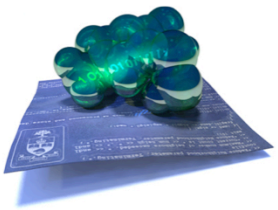
A
Bioinformatics
Course

Sequence Analysis: Composition



Boris Steipe

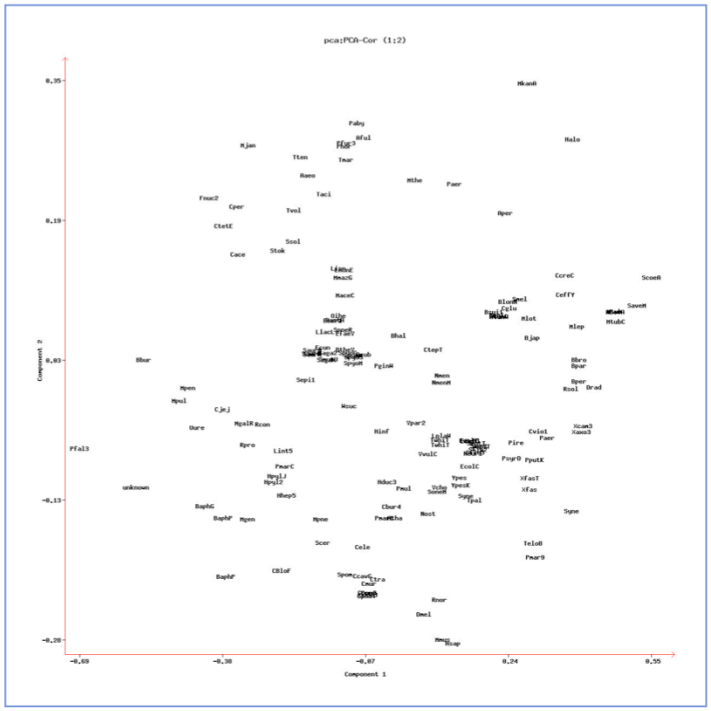
department of biochemistry – department of molecular genetics
University of Toronto

Sequence Composition

The composition of its sequences is characteristic for an organism. A large-scale analysis of 148 complete genomes confirms species-specific compositional bias in nucleotide as well as protein sequence!

This has consequences for:

- gene identification
- species-specific similarity scoring functions
- analysis of horizontal gene transfer



Principal components analysis of amino acid compositions

Dumontier M, Michalickova K, Hogue CW. (2002) Species-specific protein sequence and fold optimizations. BMC *Bioinformatics* 3:39

The composition of biological sequences is both non-random, and characteristic for individual species.

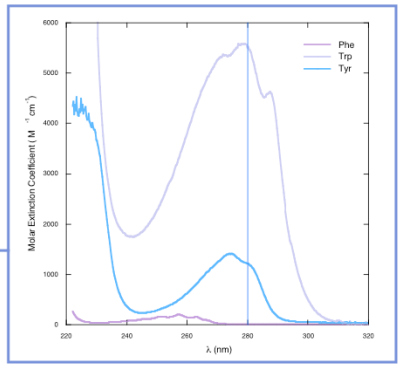
Global trends may be due to the time a particular codon was "assigned" as the genetic code evolved, the number of codons for an amino acid, and the metabolic cost of synthesizing it (resp. the fitness cost in assimilating it). These factors are not independent!

But with respect to species-specific differences, these may simply be contingent on chance characteristics of the biochemical machinery of replication and translation. Just as an enzyme has a precise an characteristic turnover rate, e.g. a polymerase assembling polynucleotides will have slightly different rate-constants for the individual nucleotides. Notwithstanding, minor phyico-chemical effects play a role as well, for example, thermophiles have somewhat increased counts of charged amino acids since salt-bridges in proteins provide stabilization that is lost to the reduced hydrophobic effect at elevated temperatures.

Sequence Composition

Some properties of biomolecules can be simply summed over their components: they depend only on the amino acid (or nucleotide) composition of the molecule.

