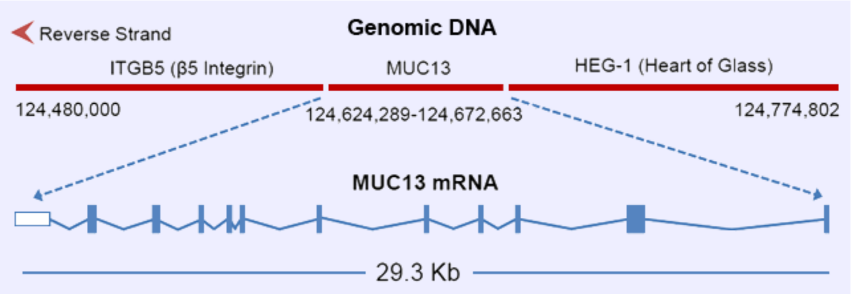
SESSION 11. A FUNCTION TO EVERY GENE

A slimy molecule

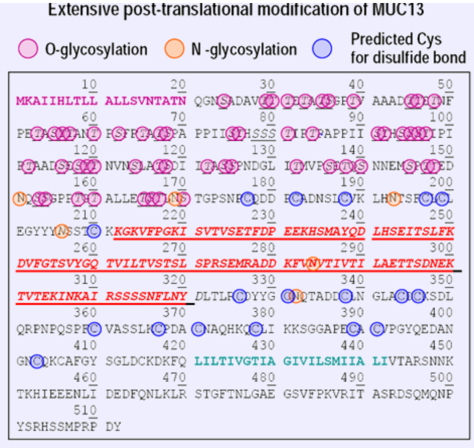
Functional similarity with low sequence similarity (by post-translational modification)

- The properties of the protein domain may not be captured on the basis of sequence alignment or by position-specific profile (PSSM).
- Typical example: Mucin (a characteristic property is to **form gel**).
  - There are many members in mucin family.
  - Mucins are a **major component of the mucous layer** that is present on the surface of epithelial cells of the lung and intestine.
  - Prevent harmful microorganisms and substances.
  - But, they have low sequence similarity and are difficult to capture by sequence alignment or profile-based search.



MUCIN with extensive sugar decoration

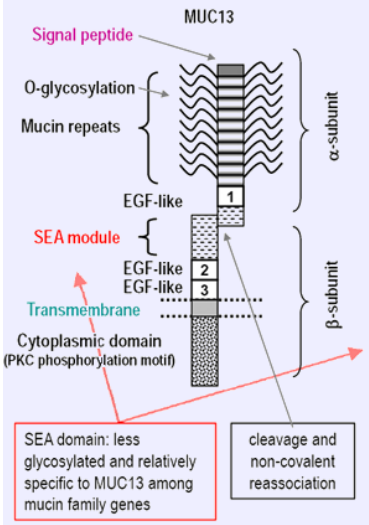
- *Very large protein*
- *Proline, threonine, and serine are enriched in a certain region → PTS domain.*
- ***Threonine and serine in PTS** are heavily **glycosylated**. It looks like **bottle brush** type of structure.*



Extensive post-translational modification of MUC13

○ O-glycosylation ○ N -glycosylation ○ Predicted Cys for disulfide bond

```
        10         20         30         40         50
MKAIIHLTLL ALLSVNTATN QGNSADAVTT TETATSGPTV AAADTTETNF
        60         70         80         90        100
PETASTTANT PSFPTATSPA PPIISTHSSS TIPTPAPPII STHSSTTIPI
       120        120        130        140        150
PTAADSESTT NVNSLATSDI ITASSPNDGL ITMVPSETTS NNEMSPTTED
       160        170        180        190        200
NQSSGPPTHT ALLETSTTNS TGPSNPCQDD PCADNSICVK LHNTSPCLCL
       210        220        230        240        250
EGYYYNSSTC KKGKVFPGKI SVTVSETFDP EEKHSMAYQD LHSEITSLFK
       260        270        280        290        300
DVFGTSVYGQ TVILTVSTSL SPRSEMRADD KFVNVTIVTI LAETTSDNEK
       310        320        330        340        350
TVTEKINKAI RSSSSNFLNY DLTLPCDYYG CNQTADDCLN GLACTCKSDL
       360        370        380        390        400
QRPNPQSPFC VASSLKCPDA CNAQHKQCLI KKSGGAPECA CVPGYQEDAN
       410        420        430        440        450
GNCQKCAFGY SGLDCKDKFQ LILTIVGTIA GIVILSMIIA LIVTARSNNK
       460        470        480        490        500
TKHIEEENLI DEDFQNLKLR STGFTNLGAE GSVFPKVRIT ASRDSQMQNP
       510
YSRHSSMPRP DY
```

