NtcA motif reconstructions: Methods and Results

May 4, 2006
John Aach

Supplemental material for

Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability

Andrew C. Tolonen, John Aach, Debbie Lindell, Zackary I. Johnson, Trent Rector, George M. Church, Sallie W. Chisholm

This document describes the methods used to attempt to reconstruct the NtcA motif from upregulated clusters 1 and 2 for MED4 and MIT9313, and the results obtained from this effort. (See main text for information on the clustering.)

METHODS

Upstream sequences from upregulated clusters 1 and 2 for MED4 and MIT9313 were analyzed for motifs by AlignACE (Hughes et.al. 2000; Roth et.al. 1998) and compared against the NtcA motif matrix derivable from the NtcA "training set" in Su et.al. (2005). High ranking motifs from MED4 and MIT9313 were also inspected and compared against each other.

Upstream sequences: 100bp upstream sequences were generated for all genes/operons from NCBI genomic sequences NC_005072.1 (MED4) and NC_005071.1 (MIT9313), and the subsets of sequences for MED4 clusters 1 and 2, and for MIT9313 clusters 1+2 were extracted. Upstream sequences were gathered using Su et.al. (2005) rules for operons, whereby tandem genes separated by <45bp are considered to be in an operon. Upstream sequences were extracted in a strand-specific manner so that nearby divergent genes (operons) may share some or all of the same upstream sequence (although with different senses).

GC content: GC content must be specified to AlignACE. GC content was calculated from each full set of upstream sequences extracted for each organism. For MED4 the GC content was 0.216762836185819. For MIT9313 the GC content was 0.478711652359087

AlignACE: AlignACE run from http://atlas.med.harvard.edu/cgi-bin/alignace.pl. Aside from specifying GC content, AlignACE may be directed to align a specific number of columns. AlignACE was run three times with the number of columns set to 8, 10, and 12, respectively Defaults were used for all other parameters. Because parts of the same upstream sequences (from different strands) may be in the input sequence set for nearby divergent genes (operons), the upstream sequences from any nearby divergent genes in the clusters were weighted more by AlignACE than the upstream sequences of other genes in the clusters. AlignACE returns a MAP score (Hughes et.al. (2000)) which reflects the degree of difference between the generated alignment from background nucleotide probabilities.

Specificity Scores: For each AlignACE motif, ScanACE (Hughes et.al. (2000) was used to find all instances in each genome whose motif scores were > 2 standard deviations below the mean of the scores of the instances comprising the AlignACE alignment itself. On an individual organism basis, strand-specific upstream sequences were converted to non-strand-specific upstream sequence regions (USRs), whereby overlapping divergent upstream sequences were merged into a common region (N_U = number of USRs). USRs containing cluster 1 and 2 upstream sequences were designated as input set regions (ISRs; N_I = number of ISRs). The number of distinct USRs, and the number of distinct ISRs, that contained at least one ScanACE-identified AlignACE were counted, yielding numbers M_U and M_I, respectively. Instances not found in any USR were ignored. Following Hughes et.al. (2000), Specificity scores (known in Hughes et..al. (2000) as Group Specificity Scores) were computed as

$$SpecificityScore = \sum_{x=M_I}^{\min(M_U, N_I)} \frac{\binom{M_U}{x}\binom{N_U - M_U}{N_I - x}}{\binom{N}{N_I}}$$

Specificity scores were computed in MatLab using the hygecdf function as

1-hygecdf(M_I–1, N_U, M_U, N_I)

where $M_I > 0$. Specificity scores are probabilities, with values near 0 representing high specificity. However, specificity scores should not be interpreted as P values for which P<.05 represents statistically significant specificity.

ScanACE was downloaded from http://atlas.med.harvard.edu/.

Motif comparisons: Motif comparisons were performed using CompareACE (Hughes et.al. (2000)), which was downloaded from http://atlas.med.harvard.edu/. Motifs from all AlignACE runs were compared to determine their similarity to each other. All motifs were also compared against an alignment of all the sites in the NtcA "training set" from Su et.al. (2005) "training set" to determine their similarity to known NtcA sites. CompareACE scores are Pearson correlation coefficients, so values near 1 represent similar motifs.

Sequence logos: All sequence logos were generated using http://weblogo.berkeley.edu/logo.cgi. (Crooks et.al. (2004); Schneider and Stephens (1990)).

