Hemocyte proteome pI MW

Polina Drozdova

1/4/2021

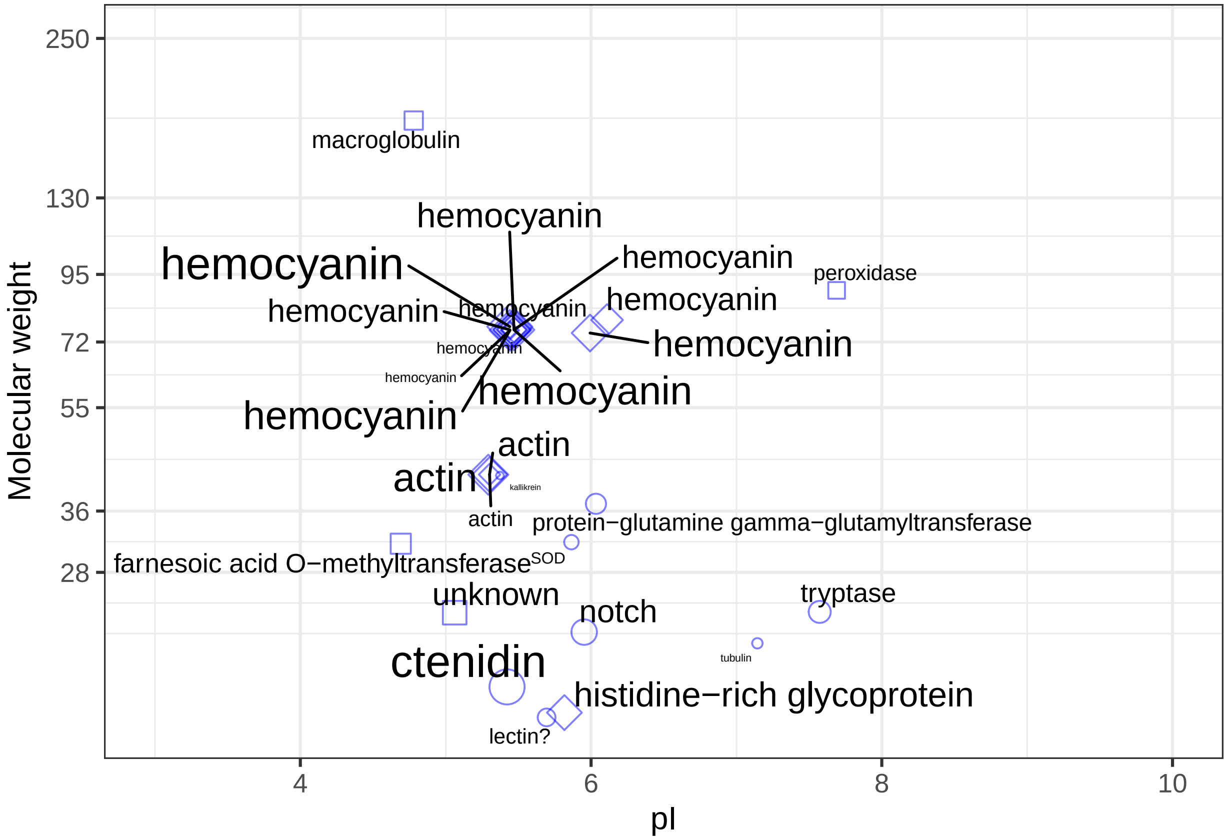
```
library(seqinr)
library(ggplot2)
library(ggrepel)
library(openxlsx)
library(knitr) ## for kable

getMedianMWpI <- function(seqvector) {
  seqdf <- data.frame(id = 1:length(seqvector), MW = NA, pI = NA)
  seqdf$MW <- sapply(seqvector, function(x) pmw(toupper(x))/1000)
  seqdf$pI <- sapply(seqvector, function(x) computePI(toupper(x)))
  kable(seqdf) ## it would be print(seqdf) in a usual script
  print(median(seqdf$MW))
  print(median(seqdf$pI))
}
```

The main part: the data behind Fig. 3B.

```
abundance_upd <- read.xlsx("Table_S1_Abundance_MW_pI.xlsx")

ggplot(data = abundance_upd[complete.cases(abundance_upd$Source) &
                              abundance_upd$Total.Quantity > 1391 &
                              abundance_upd$Reliability > 2 &
                              abundance_upd$MW_final > 15, ],
       aes(x = pI_final, y = MW_final,
           label = Short_annotation,
           shape = Source,
           size = log10(Total.Quantity)
           )) +
  geom_point(alpha = 0.5, col = "blue") +
  scale_x_continuous(limits=c(3, 10))  +
  geom_text_repel() +
  ylab("Molecular weight") + xlab("pI") +
  theme_bw(base_size = 12) + ## 16 for the figure
  #scale_size(limits = c(2.5, 4)) + ## log10(500) is 2.7
  scale_shape_manual(values = c(21, 22, 23)) + ## With letters
  #scale_shape_manual(values = c("\U25CF", "\U2735", "\U2736")) + ## With letters
  scale_y_log10(breaks = c(28, 36, 55, 72, 95, 130, 250), limits = c(15, 250)) +
  #  scale_x_binned(breaks = c(3.0, 4.0, 4.5, 5.0, 5.5, 6, 10), limits = c(3, 10)) +
  theme(legend.position = c(9, 200))
```



```
ggsave("S3B.svg", width = 21, height = 15, units = "cm")
```

The details of analyzing the top 50 proteins.

```
## 1
# >GHHK01018663_1__2_400___
# LLPAMKTILSLLLLVATAAGNRAMEYTSLDLSKVDFENVSGRGNQEYISVGDSWSDEAEA
# YRLVGVAPASAVRGGAALTYAANIGPTVGVYRELDDSYGPNYRRRVYSSQRPIWVQVEAT
# VAGGSFAVLLDD*
## Peptide coverage:  from 63 (???) LVGVAPASAVR???
## to VAGGSF...
## swissprot: nothing found / nr: only 2 unknown proteins with different lengths
## SignalP-5.0: Cleavage between positions 20 and 21
seqGHHK01018663 <- read.fasta("MW_PI_manual_seqs/GHHK01018663_output_mature.fasta")
sapply(seqGHHK01018663, function(x) pmw(toupper(x))/1000)
```

GHHK01018663_1_*2_400*___ 12.15719

```
sapply(seqGHHK01018663, function(x) computePI(toupper(x)))
```

GHHK01018663_1_*2_400*___ 4.617165

```
## Possible causes:
## 1 Either we don't see it because it's very small,
## 2 or it's the ~200/7 spot on the gel
## 3 or maybe it's just an assembly error and doesn't exist (not very plausible)
```

```r
## 2
# >GHHK01000356_1__2_361___
# QLTVRFDAERLSNHLDLVDELYWDRPIKEGFAPHATYRYGGEFPSRPDNVDFEDVDGIVR
# VRDMIIHENRIRDAIAHGYITAADGSHINIRDEHGIDHLGDIIESSLYSPNAQYYGALHN
## It's a fragment. Swissprot blastp
seqvect2 <- read.fasta("MW_PI_manual_seqs/GHHK01000356_1__2_361___10_best_hits.fasta")
getMedianMWpI(seqvect2)
```

