ABSTRACT

A COMPARATIVE STUDY OF SEQUENCE ANALYSIS TOOLS IN COMPUTATIONAL BIOLOGY

by
Wei-Jen Chuang

A biomolecular object, such as a deoxyribonucleic acid (DNA), a ribonucleic acid (RNA) or a protein molecule, is made up of a long chain of subunits. A protein is represented as a sequence made from 20 different amino acids, each represented as a letter. There are a vast number of ways in which similar structural domains can be generated in proteins by different amino acid sequences. By contrast, the structure of DNA, made up of only four different nucleotide building blocks that occur in two pairs, is relatively simple, regular, and predictable.

Biomolecular sequence alignment/string search is the most important issue and challenging task in many areas of science and information processing. It involves identifying one-to-one correspondences between subunits of different sequences. An efficient algorithm or tool is involved with many important factors, these include the following: Scoring systems, Alignment statistics, Database redundancy and sequence repetitiveness.

Sequence "motifs" are derived from multiple alignments and can be used to examine individual sequences or an entire database for subtle patterns. With motifs, it is sometimes possible to detect distant relationships that may not be demonstrable based on comparisons of primary sequences alone.

A more comprehensive solution to the efficient string search is approached by building a small, representative set of motifs and using this as a screening database with automatic masking of matching query subsequences. This technology is still under development but recent studies indicate that a representative set of only 1,000 – 3,000 sequences may suffice and such a database can be searched in seconds.

A COMPARATIVE STUDY OF SEQUENCE ANALYSIS TOOLS IN COMPUTATIONAL BIOLOGY

by
Wei-Jen Chuang

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science and Computer Science

Department of Computer and Information Science

January 1999

APPROVAL PAGE

A COMPARATIVE STUDY OF SEQUENCE ANALYSIS TOOLS IN COMPUTATIONAL BIOLOGY

Wei-Jen Chuang

Dr. Jason T.L. Wang, Thesis Advisor — Date
Associate Professor of Computer and Information Science, NJIT

Dr. James M. Calvin, Committee Member — Date
Assistant Professor of Computer and Information Science, NJIT

Dr. Franz J. Kurfess, Committee Member — Date
Assistant Professor of Computer and Information Science, NJIT

BIOGRAPHICAL SKETCH

Author: Wei-Jen Chuang

Degree: Master of Science

Date: January 1999

Undergraduate and Graduate Education:

- Master of Science in Computer Information and Science,
  New Jersey Institute of Technology, Newark, NJ 1999
- Master of Science in Phytopathology,
  National Chung-Hsing University, Taichung, Taiwan, R.O.C., 1990
- Bachelor of Science in Plant Pathology,
  National Chung-Hsing University, Taichung, Taiwan, R.O.C., 1988

Major: Computer Information and Science

Publication:

W.-J. Chuang,
"*Etiological Studies On Stem Rot and Basal Rot of Taiwan Anoetcochilus and Their Control*," NCHU, Taichung, Taiwan, R.O.C., July, 1990.

This work is dedicated to
my mother,
who is no longer with us.

ACKNOWLEDGMENT

First, I would like to express my sincere appreciation to my advisor, Professor Jason T. L. Wang, for his patience and constant guidance during the course of this research.

Also, I would like to thank Professor James M. Calvin and Professor Franz J. Kurfess for their activity participating in my committee.

Most of all, I want to express my appreciation to my family, for their supporting and devotion, and to my wife, Mei-Yi, for her love and understanding. Without their support and encouragement, my accomplishment would not be possible.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Biomolecular sequence alignment is among the most important and challenging tasks in computational biology; it involves identifying one-to-one correspondences between subunits of different sequences [33]. This procedure is essential to many tasks of biological data analysis [21,25,47], for example, retrieving a database to determine the species of unknown sequences, and discovering highly conserved subregious or patterns related to molecular structures and functions. Even by using genetic tools, forensic scientists can now examine the DNA in this biological evidence and tell almost certainly whether it came from a given individual [31].

A biomolecular object, such as a deoxyribonucleic acid (DNA), a ribonucleic acid (RNA) or a protein molecule, is made up of a long chain of subunits. In molecular sequence studies, sequence subunits are represented by characters from a domain which is denoted by ***L***. For example, the characters used to represent the nucleotides in DNA sequences are A (Adenosine), G (Guanine), C (Cytidine) and T (Thymidine), and ***L*** is the set {A,G,C,T} [52]. A protein is represented as a sequence made from 20 different amino acids (see Table 1.1), each represented as a letter.

By contrast, the structure of DNA, made up of only four different nucleotides that occur in two pairs, is relatively simple, regular, and predictable. In aligning biomolecular sequences, a function of scores (or distances) is needed to measure the goodness of alignments [20,43,48]. For a set of sequences, the optimal alignment is the one which maximizes the score (or minimizes the distance) [13,42]. To achieve a high score value, some subunits in different sequences are matched, and some sequences are considered to have insertions, deletions, and substitutions [1,40]. The following illustrates an alignment of three sequences.

Sequence 1	A	T	G	C	A	G	G	C
Sequence 2	A	T	G	A	X	G	X	C
Sequence 3	A	T	G	A	X	G	G	C
Column	1	2	3	4	5	6	7	8

In the alignment columns 1,2,3,6, and 8 indicate character matches, column 4 indicates that a substitution has occurred in the first sequence (A-C, T-G), column 5 indicates that a subunit has been inserted into the first sequence, column 7 indicates that a subunit has been deleted from the second sequence. To represent insertion and deletion, the character X is introduced.

Table 1.1 List of 20 Amino Acids

Abbreviations	Amino Acids
A	Alanine
C	Cystine
D	Aspartate
E	Glutamate
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
K	Lysine
L	Leucine
M	Methionine
N	Asparagine
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threinine
V	Valine
W	Tryptophan
Y	Tyrosine

String search is an important operation in many areas of science and information processing. It occurs naturally as part of data processing, text editing, lexical analysis and information retrieval [51]. In molecular biology this computational problem is associated with the analysis, search and comparison of biosequences, which can be considered as texts made up of only four characters in the case of nucleic acid, and twenty, in the case of proteins [34]. While the particular algorithm used is of course important, the effectiveness of database searches is dependent as well on a large number of correlative factors, many of which tend to be overlooked or dealt with an inefficient or *ad hoc* manner [5]. An efficient algorithm or tool involves many important factors, which include the following :

Scoring systems: The molecular biologist is often confronted with the task of searching a database of DNA or protein sequences for those most similar to a given one. The most straightforward definition of similarity between two sequences attributes a "score" to each of the possible ways of aligning them, including the possibility of arbitrarily long insertions/deletions at any position. Most database search algorithms rank alignments by a score, whose calculation is dependent upon a particular scoring system. Usually there is a default system, but it may not be ideal for a user's particular problem. For example, haemoglobin subunits are used to be regarded as "typical" proteins and are still often used as benchmark query sequences for evaluating new database search techniques and scoring systems [23]. However, today it is more common to encounter much larger and more complex sequences and therefore those methods developed and optimized for small, uniformly-conserved, single-domain proteins are inadequate. Optimal strategies for detecting similarities between DNA protein-coding regions differ from those for non-coding regions [4,30]. A database search program should therefore make a variety of scoring systems available and users should be aware of which ones are best suited to their problems.

Alignment statistics: Given a query sequence, most database search programs will produce an ordered list of imperfectly matching database similarities, but none of them need have any biological significance. An important question is how strong a similarity is necessary to be considered surprising.

Database: The use of an up-to-date sequence database is clearly a vital element of any similarity search. Sequence relationships critical to important discoveries have on occasion been missed because old or imcomplete databases were employed [8,50]. The variety of available databases, and their overlapping coverage, has the potential to render similarity searching cumbersome and inefficient. However, today one can download sequences from the Internet (See Table 1.2. and Table 1.3.). Timely access to complete and "nonredudant" sequence databases has become relatively simple and inexpensive.

Database redundancy and sequence repetitiveness: Surprisingly strong biases exist in protein and nucleic acid sequence database. Many of these reflect fundamental mosaic sequence properties that are of considerable biological interest in themselves, such as segments of flow compositional complexity or short-period repeats. Databases also contain some very large families of related domains, motifs or repeated sequences, in some cases with hundred of members. In other cases there has been a historical bias in the molecules that has been chosen for sequencing. In practice, unless special measures are taken, these biases very commonly confound database search methods and interfere with the discovery of interesting new sequence similarities. Problems include the occurrence of misleading, spuriously-high scores, ambiguities in the phase of sequence alignments and overwhelmingly large output lists in which interesting results may be inconspicuously buried. Failure to deal properly with the factors described above can result in chance similarities being claimed significant, or biological important relationships being overlooked.

There are a number of important issues in searching DNA and protein sequence databases, but the most important is access to a comprehensive and up-to-date data repository [3]. We will use the SDISCOVER program [10,49] to find motifs in DNA sequences and use those motifs to form a local database to try to find a better way that can speed up the database search. We will also compare the motifs found by SDISCOVER with the motifs/patterns stored in Prosite protein database.

So far, there are many tools that can do the query sequences search or alignment. One can easily find a tool that suits for his needs from the Internet (Table 1.2, Table 1.3)

Table 1.2 Selected World Wide Web Sites

Compendia of WWW Resources	
BIOSCI Newsgroups	http://www.bio.net/
EBI	http://www.ebi.ac.uk/
Pedro's Biomolecular Research Tools	http://www.public.iastate.edu/~pedro/research_tools.html
WWW Virtual Library	http://golgi.harvard.edu/htbin/biopages
Sequence Retrieval and Analysis	
ExPASy Molecular Biology Server	http://expasy.hcuge.ch/
NCBI	http://www.ncbi.nlm.nih.gov
NCBI BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
NCBI Entrez	http://www3.ncbi.nlm.nih.gov/Entrez/
PDB	http://www.pdb.bnl.gov
Organism-Specific Web Resources	
Arabidopsis (AtDB)	http://genome-www.stanford.edu/Arabidopsis/
C. elegans	http://eatworms.swmed.edu/
Flybase	http://morgan.harvard.edu/
Mouse Genome Database	http://www.informatics.jax.org/mgd.html
Saccharomyces (SacchDB)	http://genome-www.stanford.edu/Saccharomyces/
Electronic Journals	
Cell	http://www.cell.com/
Genome Research	http://www.cshl.org:80/journals/gr/
Journal of Biological Chemistry	http://www-jbc.stanford.edu/jbc/
Journal of Molecular Biology	http://www.hbuk.co.uk/jmb
Nature	http://www.nature.com/
Science	http://science-mag.aaas.org/science/home/
Pedro's List of Bio/Chemical Journals and Newsletters	http://www.public.iastate.edu/~pedro/.rt_journals.html

Table 1.3 Selected Molecular Biology FTP Servers

Database Servers		
FTP Server	Major Databases Available	FTP Server Address
NCBI	GenBank, SWISS-PROT, PIR	ncbi.nlm.nih.gov
EBI	EMBL, SWISS-PROT	ftp.ebi.ac.uk
ExPASy	Enzyme, EPD, Prosite, SeqanalRef, SWISS-PROT, SWISS-2DPAGE, SWISS-3DIMAGE	expasy.hcuge.ch
Software Servers		
FTP Server	Software Available	FTP Server Address
NCBI	BLAST, Sequin, GenInfo Software Toolbox, MACAW	ncbi.nlm.nih.gov
EBI	Mac, VAX, DOS, UNIX molecular biology software	ftp.ebi.ac.uk
IuBio	Mac, VAX, DOS, Atari software;	ftp.bio.indiana.edu

Most of those programs are required to be run on a UNIX system or need to retrieve the database from the Internet. Our goals are to combine the SDISCOVER sets into one efficient tool and run it on PC. Second, we want to evaluate a new approach that can improve the performance of database search.

CHAPTER 2

MODIFYING THE SDISCOVER PROGRAM

Sequence "motifs" are derived from multiple alignments and can be used to examine individual sequences or an entire database for subtle patterns. With motifs, it is sometimes possible to detect distant relationships that may not be demonstrable based on comparisons of primary sequences alone [15,29,41].

The SDISCOVER program is used to find motifs of the query sequences and run on UNIX system. It includes two separate C programs. These two programs making up the SDISCOVER tool are termed, the control module (including the user-interface module or command line and a similarity-score-calculation module) , and the sorting module which eliminates the substrings. The user-interface module, collects the input from the user, and the criteria used in the computation of similarity scores and then writes out the results. In the present version of SDISCOVER tool, the query sequence is read by the user-interface module. The control module receives the input data/query sequence from the user-interface module and relays the query sequence and information for the calculation of similarity scores to each of the similarity-score-calculation modules. Once all sequences have been processed, the control module sends the list of scoring sequences to the user-interface module.

In the original SDISCOVER tool, the user first enters the query sequences from the input interface to find the motifs. After motifs are found, the sorting program is used to sort/elinimate the substrings (as shown in Figure 2.1 and Figure 2.2). We combine these two steps into one to simplify the procedure but keep the original algorithm and modify it to run on PC. We do this not only because the Windows operation system is the most popular operating system but also because users may not have the access to the UNIX system (Figure 2.3).

Our test environment: CPU Pentium 200 Pro, 128 Mb RAM, Operating system Windows 95.

```
C:\PROJECT>discover
% Enter the file name of sequences
 (an example file can be found in file SAMPLE;
  maximum number of sequences in the file is 200;
  maximum length of sequences is 200) [SAMPLE]:SAMPLE

 ===> 5 sequences found in file <SAMPLE>

% Enter the form of interesting motifs 1 or 2
 (1 means *X*; 2 means *X*Y*) [1] ?

% Enter the minimum length of interesting motifs
 (default is 10) [10] ?

% Enter the minimum occurrence number for interesting motifs
 (the occurrence number of an interesting motif
  refers to the number of sequences in which
  the motif approximately occurs; default is 2) [2] ?

% Enter the number of mutations allowed in searching
 for similar motifs (default is 1; maximum number is 10) [1] ?

% Where the result should be stored (enter the file name) [data.out] ? data.out

  Occurrence number    Motif
  -----------------    -------------------
        2          *MGIVSWGEGC*
        2          *GIVSWGEGCA*
        2          *GIVSWGEGCAR*
        2          *GIVSWGEGCD*
        2          *GIVSWGEGCDR*
        2          *TGIVSWGEGC*
        2          *IVSWGEGCAR*
        2          *IVSWGEGCDR*

    16 motifs found
   350 motifs checked
```

Figure 2.1 The Input Interface of SDISCOVER

```
C:\PROJECT\discover>ssort data.out > sorted.out
Minimum length = 10
Minimum occurrence number = 2
Number of mutations allowed = 1
Total number of sequences = 3
Input file name = SAMPLE


 Occurrence number    Motif
 -----------------   -------------------
After sort...
      2          *GIVSWGEGCDR*
      2          *GIVSWGEGCAR*
      2          *TGIVSWGEGC*
      2          *MGIVSWGEGC*
```

Figure 2.2 Illustration of Executing the Sorting Program

In this example the sorting module uses the default output file from the control module as input file, in this example: data.out. After eliminating the substrings, the program write the results to output file, sorted.out.

A substring is a shorter sequence which can be found in a longer sequence and these two sequences' occurrence numbers are the same. Then we say the shorter one is a substring of the longer one.

```
C:\PROJECT\discover>discover
% Enter the file name of sequences
 (an example file can be found in file SAMPLE;
  maximum number of sequences in the file is 5000;
  maximum length of sequences is 5000) [SAMPLE]: SAMPLE

 ===> 3 sequences found in file <SAMPLE>

% Enter the form of interesting motifs 1 or 2
 (1 means *X*; 2 means *X*Y*) [1] ?

% Enter the minimum length of interesting motifs
 (default is 10) [10] ?

% Enter the minimum occurrence number for interesting motifs
 (the occurrence number of an interesting motif
 refers to the number of sequences in which
 the motif approximately occurs; default is 2) [2] ?

% Enter the number of mutations allowed in searching
 for similar motifs (default is 1; maximum number is 10) [1] ?

% Where the result should be stored (enter the file name) [data.out] ? data.out


 Occurrence number    Motif
  -----------------    -----
        2          *MGIVSWGEGC*
        2          *GIVSWGEGCA*
        2          *GIVSWGEGCAR*
        2          *GIVSWGEGCD*
        2          *GIVSWGEGCDR*
        2          *TGIVSWGEGC*
        2          *IVSWGEGCAR*
        2          *IVSWGEGCDR*
 After sorted...
        2          *GIVSWGEGCDR*
        2          *GIVSWGEGCAR*
        2          *TGIVSWGEGC*
        2          *MGIVSWGEGC*

      8 motifs found
    350 motifs checked
```

Figure 2.3 Illustration of Executing the Modified Program

CHAPTER 3

FINDING MOTIFS AND DATABASE EVALUATION

3.1. Searching for Motifs

Our test environments: CPU Pentium 200 Pro, 128 Mb RAM, Operating system Win95, and Sun Sparc Ultra-2 Pentium II 300 MHz with 512 Mb RAM.

GenBank, the EMBL nucleotide sequence database, and the DNA Database of Japan (DDBJ) are three partners in a long-standing collaboration to collect and distribute all publicly-available sequence data [6,38]. All of the sequences we use (both DNA and protein sequences) in this experiment are download from GenBank at NCBI homepage (see Appendix B for a complete list of Human DNA). There is a total number of 181423 sequences stored in NCBI until May 7. The DNA sequences we used as query sequences to find motifs from the database are as following:

nci_cgap_br7.fasta
nci_cgap_hn1.fasta
nci_cgap_hn3.fasta
nci_cgap_li5.fasta
nci_cgap_lu6.fasta
nci_cgap_mel3.fasta
nci_cgap_ov8.fasta
nci_cgap_pns1.fasta
nci_cgap_pr20.fasta

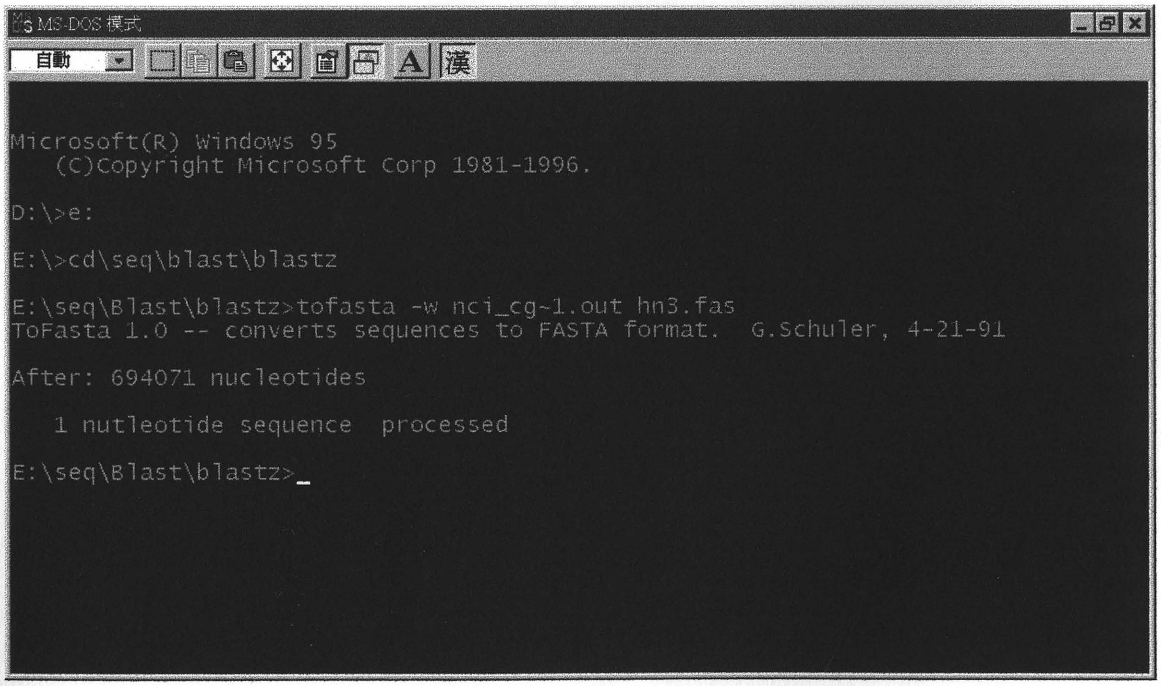
And the output of all sequences used to find motifs and query parameters after being organized are shown in Table 3.1.

Table 3.1 Lists of All Results of Finding the Motifs

	Minimum length	Minimum occurrence number	Number of mutations allowed	Total number of sequences	motifs found	motifs checked	After sorted
nci_cgap_br7	10	2	1	326	764481	850818	335614
nci_cgap_hn1	10	2	1	35	6225	14825	3072
nci_cgap_hn3	10	2	1	131	288606	333710	11208
nci_cgap_li5	10	2	1	147	96528	124704	37822
nci_cgap_lu6	10	2	1	45	174023	188003	77135
nci_cgap_mel3	10	2	1	237	482777	545209	177401
nci_cgap_ov8	10	2	1	24	54233	61720	16194
nci_cgap_pns1	10	2	1	297	418562	559749	16596
nci_cgap_pr20	10	2	1	166	625445	664218	265785
Total				1408	2910880	3342956	940827

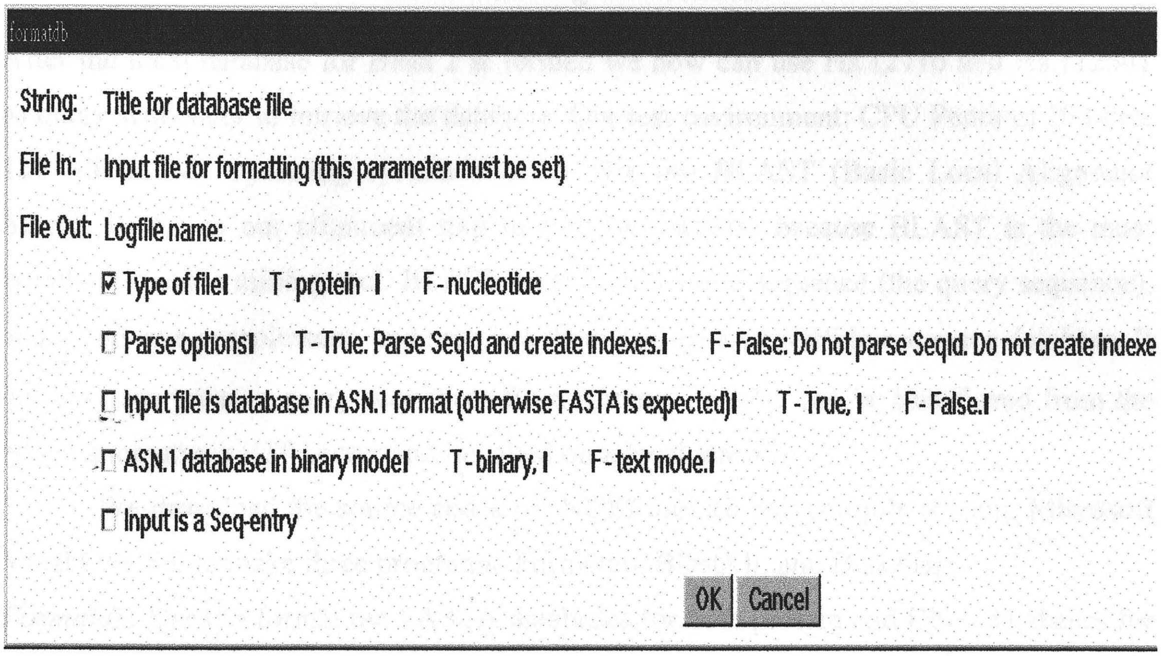
3.2. Converting the Output Into FASTA Format and Forming a Local Database

After finding the motifs of the query sequences, we convert the outputs into FASTA format (Figure 3.2.1) because the FASTA format already becomes a DNA sequence standard format (see Appendix C for a detailed description of FASTA) and the alignment tool we use can recognize this format. Then we use the motifs we found to form a local database by using NCBI Tools (Figure 3.2.2). Forming a local database has many advantages; for example, the user may not have the access to the Internet or/and can reduce the traffic of the Internet and can update the database more easily.



```
Microsoft(R) Windows 95
   (C)Copyright Microsoft Corp 1981-1996.

D:\>e:

E:\>cd\seq\blast\blastz

E:\seq\Blast\blastz>tofasta -w nci_cg~1.out hn3.fas
ToFasta 1.0 -- converts sequences to FASTA format.  G.Schuler, 4-21-91

After: 694071 nucleotides

   1 nutleotide sequence  processed

E:\seq\Blast\blastz>_
```

Figure 3.2.1 A Screen Shot of Converting Output to FASTA Format



formatdb

String: Title for database file

File In: Input file for formatting (this parameter must be set)

File Out: Logfile name:

☑ Type of fileI T - protein I F - nucleotide

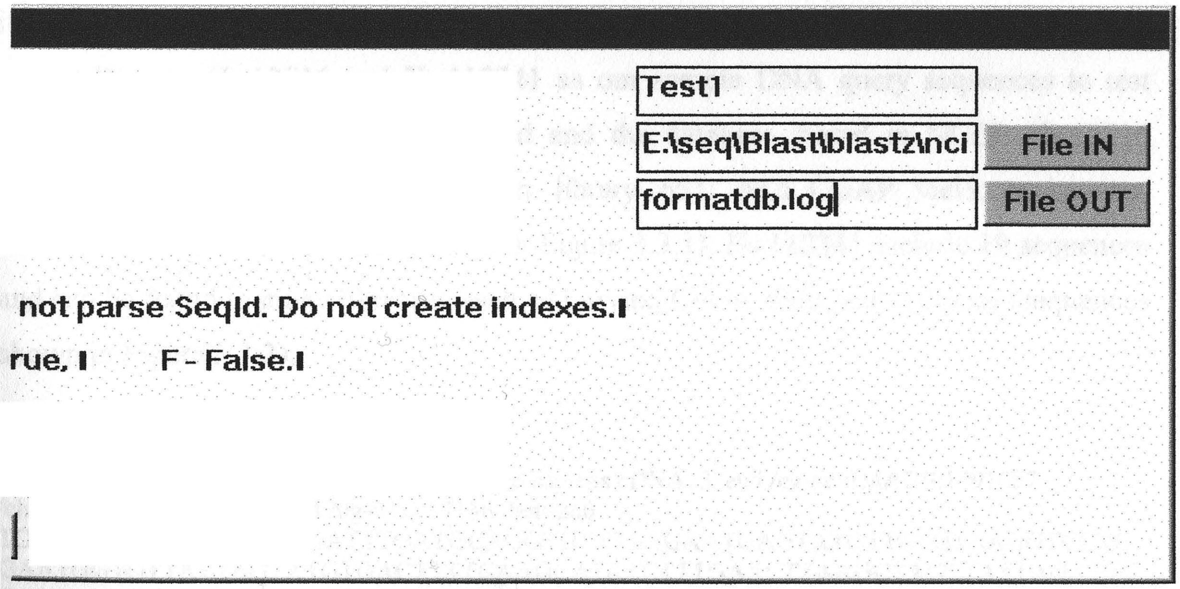
□ Parse optionsI T - True: Parse SeqId and create indexes.I F - False: Do not parse SeqId. Do not create indexe

□ Input file is database in ASN.1 format (otherwise FASTA is expected)I T - True, I F - False.I

□ ASN.1 database in binary modeI T - binary, I F - text mode.I

□ Input is a Seq-entry

Figure 3.2.2 Using NCBI Tool Formatdb to Form a Database for BLAST 2



Test1

E:\seq\Blast\blastz\nci

formatdb.log

not parse SeqId. Do not create indexes.I

rue, I F - False.I

Figure 3.2.2 Using NCBI Tool Formatdb to Form a Database for BLAST 2 (continued)

3.3. Evaluating the Database

After the local database for Blast 2 is formed we now can use Hs.12716 and Hs.112341 as query sequences to retrieve the database. Our test environment: CPU Pentium 200 Pro, 128 Mb RAM, Operating system Win 95. We use BLAST (Basic Local Alignment Search Tool) as our alignment tool to test the database because BLAST is the most popular sequence mining tool. BLAST takes a nucleotide sequence (the query sequence), and its reverse complement, and searches them against a nucleotide sequence database. It not only can process query sequences from Internet but also can be transferred from the NCBI anonymous FTP server and installed on a local machine.

