Introduction to the UniProt Database

www.ebi.ac.uk/services

The UniProt Consortium is comprised of the European Bioinformatics Institute, the Swiss Institute of Bioinformatics and the Protein Information Resource. The UniProt consortium aims to support biological research by maintaining a high quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community.

UniProt is composed of three components: the UniProt Knowledgebase (UniProtKB) is the central access point for extensive curated protein annotation, the UniProt Reference Cluster (UniRef) combines closely related sequences for speed searching, and the UniProt Archive (UniParc) is a comprehensive repository that reflects the history of all protein sequences.

This tutorial uses the UniProt Knowledgebase, which is divided into two parts: Swiss-Prot and TrEMBL. Swiss-Prot is a protein sequence database that contains core data enhanced by manually curated annotation. The core data consists of the sequence data, citation information, and taxonomic data. The annotation describes features such as function, post-translational modifications, domains and sites, secondary and quaternary structure, diseases, and sequence variants. In addition, Swiss-Prot merges separate entries covering the same sequence. The second section of UniProtKB is TrEMBL, which is the computer-annotated and contains translations of all coding regions in the DDBJ/EMBL/GenBank databases, as well as protein sequences extracted from literature or submitted to directly to UniProtKB (prior to integration into Swiss-Prot). TrEMBL allows these sequences to be made publicly available quickly, without diluting the high quality annotation found in Swiss-Prot. The quality of the TrEMBL data is dependent upon the in formation provided by the submitter of the nucleotide data. The information in TrEMBL is then enhanced through redundancy removal, evidence tags, and the use of automatic annotation procedures (transfers data from well-defined Swiss-Prot entries using InterPro domain/family predictions).

All data stored in UniProt can be downloaded from the Download Centre at http://www.ebi.uniprot.org/database/download.shtml.

Searching UniProt

UniProt can be searched in a number of different ways. The Text Search allows the database to be searched using keywords, similar to how one searches *Google*. The Power Search is also a keyword search engine, but it allows longer queries to be build using logical operators such as “and” and “butnot”, which can narrow down the results to a closely focused group. Power Search queries can be bookmarked and executed at a later time and on different data sets. Using SRS, UniProt can be searched as a database linked to a wide range of other databases. The results from several queries can be connected and analysed through the “Data Set Manager”, which allows multiple data sets to be manipulated by combining or subtracting specific portions of the results.

Alignments

UniProt incorporates several sequence search facilities, which can be queried using either protein or nucleotide sequences. A brief summary of each is given below.

BLAST (Basic Local Alignment Search Tool) is used to compare a novel sequence with those contained in a database. Through aligning the novel sequence with characterised ones, BLAST is able to find local regions of similarity that can give clues about the structure and function of the novel sequence, and about its evolutionary homology with other sequences. BLAST works by first looking for similar segments (HSSPs, or High-scoring Segment Pairs) between the query sequence and a database sequence, then evaluating the significance of such a match using defined thresholds. The advantage of using a BLAST search is speed, being one of the fastest methods, but there is some loss of sensitivity when compared to other methods. There are two versions of BLAST at the EBI, WU-Blast2 and NCBI-Blast2, which make use of distinctly different software packages that obtain their results in different ways, and which offer different features. EBI also has Blast2 EVEC, which provides a means of checking for vector contamination. (A tutorial on BLAST can be found at http://www.ebi.ac.uk/2can/tutorials/protein/blast.html).

FASTA (Fast-All) is a fast protein or nucleotide comparison tool that achieves a high level of sensitivity for similarity searching at high speed. This is achieved by identifying potential matches before attempting the more time consuming optimised searches for local alignments using a substitution matrix. FASTA is much more sensitive than BLAST, but does take more time to carry out. FASTA can be very specific when identifying long regions of low similarity, especially for highly divergent sequences. (A tutorial on the use of FASTA can be found at http://www.ebi.ac.uk/2can/tutorials/protein/fasta.html).

MPsrch is a sequence-sequence comparison tool that uses the true Smith and Waterman algorithm. MPsrch allows a rigorous search in a reasonable computational time, using an exhaustive algorithm that is recognised as the most sensitive sequence comparison available. Consequently, MPsrch is capable of identifying hits in cases where BLAST and FASTA fail, in addition to reporting fewer false-positive hits. (A tutorial on the use of MPsrch can be found at http://www.ebi.ac.uk/2can/tutorials/protein/MPsrch.html).