[1] 75.68387 [1] 5.441646

```r
## 3
# GHHK01015013_1__1_621___
# >GHHK01015013_1__1_621___
# RTLSNMMRLLLICSALLAGASGTEPVKEKETRLFGGGGGIGAPPFGSPGGGGGFGGGGKG
# KGPFGAGGANGGGGFPGGPGGAGDFGGAGDFGGAGDFGGAGDFGGAGDFGGAGGLGGGDG
# FVESGLGGGCKNFCKKHGKYTCCDNDKGSFGNVKRGECPPVRPSCPRFKNPQICDDDGEC
# GGIDKCCFDKCLGEKVCKPPSPDTGY*
# It's at least C-complete
## Then maybe guess by the covered part?
## SignalP-5.0: a cleavage between 22 and 23...
## PS: coverage from 23 to the very end. Then guess by output.
seqGHHK01015013 <- read.fasta("MW_PI_manual_seqs/GHHK01015013_1__1_621__output_mature.fasta")
sapply(seqGHHK01015013, function(x) pmw(toupper(x))/1000)
```

```
       1
17.50604
```

```r
sapply(seqGHHK01015013, function(x) computePI(toupper(x)))
```

```
       1
5.421798
```

```r
## 4
# >GHHK01000364_1__52_501___
# MDKFWTKLAAGDNHIKRKSSESSVSVPDVPSFASLIHDADAAVASGSDLHLEAFDRACGL
# PQRMLLPKGTEEGMDFLLVVAVTDGTTDSQHDALEAVDAHGHAQCGVHGEKYPDHQPMGF
# PLDRRIPDERLFLKADNVGYTIVKVFHKE*
## it's a hemocyanin again...  swissprot blastp
seqvect4 <- read.fasta("MW_PI_manual_seqs/GHHK01000364_1__52_501___10_best_hits.fasta")
getMedianMWpI(seqvect4)
```

[1] 75.68387 [1] 5.469837

```r
## 5
# >GHHK01000367_1__2_871___
# FDAERLSNHLDMVDELYWDRPIKNGFAPHTTYKYGGEFPSRPDNIEFEDVDGLVNVRDMI
# IHEGRIRDAIAHGYITAADGSHINIRDEHGIDHLGDIVESSLYSPNAQYYGALHNEAHII
# LGRQADPHGKFNLPPSVMEHFETATRDPAFFRLHKYMDNIFKEHKDSLTPYTAEEIGFPG
# VHLTSLSIEGELETFFEDFEFDLKMAVDSSEAVAEVDVKAHVSRMNHKDFAYNFDIKSDA
# ADQHAVVRVFFCPRRDNNGIIFTFEEGRWNCIEMDKFWTKLAAGDNHIKR
seqvect5 <- read.fasta(file = "MW_PI_manual_seqs/GHHK01000367_1__2_871___10_best_hits.fasta")
getMedianMWpI(seqvect5)
```

[1] 75.53213 [1] 5.441646

```
## 6
# >GHHK01011646_1__3_314___
# TLKYPIEHGIITNWDDMEKIWHHTFYNELRVAPEECPVLLTEAPLNPKVNREKMTQIMFE
# TFGCPATYVAIQAVLSLYASGRTTGIVLDSGDGVSHSVPIFEGF
## it's a fragment of actin
seqvect6 <- read.fasta(file = "MW_PI_manual_seqs/GHHK01011646_1__3_314__10_best_hits.fasta")
## the first is incomplete. Other're fine ;(
getMedianMWpI(seqvect6)
```

[1] 41.7889 [1] 5.293354

```
## 7
# >GHHK01005813_1__115_549___
# MKFTLVLVLSTVIASALAQRPQTQPPPGQNTDTRLGLLAGQIGVPAVPGAFPNTGIADFV
# PPLQGRDESCWCQAINQLCTPLRPVNLDLVTRIINRPGSPGSPGSVPSCGEDRRMCCPQD
# PPGQGFPGQGFPGQGFPGQGFPGQG
## Nothing
## what should I do if I don't know where it ends??? It ends in like a repeat?

## 8
# >GHHK01013436_1__2_475___
# AAGARGFTSMFLFVRSSRIRNNMIRAICLCVVLLAGASANVKHKRLFYGGGFAGLGDELE
# GGGAVGLGGAGGYGGVSDTCRYWCKTDVGQAYCCESGLEEPGLVGTKPGKCPPVRPTCPR
# FKAPQLCSNDGACGGIDKCCFDKCLGEHVCKPPLPGY*
## Nothing
## not really a signal peptide...
## PS: from TDVGQAYCCESGLEEPGLVGTKPGK to CLGEHVCKPPLPGY (it's the very end)
## From first methionine?
seqGHHK01013436 <- read.fasta("MW_PI_manual_seqs/GHHK01013436_first_Met.fasta")
sapply(seqGHHK01013436, function(x) pmw(toupper(x))/1000)
```

GHHK01013436_1_*2_475*___ 15.47185

```
sapply(seqGHHK01013436, function(x) computePI(toupper(x)))
```

GHHK01013436_1_*2_475*___ 8.534183

```
## Only the thing covered with peptides?
seqGHHK01013436 <- read.fasta("MW_PI_manual_seqs/GHHK01013436_covered.fasta")
sapply(seqGHHK01013436, function(x) pmw(toupper(x))/1000)
```

GHHK01013436_1_*2_475*___ 7.543684

```
sapply(seqGHHK01013436, function(x) computePI(toupper(x)))
```

GHHK01013436_1_*2_475*___ 6.432501

```
## Then we don't see it...

## 9
# >GHHK01009804_1__1_567___
# ARPWHVERPSDPLYVPRAGNTIVFRLPADQMPGGFSAQSGGTTRTVVNVETSPAVASALP
# AAVKAIPRSLPFSLNFREHRRAAKIVIDLLQQSESVTQLRDIAASIRGEINETLFVYSLS
# SVITRNPRFRTIRVPAVTETFPSRFVPSSVIDRARALTNRANSNRNTDLTQPMVINHDQD
# FSGLRGRLE
## it's hemocyanin again...
seqvect7 <- read.fasta("MW_PI_manual_seqs/GHHK01009804_1__1_567__10_best_hits.fasta")
```

```r
getMedianMWpI(seqvect7)
```

