Conservation-based methods:

- MEME_c [1]
  - 'multiple genes, multiple species' method.
  - aligns orthologous sequences (using ClustalW [2]) to identify conserved regions, masks the bases that are not conserved in 2/3 or 3/4 of the orthologous sequences, and then applies a conventional motif finder (MEME [3]).
  - ignores regions with a conservation level below the chosen threshold.
- the method of Kellis *et al.* [4]
  - 'multiple genes, multiple species' approach.
  - searches for mini-motifs (3-gap-3 motifs) that are more conserved in the bound sequences than expected, and then extends these motifs if the neighboring bases are also conserved.
  - is based on alignments of orthologous sequences. A score CC4 is used to detect motifs conserved in the intergenic regions in a category (*e.g.* coregulated genes). Let IN be the number of conserved instances of a motif in the positives (*i.e.* sequences within the category), and let OUT be the number of conserved instances in the negatives (*i.e.* sequences outside the category). The method searches for motifs of the form XYZn(0-21)UVW that have a low probability of occurring in IN out of IN+OUT cases.
- Converge [1, 5]
  - 'multiple genes, multiple species' method.
  - takes as input a set of sequences believed to share a common motif, and pair-wise alignments of these sequences to orthologous sequences from related species. It assumes the alignments are high-quality.
  - is similar to MEME [3], but modified to include conservation in the probabilistic model.
  - treats the probability of a motif occurring at a site in the alignment as the product of the probabilities of the motif occurring at the same site in each of the aligned sequences.
  - the background model is 5^{th} order Markov model.
  - requires a motif width as input.
  - the motifs found by Converge are scored by the comparing the frequency of motif occurrences in the bound versus unbound sequences, using a hypergeometric distribution.

- unlike the first version of Converge [1], the algorithm used by MacIsaac *et al.* [5] allows for different evolutionary distances between each species and the reference genome. These distances are not required as input, instead they are learned by the algorithm.

- PhyloCon (Phylogenetic Consensus) [6]:
  - 'multiple genes, multiple species' method.
  - locally aligns conserved regions of orthologous sequences into multiple sequence alignments (or profiles), keeping the optimal alignment as well as suboptimal local alignments. Then it compares profiles from non-orthologous regions and merges the common sections into a new profile, using a greedy approach, until only a few profiles are left.
  - does not take into account phylogenetic relations between species.
  - does not require aligned sequences.
  - does not need the motif width *a priori*.
  - it is slow, since it has to consider a large number of local alignments.
  - uses a 0^{th} order Markov model to describe the background.

- PhyME [7]
  - 'multiple genes, multiple species' method.
  - EM-based approach that requires as input aligned sequences, as well as a phylogenetic tree describing the distances between organisms.
  - it allows for motifs to occur in conserved as well as non-conserved regions, but when a motif occurs in a conserved region, it has to occur in the orthologous sequences as well.
  - conserved regions (blocks) between the reference species and each of the other species are computed using LAGAN [8].
  - it allows for phylogenetic trees with arbitrary topologies (*i.e.* not restricted to star topology), unlike PhyloGibbs [9].
  - the evolutionary model takes into account the binding site specificities. The model assumes that all position in a binding site evolve independently at equal rates, and the probability of fixation of a mutation to a base b is proportional to the PSSM entry of b at that position.
  - very similar to PhyloGibbs [9], with the following differences: 1) EM versus Gibbs sampling, 2) any topology versus star topology, 3) PhyloGibbs allows for several motifs to be searched simultaneously.

- FootPrinter [10]

- is a 'single gene, multiple organisms' method.
- is a Gibbs sampling algorithm that uses conservation scores computed from sequence alignments. It biases the search towards windows that are highly conserved.
- assumes a trusted phylogenetic tree is given.
- this algorithm was not selected for comparison because it is a 'single gene, multiple organisms' approach.

- CompareProspector [11]
  - 'multiple genes, multiple species' method.
  - is a Gibbs sampling algorithm that uses conservation scores computed from sequence alignments. It biases the search towards windows that are highly conserved.
  - defines two conservation thresholds: T_{ch} and T_{cl}. During the initial iterations only the positions with conservation scores $> T_{ch}$ are sampled. Then the threshold is decreased gradually until it reaches T_{cl}.
  - this algorithm is only available upon request. We did not use it in the comparison because we have not received it in time to test it on our dataset.

