Mapping Regulatory Elements in Autophagy Gene Promoters

Lena D. van der Stap, Kim Finley and Roberta A. Gottlieb

AP10-10
Mapping Regulatory Elements in Autophagy Gene Promoters.

Lena D. van der Stap1, Kim Finley2 and Roberta A. Gottlieb2
1Computational Science Research Center, 2SDSU BioScience Center
*Corresponding author: 5500 Campanile Drive, San Diego, CA 92182-4650, lvanders@sciences.sdsu.edu

Abstract: The autophagy pathway is critical for cellular homeostasis, and important for cell growth, survival, and response to pathogens. We hypothesize that the composition, number and position of multiple transcription factor binding sites (TFBS) are organized into distinctive clusters within regulatory regions that are conserved among functionally related human autophagy gene promoter sequences. TFBS clusters identified by pattern detection and pattern matching algorithms differed among subsets of autophagy genes, suggesting differential regulation of autophagy pathway components. Functional annotation, confirmation of candidate transcription factors will allow prediction of pathways and physiological stimuli affecting autophagy gene transcription.

Keywords: heart disease, stress, aging, cancer, nutrient response.

1. Introduction

Autophagy is a cellular process in which cytosolic components including harmful protein aggregates and malfunctioning organelles are engulfed by a double membrane-bound vesicle and delivered to a lysosome, where the gathered material is broken down and recycled back to the cell. Therefore, autophagy is important for the cells’ ability to adapt to environmental changes such as hormone, oxygen, or nutrient availability. Autophagy also assists in the continuous turnover of intracellular proteins required to prevent abnormal protein interactions. Therefore, a malfunctioning autophagic system may have progressive consequences, especially in non-dividing differentiated cells such as the heart and brain (Cuervo, 2004).

Currently, many factors are known to induce autophagosome formation: starvation, oxidative stress, shear stress, as well as a number of drugs that target mTOR or upstream signaling molecules such as PKA, PI3K, or the insulin signaling pathway. Conversely, aging, neurodegenerative diseases, cancer, and intracellular invasion by bacteria, viruses, or parasites are often associated with inability to maintain healthy levels of autophagosome formation or merging with the lysosome.

Until recently, most researchers have focused on the signal transduction mechanisms involved in the formation of autophagosomes, since the induction of the pathway is regulated at the protein level by kinases. However, little is known about the transcriptional regulation in the autophagy pathway. Therefore, the main motivation for this study is to determine the transcriptional regulatory profile of subsets of coregulated autophagy genes, as microarray experiments suggest, in various conditions that induce an autophagic response. Genes essential for or rate-limiting in the induction or formation of autophagosomes are the primary focus.

Transcription factors that regulate these coregulated genes may also be differentially regulated in response to cellular stress. Therefore, we hypothesize that composition, occurrence and position of multiple TFBS are organized into distinctive clusters that are conserved among functionally related human autophagy gene promoter sequences. The objective of the study is to use a multi-algorithm approach to show that multiple conserved regions containing potential regulatory binding sites reside in the promoters of subsets of experimentally validated coregulated autophagy genes.
Autophagy Pathway Genes: The tree diagram (above) shows subsets of autophagy genes that have been shown to either physically interact or share related roles in the autophagy pathway.

1.1 Previous Studies

Several studies show that subsets of autophagy genes may be transcriptionally coregulated in response to cellular stresses, regulators upstream in the autophagy pathway, or the effects of upregulating a rate-limiting essential autophagy gene in aging flies.

Dr. Kim Finley’s group has shown that changing Atg8a gene expression levels in the adult central nervous systems regulates longevity. According to Simonsen, et al., 2008, a two-way gene perturbation of just one autophagy gene showed dramatic results in flies: Over expression of Atg8a in aging Drosophila melanogaster resulted in a 56% increase in lifespan and prevented accumulation of harmful protein aggregates, while flies expressing a mutant version of the Atg8a gene displayed a 53% decrease in lifespan compared to wildtype (Simonsen, et al., 2008).

In humans, FOXO3A was shown to directly target autophagy, lysosomal, and growth arrest genes in a transcriptional response to RAS induction in oncogene-induced senescence, especially after 2 days senescence. (Young et al., 2009). In another study, promoter reporters for autophagy genes ULK2, ATG4A, ATG4B, ATG4D, ATG7, GABARAPL2, MAP1LC3A, MAP1LC3B, ATG9A, ATG10, ATG12, AND DRAM were upregulated, and based on luciferase activity induced by increased E2F1 expression in HeLa cells, whereas GABARAPL1/3 was downregulated (Kusama et al., 2009).

2. Methods

A global inventory of putative binding site abundance, type, and relative position using the pattern matching algorithm AliBaba2.1 (www.generegulation.org) was performed (Figure 3). MEME analysis was used to elicit distinct clustering patterns and cross-verify potentially coregulated subsets of autophagy genes found using the other two approaches. (http://meme.sdsc.edu/meme4_3_0/intro.html)

Pairwise comparisons of promoter sequences using a whole genome alignment-based pattern discovery algorithm called rVista (http://rvista.dcode.org/) was used to determine relative positions and number of instances of conserved aligned conserved aligned regulatory regions of the DNA sequences and the putative binding sites contained in these conserved regions. (Figure 5).

