CS229 FINAL PROJECT REPORT

Who's in Charge Here? : Using Clustering Algorithms to Infer Association of Putative Regulatory Elements and Genes

Ari Officer, Fah Sathirapongsasuti, Te Thamrongrattanarit

Background

After the release of the human genome and many other genomes, new insights have emerged about the complexity of the human genome. While only about 1.2% of the genome code for proteins [1-3], evidence from sequence conservation and expression data suggest that at least 5% of the genome is functional [4-6]. Given that these conserved non-coding elements are not genes but reside among genes, they are therefore believed to comprise the regulatory machinery that governs temporal expression of the genes. These non-gene functional genomic regions are then called putative *cis*-regulatory elements (CRE's). Although many biological experiments confirmed that CRE's do perform such task in the genome, the question remains: which genes do these putative *cis*-regulatory elements regulate?

While the most accurate way to make associations between CRE's and genes that they regulate, conducting a biological experiment can be costly in terms of time and effort and is not guaranteed to yield meaningful results. Furthermore, most experiments were motivated by the heuristic that each CRE should regulate the gene most proximal to it. Even though biologists are able to discover some gene-CRE associations based on this heuristic, the frontier is currently limited to the genes closest to the CRE's. On the other hand, a handful set of experiments [7-8] have shown that CRE's tend to cluster around genes' transcription start sites (TSS) and co-regulate target genes not necessarily closest to it but goes as far as one million base pairs away (Figure 1). Thus, machine learning algorithms, particularly clustering algorithms, can help exploring such cases, thus surpassing the frontier that impedes advancement in biological study. In this project, we would like to apply various clustering algorithms to provide alternative guidelines for further biological experiments that will identify the associations between putative CRE's and the genes that they regulate, which will lead to discovery of biological basis of genetic diseases and subsequently the cures for them.

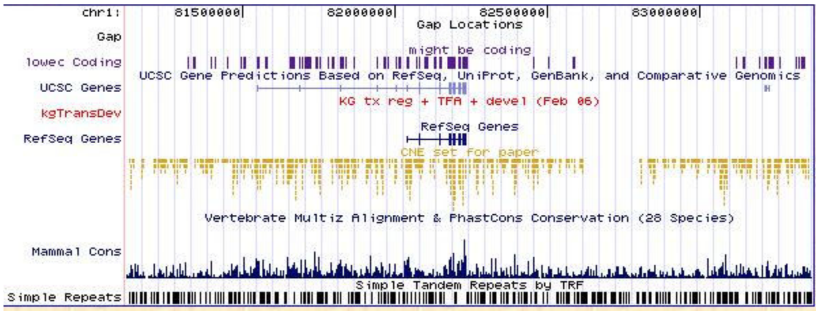
chr1:
81500000
82000000
82500000
83000000
Gap Locations
Gap
might be coding
lowec Coding
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
UCSC Genes
KG tx reg + TFA + devel (Feb 06)
kgTransDev
RefSeq Genes
RefSeq Genes
CNE set for paper
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Simple Tandem Repeats by TRF
Simple Repeats

Figure 1 : The yellow histogram shows the distribution of CRE's over the genomic regions centered at the reference sequence of gene. (Human March 2006 Assembly. chr1:81,117180-83,367,191)

Preliminary Stage

Methods

The genomic locations of CRE's and genes were extracted from UCSC Genome Browser, with courtesy to Prof. Gill Bejerano's Lab, Department of Developmental Biology and Department of Computer Science, Stanford University. Since biological findings show that CRE's that are proximally located tend to co-regulate the same genes, it is rational to cluster CRE's together and then assign each of them to a gene. Based on this heuristic, we chose k-means clustering algorithm to be one of our preliminary algorithms. The k-means algorithm was run to associate regulatory elements with genes based on their transcription starting sites (TSS). Since there are 25,552 gene loci over all chromosomes in our extracted dataset, every CRE would be assigned to one of the 25,552 centroids. For each chromosome, the centroids were initialized to the starting position of the genes. Each CRE was assigned to the closest centroid, and then centroids were updated. After the convergence, we assigned each cluster to the gene closest to its centroid.

