

9 Lesson (Paralogs)

Use uniprot.org to align the insulin protein sequences: P01325, P01326, P01322, P01323.

- (a) Which pairs of sequences are orthologs and which are paralogs? Explain.
- (b) Look at just the peptide segment of each sequence. (Check the box peptide in the left column.) Did the insulin gene duplicate before or after mouse and rat became separate species? Justify your answer.

66 Example (Google Scholar)

Use Google Scholar to search for important papers on insulin. Which papers have the most citations? Which are the most recent papers.

67 Example (PubMed)

- PubMed is probably the most popular website for searching for research in the life sciences.
- One third of users are people from the general public.
- NIH funded research has an open access mandate.
- You can create an account at NCBI to store your search results.

68 Example (PubMed)

Use PubMed to search for important papers on insulin. Which papers are review papers? Which papers have free PDF's available for download.

69 Example (Medline)

Use MeSH terms to search for articles on insulin. (MeSH terms are manually curated technical terms.)

70 Example (JabRef)

Use JabRef to create a database of papers on insulin.

spI41522) ANS_ANGAN -MALWQAFTLLVLLVLSPPAQ-SASSQHLGCSHLVDALY-MVCGRGFFYQTKRDR
spI987A8) ANS_PANBU -MAFWQAASLLVLLALSPGVD-AAAQHLGCSHLVDALY-LVCGRGFFYTPK-RD
spI04667) ANS_ONCKE
spI68989) ANS_ONCGO
spI68990) ANS_ONCKI
spI01335) ANS_CYPCA -MAVMIQAGALLFLLAVSSVNAH-AGAPQHLGCSHLVDALY-LVCGPTGFYFNPKR-
spI073727) ANS_DAMRE -MAVWLQAGALLVLLVVSSVSTN-PTGPQHLGCSHLVDALY-LVCGPTGFYFNPKR-
spI81881) ANS_P1AME -N-AGAPHLGCSHLVDALY-LVCGPSGFFYFNPK-
spI69045) ANS_LOPAM -MAALWLQSFSLVLLVVSWSGQVAPAQHLGCSHLVDALY-LVCGDRGFYFNPKR-
spI69046) ANS_LOPPI -VAPAQHLGCSHLVDALY-LVCGDRGFYFNPK-
spI81025) ANS_ORENI -MAALWLQAFSLVLLMVMSVWSGQVAGGPHQHLGCSHLVDALY-LVCGDRGFYFNPKR-
spI03399) ANS_THUTH -VAPPQHLGCSHLVDALY-LVCGDRGFYFNPK-
spI09477) ANS_PLAFE -VPPPHQCGAHLVDALY-LVCGERGFFYTPK-