In order to reduce false positives and to identify candidates for validation with high confidence, it is common to perform 1) analyses of randomly selected promoter sequences in addition to promoter sequences of interest, 2) enrichment analysis of evolutionarily conserved regions or 3) enrichment of motif position relative to start codons (Young, et al., 2008). In this study, all of such analyses were performed (or are in progress) as well as an extensive a priori analysis of the autophagy pathway coupled with validated sets of coregulated genes.

Figure 2: Autophagy Pathway Genes: The tree diagram (above) shows subsets of autophagy genes that have been shown to either physically interact or share related roles in the autophagy pathway.

Figure 3: Flow of analysis of input sequences.
2.1 Pathway and Gene Structure analysis

Starting with the KEGG pathways database, the analysis continued on to include knowledge from the HADb Human Autophagy Database (http://autophagy.lu/index.test.html), as well as literature searches to characterize genes that share known protein regulation of macroautophagy sub-processes or known protein-protein interactions (Figure 3).

2.2 Input sequences

Ensembl Genome Browser was used for gene structure analysis, and annotated promoter sequences were obtained from Ensembl (Biomart). Transcription start sites (rather than start codons) were used to define each gene’s upstream promoter sequences for motif discovery. This allowed for distinguishing between conserved elements involved in the core transcriptional complex versus regulatory elements that contain enhancer/repressor binding sites. Gene structure analysis also ensured protein coding sequences were used rather than non-coding or pseudogenes. Variability in cDNA clones representing anywhere from 1 to 40 transcripts for each gene present in Ensembl was assessed through visual inspection (Figure 3).

2.3 Prediction of cis-regulatory elements

AliBaba 2.1 - De novo putative transcription factor binding sites were detected within the empirically defined proximal promoters using AliBaba 2.1 (http://www.gene-regulation.com/), which is based on the TRANSFAC 3.5 database. AliBaba2.1 analysis begins with known binding sites instead of predefined position weight matrices. AliBaba2.1 parameters were set as follows: pairism to known sites 64, matrix width 10 bp, minimum number of sites 4-5, minimum matrix conservation 75%, similarity to sequence matrix 1%, factor class level 4-5. The algorithm operates under the assumption that each binding site possesses a context given by the role it plays in gene regulation. First, it pairwise aligns the unknown input sequence with all possible transcription factors in the TRANSFAC database. Second, sets of TFBS are classified based on position and TF class. Third, matrices are constructed from these sets of classified TFBS (context-specific process). Therefore, rather than using predefined position weight matrices (PWM), the algorithm is able to influence the conservation of the matrices.

MEME version 4.1.1 – MEME uses expectation maximization to fit a mixture model in order to detect conserved motifs in up to 30 sequences each 2000bp in length. The p-value, the probability of a random string (generated from the background letter frequencies) having the same match score or higher, was computed based on the match score comparing the site with the position specific scoring matrix for the motif. As part of the analysis, conserved motif sites are aligned with each other (not shown), including 10bp flanking each end of the motif. DNA strand is specified as either ‘+’ (sequence in the training set) or ‘-’ (reverse complement of the training set sequence). For each motif discovered, a block diagram shows the motif occurrences in each sequence in which the motif was found. A summary block diagram shows the relative positions for each motif in all of the training sequences.

Figure 4. Multiple algorithms allowed us to find short individual putative binding sites (AliBaba2.1), medium motifs containing several binding sites (MEME), and long conserved and aligned regions containing aligned conserved binding sites (zPicture/rVista). Selection of candidate binding sites that fall in the consensus of all three methods further reduced the false positives discovery rate.

zPicture – was used to dynamically generate conservation profiles and identify evolutionarily conserved regions (ECRs) between sequences.
from the same species or homologues of different species. Insert Picture is linked to the UCSC Genome Browser allowing for extraction of sequences and gene annotation for any recorded locus. A minimum sliding window length of 100bp and minimum percent identity of 70% required was used, except for cases in which the entire region was highly evolutionarily conserved between species in cross-phyla analyses. In such cases, too loose of criteria would result in an inability to distinguish between cis-regulatory regions and other non-coding sequence. Certainly regions of strong negative selection (i.e. strong functional significance as reflected by high signal/noise) and seemingly indiscriminant highly conserved non-coding regions (i.e. with low signal/noise) would be indistinguishable. In the case of high overall conservation in non-coding regions a longer window size and higher % ID (e.g. >500bp, >85%ID) will yield more reliable predictions of conserved regions.

Regulation Vista (rVista) - is used to perform pattern matching with PWM for known vertebrate or insect TFBS in the TRANSFAC database. It accounts for the linear structure of sequence conservation across a large region of DNA. Since it allows for small gaps and DNA shifts in the aligned TFBS, it maximizes the ability to identify functional TFBS in sequences that have diverged slightly but remain in highly conserved regions.