For this exercise, we will use MPsrch.

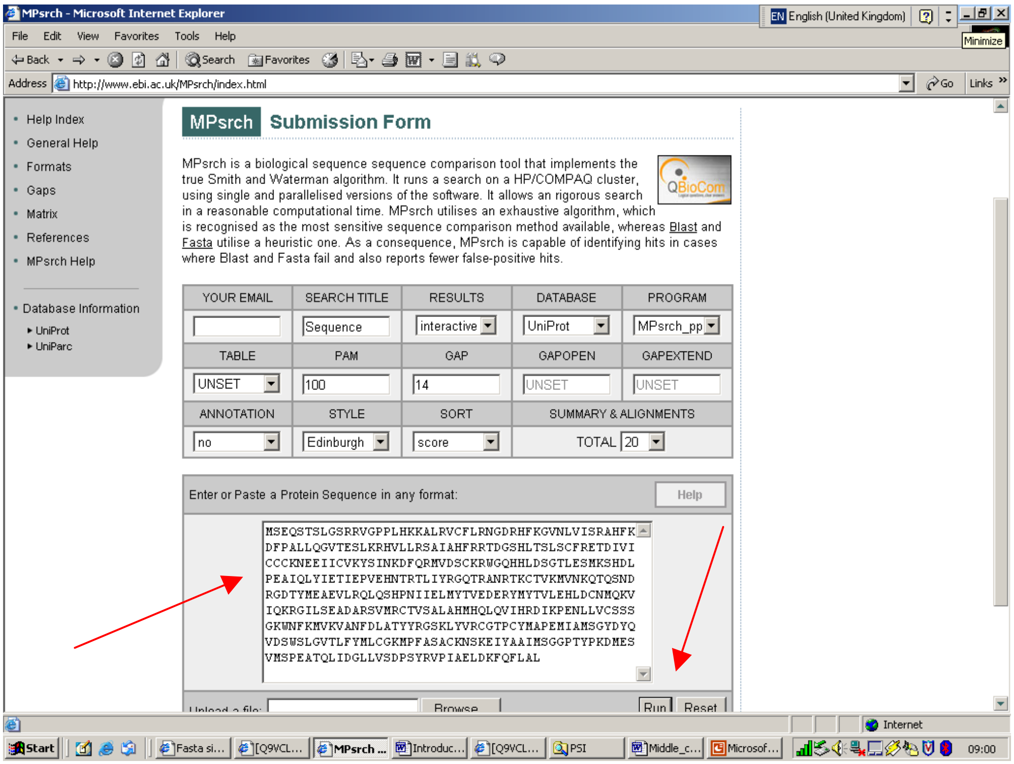
ALL THE TOOLS AND DATABASES USED IN THIS TUTORIAL CAN BE ACCESSED VIA THE PAGE: **www.ebi.ac.uk/services**

The Starting Point – a sequence (peptide or protein)

Protein A:

```
 MSEQSTSLGSRRVGPPLHKKALRVCFLRNGDRHFKGVNLVISRAHFKDFPALLQGVTESLKRHVLL
 RSAIAHFRRTDGSHLTSLSCFRETDIVICCCKNEEIICVKYSINKDFQRMVDSCKRWGQHHLDSGT
 LESMKSHDLPEAIQLYIETIEPVEHNTRTLIYRGQTRANRTKCTVKMVNKQTQSNDRGDTYMEAEV
 LRQLQSHPNIIELMYTVEDERYMYTVLEHLDCNMQKVIQKRGILSEADARSVMRCTVSALAHMHQL
 QVIHRDIKPENLLVCSSSGKWNFKMVKVANFDLATYYRGSKLYVRCGTPCYMAPEMIAMSGYDYQV
 DSWSLGVTLFYMLCGKMPFASACKNSKEIYAAIMSGGPTYPKDMESVMSPEATQLIDGLLVSDPSY
 RVPIAELDKFQFLAL
```