spI9W7R2) ANS_VERMO -MAALWLQSVSLVLLMLVSVWSGQVAPLPHQHLGCAHLVDALY-LVCGERGFFYTPKRR-
spI01336) ANS_GAMDC -MAPPHQHLGCSHLVDALY-LVCGDRGFYFNPK-
spI07453) ANS_MYOSC -ADPHQHLGCSHLVDALY-LVCGDRGFYFNPK-
spI29335) ANS_AMICA -AASQHLGCSHLVEALF-LVCGSGGFFYFNPK-
spI09476) ANS_ATRSP -AANQHLGCSHLVEALY-LVCGXGFFYFNPK-
spI21563) ANS_RODSP -MALWI-LPPLLALLVLLWGPDDPAQFVNQHLGCSHLVEALY-LVCGERGFFYTPKMSRR-
spI91X13) ANS_SPETR -MALWTRLPLPALLALLGDPDPAQFVNQHLGCSHLVEALY-LVCGERGFFYTPKMSRR-
spI01324) ANS_ACCOA -FVBQHLGCSHLVEALY-LVCGERGFFYTPKS-
spI01313) ANS_CRIL0 -NTLWMRLPLPALLVLLWEPNPAQFVNQHLGCSHLVEALY-LVCGERGFFYTPKSRGR-
spI67972) ANS_AOTTR -MALWMLLPLPALLALLWGPPEPAFQFVNQHLGCPHLEALY-LVCGERGFFYTPKTRRE-
spI67971) ANS_SAISC -FVNQHLGCPHLEALY-LVCGERGFFYAPKT-
spI03407) ANS_CHLAE -MALWMRLPLPALLALLWGPDPYPAFVNQHLGCSHLVEALY-LVCGERGFFYTPKTRRE-
spI03406) ANS_MACFA -MALWMRLPLPALLALLWGPDDPAFVNQHLGCSHLVEALY-LVCGERGFFYTPKTRRE-
spI06YK33) ANS_GORGO -MALWMRLPLPALLALLWGPDDAAAFVNQHLGCSHLVEALY-LVCGERGFFYTPKTRRE-
spI01308) ANS_HUMAN -MALWMRLPLPALLALLWGPDDAAAFVNQHLGCSHLVEALY-LVCGERGFFYTPKTRRE-
spI08XV2) ANS_PONPY -MALWMRLPLPALLALLWGPDDPAQFVNQHLGCSHLVEALY-LVCGERGFFYTPKTRRE-
spI034010) ANS_PANTR -MALWMRLPLPALLALLWGPDDASAFVNQHLGCSHLVEALY-LVCGERGFFYTPKTRRE-
spI01317) ANS_BOVIN -MALWTRLPLPALLALLWPPPARAFVNQHLGCSHLVEALY-LVCGERGFFYTPKARRE-
spI01318) ANS_SHEEP -MALWTRLVPLPALLALLWAPAPAHFVNQHLGCSHLVEALY-LVCGERGFFYTPKARRE-
spI01319) ANS_CAPHI -FVNQHLGCSHLVEALY-LVCGERGFFYTPKA-
spI01320) ANS_CAMDR -FANQHLGCSHLVEALY-LVCGERGFFYTPKA-
spI06306) ANS_FELCA -MAPWTRLPLPALLLSLWIPATRAFVNQHLGCSHLVEALY-LVCGERGFFYTPKARRE-
spI01314) ANS_BALBO -FVNQHLGCSHLVEALY-LVCGERGFFYTPKA-
spI67973) ANS_BALPH -FVNQHLGCSHLVEALY-LVCGERGFFYTPKA-
spI01315) ANS_PIG -MALWTRLPLPALLALLWAPAPAAQFVNQHLGCSHLVEALY-LVCGERGFFYTPKARRE-
spI01321) ANS_CANFA -MALWMRLPLPALLALLWAPATRAFVNQHLGCSHLVEALY-LVCGERGFFYTPKARRE-
spI67974) ANS_PHYCD -FVNQHLGCSHLVEALY-LVCGERGFFYTPKA-
spI18109) ANS_DIDVI -LVNQHLGCSHLVEALY-LVCGERGFFYTPKA-
spI01311) ANS_RABIT -MASLAAALPLPALLVLCRLDPAQFVNQHLGCSHLVEALY-LVCGERGFFYTPKMSRR-