4. Results:

Fourty-seven autophagy gene promoter sequences were analyzed for conserved regulatory elements and their relative spatial organization. Binding site abundance, type, and relative position using the pattern matching algorithm AliBaba2.1 was performed. On average, 166 response element sequences were found in each 2kb upstream flanking promoter sequence analyzed. The most abundant binding sites were primarily for constitutively active transcription factors (Table1). These are important for the overall context and for later deciphering potential interacting transcription factors. However, the binding sites found not in all autophagy gene promoters but in physiologically related subsets of autophagy gene promoters and found to share common regulatory motifs are of primary interest.

| Table 1: TFBS Abundance common to MAP1LC3B, ATG3, ATG17, ATG10, and ATG16L1 |
|-------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Gene Symbol       | MAP1LC3B        | ATG3            | ATG17           | ATG10           | ATG16L1         |
|                   | Response Sites  | Response Sites  | Response Sites  | Response Sites  | Response Sites  |
|                   | 5               | 7               | 5               | 4               | 2               |
| C/EBP beta        | 2               | 3               | 2               | 2               | 2               |
| CREB              | 2               | 2               | 2               | 2               | 2               |
| Sp1               | 1               | 1               | 1               | 1               | 1               |
| NF1               | 1               | 1               | 1               | 1               | 1               |
| TF2               | 1               | 1               | 1               | 1               | 1               |
| TF3               | 1               | 1               | 1               | 1               | 1               |
| TF4               | 1               | 1               | 1               | 1               | 1               |
| TF5               | 1               | 1               | 1               | 1               | 1               |
| TF6               | 1               | 1               | 1               | 1               | 1               |
| TF7               | 1               | 1               | 1               | 1               | 1               |
| TF8               | 1               | 1               | 1               | 1               | 1               |
| TF9               | 1               | 1               | 1               | 1               | 1               |
| TF10              | 1               | 1               | 1               | 1               | 1               |

Table 1 lists the 10 TFBS that occur in all 47 (10 shown) genes flanking sequences (Hb is in all but the ATG16L2). The most abundant binding sites are for c/EBP alpha, Sp1 (highlighted in aqua) and NF1 (AliBaba2.1).

MEME analysis revealed distinct clustering patterns and cross-verified potentially coregulated subsets of autophagy genes found using the other two approaches (Figure 5).

Figure 5: MEME Results – Motif 10. Distinctive clusters identified using MEME were found to be shared by subsets of autophagy gene promoters. Multiple instances of a cluster were present in genes known to be rate-limiting or to regulate autophagy. The (LOGO) consensus motif 10 found clustered near the transcription start site was subsequently analyzed for potential binding sites, using AliBaba2.1, SP1, AP-2alpha WT1, Krox-20, and ETF were identified in all five promoters. The ETF/Krox-20 binding site, was also found using the other two methods.

For example, a single instance of one conserved region occurs in Atg3, which may have undergone strong negative selective pressure like most protein coding regions. Meanwhile, that same conserved region is repeated six times in the 2kb upstream sequence of Map1lc3b, which is a consumed in the autophagic process and therefore may be subject to more selective pressure.
Positions of repeated conserved regions relative to the transcription start site (i.e., at base pair position 2000) of each gene are revealed (rVista). Both frequently occurring (Sp1) and relatively infrequently occurring (LyF-1) predicted TFBS were found by multiple algorithms: MEME, rVista, and AliBaba2.1. Multiple instances of a conserved region in the Map1lc3b promoter were conserved and aligned with one instance in the Atg4b promoter, and with one instance in the Atg7 promoter. The PITX2 binding site was also found repeated within a conserved region about 5 times in 18 of the 49 genes analyzed. Atg2a shares additional conserved regions with Atg10, suggesting that additional coregulation exists that is distinct from that of other subsets of autophagy genes (Kusama et al., 2009). TFBS patterns found here are consistent with previous studies in terms of tissue specific gene expression or promoter assays. E2F and EGR1 are factors shown in the literature to regulate autophagy genes (Kusama et al., 2009).

5. Conclusions

Gene promoter DNA regions were conserved among subsets of related autophagy genes, and these aligned and conserved putative regulatory sequences contained conserved potential TFBS. Promoters of autophagy genes known to act in multiple functional roles contained multiple conserved instances of the same sequence that may be conserved in promoters of other autophagy genes that play distinct roles. Using multiple algorithms to predict TFBS is useful for reducing false positives and provides a more complete view of spatial information integrated with other promoter information. In addition, reduction of false negatives could be made by expanding the promoter to include more distal regulatory elements. Currently DiRE analysis is underway for a first pass at more upstream regulators that may be missed in this analysis. By identifying putative TFBS and regulatory clusters, we can narrow the search for transcriptional regulatory mechanisms of the gene in a common pathway. This a priori computational analysis serves to: 1) predict potentially rate-limiting genes in a pathway, 2) reveal common transcriptional control patterns in subsets of functionally related genes, 3) predict subsets of genes that may share a particular physiological response, and 4) guide experimental design for gene expression/network studies.

6. Acknowledgements

NIH Road Map Initiative award T90 DK07015.

7. References


