248–250 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

INFOGENE: a database of known gene structures and predicted genes and proteins in sequences of genome sequencing projects

Victor V. Solovyev* and Asaf A. Salamov

The Sanger Centre, Hinxton, Cambridge CB10 1SA, UK

Received September 1, 1998; Revised September 30, 1998; Accepted November 3, 1998

ABSTRACT

INFOGENE is a database of known and predicted gene structures with descriptions of basic functional signals and gene components. It provides a possibility to create compilations of sequences with a given gene feature as well as to accumulate and analyze predicted genes in finished and unfinished sequences from genome sequencing projects. Protein sequence similarity searches in the database of predicted proteins is offered through the BLASTP program. INFOGENE is realized under the Sequence Retrieval System that provides useful links with the other informational databases. The database is available through the WWW server of the Computational Genomics Group at http://genomic.sanger.ac.uk/db.html

Large scale genome sequencing projects currently produce hundreds of megabases each year. The major sequencing centers are in the process of scaling up their throughput over the next few years. Shifting efforts toward sequencing gene-rich rather than random regions might provide the sequence of most of human genes during the next 3 years. Moreover, the initiative to create by 2001 a 'rough draft' of the human genome can allow other scientists to proceed more rapidly with discovering disease genes (1). However, the sequence itself does not always provide the knowledge of gene coding regions, which usually only cover a pretty small fraction of genomic DNA. Also, we cannot expect their rapid identification in the near future by pure experimental approach for such an enormous volume of sequence data. The value of sequence information for the biomedical community will strongly depend on availability of candidate genes computationally predicted in these sequences.

The aim of this work was to create the information resource of known and predicted gene structures in major model organisms as Human, Mouse, *Drosophila* and *Arabidopsis*. The structural components of the INFOGENE database are presented in Figure 1.

INFOGENE is realized under the Sequence Retrieval System (SRS), developed at the European Bioinformatics Institute (2). This system provides a possibility to connect the database with existing data resources (such as TRRD, Transfac, Swiss-Prot, GenBank, etc.) and to make complex queries over several databases using the WWW server. In SRS any retrieval command, logical operations with sets that were obtained by previous queries, links between sets of different databanks, or a combination of all can be easily expressed in the SRS query language.

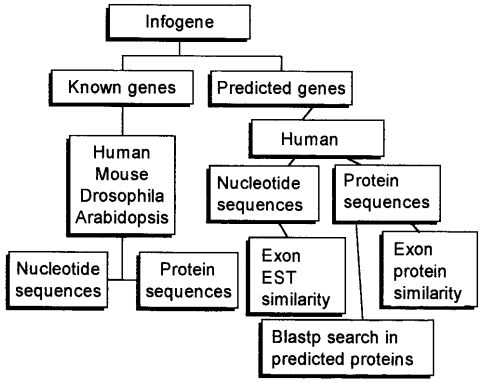

Figure 1. Structure of the INFOGENE database.

KNOWN GENE STRUCTURES DATABASE

Primary reasons for generating known gene structure databases are to: (i) have a collection of known gene structures with their main features presented in the form convenient for retrieval entries including some functional characteristics; (ii) easily create subsets of genes or exons with a given set of features; (iii) check availability of genes with particular features; (iv) have links to different informational databases providing regulatory site locations or other information for a particular gene (polymorphism or mutations underlying inherited disease, for example); and (v) provide the possibility to make links between similar genes of different model organisms.

Today the problem of reliable gene prediction in human genomic DNA is still open. The best multiple gene prediction programs such as GeneScan (3) (probabilistic approach) and Fgenes (http://genomic.sanger.ac.uk/gf/gf.html) [pattern recognition approach (4,6)] were tested mostly on short sequences containing one gene. The recent test of these programs for 660 human genes shows that the programs can correctly predict ~80%

*To whom correspondence should be addressed. Tel: +44 1223 494799; Fax: +44 1223 494919; Email: solovyev@sanger.ac.uk

```
LID    MMTNFAB      GenBank MOUSE_G
DAT 19980713
LCO     7208 bp    DNA             ROD       11-MAY-1993
LDE  Mouse complete TNF locus (TNF=tumor necrosis factor).
ORG  Mus musculus
LKE   B1 repetitive sequence; lymphotoxin; tumor necrosis factor.
LFT  mang nasp ytss nmts natp yftr npse
LGN    2  2  2     7208       0
GID GMM000399 direct
GPR  TNF-beta
IND
PSD  SWISS-PROT: P09225
GFT  nasp natp ytss nmts yftr nex5 mexo fcds rsiz nsto
GEC     4   3   0   0     916     609     202     609
TSS     1193     3207  1  be
TAT     1174  yTAT  c tata a  dba
EXO     1193     1345  f  c    gt
EXO     1709     1813  i  ag   gt
EXO     1897     1996  i  ag   gt
EXO     2221     3207  l  ag   c
CDS     1718     1813  f atg  gt I
CDS     1897     1996  i  ag  gt V
CDS     2221     2633  l  ag tag
POA     3186 c aataaa c com
SEQ   MMTNFAB     GMM000399
GRE     0 nrep dba
DWG GMM000400
GUS     1669 ctccgctacacacacacactctctctctctctctcagcaggttctccaca
GDS     2633 gattctaaagaaacccaagaattggattccaggcctccatcctgaccgtt
```

