RUMMAGE – a high-throughput sequence annotation system

Numerous bioinformatic tools are available to assist in the analysis and interpretation of genomic nucleotide sequence data. Each tool provides valuable pieces of information, which aid in the identification and characterization of functional and organizational elements, but none of these tools covers all aspects or is absolutely accurate. Therefore it would be very useful to compile the findings of different single programs, along with their heterogeneous output formats, into one homogeneous set of data. Viewing this data set in a graphical display with the detailed analysis results just one mouse click away would make it easy to retrieve the essential information and to get a comprehensive survey of large stretches of genomic sequence.

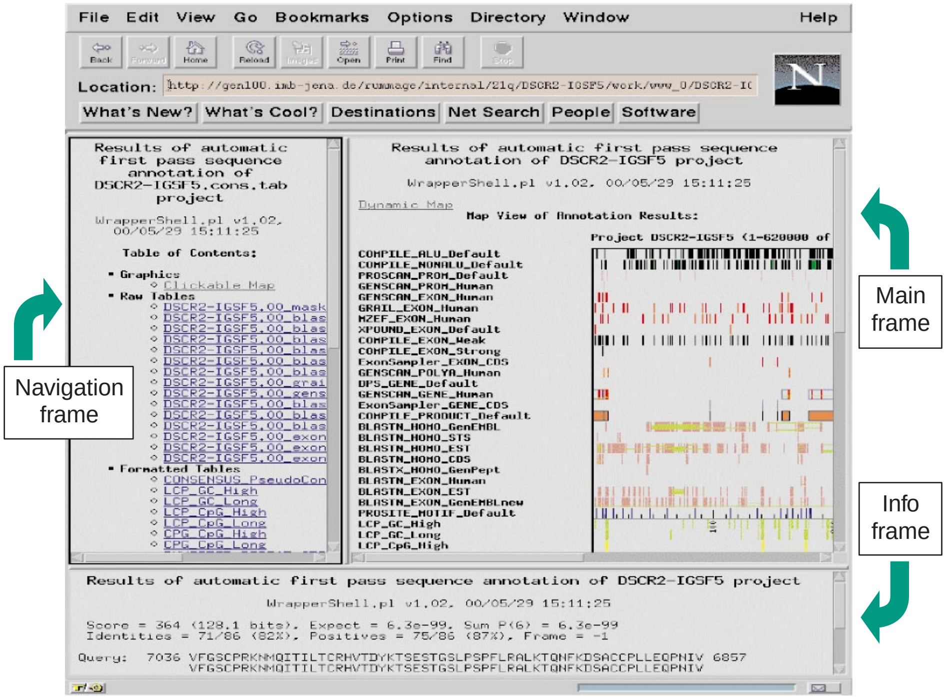
TABLE 1. Components of RUMMAGE[a]

Composition and repeats	Ref.	Exon Prediction	Ref.	Homology and motifs	Ref.	Validation of matched expressed sequence tags
LCP (local GC contents)	2	GENESCAN	4	BLASTN vs	10	EXONSAMPLER
CPG (CpG islands)	3	GRAIL	5	GENEMBL		
REPEAT Masker[b]		MZEF	6	BLASTX vs	10	EST2GENOME
(human, rodent)		FEXHB	7	GENPEPT		
INVERTED[c]		XPOUND	8	DPS vs	11	
(inverted repeats)		tRNA_SE	9	TREMBL		
TANDEM[c]				PROSITE	12	
(tandem repeats)				PROSCAN	13	

[a] RUMMAGE is a software package that incorporates the most sophisticated sequence annotation tools and compiles a homogeneous data set, the feature table, which can be transformed into a format of choice (GenBank, AceDB, HTML).
[b] A.F.A. Smit and P. Green, unpublished. REPEAT Masker is available at http://ftp.genome.washington.edu/cgi-bin/RepeatMasker
[c] Inverted repeat and Simple repeat search tools are in the AceDB package (version 4.3). R. Durbin and J. Thierry-Mieg, unpublished. Available at http://www.acedb.org/

To reach this goal we have developed the tool RUMMAGE which provides an automatic first-pass sequence annotation. RUMMAGE runs a set of diverse analysis tools (Table 1) on a given input sequence in FASTA format, parses and filters the result files and summarizes the findings in a standardized tabular format, the 'feature table'. This table forms a standard interface for several conversion tools and can be translated automatically into ASN, GenBank, AceDB and HTML format. The latter is especially convenient and allows scientists and sequencing centres to present their annotation results in tabular and graphical form on the Internet.