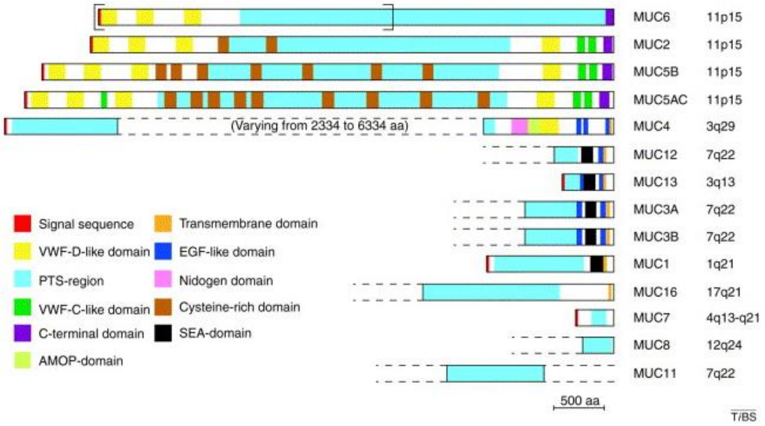


SEA domain: less glycosylated and relatively specific to MUC13 among mucin family genes

cleavage and non-covalent reassociation

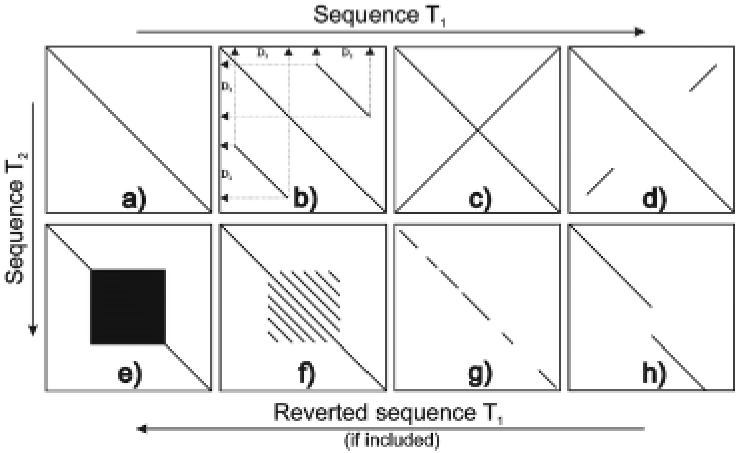
MUCIN with extensive sugar decoration

- *Membrane-bound or secreted.*
  - *Bound: MUC1, MUC3, MUC4, MUC12, MUC13, MUC16, MUC 17, and MUC20*
  - *Secreted gel forming: MUC2, MUC5B, MUC5AC, MUC6, and MUC 19*
- *Some Mucins contain SEA and VWD domains*
- *Among paralogs or orthologs, these PTS domains are not well conserved in sequence.*
  - *Actual sequence of aa is not so important for its function and* ***overall aa composition is matter****.*



Mucins and repeats

- *In addition to the a characteristic aa composition, many mucin PTS domains have **identical and near-identical repeats**.*
- *Repeats can be identified by dotplots*



Schematic overview of characteristic patterns appearing in dot plots. a-f) are self similiarity dot plots (T1=T2). g-h) are dot plots comparing two different sequences of simlar length.

a) A continuous main diagonal shows perfect similarity
b) Parallels to the main diagonal indicate repeated regions in the same reading direction on different parts of the sequences. In this case a region D is 'duplications'.
c) Lines perpendicular to the main diagonal indicate **palindromic areas**. In this case the sequence is completely palindromic in the displayed area.
d) Partially palindromic sequence
e) Bold blocks on the main diagonal indicate repetition of the same symbol in both sequences, e.g. (G)50, so called microsatellite repeats
f) Parallel lines indicate tandem repeats of a larger motif in both sequences, e.g. (AGCTCTGAC)20, so called minisatellite patterns.
g) When the diagonal is a discontinuous line this indicates that the sequences T1 and T2 share a common source.
h) Partial deletion in sequence 1 or insertion in sequence 2,