Molecular weight
Extinction coefficient
Isoelectric point



$$e_{280}\ (M^{-1}\ cm^{-1}) = (5{,}500)\#Trp + (1{,}490)\#Tyr + (125)(\#Cys\text{–}Cys)$$

Gasteiger *et al.* (2005) Protein Identification and Analysis tools on the ExPASy server. *The Proteomics Protocols Handbook*, Human Press 571-607

(# here means "number of")

A protein's isoelectric point depends on the pK values of the amino acids; the pK values characterize the propensity for an amino acid sidechain to dissociate, which in turn depends on how energetically favourable dissociation is. For example: since a negatively charged amino acid will be stabilized in a positive electrostatic field, such a field will shift a pK value **down**. This means the pH value at which the side chain will be 50% ionized is lower, or in other words, in a positive electrostatic field the concentration of protons must be higher to keep a proton associated to the sidechain.

Compositional properties of nucleic acids include hybridization temperature and helix structure.

Sequence Composition: Data

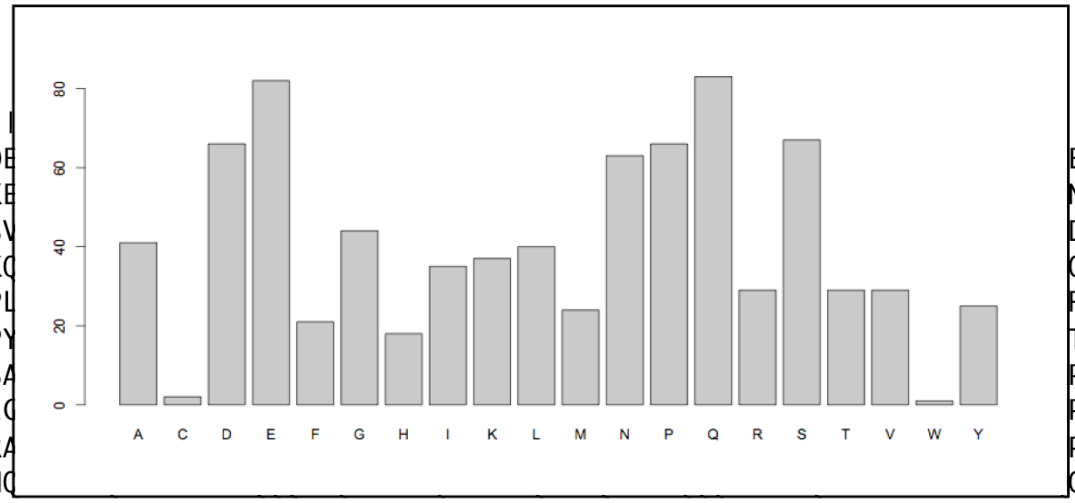
```
>gi|6325066|ref|NP_015134.1| Nab3p [Saccharomyces cerevisiae S288c]
MSDENHNSDVQDIPSPELSVDSNSNENELMNNSSADDGIEFDAPEEEREAEREEENEEQHELEDVNDEEE
EDKEEKGEENGEVINTEEEEEEEHQQKGGNDDDDDDNEEEEEEEDDDDDDDDDDDEEEEEEEEEGND
NSSVGSDSAAEDGEDEEDKKDKTKDKEVELRRETLEKEQKDVDEAIKKITREENDNTHFPTNMENVNYDL
LQKQVKYIMDSNMLNLPQFQHLPQEEKMSAILAMLNSNSDTALSVPPHDSTISTTASASATSGARSNDQR
KPPLSDAQRRMRFPRADLSKPITEEEHDRYAAYLHGENKITEMHNIPPKSRLFIGNLPLKNVSKEDLFRI
FSPYGHIMQINIKNAFGFIQFDNPQSVRDAIECESQEMNFGKKLILEVSSSNARPQFDHGDHGTNSSSTF
ISSAKRPFQTESGDMYNDDNGAGYKKSRRHTVSCNIFVKRTADRTYAIEVFNRFRDGTGLETDMIFLKPR
MELGKLINDAAYNGVWGVVLVNKTHNVDVQTFYKGSQGETKFDEYISISADDAVAIFNNIKNNRNNSRPT
DYRAMSHQQNIYGAPPLPVPNGPAVGPPPQTNYYQGYSMPPPQQQQQQPYGNYGMPPPSHDQGYGSQPPI
PMNQSYGRYQTSIPPPPPQQQIPQGYGRYQAGPPPQPPSQTPMDQQQLLSAIQNLPPNVVSNLLSMAQQQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQ
QQQQQQQQQQQQQQPAGNNVQSLLDSLAKLQK
```

```
> cat(readLines("nab3.fa"), sep = "\n")
```

Let us discuss a simple example of composition analysis for a given protein sequence, the yeast Nab3 protein.

