROPEAN JOURNAL OF BIOCHEMISTRY 198 749-760 (1991)

2. GEISOW, M.J. Annexins-forms without function but not without fun.
TIBTECH 9 180-181 (1991)

The annexins are a family of proteins that have the abillity to bind both membranes and phospholipids. These functions are both calcium dependant [1]. The role of the annexins has not yet been determined precisely, although they have been shown to be associated with regulating the membrane cytoskeleton, inhibition of phospholipase C and also to act as anti-coagulants [2]. There are eleven distinct types of annexin, each type has a primary sequence consisting of four or eight repeats of a conserved 61 residue segment. The ability to bind calcium and phospholipids is thought to reside in these repeat regions while it has been suggested that the N terminal domain is responsible for the functional specificity of each protein.

Twelve sequences were initially aligned and from this seven motifs were selected. Motifs one and two describe the first repeat while motifs three,four and five describe the first half of three further repeats. Two iterations were required using OWL version 11.0 at which point a true set of twenty eight sequences was shown to match with all the motifs.

```
SUMMARY INFORMATION
-------------------
  28 codes involving  7 elements
   0 codes involving  6 elements
   0 codes involving  5 elements
   0 codes involving  4 elements
   0 codes involving  3 elements
   0 codes involving  2 elements
```

```
COMPOUND FEATURE INDEX
----------------------

 7|  28   28   28   28   28   28   28
 6|   0    0    0    0    0    0    0
 5|   0    0    0    0    0    0    0
 4|   0    0    0    0    0    0    0
 3|   0    0    0    0    0    0    0
 2|   0    0    0    0    0    0    0
--+-----------------------------------
  |   1    2    3    4    5    6    7

True positives:
 LUHU36          LUBO36          LUMS36          ANX2$CHICK
 A35600          LUHU            !LPCH           ANX1$CAVCU
 ANX4$BOVIN      HUMP68          ANX4$PIG        ANX6$HUMAN
 ANX1$RAT        ANX6$MOUSE      S01786          ANX3$RAT
 A29250          HUMCBPE         ANX4$HUMAN      ANX1$MOUSE
 ANX5$CHICK      ANX5$HUMAN      ANX5$RAT        ANX3$HUMAN
 ANX8$HUMAN      HUMSNEXIN       ANX1$COLLI      DROANNX

 LUHU36           Calpactin I heavy chain - Human
 LUBO36           Calpactin I heavy chain - Bovine
 LUMS36           Calpactin I heavy chain - Mouse
 ANX2$CHICK      ANNEXIN II (LIPOCORTIN II) - Chicken
 A35600           Calpactin I heavy chain - Mouse
 LUHU             Calpactin I heavy chain - Human
 !LPCH            LIPOCORTIN - Human
 ANX1$CAVCU       ANNEXIN I (LIPOCORTIN I) - Guinea Pig
 ANX4$BOVIN       ANNEXIN IV (LIPOCORTIN IV) (ENDONEXIN I) - Bovine
 HUMP68           HUMP68 p68 - Homo sapiens
 ANX4$PIG         ANNEXIN IV (LIPOCORTIN IV) (ENDONEXIN I) - Pig
 ANX6$HUMAN       ANNEXIN VI (LIPOCORTIN VI) - Human
 ANX1$RAT         ANNEXIN I (LIPOCORTIN I) (CALPACTIN II) - Rat
 ANX6$MOUSE      ANNEXIN VI (LIPOCORTIN VI)(PROTEIN III) - Mouse
 S01786           Calcium-binding protein p68 - Mouse
 ANX3$RAT         ANNEXIN III (LIPOCORTIN III) - Rat
 A29250           Lipocortin III - Rat
 HUMCBPE          HUMCBPE calelectrin - Homo sapiens
 ANX4$HUMAN       ANNEXIN IV (LIPOCORTIN IV) (ENDONEXIN I) - Human
 ANX1$MOUSE       ANNEXIN I (LIPOCORTIN I) (CALPACTIN II) - Mouse
 ANX5$CHICK       ANNEXIN V (LIPOCORTIN V) (ENDONEXIN II) - Chicken
 ANX5$HUMAN       ANNEXIN V (LIPOCORTIN V) (ENDONEXIN II) - Human
 ANX5$RAT         ANNEXIN V (LIPOCORTIN V) (ENDONEXIN II) - Rat
 ANX3$HUMAN       ANNEXIN III (LIPOCORTIN III) - Human
 ANX8$HUMAN       ANNEXIN VIII (VASCULAR ANTICOAGULANT) - Human
 HUMSNEXIN        HUMSNEXIN synexin - Homo sapiens
 ANX1$COLLI       ANNEXIN I (LIPOCORTIN I) - Pigeon
 DROANNX          DROANNX annexin X - Drosophila melanogaster

 SCAN HISTORY
 ------------
 OWL11_0    2  100 NSINGLE
```

```
INITIAL MOTIF-SETS
------------------
23
motif 1
KTKGVDEVTIVNILTNRSNAQRQ              LUHU36     46     46
KTKGVDEVTIINILTNRSNEQRQ         ANX2$CHICK     46     46
TVKGVDEATIIDILTKRNNAQRQ         ANX1$CAVCU     56     56
MVKGVDEATIIDILTKRNNAQRQ               LUHU     55     55
KGIGTDEATIIDIVTHRSNAQRQ         ANX6$MOUSE    376    376
KGMGTDEETILKILTSRNNAQRQ         ANX5$CHICK     29     29
KGIGTNEQAIIDVLTKRSNTQRQ         ANX8$HUMAN     35     35
KGFGTDEQEIIDVLVGRSNQQRQ            DROANNX     29     29
RGIGTDEKMLISILTERSNAQRQ         ANX3$HUMAN     32     32
KGLGTDEDAIINVLAYRSTAQRQ         ANX4$BOVIN     27     27
KGFGTDEQAIVDVVANRSNDQRQ          HUMSNEXIN    177    177
TAKGVDEATIIDIMTTRTNAQRP         ANX1$COLLI     51     51

ANNEXIN2
17
motif 2
LKSALSGHLETVILGLL                    LUHU36     86     17
LKSALSGHLEAVILGLL               ANX2$CHICK     86     17
LKKALTGHLEEVVLALL               ANX1$CAVCU     96     17
LKKALTGHLEEVVLALL                     LUHU     95     17
LKSEISGDLARLILGLM               ANX6$MOUSE    416     17
LKSELTGKFETLMVSLM               ANX5$CHICK     69     17
LKSELSGKFERLIVALM               ANX8$HUMAN     75     17
LKDELGGKFEDVIVGLM                  DROANNX     69     17
LKGDLSGHFEHLMVALV               ANX3$HUMAN     72     17
LKSELSGNFEQVILGMM               ANX4$BOVIN     67     17
LKSELSGNMEELILALF                HUMSNEXIN    217     17
MKRVLKSHLEDVVVALL               ANX1$COLLI     91     17

ANNEXIN3
22
motif 3
LKASMKGLGTDEDSLIEIICSR               LUHU36    113     10
LKAAMKGLGTDEDTLIEIICSR          ANX2$CHICK    113     10
LRAAMKGLGTDEDTLIEILVSR          ANX1$CAVCU    123     10
LRAAMKGLGTDEDTLIEILASR                LUHU    122     10
LKKAMEGAGTDEKTLIEILATR          ANX6$MOUSE    443     10
LKHAIKGAGTNEKVLTEILASR          ANX5$CHICK     96     10
LHDAMKGLGTKEGVIIEILASR          ANX8$HUMAN    102     10
LHAAMAGIGTEEATLVEILCTK             DROANNX     96     10
LKKSMKGAGTNEDALIEILTTR          ANX3$HUMAN     99     10
LRKAMKGAGTDEGCLIEILASR          ANX4$BOVIN     94     10
LRKAMQGAGTQERVLIEILCTR           HUMSNEXIN    244     10
LRACMKGHGTDEDTLIEILASR          ANX1$COLLI    118     10

ANNEXIN4
27
motif 4
LYDAGVKRKGTDVPKWISIMTERSVPH          LUHU36    197     62
LYDAGVKRKGTDVPKWINIMTERSVPH     ANX2$CHICK    197     62
LYEAGERRKGTDVNVFITILTTRSYSH     ANX1$CAVCU    206     61
LYEAGERRKGTDVNVFNTILTTRSYPQ           LUHU    205     61
IADTPSGDKTSLETRFMTVLCTRSYPH     ANX6$MOUSE    531     66
```

LFRAGELKWGTDEETFITILGTRSVSH	ANX5$CHICK	179	61
LYAAGEKIRGTDEMKFITILCTRSATH	ANX8$HUMAN	186	62
LYSAGEAKLGTDEEVFNRIMSHASFPQ	DROANNX	180	62
LYKAGENRWGTDEDKFTEILCLRSFPQ	ANX3$HUMAN	182	61
LYEAGEKKWGTDEVKFLTVLCSRNRNH	ANX4$BOVIN	177	61
LYQAGEGRLGTDESCFNMILATRSFPQ	HUMSNEXIN	327	61
LYEAGEQKKGTDINVFVTVLTARSYPH	ANX1$COLLI	201	61

ANNEXIN5
27
motif 5

MKGKGTRDKVLIRIMVSRSEVDMLKIR	LUHU36	277	53
MKGKGTRDKVLIRIMVSRCEVDMLKIK	ANX2$CHICK	277	53
MKGAGTRHKALIRIMVSRSEIDMNDIK	ANX1$CAVCU	286	53
MKGVGTRHKALIRIMVSRSEIDMNDIK	LUHU	285	53
MKGAGTDEKTLTRVMVSRSEIDLLNIR	ANX6$MOUSE	611	53
MKGAGTDDDTLIRVMVSRSEIDLLDIR	ANX5$CHICK	259	53
MKGAGTRDGTLIRNIVSRSEIDLNLIK	ANX8$HUMAN	266	53
MNGAGTDDATLIRIIVSRSEIDLETIK	DROANNX	260	53
LKGIGTDEFTLNRIMVSRSEIDLLDIR	ANX3$HUMAN	262	53
MKGLGTDDDTLIRVMVSRAEIDMLDIR	ANX4$BOVIN	257	53
MKGAGTDDSTLVRIVVTRSEIDLVQIK	HUMSNEXIN	407	53
MKGFGTQHRDLIRIMVSRHEVDMNEIK	ANX1$COLLI	280	52

ANNEXIN6
16
motif 6

EFKRKYGKSLYYYIQQ	LUHU36	305	1
EFKRKYGKSLYYFIQQ	ANX2$CHICK	305	1
YYQKMYGISLCQAILD	ANX1$CAVCU	314	1
FYQKMYGISLCQAILD	LUHU	313	1
EFIEKYDKSLHQAIEG	ANX6$MOUSE	639	1
EFRKNFAKSLYQMIQK	ANX5$CHICK	287	1
HFKKMYGKTLSSMIME	ANX8$HUMAN	294	1
EFERIYNRTLHSAVVD	DROANNX	288	1
EFKKHYGYSLYSAIKS	ANX3$HUMAN	290	1
NFKRLYGKSLYSFIKG	ANX4$BOVIN	285	1
MFAQMYQKTLGTMIAG	HUMSNEXIN	435	1
YYKKMYGISLCQAIMD	ANX1$COLLI	308	1

ANNEXIN7
15
motif 7

DTKGDYQKALLYLCG	LUHU36	321	0
DTKGDYQRALLNLCG	ANX2$CHICK	321	0
ETKGDYEKILVALCG	ANX1$CAVCU	330	0
ETKGDYEKILVALCG	LUHU	329	0
DTSGDFMKALLALCG	ANX6$MOUSE	655	0
DTSGDYRKALLLLCG	ANX5$CHICK	303	0
DTSGDYKNALLSLVG	ANX8$HUMAN	310	0
ETSGDYKRALTALLG	DROANNX	305	1
DTSGDYEITLLKICG	ANX3$HUMAN	306	0
DTSGDYRKVLLILCG	ANX4$BOVIN	301	0
DTSGDYRRLLLAIVG	HUMSNEXIN	451	0
ELKGGYETILVALCG	ANX1$COLLI	324	0

```
FINAL MOTIF-SETS
----------------
ANNEXIN1
23
motif 1
KGIGTDEATIIDIVTHRSNAQRQ          ANX6$MOUSE    376   376
KGIGTDEATIIDIVTHRSNAQRQ              S01786    377   377
KGLGTDEDTIIDIITHRSNVQRQ          ANX6$HUMAN    376   376
KGLGTDEDTIIDIITHRSNVQRQ             HUMCBPE    377   377
KGLGTDEDTIIDIITHRSNVQRQ              HUMP68    377   377
KGIGTDEKTLINILTERSNAQRQ            ANX3$RAT     33    33
KGIGTDEKTLINILTERSNAQRQ              A29250     33    33
KTKGVDEVTIVNILTNRSNAQRQ              LUHU36     46    46
KTKGVDEVTIINILTNRSNEQRQ          ANX2$CHICK     46    46
MVKGVDEATIIDILTKRNNAQRQ                LUHU     55    55
MVKGVDEATIIDILTKRNNAQRQ               !LPCH     56    56
TVKGVDEATIIDILTKRNNAQRQ          ANX1$CAVCU     56    56
MVKGVDEATIIDILTKRTNAQRQ          ANX1$MOUSE     55    55
MVKGVDEATIIDILTKRTNAQRQ            ANX1$RAT     55    55
KTKGVDEVTIVNILTNRSNVQRQ              LUMS36     46    46
KTKGVDEVTIVNILTNRSNEQRQ              LUBO36     46    46
KGMGTDEETILKILTSRNNAQRQ          ANX5$CHICK     29    29
KGLGTDEDSILNLLTARSNAQRQ            ANX5$RAT     27    27
KTKGVDEVTIVNILTNRSMVQRQ              A35600     46    46
KGLGTDEESILTLLTSRSNAQRQ          ANX5$HUMAN     28    28
KGLGTDEDAIINVLAYRSTAQRQ          ANX4$BOVIN     27    27
KGIGTNEQAIIDVLTKRSNTQRQ          ANX8$HUMAN     35    35
RGIGTDEKMLISILTERSNAQRQ          ANX3$HUMAN     32    32
KGLGTDEDAIISVLAYRSTAQRQ            ANX4$PIG     27    27
KGFGTDEQEIIDVLVGRSNQQRQ             DROANNX     29    29
KGLGTDEDAIISVLAYRNTAQRQ          ANX4$HUMAN     27    27
TAKGVDEATIIDIMTTRTNAQRP          ANX1$COLLI     51    51
KGFGTDEQAIVDVVANRSNDQRQ           HUMSNEXIN    177   177

ANNEXIN2
17
motif 2
LKSEISGDLARLILGLM                ANX6$MOUSE    416    17
LKSEISGDLARLILGLM                    S01786    417    17
LKSEISGDLARLILGLM                ANX6$HUMAN    416    17
LKSEISGDLARLILGLM                   HUMCBPE    417    17
LKSEISGDLARLILGLM                    HUMP68    417    17
LKGDLSGHFEHVMVALI                  ANX3$RAT     73    17
LKGDLSGHFEHVMVALI                    A29250     73    17
LKSALSGHLETVILGLL                    LUHU36     86    17
LKSALSGHLEAVILGLL                ANX2$CHICK     86    17
LKKALTGHLEEVVLALL                      LUHU     95    17
LKKALTGHLEEVVLALL                     !LPCH     96    17
LKKALTGHLEEVVLALL                ANX1$CAVCU     96    17
LRKALTGHLEEVVLAML                ANX1$MOUSE     95    17
LKKALTGHLEEVVLAML                  ANX1$RAT     95    17
LKSALSGHLETVILGLL                    LUMS36     86    17
LKSALSGHLETVILGLL                    LUBO36     86    17
LKSELTGKFETLMVSLM                ANX5$CHICK     69    17
MKSELTGKFEKLIVALM                  ANX5$RAT     67    17
LKSALSGHLETVILGLL                    A35600     86    17
LKSELTGKFEKLIVALM                ANX5$HUMAN     68    17
```

LKSELSGNFEQVILGMM	ANX4$BOVIN	67	17
LKSELSGKFERLIVALM	ANX8$HUMAN	75	17
LKGDLSGHFEHLMVALV	ANX3$HUMAN	72	17
LKSELSGNFEQVILGMM	ANX4$PIG	67	17
LKDELGGKFEDVIVGLM	DROANNX	69	17
LKSELSGNFEQVIVGMM	ANX4$HUMAN	67	17
MKRVLKSHLEDVVVALL	ANX1$COLLI	91	17
LKSELSGNMEELILALF	HUMSNEXIN	217	17

ANNEXIN3
22
motif 3

LKKAMEGAGTDEKTLIEILATR	ANX6$MOUSE	443	10
LKKAMEGAGTDEKTLIEILATR	S01786	444	10
LKKAMEGAGTDEKALIEILATR	ANX6$HUMAN	443	10
LKKAMEGAGTDEKALIEILATR	HUMCBPE	444	10
LKKAMEGAGTDEKALIEILATR	HUMP68	444	10
LKKSMRGMGTDEDTLIEILTTR	ANX3$RAT	100	10
LKKSMRGMGTDEDTLIEILTTR	A29250	100	10
LKASMKGLGTDEDSLIEIICSR	LUHU36	113	10
LKAAMKGLGTDEDTLIEIICSR	ANX2$CHICK	113	10
LRAAMKGLGTDEDTLIEILASR	LUHU	122	10
LRAAMKGLGTDEDTLIEILASR	!LPCH	123	10
LRAAMKGLGTDEDTLIEILVSR	ANX1$CAVCU	123	10
LRGAMKGLGTDEDTLIEILTTR	ANX1$MOUSE	122	10
LRAAMKGLGTDEDTLIEILTTR	ANX1$RAT	122	10
LKASMKGLGTDEDSLIEIICSR	LUMS36	113	10
LKASMKGLGTDEDSLIEIICSR	LUBO36	113	10
LKHAIKGAGTNEKVLTEILASR	ANX5$CHICK	96	10
LKHALKGAGTDEKVLTEIIASR	ANX5$RAT	94	10
LKASMKGLGTDEDSLIEIICSR	A35600	113	10
LKHALKGAGTNEKVLTEIIASR	ANX5$HUMAN	95	10
LRKAMKGAGTDEGCLIEILASR	ANX4$BOVIN	94	10
LHDAMKGLGTKEGVIIEILASR	ANX8$HUMAN	102	10
LKKSMKGAGTNEDALIEILTTR	ANX3$HUMAN	99	10
LRRAMKGAGTDEGCLIEILASR	ANX4$PIG	94	10
LHAAMAGIGTEEATLVEILCTK	DROANNX	96	10
LQRAMKGAGTDEGCLIEILASR	ANX4$HUMAN	94	10
LRACMKGHGTDEDTLIEILASR	ANX1$COLLI	118	10
LRKAMQGAGTQERVLIEILCTR	HUMSNEXIN	244	10

ANNEXIN4
27
motif 4

IADTPSGDKTSLETRFMTVLCTRSYPH	ANX6$MOUSE	531	66
IADTPSGDKTSLETRFMTVLCTRSYPH	S01786	532	66
IADTPSGDKTSLETRFMTILCTRSYPH	ANX6$HUMAN	531	66
IADTPSGDKTSLETRFMTILCTRTYPH	HUMCBPE	532	66
IADTPSGDKTSLETRFMTILCTRSYPH	HUMP68	532	66
LYDAGEKKWGTDEDKFTEILCLRSFPQ	ANX3$RAT	183	61
LYDAGEKKWGTDEDKFTEILCLRSFPQ	A29250	183	61
LYDAGVKRKGTDVPKWISIMTERSVPH	LUHU36	197	62
LYDAGVKRKGTDVPKWINIMTERSVPH	ANX2$CHICK	197	62
LYEAGERRKGTDVNVFNTILTTRSYPQ	LUHU	205	61
LYEAGERRKGTDVNVFNTILTTRSYPQ	!LPCH	206	61
LYEAGERRKGTDVNVFITILTTRSYSH	ANX1$CAVCU	206	61
LYEAGERRKGTDVNVFTTILTSRSFPH	ANX1$MOUSE	205	61

LYEAGERRKGTDVNVFNTILTTRSYPH	ANX1$RAT	205	61
LYDAGVKRKGTDVPKWISIMTERSVCH	LUMS36	197	62
LYDAGVKRKGTDVPKWISIMTERSVCH	LUBO36	197	62
LFRAGELKWGTDEETFITILGTRSVSH	ANX5$CHICK	179	61
LFQAGELKWGTDEEKFITILGTRSVSH	ANX5$RAT	177	61
LYDAGVKRKGTDVPKWISIMTERSVCH	A35600	197	62
LFQAGELKWGTDEEKFITIFGTRSVSH	ANX5$HUMAN	178	61
LYEAGEKKWGTDEVKFLTVLCSRNRNH	ANX4$BOVIN	177	61
LYAAGEKIRGTDEMKFITILCTRSATH	ANX8$HUMAN	186	62
LYKAGENRWGTDEDKFTEILCLRSFPQ	ANX3$HUMAN	182	61
LYEAGEKKWGTDEVKFLTVLCSRNRNH	ANX4$PIG	177	61
LYSAGEAKLGTDEEVFNRIMSHASFPQ	DROANNX	180	62
LYEAGEKKWGTDEVKFLTVLCSRNRNH	ANX4$HUMAN	177	61
LYEAGEQKKGTDINVFVTVLTARSYPH	ANX1$COLLI	201	61
LYQAGEGRLGTDESCFNMILATRSFPQ	HUMSNEXIN	327	61