[1] 74.7033 [1] 5.993066

```r
## 10
# >GHHK01019991_1__117_497___
# MSWDEYVSGQLIGSGNIKEAAICGQDGSTWATSAGFNVSQAEALKLVAGFTDPSGLQAGG
# MNISGTKFIFLSSTDEVLRGKKEQRGVHIAKTKTAIIIAIYEEPIQPGQCAITVEALAEY
# LRGVNY*
## profilin
seqvect8 <- read.fasta("MW_PI_manual_seqs/GHHK01019991_1__117_497___3_hits.fasta")
getMedianMWpI(seqvect8)
```

[1] 13.7195 [1] 5.638842

```r
## 11
# >GHHK01000358_1__2_469___
# KAAAANFNPVADKSIYSDGGVAAQQLVDELTDHRLLEKHHWFSLFNPRQREEALLLFDVL
# MHCKTWEAALNNAAYFREQMNEGEFVYALYAAVIHSELGAGIVLPPLYEVTPHMFTNSEV
# IQKAYTAQMTQTPGNFKMDFTGSKKNPEQHVAYFGE
seqvect9 <- read.fasta("MW_PI_manual_seqs/GHHK01000358_1__2_469__10_best_seqs.fasta")
getMedianMWpI(seqvect9)
```

[1] 75.78362 [1] 5.469837

```r
## 12
# >GHHK01008003_1__1_957___
# SHDHGHSHEDHHHDHSHDHDHAHNHSHDHGDSHEDHHHDHSHDHDHAQNHSHDHGHSHED
# HHHDHDHAHNHSHDHGHSHEHHHHDHSHGHGTTEGVNVTSSNNNEEHRPPIRGHHNAEHP
# NHQDSHHHKHIDPSHHQHPRNQQLGMARATCEVKPNTGDDNSTVTGNITITQRKAGDGPV
# YFDIDLEGFDNTQVEASLYGFHIHESPVTGDDCATAGGHLNPHTTVHGGPTDDVRHVGDL
# GNIEVAADGRLSGYIVVDYVVAFSGDNNIIGKSLVVHSTKDDLGQGGDAGSLATGNAGSR
# LACCNIHIAAEGRFRFGG*
seqvect10 <- read.fasta("MW_PI_manual_seqs/GHHK01008003_1__1_957___10_best_hits.fasta")
getMedianMWpI(seqvect10)
```

[1] 15.75599 [1] 5.816555

```r
## 13
# GHHK01029156_1__39_395___
# >GHHK01029156_1__39_395___
# MYVAIQAVLSLYASGRTTGIVLDSGDGVSHTVPIYEGYALPHAILRLDLAGRDLTDYLMK
# ILTERGYTFTTTAEREIVRDIKEKLCYVALDFEQEMTTAASSSSLEKSYELPDGQVITI
seqvect11 <- read.fasta("MW_PI_manual_seqs/GHHK01029156_1__39_395___10_best_hits.fasta")
getMedianMWpI(seqvect11)
```

[1] 41.81435 [1] 5.301928

```r
## 14
# >GHHK01019871_1__27_413___
# MKTILSLLLLVATASGNRAIEYTSLDLSQVDFEKVSGPGNQEYISVEDSWSDEAEAYRLV
# GVAPASAVKGGAAWTYLANNGPTVGVYKKLDGSYPVPEYRRVSSSQRPNWVEVEATVAGG
# SFVVLLGG*
## can it be complete?
## blast (even nr): two hypothetical proteins, H. azteca and T. longiramus
#Cleavage site between pos. 16 and 17
## Peptides: NRAIEYTSLDLSQVDFEK (start 17) to DGSYPVPEYR
## The whole protein  Theoretical pI/Mw: 4.73 / 13616.27
## So, even the whole protein is unlikely to be seen...


## 15
# >GHHK01000357_1__3_440___
# HKDSLTPYTAEEIGFPGVHVTGVSIEGELETFFEDFEFDLKMAVDTSESVAEVEVKAHVN
# RLNHKDFAYNFDIKSDSADQHAVVRVFLCPRRDNNGIQFTFDEGRWNCIEMDKFWTKRKC
# VWFGSRRDVSFKFFPFILPGRVVCF*
## No signal peptide
seqvect12 <- read.fasta("MW_PI_manual_seqs/GHHK01000357_1__3_440___10_best_hits.fasta")
getMedianMWpI(seqvect12)
```

[1] 75.68387 [1] 5.441646

```r
## 16
# >GHHK01019730_1__1_324___
# SLLPALTVSLEKSCADIGGKEFVSLWHDGLKRQRIRVPETTPDPAPRTIVAPHNSAIHAL
# VPMDAPNPLYAESAEGVALRTRLLRAYMLAEAVSPEDDKLKQEGGYQA
## Definitely a fragment...
seqvect13 <- read.fasta("MW_PI_manual_seqs/GHHK01019730_1__1_324___10_best_seqs_nr.fasta")
getMedianMWpI(seqvect13)
```

[1] 23.7312 [1] 5.061709

```r
## 17
# >GHHK01029332_1__1_306___
# NYCKSRGMTLASIHSQDEQTFVEPLLPDFVWIGMDDHGREGDFRWIDGTPLDYNHFKSGQ
# PDNHLLSEHCAEMHKEIDYYWNDWLCNRVLQFLCKSAVSVQ*
## Swissprot: it's a lectin!
## only 2 matches ;( )
seqvect14 <- read.fasta("MW_PI_manual_seqs/GHHK01029332_1__1_306___2_best_seqs.fasta")
getMedianMWpI(seqvect14)
```

[1] 30.70524 [1] 5.458975

```r
## 18
## >GHHK01000350_1__2_583___
# PAFFRLHKYMDNIFKEHKDSLTPYTAEEIGFPGVHLTSFGIEGELETHFEDFEYDLKMAV
# DSSDSVKEVEIKAHVSRLNHKDFAFTFDIKSNAADQHAVVRVFLCPRKDNNGVIFTFEEG
# RWHCIEMDKFWTKLNSGNNHITRKSAESSVTVPDIPSFASLIHDADEAVASGSDLHLEEF
# DRSCGIPSRLLLPK
seqvect15 <- read.fasta("MW_PI_manual_seqs/GHHK01000350_1__2_583___10_best_hits.fasta")
getMedianMWpI(seqvect15)
```

