Correlation of SNPs with Phylogenetic Footprints

Claudia Fried[1,2], Peter Ahnert[3], Peter F. Stadler[1,2]

[1]Bioinformatik, Institut für Informatik, Universität Leipzig, Kreuzstraße 7b, D-04103
Leipzig, Germany. Phone: ++49 341 149 5120; Fax: ++49 341 149 5119;
Email: claudia@bioinf.uni-leipzig.de.
[2]Institut für Theoretische Chemie und Molekulare Strukturbiologie Universität Wien,
Währingerstraße 17, A-1090 Wien, Austria
[3]IKIT/BBZ, Medizinische Fakultät, Universität Leipzig,
Johannisallee 30, 04103 Leipzig, Germany

Introduction. Extensive polymorphism in non-coding gene-regulatory sequences was recently reported in particular for the immune system [7]. This type of genetic variation could therefore be functionally and evolutionarily highly significant. A different pattern of polymorphisms between the coding and non-coding regions seem to distinguish "introvert genes" that code for proteins dealing with self molecules, and "extrovert genes" that are targeted towards foreign molecules that enter the body [6]. Non-coding polymorphism appear to dominate in introvert immune genes.

It is likely that systematic patterns in the distribution of polymorphism between coding, non-coding but regulatory functional, and non-functional sequences can be found also in other classes of protein-coding genes. Here we describe a systematic computational approach to address this problem and discuss its limitations based on the currently available data sources. For concreteness we focus on single nucleotide polymorphisms (SNPs) and use a portfolio of immune system related genes that were of interest in another context as illustrative example.

Materials and Methods. Location, validation status and other SNP properties were retrieved from NCBI and EBI databases using `ENSMART`[1], which was also used to localize exons and introns consistently with the SNP localization. The quality of SNP data in `dbSNP` [2] depends on the SNP detection method and on independent verification. We distinguish validated and non-validated SNPs in our analysis.

DNA sequences that are regulatory active presumably cover a substantial part of the intergenic regions. On the other hand, only a small number of transcription factor binding sites or promoter sequences are experimentally verified for any given gene. Functional non-coding sequences, however, evolve much slower than the surrounding non-functional DNA because they are subject to stabilizing selection. This is true in particular in vertebrates [2], while invertebrates often show a high rate of binding site turnover. Comparative sequence analysis can therefore be used to detect functional non-coding DNA sequences in the vicinity of the genes of interest. This technique is known as *phylogenetic footprinting*.

Recently we have presented the program `tracker` as an efficient tool for surveying phylogenetic footprints in large datasets [9]. It is based on the initial computation of all pairwise `blast` alignments from the input sequences with a non-restrictive parameter setting. A hierarchy of filtering steps then removes insignificant matches. The most complicated step is the combination of overlapping alignments to maximal cliques of mutually consistent alignments, thereby producing local multiple alignments of the conserved regions. For details we refer to [9].

For the purpose of the present study we use orthologous sequences retrieved from the genome databases for *homo sapiens*, *mus musculus*, and *rattus norvegicus*. In some cases only the rat or the mouse sequence was available. In all cases the DNA sequence extending 5000nt upstream and downstream of the gene was retrieved.

[1] freely available on the world wide web at `http://www.ensembl.org/EnsMart/`

[2] URL: `http://www.ncbi.nlm.nih.gov/SNP/index.html`

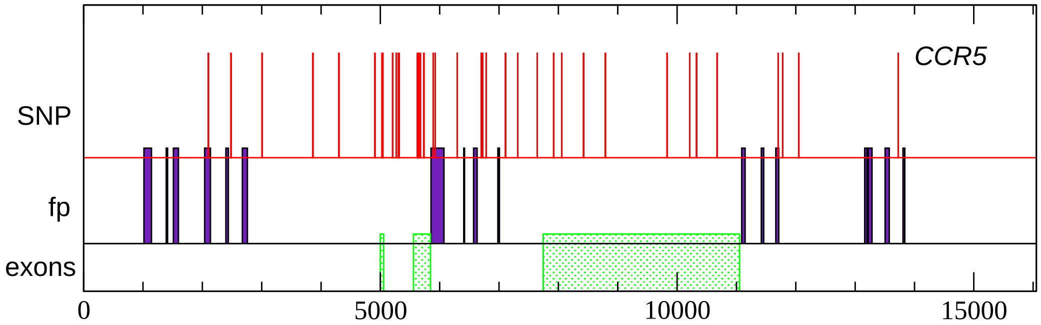

Figure 1. Graphical representation of the SNP distribution of the CCR5 gene. Acccesion number: NT_0058250, the part of the sequence that we used in our analysis spans the entire gene and surrounding region 5000 basepairs upstream and downstream of the gene

The genes analyzed here were orignally selected for a genotype-phenotype association study of rheumatoid arthritis susceptibility. Most of the selected genes play a role in cytokine balance, which has been suggested to play a significant role in rheumatoid arthritis [5]. Other genes were selected for their role in cartilage metabolism [8] or their known or suggested association with rheumatoid arthritis in other study populations [1].

First Results. Figure 1 visualizes a typical data set. We distinguish between SNPs in the exons, in phylogenetic footprints, and in the remaining DNA that presumably is non-functional. Intuitively one would expect that the rate of occurrance of SNPs in the non-functional DNA is largest since there it is not subject to selection, while mutations should be selected against in coding and regulatory sequences.

