Meenu Ravi
R Skills: PNNL Write Up
November 25 2020

Part 1: Question 1 and 2
Part 2-9: Question 3 (run consecutively)
Part 10-11: Question 4 (run consecutively)

Part 1

This file contains:

- my installation and formatting initial files given
  - Unzipped both given files
  - Read phosphopeptides into a table
- Map RefSeq IDs to gene IDs
  - Use biomart to get gene ids
  - Added that to phosphopeptides table
  - Wrote result into a text file called phosphogene.txt

Part 2

This file contains:

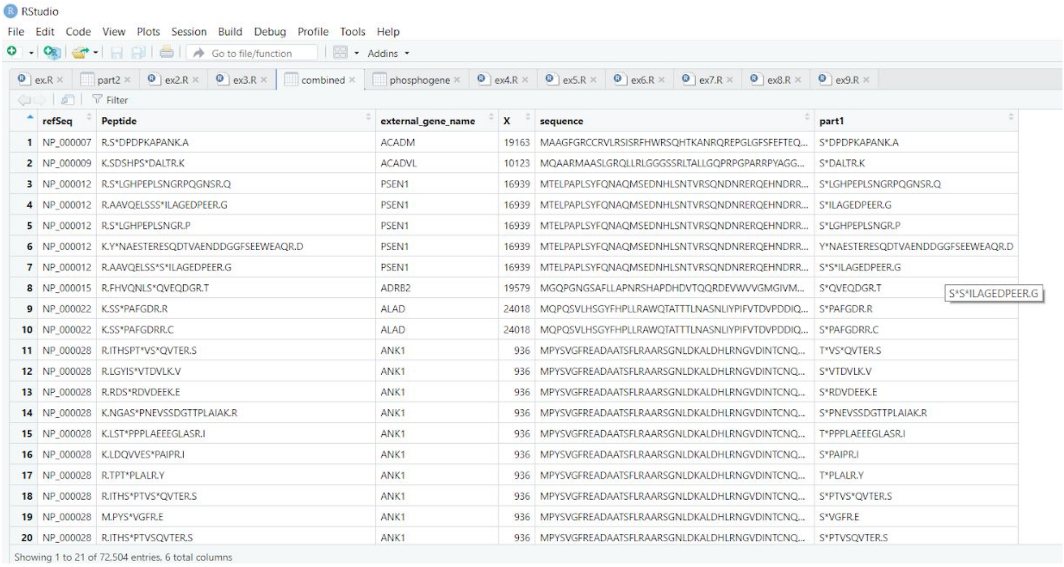
- Convert sapiens fasta file to a table by reading it in as a csv first
- Then remove ‘>’
- Save result as formattedProteins.txt

Part 3

This file contains:

- Merged dataset containing the refseq, peptide, gene id and associated protein sequence
- The substring for all peptides from right before the first *