[1] 75.68387 [1] 5.469837

```r
## 19
# >GHHK01008760_1__1_714___
# TICAPLCEATCENGGTCVAPNKCACTGAYTGDTCSEEPMLGDQPYQGFQAASLGPNQRRV
# ECEPGTRMPDGSTSVTLTFRDSSWFYPDGRLLLGVDQVTCEKATESVSSTPEETPEEEKP
# ISNLPNPEPLLAATSSPPILCHPPCQNGGQCVYTNTCSCPRGFWGPSCQVSTCTYPRFQN
# NLNASLGGTLKKMKFECHPGHHTRQGHDRVVTICQSGRWMLPGGRMLVEDDVRCIRD*
## seems to have an end...
## what about a beginning?
## no signal protein (SignalP-5.0)
## from RVECEPGTR to MLVEDDVR (4 amino acids to the end, so might as well be to the end)
## The first peptide is in 20 aa from the Met
seqGHHK01013436 <- read.fasta("MW_PI_manual_seqs/GHHK01008760_1__1_714___from_first_Met.fasta")
sapply(seqGHHK01013436, function(x) pmw(toupper(x))/1000)
```

```
GHHK01008760_1__1_714___ 21.91748
```

```
sapply(seqGHHK01013436, function(x) computePI(toupper(x)))
```

```
GHHK01008760_1__1_714___ 5.952597
```

```
## TODO maybe also try with Swissprot?

## 20
# >GHHK01009797_1__3_671___
# APSSVITRNKNFRSLPVPPLLESFPGRFVPSNVIDRARALTNMANSNMNTDRTQPLVVNH
# DQDFSGLRGRLENRVAYWREDYALNAHHWHWHIVYPTDVRTANSPDRKGELFYYMHSQII
# ARYDMERLSVGLPRVIKLENFREPILEGYFSKLRFDNADTNENRVPVPGLHPNATLWGAR
# QDNTTLSNYARLEPFIPVDIAELEMWADRLLDGIHQGFFFNRA
seqvect16 <- read.fasta("MW_PI_manual_seqs/GHHK01009797_1__3_671___10_best_hits.fasta")
getMedianMWpI(seqvect16)
```

```
[1] 78.80161 [1] 6.110083
```

```
## 21
# >GHHK01001525_1__1_366___
# MTSTVEFAAVQAAVAGSQVVILDVRNKAEVETNGFIKGSIHVPLREVEAALALEADEFKT
# KYGSDKPDEADEIITHCMLGGRAQKAGDALVAKGFSNVKVYKGSFTDWKEQGGEIIMPES
# S*
### hm... is it complete?
### peptides: GSIHVPLR (not the beginning...) to EQGGEIIMPESS (the very end)
## Swiss-prot: only 1 hit (D. melanogaster)
## Nr: many hits...
seqvect18 <- read.fasta("MW_PI_manual_seqs/GHHK01001525_1__1_366___10_best_hits_nr.fasta")
getMedianMWpI(seqvect18)
```

```
[1] 12.80909 [1] 5.90069
```

```
## Similar to the sequence itself ( Theoretical pI/Mw: 4.89 / 12946.63 )

## 22
#   >GHHK01005619_1__148_813___
# MGEELEVYDTGDEKLYRFKPFVGKTLRFMVKAAHDCHIAFTTNEGDSTPMFEVFLGGWEG
# EYSAVRFSKGDDLVKEHTPDILSADEFREFWIATDHDEVRVGRGGEFEPFLSCTLPEPVN
# PTFFGFTTGWGATGGFQFLHERNIATEDKLEYRYEPLYGDTFTFTVSCDHDAHLSFTMGP
# EQTPLMYEVFIGGWSNQHSAIRKSKETTVVKVETPDECCGDP
## C-incomplete ;(
## nothing is Swiss-Prot
seqvect17 <- read.fasta("MW_PI_manual_seqs/GHHK01005619_1__148_813___10_best_hits_nr.fasta")
getMedianMWpI(seqvect17)
```

```
[1] 31.48906 [1] 4.690845
```

```
## 23
# >GHHK01010607_1__1_858___
# FPGQGFPGQGFPGQGFPGQGFPGQGFPGQCGSRTPFGLPSAISPTADFGEYPWMAVVMGP
# GQAYMAGGVLIADGWVLTAAHKLTSNRGLIVRLGDYDVGSANDVPQFPEFEVAVSRVIVH
# PEYNSNTLANDVALLQLRRPVNRQQYRHVTPACIPAQGQQFDGQRCFVTGWGQNAFSSAQ
```

```r
# GNFQRVLQEVDVPVVDSFRCEAVLKTTRLGQAFTLDKRSFICAGGEQDKDACQGDGGSPM
# VCGGGGQGWTVAGLVAWGVGCGRQGIPSAYVNVPTYVSFIRQYVK*
## at least C-complete
## peptides: MAGGVLIADGWVLTAAHK to VNVPTYVSFIR (basically the end)
## no signal sequence
## it's the third Met
seqGHHK01010607 <- read.fasta("MW_PI_manual_seqs/GHHK01010607_1__1_858___from_third_Met.fasta")
sapply(seqGHHK01010607, function(x) pmw(toupper(x))/1000)
```

GHHK01010607_1_*1_858*___ 23.80964

```r
sapply(seqGHHK01010607, function(x) computePI(toupper(x)))
```

GHHK01010607_1_*1_858*___ 7.572873

```r
## 24
# >GHHK01005045_1__173_658___
# MSSETPQKDRPEGHATLMNQLEGFTPDKLKPAQTEEKLALPTKEDVLAEKALQEHLTSIE
# HTGKDKLKRTNTTEKFVLPSKEDIETERSHQSLFQGIEGFDKASMQHAETQEKITLPDKQ
# DIAAEKGQQALLSGIAGFDSSALKKTETHEKNPLPTKEVIEQ
## it's a fragment ;(
## peptides: DRPEGHATLMNQLEGFTPDK to LSGIAGFDSSALK...
## nr: only 1 similar seq; the other 50% identity... ;(


## 25
# >GHHK01010851_1__1_849___
# WLSSLTRRSRRIYLMFAMMFIKCVTVAVLLLGISALSSAQGYGQSIKFPSSQQQQACIGT
# RLDKISLLQDMYVQYASFSADVPTMYAFHTCLWLKLDKIYGRNAATTLNYGLDDTTNTDN
# LTIQYETSKQSWTLNINGIRIFNTKAVQVGEGRWNHFCQSWDGRTGQWNVWQNGALLDEG
# VNTKSIGLVIPGGGTMVTGQHTKTLFNGMDVLEGIIGSITLLYVSKEPIPNTSSRGTQQY
# LRLLASDCKASDRGDVVGWLRAPRKLYGGVMTELANESCGNF*
## at least C-complete...
## peptides: FPSSQQQQACIGTR to WNHFCQSWDGR (not the start and not the end)
## no signal sequence...
## 3 seqs from the same species in swissprot
# seqvect19 <- read.fasta(file = "MW_PI_manual_seqs/GHHK01010851_1__1_849___10_best_hits_nr.fasta")
# seqdf19 <- data.frame(id = 1:10, MW = NA, pI = NA)
# seqdf19$MW <- sapply(seqvect19, function(x) pmw(toupper(x))/1000)
# seqdf19$pI <- sapply(seqvect19, function(x) computePI(toupper(x)))
# median(seqdf19$MW); median(seqdf19$pI)
## Very diverse; very low identity; nope ;(

## 26
### clearly a piece
# >GHHK01016178_1__1_447___
# INETKDITFATNVAETFIQTDKYLYKAGQKVQFRVLTLQGPFFKVSTEMYPEIIVETPSG
# SRIAQWLNVSNPSGLIHLDLQLIEEPEEGMYTIKAISPASGETEMETFSIEDYVLPRFEV
# TVKPPKYLLADGEVLKIEVCATYTFGQPV
## nothing in swiss-prot
## nr. The first 2 are said to be N-term; so the rest is to be trusted
seqvect28 <- read.fasta("MW_PI_manual_seqs/GHHK01016178_1__1_447___10_best_seqs_nr.fasta")
getMedianMWpI(seqvect28)
```