Table 1 summarizes the distribution of SNPs in 20 genes. We observe a rather large variance in the density ρ of SNPs around the mean of about 2 SNPs per 1000nt.

The most surprising result is that SNPs are *overrepresented* in exons. We use Fisher's exact test[3] [3, 4] to assess the statistical signficance of the differences in SNP density in exons and footprints compared to the remaining non-functional DNA. Significant overrepresentation is found only for six genes. A p-value of $p = 0.05$ or smaller indicates a significant difference in the distribution. Almost certainly this effect is a bias in the SNP database: If the SNP overrepresentation in the exons of some immune genes were real we would have to postulate an increased mutation rate in the coding sequences compared to non-functional surrounding DNA. Since SNPs are often obtained from cDNA comparison more data are available for the exons than for the surrounding non-coding DNA.

Table 1. Distribution of SNPs in functional and non-functional sequences of several immune genes. Fisher's exact test was used to assess the significance level of over- and underrepresentation in the columns labelled FT. The column ρ lists the average number of SNPs per 1000nt.

Gene	PF	SNP		Exons			Footprints			other	
	#	#	ρ	length	SNP	FT	length	SNP	FT	length	SNP
Total	1629	820	1.92	33279	149	↑↑↑	107012	194		285823	477
Frac.				0.0781	0.1817		0.2366	0.2362		0.6708	0.5817
Fisher's	SNP overrepresented			8.40×10^{-22}			0.1176				
p-value	SNP underrepresented			1			0.845				

[3] Computations were performed using the webservice `http://www.matforsk.no/ola/fisher.htm`.

The distribution of SNPs in `tracker`-predicted phylogenetic footprints does not significantly differ from distribution in non-functional DNA on average. For six genes, however, we find a highly significant underrepresentation of SNPs in the putative regulatory sequences. As methods for SNP detection do not distinguish between functional and non-functional non-coding DNA we should expect that differences in SNP distribution between phylogenetic footprints and non-functional DNA are biological rather than caused by database biases. Interestingly, the picture does not change when only the (small) subset of validated SNPs is used instead of all database entries.

Discussion. The human genome program has enabled and spawned many studies attempting to identify the role of genetic variation in the etiology and pathogenesis of many diseases. One approach to identify genes associated with a disease is to select candidate genes according to their known or assumed role in biological processes. The association of variants of these genes with the disease are then tested by genotyping a number of SNPs distributed along their genomic DNA. These SNPs either modify the function of the gene directly or are in linkage with such a function-modifying polymorphism. Clearly, SNPs located in the coding sequence or in phylogenetic footprints of the gene are most likely to be function-modifying and are therefore the best candidates for further experimental study.

Conversely, the distribution of SNPs in exons, regulatory elements, and non-functional background potentially provides direct information on selection pressures acting on various components of a gene. In order to access this information unbiased subsets of SNP databases will have to be extracted e.g. by removing all entries that are obtained from mRNAs.

In this short contribution we have shown that the statistical analysis of the relationships between polymorphisms in the human genome and functional DNA regions is computationally feasible. Biologically interesting results, however, will have to await more extensive surveys that include larger sets of genes. This will allow the application of clustering techniques to identify protein families whose genes have particular SNP distributions among different types of surrounding non-coding DNA, e.g. GpC islands, proximal promotors, transcription factor binding sites, introns.

Acknowlegements. This work is supported in part by the *DFG* Bioinformatics Initiative.

References

[1] C. G. Baerwald, C. Mok, M. Tickly, C. S. Lau, B. P. Wordsworth, B. Ollier, G. S. Panayi, and J. S. Lanchbury. Corticotropin releasing hormone (CRH) promoter polymorphisms in various ethnic groups of patients with rheumatoid arthritis. *Z. Rheumatol.*, 59:29–34, 2000.

[2] J. W. Fickett and W. W. Wasserman. Discovery and modeling of transcriptional regulatory regions. *Current Opinion in Biotech.*, 11:19–24, 2000.

[3] R. A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society Series A*, 98:39–54, 1935.

[4] R. A. Fisher. Confidence limits for a cross-product ratio. *Australian Journal of Statistics*, 4:41, 1962.

[5] P. Miossec. Pro- and antiinflammatory cytokine balance in rheumatoid arthritis. *Clin. Exp. Rheumatol.*, 13:S13–S16, 1995.

[6] A. Mitchison. Partitioning of genetic variation between regulatory and coding gene segments: the predominance of software variation in genes encoding introvert proteins. *Immunogenetics*, 46:46–52, 1997.

[7] N. A. Mitchison. Polymorphism in regulatory gene sequences. *Genome Biology*, 2:2001.1–2001.6, 2000.

[8] G. Murphy, V. Knauper, S. Atkinson, G. Butler, W. English, M. Hutton, J. Stracke, and I. Clark. Matrix metalloproteinases in arthritic disease. *Arthritis Res.*, 4:S39–S49, 2002.

[9] S. J. Prohaska, C. Fried, C. Flamm, G. P. Wagner, and P. F. Stadler. Surveying phylogenetic footprints in large gene clusters: Applications to *Hox* cluster duplications. *Mol. Phyl. Evol.*, 2003. in press.