Systematic identification of stem-loop containing sequence families in bacterial genomes

Cozzuto L (1,2), Petrillo M (1), Silvestro G (3), Di Nocera PP (3), Paolella G (1,4)

(1) CEINGE Biotecnologie Avanzate, Napoli, Italy
(2) S.E.M.M. - European School of Molecular Medicine - Naples site, Italy
(3) DBPCM, Universita' degli Studi di Napoli Federico II, Napoli, Italy
(4) Dep. SAVA, Universita' del Molise, Campobasso, Italy

Motivation

Many bacterial genomes are known to contain families of repeated sequences of variable length and copy number. Many of them have been shown to contain a common stem-loop structure (SLS), involved in biological processes ranging from regulation of transcription to termination, to RNA stabilization. The availability of a now large number of sequenced bacterial genomes, represents an opportunity to identify novel families of such SLS containing repeated sequences. To carry out a systematic analysis of high stability stem-loop structures, predicted at both DNA and RNA level, an automatic pipeline was developed, based on Markov clustering algorithm (MCL). On each identified family, extensive annotations both at sequence and structure level were performed.

Methods

A pipeline has been previously described, able to identify, annotate and store into a relational database all potential SLSs within 40 completely sequenced bacterial genomes, representative of the bacterial world, from firmicutes to proteobacteria. From this population, SLSs predicted to fold with a free energy lower than -5 Kcal/mol were selected and filtered to eliminate those falling within tRNA/rRNA genes or known IS sequences. For each genome, SLSs were clustered according to a procedure based on BLAST and MCL programs (Altschul, S.F. and al. 1990 J. Mol. Biol. 215:403-410 and Enright A.J. et al Nucleic Acids Res. 2002 30[7]:1575-1584, respectively): an all-againstall SLSs BLAST comparison was performed for the creation of an e-value based distance matrix, which was in turn fed to MCL to produce a set of 'raw' clusters. Overlapping SLSs were fused into larger "regions". Members of the same cluster were aligned to produce a consensus sequence and the quality of the alignments was statistically evaluated. Genomic coordinates were used to classify clusters as either interspersed or tandemly repeated, and to join related raw clusters into the final refined set. BLAST analysis was used to identify clusters corresponding to repeated sequence families previously described in the literature. The families of identified sequences were analyzed to evaluate the reliability of the predicted secondary structure.

Results

Starting from a number of SLSs ranging from 6,534 (Buchnera) to 214,459 of (B. bronchiseptica) a first strand-dependent clustering step identified 717 "raw-clusters", each composed by a minimum of 7 regions. Clean-up and reclustering grouped raw-clusters in 466 second-order clusters. Clustering quality was assessed by alignment of the raw-clusters, which showed average identity above 60% for over 80% of the clusters. The established consensus was longer than 30 bp for over 98% of them. Further refinement led to a final set of about 85 clusters, including 28 known families and 29 families of not previously described repeated sequences. Analysis of predicted RNA secondary structure shows that many clusters are composed, only or predominantly, of SLSs with a low probability of random-folding ($p <= 0.005$), a clear enrichment over the original population, encouraging further studies aimed to better characterize their structure. A web interface is under development to allow public access to the SLS family database.

Contact email: cozzuto@ceinge.unina.it