Motif statistics spreadsheet: Motif MAP, specificity, and NtcA ComparACE scores were compiled in NtcA_reconstruction.motifdata.xls. Specificity, MAP, and CompareACE scores were also ranked from best to worst (best = 1), where rankings were specific to each AlignACE run (so that, e.g., there are three rank 1 specificity scores for MED4, one for each of the AlignACE column 8, 10, and 12 runs).

Motif logos ppt file: Selected motif logs (see Results) were placed in upregulated_cluster_motif_sequence_logos.ppt.

RESULTS

General

- Judged by CompareACE scores, AlignACE runs from columns 8, 10, and 12 generated many similar motifs, but also many dissimilar motifs. For instance, of the 12 column 10 motifs generated for MED4, 7 had CompareACE scores >0.7 with MED4 column 8 motifs, and of the 7 column 12 MED4 motifs, all but one had CompareACE scores > 0.7 with MED4 column 8 motifs. Similar results were obtained for MIT9313 motifs (12/18 column 10 motifs were similar at CompareACE > 0.7 with column 8 motifs, and 10/20 column 12 motifs were similar at this level with column 8 motifs. Once I found good recovery of the NtcA motif from the MED4 column 8 run (see below), I generally focused on column 8 motifs.

MED4

- For each AlignACE run (columns=8, 10, 12), the motif with the highest MAP score matched the NtcA consensus sequence by eye, and had very high CompareACE scores (>0.8 for columns = 8 and 10, >0.7 for columns=12) against the NtcA alignment derived from the Su et.al. (2005) NtcA site "training set." These motifs also had very strong specificity scores (<=1.02e-9), and were, indeed, the most specific motifs found for each run. The motif derived from columns=8 has the same length as the standard NtcA consensus, while the motifs from the other runs were longer, due mainly to appended AT strings. For this reason I generally focused on columns=8, even though the columns=10 and 12 motifs had stronger MAP and specificity scores.

- Including large upregulated cluster 3 in AlignACE runs with clusters 1 and 2 resulted in motifs that had poorer matches to NtcA (data not shown). This may suggest that clusters 1and 2 are more strongly and specifically regulated by NtcA.

- I obtained sequence logos for the NtcA motifs recovered from the column 8, 10, and 12 runs, and for the next best 9 MAP scores from the MED4 column 8 runs, and placed these in upregulated_cluster_motif_sequence_logos.ppt.

MIT9313

- The NtcA motif could not be reconstructed from clusters 1 and 2. The best CompareACE score against the Su et.al. (2005) NtcA "training set" of any motif from these upstream regions is <.5. Increasing the length of upstream sequence to 200 and 500bp also did not result in motifs similar to NtcA (data not shown). Nevertheless, several strong and specific motifs are found for MIT9313 clusters 1 and 2.

- If these results suggest that NtcA may not play a strong role in regulating clusters 1 and 2, it is possible that one of the motifs derived from these clusters represents binding sites of another regulator. On the hypothesis that this regulator might also be present in MED4, I identified the motifs in the MIT9313 AlignACE runs that were most similar to motifs generated for MED4. Only pair of such motifs had a CompareACE score of > .7 (actual score = 0.71). This motif has a relatively weak MAP score in each organism and modest specificity (<6e-5).

- I obtained sequence logos for the best 10 MIT9313 motifs from the column 8 run, for the motif with the highest similarity to the NtcA alignment based on known sites, and also the three pairs of motifs from MIT9313 and MED4 that exhibited the most similarity. All logos are in upregulated_cluster_motif_sequence_logos.ppt.

REFERENCES

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E., 2004, WebLogo: A sequence logo generator, *Genome Research* 14:1188-90

Hughes, J.D., Estep, P.W., Tavazoie S., and Church, G.M., 2000, Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*, *J Mol Biol* 296(5):1205-14.

Roth, F.R., Hughes, J.D., Estep, P.E., and Church, G.M., 1998, Finding DNA Regulatory Motifs within Unaligned Non-Coding Sequences Clustered by Whole-Genome mRNA Quantitation, *Nat Biotechnol* 16(10):939-45.

Schneider T.D. and Stephens, R.M., 1990, Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res* 18:6097-100

Su, Z. Olman, V., Fenglou, M., Xu, Y., 2005, Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. *Nucleic Acids Res* 33(16): 5156-71