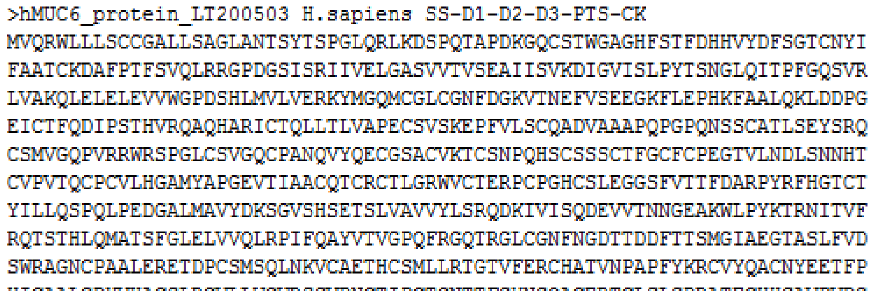
Mucins and repeats

- *Mucins are notoriously difficult to work with gene technologies because of the PTS domain repeats.*
  - *Difficult to clone as recombination events*
  - *Sequence assembly cloud have an error in length of repeats.*
- *WGA often misses the assembly of the repeat region like mucin*
- *The current human genome assembly is still lacking a complete version of the MUC5AC*
- *MUC1, 2 have many identical repeats (F13.4) but MUC6 has non-identical repeats*

Computational identification of mucin domains

- □ *SEA and VWD domains of Mucins can be searched by Pfam or HMMER*
- □ *But other proteins could contain the domains*
- □ *Mucins have to include one or more PTS domains*
- □ *But the PTS domains can not be really detected by BLAST or HMMER*

- □ ***Typical PTS domain includes more than 40% of serine and threonine and more than 5% of proline.***
- □ ***Minimum length of PTS domain is 100aa***



```
>hMUC6_protein_LT200503 H.sapiens SS-D1-D2-D3-PTS-CK
MVQRWLLLSCCGALLSAGLANTSYTSPGLQRLKDSPQTAPDKGQCSTWGAGHFSTFDHHVYDFSGTCNYI
FAATCKDAFPTFSVQLRRGPDGSISRIIVELGASVVTVSEAIISVKDIGVISLPYTSNGLQITPFGQSVR
LVAKQLELELEVVWGPDSHLMVLVERKYMGQMCGLCGNFDGKVTNEFVSEEGKFLEPHKFAALQKLDDPG
EICTFQDIPSTHVRQAQHARICTQLLTLVAPECSVSKEPFVLSCQADVAAAPQPGPQNSSCATLSEYSRQ
CSMVGQPVRRWRSPGLCSVGQCPANQVYQECGSACVKTCSNPQHSCSSSCTFGCFCPEGTVLNDLSNNHT
CVPVTQCPCVLHGAMYAPGEVTIAACQTCRCTLGRWVCTERPCPGHCSLEGGSFVTTFDARPYRFHGTCT
YILLQSPQLPEDGALMAVYDKSGVSHSETSLVAVVYLSRQDKIVISQDEVVTNNGEAKWLPYKTRNITVF
RQTSTHLQMATSFGLELVVQLRPIFQAYVTVGPQFRGQTRGLCGNFNGDTTDDFTTSMGIAEGTASLFVD
SWRAGNCPAALERETDPCSMSQLNKVCAETHCSMLLRTGTVFERCHATVNPAPFYKRCVYQACNYEETFP
```

File input as sys arguments

```
import sys

filename1 = sys.argv[1]
filename2 = sys.argv[2]
filename3 = sys.argv[3]


filein = open(filename, ‘r’)


for line in filein: print line
print filename2, filename3
```

```
python pts.py muc6.fa muc5.fa muc4.fa
```

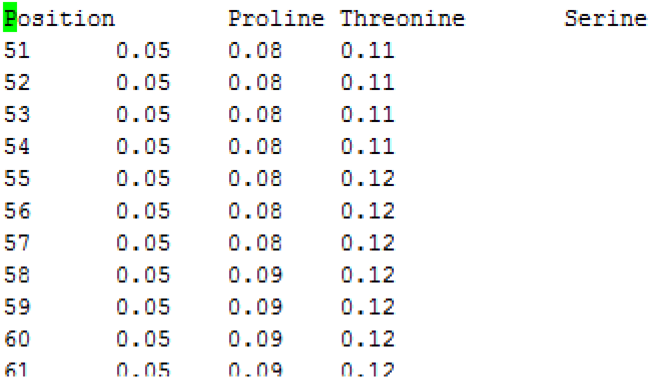
pts.py

```
#!/usr/bin/python

import re
import sys
# Basic parameters used
wid  = 100  # size of sliding window
step = 1    # size of step to move sliding window

# check if argument to the script is there.
if len(sys.argv) > 1:
    file = sys.argv[1]
else:
    exit('File in FASTA sequence format is to be used as argument to the script')
# read the sequence from the input file
seq = ''
id = ''
for line in open(file):
    line = line.rstrip()
        # in the identifier line all is captured
        # in the variable 'id' except for
        # the > character

    match = re.search('>(.*)', line)
    if match:
        id = match.group(1)
    else:
        seq = seq + line
```

