TEACHING THE GENOME GENERATION

Connecting Protein Structure and Function using PolyPhen-2, UniProt, and BLAST

The Jackson Laboratory

Leading the search for tomorrow's cures

NIH National Institutes of Health
Turning Discovery Into Health

SEPA SCIENCE EDUCATION PARTNERSHIP AWARD

SUPPORTED BY THE NATIONAL INSTITUTES OF HEALTH

Connecting Protein Structure and Function using PolyPhen-2, UniProt, and BLAST

Introduction

The researchers sequenced your daughters' DNA and identified differences between their genomes and the human reference genome. Based on the variants in your daughters' DNA, they eventually identified ten people with similar symptoms and similar variants from four other families. The researchers sequenced those patients' genomes too.

Ultimately, they identified variants in all of the patients' genomes in a gene called Autophagy related 7 (*ATG7*), which codes for the Autophagy related 7 protein (ATG7). Through DNA sequence comparison, the researchers found that the patients' DNA variants result in changes in the amino acid sequence of the patients' ATG7 proteins.

In this activity, you will follow the researchers' process to learn how they investigated whether a change in the ATG7 amino acid sequence can impact ATG7 protein function. You will specifically examine a variant from **Family 2**, which results in a change from arginine (R) to histidine (H) at amino acid #576 in the protein sequence.

Make a Prediction

Given what you know about the relationship between protein structure and function:

1. Do you think the patients' ATG7 variants will impact protein function? Why or why not?

2. What would you want to know about the ATG7 protein and the patients' variants to better predict whether the variants will impact function?

Part 1. Functional Effect Prediction

To investigate the potential impact of the *ATG7* variants on protein function, we'll use a bioinformatics tool called PolyPhen-2. This tool makes predictions about how a single amino acid change will affect protein function. Let's use this tool to investigate **Family 2 Allele 1**.

> *See PolyPhen-2 Written Tutorial*
> *Or PolyPhen-2 Tutorial Videos: Introduction, Submitting a Query using a Protein Identifier, Accessing Your Results, and Interpreting a PolyPhen-2 Report*

1. Navigate to PolyPhen-2.

2. In the field titled "Protein or SNP identifier" enter `ATG7_HUMAN`. Be sure to type in all caps.

3. The DNA variant in **Family 2 Allele 1** results in a change from Arginine (R) to Histidine (H) at amino acid #576 in the protein sequence.
    a. In the field titled "Position", enter `576`.
    b. Next to "AA_1", select `R` for Arginine.
    c. Next to "AA_2", select `H` for Histidine.
    d. In the field titled "Query description", enter `Family 2 Allele 1`.

4. Click the button labeled "Submit Query".

5. When the next page loads, wait for ~1 minute, then click the button labeled "Refresh". There should now be a link labeled "View" in the Results column. Click on that link.
    Troubleshooting notes:
    If the link does not appear, wait another minute, then click "Refresh" again.
    If you get an error, return to Step 1 and try entering the information again.

6. The results page should open either in a new tab or in the same window. When the page appears, find the "Results" section and locate the "Prediction/Confidence" report.
    a. What effect does PolyPhen-2 predict this variant will have on ATG7 protein function?

    b. Given what you know about the connection between structure and function, what information about the protein might PolyPhen-2 be using to make this prediction?

Part 2a. ATG7 Function

Next, we will look at a database called UniProt. UniProt contains information about the sequence, structure, and function of different proteins. We will be using UniProt to learn more about how the function of ATG7 is related to its structure. First, we will learn more about the function of ATG7.

See UniProt Written Tutorial: Navigating a UniProtKB Entry – Protein Function
Or UniProt Tutorial Videos: Searching for a Protein, Navigating a UniProtKB Protein Entry and Finding Information on Protein Function

1. Navigate to the entry for Human ATG7 on UniProt.

2. The first section of the main entry page is titled "Function" and includes a description of the protein's function. Read over the questions below, then read through the description. Take notes under each question as you read.

    a. What are some science words you recognize?

    b. What are some science words that are unfamiliar to you?

    c. What type of protein is ATG7? Is it a structural protein, a transport protein, a hormone, a contractile protein, a defense protein, an enzyme, or a storage protein?

    d. ATG7 is involved in a process called autophagy. What do you think autophagy means?
    Hint: What other words appear near "autophagy" and "autophagic" in the description?