ANNEXIN5
27
motif 5

MKGAGTDEKTLTRVMVSRSEIDLLNIR	ANX6$MOUSE	611	53
MKGAGTDEKTLTRVMVSRSEIDLLNIR	S01786	612	53
MKGAGTDEKTLTRIMVSRSEIDLLNIR	ANX6$HUMAN	611	53
MKGAGTDEKTLTRIMVSRSEIDLLNIR	HUMCBPE	612	53
MKGAGTDDKTLTRIMVSRSEIDLLNIR	HUMP68	612	53
LKGAGTDEFTLNRIMVSRSEIDLLDIR	ANX3$RAT	263	53
LKGAGTDEFTLNRIMVSRSEIDLLDIR	A29250	263	53
MKGKGTRDKVLIRIMVSRSEVDMLKIR	LUHU36	277	53
MKGKGTRDKVLIRIMVSRCEVDMLKIK	ANX2$CHICK	277	53
MKGVGTRHKALIRIMVSRSEIDMNDIK	LUHU	285	53
MKGVGTRHKALIRIMVSRSEIDMNDIK	!LPCH	286	53
MKGAGTRHKALIRIMVSRSEIDMNDIK	ANX1$CAVCU	286	53
MKGAGTRHKALIRIMVSRSEIDMNEIK	ANX1$MOUSE	285	53
MKGAGTRHKTLIRIMVSRSEIDMNEIK	ANX1$RAT	285	53
MKGKGTRDKVLIRIMVSRSEVDMLKIR	LUMS36	277	53
MKGKGTRDKVLIRIMVSRSEVDMLKIR	LUBO36	277	53
MKGAGTDDDTLIRVMVSRSEIDLLDIR	ANX5$CHICK	259	53
MKGAGTDDHTLIRVIVSRSEIDLFNIR	ANX5$RAT	257	53
MKGKGTRDKVLIRIMVSRSEVDMLKIR	A35600	277	53
MKGAGTDDHTLIRVMVSRSEIDLFNIR	ANX5$HUMAN	258	53
MKGLGTDDDTLIRVMVSRAEIDMLDIR	ANX4$BOVIN	257	53
MKGAGTRDGTLIRNIVSRSEIDLNLIK	ANX8$HUMAN	266	53
LKGIGTDEFTLNRIMVSRSEIDLLDIR	ANX3$HUMAN	262	53
MKGLGTDDNTLIRVMVSRAEIDMMDIR	ANX4$PIG	257	53
MNGAGTDDATLIRIIVSRSEIDLETIK	DROANNX	260	53
MKGLGTDDNTLIRVMVSRAEIDMLDIR	ANX4$HUMAN	257	53
MKGFGTQHRDLIRIMVSRHEVDMNEIK	ANX1$COLLI	280	52
MKGAGTDDSTLVRIVVTRSEIDLVQIK	HUMSNEXIN	407	53

ANNEXIN6
16
motif 6

EFIEKYDKSLHQAIEG	ANX6$MOUSE	639	1
EFIEKYDKSLHQAIEG	S01786	640	1
EFIEKYDKSLHQAIEG	ANX6$HUMAN	639	1
EFIEKYDKSLHQAIEG	HUMCBPE	640	1
EFIEKYDKSLHQAIEG	HUMP68	640	1
EFKKHYGCSLYSAIQS	ANX3$RAT	291	1

EFKKHYGCSLYSAIQS	A29250	291	1
EFKRKYGKSLYYYIQQ	LUHU36	305	1
EFKRKYGKSLYYFIQQ	ANX2$CHICK	305	1
FYQKMYGISLCQAILD	LUHU	313	1
FYQKMYGISLCQAILD	!LPCH	314	1
YYQKMYGISLCQAILD	ANX1$CAVCU	314	1
FYQKKYGISLCQAILD	ANX1$MOUSE	313	1
FYQKKYGIPLCQAILD	ANX1$RAT	313	1
EFKRKYGKSLYYYIQQ	LUMS36	305	1
EFKKKYGKSLYYYIQQ	LUBO36	305	1
EFRKNFAKSLYQMIQK	ANX5$CHICK	287	1
EFRKNFATSLYSMIKG	ANX5$RAT	285	1
EFKRKYGKSLYYYIQQ	A35600	305	1
EFRKNFATSLYSMIKG	ANX5$HUMAN	286	1
NFKRLYGKSLYSFIKG	ANX4$BOVIN	285	1
HFKKMYGKTLSSMIME	ANX8$HUMAN	294	1
EFKKHYGYSLYSAIKS	ANX3$HUMAN	290	1
NFKRLYGKSLYSFIKG	ANX4$PIG	285	1
EFERIYNRTLHSAVVD	DROANNX	288	1
HFKRLYGKSLYSFIKG	ANX4$HUMAN	285	1
YYKKMYGISLCQAIMD	ANX1$COLLI	308	1
MFAQMYQKTLGTMIAG	HUMSNEXIN	435	1

ANNEXIN7
15
motif 7

DTSGDFMKALLALCG	ANX6$MOUSE	655	0
DTSGDFMKALLALCG	S01786	656	0
DTSGDFLKALLALCG	ANX6$HUMAN	655	0
DTSGDFLKALLALCG	HUMCBPE	656	0
DTSGDFLKALLALCG	HUMP68	656	0
DTSGDYRTVLLKICG	ANX3$RAT	307	0
DTSGDYRTVLLKICG	A29250	307	0
DTKGDYQKALLYLCG	LUHU36	321	0
DTKGDYQRALLNLCG	ANX2$CHICK	321	0
ETKGDYEKILVALCG	LUHU	329	0
ETKGDYEKILVALCG	!LPCH	330	0
ETKGDYEKILVALCG	ANX1$CAVCU	330	0
ETKGDYEKILVALCG	ANX1$MOUSE	329	0
ETKGDYEKILVALCG	ANX1$RAT	329	0
DTKGDYQKALLYLCG	LUMS36	321	0
DTKGDYQKALLYLCG	LUBO36	321	0
DTSGDYRKALLLLCG	ANX5$CHICK	303	0
DTSGDYKKALLLLCG	ANX5$RAT	301	0
DTKGDYQKALLYLCG	A35600	321	0
DTSGDYKKALLLLCG	ANX5$HUMAN	302	0
DTSGDYRKVLLILCG	ANX4$BOVIN	301	0
DTSGDYKNALLSLVG	ANX8$HUMAN	310	0
DTSGDYEITLLKICG	ANX3$HUMAN	306	0
DTSGDYRKVLLILCG	ANX4$PIG	301	0
ETSGDYKRALTALLG	DROANNX	305	1
DTSGDYRKVLLVLCG	ANX4$HUMAN	301	0
ELKGGYETILVALCG	ANX1$COLLI	324	0
DTSGDYRRLLLAIVG	HUMSNEXIN	451	0

C.2 ATP sythases alpha and beta subunits

COMPOUND(6)
D.N. PERKINS 1/6/1991
FO-F1 ATP SYNTHASES

1. FUTAI, M., NOUMI, T., MAEDA, M., ATP synthase (H+-atpase): Results by combined biochemical and molecular biological approaches.
ANN. REV. BIOCHEM. 58 10541-10550 (1989)

2. AL-SHAWI, M.K., PARSONAGE, D., SENIOR, A.E. Thermodynamic analyses of the catalytic pathway of F1-ATPase from Escherichia coli.
J. BIOL. CHEM. 4402 265 (1990)

3. WALKER, J.E., SRASTE, M., RUNSWICK, M.J., GAY, N.J. Distantly related sequences in the alpha and beta subunits of ATP synthase, myosin, kinases and other ATP requiring enzymes and a common nucleotide binding fold.
EMBO JOURNAL 1 945-951 (1982)

ATP synthase catalyses the production of ATP from ADP and orthophosphate and consists of two components; the hydrophobic FO complex and the hydrophillic F1 complex. Both these complexes also consist of a number of subunits [1]. The alpha and beta chains of the F1 complex have the ability to bind both ATP and ADP. The alpha chain is thought to be involved with the regulation of ATP synthase activity whereas the beta chain contributes to the catalytic site [2]. Vacuolar ATPase is responsible for the acidification of a variety of intracellular compartments and the 60kD and 70kD subunits of these proteins show sequence similarity with the alpha and beta chains of F0-F1 ATP synthase.

Twelve sequences were used in the initial alignment, these being both alpha and beta subunits. From this alignment six motifs were selected and used to scan the OWL database. Motif two was derived from the region of the sequence shown to be responsible for the binding of ATP/ADP [3]. This region is conserved in a number of ATP binding families such as myosin and protein kinases and is also seen in GTP/GDP binding proteins, although there is now debate on the relative importance of this motif. The other five regions were chosen because of their high homology. Two iterations were required before convergence was reached. One sequence, database code RICCPCTB, was found to match with only four of the selected features. This protein is a mitochondrial beta and epsilon unit pseudogene derived from rice and shares little homology in the C terminus region with the other ATP synthases due to the mistranslation of the nucleic acid sequence. The other sequences shown to match with two features include one ATP binding protein (PR16$YEAST, PRP16 protein from yeast), the other proteins appear to constitute noise.

```
SUMMARY INFORMATION
-------------------
  76 codes involving 6 features
   0 codes involving 5 features
   1 code  involving 4 features
   0 codes involving 3 features
   5 codes involving 2 features

COMPOUND FEATURE INDEX
----------------------

 6|  76   76   76   76   76   76
 5|   0    0    0    0    0    0
 4|   1    1    1    1    0    0
 3|   0    0    0    0    0    0
 2|   1    2    2    1    3    1
--+-------------------------------
  |   1    2    3    4    5    6


True positives:
 WHTCPATPB       ATPB$YEAST      A24260          ATPB$BOVIN
 ATPB$HUMAN      ATPB$RAT        PWBSBM          ATP2$MAIZE
 ATP2$NICPL      ATPB$THEP3      HUMATPFIB       HUMATPSY2
 ATPB$IPOBA      PWNTB           ATPB$ANASP      PWSPB
 ATPB$RHOBL      BFIATPD         PWBOB           ATPB$RHORU
 ATPB$SYNP6      ATPB$CYTLY      ATP0$HELAN      ATP0$MAIZE
 ATP0$NICPL      ATP0$OENBI      ATP0$ORYSA      ATP0$PEA
 ATP0$WHEAT      PEAMTF14        !F1AB           ATPA$RHORU
 ATPB$BACFR      PWECB           ATPA$BOVIN      ATPA$RAT
 BOVATPSYN       A30245          ATPA$ANASP      PWLVA
 ATPA$XENLA      ATPA$YEAST      ATPB$VIBAL      PWRZA
 ATPA$WHEAT      ATPA$PEA        ATPA$MAIZE      PWNTA
 ATPA$SYNP6      ATPA$SPIOL      ATPA$RHOBL      ATPA$BACME
 ATPA$THEP3      ATPA$ECOLI      ATPA$VIBAL      PWECA
 SYNMTATPAA      MTPB$SULAC      VAT2$NEUCR      VAT2$YEAST
 A31487          VAT2$ARATH      MESATPAB1       VAT1$DAUCA
 VAT2$HUMAN      MESATPAB        MTPA$SULAC      VAT1$NEUCR
 VAT1$YEAST

 WHTCPATPB        ATP synthase beta subunit - Triticum aestivum
 ATPB$YEAST       ATP SYNTHASE BETA CHAIN - Yeast
 A24260           ATP synthase beta chain precursor - Yeast
 ATPB$BOVIN       ATP SYNTHASE BETA CHAIN - Bovine
 ATPB$HUMAN       ATP SYNTHASE BETA CHAIN - Human
 ATPB$RAT         ATP SYNTHASE BETA CHAIN - Rat
 PWBSBM           ATP synthase beta chain - Bacillus megaterium
 ATP2$MAIZE       ATP SYNTHASE BETA CHAIN, MITOCHONDRIAL - Maize
 ATP2$NICPL       ATP SYNTHASE BETA CHAIN, MITOCHONDRIAL - Tobacco
 ATPB$THEP3       ATP SYNTHASE BETA CHAIN - bacterium PS3
 HUMATPFIB        HUMATPFIB put. F1-beta precursor - Homo sapiens
 HUMATPSY2        ATP synthase beta subunit - Homo sapiens
 ATPB$IPOBA       ATP SYNTHASE BETA CHAIN - Sweet potato
 PWNTB            TP synthase beta chain - Common tobacco
 ATPB$ANASP       ATP SYNTHASE BETA CHAIN - Anabaena sp.
 PWSPB            ATP synthase beta chain - Spinach chloroplast
```

```
ATPB$RHOBL       ATP SYNTHASE BETA - Rhodopseudomonas blastica
BFIATPD          ATP synthase beta subunit - Bacillus firmus
PWBOB            ATP synthase, mitochondrial - Bovine
ATPB$RHORU       ATP SYNTHASE BETA CHAIN - Rhodospirillum rubrum
ATPB$SYNP6       ATP SYNTHASE BETA CHAIN - Synechococcus sp.
ATPB$CYTLY       ATP SYNTHASE BETA CHAIN - Cytophaga lytica
ATP0$HELAN       ATP SYNTHASE ALPHA CHAIN - Common sunflower
ATP0$MAIZE       ATP SYNTHASE ALPHA CHAIN - Maize
ATP0$NICPL       ATP SYNTHASE ALPHA CHAIN - Tobacco
ATP0$OENBI       ATP SYNTHASE ALPHA CHAIN - Oenothera biennis
ATP0$ORYSA       ATP SYNTHASE ALPHA CHAIN - Rice
ATP0$PEA         ATP SYNTHASE ALPHA CHAIN - Garden pea
ATP0$WHEAT       ATP SYNTHASE ALPHA CHAIN - Wheat
PEAMTF14         PEAMTF14 F-1-ATPase alpha subunit - Pisum sativum
!F1AB            F1 ATPASE, BETA SUBUNIT - E. coli
ATPA$RHORU       ATP SYNTHASE ALPHA CHAIN - Rhodospirillum rubrum
ATPB$BACFR       ATP SYNTHASE BETA CHAIN - Bacteroides fragilis
PWECB            ATP synthase beta chain - Escherichia coli
ATPA$BOVIN       ATP SYNTHASE ALPHA CHAIN - bovine
ATPA$RAT         ATP SYNTHASE ALPHA CHAIN - Rat
BOVATPSYN        alpha subunit ATP synthase isoform - Bos taurus
A30245           ATP synthase alpha chain precursor - Bovine
ATPA$ANASP       ATP SYNTHASE ALPHA CHAIN - Anabaena sp.
PWLVA            ATP synthase alpha chain - Liverwort
ATPA$XENLA       ATP SYNTHASE ALPHA CHAIN - African clawed frog
ATPA$YEAST       ATP SYNTHASE ALPHA CHAIN - Yeast
ATPB$VIBAL       ATP SYNTHASE BETA CHAIN - Vibrio alginolyticus
PWRZA            ATP synthase alpha chain - Rice chloroplast
ATPA$WHEAT       ATP SYNTHASE ALPHA CHAIN - Wheat
ATPA$PEA         ATP SYNTHASE ALPHA CHAIN - Garden pea
ATPA$MAIZE       ATP SYNTHASE ALPHA CHAIN - Maize
PWNTA            ATP synthase alpha chain - Common tobacco
ATPA$SYNP6       ATP SYNTHASE ALPHA CHAIN - Synechococcus sp.
ATPA$SPIOL       ATP SYNTHASE ALPHA CHAIN - Spinach
ATPA$RHOBL       ATP SYNTHASE ALPHA - Rhodopseudomonas blastica
ATPA$BACME       ATP SYNTHASE ALPHA CHAIN - Bacillus megaterium
ATPA$THEP3       ATP SYNTHASE ALPHA - Thermophilic bacterium PS-3.
ATPA$ECOLI       ATP SYNTHASE ALPHA CHAIN - E. coli
ATPA$VIBAL       ATP SYNTHASE ALPHA CHAIN - Vibrio alginolyticus
PWECA            ATP synthase alpha chain - Escherichia coli
SYNMTATPAA       ATP synthase alpha subunit - Artificial gene
MTPB$SULAC       ATPASE BETA CHAIN - Sulfolobus acidocaldarius
VAT2$NEUCR       VACUOLAR ATP SYNTHASE 57 KD - Neurosporra crassa
VAT2$YEAST       VACUOLAR ATP SYNTHASE SUBUNIT B - Baker's yeast
A31487           *H+-transporting ATP synthase B chain - Yeast
VAT2$ARATH       VACUOLAR ATP SYNTHASE 57 KD  - Mouse-ear cress
MESATPAB1        ATPase beta subunit - Methanosarcina barkeri
VAT1$DAUCA       VACUOLAR ATP SYNTHASE 69 KD SUBUNIT - CARROT
VAT2$HUMAN       VACUOLAR ATP SYNTHASE 58 KD SUBUNIT - Human
MESATPAB         ATPase alpha subunit - Methanosarcina barkeri
MTPA$SULAC       MEMBRANE ATPASE - Sulfolobus acidocaldarius
VAT1$NEUCR       VACUOLAR ATP SYNTHASE 67 KD - Neurosporra crassa
VAT1$YEAST       VACUOLAR ATP SYNTHASE CATALYTIC SUBUNIT A - Yeast

SCAN HISTORY
------------
OWL11_0     2     0 NSINGLE
```

```
INITIAL MOTIF-SETS
------------------
ATP1
15
motif 1
ETGIKVVDLLAPYAR              ATPB$YEAST   168   168
QTGIKAVDSLVPIGR              ATPA$BOVIN   190   190
ETGIKVVDLLAPYRR                   PWLVB   148   148
VTGIKVVDLLAPYQR              ATP2$NICPL   213   213
QTGISAIDGLNSLLR              MTPB$SULAC   133   133
QTGISPIDVMNSIAR              VAT2$HUMAN   163   163
LTGIRVLDTVFPIAK              MTPA$SULAC   211   211
QTGLIAIDSMIPIGR              ATPA$MAIZE   148   148
STGVSAIDTMNSIAR              VAT2$YEAST   152   152
QTGISTIDGTNTLVR               MESATPAB1   128   128
VTGMRILDGLFPVAK                MESATPAB   206   206
LTGQRVLDALFPSVL              VAT1$DAUCA   230   230


ATP2
20
motif 2
GGKIGLFGGAGVGKTVFIQE         ATPB$YEAST   183     0
GQRELIIGDRQTGKTSIAID         ATPA$BOVIN   205     0
GGKIGLFGGAGVGKTVLIME              PWLVB   163     0
GGKIGLFGGAGVGKTVLIME         ATP2$NICPL   228     0
GSKITDLSGSGLPANTLAAQ         MTPB$SULAC   148     0
GQKIPIFSAAGLPHNEIAAQ         VAT2$HUMAN   178     0
GGTAAIPGPFGSGKTVTLQS         MTPA$SULAC   226     0
GQRELIIGDRQTGKTAVATD         ATPA$MAIZE   163     0
GQKIPIFSASGLPHNEIAAQ         VAT2$YEAST   167     0
GQKLPIFSASGLPHNEIALQ          MESATPAB1   143     0
GGTAAIPGPFGSGKTVTQQS           MESATPAB   221     0
GGTCAIPGAFGCGKTVISQA         VAT1$DAUCA   245     0

ATP3
14
motif 3
FSVFAGVGERTREG               ATPB$YEAST   214    11
YCIYVAIGQKRSTV               ATPA$BOVIN   243    18
VSVFGGVGERTREG                    PWLVB   194    11
FSVFAGVGERTREG               ATP2$NICPL   259    11
AVVFAAIGVRYDEA               MTPB$SULAC   182    14
AIVFAAMGVNMETA               VAT2$HUMAN   220    22
VVIYVGCGERGNEM               MTPA$SULAC   254     8
ICVYVAIGQRASSV               ATPA$MAIZE   193    10
SIVFAAMGVNLETA               VAT2$YEAST   209    22
AVVFAAMGITNEEA                MESATPAB1   177    14
IVVYIGCGERGNEM                 MESATPAB   249     8
TVVYVGCGERGNEM               VAT1$DAUCA   273     8

ATP4
23
motif 4
QMNEPPGARARVALTGLTIAEYF      ATPB$YEAST   254    26
TASDAAPLQYLAPYSGCSMGEYF      ATPA$BOVIN   278    21
```

```
QMNEPPGARMRVGLTALTMAEYF              PWLVB      236     28
QMNEPPGARARVGLTGLTVAEHF         ATP2$NICPL      302     29
LANDPPSLKILTPKTALTLAEYL         MTPB$SULAC      217     21
LANDPTIERIITPRLALTTAEFL         VAT2$HUMAN      255     21
TSNMPVAARESSIYVGVTMAEYF         MTPA$SULAC      296     28
MADSPATLQYLAPYTGAALAEYF         ATPA$MAIZE      228     21
LANDPTIERIITPRLALTTAEYL         VAT2$YEAST      244     21
LADDPAVERIVTPRMALTAAEYL          MESATPAB1      212     21
TSNMPVAAREASVYTGITIAEYY           MESATPAB      291     28
TSNMPVAAREASIYTGITIAEYF         VAT1$DAUCA      318     31

ATP5
23
motif 5
RIPSAVGYQPTLATDMGLLQERI         ATPB$YEAST      307     30
RPPGREAYPGDVFYLHSRLLERA         ATPA$BOVIN      330     29
RMPSAVGYQPTLSTEMGTLQERI              PWLVB      289     30
RIPSAVGYQPTLATDLGGLQERI         ATP2$NICPL      355     30
EVPGRGGYPGYMYTDLATIYERA         MTPB$SULAC      270     30
EVPGRRGFPGYMYTDLATIYERA         VAT2$HUMAN      308     30
EMPAEEGFPSYLPSRLAEYYERA         MTPA$SULAC      348     29
RPPGREAYLGDVFYLHSRLLERA         ATPA$MAIZE      280     29
EVPGRRGYPGYMYTDLSTIYERA         VAT2$YEAST      297     30
EIPGRRGYPGYMYTDLATLYERA          MESATPAB1      265     30
EMPGEEGYPAYLSARLAEFYERA           MESATPAB      343     29
EMPADSGYPAYLAARLASFYERA         VAT1$DAUCA      370     29

ATP6
17
motif 6
LGIYPAVDPLDSKSRLL               ATPB$YEAST      375     45
KGIRPAINVGLSVSRVG               ATPA$BOVIN      402     49
KGIYPAVDPLDSTSTML                    PWLVB      357     45
LGIYPAVDPLDSTSRML               ATP2$NICPL      423     45
KGIYPPINVLMSLSRLM               MTPB$SULAC      340     47
RQIYPPINVLPSLSRLM               VAT2$HUMAN      378     47
ARHYPAINWIQGFSAYV               MTPA$SULAC      423     52
AGIRPAINVGISVSRVG               ATPA$MAIZE      352     49
KGIYPPINVLPSLSRLM               VAT2$YEAST      367     47
KGIYPPINVLPSLSRLM                MESATPAB1      335     47
RRHFPAINWLNSYSLYK                 MESATPAB      416     50
RKHFPSVNWLISYSKYS               VAT1$DAUCA      445     52

FINAL MOTIF-SETS
----------------
ATP1
15
motif 1
QTGIKAVDSLVPIGR                 ATPA$BOVIN      190     190
QTGIKAVDSLVPIGR                   ATPA$RAT      180     180
QTGIKAVDSLVPIGR                 ATPA$XENLA      191     191
QTGIKAVDSLVPIGR                  BOVATPSYN      190     190
QTGIKAVDSLVPIGR                     A30245      190     190
ETGIKVVDLLAPYAR                 ATPB$YEAST      168     168
ETGIKVVDLLAPYAR                     A24260      166     166
ETGIKVVDLLAPYRR                      PWLVB      148     148
ETGIKVVDLLAPYRR                      PWZMB      150     150
```