- **From the EBI web services page, follow the link to the *MPsrch* page under "Toolbox". Use MPsrch to find the proteins with the closest sequence identity to your search sequence.**
- **On the MPsrch page:**
  - **Paste sequence into MPsrch**
  - **Add e-mail address if you want to keep the results to look at later**
  - **Check "Database" is set at "UniProt"**
  - **Check "Program" is set at "MPsrch_pp"**
  - **Check "Results" is set at "Interactive"**
  - **Check "Summary and Alignments Total" is set at "50"**
  - **Press "Run".**


MPsrch - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ebi.ac.uk/MPsrch/index.html
Help Index
General Help
Formats
Gaps
Matrix
References
MPsrch Help
Database Information
UniProt
UniParc

MPsrch Submission Form

MPsrch is a biological sequence sequence comparison tool that implements the true Smith and Waterman algorithm. It runs a search on a HP/COMPAQ cluster, using single and parallelised versions of the software. It allows an rigorous search in a reasonable computational time. MPsrch utilises an exhaustive algorithm, which is recognised as the most sensitive sequence comparison method available, whereas Blast and Fasta utilise a heuristic one. As a consequence, MPsrch is capable of identifying hits in cases where Blast and Fasta fail and also reports fewer false-positive hits.

YOUR EMAIL	SEARCH TITLE	RESULTS	DATABASE	PROGRAM
	Sequence	interactive	UniProt	MPsrch_pp
TABLE	PAM	GAP	GAPOPEN	GAPEXTEND
UNSET	100	14	UNSET	UNSET
ANNOTATION	STYLE	SORT	SUMMARY & ALIGNMENTS	
no	Edinburgh	score	TOTAL 20	

Enter or Paste a Protein Sequence in any format:

```
MSEQSTSLGSRRVGPPLHKKALRVCFLRNGDRHFKGVNLVISRAHFK
DFPALLQGVTESLKRHVLLRSAIAHFRRTDGSHLTSLSCFRETDIVI
CCCKNEEIICVKYSINKDFQRMVDSCKRWGQHHLDSGTLESMKSHDL
PEAIQLYIETIEPVEHNTRTLIYRGQTRANRTKCTVKMVNKQTQSND
RGDTYMEAEVLRQLQSHPNIIELMYTVEDERYMYTVLEHLDCNMQKV
IQKRGILSEADARSVMRCTVSALAHMHQLQVIHRDIKPENLLVCSSS
GKWNFKMVKVANFDLATYYRGSKLYVRCGTPCYMAPEMIAMSGYDYQ
VDSWSLGVTLFYMLCGKMPFASACKNSKEIYAAIMSGGPTYPKDMES
VMSPEATQLIDGLLVSDPSYRVPIAELDKFQFLAL
```

Looking at the results from MPsrch

The results page will provide a list of proteins that match the query sequence from the UniProt database ordered by their scores, along with their description, length and scoring information. The following information is provided:

ID = brief identifier of database sequence entry.
DESCRIPTION = first 21 characters of database sequence entry.
LENGTH = length of database sequence entry in alignment.
QUERY MATCH % = percentage of the query sequence that is matched.
SCORE = alignment score representing likelihood that alignment is not random.
PRED NO. = predicted number of unrelated (random) sequences likely to show a score greater or equal to alignment score.

The list of scores for matches to the above sequence should suggest a 100% match to Q9VCL7_DROME, a *Drosophila melanogaster* protein from the TrEMBL section of UniProt. In addition, the query sequence has 38% identity to Q8N568 (DCAK2_HUMAN), a well-annotated human protein sequence from the Swiss-Prot section of UniProt. In this tutorial, we will explore the Q9VCL7_DROME protein, using DCAK2_HUMAN for comparison.

Individual pairwise alignments of the query sequence with each sequence identified by an MPsrch match can be viewed from the results page, with the ability to view only certain sequences. In each pairwise alignment, "Qy" identifies the query sequence, and "Db" identifies the database-matched sequence. Matching residues are identified by "*" above the alignment, while similar amino acids (e.g. leucine vs. methionine) are identified by ".", and mismatches are shown by a space.