Part 2b. Structural Changes

Next, we will use UniProt to learn more about the 3-dimensional (3D) structure of ATG7.

> *See UniProt Written Tutorial: Navigating a UniProtKB Entry – Protein Function*
> *Or UniProt Tutorial Videos: Searching for a Protein, Navigating a UniProtKB Protein Entry and Finding Information on Protein Function*

1. Navigate to the entry for Human ATG7 on UniProt.

2. Locate the "Function" section of the entry and then find the heading "Features". Notice the length of the protein is shown in grey from amino acid 1 to 703. Underneath the grey bar, colored shapes identify sections of the protein that are important for its function, also known as the protein's features. These features are also listed in the table below the grey bar.

3. Are there any protein features near the variant site (amino acid #576)? If so, which feature(s)?

The active site is the portion of an enzyme's structure that is directly responsible for interacting with other proteins or molecules.

4. How do you think a variant near the active site of ATG7 affect the protein's ability to complete its enzymatic function?

In addition to looking at the location of the variant within the amino acid sequence, we can also look at the protein's 3D structure. Remember that proteins aren't just flat sequences of amino acids—those amino acid strands fold up into a specific shape to make a functioning protein! We can represent a protein's 3D structure using connected spheres to represent all of the amino acids of the protein. **Figure 1** is a representation of the 3D structure of ATG7 protein from yeast.

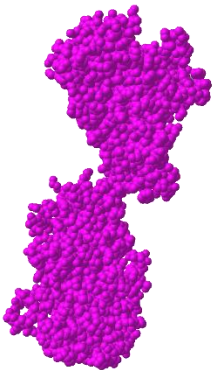

Figure 1. ATG7 3D Protein Structure[1]

Did you know? We are looking at the yeast ATG7 protein because determining the structure of protein is very difficult, and scientists have not yet determined the exact structure of human ATG7 protein. However, the yeast and human ATG7 proteins are similar enough in sequence that we can still get useful information from the 3D structure of the yeast protein.

Scientists studied ATG7 protein in model organisms, like yeast, and discovered that two ATG7 proteins need to come together and attach to each other to work properly. **Figure 2** shows two ATG7 proteins (one pink and one blue) joining together to form one functioning structure.

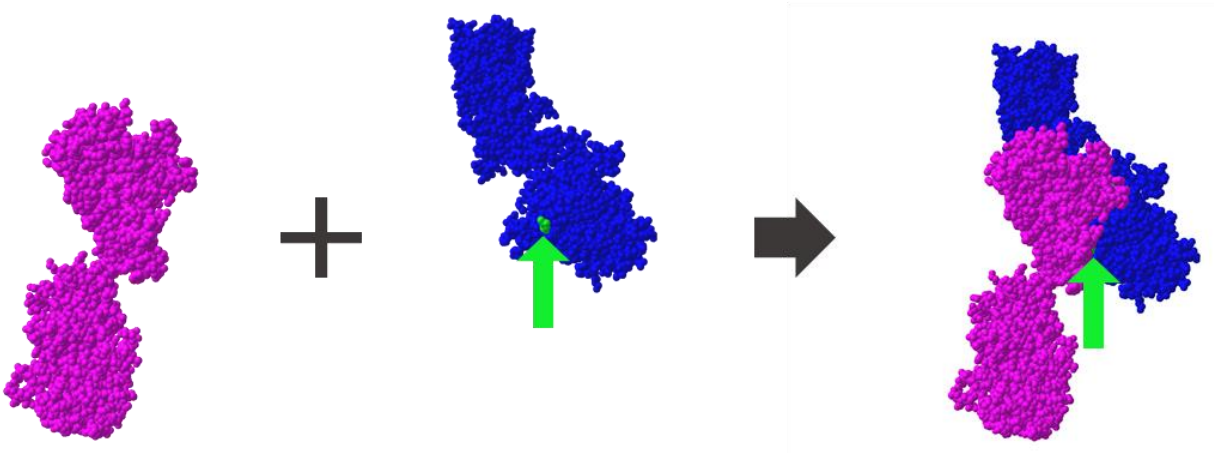

Figure 2. ATG7 Protein Interactions.[1] Two copies of the ATG7 protein interact with each other to form one functioning structure. Amino acid 576, the amino acid changed by the Family 2 Allele 1 variant, is highlighted in green.

All amino acids have unique chemical structures, leading to different protein sizes, shapes, and ways of interacting with other molecules depending on which amino acids are present. Remember that ATG7 amino acid 576 (the highlighted location in Figure 2) is typically an Arginine, but in proteins expressed from Family 2 Allele 1, that amino acid will instead be a Histidine.

5. Do you think this amino acid change will affect the ability of two ATG7 proteins to join together? Justify you answer using **Figure 2** and/or your knowledge of amino acid properties.

6. Researchers found that the two patients in Family 2 had a higher percentage of individual ATG7 proteins and a lower percentage of paired ATG7 proteins when compared to controls without the ATG7 variant. Does this evidence support your prediction in question 5? Why or why not?

[1] Structure from Kaiser SE, et al. (2012). Nat Struct Mol Biol, 19:1242-9. Data obtained through the Molecular Modeling Database (MMDB) and modeled using iCn3D.

Part 2c. Amino Acid Properties

When two ATG7 proteins join together, the R group of the arginine at amino acid #576 of one of the two proteins interacts with the R group of the glutamic acid at amino acid #589 on the other protein. Use Figure 3, which shows the structure of arginine and glutamic acid, to answer the following question.

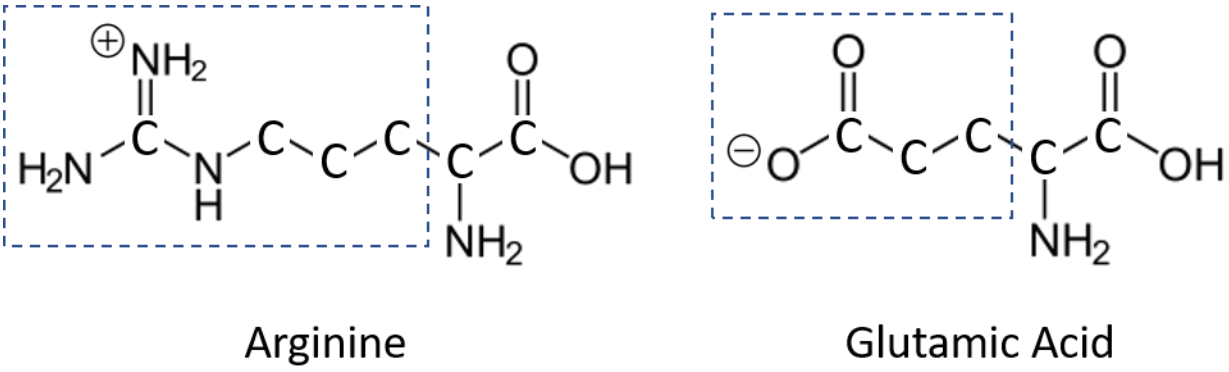

Figure 3. Structure of Arginine and Glutamic Acid. The R group, also known as the side chain, of each amino acid is highlighted by a blue box.

1. What type(s) of interaction(s) could be occurring between the R group of the arginine and the R group of the glutamic acid?

The variant in Family 2 Allele 1 results in an amino acid change from arginine to histidine at amino acid #576. Use Figure 4, which shows the structures of arginine and histidine, to answer the following questions.

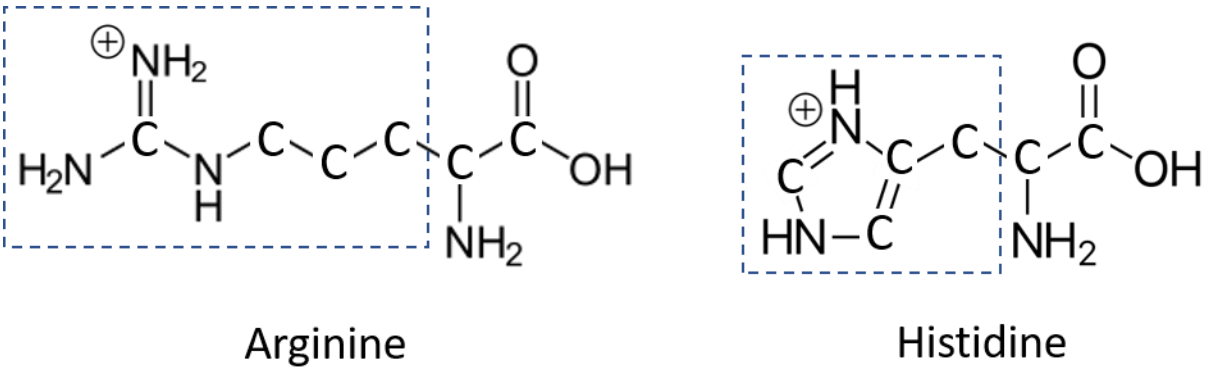

Figure 4. Structure of Arginine and Histidine. The R group, also known as the side chain, of each amino acid is highlighted by a blue box.

2. What similarities and differences do you notice in the structure of these two amino acids? You may want to think about size, shape, and charge.

3. How would you predict the interactions of amino acid #576 with the glutamic acid at amino acid #589 might change when the arginine at #576 is replaced by histidine?

Part 3. Conserved Sequences

One way to identify regions of a protein that are important for protein function is to look for conserved sequences. A conserved sequence is a protein or DNA sequence that is identical or highly similar across multiple species. If a region of a protein is conserved across many species, we can often assume that the structure of that region is important for proper protein function.

To check if the variant in Family 2 Allele 1 occurs in a conserved sequence, you will compare the human ATG7 protein reference sequence to ATG7 protein sequences from fruit fly, chimpanzee, yeast, chicken, and mouse.

Before you begin, let's make some predictions about how similar each of these species' ATG7 protein sequences is to the human ATG7 sequence.

1. Which species (chimpanzee, mouse, chicken, fruit fly, or yeast) do you think is most closely related to humans? Which species is least closely related to humans?

2. Given your answer to the previous question, which species do you expect to have the greatest percent identity for ATG7 protein sequence compared to human? What percent identity do you expect for this comparison?

3. Which species do you expect to have the lowest percent identity for ATG7 protein sequence compared to human? What percent identity do you expect for this comparison?

Now, let's check if the variant occurs in a conserved sequence by using BLAST.

See BLAST Tutorial Video: Comparing two or more protein sequences

4. Locate the ATG7 protein sequences.

5. Navigate to Protein BLAST and select the option to "align two or more sequences."

6. In the box labeled "Enter Query Sequence," paste the "Human ATG7 Protein" sequence. Copy the entire text of the sequence as it is written in the sequences file including the description line beginning with ">."

7. In the box labeled "Enter Subject Sequence," paste the chimpanzee, mouse, chicken, fruit fly, and yeast sequences. You can copy and paste all five sequences simultaneously.

8. Scroll to the bottom and click "BLAST." It may take several seconds for the results to appear.

9. Once the results load, scroll down to the table and navigate to the "Alignments" tab. In the "Alignment view" dropdown menu, select "Pairwise with dots for identities" to compare the five sequences to the human protein sequence.

10. Scroll down the page and look at each comparison. Within each comparison, the **Query** sequence is the human sequence, and the subject (**Sbjct**) sequence is either the chimpanzee, mouse, chicken, fruit fly, or yeast sequence.
    a. What do you notice about the similarities and differences between each subject sequence and the human sequence?

11. Now locate the percent identity score for each comparison. Enter the values into the second column of Table 1 (**% Identity for Whole Protein**). Then answer the questions below.

Table 1.

Comparison to Human ATG7	**% Identity for Whole Protein**	**% Identity for Variant Region**
Chimpanzee		
Mouse		
Chicken		
Fruit Fly		
Yeast		

b. What is the range of percent identity across all of the comparisons?

c. Which comparison has the highest percent identity? Which has the lowest?

d. Does this data match your predictions? Why do you think those two organisms have the highest and lowest percent identities?

Next, to see if this variant is located in a region that is conserved, we'll look at the percent identity of a 50 amino acid region around the site of the variant.

12. Copy the sequence below and search for it on your BLAST results page. On your browser select "Edit > Find" or use keyboard shortcuts Ctrl+F or Cmd+F. The location of the variant is bolded for your reference. Searching for the sequence should highlight our region of interest in all five comparisons.

    CYFCNDVVAPGDSTRDRTLDQQCTVS**R**PGLAVIAGALAVELMVSVLQHPE

    If this does not work in your browser, look for the region between amino acids 550 and 599 of the **Query** sequence. The sequence you identify should match the above sequence, starting with C and ending with E.

13. In the human reference sequence (the **Query** sequence), the amino acid at the variant site (#576) is an arginine (**R**). In the other five species, is the equivalent amino acid also an arginine? Or does it vary between the different species?

14. For each comparison, calculate the percent identity of **the 50 amino acid region only**.
    e. Calculate the percent identity of this region for:
        i. Chimpanzee

        ii. Mouse

        iii. Chicken

        iv. Fruit Fly

        v. Yeast

    f. Enter the values above into the third column of Table 1. Are these values a high or low level of percent identity?

g. How does the percent identity of the region around the variant compare to the percent identity for the whole protein?

15. Is the variant in **Family 2 Allele 1** in a conserved region of the protein? Use the data you collected to justify your answer.

Part 4. Reflection

Activity Reflection

1. At the start of this activity, you predicted whether the patients' variants would impact ATG7 protein function. After investigating the effect of one variant from Family 2, do you still agree with your initial prediction? Why or why not?

2. In Part 1 of this activity, you used PolyPhen-2 to predict the effect of the variant from Family 2 on ATG7 function. Based on what you learned from the rest of the activity, what information about the protein do you think PolyPhen-2 is using to make this prediction? Is this the same or different from what you predicted at the end of Part 1?

Guiding Questions Reflection

Revisit the following guiding questions and update your answers to include anything you've learned during this activity.

Given what you know about genome sequencing and genetic variation:

1. What can we learn from comparing genetic information across individuals and species?

Given what you know about (a) how DNA codes for proteins and (b) the connection between protein structure and function:

2. How might a DNA variant affect protein sequence, structure, or function?

ATG7 Protein Sequences

Note: If BLAST returns an error after copying sequences from this PDF, try instead to copy and paste these sequences from our ATG7 Protein Sequences Word document.

Human ATG7 Sequence

>Human ATG7 Protein
MAAATGDPGLSKLQFAPFSSALDVGFWHELTQKKLNEYRLDEAPKDIKGYYYNGDSAGLPARLTLEFSAFDM
SAPTPARCCPAIGTLYNTNTLESFKTADKKLLLEQAANEIWESIKSGTALENPVLLNKFLLLTFADLKKYHFYYW
FCYPALCLPESLPLIQGPVGLDQRFSLKQIEALECAYDNLCQTEGVTALPYFLIKYDENMVLVSLLKHYSDFFQG
QRTKITIGVYDPCNLAQYPGWPLRNFLVLAAHRWSSSFQSVEVVCFRDRTMQGARDVAHSIIFEVKLPEMAF
SPDCPKAVGWEKNQKGGMGPRMVNLSECMDPKRLAESSVDLNLKLMCWRLVPTLDLDKVVSVKCLLLGA
GTLGCNVARTLMGWGVRHITFVDNAKISYSNPVRQPLYEFEDCLGGGKPKALAAADRLQKIFPGVNARGFN
MSIPMPGHPVNFSSVTLEQARRDVEQLEQLIESHDVVFLLMDTRESRWLPAVIAASKRKLVINAALGFDTFV
VMRHGLKKPKQQGAGDLCPNHPVASADLLGSSLFANIPGYKLGCYFCNDVVAPGDSTRDRTLDQQCTVSRP
GLAVIAGALAVELMVSVLQHPEGGYAIASSSDDRMNEPPTSLGLVPHQIRGFLSRFDNVLPVSLAFDKCTACS
SKVLDQYEREGFNFLAKVFNSSHSFLEDLTGLTLLHQETQAAEIWDMSDDETI

Chimpanzee, Mouse, Chicken, Fruit Fly, and Yeast ATG7 Sequences

>Chimpanzee ATG7 Protein
MAAATGDPGLSKLQFAPFSSALDVGFWHELTQKKLNEYRLDEAPKDIKGYYYNGDSAGLPARLTLEFSAFDM
SAPTPARCCPAIGTLYNTNTLESFKTADKKLLLEQAANEIWESIKSGAALENPVLLNKFLLLTFADLKKYHFYYW
FCYPALCLPESLPLIQGPVGLDQRFSLKQIEALECAYDNLCQTEGVTALPYFLIKYDENTVLVSLLKHYSDFFQG
QRTKITIGVYDPCNLAQYPGWPLRNFLVLAAHRWSSSFQSVEVVCFRDRTMQGARDVAHSIIFEVKLPEMAF
SPDCPKAVGWEKNQKGGMGPRMVNLSECMDPKRLAESSVDLNLKLMCWRLVPTLDLDKVVSVKCLLLGA
GTLGCNVARTLMGWGVRHITFVDNAKISYSNPVRQPLYEFEDCLGGGKPKALAAADRLQKIFPGVNARGFS
MSIPMPGHPVNFSSVTLEQARRDVEQLEQLIESHDVVFLLMDTRESRWLPAVIAASKRKLVINAALGFDTFV
VMRHGLKKPKQQGAGDLCPNHLVASADLLGSSLFANIPGYKLGCYFCNDVVAPGDSTRDRTLDQQCTVSRP
GLAMIAGALAVELMVSVLQHPEGGYAIASSSDDRMNEPPTSLGLVPHQIRGFLSRFDNVLPVSLAFDKCTAC
SSKVLDQYEREGFNFLAKVFNSSHSFLEDLTGLTLLHQETQAAEVRTKQAFCSED

>Mouse ATG7 Protein
MGDPGLAKLQFAPFNSALDVGFWHELTQKKLNEYRLDEAPKDIKGYYYNGDSAGLPTRLTLEFSAFDMSAST
PAHCCPAMGTLHNTNTLEAFKTADKKLLLEQSANEIWEAIKSGAALENPMLLNKFLLLTFADLKKYHFYYWFC
CPALCLPESIPLIRGPVSLDQRLSPKQIQALEHAYDDLCRAEGVTALPYFLFKYDDDTVLVSLLKHYSDFFQGQR
TKITVGVYDPCNLAQYPGWPLRNFLVLAAHRWSGSFQSVEVLCFRDRTMQGARDVTHSIIFEVKLPEMAFSP
DCPKAVGWEKNQKGGMGPRMVNLSGCMDPKRLAESSVDLNLKLMCWRLVPTLDLDKVVSVKCLLLGAG
TLGCNVARTLMGWGVRHVTFVDNAKISYSNPVRQPLYEFEDCLGGGKPKALAAAERLQKIFPGVNARGFN
MSIPMPGHPVNFSDVTMEQARRDVEQLEQLIDNHDVIFLLMDTRESRWLPTVIAASKRKLVINAALGFDTFV
VMRHGLKKPKQQGAGDLCPSHLVAPADLGSSLFANIPGYKLGCYFCNDVVAPGDSTRDRTLDQQCTVSRPG
LAVIAGALAVELMVSVLQHPEGGYAIASSSDDRMNEPPTSLGLVPHQIRGFLSRFDNVLPVSLAFDKCTACSP
KVLDQYEREGFTFLAKVFNSSHSFLEDLTGLTLLHQETQAAEIWDMSDEETV