[1] 178.5084 [1] 4.779966

```
### 27
### another piece of hemocyanin...
# >GHHK01000360_1__1_567___
# VADKSIYSDGGVAAQHLVDELTDHRLLEKHHWFSLFNPRQREEALLLFDVLMHCKTWETA
# LNNAAYFREQMNEGEFVYALYAAVIHSKLGAGIVLPPLYEVTPHMFTNSEVIQKAYTAQM
# TQTPGTFKMDFTGSQKNPEQHVAYFGEDIGMNVHHVTWHLDFPFWWEDSYGYHLDRKGEL
# FFWAHHQLT
seqvect28 <- read.fasta("MW_PI_manual_seqs/GHHK01000360_1__1_567___10_best_hits.fasta")
getMedianMWpI(seqvect28)
```

[1] 75.68387 [1] 5.469837

```
### 28
# >GHHK01019889_1__2_307___
# PEATRFCQSEGGTLASTSTLQMKEAVVNFVNRNAPGQYWTSGRDVGSGRFMWTDTYGEIS
# ISFRGRNFRSGSCVYMCSHTRMFWDRPCDQHLGFICHTKAA*
## C-complete, but overall looks like a total fragment (or veery short)
## nr: the best identity 42%... But all the hits >150 aa. So, I can't say anything.

### 29
# >GHHK01026820_1__1_1029___
# QETSDGMYQCGPASLEAVRRGEVSLQYDVPFVLAEVNADLVRWQEDETSENGFKMINSHK
# SHIGRQLLTKAVGVLDDTSGSTADREDCTTDYKAPEGTDTERVTLYGAARNIRTARHAFR
# FPSVAEMDVVFELEKVDVVDVGQDYAVAVKITNNGSAVRTVSLSLSSSSEYYTGVKAHTV
# KRAEGTFVMQPGKEEALRMPVRYKDYITKLVEHGTMKILAIGNVKETTQSYIEDDKFQIR
# KPNITVDTPNTSVLGTEMVVRVHFNNPLQEPLTEAYIVVDGPGLTRPKRIPVPDVPAKTL
# FSHSLKLVSKRAGERSLVVTFGSKQITDIMGSSNIVVTAQEA*
## C-complete but N?
## Peptides: RGEVSLQYDVPFVLAEVNADLVR (20+) to SLVVTFGSK (almost the end)
## SignalP: no signal peptide...
seqGHHK01026820 <- read.fasta("MW_PI_manual_seqs/GHHK01026820_1__1_1029___first_Met.fasta")
sapply(seqGHHK01026820, function(x) pmw(toupper(x))/1000)
```

GHHK01026820_1_*1_1029*___ 37.08761

```
sapply(seqGHHK01026820, function(x) computePI(toupper(x)))
```

GHHK01026820_1_*1_1029*___ 6.033634

```
### 30
# >GHHK01013740_1__116_1576___
# MMLCHVSLPALLLALLAAAGCCGGEPRIDPPSPIQAEDESCNTPTSDDDLPIVRVIRQIQ
# FPGGQPNRPGRPRPGQTPPSQQLPTDPTGQCANCVPVVSCSFQLNLVQGTCQLPGGSAGV
# CCPAQPQAAVGQGDSRLFKEPRRQVSMRTLSSQEVNEACQKGINVLTEVNALEDNLIRTN
# QVVPPETPAHGHLRFFRVTRSARQQHLQALQINQASRAMMSDFSLTPAQGTHGLRQFPVR
# NSILSNNCPVPPRCNPQAKYRSVDGTCNNLENSLYGRSETSFQRILPPVYDDGVSSPRTR
# SAAGGVLPSERVIASTVLVDRDDPDQQFTLSVMQWAQFIDHDLTHAPFARLSNNEGIDCC
# PNGQEATGATRHPECWPIRLPQDDPFYAPKGRFCMNFVRSMLGLNQECAFGYAEQMNQVT
# HWLDASNVYGSGQEEANRLRQGQGGLLQVSQNNLLPVNQASQGDCTARQRGGLCYHAGDS
# RVNEQPG
## C incomplete...
## swiss-prot? The best hit 32.58%...
## nr better
seqvect31 <- read.fasta("MW_PI_manual_seqs/GHHK01013740_1__116_1576___10_best_hits_nr.fasta")
getMedianMWpI(seqvect31)
```

```
## Warning in pmw(toupper(x)): Non allowed characters in seqaa
```

[1] 88.97207 [1] 7.689324

```r
### 31
# >GHHK01022199_1__1_345___
# KCDGDNDCWDHSDEEGCSDSSSNTPDACTSDQFKCASGHCIPGRSKCDGDNDCEDLSDEE
# GCSDSSSNTSDACTSDQFRCASGDCIRGRFKCDGYNDCGDLSDEEGCSDSSSNTP
## It's clearly a small fragment
## swissprot: well, not really... About 42% best hit. Vitellogenin receptor? Serine protease?
## nr: better (59%) but super varying length....
## no hope

### 32
# >GHHK01016883_1__1_894___
# NEWLPIIVGSNFMTSFGLNPIQRGFSFDYNFVINPTMNNEFATAAFRFGHSLVQGFIRLF
# TPDNQETTIRMRDHFNSPHIFQGQAGVIDMFVRSFTRQAIQKFDSFVTDDLSNHLFQTPS
# QNFGMDLMSLNLHRGRDHGIAPYNAMREICGLRRATSFADFNDQIPTDIVTRLSQMYAHV
# DDVDFFVGGMSEKPVSGGLLGWTFLCVVGDQFARAKKGDRFFYDIGGQPGSFNEVQLQEI
# RKASWARILCDNGDNIDAVQPLAFRLASRSFNAPQPCQSNVIPRVNLAAWSGERPQA*
### Well, at least C complete...
### Peptides: GFSFDYNFVINPTMNNEFATAAFR (24th, 11th from Met) to VNLAAWSGER (3 aa to the end!)
### No signal peptide...
### Then we take from Met.
seqGHHK01016883 <- read.fasta("MW_PI_manual_seqs/GHHK01016883_1__1_894___from_Met.fasta")
sapply(seqGHHK01016883, function(x) pmw(toupper(x))/1000)
```

GHHK01016883_1__1_894___ 32.14287

```r
sapply(seqGHHK01016883, function(x) computePI(toupper(x)))
```

