Introduction to the InterPro Database

www.ebi.ac.uk/services

InterPro is an integrated protein resource that provides a classification of protein sequences into families and domains, along with their annotation. InterPro combines the major signature databases, PROSITE, PRINTS, PFAM, PRODOM, SMART, TIGRFAM, PIR Superfamily, GENE3D, SUPERFAMILY, and PANTHER, as well as structural information from PDB, MSD, CATH, SCOP, MODBASE and SWISS-MODEL, into a unified database, capitalising on their individual strengths to produce a powerful integrated diagnostic tool. This tutorial is designed to allow users to get the most value out of the database, and will focus on the type and organisation of annotated data, the different query methods possible, understanding the different visualisations of the data, as well as exploring the multiple external links and cross-references available.

Searching InterPro

InterPro can be searched in a number of different ways. The simple text search facility allows queries using keywords, UniProt accession numbers, GO terms, or InterPro entry numbers. The simple InterPro SRS search enables more complex queries, providing two field queries, one from InterPro and the other from the list of protein matches. InterPro can also be queried through SRS either directly or indirectly as a database linked to other databases, with the possibility of creating different views, as well as recovering FASTA-format sequences. There is also a sequence search facility using the web-based server of InterProScan, which permits the sequence analysis and characterisation of unknown protein sequences. Nucleotide sequences, both DNA and RNA, can also be used to query InterProScan, where the sequence used in the query is translated in all six frames. Using InterProScan, InterPro takes each sequence and analyses it against one or more of the member databases using preconfigured cut-off thresholds. Following analysis, each result is returned and combined, and then the InterPro entries and sequence signatures are returned to the submitter as a graphical view with links to both InterPro and SRS. InterPro and InterProScan are accessible for interactive use over the EBI web server (www.ebi.ac.uk/interpro), and are distributed as stand-alone copies by anonymous ftp.

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

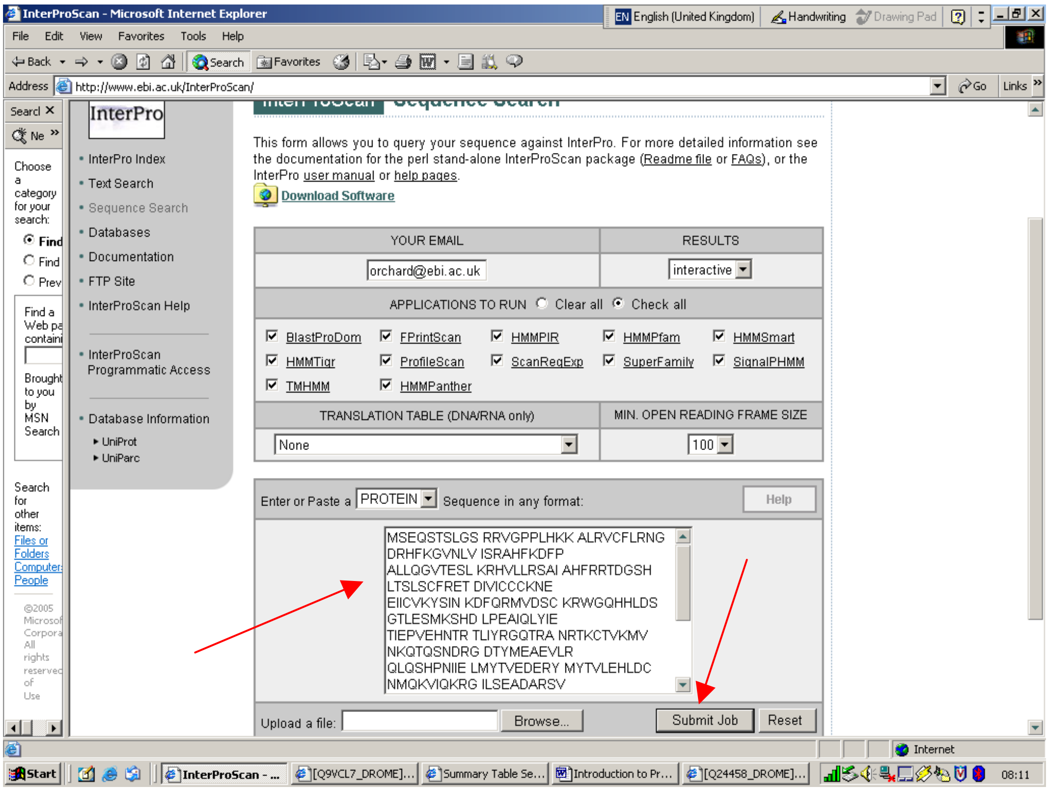
The Starting Point – a sequence (peptide or protein)