ETGIKVVDLLAPYRR	PWBHB	150	150
ETGIKVVDLLAPYRR	PWRZB	150	150
ETGIKVVDLLAPYRR	ATPB$CHLRE	150	150
ETGIKVVDLLAPYRR	ATPB$IPOBA	148	148
ETGIKVVDLLAPYRR	ATPB$PEA	150	150
ETGIKVVDLLAPYRR	RICCPCTA	150	150
ETGIKVVDLLAPYRR	WHTCPATPB	150	150
QTGIKAIDSLIPIGR	ATPA$RHORU	147	147
QTGLKAVDSLVPIGR	ATP0$HELAN	149	149
QTGLKAVDSLVPIGR	ATP0$MAIZE	149	149
QTGLKAVDSLVPIGR	ATP0$NICPL	149	149
QTGLKAVDSLVPIGR	ATP0$OENBI	149	149
QTGLKAVDSLVPIGR	ATP0$ORYSA	149	149
QTGLKAVDSLVPIGR	ATP0$PEA	149	149
QTGLKAVDSLVPIGR	ATP0$WHEAT	149	149
QTGLKAVDSLVPIGR	PEAMTF14	149	149
QTGIKAIDALVPIGR	ATPA$BACME	147	147
QTGIKAIDALVPIGR	ATPA$THEP3	147	147
VTGIKVVDLLAPYQR	ATP2$MAIZE	206	206
VTGIKVVDLLAPYQR	ATP2$NICPL	213	213
ETGIKVVDLLTPYRR	ATPB$ANASP	140	140
QTGLKAVDALVPIGR	ATPA$YEAST	184	184
VTGIKVVDLLAPYAK	ATPB$BOVIN	184	184
VTGIKVVDLLAPYAK	ATPB$HUMAN	184	184
VTGIKVVDLLAPYAK	ATPB$RAT	184	184
VTGIKVVDLLAPYAK	HUMATPFIB	194	194
VTGIKVVDLLAPYAK	HUMATPSY2	184	184
ETGIKVVDLLAPYIK	PWBSBM	136	136
ETGIKVVDLLAPYIK	ATPB$THEP3	136	136
ETGIEVVDLLAPYRR	PWNTB	150	150
QTGYKAVDSMIPIGR	PWECA	147	147
QTGYKAVDSMIPIGR	ATPA$ECOLI	147	147
ETGIKVVDLLAPYII	BFIATPD	134	134
ETGIKVIDLLAPYRQ	ATPB$SYNP6	140	140
ETGIKVVNLLAPYRR	PWSPB	150	150
VTGIKVIDLLAPYSK	ATPB$RHOBL	133	133
QTGITAIDSMIPIGR	ATPA$ANASP	149	149
VTGDKVVDLLAPYAK	PWBOB	134	134
ETGIKVIDLMCPFAR	!F1AB	128	128
QTGYKSVDSMIPIGR	ATPA$VIBAL	147	147
ATGLKAVDAMIPIGR	ATPA$RHOBL	147	147
QTGITAIDAMIPIGR	ATPA$SYNP6	148	148
FTGIKVIDLLEPYSK	ATPB$BACFR	135	135
QTGLIAIDSMIPIGR	PWLVA	148	148
QTGLIAIDSMIPIGR	PWNTA	148	148
QTGLIAIDSMIPIGR	PWRZA	148	148
QTGLIAIDSMIPIGR	ATPA$MAIZE	148	148
QTGLIAIDSMIPIGR	ATPA$PEA	148	148
QTGLIAIDSMIPIGR	ATPA$WHEAT	148	148
VTGIKVIDLIAPYTK	ATPB$RHORU	130	130
ETGIKVIDLMCPFAK	PWECB	128	128
LTGYKIVDSMLPIGR	SYNMTATPAA	150	150
FTGIKVIDLIEPYAK	ATPB$CYTLY	134	134
QTGLIAIDAMIPVGR	ATPA$SPIOL	148	148
ETGVKVIDLICPFAK	ATPB$VIBAL	127	127
QTGISAIDGLNSLLR	MTPB$SULAC	133	133
STGISAIDTMNSIAR	VAT2$NEUCR	142	142

QTGISTIDVMNSIAR	VAT2$ARATH	149	149
QTGISPIDVMNSIAR	VAT2$HUMAN	163	163
LTGIRVLDTVFPIAK	MTPA$SULAC	211	211
STGVSAIDTMNSIAR	VAT2$YEAST	152	152

ATP2
20
motif 2

GQRELIIGDRQTGKTSIAID	ATPA$BOVIN	205	0
GQRELIIGDRQTGKTSIAID	ATPA$RAT	195	0
GQRELIIGDRQTGKTSIAID	ATPA$XENLA	206	0
GQRELIIGDRQTGKTSIAID	BOVATPSYN	205	0
GQRELIIGDRQTGKTSIAID	A30245	205	0
GGKIGLFGGAGVGKTVFIQE	ATPB$YEAST	183	0
GGKIGLFGGAGVGKTVFIQE	A24260	181	0
GGKIGLFGGAGVGKTVLIME	PWLVB	163	0
GGKIGLFGGAGVGKTVLIME	PWZMB	165	0
GGKIGLFGGAGVGKTVLIME	PWBHB	165	0
GGKIGLFGGAGVGKTVLIME	PWRZB	165	0
GGKIGLFGGAGVGKTVLIME	ATPB$CHLRE	165	0
GGKIGLFGGAGVGKTVLIME	ATPB$IPOBA	163	0
GGKIGLFGGAGVGKTVLIME	ATPB$PEA	165	0
GGKIGLFGGAGVGKTVLIME	RICCPCTA	165	0
GGKIGLFGGAGVGKTVLIME	WHTCPATPB	165	0
GQRELIIGDRQTGKTAVILD	ATPA$RHORU	162	0
GQRELIIGDRQTGKTAIAID	ATP0$HELAN	164	0
GQRELIIGDRQTGKTAIAID	ATP0$MAIZE	164	0
GQRELIIGDRQTGKTAIAID	ATP0$NICPL	164	0
GQRELIIGDRQTGKTAIAID	ATP0$OENBI	164	0
GQRELIIGDRQTGKTAIAID	ATP0$ORYSA	164	0
GQRELIIGDRQTGKTAIAID	ATP0$PEA	164	0
GQRELIIGDRQTGKTAIAID	ATP0$WHEAT	164	0
GQRELIIGDRQTGKTAIAID	PEAMTF14	164	0
GQRELIIGDRQTGKTSVAID	ATPA$BACME	162	0
GQRELIIGDRQTGKTSVAID	ATPA$THEP3	162	0
GGKIGLFGGAGVGKTVLIME	ATP2$MAIZE	221	0
GGKIGLFGGAGVGKTVLIME	ATP2$NICPL	228	0
GGKIGLFGGAGVGKTVIMME	ATPB$ANASP	155	0
GQRELIIGDRQTGKTAVALD	ATPA$YEAST	199	0
GGKIGLFGGAGVGKTVLIME	ATPB$BOVIN	199	0
GGKIGLFGGAGVGKTVLIME	ATPB$HUMAN	199	0
GGKIGLFGGAGVGKTVLIME	ATPB$RAT	199	0
GGKIGLFGGAGVGKTVLIME	HUMATPFIB	209	0
GGKIGLFGGAGVGKTVLIME	HUMATPSY2	199	0
GGKIGLFGGAGVGKTVLIQE	PWBSBM	151	0
GGKIGLFGGAGVGKTVLIQE	ATPB$THEP3	151	0
GGKIGLFGGAGVGKTVLIME	PWNTB	165	0
GQRELIIGDRQTGKTRLAID	PWECA	162	0
GQRELIIGDRQTGKTALAID	ATPA$ECOLI	162	0
GGKIGLFGGAGVGKTVLIQE	BFIATPD	149	0
GGKIGLFGGAGVGKTVLIQE	ATPB$SYNP6	155	0
GGKIGLFGGAGVGKTVLIME	PWSPB	165	0
GGKIGLFGGAGVGKTVLIQE	ATPB$RHOBL	148	0
GQRELIIGDRQTGKTAIAID	ATPA$ANASP	164	0
GGKIGLFGGAGVGKTVFIME	PWBOB	149	0
GGKVGLFGGAGVGKTVNMME	!F1AB	143	0
GQRELIIGDRQIGKTALAID	ATPA$VIBAL	162	0

GQRELIIGDRQTGKTAVALD	ATPA$RHOBL	162	0
GQRELIIGDRQTGKTAIAID	ATPA$SYNP6	163	0
GGKIGLFGGAGVGKTVLIME	ATPB$BACFR	150	0
GQRELIIGDRQTGKTAVAID	PWLVA	163	0
GQRELIIGDRQTGKTAVATD	PWNTA	163	0
GQRELIIGDRQTGKTAVATD	PWRZA	163	0
GQRELIIGDRQTGKTAVATD	ATPA$MAIZE	163	0
GQRELIIGDRQTGKTAVATD	ATPA$PEA	163	0
GQRELIIGDRQTGKTAVATD	ATPA$WHEAT	163	0
GGKVGLFGGAGVGKTVLIQE	ATPB$RHORU	145	0
GGKVGLFGGAGVGKTVNMME	PWECB	143	0
GQRELIVGDRQTGKTTIAID	SYNMTATPAA	165	0
GGKIGLFGGAGVGKTVLIQE	ATPB$CYTLY	149	0
GQRELIIGDRQTGKTAVATD	ATPA$SPIOL	163	0
GGKIGLFGGAGVGKTVNMME	ATPB$VIBAL	142	0
GSKITDLSGSGLPANTLAAQ	MTPB$SULAC	148	0
GQKIPIFSAAGLPHNEIAAQ	VAT2$NEUCR	157	0
GQKIPLFSAAGLPHNEIAAQ	VAT2$ARATH	164	0
GQKIPIFSAAGLPHNEIAAQ	VAT2$HUMAN	178	0
GGTAAIPGPFGSGKTVTLQS	MTPA$SULAC	226	0
GQKIPIFSASGLPHNEIAAQ	VAT2$YEAST	167	0

ATP3
14
motif 3

YCIYVAIGQKRSTV	ATPA$BOVIN	243	18
YCIYVAIGQKRSTV	ATPA$RAT	233	18
YCIYVAIGQKRLTD	ATPA$XENLA	244	18
YCIYVAIGQKRSTV	BOVATPSYN	243	18
YCIYVAIGQKRSTV	A30245	243	18
FSVFAGVGERTREG	ATPB$YEAST	214	11
FSVFAGVGERTREG	A24260	212	11
VSVFGGVGERTREG	PWLVB	194	11
VSVFGGVGERTREG	PWZMB	196	11
VSVFGGVGERTREG	PWBHB	196	11
VSVFGGVGERTREG	PWRZB	196	11
VSVFAGVGERTREG	ATPB$CHLRE	196	11
VSVFGGVGERTREG	ATPB$IPOBA	194	11
VSVFGGVGERTREG	ATPB$PEA	196	11
VSVFGGVGERTREG	RICCPCTA	196	11
VSVFGGVGERTREG	WHTCPATPB	196	11
FCVYVAVGQKRSTV	ATPA$RHORU	201	19
YCVYVAIGQKRSTV	ATP0$HELAN	203	19
YCVYVAIGQKRSTV	ATP0$MAIZE	203	19
YCVYVAIGQKRSTV	ATP0$NICPL	203	19
YCVYVAIGQKRSTV	ATP0$OENBI	203	19
YCVYVAIGQKRSTV	ATP0$ORYSA	203	19
YCVYVAIGQKRSTV	ATP0$PEA	203	19
YCVYVAIGQKRSTV	ATP0$WHEAT	203	19
YCVYVAIGQKRSTV	PEAMTF14	203	19
VCIYVAIGQKESTV	ATPA$BACME	192	10
ICIYVAIGQKESTV	ATPA$THEP3	192	10
FSVFAGVGERTREG	ATP2$MAIZE	252	11
FSVFAGVGERTREG	ATP2$NICPL	259	11
VSVFAGVGERTREG	ATPB$ANASP	186	11
YCVYVAVGQKRSTV	ATPA$YEAST	237	18
YSVFAGVGERTREG	ATPB$BOVIN	230	11

YSVFAGVGERTREG	ATPB$HUMAN	230	11
YSVFAGVGERTREG	ATPB$RAT	230	11
YSVFAGVGERTREG	HUMATPFIB	240	11
YSVFAGVGERTREG	HUMATPSY2	230	11
ISVFAGVGERTREG	PWBSBM	182	11
ISVFAGVGERTREG	ATPB$THEP3	182	11
VSVFGGVGERTREG	PWNTB	196	11
KCIYVAIGQKASTI	PWECA	192	10
KCIYVAIGQKASTI	ATPA$ECOLI	192	10
ISVFAGVGERTREG	BFIATPD	180	11
VSVFGGVGERTREG	ATPB$SYNP6	186	11
VSVFGGVGERTREG	PWSPB	196	11
YSVFAGVGERTREG	ATPB$RHOBL	179	11
VCVYVAIGQKASTV	ATPA$ANASP	194	10
YSVFAGVGERTREG	PWBOB	180	11
YSVFAGVGERTREG	!F1AB	174	11
FSIYVAIGQKASTI	ATPA$VIBAL	192	10
YCVYVAIGQKRSTV	ATPA$RHOBL	202	20
ICVYVAIGQKASSV	ATPA$SYNP6	193	10
FSVFAGVGERTREG	ATPB$BACFR	181	11
VCVYVAIGQKASSV	PWLVA	193	10
ICVYVAIGQKASSV	PWNTA	193	10
ICVYVAIGQRASSV	PWRZA	193	10
ICVYVAIGQRASSV	ATPA$MAIZE	193	10
VCVYVAIGQKASSV	ATPA$PEA	193	10
ICVYVAIGQRASSV	ATPA$WHEAT	193	10
YSVFAGVGERTREG	ATPB$RHORU	176	11
YSVFAGVGERTREG	PWECB	174	11
YCVYVGIGQKKSSI	SYNMTATPAA	200	15
LSVFAGVGERTREG	ATPB$CYTLY	180	11
ICVYVAIGQKASSV	ATPA$SPIOL	193	10
LSVFAGVGERTREG	ATPB$VIBAL	173	11
AVVFAAIGVRYDEA	MTPB$SULAC	182	14
SIVFGAMGVNLETA	VAT2$NEUCR	203	26
AIVFAAMGVNMETA	VAT2$ARATH	210	26
AIVFAAMGVNMETA	VAT2$HUMAN	220	22
VVIYVGCGERGNEM	MTPA$SULAC	254	8
SIVFAAMGVNLETA	VAT2$YEAST	209	22

ATP4

23

motif 4

TASDAAPLQYLAPYSGCSMGEYF	ATPA$BOVIN	278	21
TASDAAPLQYLAPYSGCSMGEYF	ATPA$RAT	268	21
TASDAAPLQYLAPYSGCSMGEYF	ATPA$XENLA	270	12
TASDAAPLQYLAPYSGCSMGEYF	BOVATPSYN	278	21
TASDAAPLQYLAPYSGCSMGEYF	A30245	278	21
QMNEPPGARARVALTGLTIAEYF	ATPB$YEAST	254	26
QMNEPPGARARVALTGLTIAEYF	A24260	252	26
QMNEPPGARMRVGLTALTMAEYF	PWLVB	236	28
QMNEPPGARMRVGLTALTMAEYF	PWZMB	238	28
QMNEPPGARMRVGLTALTMAEYF	PWBHB	238	28
QMNEPPGARMRVGLTALTMAEYF	PWRZB	238	28
QMNEPPGARMRVALTALTMAEYF	ATPB$CHLRE	238	28
GQNEPPGARMRVGLTALTMAEYF	ATPB$IPOBA	235	27
QMNEPPGARMRVGLTALTMAEYF	ATPB$PEA	238	28
QMNEPPGARMRVGLTALTMAEYF	RICCPCTA	238	28

QMNEPPGARMRVGLTALTMAEYF	WHTCPATPB	238	28
TASEPAPLQFLAPYTGCTMGEFF	ATPA$RHORU	236	21
TASDPAPLQFLAPYSGCAMGEYF	ATP0$HELAN	238	21
TASDPAPLQFLAPYSGCAMGEYF	ATP0$MAIZE	238	21
TASDPAPLQFLAPYSGCAMGEYF	ATP0$NICPL	238	21
TASDPAPLQFLAPYSGCAMGEYF	ATP0$OENBI	238	21
TASDPAPLQFLAPYSGCAMGEYF	ATP0$ORYSA	238	21
TASDPAPLQFLAPYSGCAMGEYF	ATP0$PEA	238	21
TASDPAPLQFLAPYSGCAMGEYF	ATP0$WHEAT	238	21
TASDPAPLQFLAPYSGCAMGEYF	PEAMTF14	238	21
SASQPAPLLFLAPYAGVTMGEEF	ATPA$BACME	227	21
SASQPAPLLFLAPYAGVAMGEYF	ATPA$THEP3	227	21
QMNEPPGARARVGLTGLTVAEHF	ATP2$MAIZE	295	29
QMNEPPGARARVGLTGLTVAEHF	ATP2$NICPL	302	29
QMNEPPGARMRVGLSGLTMAEYF	ATPB$ANASP	217	17
TASEAAPLQYLAPFTAASIGEWF	ATPA$YEAST	272	21
QMNEPPGARARVALTGLTVAEYF	ATPB$BOVIN	271	27
QMNEPPGARARVALTGLTVAEYF	ATPB$HUMAN	271	27
QMNEPPGARARVALTGLTVAEYF	ATPB$RAT	271	27
QMNQPPGARARVALTGLTVAEYF	HUMATPFIB	281	27
QMNQPPGARARVALTGLTVAEYF	HUMATPSY2	271	27
QMNEPPGARQRVALTGLTMAEYF	PWBSBM	217	21
QMNEPPGARMRVALTGLTMAEYF	ATPB$THEP3	217	21
QMNEPPGARMRVGLTALTMAEYF	PWNTB	238	28
TASESAALQYLARMPVALMGEYF	PWECA	227	21
TASESAALQYLARMPVALMGEYF	ATPA$ECOLI	227	21
QMNEPPGARMAVALSGLTMAEHF	BFIATPD	215	21
QMNEPPGARMRVGLSALTMAEHF	ATPB$SYNP6	228	28
QMNEPPGARMRVGLTALTMAEYF	PWSPB	238	28
QMNEPPGARARVALTGLTLAEQF	ATPB$RHOBL	221	28
GASEPATLQFLAPYTGATIAEYF	ATPA$ANASP	229	21
QMNQPPGARARVALTGLTVAEYF	PWBOB	221	27
QMNQPPGNRLRVALTGLTMAEKF	!F1AB	209	21
SASESAALQYLAPYAGCAMGEYF	ATPA$VIBAL	227	21
TASDPAPMQFLAPFSGTAIGEFF	ATPA$RHOBL	237	21
NASEPATLQYLAPYAGAAIAEYF	ATPA$SYNP6	228	21
QMNEPPGARASVALSGLTVAESF	ATPB$BACFR	243	48
TANSPATLQYLAPYTGAALAEYF	PWLVA	228	21
TADSPATLQYLAPYTGAALAEYF	PWNTA	228	21
MADSPATLQYLAPYTGAALAEYF	PWRZA	228	21
MADSPATLQYLAPYTGAALAEYF	ATPA$MAIZE	228	21
TADSPATLQYLAPYTGAALAEYF	ATPA$PEA	228	21
MADSPATLQYLAPYTGAALAEYF	ATPA$WHEAT	228	21
QMNEPPGARARVALAGLTQAEYF	ATPB$RHORU	217	27
QMNEPPGNRLRVALTGLTMAEKF	PWECB	209	21
TAAQSASLQFIAPYTGCAIAEFY	SYNMTATPAA	235	21
QMNEPPGARARVALSGLTIAEYF	ATPB$CYTLY	242	48
TADSPATLQYLAPYTGAALAEYF	ATPA$SPIOL	228	21
QMNEPPGNRLRVALTGLTMAERF	ATPB$VIBAL	215	28
LANDPPSLKILTPKTALTLAEYL	MTPB$SULAC	217	21
LANDPTIERIITPRLALTTAEYY	VAT2$NEUCR	238	21
LANDPTIERIITPRIALTTAEYL	VAT2$ARATH	245	21
LANDPTIERIITPRLALTTAEFL	VAT2$HUMAN	255	21
TSNMPVAARESSIYVGVTMAEYF	MTPA$SULAC	296	28
LANDPTIERIITPRLALTTAEYL	VAT2$YEAST	244	21