Moreover, it is observed that CRE's are roughly normally distributed around a sequence of gene (Figure 1). Given this particular distribution of the CRE's, another reasonable choice of clustering algorithm is E-M with mixture of Gaussians. By fitting multiple Gaussians over genomic regions, we should be able to cluster CRE's and assign them to the gene within that region. Since this algorithm requires a lot of computational power, we could not run this algorithm on the chromosomes with a large number of genes. We only selected chromosome 11, 18, 19, and 20 because of their smaller sizes but relative abundance of known associations and insulators (Table 1) and their ratios between number of known associations and number of all CRE's are the highest.

Evaluation

In every stage of improving the clustering algorithms, we evaluate the performances of the algorithms and compare them against some baseline in the following ways:

1. Known Association Test 1: The result and the baseline will be verified with 81 experimentally confirmed associations hand-curated by T.H. Cheung et al. [9], and the error rates will be calculated. This source of associations is the most reliable because each association is supported by specific biological literature. Nevertheless, a subset of genes is more heavily studied than others, so the data from this source does not represent randomly sampled associations. This test, which relies on this set of data, might not reflect the true performance of the algorithm in question.
2. Known Association Test 2: There are a few thousand known associations in ORegAnno (Open Regulatory Annotation Database) [10-11] but the set contains some indirect association between CRE's and genes. After incomplete entries and indirect associations were screened out, we managed to extract 248 associations and would use these known associations to verify the output from the algorithms. Note that we do not combine known associations with the ones from T.H. Cheung et al. because the curation method of this second set of data appears less reliable than the first.
3. Insulator Test: Evidence from biological experiments [12] shows that a protein called CCCTC-binding factor (CTCF) acts as an insulator and blocks regulation of a gene by CRE's on the opposite side of the CTCF binding site. In other words, an association cannot be correct if the genes and the CRE's are not in between the same pair of insulators. We define the insulator test errors as the proportion of insulators intersecting the clusters. Monte Carlo Simulation was performed over random partitioning of the genomic regions with number of partitions (hence, clusters) equal to the number of genes. Using this data, we estimated the mean and variance of the insulator test error from a random partition over genomic regions and could then produce 99% confidence intervals for random partitions. Since our clustering algorithms do not take the CTCF's into account, testing their insulator test error against the random error is used to support the sensitivity of the algorithm. The null hypothesis is that the clustering algorithm functions independently from the insulators and, thus, should have random error. A low p value, then, supports the algorithm as capturing real information.

Results

For the *k*-means algorithm, the Known Association Test 1 accuracy, Known Association Test 2 accuracy, and Insulator Test error are .66, .46, and .431 respectively (Table 1). Our *k*-means algorithm did significantly better than our baseline of random partitions in every test but one (p<.0001). Although the accuracy rate is not perfect, k-means algorithm proves to be a good antecedent algorithm upon which we can improve.

The error rate of the E-M algorithm with mixtures of Gaussian from the Insulator Test is .659, which is significantly worse than k-means. Moreover, the accuracy rates from Known Associate Test 1 and Known Association Test 2 are zero. This results because E-M is inherently a clustering algorithm, but the clusters stray from one-to-one correspondence with the genes themselves. Therefore, we disregarded this algorithm from our further consideration.

Table 1 Test Results of simple *k*-means algorithm

Chr	Known Association Test 1		Known Association Test 2		Insulator Test
	#Known Associations	Accuracy rate	#Known Associations	Accuracy Rate	Error Rate
1	9	0.44	2	1.00	0.402
2	5	0.80	16	0.25	0.503
3	3	0.67	0	n/a	0.510
4	3	0.67	0	n/a	0.529
5	3	1.00	23	0.74	0.560
6	5	0.40	12	0.50	0.462
7	7	0.86	29	0.52	0.373
8	3	0.33	5	0.20	0.478
9	3	0.33	10	0.70	0.451
10	3	1.00	0	n/a	0.511
11	3	1.00	45	0.49	0.502
12	4	0.75	3	1.00	0.430
13	0	n/a	5	0.80	0.415
14	3	1.00	0	n/a	0.473
15	3	0.33	7	0.71	0.458
16	2	1.00	8	0.25	0.331
17	5	1.00	0	n/a	0.369
18	2	1.00	4	0.25	0.634
19	8	0.38	13	0.15	0.169
20	1	1.00	6	0.33	0.338
21	0	n/a	24	0.42	0.308
22	2	0.50	10	0.50	0.202
Y	0	n/a	0	n/a	0.414
X	3	0.33	26	0.23	0.402
Overall	80	0.66	248	0.46	0.431

