Correlating Regulatory Region and Genetic Evolution

Chinmay P. Rele & Laura K. Reed
Department of Biological Sciences, The University of Alabama

thegep.org

Abstract

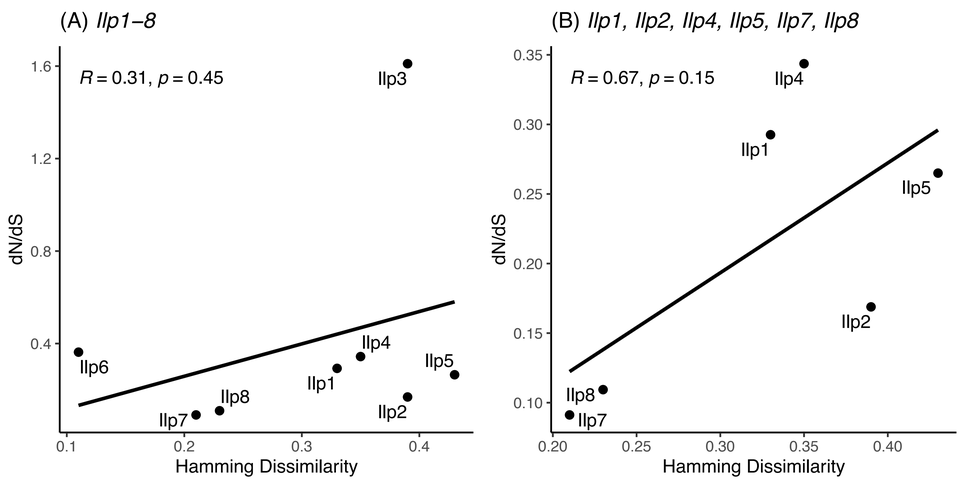
Understanding change at an evolutionary scale is primarily done through accession protein and gene evolution; however, understanding how the regulation of those loci occurs would paint a clearer picture of how the proteome and genome interacts with the environment. The regulatory regions of genes are highly variable and important, for the timing of gene expression. Understanding the dynamics between the regulatory regions of an organism with its relevant effects in the genome, transcriptome, and proteome, can inform us of how the evolution of one plays a role on the other. We are studying the evolution of the regulatory regions *of Insulin-like peptides* (*Ilps*) and how their evolution correlates with other characteristics of the genes. *Ilp* is a good gene family to examine due to high duplication rates, and different mutation rates of those duplicates that occur within the *Drosophila* phylogeny. In this study, we use the hand-curated gene models of genes within the Insulin Signaling Pathway of 28 *Drosophila* species generated by The Genomics Education Partnership, including the *Ilp* paralogs. We are using these annotations to anchor the genomic region, extract genomic regions upstream of the start codon, and run Multiple Sequence Alignments. We intend to correlate features within these islands of conservation with genomic features to be able to find an association between regulatory regions and the genes they regulate. The patterns identified within these islands of conservation upstream of genes are likely *cis*-regulatory elements, or regions where regulatory elements exist that affect the regulation of the gene. Patterns such as indels (insertions/deletions), sequence motifs, as well as overall conservation found in these islands upstream of the gene can be correlated with the metrics of the genes directly downstream, such as ratio of nonsynonymous to synonymous mutations dN/dS, intron size, indel abundance, isoform number, and expression patterns (among other genetic characteristics). We expect to see positive correlation between the evolution of this upstream regulatory region and the evolution of features of the *cis*-regulated genes. Understanding the evolutionary patterns of these upstream regions is paramount in understanding how regulation of genes occurs.

Introduction

- $H_o \rightarrow$ regulatory regions are evolving neutrally
- *cis*-regulatory elements (CREs) were used since they are easy to identify for each gene, and have a gene-specific effect.
- We expect correlation between evolution of a gene's CRE and its corresponding CDS evolution.
- We also expect correlation between the gene's CRE/CDS sequence and its connectivity in the gene network
- The Insulin/IGF Signaling pathway plays key roles in growth, metabolism, stress resistance, reproduction, and longevity.
- Invertebrate genomes usually contain multiple *Ilp* genes with functional redundancy.

- GEP is a nationwide collaboration to help hand-curate gene models.

Methods/Analyses/Results



We performed Pearson correlation of Hamming Dissimilarity Distances and dN/dS using the `ggscatter` package in R.

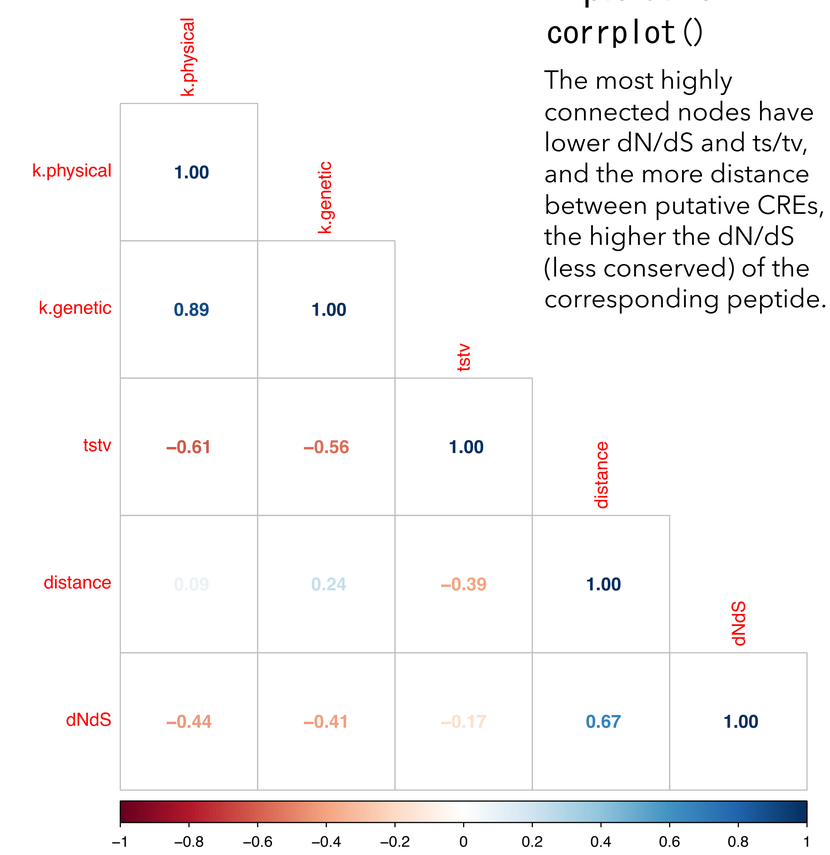
A. Positive correlation between dissimilarity distances (lower dissimilarity distances = more conserved) and dN/dS (lower dN/dS = more conserved/higher negative selection).
B. Omitting two outliers (*Ilp3* due to it not appear to be under selective constraint, and *Ilp6* due to its absence in multiple species from our analysis) revealed that there is moderate correlation between the conservation of the CDS and putative CREs of *Ilp*s in 28 species of *Drosophila*.

Pearson Correlation of Hamming Dissimilarity Distance and dN/dS

Correlating the gapless Hamming Dissimilarity distance shows that the evolution of coding sequence (acted on by selective constraint) and putative CRE sequence (assumed to be neutral) indicates pairing of evolutionary potential. (A) Including *Ilp1-8* indicates a low positive correlation, while (B) excluding *Ilp3* and *Ilp6* shows a higher correlation.

Key Findings

Moderate correlation ($\rho_P = 0.67$) between CDS evolution and putative CRE evolution.



Explorative `corrplot()`

The most highly connected nodes have lower dN/dS and ts/tv, and the more distance between putative CREs, the higher the dN/dS (less conserved) of the corresponding peptide.

- We expect some correlation with node degree (k.physical/k.genetic) with dN/dS.
- Pearson Correlation with R's corrplot, as can be seen in Explorative `corrplot()`.
- Physical degree (physical interactions) not correlated with Hamming Dissimilarity distance, but was correlated with dN/dS and ts/tv (transition/transversion bias).
- Positive correlation between Hamming Dissimilarity distance. Ho would be that there was no correlation due to the assumption that non-coding sequence evolves neutrally.
- Upstream and coding regions are maybe evolving together.

Conclusions/Significance

- Moderate positive correlation between dissimilarity and dN/dS (novel), suggesting that the CDS and putative CREs are paired in their evolutionary potential
- Suggests that the CDS and putative CREs might be sharing evolutionary history and putative CRE are not evolving neutrally

Future Directions

All three results shown will benefit from expanding to more genes and adding more characteristics for each gene.

- Explorative `corrplot()`
  - Add more gene metrics
  - CDS/Exon count
  - CDS/Exon count variation
  - Intron size
  - Indel abundance
- Add more putative CRE metrics
  - Motif abundance
  - Motif size (mean, s.d., variance)
  - Number of motif sites (mean, s.d., variance)
  - Motif width (mean, s.d., variance)
  - d_{SM} (metric that identifies motif presence in multiple replicates)
  - Sequence length (mean, s.d., variance)
  - Indel abundance
- Add network metrics
  - Node degree
  - Node connectivity
  - Neighborhood structure
- Add expression metrics
  - Life stage
  - Expression level

- CDS/Putative CRE Phylogenies
  - Run a larger set of phylogenies using MEGA/RAxML/FastTree with an additional RAxML search

References

- S. Grönke, D.-F. Clarke, S. Broughton, T. D. Andrews, and L. Partridge, "Molecular Evolution and Functional Characterization of Drosophila Insulin-Like Peptides," Plos Genet, vol. 6, no. 2, p. e1000857, 2010, doi: 10.1371/journal.pgen.1000857.
- Post S, Karashchuk G, Wade JD, Sajid W, De Meyts P, Tatar M. Drosophila Insulin-Like Peptides DILP2 and DILP5 Differentially Stimulate Cell Signaling and Glycogen Phosphorylase to Regulate Longevity. Front Endocrinol (Lausanne). 2018;9:245. Published 2018 May 28. doi:10.3389/fendo.2018.00245

Acknowledgements

We would like to thank Wilson Leung and Dr. Jeffrey D. Lozier for their valuable input and discussions concerning parts of the analysis. I would also like to appreciate all of the Reed Lab for providing an environment conducive to research. Students involved with the GEP provided input on models. This poster is based upon work supported by the National Science Foundation under Grant No. 1915544 and National Institute of General Medical Sciences of the National Institutes of Health under award number R25GM130517 to the Genomics Education Partnership (https://thegep.org/).