We download the source codes of NCBI tool kit and compile it using Microsoft Viual C++ 5.0 to make three programs: Formatdb, BlastAll, and BlastGap.

Formatdb: Used to format the FASTA databases for both protein and DNA databases for BLAST 2.0. This must be done before blastall or blastpgp can be run locally.

BlastAll: May be used to perform all five flavors of blast comparison. (See Appendix D for Blast Family)

BlastGap: Blastpgp performs gapped blastp searches and can be used to perform iterative searches in psi-blast mode.

We use Hs.12716 and Hs.112341 as our sample DNA query sequences to test both our local database we constructed and the database stored in NCBI. Hs.12716 include two sequences found only in library 651: NCI_CGAP_Mel3; melanoma, metastatic to bowel (sequences shown in Figure 3.3.1). Hs.112341 include 19 sequences and can be found in many libraries, such as Larynx, Colon, Skin, and Adipose (sequences shown in Figure 3.3.2).

```
> 996174 gnl|UG|Hs#S996174 oj03b10.s1 Homo sapiens cDNA, 3' end /clone=IMAGE:1491067
/clone_end=3' /gb=AA937378 /ug=Hs.127136 /len=260
TCATTCAAGCAGTATAGGATTTGATGCAGGTGTTTGTGAATGAGTATGTTCTGTAAGGTCCTG
GAATGGTGTTATTAGTATGTGACTTTTCAAGCATCTCTTTGAACTTAAGCTAGTTATTAGATTT
TATTACTACTATCATTTATTTTAGCAATGTTTTATAATAATGAAAGCCATTAATCTACACATTG
TCTAGGAACAGGCTGGAAGTGAAGAGTACTTGGCTATATCATAGAAATATTTCTTGGTAACCC
TCGTGC
```

Figure 3.3.1 Sequences Data of Hs.12716

```
> 996495 gnl|UG|Hs#S996495 oj03h10.s1 Homo sapiens cDNA, 3' end /clone=IMAGE:1491139
/clone_end=3' /gb=AA937699 /gi=3095810 /ug=Hs.127136 /len=260
TCATTCAAGCAGTATAGGATTTGATGCAGGTGTTTGTGAATGAGTATGTTCTGTAAGGTCCTG
GAATGGTGTTATTAGTATGTGACTTTTCAAGCATCTCTTTGAACTTAAGCTAGTTATTAGATTT
TATTACTACTATCATTTATTTTAGCAATGTTTTATAATAATGAAAGCCATTAATCTACACATTG
TCTAGGAACAGGCTGGAAGTGAAGAGTACTTGGCTATATCATAGAAATATTTCTTGGTAACCC
TCGTGC
```

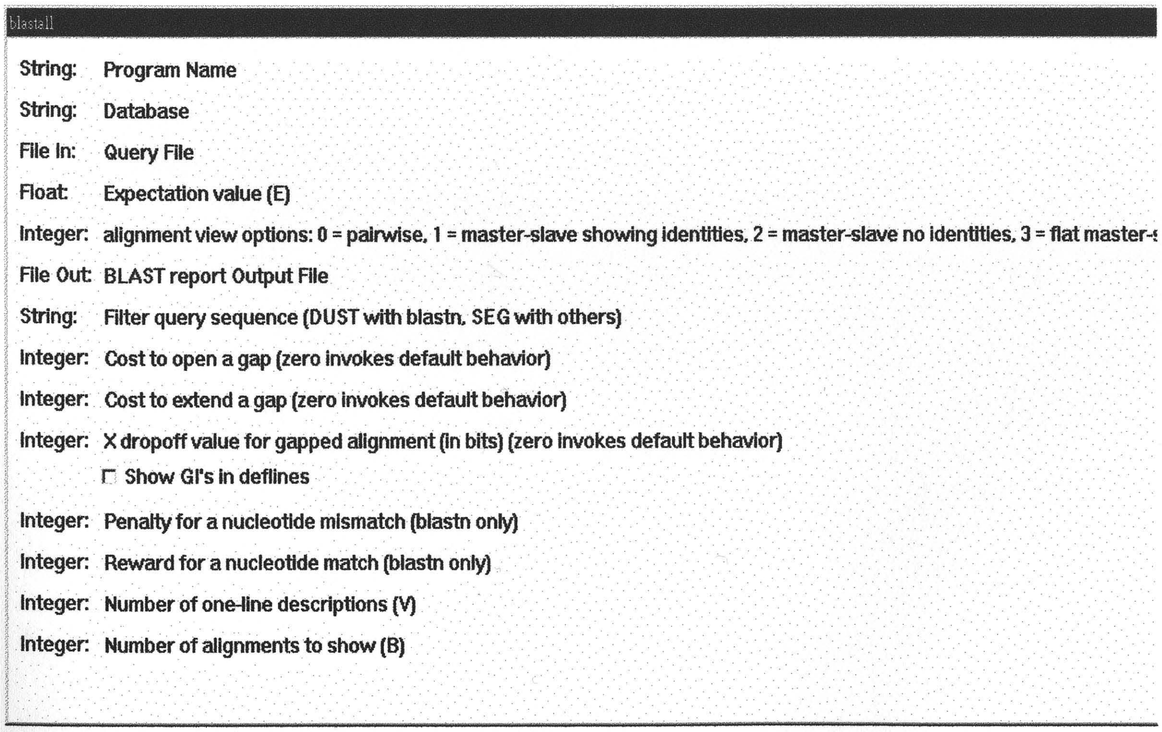
Figure 3.3.1 Sequences Data of Hs.12716 (continued)

```
> 827584 gnl|UG|Hs#S827584 nn69d08.s1 Homo sapiens cDNA /clone=IMAGE:1089135 /gb=AA586974
/gi=2397788 /ug=Hs.112341 /len=399
GGAGCAGAAGGAACTCTTTATTGGAAAGTGGATGAGAGAGGCAGCTCCAGCCGTGGGCATCC
TGAATGGGAGGAAGAATGGACAGTGTGGGAAGGGGAAGGGCAGCAGGGACTTAGGACCAGA
TGGGGCCTGTAGCTCTGGGGACGGCACAGGTGCAGCAAGGACCGGCTCCCTCTCACTGGGGA
ACGAAACAGGCCATCCCGCAAGAGCCTTCACAGCACTTCTTGATTCCTGGGCAGTCAGTATCT
TTCAAGCAGCGGTTAGGGGGATTCAACATGGCGCACCGGATCAAGATAATGGGGCAGGAGCC
AGGCTTAGTGGAGACTGGACCTTTGACTGGCTCTTGCGCTTTGACTTTAT
CTTGACCTTTAACTGAAACTTGTCCTTTAACGGGATCTT
> 341852 gnl|UG|Hs#S341852 Human gene for elafin, complete cds /cds=(66,419) /gb=D13156
/gi=219614 /ug=Hs.112341 /len=421
AGGCCAAGCTGGACTGCATAAAGATTGGTATGGCCTTAGCTCTTAGCCAAACACCTTCCTGAC
ACCATGAGGGCCAGCAGCTTCTTGATCGTGGTGGTGTTCCTCATCGCTGGGACGCTGGTTCTA
GAGGCAGCTGTCACGGGAGTTCCTGTTAAAGGTCAAGACACTGTCAAAGGCCGTGTTCCATTC
AATGGACAAGATCCCGTTAAAGGACAAGTTTCAGTTAAAGGTCAAGATAAAGTCAAAGCGCA
AGAGCCAGTCAAAGGTCCAGTCTCCACTAAGCCTGGCTCCTGCCCCATTATCTTGATCCGGTG
CGCCATGTTGAATCCCCCTAACCGCTGCTTGAAAGATACTGACTGCCCAGGAATCAAGAAGTG
CTGTGAAGGCTCTTGCGGGATGGCCTGTTTCGTTCCCCAGTGAG
> 828812 gnl|UG|Hs#S828812 nn61b01.s1 Homo sapiens cDNA /clone=IMAGE:1088329 /gb=AA583567
/gi=2368176 /ug=Hs.112341 /len=555
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGAGCAGAAGGAACTCTTTATTGGAAAGT
GGATGAGAGAGGCAGCTCCAGCCGTGGGCATCCTGAATGGGAGGAAGAATGGACAGTGTGG
GAAGGGGAAGGGCAGCAGGGACTTAGGACCAGATGGGGCCTGTAGCTCTGGGGACGGCACA
GGTGCAGCAAGGACCGGCTCCCTCTCACTGGGGAACGAAACAGGCCATCCCGCAAGAGCCTT
CACAGCACTTCTTGATTCCTGGGCAGTCAGTATCTTTCAAGCAGCGGTTAGGGGGATTCAACA
TGGCGCACCGGATCAAGATAATGGGGCAGGAGCAAGGCTTATTGGAGACTGGACCTTTTGAC
TGGCTCTTGCGCTTTTGACTTTATCTTGACCTTTAACTGGAACTTGTCCTTAAACGGGATCTTGT
CCATTGAATGGGAACACGGCCTTTGACAGTGTCTTGACCTTAAACAGGACTCCGGGAAAGTTG
CTCTAGAACAGGGTCCAGCGATGAGGACACAACACGTTCAGACTGCTGGCCC
-----------------------------------------------------------------------------------------------------
------------------------------------------ omitted ------------------------------------------------
-----------------------------------------------------------------------------------------------------
> 638961 gnl|UG|Hs#S638961 EST22235 Homo sapiens cDNA, 5' end /clone=ATCC:120504
/clone_end=5' /gb=AA319941 /gi=1972269 /ug=Hs.112341· /len=299
TGGTGTTCCTCATCGCTGGGACGCTGGTTCTAGAGGCAGCTGTCACGGGAGTTCCTGTTAAAG
GTCAAGACACTGTCAAAGGCCGTGTTCCATTCAATGGACAAGATCCCGTTAAAGGACAAGTTT
CAGTTAAAGGTCAAGATAAAGTCAAAGCGCAAGAGCCAGTCAAAGGTCCAGTCTCCACTAAG
CCTGGNTCCTGCCCCATTATCTTGATCCGGTGCGCCATGTTGAATNCCCCTAACCGCTGCTTGA
AAGATACTTGACTNCCCAGGGGATCAAGAAGTGCTGTGAAGGCTCTT
```

Figure 3.3.2 Sequences Data of Hs.112341

3.4. Results

This stand alone database we created can produce results/reports that look very similar to those generated by the original BLAST engine; however, in our case the actual results are quite different. Figure 3.4.1 shows the BLAST 2 query screen in our local machine. And Figure 3.4.2 and Figure 3.4.3 show the results from our local database for Hs.12716 and Hs.112341 respectively. Also we use Hs.12716 and Hs.112341 as query sequences to do the alignment in NCBI homepage via the Internet. Figure 3.4.4 and Figure 3.4.5 show the query screen in NCBI homepage and Appendix A.1 and Appendix A.2 show the query results for Hs.12716 and Hs.112341 respectively.



blastall

String: Program Name

String: Database

File In: Query File

Float: Expectation value (E)

Integer: alignment view options: 0 = pairwise, 1 = master-slave showing identities, 2 = master-slave no identities, 3 = flat master-

File Out: BLAST report Output File

String: Filter query sequence (DUST with blastn, SEG with others)

Integer: Cost to open a gap (zero invokes default behavior)

Integer: Cost to extend a gap (zero invokes default behavior)

Integer: X dropoff value for gapped alignment (in bits) (zero invokes default behavior)

□ Show GI's in deflines

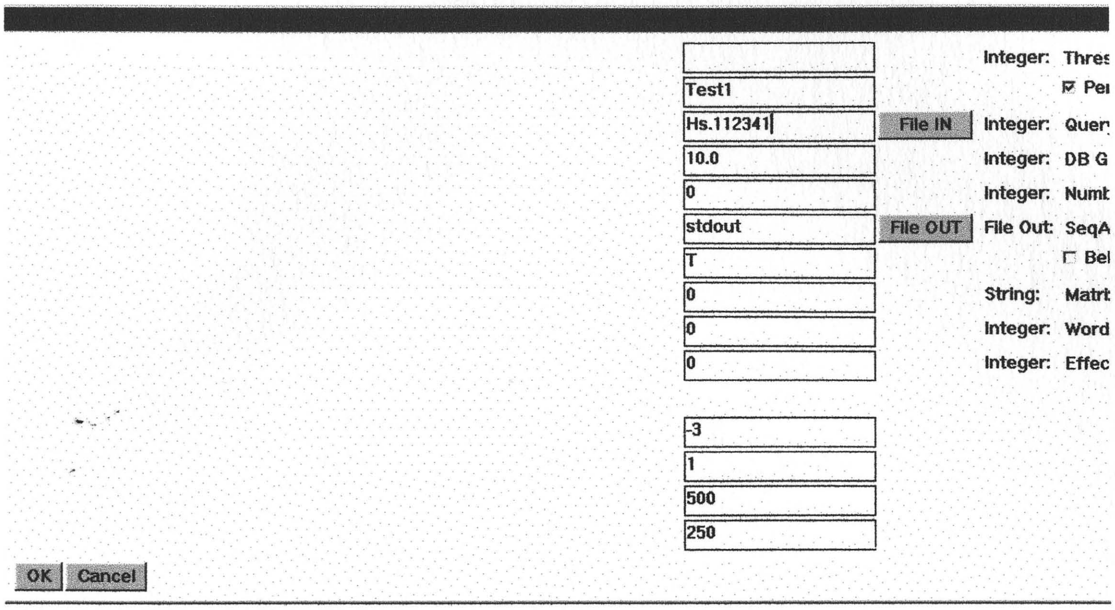
Integer: Penalty for a nucleotide mismatch (blastn only)

Integer: Reward for a nucleotide match (blastn only)

Integer: Number of one-line descriptions (V)

Integer: Number of alignments to show (B)

Figure 3.4.1 Screen Shot of BLAST 2 Query Screen



Test1

Hs.112341 File IN

10.0

0

stdout File OUT

T

0

0

0

-3

1

500

250

Integer: Thres

☑ Per

Integer: Quer

Integer: DB G

Integer: Numb

File Out: SeqA

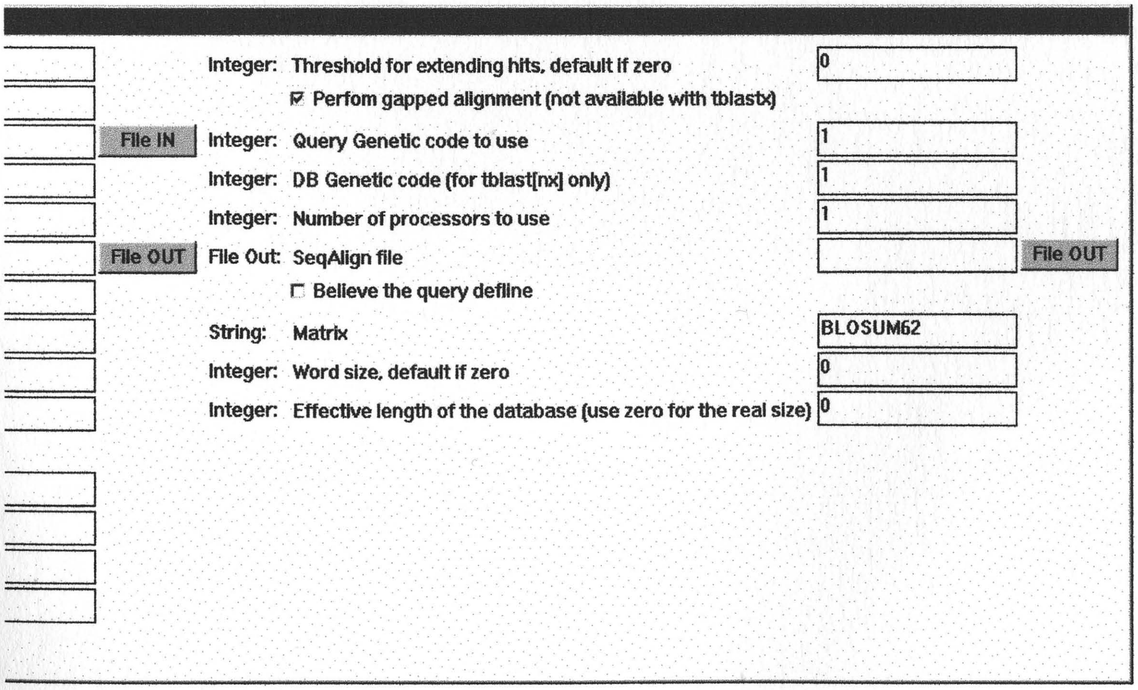
□ Bel

String: Matri

Integer: Word

Integer: Effec

Figure 3.4.1 Screen Shot of BLAST 2 Query Screen (continued)



Integer: Threshold for extending hits, default if zero 0

☑ Perfom gapped alignment (not available with tblastx)

Integer: Query Genetic code to use 1

Integer: DB Genetic code (for tblast[nx] only) 1

Integer: Number of processors to use 1

File Out: SeqAlign file File OUT

□ Believe the query defline

String: Matrix BLOSUM62

Integer: Word size, default if zero 0

Integer: Effective length of the database (use zero for the real size) 0

Figure 3.4.1 Screen Shot of BLAST 2 Query Screen (continued)

```
BLASTN 2.0.4 [Feb-24-1998]
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A.
Sch&auml;ffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

Query= 996174 gnl|UG|Hs#S996174 oj03b10.s1 Homo sapiens cDNA, 3' end
/clone=IMAGE:1491067 /clone_end=3' /gb=AA937378 /ug=Hs.127136 /len=260
      (260 letters)

Database: Test1 sequences
        9 sequences; 89,378,749 total letters

Searchingdone


                                                     Score    E
Sequences producing significant alignments:          (bits)  Value

nci_cgap_mel3.out                                      40     1.4
nci_cgap_pns1.out                                         31     2.6


> nci_cgap_mel3.out                Length = 16852185

Score = 40.1 bits (237), Expect = 2.4
Identities = 183/187 (98%), Positives = 183/187 (98%)


Query: 2   aatttatcatagaatatttcttcctaatttagatatcattaagcggtatacccattaaga 61
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 165 aatttatcatagaatatttcttcctaatttagatatcattaagcggtatacccattaaga 224

Query: 62 cattcaattatcatcaaatatcttcctaattggggaatcattatgctttatacccatcaa 121
           |||||||||||||||||||||||||||||||||||||||||| |||||||||||||||||
Sbjct: 225 cattcaattatcatcaaatatcttcctaattggggaatcattaagctttatacccatcaa 284

Query: 122 taattaatcatatcatttatctacctaggttctgcaatcatttaggcttatacgcatcac 181
           |||||||||||||||||||||| ||||||||||||| |||||||||||||||||||||||
Sbjct: 225  taattaatcatatcatttatcttcctaggttctgcagtcatttaggcttatacgcatcac 284

Query: 182 tatgtca 188
           |||| ||
Sbjct: 285  tatgcca 291
```

Figure 3.4.2 Query Results of Hs.12716 in Our Local Machine

```
> nci_cgap_pns1.out              Length = 1543428

Score = 31.1 bits (65), Expect = 2.6
Identities = 33/33 (100%), Positives = 33/33 (100%)

Query: 34 tatcattaagcggtatacccattaagacattc 66
           |||||||||||||||||||||||||||||||||
Sbjct: 117 tatcattaagcggtatacccattaagacattc 149


CPU time:    51.82 user secs.    11.02 sys. secs    62.84 total secs.

  Database: Test1
    Posted date:  Jun 21, 1998  1:41 PM
  Number of letters in database: 89,378,749
  Number of sequences in database:  9

Lambda     K      H
   1.37    0.711    1.31

Gapped
Lambda     K      H
   1.37    0.711    1.31


Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 31411
Number of Sequences: 9
Number of extensions: 31411
Number of successful extensions: 2155
Number of sequences better than 10: 2
length of query: 260
length of database: 89378749
effective HSP length: 188
effective length of query: 241
effective length of database: 284
effective search space: 3371528
T: 0
A: 0
X1: 6 (11.9 bits)
X2: 25 (49.6 bits)
S1: 0 ( 0.5 bits)
S2: 17 (34.2 bits)
```

Figure 3.4.2 Query Results of Hs.12716 in Our Local Machine (continued)

```
BLASTN 2.0.4 [Feb-24-1998]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A.
Sch&auml;ffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

Query= 827584 gnl|UG|Hs#S827584 nn69d08.s1 Homo sapiens cDNA
/clone=IMAGE:1089135 /gb=AA586974 /gi=2397788 /ug=Hs.112341 /len=399
        (399 letters)

Database: Test1 sequences
           9 sequences; 89,378,749 total letters

Searchingdone

                                                          Score    E
Sequences producing significant alignments:               (bits) Value

nci_cgap_hn3.out                                             56  4e-06
nci_cgap_hn1.out                                             42  0.061
nci_cgap_pns1.out                                            38  0.95


> nci_cgap_hn3.out          Length = 1064786

Score = 56 bits (144), Expect = 4e-06
Identities = 106/106 (100%), Positives = 106/106 (100%)


Query: 54  tgggcatcctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggact 113
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1143 tgggcatcctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggact 1202


Query: 114 taggaccagatggggcctgtagctctggggacggcacaggtgcagc 159
            ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1203 taggaccagatggggcctgtagctctggggacggcacaggtgcagc 1248
```

Figure 3.4.3 Query Results of Hs.112341 in Our Local Machine

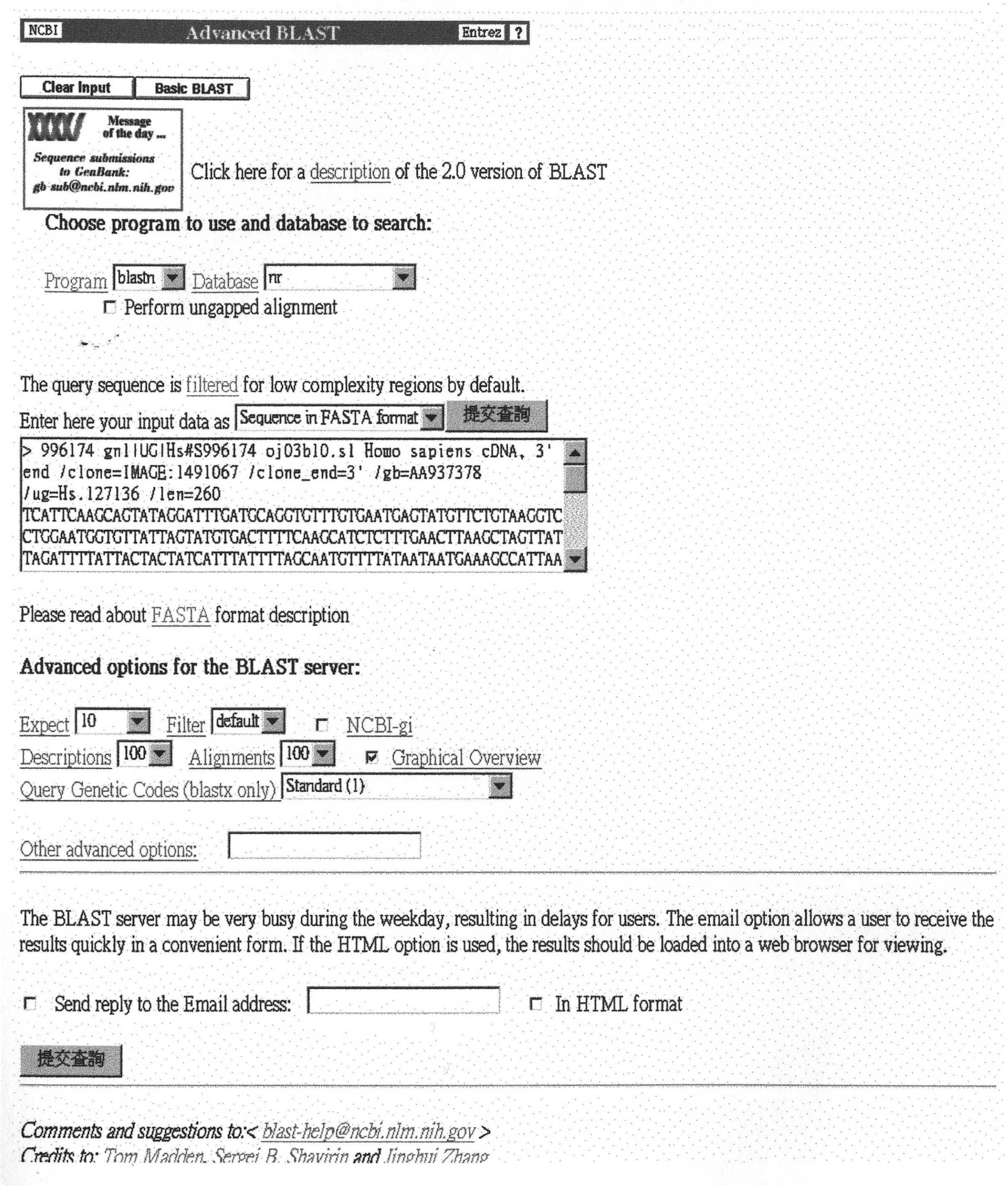
```
> nci_cgap_hn1.out          Length = 276488

Score =  42 bits (125), Expect = 0.061
Identities = 89/89 (100%), Positives = 89/89 (100%)


Query: 62  ctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggacttaggacca 121
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 29414 ctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggacttaggacca 29473


Query: 122  gatggggcctgtagctctg 140
           |||||||||||||||||||
Sbjct: 29474  gatggggcctgtagctctg   29492


> nci_cgap_pns1.out                 Length = 1543428

Score =  38 bits (108), Expect = 0.95
Identities = 72/77 (94%), Positives = 72/77 (94%)


Query: 175  ctcactggggaacgaaacaggccatcccgcaagagccttcacagcacttcttgattcctg 234
           | ||||||||||||||||||||||  ||||||||||||||||||||||||||||||||| 
Sbjct: 1132 cacactggggaacgaaacaggccatttcgcaagagccttcacagcacttcttgattccta 1191


Query: 235  ggcagtcagtatctttc 251
           |||||||||||||||||
Sbjct: 1192 ggcagtcagtatctttc  1208
```

Figure 3.4.3 Query Results of Hs.112341 in Our Local Machine (continued)

```
CPU time:   31.51 user secs.    1.10 sys. secs   32.61 total secs.

 Database: Test1 sequences
  Posted date: Jun 26, 1998  8:01 AM
 Number of letters in database: 89,378,749
 Number of sequences in database:  9

Lambda     K      H
   1.37    0.711    1.31

Gapped
Lambda     K      H
   1.37    0.711    1.31

Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 40416
Number of Sequences: 9
Number of extensions: 40416
Number of successful extensions: 3992
Number of sequences better than 10: 2
length of query: 399
length of database: 89378749
effective HSP length: 19
effective length of query: 167
effective length of database: 29492
effective search space: 4237546
T: 0
A: 0
X1: 6 (11.9 bits)
X2: 25 (49.6 bits)
S1: 0 ( 0.5 bits)
S2: 18 (34.2 bits)
```

Figure 3.4.3 Query Results of Hs.112341 in Our Local Machine (continued)

The query sequences for this searching have been filtered. Filtering eliminates low complexity regions that commonly give spuriously high scores that reflect compositional bias rather than significant position-by-position alignment. Filtering can eliminate these potentially confounding matches (e.g., hits against proline-rich regions or poly-A tails) from the blast reports, leaving regions whose blast statistics reflect the specificity of their pairwise alignment.



Sequence submissions to GenBank: gb-sub@ncbi.nlm.nih.gov

Click here for a description of the 2.0 version of BLAST

Choose program to use and database to search:

Program: blastn Database: nr

□ Perform ungapped alignment

The query sequence is filtered for low complexity regions by default.