The atypical distribution and clustering of particular amino acids suggests consequences for folding and interactions of the encoded protein.

Sequence Composition: Counts

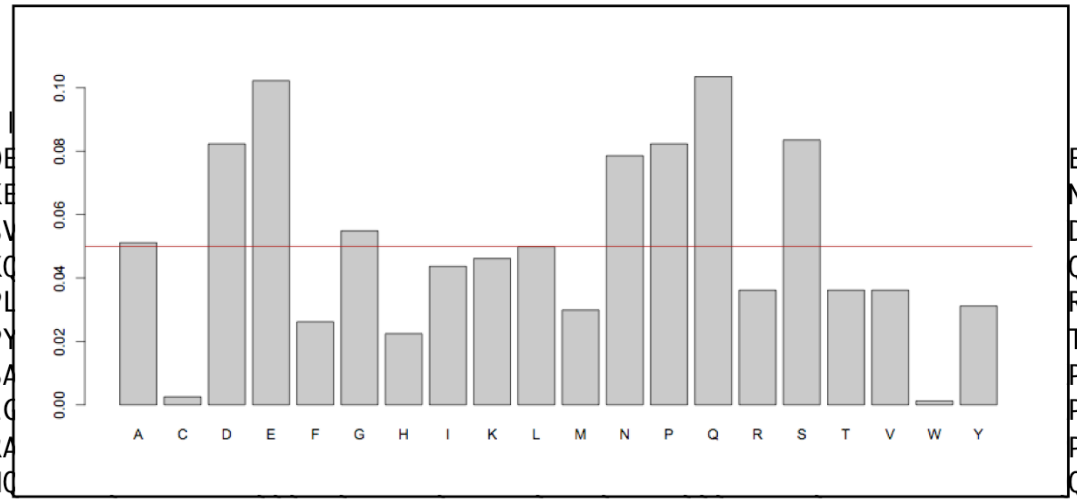



```
>gi
MSDE                                                                EE
EDKE                                                                ND
NSSV                                                                DL
LQKQ                                                                QR
KPPL                                                                RI
FSPY                                                                TF
ISSA                                                                PR
MELQ                                                                PT
DYRA                                                                PI
PMNQ                                                                QQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQ
QQQQQQQQQQQQQQPAGNNVQSLLDSLAKLQK
```

```
> nab3 <- read.fasta("nab3.fa")
> obsData <- table(nab3)
> barplot(obsData)
```

Here, **R** was used to tabulate the counts of the different amino acids in the sequence. The values are shown in a barplot, ordered by one-letter code, alphabetically. This ordering makes it hard to evaluate trends quickly.

Sequence Composition: Frequencies



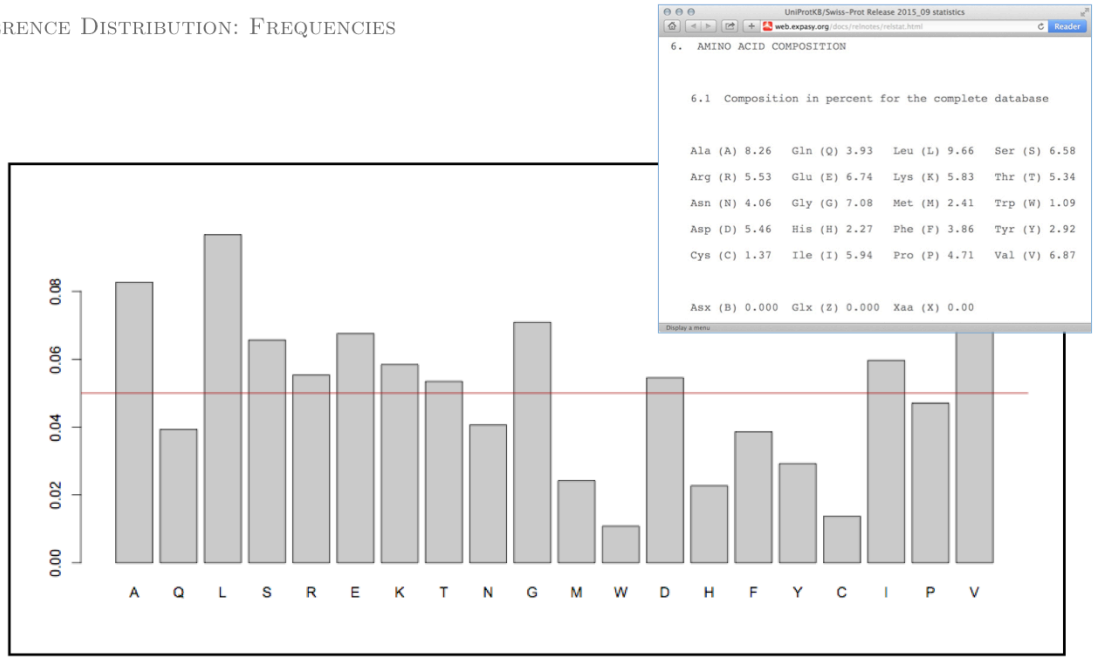

```
>gi|
MSDE                                                                EE
EDKE                                                                ND
NSSV                                                                DL
LQKQ                                                                QR
KPPL                                                                RI
FSPY                                                                TF
ISSA                                                                PR
MELQ                                                                PT
DYRA                                                                PI
PMNQ                                                                QQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQ
QQQQQQQQQQQQQQPAGNNVQSLLDSLAKLQK
```

```
> obsData = obsData / sum(obsData)
> barplot(obsData)
> abline(h=1.0/20, col="#AA0000")
```