Figure 2. Example of an INFOGENE entry corresponding to MMTNFAB GenBank locus. Description of the first gene of this locus is presented. Features of coding regions (CDS field) are: (i) start; (ii) end; (iii) start codon/acceptor splice site short consensus; (iv) stop codon/donor splice site short consensus; (v) type of exon: f, i, l, o are the initial, internal, terminal and single CDS. A table of all codes and their explanations is available at the database main page http://genomic.sanger.ac.uk/db.html

```
LID   HSCPH70   GenBank   TEST
DAT   Wed Oct 21 12:11:41 BST 1998
LCO   6711 bp    0
ORG   Homo sapiens
LKE   REPEATS  GENES  PROTEINS
LFT   oneg   ytss   mexn
LFG   1  0   1  1
GID   GHS000100
GFT   direct  mexn   ytss   fcds
GEC   4  4691 173
TSS   1584
CDS     1660     1783  atg   gt  CDSf  16.02  fgenes-h
HOP   >gi|226256|prf||1503232A peptidyl-Pro cis trans isomerase [Sus
HPP       1       24       1       24      164       51   95.0  8e-07
HOE   >gi|2032671|gb|AA380129|AA380129 EST93310 Supt cells Homo sapiens
HPE       1       71       33      103      277      141  100.0  1e-32
CDS     4318     4406   ag   gt  CDSi   2.09  both
HOP   >gi|1633054|pdb|5CYH|A Homo sapiens >gi|1633056|pdb|4CYH|A Homo
HPP       1       30       33      62      164       70  100.0  1e-12
HOE   >gi|1997506|gb|AA345272|AA345272 EST51269 Gall bladder II Homo
HPE       1       89      133      221      349      176  100.0  2e-43
CDS     4628     4800   ag   gt  CDSi  18.97  both
HOP   >gi|1633054|pdb|5CYH|A Homo sapiens >gi|1633056|pdb|4CYH|A Homo
HPP       1       57       63      119      164      125  100.0  3e-29
HOE   >gi|2568551|gb|AA643333|AA643333 nr97e05.s1 NCI_CGAP_Pr25 Homo
HPE       1      173      512      340      633      343  100.0  2e-93
CDS     6215     6350   ag  taa  CDSl  15.68  both
HOP   >gi|1633054|pdb|5CYH|A Homo sapiens >gi|1633056|pdb|4CYH|A Homo
HPP       1       44      121      164      164       92  100.0  3e-19
HOE   >gi|665138|gb|T61895|T61895 yb93d08.s1 Homo sapiens cDNA clone 78735
HPE       1      136      339      204      500      270  100.0  2e-71
POA     6538 a aataaa a
SEQ   HSCPH70   GHS000100
GUS     1660   gtgccgttttgcagacgccaccgccgaggaaaaccgtgtactattagcca
GDS     6350   gtttgacttgtgttttatcttaaccaccagatcattccttctgtagctca
LRE   9 nrep
LRL       62       360   reverse    AluSx  SINE/Alu
LRL      363       654   reverse    AluJb  SINE/Alu
LRL     1217      1246    direct    MIR  SINE/MIR
LRL     1804      1859   reverse    GC_rich  Low_complexity
LRL     1933      1986    direct    GC_rich  Low_complexity
LRL     3207      3449    direct    AluSc  SINE/Alu
LRL     5028      5314   reverse    AluSx  SINE/Alu
LRL     5328      5622   reverse    AluSx  SINE/Alu
LRL     5755      6043   reverse    AluY  SINE/Alu
```

Figure 3. Example of an INFOGENE entry corresponding to predicted gene in the HSCPH70 sequence. Description of the first gene of this locus is presented. Locus features (LFT field): (i) 'mang/oneg', if multiple (mang) or single (oneg) genes is predicted in the locus; (ii) 'ytss', if at least one transcription start site is predicted in the locus, otherwise 'ntss'; (iii) 'sexn', if at least one predicted gene consists of a single exon, otherwise 'mexn'. A table of all codes and their explanations is available at http://genomic.sanger.ac.uk/db.html

of internal exons and just ~60% of 5′-exons. The prediction of multiple genes should be even less accurate. Therefore, it is important for developing further gene prediction programs to have as much information as possible about known genes and their functional signals, that will provide the learning and testing datasets.

We have developed a GenBank (5) parser *GeneParse* which produces a flat file with some description of genes and gene features including terms corresponding to exon types, regulatory elements, processes and characteristics of genes in a given GenBank sequence. To add this information to SRS we created several files with logical structure of INFOGENE database components and files with the syntax of their entries. Using these files the information about gene structure was written to SRS with indexing of specific words in the entries.