Enter here your input data as: Sequence in FASTA format

```
> 996174 gnl|UG|Hs#S996174 oj03b10.s1 Homo sapiens cDNA, 3'
end /clone=IMAGE:1491067 /clone_end=3' /gb=AA937378
/ug=Hs.127136 /len=260
TCATTCAAGCAGTATAGGATTTGATGCAGGTGTTTGTGAATGAGTATGTTCTGTAAGGTC
CTGGAATGGTGTTATTAGTATGTGACTTTTCAAGCATCTCTTTGAACTTAAGCTAGTTAT
TAGATTTTATTACTACTATCATTTATTTTAGCAATGTTTTATAATAATGAAAGCCATTAA
```

Please read about FASTA format description

Advanced options for the BLAST server:

Expect: 10 Filter: default □ NCBI-gi

Descriptions: 100 Alignments: 100 ☑ Graphical Overview

Query Genetic Codes (blastx only): Standard (1)

Other advanced options:

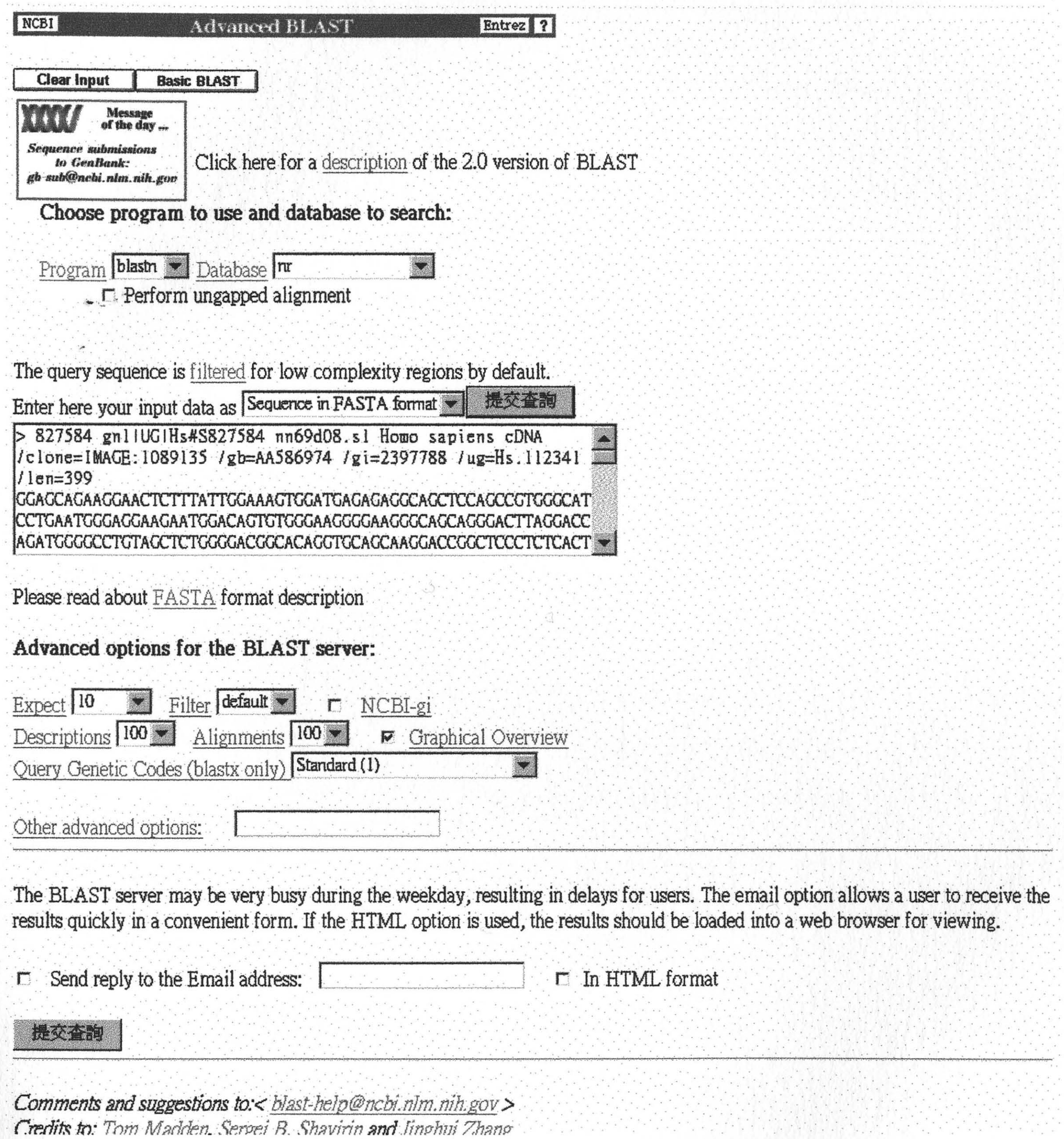
The BLAST server may be very busy during the weekday, resulting in delays for users. The email option allows a user to receive the results quickly in a convenient form. If the HTML option is used, the results should be loaded into a web browser for viewing.

□ Send reply to the Email address: □ In HTML format

Comments and suggestions to:< blast-help@ncbi.nlm.nih.gov >

Credits to: Tom Madden, Sergei B. Shavirin and Jinghui Zhang

Figure 3.4.4 Hs.12716, the NCBI HomePage Query Screen



Message of the day ...

Sequence submissions to GenBank: gb-sub@ncbi.nlm.nih.gov

Click here for a description of the 2.0 version of BLAST

Choose program to use and database to search:

Program blastn Database nr

□ Perform ungapped alignment

The query sequence is filtered for low complexity regions by default.

Enter here your input data as Sequence in FASTA format

```
> 827584 gnl|UG|Hs#S827584 nn69d08.s1 Homo sapiens cDNA
/clone=IMAGE:1089135 /gb=AA586974 /gi=2397788 /ug=Hs.112341
/len=399
GGAGCAGAAGGAACTCTTTATTGGAAAGTGGATGAGAGAGGCAGCTCCAGCCGTGGGCAT
CCTGAATGGGAGGAAGAATGGACAGTGTGGGAAGGGGAAGGGCAGCAGGGACTTAGGACC
AGATGGGGCCTGTAGCTCTGGGGACGGCACAGGTGCAGCAAGGACCGGCTCCCTCTCACT
```

Please read about FASTA format description

Advanced options for the BLAST server:

Expect 10 Filter default □ NCBI-gi

Descriptions 100 Alignments 100 ☑ Graphical Overview

Query Genetic Codes (blastx only) Standard (1)

Other advanced options:

The BLAST server may be very busy during the weekday, resulting in delays for users. The email option allows a user to receive the results quickly in a convenient form. If the HTML option is used, the results should be loaded into a web browser for viewing.

□ Send reply to the Email address: □ In HTML format

Comments and suggestions to:< blast-help@ncbi.nlm.nih.gov >

Credits to: Tom Madden, Sergei B. Shavirin and Jinghui Zhang

Figure 3.4.5 Hs.112341, the NCBI HomePage Query Screen

CHAPTER 4

COMPARING MOTIFS RETRIEVED FROM PROSITE WITH MOTIFS FOUND BY SDISCOVER

In some cases the sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by overall sequence alignment [14,45,52], but it can be identified by the occurrence in its sequence of a particular cluster of residue types which is variously known as a pattern, motif, signature, or fingerprint [37]. These motifs arise because of particular requirements on the structure of specific region(s) of a protein which may be important, for example, for their binding properties or for their enzymatic activity. These requirements impose very tight constraints on the evolution of those limited (in size) but important portion(s) of a protein sequence. The use of protein sequence patterns (or motifs) to determine the function(s) of proteins is becoming very rapidly one of the essential tools of sequence analysis.

Currently, the largest collection of sequence motifs in the world is PROSITE which contains a lot of families of protein [9,28]. PROSITE can be accessed via either the ExPASy WWW server or anonymous FTP site. In comparing the difference between motifs that are stored in Prosite database and motifs that we found by using SDISCOVER tool, we use the protein family, which include 4 protein sequences, COAGULATION FACTOR X PRECURSOR as our first sample sequences (Figure 4.1).

```
>gi|119760|sp|P25155|FA10_CHICK COAGULATION FACTOR X PRECURSOR (STUART FACTOR)
(VIRUS ACTIVATING PROTEASE) (VAP)
MAGRLLLLLLCAALPDELRAEGGVFIKKESADKFLERTKRANSFLEEMKQGNIERECNEERCSKEE
AREAFEDNEKTEEFWNIYVDGDQCSSNPCHYGGQCKDGLGSYTCSCLDGYQGKNCEFVIPKYCKI
NNGDCEQFCSIKKSVQKDVVCSCTSGYELAEDGKQCVSKVKYPCGKVLMKRIKRSVILPTNSNTN
ATSDQDVPSTNGSILEEVFTTTTESPTPPPRNGSSITDPNVDTRIVGGDECRPGECPWQAVLINEKGE
EFCGGTILNEDFILTAAHCINQSKEIKVVVGEVDREKEEHSETTHTAEKIFVHSKYIAETYDNDIALI
KLKEPIQFSEYVVPACLPQADFANEVLMNQKSGMVSGFGREFEAGRLSKRLKVLEVPYVDRSTCK
QSTNFAITENMFCAGYETEQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGKYGVYTKL
SRFLRWVRTVMRQK
```

Figure 4.1 Sequences of COAGULATION FACTOR X PRECURSOR

```
>gi|119761|sp|P00742|FA10_HUMAN COAGULATION FACTOR X PRECURSOR
(STUART FACTOR)
MGRPLHLVLLSASLAGLLLLGESLFIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEE
AREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCS
LDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAP
DSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGF
CGGTILSEFYILTAAHCLYQAKRFKVRVGDRNTEQEEGGEAVHEVEVVIKHNRFTKETYDFDIAVL
RLKTPITFRMNVAPACLPERDWAESTLMTQKTGIVSGFGRTHEKGRQSTRLKMLEVPYVDRNSCK
LSSSFIITQNMFCAGYDTKQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVT
AFLKWIDRSMKTRGLPKAKSHAPEVITSSPLK
>gi|119759|sp|P00743|FA10_BOVIN COAGULATION FACTOR X PRECURSOR (STUART FACTOR)
MAGLLHLVLLSTALGGLLRPAGSVFLPRDQAHRVLQRARRANSFLEEVKQGNLERECLEEACSLE
EAREVFEDAEQTDEFWSKYKDGDQCEGHPCLNQGHCKDGIGDYTCTCAEGFEGKNCEFSTREICS
LDNGGCDQFCREERSEVRCSCAHGYVLGDDSKSCVSTERFPCGKFTQGRSRRWAIHTSEDALDAS
ELEHYDPADLSPTESSLDLLGLNRTEPSAGEDGSQVVRIVGGRDCAEGECPWQALLVNEENEGFC
GGTILNEFYVLTAAHCLHQAKRFTVRVGDRNTEQEEGNEMAHEVEMTVKHSRFVKETYDFDIAV
LRLKTPIRFRRNVAPACLPEKDWAEATLMTQKTGIVSGFGRTHEKGRLSSTLKMLEVPYVDRSTC
KLSSSFTITPNMFCAGYDTQPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKFGVYTKV
SNFLKWIDKIMKARAGAAGSRGHSEAPATWTVPPPLPL
>gi|180336 coagulation factor X precursor
LLGESLFIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNK
YKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSV
VCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTE
NPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSEFYILTAAHCL
YQAKRFEGDRNTEQEEGGEAVHEVEVVIKHNRFTKETYDFDIAVLRLKTPITFRMNVAPACLPER
DWAESTLMTQKTGIVSGFGRTHEKGRQSTRLKMLEVPYVDRNSCKLSSSFIITQNMFCAGYDTKQ
EDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTAFLKWIDRSMKTRGLPKAK
SHAPEVITSSPLK
```

Figure 4.1 Sequences of COAGULATION FACTOR X PRECURSOR (continued)

By using SDISCOVER, the query parameters and results are as follows:

Minimum length: 10

Minimum occurrence number: 2

Number of mutations allowed: 1

Total number of sequences: 4

Motifs found: 51147

motifs checked: 52759

After sorted: 166

The motifs, after sorted (eliminating substrings), are shown in Figure 4.2. A symbol # followed by a number indicates the motif number and another number preceding the motif indicates the occurrence numbers.

```
#1        2
          *HLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYT
CTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGK
QTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQE
CKDGECPWQAL*
# 2       2
          *YKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCH
EEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAA
DLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSEFYILT
AAHCLYQAKR*
#3        2
          *YEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNC
ELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQ
ATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALL
INEENEGFCGG*
#4        2
          *WNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCD
QFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPY
DAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSEF
YILTAAHCLYQ*
#5        2
          *QANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDG
DQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCA
RGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLL
DFNQTQPERGD*
#6        2
          *NKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQF
CHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDA
ADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSEFYIL
TAAHCLYQA*
#7        2
          *NNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQ
CETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARG
YTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDF
NQTQPERGDNN*
#8        2
          *NEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGD
CDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWK
PYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILS
EFYILTAAHC*
#9        2
          *NILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQC
ETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGY
TLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFN
QTQPERGDNNL*
#10       2
          *NSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQG
KCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKA
CIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGD
NNLTRIVGGQEC*
```

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER

```
#11      2
         *TRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQ
NQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADN
GKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPE
RGDNNLTRIVGG*
#12      2
         *TNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNG
DCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITW
KPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTIL
SEFYILTAAH*
#13      2
         *TCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEG
KNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRS
VAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPW
QALLINEENEGF*
#14      2
         *SLFIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFW
NKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQN
SVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPT
ENPFDLLDFNQ*
#15      2
         *SDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSL
DNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPD
SITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCG
GTILSEFYILT*
#16      2
         *SYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKN
CELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVA
QATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQA
LLINEENEGFCG*
#17      2
         *SFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGK
CKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACI
PTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNN
LTRIVGGQECK*
#18      2
         *KGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLG
EYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYP
CGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVG
GQECKDGECPWQ*

#19      2
         *KYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFC
HEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDA
ADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSEFYIL
TAAHCLYQAK*
#20      2
         *KDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHE
EQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAAD
LDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSEFYILTA
AHCLYQAKRF*
```

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER (continued)

```
#21       2
        *KTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDN
GDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSIT
WKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGT
ILSEFYILTAA*
#22       2
        *KKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGL
GEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPY
PCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIV
GGQECKDGECPW*
#23       2
        *IRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKY
KDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVV
CSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENP
FDLLDFNQTQP*
#24       2
        *ILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCE
TSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYT
LADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQ
TQPERGDNNLT*
#25       2
        *FWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCD
QFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPY
DAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSEF
YILTAAHCLY*
#26       2
        *FEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKL
CSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGE
APDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENE
GFCGGTILSEFY*
#27       2
        *FLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKC
KDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIP
TGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNL
TRIVGGQECKD*
#28       2
        *FIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNK
YKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSV
VCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTEN
PFDLLDFNQTQ*
#29       2
        *VTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPC
QNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLAD
NGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQP
ERGDNNLTRIVG*

#30       2
        *VGDRNTEQEEGGEAVHEVEVVIKHNRFTKETYDFDIAVLRLKTPITFRMNVAPACLPERD
WAESTLMTQKTGIVSGFGRTHEKGRQSTRLKMLEVPYVDRNSCKLSSSFIITQNMFCAGYDTKQE
DACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTAFLKWIDRSMKTRGLPKAKS
HAPEVITSSPL*
```

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER (continued)

```
#31     2
        *VFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRK
LCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSG
EAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENE
GFCGGTILSEF*
#32     2
        *ETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFE
GKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKR
SVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPW
QALLINEENEG*
#33     2
        *EVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTR
KLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSS
GEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEE
NEGFCGGTILSE*
#34     2
        *EQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKD
GDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSC
ARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDL
LDFNQTQPERG*
#35     2
        *EFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDC
DQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKP
YDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSE
FYILTAAHCL*
#36     2
        *EDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLC
SLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEA
PDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGF
CGGTILSEFYI*
#37     2
        *EAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCEL
FTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQAT
SSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLIN
EENEGFCGGTI*
#38     2
        *ECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCL
EGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLE
RRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDG
ECPWQALLINE*
#39     2
        *EMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKD
GLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTG
PYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRI
VGGQECKDGEC*
#40     2
        *EETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGF
EGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRK
RSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECP
WQALLINEENE*
```

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER (continued)

```
#41      2
        *EEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCE
LFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQA
TSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLI
NEENEGFCGGT*
#42      2
        *EEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCK
DGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPT
GPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLT
RIVGGQECKDGE*
#43      2
        *ERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCT
CLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQT
LERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECK
DGECPWQALLI*
#44      2
        *ESLFIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFW
NKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQN
SVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPT
ENPFDLLDFN*
#45      2
        *EGDRNTEQEEGGEAVHEVEVVIKHNRFTKETYDFDIAVLRLKTPITFRMNVAPACLPERD
WAESTLMTQKTGIVSGFGRTHEKGRQSTRLKMLEVPYVDRNSCKLSSSFIITQNMFCAGYDTKQE
DACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTAFLKWIDRSMKTRGLPKAKS
HAPEVITSSPL*
#46      2
        *DSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCS
LDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAP
DSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFC
GGTILSEFYIL*
#47      2
        *DGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEE
QNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADL
DPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSEFYILTAA
HCLYQAKRFE*
#48      2
        *DGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEE
QNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADL
DPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTILSEFYILTAA
HCLYQAKRFK*
#49      2
        *DKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLD
NGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSI
TWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGG
TILSEFYILTA*
#50      2
        *CMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLE
GFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLER
RKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGE
CPWQALLINEE*
```

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER (continued)

```
#51     2
        *CSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGK
NCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSV
AQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQ
ALLINEENEGFC*
#52     2
        *LFIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWN
KYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNS
VVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTE
NPFDLLDFNQT*
#53     2
        *LARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCET
SPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTL
ADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQT
QPERGDNNLTR*
#54     2
        *LGESLFIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNE
FWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEE
QNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADL
DPTENPFDLLD*
#55     2
        *LEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKC
KDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIP
TGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNL
TRIVGGQECKDG*
#56     2
        *LERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTC
TCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQ
TLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQEC
KDGECPWQALL*
#57     2
        *LLGESLFIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTN
EFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHE
EQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAAD
LDPTENPFDLL*
#58     2
        *LLLGESLFIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKT
NEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCH
EEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAA
DLDPTENPFDL*
#59     2
        *RVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSP
CQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLA
DNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQ
PERGDNNLTRIV*
#60     2
        *RREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKY
KDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVV
CSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENP
FDLLDFNQTQPE*
```

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER (continued)

```
#61        2
        *REVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFT
RKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSS
SGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEE
NEGFCGGTILS*
#62        2
        *REQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYK
DGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVC
SCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPF
DLLDFNQTQPER*

#63        2
        *RECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTC
LEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTL
ERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKD
GECPWQALLIN*
#64        2
        *RANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQN
QGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNG
KACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPER
GDNNLTRIVGGQ*
#65        2
        *GHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGE
YTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPC
GKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGG
QECKDGECPWQA*
#66        2
        *GESLFIRREQANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEF
WNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQ
NSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDP
TENPFDLLDF*
#67        2
        *GDRNTEQEEGGEAVHEVEVVIKHNRFTKETYDFDIAVLRLKTPITFRMNVAPACLPERDW
AESTLMTQKTGIVSGFGRTHEKGRQSTRLKMLEVPYVDRNSCKLSSSFIITQNMFCAGYDTKQEDA
CQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTAFLKWIDRSMKTRGLPKAKSHA
PEVITSSPLK*
#68        2
        *ARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETS
PCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTL
ADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQT
QPERGDNNLTRI*
#69        2
        *AREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELF
TRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATS
SSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINE
ENEGFCGGTIL*
#70        2
        *ANNILARVTRANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGD
QCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCAR
GYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLD
FNQTQPERGDN*
```

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER (continued)

#71	2	*ANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQE*
#72	2	*MEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEEN*
#73	2	*MKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEGFEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSSSGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECP*
#74	3	*QPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGK*
#75	3	*EDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKFG*
#76	3	*EDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYG*
#77	3	*QEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGK*
#78	4	*DACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYG*
#79	3	*DACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGKYG*
#80	4	*DACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGK*
#81	2	*KDTYFVTGIVSWGEGCARKGKFGVYTK*
#82	3	*KDTYFVTGIVSWGEGCARKGKYGIYTK*
#83	4	*KDTYFVTGIVSWGEGCARKGKYGVYTK*
#84	3	*DWAEATLMTQKTGIVSGFGRTHEKGR*
#85	3	*DWAESTLMTQKTGIVSGFGRTHEKGR*
#86	3	*TLMTQKTGIVSGFGRTHEKGRLS*
#87	3	*TLMTQKTGIVSGFGRTHEKGRQS*
#88	4	*KDTYFVTGIVSWGEGCARKGKFG*
#89	3	*GECPWQALLVNEENEGFCGGTIL*
#90	3	*GECPWQALLINEENEGFCGGTIL*
#91	3	*KETYDFDIAVLRLKTPIRFR*
#92	3	*KETYDFDIAVLRLKTPITFR*
#93	3	*RFVKETYDFDIAVLRLKTPI*
#94	3	*RFTKETYDFDIAVLRLKTPI*
#95	2	*QAKRFTVRVGDRNTEQEEG*
#96	2	*QAKRFKVRVGDRNTEQEEG*
#97	3	*NEENEGFCGGTILNEFY*
#98	3	*NEENEGFCGGTILSEFY*
#99	4	*QEDACQGDSGGPHVTR*
#100	3	*QKDACQGDSGGPHVTR*
#101	3	*YTCTCAEGFEGKNCE*
#102	3	*YTCTCLEGFEGKNCE*
#103	4	*KDACQGDSGGPHVTR*
#104	2	*VRVGDRNTEQEEGNE*
#105	2	*VRVGDRNTEQEEGGE*
#106	2	*LKMLEVPYVDRSTCK*
#107	2	*LKVLEVPYVDRSTCK*

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER (continued)

#108	3	*NVAPACLPEKDWAE*
#109	3	*NVAPACLPERDWAE*
#110	3	*STRLKMLEVPYVDR*
#111	2	*LEVPYVDRSTCKLS*
#112	2	*LEVPYVDRSTCKQS*
#113	3	*STLKMLEVPYVDR*
#114	3	*VGDRNTEQEEGGE*
#115	3	*EFWSKYKDGDQCE*
#116	3	*EFWNKYKDGDQCE*
#117	3	*LKMLEVPYVDRNS*
#118	3	*LTAAHCLHQAKRF*
#119	3	*LTAAHCLYQAKRF*
#120	2	*RANSFLEEVKQGN*
#121	2	*RANSFLEEMKQGN*
#122	3	*TKRANSFLEEMK*
#123	3	*ITPNMFCAGYDT*
#124	3	*ITQNMFCAGYDT*
#125	3	*FRRNVAPACLPE*
#126	3	*FRMNVAPACLPE*
#127	3	*CSLEEAREVFED*
#128	3	*CSLDNGGCDQFC*
#129	3	*CSLDNGDCDQFC*
#130	3	*CSYEEAREVFED*
#131	4	*LKMLEVPYVDRS*
#132	3	*RANSFLEEMKKG*
#133	4	*RANSFLEEMKQG*
#134	4	*RLKMLEVPYVDR*
#135	3	*RLKVLEVPYVDR*
#136	4	*GECPWQALLINE*
#137	3	*GECPWQAVLINE*
#138	3	*GDRNTEQEEGNE*
#139	3	*NMFCAGYDTKQ*
#140	3	*TRANSFLEEMK*
#141	2	*FCGGTILNEDF*
#142	3	*EFYVLTAAHCL*
#143	3	*EFYILTAAHCL*
#144	2	*EEFCGGTILNE*
#145	3	*EGDRNTEQEEG*
#146	4	*EGFCGGTILNE*
#147	3	*EGFEGKNCELF*
#148	4	*LKVLEVPYVDR*
#149	3	*NMFCAGYDTQ*
#150	3	*KYKDGDQCEG*
#151	3	*KYKDGDQCET*
#152	4	*ITPNMFCAGY*
#153	4	*ITQNMFCAGY*
#154	4	*ITENMFCAGY*
#155	4	*FCGGTILNEF*
#156	3	*EEAREVFEDA*
#157	3	*EEAREVFEDS*

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER (continued)

#158	3	*EGECPWQALL*
#159	3	*EGFEGKNCEF*
#160	3	*DGECPWQALL*
#161	3	*CKLSSSFTIT*
#162	3	*CKLSSSFIIT*
#163	3	*CKDGLGEYTC*
#164	3	*CKDGLGSYTC*
#165	4	*LEVPYVDRNS*
#166	4	*RANSFLEEVK*

Figure 4.2 Motifs of Protein Sequences Found by SDISCOVER (continued)

We use the tools in PRATT homepage (http://www2.ebi.ac.uk) to retrieve the motifs stored in Prosite database and the query results are shown in Appendix A.3. The outputs of protein motifs are in Prosite format. Here is a brief description of Prosite format.

- The symbol `x' is used for a position where any amino acid is accepted.
- Ambiguities are indicated by listing the acceptable amino acids for a given position, between square parentheses `[ ]'. For example: [ALT] stands for Ala or Leu or Thr.
- Ambiguities are also indicated by listing between a pair of curly brackets `{ }' the amino acids that are not accepted at a given position. For example: {AM} stands for any amino acid except Ala and Met.
- Each element in a pattern is separated from its neighbor by a `-'.
- Repetition of an element of the pattern can be indicated by following that element with a numerical value or a numerical range between parenthese. Examples: x(3) corresponds to x-x-x, x(2,4) corresponds to x-x or x-x-x or x-x-x-x.
- When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern either starts with a `<' symbol or respectively ends with a `>' symbol.
- A period ends the pattern.