spI01316) ANS_ELEMA -FVNQHLGCSHLVEALY-LVCGERGFFYTP-
spI062587) ANS_PSAOB -MALWMRLPLPALLAFLLWEPSPAHAFVNQHLGCSHLVEALY-LVCGERGFFYTPKFRGR-
spI01329) ANS_CAYPO -MALWMHLLTVLALLALLWGPNTOAQFVSRHLCNSNLVETLY-SVQGD0GFFYLPKDRRE-
spI17715) ANS_OCTDE -MAPWMLLTVLALLALLWGPNSTQASVSHQVSLVETLY-MTCRSG-FYRPHDRRE-
spI01330) ANS_MYOCO -YVSGRLGCSQLVDLY-SVCRHRG-FYRPN-
spI01331) ANS_PROGU -YVSGRLGCSQLVDLY-SVCKHRG-FYRPS-
spI01327) ANS_CHICH -FVNKHLGCSHLVDALY-LVCGDRGFYTPMAXHE-
spI01328) ANS_HYSRC -FVNQHLGCSHLVEALY-LVCGNDGFFYTPKA-
spI01310) ANS_HORSE -FVNQHLGCSHLVEALY-LVCGERGFFYTPKAXHE-
spI9QTQ7) ANS_ORNAN -FPNQHLGCSHLVEALY-LVCGRGFFYIIPRM-
spI68991) ANS_CHTMO -VFQTRLGCSHLVDALY-FVCGERGFFYSPKPTR-
spI68992) ANS_HYDCO -VFQTRLGCSHLVDALY-FVCGERGFFYSPKPTR-
spI13190) ANS_CALMI -VFQTRLGCSHLVDALY-FVCGRGFFYSPKQTRD-
spI12704) ANS_SQUAC -LFSQHLGCSHLVEALY-FVCGPKGFYLLPKBZ-
spI12705) ANS_TORMA -LFSQHLGCSHLVEALY-FVCGPKGFYLLPKAXHE-
spI01334) ANS_CROAT -APNQRKLCGSHLEALF-LVCGRGFFYSPR-
spI12708) ANS_PTYDH -APNQRKLCGSHLEALF-LICERGFFYSPR-
spI12703) ANS_ALMI -AANQRKLCGSHLDALY-LVCGERGFFYSPKG-
spI67970) ANS_CHICK -MALWIRSLPALLVLLVFSGPGTSSAAANQHLGCSHLVEALY-LVCGERGFFYSPKARRD-
spI67968) ANS_MELGA -AANQHLGCSHLVEALY-LVCGERGFFYSPKA-
spI51463) ANS_SELRF -IQSLPALLALLSGPTSHAANQHLGCSHLVEALY-LVCGERGFFYSPKARRD-
spI67969) ANS_STRCA -AANQHLGCSHLVEALY-LVCGERGFFYSPKA-
spI69047) ANS_TRADO -AANQHLGCSHLVEALY-LVCGERGFFYSPKA-
spI69048) ANS_TRASC -AANQHLGCSHLVEALY-LVCGERGFFYSPKA-
spI01333) ANS_ANAPL -AANQHLGCSHLVEALY-LVCGERGFFYSPKTXD-
spI68245) ANS_ANSAN -AANQHLGCSHLVEALY-LVCGERGFFYSPKT-
spI68243) ANS_CAIMO -AANQHLGCSHLVEALY-LVCGERGFFYSPKT-
spI81423) ANS_ACTGU -AANQHLGCSHLVEALY-LVCGERGFFYTPNK-
spI0C236) ANS_POLSE -AANRHLGCSHLVEALY-LVCGNRGFFYIPSK-
spI01340) ANS_KATPE -AANPHRGCSHLVEALY-LVCGNRGFFYQPK-
spI68988) ANS_LAMFL -SALTGAGGTHLCGSHLEALY-VVCGDRGFYTPSK-
spI68987) ANS_PETMA -SALTGAGGTHLCGSHLEALY-VVCGDRGFYTPSK-
spI42633) ANS_ANGRO -ASTQHLGCSHLVEALY-LVCGSGNFFNKD-
spI01342) ANS_MYXGL -MALSPFLAAVIPLVLLLSRAPPSADTRTGHLCOKLDVNLALY-IACGVRGFFYTPKTRK