Converting the counts into frequencies, and ading a line to display an expected value, gives us a little more information. We can begin to understand which amino acids are over- and underrepresented.

But what is the “expected value”? The red line in the plot is simply taken as 5% (0.05) – the expected frequency if all 20 amino acids would occur in equal amounts.

Reference Distribution: Frequencies



Data taken from UniProt release notes. Database average over all proteins.

Note: it is often not clear what the best reference dataset is, and the information from comparison depends critically on the choice of reference.

```
> barplot(refData)
> abline(h=1.0/20, col="#AA0000")
```

Equal probability is a poor assumption. But we can access the large sequence databases to evaluate the frequency of amino acids in proteins in general.

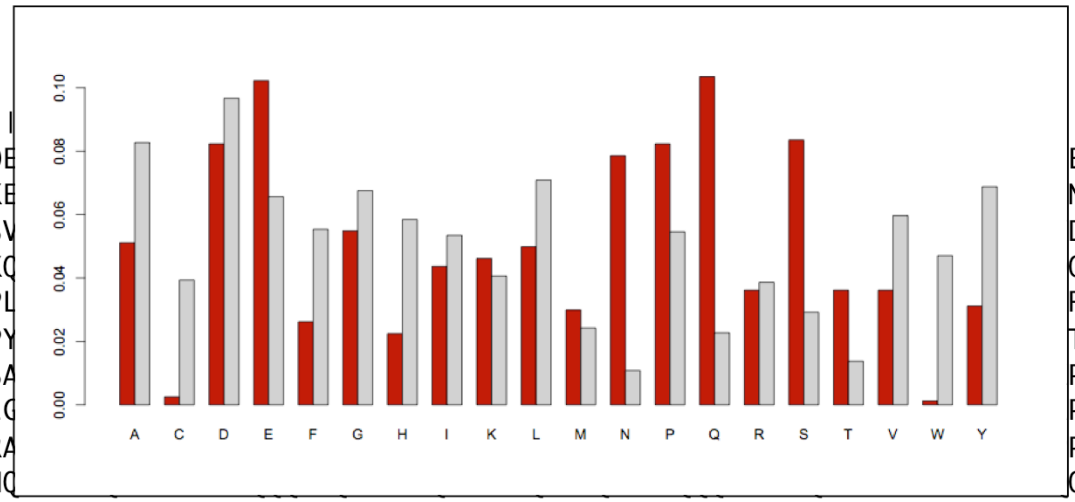
Is this a better assumption? Probably, but it has its own problems. It does not distinguish between highly- and poorly- expressed sequences.

Other considerations may be to look at an organism's total protein, or distinguish between membrane- and cytoplasmic proteins, nuclear proteins, secreted proteins. Or to take the metabolic cost of amino acids into account. Or other biologically motivated distributions we can come up with.

You should note that the definition of an expected distribution is at least as important as to compile the observations.

The inset frequencies are database averages.