- EMnEM [12]
  - 'multiple genes, multiple species' method.
  - EM approach that simultaneously learns the motif model and the evolutionary model.
  - considers observed sequences to have been generated from ancestral sequences that are two component mixture of motif and background, each with their own evolutionary model.
  - takes as input aligned sequences.
  - uses the Jukes-Cantor model of evolution, and simply assumes a slower rate of evolution in the binding sites compared to background sequences.
  - the final motif is the motif in the ancestor.
  - applicable to any group of species whose intergenic regions can be aligned.
  - time complexity: the algorithm is linear in the total length of the data, but the initialization is quadratic in the length of the data.
  - this algorithm was not selected for comparison because it is too computationally expensive.

- orthoMEME [13]

- ‘multiple genes, multiple species’ method.
- EM approach that searches the space of motifs and alignments simultaneously.
- each motif is assumed to have an orthologous copy in the other species, that could be located anywhere in the orthologous promoter.
- the model assumes a star topology for the evolutionary tree, with equal branch lengths.
- assumes the motif and its orthologs are in the same orientation.
- it can handle only two species at a time.
- time complexity: each E-step and M-step take $O(nm^2W)$, which makes the algorithm very slow (n is the number of sequences, m the length of the input sequences, and W the motif width).
- this algorithm was not selected for comparison because it is too computationally expensive, and it is designed for only two related organisms.

- PhyloGibbs [9]
  - ‘multiple genes, multiple species’ method.
  - based on Gibbs sampling.
  - uses syntenic local multiple alignments (produced by Dialign [14]). Clearly similar segments are aligned into blocks, the rest are left unaligned. Binding sites occur either in conserved regions (in which case they must be aligned) or in non-conserved regions. Occurrences in non-conserved regions are treated as independent.
  - assumes a star topology for the evolutionary tree (given as input).
  - as in PhyME [7], the evolutionary model implemented in PhyloGibbs assumes that all position in a binding site evolve independently at equal rates, and the probability of fixation of a mutation to a base b is proportional to the PSSM entry of b at that position.
  - very good performance on simulated data.
  - on real data, PhyloGibbs without phylogeny (*i.e.* treating the sequences as independent) performed almost as well as PhyloGibbs with phylogeny.
  - it can search for several motifs simultaneously.

- the sampler of [15]
  - ‘multiple genes, multiple species’ method.
  - use two substitution matrices: for motifs and for background. The background model is estimated from sequence alignments. The motif model assumes half the branch length of the background model.

- this algorithm was not selected for comparison because it has been reported [5] to perform worse than methods we do compare to in our analysis (PhyloCon and Converge).

References

[1] Harbison,C., *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104.

[2] Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G., Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497–500.

[3] Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *ISMB '94*, AAAI Press, Menlo Park, California, pp. 28–36.

[4] Kellis,M., Patterson,N., Endrizzi,M., Birren,B., and Lander,E. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **432**, 241–254.

[5] MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G., Fraenkel,E. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7**, 113.

[6] Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**, 2369–2380.

[7] Sinha,S., Blanchette,M. and Tompa,M. (2004) PhyME: A probabilistic algorithm for Finding Motifs in Sets of Orthologous Sequences. *BMC Bioinformatics* **5**, 170.

[8] Brudno,M., Do,C., Cooper,G., Kim,M.F., Davydov,E., Green,E.D., Sidow,A., and Batzoglou,S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* **13**, 721–31.

[9] Siddharthan,R., Siggia,E.D., and van Nimwegen,E. (2005) PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Computational Biology* **1**, e67.

[10] Blanchette,M. and Tompa,M. (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Research* **31**, 3840–3842.

[11] Liu,Y., Liu,X.S., Wei,L., Altman,R.B., and Batzoglou,S. (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Research* **14**, 451–458.

[12] Moses,A., Chiang,D., and Eisen,M. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp. Biocomput.* 2004, 324–335.

[13] Prakash,A., Blanchette,M., Sinha,S., and Tompa,M. (2004) Motif discovery in heterogeneous sequence data. *Pac. Symp. Biocomput.* 2004, 348–359.

[14] B. Morgenstern (2004) DIALIGN: Multiple DNA and Protein Sequence Alignment at BiBiServ. *Nucleic Acids Research* **32**, W33–W36.

[15] Li,X. and Wong,W.H. (2005) Sampling motifs on phylogenetic trees. *PNAS* **102**, 9481–9486.

[16] Elemento,O. and Tavazoie,S. (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biology* **6**, R18.

[17] Newberg,L.A., Thompson,W.A., Conlan,S., Smith,T.M., McCue,L.A., Lawrence,C.E. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for *cis*-regulatory site prediction. *Bioinformatics* **23**, 1718–1727.