Protein A:

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

- **From the EBI web services page, follow the link to the *InterPro* page under "Protein Databases", then to the link for *InterProScan*.**

- **Use InterProScan to assign family memberships, to identify functional domains etc:**
  - **Paste sequence into InterProScan**
  - **Add e-mail address if you want to keep the results to look at later**
  - **Press "Submit Job"**


InterProScan - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.ebi.ac.uk/InterProScan/
Go Links
InterPro
InterPro Index
Text Search
Sequence Search
Databases
Documentation
FTP Site
InterProScan Help
InterProScan Programmatic Access
Database Information
UniProt
UniParc
Start InterProScan - ... [Q9VCL7_DROME]... Summary Table Se... Introduction to Pr... [Q24458_DROME]... 08:11
Internet

This form allows you to query your sequence against InterPro. For more detailed information see the documentation for the perl stand-alone InterProScan package (Readme file or FAQs), or the InterPro user manual or help pages.

Download Software

YOUR EMAIL	RESULTS
orchard@ebi.ac.uk	interactive

APPLICATIONS TO RUN ○ Clear all ● Check all

☑ BlastProDom ☑ FPrintScan ☑ HMMPIR ☑ HMMPfam ☑ HMMSmart
☑ HMMTigr ☑ ProfileScan ☑ ScanRegExp ☑ SuperFamily ☑ SignalPHMM
☑ TMHMM ☑ HMMPanther

TRANSLATION TABLE (DNA/RNA only)	MIN. OPEN READING FRAME SIZE
None	100

Enter or Paste a PROTEIN Sequence in any format: Help

```
MSEQSTSLGS RRVGPPLHKK ALRVCFLRNG
DRHFKGVNLV ISRAHFKDFP
ALLQGVTESL KRHVLLRSAI AHFRRTDGSH
LTSLSCFRET DIVICCCKNE
EIICVKYSIN KDFQRMVDSC KRWGQHHLDS
GTLESMKSHD LPEAIQLYIE
TIEPVEHNTR TLIYRGQTRA NRTKCTVKMV
NKQTQSNDRG DTYMEAEVLR
QLQSHPNIIE LMYTVEDERY MYTVLEHLDC
NMQKVIQKRG ILSEADARSV
```

Upload a file: Browse... Submit Job Reset

Looking at the results from InterProScan

OVERVIEW OF THE INTERPRO ENTRY

InterPro combines a number of databases that use different methodologies and a varying degree of biological information on well-characterised proteins to derive protein signatures. InterPro member databases are:

- PROSITE uses regular expressions and profiles
- PFAM, SMART, TIGRFAM, PIRSF, PANTHER, GENE3D and SUPERFAMILY use hidden Markov models (HMM)
- PRINTS uses fingerprints (groups of aligned, un-weighted motifs)
- PRODOM uses Clustr analysis to group sequences

Signatures describing the same protein family, domain, repeat or site, are grouped into unique InterPro entries. Each combined InterPro entry has a unique accession number and name, an abstract describing the features of the proteins associated with the entry, literature references with links to PubMed, and links to the relevant member database(s). Entries are also annotated with respect to GO terms, providing information on the process, function and component of the proteins within an entry. InterPro entries contain a variety of additional external links, including those to the structural databases PDB, EMSD, CATH and SCOP, and to the databases MEROPS, PANDIT, Blocks, IntEnz, CAZy, IUPHAR, COMe and CluSTr. The taxonomic coverage of an entry is displayed by a descriptive wheel, which permits the user to select all the sequences from specific taxonomic groups.

GRAPHICAL VIEWS

All UniProt protein sequences that have matches to a particular InterPro entry are listed in the Match Table associated with that entry, and can be viewed as graphical displays in the InterPro graphical views. The graphical views, which can be sorted by UniProt accession number, structure or taxonomy, show the position of the signatures on the proteins in an entry. Mousing over individual signatures in these views will bring up a pop-box, giving the accession, name and position of the signature. The detailed view displayed all the signatures that hit the proteins, providing a comprehensive view of each protein, with links provided to the member databases and to related InterPro entries. The structural features of proteins as described by PDB, CATH, SCOP, SWISS-MODEL and MODBASE are displayed two-dimensionally in relation to the signatures, as well as three-dimensionally using AstexViewer. The InterPro overview condenses the signatures and structural features into a simplified graphical view.

