ACSESS proceedings
2010

Mapping Regulatory Elements in Autophagy Gene Promoters

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

AP10-10

CSRC an interdisciplinary gateway to the future
SAN DIEGO STATE UNIVERSITY

Mapping Regulatory Elements in Autophagy Gene Promoters.

Lena D. van der Stap[*,1], Kim Finley[2] and Roberta A. Gottlieb[2]
[1]Computational Science Research Center, [2]SDSU 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.

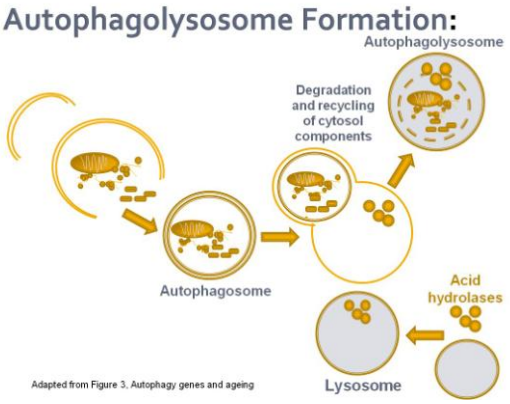
Keywords: heart disease, stress, aging, cancer, nutrient response.

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.

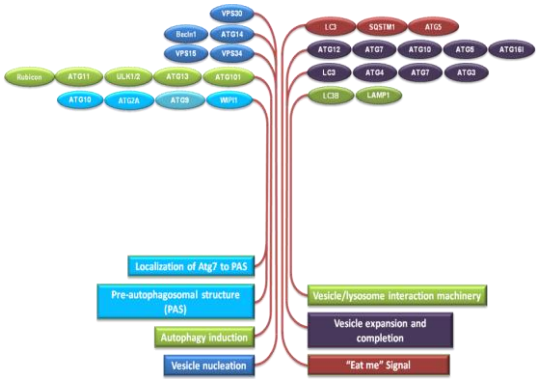


Figure2: Autophagy Pathway Genes: The tree diagram (above) shows subsets of autophagy genes that have been shown to either physically interact or share related roles in the autophagy pathway.

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 senesence. (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.

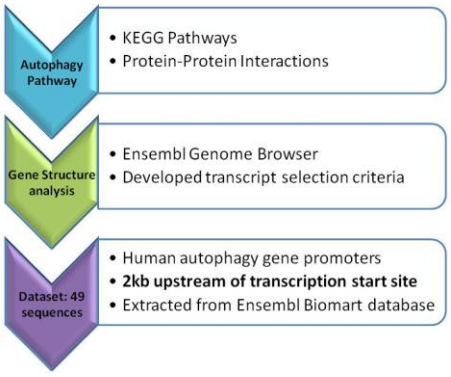


Figure 3: Flow of analysis of input sequences.

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.

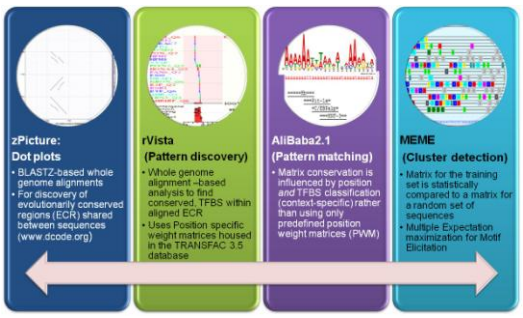


Figure 4. Multiple algorithms allowed us to find short individual putative binding sites (AliBaba2.1), medium motifs containing several binding sites (MEME), and long conserved and aligned regions containing aligned conserved binding sites (zPicture/rVista). Selection of candidate binding sites that 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. zPicture 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.

Table1: TFBS Abundance common to MAP1LC3A, ATG3, ATG7, ATG10, and ATG16L1

TFBS	MAP1LC3A	ATG3	ATG4B	ATG7	ATG9B	ATG10	ATG16L2	ATG5	ATG4A	ATG16L1
AP-1	5	2	3	1	7	1	5	2	2	3
C/EBPalpha	12	22	7	8	5	13	6	15	18	11
C/EBPbeta	3	3	1	1	2	5	1	2	1	7
GATA-1	2	6	1	4	4	1	3	3	5	4
Hb	3	1	2	2	2	3		5	3	3
NF-1	5	6	8	6	8	7	9	6	2	10
Oct-1	6	7	1	3	5	7	4	6	4	8
Sp1	30	23	71	39	27	31	51	19	29	17
TBP	4	2	2	3	2	2	2	3	4	2
USF	1	3	2	4	2	1	2	4	2	3

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).