We can use the query language and search/retrieving software of SRS that will quickly extract sets of sequences with particular biological features. For example, genes where transcription start and stop sites are known or entries with multiple genes. The query language provides an effective usage of database information in investigation of significant characteristics of genes and their regulatory elements and assists in development of methods of their recognition. Currently it might take days to collect such information from the literature and visual analysis of GenBank entries. The current release of INFOGENE contains completely sequenced genes of the following model organisms: human (1835 genes), mouse (1038), *Drosophila* (970) and *Arabidopsis* (1726).

One example of INFOGENE entry corresponding to MMTNFAB locus of GenBank is presented in Figure 2. This locus includes two neighbor genes, whose exons and coding regions are characterized as well as the locations of the start, TATA-box and stop of transcription. In the LFT (Locus Features) field we have described this sequence with special keywords: mang (locus includes many genes), nasp (no alternative splicing), nmts (no multiple starts of transcription), natp (no alternative promoters), yftr (yes full transcript), npse (no pseudogenes). For example, using the ytss keyword we can retrieve a set of genes with known start of transcription. It is described for 251 human genes with completely sequenced coding regions. The full description of all keywords is presented on the web pages of INFOGENE.

DATABASE OF PREDICTED GENES

The primary reason for generating predicted gene structure databases is to provide positional cloners, gene hunters and others with the gene candidates observed in finished and unfinished genomic sequences. Recently a broad agreement has been reached amongst major genome centers and funding agencies in the US, the Sanger Centre and the Wellcome Trust in the UK to go ahead with a plan that will deliver all of the human sequence, part finished and part in draft, into the public domain by the end of 2001. Using gene prediction the scientific community can start experimental work with most human genes during the next 3 years because gene finding programs usually correctly predict at least the major part of exons in a gene sequence. Our experience shows that the accuracy of predictions is significantly lower for long genomic sequences than in usually presented tests with single genes (decreasing in the order of 10–20% with a high rate of false positive predictions). However, exons predicted simulta-

neously by several programs based on different approaches correspond to the real ones much more often than those predicted by a single program. For example, Fgenes and GeneScan predict exactly ~80% of real exons from 38 long or multigene genomic sequences with specificity 65% (true predicted/all predicted). If we take the subset of exons predicted by both programs, then the observed specificity is 92% and this set will include ~70% of all real exons.

We have used two of our programs Fgenes-p (4,6) and Fgenes-h [HMM based approach similar to GeneScan (3)] to predict genes in genomic sequences. The Blast (7) search is used to check if some of the predicted exons have similarity with known EST and protein sequences. Possible repeats in the sequence were annotated using RepeatMasker program (Smit and Green, unpublished; http://genome.washington.edu/RM/RepeatMasker.html).

The current release of INFOGENE contains 768 finished and 3698 unfinished loci. These sequences were produced by The Sanger Centre Human Genome Project. We plan to include in the database predicted genes in finished and unfinished sequences from other sequencing centers.

An example of a description of a predicted gene is presented in Figure 3. Fgenes-h predicts four coding exons, three correct and one partially correct. All identical exons predicted by both programs are correct. The keyword 'both' in the CDS field marks such exons and they often correspond to real ones as discussed above. The other features, which increase our confidence in predicted exons are produced by searching EST and protein databases. If any significant similarity is found it is presented in HOP (for protein homology) and HOE (for EST homology) database fields. Additional fields HPP and HPE provide information about similarities found. Features of protein homology (HPP field) are: (i) and (ii) the first and the last aligned positions of exon, respectively; (iii) and (iv) the first and the last aligned positions of the database protein, respectively; (v) the length of database protein; (vi) the score of the alignment calculated by BLASTP; (vii) sequence identity; (viii) E-value from BLASTP output. Similar features are presented for EST similarity (HPE field).

The INFOGENE database is available through the WWW server of the Computational Genomics Group at http://genomic.sanger.ac.uk/db.html . Users wishing to cite INFOGENE are asked to refer to this article.

REFERENCES

1 Wadman,M. (1998) *Nature*, **393**, 399–400.
2 Etzold,T., Ulyanov,A. and Argos,P. (1996) in Doolittle,R. (ed.), *Methods in Enzymology.*, vol. 266, pp. 114–128.
3 Burge,C. and Karlin,S. (1997) *J. Mol. Biol.*, **268**, 78–94.
4 Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1994) *Nucleic Acids Res.*, **22**, 5156–5163.
5 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouelette,B.F.F. (1998) *Nucleic Acids Res.*, **26**, 1–7.
6 Solovyev,V.V. and Salamov,A.A. (1997) In Rawling,C., Clark,D., Altman,R., Hunter,L., Lengauer,T. and Wodak,S. (eds), *Proceedings of ISMB*. AAAI Press, Halkidiki, Greece, pp. 294–302.
7 Altshul,S.F., Madden,T.L., Schiffer,A.A, Zhang,J., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.