INTERPRO DOMAIN ARCHITECTURE

InterPro graphically represents the location of a protein domain and information pertaining to the origin of that domain and the proteins that contain it. Families are also defined and may contain several InterPro domains that are often, but not always, in the same order. Through the *InterPro Domain Architecture* view, the composition and order of the different domains within a family are clearly displayed for easy comparison, as well as for simple navigation between the entries for individual domains.

RELATIONSHIPS

InterPro entries are linked to one another through *PARENT/CHILD* and *CONTAINS/FOUND IN* relationships. *PARENT/CHILD* relationships indicate superfamily/family/subfamily relationships, as well as domain hierarchies, where sequences can be subdivided into more specific sub-sets. *CONTAINS/FOUND IN* relationships apply to domains, repeats and sites within families, and are used to describe the composition of protein sequences.

➢ **Look at the domain organisation of this protein.**

? How many domains does this protein have, and what are their names?

? Are the active site residues predicted?

➢ **Mouse-over the signature to bring up a pop-box on the scroll bar, giving the accession, name and position for the PF00069 signature.**

? What residues does this signature cover?

? What is the position of the signature defining the active site residues?

Notice that there are two views to choose from to view individual entries: InterPro view and SRS view.

➢ **Click on the InterPro icon for IPR000719 to see what information you can gain about this domain.**

➢ **Scroll down to the GO Mapping.**

InterPro provides mappings to three types of GO terms: (biological) process, (molecular) function, and (cellular) component.

? What GO terms can you assign to your protein on the basis of IPR000719 signature recognition?

- **Follow the link to the GO term GO:0006468.**

? What is the definition of this GO term?

- **Take a look at the hierarchical tree of GO terms from which this term was derived.**
- **Go back to the IPR000719 entry page.**
- **Choose one GO term and copy/paste the GO ID into the text search box at the top of the InterPro entry page.**

This will produce a list of all the InterPro entries associated with this GO term.

- **Go back to the IPR000719 entry page.**
- **Scroll down to the Structural Links.**

InterPro provides a list of all the PDB entries associated with an entry. There are also structural links to SCOP and CATH, which provide structural classifications of the proteins that match this entry.

- **Click on the "PDB" link to display the PDB entries of proteins that contain this domain. Close the pop-up window.**
- **Follow the "SCOP d.144.1.7" link to the SCOP database to find out the structural classification of this domain.**

? What type of structure does the protein kinase-like fold consist of?
? SCOP states that the protein kinase-like superfamily shares functional and structural similarities with which other folds?
? Which families does SCOP list as containing protein kinase-like domains?

Note that this is not an exhaustive list of families, as only those with structural information can be included.

- **Go back to the IPR000719 entry page.**
- **Scroll down to the Database links.**

InterPro provides links to several external databases, including BLOCKS multiple alignments, IntEnz enzyme information, PROSITE documentation, CAZy carbohydrate-binding enzyme information, IUPHAR receptor database, COMe bioinorganic motif information, MEROPS peptidase information, PANDIT phylogenetic trees, and MSDsite PROSITE ligand statistics.

- **Follow the "PANDIT: PF00069" link to the PANDIT database to view the phylogenetic tree of the proteins matching the PFAM signature PF00069 in this entry.**
- **Go back to the IPR000719 entry page.**
- **Scroll down to the Taxonomic coverage, which provides an at-a-glance view of the taxonomic range of the sequences associated with this InterPro entry.**

? How many Fruit Fly proteins do the signatures in this entry identify as potentially having protein kinase domains?

- **Using the taxonomic wheel, follow the link to the Fruit Fly proteins identified by this entry.**

This will produce an overview of the InterPro entries that match each of the fruit fly protein sequences. Note that there are links provided to the individual UniProt entries, to variants of each protein (where applicable), and to their GO mapping.

- **Go back to the IPR000719 entry page.**
- **Scroll up to the Relationships for this entry.**

InterPro links related signatures through special relationships, where Parent/Child relationships indicate domain/family hierarchies, and Contains/Found in relationships indicates the subdivision of domains/families into sequence regions.

? What relationships does this entry have with other InterPro entries?

- **View the PARENT/CHILD tree by following the link underneath "Child" or "Parent".**

This gives a graphical representation of the information listed on the entry page.

- **Go back to the IPR000719 entry page.**

The Parent of this entry is IPR011009 (protein kinase-like), which represents domains that have a structural fold homologous to that of protein kinases (including protein kinases themselves).

- **Follow the link to entry IPR011009.**

? What is the name of the signature that represents this entry, and from which member database does it come?

- **Go back to the IPR000719 entry page.**

? How many children are linked to the IPR000719 entry, and into what categories do they subdivide protein kinase domains?
? How do the "Contains" relationships add information to what we know from the "Children"?