- **On the MPsrch results page, make sure only the Q9VCL7_DROME and DCAK2_HUMAN sequences are checked (all are checked by default), and then click on "Show Alignments" to display aligned sequences. The alignments should show Q9VCL7_DROME as a perfect match, whereas DCAK2_HUMAN only has regions of conservation.**
- **From the alignments page, click on the hyperlinked Q9VCL7_DROME to open the UniProt entry.**

This is a UniProt/TrEMBL entry that is translated from the nucleotide sequence, and which has only automatic annotation and additional cross-referencing added. In the Q9VCL7_DROME entry, the gene name is "Orf name", which will probably change when protein is characterised.

- **Scroll down to the "Origin of the protein" section.**

There are different links provided under the taxonomic fields. In the "From" section, the species name, "*Drosophila melanogaster*", is linked to a search for all proteins in UniProtKB to this organism and will provide an extensive search results page listing all these proteins. Similarly, under the "Taxonomy" section, each taxonomic identifier will link to a UniProt search for all the proteins under that level of taxonomic classification (for e.g. "Eukaryota" will search for all eukaryotic proteins in UniProt). Note that at the lowest level is the genus name, "*Drosophila*", which will search for all *Drosophila* proteins, not only those of the species *Drosophila melanogaster*.

To gain more information on the organism itself, UniProt provides a link to the NEWT database, which is a compilation of the information in the NCBI taxonomic database together with proteins found in the Swiss-Prot and TrEMBL section of UniProt. The link to this information is contained in the TaxID number that follows the species name, here "7227". The TaxID itself is provided by the NCBI.

- **Click on the taxonomy link "7227" under the TaxID to open the relevant page in NEWT.**

For each species, NEWT displays the following taxonomy data: Swiss-Prot scientific name, Swiss-Prot common name and Swiss-Prot synonym, lineage, strain information, and the number of protein sequence entries in Swiss-Prot and TrEMBL. There are also some interesting links to external sources concerning the organism. The NEWT data is available from the European Bioinformatics Institute and is maintained by the Swiss-Prot group at the Swiss Institute of Bioinformatics.

- ***Go back to the Q9VCL7_DROME entry page.***
- ***Scroll down to the "Comments" section.***

This section provides a link to the interaction data in IntAct, the Molecular Interaction Database that provides analysis tools for protein interaction data. All interactions are derived from literature curation, or direct user submissions, and are freely available.

- ***Scroll down to the "Cross-references" section.***

There are over sixty different data collections available through UniProt via cross-reference links, a few of which can be found in this entry.

The EMBL line contains the original nucleotide submission data, which is stored as identical underlying information in EMBL, GenBank and DDBJ. Links to the relevant nucleotide data is provided to all three databases, as they display slightly different views.

HSSP (Homology-derived Secondary Structure of Proteins database) provides a link to the most similar UniProt entry that has a 3D structure for proteins lacking a PDB structure entry. A link is also provided to the PDB entry of that structure.

- ***In the "HSSP" line, click on the link to the closest UniProt entry that has a 3D structure, namely "O15075".***

? What is the name of this protein?

- ***Go back to the Q9VCL7_DROME entry page.***
- ***Scroll down to the "FlyBase" line, and click on the hyperlink.***

The FlyBase line gives a more gene-centric view of the entry. FlyBase is a comprehensive database for information on the genetics and molecular biology of *Drosophila*, and includes data from the Drosophila Genome Project and data curated from the literature. FlyBase often contains additional literature references that may be of use.

- ***Go back to the Q9VCL7_DROME entry page.***

The GO (Gene Ontology) line provides mapping to biological process, molecular function, and cellular component.

- ***In the "GO" section, click on the link to "QuickGO".***

? What GO terms have been associated with this protein?

- ***Go back to the Q9VCL7_DROME entry page.***

The InterPro line provides direct links to the InterPro entries that have signatures matching this protein. From this information, we can begin to understand the predicted domain organisation of the protein, as well as its family relationships with other proteins. There is also a link to a graphical view of the protein that displays the relative positions of the InterPro signature matches.

The PFAM, SMART and PROSITE lines provides direct links to the PFAM and SMART databases of Hidden Markov Models, and the PROSITE database of patterns and profiles that match this protein. There are also links to graphical views of the predicted domain structure for these databases.