Improvements upon *k*-means Algorithm – Weighted k-means

Methods

According to the closest-gene heuristic that biologists in this field rely on, each CRE tends to regulate the gene closest to it. Our simple k-means algorithm is oblivious of gene locations and thus fails to capture this effect. In this stage, we would like to combine the advantage of closest-gene heuristic and the advantage of k-means algorithm by implementing a weighted k-means algorithm. In this algorithm, we place actual weights at the gene locations, which are then used in the k-means algorithm as data points themselves. The only difference is that the centroid update step occurs over weighted data (with the genes now included), so rather than summing the distances, we sum the products of distance and weight. We kept the weights at 1 for all of the CRE's, then arbitrarily assigned weights of powers of 10 to the genes themselves. The assignment of the genes was returned after convergence, and then we processed the clusters into assignments in the same way as with the standard k-means algorithm. The greater the weight at the gene, the more likely that CRE's around that gene will cluster together around a centroid near the gene. Because our algorithms have been applied to a chromosome in its entirety, this weighting helps keep the clusters local, rather than being free to roam around the entire chromosome.

Results

Table 2 Comparison of the accuracy rate in predicting gene-CRE association between unweighted k-means and weighted k-means

Chr	Known Association Test 1					Known Association Test 2				
	#	Unweighted	Weighted 10^2	Weighted10^3	Weighted10^4	#	Unweighted	Weighted 10^2	Weighted10^3	Weighted10^4
1	2	1.00	1.00	1.00	1.00	9	0.44	0.78	0.89	0.89
2	16	0.25	0.44	0.56	0.63	5	0.80	1.00	0.80	1.00
3	0	n/a	n/a	n/a	n/a	3	0.67	0.67	0.67	0.67
4	0	n/a	n/a	n/a	n/a	3	0.67	0.67	0.67	0.67
5	23	0.74	0.78	0.78	0.78	3	1.00	1.00	1.00	1.00
6	12	0.50	0.50	0.50	0.50	5	0.40	1.00	1.00	1.00
7	29	0.52	0.83	0.83	0.79	7	0.86	0.86	1.00	1.00
8	5	0.20	0.80	1.00	1.00	3	0.33	0.33	0.67	1.00
9	10	0.70	0.70	0.70	0.70	3	0.33	0.67	0.67	0.67
10	0	n/a	n/a	n/a	n/a	3	1.00	0.67	1.00	1.00
11	45	0.49	0.56	0.58	0.60	3	1.00	1.00	1.00	1.00
12	3	1.00	1.00	1.00	1.00	4	0.75	0.75	0.75	0.75
13	5	0.80	0.80	0.80	0.80	0	n/a	n/a	n/a	n/a
14	0	n/a	n/a	n/a	n/a	3	1.00	1.00	1.00	1.00
15	7	0.71	0.86	0.86	0.86	3	0.33	0.67	0.67	0.67
16	8	0.25	0.25	0.25	0.25	2	1.00	1.00	1.00	1.00
17	0	n/a	n/a	n/a	n/a	5	1.00	1.00	1.00	1.00
18	4	0.25	0.75	0.75	0.75	2	1.00	1.00	1.00	1.00
19	13	0.15	0.15	0.15	0.15	8	0.38	0.50	0.50	0.50

20	6	0.33	0.33	0.33	0.50	1	1.00	1.00	1.00	1.00
21	24	0.42	0.46	0.46	0.46	0	n/a	n/a	n/a	n/a
22	10	0.50	0.50	0.50	0.50	2	0.50	0.50	0.50	0.50
Y	0	n/a	n/a	n/a	n/a	0	n/a	n/a	n/a	n/a
X	26	0.23	0.27	0.27	0.27	3	0.33	0.67	0.67	0.67

From Table 2, weighted k-means algorithm performed better than unweighted k-means algorithm. We observed that the accuracy rate rises as we increase the weight. As mentioned in the Evaluation section, the experiments that comprise the data set for Known Association Test are suggested by the closest-gene heuristic. The increase in the accuracy rate can be misleading because the data sets that the two tests depend on bias towards the closest-gene heuristic. Therefore, if we put too much weight to k-means algorithm, the algorithm will turn into the closest-gene heuristic and will miss the goal to explore the associations left uncaptured by the heuristic. By the results from Insulator Test (Figure 2), we selected 10^3-weighted k-means to be our optimal algorithm in clustering CRE's and assigning the genes, as it improves upon unweighted k-mean in terms of Insulator Test Errors yet maintains the ability to discover novel associations.