GHHK01016883_1__1_894___ 6.372311

```r
## 33
# >GHHK01005898_1__2_316___
# KLATVSLPRTPSQDIERSKCITCGLEKAIDMLESDGGSAAGGVVFLISSGSPFPLTEYDV
# NLYHNLVVPRQVQVVPVLYPMTDRSPIPATGIDQLAKITGTRFYT
## It's clearly a fragment
## swissprot: nothing.
## nr: it's clearly a chloride channel, but length variance is great.

### 34
# >GHHK01006352_1__3_1550___
# GGGLVGGSNGGQGGGYGGGSSGGSSGGQGGGYGGGSTGGSSGGQGGGYGGGSTGGSSGGQ
# AGVLGGDSVGSSSGGQGGGIGGASGASGSINRSTGGSGGLGGFRQSGSGSSTQFGGGALG
# AGGGKPPMMMFPGELTPTGYRGHSFSSSSSSNRASQQSSSSSSHHSFGTHGGQLAGVFLQ
# QHGNLHQTGLLGGGHLQSGGKLDVFPVGGVHYSGSSHSSSSASQASNSASHQSATFHTFG
# GGPLNNAAGTQQRQDTAQAVADKDAYNIHPTQNKAETAAYTDQRTQNKADKADLNPDTSQ
# AIDIGGGQEQIVKDKVDEAYGGVGSSSWADNWNSFTNWGSNAARGISDAASRTVNTIRDT
# VVAGVNKVPSIWEKFKNALKGLGTSIHQGAEYCAHVLQQKANDMKSSAFLQKLQGKVEEG
# NEDVMRLFTVLGDKISNWTDQHANEGSIDEGLSGGGGGQNGQTVVLQKTKDFEREDFPDF
# FQDAQVMGEIEKLVQGGIIEQKEADLFTQQKQERP*
### Well, at least C complete
## Peptides: STGGSGGLGGFR (93+) to EADLFTQQK (almost the end)

### 35
```

```
#GHHK01014723_1__331_801___
# >GHHK01014723_1__331_801___
# MASAGAPTLVLGLLLFVGAANAIHRCPTDYELIQNECYRVVQDRKSVADAATFCEFESGT
# LASMSTLEAKEAVVDLVNRIAPGQYWTSGADMGGRFMWSNTEENINPRFKGRQFRPQTCV
# YLCSHTRMFWDRTCNQRLGFICQKNPELDITSVEAF*
## peptides: CPTDYELIQNECYR (26...) to NPELDITSVEAF (the very end)
## yeah!!!
## SignalP-5.0: Cleavage site between pos. 22 and 23: ANA-IH. Probability: 0.9094
seqGHHK01014723 <- read.fasta("MW_PI_manual_seqs/GHHK01014723_1__331_801_mature.fasta")
## so, let's take 23 to the end
sapply(seqGHHK01014723, function(x) pmw(toupper(x))/1000)
```

GHHK01014723_1_*331_801*___ 15.44729

```
sapply(seqGHHK01014723, function(x) computePI(toupper(x)))
```

GHHK01014723_1_*331_801*___ 5.694196

```
### 36
### another hemocyanin fragment... but a big one!
# >GHHK01015873_1__2_1768___
# LRDTAASIRGAMNETLFVYSLSSVITRNPRFRTIRVPAVTETFPSRFVPSSVIDRARALT
# NRANNNRNTDLTQPLVVNHDQDFSGLRGRLENRVSYWREDYGLNAHHWHWHLVYPTDVRT
# VKSPDRKGELFYYMHSQIVARYDMERLSVGLPRVLKLDSFREPILEGYFSKLRFDNADTN
# PNRIQVPGLHPNATLWGARQDNTRLSNYTRMQTFIPVDVGELEMWSNRLLDGIHQGFFIS
# NKGERVLLSDDVDITDGQQKRGVDIIGDAFEADQNISVNYRLYGDLHNFGHVVISSCHDP
# DGTHGENLGVMSDSAVAMRDPVFYRWHKYVDWVFQQYKATQPSYTKAQLELPGVNITRIG
# VATGNLADEIHTGWNRRLFEASRGIDFGTTQSVQLNLQHLDHKPFDYHILVTNSTPGPKQ
# VYVRIFLAPKFNQNQTSVQMPLNEQRLLWTEMDKFVFNLKPGQNHIKRASSLSSVGIPGE
# LTFRQLEQGLRQPGDTRPAADAQEDFCGCGWPQHLLVPRGRPEGMIFQVFAMLTDFELDR
# IPRVSGSRQCAGAASYCGVLDERYPDKRPMGFPFDRLPPRELRTPQQQV
## swissprot: many hits; identity not great...
seqvect37 <- read.fasta("MW_PI_manual_seqs/GHHK01015873_1__2_1768___best_hits.fasta")
getMedianMWpI(seqvect37)
```

[1] 78.89189 [1] 6.110083

```
### 37
# >GHHK01001186_1__1_465___
# GHPRKTQLEMTGPGQYRATFLPDDCGKYRVGVRYNDEELPSSPFPVQVFATGKADKCEIT
# EGISHALNTGEEYCISVNAKNAGHGAVTCRIRSTSGSDLDIDITDNGDGTFSIYYTVEDA
# GDYTLAVKFGGQPVPQGFYTFTAQESSESYPAPGT
## It's a fragment...
## swissprot: only 1, Dme, 40%
## nr better, >70% identity
seqvect38 <- read.fasta(file = "MW_PI_manual_seqs/GHHK01001186_1__1_465___10_best_nr.fasta")
getMedianMWpI(seqvect38)
```

[1] 239.1079 [1] 5.757872

```
### 38
# >GHHK01021786_1__1_414___
# ASSSQLEKSYELPDGQVITIGNERFRCPETLFQPSFIGMEAAGIHETCYNSIMKCDVDIR
# KDLYANTVLSGGTTMFPGIADRMQKEISALAPPTMKIKIIAPPERKYSVWIGGSILASLS
# TFQQMWISKQEYDESGPG
## it's another actin
seqvect39 <- read.fasta(file = "MW_PI_manual_seqs/GHHK01021786_1__1_414___10_best_hits.fasta")
```

```
getMedianMWpI(seqvect39)
```