➢ **Scroll down to the "Keyword" section.**

The Keywords provide information that can be used to generate indexes of sequences based on functional, structural or other categories. The keywords chosen for an entry serve as a subject reference for the sequence. TrEMBL entries make use of the same controlled list of keywords as Swiss-Prot entries, but as these words are added during the manual curation process, TrEMBL entries tend to have fewer keywords than Swiss-Prot entries. Each keyword is linked to its definition, as well as its corresponding GO term.

➢ **Click on the keyword "Differentiation".**

? What is the corresponding GO term for this keyword?

➢ **Go back to the Q9VCL7_DROME entry page.**
➢ **Scroll down to the "Sequence information" section.**

The Sequence information provided includes the length of the protein sequence in amino acid residues, it molecular weight and the sequence itself. The molecular weight is not meant to be the mature, post-translationally modified protein, but the weight prior to any processing. There is also a CRC (Cyclic Redundancy Check) number that is a checksum on the sequence. The sequence itself can be obtained as a Fasta view, which permits an easy cut/paste of a sequence into sequence analysis forms. You can also get Fasta-formatted sequence through the link at the top of the entry page under "viewers".

➢ **Click on the "Pop-Up Fasta View" at the bottom left of the page. . Close the pop-up window.**
➢ **Scroll up to the top of the Q9VCL7_DROME entry page and click on "Extended" view. This will add additional information, such as additional cross-references (purple), and automatic annotation (green).**

? What additional information has this added to Q9VCL7_DROME?
? What information can we gain about our query protein from the UniProt TrEMBL entry of Q9VCL7_DROME?

➢ **Compare our TrEMBL entry to DCAK2_HUMAN (38% similarity by alignment).**
➢ **Open a second window, and in the "Text Search" box at the top of the UniProt page, type: DCAK2_HUMAN**

? What is the full name of this protein?

This protein contains a merged entry, Q8N399, as shown under the "Secondary accession numbers" section.

The "Protein description" section is more complete for this entry than what was in TrEMBL. The synonyms given for the protein are very useful when carrying out literature searches, and currently certain literature search engines will pull the synonym information out of UniProt in order to perform a more complete literature search. This section also contains an enzyme identifier code ("EC 2.7.1.37), as well as a link to the relevant page in the ENZYME database.

- **Follow the link to the ENZYME database entry for this protein.**

? What other alternative names are given to this type of enzyme? (Note that they are more general names).
? What is the reaction that these enzymes catalyse?

The ENZYME database lists all the Swiss-Prot entries that are known to have this ENZYME classification. Unfortunately, our Q9VCL7_DROME would not be amongst them, as it is a TrEMBL entry.

- **Go back to the DCAK2_HUMAN entry page.**
- **Scroll down to the "Comments" section.**

Notice that there are more entries under the Comments section for a Swiss-Prot entry, as these are added during the manual annotation process. There are many different possible comment fields, including allergen, alternative products, biophysicochemical properties, biotechnology, catalytic activity, caution (possible errors), cofactor, developmental stage, disease, domain, enzyme regulation, function, induction, interaction, mass spectrometry, pathway, pharmaceutical, polymorphism, post-translational modification, RNA editing, subcellular location, subunit, tissue specificity and toxic dose. The DCAK2_HUMAN entry comments provide information on its catalytic activity and its similarity to other domains or protein families.

? What domain and family information can you gain about this protein through its similarity to other proteins?

The Comments section also contains information about alternatively spliced products for this protein. The isoform that is used in the entry is marked.

? How many isoforms is this protein known to have?

- **Click on the "Display all isoforms sequences in Fasta format" link.**

Using the Fasta formatted page, the alternative splice products can be analysed in a variety of ways via the drop-down menu at the top of the page, including multiple alignments and finding conserved patterns.

- **On the Fasta formatted page, set the drop-down menu to "Clustal W (multiple alignment)", press "Select all", and then press "Submit Query".**
- **On the Clustal W query page, press "Run Clustal W".**
- **On the Clustal W query receipt page, press the "clustalW (aln)" link.**

? Is there much variation between the splice variants?

- ➢ **Go back to the DCAK2_HUMAN entry page.**
- ➢ **Scroll down to the "Keywords" section.**

? Comparing the keywords for the two entries DCAK2_HUMAN and Q9VCL7_DROME, can you find any differences?