From Appendix A.3 query results we can find out that we retrieve 50 motifs from the query sequences. And these are marked by numbers from A 1 to x 50 as following:

```
        fitness      hits(seqs)      Pattern
A  1: 199.4173     4(  4)  N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-
G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G
Occurrences: 4(4)
B  2: 198.8288     4(  4)  T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-
G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K
Occurrences: 4(4)
C  3: 198.8288     4(  4)  I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-
G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G
Occurrences: 4(4)
D  4: 198.4721     4(  4)  A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-
T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y-T-K
Occurrences: 4(4)
E  5: 198.4721     4(  4)  C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-
V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y-T
Occurrences: 4(4)
F  6: 198.4721     4(  4)  F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-
V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y
Occurrences: 4(4)
G  7: 197.5681     4(  4)  D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-
I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y-T-K-[LV]-[ST]-x-F-L-[KR]-W-[IV]
Occurrences: 4(4)
H  8: 194.6588     4(  4)  F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-
C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R
Occurrences: 4(4)
I  9: 193.3205     4(  4)  T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-
T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y-T-K-[LV]-[ST]-x-F
Occurrences: 4(4)
J 10: 192.7154     4(  4)  S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-
[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G
Occurrences: 4(4)
K 11: 188.5453     4(  4)  K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-
x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G
Occurrences: 4(4)
L 12: 188.5453     4(  4)  C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-
T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W
Occurrences: 4(4)
M 13: 188.2314     4(  4)  G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-
G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y-T-K-[LV]-[ST]-x-F-L-[KR]-W-[IV]-[DR]-x(2)-M
Occurrences: 4(4)
```

N 14: 186.8953 4(4) L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R
Occurrences: 4(4)
O 15: 186.6020 4(4) R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I
Occurrences: 4(4)
P 16: 186.6020 4(4) D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G
Occurrences: 4(4)
Q 17: 186.6020 4(4) V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T
Occurrences: 4(4)
R 18: 186.6020 4(4) Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V
Occurrences: 4(4)
S 19: 186.6020 4(4) P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F
Occurrences: 4(4)
T 20: 186.6020 4(4) V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y
Occurrences: 4(4)
U 21: 186.6020 4(4) E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T
Occurrences: 4(4)
V 22: 186.6020 4(4) L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D
Occurrences: 4(4)
W 23: 181.3531 4(4) R-I-V-G-G-[DQR]-[DE]-C-x-[DEP]-G-E-C-P-W-Q-A-[LV]-L-[IV]-N-E-[EK]-[GN]-E-[EG]-F-C-G-G-T-I-L-[NS]-E-x-[FY]-[IV]-L-T-A-A-H-C-[IL]-x-Q-[AS]-K-[ER]
Occurrences: 4(4)
X 24: 178.5552 4(4) S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H
Occurrences: 4(4)

Y 25: 175.8864 4(4) V-[AV]-P-A-C-L-P-[EQ]-x-D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R
Occurrences: 4(4)
Z 26: 174.9273 4(4) C-L-P-[EQ]-x-D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K
Occurrences: 4(4)
a 27: 174.9273 4(4) A-C-L-P-[EQ]-x-D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C
Occurrences: 4(4)
b 28: 174.3852 4(4) R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G
Occurrences: 4(4)
c 29: 174.3852 4(4) G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G
Occurrences: 4(4)
d 30: 173.9720 4(4) D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F
Occurrences: 4(4)
e 31: 173.4817 4(4) Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A
Occurrences: 4(4)
f 32: 173.4817 4(4) K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G
Occurrences: 4(4)
g 33: 173.4769 4(4) G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]
Occurrences: 4(4)
h 34: 172.5060 4(4) M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F
Occurrences: 4(4)
i 35: 172.5060 4(4) L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M
Occurrences: 4(4)

j 36: 170.5410 4(4) A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I

Occurrences: 4(4)

k 37: 170.2151 4(4) E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D

Occurrences: 4(4)

l 38: 170.1698 4(4) V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T

Occurrences: 4(4)

m 39: 168.2252 4(4) R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C

Occurrences: 4(4)

n 40: 168.2252 4(4) G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A

Occurrences: 4(4)

o 41: 168.2252 4(4) F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D

Occurrences: 4(4)

p 42: 167.5345 4(4) A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q-C

Occurrences: 4(4)

q 43: 167.5345 4(4) R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q

Occurrences: 4(4)

r 44: 164.4753 4(4) L-[AEQ]-R-[ATV]-x-R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y

Occurrences: 4(4)

s 45: 161.8011 4(4) R-[ATV]-x-R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D

Occurrences: 4(4)

t 46: 160.8212 4(4) E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q-C-[ES]-[GST]-x-P-C

Occurrences: 4(4)

u 47: 160.8212 4(4) L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q-C-[ES]-[GST]-x-P

Occurrences: 4(4)

```
v 48: 159.7174    4( 4) A-x(2)-[FIV]-L-[AEQ]-R-[ATV]-x-R-A-N-S-F-L-E-E-
[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-
[DEN]-E-F
Occurrences: 4(4)
w 49: 157.3435    4( 4) R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-
[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q-C-[ES]-[GST]-x-P-C-x(2)-[GQ]-G-x-C-
K-D-G
Occurrences: 4(4)
x 50: 157.3435    4( 4) E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-
[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q-C-[ES]-[GST]-x-P-C-x(2)-[GQ]-G-x-C-
K-D
Occurrences: 4(4)
```

In this case, all the motifs/patterns retrieved from Prosite match the motifs found by SDISCOVER. Table 4.1 lists their match numbers.

Table 4.1 Motifs Found by SDISCOVER that Match Prosite Signatures

Motifs from Prosite	Motifs from SDISCOVER
A 1	30, 45, 67
B 2	30, 45, 67
C 3	30, 45, 67
D 4	30, 45, 67
E 5	30, 45, 67
F 6	30, 45, 67
G 7	30, 45, 67
H 8	30, 45, 67
I 9	30, 45, 67
J 10	30, 45, 67
K 11	30, 45, 67
L 12	30, 45, 67
M 13	30, 45, 67
N 14	30, 45, 67

Table 4.1 Motifs Found by SDISCOVER that Match Prosite Signatures (continued)

O 15	30, 45, 67
P 16	30, 45, 67
Q 17	30, 45, 67
R 18	30, 45, 67
S 19	30, 45, 67
T 20	30, 45, 67
U 21	30, 45, 67
V 22	30, 45, 67
W 23	2, 4, 6, 8, 12, 19, 20, 47, 48
X 24	30, 45, 67
Y 25	30, 45, 67
Z 26	30, 45, 67
a 27	30, 45, 67
b 28	30, 45, 67
c 29	30, 45, 67
d 30	30, 45, 67
e 31	30, 45, 67
f 32	30, 45, 67
g 33	30, 45, 67
h 34	30, 45, 67
i 35	30, 45, 67
j 36	30, 45, 67
k 37	30, 45, 67
l 38	30, 45, 67
m 39	30, 45, 67
n 40	30, 45, 67
o 41	30, 45, 67

Table 4.1 Motifs Found by SDISCOVER that Match Prosite Signatures (continued)

p 42	5, 7, 9, 11, 14, 23, 24, 28, 29, 34, 44, 52, 53, 54, 57, 58, 59, 60, 62, 64, 66, 68, 70, 71
q 43	5, 7, 9, 11, 14, 23, 24, 28, 29, 34, 44, 52, 53, 54, 57, 58, 59, 60, 62, 64, 66, 68, 70, 71
r 44	5, 7, 9, 14, 23, 24, 28, 34, 44, 52, 53, 54, 57, 58, 60, 62, 66, 70,
s 45	5, 7, 9, 14, 23, 24, 28, 34, 44, 52, 53, 54, 57, 58, 59, 60, 62, 66, 68, 70,
t 46	1, 5, 7, 9, 10, 11, 14, 17, 23, 24, 27, 28, 29, 34, 42, 44, 52, 53, 54, 55, 57, 58, 59, 60, 62, 64, 66, 68, 70, 71
u 47	1, 5, 7, 9, 10, 11, 14, 17, 23, 24, 27, 28, 29, 34, 42, 44, 52, 53, 54, 55, 57, 58, 59, 60, 62, 64, 66, 68, 70, 71
v 48	1, 5, 7, 10, 14, 17, 23, 27, 28, 34, 42, 44, 52, 54, 55, 57, 58, 60, 62, 66, 70
w 49	1, 5, 7, 9, 10, 11, 14, 17, 18, 22, 23, 24, 27, 28, 29, 34, 38, 39, 42, 43, 44, 52, 53, 54, 55, 56, 57, 58, 59, 60, 62, 63, 64, 65, 66, 68, 70, 71,73
x 50	1, 5, 7, 9, 10, 11, 14, 17, 18, 22, 23, 24, 27, 28, 29, 34, 38, 39, 42, 43, 44, 52, 53, 54, 55, 56, 57, 58, 59, 60, 62, 63, 64, 65, 66, 68, 70, 71,73

In this second test, we choose three proteins from three different families. The first one is Coagulation Factor X Precursor, the second one is Gamma-Carboxy glutamic Acid-Containing Protein, and last one Prothrombin Precursor. See Figure 4.3 for sequences of those three proteins.

Coagulation Factor X Precursor

```
>FA10_BOVIN COAGULATION FACTOR X PRECURSOR (EC 3.4.21.6) (STUART FACTOR).
MAGLLHLVLL STALGGLLRP AGSVFLPRDQ AHRVLQRARR ANSFLEEVKQ GNLERECLEE
PHVTRFKDTY FVTGIVSWGE GCARKGKFGV YTKVSNFLKW IDKIMKARAG AAGSRGHSEA
PATWTVPPPL PL
```

Gamma-Carboxyglutamic Acid- Containing Protein

```
>OSTC_HUMAN OSTEOCALCIN PRECURSOR (GAMMA-CARBOXYGLUTAMIC ACID-CONTAINING P
RO).
MRALTLLALL ALAALCIAGQ AGAKPSGAES SKGAAFVSKQ EGSEVVKRPR RYLYQWLGAP
VPYPDPLEPR REVCELNPDC DELADHIGFQ EAYRRFYGPV
```

Prothrombin Precursor

```
>THRB_RAT PROTHROMBIN PRECURSOR (EC 3.4.21.5).
RIGKHSRTRY ERNVEKISML EKIYIHPRYN WRENLDRDIA LLKLKKPVPF SDYIHPVCLP
TDNMFCAGFK VNDTKRGDAC EGDSGGPFVM KSPYNHRWYQ MGIVSWGEGC DRNGKYGFYT
HVFRLKRWMQ KVIDQHR
```

Figure 4.3 Sequences of Three Proteins from Different Families

The motifs found by SDISCOVER are as follows: total 8 motifs (after sorted)

```
#1      2           *GIVSWGEGCDR*
#2      2           *GIVSWGEGCAR*
#3      2           *TGIVSWGEGC*
#4      2           *MGIVSWGEGC*
```

After we retrieve the motifs from Prosite database in Pratt website, they have the same reports as in the previous example. We then reorganized the results as shown in Figure 4.4.

```
fitness    hits(seqs)   Pattern
A  1:  63.3191     3(  3)  L-L-x-L-x(2)-[LP]-[ASV]-[APT]-x(4)-[GIL]-x(2)-[GPV]-
[AC]-x-[PS]-[STV]-x(4)-[CDS]-x-[AG]-x(2)-[FV]-x-[DQS]-x(2)-[ER]-x-[ADS]-[AEN]-
[CSV]-x(2)-[DER]-[EPS]
Occurrences: 3(3)
B  2:  47.4711     3(  3)  G-[AC]-[AD]-x(2)-[GS]-K-x-[EG]-x(2)-[ET]-x-V-x-[NR]-
x(2)-[KR]-[WY]-[ILM]-x(3)-[ILM]-x-[AQ]
Occurrences: 3(3)
C  3:  47.2936     3(  3)  L-x(2)-[LP]-V-[LP]-x-[PS]-[DT]-x-[IL]-x-[GP]-x(3)-[PV]-
[ACT]-[DEG]-x(4)-[ACP]-x(3)-[AV]-x-[DHR]-x(4)-[ADE]
Occurrences: 3(3)
D  4:  44.1242     3(  3)  E-[GV]-C-[ADE]-x(2)-[GP]-[DK]-x-[DG]-x(2)-[AT]-[DHK]-
x(3)-[FL]-x-[EKR]-x(2)-[DQR]-[KR]-[FIV]
Occurrences: 3(3)
E  5:  42.4325     3(  3)  S-x-[FLV]-x-K-x(3)-[HKR]-x(3)-[ANQ]-x(2)-[EG]-[AN]-
[ALP]-[DGV]-x(2)-[GIP]-x(4)-[LP]-x(3)-[TV]-[CPV]-x(2)-[DNP]
Occurrences: 3(3)
F  6:  38.0618     3(  3)  E-[GV]-[CV]-x-R-x(2)-[KR]-[FY]-[GL]-x(4)-[GV]-x(2)-
[FLV]-x(4)-[DPQ]-x(3)-[DKR]-x-[EHR]
Occurrences: 3(3)
G  7:  37.7508     3(  3)  G-x(2)-[GV]-x-Y-[PT]-[DHK]-[PV]-x-[ENR]-x(2)-[KR]-x-
[IMV]-x-[EK]-[ILV]-x(2)-[ADQ]
Occurrences: 3(3)
H  8:  35.4034     3(  3)  L-x(4)-A-x(3)-L-x(2)-[AP]-x(3)-[GNS]-[ADS]-x-[ILP]-x(2)-
[AV]-x(2)-[GPS]-x(2)-[AEN]-x(2)-[CV]
Occurrences: 3(3)
I  9:  34.8673     3(  3)  A-L-[GL]-x-L-x(2)-[LP]-[ACV]-[GIP]-x-[GSV]-x(2)-[GIP]-
x(3)-[ACS]-x(4)-[NQS]
Occurrences: 3(3)
J 10:  33.2927     3(  3)  V-S-x-[GQ]-E-G-[CS]-[ADE]-x(3)-[KR]-x(3)-Y
Occurrences: 3(3)
K 11:  30.9869     3(  3)  G-[AGS]-[APV]-F-[LV]-x-[KR]-[DQS]-[EPQ]-x(2)-[EHR]-
x(3)-[QR]
Occurrences: 3(3)
L 12:  28.3770     3(  3)  K-x-[GS]-x-[AL]-E-x(3)-[GIL]-x(4)-[NSV]-x-[QR]-x(3)-
[DET]-x(4)-[GLP]
Occurrences: 3(3)
M 13:  25.7505     3(  3)  P-R-x-[ENQ]-x(2)-[ER]-x(2)-[DPQ]-[DR]-[ACD]-x(2)-[AL]
Occurrences: 3(3)
N 14:  24.6136     3(  3)  K-[DGP]-[ATV]-x-F-[SV]-[DST]-x(3)-[GPS]-x(3)-[GPV]
Occurrences: 3(3)
O 15:  24.0702     3(  3)  D-x(4)-A-x(3)-[AGV]-x-[DGQ]-[EST]-x(3)-[DRS]-x(3)-
[AGP]-[DTV]
```

Figure 4.4 Motifs Retrieved from Prosite in Pratt WebSite

```
Occurrences: 3(3)
P 16:  19.9827    3(  3)  A-x-[AE]-x-[ADP]-[AS]-G-[AGS]Occurrences: 3(3)
Q 17:  19.3754    3(  3)  R-E-[CNV]-x-[DE]-x-[DNP]-x-[ADV]
Occurrences: 3(3)
R 18:  17.4487    3(  3)  R-[AD]-x-[AST]-[FL]-LOccurrences: 3(3)
S 19:  16.7097    3(  3)  R-[ER]-x-L-x-[EQR]-x(2)-[AGV]Occurrences: 3(3)
T 20:  16.1292    3(  3)  P-x-[PTV]-x(2)-[DSV]-P-x-[ENP]Occurrences: 3(3)
U 21:  15.1091    3(  3)  G-x-A-x(5)-G-[AGV]Occurrences: 3(3)
V 22:  14.8147    3(  3)  V-P-x-[PS]-x(2)-[IL]Occurrences: 3(3)
W 23:  14.7617    3(  3)  V-x-R-x-[KR]-[DR]Occurrences: 3(3)
X 24:  14.7213    3(  3)  P-[PV]-[CP]-x-POccurrences: 3(3)
Y 25:  14.7117    3(  3)  V-[CP]-x-P-x-[DP]Occurrences: 3(3)
Z 26:  14.7107    3(  3)  P-x-[CP]-L-[EP]Occurrences: 3(3)
a 27:  14.7088    3(  3)  P-[PV]-P-x-[PS]Occurrences: 3(3)
b 28:  14.1721    3(  3)  V-x(3)-[DST]-x-L-[EG]Occurrences: 3(3)
c 29:  14.1675    3(  3)  L-[DNP]-x-D-x-[AD]Occurrences: 3(3)
d 30:  11.6270    3(  3)  R-x(2)-[GQ]-x-VOccurrences: 3(3)
e 31:  11.5763    3(  3)  G-[AD]-AOccurrences: 3(3)
f 32:  11.5405    3(  3)  Y-[GT]-x-VOccurrences: 3(3)
g 33:  11.5281    3(  3)  G-[IP]-VOccurrences: 3(3)
h 34:  11.5110    3(  3)  V-x(2)-[QR]-x-ROccurrences: 3(3)
i 35:  11.5104    3(  3)  K-[GL]-x(3)-VOccurrences: 3(3)
j 36:  11.5102    3(  3)  D-x-I-x(4)-[AL]Occurrences: 3(3)
k 37:  11.5102    3(  3)  H-x(3)-Q-x-[AG]Occurrences: 3(3)
l 38:  10.9771    3(  3)  R-x(4)-E-x(3)-[DGT]Occurrences: 3(3)
m 39:  10.9705    3(  3)  F-x-[EGT]-x-VOccurrences: 3(3)
n 40:  10.9651    3(  3)  R-[APT]-ROccurrences: 3(3)
o 41:  10.9341    3(  3)  G-x(5)-R-x-[EQR]Occurrences: 3(3)
p 42:   8.3401    3(  3)  E-x(3)-EOccurrences: 3(3)
q 43:   8.3401    5(  3)  L-EOccurrences: 5(3)
r 44:   8.3401    3(  3)  P-x(3)-ROccurrences: 3(3)
s 45:   8.3401    4(  3)  G-x(3)-VOccurrences: 4(3)
t 46:   8.3401    4(  3)  E-GOccurrences: 4(3)
u 47:   8.3401    3(  3)  F-VOccurrences: 3(3)
```

Figure 4.4 Motifs Retrieved from Prosite in Pratt WebSite (continued)

In this example, the query protein sequences retrieve 47 motifs from the database. Comparing these 47 motifs with the 4 motifs found by SDICOVER, we only find g 33, p 42, and t 46 match all the motifs found by SDISCOVER, in our case from #1 to #4.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

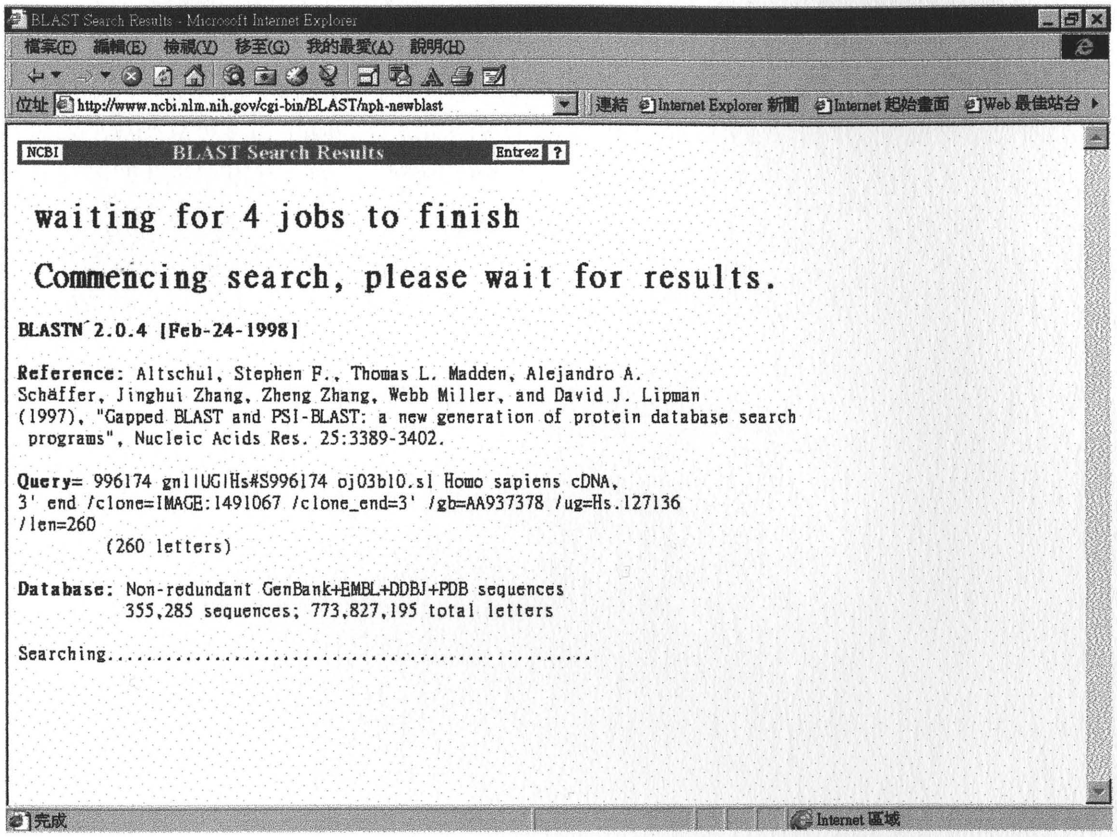
With data throughput that may soon approach hundreds of megabytes a year and sequence data that comes from a variety of sources (including the US and European Patent Offices), a major challenge will be to provide up-to-date and unique annotation for this sequence data. Next in importance to the sequence database itself is the computer program used to search it. A number of different search algorithms have been developed over the years, and further information about them may be found in Altschul et al. [1,3] , Schuler et al. [44], and references therein.

Database searching can be performed efficiently in phase, with a query first compared to a small database containing domains representative of large sequence families. Subsequences of a query that match one or more of these domains can then be masked prior to full-scale searching, thereby eliminating most of the redundant output [3]. A more comprehensive solution to the problem is approached by building a small, representative set of motifs and using this as a screening database with automatic masking of matching query subsequences. This technology is still under development but recent studies indicates that a representative set of only 1,000 – 3,000 sequences may suffice and such a database can be searched in seconds.

Computer databases, networks, and software tools are essential resources for all aspects of genome analysis [7]. The consequent abiding interest in the exhaustive alignment approach has prompted the use of powerful and expensive highly-parallel computers to make its application to sequence similarity searches through large databases feasible [39]. A cheaper alternative may be represented by the cooperative use of ordinary workstations, possibly even PCs, connected by a network; this way, the computational load mat be distributed over two or more computers, perhaps from different vendors, allowing the user to take advantage of whatever is available [49]. But, due to the bottleneck of Internet traffic, there is also a shortcoming for using Internet to do sequence querying. In our experiment, we encounter some difficulties (Figure 5.1 and Figure 5.2). An alternate approach is to construct one's own local database and download up-to-date sequence data or add one's own sequence data to the local database.

As we mentioned in Chapter 1, databases contain some very large families of related domains, motifs or repeated sequences. Unless special measures are taken, these biases very commonly confound database search methods and interfere with the discovery of interesting new sequence similarities [22]. And due to the error-prone nature of these sequence fragments, identifying redundancy in these databases is a more difficult task [24,35]. A good example is in Chapter 3. Hs.12716 include two sequences found only in library 651: NCI_CGAP_Mel3; when we used it as a query sequence to test the NCBI database, it generated many misleading results. Practically, there are two ways to avoid this problem. First, create a smaller and non-redundant database [5,36]. Second, process the query sequence for the presence of known domains and mask these prior to searching [26]. In our case, it also needs increase the database records and future studies.

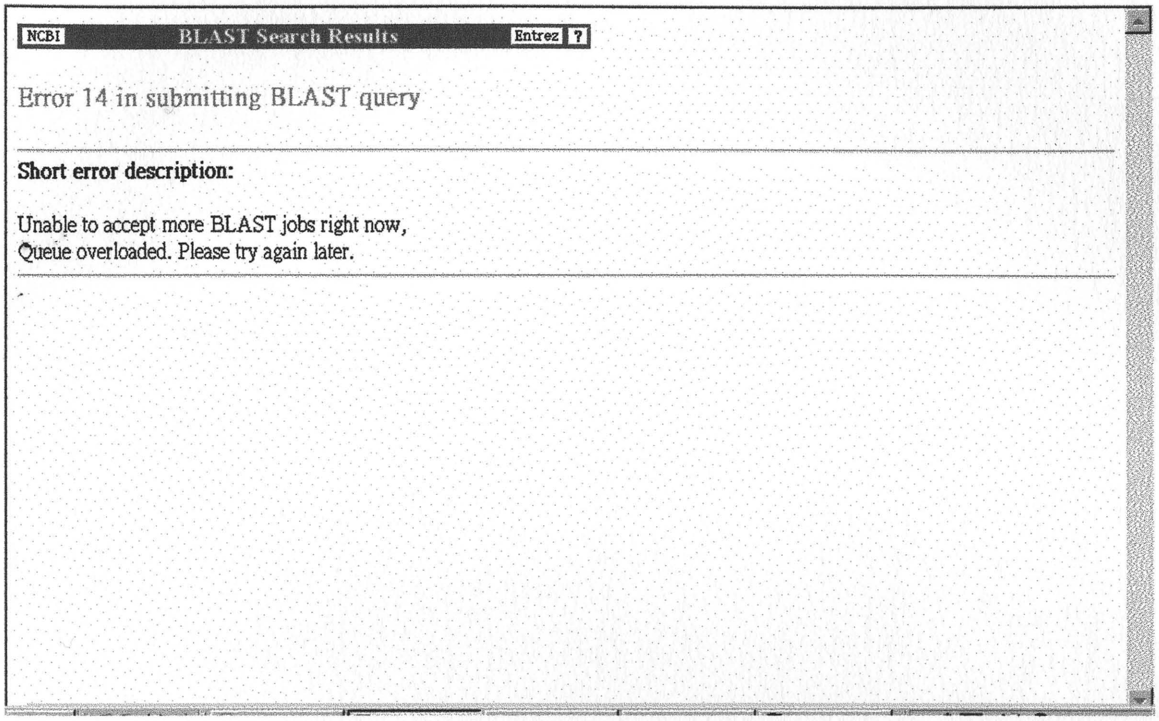
One of the most important advances in database similarity searching during the past several years has been the introduction of methods for the automatic masking of low complexity sequences [18]. Anyone who does a lot of database searching will have encountered problematic query sequences that result in hundreds (or thousands) of spurious matches to nebulous entities with names like "proline-rich protein" that may obscure more subtle but biologically interesting matches. An increasingly important use of motifs in the future will be to "preprocess" query sequences for the presence of obvious known domains and then mask these regions prior to a full-scale search [19,53]. This should simultaneously increase the speed of the search while improving the ability to detect subtle matches that would otherwise be swamped out by abundant, strong matches to other sequence regions [7,22,28].



BLAST Search Results - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 移至(G) 我的最愛(A) 說明(H)
位址 http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-newblast
連結 Internet Explorer 新聞 Internet 起始畫面 Web 最佳站台
NCBI
BLAST Search Results
Entrez ?
完成
Internet 區域

```
waiting for 4 jobs to finish

Commencing search, please wait for results.

BLASTN 2.0.4 [Feb-24-1998]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A.
Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search
 programs", Nucleic Acids Res. 25:3389-3402.

Query= 996174 gnl|UG|Hs#S996174 oj03b10.s1 Homo sapiens cDNA,
3' end /clone=IMAGE:1491067 /clone_end=3' /gb=AA937378 /ug=Hs.127136
/len=260
         (260 letters)

Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences
           355,285 sequences; 773,827,195 total letters

Searching................................................
```

Figure 5.1 Waiting Jobs to Finish



Error 14 in submitting BLAST query

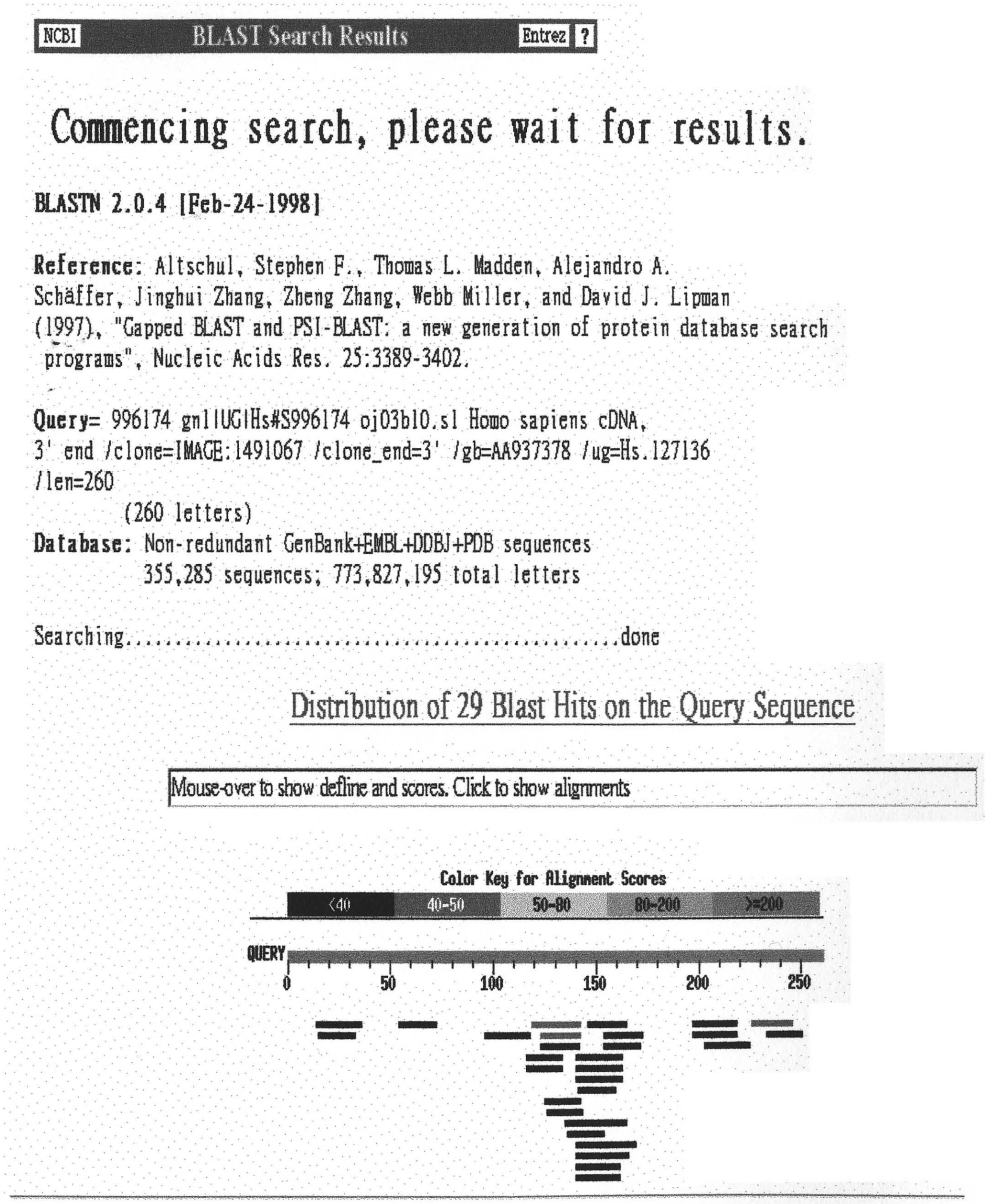
Short error description:

Unable to accept more BLAST jobs right now,
Queue overloaded. Please try again later.