pts.py

```
# Now analyze the sequence in $seq
print 'Position\tProline\tThreonine\tSerine'
for i in range(0, len(seq) - wid + 1, step):
    test = seq[i:i + wid]
    # Count proline, threonine and serine
    count_p = float(test.count('P’))
    count_t = float(test.count('T’))
    count_s = float(test.count('S'))
    pos = i + 1 + wid / 2
    print pos, '\t', count_p / wid, '\t', count_t / wid, '\t', count_s / wid
```

python pts.py muc6.fa >pts.out



```
Position        Proline Threonine       Serine
51      0.05    0.08    0.11
52      0.05    0.08    0.11
53      0.05    0.08    0.11
54      0.05    0.08    0.11
55      0.05    0.08    0.12
56      0.05    0.08    0.12
57      0.05    0.08    0.12
58      0.05    0.09    0.12
59      0.05    0.09    0.12
60      0.05    0.09    0.12
61      0.05    0.09    0.12
```

Visualization of PTS landscape with R (pts.r)

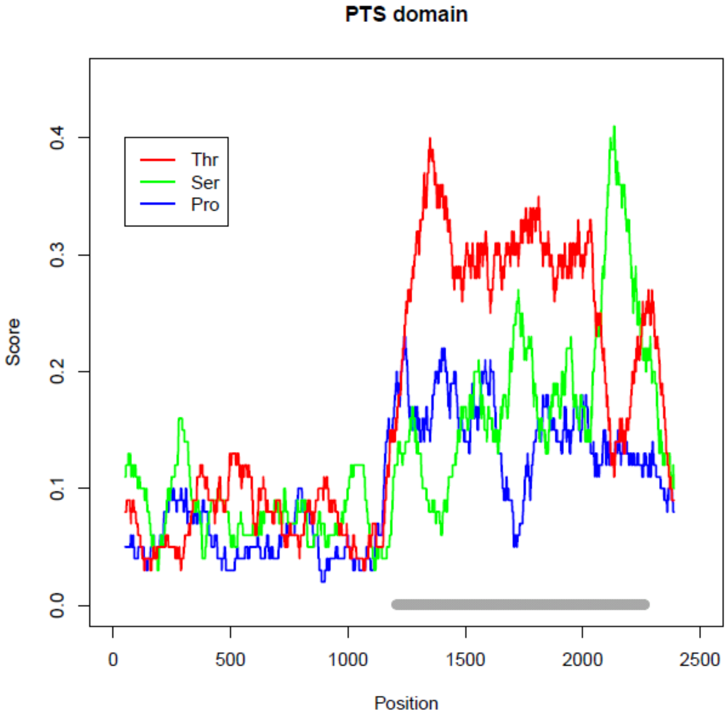
```
# read information from output from Perl script
data <- read.table("pts.out", sep = "\t", header = TRUE)

# make an empty plot
pdf("ptsL.pdf")
plot(0, type = "n", xlim = c(0, 2500), ylim = c(0,
    0.45), main = "PTS domain", xlab = "Position", ylab = "Score")

# draw lines for Proline, Serine and Threonine data
lines(data$Position, data$Proline, col = "blue", lwd = 2)
lines(data$Position, data$Serine, col = "green", lwd = 2)
lines(data$Position, data$Threonine, col = "red", lwd = 2)

# make a legend
legend(50, 0.4, c("Thr", "Ser", "Pro"), col = c("red","green", "blue"), lwd = 2)

# add a line indicating the 40% / 5% cutoff
len <- length(data$Position) # number of lines in the file
for (i in (1:len)) {
    if (((data$Serine[i] + data$Threonine[i]) > 0.4) && (data$Proline[i] >
        0.05)) {
        points(i, 0, col = "darkgrey")
    }
}
dev.off()
```

Visualization of PTS landscape

- In R,

  source("pts.r")

  open "ptsL.pdf"



Term project (by Dec-11)

- 종별 Codon Usage 비교
  - Extract coding sequences of protein-coding genes from any two of Human, Mouse, Zebrafish, Fly, C.elegans, Yeast, Arabidopsis..
  - Build codon tables with frequency
  - From Codon frequency to Codon usages (ratio by aa)
  - Comparison b/w two species
- 제출물
  - Python, R codes (Jupytor notebook 제출 가능) 자신이름 documentation 필수
  - 분석 보고서(이름, 학번, 분석방법, 분석결과, 토의, 참고문헌)