Multiple sequence alignments can provide a lot of information about proteins. Consider the multiple sequence alignment of 76 insulin sequences shown below. Observe that there are two segments that are well conserved.

```

sp|P41522|INS_ANGAN
sp|Q98T81|INS_PANBU
sp|P04667|INS_ONCKE
sp|P68989|INS_ONCGO
sp|P68990|INS_ONCKI
sp|P01335|INS_CYPCA
sp|U073727|INS_DANRE
sp|P81881|INS_PIAME
sp|P69045|INS_LOPAM
sp|P69046|INS_LOPPI
sp|P81025|INS_ORENI
sp|P01339|INS_THUTH
sp|P09477|INS_PLAFE
sp|Q9W7R2|INS_VERMO
sp|P01336|INS_GADMC
sp|P07453|INS_MYOSC
sp|P29335|INS_AMICA
sp|P09476|INS_ATRSP
sp|P21563|INS_RODSP
sp|Q91X13|INS_SPETR
sp|P01324|INS_ACOCA
sp|P01313|INS_CRILLO
sp|P67972|INS_AOTTR
sp|P67971|INS_SAISC
sp|P30407|INS_CHLAE
sp|P30406|INS_MACFA
sp|Q6YK33|INS_GORGO
sp|P01308|INS_HUMAN
sp|Q8HXV2|INS_PONPY
sp|P30410|INS_PANTR
sp|P01317|INS_BOVIN
sp|P01318|INS_SHEEP
sp|P01319|INS_CAPHI
sp|P01320|INS_CAMDR
sp|P06306|INS_FELCA
sp|P01314|INS_BALBO
sp|P67973|INS_BALPH
sp|P01315|INS_PIG
sp|P01321|INS_CANFA
sp|P67974|INS_PHYCD
sp|P18109|INS_DIDVI
sp|P01311|INS_RABIT
sp|P01316|INS_ELEMA
sp|Q62587|INS_PSAOB
sp|P01329|INS_CAVPO
sp|P17715|INS_OCTDE
sp|P01330|INS_MYOCO
sp|P01331|INS_PROGU
sp|P01327|INS_CHICH
sp|P01328|INS_HYSCR
sp|P01310|INS_HORSE
sp|Q9TQ77|INS_ORNAN
sp|P68991|INS_CHIMO
sp|P68992|INS_HYDCO
sp|P13190|INS_CALMI
sp|P12704|INS_SQUAC
sp|P12705|INS_TORMA
sp|P01334|INS_CROAT
sp|P12708|INS_PTYDH
sp|P12703|INS_ALLMI
sp|P67970|INS_CHICK
sp|P67968|INS_MELGA
sp|P51463|INS_SELRF
sp|P67969|INS_STRCA
sp|P69047|INS_TRADO
sp|P69048|INS_TRASC
sp|P01333|INS_ANAPL
sp|P68245|INS_ANSAN
sp|P68243|INS_CAIMO
sp|P81423|INS_ACTGU
sp|P0C236|INS_POLSE
sp|P01340|INS_KATPE
sp|P68988|INS_LAMFL
sp|P68987|INS_PETMA
sp|P42633|INS_ANGRO
sp|P01342|INS_MYGGL
: * * **