- Saved resulting table in proteinp1.txt



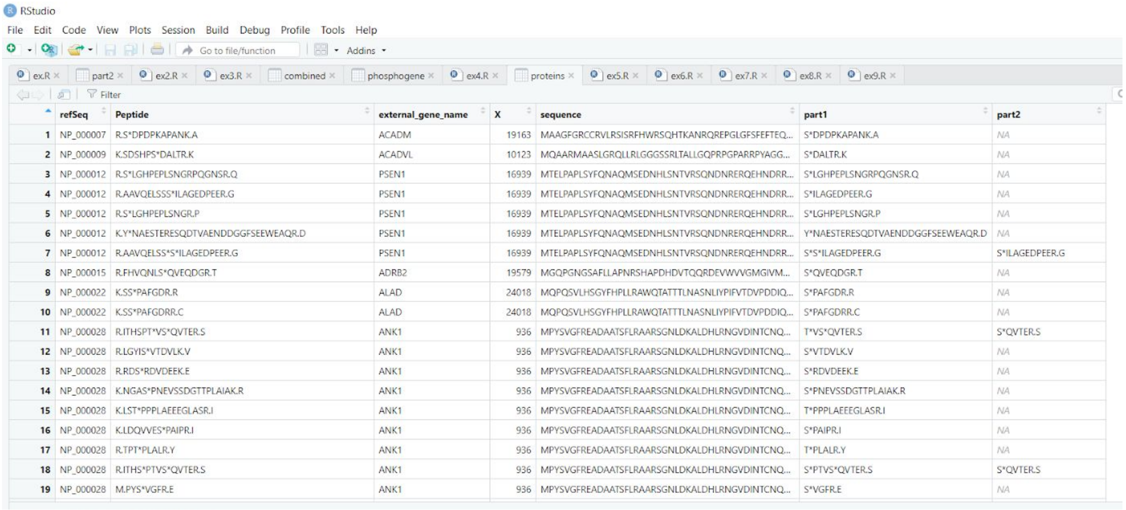
	refSeq	Peptide	external_gene_name	X	sequence	part1
1	NP_000007	R.S*DPDPKAPANK.A	ACADM	19163	MAAGFGRCCRVLRSISRFHWRSQHTKANRQREPGLGFSFEFTEQ...	S*DPDPKAPANK.A
2	NP_000009	K.SDSHPS*DALTR.K	ACADVL	10123	MQAARMAASLGRQLLRLGGGSSRLTALLGQPRPGPARRPYAGG...	S*DALTR.K
3	NP_000012	R.S*LGHPEPLSNGRPQGNSR.Q	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGRPQGNSR.Q
4	NP_000012	R.AAVQELSSS*ILAGEDPEER.G	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*ILAGEDPEER.G
5	NP_000012	R.S*LGHPEPLSNGR.P	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGR.P
6	NP_000012	K.Y*NAESTERESQDTVAENDDGGFSEEWEAQR.D	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	Y*NAESTERESQDTVAENDDGGFSEEWEAQR.D
7	NP_000012	R.AAVQELSS*S*ILAGEDPEER.G	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*S*ILAGEDPEER.G
8	NP_000015	R.FHVQNLS*QVEQDGR.T	ADRB2	19579	MGQPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVM...	S*QVEQDGR.T
9	NP_000022	K.SS*PAFGDR.R	ALAD	24018	MQPQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDR.R
10	NP_000022	K.SS*PAFGDRR.C	ALAD	24018	MQPQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDRR.C
11	NP_000028	R.ITHSPT*VS*QVTER.S	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*VS*QVTER.S
12	NP_000028	R.LGYIS*VTDVLK.V	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*VTDVLK.V
13	NP_000028	R.RDS*RDVDEEK.E	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*RDVDEEK.E
14	NP_000028	K.NGAS*PNEVSSDGTTPLAIAK.R	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PNEVSSDGTTPLAIAK.R
15	NP_000028	K.LST*PPPLAEEEGLASR.I	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PPPLAEEEGLASR.I
16	NP_000028	K.LDQVVES*PAIPR.I	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PAIPR.I
17	NP_000028	R.TPT*PLALR.Y	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PLALR.Y
18	NP_000028	R.ITHS*PTVS*QVTER.S	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PTVS*QVTER.S
19	NP_000028	M.PYS*VGFR.E	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*VGFR.E
20	NP_000028	R.ITHS*PTVSQVTER.S	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PTVSQVTER.S

Showing 1 to 21 of 72,504 entries, 6 total columns

Part 4

This file contains:

- The substring for all peptides with 2 *
- Saved this table as proteinsp2.txt

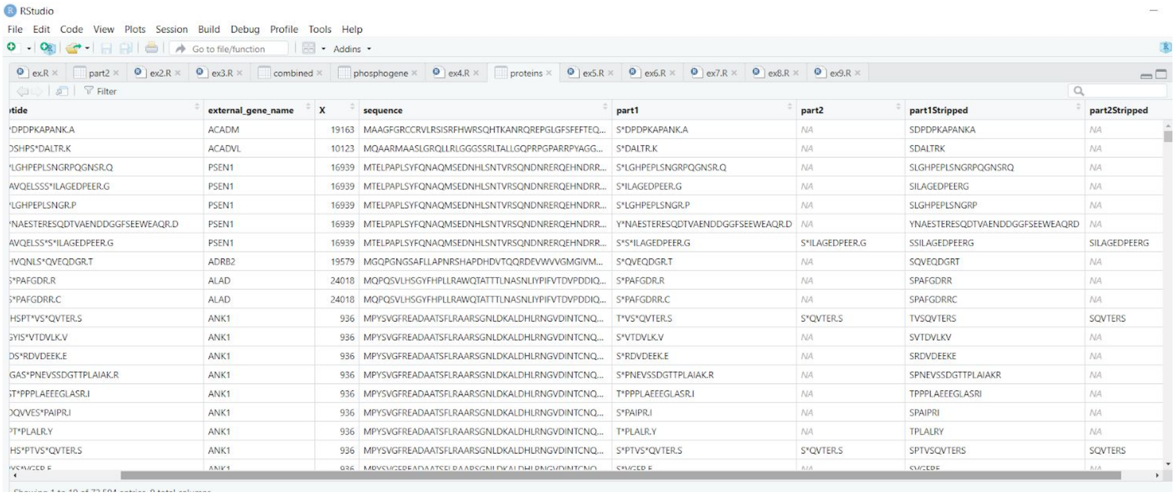


	refSeq	Peptide	external_gene_name	X	sequence	part1	part2
1	NP_000007	R.S*DPDPKAPANK.A	ACADM	19163	MAAGFGRCCRVLRSISRFHWRSQHTKANRQREPGLGFSFEFTEQ...	S*DPDPKAPANK.A	NA
2	NP_000009	K.SDSHPS*DALTR.K	ACADVL	10123	MQAARMAASLGRQLLRLGGGSSRLTALLGQPRPGPARRPYAGG...	S*DALTR.K	NA
3	NP_000012	R.S*LGHPEPLSNGRPQGNSR.Q	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGRPQGNSR.Q	NA
4	NP_000012	R.AAVQELSSS*ILAGEDPEER.G	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*ILAGEDPEER.G	NA
5	NP_000012	R.S*LGHPEPLSNGR.P	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGR.P	NA
6	NP_000012	K.Y*NAESTERESQDTVAENDDGGFSEEWEAQR.D	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	Y*NAESTERESQDTVAENDDGGFSEEWEAQR.D	NA
7	NP_000012	R.AAVQELSS*S*ILAGEDPEER.G	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*S*ILAGEDPEER.G	S*ILAGEDPEER.G
8	NP_000015	R.FHVQNLS*QVEQDGR.T	ADRB2	19579	MGQPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVM...	S*QVEQDGR.T	NA
9	NP_000022	K.SS*PAFGDR.R	ALAD	24018	MQPQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDR.R	NA
10	NP_000022	K.SS*PAFGDRR.C	ALAD	24018	MQPQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDRR.C	NA
11	NP_000028	R.ITHSPT*VS*QVTER.S	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*VS*QVTER.S	S*QVTER.S
12	NP_000028	R.LGYIS*VTDVLK.V	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*VTDVLK.V	NA
13	NP_000028	R.RDS*RDVDEEK.E	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*RDVDEEK.E	NA
14	NP_000028	K.NGAS*PNEVSSDGTTPLAIAK.R	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PNEVSSDGTTPLAIAK.R	NA
15	NP_000028	K.LST*PPPLAEEEGLASR.I	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PPPLAEEEGLASR.I	NA
16	NP_000028	K.LDQVVES*PAIPR.I	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PAIPR.I	NA
17	NP_000028	R.TPT*PLALR.Y	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PLALR.Y	NA
18	NP_000028	R.ITHS*PTVS*QVTER.S	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PTVS*QVTER.S	S*QVTER.S
19	NP_000028	M.PYS*VGFR.E	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*VGFR.E	NA

Part 5

This file contains:

- From the part 1 and part 2 that we got, I remove the non alphanumeric characters: * and . and any -
- This will be saved as proteinsStripped.txt