Sequence Composition: Frequencies Compared



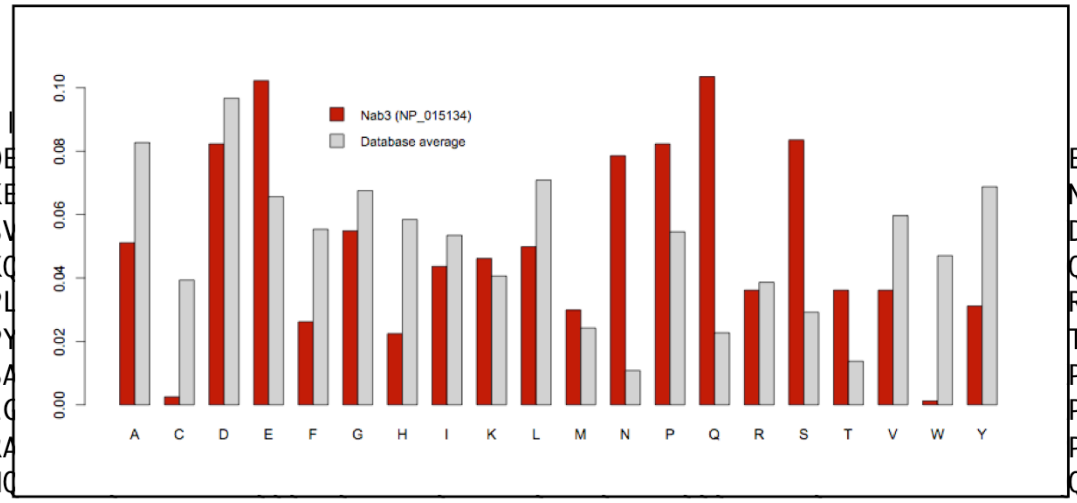

```
>gi|
MSDE                                                                EE
EDKE                                                                ND
NSSV                                                                DL
LQKQ                                                                QR
KPPL                                                                RI
FSPY                                                                TF
ISSA                                                                PR
MELQ                                                                PT
DYRA                                                                PI
PMNQ                                                                QQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQ
QQQQQQQQQQQQQQPAGNNVQSLLDSLAKLQK
```

```
> barplot(rbind(obsData, refData), beside=TRUE, col=c("#BB0000", "#CCCCCC"))
```

Given this information, we can compare database values with the values in our sequence.

Wait: what do the bars represent? Which is which?

Sequence Composition: Frequencies Compared



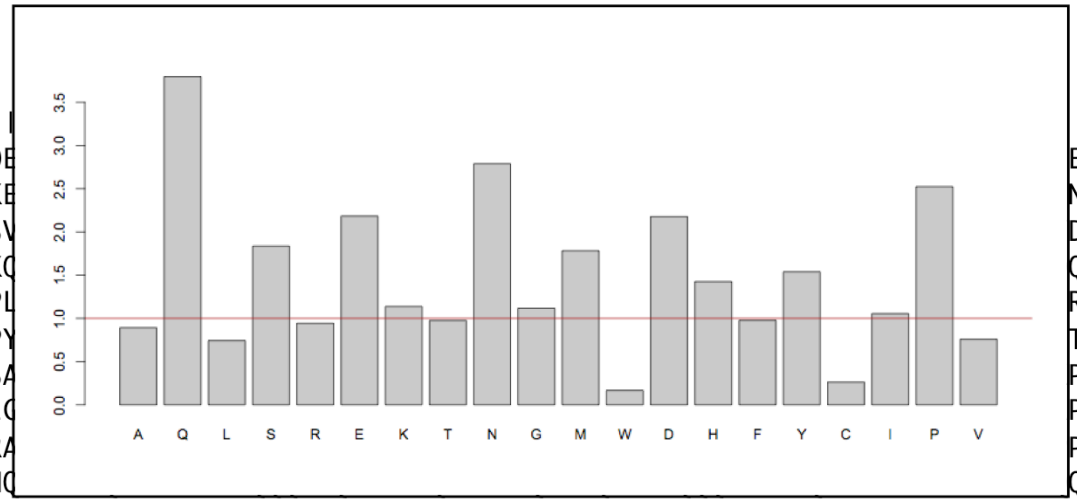

```
>gi|
MSDE                                                                    EE
EDKE                                                                    ND
NSSV                                                                    DL
LQKQ                                                                    QR
KPPL                                                                    RI
FSPY                                                                    TF
ISSA                                                                    PR
MELQ                                                                    PT
DYRA                                                                    PI
PMNQ                                                                    QQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQ
QQQQQQQQQQQQQQPAGNNVQSLLDSLAKLQK
```

```
> legend (x=14, y=0.1,
          legend=c("Nab3 (NP_015134)", "Database average"),
          bty="n", cex=0.9,
          fill=c("#BB0000", "#CCCCCC"))
```

We need to add a legend ...

