3. Hybrid approach based on a combination of methods for operon prediction

C. Hubans1,2, D. Hot2, Y. Lemoine2,4, E. Mephu-Nguifo3

1Genoscreen–CampusCalmette,1 rue du Pr Calmette59000Lille France
2Laboratoire d'Etudes Transcriptomiques et Génomiques Appliquées-Institut Pasteur de Lille, 1 rue du Pr. Calmette BP24559019Lille cedexFRANCE
3CRILs–CNRSFRE2499,IUT de Lens,Rue de l'université SP16,62307Lens cedex France
4Université de Lille I, 59655Villeneuve d'Ascq cedexFRANCE

We present a new combining method for operon prediction. The algorithm developed in this work is based on the unique information of the genome sequence on which a prediction of genetic elements is performed.
These preliminary results are quite encouraging mainly because there is still some room for prediction improvement.

MOTIVATION: To better understand transcription regulation in prokaryotic organisms, it is necessary to distinguish single transcription units from operonic structures.
Operons are complex structures composed of several elements: two or more Open Reading Frames (ORF), one or several promoter(s) and one or several terminator(s). The transcription product from an operon is a polycistronic mRNA with, classically, all its cistrons encoding proteins belonging to a common metabolic pathway and/or representing different subunits of a multisubunit protein complex [1].

ORIGINALITY:
Minimal and simplest organization for an operon is one promoter, two cistrons, and one terminator, but there are many more complex organizations with more than two ORFs or/and with internal promoters or/and terminators which hardened computing prediction. An efficient method should be able to recognize all transcribed genetic elements and their potential associations in order to identify and differentiate operonic structures from single ORF. In the literature, only a few references have described predicting algorithms to date. Some of them present either weak specificity or give efficient results only with model organisms such as E. coli [2]. Some other algorithms correctly predict operonic structures but require a lot of additional pieces of information beside sequences, such as microarray or molecular biology data [3], functional annotation [4] or neighbour genome [5] to perform analysis. A recently developed algorithm uses unsupervised method to predict if a pair of adjacent genes is included into the same operon [6]. This algorithm considers the distance between two consecutive genes, their orthologs in other genomes, and functional similarity deduced from the annotation. This kind of approach allows getting good prediction quality but, unfortunately, all these informations are not always available.
In the work presented here, we intend to develop a novel approach requiring minimal information to be able to use it on a new unannotated sequence. Our goal is to design a method for "ab initio" computational annotation of operons.

ALGORITHM:
The algorithm developed in this work is based on the unique information of the genome sequence on which a prediction of genetic elements is performed. Then, intergenic regions are deduced from the list of predicted ORF. These intergenic regions are then classified in four groups as function of the strand, size and orientation of the adjacent ORFs. On these regions the presence of promoters and terminators

are predicted, and the association of these different predicted elements allows the definition of operon position. Several methods exist to predict prokaryotic promoters such as neural network with a learning machine (NNPp), scoring matrix method (Promscan) or a linear discriminant function (Bprom). The specificity and sensitivity of these methods are evaluated (Specificity: percentage of samples labelled as "promoter" that actually are "promoter" samples. Sensitivity: measures how well the classifier can recognize "promoter" samples). The comparison of these promoter prediction methods shows that none of them give perfect results (Table 1).

method	specificity	sensitivity
NNPp	29,10%	66,73%
Bprom	56,00%	58,30%
Promscan	19,30%	50,00%
BAGGING	48,50%	66,70%

Table1: comparison of three methods to predict promoters

The idea is therefore to create a unique algorithm which combines these methods to increase the result quality. Results in table 1 led us to choose the bagging algorithm using the three experts working via web connections: NNPp www.fruitfly.org/seq_tools/promoter.html, Bprom www.bio.net/hypermail/microbio/2003-January/022004.html and PromScan www.promscan.uklinux.net/.
Supplementary tests allow defining a score function. The score function is used to compute the quality value of each prediction. Our score function is divided in two parts, the first one confers a different weight to the different experts and a second part is used in case prediction from one of the expert is missing.

Fscore= NNP *2 + Bprom *3 + PromScan *1 + (2NNP)/(NNP-0,01) + (2BProm)/(BProm-0,01) + (PromScan)/(PromScan-0,01)

All predictions are clustered by a sorting algorithm which uses the start position of each element. The aim is to identify all operonic structures with their internal promoter(s) and terminator(s) and single transcription units. The difficulty stands in the evaluation of the association of elements mainly because of the high number of unusual combinations which increases the number of possibilities.

RESULTS:
In order to evaluate the algorithm of bagging described above , a same data set corresponding to 86 transcription units of Escherichia coli K12, the composition of which is detailed in table 3A, was used without (Table 3B) or with (Table 2C) this algorithm. The first part of the implementation comprises ORF prediction, intergenic region analyses and prediction of rho-independent terminators (Table 2B). The second part uses in addition the algorithm of Bagging, which improves significantly the number of correctly predicted operon (Table 2C) demonstrating the interest of a robust method for promoter prediction. We observe that when using promoter prediction in addition of ORF and the intergenic regions prediction, the percentage of correctly detected operon increase from 32.5% to 54.5%. These preliminary results are quite encouraging mainly because there is still some room for prediction improvement. An additional method allowing better discrimination of the different parts of the genome (coding, inter-cistronic, inter operonic) is currently under testing.

	strand	forward	reverse	
A	operon	16	15	*number of the different structures in the data set.*
	single unit	33	22	
	total	49	37	
B	TU correct	17 (35%)	11 (30%)	*First test without the prediction of promoter but with prediction of optimized ORFs.*
	ORF errors	3	3	
	prom error	29	23	
C	TU correct	27 (55%)	20 (54%)	*Second test including the prediction of promoter with the algorithm of bagging.*
	ORF errors	3	3	
	prom error	19	14	

Table2: results of operon prediction using different parameters

1. Alpers, D.H. and G.M. Tomkins, The Order Of Induction And Deinduction Of The Enzymes Of The Lactose Operon In E. Coli. Proc Natl Acad Sci U S A, 1965. 53: p. 797-802.
2. Ermolaeva, M.D., O. White, and S.L. Salzberg, Prediction of operons in microbial genomes. Nucleic Acids Res, 2001. 29(5): p. 1216-21.
3. Bockhorst, J., et al., A Bayesian network approach to operon prediction. Bioinformatics, 2003. 19(10): p. 1227-35.
4. Jacob, E., R. Sasikumar, and K.N. Nair, A fuzzy guided genetic algorithm for operon prediction. Bioinformatics, 2005. 21(8): p. 1403-7.
5. Westover, B.P., et al., Operon prediction without a training set. Bioinformatics, 2005. 21(7): p. 880-8.
6. Price, M.N., et al., A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res, 2005. 33(3): p. 880-92.