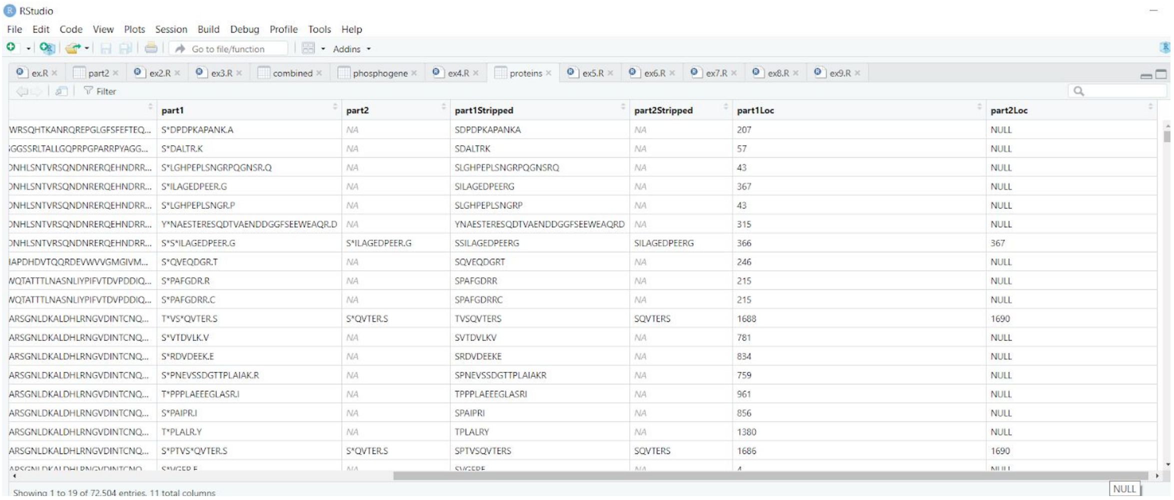


tide	external_gene_name	X	sequence	part1	part2	part1Stripped	part2Stripped
*DPDPKAPANK.A	ACADM	19163	MAAGFGRCCRVLRSISRFHWRSQHTKANRQREPGLGFSFEFTEQ...	S*DPDPKAPANK.A	NA	SDPDPKAPANKA	NA
DSHPS*DALTR.K	ACADVL	10123	MQAARMAASLGRQLLRLGGGSSRLTALLGQPRPGPARRPYAGG...	S*DALTR.K	NA	SDALTRK	NA
*LGHPEPLSNGRPQGNSR.Q	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGRPQGNSR.Q	NA	SLGHPEPLSNGRPQGNSRQ	NA
AVQELSSS*ILAGEDPEER.G	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*ILAGEDPEER.G	NA	SILAGEDPEERG	NA
*LGHPEPLSNGR.P	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGR.P	NA	SLGHPEPLSNGRP	NA
*NAESTERESQDTVAENDDGGFSEEWEAQR.D	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	Y*NAESTERESQDTVAENDDGGFSEEWEAQR.D	NA	YNAESTERESQDTVAENDDGGFSEEWEAQRD	NA
AVQELSS*S*ILAGEDPEER.G	PSEN1	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*S*ILAGEDPEER.G	S*ILAGEDPEER.G	SSILAGEDPEERG	SILAGEDPEERG
HVQNLS*QVEQDGR.T	ADRB2	19579	MGQPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVM...	S*QVEQDGR.T	NA	SQVEQDGRT	NA
S*PAFGDR.R	ALAD	24018	MQPQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDR.R	NA	SPAFGDRR	NA
S*PAFGDRR.C	ALAD	24018	MQPQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDRR.C	NA	SPAFGDRRC	NA
HSPT*VS*QVTER.S	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*VS*QVTER.S	S*QVTER.S	TVSQVTERS	SQVTERS
GYIS*VTDVLK.V	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*VTDVLK.V	NA	SVTDVLKV	NA
DS*RDVDEEK.E	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*RDVDEEK.E	NA	SRDVDEEKE	NA
GAS*PNEVSSDGTTPLAIAK.R	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PNEVSSDGTTPLAIAK.R	NA	SPNEVSSDGTTPLAIAKR	NA
ST*PPPLAEEEGLASR.I	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PPPLAEEEGLASR.I	NA	TPPPLAEEEGLASRI	NA
DQVVES*PAIPR.I	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PAIPR.I	NA	SPAIPRI	NA