- ➢ **On the DCAK2_HUMAN entry page, scroll down to the "Features" table.**

The Features table provides a precise but simple means for the annotation of sequence data. The table describes regions or sites of interest in the sequences, such as sites of post-translational modifications, binding sites, enzyme active sites, local secondary structure, as well as several other characteristics. The feature table also lists sequence conflicts between references.

? Using the Feature Table, how many domains does DCAK2_HUMAN have, and what are their residue positions?
? Which residue(s) are thought to be part of the active site of this enzyme?
? What molecule is this enzyme thought to bind (in addition to substrate), and which residues appear to be involved?
? What additional information can you find out about the different isoforms?

- ➢ **Click on the "FeatureTableViewer" link at the foot of the Feature Table.**

For each feature listed, there is a position that is linked to the entire sequence.

- ➢ **Click on the "Position" link for the first domain, namely "72-158.**

Now the entire protein sequence is shown in both one-and three-letter codes, with the relevant domain sequence highlighted in red.

- ➢ **Go back to the previous page containing the FeatureTableViewer.**
- ➢ **Scroll down to the bottom of the page to the SEView display.**

The features are displayed graphically in SEView as bars along the sequence. By clicking on a feature, its description will appear at the top of the diagram.

- ➢ **Click on the red dot on the black sequence line.**

? What is this feature and its position in the protein?
? What feature is the blue dot on the display?

- ➢ **Go back to the DCAK2_HUMAN entry page.**
- ➢ **Click on the "Feature aligner" link at the foot of the Feature Table.**

This shows the actual sequence associated with each domain, its total length in amino acids, and its position in the protein sequence. Alternatively, the position of the domain can be seen (highlighted in red) in relation to the entire protein sequence through the "Position" hyperlinks.

➢ **Go back to the DCAK2_HUMAN entry page.**

? Could annotation be transferred from human (DCAK2_HUMAN) to fly (Q9VCL7_DROME)?

Remember, the same protein in a different genetic background can function differently because of the different protein-protein interactions possible. Swiss-Prot annotators always keep this in mind when transferring any annotation.

Return to www.ebi.ac.uk/services.

This is the end of the short tour of the UniProt database. Perhaps you might like to try it again with a more relevant sequence, such as one you are currently working with. Remember that all this information is at your disposal and much of the data can be downloaded and installed in-house.

Now try and repeat some or all of these searches on the following sequences:

Protein X:

```
MPYLLPGFFCDRVIRERDRRNGEGTVSQPLKFEGQDFVVLKQRCLAQKCLFEDRVFPAGTQALGS
HELSQKAKMKAITWKRPKEICENPRFIIGGANRTDICQGDLGDCWFLAAIACLTLNERLLFRVIP
HDQSFTENYAGIFHFQFWRYGDWVDVVIDDCLPTYNNQLVFTKSNHRNEFWSALLEKAYAKLHGS
YEALKGGNTTEAMEDFTGGVTEFFEIKDAPSDMYKIMRKAIERGSLMGCSIDTIVPVQYETRMAC
GLVKGHAYSVTGLEEALFKGEKVKLVRLRNPWGQVEWNGSWSDGWKDWSFVDKDEKARLQHQVTE
DGEFWMSYDDFVYHFTKLEICNLTADALESDKLQTWTVSVNEGRWVRGCSAGGCRNFPDTFWTNP
QYRLKLLEEDDDPDDSEVICSFLVALMQKNRRKDRKLGANLFTIGFAIYEVPKEMHGNKQHLQKD
FFLYNASKARSKTYINMREVSQRFRLPPSEYVIVPSTYEPHQEGEFILRVFSEKRNLSEEAENTI
SVDRPVPRPGHTDQESEEQQQFRNIFRQIAGDDMEICADELKNVLNTVVNKHKDLKTQGFTLESC
RSMIALMDTDGSGRLNLQEFHHLWKKIKAWQKIFKHYDTDHSGTINSYEMRNAVNDAGFHLNSQL
YDIITMRYADKHMNIDFDSFICCFVRLEGMFRAFHAFDKDGDGIIKLNVLEWLQLTMYA
```