The amino acids that have a larger or smaller than average frequency are becoming apparent. But absolute values are not well suited for this type of comparison. It is much more convenient to express such differences as ratios.

Sequence Composition: Ratios



>gi|
MSDE EE
EDKE ND
NSSV DL
LQKQ QR
KPPL RI
FSPY TF
ISSA PR
MELQ PT
DYRA PI
PMNQ QQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQ
QQQQQQQQQQQQQQPAGNNVQSLLDSLAKLQK

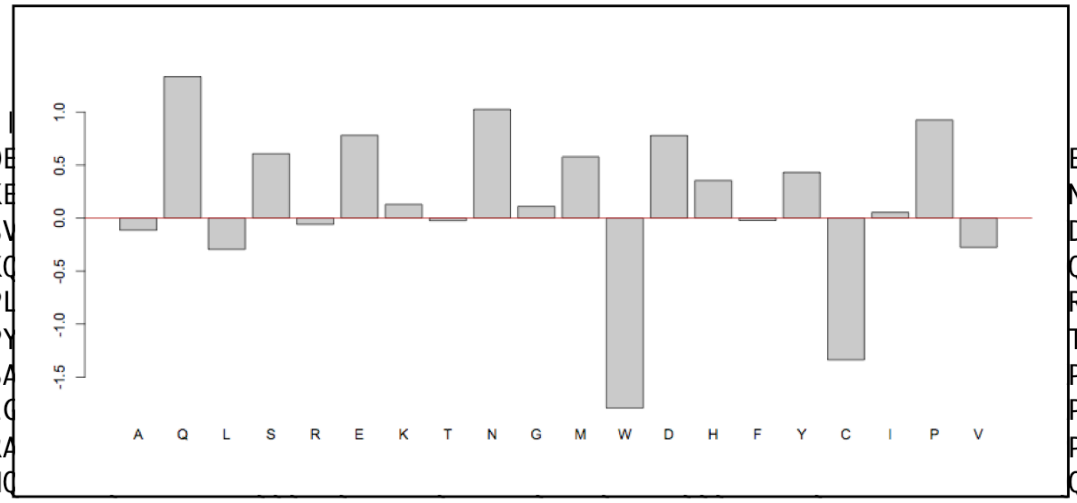
```
> barplot(obsData / refData)
> abline(h=1, col="#AA0000")
```

Now we have a ratio of one if the frequencies are the same, a ratio of 0.5 if the observed frequencies are half that of the database reference, and a ratio of 2.0 if they are double. Therefore the values mean: how much more likely do we observe an amino acid than we expect it.

But we really should make ½ and 2 times the expected frequency give the same distance on the plot – after all, which is which depends only on our arbitrary choice of which distribution should be in the numerator and denominator of the fraction.

Therefore we express such relationships as log(ratio).

Sequence Composition: Log-Ratios

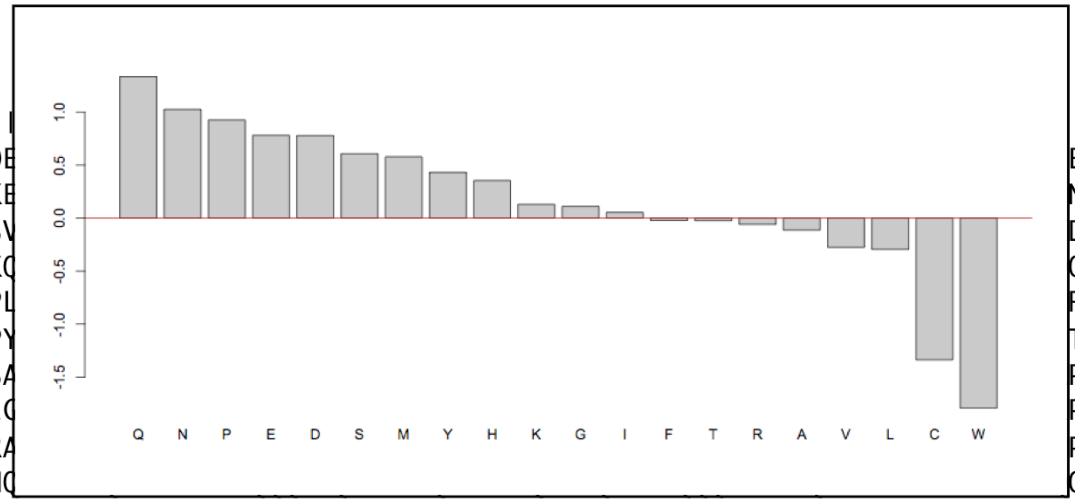



```
>gi
MSDE                                                                EE
EDKE                                                                ND
NSSV                                                                DL
LQKQ                                                                QR
KPPL                                                                RI
FSPY                                                                TF
ISSA                                                                PR
MELQ                                                                PT
DYRA                                                                PI
PMNQ                                                                QQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQ
QQQQQQQQQQQQQQPAGNNVQSLLDSLAKLQK
```

```
> barplot(log(obsData / refData))
> abline(h=0, col="#AA0000")
```

