Improving precision of core promoter predictions

Thomas Abeel, Yvan Saeys and Yves Van de Peer
Department of Plant Systems Biology, VIB, Ghent University, Technologiepark 927, 9052 Gent, Belgium

Contact: thomas.abeel@psb.ugent.be
web: http://bioinformatics.psb.ugent.be

bioinformatics Gent

Background

At the moment approximately 50 eukaryotic genomes are sequenced and more are in the pipeline. To keep up with the pace of sequencing projects, **automated annotation techniques are required**. One of the most challenging problems in genome annotation is the delineation of the core promoter.
Accurate promoter prediction may improve genome annotation and can guide experimental work.

Challenges

• Genome (3Gbp) → genes → promoter → core promoter (50 bp)
• Many genomes (350 projects)
• Little sequence preservation in core promoter
• High number of false positives
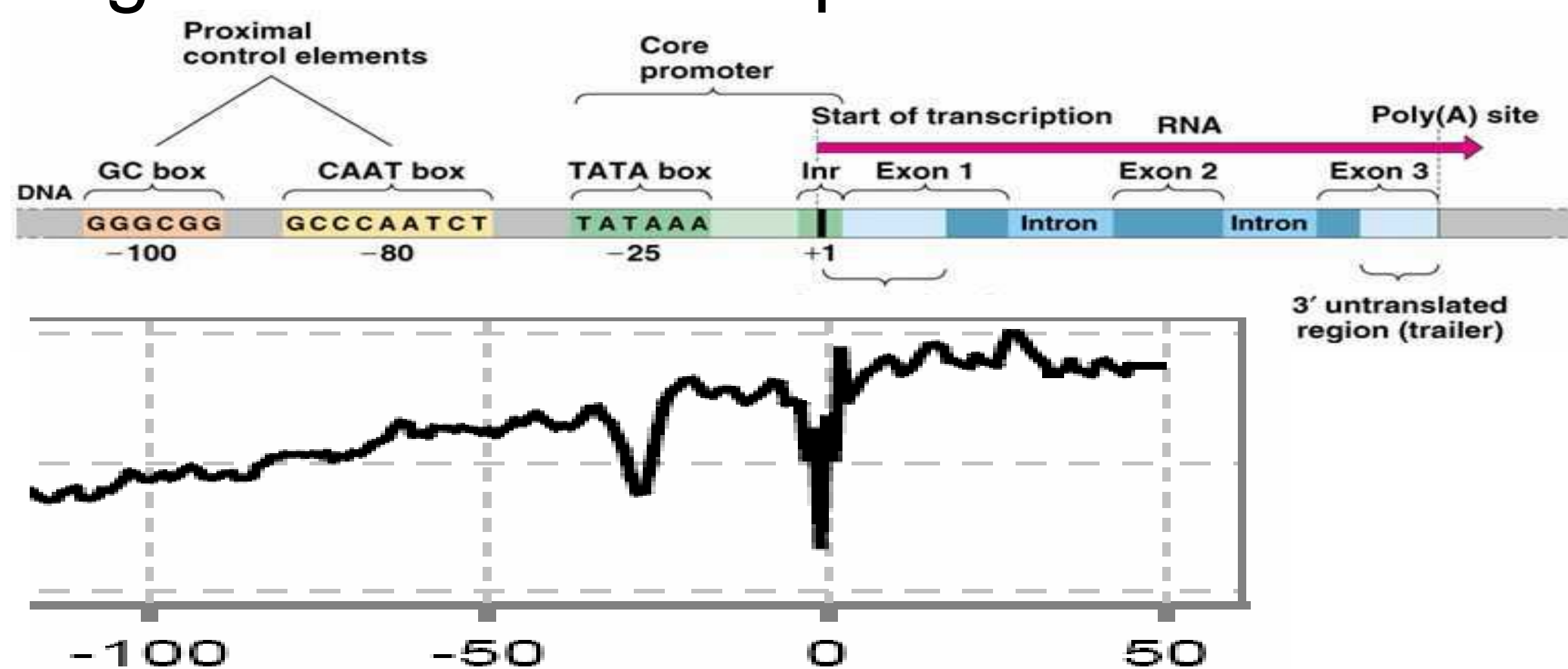
Prediction techniques