ATP5
23
motif 5

RPPGREAYPGDVFYLHSRLLERA	ATPA$BOVIN	330	29
RPPGREAYPGDVFYLHSRLLERA	ATPA$RAT	320	29
RPPGREAYPGDVFYLHSRLLERA	ATPA$XENLA	322	29
RPPGREAYPGDVFYLHSRLLERA	BOVATPSYN	330	29
RPPGREAYPGDVFYLHSRLLERA	A30245	330	29
RIPSAVGYQPTLATDMGLLQERI	ATPB$YEAST	307	30
RIPSAVGYQPTLATDMGLLQERI	A24260	305	30
RMPSAVGYQPTLSTEMGTLQERI	PWLVB	289	30
RMPSAVGYQPTLSTEMGSLQERI	PWZMB	291	30
RMPSAVGYQPTLSTEMGSLQERI	PWBHB	291	30
RMPSAVGYQPTLSTEMGSLQERI	PWRZB	291	30
RMPSAVGYQPTLATEMGGLQERI	ATPB$CHLRE	291	30
RMPSAVGYQPTLSTEMGYLQERI	ATPB$IPOBA	288	30
RMPSAVGYQPTLGTEMGTLQERI	ATPB$PEA	291	30
RMPSAVGYQPTLSTEMGSLQERI	RICCPCTA	291	30
RMPSAVGYQPTLSTEMGSLQERI	WHTCPATPB	291	30
RPPGREAFPGDVFYLHSRLLERA	ATPA$RHORU	288	29
RPPGREAFPGDVFYLHSRLLERA	ATP0$HELAN	290	29
RPPGREAFPGDVFYLHSRLLERA	ATP0$MAIZE	290	29
RPPGREAFPGDVFYLHSRLLERA	ATP0$NICPL	290	29
RPPGREAFPGDVFYLHSRLLERA	ATP0$OENBI	290	29
RPPGREAFPGDVFYLHSRLLERA	ATP0$ORYSA	290	29
RPPGREAFPGDVFYLHSRLLERA	ATP0$PEA	290	29
RPPGREAFPGDVFYLHSRLLERA	ATP0$WHEAT	290	29
RPPGREAFPGDVFYLHSRLLERA	PEAMTF14	290	29
RPPGREAYPGDVFYLHSRLLERA	ATPA$BACME	279	29
RPPGREAYPGDIFYLHSRLLERA	ATPA$THEP3	279	29
RIPSAVGYQPTLATDLGGLQERI	ATP2$MAIZE	348	30
RIPSAVGYQPTLATDLGGLQERI	ATP2$NICPL	355	30
RMPSAVGYQPTLGTDVGQLQERI	ATPB$ANASP	270	30
RPPGREAYPGDVFYLHSRLLERA	ATPA$YEAST	324	29
RIPSAVGYQPTLATDMGTMQERI	ATPB$BOVIN	324	30
RIPSAVGYQPTLATDMGTMQERI	ATPB$HUMAN	324	30
RIPSAVGYQPTLATDMGTMQERI	ATPB$RAT	324	30
RIPSAVGYQPTLATDMGTMQERI	HUMATPFIB	334	30
RIPSAVGYQPTLATDMGTMQERI	HUMATPSY2	324	30
RMPSAVGYQPTLATEMGQLQERI	PWBSBM	270	30
RMPSAIGYQPTLATEMGQLQERI	ATPB$THEP3	270	30
RMPSAVGYQPTLSTEMGSLQERI	PWNTB	291	30
RPPGREAFPGDVFYLHSRLLEML	PWECA	279	29
RPPGREAFPGDVFYLHSRLLERA	ATPA$ECOLI	279	29
RMPSAVGYQPTLATEMGQLQERI	BFIATPD	267	29
RMPSAVGYQPTLGTDVGQLQERI	ATPB$SYNP6	281	30
RMPSAVGYQPTLSTEMGSLQERI	PWSPB	291	30
RIPSAVGYQPTLATDMGQLQERI	ATPB$RHOBL	274	30
RPPGGEAYPGDVFYIHSRLLERA	ATPA$ANASP	281	29
RIPSAVGYQPTLATNMGTMQERI	PWBOB	274	30
RMPSAVGYQPTLAEEMGVLQERI	!F1AB	261	29
RPPGREAFPGDVFYLHSRLLERA	ATPA$VIBAL	279	29
RPPGREAYPGDVFYLHSRLLERS	ATPA$RHOBL	289	29
RPPGREAYPGDVFYLHSRLLERA	ATPA$SYNP6	280	29
RMPSAVGYQPTLATEMGAMQERI	ATPB$BACFR	300	34
RPPGREAYPGDVFYLHSRLLERA	PWLVA	280	29
RPPGREAYLGDVFYLHSRLLERA	PWNTA	280	29

RPPGREAYPGDVFYLHSRLLERA	PWRZA	280	29
RPPGREAYLGDVFYLHSRLLERA	ATPA$MAIZE	280	29
RPPGREAYPGDVFYLHSRLLERV	ATPA$PEA	280	29
RPPGREAYPGDVFYLHSRLLERA	ATPA$WHEAT	280	29
RIPSAVGYQPTLATDMGALQERI	ATPB$RHORU	270	30
RMPSAVGYQPTLAEEMGVLQERI	PWECB	261	29
RPLGREAFPGDVFYAHSRLLERA	SYNMTATPAA	287	29
RMPSAVGYQPTLATEMGAMQERI	ATPB$CYTLY	299	34
RPPGREAYPGDVFYLHSRLLERA	ATPA$SPIOL	280	29
RMPSAVGYQPTLAEEMGVLQERI	ATPB$VIBAL	267	29
EVPGRGGYPGYMYTDLATIYERA	MTPB$SULAC	270	30
EVPGRRGFPGYMYTDLSTIYERA	VAT2$NEUCR	291	30
EVPGRRGYPGYMYTDLATIYERA	VAT2$ARATH	298	30
EVPGRRGFPGYMYTDLATIYERA	VAT2$HUMAN	308	30
EMPAEEGFPSYLPSRLAEYYERA	MTPA$SULAC	348	29
EVPGRRGYPGYMYTDLSTIYERA	VAT2$YEAST	297	30

ATP6
17
motif 6

KGIRPAINVGLSVSRVG	ATPA$BOVIN	402	49
KGIRPAINVGLSVSRVG	ATPA$RAT	392	49
KGIRPAINVGLSVSRVG	ATPA$XENLA	394	49
KGIRPAINVGLSVSRVG	BOVATPSYN	402	49
KGIRPAINVGLSVSRVG	A30245	402	49
LGIYPAVDPLDSKSRLL	ATPB$YEAST	375	45
LGIYPAVDPLDSKSRLL	A24260	373	45
KGIYPAVDPLDSTSTML	PWLVB	357	45
KGIYPAVDPLDSTSTML	PWZMB	359	45
KGIYPAVDPLDSTSTML	PWBHB	359	45
KGIYPAVDPLDSTSTML	PWRZB	359	45
KGIYPAVDPLESTSTML	ATPB$CHLRE	359	45
KGIYPAVDPLDSTSTML	ATPB$IPOBA	356	45
KGIYPAVDPLDSTSTML	ATPB$PEA	359	45
KGIYPAVDPLDSTSTML	RICCPCTA	359	45
KGIYPAVDPLDSTSTML	WHTCPATPB	359	45
KGIRPAVNVGLSVSRVG	ATPA$RHORU	360	49
RGIRPAINVGLSVSRVG	ATP0$HELAN	362	49
RGIRPAINVGLSVSRVG	ATP0$MAIZE	362	49
RGIRPAINVGLSVSRVG	ATP0$NICPL	362	49
RGIRPAINVGLSVSRVG	ATP0$OENBI	362	49
RGIRPAINVGLSVSRVG	ATP0$ORYSA	362	49
RGIRPAINVGLSVSRVG	ATP0$PEA	362	49
RGIRPAINVGLSVSRVG	ATP0$WHEAT	362	49
RGIRPAINVGLSVSRVG	PEAMTF14	362	49
SGVRPAINAGLSVSRVG	ATPA$BACME	351	49
SGVRPAINAGLSVSRVG	ATPA$THEP3	351	49
LGIYPAVDPLDSTSRML	ATP2$MAIZE	416	45
LGIYPAVDPLDSTSRML	ATP2$NICPL	423	45
KGIYPAVDPLGSTSTML	ATPB$ANASP	338	45
KGIRPAINVGLSVSRVG	ATPA$YEAST	396	49
LGIYPAVDPLDSTSRIM	ATPB$BOVIN	392	45
LGIYPAVDPLDSTSRIM	ATPB$HUMAN	392	45
LGIYPAVDPLDSTSRIM	ATPB$RAT	392	45
LGIYPAVDPLDSTSRIM	HUMATPFIB	402	45
LGIYPAVDPLDSTSRIM	HUMATPSY2	392	45
MGIYPAVDPLASTSRAL	PWBSBM	338	45

MGIYPAVDPLVSTSRAL	ATPB$THEP3	338	45
KGIYPAVDPLDSTSTML	PWNTB	359	45
AGIRPAVNPGISVSRVG	PWECA	362	60
AGIRPAVNPGISVSRVG	ATPA$ECOLI	362	60
MGIYPAVDPLASTSRAL	BFIATPD	335	45
KGIYPAVDPLDSTSTML	ATPB$SYNP6	349	45
KGIYPAVDPLDSTSTML	PWSPB	359	45
LGIYPAVDPLDSTSRLM	ATPB$RHOBL	342	45
AGIRPAVNPGISVSRVG	ATPA$ANASP	353	49
LGIYPAVDPLDSTSRIM	PWBOB	342	45
LGIYPAVDPLDSTSRQL	!F1AB	329	45
AGVRPAVDPGISVSRVG	ATPA$VIBAL	362	60
QGIRPAVNTGLSVSRVG	ATPA$RHOBL	361	49
SGLRPAINVGISVSRVG	ATPA$SYNP6	352	49
LGIYPAVDPLESTSRIL	ATPB$BACFR	368	45
AGIRPAINVGISVSRVG	PWLVA	352	49
SGIRPAINVGISVSRVG	PWNTA	352	49
AGIRPAINVGISVSRVG	PWRZA	352	49
AGIRPAINVGISVSRVG	ATPA$MAIZE	352	49
AGIRPAINVGISVSRVG	ATPA$PEA	352	49
AGIRPAINVGISVSRVG	ATPA$WHEAT	352	49
LGIYPAVDPLDSTSRAL	ATPB$RHORU	338	45
LGIYPAVDPLDSTSRQL	PWECB	329	45
KGIRPAVNAGSSVSRVG	SYNMTATPAA	359	49
LGIYPAVDPLDSTSRIL	ATPB$CYTLY	367	45
AGIRPAINVGISVSRVG	ATPA$SPIOL	352	49
MGLYPAIDPLDSTSRML	ATPB$VIBAL	335	45
KGIYPPINVLMSLSRLM	MTPB$SULAC	340	47
RGIYPPINVLPSLSRLM	VAT2$NEUCR	361	47
RQIYPPINVLPSLSRLM	VAT2$ARATH	368	47
RQIYPPINVLPSLSRLM	VAT2$HUMAN	378	47
ARHYPAINWIQGFSAYV	MTPA$SULAC	423	52
KGIYPPINVLPSLSRLM	VAT2$YEAST	367	47

C.3 NAKATPASE

COMPOUND(9)
D.N. PERKINS 15/10/1991
E1-E2 SODIUM/POTASSIUM ATPASE

1. SHULL, G.E., LINGRELL, S.B. Molecular cloning of the rat stomach ATPase.
JOURNAL OF BIOLOGICAL CHEMISTRY 261 pp16788 (1986)

2.Sweadner, K.J., Isozymes of the Na^+/K^+-ATPase.
BIOCHIMICA ET BIOPHYSICA ACTA 988 pp185

3. WALKER, J.E., SRASTE, M., RUNSWICK, M.J., GAY, N.J. Distantly related sequences in the alpha and beta subunits of ATP synthase, myosin, kinases and other ATP requiring enzymes and a common nucleotide binding fold.
EMBO JOURNAL 1 pp945 (1982)

This compound feature describes the alpha chains of the E1-E2 sodium/potassium transporting ATPases which catalyse the hydrolysis of ATP coupled with the exchange of sodium and potassium ions across the plasma membrane. All of these proteins are located in the cell membrane and appear to consist of seven or eight transmembrane helices. The ion transport that these proteins mediate creates the electrochemical gradient which provides the energy for the active transport of various nutrients. Potassium transporting ATPases are also responsible for the production of acid in the stomach as protons and potassium ions are exchanged [1]. The Na-K ATPase consists of two subunits, alpha and beta [2]. The alpha chains contain the ATP binding site and are commonly referred to as the catalytic subunit.

Eight sequences were initially aligned and from this nine motifs were selected. Motif four corresponds to the phosphorylation site while motif five describes the ATP binding site [3]. The other seven motifs were derived from the putative transmembrane helices which were located using a consensus hydropathy plot of the alignment. Two iterations were required until convergence, at which point all the appropriate sequences in the OWL database were found to match with all nine features. One sequence, database code B27180 (a rat sodium/potassium transporting ATPase), was shown to match with only eight of the motifs. This sequence lacks the seventh probable trans-membrane helix adjacent to the C terminal (motif nine). In the four feature column two codes are found, !SPDOC and !SPDON. These two codes describe the C and N terminus of the sodium/potassium transporting ATPase from ovine kidney. !SPDOC matches with motifs one to four, while !SPDON shows motifs six to nine. This family of proteins is a subset of the E1-E2 cation transporting atpases, members of this super family were seen to match with the two features (motifs four and five) derived from the ATP binding domain and the phosphorylation site. Also shown to match with two features (motifs two and three) was the sequence JU0341 (rat intercellular adhesion molecule-1). This protein is not related to the E1-E2 atpases and can be considered as noise.

```
SUMMARY INFORMATION
-------------------
  21 codes involving  9 elements
   1 codes involving  8 elements
   0 codes involving  7 elements
   0 codes involving  6 elements
   0 codes involving  5 elements
   2 codes involving  4 elements
   0 codes involving  3 elements
  24 codes involving  2 elements
```

```
COMPOUND FEATURE INDEX
----------------------

 9|  21   21   21   21   21   21   21   21   21
 8|   1    1    1    1    1    1    1    1    0
 7|   0    0    0    0    0    0    0    0    0
 6|   0    0    0    0    0    0    0    0    0
 5|   0    0    0    0    0    0    0    0    0
 4|   1    1    1    1    0    1    1    1    1
 3|   0    0    0    0    0    0    0    0    0
 2|   0    1    1   23   23    0    0    0    0
--+------------------------------------------------
  |   1    2    3    4    5    6    7    8    9


True positives:
 ATN1$RAT         ATN1$HORSE       ATN3$PIG         ATN1$HUMAN
 ATN1$SHEEP       ATN3$HUMAN       HUMATPA23        HUMATPK14
 ATN1$PIG         A27180           ATN3$RAT         A34474
 ATN1$CHICK       ATN2$RAT         ATNA$TORCA       ATNA$DROME
 ATNA$ARTSA       HUMATPGG         ATHA$PIG         ATHA$HUMAN
 ATHA$RAT

True positives: codes involving eight elements
 B27180


 ATN1$RAT          SODIUM/POTASSIUM ATPASE ALPHA-1 CHAIN - Rat
 ATN1$HORSE        SODIUM/POTASSIUM ATPASE ALPHA-1 CHAIN - Horse
 ATN3$PIG          SODIUM/POTASSIUM ATPASE ALPHA-3 CHAIN - Pig
 ATN1$HUMAN        SODIUM/POTASSIUM ATPASE ALPHA-1 CHAIN - Human
 ATN1$SHEEP        SODIUM/POTASSIUM ATPASE ALPHA-1 CHAIN - Sheep
 ATN3$HUMAN        SODIUM/POTASSIUM ATPASE ALPHA-3 CHAIN - Human
 HUMATPA23         Na+,K+ -ATPase catalytic subunit - Homo sapiens
 HUMATPK14         LOCUS HUMATPK14 1047 bp - Homo sapiens
 ATN1$PIG          SODIUM/POTASSIUM ATPASE ALPHA-1 CHAIN - Pig
 A27180            Na+/K+-transporting ATPase alpha-1 chain - Rat
 ATN3$RAT          SODIUM/POTASSIUM ATPASE ALPHA-3 CHAIN - Rat
 A34474            Na+/K+-transporting ATPase alpha chain - Human
 ATN1$CHICK        SODIUM/POTASSIUM ATPASE ALPHA-1 CHAIN - Chicken
 ATN2$RAT          SODIUM/POTASSIUM ATPASE ALPHA-2 CHAIN - Rat
 ATNA$TORCA        SODIUM/POTASSIUM ATPASE ALPHA - Electric Ray
 ATNA$DROME        SODIUM/POTASSIUM ATPASE ALPHA CHAIN - Fruit Fly
 ATNA$ARTSA        SODIUM/POTASSIUM ATPASE ALPHA - Brine shrimp
 HUMATPGG          HUMATPGG (H+ + K+)-ATPase - Homo sapiens
 ATHA$PIG          POTASSIUM ATPASE ALPHA CHAIN (GASTRIC) - Pig
 ATHA$HUMAN        POTASSIUM ATPASE ALPHA CHAIN (GASTRIC) - Human
 ATHA$RAT          POTASSIUM ATPASE ALPHA CHAIN (GASTRIC) - Rat
 B27180            Na+/K+-transporting ATPase alpha-2 chain - Rat


 SCAN HISTORY
 ------------
 OWL12_1    2   50 NSINGLE

 INITIAL MOTIF-SETS
 ------------------
```

ATPASE1
15
motif 1

LLWIGAILCFLAYGI	ATN3$HUMAN	93	93
LLWIGALLCFLAYGI	ATN2$RAT	101	101
LLWIGAVLCFLAYGI	PWSHNA	101	101
LLWTGAILCFLAYGI	ATNA$TORCA	103	103
LLWIGSLLCFLAYGI	ATN1$CHICK	101	101
LLWIGAILCFVAYSI	ATNA$DROME	119	119
LLWIGSILCFIAYTM	ATNA$ARTSA	80	80
LMWVAAAICLIAFAI	ATHA$PIG	113	113

ATPASE2
21
motif 2

LYLGIVLAAVVIITGCFSYYQ	ATN3$HUMAN	120	12
LYLGIVLAAVVIVTGCFSYYQ	ATN2$RAT	128	12
LYLGVVLSAVVIITGCFSYYQ	PWSHNA	128	12
LYLGVVLSTVVIITGCFSYYQ	ATNA$TORCA	130	12
LYLGVVLAAVVIITGCFSYYQ	ATN1$CHICK	128	12
LYLGIVLSAVVIVTGVFSYYQ	ATNA$DROME	146	12
LYLGLALLFVVIMTGCFAYYQ	ATNA$ARTSA	107	12
LYLALALIAVVVVTGCFGYYQ	ATHA$PIG	140	12

ATPASE3
23
motif 2

LITGVAVFLGVSFFILSLILGYT	ATN3$HUMAN	284	143
LITGVAVFLGVSFFVLSLILGYS	ATN2$RAT	292	143
IITGVAVFLGVSFFILSLILEYT	PWSHNA	292	143
IITGVAVFLGVSFFILSLILGYT	ATNA$TORCA	294	143
LITGVAVFLGVSFFILSLILEYT	ATN1$CHICK	292	143
LITGVAVFLGVTFFVIAFILGYH	ATNA$DROME	309	142
IITAMAVSLAAVFAVISFLYGYT	ATNA$ARTSA	271	143
IIAGLAILFGATFFIVAMCIGYT	ATHA$PIG	304	143