In a log ratio, when observed and expected frequencies are the same, the value is zero. Excesses are positive and depletions are negative. The same relative difference (2-fold more, 2-fold less) gives the same distance on the plot, regardless of whether the log-ratio is positive or negative in absolute terms

Sequence Composition: Log-Ratios Sorted

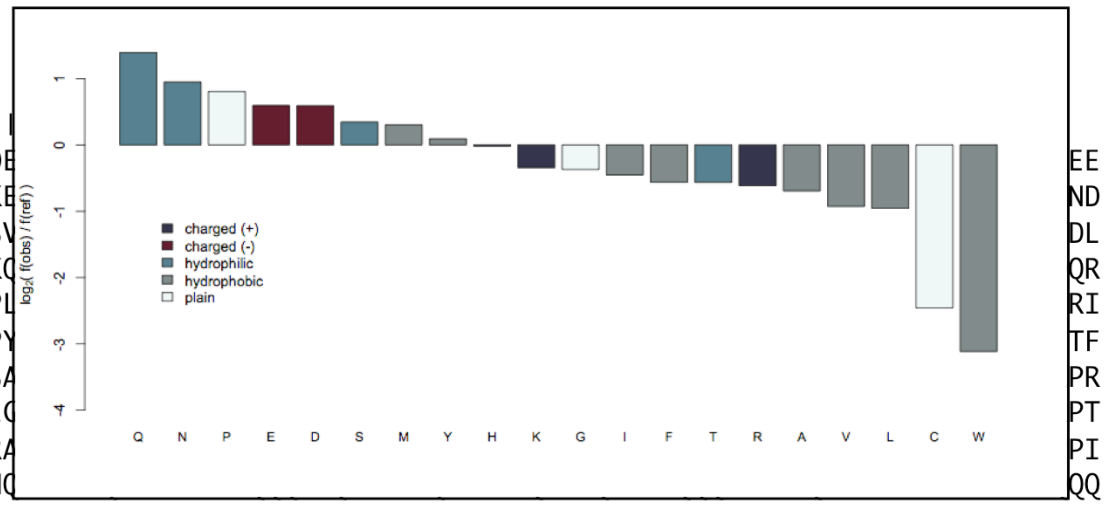


>gi|
MSDE EE
EDKE ND
NSSV DL
LQKQ QR
KPPL RI
FSPY TF
ISSA PR
MELQ PT
DYRA PI
PMNQ QQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQ
QQQQQQQQQQQQQQPAGNNVQSLLDSLAKLQK

```
> barplot(sort(log(obsData / refData), decreasing=TRUE))
> abline(h=0, col="#AA0000")
```

To further clarify what we are seeing here, we can sort by values.