ProStar: A new method for the prediction of promoter regions based on atomic molecular dynamics simulations of small oligonucleotides has been developed. (Goni et al. 2007)
EP3 is a tool for the identification of the core region of a eukaryotic gene promoter. It uses universal properties of the promoter to detect those regions in a whole genome context. (Abeel et al. 2008)
ProSOM is a promoter prediction program based on clustering of physical properties of DNA. The program uses self-organizing maps to distinguish between the structural profiles of promoter sequences and other genomic sequences.
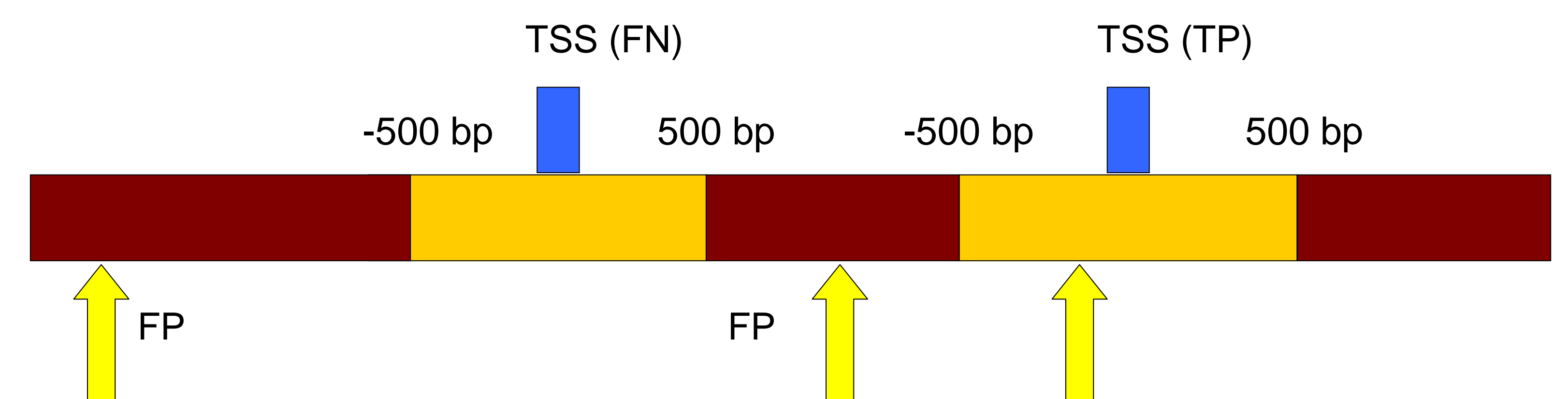
Set operation to increase precision

Single program	Recall	Precision	F-measure
ProStar	23%	60%	33%
EP3	34%	68%	46%
ProSOM	38%	69%	48%
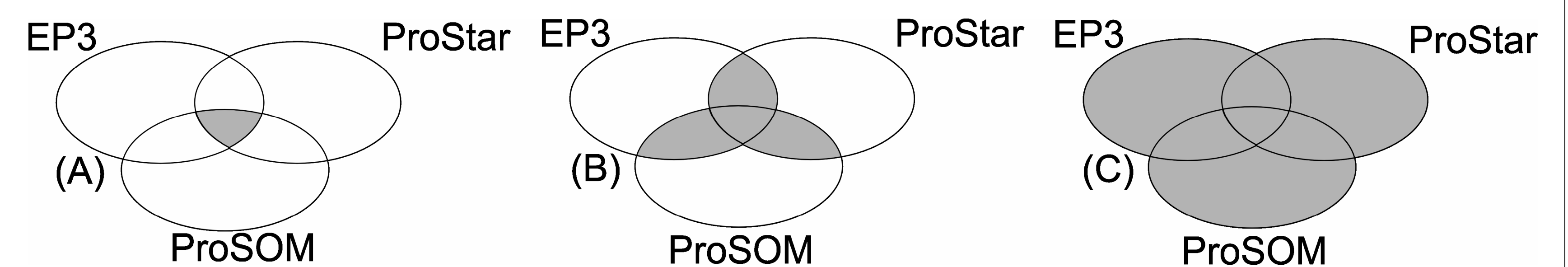
Combination	Recall	Precision	F-measure
A) ProStar ∩ EP3 ∩ ProSOM	20%	87%	31%
B) Majority vote	35%	76%	48%
C) ProStar ∪ EP3 ∪ ProSOM	40%	54%	46%

Validation

• Human genome assembly
• CAGE TSS tags for human (~120,000 TSS)
• TP: true positive: correctly predicted TSS
• FP: false positive: false prediction
• FN: false negative: unpredicted TSS

	Triple intersection	single best
• recall =	35%	38%
• precision =	**76%**	69%
• F-measure =	48%	48%

Abeel, T., **Saeys, Y.**, **Rouzé, P.**, **Van de Peer, Y.** (2008) ProSOM: Core promoter prediction based on unsupervised clustering of DNA physical profiles. Bioinformatics 24, i24-i31.
Abeel, T., **Saeys, Y.**, **Bonnet, E.**, **Rouzé, P.**, **Van de Peer, Y.** (2008) Generic eukaryotic core promoter prediction using structural features of DNA. Genome Research 18, 310-23.
J Ramon Goñi, Alberto Pérez, David Torrents and Modesto Orozco (2008) Determining promoter location based on DNA structure first-principles calculations. Genome Biol. 2007;8(12):R263.