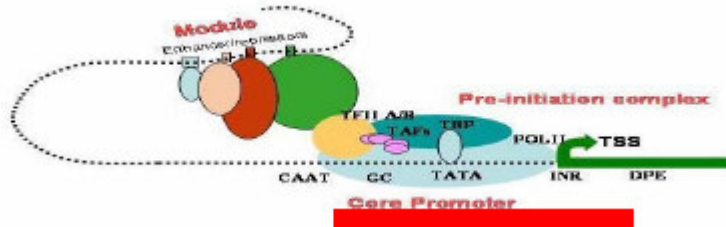


Promoter Prediction Review

(1) Biological background

Eukaryotic promoter structure



Core promoter: Few hundred base pairs 5' to the TSS, (often home to several DNA “signals” such as the TATA box, the GC box, and the CCAAT box), where Pol II localizes (assisted by other general transcription factors such as TF II) and initiates transcription.

Cis-regulatory module (CRM): Clusters of transcription factor binding sites, often several kb 5' to the TSS, that recruit protein factors which may enhance or repress the transcriptional activity.

Goal: Develop robust computational predictions of core promoters in order to understand basal transcription and also in order to better predict CRMs.

(2) Core promoter prediction I

DNA signals (core factors)

- (1) Long range sequence characteristics
 - a. DNase I derived bendability – using information for DNase I hypersensitive mapping experiments that indicate more “accessible” and “less bendable” DNA regions
- (2) Short, regional sequence characteristics
 - a. CpG islands
- (3) Specific cis elements
 - a. TATA box
 - b. GC box
 - c. CCAAT box
 - d. One or more of the above

The signals described in this list are the signals that various labs have used to identify core promoter regions. The poor performance of many of these programs stems from the fact that these signals/factors are not sufficient to describe promoters. For example, consider the table below which shows that the 4 cis-elements commonly attributed to core promoter regions are not even remotely sufficient to identify existing promoters, much less novel ones.

Percentage of promoter sequences containing 4 core-factor binding sites

	TATA	INR	GC box	CAAT box
CpG-poor	17.12%	8.87%	24.28%	7.37%
CpG-rich	14.06%	51.00%	9.01%	7.18%

Due to this insufficient amount of information, the search has been on for novel core signals by conservation analysis and/or motif discovery algorithms such as MEME and AlignACE.

(4) Core promoter prediction II

Additional core factors

Rather than the presence/absence of the various DNA signals, promoter prediction tools have more involved properties of the signals:

- (1) Known TF site distribution
- (2) Novel TF site or novel motif distribution
- (3) Cluster score of known and predicted TFs
- (4) CpG island length
- (5) Distance of CpG island from TSS
- (6) GC-rich / GC-poor nature of the sequence in the {TSS-600, TSS+600} region

Different promoter prediction tools used a subset of the above features to identify core promoter regions. Hannenhalli and Levy, in 2001, decided to incorporate all the properties in to one program and determine the relative contributions of each feature. They found that the CpG island based features (presence/absence, length, distance to TSS) are the most dominant features. In fact, the predictions based on these CpG island features alone are roughly the same as the predictions based on all of the core factors. This indicated that it will be very difficult to discern promoter regions of genes that are not flanked by CpG islands at the end 5' end (estimated to be 50% of mammalian genes). The field has been relatively stagnant since this point.