PT*PLALR.Y	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PLALR.Y	NA	TPLALRY	NA
HS*PTVS*QVTER.S	ANK1	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PTVS*QVTER.S	S*QVTER.S	SPTVSQVTERS	SQVTERS

Part 6

This file contains:

- Using the part1Stripped and part2Stripped columns that we just got, I get the amino acid position within the protein sequence
- This is saved as proteinsPositions.txt



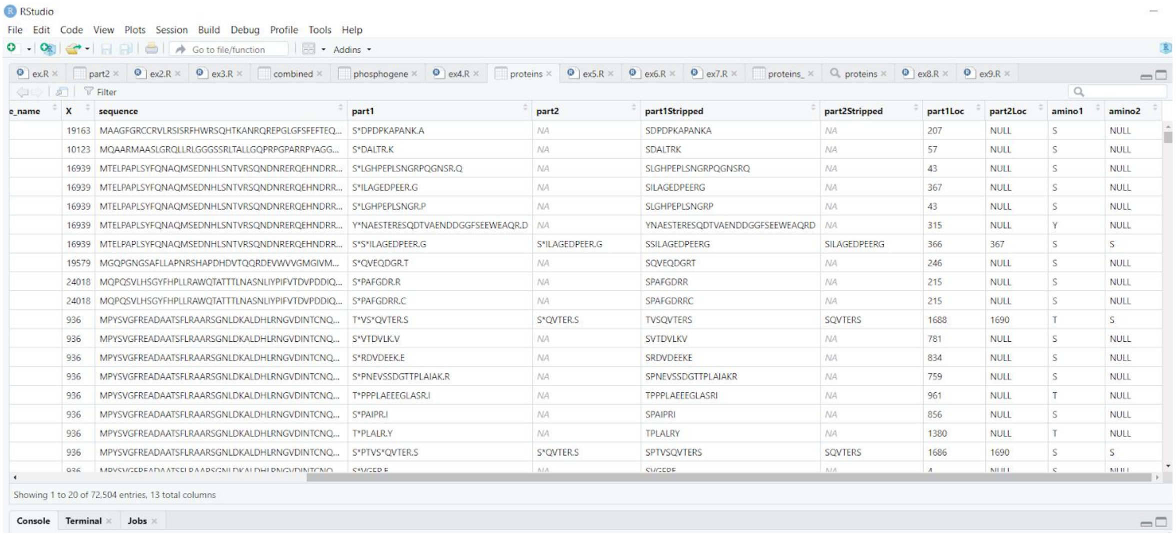
	part1	part2	part1Stripped	part2Stripped	part1Loc	part2Loc
WRSQHTKANRQREPGLGFSFEFTEQ...	S*DPDPKAPANK.A	NA	SDPDPKAPANKA	NA	207	NULL
GGSSRLTALLGQPRPGPARRPYAGG...	S*DALTR.K	NA	SDALTRK	NA	57	NULL
DNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGRPQGNSR.Q	NA	SLGHPEPLSNGRPQGNSRQ	NA	43	NULL
DNHLSNTVRSQNDNRERQEHNDRR...	S*ILAGEDPEER.G	NA	SILAGEDPEERG	NA	367	NULL
DNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGR.P	NA	SLGHPEPLSNGRP	NA	43	NULL
DNHLSNTVRSQNDNRERQEHNDRR...	Y*NAESTERESQDTVAENDDGGFSEEWEAQR.D	NA	YNAESTERESQDTVAENDDGGFSEEWEAQRD	NA	315	NULL
DNHLSNTVRSQNDNRERQEHNDRR...	S*S*ILAGEDPEER.G	S*ILAGEDPEER.G	SSILAGEDPEERG	SILAGEDPEERG	366	367
IAPDHDVTQQRDEVWVVGMGIVM...	S*QVEQDGR.T	NA	SQVEQDGRT	NA	246	NULL
WQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDR.R	NA	SPAFGDRR	NA	215	NULL
WQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDRR.C	NA	SPAFGDRRC	NA	215	NULL
ARSGNLDKALDHLRNGVDINTCNQ...	T*VS*QVTER.S	S*QVTER.S	TVSQVTERS	SQVTERS	1688	1690
ARSGNLDKALDHLRNGVDINTCNQ...	S*VTDVLK.V	NA	SVTDVLKV	NA	781	NULL
ARSGNLDKALDHLRNGVDINTCNQ...	S*RDVDEEK.E	NA	SRDVDEEKE	NA	834	NULL
ARSGNLDKALDHLRNGVDINTCNQ...	S*PNEVSSDGTTPLAIAK.R	NA	SPNEVSSDGTTPLAIAKR	NA	759	NULL
ARSGNLDKALDHLRNGVDINTCNQ...	T*PPPLAEEEGLASR.I	NA	TPPPLAEEEGLASRI	NA	961	NULL
ARSGNLDKALDHLRNGVDINTCNQ...	S*PAIPR.I	NA	SPAIPRI	NA	856	NULL
ARSGNLDKALDHLRNGVDINTCNQ...	T*PLALR.Y	NA	TPLALRY	NA	1380	NULL
ARSGNLDKALDHLRNGVDINTCNQ...	S*PTVS*QVTER.S	S*QVTER.S	SPTVSQVTERS	SQVTERS	1686	1690