ATPASE4
22
motif 4

LGSTSTICSDKTGTLTQNRMTV	ATN3$HUMAN	357	50
LGSTSTICSDKTGTLTQNRMTV	ATN2$RAT	365	50
LGSTSTICSDKTGTLTQNRMTV	PWSHNA	365	50
LGSTSTICSDKTGTLTQNRMTV	ATNA$TORCA	367	50
LGSTSTICSDKTGTLTQNRMTV	ATN1$CHICK	365	50
LGSTSTICSDKTGTLTQNRMTV	ATNA$DROME	382	50
LGSTSTICSDKTGTLTQNRMTV	ATNA$ARTSA	344	50
LGSTSVICSDKTGTLTQNRMTV	ATHA$PIG	377	50

ATPASE5
19
motif 5

LVMKGAPERILDRCSTILL	ATN3$HUMAN	495	116
LVMKGAPERILDRCSTILV	ATN2$RAT	502	115
LVMKGAPERILDRCSSILI	PWSHNA	503	116
LVMKGAPERILDRCSTILL	ATNA$TORCA	504	115
LVMKGAPERILDRCDSILI	ATN1$CHICK	503	116
LVMKGAPERILERCSTIFI	ATNA$DROME	520	116

```
LVMKGAPERILERCSTILI                 ATNA$ARTSA    480    114
LVMKGAPERVLERCSSILI                  ATHA$PIG     515    116

ATPASE6
22
motif 6
ITPFLLFIMANIPLPLGTITIL              ATN3$HUMAN    777    263
ITPFLLFIIANIPLPLGTVTIL                ATN2$RAT    784    263
ITPFLIFIIANIPLPLGTVTIL                  PWSHNA    785    263
ITPFLVFIIANVPLPLGTVTIL              ATNA$TORCA    786    263
ITPFLIFIIANIPLPLGTCTIL              ATN1$CHICK    785    263
ISPFLASILCDIPLPLGTVTIL              ATNA$DROME    802    263
LSPFLMYILFDLPLAIGTVTIL              ATNA$ARTSA    762    263
LTPYLIYITVSVPLPLGCITIL                ATHA$PIG    797    263

ATPASE7
21
motif 7
YGQIGMIQALGGFFSYFVILA               ATN3$HUMAN    844     45
YGQIGMIQALGGFFTYFVILA                 ATN2$RAT    851     45
YGQIGMIQALGGFFTYFVIMA                   PWSHNA    852     45
YGQIGMIQALGGFFSYFVILA               ATNA$TORCA    853     45
YGQIGMIQALGGFFTYFVIMA               ATN1$CHICK    852     45
YGQIGMIQAAAGFFVYFVIMA               ATNA$DROME    869     45
YGQIGVMQAFGGFFTYFVIMG               ATNA$ARTSA    827     43
YFQIGAIQSFAGFTDYFTAMA                 ATHA$PIG    864     45

ATPASE8
21
motif 8
FTCHTAFFVSIVVVQWADLII               ATN3$HUMAN    906     41
FTCHTAFFASIVVVQWADLII                 ATN2$RAT    913     41
FTCHTAFFVSIVVVQWADLVI                   PWSHNA    914     41
YTCHTSFFVSIVIVQWADLII               ATNA$TORCA    915     41
FTCHTAFFVSIVVVQWADLII               ATN1$CHICK    914     41
YTCHTAFFISIVVVQWADLII               ATNA$DROME    931     41
YTCHTAFFISIVIVQWTDLII               ATNA$ARTSA    889     41
YTCYTVFFISIEMCQIADVLI                 ATHA$PIG    926     41

ATPASE9
25
motif 9
KNKILIFGLFEETALAAFLSYCPGM           ATN3$HUMAN    940     13
KNKILIFGLLEETALAAFLSYCPGM             ATN2$RAT    947     13
KNKILIFGLFEETALAAFLSYCPGM               PWSHNA    948     13
KNKILIFGLFEETALAAFLSYTPGT           ATNA$TORCA    949     13
KNKILIFGLFEETALAAFLSYCPGM           ATN1$CHICK    948     13
RNWALNFGLVFETVLAAFLSYCPGM           ATNA$DROME    965     13
KNGTLNFALVFETCVAAFLSYTPGM           ATNA$ARTSA    923     13
RNRILVIAIVFQVCIGCFLCYCPGM             ATHA$PIG    961     14

FINAL MOTIF-SETS
----------------

ATPASE1
15
motif 1
LLWIGAILCFLAYGI                     ATN1$HORSE    101    101
```

LLWIGAILCFLAYGI	ATN1$PIG	101	101
LLWIGAILCFLAYGI	ATN1$RAT	103	103
LLWIGAILCFLAYGI	ATN3$HUMAN	93	93
LLWIGAILCFLAYGI	ATN3$PIG	101	101
LLWIGAILCFLAYGI	ATN3$RAT	93	93
LLWIGAILCFLAYGI	HUMATPA23	93	93
LLWIGAILCFLAYGI	HUMATPK14	95	95
LLWIGAILCFLAYGI	A27180	103	103
LLWIGAILCFLAYGI	A34474	101	101
LLWIGAILCFLAYSI	ATN1$HUMAN	103	103
LLWIGALLCFLAYGI	ATN2$RAT	101	101
LLWIGAVLCFLAYGI	PWSHNA	101	101
LLWTGAILCFLAYGI	ATNA$TORCA	103	103
LLWIGAILCFVAYSI	ATNA$DROME	119	119
LLWIGSLLCFLAYGI	ATN1$CHICK	101	101
LLWIGSILCFIAYTM	ATNA$ARTSA	80	80
LMWVAAAICLIAFAI	ATHA$PIG	113	113
LMWVAAAICLIAFAI	ATHA$RAT	112	112
LMWVAAAICLIAFAI	A35292	114	114

ATPASE2
21
motif 2

LYLGVVLSAVVIITGCFSYYQ	ATN1$HORSE	128	12
LYLGVVLSAVVIITGCFSYYQ	ATN1$PIG	128	12
LYLGVVLSAVVIITGCFSYYQ	ATN1$RAT	130	12
LYLGIVLAAVVIITGCFSYYQ	ATN3$HUMAN	120	12
LYLGVVLSAVVIITGCFSYYQ	ATN3$PIG	128	12
LYLGIVLAAVVIITGCFSYYQ	ATN3$RAT	120	12
LYLGIVLAAVVIITGCFSYYQ	HUMATPA23	120	12
LYLGIVLAAVVIITGCFSYYQ	HUMATPK14	122	12
LYLGVVLSAVVIITGCFSVVQ	A27180	130	12
LYLGVVLAAVVIVTGCFSYYQ	A34474	128	12
LYLGVVLSAVVIITGCFSYYQ	ATN1$HUMAN	130	12
LYLGIVLAAVVIVTGCFSYYQ	ATN2$RAT	128	12
LYLGVVLSAVVIITGCFSYYQ	PWSHNA	128	12
LYLGVVLSTVVIITGCFSYYQ	ATNA$TORCA	130	12
LYLGIVLSAVVIVTGVFSYYQ	ATNA$DROME	146	12
LYLGVVLAAVVIITGCFSYYQ	ATN1$CHICK	128	12
LYLGLALLFVVIMTGCFAYYQ	ATNA$ARTSA	107	12
LYLALALIAVVVVTGCFGYYQ	ATHA$PIG	140	12
LYLALALIAVVVVTGCFGYYQ	ATHA$RAT	139	12
LYLAIALIAVVVVTGCFGYYQ	A35292	141	12

ATPASE3
23
motif 3

IITGVAVFLGVTFFILSLILEYT	ATN1$HORSE	292	143
IITGVAVFLGVSFFILSLILEYT	ATN1$PIG	292	143
LITGVAVFLGVSFFILSLILEYT	ATN1$RAT	294	143
LITGVAVFLGVSFFILSLILGYT	ATN3$HUMAN	284	143
IITGVAVFLGVSFFILSLILEYT	ATN3$PIG	292	143
LITGVAVFLGVSFFILSLILGYT	ATN3$RAT	284	143
LITGVAVFLGVSFFILSLILGYT	HUMATPA23	284	143
LITGVAVFLGVSFFILSLILGYT	HUMATPK14	286	143
LITGVAVFLGVSFFILSLILEYT	A27180	294	143
LITGVAVFLGVSFFVLSLILGYS	A34474	292	143

IITGVAVFLGVSFFILSLILEYT	ATN1$HUMAN	294	143
LITGVAVFLGVSFFVLSLILGYS	ATN2$RAT	292	143
IITGVAVFLGVSFFILSLILEYT	PWSHNA	292	143
IITGVAVFLGVSFFILSLILGYT	ATNA$TORCA	294	143
LITGVAVFLGVTFFVIAFILGYH	ATNA$DROME	309	142
LITGVAVFLGVSFFILSLILEYT	ATN1$CHICK	292	143
IITAMAVSLAAVFAVISFLYGYT	ATNA$ARTSA	271	143
IIAGLAILFGATFFIVAMCIGYT	ATHA$PIG	304	143
IIAGLAILFGATFFVVAMCIGYT	ATHA$RAT	303	143
IIAGLAILFGATFFIVAMCIGYT	A35292	305	143

ATPASE4
22
motif 4

LGSTSTICSDKTGTLTQNRMTV	ATN1$HORSE	365	50
LGSTSTICSDKTGTLTQNRMTV	ATN1$PIG	365	50
LGSTSTICSDKTGTLTQNRMTV	ATN1$RAT	367	50
LGSTSTICSDKTGTLTQNRMTV	ATN3$HUMAN	357	50
LGSTSTICSDKTGTLTQNRMTV	ATN3$PIG	365	50
LGSTSTICSDKTGTLTQNRMTV	ATN3$RAT	357	50
LGSTSTICSDKTGTLTQNRMTV	HUMATPA23	357	50
LGSTSTICSDKTGTLTQNRMTV	HUMATPK14	359	50
LGSTSTICSDKTGTLTQNRMTV	A27180	367	50
LGSTSTICSDKTGTLTQNRMTV	A34474	365	50
LGSTSTICSDKTGTLTQNRMTV	ATN1$HUMAN	367	50
LGSTSTICSDKTGTLTQNRMTV	ATN2$RAT	365	50
LGSTSTICSDKTGTLTQNRMTV	PWSHNA	365	50
LGSTSTICSDKTGTLTQNRMTV	ATNA$TORCA	367	50
LGSTSTICSDKTGTLTQNRMTV	ATNA$DROME	382	50
LGSTSTICSDKTGTLTQNRMTV	ATN1$CHICK	365	50
LGSTSTICSDKTGTLTQNRMTV	ATNA$ARTSA	344	50
LGSTSVICSDKTGTLTQNRMTV	ATHA$PIG	377	50
LGSTSVICSDKTGTLTQNRMTV	ATHA$RAT	376	50
LGSTSVICSDKTGTLTQNRMTV	A35292	378	50

ATPASE5
19
motif 5

LVMKGAPERILDRCSSILL	ATN1$HORSE	503	116
LVMKGAPERILDRCSSILI	ATN1$PIG	503	116
LVMKGAPERILDRCSSILL	ATN1$RAT	505	116
LVMKGAPERILDRCSTILL	ATN3$HUMAN	495	116
LVMKGAPERILDRCTSILI	ATN3$PIG	503	116
LVMKGAPERILDRCATILL	ATN3$RAT	495	116
LVMKGAPERILDRCSTILL	HUMATPA23	495	116
LVMKGAPERILDRCSTILL	HUMATPK14	497	116
LVMKGAPERILDRCSSILL	A27180	505	116
LVMKGAPERILDRCSTILV	A34474	502	115
LVMKGAPERILDRCSSILL	ATN1$HUMAN	505	116
LVMKGAPERILDRCSTILV	ATN2$RAT	502	115
LVMKGAPERILDRCSSILI	PWSHNA	503	116
LVMKGAPERILDRCSTILL	ATNA$TORCA	504	115
LVMKGAPERILERCSTIFI	ATNA$DROME	520	116
LVMKGAPERILDRCDSILI	ATN1$CHICK	503	116
LVMKGAPERILERCSTILI	ATNA$ARTSA	480	114
LVMKGAPERVLERCSSILI	ATHA$PIG	515	116

LVMKGAPERVLERCSSILI	ATHA$RAT	514	116
IVMKGAPERVLERCSSIII	A35292	516	116

ATPASE6
22
motif 6

ITPFLIFIIANIPLPLGTVTIL	ATN1$HORSE	785	263
ITPFLIFIIANIPLPLGTVTIL	ATN1$PIG	785	263
ITPFLIFIIANIPLPLGTVTIL	ATN1$RAT	787	263
ITPFLLFIMANIPLPLGTITIL	ATN3$HUMAN	777	263
ITPFLIFIIANIPLPLGTVTIL	ATN3$PIG	785	263
ITPFLLFIMANIPLPLGTITIL	ATN3$RAT	777	263
ITPFLLFIMANIPLPLGTITIL	HUMATPA23	777	263
ITPFLLFIMANIPLPLGTITIL	HUMATPK14	779	263
ITPFLIFIIANIPLPLGTVTIL	A27180	787	263
ITPFLLFIIANIPLPLGTVTIL	A34474	784	263
ITPFLIFIIANIPLPLGTVTIL	ATN1$HUMAN	787	263
ITPFLLFIIANIPLPLGTVTIL	ATN2$RAT	784	263
ITPFLIFIIANIPLPLGTVTIL	PWSHNA	785	263
ITPFLVFIIANVPLPLGTVTIL	ATNA$TORCA	786	263
ISPFLASILCDIPLPLGTVTIL	ATNA$DROME	802	263
ITPFLIFIIANIPLPLGTCTIL	ATN1$CHICK	785	263
LSPFLMYILFDLPLAIGTVTIL	ATNA$ARTSA	762	263
LTPYLIYITVSVPLPLGCITIL	ATHA$PIG	797	263
LTPYLIYITVSVPLPLGCITIL	ATHA$RAT	796	263
LTPYLIYITVSVPLPLGCITIL	A35292	798	263

ATPASE7
21
motif 7

YGQIGMIQALGGFFTYFVILA	ATN1$HORSE	852	45
YGQIGMIQALGGFFTYFVILA	ATN1$PIG	852	45
YGQIGMIQALGGFFTYFVILA	ATN1$RAT	854	45
YGQIGMIQALGGFFSYFVILA	ATN3$HUMAN	844	45
YGQIGMIQALGGFFTYFVILA	ATN3$PIG	852	45
YGQIGMIQALGGFFSYFVILA	ATN3$RAT	844	45
YGQIGMIQALGGFFSYFVILA	HUMATPA23	844	45
YGQIGMIQALGGFFSYFVILA	HUMATPK14	846	45
YGQIGMIQALGGFFTYFVILA	A27180	854	45
YGQIGMIQALGGFFTYFVILA	A34474	851	45
YGQIGMIQALGGFFTYFVILA	ATN1$HUMAN	854	45
YGQIGMIQALGGFFTYFVILA	ATN2$RAT	851	45
YGQIGMIQALGGFFTYFVIMA	PWSHNA	852	45
YGQIGMIQALGGFFSYFVILA	ATNA$TORCA	853	45
YGQIGMIQAAAGFFVYFVIMA	ATNA$DROME	869	45
YGQIGMIQALGGFFTYFVIMA	ATN1$CHICK	852	45
YGQIGVMQAFGGFFTYFVIMG	ATNA$ARTSA	827	43
YFQIGAIQSFAGFTDYFTAMA	ATHA$PIG	864	45
YFQIGAIQSFAGFADYFTAMA	ATHA$RAT	863	45
YFQIGAIQSFAGFTDYFTAMA	A35292	865	45

ATPASE8
21
motif 8

FTCHTAFFVSIVVVQWADLVI	ATN1$HORSE	914	41
FTCHTPFFVTIVVVQWADLVI	ATN1$PIG	914	41
FTCHTAFFVSIVVVQWADLVI	ATN1$RAT	916	41

FTCHTAFFVSIVVVQWADLII	ATN3$HUMAN	906	41
FTCHTAFFVSIVVVQWADLVI	ATN3$PIG	914	41
FTFHTAFFVSIVVVQWADLII	ATN3$RAT	906	41
FTCHTAFFVSIVVVQWADLII	HUMATPA23	906	41
FTCHTAFFVSIVVVQWADLII	HUMATPK14	908	41
FTCHTAFFVSIVVVQWADLVI	A27180	916	41
FTCHTAFFASIVVVQWADLII	A34474	913	41
FTCHTAFFVSIVVVQWADLVI	ATN1$HUMAN	916	41
FTCHTAFFASIVVVQWADLII	ATN2$RAT	913	41
FTCHTAFFVSIVVVQWADLVI	PWSHNA	914	41
YTCHTSFFVSIVIVQWADLII	ATNA$TORCA	915	41
YTCHTAFFISIVVVQWADLII	ATNA$DROME	931	41
FTCHTAFFVSIVVVQWADLII	ATN1$CHICK	914	41
YTCHTAFFISIVIVQWTDLII	ATNA$ARTSA	889	41
YTCYTVFFISIEMCQIADVLI	ATHA$PIG	926	41
YTCYTVFFISIEMCQIADVLI	ATHA$RAT	925	41
YTCYTVFFISIEVCQIADVLI	A35292	927	41

ATPASE9
25
motif 9

KNKILIFGLFEETALAAFLSYCPGM	ATN1$HORSE	948	13
KNKILIFGLFEETALAAFLSYCPGM	ATN1$PIG	948	13
KNKILIFGLFEETALAAFLSYCPGM	ATN1$RAT	950	13
KNKILIFGLFEETALAAFLSYCPGM	ATN3$HUMAN	940	13
KNKILIFGLFEETALAAFLSYCPGM	ATN3$PIG	948	13
KNKILIFGLFEETALAAFLSYCPGM	ATN3$RAT	940	13
KNKILIFGLFEETALAAFLSYCPGM	HUMATPA23	940	13
KNKILIFGLFEETALAAFLSYCPGM	HUMATPK14	942	13
KNKILIFGLFEETALAAFLSYCPGM	A27180	950	13
KNKILIFGLLEETALAAFLSYCPGM	A34474	947	13
KNKILIFGLFEETALAAFLSYCPGM	ATN1$HUMAN	950	13
KNKILIFGLLEETALAAFLSYCPGM	ATN2$RAT	947	13
KNKILIFGLFEETALAAFLSYCPGM	PWSHNA	948	13
KNKILIFGLFEETALAAFLSYTPGT	ATNA$TORCA	949	13
RNWALNFGLVFETVLAAFLSYCPGM	ATNA$DROME	965	13
KNKILIFGLFEETALAAFLSYCPGM	ATN1$CHICK	948	13
KNGTLNFALVFETCVAAFLSYTPGM	ATNA$ARTSA	923	13
RNRILVIAIVFQVCIGCFLCYCPGM	ATHA$PIG	961	14
RNRILVIAIVFQVCIGCFLCYCPGM	ATHA$RAT	960	14
RNKILVIAIVFQVCIGCFLCYCPGM	A35292	962	14

C.4 ELONGATION
COMPOUND(5)
D.N. PERKINS 1/6/1991
ELONGATION FACTORS

1. LEBLANC, D.J., LEE, L.N., TITMAS, B.M.,SMITH, C.J., TENOVER, F.C. Nucleotide sequence analysis of tetracycline resistance gene tetO from Streptococcus mutans DLS.
JOURNAL OF BACTERIOLOGY 170 3618-3626 (1988)

2. DEVER, T.E., GLYNIAS, M.J., MERRICK, W.C., GTP binding domain: three consensus sequence elements with distinct spacing.
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCE USA 84 1814-1818 (1987)

3. BAULDAUF, S.L., MANHART, J.R., PALMER, J.D. Differrent fates of the chloroplast tufa gene following its transfer to the nucleus in Green algae.
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCE USA 87 pp5317-5321 (1990)

This family of proteins consists of elongation factors which promote the GTP-dependant binding of aminoacyl tRNA to the A site of ribosomes during protein biosynthesis and catalyse the translocation of the protein chain being synthesised from the A site of the ribosome to the P site. All of these proteins are relatively similar in the vicinity of the C-terminus and a composite discriminator from this region has been assembled. Other proteins which are highly homologous to elongation factors are also show to match with all five features including the nodulation Q protein from Rhizobium melioti, bacterial tetracycline resistance proteins [1] and the omnipotent supressor protein 2 from yeast