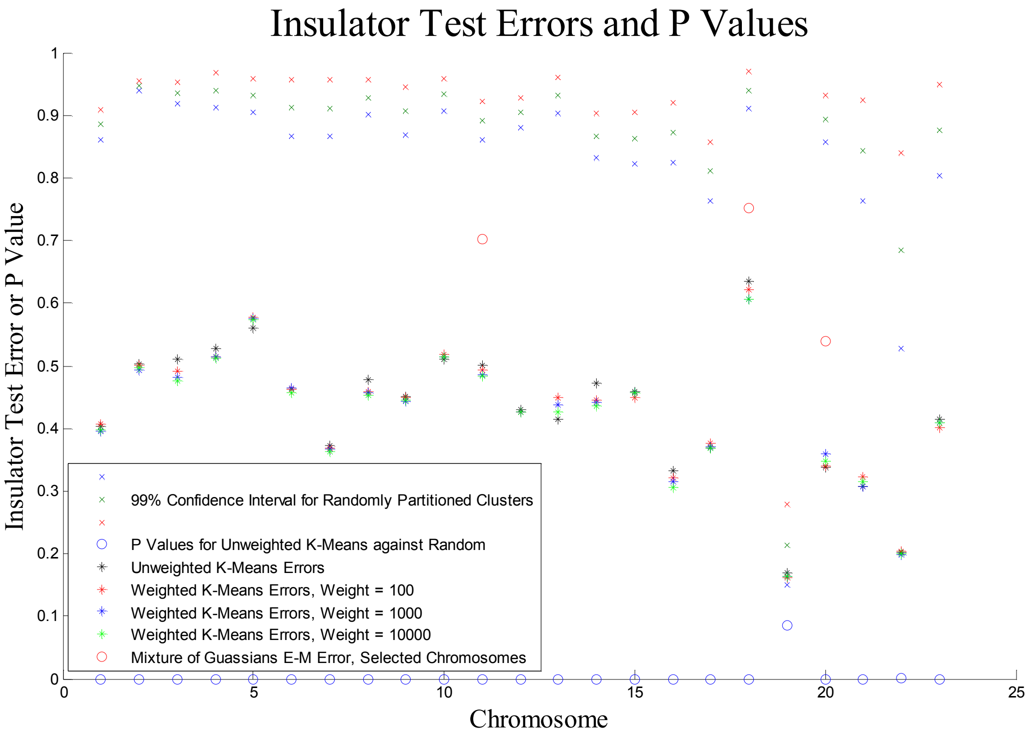

Figure 2 Summary of Insulator Test Errors for the various algorithms. The errors for k-means algorithms are significantly below the errors from random partitioning, and we find the p-values for the unweighted k-means algorithm to be near 0.

Contribution of Weighted k-means

We found 21 cases where weighted k-means algorithm made correct associations but the closest-gene heuristic made incorrect associations. For instance, at chr7:27,183,842-27,197,569, the closest-gene heuristic wrongly assigns the CRE to the gene NR_002795, which is the closest gene. On the other hand, the weighted *k*-means algorithm correctly assigns it to gene HOXA11 because the assignment takes into account the surrounding CRE's, which turn out to be clustered near HOXA11 (Figure 3). The one of most exceptional cases that we have found is that our weighted *k*-means algorithm correctly associates the gene and the CRE that are 100,000 bases apart. Because the weighted *k*-means algorithm does not depend solely on the distance, CRE's that are remotely located can be associated with the genes. An example of such case is chr11:64,103,642-64,271,641. The CRE was inaccurately assigned to SLC22A12 by the closest-gene heuristic, but the weighted *k*-means algorithm makes the correct association (Figure 4).

