# Computational discovery of composite motifs in DNA

Geir Kjetil Sandve[1,4], Osman Abul[2] and Finn Drabløs[3]

1 Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
2 Department of Computer Engineering, TOBB University of Economics and Technology, Ankara, Turkey
3 Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
4 Department of Informatics, University of Oslo, Norway

*E-mail Presenting Author:* finn.drablos@ntnu.no
*Project URL:* http://tare.medisin.ntnu.no/
*Code URL:* http://tare.medisin.ntnu.no/compo/
*Open Source License:* GPL

Computational discovery of motifs in biomolecular sequences is an established field, with applications both in the discovery of functional sites in proteins and regulatory sites in DNA. In recent years there has been increased attention towards the discovery of composite motifs, typically occurring in cis-regulatory regions of genes. Single motif discovery of transcription factor binding sites has been shown to generate a large number of false positive predictions [1]. Searching for composite motifs is interesting because it can give a more realistic prediction of regulatory binding sites by filtering out some of the false positives.

This presentation describes Compo: a discrete approach to composite motif discovery that supports richer modelling of composite motifs and a more realistic background model compared to previous methods [2]. Furthermore, multiple parameter and threshold settings are tested automatically, and the most interesting motifs across settings are selected. This avoids reliance on single hard thresholds, which has been a weakness of previous discrete methods. Comparison of motifs across parameter settings is made possible by the use of p-values as a general significance measure. Compo can either return an ordered list of motifs, ranked according to the general significance measure, or a Pareto front corresponding to a multi-objective evaluation on sensitivity, specificity and spatial clustering.

Compo performs very competitively compared to several existing methods on a collection of benchmark data sets. These benchmarks include a recently published, large benchmark suite [3] where the use of support across sequences allows Compo to correctly identify binding sites even when the relevant position weight matrices (PWMs) are mixed with a large number of noise matrices. Furthermore, the possibility of parameter-free running offers high usability, the support for multi-objective evaluation allows a rich view of potential regulators, and the discrete model allows flexibility in modelling and interpretation of motifs.

1. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ *et al*: **Assessing computational tools for the discovery of transcription factor binding sites**. *Nat Biotechnol* 2005, **23**(1):137-144.
2. Sandve GK, Abul O, Drablos F: **Compo: composite motif discovery using discrete models**. *BMC Bioinformatics* 2008, **9**:527.
3. Klepper K, Sandve GK, Abul O, Johansen J, Drablos F: **Assessment of composite motif discovery methods**. *BMC Bioinformatics* 2008, **9**:123.