An alignment of twelve sequences was prepared and from this five motifs were selected. Motifs one, three, and five correspond to the three GTP binding concensus segments [2] while the other motifs were selected because of their high homology across the family. Convergence was reached after three iterations when all the GTP binding elongation factors present in the OWL database were shown to match with all the motifs. One elongation factor was found to match with only four of the features. This sequence, database code EFTU$COLOB, from Coleochaete orbicularis, is quite different from the other elongation factors and is probably no longer functional [3]. Also found to match with four features were initiation factors, all of these proteins lack motif two. A single initiation factor was found in the three feature column and two fragments of elongation factors were found to match with only two motifs (database codes EZEC195 and !EFAS). Of the other sequences in the two features column, twenty one are GTP binding proteins, one is an ATP binding protein and the other sequences are unrelated and constitute noise.

```
SUMMARY INFORMATION
-------------------
  68 codes involving  5 elements
   7 codes involving  4 elements
   1 codes involving  3 elements
  35 codes involving  2 elements
```

```
COMPOUND FEATURE INDEX
---------------------

 5|  68   68   68   68   68
 4|   7    0    7    7    7
 3|   1    0    1    1    0
 2|   9   19   13   13   16
--+--------------------------
  |   1    2    3    4    5
```

```
True positives:
 EFSS1A          EF12$DROME      EF1A$ARTSA      EF1A$APIME
 EF1A$DICDI      EF1A$HUMAN      A32684          EFBY1A
 EF1A$CANAL      EF10$XENLA      EF11$DROME      EF11$RHIRA
 EF12$RHIRA      EF13$RHIRA      S08058          EF12$XENLA
 EF13$XENLA      XELEF1ALA       EF11$XENLA      EF1A$THECE
 EF1A$MOUSE      EF1A$ARATH      EFTU$HALMA      SHREF1A5
 EF1A$LYCES      EF1A$EUGGR      EFTU$ASTLO      EF1A$SULAC
 EFTU$METVA      EFTU$ARATH      EFTU$CYAPA      EFEGT
 EFTU$ANANI      EFTU$CHLRE      EFTU$SPIPL      EFTU$THEMA
 EFECT           EFTU$THETH      TTHTUF          EFTU$MICLU
 EFBYT           EFTU$MYCGE      1ETU            EFTU$MYCGA
 EFG$ECOLI       A28513          TETM$STRFA      EFG$ANANI
 EFG$MICLU       EFG$THETH       TETO$CAMJE      STATETOSM
 TETM$UREUR      STATETM         EF2$DICDI       EFG$SPIPL
 EF2$DROME       EF2$HALHA       EF2$CRIGR       EF2$HUMAN
 EF2$MESAU       EF2$RAT         EF2$METVA       MUSELF2PSA
 BVECLA          SUP2$YEAST      NODQ$RHIME      EFECSB

True positives: codes showing seven motifs
 EFTU$COLOB
```

EFSS1A	Elongation factor 1 alpha chain - Brine shrimp
EF12$DROME	ELONGATION FACTOR 1-ALPHA - Fruit fly
EF1A$ARTSA	ELONGATION FACTOR 1-ALPHA - Brine shrimp
EF1A$APIME	ELONGATION FACTOR 1-ALPHA - Honeybee
EF1A$DICDI	ELONGATION FACTOR 1-ALPHA - Slime mold
EF1A$HUMAN	ELONGATION FACTOR 1-ALPHA - Human
A32684	Elongation factor 1 alpha chain - Rabbit
EFBY1A	Elongation factor 1-alpha A - Yeast
EF1A$CANAL	ELONGATION FACTOR 1-ALPHA - Yeast
EF10$XENLA	ELONGATION FACTOR 1-ALPHA - African clawed frog
EF11$DROME	ELONGATION FACTOR 1-ALPHA - Fruit fly
EF11$RHIRA	ELONGATION FACTOR 1-ALPHA - *Rhizomucor racemosus*
EF12$RHIRA	ELONGATION FACTOR 1-ALPHA - *Rhizomucor racemosus*
EF13$RHIRA	ELONGATION FACTOR 1-ALPHA - *Rhizomucor racemosus*
S08058	Elongation factor - *Mucor circinelloides*
EF12$XENLA	ELONGATION FACTOR 1-ALPHA - African clawed frog
EF13$XENLA	ELONGATION FACTOR 1-ALPHA - African clawed frog
XELEF1ALA	Elongation factor-1 alpha-chain - *Xenopus laevis*
EF11$XENLA	ELONGATION FACTOR 1-ALPHA - African clawed frog
EF1A$THECE	ELONGATION FACTOR 1-ALPHA - *Thermococcus celer*
EF1A$MOUSE	ELONGATION FACTOR 1-ALPHA - Mouse
EF1A$ARATH	ELONGATION FACTOR 1-ALPHA - Mouse-ear cress
EFTU$HALMA	ELONGATION FACTOR TU - *Halobacterium marismortui*
SHREF1A5	SHREF1A5 EF-1 alpha - *Artemia salina*
EF1A$LYCES	ELONGATION FACTOR 1-ALPHA - Tomato
EF1A$EUGGR	ELONGATION FACTOR 1-ALPHA - *Euglena gracilis*
EFTU$ASTLO	ELONGATION FACTOR TU - *Astasia longa*
EF1A$SULAC	ELONGATION FACTOR - *Sulfolobus acidocaldarius*
EFTU$METVA	ELONGATION FACTOR TU - *Methanococcus vannielii*
EFTU$ARATH	ELONGATION FACTOR TU - Mouse-ear cress
EFTU$CYAPA	ELONGATION FACTOR TU - *Cyanophora paradoxa*
EFEGT	Elongation factor - *Euglena gracilis* chloroplast
EFTU$ANANI	ELONGATION FACTOR TU - *Anacystis nidulans*
EFTU$CHLRE	ELONGATION FACTOR TU - *Chlamydomonas reinhardtii*

```
EFTU$SPIPL          ELONGATION FACTOR TU - Spirulina platensis
EFTU$THEMA          ELONGATION FACTOR TU - Thermotoga maritima
EFECT               Elongation factors Tu - Escherichia coli
EFTU$THETH          ELONGATION FACTOR TU - Thermus aquaticus
TTHTUF              elongation factor Tu - Thermus thermophilus
EFTU$MICLU          ELONGATION FACTOR TU - Micrococcus luteus
EFBYT               Elongation factor Tu, mitochondrial - Yeast
EFTU$MYCGE          ELONGATION FACTOR TU - Mycoplasma genitalium
1ETU                ELONGATION FACTOR TU - Escherichia coli
EFTU$MYCGA          ELONGATION FACTOR TU - Mycoplasma gallisepticum
EFG$ECOLI           ELONGATION FACTOR G  - Escherichia coli
A28513              Elongation factor G - Escherichia coli
TETM$STRFA          TETRACYCLINE RESISTANCE - Streptococcus faecalis
EFG$ANANI           ELONGATION FACTOR G (EF-G) - Anacystis nidulans
EFG$MICLU           ELONGATION FACTOR G (EF-G) - Micrococcus luteus
EFG$THETH           ELONGATION FACTOR G (EF-G) - Thermus aquaticus
TETO$CAMJE          TETRACYCLINE RESISTANCE - Campylobacter jejuni
STATETOSM           Tetracycline-resistance - Staphylococcus mutans
TETM$UREUR          TETRACYCLINE RESISTANCE - Ureaplasma urealyticum
STATETM             STATETM tetM - Staphylococcus aureus
EF2$DICDI           ELONGATION FACTOR 2 (EF-2) - Slime mold
EFG$SPIPL           ELONGATION FACTOR G (EF-G) - Spirulina platensis
EF2$DROME           ELONGATION FACTOR 2 (EF-2) - Fruit fly
EF2$HALHA           ELONGATION FACTOR 2 - Halobacterium halobium
EF2$CRIGR           ELONGATION FACTOR 2 (EF-2) - Chinese hamster
EF2$HUMAN           ELONGATION FACTOR 2 (EF-2) - Human
EF2$MESAU           ELONGATION FACTOR 2 (EF-2) - Golden hamster
EF2$RAT             ELONGATION FACTOR 2 (EF-2) - Rat
EF2$METVA           ELONGATION FACTOR 2 - Methanococcus vannielii
MUSELF2PSA          pseudo-elongation factor 2 - Mus musculus
BVECLA              lepA protein - Escherichia coli
SUP2$YEAST          OMNIPOTENT SUPPRESSOR PROTEIN - Yeast
NODQ$RHIME          NODULATION PROTEIN Q - Rhizobium meliloti
EFECSB              Elongation factor selB - Escherichia coli
EFTU$COLOB          ELONGATION FACTOR TU - Coleochaete orbicularis

SCAN HISTORY
------------
OWL11_0    3  260 NSINGLE
INITIAL MOTIF-SETS
------------------
ELONGATION1
14
motif 1
NIVVIGHVDSGKST                        EFSS1A      9     9
NIGTIGHVDHGKTT                    EFTU$CHLRE     14    14
NIVVIGHVDSGKST                    EF1A$ARTSA      8     8
NMSVIAHVDHGKST                     EF2$HUMAN     21    21
NMSVIAHVDHGKST                     EF2$MESAU     21    21
NVGTIGHVDHGKTT                          1ETU     13    13
SLVVIGHVDSGKST                    EF1A$EUGGR      9     9
NFSIIAHIDHGKST                        BVECLA      6     6
IIATAGHVDHGKTT                        EFECSB      2     2
NIGIAAHIDAGKTT                     EFG$SPIPL     12    12
NIGIAAHIDAGKTT                     EFG$THETH     14    14
SLIFMGHVDAGKST                    SUP2$YEAST    262   262
```

ELONGATION2
9
motif 2

ERGITIDIA	EFSS1A	68	45
ERGITIDIA	EF1A$ARTSA	67	45
ARGITINTA	EFTU$CHLRE	58	30
ERCITIKST	EF2$HUMAN	65	30
ERCITIKST	EF2$MESAU	65	30
AAGITINTS	1ETU	42	15
ERCITIDIA	EF1A$EUGGR	68	45
ERGINIKAQ	BVECLA	49	29
KRGMTIDLG	EFECSB	33	17
ERGITITAA	EFG$SPIPL	58	32
ERGITITAA	EFG$THETH	60	32
NDGKTIEVG	SUP2$YEAST	321	45

ELONGATION3
11
motif 3

TIIDAPGHRDF	EFSS1A	88	11
TIIDAPGHRDF	EF1A$ARTSA	87	11
AHVDCPGHADY	EFTU$CHLRE	78	11
NLIDSPGHVDF	EF2$HUMAN	101	27
NLIDSPGHVDF	EF2$MESAU	101	27
AHVDCPGHADY	1ETU	62	11
TIIDAPGHRDF	EF1A$EUGGR	88	11
NFIDTPGHVDF	BVECLA	74	16
GFIDVPGHEKF	EFECSB	54	12
NIIDTPGHVDF	EFG$SPIPL	78	11
NIIDTPGHVDF	EFG$THETH	80	11
TILDAPGHKMY	SUP2$YEAST	341	11

ELONGATION4
12
motif 4

TGTSQADCAVLI	EFSS1A	104	5
TGTSQADCAVLI	EF1A$ARTSA	103	5
TGAAQMDGAILV	EFTU$CHLRE	94	5
AALRVTDGALVV	EF2$HUMAN	117	5
AALRVTDGALVV	EF2$MESAU	117	5
TGAAQMDGAILV	1ETU	78	5
TGTSQADAAVLV	EF1A$EUGGR	104	5
RSLAACEGALLV	BVECLA	90	5
AGVGGIDHALLV	EFECSB	70	5
RSMRVLDGVIAV	EFG$SPIPL	94	5
RSMRVLDGAIVV	EFG$THETH	96	5
GGASQADVGVLV	SUP2$YEAST	357	5

ELONGATION5
10
motif 5

LIVGVNKMDS	EFSS1A	148	32
LIVGVNKMDS	EF1A$ARTSA	147	32
VVVFLNKEDQ	EFTU$CHLRE	131	25

```
PVLMMNKMDR                           EF2$HUMAN     153      24
PVLMMNKMDR                           EF2$MESAU     153      24
IIVFLNKCDM                                1ETU     115      25
MIVATNKFDD                         EF1A$EUGGR     148      32
VVPVLNKIDL                              BVECLA     126      24
LTVALTKADR                              EFECSB     107      25
RIAFINKMDR                           EFG$SPIPL     130      24
RIAFANKMDK                           EFG$THETH     132      24
MVVVVNKMDD                         SUP2$YEAST     401      32

FINAL MOTIF-SETS
----------------
ELONGATION1
14
motif 1
NIVVIGHVDSGKST                          EFSS1A       9       9
NIVVIGHVDSGKST                     EF10$XENLA       9       9
NIVVIGHVDSGKST                     EF11$DROME       9       9
NIVVIGHVDSGKST                     EF12$DROME       9       9
NIVVIGHVDSGKST                     EF12$XENLA       9       9
NIVVIGHVDSGKST                     EF13$XENLA       2       2
NIVVIGHVDSGKST                     EF1A$APIME       9       9
NIVVIGHVDSGKST                     EF1A$ARATH       9       9
NIVVIGHVDSGKST                     EF1A$ARTSA       8       8
NIVVIGHVDSGKST                     EF1A$HUMAN       9       9
NIVVIGHVDSGKST                     EF1A$MOUSE       9       9
NIVVIGHVDSGKST                        SHREF1A5       9       9
NIVVIGHVDSGKST                          A32684       4       4
NIVVIGHVDAGKST                     EF1A$DICDI      12      12
NIGTIGHVDHGKTT                           EFEGT      14      14
NIGTIGHVDHGKTT                           EFBYT      50      50
NIGTIGHVDHGKTT                     EFTU$ANANI      14      14
NIGTIGHVDHGKTT                     EFTU$ARATH      81      81
NIGTIGHVDHGKTT                     EFTU$ASTLO      14      14
NIGTIGHVDHGKTT                     EFTU$CHLRE      14      14
NIGTIGHVDHGKTT                     EFTU$CYAPA      14      14
NIGTIGHVDHGKTT                     EFTU$MICLU      14      14
NIGTIGHVDHGKTT                     EFTU$SPIPL      14      14
NIVIIGHVDSGKST                     EF11$XENLA      12      12
NIVFIGHVDHGKST                     EF1A$THECE       9       9
NVVVIGHVDSGKST                          EFBY1A       9       9
NVVVIGHVDSGKST                     EF11$RHIRA       9       9
NVVVIGHVDSGKST                     EF12$RHIRA       9       9
NVVVIGHVDSGKST                     EF13$RHIRA       9       9
NVVVIGHVDSGKST                     EF1A$CANAL       9       9
NVVVIGHVDSGKST                          S08058       9       9
NVGTIGHVDHGKTT                           EFECT      14      14
NVGTIGHVDHGKTT                     EFTU$THETH      14      14
NVGTIGHVDHGKTT                          TTHTUF      14      14
NVGTIGHVDHGKTT                            1ETU      13      13
NLIVIGHVDHGKST                     EF1A$SULAC       8       8
SIVVIGHVDSGKST                     EF1A$LYCES       9       9
NIGTIGHIDHGKTT                     EFTU$MYCGA      14      14
KIVVIGHVDSGKST                      XELEF1ALA       9       9
NLAIIGHVDHGKST                     EFTU$HALMA       7       7
NMSVIAHVDHGKST                      EF2$CRIGR      21      21
NMSVIAHVDHGKST                      EF2$DROME      21      21
```

NMSVIAHVDHGKST	EF2$HUMAN	21	21
NMSVIAHVDHGKST	EF2$MESAU	21	21
NMSVIAHVDHGKST	EF2$RAT	21	21
NVGTIGHIDHGKST	EFTU$THEMA	14	14
NMSVIAHVDHGKTT	EF2$DICDI	21	21
NVGTIGHIDHGKTT	EFTU$MYCGE	14	14
NVAFIGHVDAGKST	EFTU$METVA	9	9
NIGVLAHVDAGKTT	TETM$STRFA	5	5
NIGVLAHVDAGKTT	TETM$UREUR	5	5
NIGVLAHVDAGKTT	STATETM	5	5
SLVVIGHVDSGKST	EF1A$EUGGR	9	9
NIAIAAHVDHGKTT	EF2$HALHA	23	23
NLGILAHVDAGKTT	TETO$CAMJE	5	5
NLGILAHVDAGKTT	STATETOSM	5	5
NFSIIAHIDHGKST	BVECLA	6	6
IIATAGHVDHGKTT	EFECSB	2	2
NIGIAAHIDAGKTT	EFG$ANANI	12	12
NIGIAAHIDAGKTT	EFG$SPIPL	12	12
NIGIAAHIDAGKTT	EFG$THETH	14	14
NIGISAHIDAGKTT	EFG$ECOLI	11	11
NIGIMAHIDAGKTT	EFG$MICLU	10	10
NIGISAHIDAGKTT	A28513	11	11
SLIFMGHVDAGKST	SUP2$YEAST	262	262
EHVSHLHVDHGKST	MUSELF2PSA	46	46
NMGICAHIAHGKTT	EF2$METVA	23	23
RFITCGSVDDGKST	NODQ$RHIME	26	26

ELONGATION2
9
motif 2

ERGITIDIA	EFSS1A	68	45
ERGITIDIS	EF10$XENLA	68	45
ERGITIDIA	EF11$DROME	68	45
ERGITIDIA	EF12$DROME	68	45
ERGITIDIS	EF12$XENLA	68	45
ERGITIDIS	EF13$XENLA	61	45
ERGITIDIA	EF1A$APIME	68	45
ERGITIDIA	EF1A$ARATH	68	45
ERGITIDIA	EF1A$ARTSA	67	45
ERGITIDIS	EF1A$HUMAN	68	45
ERGITIDIS	EF1A$MOUSE	68	45
ERGITIDIA	SHREF1A5	68	45
ERGITIDIS	A32684	63	45
ERGITIDIA	EF1A$DICDI	71	45
ARGITINTA	EFEGT	58	30
ARGITISTA	EFBYT	94	30
ARGITINTA	EFTU$ANANI	58	30
ARGITINTA	EFTU$ARATH	125	30
ARGITINTA	EFTU$ASTLO	58	30
ARGITINTA	EFTU$CHLRE	58	30
ARGITINTA	EFTU$CYAPA	58	30
QRGITINIS	EFTU$MICLU	60	32
QRGITINTA	EFTU$SPIPL	58	30
ERGITIDIS	EF11$XENLA	71	45
ERGITIDVA	EF1A$THECE	66	43
ERGITIDIA	EFBY1A	68	45
ERGITIDIA	EF11$RHIRA	68	45

ERGITIDIA	EF12$RHIRA	68	45
ERGITIDIA	EF13$RHIRA	68	45
ERGITIDIA	EF1A$CANAL	68	45
ERGITIDIA	S08058	68	45
ARGITINTS	EFECT	58	30
ARGITINTA	EFTU$THETH	59	31
ARGITINTA	TTHTUF	59	31
AAGITINTS	1ETU	42	15
ERGVTINLS	EF1A$SULAC	67	45
ERGITIDIA	EF1A$LYCES	68	45
ARGITINTA	EFTU$MYCGA	58	30
ERGITIDIS	XELEF1ALA	68	45
ERGVTIDIA	EFTU$HALMA	66	45
ERCITIKST	EF2$CRIGR	65	30
ERCITIKST	EF2$DROME	65	30
ERCITIKST	EF2$HUMAN	65	30
ERCITIKST	EF2$MESAU	65	30
ERCITIKST	EF2$RAT	65	30
ARGITINIT	EFTU$THEMA	58	30
ERGITIKSS	EF2$DICDI	65	30
ARGITINSA	EFTU$MYCGE	58	30
ERGVTIDVA	EFTU$METVA	68	45
QRGITIQTA	TETM$STRFA	51	32
QRGITIQTG	TETM$UREUR	51	32
QRGITIQTG	STATETM	51	32
ERCITIDIA	EF1A$EUGGR	68	45
ERGITIDAA	EF2$HALHA	67	30
QRGITIQTA	TETO$CAMJE	51	32
QRGITIQTA	STATETOSM	51	32
ERGINIKAQ	BVECLA	49	29
KRGMTIDLG	EFECSB	33	17
ERGITITAA	EFG$ANANI	58	32
ERGITITAA	EFG$SPIPL	58	32
ERGITITAA	EFG$THETH	60	32
ERGITITSA	EFG$ECOLI	57	32
ERGITITSA	EFG$MICLU	56	32
ERGITITSA	A28513	57	32
NDGKTIEVG	SUP2$YEAST	321	45
ERCITIKST	MUSELF2PSA	90	30
ARGITIYAA	EF2$METVA	67	30
EQGITIDVA	NODQ$RHIME	87	47

ELONGATION3
11
motif 3

TIIDAPGHRDF	EFSS1A	88	11
TIIDAPGHRDF	EF10$XENLA	88	11
TIIDAPGHRDF	EF11$DROME	88	11
TIIDAPGHRDF	EF12$DROME	88	11
TIIDAPGHRDF	EF12$XENLA	88	11
TIIDAPGHRDF	EF13$XENLA	81	11
TIIDAPGHRDF	EF1A$APIME	88	11
TVIDAPGHRDF	EF1A$ARATH	88	11
TIIDAPGHRDF	EF1A$ARTSA	87	11
TIIDAPGHRDF	EF1A$HUMAN	88	11
TIIESPGHRDF	EF1A$MOUSE	88	11
TIIDAPGHRDF	SHREF1A5	88	11