Figure 5.2 Error Message: Unable to Accept More Jobs

APPENDIX A.1

QUERY RESULTS OF HS.12716 IN NCBI HOMEPAGE



Commencing search, please wait for results.

BLASTN 2.0.4 [Feb-24-1998]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

```
Query= 996174 gnl|UG|Hs#S996174 oj03b10.s1 Homo sapiens cDNA,
3' end /clone=IMAGE:1491067 /clone_end=3' /gb=AA937378 /ug=Hs.127136
/len=260
         (260 letters)
Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences
           355,285 sequences; 773,827,195 total letters

Searching..................................................done
```

Distribution of 29 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

```
                                                              Score    E
Sequences producing significant alignments:                  (bits)  Value

emb|AL008975|PFSC03029  Plasmodium falciparum DNA *** SEQUENCING...   40  0.15
gb|AC004158|HUAC004158  Homo sapiens Chromosome 16 BAC clone CIT...   40  0.15
emb|X69067|MTAFDNA  A.franciscana complete mitochondrial DNA          40  0.15
emb|AL022477|HS172K10  Homo sapiens DNA sequence from PAC 172K10...   38  0.60
emb|AL021786|HS696H22  Human DNA sequence from PAC 696H22 on chr...   38  0.60
emb|X54106|PHADH1  P.hybrida ADH1 gene for alcohol dehydrogenase 1    38  0.60
gb|U70846|CELD1073  Caenorhabditis elegans cosmid D1073               38  0.60
dbj|D87016|D87016  Human (lambda) DNA for immunoglobin light chain    38  0.60
emb|X13812|DAADH  D.affinidisjuncta Adh gene for alcohol dehydro...   38  0.60
dbj|D87018|D87018  Human (lambda) DNA for immunogloblin light chain   38  0.60
gb|U95973|ATU95973  Arabidopsis thaliana BAC T19D16 genomic sequ...   38  0.60
emb|Z83827|HS473J6  Human DNA sequence from PAC 473J6 on chromos...   36  2.4
gb|AC004740|AC004740  Homo sapiens PAC clone DJ0631B17 from 7p21...   36  2.4
dbj|D87010|D87010  Human (lambda) DNA for immunogloblin light chain   36  2.4
gb|U23516|CELB0416  Caenorhabditis elegans cosmid B0416               36  2.4
emb|Y13334|CJY13334  Campylobacter jejuni groES, groEL genes          36  2.4
gb|M63705|XELXNF7AA  X.laeivs xnf7 protein mRNA, complete cds.        36  2.4
gb|U50058|AGU50058  Asterina gibbosa mitochondrial transfer RNAs...   36  2.4
emb|Z33348|MCAAJ  M.capricolum DNA for CONTIG MCAAJ                   36  2.4
gb|U46158|CAU46158  Candida albicans RAS-related protein (RSR1) ...   36  2.4
gb|AC002390|AC002390  Human DNA from overlapping chromosome 19-s...   36  2.4
gb|AC003026|HUAC003026  Human Chromosome 16 BAC clone CIT987SK-1...   36  2.4
gb|AC004454|AC004454  Homo sapiens PAC clone DJ0988L12 from 7q11...   36  2.4
dbj|AB006793|AB006793  Ipomoea nil DNA for dihydroflavonol 4-red...   36  2.4
gb|AF067383|HS1UBR4  Homo sapiens ubiquitin-protein ligase E3-al...   36  2.4
gb|S64515|S64515  xnf7=zinc finger nuclear phosphoprotein [Xenop...   36  2.4
emb|AL022150|HS198G23  Homo sapiens DNA sequence from PAC 198G23...   36  2.4
gb|AF039709|AF039709  Maackia amurensis 14-3-3 protein homolog m...   36  2.4
emb|Z92831|CEF22G12  Caenorhabditis elegans cosmid F22G12, compl...   36  2.4

 emb|AL008975|PFSC03029 Plasmodium falciparum DNA *** SEQUENCING IN PROGRESS *** from contig
           3-29, complete sequence [Plasmodium falciparum]
           Length = 18280

 Score = 40.1 bits (20), Expect = 0.15
 Identities = 20/20 (100%), Positives = 20/20 (100%)


Query: 227  atatcatagaaatatttctt 246
            ||||||||||||||||||||
Sbjct: 6225 atatcatagaaatatttctt 6206
```

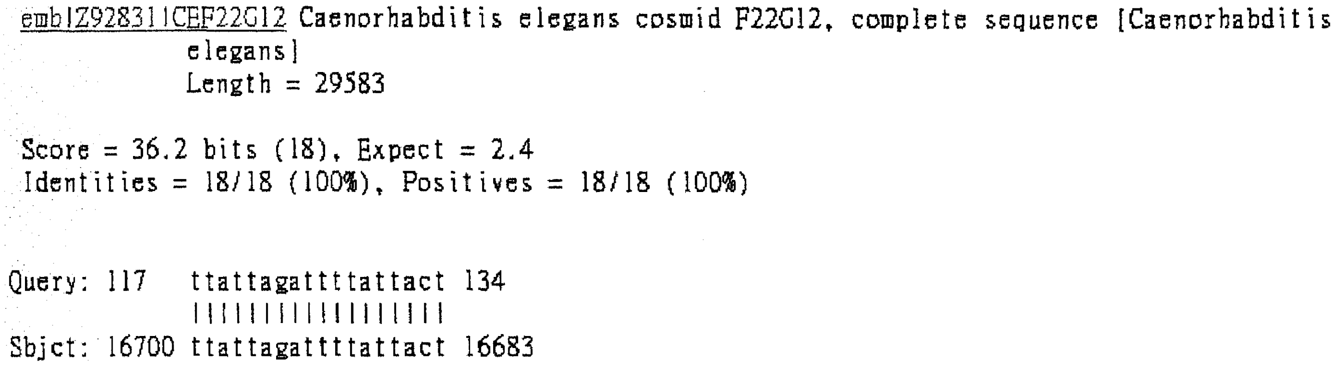
```
 gb|AC004158|HUAC004158 Homo sapiens Chromosome 16 BAC clone CIT987SK-A-10F4, complete sequence
             [Homo sapiens]
             Length = 180551

 Score = 40.1 bits (20), Expect = 0.15
 Identities = 23/24 (95%), Positives = 23/24 (95%)


Query: 120    ttagattttattactactatcatt 143
              ||||||||||||||| ||||||||
Sbjct: 119810 ttagattttattactgctatcatt 119787


 emb|X69067|MTAFDNA A.franciscana complete mitochondrial DNA
           Length = 15822

 Score = 40.1 bits (20), Expect = 0.15
 Identities = 20/20 (100%), Positives = 20/20 (100%)


Query: 124  attttattactactatcatt 143
            ||||||||||||||||||||
Sbjct: 1825 attttattactactatcatt 1844


 emb|AL022477|HS172K10 Homo sapiens DNA sequence from PAC 172K10 on chromosome 6q24. Contains
             STS, GSS and chromosome 6 fragment, complete sequence
             [Homo sapiens]
             Length = 82073




  Score = 38.2 bits (19), Expect = 0.60
  Identities = 22/23 (95%), Positives = 22/23 (95%)


 Query: 141   atttattttagcaatgttttata 163
               |||||||||| ||||||||||||
 Sbjct: 70123 atttattttatcaatgttttata 70145
```

```
  emb|AL021786|HS696H22 Human DNA sequence from PAC 696H22 on chromosome Xq21.1-21.2. Contains
            a mouse E25 like gene, a Kinesin like pseudogene and ESTs
            Length = 70665

  Score = 38.2 bits (19), Expect = 0.60
  Identities = 19/19 (100%), Positives = 19/19 (100%)


 Query: 142   tttattttagcaatgtttt 160
               |||||||||||||||||||
 Sbjct: 61900 tttattttagcaatgtttt 61918


  emb|X54106|PHADH1 P.hybrida ADH1 gene for alcohol dehydrogenase 1
            Length = 4672

  Score = 38.2 bits (19), Expect = 0.60
  Identities = 19/19 (100%), Positives = 19/19 (100%)


 Query: 124  attttattactactatcat 142
              |||||||||||||||||||
 Sbjct: 4103 attttattactactatcat 4085


  gb|U70846|CELD1073 Caenorhabditis elegans cosmid D1073
            Length = 7776

  Score = 38.2 bits (19), Expect = 0.60
  Identities = 19/19 (100%), Positives = 19/19 (100%)


Query: 155 tgttttataataatgaaag 173
           |||||||||||||||||||
Sbjct: 701 tgttttataataatgaaag 719

 dbj|D87016|D87016 Human (lambda) DNA for immunoglobin light chain
            Length = 37115

 Score = 38.2 bits (19), Expect = 0.60
 Identities = 22/23 (95%), Positives = 22/23 (95%)
 emb|Z83827|HS473J6 Human DNA sequence from PAC 473J6 on chromosome X contains STS
           Length = 135686

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 137   tatcatttattttagcaa 154
              ||||||||||||||||||
Sbjct: 95381 tatcatttattttagcaa 95398

 gb|AC004740|AC004740 Homo sapiens PAC clone DJ0631B17 from 7p21-p22, complete sequence [Homo
            sapiens]
            Length = 127270

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 24/26 (92%), Positives = 24/26 (92%)


Query: 141   atttattttagcaatgttttataata 166
              |||||||| |||| ||||||||||||
Sbjct: 26841 atttatttgagcagtgttttataata 26866

 dbj|D87010|D87010 Human (lambda) DNA for immunogloblin light chain
            Length = 40233

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 21/22 (95%), Positives = 21/22 (95%)


gb|U95973|ATU95973 Arabidopsis thaliana BAC T19D16 genomic sequence
           Length = 115641

 Score = 38.2 bits (19), Expect = 0.60
 Identities = 19/19 (100%), Positives = 19/19 (100%)


Query: 55     aaggtcctggaatggtgtt 73
              |||||||||||||||||||
Sbjct: 113952 aaggtcctggaatggtgtt 113970
```

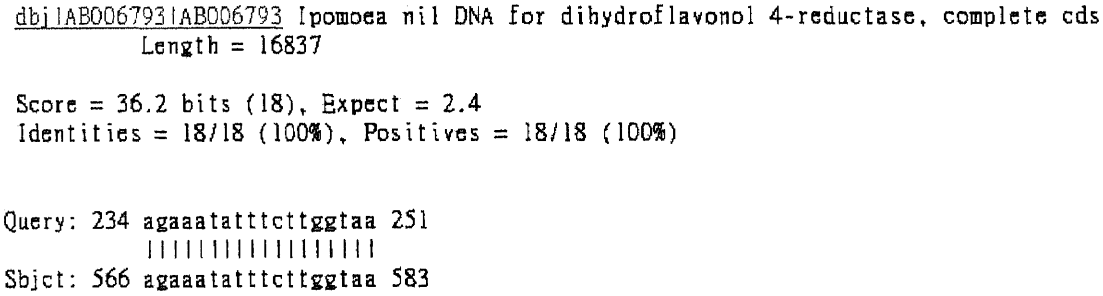
```
Query: 141   atttattttagcaatgttttat 162
             |||| |||||||||||||||||
Sbjct: 17272 atttcttttagcaatgttttat 17251


  gb|U23516|CELB0416 Caenorhabditis elegans cosmid B0416
             Length = 44797

  Score = 36.2 bits (18), Expect = 2.4
  Identities = 27/30 (90%), Positives = 27/30 (90%)


Q Query: 141   atttattttagcaatgttttataataatga 170
               |||| ||||||||||||| || ||||||||
S Sbjct: 25043 attttttttagcaatgttataaaataatga 25014

 emb|X13812|DAADH D.affinidisjuncta Adh gene for alcohol dehydrogenase
          >gi|156813|gb|M37262|DROADHAB D.affinidisjuncta alcohol
          dehydrogenase (adh) gene, exons 1-4.
          Length = 3886

 Score = 38.2 bits (19), Expect = 0.60
 Identities = 19/19 (100%), Positives = 19/19 (100%)


Query: 147 tttagcaatgttttataat 165
           |||||||||||||||||||
Sbjct: 471 tttagcaatgttttataat 453


 dbj|D87018|D87018 Human (lambda) DNA for immunogloblin light chain
           Length = 38756

 Score = 38.2 bits (19), Expect = 0.60
 Identities = 22/23 (95%), Positives = 22/23 (95%)


Query: 141   atttattttagcaatgttttata 163
             |||| ||||||||||||||||||
Sbjct: 33295 atttcttttagcaatgttttata 33273
```

```
emb|Y13334|CJY13334 Campylobacter jejuni groES, groEL genes
          Length = 2580

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 16   aggatttgatgcaggtgt 33
            ||||||||||||||||||
Sbjct: 2021 aggatttgatgcaggtgt 2038


gb|M63705|XELXNF7AA X.laeivs xnf7 protein mRNA, complete cds.
          Length = 2253

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 21/22 (95%), Positives = 21/22 (95%)


Query: 198 aacaggctggaagtgaagagta 219
           ||||||||||||| ||||||||
Sbjct: 176 aacaggctggaagagaagagta 197


gb|U50058|AGU50058 Asterina gibbosa mitochondrial transfer RNAs (Ala, Leu UAG, Asn,
          Gln, Pro) and cytochrome oxidase subunit I (COI) gene,
          complete cds. >gi|1289473|gb|U50045|PRU50045 Patiriella
          regularis mitochondrial transfer RNAs (Ala, Leu UAG,
          Asn, Gln, Pro) and cytochrome oxidase subunit I (COI)

          gene, complete cds.
          Length = 1942

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 126 tttattactactatcatt 143
           ||||||||||||||||||
Sbjct: 878 tttattactactatcatt 895


emb|Z33348|MCAAJ M.capricolum DNA for CONTIG MCAAJ
          Length = 1708
```

```
gb|AC004454|AC004454 Homo sapiens PAC clone DJ0988L12 from 7q11.23-q21.1, complete sequence
           [Homo sapiens]
           Length = 80514

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 21/22 (95%), Positives = 21/22 (95%)


Query: 141   atttattttagcaatgttttat 162
              |||||||||| |||||||||||
Sbjct: 19076 atttattttatcaatgttttat 19097


 dbj|AB006793|AB006793 Ipomoea nil DNA for dihydroflavonol 4-reductase, complete cds
          Length = 16837

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 234 agaaatatttcttggtaa 251
           ||||||||||||||||||
Sbjct: 566 agaaatatttcttggtaa 583


 gb|AF067383|HS1UBR4 Homo sapiens ubiquitin-protein ligase E3-alpha (UBR1) gene, exons 4
           through 7
           Length = 3980
 Score = 36.2 bits (18), Expect = 2.4
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 117  ttattagattttattact 134
             ||||||||||||||||||
Sbjct: 1484 ttattagattttattact 1501

 gb|U46158|CAU46158 Candida albicans RAS-related protein (RSR1) gene, complete cds
           Length = 1917

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 127 ttattactactatcattt 144
           ||||||||||||||||||
Sbjct: 179 ttattactactatcattt 196

 gb|AC002390|AC002390 Human DNA from overlapping chromosome 19-specific cosmids R30072 and
             R28588, genomic sequence, complete sequence [Homo sapiens]
             Length = 70311

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 21/22 (95%), Positives = 21/22 (95%)


Query: 198 aacaggctggaagtgaagagta 219
           ||||||||||||| ||||||||
Sbjct: 176 aacaggctggaagagaagagta 197


 emb|AL022150|HS198G23 Homo sapiens DNA sequence from PAC 198G23 on chromosome Xq21.1-q21.33.
             Contains GSS, STS, complete sequence [Homo sapiens]
             Length = 94886

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 27/30 (90%), Positives = 27/30 (90%)


Query: 136   ctatcatttattttagcaatgttttataat 165
             |||| |||||||| |||| |||||||||||
Sbjct: 63029 ctataatttatttcagcagtgttttataat 63058


 gb|AF039709|AF039709 Maackia amurensis 14-3-3 protein homolog mRNA, complete cds
           Length = 1176

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 21/22 (95%), Positives = 21/22 (95%)


Query: 15  taggatttgatgcaggtgtttg 36
            |||||||||||||| |||||||
Sbjct: 842 taggatttgatgcatgtgtttg 863
```



```
emb|Z92831|CEF22G12 Caenorhabditis elegans cosmid F22G12, complete sequence [Caenorhabditis
             elegans]
             Length = 29583

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 117   ttattagattttattact 134
              ||||||||||||||||||
Sbjct: 16700 ttattagattttattact 16683
```

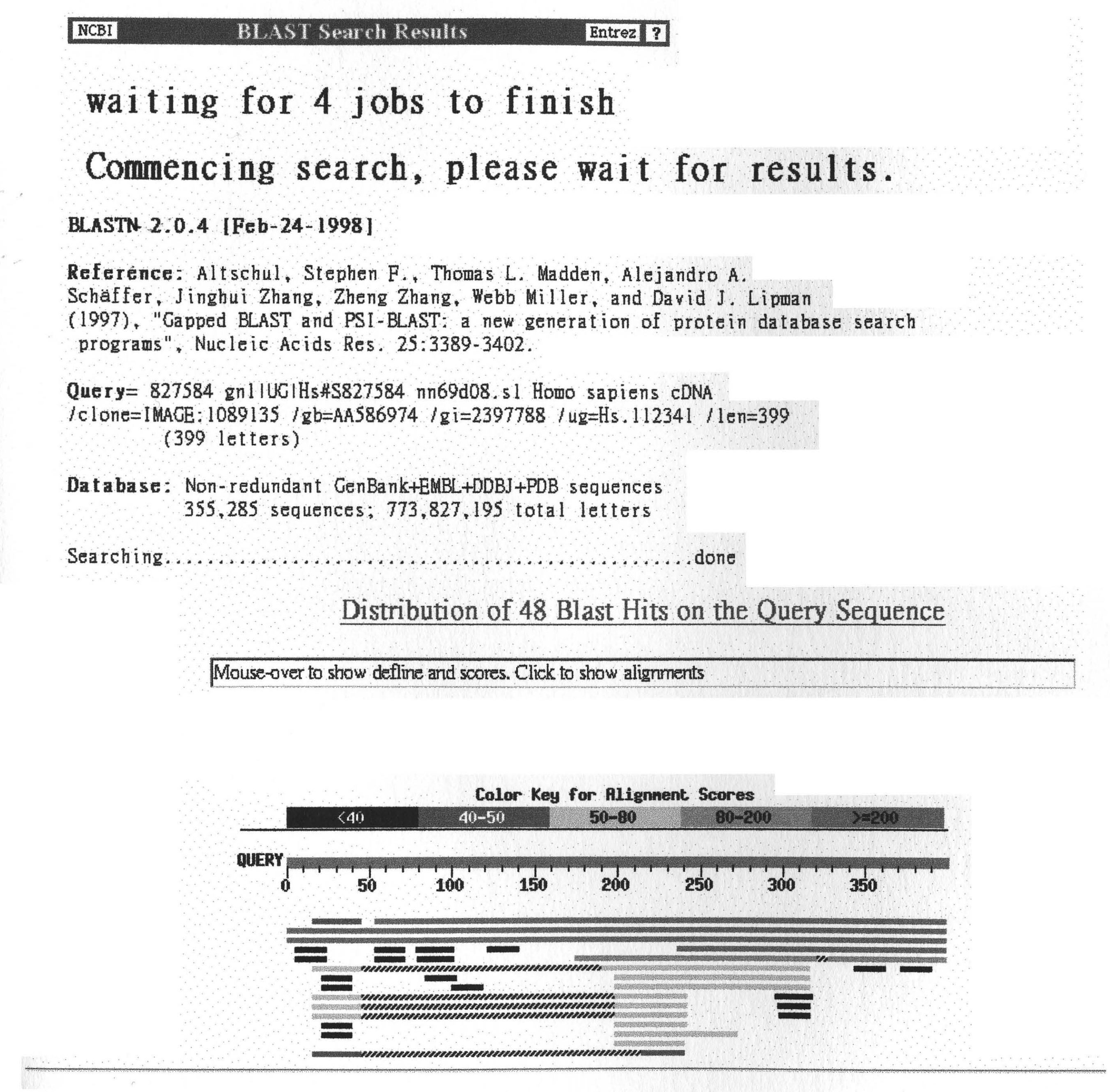


```
dbj|AB006793|AB006793 Ipomoea nil DNA for dihydroflavonol 4-reductase, complete cds
          Length = 16837

 Score = 36.2 bits (18), Expect = 2.4
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 234 agaaatatttcttggtaa 251
           ||||||||||||||||||
Sbjct: 566 agaaatatttcttggtaa 583
```

```
CPU time:     8.86 user secs.          0.84 sys. secs           9.70 total secs.

  Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences
    Posted date:  Jul 26, 1998  8:01 AM
  Number of letters in database: 773,827,195
  Number of sequences in database:  355,285

Lambda     K      H
    1.37    0.711     1.31

Gapped
Lambda     K      H
    1.37    0.711     1.31


Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 134096
Number of Sequences: 355285
Number of extensions: 134096
Number of successful extensions: 37010
Number of sequences better than 10: 38
length of query: 260
length of database: 773827195
effective HSP length: 19
effective length of query: 241
effective length of database: 767076780
effective search space: 184865503980


T: 0
A: 0
X1: 6 (11.9 bits)
X2: 25 (49.6 bits)
S1: 0 ( 0.5 bits)
S2: 17 (34.2 bits)
```

APPENDIX A.2

HS.112341, QUERY RESULT SCREEN IN NCBI HOMEPAGE



waiting for 4 jobs to finish

Commencing search, please wait for results.

```
BLASTN 2.0.4 [Feb-24-1998]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A.
Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search
 programs", Nucleic Acids Res. 25:3389-3402.

Query= 827584 gnl|UG|Hs#S827584 nn69d08.s1 Homo sapiens cDNA
/clone=IMAGE:1089135 /gb=AA586974 /gi=2397788 /ug=Hs.112341 /len=399
         (399 letters)

Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences
           355,285 sequences; 773,827,195 total letters

Searching..................................................done
```

Distribution of 48 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

```
                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Value

emb|Z18538|HSANTLEUP  H.sapiens encoding skin-derived antileukop...  686   0.0
dbj|D13156|HUMELAFIN  Human gene for elafin, complete cds            446   e-123
gb|L10343|HUMPREELAS  Huma elafin gene, complete cds. >gi|299840...  446   e-123
emb|AJ223215|MAM223215  Macaca mulatta mRNA for putative sTrappi...  276   2e-72
emb|AJ223216|BOT223216  Bos taurus mRNA for putative bTrappin-2 ...   92   7e-17
dbj|D50319|PIGWAPA  Pig DNA for elafin, complete cds                  76   4e-12
dbj|D50322|PIGWAPD  Pig mRNA for elafin family member protein, c...   68   1e-09
dbj|D83668|D83668  Sus scrofa gene for elafin homolog, exon2, pa...   68   1e-09
dbj|D50320|PIGWAPB  Pig DNA for SPAI-2, complete cds                  56   4e-06
dbj|D83667|DMY245  Wild boar; domestic pig mRNA for preproSPAI-2...   56   4e-06
dbj|D17753|PIGSPAI2A1  Porcine mRNA for SPAI-2, complete cds          56   4e-06
dbj|D17755|PIGSPAI2S2  Porcine DNA for SPAI-2, exon 2                 56   4e-06
dbj|AB011010|AB011010  Bos taurus gene for Trappin-6, partial cds     52   6e-05
dbj|D50323|PIGWAPE  Pig mRNA for elafin family member protein, c...   52   6e-05
dbj|D50321|PIGWAPC  Pig DNA for elafin family member protein, co...   42   0.061
dbj|D17756|PIGSPAI2S3  Porcine DNA for SPAI-2, exon 3                 42   0.061
emb|X05710|TBTRS16  Trypanosoma brucei DNA for trypanosome repea...   38   0.95
dbj|D90904|D90904  Synechocystis sp. PCC6803 complete genome, 6/...   38   0.95
gb|AC001228|HSAC001228  244Kb Contig from Human Chromsome 11p15....   38   0.95
gb|L19876|DROCOFACTO  Drosophila melanogaster molybdenum cofacto...   38   0.95
gb|L42568|HUMATP1G09  Homo sapiens (clone 1SW11-1) non-gastric H...   38   0.95
gb|J00306|HUMSOMI  Human somatostatin I gene and flanks.              38   0.95
gb|AF017113|AF017113  Bacillus subtilis 300-304 degree genomic s...   38   0.95
emb|Z99122|BSUB0019  Bacillus subtilis complete genome (section ...   38   0.95
gb|L10345|RICAMYBA  Oryza sativa beta-amylase gene, complete cds.     36   3.8
gb|L10346|RICAMYBB  Oryza sativa beta-amylase gene, complete cds.     36   3.8

 emb|Z49237|HSL27H9  Human DNA from cosmid L27h9, Huntington's Di...   36   3.8
 gb|M21005|HUMMRP8A  Human migration inhibitory factor-related pr...   36   3.8
 gb|AC002422|AC002422  Human Chromosome X, complete sequence [Hom...   36   3.8
 emb|X84419|HSTAX1EX1  H.sapiens TAX-1 gene (exon 1)                   36   3.8
 gb|U04855|IAU04855  Influenza A virus (H1N1) A/swine/Northern Ir...   36   3.8
 gb|U04856|IAU04856  Influenza A virus (H1N1) A/swine/Cambridge/3...   36   3.8
 gb|M30746|FLANPB  Influenza A/Wilson-Smith/33 (H1N1) nucleoprote...   36   3.8
 gb|M63769|FLANPAW  Influenza A/Swine/Cambridge/1/35 (H1N1) nucle...   36   3.8
```

```
emb|Z18538|HSANTLEUP H.sapiens encoding skin-derived antileukoproteinase
          Length = 478

 Score =  686 bits (346), Expect = 0.0
 Identities = 346/346 (100%), Positives = 346/346 (100%)

Query: 54   tgggcatcctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggact 113
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 478 tgggcatcctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggact 419

Query: 114 taggaccagatggggcctgtagctctggggacggcacaggtgcagcaaggaccggctccc 173
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 418 taggaccagatggggcctgtagctctggggacggcacaggtgcagcaaggaccggctccc 359

Query: 174 tctcactggggaacgaaacaggccatcccgcaagagccttcacagcacttcttgattcct 233
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 358 tctcactggggaacgaaacaggccatcccgcaagagccttcacagcacttcttgattcct 299

Query: 234 gggcagtcagtatctttcaagcagcggttaggggggattcaacatggcgcaccggatcaag 293
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 298 gggcagtcagtatctttcaagcagcggttaggggggattcaacatggcgcaccggatcaag 239


Query: 294 ataatggggcaggagccaggcttagtggagactggacctttgactggctcttgcgctttg 353
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 238 ataatggggcaggagccaggcttagtggagactggacctttgactggctcttgcgctttg 179

Query: 354 actttatcttgacctttaactgaaacttgtcctttaacgggatctt 399
            ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 178 actttatcttgacctttaactgaaacttgtcctttaacgggatctt 133

 dbj|D13156|HUMELAFIN Human gene for elafin, complete cds
           Length = 1878

 Score =  446 bits (225), Expect = e-123
 Identities = 225/225 (100%), Positives = 225/225 (100%)
```

```
Query: 175  ctcactggggaacgaaacaggccatcccgcaagagccttcacagcacttcttgattcctg 234
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1460 ctcactggggaacgaaacaggccatcccgcaagagccttcacagcacttcttgattcctg 1401


Query: 235  ggcagtcagtatctttcaagcagcggttaggggattcaacatggcgcaccggatcaaga 294
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1400 ggcagtcagtatctttcaagcagcggttaggggattcaacatggcgcaccggatcaaga 1341


Query: 295  taatggggcaggagccaggcttagtggagactggacctttgactggctcttgcgctttga 354
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1340 taatggggcaggagccaggcttagtggagactggacctttgactggctcttgcgctttga 1281


Query: 355  ctttatcttgacctttaactgaaacttgtcctttaacgggatctt 399
            |||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1280 ctttatcttgacctttaactgaaacttgtcctttaacgggatctt 1236


 Score =  349 bits (176), Expect = 2e-94
 Identities = 176/176 (100%), Positives = 176/176 (100%)


Query: 1    ggagcagaaggaactctttattggaaagtggatgagagaggcagctccagccgtgggcat 60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1868 ggagcagaaggaactctttattggaaagtggatgagagaggcagctccagccgtgggcat 1809


Query: 61   cctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggacttaggacc 120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1808 cctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggacttaggacc 1749


Query: 121  agatggggcctgtagctctggggacggcacaggtgcagcaaggaccggctccctct 176
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1748 agatggggcctgtagctctggggacggcacaggtgcagcaaggaccggctccctct 1693


 gb|L10343|HUMPREELAS Huma elafin gene, complete cds. >gi|299840|gb|S58717|S58717
            pre-elafin=elastase-specific inhibitor [human, placental,
            Genomic, 2309 nt]
            Length = 2309
 Score =  446 bits (225), Expect = e-123
 Identities = 225/225 (100%), Positives = 225/225 (100%)


Query: 175  ctcactggggaacgaaacaggccatcccgcaagagccttcacagcacttcttgattcctg 234
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1728 ctcactggggaacgaaacaggccatcccgcaagagccttcacagcacttcttgattcctg 1669


Query: 235  ggcagtcagtatctttcaagcagcggttaggggattcaacatggcgcaccggatcaaga 294
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1668 ggcagtcagtatctttcaagcagcggttaggggattcaacatggcgcaccggatcaaga 1609


Query: 295  taatggggcaggagccaggcttagtggagactggacctttgactggctcttgcgctttga 354
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1608 taatggggcaggagccaggcttagtggagactggacctttgactggctcttgcgctttga 1549


Query: 355  ctttatcttgacctttaactgaaacttgtcctttaacgggatctt 399
             |||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1548 ctttatcttgacctttaactgaaacttgtcctttaacgggatctt 1504


 Score =  341 bits (172), Expect = 5e-92
 Identities = 175/176 (99%), Positives = 175/176 (99%)



Query: 1    ggagcagaaggaactctttattggaaagtggatgagagaggcagctccagccgtgggcat 60
             ||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 2135 ggagcagaaggaagtctttattggaaagtggatgagagaggcagctccagccgtgggcat 2076


Query: 61   cctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggacttaggacc 120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 2075 cctgaatgggaggaagaatggacagtgtgggaaggggaagggcagcagggacttaggacc 2016


Query: 121  agatggggcctgtagctctggggacggcacaggtgcagcaaggaccggctccctct 176
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 2015 agatggggcctgtagctctggggacggcacaggtgcagcaaggaccggctccctct 1960


  emb|AJ223215|MAM223215 Macaca mulatta mRNA for putative sTrappin-2 protein, partial
           Length = 270

  Score =  276 bits (139), Expect = 2e-72
  Identities = 157/163 (96%), Positives = 157/163 (96%)
```

```
Query: 237 cagtcagtatctttcaagcagcggttagggggattcaacatggcgcaccggatcaagata 296
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 270 cagtcagtatctttcaagcagcggttagggggattcaacatggcgcaccggatcaagata 211


Query: 297 atggggcaggagccaggcttagtggagactggacctttgactggctcttgcgctttgact 356
            |||||||||||||||||||||||||||| ||||||||||||||| || || ||||||||
Sbjct: 210 ttggggcaggagccaggcttagtggagacgggacctttgactggccctcgccctttgact 151


Query: 357 ttatcttgacctttaactgaaacttgtcctttaacgggatctt 399
            ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 150 ctatcttgacctttaactgaaacttgtcctttaacgggatctt 108
```