```

```

VEPLLQ-FLSPKSAQ-----ENEVDFFPYKQQLXX--GIVEQCCHKPCNIFDLQNYC
VDPLLQ-FLSPKSAQ-----ENEADYYPKQDGLKVRGIVEQCCHPCNIFDLQNYC
VDPLIG-FLSPKSAK-----EN--EYFPKQTEMVMKRGIVEQCCHKPCNIFDLQNYC
AQHLGCSHLVDALYL-----VCEKGFFYNPK-----GIVEQCCHKPCNIFDLQNYC
AQHLGCSHLVDALYL-----VCEKGFFYNPK-----GIVEQCCHKPCNIFDLQNYC
VDPLLQ-FLPPKSAQ-----ETEVADFADKDHAEVIRKRGIVEQCCHKPCNIFDLQNYC
VEPLLQ-FLPPKSAQ-----ETEVADFADKDHAEVIRKRGIVEQCCHKPCNIFDLQNYC
-----GIVEQCCHKPCNIFDLQNYC
VDQLLQ-FLPPKSGAAAGADNEVAEAFKQDQEMVMKRGIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
VDPLLQ-FLPPKAGGAVVQGENEVT--FKDQEMVMKRGIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
VDPLLQ-FLPAKSGGAA--GGENEVAEAFKQDQEMVMKRGIVEQCCHPCNIFDLQNYC
-----GIVDQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
-----SGIVEQCCHPCNIFDLQNYC
-----VGIVEQCCHPCNIFDLQNYC
VEDPQVQVQLGAGPGAGS-----EQTLAEVARQAR--IVQQTSGICSLYQ--ENYC
VEEQGGQVQLGGPGAGL-----PQPLALEMALKRGIVEQCCTSI CSLYQLENYC
-----GIVDQCCTSI CSLYQLENYC
VEDPQVQVQLGAGPGAGD-----LQTLAEVAQKRGIVDQCCTSI CSLYQLENYC
AEDLQVQVQLGGGSI TGS-----LPPL--EGPMKRGVVDQCCTSI CSLYQLENYC
-----GIVDQCCTSI CSLYQLENYC
AEDPQVQVQLGGPGAGS-----LQPLALEGSLKRGIVEQCCTSI CSLYQLENYC
AEDPQVQVQLGGPGAGS-----LQPLALEGSLKRGIVEQCCTSI CSLYQLENYC
AEDLQVQVQLGGPGAGS-----LQPLALEGSLKRGIVEQCCTSI CSLYQLENYC
AEDLQVQVQLGGPGAGS-----LQPLALEGSLKRGIVEQCCTSI CSLYQLENYC
AEDLQVQVQLGGPGAGS-----LQPLALEGSLKRGIVEQCCTSI CSLYQLENYC
VEGPQVQVQLGGPGAGG-----L-----EGPPKRGIVEQCASVCSLYQLENYC
VEGPQVQVQLGGPGAGG-----L-----EGPPKRGIVEQCASVCSLYQLENYC
-----GIVEQCASVCSLYQLENYC
AEDLQKDAELGEAPGAGG-----LQPSALEAPLQKRGIVEQCASVCSLYQLENYC
-----GIVEQCASVCSLYQLENYC
AENPQAGAVELGG--GLGG-----LQALALEGPPKRGIVEQCCTSI CSLYQLENYC
VEDLQVQVQLGAGPGAGG-----LQPLALEGSLKRGIVEQCCTSI CSLYQLENYC
-----GIVEQCCTSI CSLYQLENYC
VEELQVQVQLGAGPGAGG-----LQPSALELALQKRGIVEQCCTSI CSLYQLENYC
-----KTGIVEQCCTSI CSLYQLENYC
VDDPQVQVQLGGPGAGD-----LQALALEVARQKRGIVEQCCTSI CSLYQLENYC
LEDPQVQVQLGAGPGAGG-----LQPLALEMALKRGIVEQCCTSI CSLYQLENYC
LEDLQVQVQLGAGPGAGG-----LQPSALEMILKRGIVEQCCTSI CSLYQLENYC
-----GIVDQCCTSI CSLYQLENYC
LEDPQVQVQADPGVPEAGR-----LQPLALEMTLQXGIVDQCCTSI CSLYQLENYC
-----GIVDQCCTSI CSLYQLENYC
AEDPQVQVQLGAGPGAGG-----LQPLALAGPQKRGIVEQCCTSI CSLYQLENYC
-----GIVEQCCHPCNIFDLQNYC
-----ELEPLL-----GIVEQCCHPCNIFDLQNYC
VGPLSAPRDLPEPLDTEMED-----RFPYRQLAGSKMRGIVEQCCHPCNIFDLQNYC
-----VGIVEQCCHPCNIFDLQNYC
VDSLQVQVQLGAGPGAGG-----LQPLALEGSLKRGIVEQCCTSI CSLYQLENYC
-----SGIVEQCCHPCNIFDLQNYC
-----TGIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
VEQPLVSSPLRGEAGV-----LPFQEEYKVKRGIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
AEHPLVNGPLHGEVGD-----LPFQEEFEKVKRGIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
VEQPLVNGPLHGEVGE-----LPFQHEEYQXGIVDQCCTSI CSLYQLENYC
-----GIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
-----VGIVEQCCHPCNIFDLQNYC
-----MGIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
-----TGIVEQCCHPCNIFDLQNYC
-----GIVEQCCHPCNIFDLQNYC
DTGALAAFLPLAYAEDNESQ-----DDESIGINEVLKSRGIVEQCCHPCNIFDLQNYC