Showing 1 to 19 of 72,504 entries, 11 total columns

Part 7

This file contains:

- I get the amino acid S,T, or Y and save it in amino 1 for the first one and amino2 for the second position
- This table is saved in proteinsAminos.txt



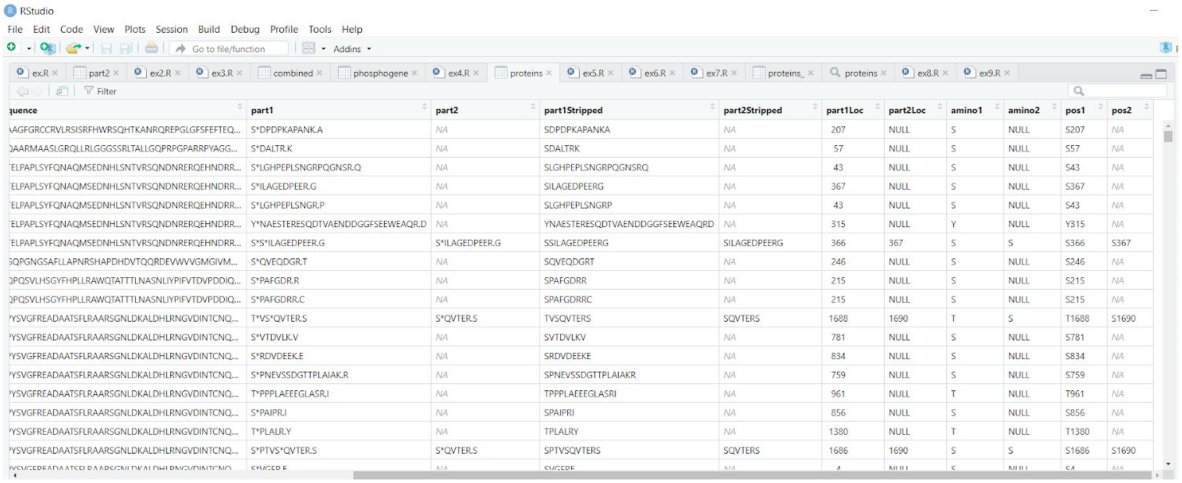
e_name	X	sequence	part1	part2	part1Stripped	part2Stripped	part1Loc	part2Loc	amino1	amino2
	19163	MAAGFGRCCRVLRSISRFHWRSQHTKANRQREPGLGFSFEFTEQ...	S*DPDPKAPANK.A	NA	SDPDPKAPANKA	NA	207	NULL	S	NULL
	10123	MQAARMAASLGRQLLRLGGGSSRLTALLGQPRPGPARRPYAGG...	S*DALTR.K	NA	SDALTRK	NA	57	NULL	S	NULL
	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGRPQGNSR.Q	NA	SLGHPEPLSNGRPQGNSRQ	NA	43	NULL	S	NULL
	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*ILAGEDPEER.G	NA	SILAGEDPEERG	NA	367	NULL	S	NULL
	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGR.P	NA	SLGHPEPLSNGRP	NA	43	NULL	S	NULL
	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	Y*NAESTERESQDTVAENDDGGFSEEWEAQR.D	NA	YNAESTERESQDTVAENDDGGFSEEWEAQRD	NA	315	NULL	Y	NULL
	16939	MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*S*ILAGEDPEER.G	S*ILAGEDPEER.G	SSILAGEDPEERG	SILAGEDPEERG	366	367	S	S
	19579	MGQPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVM...	S*QVEQDGR.T	NA	SQVEQDGRT	NA	246	NULL	S	NULL
	24018	MQPQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDR.R	NA	SPAFGDRR	NA	215	NULL	S	NULL
	24018	MQPQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDRR.C	NA	SPAFGDRRC	NA	215	NULL	S	NULL
	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*VS*QVTER.S	S*QVTER.S	TVSQVTERS	SQVTERS	1688	1690	T	S
	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*VTDVLK.V	NA	SVTDVLKV	NA	781	NULL	S	NULL
	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*RDVDEEK.E	NA	SRDVDEEKE	NA	834	NULL	S	NULL
	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PNEVSSDGTTPLAIAK.R	NA	SPNEVSSDGTTPLAIAKR	NA	759	NULL	S	NULL
	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PPPLAEEEGLASR.I	NA	TPPPLAEEEGLASRI	NA	961	NULL	T	NULL
	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PAIPR.I	NA	SPAIPRI	NA	856	NULL	S	NULL
	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PLALR.Y	NA	TPLALRY	NA	1380	NULL	T	NULL
	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PTVS*QVTER.S	S*QVTER.S	SPTVSQVTERS	SQVTERS	1686	1690	S	S
	936	MPYSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ	S*VGEP.E	NA	SVGEPE	NA	4	NULL	S	NULL