Sequence Composition: Log-Ratios Sorted And Categorized

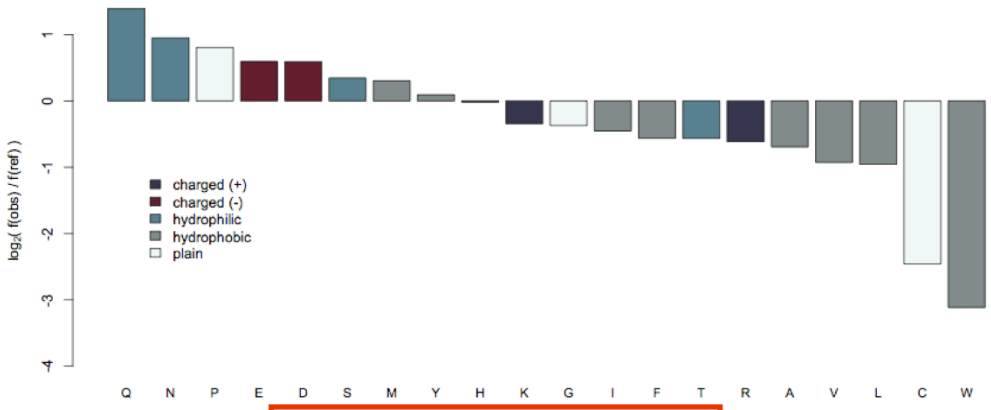


>gi|
MSDE EE
EDKE ND
NSSV DL
LQKQ QR
KPPL RI
FSPY TF
ISSA PR
MELQ PT
DYRA PI
PMNQ QQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQ
QQQQQQQQQQQQQQPAGNNVQSLLDSLAKLQK

Once we colour the resulig bars by property, we can extract important trends: our sequence has an excess of hydrophilic amino acids and of proline. This would be compatible with unstructured regions that can remain unfolded without aggregating. It also has an excess of negatively charged amino acids, with a depletion of postively charged amino acids. This would be consistent with binding to positively charged molecules. Hydrophobic (aggregation promoting) residues are depleted. That would be consistent with a protein that is natively, functionally unfolded over parts of its sequence.

From this we can build a mental image that this protein might bind to a large, positively charged polymer in a conformation-independent manner.

Sequence Composition: Interpretation



>gi|6325066|ref|NP_015134.1| Single stranded RNA binding protein; acidic ribonucleoprotein;
required for termination of non-poly(A) transcripts and efficient splicing; interacts with Nrd1p;
Nab3p [Saccharomyces cerevisiae]
MSDENHNSDVQDIPSPELSVDSNSNENELMNNSSADDGIEFDAPEEEREAEREEENEEQHELEDVNDEEEEDKEEKGEENGEVINTEEEEEEEHQQKG
GNDDDDDDNEEEEEEEEDDDDDDDDDDDDEEEEEEEEEEGNDNSSVGSDSAAEDGEDEEDKKDKTKDKEVELRRETLEKEQKDVDEAIKKITREENDN
THFPTNMENVNYDLLQKQVKYIMDSNMLNLPQFQHLPQEEKMSAILAMLNSNSDTALSVPPHDSTISTTASASATSGARSNDQRKPPLSDAQRRMRFP
RADLSKPITEEEHDRYAAYLHGENKITEMHNIPPKSRLFIGNLPLKNVSKEDLFRIFSPYGHIMQINIKNAFGFIQFDNPQSVRDAIECESQEMNFGK
KLILEVSSSNARPQFDHGDHGTNSSSTFISSAKRPFQTESGDMYNDDNGAGYKKSRRHTVSCNIFVKRTADRTYAIEVFNRFRDGTGLETDMIFLKPR
MELGKLINDAAYNGVWGVVLVNKTHNVDVQTFYKGSQGETKFDEYISISADDAVAIFNNIKNNRNNSRPTDYRAMSHQQNIYGAPPLPVPNGPAVGPP
PQTNYYQGYSMPPPQQQQQQPYGNYGMPPPSHDQGYGSQPPIPMNQSYGRYQTSIPPPPPQQQIPQGYGRYQAGPPPQPPSQTPMDQQQLLSAIQNLP
PNVVSNLLSMAQQQQQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQNISAKPSAPPMSHQPPPPQQQQQQQQQQQQQQQ
PAGNNVQSLLDSLAKLQK

These observations are entirely consistent with the annotated function of the protein, Nab3: a single-stranded nucleic acid binding protein. You would expect it to have

- Disordered regions that interact with disordered ligands;
- Negative charge that complements the positive charge of the exposed nucleobases in single-strand nucleic acid molecules;
- A reduced amount of aggregation-promoting residues to keep the disordered structure in solution in the cytoplasm.

Specifically, we can also hypothesize that the protein coats nucleobases and avoids the backbone, keeping the RNA from forming double-stranded secondary structure, perhaps even promoting melting in the first place.

It's remarkable how far we can (sometimes) get with considering composition alone.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA