



## Site2genome: locating short DNA sequences in whole genomes

Martin C. Frith<sup>1,†</sup>, Anason S. Halees<sup>1,†</sup>, Ulla Hansen<sup>1,2</sup> and Zhiping Weng<sup>1,3,\*</sup>

<sup>1</sup>Bioinformatics Program, <sup>2</sup>Biology Department and <sup>3</sup>Biomedical Engineering Department, Boston University, 44 Cummington Street, Boston, MA 02215, USA

Received on April 15, 2003; revised on December 10, 2003; accepted on January 6, 2004

Advance Access publication February 12, 2004

### ABSTRACT

**Summary:** Many biological papers describe short, functional DNA sites without specifying their exact positions in the genome. We have developed a Web server that automates the tedious task of locating such sites in eukaryotic genomes, thus giving access to the context of rich annotations that are increasingly available for genome sequences.

**Availability:** <http://zlab.bu.edu/site2genome/>

**Contact:** [zhiping@bu.edu](mailto:zhiping@bu.edu)

We performed recently an analysis of 56 previously published estrogen receptor binding sites, where among other things we were interested in the distance of each site from the start of transcription and the inter-species similarity of the sites. (O'Lone, R., Frith, M.C., Karlsson, E.K. and Hansen, U. Genomic targets of nuclear estrogen receptors. Manuscript submitted.) Since 54 of these sites were from humans, rat or mouse, for which genome assemblies with rich annotations are available, a logical approach was to locate the sites in the genomes and make use of existing transcript annotations and inter-species alignments. Most of the sites were too short (a dozen base-pairs or so) to be located uniquely in the genome using a tool such as BLAT. Therefore, we adopted the following hierarchical approach:

- (1) Identify a longer sequence that contains the site. In theory, any published site should come from a sequence which is deposited in GenBank (Benson *et al.*, 2003) and for which a GenBank record exists.
- (2) Locate the site within the longer sequence.
- (3) Align the longer sequence to the genome.
- (4) Transfer the coordinates of the site within the longer sequence to coordinates within the genome.

This procedure is quite tedious and can go awry for many reasons. The publication may not provide the GenBank accession

number for the longer sequence. The site may occur more than once in the longer sequence, or it may not be present at all. The sequence may not align to the genome, or it may align to too many locations. Finally, the sequence may only partially align to the genome, excluding the region that contains the site.

Site2genome is a Web server that automates the procedure described above. The user pastes in a list of up to 100 sites, supplying the following information for each site: a sequence identifier (GenBank/EMBL/DDBJ accession number or NCBI GI number) or a descriptive gene name, the sequence of the site and an optional description of the site. The program automatically fetches the sequences from GenBank using Efetch ([http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils\\_help.html](http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html)), locates the sites in the sequences (considering both DNA strands and handling palindromic sites as a special case) and then truncates the sequences to a suitable length for feeding into the BLAT Web server. It uses the genus from the ORGANISM field in the GenBank record to determine which genome to align each sequence to, drives the BLAT program at the UCSC website Kent (2002) and uses the BLAT alignment for each sequence to infer the coordinates of each site within the genome. It finally prints a table showing the species, chromosome number, coordinates and strand for each site.

In addition, site2genome provides a link for each site in the results table, whereby the site can be viewed in the UCSC genome browser alongside all other annotations for that genomic region, using the browser's facility for uploading external annotation tracks (Kent *et al.*, 2002).

When only a gene name is provided for the longer sequence, the user has the option to search GenBank, LocusLink, Ensembl or any combination of these resources. It is highly likely that the gene name would match multiple sequence records; therefore, the user is presented with a list of all hits, whereby one or multiple selections can be made. We facilitate the selection process by providing for each choice a hyperlink to its GenBank record. It is worth noting that searching a gene name through most databases is most likely to return records that contain the transcribed portion of the gene (cDNA or

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

mRNA sequences). More often, however, the user is looking for sites that are located in the upstream regulatory region of a genomic DNA sequence record. Site2genome will indicate the type of each record (DNA/mRNA), with DNA hits presented first. Furthermore, the user has the option of suppressing RNA records. GenBank and Ensembl searches are restricted to the titles of the records, while LocusLink searches are over the entire records and thus may retrieve sequences based on the descriptions of their functions, disease implications or other associations with the locus.

Sometimes the site occurs multiple times in the longer sequence, or the longer sequence matches with multiple locations of the genome. Site2genome will report all matches in the result table and display warning messages to indicate the non-uniqueness of matches for the sites in question. For other types of errors that site2genome cannot correct automatically (e.g. the site is not present in the GenBank sequence or the GenBank sequence does not align to the genome), an error message is displayed, stating the reason. In this way the problematic cases are rapidly identified, and the user can attempt remedies such as adding flanking bases to the site or using an alternative GenBank sequence.

Our experiences lead to the following recommendations for how authors should present data on functional sites so as to make their discoveries more readily accessible. First, an accession number of a public sequence record containing the site should be clearly specified, preferably indicating the version of the sequence (i.e. M73700.1 or GI:619784 rather than M73700). Without a version number, site2genome would retrieve the newest version of the sequence record, which may not correspond to what the user has in mind. Second, sites should be described with sufficient flanking base-pairs that there is a reasonable chance of locating them uniquely in a longer sequence. Finally, the accuracy of the site should be

checked carefully (perhaps using site2genome) as we found an alarming number of cases where the site is inconsistent with the sequence record.

Currently, site2genome supports five genomes: human, mouse, rat, *Drosophila melanogaster* and *Caenorhabditis elegans*. New genomes will be added as soon as they are included by UCSC, LocusLink and Ensembl. Site2genome is a classic bioinformatics application in that it facilitates a task that should in principle be straightforward but is in reality extremely tedious. As the human and other genomes achieve finished status, annotation efforts will increasingly be applied to and anchored upon these reference genomes, and it will become more essential to assimilate data on functional sites into this genomic framework.

## ACKNOWLEDGEMENTS

We thank members of the ZLAB for motivation and advice. We thank the anonymous journal reviewers for their suggestions. M.C.F. is a Howard Hughes Medical Institute Predoctoral Fellow. A.S.H. and Z.W. are supported in part by NSF grants DBI-0078194 and MRI DBI-0116574 and NIH grant P20GM066401. U.H. was supported in part by NIH grants R01CA081157 and P20GM066401.

## REFERENCES

- Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
- Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
- Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.