FIGURE 1. Internet presentation of RUMMAGE results



http://gen100.imb-jena.de/rummage/internal/21q/DSCR2-IGSF5/work/www_0/DSCR2-I(

Results of automatic first pass sequence annotation of DSCR2-IGSF5.cons.tab project

WrapperShell.pl v1.02, 00/05/29 15:11:25

Table of Contents:

- Graphics
  - Clickable Map
- Raw Tables
  - DSCR2-IGSF5.00_mask
  - DSCR2-IGSF5.00_blas
  - DSCR2-IGSF5.00_blas
  - DSCR2-IGSF5.00_blas
  - DSCR2-IGSF5.00_blas
  - DSCR2-IGSF5.00_blas
  - DSCR2-IGSF5.00_blas
  - DSCR2-IGSF5.00_grai
  - DSCR2-IGSF5.00_gens
  - DSCR2-IGSF5.00_blas
  - DSCR2-IGSF5.00_blas
  - DSCR2-IGSF5.00_blas
  - DSCR2-IGSF5.00_exon
  - DSCR2-IGSF5.00_exon
  - DSCR2-IGSF5.00_exon
- Formatted Tables
  - CONSENSUS_PseudoCon
  - LCP_GC_High
  - LCP_GC_Long
  - LCP_CpG_High
  - LCP_CpG_Long
  - CPG_CpG_High
  - CPG_CpG_Long

Results of automatic first pass sequence annotation of DSCR2-IGSF5 project

WrapperShell.pl v1.02, 00/05/29 15:11:25

Dynamic Map

Map View of Annotation Results:

Project DSCR2-IGSF5 (1-620000 of

COMPILE_ALU_Default
COMPILE_NONALU_Default
PROSCAN_PROM_Default
GENSCAN_PROM_Human
GENSCAN_EXON_Human
GRAIL_EXON_Human
MZEF_EXON_Human
XPOUND_EXON_Default
COMPILE_EXON_Weak
COMPILE_EXON_Strong
ExonSampler_EXON_CDS
GENSCAN_POLYA_Human
DPS_GENE_Default
GENSCAN_GENE_Human
ExonSampler_GENE_CDS
COMPILE_PRODUCT_Default
BLASTN_HOMO_GenEMBL
BLASTN_HOMO_STS
BLASTN_HOMO_EST
BLASTN_HOMO_CDS
BLASTX_HOMO_GenPept
BLASTN_EXON_Human
BLASTN_EXON_EST
BLASTN_EXON_GenEMBLnew
PROSITE_MOTIF_Default
LCP_GC_High
LCP_GC_Long
LCP_CpG_High

Results of automatic first pass sequence annotation of DSCR2-IGSF5 project

WrapperShell.pl v1.02, 00/05/29 15:11:25

```
Score = 364 (128.1 bits), Expect = 6.3e-99, Sum P(6) = 6.3e-99
Identities = 71/86 (82%), Positives = 75/86 (87%), Frame = -1

Query:  7036 VFGSCPRKNMQITILTCRHVTDYKTSESTGSLPSPFLRALKTQNFKDSACCPLLEQPNIV 6857
             VFGSCPRKNMQITILTCRHVTDYKTSESTGSLPSPFLRALKTQNFKDSACCPLLEQPNIV
```

trends in Genetics

The RUMMAGE results can be viewed with any conventional Internet browser on any computer with Internet access. For convenience, the different types of information are displayed in separate frames.

