GenoCluster: A novel platform for Comparative Genomics

Debasis Dash*, Srinivasan Ramachandran and Samir K.Brahmachari

G.N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, CSIR, Mall Road, Delhi 110 007, India

ABSTRACT

Motivation: The availability of complete sequences of more than 280 genomes provides novel opportunities for in depth understanding of various biological phenomena through *in silico* comparative genomics. Identification of novel genes, assignment of function to gene products and their evaluation as potential drug targets is considered to be of prime importance. We have developed a suite of software programs GENE'D'CFER, PROTEOME CALKULATOR, PLHOST^{FA}, and SEAPATH and porting them into LINUX cluster to harness the enhanced computational power that aids in the prediction of prokaryotic genes, functional assignment of encoded products, identification of adhesins with the help of Artificial Neural Network based algorithms.

Results: We have developed a generic and versatile new approach, designated Gene'D'cfer (GDC), for prokaryotic gene identification. Unlike other existing methods, this approach employs peptides as markers for protein coding DNA sequences. GDC determines candidate genes among all possible ORFs in a given DNA sequence through the use of Artificial Neural Network (ANN) trained on a set of known peptide library. Potential ORFs are ranked according to a scoring scheme based on the abundance and distribution pattern of heptapeptides along the ORF. ORFs identified by GDC can be overlaid with other features using complementary software programs for ribosomal binding sites, promoter sequences, transcription start sites, or codon biases for further examination. An analysis of 18 completely sequenced prokaryotic genomes has been carried out to demonstrate the capabilities of GDC. In addition, GDC has been applied on various strains of SARS virus and 4 new genes were predicted.

Delineating Conserved and Variable regions in sequences is of fundamental biological importance. Conserved regions are strong indicators for phylogenetically conserved functional roles whereas variable regions are generally implicated in auxiliary roles, often related to specific cases. The traditional approaches towards this objective involve comparing the homologous sequences using multiple sequence alignment algorithms. This approach although sound in theory is limited in terms of its speed and is not suited for high capacity. Although this limitation can be overcome in principle using powerful computers with enlarged memory, the results need careful scrutiny by the user. In most cases, users simply wish to know, in a first pass, the conserved and variable regions. PROTEOME CALKULATOR meets this need by offering a rapid approach to compare all the proteins (proteome) of a species with proteomes of other species using a peptide library approach.

Prediction of surface proteins involved in virulence from the complete sequences of proteomes of pathogens can greatly facilitate the development of ant-infectives towards eradicating infectious diseases. ANN was used to develop SEAPATH, which predicts the probability of a protein being an adhesin (Pad) based on 105 compositional properties of a sequence. SEAPATH draws upon the base algorithm SPAAN, which had optimal sensitivity of 89% and specificity of 100% and could identify 97.4% of adhesins from a wide range of bacterial pathogens causing a broad range of diseases in humans and other hosts. In the case of Severe Acute Respiratory Syndrome (SARS) associated Human corona virus, the spike glycoprotein, and nsps (nsp2, nsp5, nsp6 and nsp7) of SARS virus were identified with adhesin-like characteristics and offer new leads for rapid experimental testing.