Showing 1 to 20 of 72,504 entries, 13 total columns

Part 8

This file contains:

- The amino acid (S, T or Y) and amino acid position within the protein sequence as a single string
- This table is saved in proteinssites.txt

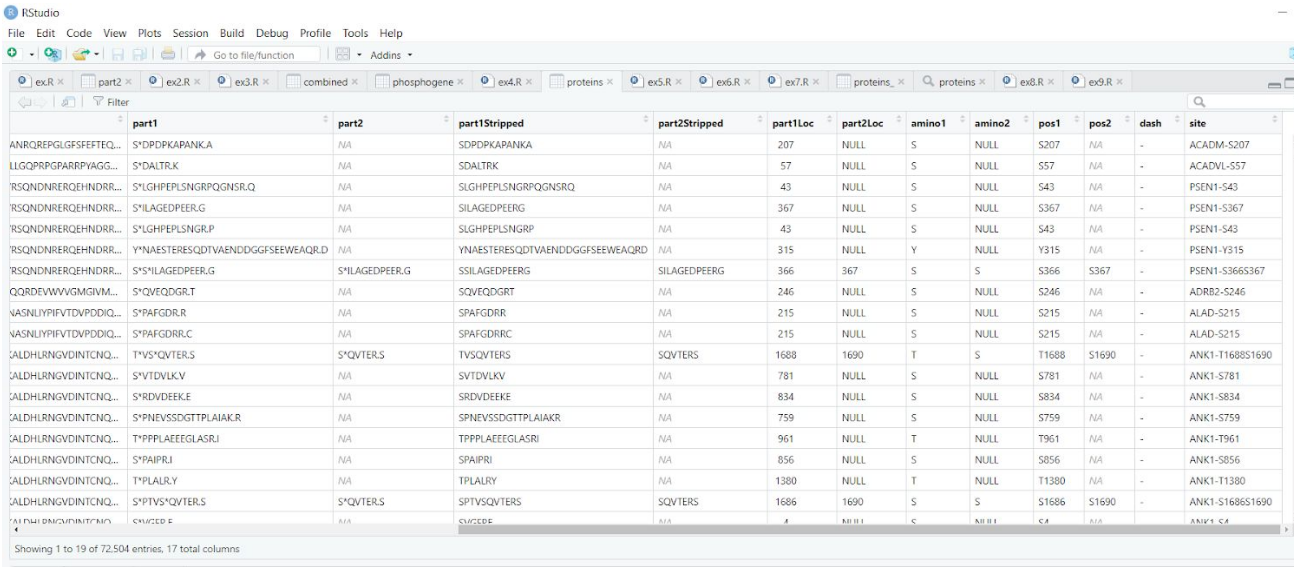


quence	part1	part2	part1Stripped	part2Stripped	part1Loc	part2Loc	amino1	amino2	pos1	pos2
AGFGRCCRVLRSISRFHWRSQHTKANRQREPGLGFSFEFTEQ...	S*DPDPKAPANK.A	NA	SDPDPKAPANKA	NA	207	NULL	S	NULL	S207	NA
QAARMAASLGRQLLRLGGGSSRLTALLGQPRPGPARRPYAGG...	S*DALTR.K	NA	SDALTRK	NA	57	NULL	S	NULL	S57	NA
ELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGRPQGNSR.Q	NA	SLGHPEPLSNGRPQGNSRQ	NA	43	NULL	S	NULL	S43	NA
ELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*ILAGEDPEER.G	NA	SILAGEDPEERG	NA	367	NULL	S	NULL	S367	NA
ELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*LGHPEPLSNGR.P	NA	SLGHPEPLSNGRP	NA	43	NULL	S	NULL	S43	NA
ELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	Y*NAESTERESQDTVAENDDGGFSEEWEAQR.D	NA	YNAESTERESQDTVAENDDGGFSEEWEAQRD	NA	315	NULL	Y	NULL	Y315	NA
ELPAPLSYFQNAQMSEDNHLSNTVRSQNDNRERQEHNDRR...	S*S*ILAGEDPEER.G	S*ILAGEDPEER.G	SSILAGEDPEERG	SILAGEDPEERG	366	367	S	S	S366	S367
QPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVM...	S*QVEQDGR.T	NA	SQVEQDGRT	NA	246	NULL	S	NULL	S246	NA
PQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDR.R	NA	SPAFGDRR	NA	215	NULL	S	NULL	S215	NA
PQSVLHSGYFHPLLRAWQTATTTLNASNLIYPIFVTDVPDDIQ...	S*PAFGDRR.C	NA	SPAFGDRRC	NA	215	NULL	S	NULL	S215	NA
YSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*VS*QVTER.S	S*QVTER.S	TVSQVTERS	SQVTERS	1688	1690	T	S	T1688	S1690
YSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*VTDVLK.V	NA	SVTDVLKV	NA	781	NULL	S	NULL	S781	NA
YSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*RDVDEEK.E	NA	SRDVDEEKE	NA	834	NULL	S	NULL	S834	NA
YSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PNEVSSDGTTPLAIAK.R	NA	SPNEVSSDGTTPLAIAKR	NA	759	NULL	S	NULL	S759	NA
YSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PPPLAEEEGLASR.I	NA	TPPPLAEEEGLASRI	NA	961	NULL	T	NULL	T961	NA
YSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PAIPR.I	NA	SPAIPRI	NA	856	NULL	S	NULL	S856	NA
YSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	T*PLALR.Y	NA	TPLALRY	NA	1380	NULL	T	NULL	T1380	NA
YSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ...	S*PTVS*QVTER.S	S*QVTER.S	SPTVSQVTERS	SQVTERS	1686	1690	S	S	S1686	S1690
YSVGFREADAATSFLRAARSGNLDKALDHLRNGVDINTCNQ	S*VGEP.E	NA	SVGEPE	NA	4	NULL	S	NULL	S4	NA

Part 9

This file contains:

- A column of -
- The complete site which includes the gene ID followed by dash, amino acid (S, T or Y) and amino acid position within the protein sequence
- This is saved in part3Final.txt