[1] 41.81435 [1] 5.301928

```
### 39
# >GHHK01005070_1__1_1632___
# AYFREKMNEGEFVYALYVAVTHSDLTEDVVLPPLYEVTPHLFTNSEVINQAYSAKMRQTP
# GRFQMDFTGSKKNPEQRVAYFGEDIGMNSHHVHWHMDFPFWWDGYKIDRKGELFFWVHHQ
# LTARFDAERLSNHLPVVDELYWDRPIYEGFAPHTTYRYGGEFPSRPDNKFFEDVDGVARI
# RDMKIIESRLHDAIDHGYIVDSEGHNINLDAEHGIDILGDVIESSAYSPNVQYYGSLHNT
# AHVMLGRQADPHGKFNMPPGVMEHFETATRDPSFFRLHKYMNNIFKEYKDTLPSYTKEEL
# GYANAEITSLGIDGELTTFFEDFEFDLINAIDDTETIDDVPITTHVSRLNHEDFTFNIEV
# KANTDEAATVRIYICPKYDANHIEYTLDEARWGCIQLDKFWTQLHAGSNTIVRKSSDSSV
# TIPDRTPFATLIKEADDAVTSGSSLPSHNSRGCGLPQRLLLPKGNTEGVDFELFVSITSG
# DDAVISDLVSNDHGGNYGYCGIKGQKYPDKRAMGYPLDRHVDDDRLFKQPNIKWTTVKVF
# FRE*
## another piece of hemocyanin
seqvect40 <- read.fasta("MW_PI_manual_seqs/GHHK01005070_1__1_1632___10_best_hits.fasta")
getMedianMWpI(seqvect40)
```

[1] 75.53213 [1] 5.441646

```
## 40
# >GHHK01010725_1__2_352___
# RDISTMLFSPVLLNIAFRATALTTLLVVASGSLTAVSNEGTSSVEKCTLTFEVVYPEDED
# KDPYPICLSHCPAKNEVWFDGVAYKQFCELNRYNMMVEATFRNGELQSCKMVPLLTD
## A small fragment...
## swissprot: nothing
## nr: the best hit 33% identity; the best identity 41%... lengths 331-555 aa.

### 41
# >GHHK01019391_1__170_1213___
# MTFDKMRAAFKNFFDECAYRLARSSDSSNVIKVEPADTNSTMYSAGKNISTFSLTNKMAE
# KDSYIAALEKKLAELSGIEVDQIRKNQLANAASEAASIQQMAKYVAGITVEQAGKALQPS
# VLHPQIGLIFDHIKAELGEEKGEHVLPPLKYDYTGLEPSISGMIMEIHHTKHHQGYINNL
# KAAVAKLNEAQANGDIAASNALVPALKFNGGGHLNHTIFWTNMAPNTSGTAPEPAGELLQ
# AINDRFGSFQDFKDQFSAASVAVKGSGWGWLGYCPVNNKLDIATCQNQDPLQLTHGLVPL
# LGLDVWEHAYYLQYKNLRPDYVKAFFNVINWDNVAERYAKARADAGN*
## Peptides:KLAELSGIEVDQIR (starts 71) - AFFNVINWDNVAER (almost the end)
## No signal peptides...
## from the closest Met?
## Swissprot: mostly query 140-170 / subject 2-17
## nr: best hits query 58 (some 54) / subject 1
## the Met closest to the first peptide is 58. Matches!!!
seqGHHK01019391 <- read.fasta("MW_PI_manual_seqs/GHHK01019391_1__170_1213___closest_Met.fasta")
sapply(seqGHHK01019391, function(x) pmw(toupper(x))/1000)
```

GHHK01019391_1_*170_1213*___ 31.69045

```
sapply(seqGHHK01019391, function(x) computePI(toupper(x)))
```

GHHK01019391_1_*170_1213*___ 5.864704

```
### 42
# >GHHK01004467_1__1_759___
# EGIGVADPSGFGSKIDSGIGGGAGFGGPGGGAGGFGGTGGGAGGFGGTGGGAGGFGGTGG
# GAGGFGGTGGGAGGLGGTGAGGYGGSSGGGAGTGGFGGTGGTGGFGGSGGGGLGGGNLGG
# GTGDLSTAIGGGGVPGVDYPTLAAVPDTGFDCSGRTPGYYADTGAEARCQVFHICQFDDR
# HDSFLCPNGTVFNQQYFVCDWWYNFSCDEAEGYYFLNEGIGVADPSGFGSKIDSGIGGGA
# GFGGPGGGAGGFG
## a fragment...
## swiss-prot: 2 hits; coverage ~25%
## nr: well, identity ~70, high length variance...
seqvect43 <- read.fasta("MW_PI_manual_seqs/GHHK01004467_1__1_759___10_best_nr.fasta")
getMedianMWpI(seqvect43)
```

[1] 30.95433 [1] 4.376396

```
### 43
# >GHHK01000352_1__3_2063___
# QTYNMRILLVCLLVAGAVAWPQFVDDITNALLSDASGPSLAKRQQDINRLVYRINEPLGF
# PELKAAADNFNPVADTSLYSDGGKAVEALVHELQDGRLLEQHHWFSLFNTRQREEALMLF
# DVFMHSKTWETAVNNAAYFREKMNEGEFIYAVYAAVIHSDLGAGIVLPPLYEVTPHMFTN
# SEVISKAYTAQMTQTPGKFNMDFTGSKKNPEQRVAYFGEDIGLNIHHVTWHLDFPFWWQD
# SYGYHLDRKGELFFWAHHQLTTRFDNERLSNHLGMVDELYWDRPIVEGFAPHTTYRYGGE
# FPARPDNVDFEDVDGEIRVRDMIIHESRIRDAIAHGYITAADGSKIDIRNNEGIDHLGDI
# IESSLYSPNIEYYGGLHNDAHIILGRQSDPHGKFNLPPGVMEHFETATRDPAFFRLHKYM
# DNLFKEYKDTLPAYTKDELEFPGISLNSVRVDGVLETFFEDYEFDLGNAVDSNPNIADVS
# VSASVSRLNHKRFALKFEVINNDSVEKHGVVRVFLCPRRDENGIIFSFEEGRWHCIEMDK
# FWTKLASGNNKISRSSRDFSVSVPDVPSFKSLINTADQAVAKGTPLGLEEFDRSCGIPDR
# LLLPKGNSRGMEYVLAVAVTDGEADIQHDLLEKSEAHSHAQCGVHGEKYPDHQPMGFPLD
# RRIEDERIMLGSPNIKYTIVSVTFKG*
# wow, it's a huge hemocyanin fragment!
seqvect44 <- read.fasta("MW_PI_manual_seqs/GHHK01000352_1__3_2063___10_best_hits.fasta")
getMedianMWpI(seqvect44)
```

[1] 75.53213 [1] 5.441646

```
### 44
# >GHHK01011502_1__3_539___
# QSKSGLAQWWTDTVSSSKMRVPATIAVTMATLVALASSTSETGPSMHVGTAGCLSWCRHR
# HNPTEFFCCKADPNSKFHHGQCPQRIVANNGMEESYCQIDHHCQPNEKCCSSDYNKQSTC
# VAAVTDRHQPRTPPPVINARFGGDPVIVGNNFDDLGFDVTYRRGNEGWGSMGDMEPTE*
## Peptide coverage: HNPTEFFCCK (61+) to RGNEGWGSMGDMEPTE (it's the very end)
## Signal peptide probability ~0.58
seqGHHK01011502 <- read.fasta("MW_PI_manual_seqs/GHHK01011502_1__3_539___from_closest_Met.fasta")
sapply(seqGHHK01011502, function(x) pmw(toupper(x))/1000)
```

