HUMAN-CHIMPANZEE PROPERTY-DEPENDANT COMPARISONS ON CHROMOSOMES 21

Deyneko I.V.*1,2, KalybaevaY.M.1, Kel A.E.3, Blöcker H.4, Kauer G.4
1 Department of Genome Analysis, GBF, Braunschweig, Germany; 2 Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; 3 BIOBASE GmbH, Wolfenbüttel, Germany; 4 University of Applied Sciences, Emden, Germany
* Corresponding authors: e-mail: ide@gbf.de; bloecker@gbf.de

Key words: comparative genomics, property-dependant similarities, FeatureScan

SUMMARY

Motivation: Identification of different functional elements and their properties is a fundamental need in biomedical research, and phylogenetic comparisons form a solid basis for this task. But being applied to close genomes they can not “feel” such small, but nevertheless phenotypically important differences in sequence.

Results: In this work we present the comparative analysis of two evolutionary close genomes – human and chimpanzee. In contrast to previous studies we focus on evolutionary differences and evaluate changes in DNA properties rather than count nucleotide mismatches. In our examples, we find that nucleotide mismatches in promoters were probably introduced in a correlated manner during the course of evolution. Such property-dependant conservation of promoters is significantly different from nucleotide conservation and shows significant functional biases.

INTRODUCTION

Comparative genomics provides a powerful approach for investigating newly sequenced genomes. To date the great advances in this area were achieved in the identification of either protein-coding or non-protein-coding functional elements (Xie et al., 2005). The application of traditional techniques for comparisons of evolutionary close species (as for example, human and chimpanzee) is not effective, since a substantial amount of functional elements are masked by prolonged conserved non-functional DNA stretches.

Genomic sequences of human and chimpanzee are roughly 99 % conserved, but obvious differences in both, appearance and behaviour are surely beyond any doubt. Therefore, phylogenetic comparisons together with novel methods for long-range sequence comparison, which are able to “feel” small differences between genomes, have the potential to reveal certain key mechanisms in evolution.

In a first test case, we decided to compare promoters of genes located on chromosomes 21 from human and chimpanzee with our novel signal-theory-based approach to long range property-dependant sequence comparison (Deyneko et al., 2005).

DATA AND METHODS

Using the Ensembl database (v. 37.3a) we built up two sets of upstream sequences of all orthologous genes on chromosome 21 from human and chimpanzee. Each sequence
spans 2Kb of the upstream region starting from the 5′-most annotated transcription start site. The total number of sequences in each set was 229.

Property-dependent similarities of orthologous promoters were calculated using our tool FeatureScan (Deyneko et al., 2004) with “melting enthalpy” as DNA characteristic. Algorithmically, FeatureScan originates from proven methodologies in image analysis and speech recognition. The current implementation is based on a convolution method and can be described briefly in three main steps (for the detailed theoretical background see our earlier publication (Kauer, Blöcker, 2003)). First is a transformation of nucleotide sequences (pattern and investigated sequence) into numerical form, which we refer to as signals. At this step users have to decide which property may play an important role in their specific cases. Second is a computation of the correlation integral (i) of two signals \( f \) and \( g \), which can be rewritten using Fourier transformants \( F \) and \( G \) yielding (iii). Assuming to have direct and inverse Fourier transformations implemented (in our case it is hardware implemented), the entire integral is reduced just to a multiplication. The final step is looking for shift values \( y \) which will define possible matches of the sequences, so that the difference between correlation (iii) and autocorrelation (ii) integrals is less than the predefined threshold.

\[
\text{Corr}(y) = \int f(x) \cdot g(x-y)dx
\]  
(i)

\[
\text{AutoCorr} = \int g(x) \cdot g(x)dx
\]  
(ii)

\[
\text{Corr}(y) = \text{InverseFourierT}\left\{F(y) \cdot G(y)\right\}
\]  
(iii)

For statistical evaluation of the results we tried to simulate the evolution of primate promoters under pure random and transition/transversion biased models with transition/transversion rate ratio of 4.31 (Rosenberg et al., 2003). The set of \( 10^4 \) random sequences were generated, which were of the same length (2Kb) and the same mononucleotide context as promoter regions. Calculation of \( p\)-values was done by summing the “tail” of the binomial distribution assuming the random value follows Bernoulli trials scheme.

RESULTS

Following the scheme described above, we calculated the similarities of chimpanzee and human promoter regions using letter-based (ClustalW) and signal-theoretical approaches (FeatureScan). The distribution of the number of promoters vs. similarity is shown in Fig. 1.

It can be easily seen that promoters of chimpanzee and human genes, which differ by less than 2 % of nucleotides, show significantly higher similarity by FeatureScan than expected (Fig. 1; white bar is higher than light grey and dark grey). If we consider the transition/transversion bias, than this increase will even be improved. We found that 139 out of 198 orthologous promoter pairs showed higher signal similarity than can be expected, which corresponds to a \( p\)-value of \( 2.43 \times 10^{-8} \).

Using the EMBL-EBI gene ontology classification, we examined the gene distribution which showed high signal similarity of their promoters. A subset of 15 genes involved in the molecular function “metal ion binding” (GO:0046872) and another subset of 11 genes involved in “nucleotide binding” (GO:0000166) were identified. These observations have an estimated \( p\)-value of \( 4.9 \times 10^{-3} \) and \( 4.52 \times 10^{-5} \), respectively. Corresponding subsets of 3 and 4 genes, can be identified in the set of genes with low promoter similarity.
DISCUSSION

The main advantage of phylogeny of close species is the ability to “see” evolutionary tendencies, which are not yet “drowned” in mutation chaos. Here we investigated the similarity of 2Kb promoter regions of human and chimpanzee genes, which is based on melting enthalpy characteristic of DNA. The observed statistically significant overrepresentation of promoters with high property-dependant similarity encouraged us to speculate, that single mutations occurring in evolution tend to compensate disturbances involved by others to retain the “original” function. As we may conclude from the presented results, a nucleotide substitution decreasing the melting temperature of a locus, may induce evolutionary pressure for further changes in the close vicinity to outbalance the first mutation.

The promising advantages of property-dependant similarity measures encourage us to tackle further interesting problems. It might be interesting to investigate changes of characteristics (melting temperature, conformation and others) caused by SNP mutations across the entire human genome. We believe that SNPs (both in coding and promoter regions) which are correlated with diseases or have phenotypic evidences, should be detectable/distinguishable in a single- or multidimensional property space.

ACKNOWLEDGEMENTS

Financial support by the German Federal Ministry for Education and Research through Projektträger Jülich (FKZ 031U210A) is gratefully acknowledged.
REFERENCES


BGRS’2006