Evaluation of the potential for secondary structure formation in repeated bacterial sequences: identification of 58 families sharing common motifs.

ID - 136

Cozzuto Luca[1,2], Petrillo Mauro[1], Silvestro Giustina[3], Di Nocera Pier Paolo[3], Paolella Giovanni[1,4]
[1]CEINGE Biotecnologie Avanzate, Napoli, Italy
[2]S.E.M.M. - European School of Molecular Medicine - Naples site, Italy
[3]DBPCM, Università degli Studi di Napoli Federico II, Napoli, Italy
[4]DBBM, Università degli Studi di Napoli Federico II, Napoli, Italy

Motivation

Detection of functional elements in genomic sequences by automated methods represents a major research goal in post-genomic analyses. Computational approaches to the analysis of secondary structure may be used to find biological entities such as regulatory hairpins, non-coding RNAs or protein binding domains otherwise not easily detectable. Systematic analysis of high stability stem-loop structures (SLSs), within a representative set of 40 bacterial genomes, showed that their number is larger than expected on the basis of sequence length and composition, and that some SLSs have strong sequence similarity [Petrillo M et al. 2006; 7:170]. Although less frequent than in eukaryotes, families of repeated short sequences have been described in several bacterial genomes and some of them, such as Pu-Bime and ERIC in E. coli, are known to contain transcribed SLSs, possibly active at the RNA level. For this reason, an automatic pipeline was developed to detect, among the available collection, SLS families defined by sequence similarity and sharing a common secondary structure.

Methods

Starting from the full SLS population previously identified, for each bacterial species, SLSs predicted to fold with a free energy lower than 5 Kcal/mol were selected and filtered to eliminate those falling within either mature stable RNA species (tRNAs, rRNAs) or known ISs. These sequences were grouped by using a BLAST-MCL procedure described by Enright A.J. et al in a very stringent way, in order to obtain homogenous clusters. Resulting clusters were then re-grouped into candidate families by using the same clustering procedure, but in a less stringent way. For each family overlapping elements were fused into SLS-containing regions (SCRs) and a combined cyclic procedure based on a Hidden Markov Model (HMM) genome search was performed in order to define to all family members and their boundaries. Finally, manual refinement was performed to combine equivalent models, which had escaped previous identification. Two different approaches were attempted to evaluate the aptitude of sequences from the detected families to fold into a secondary structure: the probability of non-random folding was tested by using RANDFOLD on SLSs contained within each family; the ability to form conserved secondary structures, was measured by using RNAz. The presence of aligned SLSs in agreement with the structures predicted by RNAz was also evaluated.

Results

The procedure led to the identification of 92 families: most of them (66) are defined by a model shorter than 200 bps, while 24 vary between 200 bps and 1 Kb; only 2 are larger. Analysis of available literature showed that 25 correspond to more or less extensively described families and that 67 appear to be novel. For 58 families, a common secondary structure was detected by RNAz: in 46 of them the predicted structure is corroborated by the stacking of the originally found SLSs. In these families, SLSs also tend to be positive to the RANDFOLD test for $p<=0.005$, a good indicator of structured sequences. The procedure mostly identified simple SLS-based families, as expected given the starting SLS populations, but also identified complex structures, such as Efa-1 in E. faecalis and Pae-1 in P. aeruginosa. Most structured families (31 of 58) are preferentially located within intergenic regions. The identified families also include 34 ones where no strong evidence was found for the formation of common secondary structures. It is interesting to note that most of them (20) are located within coding regions: this preferential localization may reflect the lower tendency of such regions to assume a structured form.

Email: cozzuto@ceinge.unina.it