TIIDAPGHRDF	A32684	83	11
TIIDAPGHRDF	EF1A$DICDI	91	11
AHVDCPGHADY	EFEGT	78	11
SHVDCPGHADY	EFBYT	114	11
AHVDCPGHADY	EFTU$ANANI	78	11
AHVDCPGHADY	EFTU$ARATH	145	11
AHVDCPGHADY	EFTU$ASTLO	78	11
AHVDCPGHADY	EFTU$CHLRE	78	11
AHVDCPGHADY	EFTU$CYAPA	78	11
AHVDAPGHADY	EFTU$MICLU	80	11
AHVDCPGHADY	EFTU$SPIPL	78	11
TIIDAPGHRDF	EF11$XENLA	91	11
TIIDAPGHRDF	EF1A$THECE	86	11
TVIDAPGHRDF	EFBY1A	88	11
TVIDAPGHRDF	EF11$RHIRA	88	11
TVIDAPGHRDF	EF12$RHIRA	88	11
TVIDAPGHRDF	EF13$RHIRA	88	11
TVIDAPGHRDF	EF1A$CANAL	88	11
TVIDAPGHRDF	S08058	88	11
AHVDCPGHADY	EFECT	78	11
SHVDCPGHADY	EFTU$THETH	79	11
SHVDCPGHADY	TTHTUF	79	11
AHVDCPGHADY	1ETU	62	11
TVIDAPGHRDF	EF1A$SULAC	87	11
TVIDAPGHRDF	EF1A$LYCES	88	11
AHVDCPGHADY	EFTU$MYCGA	78	11
TIIDAPGHRDF	XELEF1ALA	88	11
TIVDCPGHRDF	EFTU$HALMA	86	11
NLIDSPGHVDF	EF2$CRIGR	101	27
NLIDSPGHVDF	EF2$DROME	105	31
NLIDSPGHVDF	EF2$HUMAN	101	27
NLIDSPGHVDF	EF2$MESAU	101	27
NLIDSPGHVDF	EF2$RAT	101	27
AHIDCPGHADY	EFTU$THEMA	78	11
NLIDSPGHVDF	EF2$DICDI	99	25
AHVDCPGHADY	EFTU$MYCGE	78	11
TIVDCPGHRDF	EFTU$METVA	88	11
NIIDTPGHMDF	TETM$STRFA	71	11
NIIDTPGHMDF	TETM$UREUR	71	11
NIIDTPGHMDF	STATETM	71	11
TIIDAPGHRDF	EF1A$EUGGR	88	11
NLIDTPGHVDF	EF2$HALHA	91	15
NIIDTPGHMDF	TETO$CAMJE	71	11
NIIDTPGHMDF	STATETOSM	71	11
NFIDTPGHVDF	BVECLA	74	16
GFIDVPGHEKF	EFECSB	54	12
NIIDTPGHVDF	EFG$ANANI	78	11
NIIDTPGHVDF	EFG$SPIPL	78	11
NIIDTPGHVDF	EFG$THETH	80	11
NIIDTPGHVDF	EFG$ECOLI	84	18
NIIDNPGHVDF	EFG$MICLU	76	11
NIIDTPGHVDF	A28513	84	18
TILDAPGHKMY	SUP2$YEAST	341	11
NLIDSPGHVDF	MUSELF2PSA	126	27
NLIDTPGHVDF	EF2$METVA	91	15
IVADTPGHEEY	NODQ$RHIME	107	11

ELONGATION4
12
motif 4

TGTSQADCAVLI	EFSS1A	104	5
TGTSQADCAVLI	EF10$XENLA	104	5
TGTSQADCAVQI	EF11$DROME	104	5
TGTSQADCAVLI	EF12$DROME	104	5
TGTSQADCAVLI	EF12$XENLA	104	5
TGTSQADCAVLI	EF13$XENLA	97	5
TGTSQADCAVLI	EF1A$APIME	104	5
TGTSQADCAVLI	EF1A$ARATH	104	5
TGTSQADCAVLI	EF1A$ARTSA	103	5
TGTSQADCAVLI	EF1A$HUMAN	104	5
TGTSQADCAVLI	EF1A$MOUSE	104	5
GTSQVADCAVLI	SHREF1A5	105	6
TGTSQADCAVLI	A32684	99	5
TGTSQADCAVLV	EF1A$DICDI	107	5
TGAAQMDGAILV	EFEGT	94	5
TGAAQMDGAIIV	EFBYT	130	5
TGAAQMDGAILV	EFTU$ANANI	94	5
TGAAQMDGAILV	EFTU$ARATH	161	5
TGAAQMDGAILV	EFTU$ASTLO	94	5
TGAAQMDGAILV	EFTU$CHLRE	94	5
TGAAQMDGAILV	EFTU$CYAPA	94	5
TGAAQMDGAILV	EFTU$MICLU	96	5
TGAAQMDGAILV	EFTU$SPIPL	94	5
TGTSQADVALLV	EF11$XENLA	107	5
TGASQADAAVLV	EF1A$THECE	102	5
TGTSQADCAILI	EFBY1A	104	5
TGTSQADCAILI	EF11$RHIRA	104	5
TGTSQADCAILI	EF12$RHIRA	104	5
TGTSQADCAILI	EF13$RHIRA	104	5
TGTSQADCAILI	EF1A$CANAL	104	5
TGTSQADCAILI	S08058	104	5
TGAAQMDGAILV	EFECT	94	5
TGAAQMDGAILV	EFTU$THETH	95	5
TGAAQMDGAILV	TTHTUF	95	5
TGAAQMDGAILV	1ETU	78	5
TGASQADAAILV	EF1A$SULAC	103	5
TGTSQADCAVLI	EF1A$LYCES	104	5
TGAAQMDGGILV	EFTU$MYCGA	94	5
TGTSQADCAVLI	XELEF1ALA	104	5
TGASQADNAVLV	EFTU$HALMA	102	5
AALRVTDGALVV	EF2$CRIGR	117	5
AALRVTDGALVV	EF2$DROME	121	5
AALRVTDGALVV	EF2$HUMAN	117	5
AALRVTDGALVV	EF2$MESAU	117	5
AALRVTDGALVV	EF2$RAT	117	5
TGAAQMDGAILV	EFTU$THEMA	94	5
AALRVTDGALVV	EF2$DICDI	115	5
TGAAQMDGAILV	EFTU$MYCGE	94	5
TGASQADAAVLV	EFTU$METVA	104	5
RSLSVLDGAILL	TETM$STRFA	87	5
RSLSVLDGAILL	TETM$UREUR	87	5
RSLSVLDGAILL	STATETM	87	5
TGTSQADAAVLV	EF1A$EUGGR	104	5
RAMRAVDGALVV	EF2$HALHA	107	5

RSLSVLDGAVLL	TETO$CAMJE	87	5
RSLSVLDGAVLL	STATETOSM	87	5
RSLAACEGALLV	BVECLA	90	5
AGVGGIDHALLV	EFECSB	70	5
RSMRVLDGVVAV	EFG$ANANI	94	5
RSMRVLDGVIAV	EFG$SPIPL	94	5
RSMRVLDGAIVV	EFG$THETH	96	5
RSMRVLDGAVMV	EFG$ECOLI	100	5
RSLRVLDGAVAV	EFG$MICLU	92	5
RSMRVLDGAVMV	A28513	100	5
GGASQADVGVLV	SUP2$YEAST	357	5
AALRVTDGALVV	MUSELF2PSA	142	5
RAMRAIDGAVVV	EF2$METVA	107	5
TGASTADLAIIL	NODQ$RHIME	123	5

ELONGATION5
10
motif 5

LIVGVNKMDS	EFSS1A	148	32
LIVGINKMDS	EF10$XENLA	148	32
LIVGVNKMDS	EF11$DROME	148	32
LIVGVNKMDS	EF12$DROME	148	32
LIIGVNKMDS	EF12$XENLA	148	32
LIIGVNKMDS	EF13$XENLA	141	32
LIVGVNKMDM	EF1A$APIME	148	32
MICCCNKMDA	EF1A$ARATH	148	32
LIVGVNKMDS	EF1A$ARTSA	147	32
LIVGVNKMDS	EF1A$HUMAN	148	32
LIVGVNKMDS	EF1A$MOUSE	148	32
LIVGVNKMDS	SHREF1A5	149	32
LIVGVNKMDS	A32684	138	27
MIVAINKMDE	EF1A$DICDI	151	32
IVVFLNKEDQ	EFEGT	131	25
IVVFVNKVDT	EFBYT	167	25
IVVFLNKEDM	EFTU$ANANI	131	25
MVVFLNKEDQ	EFTU$ARATH	198	25
LVVFLNKEDQ	EFTU$ASTLO	131	25
VVVFLNKEDQ	EFTU$CHLRE	131	25
MVVFLNKEDQ	EFTU$CYAPA	131	25
LLVALNKSDM	EFTU$MICLU	133	25
IVVFLNKADM	EFTU$SPIPL	131	25
LIVCVNKMDL	EF11$XENLA	151	32
ILVAVNKMDM	EF1A$THECE	139	25
LIVAVNKMDS	EFBY1A	148	32
LIVAINKMDT	EF11$RHIRA	148	32
LIVAINKMDT	EF12$RHIRA	148	32
LIVAINKMDT	EF13$RHIRA	148	32
LIVAVNKMDS	EF1A$CANAL	148	32
LIVAINKMDT	S08058	148	32
IIVFLNKCDM	EFECT	131	25
IVVFMNKVDM	EFTU$THETH	132	25
IVVFMNKVDM	TTHTUF	132	25
IIVFLNKCDM	1ETU	115	25
VIVAINKMDL	EF1A$SULAC	147	32
MICCCNKMDA	EF1A$LYCES	148	32
MVVFLNKCDV	EFTU$MYCGA	131	25
LIVGINKMDS	XELEF1ALA	148	32

LIVAVNKMDL	EFTU$HALMA	139	25
PVLMMNKMDR	EF2$CRIGR	153	24
PILFMNKMDR	EF2$DROME	157	24
PVLMMNKMDR	EF2$HUMAN	153	24
PVLMMNKMDR	EF2$MESAU	153	24
PVLMMNKMDR	EF2$RAT	153	24
MIVFINKTDM	EFTU$THEMA	131	25
PVLFVNKVDR	EF2$DICDI	151	24
MVVFLNKCDI	EFTU$MYCGE	131	25
LAVAVNKMDT	EFTU$METVA	144	28
TIFFINKIDQ	TETM$STRFA	123	24
TIFFINKIDQ	TETM$UREUR	123	24
TIFFINKIDQ	STATETM	123	24
MIVATNKFDD	EF1A$EUGGR	148	32
PTLFINKVDR	EF2$HALHA	143	24
TIFFINKIDQ	TETO$CAMJE	123	24
TIFFINKIDQ	STATETOSM	123	24
VVPVLNKIDL	BVECLA	126	24
LTVALTKADR	EFECSB	107	25
RIVFVNKMDR	EFG$ANANI	130	24
RIAFINKMDR	EFG$SPIPL	130	24
RIAFANKMDK	EFG$THETH	132	24
RIAFVNKMDR	EFG$ECOLI	136	24
RICFVNKMDK	EFG$MICLU	128	24
RIAFVNKMDR	A28513	136	24
MVVVVNKMDD	SUP2$YEAST	401	32
PVLMMNKMDR	MUSELF2PSA	178	24
PVLFINKVDR	EF2$METVA	143	24
VVLAVNKIDL	NODQ$RHIME	160	25

C.5 METHYL
COMPOUND(3)
D.N. PERKINS 10/4/1991
CYTOSINE SPECIFIC METHYL TRANSFERASE

1. WU, J.C., SANTI, D.U. Kinetic and catalytic mechanism of HhaI methyltransferase.
JOURNAL OF BIOLOGICAL CHEMISTRY 262 4778-4786 (1987)

2. SULLIVAN, K.M., SAUNDER, J.R. Sequence analysis of the Ngo PII methyltransferase gene from Neisseria gonorrhoeae; homologies with other enzymes recognising the sequence GGCC.
NUCLEIC ACIDS RESEARCH 16 4369 (1988)

3. POSFAI, J, BHAGWAT, A.S., ROBERTS, R.J. Sequence Motifs for Cytosine Methyltransferases.
GENE 74 261-265 (1988)

DNA (cytosine 5) methyltransferase catalyse the methylation of cystine residues in specific sequences of DNA to produce DNA (5-methyl) cytosine. In mammlian cells, cytosine specific methyltransferases methylate certain sequences which are believed to modulate gene expression and cell differentiation. In bacteria, these enzymes are a component of restriction modification systems and serve as valuable tool for the manipulation of DNA [1]. Homology between the C-5 methyltransferases has been noted by a number of workers [2].

An alignment of eleven sequences of was prepared as the initial step in this study from which three motifs were selected. It has been suggested that there are five well conserved regions within this family, each region containing invariate residues [3].The first conserved region (FxGxG) is described in motif one, although these residues are not completely invariate. The second conserved region described by Posfai et al,(GxPCxxxSxxxG), is in fact not conserved over the whole family and was found to be of little use for discrimination. Motif two was derived from the third conserved region which is suggested to have the three invariant residues (ENV), although again these positions are not completely conserved. The fourth conserved region was used as a basis for motif three, sequence CHVCYMV (from Chlorella virus) though differs as histidine replaces the supposably conserved glutamine. The fifth region suggested by Posfai et al is not conserved across the whole family and is of no use for discrimination. Two iterations were required until convergence was reached.

```
SUMMARY INFORMATION
-------------------
  26 codes involving  3 elements
   0 codes involving  2 elements

COMPOUND FEATURE INDEX
----------------------

 3|  26   26   26
 2|   0    0    0
--+---------------
  |   1    2    3
```

```
True positives:
CTBPSR         SPRMTASE       S02599         CTBPPT
ECOMASE        S02598         CTBPRH         MTH2$HAEPA
MTB2$BACSU     MTB1$BREEP     JS0489         XYBSR1
AQUMAB         MTB1$BACSH     MTNG$NEIGO     XYECR2
DCM$ECOLI      JS0102         MTD1$DESDN     XYHIH1
MTSI$SALIN     MTSA$STAAU     MTM1$MORSP     MTDM$MOUSE
MTSI$SPISQ     CHVCYMT
```

CTBPSR	Site-specific methyltransferase - Bacteriophage
SPRMTASE	DNA methyltransferase - Bacteriophage SPR
S02599	Site-specific methyltransferase - Bacteriophage
CTBPPT	Site-specific methyltransferase - Bacteriophage
ECOMASE	Mtase protein - Artificial gene
S02598	Site-specific methyltransferase - Bacteriophage
CTBPRH	Site-specific methyltransferase - Bacteriophage
MTH2$HAEPA	METHYLTRANSFERASE - *Haemophilus parainfluenzae*
MTB2$BACSU	MODIFICATION METHYLASE BSUF I - *Bacillus subtilis*
MTB1$BREEP	METHYLTRANSFERASE - *Brevibacterium epidemidis*
JS0489	banI methylase - *Bacillus aneurinolyticus*
XYBSR1	methyltransferase BsuRI - *Bacillus subtilis*
AQUMAB	M.AquI alpha protein - *Agmenellum quadruplicatum*

```
MTB1$BACSH        METHYLTRANSFERASE - Bacillus sphaericus
MTNG$NEIGO        METHYLTRANSFERASE - Neisseria gonorrhoeae
XYECR2            Site-specific methyltransferase EcoRII - E. coli
DCM$ECOLI         DNA-CYTOSINE METHYLTRANSFERASE - Escherichia coli
JS0102            methyltransferase - Haemophilus aegyptius
MTD1$DESDN        METHYLTRANSFERASE - Desulfovibrio desulfuricans
XYHIH1            methyltransferase - Haemophilus haemolyticus
MTSI$SALIN        METHYLTRANSFERASE - Salmonella infantis
MTSA$STAAU        METHYLTRANSFERASE - Staphylococcuc aureus
MTM1$MORSP        METHYLTRANSFERASE - Morexella sp.
MTDM$MOUSE        DNA (CYTOSINE-5)-METHYLTRANSFERASE - Mouse
MTSI$SPISQ        CPG DNA METHYLASE - Spiroplasma sp.
CHVCYMT           cytosine methyltransferase - Chlorella virus

SCAN HISTORY
------------
OWL10_1     2    50 NSINGLE

INITIAL MOTIF-SETS
------------------
METHYL1
17
motif 1
KVLSLFSGCGGMDLGLE                 MTB1$BREEP      2      2
NVLSLFSGCGGLDLGFE                     XYBSR1     60     60
KIISLFSGCGGLDLGFE                 MTNG$NEIGO     13     13
RVMSLFSGIGAFEAALR                     CTBPPT      5      5
RVMSLFSGIGAFEAALR                    ECOMASE      5      5
RFIDLFAGLGGFRLALE                     XYHIH1     13     13
KFIDLFSGIGGIRQSFE                 MTM1$MORSP    106    106
RTLDVFSGCGGLSEGFH                 MTDM$MOUSE   1021   1021
KALSFFSGAMGLDLGIE                 MTSI$SALIN     76     76
RTLELFAGIAGISHGLR                    CHVCYMT      4      4
RVFEAFAGIGAQRKALE                 MTSI$SPISQ     12     12

METHYL2
15
motif 2
KPKVFIAENVKGLVT                   MTB1$BREEP    161    142
QPEIFVAENVKGMMT                       XYBSR1    187    110
QPKFFLAENVSGMLA                   MTNG$NEIGO    115     85
KPKFVILENVKGLIN                       CTBPPT    142    120
KPKFVILENVKGLIN                      ECOMASE    142    120
KPKVVFMENVKNFAS                       XYHIH1    112     82
KTPVLFLENVPGLIN                   MTM1$MORSP    206     83
RPRFFLLKNVRNFVS                   MTDM$MOUSE   1135     97
RPKYIVIENVRGLLS                   MTSI$SALIN    185     92
KPKIVFLENSHMLSH                      CHVCYMT    104     83
LPKYLLMENVGATTH                   MTSI$SPISQ    179    150

METHYL3
14
motif 3
GVAQNRERVIFIGI                    MTB1$BREEP    208     32
GVPQLRERVIIEGV                        XYBSR1    233     31
GVAQERKRVFYIGF                    MTNG$NEIGO    161     31
NVPQNRERVYIIGI                        CTBPPT    188     31
NVPQNRERVYIIGI                      ECOMASE    188     31
GIPQKRERIYMICF                       XYHIH1    158     31
GIPQKRKRFYLVAF                  MTM1$MORSP     252     31
CVAQTRRRAIIILA                  MTDM$MOUSE    1181     31
GVPQIRERVIICS                   MTSI$SALIN     252     52
GAHHQRHRWFCLAI                      CHVCYMT    147     28
GSSQARRRVFMMST                  MTSI$SPISQ     225     31

FINAL MOTIF-SETS
----------------
METHYL1
17
motif 1
RVMSLFSGIGAFEAALR                    CTBPSR      5      5
RVMSLFSGIGAFEAALR                    CTBPRH      5      5
RVMSLFSGIGAFEAALR                    CTBPPT      5      5
RFIDLFAGIGGIRKGFE                    XYECR2     97     97
RVMSLFSGIGAFEAALR                  SPRMTASE      4      4
RVMSLFSGIGAFEAALR                   ECOMASE      5      5
RVMSLFSGIGAFEAALR                    S02598      5      5
RVMSLFSGIGAFEAALR                    S02599      5      5
RFIDLFAGIGGIRRGFE                DCM$ECOLI      88     88
NVLSLFSGCGGLDLGFE                    XYBSR1     60     60
KVLSLFSGCGGMDLGLE              MTB1$BREEP        2      2
TFIDLFAGIGGIRLGFE              MTB2$BACSU      102    102
KIISLFSGCGGLDLGFE              MTNG$NEIGO       13     13
RFIDLFAGLGGFRLALE                    XYHIH1     13     13
KFVDLFAGIGGIRIGFE                    JS0489      4      4
KFIDLFSGIGGIRQSFE              MTM1$MORSP      106    106
KVVELFAGVGGFRLGLE              MTSA$STAAU        5      5
NVLSLFCGAGGLDLGFE              MTB1$BACSH       59     59
NLISLFSGAGGLDLGFQ                    JS0102      2      2
KLISLFSGAGGMDIGFH                    AQUMAB      4      4
TFIDLFAGIGGFRIAMQ              MTH2$HAEPA       33     33
NIIDLFAGCGGFSHGFK              MTD1$DESDN        2      2
RTLELFAGIAGISHGLR                   CHVCYMT      4      4
RVFEAFAGIGAQRKALE              MTSI$SPISQ       12     12
RTLDVFSGCGGLSEGFH              MTDM$MOUSE     1021   1021
KALSFFSGAMGLDLGIE              MTSI$SALIN       76     76

METHYL2
15
motif 2
QPKFFVFENVKGLIN                      CTBPSR    109     87
QPRYFVFENVKGLIN                      CTBPRH    109     87
KPKFVILENVKGLIN                      CTBPPT    142    120
KPAIFVLENVKNLKS                      XYECR2    226    112
QPKFFVFENVKGLIN                    SPRMTASE    108     87
KPKFVILENVKGLIN                     ECOMASE    142    120
KPKFVILENVKGLIN                      S02598    109     87
QPKFFVFENVKGLIN                      S02599    109     87
RPAMFVLENVKNLKS                   DCM$ECOLI    217    112
QPEIFVAENVKGMMT                      XYBSR1    187    110
KPKVFIAENVKGLVT                  MTB1$BREEP    161    142
QPKMFLLENVKGLLT                  MTB2$BACSU    201     82
QPKFFLAENVSGMLA                  MTNG$NEIGO    115     85
KPKVVFMENVKNFAS                      XYHIH1    112     82
```

RPKAFLLENVRGLVT	JS0489	107	86
KTPVLFLENVPGLIN	MTM1$MORSP	206	83
FPKYLLLENVDRLLK	MTSA$STAAU	119	97
QPEIFVAENVKGMMT	MTB1$BACSH	186	110
KPIFFLAENVKGMMA	JS0102	102	83
LPKCFVMENVKGMIN	AQUMAB	113	92
QPKAFFLENVKGLKN	MTH2$HAEPA	134	84
SPKFFVMENVLGILS	MTD1$DESDN	106	87
KPKIVFLENSHMLSH	CHVCYMT	104	83
LPKYLLMENVGATTH	MTSI$SPISQ	179	150
RPRFFLLKNVRNFVS	MTDM$MOUSE	1135	97
RPKYIVIENVRGLLS	MTSI$SALIN	185	92

METHYL3
14
motif 3

NVPQNRERLYIIGI	CTBPSR	155	31
NVPQNRERIYIIGV	CTBPRH	155	31
NVPQNRERVYIIGI	CTBPPT	188	31
FLPQHRERIVLVGF	XYECR2	280	39
NVPQNRERLYIIGI	SPRMTASE	154	31
NVPQNRERVYIIGI	ECOMASE	188	31
NVPQNRERLYIIGI	S02598	155	31
NVPQNRERLYIIGI	S02599	155	31
FLPQHRERIVLVGF	DCM$ECOLI	271	39
GVPQLRERVIIEGV	XYBSR1	233	31
GVAQNRERVIFIGI	MTB1$BREEP	208	32
GLPQRRERIVIVGF	MTB2$BACSU	247	31
GVAQERKRVFYIGF	MTNG$NEIGO	161	31
GIPQKRERIYMICF	XYHIH1	158	31
GVPQNRVRIYILGI	JS0489	153	31
GIPQKRKRFYLVAF	MTM1$MORSP	252	31
GNAQRRRRVFIFGY	MTSA$STAAU	168	34
GVPQIRERVIIVGV	MTB1$BACSH	232	31
GVAQDRKRVFYIGF	JS0102	148	31
GVPQFRERVFIVGN	AQUMAB	167	39
GVPQNRERIYIVGF	MTH2$HAEPA	182	33
GVPQSRQRVFFIGL	MTD1$DESDN	155	34
GAHHQRHRWFCLAI	CHVCYMT	147	28
GSSQARRRVFMMST	MTSI$SPISQ	225	31
CVAQTRRRAIIILA	MTDM$MOUSE	1181	31
GVPQIRERVIIICS	MTSI$SALIN	252	52

C.6 FERREDOXIN

COMPOUND(3)
D.N.PERKINS, 10-APRIL-1991
FERREDOXIN .

