Poster A-28

Delineating the Functional Components of Eukaryotic Genomes

Authors:
Jonathan M. Keith (*Department of Mathematics, University of Queensland*)
Stuart Stephen (*Institute for Molecular Bioscience, University of Queensland*)
Peter Adams (*Department of Mathematics, University of Queensland*)
John S. Mattick (*Institute for Molecular Bioscience, University of Queensland*)

Short Abstract: This poster presents an extensive collection of slowly evolving (and thus potentially functional) non-protein-coding elements in mammalian and insect genomes. These were discovered using recently developed Bayesian segmentation algorithms. An important aspect of the analysis is that it accounts for variation of GC content observed in mammalian genomes.

Long Abstract:

A surprising finding that has emerged from recent comparative genomic studies is that the proportion of conserved non-protein-coding sequence in mammalian and insect genomes is several times larger than the proportion of protein-coding sequence [1-6]. Many of these non-protein-coding elements are poorly understood, and indeed most such elements are yet to be delineated. Finding and identifying these elements is an important and difficult task.

In this poster we present new results obtained using a recently developed method for delineating the most slowly evolving (and hence probably conserved) fraction of eukaryotic genomes [7]. The method is based on segmenting pair-wise whole-genome alignments into blocks in which the proportion of single-base matches is approximately constant. A Bayesian model is developed and used to determine the posterior distribution over the space of all possible segmentations. This posterior distribution is then sampled using a Markov chain Monte Carlo technique known as the Generalized Gibbs Sampler. The result is a sample of 1000 or more plausible segmentations of the alignment. The distribution of evolutionary rate across segments is modelled as a mixture of beta distributions, and the sample is used to estimate, via Monte Carlo integration, the probability that each character position belongs to the most slowly evolving component of the mixture. An attractive browser has been developed to view the resulting profile.

The method is applied to whole-genome alignments of closely related Drosophila species. A list of elements that are consistently slowly evolving throughout these species is identified. The method is also applied to alignments of mammalian species including human, mouse and dog. Techniques for dealing with the confounding effect of variable GC content observed in mammalian genomes are described. The results of the mammalian analyses are compared to recent results obtained by Lunter et al. [8].

[1] Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.
[2] Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D (2003) The

share of human genomic DNA under selection estimated from human-mouse genomic alignments, Cold Spring Harb Symp Quant Biol, 68:245-254.
[3] Margulies EH, Blanchette M, NISC Comparative Sequencing Program, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome Research 13:2507-2518.
[4] Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.
[5] Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Research 15: 901-913.
[6] Andolfatto P (2005) Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149-1152.
[7] Keith JM, Adams P, Stephen S, Mattick JS (submitted) Delineating the slowly and rapidly evolving fractions of eukaryotic genomes via Bayesian sequence segmentation.
[8] Lunter G, Ponting CP, Hein J (2006) Genome-wide identification of human functional DNA using a neutral indel model. PLoS Computational Biology 2(1): e5.