As visualization of the RUMMAGE annotation results is based on standard HTML, the results can be viewed with any conventional Internet browser on any conventional computer system that is connected to the Internet. For convenience, the browser window is split into three frames (Fig. 1). The 'navigation frame' shows the table of contents, providing direct access to the graphical map and to each underlying table. The 'main frame' shows the visualized data or a user-selected table. The data are presented as a dynamic map, which includes convenient features such as zooming in on a desired position or fading out

Stefan Taudien*
stau@imb-jena.de

Andreas Rump*
arump@imb-jena.de

Matthias Platzer*
mplatzer@imb-jena.de

Bernd Drescher*
drescher@mwgdna.com

Ruben Schattevoy*
schattev@biochem.mpg.de

Gernot Gloeckner*
gernot@imb-jena.de

Monika Dette*
dette@imb-jena.de

Cornelia Baumgart*
baumgart@imb-jena.de

Jacqueline Weber*
jawe@mwgdna.com

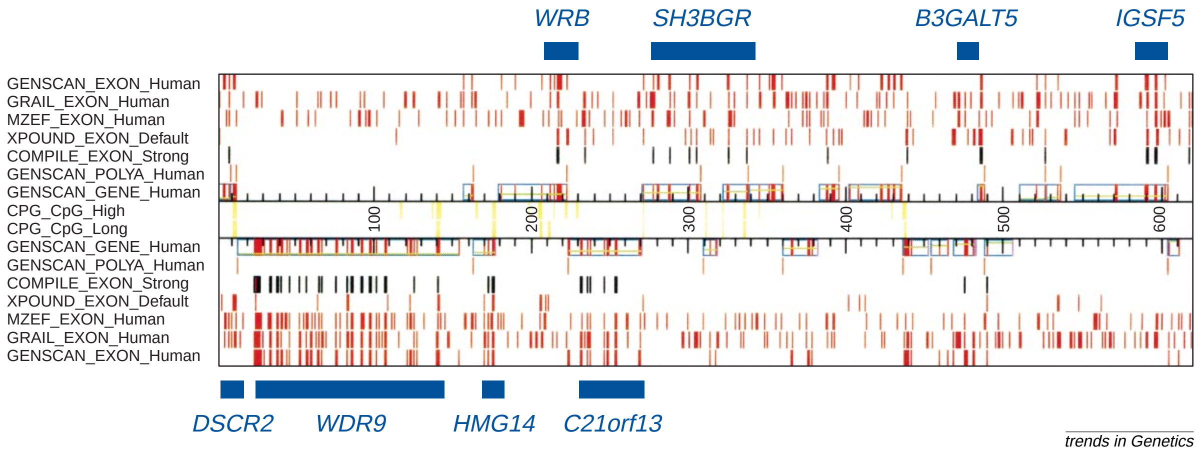
Uwe Menzel*
menzel@imb-jena.de

André Rosenthal*‡
Andre.Rosenthal@MetaGen.de

*Institute of Molecular Biotechnology, Department of Genome Analysis, Beutenbergstrasse 11, D-07743 Jena, Germany.
‡Friedrich Schiller University Jena, Am Fürstengraben 1, D-07743 Jena, Germany.

0168-9525/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(00)02127-2

FIGURE 2. Graphical output of RUMMAGE



Each row represents the hits of one single program. The number of kilobases are shown in the centre. Because of the clustering of hits, the presence of a gene can be detected easily. For each hit, detailed information is available by a mouse click. Furthermore, the user can zoom in on any desired position and fade in further results. Gene names and blue bars are not shown by RUMMAGE and were added to emphasize the clusters of hits.

unnecessary results. The third frame, the 'info frame', displays detailed information on any desired object, selected by a mouse click. Clicking on 'align' within the RUMMAGE output table, for example, displays the actual alignment of any selected genomic region with the corresponding expressed sequence tag (EST) or cDNA. Of course, all other types of alignments and database matches can be viewed within this info frame.

The performance of the exon prediction programs integrated in RUMMAGE is exemplified by a 620-kb contiguous sequence from human chromosome 21 (Ref. 1), containing eight genes with 84 exons. As the graphical output of RUMMAGE in Fig. 2 shows, none of the genes was missed and 90% of the total number of exons have been identified (Table 2). Two out of the eight genes were predicted completely, although a few exons were missed in the remaining six genes. Two of the eight genes were unknown until this analysis.

Any exon prediction program produces false-positive results, which cannot be recognized easily when only a single program is used. However, the integration of several exon prediction tools in RUMMAGE leads to an easy discrimination of false-positive predictions, as these tend not to be clustered; that is, the probability of several programs predicting the same false-positive exon is extremely low. Conversely, true exons tend to be identified by three or more prediction programs (Table 2). This leads to a clustering of hits, which can be recognized on the graphical output at first glance, so the user gets a very quick impression of the gene content within a given region of DNA.