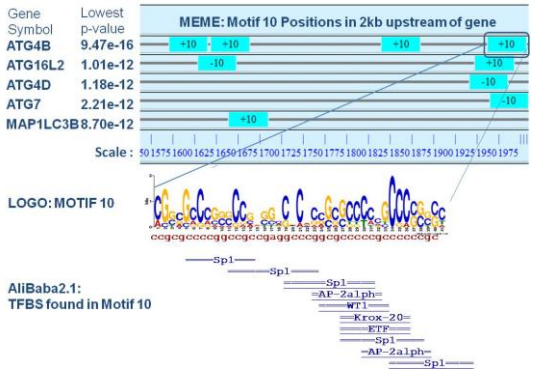


Figure 5: MEME Results – Motif 10. Distinctive clusters identified using MEME were found to be shared by subsets of autophagy gene promoters. Multiple instances of a cluster were present in genes known to be rate-limiting or to regulate autophagy. The (LOGO) consensus motif 10 found clustered near the 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.

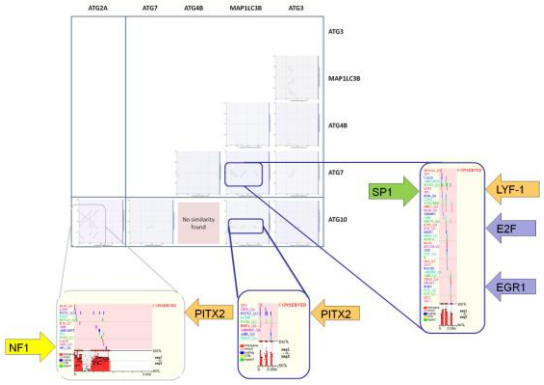


Figure 6: Conserved Regulatory Element Visualization. Dot plots show regions ranging from about 100bp to about 750bp that are conserved between promoters (zPicture). Positions of repeated conserved regions relative to the transcription start site (i.e. at base pair position 2000) of each gene are revealed (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

1. Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

2. Biobase. TRANSFAC: database on eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles. http://www.gene-regulation. com/pub/databases.html#transfac.

3. Chen Z-H, Kim HP, Sciurba FC, Lee S-J, Feghali-Bostwick C, et al. 2008 Egr-1 Regulates Autophagy in Cigarette Smoke-Induced Chronic Obstructive Pulmonary Disease. *PLoS ONE* 3(10): e3316. doi:10.1371/journal.pone.0003316

4. European Bioinformatics Institute and Wellcome Trust Sanger Institute. Ensembl: database of eukaryotic genomes. http://www.ensembl.org.

5. Elnitski L, Jin VX, Farnham PJ, Jones SJ. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res*. (2006) 16:1455–1464.

6. Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al.*(2004)* Genome sequence of the Brown

Norway rat yields insights into mammalian evolution. *Nature 428:493–521.*

7. Ji Hongkai; Wong Wing Hung, Computational biology: toward deciphering gene regulatory information in mammalian genomes. *Biometrics* (2006) 62(3):645-63.

8. KEGG pathway database: http://www.genome.jp/kegg/pathway.html

9. Kim, S. G., and S. J. Lee.. PI3K, RSK, and mTOR signal networks for the GST gene regulation. *Toxicol. Sci*.(2007) 96:206-213.

10. Tie Li ; Luo Lu; Functional role of CCCTC binding factor (CTCF) in stress-induced apoptosis, *Experimental cell research* (2007) vol.313,n14:3057-3065

11. N.Grabe. Alibaba2: Context specific identification of transcription factor. *In Silico Biology*, 2(1):S1–S15, 2002.

12. Simonsen , A. , R.C. Cumming, A. Brech , P. Isakson , D.R. Schubert , and K.D. Finley. 2008. Promoting basal levels of autophagy in the nervous system enhances longevity and oxidant resistance in adult *Drosophila. Autophagy*.4:176 – 184

13. Pavesi, G., Mereghetti, P., Mauri, G., Pesole, G.*(2004)* Weeder Web: Discovery of transcription factor binding sites in a set of sequences from co- regulated genes. *Nucleic Acids Res. 32:W199–W203.*

14. Qiu, P.*(2003)* Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochem. Biophys. Res. Commun. 309:495–501.*

15. Rombauts, S., * Florquin, K., Lescot, M., Marchal, K., Rouzé, P., Van de Peer, Y. (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol. 132, 1162-76.

16. Sinha, S., van Nimwegen, E., Siggia, E. *(2003) Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology, A probabilistic method to detect regulatory modules, (Brisbane, Australia), pp 292–301.*

17. Visel, A., S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis *et al.*, 2008 Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* **40:** 158–160.

18. Tan K, Shlomi T, Feizi H, Ideker T, Sharan R. Transcriptional regulation of protein complexes within and across species. *Proc. Natl Acad. Sci. USA* (2007) 104:1283–1288

19. Young, A. R. J.*et al.* Autophagy mediates the mitotic senescence transition. *Genes Dev.* 11 March 2009 (doi:10.1101/gad.519709)

20. Yuh C-H, et al. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* (1998) **279**:1896–1902.

21. Jinghui Zhao, Jeffrey J. Brault, Andreas Schild and Alfred L. Goldberg *Coordinate activation of autophagy and the proteasome pathway by FoxO transcription factor* Autophagy 4:3, 378-380; 1 April 2008.

22. Zhou, Q., Wong, W.H.*(2004)* CisModule: De novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. 101:12114–12119.*