- **Look at the "Found in" relationships.**

This provides information on which protein families contain the protein kinase domain signatures. In addition to these families, the IPR000719 signatures are also found in families listed under the "Found in" sections of its children.

- **Follow the links to the two children, IPR001245 and IPR002290, and compile a full list of the protein families in which IPR00719 is found.**

? How many families (InterPro entries) in total are IPR000719 signatures found in?

- **Scroll down to the Overlapping InterPro entries section of IPR000719.**

This gives a graphical representation of the number of protein matches that overlap between different entries, and the extent of the signature overlap in terms of amino acid residues.

- **In the "Overlapping InterPro Entries" section, find the data for the overlap between IPR000719 and IPR008271 (the entry for the serine/threonine active site that is sometimes contained within this domain).**

? How many proteins that contain the IPR00719 kinase domain also contain the IPR008271 serine/threonine active site?
? Are all the amino acids from this active site contained within the IPR00719 domain?

**Note that not all of the relationships determined for IPR000719 will apply to our protein.

- **Return to the InterProScan results page.**

The InterProScan results present all of the signatures that match our protein.

? Which Parent/Child and Contains/Found in relationships apply to our protein?

- **Return to the IPR000719 entry page.**

At the top of the page are a number of different views possible for the proteins within this entry. The Overview provides a summary of all the InterPro entries that match the proteins identified by IPR000719. The Detailed view expands the overview, to display all the signatures that match the proteins identified by IPR000719, along with the InterPro entry to which they belong. The Table view lists the sequence covered by each signature in IPR000719, as well as the structural coverage of the proteins in tabular format. In addition, the Overview and Detailed view can be sorted by accession number or name, and can be narrowed to just those proteins that have structural information, or those with known splice variants.

- **Follow the link to "of known structure" under "Detailed" view.**
- **Look at the protein labelled CHK1_HUMAN (O14757).**

CHK1_HUMAN has a PDB structure (green striped bar) for its kinase domain under "Structural features". This kinase domain has been classified by both CATH (pink striped bars) and SCOP (black striped bar). Frequently, SCOP will classify a functional unit composed of more than one structural domain as a single entry, whereas CATH will always split the sequence into individual structural domains. This protein also has a ModBase (yellow striped bar) homology model under "Structural predictions". Homology models are only included for regions of a sequence where there is no structure available (in which case, the entire homology model is included).

? Use the links to the domain classifications in CATH and SCOP to explain how the classification of the kinase domain differs between these two databases?

- **Have a look at the structure of the kinase domain using AstexViewer®, by clicking on the ⏏ symbol adjacent to the N-terminal CATH domain (3.30.200.20.12).**

Notice that the selected CATH domain is highlighted in yellow on the structure, with the remaining region of the PDB being in green.

- **The image can easily be rotated. Place the mouse over the image, and by holding down the left mouse button, the mouse can be used to rotate the structure to any angle you want in order to get a better view.**
- **Try circling the mouse both clockwise and anti-clockwise to rotate the image in opposite directions.**

There are many ways to view the structure. The default view shows the molecule as a cartoon structure.

- **To change the view, click on "Protein", and from the drop-down menu you can select different views. Try the "Ball&Stick" view, or the "Sphere" view for a contrast.**
- **Use "Reset view" to return the structure to the original cartoon view.**

? What is the predominant topology of this protein, alpha helix or beta sheet?

The ligand is also visible in the structure, and its view can be manipulated independently of the protein structure.

- **To change the view of the ligand, click on "Ligand", and from the drop-down menu you can select different views, such as "Sphere".**
- **Use "Reset" button under "Ligand" to return ligand to original line view.**

The chemistry involved between the active site residues and its ligand can be viewed.

- **Click on "Chemistry", and then select the structure (1,1ia8:SO4).**

A pop-up window should appear detailing the chemical interactions.

- **Change the structure of the protein to "Line" by selecting "Line" and deselecting "Cartoon" under "Protein".**

This allows specific residues to be selected more easily.

The sequence for the structure can be seen at the foot of the Astex window.

- **To move along the sequence, simply hold the cursor in the lower section that contains the sequence, and move the mouse left to scroll towards the C-terminus, and right for the N-terminus. By hovering the mouse over the sequence, the 3-letter amino acid code for the residue, as well as its position, will appear as a footnote.**
- **Scroll along the sequence to residue 42 (K - lysine) and clink on "K".**

The image will zoom in to the lysine at position 42 of the structure (it will state residue 43 in the footnote, because using the sequence it will start at position 2).