	part1	part2	part1Stripped	part2Stripped	part1Loc	part2Loc	amino1	amino2	pos1	pos2	dash	site
ANRQREPGLGFSFEFTEQ...	S*DPDPKAPANK.A	NA	SDPDPKAPANKA	NA	207	NULL	S	NULL	S207	NA	-	ACADM-S207
LLGQPRPGPARRPYAGG...	S*DALTR.K	NA	SDALTRK	NA	57	NULL	S	NULL	S57	NA	-	ACADVL-S57
'RSQNDNRERQEHNDRR...	S*LGHPEPLSNGRPQGNSR.Q	NA	SLGHPEPLSNGRPQGNSRQ	NA	43	NULL	S	NULL	S43	NA	-	PSEN1-S43
'RSQNDNRERQEHNDRR...	S*ILAGEDPEER.G	NA	SILAGEDPEERG	NA	367	NULL	S	NULL	S367	NA	-	PSEN1-S367
'RSQNDNRERQEHNDRR...	S*LGHPEPLSNGR.P	NA	SLGHPEPLSNGRP	NA	43	NULL	S	NULL	S43	NA	-	PSEN1-S43
'RSQNDNRERQEHNDRR...	Y*NAESTERESQDTVAENDDGGFSEEWEAQR.D	NA	YNAESTERESQDTVAENDDGGFSEEWEAQRD	NA	315	NULL	Y	NULL	Y315	NA	-	PSEN1-Y315
'RSQNDNRERQEHNDRR...	S*S*ILAGEDPEER.G	S*ILAGEDPEER.G	SSILAGEDPEERG	SILAGEDPEERG	366	367	S	S	S366	S367	-	PSEN1-S366S367
QQRDEVWVVGMGIVM...	S*QVEQDGR.T	NA	SQVEQDGRT	NA	246	NULL	S	NULL	S246	NA	-	ADRB2-S246
NASNLIYPIFVTDVPDDIQ...	S*PAFGDR.R	NA	SPAFGDRR	NA	215	NULL	S	NULL	S215	NA	-	ALAD-S215
NASNLIYPIFVTDVPDDIQ...	S*PAFGDRR.C	NA	SPAFGDRRC	NA	215	NULL	S	NULL	S215	NA	-	ALAD-S215
ALDHLRNGVDINTCNQ...	T*VS*QVTER.S	S*QVTER.S	TVSQVTERS	SQVTERS	1688	1690	T	S	T1688	S1690	-	ANK1-T1688S1690
ALDHLRNGVDINTCNQ...	S*VTDVLK.V	NA	SVTDVLKV	NA	781	NULL	S	NULL	S781	NA	-	ANK1-S781
ALDHLRNGVDINTCNQ...	S*RDVDEEK.E	NA	SRDVDEEKE	NA	834	NULL	S	NULL	S834	NA	-	ANK1-S834
ALDHLRNGVDINTCNQ...	S*PNEVSSDGTTPLAIAK.R	NA	SPNEVSSDGTTPLAIAKR	NA	759	NULL	S	NULL	S759	NA	-	ANK1-S759
ALDHLRNGVDINTCNQ...	T*PPPLAEEEGLASR.I	NA	TPPPLAEEEGLASRI	NA	961	NULL	T	NULL	T961	NA	-	ANK1-T961
ALDHLRNGVDINTCNQ...	S*PAIPR.I	NA	SPAIPRI	NA	856	NULL	S	NULL	S856	NA	-	ANK1-S856
ALDHLRNGVDINTCNQ...	T*PLALR.Y	NA	TPLALRY	NA	1380	NULL	T	NULL	T1380	NA	-	ANK1-T1380
ALDHLRNGVDINTCNQ...	S*PTVS*QVTER.S	S*QVTER.S	SPTVSQVTERS	SQVTERS	1686	1690	S	S	S1686	S1690	-	ANK1-S1686S1690
ALDHLRNGVDINTCNQ...	S*VGEP.E	NA	SVGEPE	NA	4	NULL	S	NULL	S4	NA	-	ANK1-S4

Showing 1 to 19 of 72,504 entries, 17 total columns

Part 10

This file contains:

- Asks user for input: The geneId
- Saves the input in a txt file called input.txt

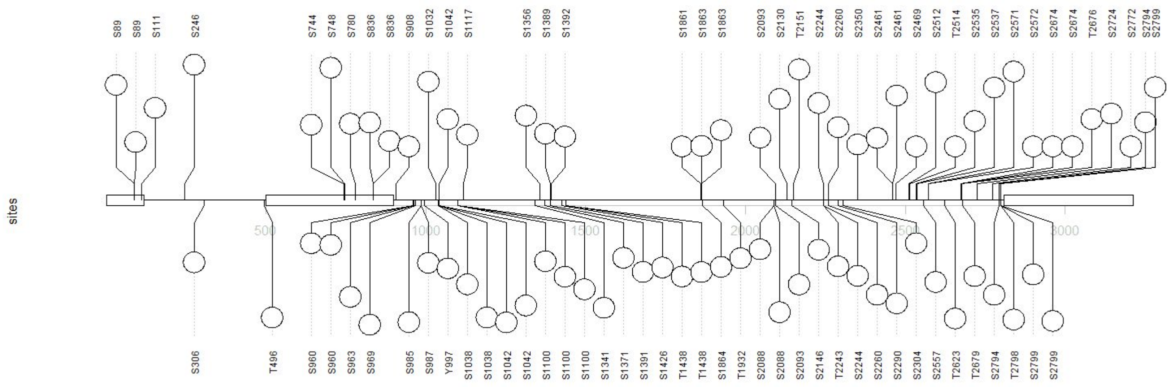
Part 11

This file contains:

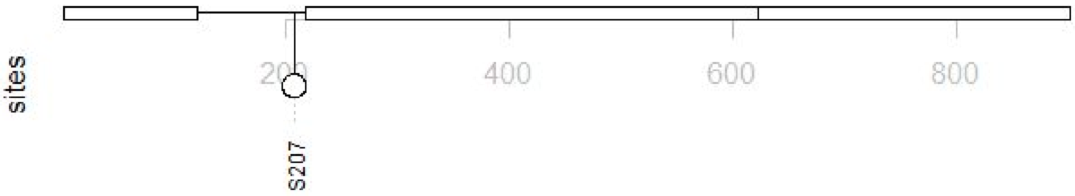
- The visualization

Below is an example for APC and ACADM (Also attached as png for clarity)

APC:



ACADM



What I learned:

- Bioconductor library for R and its various libraries and how they can be used for bioinformatics and gene mapping
- How to find position of substring in string
- How to create gene map visualizations using lolliplots
- How to handle large datasets
- I got more experience with R