1. VORST, O., VAN DAM, F., OOSTERHOFF-TEERTSTRA, R., SMEEKENS, S. and WEISBECK, P.
Tissue specific expression directed by an Arabidopsis thaliana pre-ferredoxin promoter in transgenic tobacco plants.
PLANT MOLECULAR BIOLOGY 14 491-499 (1990).

2. DUTTON, J.E., LYNDON, J.R., HASLETT, B.G., TAKRURI, I.A.H., GLEAVES, J.T. and BOULTER, D.
Comparitive studies on the properties of two ferredoxins from Pisium sativum.
JOURNAL OF EXPERIMENTAL BOTANY 31 379-391 (1980).

3. MASUI, R., WADA, K., MATSUBARU, H., WILLIAMS, M.M. and ROGERS, L.J. Characterisation, amino acid sequence and phylogenetic considerations regarding the ferrdoxin from Ochromonas danica.
PHYTOCHEMISTRY 27 2817-2820 (1988).

Ferredoxin is a low molecular weight iron-sulphur protein which is present in all photosynthetic organisms. The active centre is a 2Fe-2S cluster, chelated by four conserved cysteine residues [1]. Ferredoxin functions as an electron carrier in the photosynthetic electron transport chain of the chloroplast and also plays a central role as an electron donor to various cellular processes such as nitrate reductase, sulphite reductase and glutamate synthase [2]. There has been shown to be two types of plant ferredoxin, these differ in amino acid composition but are similar in terms of structure and function [3].

An alignment of twelve sequences was prepared from which three motifs were selected. The first motif contains the first two conserved cysteines, while motifs two and three are derived from the regions surrounding the second cysteine cluster. After two iterations convergence had been reached as all the plant type ferredoxin sequences present in the database were shown to match with all three features. There is no discrimination for the bacterial type ferredoxins that exhibit a different cysteine spacing.

```
SUMMARY INFORMATION
-------------------
  59 codes involving  3 elements
   0 codes involving  2 elements

COMPOUND FEATURE INDEX
----------------------

 3|  59   59   59
 2|   0    0    0
--+----------------
  |   1    2    3
```

FER1$SYNP7
FEKM
FER1$PEA
ANAPETF
FER$ARATH
FETA
FER1$RAPSA
FEPRR
FEDH1
FER2$RAPSA
FEFNG
FER$BUMFI
FER$BRYMA
FEAH2
FEYC2
FER1$CYAPA
FER1$PHYES
FEBQ
JX0082
FER3$RAPSA
FESG
FEFW2E
FEFW2
FERZ
FEWT
FER$MARPO
FELG
FESP1
FER$PERBI
FEHS
FENM1M
FEMW
FENM
FEKK
FEFZ1
FER$SILPR
FEYCAL
FEED
S03730
N$3FXC
FEDH2
FEAH
FEAA
FEEQ2
FEHSX
FESC
FEEF
FER1$ANAVA
FEYB6
FESP2
FERP
FEYCT
FER$APHHA
FER2$CYACA
FEPRU
FEFW1
FEEQ1
FENM2M
FER2$ANASP

FER1$SYNP7	FERREDOXIN I - *Synechococcus sp.*
FER1$CYAPA	FERREDOXIN I - *Cyanophora paradoxa*
FENM1M	Ferredoxin I - *Nostoc muscorum*
FESC	Ferredoxin - *Scenedesmus quadricauda*
FEKM	Ferredoxin - *Chlamydomonas reinhardtii*
FER1$PHYES	FERREDOXIN I - Food pokeberry
FEMW	Ferredoxin - *Fischerella sp.*
FEEF	Ferredoxin - *Chlorogloeopsis fritschii*
FER1$PEA	FERREDOXIN I - Garden pea
FEBQ	Ferredoxin - Great burdock
FENM	Ferredoxin I - *Nostoc muscorum*
FER1$ANAVA	FERREDOXIN I - *Anabaena variabilis*
ANAPETF	ANAPETF ferredoxin I - *Anabaena sp.*
JX0082	Ferredoxin L-Fd A - Radish
FEKK	Ferredoxin - *Cyanidium caldarium*
FEYB6	Ferredoxin - *Synechocystis sp.*
FER$ARATH	FERREDOXIN PRECURSOR - *Arabidopsis thaliana*
FER3$RAPSA	FERREDOXIN, LEAF L-A - Radish
FEFZ1	Ferredoxin I - *Aphanizomenon flos-aquae*
FESP2	Ferredoxin II - Spinach
FETA	Ferredoxin - Elephant's ear
FESG	Ferredoxin - *Spirulina maxima*
FER$SILPR	FERREDOXIN PRECURSOR. - White campion
FERP	Ferredoxin - Rape
FER1$RAPSA	FERREDOXIN ROOT R-B1 - Radish
FEFW2E	Ferredoxin II - Food pokeberry
FEYCAL	Ferredoxin - *Synechococcus lividus*
FEYCT	Ferredoxin - *Synechococcus sp.*
FEPRR	Ferredoxin - Red alga
FEFW2	Ferredoxin II - Common pokeberry
FEED	Ferredoxin - European elder
FER$APHHA	FERREDOXIN - *Aphanothece halophitica*
FEDH1	Ferredoxin I - *Dunaliella salina*
FERZ	Ferredoxin I - Rice
S03730	Ferredoxin I - Rice
FER2$CYACA	FERREDOXIN - *Cyanidium caldarium*
FER2$RAPSA	FERREDOXIN ROOT R-B2 - Radish
FEWT	Ferredoxin - Wheat
N$3FXC	Ferredoxin - *Spirulina platensis*

```
FEPRU              Ferredoxin - Laver
FEFNG              Ferredoxin - Urajiro
FER$MARPO          FERREDOXIN - Liverwort
FEDH2              Ferredoxin II - Dunaliella salina
FEFW1              Ferredoxin I - Common pokeberry
FER$BUMFI          FERREDOXIN - Bumilleriopsis filiformis
FELG               Ferredoxin - White popinac
FEAH               Ferredoxin - Aphanothece sacrum
FEEQ1              Ferredoxin I - Horsetail
FER$BRYMA          FERREDOXIN - Bryopsis maxima
FESP1              Ferredoxin I - Spinach
FEAA               Ferredoxin - Alfalfa
FENM2M             Ferredoxin II - Nostoc muscorum
FEAH2              Ferredoxin II - Aphanothece sacrum
FER$PERBI          FERREDOXIN - Peridinium bipes
FEEQ2              Ferredoxin II - Horsetail
FER2$ANASP         FERREDOXIN HETEROCYST - Anabaena sp.
FEYC2              Ferredoxin II (2Fe-2S) - Synechococcus sp.
FEHS               Ferredoxin - Halobacterium halobium
FEHSX              Ferredoxin - Halobacterium sp.

SCAN HISTORY
------------
OWL10_1     3     100 NSINGLE


INITIAL MOTIF-SETS
------------------
FERREDOXIN1
11
ferr_mot_1
DLPYSCRAGAC                              FESP2      34     34
DLPYSCRAGAC                              FEFNG      34     34
DLPYSCRAGAC                              FEDH2      33     33
DLPYSCRAGAC                              FEFZ1      35     35
DLPYSCRAGSC                               FEWT      34     34
DLPYSCRAGSC                               FERZ      34     34
DLPYSCRAGSC                              FEFW1      34     34
DLPYSCRAGSC                          FER$SILPR      83     83
ELPYSCRAGAC                              FEPRU      36     36
DLPLSCQAGAC                              FEEQ2      32     32
DWPFSCRAGAC                               FEHS      58     58
DLPASCLTGVC                              FEYC2      35     35

FERREDOXIN2
13
ferr mot 2
SSCAGKVTSGSVD                            FESP2      45      0
SSCTGKLLDGRVD                            FEFNG      45      0
SSCAGKVEAGTID                            FEDH2      44      0
STCAGKLVTGTID                            FEFZ1      46      0
SSCAGKLVSGEID                             FEWT      45      0
SSCAGKVVSGEID                             FERZ      45      0
SSCTGKVTAGTVD                            FEFW1      45      0
SSCAGKVVAGSVD                        FER$SILPR      94      0
STCAGKVTEGTVD                            FEPRU      47      0
STCLGKIVSGTVD                            FEEQ2      43      0
```

```
ANCASIVKEGEID                           FEHS      69      0
TTCAARILSGEVD                          FEYC2      46      0

FERREDOXIN3
8
ferr mot 3
VLTCIAYP                               FESP2      74     16
VLTCVAYP                               FEFNG      74     16
VLTCVAYA                               FEDH2      73     16
VLTCVAYP                               FEFZ1      75     16
VLTCHAYP                                FEWT      74     16
VLTCHAYP                                FERZ      74     16
VLTCVAFP                               FEFW1      74     16
VLTCAAYP                           FER$SILPR     123     16
VLTCIAYP                               FEPRU      76     16
VLTCIAIP                               FEEQ2      72     16
RLTCIGSP                                FEHS      99     17
TLLCVAYP                               FEYC2      75     16

FINAL MOTIF-SETS
----------------
FERREDOXIN1
11
ferr_mot_1
DLPYSCRAGAC                            FESP2      34     34
DLPYSCRAGAC                            FEFW2      35     35
DLPYSCRAGAC                           FEFW2E      35     35
DLPYSCRAGAC                            FEFNG      34     34
DLPYSCRAGAC                            FEPRR      35     35
DLPYSCRAGAC                             FEKM      32     32
DLPYSCRAGAC                             FESC      34     34
DLPYSCRAGAC                            FEDH2      33     33
DLPYSCRAGAC                             FEKK      36     36
DLPYSCRAGAC                            FEYB6      34     34
DLPYSCRAGAC                            FEFZ1      35     35
DLPYSCRAGAC                             FESG      36     36
DLPYSCRAGAC                             FEEF      36     36
DLPYSCRAGAC                             FEMW      36     36
DLPYSCRAGAC                             FEAH      34     34
DLPYSCRAGAC                           FENM1M      36     36
DLPYSCRAGAC                        FER$APHHA      36     36
DLPYSCRAGAC                       FER1$CYAPA      37     37
DLPYSCRAGAC                       FER1$RAPSA      36     36
DLPYSCRAGAC                       FER1$SYNP7      36     36
DLPYSCRAGAC                       FER2$CYACA      35     35
DLPYSCRAGAC                       FER2$RAPSA      36     36
DLPYSCRAGAC                           N$3FXC      36     36
DLPYSCRAGSC                            FESP1      34     34
DLPYSCRAGSC                             FETA      34     34
DLPYSCRAGSC                             FEBQ      34     34
DLPYSCRAGSC                             FERP      34     34
DLPYSCRAGSC                             FEWT      34     34
DLPYSCRAGSC                             FERZ      34     34
DLPYSCRAGSC                            FEFW1      34     34
DLPYSCRAGSC                            FEDH1      33     33
DLPYSCRAGSC                        FER$ARATH      86     86
DLPYSCRAGSC                        FER$SILPR      83     83
```

DLPYSCRAGSC	FER1$PEA	86	86
DLPYSCRAGSC	FER1$PHYES	34	34
DLPYSCRAGSC	FER3$RAPSA	34	34
DLPYSCRAGSC	S03730	34	34
DLPYSCRAGSC	JX0082	34	34
DLPFSCRAGAC	FEEQ1	33	33
DLPFSCRAGAC	FENM	36	36
DLPFSCRAGAC	FEYCAL	34	34
DLPFSCRAGAC	FEYCT	35	35
DLPFSCRAGAC	FER1$ANAVA	36	36
DLPFSCRAGAC	ANAPETF	37	37
ELPYSCRAGAC	FEPRU	36	36
ELPYSCRAGAC	FER$BUMFI	36	36
SLPYSCRAGAC	FER$MARPO	33	33
DLPSSCRAGSC	FEAH2	36	36
ELPYSCRAGSC	FELG	33	33
ELPYSCRAGSC	FER$PERBI	32	32
VLPYSCRAGSC	FEAA	34	34
DIPYSCRAGSC	FEED	34	34
DWPFSCRAGAC	FEHS	58	58
DWPFSCRAGAC	FEHSX	58	58
DLPFSCRSGSC	FENM2M	36	36
DLPLSCQAGAC	FEEQ2	32	32
DIPFSCRSGSC	FER$BRYMA	35	35
DLPASCLTGVC	FEYC2	35	35
ELPFSCHSGSC	FER2$ANASP	36	36

FERREDOXIN2
13
ferr mot 2

SSCAGKVTSGSVD	FESP2	45	0
SSCAGKVTAGAVN	FEFW2	46	0
SSCAGKVTAGSVN	FEFW2E	46	0
SSCTGKLLDGRVD	FEFNG	45	0
STCAGIVELGTVD	FEPRR	46	0
SSCAGKVAAGTVD	FEKM	43	0
SSCAGKVEAGTVD	FESC	45	0
SSCAGKVEAGTID	FEDH2	44	0
STCAGKLLEGEVD	FEKK	47	0
STCAGKITAGSVD	FEYB6	45	0
STCAGKLVTGTID	FEFZ1	46	0
STCAGKITSGSID	FESG	47	0
STCAGKIKSGTVD	FEEF	47	0
STCAGKLISGTVD	FEMW	47	0
STCAGKLVSGPAP	FEAH	45	0
STCAGKIVSGTVD	FENM1M	47	0
STCAGKIKEGEID	FER$APHHA	47	0
STCAGKVVEGTVD	FER1$CYAPA	48	0
STCAGKIEKGQVD	FER1$RAPSA	47	0
STCAGKVVSGTVD	FER1$SYNP7	47	0
STCAGKLVKGSVD	FER2$CYACA	46	0
STCAGQIVKGQVD	FER2$RAPSA	47	0
STCAGTITSGTID	N$3FXC	47	0
SSCAGKLKTGSLN	FESP1	45	0
SSCAGKVKVGDVD	FETA	45	0
SSCAGKVTAGSVD	FEBQ	45	0
SSCAGKVVSGFVD	FERP	45	0

SSCAGKLVSGEID	FEWT	45	0
SSCAGKVVSGEID	FERZ	45	0
SSCTGKVTAGTVD	FEFW1	45	0
SSCAGKVESGTVD	FEDH1	44	0
SSCAGKVVSGSVD	FER$ARATH	97	0
SSCAGKVVAGSVD	FER$SILPR	94	0
SSCAGKVVGGEVD	FER1$PEA	97	0
SSCAGKVTAGTVD	FER1$PHYES	45	0
SSCAGKVVSGSVD	FER3$RAPSA	45	0
SSCAGKVVSGEID	S03730	45	0
SSCAGKVVSGTVD	JX0082	45	0
SSCLGKVVSGSVD	FEEQ1	44	0
STCAGKLVSGTVD	FENM	47	0
STCAGKLLEGEVD	FEYCAL	45	0
STCAGKLLEGEVD	FEYCT	46	0
STCAGKLVSGTVD	FER1$ANAVA	47	0
STCAGKLVSGTVD	ANAPETF	48	0
STCAGKVTEGTVD	FEPRU	47	0
STCAGKVLSGTID	FER$BUMFI	47	0
SSCAGKVTAGEVD	FER$MARPO	44	0
STCAGKLVSGAAP	FEAH2	47	0
SSCAGKLVEGDLD	FELG	44	0
SSCAGKVLTGSID	FER$PERBI	43	0
SSCAGKVAAGEVN	FEAA	45	0
SSCAGKLVAGSVD	FEED	45	0
ANCASIVKEGEID	FEHS	69	0
ANCAAIVLEGDID	FEHSX	69	0
SSCNGILKKGTVD	FENM2M	47	0
STCLGKIVSGTVD	FEEQ2	43	0
STCAGKIEGGTVD	FER$BRYMA	46	0
TTCAARILSGEVD	FEYC2	46	0
SSCVGKVVEGEVD	FER2$ANASP	47	0

FERREDOXIN3
8
ferr mot 3

VLTCIAYP	FESP2	74	16
VLTCVAYP	FEFW2	75	16
VLTCVAYP	FEFW2E	75	16
VLTCVAYP	FEFNG	74	16
VLTCVAYP	FEPRR	75	16
VLTCVAYP	FEKM	72	16
VLTCVAYP	FESC	74	16
VLTCVAYA	FEDH2	73	16
VLTCVAYP	FEKK	76	16
VLTCVAYP	FEYB6	74	16
VLTCVAYP	FEFZ1	75	16
VLTCVAYP	FESG	76	16
VLTCVAYP	FEEF	76	16
VLTCVAYP	FEMW	76	16
ILTCVAYP	FEAH	74	16
VLTCVAYP	FENM1M	76	16
VLTCVAYP	FER$APHHA	76	16
VLTCVAYP	FER1$CYAPA	77	16
VLTCVAYP	FER1$RAPSA	76	16
VLTCVAYP	FER1$SYNP7	76	16
ILTCVAYP	FER2$CYACA	75	16

VLTCVAYP	FER2$RAPSA	76	16
VLTCVAYP	N$3FXC	76	16
VLTCAAYP	FESP1	74	16
VLTCVAYP	FETA	74	16
VLTCVAYP	FEBQ	74	16
VLTCAAYP	FERP	74	16
VLTCHAYP	FEWT	74	16
VLTCHAYP	FERZ	74	16
VLTCVAFP	FEFW1	74	16
VLTCVAYA	FEDH1	73	16
VLTCAAYP	FER$ARATH	126	16
VLTCAAYP	FER$SILPR	123	16
VLTCVAYP	FER1$PEA	126	16
VLTCVAYP	FER1$PHYES	74	16
VLTCAAYP	FER3$RAPSA	74	16
VLTCHAYP	S03730	74	16
VLTCAAYP	JX0082	74	16
VLTCIAIP	FEEQ1	73	16
VLTCVAYP	FENM	76	16
VLTCVAYP	FEYCAL	74	16
VLTCVAYP	FEYCT	75	16
VLTCVAYP	FER1$ANAVA	76	16
VLTCVAYP	ANAPETF	77	16
VLTCIAYP	FEPRU	76	16
LLTCVAYP	FER$BUMFI	76	16
VLTCIAYP	FER$MARPO	73	16
VMTCVAYP	FEAH2	77	17
VLTCAAYP	FELG	73	16
CLTCVTYP	FER$PERBI	72	16
VLTCVAYA	FEAA	74	16
VLTCVAYP	FEED	74	16
RLTCIGSP	FEHS	99	17
RLTCIGSP	FEHSX	99	17
VLTCVAYP	FENM2M	76	16
VLTCIAIP	FEEQ2	72	16
VLTCVAYP	FER$BRYMA	75	16
TLLCVAYP	FEYC2	75	16
ALLCVTYP	FER2$ANASP	76	16

Appendix D

Amino acid notation and colours used for multiple sequence alignments

Amino acid	1 letter code	3 letter code	Alignment colour
Alanine	A	Ala	grey
Aspar/agine/tate	B	Asx	grey
Cysteine	C	Cys	yellow
Aspartate	D	Asp	red
Glutamate	E	Glu	red
Phenylalanine	F	Phe	purple
Glycine	G	Gly	brown
Histidine	H	His	blue
Isoleucine	I	Ile	grey
Lysine	K	Lys	blue
Leucine	L	Leu	grey
Methionine	M	Met	grey
Asparagine	N	Asn	green
Proline	P	Pro	brown
Glutamine	Q	Gln	green
Arginine	R	Arg	blue
Serine	S	Ser	green
Threonine	T	Thr	green
Valine	V	Val	grey
Tryptophan	W	Trp	purple
Unidentified	X		grey
Tyrosine	Y	Tyr	purple
Glutam/ine/ate	Z	Glx	grey

Colour	Residues	Property
Green	S T N Q	Polar uncharged
Grey	A V L I M	Hydrophobic
Blue	H K R	Basic
Red	D E	Acidic
Purple	F Y W	Aromatic
Brown	G P	Structural oddities
Yellow	C	Cysteine/ine