- **Use the "zoom out" button to return the structure to its full view.**

Note that the residue is still highlighted in thick yellow on the structure, and it is numbered in the sequence.

? Is the lysine at residue 42 buried or on the external surface of the protein?

You can find out the type of amino acid and its position in the sequence for any residue on the structure simply by hovering the mouse over the structure. The residue you are at will appear as a footnote.

- **Click on the residue that lies adjacent to lysine 42 (residue 43, but stated as 44 in the footnote), but this time clicking on the structure itself.**

The image will zoom in to that residue, which will be highlighted in thick yellow and marked by its amino acid type and position.

- **Use the "zoom out" button to return the structure to its full view.**

Note that the residue is still highlighted and marked.

? What residue is at position 43, adjacent to the lysine?

- **Return to the InterProScan results page.**

- **To find out about the other domain found in our protein, namely the N-terminal Doublecortin domain, click on the InterPro icon for IPR003533.**
- **Scroll down to the GO terms.**

? What GO terms are associated with this domain?
? Using the GO terms from both domains, can you start to build up a picture of the possible functional role of this protein?

- **Scroll down to the Structural links.**
- **Follow the "SCOP d.15.11.1" link to the SCOP database to find out the structural classification of this domain.**

? What type of structure does the doublecortin fold consist of?
? Is the doublecortin fold related to other folds?
? Which families does SCOP list as containing doublecortin domains?

- **Go back to the IPR003533 entry page**
- **Scroll down to the Database links.**
- **Follow the "PROSITE doc: PDOC50309" link to the PROSITE documentation page.**

? What additional information can you gain about the function of this domain?

- **Go back to the IPR003533 entry page.**
- **Scroll down to the Taxonomic coverage.**

? What are the predominant organisms predicted to contain a doublecortin domain based on hits to this entry?

- **Using the taxonomic wheel, follow the link to the Fruit Fly proteins identified by this entry.**

This will provide an overview of the InterPro entries that match each of these proteins. The "index" on the left contains a list of all the Fruit Fly proteins. Our protein is Q9VCL7_DROME.

? Can you find two other Drosophila proteins with InterPro signature hits that are similar to our own protein?

- **Go back to the IPR003533 entry page.**
- **Scroll to the top of the page to the "Matches", and follow the link to "of known structure" under "Detailed" view.**
- **Look at the protein DCAK1_HUMAN (O15075).**

DCAK1_HUMAN has a PDB structure for one of its two doublecortin domains (note, our protein only has one doublecortin domain), which have been classified in CATH and SCOP. There is also a ModBase homology model predicting the structure of the protein kinase domain.

- **The structure of the doublecortin domain can be viewed using AstexViewer®, by clicking on the ⏏ symbol adjacent to the CATH or SCOP domain.**
- **On the left-hand side of the Detail view of DCAK1_HUMAN is a list of links. Follow the link marked "Variant" in order to view all the known splice variants associated with this protein.**

? How many splice variants are missing one or both of the doublecortin domains?

- **Go back to the IPR003533 entry page.**
- **At the top of the page under "Matches" is a link to "Architectures".**

This provides a summary of the domain organisation of all the proteins within this entry, which are depicted using cartoon views.

- **Follow the link to "Architectures" for IPR003533.**

The doublecortin domain is abbreviated as "DCX", and the protein kinase domain as "Prot kinase". From InterProScan, we know that our protein has one N-terminal doublecortin domain, and one C-terminal protein kinase domain.

? Is the doublecortin domain only associated with protein kinase domains?
? From the list of architectures of proteins matching IPR003533, how many proteins have the same domain architecture as our protein?

(Note that those proteins containing two doublecortin domains have their DCX domains represented only once, but are tagged IDA3533x2 to indicate the presence of two IPR003533 domains).

- **Follow the link on the left of the cartoon for the domain architecture of our protein (marked "IDA3533,719") in order to see a condensed graphical overview of all of these doublecortin-kinase containing proteins.**

Homology to related proteins is a powerful tool for gaining information on a particular protein. Using InterProScan, and by exploring the links to the relevant InterPro entries, the domain architecture of our test protein was predicted, and it has been possible to gain information regarding the function of those domains in related proteins. The predicted structure of our test protein can be gained by analogy to related proteins of known structure, and the classification of those structures and their relationship to other folds can be explored. As such, we can begin to build up a picture of the predicted structure and the possible functional roles of our protein, even though no experimental data may yet exist. This gives us powerful insights, which can be used to design more focused experiments that address the true functional role of our protein *in vivo*, and its possible interactions with other proteins.

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

This is the end of the short tour of the InterPro 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
```