There is no principal limit concerning the length of the sequence to be annotated. The constraints that are set by some of the integrated programs are overcome by splitting the sequence into pieces of appropriate size and fusion of the resulting data after annotation of all pieces is completed. Of course, total calculation time increases with sequence length. Given the current set of integrated programs and the currently available computing power of our UNIX cluster, RUMMAGE analyses approximately 20 kb h^{-1}.

RUMMAGE is very flexible with respect to hardware requirements and the kind and number of the integrated programs. As to hardware, RUMMAGE currently runs on a cluster of four Sun workstations and ten PCs, with LINUX as operating system. As RUMMAGE makes intensive use of the commercial local sharing facility (LSF), integration of additional computers can be done easily if more processing power is needed. As to software, the integrated programs are updated regularly and new analysis tools are integrated as soon as they are available. Each program is triggered by certain parameters, like stringency values or cut-off limits. The default values of these parameters can be viewed within the RUMMAGE output by following the link 'Annotation parameters'. On request, these values can be changed and adapted to the needs of the individual user. Currently, RUMMAGE is adapted to annotate genomic DNA from three species: man, mouse and *Fugu*. Adaptation of RUMMAGE to other organisms is in progress.

Since the RUMMAGE annotation service has been available on the Internet more than 40 users from all over the world have processed approximately 180 jobs on this system, encompassing around 60 MB of genomic DNA. Encouraged by the predominantly

TABLE 2. Performance of RUMMAGE[a]

Gene	Actual number of exons	Identified exons		
		Total	Using 3 or 4 programs	Using 1 or 2 programs
DSCR2	7	6	0	6
WDR9	41	39	21	18
HMG14	6	6	4	2
WRB	5	4	2	2
C21orf13	10	10	4	6
SH3BGR	7	5	5	0
B3GALT5	3	1	1	0
IGSF5	5	4	3	1
Total	**84**	**75 (90%)**	**40 (48%)**	**35 (42%)**

[a] RUMMAGE exon-prediction performance as exemplified for a 620-kb region in 21q22.2–22.3. In total, 91% of the exons belonging to the eight genes located in this interval were identified by the combination of four exon prediction programs. Six of the genes (*DSCR2*, *WDR9*, *HMG14*, *WRB*, *SH3BGR*, *B3GALT5*) are known structures with 100% identity to described mRNAs. By contrast, *C21orf13* and *IGSF5* are novel genes. Exon predictions for *C21orf13* are supported by expressed sequence tag matches, whereas *IGSF5* probably would remain undetected without exon prediction on the genomic level because of lacking any significant similarities to coding sequences in the databases. Note that the programs predicted the exact number of exons for *HMG14* and *C21orf13*. Subsequent confirmation of the two novel genes by RT–PCR experiments have underlined the performance of the gene prediction approach by RUMMAGE.

positive feedback from these external users, we recommend our RUMMAGE annotation service to everyone who wants to get a quick and comprehensive overview of genomic sequence data.

The RUMMAGE Sequence Annotation Service is available at http://gen100.imb-jena.de/~baumgart/rummage/register.html. The URL leads to a registration form that has to be submitted before the first use. This is necessary to provide a user-specific password, which ensures confidential treatment of the sequence data and the corresponding annotation results. As soon as the password is assigned, each user may run as many jobs as desired.

References

1 Hattori, M. (2000) The DNA sequence of human chromosome 21. *Nature* 405, 311–319

2 Huang, X. (1994) GC rich region search tool. *Compu. Appl. Biosci.* 10, 219–225

3 Larsen, F. (1992) CpG islands as gene markers in the human genome. *Genomics* 13, 1095–1107

4 Burge, C. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94

5 Uberbacher, E.C. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. U. S. A.* 88, 11261–11265

6 Zhang, M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. U. S. A.* 94, 565–568

7 Solovyev, V.V. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.* 22, 5156–5163

8 Thomas, A. (1994) A probabilistic model for detecting coding regions in DNA sequences. *IMA J. Math. Appl. Med. Biol.* 11, 149–160

9 Lowe, T.M. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964

10 Altschul, S.F. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

11 Huang, X. (1996) Fast comparison of a DNA sequence with a protein sequence database. *Microb. Comp. Genomics* 1, 281–291

12 Hofmann, K. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27, 215–219

13 Prestridge, D.S. (1995) Predicting Pol II promoter sequences using transciption factor binding sites. *J. Mol. Biol.* 249, 923–932

GeneNest: automated generation and visualization of gene indices

Expressed sequence tags (ESTs), introduced by Adams *et al.* in 1991 (Ref. 1), are a rapidly growing resource for analysing genes. Although ESTs might be of low sequence quality they are useful for detecting new genes, determining the genomic structure of a gene (such as exon–intron boundaries and alternative splicing sites)[2] and for gene-expression studies[3].