Protein Y:

```
QLEEEVKDLADKKESVAHWEAQITEIIQWVSDEKDARGYLQALASKMTEELEALRNSSLGTR
ATDMPWKMRRFAKLDMSARLELQSALDAEIRAKQAIQEELNKVKASNIITECKLKDSEKKNL
ELLSEIEQLIKDTEELRSEKGIEHQDSQHSFLAFLNTPTDALDQFETVDSTPLSVHTPTLRK
KGCPGSTGFPPKRKTHQFFVKSFTTPTKCHQCTSLMVGLIRQGCSCEVCGFSCHITCVNKAP
TTCPVPPEQTKGPLGIDPQKGIGTAYEGHVRIPKPAGVKKGWQRALAIVCDFKLFLYDIAEG
KASQPSVVISQVIDMRDEEFSVSSVLASDVIHASRKDIPCIFRVTASQLSASNNKCSILMLA
DTENEKNKWVGVLSELHKILKKNKFRDRSVYVPKEAYDSTLPLIKTTQAAAIIDHERIALGN
EEGLFVVHVTKDEIIRVGDNKKIHQIELIPNDQLVAVISGRNRHVRLFPMSALDGRETDFYK
LSETKGCQTVTSGKVRHGALTCLCVAMKRQVLCYELFQSKTRHRKFKEIQVPYNVQWMAIFS
EQLCVGFQSGFLRYPLNGEGNPYSMLHSNDHTLSFIAHQPMDAICAVEISSKEYLLCFNSIG
IYTDCQGRRSRQQELMWPANPSSCCYNAPYLSVYSENAVDIFDVNSMEWIQTLPLKKVRPLN
NEGSLNLLGLETIRLIYFKNKMAEGDELVVPETSDNSRKQMVRNINNKRRYSFRVPEEERMQ
QRREMLRDPEMRNKLISNPTNFNHIAHMGPGDGIQILKDLPMPGFPYPSPHHHSGLISSPIN
FEHIYHMTVNSAEKFLSPDSINPEYSPSLRSVPGTPSFMTLRNPRPQESRTVFSGSVSIPSI
TKSRPEPGRSMSASSGLSARSSAQNGSALKREFSGGSYSAKRQPMPSPSEGSLSSGGMDQGS
DAPARDFDKEDSDSPRHSTASNSSNLSSPPSPVSPRKTKSLSLESTDRGSWDP
```

