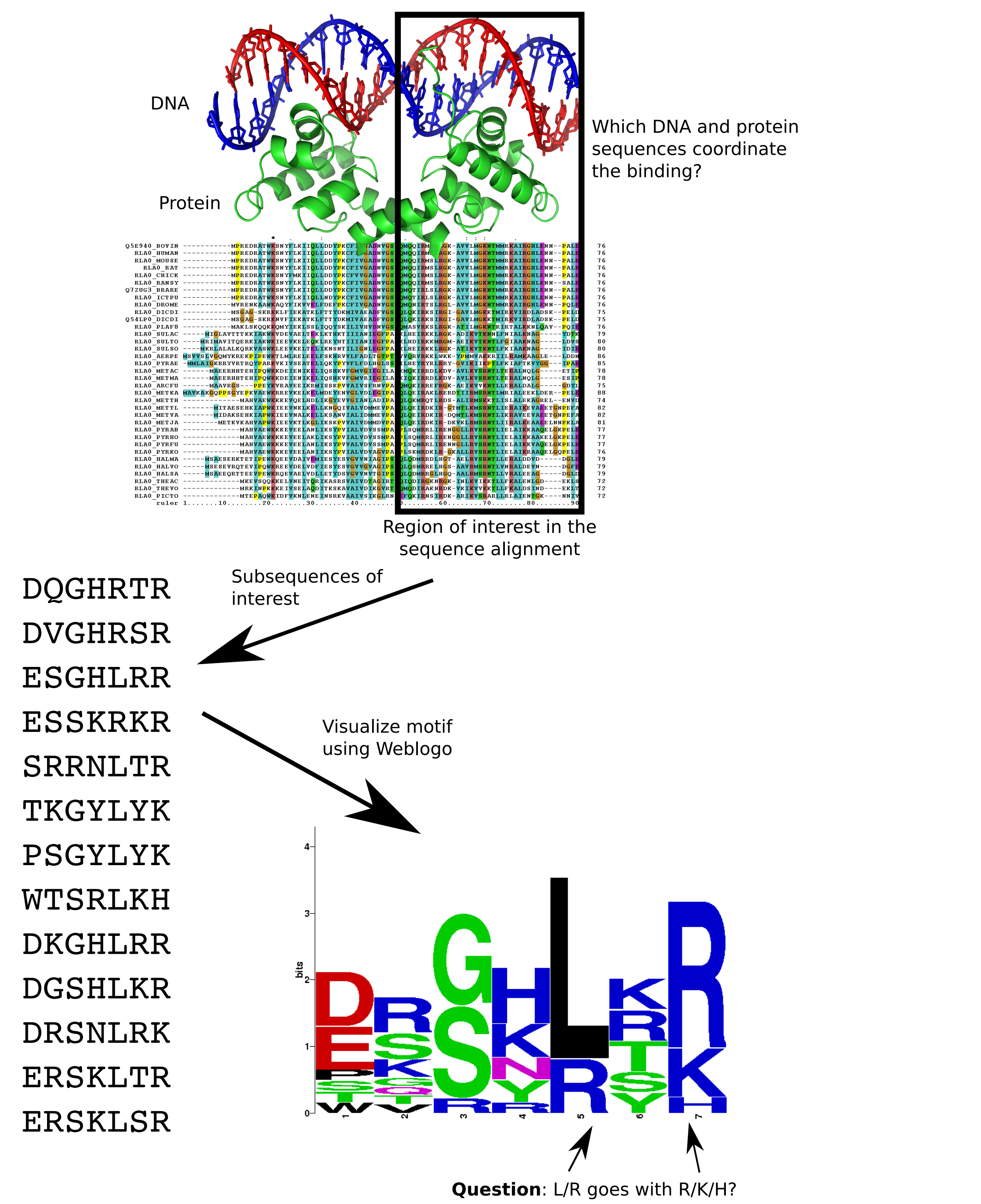


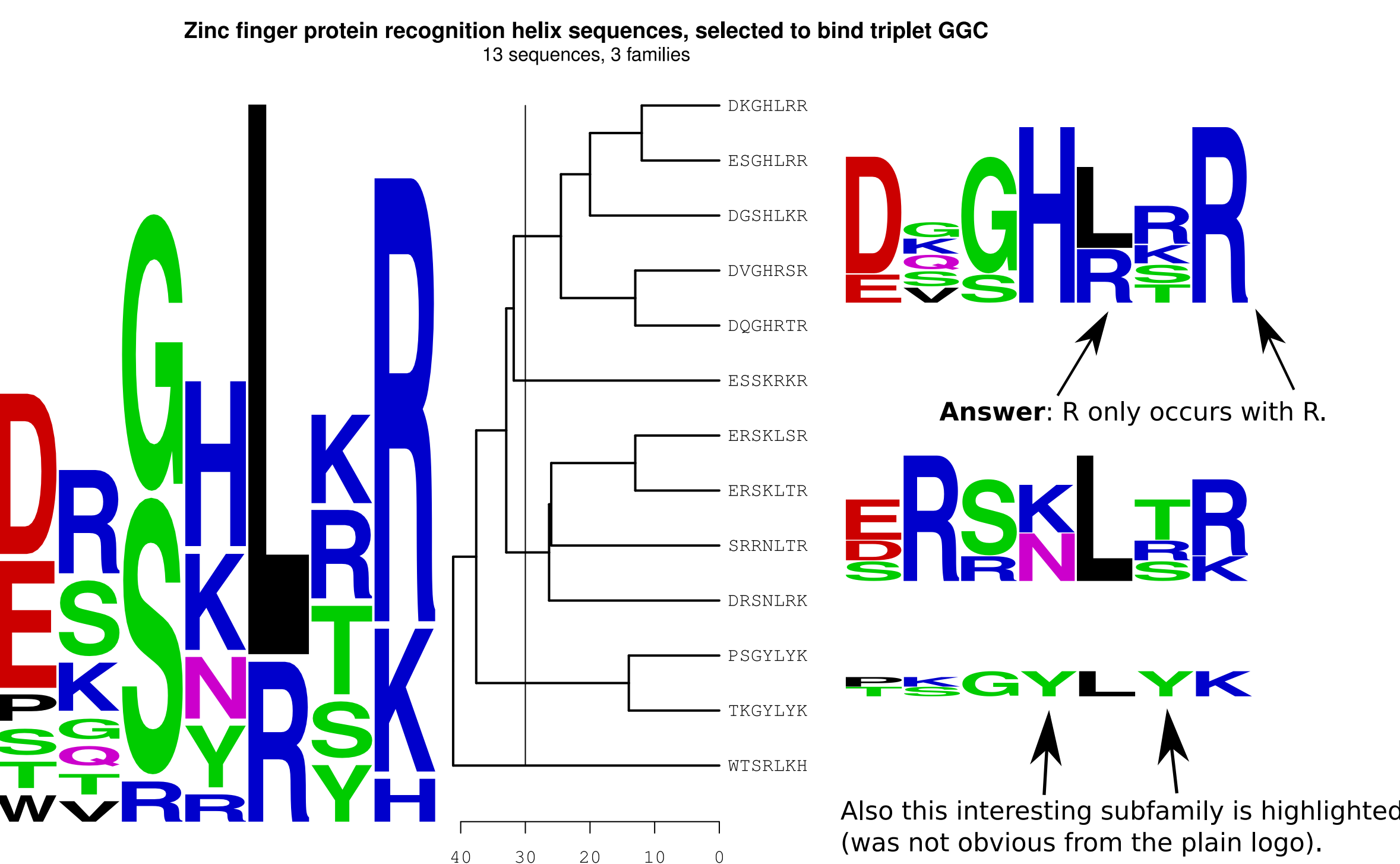
Sublogo dendrograms: visualizing correlation in biological sequence motifs

Toby Dylan Hocking - toby.hocking@etu.upmc.fr - useR! 2009 - July 7-10 - Rennes, France

Introduction: DNA-protein interactions motivate the study of sequence motifs

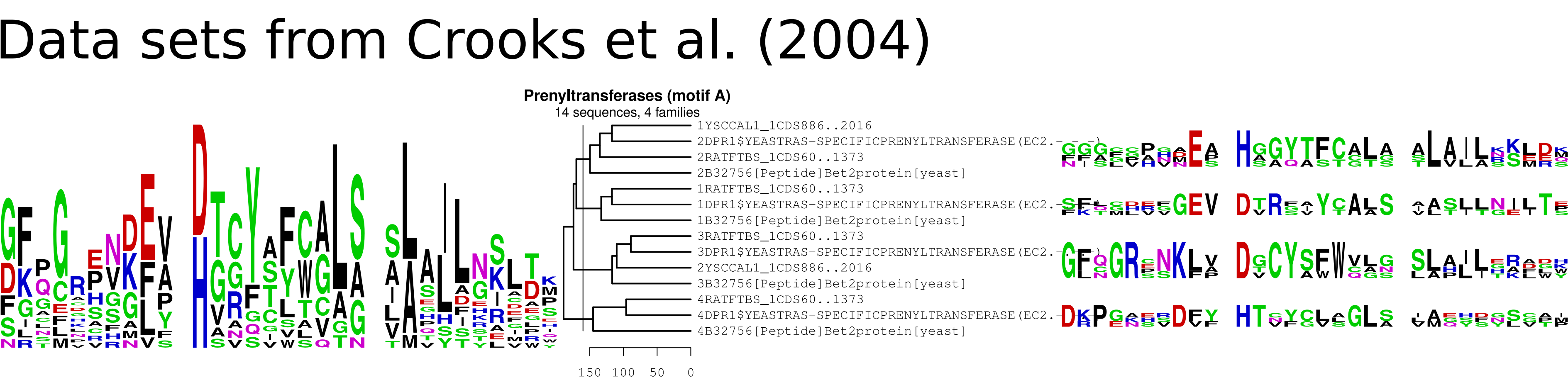
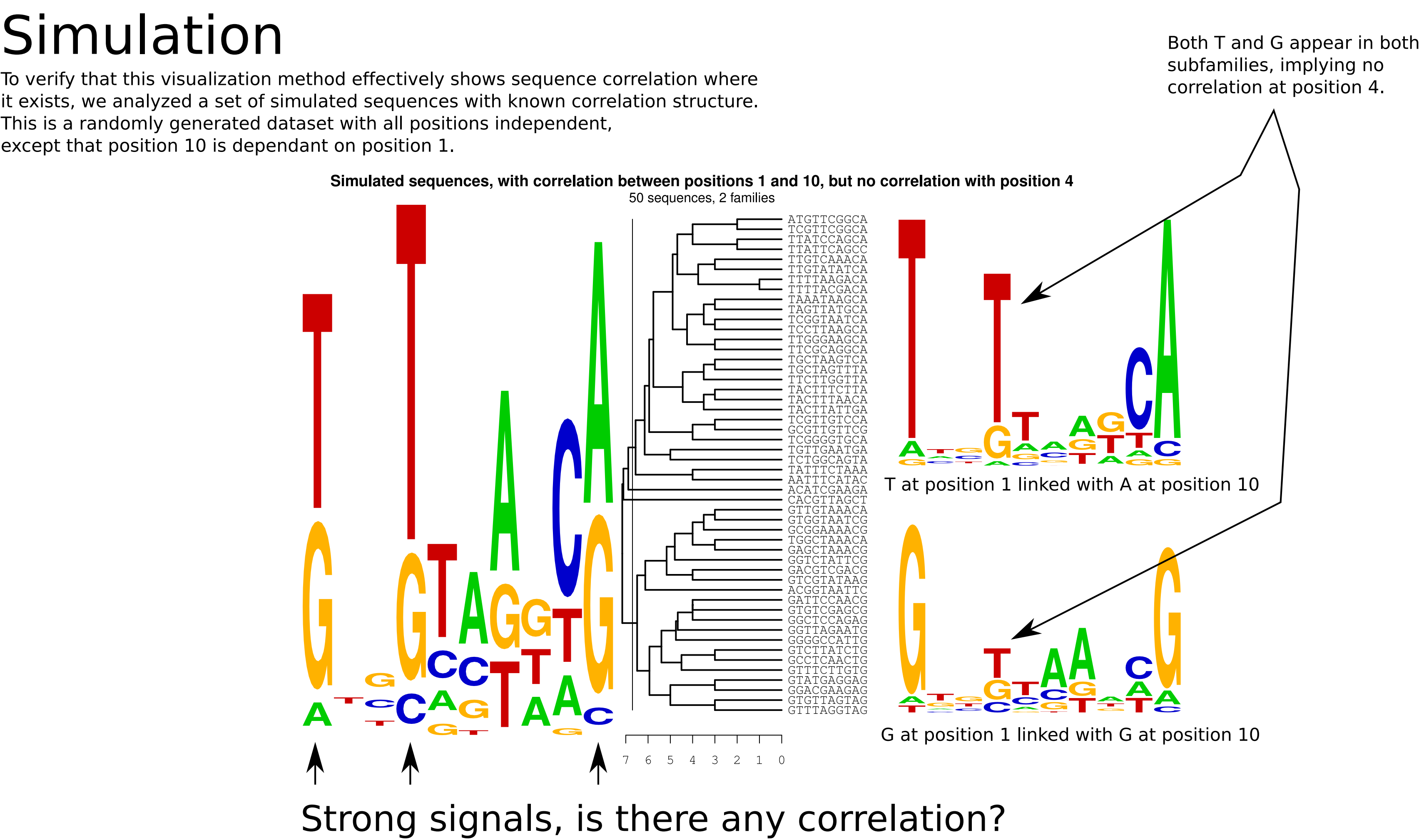


"A standard sequence logo does not provide any indication of correlations between different positions of the alignment."
Crooks et al. (2004), authors of Weblogo



Results

- Visualizes effectively subfamilies which are not readily visible from logo alone.
- Can inform about sequence correlation (or lack thereof).



Methods

Implemented in <http://sublogo.r-forge.r-project.org>

```
> sublogo(  
  c("DQGHRTTR","DVGHRSR","ESGHLRR",...),  
  cutline=30,  
  main="Zinc finger...")
```

will generate the sublogo dendrogram in the lower left.
If the vector of sequences has names, the names will be used to label the dendrogram.

0. **Alignment.** Data input to sublogo() function is a character vector of aligned input sequences.
1. **Difference matrix calculation.** A substitution matrix (BLOSUM62 for protein, identity for DNA) is used to calculate a difference between each pair of sequences.
2. **Clustering.** The difference matrix is used with hclust() to perform a hierarchical clustering.
3. **Cutting.** User selects level to cut the tree, yielding several subfamilies of sequences.
4. **Logos.** Sequences exported to plain text files for input to Weblogo, which saves PS logo images.
5. **Drawing.** R package grlImport used to import logo PS files, gridBase used to combine logos (grid graphics) with dendrogram (base graphics).

Conclusions

Sublogo dendrograms are useful for characterizing

1. subfamily structure
2. sequence correlation

Future work

- Deploy on a webserver.
- Use SeqinR package for sequence IO.
- Find or write a function for drawing dendrograms using grid graphics.
- Use seqLogo package to draw logos instead of weblogo/grlImport.
- Automatically pick cutline and family structure.
- Allow position weights, for clearer partitions.
- Adapt for large (N>100) data sets.

References

Schneider TD Stephens RM (1990). Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res., 18, 6097-6100.

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004). WebLogo: A sequence logo generator. Genome Research, 14, 1188-1190. <http://weblogo.berkeley.edu>

Paul Murrell (2009). Importing Vector Graphics: The grlImport Package for R. Journal of Statistical Software, 30(4), 1-37. URL <http://www.jstatsoft.org/v30/i04/>.