Name: _____Feng Geng_____

BIOC3010/BIOCM010/BIOCG010 Coursework

This year we have modified the coursework somewhat in order to handle the large number of students taking the course. In previous years it has been a free-form essay. This year we are asking for a more structured piece of work to make marking more straightforward.

The instructions for the coursework remain the same except that you should complete the work on this sheet rather than providing an essay – see Moodle

```
BIOC3010 -> Topic 3 -> Coursework information
```

for what is expected. This page also gives a number of tips based on questions we have been asked in previous years. See

```
http://www.bioinf.org.uk/teaching/c40/practicals/essay/
```

to obtain the 'mystery sequence'

Make sure you read through all the questions on this sheet before you start work so you don't duplicate your answers across different sections.

Ensure that you explain your answers and how you obtained them.

You may expand the boxes as required, but ensure that the completed work does not exceed 10 pages (you can remove the page breaks).

PLEASE SUBMIT YOUR WORK AS A PDF FILE

Introduction, Tools and Resources

Provide a list of all tools and resources you have employed including appropriate URLs and references. [7%]

Tool	URL	Acc.
NCBI-pBlast[5]	blast.ncbi.nlm.nih.gov	
Uniprot[6]	www.uniprot.org	P11979,P30613,P14618
RCSB PDB[7]	www.rcsb.org	1PKM, 4FXF
Swiss-model[8]	swissmodel.expasy.org	D3D7Z6
Verify3D[9]	services.mbi.ucla.edu/Verify_3D	
ProFit[10]	www.bioinf.org.uk/programs/profit/	
OMIM	www.omim.org	179050,609712
CATH[11]	www.cathdb.info	1pkma02,1pkma03
SCOPe[1]	scop.berkeley.edu	d1pkma1,d1pkma2,d1pkma3
EMBOSS water[2]	www.ebi.ac.uk/Tools/psa	
SAS[3]	www.ebi.ac.uk/thornton-srv/databases/sas/	
Rasmol[4]	www.umass.edu/microbio/rasmol/	

References:

1. Fox, N. K., Brenner, S. E. & Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42,** D304–D309 (2014).

2. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).

3. Milburn, D., Laskowski, R. A. & Thornton, J. M. Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.* **11,** 855–859 (1998).

4. Sayle, R. A. & Milner-White, E. J. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20,** 374 (1995).

5. Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7,** 12 (2012).

6. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

7. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235 (2000).

8. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42,** W252–W258 (2014).

9. Luthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356,** 83–85 (1992).

10. McLachlan, A. D. Rapid comparison of protein structures. *Acta Crystallogr. Sect. A* **38,** 871–873 (1982).

11. Dawson, N. L. *et al.* CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45,** D289–D295 (2017).

Name: _____Feng Geng_____

Initial identification of the likely nature of the sequence

State what tool(s) and databank(s) you used for a simple homology search and summarize the output from the sequence database search(es). [5%]

We will refer our mysterious sequence as the query sequence in the following.
The query has a length of 531aa.

We used pblast from NCBI website to search the non-redundant(nr) NCBI protein collection for homology. This search returns 100 hits of over 95% identity (including predicted genes). The top hit is Pyruvate kinase muscle isozyme (PKM, Uniprot acc:P11979.2) from *Felis catus* (cat), with 360/361 identity. We denote this as cat_PKM. The query differs from the PKM at position 75, where PKM holds a N but query holds an R. Different PKM isoforms from a variety of species were also returned as hits. Addtionally, the search returned a protein structure for cat_PKM (pdb:1PKM_A).

From what species did the best hit come? [5%]

Cat

Give an accession code for the best hit [4%]

P11979.2

Given the results of the search, what is the function of this protein likely to be? [5%]

It is likely to be a kinase that catalyse phosphorylation.

Further sequence analysis

Align your mystery sequence with close relatives and identify where any differences occur in the sequence. Include an image or images if it helps. [5%]

The closet relative is cat_PKM. The query differs from cat_PKM at position 65.
query: LKEMIKSGMNVARLRFSHGTHE
cat_PKM:LKEMIKSGMNVARLNFSHGTHE
75
query holds an R
cat_PKM holds an N.

Position 75 is known to involve in potassium binding.
(pos 63 in the model produced since 1-11 is missing from the structure)

Use secondary database searches to look for features of the sequence. What tool(s) did you use? [2%]

Uniprot

What information have you gained about the sequence? [5%]

The sequence is a member of the pyruvate kinase family. It is likely one of the many variants of PKM.

The sequence receives heavy post-translational modification at multiple sites. The initiator

methionine is removed post translation. The proetin has two isoform sequences (M1 and M2) and the canonical one is M1, whereas M2 is not availabe.

Functional Description

Answer the following based on the closest homologue, explaining how you obtained each piece of information.

What requirements (if any) does the protein have for co-factors? [2%]

The protein is likely to have Mg2+ and K+ as its cofactors.
Inferred from cat_PKM's uniprot page.

What is the function of the protein? (e.g. enzyme reaction catalyzed, including EC number if appropriate) [2%]

This enzyme uses PEP (phosphoenolpyruvate) to phosphorylate ADP to produce ATP and pyruvate.
PEP + ADP <=> pyruvate +ATP
EC:2.7.1.40
Inferred from cat_PKM's uniprot page.

In what pathway(s) is the protein involved? [2%]

It is part of the glycolysis pathway.
Inferred from cat_PKM's uniprot page.

Are there any variants (e.g. isoforms, alternative spliced versions) of the protein? If so what do they do? [2%]

The cat_PKM is reported to have two isoforms M1 and M2. However the M2 is not available on uniprot. We turned to human PKM page (P14618). The human_PKM has 3 varaints, M1, M2 and M3. The M2 is canonical and M1 differs from M2 at 389-433, whereas the M3 differs from M2 at 1-82. The is also a natural variant with a G→V mutation at 204.

The human_PKM is translocated to nucleus in response to apopotic stimuli. This activity is isoform-specific.
M2 interacts with EGLN3 and HF1A, enhancing transcription under hypoxia.
M2, but not M1, interacts with TRIM35 and prevent FGFP1-dependent tyrosine phosphorylation.
The distribution of M1 and M2 are also different. M1 is the main form in muscle, heart and brain. M2 is mainly found in early fetal tissues and in most cancer cells.

Apart from the M1/M2 isozyme, PKL and PKR are also pyruvate kinases but are produced from PKLR, differing by differential splicing. PKL is the isoform in liver and PKR the isoform in red blood cell.
Inferred from human_PKM uniprot page.

The M1 form is fully active intrinsically, whereas M2 is activated only by PEP and by Fru-1,6-P2

inferred from OMIM.179050

M1. 1 Publication

Disease Involvement

In what diseases, if any, are this protein known to be involved? (If the best hit is a non-human protein, then you may have to look at diseases associated with the human protein and assume that other species suffer similar diseases.) [5%]

The human_PKM plays a general role in caspase independent tumour cell death. Translocation of human_PKM into nucleus causes cell death. Since it controls glycolytic ATP production, it is also important for tumour cell proliferation and survival.

Switch from M1/M2-expressing to M2-expressing is necessary for tumour to shift to aerobic glycolysis and promotes tumorigenesis. In contrast, the M1 isoform promotes oxidative phosphorylation.

inferred from OMIM.179050

The PKLR gene associated with Pyruvate kinase deficiency(PKD) and with pyruvate kinase hyperactivity.(PKH)
PKD is the most common cause of herditary nonspherocytic hemolytic anemia.
PKH is not reported to associate with any defect.
[OMIM.609712]

Are there known mutations that lead to disease and/or benign polymorphisms ? [2%]

There are few mutations listed for PKM. There is a natural variant G204V but no physiological effect was listed.

The PKLR isozyme is extensively associated with many diesases. Thirteen varaints were reported to cause PKD, whereas only one reported to cause PKH.

The mutation caused PKH is GLY37GLN.

13 mutations for PKD include:
frameshift:
deletion in exon 6
deletion in at position 823
point mutation:
ARG132CYS R→ C
THR353MET T→ M
THR384MET T→ M
GLN421LYS Q→ K
ARG479HIS R→ H
ARG510GLN R→ Q
ARG486TRP R→ W
SER130TYR S→ Y
promoter mutations: -83 G->C
premature termination: 1318G-T
splicing defect: 1269G-A

[OMIM.609712]

Pos in cat_PKM2 (0)	Human_PKLR (+43)	Known variant?
73R	116R	no
75N	118N	no
77S	120S	no
113D	156D	no
114T	157T	no
270K	313K	no
272E	315E	no
295G	338G	no
296D	339D	no
328T	371T	no

Table 1: Residues important in PKM2

Name: _____Feng Geng_____

Sequence to Structure

Without building a homology model, explain how you can map your sequence to the structure of a related protein in the PDB. [5%]

PDB structures are associated with sequences. We can blast the query against the sequence database of existing pdb structures to find structure with a similar sequence. (advanced search in RCSB PDB).

Alternatvely, given that we have a good homologue cat_PKM, we can check the structures related to it directly on its uniprot page.

With the assumption is that a single mutation does not change the structure greatly, we can use a structure with high identity.

What is the PDB code of the closest known structure? [2%]

1PKM (chain A)

Looking at the most similar PDB file in CATH, describe the structure. What are the CATH codes of the domain(s)? What are the protein fold(s)? [5%]

###Since the CATH database is down, information from SCOPe is used instead.

Since the 1PKM contains a homo-tetramer, we will focus on the chain A.
The structure conatins alternating alpha and beta strands.(21 helices and 26 strands, contributing to 36% and 19% of total length, respectively)

More specifically, the protein contains three distinct structure domains, from N terminal to C terminal. The three domains are cleanly separated in the structure sense. However, in the sequence sense, the d1pkma1 is inserted into a loop in the middle of d1pkma2, causing the d1pkma2 to become 2 fragments in the sequence.

SCOPe annotation
d1pkma1: pyruvate kinase domain pos:116-217
family: b.58.1.1 (All beta, PK beta barrel fold, PK beta-barrel superfamily, PK beta-barrel family,)
d1pkma2: pyruvate kinase N-terminal domain
pos:12-115,218-395
family: c.1.12.1 (Alpha and beta, TIM-barrel fold, PEP/pyruvate domain superfamily, Pyruvate kinase family)
d1pkma3: pyruvate kinase C-terminal doamin
pos:396-530
family c.49.1.1 (Alpha and beta, pyruvate kinase C-terminal domian superfamily, pyruvate kinase c-terminal domain family)

http://scop.berkeley.edu/sunid=27051
http://scop.berkeley.edu/sunid=29254
http://scop.berkeley.edu/sunid=33108

[CATH annotation for reference only:
1pkma03: PK beta-barrel domain-like
1pkmA02:phosphoenolpyruvate(PEP)-binding domainsdomain/

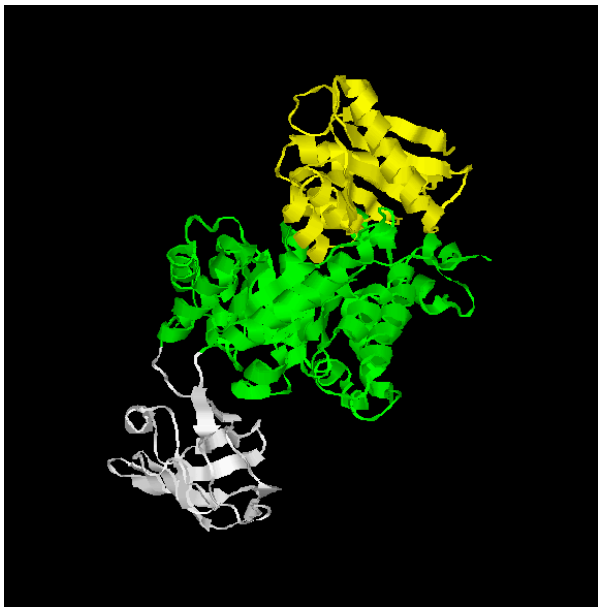


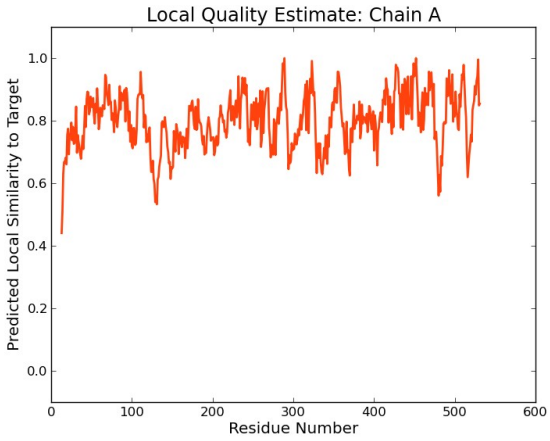
FIgure 1: Structrue of 1PKM1pkm.A annotated according to SCOP code. d1pkma1: white, d1pkma2:green ,d1pkma3:yellow Note the white domain is inserted onto a loop in green. Note the white domain is mainly beta, whereas others are α/β

The CATHdb did not assign a domian to the C-terminal region of the protein.]

http://www.rcsb.org/pdb/explore/remediatedSequence.do?structureId=1PKM
http://www.rcsb.org/pdb/explore.do?structureId=1pkm
http://www.rcsb.org/pdb/explore/macroMoleculeData.do?structureId=1PKM

Build a model of the protein and comment on the likely quality of the model [10%]

We used swiss-moodel portal to perform homology modelling. The software found 32 templates both from M1/M2 isoforms (>95% identity) and from more distant L/R isoforms (68% identity). The software automatically chose a 1pkm.1.A as its template. (99.81% identity).



The model is reported with a QMEAN at -0.70, QMEAN is an assessment function based on pseudo-energy calculated from geometrical torsion and physical interatcion. It is presented in a Z-score scheme. The negative QMEAN means the pesudo-energy is better than the average observed energy of crystal structures of a similar size.

From the swiss-model automatic quality result.
The model shows a steady similarity to the reference with an average around 0.8.This is confirmed with ProFit structure alignment, which reported an RMSD of 0.087 angstrom, meaning the model is highly similar to 1PKM.

We also used verify3D to check the structure's compatibility with its own sequence. The result shows poor model quality near pos 1-28(12-39) pos221(232), pos297(308), pos333(344), pos471(482). However, this is largely inherited from the property of 1PKM and not a result of the modelling process.

In contrast, the swiss-model compatibility descriptor only reported instability near pos(13-28),pos(130), pos(335), and pos(482). These positions are highlight with space-filling in the annotated model (figure 2), and are located on the surface of the protein, or at interface between monomers (figure not shown). We thus conclude these instability are within our explanation and the model is theoretically sensible.

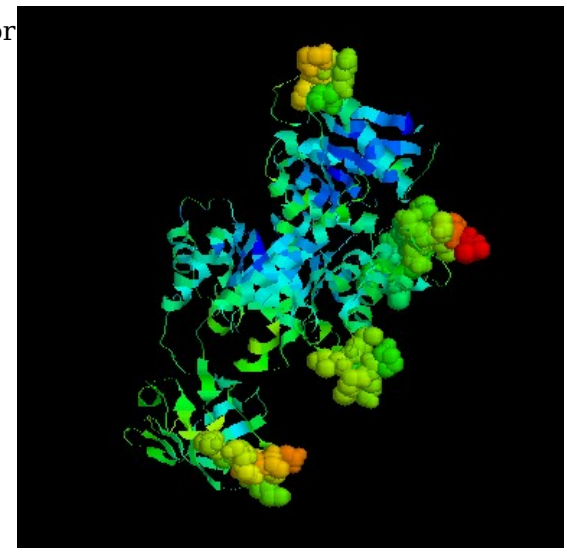


FIgure 2: Modeled structure colored by temperature/ local instability. Red correspond to high temperature, blue to low temperature.

https://swissmodel.expasy.org/interactive/D3D7Z6/models/

Name: _____Feng Geng_____

Using your model, identify where any differences are compared with the closest homologue for which a structure is available. Provide a picture highlighting these differences [5%]

Since 1PKM does not contain any ligands, we used 4FXF as our reference structure for comparision (FIgure 3). We label the ligand-interacting side chains in wireframe near the N75R mutation. This is within the d1pkma2 domain.

As from seqeuence analysis, N75R is the only mutation from cat_PKM2 to query. This mutation, however is within the active site directly responsible for catalysis of phosphorylation. ASN75 is interacting with K+ ligand and changing to ARG75 greatly affect the site morphology.

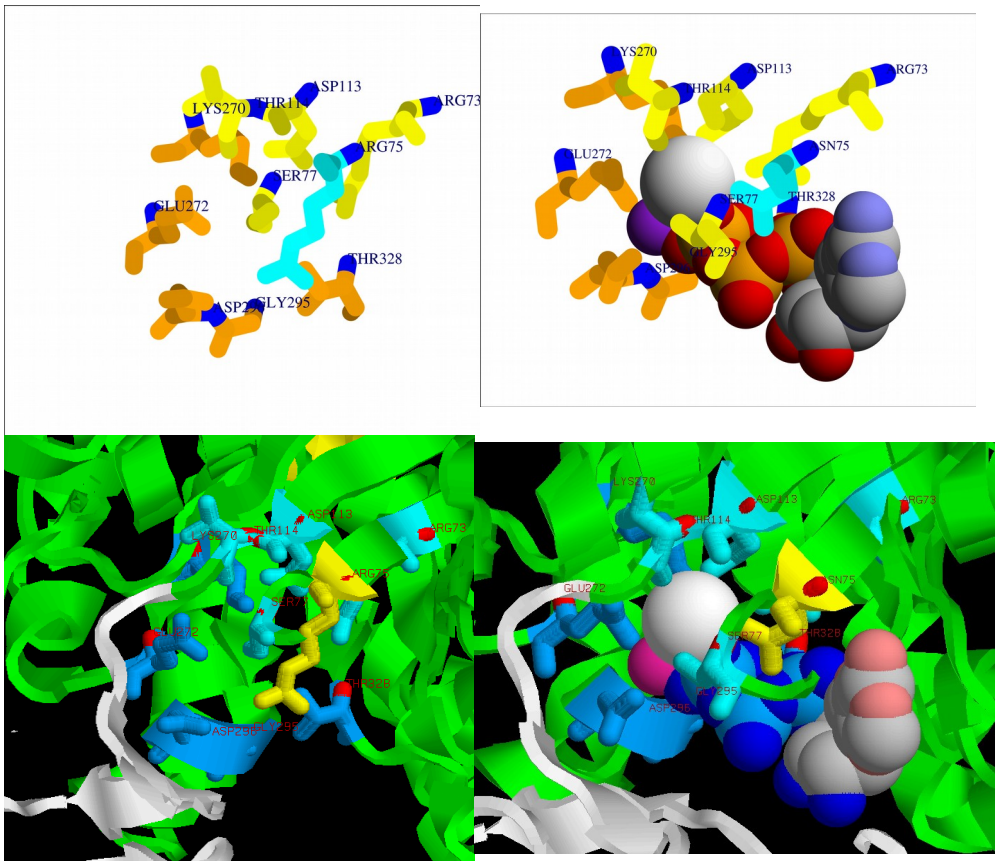


FIgure 3: Left: Model structure for query sequence. Right: 4FXF chain B (human_PKM2) green:d1pkma2, white:d1pkma1

Conclusions about your mystery sequence

Given your 3D model and the location and nature of any differences in the sequence compared with the closest homologue from species X, do you think the mystery sequence is most likely to be a mutant version of the sequence from species X, or a sequence from a related (as yet unsequenced) species? Consider where the differences occur in the structure? If you consider it more likely that this is a mutant protein, do you think the mutations are likely to affect protein function? [20%]

We report that the N75R mutation is located in the active site of PKM2 (FIgure 3), and mutation of 75Asn into a bulky Arg will likely cause abolishment of K+ binding, spatial hindrance towards the binding of ATP, and abolish the catalysis of phosphorylation.

We seacrhed SAS with the query and the resultant alignment showed that N75 is conserved in all homologs with a pdb structure (FIgure 4). Given the conservation at this site, it is unlikely a such change would keep the function of the protein, though this variant has not been reported to cause disease like PKD or PKH (Table 1). The mutation is probably too servere to produce a viable phenotype.

In conclusion, consider that the sequence is highly homologous to cat_PKM2 and carries a loss-of-function(most likely) mutation, we conclude this protein is mostly likely a mutated version of cat_PKM2, and not a homologue from related species.

Although the mutation does not change the overall structure of the protein, it greatly affected the structure of the active site. Although the mutation will abolish the function of the protein, a such mutated gene may fuel further protein evolution given that its function must have been fulfilled by redundant proteins, if we assume the sequence is obtained from a viable animal. Depending on the source of the sequence, the interpretation will be different, and this may be a translation from a pseudogene. With more mutations, the active site may be reshaped to accommodate a different substrate. Or it may serve as a signaling molecule responding to fructose 1,6-biphosphate. Anyway, it is unlikely to execute the normal function of PKM.

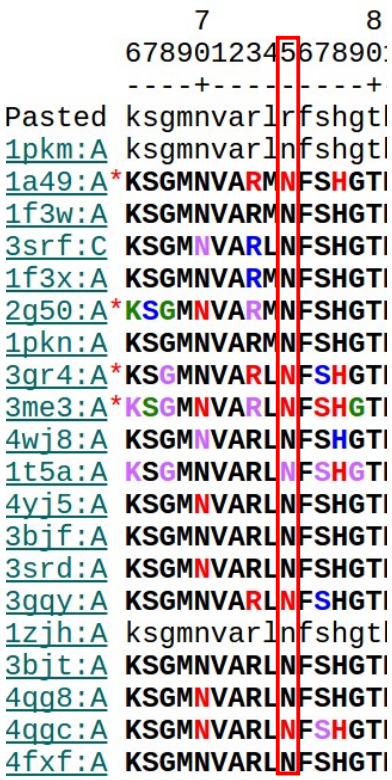


FIgure 4:Sequence alignment from SAS