Protein Z:

```
MLTDSGGGGTSFEEDLDSVAPRSAPAGASEPPPPGGVGLGIRTVRLFGEAGPASGVGSSGGGGSGS
GTGGGDAALDFKLAAAVLRTGGGGGASGSDEDEVSEVESFILDQEDLDNPVLKTTSEIFLSSTAEG
ADLRTVDPETQARLEALLEAAGIGKLSTADGKAFADPEVLRRLTSSVSCALDEAAAALTRMKAENS
HNAGQVDTRSLAEACSDGDVNAVRKLLDEGRSVNEHTEEGESLLCLACSAGYYELAQVLLAMHANV
EDRGNKGDITPLMAASSGGYLDIVKLLLLHDADVNSQSATGNTALTYACAGGFVDIVKVLLNEGAN
IEDHNENGHTPLMEAASAGHVEVARVLLDHGAGINTHSNEFKESALTLACYKGHLDMVRFLLEAGA
DQEHKTDEMHTALMEACMDGHVEVARLLLDSGAQVNMPADSFESPLTLAACGGHVELAALLIERGA
NLEEVNDEGYTPLMEAAREGHEEMVALLLAQGANINAQTEETQETALTLACCGGFSEVADFLIKAG
ADIELGCSTPLMEASQEGHLELVKYLLASGANVHATTATGDTALTYACENGHTDVADVLLQAGADL
EHESEGGRTPLMKAARAGHLCTVQFLISKGANVNRATANNDHTVVSLACAGGHLAVVELLLAHGAD
PTHRLKDGSTMLIEAAKGGHTNVVSYLLDYPNNVLSVPTTDVSQLPPPSQDQSQVPRVPTHTLAMV
VPPQEPDRTSQENSPALLGVQKGTSKQKSSSLQVADQDLLPSFHPYQPLECIVEETEGKLNELGQR
ISAIEKAQLKSLELIQGEPLNKDKIEELKKNREEQVQKKKKILKELQKVERQLQMKTQQQFTKEYL
ETKGQKDTVSLHQQCSHRGVFPEGEGDGSLPEDHFSELPQVDTILFKDNDVDDEQQSPPSAEQIDF
VPVQPLSSPQCNFSSDLGSNGTNSLELQKVSGNQQIVGQPQIAITGHDQGLLVQEPDGLMVATPAQ
TLTDTLDDLIAAVSTRVPTGSNSSSQTTECLTPESCSQTTSNVASQSMPPVYPSVDIDAHTESNHD
TALTLACAGGHEELVSVLIARDAKIEHRDKKGFTPLILAATAGHVGVVEILLDKGGDIEAQSERTK
DTPLSLACSGGRQEVVDLLLARGANKEHRNVSDYTPLSLAASGGYVNIIKILLNAGAEINSRTGSK
LGISPLMLAAMNGHVPAVKLLLDMGSDINAQIETNRNTALTLACFQGRAEVVSLLLDRKANVEHRA
KTGLTPLMEAASGGYAEVGRVLLDKGADVNAPPVPSSRDTALTIAADKGHYKFCELLIHRGAHIDV
RNKKGNTPLWLASNGGHFDVVQLLVQAGADVDAADNRKITPLMSAFRKGHVKVVQYLVKEVNQFPS
DIECMRYIATITDKELLKKCHQCVETIVKAKDQQAAEANKNASILLKELDLEKSREESRKQALAAK
REKRKEKRKKKKEEQKRKQEEDEENKPKENSELPEDEDEEENDEDVEQEVPIEPPSATTTTTIGIS
ATSATFTNVFGKKRANVVTTPSTNRKNKKNKTKETPPTAHLILPEQHMSLAQQKADKNKINGEPRG
GGAGGNSDSDNLDSTDCNSESSSGGKSQELNFVMDVNSSKYPSLLLHSQEEKTSTATSKTQTRLEG
EVTPNSLSTSYKTVSLPLSSPNIKLNLTSPKRGQKREEGWKEVVRRSKKLSVPASVVSRIMGRGGC
NITAIQDVTGAHIDVDKQKDKNGERMITIRGGTESTRYAVQLINALIQDPAKELEDLIPKNHIRTP
ASTKSIHANFSSGVGTTAASSKNAFPLGAPTLVTSQATTLSTFQPANKLNKNVPTNVRSSFPVSLP
LAYPHPHFALLAAQTMQQIRHPRLPMAQFGGTFSPSPNTWGPFPVRPVNPGNTNSSPKHNNTSRLP
NQNGTVLPSESAGLATASCPITVSSVVAASQQLCVTNTRTPSSVRKQLFACVPKTSPPATVISSVT
STCSSLPSVSSAPITSGQAPTTFLPASTSQAQLSSQKMESFSAVPPTKEKVSTQDQPMANLCTPSS
TANSCSSSASNTPGAPETHPSSSPTPTSSNTQEEAQPSSVSDLSPMSMPFASNSEPAPLTLTSPRM
VAADNQDTSNLPQLAVPAPRVSHRMQPRGSFYSMVPNATIHQDPQSIFVTNPVTLTPPQGPPAAVQ
LSSAVNIMNGSQMHINPANKSLPPTFGPATLFNHFSSLFDSSQVPANQGWGDGPLSSRVATDASFT
VQSAFLGNSVLGHLENMHPDNSKAPGFRPPSQRVSTSPVGLPSIDPSGSSPSSSSAPLASFSGIPG
TRVFLQGPAPVGTPSFNRQHFSPHPWTSASNSSTSAPPTLGQPKGVSASQDRKIPPPIGTERLARI
RQGGSVAQAPAGTSFVAPVGHSGIWSFGVNAVSEGLSGWSQSVMGNHPMHQQLSDPSTFSQHQPME
RDDSGMVAPSNIFHQPMASGFVDFSKGLPISMYGGTIIPSHPQLADVPGGPLFNGLHNPDPAWNPM
IKVIQNSTECTDAQQASLLPSVPALKGEIPSPQLTRPKKRIGRPMVASPNQRHQDHLRPKVPAGVQ
ELTHCPDTPLLPPSDSRGHNSSNSPSLQAGGAEGAGDRGRDTR
```