<u>emb|AJ223216|BOT223216</u> Bos taurus mRNA for putative bTrappin-2 protein, partial
Length = 573

```
 Score = 91.7 bits (46), Expect = 7e-17
 Identities = 121/146 (82%), Positives = 121/146 (82%)


Query: 175 ctcactggggaacgaaacaggccatcccgcaagagccttcacagcacttcttgattcctg 234
           ||||||||||| | | ||||| | |||| |||||||||||||||||||||||||  ||||
Sbjct: 406 ctcactggggatccatacaggtcttcccacaagagccttcacagcacttcttgacccctg 347


Query: 235 ggcagtcagtatctttcaagcagcggttagggggattcaacatggcgcaccggatcaaga 294
           |||| | || |||   |||  || ||||||||||| |||| |||||| ||||||||||  |
Sbjct: 346 ggcactgagcatccctcagacatcggttaggggggttcatcatggcacaccggatcagaa 287


Query: 295 taatggggcaggagccaggcttagtg 320
              |||||||||| ||||||||||||
Sbjct: 286 ccctggggcaggacccaggcttagtg 261


 Score = 50.1 bits (25), Expect = 2e-04
 Identities = 58/69 (84%), Positives = 58/69 (84%)


Query: 331 ctttgactggctcttgcgctttgactttatcttgacctttaactgaaacttgtcctttaa 390
           |||| ||||| |||||  ||||||||  |||||||||||| |||| | |||||||||| |
Sbjct: 88  ctttcactggatcttgtcctttgactggatcttgacctttgactggatcttgtcctttga 29
```

```
Query: 391 cgggatctt 399
           | |||||||
Sbjct: 28  ctggatctt 20


 Score = 36.2 bits (18), Expect = 3.8
 Identities = 57/70 (81%), Positives = 57/70 (81%)


Query: 330 cctttgactggctcttgcgctttgactttatcttgacctttaactgaaacttgtcctta 389
           ||||||||||| |||||  || |||||  || |||| |||| |||| | ||||||||||
Sbjct: 125 cctttgactggttcttgacctctgactggattttgatctttcactggatcttgtcctttg 66


Query: 390 acgggatctt 399
           || |||||||
Sbjct: 65  actggatctt 56


 Score = 48.1 bits (24), Expect = 0.001
 Identities = 60/72 (83%), Positives = 60/72 (83%)


Query: 328 gacctttgactggctcttgcgctttgactttatcttgacctttaactgaaacttgtcctt 387
           ||||||| |||||||||||  |||| |||  |||||||||||| |||| | |||| ||||
Sbjct: 199 gacctttcactggctcttgacctttcactggatcttgacctttgactggatcttgacctt 140



Query: 388 taacgggatctt 399
           | || |||||||
Sbjct: 139 tgactggatctt 128


 Score = 50.1 bits (25), Expect = 2e-04
 Identities = 52/61 (85%), Positives = 52/61 (85%)


Query: 328 gacctttgactggctcttgcgctttgactttatcttgacctttaactgaaacttgtcctt 387
           ||||||| ||||| |||||  ||||||||  |||||||||||| |||| | |||||||||
Sbjct: 181 gacctttcactggatcttgacctttgactggatcttgacctttgactggatcttgtcctt 122


Query: 388 t 388
           |
Sbjct: 121 t 121
```

```
 Score = 36.2 bits (18), Expect = 3.8
 Identities = 39/46 (84%), Positives = 39/46 (84%)


Query: 330 cctttgactggctcttgcgctttgactttatcttgacctttaactg 375
            ||||||||||| |||||  ||||||||  |||||| ||||| ||||
Sbjct: 71  cctttgactggatcttgacctttgactggatcttgtcctttgactg 26


 Score = 40.1 bits (20), Expect = 0.24
 Identities = 41/48 (85%), Positives = 41/48 (85%)


Query: 328 gacctttgactggctcttgcgctttgactttatcttgacctttaactg 375
            ||||||||||||| |||||  ||||||||  |||||| ||||| ||||
Sbjct: 163 gacctttgactggatcttgacctttgactggatcttgtcctttgactg 116


 Score = 40.1 bits (20), Expect = 0.24
 Identities = 35/40 (87%), Positives = 35/40 (87%)


Query: 328 gacctttgactggctcttgcgctttgactttatcttgacc 367
            ||||||||||||| |||||  ||||||||  |||||||||
Sbjct: 55  gacctttgactggatcttgtcctttgactggatcttgacc 16


 dbj|D50319|PIGWAPA Pig DNA for elafin, complete cds
           Length = 3693


 Score = 75.8 bits (38), Expect = 4e-12
 Identities = 104/126 (82%), Positives = 104/126 (82%)


Query: 191  acaggccatcccgcaagagccttcacagcacttcttgattcctgggcagtcagtatcttt 250
             ||||||| |||||||| |||||||||||||||||||||  |||||||| | || |||  |
Sbjct: 1947 acaggccttcccgcaaaagccttcacagcacttcttgagccctgggcactgagcatcact 1888


Query: 251  caagcagcggttagggggattcaacatggcgcaccggatcaagataatggggcaggagcc 310
             ||| || | |||||||||||| | |||   ||| |||||||  ||  |||||||||||||
Sbjct: 1887 caaacacctgttagggggattgaccatcaagcaacggatcagaatcctggggcaggagcc 1828


Query: 311  aggctt 316
             ||||||
Sbjct: 1827 aggctt 1822
```

```
 Score = 42.1 bits (21), Expect = 0.061
 Identities = 27/29 (93%), Positives = 27/29 (93%)


Query: 17   tttattggaaagtggatgagagaggcagc 45
            ||||||||||||  |||||||||||||||
Sbjct: 2585 tttattggaaagccgatgagagaggcagc 2557


 dbj|D50322|PIGWAPD Pig mRNA for elafin family member protein, complete cds
           Length = 464

 Score = 67.9 bits (34), Expect = 1e-09
 Identities = 97/118 (82%), Positives = 97/118 (82%)


Query: 199 tcccgcaagagccttcacagcacttcttgattcctgggcagtcagtatctttcaagcagc 258
           |||||||| |||||||||||||||||||||  |||||||| | || |||  |||| || |
Sbjct: 423 tcccgcaaaagccttcacagcacttcttgacccctgggcactgagcatcactcaaacacc 364


Query: 259 ggttagggggattcaacatggcgcaccggatcaagataatggggcaggagccaggctt 316
            |||||||||||| | |||   ||| |||||||  ||  |||||||||||||||||||
Sbjct: 363 tgttagggggattgaccatcaagcaacggatcagaatcctggggcaggagccaggctt 306


 dbj|D83668|D83668 Sus scrofa gene for elafin homolog, exon2, partial cds
           Length = 1034


 Score = 67.9 bits (34), Expect = 1e-09
 Identities = 97/118 (82%), Positives = 97/118 (82%)


Query: 199 tcccgcaagagccttcacagcacttcttgattcctgggcagtcagtatctttcaagcagc 258
           |||||||| |||||||||||||||||||||  |||||||| | || |||  |||| || |
Sbjct: 535 tcccgcaaaagccttcacagcacttcttgacccctgggcactgagcatcactcaaacacc 476


Query: 259 ggttagggggattcaacatggcgcaccggatcaagataatggggcaggagccaggctt 316
            |||||||||||| | |||   ||| |||||||  ||  |||||||||||||||||||
Sbjct: 475 tgttagggggattgaccatcaagcaacggatcagaatcctggggcaggagccaggctt 418


 dbj|D50320|PIGWAPB Pig DNA for SPAI-2, complete cds
           Length = 3782
 Score = 56.0 bits (28), Expect = 4e-06
 Identities = 40/44 (90%), Positives = 40/44 (90%)


Query: 199  tcccgcaagagccttcacagcacttcttgattcctgggcagtca 242
            |||||||| |||||||||||||||||||||  ||||| ||||||
Sbjct: 2010 tcccgcaaaagccttcacagcacttcttgacccctggacagtca 1967


 Score = 42.1 bits (21), Expect = 0.061
 Identities = 27/29 (93%), Positives = 27/29 (93%)


Query: 17   tttattggaaagtggatgagagaggcagc 45
            ||||||||||||  |||||||||||||||
Sbjct: 2685 tttattggaaagccgatgagagaggcagc 2657


 dbj|D83667|DMY245 Wild boar; domestic pig mRNA for preproSPAI-2, complete cds
          Length = 789

 Score = 56.0 bits (28), Expect = 4e-06
 Identities = 40/44 (90%), Positives = 40/44 (90%)


Query: 199 tcccgcaagagccttcacagcacttcttgattcctgggcagtca 242
           |||||||| |||||||||||||||||||||  ||||| ||||||
Sbjct: 601 tcccgcaaaagccttcacagcacttcttgacccctggacagtca 558


 Score = 42.1 bits (21), Expect = 0.061
 Identities = 27/29 (93%), Positives = 27/29 (93%)


Query: 17  tttattggaaagtggatgagagaggcagc 45
           ||||||||||||  |||||||||||||||
Sbjct: 766 tttattggaaagccgatgagagaggcagc 738


 dbj|D17753|PIGSPAI2A1 Porcine mRNA for SPAI-2, complete cds
          Length = 722

 Score = 56.0 bits (28), Expect = 4e-06
 Identities = 40/44 (90%), Positives = 40/44 (90%)
```

```
Query: 199 tcccgcaagagccttcacagcacttcttgattcctgggcagtca 242
           |||||||| |||||||||||||||||||||  ||||| ||||||
Sbjct: 541 tcccgcaaaagccttcacagcacttcttgacccctggacagtca 498


 Score = 42.1 bits (21), Expect = 0.061
 Identities = 27/29 (93%), Positives = 27/29 (93%)


Query: 17  tttattggaaagtggatgagagaggcagc 45
           ||||||||||||  |||||||||||||||
Sbjct: 706 tttattggaaagccgatgagagaggcagc 678


 dbj|D17755|PIGSPAI2S2 Porcine DNA for SPAI-2, exon 2
           Length = 483

 Score = 56.0 bits (28), Expect = 4e-06
 Identities = 40/44 (90%), Positives = 40/44 (90%)


Query: 199 tcccgcaagagccttcacagcacttcttgattcctgggcagtca 242
           |||||||| |||||||||||||||||||||  ||||| ||||||
Sbjct: 459 tcccgcaaaagccttcacagcacttcttgacccctggacagtca 416


 dbj|AB011010|AB011010 Bos taurus gene for Trappin-6, partial cds
           Length = 495


 Score = 52.0 bits (26), Expect = 6e-05
 Identities = 62/74 (83%), Positives = 62/74 (83%)


Query: 199 tcccgcaagagccttcacagcacttcttgattcctgggcagtcagtatctttcaagcagc 258
           |||| ||| ||||||||||||||||||||   |||||||||| || |||| || | ||
Sbjct: 306 tcccacaaaagccttcacagcacttcttggcccctgggcagtgagcatctctccaacact 247


Query: 259 ggttaggggattc 272
           ||||||||||||||
Sbjct: 246 ggttaggggattc 233
```

```
dbj|D50323|PIGWAPE Pig mRNA for elafin family member protein, complete cds
          Length = 578

 Score = 52.0 bits (26), Expect = 6e-05
 Identities = 38/42 (90%), Positives = 38/42 (90%)


Query: 199 tcccgcaagagccttcacagcacttcttgattcctgggcagt 240
            |||||||| ||||||||||||||||| |||  ||||||||||
Sbjct: 537 tcccgcaaaagccttcacagcacttcatgacccctgggcagt 496


 dbj|D50321|PIGWAPC Pig DNA for elafin family member protein, complete cds
           Length = 3670

 Score = 42.1 bits (21), Expect = 0.061
 Identities = 27/29 (93%), Positives = 27/29 (93%)


Query: 17   tttattggaaagtggatgagagaggcagc 45
            ||||||||||||  |||||||||||||||
Sbjct: 2554 tttattggaaagccgatgagagaggcagc 2526


 Score = 36.2 bits (18), Expect = 3.8
 Identities = 24/26 (92%), Positives = 24/26 (92%)


Query: 215  acagcacttcttgattcctgggcagt 240
            ||||||||||||||  ||||||||||
Sbjct: 1865 acagcacttcttgacccctgggcagt 1840


 dbj|D17756|PIGSPAI2S3 Porcine DNA for SPAI-2, exon 3
          Length = 157

 Score = 42.1 bits (21), Expect = 0.061
 Identities = 27/29 (93%), Positives = 27/29 (93%)


Query: 17  tttattggaaagtggatgagagaggcagc 45
           ||||||||||||  |||||||||||||||
Sbjct: 141 tttattggaaagccgatgagagaggcagc 113
```

```
embIX05710ITBTRSI6 Trypanosoma brucei DNA for trypanosome repeated sequence TRS 1.6
          homol. to reverse transcriptase
          Length = 6826

 Score = 38.2 bits (19), Expect = 0.95
 Identities = 19/19 (100%), Positives = 19/19 (100%)


Query: 372  actgaaacttgtcctttaa 390
            |||||||||||||||||||
Sbjct: 6355 actgaaacttgtcctttaa 6373


dbjID90904ID90904 Synechocystis sp. PCC6803 complete genome, 6/27, 630555-781448
           Length = 150894

 Score = 38.2 bits (19), Expect = 0.95
 Identities = 22/23 (95%), Positives = 22/23 (95%)


Query: 79    tggacagtgtgggaaggggaagg 101
             |||||||||||||| ||||||||
Sbjct: 71155 tggacagtgtgggagggggaagg 71177


gbIAC001228IHSAC001228 244Kb Contig from Human Chromsome 11p15.5 spanning D11S1 through
           D11S25, complete sequence [Homo sapiens]
           Length = 244254


 Score = 38.2 bits (19), Expect = 0.95
 Identities = 19/19 (100%), Positives = 19/19 (100%)


Query: 298   tggggcaggagccaggctt 316
             |||||||||||||||||||
Sbjct: 34181 tggggcaggagccaggctt 34199


gbIL19876IDROCOFACTO Drosophila melanogaster molybdenum cofactor (cin) mRNA, complete cds
           Length = 1806

 Score = 38.2 bits (19), Expect = 0.95
 Identities = 19/19 (100%), Positives = 19/19 (100%)
```

```
Query: 344  ttgcgctttgactttatct 362
             |||||||||||||||||||
Sbjct: 1774 ttgcgctttgactttatct 1792


 gb|L42568|HUMATP1G09 Homo sapiens (clone 1SW11-1) non-gastric H,K-ATPase (ATP1AL1) gene,
             exons 15-17.
             Length = 5389

 Score = 38.2 bits (19), Expect = 0.95
 Identities = 19/19 (100%), Positives = 19/19 (100%)


Query: 122  gatggggcctgtagctctg 140
             |||||||||||||||||||
Sbjct: 4978 gatggggcctgtagctctg 4996


 gb|J00306|HUMSOMI Human somatostatin I gene and flanks.
             Length = 2667

 Score = 38.2 bits (19), Expect = 0.95
 Identities = 22/23 (95%), Positives = 22/23 (95%)


Query: 296  aatggggcaggagccaggcttag 318
             |||||||||||||| ||||||||
Sbjct: 1434 aatggggcaggagcaaggcttag 1412


 gb|AF017113|AF017113 Bacillus subtilis 300-304 degree genomic sequence
              Length = 47739

 Score = 38.2 bits (19), Expect = 0.95
 Identities = 19/19 (100%), Positives = 19/19 (100%)


Query: 6     agaaggaactctttattgg 24
              |||||||||||||||||||
Sbjct: 28176 agaaggaactctttattgg 28194
```

```
 emb|Z99122|BSUB0019 Bacillus subtilis complete genome (section 19 of 21): from 3597091 to
           3809700
           Length = 212610

 Score = 38.2 bits (19), Expect = 0.95
 Identities = 19/19 (100%), Positives = 19/19 (100%)


Query: 6    agaaggaactctttattgg 24
            |||||||||||||||||||
Sbjct: 2117 agaaggaactctttattgg 2099


 gb|L10345|RICAMYBA Oryza sativa beta-amylase gene, complete cds.
           Length = 3043

 Score = 36.2 bits (18), Expect = 3.8
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 54   tgggcatcctgaatggga 71
            ||||||||||||||||||
Sbjct: 1541 tgggcatcctgaatggga 1558


 gb|L10346|RICAMYBB Oryza sativa beta-amylase gene, complete cds.
           Length = 3148

 Score = 36.2 bits (18), Expect = 3.8
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 54   tgggcatcctgaatggga 71
            ||||||||||||||||||
Sbjct: 1652 tgggcatcctgaatggga 1669


 emb|Z49237|HSL27H9 Human DNA from cosmid L27h9, Huntington's Disease Region, chromosome
           4p16.3 contains CpG island
           Length = 39324

 Score = 36.2 bits (18), Expect = 3.8
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 299  ggggcaggagccaggctt 316
            ||||||||||||||||||
Sbjct: 5972 ggggcaggagccaggctt 5955
```

```
 gb|M21005|HUMMRP8A Human migration inhibitory factor-related protein 8 (MRP8) gene,
           complete cds. >gi|2084586|gb|I38532|I38532 Sequence 1
           from patent US 5614397
           Length = 4195

 Score = 36.2 bits (18), Expect = 3.8
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 85   gtgtgggaaggggaaggg 102
            ||||||||||||||||||
Sbjct: 1842 gtgtgggaaggggaaggg 1825

 gb|AC002422|AC002422 Human Chromosome X, complete sequence [Homo sapiens]
             Length = 160091

 Score = 36.2 bits (18), Expect = 3.8
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 101    ggcagcagggacttagga 118
              ||||||||||||||||||
Sbjct: 100622 ggcagcagggacttagga 100639


 emb|X84419|HSTAX1EX1 H.sapiens TAX-1 gene (exon 1)
           Length = 5436

 Score = 36.2 bits (18), Expect = 3.8
 Identities = 21/22 (95%), Positives = 21/22 (95%)


Query: 80   ggacagtgtgggaaggggaagg 101
            |||||||||||| |||||||||
Sbjct: 2199 ggacagtgtgggcaggggaagg 2220

 gb|U04855|IAU04855 Influenza A virus (H1N1) A/swine/Northern Ireland/38 nucleoprotein
           (NP) gene, partial cds.
           Length = 1494

 Score = 36.2 bits (18), Expect = 3.8
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 22  tggaaagtggatgagaga 39
           ||||||||||||||||||
Sbjct: 303 tggaaagtggatgagaga 320
```

```
gb|U04856|IAU04856 Influenza A virus (H1N1) A/swine/Cambridge/39 nucleoprotein (NP)
          gene, partial cds.
          Length = 1494

 Score = 36.2 bits (18), Expect = 3.8
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 22  tggaaagtggatgagaga 39
           ||||||||||||||||||
Sbjct: 303 tggaaagtggatgagaga 320


gb|M30746|FLANPB Influenza A/Wilson-Smith/33 (H1N1) nucleoprotein (seg 5) mRNA,
          complete cds.
          Length = 1565

 Score = 36.2 bits (18), Expect = 3.8
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 22  tggaaagtggatgagaga 39
           ||||||||||||||||||
Sbjct: 348 tggaaagtggatgagaga 365


gb|M63769|FLANPAW Influenza A/Swine/Cambridge/1/35 (H1N1) nucleoprotein mRNA,
          complete cds.
          Length = 1565


 Score = 36.2 bits (18), Expect = 3.8
 Identities = 18/18 (100%), Positives = 18/18 (100%)


Query: 22  tggaaagtggatgagaga 39
           ||||||||||||||||||
Sbjct: 348 tggaaagtggatgagaga 365
```

```
CPU time:    33.29 user secs.          1.18 sys. secs            34.47 total secs.

  Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences
    Posted date:  Jul 26, 1998  8:01 AM
  Number of letters in database: 773,827,195
  Number of sequences in database:  355,285

Lambda     K      H
    1.37    0.711     1.31

Gapped
Lambda     K      H
    1.37    0.711     1.31


Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 157455
Number of Sequences: 355285
Number of extensions: 157455
Number of successful extensions: 40253
Number of sequences better than 10: 37
length of query: 399
length of database: 773827195
effective HSP length: 19
effective length of query: 380
effective length of database: 767076780
effective search space: 291489176400

T: 0
A: 0
X1: 6 (11.9 bits)
X2: 25 (49.6 bits)
S1: 0 ( 0.5 bits)
S2: 18 (36.2 bits)
```

APPENDIX A.3

MOTIFS RETRIEVED FROM PROSITE IN PRATT WEBSITE

PRATT results

Results of Search:

Program: PRATT 2.1

```
------------------------------------------------------------
                Pratt version 2.1, Febr. 1997
                Written by Inge Jonassen,
                   University of Bergen
                         Norway
                  email: inge@ii.uib.no
                 For more information, see
             http://www.ii.uib.no/~inge/Pratt.html
------------------------------------------------------------
                       Please quote:
             I.Jonassen, J.F.Collins, D.G.Higgins.
             Protein Science 1995;4(8):1587-1595.
------------------------------------------------------------


                Pratt version 2.1

        Analysing 4 sequences from file /data/web/home/tmp/15585.prattseq

PATTERN CONSERVATION:
     CM: min Nr of Seqs to Match                   4
     C%: min Percentage Seqs to Match          100.0

  PATTERN RESTRICTIONS :
     PP: pos in seq [off,complete,start]         off
     PL: max Pattern Length                       50
     PN: max Nr of Pattern Symbols                50
     PX: max Nr of consecutive x's                 5
     FN: max Nr of flexible spacers                0
     BI: Input Pattern Symbol File               off
     BN: Nr of Pattern Symbols Initial Search     20

  PATTERN SCORING:
     S: Scoring [info,mdl,tree,dist,ppv]        info

  SEARCH PARAMETERS:
     G: Pattern Graph from [seq,al,query]        seq
     E: Search Greediness                          3
     R: Pattern Refinement                        on
     RG: Generalise ambiguous symbols            off
```

```
OUTPUT:
    OF: Output Filename         /data/web/home/tmp/15585.prattres
    OP: PROSITE Pattern Format                          on
    ON: max number patterns                             50
    OA: max number Alignments                           50
    M: Print Patterns in sequences                      on
    MR: ratio for printing                              10
    MV: print vertically                               off


Pratt run started at Wed Jul 29 10:08:57 1998


Best Patterns before refinement:
          fitness     hits(seqs)    Pattern
     1:  183.4822      4(   4)   N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-(
     2:  183.4822      4(   4)   T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-I
     3:  183.4822      4(   4)   I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-,
     4:  179.3121      4(   4)   A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-x-G-:
     5:  179.3121      4(   4)   C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-x-(
     6:  179.3121      4(   4)   F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-:
     7:  179.3121      4(   4)   F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-(
     8:  175.1421      4(   4)   D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-x-G-x-Y-T-K-x(3)-F-I
     9:  170.9720      4(   4)   T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-x-G-x-Y-T-K-:
    10:  170.9720      4(   4)   S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S·
    11:  166.8020      4(   4)   K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I·
    12:  166.8020      4(   4)   C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G·
    13:  162.6319      4(   4)   G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-x-G-x-Y-T-K-x(3)-F-L-x-W-x(
    14:  158.4619      4(   4)   R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-Y-I
    15:  158.4619      4(   4)   D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-T-'
    16:  158.4619      4(   4)   V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-D-'
    17:  158.4619      4(   4)   Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-K-I
    18:  158.4619      4(   4)   P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-x-I
    19:  158.4619      4(   4)   V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-:
    20:  158.4619      4(   4)   E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-T-I
    21:  158.4619      4(   4)   L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-P-H-V-'
    22:  158.4619      4(   4)   L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D-S-G-G-I
    23:  150.1218      4(   4)   S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q-G-D·
    24:  145.9518      4(   4)   R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C-Q·
    25:  145.9518      4(   4)   G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D-A-C·
    26:  145.9518      4(   4)   A-N-S-F-L-E-E-x-K-x-G-x(2)-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-x-F-E-D-x(3)-T-x-E-F-W-x(2)-Y-x-I
    27:  145.9518      4(   4)   R-A-N-S-F-L-E-E-x-K-x-G-x(2)-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-x-F-E-D-x(3)-T-x-E-F-W-x(2)-Y-:
    28:  141.7817      4(   4)   E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T-x(3)-D·
    29:  141.7817      4(   4)   V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C·
    30:  141.7817      4(   4)   G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M·
    31:  141.7817      4(   4)   K-x-G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x·
    32:  141.7817      4(   4)   Q-K-x-G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T·
    33:  137.6117      4(   4)   M-x-Q-K-x-G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x·
    34:  137.6117      4(   4)   L-M-x-Q-K-x-G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F·
    35:  137.6117      4(   4)   V-x-P-A-C-L-P-x(2)-D-x-A-x(3)-L-M-x-Q-K-x-G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V·
    36:  137.6117      4(   4)   R-I-V-G-G-x(2)-C-x(2)-G-E-C-P-W-Q-A-x-L-x-N-E-x(2)-H-x-F-C-G-G-T-I-L-x-E-x(3)-L-T-A-A-H-C-x(
    37:  137.6117      4(   4)   R-x(2)-R-A-N-S-F-L-E-E-x-K-x-G-x(2)-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-x-F-E-D-x(3)-T-x-E-F-W-:
    38:  137.6117      4(   4)   L-x-R-x(2)-R-A-N-S-F-L-E-E-x-K-x-G-x(2)-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-x-F-E-D-x(3)-T-x-E-I
    39:  133.4416      4(   4)   R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x-T·
    40:  133.4416      4(   4)   G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y-x·
    41:  133.4416      4(   4)   F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S-x(2)-F-x-I-T-x-N-M-F-C-A-G-Y·
    42:  133.4416      4(   4)   C-L-P-x(2)-D-x-A-x(3)-L-M-x-Q-K-x-G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D·
    43:  133.4416      4(   4)   A-C-L-P-x(2)-D-x-A-x(3)-L-M-x-Q-K-x-G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V·
    44:  133.4416      4(   4)   R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-x-F-E-D-x(3)-T-x-E-F-W-x(2)-Y-x-D-G-D-Q-C-x(3)-P-C-x(3)-G-x-C·
    45:  133.4416      4(   4)   E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-x-F-E-D-x(3)-T-x-E-F-W-x(2)-Y-x-D-G-D-Q-C-x(3)-P-C-x(3)-G-x·
    46:  133.4416      4(   4)   E-E-x-K-x-G-x(2)-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-x-F-E-D-x(3)-T-x-E-F-W-x(2)-Y-x-D-G-D-Q-C-:
    47:  133.4416      4(   4)   L-E-E-x-K-x-G-x(2)-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-x-F-E-D-x(3)-T-x-E-F-W-x(2)-Y-x-D-G-D-Q-(
    48:  133.4416      4(   4)   A-x(3)-L-x-R-x(2)-R-A-N-S-F-L-E-E-x-K-x-G-x(2)-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-x-F-E-D-x(3)·
    49:  129.2716      4(   4)   A-x(3)-L-M-x-Q-K-x-G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K-x-S·
    50:  129.2716      4(   4)   D-x-A-x(3)-L-M-x-Q-K-x-G-x-V-S-G-F-G-R-x(2)-E-x-G-R-x-S-x(2)-L-K-x-L-E-V-P-Y-V-D-R-x(2)-C-K·
```

```
Best Patterns (after refinement phase):
          fitness     hits(seqs)    Pattern
A   1:  199.4173       4(   4)    N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-
B   2:  198.8288       4(   4)    T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-
C   3:  198.8288       4(   4)    I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-
D   4:  198.4721       4(   4)    A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-
E   5:  198.4721       4(   4)    C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-
F   6:  198.4721       4(   4)    F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-
G   7:  197.5681       4(   4)    D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y-T-
H   8:  194.6588       4(   4)    F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-
I   9:  193.3205       4(   4)    T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[F
J  10:  192.7154       4(   4)    S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-
K  11:  188.5453       4(   4)    K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-
L  12:  188.5453       4(   4)    C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-
M  13:  188.2314       4(   4)    G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y-T-K-[LV]-[
N  14:  186.8953       4(   4)    L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[
O  15:  186.6020       4(   4)    R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-
P  16:  186.6020       4(   4)    D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-
Q  17:  186.6020       4(   4)    V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-
R  18:  186.6020       4(   4)    Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-
S  19:  186.6020       4(   4)    P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-
T  20:  186.6020       4(   4)    V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-
U  21:  186.6020       4(   4)    E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-
V  22:  186.6020       4(   4)    L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-
W  23:  181.3531       4(   4)    R-I-V-G-G-[DQR]-[DE]-C-x-[DEP]-G-E-C-P-W-Q-A-[LV]-L-[IV]-N-E-[EK]-[GN]-E-[EG]-F-C-G-G-T-I-L-
X  24:  178.5552       4(   4)    S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE
Y  25:  175.8864       4(   4)    V-[AV]-P-A-C-L-P-[EQ]-x-D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-
Z  26:  174.9273       4(   4)    C-L-P-[EQ]-x-D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-
a  27:  174.9273       4(   4)    A-C-L-P-[EQ]-x-D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-
b  28:  174.3852       4(   4)    R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-
c  29:  174.3852       4(   4)    G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-
d  30:  173.9720       4(   4)    D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-
e  31:  173.4817       4(   4)    Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-
f  32:  173.4817       4(   4)    K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-
g  33:  173.4769       4(   4)    G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[S
h  34:  172.5060       4(   4)    M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-
i  35:  172.5060       4(   4)    L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[N

i  35:  172.5060       4(   4)    L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[N
j  36:  170.5410       4(   4)    A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-
k  37:  170.2151       4(   4)    E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-
l  38:  170.1698       4(   4)    V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-
m  39:  168.2252       4(   4)    R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[
n  40:  168.2252       4(   4)    G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-
o  41:  168.2252       4(   4)    F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-
p  42:  167.5345       4(   4)    A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[
q  43:  167.5345       4(   4)    R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-
r  44:  164.4753       4(   4)    L-[AEQ]-R-[ATV]-x-R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-
s  45:  161.8011       4(   4)    R-[ATV]-x-R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS
t  46:  160.8212       4(   4)    E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-
u  47:  160.8212       4(   4)    L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-
v  48:  159.7174       4(   4)    A-x(2)-[FIV]-L-[AEQ]-R-[ATV]-x-R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-
w  49:  157.3435       4(   4)    R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q-C-[E
x  50:  157.3435       4(   4)    E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q-C-


Best patterns with alignments:
          fitness     hits(seqs)    Pattern
A   1:  199.4173       4(   4)    N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   432-    481: faite NMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGKYG vytkl
gi|119761|sp|P00742|FA10_HUMAN :   410-    459: fiitq NMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYG iytkv
gi|119759|sp|P00743|FA10_BOVIN :   408-    457: ftitp NMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKFG vytkv
     gi|180336 :   380-   429: fiitq NMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYG iytkv

B   2:  198.8288       4(   4)    T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   430-    479: tnfai TENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGK ygvyt
gi|119761|sp|P00742|FA10_HUMAN :   408-    457: ssfii TQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGK ygiyt
gi|119759|sp|P00743|FA10_BOVIN :   406-    455: ssfti TPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGK fgvyt
     gi|180336 :   378-   427: ssfii TQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGK ygiyt

C   3:  198.8288       4(   4)    I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   429-    478: stnfa ITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKG kygvy
```

```
gi|119761|sp|P00742|FA10_HUMAN :   407-   456: sssfi ITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKG kygiy
gi|119759|sp|P00743|FA10_BOVIN :   405-   454: sssft ITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKG kfgvy
      gi|180336 :   377-   426: sssfi ITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKG kygiy

D  4:  198.4721       4(   4)   A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   436-   485: enmfc AGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGKYGVYTK lsrfl
gi|119761|sp|P00742|FA10_HUMAN :   414-   463: qnmfc AGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTK vtafl
gi|119759|sp|P00743|FA10_BOVIN :   412-   461: pnmfc AGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKFGVYTK vsnfl
      gi|180336 :   384-   433: qnmfc AGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTK vtafl

E  5:  198.4721       4(   4)   C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   435-   484: tenmf CAGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGKYGVYT klsrf
gi|119761|sp|P00742|FA10_HUMAN :   413-   462: tqnmf CAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYT kvtaf
gi|119759|sp|P00743|FA10_BOVIN :   411-   460: tpnmf CAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKFGVYT kvsnf
      gi|180336 :   383-   432: tqnmf CAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYT kvtaf

F  6:  198.4721       4(   4)   F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   434-   483: itenm FCAGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGKYGVY tklsr
gi|119761|sp|P00742|FA10_HUMAN :   412-   461: itqnm FCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIY tkvta
gi|119759|sp|P00743|FA10_BOVIN :   410-   459: itpnm FCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKFGVY tkvsn
      gi|180336 :   382-   431: itqnm FCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIY tkvta

G  7:  197.5681       4(   4)   D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y-T-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   444-   493: eteqk DACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGKYGVYTKLSrFLRWV rtvmr
gi|119761|sp|P00742|FA10_HUMAN :   422-   471: dtkqe DACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTaFLKWI drsmk
gi|119759|sp|P00743|FA10_BOVIN :   420-   469: dtqpe DACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKFGVYTKVSnFLKWI dkimk
      gi|180336 :   392-   441: dtkqe DACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTaFLKWI drsmk

H  8:  194.6588       4(   4)   F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   427-   476: kqstn FaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEGCAR kgkyg
gi|119761|sp|P00742|FA10_HUMAN :   405-   454: klsss FiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCAR kgkyg
gi|119759|sp|P00743|FA10_BOVIN :   403-   452: klsss FtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCAR kgkfg
```

gi|119759|sp|P00743|FA10_BOVIN :   403-   452: klsss FtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCAR kgkfg

```
      gi|180336 :   375-   424: klsss FiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCAR kgkyg

I  9:  193.3205       4(   4)   T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[F
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   440-   489: cagye TeQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEGCARKGKYGVYTKLSrF lrwvr
gi|119761|sp|P00742|FA10_HUMAN :   418-   467: cagyd TkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTaF lkwid
gi|119759|sp|P00743|FA10_BOVIN :   416-   465: cagyd TqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKFGVYTKVSnF lkwid
      gi|180336 :   388-   437: cagyd TkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTaF lkwid

J  10:  192.7154      4(   4)   S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-[FY]-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   424-   473: stckq STNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSWGEG carkg
gi|119761|sp|P00742|FA10_HUMAN :   402-   451: nsckl SSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEG carkg
gi|119759|sp|P00743|FA10_BOVIN :   400-   449: stckl SSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWGEG carkg
      gi|180336 :   372-   421: nsckl SSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWGEG carkg

K  11:  188.5453      4(   4)   K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-R-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   422-   471: drstc KqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSWG egcar
gi|119761|sp|P00742|FA10_HUMAN :   400-   449: drnsc KlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWG egcar
gi|119759|sp|P00743|FA10_BOVIN :   398-   447: drstc KlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSWG egcar
      gi|180336 :   370-   419: drnsc KlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSWG egcar

L  12:  188.5453      4(   4)   C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-G-G-P-H-V-T-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   421-   470: vdrst CKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVTGIVSW gegca
gi|119761|sp|P00742|FA10_HUMAN :   399-   448: vdrns CKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSW gegca
gi|119759|sp|P00743|FA10_BOVIN :   397-   446: vdrst CKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGIVSW gegca
      gi|180336 :   369-   418: vdrns CKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGIVSW gegca

M  13:  188.2314      4(   4)   G-D-S-G-G-P-H-V-T-R-[FY]-K-D-T-Y-F-V-T-G-I-V-S-W-G-E-G-C-A-R-K-G-K-[FY]-G-[IV]-Y-T-K-[LV]-[
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   448-   497: kdacq GDSGGPHVTRYKDTYFVTGIVSWGEGCARKGKYGVYTKLSrFLRWVRtvM rqk
gi|119761|sp|P00742|FA10_HUMAN :   426-   475: edacq GDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTaFLKWIDrsM ktrgl
gi|119759|sp|P00743|FA10_BOVIN :   424-   473: edacq GDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKFGVYTKVSnFLKWIDkiM karag
      gi|180336 :   396-   445: edacq GDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTaFLKWIDrsM ktrgl
```

```
     gi|180336 :   396-   445: edacq GDSGGPHVTRFKDTYPVTGIVSWGEGCARKGKYGIYTKVTaFLKWIDrsM ktrgl

N  14:  186.8953       4(   4)   L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   408-   457: rlskr LKVLEVPYVDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTR ykdty
gi|119761|sp|P00742|FA10_HUMAN :   386-   435: rqstr LKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTR fkdty
gi|119759|sp|P00743|FA10_BOVIN :   384-   433: rlsst LKMLEVPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTR fkdty
     gi|180336 :   356-   405: rqstr LKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTR fkdty

O  15:  186.6020       4(   4)   R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-S-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   418-   467: vpyvd RSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVTGI vswge
gi|119761|sp|P00742|FA10_HUMAN :   396-   445: vpyvd RNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTGI vswge
gi|119759|sp|P00743|FA10_BOVIN :   394-   443: vpyvd RSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTGI vswge
     gi|180336 :   366-   415: vpyvd RNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYPVTGI vswge

P  16:  186.6020       4(   4)   D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-D-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   417-   466: evpyv DRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVTG ivswg
gi|119761|sp|P00742|FA10_HUMAN :   395-   444: evpyv DRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTG ivswg
gi|119759|sp|P00743|FA10_BOVIN :   393-   442: evpyv DRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVTG ivswg
     gi|180336 :   365-   414: evpyv DRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVTG ivswg

Q  17:  186.6020       4(   4)   V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-G-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   416-   465: levpy VDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFVT givsw
gi|119761|sp|P00742|FA10_HUMAN :   394-   443: levpy VDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVT givsw
gi|119759|sp|P00743|FA10_BOVIN :   392-   441: levpy VDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFVT givsw
     gi|180336 :   364-   413: levpy VDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFVT givsw

R  18:  186.6020       4(   4)   Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-Q-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   415-   464: vlevp YVDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYFV tgivs
gi|119761|sp|P00742|FA10_HUMAN :   393-   442: mlevp YVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFV tgivs
gi|119759|sp|P00743|FA10_BOVIN :   391-   440: mlevp YVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYFV tgivs
     gi|180336 :   363-   412: mlevp YVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYFV tgivs


S  19:  186.6020       4(   4)   P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-C-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   414-   463: kvlev PYVDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTYF vtgiv
gi|119761|sp|P00742|FA10_HUMAN :   392-   441: kmlev PYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYF vtgiv
gi|119759|sp|P00743|FA10_BOVIN :   390-   439: kmlev PYVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTYF vtgiv
     gi|180336 :   362-   411: kmlev PYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTYF vtgiv

T  20:  186.6020       4(   4)   V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-A-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   413-   462: lkvle VPYVDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDTY fvtgi
gi|119761|sp|P00742|FA10_HUMAN :   391-   440: lkmle VPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTY fvtgi
gi|119759|sp|P00743|FA10_BOVIN :   389-   438: lkmle VPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDTY fvtgi
     gi|180336 :   361-   410: lkmle VPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDTY fvtgi

U  21:  186.6020       4(   4)   E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-D-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   412-   461: rlkvl EVPYVDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKDT yfvtg
gi|119761|sp|P00742|FA10_HUMAN :   390-   439: rlkml EVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDT yfvtg
gi|119759|sp|P00743|FA10_BOVIN :   388-   437: tlkml EVPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKDT yfvtg
     gi|180336 :   360-   409: rlkml EVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKDT yfvtg

V  22:  186.6020       4(   4)   L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE]-T-x-[PQ]-[EK]-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   411-   460: krlkv LEVPYVDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPHVTRYKD tyfvt
gi|119761|sp|P00742|FA10_HUMAN :   389-   438: trlkm LEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKD tyfvt
gi|119759|sp|P00743|FA10_BOVIN :   387-   436: stlkm LEVPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPHVTRFKD tyfvt
     gi|180336 :   359-   408: trlkm LEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPHVTRFKD tyfvt

W  23:  181.3531       4(   4)   R-I-V-G-G-[DQR]-[DE]-C-x-[DEP]-G-E-C-P-W-Q-A-[LV]-L-[IV]-N-E-[EK]-[GN]-E-[EG]-F-C-G-G-T-I-L-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   265-   314: pnvdt RIVGGDECrPGECPWQAVLINEKGEEFCGGTILNEdFILTAAHCInQSKE ikvvv
gi|119761|sp|P00742|FA10_HUMAN :   243-   292: dnnlt RIVGGQECkDGECPWQALLINEENEGFCGGTILSEfYILTAAHCLyQAKR fkvrv
gi|119759|sp|P00743|FA10_BOVIN :   241-   290: gsqvv RIVGGRDCaEGECPWQALLVNEENEGFCGGTILNEfYVLTAAHCLhQAKR ftvrv
     gi|180336 :   216-   265: dnnlt RIVGGQECkDGECPWQALLINEENEGFCGGTILSEfYILTAAHCLyQAKR fegdr

X  24:  178.5552       4(   4)   S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE
Occurrences: 4(4)
```

```
X 24: 176.3332      4(   4)   S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-[DE
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   405-   454: eagrl SkrLKVLEVPYVDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGGPH vtryk
gi|119761|sp|P00742|FA10_HUMAN :   383-   432: ekgrq StrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPH vtrfk
gi|119759|sp|P00743|FA10_BOVIN :   381-   430: ekgrl SstLKMLEVPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGGPH vtrfk
     gi|180336 :   353-   402: ekgrq StrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGGPH vtrfk

Y 25: 175.8864      4(   4)   V-[AV]-P-A-C-L-P-[EQ]-x-D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   369-   418: qfsey VVPACLPQaDFANEVLMNQKSGMVSGFGREFEaGRlSkrLKVLEVPYVDR stckq
gi|119761|sp|P00742|FA10_HUMAN :   347-   396: tfrmn VAPACLPErDWAESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDR nsckl
gi|119759|sp|P00743|FA10_BOVIN :   345-   394: rfrrn VAPACLPEkDWAEATLMTQKTGIVSGFGRTHEkGRlSstLKMLEVPYVDR stckl
     gi|180336 :   317-   366: tfrmn VAPACLPErDWAESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDR nsckl

Z 26: 174.9273      4(   4)   C-L-P-[EQ]-x-D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-]
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   373-   422: yvvpa CLPQaDFANEVLMNQKSGMVSGFGREFEaGRlSkrLKVLEVPYVDRSTCK qstnf
gi|119761|sp|P00742|FA10_HUMAN :   351-   400: nvapa CLPErDWAESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCK lsssf
gi|119759|sp|P00743|FA10_BOVIN :   349-   398: nvapa CLPEkDWAEATLMTQKTGIVSGFGRTHEkGRlSstLKMLEVPYVDRSTCK lsssf
     gi|180336 :   321-   370: nvapa CLPErDWAESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCK lsssf

a 27: 174.9273      4(   4)   A-C-L-P-[EQ]-x-D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-(
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   372-   421: eyvvp ACLPQaDFANEVLMNQKSGMVSGFGREFEaGRlSkrLKVLEVPYVDRSTC kqstn
gi|119761|sp|P00742|FA10_HUMAN :   350-   399: mnvap ACLPErDWAESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSC klsss
gi|119759|sp|P00743|FA10_BOVIN :   348-   397: rnvap ACLPEkDWAEATLMTQKTGIVSGFGRTHEkGRlSstLKMLEVPYVDRSTC klsss
     gi|180336 :   320-   369: mnvap ACLPErDWAESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSC klsss

b 28: 174.3852      4(   4)   R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-Y-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   403-   452: efeag RlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSGG phvtr
gi|119761|sp|P00742|FA10_HUMAN :   381-   430: thekg RqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGG phvtr
gi|119759|sp|P00743|FA10_BOVIN :   379-   428: thekg RlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSGG phvtr
     gi|180336 :   351-   400: thekg RqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSGG phvtr

c 29: 174.3852      4(   4)   G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-A-G-
Occurrences: 4(4)


Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   402-   451: refea GRlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCAGYETeQKDACQGDSG gphvt
gi|119761|sp|P00742|FA10_HUMAN :   380-   429: rthek GRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSG gphvt
gi|119759|sp|P00743|FA10_BOVIN :   378-   427: rthek GRlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGDSG gphvt
     gi|180336 :   350-   399: rthek GRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGDSG gphvt

d 30: 173.9720      4(   4)   D-[FW]-A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-]
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   378-   427: clpqa DFANEVLMNQKSGMVSGFGREFEaGRlSkrLKVLEVPYVDRSTCKqSTNF aiten
gi|119761|sp|P00742|FA10_HUMAN :   356-   405: clper DWAESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSF iitqn
gi|119759|sp|P00743|FA10_BOVIN :   354-   403: clpek DWAEATLMTQKTGIVSGFGRTHEkGRlSstLKMLEVPYVDRSTCKlSSSF titpn
     gi|180336 :   326-   375: clper DWAESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSF iitqn

e 31: 173.4817      4(   4)   Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   387-   436: evlmn QKSGMVSGFGREFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCA gyete
gi|119761|sp|P00742|FA10_HUMAN :   365-   414: stlmt QKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCA gydtk
gi|119759|sp|P00743|FA10_BOVIN :   363-   412: atlmt QKTGIVSGFGRTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCA gydtq
     gi|180336 :   335-   384: stlmt QKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCA gydtk

f 32: 173.4817      4(   4)   K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   388-   437: vlmnq KSGMVSGFGREFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCAG yeteq
gi|119761|sp|P00742|FA10_HUMAN :   366-   415: tlmtq KTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAG ydtkq
gi|119759|sp|P00743|FA10_BOVIN :   364-   413: tlmtq KTGIVSGFGRTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCAG ydtqp
     gi|180336 :   336-   385: tlmtq KTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAG ydtkq

g 33: 173.4769      4(   4)   G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[S'
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   390-   439: mnqks GMVSGFGREFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCAGYE teqkd
gi|119761|sp|P00742|FA10_HUMAN :   368-   417: mtqkt GIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYD tkqed
gi|119759|sp|P00743|FA10_BOVIN :   366-   415: mtqkt GIVSGFGRTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCAGYD tqped
     gi|180336 :   338-   387: mtqkt GIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYD tkqed

h 34: 172.5060      4(   4)   M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   385-   434: anevl MNQKSGMVSGFGREFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENMF cagye
gi|119761|sp|P00742|FA10_HUMAN :   363-   412: aestl MTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMF cagyd
gi|119759|sp|P00743|FA10_BOVIN :   361-   410: aeatl MTQKTGIVSGFGRTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNMF cagyd
     gi|180336 :   333-   382: aestl MTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMF cagyd

i  35:  172.5060       4(   4)   L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[N:
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   384-   433: fanev LMNQKSGMVSGFGREFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENM fcagy
gi|119761|sp|P00742|FA10_HUMAN :   362-   411: waest LMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNM fcagy
gi|119759|sp|P00743|FA10_BOVIN :   360-   409: waeat LMTQKTGIVSGFGRTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNM fcagy
     gi|180336 :   332-   381: waest LMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNM fcagy

j  36:  170.5410       4(   4)   A-[EN]-[AES]-[TV]-L-M-[NT]-Q-K-[ST]-G-[IM]-V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-]
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   380-   429: pqadf ANEVLMNQKSGMVSGFGREFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaI tenmf
gi|119761|sp|P00742|FA10_HUMAN :   358-   407: perdw AESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiI tqnmf
gi|119759|sp|P00743|FA10_BOVIN :   356-   405: pekdw AEATLMTQKTGIVSGFGRTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtI tpnmf
     gi|180336 :   328-   377: perdw AESTLMTQKTGIVSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiI tqnmf

k  37:  170.2151       4(   4)   E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[EPQ]-N-M-F-C-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   400-   449: fgref EaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCAGYETeQKDACQGD sggph
gi|119761|sp|P00742|FA10_HUMAN :   378-   427: fgrth EkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGD sggph
gi|119759|sp|P00743|FA10_BOVIN :   376-   425: fgrth EkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDACQGD sggph
     gi|180336 :   348-   397: fgrth EkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDACQGD sggph

l  38:  170.1698       4(   4)   V-S-G-F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   392-   440: qksgm VSGFGREFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCAGYET eqkda
gi|119761|sp|P00742|FA10_HUMAN :   370-   418: qktgi VSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDT kqeda
gi|119759|sp|P00743|FA10_BOVIN :   368-   416: qktgi VSGFGRTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCAGYDT qpeda
     gi|180336 :   340-   388: qktgi VSGFGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDT kqeda

m  39:  168.2252       4(   4)   R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-[]
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   397-   446: vsgfg REFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCAGYETeQKDAC qgdsg
gi|119761|sp|P00742|FA10_HUMAN :   375-   424: vsgfg RTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDAC qgdsg
gi|119759|sp|P00743|FA10_BOVIN :   373-   422: vsgfg RTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDAC qgdsg
     gi|180336 :   345-   394: vsgfg RTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDAC qgdsg

n  40:  168.2252       4(   4)   G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-T-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   396-   445: mvsgf GREFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCAGYETeQKDA cqgds
gi|119761|sp|P00742|FA10_HUMAN :   374-   423: ivsgf GRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDA cqgds
gi|119759|sp|P00743|FA10_BOVIN :   372-   421: ivsgf GRTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCAGYDTqPEDA cqgds
     gi|180336 :   344-   393: ivsgf GRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQEDA cqgds

o  41:  168.2252       4(   4)   F-G-R-[ET]-[FH]-E-x-G-R-x-S-x(2)-L-K-[MV]-L-E-V-P-Y-V-D-R-[NS]-[ST]-C-K-x-S-[ST]-[NS]-F-x-I-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :   395-   444: gmvsg FGREFEaGRlSkrLKVLEVPYVDRSTCKqSTNFaITENMFCAGYETeQKD acqgd
gi|119761|sp|P00742|FA10_HUMAN :   373-   422: givsg FGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQED acqgd
gi|119759|sp|P00743|FA10_BOVIN :   371-   420: givsg FGRTHEkGRlSstLKMLEVPYVDRSTCKlSSSFtITPNMFCAGYDTqPED acqgd
     gi|180336 :   343-   392: givsg FGRTHEkGRqStrLKMLEVPYVDRNSCKlSSSFiITQNMFCAGYDTkQED acqgd

p  42:  167.5345       4(   4)   A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[]
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :    66-   115: ertkr ANSFLEEMKqGnIERECnEErCSkEEAREAFEDNEkTEEFWNiYvDGDQC ssnpc
gi|119761|sp|P00742|FA10_HUMAN :    50-    99: arvtr ANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQC etspc
gi|119759|sp|P00743|FA10_BOVIN :    49-    98: qrarr ANSFLEEVKqGnLERECLEEaCSlEEAREVFEDAEqTDEFWSkYkDGDQC eghpc
     gi|180336 :    23-    72: arvtr ANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQC etspc

q  43:  167.5345       4(   4)   R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :    65-   114: lertk RANSFLEEMKqGnIERECnEErCSkEEAREAFEDNEkTEEFWNiYvDGDQ cssnp
gi|119761|sp|P00742|FA10_HUMAN :    49-    98: larvt RANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQ cetsp
gi|119759|sp|P00743|FA10_BOVIN :    48-    97: lqrar RANSFLEEVKqGnLERECLEEaCSlEEAREVFEDAEqTDEFWSkYkDGDQ ceghp
     gi|180336 :    22-    71: larvt RANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQ cetsp

r  44:  164.4753       4(   4)   L-[AEQ]-R-[ATV]-x-R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-]
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :    60-   109: sadkf LERTkRANSFLEEMKqGnIERECnEErCSkEEAREAFEDNEkTEEFWNiY vdgdq
gi|119761|sp|P00742|FA10_HUMAN :    44-    93: qanni LARVtRANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkY kdgdq
gi|119759|sp|P00743|FA10_BOVIN :    43-    92: qahrv LQRArRANSFLEEVKqGnLERECLEEaCSlEEAREVFEDAEqTDEFWSkY kdgdq
     gi|180336 :    17-    66: qanni LARVtRANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkY kdgdq
```

```
s 45: 161.8011      4(   4)   R-[ATV]-x-R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :     62-    111: dkfle RTkRANSFLEEMKqGnIERECnEErCSkEEAREAFEDNEkTEEFWNiYvD gdqcs
gi|119761|sp|P00742|FA10_HUMAN :     46-     95: nnila RVtRANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkD gdqce
gi|119759|sp|P00743|FA10_BOVIN :     45-     94: hrvlq RArRANSFLEEVKqGnLEREClEEaCSlEEAREVFEDAEqTDEFWSkYkD gdqce
     gi|180336 :     19-     68: nnila RVtRANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkD gdqce

t 46: 160.8212      4(   4)   E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W.
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :     71-    120: ansfl EEMKqGnIERECnEErCSkEEAREAFEDNEkTEEFWNiYvDGDQCSSnPC hyggq
gi|119761|sp|P00742|FA10_HUMAN :     55-    104: ansfl EEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQCETsPC qnqgk
gi|119759|sp|P00743|FA10_BOVIN :     54-    103: ansfl EEVKqGnLEREClEEaCSlEEAREVFEDAEqTDEFWSkYkDGDQCEGhPC lnqgh
     gi|180336 :     28-     77: ansfl EEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQCETsPC qnqgk

u 47: 160.8212      4(   4)   L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F.
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :     70-    119: ransf LEEMKqGnIERECnEErCSkEEAREAFEDNEkTEEFWNiYvDGDQCSSnP chygg
gi|119761|sp|P00742|FA10_HUMAN :     54-    103: ransf LEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQCETsP cqnqg
gi|119759|sp|P00743|FA10_BOVIN :     53-    102: ransf LEEVKqGnLEREClEEaCSlEEAREVFEDAEqTDEFWSkYkDGDQCEGhP clnqg
     gi|180336 :     27-     76: ransf LEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQCETsP cqnqg

v 48: 159.7174      4(   4)   A-x(2)-[FIV]-L-[AEQ]-R-[ATV]-x-R-A-N-S-F-L-E-E-[MV]-K-x-G-x-[IL]-E-R-E-C-x-E-E-x-C-S-x-E-E-,
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :     56-    105: ikkes AdkFLERTkRANSFLEEMKqGnIERECnEErCSkEEAREAFEDNEkTEEF wniyv
gi|119761|sp|P00742|FA10_HUMAN :     40-     89: irreq AnnILARVtRANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEF wnkyk
gi|119759|sp|P00743|FA10_BOVIN :     39-     88: lprdq AhrVLQRArRANSFLEEVKqGnLEREClEEaCSlEEAREVFEDAEqTDEF wskyk
     gi|180336 :     13-     62: irreq AnnILARVtRANSFLEEMKkGhLERECmEEtCSyEEAREVFEDSDkTNEF wnkyk

w 49: 157.3435      4(   4)   R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q-C-[E:
Occurrences: 4(4)
gi|119760|sp|P25155|FA10_CHICK :     80-    129: qgnie RECnEErCSkEEAREAFEDNEkTEEFWNiYvDGDQCSSnPChyGGqCKDG lgsyt
gi|119761|sp|P00742|FA10_HUMAN :     64-    113: kghle RECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQCETsPCqnQGkCKDG lgeyt
gi|119759|sp|P00743|FA10_BOVIN :     63-    112: qgnle REClEEaCSlEEAREVFEDAEqTDEFWSkYkDGDQCEGhPClnQGhCKDG igdyt
     gi|180336 :     37-     86: kghle RECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQCETsPCqnQGkCKDG lgeyt

x 50: 157.3435      4(   4)   E-R-E-C-x-E-E-x-C-S-x-E-E-A-R-E-[AV]-F-E-D-[ANS]-[DE]-x-T-[DEN]-E-F-W-[NS]-x-Y-x-D-G-D-Q-C-
Occurrences: 4(4)


gi|119760|sp|P25155|FA10_CHICK :     79-    128: kqgni ERECnEErCSkEEAREAFEDNEkTEEFWNiYvDGDQCSSnPChyGGqCKD glgsy
gi|119761|sp|P00742|FA10_HUMAN :     63-    112: kkghl ERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQCETsPCqnQGkCKD glgey
gi|119759|sp|P00743|FA10_BOVIN :     62-    111: kqgnl EREClEEaCSlEEAREVFEDAEqTDEFWSkYkDGDQCEGhPClnQGhCKD gigdy
     gi|180336 :     36-     85: kkghl ERECmEEtCSyEEAREVFEDSDkTNEFWNkYkDGDQCETsPCqnQGkCKD glgey


PATTERN MATCHES:
each . represents 10 sequence symbols
A symbol A-Z,a-z (for example A) in the place of a dot indicates the
starting point of a match to this pattern (in the example; pattern A).


-----------------------------------------------
gi|119760|sp|P25155|FA10_CHICK: .....rpt..................W.........YZekNOBAG......
gi|119761|sp|P00742|FA10_HUMAN: ...vptw.................W.........YZecNKADG.......
gi|119759|sp|P00743|FA10_BOVIN: ...vptw.................W.........YdebNJADM........
      gi|180336: .rpw.................W.........YZecNKADG.......

-----------------------------------------------
Number of patterns evaluated by Pratt:25912
Total running time:    3 seconds
```