>Chicken ATG7 Protein
MAAVSNESQNPVDPGSSKLQFAPFSSALNVGFWHELTQKKLNEYRLDETPKVIKGYYYNGDPSGFPARLTLE
YSAFDINASIPARCCPAFGTLYNTNTFETFKSCDKKSLLEKEANEIWESIKSGAALENPMLLNRFLLLTFADLKKY
HFYYWFCYPALCFPDGIHVIQKPVCLGDRFSLNQIQALQKAYDELCQTEGVTAFPYFLIKYHDNSVVVSPLKK
WDGFFQDQGGKVTVGVYDPCNLSHYPGWPLRNFLILASHKWGNILQSIEVLCFRDRTMQGVRDITHSIIFEI
KLPQGAFGPDCPKAVGWEKNQKGGMGPRVVNLSECMDPKRLAESSVDLNLKLMCWRLVPTLDLEKIVSAK
CLLLGAGTLGCSVARTLMGWGVRKITFVDNARISYSNPVRQPLYEFEDCLSGGKPKALAAAERLQKIFPGVNS
EGYNMSIPMPGHPVNFSEVTMAQARKDVATLEELIDAHDVVFLLMDTRESRWLPAVIAASKRKLVINAALG
FDTFVVMRHGLKKPKQQETGNACFSTAPGPSDLLGSSLFSNIPGYKLGCYFCNDVVAPGDSTRDRTLDQQCT
VSRPGLAMIAGALAVELMVSVLQHPEGGYAVASSSDDRMNEPPTSLGLVPHQIRGFLSRFDNVLPVSLAFDK
CTACSPKVLDQYEREGFNFLAKVFNSSHSFLEDLTGLTLLHQETQAAEIWDMSDDETV

>Fruit Fly ATG7 Protein
MSTEKEIILQFAPWESFVSPTFWHKLAELKLDHDRLSDSKRSITGHYTNRNASGCLLEVDYTAYNRMAKPPKF
SHSAIGTIYNKNTIEEFKALDKLQLLADEGKELLADMCSGGALRDPSLLTRFFVLSFADLKCHSYYYWFAFPCPL
TPTLKLQGAVQKLRDLPNSSSYIMALKALPTESQNFFILYANVEKNIFEARSLSSLDDKNVEFCYFGFADPSEYE
HPAWIMRNYAAFLLQQCPSFVGKPLKFLGLRHNQQMNIDDSLVWKVIQTEACDLSQSENIKFVGWELNKN
GKMGPRMVCMRDSMDPAKLAENSVNLNLKLMKWRLVPDLNLEIISQTKCLLFGAGTLGCAVARNLLSWGF
KHITLLDSGKVGFSNPVRQNLYTHADAVAGNRMKATTAAQRLKEINPSAETAGYVLEIPMPGHTIGESLLAQ
TKEHLKVIEKLVQDHDVIFLLTDSRESRWLPTLLGAAKEKIVINAALGFDSYLVMRHGTTRKEAGDDGQEIEGL
KCINGDQLGCYFCNDVTAPGNSLKDRTLDQQCTVTRPGVSNIAASYAVELLVALLQHPRKELAPAYYAQSGR
GRSEETEEKVPEGLLGILPHSIRGMLCNYENILPATQKFAQCIACSAAVLNEYKKEGHAFLFKTFETAKFLEDLT
GISEFKRLNSEIIDFDDEEFDMSDSDED

>Yeast ATG7 Protein
MSSERVLSYAPAFKSFLDTSFFQELSRLKLDVLKLDSTCQPLTVNLDLHNIPKSADQVPLFLTNRSFEKHNNKRT
NEVPLQGSIFNFNVLDEFKNLDKQLFLHQRALECWEDGIKDINKCVSFVIISFADLKKYRFYYWLGVPCFQRPS
STVLHVRPEPSLKGLFSKCQKWFDVNYSKWVCILDADDEIVNYDKCIIRKTKVLAIRDTSTMENVPSALTKNFL
SVLQYDVPDLIDFKLLIIRQNEGSFALNATFASIDPQSSSSNPDMKVSGWERNVQGKLAPRVVDLSSLLDPLKI
ADQSVDLNLKLMKWRILPDLNLDIIKNTKVLLLGAGTLGCYVSRALIAWGVRKITFVDNGTVSYSNPVRQALY
NFEDCGKPKAELAAASLKRIFPLMDATGVKLSIPMIGHKLVNEEAQHKDFDRLRALIKEHDIIFLLVDSRESRW
LPSLLSNIENKTVINAALGFDSYLVMRHGNRDEQSSKQLGCYFCHDVVAPTDSLTDRTLDQMCTVTRPGVA
MMASSLAVELMTSLLQTKYSGSETTVLGDIPHQIRGFLHNFSILKLETPAYEHCPACSPKVIEAFTDLGWEFVK
KALEHPLYLEEISGLSVIKQEVERLGNDVFEWEDDESDEIA