```



Figure 1: Sequence Logo of 11 aligned insulin sequences.

content of the corresponding column. Well conserved columns of letters have high information content. The color of the letters represents properties of the corresponding amino acids.

Once insulin has been translated from DNA into a protein sequence, it must be processed further before it can function as insulin. The unprocessed form of insulin is called *pre-proinsulin*.¹

Insulin (Pre-proinsulin) Schematic Diagram

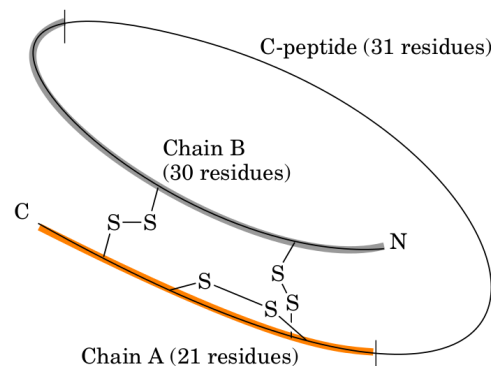


Figure 2: The post translational modification of pre-proinsulin.

71 Example (Sequence Logos)

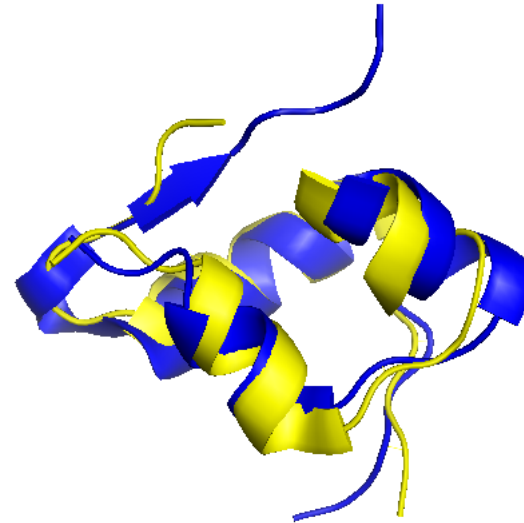
Sequence Logos provide a summary of a multiple sequence alignments. The heights of each letter represents how frequently the letter occurs in the corresponding alignment column. The stacked heights of the letters represents the *information*

First, the signal peptide tells the cell where insulin is to be transported. Then the signal peptide is removed. At this point the insulin is called *proinsulin*. The protein

¹Mark Brandt, Chemistry and Biochemistry Dept., Rose-Hulman Institute of Technology

then folds with the help of the central C-peptide section. Three disulfied bonds form, two of which hold chain A and chain B together. The third disulfied bond is within chain A. Finally the C-peptide is cut and removed. The final insulin product is only 51 amino acids long.

75 Example (Human vs Fruitfly Insulin Structure)



72 Example (pFAM)

Proteins can be organized into sequence families. pFAM is a database of over 14,800 sequence families. Multiple sequence alignment of a family can be used to identify *protein domains*, i.e. important segments of a sequence.

73 Example (pFAM)

Go to www.pfam.org and search for P01308.

- Examine the domain structure of human insulin.
 - Click on the insulin domain. Examine HMM Logo. What does the logo tell us about post-translational modifications of insulin.
-

10 Lesson (Phylogenetic Analysis)

Use uniprot.org to construct a phylogenetic tree for a group of several insulin sequences.

Go to uniprot.org

type `gene:INS`

select five interesting insulin sequences by clicking on the check box

align the sequences

select Tree under display

Discuss your results.

74 Example (Human vs Fruitfly Insulin Sequence)

Use www.uniprot.org to align *Drosophila melanogaster* insulin-like peptide 5 (DILP5) with human insulin. How similar are the sequences? Ans. 27.8% identity.

Even though the human and fruitfly insulin sequences are very different, their protein structures are very similar. Shown above is an alignment of the structure of human insulin (blue—PDB code 1MSO) with fruitfly insulin (yellow—PDB code 2WFU).

76 Example (Human vs Fruitfly Insulin Function)

Even though the human and fruitfly insulin sequences are very different, their protein functions have been shown experimentally to be very similar. For example, fruitfly insulin has been demonstrated to reduce glucose in rats.

The main point is that proteins with very different sequences can still have very similar structure and function.