Longer sequence surrounding motif distinguishes regulatory elements from false positives

Emily Rocke, [1] **James Thomas,** [2]

Keywords: motif, regulatory element, chemosensory, srh gene family, elegans

1 Overview.

The DNA sequence CASSTG[3] is overrepresented upstream of srh-family chemosensory genes in the nematode *C. elegans*. It very likely represents a regulatory binding site of interest to nematode chemosensation, although its regulatory effects and ligands are unknown.[3]

In this abstract, we make the observation that the CASSTG motif is embedded in a much longer (upwards of 40 nucleotides, although we look at only 26), stochastically conserved motif which is not palindromic. Computationally, the conservation of this longer motif allows instances of CASSTG as a regulatory motif to be distinguished from the many expected random instances of CASSTG in the genome. Because the longer motif is loosely conserved, the misidentification rate is high, but the approach is useful in identifying a number of very likely motif candidates. In this abstract we discuss the effectiveness of even a very simple algorithm which uses this additional information.

2 Introduction.

DNA regulatory elements are short DNA sequences (usually 6-20 nucleotides), typically appearing near or in the genes they regulate, that have evolved to be favorable binding sites for specific proteins or RNA molecules. The same type of protein or RNA molecule may bind to similar sites near several related genes; such patterns of similar sites, known as motifs, can often be detected computationally.

Any particular short DNA sequence is expected to occur many times in a genome by chance alone. For example, a given 6-nucleotide sequence is expected to occur about every 4,000 nucleotides, and so many thousands or even millions of times in a typical genome. These chance occurrences of a short pattern may overwhelm the number of legitimate binding sites, causing a difficult identification problem. One solution is to look at clusters of two or more nearby motif instances[1, 2], but this only works if such clusters exist in the data.

3 Method.

The nematode *C. elegans* has a large family of about a thousand genes that encode related 7-transmembrane proteins, presumed to act as chemoreceptors[5]. The large srh gene family[4] has approximately 200 genes and pseudogenes, of which 185 apparently active genes were used as a data set. The 1Kb region preceding each gene was searched for exact instances of CASSTG; 139 of the 185 regions had at least one occurrence of the motif, for 240 total motif occurrences, of which the 115 that fell closest to the genes were used to construct a weight matrix of 10 nucleotides on each side of the small motif.

[1]Genome Sciences Dept., University of Washington, Seattle. E-mail: `ecrocke@gs.washington.edu`
[2]Genome Sciences Dept., University of Washington, Seattle. E-mail: `jht@u.washington.edu`
[3]Here S, or strong, means that either nucleotide C or G may appear in this position

C. elegans chromosome V, about 21.7 million bases, was scanned for matches to either CACCTG or CAGGTG. 13,727 instances were found and sorted by the log likelihood score of the 20 surrounding nucleotides belonging to the motif weight matrix versus the background model. The 24 top-scoring instances were selected using a predetermined cutoff score, and each checked for plausibility using the Wormbase database (`http://www.wormbase.org/`). 9 of these top-scoring instances are recovered motifs from the original set, which are ignored since they were used in constructing the motif weight matrix.

4 Results.

Of the remaining 15 highest-scoring motifs, four immediately showed strong indications of being regulatory motifs. The other eleven can not yet be classified.

One interesting motif instance was juxtaposed with a gene that has been assigned to the srz family of chemosensory genes, a cousin of the srh family with about 40 member genes. This finding, while not conclusive, leads to a strong suspicion that the CASSTG motif regulates the srz family as well. We are now collecting data on the upstream regions of the srz family to test this hypothesis.

Two more high-scoring motif instances occur in front of srh-family *pseudogenes*, or inactive gene copies. The strong preservation of the regulatory motif suggests that these are very recent duplications or inactivations of active srh genes, which may be informative on the evolution of this gene family.

The fourth interesting motif instance occurs in the genome directly after the srh gene srh-120, and immediately before a nearly-perfect copy of srh-120. This copy does not appear to be annotated as a gene or pseudogene in Wormbase or in other sources. It is difficult to tell through computational means alone whether this gene is an inactive, very recent copy of srh-120, or whether it is a new active srh gene discovered through this method.

5 Conclusions.

A simple algorithm for scoring the moderately-conserved surroundings of a small, well-conserved motif allowed several new biologically interesting motif instances to be selected out of an overwhelmingly large set of false positives. This approach may help understand the elusive srh gene family, but the implications go beyond *C. elegans* biology. There is an intriguing possibility that other apparently small regulatory motifs are embedded in large, loosely conserved motifs. If this is often true, then computational methods of distinguishing "real" motif instances from false positives using surrounding sequence will become an important toolset in the arsenal of computational techniques.

References

[1] Frith, M. C., Hansen, U. and Weng, Z. 2001. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878-889

[2] GuhaThakurta, D. and Stormo, G. D. 2001. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17:608-621.

[3] McCarroll, S. and Bargmann, C. 2002. Personal communication.

[4] Robertson, H. M. 2000. The large srh family of chemoreceptor genes in Caenorhabditis nematodes reveals processes of genome evolution. *Genome Res.*, 10:192-203.

[5] Troemel, E. R., Chou, J. H., Dwyer, N. D., Colbert, H. A. and Bargmann, C. I. 1995. *Cell*, 83(2):207-218.