Because EST sequence information is highly redundant, a single gene might be covered by thousands of ESTs, each representing different parts of that gene. There have been several attempts to simplify the analysis of specific genes by clustering sequences belonging to the same gene[4–6], resulting in, so-called, gene indices. Some commonly used gene indices are Unigene (Ref. 4) at the National Center for Biotechnology Information (NCBI), the Institute for Genomic Research (TIGR) gene indices[5] and STACK (Ref. 6) at the South African National Bioinformatics Institute (SANBI). The Unigene and TIGR gene indices differ mainly in the clustering strategy used and presentation of cluster-related information[7]. Clusters of TIGR gene indices are summarized by a database of consensus sequences each reflecting a single transcript. Additionally, the relative order of sequences within a cluster is sketched roughly. In contrast, sequences in Unigene are clustered less stringently, such that alternative splice variants fall into the same cluster. Sequences derived from the same clone also may be clustered, based on their annotation. The Web presentation of Unigene at NCBI has extensive links between clusters and related information, such as mapping data or protein homologies.

Generation of gene indices

We have developed GeneNest (http://www.dkfz.de/tbi/services/GeneNest/index), a software and database for automated generation and visualization of gene indices.

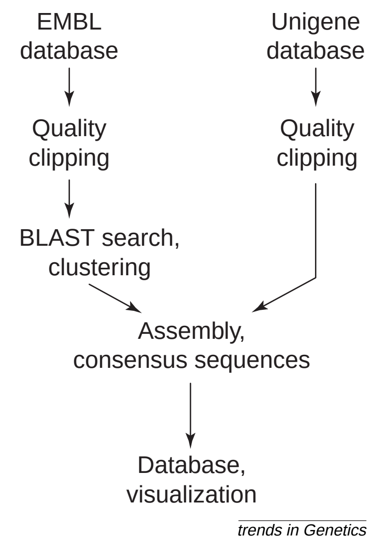
Generation of the GeneNest gene indices starts either with a database of sequences extracted from the EMBL database or from an already clustered database of ESTs from Unigene (Fig. 1). All sequences are subject to clipping, based on an extensive quality check. As a result of this step, repeats and vector sequences as well as low quality regions are masked. Similarities between these 'cleaned-up' sequences are then determined using BLAST (Ref. 8) and sequences are clustered if a near-perfect match extends over at least half of the shorter sequence. Sequences in a cluster are assembled in order to determine their relative positions and to obtain a representative consensus sequence. A cluster might be split into several contigs, each reflecting a group of sequences with global similarity. Such contigs are often caused by alternative splicing, ESTs derived from hnRNA or other artefacts such as chimeric sequences. In a final step, a Website presenting all these data is generated automatically.

In some projects, in particular for *Arabidopsis thaliana*, genomic sequences containing coding sequence annotations have been treated as potential genes. This strategy increases the number of sequences contributing to a gene index drastically, compared to Unigene or the TIGR gene indices, and, thus, also leads to an improved clustering.

Querying gene indices

The usefulness of a gene index depends strongly on its accessibility to the user. Most frequently, privately generated sequences are compared against the gene-index database. To this end,

FIGURE 1. Generation of GeneNest indices



GeneNest is a database and software package for producing and visualizing gene indices from expressed sequence tags. The processing steps involved in the automated generation of gene indices are shown.

Stefan A. Haas
s.haas@dkfz.de

Tim Beissbarth
t.beissbarth@dkfz.de

Eric Rivals
rivals@lirmm.fr

Antje Krause
a.krause@dkfz.de

Martin Vingron
m.vingron@dkfz.de

Deutsches Krebsforschungs-zentrum, Department of Theoretical Bioinformatics, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany.

0168-9525/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved. PII: S0168-9525(00)02116-8