GHHK01011502_1_*3_539*___ 14.89938

```
sapply(seqGHHK01011502, function(x) computePI(toupper(x)))
```

GHHK01011502_1_*3_539*___ 6.027165

```
## 45
# >GHHK01013689_1__145_798___
# MGGLLLISLLAAVSTAVNGQITSAHVGTSGGSSNFQGQGAALSENVGPDGSRQGECAYAD
# SNGQQIQVRYSQQQGREAQYRLIKGSSGTNPAAAYEQCLQQYRASQQAGAAFIPDFNSFN
# PFAGGGFDFGSLANAAQAFAGGFGGAVDPGFQQAAAGGVRHAHDQVGAAAHVVPSSLVEQ
# MEVLRRQNLQLQQNVFEMQQRNAELHNRLAGRFRRGL*
## Peptides: GSSGTNPAAAYEQCLQQYR (85+) to QNLQLQQNVFEMQQR (almost the end)
## Swissprot: nothing; nr: best hit 33% identity
## Cleavage site between pos. 19 and 20: VNG-QI. Probability: 0.8464
## Let's take without signal sequence...
seqGHHK01013689 <- read.fasta("MW_PI_manual_seqs/GHHK01013689_1__145_798___mature.fasta")
sapply(seqGHHK01013689, function(x) pmw(toupper(x))/1000)
```

GHHK01013689_1__*145_798*___ 20.9326

```
sapply(seqGHHK01013689, function(x) computePI(toupper(x)))
```

GHHK01013689_1__*145_798*___ 7.144122

```
### 46
# >GHHK01019869_1__63_1292___
# MFKLIIAVAALAVLTTPLQVAARSRQSRQASSGPASSIAINDLSDLAGLFEGGVNSQQGG
# DCECVPYYQCKEGVIITDGEGVIDIRFGNSLNDTGSTRLNSHSQCQNFLDVCCQHPNTAV
# TPGPGPADQYLAKCGRRNPTGVNARVSGFTATQAQFGEFPWMAAILQTEFVGAEEVNLYV
# CGGSLIYPDVVLTAAHCVQSWQNTPTVLKVRLGEWDTQRTYELYTHVDRAVSKVIVNNQY
# NPGSLSNDFAILLLETPVALTHHIDTVCLPDVYQNVEPTKCFVTGWGKNEFGKEGEFQNI
# LKKVSLPLVSHPDCEKALRTTRLGKYFNLHSTFSCAGGGIAGQDACNGDGGSPLVCPLLN
# DHATYVQVGIVAWGIGCGEAGIPGVYADVTKGITWVNQELAKLPYTRTG*
## Peptides: EGVIITDGEGVIDIR (72+) to GITWVNQELAK (it's basically the end)
## SignalP: Cleavage site between pos. 22 and 23: VAA-RS. Probability: 0.6109 (signal peptide likelihoo
## nr blastp: 84/17 for T. longiramus (88%); the others 1/1...
## Take the mature output?
seqGHHK01019869 <- read.fasta("MW_PI_manual_seqs/GHHK01019869_1__63_1292___mature.fasta")
sapply(seqGHHK01019869, function(x) pmw(toupper(x))/1000)
```

GHHK01019869_1__*63_1292*___ 41.64332

```
sapply(seqGHHK01019869, function(x) computePI(toupper(x)))
```

GHHK01019869_1__*63_1292*___ 5.373155

```
### 47
# >GHHK01010257_1__35_424___
# MALDGTFVLKSNDNYDAWLQAVGVPAELAAKMCAAKPKMTVTTTDNTLTVKTIAGEKEFD
# NTIVFGKDSVIDVAGLKYTVNVKVTDKGYSGTVAMGGKNGTLEVVADADGFTQTIVVDGV
# TGKRVYTRS*
## Is it complete?!
## Peptide coverage: ALDGTFVLK (2+) to GYSGTVAMGGK (it's not the end, actually)
seqGHHK01010257 <- read.fasta("MW_PI_manual_seqs/GHHK01010257_1__35_424___full.fasta")
sapply(seqGHHK01010257, function(x) pmw(toupper(x))/1000)
```

GHHK01010257_1__*35_424*___ 13.60335

```
sapply(seqGHHK01010257, function(x) computePI(toupper(x)))
```

GHHK01010257_1__*35_424*___ 5.334147

```
### 48
# >GHHK01006134_1__2_307___
# TCRAPNRKMKWFLVLTAVVAICAADDTAVKQQAINRLLLKVTEPIRSYFTDLKDAATKWN
# PRDHEDHCKDGGKAVAALLDEIEAGRVLQQKAIFSLFDERQR
## Another piece of hemocyanin...
## Only 2 hits, 45 and 35% identity, length 657/566...
## Nope

###49
# >GHHK01013584_1__22_708___
# MVRWLPLESNPEVMNKFLSGMGVPDSVKVCDVLGLEAELLAMVPRPVYALLLLYPLTSKS
# EEFKEQQESGIESAGQDLAEDLYYMKQFVGNACGTVALMHALANNSDKIEVADGPLKEFL
# EKTKELDPEERGHALEDDESISAVHEDCAAEGQTEAPDREHKLDTHFIALVNVGDRLYEL
# DGRKKFPINHGPTSEENFLIDGASVLRDFMDRDSDETRFAVVALTAAE*
## Is it complete?
## Peptides: WLPLESNPEVMNK (4+, so the beginning) to DFMDRDSDETR (10 aa to the end, so probably the end
seqGHHK01013584 <- read.fasta("MW_PI_manual_seqs/GHHK01013584_full_seq.fasta")
sapply(seqGHHK01013584, function(x) pmw(toupper(x))/1000)
```

GHHK01013584_1__22_708___ 25.28512

```
sapply(seqGHHK01013584, function(x) computePI(toupper(x)))
```

GHHK01013584_1__22_708___ 4.465867

```
### 50
# >GHHK01014666_1__103_879___
# MELNAILNGFVVVTVLGSYCLPLVCSTSPHRLIMPTDNVAAPSSSASYSAPLPPHSAPLP
# PLPSAFSKEAFELGASSVVNLWEHGLRLQGRERARVSISSAAEPPTPLTVVAPDSGSMLN
# LVPQDAPHPLYHDAPAYTALRHAFLLDSFVQGAVDPQDKAMSTDAGLTVENMNGRALVFK
# RDQQGTLSVNGIPVIKQQRLTDGTQLFVVDGLLFNHQEDVKKAFNRLLEENARDGSSRCP
# FGPCQPQVAPVQPVNPQD*
## looks complete...
## Peptides: EAFELGASSVVNLWEHGLR to LTDGTQLFVVDGLLFNHQEDVK
## not from beginning and not to the end
```