APPENDIX B

A COMPLETE LIST OF HUMAN DNA

Adipose
NCI_CGAP_Lip2 Bulk, liposarcoma (1730 sequences)

Adrenal gland
NCI_CGAP_AA1 Bulk, 2 pooled adenomas (3363 sequences)
NCI_CGAP_Phe1 Bulk, Pheochromocytoma (1356 sequences)

Blood
NCI_CGAP_HSC1 Flow-sorted, CD34+/CD38- hematopoietic stem cells(824 sequences)

Bone
NCI_CGAP_Ew1 Bulk, Ewing's sarcoma (4762 sequences)
NCI_CGAP_SS1 Bulk, synovial sarcoma (460 sequences)

Bowel (skin primary)
NCI_CGAP_Mel3 (no title) (237 sequences)

Brain
NCI_CGAP_CNS1 Bulk, central nervous system, substantia nigra (386 sequences)

Breast
NCI_CGAP_Br1.1 Bulk, 3 pooled invasive ductal breast tumors including well, moderately, and poorly differentiated; not normalized (normalized version is Br2) (2126 sequences)
NCI_CGAP_Br2 Bulk, 3 pooled invasive ductal tumors including well, moderately, and poorly differentiated; normalized (non-normalized version is Br1.1) (4586 sequences)
NCI_CGAP_Br3 Bulk, poorly differentiated invasive ductal breast tumor (1014 sequences)
NCI_CGAP_Br4 Microdissected, Normal breast ductal tissue (532 sequences)
NCI_CGAP_Br5 Microdissected, infiltrating breast ductal carcinoma (325 sequences)
NCI_CGAP_Br7 Bulk, normal breast (326 sequences)

Colon
NCI_CGAP_Co1 Bulk, moderately differentiated colon adenocarcinoma (300 sequences)
NCI_CGAP_Co10 Bulk, moderately differentiated colon adenocarcinoma; normalized (non-normalized version is Co9) (3587 sequences)
NCI_CGAP_Co11 Bulk, 8 pooled colon adenocarcinomas, including well, moderately, and poorly differentiated (1287 sequences)
NCI_CGAP_Co12 Bulk, 10 pooled colon adenocarcinomas, including well, moderately, and poorly differentiated (1725 sequences)

NCI_CGAP_Co2 Bulk, villous adenoma (929 sequences)
NCI_CGAP_Co3 Bulk, 12 pooled colon adenocarcinomas, including well, moderately, and poorly differentiated; normalized (non-normalized version is Co4) (8561 sequences)
NCI_CGAP_Co4 Bulk, 12 pooled colon adenocarcinomas, including well, moderately, and poorly differentiated; non-normalized (normalized version is Co3) (693 sequences)
NCI_CGAP_Co8 Bulk, 2 pooled adenocarcinomas (1988 sequences)
NCI_CGAP_Co9 Bulk, moderately differentiated colon adenocarcinoma; non-normalized (normalized version is Co10) (3726 sequences)

Germ Cell
NCI_CGAP_GC1 Bulk, germ cell, seminoma (521 sequences)
NCI_CGAP_GC2 Bulk, germ cell, yolk sac tumor (1036 sequences)
NCI_CGAP_GC3 Bulk, 3 pooled samples, including broad spectrum germ cell tumor types (1019 sequences)
NCI_CGAP_GC4 Bulk, 3 pooled samples including broad spectrum germ cell tumor types; normalized (3069 sequences)
NCI_CGAP_GC5 Bulk, 3 pooled germ cell tumors, including mixed seminoma/embryonal, teratoma with adenocarcinoma arising, and seminoma (1191 sequences)

Head and neck
NCI_CGAP_HN3 Bulk, Head and neck, squamous cell carcinoma cell line; primary site: base of tongue; (131 sequences)
NCI_CGAP_HN4 Bulk, Head and neck, squamous cell carcinoma cell line; primary site: Pharynx; non-normalized (656 sequences)

Kidney
NCI_CGAP_Kid1 Bulk, papillary renal cell carcinoma (981 sequences)
NCI_CGAP_Kid3 Bulk, 2 pooled normal samples (3827 sequences)
NCI_CGAP_Kid5 Bulk, 2 pooled tumors, clear cell type, normalized (4646 sequences)
NCI_CGAP_Kid6 Bulk, 5 pooled renal cell carcinomas, clear cell (1983 sequences)
NCI_CGAP_Kid7 Bulk, 5 pooled samples, including broad spectrum of kidney tumor types (0 sequences)

Larynx
NCI_CGAP_Lar1 Bulk, invasive larynx squamous cell carcinoma (1096 sequences)

Liver
NCI_CGAP_Li1 Microdissected, normal liver hepatocytes (matched to Li2) (502 sequences)
NCI_CGAP_Li2 Microdisected, hepatocellular carcinoma (matched to Li1) (307 sequences)
NCI_CGAP_Li5 Bulk, hepatic adenoma (147 sequences)
NCI_CGAP_Pr20 Microdissected, metastatic prostate cancer to liver (166 sequences)

Lung
NCI_CGAP_Lu1 Bulk, poorly differentiated lung neoplasm (2233 sequences)

NCI_CGAP_Lu5 Bulk, Lung, 2 pooled neuroendocrine lung carcinoids, normalized (3402 sequences)
NCI_CGAP_Lu6 Bulk, Lung, small cell carcinoma (45 sequences)

Lymph node
NCI_CGAP_Lym3 Bulk, 10 pooled samples, including broad spectrum of lymphoma tumor types (653 sequences)

Lymph node
NCI_CGAP_HN1 Bulk, squamous cell carcinoma cell line; primary site: head and neck; metastasis to the lymph node. (35 sequences)

Muscle
NCI_CGAP_AR1 Bulk, alveolar rhabdomyosarcoma (355 sequences)
NCI_CGAP_Alv1 Bulk, alveolar rhabdomyosarcoma (4832 sequences)

Neural
NCI_CGAP_Sch1 Bulk, 2 pooled schwannomas (1218 sequences)

Ovary
NCI_CGAP_Ov1 Bulk, serous ovary papillary adenocarcinoma (195 sequences)
NCI_CGAP_Ov2 Bulk, serous ovary papillary adenocarcinoma (3267 sequences)
NCI_CGAP_Ov5 Microdissected, normal ovarian epithelium (167 sequences)
NCI_CGAP_Ov6 Microdissected, normal ovarian stroma (156 sequences)
NCI_CGAP_Ov8 (no title) (24 sequences)

Peripheral nervous system
NCI_CGAP_PNS1 Bulk, dorsal root ganglion (297 sequences)

Pool
Soares NFL T GBC S1 subtracted mix of three normalized libraries (6672 sequences)

Prostate
NCI_CGAP_Pr1 Microdissected, normal prostate epithelium (5689 sequences)
NCI_CGAP_Pr10 Microdissected, invasive prostate tumor (1139 sequences)
NCI_CGAP_Pr11 Microdissected, normal epithelium from normal prostate (1376 sequences)
NCI_CGAP_Pr12 Microdissected, metastatic prostate cancer to bone (3147 sequences)
NCI_CGAP_Pr16 Microdissected, invasive prostate tumor (550 sequences)
NCI_CGAP_Pr18 Microdissected, BPH stroma (671 sequences)

NCI_CGAP_Pr2 Microdissected, low grade prostatic intraepithelial neoplasia (5688 sequences)
NCI_CGAP_Pr21 Bulk, normal prostate; non-normalized (1266 sequences)
NCI_CGAP_Pr22 Bulk, normal prostate; normalized (5867 sequences)
NCI_CGAP_Pr23 Bulk, 7 pooled prostate cancers, including well, moderately, and poorly differentiated (1011 sequences)
NCI_CGAP_Pr24 Cell line, invasive prostate tumor cell line (HPV immortalized) (991 sequences)
NCI_CGAP_Pr25 Cell line, normal prostate epithelial cell line (HPV immortalized) (1441 sequences)
NCI_CGAP_Pr3 Microdissected, invasive prostate tumor (5209 sequences)
NCI_CGAP_Pr4 Microdissected, high grade prostatic intraepithelial neoplasia (659 sequences)
NCI_CGAP_Pr4.1 Microdissected, prostatic intraepithelial neoplasia - high grade (1269 sequences)
NCI_CGAP_Pr5 Microdissected, normal prostate epithelium (805 sequences)
NCI_CGAP_Pr6 Microdissected, low grade prostatic intraepithelial neoplasia (focus #1) (1462 sequences)
NCI_CGAP_Pr7 Microdissected, low grade prostatic intraepithelial neoplasia (focus #2) (468 sequences)
NCI_CGAP_Pr8 Microdissected, invasive prostate tumor (1100 sequences)
NCI_CGAP_Pr9 Microdissected, normal prostate epithelium (1104 sequences)

Stomach
NCI_CGAP_Gas1 Bulk, 4 pooled gastric tumors (849 sequences)

Thymus
NCI_CGAP_Thym1 Bulk, thymoma (0 sequences)

Thyroid
NCI_CGAP_Thy1 Bulk, papillary thyroid carcinoma (2459 sequences)

Tonsil
NCI_CGAP_GCB0 Flow-sorted, pooled tonsil germinal B-cells; non-normalized (825 sequences)
NCI_CGAP_GCB1 Flow-sorted, pooled tonsil germinal B-cells; normalized (47620 sequences)

APPENDIX C

FASTA FORMAT DESCRIPTION

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is:

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below). Before submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue).

The nucleic acid codes supported are:

A --> adenosine
C --> cytidine
G --> guanine
T --> thymidine
U --> uridine
R --> G A (purine)
Y --> T C (pyrimidine)
K --> G T (keto)

M --> A C (amino)
S --> G C (strong)
W --> A T (weak)
B --> G T C
D --> G A T
H --> A C T
V --> G C A
N --> A G C T (any)
- gap of indeterminate length

For those programs that use amino acid query sequences (BLASTP and TBLASTN), the accepted amino acid codes are:

A	alanine
B	aspartate or asparagine
C	cystine
D	aspartate
E	glutamate
F	phenylalanine
G	glycine
H	histidine
I	isoleucine
K	lysine
L	leucine
M	methionine
N	asparagine
P	proline
Q	glutamine
R	arginine
S	serine
T	threonine
U	selenocysteine
V	valine
W	tryptophan
Y	tyrosine
Z	glutamate or glutamine
X	any
*	translation stop
-	gap of indeterminate length

APPENDIX D

THE BLAST FAMILY

The BLAST family of programs allows all combinations of DNA or protein query sequences with searches against DNA or protein databases:

Blastp: compares an amino acid query sequence against a protein sequence database.

Blastn: compares a nucleotide query sequence against a nucleotide sequence database.

Blastx: compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

Tblastn: compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).

Tblastx: compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

Some of the most commonly used blastall options are:

```
blastall   arguments:

  -p  Program Name [String]

       Input should be one of "blastp", "blastn", "blastx", "tblastn", or "tblastx".

  -d  Database [String]
   default = nr
-i  Query File [File In]
   default = stdin

       The query should be in FASTA format.  If multiple FASTA entries are in the input
       file, all queries will be searched.

  -e  Expectation value (E) [Real]
   default = 10.0

  -o  BLAST report Output File [File Out]  Optional
   default = stdout

  -F  Filter query sequence (DUST with blastn, SEG with others) [T/F]
   default = T
```

REFERENCE

1. S. F. Altschul, "Gap costs for multiple sequence alignment," *Journal of Theoretical Biology*, Vol. 138, no. 3, pp. 297-309, 1989.

2. S. F. Altschul, M. S. Boguski, W. Gish and J. C. Wootton, "Issues in searching molecular sequence databases," *Nature genetics*, Vol. 6, pp. 119-130, 1994.

3. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, Vol. 215, pp. 403-410, 1990.

4. I. Belyi and P. A. Pevzner, "Software for DNA sequencing by hybridization," *Computer Applications in the Biosciences*, Vol. 13, no. 2, pp. 205-210, 1997.

5. G. Benson and M. S. Waterman, "A Method for fast database search for all *K*-nucleotide repeats," *Nucleic Acids Research*, Vol. 22, no. 22, pp. 4828-4836, 1994.

6. J. Blazewicz, J. Kaczmarek, M. Kasprzak, W. T. Markiewicz and J. Weglarz, "Sequential and parallel algorithms for DNA sequencing," *Computer Applications in the Biosciences*, Vol. 13, no. 2, pp. 151-158, 1997.

7. A. Bolshoy, I. Ioshikhes and E. Trifonov, "Applicability of multiple alignment algorithm for detection of weak patterns: periodically distributed DNA pattern as a study case," *Computer Applications in the Biosciences*, Vol. 12, no. 5, pp. 383-389, 1996.

8. S. H. Bryant and S. F. Altschul, "Statistics of sequence-structure threading," *Current Opinion in Structural Biology*, Vol. 5, no. 2, pp. 236-244, 1995.

9. C. Caporale, C. Sepe, C. Caruso, P. Petrilli and V. Buonocore, "An algorithm to determine protein sequence alignment by utilizing data obtained from a peptide mixture and individual peptides," *Computer Applications in the Biosciences*, Vol. 10, no. 5, pp. 489-494, 1994.

10. G. J. S. Chang, J. T. L. Wang, G. W. Chirn, and C. Y. Chang, "A visualization tool for pattern matching and discovery in scientific database," in *Proceeding of the English International Conference on Software Engineering and Knowledge Engineering*, Lake Tahoe, NV, June 1996.

11. K.-M. Chao, J. Zhang, J. Ostell and W. Miller, "A local alignment tool for very long DNA sequences," *Computer Applications in the Biosciences*, Vol. 11, no. 2, pp. 147-153, 1995.

12. K.-M. Chao, J. Zhang, J. Ostell and W. Miller, "A tool for aligning very similar DNA sequences," *Computer Applications in the Bioscience*, Vol. 13, no. 1, pp. 75-80, 1997.

13. Q. K. Chen, G. Z. Hertz and G. D. Stormo, "MATRIX SEARCH 1.0: A computer program that scans DNA sequences for transcriptional elements using a database of weight matrices," *Computer Applications in the Biosciences*, Vol. 11, no. 5, pp. 563-566, 1995.

14. J. L. Cook, R. N. Re, J. F. Giardina, F. E. Fontenot, D. Y. Cheng, and J. Alam, "Distance constraints and stereospecific alignment requirements characteristic of p53 DNA-binding consensus sequence homologies," *Oncogene*, Vol. 11, no. 4, pp. 723-733, 1995.

15. L. L. Dahm, "Using the DNA profile as the unique patient identifier in the Community Health Information Network: Legal Implications," *The John Marshall Journal of Computer Information*, Vol. 15, no. 2, pp. 227-275, 1997.

16. K. Frech, K. Quandt and T. Werner, "Sortware for the analysis of DNA sequence elements of transcription," *Computer Applications in the Biosciences*, Vol. 13, no. 1, pp. 89-97, 1997.

17. O. Gotoh, "Further improvement in methods of group-to-group sequence alignment with generalized profile operations," *Computer Applications in the Biosciences*, Vol.10, no.4, pp.379-387, 1994.

18. O. Gotoh, "Optimal alignment between groups of sequences and its application to multiple sequence alignment," *Computer Applications in the Biosciences*, Vol. 9, no. 3, pp. 361-370, 1993.

19. J. A. Grice, R. Hughey and D. Speck, "Reduced space sequence alignment," *Computer Applications in the Biosciences*, Vol. 13, no. 1, pp. 45-53, 1997.

20. X. Guan and E. C. Uberbacher, "Alignments of DNA and protein sequences containing frameshift errors," *Computer Applications in the Biosciences*, Vol. 12, no. 1, pp. 31-40, 1996.

21. J. Hein, "An algorithm combining DNA and protein alignment," *Journal of Theoretical Biology*, Vol. 167, pp.169-174, 1994.

22. D. G. Higgins, A. J. Bleasby and Rainer Fuchs, "Clustal V: Improved software for multiple sequence alignment," *Computer Applications in the Biosciences*, Vol. 8, no. 2, pp. 189-191, 1992.

23. X. Huang and J. Zhang, "Methods for comparing a DNA sequence with a protein sequence," *Computer Applications in the Bioscience*, Vol. 12, no. 6, pp. 497-506, 1996.

24. X. Huang, “An algorithm for identifying regions of a DNA sequence that satisfy a content requirement,” *Computer Applications in the Biosciences*, Vol. 10, no. 3, pp. 219-225, 1994.

25. X. Huang, “On global sequence alignment,” *Computer Applications in the Biosciences*, Vol. 10, no. 3. pp. 227-235, 1994.

26. I. Ioshikhes, A. Bolshoy, E. N. Trifonov, “Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences,” *Journal of Molecular Biology*, Vol. 262, pp. 129-139, 1996.

27. M. S. Johnson and J. P. Overington, “A Structural Basis for Sequence Comparisons: An evaluation of scoring methodologies, “ *Journal of Molecular Biology*, Vol. 233, pp. 716-738, 1993.

28. D. J. Lipman and W. R. Pearson, “Rapid and sensitive protein similarity searches,” *Science*, Vol. 227, pp. 1435-1441, 1985.

29. J. T. Millard, “DNA modifying agents as tools for studying chromatin structure,” *Biochimie*, Vol. 78, no. 10, pp. 803-816, 1996.

30. S. Miyazawa, “A reliable sequence alignment method based on probabilities of residue correspondences,” *Protein Engineering*. Vol. 8, no. 10, pp. 999-1009, 1995.

31. M. D. Moody, “DNA analysis in forensic science,” *BioScience*, Vol. 39, no. 1, pp. 31-36, 1989.

32. J. D. Parsons, “Improved tools for DNA comparison and clustering,” *Computer Applications in the Biosciences*, Vol. 11, no. 6, pp. 603-613, 1995.

33. W. R. Pearson and D. J. Lipman, “Improved tools for the biological sequence comparison,” *Proc. Natl. Acad. Sci.* USA., Vol. 85, pp. 2444-2448, 1988.

34. W. R. Pearson, “Empirical statistical estimates for sequence similarity searches,” *Journal of Molecular Biology*, Vol. 276, pp. 71-84, 1998.

35. W. R. Pearson, “Identifying distantly related protein Sequences,” *Computer Applications in the Biosciences*, Vol. 13, no. 4, pp. 325-332, 1997.

36. F. E. Penotti, “A distributed system for DNA/protein database similarity searches, “ *Computer Applications in the Biosciences*, Vol. 10, no. 3, pp. 277-280, 1994.

37. D. S. Prestridge, “Signal Scan 4.0: Additional databases and sequence formats,” *Computer Applications in the Bioscience*, Vol. 12, no. 2, pp. 157-160, 1996.

38. D. S. Prestridge, "SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements," *Computer Applications in the Biosciences*, Vol. 7, no. 2, pp. 203-206, 1991.

39. N. Prunella, S. Liuni, M. Attimonelli and G. Pesole, "FASTPAT: A fast and efficient algorithm for string searching in DNA sequences," *Computer Applications in the Biosciences*, Vol. 9, no. 5, pp. 541-545, 1993.

40. E. Rivals, O. Delgrange, J.-P. Delahaye, M. Dauchet, M.-O. Delorme, A. Henaut and E. Ollivier, "Detection of significant patterns by compression algorithms: The case of approximate tandem repeats in DNA sequences," *Computer Applications in the Biosciences*, Vol. 13, no. 2, pp. 131-136, 1997.

41. J. Schafer and M. Schoniger, "Distree: A tool for estimating genetic distances between aligned DNA sequences," *Computer Applications in the Biosciences*, Vol. 13, no. 4, pp. 445-451, 1997.

42. M. Schoniger and A. von Haeseler, "Simulating efficiently the evolution of DNA sequences," *Computer Applications in the Biosciences*, Vol. 11, no. 1, pp. 111-115, 1995.

43. M. Schoniger and M. S. Waterman, "A local algorithm for DNA sequence alignment with inversions," *Bulletin of Mathematical Biology*, Vol.54, no.4, pp.521-536, 1992.

44. G. D. Schuler, S.F. Altschul and D. J. Lipman, "A workbench for multiple alignment construction and analysis," *Proteins: Structure, Function, and Genetics*, Vol. 9, no. 3, pp. 180-190, 1991.

45. G. B. Singh and S. A. Krawetz, "DNA View: A quality assessment tool for the visualization of large sequenced regions," *Computer Applications in the Biosciences*, Vol.11, no. 3, pp.317-319, 1995.

46. G. D. Smith and K. E. Bernstein, "BULLET: A computer simulation of shotgun DNA sequencing," *Computer Applications in the Biosciences*, Vol. 11, no. 2, pp. 155-157, 1995.

47. T. M. Smith, C. Abajian and l Hood, "Hopper: Software for automating data tracking and flow in DNA sequencing," *Computer Applications in the Biosciences*, Vol. 13, no. 2, pp. 175-182, 1997.

48. W. R. Taylor, "Multiple sequence threading: An analysis of alignment quality and stability," *Journal of Molecular Biology*, Vol. 2695, pp.902-943, 1997.

49. J. T. L. Wang, T. G. Marr, D. Shasha, B. Shapiro, and G. W. Chirn, "Discovering active motifs in sets of related protein sequences and using them for classification," *Nucleic Acids Research*, Vol. 22, no. 14, pp. 2769-2775, 1994.

50. M. S. Waterman, "Parametric and ensemble sequence alignment algorithms," *Bulletin of Mathematical Biology*, Vol.56, no. 4, pp. 743-767, 1994.

51. S. Widgren and C. Elvingson, "Computer simulation of DNA gel electrophoresis: influence of solid friction on linear and circular chains," *Macromolecular Theory and Simulation*, Vol. 5, no. 6, pp. 1019-1030, 1996.

52. C. Zhang and A.K.C. Wong, "A genetic algorithm for multiple molecular sequence alignment," *Computer Applications in the Bioscience*, Vol. 13, no. 6, pp. 565-581, 1997.

53. C. Zhang and A. K. C. Wong, "Toward efficient multiple molecular sequence alignment: A system of genetic algorithm and dynamic programming," *IEEE Transactions on Systems, Man and Cyberetics*, Vol. 27, no. 6, pp. 918-932. 1997.

54. P. Zhang, E. A. Schon, S. G. Fischer, E. Cayanis, J. Weiss, S. Kistler and P. E. Bourne, "An algorithm based on graph theory for the assembly of contings in physical mapping of DNA," *Computer Applications in the Biosciences*, Vol. 10, no. 3, pp. 309-317, 1994.