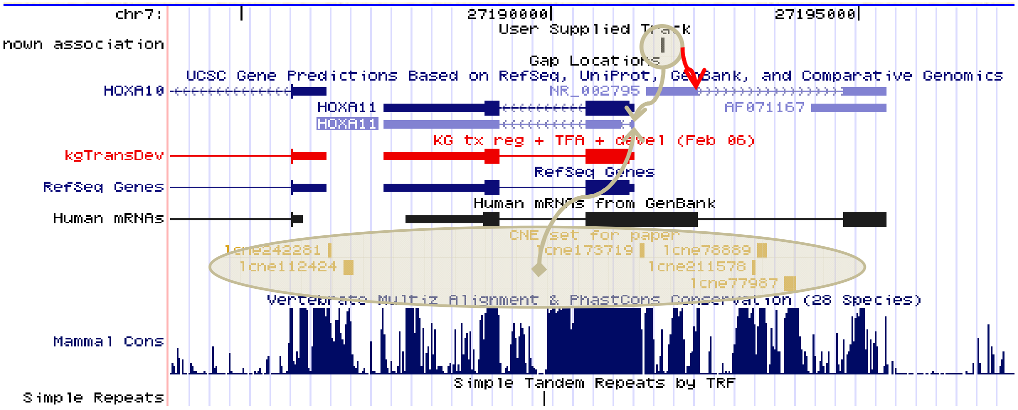

Figure 3 The closest heuristic assigns the CRE to the gene NR_002795 (red arrow), but k-means algorithm correctly assigns it to HOXA11 (tan arrow) (UCSC Human March 2006 Assembly. Chr7:27,183,842-27,197,569).

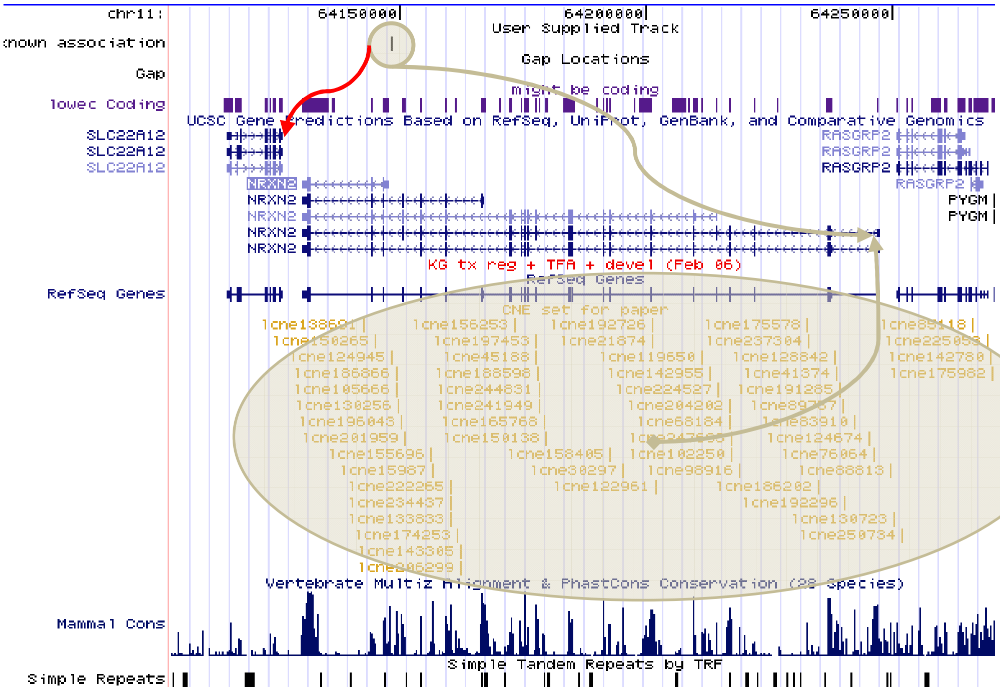

Figure 4 The closest heuristic assigns the CRE to the gene SLC22A12 (red arrow), but k-means correctly assigns it to NRXN2 (tan arrow) (UCSC Human March 2006 Assembly. Chr11:64,103,642-64,271,641).

However, because of the low accuracy rate in Known Association Tests, doubt can be cast on the validity of the result from our algorithm. As mentioned earlier, the data sets used by Known Association Tests are based on the heuristic and therefore do not represent random samples of all genes. In order to verify that our weighted k-means algorithm is a valid algorithm in clustering and assigning CRE's to the genes that they regulate, we implemented a statistical tool for this purpose. This tool is called Enrichment Test of Association (ETA). ETA is based on a study by Bejerano et al. (2004) [13] that a subset of CRE's known as the Ultraconserved Elements involve in some key transcriptional and developmental functions. Since the Ultraconserved Elements regulate genes that share some common functions, we expect to observe overrepresentation of genes with those functions being regulated (i.e. associated) by the Ultraconserved Elements. Therefore, a clustering and gene-CRE associating algorithm is valid if the algorithm associates the Ultraconserved Elements to the genes that share these key transcriptional and developmental functions. ETA uses Hypergeometric Test to assess this enrichment and report Bonferroni-corrected p value for each annotation term. If the *p* value is less than .05 for some key functions of the Ultraconserved Elements, then the algorithm is valid. We ran ETA on the weighted *k*-means algorithm that we invented, and it turns out that the algorithm is able to capture some key features of the Ultraconserved

Elements (Table 3). As a result, the algorithm can be used to predict gene-CRE association especially in the case where the CRE and the gene that it regulates are very far apart.

