Ecole Normale Supérieure
LABORATOIRE DE PHYSIQUE STATISTIQUE
associé au C.N.R.S et aux Universités Paris 6 et 7
24, rue Lhomond, 75231 Paris Cedex 05

Vincent HAKIM
Tel: +33 1 44 32 37 68
e-mail : hakim@lps.ens.fr

Dear Editor

we would like to submit the attached manuscript, "**Imogene : identification of motifs and cis-regulatory modules underlying gene co-regulation**", for publication in Nucleic Acid Research as a *Standard paper* pertaining to the *Computational Biology* subject category.

There is a strong interest in computationally decoding the cis-regulatory information that directs gene expression in a tissue or condition-specific manner. A common situation in practice is the availability of a small set of CRMs active in a condition of interest, provided either by different classical small scale studies, or by the test of a set of CRMs identified in a large-scale study. From these available CRMs, one would usually like both i) to decode the cis-regulatory information which gives rise to expression specific to the condition of interest and ii) to obtain other CRMs specifically active in the same condition. However, there are few existing algorithms that are able to deal with a small set of (10-20) characterized CRMs. Moreover, the top-performing ones are motif-blind : they address task ii) by giving up addressing task i), as assessed in a previous work (Kantorovitz et al, Dev. Cell 17, 568579 (2009)), i.e. they globally characterize the statistics of the available CRMs without providing information on transcription factor binding sites.

Our manuscript improves on this current situation. Building on our previous work (Rouault et al, PNAS 107, 1461520 (2010)), it describes and tests *Imogene*, a novel algorithm that is able to address both tasks i) and ii). Namely, it determines cis-regulatory information in multi-cellular organisms and particularly in mammals, *de novo* from a small set of well-characterized

CRMs. It then predicts from it novel CRMs with the same specificity. *Imogene* bypasses the lack of extensive data by exploiting in a systematic way the phylogenetic information now available in the sequences of multiple related genomes.

We present tests of the basic capabilities of *Imogene*, namely creating *de novo* cis-binding motifs specific to a family of CRMs and retrieving other CRMs with similar regulatory abilities, using available neural and limb mouse CRMs. *Imogene* is found to perform on these mammalian CRMs as well as the current best motif-blind methods. We also show that different classes of CRMs can be distinguished based on *Imogene* generated motifs. Notably, we show that only using genomic sequence information, *Imogene* performs as well in this classification task as the machine learning based on extensive ChIP-seq data presented in the noted work of Zinzen et al, Nature, 462, 6570 (2009).

Imogene computer code is publicly available. Additionally, a web-interface has been developed on the Pasteur Institute mobyle web platform, to render the developed ensemble of statistical tools easily usable by the biological community.

We believe that our results and the described software will be of interest to a large spectrum of readers from pure biologists to bio-informaticians and that publication in Nucleic Acid Research is particularly suited to reach this broad readership.

Yours sincerely,

F Schweisguth and V Hakim