To explore the full advantage of ETA, we then traced back the list of gene-CRE associations that yields the enrichment. We found a list of Ultraconserved Elements that are not previously known to be involved in some functions, for example sequence-specific DNA binding and RNA binding. These predicted associations can then be tested further in experiments.

Table 3 Gene Ontology terms that are enriched in Ultraconserved Elements

Gene Ontology ID	Description	P Value
Enrichment in Exonic[1] Ultraconserved Elements		
GO:0003723	RNA binding	9.680709e-06
GO:0003676	nucleic acid binding	1.172123e-05
GO:0030528	transcription regulator activity	2.165738e-05
GO:0003700	transcription factor activity	3.821301e-05
GO:0008380	RNA splicing	6.662757e-04
GO:0016070	RNA metabolism	7.088915e-04
GO:0006397	mRNA processing	1.925729e-03
GO:0043565	sequence-specific DNA binding	5.431610e-03
GO:0016071	mRNA metabolism	5.810479e-03
GO:0003677	DNA binding	9.263524e-03
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	9.686886e-03
Enrichment in Intergenic[2] Ultraconserved Elements		
GO:0043565	sequence-specific DNA binding	1.198226e-02
GO:0006350	transcription	3.720624e-02
GO:0045449	regulation of transcription	4.403441e-02
GO:0007222	frizzled signaling pathway	4.541888e-02

Conclusion and Future Direction

Our weighted k-means algorithm successfully expands the boundary of exploration. Even though the accuracy rate of this algorithm is not affirmingly high, biologists can combine facts from previous biological experiments and the prediction that the weighted *k*-means algorithm has made in order to guide further research. If the prediction and the biological facts seem to be consistent, then researchers can make a rational attempt to confirm the association by conducting biological experiments. The genomic regions that researchers afford to explore will be less restrained with the combined guidelines from biology and machine learning algorithm. Researchers can be more confident in their endeavor to dig through the fortress of CRE's that is far away from the comfort zone of the closest-heuristic.

In addition to escalating the possibilities of exploring the unexplored, results from the algorithm will lead to experiments that will potentially provide more data on known association and give a good foundation for developing the algorithm. The Known Association Tests that we used were not quite reliable because most of the known associations were discovered based on the heuristics. If the data sets contained more associations where the gene and CRE's are more than 100,000 bases apart, the Known Association Tests would become more reliable and would facilitate further development of algorithms.

Reference

1. E. S. Lander et al., Nature 409, 860 (2001).
2. J. C. Venter et al., Science 291, 1304 (2001).
3. Human Genome Sequencing Consortium, in preparation.
4. R. H. Waterston et al., Nature 420, 520 (2002).
5. K. M. Roskin, M. Diekhans, D. Haussler, in Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology (ACM, New York, NY, 2003), pp. 257–266.
6. F. Chiaromonte et al., Cold Spring Harbor Symp. Quant. Biol. 68, 245 (2003).
7. T. Vavouri and G. Elgar, Curr. Opin. Genet. Dev. 15, 395 (2005).
8. C. B. Lowe, G Bejerano, D. Haussler, PNAS, 104, 8005 (2007).
9. T. H. Cheung, K. K. B. Barthel, Y. L. Kwan, X. Lin, PNAS, 104, 10116 (2007).
10. S. B. Montgomery et al., Bioinformatics, 22, 637 (2006).
11. O. L. Griffith et al., Nucleic Acid Res., Nov 15 (2007).
12. T. H. Kim et al., Cell, 128, 1231 (2007).
13. G. Bejerano et al., Science, 304, 1321 (2004).

[1] Exonic Ultraconserved Elements are Ultraconserved Elements that overlaps protein coding portion of genes

[2] Intergenic Ultraconserved Elements are Ultraconserved Elements that locate between genes